The Power of Prediction: Predictive Analytics, Workplace Complements, and Business Performance[1]

Erik Brynjolfsson
Stanford University and NBER
erikb@stanford.edu

Wang Jin
MIT Sloan School of Management
jwangjin@mit.edu

Kristina McElheran
University of Toronto
k.mcelheran@utoronto.ca

April 15, 2021

Abstract

Working with the Census Bureau, we surveyed over 30,000 American manufacturing establishments on their use of predictive analytics and rich workplace characteristics in 2010 and 2015. We find that productivity is higher among analytics-using plants, but primarily in the presence of significant IT capital, educated workers, or monitoring-intensive management practices. Both instrumental variables estimates and timing of gains suggest a causal relationship. At a time of proliferating digitization and techniques to extract value from data, these capabilities demonstrably improve productivity. However gains appear highly contingent on slow-to-adjust workplace complements, including previously unexplored managerial practices that facilitate responsiveness to improved prediction.

Keywords: digitization, data, predictive analytics, productivity, complementarities

JEL: M2, L2, O32, O33, D2

[1] Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. All errors are our own.

1. Introduction

Greater volume, variety, and timeliness of digital information have transformed what is measurable within firms and markets (McAfee and Brynjolfsson 2012). Exponential growth in computing power, declining costs of information technology (IT), and the rise of new computational methods – notably machine learning[2] - have fueled efforts to extract more value from data (Tambe 2014, Bughin 2016, Wu et al. 2020). Worldwide revenues for "big data" and business analytics solutions is forecasted to reach $274.3 billion by the year of 2022 with a 13.2% annual growth rate (*IDC* 2019). Automation of analytics via AI is gaining attention and investment (Agrawal et al. 2018, Furman and Seamans 2019). However, these investments have yet to yield productivity gains in the aggregate (Syverson 2017, Brynjolfsson et al. 2021). At the firm level, managers struggle to close the gap between advanced technological capability and actual performance (Ransbotham et al. 2015, 2017; Wu et al. 2019). And individual workers increasingly wonder about a future transformed by data-driven automation (Autor 2015, Agrawal et al. 2019). These salient and connected concerns have been difficult to tackle empircially due to the rate of technological change, the complexity of organizational interactions, and a dearth of data.

Some of these challenges are not new. Puzzlement and concern about "productivity paradoxes," how to realize busines value from new technologies, and the "future of work" have attended every major wave of technological change.[3] However, we argue and provide evience that advances in analytics and prediction have distinct dimensions that have been unexplored, challenge existing intuitions, and require a nuanced approach. They also require new measurement.

We address this challenge by providing novel and systematic evidence on the adoption, performance impacts, and enablers of predictive analytics across diverse workplace settings. Collaborating with the U.S. Census Bureau, we collect the first direct measures of predictive

---

[2] "Machine learning," "cognitive technologies," and "artificial intelligence" (AI) are increasingly used interchangeably to denote computer systems that can perceive, adapt, and learn about the environment to make predictions, recommendations, or decisions that are aligned with human-defined objectives but do not require human intervention (OECD 2019, Christian 2020). Because this capability is often rooted in large quantities of data, it is tightly linked to "big data analytics," which denotes the more-general capability to extract insights from data to support decision making (Wu et al. 2019). This study See section 2 for more details.

[3] See Brynjolfsson and Hitt (1996), Mithas et al. (2012), Bessen (2015), and Autor (2015).

anlaytics use in a large and representative sample of U.S. firms,[4] along with critical data on tangible and intangible workplace investments (Bloom et al. 2019). Linking this survey from 2010 and 2015 to an annual panel of administrative data, we estimate the productivity impact of improved prediction in over 30,000 manufacturing plants, addressing common threats to causal identification and unearthing analytics-specifc complementarities that have been unexplored to date.

We find that, while adoption of predictive analytics is widespread, performance gains are highly contingent on workplace complements. More than 70 percent of our sample adopted predictive analytics as early as 2010, with productivity benefits of 1 to 3 percent, on average. Practically, this represents roughly $410,000 to $860,000 greater sales for adopters versus non-adopters, even accounting for production inputs[5] and a wide range of other factors. However, these productivity benefits appear *primarily* at workplaces that have also pursued high levels of IT capital, educated employees, or monitoring-intensive management practices.

The need to accumulate complementary IT infrastructure and appropriately skilled workers is perhaps conceptually unsurprising in light of prior work (e.g., Bresnahan et al. 2002, Tambe 2014). Empirically, however, heterogeneity in our representative sample highlights how costly and/or time-consuming this remains in practice, even across the five years covered by our study. Moreover, controlling for general managerial capacity, the presence of specific practices focused on monitoring changes in firm performance is integral to the business value of predictive analytics.

Notably, this complementarity between managerial inputs and automation of cognitive tasks – which extends to managerial headcount[6] - is contrary to findings associated with physical automation (Dixon et al. 2021). Beyond adding nuance to a rising "robocalypse" narrative (e.g.,

---

[4] Prior work has addressed the significant measurement challenge by triangulating on analytics use in firms via the human capital needed to adopt it, typically in smaller samples (Tambe 2014; Wu et al. 2019, 2020). Our approach is complementary, with distinct advantages and challenges. See Section 2.

[5] This is a common nonparametric approach to estimating revenue-based total-factor productivity (TFPR), and is possible due to establishment-level Census data on both expenditures and capital investment over time (e.g., Bloom et al. 2019).

[6] *Pending disclosure review.*

Autor and Salomons 2017), this finding demonstrates the need for precision in both theory and measurement about the locus of automation in firms – e.g., analysis versus execution – and the specific roles that humans and technology play in a given production function.

To this point, predictive analytics is conceptually and empirically distinct from a firm culture of being "evidence-based" (Pfeffer and Sutton 2006) or managerial practices focused on "data-driven decision making" (Brynjolfsson and McElheran 2019). While definitions for emerging technology tend to shift (a signifcant hurdle for measurement), predictive analytics is increasingly understood to be a set of techniques ranging from data mining to statistical modeling – including, most recently, AI – to analyze historical and current data to make predictions about future or unknown events.[7] A large literature has modeled how better information can improve decision-making in firms (e.g., Blackwell 1953; Raiffa 1968; March 1994). This need not be particularly automated nor forward-looking (Tambe 2014). However, advances in methods to extract insights from large data sets and a focus on prediction are believed to further reduce the cognitive costs of decision-making, improve precision, and speed execution (e.g., Agrawal et al. 2019). Importantly, these tools automate analysis and inform decision making, but they typically still stop short of automatically executing those decisions.

Adoption of management practices and techniques that leverage data has been linked to higher Tobin's q and profits (Brynjolfsson et al. 2011; Saunders and Tambe 2015), productivity (Tambe 2014, Brynjolfsson and McElheran 2019, Wu et al. 2020), and innovation (Wu et al. 2019). However, evidence for the causal benefits of analytics focused on *prediction* remains lacking. Further, while prior research explores how organizational factors may explain variation in the returns to information technologies (Melville et al. 2004 provide a review), the role of specific management practices that facilitate taking action on data-driven insights has been unexplored. This study makes progress along a few dimensions.

First, our data come from a purpose-designed mandatory survey covering more than 50% of the U.S. manufacturing economy. We directly capture early use of a fast-changing technological

[7] This interpretation is corroborated by extensive testing of the survey instrument led by Census experts on survey development. See the Data Appendix for further details.

advance, which is a persistent challenge for studying the digital frontier, particularly when "hype" far outstrips evidence of benefits or guidance for implementation (Raj and Seamans 2018, McElheran 2019, Raj and Seamans 2019). Our measure complements prior studies of analytics (Tambe 2014; Wu et al. 2019, 2020) while emphasizing prediction -- a distinction often emphasized in practice (e.g., Blum et al. 2015).

Next, we provide two types of causal evidence that predictive analytics use increases firm productivity. First, we leverage our survey question about government mandates to collect data as an instrument for greater predictive analytics use. Second, we compare the timing of adoption with the timing of productivity gains, the results of which are consistent with a causal interpretation.

Finally, linking multiple Census data sets to capture tangible and intangible workplace investment isolates first-order considerations that have gone largely unobserved to date. Because data on production inputs and outputs in this sector is well-established over a long time, we can build on earlier research into productivity dispersion (e.g., Bartelsman and Doms 2000, Syverson 2011, Collard-Wexler and deLoecker 2015, and White et al. 2018) and digitization (e.g., Doms et al. 1995, Black and Lynch 1996 & 2001) in a well-understood setting. By adding new visibility to a range of tangible and intangible workplace characteristics, we address a large number of potential confounds and show that complementary investments in IT and skilled labor are not only indispensable to the productivity of predictive analytics, but also slow to arrive. Data on management practices, in particular, reveals that a firm's capacity to act on data-driven insights is a novel and central contingency. The magnitude of these organizational interactions is large: in our sample, predictive analytics only contribute to productivity when combined with at least one of these complements.

Our findings contribute to several areas of research. We build on early research in IT productivity that emphasizes heterogeneity across industries and firms (e.g., Stiroh 2002, Brynjolfsson and Hitt 1995) and theory arguing that this heterogeneity may arise from investments in complementary assets and managerial practices (Kandel and Lazear 1992; Milgrom and Roberts 1990, 1995; Holmstrom and Milgrom 1994; Athey and Stern 1998; Brynjolfsson and Milgrom 2013; Brynjolfsson et al. 2021). A number of empirical studies have supported this theory with respect to general-purpose IT and computer use (Black and Lynch 2001; Bresnahan et al. 2002;

Aral and Weill 2007; Bloom et al. 2012), IT applications (Dranove et al. 2014), and earlier waves of data-centered management practices (Aral et al. 2012, Tambe et al. 2012, Tambe 2014; Brynjolfsson and McElheran 2019). Our study is the first to apply this line of exploration to rising improvements in prediction.

2. Setting and Data

While many technologies and management practices are information-enhancing, predictive analytics is distinct from and arguably more sophisticated than other approaches.[8] Predictive analytics leverages computer systems to rapidly analyze much larger data sets than would otherwise be humanly possible, potentially generating returns from digital information that is rapidly becoming cheaper to gather and rising in volume and complexity.

Empirical exploration of these claims has been hampered by the lack of large-scale data on predictive analytics use. To address this, we collaborated with the U.S. Census Bureau to add new, purpose-designed questions to the 2015 Management and Organizational Practice Survey (MOPS).[9] Survey response is required by law, yielding a response rate of 70.9%, with 30,000 complete establishment-level observations.[10]

Adding questions to Census surveys requires rigorous cognitive testing (Buffington et al. 2017), essential for measuring such a recent and fast-emerging technology across different industry settings. The resulting focal question asks, "How frequently does this establishment typically rely on predictive analytics (statistical models that provide forecasts in areas such as demand, production, or human resources." Respondents – typically a senior plant manager or accounting expert with the help of business function or line managers– are asked to mark all that apply among *Never, Yearly, Monthly, Weekly,* and *Daily,* with separate columns for 2015 and recall for 2010. With the recall data in 2010, we have in total 51000 observations for both years.[11]

---

[8] Prior work supports a meaningful distinction between relying on predictive analytics and being "data-driven" (Brynjolfsson and McElheran 2019), as well as differences between descriptive and predictive analytics (Berman Israeli 2020).

[9] See Bloom et al. (2019) and Buffington et al. (2017) for more details.

[10] Additional details in the Data Appendix.

[11] Note that the sample counts are rounded for disclosure reasons throughout the paper. We use the total number of observations (~51,000) as our baseline sample but all key results are robust to a sub-sample with a higher quality recall

We flag the extensive margin of analytics use, regardless of frequency, as there may be heterogeneity in inputs (such as data quality) that remain unobserved. We capture variation along the intensive margin with a numeric value ranging from 0 to 4 for each frequency category in ascending order, defaulting to the highest in cases of multiple categories. We lean on this more-continuous measure, in particular, in our instrumental variables (IV) estimation.

Predictive analytics was widely diffused among manufacturing plants across almost all states and industries as early as 2010 (Brynjolfsson et al. 2021), with average adoption well over 70% (Table 1). Among the roughly 18,000 establishments with complete data for both years,[12] we observe a small 1.4% average yearly increase.[13]

This high penetration and low rate of change have implications for our empirical approach: in particular, they forestall estimation of within-plant effects over time for two key reasons. First, focusing on changes in the subpopulation of lagging adopters would bias our inference. Later adopters of new technologies tend to be those with low anticipated returns, disproportionately high costs of adoption, and/or lagging awareness of the technology (Griliches 1957; David 1969; Bresnahan and Greenstein 1996). This is not the core population of interest for our research question, though we do explore whether organizational complements – or lack thereof – constitute barriers to adoption, as well as productivity returns. We also address selection into adoption to some degree using our IV. Even so, statistical power in the subsample of establishments that shift their predictive analytics use is severely limited, despite the overall size of our data set.

Workplace Complements and Plant Performance

Complementarities among IT investments and organizational characteristics have long been associated with differences among firms that may persist and even grow over time (Milgrom and Roberts 1990, 1995; Black and Lynch 2001; Caroli and Van Reenen 2001; Bresnahan et al. 2002; Bloom et al. 2012; Aral et al. 2012; Tambe et al. 2012; Brynjolfsson and Milgrom 2013). More

---

structured management score where respondent tenure started at least one year before the period of the recall (Bloom et al. 2019).

[12] The rotation of the ASM sample in years ending with "4" and "9" limits the number of establishments that have complete data for both years in our sample, with a core "certainty sample" of larger plants present in both.

[13] The adoption of predictive analytics rose from 73% in 2010 to 80% by 2015.

recently, heterogeneity in firm performance and workplace conditions has attracted increasing attention.[14] Rising concentration and inequality in workplace conditions and employee earnings are increasingly attributed to technology investment within industries and firms (Bessen 2017; Bennet 2020b; Lashkari et al. 2020; Barth et al. 2020). In addition, difficult-to-measure intangible features of firms and markets are argued to amplify these dynamics (Saunders and Brynjolfsson 2016; Haskel and Westlake 2018). Building on this work, we explore a few key tangible and intangible complements to predictive analytics that might shape returns to its use.

IT Capital Stock

The collection, storage, and communication of data inputs for predictive modeling all require tangible investments in things like sensors, transmission equipment, and data storage hardware. Building, training, and implementing analytics tools require corresponding data processing hardware and software. Thus, firms with existing IT capital investments that are more prepared for the industrial Internet of Things (IoT) and related "big data" innovations may possess fully-depreciated investments in infrastructure to collect and analyze data, as well as richer data inputs, giving them an advantage when it comes to analytics.

To explore potential complementarities along this dimension, we calculate IT capital stocks using capital expenditure on computer and peripheral data processing equipment from the ASM and CMF panel dating back to 2002, using a standard perpetual inventory approach and an industry-level deflator for hardware from the Bureau of Economic Analysis (BEA). We impute values for years in which values are missing and depreciate at the rate of 35% per year following Bloom et al. (2014).

Educated Workers

Prior work has established that more-skilled and better-educated workers are key drivers of growth in both manufacturing productivity (Black and Lynch 2001; Moretti 2004) and returns to

[14] While high cross-sectional heterogeneity in firm performance is long-established (e.g., Syverson 2004 & 2011; Hopenhayn 2014), recent studies point to increasing firm heterogeneity along a number of economically important dimensions (Andrews et al. 2015; Van Reenen 2018; Song et al. 2019; Decker et al. 2020; Autor et al. 2020; Bennett 2020a). This phenomenon is not restricted to the United States (e.g., Berlingieri et al. 2017).

IT (Brynjolfsson and Hitt 2000; Bresnahan et al. 2002). With increasing digitization and growing prevalence of business applications for data and predictive analytics, firms increasingly need workers that know how to deploy "smart" technologies in production settings (Helper et al. 2019) as well as those who can help translate digital information into business insights (Ransbotham et al. 2015). While competition for these workers may drive up wages to a point where there remain no excess returns to labor in a well-specified production function, we expect that worker skill will boost estimates of predictive analytics' contribution to productivity due to these complementarities.

We leverage information from the MOPS regarding the percentages of managers and non-managers with bachelor's degrees. Combined with the total number of employees (from the ASM) and the number of managers (from the MOPS), we calculate the weighted average of the percentage of employees (both managers and non-managers) with a bachelor's degree following Bloom et al. (2019). This approach is similar to prior studies using education as a proxy for human capital (Card 1999; Bresnahan et al. 2002).

Key Performance Indicator (KPI) Monitoring

Manufacturing firms have long been developing and monitoring KPIs to support a variety of operational practices such as Lean and Total Quality Management. Firms often track capacity utilization rate, manufacturing cycle time, equipment downtime, and many others. Tracking KPIs is highly correlated with the breadth and intensity of data collection and its use in decision making (Brynjolfsson and McElheran 2019). For predictive analytics, anecdotal evidence has shown that plants with well-established KPI practices will have richer inputs to feed their predictive models, potentially lowering the cost or boosting the returns to predictive analytics (Schrage and Kiron 2018). In addition, these practices reflect investments and routines that enable firms to contextualize and intepret anlatyics-driven insights, a necessary step for taking action in response to new information when execution relies on managerial intervention.

On the MOPS, respondents indicate the number of KPIs monitored at the plant, with options to select 0, 1-2, 3-9, or 10 or more. Examples given include metrics on production, cost, waste, quality, inventory, energy absenteeism, and deliveries on time. We take the top category as a proxy

for establishments with high intensity of KPI monitoring to test our hypothesis that robust KPI monitoring is a complement to predictive analytics.

Table 1 presents key summary statistics for our sample. Despite the high prevalence of some use of predictive analytics, the intensive margin is more modest, with most plants reporting only annual and/or monthly use (mean frequency is 1.12). Although our sample represents the majority of economic activity in the sector, very small establishments are underrepresented; sample mean annual sales and employment in log terms are 10.37 and 4.56, respectively, or about $32,000,000 and 96 employees. The mean plant has roughly $175,000 of IT capital stock and slightly over 15% of workers with a bachelor's degree. Around 44% of plants track 10 or more KPIs.

3. Empirical Approach

Correlation Test for Complementarity

Beyond providing the first large-scale descriptive statistics on adoption, our empirical analysis explores correlations between predictive analytics use and key potential workplace complements. If complementarities exist, we should observe higher adoption of predictive analytics for establishments with these investments and practices (Brynjolfsson and Milgrom 2013). We explore correlations with IT capital stock, educated workers, and top KPI monitoring in both linear probability and probit models, including a rich set of workplace controls. These include: general reliance on data in decision-making (Brynjolfsson and McElheran 2019), an index of other structured management practices (Bloom et al. 2019), establishment age, multi-unit status, firm headquarters status, and production process design.[15] We also control for geographic differences and industry-year fixed-effects to account for any transitory industry-specific shocks.

[15] These controls have been correlated with technology adoption and productivity in prior work. Our management index differs from that in Bloom et al. (2019) by excluding the data-related MOPS questions. See Dunne (1994) and Foster et al. (2016) for more on plant age, technology adoption, and performance. See Collis et al. (2007) for discussion of multi-unit and headquarter status. See Safizadeh et al. (1996) for more on manufacturing process designs. This set of controls is in all fully specified models for adoption and performance analysis unless stated otherwise. See data appendix for construction of the control variables.

Performance Analysis

Next, we explore both the average causal effect of predictive analytics and standard performance tests of complementarity. We take a conventional approach to modeling the plant production function (Brynjolfsson and Hitt 2003; Bloom et al. 2012) estimating the log-transformed Cobb-Douglas production function in equation (1):

$$Log(Y_{ijt}) = \beta_0 + \beta_{pa}\log(PA_{ijt}) + \beta_k \log(K_{ijt}) + \beta_l \log(L_{ijt}) + \beta_m \log(M_{ijt}) + \mu X_{ijt} + w_{ijt} + \varepsilon_{ijt} \quad (1)$$

Y_{ijt} is sales by establishment i in industry j at time t, K denotes non-IT capital stocks at the beginning of the period, PA is an indicator (or frequency measure) for use of predictive analytics, L is labor input, M is consumption of material and energy inputs, and X is a vector of abovementioned controls and the three potential complements. Both w_{ijt} - the "technical productivity" and ε_{ijt} - the "shock to productivity" are unobservable econometrically (but w_{ijt} might be observable by establishments). Our first coefficient of interest is β_{pa}, the average relationship between predictive analytics and plant productivity, all else equal.

Instrumental Variable Estimation

A standard concern with this approach is that predictive analytics use may be endogenously determined, biasing interpretation of β_{pa}.[16] To address this, we explore IV estimation using an indicator of data collection driven by government regulation or agencies.

The motivation for this instrumentation strategy rests on the so-called "Porter Hypothesis" (Porter 1991; Porter and Van der Linde 1995) arguing that well-designed government regulations can stimulate firms to innovate and adopt new technology and practices. Of relevance in our setting, data collection is often mandated by federal and local governments to demonstrate compliance

[16] This will happen if plants with higher expected returns to predictive analytics use will choose to adopt, upwardly biasing estimates of the average treatment effect. Tambe and Hitt (2012) provide a useful discussion suggesting that such concerns may be overemphasized in IT productivity studies. System GMM and other semi-structural estimation methods (see Arellano and Bond 1991; Blundell and Bond 2000; Levinsohn and Petrin 2003; Ackerberg et al. 2015) have performed well in recent studies of IT productivity (e.g., Tambe and Hitt 2012; Nagle 2019). However, our two-year panel lacks the longer lags typically required.

with environmental and safety regulations. For instance, the Environmental Protection Agency (EPA) requires manufacturing firms (e.g. pulp and paper, petroleum, and chemical manufacturing) to install continuous emission monitoring systems (CEMS) for emission data collection and monitoring. Leveraging this data in government-mandated reports requires that workers and managers be trained in systems and techniques for capturing, analyzing, and communicating data-driven conclusions. Consistent with this mechanism, firms are more likely to build infrastructure and managerial systems for data collection, storage, and analysis when required to collect new data and devise new reports by statute.[17]

Not all firms will be able to translate this into improved management of their production processes.[18] For some, however, this external "nudge" into increased investment in and awareness of data resources may shift practices on the margin among plants exposed to this additional oversight. The unexpected consequences can be striking. The case of Alcoa Corporation in the late 1980s and 90s is illustrative. When Paul O'Neil took leadership of the firm, his unexpected mandate to prioritize safety resulted in an abundance of data about accidents – but also about the performance and maintenance of infrastructure and workplace practices underlying those accidents. New data enabled new performance metrics, which were analyzed with increased frequency and linked to manager pay at the firm (*Fortune* 1991). The end result was not only improved worker safety but also improved productivity (Clark and Margolis 1991).

For this to be useful as an instrument, such oversight needs to be unrelated to the productivity of affected plants. Historically, U.S. government regulations in the manufacturing sector have fit this description. For instance, the objective of EPA CEMS requirement or OSHA's recordkeeping rule is restricted to public health and worker safety rather than plant performance. Although objections to such regulation have typically argued that they divert resources from other

[17] Abundant anecdotes support the prevalence of this phenomenon. the Occupational Safety & Health Administration (OSHA) Recordkeeping rule can serve as another example where they require about 1.5 million employers in the United States to keep records of their employees' work-related injuries and illnesses under the Occupational Safety and Health Act of 1970. For more details on OHSA Recordkeeping rule, please see the OSHA website https://www.osha.gov/recordkeeping2014/records.html

[18] Note that plants already collecting and using data extensively may be less responsive to our instrument, which we discuss below.

productivity-enhancing activities and investments (Gollop and Roberts 1983; Gray 1987), empirical evidence suggests that many well-designed regulations have had a limited negative impact on manufacturing competitiveness or overall performance (Jaffe et al. 1995; Lanoie et al. 2011; Ambec et al. 2013). Nevertheless, the standard expectation is that the direct effect will work against a positive relationship between productivity and government mandates to collect data.

Following these arguments, government-mandated data collection should satisfy both the relevance and exclusion restrictions for a valid instrument. As a practical matter, capturing this regulatory nudge at a sufficiently granular level is challenging. We addressed this by including another new question on the MOPS that captures government authority (among other decision makers) over what type of data is collected at the plant.[19]

Timing and Causality

Another threat to identifying a causal link between analytics and firm performance is the possibility that an unrelated productivity shock provides resources needed to invest in new technology or practices –not the other way around. To address this, we explore the relationship between timing of adoption and timing of productivity changes. Leveraging annual data on inputs and output from the ASM and CMF, we construct a panel from 2010 to 2016 for a large subsample of our data. We exploit the recall questions to place plants in three categories: those that had adopted predictive analytics "early" by 2010, "middle" adopters (between 2010 and 2015), and "laggard" plants that had not adopted by 2015. Leaning on evidence that many (if not most) of the organizational practice measures in the MOPS are quasi-fixed over this period (Mundlak 1961; Bloom et al. 2007), we extrapolate the organizational complements outside of our core sample window and estimate comparable yearly production functions for these differently-timed groups from 2010-2016.

If predictive analytics use causes better productivity, early adopters should have a performance premium compared to both middle adopters and laggards at or near the start of our panel. In the 2010-2016 window, middle adopters should outperform laggards.

---

[19] See Data Appendix for more details. This has been used in a related study of data-driven decision making by Brynjolfsson and McElheran (2019).

To test for this, we again rely on pooled OLS estimation with industry-year fixed-effects and rich organizational controls. It is worth noting here that, in addition to the limits on panel data estimation discussed above, the 5-year gap in our two-period panel generates additional measurement error in this undertaking. For instance, our "middle" adopters may have adopted at any time in the 2010-2015 window; thus we anticipate estimates will be considerably noisier in this analysis.

Performance Tests of Complementarity

After addressing the questions of causality in the baseline performance model, we proceed to a formal test of performance differences to explore complementarities. If complementarities exist, predictive analytics use together with a given complement should lead to higher performance than stand-alone use. Following the empirical strategy in Athey and Stern (1998) and Brynjolfsson and Milgrom (2013), our empirical specifications follow equation (2):

$$Log(Y_{ijt}) = \beta_0 + \beta_{pa}\log(PA_{ijt}) + \beta_c C_{ijt} + \beta_{interaction}\log(PA_{ijt}) \times C_{ijt} + \beta_k \log(K_{ijt}) + \beta_l \log(L_{ijt}) + \beta_m \log(M_{ijt}) + \mu X_{ijt} + w_{ijt} + \varepsilon_{ijt} \quad (2)$$

All variables in equation (2) are identical to those in equation (1) except C_{ijt}, which denotes, respectively, indicators for high IT capital stock, high percentage of educated workers, and top KPI monitoring. A positive and significant $\beta_{interaction}$ term is indicative of such performance complementarities.

4. Results

Figure 1 depicts the correlation tests for complementarity based on a linear probability model of whether or not the plant uses predictive analytics (at any frequency). Besides the potential complements and the aforementioned establishment controls, an indicator for government influence on data collection is also included, anticipating our IV estimation. See Table A2 in the Appendix for additional details.

IT infrastructure, educated workers, and reliance on KPIs are all significantly correlated with the uses of predictive analytics at the one-percent level or higher. Figure 1 is organized to show the estimated increase in probability from the benchmark of 20.6% that is associated with the

presence of each potential complement.[20] Being in the top 10% of the IT capital stock distribution by sample year increases the likelihood of predictive analytics use by 1.52%. Being in the top quartile for the percentage of employees with a Bachelor's degree is associated with a 2.75% increase; adding them together suggests a 4.27% greater likelihood. Top KPI tracking adds 4.62% more. A workplace with all three in place is 8.89% more likely to use predictive analytics than a workplace without any of these reinforcing practices and investments. This is consistent with complementarity where rational managers (e.g. consciously maximize profit and increase performance) will seek to adopt complementary practices together. Having government mandated data collection is also associated with 1.96% greater likelihood of predictive analytics adoption, satisfying the relevance condition to serve as a valid IV.

Table 2 sets up our performance tests for complementarity by first exploring the average conditional correlation between predictive analytics use and total factor productivity (TFP). Using logged sales as the dependent variable and controlling for conventional inputs (i.e. non-IT capital, labor, materials, and energy) and the above-mentioned establishment controls, we arrive at an estimate of revenue-TFP (Foster et al. 2021). All columns further include industry-year controls at the narrow 6-digit NAICS level. For example, this captures the difference between Folding Paperboard Box Manufacturing (NAICS 322212) and Setup Paperboard Box Manufacturing (322213).

Column 1 indicates that the extensive margin of predictive analytics use is associated with a roughly 2.69 percent (significant at the one-percent level) higher productivity, all else equal. This magnitude is large, representing $860,000 greater sales at the sample mean ($32M) while holding many other factors constant. Column 2 adds the potential complements. The coefficient drops significantly to around 1.91 percent, consistent with an important role for organizational complements that load onto the analytics variable when not directly accounted for. Column 3 includes controls for an indicator for plants with top quartile workplace management practices (Z_MGMT) and use of data-driven decision making (DDD). This addresses concerns that

[20] This number the based on the constant term in column 2 Table A2, which represents the average adoption controlling for all covariates in our model.

unobserved management quality or style could be affecting our estimates (Brynjolfsson and McElheran 2019); indeed they further reduce the coefficient on predictive analytics use. Nonetheless, the coefficient remains both statistically significant and economically non-trivial: 1.28 percent higher productivity is commensurate with $410,000 higher sales, on average, in excess of any costs of implementing the practice.

Column 4 explores the intensive margin of predictive analytics use. The coefficient on the frequency index is positive and significant at the one-percent level and economically meaningful. Based on its mean and standard error (see Table 1), moving from yearly to monthly use of predictive analytics is associated with 0.89 percent higher productivity, equivalent to roughly $285,000 higher sales.

We explore the robustness of these patterns to alternative measures for both the dependent and independent variables. Results are presented in Table A3 in Appendix. Our findings are robust to using labor productivity or estimated TFP[21] as the output measure, a translog production function, and alternative measures for the use frequency of predictive analytics.

Up to now, we have explored the pooled OLS regressions without considering measurement error or endogeneity in plant adoption of predictive analytics. Column 5 reports IV estimation using government-mandated data collection as an instrument for the predictive analytics index.[22]

Consistent with Figure 1, the first stage of our two-stage least square (2SLS) estimation shows that government-mandated data collection is highly correlated with the use of predictive analytics (see Appendix Table A4 for the first-stage results). In the second stage, the effect of predictive analytics on plant productivity remains large, positive, and statistically significant. The larger magnitude is consistent with downward bias in our OLS model. This is consistent with errors-in-variables arising from measurement error – something that has been found in other MOPS measures (Bloom et al. 2019).

Not mutually exclusive, this pattern is also consistent with strong local treatment effects,

[21] We employ the conventional 4-factor TFP using cost of material, energy, labor, and capital stock (Bartelsman and Gray 1996; Foster et al. 2021)

[22] Using the index for frequency of predictive analytics for the IV estimations avoids potential complications due to non-linear first stage estimation and also help capture better variation among the plants' use of predictive analytics.

whereby the subsample of workplaces that are the most receptive to the influence of the government mandate also experience the greatest productivity shift. This could arise if less-data-savvy (and likewise less-productive) plants enjoy larger indirect gains from data collection in response to requests from regulators (the "Porter Hypothesis" mechanism). In this vein, it is worth re-emphasizing that estimates in columns 3-5 control for management practices that are typically unobserved in other studies but strongly associated with higher productivity in this sector (Bloom et al. 2019; Brynjolfsson and McElheran 2019).

We further probe this causal interpretation by exploring the timing of adoption and performance in our panel (see Section 3). Figure 2 plots the coefficients for "early" and "middle" indicators in the performance model between 2010 and 2016. Consistent with our hypothesized causal relationship, early adopters perform significantly better from the start of our panel and retain their advantage vis-a-vis non-adopters through 2016. The performance for later adopters is not significantly different from that of "laggard" non-adopters through 2013. We know that these plants adopted in the 2010-2015 window, but not precisely when. Consistent with steadily increasing diffusion over time (e.g., Hall 2003 and 2004), performance in this group beings to rise and is significantly different from non-adopters by 2014. Notably, these later adopters close the performance gap, becoming statistically indistinguishable from early adopters by 2016. This could be due to the rising diffusion of predictive analytics use, if competitive convergence in these practices erodes excess returns over time. However, technical productivity compared to never-adopters seems to persist. The difference in short-run versus long-run gains to predictive analytics implied here is x%, which, given the measurement error with respect to timing, provides something of a lower bound on the phenomenon. Regardless of the reasons for convergence over time, the overall patterns are inconsistent with reverse causality.[23]

We move next to our core findings, focused on whether the factors that help explain analytics use are associated with disproportional returns, as complementarity theory would predict. These results are presented in Figure 3. The y-axis indicates the magnitude of the coefficients and the x-axis labels indicate the categories for each complement (e.g. predictive analytics with high IT

[23] Regression results for Figure 2 available upon request.

capital stock and without high IT capital stock). Confidence intervals (at 95%) are plotted to indicate statistical significance. All three interaction terms are positive and significant, consistent with strong complementarity. Significance of the differences between the interacted terms and main effects is also reported in the figure. The striking pattern that emerges is that the marginal effects of predictive analytics are never statistically different from zero, *unless* they are combined with these other tangible and intangible workplace investments.

One particular concern is that the identified complements are proxies for other relevant workplace practices (e.g. better overall management or other generic data-driven practices given their positive correlation with the adoption of predictive analytics and the complements). We conduct falsification tests by interacting the measures of structured management practices with the adoption of predictive analytics using similar specifications as above. On the contrary, no significant complementarity with the adoption of predictive analytics is found on plant performance for this measure (see results in the Appendix Table A5 column 5).[24]

Alternatively, the three identified complements could be proxies for an underlying coherent system that boosts the effect of predictive analytics (see Brynjolfsson and Milgrom 2013). We extract a principal component from principal component analysis (PCA) using these three identified complements to capture the common variation. We then interact this component with the predictive analytics use in the performance analysis for the complementarity test. Results are similar to the ones using the three complements and are also plotted in Figure 1. Along with the fact that all three complements contribute distinctly and significantly to the extract component, indicating that the identified characteristics are valid complements for predictive analytics use. Results for both falsification test and PCA are reported in Appendix Tables A5-A6.[25]

Overall, these results not only provide evidence in support of complementarities, but they provide clear boundary conditions on the phenomenon and practical guidance for managers of organizations considering these practices.

---

[24] Note that we also conduct similar exercise for the DDD-related practices and found no significant complementary effect for performance between the DDD indicator and the adoption of predictive analytics. Results are omitted to save space but available upon request.

[25] Also worth noting that our results for the performance complementary tests are largely robust when using continuous measures for the potential complements and structured management index with the only exception for IT capital stock.

5. Conclusion [under construction]

With the explosion of digital information and substantial growth in business expenditure on data and analytics, it is more crucial than ever to understand the magnitude and mechanism of effects of these practices on business performance. Although compelling anecdotal and small-sample evidence exists that predictive analytics is associated with improved performance in some settings, stories of frustration and unrealized potential also abound. Large sample microdata, controls for a growing list of potential confounds, and support for causal interpretations has heretofore been in short supply. Detailed understanding of complementarities with workplace investments in tangible and intangible infrastructure and management practices has been even more difficult to come by, even in smaller studies, due to data limitations.

To address these gaps, we worked with the U.S. Census Bureau to field a purpose-designed survey, resulting in large-scale on both data-related practices and key organizational characteristics in a sample designed to be representative of the U.S. manufacturing sector.

We find that plants reporting use of predictive analytics show almost 1% to 3% higher productivity on average, which is worth roughly $410,000 to over $860,000 in increased sales for the average plant in our sample. However, the effect is confined entirely to plants that also have high IT captial investment, educated workers, or robust monitroing practices.

These findings build on prior work exploring complementarities between organizational characteristics and IT (e.g., Bresnahan et.al. 2002, Brynjolfsson and Milgrom 2009) and shed new light on potential drivers of organizational frictions during the diffusion of new technologies.

References

Ambec, S., Cohen, M.A., Elgie, S. and Lanoie, P., 2013. The Porter hypothesis at 20: can environmental regulation enhance innovation and competitiveness?. *Review of environmental economics and policy*, 7(1), pp.2-22.

Aral, S. and Weill, P., 2007. IT assets, organizational capabilities, and firm performance: How resource allocations and organizational differences explain performance variation. *Organization science*, 18(5), pp.763-780.

Aral, S., Brynjolfsson, E. and Wu, L., 2012. Three-way complementarities: Performance pay, human resource analytics, and information technology. *Management Science*, 58(5), pp.913-931.

Arora, A. and Gambardella, A., 1990. Complementarity and external linkages: the strategies of the large firms in biotechnology. *The journal of industrial economics*, pp.361-379.

Athey, S. and Stern, S., 1998. An empirical framework for testing theories about complimentarity in organizational design (No. w6600). National Bureau of Economic Research.

Autor, D (2015) Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives*. 29(3): 3-30

Black, S.E. and Lynch, L.M., 2001. How to compete: the impact of workplace practices and information technology on productivity. *Review of Economics and statistics*, 83(3), pp.434-445.

Blackwell, D., 1953. Equivalent comparisons of experiments. The annals of mathematical statistics, pp.265-272.

Bloom, N., Sadun, R. and Van Reenen, J., 2012. Americans do IT better: US multinationals and the productivity miracle. *American Economic Review*, 102(1), pp.167-201.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R.S., Saporta Eksten, I. and Van Reenen, J., 2013. Management in America. US Census Bureau Center for Economic Studies Paper No. CES-WP-13-01.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R., Patnaik, M., Saporta-Eksten, I. and Van Reenen, J., 2019. What drives differences in management practices?. *American Economic Review*, 109(5), pp.1648-83.

Bresnahan, T.F., Brynjolfsson, E. and Hitt, L.M., 2002. Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The quarterly journal of economics*, 117(1), pp.339-376.

Brynjolfsson, E. and Hitt, L., 1995. Information technology as a factor of production: The role of differences among firms. *Economics of Innovation and New technology*, 3(3-4), pp.183-200.

Brynjolfsson, E. and Hitt, L.M., 2000. Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic perspectives*, 14(4), pp.23-48.

Brynjolfsson, E. and Hitt, L.M., 2003. Computing productivity: Firm-level evidence. *Review of economics and statistics*, 85(4), pp.793-808.

Brynjolfsson, E., Hitt, L.M. and Kim, H.H., 2011. Strength in numbers: How does data-driven decision-making affect firm performance?. Available at SSRN 1819486.

Brynjolfsson, E., & McElheran, K. 2016. The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review*. https://doi.org/10.1257/aer.p20161016

Brynjolfsson, E., & McElheran, K. 2019. Data in Action: Data-Driven Decision Making and Predictive

Analytics in U.S. Manufacturing. Rotman School of Management Working Paper No. 3422397. Available at SSRN: https://ssrn.com/abstract=3422397

Brynjolfsson, E. and Mendelson, H., 1993. Information systems and the organization of modern enterprise. *Journal of Organizational Computing and Electronic Commerce*, *3*(3), pp.245-255.

Brynjolfsson, E. and Milgrom, P., 2013. Complementarity in organizations. The handbook of organizational economics, pp.11-55.

Brynjolfsson, E., Rock, D. and Syverson, C., 2018. The productivity J-curve: How intangibles complement general purpose technologies (No. w25148). National Bureau of Economic Research.

Buffington, C., Foster, L., Jarmin, R. and Ohlmacher, S., 2017. The management and organizational practices survey (MOPS): An overview 1. Journal of Economic and Social Measurement, 42(1), pp.1-26.

Card, D., 1999. The causal effect of education on earnings. In *Handbook of labor economic*s (Vol. 3, pp. 1801-1863). Elsevier.

Caroli, E. and Van Reenen, J., 2001. Skill-biased organizational change? Evidence from a panel of British and French establishments. *The Quarterly Journal of Economics*, 116(4), pp.1449-1492.

Clark, K.B. and Margolis, J.D., 1991. Workplace safety at Alcoa (A).

Côrte-Real, N., Ruivo, P. and Oliveira, T., 2014. The diffusion stages of business intelligence & analytics (BI&A): A systematic mapping study. *Procedia Technology*, 16, pp.172-179.

Davenport, T.H., 2006. Competing on analytics. Harvard business review, 84(1), p.98.

Dixon, J, Hong B and Wu L (2021) The robot revolution: Managerial and employment consequences for firms. *Management Science*. https://doi.org/10.1287/mnsc.2020.3812

Dubey, R., Gunasekaran, A., Childe, S.J., Blome, C. and Papadopoulos, T., 2019. Big Data and Predictive Analytics and Manufacturing Performance: Integrating Institutional Theory, Resource-Based View and Big Data Culture. *British Journal of Management*, 30(2), pp.341-361.

Duhigg, C., 2012. The power of habit: Why we do what we do in life and business. Random House.

Dunne, T., 1994. Plant age and technology use in US manufacturing industries. *The RAND Journal of Economics*, pp.488-499.

Gandhi, A., Navarro, S. and Rivers, D., 2017. How heterogeneous is productivity? A comparison of gross output and value added. *Journal of Political Economy, 2017*

Enke, B., 2020. What you see is all there is. *The Quarterly Journal of Economics*, *135*(3), pp.1363-1398.

Forman, C., Goldfarb, A. and Greenstein, S., 2002. *Digital dispersion: An industrial and geographic census of commerical internet use* (No. w9287). National Bureau of Economic Research.

Foster, L., Haltiwanger, J. and Syverson, C., 2016. The slow growth of new plants: learning about demand?. *Economica*, *83*(329), pp.91-129.

Grover, V., Chiang, R.H., Liang, T.P. and Zhang, D., (2018). Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems*, 35(2), pp.388-423.

Hall, B.H., Innovation and Diffusion. In *The Oxford Handbook of Innovation*.

Haskel, J and Westlake S (2018) *Capitalism without capital: The rise of the intangible economy.* Princeton University Press.

Henderson, R.M. and Clark, K.B., 1990. Architectural innovation: The reconfiguration of existing product

technologies and the failure of established firms. *Administrative science quarterly*, pp.9-30.

Holmstrom, B. and Milgrom, P., 1994. The firm as an incentive system. *The American economic review*, pp.972-991.

Jaffe, A.B. and Palmer, K., 1997. Environmental regulation and innovation: a panel data study. *Review of economics and statistics*, 79(4), pp.610-619.

Jaffe, A.B., Peterson, S.R., Portney, P.R. and Stavins, R.N., 1995. Environmental regulation and the competitiveness of US manufacturing: what does the evidence tell us?. *Journal of Economic literature*, 33(1), pp.132-163.

Jin, W. and McElheran, K., 2018. Economies before Scale: Learning, Survival and Performance of Young Plants in the Age of Cloud Computing. SSRN: https://papers. ssrn. com/sol3/papers. cfm.

Kandel, E. and Lazear, E.P., 1992. Peer pressure and partnerships. *Journal of political Economy*, 100(4), pp.801-817.

Khan, A., Le, H., Do, K., Tran, T., Ghose, A., Dam, H. and Sindhgatta, R., 2018. Memory-augmented neural networks for predictive process analytics. arXiv preprint arXiv:1802.00938.

Khurana, U., Samulowitz, H. and Turaga, D., 2018, April. Feature engineering for predictive modeling using reinforcement learning. In Thirty-Second AAAI Conference on Artificial Intelligence.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S., 2018. Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), pp.237-293.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z., 2015. Prediction policy problems. *American Economic Review*, *105*(5), pp.491-95.

Kueng, L., Yang, M.J. and Hong, B., 2014. *Sources of firm life-cycle dynamics: differentiating size vs. age effects* (No. w20621). National Bureau of Economic Research.

Lanoie, P., Laurent-Lucchetti, J., Johnstone, N. and Ambec, S., 2011. Environmental policy, innovation and performance: new insights on the Porter hypothesis. *Journal of Economics & Management Strategy*, 20(3), pp.803-842.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S. and Kruschwitz, N., 2011. Big data, analytics and the path from insights to value. *MIT sloan management review*, 52(2), pp.21-32.

Levin, S.G., Levin, S.L. and Meisel, J.B., 1987. A dynamic analysis of the adoption of a new technology: the case of optical scanners. *The Review of Economics and Statistics*, pp.12-17.

McAfee Andrew and Erik Brynjolfsson (October 2012). Big Data: The Management Revolution. *Harvard Business Review*. Retrieved from https://hbr.org/2012/10/big-data-the-management-revolution.

Milgrom, P. and Roberts, J., 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. Econometrica: *Journal of the Econometric Society*, pp.1255-1277.

Milgrom, P. and Roberts, J., 1995. Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of accounting and economics*, 19(2-3), pp.179-208.

Moretti, E., 2004. Workers' education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review*, 94(3), pp.656-690.

Müller, O., Fay, M. and vom Brocke, J., 2018. The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of Management Information Systems*, 35(2), pp.488-509.

Novak, S. and Stern, S., 2009. Complementarity among vertical integration decisions: Evidence from automobile product development. *Management Science*, 55(2), pp.311-332.

Porter, M.E. and Van der Linde, C., 1995. Toward a new conception of the environment-competitiveness relationship. *Journal of economic perspectives*, 9(4), pp.97-118.

Raghupathi, W. and Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1), p.3.

Raiffa, H., 1968. Decision analysis: Introductory lectures on choices under uncertainty.

Raj, M and Seamans R (2018) Artificial intelligence, labor, productivity, and the need for firm-level data. *The economics of artificial intelligence: An agenda.* University of Chicago Press.

Raj, M and Seamans R (2019) Primer on artificial intelligence and robotics. *Journal of Organization Design*. 8(1): 1-14.

Ransbotham, S, Kiron D and Prentice PK (2015) *Minding the analytics gap.* MIT Sloan Management Review.

Ransbotham, S, Kiron D, Gerbert P and Reeves M (2017) Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review*. 59(1):

Rivkin, J.W., 2000. Imitation of complex strategies. *Management science*, 46(6), pp.824-844.

Sandvik, J.J., Saouma, R.E., Seegert, N.T. and Stanton, C.T., 2020. Workplace knowledge flows. *The Quarterly Journal of Economics*.

Saunders, A. and Tambe, P., 2015. Data Assets and Industry Competition: Evidence from 10-K Filings. Available at SSRN 2537089.

Saunders, A and Brynjolfsson E., 2016. Valuing it-related intangible assets. *MIS Quarterly*. 40(1): 83-110

Schlegel, G.L., 2014. Utilizing big data and predictive analytics to manage supply chain risk. *The Journal of Business Forecasting*, *33*(4), p.11.

Schrage Michael (September 2014). Learn from Your Analytics Failures. *Harvard Business Review*. Retrieved from https://hbr.org/2014/09/learn-from-your-analytics-failures

Schrage, M. and Kiron, D., 2018. Leading with next-generation key performance indicators. *MIT Sloan Management Review*, 16.

Siegel, E., 2013. Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons.

Stiroh, K.J., 2002. Information technology and the US productivity revival: what do the industry data say?. *American Economic Review*, 92(5), pp.1559-1576.

Tambe, P., 2014. Big data investment, skills, and firm value. *Management Science*, 60(6), pp.1452-1469.

Tambe, P. and Hitt, L.M., 2012. The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research*, 23(3-part-1), pp.599-617.

Tambe, P., Hitt, L.M. and Brynjolfsson, E., 2012. The extroverted firm: How external information practices affect innovation and productivity. *Management Science*, 58(5), pp.843-859.

Waller, M.A. and Fawcett, S.E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), pp.77-84.

Wang, G., Gunasekaran, A., Ngai, E.W. and Papadopoulos, T., 2016. Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of*

Production Economics, *176*, pp.98-110.

Wang, Y., Kung, L., Gupta, S., & Ozdemir, S. 2019. Leveraging Big Data Analytics to Improve Quality of Care in Healthcare Organizations : A Configurational Perspective. *British Journal Of Management*, *30*, 362–388. https://doi.org/10.1111/1467-8551.12332.

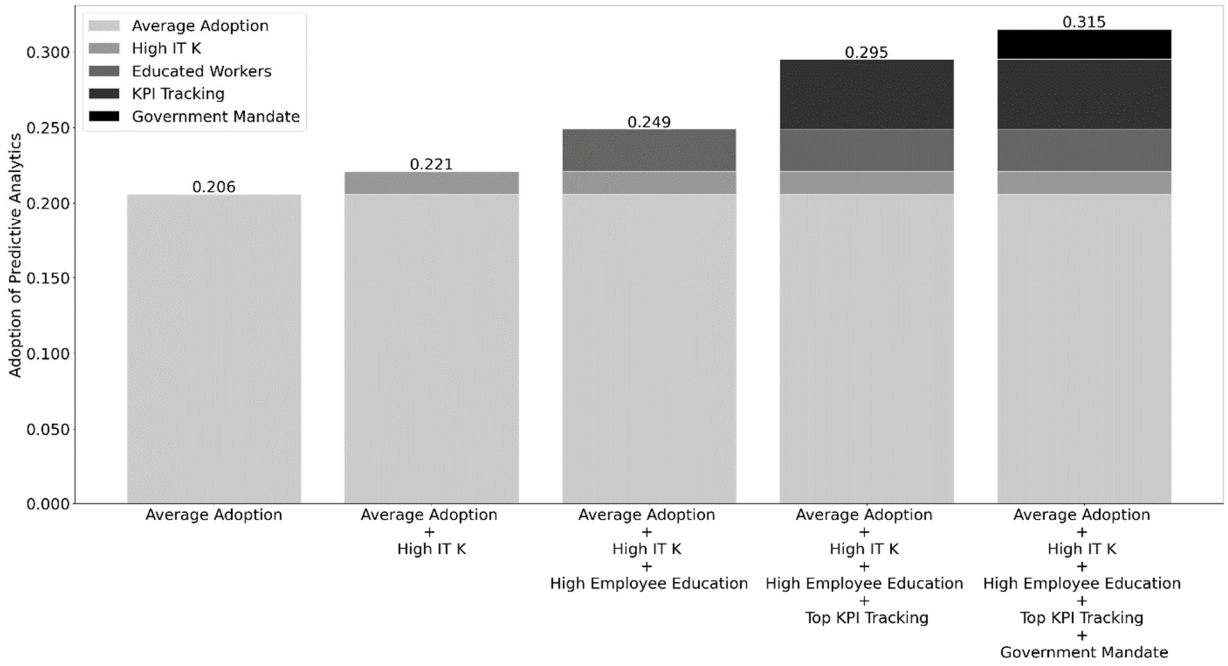
Winnig, L.W., 2016. GE's big bet on data and analytics. *MIT Sloan Management Review, 57*, 5-5.

Wu, L., Lou, B. and Hitt, L., 2019. Data Analytics Supports Decentralized Innovation. *Management Science*, 65(10), pp.4863-4877.

Wu, L, Hitt L and Lou B (2020) Data analytics, innovation, and firm productivity. *Management Science*. 66(5): 2017-2039.

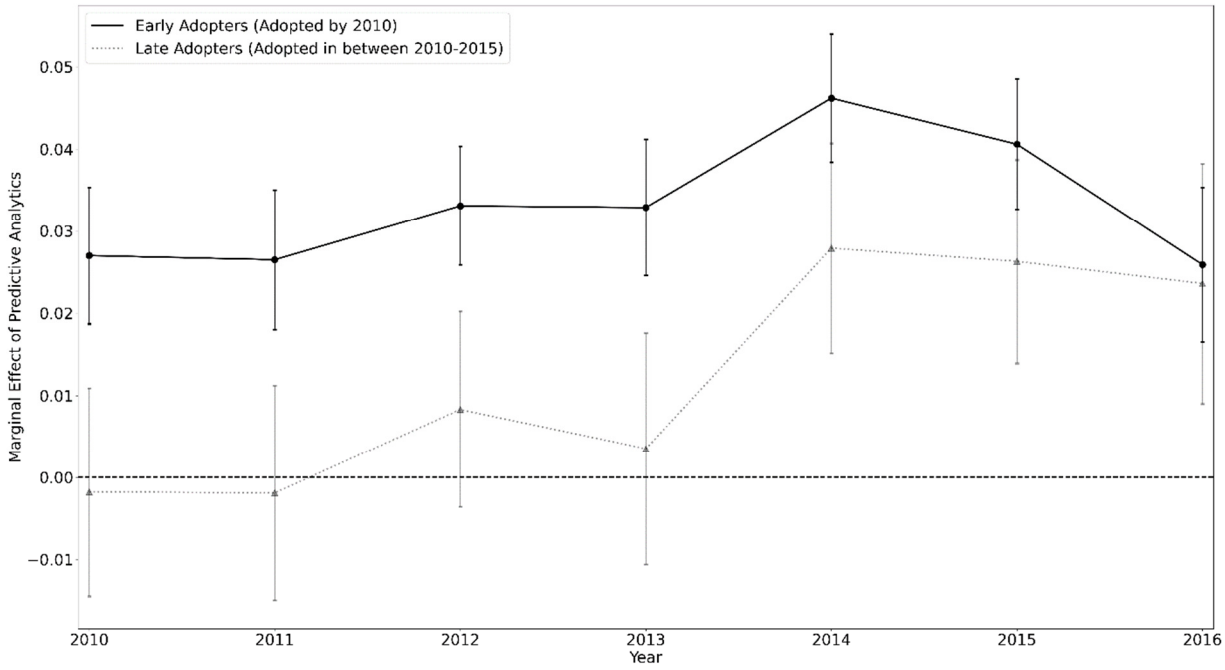
Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D.N., Buffington, C., Goldschlag, N., Foster, L. and Dinlersoz, E., 2021. *Advanced Technologies Adoption and Use by US Firms: Evidence from the Annual Business Survey* (No. w28290). National Bureau of Economic Research.

Figure 1. Conditional Correlations between Predictive Analytics and Potential Complements



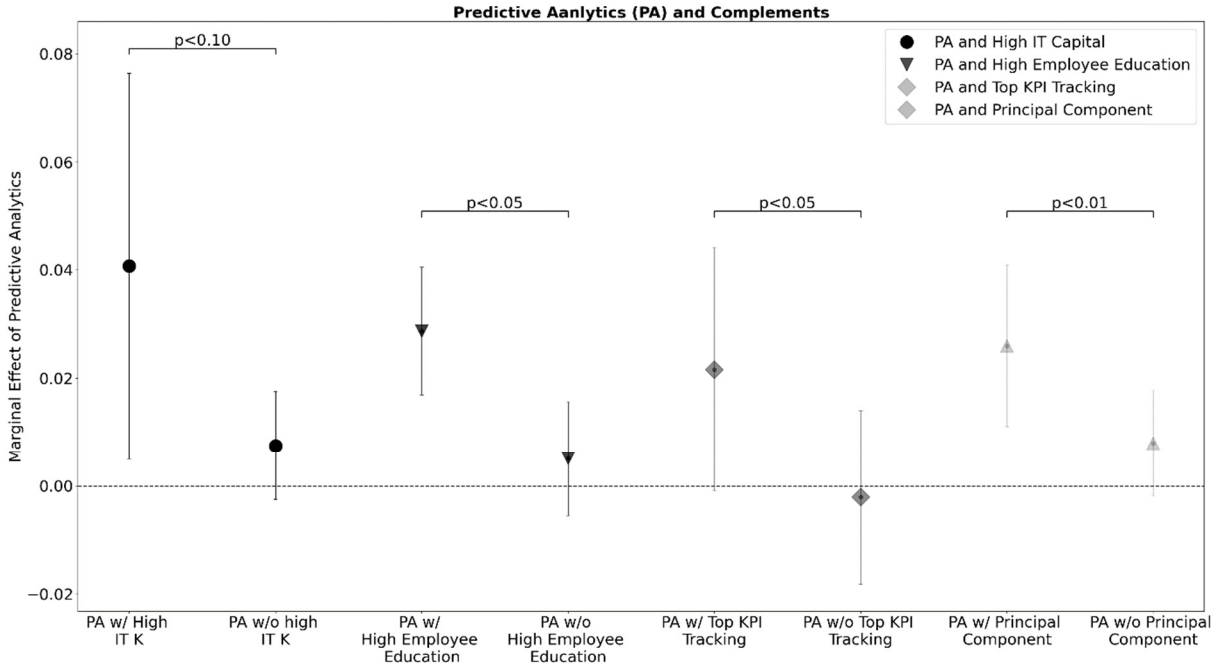
Notes: Correlates of predictive analytics use based on linear probability estimation in the baseline pooled sample. The graph depicts estimated additive marginal contributions based on a single model that includes High IT Capital Stock, High Employee Education, and High KPI Tracking, as well as an indicator for government mandated data collection. Average adoption represents the constant term as the benchmark adoption rate. High IT K is an indicator for plants in the top quintile of IT capital stock. High Employee Education is an indicator for plants with top quartile of the percentage of employees with a bachelors' degree. Top KPI Tracking is an indicator for plants that track top numbers of KPI based on question 2 in the 2015 MOPS survey (i.e. more than 10 KPIs). Finally, Government Mandate is an indicator that data collection is mandated by government regulation or agencies. All four coefficients are significant at the 1% level or higher (see Table A2 column 2 in the Appendix). Additional controls include industry (6-digit NAICS) and year, as well as plant-level employment in log terms, logged non-IT capital stock, structured management practices (index), having top data-driven decision making practices (DDD), age, plant type, multi-unit status, and headquarters status. Robust standard errors are clustered at the firm level. Findings are robust to binary (e.g., probit) estimation models (see Appendix Table A2 Column 3).

Figure 2. Effects of Predictive Analytics by Early and Late Adoption Over Time



Notes: Estimates based on a pooled OLS model with a specification similar to the model in column 3 Table 2. For this test, we construct a yearly panel using the ASM and CMF, where we have annual data on production inputs – including IT – as well as age and multi-unit status (we extrapolate the slow-moving managerial and organizational characteristics from the MOPS) and sales from 2010 to 2016. We identify the groups of establishments that had adopted predictive analytics by 2010, establishments that adopted predictive analytics between 2010 and 2015, and the remaining non-adopters ("laggards") using the 2015 MOPS data. These indicators are then interacted with year dummies to explore the differences in sales over time (using laggards as the baseline group). Histogram bars (and values on the Y-axis) represent the marginal effect of predictive analytics adoption between 2010 to 2016. Standard errors of the coefficients are plotted on the histogram bars. As in Table 2, additional controls include plant-level employment in log terms, logged non-IT capital stock, logged costs of materials and energy, structured management practices (index), an indicator for plants that adopt frontier DDD practices (i.e. identical to the one in Brynjolfsson and McElheran 2019), age, plant type, multi-unit status, and headquarters status.

Figure 3. Effects of Predictive Analytics with and without Potential Complements on Performance



Notes: Estimates based on a pooled OLS model with a specification similar to the model in column 3 Table 2 using the baseline sample. Marginal effects for the indicator of predictive analytics use and the joint marginal effects of this indicator and each potential complement are reported on the Y-axis. These categories are described on the X-axis. For these performance tests of complementarity, the indicator of predictive analytics use is interacted with each potential complement separately in three regressions, one each for high IT capital stock, high percentage of employees with bachelors' degrees, and intensive KPI tracking. High IT K is an indicator for plants with top quintile of IT capital stock. High Employee Education is an indicator for plants with top quartile of the percentage of employees with a bachelors' degree. Top KPI tracking is an indicator for plants that track top numbers of KPI based on question 2 in the 2015 MOPS survey (i.e. more than 10 KPIs). For the last two categories on the X-axis, we interact the indicator of predictive analytics use with the principal component that extract from the three aforementioned complements. All results are reported in Table A5. Error bars indicate 95% confidence intervals around the marginal effects. Statistical significance of the differential productivity of predictive analytics with and without each complement is captured by the interaction term in each model and reported using the p-value.

Table 1. Summary Statistics (Key Variables)

Variable	Definition	Mean (S.D.)	2010 Recall	2015
Log Sales	Logged total value of plant shipments ($Thousands)	10.37 (1.52)	10.68 (1.39)	10.86 (1.37)
Log L	Logged number of plant employees	4.56 (1.17)	4.79 (1.09)	4.88 (1.09)
PA Use	Indicator for plants that use predictive analytics at any frequency	0.74 (0.44)	0.73 (0.44)	0.80 (0.40)
PA Use Frequency	An index for frequency of predictive analytics use based on the highest reported value (e.g. Yearly =1, Monthly=2, Weekly=3, and/or Daily=4)	1.12 (1.06)	1.09 (1.05)	1.27 (1.12)
Log IT K	IT capital stock in log ($Thousands)	5.16 (2.41)	5.58 (2.25)	5.62 (2.18)
Log Non-IT K	Accumulated and depreciated capital investment in non-IT equipment and structures in log terms ($Thousands)	9.26 (1.47)	9.38 (1.58)	9.36 (1.61)
KPI Tracking	Indicator for monitoring 10 or more key performance indicators (KPIs) (question 2 in the MOPS 2015)	0.44 (0.50)	0.37 (0.48)	0.56 (0.50)
Employee Education	Percentage of employees (managers and non-managers) with a bachelor's degree	0.15 (0.14)	0.15 (0.13)	0.16 (0.14)
Government Mandate	Indicator that government regulations or agencies chose, at least in part, what type of data is collected at the plant	0.25 (0.43)	N/A	N/A
Z_MGMT	Normalized index for structured management practices using section A of the MOPS (excluding data-related questions 2 and 6)	0.63 (0.17)	0.60 (0.16)	0.68 (0.15)
DDD	Indicator for plants with high Data-Driven Decision-making management practices in the MOPS 2015 (i.e. questions 6 – target setting, 24 – data availability, and 25 – data usage); It equals to 1 for plants that set both short and long-term targets, having top categories for both data availability and data usage for decision-making	0.47 (0.50)	0.35 (0/48)	0.59 (0.49)
Number of Observations		~51,000 (Baseline)	~18,000 (Balanced)	

Notes: Unweighted statistics based on the baseline and balanced samples from MOPS 2015 data; Standard deviations in parentheses.

Table 2. Productivity Estimates

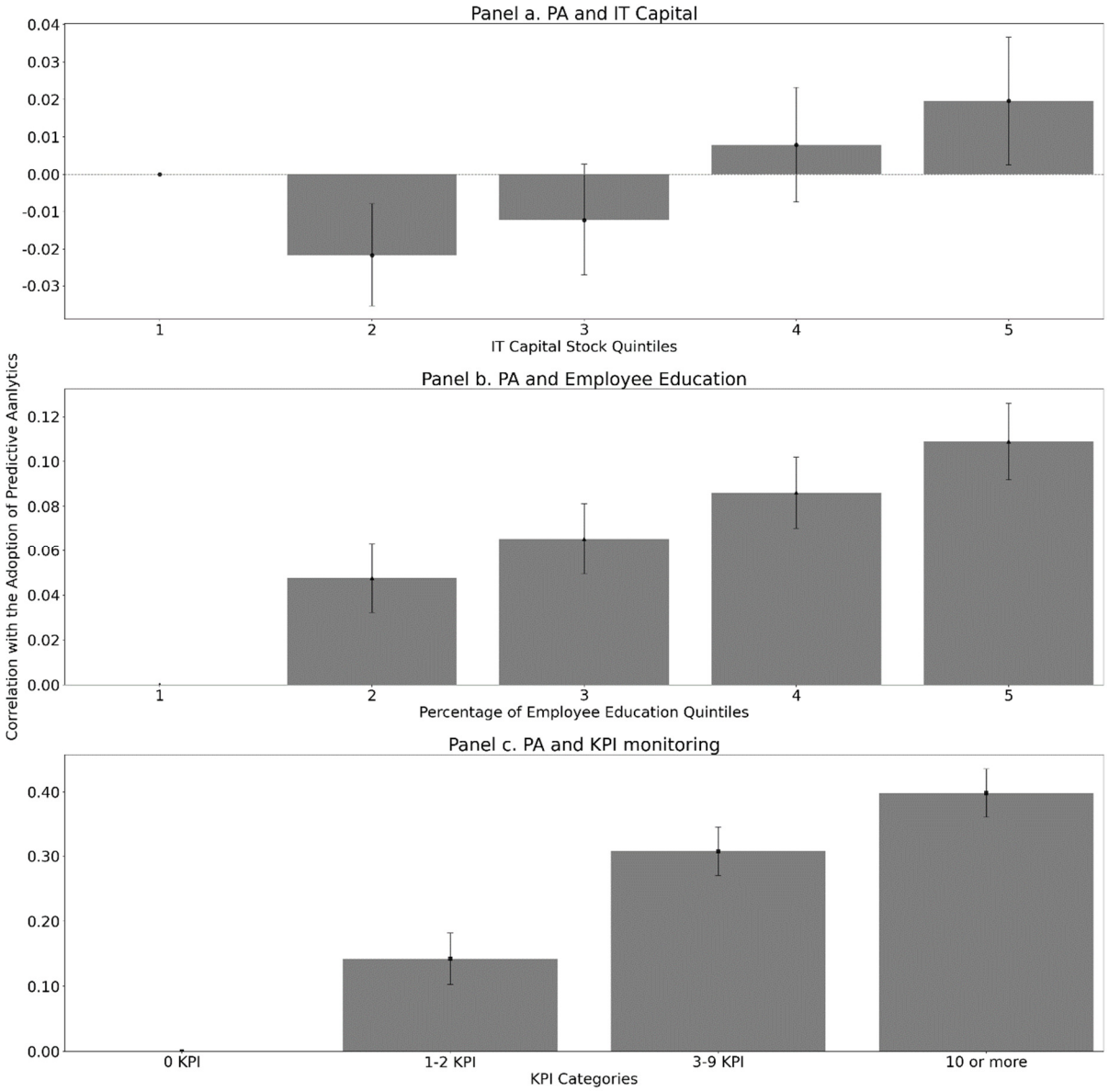
Models	(1) OLS Baseline	(2) OLS Add Complements	(3) OLS Add DDD and Mgmt	(4) OLS PA Frequency	(5) IV 2SLS
Dependent Variables			**Log Sales**		
PA Use	0.0269***	0.0191***	0.0128***		
	(0.0048)	(0.0048)	(0.0049)		
PA Use Frequency				0.0064***	0.0340**
				(0.0021)	(0.0160)
High IT K		0.1362***	0.1357***	0.1361***	0.1336***
		(0.0089)	(0.0089)	(0.0089)	(0.0088)
High Employee Education		0.0619***	0.0602***	0.0589***	0.0584***
		(0.0052)	(0.0052)	(0.0052)	(0.0052)
Top KPI Tracking		0.0299***	0.0234***	0.0202***	0.0186***
		(0.0044)	(0.0045)	(0.0045)	(0.0051)
Log L	0.3746***	0.3633***	0.3626***	0.3613***	0.3615***
	(0.0067)	(0.0067)	(0.0067)	(0.0067)	(0.0067)
Log non-IT K	0.0520***	0.0521***	0.0519***	0.0517***	0.0514***
	(0.0029)	(0.0028)	(0.0028)	(0.0028)	(0.0028)
MU	0.0265***	0.0237***	0.0223***	0.0206***	0.0190***
	(0.0061)	(0.0060)	(0.0060)	(0.0060)	(0.0062)
Industry × Year F.E.	Y	Y	Y	Y	Y
Adjusted R-Squared	0.9325	0.9335	0.9336	0.9337	0.8817
Number of Observations			~51,000		

Notes: Columns 1-4 report pooled OLS estimates. Column 5 reports results from the 2SLS estimator. All columns control for year-industry (6-digit NAICS) fixed effects using the baseline sample. Dependent variable in all columns is logged sales. Column 1 reports estimates when potential complements to predictive analytics use are not accounted for. Column 2 adds High IT capital stock, High Employee Education, and Top KPI Tracking as potential complements. High IT K is an indicator for plants with top quintile of IT capital stock. High Employee Education is an indicator for plants with top quartile of the percentage of employees with a bachelors' degree. Top KPI Tracking is an indicator for plants that track top numbers of KPI. Column 3 further adds structured management (Z_MGMT) and an indicator for plants that adopted top DDD-related practices (DDD) other than KPI tracking since they correlate with better performance (Bloom et al. 2020 and Brynjolfsson and McElheran 2019) and the adoption of predictive analytics. Column 4 uses an index of the frequency of predictive analytics use. Column 5 reports results using government mandated data collection as an instrument for the predictive analytics use categories. The results from the first stage of the IV estimation

are reported in Table A4 column 1. Unreported controls for all columns include additional production inputs (logged costs of materials and energy), plant age, plant type, and headquarters status. Robust standard errors clustered at the firm level.

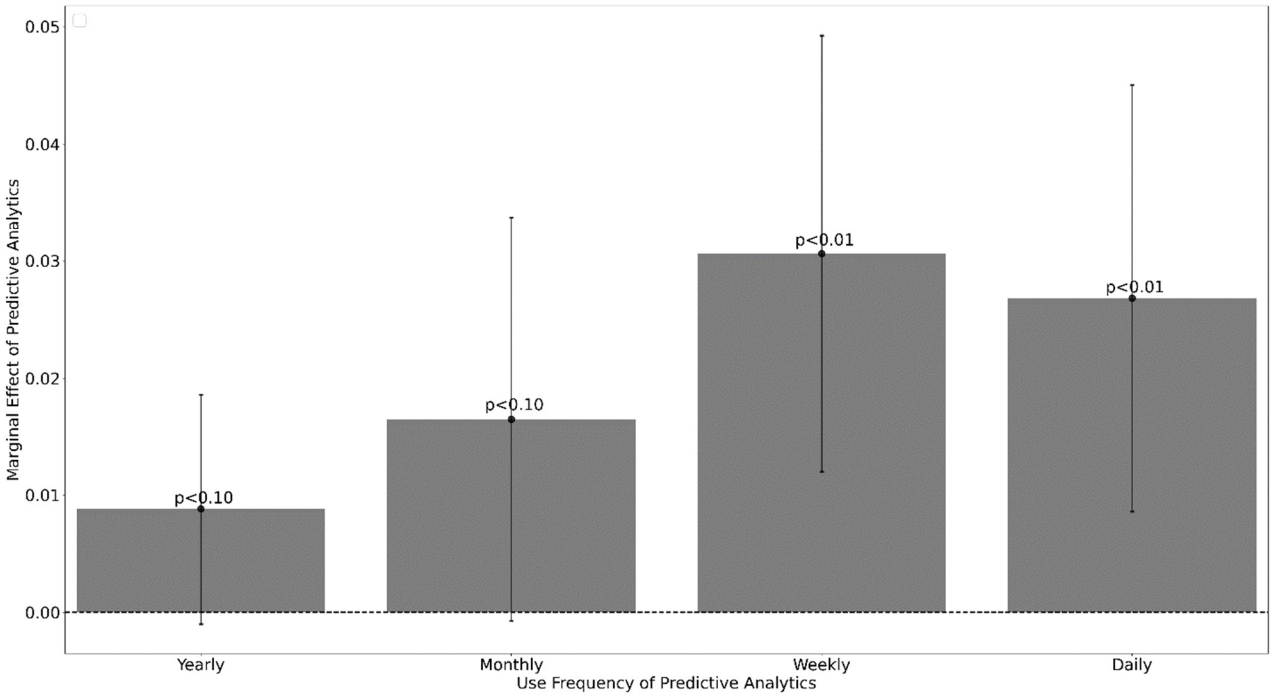
FOR APPENDIX

Figure A1. – correlations by quintile (only 3 complements) – needed to motivate the top-quintile definition of High IT



Notes: Estimates based on the baseline sample from the pooled OLS regressions controlling for plant size, plant age, and industry (6-digit NAICS) and year fixed-effects. The dependent variable is the adoption of predictive analytics use. The base group is the plants in the bottom quintile of the sample based on calculated total IT capital stock for Panel a. For Panel b, the base group is the bottom quintile of the sample based on the percentage of educated workers. For Panel c, the based groups are plants tracking zero KPIs. Histogram bars (and values on the Y-axis) represent the differences in the adoption of predictive analytics between the bottom quintile and higher quintiles (or other categories). The number of quintiles and names of the categories are labeled on the X-axis (the base group has zero value). Quintiles are used for the US Census disclosure avoidance practice and consistency across figures.

Figure A2. Marginal Effect of Predictive Analytics by Use Frequencies



Notes: Estimates based on a pooled OLS model with a specification similar to the model in column 3 Table 2 using the baseline sample with one difference: we include the dummies for each frequency of predictive analytics use in the regression instead of the extensive margin measure for predictive analytics adoption. Marginal effects are reported on the Y-axis. Error bars indicate 95% confidence intervals around the marginal effects. Statistical significance for each plot dot are reported using the p-value.

Table A1. Summary Statistics (Other Variables)

Variable	Definition	Mean (S.D.)	2010 Recall	2015
AGE	Plant age	24.5 (12.9)	24.2 (11.3)	29.2 (11.3)
MU	Indicator for establishments belonging to Multi-unit firms	0.73 (0.45)	0.78 (0.41)	0.81 (0.40)
HQ	Indicator for establishments reported being HQ or co-located with HQ	0.47 (0.50)	0.43 (0.50)	0.41 (0.49)
Production Process Design	Indicator for plants with cellular and continuous-flow production process design	0.35 (0.48)	0.38 (0.48)	0.41 (0.49)
Log Cost of Materials	Cost of material and parts in log ($Thousands)	9.54 (1.79)	9.89 (1.61)	10.1 (1.60)
Log cost of Energy	Cost of energy (fuel and electricity) ($Thousands)	5.87 (1.91)	6.37 (1.65)	6.33 (1.79)
Number of Observations		~51,000 (Baseline)	~18,000 (Balanced)	

Notes: Reported statistics are based on the baseline sample from MOPS 2015 data; Standard deviations in parentheses.

Table A2. Correlation Test for Complementarities in Predictive Analytics Use

Model	(1) Linear Probability Model Continuous Measures	(2) Linear Probability Model High Indicators	(3) Probit Model
Dependent Variable		**PA Use**	
IT K	0.0021*	0.0152**	0.0253***
	(0.0012)	(0.0068)	(0.0084)
Employee Education	0.1098***	0.0275***	0.0256***
	(0.0196)	(0.0057)	(0.0058)
KPI Tracking	0.1782***	0.0462***	0.0467***
	(0.0115)	(0.0049)	(0.0051)
Government Mandate	0.0167***	0.0196***	0.0221***
	(0.0053)	(0.0053)	(0.0058)
Log L	0.0163***	0.0193***	0.0162***
	(0.0033)	(0.0032)	(0.0031)
AGE	-0.0006***	-0.0005***	-0.0003*
	(0.0002)	(0.0002)	(0.0002)
MU	0.0598***	0.0611***	0.0525***
	(0.0070)	(0.0071)	(0.0060)
Industry × Year F.E.	Y	Y	Y
Adjusted R-Squared	0.1544	0.1489	
Number of Observations		~51,000	

Notes: Estimates in columns 2 and 3 are based on linear probability models controlling for industry (6-digit NAICS) and year fixed effects using the baseline sample. The dependent variable is the adoption of predictive analytics. Column 1 uses the continuous measures of potential complements (i.e. IT K is the Log IT Capital Stock that is accumulated and depreciated stock of IT capital expenditure at the plant in log terms; Employee Education is the percentage of employees with a bachelor's degree; KPI Tracking is the KPI tracking categorical variable). Column 2 employs the high indicators of potential complements (i.e. IT K here indicates High IT K - the indicator for plants with top quintile of log IT capital stock within the sample-year; Employee Education is the High Employee Education - the indicator for plants with top quartile of the percentage of employees with a bachelors' degree; KPI Tracking is Top KPI Tracking - the indicator for plants that track 10 or more KPIs). Mandated Data Collection is an indicator for plants that data collection that is mandated by government regulation or agencies. Results using binary estimation models (e.g., probit in Stata 16) are reported in column 3. Unreported controls include logged non-IT capital stock, structure management practice index, and indicators for having top DDD-related practices (DDD) other than KPI tracking, plant type, and headquarters status. Robust standard errors are clustered at the firm level.

Table A3. The Effect of Predictive Analytics on Plant Performance (Robustness)

Model	(1) Labor Productivity	(2) TFP	(3) Translog	(4) OLS PA Frequency (Sum)	(5) IV PA Frequency (Sum)
Dependent Variable	**Log Sales Per Employee**	**Log TFP**	**Log Sales**	**Log Sales**	**Log Sales**
PA Use	0.0123***	0.0089*	0.0167***		
	(0.0047)	(0.0048)	(0.0052)		
PA Use Frequency (Sum)				0.0018***	0.0113**
				(0.0005)	(0.0053)
High IT K	0.1366***	0.1090***	0.1046***	0.1356***	0.1343***
	(0.0086)	(0.0085)	(0.0145)	(0.0089)	(0.0088)
High Employee Education	0.0530***	0.0614***	0.0572***	0.0604***	0.0594***
	(0.0050)	(0.0051)	(0.0053)	(0.0052)	(0.0052)
Top KPI Tracking	0.0252***	0.0164***	0.0202***	0.0229***	0.0159***
	(0.0044)	(0.0045)	(0.0044)	(0.0045)	(0.0058)
Log L	-0.0443***		0.3459***	0.3626***	0.3613***
	(0.0031)		(0.0061)	(0.0067)	(0.0066)
MU	0.0219***	0.0136**	0.0508***	0.0225***	0.0189***
	(0.0058)	(0.0059)	(0.0081)	(0.0060)	(0.0063)
Industry ×Year F.E.	Y	Y	Y	Y	Y
Adjusted R-Squared	0.8231	0.1972	0.9370	0.9336	0.8812
Number of Observations			~51,000		

Notes: Estimates based on the pooled OLS models controlling industry (6-digit NAICS) and year fixed effects using the baseline sample. The dependent variable for column 1 is the logged sales per employee while the dependent variable for column 2 is logged TFP. The dependent variable for all other columns is logged sales. Column 3 estimates a translog production function and reports the calculated marginal effects. Column 4 uses an alternative measure of the frequency of predictive analytics use (i.e. the sum of the multiple choices in question 29 of MOPS 2015 instead of top counted in Table 2 columns 4 and 5) while column 5 employs the 2SLS estimator and uses government mandated data collection as an instrument for the predictive analytics use frequency. Unreported controls for column 1 include production inputs (logged non-IT capital stock, logged costs of materials and energy) per employee, plant age, structure management practice index, and indicators for having top DDD-related practices (DDD) other than KPI tracking, plant type, and headquarters status. Unreported controls for column 2 include plant age, structure management practice index, and indicators for having top DDD-related practices (DDD) other than KPI tracking, plant type, and headquarters status. Unreported controls for columns 3-5 include production inputs (logged non-IT capital stock, logged costs of materials and energy), plant age, structure management practice index, and indicators for having top DDD-related practices (DDD) other than KPI tracking, plant type, and headquarters status. Note that column 3 also includes the interactions between covariates for translog production function estimation. Robust standard errors clustered at the firm level.

Table A4. First Stage of the IV Estimation

Model	(1) For Table 2 Column 5	(2) For Table A3 Column 6
Dependent Variable	**PA Use Frequency**	**PA Use Frequency (Sum)**
Government Mandate	0.3163***	0.9523***
	(0.0166)	(0.0653)
IT K	0.0585***	0.1149
	(0.0209)	(0.0821)
Employee Education	0.0599***	0.0932*
	(0.0138)	(0.0531)
KPI Tracking	0.1451***	0.6749***
	(0.0137)	(0.0568)
Log L	0.0351***	0.1164***
	(0.0104)	(0.0425)
Under-identification Test	280.0	176.0
Weak-identification Test	1001	581.0
Other Controls	Y	Y
Industry × Year F.E.	Y	Y
Number of Observations	~51,000	

Notes: Column 1 reports the results from the first stage of the IV estimation in Table 2 column 5 while column 2 reports the results from the first stage of the IV estimation in Table A3 column 6. The dependent variable for columns 1 and 2 are the categorical measures of the frequency of predictive analytics use using top frequency and using the sum of all reported frequencies (e.g. based on the same value assignment algorithm described in Table 1 for variable PA Use Frequency) respectively. Controls are identical to those in the corresponding columns of IV estimations in Table 2 and Table A3. We reported the coefficients for key variables here. Results for other controls are available upon request.

Table A5. Organizational Complements to Predictive Analytics (Performance Test)

Model	(1) IT Capital Stock	(2) Educated Employee	(3) KPI Tracking	(4) PCA Component	(5) Falsification
Dependent Variable			**Log Sales**		
PA Use	0.0057 (0.0050)	0.0027 (0.0054)	0.0010 (0.0059)	0.0080 (0.0050)	0.0164*** (0.0052)
High IT K	0.1070*** (0.0169)				
PA × High IT K	0.0328* (0.0182)				
High Employee Education		0.0364*** (0.0101)			
PA × High Employee Education		0.0238** (0.0115)			
Top KPI Tracking			0.0021 (0.0087)		
PA × Top KPI Tracking			0.0210** (0.0097)		
Principal Component				0.0396*** (0.0038)	
PA × Principal Component				0.0156*** (0.0043)	
High Structured Management					0.0504*** (0.0108)
PA × High Structured Management					-0.0311*** (0.0116)
Joint Tests	0.0385** (0.0179)	0.0265*** (0.0105)	0.0220*** (0.0080)	0.0260*** (0.0076)	-0.0146 (0.0109)
Industry x Year Fixed Effects	Y	Y	Y	Y	Y
Adjusted R-Squared	0.9337	0.9328	0.9339	0.9336	0.9338
Number of Observations			~51,000		

Notes: Estimates based on pooled OLS models controlling industry (6-digit NAICS) and year fixed effects using the baseline sample. The dependent variable is logged sales. High IT K is an indicator for plants with top quintile of IT capital stock. High Employee Education is an indicator for plants with top quartile of the percentage of employees with a bachelors' degree. Top KPI Tracking is an indicator for plants that track top numbers of KPI. Principal Component is extracted from the PCA using all three potential complements (See Table A6 for details). High Structured Management is an indicator for plants with top quartile structured management practices index. Columns 1-5 interact the indicator of adoption of predictive analytics with each of the potential complements while controlling for all inputs and other potential complements. Joint Tests report the calculated coefficients of the adoption of predictive analytics with the presence of complements, the principal component, and high structured management (using Lincom Joint Test in Stata 16). Unreported controls for all columns include logged total number of employees, log non-IT capital stock, logged cost of material and energy, plant age, indicators for having top DDD-related practices (DDD) other than KPI tracking, plant type, multi-unit status, and headquarters status. Robust standard errors clustered at the firm level.

Table A6. Principal Components Analysis of Organizational Complements

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp 1	1.340	0.4652	0.4466	0.4466
Comp 2	0.8745	0.0877	0.2915	0.7381
Comp 3	0.7857	0	0.2619	1

Principal Components (eigenvectors)	Comp 1	Comp 2	Comp 3
Log IT K	0.6131	-0.2600	-0.7460
Employee Education	0.5923	-0.4736	0.6518
KPI Tracking Categories	0.5228	0.8415	0.1364
Number of Observations		~51,000	

Notes: Reported results based on principal component analysis of log IT capital stock, percentage of workers with a bachelor's degree, and the categories of KPI monitoring intensity (question 2 in the MOPS) using the baseline sample.

Table A7. Organizational Complements (Continuous Measures) to Predictive Analytics

Model	(1) IT Capital Stock	(2) Educated Employee	(3) KPI Tracking	(4) Falsification Mgmt.
Dependent Variable	**Log Sales**			
PA Use	0.0073 (0.0122)	-0.0046 (0.0067)	-0.0273** (0.0139)	0.0084 (0.0172)
PA × Log IT K	0.0001 (0.0022)			
PA × Employee Education		0.0869** (0.0351)		
PA × KPI Tracking			0.0490** (0.0179)	
PA × Structured Management				-0.0012 (0.0275)
Other controls	Y	Y	Y	Y
Industry x Year Fixed Effects	Y	Y	Y	Y
Adjusted R-Squared	0.9337	0.9328	0.9339	0.9339
Number of Observations	~51,000			

Notes: Estimates based on pooled OLS models controlling industry (6-digit NAICS) and year fixed effects using the baseline sample. The dependent variable is logged sales. Log IT K is an indicator for plants with top quintile of IT capital stock. Educated Workers is the percentage of employees with a bachelors' degree. KPI Tracking is a categorical variable based on question 2 in the 2015 MOPS survey. Structured Management is an index calculated based on the first 16 questions of the MOPS (excluding question 2 for KPI tracking and question 6 for target setting) following Bloom et.al. (2020). Columns 1-4 interact the indicator of adoption of predictive analytics with each of the potential complements and the structured management index while controlling for all inputs and other potential complements. Unreported controls for all columns include logged total number of employees, log non-IT capital stock, logged cost of material and energy, plant age, indicators for having top DDD-related practices (DDD) other than KPI tracking, plant type, multi-unit status, and headquarters status. Robust standard errors clustered at the firm level.