Opening the Black Box*

A Statistical Theory of the Value of Data

Giovanni Rizzi†

October 20, 2025

Abstract

This paper develops a theory of the value of data in prediction. Using a Bayesian linear regression with endogenous covariate choice, I show that returns to the number of covariates may be increasing. Training data used to learn model parameters and covariates describing the user on whom the prediction is made are complements. Covariates and observations are complements when data is scarce but substitutes when it is abundant. Prediction may be a natural monopoly, making concentration more efficient than decentralization. Access regulation may restore competition, while privacy rules may inadvertently reinforce concentration, resulting in a privacy/efficiency/competition trilemma. Observation sharing among data brokers is anticompetitive but covariates sharing is efficient as it eliminates double marginalization. Data exclusivity agreements deter entry.

JEL Classification: C11, D83, L12, O33.

Keywords: Value of data, Prediction, Bayesian regression, Natural monopoly, Competition.

*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 Insiti, Bruno Jullien, Hiroaki Kaido, Simon Loertscher, Friedrich Lucke, Leonardo Madio, Giovanni Morzenti, Juan Ortner, Christoph Reidl, Andrew Rhodes, David Salant, 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.

Better predictions attract more users, and more users generate more data. This creates a self-reinforcing loop between users and data: better predictions attract more users, who generate more data, further improving predictions. 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.

Consequently, 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 in prediction*. The framework 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.

To study the value of data, I set up a data collection problem in which an agent must choose which covariates to observe in a set of infinitely many covariates with hetereogeneous variance to run a Bayesian linear regression. I first characterize the optimal predictor and show that it relies on ridge regression estimates. Building on this, I then derive closed-form expressions for the value of data, showing that returns depend on the distribution of the signal across covariates. I then characterize the optimal covariate selection. This generates three main insights on the economies of scope and scale of data.

Firstly, there may be *increasing returns to training covariates* depending on the distribution

[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`.

of variance across covariates. Indeed, observing a new covariate reduces prediction noise, and reductions in noise have an accelerating effect on the precision of estimates. When all covariates have identical variance, the result is stark: returns to covariates are always increasing. However, when covariate variance is very heterogeneous decreasing variance deriving from the optimal 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 variable 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 three 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]

Second, sharing covariates on the same individuals among data brokers may eliminate double marginalization, generating efficiencies. By contrast, sharing the same covariates on

[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".

different individuals may facilitate collusion.

Third, the owner of prediction data may seek exclusive access to training data, even when this reduces social surplus. The reason is that exclusivity — such as OpenAI's 2024 agreement with Reddit[6]— deprives entrants of a key complement to their proprietary data; this lowers entrants' incentives to invest in proprietary data collection and hinders entry.

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 returns to scale of 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.

Finally, I develop a simple framework to study scaling laws, shedding light on the phe-

[6] "Reddit and OpenAI Announce Partnership", OpenAI Blog, May 16, 2024, `https://openai.com/index/reddit-partnership/`; see also A. Paul, "Reddit Strikes AI Content Deal With OpenAI Ahead of IPO", *Reuters*, May 16, 2024.

nomenon 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 predicting a user's outcome from covariates (user attributes). In both stages, the firm faces constraints on the number of covariates it can employ. Its objective is to maximize prediction accuracy subject to these dimensionality constraints.

2.1 Data-Generating Process

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

$$y^i = \boldsymbol{\beta}' \boldsymbol{x}^i,$$

where $\boldsymbol{x}^i \equiv (x_j^i)_{j \in \mathcal{J}} \in \mathcal{X}$, where $x_j^i \in \mathbb{R}$ and we assume $\text{Var}[y] = 1$.

Covariates Covariates are mutually independent across $j \in \mathcal{J}$ and i.i.d. across users $i \in \mathcal{I}$:

$$x_j^i \sim \mathcal{N}(0, s_j),$$

where $s_j > 0$ is the variance of covariate j and is decreasing in j.

Parameters Parameters are unknown, independent of x_j^i, and mutually independent across $j \in \mathcal{J}$:[7]

$$\beta_j \sim \mathcal{N}(0, 1).$$

Cumulative Variance For any subset $\mathcal{J}' \subseteq \mathcal{J}$, define the ***cumulative variance*** is

$$S(\mathcal{J}') \equiv \sum_{j \in \mathcal{J}'} s_j,$$

where $S(\varnothing) = 0$. Because $\boldsymbol{\beta}$ is independent of $\boldsymbol{x}^i$ and the covariates are mutually independent, the assumption $\text{Var}[y] = 1$ implies $S(\mathcal{J}) = 1$.

[7] Normalizing $\text{Var}[\beta_j] = 1$ is WLOG, as any $\text{Var}[\beta_j] = \tau^2$ can be recovered by rescaling $\tilde{s}_j = \tau^2 s_j$

2.2 Statistical Model

Training For any vector $\boldsymbol{v} \in \mathbb{R}^J$ and subset $\mathcal{J}' \subseteq \mathcal{J}$, let $\boldsymbol{v}_{\mathcal{J}'} \equiv (v_j)_{j \in \mathcal{J}'}$. Before predicting, the firm may observe a set of ***training covariates*** $\mathcal{T} \subseteq \mathcal{J}$, for a sample of past users $\{1, ..., n\} \subseteq \mathcal{I}$, which constitutes a ***training matrix***

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

The firm will learn the data-generating process using a statistical model in which for each past user $i \in \mathcal{I}$, the outcome depends on a set $\mathcal{T}$ of covariates:

$$y^i = \boldsymbol{\beta}_{\mathcal{T}}' \boldsymbol{x}_{\mathcal{T}}^i + \varepsilon_{\mathcal{T}}^i, \qquad \varepsilon_{\mathcal{T}}^i \equiv \boldsymbol{\beta}_{\mathcal{J} \setminus \mathcal{T}}' \boldsymbol{x}_{\mathcal{J} \setminus \mathcal{T}}^i. \tag{1}$$

Because covariates are mutually independent and $\mathrm{Var}[y] = 1$, it follows that

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

Furthermore, we make an assumption on the distribution of the residual.

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

$$\varepsilon_{\mathcal{T}} \sim \mathcal{N}(0, 1 - S(\mathcal{T})). \tag{3}$$

This has two justifications. Firstly, we can show that the unconditional distribution of $\varepsilon_{\mathcal{T}}$ is uniformly close to Gaussian whenever the omitted variance is spread across many covariates (see Appendix A). Furthermore, the quasi-maximum-likelihood approach of Gourieroux, Monfort, and Trognon (1984), Bollerslev and Wooldridge (1992) and White (1982), shows that a misspecified Gaussian distribution is consistent if the first two moments are known.

The firm will therefore run an ordinary regression, considering that the set of covariates $\mathcal{T}$ affects its misspecfication error.

Prediction This training stage will generate some parameter estimates $\hat{\boldsymbol{\beta}}_{\mathcal{T}}$, which are estimates for the effects of the covariates on the outcome. Then the firm will collect a set $\mathcal{P}$ of ***prediction covariates*** on the representative user on which it wants to make the prediction. Their realization is the ***prediction vector***

$$\boldsymbol{x}_{\mathcal{P}} \in \mathbb{R}^{|\mathcal{P}|}.$$

Each choice $(\mathcal{T}, \mathcal{P})$ induces a random variable $(\boldsymbol{M}_{\mathcal{T}}, \boldsymbol{x}_{\mathcal{P}})$ whose realization is a dataset of type $(\mathcal{T}, \mathcal{P})$:

Definition 1. For any choice of covariates $(\mathcal{T}, \mathcal{P})$, a ***dataset of type*** $(\mathcal{T}, \mathcal{P})$ is the tuple

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

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.

2.3 Prediction Problem

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

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

Given a choice of covariates $(\mathcal{T}, \mathcal{P})$, the firm must choose how to map a generic dataset $\boldsymbol{D} \in \mathcal{D}_{\mathcal{T},\mathcal{P}}$ into predictions:

Definition 2 (Predictor and Posterior Risk)**.** A ***predictor of type*** $(\mathcal{T}, \mathcal{P})$ is a measurable map $f : \mathcal{D}_{\mathcal{T},\mathcal{P}} \to \mathbb{R}$ that produces a prediction

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

Its ***posterior risk*** is

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

The firm solves:

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

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

Denote an ***optimal predictor*** by

$$f \in \arg\min_{f} R(f, \boldsymbol{D}).$$

The ***posterior Bayes risk*** of $\boldsymbol{D}$ is 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:

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

where $\varnothing$ denotes the empty dataset.

Remark 1. The predictor specifies how Zillow translates a dataset of house covariates and past prices into a price prediction for the target house. It thus corresponds to Zillow's prediction algorithm, and the prediction problem captures the data science team's task of selecting and fine-tuning that algorithm.

2.4 Covariate Selection Problem

The first stage problem is to choose which covariate sets to obeserve. The expected value of a dataset of type $(\mathcal{T}, \mathcal{P})$ is the expected reduction in Bayes risk of a dataset $\boldsymbol{D}_{\mathcal{T},\mathcal{P}}$ relative to the prior:

Definition 3. For any $(\mathcal{T}, \mathcal{P}) \subseteq \mathcal{J} \times \mathcal{J}$, the ***expected value of a dataset of type*** $(\mathcal{T}, \mathcal{P})$ given sample size n is

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

The optimal covariate choice is that which gives the highest expected value subject to the constraint that the number of trainign covariates must not exceed the ***scope of training*** $t \in [0, J]$; and that of prediction covariates must not exceed the ***scope of prediction*** $p \in [0, J]$**.**

Problem 2 (Covariate Selection)**.** Given dimensions (p, t),

$$\sup_{(\mathcal{T},\mathcal{P}) \subseteq \mathcal{J}\times\mathcal{J}} \hat{V}(\mathcal{T}, \mathcal{P}, n), \qquad \text{s.t.} \qquad |\mathcal{T}| \leq t, |\mathcal{P}| \leq p.$$

I define the value of a dataset of dimensions (n, p, t) as the highest value of any dataset type whose dimensions do not exceed (p, t):

Definition 4. The ***value of a dataset of dimensions*** $(n, p, t) \in \mathbb{N}^3$ is

$$V(n, p, t) \equiv \sup_{\substack{|\mathcal{T}| \leq t \\ |\mathcal{P}| \leq p}} \hat{V}(\mathcal{T}, \mathcal{P}; n).$$

The difference in the covariate constraints (p, t) reflect differences in the difficulty covariate collection: the firm can collect training covariates anonymized data, but it must collect

prediction covariates for specific users and will therefore be more affected by technological constraints and privacy regulations.

3 The Value of Data

3.1 Prediction Problem

This section characterizes the optimal predictor.

3.1.1 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 parameters have been updated using $\boldsymbol{M}_{\mathcal{T}}$. Therefore, we can state a restriction on $\mathcal{P}$:

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

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

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

3.1.2 Bayes Estimator

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

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

1. *For untrained parameters:*

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{J}\setminus\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \mathbf{0}_{J-|\mathcal{T}|};$$

2. *For trained parameters:*

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

Because parameters are independent, training $\boldsymbol{\beta}_{\mathcal{T}}$ provides no information on the untrained parameters $\boldsymbol{\beta}_{\mathcal{J}\setminus\mathcal{T}}$, whose prior mean is $\mathbf{0}_{\mathcal{J}\setminus\mathcal{T}}$.

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:

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

where $\lambda > 0$ is a penalty for the squared Euclidean distance of $\hat{\boldsymbol{\beta}}_{\mathcal{T}}$ 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 Regression)**.** *The posterior mean coincides with a ridge regression estimator with regularization*

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

Equivalently,

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

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

3.2 Covariate Selection Problem

Henceforth, we assume $\mathcal{P} \subseteq \mathcal{T}$. The quadratic loss, together with the fact the optimal prediction is the posterior mean, implies that the value of a dataset is the reduction of variance it brings forth.

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

$$\hat{V}(\mathcal{T}, \mathcal{P}) = Var_{\boldsymbol{D}_{\mathcal{T},\mathcal{P}}}\left[f^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}})\right] = \sum_{j \in \mathcal{P}} s_j Var_{\boldsymbol{M}_{\mathcal{T}}}\left[\mathbb{E}\left[\beta_j | \boldsymbol{M}_{\mathcal{T}}\right]\right].$$

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

1. A *signal* term $s_j = \text{Var}(x_j)$, and
2. A *training* term $\text{Var}(\mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{T}}])$, measuring how sensitive the posterior mean of β_j is to $\boldsymbol{M}_{\mathcal{T}}$.

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

Lemma 3 (Variance of Bayes Estimator)**.** *The posterior mean satisfies*

$$Var_{\boldsymbol{M}_{\mathcal{T}}}\left(\mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{T}}]\right) = \begin{cases} 0, & j \in \mathcal{J} \setminus \mathcal{T}, \\ \dfrac{1}{1 + \dfrac{\lambda(n,\mathcal{T})}{s_j}} + O\left(\sqrt{\frac{|\mathcal{T}|}{n}} + \frac{|\mathcal{T}|}{n}\right), & j \in \mathcal{T}, \end{cases} \qquad \text{where } \lambda(n,\mathcal{T}) \equiv \frac{1 - S(\mathcal{T})}{n}.$$

The variance of the Bayes estimator is increasing in s_j, 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. The variance of the Bayes estimator will be large if the training sample size n is large or the trained covariates are more informative, so the residual $S(\mathcal{J} \setminus \mathcal{T})$ is small. Note that the reduction of posterior variance on the parameter of an observed covariate is

$$1 - \frac{1}{\frac{\lambda(n,\mathcal{T})}{s_j} + 1},$$

which, up to a reparametrization, coincides with the "data depreciation rate" of an AR(1) process in Section 2.1 of Farboodi and Veldkamp (2025): thus, uncertainty on a covariate's parameter and temporal obsolescence of knowledge in a dynamic AR(1) cause the same the loss of information. Furthermore, reductions in the penalty $\lambda(n,\mathcal{T})$ 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. This

Putting the result in Lemma 3 inside Lemma 2 gives the value of a dataset of a given type:

Theorem 1. *The value of a dataset of type* $(\mathcal{T}, \mathcal{P})$ *is*

$$\hat{V}(\mathcal{T}, \mathcal{P}) = \sum_{j \in \mathcal{P}} \frac{s_j}{\frac{\lambda(n,\mathcal{T})}{s_j} + 1} + O\left(\sqrt{\frac{|\mathcal{T}|}{n}} + \frac{|\mathcal{T}|}{n}\right).$$

The closed form in Theorem 1 allows us to study how the choice of dataset type $(\mathcal{T}, \mathcal{P})$ affects its value. In particular, it provides the foundation for solving the covariate selection problem below.

3.2.1 Optimal Covariate Selection

Proposition 2 (Optimal covariate selection)**.** *Consider the covariate selection problem*

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

Then the optimal sets of training and prediction covariates are the covariates with the largest variances:

$$\mathcal{T}^* = \arg\max_{\substack{\mathcal{T}\subseteq\mathcal{J} \\ |\mathcal{T}|=t}} \sum_{j\in\mathcal{T}} s_j, \qquad \mathcal{P}^* = \arg\max_{\substack{\mathcal{P}\subseteq\mathcal{J} \\ |\mathcal{P}|=p}} \sum_{j\in\mathcal{P}} s_j.$$

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

Notation. Relabel s_j so that $\mathcal{P}^* = \{1,\dots,p\}$, $\mathcal{T}^* = \{1,\dots,t\}$. Redefine

$$S(t) \equiv S(\{1,...,t\}),$$

which is increasing and concave in t as $s_{(j)}$ is decreasing in j. Furthermore,

$$\lambda(n,t) \equiv \lambda(n,\{1,...,t\}),$$

which is decreasing and convex in t as $s_{(j)}$ is decreasing in j.

Corollary 3. *The value of a dataset of dimensions* $(n,t,p) \in \mathbb{N}^3$ *is*

$$V(n,t,p) = V(p,\lambda(n,t)),$$

where

$$V(t,\lambda) \equiv \sum_{j=1}^{t} \frac{s_j}{\frac{\lambda}{s_j}+1}, \qquad \lambda(n,t) \equiv \frac{1-S(t)}{n}, \qquad S(t) \equiv \sum_{j=1}^{t} s_j.$$

Assumption 2. *We will henceforth assume*

$$\frac{t}{n} \to 0.$$

This assumption rules out overparameterized regimes.[8]The asymptotic term $O\left(\frac{t}{n}\right)$ vanishes provided that the parameter number t grows slower than the sample size n, i.e., as long

[8]It 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 $\{s_j\}$. By contrast, in the high-dimensional regime $t/n \to \gamma \in (0,\infty)$, explicit

as $\frac{t}{n} \to 0$. 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.

4 The Marginal Returns to Data

I uncover three properties of the value of data:

1. ***training-prediction complementarity***: prediction covariates are more valuable when (a) there are more training covariates, or (b) there are more training observations (Proposition 5);

2. ***S-shaped Returns to Training***: Adding early covariates sharply improves prediction by reducing misspecification, but returns flatten once many are included—unless signal is extremely diffuse (Proposition 7);

3. ***Covariate-observations Interaction***: covariates and observations are complements when training data are scarce, but substitutes when data are abundant (Proposition 9).

These results highlight a general principle: the value of information is *structural,* meaning it is defined only in relation to other information.

The following lemma establishes a continuous approximation of Corollary 3 when covariates are numerous.

Proposition 3. *Suppose the variances* $\{s_j\}_{j=1}^J$ *arise from a density* $s_j = s\left(\frac{j}{J}\right)\frac{1}{J}$*, and define the asymptotic ratios*

$$\ell \equiv \lim_{J\to\infty} \frac{\ell(J)}{J} \in [0,1), \qquad p \equiv \lim_{J\to\infty} \frac{p(J)}{J} \in [0,J]$$

Let $S(t) = \int_0^t s(u)du$*. Then, treating* n *as a real number, I can write*

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

The following corollary characterizes the variance of the ridge estimator of the parameter u.

Corollary 4 (House Party Effect)**.** *Note that*

$$v^{ridge}(u;\lambda) \equiv Var\left[\hat{\beta}_u^{ridge}(\lambda)\right] = \frac{s(u)}{\lambda + s(u)}$$

is decreasing in u*, decreasing and convex in* λ*.*

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.

Note that because of the convexity in λ, marginal reductions in the penalty will have positive and increasing effect on the variance of the estimator because the marginal effect of lowering the penalty is stronger when the penalty is already small.

4.1 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 4 (Diseconomies of Scope in Prediction)**.**

$$V_t(p, \lambda) = s(p)v^{ridge}(p; \lambda) > 0$$
$$V_{pp}(p, \lambda) = \dot{s}(p)v^{ridge}(p; \lambda) + s(p)v_p^{ridge}(p; \lambda) \leq 0$$

The firm always gains from observing more covariates. By 2, the firm ranks covariates in decreasing order of informativeness, so $\dot{s}(u) \leq 0$ and the marginal value of prediction covariates decreases with p. Covariates act as inputs of heterogeneous quality, analogously to the Law of Diminishing Returns in Ricardo (1817): the firm uses higher-quality inputs (the most informative covariates) first, lower-quality ones later.

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

Proposition 5 (Training-Prediction Complementarities)**.**

$$V_{t\ell}(p, \lambda(n, t)) = s(t)v_\lambda^{ridge}(p; \lambda)\lambda_\ell(n, t) > 0$$
$$V_{tn}(p, \lambda(n, t)) = s(t)v_\lambda^{ridge}(p; \lambda)\lambda_n(n, t) > 0$$

Increasing the number of training covariates t or the number of observations n reduces the penalty λ. Both effects increase the variance of the ridge estimator for all inframarginal covariates $u \leq t$ included in the prediction set. Hence, if the firm observes a new covariate t, its estimator is more sensitive to the data and has a higher marginal value. While these complementarities are often intuited by data scientists, to my knowledge they have not been formally proven in the literature. The complementarity between t and n is consistent with the empirical findings of Schaefer and Sapi (2023) and arises from the non-rival nature of the training data: better learning, i.e., lower $\lambda(n, t)$, has a positive effect on each prediction parameter, and this effect does not depend on how many parameters are affected. This mirrors Wilson (1975): better information can be leveraged across the entire scale of production, so the non-rival nature of information generates complementarities and increasing returns.

4.2 Diseconomies of Scale to Training

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

Proposition 6 (Diseconomies of Scale to Training)**.**

$$V_n(p,\lambda) = \int_0^p s(u) v_\lambda^{ridge}(u;\lambda)\lambda_n(n,t)du > 0,$$

$$V_{nn}(p,\lambda(n,t)) = \int_0^p s(u)\left(\underbrace{v_\lambda^{ridge}(u;\lambda)\lambda_{nn}(n,t)}_{\text{Law of Large Numbers}<0} + \underbrace{v_{\lambda\lambda}^{ridge}(u;\lambda)\lambda_n^2(n,t)}_{\text{House Party Effect}>0}\right) < 0.$$

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. Two opposing forces:

1. ***Law of Large Numbers***: since the penalty $\lambda(n,t)$ decreases with n but is bounded below by zero, it is convex. Each additional observation thus eliminates less residual uncertainty, yielding diminishing returns. This pattern is consistent with Bajari et al. (2019), Schaefer and Sapi (2023), and the Law of Large Numbers, which ensures that parameter estimates converge as n grows.

2. ***House Party Effect***: the variance of the ridge estimator is convex in λ. This means that the marginal effect of reducing λ is increasing. When λ is large (little data), a small reduction has almost no effect on variance—like asking one guest to quiet down at a noisy house party. When λ is small (ample data), the same reduction sharply lowers variance—like asking one of the last talkers in an almost-silent room to stop, which dramatically improves the ability to hear.

The House Party Effect dampens but never overturns diminishing returns. Intuitively, for the HPE to be strong, the penalty λ must be small (the "room is already quiet"), which occurs only when n is large. But at that point $\lambda(n,t)$ is nearly flat, so $\frac{\partial\lambda(n,t)}{\partial n}$ is close to zero. Thus, just when variance convexity would amplify the effect of lowering λ, the penalty itself barely moves. The HPE has a negligible effect on returns to observation.

In Proposition 7 I show that it can substantially impact returns to training covariates.

Proposition 7 ((Dis-)economies of Scope in training)**.** *The marginal value of expanding the set of training covariates is given by*

$$V_t(t,\lambda(n,t)) = \int_0^p s(u) v_\lambda^{ridge}(u;\lambda)\lambda_t(n,t)du \geq 0,$$

with curvature

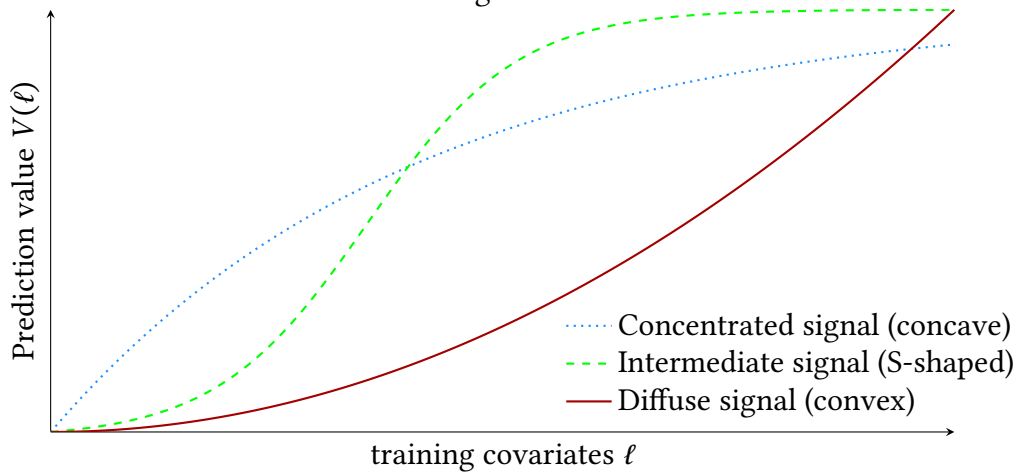
$$V_{tt}(p, \lambda(n,t)) = \int_0^p s(u) \left(\underbrace{v_{\lambda}^{ridge}(u;\lambda)\lambda_{tt}(n,t)}_{\text{Law of Large Numbers}<0} + \underbrace{v_{\lambda\lambda}^{ridge}(u;\lambda)\lambda_t^2(n,t)}_{\text{House Party Effect}>0} \right) du.$$

Proposition 8. *Returns to training covariates are increasing if and only if* $n \leq \hat{n}_{V_{tt}}(t)$*,where* $\hat{n}_{V_{tt}}(t)$ *is implicitly defined by*

$$V_{tt}(p, \lambda(\hat{n}_{V_{tt}}, t)) = 0.$$

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

Increasing the set of training covariates always raises the value of data, but the returns can be increasing or decreasing. When the number of observations is small, the penalty term λ is large, so each additional training covariate substantially relaxes the regularization. Because the House Party Effect scales quadratically with λ_t, this can outweigh the diminishing force of the Law of Large Numbers and generate *increasing returns.* As the sample grows, however, the penalty flattens and the House Party Effect loses strength, so the Law of Large Numbers dominates and returns become concave. 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 5. *If* $s(u) = 1$ *for all* $u \in [0,1]$*, then*

$$V_{tt}(p, \lambda(n,t)) > 0.$$

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.

I now show that the cross derivative of training covariates and observations can be both positive or negative depending on the size of n and t.

Proposition 9 (Complementarity/Substitutability in training). *It*

$$V_t(p, \lambda(n,t)) = \int_0^p s(u)\Big(\underbrace{v_\lambda^{ridge}(u;\lambda)\lambda_t(n,t)}_{\text{Law of Large Numbers}<0} + \underbrace{v_{\lambda\lambda}^{ridge}(u;\lambda)\lambda_t(n,t)\lambda_n(n,t)}_{\text{House Party Effect}>0} \Big)\, du.$$

In particular,

$$\lim_{n\to 0^+} V_t(p, \lambda(n,t)) > 0$$

while

$$\lim_{n\to\infty} V_t(p, \lambda(n,t)) = 0^-.$$

Hence, the returns to training covariates are increasing if and only if

$$n \le \hat{n}_{V_t}(t),$$

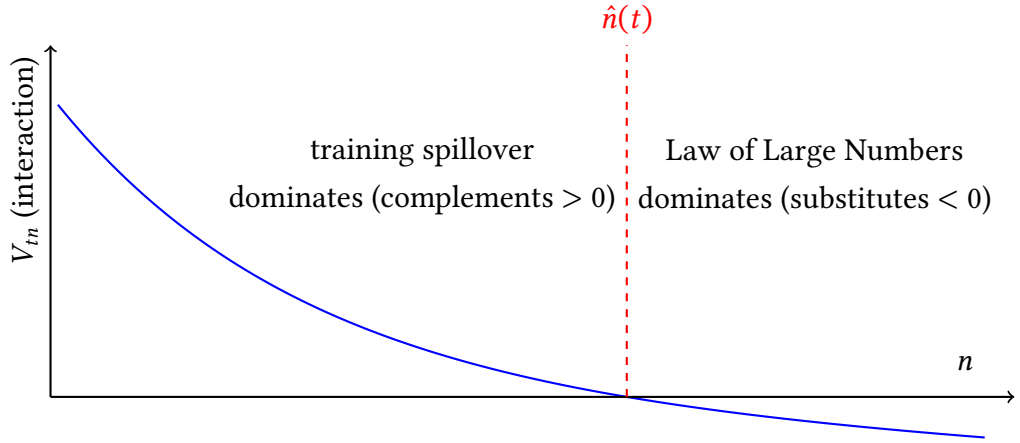
where $\hat{n}(t)$ is implicitly defined by

$$V_{tt}(p, \lambda(\hat{n}_{V_t}(t), t)) = 0.$$

Moreover, $\hat{n}_{V_t}(t)$ is decreasing in t.

The interaction between training covariates and sample size reflects two opposing forces. On the one hand, both additional covariates and additional observations reduce the penalty λ, which makes them *substitutes* (Law of Large Numbers effect). On the other hand, when λ is small, even a slight further reduction has a large payoff because of variance convexity, so the two become *complements* (House Party Effect).

Covariates and observations are complements only in data-scarce environments, when reducing the penalty is still highly valuable. Once the dataset is already rich, expanding either margin yields little extra benefit because the other has already lowered the penalty close to zero. Economically, this result shows why firms with small datasets benefit from both richer features and larger samples, whereas firms with large datasets can choose whether to collect more data or more features.



4.3 Modularity

The previous results examined marginal returns along single data dimensions—prediction covariates, training covariates, or observations—and how these interact. I now ask: what happens when whole datasets are combined?

Formally, modularity is the gap between the joint value of two datasets and the sum of their standalone values. A positive gap implies complementarity, a negative one substitutability. Its sign depends on the curvature of the value function. The Law of Large Numbers pushes toward substitutability, while the convexity of ridge variance (the House Party Effect) pushes toward complementarity. Which force dominates hinges on whether data are scarce or abundant.

The following corollaries formalize this result, establishing conditions under which datasets are complements or substitutes when combined.

Corollary 1 (Complementarity across datasets with fixed n)**.** *Fix n and p. For $t_1, t_2 \geq 0$ with $t_1 + t_2 \leq J$, define the supermodularity gap*

$$\Delta(t_1, t_2) \equiv V(n, t_1 + t_2, p) - V(n, t_1, p) - V(n, t_2, p) > 0 \iff n \leq \hat{n}_{V_{tt}}(t).$$

so datasets are complements if and only if observations are scarce. In particular, if $s(u) = 1$ for all u, datasets are always complements.

This corollary shows that whether two datasets are *complements or substitutes* depends on the curvature of the value function in t. For small n, the House Party Effect dominates: adding one dataset makes the other more valuable, so datasets are **complements**. For large n, the Law of Large Numbers dominates: the datasets overlap in value, so they are **substitutes**. The threshold $\hat{n}_{V_{tt}}(t)$ marks the switch between these regimes. In the special case of equally informative covariates ($s(u) \equiv 1$), diminishing returns vanish and datasets are **always complementary**. The result explains why *merging small datasets yields complementarities*, while *large ones add little beyond duplication.*

Corollary 2 (Substitutability across datasets with fixed t)**.** *Fix t and p. For $n_1, n_2 > 0$, define the supermodularity gap*

$$\Delta_n(n_1, n_2) \equiv V(n_1 + n_2, t, p) - V(n_1, t, p) - V(n_2, t, p) < 0,$$

so the datasets are substitutes.

This corollary establishes that datasets that differ only in their number of observations are always **substitutes**. The supermodularity gap $\Delta_n(n_1, n_2)$ is strictly negative, since $V_n > 0$ but $V_{nn} < 0$. Intuitively, once one dataset provides additional observations, the other yields less

incremental value, as both reduce the penalty λ through the same channel. Thus, unlike covariates, *observations never generate complementarities across datasets*: merging samples adds value, but less than the sum of their parts.

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 hetereogenity 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)?

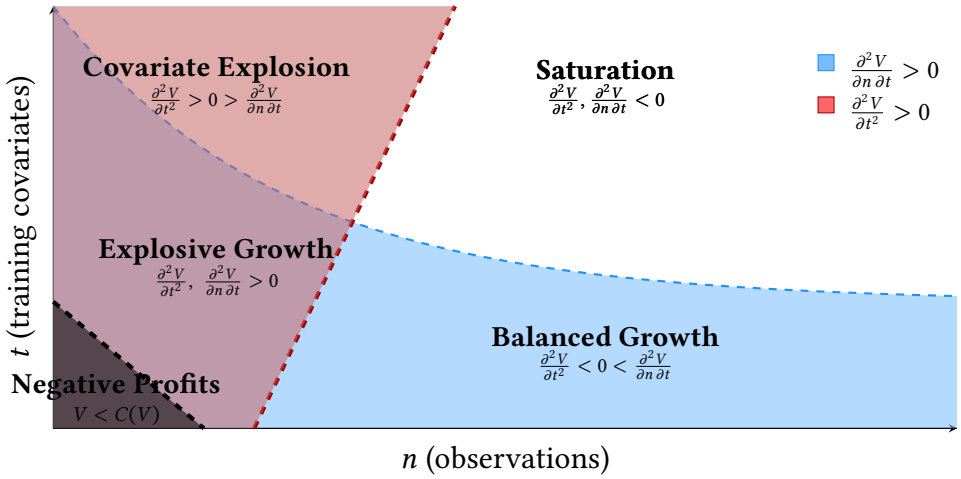
Profitability of Data Collection Corollary **??** implies prediction technologies typically require a minimum scale of data before becoming profitable, the "cold start" problem discussed in Iansiti 2021. This creates significant sunk costs: firms entering data-intensive markets must commit resources 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).

Covariate Selection Proposition 2 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 1 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

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

to a path of

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

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 AdTech and Advertisers: Natural Monopoly

6.1.1 Setup

Environment. Consider advertisers choosing which ad to show. Ads are indexed on the real line $\mathbb{R}$, and there is an unknown optimal ad $y \in \mathbb{R}$. Advertisers share a common prior on y, described in Section 2.1. Concretely, ads can be ordered along a horizontal dimension (e.g. shorter vs. longer YouTube ads), and the optimal ad depends on a vector of covariates $\boldsymbol{x}$, which capture characteristics of the ad–viewer interaction (e.g. age, product type, past

purchases). Advertisers may purchase a prediction of the best ad from an AdTech platform A. The platform can use past data on which ads were optimal to learn the mapping between y and $\boldsymbol{x}$. For simplicity, assume all covariates have the same variance.

Advertisers. An advertiser is characterized by its type $\theta \in \Theta \subseteq \mathbb{R}_+$, distributed with cumulative distribution function $G(\cdot)$ and density $g(\cdot) = G'(\cdot)$. The type θ reflects sensitivity to targeting accuracy: higher θ means larger losses when ads are mismatched. A natural interpretation is the advertiser's margin. Mistargeting on high-margin products (e.g. insurance, luxury goods, credit cards) is more costly than on low-margin ones (e.g. fast food or mass-market goods).

If an advertiser pays fee F and shows ad $\hat{y}$ while the true optimal ad is y, its Bernoulli utility is

$$\tilde{u}(\hat{y}, y) = -\theta(\hat{y} - y)^2 - F.$$

The outside option is to pick the prior mean. By Definition 4, the net expected utility of receiving a recommendation of informational value V is

$$u(V, F) = \theta V - F.$$

Hence the advertiser purchases if $u(V, F) \geq 0$. The resulting demand curve is $Q(F, V)$, which is decreasing in F and shifts upward in V.

AdTech. The AdTech platform observes $p \in [0, 1]$ prediction covariates in real time, reflecting limits on data collection (e.g. privacy regulation). It also has n past interactions available for training. The platform chooses the number of training covariates $t \in [0, 1]$, knowing that the value of its predictions is

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

Training incurs an endogenous fixed cost

$$\frac{c}{1 - t^2},$$

which is convex, reflecting that the platform collects the cheapest covariates first (all variances normalized).[9]

In practice, adtech platforms set non-negotiable prices (CPM, CPC, subscriptions), reflecting their bargaining power. We model this as a take-it-or-leave-it fee F for access to recommendations $\hat{y}$. If the platform sells to $Q(F, V(t))$ advertisers, its profit is

$$\max_{t,F} \ \Pi(t, F) = FQ(F, V(t)) - \frac{c}{1 - t^2}.$$

[9]This specification ensures interior solutions with the least convex functional form.

I proceed in two steps. First, I characterize the cost/value technology. Then I analyze demand and optimal fee choice. Finally, I combine these to study the training decision.

6.1.2 Supply

Technology and costs. With training intensity $t \in [p, 1]$, the value produced is

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

The conditional factor demand for training covariates is

$$\tilde{t}(x) \equiv V^{-1}(x) = 1 - n\left(\frac{p}{x} - 1\right).$$

We can write the cost function for V as

$$C(V) = \frac{c}{1 - \tilde{t}^2(V)}.$$

Hence, average and marginal costs are

$$\mathrm{AC}(V) \equiv \frac{C(V)}{V} = \frac{c}{V\left(1 - t^2(V)\right)}, \qquad \mathrm{MC}(V) \equiv C'(V) = \frac{2cnp\, t(V)}{V^2\left(1 - t^2(V)\right)^2}.$$

Baumol (1977) provide the classic test for natural monopoly, namely the condition under which the production of a good is more efficient when it is done by a single firm compared to when it is split among two firms. Specifically, they show that this is the case whenever the average cost is decreasing, which is equivalent to the condition under which the marginal cost lies below the average cost. In our context however the value is not per se a divisible product, rather it behaves more like a quality

Lemma 4 (Baumol test)**.**

$$\mathrm{AC}(V) \geq \mathrm{MC}(V) \quad \textit{if and only if} \quad V \leq V_{\mathrm{NM}} \equiv p\sqrt{\frac{1}{2/n + 1}}.$$

Next we turn to demand, which will allow us to characterize the feasible equilibrium.

6.1.3 Demand

The advertiser purchases the recommendation if and only if

$$\theta \;\geq\; \frac{F}{V}.$$

The resulting demand function is

$$Q(F, V) \;=\; 1 - G\left(\tfrac{F}{V}\right).$$

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

To ensure regularity of demand and rule out excessively heavy tails, assume:

Assumption 3 (Thin Tails)**.** *The relative hazard rate* $\frac{\theta g(\theta)}{1-G(\theta)}$ *is strictly increasing on the support of* θ*.*

This assumption holds, for instance, under Exponential and Uniform distributions, but not under Pareto.

Proposition 10. *Since there are no costs per advertiser, the platform chooses* F *to maximize revenue:*

$$\max_F \; FQ(F).$$

Under Assumption 3, the unique optimal fee is

$$F^*(V) \;=\; V\hat{\theta},$$

where $\hat{\theta}$ *is implicitly defined by*

$$\frac{\hat{\theta} g(\hat{\theta})}{1 - G(\hat{\theta})} = 1,$$

and the effective markup is

$$m \;\equiv\; \hat{\theta}\,[1 - G(\hat{\theta})].$$

Proof. Assumption 3 ensures that the second-order condition holds. The first-order condition is $Q(F) + FQ'(F) = 0$. Substituting $\hat{\theta} \equiv F/V$ reduces this to

$$\frac{\hat{\theta} g(\hat{\theta})}{1 - G(\hat{\theta})} = 1,$$

which uniquely defines $\hat{\theta}$ under Assumption 3. □

The optimal fee scales linearly with V, so richer datasets proportionally raise advertisers' willingness to pay and the platform's revenue extraction. The cutoff $\hat{\theta}$ is the effective marginal type, determined by the standard monopoly pricing condition: it balances the trade-off between higher fees and reduced participation, i.e. it is the type at which marginal revenue equals marginal cost of demand reduction. The effective markup $m = \hat{\theta}[1 - G(\hat{\theta})]$ is the analogue of the Lerner index.

6.1.4 Optimal Training

Define the minimum and maximum feasible values of V as

$$V_{\min} \equiv V(p) = \frac{p}{1 + \frac{1-p}{n}}, \qquad V_{\max} \equiv V(1) = p.$$

To guarantee that the optimum t^* lies strictly above p, we impose the following condition:

Assumption 4. *The training cost satisfies*

$$c \;\le\; \bar{c}(n, p, m) \;\equiv\; \frac{mn}{2}\left(\frac{1-p^2}{n+1-p}\right)^2.$$

Proposition 11. *Under Assumption 4, the AdTech platform solves*

$$\max_{V \in [V_{\min}, V_{\max}]} \Pi(V) \;=\; mV - C(V),$$

with unique solution V^* defined by the first-order condition

$$m \;=\; \mathrm{MC}(V^*; n, p, c),$$

where

$$\mathrm{MC}(V) \equiv C'(V) = \frac{2cnp\, t(V)}{V^2\big(1 - t^2(V)\big)^2}.$$

The optimal value $V^*(n, p, c, m)$ is increasing in n, p, and m, and decreasing in c.

Moreover, prediction is a natural monopoly if and only if

$$c > c_{\mathrm{NM}}(m, p, n) \;\equiv\; \frac{mp\left(1 - t_{\mathrm{NM}}^2(n)\right)^2}{2(n+2)t_{\mathrm{NM}}(n)},$$

where

$$t_{\mathrm{NM}}(n) \;\equiv\; \tilde{t}(V_{\mathrm{NM}}(p, n)) = n + 1 - \sqrt{n(n+2)}, \qquad V_{\mathrm{NM}}(p, n) \equiv p\sqrt{\tfrac{1}{2/n+1}}.$$

The threshold $c_{\mathrm{NM}}(m, p, n)$ is increasing in m, p, and n.

Proof. Existence and uniqueness. The profit function is strictly concave under Assumption 4, since the second-order condition holds at any interior optimum. Indeed,

$$C''(V) < 0 \quad\Longleftrightarrow\quad V < \tfrac{p}{2/n+1} \;<\; V_{\min},$$

which is a contradiction. Hence the SOC always holds and the unique solution is given by the FOC $m = \mathrm{MC}(V^*)$.

Natural monopoly. Because $\mathrm{MC}(V)$ is increasing, $V^* \le V_{\mathrm{NM}}$ if and only if $m \le \mathrm{MC}(V_{\mathrm{NM}})$.

At V_{NM}, we have

$$\mathrm{MC}(V_{\mathrm{NM}}) = \frac{2c(n+2)t_{\mathrm{NM}}(n)}{p\left(1 - t_{\mathrm{NM}}^2(n)\right)^2}.$$

Rearranging gives the threshold condition

$$c > c_{\mathrm{NM}}(m, p, n) \equiv \frac{mp\left(1 - t_{\mathrm{NM}}^2(n)\right)^2}{2(n+2)t_{\mathrm{NM}}(n)}.$$

□

Higher n (sample size), p (available prediction covariates), or m (advertiser markup) all increase the marginal return to training, so the platform optimally invests more and achieves a larger V^*. By contrast, higher training costs c discourage investment and lower V^*.

The natural monopoly condition reflects the Baumol test: if costs c are high relative to dataset value, average costs lie above marginal costs at the relevant scale, making it inefficient for multiple firms to coexist. Larger n, p, or m all raise the threshold c_{NM}, meaning that natural monopoly is more likely when data are scarce, valuable, or costly to collect.

6.1.5 Privacy/Efficiency/Competition Trilemma

Privacy regulation effectively raises costs c and/or reduces n and p. Policymakers therefore face a trilemma: they can achieve at most two of the following three objectives:

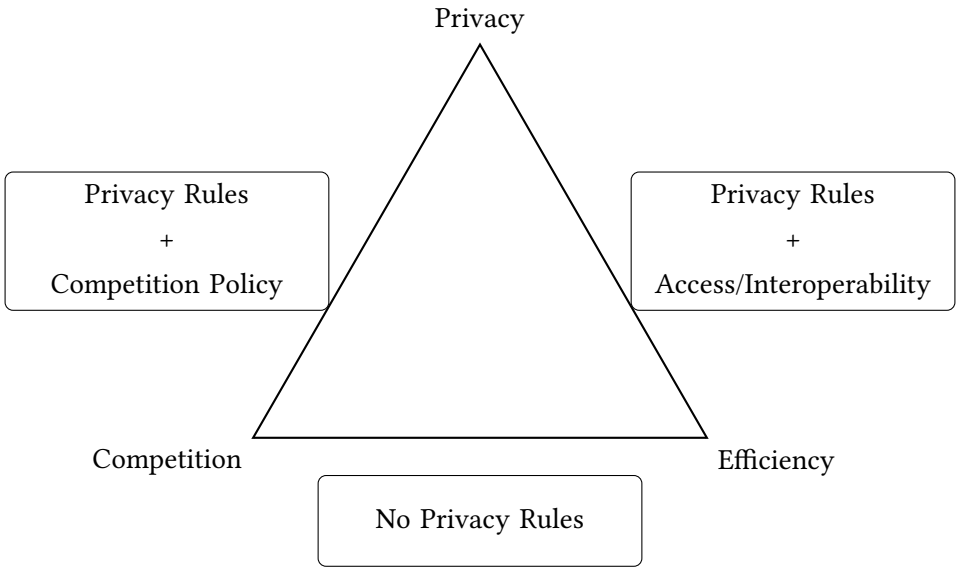
- **Privacy:** limiting data collection and sharing;
- **Competition:** preventing concentration of data in the hands of one firm;
- **Efficiency:** exploiting increasing returns in learning by concentrating data.

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*).

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 Broker Mergers

Data brokers often sign partnerships to share their data. For instance, Acxiom has partnerships with other data brokers, including Corecom (specialised in entertainment data) and Nielsen (a global data company). Gu, Madio, and Reggiani (2021) analyze the potential anticompetitive implications of these agreements, showing that they can be anticompetitive if the data products which are shared are substitutes but the agreements is not harmful if they are complements. We build on their analysis in two ways (i) we can give a clear statistical interpretation to whether datasets are compleemnts or substitutes; (ii) whereas their model uses a unit demand, we allow for a continuous demand, which allows us to model double marginal,ization arising from compelenmary products.

I assume that brokers partnerships are of two types

- User partnerships in which brokers share data lists: brokers ahare the same covariates on distinct users;
- Covariate partnerships in which brokers share data appends: brokers share distinct covariates on the same users.

I show data list broker mergers are anticompetitive, whereas data append broker mergers can be procompetitive if the available data is scarce thanks to the elimination of double marginalization. Hence, a tougher privacy policy, which restricts the amount of data available, should be accompanied by a laxer merger control policy for append brokers with scarce data, to avoid hindering efficient data consolidation, in particular if data is used in complex prediction problems, which have more diffused signal density $s(\cdot)$ and more increasing returns.

6.2.1 Data List Brokers

Proposition 1 (Data List Merger)**.** *Suppose two brokers sell lists that only increase n with ℓ fixed. Then $V_{nn} < 0$ implies the datasets are strict substitutes: for all $n_1, n_2 > 0$, $V(n_1+n_2, t, p) - V(n_1, t, p) - V(n_2, t, p) < 0$. Therefore competition between list brokers lowers prices relative to integrated monopoly. A merger of list brokers raises price and reduces adoption.*

6.2.2 Data Append Brokers

Proposition 2 (Data Append Merger)**.** *Under Assumption 3, a merger between two data append brokers can be either anticompetitive or procompetitive:*

- *If $n \leq \hat{n}_{V_{t\ell}}(t)$, the data appends are complements. Hence, the merger eliminates double marginalization so it is procompetitive.*
- *If $n > \hat{n}_{V_{t\ell}}(t)$, the datasets are substitutes. Hence, the merger weakly increases price and strictly lowers adoption, so it is anticompetitive.*

Given $\hat{n}_{V_{tt}}(t)$ is decreasing in t, if data are scarce, it is more likely the merger is procompetitive. In the special case of equal covariates, the merger is always procompetitive.

If $n \leq \hat{n}_{V_{tt}}(t)$ (so $V_{tt} > 0$), the House Party Effect dominates and the incremental value of extra attributes is increasing. Then $V_{12} > V_1 + V_2$: users value the second append more if they purchase the first append. Without the merger, both brokers would apply a margin to extract profits, resulting in the classic double marginalization problem. A merger between the brokers reestablishes efficiency by eliminating the double margin.

If $n > \hat{n}_{V_{tt}}(\ell)$ (so $V_{tt} \leq 0$), the LLN dominates and the incremental value of extra attributes is decreasing. Then $V_{12} < V_1 + V_2$: users view the second append as a weaker add-on. Competition between the two apps disciplines price; a merger eliminates this rivalry and moves the price toward monopoly bundling. A merger of append brokers on the same users is likely anticompetitive: it raises prices and lowers adoption.

6.3 Covariate Exclusivity Deals

In May 2024, Reddit signed an exclusivity agreement with OpenAI, granting privileged access to Reddit content as training data for large language models (LLMs).[10] To model this, I assume there are two sets of covariates:

- **Reddit covariate** R**:** the linguistic and conversational features embedded in user-generated discussions;
- **Proprietary covariates** P**:** other sources of text data that entrants can access.

[10] "Reddit and OpenAI Announce Partnership," OpenAI Blog, May 16, 2024; A. Paul, "Reddit Strikes AI Content Deal With OpenAI Ahead of IPO," Reuters, May 16, 2024.

In this section, I show that this deal amounts to monopolizing an informative covariate, which raises the marginal value of P for the incumbent (OpenAI), while lowering a potential entrant's incentive to collect them.

6.3.1 Environment

There are three actors: an upstream data provider (Reddit) and two downstream firms, an incumbent I and a potential entrant E.

Data and value of data. There are two informative covariates,

$$R \quad \text{("Reddit")} \qquad \text{and} \qquad P \quad \text{("proprietary")},$$

with signal strengths $s_R, s_P > 0$. In the Reddit-OpenAI case, the license grants corpus/API access that improves learning, but does not provide target-level Reddit covariates at inference. Therefore, I assume that R can only be used for learning, whereas P can be used for both learning and targeting. For any learning set $\mathcal{L} \subseteq \{\varnothing, R, P\}$,

$$V(\mathcal{L}) \;=\; \frac{s_P}{1 + \frac{\lambda(\mathcal{L})}{s_P}}, \qquad \lambda(\mathcal{L}) \equiv \frac{1 - S(\mathcal{L})}{n},$$

where $S(\mathcal{L}) = \sum_{j \in \mathcal{L}} s_j$.

Downstream profits. If firm $f \in \{I, E\}$ achieves prediction quality $q_f = V(\mathcal{L}_f)$, downstream operating profits are

$$\pi_f^D(q_f, q_{-f}) \;=\; \alpha q_f + \gamma(q_f - q_{-f}), \qquad \alpha \geq 0,\ \gamma > 0,$$

where α reflects absolute-quality rents, while γ reflects relative-quality rents deriving form business stealing.

Costs and profits. Collecting proprietary data P costs $c_P \geq 0$. To ensure it is always profitable to collect some proprietary data, I make the following assumption:

Assumption 5. *If R is available open source, it is profitable for firms to collect P*

$$c_P < \alpha V(\{R, P\}).$$

Accessing Reddit covariates requires a license fee T_f. Therefore, Firm f's total profit is

$$\Pi_f = \pi_f^D(q_f, q_{-f}) - c_P\, i_P^f - T_f,$$

where $i_P^f \in \{0, 1\}$ indicates whether P is collected.

Timing

1. Reddit offers either an exclusive contract X (selling R only to I) or a nonexclusive contract N (selling R to both).
2. Entrant E decides whether to enter.
3. Firms decide whether to collect P (i_P^f).
4. Downstream profits π^D are realized.

Bargaining Reddit is the proposer and makes take-it-or-leave-it offers.

- Under X, Reddit sets T_I^X and sells R only to I.
- Under N, Reddit sets (T_I^N, T_E^N) and sells to both.

In either case, fees can extract up to each user's incremental profit from obtaining R, relative to the outside option where neither firm observes R.

6.3.2 Entry and Investment Incentives

The marginal values of P are

$$\Delta(P \mid R) \equiv V(\{R,P\}) - V(\{R\}) = \frac{s_P}{1+\lambda^R/s_P}, \qquad \lambda^R \equiv \tfrac{1-S(\{R,P\})}{n},$$

$$\Delta(P \mid \varnothing) \equiv V(\{P\}) - V(\varnothing) = \frac{s_P}{1+\lambda^\varnothing/s_P}, \qquad \lambda^\varnothing \equiv \tfrac{1-S(\{P\})}{n}.$$

Since $s_R > 0$, I have $\lambda^\varnothing > \lambda^R$ and thus $\Delta(P \mid R) > \Delta(P \mid \varnothing)$: Reddit data raises the marginal value of proprietary data.

Best responses:

$$i_P^E = \begin{cases} 1 & \text{if } \Delta(P \mid \varnothing) \geq c_P, \\ 0 & \text{otherwise,} \end{cases} \qquad i_P^I = \begin{cases} 1 & \text{if } \Delta(P \mid R) \geq c_P, \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 3 (Entry deterrence)**.** *Suppose $s_R > 0$. If*

$$c_P \in \big(\Delta(P \mid \varnothing),\ \Delta(P \mid R)\big],$$

then the incumbent invests in P while the entrant does not, so E does not enter. Therefore, exclusivity over R allows I to deter E from entering by lowering its incentive to collect complementary data.

6.3.3 Profitability of Exclusivity

Assumption 6. *The cost*

$$c_P \in \big(\Delta(P \mid \varnothing),\ \Delta(P \mid R)\big],$$

so that I can deter E from entering if and only if I has exclusive access to R.

Qualities.

$$\begin{aligned} \textbf{Exclusive } (X): &\quad q_I^X = V(\{R,P\}), \quad q_E^X = 0, \\ \textbf{Nonexclusive } (N): &\quad q_I^N = q_E^N = V(\{R,P\}). \end{aligned}$$

Incremental downstream profits.

$$\begin{aligned} \Delta_I^X &= (\alpha+\gamma)V(\{R,P\}) - c_P, \\ \Delta_f^N &= \alpha V(\{R,P\}) - c_P, \quad f \in \{I,E\}. \end{aligned}$$

Proposition 4 (Reddit's exclusivity incentive under downstream competition)**.** *Under Assumption 5, Reddit prefers exclusivity if and only if*

$$\Pi_R^X > \Pi_R^N \quad \Longleftrightarrow \quad \gamma > \alpha - \frac{c_P}{V(\{R,P\})}.$$

Intuition. Exclusivity creates a quality gap $q_I^X - q_E^X = V(\{R,P\})$, yielding rent-shifting profits proportional to γ. Under nonexclusive access, this gap vanishes, so Reddit can only monetize absolute quality (α). When relative-quality rents dominate, exclusivity yields higher revenues even though it reduces total information (the entrant does not collect P) and lowers consumer welfare. The region where exclusivity is preferred is increasing if:

- Reddit has smaller datasets (lower $V(\{R,P\})$);
- Stronger substitutability between I and E (higher γ);
- Lower sensitivity of profits to absolute quality (lower α);
- Higher costs of proprietary data (higher c_P).

Welfare Optimum I define welfare as the sum of downstream operating profits:

$$W = \pi_I^D + \pi_E^D.$$

I exclude Reddit's licensing fees, which are pure transfers, and abstract from consumer surplus to isolate the industry-level impact of exclusivity on investment incentives and competition.

Proposition 5 (Welfare comparison)**.** *Assumption 5, exclusivity is never welfare-optimal: total downstream profits are strictly higher under nonexclusive access.*

This analysis focuses on downstream profits only. Including consumer surplus would reinforce the result, since exclusivity reduces both entry and the quality available to users. Thus, the welfare cost of exclusivity is conservative here.

Policy implications. Exclusivity creates a foreclosure channel that operates through complements: by monopolizing a highly informative covariate R, the incumbent weakens rivals' incentives to invest in complementary data sources. Even if entrants retain open access to generic corpora, their reduced incentive to clean, curate, or engineer these sources lowers the overall quality of competition. Therefore, regulators concerned with AI market concentration may treat exclusivity over highly informative datasets (e.g. Reddit, StackOverflow, PubMed, arXiv) analogously to input foreclosure in traditional industries. By locking up key complements, the incumbent amplifies its lead and undermines rivals' ability to compete downstream.

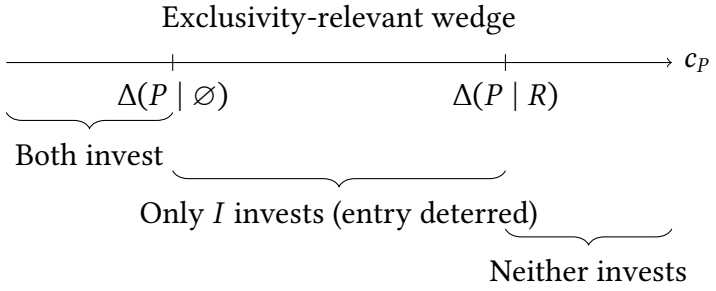


Figure 1: Investment incentives as a function of proprietary data cost c_P.

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[11] 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.[12] 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.

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

[12] 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.

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.

Baumol, William (1977). "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry." *American Economic Review* 67.5, pp. 809–22.

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.

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.

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.

Horn, Roger A. and Charles R. Johnson (2013). *Matrix Analysis*. 2nd. Cambridge; New York: Cambridge University Press.

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*.

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

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.

Vershynin, Roman (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

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

Proposition 6 (Bayes Estimator)**.** *Under the working Gaussian regression*

$$\boldsymbol{y} \mid \boldsymbol{\beta}_{\mathcal{L}} \sim \mathcal{N}(\boldsymbol{X}_{\mathcal{L}}\boldsymbol{\beta}_{\mathcal{L}},\ (1 - S(\mathcal{L}))\boldsymbol{I}_n)\,, \quad \boldsymbol{\beta}_{\mathcal{L}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\mathcal{L}}), \quad \boldsymbol{\beta}_{\mathcal{J}\setminus\mathcal{L}} \text{ independent of } \boldsymbol{\beta}_{\mathcal{L}},$$

the posterior mean satisfies

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{J}\setminus\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \boldsymbol{0}, \qquad \mathbb{E}[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \left(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + (1 - S(\mathcal{L}))\boldsymbol{I}_{\mathcal{L}}\right)^{-1} \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

Proof. Write the log posterior for $\boldsymbol{\beta}_{\mathcal{L}}$ (up to an additive constant):

$$\log p(\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{y}, \boldsymbol{X}_{\mathcal{L}}) = -\frac{1}{2(1 - S(\mathcal{L}))}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{L}}\boldsymbol{\beta}_{\mathcal{L}}\|^2 - \frac{1}{2}\|\boldsymbol{\beta}_{\mathcal{L}}\|^2 + \text{const.}$$

Collect the quadratic terms in $\boldsymbol{\beta}_{\mathcal{L}}$:

$$-\frac{1}{2}\boldsymbol{\beta}_{\mathcal{L}}'\left(\frac{1}{(1 - S(\mathcal{L}))}\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + \boldsymbol{I}_{\mathcal{L}}\right)\boldsymbol{\beta}_{\mathcal{L}} + \frac{1}{(1 - S(\mathcal{L}))}\boldsymbol{\beta}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y} + \text{const.}$$

Complete the square. The posterior is Gaussian with precision

$$\boldsymbol{\Lambda}_{\text{post}} = \frac{1}{(1 - S(\mathcal{L}))}\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + \boldsymbol{I}_{\mathcal{L}},$$

and mean

$$\boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Lambda}_{\text{post}}^{-1} \cdot \frac{1}{(1 - S(\mathcal{L}))}\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

Multiplying numerator and denominator by $(1 - S(\mathcal{L}))$ gives the stated form:

$$\boldsymbol{\mu}_{\text{post}} = \left(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + (1 - S(\mathcal{L}))\boldsymbol{I}_{\mathcal{L}}\right)^{-1} \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

For unlearned coordinates $j \in \mathcal{J}\setminus\mathcal{L}$, independence of parameters implies the posterior equals the prior, whose mean is zero: $\mathbb{E}[\boldsymbol{\beta}_{\mathcal{J}\setminus\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \boldsymbol{0}$. □

Corollary 6 (Bayes Estimator as MNLS with Shrinkage)**.** *The Bayes estimator can be written as a shrinkage transformation of the minimum-norm least-squares (MNLS) estimator:*

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \left(\underbrace{(1 - S(\mathcal{L}))}_{\textit{Shrinkage Factor}} \cdot \left(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}}\right)^{+} + \boldsymbol{I}_{\mathcal{L}}\right)^{-1} \hat{\boldsymbol{\beta}}_{\mathcal{L}}^{MNLS},$$

where $\hat{\boldsymbol{\beta}}_{\mathcal{L}}^{MNLS} = (\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}})^{+}\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}$ *denotes the MNLS estimator and* $(\cdot)^{+}$ *is the Moore-Penrose pseudoinverse.*

Proof. From Proposition **??**, the posterior mean is

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \left(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + (1 - S(\mathcal{L}))\boldsymbol{I}_{\mathcal{L}}\right)^{-1} \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

Factor out $\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}}$ using the identity

$$(\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A} = (\lambda \boldsymbol{A}^{+} + \boldsymbol{I})^{-1} \boldsymbol{A}^{+}\boldsymbol{A},$$

valid for any symmetric $\boldsymbol{A}$ with pseudoinverse $\boldsymbol{A}^{+}$. Applying this to $\boldsymbol{A} = \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}}$ and $\lambda = 1 - S(\mathcal{L})$ gives

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \left((1 - S(\mathcal{L}))(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}})^{+} + \boldsymbol{I}\right)^{-1} (\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}})^{+}\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

Recognizing the last term as the MNLS estimator completes the proof. □

Corollary 3 (Ridge Regression)**.** *The posterior mean coincides with a ridge regression estimator with optimal penalty*

$$\lambda_n^{*}(\mathcal{L}) \equiv \frac{1 - S(\mathcal{L})}{n}.$$

Equivalently,

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \hat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{ridge}}(\lambda)\Big|_{\lambda = \lambda_n^{*}(\mathcal{L})}.$$

Proof. Define the ridge estimator with sample-size normalization as

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

The first-order condition is

$$\frac{1}{n}\boldsymbol{X}_{\mathcal{L}}'(\boldsymbol{X}_{\mathcal{L}}\boldsymbol{b} - \boldsymbol{y}) + \lambda \boldsymbol{b} = \boldsymbol{0} \quad \Longleftrightarrow \quad \left(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + n\lambda\, \boldsymbol{I}_{\mathcal{L}}\right) \boldsymbol{b} = \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

Hence

$$\hat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{ridge}}(\lambda) = \left(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + n\lambda\, \boldsymbol{I}_{\mathcal{L}}\right)^{-1} \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

From Proposition **??**,

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}] = \left(\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}} + (1 - S(\mathcal{L}))\boldsymbol{I}_{\mathcal{L}}\right)^{-1} \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{y}.$$

Setting $n\lambda = 1 - S(\mathcal{L})$, i.e. $\lambda = \lambda_n^{*}(\mathcal{L}) = (1 - S(\mathcal{L}))/n$, the two expressions coincide, proving the claim. □

Lemma 5. *Assume $\mathcal{T} \subseteq \mathcal{L}$. The value of a dataset of type $(\mathcal{L}, \mathcal{T})$ is the variance of the optimal predictor*

$$v(\mathcal{L}, \mathcal{T}) = Var[f^{*}(\boldsymbol{D}_{\mathcal{L},\mathcal{T}})] = \sum_{j \in \mathcal{T}} s_j Var\left[\mathbb{E}\left[\beta_j | \boldsymbol{M}_{\mathcal{L}}\right]\right].$$

Proof. By Lemma 1, the optimal predictor is

$$f^*(\boldsymbol{D}_{\mathcal{L},\mathcal{T}}) = \mathbb{E}[y \mid \boldsymbol{D}_{\mathcal{L},\mathcal{T}}] = \sum_{j\in\mathcal{T}} x_j \,\mathbb{E}\big[\beta_j \mid \boldsymbol{M}_{\mathcal{L}}\big]\,,$$

where we used $\mathcal{T} \subseteq \mathcal{L}$. Under squared loss, the value of information equals the reduction in Bayes risk, which by the law of total variance is

$$v(\mathcal{L},\mathcal{T}) = \mathrm{Var}\left(\mathbb{E}[y \mid \boldsymbol{D}_{\mathcal{L},\mathcal{T}}]\right) = \mathrm{Var}\left(f^*(\boldsymbol{D}_{\mathcal{L},\mathcal{T}})\right).$$

Let $m_j \equiv \mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{L}}]$. Instances are independent across i, hence $\boldsymbol{x}_{\mathcal{T}}$ is independent of $\boldsymbol{M}_{\mathcal{L}}$; covariates are mutually independent with $\mathbb{E}[x_j] = 0$ and $\mathrm{Var}(x_j) = s_j$. Using these facts and $\mathbb{E}[m_j] = \mathbb{E}[\mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{L}}]] = \mathbb{E}[\beta_j] = 0$ (zero-mean prior), we obtain

$$\mathrm{Var}(f^*) = \mathrm{Var}\left(\sum_{j\in\mathcal{T}} x_j m_j\right) = \sum_{j\in\mathcal{T}} \mathrm{Var}(x_j m_j) = \sum_{j\in\mathcal{T}} \mathbb{E}[x_j^2]\,\mathbb{E}[m_j^2] = \sum_{j\in\mathcal{T}} s_j \,\mathrm{Var}(m_j),$$

where cross terms vanish because x_j and x_k are independent with mean zero for $j \neq k$, and because (x_j) is independent of (m_j). Substituting $m_j = \mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{L}}]$ yields the claimed expression. □

Proposition 12 (Variance of Bayes Estimator)**.** *Using $s_j > 0$, the posterior mean satisfies*

$$Var\left(\mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{L}}]\right) = \begin{cases} 0, & j \in \mathcal{J} \setminus \mathcal{L}, \\ \dfrac{1}{1 + \dfrac{\lambda^*(\mathcal{L})}{s_j}} + O\left(\sqrt{\frac{|\mathcal{L}|}{n}} + \frac{|\mathcal{L}|}{n}\right), & j \in \mathcal{L}, \end{cases} \qquad \text{where } \lambda^*(\mathcal{L}) \equiv \frac{\sigma^2(\mathcal{L})}{n}.$$

Proof. If $j \notin \mathcal{L}$, the parameter is never updated so the posterior mean is a.s. zero and the variance is 0.

For $j \in \mathcal{L}$, let $\Sigma_{\text{post}}(\boldsymbol{X}_{\mathcal{L}})$ denote the posterior covariance of $\beta_{\mathcal{L}}$ conditional on $\boldsymbol{X}_{\mathcal{L}}$. With Gaussian prior $\mathcal{N}(0, I)$ and noise variance $\sigma^2(\mathcal{L})$, the standard formula gives

$$\Sigma_{\text{post}}(\boldsymbol{X}_{\mathcal{L}}) = \left(I + \sigma^{-2}(\mathcal{L}) \cdot \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}}\right)^{-1} = \sigma^2(\mathcal{L})\Big(\sigma^2(\mathcal{L})I + \boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}}\Big)^{-1}.$$

By the law of total variance and $\mathrm{Var}(\beta_j) = 1$,

$$\mathrm{Var}\big(\mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{L}}]\big) = 1 - \mathbb{E}\big[\Sigma_{\text{post}}(\boldsymbol{X}_{\mathcal{L}})_{jj}\big]\,.$$

Write $\hat{\boldsymbol{S}}_{\mathcal{L}}(n) \equiv \frac{1}{n}\boldsymbol{X}_{\mathcal{L}}'\boldsymbol{X}_{\mathcal{L}}$, so

$$\Sigma_{\text{post}}(\boldsymbol{X}_{\mathcal{L}}) = \frac{\sigma^2(\mathcal{L})}{n}\left(\hat{\boldsymbol{S}}_{\mathcal{L}} + \frac{\sigma^2(\mathcal{L})}{n} \cdot \boldsymbol{I}\right)^{-1}.$$

Let $S_{\mathcal{L}} \equiv \text{diag}(s_j)_{j\in\mathcal{L}}$ and $E_{\mathcal{L}}(n) \equiv \hat{S}_{\mathcal{L}}(n) - S_{\mathcal{L}}$. Define

$$A \equiv \hat{S}_{\mathcal{L}} + \frac{\sigma^2(\mathcal{L})}{n} I, \qquad B \equiv S_{\mathcal{L}} + \frac{\sigma^2(\mathcal{L})}{n} I.$$

Since the smallest eigenvalue of $S_{\mathcal{L}} + \frac{\sigma^2(\mathcal{L})}{n} \cdot I$ is at least $s_{\min} > 0$, $\Sigma_{\text{post}}(X_{\mathcal{L}})$ is invertible and we can use the resolvent identity in Horn and Johnson (2013) Section 5.8:

$$A^{-1} - B^{-1} = A^{-1}(B - A)\, B^{-1} = -A^{-1} E_{\mathcal{L}} B^{-1}.$$

Therefore

$$\Sigma_{\text{post}}(X_{\mathcal{L}}) = \frac{\sigma^2(\mathcal{L})}{n} A^{-1} = \frac{\sigma^2(\mathcal{L})}{n} B^{-1} - \frac{\sigma^2(\mathcal{L})}{n} \left(A^{-1} - B^{-1}\right).$$

Taking the (j, j) entry and using $|(\cdot)_{jj}| \le \|\cdot\|_{\text{op}}$ and the triangle inequality,

$$\left|\Sigma_{\text{post}}(X_{\mathcal{L}})_{jj} - \frac{\sigma^2(\mathcal{L})}{n} (B^{-1})_{jj}\right| \le \frac{\sigma^2(\mathcal{L})}{n} \|A^{-1} - B^{-1}\|_{\text{op}} \le \frac{\sigma^2(\mathcal{L})}{n} \|A^{-1}\|_{\text{op}} \|B^{-1}\|_{\text{op}} \|E_{\mathcal{L}}\|_{\text{op}}.$$

The operator norm $\|A\|_{\text{op}}$ of a matrix A is the largest singular value of A and measures the maximum action A can have on any vector. Now $\lambda_{\min}(A) \ge \lambda^*(\mathcal{L}) > 0$ because $\hat{S}_{\mathcal{L}}$ is positive semi-definite, so

$$\frac{\sigma^2(\mathcal{L})}{n} \|A^{-1}\|_{\text{op}} = \lambda^*(\mathcal{L}) \|A^{-1}\|_{\text{op}} \le 1,$$

and $\|B^{-1}\|_{\text{op}} \le 1/s_{\min}$. Hence

$$\left|\Sigma_{\text{post}}(X_{\mathcal{L}})_{jj} - \frac{\sigma^2(\mathcal{L})}{n} (B^{-1})_{jj}\right| \le \frac{1}{s_{\min}} \|E_{\mathcal{L}}(n)\|_{\text{op}}.$$

Taking expectations and using the standard bound on the sample variance-covariance matrix in Vershynin (2018) Theorem 4.7.1,

$$\mathbb{E}\, \|E_{\mathcal{L}}(n)\|_{\text{op}} = O\left(\sqrt{\frac{|\mathcal{L}|}{n}} + \frac{|\mathcal{L}|}{n}\right).$$

Therefore,

$$\mathbb{E}\left[\Sigma_{\text{post}}(X_{\mathcal{L}})_{jj}\right] = \frac{\sigma^2(\mathcal{L})}{n} (B^{-1})_{jj} + O\left(\sqrt{\frac{|\mathcal{L}|}{n}} + \frac{|\mathcal{L}|}{n}\right).$$

Since B^{-1} is diagonal with $(B^{-1})_{jj} = 1/(s_j + \lambda^*(\mathcal{L}))$,

$$\frac{\sigma^2(\mathcal{L})}{n} (B^{-1})_{jj} = \frac{\lambda^*(\mathcal{L})}{s_j + \lambda^*(\mathcal{L})} = 1 - \frac{s_j}{s_j + \lambda^*(\mathcal{L})}.$$

Therefore

$$\mathrm{Var}\big(\mathbb{E}[\beta_j \mid \boldsymbol{M}_{\mathcal{L}}]\big) = 1 - \mathbb{E}\big[\Sigma_{\mathrm{post}}(\boldsymbol{X}_{\mathcal{L}})_{jj}\big] = \frac{s_j}{s_j + \lambda^*(\mathcal{L})} + O\left(\sqrt{\frac{|\mathcal{L}|}{n}} + \frac{|\mathcal{L}|}{n}\right),$$

which is $o(1)$ whenever $|\mathcal{L}| = o(n)$. □

Proposition 7 (Returns to learning covariates). *Fix $t \in (0,1]$ and let $s : [0,1] \to \mathbb{R}_+$ be continuously differentiable with s nonincreasing. Define $S(\ell) = \int_0^\ell s(u)\,du$, the ridge-variance kernel*

$$v^{ridge}(u;\lambda) \equiv \frac{s(u)}{\lambda + s(u)}, \qquad v_\lambda = -\frac{s(u)}{(\lambda + s(u))^2}, \qquad v_{\lambda\lambda} = \frac{2\,s(u)}{(\lambda + s(u))^3},$$

and the penalty

$$\lambda^*(n,\ell) \equiv \frac{1 - S(\ell)}{n}, \qquad \lambda^*_\ell = -\frac{s(\ell)}{n}, \qquad \lambda^*_{\ell\ell} = -\frac{s'(\ell)}{n}.$$

Let

$$V(t,\lambda) \equiv \int_0^t s(u)\, v^{ridge}(u;\lambda)\,du.$$

Then

$$V_\ell(t,\lambda^*(n,\ell)) = \int_0^t s(u)\, v^{ridge}_\lambda(u;\lambda^*)\,\lambda^*_\ell(n,\ell)\,du \;\geq 0,$$
$$V_{\ell\ell}(t,\lambda^*(n,\ell)) = \int_0^t s(u)\Big(v^{ridge}_\lambda(u;\lambda^*)\,\lambda^*_{\ell\ell}(n,\ell) + v^{ridge}_{\lambda\lambda}(u;\lambda^*)\,[\lambda^*_\ell(n,\ell)]^2\Big)\,du.$$

Moreover,

$$\lim_{n\to 0^+} V_{\ell\ell}(t,\lambda^*(n,\ell)) > 0 \iff -s'(\ell) < \frac{2\,s(\ell)^2}{1 - S(\ell)},$$

and

$$\lim_{n\to\infty} V_{\ell\ell}(t,\lambda^*(n,\ell)) = 0^-.$$

Consequently, for each ℓ there exists a (possibly infinite) threshold $\hat{n}(\ell) \in (0,\infty]$ such that $V_{\ell\ell}(t,\lambda^*(n,\ell)) > 0$ for $n < \hat{n}(\ell)$, $V_{\ell\ell}(t,\lambda^*(n,\ell)) < 0$ for $n > \hat{n}(\ell)$, and

$$V_{\ell\ell}(t,\lambda^*(\hat{n}(\ell),\ell)) = 0.$$

Proof. Step 1. Sign of V_ℓ. By definition,

$$V_\ell(t,\lambda^*) = \int_0^t s(u)\, v^{\mathrm{ridge}}_\lambda(u;\lambda^*)\,\lambda^*_\ell\,du.$$

For all u, $v^{\mathrm{ridge}}_\lambda(u;\lambda^*) = -s(u)/(\lambda^* + s(u))^2 < 0$ and $\lambda^*_\ell = -s(\ell)/n < 0$, hence the integrand is nonnegative and $V_\ell \geq 0$.

Step 2. Decomposition of $V_{\ell\ell}$. Differentiating once more,

$$V_{\ell\ell}(t,\lambda^*) = \int_0^t s(u)\Big(v_\lambda^{\text{ridge}}\,\lambda^*_{\ell\ell} + v_{\lambda\lambda}^{\text{ridge}}\,(\lambda^*_\ell)^2\Big)\,du,$$

with $v_{\lambda\lambda}^{\text{ridge}}(u;\lambda^*) = 2\,s(u)/(\lambda^* + s(u))^3 > 0$. Since s is nonincreasing, $s'(\ell) \le 0$, so $\lambda^*_{\ell\ell} = -s'(\ell)/n \ge 0$ and $v_\lambda^{\text{ridge}} < 0$, making the first term weakly negative ("LLN"), while the second term is strictly positive ("HPE").

Step 3. Small-n asymptotics and the positivity condition. Using $\lambda^* = (1 - S(\ell))/n$ and the identities above,

$$V_{\ell\ell}(t,\lambda^*) = \frac{s'(\ell)}{n}\int_0^t \frac{s(u)^2}{(\lambda^* + s(u))^2}\,du + \frac{2\,s(\ell)^2}{n^2}\int_0^t \frac{s(u)^2}{(\lambda^* + s(u))^3}\,du.$$

As $n \to 0^+$, we have $\lambda^* \to \infty$, so

$$\int_0^t \frac{s(u)^2}{(\lambda^* + s(u))^2}\,du = \frac{1}{(\lambda^*)^2}\int_0^t s(u)^2\,du + o((\lambda^*)^{-2}),$$

$$\int_0^t \frac{s(u)^2}{(\lambda^* + s(u))^3}\,du = \frac{1}{(\lambda^*)^3}\int_0^t s(u)^2\,du + o((\lambda^*)^{-3}).$$

Substituting $\lambda^* = (1 - S(\ell))/n$ and collecting the $O(n)$ terms yields

$$V_{\ell\ell}(t,\lambda^*(n,\ell)) = \frac{n}{(1 - S(\ell))^3}\Big(2\,s(\ell)^2 + s'(\ell)\,(1 - S(\ell))\Big)\int_0^t s(u)^2\,du + o(n).$$

Because $\int_0^t s(u)^2 du > 0$ and $(1 - S(\ell)) > 0$, the leading term is positive iff

$$2\,s(\ell)^2 + s'(\ell)\,(1 - S(\ell)) > 0 \iff -s'(\ell) < \frac{2\,s(\ell)^2}{1 - S(\ell)}.$$

Step 4. Large-n asymptotics. As $n \to \infty$, $\lambda^* \to 0$. Then $v_\lambda^{\text{ridge}}(u;\lambda^*) \to -1/s(u)$ and $v_{\lambda\lambda}^{\text{ridge}}(u;\lambda^*) \to 2/s(u)^2$. Since $\lambda^*_{\ell\ell} = -s'(\ell)/n$ and $(\lambda^*_\ell)^2 = s(\ell)^2/n^2$,

$$\begin{aligned} V_{\ell\ell}(t,\lambda^*(n,\ell)) &= \int_0^t s(u)\left[\left(-\frac{1}{s(u)} + o(1)\right)\left(-\frac{s'(\ell)}{n}\right) + \left(\frac{2}{s(u)^2} + o(1)\right)\frac{s(\ell)^2}{n^2}\right]du \\ &= \frac{s'(\ell)}{n}\,t \;+\; O\left(\frac{1}{n^2}\right). \end{aligned}$$

Since $s'(\ell) \le 0$, this shows $V_{\ell\ell}(t,\lambda^*(n,\ell)) \to 0^-$.

Step 5. Existence of a threshold. By Step 3, for n small enough the sign is determined by the condition in the statement; when it holds, $V_{\ell\ell} > 0$ for sufficiently small n. By Step 4, for n large $V_{\ell\ell} < 0$ and tends to 0^-. Continuity in n (from dominated convergence) then yields a threshold $\hat{n}(\ell) \in (0,\infty]$ with the stated sign pattern and $V_{\ell\ell}(t,\lambda^*(\hat{n}(\ell),\ell)) = 0$. □

Remark 1 (On monotonicity of the threshold). *It is convenient to rewrite the zero-curvature condition as*

$$\frac{\lambda^*_{\ell\ell}}{(\lambda^*_\ell)^2} = -\frac{\displaystyle\int_0^t s(u)\, v^{ridge}_{\lambda\lambda}(u;\lambda^*)\, du}{\displaystyle\int_0^t s(u)\, v^{ridge}_{\lambda}(u;\lambda^*)\, du} = 2\,\mathbb{E}_w\left[\frac{1}{\lambda^* + s(u)}\right],$$

where the weights are $w(u) \propto s(u)\, s(u)/(\lambda^* + s(u))^2$ *on* $[0,t]$. *Using* $\lambda^*_\ell = -s(\ell)/n$ *and* $\lambda^*_{\ell\ell} = -s'(\ell)/n$, *the left-hand side equals* $(-s'(\ell))\, n/s(\ell)^2$. *Thus at the threshold,*

$$\frac{-s'(\ell)}{s(\ell)^2}\,\hat{n}(\ell) = 2\,\mathbb{E}_w\left[\frac{1}{\lambda^*(\hat{n}(\ell),\ell) + s(u)}\right], \qquad \lambda^*(\hat{n}(\ell),\ell) = \frac{1 - S(\ell)}{\hat{n}(\ell)}.$$

The map $\ell \mapsto (-s'(\ell))/s(\ell)^2$ *is nondecreasing whenever* s *is log-concave (indeed* $d[-s'/s^2]/d\ell = (2(s')^2 - s''s)/s^3 \geq 0$ *under* $s''s - (s')^2 \leq 0$*). The right-hand side is decreasing in* λ^* *and hence increasing in* $\hat{n}(\ell)$ *and in* ℓ *(via* $1 - S(\ell)$*). Therefore, the comparative statics of* $\hat{n}(\ell)$ *depend on the relative strength of these opposing forces and, in general, need not be monotone without further restrictions beyond log-concavity. A sufficient condition ensuring that* $\hat{n}(\ell)$ *is nonincreasing is that the increase of* $(-s'(\ell))/s(\ell)^2$ *with* ℓ *dominates the induced increase in the weighted average on the right-hand side; this holds, for example, when* s *is log-concave and* t *is small enough that* $s(u)$ *varies little over* $[0,t]$.

Proposition 8 (Complementarity/Substitutability of learning breadth and observations). *Fix* $t \in (0,1]$. *Let* $s : [0,1] \to \mathbb{R}_+$ *be continuously differentiable and nonincreasing, and set*

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

Define

$$v^{ridge}(u;\lambda) = \frac{s(u)}{\lambda + s(u)}, \quad v_\lambda(u;\lambda) = -\frac{s(u)}{(\lambda + s(u))^2}, \quad v_{\lambda\lambda}(u;\lambda) = \frac{2\, s(u)}{(\lambda + s(u))^3},$$

and

$$V(t,\lambda) = \int_0^t s(u)\, v^{ridge}(u;\lambda)\, du.$$

Then the cross-partial of the value with respect to learning breadth ℓ *and sample size* n *equals*

$$V_{\ell n}(t,\lambda^*(n,\ell)) = \int_0^t s(u)\Big(v_\lambda(u;\lambda^*)\,\lambda^*_{\ell n}(n,\ell) + v_{\lambda\lambda}(u;\lambda^*)\,\lambda^*_\ell(n,\ell)\,\lambda^*_n(n,\ell)\Big)\, du,$$

with

$$\lambda^*_\ell = -\frac{s(\ell)}{n}, \qquad \lambda^*_n = -\frac{\lambda^*}{n}, \qquad \lambda^*_{\ell n} = \frac{s(\ell)}{n^2}.$$

In particular,

$$\lim_{n\to 0^+} V_{\ell n}(t,\lambda^*(n,\ell)) > 0, \qquad \lim_{n\to\infty} V_{\ell n}(t,\lambda^*(n,\ell)) = 0^-.$$

Hence there exists a unique $\bar{\lambda} > 0$ solving

$$\int_0^t \frac{s(u)^2\,(\bar{\lambda} - s(u))}{(\bar{\lambda} + s(u))^3}\,du = 0,$$

and the threshold

$$\hat{n}_{V_{\ell n}}(\ell) \;\equiv\; \frac{1 - S(\ell)}{\bar{\lambda}}$$

satisfies $V_{\ell n}(t, \lambda^*(n,\ell)) > 0$ iff $n < \hat{n}_{V_{\ell n}}(\ell)$ and $V_{\ell n}(t, \lambda^*(n,\ell)) < 0$ iff $n > \hat{n}_{V_{\ell n}}(\ell)$. Moreover $\hat{n}_{V_{\ell n}}(\ell)$ is strictly decreasing in ℓ.

Proof. Step 1. Cross-partial formula and factorization. Differentiate $V(t, \lambda^*(n,\ell))$ first in ℓ then in n:

$$V_{\ell n}(t, \lambda^*) = \int_0^t s(u)\Big(v_\lambda \lambda^*_{\ell n} + v_{\lambda\lambda}\lambda^*_\ell \lambda^*_n\Big)\,du.$$

Using $\lambda^*_\ell = -s(\ell)/n$, $\lambda^*_n = -\lambda^*/n$, and $\lambda^*_{\ell n} = s(\ell)/n^2$, pull out the common factor $s(\ell)/n^2$ and combine terms inside the integral:

$$V_{\ell n}(t, \lambda^*) = \frac{s(\ell)}{n^2}\int_0^t s(u)\Big(v_\lambda(u;\lambda^*) + \lambda^* v_{\lambda\lambda}(u;\lambda^*)\Big)\,du.$$

With $v_\lambda = -s(u)/(\lambda^* + s(u))^2$ and $v_{\lambda\lambda} = 2s(u)/(\lambda^* + s(u))^3$,

$$v_\lambda + \lambda^* v_{\lambda\lambda} = \frac{s(u)}{(\lambda^* + s(u))^3}\big(\lambda^* - s(u)\big).$$

Therefore

$$V_{\ell n}(t, \lambda^*(n,\ell)) = \frac{s(\ell)}{n^2}\,F(\lambda^*(n,\ell)), \qquad F(\lambda) \equiv \int_0^t \frac{s(u)^2\,(\lambda - s(u))}{(\lambda + s(u))^3}\,du. \tag{4}$$

Step 2. Limits as $n \to 0^+$ and $n \to \infty$. As $n \to 0^+$, $\lambda^* = (1 - S(\ell))/n \to \infty$. Then

$$F(\lambda^*) = \int_0^t \frac{s(u)^2}{(\lambda^*)^2}\frac{\lambda^* - s(u)}{\big(1 + s(u)/\lambda^*\big)^3}\,du = \frac{1}{(\lambda^*)^2}\int_0^t s(u)^2\,du + o((\lambda^*)^{-2}).$$

Using $\lambda^* = (1 - S(\ell))/n$ in (4) yields

$$\lim_{n\to 0^+} V_{\ell n}(t, \lambda^*(n,\ell)) = \frac{s(\ell)}{(1 - S(\ell))^2}\int_0^t s(u)^2\,du \;>\; 0,$$

provided $s(\ell) > 0$ and $s \not\equiv 0$ on $[0, t]$ (true under our standing assumptions).
As $n \to \infty$, $\lambda^* \to 0$. Then

$$F(\lambda^*) = -\int_0^t 1\,du + O(\lambda^*) = -t + O(\lambda^*),$$

so from (4)

$$V_{\ell n}(t, \lambda^*(n,\ell)) = -\frac{t\, s(\ell)}{n^2} + o\left(\frac{1}{n^2}\right) \xrightarrow[n\to\infty]{} 0^-.$$

Step 3. Existence and uniqueness of a $\bar{\lambda} > 0$ with $F(\bar{\lambda}) = 0$. The function F is continuous on $(0, \infty)$ by dominated convergence. Moreover

$$F(0) = -\int_0^t 1\, du = -t < 0, \qquad \lim_{\lambda\to\infty} F(\lambda) = 0^+,$$

and the large-λ expansion gives $F(\lambda) = I_2/\lambda^2 - I_3/\lambda^3 + o(\lambda^{-3}) > 0$ for λ large, where $I_k = \int_0^t s(u)^k\, du$. Thus there exists at least one $\bar{\lambda} > 0$ with $F(\bar{\lambda}) = 0$.
To see uniqueness, note that

$$F'(\lambda) = 2\int_0^t \frac{s(u)^2\,(2s(u) - \lambda)}{(\lambda + s(u))^4}\, du,$$

so $F'(0) = 2\int_0^t \frac{2}{s(u)}\, du > 0$, while for λ large enough $F'(\lambda) < 0$. Hence F is increasing on a neighborhood of 0, eventually decreasing for large λ, and $\lim_{\lambda\to\infty} F(\lambda) = 0^+$. Since $F(0) < 0$, F can cross the zero level at most once; therefore the zero is unique.

Step 4. Threshold and its monotonicity in ℓ. By (4), $\operatorname{sign} V_{\ell n}(t, \lambda^*(n,\ell)) = \operatorname{sign} F(\lambda^*(n,\ell))$ because $s(\ell)/n^2 > 0$. With the unique $\bar{\lambda}$ from Step 3, define

$$\hat{n}_{V_{\ell n}}(\ell) \equiv \frac{1 - S(\ell)}{\bar{\lambda}}.$$

Then $V_{\ell n} > 0$ iff $\lambda^*(n,\ell) > \bar{\lambda}$ iff $n < \hat{n}_{V_{\ell n}}(\ell)$; and $V_{\ell n} < 0$ iff $n > \hat{n}_{V_{\ell n}}(\ell)$. Since $S(\ell)$ is increasing in ℓ, $1 - S(\ell)$ is decreasing, so $\hat{n}_{V_{\ell n}}(\ell)$ is strictly decreasing in ℓ. □

Corollary 4 (Complementarity across datasets with fixed n)**.** *Fix n and t. For $\ell_1, \ell_2 \geq 0$ with $\ell_1 + \ell_2 \leq \bar{\ell}$, define the supermodularity gap*

$$\Delta(\ell_1, \ell_2) \equiv V(n, \ell_1 + \ell_2, t) - V(n, \ell_1, t) - V(n, \ell_2, t) > 0 \iff n \leq \hat{n}_{V_{\ell\ell}}(\ell).$$

so datasets are complements if and only if observations are scarce. In particular, if $s(u) = 1$ for all u, datasets are always complements.

Proof. Apply the mean value theorem twice to the function $\ell \mapsto V(n, \ell, t)$ to obtain $\Delta(\ell_1, \ell_2) = \frac{1}{2} V_{\ell\ell}(t, \lambda^*(n, \tilde{\ell}))(\ell_1\ell_2)$ for some $\tilde{\ell} \in (0, \ell_1 + \ell_2)$, so the sign of Δ equals the sign of $V_{\ell\ell}$. Proposition 7 gives the expression and its comparative statics in n and ℓ, yielding the threshold characterization and the constant-$s(u)$ special case. □

Corollary 5 (Substitutability across datasets with fixed ℓ)**.** *Fix ℓ and t. For $n_1, n_2 > 0$, define the supermodularity gap*

$$\Delta_n(n_1, n_2) \equiv V(n_1 + n_2, \ell, t) - V(n_1, \ell, t) - V(n_2, \ell, t) < 0,$$

so the datasets are substitutes.

Proof. Apply the mean value theorem twice to $\phi(n) \equiv V(n, \ell, t)$: for some $\tilde{n} \in (0, n_1 + n_2)$, $\Delta_n(n_1, n_2) = \frac{1}{2} V_{nn}(t, \lambda^*(\tilde{n}, \ell))\, n_1 n_2$. Proposition 6 establishes $V_{nn}(t, \lambda^*(n, \ell)) < 0$ for all $n > 0$ (diminishing returns to observations). Therefore $\Delta_n(n_1, n_2) < 0$, proving strict substitutability. □

Proposition 9 (Learning Cost)**.** *For any learning level $L \geq \frac{n}{1-S(t)}$, the required number of covariates is*

$$\tilde{\ell}(L) = S^{-1}\left(1 - \tfrac{n}{L}\right) > t,$$

and the associated cost function is

$$c(L) = \gamma\, \tilde{\ell}(L), \qquad \dot{c}(L) = \frac{\gamma n}{L^2 s(\tilde{\ell}(L))}.$$

Proof. The constraint $\bar{L} \leq L(\ell)$ is binding at the minimum cost solution, so $\bar{L} = L(\ell) = \frac{n}{1-S(\ell)}$. Solving for ℓ gives $\tilde{\ell}(L) = S^{-1}(1 - n/L)$. Multiplying by unit cost γ yields $c(L)$. Differentiation gives $\dot{c}(L)$. □

Proposition 10 (Cost of Information)**.** *For any information value $V \geq 0$, the implied learning demand is $L(V) = V^{-1}(V)$,and the information cost is $C(V) = c(L(V))$. Average and marginal costs satisfy*

$$\mathrm{AC}(V) = \frac{c(L(V))}{V}, \qquad \mathrm{MC}(V) = \frac{\dot{c}(L(V))}{\dot{V}(L(V))}.$$

Proof. By the implicit function theorem, the learning level consistent with V is uniquely defined as $L(V) = V^{-1}(V)$. Substitution into the learning cost function gives $C(V)$. Average and marginal cost formulas follow by direct division and the chain rule. □

Proposition 11 (Natural monopoly condition)**.** *At any V,*

$$\mathrm{AC}(V) \;\geq\; \mathrm{MC}(V) \;\iff\; \epsilon_V\big(L(V)\big) \;\geq\; \epsilon_c\big(L(V)\big).$$

Proof. Using $C(V) = c(L(V))$,

$$\mathrm{AC}(V) \geq \mathrm{MC}(V) \quad \iff \quad \frac{c(L(V))}{V} \;\geq\; \frac{c'(L(V))}{V'(L(V))}.$$

Proposition 12 (Profit Maximization)**.** *At the optimum V^*, the market is a natural monopoly if and only if*

$$\mathrm{AC}(V^*) \geq \mathrm{MC}(V^*) \iff M_G < \bar{M} \equiv \min_V AC\,(V)\,.$$

It follows from the natural monopoly condition in Proposition ?? evaluated at the firm's optimal choice V^* characterized in Proposition ??. □

Lemma 6 (Learning required to reach a fixed value)**.** *For any fixed $V \in (0,1)$ and feasible (n,t), the least learning depth that attains V is the unique $L(V,t)$ that solves $V(L,t) = V$. It satisfies*

$$\frac{\partial L}{\partial t}\Big|_V = -\frac{\partial_t V(L,t)}{\partial_L V(L,t)} \le 0, \qquad \frac{\partial L}{\partial n}\Big|_V = 0.$$

[Pointwise cost shifts] For any fixed L and n,

$$\frac{\partial c}{\partial L}(L;n) = \dot{c}(L;n) = \frac{\gamma n}{L^2\, s(\tilde{\ell}(L))} > 0, \qquad \frac{\partial c}{\partial n}(L;n) = -\frac{\gamma}{L\, s(\tilde{\ell}(L))} < 0,$$

where $\tilde{\ell}(L) = S^{-1}(1 - n/L)$.

Proposition 13 (Cost at a given value V)**.** *Fix V and suppose the feasibility bound does not bind at $L(V,t)$, i.e. $L(V,t) > L^{\min}(t)$. Then*

$$\frac{\partial C}{\partial n}(V;n,t) = \frac{\partial c}{\partial n}\big(L(V,t);n\big) < 0, \qquad \frac{\partial C}{\partial t}(V;n,t) = \frac{\partial c}{\partial L}\big(L(V,t);n\big)\cdot\frac{\partial L}{\partial t}\Big|_V \le 0.$$

Hence both average and marginal cost, $\mathrm{AC}(V) = C(V)/V$ and $\mathrm{MC}(V) = \dot{c}(L)/\dot{V}(L)$, shift weakly downward in n and t pointwise in V.

Proposition 13 (Effect of n and t on the natural-monopoly threshold)**.** *Let $\bar{M}(n,t) \equiv \min_{V>0} \mathrm{AC}(V;n,t)$, and assume the argmin $V^{\mathrm{AC}}(n,t)$ is interior (equivalently $\mathrm{AC} = \mathrm{MC}$ at V^{AC}). Then*

$$\frac{\partial \bar{M}}{\partial n} = \frac{1}{V^{\mathrm{AC}}}\frac{\partial C}{\partial n}\Big(V^{\mathrm{AC}};n,t\Big) < 0, \qquad \frac{\partial \bar{M}}{\partial t} = \frac{1}{V^{\mathrm{AC}}}\frac{\partial C}{\partial t}\Big(V^{\mathrm{AC}};n,t\Big) \le 0.$$

Moreover, the inequality is strict whenever $s(t)\, v^{ridge}(t, 1/L(V^{\mathrm{AC}},t)) > 0$.

Proof. By the envelope theorem for a value that minimizes $\phi(V) \equiv C(V)/V$, $d\bar{M}/d\xi = \partial\phi/\partial\xi|_{V=V^{\mathrm{AC}}} = (1/V^{\mathrm{AC}})\,\partial C/\partial\xi$ for any parameter $\xi \in \{n,t\}$. Combine with the previous proposition. □

Proposition 14 (Data List Merger)**.** *Suppose two brokers sell lists that only increase n with ℓ fixed. Then $V_{nn} < 0$ implies the datasets are strict substitutes: for all $n_1, n_2 > 0$, $V(n_1{+}n_2,\ell,t) - V(n_1,\ell,t) - V(n_2,\ell,t) < 0$. Therefore competition between list brokers lowers prices relative to integrated monopoly. A merger of list brokers raises price and reduces adoption.*

Proof. A buyer who purchases a set $S \subseteq \{1,2\}$ of brokers obtains value θV_S where I define incremental values relative to the null:

$$V_i = V(n_i,\ell,t), \quad V_{12} = V(n_1+n_2,\ell,t),$$

I have $V_{12} - V_1 - V_2 < 0$ since $V_{nn} < 0$, so the incremental value of extra observations is diminishing and datasets that differ only in n are strict substitutes. Price competition between substitutable lists pushes prices down; a merger softens this competition and raises the bundle price toward the monopoly optimum $P^M = \hat{\theta}\, V_{12}$.

Therefore, mergers between list brokers who primarily expand n (with similar ℓ) are presumptively harmful: higher prices, lower adoption, and lower consumer surplus. □

Proposition 15 (Data Append Merger)**.** *Under Assumption* **??**, *a merger between two data append brokers can be either anticompetitive or procompetitive:*

- *If $n \leq \hat{n}_{V_{\ell\ell}}(\ell)$, the data appends are complements. Hence, the merger eliminates double marginalization so it is procompetitive.*
- *If $n > \hat{n}_{V_{\ell\ell}}(\ell)$, the datasets are substitutes. Hence, the merger weakly increases price and strictly lowers adoption, so it is anticompetitive.*

Given $\hat{n}_{V_{\ell\ell}}(\ell)$ is decreasing in ℓ, if data are scarce, it is more likely the merger is procompetitive. In the special case of equal covariates, the merger is always procompetitive.

Proof. A buyer who purchases a set $S \subseteq \{1, 2\}$ of brokers obtains value θV_S where I define incremental values relative to the null:

$$V_i \equiv V(n, \ell_i, t), \qquad V_{12} \equiv V(n, \ell_1{+}\ell_2, t),$$

and the supermodularity (synergy) term

$$\Delta \equiv V_{12} - V_1 - V_2 = \Delta(\ell_1, \ell_2).$$

The buyer purchases the pair of datasets after merger iff

$$\theta \geq \frac{p_1 + p_2}{V_{12}},$$

so demand is $D(p_1{+}p_2) = 1 - G\big((p_1{+}p_2)/V_{12}\big)$. Broker i solves

$$\max_{p_i \geq 0} \; p_i\, D(p_1{+}p_2).$$

Let $z \equiv (p_1{+}p_2)/V_{12}$. The FOC for a symmetric equilibrium $p_1 = p_2 = p$ is

$$1 - G(z) = \frac{p}{V_{12}}\, g(z) = \frac{z}{2}\, g(z).$$

Define $H(\theta) \equiv \theta g(\theta)/[1 - G(\theta)]$. Under the thin-tail Assumption 1, H is increasing, so the pricing problem has a unique solution. The symmetric duopoly condition is

$$H(z^D) = 2.$$

An integrated monopolist selling the bundle chooses P with FOC

$$1 - G\Big(\frac{P}{V_{12}}\Big) = \frac{P}{V_{12}}\, g\Big(\frac{P}{V_{12}}\Big), \quad \text{so} \quad H(z^M) = 1, \qquad z^M \equiv \frac{P^M}{V_{12}}.$$

Since H is increasing, $H(z^D) = 2 > 1 = H(z^M)$ implies $z^D > z^M$ and

$$p_1^D + p_2^D = z^D V_{12} > z^M V_{12} = P^M.$$

Noncooperative duopoly exhibits double marginalization with complementary appends: the total price is higher, adoption is lower, and static welfare is below the merged outcome. If single-dataset purchases yield some value $V_i > 0$, algebra adds regions, but the same qualitative force holds when $\Delta \geq 0$. □

Proposition 16 (Welfare comparison)**.** *Assumption 5, exclusivity is never welfare-optimal: total downstream profits are strictly higher under nonexclusive access.*

Proof. Under nonexclusive access both firms enter and invest in P, yielding

$$W^N = 2\left(\alpha V(\{R, P\}) - c_P\right).$$

Under exclusivity, only the incumbent invests in P, so

$$W^X = (\alpha + \gamma) V(\{R, P\}) - c_P.$$

Since $c_P \in \big(\Delta(P|\varnothing), \Delta(P|R)\big]$, it follows that

$$W^N > W^X.$$

□

B Extensions

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

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

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

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

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

B.2 Shrinkage Interpretation

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

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

where $(\cdot)^+$ denotes the Moore–Penrose pseudo-inverse.[13] 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 esti-

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

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

mator, 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.

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 7. *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 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 8. *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[14]

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

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[15]

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

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.

[14] 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.

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

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 policy terms, data economies of scale saturate exactly when further capacity gains are information-theoretically expensive.

Table 2: Types of predictions and policy implications

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