# Data Accumulation Distorts Talent Allocation

Pablo E. Rodriguez

September 2026

Data is valuable to firms because it reduces uncertainty, but exploiting it requires talent that could instead generate new ideas. I study this tradeoff in a model of occupational choice between science, data engineering, and routine production. Scientists create ideas and, as a byproduct of innovation, data. As data accumulates, it raises the relative return to data engineering and draws talent away from science. Successful innovation can therefore reduce future scientific effort, resulting in an inefficiently small share of scientists and lower welfare. The model explains why extensive data sharing can induce misallocation despite data's non-rivalry, and why data-driven automation can harm innovation. Using career histories for 35 million US workers over the last two decades, I document three facts consistent with the theory: past scientific activity predicts subsequent data employment and data-processing infrastructure; data work has expanded most in historically innovative labor markets and is accompanied by rising returns to data occupations; and data has become an increasingly important destination for scientific workers.

Rodriguez: MIT Sloan, pablo6@mit.edu. I am indebted to Maryam Farboodi for her encouragement and many rich discussions. For many useful comments, I also thank David Autor, Marta Bilghese, Olena Bogdan, Taha Choukhmane, Xavier Gabaix, GaOn Kim, Jonathan Parker, David Thesmar, S. Vish Viswanathan, Adrien Verdelhan, Anchi Xia, and seminar participants at the MIT Finance lunch.

## 1. Introduction

In the last twenty years, data jobs - from database administration to machine learning and artificial intelligence - have become increasingly important in the United States economy. Panel A of Figure 1 attests to this fact: the employment share of data-related occupations constitutes today 3% of the workforce and has almost doubled since 2002. Panel B of Figure 1 shows that far from drying up, the supply of data workers is on the rise.

These trends may appear unambiguously beneficial. Data helps firms forecast demand, manage inventories, target customers, and make other operational decisions. There is, indeed, mounting evidence that firms using data more intensively are more productive (see, for example, Brynjolfsson and McElheran 2016; David, Hopenhayn, and Venkateswaran 2016; Brynjolfsson, Jin, and McElheran 2021; Babina et al. 2025; Asriyan and Kohlhas 2026). Yet the growth of data-intensive production may create a less obvious tradeoff. The workers needed to process and exploit data may be drawn from the same pool of talent that produces new ideas: scientists, the now canonical engine of innovation since Romer (1990) contribution.

Panels C and D of Figure 1 indicate that such a competition for talent might be in place: since 1993, the share of PhD graduates in STEM fields such as computer science, mathematics, and statistics employed in academia has sharply declined, while the share going to industry has experienced a steep rise. These movements do not by themselves establish that scientific talent is moving specifically into data work: industry also employs scientists in research and development and other innovative activities. They do, however, point to a broader change in the outside opportunities facing scientific talent. If the expansion of data-intensive production makes data careers increasingly attractive to the same workers who can produce ideas, then the rise of the data economy can affect not only how firms use information, but also who chooses to innovate.

To bring these opposing views together, I develop a simple model of talent allocation in an economy with data accumulation. The theory explains how data-intensive production shapes the allocation of talent between innovation and data activities, what this allocation implies for economic output, and how it relates to the social optimum. I find that successful innovation generates data that pulls individuals away from science toward data engineering, causing science to cannibalize itself. The resulting decentralized equilibrium exhibits an inefficiently small share of scientists and lower welfare. Using career histories for more than 30 million workers, I document several descriptive patterns that are consistent with the model's predictions.

I begin by conceptualizing the firm as an organization that wants to maximize its value. To do so, it must solve three main questions: "what can we make?" (production), "what should we make?" (innovation), and "how do we convert production into profit"? (prediction). To answer these questions, the firm hires from three different occupations: routine workers who follow defined processes or procedures to produce a good, scientists to innovate, and data engineers to make *predictions*. The firm's hiring decisions generate wages for each occupation based on

its marginal product. Agents observe these wages and choose the occupation delivering the largest utility, trading off wage premia against idiosyncratic occupational preferences, in the self-selection tradition of Roy (1951). Hence, a firm's marginal products guide occupational choices.

These marginal products depend on the accumulation of two stocks on which scientists and data engineers rely to fulfill their jobs. As in canonical growth models, scientists build on a stock of ideas to, in turn, generate new ideas and innovate, raising product quality. The novelty of my model resides in the introduction of data engineers. They use a data stock to help the firm make operational decisions. Thus, data engineers reduce *uncertainty*: they allow the firm to sell output profitably by answering questions such as "which suppliers to use?", "how much inventory to hold?", "which customers to target?". These decisions can be modeled as choosing an action as close as possible to an optimal production technique that maximizes profit given an output level. By processing the aggregate data stock, data engineers can extract signals about this optimal technique, raising the profitability of the firm.

The central link between ideas and data is that innovation itself generates data. Higher-quality products create new interactions, transactions, and uses from which firms can collect information.[1] Improvements in smartphones, for example, enabled activities ranging from digital payments to navigation and health monitoring, each generating data that earlier products could not produce. In the model, scientists therefore generate two dynamic stocks: ideas, which future scientists use, and data, which future data engineers use. Scientists therefore drive data accumulation without direct intention, and generate a stock that they do not themselves exploit.

This differential accumulation between ideas and data drives the paper's core result: the endogenous growth of the data stock reallocates talent away from science, raising the market equilibrium share of data engineers relative to scientists. The intuition is a simple form of path dependence: strong early innovation cannibalizes future scientific effort, as talent is drawn to processing the data it helped create. The novelty of the mechanism is that the competing occupation is not rent-seeking in any form, unlike in most models of occupational choice involving canonical examples such as finance or law. Instead, data engineering is highly productive, and its attractiveness grows endogenously with the data generated by prior innovation. Therefore, there is a form of *static complementarity* between science and data occupations, but that very same complementarity leads to *dynamic substitution* and hence future reallocation away from science.

How strong this reallocation is naturally depends on how scarce data is. The model endogenously generates a partition of the state space that identifies an *information scarcity region*. In this region, additional data substantially improves predictions, making data engineers' marginal product very large and accelerating talent flows away from science. In the *deep scarcity region*, in particular, this reallocation is the strongest. As data continues to accumulate, the economy may

[1] In the industry, this mechanism is commonly known as the flywheel effect. See, for example, Amazon's and Nvidia's description of this "self-improving loop".

eventually reach *information saturation*, where the marginal value of additional data engineers declines and the reallocation toward data processing peaks and partially reverses.

A natural question to ask is whether the market allocation is efficient. To consider this question, I study the allocation chosen by a constrained planner that may reassign workers but must leave each worker at her market utility. This exercise shows that the answer is "no". The intuition runs as follows. Scientists generate benefits that accrue in future periods through innovation and data accumulation. The market compensates them only for their current contribution to output, while the planner recognizes their full intertemporal value. Data engineers, by contrast, affect only current-period efficiency through uncertainty reduction: they have no dynamic spillovers, but when data is still scarce, their marginal product is very high. Therefore, the market science level and share are locally inefficiently low, and a compensated marginal reallocation by the constrained planner toward science raises welfare in every nonterminal period.

The model's implications can also be used to re-evaluate other open issues, such as the importance of data sharing in the economy, or the consequences of data-driven automation. Contrary to the intuition that the non-rival nature of data is always beneficial, I find that high levels of data sharing can have the opposite effect on welfare. While data sharing improves static efficiency by reducing operational uncertainty, it exacerbates dynamic misallocation by making data careers relatively more attractive. Therefore, a regulator limited only to a data-sharing policy may optimally restrict data sharing to preserve scientific talent, accepting higher operational uncertainty to maintain innovation capacity.

Similarly, the model provides a new channel through which data-driven automation can deepen talent misallocation. The standard displacement channel remains intact: automation reduces the marginal product of routine workers, pushing them out of the routine occupation. The novelty comes from an information channel: in information scarcity, data accumulation also raises the marginal product of data engineers, exacerbating the flow toward data rather than science occupations.

I bring the model to the data using individual-level career histories from Revelio Labs covering 35.1 million US workers between 2000 and 2022, complemented with the American Community Survey (ACS) and firm-level data on software use from Enlyft, a commercial business-to-business account-intelligence firm. I collect the empirical results in three facts. Fact 1 is that past scientific activity predicts subsequent data activity. Within firms (both public and private) and local labor markets, increases in science employment are followed by increases in data employment. Among public firms, science employment in 2005-2009 is an excellent predictor of the size of the firm's software stack in 2022, an observable measure of its capacity to process and use data. Firms with larger software stacks, in turn, employ a larger share of data workers.

Fact 2 indicates that the expansion of data work occurs primarily within existing labor markets and disproportionately in historically innovative ones. In Revelio, high-innovation

markets account for 80% of the national increase in the data share; in the ACS, they make the largest contribution at 48%. At the same time, the private return to data work rises substantially. Among mathematics and statistics graduates, the share entering data occupations in the first year after graduation increases from 3.3% for the 2000 cohort to 8.1% for the 2022 cohort, while the corresponding data wage premium rises from approximately zero to 49 log points.

Fact 3 highlights that data has become increasingly connected to the existing scientific talent pool. Among high-skill occupation switchers, the representation of scientists entering data rises from about 0.45 of its eligible-population benchmark in 2005 to approximately one by 2022. This increase is specific to data: science's representation among entrants into other major high-skill occupations (such as engineering or finance) is flat or declining. The underlying flows also become quantitatively larger. Annual science-to-data transitions rise from 0.29% of science employment in 2005-2010 to 0.96% in 2017-2022. Together, these three facts are consistent with the model's central mechanism: scientific activity precedes the expansion of data-intensive capacity, data work becomes increasingly attractive, and data becomes a more important destination for scientific talent.

The main contribution of this paper is to study how the data economy impacts talent allocation. It is hard to dispute that data has become a fundamental input in a firm's production process. This development has not left economists indifferent and has spurred important work on how firms in a data economy grow (Begenau, Farboodi, and Veldkamp 2018; Farboodi et al. 2019), compete (Prat and Valletti 2022; Eeckhout and Veldkamp 2025), deal with privacy concerns (Acemoglu et al. 2022b; Bergemann, Bonatti, and Gan 2022; Bergemann and Bonatti 2024), or contribute more generally to aggregate output (Benzell and Brynjolfsson 2019; Jones and Tonetti 2020; Farboodi and Veldkamp 2025; Asriyan and Kohlhas 2026). Yet, to the best of my knowledge, there is no contribution exploring the interaction between substantial data accumulation and occupational choices. This perspective also delivers a different source of inefficiency: while much of the literature emphasizes distortions in consumer markets (e.g. Kehoe, Larsen, and Pastorino 2018; Acemoglu et al. 2022b; Farboodi, Haghpanah, and Shourideh 2025; Asriyan and Kohlhas 2026; Farboodi, Kondor, and Kurlat 2026), this paper shows how inefficiencies arise in the market for talent. The paper thus provides a framework to think about the allocation of talent in an economy where ideas and data are two essential stocks for production.

I complement this vantage point with novel empirical evidence on the reallocation from science to data activities. The empirical analysis relates to recent work documenting the movement of scientific talent from universities toward industry, particularly in computer science and artificial intelligence (Gofman and Jin 2024; Akcigit et al. 2026). I enrich this evidence by taking a broader view of the economy and of the science-data margin. Using worker career histories, I study occupational transitions across firms, industries, and local labor markets and document the growing connection between scientific activity, data-processing capacity, and data occupations.

*Related Literature.* This paper is mainly related to three strands of the literature. First and foremost, this paper pertains to a growing literature examining the increasing use of data in contemporary markets and its implications (see Farboodi and Veldkamp 2023; Veldkamp and Chung 2024, for a review). More precisely, it connects strongly to contributions putting at the center of the reasoning a data feedback loop: more data raises output, which in turn generates more data (see Farboodi and Veldkamp 2025, and references therein). The novelty of this paper is to add an occupational-choice margin to this feedback loop. As the data stock grows, it changes the relative return to the workers who process it, so data accumulation affects not only firms' production decisions but also the composition of the labor force.

Second, this paper belongs to a tradition exploring how talent is attracted to occupations with high private returns but lower social returns. A consequence of this mismatch is the misallocation of talent: a planner would raise welfare by reallocating some agents to different occupations. Most of this literature has focused on law and financial services (Baumol 1990; Murphy, Shleifer, and Vishny 1991; Philippon 2010; Axelson and Bond 2015; Kurlat 2019; Glode and Ordoñez 2025). It is no coincidence that this paper has a similar flavor to those contributions. For instance, the parallel between data scientists and financiers has become a common representation, exemplified by Patil and Davenport (2012) when they say that "data scientists today are akin to Wall Street 'quants' of the 1980s and 1990s". What differentiates my model from this past work is that the misallocation of talent does not stem from rent-seeking or rent extraction behavior. Instead, it arises because markets do not price the ideas and data created by scientific work. Further, in contrast to models with positive feedback loops between talent and technological progress (e.g. Murphy, Shleifer, and Vishny 1991), my model shows that the opposite can occur: a highly productive pool of scientists today announces a diminished science sector tomorrow, as talent is drawn to exploit the newly produced data stock.

Third, this paper speaks to the endogenous growth literature by examining under which conditions the accumulation of data is beneficial or detrimental to welfare. As in Jones and Tonetti (2020), Cong, Xie, and Zhang (2021), Cong et al. (2022), and Farboodi and Veldkamp (2025), data is a valuable input to economic activity. Notwithstanding, by highlighting the competing accumulation of innovation and data in the economy, I put forward a tension between innovative and predictive activities. Data accumulation attracts talent toward a productive occupation while leaving the dynamic contribution of science unpriced. This ultimately leads to persistent misallocation of talent and lower welfare.

*Outline.* Section 2 introduces the model. Section 3 characterizes the decentralized market equilibrium. Section 4 analyzes the constrained planner allocation, compares it to the market equilibrium, and discusses implications on the allocation of talent and economic output. Section 5 revisits two open issues, data sharing and automation, through the lens of the model. Section 6 presents some empirical evidence. Section 7 concludes. All proofs are in the Appendix Section A.

## 2. The Model

There are three periods $t = \{0, 1, 2\}$. Each period, a generation of mass $L$ agents is born, chooses an occupation, works for one period, and then retires. There are $N$ identical firms operating in the economy, and agents work in one of these firms. Figure 2 presents the main elements of the model, which I describe in detail in this section.

### 2.1. The Firm's Problem

I begin by describing the problem faced by firm $j$. The firm's objective each period is to maximize its value $v_{j,t}$. To do so, it must answer three questions each period, which correspond to three fundamental functions: "what can we make?" (production), "what should we make?" (innovation), and "how do we convert production into profit"? (prediction). These functions map to three occupations. Routine workers ($R$) provide capacity. Scientists ($S$) improve quality through innovation. Data engineers ($D$) reduce operational uncertainty through prediction.

Since there are $N$ identical firms in the economy and workers are mobile across firms, in a symmetric equilibrium, each firm hires an equal share of the workforce. Let $S_t$, $D_t$, and $R_t$ denote the economy-wide totals of scientists, data engineers, and routine workers, respectively, where $S_t + D_t + R_t = L$. Then firm $j$ employs $S_{j,t} = S_t/N$ scientists, $D_{j,t} = D_t/N$ data engineers, and $R_{j,t} = R_t/N$ routine workers. The firm chooses how many workers of each occupation to hire, trading off their marginal products against wages. I now develop the formal expression of the firm's value by introducing each occupation's role sequentially: first, routine workers; then scientists; and finally, data engineers. This corresponds to the "demand side" component of Figure 2.

*Production.* Routine workers, $R_{j,t}$ determine production capacity. They generate the firm's output $Y_{j,t}$ by following a set of defined rules or procedures.[2] Thus, they set the scale of the firm. For example, in a phone factory, routine workers are those who assemble the phones on the production line; in a hospital, those who follow clinical protocols; in agriculture, those who operate harvesting equipment. I denote by $\psi \in (0, 1)$ the returns to scale in routine work capacity.

*Innovation.* Scientists enhance the quality, $Q_{j,t}$, of output through innovation. In this model, innovation drives up quality because it leads to more efficient production methods, a form of process innovation. Scientists' production depends on an ideas stock, $I_t$, that is common to all firms, and evolves as

$$I_{t+1} = (1 - \delta_I) I_t + \chi S_t,$$

[2] To fix ideas, one can think of routine workers as in the frameworks of Autor, Levy, and Murnane (2003) and Autor and Dorn (2013).

with $\delta_I \in (0,1)$ a depreciation parameter and $\chi > 0$ representing how many units of codified basic knowledge one scientist adds to the public stock in one period. In line with Romer (1990), scientists' innovations build on existing ideas, which can depreciate and can be transmitted with varying degrees of efficiency. These ideas are non-rival, and thus can potentially increase the productivity of all scientists. Let $\beta \in (0,1)$ denote the returns to scale parameter for scientists and $\zeta \in (0,1)$ measure the spillover elasticity of current research productivity with respect to the existing innovation stock. Therefore, output is simply

$$Y_{j,t} = Q(I_t, S_{j,t})R_{j,t}^{\psi} = S_{j,t}^{\beta} I_t^{\zeta} R_{j,t}^{\psi}.$$

I assume that $\beta + \zeta < 1$. This ensures diminishing returns in quality and prevents explosive "strong scale effects" when new ideas feed back into future research. Similarly, I assume that $\beta + \psi < 1$. This generates decreasing returns to scale in the rival inputs used in production.

At this point, the firm faces a standard production-innovation trade-off: hire more scientists to raise quality, or hire more routine workers to raise scale. The next block introduces a new margin.

*Prediction.* Higher output is valuable only if the firm can sell it profitably. To do so, the firm relies on *forecasts* to make operational decisions.[3] For example: "which suppliers to use?", "how much inventory to hold?", "which customers to target?". These decisions can be modeled as choosing an action $a_{j,t}$ when there is an unknown optimal action $\theta_{j,t}$ that maximizes profit given an output level. If the firm chooses $a_{j,t}$ when the optimal is $\theta_{j,t}$, it gets *gross* profit

$$\pi_{j,t}^{G} = Y_{j,t}(1 - \tau(a_{j,t} - \theta_{j,t})^2). \tag{1}$$

This expression tells us that profitable output is lost according to a quadratic loss $(a_{j,t} - \theta_{j,t})^2$. A perfect decision, that is $a_{j,t} = \theta_{j,t}$, implies that $\pi_{j,t}^{G} = Y_{j,t}$. Yet, wrong decisions mean that the firm loses output proportionally to how far off that decision is from the optimal one. $\tau$ governs how costly these mistakes are. Appendix Section C.2 presents a possible microfoundation for the loss function.

How does the firm choose its action? The firm doesn't know $\theta_{j,t}$ but it is Bayesian. It can therefore form a belief on how $\theta_{j,t}$ is distributed. I assume that it has a prior belief that $\theta_{j,t} \sim N(\mu_0, \sigma_0^2)$. The best the firm can do is set $a_{j,t} = \mu_0$. This represents its best guess. The variance

[3] On top of the recent evidence indicating that the use of data in the service of forecasting improves firms' productivity (see, for example Brynjolfsson and McElheran 2016, as well as other references cited in the introduction), note that this is a very old intuition in economics. For example, Bastiat writes "commerce, I say, is led by its own interests to study the seasons, to give daily statements of the state of the crops, to receive information from every part of the globe, to foresee wants, to take precautions beforehand" (Bastiat 1850). Similarly, Fisher states that "The sagacious business man represents the other extreme; he is constantly forecasting. Many great corporations, banks, and investment trusts today maintain statistical departments largely for the purpose of gauging the future developments of business" (Fisher 1930).

$\sigma_0^2$ represents the firm's uncertainty. Thus, the firm's *expected* gross profit is

$$\mathbb{E}[\pi_{j,t}^G] \equiv \pi_{j,t}^E = Y_{j,t}[1 - \tau\,\mathbb{E}[(a_{j,t} - \theta_{j,t})^2]] = Y_{j,t}[1 - \tau\sigma_0^2].$$

This expression shows explicitly how high variance $\sigma_0^2$ reduces profits. I assume $0 < \tau\sigma_0^2 < 1$. The upper bound says that baseline uncertainty is not so costly that it wipes out the firm's expected profit entirely; the lower bound ($\tau > 0$) says that forecast errors are genuinely costly, so that prediction has strictly positive value.

The role of data engineers in the firm is *to make predictions* to mitigate the costs associated with uncertainty. They do so by processing the *aggregate data stock* $\Omega_t$ and extracting signals about $\theta_{j,t}$. The intuition runs as follows. The aggregate data stock $\Omega_t$ contains raw information, such as transaction records or customer interactions *from across the economy*. Because firms face states $\theta_{j,t}$ that share common components, such as industry demand, technology trends, or input market conditions, data generated by other firms remains informative about firm $j$'s optimal decisions. Data engineers extract $n_t$ useful signals, $s_{k,t}$, from this data with $k \in \{1, \cdots, n_t\}$. Let $\alpha \in (0,1)$ be the returns to scale parameter for data engineers, then the number of signals extracted[4] depends on both the data engineers and data stock, such that

$$n_t = D_{j,t}^{\alpha}\Omega_t^{1-\alpha}.$$

Higher data stocks enable data engineers to make better predictions by building, for example, more sophisticated algorithms and models. I assume constant returns to scale in $n_t$ in line with Abis and Veldkamp (2024). This can be justified by analogy to the growth literature with idea production, where ideas display constant or even diminishing returns to scale.

Let me now expand on how data engineers reduce uncertainty by linking their actions to the firm's expected gross profit. Each signal they observe is noisy. Thus, $s_{k,t} = \theta_{j,t} + \epsilon_{k,t}$, where I assume that $\epsilon_{k,t} \sim N(0, \sigma_\epsilon^2)$. The $n_t$ independent signals can be combined into a single sufficient statistic, their average

$$\bar{s}_t = \frac{1}{n_t}\sum_{k=1}^{n_t} s_{k,t} = \theta_{j,t} + \frac{1}{n_t}\sum_{k=1}^{n_t}\epsilon_{k,t}.$$

The noise term $\frac{1}{n_t}\sum_{k=1}^{n_t}\epsilon_{k,t}$ has variance $\frac{\sigma_\epsilon^2}{n_t}$. So I can write $\bar{s}_t \sim N(\theta_{j,t}, \frac{\sigma_\epsilon^2}{n_t})$. By Bayes' rule, and using the fact that for normal distributions, the posterior precision (inverse variance) equals the sum of prior precision and data precision, the posterior variance, $\sigma_{post}^2$, is then

$$\sigma_{post}^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n_t}{\sigma_\epsilon^2}} = \frac{\sigma_0^2\sigma_\epsilon^2}{\sigma_\epsilon^2 + n_t\sigma_0^2} = \frac{\sigma_0^2\sigma_\epsilon^2}{\sigma_\epsilon^2 + D_{j,t}^{\alpha}\Omega_t^{1-\alpha}\sigma_0^2}.$$

[4] $n_t$ does not need to be an integer and should be interpreted as an effective sample-size or precision index, rather than as the literal cardinality of a finite collection of signals.

Therefore, with data engineers reducing variance from $\sigma_0^2$ to $\sigma_{post}^2$, expected gross profit becomes

$$\pi_{j,t}^E = Y_{j,t}[1 - \tau\sigma_{post}^2(D_{j,t}, \Omega_t)].$$

The term $[1 - \tau\sigma_{post}^2]$ directly shows that if $D_{j,t} = 0$, then we are back to baseline uncertainty: $[1 - \tau\sigma_0^2]$. Conversely, as $D_{j,t} \to \infty$, the expected gross profit approaches perfect information. Therefore, without data engineers, the firm operates at baseline efficiency, leaving money on the table. Data engineers thus convert raw output into realized profit by reducing operational mistakes.

*Data Accumulation.* Where does the aggregate data stock $\Omega_t$ come from? A central element of the model lies in the way the aggregate data stock is generated in the economy. There are two key features I aim to capture: the fact that data is a byproduct of economic activity (an externality) and that it is, to some extent, non-rival (Jones and Tonetti 2020; Farboodi and Veldkamp 2025).

First, data accumulates as a byproduct of the firm's economic activity. In my setting, the driver of data accumulation is quality, for two reasons. First, it separates the technical margin that raises quality (scientists) from the capacity margin (routine workers) so that staffing for production does not mechanically generate more data. This sharpens the analysis by making the race between scientists and data engineers explicit, so that comparative statics and dynamics are not confounded by changes in routine labor. Second, it emphasizes the intuition that higher-quality products create new dimensions of data that were previously uncollectable. For example, as smartphones improved through R&D, because of faster chips, better cameras, more reliable sensors, or smoother apps, users expanded what they could do on the device, such as taking photos, navigating the city, making payments, or health logging. Each new use case produces distinct, higher-quality signals, so rising product quality endogenously widens and deepens the data the firm can observe: a flywheel effect.[5] Therefore, at the *firm level*, the firm accumulates a data stock at each period, $\omega_{j,t}$ according to

$$\omega_{j,t+1} = (1 - \delta_\omega)\omega_{j,t} + \gamma Q_{j,t},$$

with $\delta_\omega \in (0,1)$ a depreciation parameter and with $\gamma > 0$ representing the rate at which quality generates new data. For instance, $\gamma$ might govern how much data the firm can mine from its customers. For robustness, Appendix Section D.1 considers an extension where the driver of

[5] A real-world example of the flywheel effect is Recursion Pharmaceuticals, a biotech company specializing in drug discovery. Its 10-K filings describe automated laboratories that run more than two million experiments per week, and the data they produce is a byproduct of that search. The firm's proprietary dataset grew from half a petabyte in 2017 to roughly sixty-five in 2024. The firm calls this process a "flywheel effect, where data generation and learning accelerate side-by-side". According to its chief executive: "Today, our wet labs primarily generate the data needed to train AI models." Interestingly, Recursion reports its workforce in three categories - life sciences, technology, and strategic operations - which can be directly mapped to the three occupations in this model. Between 2020 and 2025, the technology share rose from thirty-five to forty-three percent of employees while the life-science share fell from forty to thirty-five.

data accumulation is output instead of quality.[6]

Second, data is nonrival, to some degree. Thus, I assume that the data stocks of all other firms can also be used by firm $j$ simultaneously without reducing the total amount of available data, subject to a data sharing parameter. A simple motivation for this assumption is that $\theta_{j,t}$ might reflect industry-wide demand conditions, shared technological opportunities, or common input price shocks that affect all firms simultaneously. When many firms face similar underlying states, observing aggregated patterns from the broader market improves each firm's inference about its own position. Therefore, the aggregate data stock can be expressed as

$$\Omega_t = \omega_{j,t} + \phi \sum_{k \neq j} \omega_{k,t} = \omega_t(1 + \phi(N-1)),$$

where the second equality comes from the fact that, in the symmetric equilibrium, all firms have the same data stock, $\omega_{k,t} = \omega_t$ for all $k$. $\phi \in [0,1]$ governs the degree of data sharing in the economy. I will assume in the simple model that $\phi$ is exogenous. Formally, I interpret $\Omega_t$ as the *effective* stock of signals about the firm's own state: the common components of $\theta_{j,t}$ across firms make outside data informative about firm $j$'s state, and $\phi$ measures the fraction of other firms' data that translates into usable own-state signals, so that the extracted signals can be treated as independent and unbiased in the updating formula above. Throughout the paper, I write $\Phi \equiv 1 + \phi(N-1)$, so that $\Omega_t = \Phi\, \omega_t$.

*The Firm's Value.* Bringing the three blocks together, the firm's value depends on its workforce, the accumulated ideas, $I$, and the aggregate data stock, $\Omega$. On the cost side, the firm pays wage $w^o_{j,t}$ to each occupation $o \in \{R, S, D\}$ at time $t$. Therefore, the firm's problem is to maximize its value $v_{j,t}$, or expected net profit, by choosing the optimal number of routine workers, scientists, and data engineers period-by-period without capitalizing on the effects of current employment on future ideas or data stocks. This corresponds to

(2)
$$\max_{R_{j,t}, S_{j,t}, D_{j,t}} v_{j,t} = \underbrace{\underbrace{S_{j,t}^{\beta} I_t^{\zeta}}_{\text{science block}} \left[ 1 - \tau \underbrace{\frac{\sigma_0^2 \sigma_\epsilon^2}{\sigma_\epsilon^2 + D_{j,t}^{\alpha} \Omega_t^{1-\alpha} \sigma_0^2}}_{\text{data block}} \right] \underbrace{R_{j,t}^{\psi}}_{\text{production block}}}_{=\pi^E_{j,t}} - w^R_{j,t} R_{j,t} - w^S_{j,t} S_{j,t} - w^D_{j,t} D_{j,t}.$$

Throughout the paper, I assume that $\alpha + \beta + \psi \leq 1$. Mathematically, this assumption delivers global concavity of $\pi^E$ in $(S, D, R)$ (see Lemma A4 in the Appendix), so the firm's first-order

[6] Specifying $\omega_{j,t+1} = (1 - \delta_\omega)\omega_{j,t} + \gamma Y_{j,t}$ trades realism for transparency. It is arguably the more empirically plausible law of motion, since routine production also generates data. It also yields stronger predictions concerning the direction of misallocation, as now both scientists and routine workers generate dynamic spillovers. However, once routine workers move the data stock, the channel through which innovation cannibalizes itself is harder to isolate. The quality-based specification helps to keep that channel in view.

conditions characterize its global optimum and the competitive-equilibrium concept below is fully grounded. Economically, some decreasing returns in science and routine production are necessary to support an active, competitively priced data occupation. Appendix Section C.1 develops this point and shows that the stronger restriction above guarantees positive firm value at an interior optimum.

*Firm Ownership.* To close the firm's side of the model, I now specify its ownership. Firm value is strictly positive at any interior candidate. I assign it to a long-lived owner who holds the representative firm, consumes the residual, and has preferences

$$U^K = \sum_{t=0}^{2} \rho^t c_t^K, \tag{3}$$

with $\rho \in (0,1]$, $c_t^K$ her consumption at time $t$, and $U^K$ her discounted utility. Therefore, the owner is a passive residual claimant: she receives the firm's realized profits in every period but does not choose its employment policy. The reduced-form separation between ownership and control may capture short managerial horizons or incomplete contracting over the long-run consequences of current employment decisions, including their effects on future ideas, data, and firm value.

The owner also receives an exogenous endowment $e_t^K > 0$ of the numeraire, so that her budget reads $c_t^K \leq e_t^K + \pi_t^E - w_t^S S_t - w_t^D D_t - w_t^R R_t$. In the market equilibrium, she consumes

$$c_t^K = e_t^K + \pi_t^E - w_t^S S_t - w_t^D D_t - w_t^R R_t = e_t^K + v_t > 0. \tag{4}$$

Therefore, being a firm owner is not an occupational choice *per se*: workers choose occupations, earn wages, and an outside owner receives the residual. The endowment plays no economic role. It only guarantees positive owner consumption in a neighborhood of the market allocation. Since it is common across allocations, it cancels from every local welfare comparison. Because all quantities are in per-firm units, $c_t^K$ is per-firm owner consumption. With $N$ identical firms, aggregate owner consumption is $Nc_t^K$. This rescaling leaves every result unchanged.[7]

### 2.2. Occupational Choice

Having described the firm's demand for three occupations of workers, I now model the supply side: how agents choose among routine work, science, and data engineering. This occupational choice mechanism will determine how talent reallocates in response to data accumulation. This

[7]The residual must accrue to an agent outside the occupational-choice problem. Anticipating a bit on what follows, if profits were rebated lump sum to workers, consumption in occupation $o$ would be $w^o + T$. Agents would then compare $\ln(w^o + T) + \eta^o$, so wage premia alone would no longer summarize occupational choice. Linearity in equation (3) also makes the value of one unit of date-$t$ resources equal to $\rho^t$. With concave owner utility, marginal utilities would weight the stock values. The local welfare argument in Section 4 holds when those weights are positive.

corresponds to the "supply side" component of Figure 2.

I model occupational choice as a discrete choice with *correlated* idiosyncratic preference draws, rather than the simpler multinomial logit, for one chief reason: logit imposes independence of irrelevant alternatives, so the occupations that lose workers when the data premium rises would be pinned down by the initial shares alone. Using correlated preference draws prevents this outcome. Science and data engineering plausibly appeal to the same people: both reward quantitative reasoning, analytical thinking, and comfort with abstraction, so an agent drawn to science is more likely to be drawn to data engineering than to routine work. The intensity of this substitution depends on how correlated the draws are, which the joint normal specification below captures through a single parameter.

*Setup.* At the start of each period, agents choose between becoming routine workers ($R$), scientists ($S$), or data engineers ($D$). Let $o \in \{R, S, D\}$ index the three occupations. In what follows, I omit time subscripts except where necessary for clarity.

Agents choose occupations on two criteria: what the occupation pays, and their specific idiosyncratic preferences for a given occupation. Consider pay first. Let $w^o$ denote the wage in occupation $o$. Agents live one period and do not save, so an agent working in occupation $o$ consumes her wage, $c = w^o$. Wages are her entire income: the residual value of the firm accrues to the firm owner, who takes no part in the occupational-choice problem.

Second, agents differ in idiosyncratic occupational preferences across non-routine occupations. Let $\eta_i^S$ and $\eta_i^D$ capture agent $i$'s idiosyncratic taste for science and data engineering *relative to routine work* (for example, interest in research or applied work, tolerance for uncertainty, desire for meaning). These draws admit two readings that I leave unmodeled because neither affects a worker's marginal product: a non-pecuniary preference over occupations, or a fixed cost of acquiring occupation-specific skills. Under either reading, what matters is that the draws are correlated across the two quantitative occupations. Agent $i$ has a comparative advantage in science over data engineering when $\eta_i^S - \eta_i^D$ is large, a comparison in net terms rather than in productivity: all workers assigned to an occupation are equally productive in it. An agent of type $\eta_i$ who works in occupation $o$ and consumes $c$ therefore obtains

$$u(\eta_i, o, c) = \ln c + \eta_i^o, \qquad \eta_i^R \equiv 0, \tag{5}$$

which, evaluated at her wage, is $\ln w^o + \eta_i^o$.

Because only utility differences matter, I measure utilities relative to routine work, the outside option. Subtracting $\ln w^R$, a constant common to all three options, and defining the wage premium of $o$ relative to routine work,

$$\tilde{w}^o \equiv \ln\left(\frac{w^o}{w^R}\right), \qquad \tilde{w}^R = 0,$$

delivers the equivalent representation

$$U_i^R = 0, \qquad U_i^S = \tilde{w}^S + \eta_i^S, \qquad U_i^D = \tilde{w}^D + \eta_i^D.$$

The agent chooses the occupation with the highest utility. The two representations rank the three options identically, so occupation shares and equilibrium wages may be computed from either. The premium representation is useful to think about the decentralized market equilibrium in Section 3: it reflects the market return to each occupation, determined in equilibrium by the firm's marginal products, a signal for workers to choose the occupation delivering the highest utility. The consumption representation is rather helpful to think about the constrained planner allocation in Section 4: it states the compensation required to preserve a worker's market utility.

*Ability Distribution.* I assume that the idiosyncratic characteristics $(\eta_i^S, \eta_i^D)'$ are jointly normal

$$\begin{pmatrix} \eta_i^S \\ \eta_i^D \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_S^2 & \nu\sigma_S\sigma_D \\ \nu\sigma_S\sigma_D & \sigma_D^2 \end{pmatrix} \right).$$

I assume that $\nu \in (-1, 1)$, with $\sigma_S, \sigma_D > 0$, so that the distribution is nondegenerate. The key economic parameter is $\nu$: the correlation between the science and data engineering preference draws. It governs the intensity of talent reallocation between $S$ and $D$ as wages change. To fix ideas, consider a marginal increase in the data premium $\tilde{w}^D$, holding $\tilde{w}^S$ fixed. The new data engineers are drawn from two adjacent margins: workers previously indifferent between science and data, and workers previously indifferent between routine work and data. Both margins carry strictly positive mass for any $\nu \in (-1, 1)$, including $\nu = 0$, so a rising data premium always draws *some* talent out of science. Thus, paying close attention to $\nu$ in the decentralized equilibrium is central to understanding where the flows to data engineering come from.

*Occupation Shares.* The share of agents choosing each occupation is determined by the probability that it delivers the highest utility. It follows that

$$\begin{aligned} \xi^S &= \Pr(\tilde{w}^S + \eta_i^S > 0 \text{ and } \tilde{w}^S + \eta_i^S > \tilde{w}^D + \eta_i^D) \\ \xi^D &= \Pr(\tilde{w}^D + \eta_i^D > 0 \text{ and } \tilde{w}^D + \eta_i^D > \tilde{w}^S + \eta_i^S) \\ \xi^R &= 1 - \xi^S - \xi^D. \end{aligned}$$

These shares do not have closed-form expressions due to the bivariate normal structure, but they can be computed through numerical integration over the appropriate regions of $(\eta_i^S, \eta_i^D)$ space. The occupation shares translate directly into labor supply quantities

$$S_t = \xi_t^S L, \qquad D_t = \xi_t^D L, \qquad R_t = \xi_t^R L.$$

*Equilibrium Determination.* Equilibrium wages and occupation shares are determined jointly. The firm's first-order conditions yield wage schedules $\{w_t^S, w_t^D, w_t^R\}$ as functions of occupation quantities and state variables $(I_t, \Omega_t)$. These wages generate log premiums $\{\tilde{w}_t^S, \tilde{w}_t^D\}$, which determine occupation shares $\{\xi_t^S, \xi_t^D, \xi_t^R\}$ through the occupational-choice block. Occupation shares then determine quantities $\{S_t, D_t, R_t\}$, which feed back into wages. A rational expectations equilibrium is a fixed point of this mapping.

*Normalization.* Since the $N$ firms are identical and the equilibrium is symmetric, I express all labor quantities in per-firm units from here on: I relabel $(S_t/N, D_t/N, R_t/N)$ as $(S_t, D_t, R_t)$ and the per-firm labor mass $L/N$ as $L$, so that the labor constraint reads $S_t + D_t + R_t = L$ in per-firm units and the occupation shares $\xi_t^o$ are unchanged (they are population shares, common to the economy and to each firm by symmetry). Ideas are produced by the economy-wide scientist pool, so under this relabeling, the law of motion $I_{t+1} = (1-\delta_I)I_t + \chi S_t$ holds with $\chi$ absorbing the factor $N$. Since $\chi$ is a free parameter, no result depends on this constant. The firm-level data law of motion is already in per-firm units. With this convention, every first-order condition, wage, and posterior variance below is a statement about the representative firm, and the only remaining role of $N$ is in the aggregation map $\Omega_t = \Phi\omega_t$ with $\Phi = 1 + \phi(N-1)$.

### 2.3. Timeline and Sequence of Play

This section specifies the timeline governing how today's occupational choices affect tomorrow's opportunities through the accumulation of ideas and data stocks. The model unfolds over three periods, $t = \{0, 1, 2\}$, each consisting of two stages.

- **Stage 1**: Generation $t$ is born, observes ideas $I_t$ and data $\Omega_t$, forms expectations of wages, and chooses occupations;
- **Stage 2**: The firm chooses its mix of routine workers, scientists, and data engineers to produce. This generates ideas $I_{t+1}$ and data $\Omega_{t+1}$.

Stage 1 choices are equilibrium *supplies*: agents choose occupations given correctly anticipated wages. Stage 2 describes the firm's demand side: taking wages as given, the firm's first-order conditions determine its demanded mix. In a rational expectations equilibrium, the two coincide. The "choice" in Stage 2 is the demand consistent with the occupational supplies formed in Stage 1.

To initialize the model, I assume the economy begins with a positive quantity of ideas but no data. Thus, $I_0 > 0, \omega_0 = 0, \Omega_0 = 0$. With $I_0 > 0$, scientists are immediately productive, so the economy can begin innovating. But with $\Omega_0 = 0$, data engineers cannot extract any signals, making them worthless in period 0. This creates a natural starting point: the economy begins in a pure science-and-production equilibrium, with data engineering emerging endogenously only after innovation has generated a data stock. The assumption $I_0 > 0$ can be interpreted as inherited ideas from previous generations or as a reduced-form representation of an earlier

phase where innovation accumulated exogenously. The main point is that innovation comes first, then data.

*Solution concept.* For each period, I solve the within-period fixed point for a given state, update the state using the laws of motion, and iterate forward. I *explicitly* solve in the main text periods 0 and 1 in Sections 3.1 and 3.2, respectively. Period 2 is solved analogously to period 1, with the existence and uniqueness of the equilibrium formalized in Appendix Proposition A1. Rather than re-deriving the intra-period equilibrium, Section 3.3 focuses on comparative dynamics: it characterizes how the equilibrium allocation responds to changes in the data stock.

## 3. Market Equilibrium

This section studies the market equilibrium dynamics. It describes the economy's endogenous path when decisions are made by agents and firms. The main lesson is that the economy follows a pattern of endogenous structural change, where the economy evolves from an innovation economy to a data-intensive economy. This trajectory rests on a perhaps counterintuitive form of path dependence. An initially large scientist pool leads to a larger data stock in the future. Thus, high innovation in the beginning cannibalizes innovation later, because agents are lured to the data sector that scientists helped to produce.

### 3.1. Period 0: Initial Innovation

Period 0 is characterized by being a pure innovation phase. With no data stock yet accumulated ($\Omega_0 = 0$), data engineers are unproductive, so the economy operates as a standard two-sector model with scientists and routine workers.

#### 3.1.1. Stage 2: Firm's Production Decision

To produce, the firm takes as given the ideas and data stock, as well as the agents' occupational choices. Since $\Omega_0 = 0$, data engineers are unproductive and the firm's problem in equation (2) reduces to choosing scientists and routine workers only. Taking the first order conditions (FOC) with respect to routine workers and scientists, and rearranging yields the scientist wage premium

$$\frac{w_{j,0}^S}{w_{j,0}^R} = \frac{\beta}{\psi}\frac{R_{j,0}}{S_{j,0}}.$$

The premium is decreasing in the scientist-to-routine-worker ratio: more scientists lower the marginal product of an additional scientist due to diminishing returns ($\beta < 1$). This creates a downward-sloping demand curve for scientists, which will intersect with the upward-sloping occupational supply curve to determine equilibrium.

### 3.1.2. Stage 1: Occupation Choice Equilibrium

Agents in generation 0 must form expectations about equilibrium wages and choose occupations accordingly. Since data engineers are unproductive, the choice reduces to becoming a scientist or a routine worker. Let $\xi_0^S$ denote the share choosing science. Then in a symmetric equilibrium, $S_0 = \xi_0^S L$ and $R_0 = (1-\xi_0^S)L$.

Occupational choice generates an occupation supply curve. Given wages $\{w_0^S, w_0^R\}$, the log premium is $\tilde{w}_0^S = \ln(w_0^S/w_0^R)$, and the share choosing science is

$$\xi_0^S = \Pr(\tilde{w}_0^S + \eta_i^S > 0) = \Upsilon\left(\frac{\tilde{w}_0^S}{\sigma_S}\right),$$

where $\Upsilon(\cdot)$ is the standard normal CDF. This supply curve is increasing in the wage premium: higher relative wages for scientists attract more workers into science.

At the equilibrium, the scientist share demanded by firms (via their FOCs) equals the share supplied by agents (via occupational choice). The following proposition establishes that such an equilibrium exists and is unique.

PROPOSITION 1 (Period 0 Equilibrium). *Given initial conditions with a positive ideas stock, $I_0 > 0$, and a non-existent aggregate data stock, $\Omega_0 = 0$, there exists a unique equilibrium with occupation shares for scientists, $(\xi_0^S)^* \in (0,1)$, data engineers, $(\xi_0^D)^* = 0$, and routine workers, $(\xi_0^R)^* = 1 - (\xi_0^S)^*$.*

The equilibrium in Proposition 1 represents a "pure innovation phase". All talent allocation occurs between scientists and routine workers. Yet, this period generates two stocks that will reshape future opportunities. Scientific employment produces ideas, raising the productivity of future scientists. More subtly, scientists also generate data. Quality produces firm data, which is then aggregated across firms. The model's core tension emerges from this second channel. A large scientist pool in period 0 - the outcome that perhaps naively looks desirable from an innovation standpoint - creates a large data stock in period 1. This data stock makes data engineers productive, raising their wage premium and pulling talent away from science. The following period's analysis develops this idea.

## 3.2. Period 1: Emergence of Data Processing

Period 1 marks the transition from a pure innovation economy to a mixed economy where prediction competes with innovation for talent. This period elaborates the model's central mechanism: how data accumulation drives endogenous structural change in the composition of the workforce.

### 3.2.1. Stage 2: Firm's Production Decision

With $\Omega_1 > 0$, data engineers become productive. The firm's problem in equation (2) now features all three occupations. This has two immediate effects. First, hiring data engineers reduces $\sigma^2_{post}$ below the baseline $\sigma^2_0$, raising expected profit. The firm has a positive marginal product for data engineers and will demand $D_1 > 0$ at any finite wage. Second, the firm can now improve profitability through three channels: hire scientists to raise quality $Q$ (the innovation margin), hire routine workers to raise scale $Y$ (the capacity margin), or hire data engineers to reduce operational uncertainty $\sigma^2_{post}$ (the prediction margin). These three margins compete for the same labor supply $L$.

The FOCs for scientists and routine workers parallel period 0, but with higher productivity $[1-\tau\sigma^2_{post}]$ replacing the baseline $[1-\tau\sigma^2_0]$. Better prediction raises the marginal product of routine workers and scientists. The new margin is data engineers, with FOC

$$S_{j,1}^{\beta} I_1^{\zeta} R_{j,1}^{\psi} \tau \frac{\sigma_0^4 \sigma_\epsilon^2 \alpha D_{j,1}^{\alpha-1} \Omega_1^{1-\alpha}}{[\sigma_\epsilon^2 + D_{j,1}^{\alpha} \Omega_1^{1-\alpha} \sigma_0^2]^2} = w_{j,1}^{D} = \mathrm{MP}^{D}.$$

This expression highlights two important properties of the marginal product of data engineers, $\mathrm{MP}^D$. First, the marginal product of data engineers is proportional to $S_{j,1}^{\beta} I_1^{\zeta} R_{j,1}^{\psi}$, so higher quality or scale raises the value of reducing operational uncertainty. Better predictions are more valuable when there's more profit at stake. Second, $\alpha < 1$ ensures $\partial \mathrm{MP}^D / \partial D_{j,1} < 0$. As the firm hires more data engineers, each additional one contributes less to variance reduction. This last point means the data engineer demand curve slopes down, just like the scientist demand curve. But the level of this curve depends on the current aggregate data stock, which was determined by past innovation. This is how innovation creates its own competition.

### 3.2.2. Stage 1: Occupation Choice Equilibrium

Generation 1 agents now face a three-way choice: routine work, science, or data engineering. Occupational choice generates occupation shares $(\xi_1^S, \xi_1^D, \xi_1^R)$ as functions of the wage premiums $\{\tilde{w}_1^S, \tilde{w}_1^D\}$. As before, at the equilibrium, the occupation shares demanded by firms equal the shares supplied by agents. The following proposition establishes that such an equilibrium exists and is unique.

PROPOSITION 2 (Period 1 Equilibrium). *Given ideas stock, $I_1 > 0$, and aggregate data stock, $\Omega_1 > 0$, there exists a unique equilibrium with occupation shares $(\xi_1^S, \xi_1^D, \xi_1^R)^* \in \mathrm{relint}(\Delta^2)$ where $\Delta^2 = \{(\xi^S, \xi^D, \xi^R) \in \mathbb{R}^3_+ : \xi^S + \xi^D + \xi^R = 1\}$.*

What distinguishes this equilibrium from period 0's is the data stock inherited from it. It demonstrates how talent allocation responds endogenously to accumulated data. A positive $\Omega_1$ gives data engineers a positive marginal product and hence a positive wage. The occupational-choice block turns that premium into a positive share $\xi_1^D$. The reallocation feeds back into the

stocks: a lower $S_1$ slows ideas accumulation into $I_2$, while period-1 quality $Q_1$ determines $\Omega_2$. The accumulated data stock is therefore a state variable that coordinates labor market decisions. It makes data engineering attractive precisely when and where innovation has been successful.

### 3.3. Period 2: Mature Data Economy

Period 2 addresses a critical question left open by period 1: does the reallocation toward data engineering continue indefinitely, or does it eventually plateau? To answer this question, this section introduces the *information scarcity region* $\mathcal{S}$, which partitions the state space $(D, \Omega)$ into two regimes with qualitatively different dynamics. In scarcity, more data makes data engineers more productive; in the deep-scarcity subregion, this force is strong enough to raise the data *premium*, sustaining talent reallocation away from science in equilibrium. In saturation, the marginal value of data engineers declines, and the flow of talent reverses. The information scarcity region therefore outlines conditions to understand the strength of the reallocation toward data.

*Equilibrium.* As in period 1, the equilibrium exists and is unique with all three occupations active. Proposition A1 in the Appendix provides a formal statement. The interesting question, nonetheless, is not existence but *dynamics*: conditional on continued data accumulation, how do occupation shares evolve from period 1 to period 2 as data continues to accumulate?[8]

*Information Scarcity and Saturation.* To understand the behavior of the data sector in period 2, recall that the signal scale is $n_t \equiv D_{j,t}^{\alpha}\Omega_t^{1-\alpha}$, so the posterior variance is

$$\sigma^2_{post}(D_{j,t}, \Omega_t) = \frac{\sigma_0^2\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2 n_t}.$$

It is convenient to express the signal scale in units of the prior precision. Define the *learning scale*

$$X(D_{j,t}, \Omega_t) \equiv D_{j,t}^{\alpha}\Omega_t^{1-\alpha}\sigma_0^2 = n_t\sigma_0^2,$$

so that $\sigma^2_{post} = \sigma_0^2\sigma_\epsilon^2/(\sigma_\epsilon^2 + X)$ and every statement below can be written in terms of this single index. Two polar cases clarify the economics. First, suppose that the data precision $n_t/\sigma_\epsilon^2$ is much smaller than the prior precision $1/\sigma_0^2$. Signals are scarce and highly valuable. The posterior variance $\sigma^2_{post} \approx \sigma_0^2$ is close to the prior, meaning the firm has learned little. In this regime, an additional unit of signal scale $n_t$ generates a large reduction in uncertainty. Whether an additional data engineer commands a high marginal product then depends on how much signal scale that engineer adds. Formally, this corresponds to the derivative $\partial n_t/\partial D_{j,t} = \alpha D_{j,t}^{\alpha-1}\Omega_t^{1-\alpha}$, which tells us that the marginal product of data engineers naturally also depends on the accumulated

[8] Data necessarily rises from period 0 to period 1 because $\omega_0 = 0$, but continued growth requires $\gamma Q_1 > \delta_\omega \omega_1$.

data stock to work with. Now, imagine that the data precision far exceeds the prior precision. Signals are abundant and the posterior variance $\sigma^2_{post} \approx \sigma^2_\epsilon / n_t$ is close to zero. The firm has learned nearly everything there is to learn; operational uncertainty has been largely eliminated. Additional data engineers yield only marginal improvements, so their marginal product is low.

The transition between these regimes occurs at the saturation threshold, where data precision equals prior precision, that is, where $X = \sigma^2_\epsilon$. This defines the *information scarcity region*

$$\mathcal{S} = \{(D,\Omega) : X(D,\Omega) < \sigma^2_\epsilon\} = \{(D,\Omega) : D < D_{\text{sat}}(\Omega)\}, \qquad D_{\text{sat}}(\Omega) = \left(\frac{\sigma^2_\epsilon}{\sigma^2_0}\right)^{1/\alpha} \Omega^{(\alpha-1)/\alpha},$$

where $D_{\text{sat}}$ traces the boundary $\{X = \sigma^2_\epsilon\}$ in $(D,\Omega)$ space. This threshold has intuitive properties. Since $\alpha \in (0,1)$ implies $(\alpha-1)/\alpha < 0$, that boundary is decreasing in $\Omega$: as more data accumulates, fewer data engineers are needed to cross it. Noisier signals push it outward; a more uncertain prior pulls it inward. The threshold is not the onset of diminishing returns to information, which set in at every $n_t \geq 0$. Holding occupation counts fixed, it is where the marginal product of a data engineer stops rising and starts falling *in* $\Omega$. Further, by "saturation" I do not mean complete learning: at the threshold $\sigma^2_{post} = \sigma^2_0/2$, so half of the prior uncertainty remains.

States $(D_{j,t}, \Omega_t) \in \mathcal{S}$ are in scarcity; states with $D_{j,t} \geq D_{\text{sat}}(\Omega_t)$ are in saturation. This partition organizes the analysis of data employment dynamics: I first characterize how the marginal product of data engineers behaves in each regime at *fixed* occupation counts in the next proposition. I then convert this value statement into an exact characterization of *equilibrium* employment in Proposition 4 below.

PROPOSITION 3 (Information Scarcity and Marginal Product of Data Engineers). *Fix the returns to scale to data engineers $\alpha \in (0,1)$, firm prior variance $\sigma^2_0 > 0$, and signal variance $\sigma^2_\epsilon > 0$. Holding occupation counts $(S_t, R_t, D_t)$ constant, with $S_t, R_t, D_t > 0$, the marginal product of data engineers is single-peaked in $\Omega$: strictly increasing in scarcity, maximal at the saturation threshold, and strictly decreasing in saturation. The elasticity is*

$$\varepsilon_{\text{MP}^D,\Omega} = (1-\alpha)\frac{\sigma^2_\epsilon - D^\alpha_{j,t}\Omega_t^{1-\alpha}\sigma^2_0}{\sigma^2_\epsilon + D^\alpha_{j,t}\Omega_t^{1-\alpha}\sigma^2_0},$$

*which is positive in scarcity, zero at the threshold, and negative in saturation.*

When signals are scarce, additional data allows each data engineer to extract more information, raising the marginal value of hiring another data engineer. This suggests a positive feedback loop: data accumulation generates a higher marginal product for data engineers, higher wages for data engineers, and more workers choosing data engineering. Proposition 3 alone, however, does not establish this loop by itself: it holds occupation counts fixed, while the equilibrium inflow into data engineering is governed by the *relative* premium $\tilde{w}^D$, and a larger data stock also raises the wages of scientists and routine workers through the shared efficiency term $[1-\tau\sigma^2_{post}]$.

Once the firm has enough signal scale, $n_t$, to reduce uncertainty close to its lower bound, additional data has limited value. The firm has already hired enough data engineers to extract the signal; more data does not justify hiring even more data engineers. The feedback loop weakens or reverses.

Formalizing this reasoning in equilibrium requires keeping track of which force dominates. In terms of the learning scale $X$, the direct informational gain to data engineers dominates the shared productivity gain to routine workers exactly when $X < \sigma_\epsilon^2\big(1-\tau\sigma^2_{post}(D,\Omega)\big)$. Lemma A1 in the Appendix shows that this implicit condition has a closed form, which defines the *deep-scarcity region*

$$\mathcal{S}^{deep} \equiv \left\{(D,\Omega) : X(D,\Omega) < \sigma_\epsilon^2\sqrt{1-\tau\sigma_0^2}\right\} \subset \mathcal{S}.$$

The inclusion is strict since $\sqrt{1-\tau\sigma_0^2} < 1$, by the maintained assumption $\tau\sigma_0^2 > 0$. The deep-scarcity boundary lies strictly inside the scarcity boundary and approaches it as $\tau\sigma_0^2 \to 0$. The following proposition provides an exact sign characterization of the equilibrium reallocation and shows that the deep-scarcity region is crossed exactly once.

PROPOSITION 4 (Cannibalization and Recovery). *Fix $t \in \{1,2\}$ and $I_t > 0$, and let $\xi^*(\Omega) = (\xi^{S*}, \xi^{D*}, \xi^{R*})(\Omega)$ denote the unique equilibrium of Propositions 2 and A1 at aggregate data stock $\Omega > 0$. Write $X^*(\Omega) \equiv X\big(L\xi^{D*}(\Omega), \Omega\big)$ for the equilibrium learning scale. Then $\xi^*(\cdot)$ is continuously differentiable on $(0,\infty)$, does not depend on $I_t$, and:*

*(i) (Exact sign characterization.) For every $\Omega > 0$,*

$$\operatorname{sign}\frac{d\xi^{D*}}{d\Omega} = -\operatorname{sign}\frac{d\xi^{S*}}{d\Omega} = -\operatorname{sign}\frac{d\xi^{R*}}{d\Omega} = \operatorname{sign}\Big(\sigma_\epsilon^4\big(1-\tau\sigma_0^2\big) - X^*(\Omega)^2\Big).$$

*In particular, whenever the equilibrium lies in the deep-scarcity region $\mathcal{S}^{deep}$, the data share strictly rises with the data stock while both the science share and the routine share strictly fall, and $d\big(\xi^{D*}/\xi^{S*}\big)/d\Omega > 0$; outside the closure of $\mathcal{S}^{deep}$, all these movements are strictly reversed.*

*(ii) (Unique threshold, single peak, single trough.) There exists a unique $\bar{\Omega} \in (0,\infty)$ such that the equilibrium lies in $\mathcal{S}^{deep}$ if and only if $\Omega < \bar{\Omega}$. Consequently, $\xi^{D*}(\Omega)$ is strictly increasing on $(0,\bar{\Omega}]$ and strictly decreasing on $[\bar{\Omega},\infty)$, with $\lim_{\Omega\downarrow 0}\xi^{D*}(\Omega) = \lim_{\Omega\to\infty}\xi^{D*}(\Omega) = 0$; and $\xi^{S*}(\Omega)$ is strictly decreasing on $(0,\bar{\Omega}]$ and strictly increasing on $[\bar{\Omega},\infty)$. The threshold $\bar{\Omega}$ is the same in periods 1 and 2, and the peak of the data share occurs strictly inside the scarcity region $\mathcal{S}$ of Proposition 3.*

Proposition 4 formalizes the feedback loop reasoning when taking into consideration the equilibrium effects. In deep scarcity, more data raises the marginal product of data engineers faster than it benefits routine workers, so the data wage premium rises and workers flow into data *from both other occupations*. Past the threshold $\bar{\Omega}$, the direct benefit to data engineers is

dominated by the shared efficiency gains, so the premium falls and the flows reverse: the data share declines and scientific employment partially recovers. The reallocation path is globally single-peaked with a unique, endogenous threshold[9]. Further, equilibrium employment peaks at $X^* = \sigma_\epsilon^2\sqrt{1-\tau\sigma_0^2}$, strictly *before* the marginal product of data engineers peaks at the scarcity boundary $X = \sigma_\epsilon^2$, by Proposition 3. Thus, between the two boundaries, additional data still raises the value of prediction, but the shared productivity gain to routine work already dominates, and the data share falls. Figure 3 collects both propositions. Panel A partitions the state space $(\Omega, D)$ and traces the equilibrium path through it: the path crosses the inner boundary exactly once at its own maximum, as established by Proposition 4. Panel B plots the same equilibrium against the data stock and separates the two turning points, so that the band between them is the range over which additional data still raises the marginal product of data engineers while their equilibrium employment is already contracting.

*Where the Inflow Comes From.* A rise in the data premium draws workers from both adjacent margins. By Lemma A2 in the Appendix, it pulls $m_{SD}\, d\tilde{w}^D$ workers out of science and $m_{D0}\, d\tilde{w}^D$ out of routine work, where $m_{SD}$ and $m_{D0}$ are the densities of agents at the science-data and routine-data margins. Both are strictly positive for every $\nu \in (-1,1)$, which is why the *sign* of the science decline in Proposition 4 needs no restriction on the taste correlation. The *composition* of the inflow is a different matter. Science need not supply the majority of new data engineers, and plausibly does not: routine work is the far larger occupation, so it can cede the same headcount at much less internal strain. The comparison that survives this size effect is the proportional one: science contracts faster than routine work in percentage terms exactly when

$$\xi^R m_{SD} > \xi^S m_{D0}, \tag{6}$$

a condition made *easier* by science being the smaller occupation. Appendix Section C.3 derives both sourcing shares, gives the exact majority condition, and shows that $\nu$ signs neither on its own. Both are equilibrium objects that could instead be disciplined by the observed asymmetry of science-to-data versus routine-to-data transitions in Section 6.

The next corollary formalizes the idea that while information remains deeply scarce, innovation reduces its future talent base through the data it creates. This is the sense in which innovation cannibalizes itself.

COROLLARY 1 (Innovation Cannibalizes Its Own Talent Base). *Consider generation 1, and recall that* $\Omega_1 = \Phi\gamma S_0^\beta I_0^\zeta$ *with* $\Phi = 1+\phi(N-1)$. *Then*

$$\frac{\partial \xi_1^S}{\partial S_0} = \frac{d\xi^{S*}}{d\Omega}\big(\Omega_1\big)\cdot \Phi\gamma\beta\, S_0^{\beta-1} I_0^\zeta,$$

[9] No auxiliary monotonicity hypotheses along the path are needed

*which is strictly negative whenever $\Omega_1 < \bar{\Omega}$, that is, whenever the period-1 equilibrium lies in the deep-scarcity region; symmetrically, $\partial \xi_1^D / \partial S_0 > 0$ and $\partial \xi_1^R / \partial S_0 < 0$ there. For $\Omega_1 > \bar{\Omega}$ all signs reverse. Moreover, because the equilibrium shares are independent of the ideas stock, the reallocation operates entirely through the data byproduct of innovation: the ideas channel $\chi S_0 \mapsto I_1$ leaves period-1 occupation shares unchanged.*

*Structural Change.* With Proposition 4 and Corollary 1 we come full circle. The economy transitions from a pure *science* phase in period 0 ($D_0 = 0$) to a *mixed* economy in period 1, and potentially to a *data-intensive* economy in period 2. Whether that last transition occurs, and whether it persists, depends on the threshold $\bar{\Omega}$ of Proposition 4. The non-monotonicity also warns against a hasty conclusion suggesting the unbounded growth of data engineers. In turn, Corollary 1 clarifies how, by generating data, innovation erodes the very talent base it depends on. These dynamics offer a parsimonious theory of the rise of the employment share of data occupations and its competition for talent with scientific activities documented in Figure 1.

## 4. Constrained Planner's Allocation

A question that arises is whether the increasing attractiveness of the data sector can be harmful to the economy. This section studies what a planner would do differently with respect to the competitive equilibrium to answer this question. I study a planner that is constrained to leave every worker at least as well off as in the market: she cannot pay below the relevant market wage and must compensate workers for assignments they value less because of their occupational preferences. I solve this problem locally, at the market allocation, with no global ranking of allocation paths. The main lesson is that the market equilibrium systematically *underprices* science. Scientists generate positive externalities that propagate across time through ideas and data accumulation. However, the market only compensates them for their current contribution to output. This dynamic contribution is left unpriced in every non-terminal period. Thus, a reallocation towards science raises welfare.

*Market utilities and compensation.* Objects carrying a superscript "$m$" are evaluated along the market path of Propositions 1-A1, initialized at $I_0 > 0$ and $\omega_0 = 0$. At date $t$, a worker of type $\eta$ obtains market utility

$$\bar{u}_t(\eta) \equiv \max_{o \in \{S,D,R\}} \left\{ \ln w_t^{o,m} + \eta^o \right\} \tag{7}$$

and the market assignment satisfies $a_t^m(\eta) \in \arg\max_o \{\ln w_t^{o,m} + \eta^o\}$. At date 0, I use the extended-real convention $\ln w_0^{D,m} = -\infty$. Thus, data engineering drops out of the market choice while the compensation requirement below remains finite.

The least consumption that leaves type $\eta$ at her market utility when assigned to occupation $o$ is

$$C_t^o(\eta) \equiv \exp\{\bar{u}_t(\eta) - \eta^o\}, \tag{8}$$

Since $\bar{u}_t(\eta) \geq \ln w_t^{o,m} + \eta^o$,

$$C_t^o(\eta) \geq w_t^{o,m} \qquad \text{for every } \eta \text{ and every } o, \tag{9}$$

with equality exactly when $o$ is market optimal for type $\eta$. Hence, the market assignment costs exactly the market wage bill. Lemma A3 records both facts.

*Compensated perturbations.* Fix a date $t$. A *market-boundary perturbation* is a differentiable family of assignments $a_t(\varepsilon)$ obtained by smoothly displacing the indifference boundaries of $a_t^m$, with $a_t(0) = a_t^m$. All assignments at dates $s \neq t$ remain at their market values. The stocks evolve from the perturbed date-$t$ allocation. This is a *fixed-shares continuation*.

Every worker is held at her market utility: a date-$t$ worker assigned by $a_t(\varepsilon)$ consumes $C_t^{a_t(\varepsilon)(\eta)}(\eta)$, while workers at other dates keep their market assignments and wages. Dropping the constant endowment, define the resulting discounted residual from date $t$ onward as

$$\mathcal{W}_t(\varepsilon) \equiv \sum_{s=t}^{2} \rho^s \Big[\pi_s^E\big(z_s(\varepsilon); I_s(\varepsilon), \Omega_s(\varepsilon)\big) - L\mathbb{E}\big[C_s^{a_s(\varepsilon)(\eta)}(\eta)\big]\Big]. \tag{10}$$

Here, $a_s(\varepsilon) = a_s^m$ for $s \neq t$. Earlier dates are unaffected and omitted. At $\varepsilon = 0$, equation (10) equals the market residual from date $t$ onward. A positive change is a Pareto improvement: workers retain their market utilities and the long-lived owner is strictly better off. Appendix Section C.4 presents the full statement of the constrained planner problem.

*The dynamic wedge.* Within a period, the market is efficient under this compensation criterion. The market assignment uniquely maximizes current output net of the consumption required to hold workers at their market utilities, by Lemma A5. The omitted contribution is intertemporal. Only scientists move the two stocks:

$$I_{t+1} = (1-\delta_I) I_t + \chi S_t, \qquad \omega_{t+1} = (1-\delta_\omega)\omega_t + \gamma S_t^\beta I_t^\zeta.$$

Define the dynamic wedge on a date-$t$ scientist as

$$\kappa_t \equiv \chi \Lambda_t^I + \gamma\beta\, S_t^{\beta-1} I_t^\zeta \Lambda_t^\omega, \tag{11}$$

where $\Lambda_t^I$ and $\Lambda_t^\omega$ are the fixed-shares marginal values of next period's stocks. Lemma A6 gives their recursions. The two terms in (11) are the future value of ideas and data created by one

additional scientist.

PROPOSITION 5 (Local Underprovision of Science). *Evaluate all objects along the market path. Then:*

*(i) At every date, the market assignment uniquely maximizes current output net of the consumption required to hold every worker at her market utility.*

*(ii) Every occupation is paid its current marginal product. A scientist additionally raises discounted future output by the dynamic wedge, $\kappa_t$, where*

$$\kappa_0 > 0, \qquad \kappa_1 > 0, \qquad \kappa_2 = 0.$$

*The dynamic wedges on data engineers and routine workers are zero.*

*(iii) For every market-boundary perturbation,*

$$\frac{d\mathcal{W}_t(\varepsilon)}{d\varepsilon}\Big|_{\varepsilon=0} = \rho^t \kappa_t L \frac{d\xi_t^S(\varepsilon)}{d\varepsilon}\Big|_{\varepsilon=0}.$$

*(iv) For $t \in \{0,1\}$, there exists a science-increasing market-boundary perturbation and $\bar{\varepsilon} > 0$ such that $\mathcal{W}_t(\varepsilon) > \mathcal{W}_t(0)$ for every $\varepsilon \in (0, \bar{\varepsilon})$. Every worker remains at her market utility, and the owner is strictly better off. Thus, the market science level and share are locally inefficiently low. At $t = 2$, every market-boundary perturbation has zero first-order welfare effect.*

Proposition 5 compares private and social returns at the market allocation. Without solving for a global efficient path, it provides a robust local lesson: in every period with a future, a compensated marginal increase in science raises welfare. The economics behind Proposition 5 is a fundamental timing problem in the data economy. Scientists generate benefits that accrue in future periods through ideas and data accumulation. The market compensates them only for their current contribution to output, while the planner recognizes their full intertemporal value. Data engineers and routine workers, by contrast, affect only current-period efficiency through variance reduction and production: they have no dynamic spillovers. The two components of the wedge $\kappa_t$ formalize this intuition by separating a familiar idea's externality from the novel paper's data channel. The term $\chi \Lambda_t^I$ is present without data accumulation. The term $\gamma\beta S_t^{\beta-1} I_t^{\zeta} \Lambda_t^{\omega}$ is specific to data: scientists create a stock that raises the future productivity of data engineers. The market prices neither contribution.

## 5. Revisiting Open Issues

Sections 3 and 4 developed the model's core insights: (i) data accumulation drives endogenous structural change from science to data engineering; (ii) the market leaves the dynamic contribution of scientific labor unpriced, and employs too few scientists relative to a constrained planner. This section applies these insights to two policy-relevant debates through the lens of the model: data sharing and automation.

### 5.1. The Role of Data Sharing

Should data be shared more widely? The non-rival nature of data suggests that sharing always improves output by allowing multiple firms to use the same information simultaneously. But once we account for occupational choice, data sharing has an ambiguous effect: it improves *static* efficiency by reducing operational uncertainty while potentially worsening *dynamic* efficiency by making data careers relatively more attractive. This section characterizes this trade-off and shows that a regulator limited only to a data-sharing instrument may optimally restrict sharing to preserve scientific talent.

To compare different policies, I need to define a common welfare anchor. Fix the reservation utilities generated by full sharing, $\bar{\phi} = 1$, such that

$$\bar{u}_t^1(\eta) \equiv \max_{o \in \{S,D,R\}} \left\{ \ln w_t^{o,m}(1) + \eta^o \right\}, \qquad C_t^{o,1}(\eta) \equiv \exp\left\{ \bar{u}_t^1(\eta) - \eta^o \right\}. \tag{12}$$

Holding these requirements fixed compares policies in common consumption units. For any assignment path $a$, let

$$\mathcal{W}^A(a,\phi) \equiv \sum_{t=0}^{2} \rho^t \left[ \pi_t^E\big(z(a_t); I_t(a), \Phi(\phi)\omega_t(a)\big) - L\mathbb{E}\big[ C_t^{a_t(\eta),1}(\eta) \big] \right]. \tag{13}$$

The stocks generated by a fixed path $a$ do not depend on $\phi$. The constrained planner can choose $a$ and $\phi$. A second-best regulator chooses only $\phi$ and leaves assignments to the market. The following proposition explains how data sharing is chosen under these two mechanisms.

PROPOSITION 6 (Data Sharing and Talent Reallocation). *Assume there is more than one firm in the economy, $N > 1$, and evaluate welfare at the full-sharing anchor in equation (12). Then:*

*(i) Suppose the date-$t$ market allocation is in the deep scarcity region, $\mathcal{S}^{deep}$. Holding $\omega_t$ fixed, the market data share is strictly increasing in $\phi$. Under a permanent policy, there is no inherited-stock feedback at $t = 1$, because $\omega_1$ is predetermined. At date 2, the inherited firm-level stock responds to the policy, but the aggregate data stock $\Omega_2$ is strictly increasing in $\phi$. Consequently, the date-2 market data share strictly increases with $\phi$ if the date-2 equilibrium lies in $\mathcal{S}^{deep}$, strictly decreases outside its closure, and has zero derivative on its boundary.*

*(ii) For every fixed assignment path, $\mathcal{W}^A(a,\phi)$ is nondecreasing in $\phi$. Hence full sharing is optimal under the constrained planner allocation.*

*(iii) Let occupations be determined by the market. A maximizer $\phi^{SB} \in [0,1]$ of anchored compensated surplus exists. If the date-1 full-sharing allocation is in the deep scarcity region, $\mathcal{S}^{deep}$, then $dS_1^m/d\phi < 0$ and*

$$\left.\frac{d\mathcal{W}^{SB}}{d\phi}\right|_{\phi=1^-} = \sum_{t=0}^{2} \rho^t \underbrace{\frac{\partial \pi_t^E}{\partial \phi}}_{\text{Direct Effect}} + \rho \underbrace{\kappa_1 \frac{dS_1^m}{d\phi}}_{\text{Dynamic Terms}}. \tag{14}$$

*Consequently, every maximizer satisfies* $\phi^{SB} < 1$ *if*

$$\rho\kappa_1 \left| \frac{dS_1^m}{d\phi} \right| > \rho \frac{\partial \pi_1^E}{\partial \phi} + \rho^2 \frac{\partial \pi_2^E}{\partial \phi}, \tag{15}$$

*where each partial derivative holds occupations and the inherited firm-level data stock fixed. Thus,* $\phi^{SB} < 1$ *when returning workers to science outweighs the direct informational gains of data sharing*

In deep scarcity, Proposition 6 indicates that in the market equilibrium, greater data sharing shifts talent toward data engineering at any given inherited data stock, and unambiguously so at date 1. Thus, while data sharing raises current-period output through better variance reduction, it also draws talent away from science, which depresses welfare through slower ideas and data accumulation. The net effect on welfare is not signed in general, since more effective data also reduces future uncertainty.

When the constrained planner chooses assignments directly, data sharing is unambiguously beneficial. Fix any assignment path. The stocks it generates are independent of $\phi$, because data accrues as a byproduct of quality, which contains no aggregate data stock $\Omega$. Raising $\phi$ raises $\Omega_t = \Phi\omega_t$ and lowers posterior variance at every date, while the anchored requirements $C_t^{o,1}$ do not move at all. Every assignment path is therefore worth weakly more at higher $\phi$, and taking the maximum preserves the ranking. Her optimal assignment does respond to $\phi$ - a larger data stock raises the marginal product of data engineers, and she reallocates accordingly - but by revealed preference that response can only add to the gain.

This benchmark indicates that data sharing *per se* is not problematic. The issue is the *interaction* between data sharing and market-determined occupational choice. When workers respond to wages, higher $\phi$ raises data engineer wages and pulls talent from science, which harms welfare.

Since the inefficiency is a positive dynamic externality from scientific labor, the natural decentralized response is a direct subsidy to science that pays scientists their full marginal contribution, leaving data sharing at $\phi = 1$. The required marginal correction is $\kappa_t$ per Proposition 5. The analysis of $\phi$ below is therefore relevant for environments in which such targeted subsidies are unavailable. For instance, because science subsidies are hard to target when scientists and data engineers are drawn from the same talent pool and perform overlapping tasks, or because $\phi$, a data-governance rule, is the only lever within the regulator's mandate. Under this constraint, $\phi$ becomes an interesting second-best instrument.

Consider now that scenario: a second-best regulator can mandate data sharing by choosing $\phi$ but cannot directly control occupation choices, which are determined by the market. Condition (15) formalizes the resulting tradeoff. Its left-hand side is the dynamic loss, the science employment that sharing draws away valued at the unpriced wedge $\kappa_1$; its right-hand side is the discounted static gain from lower operational uncertainty. When the former dominates, the

regulator optimally *restricts* data sharing below the full-sharing level $\phi = 1$.[10]

Recent empirical evidence from Demirer et al. (2026) is informative about the static side of this tradeoff. Studying the European Union's General Data Protection Regulation (GDPR) - a policy that effectively restricts data collection and sharing - they estimate that the regulation raised the variable cost of data for firms by 22.4% on average, causing EU firms to reduce data storage by 26% and compute by 15% relative to comparable US firms. Despite these sizable input responses, the implied increase in total variable production costs is modest: on the order of 0.1–0.6% depending on the industry. Through the lens of Proposition 6, this suggests that the static cost of restricting $\phi$ may be modest.

### 5.2. Data Accumulation and Routine Work Automation

In many industries, the automation of routine work increasingly relies on a firm's accumulated data stock (Brynjolfsson, Rock, and Syverson 2017; Zolas et al. 2021; Acemoglu et al. 2022a; Farboodi, Koh, and Xia 2026). To the extent that data-driven automation substitutes for routine workers while simultaneously raising the value of data, it generates an additional force through which talent is drawn away from science toward data processing.

As a first pass to study this channel, I present a simple extension of the baseline model allowing routine workers to use machines, $M$, to produce. The number of machines available to all firms is an increasing function of the aggregate data stock: $M_t = M(\Omega_t)$, with $M'(\Omega) > 0$ and $M(0) = 0$. This captures the idea that accumulated data enables automation technologies, constituting perfect substitutes for routine labor. The firm's value is now

$$v_{j,t} = S_{j,t}^{\beta} I_t^{\zeta} \left[ 1 - \tau \frac{\sigma_0^2 \sigma_\epsilon^2}{\sigma_\epsilon^2 + D_{j,t}^{\alpha} \Omega_t^{1-\alpha} \sigma_0^2} \right] \left[ R_{j,t} + M(\Omega_t) \right]^{\psi} - w_{j,t}^{R} R_{j,t} - w_{j,t}^{S} S_{j,t} - w_{j,t}^{D} D_{j,t}.$$

The following proposition compares the local wedge under the two technologies.

PROPOSITION 7 (Data-Driven Automation Widens the Unpriced Contribution of Science). *Fix $t \in \{0,1\}$ and an interior allocation path. Compare the baseline economy, $M \equiv 0$, with an automated economy, $M(\Omega) > 0$ for $\Omega > 0$, at the same states and occupation shares. Then*

$$\kappa_t^{auto} > \kappa_t^{base} > 0.$$

Scientists now create data that both improves prediction and enables automation. The extra channel raises future output and therefore widens $\kappa_t$. The market still pays scientists only for current output. Nonetheless, because automation also changes current wage premia, displaced routine workers flow into both science and data. The market science share therefore need not fall.

[10] The regulator does not implement the compensating payments. The criterion asks whether the alternative regime generates enough resources to restore full-sharing utilities. It is a Hicksian, or potential-Pareto, comparison.

The next corollary examines how automation and data accumulation interact in the deep-scarcity region by reallocating labor from routine work to data engineering.

COROLLARY 2 (Automation in Deep Scarcity Reinforces the Flow to Data). *Consider the automated economy at date $t$, and suppose the equilibrium is interior with $(L\xi_t^D, \Omega_t) \in \mathcal{S}^{deep}$. A marginal increase in $\Omega_t$:*

*(i) raises both the science and the data premium relative to routine work through the automation channel, and additionally raises the data premium through the information channel;*

*(ii) in equilibrium, strictly raises the data share and strictly lowers the routine share; the science share nets an automation-driven inflow against an information-driven outflow.*

Corollary 2 captures the combined effect of two forces, both of which are statements about *relative* wages. First, automation makes routine work relatively less attractive: effective routine capacity $R_t+M(\Omega_t)$ expands with data, which raises the premia of both non-routine occupations at fixed shares. Second, in deep scarcity, the information channel raises the data premium on top of this. In equilibrium, routine employment shrinks while data employment expands, amplifying the reallocation toward data processing. By Proposition 7, the same data accumulation widens the unpriced value of science, at a common state and allocation path. Hence, data-driven automation in deep scarcity widens the wedge on science.

## 6. Empirical evidence

The model predicts that, while data remains deeply scarce, its accumulation raises the relative attractiveness of data occupations and reallocates talent toward them. This section examines whether the empirical patterns implied by this mechanism are visible in the US economy between 2000 and 2022. I first consider evidence on whether the economy operates in the deep-scarcity region described by the model. I then document three facts. First, past scientific activity predicts subsequent data activity and the development of data-processing infrastructure. Second, the expansion of data work occurs disproportionately within historically innovative labor markets and is accompanied by rising returns to data occupations. Third, data has become increasingly connected to the scientific talent pool through occupational transitions. Taken together, these three facts provide a broad empirical characterization of the growth of data-intensive work and its relationship with scientific activity and scientific talent across firms, local labor markets, cohorts, and individual career histories.

### 6.1. Data

My main data source is Revelio Labs' individual-level data on professional profiles. These files contain user-level information from LinkedIn resumes on current and historical positions, educational history, name, and demographics. The dataset can trace back an individual's work history since 1950 up to 2025 all over the world. I use two files. The file "positions", which

allows me to observe information on an individual's employment positions, including predicted salary[11], occupation, start and end dates, the employing firm, the country, the Metropolitan Statistical Area (MSA), and the city. The file "education", which gives information on the field of study, university, degree level, and the start and end dates.

I focus on individuals completing their first degree in 2000 or later and holding US positions. This period coincides with reliable Revelio coverage and the emergence of modern data infrastructure. The final sample consists of 35.1 million unique individuals across 5.3 million firms, 309 industries, and 414 MSAs, followed over the period 2000-2022. Appendix Section B.2 provides additional details on sample construction.

Building on the model, I classify workers into three occupations using Revelio Labs' standardized occupation codes: data engineers, scientists, and routine workers (all other roles). Appendix Section B.2 provides detailed definitions. I then compute the share of each occupation and its wage premium relative to routine workers at different levels of aggregation: the firm level; the public firm level, where I merge Revelio with Compustat; the local labor market, defined as the NAICS 2-MSA pair; and the cohort level, defined as the year of graduation from the *first* degree obtained by an individual in any field.

To validate the local labor market results obtained with Revelio, I use the American Community Survey (ACS). In the ACS analysis, I define local labor market as NAICS 2-commuting zones pairs. Commuting zones are mapped from PUMAs using the crosswalks developed by Autor and Dorn (2013). Appendix Section B.3 details the sample construction for the ACS.

To measure firms' capacity to use data more directly, I complement Revelio with data from Enlyft, a commercial business-to-business account-intelligence firm. Enlyft provides technology vendors and their sales, marketing, and partner organizations with data-enrichment, account-scoring, and prospecting tools. These products draw on firmographics, buying-intent signals, and, most importantly for this paper, *technographics*: information on the technologies that Enlyft detects in firms' installed technology stacks. Whereas conventional firm surveys typically record the use of broad classes of technology, Enlyft maps individual firms to specific software products. This offers an unusually granular view of firms' data-processing infrastructure. Enlyft belongs to a broader class of commercial technographic datasets that record firms' use of particular information technologies. Earlier work has mostly used the Computer Intelligence Database (CiTDB), originally produced by Harte Hanks, subsequently maintained by Aberdeen, and now part of SWZD, to measure establishment-level IT investment and the adoption of particular software and hardware products (Forman 2005; Bloom et al. 2014; Charoenwong et al. 2024; Acharya, Shao, and Xu 2026). The novelty in using Enlyft is to exploit the composition of firms' software portfolios across a broad set of functional categories to construct a measure of their capacity to process and exploit data, and link this measure to both Compustat and workers'

[11] Predicted salary in Revelio is an imputed measure that combines base salary and variable pay such as bonuses but excludes stock options. It is not observed directly. As the name indicates, the salary is predicted by Revelio using a model that takes as inputs features of the job such as role, seniority, and firm.

historical employment records in Revelio.

The extract used in this paper is a custom delivery produced for this research project. I use the February 2023 delivery, which contains 1,644,305 firm-software-product records covering 11,622 distinct products and 216 product categories for 9,802 public and private firms. The analysis below restricts the sample to the 4,295 public firms matched to Compustat. Each record indicates that Enlyft detected a firm using a particular software product and includes a proprietary confidence category for that detection. The delivery also classifies products into seven project-specific, potentially overlapping functional categories - infrastructure, customer, coordination, production, security, employee, and quality - and identifies products related to big data. I aggregate these product-level records to construct firm-level measures of software adoption and data-processing capacity. Appendix Section B.4 provides a detailed description of the data used in the paper.

### 6.2. Results

The model describes a sequence in which scientific activity generates data, the accumulated data raises the return to data-processing capacity, and talent subsequently reallocates toward data work. I organize the evidence around three empirical facts corresponding to this sequence. First, firms and local labor markets with greater past science intensity subsequently develop both more data employment and, among public firms, greater capacity to process and use data. Second, the expansion of data work occurs disproportionately within historically innovative local labor markets and is accompanied by rising private returns to data occupations. Third, data work has become increasingly connected to the scientific talent pool through occupational transitions.

*Empirical relevance of the deep-scarcity region.* Before turning to the three main empirical facts, I assess whether the deep scarcity region is empirically plausible in US and international data. Recall that the direction of the reallocation from science to data occupations in Propositions 3 and 4 depends on whether the economy lies in the deep-scarcity region $\mathcal{S}^{deep}$. This region is defined in terms of the prior and signal variances, $\sigma_0^2$ and $\sigma_\epsilon^2$, which have no direct empirical counterparts. I therefore do not attempt to identify the location of the aggregate economy in the model's state space. Instead, Appendix Section E motivates the plausibility of deep scarcity using frontier large-language-model (LLM) training data. This is a revealed-preference argument: as recently as 2022, frontier models were trained far below the compute-optimal data frontier.

Demirer et al. (2026) provide complementary evidence for the LLM argument. They estimate a production function with data and compute using high-frequency cloud-computing records for roughly 37,000 firms in the United States and the European Union. Their estimated elasticities of substitution between data and compute range from 0.32 to 0.44 across software, services, and manufacturing, indicating strong complementarity between the two inputs. This

finding is consistent with data remaining a valuable input at the margin rather than with firms having reached information saturation. It does not, however, identify the model's scarcity ratio or distinguish the scarcity region from the deep-scarcity region. I therefore treat these two arguments only as evidence that the maintained region condition is plausible and turn next to implications that can be assessed more directly in the data.

*Fact 1: Past science predicts future data activity.* The model's core mechanism - that successful science creates data which then attracts talent away from science - generates a testable implication: past science intensity should predict future data intensity within economic units. Table 1 evaluates this prediction using within-unit variation at three levels of aggregation: firms, public firms, and local labor markets. For each unit-year pair, I compute the log employment of data, science, and routine worker occupations. I then regress the contemporaneous log data employment on log science employment lagged one to five years. The model predicts that higher past science employment should predict higher current data employment. I control for the log total employment of the unit to hold its size constant. I include unit and year fixed effects to ensure identification from within-unit changes over time, ruling out confounds from stable differences across units or common time trends.

Table 1 lends broad support to the theory. To interpret magnitudes, consider column (5), which analyzes local labor markets. The cumulative effect of 0.177 implies that an increase of 10 percent in the science employment (sustained over five years) predicts a 1.77 percent increase in data employment. The joint test strongly rejects the null that all five lags are jointly zero. Columns (6), (7), and (8) subsequently saturate the specification with industry-by-year fixed effects and MSA-by-year fixed effects, controlling for national changes in industry performance and MSA cross-industry yearly shocks. While the estimated cumulative effect shrinks slightly in magnitude, it remains statistically significant, and the joint test strongly rejects the null that all five lags are jointly zero. Column (3), focusing on public firms, shows an even larger cumulative effect of 2 percent. Column (1), which includes all firms, paints a very similar picture: a cumulative effect of 1.8 percent. Both the results in columns (1) and (2) are robust to the inclusion of industry-by-year fixed effects. Across all eight specifications, the distributed lag structure indicates that the effect builds gradually. Table A1 shows that the results are robust to not weighting the regressions. Table A2 shows that the results are robust to using NAICS 4 instead of NAICS 2 industries to define local labor markets.

While Table 1 establishes a robust temporal relationship between scientific activity and subsequent data employment, it does not observe the intermediate object emphasized by the model: the data stock. To get closer to this mechanism, I now exploit Enlyft data. I measure a firm's data infrastructure using the number of distinct software products detected by Enlyft in its technology stack.[12] This measure should not be interpreted as the firm's stock of data itself.

[12]The results are robust to using a variety of other proxies such as IT spending, products per employee, and specific Enlyft scores on IT footprint, Software-as-a-Service adoption, and cloud usage.

TABLE 1. Regressions across firms, public firms, and local labor markets of contemporaneous log data employment on past log science employment.

| Level of Analysis | All firms | | Public firms | | Local labor markets | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Log (Science employment) t-1 | 0.068*** | 0.069*** | 0.062*** | 0.072*** | 0.079*** | 0.044*** | 0.062*** | 0.024*** |
| | (0.015) | (0.015) | (0.024) | (0.023) | (0.016) | (0.0096) | (0.014) | (0.0088) |
| Log (Science employment) t-2 | 0.026*** | 0.024*** | 0.014 | 0.024* | 0.022*** | 0.010* | 0.020*** | 0.0070 |
| | (0.0089) | (0.0082) | (0.015) | (0.014) | (0.0075) | (0.0060) | (0.0073) | (0.0058) |
| Log (Science employment) t-3 | 0.023*** | 0.020*** | 0.020 | 0.015 | 0.022*** | 0.0071 | 0.023*** | 0.0079 |
| | (0.0083) | (0.0069) | (0.015) | (0.015) | (0.0066) | (0.0052) | (0.0066) | (0.0053) |
| Log (Science employment) t-4 | 0.020*** | 0.013** | 0.029 | 0.022 | 0.022*** | 0.012** | 0.020*** | 0.0089 |
| | (0.0074) | (0.0064) | (0.019) | (0.015) | (0.0061) | (0.0056) | (0.0064) | (0.0057) |
| Log (Science employment) t-5 | 0.044*** | 0.034*** | 0.071*** | 0.055*** | 0.032*** | 0.024*** | 0.025*** | 0.012* |
| | (0.014) | (0.010) | (0.021) | (0.019) | (0.0085) | (0.0070) | (0.0087) | (0.0067) |
| Total employment controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Firm FE | ✓ | ✓ | — | — | — | — | — | — |
| GVKEY FE | — | — | ✓ | ✓ | — | — | — | — |
| NAICS 2-MSA FE | — | — | — | — | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | — | ✓ | — | ✓ | — | — | — |
| NAICS 2-Year FE | — | ✓ | — | ✓ | — | ✓ | — | ✓ |
| MSA-year FE | — | — | — | — | — | — | ✓ | ✓ |
| Cumulative effect (5 lags) | 0.180*** | 0.159*** | 0.197*** | 0.188*** | 0.177*** | 0.097*** | 0.150*** | 0.060*** |
| | (0.037) | (0.031) | (0.056) | (0.056) | (0.030) | (0.019) | (0.028) | (0.017) |
| Standardized coefficient | 0.377 | 0.332 | 0.380 | 0.363 | 0.391 | 0.214 | 0.333 | 0.132 |
| Mean of dependent variable | 3.850 | 3.851 | 4.712 | 4.715 | 4.824 | 4.824 | 4.829 | 4.829 |
| Joint test p-value | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.000 | 0.002 |
| Adjusted R-squared | 0.97 | 0.98 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 |
| Observations | 123,532 | 123,381 | 11,822 | 11,822 | 61,671 | 61,671 | 61,671 | 61,671 |

*Notes.* Standard errors, in parentheses, are clustered at the firm level in columns (1) and (2), the GVKEY level in columns (3) and (4), and the NAICS 2-MSA level in columns (5), (6), (7), and (8). $^{***}$, $^{**}$, and $^{*}$ correspond to rejection of the null that the coefficient is zero, at the 1%, 5%, and 10% levels, respectively. The dependent variable is a unit (firm, public firm, or local labor market) log data employment in year $t$. The independent variables are lags at $t-1$, $t-2$, $t-3$, $t-4$, and $t-5$ of the log science employment. In all regressions, I control for the log of the total employment of the unit at $t$, $t-1$, $t-2$, $t-3$, $t-4$, and $t-5$. The benchmark specification, in columns (1), (3), and (5) is $\log(\text{data employment})_{i,t} = \sum_{s=1}^{5} b_s \log(\text{science employment})_{i,t-s} + \sum_{s=0}^{5} c_s \log(\text{total employment})_{i,t-s} + g_i + d_t + u_{i,t}$. $g_i$ are unit fixed effects, $d_t$ are year fixed effects. In columns (2), (4), and (6), I include industry-year fixed effects. In columns (7) and (8) I include MSA-year fixed effects, and both industry-year fixed effects and MSA-year fixed effects, respectively. Units must have at least 10 workers to be included in columns (1), (2), (5), (6), (7), and (8). Because the dependent variable and the science-employment lags are specified in logs, the estimation sample is further restricted to unit-years with strictly positive data employment in year $t$ and strictly positive science employment in each of the preceding five years $t-1$ through $t-5$. All regressions are weighted by their employment level in the first year they appear in the panel (entry weight). The sample period is 2000-2022. Go back to main text.

Ideally, we would have access to the measure of data effectively stored and managed by the firm. Nonetheless, the measure built with Enlyft captures the software infrastructure through which firms store, process, and use data. It therefore provides a good measure of a firm's capacity to use its data.

Figure 4 presents two complementary relationships. Panel A asks whether past scientific activity predicts subsequent data infrastructure. I regress the log number of software products observed in 2022 on average log science employment over 2005-2009 period, controlling for average log assets and the average data-employment share over the same pre-period, as well as two-digit industry fixed effects. The coefficient on past science employment is 0.125 (0.019). Thus, a 10 percent increase in science employment during 2005-2009 is associated with approximately 1.25 percent more software products in 2022.

Panel B asks whether this infrastructure is in turn associated with the employment of data

workers. I regress the firm's 2022 data-employment share on the log number of software products detected by Enlyft, controlling for contemporaneous log assets and two-digit industry fixed effects. The coefficient is 0.0052 (0.0010): doubling the size of the observed software stack is associated with an approximately 0.36 percentage point increase in the data-employment share. A fractional-response specification delivers a virtually identical marginal effect.

Taken together, Table 1 and Figure 4 provide a more concrete picture of the first step in the model's mechanism. Firms with greater past scientific activity subsequently employ more data workers, and the Enlyft evidence shows that they also operate larger software infrastructures for processing data. In the cross-section, firms with larger infrastructures in turn employ relatively more data workers. These relationships are, of course, descriptive, and Enlyft provides only a 2022–2023 snapshot of software use. Nevertheless, the results reveal an observable *measurable* data object linking past science to the later expansion of data-intensive work.

*Fact 2: Data expansion is concentrated in innovative markets and accompanied by rising returns.* The previous result establishes that past science predicts future data intensity, but it does not indicate *where* this reallocation occurs. Is it driven by workers moving between firms and labor markets (between-unit reallocation), or by compositional shifts within existing units (within-unit reallocation)? The model predicts that high-innovation units, those with large past science shares, should exhibit stronger within-unit reallocation toward data roles, as accumulated data raises the marginal product of data engineers by more in those units.

To evaluate this prediction, I decompose the change in the national data share as follows. Let $\xi^D_{i,t}$ denote the data share in unit $i$ (say a local labor market, defined as a NAICS 2-digit industry-MSA pair) in year $t$, and let $p_{i,t} = e_{i,t}/\sum_j e_{j,t}$ denote unit $i$'s employment weight in the national economy. The national mean data share is $\bar{\xi}^D_t = \sum_i p_{i,t}\xi^D_{i,t}$. The change in the national mean from $t-1$ to $t$ can be decomposed as

(16)
$$\Delta\bar{\xi}^D_t = \underbrace{\sum_{i\in\mathcal{C}} p_{i,t-1}\Delta\xi^D_{i,t}}_{\text{Within}} + \underbrace{\sum_{i\in\mathcal{C}} (\xi^D_{i,t-1} - \bar{\xi}^D_{t-1})\Delta p_{i,t}}_{\text{Between}} + \underbrace{\sum_{i\in\mathcal{C}} \Delta\xi^D_{i,t}\Delta p_{i,t}}_{\text{Cross}} + \underbrace{\sum_{i\in\mathcal{E}} p_{i,t}(\xi^D_{i,t} - \bar{\xi}^D_{t-1})}_{\text{Entry}} + \underbrace{\sum_{i\in\mathcal{X}} p_{i,t-1}(\bar{\xi}^D_{t-1} - \xi^D_{i,t-1})}_{\text{Exit}},$$

where $\mathcal{C}$ denotes continuing units (active in both $t-1$ and $t$), $\mathcal{E}$ denotes entering units (active in $t$ but not $t-1$), and $\mathcal{X}$ denotes exiting units (active in $t-1$ but not $t$). The *within* component captures reallocation toward data roles within continuing units, holding their employment weights fixed. The *between* component captures shifts in employment toward units with initially high data shares. The *cross* component captures the covariance between changes in data shares and changes in employment weights. *Entry* and *exit* capture the contribution of units appearing or disappearing from the sample.

The model's main implication concerns the within component: it should be mostly driven by high-innovation units. To look at this, I classify each unit based on its science share in the

baseline period (the first year of available data), creating time-invariant innovation intensity groups. I then decompose the within component by summing separately over units in each tercile

$$\text{Within}_t = \sum_{g=1}^{3} \underbrace{\sum_{i \in \mathcal{C}_g} p_{i,t-1} \Delta \xi_{i,t}^{D}}_{\text{Within}_t^g}, \tag{17}$$

where $\mathcal{C}_g$ denotes continuing units in innovation group $g \in \{\text{Low}, \text{Medium}, \text{High}\}$. I use Revelio data for the period 2000-2022 with units defined as NAICS 2-digit industry-MSA pairs, and validate the results with the ACS for the period 2005-2022 with units defined as NAICS 2-digit industry-commuting zone pairs.

Figure 5 presents the results. Panel A shows the ACS results and Panel B shows the Revelio results. Despite differences in data sources (household surveys and online resumes), time periods (2005-2022 and 2000-2022), and geographic-industry aggregation (commuting zone by NAICS2 and MSA by NAICS 2), two patterns emerge. First, the within component (sum of all three innovation terciles) accounts for the majority of the increase in the national data share. In Revelio, within-unit reallocation represents essentially the entirety of the rise over 2000-2022 (101% of the total change), with between-unit and cross-term reallocation together contributing a net of roughly zero, as their individually modest effects (-5% and 3%, respectively) largely offset one another. In the ACS, within-unit reallocation represents 80% of the rise over 2005-2022, with between (13%) and cross (7%) playing a smaller but non-negligible secondary role.

Second, high-innovation units drive the within component. In Revelio, the contribution rises monotonically with initial science intensity: low-, medium-, and high-innovation markets account for 8%, 13%, and 80% of the total increase, respectively. The ACS exhibits a less monotonic pattern, but high-innovation markets again make the largest contribution, accounting for 48% of the total increase compared with 33% for low-innovation markets and essentially zero for the middle tercile. The gradient found in the Revelio data is what we should have expected to observe when taking the model to the letter. Together, these patterns support the model's core mechanism that reallocation within high-innovation markets is a central driver of the shift toward data occupations.

I now zoom in further to examine individual career paths. I focus on workers whose skills make them particularly likely to face the science-data margin: graduates in mathematics and statistics. These fields provide quantitative and computational training valued in both occupations, making their career choices especially informative about changes in the relative attractiveness of data work.

The model makes distinctive predictions about how these careers should evolve as data accumulates. Because the marginal product of data engineers rises with the data stock, more recent cohorts should enter data roles at higher rates directly out of school. Further, the data

wage premium at entry should rise across cohorts, reflecting the same force on the demand side and consistent with the role of the wage premium as a signal guiding workers' choices.

Figure 6 documents both patterns. Panel A of Figure 6 plots the share of workers in data roles in their first year after graduation by cohort. The data share more than doubles over time: the 2000 cohort starts at approximately 3.3% in data roles, while the 2022 cohort starts at 8.1%. Panel B of Figure 6 plots the corresponding first-year data wage premium. In line with the model's prediction, the premium rises from near zero for the 2000 cohort to approximately 49 log points for the 2022 cohort. The shift from science to data roles across cohorts at entry thus coincides with rising economic incentives.

The two series therefore move together: the period in which data occupations become increasingly attractive at career entry is also the period in which their relative compensation rises. This provides suggestive evidence that the expansion of data work operates through the occupational-choice margin emphasized by the model.

Appendix Figure A4 shows that these patterns extend beyond career entry: within mathematics and statistics cohorts, the share of workers in data occupations generally rises with years since graduation, and the data wage premium also tends to increase over the career.

*Fact 3: Data has become increasingly connected to the scientific talent pool.* The cohort evidence above shows that newer quantitative graduates increasingly choose data occupations at career entry. I now turn to the career-reallocation margin emphasized by the model: whether data has also become a more important destination for workers who are already employed in scientific and other high-skill occupations.

To isolate occupational reallocation from persistence, I focus on workers who change occupation between consecutive positions and exclude moves in which the origin and destination occupation are the same. I restrict attention to a set of high-skill occupation families - science, data, technological/software occupations, engineering, and finance - whose workers constitute plausible alternatives at this margin. Appendix Section B.2 provides the detailed occupation mappings.

A simple count of scientists entering data is difficult to interpret because science may itself become larger or more mobile over time. I therefore construct a representation ratio. For destination occupation $d$, define

$$R^{S}_{d,t} = \frac{\Pr(\text{origin} = S \,|\, \text{destination} = d,\ \text{switch})}{\Pr(\text{origin} = S \,|\, \text{origin} \neq d,\ \text{switch})}. \tag{18}$$

The numerator is the share of entrants into occupation $d$ coming from science. The denominator is the share of scientists among high-skill occupation switchers who could have entered $d$: workers already originating in $d$ are excluded because they cannot switch into their own occupation. A value of one therefore means that science supplies occupation $d$ exactly in proportion to its representation among workers eligible to enter that occupation. Values below one indicate

underrepresentation and values above one overrepresentation. The statistic therefore measures occupational affinity.

Panel A of Figure 7 plots this ratio for data occupations. The science-data representation ratio rises from approximately 0.45 in 2005 to 0.98 in 2022.[13] Thus, in the mid-2000s, science supplied workers to data at less than half the rate implied by its representation among eligible high-skill switchers. By the end of the sample, that gap had essentially disappeared. The dashed series constructs the benchmark within the MSA of the destination data job rather than nationally. The two series are nearly indistinguishable, indicating that the rising science-data connection is not explained by data jobs becoming concentrated in labor markets with unusually large scientific workforces.

Panel B asks whether this pattern is specific to data. To compare destinations using a common benchmark, I divide the share of scientists among entrants into each occupation by the share of scientists among all high-skill occupation switchers. Unlike Panel A, the denominator is held fixed across destinations. The contrast is sharp. Science's representation among workers entering data rises substantially throughout the sample, while its representation among entrants into the other high-skill destinations is flat or declining. Data therefore becomes increasingly distinctive within the set of occupations available to scientific workers, rather than simply benefiting from a general rise in scientists' occupational mobility.

The representation ratios establish that the occupational connection between science and data has strengthened substantially, but they do not indicate the magnitude of the underlying worker flows. Table 2 therefore reports science-to-data transitions over three non-overlapping six-year periods.

The reallocation grows markedly over time. Between 2005-2010, data accounts for 2.15% of observed occupation changes out of science. This share rises to 2.98% in 2011-2016 and to 5.12% in 2017-2022. Including scientists who leave the Revelio panel in the denominator produces a nearly identical pattern: the share of all science separations going to data rises from 1.99% to 2.67% and then to 4.54%.

Relative to the stock of scientific employment, the increase is similarly pronounced. Annual science-to-data transitions amount to 0.3% of science employment in 2005-2010, 0.57% in 2011-2016, and 0.96% in 2017-2022. Thus, the annual flow of scientific workers into data more than triples over the sample period.

These magnitudes reinforce the representation evidence in Figure 7. The growing science-data connection is not only a change in the composition of occupation switchers: the underlying flow of scientific talent into data occupations also becomes substantially larger over time.

The evidence so far is economy-wide and pools scientists employed across different types of

[13] The transition analysis covers only the 2005 to 2022 period. Constructing occupation changes requires observing both an origin and a destination position, and coverage of such transitions is substantially thinner in the earliest Revelio years. For example, in 2000, 92% of workers first observed entering data occupations have no prior position in the sample.

TABLE 2. Magnitude of science-to-data reallocation.

| | 2005–2010 | 2011–2016 | 2017–2022 | *2005–2022* |
|---|---|---|---|---|
| Share of science occupation changes going to data (%) | 2.15 | 2.98 | 5.12 | *3.91* |
| Share of all science separations going to data (%) | 1.99 | 2.67 | 4.54 | *3.50* |
| Annual flow as % of science employment | 0.30 | 0.57 | 0.96 | *0.71* |

*Notes.* This table reports science-to-data occupation transitions observed in Revelio. Columns (1) to (3) correspond to the non-overlapping periods 2005-2010, 2011-2016, and 2017-2022. Column (4) pools the full 2005-2022 sample. The first row reports the number of science-to-data transitions as a percentage of all observed occupation changes out of science. The second row adds exits from the Revelio panel to the denominator and reports science-to-data transitions as a percentage of all science separations. The third row reports the average annual number of science-to-data transitions divided by average science employment over the corresponding period. The sample period is 2005-2022. Go back to main text.

organizations. A natural question is whether the same pattern is visible among workers more closely associated with research and idea production. I therefore examine, in the Appendix, scientists leaving universities, defined as workers in science occupations employed in Educational Services (NAICS 61). This margin is closely related to the brain-drain patterns documented by Gofman and Jin (2024) and Akcigit et al. (2026) for computer science and AI.[14] This analysis extends that evidence to a broader set of scientific occupations and destination industries.

Appendix Figure A6 indicates that outflows from university science to Manufacturing (NAICS 31-33), Professional Services (NAICS 54), Information (NAICS 51), and Finance (NAICS 52) rise over the sample period.[15] Conditional on leaving university science for these industries, Appendix Figure A7 shows that the share entering data occupations rises substantially: between 2005 and 2022, the data-occupation index increases by a factor of two to five across the four destinations, while the corresponding indices for other major quantitative occupations are comparatively flat.[16] Therefore, the economy-wide pattern is also present among a narrower but key group closely tied to public knowledge and ideas production.

## 7. Conclusion

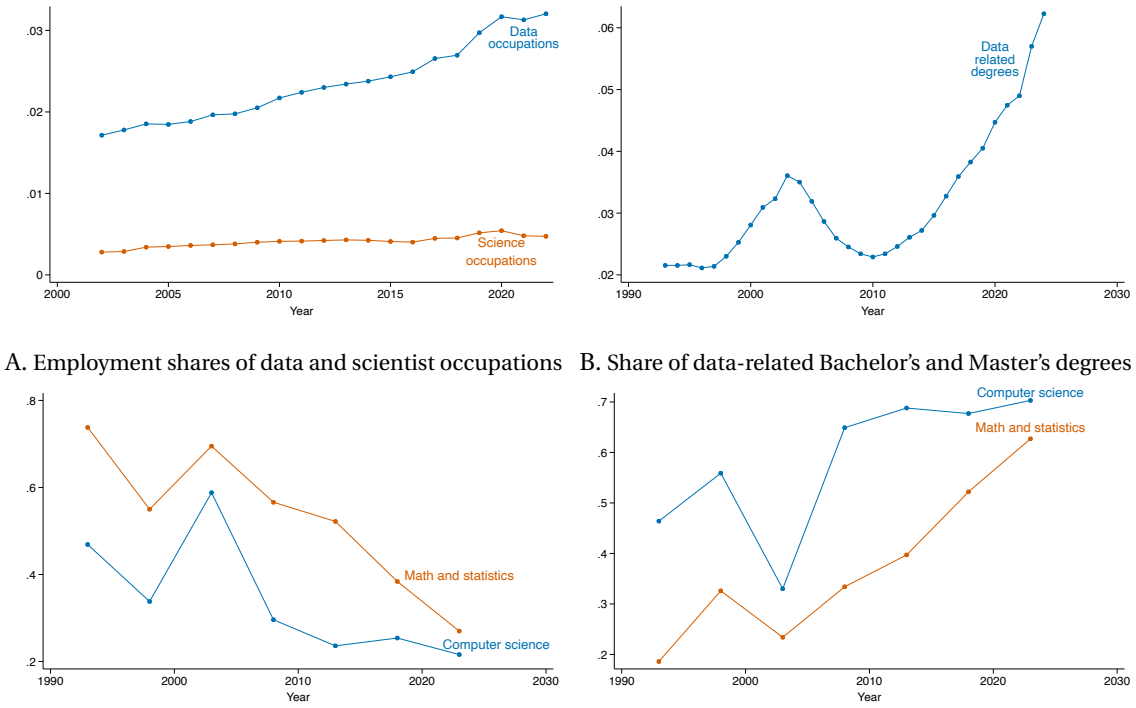
This paper studies how data accumulation shapes talent allocation. While economists have studied how data affects firm growth, competition, and economic output, the implications for occupational choice have remained unexplored. What drives the increasing importance of data jobs in the economy? And what does this mean for economic output? In a first step to answer these questions, I propose a model where occupational choice between science and data activities is driven by firms' demand to exploit an endogenously growing data stock.

[14] See also some anecdotal evidence in Ramsey and MacMillan (2015); Hernandez and King (2016); Bhaimiya (2025).

[15] Appendix Figure A5 plots the total outflows to any industry as a reference.

[16] Appendix Figure A8 reports levels as reference.

The model's main insight stems from a simple observation: unlike knowledge or physical capital, data accumulates as a *byproduct* of economic activity. Successful innovation can therefore create its own competition. Strong early innovation generates data that pulls talent toward data processing. While data remains deeply scarce, the data share rises at the expense of science. A comparison between the social planner (who may reassign workers but must leave each of them at her market utility) and the decentralized equilibrium shows that the market leaves the dynamic contribution of scientists unpriced in every period with a future. The intuition for this result is simple. Scientists generate benefits that accrue in future periods through ideas and data accumulation. The market compensates them only for their current contribution to output, while the planner recognizes their full intertemporal value. Data engineers, by contrast, affect only current-period efficiency through uncertainty reduction. In the decentralized equilibrium, talent is misallocated, and welfare is lower.




A. Employment shares of data and scientist occupations

B. Share of data-related Bachelor's and Master's degrees

C. Share of PhD graduates employed in academia

D. Share of PhD graduates employed in industry

FIGURE 1. Talent allocation and data jobs in the US between 1993 and 2024

Panel A plots the employment shares of data (blue line) and scientist (orange line) occupations in the US between 2002 and 2022. Data occupations correspond to fourteen occupations based on the National Career Clusters Framework. I specifically use two lists curated by O*NET: Data Science & Artificial Intelligence and Web & Cloud. Science occupations correspond to twenty-three occupations. The full lists are provided in Appendix Section B.1. Panel B plots the share of data-related Bachelor's and Master's degrees awarded in the US between 1993 and 2024. Data-related degrees are also identified using O*NET lists Data Science & Artificial Intelligence and Web & Cloud. Panel C plots the share of PhD graduates in computer and information sciences (blue line), and mathematics and statistics (orange line) employed in academia between 1993 and 2023. Panel D plots the share of PhD graduates in computer and information sciences (blue line), and mathematics and statistics (orange line) employed in business or industry between 1993 and 2023. Series in Panel A is created using the Occupational Employment and Wage Statistics (OEWS) program produced by the Bureau of Labor Statistics (BLS). Series in Panel B are created using the National Center for Education Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS) survey. Series in Panels C and D are created using the National Science Foundation (NSF) Survey of Earned Doctorates (SED). Appendix Section B.1 provides full details on data construction for all series. Go back to main text.

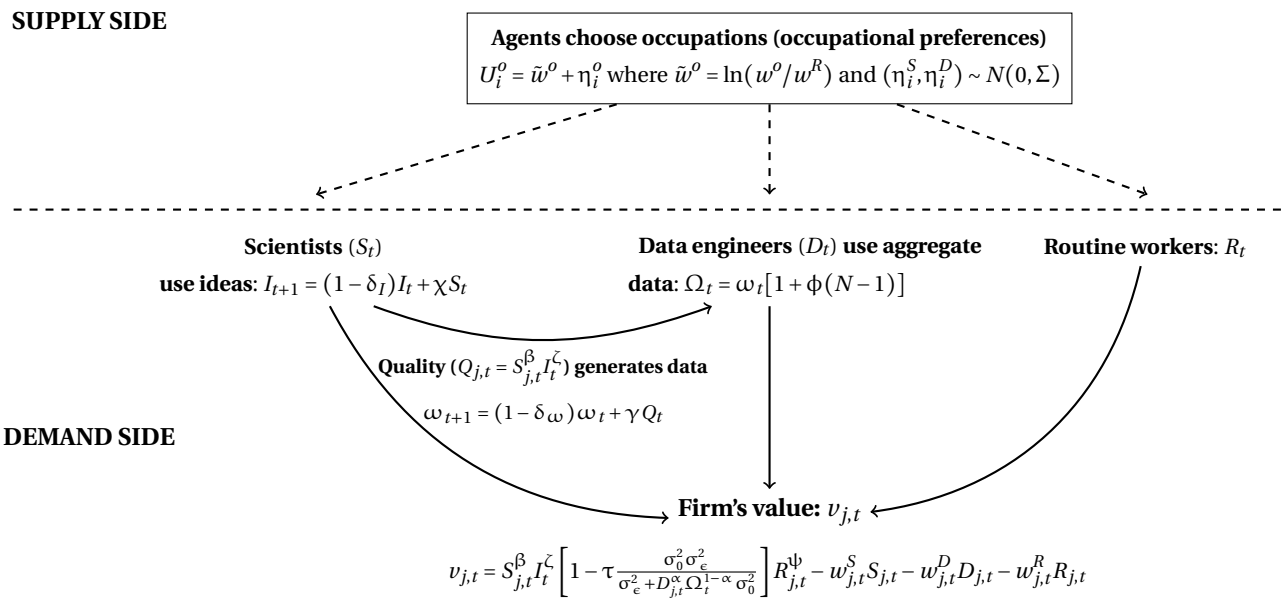
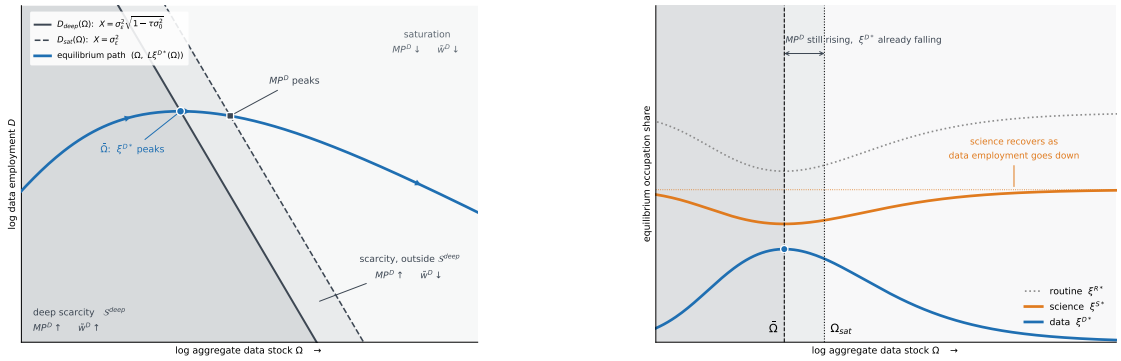



FIGURE 2. Schematic diagram of the model

*Supply side:* Agents observe wages and choose occupations based on wage premiums and idiosyncratic occupational preferences. *Demand side:* Firms hire workers to maximize value. Scientists use the ideas stock to generate quality, which produces data as a byproduct. Data engineers use the aggregate data stock to reduce operational uncertainty. Routine workers provide scale. Go back to main text.

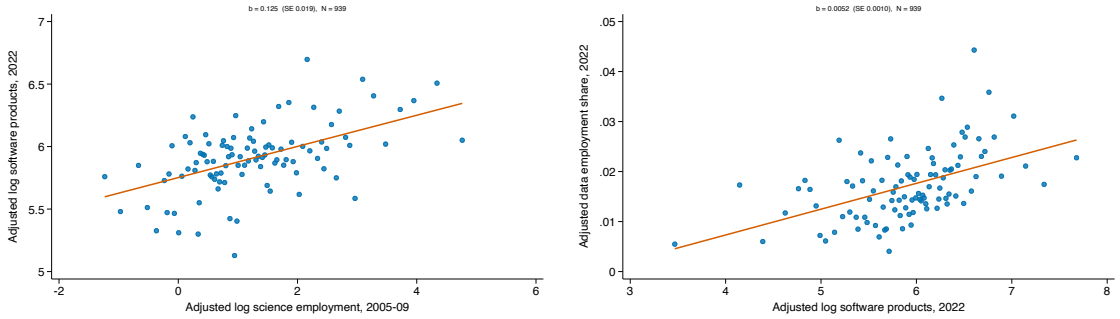



A. The scarcity region in $(\Omega, D)$ space

B. Cannibalization and recovery

FIGURE 3. The scarcity region and the path of talent reallocation

This figure illustrates illustrate Propositions 3 and 4. Panel A partitions the state space $(\Omega, D)$ by the learning scale $X = D^{\alpha}\Omega^{1-\alpha}\sigma_0^2$. The two boundaries are $D_{deep}(\Omega)$, where $X = \sigma_\epsilon^2\sqrt{1-\tau\sigma_0^2}$, and $D_{sat}(\Omega)$, where $X = \sigma_\epsilon^2$; both are decreasing in $\Omega$ and $D_{deep} < D_{sat}$. Statements about the marginal product $MP^D$ hold occupation counts fixed and are those of Proposition 3. Statements about the relative premium $\tilde{w}^D$ and the equilibrium shares $\xi^{o*}$ are equilibrium objects and are those of Proposition 4. The equilibrium path traces $(\Omega, L\xi^{D*}(\Omega))$ as the data stock accumulates. It attains its maximum exactly where it crosses $D_{deep}$, at the threshold $\bar{\Omega}$, and therefore strictly inside the scarcity region $\mathcal{S}$. Panel B plots the same equilibrium for the data, science, and routine shares against $\Omega$. The marginal product of data engineers does not peak until $\Omega_{sat} > \bar{\Omega}$, so over the shaded intermediate band additional data still raises the value of prediction while data employment is already contracting. Axes carry no numerical scale. Go back to main text.

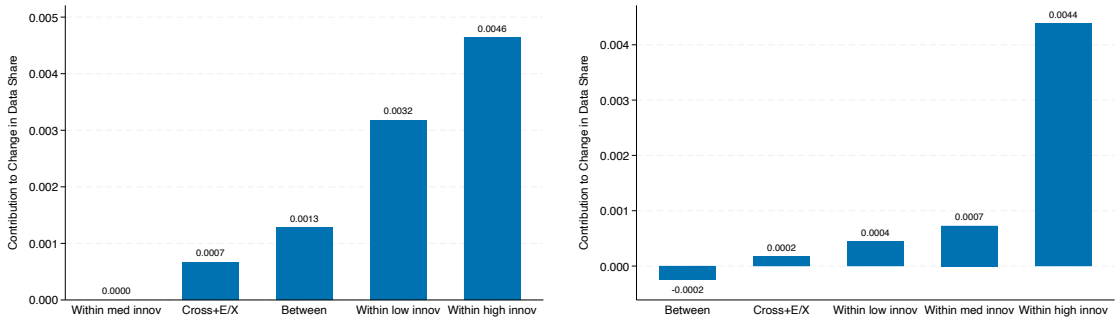



A. Past science predicts data infrastructure

B. Data infrastructure predicts data employment

FIGURE 4. Science employment, data infrastructure, and data employment

Panel A plots the relationship between average log science employment in 2005-2009 and the log number of software products detected by Enlyft in 2022. Both variables are residualized with respect to average log assets in 2005-2009, the average data-employment share in 2005-2009, and two-digit NAICS fixed effects. Panel B plots the relationship between the log number of software products detected by Enlyft and the data-employment share in 2022, residualizing both variables with respect to contemporaneous log assets and two-digit NAICS fixed effects. Dots are means within 100 equally sized bins of the residualized horizontal-axis variable. Lines are fitted using the underlying firm-level observations. Both panels use the same sample of 939 public firms. Heteroskedasticity-robust standard errors are reported in parentheses. Go back to main text.

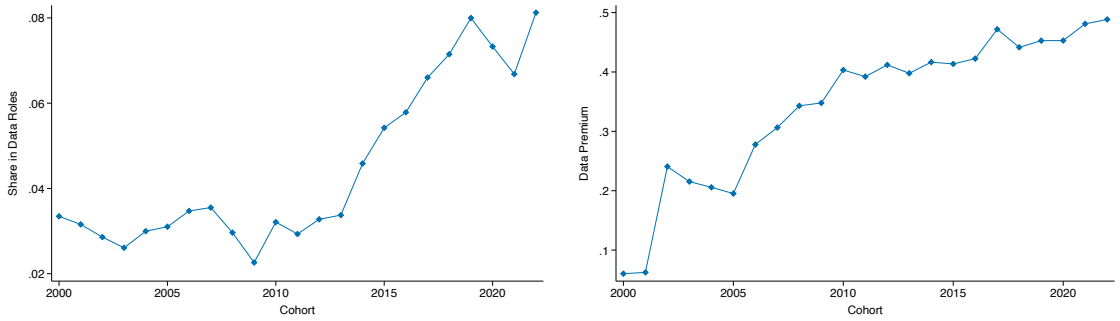



A. ACS: 2005-2022

B. Revelio: 2000-2022

FIGURE 5. Decomposition of the change in the national mean of the data share

This figure decomposes the change in the national data share into three components following equation (16): (i) Within (split by innovation intensity); (ii) Between; (iii) Cross, Entry, and Exit. Panel A uses American Community Survey (ACS) data comparing the 2005-2009 and 2018-2022 5-year samples, with units defined as NAICS 2-digit industry-commuting zone pairs. Panel B uses Revelio data for 2000-2022, with units defined as NAICS 2-digit industry-MSA pairs. In both panels, units are classified into Low, Medium, or High innovation terciles based on their science share in the baseline year (2005 for ACS, 2000 for Revelio), where the science share is science share$_{i,0}$ $=$ $\frac{\text{count of science positions}_{i,0}}{\text{count of science positions}_{i,0}+\text{count of data positions}_{i,0}+\text{count of routine positions}_{i,0}}$ for unit $i$ in the baseline period. The within component is decomposed by summing separately over units in each innovation group: $\text{Within}_t = \sum_{g=1}^{3} \underbrace{\sum_{i \in \mathcal{C}_g} p_{i,t-1} \Delta \xi^D_{i,t}}_{\text{Within}^g_t}$, where $\mathcal{C}_g$ denotes continuing units in innovation group $g \in \{\text{Low}, \text{Medium}, \text{High}\}$.
Units must have at least 10 workers in both periods to be included. Go back to main text.

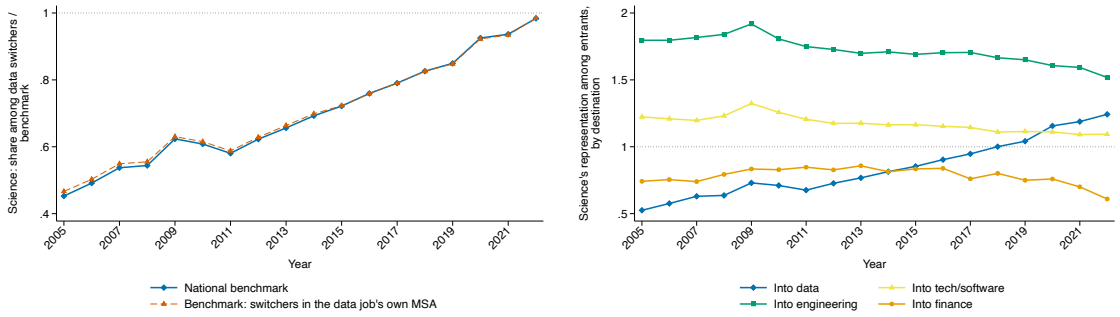



A. Share of data occupation the first year since graduation

B. Data premium the first year since graduation

FIGURE 6. Share of data occupation and compensation the first year since graduation

This figure documents cohort-level patterns in data occupation choice. A cohort is defined based on the year an individual obtains their *first* degree. In this figure, I focus on the fields of mathematics and statistics. The data share is defined as data share$_{t,x} = \frac{\text{count of data positions}_{t,x}}{\text{count of data positions}_{t,x}+\text{count of science positions}_{t,x}+\text{count of routine positions}_{t,x}}$ for cohort graduating in year $t$ and observed $x$ years after graduation. Panel A plots the share of workers in data occupations in their first year after graduation, by cohort year. Panel B plots the data wage premium in the first year after graduation, by cohort year. The data premium is defined as the mean log total compensation of data workers minus the mean log total compensation of routine workers: $\ln(w^D_{t,0}) - \ln(w^R_{t,0})$ for cohort $t$ at year 0 since graduation. Panels A and B cover cohorts graduating between 2000 and 2022. The sample period is 2000-2022. Go back to main text.




A. Science representation among data switchers

B. Science representation by destination

FIGURE 7. The rising occupational affinity between science and data

This figure uses occupation changes observed in Revelio between 2005 and 2022. An occupation change occurs when two consecutive positions held by the same worker belong to different occupation families. The high-skill pool includes science, data, tech/software, engineering, and finance occupations. Panel A plots the share of scientists among workers switching into data divided by the share of scientists among high-skill occupation switchers eligible to enter data; workers originating in data are excluded from the denominator, as defined in equation (18). The solid line uses the national pool of eligible switchers and the dashed line uses eligible switchers in the MSA of the destination data job. Panel B plots, for each destination occupation, the share of scientists among entrants divided by the share of scientists among all high-skill occupation switchers. The same denominator is used across destinations. A value of one indicates proportional representation under the corresponding benchmark. The sample period is 2005-2022. Go back to main text.

## References


Abis, Simona, and Laura Veldkamp. 2024. "The Changing Economics of Knowledge Production." *The Review of Financial Studies* 37 (1): 89–118.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022a. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." *Journal of Labor Economics* 40 (S1): S293–S340.

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. 2022b. "Too Much Data: Prices and Inefficiencies in Data Markets." *American Economic Journal: Microeconomics* 14 (4): 218–256.

Acharya, Viral V, Xuehua Shao, and Zhaoxia Xu. 2026. "AI and Productivity: The Role of Innovation." *Working Paper*.

Akcigit, Ufuk, Craig A Chikis, Emin Dinlersoz, and Nathan Goldschlag. 2026. "Attention (And Money) Is All You Need: Why Universities Are Struggling to Keep AI Talent." *Working Paper*.

Asriyan, Vladimir, and Alexander Kohlhas. 2026. "The Macroeconomics of Data: Scale, Product Choice, and Pricing in the Information Age." *Working Paper*.

Autor, David H, and David Dorn. 2013. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103 (5): 1553–1597.

Autor, David H, Frank Levy, and Richard J Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4): 1279–1333.

Axelson, Ulf, and Philip Bond. 2015. "Wall Street Occupations." *The Journal of Finance* 70 (5): 1949–1996.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. 2025. "Artificial Intelligence Makes Firm Operating Performance Less Volatile." In *AEA Papers and Proceedings*, vol. 115: 35–39, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Bastiat, Frédéric. 1850. *That which is seen, and that which is not seen.*: Guillaumin.

Baumol, William J. 1990. "Entrepreneurship: Productive, Unproductive, and Destructive." *The Journal of Political Economy* 98 (5, Part 1): 893–921.

Begenau, Juliane, Maryam Farboodi, and Laura Veldkamp. 2018. "Big Data in Finance and the Growth of Large Firms." *Journal of Monetary Economics* 97: 71–87.

Benzell, Seth G, and Erik Brynjolfsson. 2019. "Digital Abundance and Scarce Genius: Implications for Wages, Interest Rates, and Growth." *Working Paper*.

Bergemann, Dirk, and Alessandro Bonatti. 2024. "Data, Competition, and Digital Platforms." *American Economic Review* 114 (8): 2553–2595.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan. 2022. "The Economics of Social Data." *The RAND Journal of Economics* 53 (2): 263–296.

Bhaimiya, Sawdah. 2025. "Behind the AI Talent War: Why Tech Giants Are Paying Millions to Top Hires." *CNBC*.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen. 2014. "The Distinct Effects of Information Technology and Communication Technology on Firm Organization." *Management Science* 60 (12): 2859–2885.

Brynjolfsson, Erik, Wang Jin, and Kristina McElheran. 2021. "The Power of Prediction: Predictive Analytics, Workplace Complements, and Business Performance." *Business Economics* 56 (4): 217–239.

Brynjolfsson, Erik, and Kristina McElheran. 2016. "The Rapid Adoption of Data-Driven Decision-Making." *American Economic Review* 106 (5): 133–139.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2017. "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics." *Working Paper*.

Charoenwong, Ben, Zachary T Kowaleski, Alan Kwan, and Andrew G Sutherland. 2024. "RegTech: Technology-driven Compliance and its Effects on Profitability, Operations, and Market Structure." *Journal of Financial Economics* 154: 103792.

Cong, Lin William, Wenshi Wei, Danxia Xie, and Longtian Zhang. 2022. "Endogenous Growth under Multiple Uses of Data." *Journal of Economic Dynamics and Control* 141: 104395.

Cong, Lin William, Danxia Xie, and Longtian Zhang. 2021. "Knowledge Accumulation, Privacy, and Growth in a Data Economy." *Management Science* 67 (10): 6480–6492.

David, Joel M, Hugo A Hopenhayn, and Venky Venkateswaran. 2016. "Information, Misallocation, and Aggregate Productivity." *The Quarterly Journal of Economics* 131 (2): 943–1005.

Demirer, Mert, Diego J Jiménez Hernández, Dean Li, and Sida Peng. 2026. "Data, Privacy Laws and Firm Production: Evidence from the GDPR." *Working Paper*.

Eeckhout, Jan, and Laura Veldkamp. 2025. "Data and Markups: A Macro-Finance Perspective." *Working Paper*.

Farboodi, Maryam, Nima Haghpanah, and Ali Shourideh. 2025. "Good Data and Bad Data: The Welfare Effects of Price Discrimination." *Working Paper*.

Farboodi, Maryam, Andrew Koh, and Anchi Xia. 2026. "Data-Driven Automation." *Working Paper*.

Farboodi, Maryam, Peter Kondor, and Pablo Kurlat. 2026. "Equilibrium Spillover of Big Data." *Working Paper*.

Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp. 2019. "Big Data and Firm Dynamics." *AEA Papers and Proceedings* 109: 38–42.

Farboodi, Maryam, and Laura Veldkamp. 2023. "Data and Markets." *Annual Review of Economics* 15 (1): 23–40.

Farboodi, Maryam, and Laura Veldkamp. 2025. "A Model of the Data Economy." *Working Paper*.

Fisher, Irving. 1930. *The Theory of Interest as Determined by Impatience to Spend Income and Opportunity to Spend It.*: Macmillan.

Forman, Chris. 2005. "The Corporate Digital Divide: Determinants of Internet Adoption." *Management Science* 51 (4): 641–654.

Glode, Vincent, and Guillermo Ordoñez. 2025. "Technological Progress and Rent Seeking." *The Review of Financial Studies* 38 (4): 1259–1289.

Gofman, Michael, and Zhao Jin. 2024. "Artificial Intelligence, Education, and Entrepreneurship." *The Journal of Finance* 79 (1): 631–667.

Hernandez, Daniela, and Rachael King. 2016. "Universities' AI Talent Poached by Tech Giants." *The Wall Street Journal*.

Jones, Charles I, and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review* 110 (9): 2819–2858.

Kehoe, Patrick J, Bradley J Larsen, and Elena Pastorino. 2018. "Dynamic Competition in the Era of Big Data."

Kurlat, Pablo. 2019. "The Social Value of Financial Expertise." *American Economic Review* 109 (2): 556–590.

Murphy, Kevin M, Andrei Shleifer, and Robert W Vishny. 1991. "The Allocation of Talent: Implications for Growth." *The Quarterly Journal of Economics* 106 (2): 503–530.

Patil, THDJ, and T Davenport. 2012. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review* 90 (10): 70–76.

Philippon, Thomas. 2010. "Financiers versus Engineers: Should the Financial Sector be Taxed or Subsi-

dized?" *American Economic Journal: Macroeconomics* 2 (3): 158–182.

Prat, Andrea, and Tommaso Valletti. 2022. "Attention Oligopoly." *American Economic Journal: Microeconomics* 14 (3): 530–557.

Ramsey, Mike, and Douglas MacMillan. 2015. "Carnegie Mellon Reels after Uber Lures Away Researchers." *Wall Street Journal*.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5, Part 2): S71–S102.

Roy, Andrew Donald. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3 (2): 135–146.

Veldkamp, Laura, and Cindy Chung. 2024. "Data and the Aggregate Economy." *Journal of Economic Literature* 62 (2): 458–484.

Zolas, Nikolas, Zachary Kroff, Erik Brynjolfsson, Kristina McElheran, David N Beede, Cathy Buffington, Nathan Goldschlag, Lucia Foster, and Emin Dinlersoz. 2021. "Advanced Technologies Adoption and Use by US Firms: Evidence from the Annual Business Survey." *Working Paper*.

## Appendix A. Proofs for the Main Body of the Paper

### A.1. Proof of Proposition 1

PROOF. In Period 0, the aggregate data stock is $\Omega_0 = 0$. Thus, the marginal product of data engineers is zero. Hence, in a competitive equilibrium, wages must satisfy $w_0^D = 0$, and the equilibrium share of data engineers is $(\xi_0^D)^* = 0$. Formally, at $w_0^D = 0$ the data premium is the limiting value $\tilde{w}_0^D = \ln(w_0^D/w_0^R) = -\infty$, so the occupational supply of data engineers is zero; and since $D$ does not affect production at $\Omega_0 = 0$, the firm is indifferent over $D$ at zero wage, so $D_0 = 0$ belongs to its demand correspondence and the data market clears.

The equilibrium is defined by the value of $\xi$, that equates the wage premium implied by labor demand with the wage premium required by labor supply.

From the firm's first-order conditions, competitive firms equate marginal products to wages. The relative wage premium demanded by the firm is determined by the ratio of marginal products

$$\frac{w_0^S}{w_0^R} = \frac{\beta S_0^{\beta-1} I_0^{\zeta} R_0^{\psi}}{\psi S_0^{\beta} I_0^{\zeta} R_0^{\psi-1}} = \frac{\beta}{\psi}\frac{R_0}{S_0} = \frac{\beta}{\psi}\frac{(1-\xi)L}{\xi L} = \frac{\beta}{\psi}\frac{1-\xi}{\xi}.$$

Taking logs, the demand-side wage premium is

$$\tilde{w}_{\text{dem}}(\xi) \equiv \ln\left(\frac{w_0^S}{w_0^R}\right) = \ln\left(\frac{\beta}{\psi}\right) + \ln(1-\xi) - \ln(\xi).$$

Note that $\lim_{\xi\to 0^+} \tilde{w}_{\text{dem}}(\xi) = +\infty$ and $\lim_{\xi\to 1^-} \tilde{w}_{\text{dem}}(\xi) = -\infty$. Furthermore, the function is strictly decreasing

$$\tilde{w}'_{\text{dem}}(\xi) = -\frac{1}{1-\xi} - \frac{1}{\xi} < 0.$$

From occupational choice, the supply of scientists is determined by the probability that the scientist wage premium exceeds the idiosyncratic tastes. With only two active occupations, the choice probability is given by the standard normal CDF $\Upsilon(\cdot)$

$$\xi = \Upsilon\left(\frac{\tilde{w}}{\sigma_S}\right) \quad \Longleftrightarrow \quad \tilde{w}_{\text{sup}}(\xi) = \sigma_S \Upsilon^{-1}(\xi).$$

Note that $\lim_{\xi\to 0^+} \tilde{w}_{\text{sup}}(\xi) = -\infty$ and $\lim_{\xi\to 1^-} \tilde{w}_{\text{sup}}(\xi) = +\infty$. Since the inverse CDF $\Upsilon^{-1}$ is strictly increasing, the supply curve is strictly increasing. Using the inverse function theorem, I get

$$\tilde{w}'_{\text{sup}}(\xi) = \sigma_S \frac{1}{\upsilon(\Upsilon^{-1}(\xi))} > 0,$$

where $\upsilon(\cdot)$ is the standard normal PDF, which is strictly positive everywhere.

An equilibrium exists if there is a $\xi^* \in (0,1)$ such that the demand premium equals the

supply premium. Define the excess wage premium function $\Gamma(\xi)$ as

$$\Gamma(\xi) \equiv \tilde{w}_{\text{dem}}(\xi) - \tilde{w}_{\text{sup}}(\xi).$$

Substituting the expressions derived above, I have

$$\Gamma(\xi) = \left[\ln\left(\frac{\beta}{\psi}\right) + \ln(1-\xi) - \ln(\xi)\right] - \left[\sigma_S \Upsilon^{-1}(\xi)\right].$$

The function $\Gamma(\xi)$ is continuous on the interval $(0,1)$. Examining the limits gives

i. As $\xi \to 0^+$, $\tilde{w}_{\text{dem}} \to +\infty$ and $\tilde{w}_{\text{sup}} \to -\infty$, so $\Gamma(\xi) \to +\infty$.

ii. As $\xi \to 1^-$, $\tilde{w}_{\text{dem}} \to -\infty$ and $\tilde{w}_{\text{sup}} \to +\infty$, so $\Gamma(\xi) \to -\infty$.

By the Intermediate Value Theorem, there exists at least one root $\xi^* \in (0,1)$ such that $\Gamma(\xi^*) = 0$.

To show uniqueness, consider the derivative with respect to $\xi$

$$\Gamma'(\xi) = \tilde{w}'_{\text{dem}}(\xi) - \tilde{w}'_{\text{sup}}(\xi).$$

Since $\tilde{w}'_{\text{dem}}(\xi) < 0$ and $\tilde{w}'_{\text{sup}}(\xi) > 0$, it follows that $\Gamma'(\xi) < 0$ for all $\xi \in (0,1)$. The function $\Gamma(\xi)$ is strictly decreasing, which implies that the root $\xi^*$ is unique.

Thus, there exists a unique equilibrium with occupation shares $(\xi_0^S)^*, (\xi_0^R)^* \in (0,1)$ and $(\xi_0^D)^* = 0$. □

### A.2. Proof of Proposition 2

PROOF. The proof relies on a potential function approach to invoke convex optimization arguments that establish equilibrium existence and uniqueness. I proceed in eight steps. First, (i) I express the supply side, that is occupational choice, as gradients of a convex function. Second, (ii) I characterize the log wage premia using the firm's FOCs. Third, (iii) I construct the potential function of the firm. Fourth, (iv) I show that this potential function is strictly convex. Fifth, (v) I define the equilibrium as an optimization problem of some function. Sixth, (vi) I verify that the solution is interior. Seventh, (vii) I show that the solution is unique. Finally, (viii) I verify that the implied input vector globally maximizes firm profit at the equilibrium wages, so the fixed point is a competitive equilibrium.

**Part (i): Occupational choice as gradients of a convex function.**

In competitive labor markets with identical firms, wages are common across firms. Fix $I_1 > 0$ and $\Omega_1 > 0$. Let $\Delta^2 = \{(\xi^S, \xi^D) \in \mathbb{R}^2_+ : \xi^S + \xi^D \leq 1\}$ and $\xi^R = 1 - \xi^S - \xi^D$. Recall that, under the per-firm normalization of Section 3, $L$ is the per-firm mass of workers, such that $(S_1, D_1, R_1) = (L\xi^S, L\xi^D, L\xi^R)$ are the representative firm's inputs; in particular, the data term $(L\xi^D)^\alpha \Omega_1^{1-\alpha}$ below is the firm-level signal scale of Section 2.1.

Let $(\eta^S, \eta^D)$ be bivariate normal with covariance matrix $\Sigma$ and full support, and define the

expected surplus

$$G(\tilde{w}^S, \tilde{w}^D) = \mathbb{E}\Big[\max\{0, \tilde{w}^S + \eta^S, \tilde{w}^D + \eta^D\}\Big].$$

By the Williams-Daly-Zachary theorem, $G$ is convex and differentiable, with gradient equal to choice probabilities

$$\nabla G(\tilde{w}^S, \tilde{w}^D) = \big(P_S(\tilde{w}^S, \tilde{w}^D),\, P_D(\tilde{w}^S, \tilde{w}^D)\big) \in \text{int}(\Delta^2).$$

Let $G^*$ be the convex conjugate of $G$. Let $G^*$ be the convex conjugate of $G$. Parts (i)-(ii) of Lemma A2, proved directly from the Gaussian choice probabilities below, establish that $\nabla G$ is a bijection from $\mathbb{R}^2$ onto $\text{int}(\Delta^2)$ and that $G^*$ is twice continuously differentiable there. In particular, the inverse satisfies the duality relation

$$\nabla G^*(\xi^S, \xi^D) = (\tilde{w}^S, \tilde{w}^D) \quad \Longleftrightarrow \quad (\xi^S, \xi^D) = \nabla G(\tilde{w}^S, \tilde{w}^D).$$

Thus the occupational-choice equilibrium condition $(\xi^S, \xi^D) = (P_S(\tilde{w}), P_D(\tilde{w}))$ is equivalent to

$$(\tilde{w}^S, \tilde{w}^D) = \nabla G^*(\xi^S, \xi^D).$$

**Part (ii): Firm-side log wage premia.** Wages satisfy $w_1^S = \beta \pi_1^E / S_1$, $w_1^R = \psi \pi_1^E / R_1$ and

$$w_1^D = S_1^\beta I_1^\zeta R_1^\psi \cdot \tau \frac{\sigma_0^4 \sigma_\epsilon^2 \alpha D_1^{\alpha-1} \Omega_1^{1-\alpha}}{\left(\sigma_\epsilon^2 + D_1^\alpha \Omega_1^{1-\alpha} \sigma_0^2\right)^2},$$

with $\pi_1^E = S_1^\beta I_1^\zeta R_1^\psi \left[1 - \tau \sigma^2_{post}(D_1, \Omega_1)\right]$ and $\sigma^2_{post}(D, \Omega) = \frac{\sigma_0^2 \sigma_\epsilon^2}{\sigma_\epsilon^2 + D^\alpha \Omega^{1-\alpha} \sigma_0^2}$. Because $1 - \tau \sigma^2_{post}(D_1, \Omega_1) > 0$ and $\tau > 0$, all wages are strictly positive; positivity of $\tau$ is what makes $w_1^D > 0$ and the constant $\ln \tau$ in $C_D$ below well defined.

Therefore, for $\xi \in \text{int}(\Delta^2)$,

$$\tilde{w}^S(\xi) = \ln \frac{w_1^S}{w_1^R} = \ln \frac{\beta}{\psi} + \ln \frac{\xi^R}{\xi^S},$$

and, after cancelling common factors,

$$\tilde{w}^D(\xi) = \ln \frac{w_1^D}{w_1^R} = C_D + \ln \xi^R + (\alpha - 1) \ln \xi^D - 2 \ln\left(\sigma_\epsilon^2 + (L\xi^D)^\alpha \Omega_1^{1-\alpha} \sigma_0^2\right) - \ln\left(1 - \tau \sigma^2_{post}(L\xi^D, \Omega_1)\right),$$

where $C_D$ collects constants depending on $L$, $\Omega_1$, and model parameters. I have $C_D = \alpha \ln L + (1 - \alpha) \ln \Omega_1 + \ln \alpha + \ln \tau + 4 \ln \sigma_0 + 2 \ln \sigma_\epsilon - \ln \psi$.

**Part (iii): Construct a firm-side potential.** Define $U : \text{int}(\Delta^2) \to \mathbb{R}$ by

$$U(\xi^S, \xi^D) = \xi^S \ln \xi^S + \xi^R \ln \xi^R - \ln\left(\frac{\beta}{\psi}\right) \xi^S + (1 - \alpha)\, \xi^D \ln \xi^D$$

$$+\int_0^{\xi^D}\Big[2\ln\big(\sigma_\epsilon^2+(Lu)^\alpha\Omega_1^{1-\alpha}\sigma_0^2\big)+\ln\big(1-\tau\sigma^2_{post}(Lu,\Omega_1)\big)\Big]\,du+c\,\xi^D,$$

where $c=\alpha-C_D$ is chosen so that $-\partial U/\partial\xi^D$ matches $\tilde{w}^D(\xi)$.

A direct differentiation using $\partial\xi^R/\partial\xi^S=\partial\xi^R/\partial\xi^D=-1$ yields

$$-\frac{\partial U}{\partial\xi^S}(\xi)=\tilde{w}^S(\xi),$$

and

$$-\frac{\partial U}{\partial\xi^D}(\xi)=\tilde{w}^D(\xi).$$

Hence, the firm-side implied premia satisfy

$$(\tilde{w}^S(\xi),\tilde{w}^D(\xi))=-\nabla U(\xi).$$

**Part (iv): Strict convexity of $U$.** Now, I show that $U$ is strictly convex on $\mathrm{int}(\Delta^2)$ by verifying positive definiteness of its Hessian.

The full potential $U$ can be written as

$$U(\xi)=U_0(\xi)+h(\xi^D),$$

where

$$U_0(\xi)=\xi^S\ln\xi^S+\xi^R\ln\xi^R+(1-\alpha)\xi^D\ln\xi^D-\ln\Big(\frac{\beta}{\psi}\Big)\xi^S+c\,\xi^D$$

$$h(\xi^D)=\int_0^{\xi^D}f(u)\,du=\int_0^{\xi^D}\Big[2\ln(\sigma_\epsilon^2+(Lu)^\alpha\Omega_1^{1-\alpha}\sigma_0^2)+\ln(1-\tau\sigma^2_{post}(Lu,\Omega_1))\Big]\,du$$

The Hessian of $U_0(\xi)$ is

$$H_{U_0}=\begin{pmatrix}\dfrac{1}{\xi^S}+\dfrac{1}{\xi^R} & \dfrac{1}{\xi^R}\\[2ex] \dfrac{1}{\xi^R} & \dfrac{1-\alpha}{\xi^D}+\dfrac{1}{\xi^R}\end{pmatrix}.$$

This matrix is positive definite on $\mathrm{int}(\Delta^2)$. To see this, notice that the trace is positive: $\dfrac{1}{\xi^S}+\dfrac{2}{\xi^R}+\dfrac{1-\alpha}{\xi^D}>0$. Further, the determinant is positive too: $\dfrac{1-\alpha}{\xi^S\xi^D}+\dfrac{1}{\xi^S\xi^R}+\dfrac{1-\alpha}{\xi^D\xi^R}>0$, since $\alpha\in(0,1)$.

For the integral term $h(\xi^D)$, I have $h''(\xi^D)=f'(\xi^D)$. Since both arguments of the logarithms are strictly increasing in $u$, I have $f'(u)>0$, so $h$ is strictly convex in $\xi^D$.

Combining, $U=U_0+h$ is strictly convex on $\mathrm{int}(\Delta^2)$.

**Part (v): Equilibrium as an optimization problem.** Define the objective function on $\mathrm{int}(\Delta^2)$

$$J(\xi)\equiv U(\xi)+G^*(\xi).$$

Since $U$ is strictly convex, by Part (iv), and $G^*$ is convex, $J$ is strictly convex on $\mathrm{int}(\Delta^2)$.

**Part (vi): Existence and interiorness.** Consider the convex program

$$\min_{\xi \in \Delta^2} J(\xi) = U(\xi) + G^*(\xi).$$

Extend $U$ to $\Delta^2$ by the convention $x \ln x \equiv 0$ at $x = 0$. Further, because I assume $1 - \tau\sigma_0^2 > 0$, the integral term is well-defined at $u = 0$. It follows that $U$ is finite and continuous on $\Delta^2$. The convex conjugate $G^*$ is proper and lower semicontinuous. Hence $J$ is lower semicontinuous on the compact set $\Delta^2$, so $J$ attains a minimizer $\xi^* \in \Delta^2$.

Now, I show that $\xi^* \in \text{int}(\Delta^2)$ by ruling out boundary minimizers using directional derivatives. Let $\bar{\xi} \equiv \nabla G(0,0) = (P_S(0,0), P_D(0,0))$. By full support of $(\eta^S, \eta^D)$, I have $\bar{\xi} \in \text{int}(\Delta^2)$. Moreover, $G^*(\bar{\xi}) < \infty$ by the Fenchel-Young inequality. Since $\bar{\xi} \in \partial G(0,0)$, Fenchel-Young holds with equality

$$G^*(\bar{\xi}) = \bar{\xi} \cdot (0,0) - G(0,0) = -G(0,0) \in \mathbb{R}.$$

Suppose, toward a contradiction, that $\xi^* \in \partial\Delta^2$, and consider the feasible direction $d \equiv \bar{\xi} - \xi^*$, which points strictly into the simplex. If $G^*(\xi^*) = +\infty$, then $\xi^*$ cannot minimize $J$, since $J(\bar{\xi}) = U(\bar{\xi}) + G^*(\bar{\xi}) < \infty$; so assume $G^*(\xi^*) < \infty$. Convexity of $G^*$ then bounds the one-sided directional derivative *from above* by the chord slope,

$$G^{*\prime}(\xi^*; d) \leq G^*(\bar{\xi}) - G^*(\xi^*) < +\infty,$$

even though $G^{*\prime}(\xi^*; d)$ itself may equal $-\infty$ at the boundary; an upper bound is all the argument requires. Consider the following three cases.

i. If $\xi^{S*} = 0$ and $\xi^{R*} > 0$, then along any sequence $\xi \to \xi^*$ with $\xi^S > 0$, Part (ii) gives $\tilde{w}^S(\xi) = \ln(\beta/\psi) + \ln(\xi^R/\xi^S) \to +\infty$, hence $\partial U/\partial \xi^S(\xi) = -\tilde{w}^S(\xi) \to -\infty$. Because $d^S = \bar{\xi}^S - \xi^{S*} = \bar{\xi}^S > 0$, the one-sided directional derivative satisfies $U'(\xi^*; d) = -\infty$.

ii. If $\xi^{D*} = 0$ and $\xi^{R*} > 0$, then Part (ii) implies $\tilde{w}^D(\xi) \to +\infty$ as $\xi^D \to 0^+$. Recall that $(\alpha - 1)\ln \xi^D \to +\infty$ for $\alpha \in (0,1)$. Hence, $\partial U/\partial \xi^D(\xi) = -\tilde{w}^D(\xi) \to -\infty$. Because $d^D = \bar{\xi}^D > 0$, I again have $U'(\xi^*; d) = -\infty$.

iii. If $\xi^{R*} = 0$ (i.e. $\xi^{S*} + \xi^{D*} = 1$), then $d$ increases $\xi^R$ since $\bar{\xi}^R > 0$. As $\xi^R \to 0^+$, the term $\xi^R \ln \xi^R$ makes the marginal effect of increasing $\xi^R$ equal to $1 + \ln \xi^R \to -\infty$, so moving in direction $d$ yields $U'(\xi^*; d) = -\infty$.

In all cases, $U'(\xi^*; d) = -\infty$ while $G^{*\prime}(\xi^*; d)$ is bounded above by the chord slope, hence

$$J'(\xi^*; d) = U'(\xi^*; d) + G^{*\prime}(\xi^*; d) = -\infty < 0,$$

contradicting that $\xi^*$ minimizes $J$ over $\Delta^2$. Therefore $\xi^* \in \text{int}(\Delta^2)$.

**Part (vii): Uniqueness and characterization of equilibrium.** Since $\xi^* \in \text{int}(\Delta^2)$, by Part (vi), and $J$ is strictly convex on $\text{int}(\Delta^2)$, by Part (v), the minimizer is unique. At this unique interior

minimizer, $\nabla J(\xi^*) = 0$,

$$\nabla U(\xi^*) + \nabla G^*(\xi^*) = 0 \quad \Longleftrightarrow \quad -\nabla U(\xi^*) = \nabla G^*(\xi^*).$$

By Parts (i) and (iii), this is equivalent to

$$(\tilde{w}^S(\xi^*), \tilde{w}^D(\xi^*)) = \nabla G^*(\xi^*) \quad \Longleftrightarrow \quad (\xi^{S*}, \xi^{D*}) = \nabla G(\tilde{w}^S(\xi^*), \tilde{w}^D(\xi^*)),$$

which corresponds to the occupational market-clearing condition given firm-implied wages.

Setting $\xi^{R*} = 1 - \xi^{S*} - \xi^{D*}$ yields the unique equilibrium $(\xi^{S*}, \xi^{D*}, \xi^{R*}) \in \text{int}(\Delta^2)$.

**Part (viii): Global firm optimality.** At the equilibrium wages $(w_1^S, w_1^D, w_1^R)$ implied by Part (ii), the input vector $(L\xi^{S*}, L\xi^{D*}, L\xi^{R*})$ satisfies the firm's first-order conditions by construction. Because $\alpha + \beta + \psi \leq 1$, Lemma A4 shows that $z \mapsto \pi^E(z) - w \cdot z$ is concave on $\mathbb{R}^3_+$, so this stationary point is a global maximizer of profit. The fixed point is therefore a competitive equilibrium in the full sense: the firm optimizes globally, workers optimize, and the labor market clears. Moreover, equilibrium firm value is strictly positive, $v = (1-\beta-\psi)\pi^E - w^D D \geq \alpha\pi^E - w^D D > 0$, by the computation in Section 2.1. □

### A.3. Proposition A1

PROPOSITION A1 (Period 2 equilibrium). *Given ideas stock, $I_2 > 0$, and aggregate data stock, $\Omega_2 > 0$, and $\alpha \in (0,1)$, there exists a unique equilibrium with occupation shares $(\xi_2^S, \xi_2^D, \xi_2^R)^* \in \text{relint}(\Delta^2)$, where $\Delta^2 = \{(\xi^S, \xi^D, \xi^R) \in \mathbb{R}^3_+ : \xi^S + \xi^D + \xi^R = 1\}$.*

PROOF. Every step of the proof of Proposition 2 applies as before by simply replacing the subscript 1 by 2: the construction uses only $I_2 > 0$, $\Omega_2 > 0$, and the date enters the potential $J$ only through the positive constants $(I_2, \Omega_2)$. In particular, existence, interiority, and uniqueness (Parts (i)-(vii)) and global firm optimality (Part (viii), via Lemma A4) carry over unchanged. □

### A.4. Proof of Proposition 3

PROOF. The proof proceeds in two steps: (i) I show that the marginal product of data engineers is single peaked; (ii) I derive the elasticity formula. I omit the firm subscripts $j$ for simplicity.

**Part (i):** For fixed occupation counts $(S_t, R_t, D_t)$, the marginal product of data engineers is given by

$$\text{MP}_t^D = C \frac{\Omega_t^{1-\alpha}}{(\sigma_\epsilon^2 + D_t^\alpha \Omega_t^{1-\alpha} \sigma_0^2)^2},$$

where $C = S_t^\beta I_t^\zeta R_t^\psi \tau\alpha\sigma_0^4\sigma_\epsilon^2 D_t^{\alpha-1} > 0$ collects terms independent of $\Omega_t$. Let $X \equiv D_t^\alpha \Omega_t^{1-\alpha}\sigma_0^2$. I can

now substitute $\Omega_t^{1-\alpha} = \frac{X}{D_t^\alpha \sigma_0^2}$ into the expression for $\mathrm{MP}_t^D$ to obtain

$$\mathrm{MP}_t^D = \frac{C}{D_t^\alpha \sigma_0^2} \frac{X}{(\sigma_\epsilon^2 + X)^2}.$$

Since $\frac{C}{D_t^\alpha \sigma_0^2}$ is a positive constant with respect to $\Omega_t$, the sign of $\partial \mathrm{MP}_t^D / \partial \Omega_t$ depends only on the derivative of $f(X) = \frac{X}{(\sigma_\epsilon^2 + X)^2}$. Differentiating with respect to $X$ yields

$$f'(X) = \frac{1 \cdot (\sigma_\epsilon^2 + X)^2 - X \cdot 2(\sigma_\epsilon^2 + X)}{(\sigma_\epsilon^2 + X)^4} = \frac{\sigma_\epsilon^2 - X}{(\sigma_\epsilon^2 + X)^3}.$$

Since $\partial X / \partial \Omega_t > 0$, the sign of the derivative with respect to $\Omega_t$ is determined by $\sigma_\epsilon^2 - X$. I have three cases:

i. If $X < \sigma_\epsilon^2$ (scarcity), the derivative is strictly positive;
ii. If $X = \sigma_\epsilon^2$ (saturation threshold), the derivative is zero;
iii. If $X > \sigma_\epsilon^2$ (saturation), the derivative is strictly negative.

Thus, the function is single-peaked with a global maximum exactly at $X = \sigma_\epsilon^2$, which corresponds to $D_t = D_{\mathrm{sat}}(\Omega_t)$.

**Part (ii):** Taking logs of the marginal product expression

$$\ln \mathrm{MP}_t^D = \ln C + (1-\alpha) \ln \Omega_t - 2 \ln(\sigma_\epsilon^2 + D_t^\alpha \Omega_t^{1-\alpha} \sigma_0^2).$$

Differentiating with respect to $\ln \Omega_t$

$$\begin{aligned} \varepsilon_{\mathrm{MP}^D, \Omega} &= (1-\alpha) - 2 \frac{D_t^\alpha \Omega_t^{1-\alpha} \sigma_0^2}{\sigma_\epsilon^2 + D_t^\alpha \Omega_t^{1-\alpha} \sigma_0^2} (1-\alpha) \\ &= (1-\alpha) \left[ 1 - \frac{2X}{\sigma_\epsilon^2 + X} \right] \\ &= (1-\alpha) \frac{\sigma_\epsilon^2 - X}{\sigma_\epsilon^2 + X}. \end{aligned}$$

Substituting $X = D_t^\alpha \Omega_t^{1-\alpha} \sigma_0^2$ yields the expression in the proposition. □

### A.5. Proof of Proposition 4

I first establish two lemmas. Lemma A1, which provides a closed form for the deep-scarcity region. Lemma A2 which delivers the sign of every entry of the equilibrium Hessian without auxiliary assumptions. These two lemmas are the key building blocks to the proof of Proposition 4.

LEMMA A1 (Closed Form of the Deep-Scarcity Region). *For every* $(D,\Omega)$ *with* $\Omega > 0$,

$$X(D,\Omega) < \sigma_\epsilon^2\big(1-\tau\sigma^2_{post}(D,\Omega)\big) \qquad \Longleftrightarrow \qquad X(D,\Omega) < \sigma_\epsilon^2\sqrt{1-\tau\sigma_0^2}.$$

*Accordingly,* $\mathcal{S}^{deep} = \{(D,\Omega) : X(D,\Omega) < \sigma_\epsilon^2\sqrt{1-\tau\sigma_0^2}\} \subset \mathcal{S}$, *with strict inclusion since* $\sqrt{1-\tau\sigma_0^2} < 1$.

PROOF. Substituting $\sigma^2_{post} = \sigma_0^2\sigma_\epsilon^2/(\sigma_\epsilon^2 + X)$ and multiplying through by $\sigma_\epsilon^2 + X > 0$,

$$X < \sigma_\epsilon^2\Big(1-\frac{\tau\sigma_0^2\sigma_\epsilon^2}{\sigma_\epsilon^2+X}\Big) \Longleftrightarrow X\big(\sigma_\epsilon^2+X\big) < \sigma_\epsilon^2\big(\sigma_\epsilon^2+X\big) - \tau\sigma_0^2\sigma_\epsilon^4 \Longleftrightarrow X^2 < \sigma_\epsilon^4\big(1-\tau\sigma_0^2\big),$$

and the right-hand side is positive by the maintained assumption $1-\tau\sigma_0^2 > 0$. Taking square roots (both sides nonnegative) gives the claim. The same computation shows the identity

$$\sigma_\epsilon^2\big(1-\tau\sigma^2_{post}\big) - X = \frac{\sigma_\epsilon^4\big(1-\tau\sigma_0^2\big) - X^2}{\sigma_\epsilon^2 + X}, \tag{A1}$$

which is used repeatedly below. □

LEMMA A2 (Substitution Structure of Occupational Supply). *Let* $(\eta^S,\eta^D)$ *be bivariate normal with* $\sigma_S,\sigma_D > 0$ *and* $\nu \in (-1,1)$, *with density* $f > 0$ *on* $\mathbb{R}^2$, *and let* $G$ *be the occupational-choice surplus function of Appendix Section A.2, that is the proof of Proposition 2. Define, for* $\tilde{w} = (\tilde{w}^S,\tilde{w}^D) \in \mathbb{R}^2$, *the boundary masses*

$$m_{S0}(\tilde{w}) \equiv \int_{-\infty}^{-\tilde{w}^D} f\big(-\tilde{w}^S, v\big)\,dv, \qquad m_{D0}(\tilde{w}) \equiv \int_{-\infty}^{-\tilde{w}^S} f\big(u,-\tilde{w}^D\big)\,du,$$

$$m_{SD}(\tilde{w}) \equiv \int_{-\tilde{w}^S}^{\infty} f\big(u,\, u+\tilde{w}^S-\tilde{w}^D\big)\,du.$$

*These are the densities of agents indifferent between science and routine work, between data engineering and routine work, and between science and data engineering, respectively. Then,*

*(i) G is twice continuously differentiable with*

$$M(\tilde{w}) \equiv \nabla^2 G(\tilde{w}) = \begin{pmatrix} m_{S0}+m_{SD} & -m_{SD} \\ -m_{SD} & m_{D0}+m_{SD} \end{pmatrix},$$

*all three masses are strictly positive, and* $\det M = m_{S0}m_{D0} + m_{SD}\big(m_{S0}+m_{D0}\big) > 0$, *so* $M(\tilde{w})$ *is symmetric positive definite at every* $\tilde{w}$.

*(ii)* $G^*$ *is twice continuously differentiable on* $\operatorname{int}(\Delta^2)$ *with* $\nabla^2 G^*(\xi) = M\big(\tilde{w}(\xi)\big)^{-1}$, *where* $\tilde{w}(\xi) = \nabla G^*(\xi)$; *in particular*

$$\big(\nabla^2 G^*\big)_{SD} = \frac{m_{SD}}{\det M} > 0, \qquad \big(\nabla^2 G^*\big)_{SS} - \big(\nabla^2 G^*\big)_{SD} = \frac{m_{D0}}{\det M} > 0.$$

*(iii) Consequently, at every $\xi \in \mathrm{int}(\Delta^2)$ and every state, the Hessian $H = \nabla^2_{\xi\xi} J$ of the equilibrium potential of Appendix Section A.2 satisfies*

$$H > 0, \qquad H_{SD} = \frac{1}{\xi_R} + \left(\nabla^2 G^*\right)_{SD} > 0, \qquad H_{SS} - H_{SD} = \frac{1}{\xi_S} + \frac{m_{D0}}{\det M} > 0.$$

PROOF. **Part (i):** By the Williams-Daly-Zachary theorem (Appendix Section A.2, Part (i)), $\nabla G(\tilde{w}) = (P_S(\tilde{w}), P_D(\tilde{w}))$ with

$$P_S(\tilde{w}) = \int_{-\tilde{w}^S}^{\infty} \int_{-\infty}^{u+\tilde{w}^S-\tilde{w}^D} f(u,v)\, dv\, du, \qquad P_D(\tilde{w}) = \int_{-\tilde{w}^D}^{\infty} \int_{-\infty}^{v+\tilde{w}^D-\tilde{w}^S} f(u,v)\, du\, dv.$$

Since $f$ is Gaussian, differentiation under the integral sign is justified by dominated convergence (the integrands and their partial derivatives are dominated by integrable Gaussian envelopes). By the Leibniz rule,

$$\frac{\partial P_S}{\partial \tilde{w}^S} = \int_{-\infty}^{-\tilde{w}^D} f(-\tilde{w}^S, v)\, dv + \int_{-\tilde{w}^S}^{\infty} f\left(u, u+\tilde{w}^S-\tilde{w}^D\right) du = m_{S0} + m_{SD},$$

where the first term is the mass released at the science–routine boundary $\{u = -\tilde{w}^S,\ v \le -\tilde{w}^D\}$ and the second the mass at the science–data boundary $\{v = u + \tilde{w}^S - \tilde{w}^D,\ u \ge -\tilde{w}^S\}$; and

$$\frac{\partial P_S}{\partial \tilde{w}^D} = -\int_{-\tilde{w}^S}^{\infty} f\left(u, u+\tilde{w}^S-\tilde{w}^D\right) du = -m_{SD}.$$

The change of variable $u = v + \tilde{w}^D - \tilde{w}^S$ in the analogous computation for $P_D$ gives $\partial P_D/\partial \tilde{w}^S = -m_{SD}$ (confirming the symmetry of $\nabla^2 G$ directly) and $\partial P_D/\partial \tilde{w}^D = m_{D0} + m_{SD}$. Continuity of the masses in $\tilde{w}$ follows from continuity of $f$ and dominated convergence, so $G \in \mathcal{C}^2$. Each mass is strictly positive because $f > 0$ everywhere (nondegenerate bivariate normal, $\nu \in (-1,1)$) and each integral runs over a nondegenerate half-line. The determinant computation is direct: $(m_{S0} + m_{SD})(m_{D0} + m_{SD}) - m_{SD}^2 = m_{S0} m_{D0} + m_{SD}(m_{S0} + m_{D0}) > 0$, and the diagonal entries are positive, so $M > 0$.

**Part (ii):** Write $p = (p_S, p_D)$ for the log wage-premium vector. By Part (i), $G \in \mathcal{C}^2(\mathbb{R}^2)$ and $M(p) = \nabla^2 G(p) > 0$, so $G$ is strictly convex. Full support implies that all three choice probabilities are strictly positive at every finite $p$, hence $\nabla G(p) \in \mathrm{int}(\Delta^2)$.

Fix any $\xi = (s,d) \in \mathrm{int}(\Delta^2)$ and set $r = 1 - s - d > 0$. Consider $F_\xi(p) = G(p) - s p_S - d p_D$. Gaussian integrability makes $G$ finite; the maximum is Lipschitz in $p$, so $F_\xi$ is continuous. Under the maintained mean-zero taste specification, Jensen's inequality gives

$$F_\xi(p) \ge q_\xi(p) \equiv \max\{0, p_S, p_D\} - s p_S - d p_D.$$

The expression subtracted from the maximum is the weighted average of $0, p_S, p_D$ with strictly positive weights $r, s, d$. It equals the maximum only when all three numbers coincide, that is,

when $p = 0$. Thus $q_\xi(p) > 0$ for $p \neq 0$. Since $q_\xi$ is continuous and positively homogeneous, compactness of the unit circle yields

$$a_\xi \equiv \min_{\|p\|_2=1} q_\xi(p) > 0, \qquad F_\xi(p) \geq a_\xi \|p\|_2.$$

Consequently $F_\xi$ is coercive and attains a minimum. Strict convexity makes its minimizer $p(\xi)$ unique, and its first-order condition is $\nabla G(p(\xi)) = \xi$. Every interior share vector therefore has a unique preimage. Together with the range inclusion above, this proves that $\nabla G : \mathbb{R}^2 \to \text{int}(\Delta^2)$ is a bijection.

The inverse function theorem applies at every $p$, since $M(p)$ is nonsingular. The local inverses agree by uniqueness, so $p(\xi)$ is $\mathcal{C}^1$ throughout the interior simplex, with $Dp(\xi) = M(p(\xi))^{-1}$. By the definition of the convex conjugate and the minimizing property just established,

$$G^*(\xi) = \sup_{p\in\mathbb{R}^2} \{\xi \cdot p - G(p)\} = \xi \cdot p(\xi) - G(p(\xi)).$$

Differentiating this identity, the terms involving $Dp(\xi)$ cancel because $\nabla G(p(\xi)) = \xi$. Hence

$$\nabla G^*(\xi) = p(\xi), \qquad \nabla^2 G^*(\xi) = M(p(\xi))^{-1}.$$

In particular $G^* \in \mathcal{C}^2(\text{int}(\Delta^2))$.

Inverting the $2 \times 2$ matrix,

$$M^{-1} = \frac{1}{\det M} \begin{pmatrix} m_{D0} + m_{SD} & m_{SD} \\ m_{SD} & m_{S0} + m_{SD} \end{pmatrix},$$

so $(\nabla^2 G^*)_{SD} = m_{SD}/\det M > 0$ and $(\nabla^2 G^*)_{SS} - (\nabla^2 G^*)_{SD} = m_{D0}/\det M > 0$.

**Part (iii):** From the explicit Hessian of $U$ in Appendix Section A.2, Part (iv), $\partial^2 U/\partial\xi^S\partial\xi^D = 1/\xi^R$, the integral term $h(\xi^D, \Omega)$ contributes only to the $DD$ entry, and $\nabla^2 U \succ 0$ on $\text{int}(\Delta^2)$. Hence $H = \nabla^2 U + \nabla^2 G^*$ is the sum of two positive definite matrices, so $H \succ 0$ and $H_{SD} = 1/\xi^R + (\nabla^2 G^*)_{SD} > 0$,

$$H_{SS} - H_{SD} = \left(\frac{1}{\xi^S} + \frac{1}{\xi^R} + (\nabla^2 G^*)_{SS}\right) - \left(\frac{1}{\xi^R} + (\nabla^2 G^*)_{SD}\right) = \frac{1}{\xi^S} + \frac{m_{D0}}{\det M} > 0.$$

□

With the two lemmas in hand, I now prove Proposition 4. The proof proceeds in two parts, mirroring the statement.

PROOF. **Preliminaries.** The equilibrium is the unique interior minimizer of the potential $J(\xi, \Omega) = U(\xi, \Omega) + G^*(\xi)$ (Propositions 2 and A1). Neither $\tilde{w}^S(\xi)$ nor $\tilde{w}^D(\xi, \Omega)$ depends on $I_t$: the com-

mon productivity factor $S^{\beta} I^{\zeta}$ cancels from every wage ratio, as is visible from the construction of $U$ in Appendix Section A.2, which contains no $I$. Hence $\xi^*$ depends on the state only through $\Omega$. Since $H = \nabla^2_{\xi\xi} J > 0$ everywhere (Lemma A2(iii)) and $\nabla_{\xi} J$ is $\mathcal{C}^1$ in $(\xi, \Omega)$ on $\text{int}(\Delta^2) \times (0, \infty)$, the implicit function theorem applies to $\mathbf{F}(\xi^*, \Omega) \equiv \nabla_{\xi} J(\xi^*, \Omega) = \mathbf{0}$ at every $\Omega > 0$, and uniqueness of the equilibrium makes $\xi^*(\cdot)$ a single-valued $\mathcal{C}^1$ function with

$$\frac{d\xi^*}{d\Omega} = -\big[H(\xi^*, \Omega)\big]^{-1} \frac{\partial \mathbf{F}}{\partial \Omega}. \tag{A2}$$

**Part (i):** In the decomposition of the period-1 equilibrium proof, Part (iv), the term $c\xi^D$ belongs to $U_0$, but $c = c(\Omega) = \alpha - C_D(\Omega)$ and $C_D(\Omega)$ contains $(1-\alpha)\ln\Omega$. To isolate all data-stock dependence, write instead

$$\widehat{U}_0(\xi) = \xi^S \ln \xi^S + \xi^R \ln \xi^R + (1-\alpha)\xi^D \ln \xi^D - \ln(\beta/\psi)\xi^S,$$
$$\widehat{h}(d, \Omega) = h(d, \Omega) + c(\Omega) d, \qquad U(\xi, \Omega) = \widehat{U}_0(\xi) + \widehat{h}(\xi^D, \Omega),$$

where $h$ is the original integral term. Now $\widehat{U}_0$ is independent of $\Omega$, and all $\Omega$-dependence is in $\widehat{h}$, which depends on the allocation only through $\xi^D$. The added linear term has no effect on the Hessian with respect to $\xi$, but its mixed derivative $c'(\Omega) = -(1-\alpha)/\Omega$ must be included. Since $\partial_{\xi^D} U = -\tilde{w}^D$, the gradient shock is unchanged.

Consequently, the shock vector has a zero in the first component,

$$\frac{\partial \mathbf{F}}{\partial \Omega} = \begin{pmatrix} 0 \\ \frac{\partial^2 U}{\partial \xi^D \partial \Omega} \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{\partial \tilde{w}^D}{\partial \Omega} \end{pmatrix},$$

where the premium derivative is taken at fixed $\xi$. Substituting into (A2) and using the explicit inverse of the $2 \times 2$ matrix $H$,

$$\frac{d\xi^{S*}}{d\Omega} = -\frac{H_{SD}}{\det H} \frac{\partial \tilde{w}^D}{\partial \Omega}, \qquad \frac{d\xi^{D*}}{d\Omega} = \frac{H_{SS}}{\det H} \frac{\partial \tilde{w}^D}{\partial \Omega}, \qquad \frac{d\xi^{R*}}{d\Omega} = -\frac{H_{SS} - H_{SD}}{\det H} \frac{\partial \tilde{w}^D}{\partial \Omega},$$

where the third equality uses $\xi^{R*} = 1 - \xi^{S*} - \xi^{D*}$. By Lemma A2(iii), $H_{SS} > 0$, $H_{SD} > 0$, $H_{SS} - H_{SD} > 0$, and $\det H > 0$, so the three equilibrium responses carry the signs stated, all governed by $\partial \tilde{w}^D / \partial \Omega$.

It remains to sign the premium derivative at fixed shares. Because $\tilde{w}^D = \ln w^D - \ln w^R$ and both wages depend on $\Omega$,

$$\frac{\partial \tilde{w}^D}{\partial \ln \Omega} = \varepsilon_{\text{MP}^D, \Omega} - \frac{\partial \ln w^R}{\partial \ln \Omega},$$

where the first term is the elasticity of Proposition 3. For the second, $w^R = \psi S^{\beta} I^{\zeta} R^{\psi - 1} [1 - \tau \sigma^2_{post}]$, so only the efficiency term depends on $\Omega$. Writing $X = X(D, \Omega)$ and using $\sigma^2_{post} = \sigma^2_0 \sigma^2_{\epsilon} / (\sigma^2_{\epsilon} + X)$,

a direct calculation gives

$$\frac{\partial \ln w^R}{\partial \ln \Omega} = \frac{\tau\sigma^2_{post}}{1-\tau\sigma^2_{post}} \cdot \frac{(1-\alpha)X}{\sigma^2_\epsilon + X}.$$

Combining the two terms and applying identity (A1),

$$\frac{\partial \tilde{w}^D}{\partial \ln \Omega} = \frac{(1-\alpha)\left[\sigma^2_\epsilon\left(1-\tau\sigma^2_{post}\right) - X\right]}{\left(\sigma^2_\epsilon + X\right)\left(1-\tau\sigma^2_{post}\right)} = \frac{(1-\alpha)\left[\sigma^4_\epsilon\left(1-\tau\sigma^2_0\right) - X^2\right]}{\left(\sigma^2_\epsilon + X\right)^2\left(1-\tau\sigma^2_{post}\right)}, \tag{A3}$$

evaluated at $X = X^*(\Omega)$. The denominator is positive, so the sign of $\partial\tilde{w}^D/\partial\Omega$ is the sign of $\sigma^4_\epsilon(1-\tau\sigma^2_0) - X^*(\Omega)^2$, which by Lemma A1 is positive exactly on $\mathcal{S}^{deep}$. The ratio statement follows because on $\mathcal{S}^{deep}$ the numerator $\xi^{D*}$ strictly rises and the denominator $\xi^{S*}$ strictly falls (both positive at interior equilibria).

**Part (ii):** Define $\varphi(\Omega) \equiv \sigma^4_\epsilon\left(1-\tau\sigma^2_0\right) - X^*(\Omega)^2$, a $\mathcal{C}^1$ function on $(0,\infty)$ by the preliminaries.

**Step 1: $\varphi > 0$ near zero.** Since $\xi^{D*} \le 1$, $X^*(\Omega) \le L^\alpha \Omega^{1-\alpha}\sigma^2_0 \to 0$ as $\Omega \downarrow 0$, so $\varphi(\Omega) \to \sigma^4_\epsilon(1-\tau\sigma^2_0) > 0$.

**Step 2: every zero of $\varphi$ is a strict downcrossing.** Let $\varphi(\Omega_0) = 0$, so $X^*(\Omega_0) = \sigma^2_\epsilon\sqrt{1-\tau\sigma^2_0} > 0$ and, by Part (i), $d\xi^{D*}/d\Omega = 0$ at $\Omega_0$. Differentiating $X^*(\Omega) = \left(L\xi^{D*}(\Omega)\right)^\alpha \Omega^{1-\alpha}\sigma^2_0$ at $\Omega_0$,

$$\frac{dX^*}{d\Omega}\Big|_{\Omega_0} = \underbrace{\alpha\left(L\xi^{D*}\right)^{\alpha-1}\Omega^{1-\alpha}\sigma^2_0 L\frac{d\xi^{D*}}{d\Omega}}_{=0 \text{ at } \Omega_0} + (1-\alpha)\left(L\xi^{D*}\right)^\alpha \Omega^{-\alpha}\sigma^2_0 > 0,$$

hence $\varphi'(\Omega_0) = -2X^*(\Omega_0)\left(dX^*/d\Omega\right)\big|_{\Omega_0} < 0$.

**Step 3: a zero exists.** Suppose not. Then, by Steps 1 and 2, $\varphi > 0$ on all of $(0,\infty)$, so by Part (i) $\xi^{D*}$ is strictly increasing and, being bounded, satisfies $\xi^{D*}(\Omega) \ge \xi^{D*}(\Omega_a) > 0$ for all $\Omega \ge \Omega_a$ and any fixed $\Omega_a > 0$. Then $X^*(\Omega) \ge \left(L\xi^{D*}(\Omega_a)\right)^\alpha \Omega^{1-\alpha}\sigma^2_0 \to \infty$, so $\varphi(\Omega) \to -\infty$, a contradiction.

**Step 4: the zero is unique.** Suppose that $\Omega_1 < \Omega_2$ are two zeros. By Step 2, choose $a \in (\Omega_1, \Omega_2)$ sufficiently close to $\Omega_1$ that $\varphi(a) < 0$. The set $Z = \{\Omega \in [a, \Omega_2] : \varphi(\Omega) = 0\}$ is nonempty and compact, so it has a smallest element $b > a$. By continuity and the definition of $b$, $\varphi(\Omega) < 0$ for all $\Omega \in [a, b)$. For every sufficiently small $h > 0$,

$$\frac{\varphi(b) - \varphi(b-h)}{h} > 0.$$

Since $\varphi$ is differentiable at $b$, letting $h \downarrow 0$ gives $\varphi'(b) \ge 0$, contradicting Step 2. Thus there is at most one zero. Step 3 gives existence.

Hence there is a unique zero $\bar{\Omega}$, with $\varphi > 0$ on $(0, \bar{\Omega})$ and $\varphi < 0$ on $(\bar{\Omega}, \infty)$. By Lemma A1, the equilibrium lies in $\mathcal{S}^{deep}$ exactly for $\Omega < \bar{\Omega}$.

**Step 5: shape and limits.** By Part (i), $d\xi^{D*}/d\Omega > 0$ on $(0, \bar{\Omega})$, $= 0$ at $\bar{\Omega}$, and $< 0$ on $(\bar{\Omega}, \infty)$,

which gives strict monotonicity on $(0,\bar{\Omega}]$ and $[\bar{\Omega},\infty)$. The science share displays the mirror-image single trough. The limit $\lim_{\Omega\to\infty}\xi^{D*}(\Omega)=0$ follows by contradiction: if $\xi^{D*}(\Omega_k)\geq\varepsilon>0$ along $\Omega_k\to\infty$, then, using the explicit form of the data premium from the proof of Proposition 2, Part (ii),

$$\tilde{w}^D(\xi,\Omega)=C_D(\Omega)+\ln\xi^R+(\alpha-1)\ln\xi^D-2\ln\left(\sigma_\epsilon^2+X\right)-\ln\left(1-\tau\sigma_{post}^2\right),$$

with $C_D(\Omega)$ containing $(1-\alpha)\ln\Omega$ and otherwise constant, I bound each term from above: $\ln\xi^R\leq 0$; $(\alpha-1)\ln\xi^D\leq(1-\alpha)\ln(1/\varepsilon)$; $-\ln(1-\tau\sigma_{post}^2)\leq-\ln(1-\tau\sigma_0^2)$; and $-2\ln(\sigma_\epsilon^2+X)\leq -2\ln X=-2\alpha\ln(L\xi^D)-2(1-\alpha)\ln\Omega-2\ln\sigma_0^2\leq\text{const}-2(1-\alpha)\ln\Omega$. Adding the $(1-\alpha)\ln\Omega$ term of $C_D$, the total is bounded above by $\text{const}-(1-\alpha)\ln\Omega_k\to-\infty$, so $\tilde{w}^D\to-\infty$ along the sequence. But then $\xi^{D*}=P_D(\tilde{w}^{S*},\tilde{w}^{D*})\leq\Pr\left(\eta^D>-\tilde{w}^{D*}\right)\to 0$, contradicting $\xi^{D*}\geq\varepsilon$. The limit $\lim_{\Omega\downarrow 0}\xi^{D*}(\Omega)=0$ follows by the same contradiction argument: with $\xi^D\geq\varepsilon$, $X\to 0$, so $-2\ln(\sigma_\epsilon^2+X)$ stays bounded while $C_D(\Omega)\to-\infty$ through its $(1-\alpha)\ln\Omega$ term, again forcing $\tilde{w}^D\to-\infty$. That $\bar{\Omega}$ is common to periods 1 and 2 follows because the period problem, and hence $\xi^*(\cdot)$ and $\varphi$, depend on the date only through $\Omega$. Finally, the peak lies strictly inside $\mathcal{S}$ since $X^*(\bar{\Omega})=\sigma_\epsilon^2\sqrt{1-\tau\sigma_0^2}<\sigma_\epsilon^2$. □

### A.6. Proof of Corollary 1

PROOF. Since $\omega_0=0$, the data law of motion gives $\omega_1=\gamma Q_0=\gamma S_0^\beta I_0^\zeta$ and $\Omega_1=\Phi\omega_1$, so $\partial\Omega_1/\partial S_0=\Phi\gamma\beta S_0^{\beta-1}I_0^\zeta>0$. The chain rule with Proposition 4 delivers the displayed expression and all signs.[17] □

### A.7. Proofs for Section 4

This section presents four very useful lemmas repeatedly used in the proofs of Section 4. I record them here to save space in the main text.

LEMMA A3 (No Undercutting). *Fix $t$. If $(a,c)$ satisfies $\ln c(\eta)+\eta^{a(\eta)}\geq\bar{u}_t(\eta)$ almost everywhere, then*

$$c(\eta)\geq C_t^{a(\eta)}(\eta)\geq w_t^{a(\eta),m} \tag{A4}$$

*almost everywhere. The first inequality binds exactly when participation binds. The second binds exactly when $a(\eta)$ is market optimal. Consequently,*

$$L\mathbb{E}[c(\eta)]\geq w_t^m\cdot z(a), \tag{A5}$$

[17] Independence of the shares from $I_1$ is the first step of that proposition's proof.

*with equality if and only if participation binds almost everywhere and $a = a_t^m$ almost everywhere. Moreover, $\mathbb{E}[C_t^o(\eta)] < \infty$ for every occupation $o$.*

PROOF. Let $o = a(\eta)$. Participation and equation (8) give $c(\eta) \geq C_t^o(\eta)$. Also,

$$\bar{u}_t(\eta) \geq \ln w_t^{o,m} + \eta^o \quad \Longrightarrow \quad C_t^o(\eta) \geq w_t^{o,m}.$$

The equality conditions follow from the two definitions. Taking expectations gives equation (A5). Every tie set is a line, or is empty at date 0, and therefore has zero probability under the nondegenerate normal distribution. Hence equality requires $a = a_t^m$ almost everywhere. Finally, if $w_t^{\max} = \max_o w_t^{o,m}$, then

$$C_t^o(\eta) \leq w_t^{\max}\left(e^{\eta^S - \eta^o} + e^{\eta^D - \eta^o} + e^{-\eta^o}\right).$$

Each term has a finite expectation because it is the exponential of a normal variable. □

LEMMA A4 (Concavity of the Profit Function). *Fix a state $(I, \Omega)$ with $\Omega \geq 0$, and suppose $\alpha + \beta + \psi \leq 1$. Then*

$$(S, D, R) \longmapsto \pi^E = S^\beta I^\zeta R^\psi \left[1 - \tau\sigma^2_{post}(D, \Omega)\right]$$

*is concave on $\mathbb{R}^3_+$. Consequently, for any wage vector $w$, the profit $\pi^E(z) - w \cdot z$ is concave, and the first-order conditions used to construct the equilibria of Propositions 1-A1 characterize a global maximum: the market input vector maximizes profit at market wages.*

PROOF. *The case* $\Omega = 0$. Since $1 - \alpha > 0$, $D^\alpha \Omega^{1-\alpha} = 0$ for every $D \geq 0$, so $\sigma^2_{post}(D, 0) = \sigma_0^2$ and

$$\pi^E(S, D, R; I, 0) \;=\; \left(1 - \tau\sigma_0^2\right) I^\zeta S^\beta R^\psi,$$

which does not depend on $D$ and is a Cobb–Douglas of degree $\beta + \psi \leq 1$ in $(S, R)$, hence concave in $(S, R)$ and therefore concave in $(S, D, R)$ on $\mathbb{R}^3_+$. The stated consequence follows as in the general case, so the market date-0 input vector is a global profit maximizer at the date-0 wage vector. The remainder of the proof treats $\Omega > 0$.

*The case* $\Omega > 0$. Write $a \equiv \sigma_\epsilon^2$, $c \equiv \Omega^{1-\alpha}\sigma_0^2 > 0$, and $w(D) \equiv cD^\alpha$, so that $\sigma^2_{post} = \sigma_0^2\, a/(a + w(D))$ and

$$1 - \tau\sigma^2_{post}(D, \Omega) = \frac{a(1 - \tau\sigma_0^2) + w(D)}{a + w(D)} = (1 - \tau\sigma_0^2) + \tau\sigma_0^2\, u(D), \qquad u(D) \equiv \frac{w(D)}{a + w(D)} \in [0, 1).$$

Hence

$$\pi^E = I^\zeta (1 - \tau\sigma_0^2)\, S^\beta R^\psi \;+\; I^\zeta \tau\sigma_0^2\, S^\beta R^\psi\, u(D).$$

The first term is a Cobb–Douglas of degree $\beta + \psi \leq 1$, hence concave. For the second, set $k \equiv \beta + \psi$,

$p \equiv 1/(1-k) \geq 1$, and $v(D) \equiv u(D)^{\,p}$, so that

$$S^{\beta} R^{\psi} u(D) = S^{\beta} R^{\psi} v(D)^{1-k},$$

a weighted geometric mean of the nonnegative functions $S$, $R$, and $v(D)$ with weights $(\beta, \psi, 1-k)$ summing to one. Weighted geometric means of concave nonnegative functions are concave, by superadditivity of the geometric mean: for weights $q_i$ summing to one, $\prod_i (x_i + y_i)^{q_i} \geq \prod_i x_i^{q_i} + \prod_i y_i^{q_i}$, which applied to the defining inequality of concavity of each factor delivers concavity of the product. It therefore remains to show that $v$ is concave. Differentiating, with $w'(D) = \alpha w / D$,

$$v'(D) = p\,u^{p-1} u' = p\,u^{p-1} \frac{a \alpha w}{D(a+w)^2} = \frac{p a \alpha c^p D^{\alpha p - 1}}{(a + c D^{\alpha})^{p+1}} > 0, \qquad \frac{d \ln v'(D)}{dD} = \frac{(\alpha p - 1) - (p+1)\,\alpha\, u(D)}{D}.$$

Since $u > 0$ and $\alpha p \leq 1$ (this is exactly $\alpha \leq 1 - \beta - \psi$), both terms in the numerator are nonpositive and the second is strictly negative, so $v'$ is strictly decreasing: $v$ is concave on $(0, \infty)$, and by continuity at $D = 0$, on $[0, \infty)$. The sum of the two concave terms is concave, and adding the linear term $-w \cdot z$ preserves concavity, so any stationary point of profit is a global maximizer. □

LEMMA A5 (Static Efficiency of the Market). *Fix $t$ and the market state. For every assignment $a$,*

$$\pi_t^E(z(a)) - L\mathbb{E}\Big[C_t^{a(\eta)}(\eta)\Big] \leq v_t^m,$$

*with equality if and only if $a = a_t^m$ almost everywhere.*

PROOF. Lemma A3 and profit concavity give

$$\begin{aligned} \pi_t^E(z(a)) - L\mathbb{E}[\,C_t^{a(\eta)}(\eta)\,] &\leq \pi_t^E(z(a)) - w_t^m \cdot z(a) \\ &\leq \pi_t^E(z_t^m) - w_t^m \cdot z_t^m = v_t^m. \end{aligned}$$

The market assignment attains both inequalities. Conversely, equality in the first requires $a = a_t^m$ almost everywhere by Lemma A3. This also proves uniqueness. At date 0, every positive data assignment makes the first inequality strict because $w_0^{D,m} = 0 < C_0^D(\eta)$. □

LEMMA A6 (Dynamic Value of a Scientist). *Evaluate all objects along the market path and write*

$$\partial_I \pi_s^E \equiv \zeta S_s^{\beta} I_s^{\zeta - 1} R_s^{\psi} \big[1 - \tau \sigma^2_{post,s}\big], \qquad \partial_{\Omega} \pi_s^E \equiv S_s^{\beta} I_s^{\zeta} R_s^{\psi}\, \tau \frac{(1-\alpha)\,\sigma_0^4 \sigma_\epsilon^2 D_s^{\alpha} \Omega_s^{-\alpha}}{\big(\sigma_\epsilon^2 + D_s^{\alpha} \Omega_s^{1-\alpha} \sigma_0^2\big)^2},$$

*with the convention that $\partial_{\Omega} \pi_s^E = 0$ when $D_s = 0$. For $s \in \{1,2,3\}$ define the fixed-shares marginal values of the stocks,*

$$\hat{\lambda}_s^I \equiv \frac{\partial}{\partial I_s} \sum_{u=s}^{2} \rho^{\,u-s} \pi_u^E, \qquad \hat{\lambda}_s^{\omega} \equiv \frac{\partial}{\partial \omega_s} \sum_{u=s}^{2} \rho^{\,u-s} \pi_u^E,$$

*where the derivatives are taken along fixed-shares continuations and* $\hat{\lambda}_3^I = \hat{\lambda}_3^\omega = 0$. *Then,*

*(i) The marginal values obey the backward recursions*

$$\hat{\lambda}_s^I = \partial_I \pi_s^E + \rho\,(1-\delta_I)\,\hat{\lambda}_{s+1}^I + \rho\gamma\zeta\, S_s^\beta\, I_s^{\zeta-1}\,\hat{\lambda}_{s+1}^\omega,$$
$$\hat{\lambda}_s^\omega = \Phi\,\partial_\Omega \pi_s^E + \rho\,(1-\delta_\omega)\,\hat{\lambda}_{s+1}^\omega.$$

*(ii) Define* $\Lambda_t^I \equiv \rho\,\hat{\lambda}_{t+1}^I$ *and* $\Lambda_t^\omega \equiv \rho\,\hat{\lambda}_{t+1}^\omega$ *for* $t \in \{0,1,2\}$. *Then*

$$\Lambda_2^I = \Lambda_2^\omega = 0, \qquad \Lambda_1^I = \rho\,\partial_I \pi_2^E > 0, \qquad \Lambda_1^\omega = \rho\,\Phi\,\partial_\Omega \pi_2^E \geq 0,$$

*and*

$$\Lambda_0^I = \rho\,\partial_I \pi_1^E + \rho\,(1-\delta_I)\,\Lambda_1^I + \rho\gamma\zeta\, S_1^\beta\, I_1^{\zeta-1}\,\Lambda_1^\omega > 0, \qquad \Lambda_0^\omega = \rho\,\Phi\,\partial_\Omega \pi_1^E + \rho\,(1-\delta_\omega)\,\Lambda_1^\omega \geq 0.$$

*If* $\tau > 0$, *then* $\Lambda_0^\omega > 0$ *and* $\Lambda_1^\omega > 0$ *along the market path.*

*(iii) The date-t dynamic wedge on scientists,*

$$\kappa_t \equiv \chi\,\Lambda_t^I + \gamma\beta\, S_t^{\beta-1} I_t^\zeta\,\Lambda_t^\omega,$$

*equals the marginal effect of one additional date-t scientist on discounted output from* $t+1$ *onward under fixed-shares continuations. It satisfies* $\kappa_0 > 0$, $\kappa_1 > 0$, *and* $\kappa_2 = 0$. *The analogous wedges for data engineers and routine workers are identically zero at every date.*

PROOF. **Part (i):** Fix the shares $\{\xi_u\}_{u \geq s}$ at their market values, so that $(S_u, D_u, R_u)$ are constants, and let the stocks evolve according to $I_{u+1} = (1-\delta_I) I_u + \chi S_u$ and $\omega_{u+1} = (1-\delta_\omega)\omega_u + \gamma S_u^\beta I_u^\zeta$. Define $W_s(I_s, \omega_s) \equiv \sum_{u=s}^{2} \rho^{\,u-s} \pi_u^E$ along the induced path, with $W_3 \equiv 0$, so that

$$W_s(I_s, \omega_s) = \pi_s^E(I_s, \omega_s) + \rho\, W_{s+1}\big(I_{s+1}(I_s),\, \omega_{s+1}(I_s, \omega_s)\big).$$

With shares fixed, the transition maps have partial derivatives

$$\frac{\partial I_{s+1}}{\partial I_s} = 1-\delta_I, \qquad \frac{\partial \omega_{s+1}}{\partial I_s} = \gamma\zeta S_s^\beta I_s^{\zeta-1}, \qquad \frac{\partial \omega_{s+1}}{\partial \omega_s} = 1-\delta_\omega, \qquad \frac{\partial I_{s+1}}{\partial \omega_s} = 0.$$

Since $\pi_s^E$ depends on $\omega_s$ only through $\Omega_s = \Phi\omega_s$, we have $\partial \pi_s^E / \partial \omega_s = \Phi\,\partial_\Omega \pi_s^E$, where $\partial_\Omega \pi_s^E$ is as displayed (a direct computation from $\sigma^2_{post} = \sigma_0^2\sigma_\epsilon^2/(\sigma_\epsilon^2 + D_s^\alpha \Omega_s^{1-\alpha}\sigma_0^2)$). Differentiating the recursion for $W_s$ and writing $\hat{\lambda}_s^I = \partial W_s/\partial I_s$, $\hat{\lambda}_s^\omega = \partial W_s/\partial \omega_s$ delivers the two displayed recursions. Differentiability at every step is guaranteed because, along the market path, $I_s > 0$ for all $s$ and $\omega_s > 0$ for $s \geq 1$ (indeed $\omega_1 = \gamma S_0^\beta I_0^\zeta > 0$ by Proposition 1), and the continuation dates have positive data stocks and interior allocations by Propositions 2 and A1, so every composed map is smooth in a neighborhood of the path.

**Part (ii):** The terminal values follow from $\hat{\lambda}_3^I = \hat{\lambda}_3^\omega = 0$: $\Lambda_2^I = \rho\hat{\lambda}_3^I = 0$ and likewise $\Lambda_2^\omega = 0$.

Applying the recursion at $s=2$ gives $\hat{\lambda}_2^I = \partial_I \pi_2^E$ and $\hat{\lambda}_2^\omega = \Phi\,\partial_\Omega \pi_2^E$, hence the displayed $\Lambda_1^I$ and $\Lambda_1^\omega$. Applying the recursion at $s=1$ and multiplying by $\rho$ yields the displayed $\Lambda_0^I$ and $\Lambda_0^\omega$. Relative to the one-period-ahead formulas, $\Lambda_0^I$ contains the two continuation terms $\rho(1-\delta_I)\Lambda_1^I$ and $\rho\gamma\zeta S_1^\beta I_1^{\zeta-1}\Lambda_1^\omega$, which are the effect of $I_1$ on period-2 output through the ideas stock and through period-1 data creation, respectively.

Let's look at signs now. Along the market path, $S_s>0$, $R_s>0$, $I_s>0$, and $1-\tau\sigma^2_{post,s}>0$, so $\partial_I \pi_s^E>0$ for $s\in\{1,2\}$. Together with $\delta_I,\delta_\omega\in(0,1)$ and $\hat{\lambda}^\omega\geq 0$ (each term is nonnegative), this gives $\Lambda_1^I>0$ and $\Lambda_0^I\geq\rho\,\partial_I\pi_1^E>0$. For the data multipliers, $\partial_\Omega\pi_s^E\geq 0$ always. Moreover $\Omega_1>0$ (since $\omega_1>0$) and $\Omega_2>0$, so $D_1>0$ and $D_2>0$ by Propositions 2 and A1, and if $\tau>0$ then $\partial_\Omega\pi_1^E>0$ and $\partial_\Omega\pi_2^E>0$, whence $\Lambda_0^\omega,\Lambda_1^\omega>0$.

**Part (iii):** A marginal increase in $S_t$, holding all later shares fixed, moves next-period stocks by $\partial I_{t+1}/\partial S_t=\chi$ and $\partial\omega_{t+1}/\partial S_t=\gamma\,\partial Q_t/\partial S_t=\gamma\beta S_t^{\beta-1}I_t^\zeta$. It has no other effect on dates $u\geq t+1$. Its effect on $\sum_{u=t+1}^2\rho^{u-t}\pi_u^E$ is therefore $\chi\rho\hat{\lambda}_{t+1}^I+\gamma\beta S_t^{\beta-1}I_t^\zeta\rho\hat{\lambda}_{t+1}^\omega=\kappa_t$. Since neither $D_t$ nor $R_t$ enters the transition equations ($Q_t=S_t^\beta I_t^\zeta$ does not depend on them), the corresponding wedges are identically zero. Finally, $\kappa_2=0$ because $\Lambda_2^I=\Lambda_2^\omega=0$, while for $t\in\{0,1\}$, $\kappa_t\geq\chi\Lambda_t^I>0$. □

### A.8. Proof of Proposition 5

PROOF. Part (i) is Lemma A5. Competitive firms pay current marginal products. Lemma A6 gives the omitted dynamic contribution and its signs, which proves (ii).

For (iii), separate the date-$t$ derivative of (10) into its current and continuation components. The normal density is smooth and Lemma A3 gives an integrable envelope, so differentiation across the moving boundaries is valid. Consider one boundary between occupations $o$ and $o'$. Its contribution to the current derivative is its signed probability flux times

$$L\Big[\mathrm{MP}_{o,t}-\mathrm{MP}_{o',t}-\big(C_t^o-C_t^{o'}\big)\Big].$$

On that boundary, market indifference implies $C_t^o=w_t^{o,m}$ and $C_t^{o'}=w_t^{o',m}$. Firm optimality implies $\mathrm{MP}_{o,t}=w_t^{o,m}$ and $\mathrm{MP}_{o',t}=w_t^{o',m}$. Thus, every boundary contributes zero. Summing over boundaries, the current derivative vanishes. At later dates, assignments and compensation bills are fixed. Only the states move. Part (iii) of Lemma A6 therefore gives

$$\left.\frac{d\mathcal{W}_t}{d\varepsilon}\right|_{\varepsilon=0}=\rho^t\kappa_t L\left.\frac{d\xi_t^S}{d\varepsilon}\right|_{\varepsilon=0}.$$

For (iv), define

$$a_t^\varepsilon(\eta)\in\arg\max_{o\in\{S,D,R\}}\left\{\ln w_t^{o,m}+\eta^o+\varepsilon\mathbf{1}\{o=S\}\right\},$$

using $\ln w_0^{D,m}=-\infty$ at date 0. This smoothly shifts every boundary adjacent to science. Nonde-

generacy of the normal distribution implies

$$\frac{d\xi_t^S(\varepsilon)}{d\varepsilon}\Big|_{\varepsilon=0} > 0.$$

For $t \in \{0,1\}$, part (iii) and $\kappa_t > 0$ make the welfare derivative strictly positive. Differentiability then gives $\bar{\varepsilon} > 0$ such that $\mathcal{W}_t(\varepsilon) > \mathcal{W}_t(0)$ for every $\varepsilon \in (0, \bar{\varepsilon})$. Workers consume their compensation requirements and retain their market utilities. Owner consumption remains feasible for small $\varepsilon$ by continuity and its strict market slack; discounted owner consumption rises. The perturbation is therefore a local Pareto improvement. At $t = 2$, $\kappa_2 = 0$, so every boundary perturbation has zero first-order effect. □

### A.9. Proof of Proposition 6

PROOF. **Part (i):** Fix the inherited firm-level stock $\omega_t > 0$ and let $N > 1$. Then $\Omega_t = [1 + \phi(N-1)]\,\omega_t$ is strictly increasing in $\phi$, and Proposition 4(i) gives $d\xi^{D*}/d\Omega > 0$ in the deep-scarcity region, so

$$\frac{\partial \xi_t^{D*}}{\partial \phi} = \frac{d\xi^{D*}}{d\Omega}(N-1)\,\omega_t > 0.$$

Under a permanent policy, this comparative static is exact at $t = 1$: the date-0 problem does not involve $\phi$ (since $\Omega_0 = 0$), so $\omega_1 = \gamma S_0^{\beta} I_0^{\zeta}$ is independent of $\phi$, $d\Omega_1/d\phi = (N-1)\,\omega_1 > 0$, and $d\xi_1^D/d\phi > 0$ with no inherited-stock feedback to offset it, though still conditional on deep scarcity to sign $d\xi^{D*}/d\Omega$. At $t = 2$, by contrast, the inherited stock itself responds to the policy: $\omega_2 = (1-\delta_\omega)\omega_1 + \gamma S_1^{\beta} I_1^{\zeta}$ with $S_1 = S_1(\phi)$, so

$$\frac{d\Omega_2}{d\phi} = (N-1)\,\omega_2 + \Phi\,\gamma\beta\, S_1^{\beta-1} I_1^{\zeta} \frac{dS_1^m}{d\phi},$$

where the first term is the direct sharing gain and the second is strictly negative in deep scarcity, since $dS_1^m/d\phi < 0$ by the calculation in Part (iii) below.

Nevertheless, the total derivative is strictly positive: the equilibrium response of science bounds the inherited-stock feedback. To establish this, evaluate the equilibrium Hessian $H = \nabla^2_{\xi,\xi} J$ at date 1, and write $s = \xi_1^{S*}$, $d = \xi_1^{D*}$, and $r = \xi_1^{R*}$. By the occupational-supply lemma and the firm-side potential,

$$\begin{aligned}
A &\equiv H_{SS} - H_{SD} = \frac{1}{s} + \frac{m_{D0}}{\det M} > \frac{1}{s},\\
B &\equiv H_{DD} - H_{SD} = \frac{1-\alpha}{d} + h''(d) + \frac{m_{S0}}{\det M} > 0,\\
C &\equiv H_{SD} = \frac{1}{r} + \frac{m_{SD}}{\det M} > 0,
\end{aligned}$$

where $M = \nabla^2 G$ and $h''$ denotes the second derivative of the integral term in the firm-side

potential, holding the data stock fixed. Hence

$$H = \begin{pmatrix} A+C & C \\ C & B+C \end{pmatrix}, \qquad \det H = AB + C(A+B) > AC, \qquad 0 < \frac{C}{s \det H} < \frac{1}{sA} < 1.$$

Let $k \equiv \tau\sigma_0^2 \in (0,1)$ and $x \equiv X(Ld, \Omega_1)/\sigma_\epsilon^2 > 0$. The direct premium response from the proof of Proposition 4 is

$$b_1 \equiv \left. \frac{\partial \tilde{w}^D}{\partial \ln \Omega} \right|_{\xi = \xi_1^*} = (1-\alpha) \frac{1-k-x^2}{(1+x)(1+x-k)}.$$

If the date-1 equilibrium is in deep scarcity, $0 < b_1 < 1-\alpha$. The upper bound follows because the denominator of the fraction minus its numerator equals $(2-k)x + 2x^2 > 0$. The equilibrium derivative in Proposition 4 therefore implies

$$e_1 \equiv -\frac{d \ln S_1^m}{d \ln \Omega_1} = b_1 \frac{C}{s \det H}, \qquad 0 < e_1 < 1-\alpha < 1.$$

Since $I_1$ and $\omega_1$ are independent of $\phi$, and $\Omega_1 = \Phi \omega_1$, it follows that $d \ln S_1^m / d \ln \Phi = -e_1$. Define the share of the date-2 firm-level data stock generated by date-1 quality as

$$\vartheta_2 \equiv \frac{\gamma S_1^\beta I_1^\zeta}{(1-\delta_\omega)\omega_1 + \gamma S_1^\beta I_1^\zeta} \in (0,1).$$

Differentiating $\Omega_2 = \Phi \omega_2$ gives

$$\frac{d \ln \Omega_2}{d \ln \Phi} = 1 - \beta \vartheta_2 e_1 > 1 - \beta(1-\alpha) > 0.$$

If the date-1 equilibrium is on or beyond the deep-scarcity boundary, Proposition 4 instead gives $dS_1^m/d\phi \geq 0$, so the displayed expression for $d\Omega_2/d\phi$ is again strictly positive. Thus, $d\Omega_2/d\phi > 0$ at every policy value when $N > 1$.

Finally, equilibrium shares are independent of $I_2$, so

$$\frac{d\xi_2^{D,m}}{d\phi} = \frac{d\xi^{D*}}{d\Omega}(\Omega_2) \frac{d\Omega_2}{d\phi}.$$

The date-2 data share therefore rises in deep scarcity, falls outside its closure, and has zero derivative on its boundary. If only the date-1 region is specified, the date-2 employment response remains unsigned because its own region is unspecified, not because $d\Omega_2/d\phi$ is ambiguous.

**Part (ii):** Fix a feasible assignment path $\{a_t\}$ and hold it fixed while $\phi$ varies. Two observations. First, the stocks the path generates do not depend on $\phi$: the transitions are $I_{t+1} = (1-\delta_I)I_t + \chi S_t$ and $\omega_{t+1} = (1-\delta_\omega)\omega_t + \gamma Q_t$ with $Q_t = S_t^\beta I_t^\zeta$, and $\Omega_t$ enters neither. Second, the compensation bill $L\mathbb{E}[C_t^{a_t(\eta),1}(\eta)]$ does not depend on $\phi$ either, by construction of the anchor:

$C_t^{o,1}$ is built from the $\phi = 1$ market wages, which are fixed data. Only the output component of the date-$t$ surplus $\mathcal{G}_t = \pi_t^E - L\mathbb{E}[C_t^{a_t(\eta),1}(\eta)]$ reacts, and it reacts through $\Omega_t = \Phi\omega_t$ alone:

$$\pi_t^E = S_t^{\beta} I_t^{\zeta} R_t^{\psi}[1 - \tau\sigma^2_{post}(D_t, \Omega_t)], \qquad \frac{\partial \pi_t^E}{\partial \phi} = -S_t^{\beta} I_t^{\zeta} R_t^{\psi} \tau \frac{\partial \sigma^2_{post}}{\partial \phi}.$$

Fix $(S_t, D_t, R_t)$ and $(I_t, \omega_t)$. If $D_t > 0$ and $\omega_t > 0$, then

$$\sigma^2_{post}(D_t, \Omega_t) = \frac{\sigma_0^2 \sigma_\epsilon^2}{\sigma_\epsilon^2 + D_t^{\alpha} \Omega_t^{1-\alpha} \sigma_0^2}.$$

and

$$\frac{\partial \sigma^2_{post}}{\partial \Omega_t} = -\frac{\sigma_0^4 \sigma_\epsilon^2 (1-\alpha) D_t^{\alpha} \Omega_t^{-\alpha}}{(\sigma_\epsilon^2 + D_t^{\alpha} \Omega_t^{1-\alpha} \sigma_0^2)^2} < 0$$

Since $\partial\Omega_t/\partial\phi = (N-1)\omega_t > 0$, output is strictly increasing in $\phi$ when $N > 1$ and $\tau > 0$. If $D_t = 0$ or $\omega_t = 0$, output is independent of $\phi$ at that date. Thus, $\partial\pi_t^E/\partial\phi \geq 0$ everywhere.

Thus, $\mathcal{W}^A(a, \phi)$ is nondecreasing in $\phi$ for every fixed $a$. Maximizing over assignments preserves the ordering, so full sharing is optimal under direct allocation. This argument compares paths before re-optimization and requires no derivative of the optimized value.

**Part (iii):** In the second-best setting, the regulator chooses $\phi$, but assignments are determined by the market equilibrium at that $\phi$. Write the criterion at the full-sharing anchor as

$$\mathcal{W}^{SB}(\phi) \equiv \sum_{t=0}^{2} \rho^t \Big[ \pi_t^E \big( z(a_t^m(\cdot;\phi)); \phi \big) - L\mathbb{E}\big[ C_t^{a_t^m(\eta;\phi),1}(\eta) \big] \Big], \tag{A6}$$

with $a_t^m(\cdot;\phi)$ the market assignment at $\phi$ and $C_t^{o,1}$ the anchored requirements of equation (12).

At each $\phi$, the schedule $C_t^{a_t^m(\eta;\phi),1}(\eta)$ restores every worker to her full-sharing utility. Equation (A6) is output net of this compensation. At $\phi = 1$, compensation equals the market wage schedule by Lemma A3, so $\mathcal{W}^{SB}(1) = \sum_t \rho^t v_t^m$.

Differentiate equation (A6) at $\phi = 1$ from the left. The induced assignment change displaces market indifference boundaries. The static assignment derivative vanishes by the boundary argument in part (iii) of Proposition 5. Only the direct informational effect and the dynamic effect through science remain:

$$\frac{d\mathcal{W}^{SB}}{d\phi}\Big|_{\phi=1^-} = \underbrace{\sum_{t=0}^{2} \rho^t \frac{\partial \pi_t^E}{\partial \phi}}_{\text{Direct Effect}} + \underbrace{\sum_{t=0}^{2} \rho^t \kappa_t \frac{dS_t^m}{d\phi}}_{\text{Dynamic Terms}},$$

with $\kappa_t$ the dynamic wedge of Lemma A6 evaluated along the $\phi = 1$ market path, so that $\kappa_2 = 0$ and no continuation value is assigned to post-terminal stocks. The two components must be read as stated to avoid double counting: $\partial\pi_t^E/\partial\phi$ is the direct informational effect, taken holding

occupational inputs *and* the inherited stock $\omega_t$ fixed, so that it captures only the sharing channel $\Omega_t = \Phi\omega_t$. The propagation of the date-1 reallocation into $I_2$ and $\omega_2$ is already contained in $\kappa_1\, dS_1^m/d\phi$. In particular, this partial derivative is not the total derivative $d\Omega_2/d\phi$ computed in Part (i). From Part (ii), the direct effect is strictly positive for $t \in \{1,2\}$ and zero at $t = 0$ (where $\omega_0 = \Omega_0 = 0$). Moreover $dS_0^m/d\phi = 0$, since the date-0 problem does not involve $\phi$, and the $t = 2$ dynamic term vanishes with $\kappa_2 = 0$. The trade-off therefore reduces to the period-1 reallocation against the discounted direct gains. To sign it, I must sign $\frac{dS_1^m}{d\phi}$.

The market equilibrium shares $\xi^* = (\xi^{S*}, \xi^{D*})$ solve $\nabla_\xi J(\xi^*, \Omega) = 0$. By the IFT, the sensitivity of shares to the data stock $\Omega$ is

$$\frac{d\xi^*}{d\Omega} = -\big[H(\xi^*, \Omega)\big]^{-1} \frac{\partial(\nabla J)}{\partial\Omega}.$$

As shown in the proof of Proposition 4, $\Omega$ enters only through the separable term $\widehat{h}(\xi^D, \Omega) = h(\xi^D, \Omega) + c(\Omega)\xi^D$, so the gradient shock vector is $(0, -\partial\tilde{w}^D/\partial\Omega)'$. Using the formula for the inverse of a $2\times 2$ matrix, the response of the scientist share is

$$\frac{d\xi^{S*}}{d\Omega} = (H^{-1})_{12} \frac{\partial\tilde{w}^D}{\partial\Omega} = -\frac{H_{SD}}{\det(H)} \frac{\partial\tilde{w}^D}{\partial\Omega}.$$

In the deep scarcity region, $\frac{\partial\tilde{w}^D}{\partial\Omega} > 0$. By Lemma A2(iii), the off-diagonal entry of the Hessian is strictly positive at every interior allocation and every state,

$$H_{SD} = \underbrace{\frac{\partial^2 U}{\partial\xi^S \partial\xi^D}}_{\text{Firm Side}} + \underbrace{\frac{\partial^2 G^*}{\partial\xi^S \partial\xi^D}}_{\text{Supply Side}} = \frac{1}{\xi^R} + \frac{m_{SD}}{\det M} > 0:$$

the firm-side complementarity through the shared routine pool and the supply-side science-data margin push in the same direction, for any $\nu \in (-1, 1)$. Hence $\frac{d\xi^{S*}}{d\Omega} < 0$, with no auxiliary assumption.

For concreteness, focus on $t = 1$. For $t = 1$, the firm-level data stock $\omega_1$ is predetermined by choices in $t = 0$ and does not depend on $\phi$. Thus, the aggregate data stock $\Omega_1 = (1 + \phi(N-1))\omega_1$ satisfies

$$\frac{d\Omega_1}{d\phi} = \frac{\partial\Omega_1}{\partial\phi} = (N-1)\omega_1 > 0.$$

Using the chain rule, the response of the scientist stock in period 1 is unambiguously negative

$$\frac{dS_1^m}{d\phi} = L \frac{d\xi^{S*}}{d\Omega_1} \frac{d\Omega_1}{d\phi} < 0.$$

Higher data sharing raises the return to data engineering relative to science, drawing talent away from $S_1$.

Since $\kappa_1 > 0$ (Lemma A6) and $\frac{dS_1^m}{d\phi} < 0$, the dynamic term at $t = 1$ is strictly negative. Hence

$$\frac{d\mathcal{W}^{SB}}{d\phi}\Big|_{\phi=1^-} < 0 \qquad \Longleftrightarrow \qquad \rho\,\kappa_1 \left|\frac{dS_1^m}{d\phi}\right| > \rho\,\frac{\partial \pi_1^E}{\partial \phi} + \rho^2\,\frac{\partial \pi_2^E}{\partial \phi}:$$

a marginal restriction of data sharing below $\phi = 1$ strictly raises compensated surplus exactly when the dynamic reallocation loss outweighs the discounted direct informational gains. This inequality is the local condition, and I state it rather than a sufficient condition on primitives: $\kappa_1$ is itself an equilibrium object, and the primitives that raise it also move $dS_1^m/d\phi$ and the direct gains, both of which are of the same order in $\tau$.

It remains to establish existence. The market state, shares, wages, and output are continuous in $\phi$ by Propositions 2 and 4. Off the null tie set, the market assignment is also continuous. Moreover, $\max_o C_t^{o,1}(\eta)$ is integrable because it is bounded by a finite sum of exponentials of normal variables. Dominated convergence therefore makes the compensation bill continuous. Hence $\mathcal{W}^{SB}$ is continuous on $[0,1]$ and attains a maximum. The displayed inequality rules out $\phi = 1$, so every maximizer satisfies $\phi^{SB} < 1$. □

### A.10. Proof of Proposition 7

PROOF. With automation, the firm's expected profit at time $s$ is

$$\pi_s^E = S_s^{\beta}\, I_s^{\zeta}\big[1 - \tau\sigma^2_{post}(D_s, \Omega_s)\big]\big[R_s + M(\Omega_s)\big]^{\psi},$$

where $M(\Omega_s)$ is the data-enabled routine capacity available to all firms, with $M'(\Omega) > 0$ and $M(0) = 0$; the baseline economy sets $M \equiv 0$. Since all firms have access to this technology given $\Omega_s$, there is no adoption decision or cost.

**Step 1: the state paths coincide.** Fix the date-$t$ state $(I_t, \omega_t)$ and a common interior allocation path $\{\xi_s\}_{s\geq t}$ for the two economies. The transitions are $I_{s+1} = (1-\delta_I)I_s + \chi S_s$ and $\omega_{s+1} = (1-\delta_\omega)\omega_s + \gamma S_s^{\beta} I_s^{\zeta}$: machine capacity enters neither law of motion, because data is generated by quality, not output, in the baseline specification. Hence the sequences $\{(I_s, \omega_s)\}_{s\geq t}$ are identical across the two economies, and $M(\Omega_s) > 0$ for all $s \geq 1$ along the market-path states, since $\omega_s > 0$ there.

**Step 2: both multipliers rise.** Apply the fixed-shares construction of Lemma A6 to each economy. The per-period marginal values are

$$\partial_I \pi_s^E = \zeta S_s^{\beta} I_s^{\zeta-1}\big[R_s + M(\Omega_s)\big]^{\psi}\big[1 - \tau\sigma^2_{post,s}\big],$$

$$\partial_\Omega \pi_s^E = \underbrace{S_s^{\beta} I_s^{\zeta}\big[R_s + M(\Omega_s)\big]^{\psi}\left(-\tau\frac{\partial \sigma^2_{post,s}}{\partial \Omega_s}\right)}_{\text{variance reduction channel}} + \underbrace{S_s^{\beta} I_s^{\zeta}\big[1 - \tau\sigma^2_{post,s}\big]\psi\big[R_s + M(\Omega_s)\big]^{\psi-1} M'(\Omega_s)}_{\text{automation channel}},$$

where

$$\frac{\partial \sigma^2_{post}}{\partial \Omega} = -\frac{\sigma_0^4 \sigma_\epsilon^2 (1-\alpha) D^\alpha \Omega^{-\alpha}}{\left[\sigma_\epsilon^2 + D^\alpha \Omega^{1-\alpha} \sigma_0^2\right]^2} < 0,$$

so both channels are nonnegative. Since the two economies share the same states and allocations (Step 1), each of these objects can be compared term by term at every date $s \geq t+1$: whenever $M(\Omega_s) > 0$,

$$\partial_I \pi_s^{E,auto} > \partial_I \pi_s^{E,base} \qquad \text{and} \qquad \partial_\Omega \pi_s^{E,auto} > \partial_\Omega \pi_s^{E,base},$$

the first because $[R_s + M_s]^\psi > R_s^\psi$ for $\psi \in (0,1)$, the second because the variance channel inherits the same factor and the automation channel is strictly positive. The recursions of Lemma A6(i),

$$\hat{\lambda}_s^I = \partial_I \pi_s^E + \rho(1-\delta_I)\hat{\lambda}_{s+1}^I + \rho\gamma\zeta S_s^\beta I_s^{\zeta-1}\hat{\lambda}_{s+1}^\omega, \qquad \hat{\lambda}_s^\omega = \Phi\, \partial_\Omega \pi_s^E + \rho(1-\delta_\omega)\hat{\lambda}_{s+1}^\omega,$$

have identical, nonnegative coefficients in the two economies and strictly larger per-period inputs under automation. Iterating backward from $\hat{\lambda}_3 = 0$, every $\hat{\lambda}_s^{auto}$ strictly exceeds $\hat{\lambda}_s^{base}$ for $s \leq 2$. Hence, for $t \in \{0,1\}$,

$$\Lambda_t^{I,auto} > \Lambda_t^{I,base} \qquad \text{and} \qquad \Lambda_t^{\omega,auto} > \Lambda_t^{\omega,base}.$$

**Step 3: the wedge widens and remains unpriced.** Because the data law of motion is quality-based, the marginal data byproduct of a scientist, $\partial \omega_{t+1}/\partial S_t = \gamma\beta S_t^{\beta-1} I_t^\zeta$, is the same in both economies, so by Lemma A6(iii),

$$\kappa_t^{auto} = \chi \Lambda_t^{I,auto} + \gamma\beta S_t^{\beta-1} I_t^\zeta \Lambda_t^{\omega,auto} > \chi \Lambda_t^{I,base} + \gamma\beta S_t^{\beta-1} I_t^\zeta \Lambda_t^{\omega,base} = \kappa_t^{base} > 0$$

for $t \in \{0,1\}$, where positivity of the baseline wedge is Lemma A6(ii)-(iii). In the competitive equilibrium of either economy, wages equal static marginal products - under automation, scientists are paid $w_t^S = \beta S_t^{\beta-1} I_t^\zeta [1 - \tau\sigma^2_{post,t}][R_t + M_t]^\psi$ - so the dynamic contribution $\kappa_t$ is unpriced in both, and the unpriced part is strictly larger under automation. □

### A.11. Proof of Corollary 2

PROOF. **Step 1: the potential extends to the automated economy.** At the period-$t$ problem with state $(I_t, \Omega_t)$, write $\mu \equiv M(\Omega_t)/L$ for machine capacity per worker, so effective routine capacity is $R_t + M(\Omega_t) = L(\xi^R + \mu)$. The firm's first-order conditions give

$$\tilde{w}^S(\xi, \Omega) = \ln\frac{\beta}{\psi} + \ln\left(\xi^R + \mu(\Omega)\right) - \ln \xi^S,$$

$$\tilde{w}^D(\xi, \Omega) = C_D(\Omega) + \ln\left(\xi^R + \mu(\Omega)\right) + (\alpha - 1)\ln \xi^D - 2\ln\left(\sigma_\epsilon^2 + X\right) - \ln\left(1 - \tau\sigma^2_{post}\right),$$

with $X = (L\xi^D)^\alpha \Omega^{1-\alpha}\sigma_0^2$: relative to the baseline expressions in Appendix Section A.2, $\xi^R$ is everywhere replaced by $\xi^R + \mu$, and, exactly as in the baseline, the common production factor $S^\beta I^\zeta$ cancels from both premia while the efficiency term $[1-\tau\sigma^2_{post}]$ cancels from the science premium only: the data premium retains the marginal-information term $-\ln(1-\tau\sigma^2_{post})$, as displayed. The potential construction of Proposition 2 extends *mutatis mutandis* with $(\xi^R+\mu)\ln(\xi^R+\mu)$ in place of $\xi^R \ln \xi^R$ in $U$. Its Hessian replaces $1/\xi^R$ by $1/(\xi^R+\mu)$ and remains positive definite, and Lemma A2 applies unchanged to $G^*$, so at the interior equilibrium assumed in the statement,

$$H > 0, \qquad H_{SD} = \frac{1}{\xi^R+\mu} + \frac{m_{SD}}{\det M} > 0, \qquad H_{SS} - H_{SD} = \frac{1}{\xi^S} + \frac{m_{D0}}{\det M} > 0.$$

**Step 2: the two channels in the premia.** Differentiating the premia at fixed shares,

$$\frac{\partial \tilde{w}^S}{\partial \Omega}\Big|_\xi = \underbrace{\frac{M'(\Omega)/L}{\xi^R+\mu}}_{\equiv a(\Omega)>0}, \qquad \frac{\partial \tilde{w}^D}{\partial \Omega}\Big|_\xi = a(\Omega) + \underbrace{\frac{1}{\Omega}\cdot\frac{(1-\alpha)\left[\sigma_\epsilon^4(1-\tau\sigma_0^2)-X^2\right]}{(\sigma_\epsilon^2+X)^2(1-\tau\sigma^2_{post})}}_{\equiv b(\Omega)},$$

where the second term is the information channel of equation (A3) and $b(\Omega) > 0$ in the deep-scarcity region by Lemma A1. This proves part (i): the automation channel $a$ raises *both* premia, routine work becomes relatively less attractive, and the information channel $b$ additionally raises the data premium.

**Step 3: equilibrium responses.** The gradient shock vector is $\partial(\nabla_\xi J)/\partial\Omega = -(a,\ a+b)'$, so the implicit function theorem gives

$$\frac{d\xi^D}{d\Omega} = \frac{a(H_{SS}-H_{SD})+bH_{SS}}{\det H} > 0, \qquad \frac{d\xi^S}{d\Omega} = \frac{a(H_{DD}-H_{SD})-bH_{SD}}{\det H},$$

$$\frac{d\xi^R}{d\Omega} = -\frac{a(H_{SS}+H_{DD}-2H_{SD})+b(H_{SS}-H_{SD})}{\det H} < 0,$$

using $H_{SS}+H_{DD}-2H_{SD} = (1,-1)H(1,-1)' > 0$ by positive definiteness, and $H_{SS}-H_{SD} > 0$, $H_{SD} > 0$, $H_{SS} > 0$ from Step 1 (all strictly, in deep scarcity, where $a > 0$ and $b > 0$). This proves part (ii): the data share strictly rises and the routine share strictly falls, while the science share nets the automation-driven inflow $a(H_{DD}-H_{SD})$ (with $H_{DD}-H_{SD} = (1-\alpha)/\xi^D + h'' + m_{S0}/\det M > 0$) against the information-driven outflow $bH_{SD}$, and is not signed in general. □

## Appendix B. Data Construction: Details

### B.1. Introduction Figures Data

This section describes the data sources and methodology used to create the figures shown in the introduction.

*OEWS.* Panel A of Figure 1 is constructed using the Occupational Employment and Wage Statistics (OEWS) program produced by the Bureau of Labor Statistics (BLS). The OEWS program covers wage and salary workers in nonfarm establishments and produces employment and wage estimates for approximately 830 occupations based on a survey of business establishments (employers). The data is published annually with a May reference date and is issued at different levels of aggregation: national, state, metropolitan areas, or industries. I use the industry files. I harmonize all SOC codes to SOC 2010 codes to obtain a consistent panel of occupations for the 2002-2022 period.

Data occupations in Panel A correspond to fourteen SOC 2010 occupations: computer and information research scientists (15-1111); computer systems analysts (15-1121); information security analysts (15-1122); computer programmers (15-1131); software developers, applications (15-1132); software developers, systems software (15-1133); web developers (15-1134); database administrators (15-1141); network and computer systems administrators (15-1142); computer network architects (15-1143); computer user support specialists (15-1151); computer network support specialists (15-1152); computer occupations, all other (15-1199); and operations research analysts (15-2031). These occupations are clustered based on the National Career Clusters Framework. I specifically use two lists curated by O*NET: Data Science & Artificial Intelligence and Web & Cloud.

Science occupations in Panel A correspond to twenty-three SOC 2010 occupations: mathematicians (15-2021); statisticians (15-2041); animal scientists (19-1011); food scientists and technologists (19-1012); soil and plant scientists (19-1013); biochemists and biophysicists (19-1021); microbiologists (19-1022); zoologists and wildlife biologists (19-1023); biological scientists, all other (19-1029); conservation scientists (19-1031); foresters (19-1032); epidemiologists (19-1041); medical scientists, except epidemiologists (19-1042); life scientists, all other (19-1099); astronomers (19-2011); physicists (19-2012); atmospheric and space scientists (19-2021); chemists (19-2031); materials scientists (19-2032); environmental scientists and specialists, including health (19-2041); geoscientists, except hydrologists and geographers (19-2042); hydrologists (19-2043); and physical scientists, all other (19-2099).

*IPEDS.* Panel B of Figure 1 is created using the National Center for Education Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS) survey. I compute the share of Bachelor's and Master's degrees awarded in 85 fields corresponding to the curated lists by O*NET Data Science & Artificial Intelligence and Web & Cloud out of all Bachelor's and Master's degrees awarded in the same year. The data covers the period 1993 to 2024.

Data-related fields are identified by mapping the two O*NET career-cluster lists used throughout the paper, Data Science & Artificial Intelligence and Web & Cloud, into IPEDS instructional programs in three steps. First, I take the thirty O*NET-SOC 2019 occupation codes that constitute the two lists. Second, using O*NET's 2020 Classification of Instructional Programs

(CIP) to O*NET-SOC 2019 crosswalk, I keep every CIP code that maps to at least one of those thirty occupations. Third, I drop eight CIP codes that the crosswalk returns but that do not correspond to the National Career Clusters Framework classification: digital communication and media (09.0702), design and graphic design (50.0401, 50.0409), archaeology and cultural anthropology (45.0102, 45.0103), security and loss prevention services (43.0403, distinct from cyber forensics, 43.0116), logic (38.0102), and one physical-science subfield (40.0512). This leaves 85 CIP codes.

*NSF.* Panels C and D of Figure 1 are created using the National Science Foundation (NSF) Survey of Earned Doctorates (SED). The SED is an annual census of research doctorate recipients from US academic institutions that collects information on educational history, demographic characteristics, graduate funding sources, educational debts, and postgraduation plans.

### B.2. Revelio Labs Data

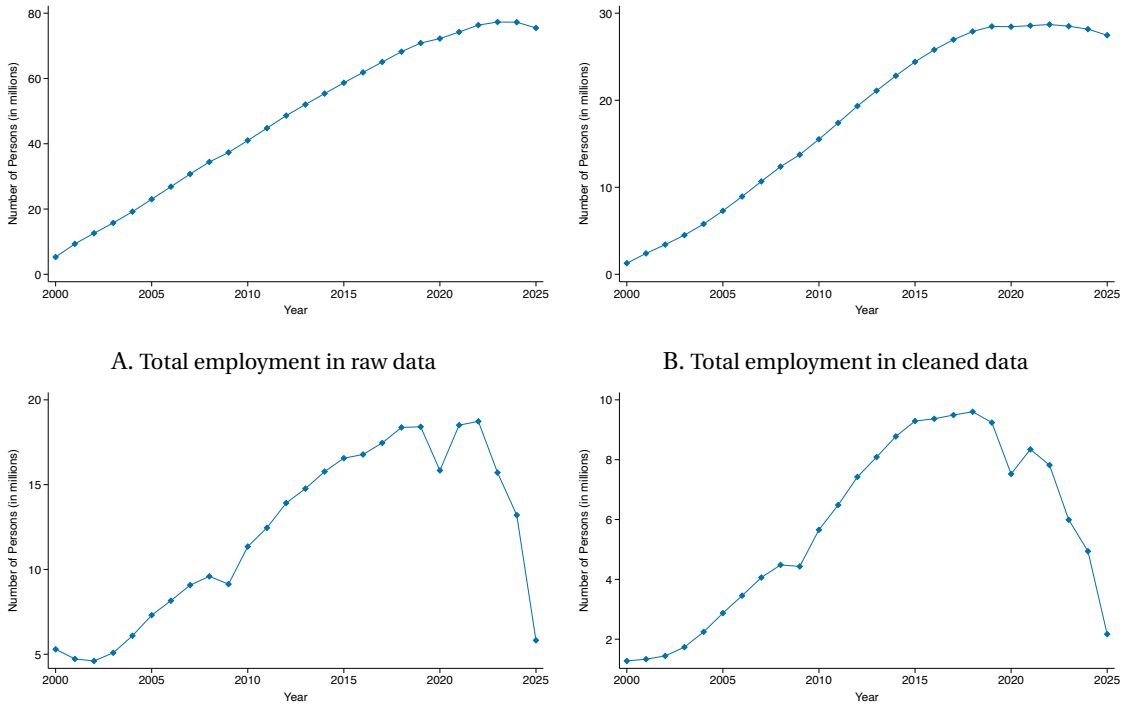
This section describes the data construction in Section 6 using Revelio Labs data.

I impose two restrictions when assembling the different Revelio datasets. First, I focus on individuals completing their first higher education degree in 2000 or later, and holding positions between 2000 and 2022. This period coincides with both reliable coverage in Revelio Labs data and the period of platform emergence and data infrastructure development. As shown in Figures A1 and A2, although Revelio's coverage extends until 2025, there is a significant drop in the number of unique positions both in the raw and the cleaned data. This decline differs from the official BLS employment trend and depicts a surprisingly steep decline for new positions, suggesting caution in using the period 2023-2025.

Second, I restrict the sample to US positions, as the United States led the global development of data infrastructure and platforms (with companies like Snowflake, Databricks, and AWS). Therefore, the US constitutes an ideal laboratory to evaluate whether data growth generates talent reallocation toward data. The final sample consists of 35,130,317 unique individuals working in 5,261,354 unique firms, of which 7,088 are public, across 414 MSAs for whom I track their US employment from 2000 through 2022.

I define occupations using Revelio Labs' standardized role fields. For data occupations, I use Revelio Labs' role_k300 and include the following roles: "Analyst", "Analytics Specialist", "Data Analyst", "Data Engineer", "Data Scientist", and "Database Administrator".

For science occupations, I use Revelio Labs' role_k500 and include the following roles: "Scientist" and "R&D Specialist". Using Revelio Labs' role_k1500, I further add the roles "geologist", "geophysicist", "scientist i", "research chemist", "chemist", "analytical chemist", "microbiologist", or "environmental scientist". I exclude from science occupations observations with role_k1000 corresponding to "Technologist" and those with role_k1500 corresponding to "research technician". All the remaining occupations are classified as routine workers.




A. Total employment in raw data

B. Total employment in cleaned data

C. New positions in raw data

D. New positions in cleaned data

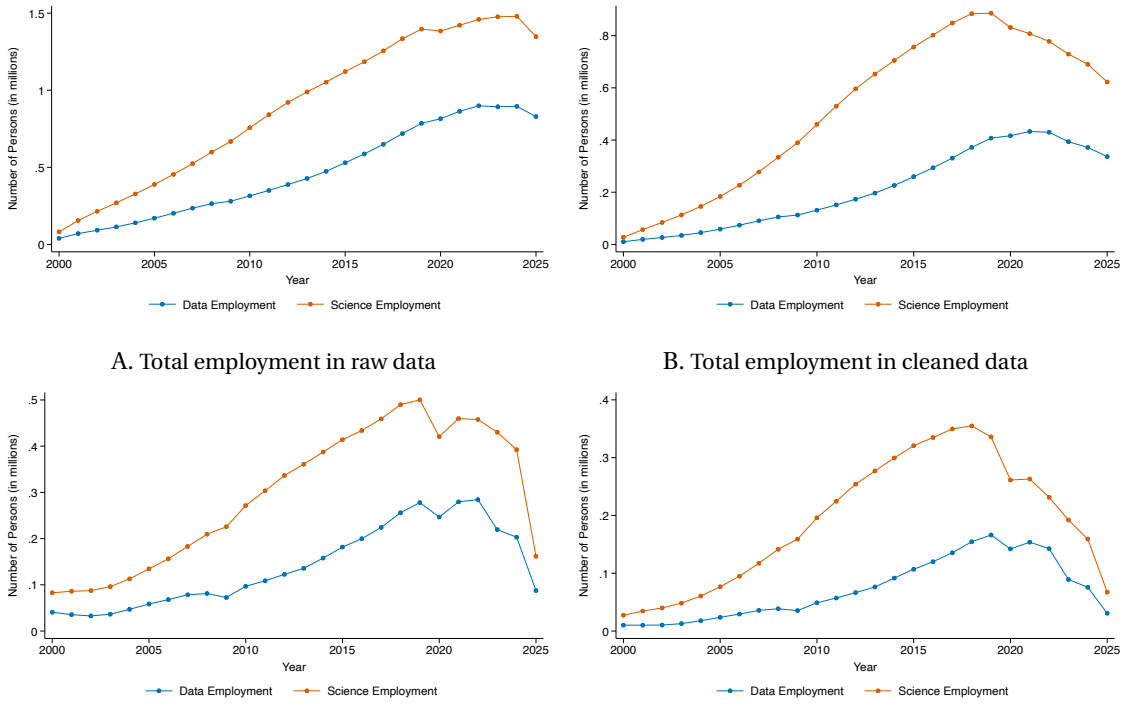
FIGURE A1. Employment counts in Revelio in the US between 2000 and 2025

This figure plots total employment counts in Revelio data for US positions. The raw data corresponds to the data directly downloaded from the WDRS server without any form of filter or cleaning. The cleaned data corresponds to the data processed by the filters described in the Appendix Section B. Panels A and B plot the total employment by year. This measure is a stock. Panels C and D plot the total new positions by year. This measure is a flow. The sample period is 2000-2025. Go back to main text.

For Figure A7 occupations besides data and science are defined as follows. For technological engineering, I use Revelio Labs' role_k300 and include the following roles: "Software Engineer", "Software Developer", "Web Developer", "Hardware Engineer", "Systems Engineer", "Java Developer", "DevOps Engineer", "Network Engineer", "Mobile Specialist", "Solutions Architect", "Test Engineer", "Software Test Engineer", "Automation Engineer", "Infrastructure Engineer", "Technical Architect", "Technology Lead", "System Administrator", "Unix System Administrator", "IT Analyst", "Technology Analyst", and "Technical Consultant".

For physical engineering, I use Revelio Labs' role_k300 and include the following roles: "Mechanical Engineer", "Electrical Engineer", "Chemical Engineer", "Civil Engineer", "Structural Engineer", "Process Engineer", "Manufacturing Engineer", "Quality Engineer", "Field Engineer", "RF Engineer", "Drilling Engineer", "Planning Engineer", "Wireline Engineer", and "Application Engineer".

For finance, I also use Revelio Labs' role_k300 and include the following roles: "Investment Banking", "Risk Analyst", "Financial Analyst", "Actuary", "Wealth Manager", "Trader", "Under-




A. Total employment in raw data

B. Total employment in cleaned data

C. New positions in raw data

D. New positions in cleaned data

FIGURE A2. Employment counts by Science and Data Occupations in Revelio in the US between 2000 and 2025

This figure plots total employment counts in Revelio data for US positions by science and data occupations, as defined in in the Appendix Section B. The raw data corresponds to the data directly downloaded from the WDRS server without any form of filter or cleaning. The cleaned data corresponds to the data processed by the filters described in the Appendix Section B. Panels A and B plot the total employment by year. This measure is a stock. Panels C and D plot the total new positions by year. This measure is a flow. The sample period is 2000-2025. Go back to main text.

writer", "Credit Specialist", "Loan Officer", "Banker", "Financial Advisor", "Fund Accountant", "Compliance", "Treasury Specialist", and "Loan Processor".

### B.3. American Community Survey Data

This section describes the data construction in Section 6 using ACS data.

I use microdata from the Integrated Public Use Microdata Series (IPUMS-USA) (Ruggles et al. 2025). The extract comprises four ACS 5-year samples, covering 2005-2009, 2010-2014, 2015-2019, and 2018-2022, together with the 1-year samples for 2001-2008. Following the IPUMS convention, each multi-year sample is labelled by the final year of its window, so that the four 5-year samples are referred to below as 2009, 2014, 2019, and 2022, respectively. Samples before 2005 are discarded because they carry no PUMA identifier and therefore cannot be mapped to commuting zones. The analysis uses the four 5-year samples throughout. I rely on 5-year rather than 1-year estimates wherever possible because the larger pooled samples are needed to

measure rare occupations (data and science workers) at the commuting zone × industry level. The sample is restricted to employed individuals.

One challenge in using ACS data is that occupation codes change over time, making it difficult to construct consistent occupation definitions across sample years. To address this, I exploit the fact that IPUMS provides two occupation variables: `OCCSOC`, which reports occupations using Standard Occupational Classification (SOC) codes, and `OCC2010`, which provides harmonized Census occupation codes consistent across all years.

I implement a two-step procedure to classify occupations:

*Step 1: Extract SOC → OCC2010 crosswalk.* For years 2010–2017, when `OCCSOC` uses SOC 2010 codes, I create an empirical crosswalk by keeping all unique (`OCCSOC`, `OCC2010`) pairs. This crosswalk maps each SOC 2010 occupation code to its corresponding Census 2010 occupation code(s).

*Step 2: Apply crosswalk to all years.* Using this crosswalk, I identify which `OCC2010` codes correspond to data and science occupations based on their SOC 2010 definitions. I then apply these `OCC2010` codes consistently across all sample years (2005-2022), ensuring time-consistent occupation definitions even though the underlying SOC system changed.

Data and science occupations are defined following the list provided in Appendix Section B.1. The `OCCSOC` variable in IPUMS sometimes uses alphanumeric codes with wildcards (e.g., "15111X") to denote aggregated occupation categories; I include these in the relevant occupation groups. All remaining occupations are classified as routine. The three occupation categories are mutually exclusive and exhaustive.

Industry is reported in IPUMS using the `INDNAICS` variable, which provides NAICS-based industry codes. These codes sometimes contain letters (e.g., "4441Z") indicating aggregated categories. I remove all non-numeric characters and extract the first two digits to create NAICS 2-digit industry codes, resulting in 20 broad sectors.

The ACS reports geographic location using Public Use Microdata Areas (PUMAs), which are statistical areas containing at least 100,000 people. To construct commuting zones (CZs), I use the PUMA-to-CZ crosswalks developed by Autor and Dorn (2013), which can be found at David Dorn's data page. Because PUMA boundaries changed in 2012, I apply different crosswalks for different time periods. For 2005-2011, I use the PUMA 2000 vintage crosswalk. I construct a PUMA 2000 identifier by concatenating the state FIPS code with the 4-digit PUMA code within state. For 2012-2022, I use the PUMA 2010 vintage crosswalk, constructing a PUMA 2010 identifier from the state FIPS code and the 5-digit PUMA code within state.

Some PUMAs span multiple commuting zones. The crosswalk files include an allocation factor (`afactor`) indicating the fraction of each PUMA's population that maps to each CZ. I adjust person weights by this allocation factor: $\text{weight}_{i,\text{CZ}} = \text{weight}_i \times \text{afactor}$, where $\text{weight}_i$ is the original ACS person weight (`PERWT`).

I aggregate the individual-level microdata to the commuting zone × NAICS 2-digit × year

level. A unit (local labor market) is defined as a CZ-industry pair.

### B.4. Enlyft Data

The Enlyft sample was assembled from a list supplied to Enlyft in 2021 comprising selected public and private firms. The February 2023 delivery includes 4,295 Compustat public firms, 5,049 venture-backed private information-technology firms, and 458 firms without a match to either source. For the public-firm analysis, I restrict attention to the 4,295 Compustat firms. I merge Enlyft to an October 2021 vintage of Compustat using the firm's bare web domain to recover its `gvkey`, and then merge the resulting firm-level data to Compustat on `gvkey`. Of the 4,298 distinct domains in the supplied Compustat crosswalk, 4,295 appear in the Enlyft delivery and carry a `gvkey`, so coverage is nearly complete conditional on the requested public-firm sample. I take industry classifications and other firm characteristics from Compustat rather than Enlyft.

Two features of Enlyft are important for interpretation. First, the data are a 2022-2023 cross-sectional snapshot, not a panel of technology adoption. Enlyft records the most recently verified date of a product's use, not when the firm adopted it, and successive deliveries largely re-observe the same firms. Second, the public-firm component is conditioned on firms that were active, US-incorporated, and had a nonmissing web domain in Compustat in October 2021. It therefore excludes firms that had already become inactive and embeds a survivorship and selection filter. In addition, product detections are proxies for adoption rather than direct measures of use intensity. Raw product counts scale strongly with firm size, and Enlyft's confidence measure mostly reflects detection strength. I therefore use normalized or share-based measures where appropriate, exclude "Business Attribute" records from software-product counts, and interpret the results as cross-sectional associations for the covered public firms.

### B.5. Epoch AI

This section describes the data construction for the scaling law analysis in Section 6 using the Epoch AI database.

I use data from the Epoch AI Machine Learning Model Database, which provides comprehensive information on notable machine learning systems. The database is publicly available at Epoch AI data and is regularly updated by Epoch AI researchers who manually collect and verify training details from technical reports, academic papers, and public announcements.

For each model the dataset allows me to observe variables such as: *Training compute (FLOP)*, which measures the total floating-point operations used during training; *Training dataset size*, reported in tokens or datapoints depending on the domain; *Training compute cost*, denominated in 2023 US dollars; *Parameters*, the total number of trainable parameters; *Publication date*; *Domain*, which categorizes models by their primary application area; and *Model*, the system name. I focus the analysis on language models by filtering to observations where the domain

equals "Language". For dataset size, I use the *Training dataset size (gradients)* variable, which reports the number of tokens for language models. Each observation in the dataset corresponds to an individual machine learning model, and all analyses operate at the model level.

In panel D of Figure A3, I compute the tokens-per-parameter ratio to assess whether models follow the Chinchilla-optimal frontier of 20 tokens per parameter (Hoffmann et al. 2022). Models with ratios below 20:1 are classified as "undertrained" in the compute-optimal sense, while models near this ratio are classified as "compute-optimal."

## Appendix C. Model: Details

### C.1. Returns to Scale and the Financing of Data Employment

The returns-to-scale restriction has a simple economic interpretation that I develop here. At any interior candidate satisfying the firm's first-order conditions,

$$w^S S = \beta \pi^E, \qquad w^R R = \psi \pi^E.$$

Hence, firm value can be written as

$$v = (1 - \beta - \psi)\pi^E - w^D D.$$

If $\beta + \psi = 1$, then

$$v = -w^D D < 0$$

whenever $D > 0$. Science and routine wages exhaust expected gross profit, leaving no residual with which to pay data engineers. Since the firm can shut down and obtain zero, an interior allocation with positive data employment cannot be optimal. Thus, $\beta + \psi < 1$ is necessary for an active data occupation in this environment.

The stronger restriction imposed in the main text, $\alpha + \beta + \psi \leq 1$, is sufficient to ensure positive firm value at any interior candidate. Since $1 - \beta - \psi \geq \alpha$,

$$v \geq \alpha \pi^E - w^D D.$$

Using the first-order condition for data employment,

$$w^D D = \frac{\alpha Y \tau \sigma_0^2 \sigma_\epsilon^2 D^\alpha \Omega^{1-\alpha} \sigma_0^2}{\left(\sigma_\epsilon^2 + D^\alpha \Omega^{1-\alpha} \sigma_0^2\right)^2},$$

and therefore

$$\alpha \pi^E - w^D D > 0$$

under $0 < \tau \sigma_0^2 < 1$. Hence $v > 0$.

### C.2. Microfoundation of the Firm's Loss Function

This section microfounds the quadratic loss function of the firm introduced in equation (1). The framework is a simple search model with random matching that is general enough to generate different desired loss functions for the firm (absolute-value, bounded losses, asymmetric ones, etc.), and in particular, the one used in the main body of the paper (quadratic).

*Setup.* After producing $Y_{j,t}$ units, the firm $j$ faces a second problem: bringing output to market. It must allocate commercial effort, such as advertising, inventory placement, sales targeting, or platform emphasis, across a continuum of customer segments $x \in \mathbb{R}$. Two objects determine how successfully it does so. The first is the firm's *commercial allocation* $q(x; a_{j,t})$: a distribution of effort across segments centered on the firm's chosen target $a_{j,t}$. The second is *profitable demand* $d(x; \theta_{j,t})$: a distribution of purchasing intent by customers across segments centered on the latent demand state $\theta_{j,t}$. Thus, $\theta_{j,t}$ represents the location in segment space where customers are most concentrated and therefore where the firm would be able to sell the most output profitably. Shifts in $\theta_{j,t}$ can capture changes in customer tastes, demographics, or willingness-to-pay driven by aggregate conditions (industry demand shocks, technological trends, or input market movements) that are common across firms and therefore make data generated elsewhere in the economy informative about $\theta_{j,t}$. The firm observes neither $\theta_{j,t}$ nor the gap between where it is targeting and where demand actually is.

*Meeting technology.* Firm and customers meet at random times, as in an idealized random search model (see, for example Duffie, Gârleanu, and Pedersen 2005). Thus, customers arrive in each segment $x$ at Poisson rate $\lambda$ and contact the firm. When the firm meets a customer in segment $x$, a transaction occurs with probability proportional to the product of the firm's effort at that segment, $q(x; a_{j,t})$, and the density of profitable demand there, $d(x; \theta_{j,t})$. The expected flow of successful transactions per unit time is therefore

$$m(a_{j,t}, \theta_{j,t}) = \lambda \int_{\mathbb{R}} q(x; a_{j,t})\, d(x; \theta_{j,t})\, dx.$$

The meeting rate $\lambda$ governs the scale of market activity, how many customers the firm encounters per period, while the overlap integral governs the quality of targeting, what fraction of those encounters result in profitable transactions.

The maximum transaction rate is attained when the firm targets perfectly at $a_{j,t} = \theta_{j,t}$.[18] Thus,

$$m(\theta_{j,t}, \theta_{j,t}) = \lambda \int_{\mathbb{R}} q(x; \theta_{j,t})\, d(x; \theta_{j,t})\, dx.$$

[18] This property is not automatic for arbitrary kernels; I impose it as a regularity condition on the class of distributions considered. It holds whenever $q$ and $d$ are symmetric unimodal location families: the overlap integral equals the convolution $(q \star d)(a_{j,t} - \theta_{j,t})$ evaluated at the targeting error, and the convolution of two symmetric unimodal densities is symmetric and unimodal, hence maximized at zero error. In particular, it holds for the Gaussian specialization adopted below, and it guarantees $\tilde{m} \in [0, 1]$.

Normalizing realized transactions by this maximum, $\lambda$ cancels, and I obtain

$$\tilde{m}(a_{j,t}, \theta_{j,t}) \equiv \frac{m(a_{j,t}, \theta_{j,t})}{m(\theta_{j,t}, \theta_{j,t})} = \frac{\int_{\mathbb{R}} q(x; a_{j,t})\, d(x; \theta_{j,t})\, dx}{\int_{\mathbb{R}} q(x; \theta_{j,t})\, d(x; \theta_{j,t})\, dx} \in [0, 1].$$

Hence $\tilde{m}$ is the fraction of transactions the firm achieves given its targeting error, independently of how fast customers arrive. Realized gross profit is

$$\pi^G_{j,t} = Y_{j,t} \tilde{m}(a_{j,t}, \theta_{j,t}).$$

When the firm targets perfectly, $\tilde{m} = 1$ and all output is sold. Mismatch between $a_{j,t}$ and $\theta_{j,t}$ causes effort to flow to segments where demand is weak. Hence, units are produced but fail to reach profitable customers.

*Gaussian specialization.* To proceed further, I now consider a particular case: normal distributions. Let both distributions be Gaussian: $q(x; a_{j,t}) = \phi_{\sigma_q}(x - a_{j,t})$ and $d(x; \theta_{j,t}) = \phi_{\sigma_d}(x - \theta_{j,t})$, where $\sigma^2_q$ captures the dispersion of the firm's targeting technology and $\sigma^2_d$ captures the concentration of demand. In this case, the normalized overlap reduces to

$$\tilde{m}(a_{j,t}, \theta_{j,t}) = \exp\left( -\frac{(a_{j,t} - \theta_{j,t})^2}{2(\sigma^2_q + \sigma^2_d)} \right).$$

The fraction of transactions depends only on the squared distance between the firm's target and the demand center, scaled by total dispersion $\sigma^2_q + \sigma^2_d$. The two variances convey distinct economic intuition. $\sigma^2_d$ reflects market structure: niche markets with concentrated customer bases have low $\sigma^2_d$, while mass markets have high $\sigma^2_d$. $\sigma^2_q$ corresponds to the precision of the firm's targeting technology, with lower values representing more precise targeting. Mismatch is costly when both are small. Indeed, in a concentrated market with very precise targeting, the firm is aiming a narrow beam of effort at a narrow band of customers, and a small targeting error means the beam misses almost entirely. When the variances are both large, the firm's effort and the customer base each spread over a wide range of segments, so even an imperfect aim leaves substantial overlap.

*Quadratic approximation.* Consider now that the firm, while uncertain about where $\theta_{j,t}$ is, does not operate completely in the wrong market. The firm operates in a market it broadly understands, and data engineers help it get the fine targeting right. To formalize this, I will introduce an approximation of the firm's action $a_{j,t}$ around $\theta_{j,t}$.

Let $z \equiv \frac{(a_{j,t} - \theta_{j,t})^2}{2(\sigma^2_q + \sigma^2_d)}$ denote the normalized squared targeting error, so $\tilde{m} = e^{-z}$. Applying the first-order expansion $e^{-z} \approx 1 - z$ around the optimum $z = 0$ (equivalently, $a_{j,t} = \theta_{j,t}$), I obtain

$$\pi^G_{j,t} \approx Y_{j,t} \left[ 1 - \tau (a_{j,t} - \theta_{j,t})^2 \right],$$

where

$$\tau \equiv \frac{1}{2(\sigma_q^2 + \sigma_d^2)}.$$

Again, the approximation holds when the squared targeting error is small relative to total dispersion, that is, when the firm is not targeting a completely wrong segment.[19] The parameter $\tau$ can now be interpreted structurally: it governs the local curvature of the fraction of transactions function at the optimum, determined jointly by demand concentration and targeting precision.

### C.3. Sourcing Shares: Where the New Data Engineers Come From

This section formalizes from which occupations the new data engineers are drawn. Throughout, the economy is in the deep-scarcity region, so $\partial \tilde{w}^D / \partial \Omega > 0$.

*Direct supply response.* By part (i) of Lemma A2, a marginal rise in the data premium at a fixed science premium draws $m_{SD}\, d\tilde{w}^D$ workers out of science and $m_{D0}\, d\tilde{w}^D$ out of routine work. The fraction of the direct inflow sourced from science is therefore

$$\lambda_S^{\text{supply}} \equiv \frac{m_{SD}}{m_{SD} + m_{D0}} \in (0,1),$$

strictly interior because both masses are strictly positive for every $\nu \in (-1,1)$. Science supplies the majority of the direct inflow if and only if $m_{SD} > m_{D0}$.

*General-equilibrium response.* From the equilibrium derivatives in the proof of Proposition 4, the shares of the equilibrium inflow sourced from science and from routine work are

$$\lambda_S \equiv \frac{-d\xi^{S*}/d\Omega}{d\xi^{D*}/d\Omega} = \frac{H_{SD}}{H_{SS}}, \qquad \lambda_R \equiv \frac{-d\xi^{R*}/d\Omega}{d\xi^{D*}/d\Omega} = \frac{H_{SS} - H_{SD}}{H_{SS}} = 1 - \lambda_S,$$

both in $(0,1)$ by part (iii) of Lemma A2: data accumulation draws workers out of *both* occupations, whatever the value of $\nu$. Science supplies the majority if and only if $2H_{SD} > H_{SS}$. Substituting the Hessian entries of part (iii) of Lemma A2, this is

$$\frac{m_{SD} - m_{D0}}{\det M} > \frac{1}{\xi^S} - \frac{1}{\xi^R}. \tag{A7}$$

Condition (A7) is demanding whenever science is the smaller occupation. With $\xi^R \gg \xi^S$, the right-hand side is large and positive, so the science-data margin must be far denser than the routine-data margin for science to supply most of the headcount. Absolute sourcing is thus governed as much by occupation size as by taste correlation.

[19] Naturally, the results in the main body of the paper and the appendixes are exact results for the approximated economy, rather than exact results for the underlying exponential-overlap economy.

*Proportional contraction.* The comparison that survives this size effect is the relative one. Science contracts faster than routine work in percentage terms when $(-d\xi^{S*}/d\Omega)/\xi^{S} > (-d\xi^{R*}/d\Omega)/\xi^{R}$, that is, when $H_{SD}\,\xi^{R} > (H_{SS} - H_{SD})\,\xi^{S}$. Substituting the Hessian entries, the firm-side terms $1/\xi^{R}$ and $1/\xi^{S}$ cancel against the shares that multiply them, leaving

$$\xi^{R} m_{SD} > \xi^{S} m_{D0}.$$

Condition (6) is markedly weaker than (A7) and, unlike it, is made *easier* by science being small. It is the formal content of the statement that data accumulation cannibalizes science.

*The role of* $\nu$. A higher $\nu$ lowers $\mathrm{Var}(\eta^{S} - \eta^{D}) = \sigma_S^2 + \sigma_D^2 - 2\nu\sigma_S\sigma_D$, but this observation does not sign either sourcing condition. The boundary masses $m_{SD}$ and $m_{D0}$ are evaluated at endogenous wage thresholds and are truncated by the availability of the third occupation. In addition, equilibrium wage premia and occupation shares respond to $\nu$. Consequently, the model implies no general monotone relationship between $\nu$ and either the science sourcing share or the proportional contraction condition. I treat these as equilibrium objects to be disciplined empirically.

### C.4. The Constrained Planner Problem Statement

This section states the constrained planner's problem underlying Section 4. The planner may reassign workers across occupations, but cannot leave any worker worse off than under the market allocation. She maximizes the utility of the long-lived owner subject to these participation constraints. This is a constrained Pareto problem.

All quantities are in the per-firm units of Section 3. At each date, the planner chooses a measurable assignment $a_t : \eta \mapsto \{S, D, R\}$, worker consumption $c_t(\eta) > 0$, and owner consumption $c_t^K \geq 0$. The assignment determines

$$z(a_t) \equiv (S_t, D_t, R_t) = L\mathbb{E}\big[\big(\mathbf{1}\{a_t(\eta) = S\}, \mathbf{1}\{a_t(\eta) = D\}, \mathbf{1}\{a_t(\eta) = R\}\big)\big],$$

so $S_t + D_t + R_t = L$. Given the market reservation utilities in (7), the planner solves

$$\max_{\{a_t, c_t, c_t^K\}_{t=0}^{2}} \sum_{t=0}^{2} \rho^t c_t^K \tag{A8}$$

subject, at every date, to the resource constraints

$$c_t^K + L\mathbb{E}[c_t(\eta)] \leq e_t^K + \pi_t^E\big(z(a_t); I_t, \Omega_t\big), \tag{A9}$$

the worker participation constraints

$$\ln c_t(\eta) + \eta^{a_t(\eta)} \geq \bar{u}_t(\eta) \qquad \text{for almost every } \eta, \tag{A10}$$

and the stock dynamics

$$\begin{aligned} I_{t+1} &= (1-\delta_I) I_t + \chi S_t, \\ \omega_{t+1} &= (1-\delta_\omega)\omega_t + \gamma S_t^{\beta} I_t^{\zeta}, \\ \Omega_t &= \left[1 + \phi(N-1)\right]\omega_t, \end{aligned} \tag{A11}$$

with $I_0 > 0$ and $\omega_0 = 0$. The transition constraints apply for $t \in \{0,1\}$. The parameter $\chi$ incorporates the per-firm normalization described in Section 3.

The market allocation is feasible. Set $a_t = a_t^m$, $c_t(\eta) = w_t^{a_t^m(\eta),m}$, and $c_t^K = e_t^K + v_t^m$. Participation and resources then bind, and $c_t^K > 0$.

For any assignment, the least consumption satisfying (A10) is $C_t^{a_t(\eta)}(\eta)$ from (8). Because the planner values owner consumption positively, participation and resources bind whenever the assignment is feasible. Hence, after subtracting the common endowment, her objective is

$$\sum_{t=0}^{2} \rho^t c_t^K - \sum_{t=0}^{2} \rho^t e_t^K = \sum_{t=0}^{2} \rho^t \left[\pi_t^E\big(z(a_t); I_t, \Omega_t\big) - L\mathbb{E}\left[C_t^{a_t(\eta)}(\eta)\right]\right]. \tag{A12}$$

Section 4 does not solve (A8) globally. It asks whether the market allocation is locally optimal for this planner. Since market owner consumption is strictly positive at every date, all sufficiently small perturbations remain feasible. Equation (10) is the change in (A12) generated by a one-date perturbation and its induced stock path. A positive change leaves every worker at her market utility and makes the owner strictly better off.

## Appendix D. Extensions

### D.1. Data as a Byproduct of Output

Suppose data accumulates from output rather than quality:

$$\omega_{t+1} = (1-\delta_\omega)\omega_t + \gamma Y_t, \qquad Y_t = S_t^{\beta} I_t^{\zeta} R_t^{\psi}.$$

All within-period results remain unchanged. Only the cross-period mapping in Corollary 1 is affected. Since $D_0 = 0$, $R_0 = L - S_0$, and $\omega_0 = 0$,

$$\Omega_1 = \Phi\gamma S_0^{\beta} I_0^{\zeta} R_0^{\psi}, \qquad \frac{d\Omega_1}{dS_0} = \Phi\gamma I_0^{\zeta} S_0^{\beta-1} R_0^{\psi-1}\left(\beta R_0 - \psi S_0\right).$$

Because period-1 equilibrium shares do not depend on $I_1$,

$$\frac{d\xi_1^S}{dS_0} = \frac{d\xi^{S*}}{d\Omega}(\Omega_1)\Phi\gamma I_0^{\zeta} S_0^{\beta-1} R_0^{\psi-1}(\beta R_0 - \psi S_0).$$

Hence, in deep scarcity, more science at date 0 reduces the date-1 science share if and only if $\beta R_0 > \psi S_0$, equivalently $S_0 < \beta L/(\beta+\psi)$ or $w_0^S > w_0^R$. If $\beta R_0 < \psi S_0$, the cross-period reallocation reverses.

The transition gives routine workers, as well as scientists, a dynamic contribution. Define

$$\kappa_t^S \equiv \chi\Lambda_t^I + \gamma\beta S_t^{\beta-1} I_t^{\zeta} R_t^{\psi}\Lambda_t^{\omega},$$
$$\kappa_t^R \equiv \gamma\psi S_t^{\beta} I_t^{\zeta} R_t^{\psi-1}\Lambda_t^{\omega}, \qquad \kappa_t^D \equiv 0.$$

The stock values are those of Lemma A6, with the output-based transition.

PROPOSITION A2 (Local Wedges with Output-Generated Data). *Adopt the output-based data law and evaluate along the market path. Then:*

*(i)* $\kappa_t^S, \kappa_t^R > 0$ *for* $t \in \{0,1\}$, *and* $\kappa_2^S = \kappa_2^R = 0$. *Moreover,*

$$\frac{\kappa_t^S}{w_t^S} - \frac{\kappa_t^R}{w_t^R} = \frac{\chi\Lambda_t^I}{w_t^S} > 0 \qquad (t \in \{0,1\}).$$

*(ii) For every one-date market-boundary perturbation with a fixed-shares continuation,*

$$\left.\frac{d\mathcal{W}_t}{d\varepsilon}\right|_{\varepsilon=0} = \rho^t L\left(\kappa_t^S \frac{d\xi_t^S}{d\varepsilon} + \kappa_t^R \frac{d\xi_t^R}{d\varepsilon}\right)\Bigg|_{\varepsilon=0}.$$

*At* $t = 1$, *every marginal flow into data engineering from science or routine work lowers welfare. Every reverse flow raises it. Thus, data engineering is locally excessive relative to both productive occupations.*

*(iii) Raise log earnings in science and routine work at rates* $s_S = \kappa_t^S/w_t^S$ *and* $s_R = \kappa_t^R/w_t^R$. *At* $t = 1$, *the induced equilibrium response raises the science share and lowers the data share. At* $t = 0$, *it raises the science share.*

PROOF. The within-period problem is unchanged, so Lemma A5 applies. Under the new transition,

$$\partial_{S_t}\omega_{t+1} = \gamma\beta S_t^{\beta-1} I_t^{\zeta} R_t^{\psi}, \quad \partial_{R_t}\omega_{t+1} = \gamma\psi S_t^{\beta} I_t^{\zeta} R_t^{\psi-1}, \quad \partial_{D_t}\omega_{t+1} = 0.$$

The fixed-shares chain rule of Lemma A6 gives the two wedges. Their signs follow from the strict positivity of the stock values at nonterminal dates. Since

$$w_t^S = \beta S_t^{\beta-1} I_t^{\zeta} R_t^{\psi}[1 - \tau\sigma^2_{\text{post},t}], \qquad w_t^R = \psi S_t^{\beta} I_t^{\zeta} R_t^{\psi-1}[1 - \tau\sigma^2_{\text{post},t}],$$

the data components cancel after division by wages, which proves (i).

For (ii), the static derivative is zero by Lemma A5. Future compensation bills use the fixed market assignments and are constant. The chain rule therefore leaves only the two displayed stock effects.

For (iii), the earnings changes shift the science and data premia by $(s_S - s_R, -s_R)$. Both $s_S - s_R$ and $s_R$ are positive. With the notation of Lemma A2, the equilibrium derivative is

$$\frac{d}{d\varepsilon}\begin{pmatrix}\xi_t^S\\ \xi_t^D\end{pmatrix}\Bigg|_{\varepsilon=0} = \frac{1}{\det H}\begin{pmatrix}H_{DD}(s_S - s_R) + H_{SD}s_R\\ -H_{SD}(s_S - s_R) - H_{SS}s_R\end{pmatrix}.$$

The Hessian satisfies $H_{SS}, H_{DD}, H_{SD}, \det H > 0$. Hence the first component is positive and the second negative. At date 0, the two-occupation excess-premium equation is strictly decreasing in the science share, so a higher relative science premium raises that share. □

I now study the model's implications for automation when data is a byproduct of output. When data accumulates from total output rather than quality alone, automation has starker implications for misallocation. The reason is that the market misses the automation externality generated by scientists and routine workers. The following results modify Proposition 7 and Corollary 2.

PROPOSITION A3 (Data-Driven Automation and the Two Unpriced Margins). *Adopt the output-based data law and compare the baseline economy ($M \equiv 0$) with the automated economy ($M(\Omega) > 0$ for $\Omega > 0$), at a common, exogenously fixed reference sequence of states and interior allocations $\{(I_s, \omega_s, \xi_s)\}_{s \geq t}$. Under the output-based law, the two economies cannot both generate the reference sequence from a common initial state, since $\omega_{s+1}^{auto} > \omega_{s+1}^{base}$ at common allocations whenever $M(\Omega_s) > 0$. Write $m_s \equiv M(\Omega_s)/R_s$ for machine intensity. For $t \in \{0, 1\}$:*

*(i) The science wedge unambiguously widens: $\kappa_t^{S,auto} > \kappa_t^{S,base}$. All three ingredients rise with automation: the marginal value of ideas, the marginal value of data, and the data byproduct of the marginal scientist, $\gamma\beta S^{\beta-1} I^{\zeta}[R+M]^{\psi}$.*

*(ii) The routine wedge nets two opposing forces: automation raises the shadow value of data, $\Lambda_t^{\omega,auto} > \Lambda_t^{\omega,base}$, but lowers the marginal data byproduct of routine labor, $\gamma\psi S^{\beta} I^{\zeta}[R+M]^{\psi-1}$. For $t = 1$, a sufficient condition for $\kappa_1^{R,auto} > \kappa_1^{R,base}$ is*

$$\psi \ln\big(1 + m_2\big) \geq (1 - \psi)\ln\big(1 + m_1\big).$$

*In particular, the routine wedge widens whenever machine intensity is nondecreasing, $m_2 \geq m_1$, and $\psi \geq 1/2$; and unambiguously so when automation only arrives next period ($m_1 = 0 < m_2$). Absent such a condition, the direction is ambiguous.*

PROOF. All objects are evaluated at the fixed reference sequence of states and allocations, so the posterior variance and its derivatives coincide across the two economies at every date. With

automation and the output-based law, $Y_s = S_s^{\beta} I_s^{\zeta}(R_s + M(\Omega_s))^{\psi}$ and $\omega_{s+1} = (1-\delta_\omega)\omega_s + \gamma Y_s$, so the per-period marginal values entering the recursions of Lemma A6 become

$$\partial_I \pi_s^E = \zeta S_s^{\beta} I_s^{\zeta-1}(R_s + M_s)^{\psi}\big[1 - \tau\sigma^2_{post,s}\big],$$

and

$$\partial_\Omega \pi_s^E = \tau\big(-\partial_\Omega \sigma^2_{post,s}\big) S_s^{\beta} I_s^{\zeta}(R_s + M_s)^{\psi} + \psi(R_s + M_s)^{\psi-1} M'(\Omega_s)\, S_s^{\beta} I_s^{\zeta}\big[1 - \tau\sigma^2_{post,s}\big],$$

where $M_s \equiv M(\Omega_s)$ and the second term of $\partial_\Omega \pi_s^E$ is the automation channel. At a common state and allocation, each of these strictly exceeds its baseline counterpart whenever $M_s > 0$ (the first because $(R_s+M_s)^{\psi} > R_s^{\psi}$, the second additionally because $M' > 0$). The recursion coefficients are weakly larger under automation as well. In the ideas recursion, the cross coefficient is $\rho\gamma\,\partial_I Y_s$ with $\partial_I Y_s = \zeta S_s^{\beta} I_s^{\zeta-1}(R_s + M_s)^{\psi} \geq \zeta S_s^{\beta} I_s^{\zeta-1} R_s^{\psi}$. In the data recursion, automation adds a term absent from the baseline: because automated output depends on the current data stock through $M(\Omega_s)$, the own-transition derivative is

$$\frac{\partial \omega_{s+1}}{\partial \omega_s} = (1-\delta_\omega) + \gamma\psi\, S_s^{\beta} I_s^{\zeta}\big[R_s + M(\Omega_s)\big]^{\psi-1} M'(\Omega_s)\,\Phi \;\geq\; 1 - \delta_\omega,$$

strictly whenever $M' > 0$: the automation feedback of data on its own accumulation. Since the recursions are monotone in their ingredients, and every coefficient is weakly larger and every per-period input strictly larger under automation, $\Lambda_t^{I,auto} > \Lambda_t^{I,base}$ and $\Lambda_t^{\omega,auto} > \Lambda_t^{\omega,base}$ for $t \in \{0,1\}$.

**Part (i):** $\kappa_t^S = \chi\Lambda_t^I + \gamma\beta\, S_t^{\beta-1} I_t^{\zeta}(R_t + M_t)^{\psi}\Lambda_t^{\omega}$. Each factor weakly rises and the multipliers strictly rise, so $\kappa_t^{S,auto} > \kappa_t^{S,base}$.

**Part (ii):** $\kappa_t^R = \gamma\psi\, S_t^{\beta} I_t^{\zeta}(R_t + M_t)^{\psi-1}\Lambda_t^{\omega}$, and $(R_t + M_t)^{\psi-1} < R_t^{\psi-1}$ for $\psi \in (0,1)$: the loading falls while the multiplier rises. For $t = 1$, $\Lambda_1^{\omega} = \rho\,\Phi\,\partial_\Omega \pi_2^E$ in both economies (the additional own-transition term above does not enter, since $\hat{\lambda}_3^{\omega} = 0$), well defined as a ratio comparison because $\tau > 0$ makes $\Lambda_1^{\omega,base} > 0$, and dropping the (positive) automation channel,

$$\frac{\Lambda_1^{\omega,auto}}{\Lambda_1^{\omega,base}} \;\geq\; \left(\frac{R_2 + M_2}{R_2}\right)^{\psi} = (1+m_2)^{\psi}, \qquad \text{so} \qquad \frac{\kappa_1^{R,auto}}{\kappa_1^{R,base}} \;\geq\; (1+m_1)^{\psi-1}(1+m_2)^{\psi}.$$

The right-hand side is at least one if and only if $\psi\ln(1+m_2) \geq (1-\psi)\ln(1+m_1)$, which delivers the sufficient condition. The two special cases follow since $\psi \geq 1/2$ and $m_2 \geq m_1$ imply $\psi\ln(1+m_2) \geq \psi\ln(1+m_1) \geq (1-\psi)\ln(1+m_1)$, and $m_1 = 0$ makes the right-hand side zero while $m_2 > 0$ makes the left-hand side strictly positive. □

The comparison is deliberately pointwise: along the two economies' *actual* paths, the automated economy accumulates strictly more data at common allocations, and since the marginal

information value $\partial_\Omega \pi^E$ eventually declines in saturation, the actual-path comparison of multipliers is not globally signed. The pointwise comparison isolates the technology effect at like-for-like states. The cleanest content free of this caveat is the date-1 statement, where $\Lambda_1^\omega = \rho\Phi\,\partial_\Omega \pi_2^E$ requires only a common date-2 evaluation point. This contrasts with the quality-based baseline of Proposition 7, where common allocations do generate a common state path.

The next corollary shows how automation and data accumulation in the deep-scarcity region reinforce the notion that routine work cannibalizes itself. Intuitively, routine work generates the data that enables its own displacement: as machines accumulate, routine workers flow toward data processing, precisely because they enable data accumulation.

COROLLARY A1 (Automation in Deep Scarcity Draws Talent from Two Unpriced Margins). *Adopt the output-based data law, and suppose the period-t equilibrium of the automated economy is interior with* $(L\xi_t^D, \Omega_t) \in \mathcal{S}^{deep}$. *Then, a marginal increase in* $\Omega_t$ *strictly raises the data share and strictly lowers the routine share. At the nonterminal interior date* $t = 1$, *workers flowing into data engineering come from routine work with certainty and possibly from science. At the terminal date* $t = 2$, *the same within-period employment comparative statics apply, but all dynamic wedges are zero.*

PROOF. The equilibrium statement is Corollary 2, whose proof involves only the within-period problem (premia, potential, and Hessian) and is therefore unaffected by the law of motion of data, which matters only for the dynamic wedges. At $t = 1$, the wedge statements follow from part (i) of Proposition A2: $\kappa_1^S, \kappa_1^R > 0$ and $\kappa_1^D = 0$. At the terminal date $t = 2$, that proposition instead gives $\kappa_2^S = \kappa_2^R = \kappa_2^D = 0$. □

### D.2. Infinite-Horizon

This section extends the baseline three-period model to an infinite-horizon environment. The purpose of the extension is to show that the paper's model mechanisms do not rely on the finite-horizon structure. Two central forces survive unchanged: (i) successful innovation increases the data stock and, through this channel, the attractiveness of data engineering; and (ii) the market continues to leave the dynamic contribution of scientists unpriced, because wages reflect only current marginal products while scientists generate intertemporal spillovers through the accumulation of ideas and data. Naturally, with an infinite horizon this wedge is strictly positive at *every* date, since every period has a future. An extension with a closed-form characterization of the entire dynamic system is left for future work.

There are infinitely many periods indexed by $t = 0, 1, 2, \ldots$. In every period, a generation of mass $L$ is born, chooses an occupation, works for one period, and then retires. Firms are competitive, identical, and operate in a symmetric equilibrium.

The key observation is that the economy can be written recursively with state $(I_t, \omega_t)$, where $I_t$ is the stock of ideas and $\omega_t$ is the firm-level data stock. I maintain the per-firm normalization

of Section 3 throughout this section: $(S_t, D_t, R_t)$ and $L$ denote per-firm quantities, and $\chi$ absorbs the number of firms in the ideas law of motion. In the symmetric equilibrium, the aggregate data stock is

$$\Omega_t = \omega_t[1 + \phi(N-1)].$$

Conditional on the inherited state $(I_t, \omega_t)$, the within-period equilibrium is exactly the same as in the baseline model: firms choose labor demand taking $(I_t, \Omega_t)$ as given, wages equal marginal products, workers choose occupations according to the occupational-choice block, and the resulting allocation determines next period's stocks,

$$I_{t+1} = (1-\delta_I)I_t + \chi S_t,$$
$$\omega_{t+1} = (1-\delta_\omega)\omega_t + \gamma S_t^{\beta} I_t^{\zeta}.$$

Since the intra-period equilibrium depends on $(I_t, \omega_t)$ only through current marginal products, Propositions 1 and 2 characterize the allocation state-by-state for any horizon.

The notion of information scarcity extends immediately to the infinite-horizon setting. It is defined identically at any state $(I, \omega)$, since the condition $D^{\alpha}\Omega^{1-\alpha}\sigma_0^2 < \sigma_\epsilon^2$ depends only on current allocations and the aggregate data stock. The same holds for the deep-scarcity region and its closed form (by Lemma A1). Further, the single-peaked path for data employment established in the main text is not a consequence of the economy ending in period 2: Proposition 4 is a statement about the within-period equilibrium map $\xi^*(\Omega)$, which does not involve the horizon at all, so it naturally applies here. The data share rises in $\Omega$ up to the reallocation threshold $\bar{\Omega}$, at the deep-scarcity boundary, and declines beyond it, at any date of the infinite-horizon economy.

The local comparison requires the discounted output derivative to be finite. Bounded stocks and uniformly interior market allocations deliver this property. Proposition 4 bounds the science and routine shares away from zero along the market path. Market wages and compensation requirements are therefore uniformly bounded.

LEMMA A7 (Uniform Bounds Along the Infinite Market Path). *Fix* $\rho \in (0,1)$ *and initial conditions* $I_0 > 0$, $\omega_0 = 0$. *Then:*

*(i) along any feasible path, the stocks are bounded:* $I_t \le \bar{I} \equiv \max\{I_0,\ \chi L/\delta_I\}$ *and* $\omega_t \le \bar{\omega} \equiv \max\{\omega_0,\ \gamma L^{\beta}\bar{I}^{\zeta}/\delta_\omega\}$*;*

*(ii) the infinite market path is well defined. The science and routine shares are uniformly positive. All three shares and wages are uniformly positive for* $t \ge 1$. *The active date-*0 *wages are positive, while* $w_0^{D,m} = 0$*;*

*(iii)* $\pi_t^E$ *and* $L\mathbb{E}[C_t^{a(\eta)}(\eta)]$ *are uniformly bounded over dates and assignments. Hence* $\sum_{t\ge 0}\rho^t[\pi_t^E - L\mathbb{E}(C_t^{a(\eta)}(\eta))]$ *converges absolutely.*

PROOF. **Part (i):** From $I_{t+1} = (1-\delta_I)I_t + \chi S_t \le (1-\delta_I)I_t + \chi L$, induction gives $I_{t+1} \le \max\{I_t, \chi L/\delta_I\} \le \max\{I_0, \chi L/\delta_I\}$. The data bound follows identically from $\gamma Q_t \le \gamma L^{\beta}\bar{I}^{\zeta}$.

**Part (ii):** The path exists by Propositions 1 and 2. Proposition 4 gives uniform positive lower bounds for the science and routine shares. Part (i) then bounds both stocks above. It also bounds $I_t$ and $\omega_t$ away from zero for $t \geq 1$. Hence $\Omega_t$ lies in a compact subset of $(0,\infty)$ for $t \geq 1$. Continuity and interiority of the equilibrium map give a uniform positive lower bound for the data share on this set. The marginal-product formulas now bound all wages above and, for $t \geq 1$, away from zero. At date 0, the science and routine wages are positive and $w_0^{D,m} = 0$.

**Part (iii):** Let $w_t^{\max} = \max_o w_t^{o,m}$. For every occupation $o$,

$$C_t^o(\eta) \leq w_t^{\max} \exp\{\max_{o'} \eta^{o'} - \eta^o\} \leq w_t^{\max}\left(e^{\eta^S - \eta^o} + e^{\eta^D - \eta^o} + e^{-\eta^o}\right).$$

The wage upper bound is uniform by part (ii). Every expectation on the right is a Gaussian exponential moment and is finite. Thus, compensation bills are uniformly bounded over dates and assignments. This includes date 0: the reservation utility excludes data engineering, but $C_0^D(\eta)$ remains finite. Output is uniformly bounded by part (i). Absolute convergence follows from $\rho < 1$. □

Anchor reservation utilities at the infinite market path. For a one-date market-boundary perturbation at date $t$, followed by fixed market shares, define

$$\mathcal{W}_{\infty,t}(\varepsilon) \equiv \sum_{s=t}^{\infty} \rho^s \left[\pi_s^E(\varepsilon) - L\mathbb{E}\left[C_s^{a_s(\varepsilon)(\eta)}(\eta)\right]\right].$$

Lemma A7 makes this series absolutely convergent.

The following proposition shows that the local result does not rely on a terminal date. Innovation creates ideas and data, these stocks shape future wage premia, and market wages omit the future benefits of scientific labor.

PROPOSITION A4 (Infinite-Horizon Local Welfare). *Evaluate along the infinite market path. Then:*

*(i) At every date, the market assignment uniquely maximizes current compensated surplus.*

*(ii) The fixed-shares stock values extend as absolutely convergent series, and*

$$\kappa_t = \chi \Lambda_t^I + \gamma\beta S_t^{\beta-1} I_t^{\zeta} \Lambda_t^{\omega} > 0 \qquad \text{for every } t \geq 0.$$

*The wedges on data engineers and routine workers are zero.*

*(iii) For every one-date market-boundary perturbation,*

$$\left.\frac{d\mathcal{W}_{\infty,t}}{d\varepsilon}\right|_{\varepsilon=0} = \rho^t \kappa_t L \left.\frac{d\xi_t^S}{d\varepsilon}\right|_{\varepsilon=0}.$$

*(iv) At every date, a small compensated increase in science raises discounted surplus.*

PROOF. **Part (i):** Lemma A5 is a date-by-date statement. It applies at every market state.

**Part (ii):** Fix market shares from date $s$ onward and let

$$W_s(I_s, \omega_s) \equiv \sum_{u=s}^{\infty} \rho^{u-s} \pi_u^E.$$

Part (ii) of Lemma A7 places a tube around the market path on which $\partial_I \pi_u^E$ and $\partial_\Omega \pi_u^E$ are uniformly bounded. On this tube, the one-step transition Jacobian is

$$T_u = \begin{pmatrix} 1-\delta_I & 0 \\ \gamma\zeta S_u^{\beta} I_u^{\zeta-1} & 1-\delta_\omega \end{pmatrix}.$$

Let $r = \max\{1-\delta_I, 1-\delta_\omega\} < 1$. Uniform boundedness of the lower-left entry gives

$$\| T_{s+n-1} \cdots T_s \| \le C(n+1) r^n$$

for some $C < \infty$. Hence the derivative of the $n$th discounted output term is bounded by $C'(n+1)(\rho r)^n$. This bound is summable. Differentiation under the series is therefore valid, and the stock values are absolutely convergent and satisfy the recursions of part (i) of Lemma A6, without a terminal condition.

One additional date-$t$ scientist changes next-period stocks by $(\chi, \gamma\beta S_t^{\beta-1} I_t^{\zeta})$. Thus,

$$\kappa_t = \chi\rho\hat{\lambda}_{t+1}^I + \gamma\beta S_t^{\beta-1} I_t^{\zeta} \rho \hat{\lambda}_{t+1}^{\omega} \ge \chi\rho\, \partial_I \pi_{t+1}^E > 0.$$

Neither data engineers nor routine workers enter a transition, so their wedges are zero.

**Part (iii):** The proof of part (iii) of Proposition 5 applies. Downstream assignments and compensation bills are fixed. Part (ii) justifies differentiation under the infinite sum.

**Part (iv):** Shift the date-$t$ science premium as in the proof of Proposition 5. Then $d\xi_t^S / d\varepsilon > 0$. Parts (ii) and (iii) imply a strictly positive derivative. □

## Appendix E. Empirical Plausibility of the Deep Scarcity Region

This section develops the revealed preference argument summarized in Section 6 for placing the US economy inside the deep-scarcity region $\mathcal{S}^{deep}$ of Section 3.3. This argument is not a direct identification of the aggregate economy's location in the model's state space. Instead, it should be seen as a suggestive exercise on the empirical plausibility of the economy being located in the deep scarcity region.

I use the Epoch AI database, to extract LLMs training compute (FLOPS), dataset size (tokens), and training costs in 2023 US dollars. Appendix Section B.5 provides additional details on the data. Appendix Figure A3 presents six forms of evidence that LLMs are in information scarcity. Panels A and C show that both training compute and dataset sizes have grown steadily since 2010. Panel B indicates that training compute increases as a power law in dataset size (slope =

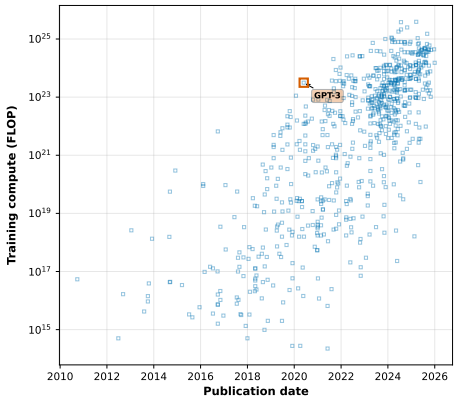
1.01) through 2020, with no evidence of saturation. This empirical relationship between inputs is consistent with the scaling-law framework of Kaplan et al. (2020). Panel D provides direct evidence that LLMs were data-starved as recently as 2022. Following Hoffmann et al. (2022), the compute-optimal ratio is approximately 20 tokens per parameter (green dashed line). Most models operated far below this frontier: GPT-3 used only 1.4 tokens per parameter, while Gopher and PaLM used 1.1 and 1.4 respectively—roughly 14 times below the optimal ratio. Panels E and F complement this interpretation through cost data. Panel E shows training costs reaching $100 million by 2024, with a mean exceeding $3.6 million for models trained since 2023. These expenditures close the revealed preference argument in the following sense: they establish that firms place a high value on frontier *scale*, of which data is one input alongside compute, parameters, and architecture. On their own, Panels E and F do not separately identify the marginal value of data. That role is played by Panels B and D, where the scaling relationship and the token-per-parameter shortfall speak to data specifically. Panel F supports this interpretation by showing a strong positive relationship between cost and compute (slope = 0.83), confirming that high-cost models are indeed frontier models pursuing greater scale.

## Appendix F. Additional Tables

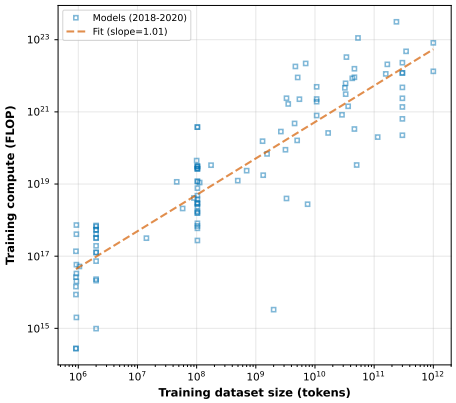
TABLE A1. Regressions across firms, public firms, and local labor markets of contemporaneous log data employment on past log science employment (no weights).

| Level of Analysis | All firms | | Public firms | | Local labor markets | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Log (Science employment) t-1 | 0.015*** | 0.015*** | 0.062*** | 0.063*** | 0.061*** | 0.046*** | 0.042*** | 0.028*** |
| | (0.0050) | (0.0049) | (0.015) | (0.015) | (0.0072) | (0.0070) | (0.0072) | (0.0070) |
| Log (Science employment) t-2 | 0.012*** | 0.012*** | 0.015 | 0.015 | 0.017*** | 0.010* | 0.0093 | 0.0027 |
| | (0.0039) | (0.0039) | (0.011) | (0.011) | (0.0057) | (0.0056) | (0.0058) | (0.0057) |
| Log (Science employment) t-3 | 0.013*** | 0.014*** | 0.010 | 0.013 | 0.023*** | 0.017*** | 0.017*** | 0.010* |
| | (0.0038) | (0.0037) | (0.010) | (0.010) | (0.0053) | (0.0053) | (0.0056) | (0.0056) |
| Log (Science employment) t-4 | 0.012*** | 0.014*** | 0.023** | 0.021* | 0.032*** | 0.028*** | 0.023*** | 0.017*** |
| | (0.0038) | (0.0037) | (0.011) | (0.011) | (0.0053) | (0.0052) | (0.0055) | (0.0055) |
| Log (Science employment) t-5 | 0.0053 | 0.011** | 0.033** | 0.036*** | 0.047*** | 0.048*** | 0.029*** | 0.018*** |
| | (0.0045) | (0.0044) | (0.013) | (0.012) | (0.0064) | (0.0063) | (0.0062) | (0.0061) |
| Total employment controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Firm FE | ✓ | ✓ | — | — | — | — | — | — |
| GVKEY FE | — | — | ✓ | ✓ | — | — | — | — |
| NAICS 2-MSA FE | — | — | — | — | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | — | ✓ | — | ✓ | — | — | — |
| NAICS 2-Year FE | — | ✓ | — | ✓ | — | ✓ | — | ✓ |
| MSA-year FE | — | — | — | — | — | — | ✓ | ✓ |
| Cumulative effect (5 lags) | 0.057*** | 0.065*** | 0.143*** | 0.148*** | 0.181*** | 0.149*** | 0.121*** | 0.077*** |
| | (0.011) | (0.011) | (0.030) | (0.029) | (0.015) | (0.014) | (0.014) | (0.014) |
| Joint test p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Adjusted R-squared | 0.91 | 0.91 | 0.96 | 0.96 | 0.94 | 0.95 | 0.95 | 0.95 |
| Observations | 123,532 | 123,381 | 11,822 | 11,822 | 61,671 | 61,671 | 61,671 | 61,671 |

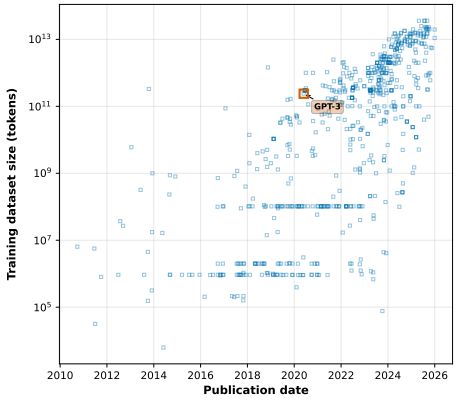
*Notes.* Standard errors, in parentheses, are clustered at the firm level in columns (1) and (2), the GVKEY level in columns (3) and (4), and the NAICS 2-MSA level in columns (5), (6), (7), and (8). $^{***}$, $^{**}$, and $^{*}$ correspond to rejection of the null that the coefficient is zero, at the 1%, 5%, and 10% levels, respectively. The dependent variable is a unit (firm, public firm, or local labor market) log data employment in year $t$. The independent variables are lags at $t-1$, $t-2$, $t-3$, $t-4$, and $t-5$ of the log science employment. In all regressions, I control for the log of the total employment of the unit at $t$, $t-1$, $t-2$, $t-3$, $t-4$, and $t-5$. The benchmark specification, in columns (1), (3), and (5) is $\log(\text{data employment})_{i,t} = \sum_{s=1}^{5} b_s \log(\text{science employment})_{i,t-s} + \sum_{s=0}^{5} c_s \log(\text{total employment})_{i,t-s} + g_i + d_t + u_{i,t}$. $g_i$ are unit fixed effects, $d_t$ are year fixed effects. In columns (2), (4), and (6), I include industry-year fixed effects. In columns (7) and (8) I include MSA-year fixed effects, and both industry-year fixed effects and MSA-year fixed effects, respectively. Units must have at least 10 workers to be included in columns (1), (2), (5), (6), (7), and (8). Regressions are unweighted. The sample period is 2000-2022. Go back to main text.



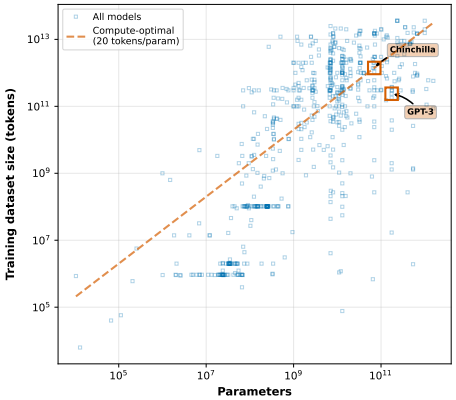

A. LLMs compute power



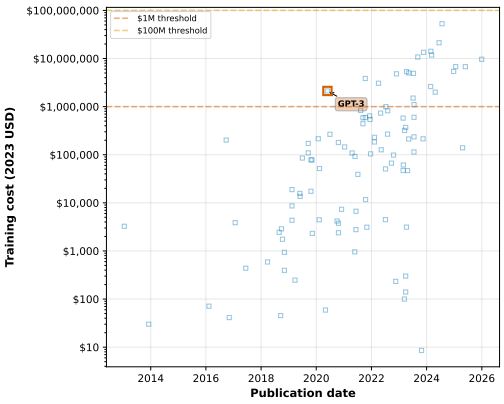

B. LLMs compute and dataset size



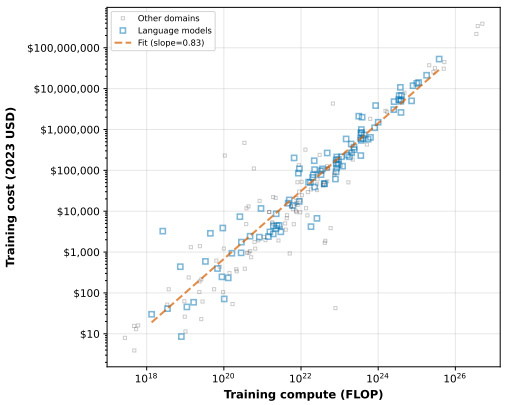

C. LLMs dataset size




D. LLMs dataset size and number of parameters




E. LLMs training costs




F. LLMs training costs and compute power

FIGURE A3. Large language models dataset size, performance, and investment

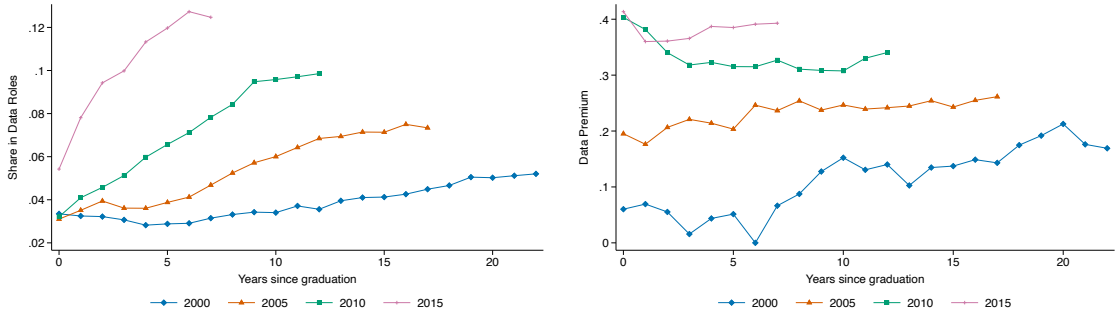
This figure documents patterns on LLMs training compute, dataset size, and training costs between 2010 and 2024. Panel A shows training compute for models published between 2010 and 2024. Panel B reproduces Kaplan et al. (2020) scaling laws for models trained between 2018 and 2020 ($n$=105). Panel C shows training dataset sizes for models published between 2010 and 2024. Panel D plots training tokens against model parameters, with the green dashed line indicating the compute-optimal ratio of 20 tokens per parameter from Hoffmann et al. (2022). Panel E shows training costs in 2023 US dollars for models published between 2012 and 2024. Panel F plots training costs against compute for all domains, with language models highlighted in blue. I specifically report GPT-3 175B (OpenAI, 2020) through several panels as reference, as well as Chinchilla 70B (DeepMind, 2022) for optimal dataset size in panel D. Go back to main text.

TABLE A2. Regressions across firms, public firms, and local labor markets of contemporaneous log data employment on past log science employment using NAICS 4 industries.

| Weighted Regression? | Yes | | | | No | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Log (Science employment) t-1 | 0.045*** | 0.028*** | 0.035*** | 0.021*** | 0.044*** | 0.032*** | 0.031*** | 0.019*** |
| | (0.0078) | (0.0057) | (0.0076) | (0.0054) | (0.0045) | (0.0044) | (0.0046) | (0.0045) |
| Log (Science employment) t-2 | 0.015*** | 0.014*** | 0.011** | 0.010** | 0.016*** | 0.014*** | 0.011*** | 0.0071* |
| | (0.0048) | (0.0040) | (0.0048) | (0.0040) | (0.0036) | (0.0036) | (0.0037) | (0.0036) |
| Log (Science employment) t-3 | 0.012*** | 0.010*** | 0.012*** | 0.0097** | 0.018*** | 0.016*** | 0.014*** | 0.010*** |
| | (0.0045) | (0.0038) | (0.0045) | (0.0039) | (0.0036) | (0.0035) | (0.0036) | (0.0036) |
| Log (Science employment) t-4 | 0.0096** | 0.0092** | 0.0064 | 0.0061 | 0.018*** | 0.017*** | 0.013*** | 0.0088** |
| | (0.0046) | (0.0039) | (0.0046) | (0.0040) | (0.0035) | (0.0035) | (0.0035) | (0.0035) |
| Log (Science employment) t-5 | 0.025*** | 0.030*** | 0.019*** | 0.021*** | 0.038*** | 0.045*** | 0.030*** | 0.027*** |
| | (0.0064) | (0.0049) | (0.0063) | (0.0048) | (0.0041) | (0.0040) | (0.0040) | (0.0040) |
| Total employment controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| NAICS 4-MSA FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | — | — | — | ✓ | — | — | — |
| NAICS 4-Year FE | — | ✓ | — | ✓ | — | ✓ | — | ✓ |
| MSA-year FE | — | — | ✓ | ✓ | — | — | ✓ | ✓ |
| Cumulative effect (5 lags) | 0.106*** | 0.091*** | 0.083*** | 0.068*** | 0.134*** | 0.123*** | 0.099*** | 0.072*** |
| | (0.019) | (0.013) | (0.018) | (0.013) | (0.010) | (0.009) | (0.010) | (0.010) |
| Joint test p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Adjusted R-squared | 0.97 | 0.98 | 0.97 | 0.98 | 0.92 | 0.93 | 0.93 | 0.93 |
| Observations | 125,916 | 125,916 | 125,916 | 125,916 | 125,916 | 125,916 | 125,916 | 125,916 |

*Notes*. Standard errors, in parentheses, are clustered at the NAICS 4-MSA level. $^{***}$, $^{**}$, and $^{*}$ correspond to rejection of the null that the coefficient is zero, at the 1%, 5%, and 10% levels, respectively. The dependent variable is a unit (local labor market) log data employment in year $t$. The independent variables are lags at $t-1$, $t-2$, $t-3$, $t-4$, and $t-5$ of the log science employment. In all regressions, I control for the log of the total employment of the unit at $t$, $t-1$, $t-2$, $t-3$, $t-4$, and $t-5$. The benchmark specification, in columns (1) and (5) is log (data employment)$_{i,t}$ = $\sum_{s=1}^{5} b_s$ log (science employment)$_{i,t-s}$ + $\sum_{s=0}^{5} c_s$ log (total employment)$_{i,t-s}$ + $g_i + d_t + u_{i,t}$. $g_i$ are unit fixed effects, $d_t$ are year fixed effects. In columns (2) and (6), I include industry-year fixed effects. In columns (3) and (7), I include MSA-year fixed effects. In columns (4) and (8), I include both industry-year fixed effects and MSA-year fixed effects. Units must have at least 10 workers to be included in the regressions. Regressions in columns (1), (2), (3), and (4) are weighted by their employment level in the first year they appear in the panel (entry weight). Regressions in columns (5), (6), (7), and (8) are unweighted. The sample period is 2000-2022. Go back to main text.

# Appendix G. Additional Figures




A. Share of data occupation by cohort and the year since graduation

B. Data premium by cohort and the year since graduation

FIGURE A4. Share of data Occupations and compensation by cohort and year since graduation

This figure documents cohort-level patterns in data occupation choice. A cohort is defined based on the year an individual obtains their *first* degree. In this figure, I focus on the fields of mathematics and statistics. The data share is defined as data share$_{t,x} = \frac{\text{count of data positions}_{t,x}}{\text{count of data positions}_{t,x}+\text{count of science positions}_{t,x}+\text{count of routine positions}_{t,x}}$ for cohort graduating in year $t$ and observed $x$ years after graduation. Panel A plots the share of workers in data occupations by cohort and years since graduation. Lines show cohorts graduating in 2000 (blue), 2005 (orange), 2010 (green), and 2015 (purple). Each line traces the data share from year 0 to year 22 post-graduation, conditional on the cohort being observable. The maximum year covered is 2022. Panel B plots the data premium by cohort and years since graduation. Lines show cohorts graduating in 2000 (blue), 2005 (orange), 2010 (green), and 2015 (purple). Each line traces the data premium from year 0 to year 22 post-graduation, conditional on the cohort being observable. The maximum year covered is 2022. The sample period is 2000-2022. Go back to main text.

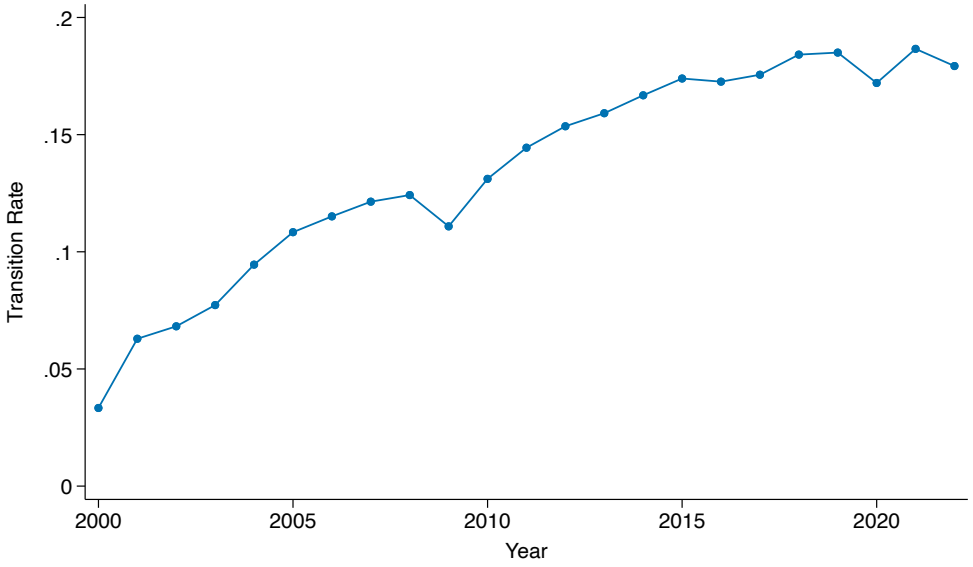



FIGURE A5. Total outflows from university science

This figure plots total annual outflows from university science - defined as positions in NAICS 61 (Educational Services) held by individuals in science occupations - to any other industry. For each year, the outflow rate is expressed as the share of the university-science stock in the prior year that had transitioned to a non-university (private-sector) position by the current year: $\text{Outflow}_t = \frac{\text{count of workers in NAICS 61 science at } t-1 \text{ observed outside NAICS 61 at } t}{\text{count of workers in NAICS 61 science at } t-1}$. The sample period is 2000-2022. Go back to main text.

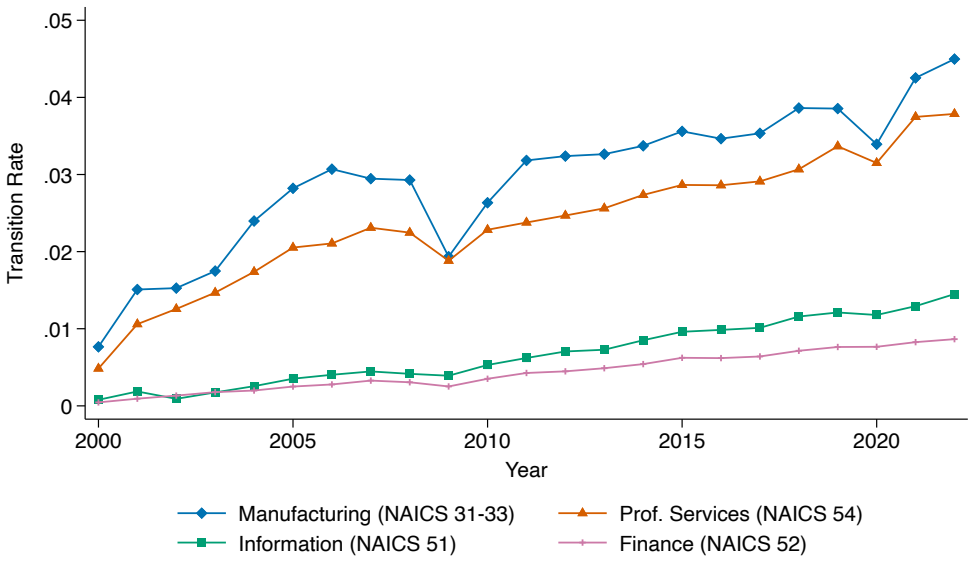
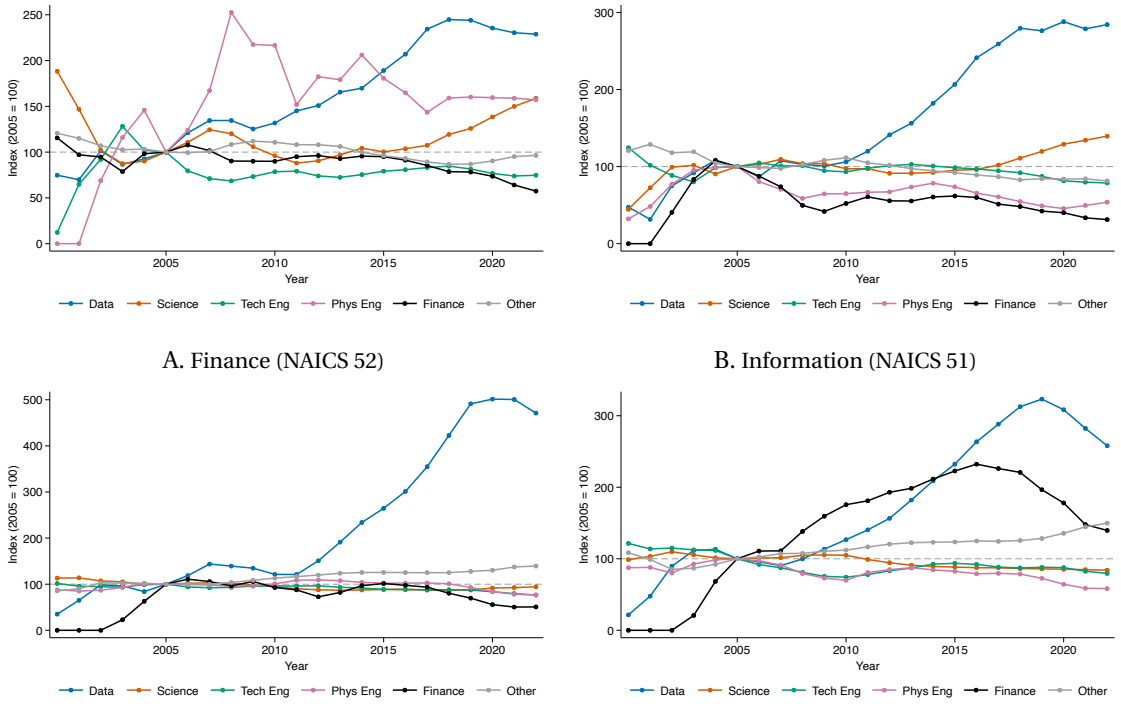



FIGURE A6. Outflows from university science to top destination industries

This figure plots annual outflows from university science - defined as positions in NAICS 61 (Educational Services) held by individuals in science occupations - to the four largest destination industries: Manufacturing (NAICS 31-33, blue), Professional Services (NAICS 54, orange), Information (NAICS 51, green), and Finance (NAICS 52, purple). For each year and destination, the outflow is expressed as a percentage of the university-science stock in that year: Outflow to $k_t = \frac{\text{count of workers in NAICS 61 science at } t-1 \text{ and NAICS } k \text{ at } t}{\text{count of workers in NAICS 61 science at } t-1}$. The sample period is 2000-2022. Go back to main text.




A. Finance (NAICS 52)

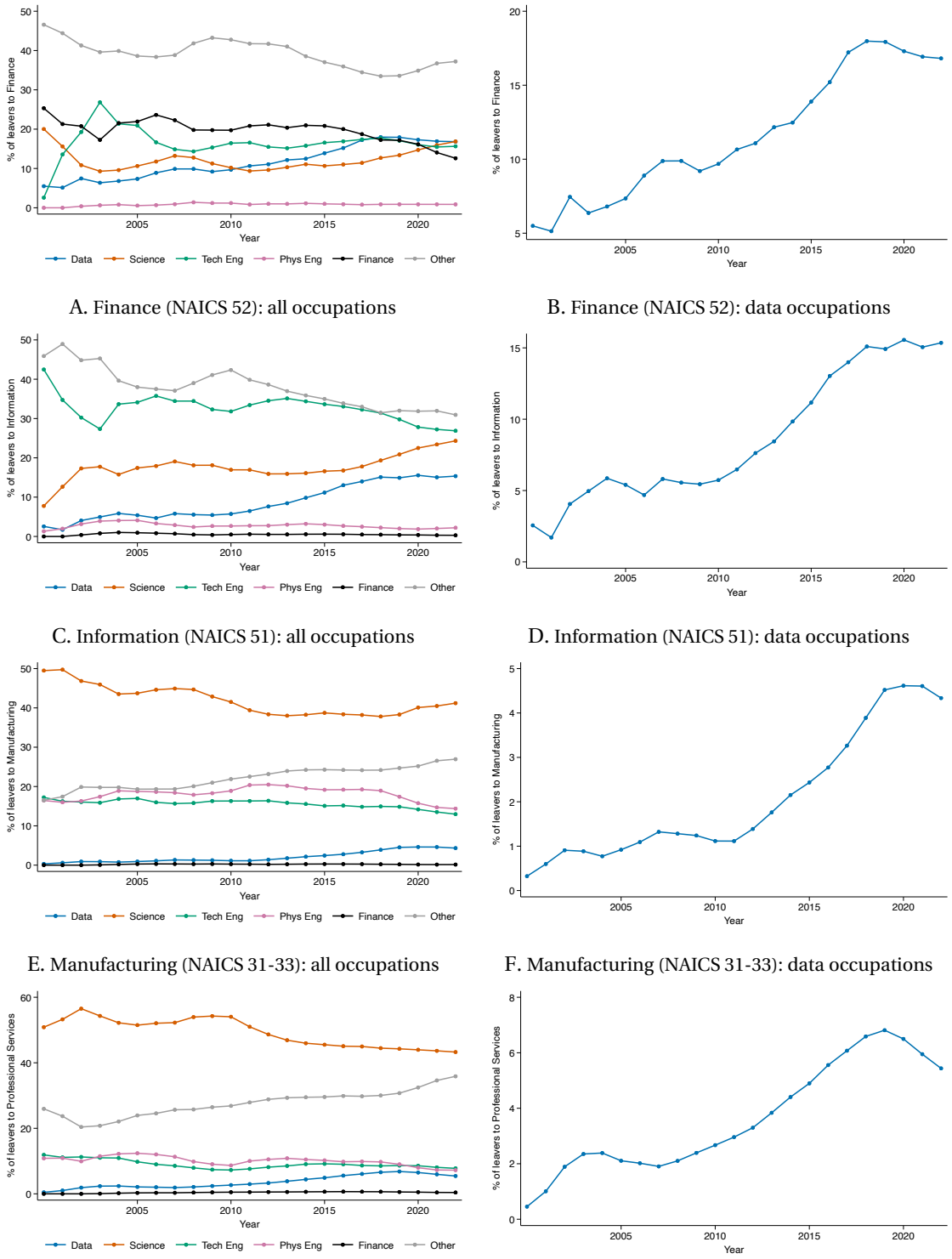
B. Information (NAICS 51)

C. Manufacturing (NAICS 31-33)

D. Professional Services (NAICS 54)

FIGURE A7. Growth in occupation mix of university-science leavers by destination industry

This figure plots the growth in the occupation mix of workers leaving university science - defined as positions in NAICS 61 (Educational Services) held by individuals in science occupations - within each of the four largest destination industries: Finance (NAICS 52), Information (NAICS 51), Manufacturing (NAICS 31-33), and Professional Services (NAICS 54). For each destination industry and each occupation category (data, science, technical engineering, physical engineering, finance, or other), I compute that occupation's share of university-science leavers entering the industry in each year, apply a 3-year centered moving average, and index the resulting series to 2005 = 100. Occupation categories are based on Revelio Labs' standardized occupation codes. Appendix Section B.2 provides detailed definitions. The sample period is 2000-2022. Go back to main text.




A. Finance (NAICS 52): all occupations

B. Finance (NAICS 52): data occupations

C. Information (NAICS 51): all occupations

D. Information (NAICS 51): data occupations

E. Manufacturing (NAICS 31-33): all occupations

F. Manufacturing (NAICS 31-33): data occupations

G. Professional Services (NAICS 54): all occupations

H. Professional Services (NAICS 54): data occupations

FIGURE A8. Occupation mix of university-science leavers by destination industry in levels

This figure plots the occupation mix of workers leaving university science - defined as positions in NAICS 61 (Educational Services) held by individuals in science occupations - within each of the four largest destination industries: Finance (NAICS 52), Information (NAICS 51), Manufacturing (NAICS 31-33), and Professional Services (NAICS 54). For each destination industry, the left panel shows the share of university-science leavers entering each occupation category (data, science, technical engineering, physical engineering, finance, or other) as a percentage of total leavers to that industry, smoothed with a 3-year centered moving average. The right panel isolates the share of university-science leavers entering data occupations in that industry. Occupation categories are based on Revelio Labs' standardized occupation codes. Appendix Section B.2 provides detailed definitions. The sample period is 2000-2022. Go back to main text.

## Appendix References


Autor, David H, and David Dorn. 2013. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103 (5): 1553–1597.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica* 73 (6): 1815–1847.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark et al. 2022. "Training Compute-Optimal Large Language Models." *Working Paper*.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." *Working Paper*.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2025. "IPUMS USA: Version 16.0 [dataset]." *Website*.