Is Dark Data Hoarding? Informational Decay and the Data Feedback Loop

Chujie Wang*

Abstract

Surveys report that more than half of the data firms store is never used, and this "dark data" share is routinely read as evidence of hoarding. We ask whether it is. We build a dynamic model in which data is a by-product of production—capital generates data, data sharpens productivity forecasts, and sharper forecasts raise the return to capital—and in which the informational value of a record decays through two compounding channels, physical obsolescence and the mean reversion of productivity. Our main result is an identification result: the share of stored data that is informationally inert depends only on the persistence of productivity and the depreciation of records, while firm scale, data-mining efficiency and managerial preferences cancel exactly. Applied to the observed 54.5% dark-data share, the model attributes 52.6 percentage points to optimal accumulation alone; the residual implies competitive intensity of 0.063 and over-accumulation of 6.4%, and even the sign of that estimate is not robust to plausible persistence and depreciation. Dark-data statistics therefore measure informational obsolescence, not hoarding. Within the model, the feedback loop raises the return to capital by only 3.1% but speeds a data-poor firm's convergence to its optimum. Competitive escalation is the most powerful of three hoarding wedges, yet the capital misallocation they cause is almost costless, with firm-value losses below 0.0013%. The resource cost lies elsewhere: in duplicated data infrastructure, about 1.2% of industry value, and, in a microfounded version of the competitive channel, in the maintenance expenditure that rivalry induces, which is of the same order. Both point to interoperability as the instrument with the most leverage.

Keywords: Dark data, Data hoarding, Data feedback loop

JEL Classification: D83, E22, L13, O33

*The University of Chicago. Email: chujie@uchicago.edu.

1 Introduction

In the digital economy era, data has become one of the most critical strategic assets for enterprises. With the rapid advancement of information technology, the volume of data accumulated by enterprises has grown exponentially. However, not all data possesses tangible commercial value. Industry surveys report that a substantial share of stored enterprise data is "dark data"—data that is collected and retained but whose value is never established. Veritas Technologies (2020) puts the figure at 54.5% for Chinese enterprises and above 50% globally. Because this estimate originates in a vendor survey rather than in audited accounts, we treat it as motivation rather than as evidence, and note that the accounting reforms now requiring Chinese firms to disclose capitalized data resources should soon permit direct measurement. We also return to this figure in Section 5.4 and argue, on the basis of our own framework, that it says considerably less about hoarding than it is usually taken to say. The natural reading is that firms hoard data far beyond its productive value. This paper asks whether that reading is warranted, by building a framework in which the optimal data stock can be computed and the observed statistic compared with it.

Theoretically, data is a tool for forecasting random outcomes and optimizing business processes, and Bajari et al. (2019) document empirically that larger data holdings improve firms' forecast accuracy, which in a model such as ours translates into a higher marginal product of capital and hence a larger optimal scale. Existing work, however, largely treats the firm's informational environment as exogenous (Eeckhout and Veldkamp 2022; Farboodi, Mihet, et al. 2019), overlooking a critical mechanism: data is not merely an input that firms purchase or invest in independently, but is fundamentally a by-product of economic activity. When a firm produces more, it generates more transactions, more customer interactions, and consequently more data. This data, in turn, improves the firm's understanding of its productivity environment, leading to better capital allocation decisions, larger scale, and even more data generation. This self-reinforcing data feedback loop lies at the heart of the digital economy, yet its implications for data-asset mismatch have not been formally

analyzed.

A second mechanism, equally important and equally neglected, concerns how informational value decays. A datum is a record of a past state of the world. Its usefulness for forecasting today's productivity therefore erodes for two distinct reasons: the record may be lost, corrupted or rendered unreadable, and the state it describes may have drifted away from the current state. The first is physical obsolescence, the second is the mean reversion of the productivity process itself. We show that under a natural age-weighting of signals these two forces compound into a single informational retention coefficient $\zeta = (1-\delta)\rho^2$, and that this coefficient—not the physical depreciation rate alone—governs both the strength of the feedback loop and the effectiveness of data-retention policy.

This paper develops a dynamic theoretical framework that incorporates the data feedback loop into firms' optimal decision-making. We model data as a form of informational capital that improves productivity forecasts through Bayesian learning. The central innovation is the dual role of physical capital: investment in capital not only generates current output but also produces data as a by-product, which enhances future knowledge precision. This creates a recursive system where real capital accumulation and informational improvement co-evolve.

This paper contributes to the literature on the economics of data in five respects. First, we integrate the data feedback loop into a firm's dynamic optimization problem, combining Bayesian learning with endogenous data generation, capital adjustment costs and fixed operating costs. Second, we derive rather than assume the depreciation of informational assets, showing that physical obsolescence and the mean reversion of productivity compound multiplicatively, and that the implied retention coefficient generates a testable prediction absent from existing work: the feedback loop is stronger in industries with more persistent productivity fundamentals. Third, we quantify the transitional dynamics from data poverty, and find that the feedback loop matters far more for the speed of that transition than for the level of the steady state. Fourth, we separate the optimal benchmark from three structural

wedges that push firms beyond it and show that they differ in magnitude by two orders of magnitude, with competitive escalation the most powerful per unit of intensity, managerial signalling second and precautionary accumulation negligible. Fifth, and in our view most substantively, we confront the competitive mechanism with data rather than assuming its intensity. The ratio of informative to stored data is a closed-form function of the persistence of productivity and the depreciation of records alone; firm scale and managerial preferences cancel from it exactly. This yields both an identification result—the dark-data share speaks to the competitive channel and to nothing else—and a negative finding: at our calibration, 52.6 of the observed 54.5 percentage points of dark data arise under fully optimal accumulation, the residual implies competitive intensity of only 0.063, and the sign of that estimate is not robust. Dark-data statistics measure informational obsolescence, and inferences about hoarding drawn from them are unsafe.

A methodological point deserves emphasis, since it bears on how the quantitative results should be read. In this class of models the marginal value of precision falls as $1/(1+\Omega)^2$, so a calibration that places the firm in the well-informed region mechanically implies a weak level effect from additional data. We find an amplification of 3.1%, an order of magnitude smaller than one might expect from the prominence the feedback loop receives in the literature. The magnitude is not robust in the way the sign is: it is governed by the scale of the information-value function, which is not separately identified from the signal-noise normalization. We therefore report it as a calibrated implication rather than a structural estimate, and locate the paper's substantive claims in the hoarding wedges of Section 5, which are identified by ratios and are correspondingly more robust.

2 Literature Review

Data assets have become emerging products which are essential for corporate development. As an asset, data possesses not only the non-rival and non-exhaustible features of

intellectual property, but also time-sensitivity, value integration and non-monetary transaction models (Ahmad and Ven 2018; Rassier et al. 2019). Data capital is in this respect closely analogous to organization capital, whose measurement raises the same difficulty of an asset that is accumulated as a by-product of operations rather than purchased (Atkeson and Kehoe 2005). Farboodi and Veldkamp (2021) develop a general equilibrium model of the data economy in which data is accumulated through production and can be traded. They show that the non-rivalry of data creates a tension between increasing and decreasing returns: data can be used simultaneously by multiple firms, generating economy-wide increasing returns, but the diminishing precision of forecasts at the firm level imposes decreasing returns. Jones and Tonetti (2020) establish the allocative counterpart, that non-rivalry makes the assignment of data property rights consequential for welfare, and Acemoglu et al. (2022) show that when one agent's data is informative about others, equilibrium data acquisition is excessive relative to the social optimum. On measurement, Xu et al. (2025) construct a data-capital stock for China using the perpetual inventory method; Goodridge et al. (2015) provide related estimates of data-related activity for the UK. Cong et al. (2022) embed multiple uses of the same data in an endogenous growth framework, and J. Li et al. (2025) argue that the capitalization of data assets can itself generate valuation bubbles—a concern directly relevant to the managerial-signalling channel we model in Section 5. What unites these approaches is their treatment of data as an input. Our paper departs from this tradition by modeling data as a by-product of production. When data is a by-product, the scale of production determines the flow of new data, which in turn affects the precision of future production decisions. This creates a feedback loop between real capital accumulation and informational improvement that is absent when data is merely an investment choice.

The idea that firms operate under imperfect information about their own productivity has a long tradition in macroeconomics. Jovanovic (1982) formalized the idea that firms learn about their efficiency through production: each period's output reveals a noisy signal, and Bayesian updating gradually narrows the firm's belief about its true type. In this

framework learning is passive: firms observe outcomes generated at whatever scale they happen to operate, and the quantity of information they receive is exogenous to their choices. This passive-learning tradition has been extended to general equilibrium settings (Hopenhayn 1992) and enriched with more general information structures, but the exogeneity of the signal flow remains its defining feature. David et al. (2016) show that dispersion in the precision of firms' productivity forecasts generates measurable misallocation and aggregate productivity losses, while Bloom (2009) demonstrates that time-varying uncertainty combined with capital adjustment costs produces wait-and-see dynamics in which firms delay investment following uncertainty shocks. Farboodi, Mihet, et al. (2019) model firms receiving private signals about future productivity and show that the resulting dispersion in capital choices generates measurable TFP differences. A firm's data stock improves the precision of its productivity forecasts, and data accumulates as a function of past production. However, in their quantitative implementation, the feedback from current investment to future data is attenuated by the assumption that data depends on lagged output rather than current capital choice, and the focus is on the cross-sectional distribution of firm sizes rather than on the within-firm transition path. Our model transforms the learning problem from passive to active: the firm can invest its way to better information. But doing so involves a tradeoff. Higher capital generates more data and better future forecasts, but it also incurs adjustment costs and commits resources that could otherwise generate current profits.

Alongside the theoretical literature on data as an asset, industry surveys document a striking fact: most of the data firms accumulate is never used. Veritas Technologies (2020) estimates that 54.5% of enterprise data in China is "dark data"—information collected and stored but whose value remains unknown—with the global share also above 50%; J. Li et al. (2025) raise the related concern that the accounting recognition of data assets may outrun their productive content. Existing economic explanations for such hoarding center on market structure. Prüfer and Schottmüller (2021) develop a dynamic duopoly model in which an incumbent accumulates data to raise the cost of entry for potential competitors, a form of

data-based predation: the incumbent over-invests in data not because the data improves its own productivity, but because withholding it from rivals preserves market power. Hagiu and Wright (2023) characterize the conditions under which data-enabled learning confers durable competitive advantage, and Bergemann and Bonatti (2019) survey the resulting markets for information. Acemoglu et al. (2022) show how informational externalities across agents push equilibrium data acquisition above the social optimum, and Koutroumpis et al. (2020) document the institutional obstacles that keep markets for data thin, which is itself a reason firms retain data rather than trade it. These frameworks share the assumption that hoarding is strategic—a weapon in competition, not a consequence of internal firm dynamics. Our paper contributes to this literature by identifying mechanisms that generate data hoarding even in the absence of market power. We introduce three channels—competitive escalation, institutional incentives, and precautionary accumulation—that each drive a wedge between the optimal data stock and the equilibrium outcome. Critically, these deviations arise within a rational-agent framework: they stem from organizational structure (agency), market structure (competition), and informational structure (uncertainty), not from behavioral biases. We use the term "structural" in this sense: whatever accumulation exceeds the optimum has identifiable roots in the design of firms, markets, and information environments, rather than in the irrationality of individual decision-makers.

A third strand concerns managerial incentives to over-invest in observable assets. Bebchuk and Stole (1993) show that when investment is observable but its return is not, managers distort investment to influence the market's inference about their type, and Holmström (1999) develops the dynamic career-concerns logic behind such distortions. We apply this logic to data: because a data stock is easy to display and its productive contribution is hard to verify, it is an unusually attractive object on which to signal.

Finally, on capital adjustment costs, Abel and Eberly (1994) provide a unified framework for investment under convex and non-convex adjustment costs, showing that the optimal policy features a region of inaction when costs are non-convex and gradual adjustment when they

are convex. Cooper and Haltiwanger (2006), using plant-level data from US manufacturing, estimate that both convex and non-convex adjustment costs are empirically important, with the former governing the speed of adjustment and the latter generating lumpy investment episodes; Hall (2004) obtains comparably small convex adjustment costs from aggregate factor data. Our model, to our knowledge, provides the first quantitative characterization of the data feedback channel's effect on transitional dynamics—in particular, the finding that forward-looking firms invest through data poverty, accepting lower current profits to generate the information that validates future expansion.

3 Model

3.1 Firm Environment

Consider a discrete-time, infinite-horizon economy populated by a representative firm that combines physical capital and stochastic productivity to produce a homogeneous output. The production technology takes a standard Cobb-Douglas form:

$$y_t = A_t \cdot k_t^{\alpha} \tag{1}$$

where k_t denotes the physical capital stock installed at the beginning of period t, and A_t represents the true, but imperfectly observed, total factor productivity (TFP). The curvature parameter $\alpha \in (0, 1)$ captures the output elasticity of capital, implying diminishing returns to capital accumulation.

The fundamental source of uncertainty in this environment stems from the unobservable nature of productivity. While the firm utilizes capital in production, it does not possess perfect knowledge of the current productivity level A_t. Instead, it must form and update beliefs based on available information. To capture the persistent nature of productivity shocks commonly documented in empirical macroeconomics, we assume that the logarithm

of productivity follows an AR(1) process:

$$a_t = \rho \cdot a_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_\theta^2) \tag{2}$$

where $a_t \equiv \ln A_t$. The persistence parameter $\rho \in (0, 1)$ governs the degree to which productivity shocks propagate over time, while σ_θ^2 measures the volatility of fundamental innovations. The normality assumption, combined with the linear Gaussian structure, facilitates analytical tractability in the subsequent Bayesian learning framework.

The key informational friction in this model arises from the fact that the firm cannot observe a_t directly. Instead, it possesses a stock of data capital, denoted D_t, which generates noisy signals about the latent productivity state. Each period, the firm extracts signals from its data stock:

$$s_{t,m} = a_t + \varepsilon_{t,m}, \quad \varepsilon_{t,m} \sim N(0, \sigma_\varepsilon^2) \tag{3}$$

where the measurement error $\varepsilon_{t,m}$ is independently and identically distributed over time and independent of the fundamental shock η_t. The precision of each signal, $1/\sigma_\varepsilon^2$, reflects the quality of the firm's data infrastructure. The number of signals extractable from the data stock depends on D_t through the data mining technology.

The firm's information set at any point in time is characterized by its posterior beliefs about a_t. For analytical convenience, we work with knowledge precision, defined as the inverse of the posterior variance:

$$\Omega_t \equiv \text{Var}[a_t \mid \mathcal{I}_t]^{-1} \tag{4}$$

This precision metric captures the firm's degree of informational awareness: higher Ω_t implies more accurate knowledge of productivity, which in turn enables more efficient capital allocation decisions.

3.2 The Data Feedback Loop and Informational Decay

A central innovation of our framework is the explicit modeling of the data feedback loop: data is a by-product of economic activity. Following Farboodi and Veldkamp (2021) and Farboodi, Mihet, et al. (2019), we assume that the volume of new data generated each period is proportional to the scale of production:

$$n_{t+1} = z \cdot k_t^{\alpha} \tag{5}$$

where $z > 0$ is the data mining efficiency parameter. This parameter captures the firm's technological and organizational capacity to harvest usable data from its production activities. Firms with higher z extract more data per unit of output, due to superior data infrastructure, more sophisticated data collection systems, or business models that naturally generate richer transactional data (Freeman et al. 2023; T. Li et al. 2020).

Physical capital k_t thus serves a dual role: it generates current output (and hence current profits), and it produces data that enhances future knowledge precision. This dual role creates a fundamental interdependence between production and learning that distinguishes our framework from standard capital accumulation models.

The stock of data capital evolves according to:

$$D_{t+1} = (1 - \delta)D_t + z \cdot k_t^{\alpha} \tag{6}$$

where $\delta \in (0, 1)$ is the rate of *physical* depreciation of data capital: records are lost, corrupted, migrated into unreadable formats, or discarded. Equation (6) describes the accounting stock—the quantity a firm would report on its balance sheet, and the quantity industry surveys of "dark data" attempt to measure.

The accounting stock is not, however, the object that determines what the firm knows. A datum is a record of a *past* state of the world. Because productivity mean-reverts, a signal drawn from a record of a_{t-j} is informative about a_t only to the extent that a_{t-j} predicts

a_t, and that predictive content decays geometrically at rate ρ^2 in the age j of the record. Two distinct forces therefore erode the informational value of the stock: the record may not survive (at rate δ), and, conditional on surviving, what it describes may no longer be relevant (at rate $1-\rho^2$). Weighting each surviving datum by its age-specific informational content and summing gives the firm's stock of *effective* data,

$$\tilde{D}_{t+1} = \zeta \cdot \tilde{D}_t + z \cdot k_t^{\alpha}, \qquad \zeta \equiv (1-\delta)\,\rho^2 \in (0,1), \tag{6'}$$

where the informational retention coefficient ζ compounds physical survival $(1-\delta)$ with informational relevance ρ^2. Appendix A motivates (6′) from the age-weighting argument and states precisely what is approximate in it.

Posterior precision is then the precision of the unconditional prior plus the accumulated precision of the effective stock:

$$\Omega_t = \frac{1}{\sigma_a^2} + \frac{\tilde{D}_t}{\sigma_\varepsilon^2}, \qquad \sigma_a^2 \equiv \frac{\sigma_\theta^2}{1-\rho^2}. \tag{7'}$$

The first term is the precision a firm with no data at all would possess, obtained from the stationary distribution of the AR(1) process; it is the floor below which precision cannot fall. The second term is the informational contribution of the data stock, with each effective unit contributing $1/\sigma_\varepsilon^2$.

Two features of (7′) deserve comment, because they distinguish it from the recursive Kalman form used in much of this literature. First, precision is written as a *level* rather than a recursion in Ω_{t-1}. This is deliberate: a recursive prior term already embeds every signal the firm has ever observed, so adding a stock term on top of it would count the same information twice—once through the carried-forward prior and once through the surviving records. Writing precision as prior-plus-accumulated-stock counts each datum once, with an age discount that is exact for a weak signal and slightly too generous for an aggregate of many (Appendix A). This rests on an explicit assumption: knowledge is embodied in the records,

so that a firm that loses a record loses what it knew from it, instead of carrying a separately stored posterior. With a stored posterior the stock would not matter, as the remark in Appendix A notes. Second, because the age discount is geometric, the accumulation in (6′) is linear and the state space stays two-dimensional, $(\tilde{D}_t, k_{t-1})$, which keeps the dynamic program tractable without sacrificing the feedback mechanism.

The joint dynamics of $(k_t, \tilde{D}_t, \Omega_t)$ thus constitute a recursive system with a feedback mechanism:

$$\text{Higher } \Omega_t \to \text{better forecasts} \to \text{larger optimal } k_t \to \text{more new data} \to \text{larger } \tilde{D}_{t+1} \to \text{higher } \Omega_{t+1},$$

and conversely,

$$\text{Lower } \Omega_t \to \text{poorer forecasts} \to \text{smaller optimal } k_t \to \text{less new data} \to \text{smaller } \tilde{D}_{t+1} \to \text{lower } \Omega_{t+1}.$$

This feedback loop between real capital accumulation and informational improvement lies at the heart of the model's economic mechanism. It amplifies the return to capital investment beyond the myopic level, as firms internalize the future data benefits of current production. Section 4 shows that the amplification is quantitatively modest at our calibration—3.1%—but that it has a pronounced effect on the *speed* at which a data-poor firm converges to its optimum.

3.3 Micro-foundations of Data Depreciation

A recurring difficulty in the economics of data is that the depreciation rate of an informational asset is typically imposed rather than derived. Our formulation avoids this, because the retention coefficient $\zeta = (1 - \delta)\rho^2$ is the product of two economically distinct objects that can in principle be measured separately. Writing total informational decay as

$1-\zeta$ and decomposing,

$$\underbrace{1-\zeta}_{\text{total decay}} \;=\; \underbrace{(1-\rho^2)}_{\text{relevance channel}} \;+\; \underbrace{\delta\,\rho^2}_{\text{obsolescence channel}}\,, \tag{7a}$$

an exact identity with no approximation and no dependence on the level of precision.

The two channels differ in what they represent and in whether policy can reach them. The *relevance channel*, $1-\rho^2$, is the rate at which the productivity state drifts away from what any record describes. It operates even if every datum ever collected survives in perfect condition, it is a property of the economic environment rather than of the firm, and it is essentially beyond the reach of either managers or regulators. The *obsolescence channel*, $\delta\rho^2$, is the additional loss from records that do not survive: format migration, schema changes, organizational forgetting as key personnel depart, and the gradual decay of undocumented data pipelines. This channel is amenable to intervention—through standardization, interoperable formats and institutional knowledge management—and it is the channel on which data-retention regulation operates.

The decomposition delivers a testable prediction that does not appear in existing work. Because ζ is increasing in ρ, the feedback loop is stronger in industries whose productivity fundamentals are more persistent: there, data retains its predictive content longer, so the informational annuity from a unit of data is larger. Section 4 confirms this quantitatively—the amplification rises from 0.9% at $\rho = 0.60$ to 4.9% at $\rho = 0.95$—and it implies that utilities and mature manufacturing should exhibit stronger data-investment complementarities than fashion retail or consumer technology, holding data-mining efficiency fixed.

A further implication concerns the interpretation of δ itself. Since only the product $(1-\delta)\rho^2$ enters the firm's problem, a retention mandate that raises δ and an environment that lowers ρ are informationally equivalent. Policy that forces deletion is therefore a substitute for turbulence in the economic environment, which is a somewhat uncomfortable way of putting the case for deletion mandates, and one reason we find below that interoperability

dominates them.

Mapping the parameters to industries. In industries where productivity fundamentals are highly persistent—utilities, mature manufacturing—the relevance channel is weak, ζ is high, and data is a durable asset. Industries characterized by rapid technological change or shifting consumer preferences face a low effective ρ, and there even perfectly preserved records lose their forecasting value quickly. Our comparative statics in δ, z and ρ can therefore be read as mapping both the structural characteristics of industries and the potential reach of data-governance policy onto equilibrium data accumulation.

3.4 Dynamic Optimization

Given the informational frictions and state dynamics described above, the firm's decision problem is dynamic and recursive. At each point in time the firm must make choices that not only affect current profits but also shape its future information set and productive capacity.

We formulate the problem as a dynamic program with two state variables: the effective data stock $\tilde{D}_t$, which by $(7')$ determines precision, and the lagged physical capital stock k_{t-1}, which enters through the adjustment cost. Both are predetermined at the beginning of period t. Note that precision is not a separate state: because $(7')$ makes Ω_t a deterministic function of $\tilde{D}_t$, carrying both would be redundant. This is a direct consequence of writing precision as a level rather than a recursion, and it reduces the state space from three dimensions to two. The firm chooses one control, the physical capital input k_t, which simultaneously determines current output, future data generation, and the adjustment cost incurred.

The firm's objective is to maximize the expected present discounted value of current and future net cash flows. In recursive form, the Bellman equation is:

$$V(\tilde{D}_t, k_{t-1}) = \max_{k_t} \left\{ \mathbb{E}[\pi(k_t, \Omega(\tilde{D}_t), k_{t-1})] + \beta \cdot V(\tilde{D}_{t+1}, k_t) \right\} \tag{8}$$

subject to the transition equations:

$$\tilde{D}_{t+1} = \zeta \tilde{D}_t + z \cdot k_t^{\alpha}, \qquad \zeta = (1-\delta)\rho^2$$

$$\Omega_{t+1} = \frac{1}{\sigma_a^2} + \frac{\tilde{D}_{t+1}}{\sigma_\varepsilon^2}$$

where $V(\cdot,\cdot)$ denotes the value function and $\beta \in (0,1)$ is the discount factor. The accounting stock D_t evolves alongside according to (6); it does not enter the firm's problem in the benchmark, but it is the object the manager values in Section 5 and the object a regulator or accountant observes.

The current-period payoff is:

$$\mathbb{E}[\pi(k_t, \Omega_t, k_{t-1})] = P \cdot \mathbb{E}[A_t \mid \Omega_t] \cdot k_t^{\alpha} - r \cdot k_t - F - \frac{\gamma}{2}(k_t - k_{t-1})^2 \tag{9}$$

where P is the output price, r is the rental rate of capital, $F \geq 0$ is a fixed operating cost, and $\gamma \geq 0$ is a capital adjustment cost parameter. The quadratic adjustment cost captures the idea that rapidly scaling up (or down) physical capital is costly—firms cannot instantaneously jump to their desired capital stock but must instead incur adjustment frictions when changing their scale of operation. This friction is essential for realistic transitional dynamics: without it, a forward-looking firm would immediately adjust to its steady-state capital level, eliminating any meaningful transition path. The fixed cost F represents the minimum expenditure required to maintain data infrastructure (servers, IT personnel, compliance). If the firm chooses $k_t = 0$, it incurs neither the fixed cost nor the adjustment cost and earns zero profit.

The term $\mathbb{E}[A_t \mid \Omega_t]$ requires care, because the way precision enters profits is easy to misstate. It is tempting to treat $\mathbb{E}[A_t \mid \Omega_t]$ as expected productivity and to assume it rises with precision. That assumption is inconsistent with the Gaussian structure of Section 3: with $a_t \mid I_t \sim N(\hat{a}_t, \Omega_t^{-1})$ and $A_t = \exp(a_t)$, we have $\mathbb{E}[A_t \mid I_t] = \exp(\hat{a}_t + 1/(2\Omega_t))$, which is *decreasing* in precision. Better information does not raise the conditional mean of

productivity; by the law of iterated expectations it cannot.

What better information does is allow the firm to match capital to the state it faces. Let $k^*(a)$ denote the full-information capital choice and let the firm instead choose k_t on the basis of its posterior. Expected profit then equals its full-information value less a loss that is second order in the forecast error, and for our purposes we write that loss as $\kappa/(1+\Omega_t)$, declining in precision and bounded as $\Omega_t \to 0$. The Ω-dependent component of the profit coefficient is therefore

$$\kappa - \frac{\kappa}{1+\Omega_t} \;=\; \kappa\,\frac{\Omega_t}{1+\Omega_t}, \tag{9a}$$

so the functional form we use below is not an assumption about expected productivity but a consequence of the forecast-loss structure, with κ interpreted as the curvature of the profit function in the forecast error rather than as a maximum productivity level. We retain the notation $\mathbb{E}[A \mid \Omega] = \bar{A}\,\Omega/(1+\Omega)$ with $\bar{A} \equiv \kappa$ for continuity with the literature, but the object is an information-adjusted profit coefficient throughout.

This distinction matters for interpretation rather than for the algebra: the transition equations, optimality conditions and quantitative results below are unchanged. It matters a great deal, however, for whether the mechanism is credible, since the direct reading would have the feedback loop running through a channel that the model's own information structure forbids. We take P and r as given, consistent with a competitive industry or a small open economy facing a world interest rate.

What depends on this functional form and what does not. The form $\kappa\,\Omega/(1+\Omega)$ is a modelling choice, so it is worth stating which results rely on it. Two sets of results do not. The dark-data share (20) is a ratio of two stocks that are both proportional to $z(k^*)^\alpha$, so it is independent of the profit function altogether and depends only on ρ and δ. The size of the competitive wedge is likewise governed almost entirely by the effective depreciation rate δ^{Nash} acting on $D^* = z(k^*)^\alpha/\delta$, since capital itself barely moves; it stays at about 30% at $\lambda = 0.30$ across the full range of innovation variances in Appendix G. What does depend on the form is the *level* of the feedback amplification $\mathcal{M}$, which is governed by how fast the

marginal value of precision declines. We bound that dependence in Appendix G rather than claim to have identified it, and we report $\mathcal{M}$ throughout as a calibrated implication.

Interpretation of the capital path. The value of information arises because a better-informed firm matches capital to the productivity state it faces, so in a full treatment capital would respond to the posterior mean and be stochastic. We instead solve for a deterministic capital path and carry the value of information as a reduced-form term in the profit coefficient, (9a). The path k_t should therefore be read as the mean (certainty-equivalent) capital stock, and (9a) as the expected profit gain from state-contingent adjustment around it. This is a reduced form, not a derivation: it is exact only if the adjustment around the mean is costless and orthogonal to the firm's choice of the mean, and the quantitative results should be read with that in mind. The results that do not run through this term, namely the dark-data share and the depreciation-driven wedges, are unaffected.

The fixed cost F plays a critical role in the equilibrium structure. Without F, the marginal product of capital approaches infinity as $k \to 0$ (since $\alpha < 1$), implying that firms always have an incentive to operate at some positive scale, however small. With Cobb-Douglas production, this eliminates the possibility of a low-level equilibrium. The fixed cost introduces a threshold: for sufficiently small k, the operating surplus $P \cdot \mathbb{E}[A \mid \Omega] \cdot k^{\alpha} - r \cdot k$ cannot cover F, making zero operation ($k = 0$) optimal. This creates an inactivity threshold: firms with low knowledge precision and low data stocks may find it optimal to remain inactive rather than invest at a scale large enough to both cover the fixed cost and generate meaningful data feedback. The fixed cost thus formalises a realistic feature of data-intensive industries, where maintaining data infrastructure requires non-trivial ongoing expenditure (Hopenhayn 1992; Melitz 2003). It also carries the social cost of hoarding in Section 4: when many firms each maintain overlapping infrastructure, F is paid repeatedly for substantially the same information.

The recursive formulation reveals the key intertemporal trade-offs facing the firm. In each period, the firm must allocate capital between three competing considerations: generating

current profits through production, managing the cost of changing its capital stock, and generating future data through the by-product mechanism. The value function $V(\tilde{D}, k_{-1})$ is increasing in both arguments: a larger effective data stock raises precision and hence the profitability of any given capital choice, and a higher lagged capital stock reduces current-period adjustment costs.

3.5 Optimality Conditions

The choice of physical capital k_t is inherently dynamic, as it affects both current profits and future states through the data generation channel. Differentiating the Bellman equation with respect to k_t yields the first-order condition:

$$\frac{\partial \mathbb{E}[\pi]}{\partial k_t} + \beta \cdot \frac{\partial V}{\partial \tilde{D}_{t+1}} \cdot \frac{\partial \tilde{D}_{t+1}}{\partial k_t} = 0 \tag{10}$$

Computing the individual derivatives (note that the fixed cost F does not appear, as it is independent of k_t):

$$\frac{\partial \mathbb{E}[\pi]}{\partial k_t} = \alpha P \cdot \mathbb{E}[A_t \mid \Omega_t] \cdot k_t^{\alpha-1} - r - \gamma(k_t - k_{t-1}), \qquad \frac{\partial \tilde{D}_{t+1}}{\partial k_t} = \alpha z \cdot k_t^{\alpha-1}.$$

Define the single shadow price

$$\lambda_t \equiv \frac{\partial V}{\partial \tilde{D}_t}.$$

It is worth pausing on why there is only one. A formulation carrying both a data stock and a precision state invites two shadow prices, λ_D and λ_Ω, and an Euler equation containing $\lambda_{D,t+1} + \lambda_{\Omega,t+1}/\sigma_\varepsilon^2$. But if precision is a deterministic function of the stock, these are not independent objects, and adding them double counts: the informational benefit of a surviving datum is credited once through the continuation value of the stock and again through its contribution to precision. Because the two terms are of comparable magnitude, the error is not small—at our calibration it inflates the data feedback premium roughly fourfold. Collapsing to a single state removes the possibility of this error by construction.

Substituting yields the Euler equation for optimal capital choice, stated here at an interior optimum with $k_t = k_{t-1}$ for clarity:

$$r = \alpha k_t^{\alpha-1} \cdot [P \cdot \mathbb{E}[A_t \mid \Omega_t] + \beta z \cdot \lambda_{t+1}] \tag{11}$$

The left-hand side is the marginal cost of capital. The right-hand side is the marginal benefit, comprising the contemporaneous marginal revenue product $P \cdot \mathbb{E}[A_t \mid \Omega_t]$ and the *data feedback premium* $\beta z \lambda_{t+1}$—the discounted value of the information that today's production will generate.

The Euler equation is independent of the fixed cost F. Since F does not depend on k_t, it vanishes from the first-order condition. The fixed cost therefore does not directly alter the intensive margin of investment—the optimal capital choice conditional on operating—but it alters the extensive margin: the firm operates only if the value of operating exceeds the value of remaining inactive. Formally, the firm chooses $k_t > 0$ only if:

$$\max_{k_t > 0} \ \mathbb{E}[\pi(k_t, \Omega_t, k_{t-1})] + \beta \cdot V(\tilde{D}_{t+1}, k_t) \geq V(\tilde{D}_t, 0)$$

where $V(\tilde{D}_t, 0)$ is the value of remaining inactive, which still depends on $\tilde{D}_t$ because the stock continues to decay at rate ζ while the firm produces nothing. The fixed cost therefore governs whether a data-poor firm finds it worthwhile to operate at all.

The data feedback premium is the distinguishing feature of our model: it implies that optimal capital investment exceeds the level a myopic firm would choose. Its magnitude depends on the data mining efficiency z and on the shadow price λ, which we now characterize.

Applying the envelope theorem to (8) with respect to $\tilde{D}_t$, and using $\partial \Omega_t / \partial \tilde{D}_t = 1/\sigma_\varepsilon^2$ from (7′) together with $\partial \tilde{D}_{t+1} / \partial \tilde{D}_t = \zeta$:

$$\lambda_t = \frac{P}{\sigma_\varepsilon^2} \cdot \frac{\partial \mathbb{E}[A_t \mid \Omega_t]}{\partial \Omega_t} \cdot k_t^\alpha + \beta \zeta \cdot \lambda_{t+1} \tag{12}$$

The first term is the flow benefit: a marginal unit of effective data raises precision by

$1/\sigma_\varepsilon^2$, which raises the information-adjusted profit coefficient by $\partial\mathbb{E}[A]/\partial\Omega$, applied to the current capital stock. The second term is the continuation value: a share ζ of the unit remains informative next period. Iterating forward and imposing a constant capital stock gives the closed form

$$\lambda^* = \frac{P}{\sigma_\varepsilon^2} \cdot \frac{\partial\mathbb{E}[A \mid \Omega^*]}{\partial\Omega} \cdot \frac{(k^*)^\alpha}{1-\beta\zeta} \tag{13}$$

so that $1/(1-\beta\zeta)$ is the *informational annuity factor*: the present value of the stream of forecasting benefits that a unit of data delivers over its informational lifetime. This single object makes transparent where the strength of the feedback loop comes from. It is increasing in ζ, hence increasing in ρ and decreasing in δ: data is a more valuable asset in industries where the environment changes slowly and where records survive.

Two properties of (13) explain the quantitative results of Section 4. First, $\partial\mathbb{E}[A \mid \Omega]/\partial\Omega = \bar{A}/(1+\Omega)^2$ declines rapidly in precision, so a firm that is already well informed assigns little value to additional data. Second, λ^* is proportional to $(k^*)^\alpha$, so large firms value data more—the feedback loop is self-reinforcing in scale as well as in information.

3.6 Steady State and Uniqueness

We define a steady state as a situation in which all endogenous variables are constant over time: $k_t = k^*$, $\tilde{D}_t = \tilde{D}^*$, $\Omega_t = \Omega^*$ and $\lambda_t = \lambda^*$ for all t.

Imposing time invariance on the accounting stock (6) gives:

$$D^* = \frac{z \cdot (k^*)^\alpha}{\delta} \tag{14}$$

This is intuitive: in steady state, new data generation exactly offsets depreciation. A firm with higher data mining efficiency z or lower depreciation δ maintains a larger equilibrium data stock for any given capital level.

From the precision equation (7′), evaluated at the stationary stock:

$$\tilde{D}^* = \frac{z(k^*)^\alpha}{1-\zeta}, \qquad \Omega^* = \frac{1}{\sigma_a^2} + \frac{\tilde{D}^*}{\sigma_\varepsilon^2} = \frac{1-\rho^2}{\sigma_\theta^2} + \frac{z(k^*)^\alpha}{\sigma_\varepsilon^2\left[1-(1-\delta)\rho^2\right]} \tag{15}$$

Equation (15) is the steady-state information technology, and it is worth comparing the two stocks it involves. The accounting stock is $D^* = z(k^*)^\alpha/\delta$, while the informative stock is $\tilde{D}^* = z(k^*)^\alpha/(1-\zeta)$. Since $1-\zeta > \delta$ whenever $\rho < 1$, we have $\tilde{D}^* < D^*$: the firm always holds more data than is informative, and the wedge $D^*/\tilde{D}^* = (1-\zeta)/\delta$ widens as productivity becomes less persistent. At our calibration this ratio is 2.11, meaning that a firm behaving fully optimally already carries roughly 111% more data on its books than contributes to its forecasts. This is a useful observation for interpreting dark-data statistics: a large gap between stored and useful data is not by itself evidence of hoarding. Hoarding must be measured against the optimum, not against the informative stock.

Several further observations follow. First, Ω^* is increasing in k^* and in z, and decreasing in δ. Second, precision cannot fall below the prior floor $(1-\rho^2)/\sigma_\theta^2$ regardless of how data-poor the firm is, which rules out the degenerate zero-information outcome. Third, Ω^* is linear rather than concave in the stock; the diminishing returns that discipline the feedback loop operate instead through $\partial\mathbb{E}[A \mid \Omega]/\partial\Omega = \bar{A}/(1+\Omega)^2$, which is where they belong economically.

Combining (13) with the steady-state Euler equation yields the core condition determining the steady-state capital stock:

$$r = \alpha(k^*)^{\alpha-1} \cdot \left[P \cdot \mathbb{E}[A \mid \Omega^*] + \beta z \lambda^*\right] \tag{16}$$

Define the feedback multiplier $\mathcal{M}(k^*)$ as:

$$\mathcal{M}(k^*) \equiv 1 + \frac{\beta z \lambda^*}{P \cdot \mathbb{E}[A \mid \Omega(k^*)]} = 1 + \frac{\beta z}{\sigma_\varepsilon^2(1-\beta\zeta)} \cdot \frac{\bar{A}/(1+\Omega^*)^2}{\mathbb{E}[A \mid \Omega^*]} \cdot (k^*)^\alpha \; > \; 1$$

Then the Euler equation can be written compactly as:

$$r = \alpha(k^*)^{\alpha-1} \cdot P \cdot \mathbb{E}[A \mid \Omega(k^*)] \cdot \mathcal{M}(k^*) \tag{17}$$

The feedback multiplier captures the amplification of the marginal revenue product of capital due to the data generation channel. Its closed form makes the determinants explicit: amplification rises with data-mining efficiency z and with the informational annuity $1/(1-\beta\zeta)$, and falls with precision through $(1+\Omega^*)^{-2}$. These forces pull in opposite directions as z rises, since more efficient data extraction raises both the annuity's payoff and the precision that erodes it, which is why $\mathcal{M}$ is non-monotone in z in Section 4. With Cobb-Douglas production the marginal product of capital diverges as $k^* \to 0$ while $\mathcal{M}$ remains bounded, so (17) admits a unique interior steady state. Appendix B proves this for $\alpha < 1/2$, which covers our calibration $\alpha = 0.33$, and shows by counterexample that uniqueness can fail for $\alpha \geq 1/2$ when the prior precision floor is small.

The role of the fixed cost in determining viability. The Euler equation (17) identifies the unique interior steady state—the point where the marginal benefit of capital equals its marginal cost. However, this candidate is economically viable only if the firm finds it optimal to operate rather than remain inactive. The fixed operating cost F creates a participation constraint:

$$P \cdot \mathbb{E}[A \mid \Omega^*] \cdot (k^*)^{\alpha} - r \cdot k^* - F \geq 0 \tag{17b}$$

When F exceeds the maximum attainable operating surplus, the firm optimally chooses $k^* = 0$ (inactivity). For the baseline calibration the left-hand side of (17b) equals 5.46 at the interior steady state—that is, operating profit net of the fixed cost is comfortably positive—so the participation constraint is slack and the unique interior steady state is the global optimum.

Data poverty and the transition to optimality. Data poverty in our framework does not refer to a distinct steady state but to a transitory condition: a firm that begins with a low

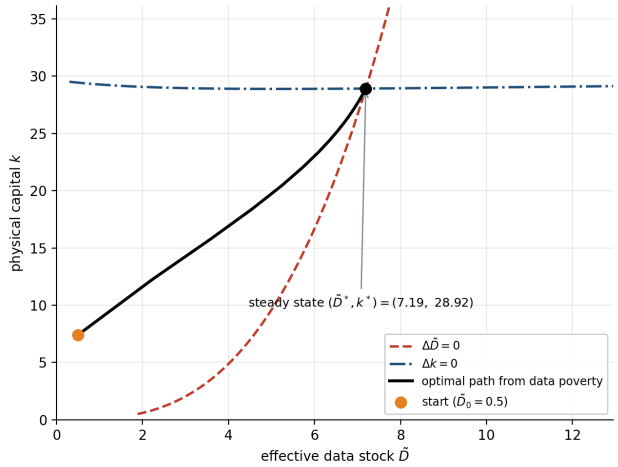


Figure 1: Phase diagram in $(\tilde{D}, k)$ space. The $\Delta\tilde{D} = 0$ locus gives combinations at which the effective data stock is stationary, $\tilde{D} = zk^{\alpha}/(1-\zeta)$; the $\Delta k = 0$ locus gives combinations satisfying the steady-state Euler equation (17). The unique intersection is the steady state $(\tilde{D}^*, k^*) = (7.19, 28.92)$. The solid curve traces the optimal transition path from data poverty, $\tilde{D}_0 = 0.5$. *Notes:* computed under the baseline calibration of Table 1 by direct maximization of the discounted profit stream over a 160-period horizon; see Appendix F for the numerical method.

data stock and hence with precision near the prior floor. The data feedback loop implies that such a firm faces a dual challenge—its low precision suppresses the expected return to capital, yet it must invest in capital to generate the data that will improve its precision. Capital adjustment costs prevent instantaneous jumps to the steady-state capital level, generating a meaningful transition path that we characterize quantitatively in Section 4.

3.7 Phase Diagram and Transitional Dynamics

Figure 1 displays the phase diagram in $(\tilde{D}, k)$ space. The $\Delta\tilde{D} = 0$ locus gives the combinations of effective data stock and capital at which the data stock is stationary; the $\Delta k = 0$ locus gives the combinations satisfying the steady-state Euler equation. The unique intersection of these loci is the steady state.

The solid curve traces the transition path from an initial condition of data poverty, $\tilde{D}_0 = 0.5$, with Ω_0 correspondingly near the prior floor. Three features characterize the transition.

Immediate investment. The firm does not wait and see. It invests substantially from the first period—initial capital $k_0 = 7.4$ is 26% of its steady-state value—because the anticipated future informational benefit justifies immediate capital deployment despite adjustment costs and despite the low current return.

Gradual data accumulation. While capital adjusts quickly, the effective stock accumulates gradually as a by-product of production, reaching 95% of its steady-state level only after 11 periods. The asymmetry between the two speeds is the substantive content of the transition: the firm front-loads the input it controls directly in order to acquire the asset it can obtain only as a by-product. This is the sense in which a forward-looking firm invests *through* data poverty.

Smooth convergence. The transition exhibits no overshooting: capital approaches the steady state from below, exceeding it by at most 0.004% along the computed path, which is within numerical tolerance. The half-life of convergence—the time required to close half the initial gap in the capital stock—is 3 periods.

Convergence from *above*, which one might expect to be slower because a firm cannot actively destroy information, is in fact no slower here. If the firm begins with an excess stock, the decay of $\tilde{D}$ is governed by ζ, so the mechanical half-life is $\ln(0.5)/\ln\zeta \approx 1.3$ periods, and we find a comparable figure along the optimal path. There is no meaningful adjustment asymmetry in this model, and hence no basis for the claim that data over-accumulation is intrinsically persistent. Persistence of hoarding, where it arises, comes from the incentives analyzed in Section 5 rather than from the physics of the stock.

4 Quantitative Analysis

4.1 Calibration and Functional Forms

As established in (9a), the information-adjusted profit coefficient takes the form

$$\mathbb{E}[A \mid \Omega] = \bar{A} \cdot \frac{\Omega}{1+\Omega} \tag{18}$$

with derivative

$$\frac{\partial \mathbb{E}[A \mid \Omega]}{\partial \Omega} = \frac{\bar{A}}{(1+\Omega)^2} > 0. \tag{19}$$

The concavity captures diminishing returns to information: initial improvements in precision yield large gains, but as precision rises further improvements contribute progressively less. The rate of decline, $(1+\Omega)^{-2}$, is steep, and this turns out to be the single most important quantitative feature of the model, since it bounds how large the feedback loop can be at any calibration that places the firm in the well-informed region.

Baseline parameter values are organized into three groups: parameters calibrated from the empirical literature (Panel A), normalization parameters fixed by model structure (Panel B), and standard macro parameters (Panel C).

Table 1: Baseline Parameter Values

Parameter	Description	Value
Panel A: Calibrated Parameters		
ρ	TFP persistence	0.85
σ_θ^2	TFP innovation variance	0.04
δ	Data depreciation rate	0.20
z	Data mining efficiency	1.00
F	Fixed operating cost	0.15
γ	Capital adjustment cost	0.05
Panel B: Normalization Parameters		
P	Output price	1.00
$\bar{A}$	Forecast-loss curvature	3.00
σ_ε^2	Signal noise variance	1.00
Panel C: Standard Macro Parameters		
α	Capital output elasticity	0.33
β	Discount factor	0.96
r	Rental rate of capital	0.10

TFP process parameters (ρ and σ_θ^2). Our baseline values $\rho = 0.85$ and $\sigma_\theta^2 = 0.04$

are drawn from the empirical literature on firm-level TFP dynamics. Brandt et al. (2012) report TFP persistence of 0.75–0.90 for Chinese manufacturing firms; the within-firm residual dispersion in their estimates implies $\sigma_\theta^2 \approx 0.04$, a magnitude consistent with the within-firm productivity dispersion documented for US manufacturers by Foster et al. (2008) and De Loecker (2011) and with the estimates surveyed in Ackerberg et al. (2015). We note that none of these papers reports an AR(1) persistence coefficient and an innovation variance as headline estimates, so our values should be read as calibrated to be consistent with the dispersion these studies document rather than as taken directly from them. Because ρ enters the informational retention coefficient ζ directly, it is no longer an incidental parameter of the productivity process but a determinant of the value of data, and we report comparative statics with respect to it below.

Calibration of δ. Our baseline physical depreciation rate $\delta = 0.20$ sits in the middle of the available range. Xu et al. (2025) construct data-capital stocks for China using the perpetual inventory method and adopt a depreciation rate near 0.31 for processed operational data. We treat $\delta = 0.20$ as a central value below that figure and report results across $\delta \in [0.02, 0.50]$, which spans it and values well below it. The reader should note that what matters for firm behavior is not δ but $\zeta = (1-\delta)\rho^2$, so the effective range of informational retention implied by these bounds is narrower than the range of δ itself.

Calibration of z and the identification of the loop's strength. The baseline $z = 1.00$ is a normalization in the sense that only z/σ_ε^2 enters the transition equation. It is important to be clear, however, that this normalization is not innocuous for the magnitude of the feedback loop. Precision enters profits through $\Omega/(1+\Omega)$, whose unit scale is fixed; setting $\sigma_\varepsilon^2 = 1$ therefore pins how informative a unit of data is relative to that scale, and hence pins $\mathcal{M}$. Put differently, $\bar{A}$ and σ_ε^2 are separately identified only given a target for the value of information, and we do not have one. We consequently regard the level of $\mathcal{M}$ reported below as a calibrated implication rather than an estimate, and we flag as the most valuable extension of this paper the use of a direct estimate—for instance the elasticity

of firm productivity with respect to data-management capability in T. Li et al. (2020)—to discipline this scale. The hoarding results of Section 5, being ratios of data stocks at a common calibration, are unaffected by this indeterminacy. Appendix G bounds the problem: varying $\bar{A} \in [1, 10]$ and $\sigma_\varepsilon^2 \in [0.25, 16]$ moves the amplification between 0.4% and 3.1%, and varying σ_θ^2 between 0.01 and 0.16 moves it from 0.5% to 7.4%, so the conclusion that the level effect is modest does not depend on the normalization even though its exact size does.

Calibration of F and γ. The fixed operating cost $F = 0.15$ corresponds to 1.8% of steady-state output at the baseline calibration, in the range of IT maintenance and data-compliance costs as a share of revenue in data-intensive industries (Gartner, Inc. 2023). The capital adjustment cost $\gamma = 0.05$ is set to generate a transition half-life of 3 periods from data poverty.

The capital output elasticity $\alpha = 0.33$ follows Gollin (2002), and β and r take conventional annual values. The parameters governing the strength of the data feedback loop are δ, z and ρ, which we vary across $\delta \in [0.02, 0.50]$, $z \in [0.10, 5.00]$ and $\rho \in [0.60, 0.95]$.

4.2 Steady State and Comparative Statics

At the baseline calibration the unique interior steady state is $k^* = 28.92$, with an effective data stock $\tilde{D}^* = 7.19$, an accounting stock $D^* = 15.18$, precision $\Omega^* = 14.13$ and a feedback multiplier $\mathcal{M} = 1.031$. The loop therefore amplifies the marginal revenue product of capital by 3.1%.

This magnitude deserves comment, since it is considerably smaller than the discussion of data feedback loops in the literature might lead one to expect. The reason is visible in the closed form for $\mathcal{M}$: the informational annuity $1/(1 - \beta\zeta) = 2.25$ is substantial, but it multiplies $\partial \mathbb{E}[A \mid \Omega]/\partial\Omega = \bar{A}/(1 + \Omega^*)^2$, which at $\Omega^* = 14.13$ is only 0.0131. A firm that already forecasts its productivity reasonably well gains little from further precision, and at any calibration placing the firm in that region the level effect must be small. We regard this as a substantive finding rather than a disappointing one: it relocates the interesting

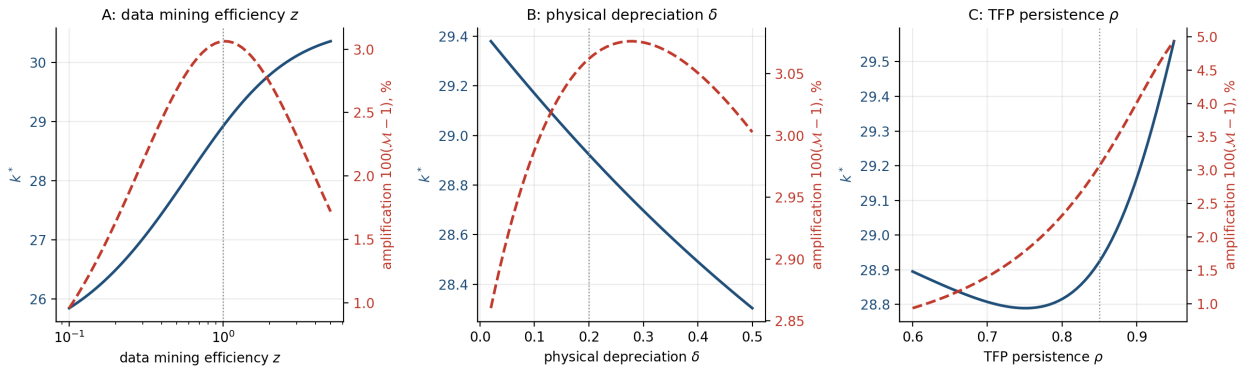


Figure 2: Comparative statics of the steady state. Panel A: steady-state capital k^* and the feedback multiplier $\mathcal{M}$ against data mining efficiency z. Panel B: against the physical depreciation rate δ. Panel C: against TFP persistence ρ, which enters through the informational retention coefficient $\zeta = (1-\delta)\rho^2$. *Notes:* each point solves the steady-state Euler equation (17) exactly; all other parameters held at their Table 1 values.

economics of data accumulation away from the level of the optimum and towards the wedges that push firms away from it.

Figure 2 collects the comparative statics.

Data mining efficiency z. k^* is increasing and concave in z, rising from 25.85 at $z = 0.10$ to 28.92 at the baseline and 30.36 at $z = 5.00$. The multiplier, by contrast, is non-monotone: it rises from 1.010 to 1.031 and then falls to 1.017. The non-monotonicity is the central mechanism of the model working in both directions at once. Higher z raises the quantity of data a given scale generates, which raises the value of the loop; but it also raises equilibrium precision, which lowers the marginal value of the next unit of information. Beyond $z \approx 1$ the second force dominates. An immediate implication is that firms with the best data infrastructure have the *weakest* marginal incentive to expand scale for informational reasons—the loop is a mechanism for escaping data poverty rather than for rewarding data abundance.

Physical depreciation rate δ. k^* is decreasing in δ, falling from 29.38 at $\delta = 0.02$ to 28.30 at $\delta = 0.50$, and the accounting stock falls far more sharply, from 152.6 to 6.0, because δ enters the stock directly. The effect on k^* is modest in proportional terms, for the reason the decomposition in (7a) makes plain: δ operates on informational retention only through

$\delta\rho^2$, and with $\rho^2 = 0.72$ a large change in δ produces a smaller change in ζ. Retention mandates are therefore a blunt instrument in this model, a conclusion we return to in the policy discussion.

TFP persistence ρ. This comparative static is new relative to the previous literature and follows directly from our treatment of informational decay. The multiplier rises steeply in ρ, from 1.009 at $\rho = 0.60$ to 1.049 at $\rho = 0.95$—a fivefold increase in the strength of the feedback loop across a plausible range of persistence. Data is a durable asset precisely when the environment it describes is stable. The prediction is testable: holding data-mining technology fixed, the complementarity between scale and data investment should be strongest in industries with persistent productivity fundamentals, and weakest in fast-moving consumer-facing sectors. It also cautions against the common presumption that data matters most in turbulent, rapidly innovating industries; in this framework the opposite holds.

Fixed operating cost F and adjustment cost γ. Increases in F reduce steady-state capital modestly, as a larger share of operating surplus must cover the fixed cost. Increases in γ do not affect the steady state, where $k_t = k_{t-1}$ and adjustment costs vanish, but slow the transition considerably.

4.3 Dynamic Analysis

We simulate the firm's optimal transition path starting from data poverty: $\tilde{D}_0 = 0.5$, so that Ω_0 sits just above the prior floor, with $k_{-1} = 0$. Capital adjustment costs ($\gamma = 0.05$) prevent instantaneous adjustment and generate a meaningful transition path. Appendix F describes the numerical method and reports accuracy diagnostics; the maximum relative Euler-equation residual along the reported path is 9.0×10^{-4}.

Three phases characterize the path. In the *data poverty* phase the firm deploys substantial capital immediately—$k_0 = 7.4$, or 26% of the steady-state level—investing through production in order to generate data, with the capital growth rate peaking at 65% in the first period. In the *accumulation* phase the arrival of data validates that early investment, sus-

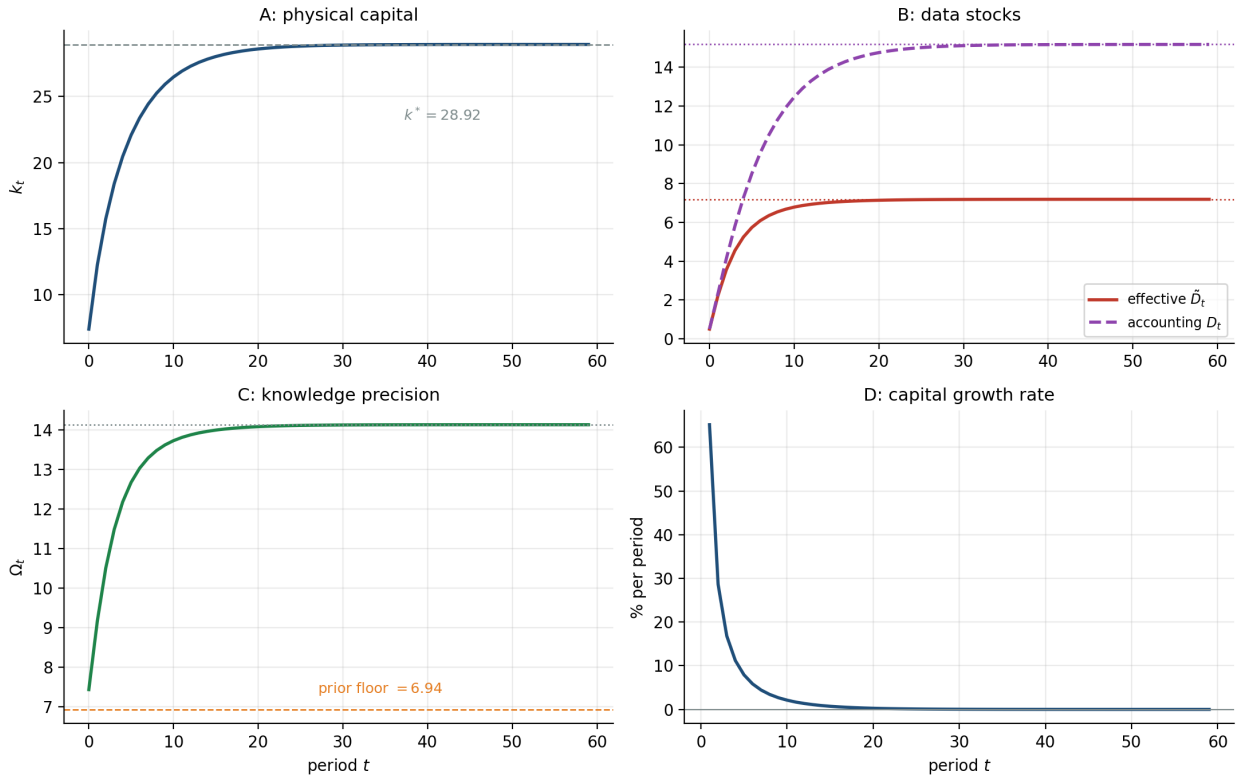


Figure 3: Optimal transition from data poverty. Panel A: capital k_t with the steady state $k^* = 28.92$ marked. Panel B: effective data stock $\tilde{D}_t$ and accounting stock D_t. Panel C: precision Ω_t, with the prior floor $(1-\rho^2)/\sigma_\theta^2 = 6.94$ marked. Panel D: growth rate of capital. *Notes:* initial condition $\tilde{D}_0 = 0.5$, $k_{-1} = 0$; baseline calibration of Table 1.

taining growth at a declining rate. In the *convergence* phase diminishing returns to precision slow growth below 1% per period. The half-life of convergence is 3 periods and the effective stock reaches 95% of its steady-state value after 11 periods. Capital approaches the steady state from below throughout: along the computed path $\max_t k_t/k^* = 1 + 0.004 \times 10^{-2}$, i.e. an excess of 0.004% at most, which is within the tolerance of the numerical solution reported in Appendix F.

The economically interesting feature of this path is not its speed as such but the ordering of the two adjustments. Capital, which the firm controls directly, moves first and fast; data, which arrives only as a by-product, follows slowly. A firm emerging from data poverty must therefore accept a period in which its capital stock is high relative to what its current information justifies—it is over-invested by a myopic standard—and it does so willingly because the informational return validates the position later. This is the precise sense in which forward-looking firms invest through data poverty, and it is a pattern that cannot arise in models where data is purchased as an input rather than generated as a by-product.

Figure 4 turns to the reverse scenario, a firm starting with $\tilde{D}_0 = 1.5 \times \tilde{D}^*$, as might arise through a data-acquiring merger. The firm does *not* cut capital below the steady state. Because the excess stock raises precision above Ω^*, the information-adjusted return to capital is temporarily higher, so the firm holds capital marginally *above* k^*—by at most 0.15%—and simply allows informational decay to bring the stock down. This is the opposite of what one might expect; the firm has no incentive to contract, because the surplus information is a free, if transient, benefit. We emphasize that the adjustment is *not* appreciably slower than convergence from below: the stock decays at rate $\zeta = 0.578$, implying a mechanical half-life of $\ln(0.5)/\ln\zeta \approx 1.3$ periods, and the optimal path delivers a half-life of 2 periods, against 3 periods for convergence from below. There is thus no adjustment asymmetry from which one could infer that data over-accumulation is intrinsically persistent. Nothing in the technology of data accumulation makes excess stocks durable; if hoarding persists in practice, the explanation must lie in the incentives of Section 5, not in the difficulty of letting data

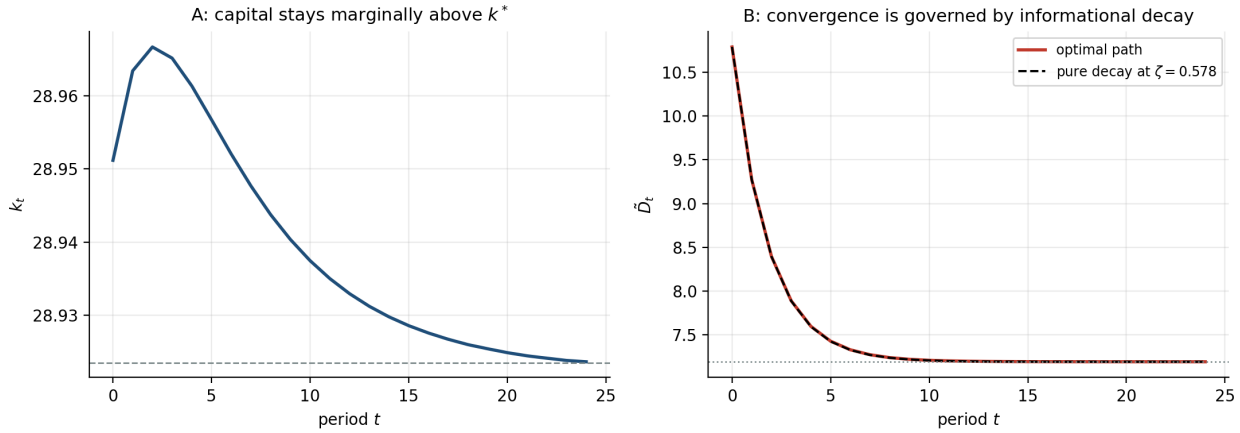


Figure 4: Reverse dynamics from an excess data stock. Panel A: capital from an initial condition $\tilde{D}_0 = 1.5\tilde{D}^*$; note that capital sits slightly *above* k^* while the excess stock persists. Panel B: convergence of the effective stock, compared with pure geometric decay at rate ζ. *Notes:* the near-coincidence of the two lines in Panel B shows that adjustment from above is governed by informational decay rather than by any optimizing friction.

decay.

4.4 The Cost of Over-Accumulation and Policy

Section 5 introduces three mechanisms—competitive escalation, institutional incentives and precautionary accumulation—that drive firms beyond the benchmark just characterized. We preview here what such deviations cost and which policy instruments address them.

A clarification of terminology is needed first, because the natural language here is misleading. Our model contains a firm and no consumer, so it cannot support a statement about welfare, and quantities such as a consumption-equivalent variation are not defined in it. What we can compute is the loss of *firm value*: the present discounted value of profits at a distorted capital stock relative to its maximized value. We report that, and label it as such.

The result is stark. Comparing the maximized commitment value of the firm with its value at the Nash capital stock, the loss is 0.0003% of firm value at $\lambda = 1.0$, where the data stock already exceeds the optimum by about 101%, and still only 0.0013% at $\lambda = 3.0$, where the excess is roughly 302%. The loss rises with λ, but from a base so close to zero

that it remains economically invisible throughout. This is the envelope theorem at work: the value function is flat in a neighbourhood of its maximum, so even large deviations in the data stock translate into negligible changes in profit. It is worth stating plainly what this implies. Holding the data stock above its optimum, taken by itself, is very nearly costless. Any account of why hoarding matters must therefore locate the cost somewhere other than in the misallocation of the hoarding firm's capital. Two candidates remain: the duplicated infrastructure below, and the resources spent maintaining the exclusive holdings that produce the excess stock, which Appendix E shows to be of comparable size.

Duplicated infrastructure. We locate it in the fixed cost. Each firm maintaining a data estate pays F per period for servers, engineering staff and compliance, and when N firms accumulate overlapping records of the same underlying economic activity, a share of that expenditure is incurred repeatedly for substantially the same information. Let ω denote the overlapping share of the industry's data. If the overlapping data is held by all N firms, the duplicated component of infrastructure cost has present value $(N-1)\,\omega\,F/(1-\beta)$; if it is held by only $\bar{m} < N$ firms, $N-1$ is replaced by $N(1-1/\bar{m})$ and the estimate falls (Appendix E.5). Taking the former, which for $N = 10$ and $\omega = 0.5$ amounts to 1.2% of industry value. Against a misallocation cost of 0.0013%, the duplication cost is larger by a factor of about 974, and by at least 108 across the grid of Appendix G: two to three orders of magnitude. Table 2 sets this beside the cost of the rivalry that produces the excess stock, all on one denominator.

This comparison is, in our view, the paper's most consequential quantitative statement, and it reframes the policy problem. The case for intervening in data accumulation does not rest on firms mismanaging their own production—they do not, to any measurable degree. It rests on an industry structure in which the same information is paid for many times over. Two caveats apply. The estimate is linear in the assumed overlap ω, which we do not measure and which is the obvious target for empirical work. Appendix G varies N and ω: the factor of 974 holds at the baseline, and over the whole grid $N \in [2, 50]$, $\omega \in [0.10, 0.75]$ the

duplication cost remains at least 108 times the misallocation cost, so the ranking is robust and the magnitude is best stated as two to three orders. And because our framework is partial equilibrium with a single firm, we cannot speak to the further effects of hoarding on concentration, entry barriers or the distribution of informational advantage (Eeckhout and Veldkamp 2022; Acemoglu et al. 2022); these plausibly matter and are outside what we can quantify here.

Table 2: Three costs of over-accumulation on a common denominator

Cost	Percent of value	Multiple of misallocation
Capital misallocation ($\lambda = 3$, stock +302%)	0.0013	1
Rivalry in maintenance (Appendix E, $\lambda = 0.25$)	0.24	190
Rivalry in maintenance (Appendix E, $\lambda = 0.48$)	1.01	794
Duplicated infrastructure ($N = 10$, $\omega = 0.5$)	1.2	977

Notes: Every entry is a percentage of the value of the non-competitive profit stream, the present value of a constant flow, so that flow ratios and value ratios coincide. The denominator is net of the baseline maintenance expenditure, which is 0.3% of operating surplus and so changes no entry materially. The misallocation entry is evaluated at $\lambda = 3$, where it is largest; at the intensities of the other rows it is smaller still, so the multiples are lower bounds.

Policy instruments. Figure 5 collects these results. Three instruments map onto the three mechanisms, and our corrected results reverse the earlier ranking of their effectiveness.

First, *interoperability requirements* reduce the private advantage of exclusive data holdings, lowering λ. Because the competitive channel is the most powerful per unit of intensity, this instrument has the greatest leverage: each 0.1 of λ removed takes roughly ten per cent off the data stock. We emphasise, however, what Section 5.4 establishes about the size of the prize. The available evidence implies $\hat{\lambda} = 0.063$ and hence excess accumulation of about 6.4%, not the 50.3% obtained at an assumed $\lambda = 0.50$, and the estimate is not robustly signed. The case for interoperability therefore cannot be built on dark-data statistics, which we show to be uninformative about hoarding; it has to rest on direct evidence about the exclusivity of data holdings and on the duplication cost below.

Second, *accounting standards* that distinguish productive data from dark data reduce the manager's ability to use stock size as a signal, addressing the μ channel and a gap of 3.5% at

$\mu = 0.05$. This is a smaller prize than interoperability but the instrument is unusually well targeted, since the distortion exists only because the stock is observable while its productive contribution is not.

Third, *retention mandates* that force deletion raise δ. Our analysis is not encouraging about these. Raising δ from 0.20 to the 0.31 of Xu et al. (2025) lowers the accounting stock substantially, but it acts on informational retention only through $\delta\rho^2$, so its effect on the firm's actual behavior is muted. More awkwardly, since only $\zeta = (1-\delta)\rho^2$ enters the firm's problem, a deletion mandate is informationally equivalent to making the economic environment more turbulent. It reduces measured hoarding partly by destroying information that was socially valuable, which is a poor trade when the real cost being targeted is duplicated infrastructure rather than the stock itself.

The instruments are complementary in that they address distinct wedges, but they are not of comparable value. If the duplication mechanism is the operative social cost, the first-best response is neither deletion nor disclosure reform but arrangements that let one collection serve many users—which is to say interoperability and data sharing, precisely the instrument our competitive results favor.

5 Data Hoarding: Quantitative Mechanisms

The benchmark model of Sections 3–4 characterizes the firm's optimal data investment: the unique steady-state accounting stock $D^* = 15.18$ is the level justified by data's contribution to productivity forecasts. Industry evidence suggests that many firms accumulate far beyond any plausible productive use. This section presents three extensions that generate systematic deviations from the benchmark, and we present them in descending order of magnitude, since that ordering is itself one of the paper's results.

Figure 6 summarizes the three mechanisms. Two broader points emerge before the details. First, data hoarding does not require abandoning the rational-agent framework: each

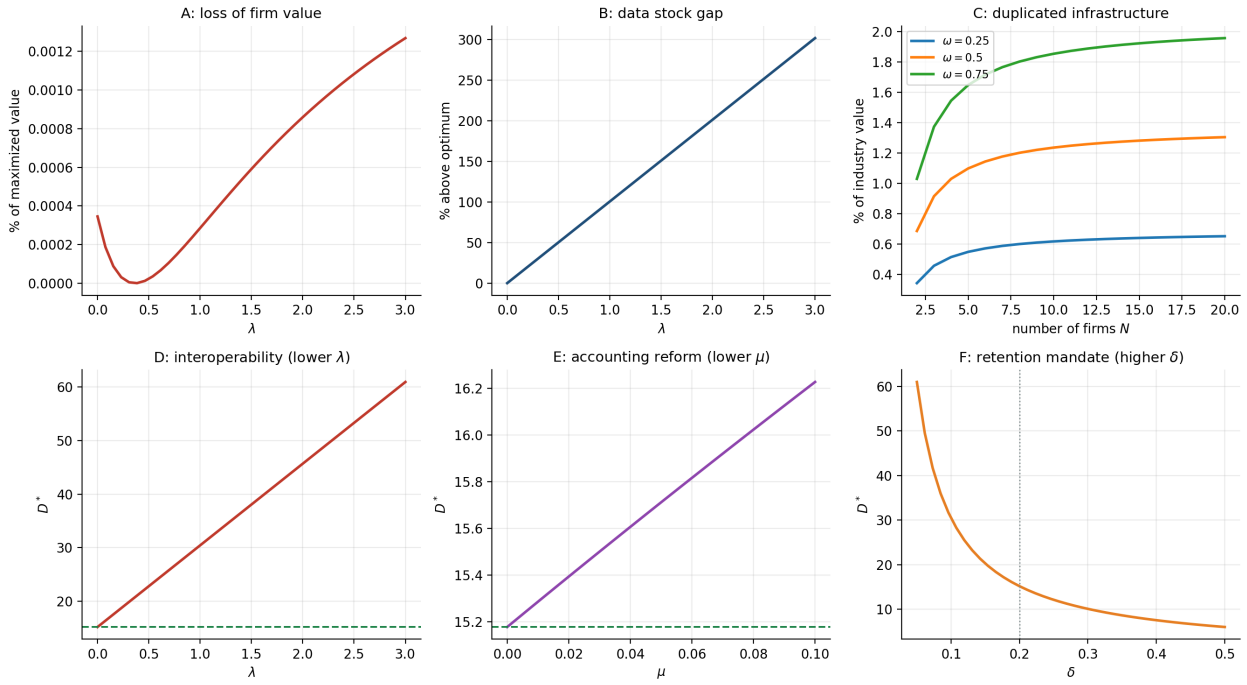


Figure 5: Cost of over-accumulation and policy counterfactuals. Panel A: loss of firm value as a share of maximized value, against competitive intensity λ. Panel B: the corresponding data-stock gap. Panel C: duplicated infrastructure cost as a share of industry value, against the number of firms N and overlap ω. Panels D–F: policy counterfactuals for interoperability (λ), accounting reform (μ) and retention mandates (δ). *Notes:* Panels A and B share a horizontal axis; note the difference in vertical scale between Panel A (ten-thousandths of a percent) and Panel C (percent): this gap of two to three orders of magnitude is the central point of this section.

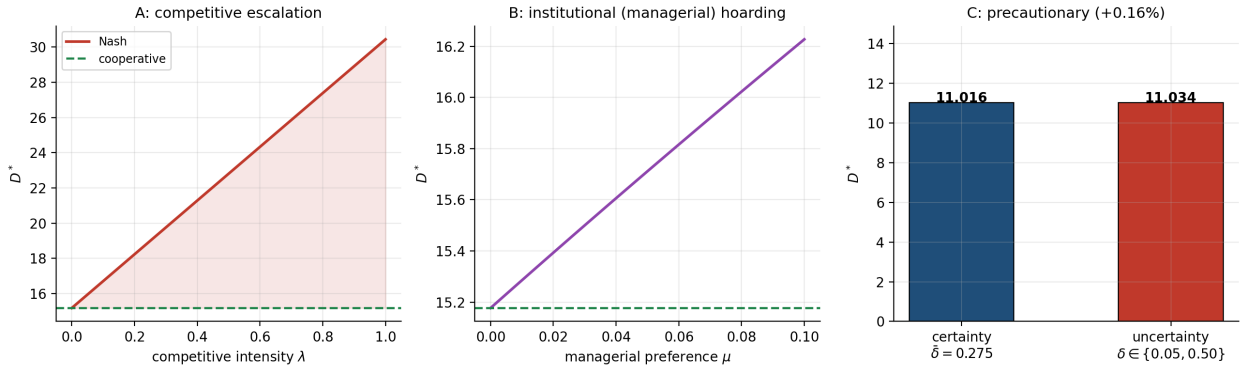


Figure 6: The three hoarding mechanisms. Panel A: competitive escalation—steady-state accounting stock against competitive intensity λ, with the cooperative benchmark $D^* = 15.18$ marked. Panel B: institutional hoarding against managerial preference μ. Panel C: precautionary hoarding, comparing the stock under certainty at the ergodic mean depreciation rate with the stock under depreciation uncertainty. *Notes:* all panels report the accounting stock D^*, evaluated at each mechanism's own effective depreciation rate; panels share no common vertical scale.

mechanism operates through a standard economic channel—Nash equilibrium, managerial signalling, and precautionary saving—applied to data as an asset. Second, the three are of very unequal importance. At the intensities conventionally used in this literature, competitive escalation raises the data stock by 30.2% and managerial signalling by 3.5%, both measured against the common baseline $D^* = 15.18$; the precautionary motive adds 0.16%, measured against its own certainty benchmark, since a change in the depreciation rate moves the level of the stock and makes a direct comparison with the baseline meaningless. We report these magnitudes because they show which channels are capable of mattering, but the reader should treat them as conditional on assumed intensities until Section 5.4, which disciplines the competitive parameter against observed data and finds a considerably smaller value. Each mechanism's stock is evaluated at the depreciation rate the firm actually faces under that mechanism, which is what makes the magnitudes comparable.

5.1 Competitive Hoarding

We model competition as an N-firm symmetric Nash equilibrium in which each firm's competitive position improves with its data stock relative to rivals. Firms compete to keep their data current and exclusive, which from the perspective of each firm's private optimization lowers the effective rate at which data ceases to be worth holding:

$$\delta^{\text{Nash}} = \frac{\delta}{1+\lambda},$$

where $\lambda \geq 0$ indexes competitive intensity. Appendix E derives it from an underlying game. The cooperative benchmark retains δ. Note that because the retention coefficient is $\zeta = (1-\delta)\rho^2$, a lower δ^{Nash} raises ζ and hence the informational annuity, so competition operates on the firm's problem through exactly the channel that makes data valuable in the first place.

The effect is large at the intensities this literature conventionally uses, though Section 5.4

will show that those intensities are hard to reconcile with the observed dark-data share. At $\lambda = 0.30$ the Nash steady state is $k^* = 29.03$ with $D^* = 19.75$, a gap of 30.2% above the cooperative stock; at $\lambda = 0.50$ the Nash steady state is $k^* = 29.09$ with $D^* = 22.81$, a gap of 50.3%, and at $\lambda = 1.00$ the gap exceeds 100%. The mechanism's power comes from the fact that it acts on the stock directly, through δ^{Nash} in the denominator of $D^* = z(k^*)^\alpha/\delta^{\text{Nash}}$, rather than only through the capital choice. Capital itself barely moves—from 28.92 to 29.03—which is worth emphasizing because it means competitive data hoarding is nearly invisible in the capital-stock data one would normally use to detect over-investment. It shows up in the data estate, not on the balance sheet for physical capital. (The game of Appendix E implies a somewhat larger capital response, 2.2% at $\lambda = 0.25$, because rivalry also raises the shadow value of the accounting stock; that is still an order of magnitude smaller than the response of the stock itself.)

5.2 Institutional Hoarding

We next introduce a manager who derives direct utility from the size of the data stock, because data assets serve as a visible signal of digital transformation to boards and investors while their productive contribution is difficult to verify. This is the configuration studied by Bebchuk and Stole (1993) and Holmström (1999): investment is observable, its return is not, and the manager consequently over-invests to influence inference about his type. The manager's per-period payoff becomes $U = \pi + \mu \cdot D$, with $\mu \geq 0$ the strength of the preference.

Two features of this specification should be flagged. It is the *accounting* stock D that the manager values, not the effective stock $\tilde{D}$, which is the economically right distinction: boards observe terabytes, not forecast precision, and that gap between what is observable and what is productive is the whole source of the distortion. And we model μ as a reduced-form preference rather than deriving it from an explicit signalling equilibrium. This is a genuine limitation, and it means μ should be read as a summary of the strength of a career-concerns distortion rather than as a primitive.

The parameter μ does not enter the first-order condition for capital directly, since D_t is predetermined, but it raises the marginal value of the accounting stock by $\mu/[1-\beta(1-\delta)]$, which enters the data feedback premium and shifts the steady state upward. Raising μ from 0 to 0.05 increases the data stock from 15.18 to 15.71, a gap of 3.5%. The feedback multiplier of (17) is essentially unchanged at 1.031, as it must be: μ does not alter the informational value of data. What rises is the manager's own valuation of the accounting stock, $\mu/[1-\beta(1-\delta)] = 0.216$, which enters the investment decision alongside the informational premium. We keep the two separate to avoid presenting an agency wedge as part of the feedback loop. At $\mu = 0.10$ the stock reaches 16.23, a gap of 6.9%, with k^* rising to 35.43 and the managerial wedge to 0.431. The effect operates through capital—k^* rises substantially—which distinguishes it observationally from the competitive channel.

5.3 Precautionary Hoarding

Finally we introduce uncertainty about the depreciation rate. The firm believes δ follows a two-state Markov process with values $\delta \in \{0.05, 0.50\}$ and symmetric transition probability $\pi = 0.15$, an ergodic mean of 0.275, and prices data using the ergodic-average informational annuity $\mathbb{E}[1/(1-\beta\zeta(\delta))]$.

The mechanism is real but quantitatively negligible. Because the annuity factor is convex in ζ, Jensen's inequality implies that uncertainty raises the expected value of a unit of data—the low-depreciation state raises the annuity by more than the high-depreciation state lowers it—and the gap in the annuity is 10.9%. But this translates into a data-stock increase of only 0.16%, from 11.02 to 11.03, because the annuity enters the Euler equation only through the feedback premium, which is itself a 3.1% correction to the marginal revenue product. A large distortion to a small term remains small. The comparison is made against the certainty benchmark at the same ergodic mean depreciation rate, so that the precautionary premium is not conflated with the level effect of a change in δ.

We report this mechanism because it is theoretically instructive rather than because it

matters: it illustrates that precautionary motives in informational assets are second-order whenever the informational premium is itself second-order, which is the generic case in calibrations that place firms in the well-informed region.

5.4 Inverting the Dark-Data Share

The three mechanisms above share a weakness: each is indexed by a parameter—λ, μ, the support of δ—that we have set by assumption. The quantitative results are then only as credible as those assumptions, and a reader is entitled to object that a data-stock gap of 30.2% at $\lambda = 0.30$ is arithmetic rather than evidence. This subsection addresses the objection for the competitive channel by turning the exercise around: rather than assuming λ and computing a data stock, we ask what λ is required to reproduce the one quantitative fact the literature actually has, the observed share of enterprise data that is never put to use (Veritas Technologies 2020).

The model's counterpart of the dark-data share. The framework delivers two stocks, and their ratio is exactly the object a dark-data survey attempts to measure. From (14) and the definition of the effective stock,

$$s \;\equiv\; \frac{\tilde{D}^*}{D^*} \;=\; \frac{z(k^*)^\alpha/(1-\zeta)}{z(k^*)^\alpha/\delta} \;=\; \frac{\delta}{(1-\rho^2)+\delta\rho^2}, \tag{20}$$

the share of the stored stock whose informational content is actually realized. Its complement $1-s$ is the share that is retained but informationally inert: data the firm holds and pays to hold, which no longer speaks to the state it is trying to forecast. We take this as the model counterpart of "dark data" and return below to how imperfect the correspondence is.

An identification result. Equation (20) has a property worth stating separately, because it determines what this moment can and cannot do. Both stocks are proportional to $(k^*)^\alpha$, so *the scale of the firm cancels.* The dark-data share therefore does not depend on z, on $\bar{A}$, on P, on F, on γ, or on k^* itself. In particular it does not depend on μ: the managerial

channel of Section 5.2 operates purely by raising the scale of operations, and so raises the informative and the accounting stock in exactly the same proportion. The precautionary channel is likewise silent in (20) except through $\bar{\delta}$.

Proposition 1 *The dark-data share is a function of ρ and the effective depreciation rate alone. It is therefore invariant to the institutional channel, which operates purely through scale; and, conditional on ρ and the physical depreciation rate, it identifies the effective depreciation rate and hence the competitive wedge.*

Two things this proposition does not say. It does not say that the moment identifies λ unconditionally: λ, the physical δ and the precautionary $\bar{\delta}$ all enter only through the single effective rate, so one moment pins down that rate and no more. Disentangling them requires ρ and δ to be known from elsewhere, which is exactly the assumption the sensitivity analysis below stresses. Nor does it say the moment is uninformative about the precautionary channel: that channel shifts the depreciation rate to $\bar{\delta}$ and so moves the dark share substantially, from 52.6% to 42.3%. It is *confounded* with the competitive channel in this moment, not absent from it. Only the invariance to μ is clean, and it is clean exactly.

This is a useful piece of news for empirical work on data assets, and to our knowledge it has not been noted. A researcher who observes only the utilization rate of a data estate can learn about the strategic motive for retention, and can learn nothing whatever about managerial signalling—not because the signalling channel is small, but because it is orthogonal to this particular moment. Distinguishing it requires a moment that speaks to scale, such as the level of data expenditure relative to output.

The inversion. Substituting the calibrated $\rho = 0.85$ and $\delta = 0.20$ into (20) gives an informative share of 47.4% and hence a dark share of 52.6% *at the optimum*, with no hoarding of any kind. The observed figure is 54.5%. The optimal benchmark therefore accounts for almost the entire observed dark-data share on its own.

Matching the remaining difference requires $\delta^{\text{eff}} = 0.188$, that is

$$\hat{\lambda} = 0.063,$$

implying over-accumulation of 6.4% relative to the cooperative stock. This is roughly a fifth of the effect we obtained in Section 5.1 at an assumed $\lambda = 0.30$, and the comparison runs the other way as well: $\lambda = 0.30$ implies a dark share of 60.4% and $\lambda = 0.50$ implies 64.3%, both well above anything reported in the survey evidence. The competitive intensities conventionally described as moderate are, on this reading, inconsistent with the data.

How much should be made of $\hat{\lambda}$. Rather little, and this is the substantive point of the subsection. Equation (20) says the dark-data share is pinned down by ρ and δ, and it is highly sensitive to both. Holding $\delta = 0.20$, the dark share at the optimum ranges from 64% at $\rho = 0.75$ to 28% at $\rho = 0.95$, so the implied $\hat{\lambda}$ ranges from -0.25 to +1.66. Holding $\rho = 0.85$, varying δ over the empirical range delivers $\hat{\lambda}$ between -0.47 and +0.65. Over plausible parameter values the implied competitive intensity is not merely imprecise; *its sign is not determined*. The crossings are sharp and can be stated exactly: $\hat{\lambda} = 0$ at $\rho^* = 0.837$ and at $\delta^* = 0.188$, and below those values the optimal benchmark over-predicts the observed dark share so that the inversion asks for negative hoarding. Our calibrated $\rho = 0.85$ and $\delta = 0.20$ lie only 0.013 and 0.012 above the respective thresholds. A positive estimate of competitive hoarding therefore survives only within about one and a half per cent of the persistence parameter, which is far inside the range over which that parameter is itself known.

We therefore read the exercise as a negative result, and we think it is the most useful thing the paper has to say about the empirical literature. The dark-data statistic that motivates almost every discussion of data hoarding, including our own introduction, is dominated by two parameters that have nothing to do with strategic behaviour: the persistence of the environment and the physical durability of records. A firm in a fast-moving industry will

report a high dark-data share while behaving entirely optimally. Conversely, a low reported share is compatible with substantial hoarding. The statistic is a measure of informational obsolescence, not of managerial or competitive failure, and inferences about hoarding drawn from it are unsafe.

A load-bearing assumption. Before drawing conclusions we must be explicit about a step that the sign of the inversion rests on. Our reduced form replaces δ by δ^{Nash} throughout, so competition slows the decay of the *physical* stock as well as raising the private value of retaining it. That is what makes the dark share increase in λ. Suppose instead that competition raises only the private valuation of retention—entering the informational retention coefficient ζ but leaving physical decay at δ, as would be the case if rivalry changed what firms are willing to keep rather than how long records physically survive. Then $s = \delta/[(1-\rho^2) + \delta^{\text{Nash}}\rho^2]$ is *increasing* in λ: competition would *lower* the dark share, to 48.5% at $\lambda = 0.30$, and no positive λ could reproduce an observed share above the optimal benchmark. The inversion would return a negative value for any λ. Which interpretation is right is an empirical question about whether data exclusivity operates through retention policy or through storage technology, and we cannot settle it here. We flag it because it is the difference between a small positive estimate and a negative one, and because the reduced form of Section 5.1 was not designed with this moment in mind.

Two further qualifications. First, the mapping from $1-s$ to a survey's dark-data figure is an interpretation, not an identity: the survey counts records or bytes whose value is unknown to the respondent, whereas $1 - s$ is the shortfall of realized informational content relative to a stock of fully current records. These coincide only if unrealized informational content and self-reported unknown value line up, which we cannot verify. Second, the inversion uses a single moment and is conditional on ρ and δ taken from elsewhere; a firm-level panel of data-asset disclosures would allow ρ and δ to be estimated jointly with λ, and would let the sign question be settled rather than bracketed. Until that is done, the honest summary is that the competitive mechanism is powerful per unit of λ—each 0.1 of competitive intensity

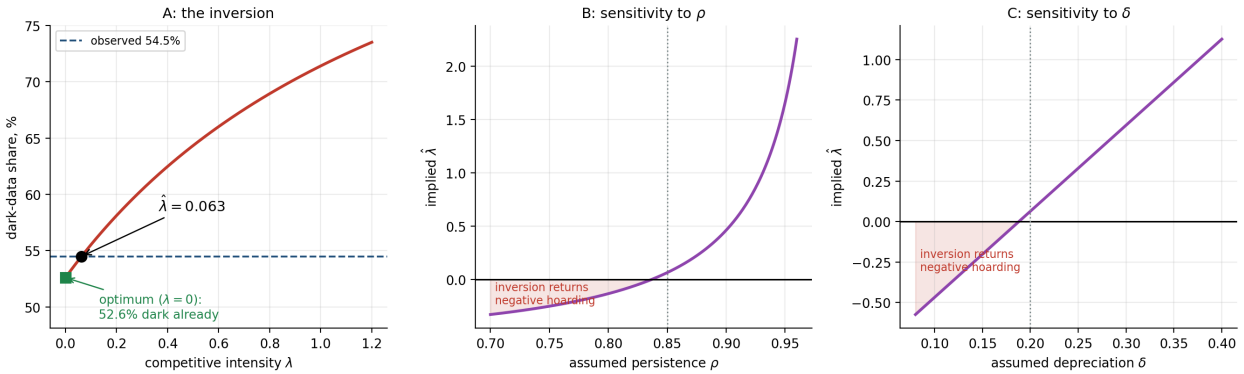


Figure 7: Inverting the dark-data share. Panel A: the dark share $1 - s$ as a function of competitive intensity λ, with the observed 54.5% marked; the intersection gives $\hat{\lambda} = 0.063$, and the value at $\lambda = 0$ shows that 52.6% of the observed share is generated by optimal accumulation alone. Panel B: the implied $\hat{\lambda}$ as a function of the assumed persistence ρ, with the zero line marked—the crossing is at $\rho^* = 0.837$, only 0.013 below the calibrated value. Panel C: the same against the assumed physical depreciation rate δ, crossing at $\delta^* = 0.188$. *Notes:* Panels B and C show that the sign of $\hat{\lambda}$ is not identified over the plausible parameter range, which is the subsection's main conclusion.

adds roughly ten per cent to the data stock—and that the available evidence does not tell us what λ is.

5.5 Scope and Extensions

The framework is a single-firm, partial-equilibrium model, and the conclusions above should be read within that scope. We state what it leaves out and in which direction each omission is likely to move the results.

Concentration and entry. Hoarding may deter entry or entrench incumbents (Prüfer and Schottmüller 2021), a cost that lies entirely outside our firm-value measure. Our private-cost estimate is therefore best read as a lower bound on the total cost of over-accumulation, not as an estimate of it, which is consistent with the paper's argument that the case against hoarding must be built elsewhere than in the hoarding firm's own productivity.

Informational externalities. When one agent's data is informative about others, equilibrium acquisition departs from the social optimum (Acemoglu et al. 2022). This operates on the social side of the ledger and does not alter the firm's problem, so it leaves the op-

timal benchmark, the dark-data identification result and the ranking of the three wedges unchanged, while adding to the social cost.

Many firms. The duplication cost in Section 4 is a reduced-form statement about N firms that each solve the single-firm problem. A model in which firms choose data holdings strategically, trade data or share it would endogenize the overlap ω, which Appendix E ties to multi-homing but does not model, and would replace the open-loop steady state of Appendix E with a Markov-perfect equilibrium. We regard this as the most valuable theoretical extension, and note that the identification result for the dark-data share survives it, since that result requires only that each firm's two stocks be proportional to its own scale.

6 Conclusion

This paper asks whether the large share of stored data that firms never use is evidence of hoarding. Our answer builds on the observation that data is not merely an input firms purchase but a by-product of production itself. Capital investment generates data; data improves productivity forecasts; better forecasts raise the return to capital. Embedding this loop in a dynamic model with Bayesian learning, capital adjustment costs and fixed operating costs, and deriving the decay of informational value from the joint action of physical obsolescence and the mean reversion of productivity, yields a unique optimal data stock.

Four findings stand out, and three of them run against the presumption one might bring to the question.

First, the feedback loop is quantitatively weak as a level effect. The multiplier contributes 3.1% of the marginal return to capital, not the double-digit figure the mechanism's prominence in the literature might suggest. The reason is structural rather than calibration-specific: the marginal value of precision declines as $(1 + \Omega)^{-2}$, so any firm that already forecasts tolerably well gains little from more data. What the loop does deliver is speed—a data-poor firm deploys 26% of its steady-state capital immediately and halves its gap within

3 periods—and an ordering, in which the input the firm controls moves first and the asset it can only harvest follows. We also find no adjustment asymmetry: excess data decays at the informational retention rate, so over-accumulation is not intrinsically persistent.

Second, among the three wedges that push firms beyond the optimum, competitive escalation is the most powerful per unit of intensity—roughly ten per cent of the data stock for each 0.1 of λ, against 3.5% for managerial signalling at $\mu = 0.05$ and 0.16% for the precautionary motive. Competition acts on the data stock almost entirely without moving physical capital, which means this form of over-accumulation is nearly undetectable in conventional investment data.

The third finding qualifies the second and is the one we would most like read. The intensity of competition cannot simply be assumed, and when we discipline it against the observed dark-data share, the mechanism turns out to be identified by that moment alone: the ratio of informative to stored data depends only on the persistence of productivity and the depreciation of records, with firm scale, data-mining efficiency and managerial preferences cancelling exactly. Inverting it, 52.6 of the observed 54.5 percentage points of dark data are produced by optimal accumulation with no hoarding whatever; the residual implies $\hat{\lambda} = 0.063$ and excess accumulation of 6.4%; and over plausible values of persistence and depreciation the implied intensity is not even robustly signed. The dark-data statistic that motivates this literature, our own introduction included, is a measure of informational obsolescence rather than of strategic or managerial failure. A firm in a fast-moving industry will report a high dark-data share while behaving optimally, and a low share is compatible with substantial hoarding. We think this is the most useful thing the paper has to say to the empirical literature, and it is a caution rather than a result.

Fourth, the cost of misallocation is essentially zero, but the cost of rivalry is not. The loss of firm value is at most 0.0013% even when the data stock exceeds the optimum by some 302%, because the value function is flat near its maximum, so hoarding is not a story about firms damaging their own productivity through misallocated capital. Appendix E shows,

however, that the maintenance expenditure through which rivalry generates an excess stock costs 0.24% to 1.01% of net flow profit at $\lambda = 0.25$ to 0.48. The second social cost we can identify lies in duplicated infrastructure: when many firms maintain overlapping records of the same activity, the fixed cost of the data estate is paid repeatedly, which at ten firms and fifty per cent overlap amounts to 1.2% of industry value—larger than the misallocation cost by two to three orders of magnitude.

The policy implication follows from these findings jointly, and it differs from the one a naive reading of the mechanisms would suggest. If the operative social cost is duplication rather than distortion, the instrument of choice is the one that lets a single collection serve many users: interoperability and mandated data sharing, which is also the instrument with the most leverage on the competitive channel. Its quantitative case, however, cannot be assembled from dark-data statistics, and building it will require direct evidence on the exclusivity of data holdings. Retention mandates are less attractive than they appear, since only the product $(1 - \delta)\rho^2$ enters the firm's problem and forced deletion is therefore informationally equivalent to environmental turbulence—it reduces measured hoarding partly by destroying information of genuine value. Accounting standards that separate productive from dark data are well targeted at the managerial channel but address a smaller distortion, and on the inverted competitive intensity the two are of comparable size.

Four limitations bound these conclusions and mark the agenda we regard as most valuable. The framework is partial equilibrium with a single firm, so it is silent on the effects of hoarding on concentration and entry, which may well dominate what we measure. The strength of the feedback loop is governed by the scale of the information-value function, which is not separately identified from the signal-noise normalization; disciplining it against a direct estimate of the productivity return to data capability is the single most useful extension. The duplication cost is linear in an overlap parameter we assume rather than measure. And the inversion of Section 5.4 uses a single moment conditional on ρ and δ taken from other studies; a firm-level panel of data-asset disclosures would let ρ, δ and λ be estimated

jointly and would settle the sign question we can currently only bracket. All four point in the same direction: towards data, and specifically towards the firm-level disclosures of capitalized data resources now becoming available, which would allow the mechanisms modeled here to be tested rather than merely calibrated.

Acknowledgements and Disclosures

This research received no specific grant or financial support from any funding agency in the public, commercial or not-for-profit sectors. The author has no financial relationships or other potential conflicts of interest to disclose with respect to this paper, consistent with the NBER research dissemination disclosure policy. Any remaining errors are the author's own.

References

Abel, Andrew B. and Janice C. Eberly (1994). "A Unified Model of Investment Under Uncertainty". In: *American Economic Review* 84.5, pp. 1369–1384.

Acemoglu, Daron et al. (2022). "Too Much Data: Prices and Inefficiencies in Data Markets". In: *American Economic Journal: Microeconomics* 14.4, pp. 218–256. DOI: 10.1257/mic.20200200.

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer (2015). "Identification Properties of Recent Production Function Estimators". In: *Econometrica* 83.6, pp. 2411–2451.

Ahmad, Nadim and Peter van de Ven (2018). "Recording and Measuring Data in the System of National Accounts". In: *Meeting of the Informal Advisory Group on Measuring GDP in a Digitalised Economy*. SDD/CSSP/WPNA(2018)5. OECD.

Atkeson, Andrew and Patrick J. Kehoe (2005). "Modeling and Measuring Organization Capital". In: *Journal of Political Economy* 113.5, pp. 1026–1053.

Bajari, Patrick et al. (2019). "The Impact of Big Data on Firm Performance: An Empirical Investigation". In: *AEA Papers and Proceedings* 109, pp. 33–37.

Bebchuk, Lucian Arye and Lars A. Stole (1993). "Do Short-Term Objectives Lead to Under- or Overinvestment in Long-Term Projects?" In: *The Journal of Finance* 48.2, pp. 719–729.

Bergemann, Dirk and Alessandro Bonatti (2019). "Markets for Information: An Introduction". In: *Annual Review of Economics* 11, pp. 85–107.

Bloom, Nicholas (2009). "The Impact of Uncertainty Shocks". In: *Econometrica* 77.3, pp. 623–685.

Brandt, Loren, Johannes Van Biesebroeck, and Yifan Zhang (2012). "Creative Accounting or Creative Destruction? Firm-level Productivity Growth in Chinese Manufacturing". In: *Journal of Development Economics* 97.2, pp. 339–351.

Cong, Lin William et al. (2022). "Endogenous Growth Under Multiple Uses of Data". In: *Journal of Economic Dynamics and Control* 141, p. 104395. DOI: 10.1016/j.jedc.2022.104395.

Cooper, Russell W. and John C. Haltiwanger (2006). "On the Nature of Capital Adjustment Costs". In: *The Review of Economic Studies* 73.3, pp. 611–633.

David, Joel M., Hugo A. Hopenhayn, and Venky Venkateswaran (2016). "Information, Misallocation, and Aggregate Productivity". In: *The Quarterly Journal of Economics* 131.2, pp. 943–1005.

De Loecker, Jan (2011). "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity". In: *Econometrica* 79.5, pp. 1407–1451.

Eeckhout, Jan and Laura Veldkamp (2022). *Data and Markups: A Macro-Finance Perspective*. Working Paper 30022. National Bureau of Economic Research.

Farboodi, Maryam, Roxana Mihet, et al. (2019). "Big Data and Firm Dynamics". In: *AEA Papers and Proceedings* 109, pp. 38–42.

Farboodi, Maryam and Laura Veldkamp (2021). *A Model of the Data Economy*. Working Paper 28427. Cambridge, MA: National Bureau of Economic Research.

Foster, Lucia, John Haltiwanger, and Chad Syverson (2008). "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" In: *American Economic Review* 98.1, pp. 394–425.

Freeman, Richard B., Buyuan Yang, and Baitao Zhang (2023). "Data Deepening and Non-balanced Economic Growth". In: *Journal of Macroeconomics* 75, p. 103503. DOI: 10.1016/j.jmacro.2023.103503.

Gartner, Inc. (2023). *IT Key Metrics Data 2023: Key Industry Measures*. Tech. rep. Gartner Benchmark Analytics.

Gollin, Douglas (2002). "Getting Income Shares Right". In: *Journal of Political Economy* 110.2, pp. 458–474.

Goodridge, Peter, Ourmane Chebli, and Jonathan Haskel (2015). *Measuring Activity in Big Data: New Estimates of Big Data Employment in the UK Market Sector*. Tech. rep. Imperial College Business School.

Hagiu, Andrei and Julian Wright (2023). "Data-Enabled Learning, Network Effects, and Competitive Advantage". In: *The RAND Journal of Economics* 54.4, pp. 638–667. DOI: 10.1111/1756-2171.12453.

Hall, Robert E. (2004). "Measuring Factor Adjustment Costs". In: *The Quarterly Journal of Economics* 119.3, pp. 899–927.

Holmström, Bengt (1999). "Managerial Incentive Problems: A Dynamic Perspective". In: *The Review of Economic Studies* 66.1, pp. 169–182.

Hopenhayn, Hugo A. (1992). "Entry, Exit, and Firm Dynamics in Long Run Equilibrium". In: *Econometrica* 60.5, pp. 1127–1150.

Jones, Charles I. and Christopher Tonetti (2020). "Nonrivalry and the Economics of Data". In: *American Economic Review* 110.9, pp. 2819–2858.

Jovanovic, Boyan (1982). "Selection and the Evolution of Industry". In: *Econometrica* 50.3, pp. 649–670.

Koutroumpis, Pantelis, Aija Leiponen, and Llewellyn D. W. Thomas (2020). "Markets for Data". In: *Industrial and Corporate Change* 29.3, pp. 645–660.

Li, Jinpu et al. (2025). "Data Asset Bubbles: A Theoretical Framework and Policy Trade-offs". In: *Management World (Guanli Shijie)* 9. In Chinese. DOI: 10.19744/j.cnki.11-1235/f.2025.0127.

Li, Tang, Qing Li, and Chuxia Chen (2020). "The Effect of Data Management Capability on Firm Productivity: New Evidence from the China Employer-Employee Survey". In: *China Industrial Economics (Zhongguo Gongye Jingji)* 6. In Chinese., pp. 174–192. DOI: 10.19581/j.cnki.ciejournal.2020.06.010.

Melitz, Marc J. (2003). "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity". In: *Econometrica* 71.6, pp. 1695–1725.

Prüfer, Jens and Christoph Schottmüller (2021). "Competing with Big Data". In: *The Journal of Industrial Economics* 69.4, pp. 967–1008. DOI: 10.1111/joie.12259.

Rassier, Dylan G., Robert J. Kornfeld, and Erich H. Strassner (2019). *Treatment of Data in National Accounts*. Tech. rep. Bureau of Economic Analysis, BEA Advisory Committee.

Veritas Technologies (2020). *The Data Iceberg Report: Global Dark Data Study*. Tech. rep. Veritas Technologies LLC.

Xu, Xianchun, Zekun Lei, and Yaru Hu (2025). "Measurement of China's Gross Data Capital Formation and Data Capital Stock". In: *Economic Research Journal (Jingji Yanjiu)* 3. In Chinese., pp. 54–68.

Appendix A: Motivation and Accuracy of the Precision Transition

This appendix motivates equations (6′) and (7′) and states precisely what is approximate in them. The firm's log productivity follows the AR(1) process

$$a_t = \rho\, a_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\theta^2), \tag{1}$$

with unconditional variance $\sigma_a^2 = \sigma_\theta^2/(1-\rho^2)$.

Step 1: The informational content of an aged record

A datum collected in period $t-j$ is a noisy observation of the state then prevailing,

$$s_{t-j,m} = a_{t-j} + \varepsilon_{t-j,m}, \qquad \varepsilon_{t-j,m} \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{2}$$

To use it in forecasting a_t, note that iterating the AR(1) process gives $a_t = \rho^j a_{t-j} + u_{t,j}$ where $u_{t,j} = \sum_{i=1}^{j} \rho^{j-i}\eta_{t-j+i}$ is independent of a_{t-j}. The record therefore carries information about $\rho^j a_{t-j}$. Its exact incremental precision about a_t is $\sigma_a^{-2}\,\rho^{2j} q/(1-\rho^{2j}q)$ with $q = \sigma_a^2/(\sigma_a^2+\sigma_\varepsilon^2)$, which equals $\rho^{2j}/\sigma_\varepsilon^2$ to first order when the signal is weak relative to the prior ($\sigma_\varepsilon^2 \gg \sigma_a^2$). We adopt the first-order form: a signal about a state j periods old is worth a fraction ρ^{2j} of a contemporaneous signal. At the baseline the approximation overstates the exact increment by 4% at $j=1$ and 12% at $j=5$.

Step 2: Aggregating surviving records

Of the $n_{t-j} = z k_{t-j-1}^{\alpha}$ records generated j periods ago, a fraction $(1-\delta)^j$ survives physically. Weighting each surviving record by its informational content ρ^{2j} and summing over vintages defines the *effective* data stock

$$\tilde{D}_t \;\equiv\; \sum_{j\geq 0} \left[(1-\delta)\rho^2\right]^j n_{t-j} \;=\; \sum_{j\geq 0} \zeta^{\,j}\, n_{t-j}, \qquad \zeta \equiv (1-\delta)\rho^2. \tag{3}$$

Because the weight is geometric, this admits the recursive representation

$$\tilde{D}_{t+1} = \zeta \tilde{D}_t + z k_t^{\alpha}, \tag{4}$$

which is equation (6′). The two channels of decay enter multiplicatively because a record must both survive and remain relevant in order to be informative, and these are independent

events.

Step 3: Posterior precision

Absent any data the firm knows only the unconditional distribution, so its prior precision is $1/\sigma_a^2 = (1-\rho^2)/\sigma_\theta^2$. Adding the age-weighted signal precision of the effective stock, which treats the information in separate records as additive,

$$\Omega_t = \underbrace{\frac{1-\rho^2}{\sigma_\theta^2}}_{\text{prior floor}} + \underbrace{\frac{\tilde{D}_t}{\sigma_\varepsilon^2}}_{\text{age-weighted signal precision}}, \tag{5}$$

which is equation (7′).

Accuracy of the additive form

Additivity is an approximation, not an implication of Bayes' rule. Records of different ages are signals about different states that are linked through the shared innovations $u_{t,j}$, so their information about a_t overlaps and the exact aggregate precision is lower than the sum. To gauge the error we compare (7′) with the exact Kalman steady state of a firm that retains its posterior and never loses a record ($\delta = 0$), at the baseline signal flow. The exact precision is 13.9, of which 7.0 is information beyond the prior floor, while (7′) gives 17.9, of which 10.9 is information: the additive form overstates the information content by a factor of about 1.6. This is quantitatively equivalent to understating σ_ε^2 by the same factor, and Appendix G.2 shows that the amplification $\mathcal{M}-1$ then falls from 3.1% to about 2% rather than changing in kind. We therefore regard (7′) as a reduced-form information technology whose scale is not identified, which is how the text treats it. The dark-data share (20), the depreciation wedges and the competitive game do not depend on its scale. The script is `check_precision.py`.

Remark: why not a recursive prior

A common alternative writes precision recursively as $\Omega_t = [\rho^2/\Omega_{t-1} + \sigma_\theta^2]^{-1} + D_t/\sigma_\varepsilon^2$. This is not equivalent to (7′), and combining a recursive prior with a stock term double counts: the prior term $[\rho^2/\Omega_{t-1} + \sigma_\theta^2]^{-1}$ already incorporates every signal the firm has observed up to $t-1$, discounted at exactly the rate ρ^2, so adding the surviving records again credits the same observations twice. The recursive form is correct if the stock term is restricted to *newly arrived* data, n_t/σ_ε^2; but in that case the stock D_t never enters the precision equation, data is not a state variable, and there is no reason for a firm to retain records at all. Our

formulation resolves this by tracking the age-weighted stock and using the unconditional prior as the floor, so each datum is counted exactly once with the correct discount. The hybrid specification would inflate the data feedback premium by a factor of approximately $1 + \beta(1-\delta)/[1-\beta(1-\delta)] \approx 4.3$ at our calibration.

Appendix B: Proof of Steady-State Uniqueness

The steady-state Euler equation (17) is

$$r = R(k) \equiv \alpha k^{\alpha-1} \left[P\,\mathbb{E}[A \mid \Omega(k)] + \beta z \lambda(k) \right], \tag{6}$$

where, substituting (13) and (15),

$$\Omega(k) = \frac{1-\rho^2}{\sigma_\theta^2} + \frac{z k^\alpha}{\sigma_\varepsilon^2(1-\zeta)}, \qquad \lambda(k) = \frac{P\bar{A}}{\sigma_\varepsilon^2(1-\beta\zeta)} \cdot \frac{k^\alpha}{(1+\Omega(k))^2}. \tag{7}$$

Claim 1 *Equation (17) admits exactly one positive solution k^*.*

Proof. It suffices to show that R is strictly decreasing on $(0,\infty)$ with $\lim_{k\to 0} R(k) = +\infty$ and $\lim_{k\to\infty} R(k) = 0$, since $r > 0$ is constant.

Write $R(k) = \alpha k^{\alpha-1}\left[P\,\mathbb{E}[A \mid \Omega(k)] + \beta z \lambda(k)\right] \equiv \alpha k^{\alpha-1} B(k)$ and take logarithms:

$$\frac{d\ln R}{d\ln k} = (\alpha - 1) + \frac{d\ln B}{d\ln k}. \tag{8}$$

We bound the second term. Let $x \equiv z k^\alpha/[\sigma_\varepsilon^2(1-\zeta)] = \Omega(k) - \Omega_0$ with $\Omega_0 \equiv (1-\rho^2)/\sigma_\theta^2 > 0$, so that $d\ln x / d\ln k = \alpha$.

First component. $\mathbb{E}[A \mid \Omega] = \bar{A}\Omega/(1+\Omega)$ gives

$$\frac{d\ln \mathbb{E}[A \mid \Omega]}{d\ln k} = \frac{\alpha x}{\Omega(1+\Omega)} \le \frac{\alpha}{1+\Omega_0} < \alpha,$$

using $x \le \Omega$.

Second component. $\lambda(k) \propto k^\alpha (1+\Omega(k))^{-2}$ gives

$$\frac{d\ln\lambda}{d\ln k} = \alpha - \frac{2\alpha x}{1+\Omega} \;<\; \alpha \qquad \text{for all } k > 0,$$

and this is negative once $2x > 1 + \Omega$, hence for all sufficiently large k.

Since B is a positive combination of the two components, its elasticity is a weighted

average of theirs with weights in $[0,1]$, so both bounds above give

$$\frac{d\ln B}{d\ln k} \;<\; \alpha,$$

and substituting into (8),

$$\frac{d\ln R}{d\ln k} \;<\; (\alpha-1)+\alpha = 2\alpha-1.$$

For $\alpha < 1/2$ this is strictly negative, establishing monotonicity. This covers the empirically relevant range for a capital elasticity, and in particular our calibration $\alpha = 0.33$.

The limits are immediate. As $k \to 0$, $B(k) \to P\bar{A}\Omega_0/(1+\Omega_0) > 0$ while $\alpha k^{\alpha-1} \to \infty$, so $R \to \infty$. As $k \to \infty$, $\mathbb{E}[A \mid \Omega] \to \bar{A}$ and $\lambda(k) \to 0$ at rate $k^{-\alpha}$, so B is bounded and $R \to 0$. Strict monotonicity and these limits give exactly one crossing of r. ■

The restriction $\alpha < 1/2$ is substantive. For $\alpha \in [1/2, 1)$ the argument fails, because the elasticity of λ approaches α as $k \to 0$ and the weighted average may then exceed $1-\alpha$. Rather than assert a bound we cannot prove, we report that uniqueness genuinely can fail there. We sweep two boxes of 2,880 parameter combinations each, both spanning $\rho \in [0.50, 0.99]$, $\sigma_\theta^2 \in [0.01, 0.20]$, $\delta \in [0.02, 0.60]$, $z \in [0.1, 5]$ and $\bar{A} \in [1, 10]$, evaluated over $k \in [10^{-5}, 10^{7}]$:

- for $\alpha \in \{0.20,\ 0.33,\ 0.45,\ 0.49\}$ there is no violation of monotonicity in any case, with $\max d\ln B/d\ln k = 0.280$, below $1-\alpha$ throughout;
- for $\alpha \in \{0.55,\ 0.70,\ 0.85,\ 0.95\}$ monotonicity fails in 684 of 2,880 cases, with $\max d\ln B/d\ln k = 0.544$. In those cases R is increasing over a median span of 1.5 decades of k, so the proof strategy breaks down.

Two qualifications keep this in proportion. The failures are systematic rather than numerical—they concentrate at high α (none occurs at $\alpha = 0.55$, and 63% occur at $\alpha = 0.95$) and are more frequent at high ρ and high σ_θ^2, though not exclusively so: about a quarter involve $\rho \le 0.85$, and the implied prior floor $\Omega_0 = (1-\rho^2)/\sigma_\theta^2$ exceeds unity in roughly a third of them. And a breakdown of monotonicity is not the same as multiplicity of equilibria: at the rental rate $r = 0.10$ used throughout, none of the 684 cases actually admits more than one crossing, because the non-monotone region of R lies well above r. What the sweep establishes is that the *proof* does not extend beyond $\alpha < 1/2$, not that the model misbehaves there. At our calibration $\Omega_0 = 6.94$ and $\alpha = 0.33$, comfortably inside the region covered by the proposition. A reader recalibrating to $\alpha \ge 1/2$ should verify uniqueness at their own r rather than presume it.

Corollary 1 *In the baseline calibration $R(1) = 0.91$, which exceeds $r = 0.10$, confirming $k^* > 1$. At the solution,*

$$k^* = 28.92, \quad \tilde{D}^* = 7.19, \quad D^* = 15.18, \quad \Omega^* = 14.13, \quad \mathcal{M} = 1.031.$$

A numerical check over 6,000 *logarithmically spaced grid points on $k \in [10^{-2}, 5\times10^3]$ confirms that R is strictly decreasing and crosses r exactly once, and the same check passes at each of the 21 discrete parameterizations tabulated in Sections 4–5 (the comparative statics in z, δ and ρ, the competitive cases, and the precautionary benchmark), as well as at the managerial and mixed cases when those are included.*

Appendix C: Decomposition of Informational Decay: Relevance vs. Obsolescence

This appendix quantifies the relative importance of the two decay channels identified in Section 3.3. The **relevance channel** captures the loss of forecasting value as the productivity state mean-reverts away from what any record describes; it is governed by ρ alone. The **obsolescence channel** captures the physical loss of records; it is governed by δ.

The decomposition is the exact identity (7a),

$$1 - \zeta \;=\; \underbrace{(1-\rho^2)}_{\text{relevance}} + \underbrace{\delta\rho^2}_{\text{obsolescence}}, \qquad \zeta = (1-\delta)\rho^2,$$

which requires no approximation and, importantly, no evaluation point in Ω. This is a substantive improvement over a formulation in which the informational decay rate is state-dependent: such a rate has to be tabulated at particular precision levels, and the shares then depend on where one chooses to evaluate. Here the shares depend only on ρ and δ.

The economic reading of Table 3 bears directly on policy. At the baseline, roughly two-thirds of the erosion of informational value comes from the channel no regulator can touch. A retention mandate that doubled δ from 0.20 to 0.40 would raise total decay from 0.422 to 0.567, a change of a third—non-trivial, but far less than the doubling of δ might suggest, because the mandate operates on the smaller of the two components. Only in industries with unusually persistent fundamentals, where the relevance channel is weak, do retention rules bite hard.

Table 3: Two-Channel Decomposition of Informational Decay

δ	Relevance $(1-\rho^2)$	Obsolescence $(\delta\rho^2)$	Total $(1-\zeta)$	Relevance share
0.02	0.2775	0.0144	0.2920	95.1%
0.05	0.2775	0.0361	0.3136	88.5%
0.10	0.2775	0.0722	0.3498	79.3%
0.20 (baseline)	**0.2775**	**0.1445**	**0.4220**	**65.8%**
0.31	0.2775	0.2240	0.5015	55.3%
0.50	0.2775	0.3612	0.6388	43.4%

Notes: Evaluated at $\rho = 0.85$, so $1-\rho^2 = 0.2775$ and $\rho^2 = 0.7225$. The relevance channel is invariant to δ by construction. At the baseline calibration the relevance channel accounts for 66% of total informational decay.

Appendix D: Mixed Hoarding Effects

Section 5 analyzes each hoarding mechanism in isolation. Table 4 reports the cumulative effect when all three operate simultaneously, under a representative mixed-hoarding scenario: moderate competitive intensity ($\lambda = 0.30$, so $\delta^{\text{Nash}} = 0.154$), moderate managerial preference ($\mu = 0.05$), and depreciation uncertainty ($\delta \in \{0.05, 0.50\}$, $\pi = 0.15$).

Table 4: Mixed Hoarding — Cumulative Effects

Scenario	$\boldsymbol{k^*}$	$\boldsymbol{D^*}$	$\boldsymbol{\Delta D^*}$ **vs. baseline**
Baseline (no hoarding)	28.92	15.18	—
Competitive only ($\lambda = 0.30$)	29.03	19.75	30.2%
Institutional only ($\mu = 0.05$)	32.12	15.71	3.5%
Precautionary only (δ uncertain)†	28.89	11.03	0.16%
All three combined	**33.01**	**20.61**	**35.8%**
Additive sum of individual effects	—	—	33.9%

Notes: Each row evaluates the accounting stock at that scenario's own effective depreciation rate, $D^* = z(k^*)^\alpha/\delta^{\text{eff}}$; this identity holds exactly in every row to machine precision. †The precautionary row is measured against its own benchmark—certainty at the ergodic mean $\bar{\delta} = 0.275$, where $D^* = 11.02$—and *not* against the baseline, since the two involve different depreciation rates and are not comparable in levels.

The combined effect (35.8%) is modestly super-additive relative to the sum of individual effects (33.9%). The interaction runs through the informational annuity: competitive escalation raises ζ, which raises $1/(1-\beta\zeta)$, which amplifies the value of the marginal unit of data and so strengthens the managerial and precautionary channels operating on top of it. The super-additivity is small because all three channels are, in the end, adjustments to a feedback premium that is itself only 3.1% of the marginal revenue product.

Appendix E: Micro-Foundation of Competitive Hoarding

Section 5 models competitive hoarding through the reduced form $\delta^{\text{Nash}} = \delta/(1+\lambda)$. This appendix derives it from a game in which firms compete on data holdings and choose how intensively to maintain them, shows exactly what the reduced form does and does not capture, and uses the game to price a cost of rivalry that the reduced form omits. The code is `micro_game.py`.

E.1 The game

Consider N symmetric firms $i = 1, \ldots, N$. Firm i chooses capital k_i and a maintenance intensity $e_i \geq 0$. Maintenance slows the depreciation of its records,

$$\delta(e) = \frac{\delta_0}{1+e}, \qquad \text{at a flow cost } \tfrac{\kappa}{2}\, e^2 D_i, \tag{9}$$

proportional to the stock maintained. The state transitions are those of the benchmark, (6) and (6′), evaluated at $\delta(e_i)$. Firm i's profit coefficient depends on its own precision and on its holdings relative to the average of its rivals,

$$\mathbb{E}[A_i \mid \Omega_i, \mathbf{D}] = \bar{A}\, \frac{\Omega_i}{1+\Omega_i} \left(1 + \eta\, \frac{D_i - \bar{D}_{-i}}{\bar{D}_{-i}}\right), \qquad \bar{D}_{-i} = \frac{1}{N-1} \sum_{j \neq i} D_j, \tag{10}$$

where $\eta > 0$ is the sensitivity of profit to relative data advantage. The relative term is written in the accounting stock D rather than the effective stock $\tilde{D}$ because what confers exclusivity is the size and breadth of the holding, not its forecasting content. The solution concept is a symmetric steady state of the open-loop Nash equilibrium: each firm takes its rivals' paths as given.

E.2 First-order conditions

Let $\tilde{\lambda}$ and λ^c denote the shadow prices of the effective and accounting stocks. At a symmetric steady state the relative term equals one, and the envelope conditions are

$$\tilde{\lambda} = \frac{P}{\sigma_\varepsilon^2} \frac{\partial \mathbb{E}[A \mid \Omega]}{\partial \Omega}\, \frac{k^\alpha}{1 - \beta\zeta(e)}, \tag{11}$$

$$\lambda^c = \frac{P\, \mathbb{E}[A \mid \Omega]\, \eta\, k^\alpha / D \;-\; \frac{\kappa}{2} e^2}{1 - \beta\,(1 - \delta(e))}, \tag{12}$$

the first being (13) at the retention coefficient implied by e. The numerator of λ^c is the marginal gain in relative position, using $\partial(\text{relative term})/\partial D_i = \eta/\bar{D}_{-i} = \eta/D$, less the maintenance cost saved by holding less. The Euler equation for capital and the first-order condition for maintenance are

$$r = \alpha k^{\alpha-1}\Big[P\,\mathbb{E}[A \mid \Omega] + \beta z\,(\tilde{\lambda} + \lambda^c)\Big], \tag{13}$$

$$\kappa\, e\, D = \beta\,\frac{\delta_0}{(1+e)^2}\Big[\lambda^c D + \rho^2\tilde{\lambda}\,\tilde{D}\Big]. \tag{14}$$

Equation (14) equates the marginal cost of maintenance with the value of the stock it preserves: a fraction $\delta_0/(1+e)^2$ of D and of $\rho^2\tilde{D}$ that would otherwise be lost.

Proposition 2 *In the symmetric steady state: (i) competition enters the capital decision only through λ^c, the shadow value of the accounting stock, which has the same structure as the managerial valuation of Section 5.2; (ii) the number of rivals N does not appear in (13)–(14); and (iii) if e_0 denotes the equilibrium maintenance intensity at $\eta = 0$ and $\delta_{\text{coop}} = \delta_0/(1+e_0)$ the corresponding depreciation rate, then the effective depreciation rate at any e is exactly*

$$\delta^{Nash} = \frac{\delta_{\text{coop}}}{1+\lambda}, \qquad 1+\lambda \equiv \frac{1+e}{1+e_0}.$$

Proof. (i) and (ii) follow by inspection: η appears only in λ^c, and because the comparator is the rivals' *average*, $\partial(\text{relative term})/\partial D_i = \eta/D$ regardless of N. (iii) $\delta(e)/\delta(e_0) = (1+e_0)/(1+e)$. ■

Part (iii) is what the reduced form of Section 5 assumes, and it shows that λ has a concrete meaning: the proportional increase in maintenance intensity that rivalry induces. Parts (i) and (ii) show what the reduced form omits. Rivalry also raises the shadow value of holding data, so it moves capital as well as depreciation, and the strength of the effect does not dilute with the number of competitors under this specification.

E.3 Quantitative content

We normalize $e_0 = 1$, so that $\delta_0 = 2\delta_{\text{coop}} = 0.40$, and choose κ so that $e = 1$ is optimal at $\eta = 0$ ($\kappa = 0.0024$); all other parameters are those of Table 1. The only free parameter is then the rivalry intensity η.

Table 5 supports four statements. First, the reduced form is a good approximation to the data-stock gap at moderate intensity: 26% in the game against 25% at $\lambda = 0.25$, and 50% against 48% at $\lambda = 0.48$, with the gap widening at higher λ. Second, the dark-data share depends on rivalry only through δ^{Nash}, so it is identical in the game and in the reduced

Table 5: The competitive game against the reduced form

η	e^*	λ	Data-stock gap (%)		Δk^* (%)	Dark share (%)	Net flow loss (%)
			Game	Reduced form	Game		
0.00	1.00	0.00	0.0	0.0	0.0	52.6	0.00
0.02	1.49	0.25	25.5	24.7	2.2	59.2	0.24
0.05	1.95	0.48	50.1	47.9	5.1	63.9	1.01
0.10	2.46	0.73	78.4	73.6	9.5	68.0	2.61
0.20	3.13	1.07	118.2	107.2	17.9	72.1	6.31
0.40	3.98	1.49	174.3	150.0	33.8	76.1	14.70

Notes: "Game" is the symmetric open-loop steady state of Appendix E; "Reduced form" is the data-stock gap implied by $\delta^{\text{Nash}} = \delta_{\text{coop}}/(1+\lambda)$ at the same λ, with capital free to respond. Δk^* is the capital response in the game. Net flow loss is the fall in operating surplus net of maintenance cost, as a percentage of the non-competitive value.

form by construction, and the identification result of Section 5.4 is unaffected. Third, capital moves more than the reduced form implies (2.2% against 0.3% at $\lambda = 0.25$), since rivalry raises the shadow value of the stock, but it still moves an order of magnitude less than the stock itself, so competitive hoarding remains largely invisible in capital data.

Fourth, and most important, the game prices a cost that the reduced form cannot. In a symmetric equilibrium the relative-advantage terms net to zero in levels, so the additional maintenance spending buys no net advantage: it is pure rent dissipation, and it is incurred by every firm. Table 5 puts it at 0.24% of net flow profit at $\lambda = 0.25$ and 1.01% at $\lambda = 0.48$. Three features of this number matter. It is two to three orders of magnitude larger than the firm-value loss from the flat misallocation of Section 4, which is the reason the two must be kept apart. It is of the same order as the duplication cost of 1.2% of industry value, so the case for interoperability does not rest on the overlap parameter alone. And it does not depend on ω or on any assumption about how much of one firm's data duplicates another's.

E.4 Limits of the derivation

Three limits should be stated. The solution concept is open-loop: a Markov-perfect equilibrium would add the effect of a firm's own stock on its rivals' subsequent choices, and we do not solve it. The normalization $e_0 = 1$ is arbitrary; it fixes the scale of λ at a given η but not the mapping between λ and the data-stock gap, which depends only on δ^{Nash}. And η is not estimated, so Table 5 says what a given intensity of rivalry implies, not what the intensity is; the dark-data share, which would otherwise be the natural moment, identifies only the combination of λ with ρ and δ that Section 5.4 shows to be weakly informative.

E.5 The overlap parameter

The duplication cost of Section 4 is linear in the overlap ω between firms' data holdings. Suppose a share ω of each firm's holdings is also held by other firms, and that the overlapping data is held by $\bar{m} \leq N$ firms on average. Of the $\bar{m}$ copies of each overlapping record only one is necessary, so the redundant infrastructure is $N\omega(1 - 1/\bar{m})F$ per period, with present value

$$\frac{N\,\omega\,(1 - 1/\bar{m})\,F}{1 - \beta},$$

which equals $(N - 1)\omega F/(1 - \beta)$ when $\bar{m} = N$ and is smaller otherwise. The expression in the text is therefore an upper bound for a given ω. In a market where customers multi-home, ω is the share of customers served by more than one firm and $\bar{m}$ is the average number of providers among them, both in principle observable from customer or platform-usage data. We have not matched them to data, and flag this as the empirical input on which the quantitative duplication estimate depends.

Appendix F: Numerical Method and Accuracy

All quantitative results are produced by a single program, `model_core.py`, which writes every reported figure to a file of LaTeX macros that this document reads. Every quantity that varies with the calibration is passed through this mechanism rather than transcribed by hand. Three classes of number are still typed directly in the source: parameter values from Table 1, the cells of Table 3 (whose entries are closed-form functions of ρ and δ alone), and a small number of rounded figures quoted in prose for readability. These are checked against the generated output but are not themselves generated. This discipline ensures that text, tables and appendices cannot report inconsistent figures for the same object.

F.1 Steady state

The steady state solves $R(k) = r$, with R as defined in Appendix B. Because Appendix B establishes that R is strictly monotone at our calibration, we use Brent's method on $k \in [10^{-4}, 5 \times 10^{3}]$ with absolute tolerance 10^{-12} and relative tolerance 10^{-14}. Precision $\Omega(k)$ and the shadow price $\lambda(k)$ are available in closed form from (15) and (13), so no inner loop or fixed-point iteration is required and no interpolation error enters the steady state.

Two checks are run. First, monotonicity and uniqueness are verified by evaluating R on 6,000 logarithmically spaced points and confirming that the differences are everywhere negative and that $\text{sign}(R - r)$ changes exactly once. This monotonicity check is also re-run

at each of the 21 discrete parameterizations tabulated in Sections 4–5, and passes at all of them. Second, the accounting identity $D^* = z(k^*)^\alpha/\delta^{\text{eff}}$ is verified to machine precision for the combined-hoarding scenario, and holds by construction in the other rows of Table 4 because D^* is computed from that expression rather than accumulated separately. This second check guards against reported stocks being inconsistent with reported capital levels.

F.2 Transition paths

The deterministic transition is computed by direct maximization over the capital sequence. Given an initial effective stock $\tilde{D}_0$ and lagged capital k_{-1}, we choose $\{k_t\}_{t=0}^{T-1}$ to maximize $\sum_{t=0}^{T-1} \beta^t \pi(k_t, k_{t-1}, \Omega_t)$ subject to (6′) and (7′), using L-BFGS-B with $T = 160$, analytic forward simulation of the state, box constraints $k_t \in [10^{-8}, 20k^*]$ and convergence tolerance 10^{-18}. The terminal condition is imposed explicitly: the final six periods are pinned at k^* by the bound constraints, and path statistics are computed over the remaining interior block. A free terminal horizon would leave boundary artifacts large enough to contaminate the reported maximum and minimum of k. With the terminal block pinned, $|k_t - k^*|$ at the end of the interior block is 7.9×10^{-5}. Initial guesses are a smooth ramp towards k^* from below and a constant at k^* from above.

Direct maximization is preferable to value function iteration here because the problem is deterministic and the state is two-dimensional, so the entire path can be optimised at once without discretising the state space or interpolating a value function. The cost is that optimality must be verified rather than imposed. We therefore check the intertemporal first-order condition

$$\alpha k_t^{\alpha-1} \left[P\, \mathbb{E}[A \mid \Omega_t] + \beta z \lambda_{t+1}\right] = r + \gamma(k_t - k_{t-1}) - \beta\gamma(k_{t+1} - k_t)$$

at every interior date of both reported paths, where λ_{t+1} is obtained by backward recursion on the envelope condition (12) along the realized path rather than from its steady-state closed form. The distinction matters: early in the transition Ω_t is far from Ω^*, and substituting the steady-state expression inflates the measured residual by roughly a factor of twenty. Taken over all interior dates of both the forward and the reverse path, the maximum relative residual is 9.0×10^{-4}. Transition statistics are accordingly reported to two significant figures; the third digit of a path quantity should not be relied upon.

F.3 Reproduction

Running `python3 model_core.py` regenerates `numbers.tex` and `results.json` and prints every table in this paper together with the diagnostic checks above. Running `python3 make_figures.py` then regenerates all seven figures from the same module; it shares every model function with `model_core.py` and computes nothing independently, so figures and tables cannot disagree. Recompiling `main.tex` reproduces the document. Running `sensitivity.py` and `micro_game.py` regenerates Appendix G and Table 5 from the same module. Total runtime is approximately two minutes on a single core, dominated by the two transition solves and by figure generation.

Appendix G: Sensitivity Analysis

This appendix collects the sensitivity checks referred to in the text. All entries are generated by `sensitivity.py`, which calls the same model functions as `model_core.py`.

G.1 Duplicated infrastructure

Table 6 reports the duplication cost $(N-1)\omega F/[(1-\beta)NV]$ as a percentage of industry value for different numbers of firms and overlap shares, holding the baseline calibration. The cost is increasing in both arguments and saturates in N. The smallest entry in the table is still 108 times the firm-value loss from misallocation at $\lambda = 3$, so the conclusion that duplication dominates misallocation does not rest on the particular values $N = 10$, $\omega = 0.5$. The table assumes that the overlapping data is held by all N firms and is therefore an upper bound for given (N, ω); if it is held by $\bar{m} = 2$ firms the entries are multiplied by $N/[2(N-1)]$, which leaves the smallest entries in the grid essentially unchanged.

Table 6: Duplication cost, percent of industry value

N	$\omega = 0.10$	$\omega = 0.25$	$\omega = 0.50$	$\omega = 0.75$
2	0.14	0.34	0.69	1.03
5	0.22	0.55	1.10	1.65
10	0.25	0.62	1.24	1.85
20	0.26	0.65	1.30	1.96
50	0.27	0.67	1.35	2.02

G.2 The level of the feedback loop

Table 7 reports the amplification $\mathcal{M}-1$ in per cent over a grid of the two parameters that are not separately identified, the loss curvature $\bar{A}$ and the signal noise variance σ_ε^2. The surface is non-monotone: raising $\bar{A}$ raises the value of information but also raises equilibrium capital and hence precision, which erodes the marginal value of the next unit of data. Across the grid the amplification never exceeds 3.1%, and the baseline lies close to the maximum, which means that the modest level effect reported in the text is if anything an upper estimate within this family.

Table 7: Feedback amplification $\mathcal{M}-1$ (%) over $(\bar{A}, \sigma_\varepsilon^2)$

$\bar{A}$	$\sigma_\varepsilon^2 = 0.25$	$\sigma_\varepsilon^2 = 1$	$\sigma_\varepsilon^2 = 4$	$\sigma_\varepsilon^2 = 16$
1	2.6	2.8	1.3	0.4
3	2.0	3.1	1.9	0.6
6	1.6	3.0	2.3	0.9
10	1.3	2.8	2.6	1.1

G.3 Innovation variance

Table 8 varies σ_θ^2, which Section 4 calibrates only to be consistent with documented dispersion. A larger innovation variance lowers the prior floor, so the firm is less well informed without data and the amplification rises, from 0.5% to 7.4% across the range. The dark-data share is exactly invariant to σ_θ^2, as equation (20) requires, and the competitive data-stock gap at $\lambda = 0.30$ is stable at about 30%. The inversion of Section 5.4 and the ranking of the hoarding wedges therefore do not depend on this parameter.

Table 8: Sensitivity to the innovation variance σ_θ^2

σ_θ^2	Floor	k^*	Ω^*	$\mathcal{M}-1$ (%)	Dark (%)	Gap, $\lambda = 0.3$ (%)
0.01	27.75	29.59	35.00	0.5	52.6	30.0
0.02	13.88	29.18	21.09	1.4	52.6	30.1
0.04	6.94	28.92	14.13	3.1	52.6	30.2
0.08	3.47	28.92	10.66	5.3	52.6	30.2
0.16	1.73	29.07	8.94	7.4	52.6	30.2