Monopolistic Data Dumping*

Kfir Eliaz and Ran Spiegler†

February 6, 2025

Abstract

A monopolist curates a database of current and historical observations for users who want to learn some parameter. "Nowcasters" ("forecasters") wish to learn its current (long-run) value. The monopolist chooses the size of each data type, facing constant marginal storage cost, and a menu of contracts, consisting of a fee and access level to each data type. The optimal menu offers full access to historical data, but discriminates access to current data: either full to all consumers or full to nowcasters and none to forecasters. Relative to social optimum, there is too much (little) historical (current) data, and sometimes too much total data.

*Financial support from UKRI Frontier Research Grant no. EP/Y033361/1 is gratefully acknowledged. We thank Tova Milo for helpful comments. We also thank Alex Clyde, Yahel Menea and Chet Geppetti for excellent research assistance.

†Eliaz: Tel Aviv University. Spiegler: Tel Aviv University and University College London

1 Introduction

Production of digital data has exploded in recent years. Data has many uses: Consumption of textual and audio-visual content, individual-specific information that facilitates targeted advertising, training predictive AI models, etc. Several commentators have pointed out that the current pace of data production may outstrip our ability to *store* it. To quote Davidson et al. (2023):

> "Although the Big Data revolution has enabled incredible advances in areas such as medicine, commerce, transportation, and science, we are facing an inflection point: The ability to collect data outstrips our ability to effectively use it and will eventually outstrip our ability to store it."

If data storage space is a scarce resource, then its allocation becomes an economic problem. How much data should society keep, and which kinds of data should it dump? How should this decision reflect the preferences of data users? Are there incentive issues that might distort the decision? How would a profit-maximizing owner of databases price and allocate access to the stored data? This paper offers a simple theoretical model that addresses this problem.

A model of data-storage management should articulate its scope by defining two aspects: (1) what the data is used for; and (2) who curates the data and controls its access, and what is their motivation?

Regarding aspect (1), demand for data in our model originates from users' interest in training predictive statistical models. Our data users do not seek information about individuals; rather, they wish to learn parameters of some predictive statistical model. Specifically, we assume that there are two types of data users: "nowcasters" and "forecasters". The former want to learn the current value of a parameter, whereas the latter want to learn its long-run value. A database consists of two random samples from two time periods:

the present and the past. Each data point has both time-specific and idiosyncratic, observation-specific noise components. Thus, all observations from some time period share the same time-specific noise realization, while having independent observation-specific noise realizations. The parameter and the noise terms follow independent Gaussian distributions. Each user type aims to minimize the mean squared error of the prediction he is interested in. This objective function induces a value that each user type attaches to a sample defined by the number of historical and current observations.

This account of user demand broadly fits real-life data usages such as training AI models; consumer research; and macroeconomic, epidemiological or political forecasts. Our distinction between nowcasters and forecasters captures the idea that data users are differentiated in terms of the time or domain specificity of their predictions. For example, a business may be interested in data about consumer behavior for the purposes of designing a campaign to market an existing product or designing a new product; the former use requires short-term prediction, while the latter requires long-term predictions. Likewise, academic researchers demand data for policy-oriented or basic research; the former aims at precise short-term predictions, while the latter aims at learning long-term fundamentals. Finally, an AI language model may be trained to "understand" general texts or texts in a specific professional domain.

As to aspect (2), data in our model is curated by a monopolistic, profit-maximizing firm, which controls users' access to the data. Storing a data point has a constant marginal cost. The firm chooses the total size of its database and its composition, between current and historical data. If the firm were perfectly informed of users' type, it would offer all users full access to the data and charge each user his ex-ante value of the information inherent in a sample of the given size and composition. However, our main model assumes that users' type is their private information. Accordingly, the firm offers a menu of data-access plans. Each plan consists of a fee as well as a level of access to the two parts of the database.

Our main results characterize the optimal menu. We first establish that nowcasters are like "high" types in a standard second-degree price discrimination model: They always have a higher willingness to pay for any sample than forecasters. However, our user typology does *not* satisfy a single-crossing property: The difference between the two types' willingness to pay increases with the size of the current sub-sample, but *decreases* with the size of the historical sub-sample.

Using this characterization, we show that the optimal menu always gives all user types full access to historical data. Nowcasters get full access to current data as well. As to forecasters, when their fraction is below some threshold, the optimal menu pools them with the nowcasters and offers both types the same full-access plan. However, if their fraction is above the threshold, then the optimal menu discriminates between the two types, such that forecasters get *no* access to current data, in return for a lower fee.

Finally, we analyze the distortions of the database size and composition that arise from second-degree price discrimination, relative to the socially efficient allocation. First, the historical sub-sample is too large and the current sub-sample is too small. In other words, the firm dumps too much new data and too little old data. This effect can be so large that we may end up having more historical than current data, whereas the opposite is always true under the social optimum. As to the total size of the database, there is no clear-cut comparison. Using a numerical illustration, we show that somewhat surprisingly, our firm may end up curating a database that is *larger* than its socially optimal level. In other words, one potential consequence of relying on users' incentives to manage data access is that too little data is dumped.

Related literature

Computer scientists have begun addressing the data-storage challenge in the age of big data (e.g., Milo (2019), Davidson et al. (2023)). This literature attempts to devise effective and computationally efficient algorithms for determining which pieces of data to delete. For examples of recent attempts to quantify the cost of training AI models (which is partly a function of training-set size), see Guerra et al. (2023) and Cottier et al. (2024).

Within economic theory and IO, our paper is closest to the growing literature on markets for information (see a review by Bergemann and Bonatti (2019)). The focus of this literature is on the buying and selling of personal data, mainly for the purpose of personalized advertising and price discrimination. By contrast, our focus is on the use of statistical data for the purpose of making general (i.e., not individual-specific) predictions. Our focus on the data-dumping problem is also new to this literature, to our knowledge.

At a high level, our model is an example of monopolistic pricing of excludable public goods (as in Brito and Oakland (1980) and Norman (2004)). What is new is that the public good in our model is statistical data. It has two dimensions (historical and current data), and users' demand for the public good originates from the informational value of statistical data, which generates a structured violation of the single-crossing property. As an example of a two-type monopolistic screening problem without single crossing, our paper is also related to Siegel and Haghpanah (2025).

2 The Model

A monopolistic *firm* designs a dataset and controls its access to *users*. The population of users has measure one. There are two types of users: "*nowcasters*" (denoted S) interested in short-term prediction, and "*forecasters*" (denoted L) interested in long-term prediction. Let $\lambda \in [0,1]$ denote the fraction of type-S users in the population.

Let $\mu \sim N(0, \sigma_\mu^2)$ be a fixed *parameter* of interest. There are two *time periods*, denoted 1 ("the present") and 0 ("the past"). A *database* is described by a pair of non-negative numbers (n_0, n_1), where n_t indicates the size of a *sample* consisting of observations from period t. For analytical convenience, we allow n_t to take any non-negative real value.

Each observation $i = 1, ..., n_t$ from the period-t sample is a realization

$$y_{t,i} = \mu + x_t + \varepsilon_{t,i}$$

where $x_t \sim N(0, 1)$ and $\varepsilon_{t,i} \sim N(0, \sigma_\varepsilon^2)$. The variance of x_t is a normalization that entails no loss of generality. The value of x_t is drawn independently for each period t, but its value is the same for all observations that belong to the period-t sample. The value of $\varepsilon_{t,i}$ is drawn independently for every t, i. Each data point in the database carries a *storage cost* of $c > 0$.

As implied by their description, the two types of users differ in what they try to learn. After learning from whatever sample he gets access to, each type chooses an action $a \in \mathbb{R}$. The two types' payoff functions are:

$$u_S(a, \mu, x_1) = -(a - \mu - x_1)^2$$
$$u_S(a, \mu) = -(a - \mu)^2$$

The interpretation is that $\mu + x_1$ is the true *current* value of a variable of interest. Nowcasters, with their short-term prediction horizon, try to learn this value. In comparison, μ is the variable's true *long-run* value. Forecasters, with their long-term prediction horizon, try to learn this value.

Users are Bayesian expected-utility maximizers. Their willingness to pay for access to a database given by (n_0, n_1) is equal to expected-utility gain that the information in the database generates. Let $V_S(n_0, n_1)$ and $V_L(n_0, n_1)$ denote this willingness to pay for the two types. We will derive exact expressions for these quantities in Section 3.

A perfect monopolist can identify user types, give them access to the database and charge them their willingness to pay. It is clear that users will receive full access, because their willingness to pay is increasing in the amount of information provided. Therefore, the perfect monopolist will choose the database (n_0, n_1) to solve the following maximization problem:

$$\max_{n_0,n_1} \{\lambda V_S(n_0,n_1) + (1-\lambda)V_L(n_0,n_1) - c(n_0+n_1)\} \tag{1}$$

We refer to a solution to this problem as the *first-best solution*, and use it as a benchmark.

The main problem we analyze is based on the assumption that users' type is their *private information*. Consequently, applying the revelation principle, the monopolist offers a menu M of *access plans* $m^k = (q_0^k, q_1^k, p^k)$, where $q_t^k \in [0, n_t]$ represents the amount access that user type k gets to the period-t sample, and $p^k \geq 0$ is the fixed access fee he pays. The usual participation and incentive constraints must hold.

Thus, our monopolist's maximization problem is

$$\max_{n_0,n_1,(q_0^k,q_1^k,p^k)_{k=S,L}} \{\lambda p^S + (1-\lambda)p^L - c(n_0+n_1)\} \tag{2}$$

subject to the constraints

$$\begin{aligned} n_t \geq q_t^k &\geq 0 \\ V_k(q_0^k, q_1^k) - p^k &\geq 0 \\ V_k(q_0^k, q_1^k) - p^k &\geq V_k(q_0^{-k}, q_1^{-k}) - p^{-k} \end{aligned}$$

for every $t = 0, 1$, $k = S, L$ ($-k$ denotes the other user type).

The first constraint means that users get potentially partial access to the database that the monopolist chooses to curate. The second constraint is user type k's participation (IR) constraint, and the third constraint is type k's incentive-compatibility (IC) constraint. We refer to a solution to (2) as the *second-best solution*.

The monopolist in our model chooses the size and composition of a database, as well as how to price user access to the database. We regard the first component as a "data dumping" decision. Our interpretation is that the monopolist controls an extremely large set of data points from both time periods. The data is prohibitively costly to store, and so the monopolist has to decide how much data from each time period to delete.

3 Preliminary Analysis: Value of Data

In this section we derive formulas for users' willingness to pay for data access, and highlight their key properties.

Let θ^k denote user type k's target — i.e., $\theta^S = \mu + x_1$ and $\theta^L = \mu$. Each type's prior belief over his target is Gaussian. Since signals are Gaussian as well, each type's posterior belief also falls into this class. Moreover, since a user's optimal action is to match the mean of his Gaussian belief over his target, it follows that his expected payoff given a Gaussian posterior distribution is the variance of this distribution.

Therefore, a user type's willingness to pay for (n_0, n_1) is equal to the reduction in the variance of his belief over his target (i.e., the difference between the posterior and prior variance). The prior variances over θ^S and θ^L are $\sigma_\mu^2 + 1$ and σ_μ^2, respectively. Let us now calculate the variance of the types' posterior beliefs.

From type L's point of view, a period-t sample generates a conditionally independent signal $\bar{y}_t = \theta^L + x_t + \bar{\varepsilon}_t$, where

$$\bar{\varepsilon}_t = \sum \frac{\varepsilon_{t,i}}{n_t}$$

is the average observational noise in the period-t sample. The variance of the period-t signal conditional on θ^L is $1 + \sigma_\varepsilon^2 / n_t$. Applying the standard Gaussian signal extraction formula to the conditionally independent signals provided by the two periods' samples (a detailed derivation is relegated to a supplementary appendix), the variance of the posterior belief of θ^L is

$$\sigma_\mu^2 - \frac{\sigma_\mu^4\left(2+\frac{\sigma_\varepsilon^2}{n_1}+\frac{\sigma_\varepsilon^2}{n_0}\right)}{\sigma_\mu^2\left(2+\frac{\sigma_\varepsilon^2}{n_1}+\frac{\sigma_\varepsilon^2}{n_0}\right)+\left(1+\frac{\sigma_\varepsilon^2}{n_1}\right)\left(1+\frac{\sigma_\varepsilon^2}{n_0}\right)}$$

Let us now turn to type S. From his point of view, the two periods' samples generate the signals $\bar{y}_1 = \theta^S + \bar{\varepsilon}_1$ and $\bar{y}_0 = \theta^S + x_0 - x_1 + \bar{\varepsilon}_0$, where $\bar{\varepsilon}_0$ is defined as before. Note that unlike the case of type L, the error term in $\bar{y}_0$ is not independent of θ^S because both include x_1. Again, using the signal-extraction formula, the variance of the posterior belief of θ^S is

$$\sigma_\mu^2 + 1 - \frac{(\sigma_\mu^2+1)\left[(\sigma_\mu^2+1)(\sigma_\mu^2+1+\frac{\sigma_\varepsilon^2}{n_0})-\sigma_\mu^4\right]+\sigma_\mu^4\frac{\sigma_\varepsilon^2}{n_1}}{\left(\sigma_\mu^2+1+\frac{\sigma_\varepsilon^2}{n_0}\right)\left(\sigma_\mu^2+1+\frac{\sigma_\varepsilon^2}{n_1}\right)-\sigma_\mu^4}$$

From now on, let us normalize $\sigma_\varepsilon^2 = 1$. Since we have already normalized the variance of x_t, this additional normalization may appear to carry a loss of generality. However, note that we can regard it as a *redefinition* of the unit of measurement of database size: n_t is effectively measured in terms of multiples of σ_ε^2. Using this normalization and simplifying the expressions for the posterior variances of θ^S and θ^L, we obtain the following result.

Remark 1 *The two user types' willingness to pay for* (n_0, n_1) *is*

$$V_L(n_0,n_1) = \frac{\sigma_\mu^4(n_1+n_0+2n_0n_1)}{\sigma_\mu^2(n_1+n_0+2n_0n_1)+(1+n_0)(1+n_1)} \tag{3}$$

$$V_S(n_0,n_1) = \frac{\sigma_\mu^4(n_1+n_0+2n_0n_1)+[3\sigma_\mu^2 n_0n_1+2\sigma_\mu^2 n_1+n_1+n_0n_1]}{\sigma_\mu^2(n_1+n_0+2n_0n_1)+(1+n_0)(1+n_1)} \tag{4}$$

While these formulas may appear unpalatable, they have simple, interpretable properties that will serve us in the sequel. The following result collects these properties.

Remark 2 *The functions V_L and V_S satisfy the following properties:*
(i) V_L and V_S are strictly increasing in both arguments.
(ii) V_L and V_S are strictly concave. In particular, $\partial^2 V_k(n_0, n_1)/\partial n_t$ and $\partial^2 V_k(n_0, n_1)/\partial n_0 \partial n_1$ are strictly negative for every type k and period t.
(iii) V_L is symmetric. In contrast, for every (n_0, n_1), $V_S(x, y) > V_S(y, x)$ if $y > x$, and

$$\frac{\partial V_S(n_0, n_1)}{\partial n_1} > \frac{\partial V_S(n_0, n_1)}{\partial n_0}$$

(iv) $V_S(0, 0) = V_L(0, 0) = 0$, and $V_S(n_0, n_1) > V_L(n_0, n_1)$ for every $(n_0, n_1) \neq (0, 0)$.
(v) For every (n_0, n_1),

$$\frac{\partial V_S(n_0, n_1)}{\partial n_1} > \frac{\partial V_L(n_0, n_1)}{\partial n_1}$$
$$\frac{\partial V_L(n_0, n_1)}{\partial n_0} > \frac{\partial V_S(n_0, n_1)}{\partial n_0}$$

Since the proofs involve no more than elementary investigation of the functions up to their second derivatives, they are relegated to a supplementary appendix. However, the intuition behind the properties is important for the subsequent analysis, and therefore we explain it here.

Parts (i) and (ii) of Remark 2 are simple consequences of V_L and V_S being value-of-information functions. First, they are strictly increasing in sample size because information always has positive marginal value in this environment. Second, the functions are strictly concave because information has diminishing marginal value in this environment: The marginal variance reduction that an additional sample point from any period generates gets smaller as we increase any period's sample size.

Part (*iii*) articulates a difference in how the two types regard sample points from each period. For type L, the two periods are symmetric: If we permute n_0 and n_1, the sample is equally informative for this type. In contrast, for type S, a present sample point is always more informative than a historical sample point, because the latter has another layer of independent noise (given by $x_0 - x_1$) relative to the former. This is unsurprising: A nowcaster, who is trying to learn something about the present, will intuitively prefer a current observation to a historical one.

Part (*iv*) means that type S is a "high" type relative to type L: His willingness to pay for non-null samples is always strictly higher. The reason is that the time-specific component x_1 is part of what type S tries to learn, whereas for type L it is mere additional noise. Therefore, even when the two types get access to the same data, type S regards it as less noisy (hence more informative) than type L. Thus, nowcasters value information more than forecasters.

However, as part (*v*) articulates, this classification of the two types into "high" and "low" does not translate to a standard single-crossing property with respect to the natural partial ordering of pairs (n_0, n_1). On one hand, both V_L and V_S increase in this order (by part (*i*) of the remark). However, while an increase in n_1 leads to an increase in the difference $V_S(n_0, n_1) - V_L(n_0, n_1)$ — as a standard single-crossing property would prescribe — an increase in n_0 leads to a *decrease* in $V_S(n_0, n_1) - V_L(n_0, n_1)$, which goes against the single-crossing property. The intuition is that a current sample point is more informative for type S than for type L, given that the term x_1 is part of what type S tries to learn while it is mere noise for type L. On the other hand, historical observations are more informative for type L, because for him the noise level of such observations is $x_0 + \varepsilon$, whereas for type 1 their noise level is $x_0 - x_1 + \varepsilon$ — i.e., they have an additional noise term.

The fact that nowcasters value statistical data more than forecasters, coupled with the two types' radically different marginal attitude to the two kinds of statistical data, will drive our results in the next section.

4 Main Results

This section characterizes the monopolist's optimal policy, including the size and composition of the database, the level of access offered to each user type, and the structure of access fees.

4.1 First-Best

As a benchmark, let us present the solution to the first-best problem (1). Since V_L and V_S are strictly concave, the optimal database (n_0^*, n_1^*) is uniquely given by first-order conditions:

$$(1-\lambda)\frac{\partial V_L(n_0,n_1)}{\partial n_1} + \lambda \frac{\partial V_S(n_0,n_1)}{\partial n_1} = c \tag{5}$$
$$(1-\lambda)\frac{\partial V_L(n_0,n_1)}{\partial n_0} + \lambda \frac{\partial V_S(n_0,n_1)}{\partial n_0} = c$$

whenever $n_0^*, n_1^* > 0$. Moreover, it is optimal for the firm to offer users full access to the database, and charge each type k his willingness to pay $V_k(n_0^*, n_1^*)$ as an access fee. The following result characterizes the composition of the optimal database.

Remark 3 *The optimal database (n_0^*, n_1^*) satisfies $n_1^* \geq n_0^*$. Moreover, the inequality is strict when $n_0^* > 0$.*

Thus, the first-best database contains more current data points than historical ones.

4.2 Second-Best

In this section we characterize the key features of the second-best solution and compare it to the social optimum.

Proposition 1 *The second-best solution has the following properties:*
(i) $q_0^S = q_0^L = n_0$.
(ii) $q_1^S = n_1$*, and there exists a threshold* $\lambda^* \in (0,1)$ *such that* $q_1^L = n_1$ *if* $\lambda \leq \lambda^*$ *and* $q_1^L = 0$ *otherwise.*
(iii) There exists a cost level $\hat{c}$ *and a threshold* $\hat{\lambda} \in (\lambda^*, 1)$ *such that for all* $c < \hat{c}$ *and* $\lambda \geq \hat{\lambda}$, n_0 *decreases in* λ, n_1 *increases in* λ, *and* $n_1 > n_0$.
(iv) The access fees paid by each user type are

$$p^L = V_L(n_0, q_1^L)$$
$$p^S = V_S(n_0, n_1) - V_S(n_0, q_1^L) + V_L(n_0, q_1^L)$$

This result highlights key features of monopolistic data dumping. First, discrimination exhibits a "bang-bang" property. When the fraction of nowcasters in the user population is below some threshold, there is no discrimination: both types are offered full access to the data for a uniform fee that extracts the forecasters' entire surplus. When the fraction of nowcasters exceeds the threshold, forecasters pay a lower in return for full access to historical data and no access to current data, while nowcasters pay a premium to get full access to all data.

In addition, the amount of data dumping is monotone in the fraction of nowcasters but goes in opposite directions for the two kinds of data. As λ increases, there is less dumping of current data and more dumping of historical data. Furthermore, when λ is high enough, more current data is stored than historical data, as in the first-best. Interestingly, for some parameter values, the *opposite* is true for a range of low λ values (e.g., when $c = 0.3$, $\sigma_\mu = 1.2$ and $\lambda \in (0.25, 0.3)$).

To understand the distortions that arise in the second-best solution, we compare it the social optimum given by the first-best solution. For brevity, we focus on the case in which the latter is interior.

Proposition 2 *Let (n_0^*, n_1^*) and (n_0', n_1') be the first-best and second-best databases, respectively. Suppose $n_t^* > 0$ for both $t = 0, 1$. Then, $n_0' > n_0^*$ and $n_1' < n_1^*$.*

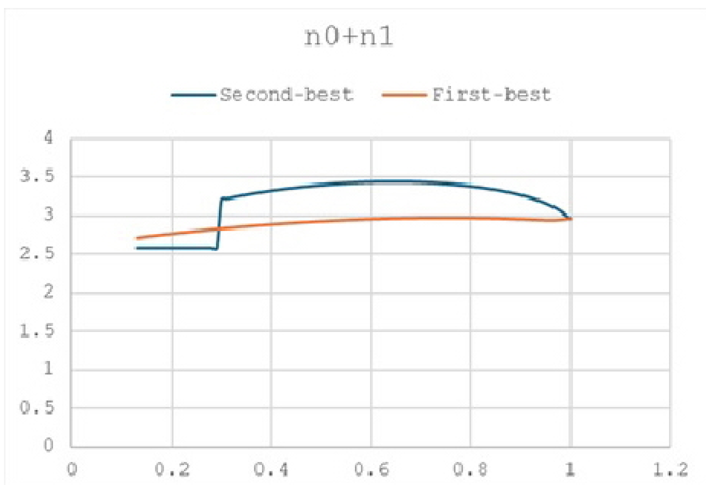
Thus, relative to the social optimum, under a profit-maximizing monopolist there are "over-dumping" of current data and "under-dumping" of historical data.

The proof of Proposition 2 proceeds by considering the two first-order conditions (one with respect to n_1 and another with respect to n_0) of the first-best problem. Each of these conditions can be thought of as an "iso-marginal value" curve that traces (n_0, n_1) pairs which yield a marginal value of c. If we graph these curves in $\mathbb{R}^2_{++}$, where n_0 and n_1 are represented by the horizontal and vertical axes, respectively, they are both downward sloping and intersect only once at (n_0^*, n_1^*).

Suppose that the second-best solution satisfies $q_1^L = 0$. Using the properties described in Remark 2, we first show that to the right of (n_0^*, n_1^*), the "iso-marginal value" curve representing the first-order condition with respect to n_1 ($FOC^{FB}(n_1)$) lies *above* the curve representing $FOC^{FB}(n_0)$. Because the two curves intersect once, this means that to the left of (n_0^*, n_1^*), the curve representing $FOC^{FB}(n_1)$ lies *below* the curve representing $FOC^{FB}(n_0)$. We then show that the properties in Remark 2 imply that the iso-marginal value curves representing the first-order condition with respect to n_1 in the *second*-best problem ($FOC^{SB}(n_1)$) can be thought of as a *downward* shift of the $FOC^{FB}(n_1)$, whereas the iso-marginal value curve representing $FOC^{SB}(n_0)$

can be thought of as an *upward* shift of the $FOC^{FB}(n_0)$ curve. This double curve shifting means that the solution (which continues to be unique, because the second-best objective function with $q_1^L = 0$ is strictly concave) satisfies $n_0' > n_0^*$ and $n_1^* > n_1'$. A separate argument, which also uses Remark 2, shows that the same conclusion holds when $q_1^L = n_1$.

Finally, let us turn to the total second-best database size $n_0^* + n_1^*$. One might expect that the cost of screening user types will lead to an efficiency loss in the form of data under-storage. It turns out that this is not the case: The comparison between the first-best and second-best total database size is not clear-cut. For instance, the following figure shows how $n_0^* + n_1^*$ varies with λ, for the parameter values $c = 0.1$ and $\sigma_\mu^2 = 2$:



This is not an isolated pattern; it arises under several configurations of (c, σ_μ^2).

The reason over-storage of data may arise in the second-best problem is that to compensate for type L's lack of access to current data, the firm inflates the historical database. This increase may be so big that it outweighs the reduction in the size of the current database. Thus, although incentive constraints dissipate the value of available data for some users (specifically, the forecasters who are interested in long-run predictions), the monopolist's reaction to this effect can result in too little data dumping relative to the social optimum.

Comment on the model's temporal interpretation

Throughout the exposition, we have interpreted n_0 and n_1 as "old" and "current" data. If we think of the interaction between the monopolist and users as a one-off event, this interpretation is airtight. However, perhaps a more realistic interpretation would be that the monopolist is a long-run player interacting with a sequence of generations of short-lived users of both types. At every time period t, there is an arbitrarily large inflow of new datapoints, and the monopolist decides how many of them to curate in the "current" database as well as how many of the previously stored datapoints to dump. Datapoints that are more than two-periods old are eliminated automatically.

Under this "Markovian" interpretation, "current" data at a time period t becomes "old" data at time period $t+1$. This means that n_0 can never exceed n_1. While this property holds anyway under the first-best solution, we saw that it can be violated under the second-best solution. Therefore, if we want our model to be consistent with the Markovian interpretation, we should add the constraint $n_0 \leq n_1$.

Of course, as we noted in the Introduction, our model has an alternative, non-temporal interpretation, according to which n_1 and n_0 represent databases that belong to narrow and broad domains, respectively. This interpretation does not pose the problem discussed here.

Appendix I: Proofs

Remark **3**

Suppose $n_0^* > n_1^*$. Suppose the firm deviates to (n_0', n_1') such that $n_0' = n_1^*$ and $n_1' = n_0^*$. By Remark 2(iii), $V_L(n_0', n_1') = V_L(n_0^*, n_1^*)$, whereas $V_S(n_0', n_1') > V_S(n_0^*, n_1^*)$. Obviously, $c(n_0' + n_1') = c(n_0^* + n_1^*)$. Therefore, the deviation increases the value of the objective function given by (1).

Now suppose $n_0^* = n_1^* > 0$. Then, the optimum is given by (5). By Remark 2(iii),

$$\frac{\partial V_L(n_0^*, n_1^*)}{\partial n_1} = \frac{\partial V_L(n_0^*, n_1^*)}{\partial n_0}$$
$$\frac{\partial V_S(n_0^*, n_1^*)}{\partial n_1} > \frac{\partial V_S(n_0^*, n_1^*)}{\partial n_0}$$

contradicting (5). ■

Proposition **1**

The proof proceeds by a series of claims about solutions to the second-best problem. Some of these claims are devoted to establishing which constraints are binding. Since $V_S - V_L$ is not increasing in (q_0, q_1), we cannot invoke standard arguments toward this end.[1]

Claim 1. IC_S binds, whereas IR_S holds with slack whenever $q^L \neq (0, 0)$.

Proof. By IC_S and part (iv) of Remark , we have:

$$V_S(q_0^S, q_1^S) - p^S \geq V_S(q_0^L, q_1^L) - p^L \geq V_L(q_0^L, q_1^L) - p^L \geq 0$$

where the last inequality follows from IR_L. When $q^L = (0, 0)$, IC_S coincides with IR_S. If it does not bind, then the monopolist can slightly raise p^S

[1] We are also unable to apply recent tools introduced by Haghpanah and Siegel (2025), because users' object of consumption q is not uni-dimensional.

without violating any of the constraints (it only relaxes IC_L, and IR_S has slack). When $q^L \neq (0,0)$, the second inequality is strict, which implies that IR_S holds with slack. Here, too, if IC_S does not bind, then the monopolist can slightly raise p^S without violating any of the constraints (it only relaxes IC_L, and IR_S has slack), contradicting optimality. □

Claim 2. IC_L holds with slack when $q^S \neq q^L$.

Proof. Suppose $q^L = (0,0)$. Then, as we saw in the proof of Claim 1, IC_S binds and coincides with IR_S, and IR_L binds. By part (*iv*) of Remark 2, $V_S\left(q^S\right) > V_L\left(q^L\right)$. Therefore, IC_L holds with slack.

Now suppose $q^L \neq (0,0)$, such that IR_S holds with slack, by Claim 1. Define

$$\Delta\left(q_0, q_1\right) = V_S\left(q_0, q_1\right) - V_L\left(q_0, q_1\right)$$

This is the difference between the two types' willingness to pay. By part (*iv*) of Remark 2, $\Delta\left(q_0, q_1\right) \geq 0$ (strictly so when $q \neq 0$). Suppose that $q^S \neq q^L$ and yet IC_L binds. Then, $\Delta(q_0^S, q_1^S) = \Delta(q_0^L, q_1^L)$. By part (*v*) of Remark 2, $\Delta\left(q_0, q_1\right)$ *decreases* in q_0 and *increases* in q_1. Therefore, it cannot be the case that either $\left(q_0^S \geq q_0^L \wedge q_1^S \leq q_1^L\right)$ (with at least one strict inequality), or $\left(q_0^S \leq q_0^L \wedge q_1^S \geq q_1^L\right)$ (with at least one strict inequality).

Suppose $\left(q_0^S \geq q_0^L \wedge q_1^S \geq q_1^L\right)$ (with at least one strict inequality). W.l.o.g, assume $q_0^S > q_0^L$. Since this means that $q_0^L < n_0$, there exist $\varepsilon, \delta > 0$ sufficiently close to zero such that $q_0^L + \varepsilon < n_0$, $V_L(q_0^L + \varepsilon, q_1^L) - \left(p^L + \delta\right) \geq 0$ and $v_S(q_0^S, q_1^S) - \left(p^S + \delta\right) \geq 0$ (because IR_S originally holds with slack). But this means that the monopolist can raise both prices without increasing its costs and without violating any of the constraints, a contradiction.

Suppose $\left(q_0^S \leq q_0^L \wedge q_1^S \leq q_1^L\right)$ (with at least one strict inequality). Since V_S increases in both of its arguments, $V_S(q_0^S, q_1^S) < V_S(q_0^L, q_1^L)$, and by IC_S, $p^S < p^L$. Hence, the monopolist can remove the contract $\left(q_0^S, q_1^S, p^S\right)$ from the menu, which raises revenues without affecting costs and without violating any of the constraints (IR_L is unaffected, IR_S holds since IR_L holds and there are no incentive constraints because the menu is a singleton). □

Claim 3. $q_t^S = n_t$ for every $t = 0, 1$.

Proof. Suppose $q^L = (0,0)$. Then, type L is effectively excluded; IC_S coincides with IR_S, such that the monopolist acts as if it only faces type S. In this case, it is clearly optimal to set $q_t^S = n_t$ for every t.

Suppose now $q^L \neq (0,0)$ and yet $q_t^S < n_t$ for some $t \in \{0,1\}$. Then, since V_S is continuous and increases in both of its arguments, and since IC_L holds with slack, there exist $\varepsilon, \delta > 0$ sufficiently close to zero such that $q_i^S + \varepsilon < n_0$ and

$$V_S(q_i^S + \varepsilon, q_{-i}^S) - V_S(q_i^S, q_{-i}^S) > \delta$$
$$V_L(q_0^L, q_1^L) - p^L > V_L(q_i^S + \varepsilon, q_{-i}^S) - \left(p^S + \delta\right)$$

This means that if the monopolist replaces the contract $\left(q_i^S, q_{-i}^S, p^S\right)$ with $\left(q_i^S + \varepsilon, q_{-i}^S, p^S + \delta\right)$, then type S will prefer $\left(q_i^S + \varepsilon, q_{-i}^S, p^S + \delta\right)$ to (q_0^L, q_1^L, p^L) but not type L (i.e., IC_S and IC_L both hold). The new S contract satisfies IR_S (because of our choice of (ε, δ) and because the original contract (q_0^L, q_1^L, p^L) satisfied IR_S with slack). But then the new menu increases revenues without affecting costs, a contradiction. □

Claim 4. IR_L binds.

Proof. If $q^S = q^L$, then $q_i^S = q_i^L = n_i$, implying that the monopolist does not discriminate between types. Hence, it is optimal for it to set a uniform access fee that is equal to $V_L(n_0, n_1)$.

Suppose next that $q^S \neq q^L$ and yet IR_L holds with slack. Then, the monopolist can raise p^L by a sufficiently small $\varepsilon > 0$ so as to still preserve IR_L and IC_L that held with slack. Since p^L increases without changing type L's data access, this only relaxes IC_S and raises profits, a contradiction. □

Claim 5. $q_0^L = q_0^S = n_0$.

Proof. By Claims 1-4, we can use the binding constraints to substitute for the access fees:

$$
\begin{aligned}
p^L &= V_L(q_0^L, q_1^L) \\
p^S &= V_L(q_0^L, q_1^L) + V_S(n_0, n_1) - V_S(q_0^L, q_1^L)
\end{aligned}
$$

and rewrite the monopolist's relaxed problem as follows:

$$
\max_{\left(n_0, n_1, q_0^L, q_1^L\right)} \left[\lambda V_S\left(n_0, n_1\right) + (1-\lambda) V_L(q_0^L, q_1^L) - c(n_0 + n_1) - \lambda \Delta(q_0^L, q_1^L)\right] \tag{6}
$$

subject to $q_i^L \in [0, n_i]$. Suppose $q_0^L < n_0$. Since raising q_0^L increases V_L and decreases $\Delta(q_0^L, q_1^L)$, raising q_0^L to n_0 improves the objective function, a contradiction. □

Claim 6. $\exists \lambda^* \in (0,1)$, such that $q_1^L = n_1$ if $\lambda \leq \lambda^*$ and $q_1^L = 0$ otherwise.

Proof. From (3) and (4), it follows that

$$
\frac{\partial V_L(q_0^L, q_1^L)/\partial q_1^L}{\partial V_S(q_0^L, q_1^L)/\partial q_1^L} = \sigma_\mu^4 \left(\frac{n_0 + 1}{n_0 + 1 + \sigma_\mu^2(2n_0 + 1)}\right)^2
$$

This ratio is independent of n_1. As a result, there exists λ^* such that the derivative of the objective function in the relaxed problem (7) is positive for $\lambda < \lambda^*$ and negative for $\lambda > \lambda^*$. This means that the optimal solution for q_1^L is extreme: $q_1^L = n_1$ for $\lambda < \lambda^*$, and $q_1^L = 0$ for $\lambda > \lambda^*$. □

Claim 7. When $q_1^L = 0$, the relaxed objective function is strictly concave.

Proof. Since $\Delta(n_0, 0) = 0$, when $q_1^L = 0$ the relaxed objective function is:

$$
(1-\lambda) V_L(n_0, 0) + \lambda V_S(n_0, n_1) - cn_0 - cn_1 \tag{7}
$$

By Remark 2, V_L and V_S are concave, hence a convex combination of them is also concave. □

Claim 8. $\exists \hat{c} > 0$ and $\exists \hat{\lambda} \in (0,1)$ such that for all $c < \hat{c}$ and $\lambda \geq \hat{\lambda}$: n_0 decreases in λ, n_1 increases in λ, and $n_1 > n_0$.

Proof. Clearly, if c is high enough, it is optimal not to store any data. For c small enough and for $\lambda > \lambda^*$ (where λ^* is as defined in the proof of Claim 6), there are positive (n_0, n_1, q_1^L) that solve the monopolist's problem. By Claim 7, this solution is unique and given by the solution to the first-order conditions,

$$\lambda \frac{\partial V_S(n_0, n_1)}{\partial n_1} = c$$

$$(1-\lambda)\frac{\partial V_L(n_0, 0)}{\partial n_0} + \lambda \frac{\partial V_S(n_0, n_1)}{\partial n_0} = c$$

which by (3) and (4) are given by

$$\frac{\lambda \left(n_0 + \sigma_\mu^2 + 2n_0\sigma_\mu^2 + 1\right)^2}{\left(n_0 + n_1 + n_0\sigma_\mu^2 + n_1\sigma_\mu^2 + n_0 n_1 + 2n_0 n_1 \sigma_\mu^2 + 1\right)^2} = c \tag{8}$$

and

$$\frac{\lambda \sigma_\mu^4}{\left(n_0 + n_1 + n_0\sigma_\mu^2 + n_1\sigma_\mu^2 + n_0 n_1 + 2n_0 n_1 \sigma_\mu^2 + 1\right)^2} + \frac{(1-\lambda)\sigma_\mu^4}{\left(n_0\sigma_\mu^2 + n_0 + 1\right)^2} = c \tag{9}$$

From these equations it follows that

$$n_1 = \sqrt{\frac{\lambda}{c}} - \frac{n_0 + n_0\sigma_\mu^2 + 1}{n_0 + \sigma_\mu^2 + 2n_0\sigma_\mu^2 + 1} \tag{10}$$

Differentiating the R.H.S. w.r.t n_0, we can see that as n_0 *decreases*, n_1 *increases*. Thus, if n_0 decreases when λ increases, then whenever $n_1 > n_0$ for some (λ, σ_μ, c), this continues to be true for $\lambda' > \lambda$.

We now show that indeed, n_0 decreases in λ. Plugging equation (10) into equation (9) and rearranging yields

$$(1-\lambda)\frac{\sigma_\mu^4}{\left(n_0\sigma_\mu^2+n_0+1\right)^2}+c\frac{\sigma_\mu^4}{\left(n_0+\sigma_\mu^2+2n_0\sigma_\mu^2+1\right)^2}=c$$

Note that the L.H.S. of this equation decreases in λ and also decreases in n_0. Hence, for (λ,σ_μ,c) such that $\lambda>\lambda^*$, if λ increases, n_0 decreases and so n_1 increases. Therefore, if $n_1>n_0$ at some $\lambda>\lambda^*$, this continues to hold for $\lambda'>\lambda$.

Finally, note that when $\lambda=1$, the monopolist's problem reduces to

$$\max_{n_0,n_1}\left[\lambda V_S(n_0,n_1)-c(n_0+n_1)\right]$$

Since the objective function is strictly concave, there is a threshold cost $\bar{c}$ such that for all $c<\bar{c}$, there is a unique interior solution given by the solution to the first-order conditions:

$$\lambda\frac{\partial}{\partial n_1}V_S(n_0,n_1)=\lambda\frac{\partial}{\partial n_0}V_S(n_0,n_1)=c$$

By properties (ii) and (iii) of Remark 2, the solution satisfies $n_1>n_0$. By continuity, there exists $\varepsilon>0$ such that for all $\lambda\in(1-\varepsilon,1]$, the solution $\left(n_0',n_1',q_1^L\right)$ to the monopolist's problem will also satisfy $n_1'>n_0'$. ■

Proposition **2**.

Throughout this proof, we take it as given that the first-best and second-best databases are strictly positive.

Let $f_1\left(n_0;x\right)$ be a function that maps each value of n_0 to a value of n_1 that solves the equation,

$$(1-\lambda)\frac{\partial V_L(n_0,n_1)}{\partial n_1}+\lambda\frac{\partial V_S(n_0,n_1)}{\partial n_1}=x \tag{11}$$

Likewise, let $f_0(n_0; y)$ be a function that maps each value of n_0 to a value of n_1 that solves the equation,

$$(1-\lambda)\frac{\partial V_L(n_0,n_1)}{\partial n_0} + \lambda\frac{\partial V_S(n_0,n_1)}{\partial n_0} = y \tag{12}$$

The L.H.S. of equations (11) and (12) are the derivatives of the first-best objective function (1) w.r.t n_1 and n_0, respectively.

By part (*ii*) of Remark 2, $f_0(n_0; x)$ and $f_1(n_0; y)$ are both decreasing in n_0 for every x and y, and there is a unique pair (n_0^*, n_1^*) (the unique solution to the first-best problem, which is interior by assumption) satisfying $n_1^* = f_1(n_0^*; c) = f_0(n_0^*; c)$. We claim that $f_1(n_0; c) < f_0(n_0; c)$ for $n_0 < n_0^*$ and $f_1(n_0^*; c) > f_0(n_0^*; c)$ for $n_0 > n_0^*$.

To see why, recall that $n_1^* > n_0^*$ for all $\lambda > 0$. By part (*iii*) of Remark 2, for $n_0 = n_1 = a$ satisfying $a = f_1(a; c)$ we have $a = f_0(a; c')$ for some $c' < c$. Hence, by part (*ii*) of Remark 2, $f_0(a; c) < a$. Since there is a unique solution to $f_1(n_0^*; c) = f_0(n_0^*; c)$, it follows that $f_1(n_0; c) > f_0(n_0; c)$ for $n_0 > n_0^*$ while $f_1(n_0; c) < f_0(n_0; c)$ for $n_0 < n_0^*$.

It will be useful to visualize $f_0(n_0; x)$ and $f_1(n_0; y)$ as downward-sloping "iso-marginal value" curves in the space $\mathbb{R}_+^2$, where the horizontal and vertical axes represent n_0 and n_1, respectively. We have thus established that the curve that represents $f_1(n_0; c)$ intersects the curve that represents $f_0(n_0; c)$ from below at a single point (n_0^*, n_1^*).

We now argue that this observation implies that the second-best database (n_0', n_1') satisfies $n_0' > n_0^*$ and $n_1' < n_1^*$ when the second-best solution satisfies $q_1^L = 0$. To see this, let $g_1(n_0; x)$ and $g_0(n_0; y)$ be the functions that map each value of n_0 to the values of n_1 that solve the equations

$$\lambda\frac{\partial V_S(n_0,n_1)}{\partial n_1} = x \tag{13}$$

and

$$(1-\lambda)\frac{\partial V_L(n_0,0)}{\partial n_0}+\lambda\frac{\partial V_S(n_0,n_1)}{\partial n_0}=y \tag{14}$$

respectively. The L.H.S. of equations (13) and (14) are the derivatives of the relaxed second-best objective function (7) w.r.t n_1 and n_0, respectively. By part (*ii*) of Remark 2, both $g_1(n_0;x)$ and $g_0(n_0;y)$ are decreasing in n_0. Thus, both are represented by downward-sloping "iso-marginal value" curves in the same $\mathbb{R}^2_{++}$ space we used to represent $f_0\,(n_0;x)$ and $f_1\,(n_0;y)$. By Claim 7 in the proof of Proposition 1, there exists a unique (n_0',n_1') satisfying $n_1'=g_1(n_0';c)=g_0(n_0';c)$. We will now show that $n_0'>n_0^*$ and $n_1'<n_1^*$ when $q_1^L=0$.

For any (n_0,n_1), the L.H.S. of (13) is lower than the L.H.S. of (11). By part (*ii*) of Remark 2, $\frac{\partial}{\partial n_1}\partial V_S(n_0,n_1)$ is decreasing in n_1. Therefore, the iso-marginal value curve that represents $g_1(n_0;x)$ lies *below* the curve that represents $f_1(n_0;x)$. In a similar vein, part (*ii*) of Remark 2 implies that $\frac{\partial}{\partial n_0}V_L(n_0,n_1)<\frac{\partial}{\partial n_0}V_L(n_0,0)$, such that the L.H.S. of (14) is higher than the L.H.S. of (12). Since $\frac{\partial}{\partial n_1}\partial V_S(n_0,n_1)$ is decreasing in n_1, it follows that the iso-marginal value curve that represents $g_0(n_0;x)$ lies *above* the curve that represents $f_0(n_0;x)$. As a result of the directions in which the curves that represent $g_1(n_0;x)$ and $g_0(n_0;y)$ are shifted relative to the curves that represent $f_1(n_0;x)$ and $f_0(n_0;y)$, the unique intersection (n_0',n_1') of the curves that represent $g_1(n_0;c)$ and $g_0(n_0;c)$ satisfies $n_0'>n_0^*$ and $n_1^*>n_1'$.

We next show that $n_0'>n_0^*$ and $n_1'<n_1^*$ also when $q_1^L=n_1$. Recall that in this case, the monopolist offers a single contract (n_0',n_1',p), where $p=V_L\,(n_0',n_1')$. Therefore, since V_L is strictly concave, (n_0',n_1') solve

$$\frac{\partial V_L}{\partial n_0}(n_0',n_1')=\frac{\partial V_L}{\partial n_1}(n_0',n_1')=c \tag{15}$$

By the symmetry of V_L, $n_0'=n_1'=b$. We claim that $n_0^*<b<n_1^*$. To see why, assume first that $b\geq n_1^*$ (which implies that $b>n_0^*$ since $n_1^*>n_0^*$).

Then,

$$c = (1-\lambda)\frac{\partial V_L(n_0^*, n_1^*)}{\partial n_1} + \lambda \frac{\partial V_S(n_0^*, n_1^*)}{\partial n_1} > \frac{\partial V_L(n_0^*, n_1^*)}{\partial n_1} > \frac{\partial V_L(b, b)}{\partial n_1}$$

where the first and second inequalities follow from parts (v) and (ii), respectively, of Remark 2. But the above inequality violates equation (15), a contradiction.

Next, assume $b \leq n_0^*$ (and hence, $b < n_1^*$). Then again by Remark 2,

$$c = (1-\lambda)\frac{\partial V_L(n_0^*, n_1^*)}{\partial n_0} + \lambda \frac{\partial V_S(n_0^*, n_1^*)}{\partial n_0} < \frac{\partial V_L(n_0^*, n_1^*)}{\partial n_0} < \frac{\partial V_L(b, b)}{\partial n_0}$$

violating equation (15). ■

Appendix II: Omitted Derivations

Derivation of Posterior Variances in Section 3

Recall the following independent Gaussian variables: $\mu \sim N(0, \sigma_\mu^2)$, $x_t \sim N(0, 1)$ and $\varepsilon_{t,i} \sim N(0, \sigma_\varepsilon^2)$, where $t = 0, 1$ and $i \in \{1, ..., n_t\}$. Also recall that an observation i from the the period t sample is a realization $y_{t,i} = \mu + x_t + \varepsilon_{t,i}$, and that types S and L are interested in forecasting $\theta^S = \mu + x_1$ and $\theta^L = \mu$, respectively. The prior variances over θ^S and θ^L are $\sigma_\mu^2 + 1$ and σ_μ^2, respectively.

From type L's point of view, a period-t sample generates a conditionally independent signal $\bar{y}_t = \theta^L + x_t + \bar{\varepsilon}_t$, where $\bar{\varepsilon}_t$ is the average observational noise in the period-t sample. The variance of the period-t signal conditional on θ^L is $1 + \sigma_\varepsilon^2 / n_t$. From S's point of view, the two periods' samples generate the signals $\bar{y}_1 = \theta^S + \bar{\varepsilon}_1$ and $\bar{y}_0 = \theta^S + x_0 - x_1 + \bar{\varepsilon}_0$. We now calculate the variance of the types' posterior beliefs.

For $c \in \{0, 1\}$, we have the following joint normal distribution (where $c = 0$ gives us the joint distribution with μ as the first variable and $c = 1$

gives us the joint distribution with $\mu + x_1$ as the first variable).

$$\begin{pmatrix} \mu + cx_1 \\ \bar{y}_0 \\ \bar{y}_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\mu^2 + c & \sigma_\mu^2 + c & \sigma_\mu^2 \\ \sigma_\mu^2 + c & \sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_0} & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 & \sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_1} \end{pmatrix}\right).$$

Denote

$$A := \begin{pmatrix} \sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_0}, \sigma_\mu^2 \\ \sigma_\mu^2, \sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_1} \end{pmatrix}$$

Then

$$\begin{aligned} det(A) &= (\sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_0})(\sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_1}) - \sigma_\mu^4 \\ &= \sigma_\mu^2(2 + \frac{\sigma_\varepsilon^2}{n_0} + \frac{\sigma_\varepsilon^2}{n_1}) + (1 + \frac{\sigma_\varepsilon^2}{n_0})(1 + \frac{\sigma_\varepsilon^2}{n_1}) \end{aligned} \tag{16}$$

and

$$A^{-1} = \frac{1}{det(A)} \begin{pmatrix} \sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_1}, -\sigma_\mu^2 \\ -\sigma_\mu^2, \sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_0} \end{pmatrix} \tag{17}$$

Therefore,

$$Var(\mu + cx_1|\bar{y}_0, \bar{y}_1) = \sigma_\mu^2 + c - \left(\sigma_\mu^2 + c, \sigma_\mu^2\right) A^{-1} \begin{pmatrix} \sigma_\mu^2 + c \\ \sigma_\mu^2 \end{pmatrix}$$

Plugging (17) into this expression yields that $Var(\mu + cx_1|\bar{y}_0, \bar{y}_1)$ reduces to

$$-\frac{[(\sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_1})c + \sigma_\mu^2(1 + \frac{\sigma_\varepsilon^2}{n_1})](\sigma_\mu^2 + c) + [-c\sigma_\mu^2 + \sigma_\mu^2(1 + \frac{\sigma_\varepsilon^2}{n_0})]\sigma_\mu^2}{det(A)}$$

When $c = 0$ we have

$$Var(\mu|\bar{y}_0, \bar{y}_1) = \sigma_\mu^2 - \frac{\sigma_\mu^4 \left(2 + \frac{\sigma_\varepsilon^2}{n_1} + \frac{\sigma_\varepsilon^2}{n_0}\right)}{\sigma_\mu^2 \left(2 + \frac{\sigma_\varepsilon^2}{n_1} + \frac{\sigma_\varepsilon^2}{n_0}\right) + \left(1 + \frac{\sigma_\varepsilon^2}{n_1}\right)\left(1 + \frac{\sigma_\varepsilon^2}{n_0}\right)}$$

When $c = 1$ we have

$$Var(\mu + x_1|s_1, s_2) = \sigma_\mu^2 + 1 - \frac{(\sigma_\mu^2 + 1)\left[(\sigma_\mu^2 + 1)(\sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_0}) - \sigma_\mu^4\right] + \sigma_\mu^4 \frac{\sigma_\varepsilon^2}{n_1}}{\left(\sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_0}\right)\left(\sigma_\mu^2 + 1 + \frac{\sigma_\varepsilon^2}{n_1}\right) - \sigma_\mu^4}$$

Proof of Remark **2.**

Proof of (i). This follows from noting that

$$\frac{\partial}{\partial n_0} V_L(n_0, n_1) = \frac{\sigma_\mu^4 (n_1 + 1)^2}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^2} > 0 \tag{18}$$

$$\frac{\partial}{\partial n_1} V_L(n_0, n_1) = \frac{\sigma_\mu^4 (n_0 + 1)^2}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^2} > 0 \tag{19}$$

$$\frac{\partial}{\partial n_0} V_S(n_0, n_1) = \frac{\sigma^4}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^2} > 0 \tag{20}$$

$$\frac{\partial}{\partial n_1} V_S(n_0, n_1) = \frac{\left(n_0 + \sigma_\mu^2 + 2\sigma_\mu^2 n_0 + 1\right)^2}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^2} > 0 \tag{21}$$

Proof of (ii). We begin by verifying that $V_L(n_0, n_1)$ is strictly concave. Its Hessian matrix is given by

$$\begin{matrix} \frac{\partial^2}{\partial(n_0)^2} V_L(n_0, n_1) & V_L(n_0, n_1) \\ \frac{\partial^2}{\partial n_1 \partial n_0} V_L(n_0, n_1) & \frac{\partial^2}{\partial(n_1)^2} V_L(n_0, n_1) \end{matrix}$$

The expressions for the terms in each cell are as follows:

$$\frac{\partial^2}{\partial (n_0)^2} V_L(n_0, n_1) = \frac{-2\sigma_\mu^4 (n_1+1)^2 \left(n_1 + \sigma_\mu^2 + 2\sigma_\mu^2 n_1 + 1\right)}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^3}$$

$$\frac{\partial}{\partial n_0 \partial n_1} V_L(n_0, n_1) = \frac{-2\sigma_\mu^6 (n_0+1)(n_1+1)}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^3}$$

$$\frac{\partial}{\partial n_1} V_L(n_0, n_1) = \frac{\sigma_\mu^4 (n_0+1)^2}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^2}$$

$$\frac{\partial^2}{\partial (n_1)^2} V_L(n_0, n_1) = \frac{-2\sigma_\mu^4 (n_0+1)^2 \left(n_0 + \sigma_\mu^2 + 2\sigma_\mu^2 n_0 + 1\right)}{\left(n_0 + n_1 + n_0 n_1 + \sigma_\mu^2 n_0 + \sigma_\mu^2 n_1 + 2\sigma_\mu^2 n_0 n_1 + 1\right)^3}$$

The function $V_L(n_0, n_1)$ is strictly concave if its Hessian matrix is negative definite. To see that this is true, note first that the first principal minor is negative: $\frac{\partial^2}{\partial(n_0)^2} V_L(n_0, n_1) < 0$. Second, note that the determinant of the Hessian matrix is positive:

$$
\begin{aligned}
&\frac{\partial^2}{\partial\left(n_0\right)^2} V_L(n_0, n_1) \cdot \frac{\partial^2}{\partial\left(n_1\right)^2} V_L(n_0, n_1) - \left(\frac{\partial^2}{\partial n_1 \partial n_0} V_L(n_0, n_1)\right)^2 \\
&= \frac{-2\sigma_\mu^4\left(n_1+1\right)^2\left(n_1+\sigma_\mu^2+2\sigma_\mu^2 n_1+1\right)}{\left(n_0+n_1+n_0 n_1+\sigma_\mu^2 n_0+\sigma_\mu^2 n_1+2\sigma_\mu^2 n_0 n_1+1\right)^3} \\
&\cdot \frac{-2\sigma_\mu^4\left(n_0+1\right)^2\left(n_0+\sigma_\mu^2+2\sigma_\mu^2 n_0+1\right)}{\left(n_0+n_1+n_0 n_1+\sigma_\mu^2 n_0+\sigma_\mu^2 n_1+2\sigma_\mu^2 n_0 n_1+1\right)^3} \\
&- \left(\frac{-2\sigma_\mu^6\left(n_0+1\right)\left(n_1+1\right)}{\left(n_0+n_1+n_0 n_1+\sigma_\mu^2 n_0+\sigma_\mu^2 n_1+2\sigma_\mu^2 n_0 n_1+1\right)^3}\right)^2 \\
&= \frac{4\sigma_\mu^8\left(n_0+1\right)^2\left(n_1+1\right)^2\left(n_1+\sigma_\mu^2+2\sigma_\mu^2 n_1+1\right)\left(n_0+\sigma_\mu^2+2\sigma_\mu^2 n_0+1\right) - 4\sigma_\mu^{12}\left(n_0+1\right)^2\left(n_1+1\right)^2}{\left(n_0+n_1+n_0 n_1+\sigma_\mu^2 n_0+\sigma_\mu^2 n_1+2\sigma_\mu^2 n_0 n_1+1\right)^6} \\
&= \frac{4\sigma_\mu^8\left(n_0+1\right)^2\left(n_1+1\right)^2\left(\left(n_1+\sigma_\mu^2+2\sigma_\mu^2 n_1+1\right)\left(n_0+\sigma_\mu^2+2\sigma_\mu^2 n_0+1\right) - \sigma_\mu^4\right)}{\left(n_0+n_1+n_0 n_1+\sigma_\mu^2 n_0+\sigma_\mu^2 n_1+2\sigma_\mu^2 n_0 n_1+1\right)^6} \\
&> 0
\end{aligned}
$$

We next turn to verifying that $V_S(n_0, n_1)$ is strictly concave. Its Hessian matrix is

$$
\begin{matrix}
\frac{\partial^2}{\partial(n_0)^2} V_S(n_0, n_1) & V_S(n_0, n_1) \\
\frac{\partial^2}{\partial n_1 \partial n_0} V_S(n_0, n_1) & \frac{\partial^2}{\partial(n_1)^2} V_S(n_0, n_1)
\end{matrix}
$$

The expressions for the terms in each cell are as follows:

$$\frac{\partial^2}{\partial\left(n_0\right)^2}V_S(n_0,n_1)=\frac{-2\sigma_\mu^4\left(n_1+\sigma_\mu^2+2\sigma_\mu^2n_1+1\right)}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^3}$$

$$\frac{\partial}{\partial n_0\partial n_1}V_S(n_0,n_1)=\frac{-2\sigma_\mu^4\left(n_0+\sigma_\mu^2+2\sigma_\mu^2n_0+1\right)}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^3}$$

$$\frac{\partial}{\partial n_1}V_S(n_0,n_1)=\frac{\left(n_0+\sigma_\mu^2+2\sigma_\mu^2n_0+1\right)^2}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^2}$$

$$\frac{\partial^2}{\partial\left(n_1\right)^2}V_S(n_0,n_1)=\frac{-2\left(n_0+\sigma_\mu^2+2\sigma_\mu^2n_0+1\right)^3}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^3}$$

$V_S(n_0,n_1)$ is strictly concave since the first principal minor is negative: $\frac{\partial^2}{\partial(n_0)^2}V_S(n_0,n_1)<0$, and the determinant of the Hessian matrix is positive:

$$\begin{aligned}
&\frac{\partial^2}{\partial\left(n_0\right)^2}V_S(n_0,n_1)\cdot\frac{\partial^2}{\partial\left(n_1\right)^2}V_S(n_0,n_1)-\left(\frac{\partial^2}{\partial n_1\partial n_0}V_S(n_0,n_1)\right)^2\\
&=\left(\frac{-2\sigma_\mu^4\left(n_1+\sigma_\mu^2+2\sigma_\mu^2n_1+1\right)}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^3}\right)\\
&\cdot\left(\frac{-2\left(n_0+\sigma_\mu^2+2\sigma_\mu^2n_0+1\right)^3}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^3}\right)\\
&-\left(\frac{-2\sigma_\mu^4\left(n_0+\sigma_\mu^2+2\sigma_\mu^2n_0+1\right)}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^3}\right)^2\\
&=\frac{4\sigma_\mu^8\left(n_0+\sigma_\mu^2+2\sigma_\mu^2n_0+1\right)^2\left[\left(n_1+\sigma_\mu^2+2\sigma_\mu^2n_1+1\right)\left(n_0+\sigma_\mu^2+2\sigma_\mu^2n_0+1\right)-1\right]}{\left(n_0+n_1+n_0n_1+\sigma_\mu^2n_0+\sigma_\mu^2n_1+2\sigma_\mu^2n_0n_1+1\right)^6}\\
&>0
\end{aligned}$$

Proof of (iii). From inspection of (3) it is easy to see that $V_L(x,y) = V_L(y,x)$. To see that $V_S(x,y) > V_S(y,x)$ for $y > x$, note that

$$V_S(x,y) - V_S(y,x) = \frac{(y-x)\left(2\sigma_\mu^2 + 1\right)}{\sigma_\mu^2(y + x + 2xy) + (1+x)(1+y)} > 0$$

The observation that $\frac{\partial V_S(n_0,n_1)}{\partial n_1} > \frac{\partial V_S(n_0,n_1)}{\partial n_0}$ follows from comparing equation (20) to equation (21).

Proof of (iv). Follows immediately from equations (3) and (4).

Proof of (v). Follows immediately from equations (18)-(21).

References

[1] Bergemann, D., & Bonatti, A. (2019). Markets for information: An introduction. Annual Review of Economics, 11(1), 85-107.

[2] Brito, D. L., & Oakland, W. H. (1980). On the monopolistic provision of excludable public goods. The American Economic Review, 70(4), 691-704.

[3] Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). The rising costs of training frontier AI models. arXiv preprint arXiv:2405.21015.

[4] Davidson, S. B., Gershtein, S., Milo, T., Novgorodov, S., & Shoshan, M. (2023). Efficiently Archiving Photos under Storage Constraints. In EDBT (pp. 591-603).

[5] Guerra, E., Wilhelmi, F., Miozzo, M., & Dini, P. (2023). The cost of training machine learning models over distributed data sources. IEEE Open Journal of the Communications Society, 4, 1111-1126.

[6] Haghpanah, N., & R. Siegel (2025). Screening Two Types. Working paper.

[7] Milo, T. (2019). Getting rid of data. Journal of Data and Information Quality (JDIQ), 12(1), 1-7.

[8] Norman, P. (2004). Efficient mechanisms for public goods with use exclusions. The Review of Economic Studies, 71(4), 1163-1188.