# Valuing Data as an Asset

Laura Veldkamp*
Columbia Business School

2022

In the twenty-first century, the most valuable firms in the world are valued primarily for their data. This makes data central to finance. Data is an important asset to price, it changes firm valuation, and it is a key consideration for an entrepreneur starting a new firm. The rise of the data economy is changing sources of revenue and sources of risk (Chiou and Tucker, 2017; Goldfarb and Tucker, 2019; Lambrecht and Tucker, 2015). The industrial-age measurement and valuation tools commonly used in finance need updating for a new era. The goal of this article is to describe a set of tools to measure and value data and point to unanswered questions, where more work is needed.

Not only is data valuation central to most research areas of finance, but finance tools are essential for the study of the data economy. Data is digitized information. Information is something that reduces the uncertainty around a prediction. In other words, data resolves uncertainty or risk. If the primary benefit of data is to resolve risk, using tools for pricing risk, tools for allocating scarce resources in a risky environment and tools for choice under risk are central. These are the tools of finance. If we try to evaluate data, but ignore its ability to resolve risk, it is like trying to price assets, ignoring their risk premium. One would miss about two-thirds of the return of risky assets. Our errors in valuing data could easily be as large, unless we adopt risk-pricing tools, from finance.

The rise of data raises asset pricing questions because data is an asset that needs to be priced. Classic asset pricing tools are not appropriate for this new asset class.

*email: lv2405@columbia.edu. This article draws on ideas developed in co-authored work. I am grateful for my co-authors who constantly inspire and challenge me: Simona Abis, Juliana Begenau, Cindy Chung, Jan Eeckhout, Maryam Farboodi, Adrien Matray, Roxana Mihet, Thomas Philippon, Dhruv Singal and Venky Venkateswaran. Thanks to Junjun Quan for his research assistance.

One reason we need to update tools is that data has a large private value component. The value of a data set to one investor or firm is not the value to another. Because valuations for the same data asset differ by many orders of magnitude, computing a return or estimating a covariance with market returns has no clear meaning or implementation.

In entrepreneurship, the data economy offers new business models and new entry barriers. In the data economy, new goods and services are increasingly bartered for data. In such an environment, monetary revenue may not accurately reveal the value that the firm is generating or accumulating. As a result, new, data-intensive firms may earn negative profits. For example, Uber and Amazon both lost money for years. Still, these firms may be extremely valuable because of the data assets they are accumulating. Firms that do manage to accumulate and monetize their data can earn a dominant market position and use that position to extract monopoly rents. However, extrapolating current profits is unlikely to capture this future market advantage. Because of this new business model, questions about competition and entry barriers for new firms loom large. Old and large firms with long histories and large numbers of transactions have large data sets generated by the information from those transactions. This gives large firms a natural advantage and requires new strategies for new entrants to succeed.

In corporate finance, data assets raise the aforementioned questions about valuation, as well as new questions about how to discount future values for risk. If firms and investors use data to make more accurate predictions, then data not only raises profits, it also resolves risk. The risk resolution could be the greatest source of data's value. Sellers require compensation for risk, in the form of higher markups, but are not better off for the higher risk and higher revenue. Consumers facing higher prices are strictly worse off. Thus, risk is like a tax on the economy. If data can reduce the deadweight loss created by business risk, that could change the investment decisions firms make and the welfare of society.

This article will not resolve all of these questions. But all are examples of questions whose answers depend on the measurement or valuation of data. The sections that follow lay out a number of approaches to measurement, that are a starting point for many research agendas. Section 1 describes what data is and how it differs from other assets and concepts. Section 2 explores the supply side of data: how data is produced, accumulated and depreciated. It introduces a distinction between raw data, structured data, and knowledge. Section 3 turns to the demand side with tools to infer a firm or an investor's quantity and value of data. There is no one-size-fits-all measurement strategy. Instead, there are a variety of different approaches to measure and value data that may work, depending on the setting and the observable empirical evidence. Approaches to valuing data include a cost approach, a revenue approach, value function estimation, and using complementary inputs. Section 4 compares data as a way to enable better matching versus data as information. Section 5 concludes with ideas for future research.

# 1 What is Data?

Data is digitized information. Of course, digitized information is a broad category that includes things like poetry, NFTs and patents. The discussion about data assets and the data economy is about a particular kind of data. It is about big data sets, used for prediction. The new data technologies, AI and machine learning, that brought data to the forefront of modern debate, are prediction technologies. They use data to forecast outcomes. This ability to forecast is what makes data different from other assets and inputs. Data is used to forecast demand, forecast cost, forecast which types of customers are mostly likely to click on an ad. In finance, much data is used to forecast asset returns. Thus, the data we consider is digitized information for forecasting.

Such forecasting data is often a byproduct of economic activity. Transactions reveal what customers want, what price they are willing to pay and various other characteristics of the customer. Traffic patterns, tweets, browser histories all leave digital footprints that firms exploit for profit. Since big data technologies need large volumes of data and transactions generate such volumes, this is the primary source for most businesses' big data sets.

## 1.1 Contrasting Data, Technology, Learning-by-doing and Intangible Capital

Data has features in common with technologies and patents, with human capital acquired from learning-by-doing, and with other forms of intangible capital. Like patents, data can be bought and sold with property rights attached to the data. Like technologies that can be licensed to multiple parties, multiple copies of data may be sold to many parties. However, data production is quite different from the production of patents and technologies. Big data sets are not discovered in a lab. They do not require intensive labor to produce, although they may require specialized skills to analyze. Instead, most large data sets are large because they are footprints left by economic activity. While storage and analysis of data may be costly, its production is typically not. The fact that data is a byproduct of economic activity distinguishes it from technology, patents and other forms of intangible capital.

Learning-by-doing is a byproduct of economic activity. It describes a type of learning by workers who improve their productivity by repeating a task over time. However, learning-by-doing creates human capital. Human capital is stored in the mind of the worker. Each worker owns their own human capital and is typically compensated for the skill it represents. Data is owned by firms. It is priced and traded. Firms and shareholders are the beneficiaries of the rents to data. This difference in ownership and tradability is enormously important for data valuation and the valuation of firms that own the data.

## 1.2 Uses of Data

The forecasts from data are input into firms' productive activity. Firms use data to advance their objectives in four ways: improving business processes, reducing risk, growing market power and innovating.

Firms improve their business processes when they do things like: procure the right inputs, allocate investment efficiently, produce the right amounts, transport goods to where they are most needed and forecast what they will need next.

For example, a firm may use data to figure out if it should produce purple shirts or blue shirts today. If purple shirts are very popular,then that purple shirt is going to be worth more. The firm that forecasts this and switches to producing purple shirts will look more productive because they predicted the purple trend. Data could help the firm manage their inventory, decide what to put on the truck, and decide when to deliver the products to customers. Data can also help firms direct advertising to better-matched customers. For example, there are people who like the color purple, and there are people who like the color blue. If the firm advertises a purple shirt to the blue-liking person, that person will not buy the shirt or not pay as much as a purple-liking person. In both cases, the firm's advertising will not be effective, as if the firm had advertised the purple shirt to the purple-liking customer. Thus, better matching can show up as more efficiency or better quality.

Data also reduces risk. Data, at its core, is digitized information, and information is a technology used to reduce uncertainty. Data is not noise; it is not a random sequence of zeros and ones. Firms use prediction technologies such as machine learning and AI, with large amounts of data, for a variety of applications. Machine learning and AI can be used as inputs into inventions and can be used to raise returns; however, fundamentally, they are about prediction. For example, machine learning is useful in classification tasks. The goal of these tasks is to predict whether an observation belongs to this set or that set. Machine learning can be used in making better predictions about uncertain consumer demands for a variety of products, costs to make certain products, or returns to a portfolio of assets that one is going to buy. Thus, not only can data increase returns, but it can also decrease uncertainty.

Firms also gain from data because it creates market power for them. Firms with more data grow bigger, and bigger firms may be able to use more price discrimination, resulting in more pricing power. Firms may use large volumes of advertising to flood the market. This strategy is a form of generating revenue, but it may not be socially efficient. Firms may also use data to extract surplus from other firms. While all these methods generate value for a firm, the equilibrium and social welfare consequences may be quite different.

Finally, data can be an input into reserach and development of new products. Babina et al. (2022) argue that this is an important use of AI. Innovation will increase firm revenue, in much the same way as the improvement of business processes. But data-driven innovation may have different consequences for social welfare and long-run growth.

## 1.3 Data Measurement vs. Data Valuation

Given this notion of what data is, there is a question of what it means to quantify or measure data. This is not obvious because data has no agreed-upon units in finance. One natural way to measure units of data is bits. But some bits are much more relevant than others. Another data measure could be the additional precision such data offers in forecasting a random variable. This is similar to Blackwell's (Blackwell, 1951) notion of an information order. Finally, the units of data could be the monetary value that the data generates. In this case, more data is defined as a data set that produces more expected revenue, or perhaps, more risk-adjusted expected revenue. With this last definition of the data metric, measuring data and valuing data are the same exercise. Thus, we will proceed to talk about measuring and valuing data somewhat interchangeably, with the understanding that there is not always a clear distinction between the two.

# 2 Accumulating Data

Now that we have established what data is and how it differs from related economic concepts, the next question is: How is data generated and accumulated? This question is not purely academic, but will inform measurement as well. Some of the strategies for data measurement and valuation discussed in the following section make use of an understanding of the data production process.

Much of the big data that firms use is transaction data that is the byproduct of economic activity. Firms may collect browsing histories, search histories, or GPS locations from their customers. But this implies that selling a good or service to a customer generates data as a byproduct. Actions that generate information are often called *active experimentation*. Perhaps the most well-known active experimentation problem is the bandit problem, where a gambler is learning about the odds of various gambling machines and needs to decide which machine to pay money to play each round, to maximize the expected profit from his visit. In the data economy, a firm's problem is somewhat different from the bandit's problem because the firm does not pay to play. It earns money and accumulates information by selling. Nevertheless, because production decisions generate useful information, some of the same tools and ideas apply.

## 2.1 Data Barter

Many modern firms offer digital services to customers "for free." Examples include Facebook, Google searches and many phone apps. These services are offered for zero price. But are they really free? These services typically collect customer data. That customer data is a valuable asset. In a way, customers are paying for the search platform or their weather app, with their data.

This is a classic barter trade. The customer is bartering their data, at a monetary price of zero, in return for a digital service that is also valuable. In this context, measuring the value of data is challenging because there is no price observed on these transactions. Barter trades, such as these data transactions, are not included in GDP.

Not only are there pure barter trades, there are potentially many more partial barter trades, where a good or service is exchanged for a monetary payment and data. For example, Whole Foods offers customers a 5-10% discount on some groceries, if they scan a QR code that links their grocery purchase to their Amazon prime account. In other words, such customers pay for 90-95% of their grocery bill with money or credit; they pay 5-10% of the bill with their data. Explicit discounts for data are still fairly rare. However, less visible forms of partial data barter could be pervasive.

Consider a firm that is eager to grow its customer data. This firm should optimally lower the price of its goods, in order to attract more customers. More customers and more transactions generate more data. In this case, there is no explicit data discount. And yet, such a firm may well sell goods below the static profit-optimizing price. They may even optimally sell goods below their marginal cost, for the purpose of generating data that will provide future revenue. The difference between the low price, that includes the data transfer and the higher price that would be optimal if data were not a consideration, is the value of the data barter.

Understanding and measuring this implicit payment for data is crucially important for policy. Many claim that firms are not paying consumers for their data. It is possible that the value of data barter trades is very small and this is close to true. But the Whole Foods example suggests otherwise. 5% of the grocery bill is not small. Firms do not need to be altruistic to compensate consumers for data. Simple dynamic profit maximization suggests that, if data is a valuable asset, the implicit payment to consumers for data should be large.

## 2.2 The Data Feedback Loop

The data feedback loop refers to the increasing returns to data that arises naturally when firms produce data as a byproduct of economic activity. Suppose that having more transactions or getting more customers generates more data. Firms find out a wealth of information about their customers, such as what they like to buy, what kind of credit card they have, where they live, their zip code, and so forth. Firms use this data to generate higher-quality goods or better-matched goods for customers; they become more efficient. Firms may use data to appropriately stock their shelves and inventory or hire the right workers in order to be more profitable. Becoming more efficient or having higher quality goods allows a firm to attract more customers and do more transactions. Higher efficiency also incentivizes the firm to invest more and grow larger. Thus, a firm with more data has greater efficiency, more customers and gathers even more data. This increasing returns feedback loop is illustrated in Figure 1. It is at the heart of the promise and concerns about the data economy.

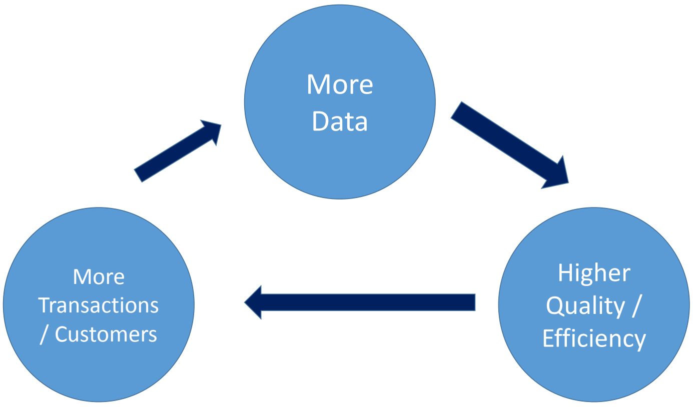


Figure 1: Data Feedback Loop

There are many ways to model or formalize this data feedback loop. A very simple one is a three-equation feedback loop.[1] Let transactions be represented by $Y_t$. The amount of data $D_t$ a firm generates today is given by their transactions today times a scaling factor $z$. The data they have tomorrow is the new data they've generated, plus a depreciated version of their data today:

$$D_{t+1} = (1-\delta)D_t + zY_t. \tag{1}$$

Here, data is assumed to depreciate at a rate $\delta$. This could be a linear rate. Or it could be the Bayesian-implied non-linear rate from the previous section. Data is used to make a firm more efficient or productive. The total factor productivity, $A_t$, is a function of the data the firm has:

$$A_t = \tilde{a}(D_t). \tag{2}$$

Finally, productivity then enters into the firm's production function and generates more output, in conjunction with another input, such as capital, $K_t$:

$$Y_t = A_t K_t^{\alpha}. \tag{3}$$

That closes the data feedback loop that generates increasing returns. The loop is that data $D_t$ increases productivity $A_t$; productivity increases output $Y_t$, which in turn, increases next period's data $D_{t+1}$.

This is a simple framework, that can be amended or extended to include more

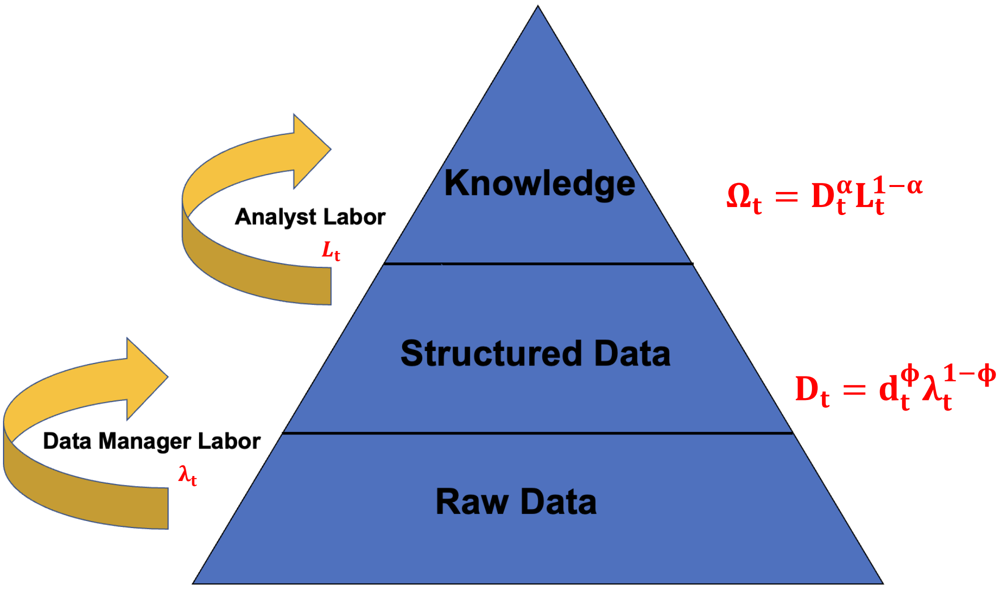
[1] This is a simplified version of Cong and Mayer (2022); Farboodi and Veldkamp (2022); Jones and Tonetti (2020)

inputs, a specific theory for how data maps into greater productivity, different forms of production functions with intermediate inputs, richer theories that link firm actions to stocks of useable data, data sales and purchases, a continuum of firms with equilibrium prices, or even a finite number of firms engaging in imperfect competition.

## 2.3 Raw Data, Structured Data and Knowledge

The data feedback loop assumes that transactions *directly* tell us something about the world that is useful. In practice, raw data does not immediately produce actionable insights. There is work that goes into taking raw data and turning it into structured data. Even if the data is already structured, there is work that goes into making it structured for a specific purpose. There is a class of workers called data managers that are involved in acquiring data, putting it into datasets, merging it with other data the firm has, maintaining the servers, and maintaining the relevant links to make sure the data updates properly. This process is represented in Figure 2. Because hiring is typically observable, measuring data-related labor will be an important clue in determining the value of data to a firm.

Figure 2: Knowledge Production Triangle



Even after the data is structured, the data still does not tell the firm what it should do. The firm may be interested in figuring out what to invest in, what color shirt to produce, or what different products to put on the truck. In order to answer these questions, the firm needs analysts. Analysts take highly structured data and make action recommendations. That is knowledge; knowledge is taking some information but using it to say: “and here is what we should do.” In sum, structured data is generated with a combination of raw data and labor input. Then, knowledge is generated by structured data and analyst labor.

## 2.4 Buying Data

Of course, firms can accumulate data by buying it. There are two different ways of selling data: direct and indirect. Financial information that is obtained by subscribing to Bloomberg or purchasing an analyst report are examples of direct data sales. The purchaser gets the data to use as they wish. In contrast, an indirect seller uses data to take actions of their clients' behalf. For example, a managed fund collects and analyzes data and uses that data to invest clients' funds. The idea is that better data results in more informed portfolio decisions and excess returns. Similarly, Google could sell the names and zip codes of the people who bought iPads – a direct information sale. More commonly, they place ads for their clients, using their information. In finance, Admati and Pfleiderer (1990) discusses indirect and direct sales of information. That idea is more relevant today, now that the sale of data is more widespread.

Regardless of how it is sold, a key characteristic of data is that data is non-rival. Consider a rival good, like a pencil. If Alice sells Bob a pencil, Bob now has the pencil. Alice cannot use that pencil; only one person can write with the pencil at a time. Data is different; Alice can sell data and keep the same data to use herself. Data contracts can make data rival: The buyer and seller may enter an exclusive-use contract, which prohibits the seller from using the data or selling it to multiple buyers. However, typically data suppliers sell data to many people. For example, many people can get subscriptions to Bloomberg terminals that provide access to the same financial data at the same time.

This raises the question: does sold data lose value, and by how much? This is an important consideration in imperfectly competitive markets. A way to model this loss is to assume that when a firm sells data, some of the data is lost. This loss can capture the loss of profits that arise from trading on information that everyone knows, versus information that only the data seller knows. For example, if a firm sells another firm 10 bits of data, the data seller is effectively losing a piece of the data. The firm is not actually losing data, but this data loss captures the notion that the data is less valuable, once it has been sold.

The law of motion for data presented in (3) can be extended to incorporate changes in data resulting from the buying and selling of data. As before, the data that the firm has tomorrow includes the depreciated past data and data the firm collects from its transactions. Two additional terms arise from data purchases and the loss from data sales.

$$\text{data}_{t+1} = (1-\delta)\text{data}_t + zK_t^{\alpha}L_t^{1-\alpha} + \underbrace{\gamma_t}_{\textit{data purchases}} - \underbrace{\iota\gamma_t \mathbb{1}_{\gamma<0}}_{\textit{loss from data sales}} \tag{4}$$

If the firm purchases data, the firm gets all the data it purchased. If the firm sells data, the firm loses a fraction of the data that it sold. Thus, the effective pirce per unit of data is higher when the firm sells than when it buys. This is negative bid-ask spread. This is similar to models with transaction costs and bid-ask spreads, but in reverse.

## 2.5 Depreciating Data

A key question for data valuation is: How quickly does data depreciate? Since big data is a forecasting technology, one should consider how forecasts depreciate. Bayes' law can inform this depreciation rate.

Suppose we use data, with normally-distributed noise, to forecast an AR(1) process, with normally-distributed innovations:

$$\theta_{t+1} = \rho\theta_t + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim N(0, \sigma_\epsilon^2). \tag{5}$$

This process could represent the return on an asset, or the demand for a good. The conditional variance of the state $\theta_t$, conditional on all information at time $t$ is $V[\theta_t|\mathcal{I}_t]$. This conditional variance is not a measure of the volatility of $\theta$. Rather, it is a measure of uncertainty. It is the expected squared forecast (or nowcast) error: $V[\theta_t|\mathcal{I}_t] := E[(\theta_t - E[\theta_t|\mathcal{I}_t])^2|\mathcal{I}_t]$. It reveals how inaccurate our forecasts are. Define $\Omega_t := V[\theta_t|\mathcal{I}_t]^{-1}$. As the inverse of inaccuracy, this represents the accuracy or precision of beliefs about the state at time $t$. We will refer to $\Omega_t$ as the stock of knowledge. A lower variance or more accurate estimate of the state means we have more knowledge about $\theta_t$.

Next, we can apply the conditional variance operator to the left and right sides of (5). This tells us the conditional variance of tomorrow's state, given today's information:

$$V[\theta_{t+1}|\mathcal{I}_t] = \rho^2\Omega_t^{-1} + \sigma_\epsilon^2. \tag{6}$$

In Bayesian language, this is a prior variance.

If data is used to forecast $\theta_{t+1}$, then Bayes' law says that we can combine that data and represent it as a signal about tomorrow's state $s_t = \theta_{t+1} + e_{st}$, with $e_{st} \sim N(0, \sigma_s^2)$.

When we combine a normal prior belief with a signal that has normally-distributed signal noise, Bayes' law says that the precision of the resulting posterior belief is the prior precision (inverse of eq. 6) plus the signal precision $\sigma_s^{-2}$:

$$\Omega_{t+1} = (\rho^2\Omega_t^{-1} + \sigma_\epsilon^2)^{-1} + \sigma_s^{-2}. \tag{7}$$

This equation maps time-$t$ stock of knowledge $\Omega_t$, in to time-$t+1$ stock of knowledge. In other words, it is a law of motion for the stock of knowledge. That law of motion says that we take the stock $\Omega_t$, depreciate it by transforming it into $(\rho^2\Omega_t^{-1}+\sigma_\epsilon^2)^{-1}$ and then add on new data, being added to the data set, with precision $\sigma_s^{-2}$. This is similar to a law of motion for a stock of capital: $k_{t+1} = (1-\delta)k_t + i_t$, where $i$ is new investment. But for data, the depreciation rate is

$$\delta = 1 - (\rho^2 + \sigma_\epsilon^2\Omega_t)^{-1}. \tag{8}$$

This depreciation rate teaches us that if the AR(1) process is highly volatile (high $\sigma_\epsilon$), then the stock of knowledge will depreciate quickly. Data about yesterday's state is less relevant to today's state because the state is changing quickly. Also, we learn that large stocks of knowledge depreciate at a faster rate than small stocks. However, in many cases, the depreciation function can be close to linear. That depends on how volatile and persistent the environment is.

This depreciation rate is measurable. It requires measuring the persistence and volatility of the object the firm is trying to forecast. We can use those estimates to create a depreciation rate for data. The depreciation rate will be context-specific. For example, data about order flow, which is highly volatile, is going to have a different depreciation rate than data about customer zip codes, which persist for years. To depreciate data, we need to know what the data will be used for. But once the persistence, volatility and data stock are known, Bayes' law can do the rest.

# 3 Measuring and Valuing Data

Understanding how firms accumulate data can help researchers make inferences about how much data firms have, and how valuable that data is. In the following section, the amount and value of data are used interchangeably. It is natural to equate these two notions of data; because data does not usually have natural units, not all data are equal, and thus, measuring the amount of data inherently requires a notion of value.

## 3.1 Cost Approach

A typical approach to valuing many assets for which transaction prices are not available is to assume that the value of the asset is the cost of its production. If the asset has traded, we would often value it at the transaction price. This approach can be applied to value traded data. Firms may purchase data at a cost; the cost of that data can be used as its value or amount.

However, lots of data is data that a firm acquired about its own customers, through transactions. This data was not purchased. There is no transaction price for this data. The value of the good or service that was sold is surely the not the same as the value of the data about that transaction. The problem is that there is also no clear cost of production for this data. This data is a byproduct of economic activity. Standard GAAP accounting rules would assign such data a book value of zero.

This lack of a clear cost is a big problem for the valuation of some of the most valuable firms in our economy. Amazon's user data and shopping history, Google's data on internet users' search histories, these are incredibly valuable assets. These firms are monetizing these assets by selling targeted advertising, among other services. Yet, the data assets themselves are typically treated as though they have zero value. The skilled labor that is hired to maintain these data assets looks like a pure expense, from a balance sheet perspective. Measures of economic activity that count production miss the value created in producing data assets.

However, there is a potential solution that could enable a cost approach to provide insight. That solution draws on the idea of data barter, introduced in Section 2. If a firm wants to grow its data set, it needs to attract customers and do more transactions. The firm does this by lowering its price. The lower price is a form of payment that the firm is giving its customers for their data. In other words, many firms are paying for data, in the form of discounts, explicit as in Whole Foods, or implicit. If one can measure this data discount, one could adopt a cost approach to measure the value of data.

The key would be to find instances where a firm charged a customer more because the customer did not provide the firm with their data. This might take the form of price changes after privacy laws are introduced or differential pricing across state or national borders with different degrees of data protection. But measuring these differences in pricing could provide us with knowledge of how much the production of data effectively costs a firm. Armed with this knowledge, the cost approach becomes a more useful tool to price data.

## 3.2 Choice Covariance

Another approach to data measurement is to measure covariances between a firm's choices and payoffs.

Data allows agents to take better actions. We call a firm's action $q_t$, which we interpret here as a quantity. It can also be a price or any other action. The quantity might covary with the payoff, which is called $r_t$. The expected profit a firm gets is equal to the expected quantity times the expected return plus the covariance between quantities and returns.

$$E[q_t r_t] = E[q_t]E[r_t] + cov(q_t, r_t) \tag{9}$$

If a firm has data that predicts $r_t$, the firm can choose the $q_t$ that covaries with $r_t$. If the firm does not know anything about $r_t$, it is not possible for the firm to choose $q_t$ to covary systemically with $r_t$. Data informs the covariance between quantities and payoffs. That is why firms value data, because it allows firms to take actions that covary with their payoffs. There will be instances in which covariances are measurable, and can be used as the value of data.

**Choice covariance in financial markets.** This idea can be applied in a finance context to portfolio choice. An investor may choose assets that systematically have high returns relative to a benchmark. A portfolio that has high returns, relative to the return of a benchmark portfolio, is called a high-alpha portfolio. For a long time, finance researchers considered the portfolio alpha to be a measure of manager skill.

But it is also a measure of the precision of the information that the manager has (Kacperczyk et al., 2016). A portfolio alpha is a measure of the covariance of the investor's portfolio choice with the realized return on the investor's assets.

A similar idea allows Bai et al. (2016) to measure the amount of price information in equity markets. They find that price informativeness is increasing and interpret this as showing that investors, in the aggregate, have more information or data that they trade on. Dávila and Parlatore (2018) and Farboodi et al. (2022a) adjust price informativeness for asset characteristics, in order to isolate changes in information from changes in asset characteristics. Farboodi et al. (2022a) find that while there is evidence of rising amounts of investor data, this data is allocated unevenly across assets. Most investor data is being used to trade large, growth stocks. For other types of assets, there is no evidence of the growing use of data. But all of these findings are premised on the idea that data enables investors to buy assets that will subsequently have higher returns. It enables a higher investment-return covariance. These approaches are simply using properties of market prices to detect this covariance.

**Choice covariance for non-financial firms.** A covariance approach could also be used for non-financial firms. Consider a firm using data to try to figure out which product to produce. If the firm wants to maximize their profits, the firm should produce high-demand (or low-cost) product. These are products that are highly profitable. Data may be used to forecast which products will be in high demand. Then, a firm should produce more of these high-profit products. Only firms that can have data to predict demand well can execute this strategy. Thus, the covariance of the firm's production and the per-unit price or profit margin is a measure of the firm's data (Eeckhout and Veldkamp, 2022).

In marketing, one can measure the covariance between advertising revenue and customer click-through. All of these covariances should tell us something about the underlying information that was used to make that decision.

Adopting a choice covariance approach may not always be possible, because firm actions or objective might be unclear or unobservable. If we think data has a clear purpose for the firm, and if both that objective and the firm's action are observable, then covariances are an important piece of evidence about the amount of data a firm has.

## 3.3 Revenue Approach

The revenue approach can value data, when we can observe or model how a firm profits from data.

The value of the data should be the present discounted value of the revenue it generates, adjusted for risk. How do we isolate data revenue from other revenue? This is the key challenge. In many cases, it may be clear; in other cases, data may be used for multiple purposes and separating data revenue may be difficult.

For young data-intensive firms, simply extrapolating or linearly forecasting revenue could be very misleading. If a firm needs a lot of customers to get more data, and they need to get more data to operate profitably and be more efficient, their main goal early on should be to get as many customers as possible, at whatever cost. In fact, Amazon was unprofitable for the first 20 quarters of its existence. That makes a lot of sense for a firm in the data feedback loop. The optimal path for the firm may involve pricing below costs early on in the life of the firm, assuming the firm is not so financially constrained they are unable to do that. Firms want to initially price below costs because its a form of costly investment in data and in transactions that will generate the data.

What is the value of this data for the firm, given that it is making a loss, quarter after quarter? It is possible to value this data with a clear idea of how data generates the revenue. Using a theory is necessary here because a counterfactual is required to value this data. How much would this firm be worth, or how much revenue would it be generating, if it did not have the data? We do not have data from the alternative world, in which the firm does not have data. Models are necessary to answer those kinds of what-if questions. We will discuss one example in one particular example of valuing data, when used for trading risky assets, using a revenue approach.

**Valuing financial data with a revenue approach.** This example is based on Farboodi et al. (2022b).[2] In this context, investors use data to purchase a portfolio of risky assets whose payoff is normally distributed. Investors have concave objectives. The solution uses a second-order approximation to the investor's utility function. After substituting in the optimal portfolio for every investor, setting the equilibrium price of all assets to clear the asset market, and taking expectations over prices and the unknown future value of data that we are valuing, expected utility takes the form:

[2] This is a simplified version of Farboodi et al. (2022b) where the price impact of an investor is zero and the investor's investment strategy does not limit the set of investable assets.

$$\begin{aligned}\text{Value of data} =& \frac{1}{r\rho_i} \underbrace{\mathbb{E}\left[R_t\right]' (\mathbb{V}\left[R_t \mid \mathcal{I}_{it}\right]^{-1} - \mathbb{V}\left[R_t\right]^{-1}) \mathbb{E}\left[R_t\right]}_{\textit{Squared Sharpe Ratio}} \\ &+ \frac{1}{r\rho_i} \mathrm{Tr}\left[\underbrace{\mathbb{V}\left[R_t\right] \mathbb{V}\left[R_t \mid \mathcal{I}_{it}\right]^{-1} - I}_{\textit{Risk Reduction}}\right]\end{aligned} \tag{10}$$

The value of data depends on expected returns, the variance of returns, and the conditional variance of returns.[3] The value of data here is not the transaction price of data, and it is a personal value of data, which depends on the investor's information and absolute risk aversion. The value of an investor's data might depend on how much the investor moves the price when the investor trades. If the market is not perfectly competitive, the expression needs to be modified to incorporate investors' price impact. The value of data to a particular investor is not the same as the price the data sells at. The transaction price depends on the intersection point between the demand and supply curve. The value of data is one investor's point on the demand curve.

The value of data, if we have $\mathcal{I}_{it}$, can be computed by figuring out the mean, the variance, and the conditional variance of profits. This approach is a step forward because previous work with this class of models solves for all the equilibrium objects. What should be the expected profit? That depends on everybody's risk aversion, their wealth, or what their expectations are, and can be mapped down to structural parameters of the model. If one is not interested in solving the model in terms of its structural parameters, it can be written in terms of a small number of sufficient statistics.

How can we estimate the value of data? Means and variances are easy-to-measure sufficient statistics. Conditional variance is the key challenge. Conditional variance measures how variable is a return conditional on what an investor knows. That is an expected squared forecast error. A linear normal Bayesian forecast is the same as an

[3] This may initially seem to not make sense because the value of information should be different if it is public or it is private. How many other investors know the information seems to be missing from the expression for the value of data. How can the value of financial data not incorporate who else knows the data? But it is in the value of data. If many investors know the information in a piece of data, then conditioning on that data should not forecast returns. If everybody knows that Tesla is going to lose value and have much lower than expected earnings, that information should be fully impounded in today's price. That piece of information should also affect tomorrow's price or dividend. The information raises the price and payoff, the numerator and the denominator of returns, and it should not correlate with returns. The data should not forecast returns. Thus, the extent to which other people know a piece of information will affect the forecastability of returns. Data that others know will not reduce the conditional variance of returns because it does not reduce the forecast error.

OLS forecast. They are both efficient linear estimators. Conditional variance is the same as an average squared residual from an OLS forecast. First, forecast returns without data and calculate the forecast errors. Second, forecast returns again using a historical sample of the dataset of interest, and calculate the forecast errors. Take the expected squared forecast errors from the two steps, and use those in the formula for the value of data in equation 10. This gives the value of the risk-adjusted return obtained from data when trading a portfolio.

This approach is valuable because it allows us to measure the value of data to a particular investor without knowing the characteristics or information of other investors in the market. The econometrician needs to only know about the investor for whom the data is being valued.

**Data as a private value asset.** The main finding from this estimation exercise is that most of the variation in the value of data comes from investor characteristics, such as wealth, the existing information that they have, and their frequency of trade. These characteristics affect how much they value data. The value of data can be estimated at a quarterly, daily, second, or even microsecond frequency. The value of data depends on price impact. Note the expression for the value of data above would have to be adjusted to take into account the effects of price impact. Investors can place very different prices on the same data asset because of very small heterogeneity in what they know, how wealthy they are, or what they intend to do with the data. That is an important result because financial assets are typically thought of as common-value assets. Alice's value for a share of GM is the same as Bob's value for a share of GM. Money is money, and all investors like it and in the same way. Investors may have different marginal utilities, but it is a common-value asset. On the other hand, data is not a common-value asset. An investor's value for data depends enormously on what they are going to do with it. A researcher may require the use of financial intermediation data, and that data is very valuable to them. Another researcher may not need it at all, and place no value on the data.

This heterogeneity in private valuations is important because this means that small changes in the price do not pick up many more customers on the margin – there is a high price elasticity of demand for data. Much discussion is focused on inelastic demand in financial markets, but data markets are also very inelastic as well because valuations for data are so different. Furthermore, we show how inelastic demand and price impact in financial markets can affect the price elasticity in data markets as well. Studying the elasticity of demand for data is important for data markets.

## 3.4 Value Function Approach

The value function approach uses the same kind of tools to value data that macroeconomists use to value capital.

The value function is a recursive equation, or Bellman equation, that maps the amount of a state variable – data in this case – into the present discounted value of future revenues of a firm. The value of data is the gross revenue a firm produces with that data, minus its costs, plus a discounted value of the data the firm will have in the next period. Farboodi and Veldkamp (2022) assume that firms produce with labor $L$ at cost $w$, capital $K$, which is rented at rate $r$, and data contributes to the firm's productivity $A$. In that case, the value function or Bellman equation for data can be expressed as:

$$v(\text{data}_t) = max_{K,L} A(\text{data}_t) K_t^{\alpha} L_t^{1-\alpha} - wL_t - rK_t + \beta v(\text{data}_{t+1}) \tag{11}$$

This value function represents an economy like the one in Section 2 on the data feedback loop. Productivity, from data, multiplies capital and labor. The discount factor here is $\beta$, which is contant in this example. It can be modified to be a stochastic discount factor. This expression can be adapted and enriched with more inputs in production, equilibrium conditions to determine the price of labor and capital, a more sophisticated mapping between data and productivity, or additional choice variables.

A theory of data inflows is required here. The law of motion for data could be analogous to equation 4, from the data feedback loop. Tomorrow's data is today's data depreciated plus some fraction of the transactions. Data purchases and sales can be added, and we can think about using labor inputs to process raw data into structured data, and structured data into knowledge. There are various ways to augment this with theories of how firms accumulate data.

The estimation procedure would be to use aggregate data to calibrate or structurally estimate the parameters $\alpha$, $w$, $r$, parameters of the productivity function $A(\cdot)$, and parameters of the data evolution equation, including data depreciation. Estimating these parameters typically involves solving (11) numerically, often with functional approximation tools like splines, grids or polynomials.

## 3.5 Complementary Inputs

The next approach is measuring data with complementary inputs. Suppose knowledge $K_{it}$ is produced using structured data and analyst data.

$$K_{it} = A_t a_i {D_{it}}^{\boldsymbol{\alpha}} {L_{it}}^{1-\boldsymbol{\alpha}}, \tag{12}$$

This equation represents the process of taking structured data and using it to make action recommendations at the top of the Knowledge Production Triangle in 2. Knowledge is the structured data and labor input multiplied by a firm-specific component to productivity, and an aggregate time-specific component. The time-specific component could arise because new technologies are invented or machine learning techniques improve over time. The evolution of structured data follows

$$D_{i,t+1} = (1-\delta)D_{it} + \lambda_{it}^{1-\phi} \tag{13}$$

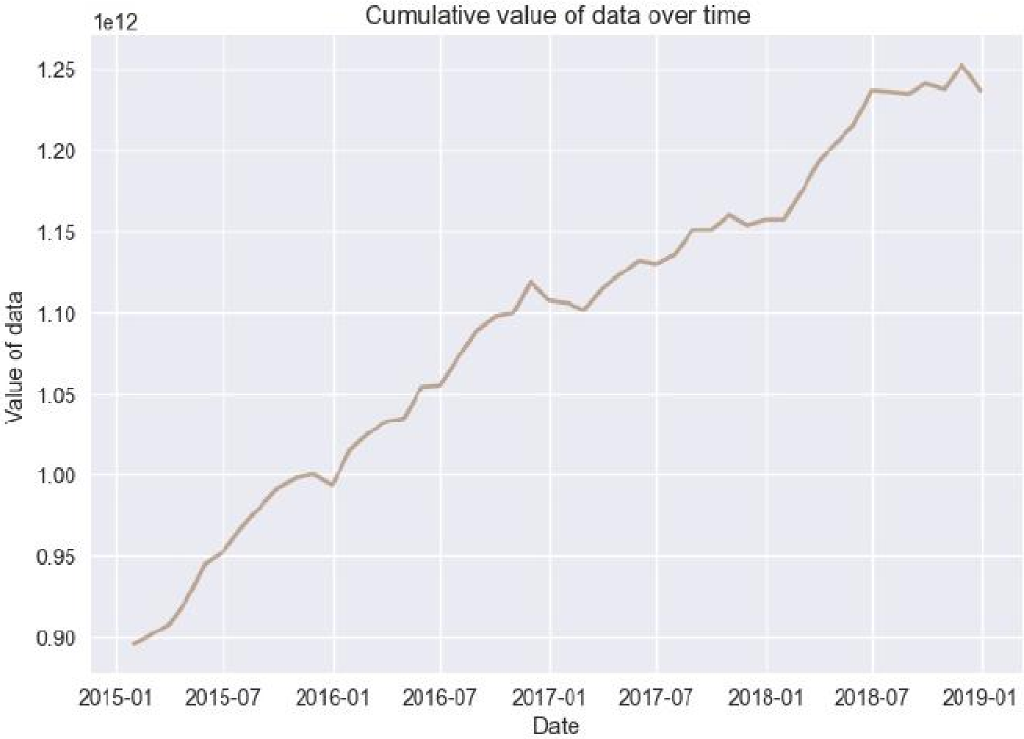
New structured data is added to the existing stock of structured data with data management labor. The existing stock of structured data depreciates, linearly.

A firm's stock of data can be estimated from measuring the hiring and the wages of these data managers, who deal with raw data, and analysts, who deal with structured data. With a structural model, if we know how many data managers and analysts the firm hires, and how much each group is paid, then we can make inferences about the extent to which there are diminishing returns to data, and how much a firm values their data. If a firm does not value their data very much, they would not hire many data managers or analysts or pay them very much. If the firm values that data a lot, then the firm would do hiring of workers who work with this data in various ways. In Abis and Veldkamp (2022), we impute the value of data for different types of firms and estimate how production functions for knowledge have changed and how different they are with and without machine learning.

Another observable complementary input is IT capital. Bresnahan et al. (2002) have studied IT capital, but the authors do not use structural estimation in their paper. If we can write down structural models, we can use any complementary input that might be used with the data. For example, if programmers consume jelly beans, then jelly beans are a complementary input to coding.

We find that the estimated value of data for firms doing financial analysis has been rising enormously over time; it has grown by more than 25% in just four years.

Figure 3: Estimated Value of the Aggregate Stock of Data, in hundreds of billions of current U.S. dollars, 2015-2018 (Abis and Veldkamp, 2022).



There are three reasons why this value is growing. First, firms are getting more data. Firms are accumulating and purchasing more data; their data managers are adding it to their stocks of data. Second, firms are hiring more workers. Many people are concerned that data is labor-replacing. In this context, data is not labor-replacing, because workers, analysts, and data managers are complementary to their data. The data shows that there is a lot more hiring. Data may still replace workers in certain contexts, but even the firms that are adopting AI and machine learning are hiring more workers. This makes each data point more valuable because if the firm pairs the data point with more labor, the marginal value of the data point is higher. Lastly, firms are becoming more productive at using new technologies like machine learning and AI. There is a positive time trend for a given amount of workers and data and the imputed valuation for that dataset. All three of these things are pushing up the value of the data at an amazing rate.

## 3.6 Intangibles Approach

Data is an example of an intangible asset, like patents, goodwill, or customer capital. Some of these items may be conflated with data. A typical approach to valuing intangible assets is to use the difference between the market value and book value of a firm. Data rarely appears on a firm's balance sheet, unless the data has been purchased from another entity. If a firm acquires a target firm, a firm may attribute some of the value of the target firm to the data the target firm owns; this may show up on the firm's balance sheet. If the data is generated internally, it cannot be listed as an asset. However, investors in the firm should know something about the firm's data and its ability to monetize that data, to produce a future revenue stream. In other words, the market value should include the value of data. Using the intangibles approach, we might argue that the difference between the market value and book value of the firm is the data value.

The difference between market and book values has been used to proxy the value of many different intangible assets; this same quantity has been called the value of the firm's branding, patents, or organizational capital. Data may contribute to each of these intangibles, but it is not equivalent to any of them. Distinguishing the value of data from the value of other intangible assets is not easy and probably requires other supporting evidence. Finally, this approach also assumes that equity market participants know precisely how to value data, which is unlikely.

# 4 Signals vs. Matching and the Importance of Risk

One of the more popular alternative approaches to model data is to model data as enabling better, more directed matches (Mihet and Philippon, 2019). Data allows customers to find and access products that were otherwise unavailable to them. The products were not really unavailable, but they may not have shown up in a customer's search, and the customer may not have known that they existed. Data can bring those products to the customer's attention and change the choice set of the customer.

The idea of data as enabling better matches has an analog in the finance literature. This idea looks like the recognition friction that was proposed by Merton (Merton, 1987) many decades ago. The recognition hypothesis argued investors do not know that many assets exist. An investors that has not heard of Tesla will not buy Tesla. Merton's hypothesis emerged prior to the rise of index funds, through which investors could own many firms that they had never heard of.

At the heart of recognition is an information friction –investors did not know that an asset existed. Data enlarges the recognized set of assets. An alternative way of representing an information friction is to assume that data is noisy information that customers or firms use to update forecasts and make more profitable choices. Under this alternative notion, a customer may know all the financial products that exist. They are simply uncertain about the quality or return of those products. Perhaps the financial product is fraudulent. In the consumer goods space, this uncertainty might take the form of concern that a shirt will fall apart on the first wash. In either case, because of this concern, because of uncertainty, the customer may not buy the product, without more data about it.

These two ways of approaching data and its relationship to demand and actions have a lot of similarities. Both notions of data improve match quality. The more a customer knows about all the products that are out there, the more likely a customer is going to get the one that is the best match, the best offer, or the asset with the highest return. Data as information and data as access both increase mean revenue. They both boost market power by making the high-data good less substitutable with others.

The key difference is that noisy signals also resolve risk. One thing we rarely see in matching settings is that matching changes uncertainty. Typical matching models features agents that can or cannot choose an option. Information is all or nothing. Risk is the in-between. Risk creates an inefficiency wedge in every transaction that hurts both parties. There is no upside to risk. It is a downside for the customer, having to bear it, and it is a downside for the firm that gets less revenue from it.

**Quantifying the risk-reduction value.** Finance moved away from data as facilitating recognition and towards data as noisy signals decades ago because risk matters. Risk matters more than twice as much as a riskless return for firm values. Of the 10% expected returns on firms, about 3% is the riskless return and 7% is the risk premium, the compensation for risk. For financial data, much of the value of financial data comes from uncertainty reduction. Farboodi et al. (2022b) uses the expected utility formula (10) to compute the value of data to various investors, with different characteristics. They break out the part of data value that comes from increasing the expected return and the part that comes from reducing uncertainty. Expected return accounts for, at most 60% of the data value. In much cases, far less than half of the value comes from a higher expected return. In most cases, the majority of data's value comes from its ability to resolve risk, to make forecasts less uncertain.

Risk also matters for firm decisions. Firms price risk and scale back investment, in the face of risk. Thus firms making real output decisions may value data for its risk-reduction properties as well. Neglecting the risk component of data's value could lead to a substantial under-valuation of non-financial data as well.

# 5 Conclusion

Data is one of the most important and highly-valued assets in the modern economy. Data is difficult to observe, measure, and put a price on. In order to value data, many different approaches are necessary. This article offered ideas, but is by no means exhaustive. Many other approaches could take hold. However, theory needs to inform the measurement. In part, theory is needed to help make inference about an asset that is difficult to observe. In part, theory is needed because the policy questions are pressing and require frameworks in which we can perform experiments with regimes that have not yet produced empirical evidence. Finally, theory is needed to interpret the measures we see.

This paper has been about the private value of data. How much is it worth to a firm, an investor or an owner? The social value is also important to quantify. There are many questions about optimal data regulation. The social value of data may be quite different from the private value because of data externalities such as privacy, competitive effects or coordination motives.

Future research could ask questions such as: Are tech firms overvalued? Does data as an asset have a factor structure? How large are the efficiency losses from various sorts of privacy protections? A combination of theory and measurement is needed to tackle these important questions and many others.

# References

Abis, S. and Veldkamp, L. (2022). The Changing Economics of Knowledge Production. *Available at SSRN 3570130.*

Admati, A. and Pfleiderer, P. (1990). Direct and Indirect Sale of Information. *Econometrica*, 58(4):901–28.

Babina, T., Fedyk, A., He, A., and Hodson, J. (2022). Artificial Intelligence, Firm Growth, and Product Innovation. Columbia University Working Papers.

Bai, J., Philippon, T., and Savov, A. (2016). Have financial markets become more informative? *Journal of Financial Economics*, 122(3):625–654.

Blackwell, D. (1951). Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. Publisher: University of California Press, Berkeley.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The quarterly journal of economics*, 117(1):339–376. Publisher: MIT Press.

Chiou, L. and Tucker, C. (2017). Search engines and data retention: Implications for privacy and antitrust. Technical report, National Bureau of Economic Research.

Cong, L. W. and Mayer, S. (2022). Antitrust and user union in the era of digital platforms and big data. *Available at SSRN.*

Dávila, E. and Parlatore, C. (2018). Identifying price informativeness. Technical report, National Bureau of Economic Research.

Eeckhout, J. and Veldkamp, L. (2022). Data and Market Power. Working Paper 30022, National Bureau of Economic Research. Series: Working Paper Series.

Farboodi, M., Matray, A., Veldkamp, L., and Venkateswaran, V. (2022a). Where has all the data gone? *The Review of Financial Studies*, 35(7):3101–3138. Publisher: Oxford University Press.

Farboodi, M., Singal, D., Veldkamp, L., and Venkateswaran, V. (2022b). Valuing financial data. Technical report, National Bureau of Economic Research.

Farboodi, M. and Veldkamp, L. (2022). A Model of the Data Economy. Technical Report Working Paper 28427, National Bureau of Economic Research.

Goldfarb, A. and Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1):3–43.

Jones, C. I. and Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9):2819–58.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626. Publisher: Wiley Online Library.

Lambrecht, A. and Tucker, C. E. (2015). Can big data protect a firm from competition? *Available at SSRN 2705530*.

Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance*, 42(3):483–510.

Mihet, R. and Philippon, T. (2019). The economics of big data and artificial intelligence. 20.