Growth, Transformation and Digital Capital: The Importance of Technological and Organizational Architecture

Ruiqing Cao
Marco Iansiti
Growth, Transformation and Digital Capital: The Importance of Technological and Organizational Architecture

Ruiqing Cao
Harvard Business School

Marco Iansiti
Harvard Business School

Working Paper 21-122
Growth, Transformation and Digital Capital: The importance of technological and organizational architecture

Ruiqing Cao
Harvard Business School
rcao@hbs.edu

Marco Iansiti
Harvard Business School
miansiti@hbs.edu

June 2021

Abstract

The benefits to data analytics and machine learning have been distributed unevenly across firms around the world. Research on IT productivity points to intangible capital as a crucial driver of value creation from innovation in computing. We argue that an important component of intangible capital is organization-wide technological architecture, which is particularly complex and idiosyncratic for large corporations. We describe a novel survey instrument to quantify the closeness of firms’ data architecture capabilities to the “best practices” of successful digital companies. Using the prevalence of third-party maintenance of data center servers as a proxy for legacy IT and an instrument for data architecture quality, we find that a 1 standard deviation improvement in data architecture leads to 15% higher productivity growth from 2016 to 2019, through increasing the intensity of machine learning deployment and reducing the costs to co-invention by technical workers. The effects of data architecture are larger among corporations with more complex IT systems, measured by the share of system software categories.

1 We are grateful to Alexandria Sheng, Ellen Wu, and the Keystone Strategy team for implementing the survey and collecting the data. We thank Tim DeStefano, Shane Greenstein, Frank Nagle, and Glen Weyl for helpful comments.
1 Introduction

Since the early 2000s, we have witnessed a decline in average firm productivity growth and a substantial increase in market concentration across the U.S. economy (Decker et al., 2014; Bessen, 2017; Van Reenen, 2018; Akicigit and Ates, 2019). At the same time, firm performance diverged across sectors and geography, and some “superstar” firms appeared to be capable of sustaining superior performance (Ayyagari et al., 2019; Covarrubias et al., 2019). Some have argued that information technology (IT) systems contributed significantly to these trends (Bessen, 2017). More recently, several authors in practitioner-oriented literature proposed the contribution of data, analytics, and artificial intelligence to economic growth (e.g., Brynjolfsson and McAfee, 2014; Iansiti and Lakhani, 2020).

Technological progress and innovation in data, analytics, and AI can play a large role in driving competitive dynamics among firms. However, the variance in IT productivity across firms appears too large to be explained by technology use alone. The missing contributor to productivity has been linked to complementary investments in intangible assets, e.g., skills training, organizational structure related to decentralized decision-making, management practices, and software customization (Brynjolfsson, Hitt and Yang, 2002; Tambe et al., 2020). These intangible assets are sometimes termed “digital capital” and have been posited to play a crucial role in realizing business value from information technology (Black and Lynch, 2001; Sculley et al., 2014; Crouzet and Eberly, 2018; Brynjolfsson, Rock and Syverson, 2021).

But what, specifically, constitute these intangible assets and how do we measure them directly? In this paper, we identify architecture as a key component of intangible digital capital that drives the productivity gains from data and machine learning technologies and describe a survey instrument to quantify it directly. The survey collects comprehensive information about firms’ closeness to frontier practices of the most successful digital companies – a framework we call “AI Intensity Scorecard”, through in-person interviews with senior executives familiar with corporation-wide technology initiatives. We transform the survey responses into several indexes measuring components of digital capital, particularly data architecture and innovation processes.

The sampling frame focuses on large traditional corporations that existed from decades to centuries before the 1990s’ IT boom. They had very different operating models from digital companies founded after the 2000s (Iansiti and Lakhani, 2020). The survey sample consists of 117 U.S.-based corporations, in a few traditional sectors including manufacturing, retail trade, financial services, and health care. Total revenues of sample corporations sum up to $4.3 trillion in 2019. Their siloed technology systems and business processes pose challenge to digital transformation.

We use an instrumental variable to identify quasi-random variation in the capabilities of an organization’s technological architecture, benchmarked against “best practices” of the most successful digital companies. The core idea of the instrument is that legacy components in a firm’s information technology (IT) systems hinder
its ability to set up architecture that can facilitate effective deployment of data tools and machine learning models. The instrument counts the number of establishments across the firm with third-party maintenance (TPM) for data center servers. Third-party maintenance (TPM) is identified when an establishment keeps outdated data center equipment running after its end-of-service-life date, no longer supported by the original equipment manufacturer (OEM).

We show that server TPM is not correlated with other firm characteristics that may directly affect revenue growth and technical innovation, and especially cloud computing adoption. After accounting for traditional IT capital measures and production factors, each 1 standard deviation increase in the data architecture index leads to 15% higher productivity growth from 2016 to 2019. The estimate is statistically significant at the 5% level and remains after controlling for the use of advanced tools, IT capital stock, and other firm characteristics, e.g., log number of establishments, log company age, and log total enterprise revenue in the baseline year.

How does data architecture lead to productivity differences among corporations? Technical innovation does not generate economic benefits by itself but requires engineers and business users to deliver value through interacting with technological systems and creating co-invention outputs (Bresnahan and Greenstein, 1996). Complementary investments into setting up data architecture with frontier capabilities allow workers to use data and build predictive models effectively, leading to greater machine learning development and deployment intensity within the firm. The survey-based machine learning (ML) deployment index measures the prevalence of ML applications in a variety of customer-facing functions and internal business processes such as production, marketing, supply chain, and employee management. Using the same instrument about legacy IT, we estimate that a 1 standard deviation improvement in data architecture capabilities leads to about 3 – 4 more ML use cases being developed and productized across the corporation.

The survey-based index measures data architecture quality, which increases in the number of capabilities associated with practices of frontier digital companies. Higher-quality data architecture reduces costs and frictions of machine learning co-invention, by making it easier for engineers, data scientists, and business analysts to query data and iterate on predictive models. Technological systems, architecture, and processes are idiosyncratic, and they adapt jointly to suit a firm’s specific business model and operational needs. It is more difficult to integrate new technologies into more complex systems, and workers transition more slowly into dealing with newly incorporated technologies in such systems.

Data architecture can be especially important for firms with more convoluted IT systems. We measure software system complexity with a variable that proxies for the prevalence of system software categories as a fraction of overall software categories across the corporation. Empirical estimates suggest that legacy IT components primarily reduce co-invention output and productivity growth for corporations with particularly complex software systems – an additional establishment reporting third-party maintenance (TPM) for data...
center servers leads to 0.08 lower ML deployment index and 7.2% lower productivity growth. Software system complexity is not correlated with data architecture or ML deployment per se. Taken together, these results suggest that data architecture (driven by legacy IT components) is more crucial to large organizations with more complicated IT systems.

Finally, we do not find evidence that data architecture best practices adversely affect employment size at the aggregate level. Workers do not appear to be displaced by frontier data architecture capabilities, and firms effectively deploying machine learning do not seem to cut more jobs. However, we do not differentiate among tasks and skills required by different jobs, nor do we consider evolving task content and new categories of work being created to accompany technical innovation. We leave these possibilities to future work.

1.1. Literature Review

This paper contributes to the IT productivity literature, especially around the complementary role of digital intangible capital to visible IT investments (Brynjolfsson, Hitt and Yang, 2002; Bartel, Ichniowski and Shaw, 2007; Tambe and Hitt, 2012; Crouzet and Eberly, 2018; Tambe et al., 2020). While past works primarily infer complementary intangibles’ quantities from financials (Brynjolfsson and Yang, 1997; Hall, 2001; Ewens, Peters and Wang, 2019; Tambe et al., 2020; Brynjolfsson, Rock and Syverson, 2021), they do not identify exact areas in which companies should devote resources and develop capabilities. Also, the literature does not distinguish between different generations of IT, e.g., the 1990s’ information processing equipment and computer software may require complementary organizational assets that are distinct from 2010s’ big data and machine learning technologies. We identify an important component of digital intangible capital for the latter as data architecture (Hannan and Freeman, 1984; Henderson and Clark, 1990; Sculley et al., 2014; Gans, 2016; Iansiti and Lakhani, 2020), and directly quantify it with a detailed survey measuring the extent to which a firm’s data architecture possesses the right types of capabilities of frontier digital companies.

This paper adds to a strand of literature on market concentration in the macroeconomy (Corrado, Hulten and Sichel, 2009; Ayyagari, Demirguc-Kunt and Maksimovic, 2019; Covarrubias, Gutiérrez and Philippon, 2019), and rising inequality driven by firm-specific implementation of IT systems (Black and Lynch, 2001; Bessen, 2017; Bessen, 2019; Acemoglu et al., 2020; Zolas et al., 2021). We identify several components of IT systems that lead to firms’ varying ability to derive value from data sets and machine learning tools. We show that outdated data center hardware (i.e., servers) constrain firms’ ability to build frontier architecture capabilities, and slow down productivity growth, particularly among corporations with complicated software systems.

The paper also speaks to the literature on survey methods quantifying within-firm practices that contribute to productivity, firm performance, and other organizational outcomes (Bloom, Sadun and Van Reenen, 2012; Bloom et al., 2012; Bloom et al., 2014; Bloom, Sadun and Van Reenen, 2016; Scur et al., 2021). The survey approach yields high-quality data and accurate measures that generalize across contexts, such as countries,
industries, and business models. While the frontier literature on using surveys to quantify firm-level practices focus on management, this paper proposes a similar approach to quantify technological and organizational architecture. The survey method distills the “best practices” from among the most successful organizations and generates indexes that capture the closeness of a firm’s technological architecture to frontier capabilities. We find comparable magnitudes of data architecture’s impact on productivity growth, relative to this literature that quantifies the importance of managerial quality using similar survey methods.

2 Theory Development

Past literature found that computer capital was associated with far greater market value compared to its input share (Brynjolfsson, Hitt and Yang, 2002; Tambe et al., 2020) and that it was not due to short-run asset mispricing or adjustment costs, but rather due to intangible capital accompanying observed investments in IT capital. Such intangible capital correlated with investments in IT is termed “digital capital”, which generates service flows like other types of intangible capital despite not being recorded on the balance sheet. Examples of digital capital assets include firm-specific worker knowledge and skills training around customized IT systems, new organizational structures and business processes required to support the use of idiosyncratic IT – including decentralized decision making, team-oriented organization of production, communication patterns and information sharing across functions, reporting and incentive systems, and corporate culture (Bresnahan, Brynjolfsson and Hitt, 2002; Brynjolfsson, Hitt and Yang, 2002; Tambe et al., 2020).

Digital intangible investments typically accompany the rise of new technologies and inventions that alter the costs or efficiency of existing tasks, e.g., lowering computing costs. Changes in digital capital often seem highly related to changes in business processes involving workers – e.g., how they communicate and interact with new tools, and organize themselves around production activities. These complementary intangibles appear slow to change, and very difficult to diffuse across firms – one firm typically cannot gain much from copying the exact practices of another firm.

IT-related complementary intangibles are firm specific, and hence measuring them directly is difficult. The predominant approach in the literature is to use financial market information to quantify these intangible investments and their contribution to productivity (Brynjolfsson and Yang, 1997; Rock and Syverson, 2020; Tambe et al., 2020). Such an approach, however, does not distinguish between different components of “digital capital” in aggregate. Another issue is that past studies around complementary intangibles focus on software services and computer hardware (Brynjolfsson and Yang, 1996), rather than new generations of cutting-edge IT – e.g., big data, cloud computing, and artificial intelligence, which may require a different set of intangible complementary investments from the past.

Two broad themes arise as we dig deeper into understanding what exactly digital capital is, when it comes to the effective adoption of big data and machine learning technologies. First, an important component of
digital capital should depend on technical and organizational architecture (Iansiti and Lakhani, 2020; Henderson and Clark, 1990), and entails structures and designs to facilitate real-time smooth flows of data between the system and users and combining data sources across locations at scale. Modularity in data assets is a very useful quality (Iansiti and Lakhani 2020), enabled by well-designed technological architecture. The rest of the system remains intact after some part of the system (e.g., one of the data sets) is changed, and the architecture incorporates such changes quickly and update in real time (Agarwal, Gans and Goldfarb, 2021). Second, digital capital consists of complementary processes that involve human capital and firm-specific skills (Bresnahan, Brynjolfsson and Hitt, 2002; Brynjolfsson, Hitt and Yang, 2002). Over time, workers figure out the processes for themselves and adapt to firm-idsyncratic implementation of new technologies (Bresnahan and Greenstein, 1996).

There is not an easy-to-implement, identical blueprint for setting up the architecture to facilitate the effective use of data and machine learning, or a uniform set of innovation practices that should be adopted by all firms. Instead, the design and implementation of an effective architecture is subject to specific business needs and traditional organizational routines that constrain the way in which companies incorporate digital technologies into their operating models, products, and services. This makes it difficult to carry over the practices of one firm to another. As we know from many examples, it takes substantial skill and focus to enable effective digital innovation across the organization (e.g., Hinterhuber, Vescovi and Checchinato, 2021).

**Hypothesis #1.** Firms vary largely in their effectiveness of deriving useful applications from data and machine learning, and as a result, creating productivity gains from adopting new technologies.

Among large established corporations, the benefits of new technology are thus slow to accumulate or diffuse. Also, architecture does not scale linearly in the amount of money spent on technical infrastructure and equipment purchases. We describe a survey instrument designed and implemented in 2020 to quantify the closeness of a corporation’s technological architecture to frontier practices of the most successful digital companies around data and machine learning. In broad strokes, an architecture with the full set of frontier capabilities should be able to process data streams flowing through multiple layers of pipelines, combine data sets across various sources, and optimize machine learning applications at scale.

What factors matter to firms’ ability to build data architecture around best practices of digital firms? Legacy IT systems lead to variation in data architecture design and implementation, and hinder firms’ ability to re-architect around frontier practices. Established large corporations are particularly likely to have convoluted legacy systems, after a long strand of mergers, spinoffs, and headquarter migrations throughout company history. The IT systems of these companies are often complex and incompatible across locations. Many years of localized investments have led to over-specialized production units and incoherent technological systems and data silos that inhibit integration with new technology.
Legacy IT systems include hardware and other technological components that are out of date and in need of upgrade or replacement, but somehow still used in core business functions and day-to-day operations. Maintenance costs for legacy systems increase over time, and more patches, compatibility layers, and increasingly intricate modifications are required as time goes by and vendor support expires. Such systems eventually become incompatible with new technologies and cannot integrate with or accommodate new tools and products when business needs change, which often requires the most updated hardware versions to run smoothly or perform normally. They also exhibit higher security and policy compliance risks. Functional users and business departments maintain separate legacy systems that create isolated data sets stuck in their silos. Many older systems were never designed to interact with each other in the first place. This poses barriers to digital transformation and limits the data processing capabilities of the technological infrastructure.

Hypothesis #2. Legacy components of existing IT systems contribute to architectural inertia and hinder corporations’ ability to accumulate the right types of digital capital to facilitate the effective use of data and machine learning technologies.

Outdated hardware thus limits the system’s ability to integrate with new technology and raises barriers to re-architecting around a data-centric operating model.

Neither technology nor architecture alone creates value without involving workers. Human capital is a crucial driver of “co-invention” (Bresnahan and Greenstein, 1996) – while increasingly powerful processors lower costs to large-scale computing and expand the possibilities for data and algorithms, technological progress does not lead to economic value without technical workers and business users interacting with technological systems to develop solutions, create output and turn these possibilities into reality.

Data and machine learning deliver value when engineers and data scientists adapt to firm-specific configurations of the data architecture to combine data sets and develop predictive models. These workers’ repeated choices, interactions, and routines shape processes that over time become embedded in the system and lower costs to carrying out job tasks. These processes are complementary assets that are truly valuable in turning technical progress into productivity gains.

Co-invention has proven to be difficult, with computer technology in the 1990s, and now with AI, machine learning, and big data. While basic technologies may be relatively widely available, the right combination of technological systems and skilled workers that lead to valuable co-invention is much harder to identify and implement. Co-invention is uncertain and difficult to create, requiring generations of workers repeatedly experimenting with idiosyncratic systems and distinct operating models. It does not diffuse easily, because organizational knowledge generated from co-invention processes tends to be firm-specific and lose value once workers depart from the firm. Legacy systems slow down co-invention and pose barriers to realizing the productivity gains from data and machine learning.
Data architecture closer to frontier practices minimizes costs and frictions around workers co-inventing around data and machine learning. Co-invention tends to be dynamic and responds to changes to the technological systems continuously, e.g., incorporating software and hardware updates, new data sets, and machine learning tools. Workers adapting to such changes in the systems incur adjustment costs, which can be minimized when they interact with the systems through data architecture that is designed to make such adjustments easy. For example, an important frontier capability of the data architecture involves being end-user friendly, e.g., having semantic layers and clean documentation that allow less technical users to easily access data sets and tools. Additionally, modular designs allow data and tools to populate across locations frictionlessly, so that workers across functional units can incorporate the latest changes in real-time.

Data architecture quality is particularly important when existing technological systems are convoluted and challenging to work with. More complex systems host a wider variety of data sets and software applications. Integrating new data sets and software updates into such systems can be more difficult, and workers will find it more daunting to adapt to changes in the systems. For example, relatively fewer technical employees can debug complex software code originally written in low-level programming languages (e.g., in many system software categories). A well-designed architecture lowers the costs and frictions associated with user co-invention even more, when workers interact with relatively complex systems.

**Hypothesis #3.** *A data architecture more distant from frontier practices lowers co-invention quality – deployment intensity of machine learning applications, and consequently missed productivity benefits from technology. Data architecture with capabilities closer to frontier practices of a data-centric operating model (e.g., coherent, and modular) reduce worker co-invention costs, particularly when existing software systems are more convoluted and hence adjustment costs associated with co-invention and applying changes to the system are higher.*

In other words, data architecture is complementary to software system complexity in facilitating co-invention.

Large corporations have IT systems of various degrees of complexity, as these corporations have unique operating models, and over time developed especially complex varieties of business processes around corporation-specific needs. As processes evolve together with idiosyncratic technologies and architectures, more complex software systems also result in more complicated processes.

3 Data and Measurement

In this section, we introduce a survey framework quantifying intangible capital that is complementary to investments in data and machine learning. The framework provides detailed measurement of organization-wide architecture and processes – intangible assets crucial to realizing the value of technical innovation, with particular emphasis on *data architecture.*
3.1 “AI Intensity Scorecard” Survey: Sampling Frame and Data

In 2020, we worked with teams at Keystone Strategy LLC\(^2\) and Microsoft Corporation to design and implement a comprehensive survey, which elicits detailed information about companies’ organizational capital and technological capabilities around data and machine learning. The survey consists of 153 questions, designed to assess firms’ closeness to frontier practices of around AI adoption.

The sampling frame targets established large corporations that operate in traditional sectors of the economy. The team reached out to Fortune 1000 companies that had operated many decades or centuries before the IT-boom of the 1990s to administer the survey.\(^3\) Many of these corporations had long histories dating back to the 1800s, and experienced mergers, acquisitions, spinoffs, name changes, and headquarter migration events throughout time. They tend to have complex and aging technological systems, which are intertwined and difficult to change or remove. The combined total revenue of all corporations in the final data sample amounts to about $4.3 trillion. They account for a substantial fraction of the total employment and market value generated by all businesses in the United States.\(^4\)

[ Insert Appendix Figure B.1 here ]

The research team conducted interviews with 1 or 2 senior technology executives, specifically targeted because of their company-wide knowledge of IT systems. The interviews were performed between June and September 2020. All survey respondents reported directly to the corporate or U.S. divisional headquarters. While 84% headed IT departments or data analytics, the rest performed senior roles central to the deployment of data centric innovations. Each interview lasted about an hour, and focused on enterprise-level questions, hence covering the entire corporation rather than specific establishments or subdivisions.

3.2 Survey-Based Indexes of Technological Intensity

We derive several indexes to quantify aspects of AI intensity from the survey. These indexes measure the extent to which firms possess frontier capabilities by adopting the best practices of successful digital companies. We start by coding survey responses as binary variables, since most questions can be answered by “yes” or

---

\(^2\) Keystone Strategy is a U.S.-based innovation strategy and economic consulting firm, working with global brands and law firms on leading-edge challenges in technology and business.

\(^3\) Fortune 1000 is an annual publication that ranks the largest U.S. corporations by revenue. Sampling took two phases: around 200 companies were contacted in phase #1, and 86 of them were interviewed; around 100 additional companies in the manufacturing sector were contacted in phase #2, and 32 of them were interviewed. The sampling frame excludes technology companies that operate around the delivery of digital services, i.e., platform companies in IT-using industries (and does not include hardware manufacturing).

\(^4\) Appendix Figure B.1 summarizes the Fortune 1000 rankings of corporations in the final data sample in 2020. Fortune 500 (i.e., top half of the Fortune 1000) represent approximately two-thirds of U.S. Gross Domestic Product. Appendix Table B.1 summarizes, by sector, the share of sample firms among all corporations on the Fortune 1000 list in 2020. Overall, the survey sample includes 14% of the Fortune 1000 firms in relevant industries, but 35% of total revenue, and 43% of total assets. The market values of the sample firms account for 63% in health care, 36% in financials, and 20% in retailing among Fortune 1000 corporations. In manufacturing sectors (e.g., Aerospace & Defense, Household Products, Apparel, and Industrial), the sample firms account for around a quarter of total counts, but two-thirds to three-quarters of total market value.
“no”, except only a few questions with three options which we combine appropriately. We average these binary variables across questions grouped around an underlying concept to derive an index, which can be interpreted as the share of “yes” responses to relevant questions.\(^5\) There are three broad categories of questions.

The first category of survey questions relates to complementary assets that facilitate data use and machine learning deployment – i.e., architecture and processes that cover the entire corporation. To use consistent taxonomy, we call the two indexes: Data Architecture and Innovation Processes associated with digital capital. The survey questions on data architecture are particularly comprehensive, hence enable us to measure the architectural component of digital capital in detail.

The overall Data Architecture index quantifies a corporation’s technological architecture that support the ingestion, transformation, and delivery of data around various locations across the entire organization. It consists of two distinct stages around processing and delivering data, for which we derive separate sub-indexes: A Data Platform covering the sourcing, ingestion, and storage of data sets; a Machine Learning (ML) System training and productizing predictive models using input data.

There are two distinct aspects to data architecture quality, for which we derive separate sub-indexes. Coherence, associated with efficiency and speed, requires the ability to automatically integrate data flows into pipelines across functions and business units. It facilitates data access among engineers and business users, enables the system to deliver high-quality products and services, and scales up machine learning applications and algorithms cheaply. Security, associated with system safety and robustness, entails capabilities around detecting data breaches, preventing cyber-attacks, and adapting to sudden changes.

The second category of questions relates to advanced tools that are present within some functional units but not necessarily across the entire organization. This captures the use of specific pieces of technologies which do not require coordination across departments and locations. We name the index Technology Use.

The third category of questions relates to machine learning applications that are developed and deployed across various business functions around the firm. These applications are co-invention generated by machine learning engineers, data scientists, and business analysts, among other end-users of data. Although specific business needs differ across firms, common ML use cases are identified generally in similar functional units. We name the index Machine Learning (ML) Deployment.

We leave detailed definitions of the sub-indexes of data architecture to Appendix A.1, along with all the other indexes briefly described above. Non-survey data sources for the paper are discussed in Appendix A.2.

\(^5\) Survey participants or research team members may choose to skip a question during the interview process, resulting in blank response. We code these blanks as missing values, and do not count them toward the calculation of indexes.
4 Empirical Strategy

This section describes an approach to testing the theories outlined in Section 2, by measuring legacy IT components that lead to quasi-random variation in architectural inertia, and software system complexity makes data architecture even more important to realizing the gains from technical innovation. We use these ideas to construct an instrumental variable and estimate the effects of data architecture on productivity growth and co-invention output, i.e., machine learning deployment at the corporation level.

4.1 Legacy IT Components and Data Architecture

Almost all large companies have information technology (IT) systems that run a variety of enterprise software applications. These software applications deliver business intelligence by generating, storing, and analyzing data. IT systems are firm-specific assets with varying capabilities and maintainability. They took shape over a long time and evolved around idiosyncratic business processes after many past decisions.

Outdated technological systems hinder architectural innovation that is required for organizations to deliver value from new data and machine learning technologies. Large corporations’ IT systems often contain legacy components such as old equipment and inconsistent software. They gather data in a distributed fashion, leading to data silos incompatible across functions and locations. This prevents the systems from integrating the latest technology and product updates which can be essential to user co-invention. It makes developing IT capabilities that require coordination across the organizations’ many departments difficult.

Legacy IT components constrain corporations’ ability to implement data architecture practices that are close to the digital frontier. However, many companies continue to operate legacy IT systems, especially if they handle important business processes, and meet the present needs. Maintaining Legacy IT components require substantial resources. Chief information officers typically dedicate 40 – 60% of their work time to managing outdated systems. Re-designing and migrating to new systems require a large amount of effort and time. These upfront investments are particularly costly to massive corporations with complex systems. Workers may have become reliant on the old systems and developed customized approaches to interacting with them. All these factors make legacy IT components difficult to remove, and hence outdated parts of the IT systems can be quasi-fixed in the short run. Hence their impact on data architecture can persist beyond a few years.

4.1.1 Third-Party Maintenance (TPM) for Data Center Servers

The Ci Technology Database (CiTDB) contains information about IT installment related to the processing and storage of data. We use a variable that is an indicator of third-party maintenance (TPM) for data center servers at the establishment level. We construct an instrumental variable for data architecture by counting the number of establishments that report the presence of server third-party maintenance (TPM) up to 2015. The instrument proxies for the prevalence of legacy IT components across the corporation.
Firms use third party vendors to provide maintenance services, after product warranty expires so that the original equipment manufacturer (OEM) is no longer responsible for keeping the servers running. Third-party maintenance (TPM) is identified typically after the equipment reaches its end-of-service-life date. Instead of choosing TPM for data center servers, firms can also purchase new servers and latest technologies from the OEM. In the short run, firms save costs by choosing TPM for old servers and avoid paying for expensive up-to-date products. In the longer run, however, this can lead to inconsistency and incompatibility issues if firms try to integrate new technology and product updates in other parts of the IT systems.

4.1.2 IV First Stage and Exclusion Restriction

For server TPM to be a valid instrument for data architecture, it should induce meaningful variation in the ability to commit to architectural innovation across firms. As discussed, legacy IT components tend to be quasi-fixed in the short run, and hence their impact on data architecture can persist for at least a few years. Equation 1 contains the OLS regression model that identifies the effect of server TPM prevalence on data architecture at the corporation level.

\[
\text{DataArchitecture}_{i} = \beta_0 + \beta_1 \cdot \text{ServerTPM}_{i,2015} + \beta_2 \cdot \text{Controls}_i + \epsilon_i
\]  

(1)

Each observation is a corporation, indexed by \(i\). We estimate Equation 1 using outcome variables that include both the overall data architecture index and sub-indexes of data architecture stages and quality. Results are presented and discussed in Section 5.2.

The validity of the instrument relies on satisfying a few conditions of exclusion restriction. Server TPM must be exogenous, and affect machine learning output and productivity growth only through its impact on data architecture. Using TPM for data center equipment implies that the firm runs outdated hardware that had expired and no longer maintained by the original equipment manufacturer (OEM). Instead of buying updated products, companies try to cut costs on IT operations by using third-party vendors to maintain old equipment.

Companies may be more likely to use third-party maintenance (TPM) when there is less urgent business need for purchasing up-to-date hardware. For example, some firms do not collect much data at all. Some businesses do not require accessing data frequently, or keeping data stored in the system for a long time. These companies have few incentives to invest in large-scale computing infrastructure. Equation 1 controls for corporation size, proxied by the log number of establishments, which accounts for the amount of existing data, the magnitude of the user base, and the scope of product offerings across the firm.

The presence of third-party maintenance (TPM) is not associated with low-quality management. Firms using TPM are often motivated by cost-saving and growth-boosting strategies. Such strategies aim at improving profit margins by cutting IT overhead costs. Server TPM can save up to 70% on equipment purchase costs, and is not related to poor management practices.
We conduct robustness checks for instrument exogeneity using Equation 2. It estimates whether the prevalence of server TPM at the corporation level is correlated with company characteristics that may directly contribute to firm performance and other outcomes such as machine learning deployment.

\[ \text{CompanyCharacteristics}_i = \gamma_0 + \gamma_1 \cdot \text{ServerTPM}_{i,2015} + \gamma_2 \text{Controls}_i + \nu_i \] (2)

Company characteristics include technology intensity indexes – technology use and innovation processes, both measured by the survey across the entire firm. Other characteristics are revenue and production factors in the baseline year of 2016.

Another major concern for instrument exogeneity is its relationship with cloud computing adoption, or the lack thereof. Both server TPM and cloud computing are strategies to help firms manage data and computing resources, which involve external providers and third-party vendors taking over (at least some part of) the infrastructural needs. However, server TPM implies that traditional data centers are still being used, but cloud computing (or at least that through a public cloud provider) is a much more flexible approach to managing data and replaces data centers altogether.

Equation 3 estimates the correlation between server TPM and cloud computing at the firm level. Cloud computing is quantified by a survey-based index, calculated as the share of “yes” responses to questions around corporation-wide adoption of cloud infrastructure, frameworks, and tools (e.g., Azure HDInsight, Amazon EMR, and Terraform).

\[ \text{CloudComputingIndex}_i = \gamma_0 + \gamma_1 \cdot \text{ServerTPM}_{i,2015} + \gamma_2 \text{Controls}_i + \nu_i \] (3)

Results from these tests on instrument validity are presented and discussed in Section 5.2.1.

4.1.3 IV Second Stage and Firm-Level Outcomes

We estimate the effects of data architecture on firm-level outcomes using Equations 4 and 5.

Equation 4 estimates the impact of data architecture on productivity growth, in a classic production function framework. Data architecture is measured by a survey-based index quantifying the organization-wide technological architecture around data and machine learning. As an important component of a firm’s digital intangible capital, it is relatively slow to change and remains quasi-fixed in the short run. The index quantifies the capabilities of the data architecture independent of its costs or monetary value, and hence does not fluctuate with the prices of technical hardware, data sets, or machine learning output.

\[ \Delta \log(\text{Revenue})_{i,2016-2019} = \alpha_0 + \alpha_1 \cdot \text{DataArchitecture}_i + \alpha_2 \text{ITCapital}_{i,2015} + \alpha_3 \Delta \log(\text{Materials})_{i,2016-2019} + \alpha_4 \Delta \log(K)_{i,2016-2019} + \alpha_5 \Delta \log(L)_{i,2016-2019} + \alpha_6 \text{OtherControls}_i + \nu_i \] (4)
The outcome variable is the growth in total enterprise revenue from 2016 to 2019. Control variables include the changes in production factors – employment, capital (property, plant, and equipment), and materials over the same time period. To distinguish between the effect of data architecture and that of traditional IT assets, the regression controls for proxies of IT capital stock derived from the CitDB database. We also account for company characteristics in the baseline year of 2016, such as log firm age, number of establishments, total enterprise revenue, as well as industry fixed effects.

Equation 5 estimates the impact of data architecture on user co-invention, measured by the machine learning (ML) deployment index derived from survey responses. The index measures the scope of typical ML applications and use cases that are developed and deployed within the corporation. These applications and use cases span multiple business functions, such as production, customer relations, sales and marketing, and employee management. They often do not involve more advanced methods than applying the basic and widely available tools to train and validate predictive models, using cumulative data collected from customers and business activities.

\[ \text{MLDeployment}_i = \alpha_0 + \alpha_1 \cdot \text{DataArchitecture}_i + \alpha_2 \text{Controls}_i + \nu_i \]  

(5)

To estimate Equation 5, we use the same server TPM instrument to identify quasi-random variations in technological architecture, and estimate the causal effect of adopting frontier data architecture practices on ML deployment at the corporation level.

4.2 IT System Complexity and Importance of Technological Architecture

Large corporations have idiosyncratic and complex IT systems. Each system is associated with a variety of business processes and user co-invention practices shaped over a long time. Some firms have more convoluted systems than others, resulting in more complicated processes that adjust more slowly. Simpler IT systems integrate new components more easily, hence less demand for frontier capabilities in the data architecture. More complex systems require data architecture equipped with more advanced capabilities, i.e., handle more complicated requests, and iterative over models more efficiently. Hence, data architecture may play an even larger role in reducing costs and frictions of accessing data and tools, among firms with more complicated IT systems.

4.2.1 Measuring the Complexity of Software Systems

To measure the complexity of computing systems, we calculate the prevalence of system software within. The idea is similar to Bresnahan and Greenstein (1996): the costs to coordinating changes across systems tend to be higher at establishments with more system software categories. System software applications tend to embed idiosyncrasies of the firm, and are particularly difficult to work with. More precisely, complexity is
defined by the ratio of the maximum number of system software categories to the maximum number of overall software categories across establishments.\(^6\)

System software is designed to operate computer hardware and manage system-wide resources. It provides a platform (e.g., operating systems) for running various components of the computer system and other types of software applications, e.g., diagnostic and communication tools, and data management programs. Users do not typically work with system software directly, but system software helps them interact with hardware, e.g., through graphical user interfaces. System software applications are typically written in complex low-level languages, and debugging them requires knowledge of the underlying hardware. Adequate data architecture designs minimize the need for technical users to directly modify or debug system software, and reduce the costs to making adjustments or committing errors from running data queries and testing predictive models.

On the other hand, complex IT systems is not necessarily associated with less advanced data architecture practices with fewer frontier capabilities. Instead, a larger degree of complexity appears to heighten adjustment costs, and increase the amount of effort required for technical workers to adapt to system-wide changes. When data users interact with a more complex IT system, they respond more slowly to updates in the underlying infrastructure, and develop new co-invention processes at larger costs.

### 4.2.2 Software System Complexity and Effect Heterogeneity

Each large corporation has its own unique configuration of IT systems. Working with these systems requires organizational assets complementary to firms’ specific needs. Optimal designs of the data architecture may entail different capabilities depending on idiosyncratic implementations of IT systems. Understanding which aspects of the systems are most relevant to data architecture can inform us about where firms should focus on re-architecting, and which practices to adopt. Equation 6 estimates the differential effects of legacy IT components (measured by server TPM) on firm-level outcomes by software system complexity.

\[
Outcome_i = \alpha_0 + \alpha_1 \cdot ServerTPM_{i,2015} + \alpha_2 \cdot ServerTPM_{i,2015} \times Complex_{i,2015} \\
+ \alpha_3 ITCapital_{i,2015} + \alpha_4 Controls_{i} + v_i
\]  

Firm-level outcomes include productivity growth and machine learning deployment. The \(Complex_{i,2015}\) variable is an indicator of software system complexity measured in 2015. Specifically, the raw complexity variable is calculated as the ratio of the maximum number of system software categories to the maximum number of overall software categories across all U.S.-based establishments within a corporation. The \(Complex_{i,2015}\) indicator is equal to 1 if more than 50% of overall software categories are system software, and 0 otherwise.

---

\(^6\) System software categories include asset management, enterprise management, identity and access, collaboration, application consolidation, enterprise resource planning, web portal, development, and workflow. Other software categories include application server, business intelligence, customer relationship management, document management, accounting, human resources, supply chain, data warehouse, security, and groupware.
Apart from estimating the main effect of server TPM, Equation 6 also includes the interaction between server TPM and the system complexity indicator as a regressor. Costs and frictions associated with co-invention may depend on the complexity of firm-specific software systems. Since software system complexity does not correlate with server TPM or data architecture (see Appendix Figure B.3), $\alpha_2$ quantifies the extent to which data architecture additionally benefits particularly complex organizations, where system software is especially prevalent across establishment-level IT systems. In other words, estimation results reflect heterogeneity in the importance of adopting data architecture frontier practices among firms with IT systems of varying complexity.

5 Empirical Results

In this section, we present descriptive statistics on the survey sample, establish the validity of using server third-party maintenance (TPM) as an instrument for data architecture, and estimate the effects of frontier data architecture capabilities on key firm-level outcomes.

5.1 Descriptive Statistics

The final data sample consists of 117 U.S.-based corporations spanning a few traditional sectors such as manufacturing, retail trade, financial services, and health care. We summarize variables in the rest of this section. The first set of variables describe survey-based data architecture index and sub-indexes. The second set of variables describe IT systems and technological intensity, as well as company characteristics and performance.

5.1.1 Data Architecture Index and Sub-Indexes

Table 1 panel (a) summarizes the overall data architecture index, as well as data platform, machine learning (ML) system, coherence, and security sub-indexes; panel (b) presents the correlations between each pair of data architecture sub-indexes. Corporations in the sample have relatively low adoption of frontier data architecture practices on average – the overall index has a median of 0.52 and an inter-decile range from 0.29 to 0.80.

[ Insert Table 1 here ]

Corporations in the sample have relatively mature data platforms. The bottom decile data platform sub-index is 0.33 across firms. Around 94% of the firms have centralized data platforms accessible to workers from across the entire organization, for performing simple queries and data analyses. On the other hand, these firms face larger difficulty in designing and implementing machine learning (ML) systems. Around 10% of the firms do not have any ML system at all. Only 64% of the firms have large-scale experimentation platforms, and 52% have automated machine learning pipelines and adopt DevOps practices.

The coherence and security aspects of data architecture quality have similar distributions in the survey sample. The average sub-index for each quality is about 0.5, and the correlation between the two sub-indexes is quite low at 0.29. Coherence and security measure distinct capabilities and practices that are not substitutes

\[ ^7 \text{A few corporations have legal headquarters outside the U.S., but operational headquarters within the country.} \]
with each other. Data architecture coherence correlations are quite large with both stages of the architecture, at 0.67 with the data platform index and 0.79 with the ML system index, respectively. Fully integrated data sets and machine learning tools are associated with efficiency and scale, across both stages of the data architecture.

Data architecture capabilities differ substantially across industries. Appendix Figure B.2 describes industry averages and inter-decile ranges of the overall index. While firms specializing in making technology hardware (e.g., computer and electronic product) are the closest to frontier implementation, firms in retail trade and other manufacturing subsectors score about 32% lower in their data architecture practices.

5.1.2 IT Systems and Firm Characteristics

Table 2 describes legacy information technology and other firm-level variables. IT installment measures are calculated from the CiTechnology Database (CiTDB), and in particular the 2014 and 2015 snapshots. Other variables come from a variety of sources including Compustat. Data sources are described in Appendix A.2.

We identify third-party maintenance (TPM) within establishments as indicators of legacy IT components. Outdated components in the IT system constrain its technical capabilities. Some parts of the system, e.g., data center servers, are linked to data processing and storage, and hence especially relevant to data architecture. About 69 percent of firms in the sample do not have an establishment that uses TPM for servers, 15 percent report server TPM in one establishment, and the rest report server TPM in at least two establishments.8

We measure the complexity of software systems using information about the different categories of software applications at the establishment level. There are primarily two types of software applications: system software (e.g., operating systems) and application software (e.g., Microsoft Word). Each type of software applications includes different products, vendors and categories. We observe the presence of each detailed software category at the establishment level. We define firm-level software system complexity by the ratio of the maximum number of system software categories to the maximum number of overall software categories across all establishments. The value of this variable lies between 0 and 1.

The average firm identifies about 9 different software categories in the establishment with the largest number of overall software categories, and about 5 different system software categories in the establishment with the largest number of system software categories (hence the average ratio is about 0.56). Software system complexity varies widely across firms, and the inter-decile range spans from 0.3 to 0.8.

8 To compare, we also summarize other indicators for legacy systems that are less related to data architecture, such as the TPM for any hardware including but not limited to servers, and for all kinds of services such as phone systems, firewall and intrusion detection services, as well as the presence of laptops and desktop computers more than three years old.
Some corporations in the survey sample have long histories dating centuries back to the Industrial Revolution. We collect data on founding year, by manually checking across company profiles and data sources such as Compustat. Company age is calculated relative to the baseline year of 2016. Founding year corresponds to when the earliest precedent of the firm was established, rather than the last time that a major event, e.g., a merger, acquisition, spinoff, or name change resulted in company restructuring or headquarter migration.\(^9\)

Firms in the sample are some of the largest corporations operating in many locations across the United States. They have sustained annual revenues of over $1 billion, and the median revenue was $16.6 billion in 2016. The median firm in the sample operates over 150 establishments within the country. Some firms span over a thousand establishments, and even the bottom-decile firms span a few dozen establishments. Total revenues grew by 16\% from 2016 to 2019 on average, with substantial variation across firms.

Machine learning (ML) deployment is measured as an index between 0 and 1, by the share of common ML applications identified within the firm. The index quantifies the scope of technical co-invention – predictive models and machine learning algorithms, developed by machine learning engineers, data scientists, and business analysts working with the technological architecture and data sets across the entire organization.

The ML deployment index has an average of 0.4 and a standard deviation of 0.2. Among other survey-based technological intensity indexes, data architecture has an average of 0.53, innovation processes (another type of digital intangible capital) has an average of 0.45, and technology use (i.e., adoption of specific tools) has an average of 0.64. These indexes quantify the technological architecture and organizational capital specifically related to data and machine learning, and not previously measured in other surveys (e.g., Census Bureau data).

Among production factors, capital stock (property, plant, and equipment) grew by 20\% on average from 2016 to 2019. Labor (enterprise-level employment) shrunk at some firms and expanded at others, with an inter-decile range between -24\% and 33\% from 2016 to 2019. The median firm in the sample has 20 computers and 2 IT staff members per establishment.\(^{10}\) These quantities proxy for traditional IT capital and correlate with general types of IT-related assets, but are distinct from digital capital, e.g., data architecture.

### 5.2 Effect of Server TPM on Data Architecture

We present regression results on the effects of legacy IT on data architecture, where legacy IT is measured by the prevalence of server third-party maintenance at the corporation level. Figure 3 plots the binned residuals from the regression and illustrates $\beta_1$, the coefficient estimate. Each additional establishment with server third-party maintenance lowers the overall data architecture index by 0.03, or 0.17 standard deviations across

\(^9\) For example, Raytheon Technologies Corporation resulted from a merger in early 2020 between two large companies – Raytheon Company founded in 1922, and United Technologies Corporation founded in 1934. The CITDB snapshots contain information on establishments associated with both companies before the merger. We link establishments of both companies to the Raytheon Technologies conglomerate, which constitutes one observation in the survey data.

\(^{10}\) IT staff estimates are based on raw data with only range buckets instead of precise quantities.
corporations in the sample. The effect is statistically significant at the 1% level. Control variables include industry fixed effects, log number of establishments, log company age, and average log number of computers per establishment.

Table 3 shows the regression results where outcome variables are the overall data architecture index, and four sub-indexes – data platform, ML system, coherence, and security, in columns 1 – 5, respectively. Server TPM in 2015 led to lower values of the overall data architecture index, and most sub-indexes measured in 2020 except security. The coefficient estimates are statistically different from zero and robust to controlling for changes in production factors from 2016 to 2019, and log total enterprise revenue in 2016.

A few interesting correlations emerge from the results. Employment growth is associated with data platform quality (column 3), and capital growth is associated with architectural coherence (column 5). Relatedly, a larger number of establishments is associated with higher-quality data architecture across all sub-indexes except security (columns 1 – 5). Data storage and processing needs tend to require more advanced architectural capabilities among particularly large firms, for hosting more data and accommodating more workers.

5.2.1 Instrument Validity and Exclusion Restriction

Appendix Table B.2 shows that server TPM is not correlated with key company characteristics that directly contribute to technology adoption and firm performance. These company characteristics include survey-based technology use and innovation processes indexes, as well as log revenue, capital, and employment in 2016.

The regressions control for log company age, log number of establishments, and average log IT staff per establishment. Another concern for instrument validity is that it may be correlated with cloud computing, but regression results in Appendix Table B.3 suggest otherwise.

All columns estimate close to zero correlation between server TPM in 2015 and cloud computing adoption in 2020. The coefficients are stable to specifications with different sets of control variables.

5.3 Effect of Data Architecture on Productivity Growth

Table 4 reports regression results from estimating the impact of data architecture on productivity growth using Equation 4. OLS results in columns 2 and 3 suggest a positive correlation between data architecture and productivity growth, after accounting for production factors – labor, capital, and materials, and robust to controlling for technology use. For comparison, column 1 estimates the coefficients on production factors only.
The OLS estimate may suffer downward bias due to reversed causality. Data architecture is measured in 2020 – corporations’ investments into digital transformation may lead to improved data architecture capabilities after 2015. On the other hand, productivity growth is measured between 2016 and 2019. Measured server TPM in 2015 or before may not reflect changes that can affect data architecture after 2016.

For example, laggards that caught up in 2018 may have improved data architecture practices by the end of 2020, but more recent investments into technological infrastructure were unlikely to yield immediate returns and may have the opposite effect on revenue growth in the short run. Early-movers that invested in developing frontier data architecture capabilities before 2016 were likely to reap productivity benefits, without putting additional resources into re-architecting after 2016. Benefits from data architecture can take time to realize, partly because it requires accumulating input data over time to tune predictive algorithms before producing sufficiently high-quality machine learning output.

Table 4 columns 3 – 7 present the causal estimates of the effect of data architecture on revenue growth, using the corporation-level prevalence of server TPM as the instrument for data architecture. A one standard deviation increase (or 0.18 in magnitude) in the data architecture index leads to about 15% higher productivity growth from 2016 to 2019. The results are robust to controlling for technology use, baseline firm characteristics, and production factors. The effect is not driven by traditional measures of IT capital. Data architecture requires specific kinds of investments into developing capabilities around data and machine learning, which are distinct from purchasing computer equipment and stacking up software applications.

Taken together, these results suggest that data architecture is crucial to boosting productivity growth and complementing relatively common technologies around data and machine learning. Productivity benefits tend to lag the adoption of new technology by at least three years (Brynjolfsson, Rock and Syverson, 2021). Our empirical approach addresses the time lag with an instrumental variable measured a few years before the firm-level outcomes are realized. The first-stage regression essentially predicts the part of current data architecture that was solely driven by past IT systems before 2016. The fitted values from the first stage regression measures pre-2016 predicted values of the data architecture index, even though the survey was not conducted until 2020.

5.4 Effect of Data Architecture on Machine Learning Deployment

Table 5 presents regression results from estimating the impact of data architecture on co-invention effectiveness using Equation 5. OLS results in columns 1 and 2 suggest that data architecture and technology use are both positively correlated with machine learning (ML) deployment. Columns 3 – 7 present IV-2SLS estimates of the effect of data architecture on ML deployment, using the same instrumental variable identification approach as in Section 5.3. A one standard deviation increase (i.e., as large as 0.18 in magnitude) in the data architecture index leads to a 0.16 increase in the ML deployment index (or between 3 – 4 additional
applications). All regression specifications control for log number of establishments, log company age, log total enterprise revenue in 2016, a measure of traditional IT capital stock, and industry fixed effects. Adding controls for headquarter regions yield similar results.

After accounting for data architecture, the coefficients on technology use – measuring the presence of scattered pieces of advanced tools – are not significantly different from zero. IT capital stock contributes positively to ML deployment intensity. Doubling IT staff per establishment is associated with a 0.09 increase in the ML deployment index, or about 2 additional applications. As expected, IT investments should be positively correlated with technological capabilities in machine learning. Policy makers may worry about the distributional impact of data architecture leading to negative side effects for workers. However, our results suggest the absence of adverse consequences in corporation-level total employment (see Appendix A.3).

### 5.5 Heterogeneous Effects by Complexity of Software Systems

Figure 2 visualizes the regression results estimated from Equation 6 on two firm-level outcomes: machine learning deployment and productivity growth. Table 7 reports coefficient estimates. Regressions control for the maximum number of overall software categories across establishments.

Figure 2 panel (a) and Table 6 columns 2 – 3 suggest that server TPM affects machine learning deployment only for firms with complex software systems. Figure 2 panel (b) and Table 7 columns 5 – 6 suggest that almost two-thirds of the productivity effects of server TPM are attributed to firms with complex software systems. These effects hold similarly after controlling for headquarter region fixed effects.

To the extent that server TPM affects firm-level outcomes only through inducing architectural inertia, these results suggest that data architecture primarily affects co-invention output and productivity growth of firms with complex IT systems, measured by the share of system software categories hosted within. This interpretation relies on an assumption that software system complexity is not correlated with outdated data center hardware and does not lead to lower-quality data architecture. Appendix Figure B.2 shows that the share of system software categories is correlated with neither server TPM nor the survey-based data architecture index. The underlying regressions control for the log number of establishments and IT capital stock.

Taken together, these results suggest that data architecture particularly reduces the costs and frictions of user co-invention when firms’ IT systems are particularly complex. The importance of data architecture to value creation from data assets grows in the complexity of software systems.
Overall, our results suggest reasons for hope and optimism for the future of machine learning and data technologies. It points to evidence that the usefulness of these new technologies may not be overblown, despite somewhat disappointing evidence of slow diffusion and ineffective adoption. It is not a trivial task to identify the right set of complementary capabilities to invest in, and the large degree of idiosyncrasy and firm-specific needs largely prevent direct copying and imitation between firms of existing best practices.

We point out a key commonality among success cases of technology adoption – in particular, that data architecture capabilities associated with “digital native” firms’ frontier practices contribute to value creation from data and machine learning. High-quality data architecture facilitates delivery of products and services and effective co-invention shaped by interactions between workers and technological systems. This eventually leads to faster productivity growth in the longer run.

Large established corporations with especially cumbersome legacy IT systems benefit the most from investing in data architecture capabilities – setting up coherent data platforms and machine learning systems. More research is needed to help firms identify unique strategies and correct approaches to re-architect parts of the technological systems that are incompatible with innovation in data analytics and machine learning.
References


Figure 1: Relationship Between Legacy IT and Data Architecture Quality. This figure shows the bincscatter plot of the relationship between the number of establishments reporting third-party server maintenance (TPM) by 2015 and the survey-based data architecture index measured in 2020. The regression controls for industry fixed effects, log number of establishments, log company age in years, and a measure of traditional IT capital stock – average log number of computers per establishment.
**Figure 2: Software System Complexity and Heterogeneous Effects of Legacy IT.** This figure shows the scatter plot of the relationship between the number of establishments reporting third-party server maintenance (TPM) by 2015 and two corporation-level outcome variables. The outcome variable in the panel (a) is the survey-based ML deployment index, measured in 2020; the outcome variable in panel (b) is productivity growth from 2016 to 2019. We plot the observations separately for corporations with relatively complex versus simple software systems. Software system complexity is measured by the ratio of the maximum number of system software categories to the maximum number of overall software categories across establishments. The regressions controls for industry fixed effects, technology use, total software categories, log company age in years, log number of establishments, log total enterprise revenue in 2016, average log IT staff estimates per establishment, and changes in log materials, capital, and labor between 2016 and 2019.

(a) Machine Learning Deployment (in 2020)

(b) Productivity Growth (from 2016 to 2019)
Table 1: Data Architecture Descriptive Statistics. This table summarizes the survey-based data architecture overall index and sub-indexes for the 117 large corporations in the sample. The indexes measure the closeness of firms’ architectural capabilities to successful digital companies’ frontier practices. Panel (a) presents the mean, standard deviation, and the 10th, 25th, 50th, 75th, and 90th percentiles of the overall index and sub-indexes related to data architecture stages – data platform and ML system, and quality – coherence and security. Panel (b) reports the correlations between pairs of sub-indexes.

(a) Overall Index and Sub-Indexes

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>P10</th>
<th>P25</th>
<th>P50</th>
<th>P75</th>
<th>P90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.53</td>
<td>0.18</td>
<td>0.29</td>
<td>0.41</td>
<td>0.52</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td>Stages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Platform</td>
<td>0.60</td>
<td>0.20</td>
<td>0.33</td>
<td>0.43</td>
<td>0.57</td>
<td>0.73</td>
<td>0.87</td>
</tr>
<tr>
<td>ML System</td>
<td>0.41</td>
<td>0.30</td>
<td>0.00</td>
<td>0.12</td>
<td>0.43</td>
<td>0.67</td>
<td>0.78</td>
</tr>
<tr>
<td>Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coherence</td>
<td>0.51</td>
<td>0.22</td>
<td>0.23</td>
<td>0.36</td>
<td>0.50</td>
<td>0.64</td>
<td>0.85</td>
</tr>
<tr>
<td>Security</td>
<td>0.49</td>
<td>0.27</td>
<td>0.20</td>
<td>0.25</td>
<td>0.50</td>
<td>0.60</td>
<td>0.80</td>
</tr>
</tbody>
</table>

(b) Correlations Among Sub-Indexes

<table>
<thead>
<tr>
<th>Index</th>
<th>Data Platform</th>
<th>ML System</th>
<th>Coherence</th>
<th>Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Platform</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML System</td>
<td>0.479</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coherence</td>
<td>0.671</td>
<td>0.794</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Security</td>
<td>0.422</td>
<td>0.376</td>
<td>0.285</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 2: Regression Sample Descriptive Statistics. This table describes key variables in the survey sample of 117 large corporations. It reports the mean, standard deviation, as well as the 10th, 25th, 50th, 75th, and 90th percentiles of each variable. Survey-based indexes are calculated as the share of “yes” responses to relevant questions and range from 0 to 1. IT system variables are derived from establishment-level data and aggregated to the firm level. The difference (Δ) variables are calculated as the changes in the logarithms of level variables (e.g., revenue, capital, labor, and materials) between 2016 and 2019.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>P10</th>
<th>P25</th>
<th>P50</th>
<th>P75</th>
<th>P90</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δ Log (Revenue)</td>
<td>0.16</td>
<td>0.21</td>
<td>-0.06</td>
<td>0.05</td>
<td>0.15</td>
<td>0.26</td>
<td>0.37</td>
</tr>
<tr>
<td>ML Deployment Index</td>
<td>0.39</td>
<td>0.21</td>
<td>0.12</td>
<td>0.23</td>
<td>0.36</td>
<td>0.53</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Survey-Based Indexes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Architecture Index</td>
<td>0.53</td>
<td>0.18</td>
<td>0.29</td>
<td>0.41</td>
<td>0.52</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td>Technology Use Index</td>
<td>0.64</td>
<td>0.18</td>
<td>0.40</td>
<td>0.50</td>
<td>0.62</td>
<td>0.75</td>
<td>0.88</td>
</tr>
<tr>
<td>Innovation Processes Index</td>
<td>0.45</td>
<td>0.21</td>
<td>0.21</td>
<td>0.29</td>
<td>0.43</td>
<td>0.57</td>
<td>0.75</td>
</tr>
<tr>
<td>Cloud Computing Index</td>
<td>0.58</td>
<td>0.26</td>
<td>0.20</td>
<td>0.33</td>
<td>0.60</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>IT System Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server TPM</td>
<td>0.70</td>
<td>1.47</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Shr. System Software Categories</td>
<td>0.48</td>
<td>0.20</td>
<td>0.31</td>
<td>0.36</td>
<td>0.44</td>
<td>0.54</td>
<td>0.80</td>
</tr>
<tr>
<td># Software Categories</td>
<td>9.20</td>
<td>3.69</td>
<td>3.00</td>
<td>7.00</td>
<td>10.00</td>
<td>12.00</td>
<td>13.00</td>
</tr>
<tr>
<td><strong>Firm Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log (# Establishments)</td>
<td>5.11</td>
<td>1.61</td>
<td>3.14</td>
<td>3.99</td>
<td>5.12</td>
<td>6.23</td>
<td>7.03</td>
</tr>
<tr>
<td>Log (# 3-Digit Zipcodes)</td>
<td>4.44</td>
<td>1.28</td>
<td>2.64</td>
<td>3.53</td>
<td>4.65</td>
<td>5.34</td>
<td>5.98</td>
</tr>
<tr>
<td>Avg Log IT Staff Per Estab</td>
<td>0.72</td>
<td>0.43</td>
<td>0.26</td>
<td>0.45</td>
<td>0.66</td>
<td>0.93</td>
<td>1.26</td>
</tr>
<tr>
<td>Avg Log # PCs Per Estab</td>
<td>3.13</td>
<td>0.63</td>
<td>2.40</td>
<td>2.71</td>
<td>3.01</td>
<td>3.42</td>
<td>3.92</td>
</tr>
<tr>
<td>Company Age</td>
<td>94.76</td>
<td>49.05</td>
<td>33.00</td>
<td>49.00</td>
<td>94.00</td>
<td>126.00</td>
<td>165.00</td>
</tr>
<tr>
<td>Log (Revenue in 2016)</td>
<td>9.65</td>
<td>1.21</td>
<td>8.04</td>
<td>8.71</td>
<td>9.65</td>
<td>10.45</td>
<td>11.18</td>
</tr>
<tr>
<td><strong>Production Factors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δ Log (Capital)</td>
<td>0.20</td>
<td>0.26</td>
<td>0.00</td>
<td>0.05</td>
<td>0.17</td>
<td>0.33</td>
<td>0.54</td>
</tr>
<tr>
<td>Missing Value - Capital</td>
<td>0.13</td>
<td>0.34</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Δ Log (# Employees)</td>
<td>0.04</td>
<td>0.36</td>
<td>-0.24</td>
<td>-0.10</td>
<td>0.00</td>
<td>0.12</td>
<td>0.33</td>
</tr>
<tr>
<td>Missing Value - Labor</td>
<td>0.03</td>
<td>0.16</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Δ Log (Materials)</td>
<td>0.14</td>
<td>0.32</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.20</td>
<td>0.57</td>
</tr>
<tr>
<td>Missing Value - Materials</td>
<td>0.50</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 3: Effect of Server TPM on Data Architecture Quality. This table reports the regression coefficients and robust standard errors on the estimated effects of the number of establishments reporting server third-party maintenance (TPM) by 2015 on organization-wide data architecture in 2020. All columns control for industry fixed effects, changes in production factors, log number of establishments, log company age, log total enterprise revenue in 2016, and average log number of computers per establishment. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Data Architecture Indexes</th>
<th>Stages</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall (1)</td>
<td>Data Plf. (2)</td>
<td>ML Sys. (3)</td>
</tr>
<tr>
<td>Server TPM</td>
<td>-0.026***</td>
<td>-0.026***</td>
<td>-0.019*</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.010)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>Δ Log (Materials)</td>
<td>-0.062</td>
<td>-0.059</td>
<td>-0.074</td>
</tr>
<tr>
<td></td>
<td>(0.058)</td>
<td>(0.061)</td>
<td>(0.082)</td>
</tr>
<tr>
<td>Δ Log (Capital)</td>
<td>0.135*</td>
<td>0.129</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td>(0.076)</td>
<td>(0.079)</td>
<td>(0.089)</td>
</tr>
<tr>
<td>Δ Log (# Employees)</td>
<td>0.020</td>
<td>0.022</td>
<td>0.085**</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.037)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Log (Company Age)</td>
<td>-0.053*</td>
<td>-0.055</td>
<td>-0.057</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.033)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Log (# Establishments)</td>
<td>0.061***</td>
<td>0.059***</td>
<td>0.049**</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.019)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Log (Revenue in 2016)</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.021)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>Avg Log # PCs Per Estab</td>
<td>-0.012</td>
<td>-0.023</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.037)</td>
<td>(0.058)</td>
</tr>
<tr>
<td>No. Obs.</td>
<td>117</td>
<td>117</td>
<td>117</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.279</td>
<td>0.280</td>
<td>0.185</td>
</tr>
</tbody>
</table>
Table 4: Effect of Data Architecture Quality on Productivity Growth. This table reports the regression coefficients and robust standard errors on the estimated effects of data architecture quality on productivity growth from 2016 to 2019. Control variables include industry fixed effects, changes in production factors, log company age, log number of establishments, and log total enterprise revenue in 2016, and average log number of computers per establishment. Columns 3, 5, and 6 control additionally for technology use. Columns 4 – 6 are estimated using server TPM prevalence by 2015 as the instrument for data architecture quality. Headquarter region fixed effects are at the Census Region level. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>OLS (1)</th>
<th>OLS (2)</th>
<th>OLS (3)</th>
<th>IV-2SLS (4)</th>
<th>IV-2SLS (5)</th>
<th>IV-2SLS (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ Log (Materials)</td>
<td>0.128</td>
<td>0.140</td>
<td>0.140</td>
<td>0.174**</td>
<td>0.181**</td>
<td>0.190**</td>
</tr>
<tr>
<td></td>
<td>(0.097)</td>
<td>(0.090)</td>
<td>(0.092)</td>
<td>(0.085)</td>
<td>(0.084)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>Δ Log (Capital)</td>
<td>0.229**</td>
<td>0.196**</td>
<td>0.196*</td>
<td>0.098</td>
<td>0.100</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>(0.093)</td>
<td>(0.096)</td>
<td>(0.099)</td>
<td>(0.117)</td>
<td>(0.108)</td>
<td>(0.126)</td>
</tr>
<tr>
<td>Δ Log (# Employees)</td>
<td>0.035</td>
<td>0.028</td>
<td>0.028</td>
<td>0.010</td>
<td>0.035</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>(0.061)</td>
<td>(0.063)</td>
<td>(0.063)</td>
<td>(0.073)</td>
<td>(0.067)</td>
<td>(0.073)</td>
</tr>
<tr>
<td>Log (Company Age)</td>
<td>0.023</td>
<td>0.035</td>
<td>0.035</td>
<td>0.068</td>
<td>0.044</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.029)</td>
<td>(0.029)</td>
<td>(0.042)</td>
<td>(0.032)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>Log (# Establishments)</td>
<td>0.012</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.040</td>
<td>-0.020</td>
<td>-0.027</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.020)</td>
<td>(0.020)</td>
<td>(0.031)</td>
<td>(0.022)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>Log (Revenue in 2016)</td>
<td>-0.028</td>
<td>-0.026</td>
<td>-0.026</td>
<td>-0.022</td>
<td>-0.024</td>
<td>-0.021</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.028)</td>
<td>(0.028)</td>
<td>(0.031)</td>
<td>(0.028)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Avg Log # PCs Per Estab</td>
<td>0.072**</td>
<td>0.076**</td>
<td>0.076**</td>
<td>0.088**</td>
<td>0.084**</td>
<td>0.089**</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.035)</td>
<td>(0.035)</td>
<td>(0.044)</td>
<td>(0.039)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Data Architecture Index</td>
<td>0.259**</td>
<td>0.258**</td>
<td>1.012**</td>
<td>0.823**</td>
<td>0.973**</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td>(0.110)</td>
<td>(0.125)</td>
<td>(0.486)</td>
<td>(0.334)</td>
<td>(0.458)</td>
<td></td>
</tr>
<tr>
<td>Technology Use Index</td>
<td>0.000</td>
<td>-0.282</td>
<td>-0.318</td>
<td>-0.040</td>
<td>-0.027</td>
<td>-0.021</td>
</tr>
<tr>
<td></td>
<td>(0.104)</td>
<td>(0.185)</td>
<td>(0.238)</td>
<td>(0.031)</td>
<td>(0.022)</td>
<td>(0.026)</td>
</tr>
</tbody>
</table>

Headquarter Region FE ✓

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>OLS</th>
<th>OLS</th>
<th>IV-2SLS</th>
<th>IV-2SLS</th>
<th>IV-2SLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Obs.</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.307</td>
<td>0.346</td>
<td>0.346</td>
<td>0.014</td>
<td>0.208</td>
<td>0.160</td>
</tr>
<tr>
<td>First Stage F-Statistic</td>
<td>7.3</td>
<td>17.7</td>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5: Effect of Data Architecture Quality on Machine Learning Deployment. This table reports the regression coefficients and robust standard errors on the estimated effects of data architecture quality on machine learning deployment measured by the survey in 2020. Columns 2, 4, 5, and 6 control for technology use. Columns 3 – 6 are estimated using server TPM prevalence by 2015 as the instrument for data architecture quality. All columns control for measures of traditional IT capital stock – average log IT staff estimates per establishment (columns 1 – 4 and 6) and average log number of computers per establishment (column 5). Headquarter region fixed effects are at the Census Region level. All columns control for industry fixed effects. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>OLS</th>
<th>OLS</th>
<th>IV-2SLS</th>
<th>IV-2SLS</th>
<th>IV-2SLS</th>
<th>IV-2SLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML Deployment Index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td></td>
</tr>
<tr>
<td>Data Architecture Index</td>
<td>0.491*</td>
<td>0.904**</td>
<td>0.833**</td>
<td>0.798**</td>
<td>0.878**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.092)</td>
<td>(0.409)</td>
<td>(0.334)</td>
<td>(0.328)</td>
<td>(0.404)</td>
<td></td>
</tr>
<tr>
<td>Log (Company Age)</td>
<td>0.012</td>
<td>0.003</td>
<td>0.030</td>
<td>0.020</td>
<td>0.020</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.032)</td>
<td>(0.037)</td>
<td>(0.031)</td>
<td>(0.031)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Log (# Establishments)</td>
<td>0.007</td>
<td>0.018</td>
<td>-0.010</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.017)</td>
<td>(0.025)</td>
<td>(0.018)</td>
<td>(0.018)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>Log (Revenue in 2016)</td>
<td>0.034*</td>
<td>0.033*</td>
<td>0.035*</td>
<td>0.034*</td>
<td>0.034*</td>
<td>0.032*</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.019)</td>
<td>(0.021)</td>
<td>(0.019)</td>
<td>(0.020)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Avg Log IT Staff Per Estab</td>
<td>0.088**</td>
<td>0.075*</td>
<td>0.090**</td>
<td>0.095**</td>
<td>0.104**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.044)</td>
<td>(0.041)</td>
<td>(0.042)</td>
<td>(0.043)</td>
<td></td>
</tr>
<tr>
<td>Technology Use Index</td>
<td>0.265**</td>
<td>0.138</td>
<td>-0.107</td>
<td>-0.154</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.111)</td>
<td>(0.176)</td>
<td>(0.174)</td>
<td>(0.206)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg Log # PCs Per Estab</td>
<td>0.063**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Headquarter Region FE</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. Obs.</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.412</td>
<td>0.310</td>
<td>0.311</td>
<td>0.354</td>
<td>0.359</td>
<td>0.347</td>
</tr>
<tr>
<td>First Stage F-Statistic</td>
<td>8.7</td>
<td>19.1</td>
<td>18.2</td>
<td>14.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6: Heterogeneous Effects of Server TPM by Software System Complexity. This table reports the regression coefficients and robust standard errors on the differential effects of server TPM prevalence on firm-level outcomes by software system complexity. Software system complexity is equal to 1 if the ratio of the maximum number of system software categories to the maximum number of overall software categories across establishments exceeds 0.5, and equal to 0 otherwise. Baseline company controls include log company age in years, log number of establishments, and log total enterprise revenue in 2016. IT investment controls include average log IT staff estimates per establishment. Production factor controls include changes in log materials, capital, and labor between 2016 and 2019. Headquarter region fixed effects are at the Census Region level. All columns control for industry fixed effects. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>ML Deployment Index</th>
<th>Revenue Growth (2016-2019)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Server TPM</td>
<td>-0.015</td>
<td>-0.009</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.013)</td>
</tr>
<tr>
<td># Software Categories</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>Share System Software</td>
<td>0.147</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>(0.106)</td>
<td>(0.108)</td>
</tr>
<tr>
<td>Server TPM x Complex</td>
<td>-0.070**</td>
<td>-0.071***</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>Company Baseline Controls</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IT Investment Controls</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Production Factor Controls</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Headquarter Region FE</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>No. Obs.</td>
<td>117</td>
<td>117</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.391</td>
<td>0.417</td>
</tr>
</tbody>
</table>
Appendices

A.1. Taxonomy of Survey-Based Indexes

**Data Platform.** This is a sub-index of data architecture measuring one of its two stages. The data platform contains the data fabric that cover many functional units across locations. It ingests and combines raw data from a variety of sources, such as user activities, business information, and government statistics. The data platform should be accessed by workers at various levels of the corporation and allow data scientists and business users to query the data and develop predictive models. For example, questions related to the data platform involve the presence and capabilities of data pipelines, application programming interface (API), elastic computing, cloud storage, among others. Application Programming Interfaces (APIs) standardize data sets and organize them into structured records, which can be sliced and retrieved by users. The data platform often needs to handle flows of fresh data rather than stocks of stale data. Data pipelines process and incorporate data streams automatically for repeat use.\(^{11}\) Data pipeline outputs need to be compatible with many different types of destinations – for storage (e.g., data lakes and warehouses), further processing (e.g., machine learning systems), and end-products (e.g., analytic dashboards). Migrating to a cloud infrastructure allows the data platform to handle huge amounts of data flexibly and robustly. Cloud computing enables the data platform to scale resource provision up and down according to real-time needs. Computing resources are spread over many servers across locations, unlike traditional data centers that must be physically located in a fixed building. Access to cloud computing is compatible with a variety of device types and networks.

**Machine Learning System.** This is a sub-index of data architecture, measuring one of its two stages. The machine learning (ML) system enables engineers to develop and deploy predictive models. The outputs enter various production-oriented, customer-facing, and internal applications. Questions related to the ML system assess the closeness between capabilities of the current system and those of gold-standard practices among successful digital companies. For example, such practices include DevOps and experimentation, which substantially speed up the development and productization of high-quality machine learning models. Workers use input data and interact with the machine learning (ML) system to train predictive algorithms. The current frontier practices center around the principle of DevOps, which originated from the development of enterprise management software applications. DevOps link the model development stage and the operations stage together, allowing algorithms and software code to loop back and forth between these two stages smoothly. The DevOps cycle aims at incorporating updates and transmitting results in an efficient and robust manner throughout the ML system.

\(^{11}\) Data pipelines are distinct from traditional batch-processing ETL (extract, transform, and load) pipelines. While ETL pipelines move data in chunks at pre-specified times, data pipelines treat data as streams or continuous flows, and divide them into smaller chunks and process them in parallel.
The system may also have an *experimentation platform*, which is able to test multiple models in parallel, accelerating model optimization and machine learning deployment.

**Coherence.** This is a sub-index measuring an aspect of data architecture quality. A coherent data architecture is an integrated system, where data resources are widely shared around the organization, and kept in sync across locations and business units. Particularly, traditional companies’ siloed architecture is at odds with coherence, often leading to isolated data sets generated by separate units and scattered around incompatible IT systems. Coherent data platforms deliver real-time capabilities in a robust manner: they can incorporate dynamic updates to data flows and return instantaneous results to end-user queries. They combine data from various sources, merge new data with existing data, and transmit them across the data platform and among users. Similarly, a coherent ML system closely connects the development and deployment stages within a tight loop. When changes to input data are applied to the system, it quickly reoptimizes to reflect these updates. New prediction models are pushed into production, and testing feedback delivered back into the development stage automatically. The loop is closed by replacing existing models with new ones that outperform.

**Security.** This is a sub-index measuring an aspect of data architecture quality. A secure data architecture should ensure that data are stored securely, and models are hosted reliably within the system, guarding large stores of data against breaches that result in data loss or theft. Corporation can suffer huge financial losses and lose consumer trust when data assets are compromised, leaked to competitors, or deleted altogether. Security features of the data architecture protect data and ML models hosted within, by assigning universal security policies that govern the entire system and designating appropriate access rights to different groups of internal users. Another crucial security feature is the ability to track data movement and model updates – keeping activity logs allows the company to detect anomalies in data access and model retrieval, and flag suspicious activities through auditing the records.

**Innovation Processes.** Digital intangible capital involves complementary processes in which workers and teams innovate using data and machine learning tools. Machine learning engineers, data scientists, and business analysts interact with the technical systems to create co-invention output, e.g., business intelligence, analytic dashboards, and predictive models. These interactions involve product-specific routines, workflows, and communication channels. They are often shaped over a long time. Questions related to innovation processes include individual practices and team dynamics related to working with data sets, technical tools, and machine learning artifacts. For example, one question asks about how easy it is for end-users to access data and models, when they are at different levels of decision-making authority (e.g., decentralized decision structures) and have various degrees of technical
expertise (e.g., citizen developer and low-code solutions\textsuperscript{12}). Another question asks about catalogues and documentation for data sets, model artifacts, and machine learning outputs. Large amounts of scattered information may pose barriers to new workers getting up to speed with available tools and resources across the organization. Clear documentation and proper catalogues can expedite learning among new workers.

**Technology Use.** Besides data architecture and innovation processes, other technology intensity factors that may also matter for the efficiency and quality of co-invention. We calculate the technology use index using questions that ask about the presence of specific tools and advanced capabilities at some units but not across the entire organization. The technology use index counts cutting-edge products in cloud computing and data security, e.g., Amazon EMR, Azure Information Protection, among others. Adopting these tools and products does not coordination across departments. Technology use measures piece-wise tools in firms’ technology stacks. Firms may invest heavily in specific tools, but still fail to develop the complementary capabilities and technological architecture to utilize them effectively.

**Machine Learning (ML) Deployment.** We quantify user co-invention by measuring the scope of deployed use cases of machine learning models. The survey contains comprehensive questions about machine learning applications, deployed in various business processes such as product development, customer interactions, operations and supply chain, employee management, and risk control. These applications serve a wide range of customer needs and internal functions. The scope of ML applications measures firms’ success of turning data and ML tools into valuable output.

### A.2. Other Data Sources

**Ci Technology Database (CiTDB).** We rely on the Ci Technology Database (CiTDB) for measuring legacy information technology (IT) systems and other control variables. The annual subscription contains information on software and hardware products at the establishment level for most large U.S. corporations. Sample corporations are linked to around 70,000 establishments overall, including divisional headquarters and major branches but not spinoffs and loosely related subsidiaries. We identify the U.S. headquarter of each corporation by fuzzy string match on company name and headquarter address\textsuperscript{13}, and link to other establishments (regional headquarters and branches). We construct IT-related variables per establishment, and aggregate values across establishment to the corporation level.

\textsuperscript{12} These concepts emerged as attempts to bridge skills gap that resulted from a rising demand for technical talent and a supply shortage in highly skilled engineers.

\textsuperscript{13} We use the corporate headquarter address in either Compustat or Duns & Bradstreet as benchmark, after cross-checking the two data sources to determine which one is more accurate, based on Google search results and Wikipedia articles. We identify the establishment associated with the operational headquarter, by matching locations after parsing the addresses into
Compustat. Compustat is a global database with information on public company profiles and financials. We obtain headquarters address, NAICS industry code, and annual revenue and production factors (capital stock, employment, and materials) variables for each sample corporation.

A.3. No Adverse Effect of Data Architecture on Total Employment

A long-standing debate over the relationship between technological progress and human capital centers on whether new technologies complement or substitute human labor in different scenarios, and whether they create more jobs than they destroy. We examine the relationship between data architecture and employment change at the corporation level. The measured employment is not restricted to IT labor, e.g., data scientists and machine learning engineers, but involve all workers across occupations and business functions. To detect any effect of new machine learning and data-driven technologies on employment, we estimate Equation 4 and focus on architectural “frontier closeness” as the key regressor.

\[
\Delta \log(\text{Employment})_{i, 2016-2019} = \alpha_0 + \alpha_1 \cdot \text{DataArchitecture}_{i, 2019} + \alpha_2 \text{Controls}_{i} + \nu_i
\]  

We run both OLS and IV-2SLS regressions consistent with empirical methods discussed in previous sections. Appendix Table B.4 presents regression results showing no evidence of technology deployment driven by data architecture displacing workers. All regressions control for log total employment in 2016, log company age, and log number of establishments. After accounting for headquarters region fixed effects, it appears that frontier data architecture capabilities positively affects employment growth. A one standard deviation increase (or 0.18 in magnitude) in the data architecture index is associated with around 2.3% larger employment growth from 2016 to 2019, although the estimates are noisy and standard errors are large.

[ Insert Appendix Table B.4 here ]

These estimates are based on aggregate data at the corporation level rather than at the worker level. We leave it to future work to scale up the measurement, and collect more detailed data on workers, to examine the causal impact of effective ML deployment on the processes and organizational structure around workers, and the distribution in wages for different occupations.

geo-coordinates using Google Places API. All sample corporations except one are in the 2015 snapshot of the CiTDB database, and the one exception is in the 2014 database. We use the latest year with available information to construct IT systems variables.
B. Appendix Tables and Figures

**Figure B.1: Fortune 1000 Rankings of Sample Corporations.** This figure shows the distribution of Fortune 1000 rankings (in year 2020) among corporations in the survey sample.
Figure B.2: Data Architecture Summary Statistics by Industry. This figure summarizes the mean and interdecile range of the data architecture index by industry, as well as the number of large companies in the sample in each industry. The NAICS codes of industries covered by the sample are: 334 (Computer & Electronic Product), 333 & 335 (Metal, Machinery, Electrical Equipment & Component), 311, 312, 315, 316, 322 & 325 (Food, Beverage, Apparel, Paper & Chemical), 336 (Transportation Equipment), 339 (Miscellaneous Manufacturing), 44 & 45 (Retail Trade), 52 (Finance & Insurance), and 62 (Health Care & Social Assistance).
Figure B.3: System Complexity Does Not Correlate with Server TPM or Data Architecture. This figure shows two sets of scatter plots: the left panel presents the relationship between system complexity and the prevalence of server third-party maintenance (TPM); the right panel presents the relationship between system complexity and the survey-based data architecture index. System complexity is measured as the share of distinct system software categories among overall (application and system) software products in the IT systems by 2015. Regressions control for industry fixed effects, total software categories, log number of establishments, log number of IT employees, log number of computers, log company age, and headquarter region fixed effects.
Table B.1: Share of Survey Sample Corporations among Fortune 1000 by Sector. This table summarizes the relative importance of corporations in the survey sample among Fortune 1000 companies in 2020. Columns (from left to right) correspond to the market value in 2020, number of firms, revenues, profits, and assets of the survey sample as percentages of Fortune 1000 totals by sector.

<table>
<thead>
<tr>
<th>Sector (Fortune 1000)</th>
<th>% Market Value in 2020</th>
<th>% Firms</th>
<th>% Revenues</th>
<th>% Profits</th>
<th>% Assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food &amp; Drug Stores</td>
<td>89.0</td>
<td>20.0</td>
<td>65.1</td>
<td>65.3</td>
<td>64.2</td>
</tr>
<tr>
<td>Aerospace &amp; Defense</td>
<td>77.9</td>
<td>27.3</td>
<td>73.7</td>
<td>69.6</td>
<td>77.5</td>
</tr>
<tr>
<td>Household Products</td>
<td>75.7</td>
<td>23.1</td>
<td>54.5</td>
<td>81.4</td>
<td>58.5</td>
</tr>
<tr>
<td>Apparel</td>
<td>73.7</td>
<td>25.0</td>
<td>52.5</td>
<td>61.2</td>
<td>51.3</td>
</tr>
<tr>
<td>Industrials</td>
<td>69.0</td>
<td>28.0</td>
<td>65.1</td>
<td>65.0</td>
<td>74.0</td>
</tr>
<tr>
<td>Healthcare</td>
<td>62.9</td>
<td>31.0</td>
<td>52.2</td>
<td>68.5</td>
<td>57.1</td>
</tr>
<tr>
<td>Financials</td>
<td>36.0</td>
<td>12.3</td>
<td>35.5</td>
<td>39.3</td>
<td>46.9</td>
</tr>
<tr>
<td>Food Beverages &amp; Tobacco</td>
<td>23.3</td>
<td>8.1</td>
<td>13.7</td>
<td>25.6</td>
<td>23.6</td>
</tr>
<tr>
<td>Chemicals</td>
<td>21.3</td>
<td>3.7</td>
<td>12.7</td>
<td>28.9</td>
<td>10.8</td>
</tr>
<tr>
<td>Retailing</td>
<td>19.6</td>
<td>13.3</td>
<td>36.2</td>
<td>29.8</td>
<td>30.1</td>
</tr>
<tr>
<td>Business Services</td>
<td>18.1</td>
<td>1.9</td>
<td>5.0</td>
<td>18.2</td>
<td>3.6</td>
</tr>
<tr>
<td>Technology</td>
<td>5.4</td>
<td>9.2</td>
<td>12.4</td>
<td>6.9</td>
<td>10.0</td>
</tr>
<tr>
<td>Materials</td>
<td>4.5</td>
<td>2.2</td>
<td>2.3</td>
<td>2.9</td>
<td>1.6</td>
</tr>
<tr>
<td>Motor Vehicles &amp; Parts</td>
<td>4.1</td>
<td>18.2</td>
<td>9.2</td>
<td>10.2</td>
<td>5.3</td>
</tr>
</tbody>
</table>
Table B.2: Robustness – Server TPM Does Not Correlate with Firm Characteristics. This table reports the regression coefficients and robust standard errors on server third-party maintenance (TPM), when the outcome variables are a set of company characteristics. The outcome variables are the technology use index (column 1), innovation processes index (column 2), log revenue in 2016 (column 3), log capital in 2016 (column 4), and log labor in 2016 (column 5), respectively. All columns control for industry fixed effects, log number of establishments, log company age in years, and average log IT staff estimates per establishment by 2015. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Tech Use (1)</th>
<th>Processes (2)</th>
<th>Log (Revenue) (3)</th>
<th>Log (Capital) (4)</th>
<th>Log (Employment) (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server TPM</td>
<td>0.014</td>
<td>-0.011</td>
<td>0.044</td>
<td>-0.241</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.018)</td>
<td>(0.058)</td>
<td>(0.221)</td>
<td>(0.063)</td>
</tr>
<tr>
<td>Log (Company Age)</td>
<td>-0.046</td>
<td>-0.051</td>
<td>-0.247</td>
<td>-0.330</td>
<td>-0.042</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.037)</td>
<td>(0.183)</td>
<td>(0.499)</td>
<td>(0.230)</td>
</tr>
<tr>
<td>Log (# Establishments)</td>
<td>0.029**</td>
<td>0.040***</td>
<td>0.574***</td>
<td>0.074</td>
<td>0.490***</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.015)</td>
<td>(0.062)</td>
<td>(0.212)</td>
<td>(0.122)</td>
</tr>
<tr>
<td>Avg Log IT Staff Per Estab</td>
<td>0.036</td>
<td>0.008</td>
<td>0.092</td>
<td>0.195</td>
<td>0.309</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.039)</td>
<td>(0.172)</td>
<td>(0.452)</td>
<td>(0.245)</td>
</tr>
<tr>
<td>No. Obs.</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
<td>117</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.202</td>
<td>0.141</td>
<td>0.553</td>
<td>0.413</td>
<td>0.259</td>
</tr>
</tbody>
</table>
Table B.3: Robustness – Server TPM Does Not Correlate with Cloud Computing. This table reports the regression coefficients and robust standard errors on server third-party maintenance (TPM) when the outcome variable is a survey-based index of cloud computing adoption. All columns control for industry fixed effects. Baseline company controls include log number of establishments and log company age by 2015. IT investment controls include average log number of computers per establishment by 2015. Headquarter region fixed effects are at the Census Region level. Production factor controls include log materials, capital (PPE) and labor in 2019. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Cloud Computing Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Server TPM (by 2015)</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
</tr>
<tr>
<td>Baseline Company Controls</td>
<td>✓</td>
</tr>
<tr>
<td>IT Investment Controls</td>
<td>✓</td>
</tr>
<tr>
<td>Production Factor Controls</td>
<td>✓</td>
</tr>
<tr>
<td>Headquarter Region FE</td>
<td>✓</td>
</tr>
<tr>
<td>No. Obs.</td>
<td>117</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.060</td>
</tr>
</tbody>
</table>
Table B.4: Relationship Between Data Architecture Quality and Enterprise Total Employment. This table reports both OLS and IV-2SLS regression coefficients and robust standard errors on the estimated effect of data architecture on corporation-wide employment growth from 2016 to 2019. Columns 3 – 5 are estimated using server TPM by 2015 as the instrument for data architecture. Columns 2, 3, and 5 control for the survey-based technology use index. Baseline company controls include log number of establishments by 2015, log company age by 2015, and log total employment in 2016. All columns control for the average log number of computers per establishment by 2015, and industry fixed effects. Headquarter region fixed effects are at the Census Region level. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Employment Growth (2016-2019)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS (1)</td>
</tr>
<tr>
<td>Avg Log # PCs Per Estab</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>(0.063)</td>
</tr>
<tr>
<td>Data Architecture Index</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>(0.172)</td>
</tr>
<tr>
<td>Technology Use Index</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Company Controls</td>
<td>✓</td>
</tr>
<tr>
<td>Headquarter Region FE</td>
<td></td>
</tr>
<tr>
<td>No. Obs.</td>
<td>117</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.075</td>
</tr>
<tr>
<td>First Stage F-Statistic</td>
<td>8.5</td>
</tr>
</tbody>
</table>