DATA SALES AND DATA DILUTION*

Ernest Liu† Song Ma‡ Laura Veldkamp§
Princeton & NBER Yale & NBER Columbia & NBER

Abstract

We explore indicators of market power in a data market. Markups cannot measure competition, because most data products' marginal cost is zero, making the markup infinite. Yet, data monopolists may not exert monopoly power because they cannot commit to restricting data sales to future customers. This limited commitment and strategic substitutability of data undermine sellers' monopoly power. But data subscriptions restore this monopoly power. Evidence from online data markets supports the model's insight that subscriptions indicate market power. Model and evidence reveal that data subscriptions are better for consumers because they sustain the incentive to invest in high-quality data.

Keywords. Market Power. Data markets. Data economy. Technological Change. Market Structure.

JEL. C6. D4. D5. L1.

*We thank Toni Whited (the Editor), an anonymous referee for feedback that significantly improved our paper. We thank Christopher Tonetti, Itay Goldstein, and Liyan Yang for discussing our paper, and Anton Korinek and participants at the FRA Conference, Econometric Society Meeting, and GSU-RFS FinTech Conference for helpful comments. We thank the Brookings Institution for financial support of this project.

†ernestliu@princeton.edu
‡song.ma@yale.edu
§lv2405@columbia.edu

One of the largest concerns that economists and policymakers have about the new digital economy is the market power of firms that sell data. The fact that data has a large fixed cost component and is free to replicate suggests the emergence of natural monopolies. However, little is known about how this market functions and prices data. We use theory to understand what indicators of market power to look for and collect new empirical evidence on data marketplaces to measure that market power and its welfare consequences.

It is not obvious how to identify market power in a market where every seller has a monopoly over their data set and where the marginal cost of producing additional data copies is zero. Therefore our empirical exploration of data markets needs to be guided by theory. We build a dynamic model of a monopolist data seller with two key features: Information that others know generates less value and sellers cannot commit not to sell more data in the future. These are realistic features of a data market. The first, commonly called the "strategic substitutability" of information (S. J. Grossman and J. E. Stiglitz 1980), arises in many settings where users of data choose quantities and market clearing determines the price. The second assumption, a classic commitment friction, is particularly relevant in data markets where one could easily transform data to make it non-identical, but functionally equivalent.

Our model teaches us that, even a data monopolist may have limited power to extract rent from their customers when the data seller cannot commit to a price schedule. The reason is that a data seller competes with its future self. If a data seller cannot commit not to sell the data to a firm's competitors, the firm's willingness to pay for the data declines. This force keeps data prices low. In this type of environment, we should worry less about the excessive profits of data monopolies and worry more about whether data is underprovided. Since we observe that many data producers sell data subscriptions, rather than data ownership, we add that feature to our model. We find that subscriptions for data allow a firm to re-capture much of its lost revenue from the lack of commitment.

But if data subscriptions allow firms to capture more surplus from consumers, why don't all firms sell data subscriptions? We use the model to identify three features of a data seller firm that make subscriptions less attractive to it: financial frictions, a small market, and high data depreciation.

Our theory thus provides us a way to understand the prevalence and force of monopoly power— it directs us to examine data sales models. Specifically, we should look for the prevalence of data sales versus data subscriptions and patterns in how these trade types are used. Therefore, how data is sold becomes the centerpiece of our empirical analysis.

To measure activity in data markets, we hand-collect a novel data set from Datarade, one of the largest online data marketplaces that connects buyers and sellers of data. The evidence about the geographic, industry, and data type coverage of this market place paints a nuanced picture of the way in which data is traded. Across over 3,600 data products, we find that 46% offer an option to buy the data for a one-time fee. However, over 90% offer a subscription or usage-based payment system. These fractions do not sum to one because many sellers offer multiple purchasing options. This finding suggests that at least half of all data providers have significant abilities to extract rents.

To test the predictions of our model, we need to merge the data marketplace evidence with company-level characteristics of the data sellers and the characteristics of their data products. Some of these data sellers are publicly listed companies, but many are private. We use a variety of data sources—Crunchbase, Pitchbook, Compustat, and CRSP to collect information on these companies background information and financing history. We use Edgar 10-K filings, combined with data product descriptions on Datarade, to fill in the characteristics of the markets in which they sell their data.

The model predicts that data sellers should choose one-time fees if they are financially constrained. If they do not urgently need cash, the subscription model of selling data is typically more profitable because it resolves the commitment problem. However, one-time fees bring in more revenue early in the life of the firm. The data confirms this prediction. We find a significant correlation between the way in which it sells its data and the age of the firm, the number of rounds of VC funding it has received, and the total amount of that funding. The older, better-funded firms are more likely to extract surplus, through the use of data subscriptions.

The model also predicts that when the market of data buyers is small, there is less scope to erode the value of data with future sales. Therefore, data that pertains to a more specialized group of potential buyers could be sold for a one-time fee, with little loss.

The data marketplace evidence also confirms this prediction. We determine the size of the market for data sales by comparing the textual similarity of data descriptions with the universe of firms' 10-K reports and then determining the industries with the greatest similarities. Then, we compute the number of publicly listed firms in those industries to determine the size of the market for the data. We find that this market size positively predicts data subscriptions and is negatively correlated with data sales.

We acknowledge that it is also possible that some settings lend themselves to complementarity in the use of data. In settings like speculative attacks or price-setting, the value of data might rise as others acquire it. In such settings, the data sellers' lack of commitment will be less costly, because data gains in value when more copies are sold. In contrast, dynamic complementarity, where an investor wants to learn data now that others will acquire later, still decreases the value of data over time.

Finally, one might object that data is not a durable good. It does depreciate. Information becomes stale. The rate of depreciation of data depends on the rate at which the environment changes. We explore the role of data depreciation in Section 2.2.

These results inform ongoing debates about data policy. The European Commission's "Free Flow of Non-Personal Data" initiative (Regulation *EU*2018/1807) argues that the development of efficient data markets that promote data mobility is essential to the development of the EU digital economy. It specifically cites "distortions of competition" as a problem to be addressed. Our paper explores competitive pricing in data markets. Traditional competition theory is designed for firms that produce physical products, with non-zero costs of replication, that are unlikely to have the strategic substitutability inherent in most information products.

Our results teach us that even a monopolist data seller has little effective market power in the market for data because it cannot commit not to compete against itself. This implies that data competition policy should err on the side of less activism. Our results further call into question even the idea that market power should be eliminated. If we do not provide some monopoly rents, the incentive to provide a high-quality product disappears. It may well make sense to regulate other harms. However, monopoly rents alone do not imply an undesirable outcome for consumers.

Related literature Our work builds on the insights of the literature on the dynamic Coase (1972) conjecture. When selling a durable good, a monopolist who lacks commitment not to lower future prices is forced to compete with its future self. As consumers become very patient, such a firm is unable to obtain any rents, despite its monopoly power (Fudenberg and Tirole (1991), Chapter 10).

What we add to this well-known problem is three-fold. First, we connect data to semi-durable goods. Second, we introduce strategic substitutability between users of data: Data that others have is less valuable. Not only is a firm competing with its own lower prices, as in a standard durable goods problem, it also suffers from its inability to commit not to sell to others. Third, we quantify the strength of this force in data marketplaces.

While the logic of debt dilution and information leakage (Brunnermeier and Oehmke 2013; Green and Liu 2021; DeMarzo and He 2021) have similarities, our mechanism has important differences as well. First, information is non-rival and can be replicated at near zero marginal cost. Second, information depreciates as the state of the world changes and old information becomes less relevant. Finally, information can be sold or licensed as a subscription, in a way that debt cannot.[1]

The literature on information sales considers when a seller should personalize data (Anat R Admati and Pfleiderer 1986), sell a data service (Anat R. Admati and Pfleiderer 1990), share data (Bergemann and Morris 2013), or offer a menu of data products (Bergemann, Bonatti, and Smolin 2018; Yang 2022). However, these are static models that miss the dynamic tension at the heart of this paper.

The data economy literature is similar in topic, but more different in its tools. Acemoglu et al. (2021) and Bergemann and Bonatti (2022) explore whether static data markets are efficient. Ichihashi (2020) show how firms can use consumer data to price discriminate. Jones and Tonetti (2020), Cong, Xie, and Zhang (2021) and Farboodi and Veldkamp (2022) build models of the data economy, but without market power in data markets. Existing work on the digital economy does explore whether data can be a source of market power (Kirpalani and Philippon 2020). Lambrecht and Tucker (2015) take a strategy perspective

[1]Externalities also arise in multilateral contracting with a principal's lack of commitment power in Arnott and Stiglitz (1991, 1993) and Segal (1999).

on whether data has the necessary features to confer market power. However, none of these consider the dynamic commitment problem of a data seller with market power, that we explore and quantify.

1 A Model of a Market for Data Purchases

Our model has two parts. The first part, describing households who purchase goods from producers that can utilize data, is there because we need households with utility functions in order to make welfare statements. This part of the model is constructed to make the willingness to pay for data decreasing in the number of other agents that buy the data. For all the non-welfare results, it would be sufficient to simply assume this relationship directly. The idea that data is a strategic substitute is an old one. It goes back to the work of S. Grossman and J. Stiglitz (1980). Of course, that paper was written about information used to choose portfolios of risky assets. But the idea of strategic substitutability in information acquisition or data purchases holds much more broadly. Hellwig, Kohls, and Veldkamp (2012) show that information is a strategic substitute in most settings where actions are strategic substitutes. Markets where quantities are chosen and prices clear markets are such a setting. If other agents demand more of a good or sell more of a product, that moves prices adversely and makes other less willing to take the same action. While we take strategic substitutability as a payoff primitive in this model, we sketch an oligopolistic goods market in the appendix to show why this form arises.

The second part of the model describes the problem of the data seller who lacks the power to commit not to engage in future data sales. This is where the model's novel ideas lie. One reason a firm might not be able to commit to restrict its sale of data is that proving the equivalence of two data sets is not easy. The seller could give the data set a different name, create linear combinations of the variables, or even add a small amount of noise to data. Although the information content of the new data set would be nearly equivalent to the original, it might be difficult to enforce a contract prohibiting the sale of identical data.

One might object that most data providers are not true monopolists. In many cases, buyers could obtain substitutable data from another source. However, since we are ex-

ploring whether market power might not be as effective as one might think, we start from a setting with an extreme degree of market power and see how much commitment problems remedy that power.

1.1 Model Assumptions

Households and data buyers Time is discrete $t = -1, 0, 1, 2, \ldots, \infty$. There are three types of players: a representative consumers, goods-producing firms, and a monopoly data supplier. The representative consumer has preferences over a measure-one continuum of goods, indexed by i

$$U = \sum_{t=0}^{\infty} \beta^t u_t, \qquad u_t = \int_0^1 \frac{\sigma}{\sigma - 1} q_{it}^{\frac{\sigma-1}{\sigma}} - p_{it} q_{it} \, \mathrm{d}i, \tag{1.1}$$

where $\sigma > 1$ governs the elasticity of substitution across goods.

There is a measure-2 continuum of goods-producing firms—twice as many firms as goods. Firms choose prices to maximize expected profit. At each date t, two firms are randomly selected to produce each good. This randomness simplifies our exposition by ensuring that firms face uncertainty in whom to compete with in the future. Once matched, two firms produce perfectly substitutable goods and compete as in standard Bertrand price competition.

Goods-producing firms use data to reduce their marginal cost of production. Let n be the measure of firms that have data. A firm without data has a marginal cost of $c = 1$. A firm with data can use the data to optimize its operations and has a marginal cost is $c = 1/z$, where $z > 1$ is the quality of the data.

The consumer and goods-producing firms are active from time $t \geq 0$ onwards. Time $t = -1$ is the ex-ante stage, where the data supplier makes investments into data, a process which we describe now.

Data supplier The data supplier is a monopolist who maximizes the expected present discounted value of profits. At the ex-ante stage $t = -1$, the data supplier chooses data quality and the duration of data access. From time $t \geq 0$ onwards, the data supplier

chooses the number of copies of the data to sell.

When the data supplier also chooses the duration of data access, they could simply sell the data, so that the buyers have permanent access to it. Alternatively, the supplier could incur a fixed cost η at the ex-ante stage $t = -1$ to setup the infrastructure for data subscription, so that buyers must pay every period. This technology limits the duration of data access.

The data supplier chooses the data quality z with a one-time, convex fixed cost $F(z)$. We assume that $F(z) = \frac{1}{2}\left(\left(\frac{\sigma-1}{\sigma}z\right)^{\sigma-1} - 1\right)^2/2$. The functional form is chosen to simplify expressions. Once produced, the data can be sold to multiple buyers with zero marginal cost.

Data sales take place over time. At each date t, the data supplier chooses how many additional copies to sell in that period. Then time moves on to $t+1$ and the game repeats.

Future payoffs are discounted. We denote the discount rate for $t \geq 0$ of data buyers as β and that of the data supplier as γ. We allow these discount rates to potentially differ. For simplicity, we assume the data supplier does not discount between the ex-ante stage ($t = -1$) and $t = 0$.

1.2 Discussion of Model Assumptions

No data resale. An important feature of the model is that data purchasers cannot resell data. In reality, most data is sold with a contract that forbids data buyers from selling the purchased data to others. But this stands in contrast to the assumption that data suppliers cannot use contracts to commit themselves.

One-sided commitment. While these assumptions comport with real features of data contracts, they do raise the question of why commitment is one-sided. One reason could be that there is one data supplier and many buyers. If a buyer violates the contract, the supplier has a strong incentive to sue. However, if the supplier were to commit to sell few copies and violated that contract, each buyer might find it optimal to wait for other buyers to sue. In other words, contract enforcement is costly. Enforcing contractual restrictions on data sales could be subject to a collective action problem.

Observable data quality and market size. We assume that data quality x and data buyer number n are observable to other buyers. In the model, both can be inferred as equilibrium choices. In practice, data quality is often conveyed by data vendors by providing a free sample of the data. Some platforms report the number of downloads. Both variables are part of the reputation of the data vendor. For example, people know that Bloomberg is used by many others, while satellite imagery is available to a smaller group, who pay its higher cost.

Up-front payments or financing of data purchases Data sellers could offer an option for the data buyers to finance their purchase, by making payments over time. Financing could be an attractive option if the data supplier is more patient than the buyer ($\gamma > \beta$). The data supplier can extract more surplus by postponing payments, but making those payments larger, and profiting from the difference in discount rates.

While most of our data market analysis assumes that financing is not a possibility, this is without loss of generality. With the exception of one case where the data seller has commitment, allowing a more general contract that allows for upfront cost and a gradual payment is formally equivalent to our model with up front payments, plus a loan from the data seller to the data buyer. These problems are separable: The optimal size of the loan is independent of the data sales choice and vice-versa.[2]

The case with data sales under commitment is different because a patient data seller can choose an increasing number of copies of data to sell, as a way of shifting data buyer profits forward in time and data seller profits back in time. This makes a particular sales path an imperfect substitute for a loan. Therefore, in this case only, we explore the interaction of data sale and financing.

1.3 Equilibrium

Equilibrium Definition: At the start of the game, the data supplier chooses the duration of data access and the data quality z to maximize the expected present value of their profits,

[2]If one party is more patient, the optimal loan may become infinite in size. That is a problem for any solution, but not one that has anything to do with the data market question at hand.

discounted at rate γ. Then, in each period t,

1. the data supplier makes take-it-or-leave-it offers to sell data to a chosen number of goods producers;

2. goods producers decide whether to buy the data or not, taking as given the others' past, current and expected future choices;

3. good producers are randomly matched and choose prices to maximize their one-period profit;

4. households choose their basket of goods to maximize (1.1), taking all prices as given;

5. time moves on to $t+1$.

Differentiating household utility (1.1) with respect to the quantity of each good and setting that to zero yields a first-order condition, which can be re-arranged in the form of a demand curve, $q_i = p_i^{-\sigma}$. The consumer surplus associated with each variety i is $\frac{\sigma}{\sigma-1} q_i^{\frac{\sigma-1}{\sigma}} - p_i q_i$.

Data buyers can generate profits from the data. The profit in each period depends on the quality of the data z and on how many firms n have access to the same data in that period. For each product variety, there are three possible market configurations in each period: 1) both producers have data; 2) one producer has data and the other does not, and 3) neither producer has data. Let n denote the measure of firms that have data in that period. Recall that the total measure of firms is 2. Thus, the probability of any firm having data is $n/2$ and the probability that two randomly selected firms have data is $n^2/4$ (case 1). Similarly, the probability that only a single firm in a matched pair has data is $(2-n)/2$. The probability that two randomly matched firms both lack data is $(2-n)^2/4$ (case 3). The fraction of varieties for which one firm has data and the other does not is $n(2-n)/2$ (case 2).

In cases 1) and 3), the two firms producing the variety have symmetric marginal costs. Symmetric firms that engage in price competition make zero profits.

In case 2), one producer has a lower marginal cost than its competitor. In this case, the firm with data maximizes its profit, $q_i(p_i - c)$, by charging price $p_d = \min\{\frac{\sigma}{\sigma-1}/z, 1\}$.

That is, there are two possible pricing regimes. In one, the firm charges the unconstrained monopolistic price $\frac{\sigma}{\sigma-1}/z$. In the other regime, i.e., when the unconstrained monopolistic price is above the competitor's marginal cost 1, the firm engages in limit pricing and charges the marginal cost of its competitor. As we show later, the functional form imposed on the data supplier's cost function $F(z)$ for improving data quality ensures that, in equilibrium, $\frac{\sigma}{\sigma-1}/z \leq 1$, thereby ensuring that we are always in the monopoly pricing regime, and $p_d = \frac{\sigma}{\sigma-1}/z$. We use the subscript d here to denote the price and quantity for a firm that has data when its competitor does not. This implies a markup of $\frac{\sigma}{\sigma-1}$. The firm without data sells nothing because its marginal cost of 1 exceeds this price.

Substituting the price p_d into the household demand curve implies that the quantity sold is $q_d = \left(z\frac{\sigma-1}{\sigma}\right)^{\sigma}$. This generates revenue for the firm with data of $p_d q_d = \left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}$. The firm's profit is $\frac{1}{\sigma}\left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}$ when its competitor does not have data.

The expected value of data, for one period, is the probability that the buyer's competitor will be uninformed, times the profit of having data when a competitor does not. We call this one-period expected value $\pi(n;z)$:

$$\pi(n;z) = \frac{1}{\sigma}\left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}(1-n/2). \tag{1.2}$$

For our subsequent analysis, it is useful to define $x \equiv (z\frac{\sigma-1}{\sigma})^{\sigma-1}$ as a monotonically transformed measure of data quality, and define $a \equiv 1/\sigma$, $b = a/2$, so that the per-period profit of each goods producer with access to data can be written as

$$\pi(n;x) = x(a-bn). \tag{1.3}$$

The substitutability of having access to data arises here because the goods producing firm makes zero profit in every case, except that case where it has data and its competitor does not. This is what makes the firm's expected value of data decline in the number of other firms that also have data. This is surely extreme. But it is a simple way of capturing an externality that is much more prevalent that this specific model mechanism. Appendix B works out a richer equilibrium model of oligopolistic firms that use data to forecast

demand, that justifies this substitutability assumption.

The reason for building out the household part of the model, rather than just assuming a π function, is to be able to derive welfare. We return to the welfare calculation in Section 5. For the rest of the model solution, we simply use the fact that the data buyers' (goods producer's) profit function π, that is increasing in quality z and decreasing in data sold n, i.e. $\pi_z > 0$, $\pi_n < 0$. The assumption that expected value is decreasing in data sold n captures the strategic substitutability of information. The parameter b governs the strength of the externality. If b is large, then data substitutability is strong. If b is close to zero, the strategic effect disappears.

1.4 Commitment Solution

We first explore a solution where the data supplier, after choosing the initial data quality, can commit to the quantity of data. It can tell customers exactly how many copies of the data will be sold in each-period. The data supplier will never sell any more copies of the data than the committed number n_t. This ability to commit will allow the data supplier to choose a higher price up front and will maximize the data supplier's revenue. After presenting this solution, we compare the price and revenue to the solution when the data supplier cannot commit.

Consider the static problem of the data supplier choosing the available quantity of data n_t to maximize the flow value it can extract from data buyers in each period t, given the initial data quality x. The maximized flow value is $\max n \cdot \pi(n; x)$. This is each buyer's maximum willingness to pay for the flow value of data, $\pi(n; x)$, times the number of copies sold n. Note that the maximizer n^* is time-invariant and equals to:

$$n^* = a/2b = 1.$$

The data supplier would thus like to commit to the same data availability at all times. Conditioning on data quality x, the value of the data producer is concave in n: having more clients n could bring more profits but could also reduces the willingness to pay by each client. There is a Laffer curve that plots the relationship between quantity and

revenue; the optimal choice n^* reflects the point at which the Laffer curve is maximized.

Figure 1 plots the data supplier's flow profit as a function of the number of subscribers. If the data supplier has no clients, it has zero revenue. But if there are $n = a/b = 2$ subscribers, it would earn a price of $x\left(a - bn\right)/\left(1 - \beta\right) = 0$ per units. This is also zero revenue. The peak revenue is achieved half way in between these two points at $n = 1$. This relationship between the data supplier's revenue and quantity is similar to the idea of a Laffer curve in public finance that describes the relationship between government taxation and government revenue.

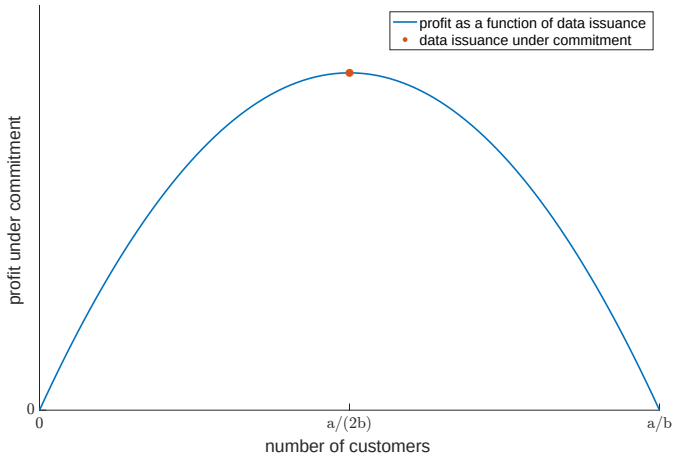


Figure 1: Equilibrium choice of n that maximizes profit under commitment

The formula for the curve is the total profit expression in (1.4), plus the one-time investment cost $F(x)$.

The maximum ex-ante value of the flow payoff for the data supplier is

$$V = \max_{x,n} \frac{n \cdot \pi\left(n; x\right)}{1 - \max\left(\beta, \gamma\right)} - F\left(x\right). \tag{1.4}$$

This is the maximum value the data supplier can extract from the data buyers. It is the present-discounted value of the maximum flow payoff for the buyers, using the discount rate of either the supplier or the buyer—whoever is more patient—net of the ex-ante fixed cost of producing data of quality x.

The value in equation (1.4) can be achieved if the data supplier can commit to selling n^* copies of data upfront in the initial period and give the data buyers an option to finance the purchase at the gross interest rate of $1/\gamma$. When the data buyers are more patient

than the supplier ($\gamma > \beta$), the buyers do not take up the financing option; they pay their valuation of data upfront, netting the supplier a revenue of $\frac{n \cdot \pi(n;x)}{1-\beta}$ in the initial period $t = 0$. On the other hand, when the supplier is more patient than the buyers, the supplier effectively lends to the buyers and receive a flow payment of $\max_n n \cdot \pi(n;x)$ each period, with a present discounted value of $\frac{n \cdot \pi(n;x)}{1-\gamma}$.

We refer to the path of data quantity $n^* = 1$ as the commitment solution. Substitute $x \equiv (z\frac{\sigma-1}{\sigma})^{\sigma-1}$ into the quadratic cost function $F(z)$, we can write the cost of data quality as $F(x) = (x-1)^2/2$. The optimal choice of data quality is thus

$$x^* = 1 + \frac{a^2}{4b\left(1 - \max\left(\beta, \gamma\right)\right)}.$$

Under the optimal choice of data quality, the value of the data supplier under commitment is

$$V^{commit} = \frac{a^2}{4b\left(1 - \max\left(\beta, \gamma\right)\right)} + \frac{1}{2}\left(\frac{a^2}{4b\left(1 - \max\left(\beta, \gamma\right)\right)}\right)^2.$$

Because the supplier can extract the statically maximal flow profits from the data buyers under the commitment solution, the supplier would never choose to incur the fixed cost η to setup the infrastructure for data subscription. Subscription is only relevant when the supplier cannot commit to the sequence of future data sales, as we show below.

1.5 Data Sales without Commitment

The problem is that after selling the data to n buyers at time t, the data supplier has the incentive to sell to more clients at $t + 1$. Doing so reduces the profitability of prior clients. Knowing that future copies will be sold reduces prior clients' willingness to pay. The lack of of commitment therefore reduces the market power of the data supplier. Note that while the data supplier cannot commit to the number of copies of data it sells in the future, it does commit to the quality of the data as the quality is chosen up-front and does not change over time.

We now study this more realistic game of data sales without commitment. In section 1.6 below we analyze the game where the supplier provides the data as a subscription,

and a buyer pays for data access in each period.

Formally, at each time t, the data supplier solves a recursive problem, with the total number of data copies already sold n and data quality x as state variables. The value function represents the supplier's present discounted value of the revenue derived from their data. The data supplier is a monopolist. So they can make a take-it-or-leave-it offer to the data buyer for the buyer's willingness to pay for the data.

The buyer, however, has a willingness to pay that is based on their rational expectation of the data supplier's future data sales. The buyer will earn $\pi(n; x)$ profits from their data in the first period. But will earn only $\pi(\tilde{n}; x)$ dollars of profit the following period, if an additional $\tilde{n} - n$ customers buy the data. Thus, at date t, the buyer's willingness to pay is $\sum_{\tau=1}^{\infty} \beta^{\tau-t} \mathbb{E}_t [\pi(n_\tau; x)]$, where n_τ is the total purchases of the data, up to and including date τ. Next, we rewrite this willingness to pay, as a function of the optimal selling strategy of the data supplier. We don't yet know that strategy. We will use a placeholder strategy function g and then solve for the optimal selling strategy, as a fixed point of the buyers' and data supplier's problem.

Let $g(n, x)$ denote the data supplier's optimal choice of new total data sales, given that n goods-producing firms have already purchased the data. As we show below, the optimal choice depends only on n and is invariant to the data quality x, so we suppress the argument and simply write $g(n)$. The number of new clients being sold to is $g(n) - n$. Let $g^2(n) \equiv g(g(n))$, $g^0(n) = n$ and define $g^k(n)$ to be the operation g performed k times on n. Then g^k represents how many total copies of the data the data supplier will choose to sell k periods from now, if there are already n total buyers today. Note that n is a stock of total past buyers and $g^k(n)$ is a new stock of buyers. If the data supplier decided to sell no new copies of data, then this would be represented as $g^k(n) = n$.

Substituting $g^{t-1}(n)$ for n_τ in the sum above, a goods-producing firm's total stream of profits from data can be expressed as

$$\bar{\pi}(n; x) = \sum_{t=1}^{\infty} \beta^{t-1} \mathbb{E}\left[\pi\left(g^{t-1}(n); x\right) | n, x\right] \tag{1.5}$$

This profit $\bar{\pi}(n; x)$ incorporates data buyers' conditional rational expectations that their

ability to extract value from this data will decline over time, given the current state variables (n, x). It anticipates the future path of data sales. This present discounted revenue from data is also the buyer's willingness to pay for data.

Since the data supplier is a monopolist, the revenue-maximizing choice is to charge each buyer their willingness to pay for data. Giving this willingness to pay, the data producer, who has already sold n data copies, chooses to sell $\tilde{n} - n$ additional copies of the data in period t. This choice earns the supplier a price of $\bar{\pi}(\tilde{n}, x)$ earned for each of the $(\tilde{n} - n)$ additional copies of the data sold that period. The supplier's optimal choice should maximize this current revenue, plus the discounted present value of future revenue. Using $V(n, x)$ to denote the data supplier's continuation value given the state variables, this choice problem can be written recursively as,

$$V(n, x) = \max_{\tilde{n}} \left\{ (\tilde{n} - n) \bar{\pi}(\tilde{n}, x) + \gamma V(\tilde{n}, x) \right\}. \tag{1.6}$$

Definition 1.1. Given data quality x, a *Markov perfect equilibrium (MPE)* is the pair of functions $\{\bar{\pi}(\cdot; x), g(\cdot)\}$ such that:

1. the goods producers' willingness to pay for data $\bar{\pi}(n; x)$ is consistent with their rational expectation of the future sequence of data sales, satisfying (1.5);

2. the policy function for the data supplier $g(n)$ solves the problem solves (1.6).

Optimal data sales without commitment In principle, the dynamic game involving a non-commitment data supplier and a sequence of data buyers (i.e., the goods producer) is difficult to analyze, as the MPE involves a fixed point in the two functions $\{\bar{\pi}(\cdot; x), g(\cdot)\}$. However, under our tractable formulation, the recursive problem of the data supplier (1.6) is quadratic in the state variable n, thereby enabling us to solve for the equilibrium in closed-form.

The solution to this model shows how commitment problems result in more data sales, lower prices and reduced profits. Importantly, the solution also tells us where to look for evidence of this commitment problem: declining data prices, over time.

Proposition 1. The data issuance policy function $g(n)$ is characterized by an equilibrium scalar δ:

$$g(n) = n + (1-\delta)\left(\frac{a}{b} - n\right) \qquad \text{with} \qquad \delta = \frac{1-\sqrt{1-\gamma}}{\gamma}. \tag{1.7}$$

Data buyers' willingness to pay is characterized by an equilibrium scalar $\xi \equiv \frac{1-\beta}{1-\beta\delta}$

$$\bar{\pi}(n,x) = \xi \frac{\pi(n,x)}{1-\beta} = \frac{1-\beta}{1-\beta\delta}\frac{a-bn}{1-\beta}x. \tag{1.8}$$

Given data quality x, the data provider's value function is

$$V(n,x) = \frac{b\delta}{2(1-\beta\delta)}\left(\frac{a}{b} - n\right)^2 x \tag{1.9}$$

All proofs are in Appendix A.

Interpretation. The Markov perfect equilibrium is captured by the two endogenous variables (δ, ξ), respectively parametrizing the data producer's and data buyers' equilibrium strategy.

Note $\bar{n} \equiv \frac{a}{b} = 2$ is the maximum total sales; data is worthless to goods producers when every potential competitor has access to it. Given existing sales n at the beginning of each period, the total sales at the end of each period $g(n)$ is a weighted average between $\bar{n}$ and n. Intuitively, $1-\delta$ captures how aggressively the data producer sells to new clients. When $\delta = 1$, $g(n) - n = 0$, meaning the data producer does not sell to new clients. A lower δ translates to more aggressive sales.

On the other hand, ξ scales how data buyers value data, reflecting their expectation about future data sales. When buyers anticipate no future data sales ($\delta = 1$), $\xi = 1$, and $\bar{\pi}(n) = \frac{a-bn}{1-\beta}$ coincides with the present discounted future flow value if no future data sales are made after the current period. Absent commitment, buyers anticipate future sales and thereby place a proportional discount ξ on the value of data.

Note that the path of total data issuance is $g^t(0) = \left(1-\delta^t\right)\frac{a}{b}$, meaning the path of new

sales at each time period is

$$g^t(0) - g^{t-1}(0) = (1-\delta)\frac{a}{b}\delta^{t-1}$$

which decays to zero exponentially at rate δ.

The path of sales price is $\bar{p}\left(g^t(0)\right) = \frac{\xi\left(a - bg^t(n)\right)}{1-\beta}$, which simplifies to

$$\bar{\pi}\left(g^t(0)\right) = \frac{\xi a \delta^t}{1-\beta},$$

which also decays to zero exponentially at rate δ.

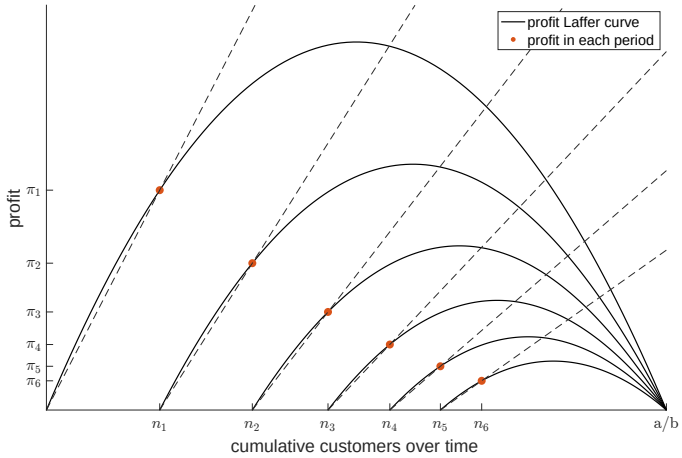


Figure 2: Equilibrium paths of data sales, prices and profit, without commitment

Figure 2 shows the equilibrium path without commitment. Specifically, the outermost curve plots $(n-0)\,\bar{\pi}(n)$, which is the initial Laffer curve. The red dot reflects the equilibrium choice n_1 in the first period and the corresponding profit. The second curve shows $(n-n_1)\,\bar{\pi}(n)$ as a function of n, which is the Laffer curve in the second period, and the red dot on the curve reflects the equilibrium choice n_2 in the second period and the corresponding profit. The slopes of the dashed lines reflect the sequence of equilibrium willingness to pay $\bar{\pi}(n_t)$.

Ex-ante stage: choosing data quality We now solve the data supplier's ex-ante problem of choosing data quality, noting that the data acquisition cost can be written as $F(x) =$

$(x-1)^2/2$, where $x := \left(\frac{\sigma-1}{\sigma}z\right)^{\sigma-1}$:

$$V^{sales} = \max_x V(0,x) - (x-1)^2/2$$

Under the data sales scheme, the data supplier's optimal choice of data quality is $x = \frac{a^2}{2b\left(1-\beta+\sqrt{1-\gamma}\right)}$. We can then plug that choice back into the data supplier's initial value function to find the value of the data to the supplier with discount rate γ:

$$V^{sales} = \frac{1}{2}\left(\frac{a^2}{2b\left(1-\beta+\sqrt{1-\gamma}\right)}\right)^2. \tag{1.10}$$

Data Buyers' Discount Rate and the Value of Commitment An important implication of the model is that the value of commitment depends crucially on the data buyer's discount rate β. Intuitively, low discounting implies that existing data buyers expect more negative externalities arising from future data sales. This expectation causes the data buyer to discount the value of the data by more. The value of data to the data buyers is lower than in the full-commitment case, because the buyers know that additional future copies will be sold.

To see this formally, consider the case where data buyers are more patient than the supplier ($\beta \geq \gamma$). This makes financing irrelevant. We compare the gap in the supplier's surplus with and without commitment, and we examine how this gap varies with the buyers' discount rate. Formally, for any given data quality x, if the data supplier were able to commit to selling a fixed number of copies of the data initially, they would have chosen $n^* = \arg\max n \cdot (a - bn) = 1$ copy of the data, which coincides with the choice under the subscription scheme. On the other hand, absent commitment, the data sales scheme features an increasing sequence of copies sold over time, with $\lim_{t\to\infty} n_t = \lim_{t\to\infty} g^t(0) = 2 > n^*$.

For a given quality of the data, the value that the data supplier can extract from data buyers is $V(0,x)$ when lacking commitment; the value is $\pi(n^*,x)/(1-\beta)$ when the data supplier can commit. Proposition 1 implies that $\frac{V(0,x)}{\pi(n^*,x)/(1-\beta)} = 2\delta\xi$, where δ and ξ are defined in the proposition. Recognizing that $\lim_{\beta\to 1}\xi = 0$, we have that the value obtained

by the data supplier who cannot commit is vanishing relative to the commitment value as $\beta \to 1$. As the data buyers anticipate future data sales, their willingness to pay relative to the commitment case declines towards zero (i.e., $\frac{\bar{\pi}(n,x)}{\pi(n,x)/(1-\beta)} = \xi \to 0$ as $\beta \to 1$).

1.6 Data Subscriptions

We now consider the data supplier's total profit under a subscription scheme. Under this scheme, a data buyer must pay for the data access each period. Because, in each period, customers observe how many subscriptions are sold, the subscription scheme allows the data supplier to commit to the optimum, period-by-period. In each period, it is optimal to sell n subscriptions, where n maximizes the flow profit in that period. We find that this n coincides with the commitment solution: $n^* = 1$.

The data supplier's ex-ante value at $t = -1$ is the present-discounted value of the flow profits, minus the ex-ante fixed costs of producing data quality x and paying for the subscription infrastructure:

$$V^{subscription} = \max_x x \cdot \frac{n^* (a - bn^*)}{1-\gamma} - F(x) - \eta.$$

The optimal choice of data quality is thus

$$x^{subscription} = 1 + \frac{a^2}{4b(1-\beta)}.$$

Under the optimal choice of data quality, the value of the data supplier choosing the subscription scheme is

$$V^{subscription} = \frac{a^2}{4b(1-\beta)} + \frac{1}{2}\left(\frac{a^2}{4b(1-\beta)}\right)^2 - \eta. \tag{1.11}$$

Data sales without commitment vs. subscription The data supplier chooses ex-ante whether to adopt a subscription scheme or sell the data for perpetual access. The data sales scheme has the disadvantage that, due to the lack of commitment power, the data supplier loses market power. The subscription scheme enables commitment. But the sub-

scription scheme also has two disadvantages: (1) it involves the ex-ante fixed cost η; (2) the cashflow arising from the subscription scheme is realized period-by-period. Since a one-time fee delivers revenue up front, it makes sense that an impatient data supplier prefers to sell data, rather than adopt a subscription model. One of the key determinants of the pricing model a data supplier selects is how impatient they are.

Formally, a data supplier prefers outright sales over subscription if $V^{sales} \geq V^{subscription}$, which is true iff $\left(\frac{1}{1-\beta+\sqrt{1-\gamma}}\right)^2 - \left(\frac{1}{2(1-\gamma)}\right)^2 + \frac{8b^2}{a^4}\eta > 0$. The left-hand side of this inequality is increasing in the discount rate of the buyers (β), decreasing in the discount rate of the supplier (γ), and decreasing in the fixed cost η of setting up subscription services.

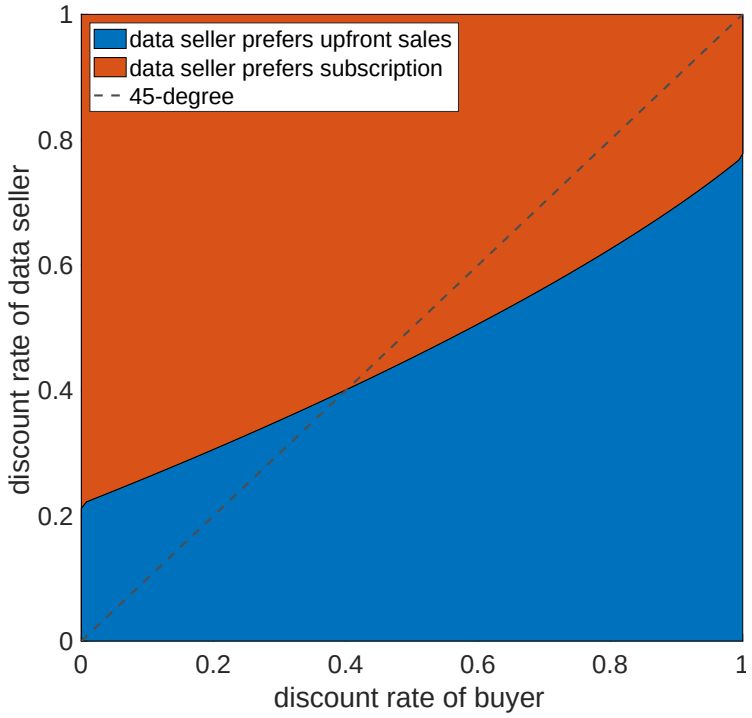


Figure 3: Data contract choices that maximize supplier revenue, without commitment. Lower discount rates mean that there is a stronger preference for early cash flows.

Figure 3 illustrates, for a given cost η, the set of discount rates for buyers and suppliers that would result in the choice of a supplier to sell the data outright, or to license it with a subscription fee. More patient suppliers wait for the subscription revenue because it is higher.

1.7 Sales and Subscriptions with Depreciating Data

In reality, data depreciates. A given data point becomes less valuable over time because it reflects some current condition that changes over time. While we use the term “duration of data access" to describe a prohibition on future use of data that may have future value, data depreciation is a property of the data, not the contract. Data depreciation is functionally like losing access to some or all of the data, but has a different economic origin.

Appendix C gives a simple micro-foundation for this depreciation and ties a depreciation rate to the volatility of the variable the data speaks to. To keep our analysis transparent, we do not model the process of depreciation. We simply consider the extremes, to inform our thinking about how depreciation could matter and to inform our empirical analysis.

The results so far pertain to the case with no depreciation. Next, we consider what happens when data fully depreciates each period.

When data depreciates fully, the sales and subscription models become identical. What subscriptions allow the firm to do is to withdraw access to the data each period. Full data depreciation implies that there is no relevant data to access in future periods. The distinction between sales and subscriptions becomes moot. What we conclude is that the higher the depreciation rate, the smaller in the different in firm payoff from choosing subscriptions over sales.

2 Testable Predictions of the Model

While the previous section laid out the solution to the data sales and data subscription models, we ultimately want to ask whether the model’s mechanism is consistent with data markets. This section derives predictions to test. These are not meant to capture all the relevant considerations in a data supplier’s choice of business model. Rather, the predictions below are the considerations that are indicative of our mechanism at work.

2.1 Financial Constraints and the Choice of Data Sales or Subscription

Financial constraints on the part of data suppliers may induce them to behave as if they were less patient. When unable to borrow, a data supplier may value immediate cash flows more highly than later ones. This type of financial constraint might explain why some suppliers choose the one-time data payment model, instead of a data subscription model.

H1: Financially constrained data suppliers are more likely to sell data, rather than license it.

2.2 Data Depreciation and Data Sales

So far, we have characterized data as being like a durable good. At the same time, data loses value over time. Next, we consider how data depreciation affects our predictions. Data depreciates for many reasons. First, data may become inaccurate over time due to changes in the environment or context in which the data was collected. For example, data on consumer behavior from five years ago may not be applicable today as consumer preferences may have evolved. Second, data can degrade over time due to corruption of electronic files. Lastly, legal requirements or regulations may require data to be deleted or destroyed after a certain period. Appendix C formalizes data depreciation and formally links it to volatility in the economic environment.

Suppose data quality depreciates exponentially at rate $\lambda < 1$. Given initial quality x, the flow payoff to data buyers is

$$x\left(a - bn_1\right), \lambda x\left(a - bn_2\right), \lambda^2 x\left(a - bn_3\right) \ldots$$

Given the anticipated sequence of future data sales, the willingness to pay by a borrower with discount rate β for the data that depreciates at rate λ is therefore identical to the willingness to pay by a borrower with discount rate $\lambda\beta$ for data that do not depreciate. Because the supplier's flow profit in equilibrium also scales proportionally with the quality of data, the equilibrium in a model with data depreciation is therefore isomorphic to

one without data depreciation, but where the data buyers have discount rate $\beta\lambda$ and the supplier has discount rate $\gamma\lambda$, where $\lambda < 1$ captures the effect of data depreciation.

If we shrink each player's rate of time preference by a common factor $\lambda < 1$, the data supplier is more likely to favor data sales over subscription. This can be seen from Figure 3: starting with a given discount rate (β, γ) in the red region and moving towards the origin, the supplier switches from preferring subscription to upfront sales.

H2: Data that depreciates faster is more likely to be sold, instead of licensed.

2.3 Market Size and Data Sales or Subscriptions

Having a larger pool of potential customers for data makes subscriptions more advantageous both because it lowers the per-user cost of setting up subscriptions and because it makes commitment more valuable. This is what we call the market size effect. Setting up a subscription service is costly. When the number of potential buyers of a data product is small, the cost is not justified by the small potential revenue. Furthermore, the presence of more potential data buyers induces a data producer to invest more in data quality. Higher quality data raises the value of commitment, achieved through subscriptions.

Consider this relationship between market size and the sales choice mathematically in the model. We can model to a larger market with measure $\mu > 1$ of goods and measure 2μ of potential data buyers. In this environment, the flow profit of the data supplier rises by the proportion $\mu > 1$. Scaling up the firm's revenue by the fixed factor μ, but leaving the fixed costs of setting up subscriptions η and the cost of investing in data quality $F(\cdot)$ unchanged is the same problem as keeping the revenue at its original level and shrinking the fixed costs by a factor of μ. Formally, the equilibrium choices in the model with larger market size are isomorphic to the baseline model with a cost of providing data quality equal to $F(\cdot)/\mu$ and the fixed cost of setting up a subscription services reduced to η/μ. Both fixed cost changes make subscriptions more advantageous.

H3: Data that is relevant to a large set of firms is more likely to be licensed, instead of sold.

2.4 The Decline in New Customers and New Revenue

For the data subscription model, the results in Section 1.6 show that the number of data copies n is time-invariant. That means that after the first period, there are no new customers. Obviously, that is unrealistic, in part because our model is one with no costs or delay from customer acquisition. In reality, it takes time to acquire the optimal number of customers. But after reaching that market size, data subscriptions should stagnate.

When data firms sell data, the customer base grows over time, because of the lack of commitment. The maximum total data that a data seller would ever sell is a/b. Proposition 1 tells us that when the number of data copies reaches $n = a/b$, data becomes worthless to goods-producing firms. When $0 < \delta < 1$, δ regulates how quickly the firm converges to this maximum. When data sellers are patient, β is close to one and therefore δ is close to 1. This makes for slow convergence. The patient data seller sells a little bit of additional data every period. But impatient sellers have β close to zero and δ close to zero. These data sellers sell most of their data quickly and approach a/b data points sold quickly. As the data market becomes saturated, the value of the data deteriorates.

H4: New customers and new revenue should grow more slowly for data firms than for comparable non-data firms.

The decline in price comes from two sources. First, is the classic force from the Coase conjecture that applies more generally to durable goods. When the most eager buyers have already purchased the data, the remaining market has a lower value of data, which makes the optimal price for a seller lower. Second, is the new force, the strategic substitutability of data. Data that others already know typically generates less profit for the user. While one or neither force could be at work for a non-data firm, both forces operate to push prices down over time for a given data set.

Of course, not all data exhibit equal strategic substitutability. Financial data is known to be nearly worthless if it is widely known. Similarly, strategic business data typically offers profits only if that market strategy is not saturated. In contrast, weather data is something everyone can benefit from, to all be better prepared for the day. The difference is not in the type of data, as much as it is the use of that data. If some trader discovered a

profitable trading strategy, based on the weather, then this would be less valuable if other acquired weather data and used it for that same purpose. Data has strategic substitutability when the uses of data are to inform actions that themselves exhibit substitutability. The finance data is a strategic substitute because buying risky assets is less profitable when others buy the same assets at the same time. Investors compete with each other. The typical weather forecast is not a strategic substitute because people do not typically compete with each other in the use of an umbrella.

3 Data and Descriptive Statistics

3.1 Data Products and Providers

We obtain data on the market for data from Datarade (https://datarade.ai/), a global data trading platform that helps companies discover, compare, and connect with data providers across the globe. As of Summer 2024, Datarade is one of the largest data markets, hosting more than 4,000 different data products, spanning a broad set of major categories, provided by more than 600 data provider firms. Datarade is Germany-based, but it hosts data providers and products around the world. Nearly half of the data providers headquarter in the US. Other popular headquarter locations for data providers are the UK, India, and Germany.[3]

Datarade provides detailed information on each data product. For each of the products hosted on the platform, the information page reports key statistics that are useful for our empirical study. Fig. A.1 in the Appendix provides an example of the information page. First, they report the data provider company, which allows us to later merge with company-level information from other data sources. Second, a menu of data transaction methods is provided, including one-off purchase, licensing (monthly or yearly), and usage-based pricing. Sometimes, the price level is available, though the coverage is very limited.

Datarade also tags each data product with up to five data categories. There are 670

[3]See https://www.crunchbase.com/organization/datarade.

Figure 4: Word Cloud of Product Descriptions on Datarade

Notes. This is a word cloud generated by the text of all the data descriptions on Datarade.

different data categories which are finely defined. In our example, the product is associated with five different tags: Location Data, Foot Traffic Data, Mobile Location Data, Raw Location Data, Mobility Data. The description of the data product is extensive (see Fig. A.2 for the continued example). It describes, not only the data contents, but also potential use cases. In our example, the data description lists uses such as consumer insights, market intelligence, advertising, and retail analytics. We use these descriptions to map each product to industries that may use it. Figure 4 presents a word cloud of the descriptions to provide an overview of how data suppliers describe their data.

The data provider firms in our sample are usually startups or small to medium sized firms, so we link these data provider firms from Datarade to company-level data sets

Crunchbase and PitchBook using a fuzzy name-matching algorithm. This allows us to extract their characteristics including the firm's geographic location, funding history, survival, and age. To measure the financial conditions of these data provider firms, we observe whether these companies obtained any funding from venture investors (as of Summer 2024), the total rounds of investments, and the total amount of investment in USD. For the unmerged companies, we impute the funding variables to be 0. The assumption that data sellers that have never received documented venture funding are unfunded, is one that we verified by hand checking and internet searches.

3.2 Measuring Potential Data Demand, Depreciation and Data Sales Growth

Identifying Relevant Industrial Firms for Each Data Product To connect data products to their potential buyers, we measure how relevant each data product is to each industrial sector. First, we create a mapping from each product to its relevant industrial users by jointly analyzing data product descriptions from Datarade and business descriptions (item 1) and Management Discussion and Analysis (MD&A, item 7) from 10-K of industrial firms. In specific, for each data product description, we calculate its textual similarities with all the Edgar 10-K filings in 2020, and find the top 30 filings with the highest similarity scores. If the overview and potential use cases outlined in a data description are more similar to an industrial firm's business description, we determine that it is more likely that the industrial firm may find the data useful.

Quantifying Potential Buyers In our theory, market size represents the potential interest from data buyers who compete in the same goods market. If there are more direct competitors interested in a data product, the data will lose more value from additional sales. Thus, if most data buyers are from the same industry, that suggests a more severe commitment problem. In contrast, if buyers' interests are dispersed across different industries, the data are less likely to be used by direct competitors, and the commitment problem is mitigated.

For each data product, we count the industries with sufficiently high relevance to that

data description. We group data products into 10 deciles based on the number of different three-digit SIC industries in the interested buyer set. A lower decile indicates more concentrated interest from potential buyers competing in product markets—we label this as high market interest.

Measuring Data Depreciation The depreciation of a data product is the weighted average of the depreciation rates for the industries relevant to that data (i.e., $j \in \mathcal{J}_i$). For each industry, data depreciation captures the ability of information from the past to predict future business performance. If past information is more predictive of future business activities in that industry, the depreciation of data is low; if past information is less predictive of future business activities, data in that industry depreciates rapidly.[4]

To implement this idea, for each Compustat firm l, we predict the firm's return on assets (ROA) using lagged ROA in the previous year, with the following model,

$$ROA_{lt} = \psi_0 + \psi_1 \, ROA_{l,t-1} + \varepsilon. \tag{3.1}$$

Denote the R^2 from this regression for firm l as R^2_l. For each data product i, we calculate the average R^2 over top potential buyer set identified above for i. We group data products into 10 deciles based on the R^2 calculated. A lower decile suggests *high* data depreciation rate. In contrast, a higher decile suggests a slow-changing environment and low data depreciation.

Measuring the Growth of a Data Seller In the model, the size of a data seller is the number of customers. When a firm attracts more customers, the firm's value rises. Venture value is the value of a start-up firm. For a firm to grow in value, it needs to attract more customers or more customer dollars. Thus, while customer growth and venture value growth are not identical, customer growth is a necessary condition for sustained venture growth.[5]

[4] In Appendix C, we formally derive a conceptual framework to formalize this measurement construction.

[5] This is not a mathematical statement: We could construct a model where some other effect raises firm value. However, in practice, while valuation effects might boost a firm's valuation, they do not sustain

The other measure of customer base we explore is Google trends. This data is only sold online. Buying an online product begins by navigating to the product website. Most people navigate to a website through a search engine and Google is the dominant search engine. Obviously, not every search results in a sale and not every sale will record a Google search. But trends in searches suggest changes in customer activity.

3.3 Descriptive Statistics

Our sample covers 3,206 data products traded on Datarade that have complete information on product description and available transaction methods. These products are offered by 445 data provider firms. Each product is tagged with four categories, on average. Table 1 details the top ten categories of data products traded on Datarade and their market shares. The category company data, B2B contact data, and similarly, B2B leads data, B2B marketing data, are among the most popular on the platform. Other business intelligence data, such as point of interest data and business website data, are also popular. These detailed data categories are further aggregated to 22 broad data categories, which we describe in Appendix Table A.1.

Table 1: Key Data Categories on the Market for Data

Category	**Product Count**	**Percentage (Out of 3,206)**
Company Data	478	14.91%
B2B Contact Data	414	12.91%
B2B Leads Data	407	12.69%
B2B Marketing Data	376	11.73%
B2B Email Data	330	10.29%
Point of Interest (POI) Data	247	7.70%
Firmographic Data	246	7.67%
Location Data	203	6.33%
Business Website Data	198	6.18%
Machine Learning (ML) Data	183	5.71%

Notes. This table presents the top ten data categories in our data set. The sample includes 3,206 data products traded on Datarade. Each product could be tagged with on average four different categories. This table presents the count of products in each of the top ten categories, and the proportion as a percentage in the full sample.

long-term growth.

These key data categories also suggest that our sample is well-suited to examine the commitment problems described by our model for three reasons. First, these data categories describe types of data that are durable, not ephemeral insights. That is important because it suggests that the threat of a seller selling this durable data to others is a relevant problem. Second, these data products are suitable for multiple users (not firm-specific). Finally, they fit the model because future data sales would likely decrease the value of data for earlier buyers. For example, if more competitors obtain the contact list of potential customers, then the earlier data buyers' ability to profit from contacting that same group of customers diminishes.

The main variable in our analysis is the transaction model adopted by each product. We report this information in Table 2 Panel (a). One-off purchase is available for 64% of the products. Licensing (either annually or monthly) is offered for 81% of the products. For 14% of the data products, one-off purchase is the only transaction model offered. The proportion of products with only licensing model offered is 31%. This also means that about 50% of the products offer one-time purchasing and licensing simultaneously.[6]

Next, we describe the key characteristics of data providers. Importantly, we measure data provider firms' access to financing. Measuring any firms' financial constraints is always imperfect (Whited and Wu 2006). In the case of data providers in our sample, the problem is compounded by the fact that most of these are not public firms. Because the firms are private and often startup companies, we use the presence of venture capital (VC) financing, the number of financing rounds and the dollar amount of VC financing as proxies of their lack of financial constraint. While not ideal, these measures are widely used when studying private firms. In addition, since (Hadlock and Pierce 2010; Ma, Murfin, and Pratt 2022) find that firm age is a useful proxy for the lack of financial constraint, we use firm age as an alternative measure.

Table 2 Panel (b) reports the funding status of the Datarade data provider firms. Out of all the firms, 16% obtained venture funding, with the average number of funding rounds being 0.56. Conditional on obtaining funding, the average number of funding rounds is

[6]In our simple model, this is only a knife-edge case. However, we could easily extend the model by adding data buyers that are heterogeneous in their rates of time preference. In such an environment, some firms will choose to offer both sales and subscriptions to segment the market.

Table 2: Financial Information about Data Suppliers: Summary Statistics

	Mean	Median	Std.Dev	25th Percentile	75th Percentile
Panel (a): Transaction Models of Data Products (N = 3,206)					
One-off Purchase (0/1)	0.64	1	0.48	0	1
Licensing (0/1)	0.81	1	0.39	0	1
Only One-off Purchase (0/1)	0.14	0	0.34	0	0
Only Licensing (0/1)	0.31	0	0.46	0	1
Panel (b): Data Providers (N = 445)					
Obtained VC Funding (0/1)	0.16	0.00	0.36	0	0
No. of Funding Rounds	0.56	0.00	1.48	0	0
Total Funding (mil. USD)	4.20	0.00	26.00	0	0
Age (as of June, 2024)	13.65	10.00	14.20	6.00	15.00

Notes. This table presents summary statistics for the transaction models of the data products on Datarade (Panel (a)), and the characteristics of data providers selling products on the platform (Panel (b)). Provider-level information is obtained from Crunchbase. Product-level transaction models are extracted from product pages from Datarade.

3.5. The total funding obtained by a provider firm is highly skewed, with an average of 26 million USD. The median age is 10 years old.

4 Testing Model Predictions: Data Subscriptions and Data Sales

Recall that the main prediction of the theory was that data subscriptions were associated with market power in the data marketplace, whereas one-time sales were likely to yield less revenue for firms. This section measures the extent of data purchases versus subscriptions and provides empirical tests of the theory's prediction. These results do not establish any causal relationships. Instead, these are novel empirical facts that inform us about the features of data markets. They are consistent with the prediction that firms choose data sales in some cases and licensing or subscriptions in others.

4.1 Financial Constraints and Data Transaction Model

The first prediction of the model (H1) that we test pertains to the relationship between data suppliers' financial conditions and the type of pricing models they adopt. To do this, we estimate

$$TransactionModel_i = \Gamma_0 + \Gamma_1 \cdot Financing_i + \theta_{category} + \varepsilon_i. \tag{4.1}$$

The dependent variables are dummy variables indicating whether the data product offers a certain transaction model (one-off purchase or licensing). The key explanatory variable *Financing* takes multiple forms. In one specification, financing measures the total number of rounds of VC financing; in another, it measures the total amount of VC financing obtained by the data provider. We also use firm age as a proxy. In this model and others in this section, we control for fixed effects at the level of primary data category as tagged on Datarade to account for similarities of transaction models among similar data products. Our model predicts a negative Γ_1 coefficient for the one-off purchase model, and positive Γ_1 coefficients for licensing.

Consistent with the model's predictions, Table 3 shows that more financially constrained firms are more likely to choose data sales. In Panel A, we present the results using the logarithm of the number of funding rounds of the provider as the explanatory variable. In columns (1) and (3), we find that the number of funding rounds is associated with a lower probability of offering one-off purchase models but a higher probability of offering licensing. In columns (2) and (4), we focus on cases where the product only has one of the two transaction models. We find consistent results: More funding rounds are associated with a lower probability of offering only one-off purchase, and higher probability of offering only licensing model.

In terms of economic magnitude, going from a firm with one round of financing to a firm with two rounds, the probability of using the one-off purchase model goes down by $0.094 \times (\ln 2 - \ln 1) = 6.5$ percentage points (pp), which is a 10.1% decrease from the base rate of 64%. Applying a similar calculation to column (3), we find that going from one round of financing to two rounds of financing is associated with a 3.7 pp (4.6% from the base) increase in the probability of using licensing. In Panel (b), we show that the results

Table 3: Providers' Financial Condition and Transaction Models

	(1)	(2)	(3)	(4)
		Only		*Only*
	One-off Purchase	One-off Purchase	Licensing	Licensing
Panel (a): Total Funding Rounds and Transaction Models				
ln(No. of Funding Rounds)	-0.094***	-0.066***	0.053***	0.080***
	(0.016)	(0.008)	(0.011)	(0.016)
Observations	3,206	3,206	3,206	3,206
R-squared	0.126	0.039	0.037	0.129
Panel (b): Total Funding Amounts and Transaction Models				
ln(Total Funding Amt)	-0.003**	-0.005***	0.004***	0.002
	(0.002)	(0.001)	(0.001)	(0.002)
Observations	3,206	3,206	3,206	3,206
R-squared	0.118	0.036	0.036	0.122
Panel (c): Provider Firm Age and Transaction Models				
Provider Firm Age	-0.032**	-0.015*	0.017*	0.034***
	(0.014)	(0.009)	(0.010)	(0.013)
Observations	1,929	1,929	1,929	1,929
R-squared	0.176	0.052	0.067	0.186

Notes. This table correlates the type of data transaction models available for each data product with the funding status of the providers. Panel (a) presents the analysis using the logarithm of total funding rounds, Panel (b) presents the analysis using the logarithm of the total funding amount received, and Panel (c) presents the analysis in which firm age is used to proxy the level of financial constraint. All specifications control for primary data category fixed effects and report robust standard errors. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

are robust to the use of the logarithm of total funding amounts as the key explanatory variable.

In Panel (c), we switch to use firm age to proxy the level of financial constraints (Hadlock and Pierce 2010; Ma, Murfin, and Pratt 2022). In this analysis, firm age is transformed using an inverse hyperbolic sine transformation. We also can only focus on the subsample of data providers with age data as this information cannot be reliably imputed. We find that older firms, which are typically considered to be less financially constrained, are less likely to offer one-off purchase as a transaction option, and they are more likely to use

licensing as the transaction model. In terms of economic magnitudes, doubling firm age is associated with a 2.2 pp decrease in the use of one-off purchase and a 1.1 pp increase in the use of licensing.

4.2 Data Depreciation

Next, we explore (H2), that data depreciation makes data sales more likely. In our model, data about environments that change quickly (fast-changing finance data vs. slow-moving consumer tastes) is like an environment with a higher discount rate. As Figure 3 shows, data providers may have more commitment power, and there is likely less loss from the one-time fee model.

To test this, we again use the transaction model as a proxy for a provider's commitment power. We use the following model:

$$TransactionModel_i = \Gamma_0 + \Gamma_1 \cdot High\ Depreciation_i + \theta_{category} + \varepsilon_i. \tag{4.2}$$

In this empirical model, the key explanatory variable is $High\ Depreciation_i$, which indicate the top decile data depreciation rates calculated using the steps outlined in Section 3.2. The high-depreciation data categories identified through this method coincides with our intuitions. For example, the most rapidly depreciatiating data category is job posting information.

It is worth noting that in this exercise, we face the following problem. While data depreciation mitigates the dynamic commitment problem, it also makes licensing convenient for buyers who need to access the data frequently. This is a force that is outside the scope of our model.

To avoid this confounding effect, we exclude a subsample of data products for which the convenience force is likely to be dominant. Specifically, we exclude products that are updated daily (or more frequently than daily).

Table 4 shows that the information depreciation rate associated with a data product positively correlates with the use of one-off purchases and negatively predicts the use of licensing models. The 0.171 in column (1) suggests that high-depreciation products are

Table 4: Data Depreciation Rate and Transaction Models

	(1)	(2)	(3)	(4)
		Only		*Only*
	One-off Purchase	One-off Purchase	Licensing	Licensing
High Depreciation	0.171***	0.157***	-0.107**	-0.121**
	(0.054)	(0.051)	(0.051)	(0.052)
Observations	709	709	709	709
R-squared	0.261	0.077	0.088	0.271

Notes. This table correlates the type of data transaction models available for each data product with the information depreciation rate of connected industries. The equation estimated is (4.1). Depreciation is defined in Section 3.2. All specifications control for primary data category fixed effects and report robust standard errors. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

17.1 percentage points more likely to use one-off purchase. Thus, data that depreciates is more likely to be sold with a single transaction, rather than a subscription, which allays fears about data seller market power. The same change is associated with a 10.7 pp decrease in the probability of using licensing.

4.3 Market Size

Next, we explore H3, that a larger pool of potential buyers makes data subscriptions or licensing more likely. Potential buyers are those that may find the data product useful and thus may make a purchase. More potential buyers make data subscription more profitable than data sales. The key explanatory variable in this model is *Market Size*, which measures the number of relevant industries from which data buyers might arise.

$$TransactionModel_i = \Gamma_0 + \Gamma_1 \cdot MarketSize_i + \theta_{category} + \varepsilon_i. \tag{4.3}$$

Table 5 shows the relationship between market size and pricing models. We find that market size negatively correlates with the use of one-off purchases and positively correlates with the use of the licensing model in transactions. The coefficient of 0.011 in column (3) means that products that are of top-decile interest to more buyers are 11 pp more likely to use the licensing model than products that are of interest to fewer data buyers. This

Table 5: Potential Market Size and Transaction Models

	(1)	(2)	(3)	(4)
		Only		*Only*
	One-off Purchase	One-off Purchase	Licensing	Licensing
Market Size (Deciles)	-0.013***	-0.013***	0.011***	0.012***
	(0.003)	(0.002)	(0.003)	(0.003)
Observations	3,206	3,206	3,206	3,206
R-squared	0.122	0.039	0.038	0.125

Notes. This table correlates the type of data transaction models available for each data product with the potential market size of each product. The equation estimated is (4.3). Market size is defined in Section 3.2. All specifications control for primary data category fixed effects and report robust standard errors. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

suggests that as the market for data grows, there is likely to be more use of subscriptions.

4.4 Slower Sales Growth

The final prediction of the model, H4, tests the mechanism in a different way. The problem a data seller faces is that they need to restrict new data sales to earn monopoly rents. This need to restrict sales is not a problem faced by most other non-data tech firms in the industry because most technologies do not exhibit strategic substitutability. A computer is not less valuable when others buy the same computer. Instead, for most technology products the reverse is true: If everyone purchases and uses Apple products, for example, using non-Apple products becomes less compatible with others and less valuable. Therefore, we test H4 by comparing the customer and revenue growth of data firms to their non-data counterparts, after controlling for time trends.

We compare growth in firm value. The comparison is between data-providing companies and companies that do not sell data, but are in related industries. Each observation is a financing round of a startup company. The firms are data providers in our Datarade sample, described above. The control sample includes companies in Crunchbase that are in top 5 industries that data providers operate in. These industries include Analytics, Software, Information Technology, Advertising, and Big Data.

Table 6: Time-Series Trend of Data Value and Potential Customers

	(1) ΔVenture Value	(2) ΔGoogle Trends Index
Data Provider	-3.646***	-0.066*
	(1.137)	(0.035)
Observation-level	Company-Round	Search Term-YearMonth
Observations	38,474	19,741
R-squared	0.001	0.012
Fixed Effects	Year, Lag Between Rounds	Year-Month

Notes. ΔVenture Value, which captures the percentage change of the current financing round from the previous round. Source: Crunchbase and PitchBook. ΔGoogle Trends Index is the percentage change in the Google Trends index from 2018 to 2023 for data and non-data tech firms. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

We estimate the following model,

$$\Delta VentureValue_{it} = \Gamma_0 + \Gamma_1 \cdot DataProvider_i + \theta_t + \varepsilon_{it}.$$

The key dependent variable is ΔVenture Value, which captures the percentage change of the current financing round from the previous round. The key explanatory variable is *DataProvider*, indicating if the company is a data provider. We control for the lag from the previous round to this round and for year fixed effects and we cluster standard errors at the year and state levels. Thus, our estimates compare that are data providers with those that are not, that raised funding in the same year, with the same funding gap from the previous round. Column (1) of Table 6 shows that data providers' value growth (Δ) is significantly lower than the comparable companies. The average ΔVenture Value in the sample is 5.81, meaning that an average startup appreciates roughly 6 times between venture rounds, a common feature for startups at this stage. The -3.646 estimate suggests that this increase is much lower for data sellers: The ΔVenture for data firms is 2.16 (the difference of $5.81 - 3.65$).

Column (2) examines the time-series change in search popularity, captured using Google Trends. This analysis compares the Google Trends index from 2018 to 2023 of two groups

of terms: data sales terms and technology terms. Data sales terms consist of all data categories extracted from Datarade (i.e., "commercial market data," "business location data"). As a control group, we use technology terms. These are top breakthrough technologies, as identified annually by MIT Technology Review during the sample period (i.e., "custom cancer vaccines," "3-D metal printing," "online privacy").[7] For each of the data and technology terms, we extract the monthly Google Trend index. We estimate the following model,

$$\Delta GoogleTrend_{it} = \Gamma_0 + \Gamma_1 \times DataProvider_i + \theta_t + \varepsilon_{it}.$$

The key dependent variable is ΔGoogleTrend, which captures the percentage change of the current Google Trends from the previous round. The key explanatory variable is *DataProvider* indicating if the term is a data sales term or a general technology term. We control for year-month fixed effects and cluster standard errors at the same level. In this thought experiment, we are comparing data sales terms with other technology terms in terms of their monthly change in search popularity, after controlling for granular time trends, using year-month fixed effects. Table 6 reveals that data related technology terms have lower growth compared to the control terms.

A potential measurement challenge is selection: Most surviving firms grow their sales over time. However, our approach of taking the difference between data and non-data firms should remove this effect. As long as data firms have a similar selection of surviving firms as their non-data counterparts, this effect should disappear in the difference. However, future work could investigate whether the failure rates are indeed similar.

5 Consumer, Goods Producer and Data Supplier Welfare

Having shown that the empirical evidence from data markets is consistent with the model, we end by investigating what this implies for welfare.

Of course, our model is a stylized one. It surely does not contain all the welfare-relevant trade-offs one would want for a thorough policy analysis. However, our results

[7]The MIT Technology Review's annual breakthrough technologies can be accessed at: https://www.technologyreview.com/supertopic/tr10-archive/.

suggest that data sales models regulate monopoly power and this can have an effect on consumer surplus and on welfare.

Data poses a trade-off for consumer surplus. Good producers without data do not have monopoly pricing power. Lower prices benefit consumers. However, data makes goods producers more efficient. Those without data have higher marginal costs, which get passed on to consumers as well. That trade-off shows up throughout our comparison of data market structures as well. To quantify the trade-off, we first derive expressions for consumer surplus from the model, in the case of data sales and data licensing.

To compute consumer surplus, we substitute in the solutions for the equilibrium price and quantity p_i and q_i for each variety. Recall that each variety has three possible market structures: 1) both firms have data, 2) one has data, the other does not, and 3) neither has data. Then, we multiply each of these surpluses times the fraction of varieties that yield that surplus. This yields a one-period consumer surplus, as a function of data supplier choices n and z. Finally, we substitute in these data producer choices and cumulate up the one-period surpluses to yield total lifetime consumer surplus. Appendix A follows these steps and prove the following result.

Proposition 2. a. Lifetime consumer surplus when data is sold is

$$\sum_{t=0}^{\infty} \beta^t u_t^{sale} = \frac{1}{\sigma - 1} \left[\frac{1 + x - 2\left(\frac{\sigma-1}{\sigma}\right)^{\sigma-1} x}{1 - \beta\delta^2} + \frac{\left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} x}{1-\beta} + \frac{2x\left(1 - \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} x\right)}{1 - \beta\delta} \right], \tag{5.1}$$

where $\delta = \frac{1-\sqrt{1-\gamma}}{\gamma}$, and $x = 1 + \frac{1}{\sigma}\frac{1}{2\left(1+\sqrt{1-\gamma}-\beta\right)}$.

b. Lifetime consumer surplus, when data is licensed as a subscription is

$$\sum_{t=0}^{\infty} \beta^t u_t^{subscription} = \frac{1}{1-\beta}\frac{1}{\sigma-1}\left[\frac{1}{4}\left(1 + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} \hat{x} + 2\hat{x}\right)\right], \tag{5.2}$$

where $\hat{x} = 1 + \frac{1}{\sigma}\frac{1}{2(1-\beta)}$.

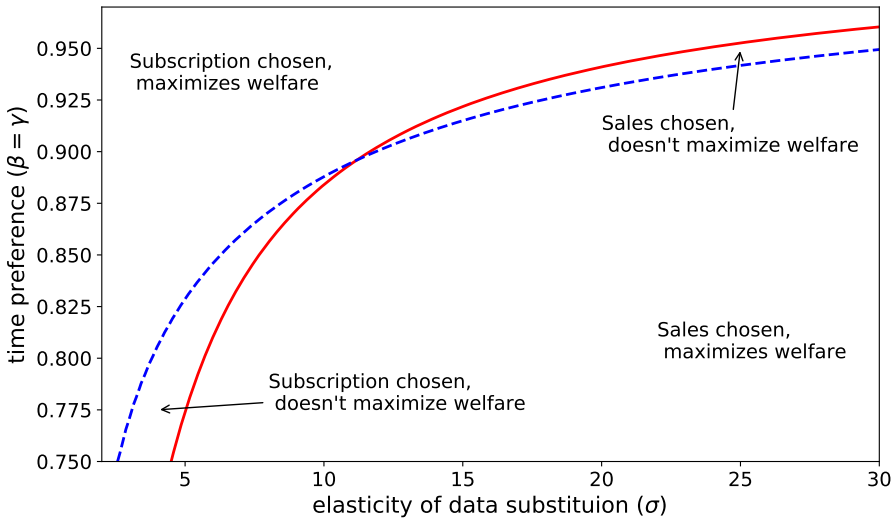
c. Data supplier surplus is $V = \max\{V^{sales}, V^{subscription}\}$, where V^{sales} follows equation (1.10) and $V^{subscription}$ follows equation (1.11).

d. Data buyer (goods producer) surplus is zero in either case.

In both cases, δ represents the rate at which data sales converge to steady state and x and $\hat{x}$ are the optimal choices of data quality.

To see how surplus depends on the model parameters, we choose some values for time preference and elasticity and plot regions where each sales model delivers higher total welfare. We set the cost of subscription setup to 0.5. But changes in this cost would just shift the choice and welfare curves out in a predictable way.

Figure 5: Do Data Sales or Data Subscriptions Maximize Welfare?



Notes. Welfare is the sum of consumer surplus (5.1) or (5.2), data seller surplus (1.10) or (1.11) and goods producer surplus, which is zero. The fixed cost of setting up the subscription model η is set to 0.5.

Understanding Consumer Welfare Finally, we can use the model to answer the policy relevant question about what sorts of data sales are better for consumers. Figure 5 shows that, unless goods are highly elastic, data licensing or subscription is often better for welfare than data sales. At first, that might seem surprising. After all, data sellers make less use of their monopoly power. Monopoly power usually creates deadweight loss that makes consumers worse off. So data sales would seem to be better for consumers. However, the two consumer surplus expressions in Proposition 2 are not easily rankable. Because data sellers lose most of their rents from data, they have little incentive to invest in data quality. Since higher quality data makes goods firms more efficient, which in turn,

makes good cheaper, consumers benefit when more data is sold. When goods are highly elastic, data sales typically dominate because, in this case, data producers have little incentive to invest in quality data anyway.

The last question we pose to the model is whether firms' incentives to choose data sales or data subscriptions aligns with consumer welfare. Figure 5 shows that firms often choose the welfare-maximizing form of data transaction. However, in some cases, firms incentives to choose the right data business model will be misaligned with the consumers' interests. When goods are inelastic, the monopoly power of subscriptions can be particularly harmful. When patience and substitutabiity are high, society might prefer that firms invest more in quality data and offer subscriptions. These results suggest that optimal regulation might be crafted in a very targeted way.

6 Conclusion

Many policy makers are concerned about the market power of data sellers. However, the inability of data sellers to commit to sell limited copies of data, combined with the fact that data's strategic value declines in the number of users, forces competitive pricing. Even if the seller is a monopolist, the inability to restrict future data sales makes the seller compete with its future self in data provision. Of course, with the loss of monopoly power comes a loss of incentive to produce quality data.

Monopoly power for data is not entirely bad. Just like patent laws protect the monopoly power of innovators to encourage innovation, copyright law could be seen as protection for data to encourage the discovery of new, high-quality data sources.

Data subscription services are a tool for firms to restore monopoly power. While subscriptions restore monopoly power, they also restore an incentive for data sellers to invest in producing high-quality data. We find that subscriptions benefit consumers when goods are not close substitutes.

Market power in data markets depends on the pricing models data sellers choose to implement. We collected data from one of the largest on-line data marketplaces to investigate data sellers' pricing strategies. We derived three predictions from the model about

firms that are most likely to choose to sell subscriptions, instead of sales and one prediction about the growth rate of data sale revenue. All four predictions are supported by new facts from data marketplaces. These findings support our hypothesis about the nature of data markets and give us new insight into the functioning of this rapidly-expanding and politically controversial market.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. 2021. "Too much data: Prices and inefficiencies in data markets." *Forthcoming, American Economic Journal: Microeconomics.*

Admati, Anat R, and Paul Pfleiderer. 1986. "A monopolistic market for information." *Journal of Economic Theory* 39 (2): 400–438. ISSN: 0022-0531. https://doi.org/https://doi.org/10.1016/0022-0531(86)90052-9. https://www.sciencedirect.com/science/article/pii/0022053186900529.

———. 1990. "Direct and Indirect Sale of Information." *Econometrica* 58 (4): 901–928. ISSN: 00129682, 14680262, accessed August 2, 2024. http://www.jstor.org/stable/2938355.

Arnott, Richard, and Joseph Stiglitz. 1991. *Equilibrium in Competitive Insurance Markets with Moral Hazard.* NBER Working Papers 3588. National Bureau of Economic Research, Inc, January.

———. 1993. "Price Equilibrium, Efficiency, and Decentralizability in Insurance Markets with Moral Hazard." *Working Paper.*

Bergemann, Dirk, and Alessandro Bonatti. 2022. *Data, Competition, and Digital Platforms.* Technical report.

Bergemann, Dirk, Alessandro Bonatti, and Alex Smolin. 2018. "The Design and Price of Information." *American Economic Review* 108, no. 1 (January): 1–48. https://doi.org/10.1257/aer.20161079. https://www.aeaweb.org/articles?id=10.1257/aer.20161079.

Bergemann, Dirk, and Stephen Morris. 2013. "Robust Predictions in Games With Incomplete Information." *Econometrica* 81 (4): 1251–1308.

Brunnermeier, Markus K., and Martin Oehmke. 2013. "The Maturity Rat Race." *Journal of Finance* 68, no. 2 (April): 483–521.

Cong, Lin William, Danxia Xie, and Longtian Zhang. 2021. "Knowledge Accumulation, Privacy, and Growth in a Data Economy." *Management Science* 67 (10): 6480–6492. ISSN: 15265501. https://doi.org/10.1287/mnsc.2021.3986.

DeMarzo, Peter, and Zhiguo He. 2021. "Leverage Dynamics without Commitment." *Journal of Finance* 76 (3): 1995–1250.

Farboodi, Maryam, and Laura Veldkamp. 2022. *A Model of the Data Economy.* Working Paper, Working Paper Series 28427. National Bureau of Economic Research. http://www.nber.org/papers/w28427.

Fudenberg, Drew, and Jean Tirole. 1991. *Game Theory.* The MIT Press.

Green, Daniel, and Ernest Liu. 2021. "A Dynamic Theory of Multiple Borrowing." *Journal of Financial Economics* 139 (2): 389–404.

Grossman, Sanford, and Joeseph Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70(3):393–408.

Grossman, Sanford J, and Joseph E Stiglitz. 1980. "On the impossibility of informationally efficient markets." *The American economic review* 70 (3): 393–408.

Hadlock, Charles J, and Joshua R Pierce. 2010. "New Evidence on Measuring Financial Constraints: Moving Beyond the KZ Index." *Review of Financial Studies* 23 (5).

Hellwig, Christian, Sebastian Kohls, and Laura Veldkamp. 2012. "Information choice technologies." *American Economic Review* 102 (3): 35–40.

Ichihashi, Shota. 2020. "Online Privacy and Information Disclosure by Consumers." *American Economic Review* 110, no. 2 (February): 569–595. https://ideas.repec.org/a/aea/aecrev/v110y2020i2p569-95.html.

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

Kirpalani, Rishabh, and Thomas Philippon. 2020. *Data sharing and market power with two-sided platforms.* Technical report. National Bureau of Economic Research.

Lambrecht, Anja, and Catherine E. Tucker. 2015. *Can Big Data Protect a Firm from Competition?* Technical report. mimeo.

Ma, Song, Justin Murfin, and Ryan Pratt. 2022. "Young firms, old capital." *Journal of Financial Economics* 146 (1): 331–356.

Segal, I. 1999. "Contracting with externalities." *Quarterly Journal of Economics* CXIV (May): 337–388.

Whited, Toni M, and Guojun Wu. 2006. "Financial constraints risk." *Review of Financial Studies* 19 (2): 531–559.

Yang, Kai Hao. 2022. "Selling Consumer Data for Profit: Optimal Market-Segmentation Design and Its Consequences." *American Economic Review* 112, no. 4 (April): 1364–93. https://doi.org/10.1257/aer.20210616. https://www.aeaweb.org/articles?id=10.1257/aer.20210616.

Appendix

A Proofs

Proofs of Proposition 1

Proof. First conjecture $\bar{\pi}(n) = \frac{\xi(a-bn)}{1-\beta}$ and we solve the data producer's problem.

$$V(n) = \max_{\widetilde{n}} \left\{ (\widetilde{n} - n) \frac{(a - b\widetilde{n})}{\sqrt{1-\beta}} + \beta V(\widetilde{n}) \right\}.$$

$$\text{FOC} \quad \frac{(a - bg(n))}{\sqrt{1-\beta}} - \frac{(g(n) - n)\, b}{\sqrt{1-\beta}} + \beta V'(n') = 0$$

$$\text{Envelope} \quad V'(n) = -\frac{(a - bg(n))}{\sqrt{1-\beta}}$$

Substitute the envelope condition into the first-order condition:

$$(a - bg(n)) - (g(n) - n)\, b - \beta \left[a - bg^2(n) \right] = 0$$

Conjecture $a - bg(n) = \delta(a - bn)$, we get

$$\left(2\delta - 1 - \beta\delta^2 \right)(a - bn) = 0$$

Given $\delta \in (0,1)$, the solution is $\delta = \frac{1-\sqrt{1-\beta}}{\beta}$.

We now solve for the data buyer firms' willingness to pay under rational expectation:

$$\begin{aligned} \bar{\pi}(n) &= \sum_{t=1}^{\infty} \beta^{t-1} \left[a - bg^{t-1}(n) \right] \\ &= \frac{a - bn}{1 - \beta\delta} \\ &= \underbrace{\frac{1-\beta}{1-\beta\delta}}_{\xi} \frac{a - bn}{1-\beta} \end{aligned}$$

Substitute $\delta = \frac{1-\sqrt{1-\beta}}{\beta}$, we can simplify $\xi = \sqrt{1-\beta}$.

Using these solutions, we can solve for the data producer's value function:

$$\begin{aligned} V(n) &= \frac{\xi}{1-\beta}\sum_{t=1}^{\infty}\beta^{t-1}\left(g^t(n)-g^{t-1}(n)\right)\left(a-bg^t(n)\right) \\ &= \frac{\xi}{1-\beta}\sum_{t=1}^{\infty}\beta^{t-1}\delta^{t-1}(1-\delta)\left(\frac{a}{b}-n\right)\delta^t(a-bn) \\ &= \frac{\xi}{1-\beta}\sum_{t=1}^{\infty}\beta^{t-1}\delta^{2t-2}\delta(1-\delta)\frac{1}{b}(a-bn)^2 \\ &= \frac{\xi\delta(1-\delta)\frac{1}{b}(a-bn)^2}{(1-\beta)(1-\beta\delta^2)} \\ &= \frac{\delta(1-\delta)b(\bar{n}-n)^2}{(1-\beta\delta)(1-\beta\delta^2)} \end{aligned}$$

where $\bar{n} \equiv a/b$. The value at time 0 is

$$V(0) = \frac{\delta(1-\delta)a^2/b}{(1-\beta\delta)(1-\beta\delta^2)} \tag{A.1}$$

$$= \frac{\delta a^2}{2b\sqrt{1-\beta}} \tag{A.2}$$

$$= \frac{a^2}{2b}\frac{\sqrt{1-\beta}-(1-\beta)}{\beta(1-\beta)} \tag{A.3}$$

□

Proof of Proposition 2 In case 1) where both firms have data, the price is the firms' marginal cost, which is $1/z$. Since demand is $q_i = p_i^{-\sigma}$, substituting these into consumer utility (1.1) yields a one-period consumer surplus of $v_1 = z^{\sigma-1}/(\sigma-1)$.

In case 2) where firms are asymmetric, the price was $z\sigma/(\sigma-1)$, which implies a quantity of $(z\sigma/(\sigma-1))^{-\sigma}$. Substituting price and quantity into consumer utility (1.1) yields a one-period consumer surplus of $v_2 = \frac{1}{\sigma-1}\left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}$.

In case 3) where neither firm has data, the price and quantity are both 1. One-period consumer surplus is $v_3 = 1/(\sigma-1)$.

Multiplying each of these three consumer surplus expressions by the fraction of varieties that have each market structure (the probabilities), we can express total consumer surplus as a function of number of copies of data sold n_t and data quality z:

$$u_t = \frac{1}{\sigma-1}\left[\left(\frac{2-n_t}{2}\right)^2 + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} z\frac{n_t^2}{4} + \frac{2n_t(2-n_t)}{4}z\right]$$

Of course, the copies of data sold and the data quality are also endogenous choices of the data provider. The next step is substitute those in.

Given the equilibrium policy function in (1.7): $n_t = (1-\delta)\frac{a}{b} + \delta n_{t-1}$ with $\frac{a}{b} = 2$, we know

$$2 - n_t = \delta(2 - n_{t-1}) = 2\delta^t$$

$$(2-n_t)^2 = 4\delta^{2t}$$

$$n_t^2 = \left(2 - 2\delta^t\right)^2 = 4 + 4\delta^{2t} - 8\delta^t$$

$$2n_t(2-n_t) = 4\left(2-2\delta^t\right)\delta^t = 8\delta^t - 8\delta^{2t}$$

$$\sum_{t=0}^{\infty}\beta^t\left(\frac{2-n_t}{2}\right)^2 = \sum_{t=0}^{\infty}\beta^t\delta^{2t} = \frac{1}{1-\beta\delta^2}$$

$$\sum_{t=0}^{\infty}\beta^t\frac{n_t^2}{4} = \frac{1}{1-\beta} + \frac{1}{1-\beta\delta^2} - \frac{2}{1-\beta\delta}$$

$$\sum_{t=0}^{\infty}\beta^t\frac{2n_t(2-n_t)}{4} = \frac{2}{1-\beta\delta} - \frac{2}{1-\beta\delta^2}$$

We can write the consumer's ex-ante surplus as

$$\begin{aligned}
&\sum_{t=0}^{\infty}\beta^t u_t \\
=\ & \frac{1}{\sigma-1}\left[\frac{1}{1-\beta\delta^2} + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}x\left(\frac{1}{1-\beta} + \frac{1}{1-\beta\delta^2} - \frac{2}{1-\beta\delta}\right) + x\left(\frac{2}{1-\beta\delta} - \frac{2}{1-\beta\delta^2}\right)\right] \\
=\ & \frac{1}{\sigma-1}\left[\frac{1+\left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}x - 2x}{1-\beta\delta^2} + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}x\left(\frac{1}{1-\beta} - \frac{2}{1-\beta\delta}\right) + x\frac{2}{1-\beta\delta}\right] \qquad \text{(A.4)} \\
=\ & \frac{1}{\sigma-1}\left[\frac{1+x-2\left(\frac{\sigma-1}{\sigma}\right)^{\sigma-1}x}{1-\beta\delta^2} + \frac{\left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}x}{1-\beta} + \frac{2x\left(1-\left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}x\right)}{1-\beta\delta}\right] \qquad \text{(A.5)}
\end{aligned}$$

Under data sales, we have $\delta = \frac{1-\sqrt{1-\gamma}}{\gamma}$, and ex-ante data choice $x = 1 + \frac{1}{\sigma}\frac{1}{2\left(1+\sqrt{1-\gamma}-\beta\right)}$. We can compute consumer surplus by substituting δ and x into (A.5).

Under subscription, $n^* = 1$, $x = 1 + \frac{a^2}{4b(1-\beta)} = 1 + \frac{1}{\sigma}\frac{1}{2(1-\beta)}$. The consumer surplus per period is

$$u_t^{sub} = \frac{1}{\sigma-1}\left[\frac{1}{4}\left(1 + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}x + 2x\right)\right].$$

The ex-ante consumer surplus is

$$\sum_{t=0}^{\infty}\beta^t u_t^{sub} = \frac{1}{1-\beta}\frac{1}{\sigma-1}\left[\frac{1}{4}\left(1 + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}x + 2x\right)\right].$$

B Market Foundations for Data Externality

Strategic substitutability of data arises in many contexts. Here is a simple one where there is imperfect competition and firms use data to forecast uncertain shocks to their profit.

FIRMS There are n_F firms, indexed by i: $i \in \{1, 2, \ldots, n_F\}$. Each firm chooses the number of units of each good they want to produce, an $N \times 1$ vector $\mathbf{q}_i$, to maximize risk-adjusted profit, where the price of risk is ρ_i.

$$U_i = \mathbf{E}\left[\pi_i | \mathcal{I}_i\right] - \frac{\rho_i}{2}\mathbf{Var}\left[\pi_i | \mathcal{I}_i\right] - g(\chi_c, \widetilde{\boldsymbol{c}}_i). \tag{B.1}$$

Firm production profit π_i depends on quantities of each good, $\mathbf{q}_i$, the market price of each good, $\mathbf{p}$, and the marginal cost of production of that good, $\boldsymbol{c}_i$:

$$\pi_i = q_i' \left(\mathbf{p} - \boldsymbol{c}_i\right). \tag{B.2}$$

PRICE Our demand system is Cournot. Therefore, the price of good i can depend on the amount every firm produces.

Each good j has an average market price that depends on an good-specific constant and on the total quantity of that good that all firms produce:

$$p_j^M = \bar{p}_j - \frac{1}{\phi}\sum_{i=1}^{n_F} q_{ij}. \tag{B.3}$$

Each firm does not receive the market price for its good, but rather has a firm-specific price that depends on a firm-specific demand shock $\mathbf{b}_i$. The demand shock $\mathbf{b}_i$ is a vector with jth element b_{ij}. This vector is random and unknown to the firm: $\mathbf{b}_i \sim N(0, I)$, which is i.i.d. across firms. The price a firm receives for a unit of good j is thus $p_j + b_{ij}$. We can express firm i's price in vector form as

$$p_i = \left[p_1^M, p_2^M, \ldots, p_N^M\right]' + b_i. \tag{B.4}$$

INFORMATION Each firm generates n_{di} data points. Each data point is a signal about the demands for each good: $\boldsymbol{s}_{i,z} = \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{i,z}$, where $\boldsymbol{\varepsilon}_{i,z} \sim N(\mathbf{0}, \Sigma_e)$ is an $N \times 1$ vector. Signal noises are uncorrelated across goods and across firms. All firms can observe all the data generated by each firm. Of course, other firms' data is not relevant for inferring b_i. But this allows firms to know what other firms will do.

EQUILIBRIUM

1. Each firm chooses a vector of marginal costs $\boldsymbol{c}_i$, taking as given other firms' cost choices. Since the data realizations are unknown in this ex ante investment stage, the objective is the unconditional

expectation of the utility in (B.1).

2. After observing the realized data, each firm updates beliefs with Bayes' law and then chooses the vector $\mathbf{q}_i$ of quantities to maximize conditional expected utility in (B.1), taking as given other firms' choices.

3. Prices clear the market for each good.

Substitutability Externality of Information To show substitutability, we now want to consider, what happens when one additional firm gets a signal with one unit of precision, about consumer demand? How does that effect the utility of another firm observing that same amount of information?

We start with the optimal production decision of a firm. Define $\boldsymbol{H}_i = \left(\rho_i \mathbf{Var}\left[\boldsymbol{b}_i|\mathcal{I}_i\right] + \frac{2}{\phi}\boldsymbol{I}_N\right)^{-1}$. Using Bayes law to replace the expectation $\mathbf{E}\left[\boldsymbol{b}_i|\mathcal{I}_i\right]$ with the weighted sum of signals $\boldsymbol{K}_i\boldsymbol{s}_i$, with $\boldsymbol{K}_i = \Sigma_{b_i}(\Sigma_{b_i} + \Sigma_{\epsilon_i})^{-1}$ yields

$$\boldsymbol{q}_i = \boldsymbol{H}_i\left(\bar{\boldsymbol{p}} + \boldsymbol{K}_i\boldsymbol{s}_i - \frac{1}{\phi}\sum_{j=1,j\neq i}^{n_F}\boldsymbol{q}_j - \boldsymbol{c}_i\right). \tag{B.5}$$

We have set the model up so that the only way one firm's information affects another firm is through the level of production. Notice that the firm's output is increasing in H_i, which itself in decreasing in conditional variance. Data reduces conditional variance. Thus, data will increase a firm's expected level of production.

A one unit increase in the precision of data increases conditional precision by one unit. The decrease in conditional variance, the inverse of conditional precision is $-\mathbf{Var}\left[\boldsymbol{b}_i|\mathcal{I}_i\right]^2$.

The size of this effect of data on the own level of production is $\partial q_i/\partial n_{di} = -\partial E[q_i]/\partial H_i \cdot \partial H_i/\partial Var \cdot \mathbf{Var}\left[\boldsymbol{b}_i|\mathcal{I}_i\right]^2$. Since $E[q_i]/\partial H_i > 0$ but $\partial H_i/\partial Var < 0$, the effect of data on own production is positive. This makes sense because a firm with more data faces less uncertainty and produces more aggressively.

The effect of firm i's data on firm i' works through the price level. When firm i produces one unit more of good j, the price of good j falls by $1/\phi$. Thus, $\partial p/\partial n_{di} = -1/\phi\partial q_i/\partial n_{di}$.

Next, we solve for the effect on expected profits. Expected profits can be expressed as:

$$\mathbf{E}\left[\boldsymbol{q}_i'\left(\boldsymbol{p}_i - \boldsymbol{c}_i\right)\right] = \mathbf{E}\left[\boldsymbol{q}_i'\left(\mathbf{E}\left[\boldsymbol{p}_i|\mathcal{I}_i\right] - \boldsymbol{c}_i\right)\right]. \tag{B.6}$$

Notice that the effect of a one unit increase in price of a good is an increase of $\boldsymbol{q}_{i'}'$ in profits. Thus, putting these effects together with the chain rule, we find that the marginal effect of an increase in data owned by firm i on firm i's profit is $\boldsymbol{q}_{i'}' \cdot (-1)/\phi\partial q_i/\partial n_{di} < 0$. So one firm's data reduces another firm's profit.

C Foundations for Data Depreciation

To understand why data depreciates and how much it depreciates, we need to model how firms derive competitive advantage from data. Data is information. Big data, used with modern big data techniques

is used for prediction. AI and machine learning are, at their core, prediction technologies. So the data we are talking about is information used to make predictions more accurate. More accurate predictions can inform more optimal or efficient actions. The greater efficiency of actions is the source of firms' competitive advantage. Understanding the role of data will allow us to deduce its depreciation rate.

Consider a firm that uses data with normally-distributed noise to forecast some profit-relevant variable that follows an AR(1) process with normal innovations:

$$\theta_{t+1} \;=\; \rho\theta_t + \zeta_{t+1}, \qquad \zeta_{t+1} \sim N(0, {\sigma_\zeta}^2), \tag{C.1}$$

for $0 < \rho < 1$.

Perhaps the cost of production of the firm is related to the distance between an action a_{it} they choose and this state, $(a_{it} - \theta_t)^2$. The optimal choice of action each period would be to choose $a_{it} = E[\theta_t|\mathcal{I}_{it}]$. This would make the marginal cost the expected squared forecast error $(E[\theta_t|\mathcal{I}_{it}] - \theta_t)^2$, which is the definition of the conditional variance $V[\theta_t|\mathcal{I}_{it}]$.

The prior mean and variance are given by $E[\theta_t|\mathcal{I}_t]$ and $V[\theta_t|\mathcal{I}_t] := \eta_t^{-1}$, where $\mathcal{I}_t$ represents whatever information set the agent has at time t. We define η with the inverse because this lends itself to interpreting η_t as the amount of data. A lower variance estimate or more accurate estimate implies more data about θ_t.

Consider the variance of tomorrow's state, given today's data. Taking the variance of both sides of (C.1), we get

$$V[\theta_{t+1}|\mathcal{I}_t] = \rho^2\eta_t^{-1} \;+\; \sigma_\zeta^2. \tag{C.2}$$

This conditional variance is the expected squared forecast error: $V[\theta_{t+1}|\mathcal{I}_t] \equiv E[(\theta_{t+1} - E[\theta_{t+1}|\mathcal{I}_t])^2|\mathcal{I}_t]$. It reveals how inaccurate the firm's prediction is, or how poor or scarce their predictive data is. In Bayesian language, this is a prior variance of θ_{t+1}.

If the data used to forecast θ_{t+1} has normally distributed noise, then according to Bayes' Law, all newly-acquired data can be combined and represented as a signal about tomorrow's state $s_t = \theta_{t+1} + e_t$, with $e_t \sim N(0, \sigma_e^2/m_{it})$, where m_{it} is the number of new data points firm i observes at time t, each with precision ${\sigma_e}^{-2}$. The $t+1$ information set is equivalent to $\mathcal{I}_{t+1} = \{\mathcal{I}_t, s_t\}$, which is the information available today, plus the signal observed at the end of period t.

According to Bayes' law, combining a normal prior belief with a normal signal yield a posterior precision that is the prior precision (the inverse of equation (C.2)), plus the precision of the new data ${\sigma_e}^{-2}m_{it}$:

$$\eta_{t+1} \;=\; (\rho^2\eta_t^{-1} \;+\; \sigma_\zeta^2)^{-1} \;+\; {\sigma_s}^{-2}. \tag{C.3}$$

This law of motion for the amount of data says that we take the existing stock of data η_t, depreciate it by transforming it into $(\rho^2\eta_t^{-1} + \sigma_\zeta^2)^{-1}$ and then add on the precision of newly-acquired data. This is similar to a law of motion for a stock of capital: $k_{t+1} = (1-\psi)k_t + i_t$, where i_t is new investment. For data that

predicts a persistent process, the depreciation rate is

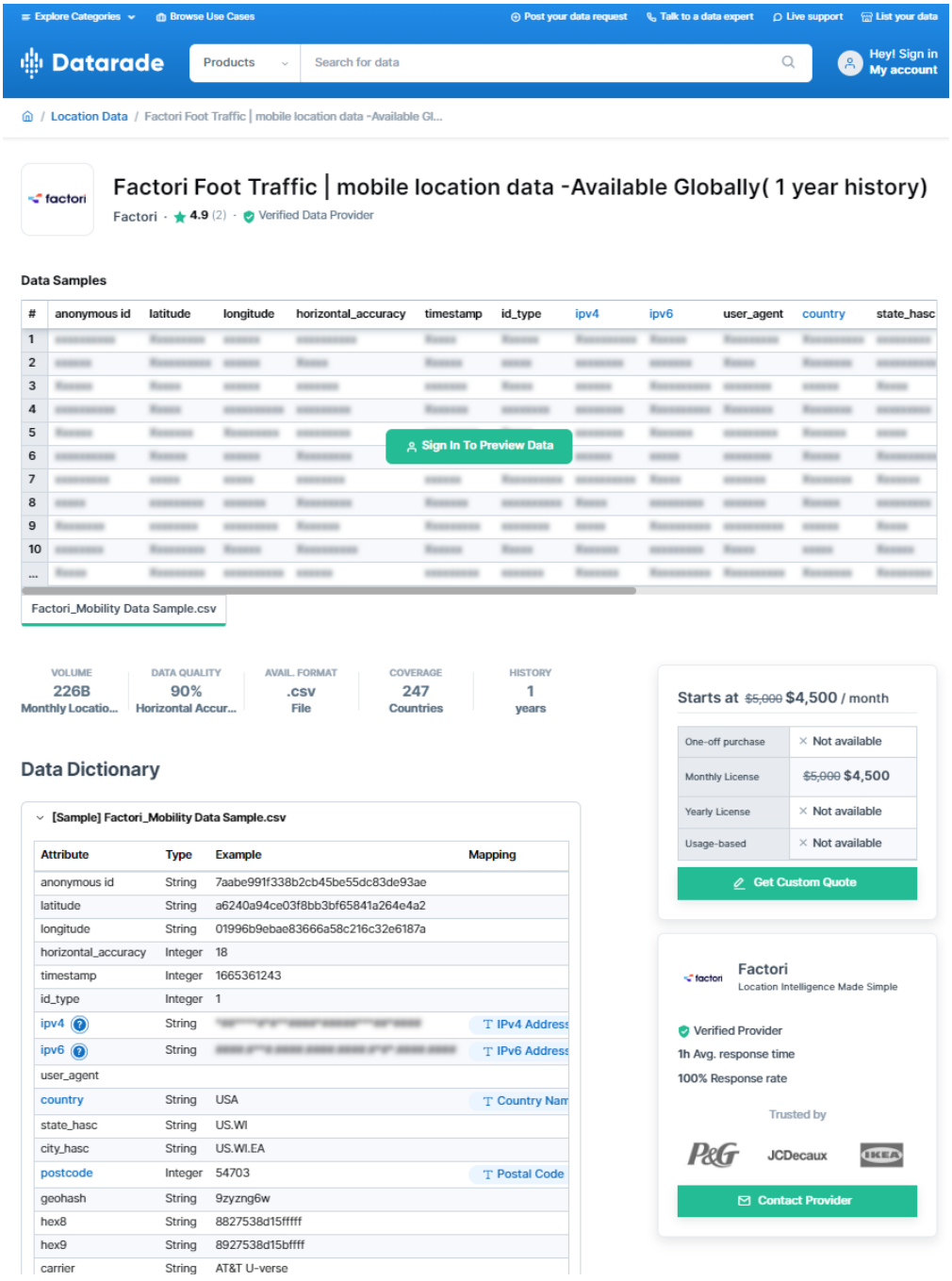
$$\psi_t \;=\; 1 - \frac{1}{\rho^2 + \sigma_\zeta^2 \eta_t}. \tag{C.4}$$

Note that if the AR(1) process is highly volatile (high σ_ζ), then the amount of data will depreciate quickly. Data about yesterday's state is less relevant to today's state because the state is changing quickly. This is the basis for our use of sector volatility as a proxy for data depreciation.

D Data Appendix

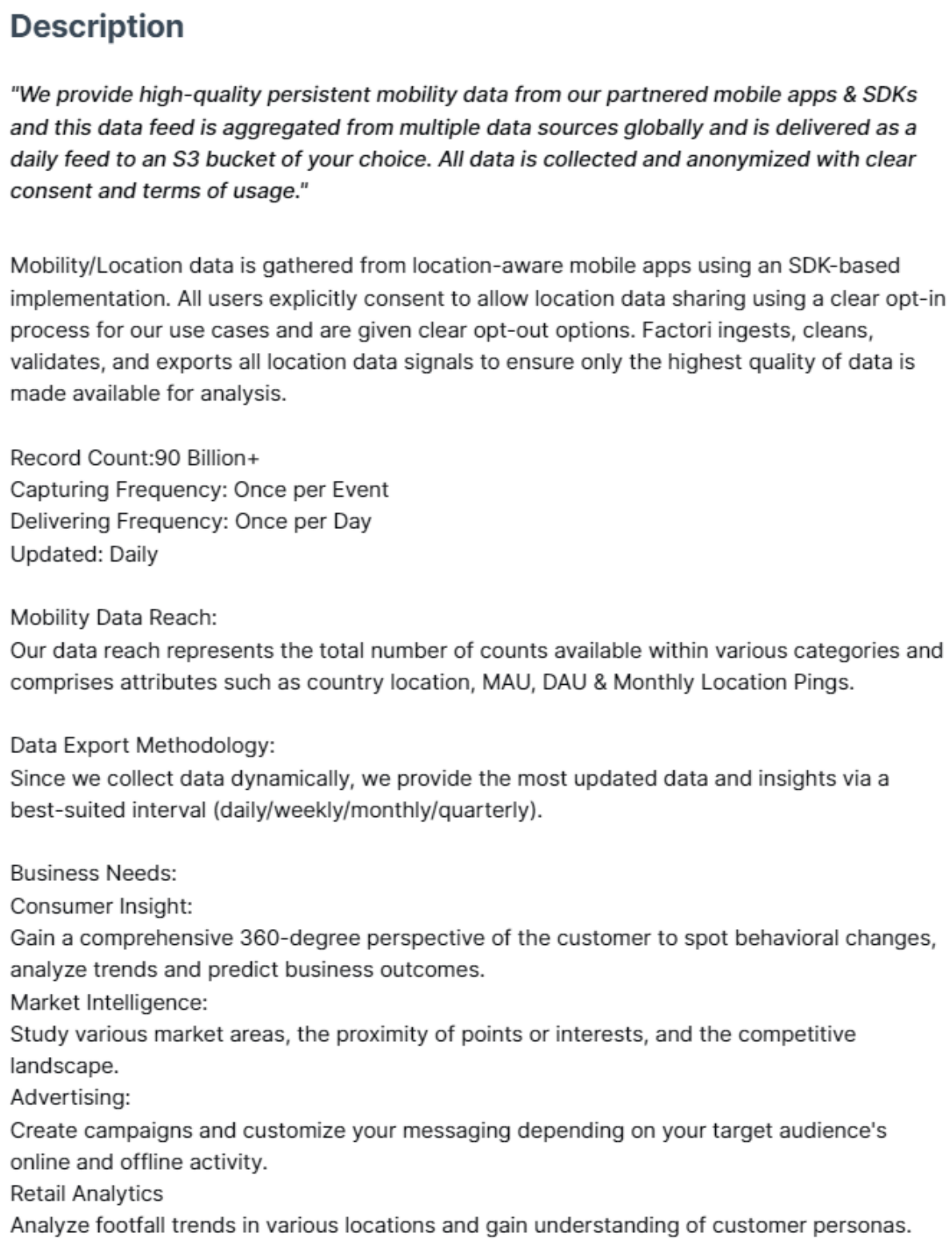
This appendix provides a richer description of our data sets. It visually illustrates the data product page, describe the data topics, the industry and geographical locations of data providers and the categories of business data.

Figure A.1: Examples of Datarade Product Page



Notes. This is a screenshot of a data product hosted on Datarade, URL address: https://datarade.ai/data-products/lifesight-foot-traffic-data-global-mobile-location-data-2-lifesight, retrieved in March, 2023.

Figure A.2: Examples of Datarade Product Page: Data Description



Description

"We provide high-quality persistent mobility data from our partnered mobile apps & SDKs and this data feed is aggregated from multiple data sources globally and is delivered as a daily feed to an S3 bucket of your choice. All data is collected and anonymized with clear consent and terms of usage."

Mobility/Location data is gathered from location-aware mobile apps using an SDK-based implementation. All users explicitly consent to allow location data sharing using a clear opt-in process for our use cases and are given clear opt-out options. Factori ingests, cleans, validates, and exports all location data signals to ensure only the highest quality of data is made available for analysis.

Record Count:90 Billion+
Capturing Frequency: Once per Event
Delivering Frequency: Once per Day
Updated: Daily

Mobility Data Reach:
Our data reach represents the total number of counts available within various categories and comprises attributes such as country location, MAU, DAU & Monthly Location Pings.

Data Export Methodology:
Since we collect data dynamically, we provide the most updated data and insights via a best-suited interval (daily/weekly/monthly/quarterly).

Business Needs:
Consumer Insight:
Gain a comprehensive 360-degree perspective of the customer to spot behavioral changes, analyze trends and predict business outcomes.
Market Intelligence:
Study various market areas, the proximity of points or interests, and the competitive landscape.
Advertising:
Create campaigns and customize your messaging depending on your target audience's online and offline activity.
Retail Analytics
Analyze footfall trends in various locations and gain understanding of customer personas.

Notes. This is a screenshot of a data product hosted on Datarade, URL address: https://datarade.ai/data-products/lifesight-foot-traffic-data-global-mobile-location-data-2-lifesight.

Table A.1: Mapping Between Broader Data Categories and Original Fine Data Categories

Broad Data Category	**Example Original Data Categories** (Up to 5 Example Categories Are Presented)
Business Contact Data	B2B Contact Data, B2B Leads Data, B2B Marketing Data, B2B Email Data, B2B Decision Maker Data
General Company Data	Company Data, Firmographic Data, Business Website Data, Company Address Data, SIC Data
Point of Interest Data	Global POI Data, Point of Interest POI Data, Business Location Data, Business Listings Data, Retail Location Data
Property, Real Estate Data	Property Data, Property Owner Data, Residential Property Data, Commercial Property Data, Property Listings Data
Weather and Geo Data	Location Data, Global Weather Data, Weather Data, Map Data, Local Weather Data
Consumer Research Data	B2B Buyer Intent Data, Demographic Data, Consumer Behavior Data, Consumer Marketing Data, Targeting Data
Financial Data	Forex Data, OTC Data, Core Financial Data, Stock Market Data, Cryptocurrency Data
ESG Data	ESG Data, Corporate Company ESG Data, ESG Equities Data, ESG Risk Data, Sustainability Data
Economic Research Data	Currency Data, Economic Data, Consumer Spending Data, Consumer Transaction Data, Research Data
Misc	Alternative Data, Infrastructure Data, Event Data, Reference Data, School Data
Internet, Cybersecurity, Software, and ML Data	AI ML Training Data, Technographic Data, Web Scraping Data, Machine Learning ML Data, Natural Language Processing NLP Data
Product Data	Ecommerce Data, Ecommerce Product Data, Amazon Data, Retail Data, Product Data
Healthcare and Medical Related Data	Healthcare Marketing Data, Healthcare Industry Leads Data, Consumer Healthcare Data, Healthcare Provider HCP Data, Insurance Data
Automobile, Aviation, Transportation, and Travel Data	Automotive Data, Mobility Data, Car Data, Vehicle Location Data, Electric Vehicle Charging Stations Data
Mobile Phone and Satelite Data	Phone Number Data, Mobile Location Data, Satellite Imagery Data, GPS Location Data, Mobile Device Location Data
Media Data	Sentiment Data, Social Media Data, News Data, Advertising Data, Brand Sentiment Data
Politics Data	Government Congressional Data, Public Sector Data, Lobbying Data, Political Risk Data, Campaign Election Data
Personal Identity Data	Consumer Identity Data, Consumer Identity Graph Data, Identity Graph Data, Individual Level Identity Data, Cross Device Identity Data
Job Posting and Employment Data	Job Postings Data, Employee Data, HR Data, Consumer Employment Data, Job Title Data
Legal and Litigation Data	Litigation Data, Intellectual Property Data, Patent Data, Court Data, Filings Data
Agricultural Data	Agricultural Data
Education Industry Data	Education Industry Data

Notes. This table presents the mapping between broad data categories (22) and the detailed original data categories (227). For each broad category, we keep up to 5 included detailed categories.