



Data Sharing Ecosystems and the Creation of Value from Data

Matthew Wysel
BE (Aerospace) (Hons) MTechMgt *UNSW*

For the award of Doctor of Philosophy

University of New England
2023

This research has been conducted with the support of the
Australian Government Research Training Program Scholarship

Certification

I certify that the ideas, experimental work, results, analyses, software and conclusions reported in this thesis are entirely my own effort, except where otherwise acknowledged. I also certify that the work is original and has not been previously submitted for any other award, except where otherwise acknowledged.



Candidate

Date

Acknowledgements

Joanne, you have stood beside me from before either of us had any idea what we were doing, and you have not waived. Through existential crises and recurrent pandemics, home-schooling first our children and then our classes and then both, tea before breakfast and whiskey before a log fire, long bike rides, long COVID, and long delays; you have been a faithful friend, a sure shelter, and a rare treasure. For your willingness to stand up for me, your patience when standing in for me, and your grace when standing aside for me: thank you. Truly this endeavour would not have been possible without you.

Samuel Alexander, Annabella Florence, and Amélie Lucy, my beloved children. May you continue to grow in knowledge and love of all that is truly good. Grow up, but never old. Thank you for your boundless questions, your faithful enthusiasm, and bless you for your insistence to resolutely search out the truth of a matter. We thank God for you.

Professor Derek Baker, my primary supervisor and mentor. Your patient encouragement, enthusiasm for inquiry, and endless belief has sustained me. Under different tutelage I would have conceded a very long time ago. You took the ingredients of intellect, enthusiasm and analytical diligence and have spent these years growing scholarship. It continues to be a privilege to work alongside you.

Associate Professor William Billingsley, your supervision, intervention during conceptual development, particularly around Shannon Entropy, and your diligence in refining my expression for an IS audience has been of considerable value. Emeritus Professor Robert Banks, your boundless enthusiasm for abstraction and relentless pursuit of what Lewin would call *good theory* has been a challenge and a joy. Thank you for your *insight into data* and for joining us as co-author in Chapter 6.

Thank you to the *Agricultural Business Research Institute* (ABRI) for your support of this research through the Arthur Rickards Scholarship for Innovation in Agribusiness. I also want to acknowledge the late Arthur Rickards who, over what was to be our last coffee, listened patiently as I outlined our intent to elucidate a pre-commercial explanation for 'why data has value', before asking just one question: "How are you going to commercialise that?". Your legacy of delivering research that makes a lasting difference in the lives of non-researchers has shaped our scholarship deeply.

David and Lynda Wysel, my parents and ever faithful grandparents to our children. You have been our companions throughout this journey, patiently editing and relentlessly baking. It continues to be our family's joy to live with you in such proximity for this season of our lives. Allan, Narelle, Roslyn, Patrick and Naomi thank you for offering the irrational encouragement that only family can provide. Ben and Ksenia, Matthew and Katrina, Andrew and Miriam, Michael and Kate, Xavier and Louise, Richard, Matthew, Stephen, Simon, James and Joanne, Paul and Faye; thank you for your constancy.

MAW.

Dedication

*to my father who has championed the pursuit
of education, no matter the trials, from before we were born.
This is for you.*

Preface

This thesis was motivated by two very different sentiments. The first was personal. I sought a task of sufficient significance that I might be preoccupied with it for a number of years, and crucially, that would keep me from work-based travel while my family was young. I sought flexibility in the composition of my day so that I could be present enough to hear tall tales from the school yard, learn to cook, and kiss each child at night. The second was strictly commercial – if also altruistic. When working in industry, I longed for an informed, rigorous alternative to the narrative that permeated the corporate world: that the value of your data was exclusively defined by what you could sell it for. I sensed such a superficial philosophy would enslave data producers who could not conceive of markets for their biological, social, or corporate data, would enable rogue behaviour by those operating analytic systems, and would eventually lead to widespread market failure – particularly in knowledge markets.

My background in aerospace engineering and technology management, as well as my practical experience in both blue-chip and startup environments, led me to recognize the potential of technology-enabled value creation. With this focus, I sought an understanding of how data-sharing platforms could be more than just tools for powering business models, but also instruments for empowering communities to make valuable decisions. I hoped that taking a step back from corporate life would permit me to orchestrate an environment where I could plunge into what I hoped was approachable theory, while being present at the school gate and in the evenings.

Fortunately, we found quick success. I could see the value of data lay not in its mere existence, but in its extensibility, which in turn was enabled by the components of these data platforms. To that end our initial goal was to discover, elucidate and chart before progressing to explicate, synthesize and apply. Like a pilgrimage, the difficulty lay in the task of extending the ideal into the everyday. Plumbing the seminal texts from economics of information to communication theory, platform economics to data management, and finding a voice in the emergence of datanomics took much longer than anticipated and required a level of analysis that seemed, at least initially, incongruous to home life.

The pandemic was the curse that brought the blessing. Home schooling and no mental space for concentration twinned with omnipresent data disclosure, their omniscient behaviour models, and policed by omnipotent systems. I had become trapped in the front room of my data sharing ecosystem, with only theory to save me.

To that end, I hope this thesis is ‘good theory’ (Lewin, 1943) in that it offers an explanation that is both rigorous and straightforward, academic and practical, where abstraction is applied only to the point where the mechanics become visible. As I penned in a graduate assignment some six months before commencing the PhD,

“With widespread good governance, Instantaneous Big Data [that is, data ecosystems] could – theoretically – usher in an era of the responsible, sustainable rise of billions of people from disconnection and isolation into a global, interconnected assembly of bourgeoisie communities. Without governance, Instantaneous Big Data will be the means by which the proletariat will be organised for the greater efficiency of a new aristocracy whose nobility is the data they control, and the middle-class will be economically harnessed, employed as non-current assets and depreciated accordingly.”

Care for my family and concern for our digital selves motivated this research. I am thankful that the path we have begun is fulfilling both mandates.

Table of Contents

CERTIFICATION	II
ACKNOWLEDGEMENTS	III
DEDICATION	V
PREFACE	VII
TABLE OF CONTENTS.....	IX
GLOSSARY OF TERMS AND PRIMARY FIGURES.....	XI
ABSTRACT.....	1
CHAPTER ONE: INTRODUCTION TO THE RESEARCH PROJECT	3
CHAPTER TWO: DATA SHARING PLATFORMS: HOW VALUE IS CREATED FROM AGRICULTURAL DATA.	13
OVERVIEW OF MANUSCRIPT	13
FULL MANUSCRIPT	16
STATEMENT OF AUTHORS' CONTRIBUTION.....	48
STATEMENT OF ORIGINALITY	49
CHAPTER THREE: HOW TO VALUE DATA: UNITING ECONOMIC AND INFORMATION THEORY TO CREATE A VALUE FRAMEWORK FOR DATA.	51
OVERVIEW OF MANUSCRIPT	51
FULL MANUSCRIPT	53
STATEMENT OF AUTHORS' CONTRIBUTION.....	95
STATEMENT OF ORIGINALITY	96
CHAPTER FOUR: PROFITING FROM DATA. HOW DATA ENABLES FIRMS TO HAVE THEIR CAKE, SELL IT, AND EAT IT TOO.....	97
OVERVIEW OF MANUSCRIPT	97
FULL MANUSCRIPT	100
STATEMENT OF AUTHORS' CONTRIBUTION.....	140
STATEMENT OF ORIGINALITY	141
CHAPTER FIVE: TAKE MY DATA... PLEASE. HOW DATA SHARING ECOSYSTEMS MAKE OVERSHARING RATIONAL.....	143
OVERVIEW OF MANUSCRIPT	143
FULL MANUSCRIPT	146

STATEMENT OF AUTHORS' CONTRIBUTION.....	186
STATEMENT OF ORIGINALITY	187
CHAPTER SIX: AGTECH, AGRICULTURAL DATA AND MARKET FAILURE. AVOIDING A TRAGEDY OF THE (DATA) COMMONS.	188
OVERVIEW OF MANUSCRIPT	188
FULL MANUSCRIPT	191
STATEMENT OF AUTHORS' CONTRIBUTION.....	215
STATEMENT OF ORIGINALITY	216
CONCLUSION	217

Glossary of Terms and Primary Figures

Master Glossary

Term	Definition
Agent	<i>Noun.</i> An individual decision-making body such as a person, firm, or computer system.
Community	<i>Noun.</i> A collection of agents identified by a congruous – but not necessarily shared – data-related goal. A community of agents which might consist of buyers and sellers of cattle in an online marketplace, both of whom desire data on cattle but for ostensibly opposite reasons.
Data	<i>Noun.</i> An observation of uncertain relevance. Data is the core asset that permits and facilitates the creation of value in data sharing ecosystems. Data can be treated as a resource, good, or currency.
Data Development	<i>Noun.</i> Equivalent to Enrichment.
Data Sharing Ecosystem	<i>Noun.</i> A community of stakeholders who share a common data-related goal, and data collected from and for that community. Includes a system that uses the data to enable and incentivize stakeholders to make valuable interactions. <i>Similar:</i> Data Sharing Platform
Data Sharing Platform	<i>Noun.</i> A specific instance of a data sharing ecosystem that operates a platform business model.
Enrichment	<i>Verb.</i> The enrichment of data is the systematic reduction of the uncertainty in data. Enrichment arises from the coordinated efforts of a system and community of agents and permits the valuation of data against the community's established goals. <i>Similar:</i> Data Development (Chapter 2)
Information	<i>Noun.</i> Strictly equivalent to Insight.
Insight	<i>Noun.</i> Enriched data. That is, data that has become at least partially relevant and therefore valuable for an agent. NB. The term <i>insight</i> has been used throughout rather than information to avoid confusion caused by the conflation of the two terms in popular media. <i>Equivalent:</i> Information
Platform	<i>Noun.</i> A socio-economic arrangement designed to create a market between two, notionally disparate, communities commonly with the

	help of technology. Agents on either side of the market participate voluntarily, that is, outside the control of the platform owner.
Stakeholder	<i>Noun.</i> An agent who is a member of a collective, such as a Data Sharing Ecosystem, community, club, etc... and carries some form of vested interest in the activities of that collective. <i>Similar:</i> Agent, Member, Community
System	<i>Noun.</i> An asset in a <i>Data Sharing Ecosystem</i> . The system presides over the data and continually develops the data towards the loci of goals specified by the community. <i>Similar:</i> Analyst (Chapter 5), Laboratory (Chapter 6)

Primary Figures

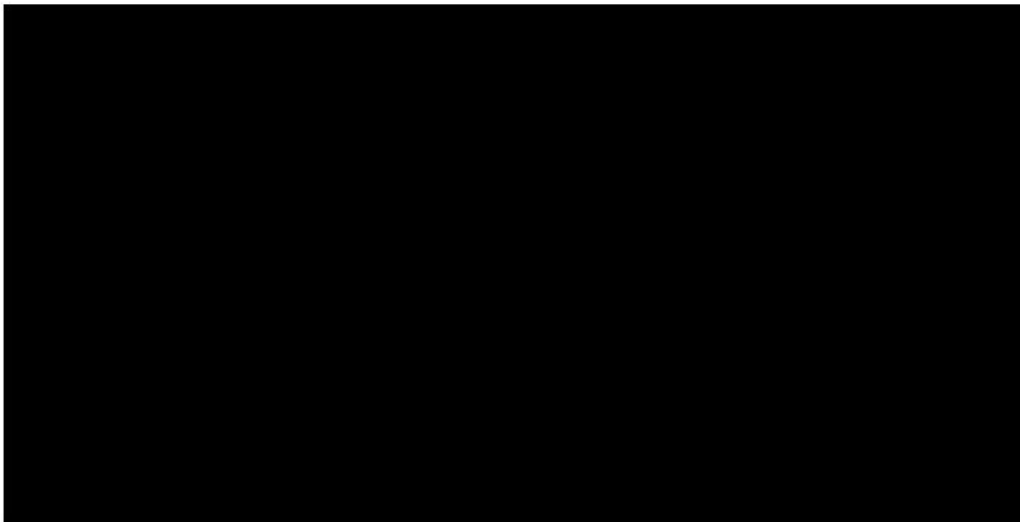


Figure 2-3. Three-layer, Data Sharing Ecosystem illustrating components necessary for the creation of value from data and their abstracted arrangement.

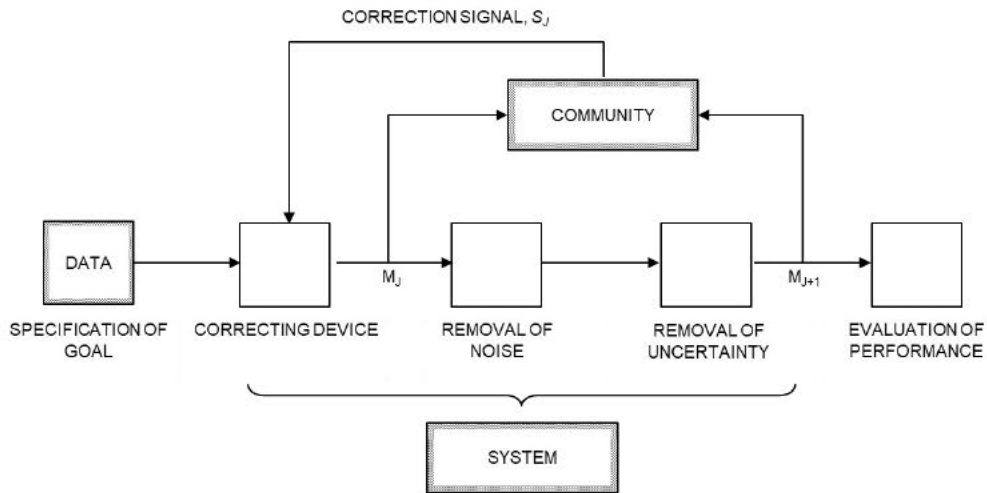


Figure 3-3. Data enrichment process – styled to reflect Shannon’s diagram of a correction system (1948, p. 22).

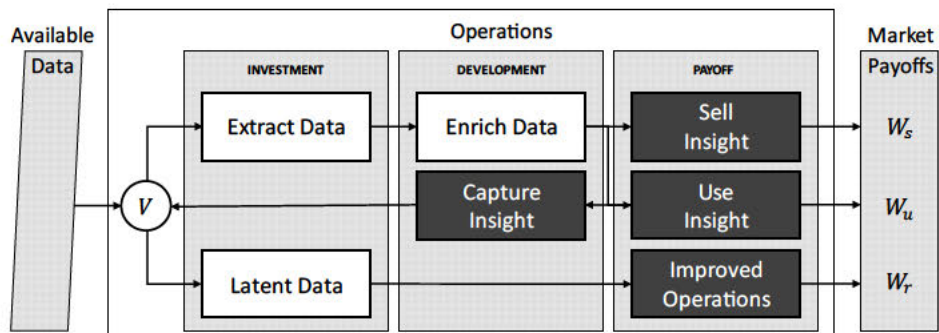


Figure 4-4. Abstraction of a data sharing platform as a firm-level, data-based production process.

Note the firm fulfills the functions of both the system and community. The firm fulfills the role of the Community when it contributes data and decides on the proportion and frequency of development (INVESTMENT, above). The firm also enriches data (DEVELOPMENT, above) and decides on allocation of value between various internal and external payoffs (dark boxes, above).

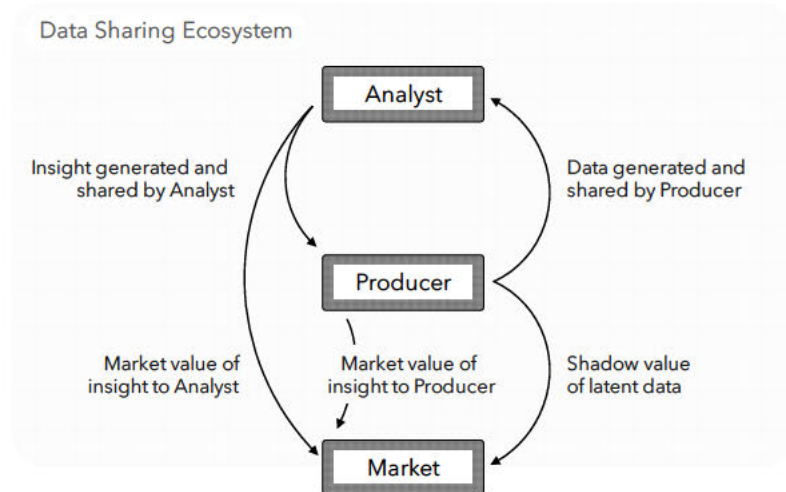


Figure 5-1. Simplified Data Sharing Ecosystem indicating how agents in a market collaborate to create value from data. The Community (PRODUCER, above) decide how much data to share with the System (ANALYST, above). Both create value from data directly and indirectly through the MARKET.

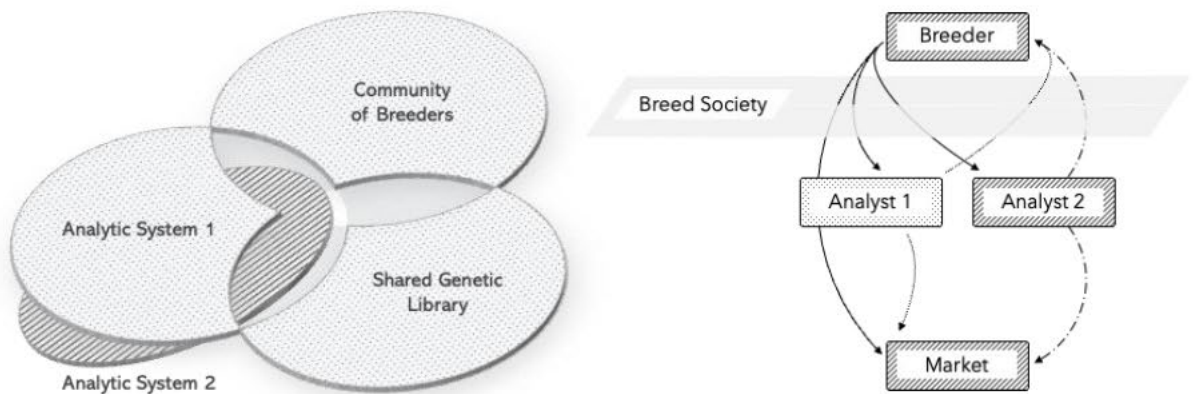


Figure 6-1. Breed societies as a data sharing ecosystem with competing analytic systems and unilateral profit maximization by breeders. (a) ecosystem view, and (b) data-flow view.

Abstract

This work centers on the mechanisms by which value is accrued to data, a field that is becoming known as *datanomics*. The research is mobilized from platform economics, the economics of information, Big Data, communication theory, the study of data as a production factor and an exchangeable service, and the emerging field of information chain failure (a vertical form of market failure). This thesis makes the case that the creation of value from data requires a broader consideration of activities and resources than has been researched previously. Essentially, value creation relies on capitalization of enrichment infrastructure, and recognition of a community of internal and external stakeholders who collaborate around the data. At issue is the fundamental question of how each stakeholder can operate around data to achieve maximum advantage, and how collaboration, not competition, is required to achieve those outcomes.

The empirical case used is of animal (specifically cattle) breeders who periodically collect performance data and then must decide both how much of it to share and how frequently to do so. The emergent model characterizes optimal strategies at firm and industry level. This novel approach provides guidance for industry and policy action. For industry, it identifies optimal data sharing actions and contextual conditions. At the policy level, it identifies and quantifies market failure, and the steps necessary to correct it.

This PhD is presented as a thesis-by-publication. The first two chapters draw from the literatures noted above to derive the ingredients required to create value from data, the attributes of the value creation process and the mechanics of its operation. The first of these chapters was published in an A* agricultural systems journal in 2021. The final three chapters apply this theory at a firm-level, across a market, and into a microeconomy, respectively, and have been through the full revision process at a management journal (A*), information systems journal (A*) and agricultural technology journal (Q1).

The primary author and student has presented his research at three annual Conferences of the Australian Agricultural and Resource Economics Society, the annual conference of the International Food and Agribusiness Management Association, and the International Congress on Modelling and Simulation. The author won first place in the Economics and Business section of UNE's 3-Minute Thesis competition and has been influential in UNE's business incubation activities that target knowledge management aspects of *Agtech* adoption.

Chapter One:

Introduction to the Research Project

1. Motivation

Organizations are awash with data but generally at a loss with how to value it (Grover et al., 2018). A recent survey of 36 companies and non-profit organizations across North America and Europe, most of whom had turnovers greater than USD1 billion, revealed most had no formal data valuation practices and any existing valuation efforts were time-consuming and complex (Short & Todd, 2017). Out of necessity, data management was focused on “storing, protecting, accessing, and analyzing [their] massive amounts of data”, all while firm data continued to grow at an average of 40% per annum. The 2018 NIST whitepaper on big data reported the growth rate of data generated and stored outpaced Moore’s Law for the growth rate of computation (Chang et al., 2018) and this capability gap has continued to broaden since then as organizations shift to real-time analysis of data (Stieglitz et al., 2018). These challenges become even more pressing when operational requirements are applied to big data (Chiang et al., 2018; Ketter et al., 2015). If ever the practice of valuing data like a standard accounting asset¹ was tenable, it is quickly becoming operationally and technically infeasible (Short & Todd, 2017).

Organizations require a method to prioritize what data to process and what data to set aside (Grover et al., 2018). An understanding of how value is created around data permits an organization to allocate scarce resources towards the generation, development, and commercialization of data (Fleckenstein et al., 2023) and enables managers to derive strategies that can target specific characteristics of the value creation process. Beyond the firm, definition of this process permits increases in the efficiency of data markets and new opportunities for cross-market, data-based partnerships (Windasari et al., 2021).

The problem of valuing data is not limited to organizations and markets. Houses, cars, webpages, and social connections create ecosystems of value (Hukal et al., 2020) as data is generated and shared with external systems (Cichy et al., 2021). Individuals trade access to their raw social, health,

¹ Record, categorize, summarize, and report.

biological or behavioral data to access subsidized services as personal devices: *wearables*, as one example, share biologic data for short-term outcomes such as entertainment or convenience, and longer-term benefits like improved health outcomes (Bardhan et al., 2020). Evidently this data constitutes a valuable exchange; these data streams offset real-world costs – usually by a proprietary value generation process (Langley & Leyshon, 2017). Indeed, data on customer’s preferences enables matchmaking platforms (Evans & Schmalensee, 2016) to amass considerable economic value even while controlling relatively few assets other than the intermediating processes and associated data (Parker, Van Alstyne, & Choudary, 2016). Data ceded by customers enables whole ecosystems whose byproduct, and often: goal, is the transformation of data from a ‘raw’ state to more valuable insight. Yet, the portions of this value that could be considered consumer surplus, producer surplus, or indeed latent economic value retained remain unapproached.

Further compounding this phenomenon of ‘drowning in data’, some industries have also experienced the rapid proliferation of data-producing sensors. This is particularly true in agriculture where the rapid adoption of Smart Farming has extended the science of Precision Agriculture by flooding the sector with data. Installation of smart machines and connected sensors within and between farms have expanded both the scope and scale of data-based interventions available to farm management. These devices digitize farms and create new opportunities for agricultural value creation through the data-driven “optimization of agricultural production systems, value chains and food systems” (Klerkx, Jakku, & Labarthe, 2019, p. 2). Semi-autonomous machinery and the incorporation of Internet of Things (IoT) devices across farms produce constantly evolving data flows (Friess, 2016) that threaten to overwhelm agribusiness managers. At an infrastructural level, the ongoing shift to off-farm, cloud-based processing and decision support tools continually increases the demands placed on farm infrastructure as data flows become critical to standard agricultural operations. Even established business models are under threat. Data generated from Smart Farming permits new operational models as traditional agricultural problems are shared with non-agricultural communities in a continual process of open innovation powered by a common belief in the value of shared data. Where value creation in agriculture was once inseparable from labor- and capital-intensive activities, value creation in Smart Farming is now unavoidably data intensive.

2. The Research Problem

What scholars, professionals and consumers need is a straightforward, rigorous framework for valuing data that is predicated on the fundamental characteristics of data that affect its value. Evidently, there exists a method for deriving value from data. However, if the principles surrounding

this method were well understood, as scholars, we ought to be able to answer fundamental questions about its nature and generic operation. What constituents must be present for data to possess value? Which characteristics of data dictate its ordinal value, and by what process is the value of data altered? Scholars and professionals ought to be able to identify the generic modes by which data may be valued within an organization and across a market. This thesis proposes timely responses to each of these questions.

Additionally, as scholars we require a theory on data management that unites those attributes of data that determine value creation, allocation, and capture with the mechanics of creating value from data (Grover et al., 2018). This endeavor requires more than simply coalescing different approaches to data (Fleckenstein et al., 2023). A generic production process that describes how value is created from agricultural data must reconcile theory from across production, innovation, and business management literature and integrate that literature within agricultural economics and information systems. For instance, under what conditions might a firm rationally refrain from further investment in data (for example, Sims (2003), Cichy et al. (2021)) even though increases in data notionally accompany increases in productive output (Gawer, 2022)? Or how might the intrinsic properties of data – such as excludability and non-rivalry – affect the specification of data as a factor of production within a firm (Pentland et al., 2021)? Indeed, treatment of data as an intangible asset has precedent (Farboodi et al., 2019) but how might conditional (Clough & Wu, 2022) data network effects (Gregory et al., 2021) be leveraged so a firm can both accumulate value from data *and* distribute value among stakeholders (Cusumano et al., 2019; Wolfert et al., 2017)? More broadly, such a theory ought to serve as *programmatic theory*, strengthening connections between deduced causalities and the real-world where these relationships may be empirically evaluated (Cronin et al., 2021). Without a theory that describes the transformation of data into value within a firm, scholars retreat to explanations for data-centric phenomena that are suitable for description, but divorced from powers of explanation, prediction, or prescription (Makadok et al., 2018).

The transformation of data into value within a firm represents a growing tension in the field of data management as scholarly theory becomes increasingly separated from the experience of practitioners (Pentland et al., 2021). Firms such as *23andMe* transform data into value as customers provide genetic data in return for insight on specific diseases. Their production process ingests data and applies proprietary technology to develop data for an agreed payoff. This process is mirrored in platform businesses. *Uber* and *LinkedIn* transform data into value through a process where behavioral data enables insight that is sold in ancillary markets or used by the firm to improve its connection of customers. In parallel, scholarly treatment of data as a resource and the management of data within a firm continues to advance. The former focuses on managing data with implications

for the firm (see for example, Hagiwara and Wright (2020)) while the latter improves management of the firm with implications for data (for example, Wysel et al. (2021)). However, mirroring the experience of scholars, leaders of even large, well-resourced firms struggle to adequately manage data, noting the absence of a united framework that articulates the production process that surrounds their data and metadata in terms of resource allocation and management interventions (Short & Todd, 2017). Both parties need a single, datanomic theory that unites data-based value production with the mechanics of data management and inform real-world data valuations (Bresciani et al., 2021), data sharing (Darnell et al., 2018), and value allocation-versus-capture (Clough & Wu, 2022; Zhang et al., 2020). Without this, the very practical theory of transforming data into value becomes like a map uncharted by explorers and unplumbed by cartographers, increasingly distant from those whose practice ought to inform it and studies ought to shape it.

3. The Research Question

The purpose of this thesis is to propose a unifying theory for the creation of value from data first by an individual agent, such as a firm, and second by groups of agents who engage across a market. The goal of this thesis is to offer a scholarly and practical explanation for how value is created from data.

Accordingly, the central research question is:

*What are the components of the economic value of data and
how do those components work together?*

This research question proposes that there is a value creation process mediated by the assembly and management of the components of a system. The research question deliberately targets an economic understanding rather than the decision-making mechanics that inform business processes, to permit analysis of both the attributes of data and the emergent commercialization models. Specifically, this thesis applies this question to the Australian Agricultural Industry. While we maintain a recurring focus on livestock breeders, the societies that represent their short- and longer-term interests, and the markets that service their creation of value from data, we develop models that are also applicable more broadly. Notwithstanding this focus, this thesis seeks to answer the research question universally seeking to apply the economic explanation at firm level, within a marketplace, and across a micro-economy.

The objectives of this research are to:

- Identify the component elements of the value creation process that surrounds data and specify how those components interact.

- Explain the mechanics of the process that brings these components together and the economic drivers of this process.
- Explain how the process works within a firm and across a marketplace.
- Evaluate what happens to value creation when one component of the process is impaired.

4. Methodology and Structure of this Thesis

This thesis uses a combination of theoretical and empirical analysis to explain how the creation of value from data is given by the operation of a data sharing platform, and that by managing the operation of the platform an agent can manage the value created from data under their control. The theoretical analysis synthesizes theory from platform economics (Parker et al., 2017), the economics of information (Arrow, 1996), production (Zellweger & Zenger, 2021) and innovation (Gomes et al., 2018) literature as well agricultural data (Wiseman et al., 2019; Wolfert et al., 2017) and the data-enabled value management (Jakku et al., 2016; Klerkx et al., 2019) that agricultural data creates.

The empirical analysis is based on a national survey of livestock breeders conducted by the Australian Meat and Livestock Association (MLA) (Banks, 2019) and addresses the data trading ecosystem that they create by sharing genetic data with industry analysts such as the Agricultural Business Research Institute (ABRI). Chapter 6 deals with the highly sensitive topic of Breed Societies' performance and the pending failure of the genetic market they preside over due to the introduction of new value-creating technologies. As breed societies are understandably reluctant to share this type of performance data, we engaged an industry expert to generate a representative dataset that covers breeder's costs, and their data-based benefits under both the nominal operating conditions and when using the new technology.

This thesis is presented as a thesis-by-publication. Accordingly, each chapter has been written to stand alone as a publishable paper and to build a research project that addresses the stated research objectives. Chapter 2 frames the analysis by bringing together the results of Wolfert et al. (2017)'s systematic literature review of Big Data in Smart Farming and platform economic literature. The latter deals with the composition (Choudary, 2015), operation (Gawer, 2022), and management (Hukal et al., 2020; Van Alstyne et al., 2016) of *platforms*, the socioeconomic constructs that utilize data to create value for the agents who participate on them (Langley & Leyshon, 2017). Chapter 2 presents data sharing platforms as the more generalized version of *platform businesses* (Täuscher & Laudien, 2018) because the former maintains a focus on both data and value created, instead of focusing primarily on the resulting value. Chapter 2 initially develops the foundational components – or *ingredients* – of the process of creating value from data before applying this framework to

connect decision making with the second-order effects of value creation. The online livestock trading platform *AuctionsPlus.com.au* is used as a sustained exemplar.

Whereas Chapter 2 approaches the process of creating value from data from two opposite and bounding perspectives, Chapter 3 uses seminal work from communication theory (Shannon, 1948) and the economics of information (Arrow, 1996) to explain how data sharing platforms operate, and therefore, how value is created from data. This manuscript maintains the value equation from Chapter 2 and following Arrow's (1996) exposition of the economics of information, specifies a single economic expression for the operation of a data sharing platform. Shannon's (1948) communication theory is used to explain how interactions between the community and system cause a coordinated enrichment of data and accretion of value.

Chapter 4 applies the valuation modes from Chapter 3 to the decision-making framework from Chapter 2 to connect firm-level decision making with the *first-order* effects of creating value from data. Specifically, this chapter develops a firm-level, objective function for the production of value from data and asks, 'how much should a firm invest in data?' and, 'how often should the firm repeat that process?'. Deliberately, other than noting data may be purchased from outside the firm and value is created via payoffs received from the market, this paper considers the firm in relative isolation from its surrounding environment leaving the complexity associated with the trade of data to Chapter 5.

Chapter 5 applies the data sharing platform model to the results of the MLA survey, to analyze value created by the genetic data that breeders share with industry analysts such as the Agricultural Business Research Institute (ABRI). This chapter extends the data-based, production process developed in Chapter 4 by applying it to evaluate how efficiently breeders and the industry analyst trade data to create value. The analysis from Chapter 5 presents a 'system-of-systems' application of the models developed in Chapters 2 through 4.

A key tenet of this chapter is that if analysts in a data sharing ecosystem adopt a *service-dominant* logic (Vargo et al., 2008) towards data trading, they can incentivize producers to share data above their notional optimum, expanding the Pareto frontier of the ecosystem and inducing a (rational) overshare from producers. While this chapter presents an analysis of a specific genetic trading market, adoption of a service-dominant view of data trading also reveals Breed Societies to be facing a pending technology-enabled market failure in the supply of genetic data. This is the research question for Chapter 6.

Chapter 6 connects technological disruption with market failure for agricultural data markets. Theory developed by this thesis shows unilateral value maximization by one party has impaired the

value co-creation that characterized the ecosystem. If the Breed Society does not act, breeders will digitally over-graze their society's genetic libraries and, eventually, create a tragedy of their shared data commons.

This thesis deliberately views breed societies and their activities as a data sharing ecosystem that comprises a community of value-maximizing members, an analytic system, and data that is jointly stewarded by both parties. This perspective is predicated on those attributes of data that distinguish it from the normal assets of capital or labor; chiefly, non-rivalry (Jones & Tonetti, 2020), excludability (Angelopoulos et al., 2021; Easley et al., 2018; Jakku et al., 2019), and conditional (Clough & Wu, 2022) data network effects (Gregory et al., 2022; Hagiu & Wright, 2020).

5. Contribution

The contribution of the analysis portion of this thesis is threefold. First, an economic valuation model is developed that accounts for: intrinsic uncertainty in the relevance of data, the non-rivalrous nature of data, uncertainty in expected payoffs, and rivalry of resources consumed in the data's valuation. Second, we extend existing economics of information research (Frankel & Kamenica, 2019) by establishing uncertainty as the sole intrinsic characteristic of data that alters its value. We also define the data enrichment process as the means by which this uncertainty is reduced. Third, the data valuation process and data enrichment process are combined to create a framework useful for understanding, managing and attributing value between all parties involved in the valuation of data. The implications of these findings are an understanding of how data may be valued according to one of three modes: the valuation of data as a resource, as an economic good, and as a currency.²

This theory is then applied at a firm level, within a marketplace, and across a micro-economy. The paper that comprises Chapter 4 is the first time a production function has been mobilized to describe the creation of value from data within a firm. Significantly, this theory encompasses a very general set of use cases, and yet can easily be applied to answer very specific questions such as, 'how much should I invest in data?' or 'how often ought I repeat that process?'. To the best of our knowledge, the proposed theory is the first time such an expansive definition of data as a production factor is brought together with the very practical exercise of creating value from data within a firm.

² We do acknowledge that a resource and currency are strictly just different projections of an economic good. Nevertheless, the distinction has been adopted as each projection reveals different aspects of the value of data, and how this value may be appropriated in different contexts.

Value can also be created from data across a marketplace. Chapter 5 relates exchange value to non-rivalry and conditional data network effects and demonstrates that a service-dominant view of data enables an expanded Pareto frontier in data trading markets as agents invert the value creation process and begin to share value to create data. This chapter also includes extended discussion on human interactions with 'generative AI', conditions for vertical integration in data supply chains, reverse subsidies, and how to create quantitative data governance thresholds.

Finally, Chapter 6 applies Club Theory to illustrate how marketplace sponsors, such as Breed Societies and agricultural platforms, can pursue technological disruption while avoiding market failure and tragedy of their shared data commons. We discuss policy implications for management of national, or supra-industry, datasets and provide recommendations for ongoing, finer grained research.

Table 1-1 includes a summary of publication status of each chapter while derivative outputs for the research contained in each paper are listed at the start of each chapter.

Table 1-1. Summary of Publication Status for Each Chapter in this Dissertation

Chapter	Status
Chapter 2	Published in <i>Agricultural Systems</i> , July 2021.
Chapter 3	Unpublished manuscript; Received back from first-round review at <i>Management Information Systems Quarterly (MISQ)</i> on 14 November 2020. Decision was made to pause development of this manuscript.
Chapter 4	Reviewed at <i>Management Science</i>
Chapter 5	Reviewed at <i>Management Information Systems Quarterly (MISQ)</i>
Chapter 6	Reviewed at <i>Computers and Electronics in Agriculture</i>

References

- Angelopoulos, S., Brown, M., McAuley, D., Merali, Y., Mortier, R., & Price, D. (2021). Stewardship of Personal Data on Social Networking Sites. *International Journal of Information Management*, 56, 102208.
- Arrow, K. J. (1996). The Economics of Information: An Exposition. *Empirica*, 23(2), 119-128.
- Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting Systems, Data, and People: A Multidisciplinary Research Roadmap for Chronic Disease Management. *MIS Quarterly*, 44(1), 185-200.

- Bresciani, S., Ciampi, F., Meli, F., & Ferraris, A. (2021). Using Big Data for Co-Innovation Processes: Mapping the Field of Data-Driven Innovation, Proposing Theoretical Developments and Providing a Research Agenda. *International Journal of Information Management*, 60, 102347.
- Chang, W. L., Roy, A., Grady, N., Reinsch, R., Underwood, M., Fox, G., Boyd, D., & von Laszewski, G. (2018). *Nist Big Data Interoperability Framework*.
- Chiang, R. H., Grover, V., Liang, T.-P., & Zhang, D. (2018). Strategic Value of Big Data and Business Analytics. *Journal of Management Information Systems*.
- Choudary, S. P. (2015). *Platform Scale: How an Emerging Business Model Helps Startups Build Large Empires with Minimum Investment*. Platform Thinking Labs.
- Cichy, P., Salge, T. O., & Kohli, R. (2021). Privacy Concerns and Data Sharing in the Internet of Things: Mixed Methods Evidence from Connected Cars. *MIS Quarterly*, 45(4).
- Clough, D. R., & Wu, A. (2022). Artificial Intelligence, Data-Driven Learning, and the Decentralized Structure of Platform Ecosystems. *Academy of Management Review*(ja).
- Cronin, M. A., Stouten, J., & van Knippenberg, D. (2021). The Theory Crisis in Management Research: Solving the Right Problem. *Academy of Management Review*, 46(4), 667-683.
- Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2019). *The Business of Platforms: Strategy in the Age of Digital Competition, Innovation, and Power*. Harper Business New York.
- Darnell, R., Robertson, M., Brown, J., Moore, A., Barry, S., Bramley, R., Grundy, M., & George, A. (2018). The Current and Future State of Australian Agricultural Data. *Farm Policy Journal*, 15(1), 41-49.
- Easley, D., Huang, S., Yang, L., & Zhong, Z. (2018). The Economics of Data. Available at SSRN 3252870.
- Evans, D. S., & Schmalensee, R. (2016). *Matchmakers: The New Economics of Multisided Platforms*. Harvard Business Review Press.
- Farboodi, M., Mihet, R., Philippon, T., & Veldkamp, L. (2019). Big Data and Firm Dynamics. AEA papers and proceedings,
- Fleckenstein, M., Obaidi, A., & Tryfona, N. (2023). A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model.
- Frankel, A., & Kamenica, E. (2019). Quantifying Information and Uncertainty. *American Economic Review*, 109(10), 3650-3680.
- Friess, P. (2016). *Digitising the Industry-Internet of Things Connecting the Physical, Digital and Virtual Worlds*. River Publishers.
- Gawer, A. (2022). Digital Platforms and Ecosystems: Remarks on the Dominant Organizational Forms of the Digital Age. *Innovation*, 24(1), 110-124.
- Gomes, L. A. d. V., Facin, A. L. F., Salerno, M. S., & Ikenami, R. K. (2018). Unpacking the Innovation Ecosystem Construct: Evolution, Gaps and Trends. *Technological Forecasting and Social Change*, 136, 30-48. <https://doi.org/10.1016/j.techfore.2016.11.009>
- Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The Role of Artificial Intelligence and Data Network Effects for Creating User Value. *Academy of Management Review*, 46(3), 534-551.
- Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2022). Data Network Effects: Key Conditions, Shared Data, and the Data Value Duality. In *Academy of Management Review*.
- Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, 35(2), 388-423. <https://doi.org/10.1080/07421222.2018.1451951>
- Hagiu, A., & Wright, J. (2020). Data-Enabled Learning, Network Effects and Competitive Advantage. In *Unpublished*.
- Hukal, P., Henfridsson, O., Shaikh, M., & Parker, G. (2020). Platform Signaling for Generating Platform Content. *MIS Quarterly*, 44(3).

- Jakku, E., Taylor, B., Fleming, A., Mason, C., Fielke, S., Sounness, C., & Thorburn, P. (2019). "If They Don't Tell Us What They Do with It, Why Would We Trust Them?" Trust, Transparency and Benefit-Sharing in Smart Farming. *NJAS-Wageningen Journal of Life Sciences*, 90-91, 100285.
- Jakku, E., Taylor, B., Fleming, A., Mason, C., & Thorburn, P. (2016). Big Data, Trust and Collaboration.
- Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9), 2819-2858.
- Ketter, W., Peters, M., Collins, J., & Gupta, A. (2015). Competitive Benchmarking: An Is Research Approach to Address Wicked Problems with Big Data and Analytics. *MIS Quarterly*.
- Klerkx, L., Jakku, E., & Labarthe, P. (2019). A Review of Social Science on Digital Agriculture, Smart Farming and Agriculture 4.0: New Contributions and a Future Research Agenda. *NJAS-Wageningen Journal of Life Sciences*, 90-91, 100315.
- Langley, P., & Leyshon, A. (2017). Platform Capitalism: The Intermediation and Capitalisation of Digital Economic Circulation. *Finance and Society*, 3(1), 11-31.
- Makadok, R., Burton, R., & Barney, J. (2018). A Practical Guide for Making Theory Contributions in Strategic Management. In (Vol. 39, pp. 1530-1545): Wiley Online Library.
- Parker, G., Van Alstyne, M., & Jiang, X. (2017). Platform Ecosystems: How Developers Invert the Firm. *MIS Quarterly*, 41(1).
- Pentland, A., Lipton, A., & Hardjono, T. (2021). *Building the New Economy: Data as Capital*. MIT Press.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379-423.
- Short, J., & Todd, S. (2017). What's Your Data Worth? *MIT Sloan Management Review*, 58(3), 17.
- Sims, C. A. (2003). Implications of Rational Inattention. *Journal of Monetary Economics*, 50(3), 665-690.
- Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social Media Analytics—Challenges in Topic Discovery, Data Collection, and Data Preparation. *International Journal of Information Management*, 39, 156-168.
- Täuscher, K., & Laudien, S. M. (2018). Understanding Platform Business Models: A Mixed Methods Study of Marketplaces. *European Management Journal*, 36(3), 319-329.
- Van Alstyne, M., Parker, G., & Choudary, S. P. (2016). Pipelines, Platforms, and the New Rules of Strategy. *Harvard Business Review*, 94(4), 54-62.
- Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On Value and Value Co-Creation: A Service Systems and Service Logic Perspective. *European Management Journal*, 26(3), 145-152.
- Windasari, N. A., Lin, F.-r., & Kato-Lin, Y.-C. (2021). Continued Use of Wearable Fitness Technology: A Value Co-Creation Perspective. *International Journal of Information Management*, 57, 102292.
- Wiseman, L., Sanderson, J., Zhang, A., & Jakku, E. (2019). Farmers and Their Data: An Examination of Farmers' Reluctance to Share Their Data through the Lens of the Laws Impacting Smart Farming. *NJAS-Wageningen Journal of Life Sciences*, 90-91, 100301.
- Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big Data in Smart Farming—a Review. *Agricultural Systems*, 153, 69-80.
- Wysel, M., Baker, D., & Billingsley, W. (2021). Data Sharing Platforms: How Value Is Created from Agricultural Data. *Agricultural Systems*, 193, 103241.
- Zellweger, T. M., & Zenger, T. R. (2021). Entrepreneurs as Scientists: A Pragmatist Approach to Producing Value out of Uncertainty. *Academy of Management Review*(ja).
- Zhang, Y., Baker, D., & Griffith, G. (2020). Product Quality Information in Supply Chains: A Performance-Linked Conceptual Framework Applied to the Australian Red Meat Industry. *The International Journal of Logistics Management*, 31(3), 697-723.

Chapter Two: Data Sharing Platforms: How Value is Created from Agricultural Data.

Overview of Manuscript

Status:	Published in <i>Agricultural Systems</i>
Date Submitted:	17 February 2021
Date Published:	26 July 2021
Suggested Citation:	Wysel, M., Baker, D., & Billingsley, W. (2021). Data Sharing Platforms: How Value Is Created from Agricultural Data. <i>Agricultural Systems</i> , 193, 103241.

Summary of Paper in Context of PhD:

This chapter unites the concept of *the platform stack* (Choudary, 2015) with the results of Wolfert et al. (2017)'s systematic literature review into Big Data in Smart Farming to propose that the creation of value from data requires the operation of a Data Sharing Platform. The data sharing platform is a pivotal concept in this PhD.

The process of creating value from data is approached from two opposite perspectives. First, the paper develops the foundational components – or *ingredients* – of the process, arranging them in three distinct layers. In essence, operation of the process occurs as each component in the asset layer interacts, while efficacy of the process is determined by the components in the management layer. The paper applies this framework to connect firm-level decision making with second-order effects on value creation. The online livestock trading platform *AuctionsPlus.com.au* is used as a sustained exemplar.

The immediate theoretical need developed by this paper is an explanation of how these components are combined to produce value from data, and what implications this creates for the treatment of data as an economic good. Chapter 3 addresses those issues.

Apart from typesetting changes and language localization, this chapter appears exactly as published by *Agricultural Systems* on 17 February 2021.

Supplementary Publications

Research that informed this chapter also appeared in the following outputs.

Type	Citation
Conference Paper	Wysel, M., Baker, D., & Conway, L. (2019). Connecting Researchers with Customers through a Common Valuation of Data. 63rd Annual Conference of the Australasian Agricultural and Resource Economics Society, Melbourne, Australia.
Market Report	Wysel, M. (2018). <i>Big Data in Australian SMEs</i> .
Conference Paper	Wysel, M. (2019). Using Platforms to Operationalize the Valuation of Information in the Red Meat Supply Chain. <i>Information in the Australian Red Meat Supply Chain</i> 63rd Annual Conference of the Australasian Agricultural and Resource Economics Society, Melbourne, Australia.
Symposium Presentation	Wysel, M. (2019). Using Platforms to Operationalize the Value of Information in the Red Meat Supply Chain. Symposium: Information in the Australian Red Meat Supply Chain, UNE Centre for Agribusiness.
Conference Paper	Burgess, S., & Wysel, M. (2020). <i>China's Social Credit System: How Robust Is the Human Rights Critique?</i> 27th Annual Australian Association for Professional and Applied Ethics, The University of New England.
Symposium Presentation	Baker, D., Cook, S., Jackson, E. L., Wysel, M., Wynn, M., & Leonard, E. (2021, 8-12 February 2021). <i>Investment in Agri-Food Digital Transformation: Avoiding the Technical Fallacy</i> 65th Annual Conference of the Australasian Agricultural and Resource Economics Society, Sydney, Australia.
Conference Paper	Wysel, M. (2021, 22 June 2021). Data Sharing Platforms in Agribusiness. International Food and Agribusiness Management Association, Costa Rica.
Book Chapter	Wysel, M. (2023). Data Sharing Platforms in Agriculture. In <i>Encyclopedia of Smart Agriculture Technologies</i> . Springer. https://doi.org/10.1007/978-3-030-89123-7_250-1

Chapter-level Glossary

Term	Definition
Community Organization	The extent the data-related goals of a community are correlated with the activities of a Data Sharing Ecosystem.
Data Development	<p><i>Noun.</i> Equivalent to <i>Enrichment</i>. This phrase is adopted to connect the concept of data enrichment to <i>Agricultural Systems</i> readers following publication of Wolfert et al. (2017)'s review into Big Data in Smart Farming</p> <p><i>Similar:</i> Enrichment (used throughout except for Chapter 2)</p>

Full Manuscript

Data Sharing Platforms: How Value is Created from Agricultural Data

Authors:	Matthew Wysel ^a (matthew.wysel@une.edu.au;)*, Derek Baker ^a (derek.baker@une.edu.au), William Billingsley ^b (wbilling@une.edu.au)
Affiliations:	^a The Centre for Agribusiness, UNE Business School, The University of New England, Armidale, Australia. ^b Computational Science, The University of New England, Armidale, Australia.
Keywords	Datanomics, platform economics, value of data, big data, data sharing platforms, smart farming
Highlights:	<ul style="list-style-type: none"> • When you create value from data, you operate a data sharing platform. • Creating value from data requires a community of stakeholders, a facilitatory system, and data on and for the community. • Agricultural datanomics is proposed as the interdisciplinary study of data as an economic good created by and for, agriculture. • Data sharing platforms enable managers to answer the question, 'how do I increase the value of Smart Farming data?' • This paper evaluates common data management techniques and proposes strategies that increase the value created from data.
Acknowledgements:	Matthew Wysel is grateful for the support of the Agricultural Business Research Institute through the Arthur Rickards Innovation in Agribusiness Scholarship. The authors declare they have no conflicts of interest relating to this manuscript. Specifically, the authors are grateful to but are not affiliated in any way with <i>AuctionsPlus Pty Ltd</i> .

Abstract³

Across agriculture, data is produced, enriched, and consumed through the centuries-old practices of producing food and fiber. The adoption of Smart Farming and its connected services and techniques accelerates agriculture's dependence on data, yet the process of creating value from data is not well understood.

What assets and management decisions comprise the process of creating value from data? What are the properties of this process, and where should resources be invested to increase the value created from agricultural data?

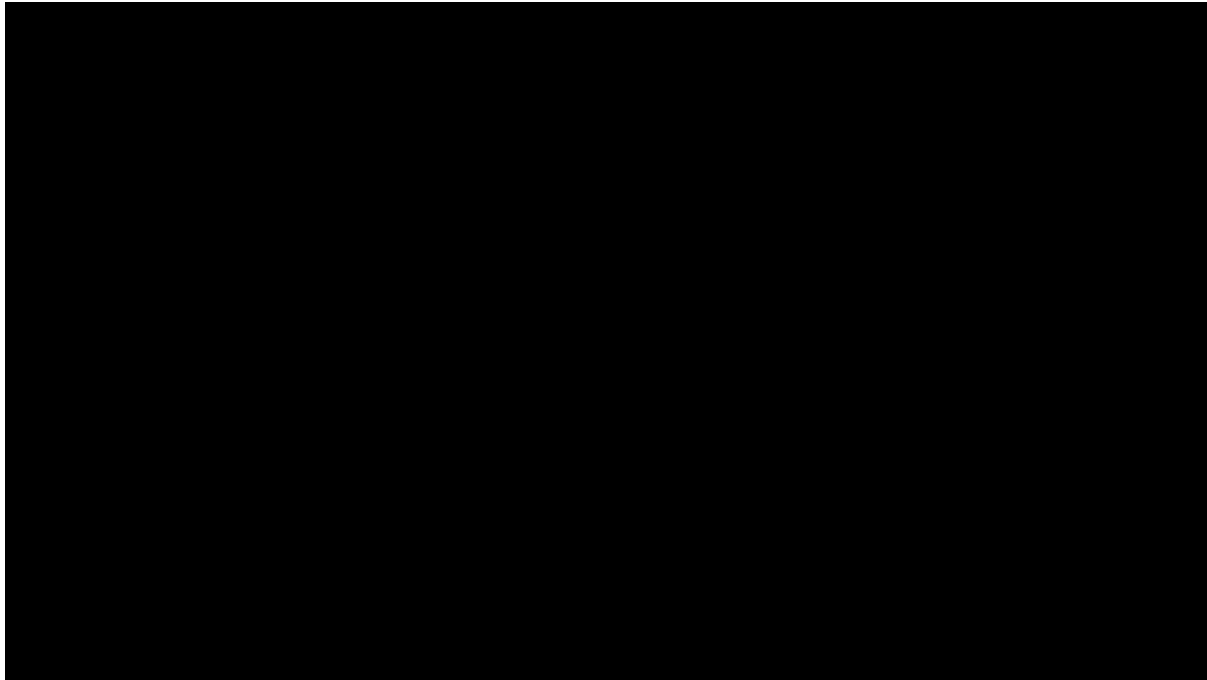
We extend platform economics theory with the results from a recent systematic literature review of Big Data in Smart Farming to show the creation of value from data occurs in Data Sharing Platforms.

Data sharing platforms are systems that connect the layers of the 'platform stack' with pertinent management tasks to create value from data. We illustrate this arrangement as a sectioned, three-circle Venn Diagram and evaluate the efficacy of common data management techniques in the creation of value from data. This paper concludes that value is created from data only when each of the components of data sharing platforms are present and that the operation of a data sharing platform describes the process that takes data as an input and produces value as an output. Further conclusions relate to commercial and institutional aspects of the creation of value from Smart Farming data.

The proposed model is useful for evaluating production processes that create value from data. This paper details several avenues for extension. Productive models of systems that rely on data as a core asset may now be assembled. Policies that trade off technical characteristics of data with social impacts of data may now be approached. Questions surrounding data ownership may be considered with greater clarity.

³ This abstract was originally prepared as a structured abstract per publishing requirements of the journal.

Graphical Abstract



1. Introduction

Agriculture 4.0 has signaled a proliferation of connected sensors across farms and throughout value chains whose streams of data offer the allure of value to those who possess the requisite skills or acumen. Staple farm equipment such as rain gauges and soil moisture probes are transformed into Internet of Things (IoT) devices that document the farming system in constant and ever-changing data flows (Friess, 2016). Meanwhile, traditional farm machinery such as tractors or livestock weighing stations are connected to non-agricultural products like smartphone apps to leverage on- and off-farm data (Barrett, 2021). These devices digitize farms and create new opportunities for agricultural value creation through the data-driven “optimization of agricultural production systems, value chains and food systems” (Klerkx, Jakku, & Labarthe, 2019, p. 2). In turn, data-based opportunities present farms with operational models that are radically different from previous generations. Farming machinery is subsumed into the sharing-economy (Daum et al., 2021; Venkataraman, 2016), robotic workforces operate alongside human laborers (Christiaensen, Rutledge, & Taylor, 2021) and provide insurance against labor shortages (Barrett, 2021), while a single farm’s access to finance can be hedged against whole-of-chain performance (AgriDigital, 2021; Jakku et al., 2019). In each case, a common belief in the value of agricultural data connects traditional agricultural problems with solutions from – often – non-agricultural communities. Where value creation in agriculture was once inseparable from labor- and capital-intensive activities, value creation in Smart Farming is now unavoidably, data intensive.

The data produced by Smart Farming intertwines agriculture and business in new ways. Where Smart Farming is the application of information technology to optimize the farming system and agribusiness, the science of making value-adding decisions regarding the agricultural assets in that system, *agricultural datanomics* can be considered as the interdisciplinary study focused on the treatment of data as a valuable resource created by, and for agriculture. Accordingly, agricultural datanomics includes consideration of the sources of value creation (Kieti, Waema, Ndemo, Omwansa, & Baumüller, 2021) through to market design (Agyekumhene, De Vries, Paassen, Schut, & MacNaghten, 2020) where stakeholders act as both producers and consumers of the value produced by Smart Farming data. Transparency is vital for embryonic digital markets (Jakku et al., 2019) as buyers and sellers often collaborate around data-products to extend value created by their exchange (Faulkner, Cebul, & McHenry, 2014). Stakeholders *multi-home* (Rochet & Tirole, 2003) with digital goods (Barrett, 2021) as their products or services participate in multiple markets simultaneously. This creates highly fragmented markets as ad hoc, agricultural platforms (Klerkx et al., 2019), whose primary advantage is access to crowd-sourced resources (Parker, Van Alstyne, & Choudary, 2016), compete with the ‘pipeline’ businesses (Choudary, 2015) that traditionally dominated agriculture. There is widespread agreement in the centrality of digitization in agriculture’s future (Jakku et al., 2019; Klerkx et al., 2019), and considerable research into the science of managing the value produced by those processes (Birner, Daum, & Pray, 2021; Wiseman, Sanderson, Zhang, & Jakku, 2019) but the issue of *how* value is created from Smart Farming data is only beginning to be addressed (Darnell et al., 2018; Rojo-Gimeno et al., 2019; Wiseman & Sanderson, 2019). This study requires consideration of both technical and economic properties of data (Birner et al., 2021; A. Fleming, Jakku, Lim-Camacho, Taylor, & Thorburn, 2018; Nikander, Manninen, & Laajalahti, 2020), in conjunction with agribusiness assets and associated management responses (Chen, Mao, & Liu, 2014; Wolfert, Ge, Verdouw, & Bogaardt, 2017). In this context, this paper aims to address two of the fundamental questions for scholars of agricultural datanomics:

- (i) what assets and managerial tasks are required to support the creation of value from Smart Farming data, and
- (ii) what properties must the process of creating value from Smart Farming data possess?

We approach both questions in parallel, initially collating extant academic literature to establish a framework of seven components required to create value from Smart Farming data before examining how their arrangement dictates that value creation process.

Value created from data, \mathbb{V} , may be understood as the difference between the benefits enabled by the data, \mathbb{B} , less the costs incurred to realize those benefits, \mathbb{C} , or, $\mathbb{V} = \mathbb{B} - \mathbb{C}$. This approach is widely adopted (Chesbrough, 2003) and is well-established in agriculture. Value created from

cropping or livestock production is the difference between the sum of market and non-market benefits less the associated costs of bringing those goods to market. Benefits enabled through the adoption of new technology or management techniques increase the value of the underlying agricultural assets through more effective observations, measurements or analysis. Likewise, a reduction in the costs of optimizing the farming system through the application of Smart Farming increases the value of farming operations by improving access to, and application of, data across the farm. Investment in key assets or accompanying management tasks increases value created from Smart Farming data as benefits are increased, or the cost of achieving those benefits is reduced.

Data produced by Smart Farming is the core resource that enables value, is often the good exchanged, and even the currency that finances interactions throughout agricultural value webs (Darnell et al., 2018). While agricultural data is *at least* an asset and “must be managed like any other [asset]” (Wiseman & Sanderson, 2019, p. 3) unlike other classes of assets, data possesses near-infinite economies of scale (Arrow, 1996). However, this property of data does not confer corresponding economies to the resulting value. Additional factors such as stakeholder’s decision rules, social intentions, and the technology used to transform data into information, all set bounds on the value created from Smart Farming data (Rojo-Gimeno et al., 2019). Value that is created from Smart Farming data is contingent on adoption and utilization of that data by a heterogeneous community of stakeholders who assemble according to a congruous goal (Wysel, 2019). Bearing some resemblance to integrated value systems (Papazoglou, Ribbers, & Tsalgatidou, 2000), stakeholders voluntarily exchange data with one another provided their individual benefits exceed both their search costs and transaction costs (Kieti et al., 2021). In this way, the farming system acts as a platform upon which a community meets to exchange data. However, stakeholder’s participation in data sharing platforms is more than transactional (Agyekumhene et al., 2020). Value can be created from Smart Farming data as stakeholders collaborate around data (Birner et al., 2021), conduct arbitrage with data (Jensen, 2007), or compete for data (Jakku et al., 2019). Additionally, the data and value created on a data sharing platform can be treated as a private good (Wiseman et al., 2019), a club good (A. Fleming et al., 2018), or a public good (Sanderson, Reeson, & Box, 2017).

Therefore, the creation of value from Smart Farming data requires more than individuals or agricultural collectives investing in digital assets or technology management. An understanding of both the complimentary assets that must exist beside data and the management tasks required to develop these assets is an important first step in investigating the process of creating value from Smart Farming data. The remainder of this paper is organized as follows: Section 2 establishes and situates this work among the current state-of-art academic and management literature on the

creation of value from data and the results from Wolfert et al. (2017)'s recent systematic literature review on Big Data in Smart Farming. Section 3 presents an intuitive framework placing each of the seven components required to create value from Smart Farming data. Section 4 examines the interplay between each component and assembles a model that describes their effect on value creation. Section 5 discusses the implications of the model for scholars and practitioners as they analyze and manage value created from Smart Farming data. Section 6 concludes with several avenues for extension.

2. Literature: Platforms, Data Sharing, and Mobilizing Value from Data

Platform business models describe organizations that broker valuable connections between typically external stakeholders (P. C. Evans & Gawer, 2016). In their simplest form, businesses that operate a platform business model, or more simply *platforms*, utilize data to create value for the participant stakeholders that assemble 'on' the platform. In platforms, managers use proprietary business processes to mobilize stakeholder's latent physical or labor assets according to some shared goal. Varying from value chain models, the community of transient stakeholders volunteer their private assets in exchange for a portion of the value created (Parker, Van Alstyne, & Jiang, 2017). In this manner, the platform provides a price structure designed to overcome a market failure through the formation and governance of a multi-sided marketplace (Rochet & Tirole, 2003). The platform distributes costs across its membership base and internalizes benefits to groups of stakeholders that would otherwise be lost as externalities (Baker et al., 2021). This allocation of the benefits and the distribution of costs occurs as stakeholders interact with one another across the platform. These activities generate additional, behavioral data which enables platforms to refine their service and remain relevant to the community's evolving priorities and sensitivities (Turland & Slade, 2020). The resulting arrangement of assets is commonly known as a 'platform stack' (Choudary, 2015) and is well developed in both the academic and management literature (D. S. Evans, 2009; Gawer & Cusumano, 2014; Parker et al., 2016; Van Alstyne, Parker, & Choudary, 2016). As illustrated in Figure 2-1, the platform stack consists of the three different types of assets common to all socio-economic platforms: first, a "network or community layer, comprised of the participants and their relationships" (Choudary, 2015, p. 61); second, an infrastructure layer that "encapsulates the tools, services, and rules" (*ibid*), and third, a data layer that "allows the platform to match supply with demand" (Choudary, 2015, p. 62). Due to the active nature of management within the infrastructure layer, we adopt the noun *the system* to refer to the entire enterprise of management working alongside the technical and business infrastructure. In commercial platforms such as the online

livestock exchange platform *AuctionsPlus Pty Ltd*⁴ the system intermediates between two sides of a market brokering valuable connections between otherwise disparate stakeholders. Such platforms come to resemble clubs (Wysel, 2019) where stakeholders act as both “producers and creators of value and generators of data, and not as [merely] consumers” (Langley & Leyshon, 2017, p. 22). As platforms extend beyond firm boundaries, brokering valuable connections between otherwise disconnected stakeholders requires the system to move beyond making one-off connections, and to actively curate persistent connectivity (Van Dijck, 2013). To this end, data is often integrated into a platform’s pricing structure as data is valued both in exchange and in use by the system and the community of stakeholders. Unlike other inter-firm arrangements such as traditional supply chains or even integrated value-webs where data is primarily a market enabler (Papazoglou et al., 2000), platforms incorporate data as an intrinsic component within their pricing structure. Access to, and concession of, data occurs in valuable ‘interactions’ (D. S. Evans, Schmalensee, Noel, Chang, & Garcia-Swartz, 2011) that become the core goal of all platform participants and the chief denominator used to measure changes in value across the platform (D. S. Evans & Schmalensee, 2016). Therefore, while platforms broker valuable exchanges for their external communities in a similar manner to firms across a value chain, platforms that preside over data-based exchanges create value by incorporating data as a tradable asset within a market structure that both enables and incentivizes stakeholders to participate.

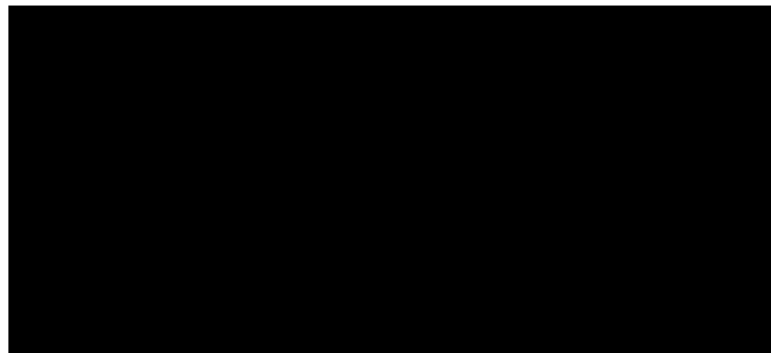


Figure 2-1: Assets from the ‘platform stack’. Adapted from Choudary (2015)

To support the clear investigation of the assets and tasks required to permit the creation of value from data, we propose the following definition of data sharing platforms:

A data sharing platform is a community of stakeholders who share a common data-related goal; data collected from, and for, the community; and a system that uses the data to enable and incentivize stakeholders to make valuable interactions.

⁴ <https://auctionsplus.com.au/>

As this paper develops, the operation of a data sharing platform describes a process that takes data as one of a number of inputs and produces value as an output. In this context, stakeholders in a data sharing platform act like individual firms in supply chains for whom collective benefit must be achieved, not by managing individual functions but by adopting an integrated approach to their separate activities (Lambert & Cooper, 2000). However, unlike a supply chain where volume is primarily driven by pricing levels, management of value created across a data sharing platform requires consideration of a multi-sided price structure (Rochet & Tirole, 2003) as stakeholders depend on same- and cross-sided network effects for the platform to create value. This is typically reflected in micro-economies that form around Smart Farming data as positive externalities generated by one part of the community are internalized to others through the use and exchange of data (Wysel, Baker, & Conway, 2019). This creation of value from data and accompanying assets is enabled by the system through its organization of the community's interactions and development of the data into information.

Wolfert et al. (2017) expand on these management tasks required to create value from Smart Farming data in their analysis of both the academic and grey Smart Farming literature. Their conceptual framework proposes how Big Data may be managed to connect industry drivers to emergent challenges, and specifically, the interventions or management tasks required to create value from Smart Farming data. Abstracting Wolfert et al.'s 'Big Data Application System' (2017, pp. 71, Figure 1) enables connection of their data application layers with each of the platform stack layers presented above. In accordance with Wolfert et al. (2017)'s framework, each management component presents the managerial variables that indicate the extent the platform's assets are developed across the data sharing platform. Figure 2-2 below, presents this abstraction and illustrates the core responsibilities agribusiness managers must administer when attempting to create value from Smart Farming data. These activities may be summarized as: Community Organization, Value Allocation, and Data Development. Figure 2-2 also accords with prior empirical work on digital service platforms by Kieti et al. (2021).

The data development task consists of farm processes, farm management and the underlying data chain (Wolfert et al., 2017) and can be broken down into four phases: data generation, data acquisition, data storage, and data analysis (Chen et al., 2014). While business processes vary considerably between different types of agricultural production, all processes share the common outcome of generating data. Once generated, data is acquired through a range of human-based and digital sensors that capture and format specific observations into storable data points. Prior to storage, raw data is categorized to support subsequent reporting and data analysis. This additional

data, or meta-data, increases the potential value created from data as fewer resources are consumed when searching for and preparing information required by future decisions.

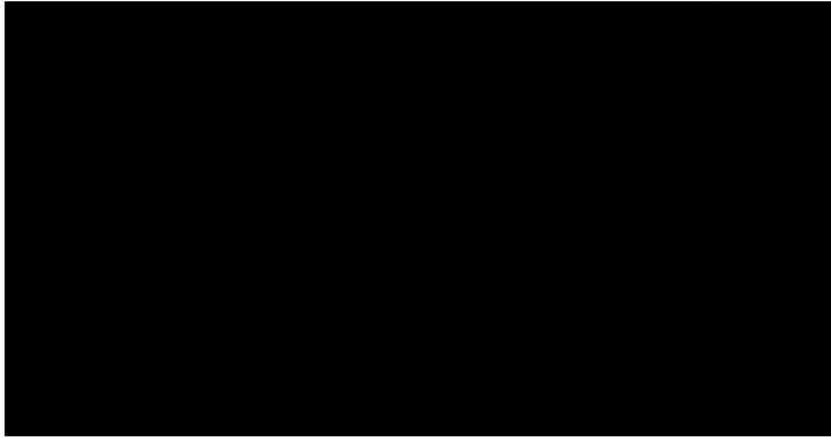


Figure 2-2 Data Management Framework. Adapted from Wolfert et al. (2017)

The second management component proposed by Wolfert et al. is community organization. The community comprises all stakeholders who pursue some goal or action that pertains to the data. This community of stakeholders includes 'users' as well as those participants who merely desire an action or outcome that is enabled by the data. Peripheral stakeholders might include other firms in the value chain such as goods and service providers, individuals such as managers or farm-based labourers, and external governance bodies such as policy makers (Wolfert et al., 2017). The value of the data to each stakeholder varies both between stakeholders and also with time. For instance, performance data of livestock may possess a high present value to a breeder but a diminished value in 12 months' time. To other stakeholders such as the breeder's insurer, the data may only hold value once aggregated. Additionally, while value may be created from data, stakeholders may only attribute that value to an action or consequence enabled by the data. In either case, value has been created from the data by the *ad hoc* community of stakeholders who have assembled on the data sharing platform.

The organization of the community and the allocation of value is also shaped by the business model adopted by the platform (Nuthall, 2018). The business model is defined by the overarching goal of the platform and determines both allocation of value between stakeholders and requirements for data development (Cook, Jackson, Fisher, Baker, & Diepeveen, 2021). The insight and tasks enabled by the data produce surplus value that is allocated across the community and the system.

Misdirected investment in the network, such as development of data towards goals not valued by stakeholders, diminishes the value created and reduces the network performance. Accordingly, platform performance is shaped by the system as it attempts to broker a *virtuous cycle* (Parker et al., 2016). In this case, the activities by one group of stakeholders cause an increase in the value created

from data by another group of stakeholders – even to the point where prospective stakeholders are enticed to begin participating in the community. This additional activity encourages even greater participation from the first group and, with it, value created from data which compounds the increase in the second group. This phenomenon describes *demand*-side economies of scale, as distinct from supply-side economies of scale which characterizes the industrial and precision revolutions in farming.⁵ When considered within a data sharing platform, the increase of data produced by Smart Farming does more than offer more efficient methods of operating. Stakeholders' self-organization and value co-creation create new networks where stakeholders who perform better, attract better performing stakeholders. This upwards cycle of value creation is enabled by the system through its allocation of value, and the development of data and organization of stakeholders across the data sharing platform.

In summary, the extant platform economics literature emphasizes the primacy of a community of stakeholders and a facilitatory system as complimentary assets for the creation of value from data. However, platform economics alone does not address how each of these assets ought to work together to create this value. Similarly, the existing grey and academic literature into Smart Farming describes the responsibilities management must consider when creating value from data but omits sufficient treatment of the underlying asset base. Therefore, the remainder of this paper aims to address this gap by joining the complementary assets required to create value from the Smart Farming data with the responsibilities management must fulfil to realize that value.

3. The Data Sharing Platform Framework: Assets, Management Tasks, and Created Value

Combining the 'platform stack' from Figure 2-1 with the interventions required to manage Smart Farming data from Figure 2-2 and valuable interactions as the nexus of value creation in socio-economic platforms, the components required to create value from Smart Farming data may be depicted as shown in Figure 2-3. Data sharing platforms consist of three assets together with three management tasks that span each pair of assets to enable valuable interactions across the platform. The three core assets data sharing platforms rely on are: a community of stakeholders, a facilitatory system, and data on and for that community. Bridging each pair of assets, the interventions proposed in Wolfert et al. (2017)'s review describe three distinct management tasks whose common

⁵ Larger, supply-side economies of scale improve profits through a reduction in unit-cost, whereas larger demand-side economies of scale create positive, closed-loop feedback in a market. See Shapiro and Varian (1998) for a full development of demand-side economies of scale in digital networks.

purpose is to create value by enabling valuable interactions across the platform. These management tasks are organization of the community, allocation of value both between stakeholders, between the system and the community, and development of data into information. Valuable interactions act as the seventh, and central component and are the desired output of the application of appropriate management to the underlying assets. Collectively these seven components must be present to create value from Smart Farming data. These seven components represent a sectioned, three-circle Venn diagram, which may be expanded across three co-dependent layers: assets, management tasks, and output. Each component will be discussed in turn.

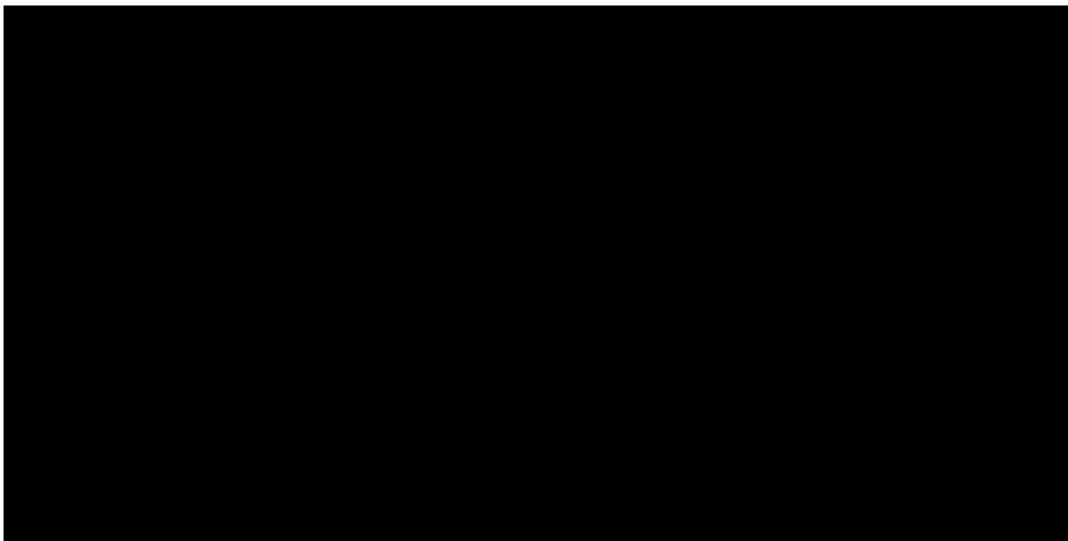


Figure 2-3: Components necessary for the creation of value from data

3.1. The Asset Layer: A Community of Stakeholders

The community exists because data sharing platforms offer stakeholders a more efficient mechanism for realizing the benefit attached to their data than if they pursued that benefit in isolation. As with farmers markets before them, platforms diminish stakeholder's marginal cost of realizing value from data by matching stakeholders with one another around congruous goals (D. S. Evans & Schmalensee, 2016). Effective arrangement of processes that guide and direct the community, enable stakeholders to achieve a benefit each desire at a cost each accepts. For example, *AuctionsPlus* dramatically reduces the search cost for stakeholders on both sides of the market by leveraging knowledge of previously successful searches to broker beneficial matches. Stakeholders on *AuctionsPlus* self-identify as 'buyers' or 'sellers' through the data they offer which the system uses to respond sale data that may be of interest. The algorithms that govern current matches are informed by data obtained during previous successful and, at times, unsuccessful matches within the community. This permits *ad hoc* communities of stakeholders that form on

AuctionsPlus to create more value from their data through more rigorous – and ostensibly more beneficial – searches. Participation in the *AuctionsPlus* data sharing platform also permits stakeholders to create value from data at a lower cost than if they attempted to create value from that same livestock data independent of the platform. Where the value created from data is the difference between the benefit derived from the data and the cost to achieve that output, the value created by each stakeholder, V_{C_i} , can be expressed as the difference between the benefit each stakeholder accrues, B_{C_i} , and the cost each incurs, C_{C_i} ,

$$V_{C_i} = B_{C_i} - C_{C_i}.$$

When taken in aggregate,

$$V_C = \sum_i^n B_{C_i} - \sum_i^n C_{C_i}$$

where each i th stakeholder participates in a community of total size, n , until their benefit of participation ceases to exceed the associated marginal cost. Therefore, *ceteris paribus*, the value created from data continues to increase while n increases.

3.2. The Asset Layer: An Enabling System

In a data sharing platform, the system presides over the data and must continually develop the data towards the loci of goals specified by the community. Observation of valuable interactions between stakeholders in the community becomes an important metric in determining system efficacy (Parker et al., 2016). However, the system must do more than just deliver valuable interactions to stakeholders and minimize their short-term search costs. Stakeholders reduce uncertainty in data as they generate, search, select and process data. Also, the system must balance investments in interrogating this activity-data with the timely delivery of potentially valuable data to stakeholders. For example, one stakeholder who carefully considers a multitude of sale lots on *AuctionsPlus* and selects one lot from an ostensibly, homogenous list strips away more uncertainty for the whole platform than if they had selected the first lot presented. In the former case, greater uncertainty was removed from the assembled data at a higher personal and system cost. However, by the same means evaluating and discarding more data points produces greater potential benefit for both the stakeholder and community. Therefore, a system seeking to create the most value from platform data may divert stakeholders away from the lowest cost set of interactions and along a path of interactions that produces the most benefit to the rest of the community. It follows the system must balance the marginal costs accrued by each stakeholder on the platform with the marginal benefit each stakeholder expects to receive from the data.

In this way, the system may choose to create value from data by conferring benefits to stakeholders at a lower cost or by deepening users' enrolment in "a participatory economic culture" (Langley & Leyshon, 2017, p. 14). While stakeholders use the data to make valuable interactions, the system may use the same data as an additional means of extracting value through the capitalization of the community's preferences. Where \mathbb{V}_S is the value created by the system from data and \mathbb{V} is the total value created from data, we can write:

$$\mathbb{V}_C = \mathbb{V} - \mathbb{V}_S.$$

However, this simultaneous extraction of benefit need not be competitive. To the extent the system can design interactions across the platform to yield value to the community and gather additional insight, both groups benefit. However, past the point of Pareto optimality, additional benefit extracted by the system will be borne as a cost by the community where the limit to the system's 'over extraction' of value is the consumer surplus each stakeholder derives from the data. Above this level, stakeholders will simply cease participating and quit the platform. Until that point, whether through greater development of data, or more effective organization of the community, the system takes investment by some means and *ceteris paribus*, the value of the data continues to increase while stakeholders in the community continue to realize the benefits from the data provided by the system.

3.3. The Asset Layer: Data on and for the Community

Data is the core asset that permits and facilitates the creation of value on data sharing platforms. Data acts as a resource when it enables a non-data outcome, such as livestock exchanged on *AuctionsPlus* or fields that are variably fertilized with data from Smart Harvesters. Data acts as an economic good when developed data *is* the outcome desired, such as clearance prices on *AuctionsPlus* or news articles in a Facebook farming community. Finally, data is valued as a currency when it acts as a store of value or facilitates an exchange of value between stakeholders or other parties. In this final case, data is neither the immediate outcome nor the stakeholder's goal *per se*. Data is valued as a currency by farmers who, for example, collect data for university research programs to benefit from government grants, or by agronomists who maintain membership in data aggregation clubs to access new clients.

Value is also created from data when the data sharing platform exchanges data with an external data marketplace (Jones & Tonetti, 2020). This valuation in exchange can be positive, where data is exchanged for benefit to the platform, \mathbb{B}_D , or negative, where data is purchased at cost from outside the data sharing platform, \mathbb{C}_D . In the former case, data that has been acquired and

developed by the combined actions of the system and community of stakeholders may be realized as a benefit and supplements the creation of value by the platform. The latter case may arise if the purchase of data constitutes a smaller cost to the platform than internal development of the same data.

Therefore, while *ceteris paribus*, an increase in the amount of data corresponds to an increase in value, this is contingent on the evolving data set not degrading in certainty and remaining aligned with the needs of the community.⁶

3.4. The Management Layer: Data Development

We now turn to consider the second layer of Figure 2-3: the management layer. Taken collectively, the three management tasks reflect the central management challenge of platform managers: how to build a community of stakeholders that share similar goals regarding data such that they remain engaged in the value creation process of the platform and satisfied with the ensuring creation of value.

The management task of developing data sits on a continuum. If data is developed towards too fine a collection of goals, it will cease to become valuable to a sufficiently large set of stakeholders who share similar but now incongruous goals. Similarly, data left under-developed will not sufficiently reduce a stakeholder's own data development costs causing stakeholders to decide that the benefit no longer outweighed the costs and that the rational action would be to cease participating in the community.

Data sharing platforms must generate, acquire, and store data to support the creation of value (Chen et al., 2014), but the creation of value from data must progress beyond those mechanics. Data must also be developed to support the needs of each stakeholder. Stakeholders use data developed by the system to lower their personal cost of accessing the same outcomes, or to achieve greater outcomes for the same cost (Jones & Tonetti, 2020). Extending the earlier example of a stakeholder who has signed in to *AuctionsPlus*, the system assumes an interest in purchasing livestock from a nearby location and at a near future time such as the next lot sales. The stakeholder confirms or corrects this assumption with their selection of fields on the *AuctionsPlus* website. This *correction signal* – of the form used in communication theory (Shannon, 1948) – permits an update to the data

⁶ A brief note on the impact the conceptualization of data sharing platforms have on both blockchains – that ostensibly do away with a centralized system – and Big Data – that is defined as *too big* for consideration by a community of stakeholders – has been included in Section 5.4.

offered to that specific stakeholder and, in aggregate, enables changes to the processes used to deliver data to all stakeholders. This improvement in network performance is the data-driven means by which one agent's value creation process is *spilled-over* to others in the community.

Data development can also happen *across* markets. In platforms that preside over multi-sided markets such as *AuctionsPlus*, stakeholders quickly begin considering data other stakeholders have provided to the system. Here the system prompts stakeholders to arrange their own data to maximize their – and other's – direct benefit. 'Suggested prices', 'suggested keywords' or 'minimum recommended fields' are examples of nudges delivered by the system to help the community create value from data. As developed in Section 4, this direct and indirect data development is the transformative step that permits demand-side economies of scale in the creation of value from data. Finally, while we leave the extension to others, the collaborative development of data by stakeholders who voluntarily share and develop data for their own, private benefit remains a core component required to assemble a production function around data.

3.5. The Management Layer: Community Organization

If the goals of the community of stakeholders are too diverse, delivery of the desired data will be too costly, and stakeholders will drift away. Conversely, if the goals of the community are too concentrated, signals arising from the successful interactions around data will fail to attract sufficient prospective stakeholders.

In a manner following Lancaster (1966), demand for data will vary between stakeholders as each maintains a different agenda for the data and therefore derives a different utility. One stakeholder may visit *AuctionsPlus* with the explicit intention of finding and purchasing Angus heifers, while another stakeholder may visit the same site only to compare prices. The former stakeholder may expect a sufficiently large utility from the data to evaluate several lots worth of livestock before bidding on a sale. Conversely, the second stakeholder may be unwilling to pay attention to more than one lot. While both stakeholders share the goal of obtaining 'data on heifers', their utility of that data varies and, with it, their willingness to pay the cost of developing that data. The literature into Rational Inattention (e.g. Sims (2003)) explains the indirect costs each stakeholder bears as they consider additional data. Practically, at a certain point it becomes rational for each stakeholder to withhold attention from additional data even if that data could have further improved their utility. While others have investigated the causes of these indirect costs (Caplin & Dean, 2015; Gentzkow & Kamenica, 2014), the consequence to the creation of value from data is that stakeholders will vary their efforts to maximize their personal value. The variance in the value created by each stakeholder

from the nominal value created by the community also permits calculation of the *correlation* of the community. This introduces a novel analytical approach where the value creation process becomes partially inverted. Where previously, data was developed to maximize the value created for stakeholders, new stakeholders may now be developed to maximize the value imputed to data. This analysis would rely on development of the production function discussed earlier, and we leave this as a novel extension of this work.

The performance of the network of stakeholders is also driven by the governance framework and the value creating mechanism adopted by the data sharing platform (Wolfert et al., 2017). *Help Wanted* listings for seasonal labor are posted in data sharing platforms that have formed for the specific purpose of sharing data outside the platform. Data kept 'on-platform' enables fewer benefits than data shared 'off-platform'. Conversely, accounting data shared beyond the accountant and business owner without both party's prior agreement, would constitute a breach of governance and likely diminish value created. Data sharing platforms that form around data gathered from Smart Harvesters, for example, are more complex; the manufacturer sees a potential for value creation through the wide-spread sharing of operational data, while the farmer prefers the data to be kept private (Faulkner et al., 2014). In all cases, consideration of the impact of changes to the network of stakeholders is vital for the sustained creation of value from data. Section 4 extends this relationship by formally relating management decisions to the productive power of the platform's assets.

3.6. The Management Layer: Value Allocation

Stakeholders participate in data sharing platforms to create value from data. This value creation requires sufficient investment by a system to permit the effective development of data that accords with stakeholders' goals. The system observes the activities of stakeholders as they search for, consider, and validate data each deems relevant. When stakeholders negotiate terms of exchange on a data sharing platform, such as *AuctionsPlus*, the system observes the full path taken by both sides of the market on all successful, and unsuccessful, exchanges. Accordingly, the system builds a more complete understanding of the community's preferences than any combination of stakeholders, past or present. When coupled with appropriate technology, this understanding enables the creation of value from data in a way the community of stakeholders could not otherwise achieve. This difference in value creation mechanisms describes two markets: an internal value creation market between the community and system, and an external market characterized by stakeholders' various alternatives.

The system can use this understanding to conduct arbitrage between these two markets (Jones & Tonetti, 2020) up to the point of Pareto optimality. Systems must appropriate some value to sustain operations, but stakeholders will prefer platforms with the lowest tax on their value creation. Under-extraction of value by a system leaves it with too few resources to gather and enroll users. However, over-extraction of value from data diminishes the benefit that holds stakeholders to the platform. Therefore, up to a Pareto efficient allocation of value, stakeholders and the system collaborate to create value from data.

3.7. The Output Layer: The Creation of Value as an Output of Data Sharing Platforms

Valuable interactions are the center of the sectioned Venn Diagram and the third layer of the data sharing platform in Figure 2-3. Production of valuable interactions requires mobilization of platform assets by the relevant management decisions as stakeholders are brought together according to the goals of the data sharing platform. As seen graphically in Figure 2-3, each management task must account for the nature of the assets that it spans, in order to support the creation of value from Smart Farming data. As the system develops data to support the goals of the community, stakeholders are enabled to interact, collaborating around, competing for, and using or exchanging data in a broader process of value creation. The design of this micro-economy determines the relative allocation of value between stakeholders, and between the community and system. Section 4 develops and then incorporates these relationships into a model that describes the process of creating value from data, culminating in a single, objective function that relates value created to each of the constituent components.

4. Operationalizing the Data Sharing Platform Model

In order to apply the proposed framework in Figure 2-3, we must first examine the value-creation properties of each of the seven components in data sharing platforms and how each component interacts. As introduced in Section 3.5, we can summarize the benefits the community receives as a function of the extent to which data enables stakeholders to achieve their goals. That is, the parameter ρ describes the organisation of the community, and μ the relative allocation of benefits and costs between the community and system. We can write,

$$\mathbb{B}_{C_i}(\rho_i, \mu_i).$$

Likewise, the benefit captured by the system, \mathbb{B}_S , will remain a function of the extent that it confers positive externalities on the community together with the additional uncertainty it strips from the data. Importantly, while the community maintains a crucial role in this data development process,

the goals possessed by the community of stakeholders will often relate to, but differ from, the goals of the system. In terms of the management layer, the benefit captured by the system is a function of the share of value retained (or allocated to itself), μ , and the value created from the development of the data, ε . From Section 3.3, each of these variables is driven by the interactions, k , the system has with stakeholders. Each of the stakeholders interact with the system and through it, the rest of the community until they either achieve their goal or give up. At either point, each stakeholder ceases participating on the platform. Therefore, the total number of interactions served by the system is the product of the number of stakeholders in the community, $n \in N$, and the maximum interactions of each of those stakeholders, K .⁷ Therefore, we can express the benefit captured by the system as,

$$\mathbb{B}_{S_k}(\varepsilon_k, \mu_k)$$

where $k \leq KN$. Finally, the benefit accrued by the data, \mathbb{B}_D , is a function of the extent the data enables value for the stakeholders, ρ , and the relative development, ε , it has received. Where there are j , data points, we can write,

$$\mathbb{B}_{D_j}(\rho_j, \varepsilon_j)$$

Therefore, summing the benefits accrued to each asset on a data sharing platform we can express the benefit created from data as,

$$\begin{aligned} \mathbb{B} &= \mathbb{B}_C + \mathbb{B}_D + \mathbb{B}_S \\ &= \sum_{i=1}^N \mathbb{B}_{C_i}(\rho_i, \mu_i) + \sum_{j=1}^J \mathbb{B}_{D_j}(\rho_j, \varepsilon_j) + \sum_{k=1}^{NK} \mathbb{B}_{S_k}(\varepsilon_k, \mu_k) \end{aligned}$$

where $j \leq J \in \mathbb{N}$. When the size of the community, number of observations and interactions all become large, we can approximate this discrete case with the continuous,

$$\mathbb{B} = \int_1^N \mathbb{B}_{C_i}(\rho_i, \mu_i) \cdot di + \int_1^J \mathbb{B}_{D_j}(\rho_j, \varepsilon_j) \cdot dj + \int_1^{NK} \mathbb{B}_{S_k}(\varepsilon_k, \mu_k) \cdot dk \quad (2-1)$$

As introduced in Section 3.2, interactions between the system and stakeholders are the primary means by which data is developed across data sharing platforms. Therefore, while Equation (2-1) contains the general statement of benefits delivered across a data sharing platform, we can make several simplifying assumptions that more clearly illustrate the centrality of interactions between the system and stakeholders as the driving force for creating value from data.

⁷ We can algebraically deal with a variance in the number of times stakeholders interact with the system by either setting K as the global average or by setting the content of any stakeholder's surplus interactions to zero.

Initially, if the significant portion of the benefit arising from the development of data comes from data provided by stakeholders during their normal activities on the platform, then $j \cong k$. The conditions required for this simplification are that data is acquired by the platform primarily through the system's interactions with stakeholders as described in Section 3.4, and that the level of development of data acquired at each interaction may be considered constant. Further, in platforms where each stakeholder's interactions with the system are characterized by a series of repeated, ostensibly equivalent interactions such as sale lots viewed, transactions processed, or messages in forums, then each stakeholder's interactions can be treated as economically equivalent to separate interactions by different stakeholders. With these simplifying assumptions, Equation (2-1) can be rewritten with interactions as the primary independent variable,

$$\mathbb{B} = \int_1^{NK} \mathbb{B}_{C_k}(\rho_k, \mu_k). dk + \int_1^{NK} \mathbb{B}_{D_k}(\rho_k, \varepsilon_k). dk + \int_1^{NK} \mathbb{B}_{S_k}(\varepsilon_k, \mu_k). dk$$

The nature of costs incurred while creating value from data within a data sharing platform proceeds similarly. Costs incurred mirror the earlier functions as management tasks drive value creation from the underlying assets. Summing the costs of each asset class we have,

$$\begin{aligned} \mathbb{C} &= \mathbb{C}_C + \mathbb{C}_D + \mathbb{C}_S \\ &= \sum_1^N \mathbb{C}_{C_i}(\rho_i, \mu_i) + \sum_1^J \mathbb{C}_{D_j}(\rho_j, \varepsilon_j) + \sum_1^{NK} \mathbb{C}_{S_k}(\varepsilon_k, \mu_k) \end{aligned}$$

Now if the platform implements a structural, systematic response to achieving a specific level of community organization, then the cost of achieving that organization may be borne by the whole platform and may be written as a price function, $P_\rho \rho$. Likewise, if the data acquired is a standard level of 'rawness' and the demands of the stakeholders require it to be developed to a common level, then the cost of data development can also be considered as a price function, $P_\varepsilon \varepsilon$. Finally, if allocation of value is constant across the community, then it will impose a constant cost rate during operation of the data sharing platform, $P_\mu \mu$. We can therefore express total costs as the sum of asset costs plus price functions,

$$\mathbb{C} = \sum_1^N \mathbb{C}_{C_i} + \sum_1^J \mathbb{C}_{D_j} + \sum_1^{NK} \mathbb{C}_{S_k} + P_\rho \rho + P_\varepsilon \varepsilon + P_\mu \mu$$

We now have a cost function where the first three terms constitute the cost of amassing platform assets while the subsequent three terms constitute the costs of managing those assets. Extending the simplifying assumptions above, if the cost each stakeholder incurs while creating value from data approaches a constant, then the cost to the platform of acquiring each dataset also remains

constant. This is the case when datasets are contributed by stakeholders at each interaction. Therefore, where the cost to the system of each interaction also remains constant – such as serving a webpage, or computing weed growth from paddock imagery – we can further simplify total costs to,

$$\mathbb{C} = N_c \mathbb{C}_{C_i} + J_d \mathbb{C}_{D_j} + K_s \mathbb{C}_{S_k} + P_\rho \rho + P_\varepsilon \varepsilon + P_\mu \mu$$

The cost incurred by stakeholders as they create value from data, \mathbb{C}_{C_i} , consists of signal costs, time costs and the unit cost of evaluating each dataset. Signal costs, such as membership or commission fees, are often associated with specific events or actions conducted by a stakeholder. Where signal costs are small compared to usage costs, stakeholders' costs become approximately proportional to the number of their interactions with the system. Likewise, where data is provided to the platform through a series of ostensibly equivalent interactions by stakeholders, then the cost of data acquisition, \mathbb{C}_D , also becomes a function of the number of interactions that occur across the platform. Finally, if the cost to the system of developing each of these datasets, \mathbb{C}_S , approaches a constant, then it too may be represented as a function of interactions. Therefore, platform costs may be expressed as the sum of the cost of amassing the assets and the cost of maintaining those assets at their current levels,

$$\mathbb{C} = K(\mathbb{C}_{C_k} + \mathbb{C}_{D_k} + \mathbb{C}_{S_k}) + P_\rho \rho + P_\varepsilon \varepsilon + P_\mu \mu$$

Uniting benefits and costs into a single objective equation yields the value created from data across a data sharing platform as a function of the number of interactions across the platform,

$$\begin{aligned} \mathbb{V} &= \mathbb{B} - \mathbb{C} \\ &= \int_0^K [\mathbb{B}_{C_k}(\rho_k, \mu_k) + \mathbb{B}_{D_k}(\rho_k, \varepsilon_k) + \mathbb{B}_{S_k}(\varepsilon_k, \mu_k)] \cdot dk \\ &\quad - \{ k[\mathbb{C}_{C_k} + \mathbb{C}_{D_k} + \mathbb{C}_{S_k}] + [P_\rho \rho + P_\varepsilon \varepsilon + P_\mu \mu] \} \end{aligned} \quad (2-2)$$

This equation accords with the standard business practice for measuring the creation of value where profit equals net revenue less the sum of variable and fixed costs. While variable costs scale with respect to output, fixed costs remain a function of the effort required to maintain current operations. We see this same relationship in Equation (2-2): the value created by data is equal to benefits from the current period less the sum of the costs that scale with interactions and the cost of maintaining current operations. We turn now to consider the implications of managing the process this model describes.

5. Implications for Managing the Value Created from Data in Smart Farming

The creation of value from data produced by Smart Farming occurs in data sharing platforms as stakeholders in a community are brought together by a facilitatory system. As interactions are the base activity from which all platforms create value, the creation of value from data is typically evaluated *per interaction*. Taking the partial derivative of Equation (2-2) with respect to interactions across the platform, k , we have,

$$\frac{\partial V}{\partial k} = \{ \mathbb{B}_{C_k}(\rho_k, \mu_k) + \mathbb{B}_{D_k}(\rho_k, \varepsilon_k) + \mathbb{B}_{S_k}(\varepsilon_k, \mu_k) \} - \{ \mathbb{C}_{C_k} + \mathbb{C}_{D_k} + \mathbb{C}_{S_k} \}. \quad (2-3)$$

This represents the net value created by data for each interaction: the sum of benefits realized by the community and the system and achieved from the data, less the cost of achieving those benefits. This relationship of net value generated per interaction could also be considered the gross profit achieved by the data sharing platform from data across each interaction (Wysel, 2021). In this context, while the gross margin rate – benefits divided by costs - exceeds 1, the data sharing platform continues to produce value from data with every interaction. Recalling the operational simplifications made in Section 4, above, if interactions on the data sharing platform are the primary means of data acquisition and all stakeholders in the community expect and receive the same treatment, then the cost rate of the platform's assets may be considered a constant. This is evident in Equation (2-3), above. Therefore, where the marginal utility of data is concave to the origin such as when the rate of change of assets is small,⁸ setting $\frac{\partial V}{\partial k}$ to zero reveals maximisation of value from the objective equation will occur where the marginal benefit realised from data equals the cost rate of the platform's assets. This has several important implications in the practical pursuit of managing data produced by Smart Farming.

If managers of Smart Farming data choose to focus on partnerships that unlock more efficient methods for managing on-farm systems and value chain data then, from Equation (2-2), this will have the effect of reducing the fixed costs incurred in the production of value from data. These are often among the first strategies explored by technology managers as they attempt to come to grips with their information systems and the requirements the management of data places on their business. Examples of these strategies include adopting cheaper technology contracts, adopting a

⁸ Selecting the community as an archetype asset: in the case where the size of community was growing rapidly with respect to interactions, then network effects such as cross-market and same-sided externalities would dominate and could reasonably produce utility curves for stakeholders whose second derivatives was positive – in the case of new, complimentary stakeholders – or negative – where new stakeholders compete for value or displace one another.

‘hands-off’ approach to stakeholder management, or reducing general managerial overheads⁹ which are analyzed in Table 2-1. While these are useful strategies in the formative stages of value creation, the reduction in the fixed costs of value created from data only increases value extracted from data under the current production regime. These strategies do not directly affect the number of interactions and the data sharing platform will continue to find value creation profitable.

Strategy	Example Actions	Desired Direct Impact	Potential (Negative) Indirect Impacts
Level 1: More Efficient value creation	<ul style="list-style-type: none"> - Reduce ICT cost e.g., ‘pursue cloud adoption strategy’ - ‘Hands-off’ approach to stakeholder management - Lower management overhead 	Reduction in cost of building assets	<ul style="list-style-type: none"> - Locked into vendor’s products, reduce ε - More dispersed community, reduce ρ - Reduction in system benefit, reduce μ
Level 2: More Sustainable value creation	<ul style="list-style-type: none"> - Make it easier for stakeholders to use platform - More powerful algorithms - Better quality ICT contracts 	Reduction in cost of managing value creation	<ul style="list-style-type: none"> - Split the community, reduce ρ - Data becomes too focused, reduce ρ - \mathbb{B}_D increased but ρ reduced, or lower ε and lower \mathbb{B}_C
Level 3: More Effective value creation	<ul style="list-style-type: none"> - More beneficial data for the community - Increase benefit acquired by system - Secondary value chains for platform data 	Increase total benefit per interaction, $\frac{\partial \mathbb{B}}{\partial k}$	<ul style="list-style-type: none"> - Increase P_ε, reduce k - Beyond Pareto-optimal position, reduce μ - Split community, reduce ρ

Table 2-1. Example strategies for creating value from data.

Alternatively, managers of Smart Farming data could choose to focus efforts on reducing the variable cost of value creation. Reduction in the cost each stakeholder experiences as they participate in the data sharing platform, such as a reduction in search cost, or a reduction in the cost to the system for providing that data, both increase value created from data by increasing the number of potentially valuable interactions across the platform. These strategies aim to reduce the rate of increase in total costs and, therefore, to increase the number of interactions the data sharing platform will continue to profitably create value.

Finally, managers of Smart Farming data could focus on improvements to marginal outputs. That is, the benefit enabled by the data sharing platform at each interaction. Increasing the marginal benefit created by data from each interaction has the effect of both improving the current value of

⁹ Each of these examples reduce the *price* of data development, community organization or the allocation of value, respectively.

production and increasing the number of potentially valuable interactions stakeholders may have on the platform. As Equation (2-3) illustrates, each set of benefits are a function of the three, core managerial responsibilities introduced by Wolfert et al. (2017).

5.1. Value Creation from Data Development

The impact of data development, ε , on the value created by data during each interaction can be determined by taking the partial derivative of Equation (2-3) with respect to ε .

$$\frac{\partial^2 \mathbb{V}}{\partial k \partial \varepsilon} = \underbrace{\left\{ \frac{\partial \mathbb{B}_D}{\partial \varepsilon} + \frac{\partial \mathbb{B}_S}{\partial \varepsilon} \right\}}_{\text{DIRECT IMPACT}} + \underbrace{\left\{ \frac{\partial \rho}{\partial \varepsilon} \left(\frac{\partial \mathbb{B}_C}{\partial \rho} + \frac{\partial \mathbb{B}_D}{\partial \rho} \right) + \frac{\partial \mu}{\partial \varepsilon} \left(\frac{\partial \mathbb{B}_C}{\partial \mu} + \frac{\partial \mathbb{B}_S}{\partial \mu} \right) \right\}}_{\text{INDIRECT IMPACT}} \quad (2-4)$$

(Refer to the Appendix for derivation).

Recalling the multi-layered, arrangement of assets and managerial functions from Figure 2-3, the two operators in the first pair of braces are the *direct* impact the pursuit of data development has on the two underlying assets. Here, the two direct consequences may be understood as ‘what change does the pursuit of the development of data bring about in the benefit realized from the outward exchange of data, and the benefit realized by the system?’ Equation (2-4) also illustrates the interdependence of the different managerial responsibilities. In this case, the ongoing pursuit of the development of data indirectly impacts assets via any correlation with community organization, ρ , and the allocation of value, μ , across the platform. This is given by the two coefficients in the second pair of braces. Comparison of the contents of each pair of parentheses with Figure 2-3 reveals each of these indirect impacts also drive changes in their underlying assets.

We now have the explanation for our earlier observation: if data is developed towards too fine a locus of goals, it will cease to become valuable for a sufficiently large set of stakeholders with congruous goals. In terms of Equation (2-4), while ongoing development of data may be justified in terms of improving \mathbb{B}_D and \mathbb{B}_S , if it serves to divide the community or adversely affect allocation of value across the platform, that is, $\frac{\partial \rho}{\partial \varepsilon}, \frac{\partial \mu}{\partial \varepsilon} < 0$, then these efforts could diminish rather than increase value created from data. From Equation (2-4), this outcome would be expected if the dominant portion of the value created from data was generated as benefits enjoyed by the community; that is, $\mathbb{B}_C \gg \mathbb{B}_S, \mathbb{B}_D$. Indeed, this hypothetical loss in value through an over-development of data was the consequence in the earlier, fictitious example where *AuctionsPlus* could have acted as if all stakeholders were only interested in steers and counted all data not pertaining to steers as irrelevant.

Irrespective of potential indirect impacts, data sharing platforms must achieve some development of data to support the creation of value. As each stakeholder only participates until their marginal utility equals zero, that is while $\mathbb{B}_{C_i} \geq \mathbb{C}_{C_i}$, the data sharing platform must develop the data sufficiently to reduce each stakeholder's cost of creating value from data to remain beneath their expected benefit. This dynamic is also present at the platform level where, from Equation (2-3), benefits must continue to outweigh costs for ongoing value creation.

5.2. Value Creation from Community Organization

Value creation from a focus on community organization proceeds in a similar manner to data development in the previous section.¹⁰ A focus on improving network performance or the goals that characterize stakeholders, directly improves the benefits realizable from data or achieved by the community but also carries the indirect impacts characterized previously.

To continue the prior example, suppose *AuctionsPlus* presumed all stakeholders participated solely to purchase livestock. Applying Wolfert et al. (2017)'s framework of managerial responsibilities to increase value created by data, the data sharing platform could make investments in network performance that targeted their closely organized community. One such development may be investment in a new data development process that improved the specificity of stakeholders' search results where a stakeholder's search may formerly have produced a prioritized list of hundreds of results, it now produces a definitive list of, say, ten results. The community would exhibit a high correlation if this increase in development efficacy resulted in an increase of benefits to stakeholders or captured within the data. Recalling stakeholders participate in a data sharing platform only while $\mathbb{B}_{C_i} \geq \mathbb{C}_{C_i}$, *AuctionsPlus* might test for this increase in benefits by testing stakeholder's willingness to accept additional platform-based costs. In the absence of competition, the system may be able to appropriate some or all of these benefits by increasing direct or indirect costs to those stakeholders. The inclusion of advertisements and the sale of anonymized activity data are two techniques commonly adopted by commercial platforms as they seek to capture value created across the platform. However, suppose a second scenario where the community was not participating to purchase livestock but simply to compare prices between livestock. An increase in the specificity of search would only partially support their goal. They would have fewer results to compare, and while they could make some comparisons more efficiently, they would evaluate less

¹⁰ For readability, Equation (2-4) has not been adapted and reinserted for Section 5.2 and 5.3. Examination of Figure 2-3 in the context of the relevant sub-heading and relationships in Equation (2-4) reveals the same principles of direct- and indirect-impacts of data management strategies.

data. Their utility may have increased, but their preparedness to accept costs could also have decreased. In this case, the correlation of the community with the system's development of data would have been lower than in the first scenario but still, conceivably, positive. Finally, suppose a third scenario, where the community was only interested in the search terms that returned the most results. The increased specificity would then work against the goals of the community as the correlation of the community with the efforts of management would be negative. In this third scenario, if tolls levied on the benefit derived by stakeholders as they traded livestock was the dominant source of revenue, investments in community organization, such as acquiring like-minded stakeholders, would improve both the value created from data and revenue more than investments in data development.

5.3. Value Creation from Value Allocation

Finally, we consider the creation of value from data through a focus on the allocation of value across the data sharing platform. From Figure 2-2, the allocation of value stems from investment in technology and organizing stakeholders. Stakeholders participate in data sharing platforms precisely because even unwitting collaboration reduces the personal cost of achieving the benefits desired from the data. *AuctionsPlus* unites buyers and sellers around data pertaining to livestock for sale. Farming groups on Facebook bring stakeholders with similar data-related goals together to swap knowledge or services. In both cases, the search cost for finding and acting on relevant data is reduced by the data sharing platform. In this manner, the system permits capitalization of a stakeholder's otherwise marginal, data development costs. To the extent data desired by stakeholders can be anticipated, the system may seek to minimize the cost of provision of that data: both C_{C_k} and C_S . However, while this shared reduction in cost increases potential value to both parties, as in the previous example, the system may also appropriate some or all of that benefit. Thus, this improvement in technology permits a shift in the allocation of value to the system through an increase in B_S .

This process is well developed in Club theory, where benefits may be extracted from stakeholders as either *anonymous* or *non-anonymous* tolls (Sandler & Tschirhart, 1997). In the former, all stakeholders pay tolls for access to the system that enriches data to a common level. Data sharing platforms such as paid subscription investment forums or commercial farmers markets are examples of this approach. In the case of non-anonymous tolls, fees are levied on stakeholders based on their activity or identity. A stakeholder's consumption of advertisements on farming newspapers, or transaction fees on exchange platforms are examples of this approach.

5.4. Value Creation from Blockchains and Big Data

The proposed framework and resulting model are also useful for understanding how value may be created from Big Data produced by Smart Farming and agricultural blockchains. The former is defined by an exclusive reliance on a system (Wolfert et al., 2017), while the latter is heralded for its ostensible independence from a system (Leng, Bi, Jing, Fu, & Van Nieuwenhuysse, 2018).

Reliance by blockchains on a growing ledger of transactions is designed to replace the trust mechanism conferred by a third party's endorsement of each dataset. The premise of the growing ledger of transactions is the creation of a dispersed system that validates current data through practically indelible links to past data-related activities. Each element in a blockchain contains references to past states that have each been affected by the process the datasets have passed through. In this way the raw data of an embryonic dataset is developed according to the goals of the community. The distributed nature of the ledger gives the data development process the veracity it requires to validate the data it produces. These processes resemble the system as defined, as they are specified by the community, defined by the conditions for valuable data, and are configured to separate valid data from invalid data. Indeed, scarcity and with it, value is defined as data that possesses those characteristics that confer veracity on each dataset.

Conversely, Big Data is defined as being *too big* for consideration by a community of stakeholders (Wolfert et al., 2017). Big data possesses such an overwhelming degree of variety, velocity, and volume that human stakeholders alone cannot value it (NIST Big Data Public Working Group, 2015). However, while the system acts as the sole creator of value in *Big Data* sharing platforms, the algorithms, networks, and processes employed are initially determined by stakeholders in their choice of training datasets and causalities. Therefore, the proposed framework with its separation of individual classes of assets, resulting management tasks and singular goal of enabling valuable interactions, enables a novel approach to assessing how value may be created from both agricultural blockchains and Big Data produced by Smart Farming.

6. Conclusions and Extensions

The rapid growth in the role of data produced by Smart Farming has centralized the task of 'creating value from data' in agriculture. There are both complimentary assets and specific managerial decisions that must be accounted for in this process. Key decisions must be made by management regarding investment in assets or expenditure of effort on managing those assets if the value created from Smart Farming data is to remain efficient. Even those firms that are aware of some

elements of the proposed framework can benefit from the systems-based model and consequential trade-offs.

This paper proposes that creation of value from agricultural data must extend beyond just managing data and account for the community of stakeholders and a facilitatory, data-management system. Creation of value from data also requires consideration of how these assets interact; specifically, effort must be invested into the development of data, organization of the community, and allocation of value across the data sharing platform. This framework provides an economic explanation for the tension found in value allocation across data sharing platforms (Agyekumhene et al., 2020; Jakku et al., 2019) and also broadens the theoretical explanation for the sensitivity of platforms to the maintenance of congruous goals among stakeholders such as observed by Fleming et al. (2018). Trade-offs between the effort invested in building and managing those assets involve both first- and second-order considerations as interventions across the platform produce both direct and indirect impacts on the productive power of the assets.

The contributions of this paper are three-fold. First, we extend platform economics and Smart Farming data management theory to provide a data sharing platform framework that explains the assets and tasks required to create value from agricultural data. This framework proposes a change to the widely adopted view of a 'platform stack' which consists of "three distinct, operational 'layers'" (Langley & Leyshon, 2017, p. 17) and illustrates how these components are more appropriately represented as three interconnected asset classes existing on the same layer. Our framework also unifies the 'layered' view with previous work by Bonchek and Choudary (2013) in their preliminary work on successful platform strategies. We also model the operation of data sharing platforms by describing a process that takes data as one of three core asset classes and with appropriate management effort creates value from data. Extending Wolfert et al. (2017)'s review we show this management effort must be focused on (i) the development of data from its raw state to a condition that is useful for stakeholders, (ii) the organization of the community so their needs correlate with the data, and (iii) the allocation of value both between stakeholders in the community, and between the community and system. Collectively, these activities comprise the middle, managerial layer, and when applied to the underlying asset layer, create value from data as graphically represented in Figure 2-3.

Secondly, the proposed model informs the current literature regarding data ownership. Current academic opinion ranges from data as an asset that can be owned like any other asset (Jones & Tonetti, 2020) through to data as a medium where ownership is primarily concerned with the interventions brought to bear on the data (A. Fleming et al., 2018). This paper adopts a middle ground, proposing that data may be fungible and offered in exchange like a traditional asset but may

also consist solely of common observations that only gain value through proprietary development. Like others (Turland & Slade, 2020; Wiseman et al., 2019), we note the function of a data sharing platform is typically to provide economic access to – or exclusion from – data. Noting this confers data with the properties of a type of club good (E. Fleming, Griffith, Mounter, & Baker, 2018; Wysel, 2019), we propose a more nuanced view of data ownership should be adopted by scholars and management. If data ownership was couched in terms of access rights and right to appropriate the value created (see for instance, Birner et al. (2021)) then many of the apparent discrepancies are reduced to issues regarding application. Data sharing platforms offer a practical framework that aids this approach by identifying each component involved in the value creation process.

Our final contribution to theory is more practical. The data sharing platform model permits analysis of existing data management efforts. We qualitatively characterize the benefits and costs associated with the development, maintenance and management of the assets and discuss how both variable and fixed costs can be balanced by benefits accrued by data. We use this model to assess existing strategies commonly adopted within Smart Farming data management, and to discuss policy implications in the efficient management of Smart Farming data. Finally, we apply this model to evaluate first, second, and third-level strategies frequently adopted by data managers in their pursuit for the efficient creation of value from data. We also propose the context each of these strategies might be best suited to.

As noted throughout, this paper introduces key theories that deserve separate, considered attention. First, the production process that governs the conversion of data to value is introduced but not explored fully. While we describe the properties this process must possess to effectively create value from data, the specification of that process in this paper would greatly complicate the current analysis. Additional research into production processes that create value from data would permit answers to persistent, real-world questions like ‘how much – otherwise private – data should a farmer share with off-farm firms?’ or ‘what happens to value creation in the presence of an independent, profit-maximizing system?’

Secondly, demand-side economies of scale are well-established in information and platform economics literature and explain why platforms can become such dominant forces in their respective industries. We illustrate that these forces serve to unite stakeholders on a data sharing platform, but we only touch on how these forces result in greater value from data. We argue that consideration of demand-side economies of scale in the context of club goods will further elucidate the intrinsic similarities in data valuation and in data sharing platforms. Finally, the construct and operation of data sharing platforms sheds new light on research into management of both Big Data and Blockchain datasets. Both are quickly gaining prominence in Smart Farming and cutting-edge

agribusiness, and we have only introduced the way in which data sharing platforms reinterpret both conceptualizations.

The promise of the value created from data produced by Smart Farming is alluring but like any agricultural resource, keen management of its interaction with complimentary assets is vital if stakeholders are to enjoy the benefits. Embodied in a field of agricultural datanomics, an interdisciplinary agenda of targeted research and effective Smart Farming management techniques is required to improve the efficacy and efficiency of agriculture's efforts to create value from data.

References

- AgriDigital. (2021). Waypath. Retrieved from <https://www.agridigital.io/products/waypath>
- Agyekumhene, C., De Vries, J., Paassen, A. v., Schut, M., & MacNaghten, P. (2020). Making smallholder value chain partnerships inclusive: Exploring digital farm monitoring through farmer friendly smartphone platforms. *Sustainability*, *12*(11), 4580.
- Arrow, K. J. (1996). The economics of information: An exposition. *Empirica*, *23*(2), 119-128.
- Baker, D., Cook, S., Jackson, E. L., Wysel, M., Wynn, M., & Leonard, E. (2021). *Investment in Agri-Food Digital Transformation: Avoiding the technical fallacy*. Paper presented at the 65th Annual Conference of the Australasian Agricultural and Resource Economics Society, Sydney, Australia.
- Barrett, C. B. (2021). Overcoming global food security challenges through science and solidarity. *American Journal of Agricultural Economics*, *103*(2), 422-447.
- Birner, R., Daum, T., & Pray, C. (2021). Who drives the digital revolution in agriculture? A review of supply-side trends, players and challenges. *Applied Economic Perspectives and Policy*. doi:10.1002/aepp.13145
- Bonchek, M., & Choudary, S. P. (2013). Three elements of a successful platform strategy. *Harvard Business Review*, *92*(1-2).
- Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, *105*(7), 2183-2203.
- Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*(2), 171-209.
- Chesbrough, H. W. (2003). *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press.
- Choudary, S. P. (2015). Platform scale: how an emerging business model helps startups build large empires with minimum investment. Platform Thinking Labs.
- Christiaensen, L., Rutledge, Z., & Taylor, J. E. (2021). Viewpoint: The future of work in agri-food. *Food Policy*, *99*, 101963. doi:10.1016/j.foodpol.2020.101963
- Cook, S., Jackson, E. L., Fisher, M. J., Baker, D., & Diepeveen, D. (2021). Embedding digital agriculture into sustainable Australian food systems: pathways and pitfalls to value creation. *International Journal of Agricultural Sustainability*, 1-22.
- Darnell, R., Robertson, M., Brown, J., Moore, A., Barry, S., Bramley, R., . . . George, A. (2018). The current and future state of Australian agricultural data. *Farm Policy Journal*, *15*(1), 41-49.
- Daum, T., Villalba, R., Anidi, O., Mayienga, S. M., Gupta, S., & Birner, R. (2021). Uber for tractors? Opportunities and challenges of digital tools for tractor hire in India and Nigeria. *World Development*, *144*, 105480.
- Evans, D. S. (2009). How catalysts ignite: the economics of platform-based start-ups. *Platforms, Markets and Innovation*, 99-128.

- Evans, D. S., & Schmalensee, R. (2016). *Matchmakers: the new economics of multisided platforms*. Harvard Business Review Press.
- Evans, D. S., Schmalensee, R., Noel, M. D., Chang, H. H., & Garcia-Swartz, D. D. (2011). Platform economics: Essays on multi-sided businesses. *Platform Economics: Essays on Multi-Sided Businesses*, David S. Evans, Ed., *Competition Policy International*.
- Evans, P. C., & Gawer, A. (2016). *The rise of the platform enterprise: a global survey*. University of Surrey.
- Faulkner, A., Cebul, K., & McHenry, G. (2014). Agriculture gets smart: the rise of data and robotics. *Cleantech Agriculture Report*, Cleantech Group.
- Fleming, A., Jakku, E., Lim-Camacho, L., Taylor, B., & Thorburn, P. (2018). Is big data for big farming or for everyone? Perceptions in the Australian grains industry. *Agronomy for Sustainable Development*, 38(3), 1-10.
- Fleming, E., Griffith, G., Mounter, S., & Baker, D. (2018). Consciously pursued joint action: Agricultural and food value chains as clubs. *International Journal on Food System Dynamics*, 9(1012-2018-4116).
- Friess, P. (2016). *Digitising the industry-internet of things connecting the physical, digital and virtual worlds*. River Publishers.
- Gawer, A., & Cusumano, M. A. (2014). Industry Platforms and Ecosystem Innovation. *Journal of Product Innovation Management*, 31(3), 417-433. doi:10.1111/jpim.12105
- Gentzkow, M., & Kamenica, E. (2014). Costly persuasion. *American Economic Review*, 104(5), 457-462.
- Jakku, E., Taylor, B., Fleming, A., Mason, C., Fielke, S., Sounness, C., & Thorburn, P. (2019). "If they don't tell us what they do with it, why would we trust them?" Trust, transparency and benefit-sharing in Smart Farming. *NJAS-Wageningen Journal of Life Sciences*, 90-91, 100285.
- Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the South Indian fisheries sector. *The Quarterly Journal of Economics*, 122(3), 879-924.
- Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9), 2819-2858.
- Kieti, J., Waema, T. M., Ndemo, E. B., Omwansa, T. K., & Baumüller, H. (2021). Sources of value creation in aggregator platforms for digital services in agriculture-insights from likely users in Kenya. *Digital Business*, 1(2), 100007.
- Klerkx, L., Jakku, E., & Labarthe, P. (2019). A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. *NJAS-Wageningen Journal of Life Sciences*, 90-91, 100315.
- Lambert, D. M., & Cooper, M. C. (2000). Issues in supply chain management. *Industrial Marketing Management*, 29(1), 65-83.
- Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2), 132-157.
- Langley, P., & Leyshon, A. (2017). Platform capitalism: the intermediation and capitalization of digital economic circulation. *Finance and Society*, 3(1), 11-31.
- Leng, K., Bi, Y., Jing, L., Fu, H.-C., & Van Nieuwenhuysse, I. (2018). Research on agricultural supply chain system with double chain architecture based on blockchain technology. *Future Generation Computer Systems*, 86, 641-649.
- Nikander, J., Manninen, O., & Laajalahti, M. (2020). Requirements for cybersecurity in agricultural communication networks. *Computers and Electronics in Agriculture*, 179, 105776.
- NIST Big Data Public Working Group. (2015). *DRAFT NIST Big Data Interoperability Framework: Volume 1, Definitions*. NIST Special Publication. doi: 10.6028/NIST.SP.1500-1
- Nuthall, P. L. (2018). *Farm business management: the human factor*. CABI.
- Papazoglou, M. P., Ribbers, P., & Tsalgatiidou, A. (2000). Integrated value chains and their implications from a business and technology standpoint. *Decision Support Systems*, 29(4), 323-342.

- Parker, G., Van Alstyne, M., & Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. WW Norton & Company.
- Parker, G., Van Alstyne, M., & Jiang, X. (2017). Platform ecosystems: How developers invert the firm. *MIS Quarterly*, *41*(1).
- Rochet, J. C., & Tirole, J. (2003). Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, *1*(4), 990-1029.
- Rojo-Gimeno, C., van der Voort, M., Niemi, J. K., Lauwers, L., Kristensen, A. R., & Wauters, E. (2019). Assessment of the value of information of precision livestock farming: A conceptual framework. *NJAS-Wageningen Journal of Life Sciences*, *90-91*, 100311.
- Sanderson, T., Reeson, A., & Box, P. (2017). Understanding and unlocking the value of public research data.
- Sandler, T., & Tschirhart, J. (1997). Club Theory: Thirty Years Later. *Public choice*, *93*(3-4), 335-355.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379-423.
- Shapiro, C., & Varian, H. R. (1998). *Information rules: a strategic guide to the network economy*. Harvard Business Press.
- Sims, C. A. (2003). Implications of Rational Inattention. *Journal of Monetary Economics*, *50*(3), 665-690.
- Turland, M., & Slade, P. (2020). Farmers' willingness to participate in a big data platform. *Agribusiness*, *36*(1), 20-36.
- Van Alstyne, M., Parker, G., & Choudary, S. P. (2016). Pipelines, platforms, and the new rules of strategy. *Harvard Business Review*, *94*(4), 54-62.
- Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.
- Venkataraman, A. (2016). How Do you Hail a Tractor in India? All it Takes Is a Few Taps on your Phone. Retrieved from <https://www.nytimes.com/2016/10/18/world/what-in-the-world/tringo-appindia.html>
- Wiseman, L., & Sanderson, J. (2019). *Agricultural Data Rules: Enabling Best Practice*. Retrieved from <https://www.crdc.com.au/sites/default/files/Fact%20Sheet%20Growing%20digital%20-%20Data%20Rules%20Sep2019.pdf>
- Wiseman, L., Sanderson, J., Zhang, A., & Jakku, E. (2019). Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. *NJAS-Wageningen Journal of Life Sciences*, *90-91*, 100301.
- Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming—a review. *Agricultural Systems*, *153*, 69-80.
- Wysel, M. (2019). *Using Platforms to Operationalize the Valuation of Information in the Red Meat Supply Chain*. Paper presented at the Australian Agricultural and Resource Economics Society Annual Conference, Melbourne, Australia.
- Wysel, M. (2021, 22 June 2021). *Data Sharing Platforms in Agribusiness*. Paper presented at the International Food and Agribusiness Management Association, Costa Rica.
- Wysel, M., Baker, D., & Conway, L. (2019). *Connecting Researchers with Customers through a Common Valuation of Data*. Paper presented at the Australian Agricultural and Resource Economics Society Annual Conference, Melbourne, Australia.

Appendix. Derivations

Equation (2-3) gives the net value created across the data sharing platform for each interaction. Our goal is to find the effect that each of Wolfert et al. (2017)'s managerial functions has on this 'per

interaction' value creation. Considering the effect of data development, first we find the partial derivative of Equation (2-3) with respect to ε ,

$$\frac{\partial^2 \mathbb{V}}{\partial k \partial \varepsilon} = \frac{\partial}{\partial \varepsilon} \{ \mathbb{B}_{C_k}(\rho_k, \mu_k) + \mathbb{B}_{D_k}(\rho_k, \varepsilon_k) + \mathbb{B}_{S_k}(\varepsilon_k, \mu_k) \} - \frac{\partial}{\partial \varepsilon} \{ \mathbb{C}_{C_k} + \mathbb{C}_{D_k} + \mathbb{C}_{S_k} \}.$$

Now, functions in the second pair of braces are constant w.r.t ε and disappear. By the chain rule we have,

$$\frac{\partial^2 \mathbb{V}}{\partial k \partial \varepsilon} = \left(\frac{\partial \mathbb{B}_C}{\partial \rho} \frac{\partial \rho}{\partial \varepsilon} + \frac{\partial \mathbb{B}_C}{\partial \mu} \frac{\partial \mu}{\partial \varepsilon} \right) + \left(\frac{\partial \mathbb{B}_D}{\partial \rho} \frac{\partial \rho}{\partial \varepsilon} + \frac{\partial \mathbb{B}_D}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right) + \left(\frac{\partial \mathbb{B}_S}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} + \frac{\partial \mathbb{B}_S}{\partial \mu} \frac{\partial \mu}{\partial \varepsilon} \right)$$

which may be re-arranged to Equation (2-4).

The impact of community organization, ρ , and value allocation, μ , on the value created by the data sharing platform at each interaction proceed by the same method.

Higher Degree Research Thesis by Publication

University of New England

Statement of Authors' Contribution

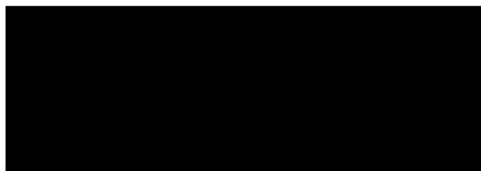
We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated in the *Statement of Originality*.

	Author's Name	Percentage of Contribution
Candidate	Matthew Wysel	80
Other Authors	Derek Baker	16
	William Billingsley	4



Candidate

Date



Principal Supervisor

Date

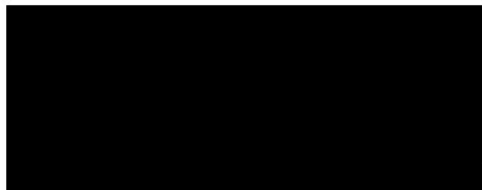
Higher Degree Research Thesis by Publication

University of New England

Statement of Originality

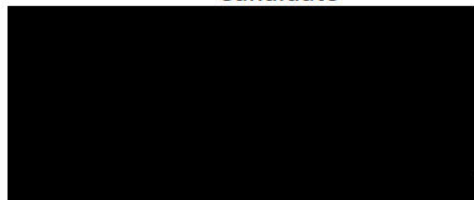
We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that the following text, figures, and diagrams are the candidate's original work.

Author	Type of Work	Page Numbers
Matthew Wysel	Conceptualization, methodology, formal analysis, writing (original draft), writing (review), editing.	Entire Document
Derek Baker	Conceptualization, writing (review), editing, supervision.	Entire Document
William Billingsley	Conceptualization.	Entire Document



Candidate

Date



Principal Supervisor

Date

Chapter Three:

How to Value Data: Uniting Economic and Information Theory to Create a Value Framework for Data.

Overview of Manuscript

Status:	Unpublished manuscript
Date Produced:	22 June 2020
Date Published:	-
Suggested Citation:	Wysel, M., Baker, D., & Billingsley, W. (2020). How to Value Data: Uniting Economic and Information Theory to Create a Value Framework for Data [Unpublished manuscript]. University of New England Business School.

Summary of Paper in Context of PhD:

Where Chapter 2 used platform economics and studies in Big Data to established data sharing platforms, Chapter 3 uses seminal work from communication theory and the economics of information to explain how data sharing platforms operate, and therefore, how value is created from data.

This manuscript maintains the value equation from Chapter 2 and following Arrow's (1996) exposition of the economics of information specifies a united economic expression for the operation of a data sharing platform. Shannon's (1948) communication theory is used to explain how interactions between the community and system cause an enrichment of data and accretion of value.

Finally, the manuscript joins the economic expression with the mechanics of the data sharing platform to explain how data is valued in use, in exchange, and subsumed as a value-carrying currency. Chapter 4 brings these apparently disparate valuation modes together to explain how value is created from data at a firm level. Chapter 5 extends this theory to a marketplace by assessing a real-world, data sharing ecosystem. Chapter 6 assesses the consequences to the process of creating value from data when operation of the data sharing ecosystem is partially impaired.

Apart from typesetting changes and language localization, this chapter appears exactly as submitted to *Management Information Systems Quarterly (MISQ)* in June 2020. Based on feedback received from the First-Round review on 14 November 2020 we decided to pause development of this manuscript and build stronger connections with extant literature. That work became the paper contained in Chapter 2.

Supplementary Publications

Research that informed this chapter also appeared in the following outputs.

Type	Citation
Conference Paper	Wysel, M., Baker, D., & Conway, L. (2019). Connecting Researchers with Customers through a Common Valuation of Data. 63rd Annual Conference of the Australasian Agricultural and Resource Economics Society, Melbourne, Australia.
Symposium Presentation	Wysel, M. (2019). Using Platforms to Operationalize the Value of Information in the Red Meat Supply Chain. Symposium: Information in the Australian Red Meat Supply Chain, UNE Centre for Agribusiness.
Conference Paper	Burgess, S., & Wysel, M. (2020). <i>China's Social Credit System: How Robust Is the Human Rights Critique?</i> 27th Annual Australian Association for Professional and Applied Ethics, The University of New England.

Chapter-level Glossary

Term	Definition
Entropy	<i>Noun.</i> 'Shannon Entropy', referred to simply as entropy, H , is the sum of the weighted probabilities that a dataset consists of relevant observations.

Full Manuscript

How to Value Data: Uniting Economic and Information Theory to Create a Value Framework for Data.

Authors:	Matthew Wysel ^a (matthew.wysel@une.edu.au;)*, Derek Baker ^a (derek.baker@une.edu.au), William Billingsley ^b (wbilling@une.edu.au)
Affiliations:	^a The Centre for Agribusiness, UNE Business School, The University of New England, Armidale, Australia. ^b Computational Science, The University of New England, Armidale, Australia.
Keywords	Data valuation, club theory, data enrichment, value framework, theory of information, theory of communication, search theory, economics of information.
Acknowledgements:	The authors wish to thank Garry Griffith, Daniel Gregg, and Phil Simmons for their helpful suggestions particularly in refining the manuscript for publication. Matthew Wysel is grateful for the support of the Agricultural Business Research Institute through the Arthur Rickards Innovation in Agribusiness Scholarship. All omissions remain our own.

Abstract

While the world drowns in data, the valuation process that surrounds data remains a poorly understood phenomenon. Existing models conflate the value of data with what can be, or has been done with the data, or have pre-supposed the relevance of data. We propose a framework useful for assessing the data valuation process independent of its context or surrounding business model. This framework accounts for both uncertainty in relevance and utility of data. By reversing the direction of Shannon's communication system, we demonstrate how the coupling of uncertainty with the value of information originates as an intrinsic characteristic of data. This process is applied to explain how data creates value as a resource, as a good, and as a currency. Contemporary examples are used throughout to illustrate both the theory and implications of these three constructs. Management and maximization of the value of data is discussed. An agenda for future research is presented.

1. Introduction and Review of Literature

The world is awash with data but generally at a loss with how to value it (Grover, Chiang, Liang, & Zhang, 2018; Gupta, Kannan, & Sanyal, 2018). A recent survey of 36 companies and non-profit organizations across North America and Europe, many with turnovers greater than USD1 billion, revealed most had no formal data valuation practices but any existing valuation efforts were time-consuming and complex and that data management was focused on "storing, protecting, accessing, and analyzing massive amounts of data" (Short & Todd, 2017, p. 17). Compounding the problem, the growth rate of data generated and stored now outpaces Moore's Law for the growth rate of computation (Chang & Boyd, 2018). Where the chief task of information systems was once to amass information against predominately defined questions, systems are increasingly required to acquire, process and report on data that exhibits both uncertain relevance and utility. Indeed, these complex datasets often produce greater operational uncertainty before yielding solutions and insights (Chiang, Grover, Liang, & Zhang, 2018; Ketter, Peters, Collins, & Gupta, 2015). If ever the practice of valuing data like a standard accounting asset¹¹ was tenable, it is quickly becoming both operationally and technically infeasible.

Clearly data *has* value. Data is the chief asset for intermediary businesses, as the exchangeable good in both ad hoc and well-established marketplaces, and as a currency used by individuals and firms to offset real-world costs. However, a rigorous, extensible framework that explains how data amasses

¹¹ That is, the ex-post process of record, categorize, summarize, report.

value is required to explain these phenomena in terms of the central variable of interest: data. In this paper, we propose a generic framework for valuing data that is independent of an operational setting or business context. This framework permits the extension, and more granular application of extant research. For instance, the value of the internal and external ecosystems of firms (G. Parker, Van Alstyne, & Jiang, 2017) might now be measured by the relative value of data contained within each system. The impact of signals produced by a platform (Hukal, Henfridsson, Shaikh, & Parker, forthcoming) could be observed as a change in the data, the fundamental resource which is controlled by the platform. Finally, the marginal change in privacy risk to consumers (Adjerid, Peer, & Acquisti, 2018) might now be measured directly when a firm begins collection of new forms of data.

The popular approach for valuing data is to adapt market valuations of comparative datasets. These events may be acquisitions or sales of data assets, mergers and acquisitions, insurance valuations or bankruptcy filings. While market valuations provide an important, external reference point for the value of a dataset, they naturally reflect the composite value of business operations with their controlled assets such as data. Without an underlying framework for value appropriation that permits assessment of the data independent of the surrounding processes, this approach will continue to produce comparisons that are subjective and contentious.¹²

The problem of valuing data is not limited to companies and markets. Individual consumers trade access to their raw social, health, biological or behavioral data for subsidized services. Data ceded by customers enables whole ecosystems whose byproduct and sometimes goal, is the transformation of data from a 'raw' state to a state of greater value. Yet, what portion of this value could be considered consumer surplus, producer surplus, or is retained as latent value within the data, remains unapproached.

While there is evidently a method for deriving value from data, the generic model remains poorly understood. As scholars, we ought to be able to answer fundamental questions about the nature and composition of the data valuation process. What constituents must be present for data to possess value? What characteristics of data dictate its ordinal value, and by what process is the value of data altered? Finally, what are the generic modes and their properties by which data may be

¹² Consider Microsoft's purchase of LinkedIn (Microsoft, 2016) and Moody's response (2016) with discussion by Short and Todd (2017). Additionally see Damodaran (2014b) and Gurley (2014) disputing and later partially resolving (Damodaran, 2014a) the value of Uber, with discussion by G. G. Parker, Van Alstyne, and Choudary (2016).

valued? This paper proposes timely responses to each of these questions and proposes a rigorous but extensible framework for understanding the value of data.

We approach this framework by first compiling an economic model that describes an agent's valuation of data. We then broaden this model to define the accompanying transformation process that happens within the data when an agent seeks its enrichment. Finally, the model and process are combined to demonstrate how data may be valued as a resource, as a good, and as a currency. Throughout, value is understood as the difference between benefits accrued and costs incurred and can be amassed by an individual or community. Valuation is a process where uncertainty regarding an object's worth is recursively reduced until the marginal utility of further evaluation is no longer positive. Club theory (for example, Sandler & Tschirhart, 1997) provides the overarching framework in which individual self-seeking agents form communities and collaborate to maximize their personal payoffs, even in the presence of rivalry and potential exclusion. Clubs offer members benefits not available outside the club and may extract tolls from their members based on their activity or identity. Therefore, agents will naturally gravitate toward communities offering data valuation services that either lower their marginal cost of valuing data or increase their potential benefit from valued data. In both cases, the result is an increase in value to the agent from their data.¹³

While a more extensive review of the economic treatment of data is included in Appendix A, the first step in developing a valuation framework for data is to establish an economic model that permits the benefits and costs of valuing data to be observed. To this end, we extend economics of information (Arrow, 1985, 1996) and rational inattention (Sims, 2003) literature to establish a model that describes a generic data valuation process. Based on agents as the core unit, this model describes the accumulation of costs and realization of benefits as agents seek value from transformed data. Access to a community that shares a congruous, data-related goal permits participation in processes that reduce uncertainty from data. The first result of this paper is to extend Arrow's treatment of information to a model that accounts for intrinsically uncertain data where agents work within a community to resolve relevance as well as maximize the utility of data.

In line with extant platform economics literature (D. S. Evans, 2003; D. S. Evans & Schmalensee, 2016; P. C. Evans & Gawer, 2016; G. G. Parker & Van Alstyne, 2002; Rochet & Tirole, 2003), we refer

¹³ Consistent with standard economic treatment, agents are treated as individual decision-making entities. These entities are commonly thought of as individual persons in a marketplace. However, consolidated decision-making bodies such as firms, and synthesized decision-making bodies such as nodes in a network also function as agents. Such a treatment permits the framework developed in this paper to remain widely extensible, even to hybrid systems – such as smart prosthetics – whose activities effect a change in the provision or value of data.

to the community's organized, data-related efforts as a *system*. The system represents the joint efforts of all agents – both past and present – as they value data. We refer to the system, the community of agents, and the data, as a *data sharing platform*. The mechanism that describes the interaction of agents with the data sharing platform is central to an understanding of the way the value of data is altered. This mechanism is reflexive as while seeking value from data, agents also accrue value to the data. Reversing the direction of Shannon's (1948) communication system permits examination of the change in the characteristics of data that affect value. Agents act as Shannon's external observer¹⁴ and interact with the system as they reduce uncertainty from data. The efficacy of each interaction is described by an enrichment efficiency, the central variable that governs the rate of change in value at each exchange. The second result of this paper is that the valuation of data, no matter how noisy, may now be explained. Our third result is to show that the reduction of the intrinsic uncertainty in data is the sole necessary and sufficient condition for the accretion of value to that data. This result illustrates that the coupling of uncertainty and the benefit of information derived by Frankel and Kamenica (2019) exists even while the relevance of data remains undefined. Importantly, the relative reduction in the uncertainty in data serves as a rationale for apportioning value between parties in data sharing platforms. This informs the 'property rights' models currently being considered (Jones & Tonetti, 2018) and permits industry bodies to design incentives for data sharing into value chains (Fleming, Griffith, Mounter, & Baker, 2018).

Finally, the combined data valuation model and data enrichment process describes the generic modes by which the value of data may be managed. Data may be valued as a resource, as a good, and as a currency.¹⁵ Drawing on contemporary examples, we apply the data valuation framework to explain how a community of agents and system seek value from, and attribute value to, data.

The analytical arc of this paper centers on an idealized scenario of agents in an ad hoc community. We posit three simplifying assumptions which are relaxed in turn permitting the introduction of greater sophistication in the data valuation framework. Crucially, the conditions that describe each mode are properties of the data valuation process and not a broader business model or organizational framework. Effective business models may now be understood in terms of which data valuation mode they employ, rather than contrariwise.

¹⁴ See *Figure 8 – Schematic diagram of a correction system*, Shannon (1948)

¹⁵ The authors acknowledge that a resource and currency are simply different projections of an economic good. Nevertheless, the distinction has been adopted as each projection reveals different insights into how the value of data may be managed.

The resulting data valuation framework explains the shared reliance and rivalry often observed between agents and providers in commercial data sharing platforms. The deliberately straightforward final equation permits observation of value creation and exchange through data by a community of agents and the system that facilitates their efforts. Indeed, the final data valuation framework may be expressed simply as the piecewise difference between the benefits and costs experienced by each constituent in a data sharing platform.

The proposed valuation model, enrichment process and framework permit both granular and high-level examination and management of the value of data. This understanding is useful for data-driven businesses, large corporate information systems, or in personal interactions with popular ecommerce or social media platforms. The mechanism by which ostensibly trivial data may be valued against fiat currency can now be understood, while questions like, ‘how ought I value my data?’ may now be approached by scholar and practitioner alike. Future work includes the mathematical adaptation of this framework to accommodate fully dynamic data streams. We conclude with discussion and implications for both scholars and professionals.

2. The Data Valuation Model

2.1. Set-up: Data and the Generic Valuation Process

The goal, or “desideratum” (Frankel & Kamenica, 2019, p. 3650), of an agent is the primary determinant of an object’s value. The goal sets the direction of value and directs all activities that encompass the ensuing valuation process. A goal of maximizing personal utility leads consumers to value the characteristics of goods while a goal of realizing valuable exchanges directs agents to participate in single- or multi-sided markets. A goal is a form of hypothetical imperative whose origin is subjective but remains testable and, therefore, universal and rational across its specified scope.¹⁶ Valuation becomes the process of taking an ex-ante goal and assembling relevant observations to evaluate the worth of an object against that goal. Observations are collected and refined to facilitate a testing process that reduces uncertainty regarding the performance of the object. This testing process is recursive as responses to historic observations affect management of current observations. In this manner, reduction in uncertainty surrounding the object will continue until performance is validated, or the process is abandoned as economically non-viable. This generic valuation process is illustrated in Figure 3-1.

¹⁶ See Kant (1870). *Grundlegung zur metaphysik der sitten*. Vol. 28: L. Heimann.

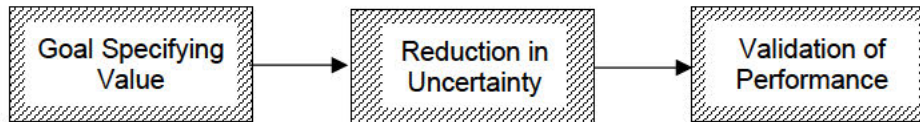


Figure 3-1. Generic valuation process

To pick a simple physical example suppose two commuters survey a train carriage for empty seats. Both share the goal of sitting in an empty seat and scan for empty seats accordingly. Both commuters use data on seat vacancy to reduce the uncertainty of potential seats and upon finding empty seats both commuters would take them. This process holds for competitive scenarios too. Assuming no societal impediment, both would move towards a single empty seat where the first commuter would sit, forcing the other to continue the process of evaluating remaining seats to determine their vacancy.

Here, commuters share a goal and value data as a resource that permits attainment of that goal. Each commuter has implicitly determined that the benefit of finding a vacant seat exceeds the associated search cost. Suppose for illustration, commuters could not evaluate seats with near zero cost but were required to trial-sit in each seat; the search cost would be significantly higher, while the benefit would remain largely unchanged. Fewer seats would be evaluated. Therefore, the value of the data to an individual who possesses a related goal is the benefit arising from the valuation minus the cost of the same. More formally,

$$V = B - C \quad (3-1)$$

where, V , B and C are the value, benefits and costs arising from the valuation, respectively.

Equation (3-1) holds where data is valued in use and in exchange. Suppose a third commuter sold observations regarding seat vacancy to the first. The exchange price confers benefit to the third commuter and cost to the first whose benefit must have exceeded that cost for the exchange to take place. This introduces an important corollary: the mere availability of data is not sufficient to occasion its valuation. The expected benefit of valuation must exceed the associated cost if the process is to proceed, and the distribution of value may be visualized as the value flow model in Figure 3-2.



Figure 3-2. Initial value flow model

To facilitate a more rigorous investigation we propose the following scenario:

There exists a set of data of indeterminate value and a number of agents who form an ad hoc community across that data; the number of datasets considered by each agent is not necessarily constant across that community, nor is the size of each dataset necessarily constant across all available data.

We may also define a *data valuation community* as a club permitting exclusion from, and shared access to, the value of data that has been contributed by agents who volunteer membership because they anticipate a net benefit. Just as club theory remains valid – if uninteresting – for clubs of a single member, data valuation communities remain useful constructs – if less interesting – even with only one member. As subsequently developed, while data is non-rivalrous, rivalry persists in this club as parties compete for the value of data. To permit efficient analysis, let us also set the following supplementary assumptions:

- (A1) The data contains only observations which are necessary and sufficient for achieving the data-related goal.
- (A2) The agents' preferences converge on, and are united around, a common data-related goal.
- (A3) The community directs the valuation process and participates in the distribution of the resulting value.

Each assumption will be relaxed in turn as we examine the valuation of data under regimes of increasing sophistication. For our present purposes, it is sufficient to note that assumptions *A1*, *A2* and *A3* guard against process-based costs arising from noisy or incomplete data; preference related inefficiencies; and structural friction in the data valuation process, respectively.

Therefore, with all assumptions standing, we may adapt Equation (3-1) and say the value of data is the benefit to the community resulting from the valuation of the data less the associated cost of the same. More formally,

$$V = \mathbb{B}_C - C_C \quad (3-2)$$

where, \mathbb{B}_C is the benefit to the community resulting from valuation of the data and, \mathbb{C}_C is the cost to the community resulting from valuation of the data.

2.2. The Model: The Economic Valuation of Data

Before relaxation of any supplementary assumptions, an important corollary of the data valuation process is worth developing. The realization of benefits by agents and concession of costs is not, in general, a binary exchange occurring at an instant of time. Intuitively, mid-way through the data valuation process agents will have incurred some costs and have accumulated some form of benefit.¹⁷ Mirroring the valuation process described above, we may say *access* by agents to a likeminded community indicates their adoption of a congruous goal, their *participation* in that community represents their efforts to reduce uncertainty surrounding that goal, and their *validation* as the culmination of the valuation process. Therefore, expanding the right-hand side of Equation (3-2) we have,

$$\mathbb{B}_C = \mathbb{B}_{C_{access}} + \mathbb{B}_{C_{participation}} + \mathbb{B}_{C_{validation}} \quad (3-3)$$

$$\mathbb{C}_C = \mathbb{C}_{C_{access}} + \mathbb{C}_{C_{participation}} + \mathbb{C}_{C_{validation}} \quad (3-4)$$

where the value of data to the community of agents is the difference between the sum of the benefits arising from that data less the sum of the associated costs.

Proposition 1: Agents maintain membership in a data valuation community to access value from that data not attainable elsewhere.

Here *access* is an instantaneous change in status occurring when an agent joins the community. The benefit to an agent arising from access to a data valuation community, $\mathbb{B}_{C_{access}}$, takes the form of a probability that the data desired by the agent is accessible by the data sharing community. To facilitate subsequent treatment, let the data desired by the agent be the set, x , and the data presided over by the community be the set, y . Accordingly, where both x and y are random variables, there also exists a specific probability, p , that $x \subset y$. Following Arrow (1996), the agent has a choice of at least two actions: to join the community or not, denoted by a , from a given opportunity set, A . Therefore, in the absence of further information about the community or data, the payoff¹⁸ to the agent of accessing the community is a function,

¹⁷ The instantaneous realization of value remains a valid corner case of the general, continual case and is contained in the subsequent derivation.

¹⁸ This payoff, and those that follow in Section 2, should be interpreted as von Neumann-Morgenstern utilities and not as monetary outcomes.

$$w(a, p(x \subset y))$$

Now, suppose the community can produce some random variable, S , that reveals to the agent the probability that $x \subset y$. We refer to S as a signal, of the type used in communication theory (Shannon, 1948). Now, the community can produce a particular $s \in S$ with a goal to influence prospective agents' choice of actions (Hukal et al., forthcoming) but creation of this signal bears a cost to the community (Gentzkow & Kamenica, 2014). Meanwhile, choice of this signal bears a cost to the prospective agent (Sims, 2003), as does the action of responding to the signal by joining the community. The cost of access to the community, $\mathbb{C}_{C_{access}}$, is therefore a function of the agent's action, a , itself a function of the signal chosen, such that $a(s)$. However, there is also no guarantee that access to x will result in a positive payoff for the agent implying that the agent must also account the probability, q , that x will produce a positive payoff. Therefore, as production and exchange of the signal is costly for both parties, notwithstanding signaling game strategies, the signal produced by the community becomes a function of p and q ,

$$S = f(p, q) \tag{3-5}$$

and a prospective agent will choose a signal and action that maximizes the expectation, E_q , of the payoff given the signal for the associated cost. In symbols,

$$\begin{aligned} \mathbb{V}_{C_{access}} &= \mathbb{B}_{C_{access}} - \mathbb{C}_{C_{access}} \\ &= E_q[w(a, p(x \subset y), s)] - \mathbb{C}_C(a(s)) \end{aligned} \tag{3-6}$$

Extending this analysis, *participation* proceeds likewise but incorporates the recursive nature of the data valuation process and therefore includes a dynamic variable for both the realization of benefits and allocation of costs. If \hat{x} is the data already found to be relevant, then the benefit of participation in the data valuation community, $\mathbb{B}_{C_{participation}}$ comprises the utility of the datasets already considered and the expected payoff that the remainder of desired data, $x - \hat{x}$, exists in the data yet to be considered, y . Similarly, cost of participation in the community, $\mathbb{C}_{C_{participation}}$ will be a function of the quantity of datasets considered and will carry an opportunity cost for attentive participation. Therefore, noting $q \in Q$ and $p \in P$, where M_j is the j th dataset valued to a maximum, m datasets, an agent will seek to maximize,

$$\begin{aligned} \mathbb{V}_{C_{participation}} &= \mathbb{B}_{C_{participation}} - \mathbb{C}_{C_{participation}} \\ &= E_Q[w(a, P((x - \hat{x}) \subset y), s)] + u_Q(\{M_j\}) - \sum_{j=1}^m \mathbb{C}_C(M_j) - \mathbb{C}_C(v(t)) \end{aligned} \tag{3-7}$$

Equation (3-7) specifies the marginal utility of all datasets considered and is of the usual concave form as the entire dataset remains of potential value. However as set forth by Sims (2003), data

cannot be considered with impunity as it remains rational to withhold attention from some data as any attention incurs a non-trivial cost. Therefore, even if the present datasets offer no utility the cost of attentive participation, $\mathbb{C}_C(v(t))$, and the cumulative cost of evaluating datasets remain significant¹⁹ and monotonically increasing as the most recent dataset always requires consideration.²⁰

Finally, each agent validates relevant data against their goal. Like access, *validation* is taken as an instantaneous change in status signaling that participation has ceased and satisfaction regarding the considered data has occurred. Validation of considered datasets confers a benefit to the agent proportional to the utility of the data evaluated and produces a signal, \acute{s} , to the community that,

$$E_Q[\acute{x}] > 0.$$

Production of this signal carries a cost to the agent and is useful to the community as it validates the process to both participating and prospective agents. Therefore, during validation agents will seek to maximize,

$$\begin{aligned} \mathbb{V}_{C_{validation}} &= \mathbb{B}_{C_{validation}} - \mathbb{C}_{C_{validation}} \\ &= u_Q(\acute{x}) - \mathbb{C}_C(\acute{s}) \end{aligned} \quad (3-8)$$

Finally, while unnecessary for our current proof but useful for subsequent discussion we can formally recognize the signal to a prospective agent, s , is at least partially derived from signals produced by satisfied agents, \acute{s} , by noting $s \in S(\acute{s})$ and therefore extending (3-5) we may write,

$$S = f(p, q, \acute{s}) \quad (3-9)$$

Combining Equations (3-6) – (3-9) into (3-3) and (3-4) while noting both x and y are corner cases of \acute{x} and \acute{y} respectively, the benefit and cost of valuing data to the i th agent becomes,

$$\mathbb{B}_{C_i} = E_Q[w(a, P((x - \acute{x}) \subset \acute{y}), S)] + u_Q(\{M_j\}) \quad (3-10)$$

$$\mathbb{C}_{C_i} = \mathbb{C}_C(a(s), v(t), \acute{s}) + \sum_{j=1}^m \mathbb{C}_C(M_j) \quad (3-11)$$

and therefore, in a community of n agents,

¹⁹ These cost functions have been variously specified elsewhere (Caplin & Dean, 2015; Sims, 2003). Our purpose is to capture the nature of these costs; specification of their function remains important, complimentary work.

²⁰ There is also an additional, small storage cost each agent must bear which is also expressed as an argument within the probability function, $P(\acute{x} \subset \acute{y})$.

$$V = \sum_{i=1}^n (\mathbb{B}_{C_i} - \mathbb{C}_{C_i}) \quad (3-12)$$

Substituting (3-10) and (3-11) into (3-12) gives us an idealized equation for the non-market value of data to a community of agents where $n \geq 1$.

Therefore, membership in a data valuation community increases an agent’s expected payoff from the data, permits greater utilization of that data, reduces the direct or indirect cost of acquiring the utility, or reduces the marginal cost of valuing the data. In all cases, the value of data to an agent – the resulting benefits less costs – is increased precisely because of an agent’s membership in a data valuation community.

2.3. The System and the Data Sharing Platform

As noted in the Introduction, in line with extant platform economics literature, we define the community’s organized, data-related efforts as a *system*. The system is a “set of assets organized in a common structure” (Gawer & Cusumano, 2014, p. 2) and includes both the infrastructure required to process data (Choudary, 2014) and the ongoing operations required to circulate value (G. Parker et al., 2017; Van Alstyne, Parker, & Choudary, 2016) among the community (Langley & Leyshon, 2017). Agents interact with the system, providing data to, and for, the system which processes the data for agents in the community.

Collectively, the community of agents and system may be considered a data sharing platform. Data sharing platforms are a generalization of Rochet and Tirole’s (2003) socio-economic platforms that exist in multi-sided markets.²¹ We may therefore define,

A data sharing platform is a community of agents, data on and for those agents, and a system that utilizes the data to enable the community to make more valuable decisions.

At this stage in our treatment of the valuation of data, the system exists solely to discharge the community’s data-related efforts and agenda. Nevertheless, separation of communal data-related efforts from individual agent’s efforts permits capitalization of system expenses and the appropriation of proportional data-related costs rather than the aggregation of an individual’s marginal costs. Therefore, the system incurs a cost, denoted \mathbb{C}_S , as it identifies noise and reduces uncertainty surrounding the community’s data. \mathbb{C}_{C_i} remains the sum of each i th agent’s individual

²¹ That is, data sharing platforms support, but are not defined by, multi-sidedness within the community.

costs as they evaluate the performance of data towards their own goal. In other words, \mathbb{C}_S creates scarcity in data while \mathbb{C}_C executes its valuation.

We now turn to examine the economic characteristics of data that affect its value and the activities required to mobilize this value.

3. The Data Enrichment Process

3.1. The Entropy of Raw Data

Equation (3-12) reveals the value of data acquired and evaluated by a community for validation against a congruous goal. Substantiating our previous assumption, data is a non-rivalrous resource (Jones & Tonetti, 2018) that is optionally excludable (Easley, Huang, Yang, & Zhong, 2018) and can therefore be treated like an imperfect public good (Fleming et al., 2018). Therefore, the value of a non-rivalrous good such as data can be summed vertically as one agent's consumption does not impinge another's consumption of the same good. *Ceteris paribus*, n agents deriving the same benefit from a single observation and incurring identical costs produce the same value as one agent valuing an observation n times. We see this reflected in Equations (3-10) and (3-11), above, and will return to treat this formally. Therefore, we can also say while repeat valuations by one agent will undoubtedly exhibit different technical characteristics to single valuations by multiple agents, they can be treated as economic equivalents at least to the extent enforced by the supplementary assumptions A1-A3.

In the preceding analysis we presumed that the desired data, x , was freely available to the community, that is $x \subset y$, and the marginal cost was focused on identification, rather than extraction, of x from y . While possible, this is certainly not a typical occurrence for desired data. Data that is of interest typically requires extraction from host or 'raw' data where data must be separated from noise and inherent uncertainty. Relaxing the first supplementary assumption A1: *The data contains only observations which are necessary and sufficient for achieving the data-related goal* the data available to the community is no longer equivalent to the data required by the community.²² Effort must be invested into the available data to enable agents to test their shared goal and access benefits.

²² A very large portion of the theory developed in this article relies on the relaxation of only A1. Assumptions A2 and A3 serve to broaden the application of this theory. Readers scanning the paper and finding the treatment of A2 and A3 near the end of the paper need not despair.

Proposition 2: The valuation of data is given by the reversal of Shannon’s communication process as agents systematically separate irrelevant data from relevant data within a data sharing platform.

Shannon (1948) pioneered the academic treatment of relative information levels across multiple stations in both ‘noiseless’ and ‘noisy’ communication channels. In the original direction, Shannon’s process models noise accretion around a central message with the consequential growth in uncertainty. When reversed, the adapted process permits modelling of the transformation of raw data by the recursive separation of noise and reduction in uncertainty. ‘Shannon Entropy’, henceforth referred to simply as entropy, H , may be measured per symbol or per second, and can be understood as the sum of the weighted probabilities that a dataset consists of relevant observations. Datasets and observations can be discrete or continuous and be of any length.

Entropy describes the probability of relevant observations;²³ the value of these relevant observations is determined by each agent’s goal and given by Equations (3-10) and (3-11), above. Agents may choose observations each with \hat{n} possibilities such that the probability of the i th observation containing an observation relevant to a specific goal is:

$$p_i = \frac{\hat{n}_i}{\sum_i \hat{n}_i}$$

and following Shannon’s (1948, Appendix 2) derivation, the entropy per dataset is:

$$H = -K \sum_i p_i \log p_i$$

where K is a positive scalar to account for chosen units and the base of the logarithm is the number of ‘observation-storing’ states each data-point could contain.²⁴

While this representation implies the nature of the random variable is discrete, rather than continuous, as observations can be discrete or continuous and be of any length the process of

²³ As entropy quantifies the inherent uncertainty in the data it also reflects the carrying capacity of relevant observations within the data. While greater entropy in a dataset permits a larger number of potentially relevant observations, this increase in capacity comes at the expense of certainty regarding those observations. Some observations will remain superfluous and be discarded as noise; which observations are relevant is the cause of uncertainty in the data. To this end, structural restrictions can be applied to the data to limit the impact of noise or uncertainty by proportionally reducing the entropic capacity of data. For example, in a dataset of strings comprising alphanumeric characters, any non-alphanumeric characters could be automatically identified and discarded as noise. However, by the same logic the potential capacity of the strings has also been reduced as intentional, non-alphanumeric characters and would require transcoding before transmission or storage.

²⁴ For example, a binary process would be represented in base-2.

developing data can be approached in the same manner. To simplify subsequent treatment the data development process will henceforth be treated as continuous.

In the absence of noise, entropy of the desired data, $H(x)$ remains equal to the entropy of the available data, $H(y)$.²⁵ This is our original data valuation scenario with all assumptions standing: all data required to evaluate the goals of agents has been supplied; no data was superfluous. Notice the process is isotropic: the entropy of the desired data, $H(x)$, equals the entropy of available data, $H(y)$. More formally we may define an isotropic valuation process as,

$$H(y) = H(x).^{26}$$

Relaxing supplementary assumption A1 permits the valuation process to become non-isotropic; noise has increased uncertainty in the data causing the entropy of the available data to exceed the entropy of the desired data:

$$H(y) > H(x).$$

$H(y)$ now has entropy surplus to requirements and requires correction to restore it to $H(x)$. This correction confers benefits to, and extracts costs from, agents and, from **Proposition 1**, is the reason agents maintain membership in a data valuation community. This correction comprises two tasks: the reduction in uncertainty and the rejection of noise. Uncertainty, $H_y(x)$, is the conditional entropy of the relevant data with knowledge of the data that is available; noise, $H_x(y)$, is the conditional entropy of the available data knowing what is relevant.

Before progressing, we also need a means of describing partially corrected data midway through the data valuation process. Recalling \hat{x} as the data considered and relevant and \hat{y} as the data not yet considered, from Appendix B we can expand Shannon's boundary states and write,

$$H(\hat{x}) + H(\hat{y}) + H_{\hat{x}}(y) + H_{\hat{y}}(x) \tag{3-13}$$

Here $H(\hat{x})$ is the entropy of the data considered and relevant; $H(\hat{y})$ the entropy of available data not yet considered; $H_{\hat{x}}(y)$ the entropy of available data knowing the data considered and relevant; and $H_{\hat{y}}(x)$ the entropy of relevant data with knowledge of the data not yet considered.

²⁵ For historical consistency, the notation $H(x)$ refers to the entropy of the random variable x and not x as an argument of H .

²⁶ Intuitively, like the equivalent thermodynamic process and as **Proposition 3** will subsequently develop, an isotropic valuation process can only exist as a theoretical construct or within appropriate confines. A hypothetical, serendipitous discovery of available data that perfectly matched data desired would still involve an entropic change – although perhaps if only considered from a broader perspective.

3.2. The Enrichment of Data and Reversal of Shannon's Communication System

As definers of the goal, the community also serves as arbiters of value. From **Proposition 1**, the expected benefit arising from participation must be greater than or equal to the associated cost before agents will participate in the valuation process. This presents agents with a problem: each agent desires immediately realizable benefit from the data but only has access to available data; that is, noisy data of potential, but uncertain, relevance and utility. Notwithstanding pre-curation by the system,²⁷ this noise and uncertainty is distributed throughout the available data; it follows, each dataset considered by an agent could reasonably contain some relevance to the agent's goals but also require further refinement. Therefore, in evaluating data, agents must determine relevance in each dataset through the identification of noise and a corresponding reduction in uncertainty. This evaluation enables a 'correction signal' which the system may use to improve the relevance of subsequent datasets.

Therefore, as illustrated in Figure 3-3, the process of enriching data proceeds as follows: upon accessing a community an agent considers a dataset, M_1 . This agent identifies relevant and irrelevant data, and produces a correction signal, S_1 . The system now has knowledge of both the dataset considered and the corresponding correction signal and will begin pre-emptively identifying noise and relevant data. The system produces a new dataset, M_2 , that accords with the agent's previous correction signal, which the agent evaluates before providing a second correction signal, S_2 . This process continues as the uncertainty of both data considered, $H_{\hat{y}}(x)$, and unconsidered, $H(\hat{y})$, is reduced by the system. In this way, the enrichment of data creates personal benefit for each agent and positive externalities for all other agents in the community.

We can therefore model a community of n agents, where the i th agent values datasets, M_j , a total of m times and $i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\}$. Datasets and correction signals have entropies of $H(M_j)$ and $H(S_j)$, respectively. The i th agent signals satisfaction with the $M_{i,1}, M_{i,2}, \dots, M_{i,m}$ datasets provided by returning a correction signal, S_m , of zero entropy and not requesting another M_j . Each condition is important: the existence of $H(S_j)$ indicates the agent has not simply withdrawn from the valuation process. $H(S_j) = 0$ indicates no further corrections are desired. The absence of a request for another M_j indicates the agent has not simply chosen to re-evaluate a

²⁷ Practically, the goal around which the community forms also serves as the distinction between valuable data and noise. Therefore, external rationale could be applied to segregate potentially valuable data from data of improbable value without direct input from agents. Additionally, structural restrictions inherent in the data also offer the basis for the autonomous identification of noise and reduction in uncertainty. This initial curation also permits the system's organization of datasets to maximize initial and ongoing value to agents.

previously supplied dataset. Valuation of the object against an ex ante goal has occurred with this i th agent validating $M_{i,m}$ datasets which constitute the joint entropy, $H(M_{i,1}, M_{i,2}, \dots, M_{i,m})$. Through this valuation, the agent has also provided correction signals $S_{i,1}, S_{i,2}, \dots, S_{i,m}$, and permitted definition of their validated exit entropy $H(x_i) = H(M_{i,1}, M_{i,2}, \dots, M_{i,m})$.

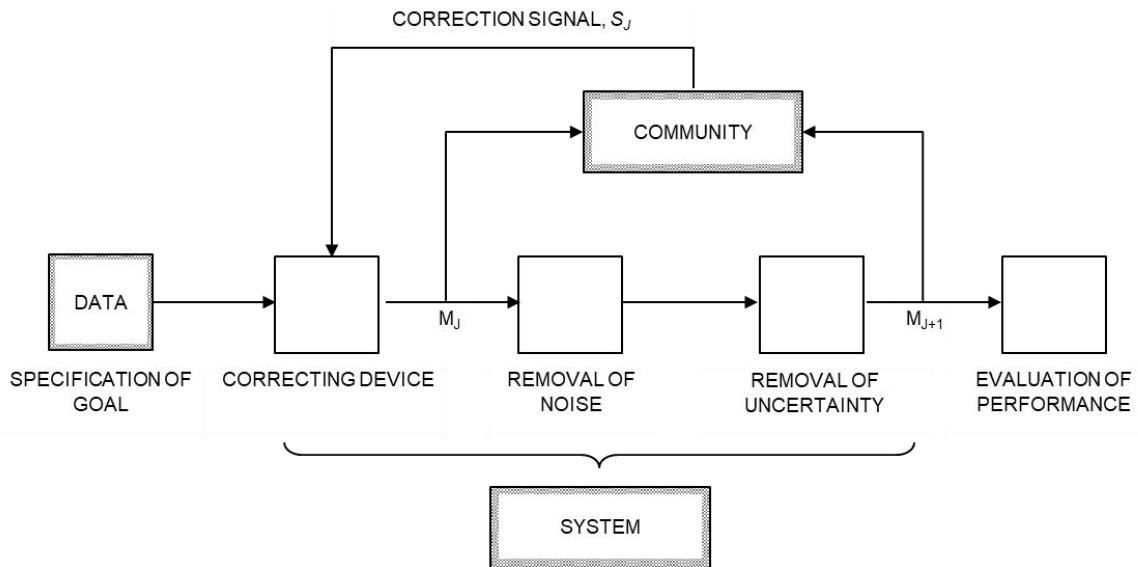


Figure 3-3. Data enrichment process – styled to reflect Shannon’s diagram of a correction system (1948, p. 22).

An illustrative example may be an agent’s progression through a popular web-based, flight booking system. The agent’s arrival signals a goal relating to data on flights and the system responds with an initial dataset $M_{i,1}$ consisting of destinations, dates, etc... with some pre-enrichment based on the current date and the agent’s location.²⁸ Here the system has automatically separated potentially valuable data from data of improbable value – such as content pertaining to other languages or historic dates. This agent responds with a preference for a destination and dates, $S_{i,1}$, which the system processes before presenting data, $M_{i,2}$, to the agent; presumably a selection of flights that accord with $S_{i,1}$. The agent’s next response might be selection of a flight with certain characteristics which constitutes the correction signal, $S_{i,2}$. The system processes this selection, and the process continues. As earlier, validation occurs when the agent signals satisfaction with all datasets offered and returns a correction signal with zero entropy, notionally by confirming purchase of tickets for a specific flight.

If the agent has multiple, complimentary goals this data valuation process may be repeated several times within the same data sharing platform. This will naturally cause an accrual of value to the data and modifying the probability distribution function, p , contained within \mathbb{B}_C as the agent’s

²⁸ The collection of all agents’ M_1 ’s would yield $H(M_{n,1})$.

expectation of finding valuable data within the data sharing platform changes. Meanwhile, the system may attempt to incorporate correction signals from previous valuations to accelerate the reduction of $H_y(x)$ for this agent – and others – in the community.

As an aside, the exit entropy of the correction signals, $H(S_{i,1}, S_{i,2}, \dots, S_{i,m})$, will be maximized only when each correction signal is statistically independent; that is, there is no way of deducing $S_{i,j}$ from knowledge of previous messages or correction signals. In such a case, the resulting set of correction signals, $\{S_{i,1}, S_{i,2}, \dots, S_{i,m}\}$, forms a Markov chain where an agent's response to each message, $M_{i,j}$, defines the exclusive circuit a system must take to proceed from $S_{i,j-1}$ to $S_{i,j}$. This is the case where the agents' correction signals will be of the highest potential benefit to the system.

3.3. The Necessity of Uncertainty in the Accretion of Value

We have now established the community and system as necessary components of a data valuation framework, with noise and uncertainty as a property and characteristic of data respectively. However, we must also explain how the data enrichment process produces an accretion of value in the dataset. Following Frankel and Kamenica (2019) in their treatment of information, this causality is central to an economic understanding of the value of data and therefore the value generation process that surrounds data.

Proposition 3: The reduction of uncertainty in data, $H_y(x)$, is necessary and sufficient for the accretion of economic value to data. The reduction of the entropy of remaining data, and the stepwise increase in the collective entropy of noise and relevant data may be sufficient – but are not necessary – for the accretion of economic value to data.²⁹

From Equations (3-10) and (3-11), while an agent's *expected* benefits are contingent on the probability remaining data contains data that will prove relevant, given by P , extant value is determined by data either already considered or currently under consideration. Therefore, the reduction of entropy of remaining data – such as might occur if a portion of the remaining data was lost – may reduce an agent's prospective costs but it would not necessarily affect the value of data. It follows that while a reduction in entropy of remaining data may be sufficient for accretion of value, it is not necessary.

²⁹ For expositional clarity, we substantiate the sole necessity of the reduction in uncertainty here, and its sufficiency for the accretion of value subsequently.

Likewise, the increase of collective entropy of noise and relevant data may be sufficient but is not necessary for the accretion of value in data. Normative frameworks permit the system to increase the efficiency of messages to agents so agents can reach satiation with lower costs. This increase in efficiency increases the value of data even while the collective and individual entropies of noise and relevant data remained unchanged. However, while the increase of the collective entropy of noise and relevant data causes an accretion of value in data, pursuit of these characteristics alone also curtails the potential value of data. For example, suppose the online flight booking system independently determined the desired flight details of all agents in the community. The system could then present all agents with datasets according to their desired flights, causing the full set of $\{M_{i,m}\}$ to constitute $H(x_i)$ and simultaneously setting all data not contained within $H(x)$ as noise. Uncertainty of unconsidered data would be set to zero even without its evaluation. Value for the community would be at a maximum as every M_j would increase each agent's personal utility at a minimum cost. However, access to the data sharing platform has become *benefitless* for prospective agents desiring even a closely related but unanticipated goal. Such a community's signal, S , will not convince prospective agents that the value of accessing the data is positive. Restated in terms of Equation (3-6), prospective agents will consider,

$$E_q[w(a, p(x \subset y), s)] \not\geq C_c(a(s))$$

and will not join the community. Therefore, the condition of certainty imposed on the system constrained the potential increase in the value of the data. Accordingly, pursuit of an increase in the entropies of noise and relevant data alone remain generally sufficient but not necessary for an accretion of value to data.

However, if sufficient uncertainty was retained in the dataset to enquire *if* $H(x)$ satisfied each agent, then at the cost of the consideration of an additional message and production of an additional correction signal, more agents would be permitted to reveal their otherwise private goals and be matched with data otherwise discarded as noise. An agent's pursuit of maximization of $u_Q(\{M_j\})$ and minimisation of $\sum_{j=1}^m C_c(M_j)$ from Equations (3-10) and (3-11), respectively, may occur at both large and small exit entropies. One agent may desire the purchase of tickets for a specific flight setting $H(x_i) \rightarrow 0$, while another may simply desire information regarding their closest airport, permitting $H(x_i)$ to remain large. In all cases, uncertainty, $H_y(x)$, remains instrumental in permitting the revelation of private goals while the recursive reduction in uncertainty permits agents to achieve their goals.

The maximization of the overall value of a dataset requires a tension in costs and benefits. While uncertainty permits an expansion in the size of the community of agents, the reduction of that

uncertainty bears a material cost to both the system and community. From Equation (3-7), agents experience the cost of participating as direct unit costs of evaluating datasets, M_j , and indirect time costs. Likewise, the system bears the cost of processing correction signals and producing additional datasets for evaluation by the community. However, these interactions also convey benefit to each party. Agents reveal their otherwise private goals to the system by issuing a series of correction signals, $H(S_j)$, to datasets. At each interaction, the system responds with an enriched dataset for the agent's valuation. Therefore, both parties possess incentives to maximize their own utility and minimize personal cost at each interaction. For agents, this takes the form of maximizing the reduction of personal uncertainty, $H_y(x_i)$, and to signal satiation, \acute{s} , once uncertainty has been sufficiently reduced to permit validation of datasets, $\{M_j\}$, against their goal. The system seeks specification of each agent's validated dataset, $H(x_i)$, to increase the rate at which uncertainty may be reduced from all available data. The ensuing process causes an accrual of value to the data as the probability that available data is valuable, p , and that the data will produce a valuable payoff, q , are both increased. Together with satiation, \acute{s} , p and q improve the quality of the signal, S , which prospective agents evaluate prior to accessing the data sharing platform. The value of the data as improved as a direct result of agents interacting with a system to value the data.

We can also measure this rate change of value to data. The data valuation process now necessarily includes the reduction in uncertainty and consequential increase in both benefit and cost of the data valued as a function of the number of interactions between the system and each agent in the community. Here, each i th agent's j th iteration creates one interaction, $k(i, j)$, where, $1 \leq k \leq mn$. Therefore, the effect of successive interactions, k , between the community and system may be expressed as a function of the interactions between each party,

$$\frac{\Delta V}{\Delta H_y(x)}(k) \tag{3-14}$$

We now have uncertainty, the chief characteristic of data that permits change in the value of data, united with the interactions, the driver of the process by which this characteristic may be altered. Uncertainty in data is altered by interactions between a system and community of agents with a consequential change in the value of data.

The last piece in this explanation of how to value data is the formal definition of the mechanism that governs this process of accumulation of value in data.

3.4. Enrichment and the Specification of Value Accretion

We have previously used the term enrichment to convey a form of mechanistic change in the intrinsic characteristics of data; we may now define it formally. The enrichment of data is the systematic reduction of the uncertainty in data; enrichment arises from the coordinated efforts of a system and community of agents and permits the valuation of data against the community's established goals.

Enrichment is, therefore, the mechanism that governs the process depicted in *Figure 3-3* and occurs as messages and signals are processed by the system and the community of agents. Each interaction offers both parties the potential to further reduce uncertainty as they identify valuable data from available data.

Proposition 4: The change of the value of data per interaction is a product of the associated enrichment process and the effect uncertainty has on the value of that data.

Let us consider an ideal scenario: every agent rationally considers each dataset offered and provides only statistically independent correction signals which the system perfectly processes to achieve a maximal reduction in overall uncertainty. Likewise, the system only produces datasets for agents' consideration that fully accord with each agent's earlier correction signals; each agent perfectly processes every dataset, unequivocally identifying noise while fully realizing the utility of their burgeoning, valuable data.

In such an ideal scenario any interaction by any agent at any iteration would provide a maximum correction of overall uncertainty. Here is the explanation for what we first postulated: that, subject to economic inefficiencies brought about by technical differences, repeat valuations by one agent may be considered as economically equivalent to single valuations by multiple agents. In such a case, formalizing our definition of enrichment above, we may express enrichment, \mathbb{E} , as a function of the interactions, $k(i, j)$, between the system and community,

$$\mathbb{E}(k) = - \frac{\Delta H_y(x)}{\Delta k} \quad (3-15)$$

The specification of a dynamic efficiency variable, $\eta_{i,j}$, on each interaction permits examination of a non-ideal scenario. Therefore, the reduction in uncertainty is equal to the discounted joint entropy of either datasets offered, or correction signals returned. For the system, we may write the reduction in uncertainty as,

$$- \Delta H_y(x) = H(S_{i,1}, S_{i,2}, \dots, S_{i,j}) \eta_{S_{i,j}} \quad (3-16)$$

or as a function of k , where $1 \leq k \leq mn$,

$$- \Delta H_y(x) = H(S_1, S_2, \dots, S_k) \eta_{S_k} \quad (3-17)$$

Analogously, for the i th agent valuing datasets to reduce their personal uncertainty, $H_y(x_i)$, we may write,

$$- \Delta H_y(x_i) = H(M_{i,1}, M_{i,2}, \dots, M_{i,j}) \eta_{C_{i,j}} \quad (3-18)$$

Therefore, reconnecting Equations (3-17) and (3-18) with Shannon's original framework of a correction system, the process of enrichment describes the recovery of the original message using the correction signals supplied by each agent. Provisioning for the relaxation of A2 and A3, the system in the data sharing platform will incorporate the unique distribution functions created by agent's unique correction signals, $H(S_j)$, such that it will attempt to optimize the enrichment function,

$$f(M_{i,j}; H_x(y)) : Y \rightarrow X$$

That is, the system provides datasets in the presence of noise attempting to map the data available to the data desired by each agent.

We now have the change in one party's uncertainty expressed as a weighted, joint-entropy of the datasets provided by the other party. The system is beholden to the relative performance of the community of agents for the valuation of data and contrariwise. Using the chain rule to adapt Equation (3-15) and rearranging for the rate of change of the value of data per interaction, we have,

$$\frac{\partial V}{\partial k} = - \mathbb{E} \cdot \frac{\partial V}{\partial H_y(x)} \quad (3-19)$$

where both parameters on the right-hand side are also functions of interactions, k .

Equation (3-19) provides an expression for the valuation of data in terms of uncertainty, enrichment and interactions, where uncertainty is the sole necessary and sufficient characteristic that permits a change in the value of data; enrichment is the mechanism that affects change in that characteristic; and interactions are the composite variable that advances that process.

Before examining the application of this Proposition, it is prudent to note the marginal value of data with respect to uncertainty, $\frac{\partial V}{\partial H_y(x)}$, must be contained within the first quadrant for sustained interactions by both the community and system. Therefore, as enrichment is only defined within the first quadrant, the marginal value of data per interaction must likewise remain both in the first quadrant and non-negative for all $k(i, j)$ provided both the system and agents maintain the freedom to unilaterally cease participation in the data valuation process. As developed in the results section

of this paper, when freedom to cease the valuation process is lost – or ceded – the value of data becomes an endogenous variable permitting value extraction from one or other of the parties.

This analysis of marginal value becomes crucial when assessing the generic modes – and their properties – by which data may be valued. We begin with an examination of the value of data when agents' preferences converge on, and are united around, a common data-related goal.

4. The Valuation of Data as a Resource

4.1. Data as a Resource: the Value of Enriched Data as a Means to an End

Data is valued as a resource when an agent specifies a goal, reduces the uncertainty in data and validates the performance of that enriched data against the goal. Enriched data is not the agent's goal *per se*, rather enriched data is a resource that enables determination of the goal. For example, agents purchasing tickets through an online flight booking system, mining cryptocurrencies, or using an internet search engine to find a specific news article each value data as a resource. Here, the endpoint of the valuation is defined; the communal enrichment of data is the means of achieving that endpoint.

The valuation of data as a resource resembles the valuation process as currently framed. The central task for the system is an optimization problem of connecting individual agents with valuable datasets while retaining sufficient adaptability to grow a community whose size maintains acceptable average costs. Characterizing this problem is the task to which we now turn.

From **Proposition 3**, we have the reduction of the uncertainty in data towards a goal as the sole necessary and sufficient property that enables accretion of value in that data. At this stage, all agents share a common, data-related goal and may freely distribute value among the community. Such a homogeneous community of agents offers complimentary correction signals, $H(S_j)$, and commence and cease valuation at the same levels of uncertainty, $H_y(x)$. Such a data sharing platform resembles a McGuire (1974) club as agents consume club goods, M_j , to cover the collective cost of the system and are drawn from a homogeneous population whose preferences and endowments are likewise homogeneous. The system functions to allocate club goods to all agents who exhibit the same utilization rate and data sharing platforms which can be replicated to meet the needs of the economy.³⁰ Crowding remains present, as scarcity of resources exists even in

³⁰ See Club Theory: Thirty Years Later (Sandler & Tschirhart, 1997) for a development and contextualization of McGuire clubs within club theory.

communities of $n = 1$ as the system incurs costs to produce club goods while agents incur costs based on the data valued, $C_C(M_j)$ and time taken, $v(t)$. While agents suffer from crowding, they also benefit from the active participation of other agents. Fleming et al. make provision for such a club, denoting partially rivalrous but excludable goods as “impure club goods” (Fleming et al., 2018, p. 169). Finally, when facilitating the valuation of data as a resource, data sharing platforms are also the natural extension of a data sharing club introduced by Easley et al. (2018) where the club of participating merchants is extended beyond two firms and a shared vendor to a scenario of open communication where any agent’s contribution carries a personal cost but facilitates both a personal benefit and benefit to the network.³¹

From **Proposition 1**, an agent’s rationale for participation in a data valuation community is to access the enrichment mechanism provided by the community. Adapting Equations (3-15), (3-16) and (3-18) gives enrichment for both the system and i th agent,

$$\mathbb{E}_S = \frac{\partial}{\partial k} (H(S_1, S_2, \dots, S_k) \eta_{S_k}) \quad (3-20)$$

$$\mathbb{E}_{C_i} = \frac{\partial}{\partial j} (H(M_{i,1}, M_{i,2}, \dots, M_{i,j}) \eta_{C_{i,j}}) \quad (3-21)$$

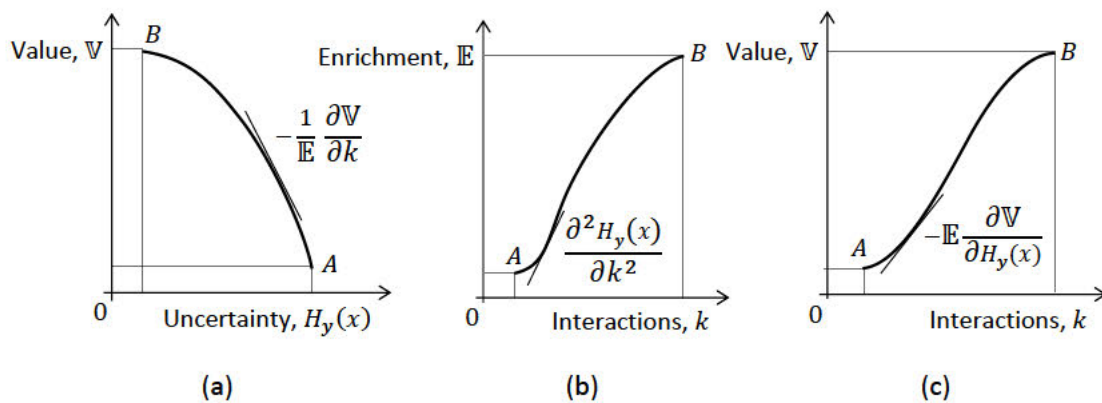
As agents evaluate datasets, $M_{i,j}$, and provide correction signals, S_k , the system may first define and then iteratively improve its enrichment efficiency, η_{S_k} with receipt of each successive correction signal. As agents value data toward a congruous goal, the exit entropy of their validated datasets remains a subset of the exit entropies previously defined. Each additional agent’s correction signals will have a correspondingly smaller impact on the joint-entropy of all extant correction signals, $H(S_1, S_2, \dots, S_k)$, diminishing the marginal value of further interactions to the system. Heuristically optimum enrichment paths will appear as probability density functions form across specific sequences of datasets, $M_{i,1}, M_{i,2}, \dots, M_{i,m}$. A system tasked with maximizing the value of contained data will ‘lead’ agents towards these sequences to reduce the cost of valuing data to both parties. Suggested travel destinations, personalized search results and proactive supply of trending news articles³² are all examples of a system responding in such a manner.

³¹ An interesting extension of the present model would be the formal integration of these two theories in the treatment of the incentives behind the provision of the correction signal to a community versus the value each agent is able to appropriate from the community.

³² Indeed, ‘trending’ monikers are simply the application of the same systematic response to valuing data applied across volatile goals.

Therefore, while new agents may remain unaware of previous agents' valuations, the prior definition of exit entropy and validation of enrichment paths permits an increased personal enrichment efficiency, $\eta_{C_{i,j}}$ as consideration of fewer datasets are required to validate each agent's data-related goal. Likewise, pre-valued datasets enable a greater per-interaction reduction of entropy further increasing each agent's personal enrichment rate, \mathbb{E}_{C_i} . This results in a vertical shift of an agent's enrichment curve in Figures 3-4(b) and permits agents access to otherwise unattainable value from data, \mathbb{V}_{C_i} , as personal uncertainty, $H_y(x_i)$ surrounding their goal is reduced for less personal cost. This valuation of data as a resource is graphically set out in Figures 3-4(a-c).

Point A in Figures 3-4(a-c) represents the position the community begins consideration of datasets, rendering the vertical offset of A as the value imputed to the data prior to the community's participation.³³ While the community is homogeneous, this investment is a fixed cost and may be defined as the inbound, or exchange cost of data to the community, \mathbb{C}_D .



Figures 3-4(a, b, c). The accretion of value in data caused by the enrichment of data through interactions between a system and community of agents³⁴

Similarly, the potential cost of the system's enrichment of data, \mathbb{C}_S , is given by the vertical offset of B from A in Figures 3-4(a, c). If an effective market exists for the purchase of appropriate data, some portion of \mathbb{C}_S could be set against the cost of acquiring equivalent data from an external data sharing platform. Similarly, the vertical offset of point B is the benefit to the community arising from the outbound transfer of data, \mathbb{B}_D , while the enrichment cost by the system, \mathbb{C}_S , is the value imputed to

³³ As previously noted, this prior imputation of value may come from external rationale or normative frameworks applied to 'raw' data or from the efforts of another community or data sharing platform.

³⁴ To the extent it aids readers: each of the graphs in Figures 4(a-c) may be thought of as addressing: a) properties of the goal, b) properties of the enrichment process, and c) properties of the combined data valuation process.

the data by the system. Figure 3-5 illustrates this value flow where costs impute value to the data and benefits realize value from data.

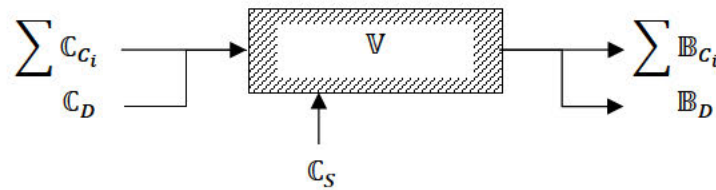


Figure 3-5. Value flow model for data as a resource

4.2. Requirements of a Marketplace for Data Valued as a Resource

These terms permit specification of an effective data marketplace as a market that facilitates the exchange of datasets by data sharing platforms between whom there exists complementary, data-related goals and a disparity in enrichment rates.

Naturally as a data marketplace matures, one would expect data sharing platforms' enrichment processes to become specialized as communities respond to past exchanges and anticipate the characteristics of future, valuable exchanges. Indeed Hartmann, Zaki, Feldmann, and Neely (2016)'s analysis of 100 data-driven businesses drawn from AngelList³⁵ found precisely this specialization. Data-driven businesses built business models to leverage competitive advantage in the collection, generation, discovery, aggregation, analysis, and "multi-source data mash-up and analysis" of data (Hartmann et al., 2016, p. 1392ff). It follows data may be valued in exchange at any stage along the enrichment pathway of a data sharing platform. Value transmission between data sharing platforms occurs when one party cedes control of a dataset of particular properties to another party.

Therefore, extending our general model defined in Equation (3-12) to include the effects of the valuation of data as a resource we have,

$$V = \sum_{i=1}^n (\mathbb{B}_{C_i} - \mathbb{C}_{C_i}) + \mathbb{B}_D - \mathbb{C}_D - \mathbb{C}_S \quad (3-22)$$

where \mathbb{B}_C and \mathbb{C}_C are given in (3-10) and (3-11), respectively; \mathbb{B}_D is the benefit derived from the outward exchange of value from data; \mathbb{C}_D is the cost incurred from the inward exchange of value from data; and \mathbb{C}_S is the system cost of enriching the data for valuation by the community.

³⁵ AngelList (www.angellist.com) is an online marketplace that permits companies to connect with investors, potential employees and interested stakeholders.

As a final observation, while the ostensible goal may be to reduce all uncertainty, ongoing enrichment where $\frac{\partial V}{\partial k} = 0$ only inflates an agent's marginal costs with direct or in-direct cost of participation in the data sharing platform. Equation (3-7) reflects the same practical limit. This transfer of cost from the system to individual agents has important implications for the treatment of both raw data and the enrichment process pursued particularly as it relates to the maximization of value to the community. We turn to these implications now.

5. The Valuation of Data as a Good

5.1. Data as a Good: the Value of Enriched Data as an End in Itself

Data is valued as an economic good when the agent's goal is enriched data. Following Lancaster (1966), agents seek utility arising from enriched data which possesses certain, desirable characteristics. This combination of characteristics changes as the data is enriched within a data sharing platform. Accordingly, in communities of $n > 1$ goals must be at least partially correlated as resources are shared in pursuit of the level of enrichment each agent's goal requires. Agents checking an online flight booking system for flight times, ticket prices or airports serviced – but not necessarily purchase tickets – value data as an economic good. Likewise, agents monitoring the prices of cryptocurrencies, or using an internet search engine to find information about recent events also value data as an economic good.

Agents join and depart data sharing platforms according to that platform's ability to enrich data with desirable characteristics. For instance, characteristics possessed by data enriched to point B in Figures 3-4(a-c), above, may enable a set of activities for one agent that constitutes an immediately realizable benefit but to another agent requires further enrichment. The community is now heterogeneous and we may relax supplementary assumption *A2: The agents' preferences converge on, and are united around, a common data-related goal*. Each agent's personal goal governs their entrance to and exit from each data sharing platform.

The data valuation process progresses as before. The system offers datasets that accord with correction signals provided by agents. However, agents now access, participate, and validate data to varying degrees as each seeks a maximization of personal value against personal goals. Indeed, we can now say with greater clarity, ad hoc data sharing platforms permit access to latent economic value because collaboration permits a reduction in the marginal costs of attaining benefits that would otherwise prove economically non-viable. However, the heterogeneity of the community comes at the expense of the system's enrichment efficiency. As Zhang, Baker, and Griffith (2019)

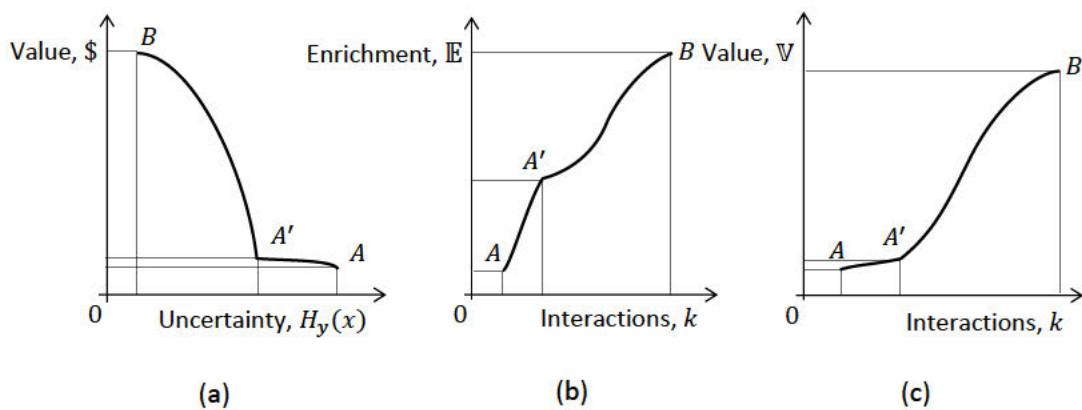
develop, maximizing the entropy of short, agent-centric chains of interactions becomes more significant than optimizing generic enrichment pathways.

The effect of a community's heterogeneity on data valuation can be best understood by examining the discrete case and then generalizing that to the continuous.

5.2. Diverging Agent's Goals and the Valuation of Data as a Good: Discrete Case

Suppose two groups of homogenous agents considered data with signal, S_A . One group decides to choose the signal and participate in the data sharing platform while the other group considers the same signal and chooses not to participate. *Early Adopters* take the data, assemble a community, and mobilize a system around it. In terms of the value flow model from Figure 3-5, above, *Early Adopters* impute value into the data through system costs with the expectation of realizing future value predicated on their system's data enrichment capabilities.

Figures 3-6(a-c) reflect the data valuation process adapted to account for discrete division in the community. *Early Adopters* proceed with enrichment and arrive at point A' by reducing uncertainty within the dataset to $H_y(x)_{A'}$. The enrichment of data to $H(x)_{A'}$ enables a signal, $S_{A'}$, which the *Majority* consider sufficient to join the community and begin valuing its data. The enlarged community continues valuing data to point B where it specifies an exit entropy, $H(x)_B$, and produces a further improved signal, S_B .



Figures 3-6(a-c). Diverging Entrance Conditions: Discrete Case

Leaving aside the investment in value made by *Early Adopters* across AA' , the effect of the *Majority* joining at point A' is a significant increase in correction signals offered to the system. From Equation (3-20), greater interactions by a broader community improves system enrichment, \mathbb{E}_S , resulting in better matching of datasets to correction signals and a corresponding increase in agents' enrichment

efficiencies, $\eta_{C_{i,j}}$. By Equation (3-22), the value of data has increased. Appendix C contains a brief, worked example.

If agents' access and participation enable an increase in the value of data, a lack of enrichment caused by non-participation or withdrawal of members will curtail value unless this behavior can be managed by the data sharing platform.

5.3. Diverging Agents' Goals and the Valuation of Data as a Good: Continuous Case

Value maximizing agents will exit the data sharing platform once satiation occurs. However, value maximizing behavior will also motivate an agent to withhold participation while the positive externalities of interactions by others accumulate.³⁶ This free-riding problem exists for data sharing platforms because while only participating agents contribute correction signals, every agent stands to benefit from all previous interactions – even from agents who have long-since exited the community.³⁷

This mobility of enriched data leaves data sharing platforms exposed to a Pareto-suboptimal outcome and presents important implications for the design and management of data sharing platforms that support agents' voluntary valuation of data as a good. Indeed, where agent benefit is principally realized at satiation, such as mining cryptocurrencies, inefficient system enrichment caused by self-serving agents may make attainment of community goals economically non-viable.

An obvious response to this problem to adjust, s , the signal supplied to prospective – and participating – agents depending on their current relationship to the community. To motivate participation, information regarding cost could be initially kept vague and only allowed to solidify once agent's expected benefits exceed remaining costs. While this approach has been suggested by Ely and Szydlowski (2020), agents must remain satisfied with the *quality* of the signal to remain participants in the community.

Universities adopt a different approach in their management of the value of data as a good by regulating the enrichment path agents must pursue to possess a completion signal. Matriculation is a signal of the type, \acute{s} , and validates completeness of the data enriched, enforcing – at least to some extent – full participation by agents. Students, as agents in university and class communities,

³⁶ While our present focus remains examination of the effect of this behavior on the value of data, we note Easley et al. (2018) provide a useful introduction to the conditions agents require for voluntary participation.

³⁷ In this way, while agents may quit the data valuation community – and may even successfully delete all data the data sharing platform maintains on them – they may not be truly 'forgotten'. Their participation has enabled enrichment by the system and enduring value for the platform.

maintain additional data valuation communities, even of size $n = 1$, to enrich data towards both transient and persistent goals. These communities – and their associated data sharing platforms – persist only while they offer the most efficient means of valuing data toward an agent’s goal. Tacit and codified data is transferred in to, and out of, each platform as necessary while the value of the platform is determined by the net change in its constituent elements as given by Equation (3-22). Like matriculation, grades are signals of the type, S , validating the data exchanged, the enrichment process, and the university as the necessary facilitator.

However, students may value a university’s data with the sole goal of possessing the matriculation signal. Where membership in a community is valued over the associated data, the value of data has become at least partially endogenous. This is the final mode we examine.

6. The Valuation of Data as a Currency

6.1. Data as a Currency: the Endogeneity of the Value of Data

Data is valued as a currency when it acts as a store of value or facilitates the exchange of value between agents or other parties. This storage and exchange of value may occur in either the short-term or long-term permitting the value of data to be treated as either a current, or non-current, asset. In both cases, the endogeneity of the value of data is required to permit the ongoing investment of costs that exceed benefits into the data sharing platform.

Thus far, central to the data valuation process has been the appropriation of all benefits by agents directly through \mathbb{B}_{C_i} or indirectly through \mathbb{B}_D . However, where agents elect not to extract benefits from the data sharing platform, preferring active participation in the enrichment process, by **Proposition 2** value continues to accumulate within the platform. We may now relax the final supplementary assumption *A3: The community directs the valuation process and participates in the distribution of the resulting value*. Under-extraction of benefit produces latent value within the data sharing platform and permits a subsequent over-extraction of benefit to either the original, or another, party.

Applying the data valuation framework, we will first consider the means by which this endogenous value is created before considering its nature.

6.2. The Temporary Under-Extraction of Value from Data

Temporary under-extraction of benefit was the deliberate strategy adopted by *Early Adopters* in the earlier valuation of data as a good. *Early Adopters* chose to defer validation of data and extraction of

benefits while they reduced uncertainty in available data and assembled their enrichment process. To the extent *Early Adopters* had reduced uncertainty in the data across AA' the value in the data had accrued according to the integral of (3-19) with respect to k . Subject to market conditions, this latent value may have been realized in exchange or use. In exchange, the benefit extracted by *Early Adopters* at point A' would have equaled the cost imputed by the *Majority*,

$$\mathbb{B}_{D_{Early\ Adopters}} = \mathbb{C}_{D_{Majority}}$$

Alternatively, the enriched data and system of enrichment may have been valued in use where *Early Adopters* extracted $\mathbb{B}_{D_{A'}}$ over the remaining enrichment process as *non-anonymous tolls*³⁸. The limit of the value carrying capacity of data between a system and agent is the benefit the agent ascribes to the reduction in uncertainty of the resulting information (Frankel & Kamenica, 2019), discounted by the uncertainty of relevance of the data to the agent's goal. A system that could resolve the relevance of data for an agent could increase the non-anonymous tolls levied on the agent beyond the value of data and unilaterally, appropriate benefit from the enrichment process.³⁹

More generally, all agents must be prepared to temporarily suspend the validation of data because, by Equation (3-6), the benefit all agents attach to *access* is predicated on *expected* payoffs. Even after *participation* begins both \acute{x} and $u_q(\{M_j\})$ remain small while agents offset ongoing costs against expected payoffs until the utility of datasets considered becomes significant.

6.3. The Sustained Under-Extraction of Value from Data

An under-extraction – or concession – of benefits also occurs when agents *prefer* active participation in the community over satiation of shared goals. This is the case where the value an agent holds in the community supersedes their valuation of enriched data. Such an agent remains less sensitive to the personal utility of data considered and more sensitive to the valuation process that permits involvement in the community. Rational agents must continue to seek a maximization of the value of participation but with a preparedness for their costs to accrue even while extant benefits do not cover these costs.

In contrast to the temporary under-extraction of value, above, where both \acute{x} and $u_q(\{M_j\})$ started small and grew together, if the former grows without significant change to the latter then the

³⁸ As Sandler and Tschirhart (1997) explain, non-anonymous tolls are a form of rent clubs may apply differentially to members based on their activity or identity. They contrast *anonymous tolls* which are levied to all (or none) of the club at a common level.

³⁹ Indeed, flight booking websites have been accused of precisely this rent-seeking behavior (Shen, 2017).

expected payoff must remain a significant component in an agent's accounting of value. Taken to the limit, as costs continue to grow as a function of time and number of datasets valued, the expected payoff must become an increasingly dominant component of an agent's perception of benefit. In terms of Equation (3-10) and (3-11):

$$E_Q[w(a, P((x - \hat{x}) \subset \hat{y}), S)] \geq C_C(a(s), v(t), \hat{s}) + \sum_{j=1}^m C_C(M_j) \geq u_Q(\{M_j\})$$

An agent's expected payoff is a composite function driven by Q, P and S . If utility from valued data does not materialize, these probabilities may avoid collapse if an agent believes their apparent progress through all data, $y - \hat{y}$, remains small compared to their potential progress and resulting payoff. In such a case, where agents remain convinced the initial signal they accepted is a subset of the desirable signals produced by others in the community they may be persuaded to accept a data valuation model where their expected payoff continues to justify the ongoing cost of enriching data in perpetuity.

6.4. The Over-Extraction of Value from Data

Finally, where the community is prepared to under-extract benefits from the valuation of data the system may also seek appropriation of benefits from data. Even where agents do not cede distribution of benefits, adversarial systems could leverage enhanced enrichment processes to reduce the uncertainty regarding the relevance of data to agents and compete with them for value.

It is important to differentiate how this case differs from one group of agents seeking value from other agents. Initially, the system differs from a group of agents because the system, uniquely, intermediates valuable exchanges with, and often between, agents. Moreover, the system has observed the collective interactions of all agents – past and present – and observed each agent's attempts to maximize their own private value. In that sense, the system not only controls one half of the means of value production but may also amass a more complete knowledge over the activities of agents than any subset of agents, past or present. To the extent this knowledge is reinvested, agents experience this advantage as a disparity between the system enrichment efficiency, η_{S_k} , and their personal enrichment efficiencies, $\eta_{C_{i,j}}$. It is precisely these characteristics of a data sharing platform that are captured in the collective signal, S , from which agents expect particular enrichment efficiencies and which permit the endogeneity of value. Therefore, the system may freely appropriate value from the community to the extent the community prefers one specific data sharing platform over any other.

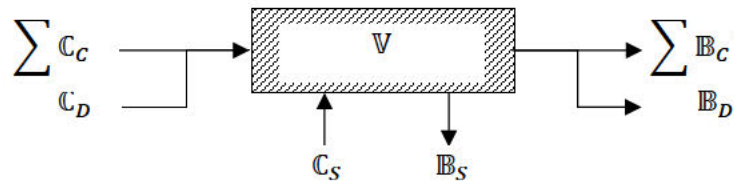


Figure 3-7. Value flow model for when the value of data is a choice variable

Figure 3-7 illustrates the multi-input and multi-output value flows that surround data where B_S is the benefit extracted by the system for the enrichment of data. Therefore, we may also expand the general valuation model to include the outward flow of benefits to the system,

$$V = \sum_{i=1}^n (B_{C_i} - C_{C_i}) + B_D - C_D + B_S - C_S \quad (3-23)$$

where B_C , C_C , B_D , C_D , and C_S are given previously and, B_S is the benefit derived by the system from the enrichment of data for valuation by the community. We have now described the mobilization of the value of data by two parties and rivalry in the appropriation of that value within, and between, them.

7. Conclusion

7.1. Summary of a Generic Value Framework for Data

With all supplementary assumptions relaxed, Equation (3-23) provides an account for how data is valued. In words, data is valued by agents who, sharing a congruous goal and desiring benefit from that data, direct the systematic reduction of uncertainty in that data. Agents are prepared to accept the cost of curating and evaluating the data to realize the resulting benefit. The coordinated activities of this ad hoc community resemble a system which acts as a mutually dependent, but potentially rivalrous, body in the creation and appropriation of the value of data. Finally, data may also be valued through direct inbound or outbound exchange. A community of at least one agent, a system and data must be present for data to have value: collectively, these three components constitute a data sharing platform.

The data valuation process takes data as an input, enriches it through the recursive separation of irrelevant data from relevant data, and validates the output against the goal. Therefore, the total value of data is the benefit of valuing that data, less the associated cost. The system and community of agents attempt to maximize their own value of data by exchanging datasets and correction signals

as each party attempts to reduce uncertainty in the data in accordance with their goal. This series of interactions is defined as the data enrichment process and, when combined with the data valuation model, explains how data may be valued as a resource, an economic good or currency. As a resource, data permits the generation of value through the systematic reduction of uncertainty surrounding a goal. As an economic good, data possesses common production paths with other data-goods but retains an infinite capacity for the variance of desirable characteristics – limited only by the enrichment efficiencies and goals of parties. Finally, as a currency, variously enriched data enables storage and communication of value to all parties within a data sharing platform, provided the value of data has become at least partially endogenous.

7.2. Open Questions

If the value of data accumulates within data sharing platforms, why do some even well-resourced businesses appear to ‘drown in data’ while others appear to thrive in deriving value from data? Clearly there are both specific and general issues to address. We have established that pursuit of the valuation of data as a currency reveals the endogeneity of value as a necessary condition for the storage, communication, or appropriation of value from data. Therefore, the management challenge becomes designing the transmission of the *value* of data through an organization, rather than merely managing the flow of data. This is consistent with, although more specific than, Short and Todd (2017)’s findings. However, the endogeneity of the value of data within a firm has an important consequence. A firm that embodies a corporate data sharing platform must account for multiple value channels: by **Proposition 3** value is created within the firm as uncertainty is stripped from data in processes that encompass and extend beyond existing production paths. Underlining G. Parker et al. (2017)’s findings, data-centric value accretion channels become an even more fundamental requirement than pursuit of emerging business models. For professionals in data driven firms, an examination of data-centric value production becomes more central than questions regarding commercialization of firm data. Such an understanding must also inform modeling and design of information systems. When data provided to the system is leveraged as a parallel currency, these systems may be mobilized to encircle stakeholders, simultaneously satisfying their goals, and enabling enrichment of meta-data to other valuable ends. This provides an alternate, data-centric foundation for the analysis of match-making platforms: intermediary platforms capitalize the variously enriched data of their communities such that the utility of the data sharing platform remains positive for all parties.

This intertwining of data-based enrichment paths with novel technology capable of dramatically altering the costs of acquiring and managing data also helps explain the pre-eminence of so-called

data-centric businesses. Their proprietary enrichment of data, provided by agents who can discern no benefit from the data in its raw form, provisions immanent systems who facilitate scarcity in a marketplace that values access to this community's enriched data. This transformation of data becomes central to the privacy debate: agents must determine the net present value of relinquishing data to a platform in exchange for access to the associated, and often short term, benefits.

If the ordinal value of data may now be understood, should we expect the apparent discrepancy in the cardinal value of data to persist? Indeed, from Proposition 3, if value accrues with the reduction of uncertainty from data, and the entropy of data can be measured, ought not the cardinal value of data be quantified as a matter of first importance? **Proposition 1** and **2** note the functional primacy of a congruous goal: a goal defines the ad hoc community and scopes the conditions for value while **Proposition 4** sets the accumulation of value as a dependent variable driven by each data sharing platform's internal enrichment process. Hence, data sharing platforms better resemble individual sovereigns than merchants in a marketplace united by fiat money. It follows, the congruity of the goal used to enrich data sets the extrinsic exchange rate of otherwise similar datasets. In other words, the more data is enriched towards a goal, the less the transfer of value away from parties who possess that goal can be taken for granted. Therefore, effective data marketplaces require ontologies of enrichment goals before quantification of value can occur.

While this framework offers an initial explanation for the relatively lack of mobility of enriched data through value chains observed by Zhang et al. (2019), more work is needed to characterize transmission of data across markets. The grounding of additional characteristics attached to data, such as properties rights (Jones & Tonetti, 2018), against each agent's marginal effect on the value of data is an important first step. Likewise, consideration of the market conditions required for sharing of data, such as Easley et al. (2018)'s game theoretic model, may now include a more granular examination of the nature of data potentially exchanged.

For professionals, we conjecture that there will be increasing tensions between data providers – often agents in data valuing communities – and the data sharing platforms that encompass them. The emergence of new commercialization models that support the exchange of value between fiat money and enriched data within data sharing platforms will continue to promulgate rapid societal and economic change. Historically, the viable collection of data has proven a formidable bottleneck, but the interconnection of Internet of Things devices and the voluntary digitization of individual's analogue activities such as social networks and health has all but removed that barrier. As both enriched data and the ensuing enrichment systems serve as organizational assets, we expect the apoptosis of data sharing platforms will be marked with liquidation, or worse the public

abandonment of sensitive data assets. This wholesale release of data will further strain the trust data providers place in data sharing platforms as data collected for one goal will become divorced from its initial terms and remain available for additional, even once incongruous, goals. As data sharing marketplaces continue to mature secondary data markets will emerge, as will derivative data characteristics such as, with some irony, proof of veracity. Self-affirming datasets such as those supported by blockchain will abound. Indeed, it seems fitting to end with the prescient remarks of Nobel Laureate Kenneth Arrow (1996, p. 127), “I would surmise that we are just beginning to face the contradictions between the systems of private property and of information acquisition and dissemination.”

References

- Adjerid, I., Peer, E., & Acquisti, A. (2018). Beyond the privacy paradox: objective versus relative risk in privacy decision making. *MIS Quarterly*, 42(2), 465-488.
- Arrow, K. J. (1985). Informational structure of the firm. *The American Economic Review*, 75(2), 303-307.
- Arrow, K. J. (1996). The economics of information: An exposition. *Empirica*, 23(2), 119-128.
- Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7), 2183-2203.
- Chang, W. L., & Boyd, D. (2018). *NIST Big Data Interoperability Framework: Volume 6, Big Data Reference Architecture*. Retrieved from
- Chiang, R. H., Grover, V., Liang, T.-P., & Zhang, D. (2018). Strategic value of big data and business analytics. In: Taylor & Francis.
- Choudary, S. P. (2014). The Platform Stack - Understanding platform business models. Retrieved from <https://artplusmarketing.com/the-platform-stack-c83f9c96e6>
- Damodaran, A. (2014a). A Disruptive Cab Ride to Riches: The Uber Payoff. Retrieved from <http://aswathdamodaran.blogspot.com/2014/06/a-disruptive-cab-ride-to-riches-uber.html>
- Damodaran, A. (2014b). Uber Isn't Worth \$17 Billion | FiveThirtyEight. Retrieved from <https://fivethirtyeight.com/features/uber-isnt-worth-17-billion/>
- Easley, D., Huang, S., Yang, L., & Zhong, Z. (2018). The Economics of Data. Available at SSRN 3252870.
- Ely, J. C., & Szydlowski, M. (2020). Moving the goalposts. *Journal of Political Economy*, 128(2), 000-000.
- Evans, D. S. (2003). Some empirical aspects of multi-sided platform industries. *Review of Network Economics*, 2(3).
- Evans, D. S., & Schmalensee, R. (2016). *Matchmakers: the new economics of multisided platforms*: Harvard Business Review Press.
- Evans, P. C., & Gawer, A. (2016). The rise of the platform enterprise: a global survey.
- Fleming, E., Griffith, G., Mounter, S., & Baker, D. (2018). Consciously pursued joint action: Agricultural and food value chains as clubs. *International Journal on Food System Dynamics*, 9(1012-2018-4116).
- Frankel, A., & Kamenica, E. (2019). Quantifying information and uncertainty. *American Economic Review*, 109(10), 3650-3680.
- Gawer, A., & Cusumano, M. A. (2014). Industry Platforms and Ecosystem Innovation. *Journal of Product Innovation Management*, 31(3), 417-433. doi:10.1111/jpim.12105

- Gentzkow, M., & Kamenica, E. (2014). Costly persuasion. *American Economic Review*, 104(5), 457-462.
- Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, 35(2), 388-423. doi:10.1080/07421222.2018.1451951
- Gupta, A., Kannan, K., & Sanyal, P. (2018). Economic experiments in information systems. *MIS Quarterly*, 42(2), 595-606.
- Gurley, B. (2014, 2014-07-11). How to Miss By a Mile: An Alternative Look at Uber's Potential Market Size. Retrieved from <http://abovethecrowd.com/2014/07/11/how-to-miss-by-a-mile-an-alternative-look-at-ubers-potential-market-size/>
- Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing value from big data—a taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management*, 36(10), 1382-1406.
- Hukal, P., Henfridsson, O., Shaikh, M., & Parker, G. (forthcoming). Platform Signaling for Generating Platform Content. *MIS Quarterly*.
- Jones, C., & Tonetti, C. (2018). *Nonrivalry and the Economics of Data*. Paper presented at the Society for Economic Dynamics: 2018 Meeting Papers.
- Kant, I. (1870). *Grundlegung zur metaphysik der sitten* (Vol. 28): L. Heimann.
- Ketter, W., Peters, M., Collins, J., & Gupta, A. (2015). Competitive benchmarking: an IS research approach to address wicked problems with big data and analytics. *MIS Quarterly*.
- Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2), 132-157.
- Langley, P., & Leyshon, A. (2017). Platform capitalism: the intermediation and capitalization of digital economic circulation. *Finance and society.*, 3(1), 11-31.
- McGuire, M. (1974). Group segregation and optimal jurisdictions. *Journal of Political Economy*, 82(1), 112-132.
- Microsoft. (2016). Microsoft to acquire LinkedIn. Retrieved from <https://news.microsoft.com/2016/06/13/microsoft-to-acquire-linkedin/>
- Moody's Investment Service. (2016). Research: Rating Action: Moody's reviews Microsoft's Aaa rating for downgrade following announced acquisition of LinkedIn - Moody's. *Global Credit Research*. Retrieved from https://www.moodys.com/research/Moodys-reviews-Microsofts-Aaa-rating-for-downgrade-following-announced-acquisition--PR_350591
- Parker, G., Van Alstyne, M., & Jiang, X. (2017). Platform ecosystems: How developers invert the firm. *MIS Quarterly*.
- Parker, G. G., & Van Alstyne, M. (2002). Unbundling in the presence of network externalities. *Mimeo*.
- Parker, G. G., Van Alstyne, M. W., & Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*: WW Norton & Company.
- Rochet, J. C., & Tirole, J. (2003). Platform Competition in Two-Sided Markets. *Journal of the European economic association*, 1(4), 990-1029.
- Sandler, T., & Tschirhart, J. (1997). Club Theory: Thirty Years Later. *Public choice*, 93(3-4), 335-355.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell system technical journal*, 27(3), 379-423.
- Shen, L. (2017). The Truth About Whether Airlines Jack Up Prices If You Keep Searching the Same Flight. *Business. Flights*. Retrieved from <https://time.com/4899508/flight-search-history-price/>
- Short, J., & Todd, S. (2017). What's Your Data Worth? *MIT Sloan Management Review*, 58(3), 17.
- Sims, C. A. (2003). Implications of Rational Inattention. *Journal of monetary Economics*, 50(3), 665-690.
- Stigler, G. J. (1961). The Economics of Information. *Journal of Political Economy*, 69(3), 213-225.

- Van Alstyne, M. W., Parker, G. G., & Choudary, S. P. (2016). Pipelines, platforms, and the new rules of strategy. *Harvard business review*, 94(4), 54-62.
- Zhang, Y., Baker, D., & Griffith, G. (2019). Measuring Information Quantity and Quality in the Supply Chain: A Systematic Literature Review and a Proposed Conceptual Framework. *Under Review*.

Appendix A. Review of Economic Literature

This work relates to understanding the value of data as distinct from the value of information. Data differs from information to the extent that the purpose of data has not been – and may never be – confirmed. Accordingly, as set forth by Shannon (1948), uncertainty moves from being a question of utility to one of a fundamental characteristic of data. Nonetheless, until recently economic treatment of data has been coincidental with the study of the economics of information.

Accordingly, Stigler (1961) assumed the goal, or “desideratum” (Frankel & Kamenica, 2019, p. 3650), of each agent, themselves buyers in a marketplace, correlated with the information provided by sellers.⁴⁰ The value of information as a resource could then be observed as the effect of buyers’ informed decisions on market prices. Arrow broadens (1985) and then extends (1996) this understanding by drawing on communication theory to establish the valuation of information as a choice variable. Agents observe signals pertaining to the nature of some information and optionally select a quantity of that information in an attempt to maximize their payoffs. The implicit assertion: choice of the signal reveals an agent’s goal for the data. The first result of this paper is to extend Arrow’s treatment of information to a model that accounts for intrinsically uncertain data where agents must resolve relevance as well as maximize utility.

Frankel and Kamenica (2019) formalize the relationship between uncertainty and information by quantifying their coupling, thereby establishing uncertainty as the chief antithesis of the benefit of information. While this paper takes a substantially different approach to theirs, it broadens the significance of their result by demonstrating the origins of this coupling are set when the relevance of that information is still unknown. We show information of uncertain relevance, or more succinctly: data, gains value as agents systematically separate irrelevant data, or noise, from relevant data in a process that resembles a reversal of Shannon’s (1948) communication system. Execution of the data valuation process confers benefits to, and extracts costs from, agents. Following rational

⁴⁰ Stigler acknowledges, but does not develop, the case of an oversupply of information (1961, p. 22); that is, the supply of data to buyers whose goal does not accord with all supplied ‘information’.

inattention literature,⁴¹ we invoke Shannon's appropriation of entropy as a useful measure of uncertainty in data but return to categories that mirror Shannon's: available data, uncertainty, relevant data and noise. The economic valuation process of data, no matter how noisy, may now be approached. This is the second result of this paper.

The economics of data has also been approached directly through examination of the value of data as a currency. As a currency, data permits the storage or transmission of value according to market conditions and the private goals of agents. Easley et al. (2018) offer a game theoretic framework that permits ordinal value questions to be addressed; their framework permits description of conditions where rational agents would voluntarily participate in data sharing practices. While insightful, this approach still attributes the value of data to its extrinsic utility. Jones and Tonetti (2018) propose the attribution of property rights to data as a partial solution. Property rights provide a rationale for the provision of rent, but efficient distribution of value relies on a mechanism for determining the relative value imputed to the data by each party (Zhang et al., 2019). Our third result is to show that the reduction of the intrinsic uncertainty in data is the sole necessary and sufficient condition for the accretion of value. This result illustrates that the coupling of uncertainty and information derived by Frankel and Kamenica (2019) exists as an independent characteristic of data. Importantly, the reduction in the uncertainty in data can therefore also be used as a rationale for apportioning value between parties.

Finally, as Jones and Tonetti (2018) develop, data is nonrival and therefore cannot be treated like an ordinary economic good. Data collected for one purpose, may be simultaneously used for other purposes. Intuitively, this permits multiple paths of attributing benefits to the data as firms can both use and sell data, giving value in both use and exchange to the data. However, the pursuit of both outcomes consumes limited resources. Implicitly, access to data remains the central property that governs value flow, but access alone does not confer benefits as data must be developed – or enriched – to create scarcity or realize some external benefit. It follows, this consumption of resources in the development of data is rational where that investment enables benefits of a greater magnitude to be extracted. Where value is understood as the difference between benefits accrued and costs incurred, our final result is the definition of the relationship between the change in value of data with respect to the activities of agents, the rate of enrichment of the data, and the relative effect of uncertainty on the value of data.

⁴¹ See Sims (2003).

Therefore, with reference to extant Information Systems and Economics literature the contribution of this paper is threefold. First, an economic valuation model is developed that accounts for: intrinsic uncertainty in the relevance of data; the non-rivalrous nature of data; uncertainty in expected payoffs; and rivalry of resources consumed in the data's valuation. Second, we establish uncertainty as the sole intrinsic characteristic that alters the economic value of data. We also define the data enrichment process as the means this uncertainty is reduced. The development of this theory borrows heavily on communication theory and becomes the method by which data is valued. Third, the data valuation process and data enrichment process are combined to create a framework useful for understanding, managing and attributing value between all parties involved in the valuation of data. Our results are that data may be valued according to one of three modes: the valuation of data as a resource, as an economic good, and as a currency.

Appendix B. Mid-process Entropic State

From Shannon, Equations (B1) and (B2) reflect the boundary conditions immediately before, and following completion of, the data valuation process. $H(x, y)$ is the joint entropy of both the relevant data and available data, and equal to the sum of the input entropies: raw data and uncertainty, and the sum of the output entropies: relevant data and noise.

$$H(x, y) = H(y) + H_y(x) \tag{B1}$$

$$= H(x) + H_x(y) \tag{B2}$$

Collectively, the sum of entropy prior to commencement of the data valuation process is equal to the sum of entropy following the theoretical completion of the same.

Prior to commencement of the data valuation process we have both raw, available data and uncertainty regarding that data. From Equation (B1) the sum of the entropy of available data and uncertainty equals the joint entropy of relevant data and available data. At this stage, relevant data and noise have not yet been defined, except that the sum of their entropies also equals the joint entropy of relevant data and available data, $H(x, y)$, and the initial entropies.

As no data has yet been selected, considered or evaluated, the probability of possessing an observation relevant to an agent's goal is zero, that is,

$$p_0 = 0.$$

Likewise, \acute{x} is an empty set and $\acute{y} = y$. Therefore,

$$H(\acute{x}) = 0$$

$$H(\hat{y}) = H(y)$$

$$H_{\hat{y}}(x) = H_y(x)$$

which sets Expression (3-13) equal to the right-hand side of Equation (B1) at the point immediately before commencement of the data valuation process.

Similarly at the completion of the data valuation process all available data will have been considered and marked as either relevant data or as noise.⁴² As unconsidered data diminishes, $\hat{y} \rightarrow 0$, the probability of finding relevant but unconsidered observations must also diminish causing $H(\hat{y}) \rightarrow 0$. Likewise, $\hat{x} \rightarrow x$ as the potential for finding additional, relevant observations has also diminished. Therefore,

$$H(\hat{x}) \rightarrow H(x)$$

$$H_{\hat{x}}(y) \rightarrow H_x(y).$$

Adapting Shannon's statement of conditional entropy (1948, p. 13), we may write the conditional entropy of x given \hat{y} as,

$$H_{\hat{y}}(x) = - \sum_{i,j} p(\hat{i}, \hat{j}) \log p_i(\hat{j})$$

where \hat{i} and \hat{j} are the probability of \hat{y} and x occurring, respectively, and $p(\hat{i}, \hat{j})$ is the probability of both events happening simultaneously. As the data valuation process approaches completion, $p(\hat{i}) \rightarrow 0$ and, therefore, $p(\hat{i}, \hat{j}) \rightarrow 0$, causing $H_{\hat{y}}(x) \rightarrow 0$. Therefore, Expression (3-13) approaches the right-hand side of Equation (B2) as the data valuation process approaches completion.

Appendix C. Valuation of Data as an Economic Good: Example

This symbiotic improvement in the data enrichment process and resulting increase in the value of data depicted in 5.2 is more familiar than we might realize. Suppose after considering a specific flight's details, $H(M_j)$, a significant number of agents issue correction signals indicating a preference for an earlier dataset, $H(M_{j-\alpha})$.⁴³ Following the second valuation by the agents of the earlier

⁴² This end state of full segregation of relevant data and noise is possible but it will not, in general, eventuate as agents cease valuing data once satiation occurs – irrespective of the current states of remaining data and uncertainty.

⁴³ For narrative simplicity we persist with the flight booking example, however in website management this simple action of 'requesting an earlier dataset' is known as a *bounce* and gives rise to the fundamental webpage metric: *bounce rate*, the proportion of viewers who, upon viewing a webpage almost immediately request a page previously viewed. This interaction is so common that web browsers – and many peripherals –

dataset, the system observes that many agents change their earlier correction signal from $H(S_{j-a})$ to $H(S_{j-a}^*)$. The updated correction signal prompts the system to produce a revised dataset $H(M_{j-a+1}^*)$ containing different flight details. Consequently, the specific probability p , that the original datasets, $H(M_{j-a+1}, M_{j-a+2}, \dots, M_j)$ constitute valuable data is reduced. Likewise, the probability q , that the original datasets will produce a valuable payoff is also reduced. As each interaction bears a cost to both agent and system, the system could reasonably manage a bias for delivering the preferred $H(M_{j-a+1}^*)$ to new agents *irrespective of their S_{j-a} correction signal* in an attempt to maximize value across the data sharing platform.

have dedicated *Back* buttons. Moreover, if the system does not automatically adjust users away from undesirable webpages, then users – even as communities of $n = 1$ – learn to adapt correction signals to suit. For example, by writing: “*Manchester, NH*” rather than just “*Manchester*”.

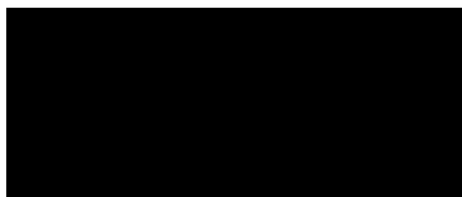
Higher Degree Research Thesis by Publication

University of New England

Statement of Authors' Contribution

We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated in the *Statement of Originality*.

	Author's Name	Percentage of Contribution
Candidate	Matthew Wysel	75
Other Authors	Derek Baker	20
	William Billingsley	5



Candidate

Date



Principal Supervisor

Date


Higher Degree Research Thesis by Publication

University of New England

Statement of Originality


We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that the following text, figures, and diagrams are the candidate's original work.

Author	Type of Work	Page Numbers
Matthew Wysel	Conceptualization, methodology, formal analysis, writing (original draft), writing (review), editing.	Entire Document
Derek Baker	Conceptualization, writing (review), editing, supervision.	Entire Document
William Billingsley	Conceptualization.	Entire Document



Candidate

Date



Principal Supervisor

Date

Chapter Four:

Profiting from data. How data enables firms to have their cake, sell it, and eat it too.

Overview of Manuscript

Status:	Reviewed at <i>Management Science</i>
Date Submitted:	9 February 2023
Date Published:	-
Suggested Citation:	Wysel, M., & Baker, D. (2023). Profiting from Data. How Data Enables Firms to Have Their Cake, Sell It, and Eat It Too. Manuscript submitted for publication.

Summary of Paper in Context of PhD:

Where Chapter 2 connected firm-level decision making with *second-order* effects on value creation, this chapter applies the valuation modes from Chapter 3 to establish a connection between firm-level decision making and *first-order* effects. Specifically, this chapter develops a firm-level, objective function for the production of value from data and asks, ‘how much should a firm invest in data?’ and, ‘how often should the firm repeat that process?’.

This paper extends the economics of information introduced in Chapter 2 to include the non-rivalry, excludability, and conditional network effects exhibited by data. This informs a firm-level, data-based production process which takes data as a choice variable and produces value as an output. This paper abstracts the data sharing ecosystem from Chapter 2 into assets and functions within a firm. From Section 2.2 in this paper,

“...where prior approaches have modelled the process as a series of ostensibly homogenous interactions involving data (Wysel et al., 2021), we adopt the opposite perspective and investigate the interactions that comprise the process. ... To that end, we model a firm that takes data as an input and develops that data into insight across a small number of discrete, operating periods. This insight creates value

internally as innovation, and externally in both use and exchange through the market.” (Chapter 4, p109)

Deliberately, other than noting data may be purchased from outside the firm and value is created via payoffs received from the market, this paper considers the firm in relative isolation from its surrounding environment, leaving the complexity associated with the trade of data to Chapter 5.

Apart from typesetting changes and language localization, this chapter appears exactly as submitted to *Management Science* on 9 February 2023.

Supplementary Publications

Research that informed this chapter also appeared in the following outputs.

Type	Citation
Conference Paper	Wysel, M. (2021, 22 June 2021). Data Sharing Platforms in Agribusiness. International Food and Agribusiness Management Association, Costa Rica.
Journal Paper	Burgess, S., & Wysel, M. (Eds.). (2021). <i>China’s Social Credit System: How Robust Is the Human Rights Critique?</i> https://doi.org/In press .
Conference Paper	Wysel, M., & Baker, D. (2021, December 2021). Sandwiches Vs. Genes. Sharing Data to Maximize Its Value. MODSIM2021, 24th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, Sydney.
Book Chapter	Fomiatti, B., ⁴⁴ & Wysel, M. (2022). Developing the Value of Smart Agriculture through Digital Twins. In <i>Encyclopedia of Smart Agriculture Technologies</i> . Springer. https://doi.org/10.1007/978-3-030-89123-7_273-1

Chapter-level Glossary

Term	Definition
Non-rivalry	<i>Noun.</i> The property of a good or service that permits its simultaneous consumption with an absence of congestion. Examples include clean air or public parks.

⁴⁴ Blake Fomiatti was a graduate student of the primary author in 2021.

Excludability	<i>Noun.</i> The property of a good or service which permits the prohibition or exclusion of its use.
Data network effects	<i>Noun.</i> The simultaneous creation of value enabled by the non-rivalry and excludability exhibited by data. Initially posited by Gregory et al. (2021) to explain how the enrichment of data by a system (such as AI), drives increases in the value of data for users such that they create more data – and therefore value – for the system.

Full Manuscript

Profiting from data. How data enables firms to have their cake, sell it, and eat it too.

Authors:	Matthew Wysel ^a (matthew.wysel@une.edu.au)*, Derek Baker ^a (derek.baker@une.edu.au).
Affiliations:	^a The Centre for Agribusiness, UNE Business School, The University of New England, Armidale, Australia.
Keywords	Value of data, non-rivalry, data network effects, data-enabled learning, data-based production function
Acknowledgements:	The authors wish to thank Alison Sheridan for her helpful suggestions particularly regarding 'the 5Cs'. Matthew Wysel is grateful for the support of the Agricultural Business Research Institute through the Arthur Rickards Innovation in Agribusiness Scholarship. All omissions remain our own.

Abstract

Data is increasingly the most valuable asset firms manage, yet both scholars and practitioners want for a generic production process that maps, in plain terms, how firms create value from that data. Without this framework, practitioners remain unable to calculate efficiencies or benchmark processes and scholars remain unable either to organize the rapidly expanding corpus of phenomena-centric data management theory, or to engage practitioners with grounded predictions, derivative causalities, or even idealized scenarios. This paper first presents, and then formally models, a data-based production process that connects the effect of management decisions with the parallel payoffs that data enables across a firm. The process is developed as an extensible framework that incorporates the non-rivalry, excludability, and conditional network effects exhibited by data as well as the fixed and variable costs incurred by the firm. We illustrate how one decision to invest in data enables the firm to simultaneously sell their data, use it to improve other products and services, and incorporate it as data-enabled learning. The model is applied to management of internal 'data projects' and externally connected, data-based business models. We show how to calculate an over- or under-investment in data and identify conditions where constantly updating insight creates sub-optimum profits. Data as a scarce resource is discussed and an alternate framework for data-enabled learning is presented. The proposed theory also acts to frame the broader management challenge of value creation versus capture between data trading platforms and their participants.

1. Introduction

The operation of any modern, competitive firm involves the production of value from data. The generalized process occurs in both tech and non-tech businesses (Kaiser et al., 2021; Rahmati et al., 2020), and in both large and small firms (Liu et al., 2020; Short & Todd, 2017) and incorporates the enrichment of data by a firm towards a valuable goal (Golman et al., 2022). For instance, sales data, once analyzed, improves inventory management, refines product mixes, and facilitates data-enabled learning (Hagiu & Wright, 2020a). The collection of customer's behavioral data enables platforms to refine pricing structures, and refining profits for both themselves, and their producers (Bhargava et al., 2022). Once enriched into insight⁴⁵, data can also be traded directly using data-based business models (Hartmann et al., 2016) that rely on the firm's ability to simultaneously sell insight, use insight to improve other products or services, and reinvest insight into the operations that power their data-based, production process.

If data was a 'normal' production factor like labor or capital, then firms would need to choose which payoff to pursue and optimize inputs and management decisions accordingly. After all, a baker cannot ice two cakes at once, nor can she eat a cake after selling it. However, data is non-rivalrous (Jones & Tonetti, 2020) and exhibits near-infinite economies of scale (Arrow, 1996). This permits firms to simultaneously sell insight even while they use the same insight to both improve ancillary products and increase the value of operations. Firms such as *23andMe* ingest customer's genetic data into proprietary technologies to produce reports on disease susceptibility which are sold back to customers. In parallel, once aggregated this insight also fuels research efforts within and beyond the firm (Hayden, 2017). Firms that operate platform business models, such as *LinkedIn* and *Uber*, enrich behavioral data for insight into customer preferences that is traded to advertisers or intermediaries in ancillary markets, used to improve existing services, and reinvested to refine their platforms (Parker & Van Alstyne, 2018).

However, while data may be non-rivalrous the resources required to invest in, and manage, data remain rivalrous. The growing centralization of data as a firm asset requires managers to justify investments in data management to increasing levels of sophistication (Pentland et al., 2021). Data-related expenses like storage, categorization, and analysis need to be justified against associated returns while investments into data such as acquisition, extraction or generation of data need to be considered against the future revenue streams they will support. Even where data-based

⁴⁵ We adopt the noun *insight* rather than *information* throughout to avoid confusion caused by the conflation of the terms data and information in popular media. Insight is data that has been processed by a system, such as a firm, towards a valuable goal.

investments stay within practical limits, firm-level policies often mandate minimum rates of return for projects and increase the pressure on managers to evaluate data projects in a similar manner to non-data projects within the firm. Likewise, as data-based production processes often operate alongside other non-data processes, the investments they require must be justified against the opportunity cost of providing those resources elsewhere within the firm. More broadly, the interfirm assessment of data and trade of data-based production processes requires benchmarking data-based operations using either common metrics or standardized approaches to data valuation (Fleckenstein et al., 2023).

Yet, leaders of even large well-resourced firms struggle to adequately manage the production of value from data within their operations. They note that while their data managers “were highly effective in storing and protecting data, they alone cannot make the key decisions that transform data into business value” (Short & Todd, 2017, p. 18). This inability to manage the data-based production process stems from the lack of a framework that articulates it in the simple terms of resource allocation, productive outputs, and management interventions (Pentland et al., 2021). Without a framework, the day-to-day decision making required by a data-based production system becomes confused (Chiang et al., 2018) and leaves practitioners with the all-too-common feeling of ‘drowning in data’ as they feel unable to prioritize investments and remain tied to the maintenance of existing data management efforts rather than empowered to adapt to new demands (Short & Todd, 2017). On the one hand, the firm-level process of producing value from data is evidently effective, but on the other hand, the multiplicity of ostensibly non-rivalrous payoffs that data enables leaves managers and scholars without a grounded framework from which standard metrics or operating strategies could be derived.

The incorporation of data as a valuable resource within a firm has received some scholarly attention. The academic treatment of data as a resource and the management of data within a firm give managers operational considerations for the treatment of data – particularly across markets (Easley et al., 2018) – and the implications of these considerations on the operation of the business (Wysel et al., 2021). The former focuses on managing data with implications for the firm (see for example, Hagiwara and Wright (2020a)) while the latter on management of the firm with implications for data (for example, Gregory et al. (2021)). However, neither offers practitioners an extensible framework that connects decisions such as, ‘how much – or how often – should I invest in data?’ to immediate payoffs or longer-term value creation. While the phenomenon of transforming data into value across a business certainly exists, standard resource management and production process theories stop short of explaining the exercise as an end-to-end process. Data managers are left standing in the

middle of this syzygy, as their practice of transforming data into value within a firm becomes increasingly isolated from the theory that ought to explain it.

As scholars, we also require a theory on data management that unites those attributes of data that determine data-based value creation, allocation, and capture with the mechanics of the process (Grover et al., 2018). This endeavor requires more than simply coalescing extant data management theory. A generic production process for firm data must reconcile theory from across production, innovation, and firm management literature as well as from within economics and information systems. For instance, under what conditions might a firm rationally refrain from further investment in data (e.g. Sims (2003) or more recently Golman et al. (2022)) even though increases in data notionally accompany increases in productive output (Gawer, 2022)? Or how might the intrinsic properties of data – such as excludability and non-rivalry – affect its specification as a factor of production within a firm (Pentland et al., 2021)? Indeed, treatment of data as an intangible asset has precedent (Farboodi et al., 2019) but how might conditional (Clough & Wu, 2022) data network effects (Gregory et al., 2021) be leveraged so a firm can both accumulate value from data *and* distribute value among stakeholders (Cusumano et al., 2019)?

We address this gap by modeling the effect of management decisions on the parallel payoffs a firm receives from data. In essence, we propose these parallel payoffs and firm-level decisions constitute a definable, data-based production process which supports the derivation of both internal and market-based operating strategies by the firm. Specifically, as depicted in Figure 4-1, the production process includes both external and internal payoffs for the firm from the development of data. External payoffs include the sale – or valuation through exchange – of insight and the application – or valuation in use – of insight. Meanwhile, the internal payoff includes the internalization of insight as data-enabled learning which has the effect of increasing the value of the firm’s data- and non-data-related operations (Shapiro & Varian, 1998). As expanded in Table 4-1, arrangement of the production process in this manner permits adaptation to a wide variety of both *tech* and *non-tech* business models and operational configurations. Firms that operate data-based business models invest in data to achieve a return from the data directly (Hartmann et al., 2016). They apply

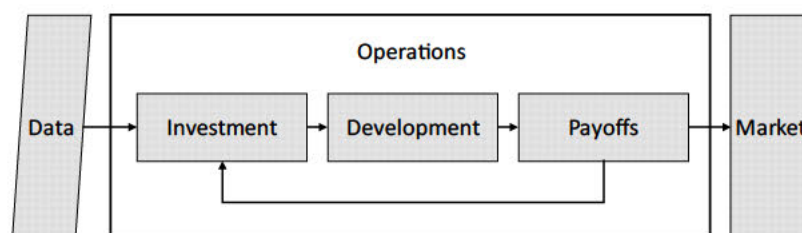


Figure 4-1. The Generic Data-based Production Process

proprietary technology to aggregate, analyze or synthesize new types of data which augments existing products or enables the creation of new products. Data analysts (e.g. *23andMe*), data traders (*Spotify*), search engines (*Google Search, Bing*) and social media platforms (*Instagram*) each operate in this manner. Alternatively, firms that operate non-data related business models, such as retailers (*Walmart*) or service providers (*The Hilton Group*), collect data to enable both data, and non-data returns (Bardhan et al., 2020). In this case, insight derived from a firm's operational data supports the ongoing improvement of non-digital products as well as the improvement of operations that deliver those products. A hotel chain's personalized marketing of services, or the analysis of complementary products conducted by retailers are examples of this interpretation of a data-based production process. In each case, managers must decide what portion of available resources to invest in data and how frequently to update their analysis to reflect newly generated data.

Incorporating this first decision, we analyze the proportion, σ , of existing operating capital a firm decides to invest in data. A decision to invest in data enables the generation, extraction, capture, synthesis, or purchase of data.⁴⁶ Only generated data can be developed by the firm into insight. Insight can be sold, used to improve other products, and captured within the firm's operations. While each of these payoffs represent potentially valuable returns on the firm's initial investment, this investment also carries an opportunity cost. If a firm invests σ in generating data, it only has $(1 - \sigma)$ to allocate to other projects as investment in generating data represent reductions in available resources for other projects across the firm. However, data that was not generated is left 'latent' in operations and reduces the potential insight a firm might have generated from an increased investment.

We also address the second decision regarding the frequency of generating new insight. The generation of new data and with it, new insight, permits new insight to be sold, applied, or incorporated within the firm. Specifically, we ask the question, when is it profit maximizing for a firm to *constantly* analyze and respond to data? Often a firm's operations may produce near continuous streams of data, but managers must still determine the frequency with which to incorporate updates or to apply insight within ancillary products. While a continuous release of up-to-the-minute insight may hold popular appeal, as noted in the opening discussion, this imposes significant costs on the production process and might reasonably even reduce potential value creation captured by the firm. To frame this decision, we define the time between consecutive releases of insight as an interval of

⁴⁶ Throughout this paper we refer to a firm that 'generates data', however our treatment proceeds unchanged if the firm *purchases* data from an external party/ies, *extracts* data from assets it controls, or expends resources to *synthesizes* data that has been amassed from other operations.

length t , where $t \geq 0$, to accommodate analysis of the continuous and episodic scenarios. Note that $t = \infty$ indicates a firm has decided to cease operation of its data-based production system.

The purpose of this paper is to first propose, and second to model, a framework that describes how firms produce value from data via internal and external payoffs that can operate in the absence of any apparent rivalry. Stated in terms of the popular metaphor, the proposed framework illustrates how data permits firms to have their cake, sell it, and eat it too, while the model explains how to optimize this ubiquitous recipe. Taken together, the data-based production process is a generic, readily adaptable model that incorporates the non-rivalry, excludability, and conditional network effects exhibited by data and offers practitioners and scholars insight into existing and complimentary value creation efforts.

To the best of our knowledge, this is the first paper that adds a production function to a model that describes the creation of value from data within a firm. We first introduce data as a factor of production before developing each of the generic steps from Figure 4-1 to describe the actions a firm takes in its production of value from data. Specifically, we examine the firm's generation of data, the decisions surrounding the development of that data, and the parallel production of value from data both internally and externally. Combining these steps, we chart the progression of data from an input to an output across a firm. The framework incorporates the non-rivalry of data through the parallel realization of normally exclusive payoffs. The firm may simultaneously trade insight in the market, apply insight to improve its own products or services, and reinvest insight by capturing the value insight unlocks via data-enabled learning. The excludability of data is incorporated through the requirement that until the firm invests in data, the data remains outside the firm's production process and beyond its control. All capitalized costs are reflected in this initial investment, while in-period, data-related expenses are introduced in Section 4. Finally, conditional data network effects affect value created from data once correlation between payoffs is introduced in Section 5. In this way, the framework serves as a quick-reference map for those looking to uncover additional value-creating paths and as a basis for the formal model.

Attaching parameters to the framework, we formally model the process a firm undertakes in the production of value from data as both an isolated data-project and as an ongoing production process nested within broader operations. We model the effect on firm profit of the firm's two central data-based, decisions: "how much should be invested in data?", and "how often should the production process be repeated?". Accordingly, the proposed model addresses three key data management challenges:

- i) the evaluation of data-based projects against other firm-level projects,

- ii) the assessment of the firm-level, data-based production process, and
- iii) the optimization of firm-level investment strategies surrounding existing data-based production processes.

The model is run over two stages to capture the effects of data-enabled innovation both within the firm and from the market. We demonstrate how data projects may be evaluated against the firm's internal cost of capital using the standard methods of net present value (NPV) and the internal rate of return (IRR). This permits practitioners to baseline data projects against other, even non-data, projects within – and beyond – their firm. We show it is possible for firms to over-invest in data and generate conditions where constantly updating insight with the most recent data creates sub-optimum profits. Given the centrality of data-enabled learning to value creation within firms, we conclude the paper with an interpretation of data-enabled learning from the perspective enabled by the proposed framework. Specifically, we propose new methods for modeling value created across the firm by data-enabled learning and the compounding effect of the increased learning on the data-derived payoffs.

Table 4-1. Generic data-related payoffs for various business models

Example Firm	Generic business model	Primary Data-related Payoff	Secondary Data-related Payoff
<i>Hotel Chain e.g. Hilton Worldwide</i>	<i>Delivery of non-digital service</i>	<i>Use insight to improve service</i>	<i>Value captured as data-enabled learning</i>
<i>Retailer e.g. Walmart Inc.</i>	<i>Delivery of non-digital product</i>	<i>Use insight to improve products</i>	<i>Value captured as data-enabled learning</i>
<i>Professional Service Agency e.g. Marsh</i>	<i>External delivery of service</i>	<i>Trade on data-derived insight</i>	<i>Use insight to improve services</i>
<i>Multisided Platform e.g. LinkedIn, Uber</i>	<i>Service provision; data generation and analysis</i>	<i>Captured as data-enabled learning, Use insight</i>	<i>Trade (access to) insight into user behavior</i>
<i>Data analyst e.g. 23andMe</i>	<i>Analytics-as-a-service</i>	<i>Trade insight derived from customer-supplied data</i>	<i>Value captured as data-enabled learning and used to power external research</i>
<i>Data trader e.g. Spotify</i>	<i>Supplier of curated data</i>	<i>Use insight to improve service</i>	<i>Value captured as data-enabled learning</i>
<i>Search engine e.g. Google Search, Bing</i>	<i>Data collector and aggregator; supplier of curated data</i>	<i>Trade (access to) insight; use insight to improve service</i>	<i>Value captured as data-enabled learning</i>
<i>Social Media Platform e.g. Instagram</i>	<i>Data knowledge discovery</i>	<i>Trade (access to) insight into user behavior</i>	<i>Use insight to improve service; value captured as data-enabled learning</i>

2. Data as a Factor of Production

The treatment of data as a firm-level factor of production builds on the foundations of datanomics, production and innovation literature as well as those of economics and information systems. Data is “one of the fundamental determinants of production” (Arrow, 1996, p. 127) and passes easily into, and out of, the firm as it undergoes a transformational change in both form and value (Frankel & Kamenica, 2019). This process of production takes both capitalized and in-period investments to enable the valuation of data as a resource (Stieglitz et al., 2018), good (Schatz & Bashroush, 2016), or even currency (Morozov, 2016). Accordingly, while data often resembles other assets, such as capital or labor, it exhibits unique properties that affect its treatment as a production factor (Pentland et al., 2021) and within a production process (Arrow, 1996).

Data is excludable (Easley et al., 2018), non-rivalrous (Jones & Tonetti, 2020), and displays conditional (Clough & Wu, 2022; Hagiú & Wright, 2020a) data network effects (Gregory et al., 2021). Study of data as a digital asset drives research in specific disciplines (Cennamo et al., 2020), applications (Wolfert et al., 2017) and phenomena (Günther et al., 2017; Hinz et al., 2020). Accordingly, the science of creating value from data may serve as a programmatic (Lakatos, 1968; Wagner & Berger, 1985) theory within management science to the extent it orients and supports the reconciliation of extant unit theory pertaining to the management of data as a firm asset (Cronin et al., 2021). However, as programmatic management theory, a production process that describes the creation of value from data ought to also support the fine-grained, applied science of managing the resource (Arend, 2022) from an input to a firm’s operations in its transformation to an – ideally – more-valuable output.

2.1. Data and Value

Establishing the terms of that transformation, the value a firm creates from data is taken as the difference between market payoffs received for the management of data and the costs incurred to achieve those payoffs. This treatment has precedent within management literature (Chesbrough, 2003; Günther et al., 2017) and accords with common experience. Insight from customers’ behavioral or social networks, data typically requires an initial investment by the firm but also enables increases in short- (Loveman, 2003) and long-term value (Hagiú & Wright, 2020a; Sharapov & MacAulay, 2022) as the data is sold, utilized, or reinvested. Value created from the management of data within a firm is a function of both assets controlled and decisions made by management regarding those assets (Horling & Kulick, 2009). Accordingly, this treatment applies to both internal (pipeline (Choudary, 2014)) and external (platform) value creation. Whether by uniting a community

of stakeholders as ‘users’ or engaging with them across a market, both approaches to production utilize an enabling system to transform data on, from, or for that community into value (Wysel et al., 2021). In this arrangement, the system utilizes data to both enable and incentivize stakeholders to make valuable interactions (Evans & Schmalensee, 2016) either within or ‘on’ the host firm. Acting as the enabling system, the firm directs the production process surrounding data, determines an effective structure for commercializing that data (Hartmann et al., 2016), and attempts to broker a sustainable distribution of value between the firm and stakeholders (Gregory et al., 2022).

The utilization of data as a factor of production carries both a capitalizable cost and an in-period expense to the firm. Operational value diverted to the former permits assembly of assets from which future revenue will be derived, while operational value diverted to the latter diminishes in-period profits. For instance, the excludable nature of data requires a firm to input data into a production process typically through extraction, generation, synthesis, or purchase⁴⁷ (Hartmann et al., 2016). As such, resources spent on generating data represent the accretion of a digital asset from which future value will be derived. Notably, latent, or unextracted, data can confer value to a firm where it becomes intertwined with the surrounding assets (Gregory et al., 2021; Short & Todd, 2017) such as operations (Short & Todd, 2017) or a community of external stakeholders (Adner & Kapoor, 2010). Either way, following an investment in data, additional, in-period, expenditure is made on data by management decisions (Zellweger & Zenger, 2021), technology such as Artificial Intelligence (Gregory et al., 2021), or organizational responses such as firm-wide innovation (Gomes et al., 2018). In all cases, the commercial allocation of resources to data enrichment is predicated on the expectation of the short- or long-term output of value that exceeds the input cost.

2.2. Enriching Data

In its full abstraction, operation of this production process begins to resemble a loosely organized, data sharing ecosystem. From Figure 4-1, this process proceeds as follows: a firm makes investments that subsume data into operations which, following a series of interventions, enable a payoff in the market. Accordingly, where prior approaches have modelled the process as a series of ostensibly homogenous interactions involving data (Wysel et al., 2021), we adopt the opposite perspective and investigate the interactions that comprise the process. That is, rather than investigating how creating value from data affects the firm – as a collection of assets, management decisions and

⁴⁷ Throughout this paper we refer to a firm that ‘generates data’, however our treatment proceeds unchanged if the firm *purchases* data from an external party/ies, *acquires* data from assets it controls, or expends resources to *assemble* data that has been amassed from other operations.

organizational responses – we investigate how operation of the firm affects the creation of value from data. To that end, we model a firm that takes data as an input and develops that data into insight across a small number of discrete, operating periods. This insight creates value internally as innovation, and externally in both use and exchange through the market.

The process of development we employ is similar to the model adopted by Parker and Van Alstyne (2018) when describing the extension of intellectual property by developers across a microeconomy. Varying from their approach, we model production decisions within a firm where value is shared (Gregory et al., 2022) and recursively invested (Arthur, 2009). Insight captured by the firm improves the value of operations (Sharapov & MacAulay, 2022) which compound the effect of subsequent data operations. This causes the future value of data to increase as data enables the firm to learn from past insight and improves the efficacy of future investments. Additionally, while resources applied to develop data exhibit diminishing returns as in Tirole (1988), the value of outputs compound as each iteration builds on prior insight retained. Like Prüfer and Schottmüller (2021), we model a dynamic feedback loop where the strategic significance of a firm's initial investment compounds, causing equivalent improvements in efficacy with diminishing relative costs. The data-based production process causes the value of digital assets to become intertwined with technologies, management intervention (Agrawal et al., 2018), and organizational models (Langley & Leyshon, 2017) as the firm strives to improve the utility of insight for its target market.

2.3. Insight is Enriched Data

Insight is data that has undergone the removal of uncertainty (Frankel & Kamenica, 2019) that surrounds it with reference to a predefined goal (Wysel, 2023) in a process akin to enrichment of minerals from ore. The process is executed by a facilitating system and for a surrounding community (Wysel et al., 2021). Insight retains data's properties of non-rivalry, excludability, and conditional data network effects. Therefore, insight also supports multiple, simultaneous valuations by the firm. Firms may simultaneously realize payoffs from the exchange and use of insight, as might occur when an advertising firm sells a benchmarking report to one client, uses the insight to substantiate consulting services to another, all while applying the insight to improve its own operations.

While these value streams may be complementary, the transformation of data into value-in-exchange and value-in-use represent different activities (Vargo et al., 2008) and warrant distinct treatment by the firm. Initially, we model these payoffs as strictly additive where the firm competes neither with itself nor with its stakeholders. We subsequently relax this assumption when examining the effect of interactions between data markets in Section 5.

Finally, the non-rivalrous nature of insight is not time-bound. Consumption in one period does not diminish the value of consumption in a subsequent period (O'Reilly, 2017). This creates an operational tension where the value of insight simultaneously decays as its relevance wanes (Pentland et al., 2021) and accumulates as insight is reinvested into the firm through data-based learning (Hagiu & Wright, 2020a). The proposed framework offers a generic treatment where the operational effectiveness of insight is characterized by prior investments in data and ongoing data management decisions⁴⁸ and serves to illustrate the mechanics of the phenomena others have postulated: that data network effects exist independently from markets and are subsets of the broader concept of data-enabled learning (Hagiu & Wright, 2020a).

As a factor of production, data is excludable, non-rivalrous, and exhibits conditional network effects which together, enable simultaneous, sometime competing payoffs for a firm. Significantly, our theory encompasses a very general set of use cases, and yet can easily be made specific. On the one hand, the theory describes the general relationships within a production function; on the other, it may be applied to answer the very specific questions, 'how much should I invest in data?' or 'how often ought I repeat that process?'. To the best of our knowledge, the proposed theory is the first time such an expansive definition of data as a production factor is brought together with the very practical exercise of creating value from data within a firm.

3. A Firm-level Framework for the Production of Value from Data

3.1. Investing in Data for Non-rivalrous Value Creation

From Figure 4-1, the production of value from data begins with a decision by the firm to invest in the generation of data from its operations. Let the total available capital the firm could invest in data be V and the amount invested in generating data from operations be X . X may be connected to σ , the proportion of capital invested in data by noting, $X = \sigma V$ where $\sigma \in [0,1]$. The decision to invest X

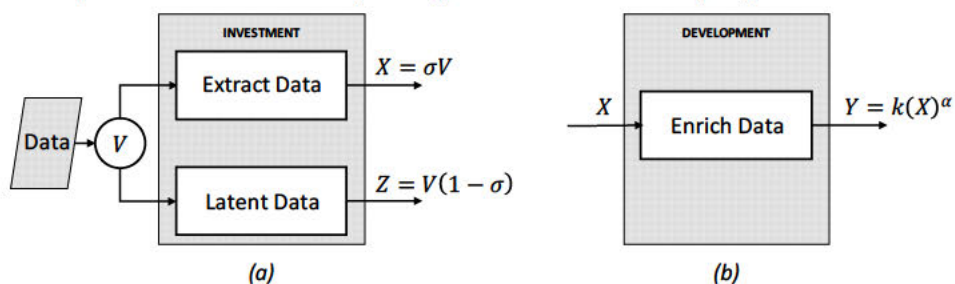


Figure 4-2(a) Investment in Data (b) Data Development

⁴⁸ This arrangement also supports the orientation and evaluation of emerging data-enabled learning literature.

into data reduces the firm's working capital to a residual, Z , such that $Z = V - X$. Here Z represents the opportunity cost of generating data from its operations. Initially X includes all data management costs and the firm receives *face value* for Z ; that is, the payoff for *not* generating data from operations, W_r , remains equal to the residual. Explicating the boundary conditions, when $\sigma = 0$, $X = 0$ and $V = Z = W_r$; that is, the firm chooses not to invest anything in data. Alternatively, when $\sigma = 1$, the firm invests all available capital into generating data and $X = V$ and $Z = W_r = 0$. This initial decision is represented in Figure 4-2(a).

The generated data serves as an input to a management activity designed to increase the value of data to the firm. This digital production process may be modeled as a single-input, Cobb-Douglas production function (Müller et al., 2018) where data is enriched into insight through the application of proprietary technologies. The transformation is illustrated in Figure 4-2(b) and takes the form $Y = k(X)^\alpha$. As in other digital production processes (Tirole, 1988), k represents real output per unit input that stem from the strategic adoption of new technologies and $\alpha \in (0,1)$ represents the diminishing effects of the technology used to enrich data into insight.

As introduced in Section 2, the non-rivalrous property of data enables the firm to generate value from insight across multiple channels simultaneously. Therefore, within restrictions imposed by the market, the firm can realize payoffs from insight *in exchange* and *in use* simultaneously. Insight valued in exchange follows a *goods-dominant* logic (Vargo et al., 2008) and provides the firm a direct valuation of enriched data from which it receives a payoff, p . Examples of the sale of insight include the sale of targeted advertising space on social media feeds or genomic sequences to research partners (Stoeklé et al., 2016).

As illustrated in Figure 4-3(a), the firm is able to simultaneously employ insight using a *service-dominant* logic by co-creating value with market participants (Vargo et al., 2008). This valuation in use occurs where the firm uses enriched data to improve a product or service. In this case, the firm receives an indirect valuation as each unit of insight confers a change in value, v , in another product. However, this marginal impact of insight on the product is transitory as the market anticipates the release of new insight by the firm next time it generates and enriches new data. Release of new insight immediately depreciates old insight, effectively bundling old insight into the outgoing product (Varian, 2018). This implies the maximum the market will be willing to pay is the difference

between the value of current insight, v , and the present value of insight from the next stage, δv .⁴⁹ Therefore, the maximum *use value* of the insight to the firm is $v - \delta v$ or $v(1 - \delta)$, where $\delta \in (0,1)$.

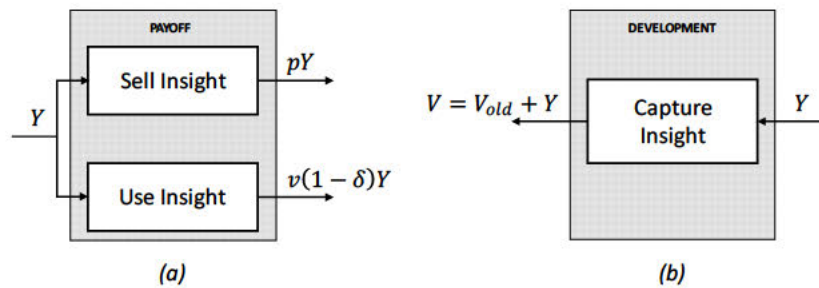


Figure 4-3(a) Immediate Payoffs, and (b) Reinvested Value from Data-derived Insight

In addition to the two external payoffs, insight also supports the simultaneous creation of value internally with the firm reinvesting insight into existing operations. This has the effect of increasing the firm’s digital capital and increasing the flow of services useful for ongoing operations (Tambe et al., 2020). As depicted in Figure 4-3(b), reinvested insight permits the firm to learn as it captures value into its current operations. Where previously the firm’s technology supported the development of data into insight, captured insight now supports the development of the firm’s technology. This recursive value creation process enables the phenomenon variously described as *data network effects* (Gregory et al., 2021) or *data-enabled learning* (Farboodi et al., 2019; Hagiu & Wright, 2020a) as the firm’s data is fused with technology causing an increase in the value of both.⁵⁰ As postulated by Hagiu and Wright (2020a), this process does not require a marketplace but can also occur within a firm’s operations. Meanwhile, the utility of this process is a “function of the scale of data-driven learning and improvements realized with AI” (Gregory et al., 2021, p. 535). Initially we represent the rate a firm internalizes the value of insight as the monotonic sum of the output of the firm’s data enrichment process. This sets the new value of a firm’s operations to,

$$\begin{aligned} V &= V_{old} + k(X)^\alpha \\ &= V_{old} + Y. \end{aligned} \tag{4-1}$$

We adopt this straightforward arithmetic form to simplify the following mathematical treatment, but the rate at which a firm learns can easily be adapted to fit scenarios of interest. For example, $V = V_{old} + \gamma Y$ where $\gamma > 1$ would reflect a firm whose operations multiply the effect of data-

⁴⁹ δv is also mathematically equivalent to the *residual book value* of insight written off by the firm when it generates another set of data at the start of the next stage.

⁵⁰ In the case of data network effects, the payoffs pursued by the firm materially affect its ability to generate new data in real time, implying $t = 0$ and correlation of outputs and inputs. We return to investigate these conditions in Section 5.2.

enabled learning. We return to the implications of different learning rates when discussing the practical implications of the proposed framework in Section 5.

There are two corollaries of the proposed data-driven production process that are worth explicating. As data is enriched by current technology, subject to the specification of the coefficient in Equation (4-1), the value of insight must increase with each successive iteration of the process. The immediate implication is that a firm's decision to generate new data must immediately depreciate the value of old insight. This confirms the earlier observation that neither the firm nor the market would choose to use old insight once newer insight was available. Second, data that is generated but not enriched amounts to a cost incurred by the firm for zero potential payoff. Therefore, the firm would attempt to enrich all data generated before restarting the production process and generating more data.

Proposition 4-1: The non-rivalry, excludability and conditional network effects of data are necessary for a firm to simultaneously create value from the exchange, use and reinvestment of enriched data.

A firm creates value from data through a production process where data is first generated, then enriched into insight, before being valued by the market and reintegrated into operations. From Section 2, the excludable nature of data implies that data remains separate from a firm's operations until resources are expended to generate it. These resources constitute a capitalizable cost for the firm and impute an initial value to the data. Ongoing data management expenses are separate to this investment as they diminish in-period returns. The non-rivalrous nature of data permits enriched data to be simultaneously sold, used, and reintegrated into operations without congestion, while the conditional data network effects exhibited by data enable insight to drive an instantaneous increase in value in current operations and also contribute to possible correlations between payoffs received by the firm. Therefore, subject to the nature of these conditional data network effects, the specific attributes of data noted in Proposition 4-1 enable a firm to operate a data-based production

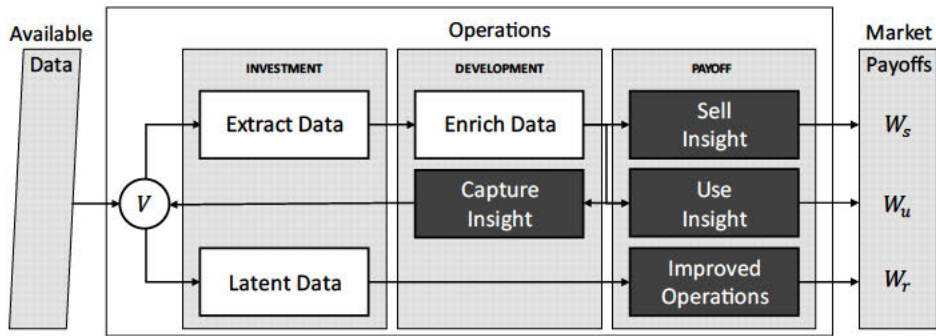


Figure 4-4 The Expanded Data-based Production Process

process where value is created from data through simultaneous, non-rivalrous payoffs. This relationship is illustrated in Figure 4-4.

In this way, a single decision to take data as an input to a production process within a firm enables up to three external payoffs, and the accrual of value internally through the reinvestment of insight within the firm. The framework illustrated in Figure 4-4 is the assembly of the steps outlined in Figures 4-2 and 4-3. This framework acts as a map for the firm-level, data-based production process and a starting point for the subsequent model development. Note, the two decisions that govern value flow within this framework – ‘how much – and how often – should the firm invest in data?’ – are accommodated by the framework. The effect of these two decisions on value creation is summarized in Table 4-2.

Table 4-2. Relationship Between Data-Related Management Decision and Operational Goals

<i>Management Decision</i>	<i>Change in Parameter</i>	<i>Operational Goal</i>
<i>Increase investment level</i>	<i>Larger σ ($\sigma \rightarrow 1$)</i>	<i>Increase insight-based payoffs and data-enabled learning</i>
<i>Decrease investment level</i>	<i>Smaller σ ($\sigma \rightarrow 0$)</i>	<i>Increase value retained in operations</i>
<i>Increase release rate (shorter stage length)</i>	<i>Smaller t (Larger δ)</i>	<i>Increase rate of data-enabled learning</i>
<i>Decrease release rate (longer stage length)</i>	<i>Larger t (Smaller δ)</i>	<i>More time to use insight</i>

We turn now to model this recursive, data-based, value creation process.

3.2. The Model: Recursive Value Creation from Firm Data

The model operates over two, sequential stages and describes a firm-level, production process that takes data as an input and produces value as an output. The firm must make two decisions: the proportion of value invested, $\sigma \in [0,1]$, and the time between consecutive releases of insight, or the

stage length, $t \in [0, \infty)$. Within each stage, value is produced from data across the same three phases introduced in Figure 4-1 and expanded in Figure 4-4.

At the beginning of the first stage the firm makes an investment, x_1 , to generate data. The firm uses Cobb-Douglas production technology to enrich x_1 to y_1 according to $y_1 = k(x_1)^\alpha$. At the end of the first stage, the firm simultaneously realizes a payoff from the sale of insight, $w_{s1} = py_1$, and the use of insight, $w_{u1} = v(1 - \delta)y_1$, in the market. In parallel, the firm also realizes the shadow value of latent data, $w_{r1} = V_1(1 - \sigma)$. Finally, the firm reinvests insight into operations, setting $V_2 = V_1 + y_1$. Therefore, at the end of the first stage the firm has created value from data through three market-based strategies and one strategy predicated on long-term operational improvement. Note that the discount rate, δ , may be formally connected to the stage length, t , and the effective interest rate of insight, r , using the form $\delta = e^{-rt}$ where $t \in (0, \infty)$ and $r \in (0, 1)$.

The second stage proceeds in the same manner as the first. Taking investment in data as a strategic decision, we hold the proportion of value invested constant across the model. Therefore, $x_2 = \sigma V_2 = \sigma(V_1 + y_1)$. The shadow value of latent data and operations has also increased, $w_{r2} = V_2(1 - \sigma)$. Insight derived from data generated during the second stage is $y_2 = k(\sigma V_2)^\alpha$ which is reflected in both the direct and indirect payoffs from the market, $w_{s2} = py_2$ and $w_{u2} = v(1 - \delta)y_2$, respectively. At this point, the firm has created value from data across two stages and the model ends. This schedule of value flows is depicted in Figure 4-6.

Profit derived by the firm from data-related operations, π , may be expressed as the sum of each market payoff,

$$\pi = W_s + W_u + W_r. \tag{4-2}$$

Figure 4-5 graphically unites the abstracted production process from Figure 4-4 with the mathematical model proposed.

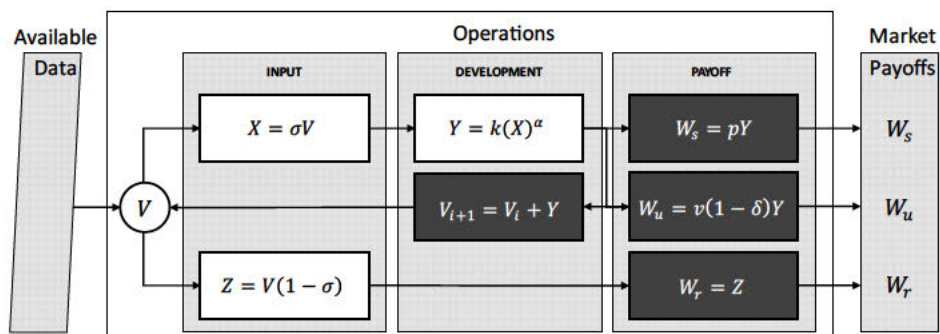


Figure 4-5. The Expanded Data Production Process (as expressions)

For tractability, let both stages be of equal length, t , and let the effective interest rate of insight, r , remain constant whether in exchange, in use, or in reinvestment.⁵¹ For practical relevance, all investments and expenses are committed at the beginning of each stage, while payoffs are realized at the end of each stage. Definitions for parameters introduced in the framework and incorporated in the model are listed in Table 4-3.

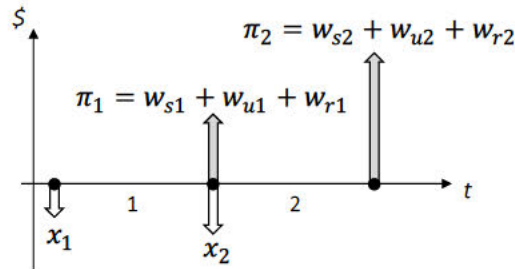


Figure 4-6. Schedule of Value Flows

Table 4-3. Variables for Production Process for the Creation of Value from Data

Parameters	Interpretation
V	Current value of firm's operations
σ	Proportion of firm's current operations invested into data
X	Size of investment in data; expressed as a proportion of current operational value
k	Coefficient of a firm's utilization of insight; the effectiveness of insight within the firm.
α	Efficacy of the enrichment technology employed by the firm
Y	Insight; enriched data; output of firm's enrichment process
F, c	Fixed and variable costs of enriching data into insight
p	Price of insight in the market
v	Marginal value of applying insight to firm's existing products or services
δ	Discount rate of insight in exchange, use or reinvestment. Isomorphic to stage length, t
$W_{s,u,r}$	Payoff from sale of insight, use of insight, or retaining latent data in operations, respectively
π	Firm profit following investment in data

⁵¹ While both constraints are useful for simplifying the following exposition, we acknowledge that variance in either does enable novel analytical frameworks such as proposed by Hagiu and Wright (2020b).

4. Results for Data Projects, Data-based Production Processes, and Optimal Investment into Data

4.1. The Evaluation of Data Projects

A project is characterized by an upfront investment that creates an episodic revenue stream. Common examples of data projects include the deployment of an analytics package or commissioning the mining of sales data for operational insight. Specification of data projects in this manner would also encapsulate activities like launching a digital currency through an initial coin offering (ICO) (Cennamo et al., 2020) or investigating data governance models that promote adoption of IoT infrastructure (Koohang et al., 2022). In each case, the central question for project managers is, can the planned investment in data generate sufficient value from each stage's data-related payoffs across the length of the project? Typically, evaluation of these projects is couched in terms of either an acceptable IRR or positive NPV. As before, the key intra-firm drivers in this production process are the efficacy of the technology used in preparing, analyzing, and reporting the generated data, α , and the utilization of the insight within the firm, k .

Proposition 4-2: The present value of a data project is more sensitive to the firm's utilization of insight, k , than to the efficacy of the enrichment technology employed, α .

Following the schedule of value flows in Figure 4-6, if the firm makes an initial investment, x_1 , and chooses never to pursue a data project which required a larger investment than the profit currently received (that is, all value flows that follow the initial investment stay net positive) then the standard equation for NPV can be adapted to include the schedule of value flows from Figure 4-6 and written,

$$NPV = \sum_{i=1}^T \frac{\pi_i - x_{i+1}}{(1 + IRR)^i} - x_1. \quad (4-3)$$

A very practical decision faced by data managers is whether to increase the firm's responsiveness to data or to improve the quality of the data asset itself (see, for instance, Hagiu and Wright (2020b)). Strategic decisions of this type occur when managers choose between investments in firm-level data-based decision making or upgrading a key data management system. The immediate implication of Proposition 4-2 is that a firm seeking to increase the present value of a data project ought to focus more on increasing the utilization of insight within the firm rather than on improving the technology responsible for enriching the data. This result extends prior frameworks that reflect the importance of both utilization of insight and data development (for example, Wolfert et al. (2017); Wysel et al. (2021)) by prioritizing the former over the latter.

The reason for this prioritization, and proof for Proposition 4-2, is that the future value of the utilization of insight compounds with every subsequent stage added to the data project. This causes the productive disparity between the two strategies to increase as the number of iterations increases within the data project. Recall that increases in the utilization of insight exhibits positive returns for all payoffs, while increases in the enrichment technology employed faces diminishing returns. As data projects account for all investments and expenses as in-period reductions in usable value, and these expenses effectively overlap profit from the previous stage, the future value of any given stage is reduced by the total expenditure committed for managing data in the *following* stage. Therefore, a production strategy that induces a constant growth rate between inputs and outputs, such as increasing utilization of insight, will create higher future value than a strategy with a diminishing effect, such as increasing enrichment technology.

We can broaden the usefulness of the model to include evaluation of the internal rate of return (IRR). For ease of exposition, let us adapt the previous example. Suppose a firm is considering a data project that involves a two-stage, recursive investment in generating data from existing operations. The firm anticipates insight will attract a payoff in two disparate markets via direct exchange – or sale – of insight, and the application – or use – of insight on existing services. Setting NPV to zero and re-arranging for IRR permits benchmarking this potential data project against capital requirements within the firm.⁵²

$$IRR = \frac{\pi_1 - (x_2 + 2x_1)}{2x_1} + \frac{\sqrt{\pi_1^2 - 2\pi_1x_2 + 4\pi_2x_1 + x_2^2}}{2x_1} \quad (4-4)$$

The appeal of adopting this approach for evaluating data projects is that data-related payoffs, investments, and expenses involved in data management are incorporated into a single calculation. Moreover, Equation (4-3) can be quickly adapted into other models as positive and negative cashflows are reconciled at future values, before being discounted to present values alongside the investment into data for the following stage.

Practically, data managers can use this expression of data-related projects to compare the IRR of prospective data projects against the firm's weighted average cost of capital (WACC) or other internal metric for projects. This approach permits data projects to be evaluated according to their

⁵² While our two-stage example permitted a simple algebraic derivation of the IRR based on the quadratic equation, analysis of data projects with more than two stages could be quickly determined by solving Equation (4-3) for IRR numerically.

relative returns and more broadly, facilitates their ready comparison against other – even non-data – projects within the firm.

4.2. The Assessment of Production Processes that Transform Data into Value

A firm creates value from data through a production process where data is first generated, then enriched into insight, and sold, used, and reinvested into operations. Our model shows how a capitalized cost, which reduces operational value, produces three external payoffs and the accrual of value within the firm. Where the capital cost dominates in-period expenses such as the cost of enriching data, calculation of gross profit offers a straightforward approximation of the value produced by the firm from data-based operations.

Gross Profit. Gross profit sets the costs of generating data against the sum of the data-enabled payoffs. Examples of this scenario include operation of *point-of-sale* systems, corporate software, or enterprise resource planning (ERP) systems where costs are set by the number of transactions or *per endpoint* – such as *per seat*, or *per device* – and not for the extent of analysis the system is required to perform.

In preparation for the calculation of gross profit we start by aligning the payoffs with investments by discounting payoffs to the beginning of each stage using the form, $\delta\pi$. Second stage investments and payoffs can also be aligned with first stage payments with additional discounting of the same form: δx_2 and $\delta(\delta\pi_2)$ respectively. Therefore, the single-stage profit a firm creates through the operation of a data-based, production process as given by Equation (4-2) can be expanded to two stages and written,

$$\begin{aligned}\pi &= \delta(\pi_1 + \delta\pi_2) \\ &= p\delta(y_1 + \delta y_2) + v(1 - \delta)\delta(y_1 + \delta y_2) + \delta(1 - \sigma)(V_1 + \delta V_2).\end{aligned}\quad (4-5)$$

Equation (4-5) says the gross profits a firm receives from data-based operations are the present value of the sum of payoffs received from the sale of insight, the use of insight and the commercialization of improved operations. In primitives, and noting the initial value of operations $V_1 \equiv V$, Equation (4-5) may be expressed,

$$\pi = (p + v(1 - \delta))\delta\left(k(\sigma V)^\alpha + \delta k(\sigma(V + k(\sigma V)^\alpha))^\alpha\right) + (1 - \sigma)\delta(V + \delta(V + k(\sigma V)^\alpha)). \quad (4-6)$$

Proposition 1 says capturing data-enabled learning is paramount for a firm that transforms data into value. While payoffs remain independent the non-rivalrous nature of data permits value created from insight to be captured internally even while the firm creates value from insight externally. Insight improves operations, compounding the effect of a firm's subsequent investments into data.

Better quality data results in further-improved insight which extends the value of each payoff and with it, the firm's data-related profit. Therefore, the more a firm incorporates data-enabled learning into the production process that surrounds data, the greater its resulting profit. Indeed, this finding highlights the strategic importance of data-enabled learning in the ongoing operation of a firm that operates with data. Finally, while the present model is restricted to two stages, it illuminates how a firm could prioritize either short- or long-term returns from their data production process. We leave the extension of this management tension for others to develop.

Net Profit. Where in-period expenses such as the cost of enriching data are also significant, Equation (4-5) may be expanded to both variable, and fixed, process-based costs. Fixed costs F are incurred where a firm pays licensing or access fees for the data, while variable costs occur where raw data requires cleaning, mining or if the ongoing production of insight required the firm to first train and then maintain algorithms. Following Parker and Van Alstyne (2018), we take variable costs as an inefficiency in the production process which suppresses the output of the enrichment process employed by the firm. Practically, variable data management costs increase with increasing production of insight but diminish with improvements in a firm's enrichment technology. Therefore, where enrichment was of the form $Y = k(X)^\alpha$, variable data management costs take the form $Y^{\frac{1}{\alpha}}$. Notice as the firm's enrichment technology improves (increasing α), the firm's variable data management costs diminish. This also accords with recent results from management literature (Agrawal et al., 2018).

Continuing with earlier treatment, expenses - like investments – occur at the beginning of each stage. Expanding Equation (4-5) to include these in-period expenses links a firm's short- and long-term data-related costs with the value the firm has created from data across two stages of investment. Therefore, single-stage, net profit may be written,

$$\pi_{net} = (p + v(1 - \delta))\delta y - cy^{1/\alpha} - F + \delta(1 - \sigma)V. \quad (4-7)$$

Substitution into Equation (4-5) gives the firm's net profit from the production of value from data across two stages,

$$\begin{aligned} \pi_{net} = & (p + v(1 - \delta))\delta y_1 - cy_1^{1/\alpha} - F \\ & + (p + v(1 - \delta))\delta^2 y_2 - c\delta y_2^{1/\alpha} - \delta F \\ & + \delta(1 - \sigma)(V_1 + \delta V_2). \end{aligned} \quad (4-8)$$

Equation (4-8) permits assessment of a firm's payoffs from data against the investments and expenses required to achieve those payoffs. Once characterized, Equation (4-8) permits a firm to balance various management strategies with payoffs from different markets and the value created

from data. For instance, faced with a decreasing payoff for insight, a firm could evaluate its changing composition of data-based value and decide between improving the efficacy of data-enabled learning, or reducing investments in data generation. Likewise, answers to practical questions such as, “should we spend more to purchase better quality data or should it be cleaned inhouse?” can now be answered quantitatively.

4.3. The Optimization of Data-based Operations

The firm must make two decisions: i) what level of investment to make in data – represented as the proportion, σ , of available value to dedicate to the generation of data, and ii) how often to release new insight, represented as stage-length, t . To optimize a firm-level, data-based production process we consider each variable in turn before discussing how the two strategies may be joined.

Optimal Investment Level. From Figure 4-4, short-term value is created from data through the use and exchange of insight, while long-term value is created through the reinvestment of insight across a firm’s operations. This management of data carries both a variable and fixed cost for the firm which is factored against profits. In parallel, value not diverted to the generation or enrichment of data is retained within the firm’s operations and commercialized directly, initially at a base level and subsequently at an increasing level as the firm improves operations through captured insight.

Proposition 4-3: Profit derived from the operation of a data-based production process is well behaved and concave with respect to the origin. There is an optimum investment level, $X^ = X(\sigma^*)$ which maximizes net profit, π_{net} .*

The optimum level of investment is given by,

$$X^* = \alpha \left[\overbrace{W_s} + \overbrace{W_u} + \overbrace{\delta(z_2 - z_1)} + \overbrace{\alpha\beta(x_2 - x_1)} - \overbrace{\frac{c}{\alpha\delta}(y_1^{1/\alpha} + \delta y_2^{1/\alpha})} \right] \quad (4-9)$$

$\overline{\hspace{2cm}}$	$\overline{\hspace{2cm}}$	$\overline{\hspace{2cm}}$	$\overline{\hspace{2cm}}$	$\overline{\hspace{2cm}}$
Investment	Payoffs	Marginal Shadow Value	Marginal Investment Cost	Variable Costs

where β is the second stage, profitability ratio for the data project:

$$\beta = (\delta w_{s2} + \delta w_{u2} - c y_2^{1/\alpha}) / x_2$$

There are two key results:

- i) *The presence of data network effects alone does not justify an increase in investment in data.*
- ii) *Strong market valuation of insight, that is $W_s, W_u \gg V$, can lead to $\sigma^* > 1$, indicating a profit maximising firm ought to seek subsidization of their own data-related operations.*

Proof: See the appendix.

Equation (4-9) states: the firm’s optimum level of investment into data is equal to the sum of the direct and indirect market payoffs from insight, the time-adjusted, marginal payoff from latent data and operations, and the product of the firm’s enrichment technology with the second-stage, profitability ratio of the project and marginal investment in data. These are exclusive of enrichment costs and are all discounted by the firm’s enrichment technology. More approachably, Equation (4-9) may be approximated as in Figure 4-7.

$$\text{Optimal Investment into Data} = \left(\text{Factored Sum of Payoffs} \right) + \left(\text{Factored Expansion Rate of Production Process} \right) - \left(\text{Variable, Data Management Costs} \right)$$

Figure 4-7. Approximated expression for a firm’s optimum investment level into data

Here, factoring includes the enrichment technology employed by the firm, and just as firm-wide operations benefit from data-enabled learning, so these data network effects impact each side of the expression, above. This creates two important considerations for a firm that operates a production process for data.

First, the presence of data network effects alone does not justify an increase in investment in data. Data network effects enable the near-instantaneous and beneficial transfer of value across an organization for capture or appropriation across a firm and its stakeholders – indeed, that is one of their key criteria (Gregory et al., 2021). Because value is created from data by the firm when it realizes a payoff from a data derivative in the market, investment in data principally for the creation of data network effects must also be accompanied by a strategy to realize that potential value. In terms posited by the model, these strategies would need to be predicated on the expectation of a long-term growth in the exchange- or use-value of insight in the market, or the expectation a firm will reduce its investment in data in subsequent stages.⁵³ Outside of these strategies, if a firm’s sole strategic objective was the accumulation of data network effects, the optimum investment in data

⁵³ While these expectations may be well-founded, they represent significant adjustments to the present model, and we leave them for others to pursue.

would be achieved when investment equals the present value of the increase in latent data and operations less enrichment costs.

Second, strong market valuations for insight may lead to an optimum investment that exceeds current optional value. So, investing all available operational value into data management, that is $\sigma = 1$, still amounts to an *under*-investment in data. In such a scenario, a profit maximizing firm could rationally seek external investment to increase the potential value of data generated to maximize the short- and long-term value generated from their data. This result offers a data-centric, theoretical explanation for the disconcertingly (Damodaran, 2014b) high valuations *Uber* received during early funding rounds. In this case, investors believed the firm’s ability to match riders with drivers – their ability to use firm-level insight – would cause a structural change in the market that justified significant expansion in the operating value of the firm (Damodaran, 2014a; Gurley, 2014). This revaluation enabled *Uber* to expand operations (Libert et al., 2014) and, with it, increase data collection and subsequent investment into data (Marr, 2015).

Optimal Investment Rate. We can now address the second question regarding the frequency of investment. We start by considering the optimal discount rate, δ^* , and proceed to the optimal stage length, t^* .

Proposition 4-4a: There is an optimum discount rate, $\delta^ \in (0,1]$ that maximizes net profit, π_{net}^* .*

$$\delta^* = \frac{[w_{s2} + \bar{w}_{u2} + z_2] \mp \sqrt{[w_{s2} + \bar{w}_{u2} + z_2]^2 - 3\bar{w}_{u2} [w_{s1} + \bar{w}_{u1} + z_1 - cy_2^{1/\alpha}]}}{3\bar{w}_{u2}} \quad (4-10)$$

where $\bar{w}_u = vy_i$, is the idealized payoff of insight valued in use.⁵⁴

The main result is that it is never profit maximizing for a firm to permanently cease enriching data.

Where the effective interest rate of insight remains the same irrespective of how value is created from data within the firm, the stage length becomes isomorphic to the discount rate. Therefore, a firm may maximize profit produced from data for any given level of investment by adopting the optimum stage length, or the release schedule for newly generated insight.

⁵⁴ Mathematically, this is equivalent to zero residual value discarded at the completion of each stage, setting $\delta v = 0$.

Proposition 4b: Where the effective interest rate of insight remains constant across the firm, there is a finite and unique stage length, t^ , that maximizes net profit, π_{net}^* .*

$$t^* = \frac{1}{r} \left[\ln \left(\frac{3\bar{w}_{u2}}{w_{s2} + \bar{w}_{u2} + z_2} \right) - \ln \left(1 - \sqrt{1 + 3\bar{w}_{u2} [w_{s1} + \bar{w}_{u1} + z_1 - cy_2^{1/\alpha}] / [w_{s2} + \bar{w}_{u2} + z_2]^2} \right) \right]. \quad (4-11)$$

The key result is that it is not profit maximizing for a firm to continually release insight unless,

$$\bar{W}_u < 2(w_{s2} + z_2) - (w_{s1} + z_1 - cy_2^{1/\alpha}). \quad (4-12)$$

Proof: See the appendix.

Proposition 4-4b says a firm may maximize profit produced from data for any given level of investment by adopting the optimum stage length, or time between release of insight. However, the results of *Proposition 4-4b* address our earlier specific question regarding the conditions when it is profit maximizing for a firm to constantly analyze and release insight. Provided the threshold given in Expression (4-12) is maintained, the profit maximizing firm would constantly release new insight. Where the inequality is maintained, $t^* < 0$ implying a sub-zero optimum stage length. In this scenario the firm would seek to generate and release insight as quickly as possible. Any delay represents lost profit. Conversely, above this threshold the firm would choose to stagger the release of insight into the market at some $t > 0$. Note that in both cases, the firm captures insight internally as data-enabled learning as soon as the insight becomes available and therefore continues to expand the value of its operations while maintaining an optimum data-derived profit.

Joining *Proposition 4-3* with *Proposition 4-4* we can now say, there exists a single, optimum investment strategy for a firm operating a data-based production process. This strategy is a function of capabilities within the firm and market forces external to the firm. This data management strategy maximizes the output of the firm's data-based production process and is achieved when the firm adopts the optimum investment level, σ^* , and the optimum time between release of new insight, t^* .

Practically, once a firm determines the optimum investment rate for its data production process, the firm must also consider the implications of maintaining that optimality even as operating conditions change. From *Proposition 4-3* and *4*, reaction to changes in a market's willingness to pay for insight will extend beyond adjustments to investment rates and also induce changes to the proportion of value created from data directly and indirectly. For instance, where a firm chooses to shorten stage

lengths in pursuit of optimum profit the proportion of revenue from the exchange of insight will increase. Noting the firm's data-enabled payoffs may be interpreted as a portfolio of investments, our theory illustrates that changes in a firm's external environment will induce changes to the composition of a firm's *data portfolio*.

The practical implications of Proposition 4-4 are that a firm's pursuit of optimum conditions will also change the proportions of value created from data directly and indirectly. Payoff from insight traded by the firm is independent of stage length while the payoff from the application of insight grows in logarithmic proportion to stage length. Therefore, where a firm chooses to increase the release rate for insight, that is shorten stage lengths, it will see the proportion of value created from data-as-a-product also increase. This has precedent within literature (Parker & Van Alstyne, 2018) and accords with common experience: a firm responding to a reduction in time to commercialize an intangible good, like a competitive advantage or intellectual property, would rationally choose direct sales rather than time-based partnerships such as licensing agreements.

Our resolution of Proposition 4 is also useful for data management within organizations – even where management decisions occur remotely from market-based payoffs. Extrinsic constraints such as operating budgets or market cycles may constrain optimization of the production process from reaching some part of the solution space described by $\pi(X, t)$. In this case, operating budgets function as a form of prior approval for an acceptable investment rate and therefore also function as a practical connection between level of investment and stage length. Noting a firm's budget or sales cycle may be expressed as a multiple of stage lengths,⁵⁵ an operating budget becomes a threshold that defines a locus of achievable value from data. Therefore, while the notional task might be to determine an optimum investment regime, extant constraints from changing markets, project budgets or operating cycles will impose boundaries on a firm's achievable payoffs from data. We leave the full development of these data management tensions for others to explore.

5. Discussion

In this paper, we present a readily adaptable framework that describes the production process used to create value from data within a firm. The resulting data-based production process explains how firms produce value from data via internal and external payoffs that can operate in the absence of any apparent rivalry. We formally model this production process and express the effect of management decisions on both immediate, and long-term, value created. Our model also permits

⁵⁵ Multiples less than 1 reflect stage lengths that are shorter than budget cycles.

treatment of the fixed and variable costs that may be incurred while managing data. The proposed production process permits assessment of data management strategies and enables firms to finally approach the question, 'how much is my data worth?'. We apply the model to evaluate data projects using conventional project assessment tools such as IRR and NPV. Finally, our model permits assessment of optimal investment levels and rates for a characterized production process and frames the impact of dynamic markets and operating budgets on the pursuit of these optimums. Both scholars and practitioners can now conduct paper-based assessments of firm-level, data-based production processes, investigate the effect of alternative data management strategies, and baseline existing data operations against either industry, or business-model exemplars.

5.1. Data as a Scarce Resource

Our model recognizes the creation of value from data may be incorporated into the process of managing scarce resources. Initially we dealt with scarce financial resources by optimizing the firm's profit function against the proportion of firm value diverted to data management. Data may now be situated among the other 'normal' firm assets of capital and labor and be treated in like manner. This also permits scholars to incorporate existing techniques and build industry-specific tools that help practitioners quickly answer, 'how much should my firm invest into data?'.

Time-based scarcity also impacts commercialization strategies. The high decay rate of 'old data' has important theoretical precedent (Shapiro & Varian, 1998) and arises from the observation that as soon as a firm releases a new stage of insight, the market value for old insight decays sharply. This decay rate implies a significant difference in the value of insight offered as either a product or a service. As a product, insight is valued in exchange and receives a point-price in the market. As a service, the market values insight in use forcing the firm to balance access to immediate returns against the need to improve the service. Longer stage lengths permit the firm to capture greater proportions of the value created from the market's use of insight, but at the cost of larger delays in generating new data from operations.

As our model incorporates time-based changes to the value of data for a variety of different types of payoffs, emerging theories on data valuation may be expanded to consider data-enabled learning models that extend payoffs beyond their initial timeframes. Theoretical data sharing models can also be examined considering private incentives and externalities, and current discussions on mobilization of data in light of value creation-versus-capture can also be oriented.

5.2. Data-Enabled Learning and Data Network Effects

Data-enabled learning is a self-reinforcing cycle (Hagi & Wright, 2020a) where firms learn from operational data (Farboodi et al., 2019) to deliver and rapidly refine their services or products (Prüfer & Schottmüller, 2021). The theory proposed in this paper permits the current data-enabled learning and (conditional) data network effect theories to be oriented such that they may be strengthened theoretically and refined empirically. To support the ensuing discussion, we propose the following three tenets as a summary of data-enabled learning. First, data-enabled learning creates value from data for the simultaneous benefit of both users – that is agents who use the firm’s services or products – and the firm. Second, demand- or supply-side factors can cause the payoffs a firm receives for data-enabled insight to become correlated. Third, the application of insight generated by data occurs rapidly enough that data contributors can also be benefactors of that insight (Gregory et al., 2022) and delays in that application materially affect ongoing contribution.

The first tenet deals with non-rivalrous value capture. In the absence of dynamic market responses and before introducing additional competition or friction, the firm shares surplus with consumers but retains all benefits from reinvested insight. Reinvested insight improves the value of operations which the firm commercializes through w_{r2} , the second-stage payoff from latent data and operations. Meanwhile, users benefit from the exchange and application of insight across both stages. Therefore, the firm’s decision to enrich data enables the non-rivalrous capture of value by the firm and by users. Note that total value realized across the ecosystem is the sum of both parties’ gains and is given by Equation (4-2). Extending the model to this first tenet sets out a decision framework for organizations that are interested in maximizing total productive output from data but are broadly indifferent to the allocation of that output. Community libraries or national laboratories are examples of this arrangement as their charter includes enrichment of data to produce public goods. In this context, a data manager could use the proposed theory to balance the effect of community-wide initiatives, such as adjusted budget cycles, against targeted investments, such as improvements to specific data-enabled learning programs.

The second tenet describes the forces that govern correlation of payoffs from data-enabled learning. Thus far, we have modelled the independent case – that is, a correlation of zero – where the reinvestment of value or realization of a payoff happens in isolation from any other value creation activity. Introducing dependent coefficients to Equation (4-2) permits assembly of data-based production process with correlated payoffs,

$$\pi = aW_s + bW_u + cW_r \tag{4-13}$$

$$= ap\delta k(X)^\alpha + bv(1 - \delta)\delta k(X)^\alpha + cZ$$

where $a, b, c \in \mathbb{R}$. Here, a demand-side correlation could resemble $a = f(v)$ while a supply-side correlation could take the form of $k = f(p, v)$. Markets that exhibit same-sided or cross-market network effects induce a positive correlation in value created from data (Parker et al., 2016) while tensions in value captured from data such as competition between users or the firm (Cusumano et al., 2019) represent a negative correlation amongst value outputs.⁵⁶ An example of the latter is managed by the genetic mapping firm *23andme* as insight sold to scientific collaborators promotes data network effects (Stoeklé et al., 2016) but also diminishes the value of its products to consumers (Hayden, 2017). The proposed theory provides a framework that permits characterization of that correlation beyond demand-side optimization by incorporating the multi-stage contribution of collaborators on the firm's data-enabled learning. For example, in the case of *23andme*, collaborators' multi-stage contribution to data-enabled learning could be set against an anticipated reduction in the exchange price of insight.

The third tenet of data-enabled learning addresses the scenario where the recursive process of enriching data and reinvesting insight occurs over a sufficiently short stage length so that users benefit from one another's data contributions (see for example, Gregory et al. (2022)). These interactions often occur in firms that operate a data-based production process as part of a platform ecosystem (Gawer, 2022). Platforms rely on a positive correlation between insight-based payoffs, as consumption of insight generates more data suitable for generation and re-enrichment by the firm. In platform economic terms: insight creates data network effects through the creation of new data and same- or cross-side increases in demand for insight. LinkedIn's Premium subscription model is one such example. LinkedIn Premium offers users enhanced insight ostensibly in aid of a *freemium* platform business model.⁵⁷ The sale of insight to Premium users amounts to differentiated product that includes enhanced direct messaging and disaggregated profile and behavioral data from other users.⁵⁸ However, this enables Premium users to generate additional data while they use this enhanced product. Increased utility from Premium users drives data network effects within LinkedIn, increasing the current and future value of insight across the platform for all users (Niemand et al., 2015). Same- and cross-sided network effects also drive these data network effects and induce

⁵⁶ We also note *forwards* or *backwards* intergenerational (pan-stage) diffusion of data-enabled value exists (for example, Hann et al. (2016)) and could be accommodated with a more sophisticated adaptation of the model.

⁵⁷ A freemium, platform business model is where levies generated from one part of one side of the market underwrite free services given to other parts of the market such that all benefit more than any user's cost of participation.

⁵⁸ Refer to <https://premium.linkedin.com/> for additional information.

positive correlation between payoffs. The sale of insight is also increased by this market multiplier while the effect of reinvested insight is increased by the increasing scope and number of interactions (Parker et al., 2017). While this heavily correlated application of data-enabled learning warrants a more sophisticated treatment of reinvested insight than the one posited here, we offer initial discussions in the appendix illustrating how the product of the number of interactions of one class of users can act as a scalar for the firm's rate of data-enabled learning.

6. Conclusion

The production of value from data permits firms to simultaneously sell their data, use it to improve other products and services, and incorporate it as data-enabled learning. One decision to invest in data enables multiple, ostensibly non-rivalrous payoffs. The production of value from data enables firms to have their cake, sell it, and eat it too.

This paper connects the two management decisions of 'how much should I invest in data?' and 'how often should I generate new insight?' to the value a firm creates from data. The process that charts the transformation of data from a factor of production to a valuable output within a firm is defined as a data-based production process. This process incorporates the non-rivalry, excludability, and conditional network effects exhibited by data through the presence of parallel payoffs, the need to invest in data before subsumption by the firm, and the potential of near-instantaneous transmission of value from data through the firm. The application of this model permits firms to assess the profit derived from data against fixed and variable inputs, evaluate data projects alongside non-data projects, and to optimize investments into data-based production processes for both pipeline and platform business models.

The proposed model permits the practice of data management to be aligned with broader optimization models such as those that treat data network effects and inform value creation vs. value capture by competing firms or stakeholders. More immediately, firms can now baseline data-based production processes against a generic form. This permits the evaluation of the efficacy, efficiency, and composition of value created from data against the output of traditional production processes that incorporate labor and capital. Indeed, in a world where the labor of bakers and their availability of cakes is sadly rivalrous, our hope is that scholars and practitioners may at least have a clearer understanding of the means by which data enables firms to have their cake, sell it, and eat it too.

References

- Adner, R., & Kapoor, R. (2010). Value Creation in Innovation Ecosystems: How the Structure of Technological Interdependence Affects Firm Performance in New Technology Generations. *Strategic Management Journal*, 31(3), 306-333.
- Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Press.
- Arend, R. J. (2022). We Are Crisis: Runtime Errors in Programmatic Theory. *Academy of Management Review*, 47(2), 331-333.
- Arrow, K. J. (1996). The Economics of Information: An Exposition. *Empirica*, 23(2), 119-128.
- Arthur, W. B. (2009). *The Nature of Technology: What It Is and How It Evolves*. Simon and Schuster.
- Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting Systems, Data, and People: A Multidisciplinary Research Roadmap for Chronic Disease Management. *MIS Quarterly*, 44(1), 185-200.
- Bhargava, H. K., Wang, K., & Zhang, X. (2022). Fending Off Critics of Platform Power with Differential Revenue Sharing: Doing Well by Doing Good? *Management Science*, 68(11), 8249-8260.
- Cennamo, C., Marchesi, C., & Meyer, T. (2020). Two Sides of the Same Coin? Decentralized Versus Proprietary Blockchains and the Performance of Digital Currencies. *Academy of Management Discoveries*, 6(3), 382-405.
- Chesbrough, H. W. (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business Press.
- Chiang, R. H., Grover, V., Liang, T.-P., & Zhang, D. (2018). Strategic Value of Big Data and Business Analytics. *Journal of Management Information Systems*.
- Choudary, S. P. (2014, 07/Feb). The Platform Stack - Understanding Platform Business Models. *The Platform Stack*. <https://artplusmarketing.com/the-platform-stack-c83f9c96e6>
- Clough, D. R., & Wu, A. (2022). Artificial Intelligence, Data-Driven Learning, and the Decentralized Structure of Platform Ecosystems. *Academy of Management Review*(ja).
- Cronin, M. A., Stouten, J., & van Knippenberg, D. (2021). The Theory Crisis in Management Research: Solving the Right Problem. *Academy of Management Review*, 46(4), 667-683.
- Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2019). *The Business of Platforms: Strategy in the Age of Digital Competition, Innovation, and Power*. Harper Business New York.
- Damodaran, A. (2014a, 08/Feb). A Disruptive Cab Ride to Riches: The Uber Payoff. *aswathdamodaran*. <http://aswathdamodaran.blogspot.com/2014/06/a-disruptive-cab-ride-to-riches-uber.html>
- Damodaran, A. (2014b). *Uber Isn't Worth \$17 Billion | Fivethirtyeight*. FiveThirtyEight. Retrieved 08/Feb from <https://fivethirtyeight.com/features/uber-isnt-worth-17-billion/>
- Easley, D., Huang, S., Yang, L., & Zhong, Z. (2018). The Economics of Data. *Available at SSRN* 3252870.
- Evans, D. S., & Schmalensee, R. (2016). *Matchmakers: The New Economics of Multisided Platforms*. Harvard Business Review Press.
- Farboodi, M., Mihet, R., Philippon, T., & Veldkamp, L. (2019). Big Data and Firm Dynamics. AEA papers and proceedings,
- Fleckenstein, M., Obaidi, A., & Tryfona, N. (2023). A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model.
- Frankel, A., & Kamenica, E. (2019). Quantifying Information and Uncertainty. *American Economic Review*, 109(10), 3650-3680.
- Gawer, A. (2022). Digital Platforms and Ecosystems: Remarks on the Dominant Organizational Forms of the Digital Age. *Innovation*, 24(1), 110-124.
- Golman, R., Loewenstein, G., Molnar, A., & Saccardo, S. (2022). The Demand for, and Avoidance of, Information. *Management Science*, 68(9), 6454-6476.

- Gomes, L. A. d. V., Facin, A. L. F., Salerno, M. S., & Ikenami, R. K. (2018). Unpacking the Innovation Ecosystem Construct: Evolution, Gaps and Trends. *Technological Forecasting and Social Change*, 136, 30-48. <https://doi.org/10.1016/j.techfore.2016.11.009>
- Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The Role of Artificial Intelligence and Data Network Effects for Creating User Value. *Academy of Management Review*, 46(3), 534-551.
- Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2022). Data Network Effects: Key Conditions, Shared Data, and the Data Value Duality. In *Academy of Management Review*.
- Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, 35(2), 388-423. <https://doi.org/10.1080/07421222.2018.1451951>
- Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating Big Data: A Literature Review on Realizing Value from Big Data. *The Journal of Strategic Information Systems*.
- Gurley, B. (2014, 08/Feb). How to Miss by a Mile: An Alternative Look at Uber's Potential Market Size. *Above the Crowd*. <http://abovethecrowd.com/2014/07/11/how-to-miss-by-a-mile-an-alternative-look-at-ubers-potential-market-size/>
- Hagiu, A., & Wright, J. (2020a). Data-Enabled Learning, Network Effects and Competitive Advantage. In *Unpublished*.
- Hagiu, A., & Wright, J. (2020b). When Data Creates Competitive Advantage. *Harvard Business Review*, 98(1), 94-101.
- Hann, I.-H., Koh, B., & Niculescu, M. F. (2016). The Double-Edged Sword of Backward Compatibility: The Adoption of Multigenerational Platforms in the Presence of Intergenerational Services. *Information Systems Research*, 27(1), 112-130.
- Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing Value from Big Data—a Taxonomy of Data-Driven Business Models Used by Start-up Firms. *International Journal of Operations & Production Management*, 36(10), 1382-1406. <https://www.emeraldinsight.com/doi/pdfplus/10.1108/IJOPM-02-2014-0098>
- Hayden, E. C. (2017). The Rise and Fall and Rise Again of 23andme. *Nature*, 550(7675), 174-177.
- Hinz, O., Otter, T., & Skiera, B. (2020). Estimating Network Effects in Two-Sided Markets. *Journal of Management Information Systems*, 37(1), 12-38. <https://doi.org/10.1080/07421222.2019.1705509>
- Horling, B., & Kulick, M. (2009). Personalized Search for Everyone. *The Official Google Blog*.
- Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9), 2819-2858.
- Kaiser, C., Stocker, A., Viscusi, G., Fellmann, M., & Richter, A. (2021). Conceptualizing Value Creation in Data-Driven Services: The Case of Vehicle Data. *International Journal of Information Management*, 59, 102335.
- Koohang, A., Sargent, C. S., Nord, J. H., & Paliszkievicz, J. (2022). Internet of Things (IoT): From Awareness to Continued Use. *International Journal of Information Management*, 62, 102442.
- Lakatos, I. (1968). Criticism and the Methodology of Scientific Research Programs. *Proceedings of the Aristotelian society*,
- Langley, P., & Leyshon, A. (2017). Platform Capitalism: The Intermediation and Capitalisation of Digital Economic Circulation. *Finance and Society*, 3(1), 11-31.
- Libert, B., Wind, Y., & Fenley, M. (2014). What Airbnb, Uber, and Alibaba Have in Common. *Harvard Business Review*, 11(1), 1-9.
- Liu, Y., Soroka, A., Han, L., Jian, J., & Tang, M. (2020). Cloud-Based Big Data Analytics for Customer Insight-Driven Design Innovation in SMEs. *International Journal of Information Management*, 51, 102034.
- Loveman, G. (2003). Diamonds in the Data Mine. *Harvard Business Review*, 81(5), 109-113.

- Marr, B. (2015). The Amazing Ways Uber Is Using Big Data Analytics. *LinkedIn*. <https://www.linkedin.com/pulse/amazing-ways-uber-using-big-data-analytics-bernardmarr> (Accessed: May. 18, 2021).
- Morozov, E. (2016). *Tech Titans Are Busy Privatising Our Data*. The Guardian. Retrieved 24/Jan from <https://www.theguardian.com/commentisfree/2016/apr/24/the-new-feudalism-silicon-valley-overlords-advertising-necessary-evil>
- Müller, O., Fay, M., & Vom Brocke, J. (2018). The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. *Journal of Management Information Systems*, 35(2), 488-509.
- Niemand, T., Tischer, S., Fritzsche, T., & Kraus, S. (2015). The Freemium Effect: Why Consumers Perceive More Value with Free Than with Premium Offers.
- O'Reilly, T. (2017, 29/Jan). Phishing for Phools. *Words That Matter 2017*. <https://medium.com/wordsthatmatter/phishing-for-phools-ad31b127cfa6>
- Parker, G., & Van Alstyne, M. (2018). Innovation, Openness, and Platform Control. *Management Science*, 64(7), 3015-3032.
- Parker, G., Van Alstyne, M., & Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. WW Norton & Company.
- Parker, G., Van Alstyne, M., & Jiang, X. (2017). Platform Ecosystems: How Developers Invert the Firm. *MIS Quarterly*, 41(1).
- Pentland, A., Lipton, A., & Hardjono, T. (2021). *Building the New Economy: Data as Capital*. MIT Press.
- Prüfer, J., & Schottmüller, C. (2021). Competing with Big Data. *The Journal of Industrial Economics*, 69(4), 967-1008.
- Rahmati, P., Tafti, A. R., Westland, J. C., & Hidalgo, C. (2020). When All Products Are Digital: Complexity and Intangible Value in the Ecosystem of Digitizing Firms. *Forthcoming, MIS Quarterly*.
- Schatz, D., & Bashroush, R. (2016). Economic Valuation for Information Security Investment: A Systematic Literature Review. *Information Systems Frontiers*, 19(5), 1205-1228. <https://doi.org/10.1007/s10796-016-9648-8>
- Shapiro, C., & Varian, H. R. (1998). *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Press.
- Sharapov, D., & MacAulay, S. C. (2022). Design as an Isolating Mechanism for Capturing Value from Innovation: From Cloaks and Traps to Sabotage. *Academy of Management Review*, 47(1), 139-161.
- Short, J., & Todd, S. (2017). What's Your Data Worth? *MIT Sloan Management Review*, 58(3), 17.
- Sims, C. A. (2003). Implications of Rational Inattention. *Journal of Monetary Economics*, 50(3), 665-690.
- Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social Media Analytics—Challenges in Topic Discovery, Data Collection, and Data Preparation. *International Journal of Information Management*, 39, 156-168.
- Stoeklé, H.-C., Mamzer-Bruneel, M.-F., Vogt, G., & Hervé, C. (2016). 23andme: A New Two-Sided Data-Banking Market Model. *BMC medical ethics*, 17(1), 1-11.
- Tambe, P., Hitt, L., Rock, D., & Brynjolfsson, E. (2020). *Digital Capital and Superstar Firms*.
- Tirole, J. (1988). *The Theory of Industrial Organization*. MIT press.
- Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On Value and Value Co-Creation: A Service Systems and Service Logic Perspective. *European Management Journal*, 26(3), 145-152.
- Varian, H. (2018). *Artificial Intelligence, Economics, and Industrial Organization*.
- Wagner, D. G., & Berger, J. (1985). Do Sociological Theories Grow? *American Journal of Sociology*, 90(4), 697-728.

- Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big Data in Smart Farming—a Review. *Agricultural Systems*, 153, 69-80.
- Wysel, M. (2023). Data Sharing Platforms in Agriculture. In *Encyclopedia of Smart Agriculture Technologies*. Springer. https://doi.org/10.1007/978-3-030-89123-7_250-1
- Wysel, M., Baker, D., & Billingsley, W. (2021). Data Sharing Platforms: How Value Is Created from Agricultural Data. *Agricultural Systems*, 193, 103241.
- Zellweger, T. M., & Zenger, T. R. (2021). Entrepreneurs as Scientists: A Pragmatist Approach to Producing Value out of Uncertainty. *Academy of Management Review*(ja).

Appendix

This paper presents a theory for a generic production process used by firms to transform data into value. It presents a formal model of that process which is used to assess the efficacy of the production process, evaluate data projects within a firm, and optimize a firm's data management decisions. Our goal was to map the production process clearly so any novel insights might remain as assessable – and above all, useful – to as broad a range of readers as possible. Following is a brief redevelopment of the model proposed in the main paper to support extension or adaption by the reader.

Modeling the Data-based Production Process within a Firm

Recall the model charts the production process over two, sequential stages of each length. The firm faces two decisions: what proportion of value is to be invested, $\sigma \in [0,1]$, and the release schedule for insight, or stage-length, $t \in (0, \infty)$. The firm makes a single decision regarding each variable so neither is changed once the model commences. Profit, π , is the sum of market payoffs and is realized at the end of each stage. All expenses occur at the start of each stage. To align value flows, profit is discounted to present values using a standard discount function, δ , which can be formally connected to stage-length and the effective interest rate of insight, $r \in (0,1)$, by noting $\delta = e^{-rt}$. The second stage may be aligned with the first stage using the same method. All other variables are summarized in Table 4-3, above.

Results for Data Projects, Data-based Production Processes, and Optimal Investment into Data

Internal Rate of Return. Equation (4-6) summarizes the output from the creation of value from data within a firm. Noting that all investments that are subsequent to the firm's initial investment in data are taken as in-period expenses we can start with the generic expression for NPV:

$$NPV = \sum_{i=1}^T \frac{(Net\ Cash\ Flow)_i}{(1 + IRR)^i}$$

and substitute in the schedule of payments depicted in Figure 4-6 to give,

$$NPV = -x_1 + \frac{w_{s1} + w_{u1} + w_{r1} - x_2}{(1 + IRR)} + \frac{w_{s2} + w_{u2} + w_{r2}}{(1 + IRR)^2}.$$

As noted above, investments from the next stage overlap with returns from the current stage, with only x_1 sitting without overlapping returns. Rearranging above, gives Equation (4-3).

The IRR that balances this 2-stage project may be determined by setting $NPV = 0$ and raising the previous equation throughout by $(1 + IRR)^2$ to give,

$$0 = -x_1(1 + IRR)^2 + (w_{s1} + w_{u1} + w_{r1} - x_2)(1 + IRR) + (w_{s2} + w_{u2} + w_{r2}).$$

In the specific case where the firm operates over two stages, this equation may be solved using the quadratic formula, as in Equation (4-4). However, in general solving Equation (4-3) numerically will be easier and certainly faster.

Gross Profit. Recalling the necessity to discount future value flows to present values, Equation (4-5) may be developed,

$$\begin{aligned} \pi &= \delta(\pi_1 + \delta\pi_2) \\ &= \delta(w_{s1} + \delta w_{s2}) + \delta(w_{u1} + \delta w_{u2}) + \delta(w_{r1} + \delta w_{r2}) \\ &= p\delta(y_1 + \delta y_2) + v(1 - \delta)\delta(y_1 + \delta y_2) + \delta(1 - \sigma)(V_1 + \delta V_2), \end{aligned} \tag{A4-1}$$

where w_s , w_u , and w_r are the payoffs the firm receives from the exchange of insight, the use of insight and data retained in operations, respectively. This may be read as the profit a firm receives from data is the present value of the sum of the payoff received from the sale of insight, the payoff from the use of insight and the payoff from commercialization of operations that house latent data. Noting the initial value of operations $V_1 \equiv V$, this can also be expressed in primitives as in Equation (4-6).

Net Profit. We can expand our treatment of profit to include both fixed, F , and variable costs, $cy^{1/\alpha}$. As previously noted, we take variable costs of enriching data to be in the form $Y^{1/\alpha}$ to reflect that increases in enrichment technology (that is, increasing α) cause increases in both efficacy and efficiency of data development. Therefore, single-stage, net profit can be written,

$$\begin{aligned} \pi &= p\delta y + v(1 - \delta)\delta y - cy^{1/\alpha} - F + \delta(1 - \sigma)V \\ &= (p + v(1 - \delta))\delta y - cy^{1/\alpha} - F + \delta(1 - \sigma)V. \end{aligned}$$

Expanding Equation (4-5) to include these in-period expenses links a firm's short- and long-term data-related costs with the payoffs a firm receives from data. The full net profit from the production of value from data across two stages within a firm can therefore be written,

$$\begin{aligned}\pi_{NET} &= (p + v(1 - \delta))\delta y_1 - c y_1^{1/\alpha} - F \\ &+ (p + v(1 - \delta))\delta^2 y_2 - c \delta y_2^{1/\alpha} - \delta F \\ &+ \delta(1 - \sigma)(V_1 + \delta V_2).\end{aligned}$$

Optimizing Investment Level. For any given stage length, the optimum investment into data, $X^* \triangleq X(\sigma^*)$ is obtained when $\frac{\partial \pi_{net}}{\partial \sigma} = 0$.

To ease the following manipulation, Equation (4-7) may be written as,

$$\pi_{net} = W_s + W_u + W_r - VC - F$$

which can be differentiated once with respect to σ and set to 0. Initially,

$$\frac{\partial \pi_{net}}{\partial \sigma} = \frac{\partial W_s}{\partial \sigma} + \frac{\partial W_u}{\partial \sigma} + \frac{\partial W_r}{\partial \sigma} - \frac{\partial VC}{\partial \sigma} - \frac{\partial F}{\partial \sigma} = 0$$

which once the requisite components within Equation (4-7) are substituted expands to,

$$\begin{aligned}0 &= \frac{\partial}{\partial \sigma} (p\delta(y_1 + \delta y_2)) + \frac{\partial}{\partial \sigma} (v\delta(1 - \delta)(y_1 + \delta y_2)) + \frac{\partial}{\partial \sigma} (\delta(1 - \sigma)(V + \delta(V + y_1))) \\ &\quad - \frac{\partial}{\partial \sigma} (c(y_1^{1/\alpha} + \delta y_2^{1/\alpha})) - \frac{\partial}{\partial \sigma} (F(1 + \delta)).\end{aligned}$$

While the differentiating process is laborious it is nonetheless procedural. Multiplying throughout by σ greatly simplifies the subsequent expression. Grouping like terms yields,

$$\begin{aligned}0 &= \left(\alpha W_s + \delta p \alpha^2 \frac{y_2}{x_2} \sigma y_1 \right) + \left(\alpha W_u + \delta v(1 - \delta) \alpha^2 \frac{y_2}{x_2} \sigma y_1 \right) \\ &\quad + (-\sigma V - \delta \sigma(V + y_1) + \delta(1 - \sigma) \alpha y_1) \\ &\quad - \left(c y_1^{1/\alpha} + \delta (c y_2^{1/\alpha} + \alpha c k^{1/\alpha} \sigma y_1) \right).\end{aligned}$$

Noting, $\sigma V + \delta \sigma(V + y_1)$, correspond to the amount invested into data by the firm across two stages and can be written as X , we can separate those terms and take them to the LHS. With X on the left, and collecting α on the right we can write,

$$X = \alpha \left[W_s + W_u + \delta(z_2 - z_1) + \alpha \left(\delta w_{s2} + \delta w_{u2} - c y_2^{1/\alpha} \right) (x_2 - x_1) / x_2 - \frac{c}{\alpha \delta} \left(y_1^{1/\alpha} + \delta y_2^{1/\alpha} \right) \right].$$

Finally, noting that $(\delta w_{s2} + \delta w_{u2} - c y_2^{1/\alpha}) / x_2$ represents the second stage profitability ratio of the process, we can define that as β and further simplify to Equation (4-9). The results for Proposition 4-3 follow quickly. If the firm invests in data solely for data network effects, then operational value is diverted, $\sigma > 0$, but $W_s = W_u = 0$ implying either $k = 0$ – which is unlikely given the firm’s strategy – or $p = v = 0$. In this case, the output of the production process is entirely reinvested. This reduces Equation (4-8) to a function of the marginal increase in payoff from latent data and operations, and enrichment costs (if any). Naturally, the firm may also be operating beyond the scope of the model and expecting an eventual growth in p or v , or to reduce σ in subsequent stages.

It is possible – and in some markets even probable – that $\sigma \in [0,1]$ will not balance Equation (4-9). In this scenario, $W_s, W_u \gg V$, can lead to $\sigma^* > 1$, indicating a profit maximising firm ought to seek subsidization of their own data-related operations. This has mathematical precedent within the management literature (Parker & Van Alstyne, 2018) and accords with common experience as discussed in the body of the paper.

Optimizing Investment Rate. Equation (4-8) can be rearranged into powers of δ and differentiated with respect to δ . As the goal is to find the value of δ that maximizes π , the first-order differential is set to zero.

$$\frac{\partial \pi_{net}}{\partial \delta} = \frac{\partial}{\partial \delta} \left\{ \left[-c y_1^{1/\alpha} - F \right] + \left[(p + v) y_1 + (1 - \sigma) V - c y_2^{1/\alpha} - F \right] \delta + \left[(p + v) y_2 + (1 - \sigma)(V + y_1) \right] \delta^2 + \left[-v y_2 \right] \delta^3 \right\} = 0.$$

Completing the derivative and simplifying yields,

$$\left[w_{s1} + \bar{w}_{u1} + z_1 - c y_2^{1/\alpha} \right] + 2\delta \left[w_{s2} + \bar{w}_{u2} + z_2 \right] - 3\delta^2 \left[\bar{w}_{u2} \right] = 0$$

where $\bar{w}_u = v y_i$, is the idealized payoff of insight valued in use. This can be solved numerically or with the quadratic equation. Choosing the latter gives Equation (4-10).

Proposition 4-4 claims the ideal time between investments is *finite*, that is, it is never profit maximizing for a firm to cease all generation and enrichment of data. That requires $\delta^* > 0$. Noting the above first-order equation is a continuous, negative quadratic and will therefore have two

solutions separated by a single maximum, we can show at least one solution is positive by proving $0 < \delta_{maxima}^* < \delta^*$.

δ_{maxima}^* may be determined by setting $\frac{\partial^2 \pi_{net}}{\partial \delta^2} = 0$ and solving for δ . Differentiating once more w.r.t δ yields the same result.

$$\delta_{maxima}^* = \frac{(p + v)y_2 + (1 - \sigma)(V + y_1)}{3vy_2}.$$

Therefore, if $V, v, k > 0$ and $\sigma \in (0,1)$ then $\delta_{maxima}^* > 0$ and therefore $\delta^* > 0$. This establishes that a profit maximizing firm would not choose $t = \infty$.

The opposite edge case occurs at $\delta^* = 1$ and corresponds to $t^* = 0$, indicating the continuous generation and enrichment of data. Note that $\delta^* > 1$ indicates $t^* < 0$ which a firm would implement by adopting $t = 0$ as the closest possible solution. To determine the conditions at this edge case, we substitute $t = 0$ into the first-order equation to find the solutions.

Incorporating the inequality makes the boundary conditions easier to track. Therefore, letting $\delta^* \geq 1$ and multiplying throughout gives,

$$2w_{s2} - \bar{w}_{u2} + 2z_2 < w_{s1} + \bar{w}_{u1} + z_1 - cy_2^{1/\alpha}.$$

Collecting terms, and noting that at the edge case, $\delta = 1$ and therefore $\bar{W}_u = \bar{w}_{u1} + \bar{w}_{u2}$ we can write,

$$\bar{W}_u < 2(w_{s2} + z_2) - (w_{s1} + z_1 - cy_2^{1/\alpha}). \quad (A4-2)$$

which states that the firm ought to adopt an episodic release schedule for insight unless the inequality is satisfied.

Finally, recalling that $\delta = e^{-rt}$, and therefore $t = -\frac{1}{r} \ln(\delta)$ the optimum time between investments in data, t^* , may be derived by rearranging Equation (4-10) for t ,

$$t^* = \frac{1}{r} \left[\ln \left(\frac{3\bar{w}_{u2}}{w_{s2} + \bar{w}_{u2} + z_2} \right) - \ln \left(1 - \sqrt{1 + 3\bar{w}_{u2} [w_{s1} + \bar{w}_{u1} + z_1 - cy_2^{1/\alpha}] / [w_{s2} + \bar{w}_{u2} + z_2]^2} \right) \right]. \quad (A4-3)$$

Use of the Data-based Production Process to Design Empirical tests for Data-Enabled Learning and Data Network Effects

The primary goal of the paper is to posit the general production process that describes the transformation of data into value. Accordingly, the paper develops and discusses the case where value production remains uncorralled; that is, where the realization of payoffs and the reinvestment of value happen in isolation from any other value creation activity undertaken by the firm. However, the proposed theory may be easily adapted to support both positive and negative correlation in value creation, such as a corporate environment where a decision to sell insight (say) reduces the value of using the insight.

Introducing dependent coefficients into Equation (4-2) permits assembly of data-based production process with correlated payoffs as in Equation (4-13). As noted above, a demand-side correlation could resemble $a = f(v)$ while a supply-side correlation could take the form of $k = f(p, v)$. For the specific example of *23andme* discussed in the body of the paper, the proposed theory would permit creation of a 'learning threshold' where organizational learning would be first characterized, such as, $V = V_{old} + \gamma Y$ where $\gamma \in \mathbb{R}$, and set against anticipated reductions in the market price of insight, p , so that the proposed data strategy might satisfy the firm's IRR.

Higher Degree Research Thesis by Publication

University of New England

Statement of Authors' Contribution

We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated in the *Statement of Originality*.

	Author's Name	Percentage of Contribution
Candidate	Matthew Wysel	90
Other Authors	Derek Baker	10



Candidate

Date



Principal Supervisor

Date

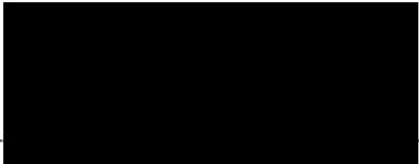
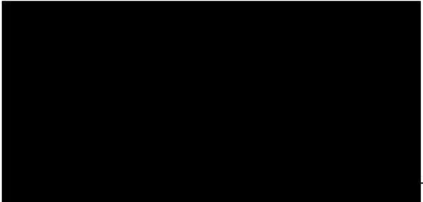
Higher Degree Research Thesis by Publication

University of New England

Statement of Originality

We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that the following text, figures, and diagrams are the candidate's original work.

Author	Type of Work	Page Numbers
Matthew Wysel	Conceptualization, methodology, formal analysis, writing (original draft), writing (review), editing.	Entire Document
Derek Baker	Conceptualization, writing (review), editing, supervision.	Entire Document

	_____
	<i>Date</i>
	_____
Principal Supervisor	Date

Chapter Five:

Take my data... please. How Data Sharing Ecosystems Make Oversharing Rational.

Overview of Manuscript

Status:	Reviewed by <i>Management Information Systems Quarterly (MISQ)</i>
Date Submitted:	April 2023
Date Published:	-
Suggested Citation:	Wysel, M., Baker, D., & Billingsley, W. (2023). Take my data... please. How Data Sharing Ecosystems Make Oversharing Rational. Manuscript submitted for publication.

Summary of Paper in Context of PhD:

This chapter takes the results from a national survey of livestock breeders conducted by the Australian Meat and Livestock Association (MLA) (Banks, 2019) and analyzes the data trading ecosystem that they create by sharing genetic data with industry analysts such as the Agricultural Business Research Institute (ABRI). We connect the exchange value of data with the non-rivalry and conditional network effects exhibited by data to explain how agents, such as breeders, can create value by giving data away. This broadens the data-based, production process developed in Chapter 4 and uses it to evaluate how efficiently both the breeders and the industry analyst create value from data.

In essence, this paper adds choice variables to each of the three tasks in the management layer of Figure 2-3 (Chapter 2) and the potential for non-unitary transformations at the junction of each box in Figure 4-4 (Chapter 4).⁵⁹ Agents in this ecosystem are represented as either *data-analysts* or *data-producers*, that is the *System*, or stakeholders in the *Community* from Chapters 2 and 3, respectively. Notice, both agents operate data-based production processes according to the model developed in

⁵⁹ Fortunately, as this paper develops, the desire each agent in the data sharing ecosystem possess to maximize either short- or longer-term value greatly simplifies the objective function each maintains.

Chapter 4, except that the data-producer outsources the enrichment function to the data-analyst. When read in context, this chapter presents a ‘system-of-systems’ application of the models developed in Chapters 2 through 4. This chapter illustrates how agents in a marketplace, each creating value from data (Chapter 4) by operating their own data sharing platform (Chapter 2), assemble as an otherwise disparate community around congruous data-related goals (Chapter 3) to form data sharing *ecosystems*.

A key tenet of this paper is that if analysts in a data sharing ecosystem adopt a *service-dominant* logic towards data trading, they can incentivize producers to share data above their notional optimum, expanding the Pareto frontier of the ecosystem and inducing a (rational) overshare from producers. While this chapter presents our analysis of a specific genetic trading market, adoption of a service-dominant view of data trading also reveals breed societies are facing a pending technology-enabled, market failure in the supply of genetic data. We return to analyze the growing tragedy of their (data) commons in Chapter 6.

Apart from typesetting changes and language localization, this chapter appears exactly as submitted to *Management Information Systems Quarterly (MISQ)* in April 2023.

Supplementary Publications

Research that informed this chapter also appeared in the following outputs.

Type	Citation
Conference Paper	Wysel, M. (2021, 22 June 2021). Data Sharing Platforms in Agribusiness. International Food and Agribusiness Management Association, Costa Rica.
Conference Paper	Wysel, M., & Baker, D. (2021, December 2021). Sandwiches Vs. Genes. Sharing Data to Maximize Its Value. MODSIM2021, 24th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, Sydney.
Conference Paper	Wysel, M., & Baker, D. (2022, 8-11 February 2022). Anonymous Tolls and the Value of Insight. 66th Annual Conference of the Australasian Agricultural and Resource Economics Society, Armidale, Australia.
Book Chapter	Fomiatti, B., & Wysel, M. (2022). Developing the Value of Smart Agriculture through Digital Twins. In <i>Encyclopedia of Smart Agriculture Technologies</i> . Springer. https://doi.org/10.1007/978-3-030-89123-7_273-1

Chapter-level Glossary

Term	Definition
Data-analyst	<i>Noun.</i> An agent whose primary, data-related task is the enrichment of data into more valuable insight. Insight is used to power the data-analyst's technologies, shared with the data-producer and be sold on the open market.
Data-producer	<i>Noun.</i> An agent whose primary, data-related task is the generation and sharing of data. Data is shared with the data-analyst in exchange for insight. Insight drives operational improvement which provide both short- and longer-term returns.
Goods-dominant logic	<i>Noun.</i> An understanding of transactions where the value created in exchange amounts to an instantaneous and – importantly – discrete transmission of value. See Vargo et al. (2008). Contrasts with a service-dominant logic.
Service-dominant logic	<i>Noun.</i> An understanding of transactions which sees value transmission as initiated by the exchange, but not exclusively contained within it. See Vargo et al. (2008). In the context of trading data, the application of each agent's skills to data currently under their control creates value for other agents in the ecosystem, even though all agents operate independently. Contrasts with a goods-dominant logic.

Full Manuscript

Take my data... please. How Data Sharing Ecosystems Make Oversharing Rational.

Authors:	Matthew Wysel ^a (matthew.wysel@une.edu.au;)*, Derek Baker ^a (derek.baker@une.edu.au), William Billingsley ^b (wbillin@une.edu.au)
Affiliations:	^a The Centre for Agribusiness, UNE Business School, The University of New England, Armidale, Australia. ^b Computational Science, The University of New England, Armidale, Australia.
Keywords	Exchange value, value of data, data platforms, economics of data, data network effects
Statement of Significance ⁶⁰	“In a world awash with data, our paper provides a straightforward framework for navigating the complexity of sharing proprietary and personal information. Empowering both data-producers and data-analysts to realize the potential of shared data, our model illuminates the fine line between rationally sharing and irrationally oversharing. Whether you share data for short-term gain or long-term growth we show you how much – and how often – to share data. Grounded in the very practical world of Australian cattle breeders, our research offers clear instructions to help you make the most out of the data you’re already giving away.”
Acknowledgements:	Matthew Wysel is grateful for the support of the Agricultural Business Research Institute through the Arthur Rickards Innovation in Agribusiness Scholarship. All errors remain our own.

⁶⁰ Per the *March 2023 Editor’s Comments*, MISQ requires authors to submit a short (<90 words) expression of their paper’s significance beyond academia.

Abstract

This paper relates exchange value with non-rivalry and conditional data network effects. Widespread treatment of data like an ordinary, private good creates problems for participants in data trading markets as they give data away to create and capture value. However, the attributes of data call for a different approach that resembles service-dominant logic rather than goods-dominant logic. Without this, both agents and the surrounding ecosystem needlessly lose value.

This paper examines an existing data sharing ecosystem that trades genetic data and comprises a data producer, data analyst and an open market. We formally model the effect that supply level and supply frequency have on value created from shared data by agents on both sides of the exchange as they pursue either short- or long-term growth strategies.

We demonstrate that a service-dominant view of data enables an expanded Pareto frontier as agents invert the value creation process by sharing value to create data. Additional applications of the model include comments on interactions with 'generative AI', vertical integration in data supply chains, reverse subsidies, and quantitative data governance thresholds.

1. Introduction

Generating and sharing otherwise private data is an intrinsic part of the operation of any modern, competitive firm. Businesses of all sizes install third-party apps on core systems, engage management consultants, or incorporate smart devices within their operations in an effort to transform raw data into valuable insight. Having received shared data, these partnering analysts return the exchange. Apps return insight for sales data, dashboards deliver understanding to users, and consultants share recommendations with clients. For the data-producer, data shared externally seeds insight beyond that which is available internally. This insight improves operations through data-enabled learning (Hagiu & Wright, 2020a) that enable efficiencies and support competitive advantage. For the data-analyst, enriched data is sold to producers as insight, used to refine the technologies that produce it (Gregory et al., 2021), or is bundled into collaborative (Bardhan et al., 2020) or competitive products (Angelopoulos et al., 2021). In this manner, one agent's decision to share data facilitates an ecosystem of data traders, each creating value from exchanged data, as data circulates between them and outwards to the market.

Data sharing permeates personal lives too. Houses, cars, webpages, and social connections create ecosystems of value (Hukal et al., 2020) as data is generated and shared with external systems (Cichy et al., 2021) and increasingly become dependent on insight returned from these systems (Kaiser et al., 2021). These ecosystems of exchanged data form new methods of value co-creation as data is created and utilized for individual and collective value creation (Constantinides et al., 2018).

Personal devices such as *wearables* share biologic data for short-term outcomes such as entertainment or convenience, and longer-term benefits like improved health outcomes (Bardhan et al., 2020) while the facilitating system uses aggregated data to provide insight on a range of health conditions (McGuire et al., 2011). In each case, sustained data sharing enrolls agents – whether firms

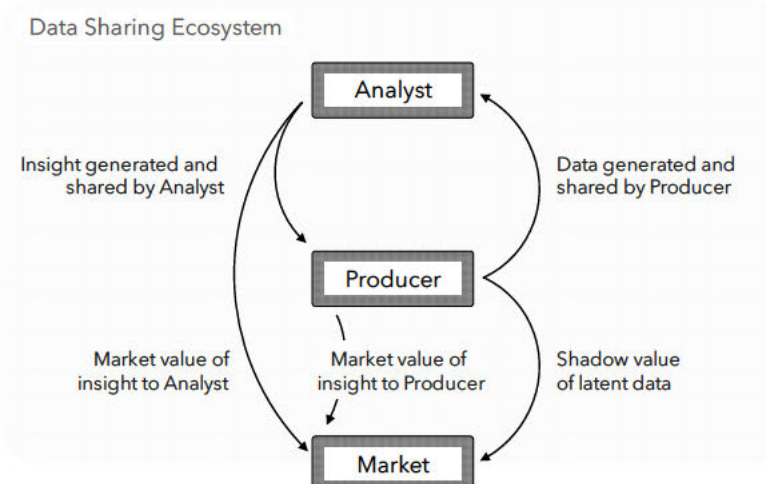


Figure 5-1. Simplified Data Sharing Ecosystem

or individuals – as data-producers and data-analysts into a system of value co-creation (Wysel et al., 2021) that circulates data within – and outside of – a data sharing ecosystem. We summarize this data-based, value circulation in Figure 5-1.

However, while data is excludable like a private good, unlike a private good it is also non-rivalrous (Jones & Tonetti, 2020) and displays conditional *data network effects* (Clough & Wu, 2022; Gregory et al., 2021). These attributes change the ‘normal’ rules of exchange-based, value creation and capture (Pentland et al., 2021). First, the non-rivalry of data means that use by one agent does not impede use by any other agent (Jones & Tonetti, 2020). For instance, an analyst can repackage and sell insight even after returning it to a client with neither agent’s use inducing congestion around the shared data. However, this lack of congestion also means shared data, once disclosed, cannot be easily recalled (Shin et al., 2022). This shifts the locus of exchange-based, value creation away from individuals and into the surrounding ecosystem (Cichy et al., 2021) as the externalities produced by the exchange circulate within the ecosystem conferring indirect value back to agents (Gregory et al., 2021). For instance, sharing a popular story on social media propagates engagement across users whose response produces additional data which the analytic system uses to expand the engaged userbase, who create yet more data in an expanding cycle of data supply and value creation. These externalities extend beyond the boundaries of the original exchange as agents appropriate value created from the use of other’s shared data. Resembling a service-dominant logic (Vargo & Lusch, 2008), this process enables mutual exposure to arm’s length value creation, fueling both privacy concerns (Saura et al., 2021) and opportunities for value co-creation (Hukal et al., 2020). This arrangement places agents in the middle of a co-opetitive relationship (Nalebuff & Brandenburger, 1997) where the value each creates from data is connected to the data-based decisions of the other.

This uncertainty in data-based value capture (Clough & Wu, 2022) is reflected in data markets (Wagner et al., 2021) whose nominal function ought to be the support of these value-creating exchanges. On the one hand, data markets enable new value streams that turn everyday activities into potentially profitable enterprises (Koohang et al., 2022). On the other hand, data markets exhibit significant information asymmetry as agents either unwittingly share data (Shin et al., 2022) or share data not realizing the value their data unlocks for others (Cichy et al., 2021). This creates markets that are characterized by data arbitrage (Angelopoulos et al., 2021), data vacuums (Holder, 2019), and questionable data equity (Jagadish et al., 2022) as data enables the co-creation of value but confusion regarding the mechanics of value creation inhibits its capture (Wagner et al., 2021). Agents on both sides of the ecosystem struggle to understand the benefits and risks of their data sharing practices (Windasari et al., 2021), forge enduring programs of co-innovation (Saura et al., 2021), or build reputations as trusted partners (Kotlarsky et al., 2023). Without a clear

understanding of how markets support the co-creation of value from data, uncertainty regarding the enduring quality of data goods will inhibit individual exchanges, incentivize inefficient behavior in agents, and ultimately suppress the size of data markets (Akerlof, 1978).

This paper explains how the sharing of data – as governed by its attributes of non-rivalry and conditional data network effects – enables agents to create value, even though that exchange amounts to giving valuable data away. Our aim is to add a formal understanding of when – and how – an agent should share data to maximize value created individually, and to describe the conditions when agents on opposite sides of the exchange might support each other’s ‘overshare’ to unlock greater levels of value for both parties. Our analysis responds to recent calls to promote awareness of value created by other’s data (Shin et al., 2022), inform data managers what actions they may take to build resilient data-communities (Tremblay et al., 2023) and to enable “users to control the level of access to their data” (Angelopoulos et al., 2021, p. 1) in order to promote fairness across data-driven ecosystems (Wagner et al., 2021).

This paper asks the question: if an agent shares data to create value, how much – and how often – should they share? We apply this question to both sides of the data trading market, that is, data-producers such as ‘normal’ firms and individuals, and data-analysts such as knowledge workers and data service providers. We consider the creation of value from data towards both short- and longer-term goals. We take profit – with its focus on market returns – as a proxy for short-term value creation, and capital growth – or growth in the intrinsic value of the underlying asset base – as an indication of longer-term value creation. For simplicity of explanation, we assume the data-producer generates raw data independently from the analyst. However, the analysis proceeds unchanged if the analyst provides the services that enable generation of raw data, or if the analyst also ‘hosts’ the entire data sharing ecosystem. Examples of the first variation include enterprise resource planning systems and other SaaS packages⁶¹. Examples of the second variation include most social media platforms. In these cases, data-producers are typically referred to as *clients* or *users*, respectively. We include discussion of the practical limitations clients and users might experience while sharing data to create value when discussing the results of the model.

We set the analysis up thus: suppose a data-producing agent desires insight regarding assets it owns but is unable to process the data itself. The producer generates and shares data with an independent analyst who enriches all data, returning it as insight. Meanwhile, the analyst also stores the data and repackages it for its own profit. The producer desires insight but must disclose the data

⁶¹ SaaS: software-as-a-service, refers to the arrangement where the vendor (analyst) provides the infrastructure, platform, and software to the client (producer), typically for an all-encompassing, recurring fee.

to achieve it. The analyst wants to maximize value captured from the data without jeopardizing the partnership with the producer that enables value creation (Schrieck et al., 2021). Therefore, all agents have the challenge to maximize value created by sharing data while minimizing the associated cost.

To ground this analysis, this paper presents a case study of an existing data trading ecosystem that involves the exchange of genetic data taken from animals. In this ecosystem the producer operates a business that selectively breeds animals, such as cattle, and elects to share genetic data – such as a sample of hair – with an independent, genetic laboratory. Insight from this laboratory enables more profitable blood lines for the producer and valuable industry reports for the analyst. While this example was chosen for its concreteness and ease of illustration rather than widespread significance, aspects of this ecosystem may prove instructive beyond their immediate use.

Systematic germline editing⁶² in humans remains highly contentious (Bergman, 2019; Farrelly, 2005) in part because while there is little doubt surrounding its commercial value, there remains significant social (Jasanoff & Hurlbut, 2018), ethical (Evans, 2021), and business model (Stoeklé et al., 2016) issues to address. This paper aims to address incentives and value allocation, and while retaining a focus on explicating the creation of value from shared data, also hopes to offer some guidance to those looking to architect markets that trade data as portentous as human genetics.

We formally model the effect that an agent's frequency of participation and level of data shared has on value captured by the agent and created across the ecosystem. The model is iterative, permitting the effects of data-enabled learning to be reflected in agents' participation decisions as they share, incorporate, and re-share data. The input decisions of the data-producing agent are optimized against first a short- and then a long-term value creation strategy.⁶³ As we subsequently develop, the decisions of the data-analyst are simpler. The analyst processes all data supplied and optimizes their decisions against immediate profit. We initially derive conditions for each agent to maximize value created when acting individually, before investigating how the non-rivalry and conditional data network effects exhibited by data enable agents to go beyond co-creation of valuable data and into cross-market collaboration around value. We illustrate how the attributes of data enable value created through its recombinant exchange to exceed value created if data were only a private good.

⁶² *Germline editing* is genetic editing that affects all cells in an organism, including gametes, so that the modified genes are passed on to future generations. This contrasts with *somatic gene editing* that affects only some of the cells in the organism being treated and is not passed on to subsequent generations.

⁶³ For a firm, these strategies are equivalent to short-term profit and long-term capital growth. For an individual, these strategies are analogous to convenience and personal development.

This confirms findings in existing datanomic literature (Easley et al., 2018; Gregory et al., 2021; Hagiu & Wright, 2020b; Jones & Tonetti, 2019).

The paper is organized as follows. Following a review of the datanomic and service-dominant literature, we assemble a model of a simplified data sharing ecosystem consisting of one data-producing agent, one data-analyst, and a market that clears goods produced by both agents. The results of applying the model to the genetic data trading ecosystem are presented alongside specification of management decisions that would enable an optimum data management strategy at the individual or ecosystem level. The discussion includes commentary on inverting the value creation process (where value is shared to create data), the training and casual use of generative AI systems, when to subsidize data exchanges, and the conditions for vertical integration in data supply chains.

2. Review of Literature

This paper builds on the literature that addresses data as a shared factor of production (Pentland et al., 2021), the self-reinforcing nature of data-based externalities (Gregory et al., 2021), and the persistence of access to value that agents have when collaborating around data (Windasari et al., 2021) even when that collaboration takes the form of *co-opetition* (Nalebuff & Brandenburger, 1997).

2.1. Data as a Shared Production Factor

Insight⁶⁴ is data that has been refined towards a goal, or *desideratum* (Frankel & Kamenica, 2019), as defined by a surrounding community of stakeholders (Wysel et al., 2021). Stakeholders are free economic agents who recursively interact with technological systems (Varian, 2014) to adjust and tailor data to their needs. These activities can take the form of generation, synthesis, analysis, or presentation of data (Hartmann et al., 2016) as they enrich the data into more valuable insight. This process includes the reduction of uncertainty (Frankel & Kamenica, 2019) and extraction of noise in a manner that resembles the restoration of a message in signal theory.

This collaborative value co-creation process may proceed without the normal allocation of value that accompanies other intangible goods such as intellectual property (Parker et al., 2017) because while data is an excludable, intangible good (Easley et al., 2018) it displays non-rivalry like a pure, public

⁶⁴ We use the term *insight* rather than *information* throughout to avoid confusion caused by the conflation of the two terms in popular media.

good (Fleming et al., 2018). Like a club good (Wysel et al., 2021) exclusion from data also excludes agents from the ability to create value from it. However, once data is shared, control over who has access to that data is also shared (Shin et al., 2022). This permits the data to move easily beyond the boundary of the firm (Arrow, 1996) or ecosystem. Meanwhile, those with access to the data collaborate (Bardhan et al., 2020) or compete (Wagner et al., 2021) in the value creation process. Therefore, participation in the process of creating value from data connects agents in a series of symbiotic partnerships (Kotlarsky et al., 2023) centered on data-based value creation as each creates, derives and provides benefit in an ecosystem of shared externalities (Parker et al., 2016). Together, these sociotechnical arrangements come to resemble a three-fold ecosystem of agents, shared data, and computation whose collective function is to produce value from continuously circulating data (Wysel et al., 2021).

2.2. Conditional Data Network Effects

Generalizing the definition proposed by Gregory et al. (2021), we take data network effects as the value creation that occurs across a data sharing ecosystem when one agent experiences a growth in value in proportion to the insight produced by another agent. Central to the creation of data network effects is that agents on both sides of the exchange act as real-time, complementors for the other's creation and capture of value (Clough & Wu, 2022) while data and more valuable insight pass between both parties. The practical implication of the presence of data network effects within a data sharing ecosystem is that the more insight the ecosystem amasses on, and for, the agents the more value the agents can capture from the data – conditional on the distribution of power across the ecosystem (Cusumano et al., 2019). The mechanics of data network effects are that data generated and shared by an agent, fuels a response by the ecosystem that encourages the generation of more data by the agent within a timespan that materially affects the agent's ongoing participation. While this relationship can be used to extract value from agents (Holder, 2019), the coercive use of data network effects is not necessary (Gregory et al., 2022) and can even produce suboptimum levels of data for the ecosystem.

Agents remain free to make decisions that maximize the value each captures (Mullins & Sabherwal, 2022) up to and including not sharing data at all. However, an agent who chooses not to share data chooses to maintain their data-based operations at a baseline rate which we take as a commodity level throughout (Parker et al., 2017). Producers choose the terms of their participation by deciding on the amount of data they wish to share and the frequency of their participation. In line with Shapiro and Varian (1998), the generation of new data effectively bundles legacy data into baseline operations or technologies. This sets the shape and governing terms of each agent's utility curve

within the ecosystem. Agents continue to participate as long as their marginal utility of membership in the ecosystem remains positive.

2.3. Data-based Co-creation, Collaboration and Co-opetition

The co-creation of value that surrounds data relies on the application and maintenance of competencies by agents as they create value for themselves and, through externalities, change the value for others in the ecosystem. This arrangement is known as a *service system* (Vargo et al., 2008) and is typified by enduring productive relationships where the recombinant exchange of a good initiates the co-creation of value in use as resources are integrated and applied towards a locus of goals. Agents in data sharing ecosystems collaborate around non-rivalrous data-goods, reducing the intrinsic uncertainty contained with the data as they create and capture value both directly from the data and indirectly through externalities delivered by the ecosystem (Wysel & Baker, 2021).

However, while agents collaborate to develop the data-good, they also compete for share of the value that proceeds from their ecosystem. In this way agents are in co-opetition (Nalebuff & Brandenburger, 1997) as each agent's data and skills contribute to the other's competitive advantage even as they build their own. Therefore, while agents may compete on value capture, they must co-operate around data management (Gregory et al., 2022).

Disentangling the allocation of value between agents based on share of contribution or value-of-interactions (Johnson et al., 2005) becomes predictably difficult (Bresciani et al., 2021; Rahmati et al., 2020; Windasari et al., 2021) with recent literature often focusing on competitive (Cichy et al., 2021; Ogbanufe, 2023) rather than integrative and co-creation outcomes (Acharya et al., 2022; Ciriello, 2021).

Therefore, while the data-producer and data-analyst participate in the same data sharing ecosystem, their individual data-based production processes are necessarily twinned but not necessarily competitive. The excludability of data requires agents to share, the non-rivalry of data enables collaboration without congestion, while conditional data network effects bind the externalities produced by one agent's data-based decisions initially to the other agent, and through that agent, back to the first. These relationships determine the operation of data sharing ecosystems and therefore the tradeoffs each agent makes when deciding on the nature of their participation in these ecosystems. Applying these relationships to the ecosystem illustrated in Figure 5-1, we assemble a model that describes the operation of a genetic data trading ecosystem.

3. The Model: Sharing Data to Create Value

The exemplar used to illustrate the model consists of both commercial and non-commercial livestock breeders, who use genetic data to improve the quality of their herd through selective breeding. Producers⁶⁵ operate commercial breeding programs designed to create genetic data from animals they own. These animals serve as ‘genetic hosts’ whose function is to create genetic material that exhibits specific performance metrics for the market (Banks, 2019). The market purchases this genetic material – either as gametes or ‘in’ the host animal – typically for incorporation into large-scale production processes that service consumer markets. To maintain competitive advantage, the producer shares data on their current operations with an industry-based analyst who returns genetic reports that inform the producer which animals to breed and which to cull. As in our motivating discussion, the producer can elect not to share data with the analyst, but this would also stop the supply of genetic insight regarding which bloodlines to maintain. Both the cost and benefit associated with sharing data are a function of the proportion of data shared by the producer. Finally, the analyst repackages and sells data shared by the producer into a parallel market, keeping all proceeds. This arrangement accords with the data sharing ecosystem as previously described and is analogous to a firm engaging management consultants or operating a network of IoT devices, a user watching an ad-supported video streaming service or sharing exercise data with a fitness app.

3.1. Trading Genetic Data to Create Value

Our model assumes the producer has data-related operations with a value V , and chooses to share data with the analyst at a cost, X , that scales in proportion to the amount of data shared, such that $X = \sigma V$. This cost reduces the value of operations to a residual level, Z , where $Z = V - X$. The analyst takes X as an input to their production process and uses technology that accords with a Cobb–Douglas production system (Müller et al., 2018) to produce an output $Y = kX^\alpha$. As in other industrial optimization models (Parker & Van Alstyne, 2018; Rochet & Tirole, 2003), k represents the efficacy of the partnership, and $\alpha \in (0,1)$ represents the diminishing returns of further technological investment in the data enrichment systems. Following Easley et al. (2018) insight is shared without loss between both parties: $Y = Y_P = Y_A$. The producer applies the insight received from the analyst to improve the quality of their genetic material amounting to a direct payoff, W_P , from use of the analyst’s insight. The insight also creates indirect improvements to operations through data-enabled

⁶⁵ Readers familiar with genetic production systems will recognize the producer in this paper as a *Breeder* and that the market comprises both *Commercial Producers* and other *Breeders*. While these subcategories carry important distinctions, they were amalgamated chiefly avoid confusion for a non-specialist audience.

learning that increases the value of the producer's operations to a new level, $V_{new} = V + Y$. Meanwhile, the analyst who receives genetic data from multiple producers, repackages and sells insight in the market for payoff, W_A . This flow of values is illustrated in Figure 5-2.

As $V = Z + X$, if producers choose to share nothing then $X = 0$ and $Z = V$. When framed in this manner, Z may be considered the *shadow value* of the unshared, or latent, data. Although livestock breeders commercialize this value, we acknowledge that in other scenarios users might not commercialize the residual value of their data, such as with driving data from vehicles (Cichy et al., 2021), health data from smart watches, or human-centered, genetic testing (Hayden, 2017).⁶⁶ Nonetheless, we maintain the face value of Z throughout the model to illustrate the importance of the shadow value of latent data to producers who share data to maximize profit.

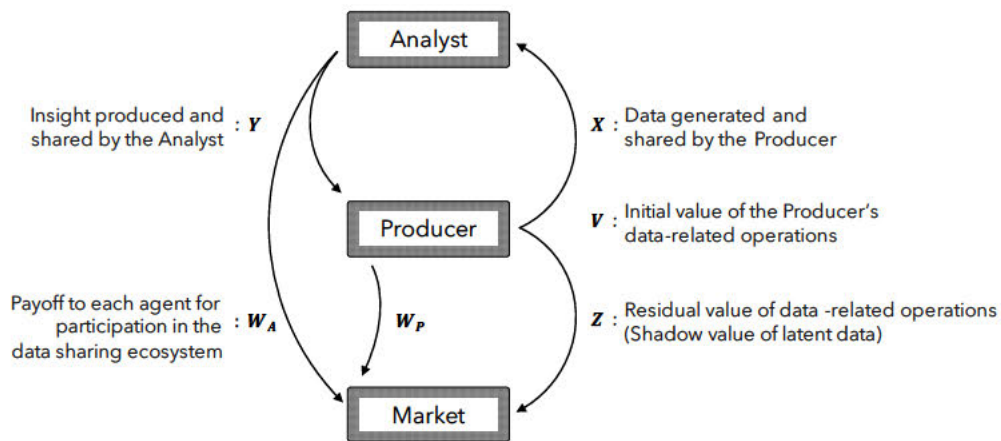


Figure 5-2. Value-flows within the data sharing ecosystem

Producers make two decisions regarding their participation in the data sharing ecosystem:

- i) the proportion of data shared, which we represent as $\sigma = X/V, \in [0, \infty)$, and
- ii) how often they participate in the ecosystem, represented as the length of time between successive sharing events, $t \in [0, \infty)$.

These decisions can be represented as a two-dimensional choice, $\langle \sigma, t \rangle$, which, for the purposes of this model serves as a persistent agreement between the producer and the ecosystem. As we subsequently develop, the analyst is motivated to process all data shared by the producer and so, always chooses to analyze and re-share all data as often as possible.

We refer to the length of time between each sharing event as a *stage*. The commencement of each stage is marked by the generation and sharing of data by the producer. To permit amalgamation of

⁶⁶ Where the residual value of data is not commercialized Z is effectively a deadweight loss.

costs and revenues the stage length can be represented as a discount factor, δ , in the usual form $\delta = e^{-rt}$ where t is the length of each stage and r is determined by the type of data shared and represents the effective interest rate of the corresponding insight produced by the analyst. As developed in the Results, producers in this data trading ecosystem must work within a constrained operating space for both variables. Note the upper bound for the proportion of data shared is not $\sigma = 1$. A producer might reasonably choose $\sigma > 1$ where they see additional benefit in subsidizing the analyst such as volunteering the supply of broader metadata. Here we assume subsidisations remain economically equivalent to value shared beneath $\sigma = 1$.⁶⁷

Agents always use the most up-to-date data available. Once either agent generates and shares data, the other agent's operations stand to benefit from the improved and newly shared data, implying agents would always choose to use the newest data into their operations. This suggests value created from ongoing participation in data sharing ecosystems is recursive rather than iterative as each stage builds on the output of the previous stage. We run the model across two stages to capture the effect of this recursive accrual of data across the ecosystem.

While the payoff each agent receives from the market is of the form, $W = pY$, the manner in which each agent captures their payoff is different. The producer uses insight to achieve an increase in genetic performance, v , which delivers value throughout the stage, while the analyst sells their benchmarking report at a price, p . Customers who purchase the producer's improved genetic material know they can wait until the following stage when new material will be generated, and the present products are depreciated. This implies that consumers are not willing to pay more than the difference between their maximum willingness to pay, v , and the present value of the data from the next stage, δv . Thus, the market's expectation of the ongoing provision of new genetic material at the start of the next stage restricts the price the producer can achieve to $v - p \geq \delta v$ which sets the maximum price for the producer's payoff to $p_p = v_p(1 - \delta)$. If the producer chooses to share data continuously, obsolescence happens immediately and they may only extract the commodity price $p = v(1 - 1) = 0$. Conversely, if the producer chooses not to generate new data, then the stage does not end, neither data nor insight expires, and the producer may charge the monopoly price $p = v(1 - 0) = v$. It follows that longer stages permit the producer to extract greater value from the use of improved genetic performance but also reduce the present value of revenue from subsequent stages.

⁶⁷ If subsidisations carried different returns, then a secondary production function could be implemented for $\sigma > 1$.

Figure 5-3 represents the two-stage expression of this model of an ecosystem of agents who share data with one another. This model tracks the relative improvements in the value of data within the data sharing ecosystem as agents share, and re-share one another's data. We now progress to evaluating the value created by each agent as they release variously improved data-products into the market.

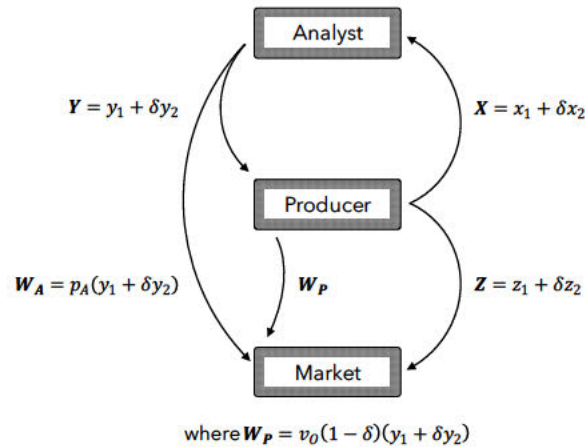


Figure 5-3. Creating value from shared data

3.2. Evaluating Success: Sharing Data for Short-term Profit or Long-term Capital Growth?

The producer must decide how to evaluate their two-dimensional decision. Specifically, should they pursue a short-term investment strategy that prioritizes market returns, a longer-term strategy that prioritizes growth in the residual value of the assets, or a combination of both strategies?⁶⁸ This arrangement is illustrated in Figure 5-4a and accords with Parker et al. (2017)'s treatment of intangible goods in a microsystem of platforms sponsors and developers.

The producer would choose a longer-term, growth-focused investment strategy if they primarily shared data for a change in the baseline value of their data-producing assets. This focus is depicted in Figure 5-4b. In this scenario, the producer is concerned with the net value created by the analyst's insight, that is, $Y - X$. Unlike the short-term strategy, the proportion of data shared by the producer is not influenced by the current market conditions but by the opportunity to increase the value of the data-producing assets in light of some non-market index.

⁶⁸ We discuss a combination of both strategies in the Results after presenting each boundary case.

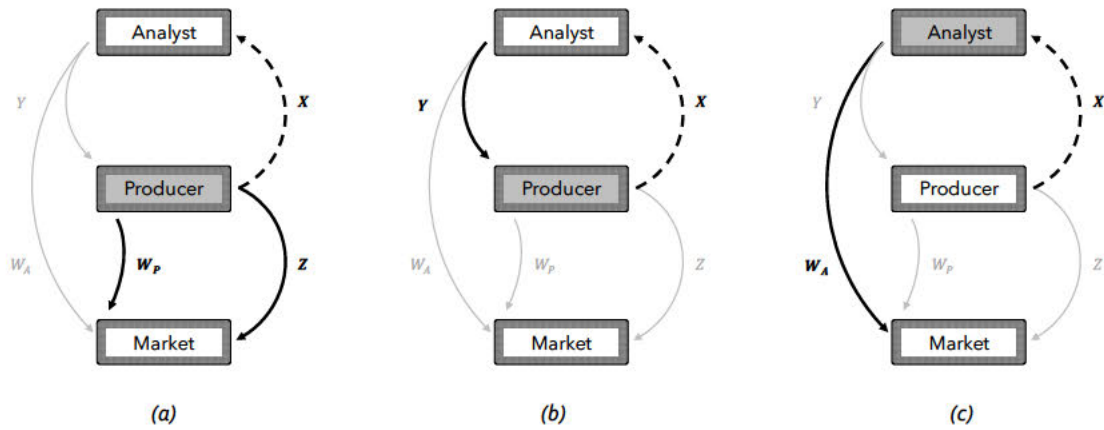


Figure 5-4. (a) Short-term vs. (b) Longer-term investment strategies for the producer. (c) Short-term investment strategy for the analyst

Finally, Figure 5-4c gives the value-flow for the analyst. Notice that while the analyst benefits from non-market improvements in their shared data, their only means of collecting a payoff comes from the direct commercialization of insight, W_A . Similarly, while the efficacy of their output increases the value of future insight via the producer's operations, these improvements must also be cleared by the market (in subsequent stages) for the analyst to realize their value. As noted by Clough and Wu (2022), the immediate implication is that sustained success for the analyst is driven by their ability to capture market, rather than non-market, returns.

3.3. Mobilizing the Model

Where the producer shares data to create short-term value, their profit is the sum of the residual value of data-based operations, Z , and the payoff from genetic improvements, W_P , across both stages of the model,

$$\begin{aligned} \pi_P &= Z + W_P \\ &= z_1 + \delta z_2 + v_P(1 - \delta)(y_1 + \delta y_2) \end{aligned} \quad (5-1)$$

Where the producer shares data to create longer-term growth, their analysis would center on the difference between the value ceded, X , and the yield returned, Y , across both stages. The net present value of the capital growth, ΔK , of the data shared by the producer across two stages is,⁶⁹

$$\begin{aligned} \Delta K_P &= Y - X \\ &= (y_1 - x_1) + \delta(y_2 - x_2). \end{aligned} \quad (5-2)$$

⁶⁹ Capital growth, ΔK , is the change in capital from an initial value of V to final value K . Therefore, the value of the producer's capital after two stages is $K = V + \Delta K$.

The producer's two value creation functions can also be expressed in primitives,

$$\pi_p = (1 - \sigma)(V + \delta(V + k(\sigma V)^\alpha)) + v_p k(1 - \delta)((\sigma V)^\alpha + \delta \sigma^\alpha (V + k(\sigma V)^\alpha)^\alpha) \quad (5-3)$$

$$\Delta K_p = (k(\sigma V)^\alpha - \sigma V) + \delta(k\sigma^\alpha (V + k(\sigma V)^\alpha)^\alpha - \sigma(V + k(\sigma V)^\alpha)) \quad (5-4)$$

Similarly, the analyst's profit can also be expressed as the sum of payoffs from both stages,

$$\pi_A = W_A = p_A(y_1 + \delta y_2) \quad (5-5)$$

and in primitives,

$$\pi_A = p_A((\sigma V)^\alpha + \delta \sigma^\alpha (V + k(\sigma V)^\alpha)^\alpha). \quad (5-6)$$

Equations (5-1), (5-2) and (5-5) give us the value created by the producer and analyst under different investment strategies as a function of insight yielded by the analyst. Note Figure 5-3 and Figure 5-4 may also be combined to produce these equations.

We can now progress to examining the effect of the producer's decisions regarding supply level and supply frequency on value created by both sides of the data sharing ecosystem.

4. Model Analysis: Sharing Data to Maximize Value

The optimal contract for either agent is the arrangement $\langle \sigma, t \rangle$ that maximizes either profit or capital growth for the producer, and profit for the analyst. As Equations (5-1) – (5-6) are also functions of δ , and where r is constant δ is isomorphic t , we will use the decision pair $\langle \sigma, \delta \rangle$ to simplify the assessment of the producer's decisions.

4.1. Maximizing Profit for the Data-Producer

Where the producer shares data to pursue a profit maximization strategy they will seek π_p^* and optimize σ and δ to that end. Intuitively, we might expect the limits to this pursuit to occur when the analyst's yield drops beneath the value ceded by the producer, that is where $y_i \not\geq x_i$, or where the payoff received from the market no longer exceeds the amount given away, that is $W_p \not\geq X$. While both conditions have important management implications, as we shall see, π_p^* may require σ and δ beyond both limits.

Proposition 5-1 *There exists a unique pair of $\langle \sigma^*, \delta^* \rangle$ that maximizes the data-producer's profits, π_p^* . Provided $t < \infty$, the data-producer's profits are concave with respect to increases in δ . The data-producer's profits are also concave with respect to increases in σ , although attainment of σ^* may*

prove impractical where the market value for the application of insight exceeds the value of latent data.⁷⁰

The discount rate that optimizes the data-producer's profit for any given amount shared is,

$$\delta^* = \begin{cases} \frac{(\bar{w}_{P2} - \bar{w}_{P1}) + z_2}{2\bar{w}_{P2}} & \text{when } \frac{y_2}{y_1} \geq \frac{\bar{w}_{P1} - z_2}{\bar{w}_{P1}} \\ 0 & \text{otherwise.} \end{cases} \quad (5-7)$$

where \bar{w}_{P_i} is the monopoly payoff for the data-producer and $\frac{y_2}{y_1}$ is the yield ratio of the data-analyst.

The proportion of shared data that optimizes the profits of the data-producer, σ^* , is achieved when,

$$X = \alpha \left[W_P + \delta(z_2 - z_1) + \delta(x_2 - x_1) \frac{\alpha}{\beta_2} \right] \quad (5-8)$$

where $\beta_2 = \frac{x_2}{w_{P,2}}$ and is the variable cost ratio for the data-producer in the second stage.

Proposition 5-1 gives two key results:

- i) There exists a unique σ that maximizes π_p^* for any $\delta \in [0,1]$
- ii) In the absence of external constraints, a data-producer can freely select their supply frequency while $\sigma > 0$ and $\frac{y_2}{y_1} > \frac{\bar{w}_{P1} - z_2}{\bar{w}_{P1}}$.
- iii) Ideal stage length is inversely proportional to the amount shared in the range $\sigma \in (0, \infty)$.

Proof. See the Appendix. ■

Equation (5-8) may be understood as the producer maximizes profit from shared data when the value of data given away equals the payoff received from the market, plus the marginal shadow value of unshared data and the marginal ceded value, all discounted by the analyst's technology. The marginal components are discounted at the appropriate rate, while the final component is further discounted by an *ecosystem factor*: the ratio of the return on investment the producer receives for sharing data.

Proposition 5-1 explains several practical phenomena. Notice the left-hand side of Equation (5-8) increases as a function of value introduced to the ecosystem by the producer: σ and V , while the right-hand side increases in properties offered by the ecosystem: α , k and v_p (contained within W_p). This confirms intuition around participation within data sharing ecosystems as the more effective the technology employed by the analyst, the more data the producer can give away before reaching satiation. Likewise, the more effective the partnership between the producer and the analyst (given

⁷⁰Where $\bar{W}_p \gg Z$, σ^* may only be reached after assumptions like the analyst operating a Cobb-Douglas production system, or a single price point from the market have broken down.

by k), or the stronger the market value for enriched data, the greater the value of data the producer can give away before reaching satiation.

Where producers have full control over both their supply level and supply frequency, a profit-maximizing producer would seek a supply level and frequency that balanced Equation (5-8), above.⁷¹ Notice the limit in Result (ii) comprises the ratio of the analyst's yields on the left and the producer's capture of value on the right. As $\alpha < 1$, the analyst's technology produces diminishing returns as the value of data shared by the producer increases. Therefore, in the absence of structural changes in the ecosystem, the marginal yield, $\frac{y_2}{y_1}$, diminishes towards unity for large values of X . While $\sigma < 1$, z_2 is positive and Result (ii) tells us the producer may continue to select their optimum supply frequency irrespective of the analyst's capabilities. However, if $\sigma > 1$ then z_2 becomes negative implying the producer must also consider the analyst's yield ratio as well as their own allocation of value in determining the optimum frequency to share data. Practically, if the analyst's yield ratio was not projected to satisfy Result (ii), the producer would immediately cease sharing any further data with the ecosystem, effectively setting $t = \infty$. This change would set $\delta^* = 0$ and reduce Equation (5-8) to $x_1 = \alpha \bar{w}_{p,1}$.⁷²

In ecosystems that value enriched data significantly greater than raw data, such that a large v_P or p_A induces $W \gg Z$, market returns inflate σ^* inferring a profit-maximizing producer ought to progress beyond 'sharing everything', even to the point of heavy subsidization of the analyst. This implication has a consequence for both the producer and analyst. First, where large market valuations sustain profit growth beyond $\sigma = 1$ additional profits from sharing data come at direct expense to the producer. Equation (2-3) illustrates this relationship clearly. Recalling $Z = V - X$, where $\sigma > 1$, $X > V$ resulting in a negative value for Z . The producer has begun subsidizing their own data sharing efforts to increase their profit. While $1 < \sigma < \sigma^*$ the producer is still participating beneath their optimum, so while this subsidy is not an 'overshare' it does illustrate the importance of additional management policies – or personal boundaries – once $X > V$. This underlines the importance of strong data stewardship and judicial monitoring of privacy policies and terms of use offered by data sharing ecosystems – especially where the data-producer is also required to pay fiat currency in addition to ceding all data – lest data-producers 'prop up' inefficient technologies or support exploitive partnerships (Shin et al., 2022).

⁷¹ σ^* may be determined directly by substitution of δ^* from Equation (5-7) into Equation (5-8) and solving numerically for σ^* .

⁷² This corner solution is also observed by Parker and Van Alstyne (2018).

4.2. Maximizing Capital Growth for the Data-Producer

Alternatively, the producer may share data for long-term growth. While the analyst continues to trade with the marketplace, the producer's decision-making framework has narrowed to a straightforward exchange with the analyst. The analyst seeks π_A^* while the producer seeks ΔK_P^* and will optimize σ and δ to that end.

Proposition 5-2 *The data-producer's capital growth, ΔK_P , from shared data is concave with respect to the proportion of data shared, σ , while $\Delta K_P(\sigma)$ is linearly proportional to the discount rate, δ , adopted by the data-producer.*

The proportion of data shared that maximizes capital growth, $\sigma_{\Delta K}^$, is achieved when,*

$$\left(\alpha \frac{y_1}{x_1} - 1\right)V + \delta \left(\alpha \frac{y_2}{x_2} - 1\right)(V + y_1(1 + \alpha)) = 0, \quad (5-9)$$

while the discount rate that optimizes capital growth for the data-producer, $\delta_{\Delta K}^$, is given by the conditions,*

$$\delta_{\Delta K}^* = \begin{cases} 1 & \text{when } x_2 < \gamma \\ 0 & \text{when } x_2 > \gamma \end{cases} \quad (5-10)$$

and anywhere in its normal range at the point $x_2 = \gamma$. Note $\gamma = k^{1/1-\alpha}$ and represents the 2-stage productive limit of a partnership with the data-analyst.

Proposition 2 gives three key results:

- i) There exists a unique σ that maximises ΔK_P for any $\delta \in [0,1]$,*
- ii) Where the data-producer's capital growth is positive for both stages, the data-producer always increases ΔK_P by shortening stage length or choosing to share data that lowers the effective interest rate of insight,*
- iii) A data-producer never maximizes capital growth by participating in a partnership that does not support positive capital growth for two stages.*

Proof. See the Appendix. ■

While Proposition 5-2 may be mobilized to develop data sharing policies in a similar manner to Proposition 5-1, the practical implications for a producer who shares data for longer-term value creation are noteworthy.

First, in the ideal case where the producer is only bound by the technology of the analyst and the efficacy of the partnership, capital growth of their data producing assets is maximized when σ balances Equation (5-9) over two stages. From Result (iii), a producer seeking maximum capital growth ought to participate in an ecosystem that supports $y_i(\sigma) > x_i(\sigma)$ for both stages. From Equation (5-10), participation across two stages sets $\delta_{\Delta K}^* = 1$, and therefore, $\langle \sigma_{\Delta K}^*, \delta_{\Delta K}^* \rangle$ reduces to $\langle \sigma_{\Delta K}^*, 1 \rangle$, confirming an analogous solution from Parker and Van Alstyne (2018). The corollary is the

producer now maximizes capital growth by sharing data as frequently as possible and choose a type of data whose insight decays immediately.

The 2-stage productive limit of a data sharing ecosystem offers an important threshold for data-producers who desire longer-term value creation but must accept conditions that restrict their supply level or frequency of data. Examples of these scenarios include clients of SaaS packages – with integrated usage monitoring – or users of health tracking apps. If the restrictions on the supply of data, represented as $\hat{\sigma}$ is greater than $\sigma_{\Delta K}^*$ – implying the producer must share more than their optimum – and $x_2 < \gamma$ the best outcome for the producer is to follow intuition: adopt the minimum supply level possible, $\hat{\sigma}$, minimize the stage length and attempt to share only data whose insight decays quickly. However, if $\hat{\sigma}$ results in $x_2 > \gamma$, the producer ought to maximize the length of each stage – even to the point of setting $t \rightarrow \infty$ – and share data whose insight decays as slowly as possible. That is, attempt to interact with the ecosystem once, and create the most value from that single interaction as possible. Where a single interaction is not possible, this strategy reduces δ , and minimize second stage losses.⁷³

The case of a specific supply rate or type of data is more straightforward. Here, the producer can select a σ that balances Equation (5-9) and maximize growth accordingly.

Finally, comparison of the two value creation strategies illustrates it is possible that a single decision pair $\langle \sigma, \delta \rangle$ could satisfy both π_p^* and ΔK_p^* . However, in general the producer's broader priorities will determine the strategy pursued and optimum supply level and rate.

4.3. Maximizing Profit for the Data-Analyst and Ecosystem

Up until this point in our analysis, we have proceeded on the basis that the analyst is a supply-taker in the data sharing ecosystem, processing all data supplied by the producer at whatever frequency the producer elects to share it. We now examine the reason for this assumption.

From both Equation (5-5) and Figure 5-4c, the value captured by the analyst from shared data is chiefly a function of the payoff received from derivative data products created from access to the producer's data.

Proposition 5-3 *A profit-maximizing data-analyst is never satiated by shared data from the producer. Proposition 5-3 gives one central result:*

⁷³ For completeness, where $\hat{\sigma} < \sigma_{\Delta K}^*$, the producer would operate as in the unconstrained case.

- i) *The data-analyst always desires the data-producer to increase both its level and rate of participation in the data sharing system.*

Proof. From Equation (5-6), a data-analyst's profit has only positive coefficients for σ and δ and that a decrease in stage length, t , induces an increase in δ . Therefore, an increase in data shared by the data-producer only ever increases the data-analyst's profit. ■

While Proposition 5-3 accords with our real-world experience of the apparently insatiable desire with which data collection platforms and analysts exhibit in real-world data sharing ecosystems, it also explains those observations in terms of ecosystem inputs and outputs. Additionally, Proposition 5-3 also validates our prior working assumption that the analyst processes all data supplied and shares all insight with the producer.

Result (i) carries important implications for the analyst's behavior towards producers in data sharing ecosystems. From service-dominant logic, analysts co-create value with producers in data sharing ecosystems, however comparison of Result (i) from both Proposition 5-1 and Proposition 5-3 illustrates a tension across the data sharing ecosystem. The producer's profit is concave with respect to increasing σ whereas the analyst's profit increases with increasing σ . To investigate the nature of this tension, let us broaden our analysis to the value created across the entire data sharing ecosystem.

Proposition 5-4 *Operation of a data sharing ecosystem at the Pareto frontier always requires the data-producer to share data at a level and/or rate that is above the amount given by their individual optimum choice.*

Proposition 5-4 gives two key results:

- i) *Collaboration between agents is necessary for a data sharing ecosystem to operate at the Pareto frontier.*
- ii) *The data-analyst always increases its data-related profit when the data-producer overshares data.*

Proof The profit produced by the data sharing ecosystem, $\pi_{eco}(\sigma, \delta)$, is the sum of the profit produced by its agents. In symbols, $\pi_{eco} = \pi_A + \pi_P$. Recall $\pi_A > 0$ and monotonically increases for all $\sigma \in (0, \infty)$ and therefore, $\pi_{A,pareto} > \pi_A^*$. Conversely, $\pi_P^* > \pi_{P,pareto}$ while $\pi_P > 0$. Therefore, a redistribution of value from analyst to producer supports the latter's increase in σ beyond σ_P^* . Proof for δ_{pareto} follows along the same lines. ■

Proposition 5-4 says optimization of the profit produced by a data sharing ecosystem requires the producer to overshare but that the producer will participate in this overshare rationally – that is, the producer will not be made worse-off for participating above their optimum amount.

Data sharing ecosystems are defined by the value co-creation that occurs among its members as agents circulate each other's productive inputs and outputs. The implication of Result (i) is optimization of the productive output across data sharing ecosystems requires agents to go beyond

co-creating around shared data and to collaborate around the shared value produced from the data. In other arrangements, this level of collaboration requires either a formal internal market between agents (Jones & Tonetti, 2020) or for the agents to address disclosure games (Easley et al., 2018; Gentzkow & Kamenica, 2014). Yet, Proposition 5-4 says a producer can be incentivized to share more data than their individual optimum – implying either the producer has incurred a marginal cost for this level of sharing or has been subsidized for their overshare. Result (ii) provides the willing benefactor. The analyst always desires the producer to share more data because, from Proposition 5-3, the analyst's profits continue to increase with increases in data shared by the producer.

However, the analyst is a supply-taker and the producer will rationally cease sharing data once they attain their optimal choice, (σ_p^*, δ_p^*) . Therefore, the analyst must incentivize the producer by at least the amount the producer has lost in value sharing data beyond their optimum,

$$\pi_P(\sigma_p^*, \delta_p^*) - \pi_P(\sigma_{over}, \delta_{over})$$

where $\sigma_p^* < \sigma_{over}$ and $\delta_p^* < \delta_{over}$. The subsidization can increase until $(\sigma_{pareto}^*, \delta_{pareto}^*)$ at which point the analyst has applied all potential surplus as subsidy and the ecosystem has reached a Pareto-efficient allocation of value.

While this subsidy represents a concession of potential value from the analyst to the producer, as we investigate when considering the results of this model, under certain market conditions the producer may reverse the subsidy and underwrite the operations of the analyst. Importantly, these cross-market subsidies implement a Pareto improvement to the ecosystem but do not necessarily expand the productive capacity of the ecosystem.

We turn now to apply this model to analyze the data sharing decisions of both data-producing and data-analyzing agents in the genetic trading ecosystem.

5. Results: Giving Away Genetic Data

5.1. Profiting from, and Improving, Genetics

The decision space available to producers in the genetic trading ecosystem corresponds to a partial constraint of both variables in the model. Participation in the ecosystem requires producers to share a non-zero proportion of either phenotype or genomic data with the analyst, setting $\sigma > 0$.

Additionally, as noted previously, each animal's gestation period sets the minimum practical stage length. This has the effect of constraining producers to a discount rate $\delta \leq 0.3867$. Further, the nature of the genetic data enables the analyst to retrospectively approximate raw data from previous stages *even if the producer chose not to share the data in those previous stages*. Therefore,

as in the model, the generation and sharing of data fully depreciates the value of all data and data-products created in the previous stage.

Note that the analyst in this ecosystem enriches data for a range of animals, with a particular focus on sheep and cattle (Banks, 2019). To simplify the following analysis, we will focus the remainder of this analysis exclusively on the commercial trade of genetic data pertaining to cattle.⁷⁴ Table A.1 summarizes values for each parameter in the model.

To establish a theoretical ceiling, let us assume for a moment that producers had free reign of their decision space setting $\sigma \in [0, \infty)$ and $\delta \in (0, 1]$. From Equations (5-7) and (5-8), the maximum two-stage profit, π_p^* , the producer could achieve from sharing data in this genetic trading ecosystem is \$387,301. This output is attained when $\langle \sigma_{\pi}^*, \delta_{\pi}^* \rangle = \langle 0.6432, 0.1673 \rangle$. Similarly, without constraints, the maximum two-stage capital growth, ΔK_p^* , the producer could achieve from sharing data is \$119,286 and is attained when the producer operates at $\langle \sigma_{\Delta K}^*, \delta_{\Delta K}^* \rangle = \langle 0.1367, 1 \rangle$. These results are summarized in Table 5-1.

However, producers' participation in this ecosystem requires they operate beneath a threshold discount rate, $\hat{\delta} = 0.3867$. Producers remain free to choose terms that result in a lower discount rate, as would occur if they elected not to generate data in a particular stage but they cannot choose terms that produce a discount rate above the threshold. Initially observe that $\hat{\delta} > \delta_{\pi}^*$ and therefore from Proposition 1 the producer may increase profits if they adopted longer stage lengths and reduced δ towards δ_{π}^* although this strategy would be difficult to implement given the biologically determined gestation period of cattle. Notice also from Proposition 2 as $x_2 < \gamma$, that is, the value of data ceded in the second stage is beneath the partnership's two-stage productive limit, $\delta_{\Delta K}^* = 1$ implying producers who seek capital gains ought to shorten stage lengths.

Therefore, producers in this data sharing ecosystem operate within conditions that support a variety of value creation strategies. Producers can decide to participate to maximize short-term returns, longer-term growth, or attempt a compromise of both strategies. While beyond the present scope of analysis, this structural tension permits specification of specific producer behavior including assessment of the effect of price elasticity across data sharing ecosystems.⁷⁵

⁷⁴ Applying Result (iii) from Proposition 5-1 to this hybrid ecosystem, as the gestation period for sheep is approximately half the gestation period for cattle the threshold discount rate, $\hat{\delta}_{sheep} \cong \sqrt{\hat{\delta}_{cattle}}$. This implies producers of genetic material for sheep would need to share a lower proportion of data (smaller σ) to maintain the same relative efficiency as producers of genetic material for cattle.

⁷⁵ Notionally, producers could share a proportion of data that causes x_2 to exceed γ , and while this strategy would remove the tension across terms, it would also reduce their capital gain from shared data.

Comparison of optimum conditions predicted by the model to a producer’s actual behavior reveals producers choose to operate at the threshold discount rate and vary the proportion of data they share with the ecosystem (Banks, 2019). This has the effect of reducing the two-dimensional decision to just the proportion of data shared. Figure 5-5 illustrates the calculated profit curve for the analyst and the calculated profit and capital growth curves for the producer over $\sigma \in (0,1]$. The average proportion of data shared by producers in this ecosystem is $\bar{\sigma}_{act} = 0.6206$, which is represented by the line AA' .

Several traits of this genetic trading ecosystem become immediately apparent in the light of *Propositions 1-3*. First, producers maximize capital growth at $(\sigma_{\Delta K}^*, \delta_{act}) = (0.1549, 0.3867)$ while profits peak at $(\sigma_{\pi}^*, \delta_{act}) = (0.4264, 0.3867)$. These positions are indicated on Figure 5-5 by the lines BB' and CC' , respectively. Therefore, producers in this genetic trading ecosystem are currently oversharing their data, or to invert the causality: producers currently *share value with the ecosystem to create data*. In the absence of any subsidization from the analyst or market, if producers were to form investment strategies optimized for two-stage returns, they would rationally reduce σ to a value that fell within the 27-point range $BB' - CC'$. Table 5-1 summarizes each decision.

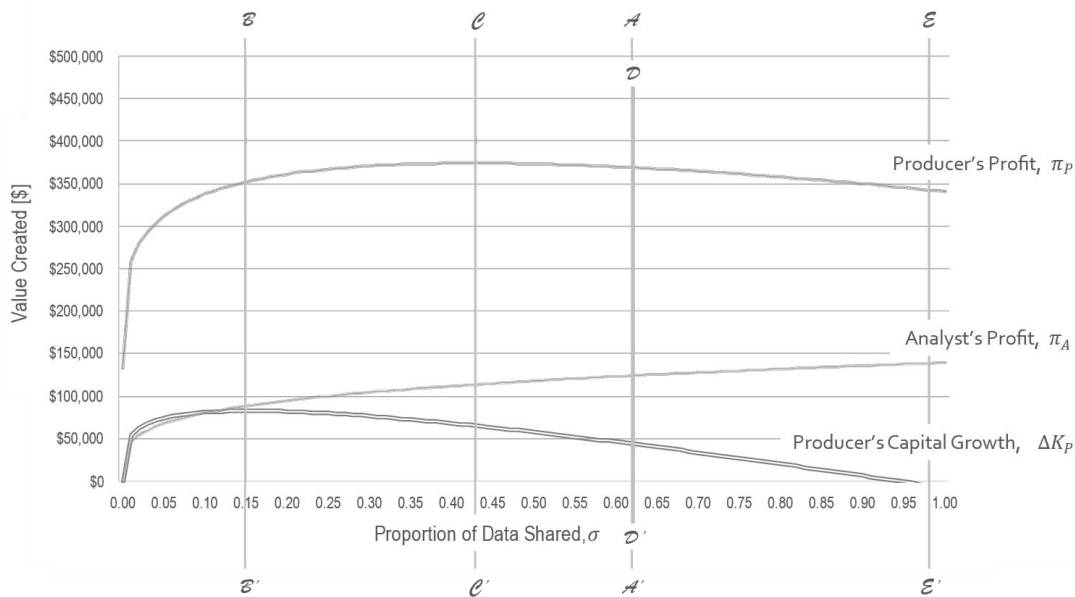


Figure 5-5. Proportion of data traded with effect on ecosystem profit and growth.

$$\sigma \in (0,1] \text{ and } \hat{\delta} = \delta_{act} = 0.3867$$

The analyst’s profits do not peak $\sigma \in (0,1]$ but continue to increase with increasing σ . This creates an ecosystem with competing incentives where the analyst desires producers to increase σ , while producers desire a σ that maximises their preferred ratio of profit and capital growth.

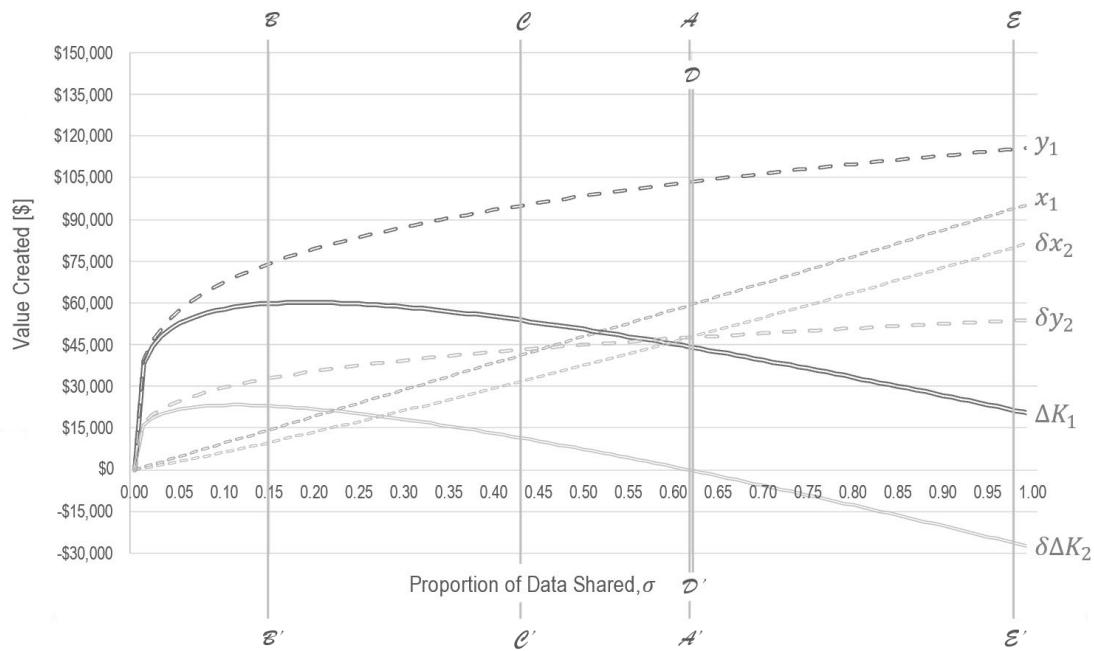


Figure 5-6. Proportion of data traded with effect on producer's multi-stage profit and capital growth

The model also permits finer-grain analysis of the producer's value creation. Figure 5-6 includes additional *per-stage* detail on the effect of σ on producers' profit and capital growth. As σ increases, second stage costs, x_2 , increase at a constant rate while second stage returns, y_2 , increase at a diminishing rate. Eventually, the second stage capital growth, ΔK_2 , becomes negative at the point where $x_2 = \gamma$ represented by the line DD' . This is the graphical representation of the finding from Proposition 2 that data sharing ecosystems possess a two-stage productive limit. Notice that $x_2(\bar{\sigma}_{actual}) \approx \gamma$, that is, on average producers in this genetic trading ecosystem have reached a short-term (two-stage) satiation for the exchange of their data-goods. This behavior is consistent with an ecosystem that treats the transaction of data according to goods-dominant logic, rather than service-dominant logic (Vargo & Lusch, 2008).

5.2. From Oversharing Irrationally to Rational Oversharing

Producers in this partially constrained genetic trading ecosystem currently participate according to, $\langle \bar{\sigma}_{act}, \delta_{act} \rangle = \langle 0.6206, 0.3867 \rangle$ and therefore, two-stage, data-based profits, $\pi_{act} = \$368,879$. From Table 5-1 producers currently overshare data and could increase two-stage profits by reducing their participation to, $\langle \sigma_{\pi}^*, \delta_{act} \rangle = \langle 0.4264, 0.3867 \rangle$. Predicted foregone profit for each producer from oversharing data across two stages is \$5,149 or 1.49%. The oversharing also creates a surplus for analysts of \$13,457 as they enjoy the increased payoffs from producers' 'irrationally shared' data.

However, Proposition 5-4 tells us analysts could incentivize the producer to share data at a level that was even further beyond their optimal. To enable a Pareto-efficient allocation of short-term value in this ecosystem, analysts would need to convince producers to increase participation to $\langle \sigma_{pareto}^*, \delta_{act}^* \rangle = \langle 0.9839, 0.3867 \rangle$. This amounts to an increase in sharing by producers of approximately 36 points beyond their current position and 56 points beyond their optimal amount. If the analyst was to subsidize the producer to move from their optimal decision it would take a subsidy of \$31,796. This subsidy would amount to a complete transfer of all additional value created by the analyst from the overshare of data to the producer. Data-producers in this ecosystem would overshare rationally – by 230% – with the analyst meeting the associated cost. From this point, any change in sharing would reduce at least one agent’s welfare, making this arrangement Pareto-efficient.

However, taking the producers’ current overshare as foregone profits and therefore a deadweight loss, the analyst would only need to provide a subsidy of \$6,104 to incentivize producers to increase sharing to the reduced Pareto frontier of $\langle \sigma_{pareto,reduced}^*, \delta_{act}^* \rangle = \langle 0.7278, 0.3867 \rangle$

Table 5-1. Summary of Producer Decisions in Both Actual and Unconstrained Genetic Trading Ecosystem

Decision Status	Symbols	Decisions Pair	Value	Line
Actual Genetic Trading Ecosystem (see Figure 5-5 and Figure 5-6)				
Producer’s Current Decision	$\langle \bar{\sigma}_{act}, \delta_{act} \rangle$	$\langle 0.6206, 0.3867 \rangle$	$\pi_{act} = \$368,879$ $\Delta K_{act} = \$43,747$	AA'
Maximum possible profit for producers	$\langle \sigma_{\pi}^*, \delta_{act} \rangle$	$\langle 0.4264, 0.3867 \rangle$	$\pi_{P,act}^* = \$374,028$	CC'
Maximum possible capital growth for producers	$\langle \sigma_{\Delta K}^*, \delta_{act} \rangle$	$\langle 0.1549, 0.3867 \rangle$	$\Delta K_{P,act}^* = \$82,757$	BB'
Limit of two-stage productive partnership	$x_2(\sigma_{\gamma}) = \gamma$	$\langle 0.6161, 0.3867 \rangle$	-	DD'
Proportion of data sharing for Pareto optimal allocation of value	$\langle \sigma_{pareto}^*, \delta_{act} \rangle$	$\langle 0.9839, 0.3867 \rangle$	$\pi_{P,pareto}^* = \$342,232$	EE'
Unconstrained Genetic Trading Ecosystem (see Figure A1 and Figure A2)				
Maximum Profit for Producer	$\langle \sigma_{\pi}^*, \delta_{\pi}^* \rangle$	$\langle 0.6432, 0.1673 \rangle$	$\pi_p^* = \$387,301$	
Maximum Capital Growth for Producer	$\langle \sigma_{\Delta K}^*, \delta_{\Delta K}^* \rangle$	$\langle 0.1367, 1 \rangle$	$\Delta K_p^* = \$119,286$	

5.3. Reverse-Subsidies: Producers Paying Analysts

In this genetic trading ecosystem profit maximization occurs before the two-stage productive limit. However, if v_A rose by approximately 40% to \$4.58, pursuit of $\hat{\sigma}_\pi^*$ would induce $x_2 > \gamma$ causing profit-maximizing producers to share data at a level that amounted to a subsidization of the analyst. Somewhat counter-intuitively, in this scenario, pursuit of a profit maximization strategy would compel the producer to incentivize the analyst to take their data through a cross-market subsidization of data assets or some other value such as fiat currency.

The characterization of the payment of services using a combination of fiat currency and data under specific terms introduces considerable opportunities for variation into the model, ranging from the introduction of an additional entrance cost, signaling costs levied against final payoffs, or non-anonymous tolls that scale in proportion to the analyst's yield (Langley & Leyshon, 2017). We leave this adaptation of the model as a novel extension of this work.

Finally, consideration of the unconstrained ecosystem offers producers who desire greater returns the opportunity to develop multidimensional strategies to increase the value of their data. Figure A1 and Figure A2 illustrate the full scope of value creation for a data-producer in this genetic trading ecosystem as they vary how much – and how often – they share data.

6. Discussion

This paper relates the non-rivalry and conditional network effects of data with the value created by the exchange of data. The exchange of data with other agents forms data sharing ecosystems that promote the co-creation of value around mutually shared data. Motivated by an existing data sharing ecosystem that trades genetic data, this paper develops and applies a formal model that evaluates the effect agents' data-sharing decisions have on the value each creates and captures.

We find data-producing agents in the data trading ecosystem overshare data while data-analysts miss the opportunity to incentivize even greater levels of data sharing and value capture. Agents on both sides of the data-exchange participate at sub-optimum levels as they trade data according to goods-dominant logic rather than service-dominant logic. This implies they also under-utilize the opportunity to circulate knowledge and skills within the enriched data, making sub-optimal decisions for themselves and inefficient decisions for the ecosystem. Adoption of service-dominant logic and cross-market data-based, collaboration would enable this data trading ecosystem to accelerate value co-creation and rationally expand data sharing by up to 230%.

Beyond this specific example, this paper demonstrates how an agent's decisions regarding the level – and rate – of data shared drive both short- and longer-term value created and captured. We describe the conditions when it is necessary to give data away to create value. We also invert that causality and use the model to describe why an agent ought to give value away to create data.

6.1. Data-based Co-opetition: Sharing *Value* to Create Data

Agents share data to capture value beyond what they could achieve themselves. A data-analyst must process and re-share all data supplied to maximize the value they capture from data. However, adoption of the same strategy produces a sub-optimum outcome for a data-producing agent who must find an internal solution to their data-based, production process to avoid an overshare of data (Mullins & Sabherwal, 2022). When agents in the ecosystem make decisions in isolation, such as when data is traded according to a goods-dominant logic, disconnected production processes incentivize competitive arrangements causing data trading markets to deliver an inefficient allocation of value (Jones & Tonetti, 2020). However, while cross-market collaboration enables increases in value created – and typically value captured – collaboration between otherwise competing agents requires an understanding of the processes that produce value from their exchanged data (Cichy et al., 2021; Shin et al., 2022). This paper provides a framework that connects the non-rivalry of data with the value created by its recursive exchange to demonstrate how agents within a data sharing ecosystem can collaborate and capture value above the level either could achieve alone.

As illustrated by the results, this collaboration requires agents to be prepared to invert the nominal process orchestrated by the data sharing ecosystem and share *value* to create data. Rather than stopping at sharing 'just enough data' to create the value they need (Mullins & Sabherwal, 2022), this paper demonstrates the conditions when it is profit-maximizing for an agent on either side of the exchange to share data beyond their individual optimum conditions. A data-producer might increase value captured by increasing the value of data shared. A data-analyst could subsidize the data-producer and incentivize them to share data beyond their nominal optimum level.

6.2. Generative AI: The Case for Collaboration and the Warning for Overshare

Generative AI platforms such as *OpenAI's ChatGPT* and *Google Bard* include data that has been 'shared' in two distinct phases: the initial training of the AI model (for example, Ouyang et al. (2022)) and users' subsequent interactions with the model once it has been released.

In the former case, data producers were not necessarily aware of the ends their shared data would be used creating criticisms of “data laundering” (Guadamuz, 2023). In this case, data producers such as content creators have inadvertently overshared, unaware that their data would be enriched in a manner that would be applied in competition with them. When their data was initially shared, it was a form of rational oversharing as content creators published their work in the expectation that publishers (analysts) would provide value to them in the form of discoverability. However, this discoverability also supported the formation of a hybrid data sharing ecosystem comprising two competing analytic systems: the original content publisher and the nascent generative AI. The introduction of this new analyst recasts content creators’ initial participation as irrational oversharing as they inadvertently subsidized the nascent analyst through provision of their data. Crucially, this development of the data sharing ecosystem constitutes more than just a technological disruption and is also different to two competing ecosystems. In this variant, analysts rely on the same dataset and ostensibly the same community for value creation. Competition between analysts represents an important extension of this model. We conjecture when ecosystems that host two competing analysts are left ungoverned, a likely market-driven outcome will be data-based market failure and eventual tragedy of the (data) commons.

In the second phase, users of generative AI resemble producers in the data sharing ecosystem as modelled. As agents they share data which consists of an initial question, or ‘prompt’, with the analytic system in exchange for an immediate increase in the utility of social or knowledge assets. Producers can increase the value co-created through recursive ‘prompt-engineering’ as they direct the system towards a specific goal. Of note, while these systems are billed as “example[s] of how imaginative humans and clever systems can work together to make new things” (OpenAI, 2022), the model proposed by this paper explicates the important nuances of the co-opetitive nature of this data-based value creation. The cooperative aspect might be “amplifying our creative potential” (ibid) but as in the previous example of content creators, while data producers participate for short-term gain, generative AI systems are currently eschewing short-term value creation for accrual of long-term, algorithmic value.

There are two other important extensions of this model that warrant discussion. The first deals with the supply of data, or rather the supply of *lower quality* data to the analyst. Throughout we have assumed the analyst processes all data received from the producer. However, where the producer supplies lower quality data, or data that inhibits rather than supports the improvement of the analyst’s technology, the analyst would rationally reduce or filter the data ingested as low-quality data would impair the efficacy of the AI model rather than improve it. This is the function engineers fulfill when determining what data to train and maintain AI on.

The second extension relates to the analyst's demand for data. Where there is an abundance of data, such as training large-scale AIs like GPT and Bard, the assumption that the analyst will consume all data will not necessarily hold. For example, generative AI is trained on an abundance of available data but not actually all available data. Inclusion of enrichment costs in the analyst's data-based, production process for the analyst (see for example, (Parker & Van Alstyne, 2018)) means at some point the payoff curve for the analyst will become concave as the net value of considering more data becomes negative.

6.3. Caveats in Data-based Subsidies

Beyond generative AI, this paper also identifies other caveats around subsidies in data sharing ecosystems. Suppose the analyst supplements the producer's market payoff directly. From Result (ii) of Proposition 5-1 the range of yield ratios the producer can maintain ideal terms within narrows. Cross-market subsidization of payoffs has made the producer *less* likely to share more data with the analyst. Alternatively, if the analyst subsidized the producer's opportunity cost of participating in the ecosystem such as either an increase in the non-market value of data-producing operations, or even the value of *non-data* related operations⁷⁶ the subsidy would expand the range of yield ratios a producer would still want to share data within. Noting that this type of cross-market subsidization has precedent in literature (Langley & Leyshon, 2017; Parker & Van Alstyne, 2002), we leave extension of the policy implications for governance of data sharing ecosystems for later development.

However, where participation in the ecosystem places constraints on the proportion of data producers must share producers may still set profit-maximizing terms while $\frac{y_2}{y_1} > \frac{\bar{w}_{P,1} - Z_2}{\bar{w}_{P,1}}$. This is the scenario users of search engines or social media platforms encounter. Producers are required to share a minimum amount of data such as search terms or email addresses in order to participate in the ecosystem. Where the minimum limit is less than the producer's optimum the producer may operate as in the unconstrained case, pursuing ideal terms and maximizing value created from their shared data. However, where the lower bound exceeds the optimum proportion shared, such as where the producer knows the analyst will compete against them, the data-producer would choose the lower bound and calculate terms for that value, as any terms at the constrained limit will be preferable to ideal terms beyond the limit.

⁷⁶ Z in the proposed model.

6.4. Governance of Data Sharing: Moving from Theoretical Guidance to Quantitative Thresholds

The implications of Proposition 5-1 also extend beyond ex-post responses by participating producers -that is, producers who are already participating in data sharing ecosystems. Potential data-producers can use Proposition 1 to develop data governance policies from theoretically determined boundary conditions to quantitative thresholds that must be satisfied prior to participation. These thresholds amount to 'approved operating envelopes' that permit agents to trade off when to share data (or withhold data) with how much data to share. Equations (5-7) and (5-8) could be applied to either specific partnering analysts or entire data sharing ecosystems as they establish thresholds based on attributes from both sides of data markets.

Broadening Wysel and Baker (2023)'s treatment of the calculation of the internal rate of return (IRR) for data-projects, where a producer has set a minimum IRR for data-sharing and stage length is determined extrinsically, such as seasonal sales cycles, a threshold discount rate, $\hat{\delta}$, could be applied to each prospective ecosystem. Comparison of $\hat{\delta}$ to the ideal discount rate provided by Equation (5-7) given by expected payoffs and shadow values provides the firm with clear quantitative decision criteria regarding what data – if any – to share. Where optimum terms are less than the threshold discount rate, above-IRR returns can be achieved by maintaining ideal terms. Conversely, where ideal terms exceed the threshold substitution of threshold terms into Equation (5-8) provides the requisite amount of data to share such that profits are maximized given the constrained IRR.

6.5. Vertically Integrating in Data Sharing Ecosystems: When to Bring Analysts Inside

Finally, Proposition 5-1 also supports derivation of criteria for when to go beyond subsidization and to internalize a data sharing ecosystem. The variable cost ratio for shared data, β , enables a producer who is able to achieve desired terms but unable to achieve desired profits to evaluate the relative contribution of the other agents in the ecosystem and respond accordingly. Following Buzzell (1983), a high variable cost ratio would lead a producer to prioritize market-based strategies attempting to increase the market value of insight, while a producer with a low variable cost ratio would focus on operational efficiencies across the ecosystem, such as increasing the analyst's yield. As the analyst operates independently from the producer, the producer may pursue this second strategy directly, through an increase in the amount of data shared, or indirectly by targeted investments in the analyst's technology or the efficacy of their partnership.

Finally, pursuit of optimum profits in a data sharing ecosystem may require a $\beta > 1$. Where data management costs exceed the market returns from insight the producer has given away more value

to the analyst than they receive as a payoff from the market. Under these conditions, if the producer is still sharing less than optimum levels of data – and so increases in sharing also increase profits – the increase in profits comes from increases in the second stage shadow value of latent data and not from the market. It follows, pursuit of increased profits in these conditions requires producers to have established a process for commercializing latent data. In the absence of this commercialization channel, the value-maximizing strategy would be to dynamically reduce the amount shared between stages. We note the analysis of a dynamic sharing model broadens Proposition 1 and leave this as a novel extension for others to pursue.

7. Conclusion

Sharing data amounts to a permanent concession of value, yet the non-rivalry and conditional network effects of data, enable agents to share data and create value – even beyond what is possible if data was ‘just’ an ordinary, private good. Our assessment of an existing data trading ecosystem illustrates that a service-dominant logic of data enables agents to expand the Pareto frontier of the ecosystem by share value in aid of creating more shared data. This paper establishes the value created by sharing a non-rivalrous but excludable good that exhibits conditional data network effects to establish the conditions when an agent should – and should not – share data.

We delineate between two types of data trading agents: *data-producers* who encapsulate ‘normal’ firms or individuals and generate data as a byproduct of their operations, and *data-analysts* who apply knowledge and skills to process generated data into insight that is valuable to data-producers. This arrangement extends to social media platforms and their users, and enterprise platform providers and their clients.

We demonstrate how data-producers can vary the supply level and frequency of shared data to optimize either short- or longer-term value created, and explain how data-analysts, such as platform owners, can create a win-win outcome by incentivizing producers to share data beyond their nominal optimum levels. We extend this cross-market collaboration to include specification of Pareto-efficient allocation of value and specification of a data-based, Pareto frontier.

In a world awash with both data and insight, this paper explains when you should share data, when you should refrain, and when sharing all your data is still not enough.

References

- Acharya, C., Ojha, D., Gokhale, R., & Patel, P. C. (2022). Managing Information for Innovation Using Knowledge Integration Capability: The Role of Boundary Spanning Objects. *International Journal of Information Management*, 62, 102438.
- Akerlof, G. A. (1978). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. In *Uncertainty in Economics* (pp. 235-251). Elsevier.
- Angelopoulos, S., Brown, M., McAuley, D., Merali, Y., Mortier, R., & Price, D. (2021). Stewardship of Personal Data on Social Networking Sites. *International Journal of Information Management*, 56, 102208.
- Arrow, K. J. (1996). The Economics of Information: An Exposition. *Empirica*, 23(2), 119-128.
- Banks, R. (2019). *Benefit and Cost of Performance Recording in the Beef and Sheep Stud*. M. a. L. A. L. (MLA). <https://www.mla.com.au/globalassets/mla-corporate/research-and-development/final-reports/2019/l.gen.1802-final-report.pdf>
- Banks, R. (2022). Email: Decay of Data Commons. In M. Wysel (Ed.), (Decay of Data Commons ed.).
- Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting Systems, Data, and People: A Multidisciplinary Research Roadmap for Chronic Disease Management. *MIS Quarterly*, 44(1), 185-200.
- Bergman, M. T. (2019, January 9). Perspectives on Gene Editing. *Harvard Gazette*. <https://news.harvard.edu/gazette/story/2019/01/perspectives-on-gene-editing/>
- Bresciani, S., Ciampi, F., Meli, F., & Ferraris, A. (2021). Using Big Data for Co-Innovation Processes: Mapping the Field of Data-Driven Innovation, Proposing Theoretical Developments and Providing a Research Agenda. *International Journal of Information Management*, 60, 102347.
- Buzzell, R. D. (1983). Is Vertical Integration Profitable. *Harv. Bus. Rev.:(United States)*, 61(1).
- Cichy, P., Salge, T. O., & Kohli, R. (2021). Privacy Concerns and Data Sharing in the Internet of Things: Mixed Methods Evidence from Connected Cars. *MIS Quarterly*, 45(4).
- Ciriello, R. F. (2021). Tokenized Index Funds: A Blockchain-Based Concept and a Multidisciplinary Research Framework. *International Journal of Information Management*, 61, 102400.
- Clough, D. R., & Wu, A. (2022). Artificial Intelligence, Data-Driven Learning, and the Decentralized Structure of Platform Ecosystems. *Academy of Management Review*(ja).
- Constantinides, P., Henfridsson, O., & Parker, G. G. (2018). Introduction—Platforms and Infrastructures in the Digital Age. *Information Systems Research*, 29(2), 381-400. <https://doi.org/10.1287/isre.2018.0794>
- Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2019). *The Business of Platforms: Strategy in the Age of Digital Competition, Innovation, and Power*. Harper Business New York.
- Department of Agriculture and Water Resources. (2018). *Report to Levies Stakeholders 2017–18*. <https://www.agriculture.gov.au/sites/default/files/sitecollectiondocuments/ag-food/levies/documentsandreports/report-levies-stakeholders-2017-18.pdf>
- Department of Agriculture and Water Resources. (2022, 22 January 2021). *Cattle and Livestock Transaction Levy*. <https://www.agriculture.gov.au/ag-farm-food/levies/rates/cattle-livestock-transaction#exemptions-from-paying-the-cattle-transaction-levy>
- Easley, D., Huang, S., Yang, L., & Zhong, Z. (2018). The Economics of Data. *Available at SSRN* 3252870.
- Evans, J. H. (2021). Setting Ethical Limits on Human Gene Editing after the Fall of the Somatic/Germline Barrier. *Proceedings of the National Academy of Sciences*, 118(22), e2004837117.
- Farrelly, C. P. (2005). Justice in the Genetically Transformed Society. *Kennedy Institute of Ethics Journal*, 15(1), 91-99.
- Fleming, E., Griffith, G., Mounter, S., & Baker, D. (2018). Consciously Pursued Joint Action: Agricultural and Food Value Chains as Clubs. *International Journal on Food System Dynamics*, 9(1012-2018-4116).

- Frankel, A., & Kamenica, E. (2019). Quantifying Information and Uncertainty. *American Economic Review*, 109(10), 3650-3680.
- Gentzkow, M., & Kamenica, E. (2014). Costly Persuasion. *American Economic Review*, 104(5), 457-462.
- Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The Role of Artificial Intelligence and Data Network Effects for Creating User Value. *Academy of Management Review*, 46(3), 534-551.
- Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2022). Data Network Effects: Key Conditions, Shared Data, and the Data Value Duality. In *Academy of Management Review*.
- Guadamuz, A. (2023). A Scanner Darkly: Copyright Infringement in Artificial Intelligence Inputs and Outputs. Available at SSRN 4371204.
- Hagi, A., & Wright, J. (2020a). Data-Enabled Learning, Network Effects and Competitive Advantage. In *Unpublished*.
- Hagi, A., & Wright, J. (2020b). When Data Creates Competitive Advantage. *Harvard Business Review*, 98(1), 94-101.
- Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing Value from Big Data—a Taxonomy of Data-Driven Business Models Used by Start-up Firms. *International Journal of Operations & Production Management*, 36(10), 1382-1406.
<https://www.emeraldinsight.com/doi/pdfplus/10.1108/IJOPM-02-2014-0098>
- Hayden, E. C. (2017). The Rise and Fall and Rise Again of 23andme. *Nature*, 550(7675), 174-177.
- Holder, S. (2019). For Ride-Hailing Drivers, Data Is Power. *CityLab Online*, 22.
<https://www.citylab.com/transportation/2019/08/uber-drivers-lawsuit-personal-data-ride-hailing-gig-economy/594232/>
- Hukal, P., Henfridsson, O., Shaikh, M., & Parker, G. (2020). Platform Signaling for Generating Platform Content. *MIS Quarterly*, 44(3).
- Jagdish, H., Stoyanovich, J., & Howe, B. (2022). The Many Facets of Data Equity. *ACM Journal of Data and Information Quality*, 14(4), 1-21.
- Jasanoff, S., & Hurlbut, J. B. (2018). A Global Observatory for Gene Editing. *Nature*, 555(7697), 435-437.
- Johnson, B. C., Manyika, J. M., & Yee, L. A. (2005). The Next Revolution in Interactions. *McKinsey Quarterly*, 4(25-26).
- Jones, C. I., & Tonetti, C. (2019). *Nonrivalry and the Economics of Data* (0898-2937).
- Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9), 2819-2858.
- Kaiser, C., Stocker, A., Viscusi, G., Fellmann, M., & Richter, A. (2021). Conceptualising Value Creation in Data-Driven Services: The Case of Vehicle Data. *International Journal of Information Management*, 59, 102335.
- Koohang, A., Sargent, C. S., Nord, J. H., & Paliszkievicz, J. (2022). Internet of Things (Iot): From Awareness to Continued Use. *International Journal of Information Management*, 62, 102442.
- Kotlarsky, J., Rivard, S., & Oshri, I. (2023). Building a Reputation as a Business Partner in Information Technology Outsourcing. *The University of Auckland Business School Research Paper Series, Forthcoming, MIS Quarterly (Open Access)*. DOI, 10.
- Langley, P., & Leyshon, A. (2017). Platform Capitalism: The Intermediation and Capitalisation of Digital Economic Circulation. *Finance and Society*, 3(1), 11-31.
- McGuire, A. L., Oliver, J. M., Slashinski, M. J., Graves, J. L., Wang, T., Kelly, P. A., Fisher, W., Lau, C. C., Goss, J., & Okcu, M. (2011). To Share or Not to Share: A Randomized Trial of Consent for Data Sharing in Genome Research. *Genetics in Medicine*, 13(11), 948-955.
<https://www.nature.com/articles/gim2011159>
- Müller, O., Fay, M., & Vom Brocke, J. (2018). The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. *Journal of Management Information Systems*, 35(2), 488-509.

- Mullins, J. K., & Sabherwal, R. (2022). Just Enough Information? The Contingent Curvilinear Effect of Information Volume on Decision Performance in Is-Enabled Teams. *MIS Quarterly*, 46(4).
- Nalebuff, B. J., & Brandenburger, A. M. (1997). Co-Opetition: Competitive and Cooperative Business Strategies for the Digital Economy. *Strategy & leadership*.
- Ogbanufe, O. (2023). Securing Online Accounts and Assets: An Examination of Personal Investments and Protection Motivation. *International Journal of Information Management*, 68, 102590.
- OpenAI. (2022). *Dall.E 2 Promotional Video* [Online]. <https://openai.com/product/dall-e-2>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
- Parker, G., & Van Alstyne, M. (2002). Unbundling in the Presence of Network Externalities. *Mimeo*.
- Parker, G., & Van Alstyne, M. (2018). Innovation, Openness, and Platform Control. *Management Science*, 64(7), 3015-3032.
- Parker, G., Van Alstyne, M., & Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. WW Norton & Company.
- Parker, G., Van Alstyne, M., & Jiang, X. (2017). Platform Ecosystems: How Developers Invert the Firm. *MIS Quarterly*, 41(1).
- Pentland, A., Lipton, A., & Hardjono, T. (2021). *Building the New Economy: Data as Capital*. MIT Press.
- Rahmati, P., Tafti, A. R., Westland, J. C., & Hidalgo, C. (2020). When All Products Are Digital: Complexity and Intangible Value in the Ecosystem of Digitizing Firms. *Forthcoming, MIS Quarterly*.
- Rochet, J. C., & Tirole, J. (2003). Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1(4), 990-1029.
- Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2021). From User-Generated Data to Data-Driven Innovation: A Research Agenda to Understand User Privacy in Digital Markets. *International Journal of Information Management*, 60, 102331.
- Schreieck, M., Wiesche, M., & Krcmar, H. (2021). Capabilities for Value Co-Creation and Value Capture in Emergent Platform Ecosystems: A Longitudinal Case Study of Sap's Cloud Platform. *Journal of Information Technology*, 36(4), 365-390.
- Shapiro, C., & Varian, H. R. (1998). *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Press.
- Shin, D., Kee, K. F., & Shin, E. Y. (2022). Algorithm Awareness: Why User Awareness Is Critical for Personal Privacy in the Adoption of Algorithmic Platforms? *International Journal of Information Management*, 65, 102494.
- Stoeklé, H.-C., Mamzer-Bruneel, M.-F., Vogt, G., & Hervé, C. (2016). 23andme: A New Two-Sided Data-Banking Market Model. *BMC medical ethics*, 17(1), 1-11.
- Tremblay, M. C., Kohli, R., & Rivero, C. (2023). Data Is the New Protein: How the Commonwealth of Virginia Built Digital Resilience Muscle and Rebounded from Opioid and Covid Shocks.Pdf. *MIS Quarterly*, 47(1).
- Vargo, S. L., & Lusch, R. F. (2008). Service-Dominant Logic: Continuing the Evolution. *Journal of the Academy of marketing Science*, 36, 1-10.
- Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On Value and Value Co-Creation: A Service Systems and Service Logic Perspective. *European Management Journal*, 26(3), 145-152.
- Varian, H. R. (2014). Beyond Big Data. *Business Economics*, 49(1), 27-31.
- Wagner, A., Wessels, N., Brakemeier, H., & Buxmann, P. (2021). Why Free Does Not Mean Fair: Investigating Users' Distributive Equity Perceptions of Data-Driven Services. *International Journal of Information Management*, 59, 102333.

- Windasari, N. A., Lin, F.-r., & Kato-Lin, Y.-C. (2021). Continued Use of Wearable Fitness Technology: A Value Co-Creation Perspective. *International Journal of Information Management*, 57, 102292.
- Wysel, M., & Baker, D. (2021, December 2021). Sandwiches Vs. Genes. Sharing Data to Maximise Its Value. MODSIM2021, 24th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, Sydney.
- Wysel, M., & Baker, D. (2023). Profiting from Data. How Data Enables Firms to Have Their Cake, Sell It, and Eat It Too. *Manuscript submitted for publication*.
- Wysel, M., Baker, D., & Billingsley, W. (2021). Data Sharing Platforms: How Value Is Created from Agricultural Data. *Agricultural Systems*, 193, 103241.

Appendix

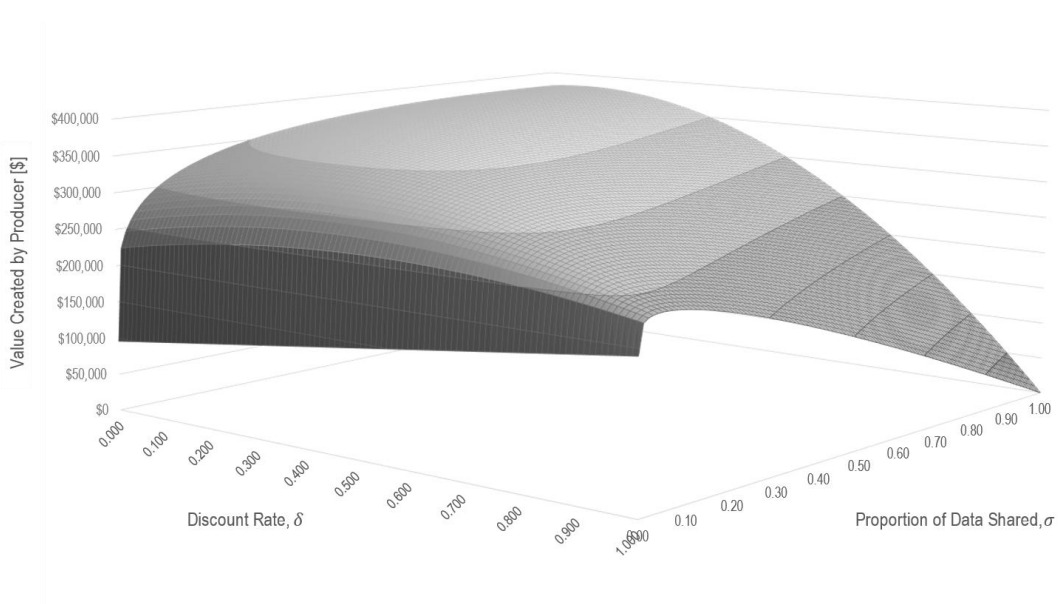


Figure A1. Two-stage profit for the data-producer where $\sigma \in [0, \infty)$ and $\delta \in [0,1]$.

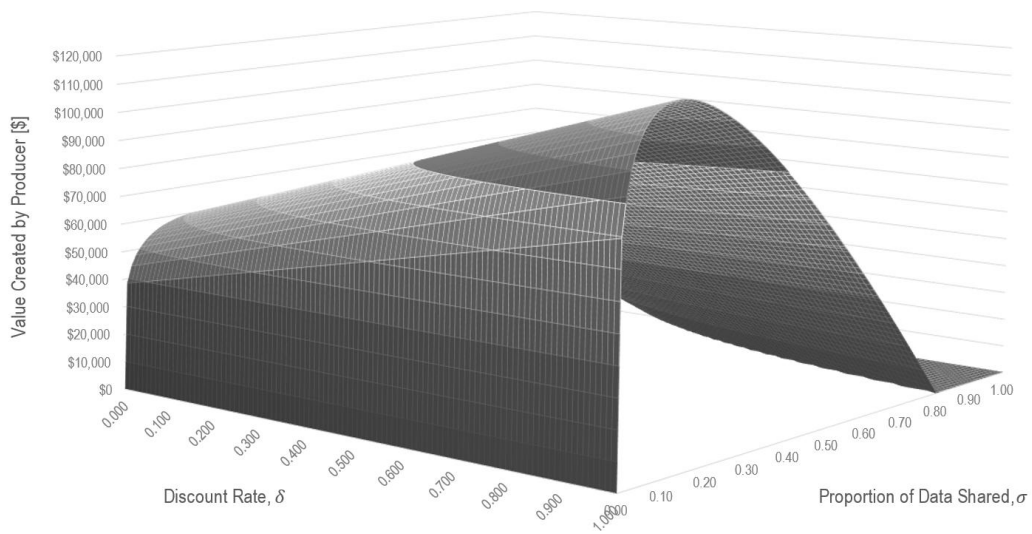


Figure A2. Two-stage capital growth for the data-producer where $\sigma \in [0, \infty)$ and $\delta \in [0,1]$.

Note the inflexion point of $\frac{\partial \Delta K}{\partial \delta}$ at $\sigma \cong 0.62$

Table A.1 Values for each parameter in the genetic trading ecosystem

Parameter	Meaning in Model	Value in Ecosystem	Meaning in Genetic Trading Ecosystem	Comments	Source
π_0	Producer's profit	\$217,243			<i>Calculated.</i> See Equation (5-3).
ΔK_0	Capital growth of producer's assets	\$43,747			<i>Calculated.</i> See Equation (5-4).
σ	Proportion of data shared	0.62	Median cost of sales ratio for producers who participate in data sharing ecosystem	This also corresponds to the average <i>Completeness of Performance</i> (COP) that large breed societies report.	Banks (2019, p. 18)
δ	Discount coefficient	0.386741	-	Calculated. $\delta = e^{-rt}$	
V	Value of data-producing assets	\$95,833	Average, annual sales for producers who participate in data sharing ecosystem	Across-years average total sales value.	Banks (2019, p. 18)
v_0	Value per unit output of producer's improved genetic data	\$3.27	Normalized improvement of bull sales due to participation in data sharing ecosystem		Banks (2019, p. 18)
p_A	Unit price of analyst's benchmarking report	\$1.02	Weighted, average reimbursement of analyst per interaction with producer	<i>Calculated.</i> <i>Median stock level * weighted average levy * factor applied to breeding services</i>	Banks (2019); Department of Agriculture and Water Resources (2018, 2022)
t	Length of stage	1	Normalized gestational period of cattle		
r	Effective interest rate of insight	95%	Percentage drop in value of enriched data between stages		Banks (2022)
k	AGBU's Coefficient of enrichment	7833.25	Coefficient of efficacy of genetic trading ecosystem	Represents real output per unit input from the partnership. <i>Calculated.</i> Polynomial regression.	Banks (2019, p. 18)
α	AGBU's Enrichment technology	0.234815	Coefficient of analyst's technology	Represents diminishing returns of further technological investment. <i>Calculated.</i> Polynomial regression.	Banks (2019, p. 18)

Proof of Propositions

Summarizing the prior treatment of variables, a firm may vary both the proportion of operational value invested into data, $\sigma \in [0, \infty)$, and the length of time between each investment, referred throughout as the stage length, $t \in [0, \infty)$. The stage length adopted by the producer will be influenced by several operational considerations, including the effective interest rate of insight, $r \in (0,1)$. We can formally connect the discount rate, δ , to both t and r by noting $\delta = e^{-rt}$. Note that δ moves in the opposite direction to both r and t ; for example, while r is constant, a *decreasing* δ implies an *increasing* t .

Proposition 5-1 deals with variation of σ and δ by the producer to maximize the profit from sharing data into the ecosystem across two stages.

The discount rate that maximizes profit over two stages is determined by setting $\frac{\partial \pi}{\partial \delta} = 0$. This can be done quickly by rearranging π into different powers of δ and differentiating throughout w.r.t δ :

$$\begin{aligned} \frac{\partial \pi}{\partial \delta} &= \frac{\partial}{\partial \delta} \{[(1 - \sigma)V + vy_1] + [(1 - \sigma)(V + y_1) + v(y_2 - y_1)]\delta + [-vy_2]\delta^2\} \\ 0 &= [(1 - \sigma)(V + y_1) + v(y_2 - y_1)] - 2[vy_2]\delta \end{aligned}$$

Solving for δ gives,

$$\delta^* = \frac{(1 - \sigma)(V + y_1) + v(y_2 - y_1)}{2(vy_2)}.$$

Note that: $z_1 = (1 - \sigma)(V + y_1)$, $\bar{w}_{P,2} - \bar{w}_{P,1} = v(y_2 - y_1)$, and $\bar{w}_{P2} = vy_2$ which produces Equation (5-7).

We must also establish the conditions which maintain $\delta^* > 0$. Imposing that condition on Equation (5-7) and simplifying gives $\bar{w}_{P2} - \bar{w}_{P1} + z_2 > 0$. Dividing throughout by y_1 and rearranging for y_2/y_1 gives the condition on Equation (5-7).

To show that δ^* is a maxima we can take the second partial derivative of profit which by inspection from above produces, $\frac{\partial^2 \pi}{\partial \delta^2} = -2vy_2$. As both v and y_2 are always positive, $\frac{\partial^2 \pi}{\partial \delta^2} < 0 \forall \delta$ indicating δ^* creates a maxima in π .

Derivation of σ^* follows along similar lines. Using the expression for profit from Equation (2-3) and differentiating w.r.t σ gives,

$$\begin{aligned}\frac{\partial \pi}{\partial \sigma} &= \frac{\partial Z}{\partial \sigma} + \frac{\partial W_p}{\partial \sigma} \\ &= \frac{\partial}{\partial \sigma} \left((1 - \sigma)(V + \delta(V + y_1)) \right) + \frac{\partial}{\partial \sigma} (v(1 - \delta)(y_1 + \delta y_2))\end{aligned}$$

At σ^* , $\frac{\partial \pi}{\partial \sigma} = 0$, while multiplying throughout by σ greatly simplifies the treatment. Completing the derivative from above,

$$0 = (-\sigma V - \delta\sigma(V + y_1)) + \delta(1 - \sigma)\alpha y_1 + v(1 - \delta) \left(\alpha y_1 + \delta\alpha y_2 + \delta\alpha^2\sigma \frac{y_1}{x_2} y_2 \right). \quad (\text{A5-1})$$

Recalling that $X = \sigma V + \delta\sigma(V + y_1)$, $z_2 - z_1 = (1 - \sigma)y_1$, $x_2 - x_1 = \sigma y_1$, $W_p = w_1 + \delta w_2 = v(1 - \delta)(y_1 + \delta y_2)$ and for completeness, $w_2 = v(1 - \delta)y_2$, Expression (A5-1) may be written,

$$X = \alpha[W_p + \delta(z_2 - z_1) + \alpha\delta w_{p2} (x_2 - x_1)/x_2].$$

Setting $\beta_2 = x_2/w_{p2}$ which is equivalent to variable cost ratio for the producer in the second stage, gives us Equation (5-8).

Establishing that $X(\sigma^*)$ produces a maximum profit for the owner is laborious to do manually but nonetheless procedural. Dividing Equation (A5-1) by σ (recall we multiplied throughout to ease manipulation) and taking the derivative w.r.t σ gives,

$$\frac{\partial^2 \pi}{\partial \sigma^2} = \alpha\delta \frac{\partial}{\partial \sigma} (1 - \sigma)y_1 - \delta \frac{\partial}{\partial \sigma} (y_1) + \alpha v(1 - \delta) \frac{\partial}{\partial \sigma} \left(\frac{y_1}{\sigma} + \delta \left(\frac{1}{\sigma} + \alpha \frac{y_1}{x_2} \right) y_2 \right)$$

which can be followed through to:

$$\begin{aligned}\frac{\partial^2 \pi}{\partial \sigma^2} &= \frac{\alpha\delta}{\sigma^2} y_1 (1 - \sigma)(1 - \alpha) \\ &\quad + \alpha v(1 - \delta) \left(\frac{y_1}{\sigma} (\alpha - 1) + \delta \frac{y_2}{\sigma^2} \left(\alpha^2 \sigma \frac{y_1}{x_2} + \alpha - 1 \right) \right. \\ &\quad \left. + \alpha\delta \frac{y_1 y_2}{x_2^2} \left((2\alpha - 1)(V + y_1) + \alpha(\alpha - 1)y_1 \right) \right)\end{aligned}$$

Which first reapplying the substitutions following Equation (A5-1), Equation (5-8) and simplifying gives,

$$\frac{\partial^2 \pi}{\partial \sigma^2} = -\frac{(1 - \alpha)}{\sigma^2} X - \frac{\delta(x_2 - x_1)}{\sigma^2} \frac{\alpha}{\beta_2} \left(\frac{x_1\alpha(\alpha - 1) - x_2(\alpha + 1)(2\alpha - 1)}{x_2} \right)$$

Which can be shown numerically to remain negative for all σ . Inspection of Figure A1 also indicates the same properties.

Proposition 5-2 deals with the decision pair that maximize a data-producer's capital growth.

$\sigma_{\Delta K}^*$ is determined by setting $\frac{\partial \Delta K}{\partial \sigma} = 0$ and solving for σ . From Equation (5-2), $\Delta K_P = Y - X = (y_1 - x_1) + \delta(y_2 - x_2) = (kx_1^\alpha - x_1) + \delta(kx_2^\alpha - x_2)$ and therefore,

$$\frac{\partial \Delta K_P}{\partial \sigma} = (\alpha V k x_1^{\alpha-1} - V) + \delta \left(\alpha k x_2^{\alpha-1} (V + y_1(1 + \alpha)) - (V + y_1(1 + \alpha)) \right),$$

which by reverse substitution of $y = kx^\alpha$ gives Equation (5-9).

Completing the second order derivative indicates if the equality from Equation (5-9) is a maxima or minima. In two steps,

$$\begin{aligned} \frac{\partial^2 \Delta K_P}{\partial \sigma^2} = & V^2 \left(\alpha(\alpha - 1) \frac{y_1}{x_1^2} \right) \\ & + \delta \left(\alpha(\alpha + 1) \frac{y_1}{\sigma} \left(\alpha \frac{y_2}{x_2} - 1 \right) + (V + y_1(1 + \alpha))^2 \left(\alpha(\alpha - 1) \frac{y_2}{x_2^2} \right) \right) \end{aligned}$$

Which is categorically negative when $x_2 > (\alpha k)^{1/1-\alpha}$, and must be solved numerically otherwise.

Cursory inspection of Figure A2 reveals $\sigma_{\Delta K}^*$ to be a maxima and therefore this expression to be negative in σ .

The discount rate that optimizes capital growth for a data-producer is likewise found by setting the first order derivative to zero. From Equation (5-2),

$$\begin{aligned} \frac{\partial \Delta K_P}{\partial \delta} &= (y_2 - x_2) \\ &= (kx_2^\alpha - x_2) \end{aligned} \tag{A5-2}$$

Notice that $\frac{\partial \Delta K_P}{\partial \delta}$ is independent of δ and therefore, $\frac{\partial^2 \Delta K_P}{\partial \delta^2} = 0 \forall (\sigma, \delta)$. Equation (A5-2) says the slope of a data producer's capital growth is their second-stage marginal return, and that for any given σ $\frac{\partial \Delta K_P}{\partial \delta}$ is constant. Therefore, $\delta_{\Delta K}^*$ will be pushed towards either its maximum or minimum bound.

Simplifying Equation (A5-2) gives the limit at $x_2 = k^{1/1-\alpha}$ and therefore the limits given in Equation (5-10).

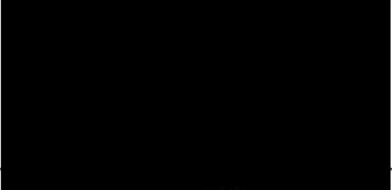
Higher Degree Research Thesis by Publication

University of New England

Statement of Authors' Contribution


We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated in the *Statement of Originality*.

	Author's Name	Percentage of Contribution
Candidate	Matthew Wysel	80
Other Authors	Derek Baker	15
	William Billingsley	5



Candidate

Date



Principal Supervisor

Date

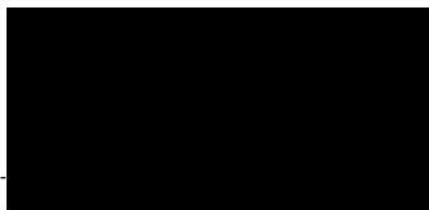
Higher Degree Research Thesis by Publication

University of New England

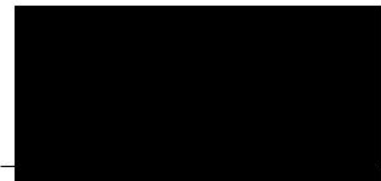
Statement of Originality

We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that the following text, figures, and diagrams are the candidate's original work.

Author	Type of Work	Page Numbers
Matthew Wysel	Conceptualization, methodology, formal analysis, writing (original draft), writing (review), editing.	Entire Document
Derek Baker	Conceptualization, writing (review), editing, supervision.	Entire Document
William Billingsley	Conceptualization, writing (review), editing, supervision.	Entire Document



Date



Principal Supervisor

Date

Chapter Six:

Agtech, Agricultural data and Market Failure.

Avoiding a Tragedy of the (Data) Commons.

Overview of Manuscript

Status:	Reviewed by <i>Computers and Electronics in Agriculture</i>
Date Submitted:	April 2023
Date Published:	-
Suggested Citation:	Wysel, M., Baker, D., & Banks, R. (2023). Agtech, Agricultural data and Market Failure. Avoiding a Tragedy of the (Data) Commons. Manuscript submitted for publication.

Summary of Paper in Context of PhD:

This paper closes out several themes introduced in the PhD and addresses the final research objective by empirically testing the assertion from Chapter 2 that “the creation of value from data requires the operation of a Data Sharing Platform” (p.13) by investigating what happens to the value creation process when one component of the data sharing ecosystem is impaired. The result is a slow decline in value and the eventual, functional collapse of the ecosystem.

By continuing the analysis of breeders and systems from Chapter 5, this paper demonstrates that microeconomies such as Breed Societies that benefit from shared data must consider the effect new technologies have on the health of their data libraries, and not just the value created for members. Specifically, genotyping is a new technology that offers breeders an increase in the value they create from genetic data and more broadly, an increase the value created from data by the whole Breed Society. However, adoption of a service-dominant view of data trading reveals this new technology does not allow the partnering laboratory to maintain the *use-value* of insight (Chapter 4) that the former technology permitted. Unilateral value maximization by the breeders has impaired the value co-creation that characterized the ecosystem. Without intervention from the Breed Society, each breeder rationally maximizes the value created from their data, creating a failure in the market to

deliver premium phenotypic data. Market failure enables breeders to digitally over-graze their society's genetic libraries and, eventually, creates a tragedy of their shared data commons.

While beyond the scope of this paper, this effect can also be understood in terms of the mechanics of the data sharing platform from Chapter 3. Breeders begin providing the system with a new form of correction signal that enables a sufficient decrease in entropy for the specific exchange but lacks sufficient entropy for the ongoing restoration of the database. Where Chapter 3 proposed a static model, this chapter introduces dynamic effects as each message also brings additional, unrestored genotypic data. Appropriating the popular metaphor, breeders are introducing more water into their sinking boat than they are bailing out.

Notice that breeders interpret this phenomenon as a degradation in the value of their shared data library, which while correct, is only a partial explanation as communication theory reveals the increasing entropy of their data library comes from an increase in uncertainty (and not noise) and is therefore a decrease in *specificity* of their library. This produces a drop in accuracy (as observed), but also lowers the barrier to entry for a rogue genetic service provider, such as a foreign data platform, who being unencumbered by political pressure, could enter the market with a substitute genetic library. We substantiate the first claim and offer the second as conjecture in the conclusions to avoid distracting the reader with communication theory.

Apart from typesetting changes and language localization, this chapter appears exactly as submitted to *Computers and Electronics in Agriculture* in April 2023.

Supplementary Publications

Research that informed this chapter also appeared in the following outputs.

Type	Citation
Conference Paper	Wysel, M., & Baker, D. (2022, 8-11 February 2022). Anonymous Tolls and the Value of Insight. 66th Annual Conference of the Australasian Agricultural and Resource Economics Society, Armidale, Australia.
Book Chapter	Fomiatti, B., & Wysel, M. (2022). Developing the Value of Smart Agriculture through Digital Twins. In <i>Encyclopedia of Smart Agriculture Technologies</i> . Springer. https://doi.org/10.1007/978-3-030-89123-7_273-1

Chapter-level Glossary

Term	Definition
Breed Association	The formal organization tasked with preservation and continuation of a breed society's genetic stock (material) and operations.
Breed Society	A group of breeders who own a specific type (breed) of animal. Includes the Breed Association as the management layer.
Breeder	Owner of livestock. Operates farming operations that produce genetic material, such as gametes, bulls, or heifers. Collects raw data as phenotypes or genotypes for internal operations and to facilitate the supply of insight from the Laboratory.
Laboratory	Partnering data enrichment service that turns Breeder's raw data into valuable insight. Equivalent to an <i>Analyst</i> (Chapter 5)
Genetic Insight	Information on the total genetic value of an animal, usually expressed in terms of Estimated Breeding Values (EBVs), specific indices that rank animals against established baselines. Broadly considered a proxy for the worth of an animal to a breeder.
Genotypic Data	Genetic data pertaining to livestock and collected by a breeder. E.g., a tail hair from an animal.
Phenotypic Data	Measurement data pertaining to livestock and generated by a breeder. E.g., 100-day weight.

Full Manuscript

Agtech, Agricultural data and Market Failure. Avoiding a Tragedy of the (Data) Commons.

Authors:	Matthew Wysel ^a (matthew.wysel@une.edu.au;)*, Derek Baker ^a (derek.baker@une.edu.au), Robert Banks ^b (rbanks@une.edu.au)
Affiliations:	^a The Centre for Agribusiness, UNE Business School, The University of New England, Armidale, Australia. ^b The Animal Genetics Breeding Unit (AGBU), The University of New England, Armidale, Australia.
Keywords	Genetic trading, data sharing, club goods, market failure, value of data, genotyping
Highlights:	<ul style="list-style-type: none">• Technology disruption can cause market failure for agricultural data markets.• New tech can increase value of data now but cause digital over-grazing tomorrow.• Agricultural platforms must consider data health not just value created from data.• Breed Societies are data platforms who connect data generators to data consumers.• Avoiding tragedy of data commons requires more than just maximizing the value of data now.
Acknowledgements:	Matthew Wysel is grateful for the support of the Agricultural Business Research Institute through the Arthur Rickards Innovation in Agribusiness Scholarship. All errors remain our own.

Abstract

This paper connects technological disruption with market failure for agricultural data markets. On the one hand, new *agtech* services offer consumers a greater share of the value created from their data. On the other hand, the ecosystems that form around agricultural data rely on the appropriate allocation of externalities to maintain the health of their common data libraries. New technologies change the value each member receives when sharing data, threatening the short-term supply of data, long-term membership of stakeholders, and eventually: viability of the platform.

This paper uses the growing adoption of genotypic technologies by livestock breeders to illustrate the adverse effect even labor-saving technologies can have on data markets. We show that, left ungoverned, widespread adoption of this technology can cause an over-grazing of the Breed Society's shared data library and subsequent tragedy of their data commons. Using a straight-forward analytic model, we evaluate three responses the Breed Society might pursue to avoid market failure.

This paper uses Club Theory to illustrate how marketplace sponsors, such as Breed Associations or platform owners, can balance adoption of Smart Farming technologies with the health of their data assets. We discuss policy implications for management of national, or supra-industry, datasets and provide recommendations for ongoing, finer-grained research.

1. Introduction

Historically, livestock breeders receive insight regarding their herd's genetics by generating and sharing phenotypic – or measurement – data with their Breed Society and industry-based analysts such as the Animal Genetics and Breeding Unit (AGBU).⁷⁷ These analysts use algorithms refined from decades of similar genetic data to enrich breeders' data into insight⁷⁸. That insight enables short- and long-term benefits for breeders as they trade genetic material above commodity rates and achieve long term genetic improvement in their herd (Georges et al., 2019; Miller, 2010). The broader Breed Society is the custodian of the shared genetic asset base. Genetic improvement in an individual herd can result in market gains for each breeder, while genetic improvement in the national or collective breed further differentiates it in commercial markets (Hine et al., 2021).⁷⁹

Breed Societies function as clubs (Sandler & Tschirhart, 1997) as they internalize the positive externalities of sharing data amongst producers of the same breed. Club theory outlines clubs' membership fees and usage licenses as being anonymous, and non-anonymous tolls, respectively. The data goods which Breed Societies produce such as access to their members' shared genetic library, are analogous to club goods as membership is required for access, access is necessary to draw a benefit, and the extraction of benefits is not typically characterized by rivalry (Romano, 1999). In addition to membership fees, breeders also incur direct costs for generating phenotypic data. These costs arise from the labor-intensive measurement of hundreds – and sometimes thousands – of animals. Although these costs can increasingly be capitalized through the adoption of *agtech* products and their accompanying business models (Daum et al., 2021; Fomiatti & Wysel, 2022), they are likely to remain a significant component of data collection activities by breeders and a decisive factor in the management of the aggregate genetic value of herds (Banks, 2019).

However, recent advances in tissue sampling technologies have lowered the cost of genotyping, that is, the cost associated with the direct collection of genetic data from the animal (Amer et al., 2015). These technologies allow breeders to generate genetic insight about an animal by sending a single hair from the animal to a laboratory that queries the breed society's genetic library and returns insight to the breeder analogous to that returned from phenotypic data. Development of insight from genotypes depends on the existence of a relevant 'reference dataset' consisting of phenotypic

⁷⁷ The Animal Genetics Breeding Unit (AGBU) is a joint venture between the NSW Department of Primary Industries (DPI) and University of New England (UNE).

⁷⁸ We use the term *insight* rather than *information* throughout to avoid confusion caused by the conflation of the two terms *data* and *information* by popular media.

⁷⁹ A brief glossary of terms is included in Table A1 in the Appendix.

and genetic data on sufficiently sized reference population. If the reference dataset is not maintained at sufficient size and representativeness of the current population, it no longer supports the generation of reliable insight, and the value of the data drops for all users. Therefore, while the generation of genotypes still involves the contribution of data to the reference library, if the more costly phenotypic data is not also generated, then the reference library becomes increasingly separated from each successive generation of animal and suffers an associated drop in value.

If data were a 'normal' private good, then the introduction of genotyping would be just a technological disruption that channels private surplus to breeders. However, the creation of value from data relies on the active participation of a community of stakeholders, a centralized analytic system, and data about which both collaborate (Wysel et al., 2021). Going further, incorporation of data into a market based production process requires a service-dominant logic where members of the data sharing ecosystem co-create value around *each other's data* (Vargo et al., 2008). The non-rivalrous nature of data makes this persistent pan-market collaboration possible as the use of data in one member's production process neither excludes it from, nor diminishes its value for, other member's production processes (Jones & Tonetti, 2020). The corollary is that the genetic library presided over by Breed Associations becomes a shared, or common, data good which each member – and all members together – contribute to and rely on for the derivation of data-based, value. The capture and allocation of value created from shared data is therefore vitally important for the long-term health of the ecosystem and value-creating activities of its members (Gregory et al., 2021).

Standing in the middle of this ecosystem, industry laboratories enable the transformation of raw data collected by breeders into genetic insight which breeders use to improve the value of their herds. In this context, these laboratories function as data analysts (Wysel & Baker, 2023) who ingest raw data and produce valuable insight for the Breed Society. As analysts, they process all data they receive but they cannot generate data themselves. Therefore, while the laboratories facilitate value creation across the data sharing ecosystem, they do not govern which data is collected, nor participate in the allocation of the resulting value. Rather the laboratory is motivated to continue providing breeders with the most valuable insight for the lowest cost. Data trading across the ecosystem comes to resemble a club where each member acts to maximize the value they take from data but also relies on the contributions of all other members for maintenance of that value creation process. In terms of Club Theory (for example, Sandler and Tschirhart (1997)), all members use the insights generated from data but if all users of the insights don't contribute to the cost of data generation appropriately, the proportion of members who undertake the cost of generating costly data will diminish permitting the erosion of value of the reference library as each new generation of animals is produced. Therefore, as we develop in this paper, Breed Associations must consider

careful economic management to curtail the actions of ‘genetic free riders’ and avoid a tragedy of their shared data commons.

The purpose of this paper is to demonstrate that communities that benefit from mutually shared data must consider the effect of new technologies on the long-term health of their data, and not just the short-term value created for members. To that end, this paper extends earlier models of the data sharing ecosystem to accommodate two competing laboratories – or analytic systems. As described, the breeder generates either phenotypic or genotypic data from their livestock before sharing it through their Breed Society to an analyst. The resulting data sharing ecosystem is illustrated in Figure 6-1(a). The chosen analyst processes the raw data, enriching it to insight which is returned to the breeder. This process increases both the breeder’s short- and long-term operational value while enabling the analyst to operate their technology and – within the limitations of the data contributed – maintain the shared genetic library. The non-rivalry of data also enables the analyst to generate parallel value streams from the breeder’s data by repackaging it into derivative data-products, such as benchmarking reports, which are sold into the market (Wagner et al., 2021). This ecosystem is represented as a simplified microeconomy in Figure 6-1(b).

This paper characterizes the incentives around the data commons as an application of datanomics and assesses the long run consequences of centralized interventions following widespread availability of genotypic technology to breeders. Using representative data, we first project the impact on genetic libraries and secondly evaluate three different economic strategies that Breed Societies might adopt. We propose a decision framework that enables Breed Societies to promote

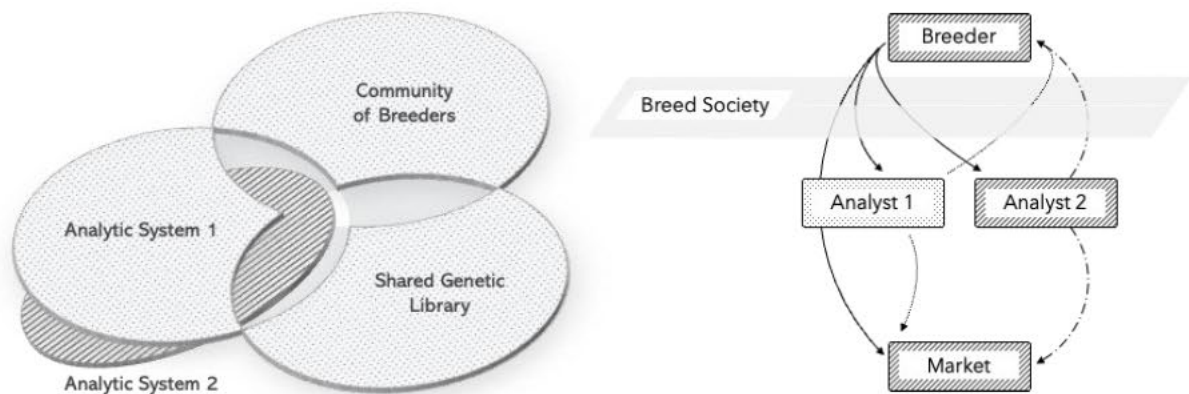


Figure 6-1. Breed societies as a data sharing ecosystem with competing analytic systems

(a) ecosystem view, and (b) data-flow view.

Source: (a) Adaption of data ecosystems proposed by Wysel et al. (2021).

(b) Authors’ representation

adoption of new data management technologies while assuring long-term governance of their shared data assets. For individual breeders, this paper supports the increase in consumer surplus offered by genotyping while offering assurance that the shared genetic libraries breeders depend on are safeguarded against digital over-grazing or worse, genetic cattle rustling.

2. Background Literature

The analysis presented in this paper draws on the three areas of literature about data as a production factor (Pentland et al., 2021), the management of agricultural data (Bahlo & Dahlhaus, 2021; Wolfert et al., 2017), and agricultural *datanomics* – or the study of the attachment of value to agricultural data (Klingenberg et al., 2022; Wysel et al., 2021).

2.1. Club Goods: Insight as Enriched Shared Data

The incorporation of agricultural data into a firm's production process unlocks both short- and long-term benefits (Kosior, 2020, p. 61). Short term value arises as firms collect information regarding their operations, either applying it in production directly or selling it into the marketplace (Liu et al., 2020; Saura et al., 2021). Breeders collect operational data on size of herd, feed levels and growth rates to improve their operations and use observed data on herd characteristics to improve livestock resale value. Enriching data within the ecosystem also confers longer-term value indirectly through data-enabled learning (Hagiu & Wright, 2020). The data that breeders collect on attributes and the performance of their herd enables insight that drives continuous and compounding increases in the genetic value of their animals. Breeders participate in these data-based production processes to achieve an accrual of value to their livestock (Zhao et al., 2019) as non-digital products with aggregated increases in value for both their individual operations (Ramsbottom et al., 2012) and the whole breed society (Amer et al., 2015).

Genetic insight is a club good mediated exclusively by each Breed Society. Accordingly, breeders remain free to retain all value created from their society's genetic library and are dependent on their society to produce that insight. Breeders choose their level of participation in the data sharing ecosystem by varying the proportion and type of data shared and electing what subset of data-related services to subscribe to (Banks, 2019). Where members choose to generate and share phenotypic data, the Breed Society is able continually to re-baseline its data libraries ensuring they remain relevant to all members' herds. In terms of club theory, sharing either phenotypic or genotypic data produces externalities that are captured and allocated by the Breed Society – typically in the form of genetic reports that rank each animal (and indirectly, each breeder) – but

only phenotypic data enables updating of the shared database, so it maintains accuracy for each generation of animals. As noted earlier, this generational externality is vital for the continued 'health' of the shared data commons.

2.2. The Trade of Data: Costs and Ownership

The trade of phenotypic or genotypic data becomes necessary when breeders are not able to generate the required levels of insight internally (Picard et al., 2015). However, this trade creates indirect data-related costs for breeders as data that has been shared is more appropriately considered disclosed (Shin et al., 2022). Many studies appropriate the implied, indirect cost as principally a unilateral extraction of value, such as privacy concerns (Cichy et al., 2021; Wagner et al., 2021), but there is growing research that draws attention to 'sanctioned' value captured by service providers (Klingenberg et al., 2022; Zhang et al., 2020). These service providers range from private companies (Birner et al., 2021) to social co-operatives (Jakku et al., 2016) each bringing different motivations for participation and requirement for governance (Wiseman et al., 2019).

Easley et al. (2018) evaluate the effect of different ownership structures on the value captured by both a supra-industry body such as a breed society and market participants such as breeders. They show that where the governing analyst cares only for its own profit maximization, the overall ecosystem profit is maximized; but that maximization comes at the expense of benefits to market participants. Alternatively, where analysts are owned by, and chartered to the prosperity of, market participants – as is the case of Breed Societies (see for example, Herefords Australia Limited (2021)) – value creation for the whole ecosystem might be lower than in the previous scenario but members maximize value captured from their data under certain conditions. Finally, where there were no data-sharing arrangements – that is, where the analyst avoids active intervention – members eventually chose to withhold data, creating inefficient outcomes at both the individual and collective level (Wu, 2022).

Trade of phenotypic or genotypic data by breeders through Breed Societies to industry analysts also introduces an unavoidable time separation between when a cost is incurred and when the corresponding benefit is received. This arrangement requires long-term relational trust between breeders (Cao et al., 2021) who incur the cost, and their societies who act as custodians of members' benefits (Jakku et al., 2019). Therefore, in a microeconomy where breeders are individual profit-maximizers, the ongoing maintenance of that data-based trust – and ultimately provision of data services by Breed Societies – relies on their efficient allocation of value that is commensurate with perceived value contributions by their members (Banks, 2020).

2.3. Interventions: Subsidies, Anonymous and Non-Anonymous Tolls

Breed Societies act as *ex officio* guardians of the breed's genetic data and must raise funds from their membership to operate schemes that assist all members to derive benefits and to safeguard their members' shared, club goods. While members elect whether to generate phenotypic or genotypic data, they access the services of the nominated data analyst by sharing data with their society, who then share the data with the third-party analyst. Therefore, the Breed Society is uniquely placed to intervene in that exchange, either to divert or allocate funds to breeders as individuals, groups or a collective, based on the type of data shared or quality of insight returned.

The charter of Breed Societies generally resembles the maintenance of "general integrity, genetic integrity, [and] commercial and noncommercial benefits" (Herefords Australia Limited, 2021, p. 3) for members, and they raise funds to those ends. From club theory, these fund-raising schemes can be implemented as either *anonymous*, or *non-anonymous* tolls (Sandler & Tschirhart, 1997) and provide protection against free-riding of the common good (Marciano, 2021) and maintenance of the club's shared commons. Anonymous tolls are tied to factors that affect all members equally, such as membership status, and are useful for defining the boundaries of the club. Non-anonymous tolls are fees levied on members based on their activities within the club or contributions to a common good – such as the shared genomic reference library. These tolls can be combined or inverted and expressed as subsidies if deemed necessary by the Breed Society.

Any intervention levied by Breed Societies is additive to breeders' private costs of generating either phenotypic or genetic data. As membership in the club is strictly voluntary, the society must ensure that the utility breeders derive from membership remains positive. While breeders maintain membership for both public and private benefit, their preparedness to accept marginal costs, such as the effort of generating and sharing phenotypic data, relies on an appropriate allocation of private benefit from that effort. Alternatively, Breed Societies could also subsidize members' costs of participating in the data sharing club if they anticipated a need to do so.

Our analysis proceeds with the assumption that the Breed Society will intervene only to the extent required to protect their (shared) data commons. This suggests that it will not use the technological advancement to reshape the internal data market that exists between members of the ecosystem. For tractability we assume a homogeneous membership in the club which is responsive to change in the price of data services. Finally, as noted earlier, we assume each breeder acts to maximize their own utility without consideration for longer-term societal considerations such as the health of the breed society's common data library.

3. The Theory: Background and Model

We extend the three-layer Venn construction for assessing the value created from agricultural data proposed by Wysel et al. (2021). As described earlier, they define a data sharing ecosystem as the assembly of three constituent parts: a community of stakeholders; a facilitating analytic system; and a set of data. In our case of trading genetic data, the production process entails an analyst who refines data for breeders by reducing uncertainty from raw data (Frankel & Kamenica, 2019). Breeders use insight returned from the analyst to improve bloodlines through selective breeding programs to increase the value of their operations. Breeders sell their improved genetic material in the form of either gametes or within a 'host' animal into an open market that comprises both other breeders and commercial producers. The latter take breeders' genetic material and scale it in large-scale, commercial operations.

Representing this arrangement as a series of transactions provides a model where value generated from data circulates within an ecosystem and flows from the ecosystem to the market (Figure 6-1(b)). Breeders and analysts co-create successively improved genetic datasets as each stakeholder improves the output of the previous owner of the data. It is notable that these exchanges are recursive, causing value created from data to both accrue within the ecosystem and permeate across the ecosystem. A breeder might initially choose to generate phenotypic data with the relevant analyst, who generates insight permitting an increase in the value of the breeder's operations. If the breeder then decides to generate genotypes, the improved data is shared with the other analyst who benefits from the inputs all other parties have previously made. Additionally, no agent is locked into a particular course of action beyond their current exchange. Here the value co-created across the ecosystem has resulted from all members of the ecosystem collaborating around a shared data asset. In the ideal case, this collaboration has the effect of maintaining data in the ecosystem and offering each member opportunities to capture private gains from either new strategies or new technologies (Parker & Van Alstyne, 2018).

Following Lancaster (1966), we assume each breeder derives utility from a bundle of different attributes conferred by insight, rather than the insight *per se*. Breeders' motivation to accept a higher cost for better quality insight depends on attributes of that data that produce differentiated payoffs from insight within their operations. The size of breeders' operations, both in scale and level of investment, and the deployment of supporting infrastructure breeders have amassed to incorporate data-enabled learning within operations, are factors that influence a breeder's willingness to incur greater costs for higher quality insight (Banks, 2019; Zhang et al., 2020). Therefore, we can express the value each breeder captures from sharing data as the sum of the payoff of insight returned from the analyst, W_B , the current value of a breeder's operations, V , less

the cost of supplying data to the analyst, x , and any anonymous tolls associated with membership, F ,

$$\pi_B = W_B + (V - x) - F. \quad (6-1)$$

Note that W_B is a product of insight returned from the analyst, y , and the price each breeder can capture in the market, p (Wysel & Baker, 2023).⁸⁰ In genetic trading markets, insight returned by the analyst is a composite function containing an estimate and an accuracy. The estimate reflects what is known about the animal while the accuracy describes how well that estimate is known. More specifically, accuracy is a compound product that is derived from the variation of actual breeding values from predicted breeding values of each animal. This index is commonly taken as a proxy for the value of insight returned by the analyst. Estimated breeding values are derived from heuristically defined algorithms that tie the measurement points of decades of animals within a breed to a variety of indices that include specific categories of market value and long-term genetic improvement. The phenotypic measurement of each successive generation of animal and its (delayed) market performance permits refinement of the analyst's technology producing benefits for the breeder and positive externalities for all members of the breed society.

The generation, or collection of phenotypic data represents a combination of fixed F and variable x costs to the breeder. Fixed costs are a combination of licensing fees imposed by the Breed Association and analyst plus operating costs from Breeder's activities, while variable costs are tied to the labor cost of measuring each animal at several points across its life. The amount each breeder invests in data acquisition varies and both the Breed Society and analyst can track the relative completeness of measurements each breeder maintains (Gouws, 2016). Breeders understand that the accuracy of predictions returned from the analyst remain proportional to their data-collection ranking within their Society and that the more measurements a breeder makes, the better the quality the insight returned is likely to be. The enrichment technology applied by the analyst conforms with a standard Cobb-Douglas production function $y = kx^\alpha$ (Parker et al., 2017) where k represents the efficacy of the partnership between the breeder and analyst, and $\alpha \in (0,1)$ represents the analyst's diminishing returns of further technological investment.

In contrast, the generation of genotypes confers a fixed cost to the breeder which we take as the sum of analyst's fees and the modest labor of collecting and dispatching the physical genetic sample.

⁸⁰ Strictly W_B is the agent's (breeder's) payoff from insight is a product of insight returned from the analyst, y , and a marginal value, v , achieved by the breeder until they decide to generate new data. This duration gives rise to a discount factor, δ , and the payoff of insight becomes, $W_B = v(1 - \delta)y$. As the present analysis assumes a constant – if binary – participation rate for breeders, we set $v(1 - \delta) = p$ throughout.

Importantly, where phenotypic data returned insight proportional to the amount of data a breeder collected, genotyping returns insight proportional to the quality of the shared genetic library. Therefore, breeders evaluating transactions with analysts in the data sharing ecosystem will make a choice that maximizes the value each individually creates from the exchange by way of:

$$MU = p(y) - x. \quad (6-2)$$

We can now establish a preference matrix for breeders. Where a breeder's payoff for insight is less than the cost of generating genotypes, or where desired accuracy exceeds what is available from genotypes, they would choose to collect phenotypic data. Conversely, where a breeder's requirement for accuracy is below that returned by genotypic data and their payoff for that insight is above the fixed cost of genotyping, they would choose to collect genotypes.

4. Data

Data on the costs and benefits of generating phenotypic and genotypic livestock data is highly commercially sensitive and as a result, generally not available. However, a unique set of representative data was generated with the assistance of the Animal Genetics and Breeding Unit (AGBU) in 2022 (Banks, 2022). The data is summarized in Table A2. The data points pertain to costs typically incurred by breeders during their collection of phenotypic data and includes representative data for reproductive and hereditary data. The data has been adjusted to reflect an effective cost-per-animal.

The data shows breeders operating in three distinct groups. Group 1 from Figure 6-2 operates smaller-scale operations typified by wide ranges in operating costs and low thresholds for accuracy. Group 2 operates with significantly greater efficiency as they focus on cost-reductions within data generation. Group 3 prioritize quality of insight from the analyst and a preparedness to invest in completeness of measurement in order to achieve desired accuracy levels. We return to discuss the implications of economic intervention for each group after presenting the results.

The current cost-per-animal for genotyping is \$45 and returns an equivalent accuracy of 65%. This is represented as *Point G* on Figure 6-2.

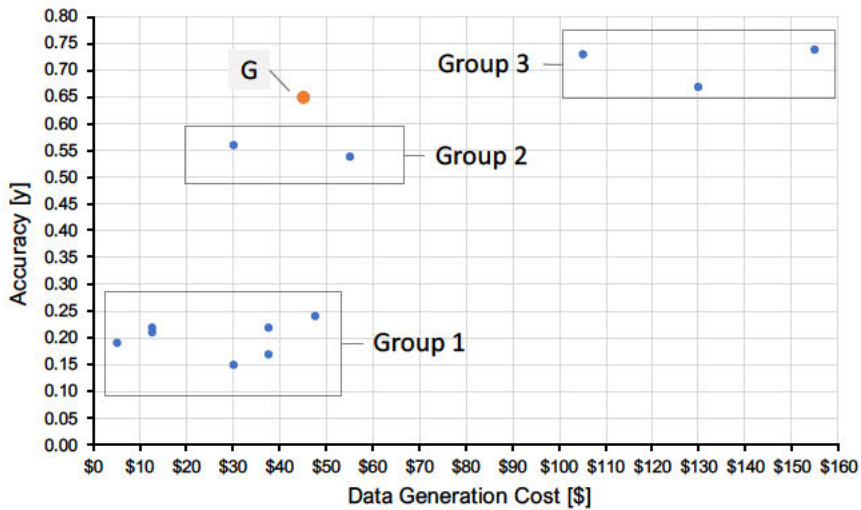


Figure 6-2. Data Collection Costs vs. Accuracy. Source: Authors' representation

5. Results

Applying the Cobb Douglas production function across the phenotypic data gives the benefits curve *AD* with a diminishing marginal utility illustrated in Figure 6-3. Note that the insight returned from genotypes, *G*, is above curve *AD*, which reflects the higher quality insight and better private returns than any phenotypic insight in the range *BC*. Therefore, any profit-maximizing breeder who collected phenotypic data in the range *BC* before the introduction of genotyping would rationally switch to generating and sharing genotypic data from their herd and enjoy a reduction in costs to *B*. Additionally, notice that breeders are unlikely to operate in the range *CD* either, as the step change

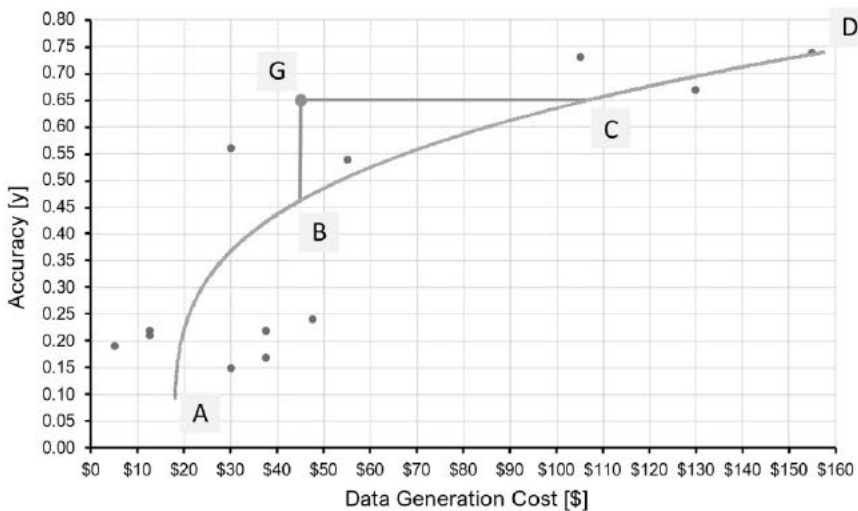


Figure 6-3. Cost/Benefit curve for breeders in 'two-analyst' ecosystem.

Source: Authors' representation.

from G to anywhere in CD represents a significantly lower marginal utility than any breeder had adopted before the introduction of genotyping.

From Figure 6-4, the area BGC represents a pure private surplus for breeders (Fleming et al., 2018), while the area $G\check{G}DC$ represents a derived surplus for the breed society as those who would have operated above C will revert to operating at point G , while benefits accrue up to D . Finally, the change in behavior about C represents a loss for the whole Breed Society as accuracies above G will not be generated without additional cost, and therefore the area $CD\check{D}$ represents a public cost to the society.

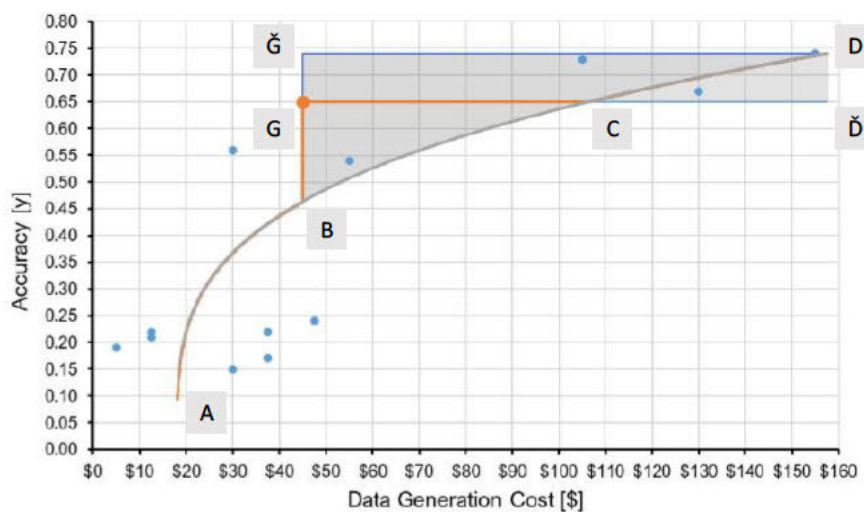


Figure 6-4. Private Surplus and Public Cost Caused by Genetic Sampling.

Source: Authors' representation

Recall that this ecosystem trades data and not 'normal' goods. If the ecosystem traded normal goods, then all we have is the consequence of technical disruption. A new technology has entered the market offering breeders a radically different value proposition and has displaced the incumbent technology. Practically, breeders operating in the range BD now shift to G to capture the private surplus enabled by genotyping. Meanwhile, no one in the ecosystem consumes above B . However, both analysts rely on the sustained maintenance of a shared data library to operate their technology, and this data library decays at a known rate unless it is re-baselined with the provision of phenotypic data (Banks, 2020). As Figure 6-5 illustrates, if all breeders pursue the acquisition and sharing of genotypic data alone, there is universal degradation in the operational value of the breed society's library. The half-life for the breed's genetic library is approximately five years (Banks, 2022), but the operational value of their shared library increases when both phenotypic and genotypic data is circulated throughout the ecosystem. Although this degradation in operational value is caused by a gradual separation of the library from the specific attributes exhibited by animals in breeders' herds,

the reduction in accuracy also resembles the properties of a substitute database – not just a substitute technology. We return to the implications of substitute databases when discussing the results.

As illustrated in Figure 6-5, where the genetic trading ecosystem operates like a free market each breeder acts to maximize their individual profit and partners with whichever analyst returns the best accuracy for the lowest cost. The breed society sees widespread over-grazing of their shared data assets, and a reduction in accuracy and therefore value, of their genetic library. A reduction in the genetic yield for all members follows. Crucially, the reduction in value of the database affects the accuracy that either data sampling technique delivers, making this problem different from a simple short-run degradation of one technique. The consequence is that in the absence of any intervention, the non-rivalry of traded data turns the introduction of a new analytic technology into a pending market failure.

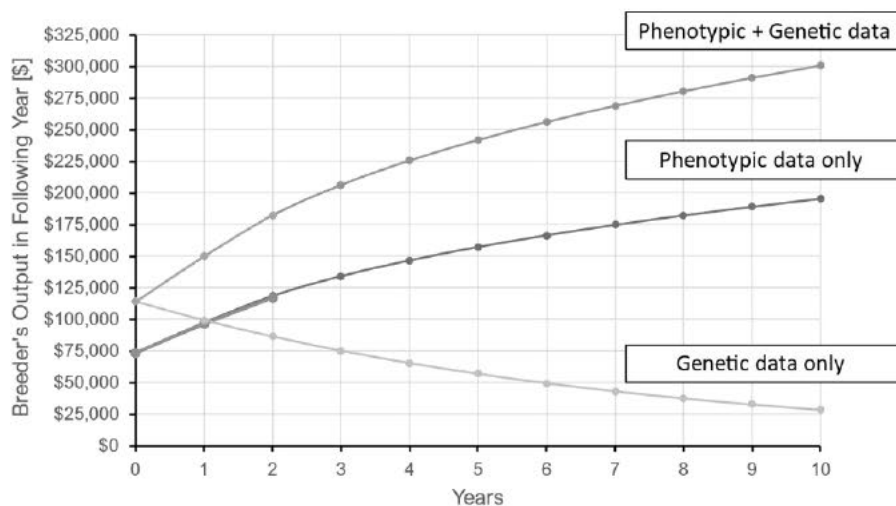


Figure 6-5. Decay of Data Libraries from Rational Over-Grazing.

Source: Authors' representation

Our model depicts Breed Societies which are able to intervene on behalf of their members, according to several strategies. The remainder of this paper investigates the effects of an output tax based on the value of insight produced by the analyst; an input tax based on the type of data generated; and a subsidy for sharing data that acts to maintain the society's common data good. Note that as the breed society brokers only data trading for its members and so while a toll might be applied in isolation with funds held in escrow, any subsidy offered must first be funded by tolls applied to members.

The first intervention is a tariff that is a form of non-anonymous toll the breed society can levy on the output of all breeders who receive insight from sharing genotypic data. As illustrated in Figure 6-6(a) the effect of this tariff is a reduction in value captured by any breeder that shares genotypic data. Unlike other output taxes such as consumption taxes, a tariff on the yield of the analyst that enriches genotypes only suppresses the benefit breeders might otherwise have accessed without capturing value for use elsewhere within the ecosystem. Therefore, as illustrated in Figure 6-6(a), in the short term an output tax on the accuracy provided by genotypic analysis reduces efficiencies across the whole ecosystem and creates a deadweight loss for all breeders as the value of insight return has been suppressed. While acknowledging the deadweight loss, this intervention does create the desired incentive to collect the requisite phenotypic data because while a breeder's input costs remain the same their output has diminished, and therefore, from Equation (6-2) their marginal utility has diminished.

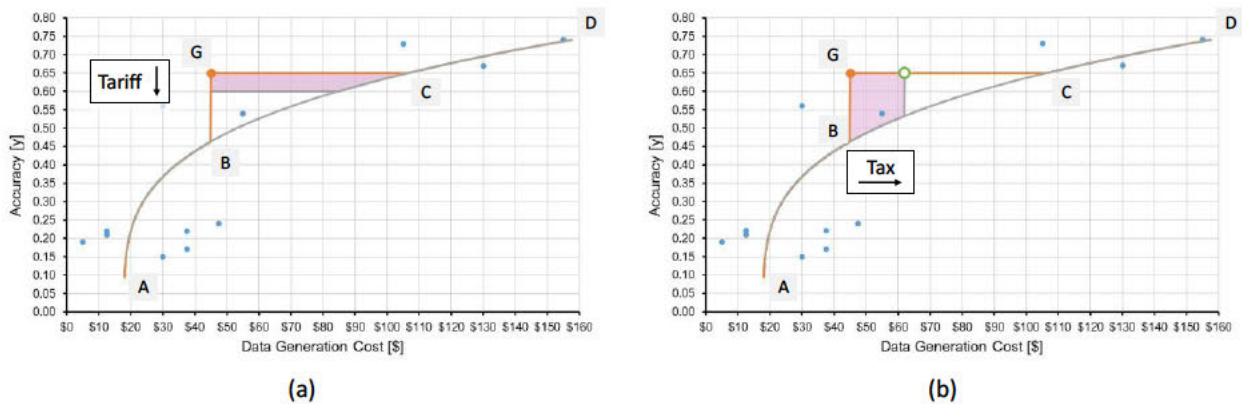


Figure 6-6. Impact of on Interventions on Production of Insight for Data Sharing Ecosystem. (a) Non-anonymous tariff (b) anonymous tax

Source: Authors' representation

Alternatively, the Breed Society could implement an input tax based on the generation of genotypic data which increases the input cost, x , for all breeders who desire genetic insight but do not collect phenotypic data. This toll maintains the value of the output but increases input costs for these breeders. As illustrated in Figure 6-6(b) this form of anonymous toll retains incentives for further genotypic development and permits all breeders to continue receiving the full accuracy available from whichever analyst with whom they partner. This tax also enables the breed society to capture value from breeders using the new data product which, as we develop shortly, enables the breed society to incentivize the requisite collection of phenotypic data. However, in isolation, unless this toll is applied to a level that removes all private surplus for breeders who collect genetic data, this intervention still does not incentivize the collection of requisite phenotypic data.

However, value created from the input tax can then be used to subsidize production of ‘premium’ phenotypic data in the region CD. As illustrated in Figure 6-7, in this specific example a tax of 62% on the cost of submitting genotypic data creates sufficient surplus to maintain the original utility curve of breeders.

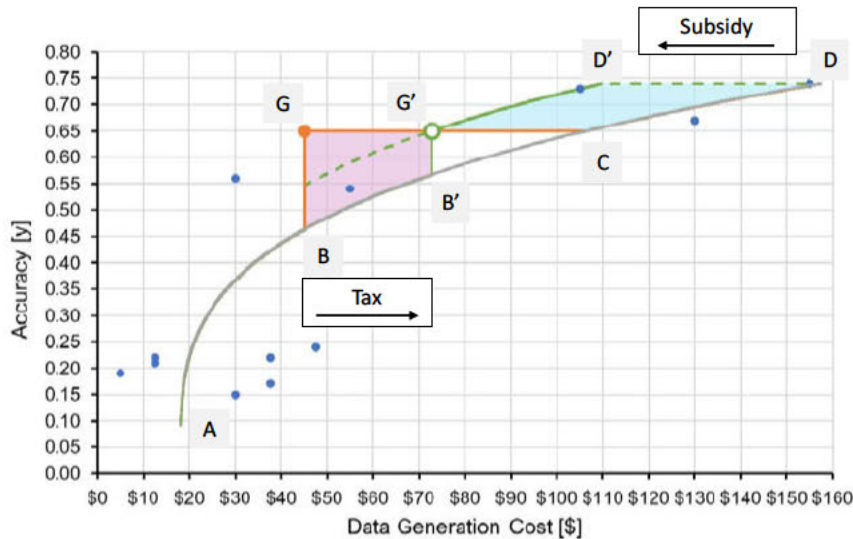


Figure 6-7. Adjusted Cost/Benefit curve where the subsidy for premium phenotypic data is met by the tax on genotyping. Source: Authors’ representation.

Geometrically, the area $BGG'B'$ is the same as the area $G'D'DC$ setting value created from taxes equal to value spent on subsidies. In effect, this intervention permits the Breed Society to allocate a portion of the surplus created by the introduction of genotyping, to the maintenance of the common data library that all breeders and analysts rely on.

The resulting tax and subsidy have the effect of changing the aggregated yield curve that breeders interact with to enrich raw data from $ABCD$ to $AB'G'D'$. This adjusted yield curve is illustrated in Figure 6-8. Importantly, no section of the Breed Society is worse off than when the data sharing ecosystem operated with only one analyst. The proposed price structure permits an increase in the Pareto frontier of this data sharing ecosystem by 6.2%. That is, the proactive management of the introduction of a competing analytic technology by the Breed Society has resulted in an increase in overall productive output and maintained the long-term quality of the shared data assets.

6. Discussion, Extensions and Limitations

This paper addresses the challenge of tracking and allocating value creation in the provision of a common data good whose long-run health determines the welfare of all agents who interact with it.

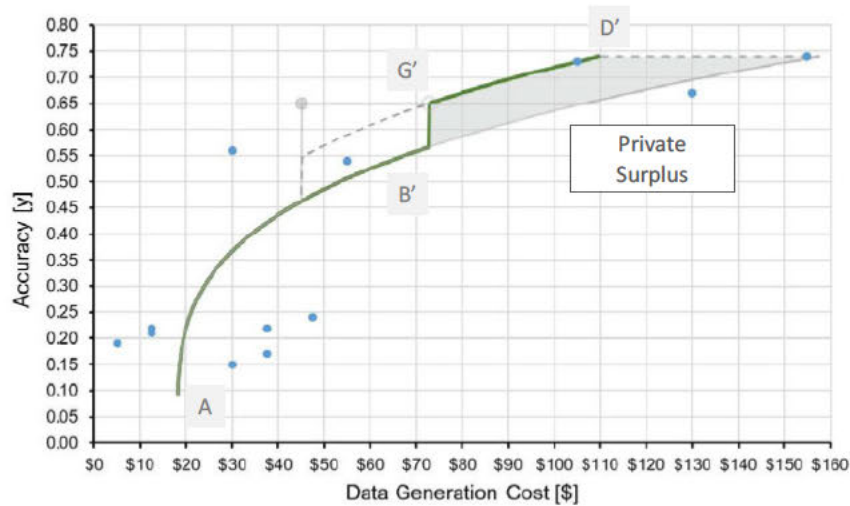


Figure 6-8. Adjusted Cost/Benefit Curve for breeders in a 'two-analyst' ecosystem showing total private surplus. Source: Authors' representation

By interpreting the genetic trading market as a data sharing ecosystem, and identifying data as a non-rival good, this paper demonstrates an emerging technology – that promises the provision of greater consumer surplus – which will actually become a threat to the long-term health of the community's shared data and ultimately, viability of the market. This paper proposes a data-centric response that enables the Breed Society to retain much of the benefit created by the technology and increases the society's total productivity. Crucially, this increase in productive output is stable and permits the Breed Society to safeguard its data assets from genetic over-grazing.

6.1. Increasing the Power of the Model by Relaxing the Assumptions

The envisaged ideal result is that the club solution pushes all breeders towards the top end of the benefit curve. Practically, the effectiveness of this intervention will depend on several factors that include first, the heterogeneity of the Breed Society and, second the elasticity of their response. Both assumptions were made to avoid unnecessary complication of the core message of the importance of maintaining a data commons. Firstly, the heterogeneity of the society is a measure of how much breeders' preferences vary within the Breed Society. Chiefly, this says all members of the Breed Society share a common value for accuracy and are evenly distributed along the benefit curve. This assumption will hold more naturally for larger breed societies than for smaller societies. The second assumption says breeders desire a specific accuracy of genetic insight even if the cost of that accuracy changes. While this assumption extends the first assumption to support the implementation of each of the interventions, notice that the groups observed in Figure 6-2 support this assumption as they are vertically compressed and horizontally (left-to-right) extended. This

indicates breeders in this representative dataset exhibit a range of operating efficiencies as they target a specific accuracy and are willing to accept variances in costs x or F . Variation between groups is noticeably associated with different benefits generated, that is a vertical distance with a corresponding variance in the position occupied along the benefits curve. This variation may be due to groups` collecting different types of phenotypic data, or managing different sized operations (Banks, 2019). Even without the second assumption, tax and subsidy arrangements as proposed would impact each breeder in a different way. Interventions could be targeted to deliver the minimum amount of phenotypic data and could engage and compensate only the most efficient providers of data.⁸¹ This would create additional increases in productive output across the ecosystem as subsidies would not be `wasted` generating redundant phenotypic data. However, analysts are unlikely to be indifferent between breeders inhabiting the different groups and they may seek to influence breeders by charging them spot prices. While this would enable the Breed Society to further increase data collection, with the dynamic effect of boosting benefits to both producers and analysts, they would need an output-based, non-anonymous subsidy to restore the market.

6.2. Extensions

This paper`s research contribution is the modeling of costs and benefits of club action, as strategies to avoid a tragedy of the (data) commons. The application uses an indicative set of secondary data, due to a general paucity of available primary data of this form. An obvious extension of this work is the collection of such data, in a manner that preserves relationships between breeders and Breed Societies. However, we acknowledge that generation of the benefit curve using Breed Societies` members data would, on the one hand, enable personalized service delivery and, on the other hand identify fault lines within the membership and encourage fragmentation.

In the long term, the disruption generated by ever cheaper genomic technologies will impact an ever-larger number of breeders and so remove the fragmentation incentive. Future research would ideally characterize the nature of data provision costs (beyond our indicative fixed and variable components) and relate Breed Societies` club-like needs and capabilities to these challenges.

This research could also be extended technically. Where this paper has assessed the interventions required to capture and allocate positive externalities created by the club within each generational

⁸¹ Interestingly, while appealing from a technical and economic perspective, this outcome would introduce sub-clubs within the Breed Society, which may not be ideal for social reasons.

cycle, as the ecosystem functions recursively, optimization could be conducted across multiple data generation cycles. Recall the collection of new data immediately depreciates old data in the ecosystem and that generation of sufficient phenotypic data permits the restoration of the Breed Society's genetic library. A Breed Society may be able to decrease the multi-year value spent on subsidies (and therefore reduce taxes on genotypic data) by intentionally permitting degradation of their genetic library for multiple cycles, before subsidizing an extraordinary phenotypic campaign to re-baseline the library.

Additionally, in line with our assumption of homogeneity, this analysis assumes preservation of the original benefit curve was the goal. The specific nature of a breed's genomic material and the algorithms dispensed by the analyst may make it possible to subsidize the collection of phenotypic data from specific livestock, rather than subsidizing the operations of the breeders who currently own them. While returning to the risk of encouraging spot-prices on specific livestock and fragmented clubs within the society, this strategy would greatly increase the precision of taxes and subsidies applied.

7. Conclusion

This paper deals with maintenance of data commons. We show these shared data assets both rely on, and are threatened by, technological advancement. Specifically, this paper looks at a Breed Society that stands to benefit from the adoption of genotyping but must also intervene if it is to avoid a market failure and subsequent tragedy of its data commons.

Importantly, these findings are relevant beyond Breed Societies and genetics. Widespread adoption of digital twins in agriculture (Fomiatti & Wysel, 2022) unites existing, if latent, data libraries with new analytical models of creating value from shared data. Likewise, the adoption of generative AI by users who outsource an increasing scope of analytic tasks, from writing emails to augmenting images (Lu et al., 2022), also create communities that derive benefit from mutually shared data and, therefore, must consider the long-term health of their data and not just the short-term value created.

This paper demonstrates that the non-rivalrous nature of data means that new, agricultural data-based technologies cannot be considered in isolation from the data trading ecosystems they support. We demonstrate that any data sharing ecosystem must consider the long-run effects on the value of their data commons when evaluating new enrichment technologies. We show that technological disruption in distributed production processes that deal with the exchange of data extend beyond simply conferring consumer surplus to a subset of agents and – without adequate

data governance – can extend to a failure of the data trading market. Tracking and allocating the creation of value from data is a necessary task for industry bodies such as Breed Societies who preside over data sharing ecosystems and act as *ex officio* governing bodies.

Without collective action, the private benefits that consumers capture from genotyping create an unstable market structure because value creation from various data types are causally linked via a common data platform. This means, ‘doing nothing’ is not an option for platform owners and operators as inaction will simultaneously create increasing technological inefficiencies for all members and decreasing yields at both macro- and micro-decision levels. Finally, in the absence of an adequate response, we conjecture the marginal value contributed by the platform owners such as Breed Societies will diminish even to the extent their shared data asset will become ripe for arbitrage by rogue data platforms who, competing with the whole Breed Association, will rustle first the data and then the whole club.

References

- Amer, P., Byrne, T., Fennessy, P., Jenkins, G., Martin-Collado, D., & Berry, D. (2015). Review of the Genetic Improvement of Beef Cattle and Sheep in the UK with Special Reference to the Potential for Genomics.
- Bahlo, C., & Dahlhaus, P. (2021). Livestock Data—Is It There and Is It Fair? A Systematic Review of Livestock Farming Datasets in Australia. *Computers and Electronics in Agriculture*, *188*, 106365.
- Banks, R. (2019). *Benefit and Cost of Performance Recording in the Beef and Sheep Studs*. M. a. L. A. L. (MLA). <https://www.mla.com.au/globalassets/mla-corporate/research-and-development/final-reports/2019/l.gen.1802-final-report.pdf>
- Banks, R. (2020). *Ethical Dimensions of Big Data in Agriculture* 27th Annual Australian Association for Professional and Applied Ethics, The University of New England.
- Banks, R. (2022). Email: Decay of Data Commons. In M. Wysel (Ed.), (Decay of Data Commons ed.).
- Birner, R., Daum, T., & Pray, C. (2021). Who Drives the Digital Revolution in Agriculture? A Review of Supply-Side Trends, Players and Challenges. *Applied Economic Perspectives and Policy*. <https://doi.org/10.1002/aapp.13145>
- Cao, S., Powell, W., Foth, M., Natanelov, V., Miller, T., & Dulleck, U. (2021). Strengthening Consumer Trust in Beef Supply Chain Traceability with a Blockchain-Based Human-Machine Reconcile Mechanism. *Computers and Electronics in Agriculture*, *180*, 105886.
- Cichy, P., Salge, T. O., & Kohli, R. (2021). Privacy Concerns and Data Sharing in the Internet of Things: Mixed Methods Evidence from Connected Cars. *MIS Quarterly*, *45*(4).
- Daum, T., Villalba, R., Anidi, O., Mayienga, S. M., Gupta, S., & Birner, R. (2021). Uber for Tractors? Opportunities and Challenges of Digital Tools for Tractor Hire in India and Nigeria. *World Development*, *144*, 105480.
- Easley, D., Huang, S., Yang, L., & Zhong, Z. (2018). The Economics of Data. *Available at SSRN* 3252870.
- Fleming, E., Griffith, G., Mounter, S., & Baker, D. (2018). Consciously Pursued Joint Action: Agricultural and Food Value Chains as Clubs. *International Journal on Food System Dynamics*, *9*(1012-2018-4116).

- Fomiatti, B., & Wysel, M. (2022). Developing the Value of Smart Agriculture through Digital Twins. In *Encyclopedia of Smart Agriculture Technologies*. Springer. https://doi.org/10.1007/978-3-030-89123-7_273-1
- Frankel, A., & Kamenica, E. (2019). Quantifying Information and Uncertainty. *American Economic Review*, 109(10), 3650-3680.
- Georges, M., Charlier, C., & Hayes, B. (2019). Harnessing Genomic Information for Livestock Improvement. *Nature Reviews Genetics*, 20(3), 135-156.
- Gouws, A. (2016). Select Your Breeder, Then Your Bull. *Stockfarm*, 6(12), 34-35.
- Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The Role of Artificial Intelligence and Data Network Effects for Creating User Value. *Academy of Management Review*, 46(3), 534-551.
- Hagiu, A., & Wright, J. (2020). Data-Enabled Learning, Network Effects and Competitive Advantage. In *Unpublished*.
- Herefords Australia Limited. (2021). *Constitution of Herefords Australia Limited*. <https://www.herefordsaustralia.com.au/wp-content/uploads/2021/05/Herefords-Australia-Limited-Constitution.May2021.pdf>
- Hine, B. C., Duff, C. J., Byrne, A., Parnell, P., Porto-Neto, L., Li, Y., Ingham, A. B., & Reverter, A. (2021). Development of Angus Steerselect: A Genomic-Based Tool to Identify Performance Differences of Australian Angus Steers During Feedlot Finishing: Phase 1 Validation. *Animal Production Science*.
- Jakku, E., Taylor, B., Fleming, A., Mason, C., Fielke, S., Sounness, C., & Thorburn, P. (2019). "If They Don't Tell Us What They Do with It, Why Would We Trust Them?" Trust, Transparency and Benefit-Sharing in Smart Farming. *NJAS-Wageningen Journal of Life Sciences*, 90-91, 100285.
- Jakku, E., Taylor, B., Fleming, A., Mason, C., & Thorburn, P. (2016). Big Data, Trust and Collaboration.
- Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9), 2819-2858.
- Klingenberg, C. O., Júnior, J. A. V. A., & Müller-Seitz, G. (2022). Impacts of Digitalization on Value Creation and Capture: Evidence from the Agricultural Value Chain. *Agricultural Systems*, 201, 103468.
- Kosior, K. (2020). Economic, Ethical and Legal Aspects of Digitalization in the Agri-Food Sector. *Zagadnienia Ekonomiki Rolnej/Problems of Agricultural Economics*.
- Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74(2), 132-157.
- Liu, Y., Soroka, A., Han, L., Jian, J., & Tang, M. (2020). Cloud-Based Big Data Analytics for Customer Insight-Driven Design Innovation in Smes. *International Journal of Information Management*, 51, 102034.
- Lu, Y., Chen, D., Olaniyi, E., & Huang, Y. (2022). Generative Adversarial Networks (Gans) for Image Augmentation in Agriculture: A Systematic Review. *Computers and Electronics in Agriculture*, 200, 107208.
- Marciano, A. (2021). Retrospectives: James Buchanan: Clubs and Alternative Welfare Economics. *Journal of Economic Perspectives*, 35(3), 243-256.
- Miller, S. (2010). Genetic Improvement of Beef Cattle through Opportunities in Genomics. *Revista Brasileira de Zootecnia*, 39, 247-255.
- Parker, G., & Van Alstyne, M. (2018). Innovation, Openness, and Platform Control. *Management Science*, 64(7), 3015-3032.
- Parker, G., Van Alstyne, M., & Jiang, X. (2017). Platform Ecosystems: How Developers Invert the Firm. *MIS Quarterly*, 41(1).
- Pentland, A., Lipton, A., & Hardjono, T. (2021). *Building the New Economy: Data as Capital*. MIT Press.

- Picard, B., Lebret, B., Cassar-Malek, I., Liaubet, L., Berri, C., Le Bihan-Duval, E., Hocquette, J.-F., & Renand, G. (2015). Recent Advances in Omic Technologies for Meat Quality Management. *Meat Science*, *109*, 18-26.
- Ramsbottom, G., Cromie, A., Horan, B., & Berry, D. (2012). Relationship between Dairy Cow Genetic Merit and Profit on Commercial Spring Calving Dairy Farms. *Animal*, *6*(7), 1031-1039.
- Romano, D. (1999). Genetic Resource Valuation Methodologies, Their Strengths and Weaknesses in Application: The Contingent Valuation Method. *Economic valuation of animal genetic resources*, 47.
- Sandler, T., & Tschirhart, J. (1997). Club Theory: Thirty Years Later. *Public Choice*, *93*(3-4), 335-355.
- Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2021). From User-Generated Data to Data-Driven Innovation: A Research Agenda to Understand User Privacy in Digital Markets. *International Journal of Information Management*, *60*, 102331.
- Shin, D., Kee, K. F., & Shin, E. Y. (2022). Algorithm Awareness: Why User Awareness Is Critical for Personal Privacy in the Adoption of Algorithmic Platforms? *International Journal of Information Management*, *65*, 102494.
- Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On Value and Value Co-Creation: A Service Systems and Service Logic Perspective. *European Management Journal*, *26*(3), 145-152.
- Wagner, A., Wessels, N., Brakemeier, H., & Buxmann, P. (2021). Why Free Does Not Mean Fair: Investigating Users' Distributive Equity Perceptions of Data-Driven Services. *International Journal of Information Management*, *59*, 102333.
- Wiseman, L., Sanderson, J., Zhang, A., & Jakku, E. (2019). Farmers and Their Data: An Examination of Farmers' Reluctance to Share Their Data through the Lens of the Laws Impacting Smart Farming. *NJAS-Wageningen Journal of Life Sciences*, *90-91*, 100301.
- Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big Data in Smart Farming—a Review. *Agricultural Systems*, *153*, 69-80.
- Wu, J. (2022). Secondary Market Monetization and Willingness to Share Personal Data. *Available at SSRN 4269334*.
- Wysel, M., & Baker, D. (2023). Profiting from Data. How Data Enables Firms to Have Their Cake, Sell It, and Eat It Too. *Manuscript submitted for publication*.
- Wysel, M., Baker, D., & Billingsley, W. (2021). Data Sharing Platforms: How Value Is Created from Agricultural Data. *Agricultural Systems*, *193*, 103241.
- Zhang, Y., Baker, D., & Griffith, G. (2020). Product Quality Information in Supply Chains: A Performance-Linked Conceptual Framework Applied to the Australian Red Meat Industry. *The International Journal of Logistics Management*, *31*(3), 697-723.
- Zhao, J., Xu, Z., You, X., Zhao, Y., He, W., Zhao, L., Chen, A., & Yang, S. (2019). Genetic Traceability Practices in a Large-Size Beef Company in China. *Food chemistry*, *277*, 222-228.

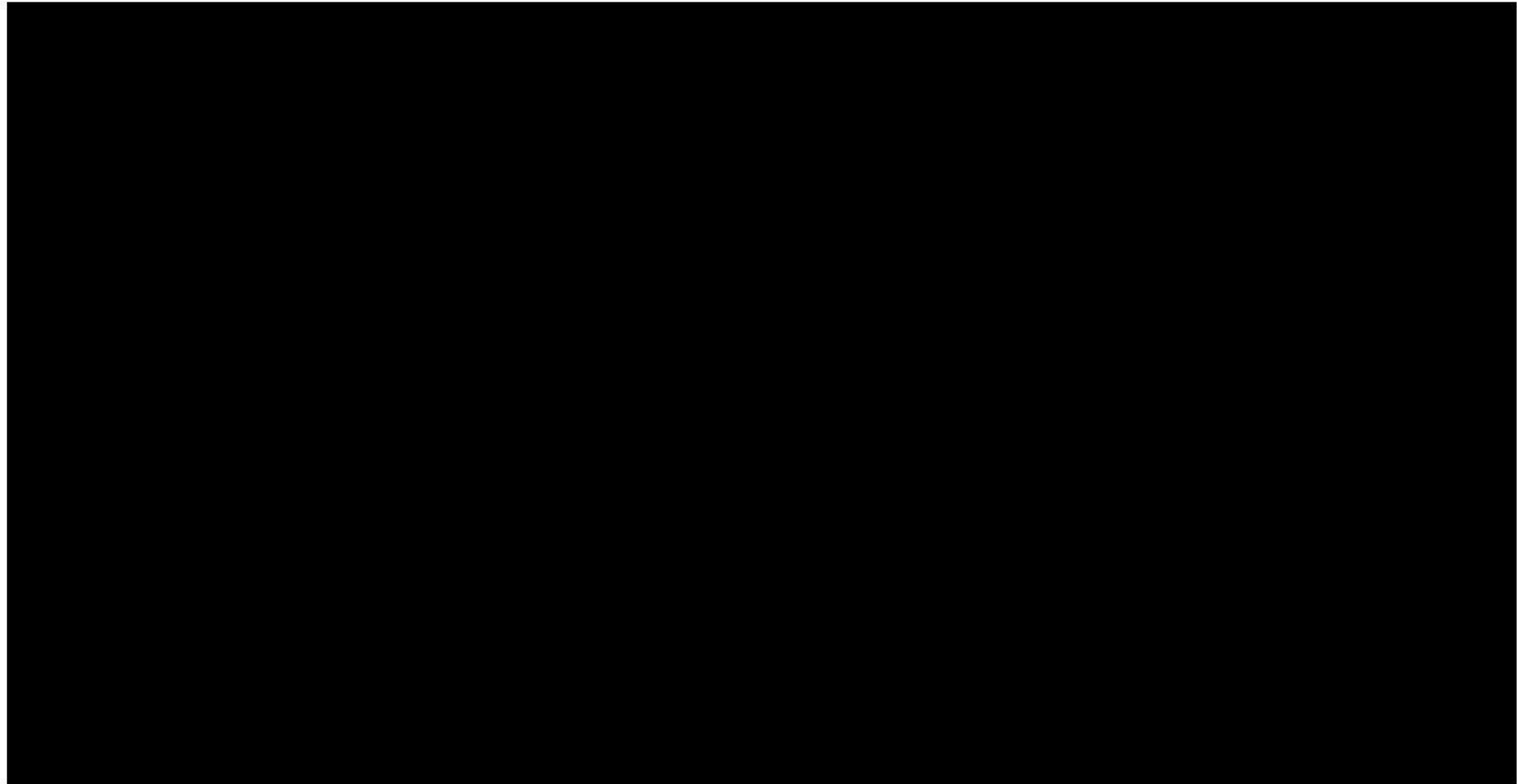
Appendix

Table A1. Brief Glossary of Terms

Term	Definition
Breed Association	The formal organization tasked with preservation and continuation of a breed society's genetic stock (material) and operations.
Breed Society	A group of breeders who own a specific type (breed) of animal. Includes the Breed Association as the management layer.
Breeder	Owner of livestock. Operates farming operations that produce genetic material, such as gametes, bulls, or heifers. Collects raw data as phenotypes or genotypes for internal operations and to facilitate the supply of insight from the Laboratory.
Laboratory, Analytic System	Partnering data enrichment service that turns Breeder's raw data into valuable insight.
Genetic Insight	Information on the total genetic value of an animal, usually expressed in terms of Estimated Breeding Values (EBVs), specific indices that rank animals against established baselines. Broadly considered a proxy for the worth of an animal to a breeder.
Genotypic Data	Genetic data pertaining to livestock and collected by a breeder. E.g., a tail hair from an animal.
Phenotypic Data	Measurement data pertaining to livestock and generated by a breeder. E.g., 100-day weight.

Table A2. Representative Data for Breeders Collecting Phenotypic or Genetic Data. Source: Banks (2022)

NB. Assumes an average herd size of 100 animals.

A large black rectangular area covering the majority of the page, indicating that the table content has been redacted.


Higher Degree Research Thesis by Publication

University of New England

Statement of Authors' Contribution


We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated in the *Statement of Originality*.

	Author's Name	Percentage of Contribution
Candidate	Matthew Wysel	80
Other Authors	Derek Baker	10
	Robert Banks	10



Candidate

Date



Principal Supervisor

Date



Higher Degree Research Thesis by Publication

University of New England

Statement of Originality

We, Matthew Wysel, the PhD candidate, and Derek Baker, the Principal Supervisor, certify that the following text, figures, and diagrams are the candidate's original work.

Author	Type of Work	Page Numbers
Matthew Wysel	Conceptualization, methodology, formal analysis, writing (original draft), writing (review), editing.	Entire Document
Derek Baker	Conceptualization, writing (review), editing, supervision.	Entire Document
Robert Banks	Conceptualization, writing (review), editing, provision of data.	Entire Document

	_____
<i>Candidate</i>	<i>Date</i>
	_____
Principal Supervisor	Date

Conclusion

1. The Problem of Creating Value from Data

The proliferation of data at a personal, organizational, and societal level has created widespread confusion, uncertainty, and inefficiency. This thesis proposes a straightforward and academically rigorous explanation for the process by which value is created from data within a firm, across a marketplace, and within a microeconomy. This research has built on the literatures of platform economics, the economics of information, production and innovation literature, as well agricultural datanomics to derive the components of the economic value of data, and how these components come together within agriculture and more broadly.

Unravelling the process of creating value from data is an important problem to solve because when data is surrounded by enriching technologies the resulting insight retains the near-infinite economies of scale of data (Arrow, 1996) but continues to confer increasing value to the data sharing systems that control access to it. This thesis explains how the simple absence of an understanding of 'how my data could be valuable' does not preclude other agents in *ad hoc* data sharing ecosystems from amassing value around that data and applying it to cooperative, competitive, and co-opetitive ends. This information asymmetry is replicated in data markets and across data microeconomies such as those that Breed Societies mediate. Indeed, without a clear understanding of how markets support the co-creation of value from data, uncertainty regarding the enduring quality of data goods will inhibit individual exchanges (Cichy et al., 2021), incentivize inefficient behavior in agents (Wagner et al., 2021), and ultimately suppress the size of data markets (Koohang et al., 2022).

The problem of creating value from data is that the solution cannot be simply abstracted from somewhere else. Firms do not have an answer (Short & Todd, 2017), neither do markets (Kenney & Zysman, 2016), nor individuals (Shin et al., 2022). Data producers and data analysts struggle to define a trustworthy process whether together (Windasari et al., 2021) or individually (Kotlarsky et al., 2023; Libert et al., 2014). We cannot simply appropriate an answer from management (Pentland et al., 2021), economics (Fleckenstein et al., 2023), or information systems (Angelopoulos et al., 2021). The development of an explanation for the mechanism whereby value is accrued to data necessarily requires an interdisciplinary approach as theories from each discipline and field of practice are analyzed, synthesized, and harmonized into a coordinated and coherent whole.

2. The Objectives

The first objective was the identification and explanation of the component parts of the value creating process that surrounds data and an explanation of how those components interacted.

Chapter 2 derived a decision framework that connected the three asset classes of data, a community of agents and a mobilizing system, together with the relevant management interactions and presented the model as an approachable three-layer Venn diagram.

The second objective was to explain the interactions of the mechanics of how value is created from data and to identify the economic drivers. Chapter 3 compiled a single economic expression for the creation of value from data that reflected the sustained engagement that agents and systems must maintain when collaborating to accrue value to data and strip uncertainty from it (Frankel & Kamenica, 2019). We reversed Shannon (1948)'s theory of communication in a noisy pipe to assemble an entropy-based explanation for the effect of the exchange of prompts and replies on the average entropy of a dataset.

We assembled these theories to explain how data could be valued in use as a resource, valued in exchange as a good, and utilized as a value accrual medium when valued as a currency. The third objective was to show how these processes operated within a firm and across a marketplace.

Chapter 4 projected those mechanics onto a firm-level production process, while Chapter 5 applied these concepts to the exchange of value across a marketplace. The former leveraged the non-rivalrous nature of data to explain how firms could simultaneously realize the value created from data as three different types of market-based payoffs and one internal value creation modality.

Chapter 5 empirically applied the consequence of the same properties of data to first derive and then investigate the service-dominant logic that the creation of value around data enables. When agents in an ecosystem adopt a service-dominant view of data both benefits, and their collective productive capacity is also expanded.

Finally, using a representative dataset, we assessed the implications of a partial impairment of a data sharing ecosystem. In this case, unilateral value appropriation by data producing breeders was causing a collapse in their market's allocation of value throughout their community. The immediate consequence was the rational adjustment of participants in the microeconomy to maximize the value each created from their own exchange of data. This created a disproportionately large marginal cost for any one agent to produce the premium phenotypic data they all relied on. By deliberately viewing Breed Societies as data sharing ecosystems, we showed that as the entropy contained in their shared data library continued to increase with each passing donation of genotypic data, the value of their data library first diminished, and then collapsed. Using Club Theory, Chapter 6

assessed three potential interventions and proposed a cost-neutral policy the Breed Society could use to repair and sustain the operation of their data sharing ecosystem.

3. Contributions and Extensions

The purpose of this thesis was to propose a unifying theory for the creation of value from data that could be abstracted into a variety of contexts of increasing economic complexity. We began by offering a data-centric, decision framework (Figure 2-3) that explicates the components in the value creation process, how they interact, and specifying what the characteristics of that process are.

Following Arrow (1996)'s exposition of the economics of information, the economic circulation that occurs across data sharing platforms was represented within a single economic expression (Equation (3-10)), while the mechanics of how the two asset classes with agency (the system, and the community of agents) interacted to create value from the third asset class (data) was explained in terms of the exchange of Shannon Entropy (Figure 3-3). This exchange informed the modes of valuing data (Chapter 3, Section 4, 5, 6) which were then identified within a firm in Chapter 4. More generally, the firm was represented as the compilation of the three asset classes but also permitted to decide how much, and how often it would generate value from shared data.

Chapter 5 lifted these functions from the firm level to permit the specialization that typically happens within data sharing ecosystems (Hartmann et al., 2016). Collected into two firms that embodied the data-analyzing system and the data-producing community, the persistent, self-reinforcing and co-opetitive nature of data sharing ecosystems was revealed in the way breeders organized the trade of data and managed themselves (Figure 5-2). The theory was applied to this real-world ecosystem to assess how the data sharing ecosystem might increase its productive output (Chapter 5, Section 5.3, 6.1). Finally, the causality was inverted to permit investigation of a partial impairment to one part of the data sharing ecosystem in Chapter 6. This theory forecasts a technological market failure that leads to a tragedy of the agricultural data commons, draws parallels with contemporary competitive agents in data sharing ecosystems (Chapter 5, Section 6.2; Chapter 6, Section 7) and informs policy makers what actions are required to avoid that failure.

Throughout this thesis, wherever theory is advanced it is kept at a very high level to make it readily applicable to a very broad set of use cases. However, we also repeatedly illustrate how the theory can be tested by specific real-world use-cases and how these specific examples may also be tested by the applied theory. For instance, on the one hand, the theory describes the general relationships within a production function or across a microeconomy while on the other, it may be applied to answer the very specific questions, 'how much should I invest in data?' or 'how often ought I repeat

that process?'. This is the first time such an expansive definition of data as a productive asset has been brought together with the very practical exercise of creating value from data.

The most topical extension of this work is the direct analysis of the effect of generative AI on societal knowledge assets (Chapter 5, Section 6.2; Chapter 6, Section 7). Analysis of such a use-case could proceed in almost precisely the same manner as Chapter 6. 'Thinkers', as producers of knowledge-assets, currently exhibit a large variety in their willingness to accept the cost of applying these knowledge-assets to derive knowledge-based benefits of varying levels of payoffs (Figure 6-2). The ensuing analysis could even group Thinkers into various groups according to education or explicitness of interactions with the engine (for example, *exclusively via third party apps* through to *paying members of the ecosystem*). As in Chapter 6, the long-term viability of this knowledge-based ecosystem relies on the ongoing willingness of Thinkers to apply their knowledge-assets *outside* the auspices of generative AI systems. However, the marginal cost of this strategy will become increasingly prohibitive. Therefore, the overall health of the knowledge-assets in the knowledge-based ecosystem will diminish unless there is either economic intervention or fracturing within the community (Chapter 6).

The corollary of this thesis is the specification of a value generation engine whose operation could be inverted so as to value the contributions of agents based on the change each instigates on the data contained within the ecosystem (Chapter 2). This extension would open several new avenues for research as agents in data sharing ecosystems could now be treated like individual stocks in a portfolio of assets. Where the system desires a reduction in entropy of data, differing messages of the form described in Figure 3-3 might be offered to an otherwise homogenous community of agents, whose response would confer value to the ecosystems data and also rank the agents. Communities of agents could be described according to systematic and unsystematic 'risk' which would be reflected in terms of correlation with established responses. Indeed, 'A/B testing' in online marketing is well-established and represents a rudimentary form of this *inverted value creation* where data values those *ad hoc* communities.

This dissertation began with an observation that organizations – and by extension, the world around them – was fast becoming 'awash with data', and that want for a valuation framework inhibited individuals, firms, and even economies from prioritizing the allocation of scarce resources to first establish, and then to assess, the efficiency and efficacy of a process that seemed as intractable as it was ubiquitous. Our hope is that this thesis lays the foundations for what will one day become a collection of commonplace tenets in a society that has learnt to integrate non-rivalrous exchange with value co-creation, data-based contribution with governance, and finally transparency with equity.

References

- Angelopoulos, S., Brown, M., McAuley, D., Merali, Y., Mortier, R., & Price, D. (2021). Stewardship of Personal Data on Social Networking Sites. *International Journal of Information Management*, 56, 102208.
- Arrow, K. J. (1996). The Economics of Information: An Exposition. *Empirica*, 23(2), 119-128.
- Cichy, P., Salge, T. O., & Kohli, R. (2021). Privacy Concerns and Data Sharing in the Internet of Things: Mixed Methods Evidence from Connected Cars. *MIS Quarterly*, 45(4).
- Fleckenstein, M., Obaidi, A., & Tryfona, N. (2023). A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model.
- Frankel, A., & Kamenica, E. (2019). Quantifying Information and Uncertainty. *American Economic Review*, 109(10), 3650-3680.
- Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing Value from Big Data—a Taxonomy of Data-Driven Business Models Used by Start-up Firms. *International Journal of Operations & Production Management*, 36(10), 1382-1406.
<https://www.emeraldinsight.com/doi/pdfplus/10.1108/IJOPM-02-2014-0098>
- Kenney, M., & Zysman, J. (2016). The Rise of the Platform Economy. *Issues in Science and Technology*, 32(3), 61-69. <Go to ISI>://WOS:000384706800030
- Koohang, A., Sargent, C. S., Nord, J. H., & Paliszkievicz, J. (2022). Internet of Things (Iot): From Awareness to Continued Use. *International Journal of Information Management*, 62, 102442.
- Kotlarsky, J., Rivard, S., & Oshri, I. (2023). Building a Reputation as a Business Partner in Information Technology Outsourcing. *The University of Auckland Business School Research Paper Series, Forthcoming, MIS Quarterly (Open Access). DOI, 10.*
- Libert, B., Wind, Y., & Fenley, M. (2014). What Airbnb, Uber, and Alibaba Have in Common. *Harvard Business Review*, 11(1), 1-9.
- Pentland, A., Lipton, A., & Hardjono, T. (2021). *Building the New Economy: Data as Capital*. MIT Press.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379-423.
- Shin, D., Kee, K. F., & Shin, E. Y. (2022). Algorithm Awareness: Why User Awareness Is Critical for Personal Privacy in the Adoption of Algorithmic Platforms? *International Journal of Information Management*, 65, 102494.
- Short, J., & Todd, S. (2017). What's Your Data Worth? *MIT Sloan Management Review*, 58(3), 17.
- Wagner, A., Wessels, N., Brakemeier, H., & Buxmann, P. (2021). Why Free Does Not Mean Fair: Investigating Users' Distributive Equity Perceptions of Data-Driven Services. *International Journal of Information Management*, 59, 102333.
- Windasari, N. A., Lin, F.-r., & Kato-Lin, Y.-C. (2021). Continued Use of Wearable Fitness Technology: A Value Co-Creation Perspective. *International Journal of Information Management*, 57, 102292.