Overreliance on Data in Forecasting*

October 16, 2023

Lindsey Gallo[a], Eva Labro[b], James D. Omartian[a,*]

[a]*Ross School of Business, University of Michigan, 701 Tappan Avenue, Ann Arbor, MI 48109, USA*
[b]*Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Campus Box 3490, McColl Building, Chapel Hill, NC 27599, USA*

Abstract

This paper examines the reliance on data in internal forecasting. Using US Census microdata on plant-level sales growth expectations, we find that plants with higher data intensity make forecasts that are both overly precise and less predictive of actual sales. These effects are strongest for plants that have recently increased their data intensity, suggesting there is a learning curve when it comes to efficiently using data for forecasting. Additionally, high data intensity plants issue forecasts that are less idiosyncratic and more like other plants within the firm as opposed to geographic peers, consistent with overreliance on readily available data crowding out incentives for managers to gather relevant local information. Finally, although high data intensity plants suffer from worse forecasting outcomes, they appear to respond more nimbly to unexpected sales growth patterns.

Keywords: Forecasting, uncertainty, data systems, data intensity

*Any views expressed are those of the authors and not those of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 1532. (CBDRB-FY23-P1532-R10716). We thank Salman Arif (discussant), Venky Nagar, Uday Rajan, Kellogg Accounting Conference participants, and seminar participants at Michigan State University, New York University and University of Michigan for very helpful comments.

*Corresponding Author. Stephen M. Ross School of Business, University of Michigan, 710 Tappan Ave., Ann Arbor, MI 48109, USA. Phone: (734) 763-0214.

Email addresses: gallol@umich.edu (Lindsey Gallo), eva_labro@unc.edu (Eva Labro), omartian@umich.edu (James D. Omartian)

1. Introduction

Internal forecasts are a key input for decision making around production planning and resource allocation within firms. Practitioners state, however, that forecasting difficulty increased substantially during the pandemic and they expect it to continue to become increasingly hard even post-pandemic (AICPA and CIMA, 2022). Despite the importance of these internal forecasts and the difficulties practitioners face in formulating them, little is known about how managers form their expectations. It is likely that managers draw from sources inside and outside of the firm and combine this information with their own intuition when forecasting future performance. In recent years, the increased ability to store, access, and process large amounts of data from throughout an organization has led to the proliferation of data-driven decision making (Brynjolfsson and McElheran, 2016b). Practitioner outlets advocate for building data-driven organizational cultures where any new ideas are backed up with solid data, data analysis is performed throughout the entire organization, and top management leads by example in anchoring decisions in data. The same outlets provide advice on how to overcome obstacles in building such data-driven organizational culture (e.g., Waller, 2020; Subrahmanyan and Jalona, 2020; Bean, 2022; AlOwaish and Redman, 2023). Still, it is not clear how this data is used for internal forecasting. On one hand, data availability should reduce the cost to acquire useful information and analytical capabilities can help managers integrate this data more readily into their forecasts. On the other hand, data availability may push metrics to the manager that are ultimately not useful for forecasting and managers may overly rely on readily available data at the expense of local knowledge and experience. In this study we examine the role of data intensity in internal forecasting behavior.[1]

[1] We define data intensity in an organization as the availability and use of numeric information in the plant, as well as the weight placed on such numeric information in the management of the plant such as in communication norms

Most of what we know about managerial forecasting behavior comes from studies examining *external* forecasts designed for capital market consumption. Despite limited research looking directly at them[2], there is little doubt that internal firm forecasts are critically important. The effectiveness of the planning and budgeting process relies heavily on these internal forecasts, and inaccurate forecasts are costly. Overestimating sales growth can lead to poor investment decisions, inventory obsolescence, and holding costs, whereas underestimating sales growth can result in lost revenue opportunities and disappointed customers. Scenario planning that does not reflect the actual underlying uncertainty can similarly lead to inefficient investment (Ben-David et al., 2013). Prior literature finds that better-managed firms make more accurate forecasts, suggesting that forecasting is a core ability of good managers (Bloom et al., 2021).

Given the importance of forecasting, it is perhaps unsurprising that firms would incorporate large amounts of data into the process. Improvements in information systems ease the transmission of data across the organization, reducing asymmetries across plants as well as between plants and headquarters. Increased use of data facilitates centralized decision making by reducing the need for plant managers to interpret local information (Labro et al., 2023). Data-driven modeling can also potentially overcome cognitive biases present in forecasts. However, managers tend to over-extrapolate from prior performance when forecasting (Barrero, 2022) which can be exacerbated by the use of data that typically includes a large amount of historical information. This overreliance on readily available data can come at the expense of useful local insights and experience. Thus, most firms rely on human judgement either alone or in conjunction with statistical modeling to

around decision-making.

[2]Notable exceptions are Brüggen et al. (2021), Forker et al. (2022), Altig et al. (2022), Barrero (2022), Bloom et al. (2021), and Forker (2023).

forecast demand (i.e., Fildes et al., 2006). In short, whether data intensity helps or hinders the accuracy of internal forecasts is an open question.

To address this question, we employ data collected by the US Census Bureau through the Management Organization and Practices Survey (MOPS) which queries roughly 35,000 manufacturing plants. For the first time, the 2015 MOPS survey collected information about managers' expectations of future sales. Specifically, respondents provided a distribution of 2017 plant-level sales expectations in the form of five growth scenarios. For each growth scenario (lowest, low, medium, high, and highest), the respondent provides both a point estimate of sales and a percent likelihood (i.e., probability), where the likelihoods must total 100%. The same survey also asked plant managers about the availability and use of data at the establishment level, from which we derive a measure of data intensity. We exploit this survey data to provide insights into how data intensity is associated with internal forecasting behavior.

Using MOPS data has significant advantages over traditional measures of managerial forecasts. First, we can observe a distribution of forecasted outcomes for each plant rather than simply a point or range. This allows us to examine how the realization maps into the distribution of expected outcomes. Second, the data allow us to observe multiple plants within the same firm, permitting comparisons of internal forecasting behavior while holding firm-level attributes constant. Third, and perhaps most important, the forecasts we observe should be relatively free of bias. Whereas external forecasts are voluntary in nature, participation in the Census survey is mandatory, considerably mitigating the selection bias that results from incentives in choosing to disclose. Further, Census has strict confidentiality requirements and individual responses are not visible to market participants, competitors, or even others within the same firm, removing incentives to shade expectations or misreport. These features of the data permit much more confidence in our inferences.

After validating the forecasts and our data intensity construct, we examine the role of data intensity in shaping internal forecasts. We observe that managers of data-intensive plants are more precise in their expectations of sales growth; the distribution is narrower and the probability of the middle scenario is greater for these plants although the overall expected value is similar. The difference between high and low data intensity plants is driven largely by the low end of the distribution—less data-intensive plants are more pessimistic; the bottom two scenarios forecast lower sales growth and assign a higher probability to these outcomes. Low data intensity plants also forecast higher sales growth for the top two scenarios, although the probabilities are not different from high data intensity plants. The results are similar incorporating industry and geography fixed effects.

We next explore the appropriateness of this additional precision by testing where realized sales fall in the forecasted distribution. Consistent with Barrero (2022), we find that realized sales are much more dispersed than managers expect; 43.5% of plants have actual sales growth that falls outside the forecasted scenarios. Further, we find that relative to forecasted sales, the dispersion in actual sales is *higher* for plants with high data intensity. A one-standard-deviation increase in data intensity is associated with a 5% reduction in the probability of realized sales growth appearing in the middle three forecast quintiles, with most of this probability mass shifted outside the distribution. Taken together, the results suggest that managers of high data intensity plants are overly precise in their forecasts as a result of overreliance on readily available data, making them less attuned to the possibility of extreme outcomes or black swan scenarios.

While managers of high data intensity plants may be overly precise in their forecasts, it may be that better information improves the mapping of expected into actual sales growth. Thus, we next examine whether forecasts better predict actual growth in the presence of high data intensity.

We find that this is not the case; despite increased precision of their forecasts, high data intensity plants produce forecasts that are less predictive. For a one standard deviation increase in our data intensity measure, the mapping of the expected value of the forecast distribution into actual growth declines by 22%. While data intensity is associated with less predictive forecasting of the expected value, we next investigate if the tails of the distribution from high data intensity plants have better predictive power for actual growth. To test this, we focus on the 10^{th}, 50^{th}, and 90^{th} percentile points in the distribution rather than the mean and standard deviation. We find that variation in the low-end of the distribution is most predictive of realized growth, but the predictive power does not vary with data intensity. Instead, greater data intensity significantly reduces the predictive power of variation in the right tail of the distribution. The predictive power of the median is not affected by data intensity. Overall, our results point to a negative relationship between data intensity and the predictive power of forecasts that affects both the expected value and tail scenarios.

If data intensity has a negative effect on managers' ability to forecast future sales, we expect the deleterious effects would be most pronounced when managers have less experience with the data and their limitations. Exploiting variation in recent increases in data intensity, we test for and find that plants that recently increased data intensity have forecasts that are less predictive of actual growth. This finding suggests managers do seem to learn over time how to better incorporate data into their forecasts, and helps bolster our inferences about the relation between data intensity and forecasting.

Forker et al. (2022) finds that a "learning-by-doing" effect exists in that forecasts become more accurate with experience. We next benchmark the influence of data intensity on internal forecasting behavior relative to manager education and tenure. We find that like data intensity, both education and tenure are associated with greater forecast precision. However, we observe divergent patterns

in their relation with forecasts' predictive power. Specifically, while data intensity is associated with weaker mapping of forecasts into actual sales growth, education is associated with improved mapping; manager tenure has no moderating relationship. Taken together these results indicate that more educated managers produce better forecasts while experienced managers are able to be more precise in their forecasts without sacrificing quality. These results also provide assurance that data intensity is not simply a proxy for manager sophistication.

Although our results are consistent with overreliance on data leading to overly precise internal forecasts with diminished predictive power, it is not clear if this pattern ultimately harms firm productivity. On the one hand, if data intensity increases certainty about future sales or the lack thereof, it will likely result in inefficient over- or under-investment. On the other hand, strong data intensity may make managers aware of deviations from expectations more quickly, allowing them to intervene sooner. Consistent with the latter, in a model interacting data intensity with realized sales scenarios, we find that while total factor productivity (TFP) is negatively impacted by unexpectedly weak sales, stronger data intensity significantly attenuates this productivity loss. Interestingly, the main effect of data intensity on TFP is insignificant, suggesting that the primary benefit of data intensity with respect to productivity is in adjusting to unforeseen shocks.

In our final set of analyses, we explore the mechanism by which forecasts differ in high data intensity plants. It is possible that data intensity can cause managers in multi-plant firms to focus on firm- or economy-wide forces rather than local, plant-specific factors. For example, prior literature finds that the use of predictive analytics is associated with more centralized decision making (Labro et al., 2023). Alternatively, data intensity might reflect more robust local information collection, allowing managers to be more confident forecasting idiosyncratic plant performance. Survey evidence suggests that managers often focus on local factors when forecasting economic

outcomes (Andrade et al., 2022). We investigate these alternatives by comparing the similarity of forecast characteristics between plants. We employ the first-order Wasserstein distance to measure the difference between probabilities. We find that greater data intensity is associated with less idiosyncratic forecasts and more similarity in forecasts between plants within the same firm. However, companies may employ more robust data systems because of a need to coordinate across plants, which may drive these results. To rule out this scenario, we control for data intensity at the plant's corporate headquarters and find that while it too is positively associated with same-firm forecast similarity, plant-level data intensity remains a significant predictor of similarity between same-firm peers. Focusing on different-firm industry and geographic peers, we find no significant relation between data intensity and forecast similarity with industry peers, but a negative relation to forecast similarity with geographic peers. Taken together, these results point to data systems focusing more on company-level data and less on the local economy, which may help explain why high-data intensity plants have worse forecast accuracy.

Our study makes several contributions. First, we add to the literature examining the consequences for firms of data availability and use by providing evidence on how data intensity impacts forecasting behavior. Despite the recent explosion of big data within firms, little is known about its influence on forecast characteristics in practice. In practice, forecasting often relies on both data and managerial judgment (Fildes and Petropoulos, 2015; Petropoulos et al., 2022). A rapidly growing experimental literature (e.g., Commerford et al., 2022; Dietvorst and Bharti, 2020; Chen et al., 2022) studies the interaction between man and machine in coming up with estimations. Casas-Arce et al. (2022) call for research on the over- versus under-reliance on (data-driven) decision algorithms. Our findings indicate that data intensity is associated with forecasts that are less predictive of future sales growth, and that these forecasts are overly precise. This pattern arises

because managers overly rely on (readily available) data. Our results also point to a potential benefit of data intensity by demonstrating that these plants are able to adequately adjust production despite their relative forecast inaccuracy.

We also contribute to a nascent literature that uses surveys to examine individual forecasting behavior. Barrero (2022) finds that managers are, on average, overly precise when forecasting sales growth and over-extrapolate from recent performance. Our findings point to an overreliance on data as a potential driver of overprecision in forecasting. Recent studies find that managers may focus on local signals when forecasting aggregate outcomes (Andrade et al., 2022; Coibion and Gorodnichenko, 2015; Cavallo et al., 2017; Dovern et al., 2023). Our results suggest that data intensity can potentially attenuate the reliance on local information. However, results also indicate that managers may overly rely on firm-level data at the expense of relevant local information when forecasting plant-level outcomes.

Finally, we add to the accounting literature on internal forecasts. Prior literature highlights the relationship between internal and external forecasts (Hemmer and Labro, 2008; Kroos et al., 2018; Ittner and Michels, 2017), and external forecasts are sometimes used to gauge the quality of firms' internal information systems (Gallemore and Labro, 2015). Despite the importance of internal forecasting for firm decision-making, little is known about how managers develop these forecasts or what drives variation in their accuracy. The management accounting literature has recently started to tackle this important question using field study data (e.g., Brüggen et al., 2021; Forker et al., 2022). Our study is able to speak directly to the characteristics of internal forecasts on a large scale. In particular, we examine both the first and second moment of sales growth forecasts which allow us to make observations both about forecast accuracy and precision.

Our study is subject to important caveats. First, data-intensity develops endogenously and

while we use firm, industry, and geography fixed effects to control for drivers of data intensity, we cannot definitively claim a causal relationship. Second, the forecasts we observe may not be identical to the forecasts being used by managers and communicated to headquarters, particularly because the response must conform to the Census Bureau's format. However, even if the reported forecasts are not in the same format as what is used by the organization, they allow us to observe managers' underlying growth expectations, which serve as the foundation of the forecast regardless of the final format. Holding constant the format of the forecast allows for easier comparison of forecast characteristics across plants and firms. Finally, our sample is limited to manufacturing firms. While we have no reason to believe our results are specific to manufacturers, it is possible that the results may not generalize to other industries.

2. Prior Literature and Theoretical Framework

Internal forecasts are a key input for production planning, and many firms spend several weeks if not months each year to come up with accurate sales forecasts during the budgeting process. Accurate sales forecasts are pivotal to orderly planning of input acquisition, scheduling , processing, inventory replenishment and financing (Cassar and Gibson, 2008). Overestimating sales can lead to inefficient investment and costly inventory obsolescence whereas underestimating sales can lead to missed profit opportunities and disappointed customers. Internal forecasts aid headquarters in determining resource allocation across the organization. Managers rely on expectations for budgeting and these forecasts can be the basis for performance measurement.[3] Despite the im-

[3]Though not the focus of this study, these various uses of internal forecasts can introduce biases into their measurement. Our research design helps us to abstract away from these biases. The U.S. Census's MOPS survey is kept confidential and hence can be reasonably expected to be free of bias that may enter into forecasts submitted to headquarters. Additionally, to the extent that biases are firm-specific as opposed to plant-specific, our firm fixed effects can control for these effects.

portance of internal forecasts both for individual managers and the organization as a whole, few studies are able to directly examine the characteristics of internal forecasts of firm performance. Notable exceptions are Brüggen et al. (2021) and Forker et al. (2022) who examine the impact of disaggregating demand forecasts within a multi-national agricultural chemical manufacturer, Altig et al. (2022) and Barrero (2022) who utilize the Survey of Business Uncertainty, and Bloom et al. (2021) who link uncertainty measured from the Management and Organizational Practices Survey (the same data used in this study) to investment. Most of the accounting and finance research on forecasting focuses on *external* forecasts intended for *external* stakeholders. Prior literature does link the quality of internal and external information (Hemmer and Labro, 2008). Cassar and Gibson (2008) and Ittner and Michels (2017) demonstrate that better internal information leads managers to make more accurate external forecasts. However, internal information can be difficult to measure and prior large sample literature often has had to rely on publicly available proxies (e.g., Gallemore and Labro, 2015).

The exercise of forecasting frequently involves the combination of data and human judgement (i.e., Fildes and Petropoulos, 2015; Petropoulos et al., 2022). For example, the process may involve first building a model and selecting the relevant inputs, then interpreting the results and possibly adjusting them based on qualitative input from the manager. Historically, the collection, processing, assimilation, and analysis of data could be prohibitively costly, but in recent years the monetary cost of computing has rapidly declined, and with it, firms have ramped up their use of data in decision making (Brynjolfsson and McElheran, 2016b). Automated data entry and consolidation possibilities may help processing data for forecasting purposes (Eichholz et al., 2023). This relatively "low cost" data is in contrast to data that must be actively acquired by a plant manager—for example collecting information about the local economy or inquiring with customers about future

orders. Managers who are faced with the choice of whether to collect new data or rely on readily available data will balance the costs of acquisition against the potential benefits.

Given the perceived benefits of data availability and usage, we may expect that greater data intensity would result in better forecasts by plant managers. By reducing the cost of acquiring and integrating data, managers may be able to optimize the information used in forecasting future demand. Better data may also help managers move away from simple heuristics (i.e., Harvey, 2020) that limit their ability to make high quality forecasts. Data may also reduce information asymmetry in the firms, allowing plant managers to incorporate data from other plants in their forecast models. Further, managers can learn from and adjust models over time, even if the models themselves are complicated (De Baets and Harvey, 2020). Increased computing capacity provides the opportunity to use sophisticated forecasting techniques that may ultimately require little adjustment from managers. Importantly, if data is perceived to be not useful, it can always be ignored.

While it may seem intuitive that more information is always better, there are reasons why data intensity could harm forecast quality. The models themselves can be complicated, which may make it more difficult for the manager to understand the types of qualitative adjustments that are appropriate. For example, manipulating big data sets can require advanced machine learning techniques that are almost certainly outside the expertise of a plant manager (i.e. Varian, 2014).[4] It may be the case that simple models best describe future growth, but managers are motivated to use all available data resulting in an overfitted model that describes the past but does a poor job predicting the future (Brighton and Gigerenzer, 2015). If the most important part of the data relates to managers' own firms, they cannot apply techniques to correct for overfitting, such as

[4]For an overview of the many theoretical and empirical approaches to forecasting see Petropoulos et al. (2022).

using holdout samples. Consistent with a data overload problem, Eichholz et al. (2023) find in a survey of German companies during the COVID-19 pandemic that lower data volume positively correlated with a higher satisfaction with the forecasting process during this crisis. Managers may falsely equate the amount of data and computation used with how good the model is at predicting future growth, such as shown in Zacharakis and Shepherd (2001) where venture capitalists become overconfident when they have more information to make decisions. Managers may also feel less responsible for missed forecasts if they rely on data, particularly if the importance of data analytics has been communicated by headquarters, leading to reduced incentives to make adjustments to machine-generated forecasts.

These competing arguments around whether data intensity can be expected to help or hinder the forecasting process underscore the importance of examining the impact of data intensity on managerial forecast quality. In particular, companies continue to invest in expanding the availability and use of data. Ultimately, whether and how data intensity impacts forecasting is an empirical question.

3. Data

Lack of data has been a key impediment to researching how managers form expectations of the future. While a large literature analyzes management-issued guidance, it typically relies on the content of public forecast disclosures. Because the market responds to these public forecasts, managers make strategic considerations when deciding whether to issue a forecast, how optimistic it should be, and how precise to make it (e.g., point or range forecasts). As a result, unpacking managers' underlying expectations from disclosed forecasts is problematic.

To overcome this obstacle, we exploit new data about managers' expectations from the US

Census Bureau.[5] Census conducts an Annual Survey of Manufactures (ASM) collecting detailed plant-level operational data (e.g., sales, cost of materials, employees) from a large, representative sample of US manufacturing plants.[6] These data underpin macro indicators like the Federal Reserve's Index of Industrial Production and the Bureau of Economic Analysis' calculation of Gross Domestic Product (GDP). Plants are required by law to respond, and Census follows up to ensure a high response rate of 70-80%. Census also validates responses against other administrative data (e.g., tax returns). The confidentiality of individual responses, even from other government branches, helps minimize selection biases in the sample.[7] In 2015 Census sent the Management and Organizational Practices Survey (MOPS) as a supplement to the ASM, asking respondents detailed questions about data, sales forecasts, targets, incentives, and decision rights.[8] We rely on this survey for our forecasting data and our measures of plant-level data intensity.

3.1. Managerial Forecasts

The 2015 MOPS, sent out in May 2016, asks about sales growth expectations for 2017. The form directs managers to predict the total value of goods shipped for five growth scenarios—lowest, low, medium, high, and highest—and assign a probability to each.[9] The five probabilities must sum to 100%. This approach allows managers to express their expectations and certainty

[5]Bloom et al. (2019) describe that the respondents are typically plant managers, financial controllers, CEOs, CFOs or general managers.

[6]In years ending in 2 or 7, the census conducts a full Census of Manufacturers (CMF) and sends forms to all but the smallest establishments. Based on responses to the full census, the ASM is sent to a stratified random sample of roughly 50,000 plants every year over a five-year window, ensuring adequate coverage across industry, geography, and establishment size. For additional details on the ASM methodology, see `https://www.census.gov/programs-surveys/asm/technical-documentation/methodology.html`

[7]Researchers can apply to use the confidential microdata for approved projects. All analysis is performed in a secure facility on Census servers. Census reviews all output to ensure that no individual responses can be inferred from the results.

[8]For a detailed description of the design choices behind the MOPS, see Buffington et al. (2017). The actual survey form can be found at `https://www2.census.gov/programs-surveys/mops/technical-documentation/questionnaires/ma-10002_15_final_3-2-16.pdf`.

[9]Question 31 asks for the five sales scenarios and their associated probabilities for 2017.

without the constraints of a standard parametric distribution. Using the ASM data from 2017, we can compare the forecasted value to actual realized sales. For admission to our main sample, we require valid responses to the sales forecast questions along with actual sales from the 2015 and 2017 ASM. Our resulting sample includes 24,500 plants across 15,000 firms. Because the unit of observation is the plant and there are many multi-plant firms, we can conduct within-firm analyses.

A distinct advantage of our data is that they should be largely free from reporting bias. While managers have strong strategic incentives to shade their publicly disclosed forecasts, the forecasts in our data are confidential. Further, the surveys are sent directly to the plants and responses are not shared by Census with other plants in the firm or corporate headquarters. Thus, there should be no reason for managers to shade their forecasts for internal-firm audiences. Even if managers typically strategically shade their internal forecasts, the uniqueness and specificity of the MOPS response format makes it unlikely that managers simply re-purpose a preexisting biased internal forecast.[10] Thus, the forecasts we observe should be an accurate depiction of managers' growth expectations.

[Fig. 1 about here.]

Figure 1 summarizes the distributions of forecasted sales growth and associated probabilities for each of the five scenarios. In general, managers are optimistic about the future—the median "medium growth" scenario is positive, and the higher growth scenarios depict greater growth than the estimated contraction in lower growth scenarios. There is considerable variation in the amount of growth, however, with some plants forecasting contractions in the rosiest of scenarios and oth-

[10]One downside of the forecasts likely being in a different format is that managers may not be used to producing forecasts in this format. While this unfamiliarity may make the responses a noisy measure of management's expectations, it is unlikely to systematically bias the responses.

ers predicting growth in even the most pessimistic scenarios. On the probability side, managers typically assign the highest likelihood to the middle scenarios (the median plant assigns this scenario a 50% likelihood), and allocate a higher likelihood to the higher-growth scenarios than the lower-growth ones.

[Table 1 about here.]

In Table 1, Panel A, we validate these forecasts by mapping them into actual sales growth, using a sample that excludes singleton plants—plants that would uniquely identify a geographic (CBSA), industry (6-digit NAICS) and/or firm fixed effect dummy variable.[11] Using this sample of 12,500 plant observations, in column (1) we regress actual sales growth directly on the forecasted growth scenarios.[12] We standardize all variables to mean zero and unit variance to be able to compare coefficients and determine relative strength of each variable. Overall, we find that these forecasts are weakly predictive of actual sales growth. The R^2 value suggests that a linear model using these forecasted values as inputs explains 3.7% of the variation in actual sales growth. Variation in the medium sales growth scenario has the greatest influence on predicting actual sales growth, consistent with managers having a general sense of where overall sales will likely end up. On the low side of the distribution, we find that variation in the sales growth of the lowest scenario is significantly predictive of actual sales. However, there is no significant relation for probabilities on the low end of the distribution to actual sales. On the high side of the distribution, we observe competing forces: placing greater likelihood on high- and highest-growth scenarios is predictive of

[11] In all tables where we report coefficients on independent variables based on regressions with firm fixed effects, we use this sample that excludes singletons so as to not identify plants for Census disclosure purposes. In analyses where we do not include firm fixed effects, we use the larger sample that also includes singletons.

[12] We exclude the probability of the middle growth scenario as it would be perfectly collinear with the other probabilities given the adding-up constraint.

stronger realized growth, but variation in the magnitude of the highest growth scenario is actually *negatively* correlated with sales growth.

To better understand the competing forces at the high end of the distribution, in column (2) we summarize the forecast distribution by percentiles.[13] While the coefficient on the median is positive, it falls short of significance at traditional levels. Instead, variation in the 10th percentile is the strongest predictor of growth. The competing forces at the top end of the distribution from column (1) appear to largely cancel each other out, making the 70th and 90th percentiles of the distribution not predictive of actual growth. In column (3) we add geography, industry, and firm fixed effects and find qualitatively similar results to column (2).

In Panel B we explore the extent to which forecasts predict idiosyncratic growth beyond industry, local geography, and firm-specific factors. For parsimony, we distill the forecast distribution to the expected value and standard deviation, and regress actual growth on these variables and the interaction of the two. Column (1) presents results with no fixed effects, whereas in columns (2)-(4) we employ geography (CBSA), industry (6-digit NAICS), and firm fixed effects. Column (5) adds all three sets of fixed effects.[14] Consistent with the prior panel and with Bloom et al. (2021), the positive and significant coefficient on $ForecastSalesGrowthEV_p$ indicates variation in forecasts predict variation in actual sales growth.[15] While uncertainty is not significantly related

[13]For example, if a firm lists growth scenarios of -6%, -3%, 0%, 3%, and 6%, with associated likelihoods of 5%, 20%, 60%, 10%, and 5% respectively, the 10th percentile would be -3%, the 30th would be 0%, the 50th would be 0%, the 70th would be 0%, and the 90th would be 3%.

[14]To meaningfully compare R^2 values across specifications, we exclude singleton plants that uniquely identify a fixed effect dummy. This results in a sample of 12,500 observations, which for consistency we maintain across both panels of this table and in the other tables with similar model structures.

[15]As with our other specifications, we standardize all variables to zero mean an unit variance before estimating the model. Standardizing variables has an advantage in allowing us to compare relative coefficient sizes for a standard deviation change in the variable. However, the downside is it precludes us from interpreting deviation from 1 in the coefficient of $ForecastSalesGrowthEV_p$ as evidence of average forecast bias. Because, as indicated in results in subsequent tables, the across-plant variation in forecasts is smaller than the variation in actual realized sales, the standardization scaling factor for $ForecastSalesGrowthEV_p$ is smaller than for $ActualSalesGrowthEV_p$. The net result

to sales growth, consistent with Bloom et al. (2021), we do see it linked with a dampening of the expected value's predictive power, as indicated by the negative interaction coefficient. This interaction suggests that managers on average are aware of at least some relative uncertainty. Comparing R^2 values across specifications, we see that 6% of the variation in actual growth can be attributed to local economic factors, 7% to industry factors, and 31% to firm-specific factors. However, even after including all these fixed effects in column (5), we still find strong predictive power in the forecast, indicating that these forecasts contain a significant plant-specific component.

3.2. Plant-level Data Intensity Measurement

Our research question centers on the role that data plays in forming managerial expectations of the future. As a result, we construct our primary independent variable of interest, $DataIntensity_p$, to capture how important data is to managerial decision-making at plant p. Specifically, we combine responses from MOPS question 24: "What best describes the availability of data to support decision making at this establishment?" and 25: "What best describes the use of data to support decision making at this establishment?" For each of these questions, respondents check one of five boxes on similar scales, ranging from "Data to support decision making are not available" to "All the data we need to support decision making is available" for question 24 and from "Decision making does not use data" to "Decision making relies entirely on data" for question 25. We record the response to each question on a uniform 0-1 scale (with 1 being highest data intensity), sum the responses by plant, and standardize the sum to mean 0 and unit variance. Combining responses to both questions allows us to capture more variation in the underlying construct and acknowledges that both the availability and use of data are key components of the overall importance of data in

of this scaling is to depress the coefficient magnitude (but not change its statistical significance).

the management of the plant. While we cannot know exactly how respondents interpreted these survey questions, we believe a common interpretation would be that the answers to these questions reflect the availability and use of numeric information in the plant, as well as the weight placed on such numeric information in the management of the plant such as in communication norms around decision-making.

Table 2, Panel A, reports descriptive statistics for the two component questions, both for the 2015 level and for the respondent's 2015 recall of 2010. In general, data appears to be an important part of managerial decision-making, with the average plant reporting a value just shy of "a great deal of data to support decision making is available." Use of data is slightly less but still quite strong; the average plant reports somewhere between a moderate to heavy reliance on data for decision-making. The large standard deviation for each question suggests considerable heterogeneity in data intensity across plants. Further, data availability and use have both increased considerably since 2010. Over that five-year window, the average plant's data availability and use scores each increased by nearly 20%.

[Table 2 about here.]

To validate our measure, in Panel B we look to see if stronger data intensity is related to more information technology investments at the plant. Specifically, we correlate $DataIntensity_p$ with the logged stock of IT capital investments at the plant—both in levels (2015) and in changes (from 2010 to 2015). We calculate IT capital investments using ASM and CMF responses from the plant since 2002 using the approach in Brynjolfsson and McElheran (2016a). We find strong support for a link between IT investment and data intensity, both with and without firm fixed effects.

Our ability to observe a valid and precise measure of data intensity is a strength of our study. However, we acknowledge that we observe an equilibrium outcome. Data availability and use develops endogenously, and as a result, we cannot take relations between data intensity and the characteristics of plant-level expectations as being definitive proof of causality. Factors correlated with data intensity may drive forecasting patterns. We address this challenge by exploiting a battery of fixed effects, controlling for industry-, geography-, and firm-specific variation. Relatedly, reverse causality could be at play: forecasting difficulty may drive the choice to rely more or less intensively on data. While exploiting an exogenous shock to data intensity could help pinpoint causality more definitively, such a shock is not readily available and it would likely reduce the generalizability of our findings (Glaeser and Guay, 2017). Instead, we present a mosaic of relations that are consistent with a causal story. Our study provides a needed first step in understanding how data impacts managers' forecasting behavior and our findings can motivate other studies to expand upon our analysis.

4. Results

4.1. Data Intensity and Forecast Characteristics

We begin our analysis of the role of data in forecasting by correlating differences in data intensity across plants with differences in their sales forecasts. We document these relations in two ways: graphically by splitting the sample at the median into high and low data-intensity plants, and in regressions that exploit the full continuous variation of our data intensity measure. Figure 2 plots the average growth and probability of each scenario, along with error bars showing 99% confidence intervals, comparing between high and low data intensity plants. Two themes emerge. First, managers of data-intensive plants are more precise in their expectations. The spread between

the lowest and highest growth scenarios is smaller, and the probability of the middle scenario is higher. Second, there is asymmetry in this precision. For plants with low data intensity, the top two scenarios forecast higher growth, but the probability assigned to these scenarios is no different based on data intensity. In contrast, for the bottom two scenarios, low data-intensity plants exhibit more pessimism—both in forecasting lower scenario growth and assigning a higher likelihood of occurrence to the two bottom scenarios.

[Fig. 2 about here.]

In Table 3 we confirm these results by individually testing parameters of the forecasted distribution using regressions. Each cell reports the coefficient and standard error of β from a regression in the form:

$$\text{Dependent Variable}_p = \beta \mathit{DataIntensity}_p \, [+\lambda_i + \mu_g] \, \{+\nu_f\} + \varepsilon_p. \tag{1}$$

Column (1) reports coefficients from an estimation without fixed effects. We begin with the magnitude of growth in each of the five scenarios as dependent variables and find data intensity is correlated with moderation of extremes; with more data intensity the lowest and low scenarios predict stronger sales whereas the high and highest scenarios predict weaker sales. The moderation is not symmetric, however, with higher growth scenarios exhibiting more moderation than lower growth scenarios. We more formally test for a change in asymmetry using $\mathit{AsymmetrySalesGrowth}_p$ as a dependent variable—defined as the distance between medium and highest sales growth scenarios minus the distance between lowest and medium scenarios—and find a significant leftward shift as data intensity increases.

[Table 3 about here.]

For the probability components of the forecasts as dependent variables, we observe that data intensity is correlated with a re-allocation from low and lowest growth scenarios to the middle growth scenario probabilities, representing a significant asymmetrical rightward shift in probability mass, as observed in the significantly positive coefficient on $AsymmetryProbability_p$. Together, the differences in magnitudes and probability weights result in greater precision of sales forecasts associated with higher levels of data intensity. The standard deviation of the probability distribution ($ForecastUncertainty_p$) is significantly lower with greater data intensity, and the forecast range ($RangeSalesGrowth_p$), defined as the highest growth magnitude minus the lowest, is tighter. The net result of the two asymmetric shifts is that the expected value of the forecast distribution ($ForecastSalesGrowthEV_p$) becomes slightly more conservative with increased data intensity. Interestingly, this conservatism appears to be unwarranted; we observe greater data intensity correlated with stronger realized sales growth ($ActualSalesGrowth_p$). Column (2) presents the signs and significance of these effects, adding industry and geography fixed effects. Qualitatively the results are similar across these columns, suggesting differences are not driven by industry- or geography-specific economic factors. However, with the introduction of firm fixed effects (column (3)) we find attenuation of the differences in forecasting associated with variation in data intensity.

The aforementioned results show that plants with higher data intensity have more precise forecasts. In Table 4 we test the appropriateness of this precision by exploring where actual sales fall relative to the forecasted distribution. For each plant, we take the five forecasted scenarios and associated probabilities and form a probability distribution function by interpolation. Precise details of this interpolation process can be found in Appendix B. We then calculate seven indicator variables per plant based on which quintile of the forecast distribution actual growth falls: below the distribution, quintiles 1-5, or above the distribution. We use these indicators as dependent vari-

ables and regress them on $DataIntensity_p$. The intercept of each regression is an estimate for the portion of plants whose actual sales growth fall in a specific quintile of the forecast distribution (or above/below the distribution) given average data intensity in the sample, and the coefficient on $DataIntensity_p$ represents the additional probability of a plant falling in the specified part of the distribution associated with a one standard deviation increase in data intensity.

[Table 4 about here.]

We find that actual sales growth is much more disperse than managers expect. 43.5% of plants have sales realizations that fall completely outside the forecast distribution, with slightly more falling below than above. This indicates that a large proportion of actual sales realizations are black swan scenarios—growth realizations to which management assigns no positive probability and hence come completely as a surprise. The bulk of plants appearing outside the distribution can be attributed to an under-representation of plants in the middle three quintiles.[16] This excess dispersion in actual sales is consistent with prior work (Barrero, 2022). Turning to the coefficients on $DataIntensity_p$, we find that plants with higher data intensity have higher dispersion in realized growth relative to their forecast distribution. We observe significant reductions in the probability of appearing in the middle three forecast quintiles—on the order of a 5% increase in the unconditional probabilities for a one standard deviation increase in data intensity—and this probability mass is almost entirely shifted outside the distribution. The probability of appearing above the forecast distribution increases by one percentage point (a 5% increase in the unconditional probability) for a one standard deviation increase in data intensity, and the result is highly statistically significant. On

[16] If realized sales across plants were IID draws, we would expect, for an appropriate level of uncertainty, no plants falling outside the forecast distribution and roughly 20% of plants falling in each quintile.

the low side, the magnitude of the increase in likelihood of being below the distribution is roughly half that of being above the distribution, but the statistical significance is just shy of conventional levels. We find virtually no change in the likelihood of being within the predicted distribution but in the tails (i.e., 1^{st} or 5^{th}). Collectively, these results suggest that the forecast precision that managers in high data intensive plants exhibit makes them less attuned to the possibility of extreme scenarios.

In practice, plants are subject to systematic shocks that may drive bunching in certain portions of the distribution of actual sales growth. Observing so much mass both above and below the distribution suggests that a single economy-wide shock is not driving our results. Industry and/or geographic shocks correlated across plants may also contribute to systemic bunching at various points in the distribution. Below the line in Table 4 we report the signs and significance of the coefficients on $DataIntensity_p$ including industry and geographic fixed effects. While the statistical significance is weaker, the results follow a consistent pattern. Thus, we conclude that systemic shocks to sales are unlikely to be the main driver of excess dispersion in actual sales relative to forecast in the presence of higher data intensity.

4.2. Data Intensity and the Predictive Power of Forecasts

Having observed what appears to be excess forecast precision associated with greater data intensity, we turn to exploring the link between data intensity and the forecast's ability to predict actual growth. Specifically, we assess the strength of the mapping between forecasted and actual growth as a function of data intensity. In Table 5, columns (1) and (2), we take models from Table 1, Panel B and additionally interact $ForecastSalesGrowthEV_p$ with our plant-level measure of data intensity. Odd columns include no fixed effects whereas even columns include geography,

industry, and firm fixed effects; the results are very similar regardless of their inclusion, so we report the coefficients for the fixed effects models and the signs and significance only for the odd columns. All variables are scaled to mean zero and unit variance to facilitate interpretation of the main effects. We find that with greater data intensity, the mapping of expected into actual growth erodes considerably—the -0.063 coefficient equates to a 22% reduction in the correlation between forecasted and actual growth for a one standard deviation increase in data intensity. In columns (3) and (4) we swap out the covariates for the expected value and forecast uncertainty for the 10^{th}, 50^{th}, and 90^{th} percentile points in the forecast distribution. Consistent with results in Table 1, Panel A, variation in the low end of the distribution is most predictive of actual growth, but the predictive power of this left tail is not significantly different at different levels of data intensity. However, we find that data intensity is associated with variation in the right tail that is *negatively* related to actual growth. Collectively, the results in this table suggest that managers of data-intense plants have a harder time forecasting sales growth, despite their forecasts indicating more precision. With a stronger reliance on data, managers may be lulled into a false sense of security that estimators of historical sales obtained from available data will predict future sales growth. That is, their forecasting models overfit historical data. [17]

[Table 5 about here.]

If data intensity results in overprecise and less predictive forecasts, we expect that, over time, managers will gain greater familiarity with the limitations and pitfalls of the data. As a result, we

[17] An alternative explanation for our Table 5 results could be that companies adopt greater data intensity when growth is more unpredictable. Thus, the correlations we observe between data intensity and inferior prediction of actual growth are driven by selection. However, such an explanation would be inconsistent with our findings that data intensity is correlated with greater forecast precision. Furthermore, we also find that actual dispersion in sales is not greater for high data intensity plants than it is for low data intensity plants. Thus, while selection may partially contribute to the associations we observe, it is highly unlikely to be the driving force.

predict that forecasts from plants which have recently increased data intensity will exhibit greater overreliance on data. While the forecasts we observe are from a single point in time and we cannot track the progression of a plant's forecasting longitudinally, we can exploit time-series variation in data intensity to compare in the cross-section plants that recently expanded data intensity to those that have not. Instead of interacting the forecasted expected value with the 2015 level of data intensity (as in column (2) of Table 5), in Table 6 we interact the expected value of sales growth with data intensity from 2010 and the change in data intensity from 2010 to 2015. Our main coefficient of interest is on the interaction $ForecastSalesGrowthEV_p \times ChangeDataIntensity_{2015,p}$; we expect and find adverse effects to be strongest in plants that recently increased their data intensity. The negative coefficient is consistent with learning about the limitations of data—plants that most recently increased data intensity suffer in terms of the predictive power of their forecasts' expected value. This learning result also helps bolster the validity of our inference that data intensity, rather than a factor correlated with it, influences forecast quality.

[Table 6 about here.]

Having established that high data intensity is linked with overly precise, less-predictive forecasts, we compare this relation to the impact of managerial experience on forecasts. Specifically, we compare the relations between data intensity and forecasting to those of managerial education and respondent tenure and forecasting. Forker et al. (2022) find that learning-by-doing improves forecasting ability. Hence, education and experience may make managers more confident in their predictions and more proficient at forecasting, but may also make them more cognizant of uncertainty. Table 7 presents results. In the first two columns, similar to data intensity, we find that management education and respondent tenure are associated with greater precision in sales fore-

casts. The magnitudes of the relations across the three variables are similar, though the education and tenure coefficients become statistically insignificant with the addition of geography, industry, and firm fixed effects in column (2). However, in columns (3) and (4) when we interact the forecasted expected value with data intensity, management education, and tenure, we observe divergent patterns. While data intensity is associated with weaker mappings between forecasted and actual sales growth, manager education is associated with improved mappings; respondent tenure has no moderating relation. Thus, while data intensity is associated with less predictive forecasts, more highly educated managers appear to produce more predictive forecasts, and more experienced managers produce more precise forecasts without sacrificing predictive power. These divergent results bolster the validity of our prior inferences, showing that data intensity is not simply a proxy for manager sophistication.

[Table 7 about here.]

4.3. Productivity Implications

Collectively, our results suggest that data intensity is correlated with plant forecasts that are more precise yet less predictive of future events. A natural follow-on question is whether this seemingly misguided precision has productivity implications. Unexpected realizations of sales likely have significant productivity implications. Realizations below expectations may result in productivity declines due to excess capacity that cannot be shed quickly. For realizations above expectations the net effect is unclear—strong sales may allow the plant to better utilize available capacity, but may also require capacity increases, which may be more costly when made on short notice (Banker and Hughes, 1994). Data intensity likely influences the link between unexpected sales and productivity, but the direction is unclear. On one hand, if data intensity increases certainty

about future sales, it will likely increase investment, which may be inefficient. On the other hand, strong data intensity may make managers more attuned to deviations from expectations, allowing the manager to intervene earlier. Data intensity may also make operations more nimble, facilitating a more graceful response to unforeseen sales patterns.

Using establishment-level data, the Census Bureau calculates (logged) total factor productivity (TFP) based on plant outputs and inputs in relation to industry averages.[18] We test to see if TFP changes as a function of unexpected realized sales growth, and if data intensity at the plant moderates or amplifies this relation. In Table 8, column (1), we regress the change in $\ln TFP_p$ from 2015 to 2017 on industry and geography fixed effects, and indicators for if realized sales growth is in the extremes or outside the 2017 forecast distribution. Specifically, $AboveHighest_p$ indicates if actual sales are higher than the "highest" forecast scenario, $HighToHighest_p$ indicates if realized sales are in the right tail of the distribution (between the "high" and "highest" scenarios), $LowestToLow_p$ indicates sales in the left tail of the distribution (between the "low" and "lowest" scenarios), and $BelowLowest_p$ indicates sales below the forecast distribution entirely (below the "lowest" forecast scenario). As predicted, the results show that unexpected growth has a significant correlation with changes in logged TFP. We observe significant declines in productivity associated with unexpectedly weak sales. For unexpectedly strong sales it appears that better utilization of capacity dominates any productivity cost associated with adding capacity on short notice—realizations in the right tail and beyond the forecast distribution are associated with productivity increases.

[18] See Foster et al. (2016) for a detailed explanation of the methodology. At a high level, productivity is calculated as the difference between output and inputs scaled by factor elasticities: $\ln TFP_{p,t} = \ln Q_{p,t} - \alpha_K \ln K_{p,t} - \alpha_L \ln L_{p,t} - \alpha_M \ln M_{p,t}$, where $Q_{p,t}$ is plant p's output for time t, $K_{p,t}$ is capital stock, $L_{p,t}$ is labor, and $M_{p,t}$ is materials. α_K, α_L, and α_M are industry-level cost shares. Our sample is slightly smaller for this analysis because the TFP calculation is not available for all plants in our main sample.

[Table 8 about here.]

In column (2) we test if data intensity is associated with productivity changes. We find evidence of a weak positive correlation, consistent with the notion that on average, data helps managers improve operations. In column (3) we interact the data intensity variable with the indicators from column (1). We find that data intensity is linked with positive changes in productivity in the tails of the distribution. For sales realizations below the forecast distribution, we find a one standard deviation increase in data intensity is associated with a 21% reduction in the productivity loss. For realizations above the forecast distribution, the economic magnitude of the productivity boost is similar, though the statistical significance falls short of conventional levels. Interestingly, the main effect of $DataIntensity_p$ is insignificantly negative, suggesting that the main benefit of high data intensity on productivity is in dealing with unforeseen circumstances. Given we only have forecasts from a single point in time, we cannot disentangle if the productivity improvements we see from data-intense plants in unforeseen scenarios obtain from data alerting managers to changing environments sooner or facilitating smoother re-tooling of operations for different demand patterns. We leave this question to further research.

4.4. Forecast Similarity Mechanism Tests

Our results are consistent with managers over-relying on available data, resulting in overprecise expectations of future growth. To bolster this explanation, we attempt to unpack some of the key channels through which data influences forecasting. The data available and used at plants may highlight plant-specific information, resulting in over-reliance and over-precision when plant managers form idiosyncratic expectations. Alternatively, data may focus on and highlight economy-, industry- and firm-wide economic forces, distracting managers from generating and impounding

plant-specific information. As a result, with higher data intensity, forecasts may more closely resemble certain peers. To explore this tension, we compare the similarity (or lack thereof) of forecasts between plants.

Because each plant provides a distribution, there are numerous dimensions on which to compare similarity or divergence in forecasts across plants. For dimensional reduction we employ the first-order Wasserstein distance, a well-established, parsimonious approach for measuring the difference between two probability distributions (see Panaretos and Zemel, 2019, for a review). Often called the earth mover's distance, intuitively this measure captures the minimum cost to move piles of dirt in the shape of the first probability distribution into the shape of the second distribution, where cost is measured as mass times distance.[19] The resulting non-negative scalar distance is symmetric with respect to the direction in which we compare the two distributions.

In Table 9, we evaluate how forecast distribution similarity between plants correlates with data intensity. We begin by creating a sample of unordered plant pairs, where both plants belong to the same firm (i.e. $\{p, p'\} \subseteq f$).[20] We then create an equal number of control pairs not matched by firm by shuffling plant p' across the pairs. By shuffling the matched plants, we ensure that the population of plants represented in the sample is consistent in both treated and control pairs. We calculate the Wasserstein distance for each pair ($WDist_{p,p'}$), and regress it on an indicator if p and p' belong to the same firm ($SameFirm_{p,p'}$), the minimum data intensity score of the two plants

[19]For instance, assume plant p has sales growth scenarios of (-0.2, -0.1, 0, 0.1, 0.2) with uniform probabilities (0.2, 0.2, 0.2, 0.2, 0.2), and plant p' has growth scenarios of (-0.3, -0.1, 0, 0.1, 0.2) with probabilities (0.05, 0.35, 0.2, 0.2, 0.2) respectively. The Wasserstein distance would be 0.02: calculated as converting the first distribution to the second by moving 0.05 of mass from the -0.2 scenario a distance of 0.1 to the left plus moving 0.15 of mass from the -0.2 scenario a distance of 0.1 to the right.

[20]Because the Wasserstein distance is symmetric, using ordered pairs would result in duplicate observations. For instance, if plant A is matched with plant B in our sample, we do not include an observation for plant B matching with plant A.

($DataIntensity_{p,p'}$), and the interaction of these two variables. We double cluster by plant p and plant p' to account for the correlation across pairs when one of the members is the same. The results in column (1) show that when two plants are from the same firm, their forecasts are more similar. Compared to the intercept which captures the distance between different-firm pairs with average data intensity, same-firm pairs forecast distributions are 3.8% more similar (a reduction of 0.011 from 0.287). With higher data intensity, plants are significantly more similar to different-firm peers; benchmarking magnitudes, the coefficient of -0.011 on $DataIntensity_{p,p'}$ indicates that the increased similarity associated with increasing data intensity by one standard deviation is equivalent to the the increased similarity from being part of the same firm. This increased forecast similarity between unaffiliated plants suggests that at least part of the data in data-intense plants captures economy-wide information. Further, the interaction of these two variables is significantly negative, indicating that in higher-data intensity environments, plants' forecasts look more similar to their same-firm peers.

[Table 9 about here.]

The link between data intensity and intra-firm forecast similarity may be a result of individual plant managers' heavier reliance on data systems. However, if companies have more robust data systems because of the need for coordination across plants, a similar correlation would likely obtain. To help rule out this possibility being the only operative channel, we isolate company-wide data effects from plant-specific ones by controlling for the intensity of data at corporate headquarters. In column (2) we restrict our sample to include only same-firm pairs, and further exclude pairs where either p or p' are co-located with corporate headquarters. Because our sample is restricted to same-firm pairs, the intercept becomes analogous to the intercept plus the $SameFirm_{p,p'}$ coefficient

in column (1), and the $DataIntensity_{p,p'}$ coefficient becomes analogous to the sum of column (1)'s $DataIntensity_{p,p'}$ coefficient plus the $SameFirm_{p,p'} \times DataIntensity_{p,p'}$ coefficient. We also include in our model $HQDataIntensity_f$, a firm-level measure of the data intensity at corporate headquarters, which allows us to control for corporate-level information systems. We find that corporate headquarters' use of data is associated with significantly greater same-firm forecast similarity between remote plants within the firm, but that even after controlling for corporate headquarters' data intensity, local managers' use of data is still linked with greater forecast similarity across plants.

Some of the overall reduction in forecast idiosyncrasy associated with data intensity may be a result of formal data systems providing managers with general information about local economic conditions, industry trends, or macroeconomic forces. To test for these possibilities, we return to the sample from column (1) but add additional same-industry and same-geography peer pairings, and look to see if data intensity is associated with stronger forecast similarity within these groups. Specifically, we take the plants from column (1) and form all pairs of plants that are from different firms but either in the same industry (6-digit NAICS) or the same geography (CBSA). From each of these sets of possible pairs we select at random a roughly equivalent number of pairs equal to our original column (1) sample. On this sample we estimate the model from column (1) but add indicators if the pair is from the same industry or the same geography, along with interactions with data intensity. In column (3), we find that plants in the same industry have more similar forecasts; the same-industry similarity estimate is actually stronger than plants within the same firm but different industries, though the difference is insignificant. Plants in the same geography appear to be no less similar in their forecasts than plants in different geographies, suggesting that local economic effects play a very small role in formulating plant-level growth expectations. When we interact these indicators with data intensity, we find that higher data intensity is not associated with

greater industry forecast similarity and is correlated with significantly greater dissimilarity from local peer plants. Thus, it appears that with greater data intensity, managers do not incorporate additional industry-specific information into their forecasts and impound less information about the local economy.

5. Conclusion

Using new plant-level data from the US Census Bureau, we unpack the relations between forecasted growth expectations and the intensity of data use and availability in a plant. After validating our measures of data intensity and managerial expectations, we find that in plants with higher data intensity, forecasts of future sales growth exhibit greater precision. This precision comes from a reduction in the extremity of forecasted growth—particularly on the high-growth side of the distribution, and a significant reduction in the probability assigned to lower growth scenarios. The net result of these movements are slightly more conservative expected values of growth for plants with higher data intensity. These results are robust to controlling for industry- and geography-specific economic shocks using a battery of fixed effects. However, by examining the relations between the forecast distributions and actual growth, we find the additional precision associated with higher data intensity is not warranted. Forecast predictive power appears to be negatively associated with data intensity—both in terms of the central tendency of the forecast distribution and the tails. This evidence is consistent with plant management's over-reliance on data in forecasting. We find evidence consistent with learning: the greatest decline in forecast predictive power appears for plants that recently increased data intensity.

We find that plants with higher data intensity experience somewhat higher productivity growth, and that productivity growth is concentrated in scenarios where sales growth is unexpectedly low.

This finding suggests that managers of plants with high data intensity are better able to handle unforeseen sales realizations. They may be able to more quickly identify and respond to growth not going according to plan. Alternatively, high data intensity may make plants more flexible and nimble in responding to unforeseen sales growth scenarios. Overall, while plant managers may over-rely on data in forecasting resulting in too precise and lower-predictive forecasts, as unforeseen sales realizations occur, data seems to help plant managers respond more appropriately.

Lastly, we explore the mechanisms by which data intensity influences forecast behavior by studying the similarity in forecasting across plants. We find plant-level data intensity is associated with greater similarity to other plants' forecasts in general, and specifically between plants of the same firm. These results are consistent with the notion that data systems highlight economy-wide and firm-specific information, crowding out managerial incentives to produce and impound plant-specific information.

Our paper contributes to the forecasting and data-driven decision-making literatures. We find that data is not a panacea in resolving uncertainty, but rather may facilitate developing overfitted models of the future. This finding is important for companies weighing trade-offs between specialized managers with deep tacit knowledge and generalist managers with strong data skills. It also is relevant for companies contemplating investment in sophisticated data systems to support decision-making.

Our findings are not without limitations. Because we do not have exogenous variation in data intensity, we cannot definitively prove a causal relation on forecast behaviors. Other factors related to forecast precision may drive the decision to adopt greater data intensity, producing the associations we observe. However, our collective results produce a mosaic consistent with data strengthening managers' certainty in their forecasts—to an excessive degree.

References

AICPA, CIMA, 2022. AICPA Business and Industry Economic Outlook Survey Detailed survey results: 4Q 2022 Management Accounting & Finance. Tech. rep., Association of International Certified Professional Accountants.

AlOwaish, M. B., Redman, T. C., 2023. What does it actually take to build a data-driven culture? Harvard Business Review May 23.

Altig, D., Barrero, J. M., Bloom, N., Davis, S. J., Meyer, B., Parker, N., 2022. Surveying business uncertainty. Journal of Econometrics 231, 282–303.

Andrade, P., Coibion, O., Gautier, E., Gorodnichenko, Y., 2022. No firm is an island? How industry conditions shape firms' expectations. Journal of Monetary Economics 125, 40–56.

Banker, R. D., Hughes, J. S., 1994. Product Costing and Pricing. The Accounting Review 69, 479–494.

Barrero, J. M., 2022. The micro and macro of managerial beliefs. Journal of Financial Economics 143, 640–667.

Bean, R., 2022. Why becoming a data-driven organization is so hard. Harvard Business Review .

Ben-David, I., Graham, J. R., Harvey, C. R., 2013. Managerial miscalibration. Quarterly Journal of Economics 128, 1547–1584.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R. S., Patnaik, M., Saporta-Eksten, I., Van Reenen, J., 2019. What Drives Differences in Management. American Economic Review 109, 1648–1683.

Bloom, N., Davis, S., Foster, L., Lucking, B., Ohlmacher, S., Saporta-Eksten, I., 2021. Business-level expectations and uncertainty. NBER Working Paper Series .

Brighton, H., Gigerenzer, G., 2015. The bias bias. Journal of Business Research 68, 1772–1784.

Brüggen, A., Grabner, I., Sedatole, K. L., 2021. The folly of forecasting: The effects of a disaggregated demand forecasting system on forecast error, forecast positive bias, and inventory levels. Accounting Review 96, 127–152.

Brynjolfsson, E., McElheran, K., 2016a. Data in Action: Data-Driven Decision Making in U.S. Manufacturing.

Brynjolfsson, E., McElheran, K., 2016b. The Rapid Adoption of Data-Driven Decision-Making. American Economic Review: Papers & Proceedings 106, 133–139.

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

Casas-Arce, P., Cheng, M. M., Grabner, I., Modell, S., 2022. Managerial Accounting for Decision-Making and Planning. Journal of Management Accounting Research 34, 1–7.

Cassar, G., Gibson, B., 2008. Budgets, internal reports, and manager forecast accuracy. Contemporary Accounting Research 25, 707–738.

Cavallo, A., Cruces, G., Perez-Truglia, R., 2017. Inflation Expectations, Learning, and Supermar-

ket Prices: Evidence from Survey Experiments. American Economic Journal: Macroeconomics 9, 1–35.

Chen, C. X., Hudgins, R., Wright, W. F., 2022. The Effect of Advice Valence on the Perceived Credibility of Data Analytics. Journal of Management Accounting Research 34, 97–116.

Coibion, O., Gorodnichenko, Y., 2015. Is the Phillips Curve Alive and Well after All? Inflation Expectations and the Missing Disinflation. American Economic Journal: Macroeconomics 7, 197–232.

Commerford, B. P., Dennis, S. A., Joe, J. R., Ulla, J. W., 2022. Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence. Journal of Accounting Research 60, 171–201.

De Baets, S., Harvey, N., 2020. Using judgment to select and adjust forecasts from statistical models. European Journal of Operational Research 284, 882–895.

Dietvorst, B. J., Bharti, S., 2020. People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. Psychological Science 31, 1302–1314.

Dovern, J., Müller, L. S., Wohlrabe, K., 2023. Local information and firm expectations about aggregates. Journal of Monetary Economics .

Eichholz, J., Knauer, T., Winkelmann, S., 2023. Digital maturity of forecasting and its impact in times of crisis.

Fildes, R., Goodwin, P., Lawrence, M., 2006. The design features of forecasting support systems and their effectiveness. Decision Support Systems 42, 351–361.

Fildes, R., Petropoulos, F., 2015. Improving Forecast Quality in Practice. Foresight: The International Journal of Applied Forecasting 36, 5–12.

Forker, E., 2023. The Informativeness of Dark Data for Future Firm Performance.

Forker, E., Grabner, I., Sedatole, K., 2022. Sooner is better than later: The effect of forecast disaggregation on the year-over-year improvement in demand forecast revisions.

Foster, L., Grim, C., Haltiwanger, J., 2016. Reallocation in the Great Recession: Cleansing or Not? Journal of Labor Economics 34, S293–S331.

Gallemore, J., Labro, E., 2015. The importance of the internal information environment for tax avoidance. Journal of Accounting and Economics 60, 149–167.

Glaeser, S., Guay, W. R., 2017. Identification and generalizability in accounting research: A discussion of Christensen, Floyd, Liu, and Maffett (2017). Journal of Accounting and Economics 64, 305–312.

Harvey, N., 2020. Use of heuristics: Insights from forecasting research. Thinking & Reasoning pp. 5–24.

Hemmer, T., Labro, E., 2008. On the optimal relation between the properties of managerial and financial reporting systems. Journal of Accounting Research 46, 1209–1240.

Ittner, C. D., Michels, J., 2017. Risk-based forecasting and planning and management earnings

forecasts. Review of Accounting Studies 22, 1005–1047.

Kroos, P., Schabus, M., Verbeeten, F., 2018. Voluntary clawback adoption and the use of financial measures in CFO bonus plans.

Labro, E., Lang, M., Omartian, J. D., 2023. Predictive analytics and centralization of authority. Journal of Accounting and Economics 75.

Panaretos, V. M., Zemel, Y., 2019. Statistical Aspects of Wasserstein Distances. Annual Review of Statistics and Its Application 6, 405–431.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P. H., Frazier, D. T., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Trapero Arenas, J. R., Wang, X., Winkler, R. L., Yusupova, A., Ziel, F., 2022. Forecasting: theory and practice. International Journal of Forecasting 38, 705–871.

Subrahmanyan, S. N., Jalona, S., 2020. Building a Data-Driven Culture from the Ground Up. Harvard Business Review .

Varian, H. R., 2014. Big data: New tricks for econometrics. Journal of Economic Perspectives 28, 3–28.

Waller, D., 2020. 10 steps to creating a data-driven culture. Harvard Business Review .

Zacharakis, A. L., Shepherd, D. A., 2001. The nature of information and overconfidence on venture capitalists' decision making. Journal of Business Venturing 16, 311–332.

Fig. 1
Distributional Statistics of Forecast Scenarios.

This figure plots distributional statistics of the five sales forecast scenarios for 2017. The left pane presents box-and-whisker plots for the amounts of sales growth in each forecast scenario, calculated with respect to 2015 sales for the plant. The right pane presents box-and-whisker plots of the probabilities that managers assign to each forecast scenario. The box depicts the 25th, 50th, and 75th percentiles of the distribution, whereas the ends of the whiskers show the 5th and 95th percentiles.

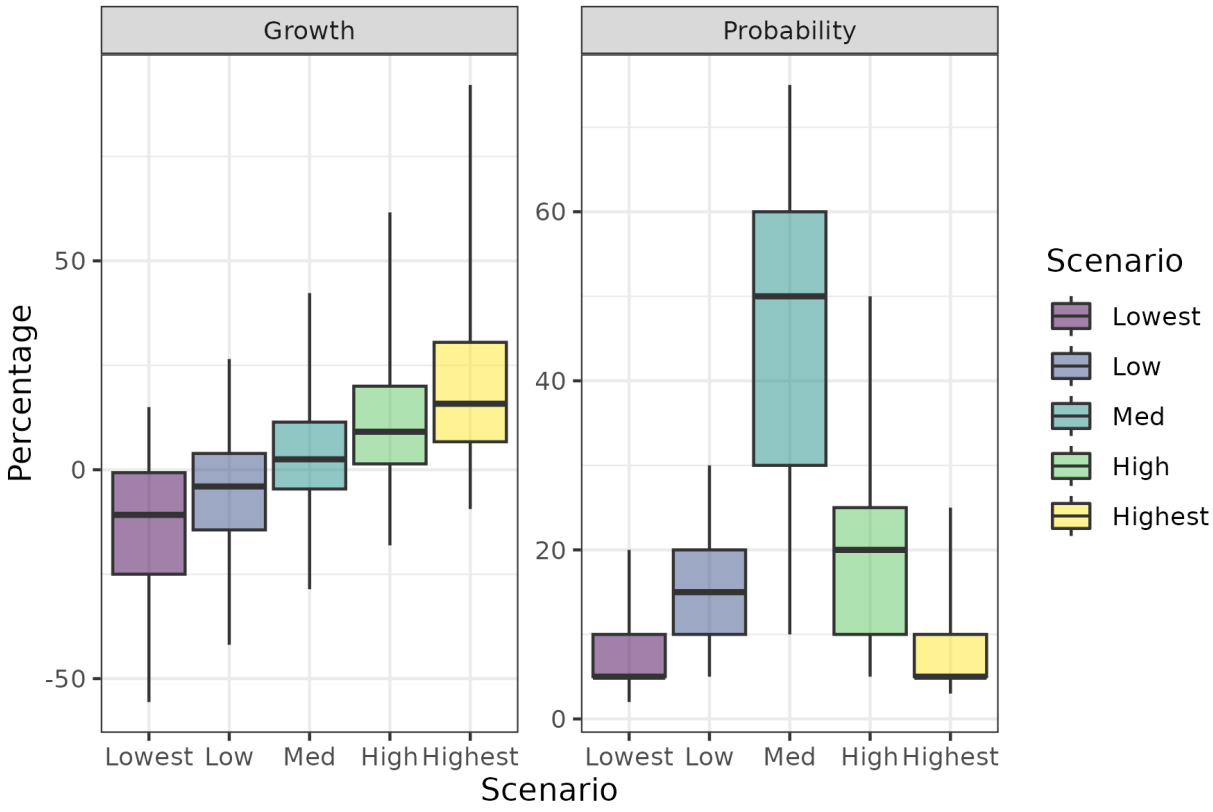

Fig. 2
Comparing Average Forecast Distributions Between High and Low Data Intensity Plants.

This figure plots distributional statistics of the five sales forecast scenarios for 2017, partitioned by whether or not the plant has a $DataIntensity_p$ value above the sample mean. The horizontal axis records the scenario growth and the vertical axis records the scenario probability. Error bars indicate 99% confidence intervals, clustering by industry and geography.

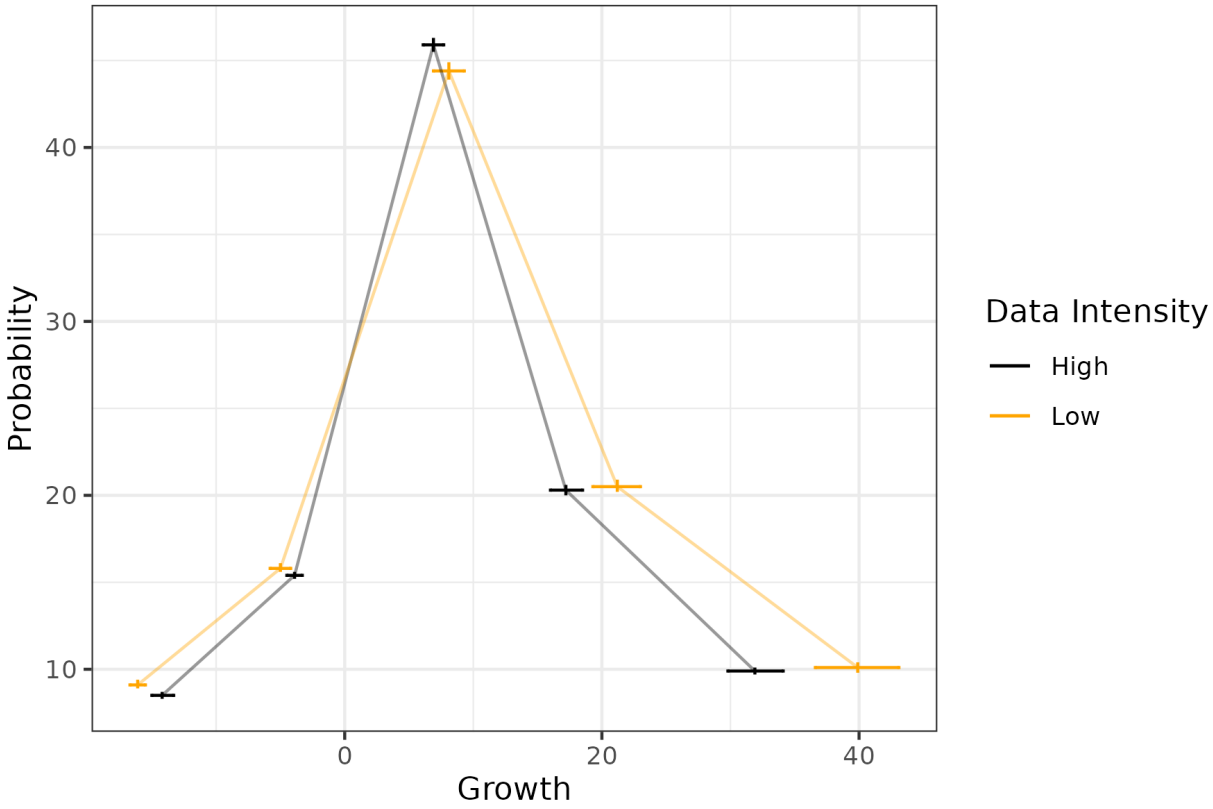

Table 1
Mapping the Sales Forecast Distribution into Actual Sales Growth.

This table presents regressions of actual sales growth in the distribution of the plant's sales forecast scenarios. The sample excludes singleton plants—plants that would uniquely identify a geographic (CBSA), industry (6-digit NAICS), and/or firm fixed effect dummy variable. Variable definitions can be found in Appendix A. Panel B focuses on the first two moments of the forecast distribution; Panel A focuses on either the reported scenarios or percentile points in the forecasted distribution. Column (3) of Panel A, which layers on the full set of fixed effects, reports the coefficient signs and significance only for Census disclosure avoidance purposes. All variables have been scaled to mean 0 and unit variance for ease of interpretation. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

Panel A: Mapping Components of the Forecast Distribution

	$ActualSalesGrowth_p$		
	(1)	(2)	(3)
$LowestSalesGrowth_p$	0.068**		
	(0.027)		
$LowSalesGrowth_p$	0.027		
	(0.049)		
$MedSalesGrowth_p$	0.209**		
	(0.096)		
$HighSalesGrowth_p$	−0.067		
	(0.089)		
$HighestSalesGrowth_p$	−0.053*		
	(0.027)		
$LowestProbability_p$	−0.012		
	(0.009)		
$LowProbability_p$	−0.012		
	(0.011)		
$HighProbability_p$	0.018**		
	(0.008)		
$HighestProbability_p$	0.024***		
	(0.008)		
$ForecastSalesGrowth10Pctl_p$		0.152***	+***
		(0.025)	
$ForecastSalesGrowth30Pctl_p$		−0.043	−
		(0.049)	
$ForecastSalesGrowth50Pctl_p$		0.076	+
		(0.051)	
$ForecastSalesGrowth70Pctl_p$		−0.010	−
		(0.043)	
$ForecastSalesGrowth90Pctl_p$		−0.003	−
		(0.033)	
Intercept	0.000	0.000	
	(0.017)	(0.016)	
Geography FE	No	No	Yes
Industry FE	No	No	Yes
Firm FE	No	No	Yes
N	12,500	12,500	12,500
R^2	0.037	0.027	

Panel B: Industry, Geography, and Firm Components

	$ActualSalesGrowth_p$				
	(1)	(2)	(3)	(4)	(5)
Intercept	0.024				
	(0.018)				
$ForecastSalesGrowthEV_p$	0.338***	0.313***	0.339***	0.286***	0.288***
	(0.027)	(0.028)	(0.028)	(0.033)	(0.034)
$ForecastUncertainty_p$	−0.008	−0.009	−0.011	−0.020	−0.016
	(0.028)	(0.025)	(0.030)	(0.028)	(0.029)
$ForecastSalesGrowthEV_p \times ForecastUncertainty_p$	−0.031***	−0.027***	−0.030***	−0.024***	−0.024***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Geography FE	No	Yes	No	No	Yes
Industry FE	No	No	Yes	No	Yes
Firm FE	No	No	No	Yes	Yes
N	12,500	12,500	12,500	12,500	12,500
R^2	0.037	0.097	0.105	0.352	0.440

Table 2
Validation of Data Intensity Measure.

Panel A provides summary statistics of the two component variables to our data intensity construct. Panel B reports regressions of data intensity on logged IT capital stock, and 2010 to 2015 changes in data intensity on 2010 to 2015 changes in logged IT capital stock. Odd-numbered columns include industry and geography fixed effects; even-numbered columns layer on firm fixed effects but again only report the signs and significance of the coefficients. All variables are defined in Appendix A, and each has been scaled to mean 0 and unit variance prior to estimating the regression for ease of interpretation. Standard errors, listed in parentheses, are double-clustered by industry and geography. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

Panel A: $DataIntensity_p$ Components

	2015 Value		Recall of 2010 Value	
	Mean	Std. Dev.	Mean	Std. Dev.
$DataAvailability_p$	0.728	0.210	0.612	0.266
$DataUse_p$	0.666	0.167	0.563	0.217

Panel B: Data Intensity and IT Capital

	$DataIntensity_{p,2015}$		$DataIntensity_{p,2015} - DataIntensity_{p,2010}$	
	(1)	(2)	(3)	(4)
$\ln\left(ITCapital_{p,2015}\right)$	0.114***	+**		
	(0.010)			
$\ln\left(ITCapital_{p,2015}\right) - \ln\left(ITCapital_{p,2010}\right)$			0.055***	+**
			(0.008)	
Geography FE	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
N	24,500	24,500	24,500	24,500
R^2	0.101		0.074	

Table 3
Differences in Forecasting Associated with Data Intensity.

Each cell reports the coefficient and its associated standard error from a separate regression of various dependent variables on $DataIntensity_p$. Specifications in column (1) include no fixed effects, column (2) includes industry (6-digit NAICS) and geography (CBSA) fixed effects, and column (3) includes firm fixed effects. Column (2) limits reporting to the signs and significance of the coefficients for Census disclosure avoidance purposes. The models are estimated on the full sample of 24,500 plants. Variable definitions can be found in Appendix A. Standard errors, listed in parentheses, are double-clustered by geography and industry. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

	Fixed Effects					
	None		Industry & Geography (i, g)		Firm (f)	
Dependent Variable	(1)		(2)		(3)	
$LowestSalesGrowth_p$	0.014^{***}	(0.002)	$+^{***}$		0.006	(0.005)
$LowSalesGrowth_p$	0.005^{***}	(0.002)	$+^{*}$		0.001	(0.006)
$MedSalesGrowth_p$	-0.005^{*}	(0.003)	$-^{*}$		−0.005	(0.009)
$HighSalesGrowth_p$	-0.019^{***}	(0.005)	$-^{***}$		−0.010	(0.013)
$HighestSalesGrowth_p$	-0.039^{***}	(0.008)	$-^{***}$		−0.033	(0.023)
$AsymmetrySalesGrowth_p$	-0.014^{***}	(0.004)	$-^{***}$		−0.016	(0.013)
$LowestProbability_p$	-0.003^{***}	(0.001)	$-^{***}$		0.000	(0.002)
$LowProbability_p$	-0.003^{***}	(0.001)	$-^{***}$		-0.005^{***}	(0.002)
$MedProbability_p$	0.008^{***}	(0.001)	$+^{***}$		−0.003	(0.005)
$HighProbability_p$	−0.001	(0.001)	+		0.003	(0.003)
$HighestProbability_p$	−0.001	(0.001)	+		0.005^{**}	(0.002)
$AsymmetryProbability_p$	0.004^{**}	(0.002)	$+^{***}$		0.014^{***}	(0.005)
$ForecastUncertainty_p$	-0.048^{***}	(0.008)	$-^{***}$		−0.031	(0.020)
$RangeSalesGrowth_p$	-0.052^{***}	(0.008)	$-^{***}$		-0.039^{*}	(0.022)
$ForecastSalesGrowthEV_p$	-0.010^{**}	(0.004)	$-^{**}$		−0.007	(0.011)
$ActualSalesGrowth_p$	0.006^{*}	(0.003)	$+^{**}$		0.004	(0.008)

Table 4
Data Intensity and Location of Realized Sales Growth in Forecast Distribution.

This table presents linear probability models of the likelihood of realized sales growth appearing within certain portions of the forecast distribution. The dependent variable in column (1) takes on a value of 1 if plant p's actual growth is below the forecasted distribution; columns (2)-(6) have indicators if the actual growth is in the first (bottom) through fifth (top) quintiles, and (7) is an indicator if p's growth was above the distribution. Precise details of the conversion from a discrete 5-point forecast distribution to a probability distribution function admitting continuous growth values can be found in Appendix B. The tabulated coefficients come from regressing distributional indicators on the data intensity measure using the full sample; below the line presents the signs and significance of re-estimating the regressions with geography (CBSA) and industry (6-digit NAICS) fixed effects. All specifications are estimated on the full sample of 24,500 plants. Standard errors are double-clustered by industry and geography. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

		Quintile of Forecast Distribution					
	$Below_p$	$1^{st}{}_p$	$2^{nd}{}_p$	$3^{rd}{}_p$	$4^{th}{}_p$	$5^{th}{}_p$	$Above_p$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	0.231^{***}	0.203^{***}	0.085^{***}	0.067^{***}	0.068^{***}	0.142^{***}	0.204^{***}
	(0.005)	(0.004)	(0.002)	(0.002)	(0.002)	(0.003)	(0.005)
$DataIntensity_p$	0.005	−0.001	-0.004^{**}	-0.003^{*}	-0.005^{***}	−0.002	0.010^{***}
	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)	(0.003)	(0.003)
Including Industry and Geography Fixed Effects:							
$DataIntensity_p$	+	+	−	−	$-^{***}$	−	$+^{**}$

Table 5

Data Intensity and Mapping the Sales Forecast Distribution into Actual Sales Growth.

This table presents links between data intensity and the predictive power of forecast distributions. Columns (1) and (2) present the influence of data on the mapping between the expected value of forecasted growth and actual growth; columns (3) and (4) present different points of the forecast distribution. Variable definitions can be found in Appendix A. Even columns include geography (CBSA), industry (6-digit NAICS), and firm fixed effects, whereas odd columns include no fixed effects. Each model is estimated on the sample of plants that excludes singletons. Standard errors are double-clustered by industry and geography. All variables are standardized to mean 0 and unit variance for ease of interpretation. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

	$ActualSalesGrowth_p$			
	(1)	(2)	(3)	(4)
Intercept	+		−	
$ForecastSalesGrowthEV_p$	$+^{***}$	0.284^{***}		
		(0.033)		
$ForecastUncertainty_p$	−	−0.016		
		(0.029)		
$ForecastSalesGrowthEV_p \times ForecastUncertainty_p$	$-^{***}$	-0.024^{***}		
		(0.005)		
$DataIntensity_{2015,p}$	−	0.008	−	0.009
		(0.012)		(0.012)
$ForecastSalesGrowthEV_p \times DataIntensity_{2015,p}$	$-^{***}$	-0.063^{***}		
		(0.019)		
$ForecastSalesGrowth10Pctl_p$			$+^{***}$	0.119^{***}
				(0.024)
$ForecastSalesGrowth50Pctl_p$			+	0.036
				(0.031)
$ForecastSalesGrowth90Pctl_p$			−	−0.006
				(0.022)
$ForecastSalesGrowth10Pctl_p \times DataIntensity_{2015,p}$			+	0.020
				(0.022)
$ForecastSalesGrowth50Pctl_p \times DataIntensity_{2015,p}$			−	−0.027
				(0.036)
$ForecastSalesGrowth90Pctl_p \times DataIntensity_{2015,p}$			$-^{*}$	-0.056^{*}
				(0.031)
Geography FE	No	Yes	No	Yes
Industry FE	No	Yes	No	Yes
Firm FE	No	Yes	No	Yes
N	12,500	12,500	12,500	12,500
R^2		0.442		0.439

Table 6
Managerial Learning and Forecast Mapping.

This table explores the influence of data intensity on the mapping of forecasted growth into actual growth based on the recency of investments in data intensity. Variable definitions can be found in Appendix A. Even columns include geography (CBSA), industry (6-digit NAICS), and firm fixed effects, whereas odd columns include no fixed effects. Each model is estimated on the sample of plants that excludes singletons. Standard errors are double-clustered by industry and geography. All variables are standardized to mean 0 and unit variance for ease of interpretation. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

	$ActualSalesGrowth_p$	
	(1)	(2)
Intercept	+	
$ForecastSalesGrowthEV_p$	$+^{***}$	0.282^{***}
		(0.032)
$ForecastUncertainty_p$	−	−0.017
		(0.029)
$ForecastSalesGrowthEV_p \times ForecastUncertainty_p$	$-^{***}$	-0.024^{***}
		(0.005)
$DataIntensity_{2010,p}$	−	0.006
		(0.015)
$ChangeDataIntensity_{2015,p}$	+	0.014
		(0.015)
$ForecastSalesGrowthEV_p \times DataIntensity_{2010,p}$	−	-0.077^{***}
		(0.022)
$ForecastSalesGrowthEV_p \times ChangeDataIntensity_{2015,p}$	$-^{*}$	-0.041^{*}
		(0.022)
Geography FE	No	Yes
Industry FE	No	Yes
Firm FE	No	Yes
N	12,500	12,500
R^2		0.443

Table 7
Benchmarking Data Intensity with Managerial Education and Tenure.

This table benchmarks the relation of managerial education and respondent tenure on forecast precision and information content against data intensity. Variable definitions can be found in Appendix A. Even columns include geography (CBSA), industry (6-digit NAICS), and firm fixed effects, whereas odd columns include no fixed effects. Each model is estimated on the sample of plants that excludes singletons. Standard errors are double-clustered by industry and geography. All variables are standardized to mean 0 and unit variance for ease of interpretation. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

	$ForecastUncertainty_p$		$ActualSalesGrowth_p$	
	(1)	(2)	(3)	(4)
Intercept	0.000		−0.001	
	(0.011)		(0.017)	
$DataIntensity_p$	−0.028***	−0.027*	0.001	0.011
	(0.009)	(0.014)	(0.013)	(0.012)
$MgrEducation_p$	−0.030***	−0.019	−0.017	−0.003
	(0.012)	(0.016)	(0.012)	(0.013)
$\ln(Tenure_p)$	−0.022**	−0.025	−0.017*	−0.009
	(0.010)	(0.016)	(0.010)	(0.011)
$ForecastSalesGrowthEV_p$			0.126***	0.115***
			(0.015)	(0.016)
$ForecastSalesGrowthEV_p \times DataIntensity_p$			−0.049***	−0.063***
			(0.016)	(0.019)
$ForecastSalesGrowthEV_p \times MgrEducation_p$			0.031*	0.034**
			(0.016)	(0.016)
$ForecastSalesGrowthEV_p \times \ln(Tenure_p)$			−0.008	0.000
			(0.016)	(0.016)
Geography FE	No	Yes	No	Yes
Industry FE	No	Yes	No	Yes
Firm FE	No	Yes	No	Yes
N	12,500	12,500	12,500	12,500
R^2	0.002	0.346	0.022	0.435

Table 8
Total Factor Productivity and Data Intensity in Unexpected Scenarios.

This table presents the change in logged total factor productivity at plant p as a function of realized sales growth and data intensity. Total factor productivity is calculated at the plant level by Census according to the methodology outlined in Foster et al. (2016). $AboveHighest_p$ is an indicator if realized sales growth is above the highest forecast scenario. $HighToHighest_p$ is an indicator if realized sales are greater than the high forecast scenario but lower than the highest scenario. $LowestToLow_p$ is an indicator if realized sales is greater than the lowest forecast scenario but lower than the low scenario. $BelowLowest_p$ is an indicator if realized sales growth is below the lowest forecast scenario. All variables are defined in Appendix A. With the exception of the indicator variables, all variables are standardized to mean 0 and unit variance for ease of interpretation. Standard errors, listed in parentheses, are double-clustered by industry and geography. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

	$ChangeLnTFP_p$		
	(1)	(2)	(3)
$AboveHighest_p$	0.161***		0.161***
	(0.018)		(0.018)
$HighToHighest_p$	0.094***		0.095***
	(0.020)		(0.020)
$LowestToLow_p$	−0.020		−0.020
	(0.022)		(0.022)
$BelowLowest_p$	−0.179***		−0.179***
	(0.019)		(0.019)
$DataIntensity_p$		0.013*	−0.007
		(0.007)	(0.011)
$DataIntensity_p \times AboveHighest_p$			0.029
			(0.022)
$DataIntensity_p \times HighToHighest_p$			0.008
			(0.026)
$DataIntensity_p \times LowestToLow_p$			0.017
			(0.021)
$DataIntensity_p \times BelowLowest_p$			0.037**
			(0.019)
Geography FE	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
N	23,000	23,000	23,000
R^2	0.099	0.085	0.099

Table 9
Interplant Forecast Similarity and Data Intensity.

This table presents regressions of the first order Wasserstein distance between two plants' forecasted sales growth distributions (focal plant denoted p, peer plant p'). The sample for column (1) is constructed by creating unordered same-firm plant pairs ($\{p, p'\} \in$ firm f). To these unordered pairs, we add an equal number of plant pairs, unmatched by firm, constructed by shuffling the second plant in the original set of plant pairs. In column (2) we limit the sample to same-firm pairs and exclude pairs where either p or p' is co-located with corporate headquarters or where we do not have a data intensity measure for corporate headquarters. In column (3), we return to the sample for column (1) but add to the sample a random sample different-firm plant pairs from the same industry or same geography. We create these unordered pairs by taking the set of all plants where we have multiple plants per firm, and matching those plants to all other plants within the set that have the same 6-digit NAICS industry or CBSA. All variables are defined in Appendix A. Standard errors, listed in parentheses, are double-clustered by plant p and plant p'. *, **, and *** denote significance at better than 10%, 5%, and 1% respectively.

	$WDist_{p,p'}$		
	(1)	(2)	(3)
Intercept	0.287***	0.258***	0.292***
	(0.010)	(0.011)	(0.010)
$SameFirm_{p,p'}$	−0.011***		−0.012**
	(0.002)		(0.005)
$DataIntensity_{p,p'}$	−0.011*	−0.015*	−0.011*
	(0.006)	(0.007)	(0.006)
$SameFirm_{p,p'} \times DataIntensity_{p,p'}$	−0.008***		−0.010***
	(0.002)		(0.003)
$HQDataIntensity_f$		−0.031**	
		(0.013)	
$SameIndustry_{p,p'}$			−0.018*
			(0.011)
$SameGeography_{p,p'}$			0.002
			(0.012)
$SameIndustry_{p,p'} \times DataIntensity_{p,p'}$			0.001
			(0.007)
$SameGeography_{p,p'} \times DataIntensity_{p,p'}$			0.016***
			(0.006)
N	165,000	53,000	499,000
R^2	0.001	0.004	0.001

Appendix A. Variable Definitions

AboveHighest$_p$ indicator if actual sales for 2017 is above the forecasted "highest" scenario

ActualSalesGrowth$_p$ Difference in plant p's 2017 total value of shipments from 2015 total value of shipments, scaled by 2015 total value of shipments; winsorized at 1% and 99%. From 2017 CMF and 2015 ASM

AsymmetryProbability$_p$ $\left(HighestProbability_p + HighProbability_p\right) - \left(LowProbability_p + LowestProbability_p\right)$

AsymmetrySalesGrowth$_p$ $\left(HighestSalesGrowth_p - MedSalesGrowth_p\right) - \left(MedSalesGrowth_p - LowestSalesGrowth_p\right)$

BelowLowest$_p$ indicator if actual sales for 2017 is lower than the forecasted "lowest" scenario

ChangeDataIntensity$_{2015,p}$ $DataIntensity_{2015,p} - DataIntensity_{2010,p}$

ChangeLnTFP$_p$ Change in logged total factor productivity from 2015 to 2017 using the Census Bureau's estimation of plant level TFP, calculated as the difference between logged outputs and logged inputs using industry elasticities for capital stock, labor, and materials. The TFP estimation process is described extensively in Foster et al. (2016)

DataAvailability$_p$ Response to question 24 of MOPS, encoded on a 1–5 scale with 1 being "Data to support decision making are not available" and 5 being "All the data we need to support decision making is available"

DataIntensity$_{y,p}$ Combination of year y responses to MOPS questions 24 and 25 (availability and use of data for decision-making). Each question is encoded on a 1–5 scale with 1 being the lowest amount and 5 being the top amount, and then standardized to mean 0 and unit variance. The two standardized responses are then added and the sum is standardized to mean 0 and unit variance. When no year subscript is specified, the year is 2015

DataIntensity$_{p,p'}$ Minimum of $DataIntensity_p$ and $DataIntensity_{p'}$

DataUse$_p$ Response to question 25 of MOPS, encoded on a 1–5 scale with 1 being "Decision making does not use data" and 5 being "Decision making relies entirely on data"

ForecastSalesGrowthEV$_p$ Expected value of plant p's sales growth from 2015 to 2017, measured from 2015 MOPS question 31 and realized sales from 2015 ASM. Calculated using the formula:

$$\sum_{X\in\{\text{Lowest,Low,Medium,High,Highest}\}} \left(XSalesGrowth_p \times XProbability_{p,n}\right)$$

ForecastSalesGrowthXXPctl$_p$ XXth percentile of the forecasted probability distribution of growth in sales for plant p from 2015 to 2017, measured from 2015 MOPS question 31 and realized sales from 2015 ASM

ForecastUncertainty$_p$ Standard deviation of the distribution of plant p's forecasted sales growth from 2015 to 2017, measured from 2015 MOPS question 31 and realized sales from 2015 ASM. Calculated using the formula:

$$\sqrt{\sum_{X\in\{\text{Lowest,Low,Medium,High,Highest}\}} \frac{(ForecastSalesGrowthEV_p - XSalesGrowth_p)^2}{XProbability_p}}$$

HighProbability$_p$ probability associated with the "high" sales growth scenario from question 31 of MOPS

HighSalesGrowth$_p$ 2015 to 2017 growth in products shipped for the "lowest" scenario, calculated from question 31 using 2015 products shipped from question 30 as the baseline

HighToHighest$_p$ indicator if actual sales growth for 2017 is greater than the forecasted "high" scenario but less than or equal to the "highest" scenario

HighestProbability$_p$ probability associated with the "highest" sales growth scenario from question 31 of MOPS

HighestSalesGrowth$_p$ 2015 to 2017 growth in products shipped for the "highest" scenario, calculated from question 31 using 2015 products shipped from question 30 as the baseline

HQDataIntensity$_f$ $DataIntensity_p$ of the plant at firm f that is listed as co-located with corporate headquarters

ITCapital$_{p,y}$ Stock of IT capital investment by plant p in year y. Measured by aggregating capital expenditures in computing hardware from 2002 and software from 2006, deflated by the BEA CPI for computer equipment and depreciated by 35% per year using a perpetual inventory method (Brynjolfsson and McElheran, 2016a).

LowestProbability$_p$ probability associated with the "lowest" sales growth scenario from question 31 of MOPS

LowestSalesGrowth$_p$ 2015 to 2017 growth in products shipped for the "lowest" scenario, calculated from question 31 using 2015 products shipped from question 30 as the baseline

LowProbability$_p$ probability associated with the "low" sales growth scenario from question 31 of MOPS

LowSalesGrowth$_p$ 2015 to 2017 growth in products shipped for the "lowest" scenario, calculated from question 31 using 2015 products shipped from question 30 as the baseline

LowestToLow$_p$ indicator if actual sales growth for 2017 is greater than or equal to the forecasted "lowest" scenario but less than the "low" scenario

MedSalesGrowth$_p$ 2015 to 2017 growth in products shipped for the "lowest" scenario, calculated from question 31 using 2015 products shipped from question 30 as the baseline

MedProbability$_p$ probability associated with the "medium" sales growth scenario from question 31 of MOPS

MgrEducation$_p$ response to question 40 of MOPS from 2015: "what was the percent of managers at this establishment with a bachelor's degree?" Encoded "20% or less" = 0, "21-40%" = 0.25, "41-60%" = 0.5, "61-80%" = 0.75, or "More than 80%" = 1

RangeSalesGrowth$_p$ $HighestSalesGrowth_p - LowestSalesGrowth_p$

SameFirm$_{p,p'}$ Indicator if plants p and p' are part of the same company

SameGeography$_{p,p'}$ Indicator if plants p and p' belong to the same core-based statistical area (CBSA). If the plant is not part of any CBSA, consider the remainder of the state in which the plant is located as its CBSA

SameIndustry$_{p,p'}$ Indicator if the primary 6-digit NAICS is the same for plants p and p'

Tenure$_p$ Number of years respondent has worked at the firm

WDist$_{p,p'}$ Wasserstein distance between forecasted growth and plants p and p'. Calculated as the minimum of the probability mass times the distance in sales growth necessary to convert the forecast distribution for plant p into the forecast distribution for plant p'

Appendix B. Inferring Percentiles from 5-Point Discrete Distributions of Expectations

The forecast distributions provided by respondents are five-point discrete distributions. When determining where in the forecast distribution actual sales falls (e.g. in Table 4), we need to convert the response into a continuous probability distribution function. We do so using the following algorithm:

1. Determine the midpoint in sales growth between each adjacent forecast scenarios. These are the interior cutpoints.

2. Calculate a left endpoint as the "lowest" scenario minus the distance to the midpoint between the "lowest" and the "low" scenario. Similarly, calculate a right endpoint as the "highest" scenario plus the distance to the midpoint between the "high" and "highest" scenarios.

3. Allocate probability mass between cutpoints (or endpoints for the lowest and highest scenarios) based on the probability assigned to each scenario. Assign half of the probability mass between the left cutpoint and the midpoint (i.e., the scenario's growth value), and half of the probability mass between the midpoint and the right cutpoint.

For example, if a plant gave the five forecast scenarios represented by the gray dots below, we would transform it to the superimposed probability distribution function. Note that there are two "columns" associated with each reported scenario—one to the left and one to the right—each with the same area. The differing heights between these two columns are a result of different widths, driven by the spacing of the 5 scenarios on the growth axis.