Scale, Scope, and the Value of Data*

Giovanni Colla Rizzi†

November 2026

Abstract

Data that cannot be used for prediction can nevertheless make additional observations more valuable. I study a firm that learns the effects of many independent covariates from past individuals to predict outcomes for new ones. When all covariates are available at prediction, returns to observations diminish. But sufficiently informative covariates observed only after prediction can generate increasing returns in small samples. The mechanism is a learning spillover: each covariate reduces the noise in estimating the effects of others, making covariates complements despite their statistical independence. This complementarity also affects the division of gains from combining data. If separate owners of covariates charge their marginal contributions, their claims exceed the value of the combined dataset. A learning multiplier, measurable from the firm's validation learning curve, quantifies these spillovers.

JEL Classification: C55, C78, D24, D83, L41, L86.

Keywords: value of data, prediction, returns to scale, learning curves, data markets, bargaining, property rights over data.

*First public draft: "Increasing Returns to Data for a Bayesian Platform," June 2023. I am particularly indebted to Patrick Rey for his invaluable guidance and support throughout this project. I am also grateful to Jad Beyhum, Michele Bisceglia, Emilio Calvano, Doruk Cetemen, Zhijun Chen, Alexandre de Cornière, Matteo Escudé, Jean-Pierre Florens, Eric Gautier, Andrei Hagiu, Johannes Hörner, Doh-Shin Jeon, Garrett Johnson, Bruno Jullien, Hiroaki Kaido, Gastón Llanes, Simon Loertscher, Friedrich Lucke, Leonardo Madio, Thierry Magnac, Nour Meddahi, Giovanni Morzenti, Juan Ortner, Andrew Rhodes, Christoph Riedl, François Salanié, Maximilian Schäfer, Sara Shahanaghi, Tim Simcoe, Alex Smolin, Marshall Van Alstyne and Davide Viviano for helpful comments and discussions. I thank the seminar participants at the Toulouse School of Economics IO workshop (2022–2026), the Questrom Digital Platforms Seminar (2025), Boston University (2025), EARIE 2025 (Valencia) and LUISS (2026). I gratefully acknowledge financial support from the Horizon Europe – European Research Council (ERC) under Grant Agreement No. 101055280-DIGECON.

†LUISS Guido Carli University, Rome (email: grizzi@luiss.it).

1 Introduction

Suppose a search engine buys data to improve its predictions of which results users will click. It can buy more search records or more characteristics of users already in its records. If each supplier demands what the engine would lose without its data, the engine may afford to buy from every supplier of records but not from every supplier of characteristics. The reason is that observations can be substitutes while covariates are complements. The distinction also matters for pricing data access. Under the Digital Markets Act, gatekeeper search engines must share query, click and view data with rivals on fair, reasonable and non-discriminatory terms. Setting these terms requires knowing how much recipients gain from additional data and how those gains vary with the data they already hold.[1]

The evidence on returns to data is divided. Demand forecasts and image classifiers improve at a diminishing rate with the number of observations they learn from (Bajari et al. 2019; Varian 2019), and the UK competition authority found that the returns to click-and-query data diminish rapidly (Competition and Markets Authority 2020). By contrast, adding covariates to a model that predicts health outcomes can raise accuracy at an increasing rate (Carballa-Smichowski et al. 2025), and on a search engine, the gain from further searches rises with the length of user histories while these are short (Schäfer and Sápi 2023). Moreover, a US court found that search engine rivals given access to Google's data would "keep getting better at a faster and faster rate up to some point."(United States v. Google LLC 2025),

This paper reconciles these findings by distinguishing data by when they become available. A firm learns from records of past individuals and predicts a target variable for a new one. Most covariates are available when the firm predicts, so it observes them for the new individual too. Some become available only after the prediction, so the firm observes them only in its records; I call them *post-prediction covariates*. In search, the query and a user's past searches are covariates, and her next search is a post-prediction covariate. A dataset thus has three margins: *scale*, the number of observations; *scope*, the number of covariates; and *post-prediction scope*, the number of post-prediction covariates. Economic models of the value of data treat data as one-dimensional. Yet two datasets with the same number of entries can differ widely in value (Figure OA.5 in the Online Appendix), and post-prediction covariates change the sign of the returns to observations.

Learning noise and the learning multiplier. The target variable is a linear combination of many independent covariates, some of which no firm observes. The firm learns the weight of each covariate from its sample as a Bayesian and knows its *signal*, the share of the variance of the target variable that it explains. The value of a dataset is the sum, over the covariates the firm observes for the new individual, of each signal times the fraction of the weight that the firm learns, so a post-prediction covariate adds value only by helping the firm learn the

[1] European Union (2022, Article 6(11)).

other weights. When covariates are many, the firm learns each weight as if it were the only unknown, against Gaussian noise whose variance, *learning noise*, is the validation error of its model. Lower noise lets the firm learn every weight better, which lowers noise again, so every change in the data moves learning noise by its first effect times the *learning multiplier*, $1 + \alpha$. The firm can measure it: a sample one percent larger lowers validation error by α percent, so α is the elasticity of its learning curve, a sufficient statistic in the sense of Chetty (2009).

Returns to data. When every covariate is available at prediction, *returns to scale*, the returns to observations with the covariates held fixed, diminish at every scale. Each observation lowers the noise against which the firm learns from the next, but once the firm has learned a weight, it spends further observations on weights with less signal and on averaging out what its covariates leave unexplained. *Returns to scope*, the returns to covariates with the sample held fixed, can increase: each covariate is worth its learned signal plus its *spillover*, the value it adds by helping the firm learn every other weight, and the spillover can rise faster than signal declines, which can account for the increasing returns to health covariates. Groups of covariates exhibit *economies of scope* at every scale, even when returns to scope diminish. Scale and scope are complements while the firm has learned, on average, at most half of the weights it applies, which accounts for the complementarity between searches and the length of user histories.

Post-prediction covariates. Post-prediction covariates overturn diminishing returns to observations. When they carry more signal than the covariates, the firm spends its first observations mostly on their weights, which it learns fast but cannot apply. Value is then either concave or S-shaped in scale, and it is S-shaped if and only if the squared signal of the post-prediction covariates exceeds a *critical mass*, which is low when each covariate carries little signal. In search, each term of a query carries little signal about which result the user will click, and her next search carries much. Increasing returns in small samples can account for the court's finding(United States v. Google LLC 2025), and diminishing returns in large samples for the evidence that the competition authority reviewed. Post-prediction covariates also complement covariates, so data that the firm cannot use when it predicts can create increasing returns to the data that it can. A purpose limitation, which lets a firm learn from a covariate but bars its use at prediction, costs the firm only the covariate's learned signal, while a ban on collecting it also costs its spillover.

The market for data. The same curvature determines who captures the value of data. Spillovers make covariates complements, so when each of many small holders is paid its marginal contribution, holders of covariates together claim $1 + \alpha$ times the value of the data. Bilateral bargaining with them breaks down when holders are strong, an anticommons (Heller 1998) that α measures, and covariates are worth most when generated under common own-

ership, as in ecosystems of services to the same users. Observations without post-prediction covariates are substitutes: their holders together receive the share $\alpha(1-R^2)/R^2 < 1$ of value, where R^2 is the validation R^2, and gain by bargaining together, as proposals for data unions envisage. A learning curve and a held-out sample therefore measure how far data depart from the constant returns under which marginal payments exhaust value (Wicksteed 1894). Post-prediction covariates often travel with observations, as the next search does with a search session, and observations that carry a critical mass of them are complements while they are scarce: their holders claim more than value, and an entrant that must buy them faces the anticommons that a buyer of covariates faces. Three policy conclusions follow. Vesting in each individual the rights over her observation, including every covariate on her, as portability rights do, integrates complements and ensures that bargaining closes, unless her observation carries a critical mass of post-prediction covariates. When regulation mandates access, paying each holder of covariates the learned signal of its group divides value without exceeding it and lies in the core. Anonymization that strips each user's next search from shared sessions makes them substitutes, which help the smallest rival most. These are effects on how value is divided; their effects on prices and welfare depend on how firms compete.

Related literature. Economic models specify the value of data in reduced form to study competition (de Cornière and Taylor 2025; Hagiu and Wright 2023; Prüfer and Schottmüller 2021) and growth (Jones and Tonetti 2020), or model data as signals about one unknown state (Farboodi and Veldkamp 2026; Veldkamp and Chung 2024); data have one dimension, and increasing returns, when present, are assumed or come from feedback between data and demand. To my knowledge, this is the first theory in economics in which increasing returns to data, in both scale and scope, arise from learning alone, with independent covariates and an optimal predictor, and the first to tie increasing returns to scale to a class of data. In independent and contemporaneous work, Schäfer (2025) obtains increasing returns to covariates from their correlation, with known weights. Models of data and information markets take the value of data as a primitive (Acemoğlu et al. 2022; Bergemann et al. 2022; Jones and Tonetti 2020), treat data as general information structures (Bergemann and Bonatti 2024; Bergemann et al. 2018), or study externalities among individuals whose data are correlated (Acemoğlu et al. 2022; Bergemann et al. 2022; Choi et al. 2019). Here value follows from how the firm learns, and its curvature determines how bargaining divides it: the division depends on whether the parts are substitutes or complements (Horn and Wolinsky 1988), machine-learning practice values data by marginal contribution (Ghorbani and Zou 2019), and the learning multiplier gives both a measurable form. Consent-based privacy regulation can favor large firms (Campbell et al. 2015; Johnson et al. 2023); here it weighs most on small firms when behavioral data recorded after the prediction reach the critical mass.

Radner and Stiglitz (1984) show that the marginal value of information is zero at no information, so the value of information is not concave. Here the marginal value of an observation

is positive at no information, and the value of observations can be convex because the firm learns many weights from the same observations and applies only some of them. Börgers et al. (2013) characterize when two signals are complements in every decision problem, for a decision maker who knows how they relate to the state; here the firm learns that relation. The fixed point that defines learning noise is the Silverstein–Bai equation for the spectrum of a sample covariance matrix (Silverstein and Bai 1995), which underlies the asymptotics of ridge regression (Dobriban and Wager 2018; Hastie et al. 2022). I prove the representation directly, from how the firm learns each weight at a given noise and an identity for the degrees of freedom of the sample, which allows post-prediction covariates and makes the multiplier explicit (Online Appendix OA.4.1). Theorists of scaling laws use the same features to learn and to predict (Bahri et al. 2024; Bordelon et al. 2024), so returns to scale diminish (Proposition 1(i)). Machine learning calls post-prediction covariates privileged information and shows that it can speed up learning (Vapnik and Vashist 2009); here it makes returns to scale increase, and the composition of the data matters, as researchers in machine learning have argued (Sorscher et al. 2022).

Roadmap. Section 2 presents the model. Section 3 characterizes the value of every dataset and derives the learning multiplier when every covariate is available at prediction, and Section 4 studies the returns to scale and scope in that baseline. Section 5 adds post-prediction covariates, and Section 6 studies how bargaining divides the value of data among its holders. The Appendix contains the proofs; the Online Appendix contains simulations, further results and connections to statistics.

2 Model

A firm predicts a scalar target variable for a target individual. To do so, it observes the target variable and some covariates for a sample of other individuals, the training sample, and some of these covariates for the target individual. Some covariates become available only after the firm predicts, so the firm observes them in the training sample but not for the target individual. The firm chooses the size of the sample and which covariates of each kind to observe. I ask how much the firm lowers its expected prediction error with each choice.

2.1 Data-Generating Process

There are K potential covariates. Individuals are indexed by $i = 0, 1, 2, \ldots$: the firm predicts for individual 0, the target individual, and learns from individuals $1, \ldots, N$, for a sample size N that the firm chooses (Section 2.2). For each individual i, the target variable is a linear combination

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

where $\boldsymbol{x}_i \in \mathbb{R}^K$ is the covariate vector of individual i and $\boldsymbol{\beta} \in \mathbb{R}^K$ is a vector of unknown weights common to all individuals.[2]

Weights. The weights are independent of the covariates, and they are mutually independent zero-mean Gaussians, $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \sigma^2 \cdot \boldsymbol{I}_K\right)$. The variance $\sigma^2 > 0$ measures prior uncertainty about how covariates map into the target variable.

Covariates. Covariates are independent across individuals and across coordinates:

$$x_{ik} \sim \mathcal{N}\left(0, \frac{\omega_k}{K}\right), \qquad \omega_k \geq 0, \qquad \frac{1}{K}\sum_{k=1}^{K} \omega_k = 1.$$

The target variable therefore has variance $\mathbb{V}(y_i) = \sigma^2$, and covariate k explains the fraction ω_k/K of it. I call ω_k the *signal* of covariate k.[3] The signal averages one whatever the number of covariates, and the firm knows the signal of each covariate.

Unknown weights, independent covariates. Weights have zero mean, so the firm cannot tell whether the target variable rises or falls with a covariate.[4] With independent covariates, the firm cannot use one covariate as a proxy for another, so all interactions between covariates arise from learning about the weights. Simulations with correlated covariates reproduce the patterns of the returns in Section 4 (Online Appendix OA.2.5).

2.2 Datasets

Covariates differ in when they become available. The first $\bar{S}$ covariates, $k = 1, \ldots, \bar{S}$, are available when the firm predicts, so the firm can observe them for every individual, in the training sample and for the target individual. The next $\bar{P}$ covariates, $k = \bar{S}+1, \ldots, \bar{S}+\bar{P}$, become available only after the firm predicts, so the firm can observe them for past individuals in the training sample but not for the target individual. I call them *post-prediction covariates*, and I reserve the word *covariates*, without qualifier, for the first kind when the two are contrasted. What matters is availability at the time of prediction: a post-prediction covariate can be realized before or after the target variable. In search, a user's next search is a post-prediction covariate: the engine records it for past queries, but it is unavailable when the engine ranks results for a new one. The remaining covariates are never observed.

[2]The model has no idiosyncratic error. With many covariates, every covariate outside the sample acts as Gaussian noise whose variance is its signal (Step 2 of the proof of Theorem 1), so an error independent of the covariates enters exactly as covariates that the firm never observes.

[3]Heterogeneous prior variances add nothing. If $x_{ik} \sim \mathcal{N}(0, a_k/K)$ and $\beta_k \sim \mathcal{N}(0, c_k)$, rescaling covariate k by the known constant $\sqrt{c_k}$ yields this model with signal proportional to $a_k c_k$ and leaves value unchanged.

[4]With a known nonzero prior mean, value gains a term independent of the training data, and the noise that the firm faces when it learns the weights rises. I measure the value that data add to prior knowledge.

Within each kind, covariates become available in the order of their index, for instance as the firm acquires sources or as its engineers add features to a model. The order places no restriction on how signal varies with the index. A firm free to choose the order would take the covariates of each kind in decreasing order of signal, which weakly raises the value of every dataset (Proposition OA.1 in Online Appendix OA.3.2).

The firm chooses a dataset $D \equiv (N, S, P) \in \mathbb{N}_0^3$, with $S \leq \bar{S}$ and $P \leq \bar{P}$: it observes the target variable of N training individuals, the first S covariates of the training individuals and of the target individual, and the first P post-prediction covariates of the training individuals. The firm therefore observes the data

$$\boldsymbol{d}_D \equiv \left(\left(y_i,\ x_{i1}, \ldots, x_{iS},\ x_{i,\bar{S}+1}, \ldots, x_{i,\bar{S}+P} \right)_{i=1}^{N},\ \left(x_{01}, \ldots, x_{0S} \right) \right),$$

which are empty for the empty dataset $D_\varnothing \equiv (0, 0, 0)$.[5] Figure 1 depicts a dataset.

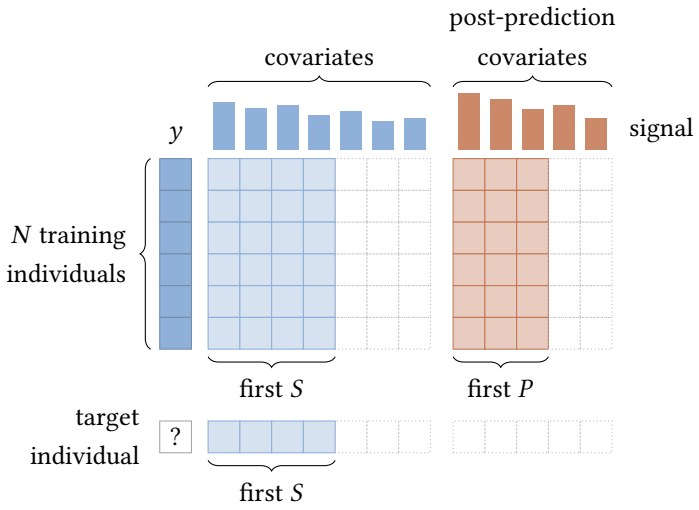


Figure 1. A dataset $D = (N, S, P)$ with $K = 12$: seven covariates and five post-prediction covariates, each kind in the order in which its covariates become available; bars show their signal. Shaded cells are observed: the target variable y, the first S covariates and the first P post-prediction covariates of each of the N training individuals, and the first S covariates of the target individual, whose target variable the firm predicts.

2.3 Prediction and the Value of a Dataset

Predictors and risk. A *predictor* for dataset D is a measurable function $\hat{y}$ of the data $\boldsymbol{d}_D$. The firm incurs quadratic loss $(y_0 - \hat{y})^2$. The *Bayes risk* of D is the lowest expected loss any predictor for D achieves,

$$R(D) \equiv \inf_{\hat{y}} \mathbb{E}\left[(y_0 - \hat{y}(\boldsymbol{d}_D))^2 \right].$$

[5] The firm could also observe for the target individual a covariate absent from its training sample. Such a covariate adds no value: weights and covariates are symmetric about zero, so the posterior mean of its weight is zero (Remark 1 in the Appendix).

Value. The *value* of dataset D is the fall in risk relative to predicting without data,

$$V(D) \equiv R(D_\varnothing) - R(D).$$

Validation error. A firm can measure its error only on individuals whose target variable it observes, such as an individual that it sets aside from its training sample. For such an individual, it observes every covariate in the sample, post-prediction covariates included. Write $\boldsymbol{d}_D^{\text{val}}$ for the data $\boldsymbol{d}_D$ together with $x_{0,\bar{S}+1}, \dots, x_{0,\bar{S}+P}$. The *validation error* of D is the Bayes risk with these data,

$$R^{\text{val}}(D) \equiv \inf_{\hat{y}} \mathbb{E}\big[(y_0 - \hat{y}(\boldsymbol{d}_D^{\text{val}}))^2\big],$$

and $1 - R^{\text{val}}(D)$ is the *validation* R^2.

Order on datasets. Datasets are ordered componentwise: $D \leq D'$ if and only if $N \leq N'$, $S \leq S'$ and $P \leq P'$. Because the covariates of each kind enter every dataset in the same order, the firm observes under a larger dataset everything it observes under a smaller one.

3 Value and the Learning Multiplier

This section characterizes the value of every dataset and then specializes to datasets without post-prediction covariates, in which one feedback loop, the learning multiplier, governs how the firm learns.[6] Proofs are in the Appendix.

3.1 Learning and Prediction

Under quadratic loss, the best predictor is the posterior mean, $\hat{y}_D^\star \equiv \mathbb{E}[y_0 \mid \boldsymbol{d}_D]$, and the value of a dataset is the variance of the target variable that this predictor explains:

$$V(D) = \mathbb{V}(y_0) - \mathbb{E}\big[\mathbb{V}(y_0 \mid \boldsymbol{d}_D)\big] = \mathbb{V}(\hat{y}_D^\star). \tag{1}$$

Value rises with the dataset, because a larger dataset reveals everything a smaller one does.[7] Write $\boldsymbol{y}_N \in \mathbb{R}^N$ and $\boldsymbol{X}_D \in \mathbb{R}^{N\times(S+P)}$ for the target variables and the covariates that the firm observes for the training individuals, and $\boldsymbol{m}_D \equiv \mathbb{E}[\boldsymbol{\beta} \mid \boldsymbol{y}_N, \boldsymbol{X}_D]$ for the posterior mean of the weights, with $\boldsymbol{m}_D = \boldsymbol{0}$ when $N = 0$.

[6] Table OA.1 in the Online Appendix lists every object with the page on which it is defined.

[7] Both facts hold for any square-integrable target variable (Online Appendix OA.3.1), so the optimal predictor never overfits: its expected error is nonincreasing in the sample size and in the number of covariates the firm observes.

Lemma 1 (Learning and prediction). *For every dataset* $D = (N, S, P)$,

$$\hat{y}^{\star}_{D} = \sum_{k=1}^{S} x_{0k}\, m_{D,k} \qquad \textit{and} \qquad V(D) = \sum_{k=1}^{S} \mathbb{V}(x_{0k})\, \mathbb{V}(m_{D,k}) = \frac{1}{K} \sum_{k=1}^{S} \omega_k\, \mathbb{V}(m_{D,k}),$$

and each $\mathbb{V}(m_{D,k})$ *is nondecreasing in* D.

The variance of a learned weight, as a share of the prior variance, $\mathbb{V}(m_{D,k})/\sigma^2 \in [0, 1]$, is the *fraction learned* of the weight. The sample, post-prediction covariates included, determines the fractions learned; the covariates observed for the target individual determine which learned weights the firm applies. A post-prediction covariate therefore adds value as an observation does, by improving what the firm learns about the weights it applies.[8]

3.2 Learning with Many Covariates

At finite K, what the firm learns about one weight depends on what it learns about every other. When covariates are many, the firm learns each weight as if that weight were the only unknown and the rest of the target variable were Gaussian noise whose variance, learning noise, is the same for every weight and is the limit of the validation error.

Units, profiles and margins. From now on, $\sigma^2 = 1$, so value is the expected out-of-sample R^2 of the firm's predictions and the fraction learned of covariate k is $\mathbb{V}(m_{D,k})$.[9] I let the number of potential covariates grow, holding fixed the shares of covariates, $\bar{s}$, and of post-prediction covariates, $\bar{p}$, and how signal varies along each order. Two *signal profiles*, ω_s on $[0, \bar{s}]$ and ω_p on $[0, \bar{p}]$, describe the signals: the covariate at position u of its order has signal $\omega_s(u)$, and the post-prediction covariate at position u has signal $\omega_p(u)$, so that $\omega_k \approx \omega_s(k/K)$ for $k \leq \bar{S}$ and $\omega_{\bar{S}+j} \approx \omega_p(j/K)$ for $j \leq \bar{P}$. A decreasing profile describes covariates that become available in decreasing order of signal, and a step profile describes groups of covariates of similar signal; Appendix A states the construction precisely. *Scale* n, *scope* s and *post-prediction scope* p approximate N/K, S/K and P/K. A sample that contains the first s covariates and the first p post-prediction covariates leaves out the signal

$$\underline{E}(s, p) \equiv 1 - \int_0^s \omega_s(u)\, du - \int_0^p \omega_p(u)\, du,$$

the prediction error of a firm that knew the weights of all the covariates in its sample and observed them when it predicts.

[8] Lemma 1 uses only that weights and covariates are independent, centered and square-integrable, and it holds for every set of covariates in the sample and every subset of them observed for the target individual.

[9] Dividing the target variable by σ divides every weight by σ. The firm knows σ, so the rescaled data carry the same information, and value, a variance, is divided by σ^2.

Assumption 1 (Many covariates). *The firm chooses a dataset $D_K = (N, S, P)$ for each K, and as $K \to \infty$,*

$$\frac{N}{K} \to n \in (0, \infty), \qquad \frac{S}{K} \to s \in [0, \bar{s}], \qquad \frac{P}{K} \to p \in [0, \bar{p}], \qquad \underline{E}(s, p) > 0.$$

The condition $\underline{E}(s, p) > 0$ states that the sample leaves out some signal. It keeps the noise that the firm faces when it learns the weights away from zero.

Learning a single weight. Suppose that the firm observes the target variable and one covariate, of signal ω, for $N = nK$ individuals, and that everything else in the target variable is Gaussian noise of variance E. Each observation carries precision x_{ik}^2/E about the weight, which averages $\omega/(KE)$, so the sample carries precision $\tau\omega$, where $\tau \equiv n/E$ is the *precision* per unit of signal (the prior has unit precision). The fraction learned of the weight is then $\ell(\tau\omega)$, where $\ell(x) \equiv x/(1+x)$ for $x \geq 0$ is the share of the posterior precision that comes from the data. At a given precision, the firm learns the weights of covariates with more signal better. Once the firm applies its learned weight, the covariate explains the share $\omega\,\ell(\tau\omega)$ of the variance of the target variable, its *learned signal*.

Learning noise. The noise on each weight is the signal of the covariates outside the sample plus the signal of the other covariates in the sample that the firm has not learned. Each covariate carries a vanishing share of the signal, so the noise is the same for every weight. If the firm learned every weight at precision τ, the value of the covariates that it applies would be

$$v(\tau, s) \equiv \int_0^s \omega_s(u)\,\ell\big(\tau\omega_s(u)\big)\,du,$$

and the noise on each weight would be

$$e(\tau; s, p) \equiv \underline{E}(s, p) + \int_0^s \omega_s(u)\big[1-\ell\big(\tau\omega_s(u)\big)\big]\,du + \int_0^p \omega_p(u)\big[1-\ell\big(\tau\omega_p(u)\big)\big]\,du = 1 - v(\tau, s) - \int_0^p \omega_p(u)\,\ell\big(\tau\omega_p \tag{2}$$

the prediction error of a model whose weights the firm learns at precision τ, for an individual whose covariates in the sample, post-prediction covariates included, the firm observes. The error falls as precision rises, from $e(0; s, p) = 1$ toward $\underline{E}(s, p)$; write $e_\tau \equiv \partial e/\partial\tau < 0$. This error is also the noise on each weight, and precision is scale per unit of noise, so noise and precision must satisfy $E = e(\tau; s, p)$ and $\tau = n/E$. An individual that the firm sets aside from its sample is such an individual, so this error is also the validation error of Section 2. Value, validation error and the fractions learned are expectations over the draw of the sample, so the limits below are limits of numbers.

Theorem 1 (Learning noise). *Under Assumption 1:*

(i) the system

$$\tau^\star = n/E^\star, \qquad E^\star = e(\tau^\star; s, p), \qquad \mathcal{V} = v(\tau^\star, s) \tag{3}$$

has a unique solution $\big(\tau^\star(n,s,p), E^\star(n,s,p), \mathcal{V}(n,s,p)\big)$*, at which* $E^\star \in [\underline{E}(s,p), 1]$*;*

(ii) as $K \to \infty$*, the fraction learned of every covariate in the sample, of either kind, converges, uniformly, to its single-weight value at precision* $\tau^\star$*,*

$$\max_{k \le S \text{ or } \bar{S} < k \le \bar{S}+P} \left| \mathbb{V}(m_{D_K,k}) - \ell(\tau^\star \omega_k) \right| \longrightarrow 0,$$

so $V(D_K) \to \mathcal{V}$ *and* $R^{\text{val}}(D_K) \to E^\star$*.*

I call $E^\star(n,s,p)$ the *learning noise* and $\tau^\star(n,s,p)$ the precision at learning noise. When the margins are fixed, I suppress their arguments; a star marks an object evaluated at learning noise.[10] Noise is the signal that the firm has not learned, of covariates of either kind, and value is the learned signal of the covariates that the firm applies. Precision therefore acts as an intermediate input: the firm produces it from observations and, by lowering noise, from covariates of either kind, and it uses it to learn the weights of the covariates it applies. Validation R^2, $1 - E^\star$, exceeds value by the learned signal of the post-prediction covariates, which the firm learns but cannot apply (Section 5).

The baseline. Until Section 5, the firm's sample contains no post-prediction covariates, $p = 0$, and I drop the argument p: $e(\tau; s)$, $\underline{E}(s)$, $\tau^\star(n,s)$, $E^\star(n,s)$ and $\mathcal{V}(n,s)$. Then $e(\tau; s) = 1 - v(\tau, s)$, so value and learning noise sum to one: value is the validation R^2, and the learning curve is one minus value. Equal signals illustrate the system. If every covariate has signal one, every weight in the sample has the same fraction learned, $x \equiv \ell(\tau^\star)$; noise is $E^\star = 1 - sx$ and $x = n/(n + E^\star)$, so

$$x = \frac{2n}{n + 1 + \sqrt{(n+1)^2 - 4ns}} \qquad \text{and} \qquad \mathcal{V}(n,s) = s\,x$$

(Corollary OA.1 in the Online Appendix). The fraction learned of each weight rises with scale and with scope: a further covariate lowers the noise against which the firm learns every other weight.

3.3 The Learning Multiplier

From here on, I fix $n > 0$, $s > 0$ and p with $\underline{E}(s,p) > 0$.[11] Every margin acts on learning noise through a feedback loop: lower noise means more precision, which raises the fraction learned

[10] I treat the limit as an approximation to a large but finite environment; in the simulations of Online Appendix OA.2.2, it is close to simulated value already with 100 covariates.

[11] The profiles are positive, and each is continuously differentiable near the marginal position, s or p, whenever I differentiate with respect to that margin; a step profile satisfies this away from its jumps.

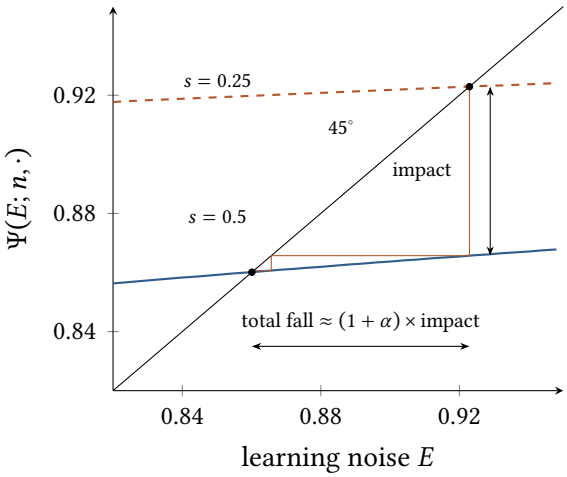


Figure 2. Learning noise as a fixed point. Signal $\omega_s(u) = c\, e^{-0.75u}$ on $[0, \bar{s}]$ for covariates, with $\bar{s} = 0.65$; never-observed covariates fill the remaining positions, with signal 2 on the next 0.25 positions and 1 on the last 0.1, and c sets the total signal to one. Here $n = 1$. The map Ψ at scopes 0.25 and 0.5, its fixed points, and the iterations that follow a rise in scope from 0.25 to 0.5. Section 5 turns the covariates of signal 2 into post-prediction covariates.

of every weight in proportion to the part still to learn, and the newly learned signal leaves the noise. The loop multiplies the effect of every margin, and the elasticity of the firm's learning curve measures it.

In the baseline, write $\Psi(E; n, s) \equiv e(n/E; s)$ for the error at the precision that noise E implies. Learning noise is the fixed point of this map, $E^\star = \Psi(E^\star; n, s)$ (Figure 2). When the noise at which the firm learns falls by one unit at fixed scale, precision rises by $n/E^{\star 2}$, and each unit of precision lowers the error by $-e_\tau$. The error therefore falls by the slope of the map,

$$\Psi' \equiv \frac{\partial \Psi(E^\star; n, s)}{\partial E} = \frac{n}{E^{\star 2}}\,(-e_\tau) = \frac{1}{E^\star}\int_0^s \omega_s(u)\,\ell\big(\tau^\star \omega_s(u)\big)\big[1 - \ell\big(\tau^\star \omega_s(u)\big)\big]\,du \;:$$

the learned signal of a covariate of signal ω rises by $\omega\ell(\tau^\star\omega)\big[1 - \ell(\tau^\star\omega)\big]/E^\star$, the *speed* at which the firm learns its weight (Lemma 2), and the error falls by the sum of the speeds over the sample. The slope is the *strength* of the feedback loop.

The strength of the loop determines the *learning multiplier* $1+\alpha$, where $\alpha \equiv \Psi'/(1-\Psi') \in [0, \infty)$ is the *amplification*: a unit fall in noise lowers noise by a further Ψ', then by Ψ'^2, and the falls sum to $1 + \Psi' + \Psi'^2 + \cdots = 1 + \alpha$.[12] Differentiating $E^\star = \Psi(E^\star; n, s)$ with respect to $x \in \{n, s\}$ gives $\partial E^\star/\partial x = (1 + \alpha)\,\partial\Psi/\partial x$: the total effect of a margin on learning noise, and through $\tau^\star = n/E^\star$ on precision, is its *impact*, the effect at given noise, times the multiplier. The staircase of Figure 2 shows these falls after a rise in scope. With equal signals, $\Psi' = sx^2/n$, so the loop strengthens with scope.

Write $\ell_s \equiv \ell(\tau^\star\omega_s(s))$ for the fraction learned of the marginal covariate, and $v_\tau \equiv \partial v(\tau^\star, s)/\partial\tau$ for the *value of precision*. In the graph of the system in Figure 3, scale enters through precision: at given noise, a sample larger by one percent is one percent more precise. Scope enters through noise: once the firm learns the weight of a new covariate, its learned signal leaves the noise. Precision raises value, scope also raises value directly, and noise and

[12] The sum is finite because $\Psi' < 1$: the right side of $n = \tau e(\tau; s)$ rises with precision, at the rate $E^\star(1 - \Psi')$ at $\tau^\star$.

precision feed back on each other.

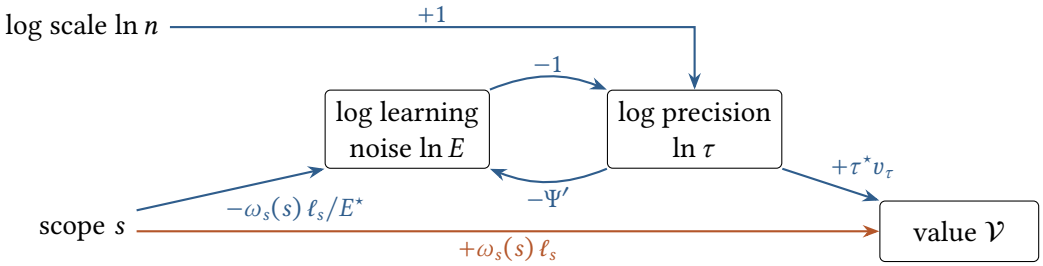


Figure 3. How the margins affect value. The graph represents Equation (3), with scale, noise and precision in logs. Each arrow shows the partial derivative of the variable at its head with respect to the variable at its tail, with the other variables fixed. The arrow from noise to precision is the identity $\ln\tau = \ln n - \ln E$; the loop has strength Ψ', the product of its two arrows, and multiplies the impact of every margin on noise and precision by $1/(1-\Psi')$ (Section 3.3).

Measurement. By Theorem 1(ii), learning noise is the limit of validation error, which the firm can estimate from its training sample by cross-validation; a firm that fits ridge regression with a cross-validated penalty, or runs approximate message passing, attains the value in Theorem 1 (Online Appendix OA.4.3). The validation error as a function of the sample size, $N \mapsto R^{\text{val}}(D)$, is the firm's *learning curve*. In Figure 3, the path from scale to noise has arrows 1 and $-\Psi'$, so, with the multiplier, a sample one percent larger lowers learning noise by $(1+\alpha)\Psi' = \alpha$ percent: $\alpha = -\partial \ln E^\star / \partial \ln n$ is the elasticity of the learning curve. The firm can estimate it by refitting its model on nested subsamples of its training sample and regressing the log of the cross-validated error on the log of the subsample size near the full sample. The multiplier is a sufficient statistic in the sense of Chetty (2009): whatever the signal profile, precision rises by $1+\alpha$ percent when the sample grows by one percent.

4 Returns to Data

When the firm applies every weight it learns, returns to observations diminish at every scale, while spillovers on the weights of other covariates can make returns to scope increase and make any two fixed groups of covariates complements.

By Theorem 1, $\mathcal{V}(n,s) = v\big(\tau^\star(n,s), s\big)$; subscripts denote partial derivatives of v, evaluated at $\tau^\star$. Hence

$$\frac{\partial \mathcal{V}}{\partial n} = v_\tau \frac{\partial \tau^\star}{\partial n}, \qquad \frac{\partial \mathcal{V}}{\partial s} = \omega_s(s)\,\ell_s + v_\tau \frac{\partial \tau^\star}{\partial s}, \qquad \frac{\partial \tau^\star}{\partial n} = \frac{1+\alpha}{E^\star}, \qquad \frac{\partial \tau^\star}{\partial s} = \frac{n}{E^{\star 2}}\,\omega_s(s)\ell_s\,(1+\alpha).$$

The first term of $\partial\mathcal{V}/\partial s$ is the *direct effect* of a covariate, its learned signal at given precision. I call the second term its *spillover*, because it is the value of learning, at higher precision, the weights of the other covariates. The precision that each margin adds is its impact times the

multiplier (Section 3.3). Precision has positive and diminishing value:

$$\begin{aligned} v_\tau &= \int_0^s \omega_s(u)^2\big(1-\ell(\tau^\star\omega_s(u))\big)^2 du > 0, \\ v_{\tau\tau} &= -\frac{2\bar{\ell}_s}{\tau^\star}\, v_\tau < 0, \qquad \text{where } \bar{\ell}_s \equiv \frac{\int_0^s \ell\big(\tau^\star\omega_s(u)\big)^3 du}{\int_0^s \ell\big(\tau^\star\omega_s(u)\big)^2 du} \end{aligned} \tag{4}$$

is the average fraction learned of the weights that the firm applies, each weighted by its contribution to v_τ, $\omega^2(1-\ell)^2$, which is proportional to its learning speed. One more unit of precision raises the learned signal of a weight by the square of its signal still to learn, and this rise falls as the firm learns the weight. The assumptions of Section 3.3 hold throughout.

4.1 Returns to Scale

Returns to observations diminish at every scale (Proposition 1). Two forces compete: the loop makes precision rise at an increasing rate with scale, and value rises with precision at a decreasing rate. When the firm applies every weight it learns, the second force always wins.

Precision accelerates because each observation lowers noise, which raises the impact of the next one, and because it changes the amplification. How it changes the amplification depends on the learning speeds. At given noise, the speed of a weight, $\omega\ell(1-\ell)/E^\star$, is highest when the firm has learned half of it, so one more observation speeds up the learning of the weights learned less than half and slows down that of the others. Lower noise speeds up the learning of every weight: a further unit fall in noise yields more precision the lower the noise. In small samples, the firm has learned less than half of every weight, and the amplification rises with scale; in large samples, it has learned most of every weight, and the amplification falls toward zero (Appendix B). Write $\bar{\ell} \equiv \alpha(1-\bar{\ell}_s) < \alpha$.

Proposition 1 (Scale)**.**

(i) Returns to scale diminish at every scale: $\partial^2\mathcal{V}/\partial n^2 < 0$.

(ii) The marginal value of an observation is mediated by precision:

$$\frac{\partial\mathcal{V}}{\partial n} = \underbrace{v_\tau}_{\substack{\text{value of}\\ \text{precision (+)}}} \underbrace{\frac{1}{E^\star}}_{\text{impact (+)}} \underbrace{(1+\alpha)}_{\text{multiplier (+)}} = \frac{\alpha E^\star}{n} > 0. \tag{5}$$

(iii) Precision is convex in scale, $\partial^2\tau^\star/\partial n^2 = (2\bar{\ell}/\tau^\star)(\partial\tau^\star/\partial n)^2 > 0$*, and the curvature of value in scale is*

$$\frac{\partial^2\mathcal{V}}{\partial n^2} = \frac{2}{\tau^\star}\, v_\tau\Big(\frac{\partial\tau^\star}{\partial n}\Big)^2\big(\bar{\ell}-\bar{\ell}_s\big) = \frac{2}{\tau^\star}\, v_\tau\Big(\frac{\partial\tau^\star}{\partial n}\Big)^2(1+\alpha)\big(\Psi'-\bar{\ell}_s\big). \tag{6}$$

Since value is one minus learning noise, $n\,\partial\mathcal{V}/\partial n$ is the slope of the learning curve against log scale (part (ii)). A one percent rise in precision raises the precision that an observation adds by $2\bar{\ell}$ percent and lowers the value of precision by $2\bar{\ell}_s$ percent (part (iii)). Without the loop, $\alpha = 0$, and returns to scale diminish. With it, the precision that an observation adds accelerates through the same weights whose value of precision falls. The strength of the loop, Ψ', is the average fraction learned over the degrees of freedom that the firm has spent: it spends its sample on the weights in the sample, in proportion to the fractions it has learned, and on averaging out the signal outside the sample, which it learns nothing about (Appendix B). The value of precision falls at the rate $\bar{\ell}_s$, which weights the same fractions by learning speed and so favors the weights that the firm has learned most. The second average exceeds the first (part (i)). With equal signals, value $s\,x$ is concave in scale. Observations are therefore substitutes whenever the firm applies every weight it learns, as demand forecasts and image classifiers do (Section 6).

4.2 Returns to Scope

A covariate is worth its direct effect plus its spillover, and spillovers make up the share $\alpha/(1+\alpha)$ of its marginal value. Along an order of acquisition, returns to scope increase when signal does not decline, and declining signal makes them diminish in small and in large samples. These are marginal comparisons: they compare the next covariate in the order with the last.

Proposition 2 (Scope)**.**

(i) The marginal value of a covariate is its direct effect plus its spillover:

$$\frac{\partial \mathcal{V}}{\partial s} = \underbrace{\omega_s(s)\,\ell_s}_{\textit{direct effect}} + \underbrace{\alpha\,\omega_s(s)\ell_s}_{\textit{spillover}} = \omega_s(s)\ell_s\,(1+\alpha) > 0. \tag{7}$$

(ii) One more covariate raises the amplification, $\partial\alpha/\partial s > 0$, and the curvature of value in scope exceeds the change in the marginal value due to the signal of the next covariate: $\partial^2\mathcal{V}/\partial s^2 > (1+\alpha)\,\ell_s(2-\ell_s)\,\omega_s'(s)$.

(iii) Scale and scope are complements whenever $\bar{\ell}_s \leq 1/2$.

The spillover is the fall in noise that the covariate causes, which is equal to its learned signal times the multiplier, valued at Ψ' per unit: a unit fall in noise teaches the firm every other weight better and raises value by the strength of the loop. Hestness et al. (2017) estimate α between 0.07 and 0.35 for deep networks on most translation, language, image and speech tasks, which use every input both in training and at prediction. If the model described them, spillovers would account for 7 to 26 percent of the marginal value of scope. As scope rises, the next covariate adds its learning speed to the loop and lowers noise, which raises the amplification and lets the firm learn the next weight better. Along an order of declining signal, however,

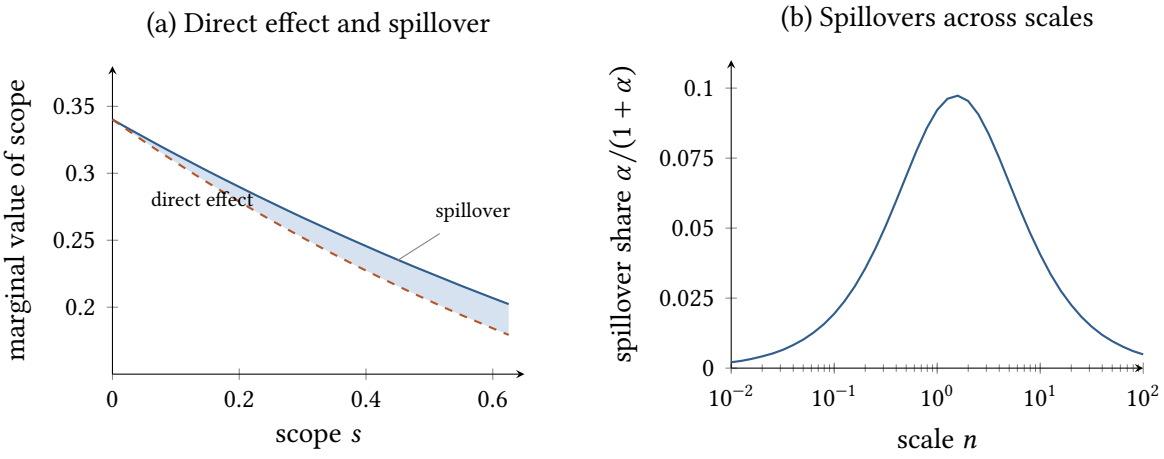


Figure 4. The marginal value of scope. Example of Figure 2. (a) At $n = 1$, the marginal value of scope, split into the direct effect and the spillover (shaded). (b) At $s = 0.5$, the share of the spillover in the marginal value of scope.

the next covariate carries less signal, and this is the only force that can make returns to scope diminish. A covariate also adds a weight that the firm learns from every further observation, and lower noise raises the impact of an observation while lowering the value of precision; the first effect wins whenever $\bar{\ell}_s \leq 1/2$ (part (iii); Appendix B gives the decomposition). Schäfer and Sápi (2023) and Lee and Wright (2025) document such complementarities: in search, the gain from further searches rises with the length of user histories.

Corollary 1 (Returns to scope)**.**

(i) If signal does not decline at s, $\omega_s'(s) \geq 0$, returns to scope increase at s.

(ii) If signal declines at s, $\omega_s'(s) < 0$, there are $0 < \underline{n} \leq \bar{n}$ such that returns to scope diminish at s at every $n < \underline{n}$ and every $n > \bar{n}$.

With equal signals, value $s\,x$ is strictly convex in scope at every scale.[13] In small samples, a new covariate barely lowers noise; in large samples, the firm has nearly learned the weights it applies, and the multiplier tends to one. The precision terms therefore vanish at both ends, while declining signal remains; in between, they can outweigh it. In the example of Figure 4, the spillover share peaks at intermediate scale. In large samples, scale and scope are substitutes if the covariates outside the sample carry less signal than scope times the signal of the marginal covariate (Online Appendix OA.3.6).

[13]Decline also vanishes if the firm adds covariates in random order: each new covariate is then a random draw from the same distribution of signal, and value is strictly convex in s at every scale (Online Appendix OA.3.3). Carballa-Smichowski et al. (2025) add health covariates in random order and find returns to scope that first increase and then diminish. Spillovers can account for the increase; under a random order, the model predicts no decline.

4.3 Economies of Scope

Two fixed groups of covariates add more value together than apart, at every scale, even when returns to scope diminish (Proposition 3). This comparison concerns finite groups, as when a firm buys a broker's file, rather than the next covariate in an order. Let $V_N(\mathcal{S})$ be the value when the firm observes the covariates in a group $\mathcal{S} \subseteq \{1, \dots, \bar{S}\}$ for N training individuals and for the target individual. In the limit, $\mathcal{S}$ collects the covariates whose positions lie in a finite union of intervals of $[0, s]$ with positive length, also written $\mathcal{S}$, and $\mathcal{V}(\mathcal{S})$ is the limit of $V_N(\mathcal{S})$ at scale n (Appendix B). Write $|\mathcal{S}|$, $\int_{\mathcal{S}} \omega_s$ and $\int_{\mathcal{S}} \omega_s^2$ for the *size*, *cumulative signal* and *squared signal* of a group. Two disjoint groups $\mathcal{S}_1$ and $\mathcal{S}_2$ exhibit *economies of scope* at scale n if $G \equiv \mathcal{V}(\mathcal{S}_1 \cup \mathcal{S}_2) - \mathcal{V}(\mathcal{S}_1) - \mathcal{V}(\mathcal{S}_2) > 0$ (Baumol et al. 1982).

Proposition 3 (Economies of scope)**.**

(i) The value that $\mathcal{S}_2$ adds to $\mathcal{S}_1$ is at least its standalone value: for every K and N, $V_N(\mathcal{S}_1 \cup \mathcal{S}_2) - V_N(\mathcal{S}_1) \geq V_N(\mathcal{S}_2)$.

(ii) In the limit, the value that $\mathcal{S}_2$ adds strictly increases when $\mathcal{S}_1$ grows, disjointly from $\mathcal{S}_2$. Hence the two groups exhibit economies of scope at every scale, and G strictly increases when either group grows.

(iii) In small samples, $G/n^2 \to 2 \int_{\mathcal{S}_1} \omega_s^2 \int_{\mathcal{S}_2} \omega_s^2$.

(iv) In large samples, $nG \to |\mathcal{S}_1| \int_{\mathcal{S}_2} \omega_s + |\mathcal{S}_2| \int_{\mathcal{S}_1} \omega_s$.

The direct effects of two groups add up, and combining them adds the spillovers of each group on the weights of the other, which are positive at every scale. In small samples, each spillover is the fall in noise that one group causes, about scale times its squared signal, times the value of lower noise on the other group's weights, about scale times that group's squared signal. In large samples, the firm learns each weight it applies up to an error of $\underline{E}/n$, as least squares does, and combining the groups removes the cumulative signal of each from the noise against which the firm learns the weights of the other, so size and cumulative signal are sufficient statistics. The result uses the independence of the covariates: a group perfectly correlated with the covariates a firm holds adds nothing to it.

5 Post-Prediction Covariates

Post-prediction covariates improve learning but cannot enter the prediction. By Theorem 1, they change one thing in the learning mechanism: their learned signal leaves the noise without adding to value. This is enough to overturn the central conclusion of Section 4.1, diminishing returns to observations.

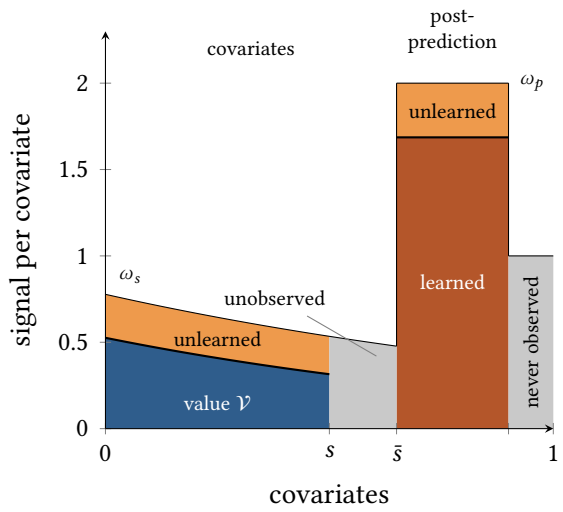


Figure 5. Value and learning noise with post-prediction covariates. Example of Figure 2, with the covariates of signal 2 turned into post-prediction covariates, $\omega_p(u) = 2$ on $[0, \bar{p}]$ and $\bar{p} = 0.25$; here $n = 1$, $s = 0.5$ and $p = \bar{p}$. Value is the area under the learned signal of the first s covariates. Learning noise, the limit of validation error, is the sum of the unlearned areas and the gray areas, which are the signal outside the sample. The terracotta area, the learned signal of the post-prediction covariates, lowers learning noise but cannot enter a prediction, so value falls short of one minus learning noise by this area.

5.1 Learning Noise and Predictive Value

With $p > 0$, value falls short of the validation R^2, $1 - E^\star$, by the learned signal of the post-prediction covariates (Figure 5): validation error, measured on held-out individuals whose post-prediction covariates the records contain, differs from the prediction error for target individuals, $1 - \mathcal{V}$.

The strength of the loop, $\Psi' = (n/E^{\star 2})(-e_\tau)$ with e from Equation (2), now sums the learning speeds over both kinds of weights in the sample, and only part of it raises value. At fixed scale, the *value of lower noise*, the rise in value when the noise at which the firm learns falls by one unit, is

$$\Psi'_s(E) \equiv -\frac{\partial v(n/E, s)}{\partial E} = \frac{1}{E}\int_0^s \omega_s(u)\,\ell\Big(\frac{n\omega_s(u)}{E}\Big)\Big[1 - \ell\Big(\frac{n\omega_s(u)}{E}\Big)\Big]\,du,$$

the sum of the learning speeds of the weights that the firm applies, and $\Psi'_s \equiv \Psi'_s(E^\star) = (n/E^{\star 2})\,v_\tau$; it rises as noise falls, because lower noise speeds up the learning of every weight. Write $\alpha_s \equiv (1 + \alpha)\Psi'_s$ for the value of an initial unit fall in noise, once the loop has amplified it; $\alpha_s \leq \alpha$, with equality at $p = 0$. In Figure 3, post-prediction scope adds a second arrow into noise, $-\omega_p(p)\ell_p/E^\star$, where $\ell_p \equiv \ell(\tau^\star \omega_p(p))$, and no arrow into value.

The marginal values of Section 4 carry over with α_s in place of α: $\partial\mathcal{V}/\partial n = \alpha_s E^\star/n$ and $\partial\mathcal{V}/\partial s = \omega_s(s)\ell_s(1 + \alpha_s)$ (Lemma 3). One more observation therefore lowers the prediction error for target individuals by $\alpha_s E^\star/n$, while it lowers validation error by $\alpha E^\star/n$. A post-prediction covariate is all spillover:

$$\frac{\partial \mathcal{V}}{\partial p} = \underbrace{\Psi'_s}_{\substack{\text{value of}\\ \text{lower noise}}}\ \underbrace{\omega_p(p)\,\ell_p}_{\text{impact}}\ \underbrace{(1+\alpha)}_{\text{multiplier}} = \alpha_s\,\omega_p(p)\,\ell_p > 0. \tag{8}$$

Post-prediction covariates and observations are two inputs into one product, precision, and

differ only in their impacts. A firm can therefore price a post-prediction covariate in observations without knowing any signal: it is worth as many observations as the sample size times the rise in validation error that deleting its term from the model, without refitting, causes, divided by validation error (Corollary 5(i) in Appendix E).

A rule that lets a firm learn from a covariate in its sample but bars its use at prediction, such as a purpose limitation, turns a covariate into a post-prediction covariate. It leaves the sample, and with it learning noise, unchanged, so it costs the firm exactly the direct effect of the covariate; a ban on collecting the covariate also costs its spillover. A restriction on use therefore costs the fraction $1/(1+\alpha_s)$ of what a ban on collection costs. The comparison is marginal: it concerns one covariate.

5.2 Increasing Returns to Observations

Post-prediction covariates can overturn diminishing returns to observations. In small samples, returns to scale increase if the squared signal of the post-prediction covariates exceeds a critical mass, and diminish if it falls short. If every post-prediction covariate carries at least as much signal as every covariate, value is S-shaped in scale when the critical mass is reached and concave otherwise (Proposition 4).

The competition of Section 4.1 still decides the curvature: Proposition 1(iii) holds with $\bar{\ell} \equiv \alpha(1-\tilde{\ell}) < \alpha$, where

$$\tilde{\ell} \equiv \frac{\int_0^s \ell\big(\tau^\star \omega_s(u)\big)^3 du + \int_0^p \ell\big(\tau^\star \omega_p(u)\big)^3 du}{\int_0^s \ell\big(\tau^\star \omega_s(u)\big)^2 du + \int_0^p \ell\big(\tau^\star \omega_p(u)\big)^2 du}$$

is the average fraction learned of all the weights in the sample, each weighted by its learning speed (Appendix B), so returns to scale increase if and only if $\bar{\ell} > \bar{\ell}_s$. Post-prediction covariates add their speeds to the loop, and so raise the amplification, and add nothing to the fall in the value of precision, because the firm does not apply their weights. Since $\bar{\ell} < \alpha$, returns to scale increase only while the firm has learned, on average, less than the fraction α of the weights that it applies, and they diminish in large samples, where the amplification falls toward zero.

In small samples, learning noise is close to one, so the firm learns a weight of signal ω at the speed $n\omega^2$, and the strength of the loop is about nQ, where $Q \equiv \int_0^s \omega_s(u)^2 du + \int_0^p \omega_p(u)^2 du$ is the squared signal of the sample: each observation raises the amplification, and the impact, by about Q. The value of precision starts at $\int_0^s \omega_s(u)^2 du$ and falls by $2\int_0^s \omega_s(u)^3 du$ per unit of precision, because the firm learns first the weights with more signal. Hence $\partial^2 \mathcal{V}/\partial n^2 \to 2\big(Q\int_0^s \omega_s(u)^2 du - \int_0^s \omega_s(u)^3 du\big)$ (Appendix B), and returns to scale increase in small samples if the post-prediction covariates reach the *critical mass*,

$$\int_0^p \omega_p(u)^2 du > \kappa(s) \equiv \frac{\int_0^s \omega_s(u)^3 du}{\int_0^s \omega_s(u)^2 du} - \int_0^s \omega_s(u)^2 du, \tag{9}$$

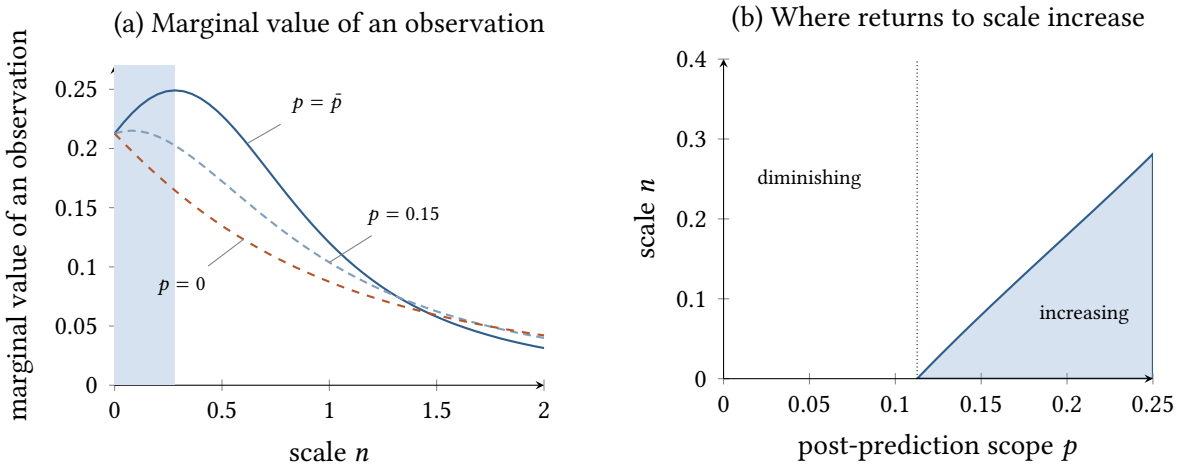


Figure 6. Returns to scale and post-prediction covariates. Example of Figure 5, $s = 0.5$. (a) Marginal value of an observation without post-prediction covariates, with $p = 0.15$ and with all of them; shaded, increasing returns to scale at $p = \bar{p}$. (b) The combinations of post-prediction scope and scale at which returns to scale increase; dotted, the post-prediction scope $\hat{p}$ above which they increase at small scale.

which is positive by the Cauchy–Schwarz inequality, because the covariates carry less than all the signal of the target variable.

Proposition 4 (Increasing returns to observations)**.**

(i) If $\int_0^p \omega_p(u)^2 du > \kappa(s)$*, returns to scale increase at every sufficiently small* n*; if* $\int_0^p \omega_p(u)^2 du < \kappa(s)$*, they diminish at every sufficiently small* n*.*

(ii) Returns to scale diminish at every sufficiently large n*.*

(iii) Let every post-prediction covariate in the sample carry at least as much signal as every covariate in the sample. If $\int_0^p \omega_p(u)^2 du \leq \kappa(s)$*, returns to scale diminish at every* n*. If* $\int_0^p \omega_p(u)^2 du > \kappa(s)$*, there is* $\hat{n} > 0$ *such that returns to scale increase at every* $n < \hat{n}$ *and diminish at every* $n > \hat{n}$*.*

Under the reverse ranking, returns to scale diminish at every scale, as in the baseline, because the firm then learns first the weights it applies (Lemma 4(i)).[14] Increasing returns to scale therefore require post-prediction covariates that carry more signal than some covariates the firm applies, and, under the ranking of part (iii), enough of them. The critical mass is low when each covariate carries little signal: if the first s covariates all have signal c, then $\kappa(s) = c(1 - cs)$. Since $\int_0^p \omega_p(u)^2 du$ rises with post-prediction scope, the critical mass is reached once p exceeds a threshold $\hat{p}$, if all post-prediction covariates together exceed it. In Figure 6, $\hat{p} = 0.11$; with $p = 0.15$, returns to scale increase up to $n = 0.08$, and with all post-prediction covariates up to $n = 0.28$.

[14]Without the ranking of part (iii), parts (i) and (ii) still hold, but returns to scale can change sign more than once: in the example of Online Appendix OA.3.5, they diminish, increase and diminish again. That appendix also gives a sufficient condition for diminishing returns to scale at every scale that does not rank signals.

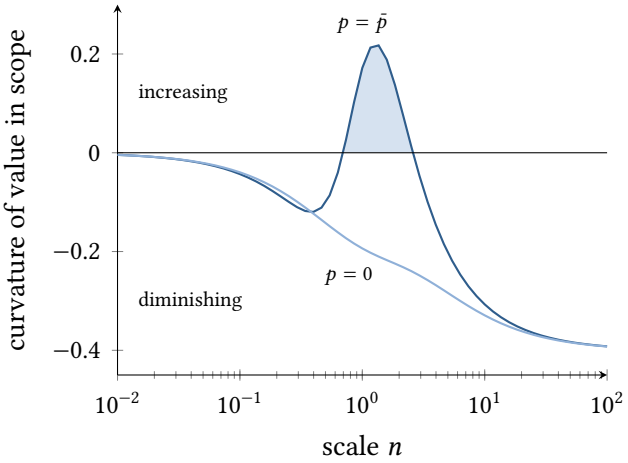


Figure 7. Post-prediction covariates and returns to scope. Example of Figure 5, $s = 0.5$: the curvature of value in scope, $\partial^2 \mathcal{V}/\partial s^2$, with all post-prediction covariates and without them (light); shaded, returns to scope increase.

In search, each term of a query carries little signal about which result the user will click, and the next search carries much: it often reveals what the user wanted. The ranking of part (iii) is then natural, and the critical mass is low. The next search is correlated with the query, so in the model the post-prediction covariates stand for the part of the user's later behavior that the query does not predict; residualizing the covariates in this way can correlate their weights, which the model rules out. The mapping shows that the mechanism can operate in search, and whether it does is an empirical question. In simulations, increasing returns to scale appear already with a few hundred covariates and hold with non-Gaussian data (Online Appendix OA.2.4).

5.3 Post-Prediction Covariates and Scope

Proposition 5 (Scope and post-prediction scope)**.** *Scope and post-prediction scope are complements,* $\partial^2 \mathcal{V}/\partial s\, \partial p > 0$.

Every factor of Equation (8) rises with scope: a new covariate adds a weight that gains from lower noise; it lowers noise, which raises the value of a further fall and the learned signal of the post-prediction covariate; and it raises the amplification. A search engine with long user histories gains more from recording what each user searched for next. Proposition 2(ii) and Corollary 1 hold with α_s in place of α (Appendix B): declining signal still wins at both ends of the scale, and post-prediction covariates help the precision terms in between, because they raise the multiplier. In Figure 7, at $s = 0.5$, returns to scope diminish at every scale without post-prediction covariates and increase for n between 0.70 and 2.6 with all of them. Data that the firm cannot use when it predicts can therefore create increasing returns to the data that it can. Appendix E collects the returns to post-prediction scope, which increase more easily than returns to scope, their interaction with scale, and economies of scope among groups of post-prediction covariates.

6 The Market for Data

A firm rarely owns all the data it learns from. Individuals hold their own observations, and other firms hold covariates on the firm's users, such as a broker's demographic file or a bank's transaction history. How the value of a dataset is divided among those who hold its parts depends on whether the parts are substitutes or complements (Horn and Wolinsky 1988), and Sections 4 and 5 supply the answer. Groups of covariates are complements, and they exhibit economies of scope (Proposition 3). Observations without post-prediction covariates are substitutes, because returns to scale then diminish (Proposition 1(i)); observations that carry a critical mass of post-prediction covariates are complements while they are scarce, because returns to scale then increase (Proposition 4). Post-prediction covariates are often recorded with the target variable, so they travel with observations: a search session records the query, the result the user clicked and the user's next search.[15]

6.1 Holders and Bargaining

I work in the limit of Section 3.2 under the assumptions of Section 3.3. There are $m \geq 2$ holders, and the firm earns one unit per unit of value.[16] Write $R^2 \equiv 1 - E^\star$ for the *validation* R^2, the expected out-of-sample R^2 of the fitted model (Theorem 1(ii)); without post-prediction covariates, value equals it and $\alpha_s = \alpha$.

Three economies. In the economy of observations, the firm observes the covariates of the target individual and obtains its training sample from m *holders of observations*: holder h holds the target variables, covariates and post-prediction covariates of a set of individuals of scale $n_h > 0$, with $\sum_h n_h = n$, as an individual holds her own observation, a hospital its patients' observations and a publisher its texts. In the economy of covariates, the firm holds the target variable of individuals at scale n, without post-prediction covariates, and obtains covariates from m *holders of covariates*: holder h holds, for the training individuals and the target individual, the covariates whose positions lie in a finite union of intervals $\mathcal{S}_h$, as a broker holds a file on the firm's customers. The sets $\mathcal{S}_h$ are disjoint, each has positive length, and their union is $[0, s]$. In the joint economy, without post-prediction covariates, holders of observations hold the target variables of their individuals and holders of covariates hold the covariates; the firm can use individual i's covariate k only if it reaches agreement both with the holder of i's observation and with the holder of k, as when data-protection law requires an individual's consent for the use of a broker's file on her.

[15] A lender, by contrast, learns about repayments to other lenders from a credit bureau, which then holds the post-prediction covariates.

[16] If the firm serves a mass μ of target individuals and earns θ per unit of value, every value, payment and payoff below is multiplied by $\theta\mu$. Data are nonrival, so the firm uses one dataset for every target individual.

Value and marginal contributions. A coalition $\mathcal{H}$ of holders supplies a dataset. Its value $W(\mathcal{H})$ is $\mathcal{V}(n_{\mathcal{H}}, s, p)$, where $n_{\mathcal{H}} \equiv \sum_{h\in\mathcal{H}} n_h$, in the economy of observations; $\mathcal{V}\big(\bigcup_{h\in\mathcal{H}} \mathcal{S}_h\big)$, the value at scale n of observing these covariates for the training and the target individuals (Section 4.3), in the economy of covariates; and the value at the scale of the observations in $\mathcal{H}$ of the covariates in $\mathcal{H}$ in the joint economy. The *marginal contribution* of holder h is

$$\Delta_h \equiv W(\mathcal{H}_{\text{all}}) - W(\mathcal{H}_{\text{all}} \setminus \{h\}),$$

where $\mathcal{H}_{\text{all}}$ is the set of all holders. Holdings are *atomistic* in the limit as the largest holding, $\max_h n_h$ or $\max_h \int_{\mathcal{S}_h} \omega_s(u)\, du$, tends to zero, taken after the large-K limit.

Bargaining. The firm negotiates with every holder simultaneously and bilaterally. As in Collard-Wexler et al. (2019), each pair splits the gains from its own agreement by the generalized Nash bargaining solution, taking the other agreements as given. The weight on the holder is $\varphi \in (0, 1]$, and its disagreement payoff is zero, because data are nonrival and costless to supply.[17] Holder h therefore receives $\varphi\Delta_h$, and the firm's payoff is $\Pi \equiv W(\mathcal{H}_{\text{all}}) - \varphi \sum_h \Delta_h$. With $\varphi = 1$, every holder captures its full marginal contribution, as when it makes a take-it-or-leave-it offer. Payment by marginal contribution is also how machine-learning practice values data (Ghorbani and Zou 2019).

6.2 Product Exhaustion

Under constant returns, factors paid their marginal products exactly exhaust output (Wicksteed 1894). Data do not have constant returns, and the direction in which they depart from them depends on who holds the data and on what the data carry. When the post-prediction covariates reach the critical mass, write $\hat{n}$ for the scale up to which returns to scale increase, as in Proposition 4(iii).[18]

Proposition 6 (Holders of observations)**.** *Without post-prediction covariates:*

(i) marginal contributions fall short of value,

$$\alpha E^\star < \sum_{h=1}^{m} \Delta_h < \mathcal{V};$$

(ii) their sum falls strictly when a holding is split between two holders;

[17] A holder that also sells its data to other firms keeps those sales whether or not it agrees, so they cancel from its gains. A cost c_h of supplying the data, such as a privacy cost, raises the payment to $c_h + \varphi(\Delta_h - c_h)$.

[18] Without the ranking of Proposition 4(iii), returns to scale still increase at every scale below some $\hat{n} > 0$ when the critical mass is reached (Proposition 4(i)), and the results below hold with that $\hat{n}$.

(iii) as holdings become atomistic, the sum converges to $\alpha E^\star$, the share

$$\frac{\alpha E^\star}{\mathcal{V}} = \alpha \frac{1 - R^2}{R^2} < 1$$

of value.

If the observations carry post-prediction covariates that reach the critical mass, Inequality (9)*, then at every $n < \hat{n}$:*

(iv) as holdings become atomistic, marginal contributions sum to $\alpha_s E^\star > \mathcal{V}(n, s, p)$;

(v) when two holders of observations merge, their payment falls strictly.

Without post-prediction covariates, observations are substitutes: each is worth less to a firm that holds more of them, so a holder's marginal contribution falls short of its average contribution, and the firm keeps $\mathcal{V} - \alpha E^\star$ as a rent. Atomistic holders receive $n\,\partial\mathcal{V}/\partial n$, so their share of value is the elasticity of value with respect to scale, and the learning multiplier pins it down: when the sample grows by one percent, validation error falls by α percent, and value rises by $0.01\,\alpha E^\star$. The share tends to one as $n \to 0$, where value is proportional to scale, and to zero as $n \to \infty$: an individual's observation is worth little at the margin of a large sample, although all observations together are worth all of the value. With post-prediction covariates, value rises by $0.01\,\alpha_s E^\star$, the fall in the prediction error for target individuals, and the elasticity exceeds one when value is strictly convex in scale up to n. Below $\hat{n}$, observations that carry a critical mass of post-prediction covariates are therefore complements: a search query with its clicked result is a substitute for other queries, and a session that also records the user's next search is a complement.

Proposition 7 (Holders of covariates)**.**

(i) Marginal contributions exceed value:

$$\mathcal{V} < \sum_{h=1}^{m} \Delta_h < (1 + \alpha)\,\mathcal{V}.$$

(ii) Their sum rises strictly when a holding is split in two.

(iii) As holdings become atomistic, the sum converges to $(1+\alpha)\mathcal{V}$, the share $1+\alpha > 1$ of value.

A group of covariates is worth more to a firm that holds more groups (Proposition 3), so marginal contributions sum to more than value. By Lemma 3(ii), the marginal contribution of an atomistic group is its learned signal plus its spillover. Each group claims the spillovers it creates, yet the firm learns every weight at the precision that all groups raise together, so the claims add up to $(1 + \alpha)\mathcal{V}$, and the excess, $\alpha\mathcal{V}$, is the sum of the spillovers. The excess

vanishes at small scale, where a group barely raises precision, and at large scale, where $\alpha \to 0$ (Section 4.2), so it is largest at intermediate scale.[19]

Corollary 2 (Observations and covariates)**.** *In the joint economy, as both kinds of holdings become atomistic, marginal contributions sum to*

$$\alpha E^\star + (1+\alpha)\mathcal{V} = \mathcal{V} + \alpha = \left(1 + \frac{\alpha}{R^2}\right)\mathcal{V}.$$

When individuals and holders of covariates must both agree, marginal contributions exceed value by α: the elasticity of the learning curve times the variance of the target variable, which is the firm's prediction error without data.

6.3 Payments, Unions, and Integration

Proposition 8 (Division of value)**.** *Let holdings be atomistic.*

(i) When it buys observations without post-prediction covariates, the firm keeps the share $1 - \varphi\,\alpha(1-R^2)/R^2$ *of value, which exceeds* $1-\varphi$*.*

(ii) When it buys covariates, the firm keeps the share $1 - \varphi(1+\alpha)$*, which is negative if* $\varphi > 1/(1+\alpha)$*.*

(iii) When it buys observations that carry post-prediction covariates of critical mass, at $n < \hat{n}$*, the firm keeps the share* $1 - \varphi\,\alpha_s E^\star/\mathcal{V}$ *of value, which is below* $1-\varphi$*.*

Moreover, with $\varphi = 1$*, payments to holders of covariates exceed value for every* m*.*

When $\varphi(1+\alpha) > 1$, the payments that bilateral bargaining assigns exceed the value they divide, so bargaining with every holder cannot close: this is the anticommons of Heller (1998) and Buchanan and Yoon (2000), measured by α. A firm that buys observations that carry a critical mass of post-prediction covariates faces it in small samples, once $\varphi\,\alpha_s E^\star > \mathcal{V}$. At the values of α that Hestness et al. (2017) estimate, between 0.07 and 0.35 (Section 4.2), bargaining with atomistic holders of covariates would leave the firm a nonnegative payoff only if φ were below a threshold between 0.74 and 0.93. At $\alpha = 0.2$ and $R^2 = 0.5$, the firm keeps $1 - 0.2\,\varphi$ of value when it buys observations without post-prediction covariates and $1 - 1.2\,\varphi$ when it buys covariates: 90 and 40 percent at equal bargaining weights. Table 1 summarizes.

Corollary 3 (Unions and intermediaries)**.**

[19] With $\varphi = 1$, the payments are those of the Vickrey–Clarke–Groves mechanism. By the procurement analogue of Ausubel and Milgrom (2006), these payments lie in the core if and only if value is submodular in holders, as it is for observations without post-prediction covariates and is not for covariates. The game among the firm and the holders of covariates is convex, so its core is nonempty and contains the Shapley value (Shapley 1971) (Appendix D). A stable division exists; only bilateral bargaining fails.

Table 1. The market for data

	Observations		
	without post-prediction	with post-prediction, $n < \hat{n}$	Covariates
Holdings are	substitutes	complements	complements
Atomistic contributions / value	$\alpha(1-R^2)/R^2 < 1$	$\alpha_s E^\star/\mathcal{V} > 1$	$1+\alpha > 1$
Firm's share of value	$1-\varphi\,\alpha(1-R^2)/R^2$	$1-\varphi\,\alpha_s E^\star/\mathcal{V}$	$1-\varphi(1+\alpha)$
Payment after a merger of two	rises	falls	falls
Payment as one / atomistic	$R^2/[\alpha(1-R^2)] > 1$	$\mathcal{V}/(\alpha_s E^\star) < 1$	$1/(1+\alpha) < 1$

Notes: α is the elasticity of the learning curve, R^2 the validation R^2 and φ the holders' bargaining weight. "With post-prediction" refers to observations that carry a critical mass of post-prediction covariates, Inequality (9), at scales below $\hat{n}$ (Proposition 6(iv)–(v)); $\alpha_s E^\star$ is the slope, against log scale, of the firm's prediction error for target individuals. The last row compares what all holders receive when they bargain as one with what atomistic holders receive.

(i) When two holders of observations without post-prediction covariates merge, their payment rises strictly and the firm's payoff falls by the same amount. If all of them bargain as one, they receive $\varphi\mathcal{V}$, which exceeds what atomistic holders receive by the factor $R^2/[\alpha(1-R^2)]$.

(ii) When two holders of covariates merge, their payment falls strictly and the firm's payoff rises by the same amount. If all holders of covariates bargain as one, they receive $\varphi\mathcal{V}$, a fraction $1/(1+\alpha)$ of what atomistic holders receive, and the firm gains $\varphi\alpha\mathcal{V}$.

Corollary 3 is the result of Horn and Wolinsky (1988) applied to data: suppliers of substitutes gain by bargaining together, and suppliers of complements lose. Holders of observations gain from a union by a factor that is largest when the firm has learned much: at $\alpha = 0.1$ and $R^2 = 0.8$, a union multiplies their payment by 40. This is the economics behind proposals for data unions and for collective licensing of training data (Arrieta-Ibarra et al. 2018). The case for a union holds only for observations without post-prediction covariates, or in large samples: below $\hat{n}$, holders of observations that carry a critical mass of post-prediction covariates lose by merging (Proposition 6(v)), as holders of covariates do.

Holders of covariates gain by staying apart, unless $\varphi(1+\alpha) > 1$, in which case separate bargaining fails and merging lets them trade. An intermediary that buys covariates from separate holders faces the same over-claim, so the over-claim is avoided only when the covariates are generated through common ownership or coordinated licensing. This offers one reason for ecosystems of apparently unrelated services: a firm that provides several services to the same users, as Google does with Search, Gmail, Maps, Android and Google Pay, generates in-house covariates that are complements. When $\varphi(1+\alpha) > 1$, integration avoids the breakdown of bargaining; when $\varphi(1+\alpha) \leq 1$, it shifts $\varphi\alpha\mathcal{V}$ to the firm without raising value. For the same reason, increasing returns to scale are a barrier to entry: an entrant that must buy search sessions with each user's next search faces, below $\hat{n}$, the anticommons that a buyer of covariates faces, while the first sessions it records itself are worth less than later ones.

6.4 Property Rights

Rights can be vested in *holders of observations*, each individual holding her entire observation, including every covariate on her, as rights to data portability would give her; in *holders of covariates*, each source holding its group; or in *both*, so that the firm must agree with each.

Corollary 4 (Property rights over data)**.** *With atomistic holdings and without post-prediction covariates, bilateral bargaining gives holders together a share of value:*

(i) the share $\varphi\alpha(1-R^2)/R^2$ *when rights are vested in holders of observations, which is below one for every* $\varphi \in (0,1]$*;*

(ii) the share $\varphi(1+\alpha)$ *when they are vested in holders of covariates, which is at most one if and only if* $\varphi \leq 1/(1+\alpha)$*;*

(iii) the share $\varphi(1+\alpha/R^2)$ *when they are vested in both, which is at most one if and only if* $\varphi \leq R^2/(R^2+\alpha)$*.*

Bargaining closes when the share is at most one. Since $1 > 1/(1+\alpha) > R^2/(R^2+\alpha)$*, each assignment closes whenever the next one does.*

Vesting in an individual the rights over all the covariates on her integrates complements at the level of the individual, as common ownership does at the level of the firm, and bargaining with holders of observations without post-prediction covariates always closes. Portability rights, such as Article 20 of the GDPR and Article 6(9) of the Digital Markets Act, move rights from sources toward individuals; regimes that require both the individual's consent and the source's license are the most prone to breakdown. Rights vested in individuals protect efficiency but transfer little: a union of holders of observations receives $\varphi\mathcal{V}$, which never exceeds value, so collective management of individual rights redistributes without causing breakdown. When an individual's observation includes a critical mass of post-prediction covariates, such as what she did after the prediction, rights vested in her can fail as well: below $\hat{n}$, bargaining breaks down once φ exceeds $\mathcal{V}/(\alpha_s E^\star)$ (Proposition 8(iii)). A requirement of individual consent for the use of behavioral data therefore weighs most on small firms, consistent with evidence that the GDPR raised concentration among web technology vendors (Johnson et al. 2023).

6.5 The Price of Mandated Access

Rules that oblige holders to make data available on fair and reasonable terms need a compensation rule. Paying each holder of covariates its marginal contribution stacks claims, as patent royalties stack when many holders own complementary patents (Lemley and Shapiro 2007). A rule that deflates by the learning multiplier avoids this.

Proposition 9 (Learned-signal pricing)**.** *In the economy of covariates, pay holder h the amount $\lambda \int_{S_h} \omega_s(u)\, \ell\big(\tau^\star \omega_s(u)\big)\, du$, for some $\lambda \in [0,1]$.*

(i) Payments sum to $\lambda \mathcal{V}$ for every partition.

(ii) The allocation lies in the core of the game among the firm and the holders.

(iii) As holdings become atomistic, each payment approaches $\lambda/(1+\alpha)$ times the holder's marginal contribution: $\sum_h \left|\lambda \int_{S_h} \omega_s(u)\, \ell(\tau^\star \omega_s(u))\, du - \lambda \Delta_h/(1+\alpha)\right| \to 0$.

A group's learned signal is the rise in the fitted model's validation error when the group's terms are deleted without refitting, a quantity that machine-learning practice already computes. The rule is the top-down approach to fair, reasonable and non-discriminatory royalties: cap the aggregate and apportion it, here by the model. For observations without post-prediction covariates, marginal contributions already lie in the core, because these observations are substitutes.

What shared observations are worth depends on what they carry. Search sessions that keep each user's next search are complements below $\hat{n}$, so the value of a shared sample is convex in its size, and they help a rival that already holds some sessions of the same kind more than an entrant. Anonymization that severs the link between a query and the user's later searches strips the next search, so the sessions become substitutes and the first shared sessions are worth the most. The Digital Markets Act requires that the shared query, click and view data that constitute personal data be anonymized (European Union 2022, Article 6(11)), so the way the data are anonymized determines which of the two cases applies.

7 Conclusion

Do firms with more data benefit more from additional data, and who captures the value of data? The answers depend on the margin and on when the data become available. A group of covariates on the same users is worth more to a firm that already holds more covariates on these users, at every scale. A further observation can be worth more to a firm with a larger sample only if the firm's records carry post-prediction covariates, as a search engine's logs carry each user's next search. When these carry more signal than the covariates, this happens if and only if they reach a critical mass, and only while the sample is small: the pattern the court found for Google's rivals.(United States v. Google LLC 2025, p. 131).

Both answers come from learning. Each covariate in the sample lowers learning noise, the validation error of the model, so the firm learns every weight better, and the learning multiplier, one plus the elasticity of the learning curve, measures this feedback. Post-prediction covariates act only through learning noise, as observations do: indeed, they can substitute for observations in large samples, complement covariates, and are worth most at intermediate scale.

The same curvature divides the value of data among those who hold its parts. Covariates are complements: their holders claim more than value, and bargaining with them breaks down when they are strong, unless the services that generate the covariates are under common ownership or holders are paid the learned signal of their groups. Observations without post-prediction covariates are substitutes: their holders receive less than value and gain by bargaining together, and rights that give each individual her entire observation ensure that bargaining closes. Observations that carry a critical mass of post-prediction covariates are complements in small samples, so an entrant that must buy them faces the breakdown that a buyer of covariates faces, and the way shared data are anonymized decides whether a sharing mandate helps the smallest rivals most.

Future work could price observations and covariates, so that the value function becomes the benefit side of a firm's choice of dataset, and let better predictions attract users who supply further observations, to show how these returns shape market structure. Learning curves across prediction tasks would measure the learning multiplier and test a prediction of the theory: prediction error falls at an increasing rate only for firms whose samples carry post-prediction covariates.

Appendix: Proofs

This appendix proves, in order of appearance, the results of Sections 3 to 6 that the text does not prove, and states the supporting results that the proofs use where they first need them; Appendix E states and proves further results on post-prediction covariates. The proofs allow post-prediction covariates throughout. The results of Section 4, and those of Section 6 stated without post-prediction covariates, are the case $p = 0$, in which $\mathcal{V}(n, s) = \mathcal{V}(n, s, 0)$, $\Psi'_s = \Psi'$ and $\alpha_s = \alpha$. Table OA.1 in the Online Appendix collects the notation. Throughout, $\bar{\omega}$ is the bound on the signal of every potential covariate stated at the start of Appendix A.

A Proofs for Section 3

Signal profiles. In Section 3.2, $\bar{S} = \lfloor \bar{s}K \rfloor$ and $\bar{P} = \lfloor \bar{p}K \rfloor$, the profiles are measurable and bounded by some $\bar{\omega}$, with $\int_0^{\bar{s}} \omega_s(u)\,du + \int_0^{\bar{p}} \omega_p(u)\,du \leq 1$, and covariate $k \leq \bar{S}$ has signal $K \int_{(k-1)/K}^{k/K} \omega_s(u)\,du$, the average of the profile over the k-th cell of width $1/K$; post-prediction covariates are defined in the same way from ω_p. The never-observed covariates, each with signal at most $\bar{\omega}$, carry the remaining signal, and only their total enters the results.

Equalities between random variables hold almost surely.

Proof of Lemma 1. The predictor. Write $y_0 = \sum_k \beta_k x_{0k}$ and take expectations given the data term by term. For $k > S$, the firm does not observe x_{0k}, which has mean zero and is independent of β_k and of the data, so the term contributes zero. For $k \leq S$, the firm observes x_{0k}, and the covariates of the target individual are independent of the weights and of the training sample, so $\mathbb{E}[\beta_k x_{0k} \mid \boldsymbol{d}_D] = x_{0k}\,\mathbb{E}[\beta_k \mid \boldsymbol{y}_N, \boldsymbol{X}_D] = x_{0k} m_{D,k}$. Summing gives the predictor.

Value. By Equation (1) and the predictor, $V(D) = \mathbb{V}\big(\sum_{k=1}^{S} x_{0k} m_{D,k}\big)$. The covariates of the target individual are mutually independent, independent of the training sample, with mean zero and $\mathbb{E}[x_{0k}^2] = \omega_k / K$, so the cross terms vanish and

$$V(D) = \frac{1}{K} \sum_{k=1}^{S} \omega_k\, \mathbb{E}\big[m_{D,k}^2\big] = \frac{1}{K} \sum_{k=1}^{S} \omega_k\, \mathbb{V}(m_{D,k}),$$

because $\mathbb{E}[m_{D,k}] = \mathbb{E}[\beta_k] = 0$. For monotonicity, let $D \leq D'$. The training sample under D is a subvector of that under D'. The expected posterior variance of β_k is the lowest mean squared error of any estimate of β_k that is a function of the training sample, and the lowest error over a larger set of functions is weakly smaller. Since $\mathbb{V}(m_{D,k}) = \sigma^2 - \mathbb{E}\big[\mathbb{V}(\beta_k \mid \boldsymbol{y}_N, \boldsymbol{X}_D)\big]$, $\mathbb{V}(m_{D,k}) \leq \mathbb{V}(m_{D',k})$.

The argument uses the covariates in the sample and the covariates observed for the target individual only through the independence of the primitives, never through their order, so it holds for every set of covariates in the sample and every subset $\mathcal{A}$ of them observed for the target individual, with the sum over $\mathcal{A}$ in place of the sum over $k \leq S$. □

Remark 1 (Covariates observed only at prediction). Let the firm observe x_{0k} for a covariate k that is not in its sample. Flip the signs of β_k and of x_{ik} for every training individual $i \leq N$. The flip leaves the target variable y_i of every training individual unchanged, because it leaves $\beta_k x_{ik}$ and every other term unchanged, and it leaves the covariates in the sample unchanged, because covariate k is not among them. Weights and covariates are independent and symmetric about zero, so the flip preserves their joint distribution. Hence $(\beta_k, \boldsymbol{y}_N, \boldsymbol{X}_D)$ and $(-\beta_k, \boldsymbol{y}_N, \boldsymbol{X}_D)$ have the same distribution, so $\mathbb{E}[\beta_k \mid \boldsymbol{y}_N, \boldsymbol{X}_D] = 0$, and the term $x_{0k}\beta_k$ contributes $x_{0k} \cdot 0$ to the posterior mean of y_0: observing x_{0k} adds no value.

Proof of Theorem 1. At $p = 0$, $e(\tau; s, 0) = 1 - v(\tau, s)$, so the limit of value, $v(\tau^\star, s)$, equals $1 - E^\star$, and validation error equals the Bayes risk.

Notation. Fix K and suppress it when no confusion arises. Write C for the set of the $S + P$ covariates in the sample, the first S covariates and the first P post-prediction covariates, indexed $k \in C$ in matrices, $\boldsymbol{X} \equiv \boldsymbol{X}_{D_K}$, $\boldsymbol{G} \equiv \boldsymbol{X}^\top \boldsymbol{X}$, $\boldsymbol{x}_k$ for the column of $\boldsymbol{X}$ of covariate k, $\boldsymbol{X}_{-k}$ for $\boldsymbol{X}$ without that column, and $\boldsymbol{z}_C \equiv (z_k)_{k \in C}$ for the coordinates of a vector $\boldsymbol{z}$ in C. Let $n_K \equiv N/K$, and let $\bar{c} \equiv \frac{1}{K} \sum_{k \notin C} \omega_k$ be the signal outside the sample at finite K. Step 5 shows that $\bar{c} \to \underline{E}(s, p) > 0$, so $\bar{c} \geq c_0 \equiv \underline{E}(s, p)/2$ and $n_K \in [n/2, 2n]$ for K large. The letter C denotes a positive constant that depends only on σ^2, c_0, $\bar{\omega}$ and n; its value may change from line to line. Steps 2 to 4 keep σ^2 general; the limits of value and validation error in part (ii) take $\sigma^2 = 1$.

Step 1: the limit equation has a unique solution, and $\Psi' < 1$ there. Write $\int_C f(\omega)\, du \equiv \int_0^s f\big(\omega_s(u)\big)\, du + \int_0^p f\big(\omega_p(u)\big)\, du$ within this proof. The error $e(\tau; s, p) = \underline{E}(s, p) + \int_C \omega \big(1 - \ell(\tau\omega)\big)\, du$ is continuously differentiable in τ on $[0, \infty)$, with $e_\tau = -\int_C \omega^2/(1 + \tau\omega)^2\, du \leq 0$, $e(0; s, p) = 1$ and $e(\tau; s, p) \to \underline{E}(s, p)$ as $\tau \to \infty$ by dominated convergence, where $\underline{E}(s, p) > 0$ by Assumption 1; so e maps into $[\underline{E}(s, p), 1]$. Since $\tau\omega\big(1 - \ell(\tau\omega)\big) = \ell(\tau\omega)$,

$$\tau\, e(\tau; s, p) = \underline{E}(s, p)\, \tau + \int_C \ell(\tau\omega)\, du,$$
$$\frac{d}{d\tau}\big[\tau\, e(\tau; s, p)\big] = \underline{E}(s, p) + \int_C \omega\big(1 - \ell(\tau\omega)\big)^2 du \geq \underline{E}(s, p),$$

so $\tau e(\tau; s, p)$ rises strictly and continuously from zero without bound, and $n = \tau e(\tau; s, p)$ has a unique solution $\tau^\star(n, s, p) > 0$. Then $E^\star(n, s, p) \equiv n/\tau^\star = e(\tau^\star; s, p)$ lies in $[\underline{E}(s, p), 1]$, and the system of Theorem 1(i) has the unique solution $\big(\tau^\star, E^\star, v(\tau^\star, s)\big)$. In terms of noise, $\Psi(E; n, s, p) = e(n/E; s, p)$ for $E > 0$ is continuously differentiable, with

$$\Psi'(E; n, s, p) = \frac{n}{E^2}\big(-e_\tau\big) = \int_C \frac{n\omega^2}{(n\omega + E)^2}\, du \geq 0,$$
$$\frac{\partial^2 \Psi(E; n, s, p)}{\partial E^2} = -\int_C \frac{2n\omega^2}{(n\omega + E)^3}\, du \leq 0,$$

and it extends continuously to $E = 0$ by $\Psi(0; n, s, p) = \underline{E}(s, p)$, the limit of e as $\tau \to \infty$; so

$\Psi(\cdot; n, s, p)$ is nondecreasing and concave on $[0, \infty)$. At $\tau^\star$, the derivative of $\tau e(\tau; s, p)$ equals $e + \tau e_\tau = E^\star\big(1 - \Psi'(E^\star; n, s, p)\big)$, which is positive, so $\Psi'(E^\star; n, s, p) < 1$. Since $E < E^\star$ if and only if $n/E > \tau^\star$, that is, $(n/E)\, e(n/E; s, p) > n$,

$$E < \Psi(E; n, s, p) \;\text{ for } E < E^\star(n, s, p), \qquad E > \Psi(E; n, s, p) \;\text{ for } E > E^\star(n, s, p). \tag{10}$$

The step uses only that the profiles are bounded and nonnegative and that the signal outside the sample is positive.

Step 2: the covariates outside the sample act as Gaussian noise whose variance is their signal. The target variable of each training individual is the part that the covariates in the sample explain plus a residual that collects the covariates outside C: covariates and post-prediction covariates outside the sample and never-observed ones,

$$y_i = \boldsymbol{x}_{iC}^\top \boldsymbol{\beta}_C + w_i, \qquad w_i \equiv \sum_{k \notin C} \beta_k x_{ik}.$$

Given the weights outside C, the residuals are i.i.d. $\mathcal{N}(0, \sigma^2 c)$, independent of the weights in C and of $\boldsymbol{X}$, with

$$c \equiv \frac{1}{K} \sum_{k \notin C} \omega_k \frac{\beta_k^2}{\sigma^2}, \qquad \mathbb{E}[c] = \bar{c}.$$

A firm that knew c would face a Gaussian linear regression with known noise variance, and its posterior mean of the weights in C would be ridge regression with penalty c: for $\lambda > 0$, let $\tilde{\boldsymbol{m}}(\lambda) \equiv (\lambda \boldsymbol{I} + \boldsymbol{G})^{-1} \boldsymbol{X}^\top \boldsymbol{y}_N$ and $f_k(\lambda) \equiv \big[(\boldsymbol{I} + \boldsymbol{G}/\lambda)^{-1}\big]_{kk}$; then

$$\mathbb{E}[\boldsymbol{\beta}_C \mid \boldsymbol{y}_N, \boldsymbol{X}, c] = \tilde{\boldsymbol{m}}(c), \qquad \mathbb{V}(\beta_k \mid \boldsymbol{y}_N, \boldsymbol{X}, c) = \sigma^2 f_k(c).$$

The firm does not know c, but c averages $K - |C|$ independent terms and concentrates at $\bar{c}$: $c - \bar{c} = \frac{1}{K} \sum_{k \notin C} \omega_k (\beta_k^2/\sigma^2 - 1)$ averages independent centered terms with bounded moments, so $\mathbb{E}[(c - \bar{c})^4] \le C/K^2$, and the event $\mathcal{G} \equiv \{c \ge \bar{c}/2\}$ has $\Pr(\mathcal{G}^c) \le 16\, \mathbb{E}[(c - \bar{c})^4]/\bar{c}^4 \le C/K^2$. By the law of total variance, the expected posterior variance of β_k given the sample exceeds that given the sample and c by $\varepsilon_k \equiv \mathbb{E}\big[(\mathbb{E}[\beta_k \mid \boldsymbol{y}_N, \boldsymbol{X}, c] - m_{D,k})^2\big]$. The posterior mean $m_{D,k}$ is the best estimate of $\mathbb{E}[\beta_k \mid \boldsymbol{y}_N, \boldsymbol{X}, c]$ among functions of the sample, and $\tilde{m}_k(\bar{c})$ is one such function, so $\varepsilon_k \le \mathbb{E}\big[(\tilde{m}_k(c) - \tilde{m}_k(\bar{c}))^2\big]$.

On $\mathcal{G}$, both penalties are at least $\bar{c}/2 \ge c_0/2$. The derivative of the ridge estimate with respect to the penalty is $\partial_\lambda \tilde{\boldsymbol{m}}(\lambda) = -\boldsymbol{q}(\lambda)$, with $\boldsymbol{q}(\lambda) \equiv (\lambda \boldsymbol{I} + \boldsymbol{G})^{-2} \boldsymbol{X}^\top \boldsymbol{y}_N$. Given $(\boldsymbol{X}, c)$, $\boldsymbol{y}_N \sim \mathcal{N}\big(0, \sigma^2(\boldsymbol{X}\boldsymbol{X}^\top + c\boldsymbol{I}_N)\big)$, so $q_k(\lambda)$ is centered Gaussian with variance at most

$$\sigma^2 \max_{x \ge 0} \frac{x^2 + cx}{(\lambda + x)^4} \le \sigma^2 \Big(\frac{1}{16\lambda^2} + \frac{27c}{256\lambda^3} \Big) \le C(1 + c) \qquad \text{for } \lambda \ge c_0/2.$$

By the Cauchy–Schwarz inequality along the segment between c and $\bar{c}$, $(\tilde{m}_k(c) - \tilde{m}_k(\bar{c}))^2 \le$

$|c-\bar c|\,\big|\int_{\bar c}^{c} q_k(\lambda)^2\,d\lambda\big|$, so, taking expectations given $(\boldsymbol{X}, c)$ and then unconditionally,

$$\mathbb{E}\big[(\tilde m_k(c)-\tilde m_k(\bar c))^2\mathbf{1}_{\mathcal{G}}\big] \le C\,\mathbb{E}\big[(c-\bar c)^2(1+c)\big] \le C\big(\mathbb{E}[(c-\bar c)^4]\big)^{1/2}\big(\mathbb{E}[(1+c)^2]\big)^{1/2} \le \frac{C}{K}.$$

Off $\mathcal{G}$, $(\tilde m_k(c)-\tilde m_k(\bar c))^2 \le 2\tilde m_k(c)^2+2\tilde m_k(\bar c)^2$. Since $\tilde m_k(c)=\mathbb{E}[\beta_k \mid \boldsymbol{y}_N, \boldsymbol{X}, c]$, Jensen's inequality gives $\mathbb{E}[\tilde m_k(c)^4] \le \mathbb{E}[\beta_k^4] = 3\sigma^4$. Given $(\boldsymbol{X}, c)$, $\tilde m_k(\bar c)$ is centered Gaussian with variance at most $\sigma^2 \max_{x\ge 0}(x^2+cx)/(\bar c+x)^2 \le \sigma^2(1+c/\bar c)$, so $\mathbb{E}[\tilde m_k(\bar c)^4] \le C$. By the Cauchy–Schwarz inequality, both terms contribute at most $C\Pr(\mathcal{G}^c)^{1/2} \le C/K$, so $\varepsilon_k \le C/K$. For the posterior variance, $f_k(\lambda) \in [0,1]$ and $f_k'(\lambda) = \big[\boldsymbol{G}(\lambda \boldsymbol{I}+\boldsymbol{G})^{-2}\big]_{kk} \in [0, 1/(4\lambda)]$, so $|\mathbb{E}f_k(c) - \mathbb{E}f_k(\bar c)| \le \mathbb{E}|c-\bar c|/(2c_0) + \Pr(\mathcal{G}^c) \le C/\sqrt{K}$. Since $\mathbb{V}(m_{D,k})/\sigma^2 = 1 - \mathbb{E}[f_k(c)] - \varepsilon_k/\sigma^2$,

$$\max_{k\in\mathcal{C}} \big|\mathbb{V}(m_{D,k})/\sigma^2 - \ell_k^\circ\big| \le \frac{C}{\sqrt{K}}, \qquad \ell_k^\circ \equiv 1-\mathbb{E}\big[f_k(\bar c)\big]. \tag{11}$$

The same bounds show that ridge regression with penalty $\bar c$ is within $CK^{-1/2}$ of the posterior mean of each weight in $\mathcal{C}$ in root mean square.

Step 3: every weight in $\mathcal{C}$ faces the same learning noise. Let $\boldsymbol{R} \equiv (\bar c \boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}$ and $\boldsymbol{R}_{-k} \equiv (\bar c \boldsymbol{I}_N + \boldsymbol{X}_{-k}\boldsymbol{X}_{-k}^\top)^{-1}$, both with operator norm at most $1/\bar c$. By the Schur complement and the push-through identity,

$$f_k(\bar c) = \frac{1}{1+\tau_k}, \qquad \tau_k \equiv \boldsymbol{x}_k^\top \boldsymbol{R}_{-k}\boldsymbol{x}_k, \qquad \text{so} \qquad \ell_k^\circ = \mathbb{E}[\ell(\tau_k)].$$

The column $\boldsymbol{x}_k \sim \mathcal{N}(0, \frac{\omega_k}{K}\boldsymbol{I}_N)$ is independent of $\boldsymbol{R}_{-k}$, so τ_k concentrates around its conditional mean: with $g_{-k} \equiv \frac{1}{N}\operatorname{tr}\boldsymbol{R}_{-k}$,

$$\mathbb{E}[\tau_k \mid \boldsymbol{X}_{-k}] = \omega_k\, n_K\, g_{-k}, \qquad \mathbb{V}(\tau_k \mid \boldsymbol{X}_{-k}) = 2\Big(\frac{\omega_k}{K}\Big)^2 \operatorname{tr}\boldsymbol{R}_{-k}^2 \le \frac{2\bar\omega^2 N}{K^2\bar c^2} \le \frac{C}{K}.$$

The normalized trace $g \equiv \frac{1}{N}\operatorname{tr}\boldsymbol{R}$ is common to all weights, and it concentrates at its mean $\bar g_K \equiv \mathbb{E}[g]$. Removing one column changes $\operatorname{tr}\boldsymbol{R}$ by at most $1/\bar c$, so $|g-g_{-k}| \le 1/(N\bar c)$. Replacing one column by an independent copy changes $\operatorname{tr}\boldsymbol{R}$ by at most $2/\bar c$, so the Efron–Stein inequality gives $\mathbb{V}(g) \le 2|\mathcal{C}|/(N^2\bar c^2) \le C/K$. Since ℓ is 1-Lipschitz on $[0,\infty)$, the three bounds give, uniformly in $k \in \mathcal{C}$,

$$\big|\ell_k^\circ - \ell(n_K\omega_k/E_K)\big| \le \frac{C}{\sqrt{K}}, \qquad E_K \equiv \frac{1}{\bar g_K}. \tag{12}$$

Since $\boldsymbol{R} \preceq \bar c^{-1}\boldsymbol{I}$, $E_K \ge \bar c$. Since $x \mapsto 1/(\bar c + x)$ is convex, Jensen's inequality, applied to the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$ and then to the expectation, gives $\bar g_K \ge 1/\big(\bar c + \mathbb{E}[\frac{1}{N}\operatorname{tr}\boldsymbol{X}\boldsymbol{X}^\top]\big) = 1/\big(\bar c + \frac{1}{K}\sum_{k\in\mathcal{C}}\omega_k\big) = 1$, so $E_K \le 1$.

Step 4: counting degrees of freedom pins down the learning noise. Since $(\bar c\boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{R} = \boldsymbol{I}_N$, taking traces gives $\bar c \operatorname{tr}\boldsymbol{R} + \sum_{j\in\mathcal{C}} \boldsymbol{x}_j^\top\boldsymbol{R}\boldsymbol{x}_j = N$. By the Sherman–Morrison formula, $\boldsymbol{R} = \boldsymbol{R}_{-j} - \boldsymbol{R}_{-j}\boldsymbol{x}_j\boldsymbol{x}_j^\top\boldsymbol{R}_{-j}/(1+\tau_j)$, so $\boldsymbol{x}_j^\top\boldsymbol{R}\boldsymbol{x}_j = \ell(\tau_j)$: the degrees of freedom spent on weight j equal the fraction of it that the firm learns. Taking expectations and dividing by N gives the exact

identity

$$\frac{\bar{c}}{E_K} = 1 - \frac{1}{N}\sum_{j\in C} \ell_j^\circ. \tag{13}$$

Its limit is Equation (18) below: multiplied by n_K, it reads $\bar{c}\, n_K/E_K + \frac{1}{K}\sum_{j\in C} \ell_j^\circ = n_K$. Substituting (12), multiplying by $E_K \le 1$, and using $\ell(\tau\omega)/\tau = \omega\big(1-\ell(\tau\omega)\big)$ gives

$$\big|E_K - \Psi_K(E_K)\big| \le \frac{C}{\sqrt{K}}, \qquad \Psi_K(E) \equiv \bar{c} + \frac{1}{K}\sum_{j\in C} \omega_j\big(1-\ell(n_K\omega_j/E)\big).$$

Step 5: the finite fixed point converges to the limit one. It remains to show that Ψ_K converges to Ψ uniformly and that $\bar{c} \to \underline{E}(s,p)$. Let $\omega_{s,K}$ be the step function equal on the j-th cell to the signal of the j-th covariate, and define $\omega_{p,K}$ likewise for the post-prediction covariates, so that $\frac{1}{K}\sum_{j=1}^{S} f(\omega_j) = \int_0^{S/K} f(\omega_{s,K}(u))\,du$ for every f, and likewise for the post-prediction covariates. Averaging over the cells of a uniform partition is a contraction in L^1 and converges uniformly for continuous functions, which are dense in L^1; hence $\|\omega_{s,K} - \omega_s\|_{L^1} \to 0$ and $\|\omega_{p,K} - \omega_p\|_{L^1} \to 0$. With $f(\omega) = \omega$, the signal of the covariates in the sample converges to $\int_0^s \omega_s(u)\,du + \int_0^p \omega_p(u)\,du$, and the normalization gives $\bar{c} \to \underline{E}(s,p)$. The functions $f_{E,m}(\omega) \equiv \omega\big(1-\ell(m\omega/E)\big) = \omega E/(m\omega+E)$ satisfy $|\partial_\omega f_{E,m}| \le 1$ and are Lipschitz in (E,m) uniformly over $\omega \in [0,\bar{\omega}]$, $E \in [c_0,1]$ and $m \in [n/2, 2n]$. Therefore

$$\sup_{E\in[c_0,1]} \left| \int_0^{S/K} f_{E,n_K}\big(\omega_{s,K}(u)\big)\,du - \int_0^s f_{E,n}\big(\omega_s(u)\big)\,du \right| \le \|\omega_{s,K}-\omega_s\|_{L^1} + \bar{\omega}\left|\frac{S}{K} - s\right| + C|n_K - n| \longrightarrow 0,$$

and likewise for the post-prediction covariates. With $\bar{c} \to \underline{E}(s,p)$, this gives

$$\sup_{E\in[c_0,1]} |\Psi_K(E) - \Psi(E;n,s,p)| \to 0,$$

and hence $\Phi(E_K) \to 0$. Since $E_K \in [c_0,1]$ for K large and Φ is continuous with the unique zero $E^\star(n,s,p)$, every limit point of (E_K) equals $E^\star(n,s,p)$, so $E_K \to E^\star(n,s,p)$.

Part (ii). The convergence of the fractions learned follows from (11), (12), $E_K \to E^\star(n,s,p)$ and $n_K \to n$, because $(\omega,m,E) \mapsto \ell(m\omega/E)$ is uniformly Lipschitz on $[0,\bar{\omega}]\times[n/2,2n]\times[c_0,1]$. For value, where $\sigma^2 = 1$, Lemma 1 gives $V(D_K) = \frac{1}{K}\sum_{k=1}^{S} \omega_k \mathbb{V}(m_{D_K,k})$. By the convergence of the fractions learned,

$$\Big|V(D_K) - \frac{1}{K}\sum_{k=1}^{S} \omega_k\, \ell(\tau^\star\omega_k)\Big| \le \bar{\omega}\, \max_{k\in C} \big|\mathbb{V}(m_{D_K,k}) - \ell(\tau^\star\omega_k)\big| \longrightarrow 0.$$

Since $|\partial_\omega[\omega\,\ell(\tau\omega)]| \le 1$, the argument of Step 5 shows that the sum converges to $\int_0^s \omega_s(u)\,\ell(\tau^\star\omega_s(u))\,du$. The same argument applies to any set of covariates observed at prediction whose positions form a finite union of intervals within the positions of the sample, such as the whole sample, which the firm observes for a held-out individual. By Lemma 1 for

that set, which holds for every subset of the covariates in the sample observed for the individual whose target variable the firm predicts, one minus validation error is $\frac{1}{K}\sum_k \omega_k \mathbb{V}(m_{D_K,k})$, the sum over the $S+P$ covariates in the sample, so it converges to $1-\Psi(E^\star; n, s, p) = 1-E^\star$. □

B Proofs for Section 4

From here on, $\sigma^2 = 1$, and the assumptions of Section 3.3 hold unless stated otherwise; when $p > 0$, the post-prediction profile is also positive and continuously differentiable near p whenever I differentiate with respect to p. The results hold for every p; at $p = 0$, the averages $\langle\cdot\rangle$ and $\langle\cdot\rangle_s$ below coincide, $\Psi'_s = \Psi'$ and $\alpha_s = \alpha$. In integrals over positions, ω stands for $\omega_s(u)$ on $[0, s]$ and for $\omega_p(u)$ on $[0, p]$. In proofs, $M \equiv 1 + \alpha$ denotes the learning multiplier. Then $\mathcal{V}(n, s, p) = \int_0^s \omega\, \ell\big(n\omega/E^\star(n, s, p)\big)\, du$ is continuously differentiable in n, and continuously differentiable in (n, s, p) on every open set of margins on which ω_s is continuous at s and ω_p at p (proof of Lemma 3). The next lemma collects the derivatives of the fraction learned and of the learned signal at noise E, and that of Ψ'; in it, ℓ stands for $\ell(n\omega/E)$.

Lemma 2 (Derivative identities)**.** *For $n > 0$ and $E > 0$, with $\omega \geq 0$ in (a), (c) and (e) and $\omega > 0$ in (b):*

(a) $\dfrac{\partial \ell}{\partial E} = -\dfrac{\ell(1-\ell)}{E}$ *and* $\dfrac{\partial \ell}{\partial n} = \dfrac{\ell(1-\ell)}{n}$;

(b) $\dfrac{\partial \ln(\omega\ell)}{\partial \ln \omega} = 2 - \ell$ *and* $\dfrac{\partial \ln(\omega\ell)}{\partial E} = -\dfrac{1-\ell}{E}$;

(c) $\dfrac{\ell^2}{n} = \dfrac{\omega\,\ell(1-\ell)}{E}$;

(d) $\dfrac{\partial \Psi'(E; n, s, p)}{\partial E} = -\dfrac{2}{nE}\Big[\int_0^s \ell\Big(\dfrac{n\omega}{E}\Big)^2\Big(1-\ell\Big(\dfrac{n\omega}{E}\Big)\Big)du + \int_0^p \ell\Big(\dfrac{n\omega}{E}\Big)^2\Big(1-\ell\Big(\dfrac{n\omega}{E}\Big)\Big)du\Big]$;

(e) the learned signal satisfies

$$\frac{\partial(\omega\ell)}{\partial E} = -\frac{\ell^2}{n}, \qquad \frac{\partial(\omega\ell)}{\partial n} = \frac{E}{n}\cdot\frac{\ell^2}{n}. \tag{14}$$

Proof. With $\tau = n/E$, $d\ell(\tau\omega)/d\tau = \omega/(1+\tau\omega)^2 = \ell(1-\ell)/\tau$, and $\partial \ln\tau/\partial E = -1/E$, $\partial \ln\tau/\partial n = 1/n$, which gives (a). For (b), $\ln(\omega\ell) = \ln\tau + 2\ln\omega - \ln(1+\tau\omega)$. For (c), both sides equal $n\omega^2/(n\omega+E)^2$. For (d), $\Psi'(E; n, s, p)$ is $1/n$ times the integrals of ℓ^2 over the two kinds of covariates; differentiate under the integrals and use (a). For (e), $\partial(\omega\ell)/\partial E = -n\omega^2/(n\omega+E)^2$ and $\partial(\omega\ell)/\partial n = \omega^2 E/(n\omega+E)^2$. □

The next lemma collects the marginal values of the three margins.

Lemma 3 (Marginal values)**.**

(i) Precision rises with every margin:

$$\frac{\partial \ln \tau^\star}{\partial \ln n} = 1 + \alpha, \qquad \frac{\partial \ln \tau^\star}{\partial s} = (1+\alpha)\,\frac{\omega_s(s)\,\ell_s}{E^\star}, \qquad \frac{\partial \ln \tau^\star}{\partial p} = (1+\alpha)\,\frac{\omega_p(p)\,\ell_p}{E^\star}.$$

(ii) Value rises with every margin:

$$\frac{\partial \mathcal{V}}{\partial \ln n} = \alpha_s\, E^\star, \qquad \frac{\partial \mathcal{V}}{\partial s} = \omega_s(s)\,\ell_s\,(1+\alpha_s), \qquad \frac{\partial \mathcal{V}}{\partial p} = \omega_p(p)\,\ell_p\,\alpha_s.$$

Proof of Lemma 3. (i). The map $\Psi(E; n, s, p)$ is continuously differentiable in (E, n) for $E > 0$, with $\partial\Psi/\partial n = -(E/n)\,\Psi'(E; n, s, p)$ by (14). On an open interval of positions on which ω_s is continuous, it is also continuously differentiable in s, with $\partial\Psi/\partial s = -\omega_s(s)\,\ell(n\omega_s(s)/E)$, and likewise in p. Because $1 - \Psi'(E^\star; n, s, p) > 0$ (Step 1 of the proof of Theorem 1), the implicit function theorem applies to $E = \Psi(E; n, s, p)$ and gives $\partial E^\star/\partial x = M\,\partial\Psi/\partial x$, for $x = n$ at every (s, p) and for $x = s$ or $x = p$ wherever the profile is continuous on a neighborhood of the marginal position. With $\Psi' M = M - 1 = \alpha$, this gives the derivatives of $E^\star$, and $\tau^\star = n/E^\star$ gives $\partial \ln \tau^\star/\partial \ln n = 1 - \partial \ln E^\star/\partial \ln n = 1 + \alpha$ and $\partial \ln \tau^\star/\partial x = -\partial \ln E^\star/\partial x$ for $x = s, p$. The arrows of Figure 3 are the partial derivatives of Equation (3) with scale, noise and precision in logs: $\ln \tau = \ln n - \ln E$ gives 1 and -1; $\partial \ln e/\partial \ln \tau = \tau e_\tau/e$, which is $-\Psi'$ at $\tau^\star$ by Step 1 of the proof of Theorem 1; at $\tau^\star$, $\partial \ln e/\partial s = -\omega_s(s)\,\ell_s/E^\star$ and $\partial \ln e/\partial p = -\omega_p(p)\,\ell_p/E^\star$, the arrow that Section 5.1 adds; and scope moves the upper limit of the integral that defines value, which adds $\omega_s(s)\,\ell(\tau\omega_s(s))$ at fixed precision.

(ii). The fraction learned of each weight, $\ell(\tau^\star\omega)$, depends on scale and on post-prediction scope only through $\tau^\star$, with elasticity $1 - \ell$. Differentiating $\mathcal{V} = \int_0^s \omega\,\ell(\tau^\star\omega)\,du$ under the integral, which the bounded derivatives in (14) permit because $E^\star \geq \underline{E}(s, p) > 0$, and using Lemma 2(c), the effect of precision on value at fixed scope is

$$\Gamma \equiv \int_0^s \omega\,\ell(\tau^\star\omega)\big(1 - \ell(\tau^\star\omega)\big)\,du = \frac{E^\star}{n}\int_0^s \ell(\tau^\star\omega)^2\,du = E^\star\Psi'_s = \frac{\alpha_s E^\star}{M}, \tag{15}$$

which is the arrow from precision to value in Figure 3. The chain rule with (i) gives

$$\frac{\partial \mathcal{V}}{\partial \ln n} = M\Gamma = \alpha_s E^\star, \qquad \frac{\partial \mathcal{V}}{\partial p} = \Gamma\,\frac{\partial \ln \tau^\star}{\partial p} = \omega_p(p)\,\ell_p\,\alpha_s.$$

Scope also moves the upper limit of the integral, which adds $\omega_s(s)\,\ell_s$, so $\partial\mathcal{V}/\partial s = \omega_s(s)\,\ell_s(1 + \alpha_s)$. □

Averages. For a function g on $[0, 1]$, write

$$\langle g(\ell)\rangle \equiv \frac{\int_0^s \ell\big(\tau^\star\omega_s(u)\big)^2\, g\big(\ell(\tau^\star\omega_s(u))\big)\,du + \int_0^p \ell\big(\tau^\star\omega_p(u)\big)^2\, g\big(\ell(\tau^\star\omega_p(u))\big)\,du}{\int_0^s \ell\big(\tau^\star\omega_s(u)\big)^2\,du + \int_0^p \ell\big(\tau^\star\omega_p(u)\big)^2\,du}$$

for the average of g of the fraction learned over the sample, each covariate weighted by its contribution to the slope Ψ', and $\langle g(\ell)\rangle_s$ for the average over the covariates observed for the target individual, which keeps only the first integral in the numerator and in the denominator. Thus $\langle \ell\rangle$ and $\langle \ell\rangle_s$ are $\tilde{\ell}$ and $\bar{\ell}_s$ of Section 4 (Steps 1 and 2 below), and $\bar{\ell} = \alpha\langle 1-\ell\rangle$.

Returns to scale. Throughout the next results, write $\underline{E} \equiv \underline{E}(s,p) > 0$; for a precision τ and a signal ω, write $\ell \equiv \ell(\tau\omega)$, which is the fraction learned at $\tau = \tau^\star$, and $\psi \equiv \omega(1-\ell) = \omega/(1+\tau\omega)$ for the unlearned signal of a weight, so that $\tau\psi = \ell$ and $\partial\psi/\partial\tau = -\psi^2$. Write $\int_C f(\omega)\,du \equiv \int_0^s f\big(\omega_s(u)\big)\,du + \int_0^p f\big(\omega_p(u)\big)\,du$ for an integral over all the covariates in the sample, so that $e(\tau; s, p) = \underline{E} + \int_C \psi\,du$.

Step 1: the value of precision. At fixed scope, write $F(\tau) \equiv v(\tau, s) = \int_0^s \omega\ell\,du$. Since $\partial(\omega\ell)/\partial\tau = \omega^2(1-\ell)^2 = \psi^2$ and $\partial\psi^2/\partial\tau = -2\psi^3$, differentiating under the integral gives $v_\tau = F'(\tau) = \int_0^s \psi^2 du > 0$, because $\int_0^s \omega\,du > 0$, and $v_{\tau\tau} = F''(\tau) = -2\int_0^s \psi^3 du < 0$, which is Equation (4). Since $\psi = \ell/\tau$, $F'' = -(2/\tau)\int_0^s \psi^2\ell\,du = -(2/\tau)\langle\ell\rangle_s F'$, because ψ^2 is proportional to ℓ^2; and $F' = \Gamma/\tau = E^{\star 2}\Psi'_s/n$ by (15).

Step 2: the marginal value of an observation and the loop. By Lemma 3(i), $\partial\tau^\star/\partial n = M\tau^\star/n = M/E^\star$, so $\partial\mathcal{V}/\partial n = F'(\tau^\star)\,M/E^\star = \alpha_s E^\star/n$, which is Equation (5), and $\partial E^\star/\partial n = E^\star(1-M)/n = -\alpha E^\star/n$. Write $\Psi'' \equiv \partial^2\Psi(E^\star; n, s, p)/\partial E^2$ for the curvature of Ψ at learning noise. By Lemma 2(d), $\Psi'' = -\frac{2}{nE}\int_C \ell^2(1-\ell)\,du < 0$: lower noise speeds up the learning of every weight. By Lemma 2(c), $\Psi' = \frac{1}{E}\int_C \omega\ell(1-\ell)\,du$ at every noise E, and by Lemma 2(a), $\partial\ell/\partial n = \ell(1-\ell)/n$ at fixed noise, so $\partial\Psi'/\partial n = \frac{1}{nE}\int_C \omega\ell(1-\ell)(1-2\ell)\,du = \frac{1}{n^2}\int_C \ell^2(1-2\ell)\,du$: at given noise, each weight contributes its speed, $\omega\ell(1-\ell)/E = \ell^2/n$, times $(1-2\ell)/n$. Since $\partial\alpha/\partial\Psi' = 1/(1-\Psi')^2 = M^2$, the chain rule along $E^\star(n,s,p)$ splits the change in the amplification into a shift of the map at given noise and a move along it, through lower noise:

$$\frac{\partial\alpha}{\partial n} = M^2\Big[\underbrace{\frac{\partial\Psi'}{\partial n}}_{\substack{\text{at given}\\ \text{noise } (\pm)}} + \underbrace{\Psi''\frac{\partial E^\star}{\partial n}}_{\text{lower noise } (+)}\Big]. \tag{16}$$

At $E = E^\star$,

$$\frac{\partial\alpha}{\partial n} = \frac{M^2}{n^2}\Big[2\alpha\int_C \ell^2(1-\ell)\,du + \int_C \ell^2(1-2\ell)\,du\Big] = \frac{M\alpha}{n}\Big[2\alpha\langle 1-\ell\rangle + 1 - 2\langle\ell\rangle\Big],$$

using $\int_C \ell^2 du = n\Psi'$ and $M\Psi' = \alpha$. The bracket equals $2M\langle 1-\ell\rangle - 1$, which is positive if and only if $\langle\ell\rangle < 1 - 1/(2M) = (1+\Psi')/2$. As $n\to\infty$, $\tau^\star\to\infty$ and $\ell\to 1$ wherever $\omega > 0$, so $\langle\ell\rangle\to 1$ by dominated convergence and the bracket tends to -1, while $\alpha = M\Psi'$ and $\Psi' \le (s+p)/n \to 0$.

Step 3: returns to scale. Since $\partial(1/E^\star)/\partial n = \alpha/(nE^\star)$, differentiating $\partial\tau^\star/\partial n = M/E^\star$ gives $\partial^2\tau^\star/\partial n^2 = M\alpha/(nE^\star) + (1/E^\star)\,\partial\alpha/\partial n$, and the chain rule splits the curvature of value in scale

into three terms:

$$\frac{\partial^2 \mathcal{V}}{\partial n^2} = \underbrace{v_{\tau\tau}\left(\frac{\partial \tau^\star}{\partial n}\right)^2}_{\substack{\text{value of}\\ \text{precision } (-)}} + \underbrace{v_\tau M \frac{\alpha}{nE^\star}}_{\text{impact } (+)} + \underbrace{v_\tau \frac{1}{E^\star}\frac{\partial \alpha}{\partial n}}_{\text{amplification } (\pm)} . \tag{17}$$

Without the loop, $\alpha = 0$, and the last two terms vanish. By Step 2

$$\frac{\partial^2 \tau^\star}{\partial n^2} = \frac{M\alpha}{nE^\star}\Big[1 + 2\alpha\langle 1-\ell\rangle + 1 - 2\langle\ell\rangle\Big] = \frac{2M^2\alpha\langle 1-\ell\rangle}{nE^\star} = \frac{2\bar{\ell}}{\tau^\star}\left(\frac{\partial \tau^\star}{\partial n}\right)^2 > 0,$$

with $\bar{\ell} = \alpha\langle 1-\ell\rangle < \alpha$ and $\tilde{\ell} = \langle\ell\rangle$. With Step 1, this gives Equation (6), with $\bar{\ell}_s = \langle\ell\rangle_s$. Steps 2 and 3 prove Proposition 1(ii)–(iii), in which $p = 0$, so that $\tilde{\ell} = \bar{\ell}_s$ and $\bar{\ell} = \alpha(1-\bar{\ell}_s)$, and the claims of Sections 4.1 and 5.2 on $\bar{\ell}$. The multiplier rises with scale in small samples and falls in large ones by the continuity of the bracket in Step 2 and its limits, 1 as $n \to 0$, because $\alpha \to 0$ and $\langle\ell\rangle \to 0$ (below), and -1 as $n \to \infty$; and returns to scale increase only if $\bar{\ell}_s < \alpha$ by Equation (6), since $\bar{\ell} < \alpha$. As $n \to 0$, $E^\star \to 1$ because $\Psi(E; n, s, p) \to 1$ for every $E > 0$, so $\tau^\star \to 0$ and $\ell/\tau^\star \to \omega$ uniformly. Hence $F' \to \int_0^s \omega^2 du$, $F'' \to -2\int_0^s \omega^3 du$, $\partial\tau^\star/\partial n \to 1$, $\alpha/n \to Q$, $\langle\ell\rangle \to 0$ and $\partial^2\tau^\star/\partial n^2 \to 2Q$, which gives the limit of $\partial^2\mathcal{V}/\partial n^2$ in Section 5.2.

Step 4: a representation for the proofs below. Multiplying the fixed point $E^\star = \Psi(E^\star; n, s, p)$ by $\tau^\star = n/E^\star$ and using $\tau\omega(1-\ell) = \ell$ gives $n = C(\tau^\star)$, where

$$C(\tau) \equiv \underline{E}\tau + \mathrm{df}(\tau), \qquad \mathrm{df}(\tau) \equiv \int_C \ell(\tau\omega)\, du, \tag{18}$$

is the scale at which the firm attains precision τ. The term df counts one degree of freedom for each weight in the sample, of either kind, weighted by the fraction of it that the firm has learned. Since $\partial\ell/\partial\tau = \omega(1-\ell)^2$, $C'(\tau) = \underline{E} + \int_C \omega(1-\ell)^2 du \geq \underline{E}$, and differentiating $n = C(\tau^\star)$ gives $\partial\tau^\star/\partial n = 1/C'$, so $\tau^\star C'(\tau^\star) = n/M$ by Step 2. Let

$$\bar{\psi}_F(\tau) \equiv \frac{\int_0^s \psi^2 \cdot \psi\, du}{F'(\tau)}, \qquad \bar{\psi}_C(\tau) \equiv \frac{\underline{E}\cdot 0 + \int_C \omega(1-\ell)^2 \cdot \psi\, du}{C'(\tau)}.$$

The first averages the unlearned signal of the weights that the firm applies, each weighted by its term in F'. The second averages it over the weights in the sample, each weighted by its term in C', and over the signal outside the sample, weighted by $\underline{E}$, at $\psi = 0$. Since $\tau\psi = \ell$, $\tau\bar{\psi}_F = \langle\ell\rangle_s = \bar{\ell}_s$. Multiplying the numerator and the denominator of $\tau\bar{\psi}_C$ by τ and using $\tau\omega(1-\ell) = \ell$,

$$\tau\,\bar{\psi}_C = \frac{\int_C \ell^2(1-\ell)\, du}{\underline{E}\tau + \int_C \ell(1-\ell)\, du} = \frac{n\Psi'\langle 1-\ell\rangle}{n/M} = \alpha\,\langle 1-\ell\rangle = \bar{\ell},$$

because $\tau C'(\tau) = n/M$, $n\Psi' = \int_C \ell^2 du$ and $M\Psi' = \alpha$. By Step 3, $\partial^2\mathcal{V}/\partial n^2$ has the sign of $D \equiv \bar{\psi}_C - \bar{\psi}_F$.

Proof of Proposition 1(i). At $p = 0$, $\langle \ell \rangle = \langle \ell \rangle_s = \bar{\ell}_s$, so by Step 3 and $\alpha = M\Psi'$, $\bar{\ell} - \bar{\ell}_s = \alpha(1 - \bar{\ell}_s) - \bar{\ell}_s = M(\Psi' - \bar{\ell}_s)$, which is the second equality in Equation (6). By Equation (18), $n = \underline{E}\tau^\star + \int_0^s \ell \, du$, and $n\Psi' = \int_0^s \ell^2 du$ (Lemma 2(c)), so $\Psi' < \bar{\ell}_s$ if and only if

$$\Big(\int_0^s \ell^2 du \Big)^2 < \Big(\underline{E}\tau^\star + \int_0^s \ell \, du \Big) \int_0^s \ell^3 du.$$

By the Cauchy–Schwarz inequality, $\big(\int_0^s \ell^2 du \big)^2 \leq \int_0^s \ell \, du \int_0^s \ell^3 du$, and $\underline{E}\tau^\star \int_0^s \ell^3 du > 0$ because $\underline{E} > 0$, $\tau^\star > 0$ and the profile is positive. Hence $\partial^2 \mathcal{V} / \partial n^2 < 0$ at every n. Lemma 4(i) below reaches the same conclusion by another route. □

Lemma 4 (Scale in small and large samples)**.**

(i) Returns to scale diminish at every n if every covariate in the sample carries at least as much signal as every post-prediction covariate in the sample, in particular if $p = 0$.

(ii) The critical mass is positive, $\kappa(s) > 0$. If Inequality (9) *holds, there is $\hat{n} > 0$ such that returns to scale increase at every $n < \hat{n}$; if the reverse strict inequality holds, there is $\hat{n} > 0$ such that they diminish at every $n < \hat{n}$.*

(iii) There is $n^\dagger > 0$ such that returns to scale diminish at every $n > n^\dagger$.

Proof. (i). The average $\bar{\psi}_C$ runs over three groups: the weights that the firm applies, the post-prediction weights and the signal outside the sample. On the weights that the firm applies, $\bar{\psi}_C$ weights ψ by $\omega(1-\ell)^2$ and $\bar{\psi}_F$ by $\psi^2 = \omega \cdot \omega(1-\ell)^2$; since ψ increases with ω, Chebyshev's integral inequality makes the $\bar{\psi}_C$-weighted average of ψ over these weights at most $\bar{\psi}_F$. On the post-prediction weights, every value of ψ is at most every value on the weights that the firm applies, because ψ increases with ω and these covariates carry at most as much signal, and hence at most $\bar{\psi}_F$; without post-prediction covariates, there are no such weights. The signal outside the sample has positive weight $\underline{E}/C'$ and $\psi = 0 < \bar{\psi}_F$. Hence $\bar{\psi}_C < \bar{\psi}_F$ at every $n > 0$.

(ii). As $n \to 0$, $\tau \to 0$ and $\psi \to \omega$ uniformly. Hence $\bar{\psi}_F \to \int_0^s \omega^3 du / \int_0^s \omega^2 du$ and $\bar{\psi}_C \to \int_C \omega^2 du / \big(\underline{E} + \int_C \omega \, du \big) = \int_0^s \omega_s(u)^2 du + \int_0^p \omega_p(u)^2 du$, so $D \to \int_0^p \omega_p(u)^2 du - \kappa(s)$, which is positive if and only if Inequality (9) holds. Both averages are continuous in n, so if the limits differ, there is $\hat{n} > 0$ such that $\bar{\psi}_C - \bar{\psi}_F$ has the sign of their difference at every $n < \hat{n}$. By the Cauchy–Schwarz inequality, $\big(\int_0^s \omega^2 \big)^2 \leq \int_0^s \omega \int_0^s \omega^3$, so $\int_0^s \omega^3 / \int_0^s \omega^2 \geq \int_0^s \omega^2 / \int_0^s \omega > \int_0^s \omega^2$, because $0 < \int_0^s \omega < 1$; hence $\kappa(s) > 0$. Since $\omega_p > 0$, the map $p \mapsto \int_0^p \omega_p(u)^2 du$ is continuous and strictly increasing from zero, so it crosses $\kappa(s)$ at a unique $\hat{p} \in (0, \bar{p})$ when $\int_0^{\bar{p}} \omega_p(u)^2 du > \kappa(s)$, and Inequality (9) holds if and only if $p > \hat{p}$.

(iii). As $n \to \infty$, $\tau \to \infty$ and $\ell(\tau^\star \omega) \to 1$ wherever $\omega > 0$, so $\langle \ell \rangle_s \to 1$ by dominated convergence, while $\alpha\langle 1 - \ell \rangle = \tau \bar{\psi}_C \leq \int_C \ell(1-\ell) du / (\underline{E}\tau) \leq (s+p)/(4\underline{E}\tau) \to 0$. Hence there is $n^\dagger > 0$ such that $\partial^2 \mathcal{V} / \partial n^2 < 0$ at every $n > n^\dagger$.

□

Proof of Propositions 2, 5 and 10. Let x be scope or post-prediction scope, s or p, and $\omega_x(x)$ and ℓ_x the signal and the fraction learned of the marginal covariate of that kind. By Section 3.3, $\partial E^\star/\partial x = -M\omega_x(x)\ell_x < 0$. The strength of the loop, $\Psi'(E; n, s, p) = \frac{1}{n}\int_C \ell(n\omega/E)^2\,du$, gains the term $\ell_x^2/n = \omega_x(x)\ell_x(1-\ell_x)/E$ (Lemma 2(c)) at the upper limit of the integral over positions of kind x, and $\Psi'_s = \frac{1}{n}\int_0^s \ell^2\,du$ gains ℓ_s^2/n when $x = s$ only. By Lemma 2(a), $\partial\Psi'_s/\partial E = -\frac{2}{nE}\int_0^s \ell^2(1-\ell)\,du < 0$, and $\Psi'' < 0$ (Step 2 of the returns to scale). Since $\partial\alpha/\partial\Psi' = M^2$, the chain rule along $E^\star(n, s, p)$ gives, for $x = s, p$,

$$\frac{\partial\alpha}{\partial x} = M^2\Big[\underbrace{\frac{\omega_x(x)\ell_x(1-\ell_x)}{E^\star}}_{\substack{\text{own}\\ \text{speed (+)}}} + \underbrace{\Psi''\,\frac{\partial E^\star}{\partial x}}_{\text{lower noise (+)}}\Big] > 0, \tag{19}$$

which is the first claim of part (ii) of each proposition, and

$$\frac{\partial\Psi'_s}{\partial p} = \frac{\partial\Psi'_s}{\partial E}\,\frac{\partial E^\star}{\partial p} > 0, \qquad \frac{\partial\Psi'_s}{\partial s} = \frac{\omega_s(s)\ell_s(1-\ell_s)}{E^\star} + \frac{\partial\Psi'_s}{\partial E}\,\frac{\partial E^\star}{\partial s} > 0.$$

Parts (i) are Lemma 3(ii), with $\alpha_s = M\Psi'_s$, which equals α at $p = 0$. The learned signal $\omega_y(y)\ell(n\omega_y(y)/E)$ of the marginal covariate of kind y changes with its own position through its signal, with $\partial(\omega\ell)/\partial\omega = \ell(2-\ell)$ by Lemma 2(b), and with noise, with $\partial(\omega\ell)/\partial E = -\ell^2/n$ by (14). Hence

$$\frac{\partial\big(\omega_p(p)\ell_p\big)}{\partial p} = \underbrace{\ell_p(2-\ell_p)\,\omega'_p(p)}_{\substack{\text{signal of the}\\ \text{next covariate (±)}}} + \underbrace{\frac{\omega_p(p)\ell_p(1-\ell_p)}{E^\star}\Big(-\frac{\partial E^\star}{\partial p}\Big)}_{\substack{\text{learned}\\ \text{better (+)}}}, \tag{20}$$

the same holds with s in place of p, and $\partial\big(\omega_p(p)\ell_p\big)/\partial s = \big(\omega_p(p)\ell_p(1-\ell_p)/E^\star\big)(-\partial E^\star/\partial s) > 0$. The product rule, applied to parts (i), gives the curvatures

$$\frac{\partial^2\mathcal{V}}{\partial p^2} = \underbrace{\frac{\partial\Psi'_s}{\partial p}\,\omega_p(p)\ell_p\,M}_{\substack{\text{value of}\\ \text{lower noise (+)}}} + \underbrace{\Psi'_s\,\frac{\partial\big(\omega_p(p)\ell_p\big)}{\partial p}\,M}_{\text{impact (±)}} + \underbrace{\Psi'_s\,\omega_p(p)\ell_p\,\frac{\partial\alpha}{\partial p}}_{\text{amplification (+)}}, \tag{21}$$

$$\frac{\partial^2\mathcal{V}}{\partial s^2} = \underbrace{(1+\alpha_s)\,\frac{\partial\big(\omega_s(s)\ell_s\big)}{\partial s}}_{\substack{\text{direct effect}\\ \text{and impact (±)}}} + \underbrace{\frac{\partial\Psi'_s}{\partial s}\,\omega_s(s)\ell_s\,M}_{\substack{\text{value of}\\ \text{lower noise (+)}}} + \underbrace{\Psi'_s\,\omega_s(s)\ell_s\,\frac{\partial\alpha}{\partial s}}_{\text{amplification (+)}}, \tag{22}$$

and the cross effect of the two kinds of scope,

$$\frac{\partial^2\mathcal{V}}{\partial s\,\partial p} = \underbrace{\frac{\partial\Psi'_s}{\partial s}\,\omega_p(p)\ell_p\,M}_{\substack{\text{value of}\\ \text{lower noise (+)}}} + \underbrace{\Psi'_s\,\frac{\partial\big(\omega_p(p)\ell_p\big)}{\partial s}\,M}_{\text{impact (+)}} + \underbrace{\Psi'_s\,\omega_p(p)\ell_p\,\frac{\partial\alpha}{\partial s}}_{\text{amplification (+)}} > 0, \tag{23}$$

which is Proposition 5. In Equations (21) and (22), every term is positive except the part of the impact term due to the signal of the next covariate, $\Psi'_s M\,\ell_p(2-\ell_p)\,\omega'_p(p)$ and $(1+\alpha_s)\,\ell_s(2-\ell_s)\,\omega'_s(s)$, which proves the second claim of part (ii) of each proposition.

For the cross effects with scale, $v_\tau = \int_0^s \omega^2(1-\ell)^2\,du$ gains $v_{\tau s} = \omega_s(s)^2(1-\ell_s)^2$ at the upper limit when $x = s$ and depends on x otherwise through $\tau^\star$, and $\partial\tau^\star/\partial n = M/E^\star$ depends on x through $E^\star$ and M. The product rule, applied to Equation (5), gives

$$\frac{\partial^2 \mathcal{V}}{\partial n\,\partial x} = \mathbf{1}_{\{x=s\}}\, v_{\tau s}\,\frac{\partial \tau^\star}{\partial n} + \underbrace{v_{\tau\tau}\,\frac{\partial \tau^\star}{\partial x}\,\frac{\partial \tau^\star}{\partial n}}_{\substack{\text{value of}\\ \text{precision } (-)}} + \underbrace{v_\tau M\,\frac{-\partial E^\star/\partial x}{E^{\star 2}}}_{\text{impact } (+)} + \underbrace{v_\tau\,\frac{1}{E^\star}\,\frac{\partial \alpha}{\partial x}}_{\text{amplification } (+)}.$$

Since $v_{\tau\tau} = -(2\bar{\ell}_s/\tau^\star)v_\tau$ and $\partial\tau^\star/\partial x = (\tau^\star/E^\star)(-\partial E^\star/\partial x)$, the term on the value of precision and the term on the impact sum to $Mv_\tau(-\partial E^\star/\partial x)(1-2\bar{\ell}_s)/E^{\star 2}$. For $x = p$, this gives Equation (31); for $x = s$, it gives

$$\frac{\partial^2 \mathcal{V}}{\partial n\,\partial s} = \underbrace{v_{\tau s}\,\frac{\partial \tau^\star}{\partial n}}_{\substack{\text{own weight}\\ \text{applied } (+)}} + \underbrace{v_\tau(1+\alpha)\,\frac{-\partial E^\star/\partial s}{E^{\star 2}}\left(1-2\bar{\ell}_s\right)}_{\substack{\text{impact and}\\ \text{value of precision } (\pm)}} + \underbrace{v_\tau\,\frac{1}{E^\star}\,\frac{\partial \alpha}{\partial s}}_{\text{amplification } (+)} : \tag{24}$$

a covariate adds a weight that the firm learns from every further observation, lower noise raises the impact of an observation and lowers the value of precision, and the covariate raises the amplification. The middle term is nonnegative if $\bar{\ell}_s \le 1/2$, and the other terms are positive, which proves Proposition 2(iii) and Corollary 5(v).

Proportional growth of the factors. The proofs below divide each second derivative by the marginal value, which leaves its sign unchanged. The marginal covariate of that kind has signal $\omega_x(x)$ and fraction learned ℓ_x, and it raises precision at the rate $r_x \equiv \partial \ln \tau^\star/\partial x = M\omega_x(x)\ell_x/E^\star$ (Lemma 3(i)). Inside integrals, ℓ stands for $\ell(\tau^\star\omega)$. Since $\ell(x) = x/(1+x)$, ℓ and $1-\ell$ have elasticities $1-\ell$ and $-\ell$ with respect to $\tau^\star$. The quantity Γ of (15) depends on post-prediction scope only through $\tau^\star$; each term of its integrand has elasticity $1-2\ell$, and the terms are proportional to ℓ^2 (Lemma 2(c)), the weights of $\langle\cdot\rangle_s$. Hence

$$\frac{\partial \ln \Gamma}{\partial p} = r_p\,\langle 1-2\ell\rangle_s, \qquad \frac{\partial \Gamma}{\partial s} = \omega_s(s)\ell_s(1-\ell_s) + r_s\,\Gamma\,\langle 1-2\ell\rangle_s, \tag{25}$$

where the first term of $\partial\Gamma/\partial s$ comes from the upper limit of the integral. Since $M = 1/\big(1-\Psi'(E^\star; n, s, p)\big)$ and the new covariate adds ℓ_x^2/n to the slope at fixed noise,

$$\frac{\partial \ln M}{\partial x} = M\left[\frac{\ell_x^2}{n} + \frac{\partial \Psi'(E^\star; n, s, p)}{\partial E}\,\frac{\partial E^\star}{\partial x}\right] = r_x\,(1-\ell_x) + 2r_x\,\alpha\,\langle 1-\ell\rangle = r_x\left[(1-\ell_x) + 2\bar{\ell}\right], \tag{26}$$

where the first term uses Lemma 2(c), the second uses Lemma 2(d), Lemma 3(i) and $M\Psi' = \alpha$, and the last equality uses $\bar{\ell} = \alpha\langle 1-\ell\rangle$ (Step 3 of the returns to scale). In the same way, the value of lower noise, $\Psi'_s = \frac{1}{n}\int_0^s \ell^2\, du$, has elasticity $2\langle 1-\ell\rangle_s = 2(1-\bar{\ell}_s)$ with respect to $\tau^\star$, and only scope adds a term, $\ell_s^2/n = r_s(1-\ell_s)/M = r_s\Psi'_s(1-\ell_s)/\alpha_s$. Hence

$$\frac{\partial \ln \Psi'_s}{\partial x} = 2r_x(1-\bar{\ell}_s) + \mathbf{1}_{\{x=s\}}\frac{r_s(1-\ell_s)}{\alpha_s}. \tag{27}$$

Finally, the learned signal of the marginal covariate of either kind changes with its own scope through its signal and through precision, and with the other scope through precision only. By Lemma 2(b) and Lemma 3(i),

$$\frac{d\ln(\omega_x(x)\ell_x)}{dx} = (2-\ell_x)\frac{\omega'_x(x)}{\omega_x(x)} + r_x(1-\ell_x), \qquad \frac{\partial \ln(\omega_p(p)\ell_p)}{\partial s} = r_s(1-\ell_p). \tag{28}$$

Dividing Equations (21) and (22) by the marginal values, and using (26), (27) and (28),

$$\frac{\partial^2\mathcal{V}/\partial x^2}{\partial\mathcal{V}/\partial x} = (2-\ell_x)\frac{\omega'_x(x)}{\omega_x(x)} + r_x B_x, \tag{29}$$

where

$$B_p \equiv 2(1-\bar{\ell}_s) + 2(1-\ell_p) + 2\bar{\ell},$$
$$B_s \equiv (1-\ell_s) + \frac{\alpha_s}{1+\alpha_s}\Big[\frac{1-\ell_s}{\alpha_s} + 2(1-\bar{\ell}_s) + (1-\ell_s) + 2\bar{\ell}\Big].$$

If $\omega_p(p) = \omega_s(s)$ and $\omega'_p(p)/\omega_p(p) = \omega'_s(s)/\omega_s(s)$, the two marginal covariates have the same fraction learned ℓ_x and the same rate r_x, and $B_p - B_s = \big[2\bar{\ell} + 2(1-\bar{\ell}_s)\big]/(1+\alpha_s) > 0$; so if $\partial^2\mathcal{V}/\partial s^2 > 0$, the right side of (29) is positive for $x = s$, hence for $x = p$, and $\partial^2\mathcal{V}/\partial p^2 > 0$.

Lemma 5 (Scope in small and large samples)**.**

(i) As $n \to 0$, $r_x \to 0$ and B_x stays bounded for $x = s, p$, and $\partial\mathcal{V}/\partial p \to 0$.

(ii) As $n \to \infty$, r_x stays bounded and $B_x \to 0$, while $\partial\mathcal{V}/\partial p \to 0$ and $\partial\mathcal{V}/\partial s \to \omega_s(s)$.

Proof. Recall that $E^\star \in [\underline{E}, 1]$, with $\underline{E} \equiv \underline{E}(s,p) > 0$, and $\omega \le \bar{\omega}$. The averages $\langle\cdot\rangle$ and $\langle\cdot\rangle_s$ of quantities in $[0,1]$ lie in $[0,1]$, $\bar{\ell} = \alpha\langle 1-\ell\rangle \le \alpha$ and $\alpha_s \le \alpha$, so every term of B_x is nonnegative and $B_x \le 5 + 2\alpha$.

(i). As $n \to 0$, every fraction learned is at most $n\bar{\omega}/\underline{E}$, so $\omega_x(x)\ell_x \le n\bar{\omega}^2/\underline{E}$ and $\Psi'(E^\star; n, s, p) \le n(s+p)\bar{\omega}^2/\underline{E}^2$. Hence $M \to 1$, $\alpha \to 0$, $r_x \to 0$, $B_x \le 5 + 2\alpha$ stays bounded, and $\partial\mathcal{V}/\partial p = \alpha_s\omega_p(p)\ell_p \to 0$.

(ii). As $n \to \infty$, $\tau^\star = n/E^\star \ge n \to \infty$, so $1 - \ell_x = 1/(1+\tau^\star\omega_x(x)) \le 1/(n\omega_x(x)) \to 0$, because $\omega_x(x) > 0$, and $\ell \to 1$ wherever $\omega > 0$, so $1 - \bar{\ell}_s = \langle 1-\ell\rangle_s \to 0$ by dominated convergence, because the profiles are positive. Since $\Psi'(E^\star; n, s, p) \le (s+p)/n$, $M \to 1$, and α, α_s and $\bar{\ell}$ tend to zero. Hence $B_x \to 0$, because the term $(1-\ell_s)/(1+\alpha_s)$ of B_s is at most $1-\ell_s$,

while $r_x = M\omega_x(x)\ell_x/E^\star \le M\bar{\omega}/\underline{E}$ stays bounded. Finally, $\partial\mathcal{V}/\partial p = \alpha_s\omega_p(p)\ell_p \le \alpha\bar{\omega} \to 0$ and $\partial\mathcal{V}/\partial s = \omega_s(s)\ell_s(1+\alpha_s) \to \omega_s(s)$. □

Proof of Corollaries 1(i)–(ii), 5(iii)–(iv) and 6. Let x be p or s. Corollary 6 is the comparison that follows Equation (29). Since $\partial\mathcal{V}/\partial x > 0$, $\partial^2\mathcal{V}/\partial x^2$ has the sign of the right side of Equation (29), in which $r_x B_x > 0$.

Signal that does not decline. If $\omega_x'(x) \ge 0$, the first term is nonnegative, so $\partial^2\mathcal{V}/\partial x^2 > 0$.

Declining signal. If $\omega_x'(x) < 0$, the first term is at most $\omega_x'(x)/\omega_x(x) < 0$, because $2-\ell_x \ge 1$, while $r_x B_x \to 0$ as $n \to 0$ and as $n \to \infty$ (Lemma 5). Hence $\partial^2\mathcal{V}/\partial x^2 < 0$ at all sufficiently small and all sufficiently large n. □

Any part of the covariates, in the limit. For the next results, the sample contains no post-prediction covariates. For a measurable $g : [0,\bar{s}] \to [0,1]$, read $g(u)$ as the fraction of the covariates near position u that the firm observes, in its sample and for the target individual, and let

$$\Psi_g(E) \equiv 1 - \int g\,\omega\,\ell\Big(\frac{n\omega}{E}\Big)du, \qquad \Psi_g'(E) = \frac{1}{n}\int g\,\ell\Big(\frac{n\omega}{E}\Big)^2 du,$$

with integrals over $[0,\bar{s}]$, in which ω stands for $\omega_s(u)$. Fix $\bar{g}$ with $\int \bar{g}\,\omega < 1$, let $E_0 \equiv 1 - \int \bar{g}\,\omega > 0$, and consider $g \le \bar{g}$. Since $\omega\,\ell(n\omega/E) = n\omega^2/(n\omega+E)$ is decreasing and convex in E, Ψ_g is nondecreasing and concave, with $\Psi_g(0) = 1 - \int g\,\omega \ge E_0$ and $\Psi_g \le 1$. It therefore has a unique fixed point $E^\star(g) \in [E_0, 1]$, and concavity gives $1 - \Psi_g'(E^\star(g)) \ge \Psi_g(0)/E^\star(g) \ge E_0$. Write

$$M(g) \equiv \frac{1}{1-\Psi_g'\big(E^\star(g)\big)}, \qquad \mu(u;g) \equiv \omega_s(u)\,\ell\Big(\frac{n\omega_s(u)}{E^\star(g)}\Big)\,M(g),$$
$$\mathcal{V}(g) \equiv \int g\,\omega\,\ell\Big(\frac{n\omega}{E^\star(g)}\Big)du = 1 - E^\star(g).$$

For a set $\mathcal{S}$ of positions that is a union of J disjoint intervals, $\mathcal{V}(\mathbf{1}_\mathcal{S})$ is the limit of the value of observing, in the sample and for the target individual, the covariates whose positions lie in $\mathcal{S}$, those whose cells lie in $\mathcal{S}$. Lemma 1 holds for every set of covariates in the sample, and Step 5 of the proof of Theorem 1 uses the signal only through the L^1 convergence of its step function. Let $\omega_\mathcal{S}$ be the profile ω_s restricted to $\mathcal{S}$ and translated, interval by interval, onto $[0,|\mathcal{S}|]$, so that $\int_0^{|\mathcal{S}|} f(\omega_\mathcal{S})\,du = \int_\mathcal{S} f(\omega)\,du$ for every bounded f. List the covariates in $\mathcal{S}$ interval by interval. On each interval, their step function is that of their order translated by a multiple of $1/K$, and at most $2J$ cells straddle an endpoint of $\mathcal{S}$, so the two translations differ by at most $2J/K$. The L^1 distance between this step function and $\omega_\mathcal{S}$ therefore vanishes as $K \to \infty$, because translation is continuous in L^1. Step 5 applies with $\omega_\mathcal{S}$ in place of ω_s, $|\mathcal{S}|$ in place of s and no post-prediction covariates, and gives the limit $\mathcal{V}(\mathbf{1}_\mathcal{S})$.

Lemma 6 (Combining covariates in the limit)**.** *Let $g, g', \eta : [0,\bar{s}] \to [0,1]$ be measurable with $g \le g'$ and $g' + \eta \le \bar{g}$. Then:*

(i) the value that η adds integrates the marginal values along the path from g to $g+\eta$: $\mathcal{V}(g+\eta)-\mathcal{V}(g)=\int_0^1\int \eta(u)\,\mu(u;g+t\eta)\,du\,dt$;

(ii) more covariates lower learning noise and raise the multiplier, $E^\star(g')\le E^\star(g)$ and $M(g')\ge M(g)$, so $\mu(u;g')\ge\mu(u;g)$ for every u;

(iii) the value that η adds rises with the covariates already held, $\mathcal{V}(g'+\eta)-\mathcal{V}(g')\ge\mathcal{V}(g+\eta)-\mathcal{V}(g)$;

(iv) if, in addition, $\int(g'-g)\,\omega>0$, then $E^\star(g')<E^\star(g)$ and $\mu(u;g')>\mu(u;g)$ for every u with $\omega_s(u)>0$;

(v) there is $L>0$, which depends only on E_0, such that $|\mu(u;g')-\mu(u;g)|\le L\,\omega_s(u)\int|g'-g|\,\omega$ for every u.

Proof. Write $\rho(E;\omega)\equiv\omega\,\ell(n\omega/E)$ for the learned signal at noise E.

(i). The function $(E,t)\mapsto E-\Psi_{g+t\eta}(E)$ is continuously differentiable, with derivative $1-\Psi'_{g+t\eta}(E)\ge E_0$ in E at the fixed point and $\int\eta\,\rho(E;\omega)\,du$ in t. The implicit function theorem gives $dE^\star(g+t\eta)/dt=-M(g+t\eta)\int\eta\,\rho(E^\star(g+t\eta);\omega)\,du$, and $\mathcal{V}=1-E^\star$ gives (i).

(ii). Since $v\ge0$, $\Psi_{g'}\le\Psi_g$, so $\Psi_{g'}(E^\star(g))\le E^\star(g)$. A concave function $\Psi_{g'}$ with $\Psi_{g'}(0)>0$ lies above the diagonal below its fixed point and below it above, so $E^\star(g')\le E^\star(g)$. Since $\ell(n\omega/E)$ decreases in E and $g'\ge g$, $\Psi'_{g'}(E^\star(g'))\ge\Psi'_g(E^\star(g'))\ge\Psi'_g(E^\star(g))$, so $M(g')\ge M(g)$; and v decreases in E.

(iii). Apply (i) to both differences and (ii) to $g'+t\eta\ge g+t\eta$.

(iv). Since $\rho(E;\omega)>0$ whenever $\omega>0$ and $E>0$, $\Psi_{g'}(E)<\Psi_g(E)$ for every $E>0$, so $\Psi_{g'}(E^\star(g))<E^\star(g)$, and the argument of (ii) gives $E^\star(g')<E^\star(g)$. Since $\rho(E;\omega)$ strictly decreases in E wherever $\omega>0$, and $M(g')\ge M(g)$ by (ii), $\mu(u;g')>\mu(u;g)$ wherever $\omega_s(u)>0$.

(v). Write $\epsilon\equiv\int|g'-g|\,\omega$. Since $0\le\rho(E;\omega)\le\omega$, $|\Psi_{g'}(E)-\Psi_g(E)|\le\epsilon$ at every E. The function $\Phi_{g'}(E)\equiv E-\Psi_{g'}(E)$ is convex, at most $-E_0$ at $E=0$ and zero at $E^\star(g')\le1$, so the chord from $(0,\Phi_{g'}(0))$ to $(E^\star(g'),0)$ has slope at least E_0. Below $E^\star(g')$, convexity puts $\Phi_{g'}$ below this chord; above $E^\star(g')$, it puts $\Phi_{g'}$ above its tangent at $E^\star(g')$, whose slope is at least the chord's. Hence $|E-\Psi_{g'}(E)|\ge E_0|E-E^\star(g')|$ for every $E\ge0$, and evaluating at $E=E^\star(g)$ gives $|E^\star(g')-E^\star(g)|\le\epsilon/E_0$. On $[E_0,1]$, $|\partial\rho(E;\omega)/\partial E|\le\omega/(4E_0)$, $\ell(n\omega/E)^2\le n\omega/E_0$ and, by Lemma 2(d), $|\partial\Psi'_g/\partial E|\le2/E_0^2$. Hence $|\Psi'_{g'}(E)-\Psi'_g(E)|\le\epsilon/E_0$, $|\Psi'_{g'}(E^\star(g'))-\Psi'_g(E^\star(g))|\le C\epsilon$ and, since both multipliers are at most $1/E_0$, $|M(g')-M(g)|\le C\epsilon$, with C depending only on E_0. Since $v\le\omega$ and $M\le1/E_0$, the bound on $\mu=v\,M$ follows. □

Groups of post-prediction covariates. Now let the sample contain the first s covariates and the post-prediction covariates whose positions lie in a set $\mathcal{P}$, a union of finitely many disjoint intervals of $[0,\bar{p}]$. The argument above, applied to the post-prediction covariates,

with $\omega_{\mathcal{P}}$ in place of ω_p and $|\mathcal{P}|$ in place of p, shows that value converges to $\mathcal{V}(\mathcal{P})$, the value of Theorem 1 with that post-prediction profile. For a measurable $h : [0, \bar{p}] \to [0, 1]$ with $\int_0^s \omega_s \, du + \int h \, \omega_p \, du < 1$, let

$$\Phi_h(E) \equiv 1 - \int_0^s \omega_s \, \ell\Big(\frac{n\omega_s}{E}\Big) du - \int h \, \omega_p \, \ell\Big(\frac{n\omega_p}{E}\Big) du,$$

with integrals over positions, and let $E^\star(h)$ be its fixed point, $M(h) \equiv 1/\big(1 - \Phi_h'(E^\star(h))\big)$, $\alpha_s(h) \equiv M(h)\,\Psi_s'(E^\star(h))$ with $\Psi_s'(E) \equiv \frac{1}{n}\int_0^s \ell(n\omega_s/E)^2 du$, and $\mathcal{V}(h) \equiv \int_0^s \omega_s \, \ell\big(n\omega_s/E^\star(h)\big) du$, so that $\mathcal{V}(\mathbf{1}_{\mathcal{P}}) = \mathcal{V}(\mathcal{P})$. As for Ψ_g, Φ_h is nondecreasing and concave with $\Phi_h(0) > 0$, so the fixed point is unique and $1 - \Phi_h'(E^\star(h)) > 0$.

Lemma 7 (Post-prediction groups in the limit)**.** *Let $h, h', \eta : [0, \bar{p}] \to [0, 1]$ be measurable, with $h \le h'$ and $\int_0^s \omega_s \, du + \int (h' + \eta) \, \omega_p \, du < 1$.*

(i) The value that η adds accumulates spillovers along the path from h to $h + \eta$: $\mathcal{V}(h + \eta) - \mathcal{V}(h) = \int_0^1 \alpha_s(h + t\eta) \int \eta \, \omega_p \, \ell\big(n\omega_p/E^\star(h + t\eta)\big) \, du \, dt$.

(ii) If $\int (h' - h) \, \omega_p \, du > 0$ and $\int \eta \, \omega_p \, du > 0$, then $\mathcal{V}(h' + \eta) - \mathcal{V}(h') > \mathcal{V}(h + \eta) - \mathcal{V}(h)$.

Proof. (i). As in the proof of Lemma 6(i), the implicit function theorem gives $dE^\star(h+t\eta)/dt = -M(h+t\eta) \int \eta \, \omega_p \, \ell \, du$. Value depends on h only through $E^\star(h)$, and by (14) its derivative with respect to E is $-\Psi_s'(E)$, so $d\mathcal{V}(h + t\eta)/dt = \alpha_s(h + t\eta) \int \eta \, \omega_p \, \ell \, du$.

(ii). Fix $t \in [0, 1]$, and write $k \equiv h + t\eta$ and $k' \equiv h' + t\eta$. The argument of the proof of Lemma 6(ii) and (iv), with Φ in place of Ψ_g, gives $E^\star(k') < E^\star(k)$. The value of lower noise $\Psi_s'(E)$ strictly decreases in E, because $s > 0$ and $\omega_s > 0$. The slope $\Phi_k'(E) = \Psi_s'(E) + \frac{1}{n}\int k \, \ell(n\omega_p/E)^2 du$ decreases in E and rises with k, so $\Phi_{k'}'(E^\star(k')) \ge \Phi_k'(E^\star(k))$ and $M(k') \ge M(k)$. Hence $\alpha_s(k') > \alpha_s(k)$, and $\ell(n\omega_p/E)$ is larger at $E^\star(k')$ wherever $\omega_p > 0$. The integrand of (i) is therefore strictly larger at k' than at k for every t, and (i) gives the claim. □

For the next lemma, let the sample contain the first s covariates and the post-prediction covariates in a group $\mathcal{P}$ that leaves out the signal $\underline{E}_{\mathcal{P}} \equiv 1 - \int_0^s \omega_s - \int_{\mathcal{P}} \omega_p > 0$. Write Q_2, Q_3 and Q_4 for the integrals of ω_s^2, ω_s^3 and ω_s^4 over $[0, s]$, $\hat{Q}_2 \equiv Q_2 + \int_{\mathcal{P}} \omega_p^2$ and $\hat{Q}_3 \equiv Q_3 + \int_{\mathcal{P}} \omega_p^3$ for the same integrals over the whole sample, and $I_s \equiv \int_0^s du/\omega_s(u)$.

Lemma 8 (Groups in small and large samples)**.**

(i) As $n \to 0$, $\mathcal{V}(\mathcal{P}) = nQ_2 + n^2(Q_2\hat{Q}_2 - Q_3) + n^3(2Q_2\hat{Q}_2^2 - Q_2\hat{Q}_3 - 2Q_3\hat{Q}_2 + Q_4) + O(n^4)$.

(ii) As $n \to \infty$, $n\big[\int_0^s \omega_s - \mathcal{V}(\mathcal{P})\big] \to s\,\underline{E}_{\mathcal{P}}$.

(iii) If the signals of the covariates in the sample are bounded away from zero, then, as $n \to \infty$,

$$\mathcal{V}(\mathcal{P}) = \int_0^s \omega_s - \frac{s\,\underline{E}_{\mathcal{P}}}{n} - \frac{s\,(s + |\mathcal{P}|)\,\underline{E}_{\mathcal{P}}}{n^2} + \frac{I_s\,\underline{E}_{\mathcal{P}}^2}{n^2} + O(n^{-3}).$$

The bounds in the O-terms depend only on $\bar{\omega}$, $s + |\mathcal{P}|$, a positive lower bound on $\underline{E}_{\mathcal{P}}$ and, in (iii), the lower bound on signal.

Proof. Write E for the learning noise, which lies in $[\underline{E}_{\mathcal{P}}, 1]$, $x \equiv n\omega/E$ and $\int_C$ for an integral over all the covariates in the sample.

(i). Since $x \leq n\bar{\omega}/\underline{E}_{\mathcal{P}}$, $\ell(x) = x - x^2 + x^3 + O(x^4)$ uniformly. Hence $1 - E = \int_C \omega\,\ell = n\hat{Q}_2/E - n^2\hat{Q}_3/E^2 + O(n^3)$, so $1 - E = O(n)$, $1/E = 1 + (1-E) + (1-E)^2 + O(n^3)$, and substituting twice gives $1/E = 1 + nQ + n^2(2Q^2 - C) + O(n^3)$. Substituting this into $\mathcal{V} = nQ_2/E - n^2Q_3/E^2 + n^3Q_4/E^3 + O(n^4)$ gives (i).

(ii). Since $\omega(1 - \ell(x)) = \omega E/(E + n\omega) \leq E/n$, $E - \underline{E}_{\mathcal{P}} = \int_C \omega(1-\ell) \leq (s + |\mathcal{P}|)/n \to 0$. On $[0, s]$, $n\omega(1-\ell) = n\omega E/(E+n\omega)$ is at most one and tends to $\underline{E}_{\mathcal{P}}$ wherever $\omega > 0$, so dominated convergence gives $n\int_0^s \omega(1-\ell) \to s\,\underline{E}_{\mathcal{P}}$, which is (ii).

(iii). Let $\omega_{\min} > 0$ bound from below the signals in the sample. Since $x \geq n\omega_{\min}$, $1 - \ell(x) = 1/(1+x) = 1/x - 1/x^2 + O(x^{-3})$, so $\omega(1-\ell) = E/n - E^2/(n^2\omega) + O(n^{-3})$ uniformly in the sample. Hence $E = \underline{E}_{\mathcal{P}} + (s + |\mathcal{P}|)E/n + O(n^{-2})$, so $E = \underline{E}_{\mathcal{P}}\big(1 + (s + |\mathcal{P}|)/n\big) + O(n^{-2})$, and

$$\int_0^s \omega_s - \mathcal{V}(\mathcal{P}) = \int_0^s \omega(1-\ell) = \frac{sE}{n} - \frac{I_s E^2}{n^2} + O(n^{-3}) = \frac{s\,\underline{E}_{\mathcal{P}}}{n} + \frac{s\,(s + |\mathcal{P}|)\,\underline{E}_{\mathcal{P}}}{n^2} - \frac{I_s\,\underline{E}_{\mathcal{P}}^2}{n^2} + O(n^{-3}). \quad \square$$

Proof of Proposition 3. (i). By Lemma 1, with $\mathcal{S}$ as the group of covariates in the sample and of covariates observed for the target individual, $V_N(\mathcal{S}) = \frac{1}{K}\sum_{k\in\mathcal{S}} \omega_k\,\mathbb{V}\big(m_k(N, \mathcal{S})\big)$, where $m_k(N, \mathcal{S})$ is the posterior mean of covariate k's weight when the firm observes $\mathcal{S}$ for N individuals. With $\mathcal{S}_1 \cup \mathcal{S}_2$, the firm observes everything it observes with $\mathcal{S}_i$ alone, so the monotonicity argument in the proof of Lemma 1 gives $\mathbb{V}\big(m_k(N, \mathcal{S}_1 \cup \mathcal{S}_2)\big) \geq \mathbb{V}\big(m_k(N, \mathcal{S}_i)\big)$ for $k \in \mathcal{S}_i$, $i = 1, 2$. Splitting the sum over $\mathcal{S}_1 \cup \mathcal{S}_2$ into its parts over $\mathcal{S}_1$ and over $\mathcal{S}_2$ gives $V_N(\mathcal{S}_1 \cup \mathcal{S}_2) \geq V_N(\mathcal{S}_1) + V_N(\mathcal{S}_2)$, which is the claim.

(ii). Let $\mathcal{S}_1$ grow to $\mathcal{S}_1'$, disjoint from $\mathcal{S}_2$, with $\int_{\mathcal{S}_1' \setminus \mathcal{S}_1} \omega > 0$ and $\int_{\mathcal{S}_1' \cup \mathcal{S}_2} \omega < 1$, and write $g \equiv \mathbf{1}_{\mathcal{S}_1}$, $g' \equiv \mathbf{1}_{\mathcal{S}_1'}$, $\eta \equiv \mathbf{1}_{\mathcal{S}_2}$ and $\bar{g} \equiv g' + \eta$. By Lemma 6(i), the value that $\mathcal{S}_2$ adds rises by

$$\int_0^1 \int \eta(u)\,\big[\mu(u; g' + t\eta) - \mu(u; g + t\eta)\big]\,du\,dt,$$

and by Lemma 6(iv), applied to $g + t\eta \leq g' + t\eta$, the integrand is positive wherever $\eta(u)\,\omega_s(u) > 0$. Since $\int_{\mathcal{S}_2} \omega > 0$, the rise is positive. With $\mathcal{S}_1 = \varnothing$ growing to $\mathcal{S}_1$, the value that $\mathcal{S}_2$ adds to $\mathcal{S}_1$ therefore exceeds $\mathcal{V}(\mathcal{S}_2)$, so $G > 0$. Since $G = \big[\mathcal{V}(\mathcal{S}_1 \cup \mathcal{S}_2) - \mathcal{V}(\mathcal{S}_1)\big] - \mathcal{V}(\mathcal{S}_2)$ and $\mathcal{V}(\mathcal{S}_2)$ does not depend on $\mathcal{S}_1$, G strictly increases when $\mathcal{S}_1$ grows. The case in which $\mathcal{S}_2$ grows is symmetric.

(iii) and (iv). The limit $\mathcal{V}(\mathcal{S})$ is the value of a sample with profile $\omega_{\mathcal{S}}$ on $[0, |\mathcal{S}|]$ and no post-prediction covariates, so Lemma 8 applies with $\omega_{\mathcal{S}}$ in place of ω_s, $|\mathcal{S}|$ in place of s and $\mathcal{P} = \varnothing$; since $\int_{\mathcal{S}_1 \cup \mathcal{S}_2} \omega \leq \int_0^s \omega < 1$, its bounds hold uniformly over the three groups. Size, cumulative signal and the integrals $\int_{\mathcal{S}} \omega^k$ are additive over disjoint groups. Write $a_i \equiv |\mathcal{S}_i|$,

$b_i \equiv \int_{S_i} \omega$ and $q_i \equiv \int_{S_i} \omega^2$. By Lemma 8(i), $\mathcal{V}(S) = nq + n^2(q^2 - c) + O(n^3)$ for a group with squared signal q and $c \equiv \int_S \omega^3$, so the terms in nq and $n^2 c$ cancel in G, and $G = n^2\big[(q_1 + q_2)^2 - q_1^2 - q_2^2\big] + O(n^3) = 2n^2 q_1 q_2 + O(n^3)$, which is (iii). By Lemma 8(ii), $n\big[\int_S \omega - \mathcal{V}(S)\big] \to a(1-b)$ for a group of size a and cumulative signal b. Cumulative signal cancels in G, so $nG \to -\big[(a_1+a_2)(1-b_1-b_2) - a_1(1-b_1) - a_2(1-b_2)\big] = a_1 b_2 + a_2 b_1$, which is (iv). □

C Proofs for Section 5

Proposition 5 is Equation (23) in Appendix B.

Proof of Proposition 4. Parts (i) and (ii) are Lemma 4(ii) and (iii). For part (iii): Precision $\tau^\star = n/E^\star$ strictly increases with n (Lemma 3(i)), from zero as $n \to 0$ to infinity as $n \to \infty$, so it suffices to study the sign of $D(\tau) \equiv \bar\psi_C(\tau) - \bar\psi_F(\tau)$ on $[0, \infty)$. With $m_k \equiv \int_0^s \psi^k \, du$ and $c_k \equiv \int_C \psi^k/\omega \, du$, which are finite because $\psi \le \omega \le \bar\omega$ and the profiles are positive, $\bar\psi_F = m_3/m_2$ and $\bar\psi_C = c_3/(\underline{E} + c_2)$. Since $\partial\psi/\partial\tau = -\psi^2$, $dm_k/d\tau = -k\, m_{k+1}$ and $dc_k/d\tau = -k\, c_{k+1}$, so D is continuously differentiable, with

$$\bar\psi_F' = 2\bar\psi_F^2 - 3\,\frac{m_4}{m_2}, \qquad \bar\psi_C' = 2\bar\psi_C^2 - 3\,\frac{c_4}{\underline{E} + c_2}.$$

By the proof of Lemma 4(ii), $D(0) = \int_0^p \omega_p(u)^2 du - \kappa(s)$, and by that of Lemma 4(iii), $D(\tau) < 0$ at every large τ.

Step 1: $D' < 0$ at every zero of D. Let μ be the distribution of ψ that weights each weight that the firm applies by ψ^2/m_2, and ν the distribution of ψ that puts mass $\underline{E}/(\underline{E} + c_2)$ at $\psi = 0$ and weights each weight in the sample, of either kind, by $(\psi^2/\omega)/(\underline{E} + c_2)$. The means of ψ under μ and ν are $\bar\psi_F$ and $\bar\psi_C$, and the means of ψ^2 are m_4/m_2 and $c_4/(\underline{E}+c_2)$. At a zero of D, the two means of ψ coincide and $D' = 3\big[m_4/m_2 - c_4/(\underline{E}+c_2)\big]$, so it suffices to show that the mean of ψ^2 is strictly larger under ν. Let $h \equiv \nu - \mu$, a signed measure with $\int dh = 0$ and $\int \psi \, dh = 0$. On the weights that the firm applies, ν has density $\hat\omega/\omega$ relative to μ, with $\hat\omega \equiv m_2/(\underline{E} + c_2)$, so $h \ge 0$ on the weights with $\omega \le \hat\omega$ and $h \le 0$ on those with $\omega \ge \hat\omega$; ν also puts positive mass at $\psi = 0$ and on the post-prediction weights, where μ puts none. Since $h \ne 0$ and $\int dh = 0$, h is negative somewhere, so $\hat\omega$ lies below the largest signal $\omega^+ \equiv \operatorname{ess\,sup}_{u \le s} \omega_s(u)$ of a covariate that the firm applies. Let a and b be the unlearned signals of weights of signal $\hat\omega$ and ω^+, so that $0 < a < b$. By the ranking, every post-prediction weight has $\psi \ge b$. Hence $h \ge 0$ on $[0, a)$ and on (b, ∞), and $h \le 0$ on $[a, b)$. Since $\int dh = \int \psi \, dh = 0$,

$$\int \psi^2 \, dh = \int (\psi - a)(\psi - b) \, dh.$$

The integrand is positive outside $[a, b]$, where $h \ge 0$, nonpositive on $[a, b)$, where $h \le 0$, and zero at b, so every part of the integral is nonnegative, and the mass of ν at $\psi = 0$ contributes $ab\, \underline{E}/(\underline{E} + c_2) > 0$. Hence $D' < 0$ at every zero of D.

Step 2: the two cases. By Step 1, D crosses zero only from above: if $D(\tau_0) \le 0$, then $D < 0$ just after τ_0, by Step 1 if $D(\tau_0) = 0$ and by continuity otherwise, and a first return of D to zero at some $\tau_1 > \tau_0$ would require $D'(\tau_1) \ge 0$, against Step 1; so $D < 0$ on (τ_0, ∞). First, if $\int_0^p \omega_p(u)^2 du \le \kappa(s)$, then $D(0) \le 0$, so $D < 0$ at every $\tau > 0$, and returns to scale diminish at every n. Second, if $\int_0^p \omega_p(u)^2 du > \kappa(s)$, then $D(0) > 0$ and $D < 0$ at large τ, so D has a smallest zero $\hat{\tau}$, $D > 0$ on $[0, \hat{\tau})$ and $D < 0$ on $(\hat{\tau}, \infty)$; set $\hat{n} \equiv C(\hat{\tau})$. □

D Proofs for Section 6

Two facts recur; ω stands for $\omega_s(u)$ in integrals, and $p = 0$ unless stated otherwise. First, $\mathcal{V}(\cdot, s)$ is continuously differentiable and strictly concave on $(0, \infty)$ (Proposition 1(i)), and $\mathcal{V}(v, s) \to 0$ as $v \to 0$, because $\omega\, \ell(v\omega/E) \le v\omega^2/\underline{E}(s)$. For $x \in [0, n]$, let $f(x) \equiv \mathcal{V}(n, s) - \mathcal{V}(n - x, s)$, the value lost when a holding of scale x leaves. Then $f(0) = 0$ and f is strictly convex, so $f(x + y) > f(x) + f(y)$ for $x, y > 0$ with $x + y \le n$. Second, in the economy of covariates, $W(\mathcal{H}) = \mathcal{V}(\mathbf{1}_{\mathcal{S}_\mathcal{H}})$ in the notation of Lemma 6, with $\mathcal{S}_\mathcal{H} \equiv \bigcup_{h \in \mathcal{H}} \mathcal{S}_h$, $\bar{g} \equiv \mathbf{1}_{[0,s]}$ and $E_0 = \underline{E}(s)$; write $g_{\text{all}} \equiv \mathbf{1}_{[0,s]}$, so that $\mu(u; g_{\text{all}}) = \omega\, \ell(\tau^\star \omega) M$.

Proof of Proposition 6. (i). Since $\Delta_h = f(n_h) = \int_{n-n_h}^{n} \partial\mathcal{V}(v, s)/\partial v \, dv$ and $\partial\mathcal{V}/\partial v$ strictly decreases, $\Delta_h > n_h\, \partial\mathcal{V}(n, s)/\partial n$, and summing gives $\sum_h \Delta_h > n\, \partial\mathcal{V}/\partial n = -\partial E^\star/\partial \ln n = \alpha E^\star$. For the upper bound, order the holders and let $v_h \equiv n_1 + \cdots + n_h$. Then $\mathcal{V}(n, s) = \sum_h \left[\mathcal{V}(v_h, s) - \mathcal{V}(v_{h-1}, s)\right]$, and since $v_{h-1} \le n - n_h$, strict concavity makes each increment at least $f(n_h) = \Delta_h$, strictly for $h = 1$ because $m \ge 2$.

(ii). Splitting holding h into holdings x and $n_h - x$ replaces $f(n_h)$ by $f(x) + f(n_h - x) < f(n_h)$, and leaves every other marginal contribution unchanged, because it leaves the scale of every coalition that contains both parts unchanged.

(iii). By the mean value theorem, $0 < \Delta_h - n_h\, \partial\mathcal{V}(n, s)/\partial n \le n_h\, \varpi(\max_j n_j)$, where ϖ is the modulus of continuity of $\partial\mathcal{V}/\partial n$ on $[n/2, n]$. Summing, the gap is at most $n\, \varpi(\max_j n_j) \to 0$. Since $\mathcal{V} = R^2$ and $E^\star = 1 - R^2$, the share is $\alpha(1 - R^2)/R^2$, and it is below one by (i).

(iv)–(v). Let $p > 0$. The argument of (iii) shows that, for every dataset, the atomistic sum converges to $n\, \partial\mathcal{V}/\partial n$, which equals $\alpha_s E^\star$ (Lemma 3(ii)). By the definition of $\hat{n}$, $\mathcal{V}(\cdot, s, p)$ is strictly convex on $(0, \hat{n})$, and $\mathcal{V}(v, s, p) \to 0$ as $v \to 0$. At $n < \hat{n}$, strict convexity gives (iv): $n\, \partial\mathcal{V}/\partial n > \int_0^n \partial\mathcal{V}(v, s, p)/\partial v \, dv = \mathcal{V}(n, s, p)$. For (v), the value lost when a holding of scale x leaves, $\mathcal{V}(n, s, p) - \mathcal{V}(n - x, s, p)$, is strictly concave in x on $[0, n]$ and zero at $x = 0$, so the merged holders' marginal contribution is below the sum of their separate contributions. □

Proof of Proposition 7. (i). Order the holders and let $\mathcal{S}_{<h} \equiv \bigcup_{j<h} \mathcal{S}_j$. Then $\mathcal{V} = \sum_h \left[\mathcal{V}(\mathcal{S}_{<h} \cup \mathcal{S}_h) - \mathcal{V}(\mathcal{S}_{<h})\right]$, and by Lemma 6(iii), each increment is at most $\mathcal{V}([0, s]) - \mathcal{V}([0, s] \setminus \mathcal{S}_h) = \Delta_h$, strictly for $h = 1$ by Lemma 6(i) and (iv), because $[0, s] \setminus \mathcal{S}_1$ has positive signal when $m \ge 2$.

For the upper bound, Lemma 6(i) with $g = g_{\text{all}} - \mathbf{1}_{S_h}$ and $\eta = \mathbf{1}_{S_h}$ gives

$$\Delta_h = \int_0^1 \int_{S_h} \mu\big(u; g_{\text{all}} - (1-t)\mathbf{1}_{S_h}\big)\, du\, dt < \int_{S_h} \mu(u; g_{\text{all}})\, du, \tag{30}$$

where the inequality follows from Lemma 6(iv), because S_h has positive signal. Summing over holders, $\sum_h \int_{S_h} \omega\, \ell(\tau^\star \omega) M\, du = M\mathcal{V} = (1+\alpha)\mathcal{V}$.

(ii). Let holding h split into S' and S'', both with positive signal, and let $\mathcal{G} \equiv [0,s] \setminus S_h$. The sum of marginal contributions changes by

$$\big[W(\mathcal{G} \cup S' \cup S'') - W(\mathcal{G} \cup S'')\big] - \big[W(\mathcal{G} \cup S') - W(\mathcal{G})\big],$$

the value that S' adds to $\mathcal{G} \cup S''$ minus the value it adds to $\mathcal{G}$, which is positive by Lemma 6(i) and (iv). Every other marginal contribution is unchanged.

(iii). By (30) and Lemma 6(v), $\big|\Delta_h - \int_{S_h} \mu(u; g_{\text{all}})\, du\big| \le L\big(\int_{S_h} \omega\big)^2$, so $\big|\sum_h \Delta_h - (1+\alpha)\mathcal{V}\big| \le L \int_0^s \omega \max_h \int_{S_h} \omega \to 0$. □

Proof of Corollary 2. Every holder of observations faces all covariates, so its marginal contribution is that of the economy of observations; every holder of covariates faces all observations, so its marginal contribution is that of the economy of covariates. By Propositions 6(iii) and 7(iii), the sum converges to $\alpha E^\star + (1+\alpha)\mathcal{V} = \mathcal{V} + \alpha(E^\star + \mathcal{V}) = \mathcal{V} + \alpha$, because $E^\star + \mathcal{V} = 1$; and $\alpha = (\alpha/R^2)\mathcal{V}$. □

Proof of Proposition 8. The firm keeps the share $\Pi / W(\mathcal{H}_{\text{all}}) = 1 - \varphi \sum_h \Delta_h / W(\mathcal{H}_{\text{all}})$ of value, and Propositions 6(iii) and 7(iii) give the shares in (i) and (ii); $\alpha(1-R^2)/R^2 < 1$ gives (i). By Proposition 6(iv), the share in (iii) converges to $1 - \varphi\, \alpha_s E^\star / \mathcal{V} < 1 - \varphi$. With $\varphi = 1$, Proposition 7(i) gives the last claim. □

Proof of Corollary 3. A merger of holders 1 and 2 leaves $W(\mathcal{H}_{\text{all}})$ and the marginal contribution of every other holder unchanged, so the firm's payoff changes by φ times the fall in the merged holders' marginal contribution.

(i). The merged holder's marginal contribution is $f(n_1+n_2) > f(n_1) + f(n_2)$. A single holder of all observations has marginal contribution $\mathcal{V}$, which exceeds the atomistic sum $\alpha E^\star$ by the factor $\mathcal{V}/(\alpha E^\star) = R^2/[\alpha(1-R^2)]$.

(ii). With $\mathcal{G} \equiv [0,s] \setminus (S_1 \cup S_2)$, the merged holder's marginal contribution minus $\Delta_1 + \Delta_2$ is

$$-\Big\{\big[W(\mathcal{G} \cup S_1 \cup S_2) - W(\mathcal{G} \cup S_2)\big] - \big[W(\mathcal{G} \cup S_1) - W(\mathcal{G})\big]\Big\} < 0$$

by Lemma 6(i) and (iv), because S_2 has positive signal. A single holder of all covariates has marginal contribution $\mathcal{V}$, the fraction $1/(1+\alpha)$ of the atomistic sum, so the firm gains $\varphi\alpha\mathcal{V}$. □

Proof of Corollary 4. By Propositions 6(iii) and 7(iii) and Corollary 2, holders receive φ times the shares $\alpha(1-R^2)/R^2$, $1+\alpha$ and $1+\alpha/R^2$ of value. Payments do not exceed value if and only if these products are at most one. The first is below $\varphi \le 1$ by Proposition 6(iii), and the other two give the thresholds. Since $\alpha > 0$ and $R^2 < 1$, $1+\alpha/R^2 > 1+\alpha > 1$, which orders the thresholds. □

Proof of Proposition 9. *(i).* The payments sum to $\lambda \int_0^s \omega\, \ell(\tau^\star\omega)\, du = \lambda\mathcal{V}$.

(ii). A coalition that does not contain the firm is worth zero, and payments are nonnegative. For a coalition of the firm with the holders in $\mathcal{H}$, the allocation gives it $\mathcal{V} - \lambda \int_{[0,s]\setminus S_{\mathcal{H}}} \omega\, \ell(\tau^\star\omega)\, du$, which must be at least $W(\mathcal{H})$. By Lemma 6(i), with $g = \mathbf{1}_{S_{\mathcal{H}}}$ and $\eta = g_{\text{all}} - g$, $\mathcal{V} - W(\mathcal{H}) = \int_0^1 \int \eta\, \mu(u; g+t\eta)\, du\, dt$. By Lemma 6(ii), $E^\star(g+t\eta) \ge E^\star(g_{\text{all}})$, so $\omega\, \ell\big(n\omega/E^\star(g+t\eta)\big) \le \omega\, \ell(\tau^\star\omega)$, and $M(g+t\eta) \ge 1$. Hence $\mathcal{V} - W(\mathcal{H}) \ge \int_{[0,s]\setminus S_{\mathcal{H}}} \omega\, \ell(\tau^\star\omega)\, du \ge \lambda \int_{[0,s]\setminus S_{\mathcal{H}}} \omega\, \ell(\tau^\star\omega)\, du$.

(iii). Since $\int_{S_h} \omega\, \ell(\tau^\star\omega)\, du = \int_{S_h} \mu(u; g_{\text{all}})\, du/(1+\alpha)$, the sum is at most $\lambda/(1+\alpha)$ times the bound in the proof of Proposition 7(iii), which vanishes. □

The convex game. The footnote in Section 6.2 uses the game among the firm and the holders of covariates, in which a coalition C is worth $W(C \setminus \{\text{firm}\})$ if it contains the firm and zero otherwise. The game is convex: for $C \subseteq C'$ and a player $i \notin C'$, if i is a holder and C contains the firm, the value that i adds rises from C to C' by Lemma 6(iii); if C does not contain the firm, i adds nothing to C and a nonnegative amount to C'; and if i is the firm, it adds $W(C) \le W(C')$. The Shapley value of a convex game lies in its core (Shapley 1971).

E Further Results on Post-Prediction Covariates

This appendix collects the results on post-prediction covariates that Section 5 uses only in passing: the returns to post-prediction scope and their interaction with scale, and economies of scope among groups of post-prediction covariates. The assumptions of Section 3.3 hold, and the post-prediction profile is positive and continuously differentiable near p whenever I differentiate with respect to p.

Returns to post-prediction scope. A post-prediction covariate is worth a number of observations that the firm can compute from its validation error; it is worth most at intermediate scale; and its returns increase unless signal declines along the order, in which case they diminish in small and in large samples.

Proposition 10 (Post-prediction scope)**.**

(i) The marginal value of a post-prediction covariate is all indirect, as in Equation (8)*.*

(ii) One more post-prediction covariate raises the amplification, $\partial\alpha/\partial p > 0$, and the curvature of value in post-prediction scope exceeds the change in the marginal value due to the signal of the next covariate:

$$\frac{\partial^2 \mathcal{V}}{\partial p^2} > \Psi'_s\,(1+\alpha)\,\ell_p(2-\ell_p)\,\omega'_p(p).$$

(iii) The cross effect of scale and post-prediction scope is

$$\frac{\partial^2 \mathcal{V}}{\partial n\,\partial p} = \underbrace{v_\tau(1+\alpha)\,\frac{-\partial E^\star/\partial p}{E^{\star 2}}\left(1-2\bar{\ell}_s\right)}_{\substack{\text{impact and}\\ \text{value of precision } (\pm)}} + \underbrace{v_\tau\,\frac{1}{E^\star}\,\frac{\partial \alpha}{\partial p}}_{\text{amplification } (+)}. \tag{31}$$

The learned signal of the new covariate leaves the noise, the multiplier amplifies this impact, and each unit of lower noise is worth Ψ'_s on the weights that the firm applies (part (i)). As post-prediction scope rises, the new covariate adds its learning speed to the loop, and the lower noise speeds up the learning of every weight; both raise the amplification (part (ii)). The lower noise also raises the value of lower noise and lets the firm learn the weight of the next covariate better. Along an order of declining signal, however, the next covariate carries less signal, and this is the only force that can make returns to post-prediction scope diminish (Equations (20) and (21) in Appendix B). One more post-prediction covariate also changes the marginal value of an observation (part (iii)). The lower noise raises the impact of an observation, because precision is scale per unit of noise, and lowers the value of precision, because value is concave in precision; the first effect is at least the second whenever $\bar{\ell}_s \leq 1/2$. The covariate also raises the amplification.

Corollary 5 (Post-prediction covariates)**.**

(i) One more post-prediction covariate is worth as many observations as the sample size times the rise in validation error that deleting its term from the model, without refitting, causes, divided by validation error.

(ii) The marginal value of post-prediction scope tends to zero as $n \to 0$ and as $n \to \infty$, whereas that of scope tends to $\omega_s(s)$ as $n \to \infty$.

(iii) If signal does not decline at p, $\omega'_p(p) \geq 0$, returns to post-prediction scope increase at p.

(iv) If signal declines at p, $\omega'_p(p) < 0$, there are $0 < \underline{n} \leq \bar{n}$ such that returns to post-prediction scope diminish at p at every $n < \underline{n}$ and every $n > \bar{n}$.

(v) Scale and post-prediction scope are complements whenever $\bar{\ell}_s \leq 1/2$.

Part (i) gives the exchange rate between post-prediction covariates and observations. The value of precision and the multiplier are common to both and cancel, so the rate is the ratio of the impacts: along an isoquant of value at fixed scope, $dn/dp = -n\,\omega_p(p)\ell_p/E^\star$. In the

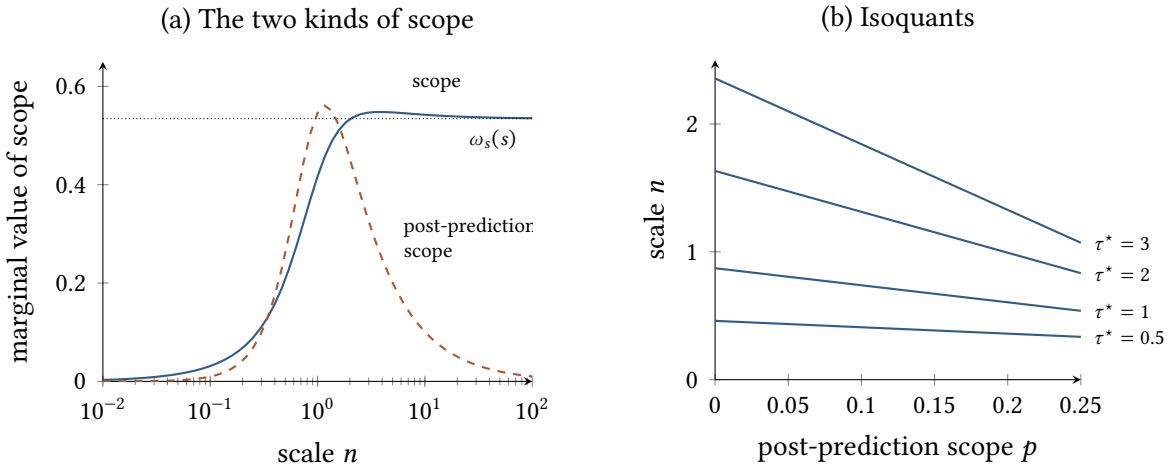


Figure 8. Post-prediction covariates. Example of Figure 5, $s = 0.5$. (a) At $p = \bar{p}$, the marginal values of scope and post-prediction scope as functions of scale; dotted, the signal of the marginal covariate. (b) Contours of precision, which are the isoquants of value: combinations of post-prediction scope and scale that yield the same value.

example of Figure 8, at $n = 1$ and $p = 0.125$, one post-prediction covariate is worth 2.3 observations. Along an isoquant, precision is fixed, so the rate varies only with the signal of the marginal post-prediction covariate; with equal signals, as in panel (b), isoquants are straight lines.

Part (ii) states that post-prediction covariates are worth most at intermediate scale (Figure 8(a)). In small samples, the firm has learned little of any weight, so both the impact and the value of lower noise vanish. In large samples, the firm has nearly learned the weights it applies: the value of lower noise vanishes and the multiplier tends to one, while the direct effect of a covariate tends to its full signal. For the same reasons, declining signal wins at both ends of the scale (part (iv)); in between, the rise in the value of lower noise, in the multiplier and in the impact can outweigh it. Part (v) follows from Proposition 10(iii). In large samples, scale and post-prediction scope are substitutes (Online Appendix OA.3.6).

Corollary 6 (Post-prediction scope and scope)**.** *If* $\omega_p(p) = \omega_s(s)$ *and* $\omega_p'(p)/\omega_p(p) = \omega_s'(s)/\omega_s(s)$*, returns to post-prediction scope increase at* p *whenever returns to scope increase at* s*.*

Declining signal weighs more on returns to scope than on returns to post-prediction scope, because it scales the whole marginal value of a covariate, with weight $1 + \alpha_s$, but only the indirect effect of a post-prediction covariate, with weight α_s.

Economies of post-prediction scope. A firm can also acquire post-prediction covariates in groups that are not adjacent in the order, as when a partner service shares what users did next. Two such groups always add more value together than apart. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be disjoint groups of post-prediction covariates, each collecting the post-prediction covariates whose positions lie in a finite union of intervals of $[0, \bar{p}]$ with positive length, also written $\mathcal{P}_i$, such

that $\int_0^s \omega_s + \int_{\mathcal{P}_1 \cup \mathcal{P}_2} \omega_p < 1$. Write $\mathcal{V}(\mathcal{P})$ for the limit of the value when the sample contains the first s covariates and the post-prediction covariates in $\mathcal{P}$ (Appendix B); size, cumulative signal and squared signal are defined as for groups of covariates (Section 4.3), with ω_p in place of ω_s. The two groups exhibit *economies of post-prediction scope* at scale n if together they add more value than the sum of what each adds alone,

$$G_p \equiv \big[\mathcal{V}(\mathcal{P}_1 \cup \mathcal{P}_2) - \mathcal{V}(\varnothing)\big] - \big[\mathcal{V}(\mathcal{P}_1) - \mathcal{V}(\varnothing)\big] - \big[\mathcal{V}(\mathcal{P}_2) - \mathcal{V}(\varnothing)\big] > 0.$$

Proposition 11 (Economies of post-prediction scope)**.**

(i) They exhibit economies of post-prediction scope at every scale, and G_p strictly increases when either group grows, disjointly from the other.

(ii) In small samples, $G_p/n^3 \to 4 \int_0^s \omega_s^2 \int_{\mathcal{P}_1} \omega_p^2 \int_{\mathcal{P}_2} \omega_p^2$.

(iii) In large samples, if the signals of the covariates in the sample are bounded away from zero,

$$n^2 G_p \to s\Big(|\mathcal{P}_1| \int_{\mathcal{P}_2} \omega_p + |\mathcal{P}_2| \int_{\mathcal{P}_1} \omega_p\Big) + 2 \int_{\mathcal{P}_1} \omega_p \int_{\mathcal{P}_2} \omega_p \int_0^s \frac{du}{\omega_s(u)}.$$

A group of post-prediction covariates adds only spillovers, each the product of the three factors of Equation (8). The other group raises every factor, as one more post-prediction covariate does in Proposition 10(ii): it lowers noise, which raises the value of lower noise, the amplification and the impact of every covariate in the first group, and it adds the learning speeds of its own covariates to the strength of the loop, which raises the amplification further. The signal of the next covariate plays no role, because the groups are fixed.

In small samples, the value that $\mathcal{P}_2$ adds is about $n^2 \int_0^s \omega_s^2 \int_{\mathcal{P}_2} \omega_p^2$, and $\mathcal{P}_1$ raises it by about $4n^3 \int_0^s \omega_s^2 \int_{\mathcal{P}_1} \omega_p^2 \int_{\mathcal{P}_2} \omega_p^2$: half through the value of lower noise, a quarter through the amplification and a quarter through the impact (proof below). The gain is therefore of third order in scale and depends on each group only through its squared signal. In large samples, the error on the weights that the firm applies is about $s\,\underline{E}/n$. A correction of order n^{-2} is proportional to the number of weights in the sample times the signal outside the sample, as in least squares, in which the error on each weight rises with the number of regressors. Each group adds its size to the first factor and removes its cumulative signal from the second, so the cumulative signal of one group is worth more when the other group is larger. The second term of part (iii) comes from the convexity of value in noise. Size and cumulative signal are then sufficient statistics. As for covariates, the two statistics suffice at both ends if every covariate in a group has the same signal; in between, the whole distribution of signal within each group matters. Economies of post-prediction scope are an order smaller at both ends than economies of scope (Proposition 3), because post-prediction covariates act only through precision.

Proofs. Proposition 10, Corollary 5(iii)–(v) and Corollary 6 are proved in Appendix B, with the decompositions behind them.

Proof of Corollary 5, parts (i) and (ii). By Lemma 3(ii), $\partial \mathcal{V}/\partial p = \alpha_s\, \omega_p(p)\ell_p > 0$, because $\int_0^s \omega\, du > 0$ and $\omega_p(p) > 0$.

(i). At fixed scope, value depends on scale and post-prediction scope only through $\tau^\star$, and $\Gamma > 0$, so the isoquants of value are the contours of $\tau^\star$. By Lemma 3(i), their slope is $dn/dp = -r_p/(\partial \ln \tau^\star/\partial n) = -r_p n/M = -\omega_p(p)\ell_p\, n/E^\star = -\tau^\star \omega_p(p)\ell_p$. One more post-prediction covariate raises p by $1/K$, so it is worth $\tau^\star \omega_p(p)\ell_p/K$ units of scale, that is, $N\, \omega_p(p)\ell_p/(KE^\star)$ observations. By Lemma 1, applied to all the covariates in the sample as for validation error in Theorem 1(ii), deleting the term of covariate k from the model without refitting lowers the value of observing them by $\omega_k \mathbb{V}(m_{D,k})/K$ and so raises validation error by that amount, which for the marginal post-prediction covariate is $\omega_p(p)\ell_p/K$ up to a term of smaller order by Theorem 1(ii). Since validation error converges to $E^\star$, the covariate is worth the sample size times this rise divided by validation error, in the limit.

(ii). Lemma 5. □

Proof of Proposition 11. Write $a_i \equiv |\mathcal{P}_i|$, $b_i \equiv \int_{\mathcal{P}_i} \omega_p$ and $q_i \equiv \int_{\mathcal{P}_i} \omega_p^2$, which are positive because the profile is positive and each group has positive length.

(i). In the notation of Lemma 7, $G_p = \big[\mathcal{V}(\mathbf{1}_{\mathcal{P}_1} + \mathbf{1}_{\mathcal{P}_2}) - \mathcal{V}(\mathbf{1}_{\mathcal{P}_1})\big] - \big[\mathcal{V}(\mathbf{1}_{\mathcal{P}_2}) - \mathcal{V}(0)\big]$, which is positive by Lemma 7(ii) with $h = 0$, $h' = \mathbf{1}_{\mathcal{P}_1}$ and $\eta = \mathbf{1}_{\mathcal{P}_2}$. If $\mathcal{P}_1$ grows to $\mathcal{P}_1'$, disjoint from $\mathcal{P}_2$, G_p rises by $\big[\mathcal{V}(\mathcal{P}_1' \cup \mathcal{P}_2) - \mathcal{V}(\mathcal{P}_1')\big] - \big[\mathcal{V}(\mathcal{P}_1 \cup \mathcal{P}_2) - \mathcal{V}(\mathcal{P}_1)\big]$, which is positive by Lemma 7(ii) with $h = \mathbf{1}_{\mathcal{P}_1}$ and $h' = \mathbf{1}_{\mathcal{P}_1'}$. Since G_p is symmetric in the two groups, the same holds when $\mathcal{P}_2$ grows.

(ii). In Lemma 8(i), $\hat{Q}_2$ and $\hat{Q}_3$ are affine in the group of post-prediction covariates and additive over disjoint groups, so every term that is affine in $(\hat{Q}_2, \hat{Q}_3)$ cancels in G_p, and only $2n^3 Q_2 \hat{Q}_2^2$ remains: $G_p = 2n^3 Q_2\big[(Q_2 + q_1 + q_2)^2 - (Q_2 + q_1)^2 - (Q_2 + q_2)^2 + Q_2^2\big] + O(n^4) = 4n^3 Q_2 q_1 q_2 + O(n^4)$.

The split of this gain in Section E follows from Lemma 7(i): the value that $\mathcal{P}_2$ adds integrates over $\mathcal{P}_2$ the spillover $M\Psi_s' \omega_p \ell(n\omega_p/E^\star)$ along the path. As $n \to 0$, learning noise is $E^\star = 1 - n\hat{Q}_2 + O(n^2)$, with $\hat{Q}_2$ the squared signal of the sample, so $M = 1 + n\hat{Q}_2 + O(n^2)$, $\Psi_s' = nQ_2/E^{\star 2} + O(n^2)$ and $\omega_p \ell(n\omega_p/E^\star) = n\omega_p^2/E^\star + O(n^2)$, where the last two $O(n^2)$ terms depend on $\mathcal{P}_1$ only at order n^3. Adding $\mathcal{P}_1$ raises $\hat{Q}_2$ by q_1 and lowers $E^\star$ by $nq_1 + O(n^2)$, so it raises Ψ_s' by $2n^2 Q_2 q_1$, M by nq_1 and the learned signal by $n^2\omega_p^2 q_1$, up to terms of higher order. Multiplied by the other two factors at leading order and integrated over $\mathcal{P}_2$, these rises give $2n^3 Q_2 q_1 q_2$ through the value of lower noise, $n^3 Q_2 q_1 q_2$ through the amplification and $n^3 Q_2 q_1 q_2$ through the impact.

(iii). In Lemma 8(iii), $\underline{E}_{\mathcal{P}} = \underline{E}(s, 0) - b$ and $s + |\mathcal{P}| = s + a$ for a group of size a and cumulative

signal b, so the terms of orders one and $1/n$ cancel in G_p, and

$$n^2 G_p \to -s\big[(a_1+a_2)(-b_1-b_2)-a_1(-b_1)-a_2(-b_2)\big]+2I_s b_1 b_2 = s(a_1 b_2+a_2 b_1)+2I_s b_1 b_2. \quad \square$$

References

Acemoğlu, D., A. Makhdoumi, A. Malekian, and A. Özdağlar (2022). "Too Much Data: Prices and Inefficiencies in Data Markets." *American Economic Journal: Microeconomics* 14 (4): 218–56.

Arrieta-Ibarra, I., L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl (2018). "Should We Treat Data as Labor? Moving beyond Free." *AEA Papers and Proceedings* 108: 38–42.

Ausubel, L. M. and P. Milgrom (2006). "The Lovely but Lonely Vickrey Auction." In *Combinatorial Auctions*. Ed. by P. Cramton, Y. Shoham, and R. Steinberg. Cambridge, MA: MIT Press, 17–40.

Bahri, Y., E. Dyer, J. Kaplan, J. Lee, and U. Sharma (2024). "Explaining Neural Scaling Laws." *Proceedings of the National Academy of Sciences* 121 (27): e2311878121.

Bajari, P., V. Chernozhukov, A. Hortaçsu, and J. Suzuki (2019). "The Impact of Big Data on Firm Performance: An Empirical Investigation." *AEA Papers and Proceedings* 109: 33–37.

Baumol, W. J., J. C. Panzar, and R. D. Willig (1982). *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich.

Bergemann, D. and A. Bonatti (2024). "Data, Competition, and Digital Platforms." *American Economic Review* 114 (8): 2553–95.

Bergemann, D., A. Bonatti, and T. Gan (2022). "The Economics of Social Data." *RAND Journal of Economics* 53 (2): 263–96.

Bergemann, D., A. Bonatti, and A. Smolin (2018). "The Design and Price of Information." *American Economic Review* 108 (1): 1–48.

Bordelon, B., A. Atanasov, and C. Pehlevan (2024). "A Dynamical Model of Neural Scaling Laws." *International Conference on Machine Learning*, 4345–82.

Börgers, T., Á. Hernando-Veciana, and D. Krähmer (2013). "When Are Signals Complements or Substitutes?" *Journal of Economic Theory* 148 (1): 165–95.

Buchanan, J. M. and Y. J. Yoon (2000). "Symmetric Tragedies: Commons and Anticommons." *Journal of Law and Economics* 43 (1): 1–13.

Campbell, J., A. Goldfarb, and C. Tucker (2015). "Privacy Regulation and Market Structure." *Journal of Economics & Management Strategy* 24 (1): 47–73.

Carballa-Smichowski, B., N. Duch-Brown, S. Höcük, P. Kumar, B. Martens, J. Mulder, and P. Prüfer (2025). "Economies of Scope in Data Aggregation: Evidence from Health Data." *Information Economics and Policy* 71: 101146.

Chetty, R. (2009). "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics* 1: 451–88.

Choi, J. P., D.-S. Jeon, and B.-C. Kim (2019). "Privacy and Personal Data Collection with Information Externalities." *Journal of Public Economics* 173: 113–24.

Collard-Wexler, A., G. Gowrisankaran, and R. S. Lee (2019). "Nash-in-Nash Bargaining: A Microfoundation for Applied Work." *Journal of Political Economy* 127 (1): 163–95.

Competition and Markets Authority (2020). *Online Platforms and Digital Advertising Market Study, Appendix I: Search Quality and Economies of Scale*. Tech. rep. London: Competition and Markets Authority.

de Cornière, A. and G. Taylor (2025). "Data and Competition: A Simple Framework." *RAND Journal of Economics* 56 (4): 494–510.

Dobriban, E. and S. Wager (2018). "High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification." *Annals of Statistics* 46 (1): 247–79.

European Union (2022). *Digital Markets Act, Regulation (EU) 2022/1925*. Official Journal of the European Union, L 265.

Farboodi, M. and L. Veldkamp (2026). "A Model of the Data Economy." *Review of Economic Studies*. Forthcoming.

Ghorbani, A. and J. Zou (2019). "Data Shapley: Equitable Valuation of Data for Machine Learning." *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2242–51.

Hagiu, A. and J. Wright (2023). "Data-Enabled Learning, Network Effects, and Competitive Advantage." *RAND Journal of Economics* 54 (4): 638–67.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). "Surprises in High-Dimensional Ridgeless Least Squares Interpolation." *Annals of Statistics* 50 (2): 949–86.

Heller, M. A. (1998). "The Tragedy of the Anticommons: Property in the Transition from Marx to Markets." *Harvard Law Review* 111 (3): 621–88.

Hestness, J. et al. (2017). "Deep Learning Scaling is Predictable, Empirically." arXiv:1712.00409.

Horn, H. and A. Wolinsky (1988). "Bilateral Monopolies and Incentives for Merger." *RAND Journal of Economics* 19 (3): 408–19.

Johnson, G. A., S. K. Shriver, and S. G. Goldberg (2023). "Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR." *Management Science* 69 (10): 5695–721.

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

Lee, G. and J. Wright (2025). "Recommender Systems and the Value of User Data." SSRN Working Paper 5690545.

Lemley, M. A. and C. Shapiro (2007). "Patent Holdup and Royalty Stacking." *Texas Law Review* 85 (7): 1991–2049.

Prüfer, J. and C. Schottmüller (2021). "Competing with Big Data." *Journal of Industrial Economics* 69 (4): 967–1008.

Radner, R. and J. E. Stiglitz (1984). "A Nonconcavity in the Value of Information." In *Bayesian Models in Economic Theory*. Ed. by M. Boyer and R. E. Kihlstrom. Amsterdam: North-Holland, 33–52.

Schäfer, M. (2025). "When Should We Expect Non-Decreasing Returns from Data in Prediction Tasks?" arXiv:2503.03602.

Schäfer, M. and G. Sápi (2023). "Complementarities in Learning from Data: Insights from General Search." *Information Economics and Policy* 65: 101063.

Shapley, L. S. (1971). "Cores of Convex Games." *International Journal of Game Theory* 1 (1): 11–26.

Silverstein, J. W. and Z. Bai (1995). "On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices." *Journal of Multivariate Analysis* 54 (2): 175–92.

Sorscher, B., R. Geirhos, S. Shekhar, S. Ganguli, and A. S. Morcos (2022). "Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning." *Advances in Neural Information Processing Systems 35*, 19523–36.

United States v. Google LLC (2025). *Memorandum Opinion on Remedies*. No. 20-cv-3010 (APM) (D.D.C. Sept. 2, 2025).

Vapnik, V. and A. Vashist (2009). "A New Learning Paradigm: Learning Using Privileged Information." *Neural Networks* 22 (5–6): 544–57.

Varian, H. (2019). "Artificial Intelligence, Economics, and Industrial Organization." In *The Economics of Artificial Intelligence: An Agenda*. Ed. by A. Agrawal, J. Gans, and A. Goldfarb. Chicago: University of Chicago Press, 399–419.

Veldkamp, L. and C. Chung (2024). "Data and the Aggregate Economy." *Journal of Economic Literature* 62 (2): 458–84.

Wicksteed, P. H. (1894). *An Essay on the Co-ordination of the Laws of Distribution*. London: Macmillan.