Opening the Black Box*

A Theory of the Value of Data

Giovanni Colla Rizzi†

November 5, 2025

Abstract

This paper develops a theory of the value of data for prediction purposes. An agent uses training data (a sample of covariates and the target variable) to learn about the parameters of a statistical model, and prediction data (covariates) about target individuals. The main findings are that: (i) training covariates exhibit economies of scope; (ii) training covariates and sample size are complements when data are scarce and otherwise substitutes; and (iii) training and prediction data are complements.

These patterns have several policy implications. First, data-driven acquisitions may lead to data concentration, all the more so when sample sizes are small due to, e.g., privacy regulation. Second, pooling covariates is always procompetitive, whereas pooling observations can be anticompetitive when data is abundant. Third, a data seller may profitably conclude exclusivity agreements with a firm selling predictions even if this reduces total welfare, especially when competition is fierce and datasets are rich in covariates.

JEL Classification: C11, D83, L12, O33.

Keywords: Value of data, Prediction, Economies of Scope, Data-driven Acquisitions, Data Pools.

*I thank Patrick Rey for his guidance, and Jean-Pierre Florens and Doh-Shin Jeon for their suggestions. I also thank Jad Beyhum, Michele Bisceglia, Zhijun Chen, Krishna Dasaratha, Alexandre de Cornière, Eric Gautier, Andrei Hagiu, Johannes Hörner, Marco Iansiti, Bruno Jullien, Hiroaki Kaido, Simon Loertscher, Friedrich Lucke, Leonardo Madio, Giovanni Morzenti, Juan Ortner, Christoph Reidl, Andrew Rhodes, David Salant, Maximilian Schaefer, Sara Shahanaghi, Tim Simcoe, Alex Smolin, Emanuele Tarantino, Ehsan Valavi, Marshall Van Alstyne, Davide Viviano, Julian Wright and the participants at the European Association of Industrial Economics Conference (Valencia, 2025), Questrom Digital Platform Seminars, TSE and Boston University.

†Toulouse School of Economics, University of Toulouse Capitole, France. E-mail: giovanni.rizzi@tse-fr.eu

1 Introduction

Data is a key source of competitive advantage in digital markets, as it allows firms to make better predictions: Google and Meta predict ad clicks, Amazon and Uber predict demand, and Spotify and Netflix predict content choices. Over the last decade, a small number of firms have come to dominate these markets, raising concerns that data concentration creates barriers to entry. However, major policy reports[1] and the EU Data Act proposal[2] warn that data concentration can restrict competition. The 2021 U.S. House Report similarly notes that "data advantages [...] can reinforce dominance and serve as a barrier to entry." [3]

Initially, industry leaders emphasized the strategic importance of scale, with former Google CEO Eric Schmidt noting in 2009 that "Scale is the key. We just have so much scale in terms of the data we can bring to bear".[4] By contrast, more recently tech firms have invoked the Law of Large Numbers to argue that additional data improve predictions less and less as datasets grow, implying data have diminishing returns. In that case, the concentration of data in the hands of few firms reflects their superior technology rather than the presence of entry barriers.[5]

To assess the merits of these arguments, we must understand whether there are economies of scope and scale of data. To this end, I develop a *theory of the value of data* for the purpose of prediction. The framework is based on a statistical model in which a target variable is the outcome of a linear process of infinitely many covariates with heterogeneous, each of which has an unknown effect on the target variable, captured by a linear parameter. Critically, it distinguishes between the value of additional *observations* (e.g., individuals) and additional *covariates* (e.g., attributes of those individuals), as well as between *training covariates*, used to train algorithms, and *prediction covariates*, used to apply trained algorithms to predict outcomes (e.g., willingness to pay) for specific individuals.

Specifically, I set up a data collection problem in which an agent must choose how many observations and which training and prediction covariates to observe. I first characterize the optimal predictor and show it can be interpreted as a ridge estimator. Building on this, I then derive closed-form expressions for the value of data, showing that returns depend on the distribution of the variance across covariates. This generates three main insights on the economies of scope and scale of data.

Firstly, there are always *economies of scope in training covariates.* This is because collecting

[1]See the Stigler Committee's Final Report on Digital Platforms and the UK Competition and Markets Authority's Digital Competition Expert Panel Report.

[2]See `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52023PC0193`.

[3]U.S. House of Representatives, Committee on the Judiciary, Subcommittee on Antitrust, Commercial and Administrative Law (2020), "Investigation of Competition in Digital Markets," Committee Print 117-40, at 36-38. See the House Committee Print: `https://www.congress.gov/committee-print/117th-congress/house-committee-print/47832`.

[4]See `https://www.bloomberg.com/news/articles/2009-10-02/how-google-plans-to-stay-ahead-in-search`.

[5]See Varian (2018) and Bajari et al. (2019).

a new covariate reduces prediction noise on the estimates of all the parameters of the other covariates and reductions in noise have an accelerating effect on the precision of estimates. As long as covariates have similar variance, this has stark implications: returns to covariates are always increasing. However, when the distribution of variance across covariates is very concentrated, decreasing marginal variance deriving from the covariate selection implies diminishing returns may dominate.

Secondly, *training covariates and observations are complements when data is scarce and substitutes when it is abundant.* In the former case, collecting more covariates makes each additional observation more valuable: since both dimensions reduce noise, which has an accelerating effect on the precision of estimates, an additional observation will be more valuable if an additional covariate is observed (and vice versa). However, once datasets become large, they become substitutes: more data has a diminishing impact on the noise reduction.

Finally, *training data and prediction covariates are complements.* This is because adding covariates or observation to the training dataset reduces estimation noise, which improves the value of prediction covariates. Intuitively, the data an app collects on its own users becomes more valuable when it is embedded in a larger ecosystem with richer data, since the broader dataset improves the precision of parameter estimates that make those individual covariates informative.

I then explore the implications of the insights above for firms' data collection strategies. Firms must collect a minimum scale of data before prediction is profitable, so there are sunk costs. In the early stages, firms should balance marketing (acquiring users/observations) and product development (collecting attributes/covariates) to exploit complementarities between observations and covariates. However, once datasets are large enough, firms should specialize either marketing or product development, depending on where marginal returns are highest. Since fragmented datasets reduce predictive accuracy, integrating user profiles is essential—especially for scale-ups with rich user profiles and mid-sized user bases.

Finally, to study the policy implications, I develop three applications. First, I study a market where buyers purchase predictions whose value depends on the informativeness of a seller's dataset. An incumbent (Google) and an entrant (Fitbit) each own proprietary covariates and sell to separate markets. and can sell predictions separately or merge their data through acquisition. Even if their demands are unrelated Google may find it profitable to purchase Fitbit, all the more so if observations are scarce and covariates are abundant. In this case, blocking the acquisition may reduce total surplus by increasing the cost of a given reduction of prediction error. Furthermore, by limiting the number of observations, privacy regulation makes data concentration more likely. This effect gives rise to a policy trilemma: regulators can at most achieve two of three objectives — privacy, competition, or efficiency; as a result, decentralizing data may exacerbate the trade-off between privacy and efficiency. Ex-ante access regulation, such as federated learning or regulated Application Programming Interfaces

(APIs) for training data, may constitute a more promising avenue.[6]

Second, sharing covariates on the same individuals among data brokers eliminates double marginalization, generating efficiencies. By contrast, sharing the same covariates on different individuals may facilitate collusion when the benefits from eliminating double marginalization are small, all the more so if datasets are large.

Third, data sellers may find it profitable to conclude exclusive licensing agreements with firms selling predictions, such as the 2024 Reddit–OpenAI deal, to commit not to engage in opportunistic behavior. However, when firms complement seller data with their own proprietary data, such agreements may harm social welfare by disincentivizing investment in proprietary data by the excluded prediction firm, all the more so datasets are large and competition between prediction firms is fierce.

Related Literature. There is a rich information design literature on the value of data (Jones and Tonetti (2020), Bergemann, Bonatti, and Gan (2022), Bergemann and Bonatti (2024), and Acemoğlu et al. (2022)). While this literature values information through the choice of a probability distribution, I link the value of data directly to the realization of a dataset, illuminating the link between statistical properties of sets of random variable and their economic value. Methodologically, my work is related to Montiel Olea et al. (2022), Iyer and Ke (2024), and Dasaratha, Ortner, and Zhu (2025), who analyze competition between models with different covariates. In contrast, I jointly model covariates and observations, and distinguish training from prediction data, which allows me to derive structural non-convexities generating increasing returns and complementarities.

I also contribute to the broad literature on economies of scope and scale to data. Whereas most models fix covariates and study returns to observations as Bajari et al. (2019) and Goldfarb and Tucker (2011), my framework endogenizes covariate collection. This extension rationalizes empirical findings on complementarities in Schaefer and Sapi (2023) and economies of scope in Carballa-Smichowski, Duch-Brown, et al. (2025). Schaefer (2025) develops a complementary frequentist approach and shows that the distribution of covariates shapes returns to scale. Allcott et al. (2025) run a structural model to estimate returns to scale of additional observations in search. They finding diminishing returns and evidence of limited complementarities across different queries. Radner and Stiglitz (1984) attributes increasing returns to information costs, while I show they can emerge independently of costs.

My work provides microfoundations for two strands of literature that take increasing returns to data as assumptions: the IO literature on platforms and the macroeconomics literature on data as a production input. Prior work explains increasing returns through feedback between data and demand (Hagiu and Wright (2023), Prüfer and Schottmüller (2021), Farboodi

[6] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonized rules on fair access to and use of data ("Data Act"), esp. Chapter IV (Articles 30-34), which require that compensation for data access be "fair, reasonable and non-discriminatory".

and Veldkamp (2025), Aral, Brynjolfsson, and Wu (2008), and Cong, He, and Yu (2021)) or by assuming complementarities across datasets Carballa-Smichowski, Lefouili, et al. (2025), De Corniere and Taylor (2025), and Calzolari, Cheysson, and Rovatti (2025). I show instead that prediction accuracy alone generates increasing returns due to the statistical structure of data, independent of demand feedback.

The applications of the model contribute more broadly to the IO literature on digital markets. De Corniere and Taylor (2025) shows that the pro- or anti-competitive effect of collecting more data only depends on the supply of data-driven services rather than its demand, implying that our supply-side model of prediction is in many cases sufficient to characterize the welfare implications of more data collection. Furthermore, Cornière and Taylor (2024) studies data-driven mergers developing a theory of harm of mergers which rely on cross-market effects. Both works complement the applications of my paper as they treat data as an undifferentiated good to which my results readily apply. Several papers on the economics of patents can be applied to datasets and motivate the application of my model to data pools and exclusivity in data sale. Lerner and Tirole (2004) deals with complementarity/substitutability of patents and the private and social value of commercializing them jointly in pools. We use it to characterize when data pooling by brokers is anticompetitive. Gu, Madio, and Reggiani (2021) also study broker pools in a context without double marginalization and conclude they are anticompetitive when datasets are substitutes. Katz and Shapiro (1986) show exclusive deals are optimal when selling patents to competing firms and this results in a suboptimal dissemination of patents. Aghion and Bolton (1987) show that exclusive contracts can serve as a strategic commitment device that deters entry by raising rivals' costs, allowing incumbents to extract rents from buyers at the expense of social welfare. My model extends insights from both papers to non-rival data markets with complementarities and endogenous investment in proprietary data.

Finally, I develop a simple framework to study scaling laws, shedding light on the phenomenon of double descent, i.e., that maximum likelihood-based algorithms generalize well even when overparametrized as explored in Hastie et al. (2020), Nakkiran et al. (2021), and Belkin et al. (2019).

2 Model Setup

This section models a firm that predicts an individual's target variable from covariates (individual attributes). The firm faces dimensionality constraints in both training and prediction stages and aims to maximize prediction accuracy subject to these constraints.

2.1 Data-Generating Process

A firm must predict a random *target variable* $y \in \mathbb{R}$ for a *target individual* drawn from a population $\mathcal{I}$. For each $i \in \mathcal{I}$, the target variable depends on a countably infinite set of *covariates*:

$$y^i = \sum_{j \in \mathbb{N}} \beta_j x_j^i,$$

where $x_j^i \in \mathbb{R}$.

Covariates Covariates are mutually independent across $k \in \mathbb{N}$ and i.i.d. across individuals $i \in \mathcal{I}$:

$$x_k^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, s_k),$$

where $s_k > 0$ is the signal of covariate k. [7]

Parameters Parameters are unknown, independent of x_k^i, and mutually independent across $k \in \mathcal{K}$:[8]

$$\beta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

Cumulative Variance For any subset of covariates $\mathcal{K} \subseteq \mathbb{N}$, define the ***cumulative signal***

$$S(\mathcal{K}) \equiv \sum_{k \in \mathcal{K}} s_k,$$

where $S(\varnothing) = 0$. We normalize $\text{Var}[y] = 1$, so since β_k are independent of x_j^i for all $k, j \in \mathbb{N}$ and since x_j^i are mutually independent, it follows that $S(\mathbb{N}) = 1$ and $S : 2^{\mathbb{N}} \to [0, 1]$.

We denote the covariate vector by $\boldsymbol{x}^i \equiv (x_k^i)_{k \in \mathbb{N}}$ and the parameter vector by $\boldsymbol{\beta}^i \equiv (\beta_k^i)_{k \in \mathbb{N}}$. For any vector $\boldsymbol{v} \in \mathbb{R}^{|\mathbb{N}|}$ and subset $\mathcal{K} \subseteq \mathbb{N}$, let $\boldsymbol{v}_{\mathcal{K}} \equiv (v_k)_{k \in \mathcal{K}}$.

2.2 Training and Prediction Datasets

Conditional on knowing $\boldsymbol{\beta}$, the target variable is independently distributed across individuals. This implies that we can distinguish a *training* step in which the firm learns about $\boldsymbol{\beta}$ and a *prediction* step in which it applies that knowledge to make predictions on a specific individual.

Training Data Before predicting, the firm may observe a set of ***training covariates*** $\mathcal{T} \subseteq \mathbb{N}$, for a sample of n individuals in $\mathcal{I}$, which constitutes the ***training data*** which is a matrix

$$\boldsymbol{M}_{\mathcal{T}}^n \equiv \left\{ (y^i, \boldsymbol{x}_{\mathcal{T}}^i) \right\}_{i=1}^n = \begin{pmatrix} \boldsymbol{y} & \boldsymbol{X}_{\mathcal{T}} \end{pmatrix}, \qquad \boldsymbol{y} \in \mathbb{R}^n, \; \boldsymbol{X}_{\mathcal{T}} \in \mathbb{R}^{n \times t},$$

[7]In Appendix B we relax the independence assumption by allowing correlated covariates. We show that, after applying Principal Component Analysis to orthogonalize the covariate space, all the results carry through identically.

[8]Normalizing $\text{Var}[\beta_k] = 1$ is WLOG, as any $\text{Var}[\beta_k] = \tau^2$ can be recovered by rescaling $\tilde{s}_k = \tau^2 s_k$.

where $t \equiv |\mathcal{T}|$.

Prediction Data Then the firm will collect a set $\mathcal{P} \subseteq \mathbb{N}$ of ***prediction covariates*** on the target individual. Their realization is the ***prediction data*** which is a vector

$$\boldsymbol{x}_{\mathcal{P}} \in \mathbb{R}^{p},$$

where $p \equiv |\mathcal{P}|$.

Prediction Models

Definition 1. The training covariates and the prediction covariates $(\mathcal{T}, \mathcal{P})$ determine the ***prediction model***.

A prediction model $(\mathcal{T}, \mathcal{P})$ induces a random variable whose realization $(\boldsymbol{M}^{n}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}})$ is a ***dataset of type*** $(\mathcal{T}, \mathcal{P})$.

Definition 2. An ***dataset of type*** $(\mathcal{T}, \mathcal{P})$ is a collection of a training data and a prediction data

$$\boldsymbol{D}^{n}_{\mathcal{T},\mathcal{P}} \equiv (\boldsymbol{M}^{n}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}) \in \mathcal{D}^{n}_{\mathcal{T},\mathcal{P}} \equiv \mathbb{R}^{n\times(1+t)} \times \mathbb{R}^{p}.$$

2.3 Firm's Problem

Given a prediction $\hat{y} \in \mathbb{R}$, the loss is the squared-error

$$L(y, \hat{y}) = (y - \hat{y})^{2}.$$

Taking N as given, the firm must solve

$$\min_{\mathcal{T},\mathcal{P},\hat{y}} L(y, \hat{y}) + C\left(\mathcal{T}, \mathcal{P}, n\right).$$

We decompose the problem in two sequential subproblems, a covariate selection problem in which the firm chooses $(\mathcal{T}, \mathcal{P})$ and an inference problem in which the firm chooses the optimal prediction $\hat{y}$ given two fixed covariate sets. As customary we solve the problem backwards starting with the inference problem.

2.3.1 Inference Problem

Given prediction model $(\mathcal{T}, \mathcal{P})$ and a sample size n, the firm must choose how to map a generic dataset $\boldsymbol{D} \in \mathcal{D}$ into predictions. The optimal prediction conditional on a dataset will determine its value.

Definition 3 (Predictor and Posterior Risk)**.** A ***predictor*** is a measurable map $f : \mathcal{D} \to \mathbb{R}$ that produces a prediction

$$\hat{y} = f(\boldsymbol{D}).$$

Its ***posterior risk*** of a predictor f conditional on $\boldsymbol{D}$ is the expected prediction loss

$$R(f, \boldsymbol{D}) \equiv \mathbb{E}_{y|\boldsymbol{D}} \left[L\left(y, f(\boldsymbol{D})\right) \mid \boldsymbol{D} \right].$$

The firm will choose a predictor to minimize the posterior risk.

Problem 1 (Inference)**.** Given a dataset $\boldsymbol{D}$,

$$\min_{f : \mathcal{D} \to \mathbb{R}} R(f, \boldsymbol{D}).$$

Denote an ***optimal predictor*** by[9]

$$f^* \in \arg\min_f R(f, \boldsymbol{D}).$$

The ***posterior Bayes risk*** of $\boldsymbol{D}$ is the risk of the optimal predictor, viz. the minimum risk attainable after observing $\boldsymbol{D}$:

$$R^*(\boldsymbol{D}) \equiv R\left(f^*(\boldsymbol{D}), \boldsymbol{D}\right).$$

The ***prior Bayes risk*** is the minimum risk attainable without observing any data:

$$R^*(\varnothing) = \mathrm{Var}[y] = 1,$$

where $\varnothing$ denotes the empty dataset.

The value of a dataset is the expected reduction of expected prediction error (the Bayes risk) when optimally using the dataset to make predictions relative to what could be achieved by making the best prior prediction:

Definition 4. The ***(prior) value of a dataset of type*** $(\mathcal{T}, \mathcal{P})$ ***and scale*** n is a function $V : \mathbb{R} \times 2^{\mathbb{N}} \times 2^{\mathbb{N}} \to [0, 1]$ defined as

$$V(n, \mathcal{T}, \mathcal{P}) \equiv 1 - \mathbb{E}_{\boldsymbol{D}^n_{\mathcal{T},\mathcal{P}}} \left[R^*(\boldsymbol{D}^n_{\mathcal{T},\mathcal{P}}) \right].$$

The best prior prediction, the prior mean $\hat{y} = 0$, gave a loss equal to the prior variance $\mathrm{Var}[y] = 1$.

[9]Convexity of the quadratic loss function ensures its existence and uniqueness.

2.3.2 Model Selection Problem

Knowing the expected value of a dataset of a given type, the firm chooses the prediction model $(\mathcal{T}, \mathcal{P})$ and the sample size n.

Problem 2 (Model Selection)**.** The firm must choose the prediction model $(\mathcal{T}, \mathcal{P})$ and sample size n to solve

$$\sup_{n,\mathcal{T},\mathcal{P}} V(n, \mathcal{T}, \mathcal{P}) - C\left(n, \mathcal{T}, \mathcal{P}\right).$$

We will assume additively separable costs

$$C\left(n, \mathcal{T}, \mathcal{P}\right) = C_n(n) + C_t(t) + C_p(p),$$

where $t \equiv |\mathcal{T}|$ and $p \equiv |\mathcal{P}|$ so that the comparative statics do not depend on the properties of the value function.

3 Inference

In this section we characterize the optimal predictor for a given prediction model $(\mathcal{T}, \mathcal{P})$ and sample size n. This will allow us to characterize the value of a dataset in Theorem 2.

3.1 The Optimal Predictor

We recall a standard Bayesian result: the predictor minimizing expected squared error is the posterior mean.

Lemma 1 (Optimal Predictor)**.** *The optimal predictor is*

$$f^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}}) = \mathbb{E}[y \mid \boldsymbol{D}_{\mathcal{T},\mathcal{P}}] = \boldsymbol{x}'_{\mathcal{T}\cap\mathcal{P}}\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}\cap\mathcal{P}} \mid \boldsymbol{M}_{\mathcal{T}}\right].$$

The optimal predictor is a linear combination of the prediction covariates weighted by the posterior mean of their parameters. Because instances are independent conditional on $\boldsymbol{\beta}$, the training matrix $\boldsymbol{M}_{\mathcal{T}}$ influences predictions only through the posterior beliefs on $\boldsymbol{\beta}$. Untrained parameters have posterior mean zero, so the predictor only uses covariates whose parameter priors have been updated using $\boldsymbol{M}_{\mathcal{T}}$.

Corollary 1 (Prediction Requires Training)**.** *The firm should never use a prediction covariate whose parameter it has not updated:*

$$\mathcal{P}^* \subseteq \mathcal{T}^*.$$

As the prior mean of $\boldsymbol{\beta}$ is $\mathbf{0}$, the firm expects that any covariate $k \in \mathcal{P} \setminus \mathcal{T}$ will not affect y.[10]

Quasi-maximum Likelihood The firm will learn the data-generating process using a regression of the target variable on a set $\mathcal{T}$ of covariates:[11]

$$y^i = \boldsymbol{\beta}'_{\mathcal{T}} \boldsymbol{x}^i_{\mathcal{T}} + \varepsilon^i_{\mathcal{T}}, \qquad \varepsilon^i_{\mathcal{T}} \equiv \boldsymbol{\beta}'_{\mathcal{T}^c} \boldsymbol{x}^i_{\mathcal{T}^c}. \tag{1}$$

Because covariates are mutually independent and $\mathrm{Var}[y] = 1$, it follows that, for all $i \in \mathcal{I}$,

$$\mathbb{E}[\varepsilon^i_{\mathcal{T}}] = 0, \qquad \mathrm{Var}(\varepsilon^i_{\mathcal{T}}) = 1 - S(\mathcal{T}). \tag{2}$$

Although $\varepsilon^i_{\mathcal{T}}$ is not Gaussian, the following assumption defines a working likelihood used to compute the posterior distribution of $\boldsymbol{\beta}_{\mathcal{T}}$.

Assumption 1 (*Gaussian approximation*). *We approximate the misspecification error as an i.i.d. Gaussian:*

$$\varepsilon^i_{\mathcal{T}} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1 - S(\mathcal{T})). \tag{3}$$

Assumption 1 therefore entails no loss of generality for the results that follow.[12]

Under 1, the prediction model is a generalization of the classic Bayesian linear model analyzed in DeGroot (2005) and Berger (1990), with an endogenous set of covariates $\mathcal{T}$ affecting its misspecification error. This training stage will generate some parameter estimates $\hat{\boldsymbol{\beta}}_{\mathcal{T}}$, which are estimates for the effects of the covariates on the target variable.

By Lemma 1, the training matrix affects prediction exclusively through the posterior mean of the parameters, which we call the ***Bayes Estimator***.

Proposition 1 (Optimal Predictor). *The Bayes Estimator is the posterior mean $\boldsymbol{\beta}$ and satisfies:*

1. *For untrained parameters:*

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}^c} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \mathbf{0}_{|\mathbb{N}|-t};$$

2. *For trained parameters:*

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \left(\boldsymbol{X}'_{\mathcal{T}} \boldsymbol{X}_{\mathcal{T}} + (1 - S(\mathcal{T})) \cdot \boldsymbol{I}_t\right)^{-1} \boldsymbol{X}'_{\mathcal{T}} \boldsymbol{y}.$$

[10]This result relies on uninformative priors. If priors are informative, i.e.,$\mathbb{E}\boldsymbol{\beta} \neq \mathbf{0}$, the firm could find it optimal to predict using covariates it has not trained.

[11]For any subset $\mathcal{S} \subseteq \mathcal{K}$, we denote its complement by $\mathcal{S}^c \equiv \mathcal{K} \setminus \mathcal{S}$.

[12]Following the quasi-maximum-likelihood (QML) framework of Gourieroux, Monfort, and Trognon (1984), Bollerslev and Wooldridge (1992) and White (1982), consistency of the corresponding estimators—and hence of the moments that enter the Bayesian updating rule—requires only that the conditional mean and variance be correctly specified. When these first two moments are correct, the Gaussian likelihood is moment-equivalent to the true sampling model. Because all subsequent analytical results depend only on the posterior mean and variance—quantities determined by the first two moments—the analysis remains valid for any distribution of with matching first and second moments.

Because parameters are independent, training $\boldsymbol{\beta}_{\mathcal{T}}$ provides no information on the untrained parameters $\boldsymbol{\beta}_{\mathcal{T}^c}$, whose prior mean is $\mathbf{0}_{|\mathbb{N}|-t|}$.

Ridge Regression Interpretation It is well known in the Bayesian statistics literature (see DeGroot (2005)) that estimators like that in Proposition 1 have a frequentist counterpart in the ridge regression estimator defined as:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}_{\boldsymbol{M}_{\mathcal{T}}}(\lambda) \equiv \arg\min_{\boldsymbol{b}\in\mathbb{R}^t} \left\{ \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_2^2 \right\},$$

where $\lambda \geq 0$ is a penalty for the squared Euclidean distance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}_{\mathcal{T}}}$ from the origin. In practice, λ is typically chosen by cross-validation to minimize the prediction error. The following result bridges the gap between the practical applications and our theoretical results by establishing a link between the set of training covariates $\mathcal{T}$ and the regularization λ implied by the prior.

Corollary 2 (Ridge Estimator Interpretation)**.** *The posterior mean coincides with a ridge regression estimator with regularization*

$$\lambda(n, \mathcal{T}) \equiv \frac{1 - S(\mathcal{T})}{n}.$$

Equivalently,

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \hat{\boldsymbol{\beta}}^{ridge}_{\boldsymbol{M}_{\mathcal{T}}}(\lambda)\Big|_{\lambda=\lambda(n,\mathcal{T})}.$$

Lindley and Smith (1972) establishes that the regularization λ is the ratio of the residual variance to the parameter variance (which is $\text{Var}[\beta_j] = 1$). By putting structure on the regression residual, we can study the dependence of λ on the training covariates $\mathcal{T}$.

Remark 1. Note that, contrary to the common presumption of statistics practitioners, the cumulative rate of regularization defined as $\Lambda \equiv t\lambda(n, \mathcal{T})$ is not increasing in the number of parameters t. Specifically, adding covariate k to a set of covariates $\mathcal{T}$ decreases the overall regularization if and only if

$$\lambda(n, \mathcal{T} \cup \{i\}) \leq \lambda(n, \mathcal{T}) \iff t \geq \frac{1 - S(\mathcal{T} \cup \{i\})}{s_i}.$$

The intuition is that if there are many training covariates, the benefit from the reduction of s_j on each of their regularization coefficients is larger than the cost of increasing the noise of $1 - S(\mathcal{T} \cup \{i\})$ by requiring the estimate of parameter k.

3.2 The Value of a Dataset

The quadratic loss implies that the value of a dataset is the variance of the posterior mean. We define the weight of covariate k as

$$w_k(\lambda) \equiv \mathrm{Var}_{\boldsymbol{M}_{\mathcal{T}}}\left[\hat{\boldsymbol{\beta}}^{\text{ridge}}_{\boldsymbol{M}_{\mathcal{T}}}(\lambda)\right].$$

Lemma 2. *Assume $\mathcal{P} \subseteq \mathcal{T}$. The value of a dataset of covariates $(\mathcal{T}, \mathcal{P})$ is the variance of the optimal predictor*

$$V(n, \mathcal{T}, \mathcal{P}) = Var_{\boldsymbol{D}^n_{\mathcal{T},\mathcal{P}}}\left[f^*(\boldsymbol{D}^n_{\mathcal{T},\mathcal{P}})\right] = \sum_{k \in \mathcal{P} \cap \mathcal{T}} s_k \; w_k(\lambda)|_{\lambda = \lambda(n,\mathcal{T})}.$$

The value decomposes additively across prediction covariates and, for each prediction covariate, multiplicatively into:

1. A *signal* term $s_k \equiv \mathrm{Var}(x_k)$, and
2. A *parameter training* term $w_k(\lambda)$, measuring how sensitive the posterior mean of β_k is to the training data.

With no training data, $\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}] = 0$ a.s., so the sample variance is $w_k(\lambda) = 0$, and the value is zero. With infinitely informative data, $\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}] \to \beta_k$, so $w_k(\lambda) \to 1$, since $\mathrm{Var}(\beta_k) = 1$; hence the maximal contribution of a prediction covariate k is s_k. Thus $V(n, \mathcal{T}, \mathcal{P})$ increases both when we predict using high-variance covariates and when $\boldsymbol{M}_{\mathcal{T}}$ is more informative about their parameters.

Lemma 3. *[House Party Effect] The posterior mean of β_k satisfies*

$$w_k(\lambda) = \begin{cases} 0, & j \in \mathcal{T}^c, \\ \dfrac{1}{1 + \dfrac{\lambda}{s_k}} + O\left(\sqrt{\frac{t}{n}}\right), & j \in \mathcal{T}. \end{cases}$$

The variance of the Bayes estimator is increasing in s_k, since if the covariate is more variable, a greater fraction of the variance along its direction is due to its parameter rather than the regression residual. Furthermore, it is decreasing in the penalty λ: the penalty pulls the estimates towards the prior mean, which 0, thereby reducing the variance. [13]Furthermore, reductions in the penalty λ affect the parameters of all prediction covariates: training covariates are non-rival, as each covariate contributes to better estimates of all the prediction parameters, independently from how many prediction parameters are affected.

Lemmas 2 and 3 allow me to characterize the value of a dataset of a given type.

[13]Note that $w_k(\lambda)$ coincides with the "data depreciation rate" in Section 2.1 of Farboodi and Veldkamp (2025), up to a reparametrization.

Proposition 2 (Value of a Dataset)**.** *The value of a dataset of n observations and covariates* $(\mathcal{T}, \mathcal{P})$ *is*

$$V(n, \mathcal{T}, \mathcal{P}) = \sum_{k \in \mathcal{P} \cap \mathcal{T}} \frac{s_k}{\frac{\lambda(n,\mathcal{T})}{s_k} + 1} + O\left(\sqrt{\frac{t}{n}}\right),$$

where

$$\lambda(n, \mathcal{T}) \equiv \frac{1 - S(\mathcal{T})}{n}.$$

The asymptotic term $O\left(\frac{t}{n}\right)$ vanishes provided that the parameter number t grows slower than the sample size n.

Assumption 2. *We assume* $\frac{t}{n} \to 0$ *so as to eliminate the asymptotic term.*

The intuition is that, provided the dimensionality does not explode relative to the sample size, the empirical covariance matrix converges to the population covariance matrix. This assumption rules out overparameterized regimes.[14]

3.3 Properties of Data

3.3.1 Diminishing Returns to n

The following results establish that training observations n yield positive but diminishing returns.

Corollary 3 (Diminishing Return to Observations)**.** *The value of a dataset is strictly increasing and strictly concave in n.*

A larger training sample decreases the penalty λ, thereby increasing the variance of the estimators of all prediction covariates. The gains, however, decline with sample size. This is due to the Law of Large Numbers: since the penalty $\lambda(n, \mathcal{T})$ decreases with n but is bounded below by zero, it is convex. Each additional observation thus eliminates less residual uncertainty, yielding diminishing returns. This finding is consistent with Goldfarb and Tucker (2011), Bajari et al. (2019) and Schaefer and Sapi (2023).

3.3.2 Economies of Scope in Training

Equipped with the formula in Theorem 2 we can now study whether data combination increases the marginal value of data or not. Proposition 3 implies that merging two datasets of the same covariates but different observations and the same covariates *never generates complementarities across datasets*: the merged dataset has less value than the sum of its parts.

[14]This assumption lets us avoid random-matrix tools (e.g., Marchenko–Pastur limits for the empirical spectral distribution) needed when $t/n \to \gamma > 0$. The payoff of this assumption is that we can derive closed-form expressions while allowing heterogeneous variances $\{s_k\}$. By contrast, in the high-dimensional regime $t/n \to \gamma \in (0, \infty)$, explicit formulas are typically available only under homoskedastic designs; with heteroskedasticity, one usually solves fixed-point equations numerically. For a version with $t/n \to \gamma > 0$ under homoskedasticity, see Appendix B.

We define the ***marginal contribution of covariate*** k ***to set*** $\mathcal{K} \in \{\mathcal{T}, \mathcal{P}\}$

$$V(\mathcal{K} \cup \{i\}) - V(\mathcal{K}).$$

We will say that the value of the dataset is ***increasing*** in the training covariates $\mathcal{T} \subseteq \mathbb{N}$ if the marginal contribution of any covariate $k \in \mathbb{N}$ is positive.

The change marginal contribution of covariate k to set $\mathcal{K}$ brought forth by adding covariate j to $\mathcal{K} \in \{\mathcal{T}, \mathcal{P}\}$ is

$$V(\mathcal{K} \cup \{i, j\}) - V(\mathcal{K} \cup \{j\}) - (V(\mathcal{K} \cup \{i\}) - V(\mathcal{K}))$$

We will say that the value of the dataset is ***supermodular (additive)*** in the covariates $\mathcal{K} \subseteq \mathbb{N}$ if the marginal contribution of any covariate $k \in \mathbb{N}$ is strictly increasing (constant) in any covariate $j \neq i$.

What happens if instead we merge two datasets on the same individuals but datasets that have different covariates? To answer this question, assume for simplicity $\mathcal{T} = \mathcal{P}$, as is the case in most applied settings.

Proposition 3 (Economises of Scope in Training)**.** *The value of a dataset is strictly increasing and supermodular in* $\mathcal{T}$*.*

Therefore, merging two datasets of different covariates on the same individuals results in synergies: merging covariates adds value, and more than the sum of their parts.

Proposition 4 (Additivity in Prediction)**.** *The value of a dataset is strictly increasing and additive in* $\mathcal{P}$*.*

3.3.3 Complementarity

We will say that the sample size and a set $\mathcal{K} \in \{\mathcal{T}, \mathcal{P}\}$ of covariates are ***complements (substitutes)*** if

$$V(n+1, \mathcal{K} \cup \{k\}) - V(n+1, \mathcal{K}) - (V(n, \mathcal{K} \cup \{k\}) - V(n, \mathcal{K})),$$

is positive (negative), meaning that that increasing the sample size by a unit increases the marginal contribution of covariate k.

We will say that the training and prediction covariates are ***complements*** if

$$V(\mathcal{K} \cup \{k\}, \mathcal{P} \cup \{j\}) - V(\mathcal{K}, \mathcal{P} \cup \{j\}) - (V(\mathcal{K} \cup \{k\}, \mathcal{P}) - V(\mathcal{K}\mathcal{P})),$$

is positive, meaning that adding a prediction covariate j increases the marginal contribution of any training covariate $k \neq j$ (and vice versa).

Proposition 5 (Complementarity/Substitutability of n and Training Scope)**.** *Fix a training set $\mathcal{T}$ and prediction set $\mathcal{P} \subseteq \mathcal{T}$. For any covariate $k \notin \mathcal{T}$ with signal $s_k > 0$, define the cross-effect*

$$\Delta(n;\mathcal{T},k) := \partial_n V\big(n,\mathcal{T}\cup\{k\},\mathcal{P}\big) - \partial_n V\big(n,\mathcal{T},\mathcal{P}\big).$$

Then

$$\Delta(n;\mathcal{T},k) = s_k \sum_{j\in\mathcal{P}} \frac{s_j^2\Big((1-S(\mathcal{T}))(1-S(\mathcal{T})-s_k) - n^2 s_j^2\Big)}{\big(ns_j + 1 - S(\mathcal{T})\big)^2\big(ns_j + 1 - S(\mathcal{T}) - s_k\big)^2}. \tag{4}$$

In particular, there exist thresholds

$$\underline{n} = \frac{\sqrt{(1-S(\mathcal{T}))(1-S(\mathcal{T})-s_k)}}{\max_{j\in\mathcal{P}} s_j}, \qquad \overline{n} = \frac{\sqrt{(1-S(\mathcal{T}))(1-S(\mathcal{T})-s_k)}}{\min_{j\in\mathcal{P}} s_j},$$

such that

$$\Delta(n;\mathcal{T},k) > 0 \;\; \textit{for all} \;\; n < \underline{n} \quad \textit{and} \quad \Delta(n;\mathcal{T},k) < 0 \;\; \textit{for all} \;\; n > \overline{n}.$$

Hence, when data are scarce (small n) sample size and training scope are complements*, while for abundant data (large n) they are* substitutes.

Furthermore,

Proposition 6 (Sample Size and Prediction Scope are Complements)**.** *Fix a training set $\mathcal{T}$ and let $\mathcal{P} \subseteq \mathcal{T}$. For any $i \in \mathcal{T}\setminus\mathcal{P}$, consider the cross-effect*

$$\Delta^{\mathcal{P}}(n;\mathcal{T},i) := \partial_n V\big(n,\mathcal{T},\mathcal{P}\cup\{i\}\big) - \partial_n V\big(n,\mathcal{T},\mathcal{P}\big).$$

Then

$$\Delta^{\mathcal{P}}(n;\mathcal{T},i) = \frac{\lambda_{\mathcal{T}}}{n}\cdot\frac{1}{\left(1+\dfrac{\lambda_{\mathcal{T}}}{s_i}\right)^2} = \frac{(1-S(\mathcal{T}))\,s_i^2}{n^2\big(s_i + \frac{1-S(\mathcal{T})}{n}\big)^2} \;>\; 0 \quad \textit{for all } n \geq 1. \tag{5}$$

Hence sample size n and prediction scope (adding a used covariate i) are strict complements *for all n. Moreover, the complementarity strength is decreasing in n.*

Proposition 7 (Training–Prediction Cross-Effect)**.** *Fix $n \geq 1$ and a training set $\mathcal{T}$; let $\mathcal{P} \subseteq \mathcal{I}$ be the (possibly strict) set of prediction covariates. The value of a dataset is*

$$V(n,\mathcal{T},\mathcal{P}) = \sum_{j\in\mathcal{P}\cap\mathcal{T}} \frac{s_j}{1+\dfrac{\lambda_{\mathcal{T}}}{s_j}}, \qquad \lambda_{\mathcal{T}} := \frac{1-S(\mathcal{T})}{n}.$$

Pick indices $i \notin \mathcal{P}$ and $k \notin \mathcal{T}$, and define the discrete cross-effect of adding i to prediction and k

to training:

$$\Delta_{i,k} := \Big[V(n, \mathcal{T} \cup \{k\}, \mathcal{P} \cup \{i\}) - V(n, \mathcal{T} \cup \{k\}, \mathcal{P})\Big] - \Big[V(n, \mathcal{T}, \mathcal{P} \cup \{i\}) - V(n, \mathcal{T}, \mathcal{P})\Big].$$

Then:

1. *If $i = k$ and $i \notin \mathcal{T}$, the cross-effect is strictly positive:*

$$\Delta_{i,i} = \frac{s_i}{1 + \dfrac{\lambda_{\mathcal{T}\cup\{i\}}}{s_i}} > 0.$$

2. *If $i \in \mathcal{T}$ and $k \neq i$, the cross-effect is strictly positive:*

$$\Delta_{i,k} = \frac{s_i}{1 + \dfrac{\lambda_{\mathcal{T}\cup\{k\}}}{s_i}} - \frac{s_i}{1 + \dfrac{\lambda_{\mathcal{T}}}{s_i}} > 0.$$

3. *If $i \notin \mathcal{T}$ and $k \neq i$, the cross-effect is zero:*

$$\Delta_{i,k} = 0.$$

Hence, training scope and prediction scope are supermodular *along each coordinate: adding a covariate to training strictly increases the marginal value of adding that* same *covariate to prediction; if a different covariate is trained, the marginal value of adding an untrained covariate to prediction is unaffected (and remains zero).*

Proof. Note that

$$V(n, \mathcal{T}, \mathcal{P}) = \sum_{j \in \mathcal{P} \cap \mathcal{T}} f_j(\lambda_{\mathcal{T}}), \qquad f_j(\lambda) := \frac{s_j}{1 + \lambda/s_j}.$$

For any set X, the difference

$$V(n, X, \mathcal{P} \cup \{i\}) - V(n, X, \mathcal{P}) = \mathbf{1}_{\{i \in X\}}\, f_i(\lambda_X)$$

because $(\mathcal{P} \cup \{i\}) \cap X$ and $\mathcal{P} \cap X$ differ only by the possible inclusion of i. Apply this identity twice with $X = \mathcal{T} \cup \{k\}$ and $X = \mathcal{T}$:

$$\Delta_{i,k} = \mathbf{1}_{\{i \in \mathcal{T}\cup\{k\}\}}\, f_i(\lambda_{\mathcal{T}\cup\{k\}}) - \mathbf{1}_{\{i \in \mathcal{T}\}}\, f_i(\lambda_{\mathcal{T}}).$$

We now analyze the three cases.

(i) If $i = k$ and $i \notin \mathcal{T}$, then $\mathbf{1}_{\{i \in \mathcal{T}\cup\{k\}\}} = 1$ and $\mathbf{1}_{\{i \in \mathcal{T}\}} = 0$, yielding $\Delta_{i,i} = f_i(\lambda_{\mathcal{T}\cup\{i\}}) > 0$.

(ii) If $i \in \mathcal{T}$ and $k \neq i$, both indicators are 1 and

$$\Delta_{i,k} = f_i(\lambda_{\mathcal{T}\cup\{k\}}) - f_i(\lambda_{\mathcal{T}}).$$

Since $\lambda_{\mathcal{T}\cup\{k\}} = \lambda_{\mathcal{T}} - \frac{s_k}{n} < \lambda_{\mathcal{T}}$ and $f_i'(\lambda) = -\frac{1}{(1+\lambda/s_i)^2} < 0$, we have $\Delta_{i,k} > 0$.

(iii) If $i \notin \mathcal{T}$ and $k \neq i$, both indicators are 0, so $\Delta_{i,k} = 0$.

These cases exhaust all possibilities, establishing the claims. Supermodularity along each coordinate follows directly: the marginal value of adding i to prediction weakly increases when the training set expands, and strictly increases when the expansion either trains i itself or (if i is already trained) improves precision by lowering λ. □

4 Model Selection

4.1 Optimal Model Selection

Problem 3 (Model Selection)**.** The firm must choose the penalty $\lambda = \lambda(n, \mathcal{T})$ and the prediction covariates $\mathcal{P}$ to solve

$$\sup_{n,\mathcal{T},\mathcal{P}} \Pi(n, \mathcal{T}, \mathcal{P}) = \sum_{k\in\mathcal{P}} \frac{s_k}{\frac{1-S(\mathcal{T})}{ns_k} + 1} - C(n, t, p)\,.$$

where $t = |\mathcal{T}|$ and $p = |\mathcal{P}|$.

We will solve the model in two subproblems. Firstly, we find the optimal sets of covariates of given sizes t and p.

Problem 4 (Constrained Covariate Selection)**.** The firm chooses covariate sets of dimensions t and p

$$\begin{aligned} &\max_{\mathcal{P},\mathcal{T}} V(n, \mathcal{P}, \mathcal{T}), \\ &\text{s.t.}\, |\mathcal{T}| = t, \\ &\quad\;\; |\mathcal{P}| = p. \end{aligned}$$

This will yield the optimized value of given dimensions.

Definition 5. The ***optimized value*** of n observations, t training covariates and p prediction covariates is

$$v(n, t, p) \equiv \max_{\mathcal{P},\mathcal{T}} V(n, \mathcal{P}, \mathcal{T}).$$

The optimized value is of independent interest as it shows the interplay between the economies of scope highlighted in the previous section and the diminishing returns deriving from the optimal covariate selection (more valuable covariates will be collected first). This is the statistical equivalent of the Law of Diminishing Returns in Ricardo (1817): the firm uses higher-quality inputs (the most informative covariates) first, lower-quality ones later.

Secondly, we find the optimal dimensions subject to costs.

Problem 5 (Optimal Dimensions)**.** The firm solves

$$\max_{n,t,p} v(n,t,p) - \left[C_n(n) + C_t(t) + C_p(p) + F\right].$$

If we assume the cost function is additively separable, the comparative statics of the optimal dimensionality (n^*, t^*, p^*) will only depend on the value function $v(n,t,p)$. We will study them by analyzing the properties of $v(n,t,p)$.

4.2 Constrained Covariate Selection

Proposition 8 (Optimal Statistical Model)**.** *The model selection problem*

$$\begin{aligned} &\max_{\mathcal{P},\mathcal{T}} V(n,\mathcal{P},\mathcal{T}),\\ &s.t. |\mathcal{T}| \le t,\\ &\quad |\mathcal{P}| \le p, \end{aligned}$$

has the covariates with the largest signals:

$$\mathcal{T}_t^* \in \arg\max_{\substack{\mathcal{T}\subset\mathbb{N}\\ |\mathcal{T}|=t}} \sum_{k\in\mathcal{T}} s_k, \qquad \mathcal{P}^* \in \arg\max_{\substack{\mathcal{P}\subset\mathbb{N}\\ |\mathcal{P}|=p}} \sum_{k\in\mathcal{P}} s_k.$$

It is sufficient to observe that $V(n,\mathcal{T},\mathcal{P})$ is increasing in all s_k, directly for covariate $k \in \mathcal{P}$ and through the penalization $\lambda(n,\mathcal{T})$ for covariate $k \in \mathcal{T}$. The firm treats covariates as production factors of heterogeneous quality: each additional covariate's marginal productivity (signal) falls as one moves down the ordered list.

4.2.1 Continuum Formulation

To simplify notation and obtain differentiable comparative statics, it is convenient to move from a discrete set of covariates $\mathcal{T}, \mathcal{P} \subset \mathbb{N}$ to a continuum representation.

Let $s : [0,\infty) \to \mathbb{R}_+$ denote a weakly decreasing function describing the informativeness (signal) of covariate $u \in [0,\infty)$. Selecting t training covariates and p prediction covariates corresponds to choosing measurable subsets $\mathcal{T}, \mathcal{P} \subset [0,\infty)$ of measure t and p, respectively. Because the value function $V(n,\mathcal{T},\mathcal{P})$ is increasing in each signal s_k, the optimal choice is to select the covariates with the highest signals:

$$\mathcal{T}^* = [0,t), \qquad \mathcal{P}^* = [0,p).$$

Define the cumulative signal and the implied regularization term as

$$S(t) \equiv \int_0^t s(u)\,du, \qquad \lambda(n,t) \equiv \frac{1 - S(t)}{n}.$$

Replacing finite sums by their integral counterparts gives the continuous representation of the optimized value:

$$v(n,t,p) = \int_0^p s(u)\; w(u;\lambda)\big|_{\lambda=\lambda(n,t)}\, du, \qquad w(k;\lambda) = \frac{1}{1+\frac{\lambda}{s(k)}}$$

This continuous formulation is the natural relaxation of the discrete model where s_k denotes the informativeness of the k-th covariate. If $s(\cdot)$ is taken as the step function $s(u) = s_{\lceil u \rceil}$, the integral coincides exactly with the discrete sum $\sum_{k=1}^{p} \frac{s_k}{1+\lambda(n,t)/s_k}$. Smoothing $s(\cdot)$ allows treating t and p as real variables, which facilitates differentiation and yields the same qualitative results. Economically, t and p represent the *breadth* of training and prediction dimensions, while $s(u)$ captures their declining marginal quality.

The marginal returns will be the result of five forces:

1. ***Covariate Spillovers.*** $w(k;\lambda)$ is decreasing in $\lambda(n,t)$, which is decreasing in t, intuitively each covariate on which training occurs improves the estimates of all the parameters of the prediction covariates in p;

2. ***Joint Observations.*** $w(k;\lambda)$ is decreasing in $\lambda(n,t)$, which is decreasing in n, intuitively each observation improves the estimates of all the parameters of the prediction covariates in p;

3. ***House Party Effect***. $w(k;\lambda)$ is convex in λ,intuitively reductions in noise have a positive and convex effect on estimator variance;

4. ***Selection Effect.*** $s'(k) < 0$ (equivalently, $\lambda_{tt}(n,t) > 0$ and $w_k(k;\lambda) < 0$), intuitively each new covariate improves estimates less and less.

4.3 Economies of Scope to Prediction

I now study the economies of scope to prediction by analyzing the marginal value of prediction covariates p and its interactions with training covariates t and sample size n.

Proposition 9 (Diminishing Returns to Prediction Breadth)**.**

$$v_{pp}(n,t,p) = s'(p)w(p;\lambda) + s(p)w_p(p;\lambda) \le 0$$

Selection effect implies the firm ranks covariates in decreasing order of informativeness and the marginal value of prediction covariates decreases with the number of prediction covariates p.

However, the firm can mitigate the decline in returns to t by increasing the number of training observations n or training covariates t.

Proposition 10 (Training-Prediction Complementarities)**.**

$$v_{tp}(t,p) = s(t)w_\lambda(p;\lambda)\lambda_t(t) > 0,$$
$$v_{np}(n,p) = s(t)w_\lambda(p;\lambda)\lambda_n(n) > 0.$$

Covariate spillovers imply that increasing the number of training covariates t reduces the penalty λ for all p prediction covariates, as does increasing the sample size n due to joint observations. Consequently, training and prediction data are complements. To my knowledge this is the first proof of this type of complementarity in the literature. The complementarity between p and n is consistent with the empirical findings of Schaefer and Sapi (2023). The findings are coherent withWilson (1975): better information can be leveraged across the entire scale of production, so the non-rival nature of information generates complementarities.

The following result shows that there can be increasing returns to scale to the number of training covariates t if the selection effect is weak.

Proposition 11 (Returns to Training Breadth)**.** *The marginal value of the number of training covariates t is*

$$v_t(t) = \int_0^p s(u)w_\lambda(u;\lambda)\lambda_t(t)du \geq 0,$$

with curvature

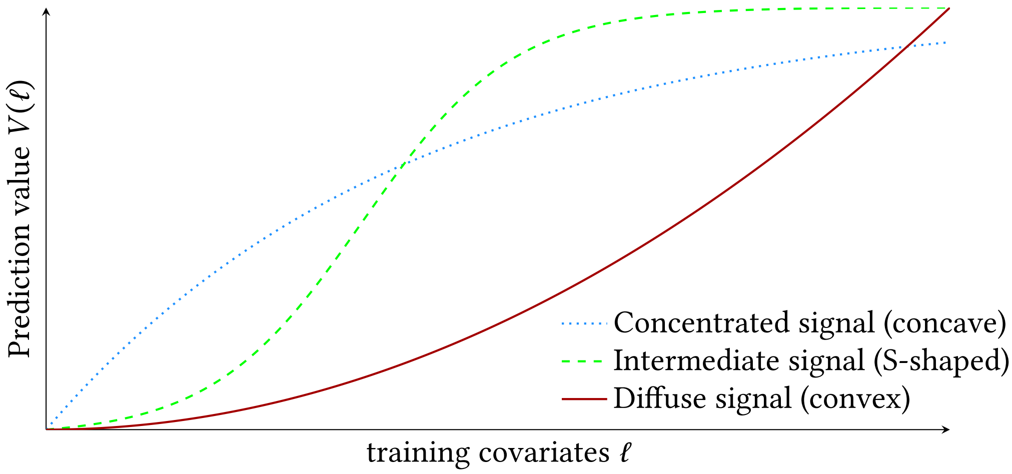
$$v_{tt}(t) = \int_0^p s(u)\left(\underbrace{w_\lambda(u;\lambda)\lambda_{tt}(t)}_{SE<0} + \underbrace{w_{\lambda\lambda}(u;\lambda)\lambda_t^2(t)}_{HPE+CS>0}\right)du.$$

Returns to training covariates are increasing if and only if $n \leq \hat{n}_{tt}(t,p)$,where $\hat{n}_{tt}(t,p)$ is implicitly defined by $\hat{n}_{tt}(t,p) = n$ such that

$$v_{tt}(n,t,p) = 0.$$

Moreover, if $s(u)$ is log-concave $\hat{n}_{tt}(t,p)$ is decreasing in t.

This result is the interplay of two opposing forces: the ***selection effect*** implies that the improvements deriving from training are decreasing as each new variable is less informative than the former; the ***training spillovers*** go in the opposite direction as the spillover is larger if t is already large so more covariates benefit from the economies of scope. The sample size n reduces the importance of having many training covariates so it tilts the balance in favor of the selection effect. Log-concavity of $s(\cdot)$ ensures returns go from increasing to decreasing (the S-shape found Carballa-Smichowski, Duch-Brown, et al. (2025)): with thin-tailed signal distributions, each additional covariate captures a decreasing fraction of the residual variance reducing the HPE effect.



Corollary 4. *If $s(u) = 1$ for all $u \in [0, 1]$, then*

$$v_{tt}(n, t, p) > 0.$$

In the limit case in which all covariates are equally informative there are no Ricardian diminishing returns: the house party effects always dominates and there are always increasing returns.

5 Managerial Implications

My findings have implications for managers, which face increasingly complex decisions on how much and what data to collect. Iansiti (2021) recognizes this difficulty highlighting the importance of accounting for the heterogeneity in data types and the complementarities they generate.Managers must make three decisions

1. **Profitability of data collection**: Is it worthwhile to invest in building a data infrastructure?

2. **Covariate selection**: If so, which user attributes should be prioritized for collection?

3. **Depth vs. breadth of data**: Should the firm focus on expanding the user base (more users) or on increasing engagement (more data per user)?

Profitability of Data Collection Corollary **??** implies prediction technologies typically require a minimum scale of data before becoming profitable, the "cold start" problem discussed in Iansiti 2021. This creates significant sunk costs: firms entering data-intensive markets must commit resources upfront to both user acquisition and data infrastructure before returns can materialize. The challenge is most pronounced when predicting outcomes that:

- Depend on a large number of user attributes (e.g. genomics),

- Exhibit high intrinsic unpredictability (e.g. financial markets), or
- Face elevated data costs due to regulation (e.g. healthcare or privacy-sensitive sectors).

Data Silos Corollary **??** implies that data silos are a concern only if the separated data contain distinct covariates. If the type of information contained in each silo is the same there are no synergies. This implies that firms should be cautions when pursuing data integration: spending to integrate data of the same type will not yield additional value.

Covariate Selection Proposition 8 shows that when choosing which variables to collect managers should balance two aspects:

1. **Relevance**: How strongly the covariate is related to the outcome of interest (i.e. how much predictive power it provides).
2. **Heterogeneity**: How much the covariate varies across the population, since greater variation yields more information for distinguishing between users.

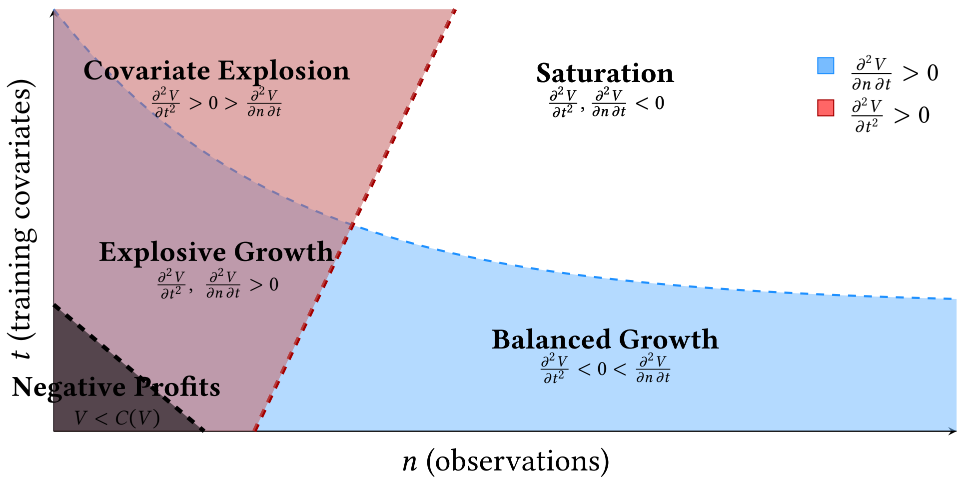
For example, suppose a streaming platform wants to predict churn probability. Age may be more directly correlated with churn than preferred device type. However, if nearly all users fall into a narrow age range (e.g. 25–35), then age offers little information for prediction. By contrast, device type (mobile, tablet, smart TV, console) might be less correlated with churn on average, but because it is much more heterogeneous, collecting it can yield greater predictive gains. Thus, managers should not focus solely on variables with the strongest average correlation, but instead prioritize those that combine relevance with heterogeneity in the user base.

Value of Data Integration Proposition **??** establishes that distinct covariates are complementary, which formalizes the benefits from data integration discussed in Goodhue, Wybo, and Kirsch 1992

Depth vs. breadth of data To choose their firm's data strategy, managers must know where there firm is in the training data space

To choose the right data strategy, managers must understand where their firm is located in the training space defined by the number of observations (n) and the number of covariates trained (t). The trade-off between expanding the user base (more n) or collecting richer attributes (more t) depends critically on this position:

- Loss ($V < C(V)$)
- 6: the amount of data acquired is insufficient to make predictions. Firms need to pass through this phase to accumulate enough data to start making profitable predictions.



- Explosive Growth ($\partial^2 V/\partial t^2, \partial^2 V/\partial n \partial t > 0$): Both adding users and collecting new covariates reinforce each other, producing rapid gains. Startups in early stages of ad-targeting or recommendation may be here, where every new user and attribute dramatically boosts prediction quality.

- Covariate Explosion ($\partial^2 V/\partial t^2 > 0 > \partial^2 V/\partial n \partial t$): Gains come mainly from richer user data, not more users. For instance, a medical AI firm with limited patients benefits more from expanding the range of biomarkers collected per patient than from recruiting a few extra patients.

- Balanced Growth ($\partial^2 V/\partial t^2 < 0 < \partial^2 V/\partial n \partial t$): Returns to new covariates diminish, but expanding the user base still boosts the value of existing attributes. Social media platforms at scale often fall here, where growth in users is more valuable than adding more features per user.

- Saturation ($\partial^2 V/\partial t^2, \partial^2 V/\partial n \partial t < 0$): Both margins yield diminishing returns; prediction performance has plateaued. At this stage, further data collection may not be cost-effective, and firms should shift attention to algorithmic innovation or new products.

The framework also sheds light on firm growth. Early investment decisions shape a firm's long-run trajectory in the training–data space. If, during the explosive growth phase, the firm directs slightly more resources toward covariate collection, its path may shift from

$$\text{Explosive Growth} \to \text{Balanced Growth} \to \text{Saturation}$$

to a path of

$$\text{Explosive Growth} \to \text{Covariate Explosion} \to \text{Saturation}.$$

The latter path stabilizes at a much higher overall data scale, even though the initial difference in investment is small. In other words, modest early increases in the collection of user attributes can push firms from balanced growth toward covariate explosion, ultimately leading them to operate with much larger models in the long-run equilibrium. These two paths are coherent with Farboodi and Veldkamp (2025) which highlights that the presence of economies of scale when data are limited implies that small firms face substantial sunk costs before becoming productive but once they reach the explosive growth phase they either scale up quickly (Covariate explosion) or get caught into a data-poor trap (balanced growth).

6 Applications

We will assume that covariates have identical variance throughout and that prediction and training covariates are the same. Therefore the value of data will have the following form

$$v(n,k) = \frac{k}{\frac{1-k}{n} + 1}.$$

6.1 Natural Monopsony and Data-driven Acquisitions

6.1.1 Setup

Demand. Consider prediction buyers who must take a decision $\hat{y} \in \mathbb{R}$. Their payoff depends on the distance between their chosen decision $\hat{y}$ and the unknown optimal decision $y \in \mathbb{R}$. Buyers share a common prior on y, described in Section 2.1. The optimal decision depends on a vector of covariates $\boldsymbol{x}$, which capture the factors influencing the outcome. Buyers may purchase a prediction from a seller that uses data $\boldsymbol{D}$ as defined in Section 2.2.

If purchasing from seller $i \in \{I, E\}$ at price p_i, the buyer obtains expected utility

$$\tilde{u}(p_i, v_I) = -\mathbb{E}_D\left[(\hat{y}^*(\boldsymbol{D}) - y)^2\right] - p_i = -(1 - v_I) - p_i,$$

where $\hat{y}^*(\boldsymbol{D})$ is the optimal prediction characterized in Lemma 1 and v_I is the value of the seller's dataset.[15]

The outside option is to make the prior mean decision, yielding $\bar{u} = -1$. The net surplus from purchasing a prediction is therefore

$$u(p_i, v_I) \equiv \tilde{u}(p_i, v_I) - \bar{u} = v_I - p_i.$$

Supply An incumbent firm I (Big Tech) sells predictions to a unit mass of potential buyers. An entrant firm E (Startup) can either: (i) sell predictions to a distinct unit mass of buyers, or (ii) accept a take-it-or-leave-it acquisition offer from I for an acquisition price P.

[15] In Appendix B, I provide a microfoundation of α as the fraction of surplus which can be appropriated by a prediction seller when there is a continuum of decision buyers with heterogeneous quality.

Both firms have observed n past decisions. I observes $k\iota$ proprietary training covariates with $k \in [0,1)$ and $\iota \in [0,1]$, and E observes $k(1-\iota)$ proprietary training covariates distinct from those of I. If I acquires E, it gains access to E's additional training covariates. The prediction value generated by the combined data of seller $i \subseteq \{I, E\}$ is

$$v_i \equiv v(n, k_i),$$

where

$$(k_I, k_E) = \begin{cases} (k, 0), & \text{if } I \text{ acquires } E, \\ (k\iota, k(1-\iota)), & \text{otherwise.} \end{cases}$$

Denoting demand of seller i as $D(p_i, v_i)$, the payoff of I is

$$\Pi_I(F, p_I, p_{IE}) = \begin{cases} p_{IE} D(p_{IE}, v_{IE}) - P, & \text{if } I \text{ acquires } E, \\ p_I D(p_I, v_I), & \text{otherwise.} \end{cases}$$

The payoff of E is

$$\Pi_E(p_E) = \begin{cases} F, & \text{if } I \text{ acquires } E, \\ p_E D(p_E, v_E), & \text{otherwise.} \end{cases}$$

Planner Welfare The planner has an objective function

$$W = \Pi_I + \Pi_E + CS - \xi \mathbf{1}(I \text{ acquires } E),$$

where CS denotes the consumer surplus and ξ is the social cost of reduction of entry, due to , e.g., the loss of positive knowledge spillovers as documented by Bloom, Schankerman, and Reenen (2013).

Timing The game has two phases

1. **Potential Acquisition:** I offers P to buy E which accepts or rejects;
2. **Pricing:** if acquisition occurs, I sets p_{IE}; otherwise, I and E set (p_I, p_E) simultaneously. Payoffs are realized.

The equilibrium concept is therefore Subgame Perfect Nash Equilibrium, solved by backward induction.

6.1.2 Buyer Purchase

Buyers purchase the prediction if and only if $v \geq p$. The resulting demand function is

$$D(p, v) = \mathbf{1}(v \geq p).$$

The Seller can extract the whole revenue choosing

$$p^* = v$$

In order to acquire E, I must offer her the value of its outside option, which is charging $p_E^* = v_E$. Therefore, the lowest possible acquisition price is

$$P^* = v_E.$$

The following result characterizes when the acquisition price is lower than the increase in revenue I gets from purchasing E.

Proposition 12. *The incumbent I will always acquire the entrant E.*

Proof. It is sufficient that by convexity of $v(n, k)$ in k we have $v_{EI} \geq v_I + v_E$. □

6.1.3 Planner's Problem

The planner solves

$$W = \begin{cases} \Pi_{IE} - \xi, & I \text{ acquires } E, \\ \Pi_I + \Pi_E, & \text{otherwise}, \end{cases}$$

where we can omit CS as consumer surplus is zero.

Proposition 13. *The acquisition of E by I improves welfare if and only if*

$$\tilde{\xi}(n, k, \iota) \equiv v_{EI} - (v_I + v_E) \geq \xi,$$

where

$$\tilde{\xi}(n, k, \iota) = \frac{k}{\frac{1-k}{n} + 1} - \left(\frac{\iota k}{\frac{1-\iota k}{n} + 1} + \frac{(1-\iota)k}{\frac{1-(1-\iota)k}{n} + 1} \right)$$

is increasing in k, increasing in ι if and only if $\iota \leq 1/2$, increasing in n if and only $n \leq \tilde{n}(\iota, k)$, which is decreasing in k and increasing in ι if and only if $\iota \leq 1/2$.

The result highlights a tension between private and social incentives for acquisition. Privacy worsen this tension. For example, the EU's General Data Protection Regulation (GDPR) emphasizes privacy, but by raising compliance costs it lowers k, thereby dampening economies of scope and reducing the interval $[0, \tilde{\xi}(n, k, \iota)]$ in which the acquisition is socially desirable even though it kills the knowledge spillovers due to entry. Conversely, open-data initiatives raise k and broaden the interval in which economies of scope make acquisition desirable.

6.2 Data Pools

Data owners often form partnerships to pool their datasets and sell access jointly. For example, BMW, Mercedes-Benz, and Audi co-founded the platform *Here Mobility Data Marketplace*, which aggregates GPS, speed, and road-condition data from connected cars. Gu, Madio, and Reggiani (2021) study two data brokers selling datasets that can be either complements or substitutes, showing that pooling is neutral when datasets are complements but collusive when they are substitutes. We extend their analysis by developing a model in the spirit of Lerner and Tirole (2004), building on the fundamental *complements problem* originally noted by Cournot (1838): when complementary goods are sold by independent monopolists, the resulting double marginalization leads to inefficiently high prices.

6.2.1 Setup

Data Owners Consider two data owners, each holding a dataset of identical informativeness. The complete dataset consists of n observations and k covariates, with the training and prediction covariates coinciding. However, the data may be split between the two owners either along the n (sample size) dimension or along the k (covariate) dimension. All parties are symmetrically informed about the informativeness of each dataset.

If the data are split along the sample dimension, each owner holds n observations and all k covariates. Pooling thus doubles the sample size. Conversely, if the data are split along the covariate dimension, each owner holds all n observations but only half the covariates, so pooling expands the covariate space. These two cases capture two distinct sources of complementarity: statistical precision (more n) and informational richness (more k).

Data Buyers There is a continuum of potential buyers (e.g., prediction firms) who can purchase access to one or both datasets and combine them without cost. Buyers are heterogeneous and indexed by $\theta \in [\underline{\theta}, \bar{\theta}]$, representing their adoption cost or opportunity cost of using the prediction technology.

A buyer of type θ who purchases access to $q \in \{1, 2\}$ datasets obtains a gross surplus

$$U_q = V_q - \theta,$$

where V_q denotes the predictive value of having access to q datasets. Specifically,

$$V_q \equiv \begin{cases} V\left(\dfrac{2n}{q}, k\right), & \text{if the data are split along the } n \text{ dimension,} \\ V\left(n, \dfrac{2k}{q}\right), & \text{if the data are split along the } k \text{ dimension.} \end{cases}$$

Since $V(n, k)$ is increasing in both arguments, combining datasets strictly improves predictive accuracy. Intuitively, pooling along the n dimension increases the number of observations available for training, which reduces estimation error, whereas pooling along the k dimension increases the number of predictive features, which broadens the scope of prediction.

The heterogeneity parameter θ is distributed according to

$$G(\theta) = \theta^{\alpha}, \qquad \alpha \in [0, 1],$$

implying a strictly increasing hazard rate $\frac{g(\theta)}{1-G(\theta)}$ with $g = G'$. The parameter α thus governs the curvature of demand: when α is low, heterogeneity is large and demand is relatively inelastic; when α is high, buyers are more homogeneous and demand becomes more elastic.

Given a price P for access to a bundle of $q \in \{1, 2\}$ datasets, only buyers with $V_q - \theta \geq P$ make the purchase. Hence, the corresponding demand function is

$$D(P; V_q) = \Pr(V_q - \theta \geq P) = (V_q - P)^{\alpha}.$$

This formulation implies that the elasticity of demand increases in α: higher α corresponds to a market in which adoption falls faster as price rises. Equivalently, α can be interpreted as the *semi-elasticity of demand*, describing how responsive adoption is to changes in price.

6.2.2 Pooling Optimum

When data owners form a pool, they coordinate pricing and behave as a single monopolist offering a bundled dataset. This situation mirrors the *pool pricing benchmark* in [Section I(b)]LernerTirole2004, where a patent pool chooses the package price P to maximize joint revenue given demand $D(P - V_2)$. The logic is analogous here: the pooled data bundle yields predictive value V_2, and all buyers face a single posted price P.

Lemma 4 (Optima Pool Price)**.** *The optimal pool price is*

$$P^* = \frac{V_2}{\alpha + 1}.$$

The pool acts as a monopolist setting the joint-access price that equates marginal revenue to zero, just as in Equation (1) in Lerner and Tirole (2004). The resulting price is decreasing in the demand semi-elasticity α: when α is high, buyers are more sensitive to price, leading the pool to charge less. Conversely, when α is low, demand is inelastic, so the pool can extract a larger fraction of the total value V_2.

Economically, pooling internalizes the complementarity between the dataset, analogous to Cournot's double marginalization problem in complementary goods. When data are sold separately, each owner fails to account for the positive externality that lower prices have on the other's sales. By setting a single joint price, the pool eliminates this inefficiency and behaves as an integrated monopolist.

6.2.3 Observation Fragmentation

Section II in Lerner and Tirole (2004) characterizes the unique symmetric equilibrium in the case in which the brokers do not form a pool.

Lemma 5. *If the brokers have different observations on the same covariates,*

$$p_i = \min\left\{ V_2 - V_1, \frac{V_2}{2+\alpha} \right\}$$

and the buyers will buy from both brokers.

Proof. We follow Lerner and Tirole (2004). **Demand Margin Binds.** Suppose that the brokers offer prices $\mathcal{P} \equiv (p_1, p_2)$, and wlog $p_1 \leq p_2$. Prediction Sellers decide how many datasets to buy.

$$\mathcal{V}(\mathcal{P}) = \max_{q\in\{1,2\}} \left\{ V_q - p_1 - p_2 \mathbf{1}(\{q=2\}) \right\}$$

Second, the user adopts the technology if and only if

$$\mathcal{V}(\mathcal{P}) \geq \theta.$$

Lerner and Tirole (2004) demonstrate the existence of a symmetric equilibrium. Individual data sellers solve

$$\hat{p} = \arg\max_{p_i} \left\{ p_i D\left(p_i + \hat{p} - V_2\right) \right\}$$

which has FOC

$$\hat{p} D'\left(2\hat{p} - V_2\right) + D\left(2\hat{p} - V_2\right) = 0$$

which has a unique solution buy hazard-rate monotonicity. It can be seen as selling the whole pool setting total price P and keeping $p_i = P - \hat{p}$ for itself. Therefore

$$\hat{P} = \arg\max_{P} \left\{ \left(P - \hat{p}\right) D\left(P - V_2\right) \right\}.$$

The term $\hat{p}$ can be seen as a marginal cost $\hat{c} = \hat{p}$. In this interpretation when there is the pool $c^* = 0$ so by revealed preference

$$\hat{P} \geq P^*.$$

If demand margin binds in the absence of a pool then the pool reduces price paid by data buyers. This means that if all datasets can increase the price marginally without being excluded, the pool is pro-competitive.

With our CDF G,

$$p_{\text{dem}} = \frac{V_2}{2+\alpha}.$$

Competition Margin Binds. Define the price when the competition margin binds will be p_{comp} defined by

$$V_2 - 2p_{\text{comp}} = \max_{q\in\{0,1\}} \left\{ V_q - q p_{\text{comp}} \right\}.$$

If $V_1 - p_{\text{comp}} \geq 0$, then $p_{\text{comp}} = V_2 - V_1$. This is consistent because $V_1 - (V_2 - V_1) > 0$ by concavity of $V(n,t)$ in n. Otherwise, if $V_1 - p_{\text{comp}} < 0$, then $p_{\text{comp}} = V_2/2$. This is not consistent because $V_1 - V_2/2 > 0$ by concavity of $V(n,t)$ in n. □

Intuitively, as the joint value V_2 increases, the price increases but there is a kink in $V_2 = V_1 \cdot \frac{\alpha+2}{\alpha+1}$ above which part of the increment is appropriated by buyers. Then Proposition 1 in Lerner and Tirole (2004) directly implies the following result

Proposition 14. *A pool of observations is procompetitive if and only if*

$$\alpha > \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1} = \frac{n}{2(1-k)}.$$

Proof. Pools are procompetitive if the demand margin binds i.e.

$$p_{\text{comp}} > p_{\text{dem}} \iff V_2 - V_1 > \frac{V_2}{2+\alpha} \iff \alpha > \alpha_{\text{marg}} \equiv \frac{2V_1 - V_2}{V_2 - V_1}.$$

In this case

$$p_i = \frac{V_2}{2+\alpha}.$$

This implies that observation pooling can be procompetitive when n is not too large and k is not too small, meaning data is relatively abundant and models are relatively complex. Furthermore, as the RHS is increasing in Q, data pools are more likely to be competitive if Q is small meaning if data fragmentation is limited.

Otherwise if the competition margin binds,

$$p_i = V_2 - V_1.$$

the pool is procompetitive if the pool price is lower than the competition price, i.e.

$$P^* < Qz(Q) \iff \frac{V_2}{\alpha+1} < 2\left(V_2 - V_1\right) \iff \alpha > \alpha_{\text{comp}} \equiv \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1}.$$

As $\alpha_{\text{marg}} > \alpha_{\text{comp}}$, the relevant threshold is α_{comp} and so a pool of observations is procompetitive if and only if

$$\alpha > \alpha_{\text{comp}} = \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1}.$$

This implies that as k and n increase it becomes less likely that the pool is procompetitive. When data is abundant and demand is inelastic observation pools are anticompetitive. Direct application of Proposition 5 in Lerner and Tirole (2004) implies that the pool is strongly unstable, therefore enforcing independent licensing of datasets will prevent pooling if and only if the pool is welfare-reducing. In this case each data broker will charge $p_{\text{comp}} = V_2 - V_1$. $\square$

Lemma 6. *If the brokers have distinct covariates each B_i prices at*

$$p_i = \frac{V_2}{2 + \alpha}$$

and the buyers will buy from both brokers.

Proof. If $V_1 - p_{\text{comp}} \geq 0$, then $p_{\text{comp}} = V_2 - V_1$. This is not consistent because $V_1 - (V_2 - V_1) < 0$ by convexity of $V(n, t)$ in t. Otherwise, if $V_1 - p_{\text{comp}} < 0$, then $p_{\text{comp}} = V_2/2$. This is consistent because $V_1 - V_2/2 < 0$ by convexity of $V(n, t)$ in t. Demand margin always binds as

$$p_{\text{comp}} > p_{\text{dem}} \iff \frac{V_2}{2} > \frac{V_2}{2 + \alpha}.$$

Applying Proposition 1 in Lerner and Tirole (2004) implies the following. $\square$

Proposition 15. *A pool of buyers with distinct covariates is always procompetitive*

Following Lerner and Tirole (2004), there may be asymmetric equilibria but they result in lower industry profit than the symmetric one so we focus on the symmetric equilibrium.

6.3 Data Exclusivity

Recent exclusive data-licensing agreements—such as Reddit's 2024 deals with OpenAI and Google—highlight broader concerns that proprietary access to datasets may distort competition in AI and prediction markets. Because data are non-rival, a data seller faces a dynamic commitment problem akin to that of a durable-good monopolist: once it has licensed the dataset to one firm, it is tempted to also license it to the rival, eroding the first buyer's advantage. As in Katz and Shapiro (1986) and Aghion and Bolton (1987), exclusivity can serve as a contractual commitment device that mitigates opportunism by the seller but introduces a welfare trade-off: it softens business-stealing while depressing investment by excluded firms when data and proprietary inputs are complements. The model predicts that profitable exclusivity may become socially undesirable when datasets are abundant and product-market rivalry is intense. In such settings, exclusivity amplifies incumbency advantages and can deter entry—offering a micro-foundation for regulatory scrutiny of data-sharing agreements such as the Reddit–OpenAI deal.

6.3.1 Setup

There are three players: a data seller S (e.g., a platform holding user-generated data) and two prediction firms F_1 and F_2. The firms compete to sell predictions to consumers whose utility depends on the accuracy of the prediction.

Prediction Demand. There is a unit mass of consumers, divided into:

- a mass $\sigma \in [0, 1]$ of *shoppers*, who can buy from either firm;
- a mass $(1 - \sigma)/2$ of *captive consumers* for each firm, who can only buy from that firm.

The parameter σ captures the intensity of competition: when $\sigma = 0$, all consumers are captive and firms behave as local monopolists; when $\sigma = 1$, all consumers are shoppers and the market is fully competitive. This simple structure captures the idea that data quality matters only in relative terms, since shoppers migrate toward the firm offering the more accurate prediction.

Consumer utility from firm $i \in \{1, 2\}$ is

$$u_i = v_I - p_i,$$

where v_I denotes the quality of the prediction (the *value of data*) and p_i is the price charged. Shoppers buy from the firm with the highest net utility $v_i - p_i$, while captive consumers always purchase from their incumbent firm.

Prediction Supply. Each firm $i \in \{1, 2\}$ starts with no covariates and can improve its prediction quality through two channels:

1. *Licensing.* Firm i can license a share $k\lambda$ of the seller's covariates, paying a fee f. This choice is represented by a binary variable $\ell_i \in \{0, 1\}$.
2. *Proprietary Data Collection.* Firm i can also collect a share $k(1 - \lambda)$ of proprietary covariates at a fixed cost $c > 0$, represented by a binary choice $r_i \in \{0, 1\}$.

The parameter $\lambda \in [0, 1]$ measures the importance of the licensed data relative to proprietary data, while $k \in [0, 1)$ scales the overall potential richness of covariates. We assume that the same covariates are relevant for both training and prediction.

A firm's data strategy is $(\ell_i, r_i) \in \{0, 1\}^2$, and its resulting data stock is

$$k_i(\ell_i, r_i) \equiv k\,[\lambda\ell_i + (1 - \lambda)r_i].$$

Under the assumption that licensed and proprietary covariates are distinct and i.i.d., the prediction quality of firm i is

$$v_i \equiv v(\ell_i, r_i) = \frac{k_i(\ell_i, r_i)}{1 + \frac{1 - k_i(\ell_i, r_i)}{n}},$$

where n denotes the number of observations available for training. Intuitively, prediction quality increases with the amount of available data but at a diminishing rate, reflecting decreasing returns to data scale.

Pricing and Profits. Firms can price discriminate between captive consumers and shoppers. Denote prices by p_i^c (captives) and p_i^s (shoppers). Let D_i^s denote firm i's share of shoppers, with $D_1^s + D_2^s = 1$. Firm i's revenue from prediction is then

$$\pi_i = p_i^c \frac{1-\sigma}{2} + p_i^s \sigma D_i^s,$$

and total profit is

$$\Pi_i = \pi_i - c\, r_i - f\, \ell_i.$$

The cost terms capture the trade-off between acquiring data and improving prediction quality: collecting proprietary data entails the fixed cost c, while licensing requires the payment f to the data seller.

Data Supply. The data seller S has a monopoly over k_ℓ covariates and sets a take-it-or-leave-it license fee f.[16] The seller's profit is

$$\Pi_S = f \sum_{i \in \{1,2\}} \ell_i.$$

Because data are non-rival, S could in principle sell to both firms without loss of quality, but doing so may erode the exclusivity premium paid by the first buyer. This creates the central commitment problem: after selling to one firm, the seller is tempted to license to the rival as well, thereby reducing the first buyer's willingness to pay ex ante.

Social Planner. A benevolent social planner evaluates total welfare as

$$W = \Pi_S + \Pi_1 + \Pi_2 + CS,$$

where consumer surplus is given by

$$CS = \sum_{i \in \{1,2\}} \left[(v_i - p_i^c)\frac{1-\sigma}{2} + \sigma (v_i - p_i^s) D_i^s \right].$$

This allows us to assess how exclusivity affects welfare through both prices and investment incentives.

Timing. Information is complete. The game unfolds in three stages:

[16] Allowing S to set different fees for F_1 and F_2 would not affect the analysis.

1. **License allocation.** The data seller S sets a publicly observed license fee f;
2. **Data collection.** Each firm F_i chooses $(\ell_i, r_i) \in \{0, 1\}^2$, which are publicly observed;
3. **Price competition.** Firms simultaneously set (p_i^c, p_i^s). Consumers choose the firm offering higher utility, and profits are realized for S, F_1, and F_2.

We solve the game by backward induction and characterize the Subgame Perfect Equilibrium (SPE). Intuitively, exclusivity will emerge when the seller can extract a higher licensing fee from one buyer by committing not to sell to the other, even though such a restriction reduces total welfare by suppressing the excluded firm's incentive to invest.

6.3.2 Bertrand Competition in Prediction

We analyze the pricing subgame, taking prediction qualities (v_1, v_2) as given from the data stage. Firms can price-discriminate between captive consumers and shoppers.

Captive pricing. Let the outside option yield utility 0. Captive consumers of F_i can only buy from F_i and do so iff $v_i - p_i \geq 0$. Since captives of F_i will buy only if $p_i^c \leq v_i$, and F_i prefers higher prices, the unique best response is

$$p_i^c = v_i \qquad \text{for each } i \in \{1, 2\}. \tag{6}$$

This extracts the full surplus of captives and yields zero captive consumer surplus.

Shopper pricing. Shoppers buy from the firm that offers the higher net utility; if $v_i - p_i > v_j - p_j^s$ they all buy from F_i, if $v_i - p_i < v_j - p_j^s$ they all buy from F_j, and if equal they split equally.[17] Prices are restricted to be nonnegative. Write $(x)^+ \equiv \max\{x, 0\}$. The next lemma characterizes the unique trembling-hand perfect equilibrium for shoppers.

Lemma 7 (Shopper-price equilibrium). *Fix (v_1, v_2). The pricing subgame in shopp[Shopper-price equilibrium]er prices admits a unique equilibrium under trembling-hand perfection:*

$$p_i^s = (v_i - v_j)^+, \qquad p_j^s = (v_j - v_i)^+, \qquad i \neq j, \tag{7}$$

so that (i) if $v_i > v_j$, firm i sets $p_i^s = v_i - v_j$, firm j sets $p_j^s = 0$, and all shoppers buy from i; (ii) if $v_i = v_j$, both set $p_i^s = p_j^s = 0$ and shoppers can be split arbitrarily.

Combining Equation (6) and Lemma 7, the equilibrium profit of firm i given data strategies (ℓ_i, r_i) and (ℓ_j, r_j) (which pin down (v_I, v_j)) is

$$\pi(\ell_i, r_i; \ell_j, r_j) \;=\; \frac{1-\sigma}{2}\, v_i \;+\; \sigma\,(v_i - v_j)^+, \qquad j \neq i. \tag{8}$$

[17] This tie-breaking rule is without loss for the equilibrium characterization; a trembling-hand refinement will select the limit outcome stated below.

The first term is captive revenue at the value price; the second term is shopper revenue, which equals the quality advantage when the firm is better and zero otherwise. Captives are locked in, so each firm posts $p_i^c = v_i$ to extract their entire surplus. Shoppers create a Bertrand-style race in *net utility.* The higher-quality firm can just match the rival's net utility by setting $p_i^s = v_i - v_j$, thereby winning all shoppers while extracting their full willingness to pay. The lower-quality firm cannot profitably win shoppers without pricing below zero, so its optimal (refined) shopper price is 0. Hence, a one-point quality lead converts into a one-for-one price premium for shoppers, delivering the surplus formula in Equation (8). As σ increases, the return to even small quality advantages grows proportionally, sharpening investment incentives upstream.

6.3.3 Data Collection

We now analyze the data-collection subgame, taking the pricing behavior characterized above as given. Recall that with discriminatory pricing the profit of firm i when qualities are (v_i, v_j) is

$$\pi(\ell_i, r_i; \ell_j, r_j) \;=\; \frac{1-\sigma}{2}\, v_I \;+\; \sigma\,(v_i - v_j)^+, \qquad j \neq i,$$

so the marginal private benefit from acquiring proprietary data is the induced increase in v_I times the relevant captive and shopper weights. Because licensed and proprietary data are *complements* in quality, incentives to collect proprietary covariates are strictly stronger when the firm holds a license.

Assumption 3.

$$c \in [\,\underline{c},\ \bar{c}\,], \qquad \underline{c} \equiv \frac{1-\sigma}{2}\, v(0,1), \qquad \bar{c} \equiv \frac{1-\sigma}{2}\,[\,v(1,1) - v(1,0)\,].$$

The lower bound $\underline{c}$ is equivalent to

$$\pi(0,1;\ell_j,1) - \pi(0,0;\ell_j,1) < c,$$

meaning that if the rival invests a firm will not invest if it has not purchased the license. The upper bound $\bar{c}$ is equivalent to

$$\pi(1,1;\ell_j,1) - \pi(1,0;\ell_j,1) > c,$$

meaning that if the rival invests a firm will invest if it has purchased the license. Complementarity implies $\bar{c} > \underline{c}$, so the interval is non-empty.

Lemma 8 (Monotone best response in r_i)**.** *Under Assumption 3, the proprietary-collection best response of F_i is monotone in its license choice:*

$$r_i^{BR}(\ell_i;\ \ell_j, r_j) \;\geq\; \ell_i \qquad \textit{for all } (\ell_j, r_j).$$

Equivalently: if F_i purchases the license ($\ell_i = 1$), then collecting proprietary data ($r_i = 1$) is (weakly) strictly optimal for any opponent action; if i does not license ($\ell_i = 0$), collecting may fail to be profitable in particular opponent configurations.

With a license in place, adding proprietary covariates raises quality from $v(1, 0)$ to $v(1, 1)$. Even in the worst case where the rival already serves all shoppers (so the positive-part term is zero), the firm still extracts the captive gain $\frac{1-\sigma}{2}[v(1, 1) - v(1, 0)]$, which covers c by Assumption 3. Without a license, the best the firm can guarantee is the captive gain $\frac{1-\sigma}{2}v(0, 1)$; if the rival is strong (e.g., at $(1, 1)$), no shoppers can be won at the margin, so collecting can be unprofitable at cost $c \geq \underline{c}$. Complementarity thus makes r_i strictly more attractive when $\ell_i = 1$.

Lemma 9 (Data-collection equilibrium by license allocation). *Under Assumption 3, let $\ell_1 + \ell_2 \geq 1$. Then any Nash equilibrium of the data-collection subgame satisfies*

$$(r_1^*, r_2^*) \; = \; (\ell_1, \ell_2)$$

Assumption 3 isolates an intermediate-cost region in which (i) licensing strictly strengthens the incentive to collect proprietary data (complementarity), but (ii) without a license, collection can be privately unprofitable when facing a rival who invests and collects data. When both firms are licensed, duplication of the fixed cost c can occur because each firm ignores the business-stealing externality on shoppers; when only one is licensed, the excluded firm typically forgoes investment, preserving its captive market but limiting overall data acquisition.

6.3.4 License Allocation

In this stage, the data seller S sets the license fees (f_1, f_2) anticipating the firms' data-collection and pricing responses. The seller chooses these fees to maximize its profit, taking into account that the fees determine which firms choose to license the data.

Proposition 16 (Seller's optimal licensing policy). *Under Assumption 3, the data seller's optimal licensing fee satisfies*

$$f^* = \begin{cases} f^{NE}, & \text{if } \sigma \leq 1/3 \wedge XXX, \\ f^{NE} + \sigma v(1, 1), & \text{if } \sigma > 1/3, \end{cases} \qquad L^* = \begin{cases} 2, & \text{if } \sigma \leq 1/3, \\ 1, & \text{if } \sigma > 1/3, \end{cases}$$

where

$$f^{NE} \equiv \frac{1-\sigma}{2} v(1, 1) - c.$$

When σBy setting asymmetric fees, S can credibly implement an exclusive deal and extract the exclusivity premium $\sigma v(1, 1)$ the additional willingness to pay stemming from being the sole holder of the licensed covariates. Intuitively, the more intense competition among shoppers (higher σ), the greater the advantage from exclusivity, as a small quality edge allows the licensee to capture the entire shopper market.

6.3.5 Planner's Problem

Lemma 10 (Planner's optimum)**.** *Under Assumption 3, it is socially optimal to sell licenses to both firms ($L = 2$).*

Corollary 5 (Conflict between private and social incentives)**.** *Under Assumptions 3 and* **??***, the data seller finds exclusivity profitable even though total welfare would be higher under non-exclusive licensing.*

The difference between the cost bounds $\bar{c} - \underline{c}$ captures the parameter range where exclusivity can arise:

$$\bar{c} - \underline{c} = \frac{1-\sigma}{2}\left[\, v(1,1) - v(1,0) - v(0,1) \,\right] = \frac{k^2(\lambda-1)\lambda n(1-\sigma)(k-2(n+1))}{2(k-n-1)(k(\lambda-1)+n+1)(k\lambda-n-1)}.$$

This length is increasing in k (the overall richness of covariates), and inverted-U-shaped in both the number of observations n and the share of licensed data λ, with a maximum at $\lambda = 1/2$. Hence, anticompetitive exclusivity is most likely when data are *abundant* (k large) and evenly split between proprietary and licensed sources ($\lambda \simeq 1/2$).

The effect of n is non-monotonic: when data are scarce, adding observations exacerbates exclusivity concerns by strengthening the complementarity between licensed and proprietary data; beyond a threshold, further increases in n reduce these incentives. Overall, the model predicts that profitable exclusivity becomes *socially undesirable* when covariates are abundant, observations are intermediate, and competition over shoppers is fierce—providing a microfoundation for antitrust scrutiny of data-sharing deals such as the Reddit-OpenAI agreement, where exclusivity enhances incumbency advantages and deters market entry.

7 Conclusion

This paper develops a general framework for understanding the value of data in prediction by explicitly modeling covariates. The analysis shows how economies of scope across covariates, interactions between covariates and observations, and complementarities between training and prediction can generate increasing returns, offering a microfoundation for the rich-get-richer effects often observed in data-driven markets.

These forces have direct implications for policy and strategy. Prediction technologies may display natural monopsony characteristics, as concentrating covariates within one firm can raise efficiency. Privacy regulation that fragments data supply may inadvertently reinforce monopsony power, creating a trilemma between privacy, competition, and efficiency. The framework also highlights that not all data pooling agreements are alike: pooling lists of users with the same covariates can be anticompetitive, where as pooling different covariates on similar users raise welfare by eliminating double marginalization. Exclusivity deals, such as those signed between AI labs and data providers, may profitably foreclose entry by depriving rivals of essential complements. For firms, the results imply that prediction entails substantial sunk costs: early on, investment should balance user acquisition and attribute enrichment, while specialization and integration become optimal at a larger scale.

More broadly, the analysis cautions against treating data as homogeneous. Policies promoting open data without regard to dataset composition may miss crucial efficiency margins, whereas access remedies such as FRAND-priced APIs or federated learning preserve economies of scope.

My work opens two natural avenues for future research. The first is empirical. I aim to develop a methodology to test my results on real datasets. While the existing empirical literature[18] provides partial support to my findings, it suffers from two limitations: (i) most studies focus on a single dataset, whereas uncovering general properties requires comparing multiple datasets along common dimensions; and (ii) no existing work systematically tests all the properties identified in my model. Once these empirical properties are validated, my framework could serve as the foundation for a practical formula for data valuation, in the spirit of the Black–Scholes–Merton formula for derivatives.[19] The second avenue is theoretical. Embedding my static model into a dynamic Wald sampling framework would allow me to microfound data-enabled learning and analyze when feedback loops generate convergent data-collection strategies versus when they diverge.

Finally, the framework invites a broader research agenda: in his seminal critique of central planning, Hayek 1945 emphasized that “knowledge... never exists in concentrated form but solely as the dispersed bits... which all the separate individuals possess”. Today, users' online activity transforms such dispersed knowledge into datasets that can be centralized, recombined, and monetized. My analysis shows that statistical properties of prediction create intrinsic incentives for such concentration. The concentration of data in servers controlled by a few large firms raises a broader question: do prediction algorithms substitute for, or complement, the market mechanism? Is the rise of data the panacea to market failures deriving from asymmetric information and search frictions, or is it the first step to the fall of the market? I leave this foundational question open to future research.

[18] See Bajari et al. 2019; Schaefer and Sapi 2023; Lee and Wright 2023; Yoganarasimhan 2020; Carballa-Smichowski, Duch-Brown, et al. 2025

[19] See Black and Scholes 1973, Merton 1973

References

Acemoğlu, Daron et al. (2022). “Too much data: Prices and inefficiencies in data markets.” *American Economic Journal: Microeconomics* 14.4, pp. 218–256.

Aghion, Philippe and Patrick Bolton (1987). “Contracts as a Barrier to Entry.” *The American Economic Review* 77.3, pp. 388–401.

Allcott, Hunt et al. (2025). *Sources of market power in web search: Evidence from a field experiment.* Tech. rep. National Bureau of Economic Research.

Aral, Sinan, Erik Brynjolfsson, and DJ Wu (2008). “Which came first, IT or productivity? The virtuous cycle of investment and use in enterprise systems.”

Bajari, Patrick et al. (2019). “The impact of big data on firm performance: An empirical investigation.” *AEA papers and proceedings* 109, pp. 33–37.

Belkin, Mikhail et al. (2019). “Reconciling modern machine-learning practice and the classical bias–variance trade-off.” *Proceedings of the National Academy of Sciences* 116.32, pp. 15849–15854.

Bergemann, Dirk and Alessandro Bonatti (Aug. 2024). “Data, Competition, and Digital Platforms.” *American Economic Review* 114.8, pp. 2553–2595.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2022). “The economics of social data.” *The RAND Journal of Economics* 53.2, pp. 263–296.

Berger, James O. (1990). *Statistical decision theory.* Springer, pp. 277–284.

Black, Fischer and Myron Scholes (1973). “The Pricing of Options and Corporate Liabilities.” *Journal of Political Economy* 81.3, pp. 637–654.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen (2013). “IDENTIFYING TECHNOLOGY SPILLOVERS AND PRODUCT MARKET RIVALRY.” *Econometrica* 81.4, pp. 1347–1393.

Bollerslev, Tim and Jeffrey M. Wooldridge (1992). “Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances.” *Econometric Reviews* 11, pp. 143–172.

Calzolari, Giacomo, Anatole Cheysson, and Riccardo Rovatti (2025). “Machine data: market and analytics.” *Management Science.*

Carballa-Smichowski, Bruno, Néstor Duch-Brown, et al. (2025). “Economies of scope in data aggregation: Evidence from health data.” *Information Economics and Policy* 71, p. 101146.

Carballa-Smichowski, Bruno, Yassine Lefouili, et al. (Feb. 2025). *Data Sharing or Analytics Sharing?* TSE Working Paper 25-1615. Toulouse School of Economics.

Cong, Lin William, Zhiguo He, and Changhua Yu (2021). “Data as Capital.” *Review of Financial Studies* 34.6, pp. 2895–2936.

Cornière, Alexandre de and Greg Taylor (2024). “Data-Driven Mergers.” *Management Science* 70.9, pp. 6473–6482.

Cournot, Antoine Augustin (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. English translation: *Researches into the Mathematical Principles of the Theory of Wealth*, Macmillan, 1897. Paris: Hachette.

Dasaratha, Krishna, Juan Ortner, and Chengyang Zhu (2025). "Markets for Models." *arXiv preprint arXiv:2503.02946*.

De Corniere, Alexandre and Greg Taylor (2025). "Data and Competition: A Simple Framework." *Forthcoming, RAND Journal of Economics*.

DeGroot, Morris H. (2005). *Optimal statistical decisions*. John Wiley & Sons.

Farboodi, Maryam and Laura Veldkamp (2025). *A model of the Data Economy*. Tech. rep. R&R, Review of Economic Studies.

Goldfarb, Avi and Catherine Tucker (2011). "Privacy Regulation and Online Advertising." *Management Science* 57.1, pp. 57–71.

Goodhue, Dale L., Michael D. Wybo, and Laurie J. Kirsch (1992). "The Impact of Data Integration on the Costs and Benefits of Information Systems." *MIS Quarterly* 16.3, pp. 293–311.

Gourieroux, C., A. Monfort, and A. Trognon (1984). "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52.3, pp. 681–700.

Gu, Yiquan, Leonardo Madio, and Carlo Reggiani (Sept. 2021). "Data brokers co-opetition." *Oxford Economic Papers* 74.3, pp. 820–839.

Hagiu, Andrei and Julian Wright (2023). "Data-enabled learning, network effects, and competitive advantage." *The RAND Journal of Economics* 54.4, pp. 638–667.

Hastie, Trevor et al. (2020). "Surprises in High-Dimensional Ridgeless Least Squares Interpolation."

Hayek, Friedrich A. (1945). "The Use of Knowledge in Society." *American Economic Review* 35. Reprinted in F.A. Hayek (ed.), Individualism and Economic Order. London: Routledge and Kegan Paul, pp. 519–530.

Iansiti, Marco (2021). "The Value of Data in the Age of AI." *Working Paper*.

Iyer, Ganesh and Tianshu Ke (2024). "Competition and Algorithmic Complexity in Predictive Analytics." *Marketing Science* 43.2, pp. 215–233.

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. New York: Springer-Verlag.

Jones, Charles I. and Christopher Tonetti (Sept. 2020). "Nonrivalry and the Economics of Data." *American Economic Review* 110.9, pp. 2819–58.

Kaplan, Jared et al. (2020). "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361*.

Katz, Michael L. and Carl Shapiro (Aug. 1986). "How to License Intangible Property*." *The Quarterly Journal of Economics* 101.3, pp. 567–589.

Lee, Gunhaeng and Julian Wright (2023). “Recommender systems and the Value of User Data.” *National University of Singapore Working Paper.*

Lerner, Josh and Jean Tirole (June 2004). “Efficient Patent Pools.” *American Economic Review* 94.3, pp. 691–711.

Lewis, Patrick et al. (2020). “Retrieval-augmented generation for knowledge-intensive nlp tasks.” *Advances in neural information processing systems* 33, pp. 9459–9474.

Lindley, D. V. and A. F. M. Smith (1972). “Bayes Estimates for the Linear Model.” *Journal of the Royal Statistical Society: Series B (Methodological)* 34.1, pp. 1–18.

Liu, Nelson F et al. (2023). “Lost in the middle: How language models use long contexts.” *arXiv preprint arXiv:2307.03172.*

MacKay, David J. C. (May 1992). “Bayesian Interpolation.” *Neural Computation* 4.3, pp. 415–447.

Merton, Robert C. (1973). “Theory of Rational Option Pricing.” *The Bell Journal of Economics and Management Science* 4.1, pp. 141–183.

Montiel Olea, José Luis et al. (Apr. 2022). “Competing Models.” *The Quarterly Journal of Economics* 137.4, pp. 2419–2457.

Nakkiran, Preetum et al. (2021). “Deep double descent: Where bigger models and more data hurt.” *Journal of Statistical Mechanics: Theory and Experiment* 2021.12, p. 124003.

Prüfer, Jens and Christoph Schottmüller (2021). “Competing with big data.” *The Journal of Industrial Economics* 69.4, pp. 967–1008.

Radner, Roy and Joseph Stiglitz (1984). “A Nonconcavity in the Value of Information.” *Bayesian models in economic theory* 5, pp. 33–52.

Ricardo, D. (1817). *On the Principles of Political Economy and Taxation.* John Murray.

Schaefer, Maximilian (2025). *When Should we Expect Non-Decreasing Returns from Data in Prediction Tasks?*

Schaefer, Maximilian and Geza Sapi (2023). “Complementarities in learning from data: Insights from general search.” *Information Economics and Policy* 65, p. 101063.

Varian, Hal (Mar. 2018). “Artificial Intelligence, Economics, and Industrial Organization.” *The Economics of Artificial Intelligence: An Agenda.* NBER Chapters. National Bureau of Economic Research, Inc, pp. 399–419.

White, Halbert (1982). “Maximum Likelihood Estimation of Misspecified Models.” *Econometrica* 50.1, pp. 1–25.

Wilson, Robert (1975). “Informational Economies of Scale.” *Bell Journal of Economics* 6.1, pp. 184–195.

Yoganarasimhan, Hema (2020). “Search personalization using machine learning.” *Management Science* 66.3, pp. 1045–1070.

A Proofs

Lemma 1 (Optimal Predictor)**.** *The optimal predictor is*

$$f^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}}) = \mathbb{E}[y \mid \boldsymbol{D}_{\mathcal{T},\mathcal{P}}] = \boldsymbol{x}'_{\mathcal{T}\cap\mathcal{P}}\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}\cap\mathcal{P}} \mid \boldsymbol{M}_{\mathcal{T}}\right].$$

Proof. Under squared loss, the Bayes optimal predictor is the conditional mean:

$$f^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}}) = \mathbb{E}[y \mid \boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}].$$

Write $y = \sum_{k\in\mathcal{K}} \beta_k x_k$. By the law of iterated expectations and independence of $\boldsymbol{x}$ and $\boldsymbol{\beta}$,

$$\mathbb{E}[y \mid \boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}] = \sum_{k\in\mathcal{K}} \mathbb{E}[\beta_k x_k \mid \boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}] = \sum_{k\in\mathcal{P}} x_k\, \mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}] + \sum_{k\notin\mathcal{P}} \mathbb{E}[\beta_k x_k \mid \boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}].$$

For $k \notin \mathcal{P}$, x_k is mean zero and independent of $(\boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}, \boldsymbol{\beta})$, so $\mathbb{E}[\,\beta_k x_k \mid \boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}] = 0$. Thus

$$\mathbb{E}[y \mid \boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}] = \sum_{k\in\mathcal{P}} x_k\, \mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}].$$

If $k \notin \mathcal{T}$, then β_k is not updated by $\boldsymbol{M}_{\mathcal{T}}$ and $\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}] = \mathbb{E}[\beta_k] = 0$. Hence only indices in $\mathcal{T} \cap \mathcal{P}$ contribute, giving

$$\mathbb{E}[y \mid \boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}}] = \boldsymbol{x}'_{\mathcal{T}\cap\mathcal{P}}\, \mathbb{E}[\boldsymbol{\beta}_{\mathcal{T}\cap\mathcal{P}} \mid \boldsymbol{M}_{\mathcal{T}}],$$

which proves the claim. □

Proposition 1 (Optimal Predictor)**.** *The Bayes Estimator is the posterior mean $\boldsymbol{\beta}$ and satisfies:*

1. *For untrained parameters:*
$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}^c} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \boldsymbol{0}_{|\mathbb{N}|-t};$$

2. *For trained parameters:*

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} + (1 - S(\mathcal{T})) \cdot \boldsymbol{I}_t\right)^{-1} \boldsymbol{X}'_{\mathcal{T}}\boldsymbol{y}.$$

Proof. Because the prior is $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ and only $\boldsymbol{\beta}_{\mathcal{T}}$ enters the likelihood, the joint posterior factorizes as

$$p(\boldsymbol{\beta} \mid \boldsymbol{M}_{\mathcal{T}}) = p(\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}})\, p(\boldsymbol{\beta}_{\mathcal{T}^c} \mid \boldsymbol{M}_{\mathcal{T}}),$$

with $p(\boldsymbol{\beta}_{\mathcal{T}^c} \mid \boldsymbol{M}_{\mathcal{T}}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\mathcal{T}^c})$ since $\boldsymbol{\beta}_{\mathcal{T}^c}$ does not appear in the likelihood and the prior mean is zero. This proves part 1.

For part 2, the likelihood is

$$\boldsymbol{y} \mid \boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}\left(\boldsymbol{X}_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}},\, \sigma^2_{\mathcal{T}}\boldsymbol{I}_N\right), \quad \sigma^2_{\mathcal{T}} = 1 - S(\mathcal{T}),$$

and the prior is $\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{\mathcal{T}})$. By conjugacy (or completing the square), the posterior is Gaussian

$$\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{T}}, \Sigma_{\mathcal{T}}),$$

with precision

$$\Sigma_{\mathcal{T}}^{-1} = \frac{1}{\sigma_{\mathcal{T}}^2} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \boldsymbol{I}_{\mathcal{T}},$$

and mean

$$\boldsymbol{\mu}_{\mathcal{T}} = \Sigma_{\mathcal{T}} \left(\frac{1}{\sigma_{\mathcal{T}}^2} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y} \right) = \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \sigma_{\mathcal{T}}^2 \boldsymbol{I}_{\mathcal{T}} \right)^{-1} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y}.$$

Therefore the Bayes estimator (posterior mean) equals the stated ridge form, which proves part 2. □

Corollary 2 (Ridge Estimator Interpretation)**.** *The posterior mean coincides with a ridge regression estimator with regularization*

$$\lambda(n, \mathcal{T}) \equiv \frac{1 - S(\mathcal{T})}{n}.$$

Equivalently,

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \hat{\boldsymbol{\beta}}_{\boldsymbol{M}_{\mathcal{T}}}^{ridge}(\lambda) \Big|_{\lambda = \lambda(n, \mathcal{T})}.$$

Proof. The ridge objective is strictly convex; its unique minimizer solves the first-order condition:

$$\frac{2}{N} \boldsymbol{X}_{\mathcal{T}}' \left(\boldsymbol{X}_{\mathcal{T}} \hat{\boldsymbol{b}} - \boldsymbol{y} \right) + 2\lambda \hat{\boldsymbol{b}} = \mathbf{0}.$$

Hence

$$\left(\tfrac{1}{N} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \lambda \boldsymbol{I}_{\mathcal{T}} \right) \hat{\boldsymbol{b}} = \tfrac{1}{N} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y}, \quad \text{so} \quad \hat{\boldsymbol{b}} = \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + N\lambda \boldsymbol{I}_{\mathcal{T}} \right)^{-1} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y}.$$

Comparing with the posterior mean from Proposition **??**,

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}] = \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \sigma_{\mathcal{T}}^2 \boldsymbol{I}_{\mathcal{T}} \right)^{-1} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y},$$

we obtain equality when $N\lambda = \sigma_{\mathcal{T}}^2$, i.e. $\lambda = \frac{\sigma_{\mathcal{T}}^2}{N} = \frac{1 - S(\mathcal{T})}{N}$. This proves the claim. □

Lemma 2. *Assume $\mathcal{P} \subseteq \mathcal{T}$. The value of a dataset of covariates $(\mathcal{T}, \mathcal{P})$ is the variance of the optimal predictor*

$$V(n, \mathcal{T}, \mathcal{P}) = Var_{\boldsymbol{D}_{\mathcal{T},\mathcal{P}}^n} \left[f^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}}^n) \right] = \sum_{k \in \mathcal{P} \cap \mathcal{T}} s_k \; w_k(\lambda)|_{\lambda = \lambda(n, \mathcal{T})}.$$

Proof. Under squared loss, the posterior mean minimizes posterior risk, so the ex-ante (expected) value of the dataset equals the prior variance minus the posterior variance of y. Because $y = \boldsymbol{x}_{\mathcal{P}}' \boldsymbol{\beta}_{\mathcal{P}}$ and $\mathcal{P} \subseteq \mathcal{T}$,

$$f^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}}^N) = \boldsymbol{x}_{\mathcal{P}}' \mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}} \mid \boldsymbol{M}_{\mathcal{T}}].$$

Taking the variance over the joint distribution of $\boldsymbol{x}_{\mathcal{P}}$ and $\boldsymbol{M}_{\mathcal{T}}$ gives

$$\operatorname{Var}\left[f^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}}^N)\right] = \operatorname{Var}\left(\boldsymbol{x}_{\mathcal{P}}'\mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}} \mid \boldsymbol{M}_{\mathcal{T}}]\right) = \sum_{k\in\mathcal{P}} \operatorname{Var}(x_k)\operatorname{Var}(\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}]),$$

using independence of covariates. Since $\operatorname{Var}(x_k) = s_k$ and $\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}] = \hat{\beta}^{\text{ridge}}_{\boldsymbol{M}_{\mathcal{T}},k}(\lambda(N,\mathcal{T}))$ by Corollary **??**, we obtain

$$V(N,\mathcal{T},\mathcal{P}) = \sum_{k\in\mathcal{P}} s_k \operatorname{Var}_{\boldsymbol{M}_{\mathcal{T}}}\left[\hat{\beta}^{\text{ridge}}_{\boldsymbol{M}_{\mathcal{T}},k}(\lambda(N,\mathcal{T}))\right] = \sum_{k\in\mathcal{P}} s_k\, w_k(\lambda(N,\mathcal{T})),$$

which proves the result. □

Lemma 3. *[House Party Effect] The posterior mean of β_k satisfies*

$$w_k(\lambda) = \begin{cases} 0, & j\in\mathcal{T}^c, \\ \dfrac{1}{1+\dfrac{\lambda}{s_k}} + O\left(\sqrt{\frac{t}{n}}\right), & j\in\mathcal{T}. \end{cases}$$

Proof. For $k\notin\mathcal{T}$, the parameter β_k is never updated by the training sample $\boldsymbol{M}_{\mathcal{T}}$, so $\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}] = 0$ almost surely. Hence $w_k(\lambda) = \operatorname{Var}_{\boldsymbol{M}_{\mathcal{T}}}[\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}]] = 0$.

For $k\in\mathcal{T}$, the posterior mean from Proposition **??** (or equivalently from ridge regression) is

$$\hat{\boldsymbol{\beta}}_{\mathcal{T}} = \left(\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}} + \lambda N\,\boldsymbol{I}_{\mathcal{T}}\right)^{-1}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{y}, \qquad \lambda(N,\mathcal{T}) = \frac{1-S(\mathcal{T})}{N}.$$

When regressors are mutually independent with $\operatorname{Var}(x_k) = s_k$, the sample covariance matrix satisfies

$$\tfrac{1}{N}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}} = \operatorname{diag}(s_k)_{k\in\mathcal{T}} + O_p\left(\sqrt{\tfrac{|\mathcal{T}|}{N}}\right),$$

where $O_p(\cdot)$ denotes a stochastic bound: a sequence $Z_N = O_p(a_N)$ means that Z_N/a_N is bounded in probability. The random perturbation $O_p\left(\sqrt{|\mathcal{T}|/N}\right)$ propagates through the ridge estimator, but once we take the variance over datasets, the stochastic deviation integrates out, leaving a deterministic approximation error of order $O\left(\sqrt{|\mathcal{T}|/N}\right)$.

Plugging this into the expression for $\hat{\beta}_k$ and taking variance across datasets yields

$$w_k(\lambda) = \operatorname{Var}_{\boldsymbol{M}_{\mathcal{T}}}\left[\hat{\beta}^{\text{ridge}}_k(\lambda)\right] = \frac{s_k}{s_k+\lambda} + O\left(\sqrt{\tfrac{|\mathcal{T}|}{N}}\right) = \frac{1}{1+\frac{\lambda}{s_k}} + O\left(\sqrt{\tfrac{|\mathcal{T}|}{N}}\right).$$

Thus $w_k(\lambda)$ smoothly interpolates between 0 (no learning) and 1 (perfect information), depending on the relative magnitude of the regularization λ and the signal s_k, which proves the claim. □

Proposition 2 (Value of a Dataset)**.** *The value of a dataset of n observations and covariates*

$(\mathcal{T}, \mathcal{P})$ *is*

$$V(n, \mathcal{T}, \mathcal{P}) = \sum_{k \in \mathcal{P} \cap \mathcal{T}} \frac{s_k}{\frac{\lambda(n, \mathcal{T})}{s_k} + 1} + O\left(\sqrt{\frac{t}{n}}\right),$$

where

$$\lambda(n, \mathcal{T}) \equiv \frac{1 - S(\mathcal{T})}{n}.$$

Proof. By Lemma ??, the value of a dataset equals

$$V(N, \mathcal{T}, \mathcal{P}) = \sum_{k \in \mathcal{P}} s_k \, w_k(\lambda(N, \mathcal{T})).$$

Lemma ?? gives

$$w_k(\lambda) = \begin{cases} 0, & k \in \mathcal{T}^c, \\ \dfrac{1}{1 + \dfrac{\lambda}{s_k}} + O\left(\sqrt{\frac{|\mathcal{T}|}{N}}\right), & k \in \mathcal{T}. \end{cases}$$

Substituting this expression into the value formula and using that $\mathcal{P} \subseteq \mathcal{T}$ yields

$$V(N, \mathcal{T}, \mathcal{P}) = \sum_{k \in \mathcal{P}} \frac{s_k}{1 + \dfrac{\lambda(N, \mathcal{T})}{s_k}} + O\left(\sqrt{\frac{|\mathcal{T}|}{N}}\right).$$

The term $\frac{s_k}{1+\lambda(N,\mathcal{T})/s_k}$ increases in the signal s_k and decreases in the penalty λ, capturing the tradeoff between signal strength and regularization. This establishes the theorem. □

Proposition 3 (Economises of Scope in Training)**.** *The value of a dataset is strictly increasing and supermodular in $\mathcal{T}$.*

Proof. We show the marginal contribution is positive. Adding i to $\mathcal{T}$ reduces the penalty $\lambda(\mathcal{T})$ by s_i/n. Hence

$$\Delta_i V(\mathcal{T}) = \sum_{k \in \mathcal{P}} \left[\frac{s_k^2}{s_k + \lambda - \frac{s_i}{n}} - \frac{s_k^2}{s_k + \lambda} \right] = \frac{s_i}{n} \sum_{k \in \mathcal{P}} \frac{s_k^2}{(s_k + \lambda)(s_k + \lambda - \frac{s_i}{n})} > 0.$$

For two covariates i, j, denote $f_k(\lambda) = s_k^2/(s_k + \lambda)$. Then

$$\Delta_{i,j}^2 V(\mathcal{T}) = \sum_{k \in \mathcal{P}} \left[f_k(\lambda - \tfrac{s_i + s_j}{n}) - f_k(\lambda - \tfrac{s_i}{n}) - f_k(\lambda - \tfrac{s_j}{n}) + f_k(\lambda) \right].$$

Because $f_k''(\lambda) = 2s_k^2/(s_k + \lambda)^3 > 0$, each f_k is convex in λ. As adding a covariate reduces λ, the discrete second difference above is strictly positive. Hence $\Delta_{i,j}^2 V(\mathcal{T}) > 0$, proving supermodularity in $\mathcal{T}$. □

Proposition 4 (Additivity in Prediction)**.** *The value of a dataset is strictly increasing and additive in $\mathcal{P}$.*

Proof. Write $f_k(\lambda) = s_k^2/(s_k + \lambda)$, so $V(n, \mathcal{T}, \mathcal{P}) = \sum_{k\in\mathcal{P}} f_k(\lambda(\mathcal{T}))$. Because $\lambda(\mathcal{T})$ depends only on $\mathcal{T}$, changing $\mathcal{P}$ leaves λ unchanged.

(i) Adding i simply adds the term $f_i(\lambda(\mathcal{T}))$, giving $\Delta_i^{\mathcal{P}} V(\mathcal{T}) = f_i(\lambda(\mathcal{T})) = \frac{s_i^2}{s_i+\lambda(\mathcal{T})} > 0$.

(ii) By additivity over $k \in \mathcal{P}$,

$$V(n, \mathcal{T}, \mathcal{P} \cup \{i, j\}) = V(n, \mathcal{T}, \mathcal{P}) + f_i(\lambda(\mathcal{T})) + f_j(\lambda(\mathcal{T})),$$

so the inclusion-exclusion combination cancels exactly and equals zero. □

Proposition 5 (Complementarity/Substitutability of n and Training Scope)**.** *Fix a training set $\mathcal{T}$ and prediction set $\mathcal{P} \subseteq \mathcal{T}$. For any covariate $k \notin \mathcal{T}$ with signal $s_k > 0$, define the cross-effect*

$$\Delta(n; \mathcal{T}, k) := \partial_n V\big(n, \mathcal{T} \cup \{k\}, \mathcal{P}\big) - \partial_n V\big(n, \mathcal{T}, \mathcal{P}\big).$$

Then

$$\Delta(n; \mathcal{T}, k) = s_k \sum_{j\in\mathcal{P}} \frac{s_j^2\Big((1 - S(\mathcal{T}))(1 - S(\mathcal{T}) - s_k) - n^2 s_j^2\Big)}{\big(ns_j + 1 - S(\mathcal{T})\big)^2 \big(ns_j + 1 - S(\mathcal{T}) - s_k\big)^2}. \tag{4}$$

In particular, there exist thresholds

$$\underline{n} = \frac{\sqrt{(1 - S(\mathcal{T}))(1 - S(\mathcal{T}) - s_k)}}{\max_{j\in\mathcal{P}} s_j}, \qquad \overline{n} = \frac{\sqrt{(1 - S(\mathcal{T}))(1 - S(\mathcal{T}) - s_k)}}{\min_{j\in\mathcal{P}} s_j},$$

such that

$$\Delta(n; \mathcal{T}, k) > 0 \ \textit{ for all } \ n < \underline{n} \quad \textit{and} \quad \Delta(n; \mathcal{T}, k) < 0 \ \textit{ for all } \ n > \overline{n}.$$

Hence, when data are scarce (small n) sample size and training scope are complements*, while for abundant data (large n) they are* substitutes.

Let $a := 1 - S(\mathcal{T})$ and $x_j := ns_j$. Write

$$\partial_n V(n, \mathcal{T}, \mathcal{P}) = \sum_{j\in\mathcal{P}} s_j \cdot \partial_n \left(\frac{1}{1 + \frac{a}{ns_j}} \right) = \frac{a}{n^2} \sum_{j\in\mathcal{P}} \frac{1}{\left(1 + \frac{a}{ns_j}\right)^2} = \sum_{j\in\mathcal{P}} \frac{a\, x_j^2}{n^2(x_j + a)^2}.$$

If we add k to training, a becomes $b := a - s_k$ while x_j is unchanged. Hence

$$\partial_n V(n, \mathcal{T} \cup \{k\}, \mathcal{P}) = \sum_{j\in\mathcal{P}} \frac{b\, x_j^2}{n^2(x_j + b)^2}.$$

Subtracting, for each j we obtain

$$\frac{b\, x_j^2}{n^2(x_j + b)^2} - \frac{a\, x_j^2}{n^2(x_j + a)^2} = \frac{x_j^2}{n^2} \cdot \frac{b(x_j + a)^2 - a(x_j + b)^2}{(x_j + b)^2(x_j + a)^2}.$$

The numerator simplifies as

$$b(x_j + a)^2 - a(x_j + b)^2 = ba^2 - ab + (b - a)x_j^2 = ab(a - b) - s_k x_j^2 = s_k\big(ab - x_j^2\big),$$

since $b - a = -s_k$. Therefore

$$\Delta(n; \mathcal{T}, k) = \sum_{j\in\mathcal{P}} \frac{x_j^2}{n^2} \cdot \frac{s_k\big(ab - x_j^2\big)}{(x_j + b)^2(x_j + a)^2} = s_k \sum_{j\in\mathcal{P}} \frac{s_j^2\big(ab - n^2 s_j^2\big)}{\big(ns_j + b\big)^2\big(ns_j + a\big)^2},$$

which is (4) after substituting back $a = 1 - S(\mathcal{T})$ and $b = 1 - S(\mathcal{T}) - s_k$.

Each denominator in (4) is strictly positive, as are s_k and s_j^2. Hence the sign of each summand is the sign of

$$ab - n^2 s_j^2 = \big(1 - S(\mathcal{T})\big)\big(1 - S(\mathcal{T}) - s_k\big) - n^2 s_j^2.$$

If $n < \sqrt{ab}/\max_{j\in\mathcal{P}} s_j$ then $ab - n^2 s_j^2 > 0$ for all j, so every summand is positive and $\Delta(n; \mathcal{T}, k) > 0$. If $n > \sqrt{ab}/\min_{j\in\mathcal{P}} s_j$ then $ab - n^2 s_j^2 < 0$ for all j, so every summand is negative and $\Delta(n; \mathcal{T}, k) < 0$. This proves that the cross-effect is positive for sufficiently small n and negative for sufficiently large n, i.e., complements when data are scarce and substitutes when data are abundant.

Proposition 6 (Sample Size and Prediction Scope are Complements)**.** *Fix a training set $\mathcal{T}$ and let $\mathcal{P} \subseteq \mathcal{T}$. For any $i \in \mathcal{T} \setminus \mathcal{P}$, consider the cross-effect*

$$\Delta^{\mathcal{P}}(n; \mathcal{T}, i) := \partial_n V\big(n, \mathcal{T}, \mathcal{P} \cup \{i\}\big) - \partial_n V\big(n, \mathcal{T}, \mathcal{P}\big).$$

Then

$$\Delta^{\mathcal{P}}(n; \mathcal{T}, i) = \frac{\lambda_{\mathcal{T}}}{n} \cdot \frac{1}{\left(1 + \dfrac{\lambda_{\mathcal{T}}}{s_i}\right)^2} = \frac{(1 - S(\mathcal{T}))\, s_i^2}{n^2\big(s_i + \frac{1-S(\mathcal{T})}{n}\big)^2} > 0 \quad \text{for all } n \geq 1. \tag{5}$$

Hence sample size n and prediction scope (adding a used covariate i) are strict complements *for all n. Moreover, the complementarity strength is decreasing in n.*

Proof. Write $a := 1 - S(\mathcal{T})$ so that $\lambda_{\mathcal{T}} = a/n$. Because $\mathcal{P} \subseteq \mathcal{T}$, changing $\mathcal{P}$ does not affect $\lambda_{\mathcal{T}}$. Adding i to prediction simply adds its term to the value function:

$$V(n, \mathcal{T}, \mathcal{P} \cup \{i\}) - V(n, \mathcal{T}, \mathcal{P}) = \frac{s_i}{1 + \lambda_{\mathcal{T}}/s_i}.$$

Differentiate w.r.t. n. With $f(\lambda) := \dfrac{s_i}{1 + \lambda/s_i}$, we have

$$\frac{\partial f}{\partial \lambda} = -\frac{1}{\big(1 + \lambda/s_i\big)^2}, \qquad \frac{\partial \lambda_{\mathcal{T}}}{\partial n} = -\frac{a}{n^2} = -\frac{\lambda_{\mathcal{T}}}{n}.$$

By the chain rule,

$$\Delta^{\mathcal{P}}(n;\mathcal{T},i)=\frac{\partial f}{\partial \lambda}\cdot\frac{\partial \lambda_{\mathcal{T}}}{\partial n}=\frac{\lambda_{\mathcal{T}}}{n}\cdot\frac{1}{\left(1+\lambda_{\mathcal{T}}/s_i\right)^2},$$

which yields (5). The right-hand side is strictly positive because $\lambda_{\mathcal{T}}>0$ and the denominator is positive. Hence n and prediction scope are complements for all n.

Finally, use $a=1-S(\mathcal{T})$ to rewrite

$$\Delta^{\mathcal{P}}(n;\mathcal{T},i)=\frac{a\,s_i^2}{n^2\left(s_i+a/n\right)^2}.$$

This expression is strictly decreasing in n (its derivative in n is negative), so the complementarity is stronger when data are scarce and weakens as n grows. □

Lemma 4 (Optima Pool Price)**.** *The optimal pool price is*

$$P^*=\frac{V_2}{\alpha+1}.$$

Proof. The pool maximizes revenue

$$\max_P\ P\,D(P-V_2)=P(V_2-P)^{\alpha},$$

which yields the first-order condition

$$(V_2-P)^{\alpha}+\alpha P(V_2-P)^{\alpha-1}=0.$$

Solving gives $P^*=V_2/(\alpha+1)$. □

Proof. We follow Lerner and Tirole (2004). **Demand Margin Binds.** Suppose that the brokers offer prices $\mathcal{P}\equiv(p_1,p_2)$, and wlog $p_1\le p_2$. Prediction Sellers decide how many datasets to buy.

$$\mathcal{V}(\mathcal{P})=\max_{q\in\{1,2\}}\left\{V_q-p_1-p_2\mathbf{1}(\{q=2\})\right\}$$

Second, the user adopts the technology if and only if

$$\mathcal{V}(\mathcal{P})\ge\theta.$$

Lerner and Tirole (2004) demonstrate the existence of a symmetric equilibrium. Individual data sellers solve

$$\hat{p}=\arg\max_{p_i}\left\{p_iD\left(p_i+\hat{p}-V_2\right)\right\}$$

which has FOC

$$\hat{p}D'\left(2\hat{p}-V_2\right)+D\left(2\hat{p}-V_2\right)=0$$

which has a unique solution buy hazard-rate monotonicity. It can be seen as selling the whole pool setting total price P and keeping $p_i = P - \hat{p}$ for itself. Therefore

$$\hat{P} = \arg\max_{P} \left\{ \left(P - \hat{p}\right) D\left(P - V_2\right) \right\}.$$

The term $\hat{p}$ can be seen as a marginal cost $\hat{c} = \hat{p}$. In this interpretation when there is the pool $c^* = 0$ so by revealed preference

$$\hat{P} \geq P^*.$$

If demand margin binds in the absence of a pool then the pool reduces price paid by data buyers. This means that if all datasets can increase the price marginally without being excluded, the pool is pro-competitive.

With our CDF G,

$$p_{\text{dem}} = \frac{V_2}{2+\alpha}.$$

Competition Margin Binds. Define the price when the competition margin binds will be p_{comp} defined by

$$V_2 - 2p_{\text{comp}} = \max_{q\in\{0,1\}} \left\{ V_q - q p_{\text{comp}} \right\}.$$

If $V_1 - p_{\text{comp}} \geq 0$, then $p_{\text{comp}} = V_2 - V_1$. This is consistent because $V_1 - (V_2 - V_1) > 0$ by concavity of $V(n,t)$ in n. Otherwise, if $V_1 - p_{\text{comp}} < 0$, then $p_{\text{comp}} = V_2/2$. This is not consistent because $V_1 - V_2/2 > 0$ by concavity of $V(n,t)$ in n. □

Proposition 17. *A pool of observations is procompetitive if and only if*

$$\alpha > \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1} = \frac{n}{2(1-k)}.$$

Proof. Pools are procompetitive if the demand margin binds i.e.

$$p_{\text{comp}} > p_{\text{dem}} \iff V_2 - V_1 > \frac{V_2}{2+\alpha} \iff \alpha > \alpha_{\text{marg}} \equiv \frac{2V_1 - V_2}{V_2 - V_1}.$$

In this case

$$p_i = \frac{V_2}{2+\alpha}.$$

This implies that observation pooling can be procompetitive when n is not too large and k is not too small, meaning data is relatively abundant and models are relatively complex. Furthermore, as the RHS is increasing in Q, data pools are more likely to be competitive if Q is small meaning if data fragmentation is limited.

Otherwise if the competition margin binds,

$$p_i = V_2 - V_1.$$

the pool is procompetitive if the pool price is lower than the competition price, i.e.

$$P^* < Qz(Q) \iff \frac{V_2}{\alpha+1} < 2\left(V_2 - V_1\right) \iff \alpha > \alpha_{\text{comp}} \equiv \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1}.$$

As $\alpha_{\text{marg}} > \alpha_{\text{comp}}$, the relevant threshold is α_{comp} and so a pool of observations is procompetitive if and only if

$$\alpha > \alpha_{\text{comp}} = \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1}.$$

This implies that as k and n increase it becomes less likely that the pool is procompetitive. When data is abundant and demand is inelastic observation pools are anticompetitive. Direct application of Proposition 5 in Lerner and Tirole (2004) implies that the pool is strongly unstable, therefore enforcing independent licensing of datasets will prevent pooling if and only if the pool is welfare-reducing. In this case each data broker will charge $p_{\text{comp}} = V_2 - V_1$. □

Lemma 11. *If the brokers have distinct covariates each B_i prices at*

$$p_i = \frac{V_2}{2+\alpha}$$

and the buyers will buy from both brokers.

Proof. If $V_1 - p_{\text{comp}} \geq 0$, then $p_{\text{comp}} = V_2 - V_1$. This is not consistent because $V_1 - (V_2 - V_1) < 0$ by convexity of $V(n,t)$ in t. Otherwise, if $V_1 - p_{\text{comp}} < 0$, then $p_{\text{comp}} = V_2/2$. This is consistent because $V_1 - V_2/2 < 0$ by convexity of $V(n,t)$ in t. Demand margin always binds as

$$p_{\text{comp}} > p_{\text{dem}} \iff \frac{V_2}{2} > \frac{V_2}{2+\alpha}.$$

□

Lemma 7 (Shopper-price equilibrium). *Fix (v_1, v_2). The pricing subgame in shopp[Shopper-price equilibrium]er prices admits a unique equilibrium under trembling-hand perfection:*

$$p_i^s = (v_i - v_j)^+, \qquad p_j^s = (v_j - v_i)^+, \qquad i \neq j, \tag{7}$$

so that (i) if $v_i > v_j$, firm i sets $p_i^s = v_i - v_j$, firm j sets $p_j^s = 0$, and all shoppers buy from i; (ii) if $v_i = v_j$, both set $p_i^s = p_j^s = 0$ and shoppers can be split arbitrarily.

Proof. We first show that (7) is an equilibrium. Consider $v_i > v_j$. If i sets $p_i^s = v_i - v_j$ and j sets $p_j^s = 0$, shoppers are indifferent and (by the refinement) all go to i.[20] Any deviation by i:

[20]Formally, with tiny perturbations (trembles) that give i an ε quality advantage or j an ε higher price with positive probability, the unique limit assigns the shoppers to i.

If $p_i^s > v_i - v_j$, then $v_i - p_i^s < v_j - p_j^s = v_j$ and i loses all shoppers, strictly reducing its shopper revenue to 0.

If $p_i^s < v_i - v_j$, i still serves all shoppers but leaves revenue on the table; since demand is inelastic at one (all shoppers) at the margin, profit increases by raising p_i^s up to $v_i - v_j$.

Any deviation by j:

If $p_j^s > 0$, then $v_j - p_j^s < v_j = v_i - p_i^s$ and j serves no shoppers with the same zero shopper revenue; under trembling hand, the weakly dominated positive price is eliminated in the limit, selecting $p_j^s = 0$.

If $p_j^s < 0$ is infeasible; if $p_j^s = 0$ already, no profitable deviation exists.

Thus (7) is an equilibrium when $v_i > v_j$. The case $v_i = v_j$: for any $p_i^s = p_j^s$, shoppers are indifferent and each firm earns $\sigma p_i^s/2$ from shoppers; any unilateral increase loses all shoppers, any decrease reduces price with the same demand, so $p_i^s = p_j^s = 0$ is the unique trembling-hand limit (positive common prices are not robust to small payoff perturbations). Symmetry covers $v_j > v_i$.

Uniqueness under trembling-hand perfection follows from the standard Bertrand-vertical-differentiation logic: if $v_i > v_j$, any equilibrium must have the higher-quality firm serving all shoppers; then the highest sustainable price for i that keeps all shoppers is $v_i - v_j$, and the lower-quality firm's best response is any price with zero demand and zero revenue, refined to 0. □

Lemma 8 (Monotone best response in r_i). *Under Assumption 3, the proprietary-collection best response of F_i is monotone in its license choice:*

$$r_i^{BR}(\ell_i;\, \ell_j, r_j) \;\geq\; \ell_i \qquad \textit{for all } (\ell_j, r_j).$$

Equivalently: if F_i purchases the license ($\ell_i = 1$), then collecting proprietary data ($r_i = 1$) is (weakly) strictly optimal for any opponent action; if i does not license ($\ell_i = 0$), collecting may fail to be profitable in particular opponent configurations.

Proof. Fix the opponent (ℓ_j, r_j).

Step 1 (Licensed case). Consider the marginal gain from collecting when $\ell_i = 1$:

$$\Delta^L(\ell_j, r_j) \;\equiv\; \pi(1, 1; \ell_j, r_j) - \pi(1, 0; \ell_j, r_j).$$

By the profit formula,

$$\Delta^L(\ell_j, r_j) = \frac{1-\sigma}{2}\,[\,v(1,1) - v(1,0)\,] \;+\; \sigma\Big((v(1,1) - v_j)^+ - (v(1,0) - v_j)^+\Big).$$

The second term is weakly nonnegative and is minimized at 0 when v_j is high enough that both positive parts are zero (e.g., when the rival also plays $(1,1)$). Therefore

$$\Delta^L(\ell_j, r_j) \;\geq\; \frac{1-\sigma}{2}\,[\,v(1,1) - v(1,0)\,] \;=\; \bar{c}.$$

Under Assumption 3, $\bar{c} \geq c$, so $\Delta^L(\ell_j, r_j) \geq c$ for all (ℓ_j, r_j). Hence, if $\ell_i = 1$ it is (weakly) strictly optimal to choose $r_i = 1$.

Step 2 (Unlicensed case). Consider the marginal gain from collecting when $\ell_i = 0$:

$$\Delta^U(\ell_j, r_j) \;\equiv\; \pi(0,1;\ell_j, r_j) - \pi(0,0;\ell_j, r_j) = \frac{1-\sigma}{2}\,v(0,1) + \sigma\Big((v(0,1) - v_j)^+ - (0 - v_j)^+\Big).$$

Since $(0 - v_j)^+ = 0$ and the second term is minimized at 0 when $v_j \geq v(0,1)$ (e.g., when the rival plays $(1,1)$), we have

$$\Delta^U(\ell_j, r_j) \;\geq\; \frac{1-\sigma}{2}\,v(0,1) \;=\; \underline{c}.$$

Under Assumption 3, $c \geq \underline{c}$, so there exist opponent actions (e.g., $(\ell_j, r_j) = (1,1)$) for which $\Delta^U(\ell_j, r_j) \leq c$ and collecting is *not* profitable without a license. Therefore, $r_i = 0$ can be optimal when $\ell_i = 0$.

Combining the two steps establishes $r_i^{BR}(\ell_i; \ell_j, r_j) \geq \ell_i$ for all (ℓ_j, r_j). □

Lemma 9 (Data-collection equilibrium by license allocation). *Under Assumption 3, let $\ell_1 + \ell_2 \geq 1$. Then any Nash equilibrium of the data-collection subgame satisfies*

$$(r_1^*, r_2^*) \;=\; (\ell_1, \ell_2)$$

Proof. If $L = 2$, each firm has $\ell_i = 1$ and by Lemma 8 each strictly prefers $r_i = 1$, giving $(1,1)$. If $L = 1$, exactly one firm has $\ell_i = 1$; by Lemma 8 that firm chooses $r_i = 1$, while the unlicensed rival (with $\ell_j = 0$) has an opponent profile available (namely the licensed firm choosing $r_i = 1$) under which collecting is not profitable at $c \geq \underline{c}$, so $r_j = 0$ is optimal in equilibrium. Symmetry yields the two possibilities $(1,0)$ or $(0,1)$. □

Proposition 16 (Seller's optimal licensing policy). *Under Assumption 3, the data seller's optimal licensing fee satisfies*

$$f^* = \begin{cases} f^{NE}, & \text{if } \sigma \leq 1/3 \wedge XXX, \\ f^{NE} + \sigma v(1,1), & \text{if } \sigma > 1/3, \end{cases}, \qquad L^* = \begin{cases} 2, & \text{if } \sigma \leq 1/3, \\ 1, & \text{if } \sigma > 1/3, \end{cases}$$

where

$$f^{NE} \equiv \frac{1-\sigma}{2} v(1,1) - c.$$

Proof. The data seller can indirectly choose the number of licenses $L \in \{1,2\}$ sold by setting (f_1, f_2).

Case 1 (Two licenses). Suppose both firms purchase the license ($L = 2$). For firm i to be willing to license, the gain from doing so must cover the cost of licensing and data collection:

$$\pi(1,1;1,1) - \pi(0,0;1,1) \geq c + f_i.$$

Substituting the profit function gives

$$f_i \leq \frac{1-\sigma}{2} v(1,1) - c \;\equiv\; f^{NE}.$$

The term f^{NE} is thus the maximum fee consistent with both firms licensing (*non-exclusive* regime).

Case 2 (One license). Suppose S intends to sell the license only to firm i ($L = 1$). Then F_i compares profits from licensing versus not licensing when the rival does not license:

$$\pi(1,1;0,0) - \pi(0,0;0,0) \geq c + f_i.$$

Substituting yields

$$f_i \leq \frac{1+\sigma}{2} v(1,1) - c = f^{NE} + \sigma v(1,1).$$

Hence, if S sets

$$f_i^* = f^{NE} + \sigma v(1,1), \qquad f_j^* \in \left(f^{NE} + \sigma v(1,1), \infty\right),$$

only one firm will find licensing profitable. The seller thus ensures a de facto exclusive deal ($L = 1$).

Seller's choice between regimes. Selling to both firms yields revenue $2f^{NE}$, while selling to one yields $f^{NE} + \sigma v(1,1)$. The seller prefers exclusivity if and only if

$$f^{NE} + \sigma v(1,1) > 2f^{NE} \quad \Longleftrightarrow \quad c > \tfrac{1-3\sigma}{2} v(1,1) \equiv c^E.$$

Under Assumption **??**, $\sigma \geq 1/3$ implies $c^E < 0$, and since $c > 0$, the inequality is always satisfied. Hence, the seller strictly prefers to sell the license to a single firm. □

Lemma 10 (Planner's optimum)**.** *Under Assumption 3, it is socially optimal to sell licenses to both firms ($L = 2$).*

Proof. Because licensing fees and prediction prices are transfers, total welfare depends only on prediction quality and data-collection costs.

Case 1 ($L = 2$): Both firms license and collect proprietary data, so both reach quality $v(1,1)$. Total welfare is

$$TW^{NE} = v(1,1) - 2c,$$

as each firm serves its captives and shoppers split according to quality (here equal).

Case 2 ($L = 1$): Only one firm licenses and invests, attaining quality $v(1,1)$, while the other remains at $(0,0)$ and serves only its captives. Welfare is then

$$TW^E = \frac{1+\sigma}{2} v(1,1) - c.$$

Exclusivity is socially desirable only if $TW^E > TW^{NE}$, i.e.

$$c > \frac{1-\sigma}{2} v(1,1).$$

Under Assumption 3, $c \leq \bar{c} = \frac{1-\sigma}{2}[v(1,1) - v(1,0)] < \frac{1-\sigma}{2} v(1,1)$, so this condition cannot hold. Therefore, exclusivity is never socially optimal. $\square$

Corollary 5 (Conflict between private and social incentives)**.** *Under Assumptions 3 and* **??***, the data seller finds exclusivity profitable even though total welfare would be higher under non-exclusive licensing.*

B Extensions

House i	y^i	x^i_{size}	x^i_{year}	x^i_{dist}	x^i_{sun}	
0	?	x^0_{size}	NA	x^0_{dist}	NA	Prediction Vector: $\boldsymbol{x}'_{\mathcal{P}}$
1	y^1	x^1_{size}	x^1_{year}	NA	x^1_{sun}	Training Matrix: $\boldsymbol{M}^{(n)}_{\mathcal{T}}$
2	y^2	x^2_{size}	x^2_{year}	NA	x^2_{sun}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	y^n	x^n_{size}	x^n_{year}	NA	x^n_{sun}	

Table 1: Example of Zillow dataset with prediction covariates $\mathcal{P} = \{\text{size}, \text{dist}\}$ and training covariates $\mathcal{T} = \{\text{size}, \text{year}, \text{sun}\}$, where *size* denotes square meters, *dist* the distance to the nearest supermarket, *year* the construction year, and *sun* the daily sunlight exposure.

B.1 Scope as Model Complexity and LLMs

Scope as Model Complexity. Instead, make no restriction on Σ. Furthermore suppose the firm observes all covariates for all individuals but faces constraints on the number of covariates it can effectively use in the learning and targeting steps. The scope of learning, ℓ, is the number of principal components the firm can use in learning. The scope of targeting, t, is the number of principal components that can be used in targeting. This interpretation captures the *model complexity*, which reflects the higher computing cost deriving from analyzing more covariates.

To reduce the dimensionality whilst extracting the maximum information in the constraints, Jolliffe (2002) shows that the optimal procedure is Principal Component Analysis

(PCA). Let the eigendecomposition of the variance/covariance matrix be

$$\Sigma \;=\; \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}', \qquad \boldsymbol{S} = \operatorname{diag}(s_1 \geq \cdots \geq s_{\bar{\ell}} \geq 0), \qquad \boldsymbol{U} \text{ orthonormal.}$$

Define principal components $\boldsymbol{z}^i \;\equiv\; \boldsymbol{x}^i\boldsymbol{U}$. Then

$$\boldsymbol{z}^i \sim \mathcal{N}(0, \Lambda), \qquad z^i_j \text{ are uncorrelated with variances } s_j.$$

Remark 2 (Application to Large Language Models (LLMs)). Although LLMs are trained with cross-entropy loss, near a trained solution their behavior can be well approximated by a linear predictor under squared loss in a suitable linear transformation of the covariates (MacKay (1992); Jacot, Gabriel, and Hongler (2018)). In this local view, our primitives map directly: the scale of learning n corresponds to the amount of training information (e.g., the number of training observations/tokens), the scope of learning ℓ captures the effective number of informative directions used at the learning stage, and the scope of targeting t captures the amount of information observed at the targeting stage for specific instances. Under this mapping, comparative statics in (n, ℓ, t) align with empirical scaling laws for language models (Kaplan et al. (2020)). Supplying richer information at prediction time corresponds to increasing t via retrieval-augmented inputs (P. Lewis et al. (2020)), with benefits contingent on relevance and known long-context effects (Liu et al. (2023)).

B.2 Shrinkage Interpretation

We express the Bayes estimator in terms of a generalization of the ordinary least-squares (OLS) estimator — the minimum-norm least-squares (MNLS) estimator, defined as

$$\hat{\boldsymbol{\beta}}^{\text{MNLS}}_{\mathcal{T}} \equiv \left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}}\right)^{+}\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{y} = \begin{cases} \hat{\boldsymbol{\beta}}^{\text{OLS}}_{\mathcal{T}}, & \text{if } |\mathcal{T}| \leq n, \\ \min_{\boldsymbol{b}_{\mathcal{T}}} \left\{ \|\boldsymbol{b}_{\mathcal{T}}\|_2 \,:\, \boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}_{\mathcal{T}} = \boldsymbol{y} \right\}, & \text{if } |\mathcal{T}| > n, \end{cases}$$

where $(\cdot)^{+}$ denotes the Moore–Penrose pseudo-inverse.[21] The MNLS is the estimator that the firm would adopt if the residual variance were approximately zero (i.e., the cumulative signal $S(\mathcal{T}) \approx 1$). It comes in two flavors, depending on whether the number of parameters is greater than the sample size:

- Underparametrized regime ($n \geq |\mathcal{T}|$): the MNLS estimator coincides with the OLS estimator, which is uniquely defined because $\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}}$ is invertible.

[21]For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, the Moore–Penrose pseudo-inverse is the unique matrix $\boldsymbol{A}^{+} \in \mathbb{R}^{m\times n}$ satisfying

$$\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{A}, \quad \boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{A}^{+}, \quad (\boldsymbol{A}\boldsymbol{A}^{+})' = \boldsymbol{A}\boldsymbol{A}^{+}, \quad (\boldsymbol{A}^{+}\boldsymbol{A})' = \boldsymbol{A}^{+}\boldsymbol{A}.$$

- Overparametrized regime ($n < |\mathcal{T}|$): the OLS estimator is not defined because the system $\boldsymbol{X}_{\mathcal{T}} \boldsymbol{b}_{\mathcal{T}} = \boldsymbol{y}$ has infinitely many solutions; the MNLS chooses the solution with the smallest Euclidean norm.

The MNLS is useful because it is well-defined in both regimes and coincides with the maximum-likelihood estimator. The Bayes estimator is a shrinkage of the MNLS estimator towards the prior mean $\boldsymbol{0}_{|\mathcal{T}|}$

Corollary 6. *The Bayes Estimator is the MNLS estimator with shrinkage:*

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \left(\underbrace{(1 - S(\mathcal{T}))}_{\textit{Shrinkage Factor}} \cdot \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}}\right)^{+} + \boldsymbol{I}_{\mathcal{T}} \right)^{-1} \hat{\boldsymbol{\beta}}_{\mathcal{T}}^{MNLS}.$$

Because it is the maximum likelihood estimator, the MNLS estimator attributes all the variation in the learning matrix $\boldsymbol{M}_{\mathcal{T}}$ to the parameters $\boldsymbol{\beta}_{\mathcal{T}}$. In reality, a fraction $1 - S(\mathcal{T})$ of the variation in $\boldsymbol{y}$ is residual variance and not due to $\boldsymbol{\beta}_{\mathcal{T}}$. The posterior mean corrects for this by shrinking $\hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\text{MNLS}}$towards the prior mean $\boldsymbol{0}_{|\mathcal{T}|}$ with a shrinkage factor equal to the residual variance $1 - S(\mathcal{T})$. Adding a new covariate $j \notin \mathcal{T}$ reduces the residual variance by s_j, the variance of x_j, thereby lowering the shrinkage factor and the weight of the prior mean. Hence, the posterior mean moves closer to the MNLS estimator. Hence, covariates lend precision to each other: observing a new variable improves the accuracy of the estimated parameters of the others.

B.3 Double Descent

Corollary 7. *If covariates in $\mathcal{L}$ are highly informative, the Bayes Estimator is equivalent to the ridgeless estimator and the MNLS estimator*

$$\lim_{S(\mathcal{L}) \to 1^-} \mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}\right] = \lim_{\lambda \to 0^+} \hat{\boldsymbol{\beta}}_{\mathcal{L}}^{ridge}(\lambda) = \hat{\boldsymbol{\beta}}_{\mathcal{L}}^{MNLS}.$$

In general, sophisticated algorithms are needed to compute or approximate the posterior mean $\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}\right]$. Instead, the MNLS can be obtained by a simple machine learning algorithm, *gradient descent*. This equivalence therefore shows that once the data is sufficiently rich, even such a rudimentary algorithm approximates the Bayes estimator arbitrarily well. When data is linear-separabler, prediction accuracy is driven almost entirely by data, not by algorithms.

Remark 3. The result also sheds light on a central puzzle in modern statistics and machine learning: the double descent phenomenon first discusssed in Belkin et al. (2019). Classical statistics tells us the prediction error of gradient descent is U-shaped in the number of parameters $|\mathcal{L}|$: with too few parameters the model underfits, while beyond the optimum $|\mathcal{L}|^* \in (0, n)$ prediction error increases due to overfitting, as residual variation $\boldsymbol{\varepsilon}$ is mistakenly attributed to

$\boldsymbol{\beta}_{\mathcal{L}}$. However, empirical work shows that expanding $\mathcal{L}$ further can reduce the error again—the second descent in the error. Double descent is not yet fully understood: the dominant explanations rely on intricate properties of high-dimensional geometry (see Hastie et al. (2020)). Our model offers a simpler account that also applies to low-dimensions. As the learning set $\mathcal{L}$ expands, the residual variance $1 - S(\mathcal{L})$ decreases, and the shrinkage operator in the Bayes estimator vanishes. When $S(\mathcal{L}) \approx 1$, the Bayes estimator is arbitrarily close to the MNLS even in finite samples, so gradient descent is approximately optimal.

B.4 Connection with Shannon's Information Theory

Remark 4. Let a real-valued additive white Gaussian residual variance (AWGN) channel be given by

$$y = w + z, \qquad z \sim \mathcal{N}(0, \sigma^2),$$

with an input power conslearnt $\mathbb{E}[w^2] \leq P$. Classical results due to Shannon (1948) show that the mutual information between w and y is[22]

$$I(w; y) = \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right) \quad \text{nats.} \tag{R.1}$$

If the channel is decomposed into independent "frequency" slices indexed by $j \in \mathcal{T}$ that each carry an SNR of

$$\mathrm{SNR}_j = \frac{s_j}{\lambda^*},$$

then (R.1) adds up across slices by orthogonality. The total mutual information revealed by a learning sample of *strength* t is therefore[23]

$$I_{\mathcal{T}}(\lambda^*) = \frac{1}{2} \sum_{j \in \mathcal{T}} \log_2 \left(1 + \frac{s_j}{\lambda^*}\right). \tag{R.2}$$

Equation (R.2) is exactly the functional that appears in our model. Thus the economic value function I study,

$$v(t) = \sum_{j \in \mathcal{T}} \frac{t\,\lambda_j}{1 + t\,\lambda_j},$$

equals

$$v(\mathcal{L}, \mathcal{T}) = 2 \left(\frac{I'_{\mathcal{T}}\left(\lambda^*(\mathcal{L})\right)}{\lambda^*(\mathcal{L})} - I_{\mathcal{T}}\left(\lambda^*(\mathcal{L})\right) \right),$$

linking our "value of accuracy" directly to the canonical Shannon measure of information. Two substantive insights follow:

[22] See C. E. Shannon, *Bell System Technical Journal*, 1948, eq. (26); or T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., §9.1.

[23] This integral form follows immediately from Gallager, *Information Theory and Reliable Communication*, 1968, Ch. 8, where parallel Gaussian sub-channels are treated.

1. **Capacity-driven diminishing returns.** Because $I''(t) < 0$ by Shannon's law, marginal economic value $v'(t) = 2I'(t)$ must also fall. No additional curvature assumption is needed; the concavity of v is pinned down by fundamental information limits. In policy terms, data economies of scale saturate exactly when further capacity gains are information-theoretically expensive.

Table 2: Types of predictions and policy implications

Type of prediction	**Data abundant?**	**Tails thick?**	**Monopoly Remedy**
Genomic risk prediction (health)	No	Yes	Access regulation
Clinical decision support for rare diseases	No	Yes	Access regulation
Credit scoring / SME default probability	No	Yes	Access regulation
Fraud / AML detection	No	Yes	Access regulation
Industrial predictive maintenance (OEM IoT)	No	Yes	Access regulation
Smart grid anomaly detection (critical infra)	No	Yes	Access regulation
Autonomous driving safety edge cases	Yes	Yes	Hybrid
Weather nowcasting for extremes	Yes	Yes	Hybrid
E-commerce CTR / product recommendation	Yes	No	Competition policy
Targeted Ads	Yes	No	Competition policy
Media streaming recommendation	Yes	No	Competition policy
Web search ranking	Yes	No	Competition policy