Opening the Black Box*

A Theory of the Value of Data

Giovanni Colla Rizzi†

November 3, 2025

Abstract

This paper develops a theory of the value of data for prediction purposes. An agent uses training data (observations of covariates and 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 observations are complements when data is scarce and otherwise substitutes; and (iii) training and prediction data are complements.

These patterns have several implications. First, data-driven acquisitions may lead to data concentration, all the more so when sample sizes due, e.g., to privacy rules. Second, pooling covariates is always pro-competitive, whereas pooling observations can be either pro- or anticompetitive. Thirdly, a data seller may profitably conclude exclusivity agreements with a firm selling predictions even if this harms social welfare.

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: Amazon and Uber predict where demand will arise, Google and Meta predict which ad a user will click, and Spotify and Netflix predict what content a user will enjoy most.

Furthermore, better predictions attract more users, and more users generate more data: if more data substantially improves predictions, a self-reinforcing loop between users and data emerges. As former Google CEO Eric Schmidt noted, "Scale is the key. We just have so much scale in terms of the data we can bring to bear".[1]

Policymakers increasingly view this feedback loop as a potential source of barriers to entry.[2] In 2022, the EU's Data Act proposal warned that "market imbalances arising from the concentration of data restrict competition and increase barriers to entry".[3] In the United States, the 2021 House Report concluded that "data advantages [...] can reinforce dominance and serve as a barrier to entry".[4] By contrast, tech firms argue that there are diseconomies of scale in data, so that the ability of additional data to improve predictions declines rapidly as datasets grow. In that case, the feedback loop between users and data would break down once datasets are large: new users add little informational value, and prediction quality no longer improves.

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

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

[2] In 2019, the Stigler Committee's Final Report on Digital Platforms and the UK Competition and Markets Authority's Digital Competition Expert Panel Report argued that data concentration can be a barrier to entry.

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

[4] 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`.

economies of scope and scale of data.

Firstly, there are always *economies of scope in training covariates.* This is because collecting 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 degrade predictive accuracy, integrating user profiles is essential—especially for scale-ups with mid-sized user bases.

Finally, to study the policy implications, I develop two applications. First, when prediction can be a natural monopoly. In that case, dividing data across firms may reduce total surplus by increasing the cost of a given reduction of prediction error. Furthermore, by limiting data availability, privacy regulation can make natural monopoly outcomes more likely, giving rise to a 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.[5]

[5]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".

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 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 infinitely countable 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.

Parameters Parameters are unknown, independent of x_k^i, and mutually independent across $k \in \mathcal{K}$:[6]

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

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

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

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

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

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 $(\boldsymbol{M}_{\mathcal{T}}^n, \boldsymbol{x}_{\mathcal{P}})$ 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

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

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

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

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

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

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

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

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

Denote an ***optimal predictor*** by[7]

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

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

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

The best prior prediction, the prior mean $\hat{y} = 0$, gave a loss equal to the prior variance $\mathrm{Var}[y] = 1$. We will treat $n \in \mathbb{R}$ as a scalar throughout.

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

$$C\left(n,\mathcal{T},\mathcal{P}\right) = np_n + \gamma_t\sum_{k\in\mathcal{T}} p_k + \gamma_n\sum_{k\in\mathcal{P}} p_k,$$

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

for tractability, where np_n is the sample cost $\gamma_t \sum_{k\in\mathcal{T}} p_k$ is the training cost and $\gamma_p \sum_{k\in\mathcal{P}} p_k$ is the prediction cost.

The solution of the problem will be the ***optimal sample size*** and the ***optimal prediction model***:

$$(n^*, \mathcal{T}^*, \mathcal{P}^*) \in \arg\sup_{n,\mathcal{T},\mathcal{P}} V(n, \mathcal{T}, \mathcal{P}).$$

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.

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:[8]

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

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

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

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

Although $\varepsilon^i_{\mathcal{T}}$ is not Gaussian, we follow the quasi-maximum-likelihood approach (Gourieroux, Monfort, and Trognon (1984), Bollerslev and Wooldridge (1992) and White (1982)), which ensures consistency when first and second moments are correctly specified.

Assumption 1 (*Gaussian approximation*)**.** *We approximate the misspecification error as a homoskedastic Gaussian:*

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

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 misspecfication 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:*

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

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

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

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 $s_k \equiv \text{Var}(x_k)$, 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 $\text{Var}(\beta_k) = 1$; 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. [9]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.

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.[10]

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.

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

[10]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.

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.

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

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

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

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

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

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

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

$$\max_{n,t,p} v(n,t,p) - C(n,t,p).$$

4.2 Constrained Covariate Selection

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

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

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:

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

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)**.**

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

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

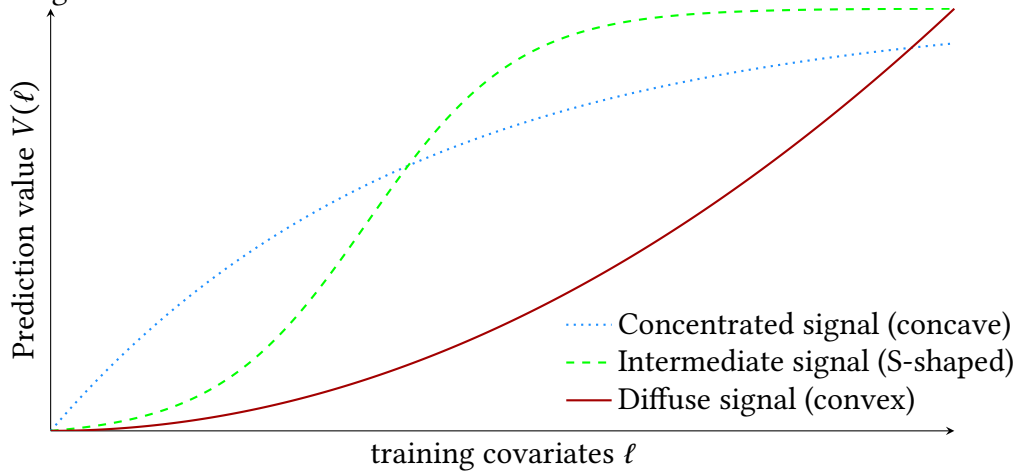
$$v_{tt}(t) = \int_0^p s(u)\left(\underbrace{w_\lambda(u;\lambda)\lambda_{tt}(t)}_{SE<0} + \underbrace{w_{\lambda\lambda}(u;\lambda)\lambda_t^2(t)}_{HPE+CS>0}\right)du.$$

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.

4.4 Optimal Dimensionality

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.

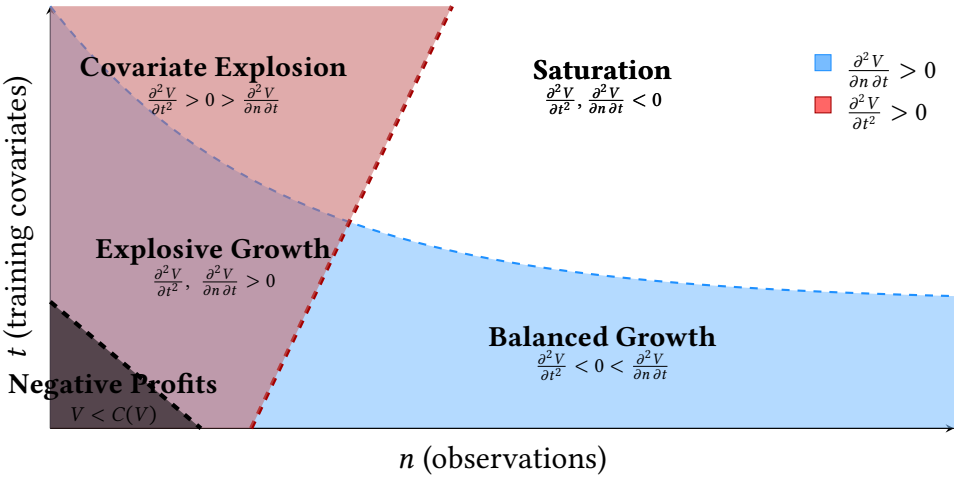
Covariate Selection Proposition 8 shows that when choosing which variables to collect managers should balance two aspects:

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

6.1 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 $s \in \{I, E\}$ at price P_s, the buyer obtains expected utility

$$U(P_s, V_s) = -\alpha \mathbb{E}_D\left[(\hat{y}^*(\boldsymbol{D}) - y)^2\right] - P_s = -(1 - V_s) - P_s,$$

for $\alpha > 0$ is the importance of the decision, where $\hat{y}^*(\boldsymbol{D})$ is the optimal prediction characterized in Lemma 1 and V_s is the value of the seller's dataset.[11]

The outside option is to make the prior mean decision, yielding $\bar{U} = -\alpha$. The net surplus

[11] 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.

from purchasing a prediction is therefore

$$u(P_s, V_s) \equiv U(P_s, V_s) - \bar{U} = \alpha V_s - P_s.$$

For simplicity, assume all covariates have identical variance, so that

$$V(n, t, p) = \frac{\min\{p, t\}}{n + 1 - t}.$$

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 a fee F.

There is a publicly available dataset of p covariates available for both training and prediction on all potential users. Both firms have observed n past decisions. In addition, I observes $t_I \in [0, 1-p]$ proprietary training covariates, and E observes $t_E \in [0, 1-p-t_I]$ distinct proprietary training covariates. If I acquires E, it gains access to E's additional training covariates.

The prediction value generated by seller $s \in \{I, E\}$ is

$$V_s \equiv V(n, p + t_s, p),$$

and under acquisition the combined value is

$$V_{IE} \equiv V(n, p + t_I + t_E, p).$$

Demand for predictions from seller s is denoted $Q(P_s, V_s)$, representing the measure of buyers with nonnegative net surplus $u(P_s, V_s) \geq 0$.

The incumbent's payoff is

$$\Pi_I(F, P_I, P_{IE}) = \begin{cases} P_{IE}\, Q(P_{IE}, V_{IE}) - F, & \text{if } a = 1 \text{ (acquisition)}, \\ P_I\, Q(P_I, V_i)\,, & \text{if } a = 0 \text{ (no acquisition)}. \end{cases}$$

The entrant's payoff is

$$\Pi_E(a, P_E) = \begin{cases} F, & \text{if } a = 1 \text{ (acquisition)}, \\ P_E\, Q(P_E, V_E)\,, & \text{if } a = 0 \text{ (no acquisition)}. \end{cases}$$

Timing The game has two phases

Step 1 (Bid): I decides a bid F and E decides whether to accept $a = 1$ or reject $a = 0$.

tep 2 (Prediction): if $a = 1$, I sets P_{IE}. If $a = 0$, 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

The buyer of type θ purchases the prediction if and only if

$$\alpha \geq \frac{P}{V}.$$

The resulting demand function is

$$Q(P, V) = \mathbf{1}(\alpha V \geq P).$$

Demand decreases in the price P and increases in the dataset value V: richer datasets raise willingness to pay, while higher prices exclude lower-θ types. The inverse demand curve $P(V, Q)$ highlights that the platform can charge higher prices when predictions are more informative or when serving fewer advertisers.

6.1.3 Optimal Prediction Price

Proposition 12. *Since there are no costs per buyer, the Seller chooses P to maximize revenue:*

$$P = \alpha V$$

In order to acquire E, I must offer her the value of its outside option. Therefore

$$F^* = \alpha V_E.$$

Therefore I acquires E if and only if

$$V_{EI} \geq V_i + V_E \iff n \leq \bar{n} \equiv \sqrt{t_E t_I} + p + t_E + t_I - 1$$

Policies that restrict n and expand p will make it more likely that I buys E.

6.1.4 Privacy/Efficiency/Competition Trilemma

Suppose that the regulator assigns a positive value to entry because of positive externalities (either due to dynamic effects or to spillovers). This reflects a positive value of ***competition***: entry is good. Privacy regulation limiting observations will lower n or p. Efficiency implies acquisition should happen if $V_{EI} \geq V_i + V_E \iff n \leq \bar{n} \equiv \sqrt{t_E t_I} + p + t_E + t_I - 1$. If policymakers act on n they therefore face a trilemma: they can achieve at most two of the following three objectives:

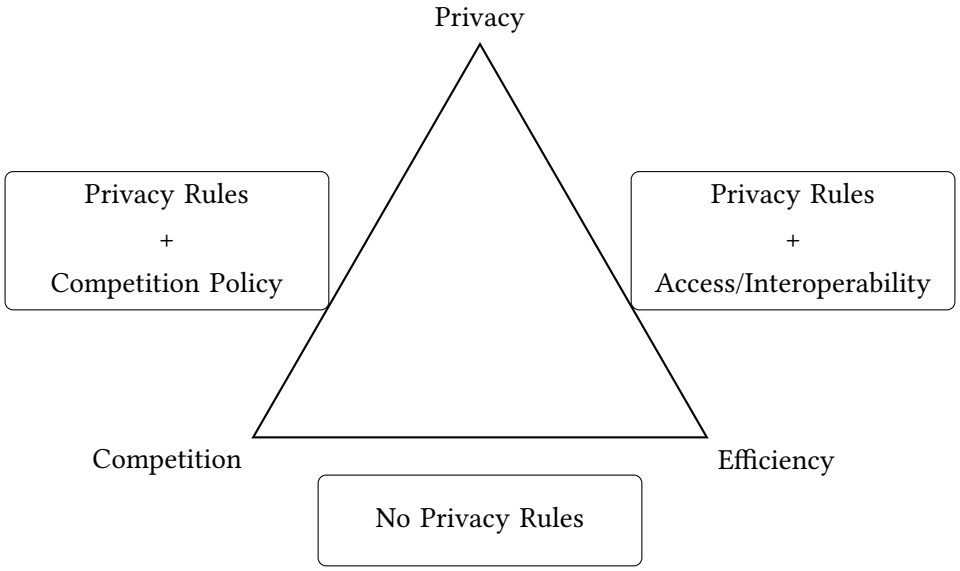
- **Privacy:** limiting data collection and sharing;
- **Competition:** preventing concentration of data in the hands of one firm;
- **Efficiency:** exploiting increasing returns in learning by concentrating data.

Prioritizing privacy forces a trade-off. Policymakers may either:

1. allow data concentration and preserve efficiency through regulated access (*privacy + efficiency*); or
2. limit concentration to preserve competition, at the cost of efficiency (*privacy + competition*).

Instead, by restricting p, policymakers can protect privacy while making it more efficient for the entrant to refuse to be acquired.

Regulatory Tensions. This trilemma reflects real-world policy debates. For example, the EU's General Data Protection Regulation (GDPR) emphasizes privacy, but by raising compliance costs and limiting data flows, it may unintentionally strengthen dominant incumbents that can absorb these costs, thereby reducing competition. Conversely, open-data or portability initiatives aim to enhance competition by lowering barriers to entry, but they risk undermining privacy protections. Finally, efficiency considerations often push regulators to tolerate concentration, as in cases where large integrated datasets (e.g. for health or financial markets) are needed to achieve socially valuable predictions. The trilemma thus captures the inherent difficulty of reconciling these three objectives simultaneously.



6.2 Data Pools

Data owners often sign partnerships to pool their data and sell access to it jointly. For instance, BMW, Mercedes-Benz, and Audi jointly founded the platform Here Mobility Data Marketplace, which aggregates GPS, speed, and road condition data from connected cars. Gu, Madio, and Reggiani (2021) analyze the problem of two brokers selling datasets which can be complements or substitutes, finding that agreements are neutral when they are complements and collusive if they are substitutes. We extend their analysis, developing a model in the spirit of Lerner and Tirole (2004). All these works can be seen as extensions of the fundamental complements problem highlighted in Cournot 1838.

6.2.1 Setup

Data Owners Data is divided in datasets of identical informativeness. Suppose there is a dataset of n observations and t covariates. However it is split in two pieces, either along the n dimension or the t dimension. All parties are symmetrically informed.

Data Buyers Data combination involves no costs. There is a continuum of buyers who can buy the data and combine it to make better predictions. Buyers are heterogeneous and are indexed by a parameter $\theta \in [\underline{\theta}, \bar{\theta}]$. Buyer θ's gross surplus from using q datasets, for $q \in \{1, 2\}$, is[12]

$$V_q - \theta \qquad \text{where } V_q \equiv \begin{cases} V\left(\frac{2n}{q}, t\right), & \text{if split along } n \text{ dimension,} \\ V\left(n, \frac{2t}{q}\right), & \text{if split along } t \text{ dimension.} \end{cases}$$

Since we have shown $V(\cdot, \cdot)$ is increasing in both arguments, prediction is optimal when we combine as many datasets as possible. The parameter θ reflects the heterogeneity in the fixed cost for the buyer to adopt the prediction technology, (b) other benefits from prediction. Sppose the CDF of θ is $G(\theta) = \theta^\alpha$ with $\alpha \in [0, 1]$, to guarantee that the hazard-rate $g/(1-G)$ is strictly increasing. The demand for the bundle of 2 datasets at price $P$is

$$D(P - V_2) = \Pr(V_2 - \theta \geq P) = (V_2 - P)^\alpha .$$

6.2.2 Pooling

Let P^*denote the optimum

$$P^* = \arg\max_P \{PD(P - V_2)\},$$

The pool price will be

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

[12]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.

6.2.3 Fragmentation

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

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

6.2.4 Observation Fragmentation

Lemma 4. *If the brokers have different observations on the same covariates, $p_{comp} = V_2 - V_1$*

Proof. 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. □

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

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

This implies that observation pooling can be procompetitive when n is not too large and t 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, the pool is procompetitive if its price is lower than the competition price, i.e.

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

As $\alpha_{\text{comp}} < \alpha_{\text{dem}}$, a pool 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 t and n increase it becomes less likely that the pool is procompetitive. When data is abundant and highly fragmented, 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 $p_{\text{comp}} = V_2 - V_1$.

6.2.5 Covariate Fragmentation

Lemma 5. *If the brokers have different observations on the same covariates,* $p_{comp} = \frac{V_2}{2}$

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

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

Therefore the pool 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

The 2024 Reddit–OpenAI agreement, which granted OpenAI preferential access to Reddit's user-generated content for training its models, illustrates growing concerns that exclusive data deals may distort competition in AI markets. To formalize these concerns, I develop a model of a data seller that can license its dataset to two competing firms selling predictions, each of which can also invest in collecting proprietary covariates. As data is non rival, the data seller faces a similar predicament to a durable-good monopolist of the Coase conjecture: once it has contracted with one of the firms, it has an incentive to provide access to the other firm, even though the latter will compete with the former, reducing its profits. Similarly to what Katz and Shapiro (1986) establishes for licensing of patents, exclusivity prevents opportunistic behavior on part of data sellers and might therefore raise their revenue from selling data. However, exclusivity entails a trade-off from a social welfare perspective: it alleviates business-stealing between firms but reduces the investment incentives of excluded firm because its proprietary covariates and those of the data seller are complements. I characterize the conditions under which exclusivity is profitable for the data seller yet socially undesirable—showing that this occurs when data are abundant and competition between prediction firms is fierce.

6.3.1 Setup

There are three players: a data seller S (e.g., Reddit) and two prediction firms F_1 and F_2. Firms compete to sell predictions to consumers whose utility depends on the quality 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 controls the intensity of competition: when $\sigma = 0$, all consumers are captive and firms act as local monopolists; when $\sigma = 1$, all consumers are shoppers and the market becomes fully competitive. This simple structure captures the idea that data quality matters only through its relative level across firms, since shoppers migrate toward the firm that offers more accurate predictions. Consumer utility from firm $i \in \{1, 2\}$ is $V_i - p_i$,where V_i is the quality of the prediction (the value of data) and p_i is the price charged to that consumer group. Shoppers purchase from the firm offering the higher net utility $V_i - p_i$.

Prediction Supply

Each firm $i \in \{1, 2\}$ chooses its data strategy and prices. Firm i can improve its prediction quality by collecting covariates in two ways:

1. **R&D investment.** The firm acquires k proprietary covariates at fixed cost $c > 0$, represented by a binary choice $r_i \in \{0, 1\}$.

2. **Data licensing.** The firm can license $k_S \in [0, 1 - k)$ covariates of S by paying a license fee f, represented by a binary choice $\ell_i \in \{0, 1\}$.

We assume throughout that training covariates are also prediction covariates and vice versa.

A data strategy is $(\ell_i, r_i) \in \{0, 1\}^2$ the resulting data stock is

$$k_i = k_S \ell_i + k r_i.$$

By Corollary **??**, the prediction quality function is

$$V_i = v(k_i), \qquad v(k_i) = \frac{n k_i}{n + 1 - k_i},$$

assuming covariates in k and k_S are distinct and i.i.d. and denoting by n the number of observations. We denote

$$V_\ell \equiv v(k_S), \quad V_r \equiv v(k), \quad V_{\ell r} \equiv v(k_S + k).$$

Firms can price discriminate between shoppers and captive consumers:

$$p_i^s \text{ (price to shoppers)}, \qquad p_i^c \text{ (price to captives)}.$$

Let D_i^s denote firm i's share of shoppers, so that $D_1^s + D_2^s = \sigma$. We denote the profit from prediction as

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

Firm i's profit is

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

Data Demand The data seller S has a monopoly on k_S covariates and can choose:

- whether to sell the covariates to both firms or *exclusively* to one firm, represented by a binary choice $e \in \{N, E\}$;
- the license fee $f_e \geq 0$, possibly differing under exclusivity and non-exclusivity.

The data seller's profit is

$$\Pi_S = \begin{cases} f_E, & \text{if exclusive and one firm buys,} \\ 2 f_N, & \text{if non-exclusive and both buy.} \end{cases}$$

Note that economic exclusion of one of the firms does not require that S refrain from dealing therewith but could take the form of an extravagant price, amounting to constructive

refusal. Exclusivity may also be a means of achieving a de facto merger between S and one of the firms when doing so de iure would alarm regulators.

Timing We assume the game takes place in four steps:

1. **Data seller's offer:** S chooses whether to sell the dataset exclusively or non-exclusively, and sets the license fee f_e. Offers are take-it-or-leave-it.
2. **Data collection:** Each firm i decides $(\ell_i, r_i) \in \{0,1\}^2$.
3. **Price competition:** Firms simultaneously set prices (p_i^c, p_i^s). Consumers choose the firm maximizing utility; profits are realized for firms and S.

We characterize a Subgame-Perfect Equilibrium (SPE) of the four-stage game, solving by backward induction.

6.3.2 Bertrand Competition in Prediction

The firms will price at value on non-shoppers:

$$p_i^c = V_i.$$

As non-shoppers have no outside option, the firms can extract all the value by doing price discrimination.

Firms will compete a la Bertrand for shoppers, leading to an equilibrium price and demand given by

Lemma 6. *The pricing subgame has a unique Nash Equilibrium*

$$p_i^s = (V_i - V_j)^+ \quad \textit{and} \quad D_i = \begin{cases} \sigma, & \textit{if } V_i > V_j, \\ \frac{\sigma}{2}, & \textit{if } V_i = V_j, \\ 0, & \textit{if } V_i < V_j. \end{cases}$$

This will imply that the equilibrium profit of F_i if it plays data strategy S_i and F_j plays S_j is

$$\pi(S_i, S_j) = \frac{1-\sigma}{2} V_i + \sigma(V_i - V_j)^+.$$

Intuitively, the fraction of shoppers increases the payoff of the firm with prediction of the highest quality.

6.3.3 Data Collection

Having characterized the continuation payoff $\pi(S_i, S_j)$ we now find the Nash Equilibrium of the data collection subgame. Incentives to invest in data collection depend on whether

the firm already holds a license to use S's covariates. If a firm has acquired the license, the marginal return from adding proprietary data is high because the two sets of covariates k and k_S are complementary. Without the license, the marginal return is low because the proprietary dataset alone offers limited predictive improvement. Assumption 3 restricts attention to values of c for which firms invest if and only if they have access to the licensed covariates.

Assumption 3. *Firms invest in proprietary data if and only if they have purchased a license, i.e.,* $c \in [\underline{c}, \bar{c}]$*, with*

$$\underline{c} \equiv \frac{1+\sigma}{2} V_r, \qquad \text{and} \qquad \bar{c} \equiv \frac{1-\sigma}{2} \left(V_{\ell r} - V_\ell\right).$$

Under condition

$$\pi\left((\ell, r); (\ell, r)\right) - \pi\left((\ell, 0); (\ell, r)\right) \geq c \iff \bar{c} \equiv \frac{1-\sigma}{2} \left(V_{\ell r} - V_\ell\right) \geq c$$

investment in proprietary data is always profitable if a firm has purchased a license.

Under condition

$$\pi\left((0, r); (0, 0)\right) - \pi\left((0, 0); (0, 0)\right) \leq c \iff \underline{c} \equiv \left(\frac{1-\sigma}{2} + \sigma\right) V_r \leq c$$

investment in proprietary data is profitable only if a firm has purchased the license. The interval $[\underline{c}, \bar{c}]$ is non-empty under $V_i = v(k_i)$. Note that Assumption 3 implies that when both firms purchase the license, there is the risk of inefficient duplication of investment costs c due to the presence of shoppers: as the shoppers have unit demand, when a firm invests in proprietary covariates it does not internalize the business-stealing effect of luring the mass σ of shoppers away from the competitor. Therefore, exclusivity might be socially desirable when c and σ are large. On the other hand, exclusivity harms welfare by preventing the investment in proprietary data by the excluded firm, thereby killing off its captive market of mass $\frac{1-\sigma}{2}$. Therefore, exclusivity might be socially harmful when σ is small.

One could think of a distinct reason for which exclusivity might be socially desirable: if c and σ are even larger, specifically $c \in \left[\frac{1-\sigma}{2} V_{\ell r}, \frac{1+\sigma}{2} V_{\ell r}\right]$, then investment in proprietary covariates will occur if and only if only one firm purchases the license. Without exclusivity, the market for k_S will break down and as $\frac{1-\sigma}{2} V_{\ell r} > \bar{c} > \underline{c}$, no prediction firm will investment in proprietary covariates. Therefore the prediction market will break down. Exclusivity will achieve proprietary data collection by one firm which will monopolize the prediction market also for shoppers. S will extract the whole social surplus, therefore in this case exclusivity will be profitable if and only if it maximizes social welfare. By contrast, we will show that under Assumption 3 exclusivity might be profitable for S but socially harmful.

Lemma 7. *Under Assumption 3, if* $e = N$*, there is a unique equilibrium in the data strategy subgame* $(\ell, r); (\ell, r)$ *in which both firms invest in proprietary data and purchase the license*

from S for a fee

$$f_N^* = \frac{1-\sigma}{2} V_{\ell r} - c.$$

Proof. Firm i will purchase the license for any f_N charged by S such that

$$\pi\left((\ell, r); (\ell, r)\right) - c - f_N \geq \pi\left((0,0); (\ell, r)\right) \iff \frac{1-\sigma}{2} V_{\ell r} - c - f_N \geq 0$$

This will lead to a maximum price of

$$f_N^* = \frac{1-\sigma}{2} V_{\ell r} - c.$$

□

If there is exclusivity and the firms which purchases the license invests and the other does not. Under non-exclusive licensing, both firms symmetrically acquire the data and invest. The fee f_N^* equals the incremental value of data net of the cost of proprietary investment. The data seller extracts all surplus from each firm's willingness to pay, which depends solely on the incremental improvement in prediction quality $V_{\ell r}$. This outcome represents the benchmark of parallel adoption with potentially redundant investment.

Lemma 8. *Under Assumption 3, if $e = E$, there is a unique equilibrium in the data strategy subgame $(\ell, r); (0,0)$ in which only the licensee invests in proprietary data and purchase the license from S for a fee*

$$f_E^* = \left(\frac{1-\sigma}{2} + \sigma\right) V_{\ell r} - c$$

Proof. By Assumption 3, there is a unique equilibrium $(\ell, r); (0,0)$. The firm acquiring the license will be willing to pay any f_E charged by S such that

$$\pi\left((\ell, r); (0,0)\right) - c - f_E \geq \pi\left((0,0); (\ell, r)\right) \iff \left(\frac{1-\sigma}{2} + \sigma\right) V_{\ell r} - c - f_E \geq 0$$

This will lead to a maximum price of

$$f_E^* = \left(\frac{1-\sigma}{2} + \sigma\right) V_{\ell r} - c.$$

□

Under exclusivity, only one firm licenses and invests, while its rival remains covariate-poor. The exclusive licensee's willingness to pay rises because it can capture the whole mass of shoppers. Exclusivity thus allows the data seller to internalize the value of market power created by data asymmetry. This can be good because it reduces the concerns for potential overinvestment but excludes the captive buyers of the firm which does not enter because it has failed to gain access to S 's covariates.

6.3.4 Exclusivity

Comparing the payoffs in Lemmas 7 and 8, we can characterize the condition under which S finds it optimal to offer exclusivity.

Lemma 9. *If $\sigma \geq 1/3$ or $\sigma < 1/3$ and $k \geq \underline{k}$, there exists $c_E^* < \bar{c}$ such that selling with exclusivity is profitable for S if and only if*

$$c > c_E^* \equiv \left(\frac{1-3\sigma}{2}\right) V_{\ell r},$$

with $\underline{k} \equiv \frac{n+1-k_S}{1+\frac{2\sigma(n+1)}{(1-3\sigma)k_S}}$.

Proof. It is profitable to offer exclusivity if and only if

$$f_E^* \geq 2f_N^* \iff \left(\frac{1-\sigma}{2}+\sigma\right) V_{\ell r} - c \geq (1-\sigma)\, V_{\ell r} - 2c \iff c_E^* \equiv \left(\frac{1-3\sigma}{2}\right) V_{\ell r} \leq c$$

Let us check that

$$c_E^* \leq \bar{c} \iff V_{\ell r} \geq \frac{1-\sigma}{2\sigma} V_{\ell},$$

which is always satisfied for $\sigma \in [1/3, 1]$ and otherwise is satisfied provided

$$k \geq \frac{n+1-k_S}{1+\frac{2\sigma(n+1)}{(1-3\sigma)k_S}}$$

which is decreasing in σ, n and inverted U-shaped in k_S. □

Analogously to Katz and Shapiro (1986), the seller trades off two effects: non-exclusive sales double the number of captive buyers but trigger competition for shoppers that erodes firms' willingness to pay; exclusivity restricts access but makes the license more valuable to the single buyer. When the fraction of shoppers σ is large or proprietary investment costs c are high, the competitive erosion effect dominates, making exclusivity more profitable in the region $[c_E^*, \bar{c}]$.

Lemma 10. *Exclusivity maximizes total welfare if and only if*

$$c_E^W \equiv \left(\frac{1-\sigma}{2}\right) V_{\ell r} \leq c.$$

Proof. The total welfare under non exclusivity is

$$W_{NE} = V_{\ell r} - 2c.$$

The total welfare under exclusivity is

$$W_E = \left(\frac{1-\sigma}{2} + \sigma \right) V_{\ell r} - c.$$

The condition for optimality of exclusivity is

$$W_E \geq W_{NE} \iff \left(\frac{1-\sigma}{2} + \sigma \right) V_{\ell r} - c \geq V_{\ell r} - 2c$$

□

From a social perspective, exclusivity can reduce wasteful duplication of proprietary data and may raise total welfare when investment incentives are excessive. Yet it may also harm consumers when it grants excessive market power. The comparison between c_E^* and c_E^W highlights that the data seller's private incentives are generally misaligned with welfare: exclusivity is chosen too often because the seller captures monopoly rents but does not internalize consumer surplus losses.

Proposition 13. *Whenever $c \in \left[c_E^*, c_E^W\right]$, exclusivity is profitable but reduces social welfare.*

Proof. Follows from observing that $c_E^* < c_E^W$. □

The length of the interval in which the cost of investment in proprietary covariates is profitable but socially harmful is $\sigma V_{\ell r}$: the higher the fraction of shoppers σ, the more abundant the data (i.e., the larger k, k_S and n) the more likely it is that exclusivity is profitable and anticompetitive.

The model shows that profitable exclusivity can be socially undesirable when data is abundant and competition for shoppers is fierce. This provides a micro-foundation for antitrust concerns surrounding data-sharing agreements: when datasets are strong complements to proprietary information, exclusivity amplifies incumbency advantages and can hinder market entry, echoing the rationale for regulatory scrutiny of real-world cases such as the Reddit–OpenAI deal.

7 Conclusion

This paper develops a general framework for understanding the value of data in prediction by explicitly modeling covariates. The analysis shows how complementarities between training and prediction, economies of scope across covariates, and interactions between covariates and observations 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 monopoly characteristics, as concentrating covariates within one firm

can raise efficiency. Privacy regulation that fragments data supply may inadvertently reinforce monopoly power, creating a trilemma between privacy, competition, and efficiency. The framework also highlights that not all data mergers are alike: list mergers, which combine the same covariates across users, are anticompetitive, while append mergers, which combine different covariates on the same users, can 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[13] 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.[14] 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.

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

[14] 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.

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.

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.

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:*
$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}^c} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \mathbf{0}_{|\mathbb{N}|-t};$$
2. *For trained parameters:*
$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} + (1 - S(\mathcal{T}))\cdot \boldsymbol{I}_t\right)^{-1} \boldsymbol{X}'_{\mathcal{T}}\boldsymbol{y}.$$

Proof. Because the prior is $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ 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 $p(\boldsymbol{\beta}_{\mathcal{T}^c} \mid \boldsymbol{M}_{\mathcal{T}}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{\mathcal{T}^c})$ 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*

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

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

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

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

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

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 $w_k(\lambda) = \mathrm{Var}_{\boldsymbol{M}_{\mathcal{T}}}[\mathbb{E}[\beta_k \mid \boldsymbol{M}_{\mathcal{T}}]] = 0$.

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 $\mathrm{Var}(x_k) = s_k$, the sample covariance matrix satisfies

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

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 $O_p\big(\sqrt{|\mathcal{T}|/N}\big)$ 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 $O\big(\sqrt{|\mathcal{T}|/N}\big)$.

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

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

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

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

since $b - a = -s_k$. Therefore

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

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

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

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.*

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. □

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}	
⋮	⋮	⋮	⋮	⋮	⋮	
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

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

Define principal components $z^i \equiv x^i U$. Then

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

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

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

where $(\cdot)^+$ denotes the Moore–Penrose pseudo-inverse.[15] 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 $X'_{\mathcal{T}} X_{\mathcal{T}}$ is invertible.
- Overparametrized regime ($n < |\mathcal{T}|$): the OLS estimator is not defined because the system $X_{\mathcal{T}} b_{\mathcal{T}} = y$ has infinitely many solutions; the MNLS chooses the solution with the smallest Euclidean norm.

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

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad (AA^+)' = AA^+, \quad (A^+A)' = A^+A.$$

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 $\mathbf{0}_{|\mathcal{T}|}$

Corollary 5. *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 $\mathbf{0}_{|\mathcal{T}|}$ 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 6. *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[16]

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

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

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

then (R.1) adds up across slices by orthogonality. The total mutual information revealed by a learning sample of *strength* t is therefore[17]

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

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:

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 pol-

[16] 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.

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

icy 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