Creative Destruction in a Data Economy*

Erwan Morellec† Francesca Zucchi‡

June 9, 2026

Abstract

We study how data and artificial intelligence (AI) shape innovation, firm dynamics, and growth. We develop a Schumpeterian growth model in which incumbent firms generate proprietary data through production, while entrants must acquire data from incumbents or source it independently. Data increases R&D productivity when combined with computing capabilities, creating complementarities between production, innovation, and AI investment that give rise to larger and more innovative incumbents. Although data can act as a barrier to entry by reducing the number of entrants, it may simultaneously increase creative destruction by raising entrants' innovation intensity. Thus, in a data economy, entry and creative destruction need not move together. The overall impact of data and AI on growth and welfare depend on how entrants gain access to data.

Keywords: Data economy, creative destruction, AI, innovation, endogenous growth.
JEL Classifications: E22; G31; L22; O30; O40.

---

*We are grateful to Thierry Foucault and Peter Hoffmann for insightful conversations. The views expressed are those of the authors and should not be interpreted as representing the European Central Bank, the Eurosystem, or their staff.

†EPFL, CEPR, and Swiss Finance Institute. Email: erwan.morellec@epfl.ch.
‡European Central Bank and CEPR. Email: francesca.zucchi@ecb.europa.eu.

The growing importance of data and artificial intelligence (AI) is reshaping innovation and competition. Firms increasingly rely on proprietary data to improve prediction, automate decision-making, and accelerate research and development (R&D), while advances in AI expand the economic value that can be extracted from large and complex datasets (Varian, 2019; Cockburn, Henderson, and Stern, 2019). As a result, access to data has become a central determinant of innovative capacity and competitive advantage. A defining feature of the data economy is that incumbent firms naturally generate vast quantities of proprietary data through production and customer interactions, whereas entrants lack comparable access. Data therefore emerges simultaneously as a productive input into innovation and as a potential source of market power creating endogenous barriers to entry (Farboodi et al., 2019; Jones and Tonetti, 2020).

These developments raise a fundamental question: does the data economy foster innovation and creative destruction, or does it merely entrench incumbents and stifle market dynamism? On one hand, proprietary data can build high barriers to entry, as incumbents naturally accumulate data through ongoing production and customer interactions. Growing concerns over market concentration and the rise of "superstar" firms in data-intensive sectors (Autor et al., 2020; Akcigit and Ates, 2021; Farboodi et al., 2019) suggest that data advantages may favor incumbents over new entrants, ultimately weakening competitive pressure. On the other hand, data boosts R&D productivity—especially when paired with AI technologies that accelerate experimentation and the recombination of ideas (Cockburn, Henderson, and Stern, 2019; Agrawal, Gans, and Goldfarb, 2019; Bontadini et al., 2026; Acharya, Shao, and Xu, 2026).[1] When data directly enhances innovative capacity, the link between entry and creative destruction becomes more nuanced; a decline in new entrants does not necessarily mean weaker creative destruction if access to data substantially increases the innovation intensity of active firms.

This paper develops a dynamic model of innovation and endogenous growth to study how data access and AI-enabled data processing affect firm behavior, market structure, creative

[1]For example, using data on granted patents filed by publicly listed U.S. firms to measure innovation outcomes, Acharya, Shao, and Xu (2026) find that AI-adopting firms experience subsequent increases in both the quantity and quality of their patents relative to non-adopting firms. Furthermore, AI-adopting firms produce patents that exhibit greater originality and span more technologically distant fields, suggesting that AI enables firms to explore broader and less familiar technological domains.

destruction, and welfare. We extend a standard quality-ladder framework by introducing data as an endogenous byproduct of production that increases the productivity of innovation when combined with investment in computing capabilities (Agarwal, McHale, and Oettl, 2019; Restrepo, 2025; Lane, 2026).[2] Incumbents generate data through production and can exploit it directly in their innovation activities. Entrants, by contrast, do not produce and therefore must either acquire data from incumbents (also referred to as *data trading*) or invest in alternative *data sourcing technologies* such as web scraping or synthetic data generation. The model therefore captures a central feature of the data economy: data simultaneously acts as a productive input into innovation and as a potential barrier to entry.

The model reveals the existence of complementarities between production, data accumulation, AI investment, and innovation. Incumbent firms that produce at a larger scale generate more data. Data, in turn, raises the productivity of innovation, and higher engagement in innovation increases the returns to investment in AI and computing capabilities. These complementarities create a self-reinforcing mechanism through which successful incumbent firms expand production, innovate more, and further strengthen their data advantage. The model therefore provides a micro-foundation for the emergence of large and highly innovative firms in data-intensive industries.

Our main result is that the effect of data on business dynamism depends on the interaction between the extensive and intensive margins of innovation. On the one hand, data can reduce entry by strengthening incumbents' competitive advantages and raising barriers to entry. On the other hand, when entrants obtain access to data through trading or sourcing, data increases the productivity and intensity of their innovation efforts. As a result, the overall rate of creative destruction may increase even if the equilibrium mass of entrants declines. In the data economy, entry and creative destruction therefore need not move together.

The mechanism through which entrants access data further shapes equilibrium outcomes. Under data trading, incumbents internalize the value of selling data and retain partial control over access conditions. Data trading can therefore stimulate innovation on both sides of

[2] While the application to pharmaceuticals is perhaps the most famous (see the example of AtomWise)—harnessing AI and data to speed up drug discovery—AI-driven innovation is transforming sectors as diverse as materials science (Citrine Informatics, Kebotix), agriculture (Plenty, Benson Hill), or energy (Fervo Energy), just to name some examples. In each case, AI reduces the experimentation cycle and expands the feasible search space, then easing the attainment of technological breakthroughs.

the market while preserving incumbent advantages. By contrast, when entrants rely on alternative sourcing technologies, incumbents face greater competitive pressure because they cannot restrict entrants' access to data. Consequently, these alternative access regimes affect not only the aggregate growth rate but also the relative contributions of incumbents and entrants to economic growth. The mode of data access is also central to welfare outcomes. Alternative data sourcing effectively duplicates incumbents' data-generation efforts and may therefore generate lower welfare. By contrast, data trading exploits the non-rival nature of data, enabling its more efficient use across firms. The model thus highlights that welfare depends not only on the quantity of available data but also on the institutional arrangements governing data access, sharing, and portability.

We further show that allowing for mergers between incumbents and successful entrants changes the nature of creative destruction itself. In standard Schumpeterian environments, successful innovation displaces incumbents. In practice, however, incumbents may instead acquire innovative entrants and preserve market leadership despite continued technological progress and active startup formation. Hence, the framework reconciles the coexistence of rapid technological progress and dynamic entrepreneurship with the persistent market dominance and high acquisition volumes seen in data-intensive markets.

More broadly, the paper provides a different perspective on the relationship between concentration and innovation in the data economy. Data can simultaneously stimulate innovation and reinforce incumbent dominance because it affects not only market access, but also the productivity of innovation itself. Policies governing data access, portability, and mergers may therefore have complex effects on innovation and market structure that depend critically on how data flows between incumbents and entrants.

Literature review This paper contributes to the literature on innovation, firm dynamics, and endogenous growth. It complements existing work on data as an economic input (Varian, 2019; Jones and Tonetti, 2020) by endogenizing both the access to data and its role in the innovation process. More broadly, it speaks to ongoing policy debates on competition in digital markets by showing that policies restricting or facilitating data access can have complex and sometimes countervailing effects on innovation and welfare.

The paper also contributes to the literature on innovation and endogenous growth by introducing data and AI-enabled processing as determinants of innovation efficiency and firm dynamics. Existing work studies how financial frictions, industry structure, mergers, and market power shape innovation incentives, including financing frictions (Malamud and Zucchi, 2019; Brown, Fazzari, and Petersen, 2009), debt financing (Geelen, Hajda, and Morellec, 2022), soft budget constraints (Aghion et al., 2025), and aggregate discount rates (Bustamante, D'Acunto, and Zucchi, 2026). We complement this literature by emphasizing how access to data alters both the organization and the productivity of innovation.

Third, the paper relates to the growing literature on AI and data as drivers of technological change. Prior work argues that AI functions as a "method of invention" that accelerates the discovery process and reduces the cost of experimentation (Cockburn, Henderson, and Stern, 2019; Agarwal, McHale, and Oettl, 2019; Bontadini et al., 2026). Emerging empirical evidence suggests that AI adoption increases innovation and firm performance (Babina et al., 2024; Acharya, Shao, and Xu, 2026; Aldasoro et al., 2025; Lebastard and Sondermann, 2026). We formalize these mechanisms within an endogenous growth framework and show how they reshape the allocation of innovation across incumbents and entrants.

Fourth, our analysis contributes to the literature on data as an economic input. Jones and Tonetti (2020) emphasize the non-rival nature of data and study the trade-off between data usage and privacy, while Cong, Xie, and Zhang (2021) examine privacy considerations in a growth framework with data. Farboodi and Veldkamp (2026) show that data improves productivity through better prediction but argue that sustained growth requires data to affect innovation directly. We build on these insights by endogenizing both access to data and its role in the innovation process. More broadly, our paper relates to recent work studying data-driven learning and automation (Farboodi, Koh, and Xia, 2025) and complementarities between data capabilities and innovation (Fedyk et al., 2026).

Section I presents the model. Section II characterizes firm policies, creative destruction, and growth in a data economy. Section III analyzes how data trading and data sourcing affect entry and creative destruction. Section IV studies welfare implications. Section V introduces mergers between incumbents and successful entrants. Section VI concludes. Technical developments are gathered in the Appendix.

I The model

We develop a continuous-time, infinite-horizon equilibrium model of endogenous growth in a data economy. The economy admits a representative agent with logarithmic preferences and discount rate ρ:

$$E\left[\int_0^\infty e^{-\rho t}\ln\mathcal{C}_t dt\right]$$

where $\mathcal{C}_t$ denotes consumption at time t. The representative agent is endowed with one unit of labor, which is supplied inelastically and yields a competitive wage denoted by W_t. Population size is denoted by L, which we normalize to one.

A Consumption good sector

There is a unique consumption good (or final good) that serves as the numéraire of the economy, whose output is denoted by $\mathcal{X}_t$. This final good is produced competitively using a continuum of inputs $j \in [0,1]$ and labor through the following technology:

$$\mathcal{X}_t = \frac{1}{1-\beta}\int_0^1 L^\beta Y_{jt}^{1-\beta} q_{jt}^\beta dj.$$

In this equation, Y_{jt} denotes the quantity of a given input and q_{jt} is the associated quality at time t. Initial quality q_{j0} is normalized to one for all inputs. Improvements in quality follow from innovation. As standard in quality ladder models, only the highest-quality ("frontier") version of each input is used by the consumption good sector. Because population is normalized to one, so are labor units. We thus omit the parameter L in the rest of the paper.

B Input sector

The input sector is populated by two types of firms, incumbents and entrants. There is a continuum of each type of firms, with an endogenous mass m_E of entrants.

Incumbent Each incumbent produces one type of input and invests in innovation to keep improving its quality, with the objective of maximizing its market value. Following the literature, the latest innovator in a given product line j becomes the monopolist of that product line. Each incumbent takes the demand schedule of the consumption good sector as given and optimally sets the production rate for input j and pins down the associated selling price. These quantities are denoted respectively by Y_{jt} and p_{jt}, which are therefore endogenously derived. The marginal cost of production is normalized to 1.

As in Jones and Tonetti (2020) and Farboodi and Veldkamp (2026), production generates data as a byproduct. For a given production rate, a flow of raw data

$$D_{jt} = \psi Y_{jt}$$

is generated. The parameter $\psi \geq 0$ can be interpreted as the degree of data intensity of the industry—for instance, a digital service firm would create more data than a construction firm. This parameter can also be inversely related to the degree of privacy enforcement in the economy—i.e., if $\psi = 0$, privacy enforcement is the tightest and no data is generated.

We denote by z_{jt} the innovation intensity of incumbent j at time t. Innovation is costly, and its outcome is uncertain. Namely, if the firm spends the flow cost:[3]

$$-\zeta \frac{z_{jt}^2}{2} q_{jt}, \quad \zeta > 0,$$

it increases the quality of input j at Poisson rate ϕz_{jt} when abstracting from the role of data and AI thereof (more details in the paragraph below). These Poisson events represent technological breakthroughs. When a breakthrough happens, input quality increases from q_{jt-} to $q_{jt} = \lambda q_{jt-}$, with the parameter $\lambda > 1$ representing the size of the quality improvement.

The novelty of our setting is that data supports innovation by reducing the runway towards attaining a breakthrough, consistent with Agarwal, McHale, and Oettl (2019) and Cockburn, Henderson, and Stern (2019). Namely, data supports the innovation process if the firm invests in computing—i.e., AI—then turning the raw data into processed data. In the following, we use AI, data processing, and computing capabilities interchangeably. To support a given level a_{jt} of computing capabilities, the firm bears the flow cost $\alpha \frac{a_{jt}^2}{2} q_{jt}$ with

[3]The fact that the cost of innovation increases with q_{jt} is consistent with the innovation literature, and the idea that improvements are harder if the technology has higher quality.

$\alpha > 0$, consistent with the evidence that AI and data processing abilities require recurrent investment in infrastructure and maintenance. For a given innovation rate z_{jt}, the combination of data and computing capabilities increases the arrival rate of a technological breakthrough from ϕz_{jt} to

$$\left(\phi + \frac{D_{jt} a_{jt}^{\xi}}{q_{jt}}\right) z_{jt}.$$

That is, raw data together with computing capabilities boost the probability of a breakthrough.[4] We adopt a conservative approach and assume that the higher the quality of the technology, the lower the incremental ability of new raw data in pushing the frontier further. Moreover, we assume that data processing can exhibit returns other than linear through the parameter $\xi > 0$ (as a baseline, we assume $\xi \in (0, 1)$). Importantly, raw data D_{jt} and computing capabilities a_{jt} in isolation have no impact on innovation.

Incumbents can sell their data D_{jt} to entrants. As data is non-rival, selling it to entrants does not prevent the incumbent from exploiting it too. Selling data results in greater revenues from operations, increasing further incentives to invest and innovate and creating a positive feedback loop. Yet, when deciding on whether to sell its data, the incumbent balances the direct revenue from data sale against the resulting potential competitive advantage acquired by entrants (described in the next paragraph). Weighting these considerations, the incumbent endogenously set the price of data, denoted by θ_{jt}, to maximize its value.

An incumbent j is subject to creative destruction and exits if an entrant innovates on product line j. We denote the endogenous rate of creative destruction by x_d. When creative destruction hits, the incumbent recovers a fraction ω of its value and exits.

Entrants A mass of entrants invests in innovation with the aim of displacing incumbents if they achieve technological breakthroughs that improve the quality of existing inputs. Since entrants engage only in innovation activities and do not produce goods themselves, they can

[4]This specification captures the idea discussed e.g. in Lane (2026) that what "distinguishes AI from earlier revolutionary technologies is its scope [...] AI has the potential to raise the productivity of the innovation process itself. AI systems can meaningfully accelerate scientific discovery, shorten research and development cycles, and compress the time between knowledge creation and commercial application. The technology is set to not just shift the level of productive capacity but shift the rate at which productive capacity grows."

be interpreted as startups. The mass of entrants is endogenous and denoted by m_E.

Because entrants seek to innovate inputs they do not currently lead, their innovation efforts have wide breadth and uncertain applications. To capture this uncertainty, we follow Akcigit and Kerr (2018) and Malamud and Zucchi (2019) and assume that entrants' R&D investment result in breakthroughs in any industry j with equal probability. In expectation, entrant breakthroughs lead to an increase in average input quality, denoted by

$$\bar{q}_t = \int_0^1 q_{jt} dj.$$

For an entrant innovation rate by z_E, the innovation cost then depends on average quality

$$\zeta_E \frac{z_E^2}{2} \bar{q}_t,$$

where we allow the cost coefficient ζ_E to differ from that borne by the incumbents (ζ), as in previous models. As for the incumbents, the innovation rate regulates the Poisson rate of technological breakthroughs, which is ϕz_E on any time interval absent the entrants' use of data to support the innovation process. When an entrant achieves a breakthrough, uncertainty dissipates and it becomes clear which product lines j the successful entrant takes over. Thus, the quality of input j jumps from q_{jt-} to $q_{jt} = \Lambda q_{jt-}$, with the parameter $\Lambda > 1$ representing the size of quality improvement.

Entrants do not produce. Thus, to harness data in the innovation process, they can use two approaches. First, they can purchase the incumbents' data D_t. Second, they can invest in data sourcing, relying on AI technology.

Data trading. Consider the first approach, referred to as data trading. We assume that the entrant can purchase a share $f_{jt} \in [0, 1]$ of an incumbent's data D_{jt} in product line j, where f_{jt} is an endogenous choice. The transaction entails a proportional cost θ_{jt}, endogenously set by the incumbent. Consistent with the broad scope and uncertain ex ante applications of entrants' innovations, entrants can acquire data across different product lines. However, as with incumbents, access to raw data alone is insufficient to support the innovation process; entrants must also invest in data-processing and computing capabilities. Like incumbents, entrants incur a flow cost $\alpha \frac{a_{Et}^2}{2} \bar{q}_t$ when investing in data processing. For a given innovation rate z_{Et}, the flow of data and the computing capabilities lead to an increase in the arrival rate of a breakthrough from ϕz_{Et} to

$$\left(\phi + a_{Et}^{\xi} \int_0^1 \frac{f_{jt} D_{jt}}{q_{jt}} dj \right) z_{Et} = \left(\phi + a_{Et}^{\xi} \frac{D_{Et}}{\bar{q}_t} \right) z_{Et} \tag{1}$$

where $D_{Et} = \int_0^1 f_{jt} D_{jt} dj$ represents the amount of data purchased across different product lines. As for the incumbent, we consider that the higher the technological frontier, the harder it is for existing data to enable a breakthrough. Moreover, we allow for returns to data processing other than linear through the parameter $\xi \geq 0$. As with incumbents, neither raw data nor computing capabilities alone affect the probability of achieving a breakthrough; however, when combined, they increase the likelihood of attaining one.

Data sourcing. When direct data trading is infeasible—either because incumbents do not want to share their proprietary data or because the price of data access exceeds the expected gains for entrants—firms can instead rely on data sourcing, denoted by L_{Ejt}. This encompasses a range of strategies through which entrants independently obtain or construct the data necessary to compete with incumbents.

Data sourcing can take several distinct forms. First, data extraction involves collecting information from existing, often publicly accessible, digital environments. A common example is web scraping, where firms deploy automated tools—frequently powered by advanced AI techniques—to systematically gather large volumes of data from online platforms. Second, entrants may engage in synthetic data generation, whereby data are artificially created rather than directly observed. This can occur through traditional methods such as surveys or panels, but increasingly involves the use of AI-driven models to simulate realistic datasets. Firms may also design engineering processes or controlled environments that produce data tailored to their specific analytical needs, thereby reducing reliance on external sources.[5] Finally, alternative data refers to non-traditional datasets that can serve as proxies for variables of interest. Examples include satellite imagery used to approximate economic activity,

[5]A relevant example comes from Jones and Tonetti (2020) in the context of innovation in the self-driving space, whereby firms might set up artificial towns to collect driving data, and then process it through AI.

mobility data to infer consumer behavior, or environmental indicators such as CO2 emissions to estimate industrial output. These unconventional sources can provide valuable insights, particularly when standard data are unavailable or prohibitively costly.

Together, these data sourcing strategies allow entrants to circumvent barriers imposed by incumbents' control over proprietary data, albeit often at the cost of increased technological investment and methodological complexity.

Data sourcing and processing are closely intertwined as they both rely heavily on the A.I. technology. We thus take a parsimonious approach and capture all related expenditures through the investment margin I_{Ejt}. As entrants are ex-ante uncertain as to which production line will manage to improve on, data sourcing spans multiple product lines j and involves the cost:

$$\frac{\delta}{2}\int_0^1 \frac{L_{Ejt}^2}{(1+\psi)q_{jt}}dj$$

across the different product lines j, with $\delta > 0$ being a cost coefficient. This cost function is purposely defined to have the same form as the cost functions related to raw data, in order to focus comparison between the relative costs ζ, α, and δ. This formulation implies that the greater the average quality of inputs, the more data is available for web scraping, then lowering the cost of investment in data extraction (as captured by the term q_{jt} at the denominator).[6] If a sector is more data-intensive or privacy is less enforced, the cheaper data sourcing is, as captured by the parameter ψ at the denominator. When entrants invest in data extraction, the likelihood of attaining a breakthrough increases from ϕz_{Et} to:

$$\left(\phi + \int_0^1 \left(\frac{\gamma L_{Ejt}}{q_{jt}}\right)^{\mu} dj\right) z_{Et}. \tag{2}$$

where $\gamma \in [0, 1]$ captures the quality of sourced data. As for the usage of actual raw data, we continue to make the conservative assumption that the greater the average quality of inputs, the harder it is for synthetic data to increase the probability of attaining a breakthrough (recall that I_E captures both the investment in data creation and processing through AI

[6]E.g., richer data is available for web scraping, or more information is available for setting up an infrastructure for synthetic data generation.

technology). We also leave the model flexible and allow for returns to scale other than linear through the parameter $\mu \geq 0$. Lastly, the specification in equation (2) mirrors that in equation (1) to facilitate comparisons.

At the outset, entrants need to incur the entry cost K_t before being able to start investing in innovation. As previous contributions, we assume that the entry cost increases with the average technological frontier, i.e., $K_t = \kappa \bar{q}_t$. The industry is competitive, in that the free-entry condition is binding at any time.

C Balanced growth path

We focus on a balanced growth path equilibrium, in which aggregate quantities grow at a constant endogenous rate g, shaped by the innovative efforts of incumbents and entrants. An equilibrium is an allocation whereby

- Incumbents set production, innovation, and investment in data processing to maximize their value;
- Entrants set innovation, data acquisition from the incumbents, and investment in data processing, or data extraction to maximize their value;
- The free-entry condition is binding, determining the mass of active entrants m_E;
- The consumption good sector maximizes profits;
- The representative agent maximizes utility from consumption; all markets clear.

D Discussion of the assumption

Our model is grounded in the idea that the growing availability of raw data—combined with AI-enabled analysis—is fundamentally reshaping R&D by shortening the path to breakthrough discoveries. Although pharmaceuticals provide the most prominent example, where AI and large datasets are accelerating drug discovery, this transformation extends far beyond that field. AI-driven innovation is influencing areas as varied as materials science, agriculture, energy, and engineering, just to name a few. In each of these domains, AI reduces the time required for experimentation while simultaneously broadening the range of possibilities that can be explored, thereby making technological breakthroughs more attainable.[7]

Our model is flexible in that it accounts for the two main methods through which entrants can gain access to data, i.e., data acquisition from incumbents or data sourcing harnessing AI technologies. Irrespective of the method, our modeling emphasizes that entrants engage with multiple product lines. As noted by Agarwal, McHale, and Oettl (2019), mixing different technologies is key for innovation, as reminded among others by Romer (1993): “to appreciate the potential for discovery, one needs to consider the possibility that an extremely small fraction of the large number of potential mixtures may be valuable.” Deep learning indeed eases the selection of promising “potential mixtures” by crunching big data, as also posited by the *Nature* article by LeCun, Bengio, and Hinton (2015): “Deep learning [...] has turned out to be very good at discovering intricate structure in high-dimensional data and is therefore applicable to many domains of science, business and government.”[8]

As in Jones and Tonetti (2020) or Cong, Xie, and Zhang (2021), we consider the flow of data as opposed to the stock, as it is the fresh, incremental data that actually provide new information. This modeling is subject to two interpretations: (1) old data has already been processed and learned from, so it is the new, fresh data that matter; or (2) data depreciates fully every period, as today’s data will supersede yesterday’s.

II Firm policies, creative destruction, and growth

We solve the model in the general case which encompasses the incumbent’s data monopoly (when only incumbents harness their proprietary data from production), the data trading equilibrium (in which entrants acquire the incumbents’ data), and the data sourcing equilibrium (in which entrants engage in data sourcing). We use the superscript $i \in \{M, T, S\}$ to denote the endogenous quantities in these three settings.

[7]Recent research by Acharya, Shao, and Xu (2026) provides large sample evidence supporting the view that AI affects productivity via the innovation channel.

[8]Agarwal, McHale, and Oettl (2019) report ”One reason to be hopeful about the future is the recent explosion of data availability under the rubric of big data and computer-based advanced in capabilities to discover and process those data. We can view these technologies as meta technologies, i.e. technologies for the production of new knowledge” and report the Atomwise example.

A Incumbent

We start by solving incumbent's optimization problem, which consists in choosing how much to produce, invest in innovation, and invest in computing capabilities. The maximization of the final good sector gives the following demand function for each frontier input j:

$$Y_{jt} = \left(q_{jt}^{\beta}/p_{jt}\right)^{\frac{1}{\beta}} \tag{3}$$

which enters the maximization problem of the incumbents. Because of symmetry across product lines, we drop the subscript j in the following when it does not create confusion.

Standard derivations imply that the value of an incumbent firm solves the Hamilton-Jacobi-Bellman (HJB) equation:

$$\begin{aligned} r^i V^i(q) = \max_{Y^i, z^i, a^i} \Bigg\{ & Y^i\left(p^i - 1\right) + m_E^i \theta f D^i 1_{\{i=T\}} - \alpha \frac{(a^i)^2}{2} q - \zeta \frac{(z^i)^2}{2} q \\ & + z^i \left(\phi + \frac{D^i (a^i)^{\xi}}{q}\right) \left[V^i(\lambda q) - V^i(q)\right] + x_d^i(\omega - 1) V^i(q) \Bigg\}, \end{aligned} \tag{4}$$

where $1_{\{i=T\}} = 1$ if $i = T$ and $1_{\{i=T\}} = 0$ otherwise. The left-hand side of this equation is the return required by the investors, with the equilibrium interest rate r^i being endogenously determined in equilibrium. The right-hand side is the expected change in the value of the incumbent on each time interval. The first term on the right-hand side represents the revenues from production net of the associated cost. The second term represents the revenues from the sale of data to entrants, which occurs in the data trading setting. The greater the mass of active entrants m_E^T, the more interested parties an incumbent can sell its data to: Being non-rival, data can be sold multiple times. The third term captures the investment in data processing. The fourth term represents the investment in innovation. Lastly, the fifth term represents the probability-weighted impact of a technological breakthrough. The term $z^i\left(\phi + \frac{D^i(a^i)^{\xi}}{q}\right)$ represents the endogenous rate of arrival of technological breakthroughs—increasing not only with the incumbent's innovation rate but also with AI-processed data—whereas the term $[V^i(\lambda q) - V^i(q)]$ represents the surplus created upon attaining a breakthrough, in which case input quality increases by λ. The last term represents the impact of creative destruction, driven by entrants' breakthroughs. When creative destruction hits, the firm only recovers a fraction ω of its value and exits.

To solve this equation, we conjecture and verify that the value of an incumbent scales with q_t, i.e., $V^i(q_t) = q_t v^i$. We also define the scale of production normalized by quality as $y^i = Y_t^i / q_t$. Maximizing the resulting scaled HJB equation gives the optimal firm policies (details are reported in the Appendices).

Proposition 1. *In a data economy, the incumbent's optimal production rate y^i, innovation rate z^i, and investment in computing capabilities a^i satisfy*

$$y^i = \left(\frac{1-\beta}{1 - m_E^i f \theta \psi 1_{\{i=T\}} - (a^i)^\xi \psi z^i (\lambda - 1) v^i} \right)^{\frac{1}{\beta}}, \tag{5}$$

$$z^i = \frac{\left[\phi + (a^i)^\xi \psi y^i\right] (\lambda - 1) v^i}{\zeta}, \tag{6}$$

$$a^i = \left(\frac{\xi \psi y^i z^i (\lambda - 1) v^i}{\alpha} \right)^{\frac{1}{2-\xi}}.$$

Equation (5) shows that in a data economy, the optimal production rate is not only driven by the demand of the final good sector, but also by how much production translates into data that can support the innovation process as well as the entrants' demand for data. The greater the investment in innovation z^i or the investment in data processing capabilities a^i, the more data helps the innovation process—and, thus, the greater the firm production rate. Our model then highlights complementarities among production, innovation, and investment in computing capabilities. In addition, the higher the revenues from selling data, the higher the benefits from producing, and the greater firm size.

The incumbent's optimal innovation rate z^i increases with the surplus upon attaining a breakthrough (the term $(\lambda - 1)v^i$) and decreases with the cost parameter ζ. When data supports the innovation process, the optimal innovation rate further increases with raw data and computing capabilities, ψy^i and a^i. Last, optimal investment in computing capabilities a^i positively depends on how much production translates into raw data as well as how innovation-intensive the firm is.

The incumbents' optimal policies then reveal that the scale of production y^i, the investment in computing a^i, and the innovation rate z^i are interconnected and complementary. All else equal, a greater scale of production produces more raw data as a byproduct, which can be processed to further support the innovation process. This is more effective if the firm exhibits greater investment in computing or if it is more R&D intensive, in turn leading to a greater production rate. Notably, the emergence of larger firms (and "superstar firms") arises naturally in a data economy, as a greater scale of production begets more innovation and AI investment, which further expands the scale of production. A second channel through which data can lead to superstars is through the revenues from selling data, which increase the benefits of operating at a larger scale.

B Entrant

The value of entrants satisfies the Hamilton-Jacobi-Bellman equation:

$$
\begin{aligned}
r^i V_E^i(\bar{q}) = \max_{\{z_E^i, f, a_E^i, L_E\}} \Bigg\{ & V_{Et}^i - \zeta_E \frac{(z_E^i)^2}{2}\bar{q} + z_E^i \phi \left[V^i(\Lambda\bar{q}) - V_E^i\right] \\
& + 1_{\{i=T\}} \left[z_E^i (a_E^i)^\xi \int_0^1 \frac{f_j D_j}{q_j} dj \left[V^i(\Lambda\bar{q}) - V_E^i\right] - \left(\int_0^1 \theta_j f_j D_j dj + \alpha \frac{(a_E^i)^2}{2}\bar{q} \right) \right] \\
& + 1_{\{i=S\}} \left[z_E^i \int_0^1 \left(\frac{\gamma L_{Ej}}{q_j} \right)^\mu dj \left[V^i(\Lambda\bar{q}) - V_E^i\right] - \frac{\delta}{2} \int_0^1 \frac{L_{Ej}^2}{(1+\psi) q_j} dj \right] \Bigg\}
\end{aligned} \tag{7}
$$

The left-hand side of equation (7) is the return required by investors. The first term on the right-hand side is the change in value on a small time interval. The second and third terms capture the cost and benefits of innovation when entrants have no access to data. In particular, the third term is the probability-weighted impact of a technological breakthrough, in which case the successful entrant takes over an incumbent's market position and $V^i(\Lambda\bar{q}) - V_E^i$ is the associated surplus.[9] The second line captures the incremental effect of the entrants' purchase of the incumbent data in product line j, coupled with investment in computing. Namely, the first term on this second line is the incremental probability-weighted impact of a technological breakthrough—where the likelihood is affected by the amount of data acquired from incumbents and by investment in compute—whereas the second term is the cost of acquiring incumbents' data and the cost of investing in data processing. The third

[9] As in Malamud and Zucchi (2019), entrants' breakthroughs have uncertain applications, so the change in value satisfies $\int_0^1 V^M(\Lambda q_{jt})dj = V^M(\Lambda\bar{q})$.

line captures the effects of data sourcing and admits a similar interpretation, namely, the first term is the incremental probability-weighted impact of data sourcing on a technological breakthrough, whereas the second term is the associated cost.

To solve for the entrants' problem, we conjecture that the value of entrants scales with the average quality: $V_E^i(\bar{q}_t) = \bar{q}_t v_E^i$ (see also Malamud and Zucchi (2019)). Because of symmetry across product lines, the share of data purchased (if positive) is the same across product and referred to as f in the following. Maximizing the resulting scaled Hamilton-Jacobi-Bellman equation with respect to z_E^i, f, and a_E^i yields the following result.

Proposition 2. *Free entry requires that entrant value equals the entry cost, $v_E^i = \kappa$. In a data economy, the entrant's optimal innovation rate z_E^i, the fraction of data purchased f, investment in computing capabilities a_E^i, and investment in data sourcing satisfy:*

$$z_E^i = \frac{\left(\phi + 1_{\{i=T\}}(a_E^i)^\xi f\psi y^i + 1_{\{i=S\}}\ell_E^\mu \gamma^\mu\right)\left[v^i\Lambda - v_E^i\right]}{\zeta_E}, \tag{8}$$

$$f = 1_{\{i=T\}} 1_{\{z_E^T (a_E^T)^\xi \left[\Lambda v^T - v_E^T\right] > \theta\}},$$

$$a_E^i = 1_{\{i=T\}} \left(\frac{z_E^i f\psi y^i \left[v^i\Lambda - v_E^i\right]\xi}{\alpha}\right)^{\frac{1}{2-\xi}}$$

$$\ell_E = 1_{\{i=S\}} \left(\frac{z_E^S \left[v^S\Lambda - v_E^S\right]\mu(1+\psi)\gamma^\mu}{\delta}\right)^{\frac{1}{2-\mu}}. \tag{9}$$

Equation (8) shows that because an entrant becomes an incumbent once making a breakthrough, incumbent value v^i affects the entrants' innovation rate. In this equation, the numerator is the marginal benefit of increasing the innovation rate. The term $\zeta_E z_E^i$ is the marginal cost of doing so. Data access increases the likelihood of innovating, boosting the intensive margin of innovation for entrants.

Entrants buy data if their price is sufficiently low—i.e., lower than the probability-weighted gain associated with an improvement in the innovation process. The solution is bang-bang: If $z_E^T(a_E^T)^\xi\left[\Lambda v^T - v_E^T\right] > \theta$, entrants buy the entire flow of new data from incumbents, i.e. $f = 1$ and $1_{\{i=T\}} = 1$; if it does not hold, entrants do not purchase the incumbents' data (so $f = 0$ and $1_{\{i=T\}} = 0$). If $f = 0$, entrants do not invest in data processing capabilities ($a_E^T = 0$). That is, the entrants' innovation rate z_E^T and investment in data processing a_E^T

are complementary if entrants engage in data acquisition $f > 0$ and $1_{\{i=T\}} = 1$: The greater the investment in computing, the greater the innovation rate, and vice versa.

Lastly, equation (9) shows that the optimal level of data sourcing equates the marginal benefits of data sourcing to its marginal cost.

C Aggregate quantities

The rate of creative destruction represents the likelihood that entrants, on aggregate, attain a technological breakthrough. Thus, aggregating optimal innovation rates across the mass of entrants m_E^i gives the entrants' aggregate innovation rate or, equivalently, the rate of creative destruction

$$x_d^i = m_E^i \left[\phi + 1_{\{i=T\}}(a_E^T)^\xi f \psi y^T + 1_{\{i=S\}} \ell_E^\mu \gamma^\mu\right] z_E^i.$$

This expression highlights two endogenous margins of creative destruction. The extensive margin is given by the mass of active entrants, (m_E^i), while the intensive margin is given by the innovation arrival rate of an individual entrant, $\left[\phi + 1_{\{i=T\}}(a_E^T)^\xi f \psi y^T + 1_{\{i=S\}} \ell_E^\mu \gamma^\mu\right] z_E^i$—both of which are endogenous. Changes in the rate of creative destruction may therefore arise from variations in either the number of entrants or the innovation intensity of each entrant.

Economic growth is the sum of the incumbents' and the entrants' contributions

$$g^i = g_I^i + g_E^i.$$

Incumbents' contribution to growth g_I^i is the probability-weighted jump in quality due to breakthroughs by incumbents (whose mass is one):

$$g_I^i = (\lambda - 1) \left[\phi + (a^i)^\xi \psi y^i\right] z^i. \tag{10}$$

The term $(\lambda-1)$ represents the increase in quality following a breakthrough by an incumbent. The remaining terms capture the arrival rate of breakthroughs. In contrast to standard growth models, growth depends not only on the rate of innovation, but also on the rate of production. The reason is that production generates data, and the availability of data increases the likelihood of achieving future breakthroughs when combined with investment in data processing. Indeed, equation (10) further shows how investment in computing supports growth by enabling data processing that, in turn, raises the arrival rate of new technologies.

Similarly, the entrants' contribution to growth is the probability-weighted jump in quality due to breakthroughs by entrants

$$g_E^i = (\Lambda - 1) \underbrace{m_E^i \left[\phi + 1_{\{i=T\}} (a_E^T)^\xi f \psi y^T + 1_{\{i=S\}} \ell_E^\mu \gamma^\mu\right] z_E^i}_{x_d^i}. \tag{11}$$

The first term in parentheses captures the increase in quality associated with entrant breakthroughs, while the remaining terms reflect the arrival rate of such breakthroughs across the pool of potential entrants—that is, the rate of creative destruction. Entrants' access to incumbents' data directly affects the pace of creative destruction. Incumbents account for this effect when determining the equilibrium price, θ, which in turn influences the extensive margin, m_E^T. In particular, a higher value of θ raises barriers to data access and is therefore expected to reduce the mass of active entrants. As a result, the impact of data trading on the rate of creative destruction—which we further investigate in the next section—is ambiguous.[10] In the data sourcing equilibrium, the rate of creative destruction is further supported by entrants' investment in data sourcing.

The growth rate g^i enters the standard Euler equation associated with the households' maximization problem, which gives the equilibrium interest rate $r^i = \rho + g^i$.

D Benchmark: The no-data economy

To develop intuition and disentangle the economic strengths at play, we compare outcome variables in the data economy to those of a benchmark environment in which data do not matter, in the spirit of previous models of innovation and growth. We refer to the quantities associated with this case with the superscript *. In the no-data economy, the incumbents'

[10] Conversely, Jones and Tonetti (2020) assume that incumbents' data sharing with the entrants has a mechanical positive impact on the rate of creative destruction.

optimal scale of production and innovation rate are (see Appendix A)

$$y^* = (1-\beta)^{\frac{1}{\beta}}, \quad z^* = \frac{\phi(\lambda-1)v^*}{\zeta} \tag{12}$$

and investment in data processing is zero, i.e. $a^* = 0$. Absent data, there is no direct interaction between production and innovation decisions, as in previous quality ladder models.

In this setting, the optimal innovation rate of entrants does not depend on data either. Thus, the entrants' optimal innovation rate satisfies

$$z_E^* = \frac{\phi(\Lambda v^* - v_E^*)}{\zeta_E} \tag{13}$$

This equation shows that the optimal innovation rate equates the marginal cost of increasing the innovation rate $\zeta_E z_E^*$ to its marginal benefit $\phi(\Lambda v^* - v_E^*)$.

Economic growth continues to be the sum of the incumbents' and the entrants' contributions, denoted by g_I^* and g_E^* respectively, i.e. $g^* = g_I^* + g_E^*$. Entrants' contribution to growth satisfies $g_E^* = (\Lambda - 1)x_d^* = (\Lambda - 1)m_E \phi z_E^*$. Incumbents' contribution to growth is simply given by $g_I^* = (\lambda - 1)\phi z^*$, i.e., it depends on the incumbents' optimal innovation rate.

III Creative destruction and growth in a data economy

A Incumbents' data monopoly: Data as a barrier to entry

We start the analysis of the effects of data on growth by comparing optimal quantities in the incumbents' data monopoly and in the no-data economy, which allows to show how data and computing—used solely by the incumbents—affect firm-level and equilibrium quantities and effectively serve as a barrier to entry, entrenching incumbents and supporting their growth. In this data monopoly equilibrium, the value of incumbents satisfies

$$v^M = \frac{1}{\Lambda}\left[\kappa + \frac{\sqrt{2\zeta_E \kappa \rho}}{\phi}\right] \tag{14}$$

and we have the following result.

Proposition 3 (Data monopoly)**.** *Relative to the no-data economy, the incumbent data monopoly exhibits the following characteristics:*

1. *Incumbents increase their scale of production and R&D investment,* $y^M > y^*$ *and* $z^M > z^*$*, so they contribute to growth by a greater extent,* $g_I^M > g_I^*$*;*

2. *Entrants' innovation rate does not change,* $z_E^M = z_E^*$*. The mass of active entrants and the rate of creative destruction both decrease if incumbents' net profits decrease i.e.,* $m_E^M < m_E^*$ *and* $x_d^M < x_d^*$ *if*

$$y^M\left[(y^M)^{-\beta}-1\right]-\frac{\zeta}{2}(z^M)^2-\frac{\alpha}{2}(a^M)^2 < y^*\left[(y^*)^{-\beta}-1\right]-\frac{\zeta}{2}(z^*)^2.$$

When this is the case, entrants' contribution to growth decreases, i.e. $g_E^M < g_E^*$*.*

Compared to the no-data economy, incumbents unambiguously produce more in the data monopoly case. In fact, because production generates data as a byproduct, and data support the innovation process, incumbents find it optimal to increase their scale of production and generate more data. In addition, because production and innovation decisions are intertwined, incumbents also exhibit a greater innovation rate compared to the no-data economy. Being complementary, investment in data processing is also positive, $a^M > a^* = 0$. Overall, the greater investment in innovation, together with the efficiency push coming from data and AI investment, lead to an increase in the incumbents' contribution to growth:

$$g_I^M = (\lambda-1)\left[\phi+(a^M)^{\xi}\psi y^M\right]z^M > g_I^* = (\lambda-1)\phi z^*.$$

When entrants do not have access to data, their innovation rate z_E remains unchanged compared to the no-data economy. In equilibrium, incumbents' net profits are a key driver of the equilibrium mass of active entrants and, in turn, of the rate of creative destruction. If incumbents spend more resources on innovation and data processing, which drain profits, the equilibrium mass of entrants decreases. Because the rate of creative destruction satisfies $x_d^M = m_E^M \phi z_E^M$, it decreases too, and so does the entrants' contribution to growth.

To illustrate these effects, we calibrate the model parameters building on previous models

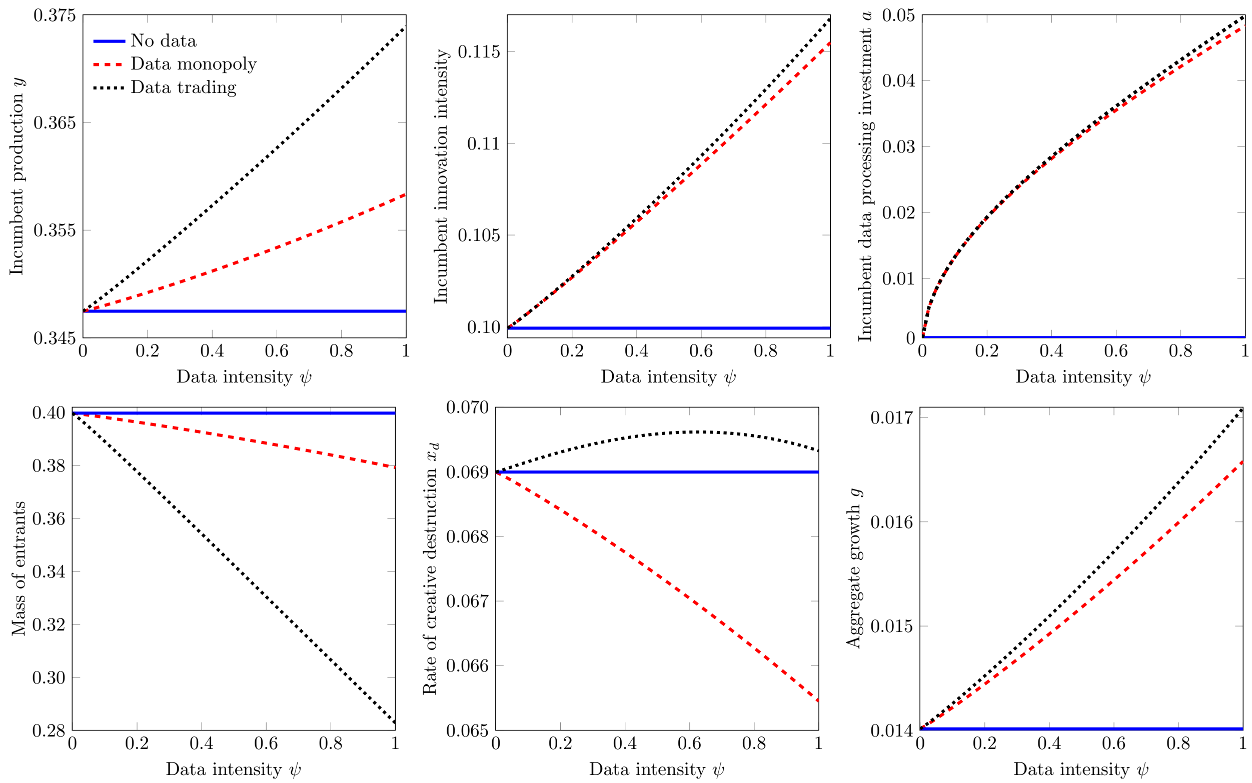

Figure 1: **Optimal policies in a data economy.** The top row shows the incumbent optimal production rate y, innovation z, investment in data processing a. The bottom row shows the mass of entrants m, the rate of creative destruction x_d and growth g as a function of the data intensity of the economy ψ. The solid blue, dashed red, and dotted black lines respectively represent outcome variables in the no-data, data monopoly, and data trading cases.

of innovation and endogenous growth, as detailed in Appendix F. We then examine numerically the effects of data on firm policies, creative destruction and growth in Figure 1.

The bottom row of Figure 1 shows that the mass of active entrants is always lower in the data economy compared to the no-data economy (left panel). This result is consistent with the idea of data acting as entry barriers. As a result of the decrease in the mass of entrants (and since $z_E^M = z_E^*$), the rate of creative destruction unambiguously decreases (middle panel), leading to a greater entrenchment of incumbents. The lower rate of creative destruction leads to a lower contribution to aggregate growth by entrants (equation (11) and Proposition 3). As illustrated by Figure 1, this effect can be offset by the increase in the contribution to growth of incumbents (equation (10) and Proposition 3), potentially leading to larger aggregate growth rate in a data economy.

Figure 1 plots firm-level optimal policies of both incumbents and entrants as functions of the data intensity of the economy ψ. Confirming our analytical results, the figure shows that when data supports the innovation process, incumbents exhibit a greater scale of production. The increase in the production rate is even greater in data intensive industries (for which ψ is larger). This result aligns with the idea that the data economy spurs the emergence of larger incumbents. Due to the complementarity between data and computing, investment in data processing a and in innovation z also increase with ψ.

B Data trading as a catalyst for creative destruction

We now examine how data trading influences endogenous outcomes relative to the incumbents' data monopoly and the no-data economy. The main result of this section is that *creative destruction can be greater in a data economy* than in one without data, provided entrants can access data to support their expansion. We further show in section C that this conclusion continues to hold when data sourcing serves as a substitute for data trading.

We start by the following analytical characterization of the effects of data trading on production and innovation rates.

Proposition 4 (Data trading)**.** *In a trading equilibrium in which incumbents sell data to entrants and endogenously set the price of data θ^*:*

1. *Incumbent firms have greater value ($v^T \geq v^M = v^*$), produce more ($y^T \geq y^M > y^*$) and innovate more ($z^T > z^M$ and $z^T > z^*$) compared to both the data monopoly and the no-data benchmark equilibria. Data trading unambiguously increases incumbents' contribution to growth, i.e., $g_I^T \geq g_I^M > g_I^*$;*

2. *Entrants invest more in innovation compared to both the incumbent data monopoly and the no-data benchmark, $z_E^T \geq z_E^* = z_E^M$.*

Recall that the price of data θ^* in a data trading economy is endogenous. Consider first the entrants' incentive to buy data. Differentiating the entrants' Hamilton-Jacobi-Bellman equation with respect to the share f of incumbent data purchased gives

$$z_E^T(a_E^T)^\xi \left[\Lambda v^T - v_E^T\right] > \theta, \tag{15}$$

suggesting that entrants are willing to buy data if the associated price (the right-hand side) is sufficiently low—i.e., lower than the marginal gain associated with buying data (the left-hand side), making the arrival rate of a breakthrough more likely.

Next, consider incumbents' incentive to sell data. Incumbents set θ to maximize their value, subject to constraint (15) and $v^T \geq v^M$, i.e., incumbent value with trading should not be lower than in the incumbent data monopoly case, otherwise incumbents would refrain from selling data. On the one hand, selling data generates additional cash flows to incumbents, increasing the benefits of operating at a larger scale and of innovating. On the other hand, selling data to entrants improves the efficiency of entrants' innovation technology, increasing incumbent's likelihood of being hit by creative destruction. Incumbents are willing to sell data if their value increases compared to the incumbent data monopoly, i.e., if $v^T \geq v^M$. If, instead, this inequality does not hold for any θ satisfying (15), the incumbent does not to sell data, and the model solution falls back to the incumbents' data monopoly case.

The higher incumbent value in the data trading economy also makes it more appealing for entrants to become incumbents, which occurs if they achieve a breakthrough. This effect, combined with entrants' increased chances of achieving a breakthrough due to their use of purchased data, incentivizes them to pursue a higher rate of innovation compared to both the no-data scenario and the incumbents' data monopoly case. Although we cannot derive analytical results for the the mass of active entrants or the rate of creative destruction, our numerical analysis below shows that data trading can actually increase creative destruction, challenging the notion that data serves as a barrier to entry.

Figure 1 plots firm-level optimal policies of incumbents and entrants as functions of the data intensity ψ in the data trading case, in addition to the no data and data monopoly cases. Confirming our analytical results, it shows that when data support the innovation process, incumbents exhibit a greater scale of production, and more so with data trading due to the entrants' additional demand for data. The increase in the production rate is again greater in data intensive industries. Consistent with Proposition 5, the innovation rate of incumbents is the largest in the data trading case. Moreover (and confirming our analytical result in Proposition 4), Figure 2 shows that the innovation rate of entrants is greater when they can acquire data from incumbents to support their innovation process.

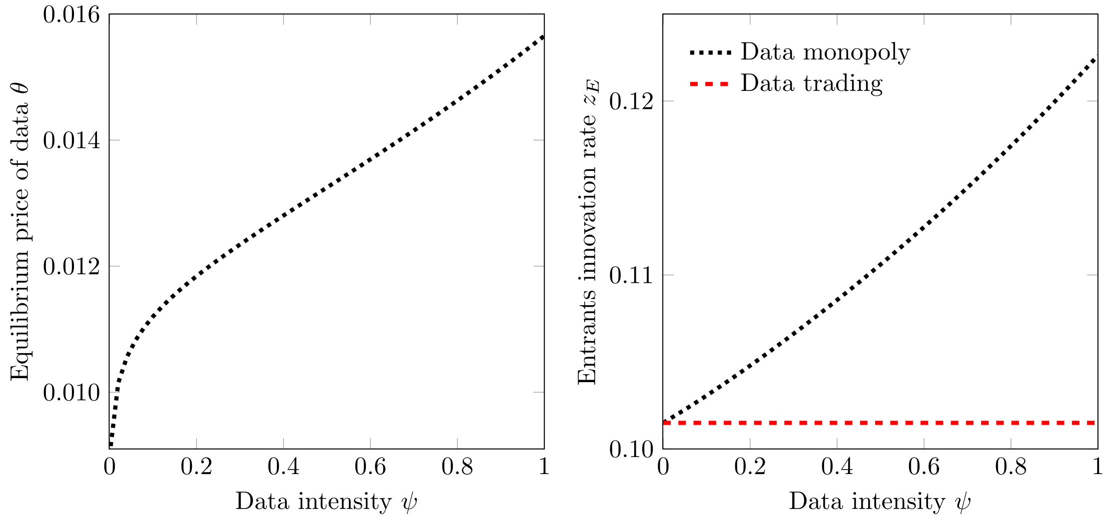

Figure 2: **The effects of data trading.** The left panel plots the equilibrium price of data, the middle panel plots the value of incumbents and the right panel plots entrants' innovation rate as functions of the data intensity of the economy ψ. The dashed red and dotted black lines respectively represent outcome variables in the data monopoly and data trading cases.

Figure 1 also shows that the mass of active entrants is the lowest in the data trading economy, consistent with data deterring entry. Yet, in contrast with the common view that data stifles business dynamism as a consequence, Figure 1 shows that the rate of creative destruction need not decrease when entrants can buy the data of incumbents. Recall that the rate of creative destruction aggregates the likelihood that entrants reach a technological breakthrough (the intensive margin that increases as shown in Proposition 4 and depicted in the right panel of Figure 2) across the mass of active entrants (the extensive margin, depicted in the bottom left panel of Figure 1). Because the intensive margin increases substantially in the data trading case compared to the no-data and incumbent data monopoly cases, it can more than offset the decline in the mass of entrants along the extensive margin. Thus, overall, the rate of creative destruction can increase in the data trading case.

Despite the higher rate of creative destruction, incumbents benefit from selling their data to potential entrants (not shown in the figure). The intuition is that incumbents sell data only when they can charge a price that outweighs the adverse effect of enhancing entrants' innovation efficiency.

These results highlight a more nuanced view than narrative in which data mechanically entrenches incumbents. Whereas data as a barrier to entry implies a smaller fringe of entrants, this need not translate into less creative destruction thanks to the greater engagement in innovation of these active, albeit fewer, entrants. As a result, entrants' contribution to growth need not decrease with data trading, supporting a higher overall growth rate.

C Data sourcing

The possibility for entrants to engage in data sourcing, i.e. strategies through which entrants independently obtain or generate the data to compete with incumbents, challenges incumbent's control over data usage and allows entrants to attain a greater innovation efficiency. The next proposition examines outcome variables when entrants can engage in data sourcing and provides analytical results in the special case in which $\mu = 1$ as to how data extraction affects endogenous quantities.

Proposition 5 (Data sourcing)**.** *Assume that the relation between investments in data sourcing and innovation rates exhibits linear returns to scale in that $\mu = 1$. When entrants invest in data sourcing:*

1. *The value of incumbents decreases below that in a no-data economy: $v^S \leq v^* = v^M < v^T$, the first inequality being strict when sourced data fosters innovation in that $\gamma > 0$.*
2. *Incumbents produce more than in the economy with no data $y^S > y^*$ but less than in the data trading case $y^S \leq y^M < y^T$, the first inequality being strict when sourced data fosters innovation in that $\gamma > 0$. Incumbents innovate less than in the data trading or data monopoly cases $z^S \leq z^M < z^T$, the first inequality being strict when $\gamma > 0$.*
3. *Entrants invest more in innovation $z_E^S > z_E^M = z_E^*$ compared to the no-data or the incumbent monopoly cases. Data sourcing may lead to greater or lower innovation rates than data trading in that the ordering of z_E^S and z_E^T is ambiguous.*

Proposition 5 shows that entrants' investment in data sourcing decreases the value of incumbent firms when $\gamma > 0$ and sourced data have value, as data sourcing makes entrants more innovative. While in the data trading case incumbents can impact the degree of competitive pressure by optimally setting the price of data, they do not have such lever when entrants access data via data sourcing. The resulting lower incumbent value reduces the surplus from technological breakthroughs and leads to a decrease in the incumbent's innovation rate compared to both the data trading and the incumbent data monopoly cases. Due to the complementarity between production and innovation, production is also lower (although greater than in the no-data economy).

Proposition 5 further demonstrates that access to sourced data induces entrants to invest more in innovation than in settings where such data is unavailable—namely, under both a data monopoly and a no-data economy. In other words, even though the returns to innovation decline due to the reduced value of incumbents, the enhanced efficiency of innovation enabled by data sourcing ultimately encourages greater innovative effort by entrants.

Figure 3 plots the incumbent's innovation rate, the mass of entrants, the rate of creative destruction, and the aggregate growth rate as functions of the data intensity of the economy ψ, the cost of sourcing data δ, and the quality of sourced data γ. The figure compares the data sourcing case against the no data, the incumbents' data monopoly, and the data trading cases. Confirming our analytical results, the top panel shows that the incumbents' innovation rate is lower in the data sourcing case compared to the data trading case and, moreover, can also be lower than in the no-data economy. That is, in the data sourcing case, incumbents' innovation can be lower than in the no-data economy, because incumbents have no control over entrants' data access. Namely, $z^S < z^*$ if ψ is relatively low, in which case the extent to which production translates into useful data supporting innovation is lower. Consistent with economic intuition, the innovation rate of incumbents is greater when the cost of sourcing data is greater and when the quality of sourced data is lower.

The second row of Figure 3 shows that the mass of entrants is always lower in a data economy than in the no-data economy, independently of the data regime. In addition, the decrease in the mass of entrants is larger when the data intensity ψ and the quality of sourced data γ are larger or when the the cost of sourcing data δ is lower. This result confirms the received wisdom that, no matter how data is produced or exchanged, data act as entry barriers and make the fringe of entrants more concentrated.

Yet, as shown by the third row of the figure, when entrants have access to data either through trading or sourcing, the rate of creative destruction can be higher in a data economy

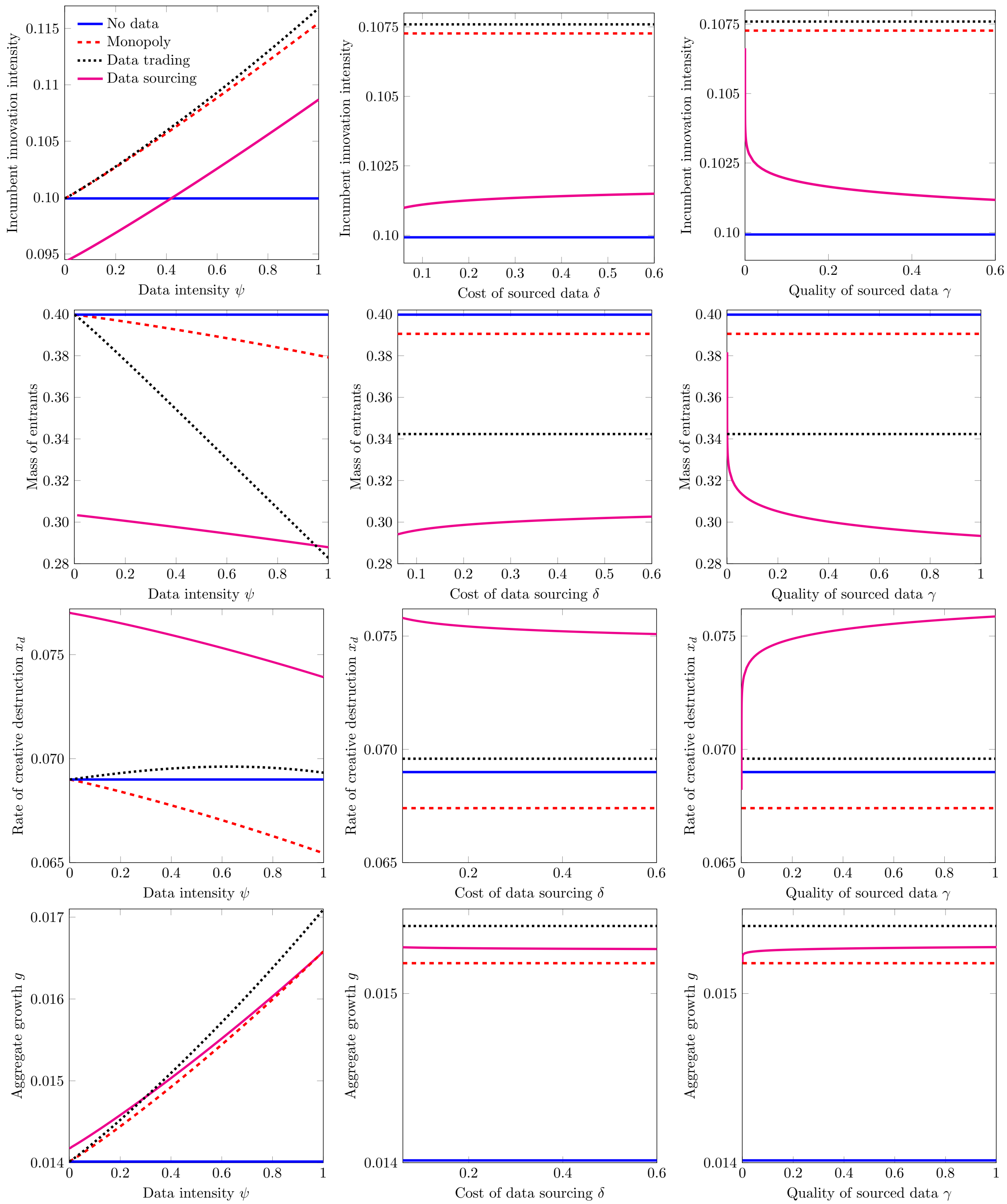

Figure 3: **Incumbent innovation, creative destruction, and growth in a data economy.** The figure shows the incumbent optimal innovation rate z, the mass of entrants, the rate of creative destruction, and the aggregate growth rate as functions of the data intensity of the economy ψ, the cost of sourcing data δ, and the quality of sourced data γ. The solid blue line represents the benchmark case with no data. The dashed red and dotted black lines respectively represent outcome variables in the data monopoly and data trading cases. The solid magenta line represents outcome variables in the data sourcing case.

than in an economy with no data. This is due to the higher innovation rates of entrants, as established in Propositions 4 and 5. Indeed, the rate of creative destruction results from aggregating the likelihood of entrants attaining a breakthrough (the intensive margin) across the mass of entrants (the extensive margin). The increase over the intensive margin can more than offset the decline over the extensive margin and, thus, creative destruction increases in this case. That is, in spite of fewer entrants, these fewer entrants are more innovative and, thus, the data economy need not discourage creative destruction.

The last row of Figure 3 shows that the equilibrium growth rate can be higher in a data economy than in a no-data economy. In addition, growth can be greater in the data trading or in the data sourcing case depending on the degree of data intensity ψ.

IV Welfare implications

Our analysis suggests that the growth rate in a data economy can be greater than in a no-data economy and that, in spite of a more concentrated fringe of entrants, the rate of creative destruction can also be greater, going against the conventional wisdom of lower business dynamism due to the entry barrier brought along by data.

In this section, we study the welfare implications of data access/usage. Welfare is the intertemporal utility of the representative agent which, in the balanced growth path equilibrium, is given by

$$\int_0^\infty e^{-\rho t} \ln(\mathcal{C}_t) = \frac{1}{\rho}\left[\ln(\mathcal{C}_0) + \frac{g}{\rho}\right],$$

weighting in the initial consumption of the representative agent and its growth rate.[11] By the budget constraint of the representative agent, consumption $\mathcal{C}_t$ is the sum of labor wage $\mathcal{W}_t$ and net dividends from incumbents $\mathcal{D}_t$, net of the financing cost to the entrants $\mathcal{F}_t$. Along the balanced growth path, consumption and all of these components grow at the endogenous rate g. We therefore have that

$$\mathcal{C}_t = \mathcal{W}_t + \mathcal{D}_t - \mathcal{F}_t = \mathcal{C}_0 e^{gt}$$

[11]We omit superscripts as the equations hold across the different data environments.

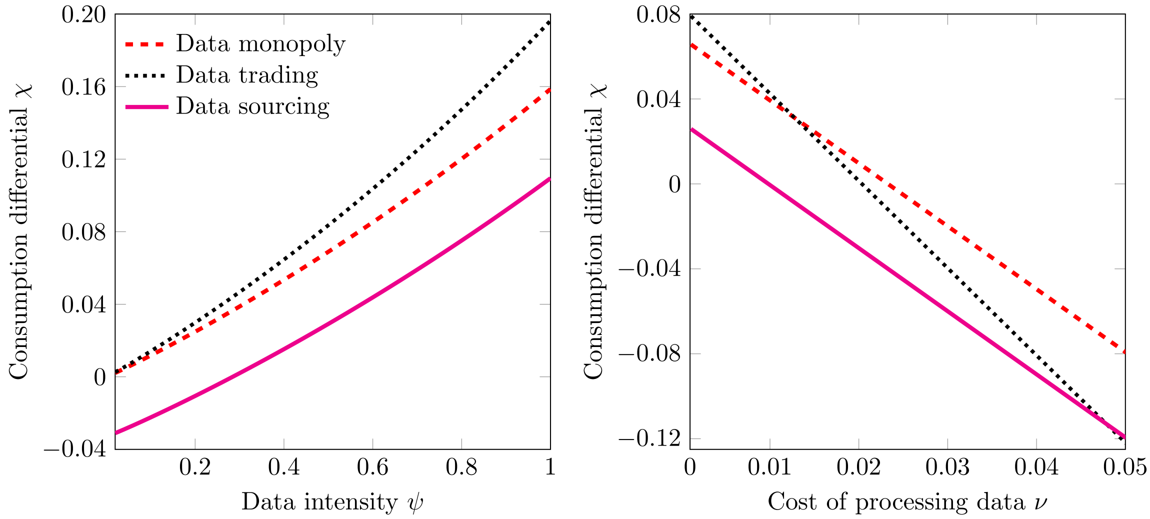

Figure 4: **Welfare.** The figure shows the consumption differential χ relative to the no-data economy in the incumbent data monopoly (dashed red line), in data sourcing economy (solid magenta line) and in data trading economy (dotted black line).

whose components are derived in the Appendix (and for the different data settings considered so far). It is worth emphasizing that only the expenditures that are sunk (i.e., those that are not just a transfer of funds) matter for consumption. For example, investment in data sourcing is lost to the consumption of the representative agent. Instead, entrants' cost of acquiring the incumbents' data correspond to an exchange of funds between incumbents and entrants and, thus, do not affect the consumption of the representative agent. That said, data trading affects the optimal choices of incumbents and entrants (including investment in innovation, which is sunk), so the comparison between the two cases is not obvious.

To compare welfare in the data and non-data economies, we calculate the differential in terms of initial consumption that equalizes welfare in the no-data economy to those in the different data configurations analyzed. That is, we calculate how much initial consumption would need to be raised or reduced in the no-data economy so that the representative agent has the same welfare level as in the different configurations of the data economy—considering in turn the incumbents' data monopoly, the data trading, and the data sourcing cases. We denote this measure by χ and report calculations in Appendix D.

The left panel of Figure 4 reports the corresponding results as a function of the degree of data intensity ψ. It shows that welfare is the highest with data trading and the lowest with data sourcing. Because, as highlighted in the previous section, the growth rate can be higher in the data trading or in the data sourcing cases depending on the degree of data intensity ψ, the wedge in welfare emerging from Figure 4 is largely driven by the consumption in the two settings. Indeed, data sourcing requires a duplication of investment (precisely, the entrants' investment in data sourcing) that is lost to the economy, whereas data trading largely involves an exchange between incumbents and entrants (so, it is just an exchange of funds and does not weigh on the consumption of the representative agent).

Privacy consideration Because our analysis focuses on innovation and industry dynamics, we abstract from privacy considerations in the baseline model. Privacy concerns, however, may affect welfare differently across data-access regimes. To examine this issue, we extend the framework to incorporate privacy costs. We assume that the use of production-generated data entails a privacy cost because such data originate from actual transactions. By contrast, data sourcing does not generate privacy costs, as it relies on publicly available information or synthetic data. Specifically, whenever production-generated data are processed, the representative agent incurs a proportional privacy cost. Processing an amount (ψy) of data therefore generates a cost equal to $(\nu\psi y)$, where (ν) measures the privacy cost per unit of data processed. We further assume that privacy costs increase with the number of parties processing a given body of data. Consequently, data sharing through trade amplifies privacy costs relative to regimes in which data remain exclusively under the control of the incumbent.

The right panel of Figure 4 reports the corresponding results, showing the consumption differential χ as a function of the privacy cost ν across the incumbent data monopoly, the data trading, and the data sourcing cases. While the economy with data trading continues to exhibit the greatest welfare if ν is sufficiently low (consistent with the analysis in the left panel, in which $\nu = 0$), this relation reverts if ν is sufficiently large. In fact, with data trading, data stemming from actual transactions are used by multiple parties (i.e., incumbents and entrants)—instead, only incumbents use such data in the data monopoly and data sourcing cases. Thus, if ν is sufficiently large, welfare can become the lowest with data trading.

V Allowing for M&As

The main takeaway of the model is that, in spite of leading to a greater industry concentration, the data economy can spur investment in innovation and growth. Absent merger activity, this implies a greater industry turnover in the data economy in spite of a greater concentration. In practice, we see established incumbent firms maintaining their leadership position on a prolonged basis, even when entrants successfully innovate. The reason is that these firms often engage in acquiring entrants achieving breakthroughs.[12]

We now extend the model by assuming that, whenever creative destruction hits, the successful entrant and the incumbent can merge with positive probability η.[13] In this case, the incumbent and the successful entrant negotiate over the terms of the deal—we assume Nash bargaining where $b \in (0,1)$ represents the bargaining power of the incumbent, and denote the endogenous acquisition price by N_{jt}. Conversely, with probability $1-\eta$, the entrant becomes the new incumbent whereas the former incumbent exits, as in our baseline model without mergers. In this case, the entrant pays a setup cost H_{jt} upon taking over the initiators' market position—whereas it bears no cost if it is acquired instead.

We build on the data sourcing case (and drop the superscript S for simplicity), where incumbents cannot control entrants' access to data. In this case, when the incumbent might take over the entrant, the valuation equation of the incumbent becomes (see Appendix E):

$$rV(q) = \max_{Y,z,a} \Bigg\{ Y(p-1) - \alpha \frac{a^2}{2} q - \zeta \frac{z^2}{2} q + z\left(\phi + \frac{a^{\xi} D}{q}\right)\left[V(\lambda q) - V(q)\right] \\ + x_d\left[(1-\eta)\left(\omega V(q) - V(q)\right) + \eta\left(V(\Lambda q) - V - N\right)\right] \Bigg\},$$

where the last term on the right-hand side captures probability-weighted payoff of incumbents when creative destruction hits. Notably, incumbents exit with probability $1-\eta$ or acquire the entrant with probability η. In the latter case, the incumbent gets the new technology with quality Λq_{jt-} and pays the price $N_{jt} = nq_{jt} = n\Lambda q_{jt-}$, endogenously derived below.

[12] The exit-via-merger outcome should not be seen as preventing entrants' innovation—rather, it might support it; see, e.g., Phillips and Zhdanov (2013) and Fons-Rosen, Roldan-Blanco, and Schmitz (2026).

[13] The merger rate can be endogenized by allowing the incumbent to search for targets, along the lines of Bustamante, D'Acunto, and Zucchi (2026).

Similarly, the Hamilton-Jacobi-Bellman equation of the entrant now satisfies:

$$
\begin{aligned}
rV_E(\bar{q}) = \max_{z_E, L_E} \Bigg\{ & V_{Et} - \zeta_E \frac{z_E^2}{2}\bar{q} - \frac{\delta}{2(1+\psi)} \int_0^1 \frac{L_{Ej}^2}{q_j} dj \\
& + z_E \left(\phi + \int_0^1 \left(\frac{\gamma L_{Ej}}{q_j} \right)^{\mu} dj \right) \Big[(1-\eta) \left(V(\Lambda \bar{q}) - V_E - H \right) + \eta \left(N - V_E \right) \Big] \Bigg\}.
\end{aligned}
$$

This equation differs from that in the baseline model in the last term on the right-hand side. Entrants continue to take over the incumbents' market position with probability $1-\eta$ (if the acquisition does not happen) but, with probability η, they merge with incumbents, in which case they receives the acquisition price. Because the outcome of an innovation breakthrough can result in a merger, this affects the optimal innovation rate, which then satisfies

$$
z_E = \frac{(\phi + \gamma^{\mu}\ell^{\mu}) \left[\Lambda \left((1-\eta)(v-h) + \eta n \right) - v_E \right]}{\zeta_E}.
$$

In turn, optimal investment in sourcing satisfies

$$
\ell_E = \left(\frac{z_E \left[\Lambda \left((1-\eta)(v-h) + \eta n \right) - v_E \right] \mu (1+\psi) \gamma^{\mu}}{\delta} \right)^{\frac{1}{2-\mu}}
$$

where the term in square brackets at the numerator suggests that the outcome of an entrant's innovation breakthrough affects its investment in data sourcing.

The incumbent and the entrant negotiate over the terms of the deal through Nash bargaining, where the endogenous acquisition price $N_{jt} = nq_{jt}$ satisfies (see Appendix (E)):

$$
n = v \left(1 - \frac{\omega(1-b)}{\Lambda} \right) - bh. \tag{16}
$$

The acquisition adds value to both parties if $\frac{v}{\Lambda}\omega < h$.

Figure 5 plots the optimal innovation rate of incumbents z, the mass of entrants m_E, the rate of creative destruction x_d, and the aggregate growth rate g as functions of the data intensity of the economy ψ and the probability of a merger η. The solid blue line depicts equilibrium outcomes in the no-data economy. The solid magenta line depicts outcomes in the data-sourcing economy. We use the baseline parameterization in Appendix F. Following

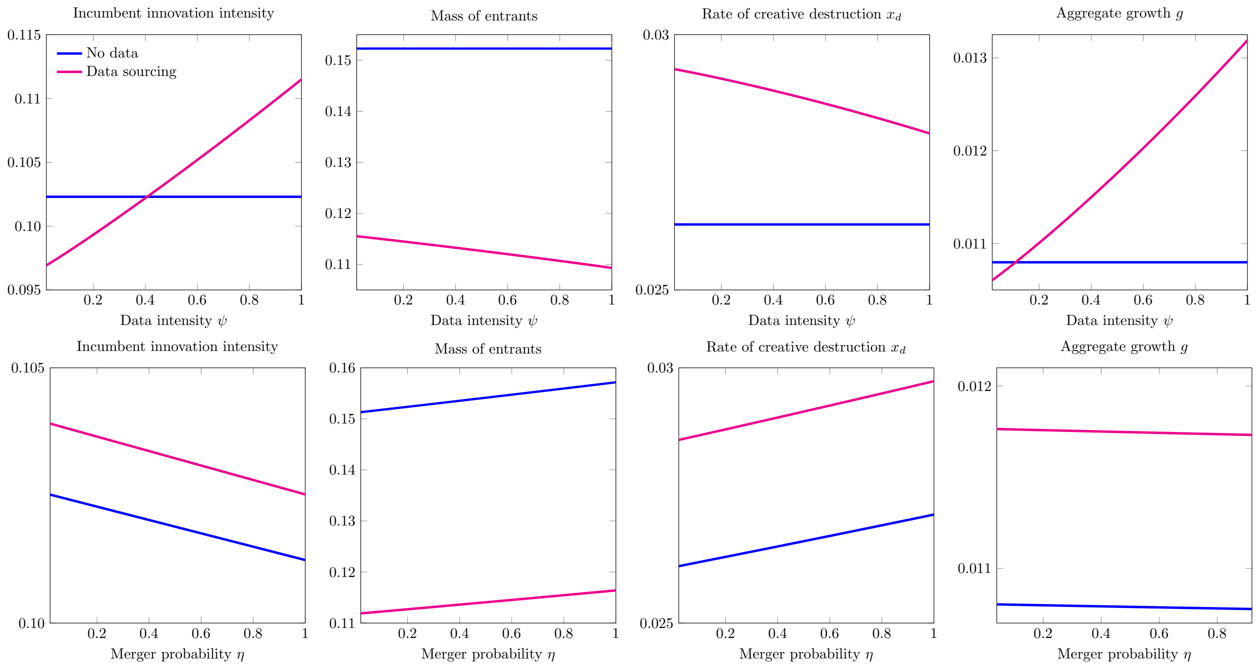

Figure 5: **Allowing for mergers.** The figure shows the incumbent optimal innovation rate z, the mass of entrants, the rate of creative destruction, and the aggregate growth rate as functions of the data intensity ψ and the probability of a merger η. The solid blue line represents outcome variables in the no data economy. The solid magenta line represents outcome variables in the data sourcing economy.

Paron (2025) and Fons-Rosen, Roldan-Blanco, and Schmitz (2026), we set the incumbent's bargaining power to $b = 0.5$. Following Phillips and Zhdanov (2013), we set the acquisition probability to $\eta = 0.186$.

The figure confirms that the main insights from our benchmark model remain robust when mergers between incumbents and entrants are allowed. In particular, the data sourcing economy continues to display a lower mass of entrants relative to the no-data economy. At the same time, the rate of creative destruction remains higher in the data economy, as illustrated in the third panel. Moreover, because incumbents cannot restrict entrants' access to data in the data sourcing economy, the incumbents' optimal innovation rate, z, may be either higher or lower than in the no-data economy. A higher innovation rate becomes more likely as data intensity, ψ, increases (as shown in the top-left panel). Consequently, when ψ is sufficiently low, the growth rate in the data economy can fall below that of the no-data economy, as shown in the bottom-left panel.

Consistent with Fons-Rosen, Roldan-Blanco, and Schmitz (2026), our model captures the idea that the prospect of a merger can increase entry and the rate of creative destruction. At the same time, incumbents might reduce their innovation rate when they can acquire technological breakthroughs from successful entrants, consistent with the evidence in Phillips and Zhdanov (2013).

VI Conclusion

This paper develops a dynamic framework to study how data access and AI-enabled processing shape innovation, firm dynamics, and economic growth. In our model, production generates data that can be transformed into a key input in the innovation process through investment in computing capabilities. Because incumbents naturally accumulate data through production while entrants must either acquire or source it, the data economy creates a novel interaction between market structure, innovation incentives, and creative destruction.

Our analysis delivers three main insights. First, data creates complementarities between production, innovation, and computing investment. Firms that produce more generate more data, which increases the productivity of R&D and strengthens incentives to invest further in both innovation and AI capabilities. This mechanism leads to the endogenous emergence of larger and more innovative incumbent firms and provides a micro-foundation for the rise of "superstar" firms in data-intensive industries.

Second, the effects of data on business dynamism are more nuanced than the conventional view that data mechanically entrenches incumbents and suppresses innovation. While data can act as a barrier to entry by reducing the mass of entrants, it simultaneously increases the productivity of innovation. As a result, the intensive margin of innovation among active entrants can offset the decline in entry, so that the overall rate of creative destruction may increase. Whether the data economy ultimately dampens or stimulates dynamism critically depends on how data access affects both the extensive and intensive margins of innovation.

Third, the mechanism through which entrants access data matters for equilibrium outcomes. Data trading can stimulate innovation on both sides of the market by allowing incumbents to monetize data while improving entrants' innovation efficiency. By contrast, data sourcing technologies intensify competitive pressure on incumbents but weaken their incentives to innovate because incumbents cannot control entrants' access to data. These differences have important implications for growth and welfare. In particular, absent large privacy concerns, data trading tends to dominate data sourcing because it avoids socially wasteful duplication of data acquisition efforts.

More broadly, the paper highlights that the economic effects of AI and data cannot be understood solely through the lens of concentration or entry barriers. Data reshapes not only who innovates, but also how innovation is produced and scaled. Policies governing data access, portability, privacy, and ownership therefore have potentially complex and countervailing effects on innovation, growth, and welfare.

A Appendix

A Proofs for the data monopoly case

The maximization problem of the final good sector yields the demand curve for each frontier input as in equation (3) and feeds into the Hamilton-Jacobi-Bellman of the incumbent, as per equation (4). Substituting our scaling conjecture gives:

$$r^M v^M = \max_{\{y^M, z^M, a^M\}} \Big\{ y^M \left[(y^M)^{-\beta} - 1 \right] - \alpha \frac{(a^M)^2}{2} - \zeta \frac{(z^M)^2}{2} + z^M \left(\phi + (a^M)^{\xi} \psi y^M \right) [\lambda - 1] \, v^M + x_d^M \, [\omega - 1] \, v^M . \Big\} \tag{17}$$

Maximizing this equation with respect to y^M, z^M and a^M gives the expression of the optimal policies as reported in the main text.

Consider next entrants. Substituting our scaling conjecture into the entrants' HJB equation gives:

$$(r^M - g^M) v_E^M = \max_{z_E^M} \left\{ - \zeta_E \frac{(z_E^M)^2}{2} + z_E^M \phi \left[v^M \Lambda - v_E^M \right] \right\} \tag{18}$$

where the left-hand side accounts for $\bar{q}$ growing smoothly at rate g^M along the balanced growth path. We maximize this equation with respect to z_E^M to obtain the optimal policy reported in the main text. Substituting the optimal z_E^M into (18) gives:

$$\rho v_E^M = \frac{\phi^2 (\Lambda v^M - v_E^M)^2}{2 \zeta_E}$$

Imposing the free-entry condition and solving the resulting equation with respect to v^M gives the expression in equation (14).

The expected increase in quality on each time interval is given by

$$E_{t-}[dq_{jt}] = q_{jt-} \left[(\lambda - 1)(\phi + (a^M)^{\xi} \psi y^M) z^M + (\Lambda - 1) \phi z_E^M m_E^M \right]$$

where the first and the second terms are the incumbents' and entrants' contribution to the quality improvement of input j. Improvements occur at independent Poisson times in different industries j. By the law of large numbers, the growth rate then satisfies

$$\int_0^1 q_{jt} dj = e^{g^M t}$$

where $g^M = g_I^M + g_E^M$ as reported in the main text. Aggregate quantities grow at rate g^M along the balanced growth path, and so does the consumption of the representative agent. The maximization problem of the representative agent then satisfies the standard Euler equation $g^M = r^M - \rho$. In turn, the resulting equilibrium interest rate feeds back into firms' decisions.

Substituting the expression for r^M into equation (17) gives

$$\left[\rho + x_d^M(\Lambda - \omega)\right] v^M = y^M((y^M)^{-\beta} - 1) - \zeta\frac{(z^M)^2}{2} - \alpha\frac{(a^M)^2}{2}$$

which we solve with respect to x_d^M and obtain

$$x_d^M = \frac{1}{\Lambda - \omega}\left[\frac{y^M\left[(y^M)^{-\beta} - 1\right] - \zeta\frac{(z^M)^2}{2} - \alpha\frac{(a^M)^2}{2}}{v^M} - \rho\right]. \tag{19}$$

The representative agent supplies one unit of labor inelastically, for which it receives labor income. In addition, the household collects dividends from incumbents and provides financing to entrants. Notably, only the financing that is sunk—i.e., that is lost to the economy (and is not just a transfer of funds)—matters for consumption.

By the representative agent's budget constraint, we have that consumption $\mathcal{C}_t^M$ is the sum of labor wages and dividends by the incumbents in the intermediate good sector net of financing to the entrants. Substituting the optimal production rate and the price of inputs into the production function of the final good sector gives

$$\mathcal{X}_t^M = \frac{(y^M)^{1-\beta}}{1-\beta}\int_0^1 q_{jt}dj = \mathcal{X}_0 e^{g^M t} \tag{20}$$

This, together with the price and the optimal quantities of the intermediate inputs yields the competitive labor wage

$$\mathcal{W}_t^M = \mathcal{X}_t^M - p(y^M)y^M\int_0^1 q_{jt}dj = (y^M)^{1-\beta}\left[\frac{\beta}{1-\beta}\right]e^{g^M t} = \mathcal{W}_0^M e^{g^M t} \tag{21}$$

Next, the representative agent receives the incumbents' profits as dividends,[14] plus the liquidation value of incumbents that are hit by creative destruction:

$$\mathcal{D}_t^M = \int_0^1\left[y^M\left[(y^M)^{-\beta} - 1\right] - \zeta\frac{(z^M)^2}{2} - \alpha\frac{(a^M)^2}{2} + x_d^M\omega v^M\right]q_{jt}dj = \mathcal{D}_0 e^{g^M t}$$

where the first three terms in the square brackets are the net cash flows of the incumbents on any time interval, whereas the last term is the liquidation value of incumbents when hit

[14]We require incumbents' net profits to be positive to avoid that all firm types make losses.

by creative destruction. In turn, the flow of financing to the entrants is

$$\mathcal{F}_t^M = x_d^M \Lambda \kappa \bar{q}_t + m_E^M \zeta_E \frac{(z_E^M)^2}{2} \bar{q}_t = \mathcal{F}_0^M e^{g^M t}$$

Because consumption (and the components above) grows at the balanced growth path g^M, it follows that $\mathcal{C}_0^M = \mathcal{W}_0^M + \mathcal{D}_0^M - \mathcal{F}_0^M$.

Benchmark: No-data economy. Results are obtained by setting $\psi = 0$ in the proofs of the previous section. Substituting the optimal z_E into the entrants' scaled HJB, imposing the free-entry condition, and solving the resulting equation with respect to v^* gives

$$v^* = \frac{1}{\Lambda}\left[\kappa + \frac{\sqrt{2\zeta_E \kappa \rho}}{\phi}\right].$$

Following the steps as in the data monopoly case, we get the following expression for the rate of creative destruction:

$$x_d^* = \frac{1}{\Lambda - \omega}\left[\frac{y^*((y^*)^{-\beta} - 1) - \zeta \frac{(z^*)^2}{2}}{v^*} - \rho\right] \tag{22}$$

We now prove Proposition 3, comparing optimal quantities in the incumbent data monopoly case against the no-data economy.

Proof of Proposition 3. Consider claim (1). The optimal production rate in the no-data economy y^* is in equation (12), whereas y^M is in equation (5). As z^M, a^M, v^M are all positive, and $\lambda > 1$ by assumption, then $y^M > y^*$. The optimal innovation rate in the no-data economy z^* is in equation (12), and z^M is in equation (6). As $a^M, y^M > 0$, $v^* = v^M$ as shown above, and $\psi > 0$ and $\lambda > 1$ by assumption, then it follows that $z^M > z^*$. Last, we have $g_I^* = ($ $lambda - 1)\phi z^*$, whereas $g^M = (\lambda - 1)\left[\phi + (a^M)^\xi \psi y^M\right] z^M$. Because $z^M > z^*$, and $a^M, y^M > 0$, it follows that $g_I^M > g_I^*$ holds.

Consider claim (2). Note that z_E^M and z_E^* satisfy a similar expression, see equations (8) and (13), respectively. As $v^* = v^M$ (as shown above) and $v_E^* = v_E^M = \kappa$ by the free-entry condition, then $z_E^M = z_E^*$ follows. As a last step, comparing the expressions for x_d^M in equation (19) and x_d^* in equation (22) gives that $x_d^M < x_d^*$ if the inequality in the proposition holds—i.e., it is solely driven by the relative magnitude of net cash flows in the two cases. Now, the rate of creative destruction is given by aggregating the innovation rates across the mass of entrants, so $x_d^M = m_E^M \phi z_E^M$ and $x_d^* = m_E^* \phi z_E^*$. Given that $z_E^M = z_E^*$ as just shown, the inequality $x_d^M < x_d^*$ implies that $m_E^M < m_E^*$ holds too. Lastly, as the entrants' contribution to growth satisfies $g_E^* = x_d^* (\Lambda - 1)$ and $g_E^M = x_d^M (\Lambda - 1)$ in the no-data and in the incumbent data monopoly case, the claim follows.

B Proof for the data trading case

To solve the model with data trading, we again exploit the scaling property $V(q) = v^T q$ to get the unscaled HJB equation:

$$r^T v^T = \max_{y^T, z^T, a^T} \Bigg\{ y^T \left[(y^T)^{-\beta} - 1 \right] + m_E^T \theta s \psi y^T - \alpha \frac{(a^T)^2}{2} - \zeta \frac{(z^T)^2}{2} + z^T \left(\phi + (a^T)^\xi \psi y^T \right) [\lambda - 1] \, v^T + x_d^T \left[\omega v^T - v^T \right] \Bigg\}.$$

Maximizing with respect to y^T gives (8), whereas z^T and a^T satisfy respectively

$$z^T = \frac{\left[\phi + (a^T)^\xi \psi y^T \right] (\lambda - 1) v^T}{\zeta} \text{ and } a^T = \left(\frac{\xi \psi y^T z^T (\lambda - 1) v^T}{\alpha} \right)^{\frac{1}{2-\xi}}.$$

Consider next entrants. Because of symmetry across product lines, (7) boils down to:

$$r^T V_E^T(\bar{q}) - V_{Et}^T = \max_{z_E^T, s, a_E^T} \left\{ -\zeta_E \frac{(z_E^T)^2}{2} \bar{q} - \theta s \psi y^T \bar{q} - \alpha \frac{(a_E^T)^2}{2} \bar{q} + z_E^T \left(\phi + (a_E^T)^\xi s \psi y^T \right) \left[V^T(\Lambda \bar{q}) - V_E^T \right] \right\}$$

Exploiting again the scaling property $V_E(\bar{q}) = \bar{q} v_E$ in turn gives the scaled HJB equation:

$$(r^T - g^T) v_E^T = \max_{z_E^T, s, a_E^T} \left\{ -\zeta_E \frac{(z_E^T)^2}{2} - \theta s \psi y^T - \alpha \frac{(a_E^T)^2}{2} + z_E^T \left[\phi + (a_E^T)^\xi s \psi y^T \right] \left[v^T \Lambda - v_E^T \right] \right\}$$

The above equation is maximized to obtain the entrants' optimal policies as reported in the main text. Note here that the incumbent production rate enters the valuation equation of the entrant, barring closed-form solutions in this setting.

By the law of large numbers, the growth rate satisfies $\int_0^1 q_{jt} dj = e^{g^T t}$, being the sum of the contribution of incumbents and entrants $g^T = g_E^T + g_I^T$. Entrants contribution satisfies equation (11), whereas the incumbent contribution satisfies equation (10). The maximization problem of the representative agent continues to satisfy the standard Euler equation.

Consider now the components of aggregate consumption. In the data trading economy, consumption continues to be the same of labor wage, dividends from the incumbents, and financing to the entrants. By similar calculations, we obtain that $\mathcal{X}_t^T$ and $\mathcal{W}_t^T$ give equations analogous to (20) and (21) (but substituting y^T with y^M). Next, the representative agent receives the incumbents' profits as dividends:

$$\mathcal{D}_t^T = \int_0^1 \left[y^T \left((y^T)^{-\beta} - 1 \right) + m_E^T \theta s \psi y^T - \zeta \frac{(z^T)^2}{2} - \alpha \frac{(a^T)^2}{2} + x_d^T \omega v^T \right] q_{jt} dj = \mathcal{D}_0^T e^{g^T t}$$

which, compared to the case with data extraction, is augmented by the flow coming from selling data to entrants. In turn, the flow of financing to the entrants is

$$\mathcal{F}_t^T = x_d^T \Lambda \kappa \bar{q}_t + m_E^T \left[\zeta_E \frac{(z_E^T)^2}{2} + \alpha \frac{(a_E^T)^2}{2} \right] \bar{q}_t + m_E^T \theta s \psi y^T \bar{q}_t = \mathcal{F}_0^T e^{g^T t}$$

which is, instead, augmented by the cost of acquiring data from the incumbents. The flow associated with trading is an exchange between the agents, not lost to the economy.

Next, we prove Proposition 4.

Proof of Proposition 4. Consider claim (1). Incumbents set the trading price to maximize their value. Firm value then cannot go lower than v^M, as otherwise incumbents would just decide not to sell data. Thus, the incumbent value in the no-data and incumbent monopoly cases $v^M = v^*$ represent a lower bound for v^T. Because of the expression for z^* reported in equation (12) and

$$z^T = \frac{\left[\phi + (a^T)^\xi \psi y^T\right] (\lambda - 1) v^T}{\zeta},$$

the claim that $z^T > z^*$ follows from $v^T \geq v^*$ and noting that $a^T, y^T > 0$. It is also straightforward to show that the inequality $y^T > y^*$ holds.

Consider now the relation between z^M and z^T, as well as between y^M and y^T. Recall from Proposition 1 that the incumbent policies satisfy

$$\begin{aligned}
y^i &= \left(\frac{1-\beta}{1 - m_E^i f \theta \psi 1_{\{i=T\}} - (a^i)^\xi \psi z^i (\lambda - 1) v^i} \right)^{1/\beta}, \\
z^i &= \frac{\left(\phi + \psi y^i (a^i)^\xi\right) (\lambda - 1) v^i}{\zeta}, \\
a^i &= \left(\frac{\xi \psi y^i z^i (\lambda - 1) v^i}{\alpha} \right)^{1/(2-\xi)}, \qquad i \in \{M, T\}.
\end{aligned}$$

Relative to the monopoly case, the incumbent's HJB equation in the trading equilibrium contains the additional revenue term

$$m_E^T \theta \psi y^T > 0,$$

which raises the marginal payoff from production. Holding (z, a, v) fixed, the optimal production policy is therefore strictly higher under data trading:

$$y^T > y^M.$$

Because production generates data that enhances innovation efficiency, a larger production scale increases the marginal return to both innovation and computing investment. The policy functions above imply that y, z, and a are strategic complements: z^i is increasing in y^i and a^i, while a^i is increasing in y^i and z^i. Moreover, incumbents only agree to sell data if doing so weakly increases firm value:

$$v^T \geq v^M.$$

Combining these inequalities implies

$$a^T \geq a^M$$

and therefore

$$z^T = \frac{\left(\phi + \psi y^T (a^T)^\xi\right)(\lambda - 1) v^T}{\zeta} \geq \frac{\left(\phi + \psi y^M (a^M)^\xi\right)(\lambda - 1) v^M}{\zeta} = z^M.$$

The inequalities are strict whenever data trading is active ($\theta > 0$ and $m_E^T > 0$), because the additional revenues from data sales strictly raise the marginal value of production and amplify the complementarity between production, data processing, and innovation. Hence,

$$y^T > y^M, \qquad a^T > a^M, \qquad z^T > z^M.$$

Consider now claim (2). Recall that $v_E^T = \kappa$ due to the free entry condition. Because $v^T \geq v^*$ and $a^T \geq 0$ (with the inequalities becoming binding if trading is suboptimal), it follows from the expression for z_E^T in equation (8) that $z_E^T \geq z_E^* = z_E^M$ from the corresponding expression for these endogenous quantities (where $z_E^* = z_E^M$ stems from Proposition 3).

C Data sourcing

We omit the derivations for the incumbents' optimal policies, as they follow steps similar to those in Appendix A. Consider the entrants. Substituting our scaling conjecture into the entrants' HJB equation gives:

$$(r^S - g^S) v_E^S = \max_{z_E^S, \ell_E} \left\{ -\zeta_E \frac{(z_E^S)^2}{2} - \delta \frac{(\ell_E)^2}{2(1+\psi)} + z_E^S (\phi + \ell_E^\mu \gamma^\mu) \left[v^S \Lambda - v_E^S \right] \right\}$$

which we maximize with respect to z_E^S and ℓ_E to obtain optimal policies.

As in the other environments, improvements occur at Poisson times in different industries j. By the law of large numbers, the growth rate then satisfies $\int_0^1 q_{jt} dj = e^{g^S t}$, and $g = g_I^S + g_E^S$ with $g_I^S = (\lambda - 1)(\phi + a^S \psi y^S) z^S$ and g_E^S as reported in equation (11). In turn, the equilibrium interest rate continues to satisfy $g^S = r^S - \rho$.

By the representative agent's budget constraint, we have that consumption $\mathcal{C}_t^S$ continues to be the same of labor wage, dividends from the incumbents, net of the financing to the entrants. Substituting the optimal production rate and the price of inputs into the production function of the final good sector gives

$$\mathcal{X}_t^S = \frac{(y^S)^{1-\beta}}{1-\beta}\int_0^1 q_{jt}dj = \mathcal{X}_0^S e^{g^S t}$$

whereas the competitive labor wage is given by

$$\mathcal{W}_t^S = (y^S)^{1-\beta}\left[\frac{\beta}{1-\beta}\right]e^{g^S t} = \mathcal{W}_0^S e^{g^S t}$$

Next, the representative agent receives the incumbents' profits as dividends:

$$\mathcal{D}_t^S = \int_0^1 \left[y^S\left[(y^S)^{-\beta}-1\right] - \zeta\frac{(z^S)^2}{2} - \alpha\frac{(a^S)^2}{2} + x_d^S \omega v^S\right] q_{jt}dj = \mathcal{D}_0^S e^{g^S t}$$

where the first three terms in the square brackets are the net cash flows of the incumbents on any time interval, whereas the last term is the liquidation value of incumbents when hit by creative destruction. Lastly, the flow of financing to the entrants is

$$\mathcal{F}_t^S = x_d^S \Lambda\kappa\bar{q}_t + m_E^S\left[\zeta_E\frac{(z_E^S)^2}{2} + \delta\frac{\ell_E^2}{2(1+\psi)}\right]\bar{q}_t = \mathcal{F}_0^S e^{g^S t}$$

Because consumption and its components all grow at the balanced growth rate g^S, it follows that $\mathcal{C}_0^S = \mathcal{W}_0^S + \mathcal{D}_0^S - \mathcal{F}_0^S$, using the expression just derived.

We now prove Proposition 5, which compares the different cases analyzed so far in the specific case $\mu = 1$, for which we can obtain closed-form expressions for z_E^S and ℓ_E.

Proof of Proposition 5. When $\mu = 1$, we can solve for the optimal policies of the entrant in closed form as:

$$\begin{aligned} z_E^S &= \frac{\phi\delta(\Lambda v^S - v_E^S)}{\delta\zeta_E - \gamma^2\psi(\Lambda v^S - v_E^S)^2}, \\ \ell_E^S &= \frac{\phi(\Lambda v^S - v_E^S)^2\gamma(1+\psi)}{\delta\zeta_E - \gamma^2(1+\psi)(\Lambda v^S - v_E^S)^2}. \end{aligned} \tag{23}$$

Substituting these expression into the scaled HJB equation of entrants, imposing the free-entry condition, and solving for v^S gives

$$v^S = \frac{1}{\Lambda}\left[\kappa + \sqrt{\frac{2\delta\zeta_E\rho\kappa}{\delta\phi^2 + 2\kappa\rho\gamma^2(1+\psi)}}\right]$$

Consider first claim (1). First of all, consider the relation between v^S and $v^* = v^M < v^T$

(where the latter inequality stems from Proposition 4). Substituting the relevant quantities into the inequality $v^S \leq v^*$ gives

$$v^S = \frac{1}{\Lambda}\left[\kappa + \frac{1}{\phi}\sqrt{\frac{2\delta\zeta_E\rho\kappa}{\delta + 2\kappa\rho\gamma^2(1+\psi)/\phi^2}}\right] \leq \frac{1}{\Lambda}\left[\kappa + \frac{\sqrt{2\zeta_E\kappa\rho}}{\phi}\right] = v^*$$

which boils down to

$$\sqrt{\frac{\delta}{\delta + 2\kappa\rho\gamma^2(1+\psi)/\phi^2}} \leq 1$$

which indeed holds as κ, ρ, ψ, γ, and ϕ are positive and holds with a strict inequality when $\gamma > 0$, i.e. when sourcing data is valuable. Thus, $v^S \leq v^* = v^M < v^T$.

Consider next claim (2). Using the relevant expressions, we have that incumbents' optimal production rate satisfies

$$y^* = (1-\beta)^{\frac{1}{\beta}} < y^S = \left(\frac{1-\beta}{1-(a^S)^\xi \psi z^S(\lambda-1)v^S}\right)^{\frac{1}{\beta}}.$$

as $1-(a^S)^\xi\psi z^S(\lambda-1)v^S \leq 1$. Now, consider the relation between y^S and y^M, z^S and z^M, and a^S and a^M—recall that these expressions share the same functional form, although $v^S \neq v^M$, so $z^S \neq z^M$, $y^S \neq y^M$, and $a^S \neq a^M$. As a direct consequence of the inequality $v^S \leq v^* = v^M$ shown above, the following inequalities should follow $z^S \leq z^M$, $y^S \leq y^M$, and $a^S \leq a^M$. Moreover, due to the complementarity, y^S, z^S, and a^S are lower, as a "second round effect" impacts these quantities. Last, recall that $y^M < y^T$ and $z^M < z^T$ stem from Proposition 4. The full inequalities $y^X \leq y^M < y^T$ and $z^X \leq z^M < z^T$ follow.

Last, consider claim (3). Recall first that $z_E^M = z_E^* = \frac{\phi(\Lambda v^* - \kappa)}{\zeta_E}$. Using the expression for z_E^S in equation (23) gives that $z_E^S \geq z_E^*$ if the following inequality holds:

$$\frac{\delta(\Lambda v^S - \kappa)\phi}{\delta\zeta_E - \gamma^2(1+\psi)\left(\Lambda v^X - \kappa\right)^2} \geq \frac{(\Lambda v^* - \kappa)\phi}{\zeta_E}$$

By calculations, we have

$$(\Lambda v^X - \kappa)^2(\Lambda v^* - \kappa)\gamma^2(1+\psi) > \delta\zeta_E\Lambda(v^* - v^X)$$

and, substituting the relevant expressions for v^X and v^*, the inequalities boil down to

$$\frac{2\kappa\rho\gamma^2(1+\psi)/\phi^2 + \sqrt{\delta(\delta + 2\kappa\rho\gamma^2(1+\psi)/\phi^2)}}{\delta + 2\kappa\rho\gamma^2(1+\psi)/\phi^2} > 1$$

An inspection of the left-hand side confirms that the inequality indeed holds—as the term under the square root is greater than δ. The claim holds. $\square$

D Welfare

Welfare is the intertemporal utility of the representative agent, which satisfies the expression reported in the main text. To compare welfare in consumption-equivalent terms across the different cases throughout, we calculate the quantity χ that solves:

$$\ln\left((1+\chi)\mathcal{C}_0^*\right) + \frac{g^*}{\rho} = \ln\left(\mathcal{C}_0\right) + \frac{g}{\rho}$$

where we take as a benchmark of comparison the no-data economy (whose endogenous quantities are labeled with $*$). On the right-hand side, we consider in turn the incumbent data monopoly, the data trading case, and the entrants' data sourcing.

By calculations, we then have

$$\chi = \frac{\mathcal{C}_0}{\mathcal{C}_0^*} e^{\frac{g-g^*}{\rho}} - 1$$

which is then the quantity we analyze numerically in the main chart. When accounting for the privacy cost described in the main text, it drains the consumption of the representative agents, so the endogenous quantities are adjusted accordingly to take this cost into account.

E Allowing for mergers

Given the HJB equations reported in the main text, we follow the same steps as in Appendix C to back out the scaled equations. Namely, the entrants' HJB equation boils down to

$$(r-g)v_E = \max_{z_E,\ell_E} -\zeta_E\frac{(z_E)^2}{2} - \delta\frac{(\ell_E)^2}{2(1+\psi)} + z_E(\phi+\ell_E^{\mu}\gamma^{\mu})\left[(1-\eta)(\Lambda v - v_E - \Lambda h) + \eta(\Lambda n - v_E)\right]$$

Maximizing with respect to z_E and ℓ_E gives the optimal policies reported in the main text, where we note that these endogenous choices become functions of the probability of a merger happening and the associated outcome. In turn, the incumbent equation satisfies:

$$\begin{aligned} rv = \max_{y,z,a}\, & y\left[y^{-\beta}-1\right] - \alpha\frac{a^2}{2} - \zeta\frac{z^2}{2} + z\left(\phi + a^{\xi}\psi y\right)\left[\lambda - 1\right]v \\ & + x_d\left[(1-\eta)(\omega-1)v + \eta(\Lambda v - v - \Lambda n)\right] \end{aligned}$$

Maximizing this equation with respect to y, z, and a returns the incumbent optimal policies, which satisfy expressions analogous to those in the data sourcing case.

When a successful entrant (who has attained a breakthrough) and an incumbent meet, they negotiate over the terms of the deal through Nash bargaining, where the endogenous acquisition price $N_{jt} = nq_{jt} = n\Lambda q_{jt-}$ is the solution to:

$$\arg\max(\Lambda v - \omega v - \Lambda n)^b(\Lambda n - \Lambda v + h\Lambda)^{1-b}$$

where recall that $H_{jt} = hq_{jt} = h\Lambda q_{jt-}$ The first term is the incremental value to the initiator stemming from the acquisition as opposed to exiting due to creative destruction. The second term is the incremental value for the entrant stemming from being acquired as opposed to becoming initiator. Solving this expression gives the expression for n reported in equation (16) in the main text.

F Baseline calibration

We calibrate the model parameters building on previous models of innovation and endogenous growth. The discount rate ρ is set to 0.02. We set the size of quality improvements due to innovations to $\lambda = 1.05$ and $\Lambda = 1.08$ following Akcigit and Kerr (2018), Malamud and Zucchi (2019) and Akcigit, Hanley, and Serrano-Velarde (2021)). This specification captures the observation that innovations by entrants tend to be more path-breaking than those of incumbents. Following these contributions, we also set the innovation cost coefficient $\zeta = 0.65$ and $\zeta_E = 2.6$. We set the scaling parameter ϕ in the hazard rate of innovation to 1.7 to get a growth rate in the economy with no data similar to that in Akcigit, Hanley, and Serrano-Velarde (2021).

We assume the recovery rate in liquidation $\omega = 0.65$, which lies between the estimates of Korteweg (2010) and Glover (2016). (In the extension with mergers, we assume that $\omega = 0$, as this allows us to run comparative statics on a larger parametric range.) Moreover, we set $\beta = 0.106$ as Akcigit and Kerr (2018). The set the cost coefficient $\alpha = 0.05$ and $\delta = 0.1$, to acknowledge that data generation is more costly than data processing—moreover, note that these costs are markedly lower than the innovation cost coefficient. Lastly, we set the magnitude of the entry cost to $\kappa = 0.67$, which enables us to get an entrants' contribution to growth between 30 and 40 percent, consistent with Akcigit, Hanley, and Serrano-Velarde (2021). Last, we conservatively set the values of ξ and μ to 0.1, and set the parameter γ to 0.7.

References

Acharya, V., X. Shao, and Z. Xu. 2026. "AI and Productivity: The Role of Innovation." Working Paper New York University.

Agarwal, A., J. McHale, and A. Oettl. 2019. "Finding Needles in Haystacks: Artificial intelligence and Recombinant Growth." in *The Economics of Artificial Intelligence: An Agenda* by A. Agarwal, J. Gans, and A. Goldfarb, eds, University of Chicago Press.

Aghion, P., A. Bergeaud, M. Dewatripont, and J. Matt. 2025. "Firm Dynamics and Growth with Soft Budget Constraints." CEPR Discussion Paper.

Agrawal, A., J.S. Gans, and A. Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." *Journal of Economic Perspectives* 33:31–50.

Akcigit, U., and S. Ates. 2021. "Ten Facts on Declining Business Dynamism and Lessons from Endogenous Growth Theory." *American Economic Journal: Macroeconomics* 13:257–298.

Akcigit, U., D. Hanley, and N. Serrano-Velarde. 2021. "Back to Basics: Basic Research Spillovers, Innovation Policy and Growth." *Review of Economic Studies* 88:1–43.

Akcigit, U., and W. Kerr. 2018. "Growth Through Heterogenous Innovation." *Journal of Political Economy* 126:1374–1443.

Aldasoro, I., L. Gambacorta, R. Pal, D. Revoltella, C. Weiss, and M. Wolski. 2025. "AI Adoption, Productivity and Employment: Evidence from European Firms." EIB Working Paper.

Autor, D., D. Dorn, L.F. Katz, C. Patterson, and J.V. Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." *Quarterly Journal of Economics* 135:645–709.

Babina, T., A. Fedyk, A. He, and J. Hodson. 2024. "Artificial Intelligence, Firm Growth, and Product Innovation." *Journal of Financial Economics* 151:103745.

Bontadini, F., C. Corrado, J. Haskel, and C. Jona-Lasinio. 2026. "AI as an Innovation in the Method of Innovation: Implications for Productivity Growth." *AEA Papers and Proceedings* 116:36–40.

Brown, J., S. Fazzari, and B. Petersen. 2009. "Financing Innovation and Growth: Cash Flow, External Equity, and the 1990s R&D Boom." *Journal of Finance* 64:151–184.

Bustamante, M., F. D'Acunto, and F. Zucchi. 2026. "Innovation, Industry Equilibrium, and Discount Rates." Working Paper.

Cockburn, I., R. Henderson, and S. Stern. 2019. "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis." in *The Economics of Artificial Intelligence: An Agenda* by A. Agarwal, J. Gans, and A. Goldfarb, eds, University of Chicago Press.

Cong, W., D. Xie, and L. Zhang. 2021. "Knowledge Accumulation, Privacy, and Growth in a Data Economy." *Management Science* 67:5969–6627.

Farboodi, M., A. Koh, and B. Xia. 2025. "Data-Driven Automation.", pp. .

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp. 2019. "Big Data and Firm Dynamics." *AEA Papers and Proceedings* 109:38–42.

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

Fedyk, A., R. Mihet, K. Rishabh, and O. Gomes. 2026. "Data Innovation Complementarity and Firm Growth." Working Paper.

Fons-Rosen, C., P. Roldan-Blanco, and T. Schmitz. 2026. "The Effects of Startup Acquisitions on Innovation and Economic Growth." Working Paper.

Geelen, T., J. Hajda, and E. Morellec. 2022. "Debt, Innovation, and Growth." *Review of Financial Studies* 35:4152–4200.

Glover, B. 2016. "The Expected Cost of Default." *Journal of Financial Economics* 119:284–299.

Jones, C., and C. Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review* 110:2819–2858.

Korteweg, A. 2010. "The Net Benefits to Leverage." *Journal of Finance* 65:2137–2170.

Lane, P. 2026. "AI and the Euro Area Economy." Keynote Speech at the ECB-SAFE-RCEA International Conferencce on the Climate-Macro-Finance Interface.

Lebastard, L., and D. Sondermann. 2026. "Arificial Intelligence: Friend or Foe for Hiring in Europe Today." ECB Blog.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521:436–444.

Malamud, S., and F. Zucchi. 2019. "Liquidity, Innovation, and Endogenous Growth." *Journal of Financial Economics* 132:519–541.

Paron, J. 2025. "The wealth of stagnation: Falling growth, rising valuations." Working Paper Stanford University.

Phillips, G., and A. Zhdanov. 2013. "R&D and the Incentives from Merger and Acquisition Activity." *Review of Financial Studies* 26:34–78.

Restrepo, P. 2025. "We Won't Be Missed: Work and Growth in the AGI World." NBER conference volume on *The Economics of Transformative AI.*

Romer, P. 1993. “Two Strategies for Economic Development: Using and Producing Ideas.” in *Proceedings of the World Bank Annual Conference on Development Economics, 1992* Washington DC: The World Bank.

Varian, H. 2019. “Artificial Intelligence, Economics, and Industrial Organization.” in *The Economics of Artificial Intelligence: An Agenda* by A. Agarwal, J. Gans, and A. Goldfarb, eds, University of Chicago Press.