Opening the Black Box*

A Theory of the Value of Data

Giovanni Rizzi†

Latest Draft

Abstract

This paper develops a theory of the value of data for prediction. An agent observes an outcome and some characteristics (covariates) of a sample of individuals to predict the outcome for a target individual based on her characteristics. The main findings are: (i) covariates exhibit economies of scope, as the value of one covariate is higher when others are also observed; (ii) covariates and observations are complements when data are scarce but become substitutes when data are abundant. These findings have three policy implications. Mergers between firms having different covariates can be privately profitable yet reduce welfare, especially when data are scarce (e.g., under strict privacy rules). Allowing firms to pool covariates is always procompetitive because it removes double marginalization, whereas pooling observations can be anticompetitive when data are abundant. Finally, a data seller may profitably exclude one of several competing prediction providers even when this lowers total welfare.

JEL Classification: C11, D83, L12, L40.

Keywords: Value of Data, Prediction, Economies of Scope, Data Markets, Digital Platforms.

*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, Alexandre de Cornière, Andrei Hagiu, Johannes Horner, Bruno Jullien, Hiroaki Kaido, Simon Loertscher, Friedrich Lucke, Leonardo Madio, Thierry Magnac, Nour Meddahi, Giovanni Morzenti, Juan Ortner, Christoph Reidl, Andrew Rhodes, Maximilian Schaefer, Sara Shahanaghi, Tim Simcoe, Alex Smolin, Marshall Van Alstyne, Davide Viviano, as well as seminar participants at the European Association for Research in Industrial Economics Conference (Valencia, 2025), the Questrom Digital Platforms Seminar, and at TSE and Boston University.

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

1 Introduction

In digital markets, data drive competitive advantage because they allow firms to make better predictions: ChatGPT and Gemini predict the next response a user is likely to find helpful given a prompt and context; Amazon and Uber predict where demand will arise; Google and Meta predict which ad a user will click; and Netflix predicts what content a user will enjoy most. Over the last decade, many digital markets have become dominated by a few firms that control vast amounts of data, attracting the attention of competition authorities. For instance, according to the 2021 U.S. House Report, "data advantages [...] can reinforce dominance and serve as a barrier to entry,"[1] a warning echoed in the EU Data Act proposal[2] and in major policy reports.[3]

These concerns are partly fueled by earlier statements from industry leaders: as Google's CEO declared in 2009, "Scale is the key. We just have so much scale in terms of the data we can bring to bear."[4] However, more recently, technology firms have downplayed policymakers' claims, arguing that data exhibit diminishing returns to scale by the Law of Large Numbers, and that data concentration in the hands of few firms reflects their technological advantage on rivals rather than entry barriers.[5]

To evaluate these competing narratives, we must consider two distinct ways in which firms acquire data: either by observing more individuals (i.e., increasing sample size) or by observing more attributes of each individual (i.e., increasing the covariates). This paper asks a simple but unresolved question: how do predictions improve when a firm acquires (i) more observations and/or (ii) more covariates? Empirically and conceptually, these two channels—*economies of scale* and *economies of scope*—are often conflated. My goal is to disentangle them.

I develop a Bayesian model that cleanly disentangles these two dimensions. A firm predicts an outcome for a target individual—for example, the price of a house. The outcome depends on many potential covariates (observable attributes of the house, e.g., size, number of rooms, year of construction), each of which has an unknown impact on the outcome. The firm begins with prior uncertainty about these impacts and can collect two types of data: (1) *Training data*: covariates and outcomes for a sample of individuals, used to learn the relationship between covariates and outcomes; (2) *Prediction covariates*: a subset of the covariates of the target individual, used to apply what the firm has learned from the training data.

Consequently, the firm (Zillow) designs a dataset by choosing three objects: (i) the number of observations in the training data N (the number of past house prices), (ii) the set of *training covariates* $\mathcal{T}$ (which house characteristics on past purchases), and (iii) the set of *prediction*

[1] U.S. House of Representatives (2020), Investigation of Competition in Digital Markets, 117-40, pp. 36–38.
[2] European Commission (2023), Data Act proposal, COM(2023) 193 final.
[3] Digital Platforms (2019) and UK Competition and Markets Authority (2019).
[4] Schmidt (2009), "How Google Plans to Stay Ahead in Search," Bloomberg, 2 October.
[5] Varian (2018) and Bajari et al. (2019).

covariates $\mathcal{P}$ to observe for the target individual (which characteristics to observe on the house whose price must be predicted).

Because collecting either type of data is costly, the dataset design choice reflects the trade-off between better predictions and higher data collection costs.

Given any dataset design, I derive the optimal Bayesian predictor and I characterize the value of the data, defined as the expected reduction of out-of-sample mean squared error achieved by observing the data. In this framework, each training covariate improves prediction by reducing uncertainty about its effect, while each prediction covariate allows the firm to apply what it has learned to the target individual. The key quantities are therefore: (i) how much the training data teach the firm about each covariate, and (ii) how much each prediction covariate matters for the outcome.

I show how these components combine into a closed-form expression for the value of any dataset design. The value rises when (i) training covariates reduce residual noise, (ii) additional observations sharpen parameter estimates, and (iii) prediction covariates allow the firm to apply the learned relationship between covariates and outcome to the target individual. I show that there are two regimes depending on whether observations are scarce or abundant.

In the observation-scarce regime, the same covariates are used for training and prediction, and covaraites have decreasing marginal returns. Furthermore, covariates and observations are complements.

In the observation-rich regime, the prediction covariates will be a subset of the training covariates (in machine-learning jargon, the firm will "train large deploy small"). In this regime two surprising findings:

First, *training covariates exhibit economies of scope*: adding a new training covariate is more valuable when many other covariates are already observed for the sample. Learning one relationship improves the firm's understanding of all the others because it reduces the residual noise in the regression. This implies there are increasing returns to training covariates.

Second, *training covariates and observations are complements when data are scarce, but become substitutes when data are abundant.* With small datasets, adding a covariate raises the marginal value of an observation; with large datasets, extra covariates mainly reduce residual noise, lowering the marginal value of further observations.

Third, *training data and prediction covariates are complements.* Any improvement in how well the model is trained raises the predictive value of each covariate observed on the target individual, generating positive spillovers from training data to prediction covariates.

I show that these findings generate policy-relevant implications in three applications. First, I study data-driven acquisitions. First, incumbents may acquire entrants even without product or user overlaps, and even when entry by the smaller firm would be socially valuable. Even when firms operate in independent prediction markets, an incumbent with a rich set of covariates (e.g., Google) optimally acquires an entrant with distinct covariates (e.g., Fitbit). This incentive arises purely from economies of scope in covariates: merging datasets

raises the marginal value of every covariate and creates a statistical analogue of a natural monopsony in data.

Second, I study data pools, a type of agreement between firms explicitly encouraged in the EU Data Act and in European common data-space initiatives. In the model, datasets can be fragmented either in observations (firms hold the same covariates on different users) or in covariates (firms hold different covariates on the same users). Pooling along these two dimensions has fundamentally different welfare effects. Pooling covariates on the same individuals combines complementary features and always generates procompetitive efficiencies by eliminating a Cournot-style double marginalization between data brokers. By contrast, pooling observations combines substitute datasets: when data are abundant, competition between brokers is already intense and pooling becomes effectively collusive.

Third, I apply the framework to licensing of training datasets in prediction markets, motivated by recent agreements such as the 2024 Reddit–OpenAI deal. A data seller owns a non-rival training dataset and may license it to two competing prediction firms, each of which can combine the licensed training data with its own proprietary covariates on target users. Licensed and proprietary data are strict complements in generating prediction accuracy, so sellers my raise the buyer's willingness to pay by charging a price so high the rival would not want to buy the license (de facto exclusivity). However, de facto exclusivity also discourages the excluded firm from investing in proprietary data. The model shows that de facto exclusivity is privately profitable whenever competition for prediction buyers is sufficiently fierce, yet it can be socially harmful in an intermediate region where both firms would invest under non-exclusive access, because it precludes investment by the excluded buyer.

Roadmap. Section 2 introduces the data-generating process and the firm's dataset-design problem. Section 3 derives the optimal predictor, highlighting the role of training information and the misspecification penalty. Section 4 analyzes the value of a dataset design, develops the core comparative statics on economies of scale and scope, and studies optimal covariate selection under data-collection costs. Section 5 applies the framework to three policy-relevant environments—data-driven acquisitions, data pools, and exclusive licensing. Section 6 concludes. Proofs are collected in Appendix A and extensions are in Appendix B.

Related Literature. There is a rich information-design literature on the value of data, starting with Bergemann, Bonatti, and Smolin (2018) and continuing with Jones and Tonetti (2020), Bergemann, Bonatti, and Gan (2022), Bergemann and Bonatti (2024), and Acemoğlu et al. (2022). These papers characterize how much an agent is willing to pay for more informative experiments by modeling the choice of an information structure. In contrast, my setting is more structured than a Blackwell experiment and allows me to distinguish observations, training covariates, and prediction covariates. This lets me map the value of data to transparent statistical objects (covariate variances, posterior variances, and a misspecification penalty)

and separate the roles of sample size, scope, and covariate selection. Methodologically, the paper is related to Montiel Olea et al. (2022), Iyer and Ke (2024), and Dasaratha, Ortner, and Zhu (2025), who study competition between models with different covariates; here, covariates and observations are modeled jointly and training and prediction data are separated, which generates structural non-convexities and complementarities in the value of data. Finally, Strzalecki (2024) interpret deviations from Bayesian updating via a primitive misspecification parameter; in my framework, an analogous distortion arises endogenously as standard Bayesian behavior under incomplete data collection, providing a microfoundation for likelihood discounting.

I also contribute to the broader literature on economies of scope and scale in data. Most models fix the set of covariates and study returns to observations, as in Bajari et al. (2019) and Goldfarb and Tucker (2011). By endogenizing which covariates are collected, my framework rationalizes empirical evidence on complementarities and economies of scope in Schaefer and Sapi (2023) and Carballa-Smichowski et al. (2025b). Schaefer (2025) develops a complementary frequentist approach and show that the eigenspectrum of the covariate matrix shapes returns to scale, in line with my comparative statics. Allcott et al. (2025) estimate diminishing returns to additional observations and limited complementarities across search queries. While Radner and Stiglitz (1984) attribute increasing returns to information costs, here non-convexities and increasing returns arise purely from the statistical structure of prediction, even under separable reduced-form data-collection costs. These results provide microfoundations for work in IO and macroeconomics that takes increasing returns to data as primitive, either via feedback between data and demand or by assuming complementarities across datasets (Hagiu and Wright 2023; Prüfer and Schottmüller 2021; Farboodi and Veldkamp 2025; Aral, Brynjolfsson, and Wu 2008; Cong, He, and Yu 2021; Carballa-Smichowski et al. 2025a; De Corniere and Taylor 2025; Calzolari, Cheysson, and Rovatti 2025): in my framework, economies of scale and scope emerge from posterior-variance reduction and the House Party Effect, without relying on demand-side feedback or exogenous complementarities.

The applications connect these statistical forces to the IO literature on digital markets, data pools, exclusive contracts, and scaling in machine learning. De Corniere and Taylor (2025) show that the pro- or anti-competitive effect of more data depends on the supply of data-driven services rather than demand, while Cornière and Taylor (2024) develop a theory of harm for data-driven mergers based on cross-market effects. Both treat data as an undifferentiated input; my analysis shows how their insights extend when "data" is unpacked into observations and distinct covariates. The pooling application relates to the patent-pool and broker literatures: Lerner and Tirole (2004) study the private and social value of commercializing complementary patents in pools, and Gu, Madio, and Reggiani (2021) analyze broker pools when datasets are substitutes. By endogenizing the statistical substitutability or complementarity of datasets, I microfound welfare implications that complement Nocke, Peitz, and Stahl (2007) on vertical integration and platform fragmentation. The analysis of exclusive data licenses builds on classic models in which exclusivity serves either to extract rents or to

deter entry by raising rivals' costs (Katz and Shapiro 1986; Aghion and Bolton 1987), adapting these insights to non-rival training datasets that are strict complements to proprietary covariates.

2 Model Setup

An agent (she) seeks to predict a target variable of a target individual (he). To do so, she can collect (i) training data, consisting of covariates and past target variables for a sample of individuals; and (ii) covariates for the target individual. The agent chooses which covariates to observe, trading off prediction accuracy and data collection costs.

2.1 Data-Generating Process

The agent must predict a *target variable* $y \in \mathbb{R}$ for a *target individual*. For each individual i, the target variable is generated by $Z < \infty$ observable covariates $\{x_k^i\}_{k=1}^Z$ according to

$$y^i = \sum_{k=1}^{Z} \beta_k x_k^i + u^i,$$

where $x_k^i \in \mathbb{R}$ is the realization of covariate k for individual i, and β_k is its (unknown) effect on y^i and u^i is an error term.

Parameters. Parameters are mutually independent and independent of covariates, with prior distribution

$$\beta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \tau^2\right),$$

where $\tau^2 \in [0, \infty)$ reflects the agent's prior uncertainty about the parameters.

Error. The error u^i is i.i.d. across individuals and such that

$$u^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2 \underline{\sigma}^2).$$

The scalar $\underline{\sigma}^2 \in [0, 1]$ denotes the irreducible fraction of variance that cannot be explained even when all covariates are observed.

Covariates. Covariates are mutually independent and i.i.d. across individuals, with prior distribution

$$x_k^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1 - \underline{\sigma}^2}{Z}\right).$$

The $1/Z$ scaling ensures that the total signal variance remains bounded as Z grows.

Under these conditions, $\text{Var}[y] = \tau^2 < \infty$, and the linear prediction problem is well-defined.

Notation For future reference, let $\bar{\mathcal{K}} \equiv \mathbb{N} \setminus \mathcal{K}$ denote the complement of a set $\mathcal{K}$, $\boldsymbol{x}^i \equiv (x_k^i)_{k \in \mathbb{N}}$ denote the covariate vector, and $\boldsymbol{\beta} \equiv (\beta_k)_{k \in \mathbb{N}}$ denote the parameter vector. Furthermore, for any vector $\boldsymbol{v}$ and set $\mathcal{K}$, let $\boldsymbol{v}_{\mathcal{K}} \equiv (v_k)_{k \in \mathcal{K}}$ denote the subvector of $\boldsymbol{v}$ with indices in $\mathcal{K}$.

2.2 Observed Data

The agent chooses three objects which determine the characteristics of the observed dataset:

- a ***sample size*** $N \in \mathbb{N}$ of individuals on which to observe the target variable and resulting in a ***target vector*** $\boldsymbol{y} \in \mathbb{R}^N$;
- a ***set of training covariates*** $\mathcal{T} \subset \mathbb{N}$ observed on the N individuals in the sample and resulting in a ***design matrix*** $\boldsymbol{X}_{\mathcal{T}} \in \mathbb{R}^{N \times |\mathcal{T}|}$; and
- a ***set of prediction covariates*** $\mathcal{P} \subset \mathbb{N}$ observed on the target individual and resulting in an ***input vector*** $\boldsymbol{x}_{\mathcal{P}} \in \mathbb{R}^{|\mathcal{P}|}$.

To summarize, the agent chooses a dataset design defined as follows:

Definition 1. A ***dataset design*** is a triple

$$\mathcal{D} \equiv (N, \mathcal{T}, \mathcal{P}),$$

where N is the sample size, $\mathcal{T}$ is the set of training covariates and $\mathcal{P}$ is the set of prediction covariates.

The choice of dataset design is the data collection strategy of the agent and reflects the choice of the experiments, sampling techniques and sensors which generate the data. We let the ***training data*** be the matrix

$$(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}}),$$

which for each individual (row) records the target variable and the training covariates (columns). We let a ***dataset*** be a collection of training data and target vector:

$$\boldsymbol{D} \equiv ((\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}}), \boldsymbol{x}_{\mathcal{P}}).$$

Notation. For any design $\mathcal{D}$, let $\mathbb{E}_{\mathcal{D}}[\cdot]$ denote expectation under the prior induced by $\mathcal{D}$ and the data-generating process in Section 2.1. Thus, for any integrable function h and any random object Z generated by the design $\mathcal{D}$, such as $\boldsymbol{x}_{\mathcal{P}}$, $(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})$ or $\boldsymbol{D}$, I write $\mathbb{E}_{\mathcal{D}}[h(Z)]$.

Predictor. The agent's predictions will depend on the realization of the dataset. A ***predictor*** is a measurable map $\hat{y} : \mathbb{R}^{N\times(1+|\mathcal{T}|)} \times \mathbb{R}^{|\mathcal{P}|} \to \mathbb{R}$ that maps a dataset $\boldsymbol{D}$ to a prediction

$$\boldsymbol{D} \mapsto \hat{y}(\boldsymbol{D}).$$

Equivalently, the prediction is the realization of a random variable which is a deterministic function of the dataset, i.e. the predictor. The choice of a predictor reflects how the agent plans to use the data it will observe to make a prediction.

2.3 Agent's Problem

Agent Utility Denoting the agent's prediction by $\hat{y} \in \mathbb{R}$, the realized prediction error is $y - \hat{y}$, and the agent suffers a quadratic loss so that

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

Therefore, the agent is penalized proportionally more for larger prediction errors. Because of its tractability, quadratic loss is a standard assumption in econometrics and machine learning. Furthermore it has desirable theoretical properties discussed in Brier (1950) and axiomatized in Selten (1998).[6] The agent will attempt to minimize her expected loss based on her knowledge on the data-generating process at the time of prediction. To increase her knowledge, she will choose a set of covariates and observe the resulting dataset.

The total cost of a design depends only on the size of the sample and the covariate sets, and is additively separable:

$$C(N, \mathcal{T}, \mathcal{P}) = C_n\left(\frac{N}{Z}\right) + C_t\left(\frac{|\mathcal{T}|}{Z}\right) + C_p\left(\frac{|\mathcal{P}|}{Z}\right).$$

This reduced form cost imposes a minimal structure on results, making the qualitative predictions of the model depend exclusively on the improvements in prediction accuracy stemming from data collection rather than from the function form of the cost of design. We assume that the choice of a predictor is costless.

Timing The agent's objective is to maximize expected utility minus a design cost. The predictor is chosen ex ante as a mapping from datasets to predictions. Therefore, the model unfolds in two steps:

1. ***Design Problem***: the agent chooses a design $\boldsymbol{D}$ and pays the associated cost;

[6] Namely, among all incentive-compatible scoring rules—that is, rules for which truthful probabilistic forecasts maximize expected scores—the quadratic rule is uniquely characterized (up to a positive linear transformation) by three axioms: symmetry, elongation invariance, and neutrality.

2. ***Prediction Problem***: the agent chooses a predictor $\hat{y}(\boldsymbol{D})$, nature draws a dataset $\boldsymbol{D}$ and an outcome y and losses occur.

As customary, we will solve the model backwards. I will give a broad overview of the problem the agent faces, and will defer the solution to Sections 3 and 4.

2.3.1 Prediction Problem

The agent will choose the predictor so that conditional on the realization of a dataset $\boldsymbol{D}$, minimizes her posterior risk, defined as the expected loss conditional on $\boldsymbol{D}$:

$$\rho_D(\hat{y}) \equiv \mathbb{E}_{y|D}\left[\ell(y, \hat{y}(\boldsymbol{D})) \mid \boldsymbol{D}\right].$$

Therefore, the agent chooses the ***optimal predictor***

$$\hat{y}^*(\boldsymbol{D}) \equiv \arg\min_{\hat{y}} \rho_D(\hat{y}).$$

For quadratic loss, $\hat{y}^*(\boldsymbol{D})$ exists and is unique. Having characterized the optimal predictor, we can define the value ***value of a dataset*** $\boldsymbol{D}$ as

$$v(\boldsymbol{D}) \equiv \underbrace{\rho_{\varnothing}\left(\hat{y}^*(\varnothing)\right)}_{\text{Minimum Prior Risk}} - \underbrace{\rho_D\left(\hat{y}^*(\boldsymbol{D})\right)}_{\text{Minimum Posterior Risk } |\boldsymbol{D}}.$$

The value of a dataset $\boldsymbol{D}$ is the reduction of risk which the agent achieves by observing it and using it optimally by predicting $\hat{y}^*(\boldsymbol{D})$, relative to the outside option of not collecting any data and making the ex-ante optimal prediction $\hat{y}^*(\varnothing)$.

2.3.2 Dataset Design Problem

The value of a dataset design is the ex-ante expected value of the dataset it induced.

Definition 2. The ***value of a design*** $\mathcal{D}$ is

$$V(\mathcal{D}) \equiv \mathbb{E}_{\mathcal{D}}\left[v(\boldsymbol{D})\right].$$

The value of a design $\mathcal{D}$ is the maximal increase in ex-ante expected utility that agent can achieve when using the optimal predictor $\hat{y}^*(\boldsymbol{D})$, given that design.

Maximizing the utility minus data collection costs amounts to maximizing the value of a design minus its costs. Therefore, the agent will solve

$$\max_{\mathcal{D}} \left\{V(\mathcal{D}) - C(\mathcal{D})\right\},$$

which, I will show, always has a solution.

3 Prediction

In this section we characterize the Bayesian optimal predictor. A Bayesian agent chooses how to use the input vector $\boldsymbol{x}_{\mathcal{P}}$ of the target individual and the training data $(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})$ to form a prediction. I will show that learning can be separated into two forms of learning:

1. ***Across-individual learning*** about the parameters $\boldsymbol{\beta}_{\mathcal{T}}$ using the training data;
2. ***Within-individual learning*** using the realized input vector $\boldsymbol{x}_{\mathcal{P}}$ of the target individual.

Fix a training set $\mathcal{T}$. Conditional on $\mathcal{T}$, the training data satisfy the regression

$$y^i = \sum_{k \in \mathcal{T}} \beta_k x_k^i + \varepsilon_{\mathcal{T}}^i, \qquad \varepsilon_{\mathcal{T}}^i \equiv u^i + \sum_{k \notin \mathcal{T}} \beta_k x_k^i.$$

This equation is a reduced-form projection: the residual $\varepsilon_{\mathcal{T}}^i$ collects both irreducible noise and the contribution of omitted covariates.

I now show that, in a high-dimensional regime, the regression residual is approximately Gaussian. For each $Z \geq 1$, consider a model with Z potential covariates, as in Section 2.1, and a training set $\mathcal{T}(Z) \subseteq \{1, \dots, Z\}$, a prediction set $\mathcal{P}(Z) \subseteq \{1, \dots, Z\}$.

Assumption 1 (Proportional-growth regime)**.** *Fix constants $(n, t, p) \in [0, \infty) \times [0, 1]^2$. We consider a sequence of designs $(N(Z), \mathcal{T}(Z), \mathcal{P}(Z))_{Z \geq 1}$ such that*

$$n = \lim_{Z \to \infty} \frac{N(Z)}{Z}, \qquad t = \lim_{Z \to \infty} \frac{|\mathcal{T}(Z)|}{Z}, \qquad p = \lim_{Z \to \infty} \frac{|\mathcal{P}(Z)|}{Z}.$$

This assumption is reasonable because modern machine learning operates in regimes with both many observations and many features; proportional-growth asymptotics are widely used to model such settings as in Hastie et al. (2020).

We now define the relative regression noise which is the fraction of the overall variance of y which is due to covariates not included in $\mathcal{T}$.

Definition 3 (Relative residual noise)**.** For each $Z \geq 1$, define the relative regression noise induced by the training set $\mathcal{T}(Z)$ as $Z \to \infty$ as

$$\sigma^2(t) \equiv \lim_{Z \to \infty} \frac{\mathrm{Var}\left[\varepsilon_{\mathcal{T}(Z)}^i \mid \mathcal{T}(Z)\right]}{\tau^2} \in [\underline{\sigma}^2, 1].$$

The following result shows that the residual $\varepsilon_{\mathcal{T}(Z)}^i$ is distributed as a Gaussian random variable when Z is large.

Lemma 1. *Under Assumption 1,*

$$\sigma^2(t) = \underline{\sigma}^2 + (1 - t)(1 - \underline{\sigma}^2) \in [\underline{\sigma}^2, 1].$$

Furthermore, conditional on $\mathcal{T}(Z)$, as $Z \to \infty$

$$\varepsilon^i_{\mathcal{T}(Z)} \Longrightarrow \mathcal{N}(0, \tau^2\sigma^2(t).)$$

The relative regression noise $\sigma^2(t)$ is the sum of the irreducible noise $\underline{\sigma}^2$ and a misspecification noise $(1-t)(1-\underline{\sigma}^2)$, which is decreasing in the fraction of covariates included in the training set. By collecting additional covariates, the firm can therefore reduce its misspecification noise down to the lower bound $\underline{\sigma}^2$.

Lemma 1 shows that the residual converges in distribution to a Gaussian random variable when it is made up of numerous covariates due to the Lindeberg central limit theorem. This asymptotic Gaussianity allows us to apply standard Bayesian regression tools as an approximation in high-dimensional environments, following DeGroot (2005).

In this section we fix the training covariate set $\mathcal{T}$, so we write $\sigma^2 = \sigma^2(t)$.

3.1 The Optimal Predictor

The agent chooses a predictor, which specifies how to use any realized dataset $\boldsymbol{D}$ to make a prediction. Under quadratic loss, the optimal predictor is the posterior mean of the target variable, a well-known Bayesian result that motivates the following result.

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

$$\hat{y}^*(\boldsymbol{D}) = \mathbb{E}\left[y \mid \boldsymbol{D}\right] = \boldsymbol{x}'_{\mathcal{P}\cap\mathcal{T}} \underbrace{\left(\left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} + \sigma^2 \cdot \boldsymbol{I}_{|\mathcal{T}|}\right)^{-1} \boldsymbol{X}'_{\mathcal{T}}\boldsymbol{y}\right)_{\mathcal{P}\cap\mathcal{T}}}_{=\mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}}|(\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})]}.$$

Because the target variables $\{y^i\}$ are independent across individuals conditional on $\boldsymbol{\beta}$, the conditional expectation $\mathbb{E}[y \mid \boldsymbol{D}]$ depends on the training data $(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})$ only through the posterior mean of the coefficients of the prediction covariates, $\mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}} \mid (\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})]$. Prediction covariates that are not observed in the training sample never enter this posterior mean: their coefficients are not updated and remain at their prior mean, which is zero, so they do not affect the optimal predictor.

The posterior mean of the parameters shrinks the OLS estimator

$$\left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}}\right)^{-1} \boldsymbol{X}'_{\mathcal{T}}\boldsymbol{y}$$

towards the prior mean $\boldsymbol{0}_{|\mathcal{T}|}$. The strength of this shrinkage is governed by σ^2, which captures the noise arising from unobserved covariates through the unidentified parameters $\boldsymbol{\beta}_{\mathcal{T}}$. Intuitively, if $\sigma^2 \approx 0$, the coefficients $\boldsymbol{\beta}_{\mathcal{T}}$ are well identified, so the posterior mean coincides with the maximum-likelihood estimator, namely the OLS estimate. As σ^2 increases, the agent becomes more cautious in using the training data, and the posterior mean shrinks more strongly towards the prior mean.

Remark 1. The noise σ^2 is equivalent to the misspecification parameter of Strzalecki (2024). In contrast to his model, where the misspecification parameter is exogenous, my framework endogenizes σ^2 as the weight a Bayesian agent assigns to the prior because of the omitted covariates and the nuisance parameters. The agent's implicit discounting of the likelihood reflected in the shrinkage away from the OLS thus emerges endogenously from the choice of covariates, rather than exogenously because of a preference for robustness as in Strzalecki (2024).

3.2 Value of a dataset

Given the loss is quadratic, Proposition 1 implies that the value of a dataset is the reduction in the variance of the target variable brought forth by its observation, a result which allows us to state the following result.

Proposition 2. *Value of a dataset* $\boldsymbol{D}$ *is*

$$v(\boldsymbol{D};\sigma^2) = \operatorname{Var}[y] - \operatorname{Var}[y \mid \boldsymbol{D};\sigma^2] = \boldsymbol{x}'_{\mathcal{P}\cap\mathcal{T}} \underbrace{\left(\tau^2 \cdot \boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} \left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} + \sigma^2 \cdot \boldsymbol{I}_{|\mathcal{T}|}\right)^{-1}\right)_{\mathcal{P}\cap\mathcal{T}}}_{= \operatorname{Var}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}}] - \operatorname{Var}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}} \mid (\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})]} \boldsymbol{x}_{\mathcal{P}\cap\mathcal{T}}.$$

The value of a dataset is driven by how much it shrinks uncertainty about the parameters of the observed covariates. The braced matrix is the reduction in the covariance of $\boldsymbol{\beta}_{\mathcal{T}}$ induced by observing $(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})$. The quadratic form $\boldsymbol{x}'_{\mathcal{P}\cap\mathcal{T}}(\cdot)\boldsymbol{x}_{\mathcal{P}\cap\mathcal{T}}$ then maps this reduction in parameter uncertainty into a reduction in the uncertainty of y for the specific individual under consideration. Intuitively, the dataset is most valuable when the input vector $\boldsymbol{x}_{\mathcal{P}\cap\mathcal{T}}$ loads heavily on covariates whose parameters are estimated most precisely from the training data, and its value vanishes when either $\boldsymbol{x}_{\mathcal{P}\cap\mathcal{T}} = \boldsymbol{0}$ or the posterior coincides with the prior.

Note that, for any fixed realized dataset $\boldsymbol{D}$, its value is decreasing and strictly convex in the noise σ^2. Thus, the same dataset is more valuable if the prior attributes a smaller share of the total variance of y to the unobserved covariates rather than to the observed covariates. Moreover, convexity implies that the marginal value of decreasing the noise is larger when it is already low: reallocating a bit more prior variance from the omitted covariates to the observed covariates is especially valuable when the unobserved covariates are already believed to account for a small fraction of the signal. This feature will be a key driver of the increasing returns to covariates that we explore in the next section.

4 Design Choice

We now study the optimal design choice. To do so, we first characterize how informative a design is about the parameters, and then translate this into the value of a covariate set and the marginal value of individual covariates.

4.1 Training Information

We first study how informative the training data $(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})$ is about the parameter vector $\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}}$. In this subsection we fix the relative residual variance $\sigma^2 > 0$, meaning that changes in $\mathcal{T}$ will not affect the regression residual but only the degrees of freedom of the training regression.

Definition 4. Fix $\sigma^2 > 0$. The ***training information*** is

$$\mathcal{I}(N, \mathcal{T}; \sigma^2) \equiv \frac{1}{|\mathcal{T}|}\mathrm{tr}\left[\mathrm{Var}(\boldsymbol{\beta}_{\mathcal{T}}) - \mathbb{E}_{N,\mathcal{T}}\left[\mathrm{Var}\left(\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}}; \sigma^2\right)\right]\right].$$

The expectation $\mathbb{E}_{N,\mathcal{T}}$ is taken under the data-generating process induced by sample size N, covariate set $\mathcal{T}$, and relative noise level σ^2 (equivalently, the posterior variance is computed under that same model).

Intuitively, $\mathcal{I}$ is the expected reduction of uncertainty about the parameters of training covariates on a design with N observations and covariate set $\mathcal{T}$. Equivalently, $\mathcal{I}$ measures the average fraction of prior variance about each trained coefficient that is eliminated by observing the training sample.

The following result characterizes the training information.

Lemma 2 (Information on a trained parameter)**.** *Fix $\sigma^2 > 0$. Under the assumptions of Section 2.1,*

$$\mathcal{I}(N, \mathcal{T}; \sigma^2) = \frac{\tau^2}{|\mathcal{T}|} \times \mathrm{tr}\left[\mathbb{E}_{N,\mathcal{T};\sigma^2}\left[\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}\left(\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}} + \sigma^2 \cdot \boldsymbol{I}_{|\mathcal{T}|}\right)^{-1}\right]\right] \in [0, \tau^2]$$

Lemma 2 expresses information as an expected shrinkage factor that depends on the design matrix $\boldsymbol{X}_{\mathcal{T}}$ and on the relative noise level σ^2. Throughout this subsection we treat $\sigma^2 > 0$ as fixed: varying $\mathcal{T}$ changes the degrees of freedom and the geometry of the regression, but not the residual variance. This isolates the purely statistical effect of the training design on parameter learning. In the next subsection we endogenize σ^2 through the choice of observed covariates and translate parameter information into the economic value of a design.

Proposition 3 (High-dimensional information)**.** *Under Assumption 1,*

$$\lim_{Z\to\infty} \mathcal{I}(N(Z), \mathcal{T}(Z); \sigma^2) = \mathcal{I}^{\mathrm{hd}}(n, t; \sigma^2),$$

*where $\mathcal{I}^{\mathrm{hd}}(n, t; \sigma^2) \in (0, 1)$ is the **high-dimensional information** defined by*

$$\mathcal{I}^{\mathrm{hd}}(n, t; \sigma^2) \equiv \frac{\sigma^2 + n + t}{2t}\left(1 - \sqrt{1 - \frac{4nt}{(\sigma^2 + n + t)^2}}\right),$$

and, equivalently, as the unique solution to the equation

$$\mathcal{I}^{\text{hd}} = \frac{1}{1 + \underbrace{\dfrac{\sigma^2 + t(1 - \mathcal{I}^{\text{hd}})}{n}}_{\textit{noise-to-signal ratio}}}.$$

The fixed-point formulation makes clear that $\mathcal{I}^{\text{hd}}$ can be interpreted as a signal share in the training problem. The high-dimensional information is the fraction of total variation in the training regression that is attributable to true signal rather than noise: the signal term is the sample size n, which captures the amount of across-individual variation. The noise term has two components. First, σ^2 is the residual noise. Second, $t(1 - \mathcal{I}^{\text{hd}})$ is the estimation noise due to the dimensionality of the training problem: among the t training covariates, only a fraction $\mathcal{I}^{\text{hd}}$ is effectively learned, while the remaining share $1 - \mathcal{I}^{\text{hd}}$ is noise. Because estimation noise itself depends on how informative the data are, the signal share must satisfy a self-consistency condition, which gives rise to the fixed point.

In this sense, $\mathcal{I}^{\text{hd}}$ summarizes a balance between across-individual learning (through n) and within-individual complexity (through t), with higher dimensionality endogenously increasing noise unless it is offset by a reduction of the residual noise.

We now define the ***effective information*** as

$$\mathcal{I}^{\text{eff}}(n, t) \equiv \mathcal{I}^{\text{hd}}(n, t; \sigma^2(t)),$$

where $\sigma^2(t) = \underline{\sigma}^2 + (1 - \underline{\sigma}^2)(1 - t)$.

Corollary 1 (Marginal Information of Observations)**.**

$$\frac{\partial}{\partial n}\mathcal{I}^{\text{eff}}(n, t) > 0, \qquad \frac{\partial^2}{\partial n^2}\mathcal{I}^{\text{eff}}(n, t) < 0,$$

Information is increasing and concave in the sample size n: additional observations improve learning, but at diminishing marginal returns, consistent with the Law of Large Numbers. This is because if we keep the covariate set fixed, we are collecting more samples of a process with an identical distribution.

Corollary 2 (Marginal Information of Covariates)**.**

$$\frac{\partial}{\partial t}\mathcal{I}^{\text{eff}}(n, t) = \underbrace{\frac{\partial \mathcal{I}^{\text{hd}}}{\partial t}\Big(n, t; \sigma^2(t)\Big)}_{<0\,:\,\textit{dimensionality}} + \underbrace{\frac{\partial \mathcal{I}^{\text{hd}}}{\partial \sigma^2}\Big(n, t; \sigma^2(t)\Big) \cdot \sigma_t^2(t)}_{>0\,:\,\textit{residual noise reduction}} > 0 \iff \mathcal{I}^{\text{eff}}(n, t) > \underline{\sigma}^2 \iff n \geq \frac{\underline{\sigma}^2}{1 - \underline{\sigma}^2}.$$

Furthermore, if $\frac{\partial}{\partial t}\mathcal{I}^{\text{eff}}(n,t) > 0$*, then*

$$\frac{\partial^2}{\partial t^2}\mathcal{I}^{\text{eff}}(n,t) = \underbrace{\frac{\partial^2 \mathcal{I}^{\text{hd}}}{\partial t^2}\left(n,t;\sigma^2(t)\right)}_{\textit{direct (dimensionality)}} + 2\underbrace{\frac{\partial^2 \mathcal{I}^{\text{hd}}}{\partial t\,\partial \sigma^2}\left(n,t;\sigma^2(t)\right)}_{\textit{interaction}}\sigma_t^2(t) + \underbrace{\frac{\partial^2 \mathcal{I}^{\text{hd}}}{\partial (\sigma^2)^2}\left(n,t;\sigma^2(t)\right)}_{>0:\,\textit{curvature in noise}}\left(\sigma_t^2(t)\right)^2 > 0$$

Increasing the number of covariates has two opposing effects on the information: on the one hand, it has a negative effect because, holding constant the residual noise σ^2 it increases the dimensionality and consequently the number of parameters which the estimation depends on; on the other, it has a positive effect because, holding constant the dimensionality t, it reduces the residual noise by $\sigma_t^2(t)$. The proposition shows that the latter effect dominates if $\mathcal{I}^{\text{eff}}(n,t) > \underline{\sigma}^2$ which is equivalent to

$$n \geq \tilde{n} \equiv \frac{\underline{\sigma}^2}{1-\underline{\sigma}^2}.$$

The threshold does not depend on the dimensionality t because the latter affects both the marginal cost and the marginal benefit of adding covariates, so it cancels out of the sign condition. What matters is whether the sample is large enough for reductions in residual noise to dominate the curse of dimensionality.

This implies that there are two regimes: in the information-poor regime, $n \leq \tilde{n}$, larger models are less precise; in the observation-rich regime $n \geq \tilde{n}$ larger models are more precise. If the agent has a large enough number of observations it can switch to a regime where increases in t generate additional information.

Furthermore, in the observation-rich regime, there are increasing returns to t: the intuition is that there is a convexity in σ^2 because reductions in residual noise have increasing positive impact on the information.

The following result studies the interaction of the marginal information of observations and covariates in the observation-rich regime.

Corollary 3 (Interaction of Observations and Covariates)**.** *If* $\mathcal{I}_t^{eff}(n,t) > 0$*, then*

$$\frac{\partial^2}{\partial n \partial t}\mathcal{I}^{\text{eff}}(n,t) = \frac{\mathcal{I}_t^{eff}(n,t)}{n}(1 - 2\mathcal{I}^{\text{eff}}(n,t)) > 0 \iff \mathcal{I}^{\text{eff}}(n,t) < \frac{1}{2}$$

In the observation-rich regime, covariates and observations are complements when $\mathcal{I}^{\text{eff}}(n,t) < \frac{1}{2}$ but become substitutes when $\mathcal{I}^{\text{eff}}(n,t) > \frac{1}{2}$.

4.2 The Value of a Design

I will now characterize the value of a dataset using the training information and the fact that the residual noise is endogenous and depends on the set of training covariates since $\sigma^2 = \sigma^2(\mathcal{T}(Z))$.

The following result, uses the training information matrix to characterize the value of a dataset design $(N(Z), \mathcal{T}(Z), \mathcal{P}(Z))$.

Proposition 4 (Low-dimensional Value of Design)**.** *Fix $Z \geq 1$ and a dataset design $\mathcal{D}(Z) = (N(Z), \mathcal{T}(Z), \mathcal{P}(Z))$ with $\mathcal{P}(Z) \subseteq \mathcal{T}(Z)$. Then the value of the design satisfies*

$$V(\mathcal{D}(Z)) = \underbrace{|\mathcal{P}(Z) \cap \mathcal{T}(Z)| \times \frac{1-\underline{\sigma}^2}{Z}}_{\textit{Within-individual signal}} \times \underbrace{\mathcal{I}\left(N(Z), \mathcal{T}(Z); \sigma^2(\mathcal{T}(Z))\right)}_{\textit{Across-individual information}},$$

where $\sigma^2(\mathcal{T}(Z))$ is the relative regression noise induced by the training set.

Intuitively, $|\mathcal{P}(Z) \cap \mathcal{T}(Z)| \times \frac{1-\underline{\sigma}^2}{Z}$ measures how much of the predictable variation in the target outcome is spanned by the prediction covariates which have also been trained on. The term $\mathcal{I}(N(Z), \mathcal{T}(Z); \sigma^2)$ measures how well the trained coefficients are learned from the training sample.

Since τ^2 enters multiplicatively in $\mathcal{I}(\cdot)$, we normalize $\tau^2 = 1$ for the rest of the section.

Corollary 4 (Optimal overlap under cardinality costs)**.** *For any $Z \geq 1$, any optimal design $\mathcal{D}(Z)$ with $V(\mathcal{D}(Z)) > 0$ satisfies*

$$|\mathcal{P}(Z) \cap \mathcal{T}(Z)| = \min\{|\mathcal{P}(Z)|, |\mathcal{T}(Z)|\}.$$

The key observation is that the optimal predictor uses only covariates in $\mathcal{P}(Z) \cap \mathcal{T}(Z)$, since coefficients of covariates not observed in the training sample are not updated and remain at their prior mean. Because costs depend only on cardinalities, the agent can relabel covariates without changing costs and thus choose $\mathcal{P}(Z)$ to overlap as much as possible with $\mathcal{T}(Z)$. Therefore, without loss of generality, an optimal design maximizes overlap.

Using Corollary 4, we now characterize the high-dimensional limit of $V(\mathcal{D}(Z))$, which has a closed form as it is the expectation of a random matrix with a Marchenko-Pastur distribution.

Theorem 1 (High-dimensional Value of Design)**.** *Under Assumption 1,*

$$\lim_{Z \to \infty} V(\mathcal{D}(Z)) = \min\{p, t\}(1 - \underline{\sigma}^2)\, \mathcal{I}^{\text{hd}}(n, t; \sigma^2(t)).$$

Define

$$\bar{V}(n, t, p) \equiv \lim_{Z \to \infty} V(\mathcal{D}(Z)).$$

The value of a dataset design is the product of the fraction of covariates of the target individual which are used to make predictions $\min\{p, t\}$ with the fraction of variance that can be explained by the parameters $(1 - \underline{\sigma}^2)$ and the fraction of that variance which is learned $\mathcal{I}^{\text{hd}}(n, t; \sigma^2(t))$.

At this point, I can analyze the returns to data by differentiation.

Marginal Value of Observations. I first analyze the value of increasing the sample size by differentiating the value of a dataset design with respect to n.

Corollary 5 (Returns to Observations)**.** *The value of additional observations is positive and concave:*

$$\frac{\partial}{\partial n}\bar{V}(n,t,p) > 0 \qquad \textit{and} \qquad \frac{\partial^2}{\partial n^2}\bar{V}(n,t,p) < 0.$$

This follows directly from the fact the information is increasing and concave in n due to the Law of Large Numbers: as the sample grows, posterior uncertainty falls, but each additional observation reduces uncertainty by less than the previous one. The prediction value therefore exhibits decreasing marginal returns in n, consistent with empirical evidence on diminishing returns to data in settings with fixed covariate sets (e.g., Goldfarb and Tucker (2011), Bajari et al. (2019), and Schaefer and Sapi (2023)).

This curvature has an immediate implication for data allocation. When covariate sets are fixed, concavity in n implies a tendency toward decentralization of observations: the marginal value of one more observation is higher for an agent with a smaller dataset. In a simple acquisition environment (e.g., a second-price auction for an additional observation), the smaller-data agent bids more and is more likely to acquire the observation, reducing asymmetries in sample size over time. In this sense, diminishing returns in n make extreme concentration of observations harder to sustain absent additional forces (such as complementarities across covariates, scale economies in data collection, or strategic barriers).

Marginal Returns to Training Covariates. I then analyze the value of increasing the number training covariates by differentiating the value of a dataset design with respect to t.

Corollary 6 (Value of Training Covariates)**.** *There are two regimes:*

1. *If* $n \leq \frac{\underline{\sigma}^2}{1-\underline{\sigma}^2}$ *(observation-poor regime),*

$$\frac{\partial}{\partial t}\bar{V}(n,t,p) > 0 \iff t \leq p, \quad \frac{\partial^2}{\partial t^2}\bar{V}(n,t,p) < 0.$$

2. *If* $n > \frac{\underline{\sigma}^2}{1-\underline{\sigma}^2}$ *(observation-rich regime),*

$$\frac{\partial}{\partial t}\bar{V}(n,t,p) > 0, \quad \frac{\partial^2}{\partial t^2}\bar{V}(n,t,p) > 0.$$

There are two regimes: in the observation-poor regime, collecting training covariates is useful because it allows the agent to estimate the parameters of the prediction covariates. Furthermore, there are decreasing returns to covariates: the intuition is that there is an increasing dimensionality cost because as more parameters are estimated the precision with which each one is estimated decreases.

In the observation-rich regime, the agent might find it optimal to collect some training covariates even if they are not observed for the target individual. This rationalizes the machine-learning practice of "train large, deploy small": often algorithms estimate models on rich sets of covariates but rely on a smaller set of covariates for prediction. The proposition implies that such architectures are attractive only when n is large enough.

Furthermore, in these "train large, deploy small" regimes, the value is increasing and convex in t because the information itself is increasing and convex in t. Therefore, there are increasing marginal returns to covariates, consistently with the empirical findings in Carballa-Smichowski et al. (2025b). This finding suggests that data can generate entry barriers in observation-rich regimes. It also provides a mechanism for the emergence of ecosystems: if distinct services (e.g., email, maps, social media) generate distinct covariates, increasing returns in t imply economies of scope across services, strengthening incentives for firms to expand into multi-service ecosystems.

Marginal Value of Prediction Covariates. I then analyze the value of increasing the number prediction covariates by differentiating the value of a dataset design with respect to p.

Corollary 7 (Marginal Value of Prediction Covariates)**.** *The value of training covariates satisfies:*

$$\frac{\partial}{\partial p}\bar{V}(n,t,p) > 0 \iff t > p, \quad \frac{\partial^2}{\partial p^2}\bar{V}(n,t,p) = 0.$$

It is useful to collect additional prediction covariates only if they are also trained on. Furthermore, there are no spillovers across prediction covariates, so they have constant returns.

The following result highlights that prediction covariates and training covariates and observations are complements:

Corollary 8 (Complementarity Prediction-Training)**.** *The value of training covariates satisfies:*

$$\frac{\partial^2}{\partial t \partial p}\bar{V}(n,t,p) > 0 \iff p < t \wedge n > \tilde{n}, \quad \frac{\partial^2}{\partial t \partial p}\bar{V}(n,t,p) = 0 \iff p > t.$$

and

$$\frac{\partial^2}{\partial p \partial n}\bar{V}(n,t,p) > 0 \iff t > p.$$

Whenever $p < t$, increasing p allows the agent to use information on additional parameters.

The following result analyzes the interaction of training covariates and observations.

Corollary 9 (Complements and Substitutes in Training)**.** *Suppose $\mathcal{I}_t^{eff}(n,t) > 0$ (equivalently $\mathcal{I}^{\text{eff}}(n,t) > \underline{\sigma}^2$), then*

$$\bar{V}_{nt}(n,t,p) = \begin{cases} (1-\underline{\sigma}^2)\, p\, I_{nt}(n,t), & t > p, \\ (1-\underline{\sigma}^2)\big(I_n(n,t) + tI_{nt}(n,t)\big), & t < p. \end{cases}$$

In particular,

$$\bar{V}_{nt}(n,t,p) > 0 \text{ if } \mathcal{I}^{\text{eff}}(n,t) < \frac{1}{2},$$

whereas for $\mathcal{I}^{\text{eff}}(n,t) > \frac{1}{2}$ *the sign is negative when* $t > p$ *and otherwise ambiguous.*

We will say that the agent is information-rich if $\mathcal{I}^{\text{eff}}(n,t) > \frac{1}{2}$ and information poor otherwise. The result can be summarized in the following figure.

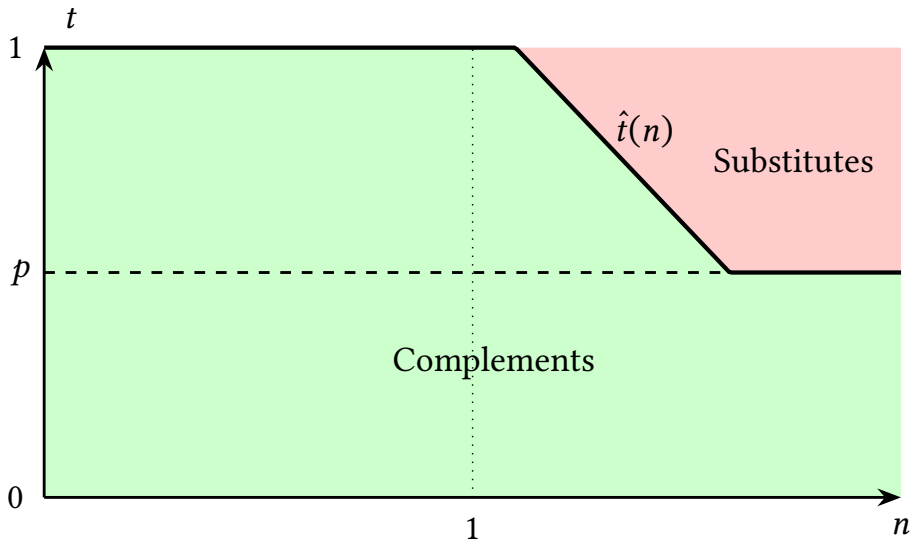

Figure 1: Covariates and observations are complements for all t below the decreasing threshold $t = \hat{t}(n, \underline{\sigma}^2)$, where additional observations raise the marginal value of training covariates. Above this curve, the two inputs become substitutes, as further observations reduce the marginal value of additional training covariates.

Therefore when the models are underparametrized (i.e. $t < n$) and $t > p$, observations and covariates can be substitutes. This implies that the complementarity between covariates and observations empirically documented in Schaefer and Sapi (2023) and Lee and Wright (2023) may not hold when datasets are large.

4.3 Comparative Statics

Assuming an interior solution, comparative statics are governed by the patterns of complementarity and substitutability characterized in this section.

Regime selection. The cost of observations $C_n(n)$ determines whether the agent operates in the observation-poor or observation-rich regime. If observation costs are high, the optimal sample size satisfies

$$n^* < \tilde{n} \equiv \frac{\underline{\sigma}^2}{1 - \underline{\sigma}^2},$$

so the agent is in the observation-poor regime. In this case, the optimal design satisfies $t^* = p^*$: training scope coincides with prediction scope. Moreover, observations and covariates are complements. A reduction in the cost of observations reduces n^* and, through complementarity, increases $t^* = p^*$. Symmetrically, a reduction in the cost of covariates reduces $t^* = p^*$ and increases the optimal sample size n^*.

If observation costs are sufficiently low so that

$$n^* > \tilde{n},$$

the agent operates in the observation-rich regime. In this regime, provided the marginal cost of training covariates is small relative to the marginal cost of prediction covariates, the optimal design features a "train large, deploy small" strategy with $t^* > p^*$. Furthermore, the interaction between training covariates and observations is state-dependent: when both t and n are small, they are complements, while when both are sufficiently large—specifically when $\mathcal{I}^{\text{eff}}(n, t) > \frac{1}{2}$—they become substitutes. As a result, reductions in observation costs can lead to a non-monotonic response of training scope: as C_n falls, t^* initially increases but may eventually decrease once the system enters the substitution region.

Cost shocks. An increase in the cost of prediction covariates $C_p(\cdot)$ reduces the optimal number of prediction covariates. Because prediction and training covariates are complements, this also lowers the optimal training scope and, through complementarity, reduces the optimal sample size.

An increase in the cost of training covariates $C_t(\cdot)$ reduces the optimal training scope. Through the optimality condition linking training and prediction scope, this also reduces the number of prediction covariates and increases the marginal variance of the last prediction covariate deployed. The response of observations depends on the regime. Firms operating with small t and n, where training covariates and observations are complementary, optimally reduce both margins. By contrast, firms operating with large t and n, where the two inputs are substitutes, reduce training scope but partially offset the cost increase by expanding their collection of observations.

An increase in the cost of observations $C_n(\cdot)$ has analogous effects. It reduces the optimal number of observations and, via complementarity, lowers prediction scope. The response of training covariates again depends on the regime: information-poor firms cut both observations and training covariates, while information-rich firms substitute away from observations toward relatively more training covariates.

Implications. Increases in data collection costs—whether through $C_n(\cdot)$ or $C_t(\cdot)$—amplify disparities between information-rich and information-poor agents. information-poor agents contract on all margins simultaneously, while observation-rich agents reallocate across mar-

gins in response to substitution, thereby reinforcing their informational advantage compared to information-poor agents.

5 Applications

This section applies the model in an observation-rich environment in which $n \gg t$, so that estimation noise is negligible. Accordingly, we work with the following approximation to the value of a dataset design.

Lemma 3 (Observation-rich approximation for value)**.** *Fix $p \in (0, 1]$ and suppose that $t/n \to 0$. Then, with endogenous residual variance $\sigma^2(t) = 1 - t(1 - \underline{\sigma}^2)$,*

$$\bar{V}(n, t, p) = (1 - \underline{\sigma}^2) \frac{\min\{t, p\}}{1 + \sigma^2(t)/n} + o(1). \tag{1}$$

Throughout the applications, we restrict attention to balanced dataset designs in which the training and prediction covariates coincide. That is, we assume $\mathcal{T} = \mathcal{P} = \mathcal{K}$, and define

$$k \equiv \lim_{Z \to \infty} \frac{|\mathcal{K}(Z)|}{Z},$$

so that $t = p = k$. Under this restriction, the value of a dataset design is given by $\bar{V}(n, k)$.

5.1 Natural Monopsony and Data-driven Acquisitions

In this application we show how a "natural monopsony" in data can emerge: a firm with more covariates has a strictly increasing and convex advantage in acquiring additional covariates. This leads to inevitable acquisitions even when products, users, and markets do not overlap. Mergers may be anti-competitive even when the merging firms do not compete for users, attention, advertising, or any traditional IO market.

5.1.1 Setup

Demand. For each prediction seller there is a unit mass of captive prediction buyers.[7] Each buyer must choose an action $\hat{y} \in \mathbb{R}$, while the payoff-relevant state is $y \in \mathbb{R}$. Payoffs depend on the squared error $(\hat{y} - y)^2$. Buyers share a common prior on y as specified in Section 2. If they do not buy any prediction, they choose the prior mean and obtain expected utility

$$\bar{u} = -1,$$

because of the normalization $\mathrm{Var}(y) = 1$.

[7] I assume there is no downstream competition to abstract from the usual "killer acquisition" motive: acquisitions are solely motivated by the acquisition of data.

If a buyer purchases from seller i at price p_i, her *net* surplus relative to the outside option is

$$u(p_i, V_i) = V_i - p_i,$$

where $V_i \in [0, 1]$ is the value of the dataset design of seller i, which is equivalent to the buyers' willingness-to-pay for seller i's prediction: each buyer strictly prefers i whenever $p_i < V_i$. We summarize individual and competitive demand by a reduced-form function $D(p_i, V_i)$,which is increasing in V_i and decreasing in p_i.

Supply. There are two potential sellers: an incumbent I ("Big Tech") and an entrant E ("Fitbit"). Both observe the same sample size n. Seller $i \in \{I, E\}$ has access to a set $\mathcal{K}_i$ of training and prediction covariates, with disjoint covariate sets $\mathcal{K}_I \cap \mathcal{K}_E = \varnothing$. To obtain closed-form expressions, assume all covariates have identical informativeness and impose symmetry by letting

$$|\mathcal{K}_I| = |\mathcal{K}_E| = k/2,$$

so that firms are symmetric.

The standalone value of seller i's design is $V_1 = V(n, \mathcal{K}_i)$.

The entrant E can either: (i) remain active and sell its own predictions, or (ii) accept a take-it-or-leave-it acquisition offer P from I. If I acquires E, it gains access to E's covariates. The merged firm operates with sample size n and combined covariate sets, so post-merger design value is

$$V_2 = V(n, \mathcal{K}_I \cup \mathcal{K}_E).$$

Profits. Let $D(p, V)$ denote the demand faced by a monopolist offering a prediction with design value V at price p. There are two possible cases

- *If I acquires E:* Only the merged seller operates. Setting price p_2, the incumbent's profit is

  $$\Pi_I^{\text{acq}}(P, p_2) = p_2 D\big(p_2, V_2\big) - P,$$

  while the entrant receives the acquisition payment

  $$\Pi_E^{\text{acq}}(P) = P.$$

- *If there is no acquisition:* Both firms are active. As demands are not related because each firm has a unit mass of captive customers, let $D_i(p_i, V_i)$ be seller i's demands. Profits are

  $$\Pi_i^{\text{no}}(p_i) = p_i D_i(p_i, V_i).$$

Planner welfare. The planner's objective is

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

where CS is customer surplus and $\xi \geq 0$ is the social cost of reduced entry. Intuitively, ξ reflects positive spillovers, e.g., new data, experimentation on alternative prediction tasks, technological diversity, knowledge spillovers.

Timing. The game unfolds in two stages:

1. **Acquisition stage.** The incumbent I offers P to acquire E. The entrant accepts or rejects.
2. **Pricing stage.** If the acquisition occurs, the merged firm chooses p_2. Otherwise, I and E simultaneously set (p_I, p_E).

Payoffs are realized at the end. The solution concept is subgame perfect Nash equilibrium, obtained by backward induction.

5.1.2 Buyer Purchase

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

$$D(p, V) = \mathbf{1}\left(p \leq V\right).$$

The seller therefore sets

$$p^* = V,$$

sells to all buyers, and obtains profit equal to the value of its dataset design:

$$\Pi_i^{\text{no}} = V_i.$$

Intuitively, with homogeneous buyers and no frictions, the platform can fully extract the expected gains from improved prediction.

Acquisition For the entrant E, the standalone outside option is the expected gains from improved prediction when it operates separately from I, earning

$$\Pi_E^{\text{no}} = V_1.$$

To acquire E, it is sufficient that the incumbent I offer it this amount. When I acquires E, it combines covariates and gains an increase of value

$$V_2 - V_1$$

compared to its outside option V_1.

Proposition 5 (Harmful Acquisition)**.** *The incumbent I always acquires the entrant E.*

Intuitively, the economies of scope stemming from Corollary 6 implies that

$$V_2 - \sum_{i \in \{I,E\}} V_i \geq 0.$$

Combining disjoint datasets makes each covariate more informative. This is due to two things: first, as far as prediction covariates $\boldsymbol{\mathcal{K}}_i$ can benefit from the covariate spillovers of the training covariates $\boldsymbol{\mathcal{K}}_{-i}$ (and vice versa); second, due to the House Party Effect, pushing training covariates from $\boldsymbol{\mathcal{K}}_i$ to $\boldsymbol{\mathcal{K}}_I \cup \boldsymbol{\mathcal{K}}_E$ yields supermodular gains. As the incumbent can fully exploit these complementarities, it will prevent entry by E even if demands are independent. This is the natural-monopsony effect: data-driven economies of scope push data into the hands of a single firm.

5.1.3 Planner's Problem

With homogeneous buyers and full surplus extraction, customer surplus is zero in equilibrium. Social welfare therefore equals industry profit net of any cost of foregone entry and spillovers:

$$W = \begin{cases} \Pi_2 - \xi, & \text{if } I \text{ acquires } E, \\ \Pi_I + \Pi_E, & \text{otherwise,} \end{cases}$$

where $\xi \geq 0$ denotes the social cost of eliminating the entrant as an independent firm.

Proposition 6. *The acquisition is harmful whenever*

$$V_2 - 2V_1 \geq \xi.$$

This condition is equivalent to

$$k \leq \tilde{k}(n, \xi) \equiv \frac{n+1}{2} \frac{\sqrt{\xi(8n+\xi)} - 3\xi}{n - \xi},$$

which is U-shaped in n and increasing in ξ.

The social planner compares the economies of covariate integration reflected in $V_2 - 2V_1$ and the spillovers ξ to evaluate whether the loss of spillovers is worth the gains in prediction value it generates. However, the incumbent does not internalize the loss of positive spillovers, generating a tension between private and social incentives for acquisition. Although scope economies make data combinations privately valuable, they also cause the incumbent to systematically over-acquire entrants. The incumbent captures all private complementarities but

does not internalize the entrant's contribution to future innovation, data generation, or technological diversity. The resulting acquisitions are therefore privately efficient but potentially socially harmful—too many mergers occur from society's perspective.

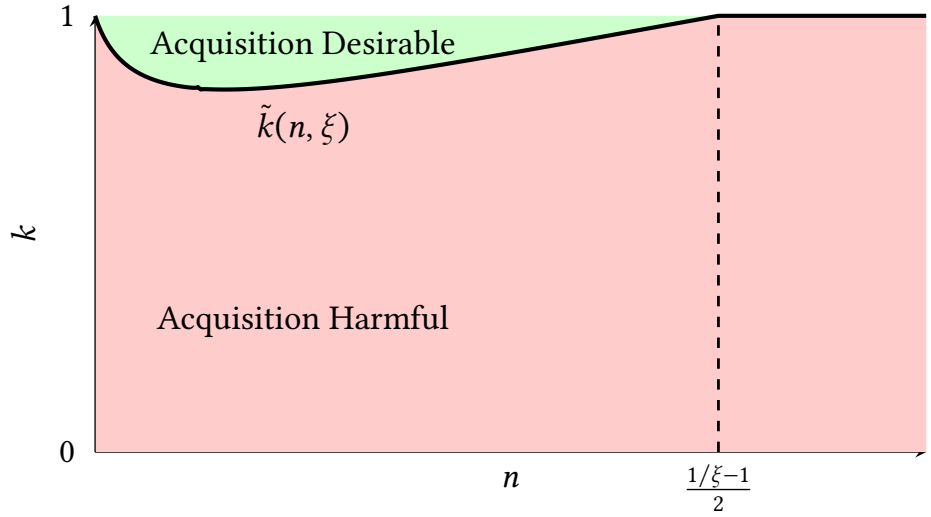

Figure 2: An acquisition is harmful for all (n, k) below the U-shaped threshold $k = \tilde{k}(n, \xi)$, where the loss of innovation spillovers exceeds the scope economies from combining datasets.

Regulation and Remedies Privacy regulation can worsen this tension. By raising compliance costs, the EU's General Data Protection Regulation (GDPR) lowers k, thereby dampening economies of scope and reducing the interval $[0, \tilde{k}(n, \xi)]$ in which the acquisition is socially desirable. Conversely, open-data initiatives raise k and broaden the interval in which economies of scope make acquisition desirable. These results suggest that privacy law, data governance, and merger policy must be coordinated: measures that restrict data access may unintentionally increase the likelihood of harmful data-driven acquisitions.

A potential remedy to potentially harmful acquisitions are FRAND APIs (Fair, Reasonable, and Non-Discriminatory Application Programming Interfaces). FRAND APISs are a regulated mechanism that guarantees standardized, non-exclusive access to a firm's data on fair technical and pricing terms. It preserves interoperability by allowing rivals to continue accessing the entrant's data after a merger. In this model, a FRAND API mitigates the dynamic loss ξ by keeping the entrant's data and experimentation available to the ecosystem, while still allowing the incumbent to capture the scope economies $\Delta(n, k)$. Therefore, the model provides support to the FRAND access obligations adopted in EU digital regulation as merger remedies, most notably in the Digital Markets Act and the Data Act.

5.2 Data Pools

Data owners often form partnerships to pool their datasets and sell access jointly. For example, BMW, Mercedes-Benz, and Audi co-founded the platform *Here Mobility Data Marketplace*, which aggregates GPS, speed, and road-condition data from connected cars. These agreements can benefit society as they address an instance fo the fundamental *complements problem* originally noted by Cournot (1838): if independent monopolists sell datasets that are complements, the resulting double marginalization leads to inefficiently high prices. Consequently, policymakers have encouraged pooling or sharing of datasets to overcome these frictions in the EU Data Act and in the European strategy for common data spaces. However, several papers have warned that brokers might use data sharing agreements to collude.

To assess the merits of these arguments, we develop a model in the spirit of Lerner and Tirole (2004): it turns out that the economics of combining datasets closely matches the economics of combining patents.

5.2.1 Setup

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

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

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

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

$$U_q = V_q - \theta,$$

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

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

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

The heterogeneity parameter θ is distributed according to

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

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

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

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

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

5.2.2 Pooling Price

When data owners form a pool, they coordinate pricing and behave as a single monopolist offering a pooled dataset. This situation mirrors the *pool pricing benchmark* in Section I.b of Lerner and Tirole (2004), where a patent pool chooses the package price P to maximize joint revenue given demand $D(P - V_2)$. The logic is analogous here: the pooled data pool yields predictive value V_2, and all buyers face a single posted price P.

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

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

The pool acts as a monopolist setting the joint-access price that equates marginal revenue to zero, just as in Equation (1) in Lerner and Tirole (2004). The resulting price is decreasing in the demand semi-elasticity α: when α is high, buyers are more sensitive to price, leading the

pool to charge less. Conversely, when α is low, demand is inelastic, so the pool can extract a larger fraction of the total value V_2.

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

5.2.3 Covariate Silos

If the datasets are split in two sets of covariates, the datasets are complements by Corollary 6. We can use the results in Section II in Lerner and Tirole (2004) to characterize the unique symmetric equilibrium in the case in which brokers selling complementary data do not form a pool.

Lemma 5 (Covariate Fragmentation Price)**.** *If the brokers have distinct covariates each B_i prices at*

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

and the buyers will buy from both brokers.

Applying Proposition 1 in Lerner and Tirole (2004) directly yields the following result.

Proposition 7. *A pool of buyers with distinct covariates is always procompetitive.*

Covariate silos are an instance of the classic complements problem of Cournot (1838). Each broker prices without internalizing that lowering its price increases the value of the other dataset. Pooling eliminates this double marginalization and behaves exactly like integration in complementary patents in Lerner and Tirole (2004).

5.2.4 Observation Silos

If the datasets are split in two sets of covariates, the datasets are substitutes by Corollary 5. We can use the results in Section II in Lerner and Tirole (2004) characterizes the unique symmetric equilibrium in the case in which the brokers selling substitute data do not form a pool.

Lemma 6 (Sample Fragmentation Price)**.** *If the brokers have different observations on the same covariates,*

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

and the buyers will buy from both brokers.

Applying Proposition 1 in Lerner and Tirole (2004) directly yields the following characterization of when fragmented sample pooling is procompetitive.

Proposition 8 (Welfare of Observation Pooling). *A pool of observations is procompetitive if and only if*

$$k < \hat{k}(n, \alpha) \equiv 1 - \frac{n}{2\alpha}.$$

The condition $k \leq \hat{k}(n, \alpha)$ mirrors the distinction between the *competition margin* and the *demand margin* in Lerner and Tirole (2004). When k is low, the incremental predictive value of the second block of observations, $V_2 - V_1$, remains substantial. Buyers therefore strictly prefer to acquire both datasets, so each broker must set a price low enough not to be excluded. Fragmented pricing is thus constrained by the competition margin, and the two datasets act as effective complements at the equilibrium prices. Separate pricing then reproduces the classic Cournot double-marginalization problem, and pooling internalizes this externality, lowering the total price.

When $k > \hat{k}(n, \alpha)$, the incremental value $V_2 - V_1$ becomes small. Buyers are nearly indifferent between acquiring one or two datasets, so each broker can raise its price without being displaced. Pricing is now governed by the demand margin, and a pool relaxes this competitive pressure. In this region, pooling raises the joint-access price and is therefore anti-competitive.

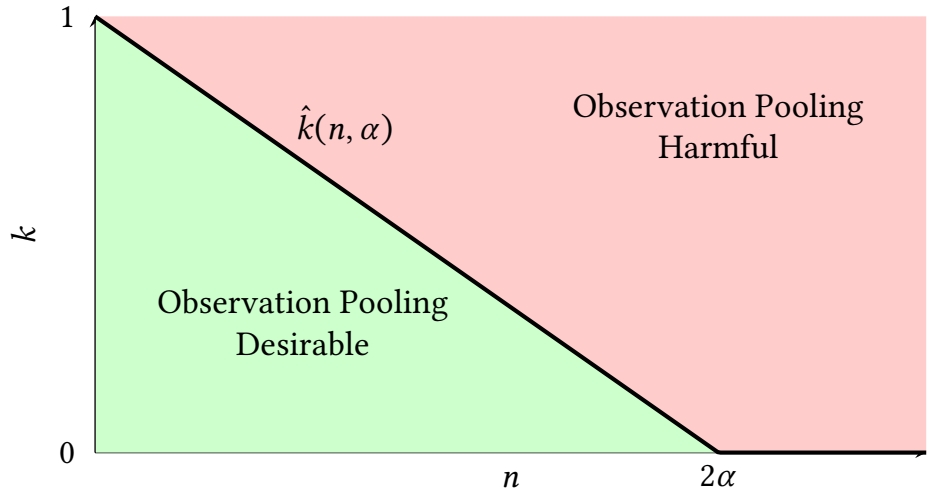

Figure 3: The threshold $k = \hat{k}(n, \alpha)$ separates the region in which fragmented pricing is constrained by the competition margin (pooling lowers prices) from the region in which it is constrained by the demand margin (pooling raises prices).

Because $\hat{k}(n, \alpha)$ is decreasing in n, the model offers a sharp statistical insight: observation pooling becomes anti-competitive when datasets are large. As n grows, posterior variance is already small, so doubling the sample yields only a negligible reduction in prediction error. The marginal value $V_2 - V_1$ of the second dataset therefore collapses. In this case, fragmentation disciplines prices through the competition margin—each broker must keep its price close to $V_2 - V_1$ to avoid being dropped—while pooling removes this discipline. As a result, pooling

observations raises the price of access and is anti-competitive in the upper-right corner of the (n, k) plane.

Regulation of Data Pools The model qualifies current policy discussions on data sharing. European legislation such as the EU Data Act and the European strategy for common data spaces broadly promotes data pooling and data spaces, which can overcome the Cournot complements problem in fragmented data markets. However, these policy initiatives do not distinguish between sources of dataset heterogeneity (covariates vs. observations) as a condition for pooling to be procompetitive. The model shows that this distinction is essential. Pooling heterogeneous user attributes (different covariates) always resolves a pure complements problem and is therefore procompetitive, fully consistent with the intuition in the Data Act Recitals that data portability and interoperability reduce market power. In contrast, pooling additional observations on the same attributes may be anticompetitive when $k > \hat{k}(n, \alpha)$: in this region the second dataset adds little predictive value, fragmented pricing would be disciplined by the competition margin, and a pool merely relaxes competitive pressure.

These results imply that policies encouraging sharing of data on similar attributes may unintentionally raise prices in markets where data is abundant. Conversely, when different attributes are shared, the model confirms the procompetitive rationale underlying EU data-space initiatives. More generally, the model highlights the need for analytical tools within merger control and data-access regulation to understand the level of heterogeneity of the shared data, a distinction absent from existing regulatory guidance.

5.3 Data Exclusivity

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

5.3.1 Setup

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

Prediction Buyers. There is a unit mass of customers, divided into:

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

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

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

$$u_i = V_i - p_i,$$

where V_i denotes the acccuracy of the prediction and p_i is the price charged. Shoppers buy from the firm with the highest net utility $V_i - p_i$, while captive customers always purchase from their incumbent firm.

Data Buyers. We assume covariates are ex-ante identical and that training and prediction covariates coincide. Each firm $i \in \{1, 2\}$ starts with no data and can improve its prediction quality through two channels:

1. *Training Data Licensing.* Firm i can buy a license for the seller's training dataset of k covariates and the realization of the target variable for n observations, paying a fee f. This choice is represented by a binary variable $\ell_i \in \{0, 1\}$.
2. *Proprietary Data Collection.* Firm i can collect the k covariates for the target user, paying a fixed cost $c > 0$, represented by a binary choice $r_i \in \{0, 1\}$.

A firm's data strategy is $(\ell_i, r_i) \in \{0, 1\}^2$, the prediction quality of firm i is

$$V_i \equiv V(\ell_i, r_i) = \hat{V} \times \mathbf{1}(\ell_i r_i = 1),$$

where

$$\hat{V} \equiv \frac{k}{\frac{1-k}{n} + 1}.$$

Intuitively, in order to make a prediction, the firm must license the training data and invest to collect proprietary data on the target individual.

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

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

and total profit is

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

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

Data Seller. The data seller S has a monopoly over the training dataset and sets a take-it-or-leave-it license fee f.[8] The seller's profit is

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

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

Social Planner. A benevolent social planner evaluates total welfare as

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

where customer surplus is given by

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

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

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

1. **Data Pricing.** The data seller S sets a publicly observed license fee f;
2. **Data Collection.** Each firm F_i chooses $(\ell_i, r_i) \in \{0,1\}^2$, which are publicly observed;

[8] Allowing S to set different fees for F_1 and F_2 would not affect the analysis as far s exclusivity is concerned.

3. **Prediction Pricing.** Firms simultaneously set (p_i^c, p_i^s). Customers choose the firm offering higher utility,

4. **Realization.** Nature draws the data and the target variable, prediction losses and profits are realized for S, F_1, and F_2.

We solve the game by backward induction and characterize the Subgame Perfect Equilibrium (SPE).

5.3.2 Prediction Pricing

We analyze the pricing subgame, taking prediction accuracies (V_1, V_2) as given from the data collection. Firms can price-discriminate between captive customers and shoppers.

Captive pricing. Let the outside option yield utility 0. Assuming they are expected utility maximizers, captive customers of F_i buy if and only if $V_i - p_i \geq 0$. Therefore, the unique optimal price is

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

The firm extracts the full surplus of captives.

Shopper pricing. Shoppers buy from the firm that offers the higher net utility; if $V_i - p_i > V_j - p_j^s$ they all buy from F_i, if $V_i - p_i < V_j - p_j^s$ they all buy from F_j, and if equal any split yields the same payoff for firms and customers.[9] Write $(x)^+ \equiv \max\{x, 0\}$. The next lemma characterizes the unique trembling-hand perfect equilibrium for shoppers.

Lemma 7 (Shopper-price equilibrium)**.** *Fix (V_1, V_2). The pricing subgame in shopp[Shopper-price equilibrium]er prices admits a unique equilibrium under trembling-hand perfection:*

$$p_i^s = (V_i - V_j)^+, \qquad i \neq j, \tag{3}$$

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

Combining Equation (2) and Lemma 7, the equilibrium profit of firm i given data strategies (ℓ_i, r_i) and (ℓ_j, r_j) (which pin down (V_i, V_j)) is

$$\pi(\ell_i, r_i; \ell_j, r_j) \;=\; \frac{1-s}{2}\, V_i \;+\; s\,(V_i - V_j)^+, \qquad j \neq i. \tag{4}$$

The first term is captive revenue at the value price; the second term is shopper revenue, which equals the quality advantage when the firm is better and zero otherwise.

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

5.3.3 Cost Regions

The trade-off between exclusivity and non-exclusivity depends on the equilibirum of the proprietary data investment subgame, which in turn depends on the cost of collecting proprietary data c relative to the value $\hat{V}$. Two regimes arise:

- **Low-cost region:** $c \in [0, \bar{c})$, where

$$\bar{c} \equiv \frac{1-s}{2}\hat{V}.$$

  Here proprietary data are cheap enough that any licensed firm will always invest.

- **High-cost region:** $c \in [\bar{c}, \bar{\bar{c}})$, where

$$\bar{\bar{c}} \equiv \frac{1+s}{2}\hat{V}.$$

  Here investment occurs only when a firm obtains a monopoly over shoppers, leading to mixed strategies when both license.

- **No investment region:** $c \geq \bar{\bar{c}}$. Here investment never occurs.

These regions generate two distinct exclusivity incentives for the seller and two distinct welfare benchmarks for the planner, which we will analyze separately.

5.3.4 Low-cost Region

Proprietary Data Investment. If $c < \bar{c}$, the cost of investing is less than the increase in revenue on a firm's captive customers. Therefore investment will occur if and only if the license is purchased.

Proposition 9. *If $c < \bar{c}$, for any licensing choice (ℓ_1, ℓ_2) there is a unique Nash Equilibrium*

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

The intuition is straightforward: the rival's competition in the shopper segment is irrelevant because the incentives to invest in the captive segment are sufficiently large relative to the cost of investing. Furthermore, because investing in proprietary data is costly but gives no benefit if the license is acquired, in case a firm does not license it will also avoid investing.

Data Pricing. The seller chooses these fees to maximize its profit, taking into account that the fees determine how many firms choose to license the data. The following result characterizes the maximal fee f_d compatible with $d \in \{1, 2\}$ firms purchasing the data.

Lemma 8. *If $c < \bar{c}$,*

$$f_2 = \frac{1-s}{2}\hat{V} - c, \quad \textit{and} \quad f_1 = \frac{1+s}{2}\hat{V} - c.$$

Intuitively, if both firm license the training data they will both invest and will only be able to extract the surplus of the captive customers $\frac{1-s}{2}\hat{V}$. Instead, if only one licenses, it will extract the full surplus of the shoppers as well, giving it an additional demand of sA, because by Proposition 9 it will invest, and since $f_1 > f_2$ the rival will not license as it is only profitable for a single firm to license.

If the profit from selling to one firm is higher than selling to both, i.e., $f_1 > 2f_2$, the data seller will pursue *de facto exclusivity*, setting a price so high hat only one firm will choose to purchase the data and the other will avoid entry.

Proposition 10. *The data seller will choose $f^* = f_1$, i.e., it will pursue de facto exclusivity, if $c \in [\underline{c}, \bar{c})$, where*

$$\underline{c} \equiv \frac{1-3s}{2}\hat{V} \in [0, \bar{c}].$$

Therefore, whenever competition is fierce, i.e. $s > 1/3$, the seller will always prefer de facto exclusivity. If $s \leq 1/3$, the seller will prefer de facto exclusivity if the cost of collecting prediction covariates is large compared to $\hat{V}$. Since $\hat{V} = V(n,k)$, which is increasing in both arguments, exclusivity is profitable if datasets are sufficiently small.

Social Welfare. I assume the planner can choose between de facto exclusivity and real non exclusivity but cannot influence the proprietary data investment equilibrium.

Proposition 11. *If $c < \bar{c}$, the planner prefers real non-exclusivity to de facto exclusivity.*

The cost is too low for the motive of avoiding excessive entry and overinvestment in data collection ot make it socially desirable to pursue de facto exclusivity. The following result characterizes an interval in which de facto excusivity is profitable for the data seller even if it harms social welfare.

Corollary 10. *The seller chooses de facto exclusivity even though it is harmful if and only if $c \in (\underline{c}, \bar{c})$.*

The length of the interval is $\bar{c} - \underline{c} = s\hat{V}$, so harmful exclusivity is more likely if competition is fierce and, given $\hat{V} \equiv V(n,k)$, which is increasing in both arguments, if data is abundant.

5.3.5 High-cost Region

I now consider what happens when the cost of proprietary data collection is $c \in (\bar{c}, \bar{\bar{c}})$. This means the cost of investing is more than the surplus of a firm's captive customers but possibly less than the cumulative surplus of a firm's captive customers and the shoppers.

Proprietary Data Investment. If $c \in (\bar{c}, \bar{\bar{c}})$, a firm will collect prediction covariates with probability one only if it can sell as a monopolist to the shoppers, meaning that it benefits from de facto exclusivity. If both firms purchase the license, there is no pure-strategy equilibrium in the proprietary data-collection subgame.

Proposition 12. *We distinguish two subcases:*

- ***De facto exclusivity:*** *Without loss of generality, let $\ell_1 = 1$, $\ell_2 = 0$, the unique NE is*

$$r_1 = 1, \qquad r_2 = 0.$$

- ***Both firms license:*** *$\ell_1 = \ell_2 = 1$. The unique symmetric mixed strategy equilibrium has each firm investing with probability*

$$\xi^* = \frac{V(1+s) - 2c}{2As}, \qquad \xi^* \in [0, 1].$$

  In equilibrium, profits are zero: $\Pi_i^{\text{mix}} = 0$.

The mixed strategy equilibrium investment probability is increasing in s if and only if $c > \hat{V}/2$ (i.e., when proprietary data are relatively costly) and decreasing in s when $c < \hat{V}/2$.

Data Pricing. As the profit of the mixed strategy equilibirum is zero, firms will have no willingness to pay for the license unless there is de facto exclusivity. Under de facto exclusivity, only one firm licenses and invests, while the rival neither licenses nor invests.

Proposition 1. *If $c \in (\bar{c}, \bar{\bar{c}})$, the data seller will always choose $f^* = f_1$, i.e., it will always pursue de facto exclusivity, where*

$$f_1 = \frac{1+s}{2} V_2 - c.$$

Social welfare. The following result characterizes the social planner optimum, by comparing the pure strategy equilibrium of the investment subgame and the mixed strategy non-exclusive one.

Proposition 13. *If $c \in (\bar{c}, \bar{\bar{c}})$, the planner prefers de facto exclusivity to real non-exclusivity.*

The overall investment in the real non exclusive case is so low that the planner prefers to gran exclusivity in order to ensure that investment occurs.

5.3.6 Summary.

The equilibrium interaction between exclusivity, investment, and welfare is governed by the fixed cost of proprietary data collection c, relative to the value

$$\hat{V} \equiv \frac{k}{\frac{1-k}{n} + 1}.$$

Three cost thresholds determine firms' willingness to collect proprietary covariates:

$$\underline{c} \equiv \frac{1-3s}{2}\hat{V}, \qquad \bar{c} \equiv \frac{1-s}{2}\hat{V}, \qquad \bar{\bar{c}} \equiv \frac{1+s}{2}\hat{V}.$$

For fixed (c, s) we can invert these equations and obtain three endogenous thresholds in covariate richness,

$$\underline{k}(n, c, s), \qquad \bar{k}(n, c, s), \qquad \bar{\bar{k}}(n, c, s),$$

defined implicitly by

$$c = \frac{1-3s}{2}V(n, \underline{k}), \qquad c = \frac{1-s}{2}V(n, \bar{k}), \qquad c = \frac{1+s}{2}V(n, \bar{\bar{k}}).$$

For $s < 1/3$ these satisfy

$$\underline{k}(n, c, s) < \bar{k}(n, c, s) < \bar{\bar{k}}(n, c, s) \qquad \text{for all } (n, c).$$

These thresholds partition the (n, k)-space into four regions:

- **No investment:** If $k < \underline{k}(n, c, s)$, proprietary data collection is never profitable, even under exclusivity. No firm invests, and licensing decisions have no effect.
- **Beneficial exclusivity:** If $k \in [\underline{k}(n, c, s), \bar{k}(n, c, s))$, investment occurs only under exclusivity, because a firm collects proprietary data if and only if it can monopolize shoppers. Both the seller and the planner prefer exclusivity, as it ensures that at least one firm invests.
- **Harmful exclusivity:** If $k \in [\bar{k}(n, c, s), \bar{\bar{k}}(n, c, s))$, both firms would invest even without exclusivity, but the seller prefers to restrict access to extract a higher fee. The planner prefers non-exclusivity to keep both firms active. Exclusivity is privately profitable but socially harmful.
- **No exclusivity:** If $k \geq \bar{\bar{k}}(n, c, s)$, proprietary data are sufficiently cheap relative to dataset quality that both firms always invest. Both the seller and the planner prefer non-exclusivity, so broad data access is jointly optimal.

The figure below plots these three thresholds and the resulting four regions for a fixed pair (c, s) with $s < 1/3$.

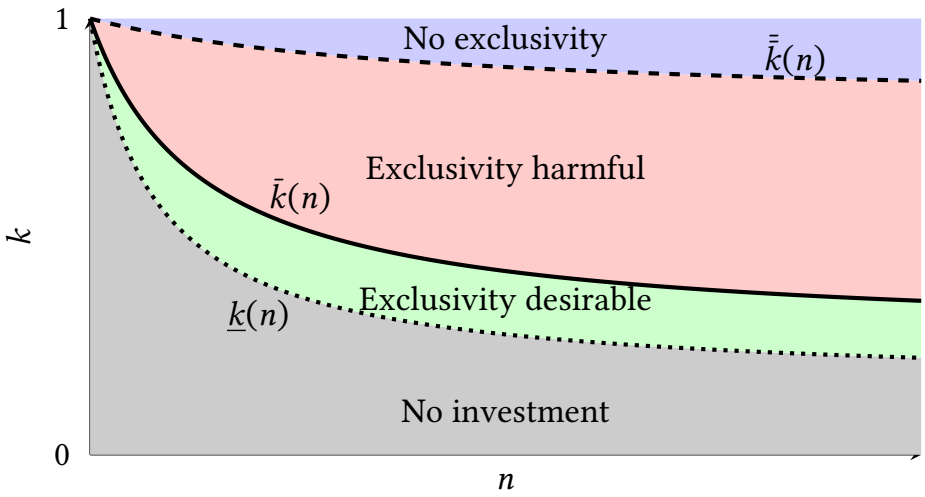

Figure 4: The three thresholds $k = \underline{k}(n)$, $k = \bar{k}(n)$, and $k = \bar{\bar{k}}(n)$ partition the (n, k)-space into four regions. Below $\underline{k}(n)$, investment never occurs. Between $\underline{k}(n)$ and $\bar{k}(n)$, exclusivity is jointly optimal because it induces investment. Between $\bar{k}(n)$ and $\bar{\bar{k}}(n)$, exclusivity is privately profitable but socially harmful. Above $\bar{\bar{k}}(n)$, both the seller and the planner prefer non-exclusivity.

6 Conclusion

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

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

More broadly, the analysis cautions against treating data as homogeneous. Policies promoting open data without regard to dataset composition may miss crucial efficiency mar-

gins, 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[10] 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.[11] 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.

[10] See Bajari et al. (2019), Schaefer and Sapi (2023), Lee and Wright (2023), Yoganarasimhan (2020), and Carballa-Smichowski et al. (2025b)

[11] See Black and Scholes (1973), Merton (1973)

References

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

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

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

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

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

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

Bergemann, Dirk and Alessandro Bonatti (2024). "Data, Competition, and Digital Platforms." In: *American Economic Review* 114.8, pp. 2553–2595.

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

Bergemann, Dirk, Alessandro Bonatti, and Alex Smolin (2018). "The Design and Price of Information." In: *American Economic Review* 108.1, pp. 1–48.

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

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

Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability." In: *Monthly Weather Review* 78.1, pp. 1–3.

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

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

Carballa-Smichowski, Bruno et al. (2025b). "Economies of scope in data aggregation: Evidence from health data." In: *Information Economics and Policy* 71, p. 101146.

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

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

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

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

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

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

Digital Platforms, Stigler Committee on (2019). *Final Report*. Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business.

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." In: *Management Science* 57.1, pp. 57–71.

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

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

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

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

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

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." In: *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 (2020). "Nonrivalry and the Economics of Data." In: *American Economic Review* 110.9, 2819–58.

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

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

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

Lerner, Josh and Jean Tirole (2004). "Efficient Patent Pools." In: *American Economic Review* 94.3, 691–711.

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

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

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

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

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

Nocke, Volker, Martin Peitz, and Konrad Stahl (Dec. 2007). "Platform Ownership." In: *Journal of the European Economic Association* 5.6, pp. 1130–1160.

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

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

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." In: *Information Economics and Policy* 65, p. 101063.

Schmidt, Eric (Oct. 2, 2009). *How Google Plans to Stay Ahead in Search*. Accessed: 2025-11-12. Bloomberg. URL: https://www.bloomberg.com/news/articles/2009-10-02/how-google-plans-to-stay-ahead-in-search.

Selten, Reinhard (1998). "Axiomatic Characterization of the Quadratic Scoring Rule." In: *Experimental Economics* 1.1, pp. 43–62.

Strzalecki, Tomasz (2024). *Variational Bayes and non-Bayesian Updating*.

UK Competition and Markets Authority (2019). *Unlocking Digital Competition: Report of the Digital Competition Expert Panel*. London: Competition and Markets Authority.

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

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

A Proofs

Theorem 1 (High-dimensional Value of Design)**.** *Under Assumption 1,*

$$\lim_{Z\to\infty} V(\mathcal{D}(Z)) = \min\{p, t\}(1-\underline{\sigma}^2)\,\mathcal{I}^{\mathrm{hd}}(n, t; \sigma^2(t)).$$

Define

$$\bar{V}(n, t, p) \equiv \lim_{Z\to\infty} V(\mathcal{D}(Z)).$$

Proof. Fix $Z \geq 1$ and a training set $\mathcal{T}(Z) \subseteq \{1, \ldots, Z\}$. Recall

$$\varepsilon^i_{\mathcal{T}(Z)} = u^i + \sum_{k\notin\mathcal{T}(Z)} \beta_k x^i_k.$$

By the DGP in Section 2.1, we have $u^i \sim \mathcal{N}(0, \tau^2\underline{\sigma}^2)$, independent of $\{(\beta_k, x^i_k)\}_{k=1}^Z$, and for each k,

$$\beta_k \sim \mathcal{N}(0, \tau^2), \qquad x^i_k \sim \mathcal{N}\left(0, \frac{1-\underline{\sigma}^2}{Z}\right),$$

with mutual independence across k and independence between β_k and x^i_k.

Step 1: Limiting relative regression noise. Conditional on $\mathcal{T}(Z)$, the variance of the residual is

$$\mathrm{Var}\left[\varepsilon^i_{\mathcal{T}(Z)} \mid \mathcal{T}(Z)\right] = \mathrm{Var}[u^i] + \mathrm{Var}\left[\sum_{k\notin\mathcal{T}(Z)} \beta_k x^i_k \;\middle|\; \mathcal{T}(Z)\right],$$

where the cross term is zero because u^i is independent of the sum and has mean zero. Moreover, for $k \neq \ell$, independence across covariates implies $\mathrm{Cov}(\beta_k x^i_k, \beta_\ell x^i_\ell) = 0$, and therefore

$$\mathrm{Var}\left[\sum_{k\notin\mathcal{T}(Z)} \beta_k x^i_k \;\middle|\; \mathcal{T}(Z)\right] = \sum_{k\notin\mathcal{T}(Z)} \mathrm{Var}(\beta_k x^i_k).$$

Since β_k and x^i_k are independent and centered,

$$\mathrm{Var}(\beta_k x^i_k) = \mathbb{E}[\beta_k^2]\mathbb{E}[(x^i_k)^2] = \tau^2 \cdot \frac{1-\underline{\sigma}^2}{Z}.$$

Hence,

$$\mathrm{Var}\left[\varepsilon^i_{\mathcal{T}(Z)} \mid \mathcal{T}(Z)\right] = \tau^2\underline{\sigma}^2 + |\overline{\mathcal{T}(Z)}| \cdot \tau^2 \frac{1-\underline{\sigma}^2}{Z},$$

where $\overline{\mathcal{T}(Z)} \equiv \{1, \ldots, Z\} \setminus \mathcal{T}(Z)$ and $|\overline{\mathcal{T}(Z)}| = Z - |\mathcal{T}(Z)|$. Dividing by τ^2 yields

$$\sigma^2(\mathcal{T}(Z)) = \frac{\mathrm{Var}[\varepsilon^i_{\mathcal{T}(Z)} \mid \mathcal{T}(Z)]}{\tau^2} = \underline{\sigma}^2 + \left(1 - \frac{|\mathcal{T}(Z)|}{Z}\right)(1-\underline{\sigma}^2).$$

Under Assumption 1, $|\mathcal{T}(Z)|/Z \to t$, so

$$\lim_{Z\to\infty} \sigma^2(\mathcal{T}(Z)) = \underline{\sigma}^2 + (1-t)(1-\underline{\sigma}^2) \;\equiv\; \sigma^2(t) \in [\underline{\sigma}^2, 1].$$

Step 2: Asymptotic Gaussianity of the residual. Conditional on $\mathcal{T}(Z)$, the error term u^i is Gaussian and independent of $S_Z^i \equiv \sum_{k\notin\mathcal{T}(Z)} \beta_k x_k^i$. We show that S_Z^i converges in distribution to a centered Gaussian with variance $\tau^2(1-t)(1-\underline{\sigma}^2)$, and then add u^i.

Let $m_Z \equiv |\overline{\mathcal{T}(Z)}|$ and enumerate the omitted indices as $\{k_{1,Z}, \dots, k_{m_Z,Z}\} = \overline{\mathcal{T}(Z)}$. Define the triangular array

$$\xi_{j,Z}^i \equiv \beta_{k_{j,Z}} x_{k_{j,Z}}^i, \qquad j = 1, \dots, m_Z.$$

Conditional on $\mathcal{T}(Z)$, the random variables $\{\xi_{j,Z}^i\}_{j=1}^{m_Z}$ are independent, centered, and have common variance

$$\mathrm{Var}(\xi_{j,Z}^i) = \tau^2 \frac{1-\underline{\sigma}^2}{Z}.$$

Therefore,

$$\mathrm{Var}(S_Z^i \mid \mathcal{T}(Z)) = \sum_{j=1}^{m_Z} \mathrm{Var}(\xi_{j,Z}^i) = m_Z\, \tau^2 \frac{1-\underline{\sigma}^2}{Z} = \tau^2 \left(1 - \frac{|\mathcal{T}(Z)|}{Z}\right)(1-\underline{\sigma}^2) \;\to\; \tau^2(1-t)(1-\underline{\sigma}^2).$$

To establish a CLT, we verify the Lindeberg condition. Fix $\epsilon > 0$ and let $s_Z^2 \equiv \mathrm{Var}(S_Z^i \mid \mathcal{T}(Z))$. Using Cauchy–Schwarz,

$$\mathbb{E}\left[(\xi_{j,Z}^i)^2\, \mathbf{1}\{|\xi_{j,Z}^i| > \epsilon s_Z\}\right] \le \sqrt{\mathbb{E}[(\xi_{j,Z}^i)^4]}\, \sqrt{\mathbb{P}(|\xi_{j,Z}^i| > \epsilon s_Z)}.$$

Because β_k and x_k^i are independent Gaussian, their fourth moments are finite and

$$\mathbb{E}[(\xi_{j,Z}^i)^4] = \mathbb{E}[\beta_{k_{j,Z}}^4]\mathbb{E}[(x_{k_{j,Z}}^i)^4] = (3\tau^4) \cdot \left(3\left(\frac{1-\underline{\sigma}^2}{Z}\right)^2\right) = \frac{9\tau^4(1-\underline{\sigma}^2)^2}{Z^2}.$$

Moreover, by Markov's inequality,

$$\mathbb{P}(|\xi_{j,Z}^i| > \epsilon s_Z) \le \frac{\mathbb{E}[(\xi_{j,Z}^i)^2]}{\epsilon^2 s_Z^2} = \frac{\tau^2(1-\underline{\sigma}^2)/Z}{\epsilon^2 s_Z^2} = O\left(\frac{1}{Z}\right),$$

since $s_Z^2 \to \tau^2(1-t)(1-\underline{\sigma}^2) > 0$ whenever $t < 1$ (and when $t = 1$ the sum is identically zero for all Z, so the claim is trivial). Hence

$$\mathbb{E}\left[(\xi_{j,Z}^i)^2\, \mathbf{1}\{|\xi_{j,Z}^i| > \epsilon s_Z\}\right] = O\left(\frac{1}{Z^{3/2}}\right),$$

uniformly in j, and therefore

$$\frac{1}{s_Z^2}\sum_{j=1}^{m_Z}\mathbb{E}\left[(\xi_{j,Z}^i)^2\,\mathbf{1}\{|\xi_{j,Z}^i|>\epsilon s_Z\}\right]=\frac{m_Z}{s_Z^2}\,O\!\left(\frac{1}{Z^{3/2}}\right)=O\!\left(\frac{1}{\sqrt{Z}}\right)\to 0.$$

Thus the Lindeberg–Feller CLT applies and, conditional on $\mathcal{T}(Z)$,

$$S_Z^i \Longrightarrow \mathcal{N}\left(0,\,\tau^2(1-t)(1-\underline{\sigma}^2)\right).$$

Finally, since u^i is independent Gaussian with variance $\tau^2\underline{\sigma}^2$, the sum $\varepsilon^i_{\mathcal{T}(Z)}=u^i+S_Z^i$ converges to a Gaussian with variance equal to the sum of the limiting variances:

$$\varepsilon^i_{\mathcal{T}(Z)} \Longrightarrow \mathcal{N}\left(0,\,\tau^2[\underline{\sigma}^2+(1-t)(1-\underline{\sigma}^2)]\right)=\mathcal{N}\left(0,\,\tau^2\sigma^2(t)\right).$$

This concludes the proof. □

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

$$\hat{y}^*(\boldsymbol{D})=\mathbb{E}\left[y\mid \boldsymbol{D}\right]=\boldsymbol{x}'_{\mathcal{P}\cap\mathcal{T}}\underbrace{\left(\left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}}+\sigma^2\cdot\boldsymbol{I}_{|\mathcal{T}|}\right)^{-1}\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{y}\right)_{\mathcal{P}\cap\mathcal{T}}}_{=\mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}}\mid(\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})]}.$$

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

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

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

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

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

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

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

$$\mathbb{E}[y\mid(\boldsymbol{y},\boldsymbol{X}),\boldsymbol{x}_{\mathcal{P}}]=\boldsymbol{x}'_{\mathcal{P}}\,\mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}}\mid(\boldsymbol{y},\boldsymbol{X})],$$

which proves the claim. □

Proposition 2. *Value of a dataset $\boldsymbol{D}$ is*

$$v(\boldsymbol{D};\sigma^2) = \operatorname{Var}[y] - \operatorname{Var}[y \mid \boldsymbol{D};\sigma^2] = \boldsymbol{x}'_{\mathcal{P}\cap\mathcal{T}} \underbrace{\left(\tau^2 \cdot \boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} \left(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} + \sigma^2 \cdot \boldsymbol{I}_{|\mathcal{T}|}\right)^{-1}\right)_{\mathcal{P}\cap\mathcal{T}}}_{= \operatorname{Var}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}}] - \operatorname{Var}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}}|(\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})]} \boldsymbol{x}_{\mathcal{P}\cap\mathcal{T}}.$$

Proof. Since the residual is Gaussian, we can apply the results in DeGroot (2005) or Berger (1990). □

Lemma 2 (Information on a trained parameter)**.** *Fix $\sigma^2 > 0$. Under the assumptions of Section 2.1,*

$$\mathcal{I}(N,\mathcal{T};\sigma^2) = \frac{\tau^2}{|\mathcal{T}|} \times \operatorname{tr}\Big[\mathbb{E}_{N,\mathcal{T};\sigma^2}\left[\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}}(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}} + \sigma^2 \cdot \boldsymbol{I}_{|\mathcal{T}|})^{-1}\right]\Big] \in [0, \tau^2]$$

Proof. The proof follow directly from the definition of the training information and Proposition 2. □

Proposition 3 (High-dimensional information)**.** *Under Assumption 1,*

$$\lim_{Z\to\infty} \mathcal{I}(N(Z),\mathcal{T}(Z);\sigma^2) = \mathcal{I}^{\mathrm{hd}}(n,t;\sigma^2),$$

*where $\mathcal{I}^{\mathrm{hd}}(n,t;\sigma^2) \in (0,1)$ is the **high-dimensional information** defined by*

$$\mathcal{I}^{\mathrm{hd}}(n,t;\sigma^2) \equiv \frac{\sigma^2+n+t}{2t}\left(1 - \sqrt{1 - \frac{4nt}{(\sigma^2+n+t)^2}}\right),$$

and, equivalently, as the unique solution to the equation

$$\mathcal{I}^{\mathrm{hd}} = \frac{1}{1 + \underbrace{\dfrac{\sigma^2 + t(1-\mathcal{I}^{\mathrm{hd}})}{n}}_{\textit{noise-to-signal ratio}}}.$$

Proof of Proposition 3 (High-dimensional information). Fix $\sigma^2 > 0$. For each $Z \geq 1$, write

$$N_Z \equiv N(Z), \qquad m_Z \equiv |\mathcal{T}(Z)|, \qquad \boldsymbol{X}_Z \equiv \boldsymbol{X}_{\mathcal{T}(Z)} \in \mathbb{R}^{N_Z \times m_Z}, \qquad \boldsymbol{S}_Z \equiv \boldsymbol{X}'_Z\boldsymbol{X}_Z \in \mathbb{R}^{m_Z \times m_Z}.$$

Under Assumption 1, $N_Z/Z \to n$ and $m_Z/Z \to t$, hence the aspect ratio satisfies

$$\frac{m_Z}{N_Z} \longrightarrow \frac{t}{n} \in (0,\infty) \qquad (\text{for } n > 0 \text{ and } t > 0).$$

Step 1: Rewrite information as a resolvent trace. By Lemma 2 and cyclicity of the trace,

$$\frac{1}{\tau^2}\,\mathcal{I}(N_Z,\mathcal{T}(Z);\sigma^2) = \frac{1}{m_Z}\,\mathbb{E}\Big[\operatorname{tr}\Big(\boldsymbol{S}_Z(\boldsymbol{S}_Z + \sigma^2\boldsymbol{I}_{m_Z})^{-1}\Big)\Big].$$

Use the identity $S(S+\sigma^2 I)^{-1} = I - \sigma^2(S+\sigma^2 I)^{-1}$ to get

$$\frac{1}{\tau^2}\,\mathcal{I}(N_Z, \mathcal{T}(Z); \sigma^2) = 1 - \frac{\sigma^2}{m_Z}\,\mathbb{E}\Big[\mathrm{tr}\Big((S_Z + \sigma^2 I_{m_Z})^{-1}\Big)\Big]. \tag{5}$$

Define the (random) Stieltjes transform of the empirical spectral distribution (ESD) of S_Z by

$$m_Z(z) \;\equiv\; \frac{1}{m_Z}\,\mathrm{tr}\left((S_Z - zI_{m_Z})^{-1}\right), \qquad z \in \mathbb{C}\setminus\mathbb{R}_+.$$

Then (5) becomes, with $z = -\sigma^2 < 0$,

$$\frac{1}{\tau^2}\,\mathcal{I}(N_Z, \mathcal{T}(Z); \sigma^2) = 1 - \sigma^2\,\mathbb{E}\big[m_Z(-\sigma^2)\big]. \tag{6}$$

Step 2: Apply the Marchenko–Pastur limit for the resolvent. By the DGP in Section 2.1, the entries of X_Z are i.i.d. Gaussian with variance

$$\mathrm{Var}(x_k^i) = \frac{1-\underline{\sigma}^2}{Z}.$$

Write $X_Z = \sqrt{\frac{1-\underline{\sigma}^2}{Z}}\,G_Z$ where G_Z has i.i.d. $\mathcal{N}(0,1)$ entries. Then

$$S_Z = X_Z' X_Z = \frac{1-\underline{\sigma}^2}{Z}\,G_Z' G_Z.$$

Let $W_Z \equiv \frac{1}{N_Z} G_Z' G_Z$ be the (normalized) sample covariance matrix. The classical Marchenko–Pastur theorem implies that the ESD of W_Z converges almost surely to the MP law with aspect ratio $c = \lim m_Z/N_Z = t/n$, and moreover the Stieltjes transform $m_{W_Z}(z)$ converges (almost surely, and also in expectation for fixed $z<0$) to the deterministic limit $m_{\mathrm{MP}}(z)$ characterized as the unique solution with $m_{\mathrm{MP}}(z) > 0$ for $z < 0$ of

$$m_{\mathrm{MP}}(z) = \frac{1}{-z + \frac{t}{n}\cdot\frac{1}{1+m_{\mathrm{MP}}(z)}}. \tag{7}$$

(See, e.g., standard random matrix references on sample covariance matrices and the Silverstein equation.)

Because $S_Z = \alpha_Z W_Z$ with $\alpha_Z \equiv (1-\underline{\sigma}^2)N_Z/Z \to (1-\underline{\sigma}^2)n$, the Stieltjes transforms are linked by the scaling relation

$$m_Z(z) = \frac{1}{m_Z}\,\mathrm{tr}\left((\alpha_Z W_Z - zI)^{-1}\right) = \frac{1}{\alpha_Z}\,m_{W_Z}(z/\alpha_Z).$$

Therefore, for fixed $\sigma^2 > 0$,

$$m_Z(-\sigma^2) \;\longrightarrow\; m(-\sigma^2) \;\equiv\; \frac{1}{\alpha}\,m_{\mathrm{MP}}\Big(-\frac{\sigma^2}{\alpha}\Big), \qquad \alpha \equiv (1-\underline{\sigma}^2)n.$$

Substituting this into (6) yields the existence of a deterministic limit

$$\lim_{Z\to\infty} \frac{1}{\tau^2}\, \mathcal{I}(N_Z, \mathcal{T}(Z); \sigma^2) = 1 - \sigma^2\, m(-\sigma^2) \;\equiv\; \mathcal{I}^{\mathrm{hd}}(n, t; \sigma^2) \in (0, 1).$$

Step 3: Derive the fixed-point characterization for $\mathcal{I}^{\mathrm{hd}}$. Let $m \equiv m(-\sigma^2) > 0$ denote the limiting resolvent trace at $z = -\sigma^2$. Define $\mathcal{I}^{\mathrm{hd}}$ by the limiting identity corresponding to (6):

$$\mathcal{I}^{\mathrm{hd}} \equiv 1 - \sigma^2 m.$$

Evaluating the Silverstein equation (7) at $z = -\sigma^2/\alpha$ and translating back to m yields an algebraic relation between m and (n, t, σ^2). Rewriting this relation in terms of $\mathcal{I}^{\mathrm{hd}} = 1 - \sigma^2 m$ gives the scalar fixed-point equation

$$\mathcal{I}^{\mathrm{hd}} = \frac{1}{1 + \frac{\sigma^2 + t(1-\mathcal{I}^{\mathrm{hd}})}{n}}. \tag{8}$$

(The term σ^2/n is the usual ridge noise-to-signal term, while $t(1-\mathcal{I}^{\mathrm{hd}})/n$ captures the effective high-dimensional inflation of estimation noise coming from fitting tZ parameters with nZ observations.)

Step 4: Solve the fixed point and select the admissible root. Equation (8) is equivalent to the quadratic

$$t(\mathcal{I}^{\mathrm{hd}})^2 - (n + t + \sigma^2)\mathcal{I}^{\mathrm{hd}} + n = 0,$$

whose two roots are

$$\mathcal{I}^{\mathrm{hd}}_{\pm} = \frac{n + t + \sigma^2 \pm \sqrt{(n + t + \sigma^2)^2 - 4nt}}{2t}.$$

Since $\mathcal{I}^{\mathrm{hd}} \in (0, 1)$ and $\mathcal{I}^{\mathrm{hd}}_{+} > 1$, the unique admissible solution is

$$\mathcal{I}^{\mathrm{hd}}(n, t; \sigma^2) = \frac{n + t + \sigma^2 - \sqrt{(n + t + \sigma^2)^2 - 4nt}}{2t} = \frac{\sigma^2 + n + t}{2t}\left(1 - \sqrt{1 - \frac{4nt}{(\sigma^2 + n + t)^2}}\right).$$

This matches the closed-form expression in the statement and establishes the limit.

□

Corollary 2 (Marginal Information of Covariates)**.**

$$\frac{\partial}{\partial t}\mathcal{I}^{\mathrm{eff}}(n, t) = \underbrace{\frac{\partial \mathcal{I}^{\mathrm{hd}}}{\partial t}\big(n, t; \sigma^2(t)\big)}_{<0\,:\,\textit{dimensionality}} + \underbrace{\frac{\partial \mathcal{I}^{\mathrm{hd}}}{\partial \sigma^2}\big(n, t; \sigma^2(t)\big) \cdot \sigma_t^2(t)}_{>0\,:\,\textit{residual noise reduction}} > 0 \iff \mathcal{I}^{\mathrm{eff}}(n, t) > \underline{\sigma}^2 \iff n \geq \frac{\underline{\sigma}^2}{1 - \underline{\sigma}^2}.$$

Furthermore, if $\frac{\partial}{\partial t}\mathcal{I}^{\mathrm{eff}}(n,t) > 0$*, then*

$$\frac{\partial^2}{\partial t^2}\mathcal{I}^{\mathrm{eff}}(n,t) = \underbrace{\frac{\partial^2 \mathcal{I}^{\mathrm{hd}}}{\partial t^2}\big(n,t;\sigma^2(t)\big)}_{\text{direct (dimensionality)}} + 2\,\underbrace{\frac{\partial^2 \mathcal{I}^{\mathrm{hd}}}{\partial t\,\partial \sigma^2}\big(n,t;\sigma^2(t)\big)}_{\text{interaction}}\,\sigma_t^2(t) + \underbrace{\frac{\partial^2 \mathcal{I}^{\mathrm{hd}}}{\partial (\sigma^2)^2}\big(n,t;\sigma^2(t)\big)}_{>0\,:\,\text{curvature in noise}}\,\big(\sigma_t^2(t)\big)^2 > 0$$

Proof of Corollary 2. Fix $(n,t) \in (0,\infty)\times(0,1]$ and write $\underline{\sigma}^2 \equiv \underline{\sigma}^2 \in (0,1)$. Recall the *effective information*

$$\mathcal{I}^{\mathrm{eff}}(n,t) \;\equiv\; \mathcal{I}^{\mathrm{hd}}\big(n,t;\sigma^2(t)\big), \qquad \sigma^2(t) = \underline{\sigma}^2 + (1-\underline{\sigma}^2)(1-t) = 1-(1-\underline{\sigma}^2)t.$$

Throughout the proof, abbreviate

$$I \;\equiv\; \mathcal{I}^{\mathrm{eff}}(n,t) \in (0,1).$$

Step 1: A convenient implicit equation for I. From Proposition 3, $\mathcal{I}^{\mathrm{hd}}$ is the unique solution in $(0,1)$ to

$$I = \frac{1}{1+\frac{\sigma^2+t(1-I)}{n}}.$$

Plugging $\sigma^2 = \sigma^2(t)$ yields

$$I = \frac{1}{1+\frac{\sigma^2(t)+t(1-I)}{n}} = \frac{1}{1+\frac{\underline{\sigma}^2+(1-\underline{\sigma}^2)(1-t)+t(1-I)}{n}} = \frac{1}{1+\frac{1-\underline{\sigma}^2\,t+t(1-I)}{n}}.$$

Equivalently, I is the unique root in $(0,1)$ of the quadratic equation

$$F(I,t) \equiv tI^2 - (n+1+\underline{\sigma}^2 t)I + n = 0. \tag{9}$$

(Indeed, multiplying out the fixed point and collecting terms yields (9).)

Step 2: First derivative and sign. Differentiate (9) implicitly with respect to t:

$$F_I(I,t)\,I_t + F_t(I,t) = 0 \quad \Rightarrow \quad I_t = -\frac{F_t}{F_I}.$$

Compute the partial derivatives:

$$F_t(I,t) = I^2 - \underline{\sigma}^2 I = I(I-\underline{\sigma}^2), \qquad F_I(I,t) = 2tI - (n+1+\underline{\sigma}^2 t).$$

Hence

$$\frac{\partial}{\partial t}\mathcal{I}^{\mathrm{eff}}(n,t) = I_t = -\frac{I(I-\underline{\sigma}^2)}{2tI-(n+1+\underline{\sigma}^2 t)}. \tag{10}$$

We now sign the denominator. The polynomial $F(\cdot,t)$ is strictly convex in I because its leading coefficient is $t > 0$. Moreover, $F(0,t) = n > 0$ and $F(1,t) = t-(n+1+\underline{\sigma}^2 t)+n =$

$(1-\underline{\sigma}^2)t - 1 < 0$ for all $t \in (0,1]$, so there is a unique root in $(0,1)$, which is I. For a strictly convex function, the derivative at its (leftmost) root is negative when $F(0,t) > 0$ and $F(1,t) < 0$, hence

$$F_I(I,t) < 0.$$

Using (10) and $F_I(I,t) < 0$, we obtain

$$I_t > 0 \quad \Longleftrightarrow \quad I(I-\underline{\sigma}^2) > 0 \quad \Longleftrightarrow \quad I > \underline{\sigma}^2.$$

This proves the first equivalence:

$$\frac{\partial}{\partial t}\mathcal{I}^{\text{eff}}(n,t) > 0 \quad \Longleftrightarrow \quad \mathcal{I}^{\text{eff}}(n,t) > \underline{\sigma}^2.$$

Step 3: Threshold condition $I > \underline{\sigma}^2 \iff n \geq \underline{\sigma}^2/(1-\underline{\sigma}^2)$. Define the fixed-point map (with effective noise)

$$g(I) \equiv \frac{1}{1+\frac{\sigma^2(t)+t(1-I)}{n}},$$

so that I is the unique fixed point of g on $(0,1)$. The map g is strictly increasing in I. Therefore,

$$I > \underline{\sigma}^2 \quad \Longleftrightarrow \quad g(\underline{\sigma}^2) > \underline{\sigma}^2.$$

Compute $g(\underline{\sigma}^2)$ using the key cancellation:

$$\sigma^2(t) + t(1-\underline{\sigma}^2) = \underline{\sigma}^2 + (1-\underline{\sigma}^2)(1-t) + t(1-\underline{\sigma}^2) = \underline{\sigma}^2 + (1-\underline{\sigma}^2) = 1.$$

Hence

$$g(\underline{\sigma}^2) = \frac{1}{1+\frac{1}{n}} = \frac{n}{n+1}.$$

Thus,

$$I > \underline{\sigma}^2 \quad \Longleftrightarrow \quad \frac{n}{n+1} > \underline{\sigma}^2 \quad \Longleftrightarrow \quad n > \underline{\sigma}^2(n+1) \quad \Longleftrightarrow \quad n(1-\underline{\sigma}^2) > \underline{\sigma}^2 \quad \Longleftrightarrow \quad n > \frac{\underline{\sigma}^2}{1-\underline{\sigma}^2}.$$

This proves the second equivalence (weak inequality versions follow by continuity):

$$\mathcal{I}^{\text{eff}}(n,t) > \underline{\sigma}^2 \quad \Longleftrightarrow \quad n \geq \frac{\underline{\sigma}^2}{1-\underline{\sigma}^2}.$$

Step 4: Second derivative when $I_t > 0$. Differentiate $F_I(I,t)I_t + F_t(I,t) = 0$ once more with respect to t:

$$F_I I_{tt} + F_{II} I_t^2 + 2F_{tI} I_t + F_{tt} = 0 \quad \Rightarrow \quad I_{tt} = -\frac{F_{II} I_t^2 + 2F_{tI} I_t + F_{tt}}{F_I}.$$

Compute the remaining partial derivatives from (9):

$$F_{II}(I,t) = 2t, \qquad F_{tI}(I,t) = 2I - \underline{\sigma}^2, \qquad F_{tt}(I,t) = 0.$$

Therefore,

$$I_{tt} = -\frac{2tI_t^2 + 2(2I - \underline{\sigma}^2)I_t}{F_I} = -\frac{2I_t\big(tI_t + 2I - \underline{\sigma}^2\big)}{F_I}. \tag{11}$$

We already established $F_I(I,t) < 0$. If $I_t > 0$ (equivalently $I > \underline{\sigma}^2$), then

$$2I - \underline{\sigma}^2 > 0 \quad \text{and} \quad tI_t > 0,$$

so the bracket in (11) is strictly positive. Hence the numerator in (11) is strictly negative, and dividing by $F_I < 0$ implies

$$I_{tt} > 0.$$

This proves the claimed convexity of $\mathcal{I}^{\text{eff}}(n,t)$ in t conditional on $\partial\mathcal{I}^{\text{eff}}/\partial t > 0$.

Step 5: Chain-rule decomposition (as stated in the corollary). Finally, the decomposition

$$\frac{\partial}{\partial t}\mathcal{I}^{\text{eff}}(n,t) = \frac{\partial\mathcal{I}^{\text{hd}}}{\partial t}\big(n,t;\sigma^2(t)\big) + \frac{\partial\mathcal{I}^{\text{hd}}}{\partial\sigma^2}\big(n,t;\sigma^2(t)\big)\cdot\sigma_t^2(t)$$

is immediate from the chain rule. Moreover, $\frac{\partial\mathcal{I}^{\text{hd}}}{\partial t} < 0$ and $\frac{\partial\mathcal{I}^{\text{hd}}}{\partial\sigma^2} < 0$ (following from implicit differentiation of the fixed point holding the other argument fixed), while $\sigma_t^2(t) = -(1-\underline{\sigma}^2) < 0$, so the second term is strictly positive.

Collecting Steps 2–4 yields the result. □

Proof of Corollary 3. Fix $(n,t) \in (0,\infty)\times(0,1]$ and write

$$I \equiv \mathcal{I}^{\text{eff}}(n,t) = \mathcal{I}^{\text{hd}}\big(n,t;\sigma^2(t)\big) \in (0,1).$$

As in the previous corollary, I is characterized implicitly by

$$F(I,n,t) \equiv tI^2 - (n+1+\underline{\sigma}^2 t)I + n = 0, \tag{12}$$

where $\underline{\sigma}^2 \in (0,1)$ is fixed.

Step 1: First derivatives. Differentiate (18) with respect to n and t:

$$F_I I_n + F_n = 0, \qquad F_I I_t + F_t = 0,$$

so

$$I_n = -\frac{F_n}{F_I}, \qquad I_t = -\frac{F_t}{F_I}. \tag{13}$$

Compute the partial derivatives:

$$F_I = 2tI - (n + 1 + \underline{\sigma}^2 t), \qquad F_n = 1 - I, \qquad F_t = I^2 - \underline{\sigma}^2 I = I(I - \underline{\sigma}^2).$$

In particular, under the maintained condition $I_t > 0$ we have $F_I < 0$ (as shown previously), hence (19) is well-defined.

Step 2: A useful ratio identity. From (19) and the expressions for F_n and F_t,

$$\frac{I_n}{I_t} = \frac{F_n}{F_t} = \frac{1 - I}{I(I - \underline{\sigma}^2)}. \tag{14}$$

This implies

$$I_n = I_t \cdot \frac{1 - I}{I(I - \underline{\sigma}^2)}. \tag{15}$$

Step 3: Cross-partial I_{nt}. Differentiate $I_n = -F_n/F_I$ with respect to t (holding n fixed):

$$I_{nt} = -\frac{(F_n)_t F_I - F_n (F_I)_t}{F_I^2}.$$

Compute the needed derivatives:

$$(F_n)_t = -(I_t), \qquad (F_I)_t = 2I + 2tI_t - \underline{\sigma}^2.$$

Therefore,

$$\begin{aligned} I_{nt} &= -\frac{(-I_t)F_I - (1 - I)\big(2I + 2tI_t - \underline{\sigma}^2\big)}{F_I^2} \\ &= \frac{I_t F_I + (1 - I)\big(2I + 2tI_t - \underline{\sigma}^2\big)}{F_I^2}. \end{aligned} \tag{16}$$

Step 4: Eliminate F_I and t using the implicit system. Use $F_I I_t + F_t = 0$, i.e.

$$F_I = -\frac{F_t}{I_t} = -\frac{I(I - \underline{\sigma}^2)}{I_t}.$$

Plug this into the first term of the numerator in (22):

$$I_t F_I = -I(I - \underline{\sigma}^2).$$

Next, use the identity obtained by rearranging (18):

$$tI^2 - \underline{\sigma}^2 tI = (n + 1)I - n \quad \Rightarrow \quad tI(I - \underline{\sigma}^2) = (n + 1)I - n = n(I - 1) + I.$$

Hence,

$$1 - \underline{\sigma}^2 = \frac{1 - \underline{\sigma}^2}{1} \quad \text{(kept as primitive)},$$

and more importantly we can express t only through I, n inside combinations of the form $tI(I - \underline{\sigma}^2)$. Now rewrite the remaining piece of the numerator in (22) as

$$(1 - I)\big(2I - \underline{\sigma}^2\big) + 2t(1 - I)I_t.$$

Use (21) to express $(1 - I)$ times I_t in terms of I_n:

$$(1 - I)I_t = I_n\, I(I - \underline{\sigma}^2).$$

Therefore,

$$2t(1 - I)I_t = 2t\, I_n\, I(I - \underline{\sigma}^2) = 2I_n\big(n(I - 1) + I\big),$$

where we used $tI(I - \underline{\sigma}^2) = n(I - 1) + I$.

Substituting these simplifications into (22) yields

$$I_{nt} = \frac{-I(I - \underline{\sigma}^2) + (1 - I)(2I - \underline{\sigma}^2) + 2I_n\big(n(I - 1) + I\big)}{F_I^2}.$$

Collect the first two terms:

$$-I(I - \underline{\sigma}^2) + (1 - I)(2I - \underline{\sigma}^2) = 2I(1 - I) - \underline{\sigma}^2.$$

Also note that $n(I - 1) + I = I - n(1 - I)$, so

$$2I_n\big(n(I - 1) + I\big) = 2I_n\big(I - n(1 - I)\big).$$

Now use $I_n = (1 - I)/(-F_I)$ from (19) to eliminate I_n and F_I:

$$2I_n\big(I - n(1 - I)\big) = \frac{2(1 - I)}{-F_I}\big(I - n(1 - I)\big).$$

Substituting and simplifying shows that all $\underline{\sigma}^2$-terms cancel, and one obtains the compact expression

$$I_{nt} = \frac{I_t}{n}\,(1 - 2I). \tag{17}$$

(Algebraic simplification is routine once everything is expressed in terms of I, I_t, n, and F_I, and the identity $tI(I - \underline{\sigma}^2) = n(I - 1) + I$ is used to remove t.)

Step 5: Sign. Under the maintained condition $I_t > 0$ and since $n > 0$, (23) implies

$$I_{nt} > 0 \quad \Longleftrightarrow \quad 1 - 2I > 0 \quad \Longleftrightarrow \quad I < \frac{1}{2},$$

which is the desired condition.

This proves the corollary. □

Corollary 3 (Interaction of Observations and Covariates). *If* $\mathcal{I}_t^{eff}(n,t) > 0$*, then*

$$\frac{\partial^2}{\partial n \partial t}\mathcal{I}^{\text{eff}}(n,t) = \frac{\mathcal{I}_t^{eff}(n,t)}{n}(1 - 2\mathcal{I}^{\text{eff}}(n,t)) > 0 \iff \mathcal{I}^{\text{eff}}(n,t) < \frac{1}{2}$$

Proof of Corollary 3. Fix $(n,t) \in (0,\infty) \times (0,1]$ and write

$$I \equiv \mathcal{I}^{\text{eff}}(n,t) = \mathcal{I}^{\text{hd}}\big(n,t;\sigma^2(t)\big) \in (0,1).$$

As in the previous corollary, I is characterized implicitly by

$$F(I,n,t) \equiv tI^2 - (n+1+\underline{\sigma}^2 t)I + n = 0, \tag{18}$$

where $\underline{\sigma}^2 \in (0,1)$ is fixed.

Step 1: First derivatives. Differentiate (18) with respect to n and t:

$$F_I I_n + F_n = 0, \qquad F_I I_t + F_t = 0,$$

so

$$I_n = -\frac{F_n}{F_I}, \qquad I_t = -\frac{F_t}{F_I}. \tag{19}$$

Compute the partial derivatives:

$$F_I = 2tI - (n+1+\underline{\sigma}^2 t), \qquad F_n = 1 - I, \qquad F_t = I^2 - \underline{\sigma}^2 I = I(I - \underline{\sigma}^2).$$

In particular, under the maintained condition $I_t > 0$ we have $F_I < 0$ (as shown previously), hence (19) is well-defined.

Step 2: A useful ratio identity. From (19) and the expressions for F_n and F_t,

$$\frac{I_n}{I_t} = \frac{F_n}{F_t} = \frac{1-I}{I(I-\underline{\sigma}^2)}. \tag{20}$$

This implies

$$I_n = I_t \cdot \frac{1-I}{I(I-\underline{\sigma}^2)}. \tag{21}$$

Step 3: Cross-partial I_{nt}. Differentiate $I_n = -F_n/F_I$ with respect to t (holding n fixed):

$$I_{nt} = -\frac{(F_n)_t F_I - F_n (F_I)_t}{F_I^2}.$$

Compute the needed derivatives:

$$(F_n)_t = -(I_t), \qquad (F_I)_t = 2I + 2tI_t - \underline{\sigma}^2.$$

Therefore,

$$\begin{aligned} I_{nt} &= -\frac{(-I_t)F_I - (1-I)\big(2I + 2tI_t - \underline{\sigma}^2\big)}{F_I^2} \\ &= \frac{I_t F_I + (1-I)\big(2I + 2tI_t - \underline{\sigma}^2\big)}{F_I^2}. \end{aligned} \tag{22}$$

Step 4: Eliminate F_I and t using the implicit system. Use $F_I I_t + F_t = 0$, i.e.

$$F_I = -\frac{F_t}{I_t} = -\frac{I(I - \underline{\sigma}^2)}{I_t}.$$

Plug this into the first term of the numerator in (22):

$$I_t F_I = -I(I - \underline{\sigma}^2).$$

Next, use the identity obtained by rearranging (18):

$$tI^2 - \underline{\sigma}^2 tI = (n+1)I - n \quad \Rightarrow \quad tI(I - \underline{\sigma}^2) = (n+1)I - n = n(I-1) + I.$$

Hence,

$$1 - \underline{\sigma}^2 = \frac{1 - \underline{\sigma}^2}{1} \quad \text{(kept as primitive)},$$

and more importantly we can express t only through I, n inside combinations of the form $tI(I - \underline{\sigma}^2)$. Now rewrite the remaining piece of the numerator in (22) as

$$(1-I)\big(2I - \underline{\sigma}^2\big) + 2t(1-I)I_t.$$

Use (21) to express $(1-I)$ times I_t in terms of I_n:

$$(1-I)I_t = I_n\, I(I - \underline{\sigma}^2).$$

Therefore,

$$2t(1-I)I_t = 2t\, I_n\, I(I - \underline{\sigma}^2) = 2I_n\big(n(I-1) + I\big),$$

where we used $tI(I - \underline{\sigma}^2) = n(I-1) + I$.

Substituting these simplifications into (22) yields

$$I_{nt} = \frac{-I(I - \underline{\sigma}^2) + (1-I)(2I - \underline{\sigma}^2) + 2I_n\big(n(I-1) + I\big)}{F_I^2}.$$

Collect the first two terms:

$$-I(I-\underline{\sigma}^2)+(1-I)(2I-\underline{\sigma}^2)=2I(1-I)-\underline{\sigma}^2.$$

Also note that $n(I-1)+I=I-n(1-I)$, so

$$2I_n\big(n(I-1)+I\big)=2I_n\big(I-n(1-I)\big).$$

Now use $I_n=(1-I)/(-F_I)$ from (19) to eliminate I_n and F_I:

$$2I_n\big(I-n(1-I)\big)=\frac{2(1-I)}{-F_I}\big(I-n(1-I)\big).$$

Substituting and simplifying shows that all $\underline{\sigma}^2$-terms cancel, and one obtains the compact expression

$$I_{nt}=\frac{I_t}{n}\,(1-2I). \tag{23}$$

(Algebraic simplification is routine once everything is expressed in terms of I, I_t, n, and F_I, and the identity $tI(I-\underline{\sigma}^2)=n(I-1)+I$ is used to remove t.)

Step 5: Sign. Under the maintained condition $I_t>0$ and since $n>0$, (23) implies

$$I_{nt}>0 \quad\Longleftrightarrow\quad 1-2I>0 \quad\Longleftrightarrow\quad I<\frac{1}{2},$$

which is the desired condition.

This proves the corollary. □

Proposition 4 (Low-dimensional Value of Design)**.** *Fix $Z\geq 1$ and a dataset design $\mathcal{D}(Z)=(N(Z),\mathcal{T}(Z),\mathcal{P}(Z))$ with $\mathcal{P}(Z)\subseteq\mathcal{T}(Z)$. Then the value of the design satisfies*

$$V(\mathcal{D}(Z))=\underbrace{|\mathcal{P}(Z)\cap\mathcal{T}(Z)|\times\frac{1-\underline{\sigma}^2}{Z}}_{\textit{Within-individual signal}}\times\underbrace{\mathcal{I}\big(N(Z),\mathcal{T}(Z);\sigma^2(\mathcal{T}(Z))\big)}_{\textit{Across-individual information}},$$

where $\sigma^2(\mathcal{T}(Z))$ is the relative regression noise induced by the training set.

Proof of Proposition 4. Fix $Z\geq 1$ and a design $\mathcal{D}(Z)=(N,\mathcal{T},\mathcal{P})$ with $\mathcal{P}\subseteq\mathcal{T}$. Write $m\equiv|\mathcal{T}|$ and $q\equiv|\mathcal{P}|=|\mathcal{P}\cap\mathcal{T}|$.

Step 1: Start from the definition of design value. By definition,

$$V(\mathcal{D}(Z))=\mathbb{E}_D[v(\boldsymbol{D})]\,,\qquad v(\boldsymbol{D})=\mathrm{Var}[y]-\mathrm{Var}[y\mid\boldsymbol{D}].$$

Under the Gaussian-linear model and quadratic loss, Proposition 2 gives

$$v(\boldsymbol{D};\sigma^2(\mathcal{T})) = \boldsymbol{x}_{\mathcal{P}}' \left(\tau^2 \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \sigma^2(\mathcal{T}) \boldsymbol{I}_m \right)^{-1} \right)_{\mathcal{P}} \boldsymbol{x}_{\mathcal{P}}, \tag{24}$$

where $\sigma^2(\mathcal{T})$ is the relative regression noise induced by $\mathcal{T}$ and we used $\mathcal{P} \subseteq \mathcal{T}$ to write the relevant block simply as $(\cdot)_{\mathcal{P}}$.

Taking expectation under the design,

$$V(\boldsymbol{D}) = \mathbb{E}_D \left[\boldsymbol{x}_{\mathcal{P}}' \, \boldsymbol{M}(\boldsymbol{X}_{\mathcal{T}})_{\mathcal{P}} \, \boldsymbol{x}_{\mathcal{P}} \right], \qquad \boldsymbol{M}(\boldsymbol{X}_{\mathcal{T}}) \equiv \tau^2 \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \sigma^2(\mathcal{T}) \boldsymbol{I}_m \right)^{-1}. \tag{25}$$

Step 2: Integrate out the prediction covariates $\boldsymbol{x}_{\mathcal{P}}$. Under the DGP in Section 2.1, covariates are independent across individuals and independent of $(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})$. In particular, conditional on $\boldsymbol{X}_{\mathcal{T}}$,

$$\boldsymbol{x}_{\mathcal{P}} \perp \boldsymbol{X}_{\mathcal{T}}, \qquad \mathbb{E}[\boldsymbol{x}_{\mathcal{P}}] = \boldsymbol{0}, \qquad \mathrm{Var}(\boldsymbol{x}_{\mathcal{P}}) = \frac{1 - \underline{\sigma}^2}{Z} \boldsymbol{I}_q.$$

Therefore, conditional on $\boldsymbol{X}_{\mathcal{T}}$,

$$\mathbb{E}\left[\boldsymbol{x}_{\mathcal{P}}' \, \boldsymbol{M}(\boldsymbol{X}_{\mathcal{T}})_{\mathcal{P}} \, \boldsymbol{x}_{\mathcal{P}} \,\middle|\, \boldsymbol{X}_{\mathcal{T}} \right] = \mathrm{tr}(\boldsymbol{M}(\boldsymbol{X}_{\mathcal{T}})_{\mathcal{P}} \, \mathrm{Var}(\boldsymbol{x}_{\mathcal{P}})) = \frac{1 - \underline{\sigma}^2}{Z} \, \mathrm{tr}(\boldsymbol{M}(\boldsymbol{X}_{\mathcal{T}})_{\mathcal{P}}).$$

Taking unconditional expectation over $\boldsymbol{X}_{\mathcal{T}}$ gives

$$V(\boldsymbol{D}) = \frac{1 - \underline{\sigma}^2}{Z} \, \mathbb{E}_D[\mathrm{tr}(\boldsymbol{M}(\boldsymbol{X}_{\mathcal{T}})_{\mathcal{P}})]. \tag{26}$$

Step 3: Use exchangeability across covariates to reduce to an average diagonal element. Because the columns of $\boldsymbol{X}_{\mathcal{T}}$ are i.i.d. and the prior is symmetric across indices, the random matrix

$$\boldsymbol{A} \equiv \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \sigma^2(\mathcal{T}) \boldsymbol{I}_m \right)^{-1} \in \mathbb{R}^{m \times m}$$

is invariant in distribution under any permutation of coordinates. In particular, for any $k \in \mathcal{T}$, $\mathbb{E}[V_{kk}]$ is the same constant, and thus for any fixed subset $\mathcal{P} \subseteq \mathcal{T}$ of size q,

$$\mathbb{E}[\mathrm{tr}(\boldsymbol{V}_{\mathcal{P}})] = \sum_{k \in \mathcal{P}} \mathbb{E}[V_{kk}] = \frac{q}{m} \sum_{k \in \mathcal{T}} \mathbb{E}[V_{kk}] = \frac{q}{m} \, \mathbb{E}[\mathrm{tr}(\boldsymbol{A})].$$

Since $\boldsymbol{M} = \tau^2 \boldsymbol{A}$, we obtain from (26)

$$V(\boldsymbol{D}) = \frac{1 - \underline{\sigma}^2}{Z} \, \tau^2 \, \frac{q}{m} \, \mathbb{E}_D \left[\mathrm{tr}\left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \sigma^2(\mathcal{T}) \boldsymbol{I}_m \right)^{-1} \right) \right]. \tag{27}$$

Step 4: Recognize the training information. By Definition (Training Information) and Lemma 2,

$$\mathcal{I}\left(N, \mathcal{T}; \sigma^2(\mathcal{T})\right) = \frac{\tau^2}{m}\ \mathrm{tr}\left[\mathbb{E}_D\left[X_{\mathcal{T}}' X_{\mathcal{T}}\left(X_{\mathcal{T}}' X_{\mathcal{T}} + \sigma^2(\mathcal{T}) I_m\right)^{-1}\right]\right] = \frac{\tau^2}{m}\,\mathbb{E}_D[\mathrm{tr}(A)]\,.$$

Substituting this into (27) yields

$$V(\mathcal{D}) = \frac{1-\underline{\sigma}^2}{Z}\, q\, \mathcal{I}\left(N, \mathcal{T}; \sigma^2(\mathcal{T})\right).$$

Since $q = |\mathcal{P} \cap \mathcal{T}|$ (and $\mathcal{P} \subseteq \mathcal{T}$ by assumption), this is exactly

$$V(\mathcal{D}(Z)) = |\mathcal{P}(Z) \cap \mathcal{T}(Z)| \times \frac{1-\underline{\sigma}^2}{Z}\ \times\ \mathcal{I}\left(N(Z), \mathcal{T}(Z); \sigma^2(\mathcal{T}(Z))\right),$$

which proves the proposition. □

Lemma 3 (Observation-rich approximation for value)**.** *Fix $p \in (0,1]$ and suppose that $t/n \to 0$. Then, with endogenous residual variance $\sigma^2(t) = 1 - t(1-\underline{\sigma}^2)$,*

$$\bar{V}(n,t,p) = (1-\underline{\sigma}^2)\,\frac{\min\{t,p\}}{1+\sigma^2(t)/n} + o(1). \tag{1}$$

Proof. By Proposition 3, $\mathcal{I}^{\mathrm{hd}}$ is the unique solution in $(0,1)$ to

$$\mathcal{I}^{\mathrm{hd}} = \frac{1}{1 + \frac{\sigma^2}{n} + \frac{t}{n}\left(1 - \mathcal{I}^{\mathrm{hd}}\right)}.$$

Since $0 < \mathcal{I}^{\mathrm{hd}} < 1$, the estimation-noise term satisfies

$$0 \le \frac{t}{n}\left(1 - \mathcal{I}^{\mathrm{hd}}\right) \le \frac{t}{n} = o(1),$$

so

$$\mathcal{I}^{\mathrm{hd}}(n,t;\sigma^2) = \frac{1}{1+\sigma^2/n + o(1)} = \frac{1}{1+\sigma^2/n} + o(1).$$

Substituting $\sigma^2 = \sigma^2(t)$ and the value formula in Theorem 1,

$$\bar{V}(n,t,p) = (1-\underline{\sigma}^2)\ \min\{t,p\}\,\mathcal{I}^{\mathrm{hd}}(n,t;\sigma^2(t)),$$

yields (1). □

Proposition 5 (Harmful Acquisition)**.** *The incumbent I always acquires the entrant E.*

Proof. If the acquisition does not occur, each firm $i \in \{I, E\}$ sets $p_i = V_1$ and sells to its unit

mass of captive buyers, so profits are $\Pi_I^{\text{no}} = \Pi_E^{\text{no}} = V_1$ and

$$W^{\text{no}} = \Pi_I^{\text{no}} + \Pi_E^{\text{no}} = 2V_1.$$

If the acquisition occurs, the merged firm sets $p_2 = V_2$ and earns profit $\Pi_2 = V_2$, while the planner incurs the dynamic loss ξ. Hence

$$W^{\text{acq}} = \Pi_2 - \xi = V_2 - \xi.$$

The acquisition is (weakly) harmful whenever $W^{\text{acq}} \leq W^{\text{no}}$, that is,

$$V_2 - \xi \;\leq\; 2V_1 \quad \Longleftrightarrow \quad V_2 - 2V_1 \;\leq\; \xi.$$

Under the symmetry and equal-informativeness assumptions, Lemma 3 implies that

$$V_2 - 2V_1 \;=\; \Delta(n,k) = \frac{n}{\left(\frac{1+n}{k} - 1\right)\left(\frac{2(n+1)}{k} - 1\right)},$$

which is increasing in k. Define $\tilde{k}(n,\xi)$ as the (unique) solution to $\Delta(n,k) = \xi$. By monotonicity, $V_2 - 2V_1 \leq \xi$ holds if and only if $k \leq \tilde{k}(n,\xi)$, and solving $\Delta(n,k) = \xi$ for k yields

$$\tilde{k}(n,\xi) = \frac{n+1}{2}\,\frac{\sqrt{\xi(8n+\xi)} - 3\xi}{n-\xi}.$$

This proves the claim. □

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

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

Proof. The pool price is

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

which can be solved using the first-order condition

$$D\,(P - V_2) + PD'\,(P - V_2) = 0.$$

□

Lemma 6 (Sample Fragmentation Price)**.** *If the brokers have different observations on the same covariates,*

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

and the buyers will buy from both brokers.

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

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

Second, the user adopts the technology if and only if

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

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

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

which has FOC

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

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

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

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

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

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

With our CDF G,

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

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

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

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

Proposition 8 (Welfare of Observation Pooling). *A pool of observations is procompetitive if and only if*

$$k < \hat{k}(n, \alpha) \equiv 1 - \frac{n}{2\alpha}.$$

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

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

In this case

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

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

Otherwise if the competition margin binds,

$$p_i = V_2 - V_1.$$

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

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

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

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

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

Lemma 5 (Covariate Fragmentation Price)**.** *If the brokers have distinct covariates each B_i prices at*

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

and the buyers will buy from both brokers.

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

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

□

Lemma 7 (Shopper-price equilibrium)**.** *Fix* (V_1, V_2)*. The pricing subgame in shopp[Shopper-price equilibrium]er prices admits a unique equilibrium under trembling-hand perfection:*

$$p_i^s = (V_i - V_j)^+, \qquad i \neq j, \tag{3}$$

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

Proof. We first show that (3) is an equilibrium. Consider $V_i > V_j$. If i sets $p_i^s = V_i - V_j$ and j sets $p_j^s = 0$, shoppers are indifferent and (by the refinement) all go to i.[12] Any deviation by i:

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

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

Any deviation by j:

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

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

Thus (3) is an equilibrium when $V_i > V_j$. The case $V_i = V_j$: for any $p_i^s = p_j^s$, shoppers are indifferent and each firm earns $\sigma p_i^s/2$ from shoppers; any unilateral increase loses all shoppers, any decrease reduces price with the same demand, so $p_i^s = p_j^s = 0$ is the unique trembling-hand limit (positive common prices are not robust to small payoff perturbations). Symmetry covers $V_j > V_i$.

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

Proposition 11. *If* $c < \bar{c}$*, the planner prefers real non-exclusivity to de facto exclusivity.*

Proof.

$$W^e \equiv \frac{1+s}{2}\hat{V} - c < W^{ne} \equiv \hat{V} - 2c \iff c < \frac{1-s}{2}\hat{V}.$$

which contradicts $c < \bar{c} \equiv \frac{1-s}{2}\hat{V}$. □

Proposition 12. *We distinguish two subcases:*

- ***De facto exclusivity:*** *Without loss of generality, let* $\ell_1 = 1$*,* $\ell_2 = 0$*, the unique NE is*

$$r_1 = 1, \qquad r_2 = 0.$$

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

- ***Both firms license:*** $\ell_1 = \ell_2 = 1$. *The unique symmetric mixed strategy equilibrium has each firm investing with probability*

$$\xi^* = \frac{V(1+s) - 2c}{2As}, \qquad \xi^* \in [0,1].$$

In equilibrium, profits are zero: $\Pi_i^{\text{mix}} = 0$.

Proof. We condition throughout on the licensing profile (ℓ_1, ℓ_2) and characterize the equilibrium in proprietary investment $r_i \in \{0,1\}$.

De facto exclusivity. Suppose without loss of generality that $\ell_1 = 1$ and $\ell_2 = 0$. Firm 2 cannot profitably invest in proprietary covariates, so its best response is

$$r_2 = 0.$$

If firm 1 invests, it obtains monopoly access to shoppers. By construction of the threshold $\bar{c}$, monopoly investment yields strictly positive net profit when $c > \bar{c}$, whereas not investing yields zero. Hence firm 1 strictly prefers to invest:

$$r_1 = 1.$$

Thus $(r_1, r_2) = (1,0)$ is the unique Nash equilibrium in proprietary investment when $(\ell_1, \ell_2) = (1,0)$.

Both firms license. Now suppose $\ell_1 = \ell_2 = 1$. Let ξ denote the probability with which firm $j \neq i$ invests, so that firm i's expected payoff from choosing $r_i = 1$ is

$$\begin{aligned}\Pi_i(r_i = 1) &= (1-\xi)\left[\frac{1-s}{2}\hat{V} + s\hat{V}\right] + \xi\left[\frac{1-s}{2}\hat{V}\right] - c \\ &= \frac{1-s}{2}\hat{V} + s\hat{V}(1-\xi) - c,\end{aligned}$$

while the payoff from not investing is

$$\Pi_i(r_i = 0) = 0.$$

First, no symmetric pure-strategy equilibrium exists. If $(r_1, r_2) = (0,0)$, then $\xi = 0$ and

$$\Pi_i(1) = \frac{1-s}{2}\hat{V} + s\hat{V} - c > 0$$

by the definition of $\bar{c}$, so each firm has a profitable deviation to $r_i = 1$. If $(r_1, r_2) = (1,1)$, then $\xi = 1$ and

$$\Pi_i(1) = \frac{1-s}{2}\hat{V} - c < 0$$

for $c > \bar{c}$, so each firm has a profitable deviation to $r_i = 0$. Hence neither $(0,0)$ nor $(1,1)$ is a symmetric equilibrium.

Consider now a symmetric mixed-strategy equilibrium in which each firm invests with probability $\xi^* \in (0,1)$. Symmetry requires that firm i be indifferent between $r_i = 1$ and $r_i = 0$, so

$$\Pi_i(r_i = 1) = \Pi_i(r_i = 0) = 0.$$

Substituting the expression for $\Pi_i(r_i = 1)$ and solving for ξ yields

$$\frac{1-s}{2}\hat{V} + s\hat{V}(1-\xi^*) - c = 0 \quad \Longrightarrow \quad \xi^* = \frac{\hat{V}(1+s) - 2c}{2\hat{V}s}.$$

If we are in the high-cost region where $c \in [\bar{c}, \bar{\bar{c}})$, with $\bar{c} \equiv \frac{1-s}{2}\hat{V}$ and $\bar{\bar{c}} \equiv \frac{1+s}{2}\hat{V}$, this mixing probability satisfies $\xi^* \in [0,1]$. Moreover, because $\Pi_i(r_i = 1)$ is strictly decreasing in ξ, the indifference condition can hold for at most one value of ξ, so the symmetric mixed strategy equilibrium is unique.

Finally, since firm i is indifferent between $r_i = 1$ and $r_i = 0$ at ξ^*, its expected equilibrium profit (before paying the license fee) equals the payoff from not investing:

$$\Pi_i^{\text{mix}} = 0.$$

This proves the two cases stated in the proposition. □

Proposition 13. *If $c \in (\bar{c}, \bar{\bar{c}})$, the planner prefers de facto exclusivity to real non-exclusivity.*

Proof. Under exclusivity:

$$W^e = \frac{1+s}{2}\hat{V} - c.$$

Under non-exclusivity (mixed equilibrium):

$$W^{ne} = \xi^2(\hat{V} - 2c) + 2\xi(1-\xi)\left(\frac{1+s}{2}\hat{V} - c\right) = \frac{c\left(\frac{c}{\hat{V}} - s - 1\right) + \hat{V}\left(\frac{s+1}{2}\right)^2}{s}.$$

One can verify that

$$W^e > W^{ne} \qquad \text{for } c \in (\bar{c}, \bar{\bar{c}}).$$

□

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 covari-

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.

ates it can effectively use in the learning and targeting steps. The scope of learning, ℓ, is the number of principal components the firm can use in learning. The scope of targeting, t, is the number of principal components that can be used in targeting. This interpretation captures the *model complexity*, which reflects the higher computing cost deriving from analyzing more covariates.

To reduce the dimensionality whilst extracting the maximum information in the constraints, Jolliffe (2002) shows that the optimal procedure is Principal Component Analysis (PCA). Let the eigendecomposition of the variance/covariance matrix be

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

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

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

Remark 1 (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 (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

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

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

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

Corollary 11. *The Bayes Estimator is the MNLS estimator with shrinkage:*

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

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

[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

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

B.3 Double Descent

Corollary 12. *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 2. 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 3. 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

[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.

each carry an SNR of

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

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

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