Wharton
University of Pennsylvania

Alina Jiajing Song
2422 Steinberg-Dietrich Hall
Philadelphia, PA 19104
574-440-3856
songjj@wharton.upenn.edu

Dear PhD student Workshop on Climate Finance Committee,

I am writing to apply for the PhD Student Workshop on Climate Finance, to be held on October 9, 2025 in Cambridge, Massachusetts. I am a fifth-year PhD candidate in Finance at the Wharton School, University of Pennsylvania. I expect to complete my dissertation in May 2027. My research sits at the intersection of corporate finance, energy economics, and the digital economy, focusing on how the physical infrastructure that powers data-intensive technologies reshapes firms' climate-relevant investment choices and valuations.

Enclosed is my working paper, "Data-Gravity". The study examines how data center geography affects firm outcomes. Combining the first-ever dataset linking the near-universe of data center locations with online job postings, I find that establishments of the same firm located in regions with greater data center capacity increase investments in data-intensive (AI) human capital while reducing traditional technology investments. Data centers are power-intensive infrastructure. I therefore exploit variation in local electricity slack to instrument the construction of data centers. Firms more exposed to data center growth experience larger increases in sales, employment, assets, and market value.

I believe my project is of particular interest to the workshop because it identifies an energy-supply channel through which exogenous variation in local electricity supply reallocates real economic activities: lower marginal power costs induce data-center entry, and the resulting "data gravity" draws labor and capital into the same counties. Furthermore, I plan on exploring private-equity financing as an additional, external financial shock that shapes the equilibrium trajectory of data-center build-outs and therefore affects labor demand in different types of technologies.

By revealing the electricity-slack channel that guides data center geography and firm-level AI investment, my work highlights a rapidly growing, electricity-hungry sector whose carbon impact and financing needs remain under-studied in climate finance. I look forward to discussing how integrating explicit emissions data and dynamic energy pricing can sharpen policy prescriptions—whether through differentiated power tariffs, clearer Scope-2 accounting, or incentives for renewable-powered "green data centers."

I would be honored to present my research, receive feedback, and engage with scholars who are shaping the climate-finance field. Thank you for considering my application. Please feel free to contact my adviser, Professor Dan Garrett, for further information.

Sincerely,

Alina Jiajing Song

ALINA JIAJING SONG

Phone: 574-440-3856 Email: songjj@wharton.upenn.edu

EDUCATION

The Wharton School, University of Pennsylvania 2021 - Present
Ph.D. Candidate in Finance

University of Notre Dame 2017 - 2021
B.A. in Economics & Applied and Computational Mathematics, *summa cum laude*

RESEARCH INTERESTS

Corporate Finance

WORKING IN PROGRESS

Data Gravity

This paper challenges the conventional wisdom of viewing data as purely intangible by examining how data center geography affects firm outcomes. Combining the first-ever dataset linking the near-universe of data center locations with online job postings, I find that establishments of the same firm located in regions with greater data center capacity increase investments in data-intensive (AI) human capital while reducing traditional technology investments. Further, firms systematically reallocate AI roles toward establishments close to data centers. To establish causality, I use an instrumental variables approach exploiting regional electricity slack. Firms more exposed to data center growth experience larger increases in sales, employment, assets, and market value. Consistent with a data-gravity mechanism, in which heavy data pull in resources and activities, the effects are more pronounced for digital firms and data-intensive establishments.

M&A, Skill Demand, and Labor Market Power with Elmer Li and Marina M. Tavares

RESEARCH EXPERIENCE

The Wharton School

Research Assistant to Prof. Huan Tang *2024*
The Network Structure of Data Economy
Breaking the Data Chain: The Ripple Effect of Data Sharing Restrictions on Financial Markets

Research Assistant to Prof. Joao Gomes *2022 - 2023*
Oligopoly Dynamics with Financing Frictions
Monetary Policy and Financial Stability

Wilson Sheehan Lab for Economic Opportunities 2018 - 2019
Research Assistant, randomized controlled trials (RCT) for several homelessness prevention programs

TEACHING EXPERIENCE

Macroeconomics and the Global Economic Environment, EMBA, MBA 2022 - 2024
Teaching assistant to Prof. Stephen Meyer, Tim Landvoigt, Tom Winberry

Central Banks, Macroeconomic Policy and Financial Markets, MBA 2023 - 2024
Teaching assistant to Prof. Joao Gomes and Zvi Eckstein

Corporate Finance, Undergraduate 2022
Teaching assistant to Prof. Howard Kaufold

HONORS AND AWARDS

Wharton ESG Initiative Grant ($5000)	2025
AI & Analytics at Wharton Research Grant ($5500)	2025
Mack Institute of Innovation ($5000)	2024
AI & Analytics at Wharton Research Grant ($4500)	2024
AI & Analytics at Wharton Research Grant ($6700)	2024
Bernoulli Award First Prize	2021
John Harold Sheehan Award for the best senior honors thesis	2021
Phi Beta Kappa	2021

OTHER WORK EXPERIENCE

Jefferies, New York, NY 2020
Equity Research Summer Associate

CODING & SOFTWARE SKILLS

Python, STATA, Matlab, SAS

REFERENCES

Prof. Daniel Garrett
Assistant Professor of Finance
The Wharton School
University of Pennsylvania
danielgg@wharton.upenn.edu

Prof. Joao Gomes
Howard Butcher III Professor of Finance
The Wharton School
University of Pennsylvania
gomesj@wharton.upenn.edu

Prof. Nikolai Roussanov
Moise Y. Safra Professor of Finance
The Wharton School
University of Pennsylvania
nroussan@wharton.upenn.edu

Prof. Huan Tang
Assistant Professor of Finance
The Wharton School
University of Pennsylvania
huantang@wharton.upenn.edu

Data Gravity

Alina Song

April, 2025

Preliminary and Incomplete.
Please do not cite or circulate.

Abstract

This paper challenges the conventional wisdom of viewing data as purely intangible by examining how data center geography affects firm outcomes. Combining the first-ever dataset linking the near-universe of data center locations with online job postings, I find that establishments of the same firm located in regions with greater data center capacity increase investments in data-intensive (AI) human capital while reducing traditional technology investments. Further, firms systematically reallocate AI roles toward establishments close to data centers. To establish causality, I use an instrumental variables approach exploiting regional electricity slack. Firms more exposed to data center growth experience larger increases in sales, employment, assets, and market value. Consistent with a data-gravity mechanism, in which heavy data pull in resources and activities, the effects are more pronounced for digital firms and data-intensive establishments.

1 Introduction

Data centers have emerged as the backbone of the global digital economy. The rapid growth of mobile technologies and artificial intelligence (AI) has propelled firms and investors to channel substantial capital into the development of specialized facilities to store and process vast amounts of data[1]. Yet the broader economic payoff is debated: data centers are extremely capital-intensive, but once operational they employ few workers[2]. Do data centers enhance productivity? How do data centers reshape economic activities?

This paper tackles these questions by analyzing within-firm responses to regional expansions in data center capacity. Does regional data center construction cause firms to shift labor both geographically and across technologies? How does it affect firm performance and value? To answer these questions, I assemble the *first-ever* dataset linking the near-universe of US data center locations and online job postings. I document that first, in regions where data center capacity is high, firms expand AI-related job postings while cutting back on traditional analytic ("OldTech") roles. Second, firms reallocate AI workers into the affected regions and, in parallel, move OldTech labor away (Figures 4 and 6). Ultimately, firms with greater exposure to data center growth experience faster growth in sales, employment, assets, and market value.

Data centers are themselves intriguing economic objects beyond their indispensable role in powering today's digital society. Unlike traditional infrastructure such as roads or ports, data centers are, in principle, location-agnostic: users can access data storage and compute resources from virtually anywhere. Accordingly, a natural null hypothesis is that within-firm decisions remain unchanged in response to region-specific data center expansion. The results in this paper reject this view, suggesting that data has *gravity*: large volumes of data exert a "gravitational pull" on services and applications.

[1] A recent McKinsey report projects that global demand for data center capacity could more than triple by 2030 (Srivathsan et al., 2024). Private-equity funds have accordingly poured billions into data center developers, treating data centers as a high-return, strategic investment (Thomas, 2023).

[2] For example, Google's large-scale build-out in Loudoun County, Virginia created roughly 150 direct jobs despite sizable tax incentives and public-sector infrastructure spending (Merica and Bedayn, 2024).

I formalize this hypothesis by presenting a conceptual framework in Section 3 where a multi-region firm produce with AI labor, but AI labor is constrained by firm-wide data capacity. If local data center expansion raises productivity in firms' locations uniformly, all regions should see a similar increase in the demand for AI labor, and there is no reallocation effect. However, as long as the local data center expansion raises local productivity the most (compared with peer locations within the same firm), there will be heterogeneity across regions as well as reallocation effect. The findings in this paper reject the null hypothesis, showing instead that proximity to data centers significantly reshapes both the composition and geography of firms' labor demand.

Empirically testing the hypothesis is challenging for several reasons. Data availability is the primary issue. Most data centers are privately owned, so no public dataset discloses their locations or sizes; likewise, no dataset tracks firms' technology portfolios, let alone distinguishes AI tools from OldTech. I fill this gap by assembling a facility-level panel with information on the addresses, sizes, and year operational of the near-universe of data centers in the US. To the best of my knowledge, this is the first in the literature. On the technology side, I measure technology adoption for AI and OldTech using the human-capital approach (Acemoglu et al., 2022; Babina et al., 2024; Abis and Veldkamp, 2024). The heavy reliance of AI and analytics on human expertise makes the human-capital-based approach particularly well-suited in this setting. The combined dataset provides a comprehensive overview of firm-region(henceforth, establishment)-year level vacancies and regional data center expansions. The master dataset allows me to measure each establishment's exposure to regional data center capacity and its investments in AI v.s. OldTech human capital (measured by the fraction of AI and OldTech job postings) at the annual frequency.

Moreover, even with data, identification remains difficult due to potential reverse causality and omitted variable issues. To be more specific, data center operators may strategically choose to build facilities in regions where firms are already intensifying their AI activities. Additionally, unobserved regional factors could simultaneously influence both data center

construction and AI adoption. To address these concerns, I first employ establishment-year level regressions with a battery of granular fixed effects, controlling for time-varying firm-level shocks and time-invariant establishment and region-specific characteristics that could otherwise bias the results. Furthermore, I implement an instrumental variables strategy that leverages regional *electricity slack*, defined as the pre-period share of unused electricity capacity (the gap between installed nameplate capacity and actual utilization) multiplied by national data center capacity. Because data centers require substantial electricity, regions with greater ex-ante electricity slack are more attractive for such investments, supporting the relevance condition. Additionally, since this slack measure predates the rise in data center expansion and AI adoption, it is plausibly uncorrelated with contemporaneous error terms. Moreover, national capacities are unlikely to be influenced by regional shocks. Together, these arguments provide strong support for the validity of the instrument's exclusion restriction.

I begin by comparing establishments within the same firm across different locations. I conduct two types of analysis: establishment-year panel regressions and long-difference regressions, employing both ordinary least squares (OLS) and two-stage least squares (2SLS) instrumental variable (IV) approaches. While long-difference regressions are in particular helpful for slow-moving technological progresses, panel regressions allow me to utilize the full set of observations without requiri ng it to be present at both the beginning and the end of the sample period.

The main findings are as follows. Table 2 presents the establishment-year panel OLS regression results, while Table 5 reports the corresponding 2SLS IV estimates. The preferred specifications (Columns 5–7) suggest that doubling data center capacity increases the share of AI job postings by 0.20–0.56 percentage points (p.p.), concurrently reducing traditional technology (OldTech) postings by 0.03–0.31 p.p. Table 3 shows results from the long-difference OLS regressions, and Table 6 provides the IV estimates. The long-difference estimates indicate even larger cumulative effects: doubling data center capacity raises the

AI postings share by approximately 0.4 p.p. and reduces OldTech postings by about 0.47 p.p. These magnitudes are economically significant; for example, the changes in AI postings represent about 23 to 65% of the sample mean.

Next, I provide evidence of between-region employment reallocation in multi-location firms responding to data center expansion, following an approach similar to Giroud and Mueller (2019). Specifically, I construct a measure of "peer location data center exposure" for a given establishment, defined as the total megawatts (MW) of data center capacity that peer establishments within the same firm experience in a given year, scaled by peer establishments' job postings relative to the given establishment. This strategy compares the labor demand of two firms in the same CBSA-year exposed to different shocks due to variations in their internal establishment networks. This specification effectively controls for time-varying local economic shocks affecting local employment growth. The findings reveal a negative spillover effect on AI job postings and a positive spillover effect on OldTech postings. To interpret the results, consider this simplified example: if a firm employs an equal number of employees in regions A and B, doubling data center capacity in region A reduces the fraction of AI postings by 0.05 p.p. in region B while increasing OldTech postings by 0.04 p.p.

Are firms benefiting from being located close to data centers? I show that firms with greater exposure to data center expansions exhibit significantly higher growth in sales, employment, assets, and market value. Building on existing literature that highlights data's pivotal role in enhancing firms' prediction and decision-making processes (Veldkamp and Chung (2024)), I provide consistent empirical support for this mechanism underlying superior firm performance. Specifically, I demonstrate that managers in firms more exposed to data center growth make more accurate revenue forecasts. In 2023, firms experiencing greater increases in data center capacity between 2010 and 2023 are 25% more likely to have their sales forecasts for the subsequent fiscal year fall within the projected range. A placebo test confirms this effect was not present in 2013.

Why does the physical locations of data centers matter? What frictions prevent the firms from accessing the non-local data? I argue that data possesses *gravity*, suggesting that substantial data volumes naturally attract services and applications.[3]. Since moving heavy data is costly and complex, it becomes more efficient to bring compute and services to the data. In other words, the frictions of accessing and utilizing data grow as data gets heavier. If this hypothesis holds, we should observe stronger economic effects in regions where firms store substantial volumes of frequently accessed data. However, this is unobservable. Thus, I proxy data usage intensity in several ways.

First, I distinguish digital from non-digital firms based on whether they have any mobile apps or their industry classification (e.g., tech versus non-tech sectors). Second, I proxy establishment-level data intensity using establishment size, capturing the idea that data is a byproduct of economic activity (more output, more data). Third, I use population size as a proxy for consumer data, under the assumption that consumer data accumulates more substantially in populated regions. Finally, considering that one function of data centers is to facilitate data exchange along the supply chain, I test whether effects intensify for firms with suppliers or customers in that region. Consistent with the hypothesis, we find that the results are stronger for digital firms and data-intensive establishments. Consistent with the data gravity hypothesis, my empirical findings reveal stronger outcomes for digital firms and data-intensive establishments across these various proxies.

Taken together, the findings in this paper underscore a largely overlooked aspect of data: the necessity of dedicated physical spaces for its storage and processing. I propose that data exerts a form of *gravity*, attracting economic activities that heavily rely on data closer to this infrastructure. By empirically quantifying both direct local impacts and within-firm geographic reallocations, this paper emphasizes the critical role that data-center geography plays in shaping the digital economy.

Related Literature: This paper contributes to the rapidly growing literature on the data

[3]https://www.crowdstrike.com/en-us/cybersecurity-101/next-gen-siem/data-gravity/

economy, where researchers increasingly treat data as a key production input and a valuable asset to firms. Classic work in financial economics models the acquisition and sale of information, thereby laying a foundation for modeling data (Admati and Pfleiderer, 1990; Veldkamp, 2005). More recent models emphasize data's unique properties: for instance, Jones and Tonetti (2020) highlight that data is nonrival and treat it as an input that directly boosts productivity. Others underscore data's role in improving predictions and decision-making (see Farboodi and Veldkamp (2023), Veldkamp and Chung (2024), and Baley and Veldkamp (2025) for an overview), changing the sources of firms' revenues and risks. The literature also emphasizes that data may need to be combined with labor input, physical capital (such as computing hardware), or other complementary resources (Abis and Veldkamp, 2024). Further, recent research shows that the exchange and use of data generates linkage among firms Bian et al. (2025). This paper builds on the data-as-input literature by providing novel evidence that even though data is intangible and nonrival, its effective use depends on tangible infrastructure, and this feature of data has implications for the spatial distribution of economic activities.

This paper also contributes to the growing literature examining how technology, particularly artificial intelligence (AI) and data analytics, impacts firm performance, valuation, and growth. For broad, investment-specific technology shocks, Papanikolaou (2011) and Kogan and Papanikolaou (2014) capital-embodied technology shocks are priced in equity markets. Turning to AI, Webb (2019) shows that the rise of AI shapes labor demand and job reallocation, and Acemoglu et al. (2022) provide evidence on how firm exposure to AI affects hiring and skill demand. Furthermore, Babina et al. (2024) document that firms investing in AI exhibit faster growth in sales, employment, and market value, primarily through increased product innovation. Eisfeldt et al. (2023) extend this line of work to generative AI, finding similar positive effects on firm valuation. These papers take firms' adoption of technology as given, and then explores the impact on firms. This paper emphasizes a new perspective: how the geographical distribution of data infrastructure underpins AI adoption decisions within

multi-establishment firm and its implications on firm valuation and performance.

Furthermore, this paper also connects to research on firms' establishment networks. A long line of work on internal capital markets demonstrates that multi-unit firms can allocate resources across divisions or plants to maximize overall returns (Stein (1997)). Studies find that when one unit of a conglomerate is flush with cash or hit by a shock, other units' investment and hiring are adjusted accordingly (see, for example, Lamont (1997); Shin and Stulz (1998)). Recent research has emphasized the spatial dimension of these internal networks. For example, Giroud and Mueller (2019) studied how local demand shocks spread across regions through the internal networks of multi-region, multi-establishment firms (see also Giroud and Mueller (2015), Giroud and Mueller (2017), Giroud et al. (2024)). A common feature of the geographical reallocation is firm-wide financial constraints as a mechanism that generates linkages between a firm's establishments, and ultimately between regions. This paper extends this intra-firm networks literature by uncovering a new channel of internal resource allocation: binding data constraint instead of financial constraint. Further, this paper examines the heterogeneity of technological workers and show that the reallocation depends on the type of work.

Finally, this paper study relates to the literature on place-based industrial policies. Greenstone et al. (2010) exploit "million dollar plant" openings to show that when a large manufacturing plant is attracted to a county, incumbent firms in the winning county become significantly more productive. More recently, Giroud et al. (2024) demonstrate how productivity spillovers from such large investments propagate through the internal networks of multi-establishment firms, potentially amplifying the local economic gains. This paper demonstrates that digital infrastructure can generate significant externalities for local economies, despite its access being location-independent. In this sense, the results have direct policy implications: subsidizing the construction of data centers or investing in underlying infrastructure may serve as a catalyst for local innovation, attract high-tech talent, ultimately bolstering both firm performance and regional development.

2 Institutional Background

2.1 Data Centers

History. Data centers emerged alongside advances in computing and networking, but have evolved significantly since the 1950s and 60s, when they were known as mainframes—large, centralized computers housed in specialized rooms. The advent of personal computers in the 1980s led to decentralized computing, but the rise of the internet in the 1990s necessitated centralized data storage and processing, giving birth to modern data centers. The 2000s then saw the advent of cloud computing , spurring the development of ever-larger hyperscale data centers[4].

Function. A modern data center is a specialized facility that houses an organization's core IT resources, including servers, storage systems, networking equipment, and security devices. The primary functions of a data center are twofold: (1) Store and Compute, encompassing the large-scale storage, processing, and analysis of data; and (2) Exchange/Network, which includes high-bandwidth connections to enable efficient data transfers across both internal and external networks. Modern data centers are designed to ensure high availability, robust security, efficient cooling systems, and backup power supplies.

Characteristics. Data centers are extremely energy-intensive. They require substantial electricity both to power servers and to cool equipment, making reliable power supply a paramount concern. A single large data center can demand on the order of 50–100 megawatts (MW) of power. For perspective, a 100 MW facility running continuously consumes about 876,000 MWh per year – roughly equivalent to the annual electricity usage of 62,500 average U.S. homes [5].

Unlike traditional factories, data centers can be accessed and managed remotely. This means companies can utilize off-site computing resources without needing physical proximity to the facility. A firm's servers can be located in a data center hundreds of miles away yet

[4] https://www.digitalrealty.com/resources/articles/a-brief-history-of-data-centers.
[5] https://www.fxbgadvance.com/p/environmental-cents-how-much-power.

be monitored and controlled in real time.

2.2 Artificial Intelligence

History. Artificial intelligence (AI) originated in the mid-1950s, marked by seminal conferences such as the 1956 Dartmouth Workshop, which convened a small group of researchers to explore the possibility of making machines "think." Early AI research focused on symbolic methods, rule-based systems, and logical reasoning. Computers of the era had very limited memory and processing power, so AI experiments were constrained to relatively narrow tasks—like solving puzzles or proving mathematical theorems—often under highly controlled settings. The 1990s marked the rise of machine learning, which emphasized algorithms that learn from data rather than relying solely on predefined rules. However, the defining moment doesn't arrive until 2012, epitomized by systems that significantly improved performance on tasks such as image classification (e.g., the ImageNet competition), and the commercialization potential of AI for recommendation systems, targeted advertising, fraud detection, and more was quickly recognized.

Characteristics. AI's key distinction from traditional data analysis methods lies in their capacity to autonomously learn intricate patterns from large volumes of high-dimensional data (Goodfellow et al., 2016; Hauptmann et al., 2015; Babina et al., 2024). Deep neural networks, for example, use many layers of nonlinear transformations to approximate highly complex functions that would be intractable for traditional models, allowing them to capture subtle interactions without manual feature engineering. These models are also scale-adaptive: accuracy tends to increase as computing power and data volume grow, enabling capabilities—such as real-time speech translation or generative image creation—that conventional statistics cannot match (Kaplan et al., 2020; Jordan and Mitchell, 2015).

3 Conceptual Framework

3.1 Set-up and Firm's Problem

A firm operates in N regions indexed by $i = 1, \ldots, N$. Regional unit i produce output using labor input according to the production function

$$f_i(L_i) \;=\; z_i(D_1, \ldots, D_N)^{1-\alpha}\, L_i^{\alpha}, \qquad 0 < \alpha < 1,$$

where L_i represents AI labor. D_i is the data capacity in each region. $z_i(\cdot)$ is a productivity index. We assume that data raises AI labor's productivity $\partial z_i / \partial D_i > 0$. that can depend on data capacity anywhere. The hiring of AI labor is constrained by the availability of data and computing. The constraint is at the *firm* level. This setup captures the idea that AI labor does not have to use the data and computing power

$$b \sum_{i=1}^{N} L_i \;=\; \sum_{i=1}^{N} D_i, \qquad b > 0.$$

where b captures the tightness of the constraint.

3.2 Optimization

Let the firm take the prices as given and choose the vector $\{L_i\}$ to maximize profits:

$$\max_{\{L_i\}} \; \sum_{i=1}^{N} \Big[p_i \, z_i(D_1, \ldots, D_N)^{\,1-\alpha}\, L_i^{\alpha} - w L_i \Big] \quad \text{s.t.} \quad b \sum_i L_i = \sum_i D_i.$$

Let $s_i \equiv \alpha \, p_i^{1/(1-\alpha)}$ and $Z \equiv \sum_{n=1}^{N} s_n \, z_n(D)$. The first-order conditions yield

$$L_i^{*} \;=\; R_i \, T, \qquad R_i \;=\; \frac{s_i \, z_i(D)}{Z}, \qquad \sum_{i=1}^{N} R_i = 1.$$

How does the optimal labor in region i react to data shocks in region k? We derive

$$\frac{\partial L_i^*}{\partial D_k} = \underbrace{\frac{\partial R_i}{\partial D_k} T}_{\text{share (reallocation)}} + \underbrace{R_i \frac{\partial T}{\partial D_k}}_{\text{budget}}, \qquad k = 1, \ldots, N,$$

Budget effect Each extra unit of data anywhere enlarges the firm-level data constraint budget by $1/b$; region i captures a fixed fraction R_i of that expansion. This part is *always* positive.

Share (reallocation) effect Shocks to D_k shifts the *composition* of AI labor across regions via R_i. We discuss the own-region($k = i$) v.s. cross-region($k \neq i$) effects below.

3.3 Responses to data shocks

Let $\varepsilon_{kn} = \dfrac{\partial z_n / \partial D_k}{z_n / D_k}$ denote the elasticity of region n's productivity with respect to D_k.

Lemma 1 (Own-region v.s. cross-region) *For $j \neq i$*

$$\frac{\partial R_i}{\partial D_i} = \frac{s_i z_i}{Z^2 D_i} \sum_{n \neq i} s_n z_n \big(\varepsilon_{ii} - \varepsilon_{in}\big), \tag{1}$$

$$\frac{\partial R_i}{\partial D_j} = -\frac{s_i z_i}{Z^2 D_j} \sum_{n=1}^{N} s_n z_n \big(\varepsilon_{jn} - \varepsilon_{ji}\big) \tag{2}$$

Case A: Local data shocks raise productivity raises local productivity the most. Assume $\varepsilon_{ii} > \varepsilon_{ji}$ for all $j \neq i$.

Proposition 1 (Own-region data shock) *Increasing D_i raises L_i^*.*

$$\frac{\partial L_i^*}{\partial D_i} > 0.$$

Proposition 2 (Cross-region data shock) *Increasing D_j for $j \neq i$ has an ambiguous effect on L_i^*: a negative share effect competes with a positive budget effect.*

Case B: Local data shocks raise productivity raises productivity uniformly. Assume $\varepsilon_{ij} = \varepsilon$ for all i, j.

Proposition 3 (Uniform expansion) *Any increase in any D_k raises every L_i^* proportionally: $\partial R_i/\partial D_k = 0$ and $\partial L_i^*/\partial D_k = R_i/b > 0$.*

The formal proofs are deferred to Appendix C.

4 Data and Measurement

4.1 Data Sources

4.1.1 Data Center Locations and Capacity

I obtain the data center locations in the United States from two proprietary sources: Baxtel[6] and 451 Research (a unit of S&P Global Market Intelligence) Baxtel is a data center consulting firm that helps clients select the most suitable data center options. Meanwhile, 451 Research aims to provide market analyses and in-depth reports on data center facilities. Both aim to provide a thorough picture of the data center industry, including metrics such as location, installed power capacity, total floor space, and facility ownership or operators. To the best of my knowledge, Greenstein and Fang (2020) is the only academic research paper in economics that utilized datasets on locations of data centers. They assembled a cross-sectional view of the supply of data centers by scraping data from Baxtel.com and Datacenters.com in February 2018 and February 2019 [7].

To form my master dataset, I take the union of these two sources and manually verify each data center's address and power capacity. An additional step involves determining the year each data center became operational. Here, I combine information from Baxtel and

[6] https://baxtel.com.

[7] I chose Baxtel.com over Datacenters.com because, unlike Datacenters.com's self-reporting model which allows data center suppliers to input their own information—potentially resulting in facility specification misrepresentations and outdated data—Baxtel.com independently validates all location information for greater accuracy and reliability.

451 Research with supplementary evidence from Google Earth imagery, news articles, and historical company websites (using the Wayback Machine). Through these cross-checks, I successfully collect an operational start year for data centers representing about 80% of the total US power capacity.

4.1.2 Lightcast

Lightcast[8] (formerly Burning Glass) compiles over 400 million job postings from 5.8 million companies. The company aggregates data from company websites and online job boards. A proprietary de-duplication algorithm ensures that repeated postings of the same position on multiple platforms are consolidated into a single record, reducing double-counting. A great advantage of the BG data over the JOLTS is that they allow us to categorize jobs by occupation at a detailed level. Specifically, it provides rich information on the job text, and each job posting entry is parsed to extract detailed information on: employer, detailed geographic codes, occupation and industry, and a rich taxonomy of skills. Existing research such as Acemoglu et al. (2022) and Hershbein and Kahn (2018) show that Burning Glass data closely track the evolution of overall vacancies in the US economy as recorded by the nationally representative Bureau of Labor Statistics (BLS) Job Openings and Labor Turnover Survey (JOLTS). [9].

Following Abis and Veldkamp (2024), Acemoglu et al. (2022), Babina et al. (2024), I use the dataset to identify specific categories of job postings. The details are described in the next section. A novelty in my analysis is to include the geographical dimension to capture the within-firm allocation of skills.

[8] https://lightcast.io.

[9] BG postings are shown to be disproportionately concentrated in occupations and industries that typically require greater skill. However, this is not a severe concern in this paper because 1) I focus on the heterogeneity within technology skills within firms and 2) the distributions of the postings are relatively stable across time.

4.1.3 Electricity

To capture the availability and usage of electricity resources across the United States, I draw on data from the U.S. Energy Information Administration (EIA)[10]. The EIA collects and publishes extensive information on electricity generation facilities. Existing plants are required to respond to the EIA surveys if the plant's total generator nameplate capacity is 1 Megawatt (MW) or greater. EIA Form 860 collects the nameplate capacity, which refers to the maximum intended or design capacity of a power plant or generator, measured in megawatts (MW). EIA Form 923 collects the total amount of electricity actually produced by each generator over a given period (annual or monthly), measured in megawatt-hours (MWh). The variables are reported at the plant-generator-year level. I aggregate the plants to the CBSA-year level.

4.1.4 Other Data Sources

I obtain accounting and financial data on US public companies from Compustat North America Fundamentals Annual and CRSP. I source demographic information from the Census American Community Surveys (ACS). For firm-level management guidance, I use IBES Guidance.

4.2 Measurement

The main goal of our analysis is to estimate the impact of data center establishment on labor demand from nearby firms. Our independent variable is the exposure of firms to data center construction. To capture variation in data center size, we follow industry convention by designating a data center's size by its electricity consumption capacity. This approach maintains flexibility in our capacity measure to accommodate different server applications, including storage, computation, and other uses.

[10] https://www.eia.gov/

For a region r in year t, we define data center capacity as the cumulative sum of all data centers' capacity (existing capacity from previously built data centers plus newly constructed data centers in that year). Further, we define a firm f's establishment in CBSA r's exposure to data centers in year t as:

$$\text{DC Exposure}_{r,t} = \log\big(1 + \text{Data Center Capacity}_{r,t}\big)$$

This formulation means higher electricity capacity in a region's data centers translates into a larger exposure measure for all firms in that region.

To examine the effect of such exposure on labor demand, we construct two distinct measures that differentiate between AI-related hiring and traditional statistical (OldTech) hiring. Our hypothesis is that AI-related tasks are substantially more data- and computing-intensive, and data center construction relaxes the computational constraints for AI work. To classify jobs, we follow the methodology developed by Abis and Veldkamp (2024), which relies on the frequency of relevant keywords throughout job descriptions.

Our approach differs from earlier studies such as Acemoglu et al. (2022) in two critical ways. First, we employ a finer partition of skill types: postings containing predominantly AI-specific terminology (e.g., "deep neural networks," "reinforcement learning") are classified as AI, whereas postings emphasizing more traditional statistical methods (e.g., "regression modeling," "time-series analysis") are classified as OldTech. Second, rather than rely solely on structured skill lists, we analyze the overall textual content of each job description—accounting for the relative frequency of different skill mentions—thus avoiding situations where overlapping references might conflate older and newer skill sets. Full details on the classification keywords are provided in the Appendix.[11]

[11]For example, a posting mentioning "deep neural networks" eight times and "familiarity with regression models" once would be classified under AI, whereas a simpler keyword-matching approach might incorrectly label it as both AI and OldTech.

Based on this classification methodology, we define:

$$\text{Fraction of AI Postings}_{r,t} = \frac{\text{Number of AI postings}_{r,t}}{\text{Total job postings}_{r,t}},$$

$$\text{Fraction of OldTech Postings}_{r,t} = \frac{\text{Number of OldTech postings}_{r,t}}{\text{Total job postings}_{r,t}}.$$

5 Empirical Approach

5.1 Establishment-level Panel Regressions

To estimate the impact of the construction of data centers on firms' labor demand across regions and technologies, I run the following regression at the firm-region-year level.

$$\mathrm{Y}_{f,r,t} \;=\; \beta\,\text{DC Exposure}_{r,t} \;+\; FEs \;+\; \varepsilon_{f,r,t}, \tag{3}$$

where

$$\text{DC Exposure}_{r,t} = \log\big(1 + \text{Data Center Capacity}_{r,t}\big)$$

where $\mathrm{Y}_{f,r,t}$ is the establishment-level labor demand (fraction of AI or OldTech job postings) and FEs represent a battery of specifications with different levels of fixed effects.

Table 3 presents the regression results. Across all columns, the estimated coefficient on data center exposure are positive and statistically significant for the share of AI postings (Panel A) and negative for the share of OldTech postings (Panel B). Our preferred specifications (Columns 5–7) suggest that doubling data center capacity is associated with approximately a 0.02 to 0.08 p.p. increase in the fraction of AI postings and a corresponding 0.05 to 0.10 p.p. decrease in the fraction of OldTech postings [12]. These estimates are

[12]Note that doubling capacity raises the log term by roughly $\ln(2) = 0.693$. Multiplying that increment by the estimated coefficients gives $\text{AI}_{(5)} = 0.0927 \times 0.693 = 0.064$, $\text{AI}_{(6)} = 0.0222 \times 0.693 = 0.015$, $\text{AI}_{(7)} = 0.122 \times 0.693 = 0.085$; $\text{OldTech}_{(5)} = -0.148 \times 0.693 = -0.103$, $\text{OldTech}_{(6)} = -0.0782 \times 0.693 = -0.054$, $\text{OldTech}_{(7)} = -0.0867 \times 0.693 = -0.060$.

economically significant since the average annual change in the fraction of AI and OldTech workers at the firm level is around 0.2 p.p. Additionally, this represents 2 to 10% of the sample mean of the fraction of AI postings, but only 0.5 to 1% of the sample mean of the fraction of OldTech postings. Further, we note that endogeneity issues might drive the coefficient small, since firms (and regions) likely expand capacity where latent AI demand is already strong, which biases OLS downward if the true causal effect is larger. We further address this issue in the next section.

The magnitudes differ once we incorporate different sets of fixed effects. We discuss why these coefficients vary across specifications, focusing specifically on Panel A. Column 1 presents the unconditional correlation. The results mimic the graphical representation in Figure 4, showing a strong positive correlation between the fraction of AI postings and data center capacity and a strong negative correlation between the fraction of OldTech postings and data center capacity. Column 2 adds firm fixed effects and eliminates time-invariant heterogeneity across firms. This shrinks the coefficient by half suggesting that firm-specific characteristics explain a significant portion of the variation in labor demand. For instance, large and data-savvy tech firms systematically hire more AI talent everywhere. Column 3 incorporates region-level (CBSA) fixed effects, removing region-specific characteristics such as labor markets and industry composition. The fact that it triples the coefficient indicates that data centers tend to be constructed in regions initially exhibiting lower AI intensity. This makes sense if we think data centers are constructed in distant regions with abundant land and power. Column 4 year adds year fixed effects and exhibits a smaller coefficients, indicating that certain years experiencing aggregate AI hiring booms inflate the unconditional correlation.

Our preferred specification starts from column 5 because we want to isolate variation that is within-firm-within-region over time. This triple-difference structure removes (i) time-invariant heterogeneity across firms, (ii) persistent differences across labor markets, and (iii) aggregate shocks common to all establishments in a given year, which is consistent

with the paper's goal to examine how a firm reallocates labor across its own establishments as local data center capacity expands. Column 6 further tightens the specification by using firm-year fixed effects, absorbing any unobserved, time-varying shocks specific to each firm in each year (for example, demand shocks, new product launches). The coefficient is identified from cross-regional deviations in capacity within the same firm and year. Finally, Column 7 uses firm-CBSA fixed effects (i.e., establishment fixed effects), and the identification becomes within-establishment changes over time; this specification strips away all cross-sectional variation.

5.1.1 Instrumental Variables

Construction By writing equation 3, we allow for the possibility of treatment endogeneity: that is, the independent variable and the error term to be correlated. This correlation can arise from various sources; for instance, regional economic conditions may simultaneously attract data-center investments and increase labor demand for AI roles, while potentially reducing traditional technology roles through wage competition or shifting industry composition. These omitted variables, captured in $\varepsilon_{f,r,t}$, would bias OLS estimates of β. To address this challenge, we implement an instrumental variables (IV) approach.

Specifically, we construct an instrumental variable, $z_{r,t}$, that exploits the interaction between historical regional electricity capacity and national data center growth. We begin by defining the *electricity slack* in region r at time t as follows:

$$\text{Electricity Slack}_{r,t} = \text{Nameplate Capacity}_{r,t}\ (\text{MW}) \times \text{Hours in the Period} - \text{Total Net Generation}_{r,t}\ (\text{MWh}),$$

which captures the total utilization of electricity in region r relative to its maximum potential

capacity [13] [14]. Next, denote our instrument as

$$z_{r,t} \;=\; \text{Slack Share}_{r,2005} \;\times\; \text{National Data Center Capacity}_t, \tag{4}$$

where

$$\text{Slack Share}_{r,2005} \;=\; \frac{\text{Electricity Slack}_{r,2005}}{\text{Electricity Slack}_{\text{Total},2005}}$$

is region r's proportion of total national "energy slack" at the start of the sample period (2005), and National Data Center Capacity$_t$ indicates the aggregate national data-center capacity in year t. Intuitively, data centers require substantial power resources, and regions endowed with greater initial electricity slack are more attractive for expansions. This instrument borrows the idea from a classical shift-share instrument (Bartik, 1991; Autor et al., 2013), which typically allocate a national industry-level shock (e.g., an import shock or demand shift) across local labor markets in proportion to those markets' baseline employment shares in that industry. The intuition behind this instrument is that national data-center capacity—the "shift" component—is taken as largely exogenous to local economic shocks. We then "share" this national shift across regions based on their initial electricity slack. From the perspective of shift-share designs, each region's Slack Share$_{r,2005}$ acts as its exposure to the aggregate shock (in this case, the growth in total data center capacity).

It is worthwhile to point out the importance of using the "slack" measure, rather than raw nameplate capacity. *Nameplate capacity* indicates a power plant's maximum potential output

[13] For each region r, let $\mathcal{J}_r$ be the set of power plants in r. Then the total net generation in region r at time t is:

$$\text{Total Net Generation}_{r,t} \;=\; \sum_{j \,\in\, \mathcal{J}_r} \text{NetGen}_{j,t},$$

and the total nameplate capacity in region r is:

$$\text{Nameplate Capacity}_{r,t} \;=\; \sum_{j \,\in\, \mathcal{J}_r} \text{NameCap}_j.$$

Here, NetGen$_{j,t}$ is the net generation from plant j at time t (in MWh), and NameCap$_j$ is the nameplate capacity of plant j (in MW).

[14] Since the frequency of observation is annual, hours in the period is $365 \times 7 \times 24 = 61320$.

under ideal conditions, but electricity supply responds to local demand. Two problems thus arise if we rely on raw electricity supply. First, a region with equally high supply and demand may exhibit little surplus capacity, making it *less* appealing for new data centers, even if total capacity appears large. Second, electricity supply may correlate with local economic conditions that also influence labor demand, thereby violating the exclusion restriction. By focusing on slack capacity—which removes the effect of local demand—we capture exogenous variation stemming from historical infrastructure investments and natural resource endowments rather than contemporaneous local economic factors.

First-Stage Regression. We replace the endogenous variable DC Exposure$_{r,t}$ with its predicted component from the instrument:

$$\text{DC Exposure}_{r,t} \;=\; \pi_1 \, z_{r,t} \;+\; FEs \;+\; \eta_{r,t}.$$

The parameter π_1 reflects the relevance of the instrument: a positive and statistically significant estimate of π_1 indicates that regions with greater ex ante slack share attract more data-center capacity as total national capacity expands.

Second-Stage Regression. In the second stage, we regress the establishment-level labor demand outcome on the fitted values of the endogenous regressor:

$$\text{Y}_{f,r,t} \;=\; \beta \, \text{DC } \widehat{\text{Exposure}}_{r,t} \;+\; FEs \;+\; \varepsilon_{f,r,t}.$$

Identification Assumptions We require two main assumptions for valid identification via $z_{r,t}$:

1. **Relevance:**

$$\text{Cov}\big(z_{r,t}, \text{DC Exposure}_{r,t}\big) \;\neq\; 0.$$

That is, the instrument must be strongly correlated with the actual data-center capac-

ity in region r. The development of large-scale data centers relies heavily on adequate and affordable electricity supplies. Consequently, regions with ex-ante higher electricity slack are more attractive locations for data-center investments.

2. **Exclusion Restriction**

$$\mathrm{Cov}\big(z_{r,t}, \varepsilon_{f,r,t}\big) \;=\; 0.$$

The instrument must not be correlated with unobserved factors that directly affect labor demand in region r. The regional slack share $\mathrm{SlackShare}_{r,0}$ is measured at the beginning of the sample and is largely determined by historical infrastructure investments and resource endowments. These factors are unlikely to be driven by (or strongly correlated with) unobserved current demand shocks that affect $\mathrm{Y}_{f,r,t}$. Moreover, the national data-center expansion, $\mathrm{National\ Data\ Center\ Capacity}_t$, is shaped by macro trends rather than local shocks in individual CBSAs, which helps ensure that $z_{r,t}$ is not picking up local economic cycles or firm-specific hiring demands.

Table 5 reports the two–stage least-squares results. The first-stage estimates in Panel A show that an electricity-slack increase of 1 percent is linked to a 0.63 percent rise in data center capacity. The Kleibergen–Paap F-statistics (37–74) comfortably exceed the conventional threshold of a weak instrument. Panel B reports the second-stage estimates. Across all specifications, the coefficient on instrumented data center exposure is positive and highly significant for the share of AI postings (Panel A) and negative for the share of OldTech postings (Panel B). Our preferred specifications (Columns 5–7) imply that doubling regional DC capacity is associated with roughly a 0.20–0.56 p.p. increase in the fraction of AI postings and a 0.03–0.31 p.p. decrease in the fraction of OldTech postings.[15] This represents a 23-65% of the mean for AI and 3% of the mean of OldTech.

[15]Doubling capacity raises the log term by $\ln(2) = 0.693$. Multiplying this increment by the estimated coefficients yields $\mathrm{AI}_{(5)} = 0.682 \times 0.693 = 0.473$, $\mathrm{AI}_{(6)} = 0.283 \times 0.693 = 0.196$, $\mathrm{AI}_{(7)} = 0.811 \times 0.693 = 0.562$; $\mathrm{OldTech}_{(5)} = -0.444 \times 0.693 = -0.308$, $\mathrm{OldTech}_{(6)} = -0.0411 \times 0.693 = -0.029$, $\mathrm{OldTech}_{(7)} = -0.440 \times 0.693 = -0.305$.

5.2 Long-difference Regressions

In addition to panel regressions, we also employ a long-difference approach to capture the long-term impact of data center construction:

$$\Delta Y_{f,r,2010-2023} \;=\; \beta\, \Delta\text{DC Exposure}_{r,2010-2023} \;+\; \varepsilon_{f,r},$$

where ΔDC Exposure$_{r,2010-2023}$ is the change in data center capacity in region r from 2010 to 2023 and ΔY$_{f,r,2010-2023}$ is the change in the fraction of AI or OldTech job postings at establishment f, r from 2010 to 2023.

The long-difference approach is especially relevant for processes like data-center expansions, which often occur slowly and have effects that may unfold over many years (Acemoglu and Restrepo, 2020; Babina et al., 2024). It exploits the long-run changes in the fraction of AI and OldTech postings and data center capacity over a substantial horizon, mitigating concerns about time-invariant unobserved heterogeneity and short-term fluctuations.

Table 3 demonstrates the long-difference estimates on the impact of the change in data center capacity on labor demand. Doubling regional data center capacity is associated with roughly a 12 p.p. rise in the fraction of AI job postings at the establishment level, while fraction of OldTech job postings falls by about 14 p.p.[16]. Both effects are highly significant. Further, the coefficients indicate that there is almost one-for-one reallocation of labor demand away from legacy technologies. Establishments in regions with larger increase in data center capacity substitute more traditional technology with AI technology.

A limitation of the long-difference design is that it relies on establishments observed in both 2010 and 2023, raising concerns about selection bias. To address this concern, we conduct robustness checks to compare the the mean outcomes over the first five years (2010–2014) and the last five years (2019–2023), which requires an establishment to appear

[16]Note that doubling capacity raises the log term by roughly $\ln(2) = 0.693$. Multiplying that increment by the estimated coefficients (0.181 for AI and –0.205 for OldTech) gives the implied percentage-point shifts reported in the text.

at least once in each window. The results are very similar. Further, we note that the establishment-level panel regressions make use of all the sample. The consistency between the two approaches strengthens our conclusions.

5.2.1 Instrumental Variables

We construct a similar IV for long-difference regressions. Since by construction, long-difference regressions are cross-sectional regressions, we use cross-sectional variation from Slack Share$_{r,2005}$ as an instrument.

First-Stage Regression. We instrument the endogenous long-run change in regional data-center capacity with the region's initial electricity slack share:

$$\Delta\text{DC Exposure}_{r,2010\text{–}2023} \;=\; \pi_1\,\text{Slack Share}_{r,2005} \;+\; \text{FEs} \;+\; \eta_r.$$

A positive and statistically significant estimate of π_1 indicates that regions endowed with greater ex-ante electricity slack experienced larger subsequent expansions in data-center capacity.

Second-Stage Regression. In the second stage, we regress the long-run change in establishment-level labor demand outcome on the fitted values of the endogenous regressor:

$$\Delta Y_{f,r,2010\text{–}2023} \;=\; \beta\,\Delta\text{DC}\,\widehat{\text{Exposure}}_{r,2010\text{–}2023} \;+\; \text{FEs} \;+\; \varepsilon_{f,r}.$$

Identification Assumptions

1. **Relevance.**

$$\text{Cov}\Big(\text{Slack Share}_{r,2005}, \Delta\text{DC Exposure}_{r,2010\text{–}2023}\Big) \;\neq\; 0.$$

Pre-period electricity slack is a strong predictor of where new data centers are built.

2. **Exclusion Restriction.**

$$\text{Cov}\Big(\text{Slack Share}_{r,2005}, \varepsilon_{f,r}\Big) \;=\; 0.$$

Conditional on fixed effects and controls, baseline slack—determined by historical infrastructure and resource endowments—should not correlate with unobserved shocks to labor-demand growth between 2010 and 2023.

Table 6 reports the second-stage results of the long-difference IV specification estimating the impact of changes in data center capacity on changes in the fraction of AI and OldTech Postings from 2010 to 2023. Panel A reports the first-stage estimates, which strongly support the relevance of the instrument. The slack share in 2005 is significantly positively associated with the changes in data center capacity. The Kleibergen-Paap F-statistics is 326.40, far exceeding conventional thresholds against weak instruments. Panel B presents the second-stage estimates, which capture the cumulative effect of data center construction on firms' labor demand. The IV estimates indicate that firms in CBSAs where data center expansion was exogenously driven by the instruments experienced a substantially greater long-term increase in the fraction of AI job postings but a decrease in the fraction of OldTech job postings. Note that the long-difference IV estimates yield larger coefficients (in absolute terms) than the establishment-level panel regression estimates. This discrepancy arises because the IV approach captures the *cumulative* effect of data center construction over the sample period, whereas the panel regression approach identifies the *average* effect on an annual basis.

5.3 Peer Location Effects

Our empirical analysis closely follows prior studies on firm internal networks (e.g. Giroud and Mueller (2019)). In particular, we look at the labor demand growth in one establishment after its peer establishments owned by the same firm experience growth in exposure to data center capacity.

For firm f, CBSA r, and year t, define an establishment's peer location exposure to data center as

$$\text{Peer Location DC Exposure}_{f,r,t} \;=\; \log\Big(1 \;+\; \text{Data Center Capacity}^{\text{Other}}_{f,r,t}\Big)$$

where

$$\text{Data Center Capacity}^{\text{Other}} \;=\; \sum_{r' \neq r} \left(\frac{\text{Postings}_{f,r',t-1}}{\text{Postings}_{f,r,t-1}} \right) \times \text{DC Capacity}_{r',t}.$$

Here, $\dfrac{\text{Postings}_{f,r',t-1}}{\text{Postings}_{f,r,t-1}}$ represents how large location r' was relative to r within firm f at time $t-1$, which captures the *relative importance* of other locations to the focal location.

I run the following specification:

$$\Delta \text{Y}_{f,r,t \to t+k} \;=\; \beta\, \text{PeerExposure}_{r,t} \;+\; \alpha_f \;+\; \gamma_{r,t} \;+\; \varepsilon_{f,r,t},$$

5.4 Firm-level Outcomes

$$\Delta Y_f \;=\; \beta\, \Delta\text{DC Exposure}_f \;+\; \varepsilon_f, \quad \text{where} \quad \Delta Y_f = log(Y_{f,t_1}) - log(Y_{f,t_0}).$$

where Y_f is firm-level financial outcomes. DC Exposure$_f$ is the posting-weighted average of CBSA-level data center capacity.

6 Mechanisms

6.1 Digital Firms

The estimated coefficients are expected to be stronger among firms that heavily on data. To test this hypothesis, I split the sample into data-intensive and non-data-intensive firms, categorizing data intensity using two distinct measures. The first measure classifies firms

based on whether they have mobile apps (see Bian et al. (2021, 2023) for further discussion on using mobile app presence as an indicator of data intensity). The second measure identifies firms operating within the information and technology sector.

Table 8 reports the results of these analyses. Panels A vs. B indicate that firms with mobile apps exhibit substantially larger coefficients across all specifications. Panels C vs. D similarly show that technology-sector firms (NAICS2 = 51) exhibit larger coefficients. These results strongly support the hypothesis that data-intensive firms benefit more from regional data infrastructure.

6.2 Data as a Prediction Technology

One distinguishing data from technology, human capital, and other intangible capital is that it improves prediction (Baley and Veldkamp (2025)). Specifically, firms with bigger datasets can predict which ads and products will be most profitable, and they can make choices that covary more with the realized return on each choice. To test this prediction, I ran the following regression:

$$\mathbb{1}\{AccurateForecast\}_f \;=\; \alpha \;+\; \beta\,\Delta\text{Exposure}_f \;+\; \varepsilon_f,$$

where the independent variable is a dummy variable = 1 if realized sales falls within the range of management guidance.

7 Robustness

7.1 What about other infrastructure?

A natural follow-up question is whether the observed results are unique to data centers. As noted, data centers represent a distinct category of infrastructure because users do not have to be physically present to utilize them. Nonetheless, it is worth investigating whether the

construction of data centers is closely tied to other types of infrastructure, potentially leading us to conflate effects specific to data centers with those of other facilities. For instance, if data center construction is accompanied by broader regional infrastructure upgrades, such as broadband and power plants, these infrastructure investments may introduce biases in the estimated effects of data centers on labor demand.

To explore this possibility, I focus on electricity plants, which are presumably the infrastructure most likely to correlate with data centers' construction. I replicate the analysis from Equation 3, substituting the capacity of electricity plants for that of data centers as the independent variable. The results, shown in Table 11, reveal that the within-firm estimates become reversed for the fraction of AI postings and become statistically insignificant for the fraction of OldTech postings.

These results support the idea that it is not simply the availability of electricity power that matters for workforce composition—it's the specialized capacity for data processing, storage, and connectivity found in data centers.

7.2 IV Construction

One of the exclusion restriction of my IV relies on the assumption that each single region is too small to influence the national driver. In other words, national data center capacity is orthogonal to regional error terms. To completely rule out the idea that any single region is driving national levels, we can construct a leave-one-out IV [17]:

$$z_{r,t}^{\text{LOO}} \;=\; \text{SlackShare}_{r,2010} \times \text{National } \widetilde{\text{Data Center}} \text{ Capacity}_{-r,t}.$$

[17] Borusyak et al. (2022) demonstrate that constructing national growth rates using leave-one-out averages mitigates the finite-sample bias that arises mechanically from including own-observation information. This method parallels the jackknife instrumental variable estimator introduced by Angrist et al. (1999), which corrects the bias inherent in 2SLS estimation when numerous instruments are employed.

where

$$\widetilde{\text{National Data Center Capacity}}_{-r,t} = \sum_{r' \neq r} \text{DC Capacity}_{r',t},$$

By construction, the national capacity is independent to any single region's error term.

References

Abis, S. and Veldkamp, L. (2024), 'The changing economics of knowledge production', *The Review of Financial Studies* **37**(1), 89–118.

Acemoglu, D., Autor, D., Hazell, J. and Restrepo, P. (2022), 'Artificial intelligence and jobs: Evidence from online vacancies', *Journal of Labor Economics* **40**(S1), S293–S340.

Acemoglu, D. and Restrepo, P. (2020), 'Robots and jobs: Evidence from us labor markets', *Journal of political economy* **128**(6), 2188–2244.

Admati, A. R. and Pfleiderer, P. (1990), 'Direct and indirect sale of information', *Econometrica* **58**(4), 901–928.

Angrist, J. D., Imbens, G. W. and Krueger, A. B. (1999), 'Jackknife instrumental variables estimation', *Journal of Applied Econometrics* **14**(1), 57–67.

Autor, D. H., Dorn, D. and Hanson, G. H. (2013), 'The china syndrome: Local labor market effects of import competition in the united states', *American economic review* **103**(6), 2121–2168.

Babina, T., Fedyk, A., He, A. and Hodson, J. (2024), 'Artificial intelligence, firm growth, and product innovation', *Journal of Financial Economics* **151**, 103745.

Baley, I. and Veldkamp, L. L. (2025), *The Data Economy: Tools and Applications*, Princeton University Press.

Bartik, T. J. (1991), 'Who benefits from state and local economic development policies?'.

Bian, B., Huang, Q., Li, Y. and Tang, H. (2025), 'Data as a networked asset'. Working Paper.

Bian, B., Ma, X. and Tang, H. (2021), 'The supply and demand for data privacy: Evidence from mobile apps', *SSRN Electronic Journal* . Available at SSRN: https://ssrn.com/abstract=3987541.

Bian, B., Pagel, M., Raval, D. and Tang, H. (2023), Consumer surveillance and financial fraud, Working Paper 31692, National Bureau of Economic Research. Revised April 2024. **URL:** *https://www.nber.org/papers/w31692*

Borusyak, K., Hull, P. and Jaravel, X. (2022), 'Quasi-experimental shift-share research designs', *The Review of economic studies* **89**(1), 181–213.

Eisfeldt, A. L., Schubert, G. and Zhang, M. B. (2023), Generative ai and firm values, Technical report, National Bureau of Economic Research.

Farboodi, M. and Veldkamp, L. (2023), 'Data and markets', *Annual Review of Economics* **15**(1), 23–40.

Giroud, X., Lenzu, S., Maingi, Q. and Mueller, H. (2024), 'Propagation and amplification of local productivity spillovers', *Econometrica* **92**(5), 1589–1619.

Giroud, X. and Mueller, H. M. (2015), 'Capital and labor reallocation within firms', *The Journal of Finance* **70**(4), 1767–1804.

Giroud, X. and Mueller, H. M. (2017), 'Capital and labor reallocation within firms', *The Journal of Finance* **72**(4), 1767–1804.

Giroud, X. and Mueller, H. M. (2019), 'Firms' internal networks and local economic shocks', *American Economic Review* **109**(10), 3617–3639.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016), *Deep learning*, Vol. 1, MIT press Cambridge.

Greenstein, S. and Fang, T. P. (2020), Where the cloud rests: The location strategies of data centers, Technical report, Harvard Business School Working Paper.

Greenstone, M., Hornbeck, R. and Moretti, E. (2010), 'Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings', *Journal of political economy* **118**(3), 536–598.

Hauptmann, A. G., Hodson, J., Li, J., Sebe, N. and Rettinger, A. (2015), 'Cross-lingual cross-media content linking: Annotations and joint representations (dagstuhl seminar 15201)', *Dagstuhl Reports* **5**(5), 43–56.

Hershbein, B. and Kahn, L. B. (2018), 'Do recessions accelerate routine-biased technological change? evidence from vacancy postings', *American Economic Review* **108**(7), 1737–1772.

Jones, C. I. and Tonetti, C. (2020), 'Nonrivalry and the economics of data', *American Economic Review* **110**(9), 2819–2858.

Jordan, M. I. and Mitchell, T. M. (2015), 'Machine learning: Trends, perspectives, and prospects', *Science* **349**(6245), 255–260.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D. (2020), 'Scaling laws for neural language models', *arXiv preprint arXiv:2001.08361* .

Kogan, L. and Papanikolaou, D. (2014), 'Growth opportunities, technology shocks, and asset prices', *The Journal of Finance* **69**(2), 675–718.

Lamont, O. A. (1997), 'Cash flow and investment: Evidence from internal capital markets', *The Journal of Finance* **52**(1), 83–109.

Merica, D. and Bedayn, J. (2024), 'As data centers proliferate, conflict with local communities follows', *Associated Press* . Accessed 18 April 2025.
URL: *https://apnews.com/article/data-centers-artificial-intelligence-technology-amazon-google-56b84cbb94942039754282afb076a87b*

Papanikolaou, D. (2011), 'Investment shocks and asset prices', *Journal of Political Economy* **119**(4), 639–685.

Shin, H.-H. and Stulz, R. M. (1998), 'Efficiency of internal capital allocation: Market versus division manager', *The Journal of Financial Economics* **51**(2), 263–289.

Srivathsan, B., Sorel, M. and Sachdeva, P. (2024), 'AI power: Expanding data center capacity to meet growing demand'. Accessed 18 April 2025.
URL: *https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand*

Stein, J. C. (1997), 'Internal capital markets and the competition for corporate resources', *The Journal of Finance* **52**(1), 111–133.

Thomas, D. (2023), 'Private equity bets on AI gold rush with billions pumped into datacenters'. Accessed 18 April 2025.
URL: *https://www.spglobal.com/market-intelligence/en/news-insights/articles/2023/12/private-equity-bets-on-ai-gold-rush-with-billions-pumped-into-datacenters-79666528*

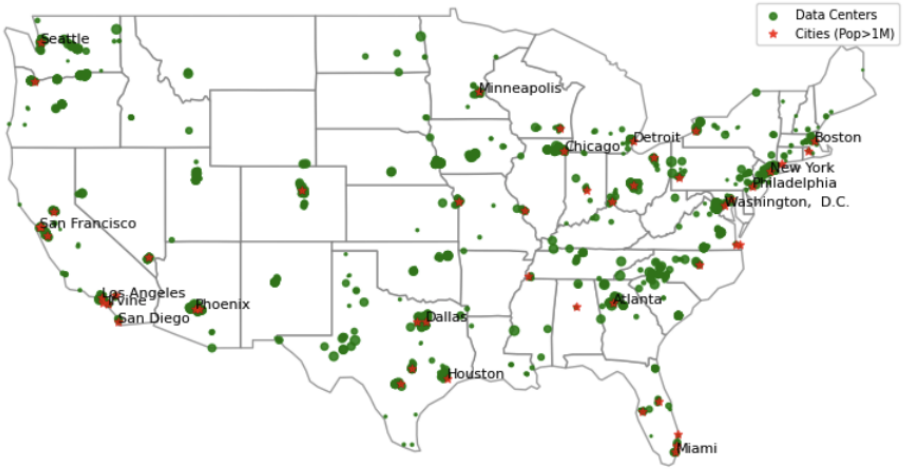
Veldkamp, L. (2005), 'Information markets and the comovement of asset prices', *The Review of Economic Studies* **72**(4), 823–845.

Veldkamp, L. and Chung, C. (2024), 'Data and the aggregate economy', *Journal of Economic Literature* **62**(2), 458–484.

Webb, M. (2019), 'The impact of artificial intelligence on the labor market: evidence from online job postings', *Stanford Working Paper* .

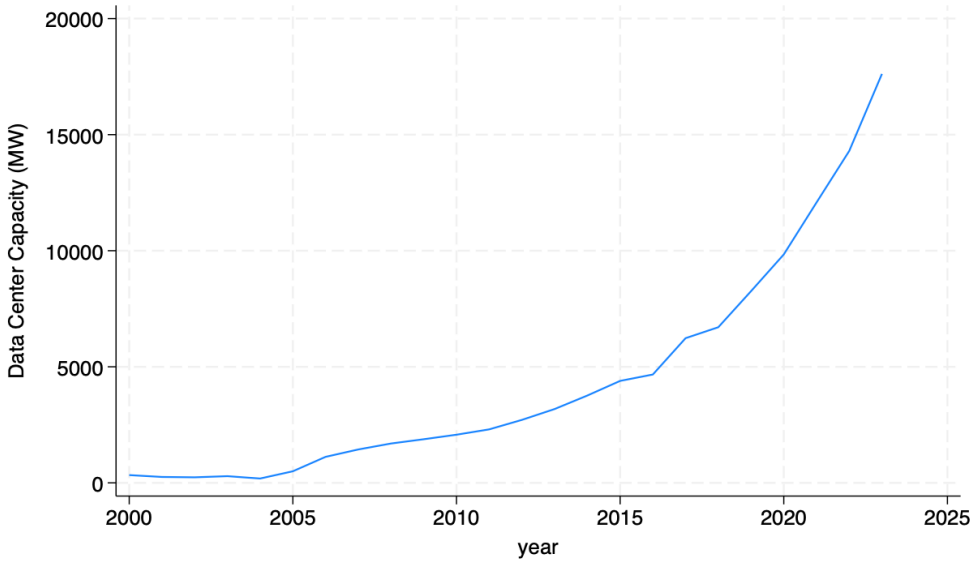
8 Figures

Figure 1: Data Center Locations



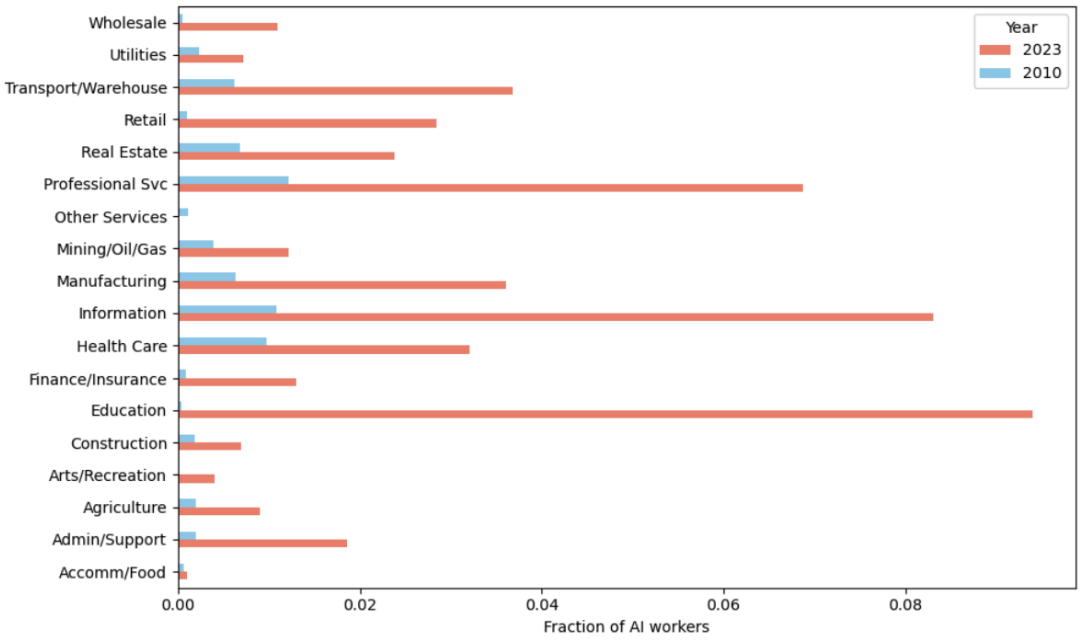
Notes: This figure plots the locations of operating data centers in the US in 2024. The size of the dots are proportional to the log of the capacity of the data center.

Figure 2: Data Center Capacity over Time

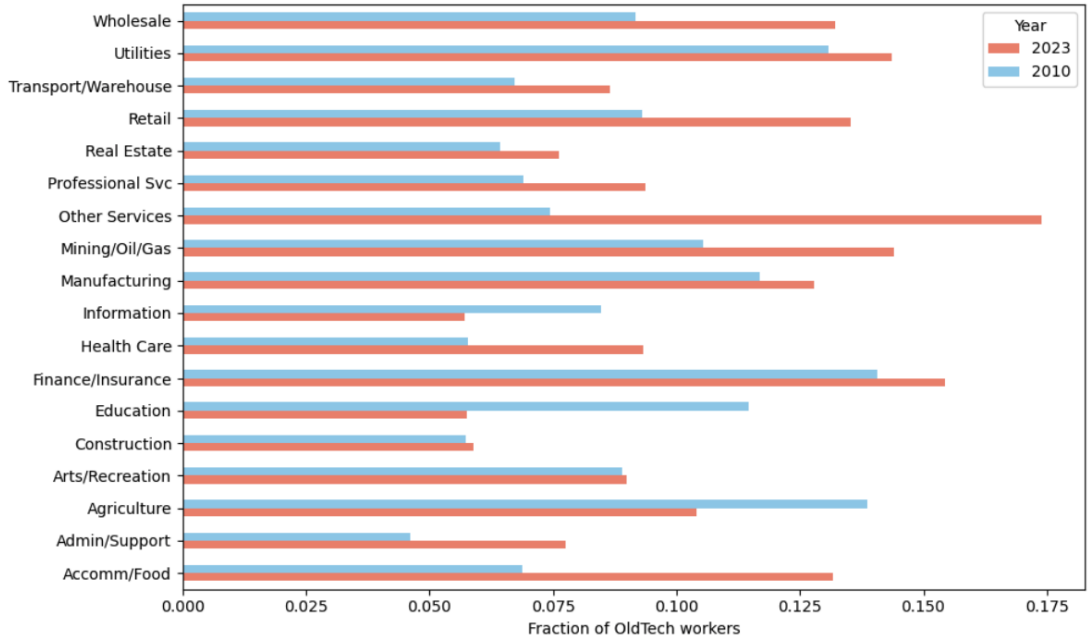


Notes: This figure plots data center capacity (MW) at the national level over time.

Figure 3: Industry-level Average of the Fraction of AI/OldTech Workers



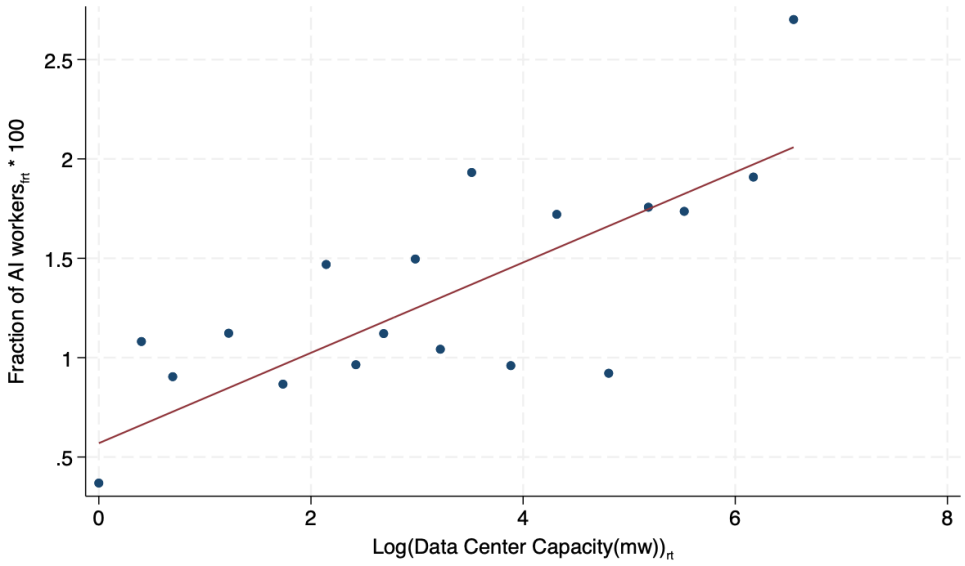
(a) Fraction of AI workers



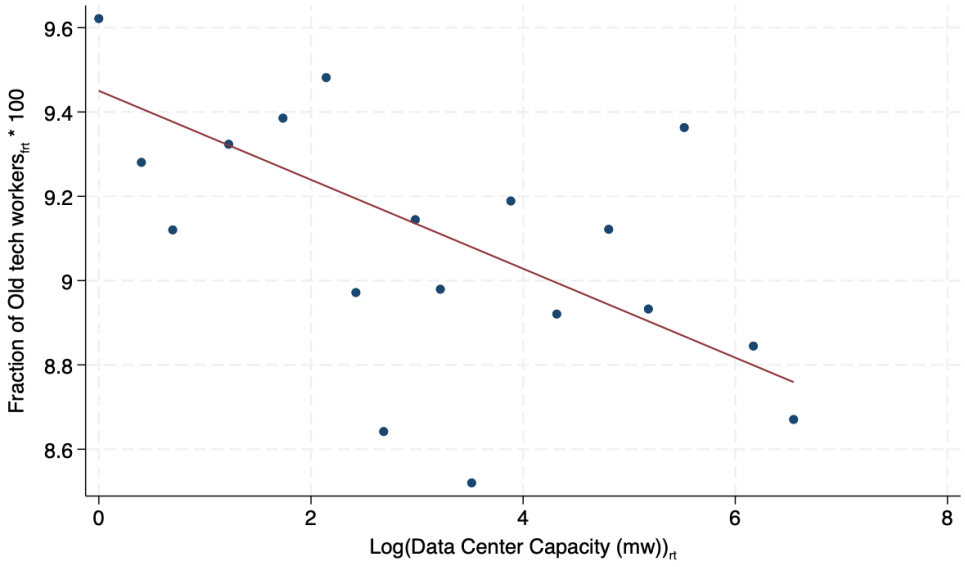
(b) Fraction of OldTech workers

Notes: This figure plots the mean of the establishment-level fraction of AI and OldTech workers at the NAICS2 level.

Figure 4: Establishment-level labor outcomes and data center power capacity



(a) Fraction of AI postings



(b) Fraction of OldTech postings

Notes: Note that the figure excludes CBSAs that never have any data centers for illustration purpose.

Figure 5: Pre-period Slack Share v.s. Growth in Data Center Exposure

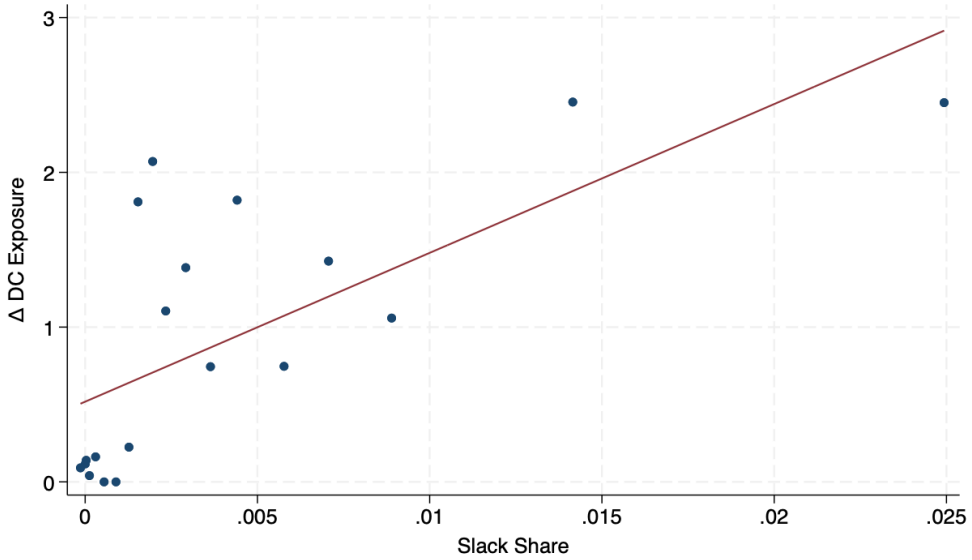
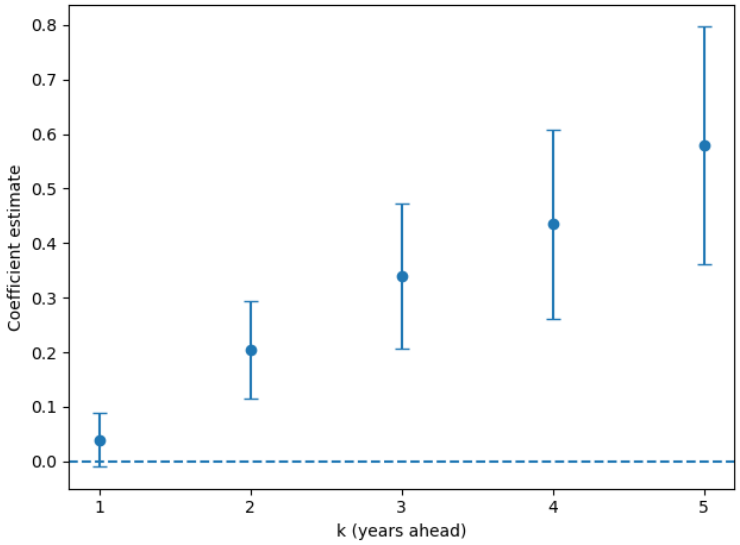
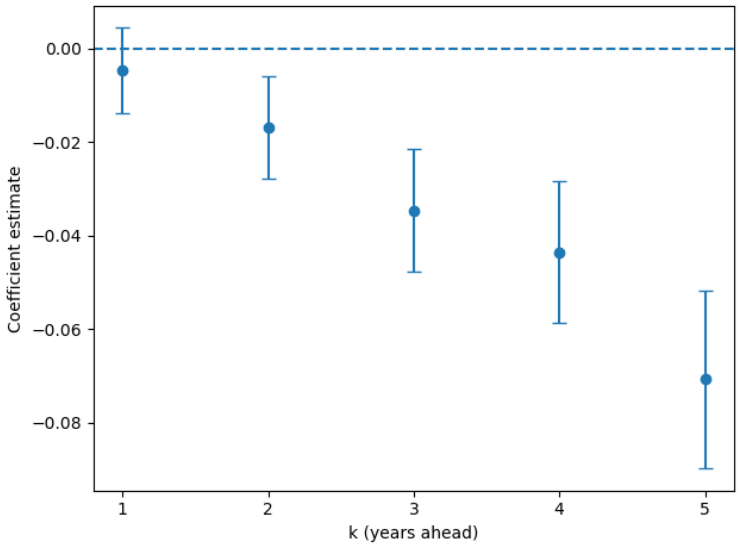


Figure 6: Own-region v.s. Peer-region



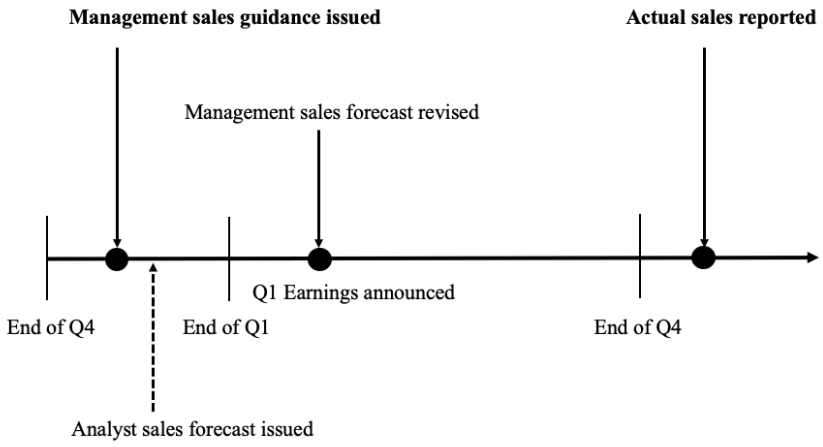
(a) Own-region Effect of the Fraction of AI postings



(b) Peer-region Effect of the Fraction of AI postings

Notes:

Figure 7: Management Guidance Timeline



9 Tables

Table 1: **Summary Statistics**

Panel A: Establishment-Year Level								
Variable	Obs	Mean	Std. Dev.	1%	25%	Median	75%	99%
Fraction of AI workers × 100	1,681,996	0.86	5.99	0	0	0	0	33.33
Fraction of OldTech workers × 100	1,681,996	10.31	24.38	0	0	0	4.76	100.00
Job postings	1,681,996	37.73	271.49	1	2	5	19	541
Job postings/L.EMP × 100	1,143,383	0.37	1.16	0	0.01	0.05	0.20	8.90
Δ Fraction of AI workers × 100	1,150,515	0.12	2.73	-6.25	0	0	0	10.71
Δ Fraction of OldTech worker × 100	1,150,515	0.07	15.19	-57.14	0	0	0	55.56
Δ Job postings/L.EMP × 100	878,727	0	0.37	-1.49	-0.02	0	0.02	1.5
DC Exposure	1,681,996	0.92	1.77	0	0	0	0.69	6.34
Δ DC Exposure	1,150,515	0.05	0.26	0	0	0	0	1.38
DC $\widehat{\text{Exposure}}$	1,681,996	2.07	1.59	0	0.55	2.09	3.27	5.60
Panel B: Establishment-Year long-difference								
Variable	Obs	Mean	Std. Dev.	1%	25%	Median	75%	99%
Δ Fraction of AI workers × 100	43,821	1.36	6.32	-2.40	0	0	0	48.52
Δ Fraction of OldTech worker × 100	43,821	0.97	27.13	-100	0	0	5	95.77
Δ DC Exposure	43,821	0.89	1.63	0	0	0	1.24	6.0
Δ DC $\widehat{\text{Exposure}}$	43,821	1.74	0.82	0	1.45	2.16	2.32	2.41
Panel C: Firm-Year Level								
Variable	Obs	Mean	Std. Dev.	1%	25%	Median	75%	99%

All variables are winsorized at 1%.

Table 2: Impact of Data Center Capacity on Job Postings

Panel A: Fraction of AI Postings × 100							
Exposure	0.172***	0.0893***	0.541***	0.128***	0.0927***	0.0222**	0.122***
	(0.0175)	(0.0111)	(0.0414)	(0.0163)	(0.0175)	(0.0104)	(0.0191)
N	1681996	1681996	1681996	1681996	1681996	1679487	1573510
R^2	0.003	0.355	0.007	0.017	0.363	0.568	0.528
Firm FE		✓			✓		
CBSA FE			✓		✓	✓	
Year FE				✓	✓		✓
Firm × CBSA FE							✓
Firm × Year FE						✓	
$\mathbb{E}$(LHS)	0.856	0.856	0.856	0.856	0.856	0.856	0.845
Panel B: Fraction of Oldtech Postings × 100							
Exposure	-0.392***	-0.0904***	-0.106***	-0.404***	-0.148***	-0.0782**	-0.0867*
	(0.0286)	(0.0175)	(0.039)	(0.0291)	(0.0420)	(0.0343)	(0.0486)
N	1681996	1681996	1681996	1681996	1681996	1679487	1573510
R^2	0.001	0.474	0.006	0.001	0.475	0.584	0.660
Firm FE		✓			✓		
CBSA FE			✓		✓	✓	
Year FE				✓	✓		✓
Firm × CBSA FE							✓
Firm × Year FE						✓	
$\mathbb{E}$(LHS)	10.312	10.312	10.312	10.312	10.312	10.312	10.489

Notes: Standard errors in parentheses. Standard errors are clustered at the CBSA level. $*p < 0.10, **p < 0.05, ***p < 0.01$. The results are robust to clustering at the CBSA-year level. I ran the following regression:

$$\mathrm{Y}_{f,r,t} \;=\; \beta\,\mathrm{DC\ Exposure}_{r,t} \;+\; FEs \;+\; \varepsilon_{f,r,t},$$

where

$$\mathrm{DC\ Exposure}_{r,t} = \log\bigl(1 + \mathrm{Data\,Center\,Capacity}_{r,t}\bigr)$$

and $\mathrm{Y}_{f,r,t}$ is the establishment-level labor outcome (fraction of AI/Oldtech job postings).

Table 3: Impact of Data Center Capacity on Job Postings: Long-difference

	Δ Fraction of AI Postings	Δ Fraction of OldTech Postings
Δ Exposure	0.181***	-0.205***
	(0.019)	(0.080)
N	43,821	43,821
R2	0.0022	0.0002

Notes: Standard errors in parentheses. Standard errors are clustered at the CBSA level. $*p < 0.10, **p < 0.05, ***p < 0.01$. The results are robust to clustering at the CBSA-year level. I ran the following regression

$$\Delta Y_{f,r,2010-2023} \;=\; \beta\, \Delta\text{DC Exposure}_{r,2010-2023} \;+\; \varepsilon_{f,r},$$

where

$$\Delta\text{DC Exposure}_{r,2010-2023} = \log\big(1+\text{Data Center Capacity}_{r,2023}\big) - \log\big(1+\text{Data Center Capacity}_{r,2010}\big)$$

and $\Delta \text{Y}_{f,r,2010-2023}$ is the change establishment-level labor outcome from 2010 to 2023 (fraction of AI/Oldtech job postings).

Table 4: Impact of Data Center Capacity on Job Postings: IV first stage

	DC Exposure						
ElectricitySlack	0.634***	0.563***	0.492***	0.653***	0.499***	0.421***	0.531***
	(0.074)	(0.070)	(0.086)	(0.078)	(0.080)	(0.070)	(0.133)
N	1635333	1635331	1635333	1635333	1635331	1632844	1528677
Kleibergen-Paap F-stat	73.91	65.77	56.53	69.29	38.54	36.14	38.33
Firm FE		✓			✓		
CBSA FE			✓		✓	✓	
Year FE				✓	✓		✓
Firm × CBSA FE							✓
Firm × Year FE						✓	

Table 5: Impact of Data Center Capacity on Job Postings: IV second stage

	Panel A: Fraction of AI Postings × 100						
Exposure	0.408***	0.277***	2.408***	0.209***	0.682***	0.283***	0.811***
	(0.0500)	(0.0366)	(0.314)	(0.0246)	(0.115)	(0.0610)	(0.133)
N	1635333	1635331	1635333	1635333	1635331	1632844	1528677
Firm FE		✓			✓		
CBSA FE			✓		✓	✓	
Year FE				✓	✓		✓
Firm × CBSA FE							✓
Firm × Year FE						✓	
	Panel B: Fraction of Oldtech Postings × 100						
Exposure	-0.863***	-0.224***	-0.651***	-0.867***	-0.444**	-0.181**	-0.440**
	(0.118)	(0.0382)	(0.146)	(0.123)	(0.188)	(0.090)	(0.195)
N	1635333	1635331	1635333	1635333	1635331	1632844	1528677
Firm FE		✓			✓		
CBSA FE			✓		✓	✓	
Year FE				✓	✓		✓
Firm × CBSA FE							✓
Firm × Year FE						✓	

Notes: Standard errors in parentheses. Standard errors are clustered at the CBSA level. $*p < 0.10, **p < 0.05, ***p < 0.01$. The results are robust to clustering at the CBSA-year level. I ran the following regression:

$$\mathrm{Y}_{f,r,t} \;=\; \beta\, \mathrm{DC}\ \widehat{\mathrm{Exposure}}_{r,t} \;+\; FEs \;+\; \varepsilon_{f,r,t},$$

where $\mathrm{DC}\ \widehat{\mathrm{Exposure}}_{r,t}$ is estimated from the first stage.

Table 6: IV: Long-difference estimation

Panel A: First-stage		
	Δ DC Exposure	
Slack share	75.775***	
	(4.194)	
N	43,821	
Kleibergen-Paap F-stat	326.40	
Panel B: Second-stage		
	Δ Fraction of AI Postings	Δ Fraction of OldTech Postings
$\Delta \widehat{DCExposure}$	0.583***	-0.683***
	(0.240)	(0.261)
N	43,821	43,821

Notes: Standard errors in parentheses. Standard errors are clustered at the CBSA level. $*p < 0.10, **p < 0.05, ***p < 0.01$. The results are robust to clustering at the CBSA-year level. The first-stage regression is :

$$\Delta\text{DC Exposure}_{r,2010\text{–}2023} \;=\; \pi_1 \,\text{Slack Share}_{r,2005} \;+\; \text{FEs} \;+\; \eta_r.$$

and the second-stage equation

$$\Delta Y_{f,r,2010\text{–}2023} \;=\; \beta \,\Delta\widehat{\text{DC Exposure}}_{r,2010\text{–}2023} \;+\; \text{FEs} \;+\; \varepsilon_{f,r}.$$

is $\Delta\widehat{\text{DC Exposure}}_{r,2010\text{–}2023}$ is estimated from the first stage and $\Delta \text{Y}_{f,r,2010-2023}$ is the change establishment-level labor outcome from 2010 to 2023 (fraction of AI/Oldtech job postings).

Table 7: Peer Location Effect

Panel A: OLS Results

Δ Fraction of AI ×100

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Peer DC Exposure	-0.00353	-0.0208***	-0.0421***	-0.0613***	-0.0919***
	(0.00458)	(0.00548)	(0.00657)	(0.00752)	(0.00951)
R^2	0.031	0.072	0.093	0.111	0.137
N	973772	824793	699942	592355	494187

Δ Fraction of OldTech ×100

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Peer DC Exposure	0.0287**	0.0755***	0.0628***	0.0163	0.0284
	(0.0113)	(0.0144)	(0.0191)	(0.0206)	(0.0243)
R^2	0.025	0.052	0.085	0.127	0.171
N	973772	824793	699942	592355	494187

Panel B: IV Results

Δ Fraction of AI ×100

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Peer DC Exposure	-0.00476	-0.0170***	-0.0346***	-0.0436***	-0.0707***
	(0.00465)	(0.00558)	(0.00671)	(0.00773)	(0.00971)
N	976980	827712	702563	594696	496263

Δ Fraction of OldTech ×100

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Peer DC Exposure	0.0238**	0.0564***	0.0573***	0.0232	0.0237
	(0.0115)	(0.0145)	(0.0191)	(0.0208)	(0.0246)
N	976980	827712	702563	594696	496263

Table 8: Subsample(tech): Impact of Data Center Capacity on Job Postings

Independent Variable: Fraction of AI Postings × 100

Panel A: Firms with mobile apps							
Exposure	0.273***	0.139***	0.632***	0.230***	0.191***	0.0410**	0.223***
	(0.0269)	(0.0191)	(0.0512)	(0.0263)	(0.0297)	(0.0178)	(0.0307)
N	756063	756063	756063	756063	756063	755839	722317
R^2	0.006	0.327	0.013	0.018	0.338	0.552	0.497
Panel B: Firms with no mobile apps							
Exposure	0.107***	0.0562***	0.476***	0.0602***	0.0269*	0.00744	0.0359**
	(0.0130)	(0.00698)	(0.0372)	(0.0116)	(0.0141)	(0.0102)	(0.0144)
N	911190	911190	911190	911190	911190	908952	837769
R^2	0.001	0.379	0.006	0.017	0.387	0.583	0.556
Panel C: Tech firms (NAICS2 = 51)							
Exposure	0.421***	0.226***	1.306***	0.318***	0.222***	0.0420	0.372***
	(0.0473)	(0.0345)	(0.117)	(0.0458)	(0.0545)	(0.0529)	(0.0705)
N	133139	133139	133139	133139	133139	132982	121137
R^2	0.008	0.321	0.025	0.039	0.350	0.585	0.497
Panel D: Non-Tech firms (NAICS2 != 51)							
Exposure	0.134***	0.0745***	0.459***	0.0954***	0.0773***	0.0191**	0.0971***
	(0.0138)	(0.00910)	(0.0340)	(0.0124)	(0.0155)	(0.00931)	(0.0162)
N	1534114	1534112	1534114	1534114	1534112	1531809	1437772
R^2	0.002	0.363	0.005	0.015	0.371	0.564	0.540
Firm FE		✓			✓		
CBSA FE			✓		✓	✓	
Year FE				✓	✓		✓
Firm × CBSA FE							✓
Firm × Year FE						✓	

Notes: Standard errors in parentheses. Standard errors are clustered at the CBSA level. $*p < 0.10, **p < 0.05, ***p < 0.01$. The results are robust to clustering at the CBSA-year level. I ran the following regression:

$$\mathrm{Y}_{f,r,t} \;=\; \beta\,\mathrm{DC\ Exposure}_{r,t} \;+\; FEs \;+\; \varepsilon_{f,r,t},$$

where

$$\mathrm{DC\ Exposure}_{r,t} = \log\bigl(1 + \mathrm{Data\,Center\,Capacity}_{r,t}\bigr)$$

where $\mathrm{Y}_{f,r,t}$ is the establishment-level labor outcome (fraction of AI/Oldtech jobs). Firms with mobile apps are defined statically as firms that have released at least one mobile application at any point.

Table 9: Firm-level Outcomes

	Δ Log(sale)	Δ Log(assets)	Δ Log(emp)
Δ DC Exposure	0.0562^{***}	0.0639^{***}	0.0480^{***}
	(0.0161)	(0.0181)	(0.0153)
N	1230	1239	1225
R2	0.0647	0.0399	0.0573

Notes: Standard errors in parentheses. Standard errors are clustered at the CBSA level. $*p < 0.10, **p < 0.05, ***p < 0.01$. I ran the following regression:

$$\Delta Y_f \;=\; \beta\, \Delta\text{DC Exposure}_f \;+\; \varepsilon_f, \quad \text{where} \quad \Delta Y_f = log(Y_{f,t_1}) - log(Y_{f,t_0}).$$

Table 10: Management Guidance Accuracy

	$\mathbb{1}$ {Accurate Forecast in 2023}		$\mathbb{1}$ {Accurate Forecast in 2013}	
Δ Exposure	0.045**	0.053***	-0.004	-0.002
	(0.022)	(0.017)	(0.028)	(0.022)
N	382	378	350	347
R2	0.011	0.054	0.000	0.068
NAICS2 FE		✓		✓

I ran the following regression:

$$\mathbb{1}\{AccurateForecast\}_f \;=\; \alpha \;+\; \beta\,\Delta\text{Exposure}_f \;+\; \varepsilon_f,$$

where the independent variable is a dummy variable = 1 if realized sales falls within the range of management guidance.

Table 11: Effect of Electricity Exposure on Establishment-Level Job Postings

	Fraction AI Postings			Fraction OldTech Postings		
	(1)	(2)	(3)	(1)	(2)	(3)
Electricity Exposure	-0.0652***	-0.0259***	-0.0750***	0.0206	-0.0188	0.0025
	(0.0115)	(0.0078)	(0.0125)	(0.0398)	(0.0363)	(0.0473)
$R2$	0.363	0.568	0.528	0.463	0.577	0.637
N	1681996	1679487	1573510	1681996	1679487	1573510
Firm FE	✓			✓		
CBSA FE	✓	✓		✓	✓	
Year FE	✓		✓	✓		✓
Firm × CBSA FE			✓			✓
Firm × Year FE		✓			✓	

Notes: Standard errors in parentheses. Standard errors are clustered at the CBSA level. $*p < 0.10, **p < 0.05, ***p < 0.01$. The results are robust to clustering at the CBSA-year level. I ran the following regression:

$$\mathrm{Y}_{f,r,t} \;=\; \beta\,\text{Electricity Exposure}_{r,t} \;+\; FEs \;+\; \varepsilon_{f,r,t},$$

where

$$\text{Electricity Exposure}_{r,t} = \log\big(1 + \text{Electricity Capacity}_{r,t}\big)$$

and $\mathrm{Y}_{f,r,t}$ is the establishment-level labor outcome (fraction of AI/Oldtech jobs).

A Variable Construction

A.1 Keywords

Old Technologies Keywords: Linear Regression, Logistic Regression, Statistics, STATA, Emacs, Technical Analysis, Qualitative Analysis, Qualitative Portfolio Management, Data Trending, Stochastic Optimization, Multivariate Testing, Bootstrapping, Factor Analysis, Duration Analysis, Markov, HMM, Econometrics, Stochastic Processes, Calculus, Statsmodels, Linear Algebra, Mathematics, Monte Carlo Simulation, Generalized Linear Model, GLM, Linear Programming, Bayesian, Analysis of Variance, ANOVA, Behavioral Modeling, Black-Scholes, Behavioral Analysis, Discounted Cashflow, Correlation Analysis, E-Views, Differential Equations, Value at Risk, Asset Pricing Models, Statistician, Mathematician, Econometrician.
AI Keywords: Artificial Intelligence, Machine Learning, Natural Language Processing, NLP, Speech Recognition, Gradient Boosting, DBSCAN, Nearest Neighbor, Supervised Learning, Unsupervised Learning, Deep Learning, Automatic Speech Recognition, Torch, scikit-learn, Conditional Random Field, TensorFlow, Platfora, Neural Network, CNN, RNN, Decision Trees, Random Forest, Support Vector Machine, SVM, Reinforcement Learning, Lasso, Stochastic Gradient Descent, SGD, Ridge Regression, Elastic-Net, Text Mining, Classification Algorithms, Image Processing, Natural Language Toolkit, NLTK, Pattern Recognition, Computer Vision, Long Short-Term Memory, LSTM, K-Means, Geospatial Intelligence, Big Data Analytics, Latent Dirichlet Allocation, LDA, Backpropagation, Machine Translation, Caffe Deep Learning Framework, Word2Vec, Genetic Algorithm, Evolutionary Algorithm, Data Science, Sentiment Analysis, Maximum Entropy Classifier, Neuroscience, Computational Linguistics, Semi-Supervised Learning, Data Scientist.

B Tables

Table 12: IV reg: demo controls

	Fraction AI Postings			Fraction OldTech Postings		
	(1)	(2)	(3)	(1)	(2)	(3)
DC Exposure	0.623***	0.238***	0.730***	-0.519**	-0.0809**	-0.510*
	(0.130)	(0.0725)	(0.143)	(0.263)	(0.040)	(0.268)
log(population)	-0.305	-0.0921	-0.388	0.303	0.549	0.196
	(0.244)	(0.119)	(0.273)	(0.523)	(0.444)	(0.504)
log(median income)	0.592	0.530*	0.676	-2.670***	-1.582**	-3.354***
	(0.540)	(0.274)	(0.604)	(0.863)	(0.676)	(0.899)
log(median home value)	-0.862**	-0.479***	-0.884**	1.114**	0.137	1.595***
	(0.340)	(0.165)	(0.377)	(0.511)	(0.401)	(0.506)
% comp occupations	0.434*	0.280**	0.523**	0.179	0.0398	0.173
	(0.233)	(0.124)	(0.247)	(0.294)	(0.247)	(0.292)
N	1621718	1619222	1515554	1621718	1619222	1515554

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C Proofs

Proof of Lemma 1

Differentiating $R_i = s_i z_i / Z$ w.r.t. D_k gives

$$\frac{\partial R_i}{\partial D_k} = \frac{s_i}{Z^2}\big(Z\,\partial z_i/\partial D_k - z_i \sum_{n=1}^{N} s_n\,\partial z_n/\partial D_k\big).$$

Insert $\partial z_n/\partial D_k = \varepsilon_{kn} z_n / D_k$ and collect terms to obtain the stated formulas.

Proof of Proposition 1

Under Case A, Lemma 1 implies $\partial R_i/\partial D_i > 0$. Both the share term and the budget term R_i/b in $\partial L_i^*/\partial D_i$ are positive, hence the total derivative is positive.

Proof of Proposition 2

For $j \neq i$, Case A gives $\partial R_i/\partial D_j < 0$ while the budget term R_i/b is positive. The sign of their sum is therefore indeterminate.

Proof of Proposition 3

With $\varepsilon_{kn} = \varepsilon$ for all k, n, Lemma 1 yields $\partial R_i/\partial D_k = 0$. Consequently $\partial L_i^*/\partial D_k = R_i/b > 0$ for every i, k.