Data Analytics Supports Decentralized Innovation

Lynn Wu

Bowen Lou

Lorin Hitt

The Wharton School

University of Pennsylvania

Abstract

Modern analytics technology can accelerate the innovation process by increasing the rate at which existing knowledge can be identified, accessed, combined and deployed to address new problem domains. However, like prior advances in information technology, the ability of firms to exploit these opportunities depends on a variety of complementary human capital and organizational capabilities. We focus on how the benefits of analytics to innovation may differ depending on how firms organize their innovative activities. Our analysis draws on prior work that has measured firm analytics capability using detailed employee-level data and matches these data to metrics on intra-firm inventor networks that reveal whether a firm has a centralized or decentralized innovation structure. In a panel of large firms from the years 1988 to 2013, we find that firms with a decentralized innovation structure have a greater demand for analytics skills and receive greater productivity benefits from their analytics capabilities, consistent with a complementarity between analytics and decentralized innovation. Furthermore, we find the complementarity is strongest for innovation involving the recombination of existing technologies. Our results suggest that organizational structures, such as decentralization, have a substantial influence on the ability of firms to capture the benefits of analytics for innovation, primarily because analytics can facilitate the linking of distant technologies between different decentralized groups.

Introduction

Innovation is critical to the growth of advanced economies. Technology, especially information technology, has always been closely linked to innovation as both an enabler and a result of the innovative process over the last several decades. The recent rise of analytics technology may have an especially important role in supporting innovation given that the production of new knowledge is closely related to the ability to exploit the existing stock of knowledge (Joshi et al. 2010, Majchrzak and Malhotra 2016). Analytics capabilities, as defined by the ability to detect hidden patterns in large-scale data, have substantially been improved by advances in artificial intelligence and digitization. By accessing a wide range of knowledge, both within the firm and external to it, analytics can accelerate the rate at which different ideas can be combined and applied in new problem domains. The ability of analytics to support certain types of innovative activity is thus a potentially important mechanism for creating economic value (Brynjolfsson and McElheran 2016, McElheran and Brynjolfsson 2015).

However, not all firms have been able to take advantage of the opportunities created by the increased availability of data. A recent study revealed 59% of firms fail to use advanced analytics despite possessing the necessary data (Bradstreet 2017). One possible explanation is that not all firms have strategies or other organizational structures that are suited to the use of analytics. Prior work has suggested that IT-organizational complementarities are a particularly important determinant of both the demand for and benefits reaped from IT investment (e.g. Bharadwaj et al. (2007), Bresnahan et al. (2002), Nagle (2017), Pang (2016), Tafti et al. (2013)). This relationship between IT investments and benefits attained is likely to hold true for investments in analytics capabilities, although the complements critical to leveraging analytics appear to differ from those needed for leveraging general IT.

Prior work has also suggested that IT plays a role in driving the innovation process (Gao and Hitt 2012, Kleis et al. 2012), and this may be especially true for analytics technology due to its close relationship with the acquisition and production of information. However, these studies have not addressed the organizational processes (e.g., allocation of decisions, incentives, human capital, structure of formal and informal pathways for the flow of information or the exercise of authority) that potentially enhance or inhibit these

relationships. As a first step, we focus specifically on one important form of investment in analytics, the hiring of employees with data analytics skills (Tambe 2014, Wu et al. 2017) and on one structural characteristic of firms, the choice of using a centralized versus decentralized organizing principle that is known to have an important influence on the nature and outcome of innovation (Argyres and Silverman 2004, Siggelkow and Rivkin 2006)[1]. Our goal is to discover whether there is empirical support for a relationship between innovation and analytics, based on their respective (and mostly independent) literatures, and to demonstrate that examining the relationship between analytics and innovation is important for fostering further research. Moreover, complementarities between the particular practices we found are unlikely to the only ones associated with data analytics. As analytics skills become more widespread across firms it is important to identify other complementary assets that facilitate innovation and productivity.

Motivating Examples and Supporting Theory

Our analysis is motivated by the observation that successful firms structure their innovation activity in substantially different ways and that the choice of structure may potentially affect the ability to gain benefits from the increase in available data and associated analytics tools. For instance, Google and Apple, two of the most innovation-driven firms in the information economy,[2] organize their innovation processes differently, as shown in their respective networks of patent co-authorship (Figure 1). Google displays a more decentralized innovation structure with many small groups of loosely connected inventors and some larger clusters with ties extending throughout the firm. In such a structure where information is often hard to transfer across organizational boundaries (Von Hippel 1994), individual groups directly involved in a firm's operations are often better at understanding the nature of operational problems and at creating and implementing solutions. By contrast, at Apple, much of the innovation output is centralized in a few tightly-

[1] The IT value literature, largely independent from the innovation literature) documented IT to be complementary to decentralized decision making (Bresnahan et al. 2002).

[2] According to 2016 data, Google and Apple both received many patents, ranked 5th and 12th respectively in the number of patents granted by the US Patent and Trademark Office. (http://www.ipo.org/wp-content/uploads/2017/05/2016_Top-300-Patent-Owners.pdf)

linked clusters whose connections to other groups in the firm are limited. This centralized structure has the advantage of enabling the conception and development of foundational technologies that are applicable beyond the confines of a specific group or the immediate needs of current customers and local markets. This may create an advantage and a weakness: it can facilitate the search for external information and technologies which may not be of direct interest to any particular internal group (Argyres and Silverman 2004) but it may be less effective in exploiting information already known inside a group.

The structural difference between Apple and Google appears to be associated with differences in the results of innovation. The differences in innovation can be classified along two dimensions: (1) novelty of foundational ideas or the cognitive dimension of breakthroughs (2) the realization of value or the economic dimensions (Amabile and Pillemer 2012, Audia and Goncalo 2007). While the two dimensions are often correlated, they are distinct from each other (e.g. not all cognitive novel innovations are a commercial success, and not all commercial successful innovation are cognitive novel). Prior research suggests that novel ideas are more likely to be created by a centralized innovation structure such as in Apple – a company known for the creation of cognitively novel, breakthrough products that often introduce a potentially new technological trajectory (Argyres and Silverman 2004). On the other hand, Google, having a more decentralized structure, has been commercially successful in generating a steady stream of improvements to their search and targeted advertising technology that often follow on existing trajectories. As data availability has grown exponentially and data analytics is increasingly adopted by firms, we ask whether a centralized or decentralized innovation structure is better suited for leveraging the new capabilities that analytics can bring. The rapid diffusion of analytics in firms with different ex-ante innovation structures, structures that change much more slowly than analytics tools, provides the opportunity to identify the relationship between analytics and innovation.

We hypothesize that decentralized structures such as those found at Google are complementary to analytics capability for creating innovations: firms that combine analytics with decentralized innovation are likely to receive greater benefits than firms that combine analytics with centralized innovation structures. A main driver for the complementarities relationship is that analytics capabilities can mitigate a central

weakness associated with decentralization: the lack of search capabilities for acquiring diverse knowledge from many different areas (Argyres and Silverman 2004). By collecting digital traces from a variety of business processes and user behaviors both inside and outside of the department or functional area, analytics can help firms assimilate information from divergent sources that is necessary to generate new solutions. Similarly, it has been noted that as a product's design matures, informal communication channels that support the development process become deeper and narrower, reducing knowledge sharing between groups and thus limiting opportunities to link innovative ideas across different areas (Henderson and Clark 1990). Analytics can mitigate this disadvantage to some extent by automatically detecting hidden patterns among innovations from different communities, and thereby facilitate the transfer and use of knowledge across organizational boundaries.

This effort is accelerated by recent advances in data analytics such as machine learning that has become increasingly effective at discovering hidden patterns and ideas across different domains. For example, IBM's Watson digested 23 million medical papers across many different disciplines to find information related to a tumor suppressor known as p53 that is associated with half of all cancers. In a short amount of time, Watson was able to identify six previously-unknown proteins that interact with p53, a feat that would have taken researchers more than 6 years to accomplish (Chen et al. 2016), and may generate a substantial financial return (Zafar et al. 2013). Essentially, to find these proteins, Watson has employed data analytics to substantially reduce the cost of conducting a broad search and then linking distant innovation areas to create new solutions. Similarly, the British AI firm BenevolentAI was able to identify five potential hypotheses for the treatment of ALS (Amyotrophic lateral sclerosis) in less than a week, one of which is linked to preventing the death of motor neurons (Smalley 2017), overcoming a critical difficulty in treating a disease that has no known cure. Like the IBM Watson example, BenevolentAI found new treatments by linking vast quantities of complex, often unstructured scientific information including journal articles, clinical trials, and medical records across diverse disciplines and disease areas. The ability to detect subtle patterns across a wide volume of diverse knowledge accelerated the rate of innovation through combining known insights to provide new solutions.

Data analytics has impacted innovation beyond the healthcare industry. Autodesk teamed with race car drivers and engineers to use cheap sensors to collect vast amount of data on how a car's chassis responds to stresses, strains, temperature, displacement and all other factors that might affect performance. Combing the sensor data with existing chassis design knowledge from various sources, analytics helped to create a substantially improved chassis over traditional ones. For example, unlike traditional designs, the new chassis is asymmetric, mirroring the fact that a race car turn more often in one direction than the other and thus subject the chassis to different forces. The need for designing a deeply asymmetric chassis would not have been detected without using data analytics to uncover hidden patterns embedded in the new sensor data and traditional data (McAfee and Brynjolfsson 2017). As decentralized innovation structures often face difficulty in sharing information between different departments, analytics can be particularly helpful for integrating and uncovering common patterns in data across different groups, allowing a decentralized structure to recombine existing knowledge to create new inventions. Thus, we expect that analytics and decentralization are complementary (e.g., analytics are increasingly more valuable in firms that are more decentralized).

The relationship between data analytics and centralized innovation is less clear than it is for data analytics and decentralized innovation. On the one hand, the preceding arguments suggest that analytics can support the gathering of external information which could benefit innovation output generally. On the other hand, there may be limits to the benefits of analytics for centralized structures for at least two reasons. First, centralized structures have existing mechanisms for sharing internal knowledge and gathering external information, limiting the marginal benefit in these areas for any new technology (including analytics). Second, centralized structures are disproportionately used in firms engaged in foundational or cognitive novel innovation (Arora et al. 2014, Siggelkow and Rivkin 2006), that could serve as the basic building block for future recombination in innovation (such as discovering a new type of p53 kinases and an asymmetric chassis). This type of foundational innovations has less to gain from the current state of analytics technology since there is often little or no available data for such innovation because they do not

have existing producers or customers prior to their initial deployment. Despite these critical knowledge-sharing constraints, it has been noted that much of the relevant information for creating foundational innovation is tacit and shared through informal communications channels (Avery and Norton 2014), thus making it less amenable to digitization or automated analysis. Firms that adopt analytics under these conditions should expect less than full benefits but still bear the full costs of their analytics investments. Thus, we would expect a limited interaction between analytics and centralization.

Measurement

Centralization and Decentralization. We use network analysis of patent co-authorship relationships to construct an intra-firm patent network for each firm in our sample between 1988 and 2013 with one network for each year. A node in a network represents an employee inventor and an edge (link) indicates the presence of one or more coauthored patents between two inventors working in the same firm. To measure decentralization, we apply machine learning-based community detection algorithms (Multilevel) to these networks (Blondel et al. 2008) to identify distinct innovation communities and calculate a Herfindahl-based metric to measure how widely the innovation communities are dispersed for the firm (see Appendix A). Our metric has three advantages: it measures the actual structure of innovation regardless of the formal hierarchy (e.g. org chart); it is entirely data driven requiring no arbitrary definitions of organizational boundaries; and it will not be biased if the formal and informal organization of a firm diverge, an important issue noted in the knowledge management literature (Cross et al. 2001). The distribution of the innovation structure is shown in Figure 2.[3]

Innovation. Our measures of innovation output are also based on patent data consistent with prior work on R&D productivity (Griliches 1990, Hall et al. 2001). While patent data cannot cover all types of innovative activity (e.g. internal organizational processes and trade secrets), they have the advantage that

[3] The chart shows a mass at zero (maximally centralized, likely due to a limited number of inventors) and rest of the distribution is proximately lognormal. Including dummy variables for the mass at zero in our analysis did not qualitatively change the results (see Appendix A).

they can be consistently measured[4] and there is a large and robust literature on patent-related measurement approaches. We also measure novelty in patents: whether an innovation involves the creation of a new technology class (a subclass in patent classification). We also distinguish the reuse of existing combinations from the creation of new combinations (Akcigit et al. 2013). These innovation quality metrics can be measured at both the firm level (the innovation is new to the firm, but other firms have used similar combinations or technologies) or at a global level (an original technology or combination that no one else has created). Thus, we have a total of 6 ways to classify innovation along two dimensions: (1) new technology, new combination and reuse, and (2) local (firm) versus global (see Appendix B).

Analytics. For the purpose of our analysis we define analytics as the ability to process data and find patterns within data. To measure a firm's data analytics capabilities, we use six million resumes from 1988 to 2007, and 3.7 million job reviews from 2008 to 2013 to calculate the total number of employees possessing data analytics skills. We apply natural language processing techniques on free-form text (when available) and job titles to identify the analytics skills of each employee and aggregating all employees with data analytics skills for each firm in each year after adjusting for a sampling rate (based on the fraction of employees found in the data for each firm) (see Appendix C). We then link these data to financial metrics such as physical assets, employees, and sales using the Compustat Industrial Annual files. The summary statistics for the financial variables, data analytics and patent-related variables are shown in Table 1. Trends in our data analytics measure are shown in Figure 3.

Our primary analysis tests for complementarities between data analytics and innovation structure using (1) correlations (adoption or demand equations) and (2) performance differences (productivity equations). Although the demand equations have the advantage that they are relatively simple and provide the greatest power if firms are matching complementary practices optimally, they have the disadvantage that the simplicity makes them vulnerable to unobserved heterogeneity. In addition, such an analysis will tend to

[4] Consistent measurement combined with firm fixed effects or industry controls can partially address the concern that different industries may have a different mix of patents and other types of innovative outputs. Moreover, since different types of innovation are likely to be at least weakly positively correlated, patent measures may provide a useful indicator of innovative activity more broadly.

understate the strength of complements if not all firms are endowed with or able to change to the optimal match between complements. The productivity test has the advantage of a direct tie to an important firm outcome (performance), and the effects are most powerful statistically when not all firms have found the optimal match, which is likely when implementing new business practices. Over time, as the complementarities system diffuses to other firms, the correlation would increase but the productivity premium would decrease because the relative advantage of using a complementary system diminishes as its adoption spreads.

For econometric identification, we take the view that data analytics skills are rapidly changing (and potentially endogenous) while the organization of innovation (as centralized or decentralized) is quasi-fixed and (Hannan and Freeman 1984, Lam 2005) and therefore exogenous. Identification using a combination of a fast-changing practice along with a slow-changing organizational complement is consistent with prior work (Autor et al. 2003, Bresnahan et al. 2002, Brynjolfsson and Hitt 1996, Milgrom and Roberts 1990). The fact that firms may not be able to instantaneously adapt to the diffusion of analytics provides data variation that enables productivity differences to be observed. Table 2 shows the demand for data analytics when firms have a decentralized innovation structure. After controlling for firm and year fixed effects, as well as the level of R&D and patent activity, the demand for data analytics is positively correlated with a decentralized innovation structure (this measurement is consistent and robust using alternative community detection algorithms, and we use the Multilevel algorithm for our main results (see Appendix A). The difference in demand for data talent is also reflected in the difference between Google and Apple, with Google having 34% more data talent than Apple.[5]

Results

If data analytics and the decentralization of innovation are complementary, firms that possess both should experience a greater return than firms that have only one of the two complements. We first estimate a baseline regression specification in firm-level fixed effects that shows that our regression analysis yields

[5] Data talent constitutes 20.6% of Google's workforce and 13.5% of Apple's workforce.

similar estimates to those found in prior work for IT labor, as well as other productivity inputs (Table 3, Column 1). The regression continues to show reasonable properties when we add metrics for analytics and interactions with dispersion (Columns 2-4). Our key result (Column 2) shows that while having decentralized innovation structures is negatively correlated with productivity (albeit not significantly so), the interaction with data analytics is positive. On average, a firm that is more decentralized than the average (one standard deviation above the mean in our decentralization metric) receives a 2.2% increase in productivity for every standard deviation increase in data analytics skills.

To examine whether our results are affected by the potential endogeneity of analytics skills we repeat our main analysis using 2SLS and GMM. We treat innovation structure in firms as quasi-fixed, and instrument analytics and IT skills with a metric based on the adoption of enterprise systems and the flow of analytics skills in neighboring firms in the labor supply network (i.e., firms that the focal firm hires from). Our instruments are motivated by the argument that adoption of large-scale IT innovations changes the relative availability of skills in the market which affects the cost of acquiring data talent. This shift, driven by choices external to the firm, is not directly reflected in free cash flow or management characteristics that would cause data analytics skills to be influenced by prior performance (see Appendix D). The first-stage regression is shown in Column 5 of Table 2 and the associated first-stage F-statistics are above the threshold needed to pass the weak instrument test. In the GMM analysis, we also use instruments derived from information within the panel. The results on the key coefficients continue to be consistent, suggesting that the potential endogeneity of data analytics is not leading to an upward bias of our complementarities measure. We then explore how the interaction effect changes over time by separating our sample into two periods: 1988-2007 (the pre-cloud computing period) and 2008-2013 (the post-cloud computing period). We believe that this technological change may have provided a structural shift in the cost of large-scale analytics to firms (Hashem et al. 2015). We find that the interaction between analytics and dispersion is positive for both periods but the effect is stronger in later years, as shown in Columns 5 & 6.

We then use the same variables from our previous models (data analytics, decentralization, their interactions, and other controls) to determine their relationship to the production of different types of

innovation (see Table 4). First, we find that none of our metrics has a substantial effect on the production of completely new knowledge overall (Column 1), but our hypothesized complementarity is present for within-firm novel innovation, indicating that analytics can enable the decentralized innovation structure to better acquire knowledge about technologies that already exist outside the firm (Column 2). Next, we examine the effect of analytics-decentralization complementarities on innovation through the recombination of existing technologies. We distinguish new combinations from the reuse of existing combinations and find that our data-decentralization complementarity facilitates both combinations that are new to the overall market (Column 3) and those new to the firm (Column 4). On average, the interaction term implies that a one standard deviation increase in data analytics increases the number of new combination patents by 4.12% in firms that are one standard-deviation above the mean in innovation decentralization. Consistent with the prior literature, we also find that decentralization promotes innovations that constitute the reuse of existing combinations, but there appears to be no benefit of data analytics (direct or complementary) to this type of innovation (Columns 5 & 6).

To further explore the extent to which decentralization and data analytics can support the ability to integrate available external information, we use an alternative measure to capture the integration of new technologies into a firm's patent portfolio. Following Lin et al. (2016) we use the portion of citations to a firms' own prior art to characterize firm innovation, which captures the inverse of the amount of external innovation in use. In Table 5 we repeat our main analysis, counting only innovations that cite a certain fraction of a firms' prior art (from 0 to 100% in 10% increments). We find a complementarity between analytics and in-house based innovation, and the complementarity strengthens as the proportion of internal information decreases but external information increases, peaking at having 50%-60% internal citations and 40%-50% external citations. However, the complementarity becomes statistically insignificant and even turns negative (although not significantly so) when prior citations to firm's patent stock are below 20%. These results suggest that analytics is most useful when there is potentially useful external information that can be combined with a firm's own stock of patents, but not when the innovation is completely new or mostly incremental.

Finally, all our prior metrics focus on analytics of the firm broadly across any business purpose, which raises the possibility that our results are driven by the characteristics of data-intensive firms rather than by the specific use of analytics to support the innovation process. To examine whether there exists a direct link between analytics and innovators, we repeat our main analyses focusing specifically on the data analytics skills of named inventors. While named inventors are only a subset of all employees with data analytics skills and matching limits the number that we can identify, we can still conduct an exploratory analysis where we divide those with analytics skills into two groups: inventors and other employees. Overall, we find that this narrower measure, despite the likely presence of considerable measurement error, behaves similarly to our broader data skill measure. The correlation test results are straightforward: decentralization is associated with a having more innovators with data analytics skills (Table 6). The productivity test results are also similar. Furthermore, the interaction of data analytics skills for non-inventors shows a significantly positive effect, and the magnitude is substantially greater than the measure for inventors (see Table 7, column 2). This result suggests that while the complementarities associated with data analytics come from both types of employees, the complementarities effect is stronger for employees who provide infrastructure and extensive support to the inventors than it is for the inventors who have analytics skills themselves. Replication of our other results using this inventor-based measure is similar in magnitude but does not show consistently significant results, likely due to a lack of power.

Overall, this pattern of results is consistent with the idea that data analytics complements decentralized innovation to better integrate external information and to reuse novel information from within the firm. This assistance from data analytics is significant because both integrating external information and reusing novel information are known to be difficult tasks for decentralized structures. However, data analytics provides limited help in improving completely novel innovation or in the reuse of existing technologies. This finding also appears reasonable because in these situations data may have limited marginal benefit either because there is no data to integrate (in the case of a completely novel innovation) or such data already exists locally (in the case of reuse). Our result is robust for the alternative measurement of

innovation, decentralization, and whether we consider data analytics skills of inventors or employees overall.

Conclusion

As data availability has grown exponentially and data analytics is increasingly adopted by firms, our results suggest that the diffusion of analytics skills in the short run will favor the innovation process for some firms and some types of innovations over others; in the longer term, it suggests that firms can potentially benefit by shifting to a more decentralized innovation structure if they can also benefit from the types of innovations these structures produce. With data analytics becoming even more important over time, an acceleration fueled by advances in artificial intelligence and new ways of gathering data (e.g. using robotics to capture optimal machine toolpaths by analyzing the motion of human laborers), a firm's innovation structure would also likely evolve substantially to take advantage of the capabilities these technologies could bring. Those firms that can best match their existing capability to opportunities for innovation, or that can acquire such capability would likely benefit substantially from using data analytics.

Our results cover an important structure characteristic of firms that may affect the nature and outcome of firms' internal innovative processes as analytics technology becomes more widely deployed, and show that the effects on both aggregate innovation and different types of innovation are measurably influenced by complementarities between analytics and the organization of innovation. Future work could productively identify additional complements (specific technologies, additional organizational practices) or extend the investigation into other types of innovative output (process innovation, new products) which may not be as well captured by patents. Ultimately a better understanding of the relationship between analytics and innovation can enable better investment decisions in analytics capability and provide guidance on the other concurrent investments needed to capture the full benefit. As Bloom et al. (2017) pointed out, a decline in innovative output has been observed recently. Improved understanding of the relationship between analytics and innovation could explain and perhaps mitigate that decline.

Table 1. Summary Statistics

VARIABLES	(1) # obs.	(2) Mean	(3) Std. dev.	(4) Min	(5) Max
ln(Sales ($ million))	14,960	5.587	2.268	-6.830	13.00
ln(Materials ($ million))	14,960	5.240	2.015	-2.315	12.74
ln(Capital ($ million))	14,960	4.913	2.330	-1.427	12.41
ln(IT labor)	14,960	3.845	1.681	0	11.68
ln(Other labor)	14,960	6.805	2.396	0	14.60
ln(Dispersion)	14,960	0.514	0.163	0	0.690
ln(Data)	14,960	3.696	3.508	0	11.66
ln(R&D ($ million))	14,721	2.556	1.953	-4.542	9.513
New-tech Global	14,960	0.0369	0.348	0	16.50
New-combination Global	14,960	18.17	88.30	0	1,832
Reuse Global	14,960	7.537	34.09	0	900
New-tech Local	14,960	3.615	8.489	0	171.5
New-combination Local	14,960	16.63	88.65	0	1,871
Reuse Local	14,960	5.504	27.18	0	745.5

Correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. ln(Sales)	1													
2. ln(Materials)	0.946	1												
3. ln(Capital)	0.899	0.909	1											
4. ln(IT labor)	0.425	0.437	0.432	1										
5. ln(Other labor)	0.813	0.781	0.795	0.607	1									
6. ln(Dispersion)	0.0945	0.110	0.106	0.0249	0.0747	1								
7. ln(Data)	0.430	0.430	0.416	0.348	0.473	0.0387	1							
8. ln(R&D)	0.376	0.457	0.400	0.260	0.307	0.216	0.259	1						
9. New-tech Global	0.146	0.159	0.149	0.109	0.136	0.00191	0.123	0.186	1					
10. New-combination Global	0.277	0.299	0.291	0.215	0.253	0.00005	0.211	0.345	0.523	1				
11. Reuse Global	0.279	0.298	0.287	0.217	0.240	0.0143	0.200	0.364	0.421	0.899	1			
12. New-tech Local	0.357	0.386	0.376	0.259	0.338	0.0793	0.248	0.416	0.444	0.792	0.711	1		
13. New-combination Local	0.268	0.288	0.280	0.209	0.242	-0.00457	0.205	0.339	0.517	0.998	0.909	0.760	1	
14. Reuse Local	0.267	0.285	0.275	0.210	0.232	0.00825	0.192	0.346	0.416	0.881	0.994	0.681	0.889	1

Table 2. The Correlation Test: Data Analytics Skills and Innovation Community Structure

DV: ln(Data)	(1)	(2)	(3)	(4)	(5)
Model	FE	FE	FE	FE	FE
ln(Sales)	0.438***	0.437***	0.436***	0.436***	0.493***
	(0.0604)	(0.0605)	(0.0604)	(0.0604)	(0.0750)
std(Patents)	0.160	0.155	0.162	0.160	0.128
	(0.136)	(0.139)	(0.136)	(0.136)	(0.140)
ln(R&D)	0.173***	0.175***	0.172***	0.173***	0.183***
	(0.0534)	(0.0534)	(0.0534)	(0.0534)	(0.0572)
ln(Dispersion: multilevel)	0.918**				0.827**
	(0.403)				(0.408)
ln(Dispersion: infomap)		0.628*			
		(0.337)			
ln(Dispersion: leading eigenvector)			1.035***		
			(0.391)		
ln(Dispersion: fastgreedy)				0.980**	
				(0.395)	
ln(Total Neighbor Data)					-0.0530
					(0.216)
ln(Total Neighbor ERP)					0.986**
					(0.433)
ln(Total Neighbor HCM)					0.171
					(0.187)
Lagged ln(Total Neighbor Data)					0.221
					(0.214)
Lagged ln(Total Neighbor ERP)					-0.541
					(0.357)
Lagged ln(Total Neighbor HCM)					-0.137
					(0.159)
Observations	14,721	14,721	14,721	14,721	11,012
R-squared	0.658	0.658	0.658	0.658	0.684
# of firms	1,856	1,856	1,856	1,856	1,592
Industry	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES

Note:
(1) Column 1-4 are results using different community detection algorithms.
(2) Column 5 includes instrumental variables for data analytics skills, and can thus be interpreted as the first stage results in 2SLS estimation used in Column 3 of Table 3.
(3) Firm's patent stock is centered and controlled. For a total stock of patents, we capture all patents issued to the firm in the 20 years prior to the observation year, and we form a stock measure by accumulating the number of patents and then assuming an annual 15% depreciation, a value consistent with prior work in R&D productivity. We also test the yearly number of patents filed by the firm, which yields similar effect.
(4) Data source dummy for measuring data analytics skills is also controlled.
(5) Robust (clustered by firm) standard errors are reported in parentheses.
(6) *** p<0.01, ** p<0.05, * p<0.1

Table 3. The Productivity Test: Data Analytics Skills and Innovation Community Structure

DV: ln(Sales)	(1)	(2)	(3)	(4)	(5)	(6)
Model	FE	FE	2SLS	GMM/IV	FE 1988-2007	FE 2008-2013
ln(Materials)	0.641***	0.642***	0.583***	0.547***	0.598***	0.606***
	(0.0223)	(0.0224)	(0.0244)	(0.0554)	(0.0233)	(0.0614)
ln(Capital)	0.121***	0.125***	0.0822***	0.215***	0.109***	0.107***
	(0.0208)	(0.0204)	(0.0215)	(0.0564)	(0.0190)	(0.0364)
ln(IT labor)	0.0191***	0.0197***	0.0445	0.00503	0.00820**	0.0121***
	(0.00552)	(0.00556)	(0.0555)	(0.0308)	(0.00412)	(0.00460)
ln(Other labor)	0.212***	0.212***	0.259***	0.248***	0.261***	0.123**
	(0.0251)	(0.0251)	(0.0274)	(0.0671)	(0.0181)	(0.0502)
ln(Emp w/ college+)	0.00627**	0.00613**	-0.00320	0.0433*	0.00423*	-0.00265
	(0.00251)	(0.00256)	(0.00655)	(0.0244)	(0.00238)	(0.00233)
ln(Data)		-0.000381	0.0371	-0.00383	0.000263	-0.00335
		(0.00135)	(0.0439)	(0.00538)	(0.00127)	(0.00244)
ln(Dispersion)		-0.0794	-0.0793	-0.0834	-0.0422	0.0293
		(0.0502)	(0.0677)	(0.137)	(0.0492)	(0.0891)
Data X Dispersion		0.0224***	0.0300**	0.0238*	0.0186***	0.0302***
		(0.00400)	(0.0130)	(0.0135)	(0.00371)	(0.00821)
Observations	14,960	14,960	11,012	11,012	12,994	1,966
R-squared	0.985	0.985	0.733		0.986	0.997
# of firms	1,864	1,864	1,592	1,592	1,832	518
Industry	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES

Note:
(1) The Multilevel algorithm is used to measure the dispersion of innovation communities. All interactions are appropriately centered. We also replicate Column 2 of Table 3 with other measurements of decentralization. They show similar results for the estimation on interaction term (0.0212 for leading eigenvector, 0.0149 for infomap, 0.0207 for fastgreedy, see Appendix A). All are significant at $p<0.01$ (against the null hypothesis of zero).
(2) Six instrumental variables are used: 1) total number of employees with data analytics skills in neighboring firms; 2) one-period lagged values of 1); 3-4) total number of neighbors that adopt an enterprise resource planning (ERP) Human Capital Management (HCM) systems; 5-6) one-period lagged values for ERP or HCM variables in what are calculated for 3-4). The first-stage F-statistics ($F(12, 10958) = 299.88$) passes the weak instrument test.
(3) We use the Blundell and Bond (1998) SYS-GMM estimator which derives additional instruments using the lagged level and differences from production inputs. This method was originally developed specifically for micro-productivity applications, especially for measuring the productivity of R&D. We use three-period lags of the endogenous variables as instruments in addition to the external instruments used in 2SLS. The Arellano-Bond test for AR(2) autocorrelation in first differences suggests that serial correlation is not a concern in the first-differencing equations (p=0.176). Neither the Hansen test of over-identification (p= 0.277) nor the difference-in-Hansen test of the system GMM instruments (p= 0.464) rejects the null that the instruments are uncorrelated with the error term, ensuring the validity of the instruments used in the GMM estimation. We also test other lags and find they don't qualitatively change our results.
(4) Data source dummy for measuring data analytics skills, and patent variables including firm's patent stock and number of inventors are also controlled.
(5) Robust (clustered by firm) standard errors are reported in parentheses.
(6) *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 4. Data Analytics Skills, Innovation Community Structure, and Types of Innovation

	(1)	(2)	(3)	(4)	(5)	(6)
DV	ln(New-tech Global)	ln(New-tech Local)	ln(New-combination Global)	ln(New-combination Local)	ln(Reuse Global)	ln(Reuse Local)
Model	FE	FE	FE	FE	FE	FE
ln(Data)	-9.84e-05	-0.00111	-0.00211	-0.00296	-0.000663	0.00212
	(0.000395)	(0.00327)	(0.00393)	(0.00379)	(0.00351)	(0.00331)
ln(Dispersion)	0.0128	0.932***	1.274***	0.933***	0.742***	0.462***
	(0.0104)	(0.0902)	(0.123)	(0.116)	(0.0967)	(0.0883)
Data X Dispersion	-0.000263	0.0296***	0.0412***	0.0350**	0.0113	0.0167
	(0.00179)	(0.0113)	(0.0144)	(0.0138)	(0.0121)	(0.0113)
ln(Sales)	0.000439	0.0392*	0.0514**	0.0418*	0.0539**	0.0593***
	(0.00214)	(0.0200)	(0.0243)	(0.0227)	(0.0219)	(0.0199)
ln(Employees)	0.00833**	0.177***	0.257***	0.265***	0.238***	0.197***
	(0.00419)	(0.0304)	(0.0407)	(0.0386)	(0.0340)	(0.0314)
ln(Emp w/ college+)	0.000638	0.00635	0.00699	0.00187	0.00238	0.00110
	(0.000481)	(0.00564)	(0.00728)	(0.00682)	(0.00592)	(0.00564)
Observations	14,960	14,960	14,960	14,960	14,960	14,960
R-squared	0.471	0.694	0.809	0.827	0.790	0.795
# of firms	1,864	1,864	1,864	1,864	1,864	1,864
Industry	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES

Note:
(1) The Multilevel algorithm is implemented to measure the dispersion of innovation community.
(2) Data source dummy for measuring data analytics skills is also controlled.
(3) Robust (clustered by firm) standard errors are reported in parentheses.
(4) *** p<0.01, ** p<0.05, * p<0.1

Table 5. Innovative Output Relative to A Firm's Own Patent Stock

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
DV	90 - 100	80 - 90	70 - 80	60 - 70	50 - 60	40 - 50	30 - 40	20 - 30	10 - 20	0 - 10
Model	FE	FE	FE	FE	FE	FE	FE	FE	FE	FE
ln(Data)	0.00110	-0.00365	-0.00230	-0.000965	-0.00180	-0.00140	0.00135	2.14e-05	0.00212	-0.00143
	(0.00263)	(0.00240)	(0.00259)	(0.00290)	(0.00317)	(0.00328)	(0.00305)	(0.00326)	(0.00305)	(0.00330)
ln(Dispersion)	0.175**	0.0805	0.132	0.234***	0.318***	0.396***	0.343***	0.433***	0.402***	0.579***
	(0.0873)	(0.0780)	(0.0808)	(0.0891)	(0.0909)	(0.0924)	(0.0876)	(0.0852)	(0.0845)	(0.0847)
Data X Dispersion	0.0253***	0.0374***	0.0355***	0.0392***	0.0422***	0.0379***	0.0367***	0.0220*	0.0122	-0.00376
	(0.00904)	(0.00889)	(0.00920)	(0.0104)	(0.0110)	(0.0114)	(0.0107)	(0.0113)	(0.0110)	(0.0112)
ln(Sales)	0.0181	0.0100	0.0221	0.0240	0.0382**	0.0457***	0.0486***	0.0411**	0.0331*	0.0444**
	(0.0152)	(0.0114)	(0.0139)	(0.0148)	(0.0175)	(0.0175)	(0.0168)	(0.0177)	(0.0180)	(0.0193)
ln(Employees)	0.123***	0.0973***	0.0993***	0.125***	0.147***	0.152***	0.150***	0.156***	0.158***	0.193***
	(0.0255)	(0.0235)	(0.0256)	(0.0268)	(0.0286)	(0.0285)	(0.0266)	(0.0272)	(0.0271)	(0.0293)
ln(Emp w/ college+)	-0.00475	-0.00347	-0.00359	-0.00292	-0.00282	0.00228	0.000953	-0.00128	0.00372	0.00599
	(0.00571)	(0.00560)	(0.00586)	(0.00584)	(0.00593)	(0.00588)	(0.00555)	(0.00475)	(0.00508)	(0.00556)
Observations	14,960	14,960	14,960	14,960	14,960	14,960	14,960	14,960	14,960	14,960
R-squared	0.769	0.719	0.739	0.747	0.764	0.762	0.751	0.740	0.740	0.777
# of firms	1,864	1,864	1,864	1,864	1,864	1,864	1,864	1,864	1,864	1,864
Industry	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note:
(1) The Multilevel algorithm is implemented to measure the dispersion of innovation community.
(2) We take the logarithm of the dependent variable that captures the percentage of citations that come from a firm's own previous patent stock.
(3) Data source dummy for measuring data analytics skills is also controlled.
(4) Robust (clustered by firm) standard errors are reported in parentheses.
(5) *** p<0.01, ** p<0.05, * p<0.1

Table 6. The Correlation Test using Data Analytics Skills of Inventors

DV: ln(Data & Inventor)	(1)
Model	FE
ln(Dispersion)	0.265***
	(0.0903)
ln(Sales)	0.0347**
	(0.0172)
std(Patents)	0.146*
	(0.0835)
ln(R&D)	0.0388***
	(0.0138)
Observations	12,994
R-squared	0.392
# of firms	1,832
Industry	YES
Year	YES

Note: The Multilevel algorithm is implemented to measure the dispersion of innovation community. Robust (clustered by firm) standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7. The Productivity Test using Data Analytics Skills of Inventors

DV: ln(Sales)	(1)	(2)
Model	FE	FE
ln(Materials)	0.598***	0.598***
	(0.0233)	(0.0232)
ln(Capital)	0.107***	0.109***
	(0.0191)	(0.0190)
ln(IT labor)	0.00848**	0.00820**
	(0.00410)	(0.00413)
ln(Other labor)	0.262***	0.261***
	(0.0182)	(0.0181)
ln(Emp w/ college+)	0.00446*	0.00425*
	(0.00237)	(0.00238)
ln(Data & Inventor)	-0.0255	-0.0256
	(0.0177)	(0.0177)
ln(Data & Non-Inventor)		0.000274
		(0.00127)
ln(Dispersion)	-0.00370	-0.00313
	(0.0504)	(0.0498)
(Data & Inventor) X Dispersion	0.0213*	0.0208
	(0.0129)	(0.0129)
(Data & Non-Inventor) X Dispersion		0.0179***
		(0.00362)
Observations	12,994	12,994
R-squared	0.986	0.986
# of firms	1,832	1,832
Industry	YES	YES
Year	YES	YES

Note: The Multilevel algorithm is implemented to measure the dispersion of innovation community. All interactions are appropriately centered. Patent variables including firm's patent stock and number of inventors are also controlled. Robust (clustered by firm) standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

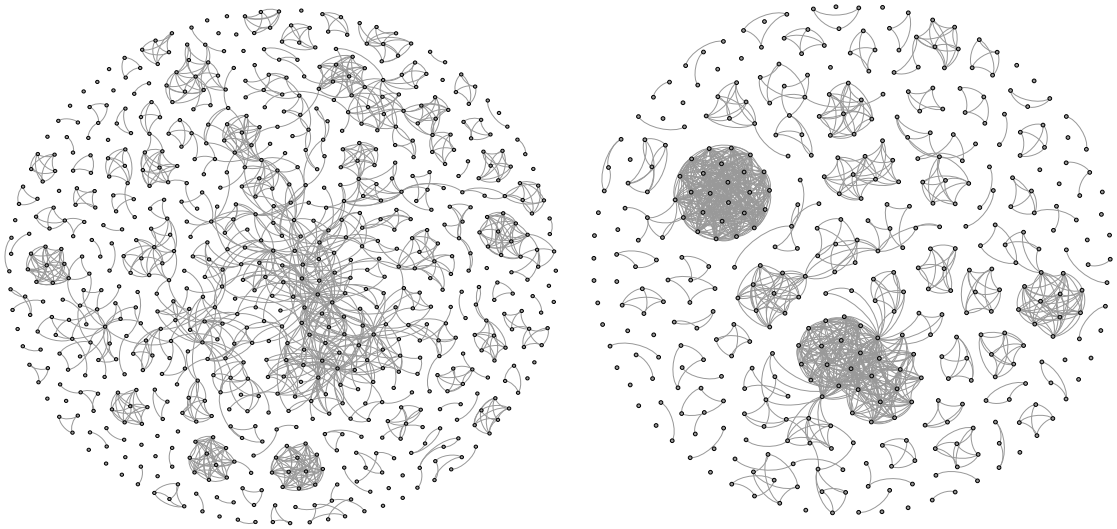

Figure 1. Comparing innovation structures between Google (left) and Apple (right). Each node on the graph is a particular inventor, while the edges are inventors appearing on the same patent. This graph is generated from the patent data in our sample.

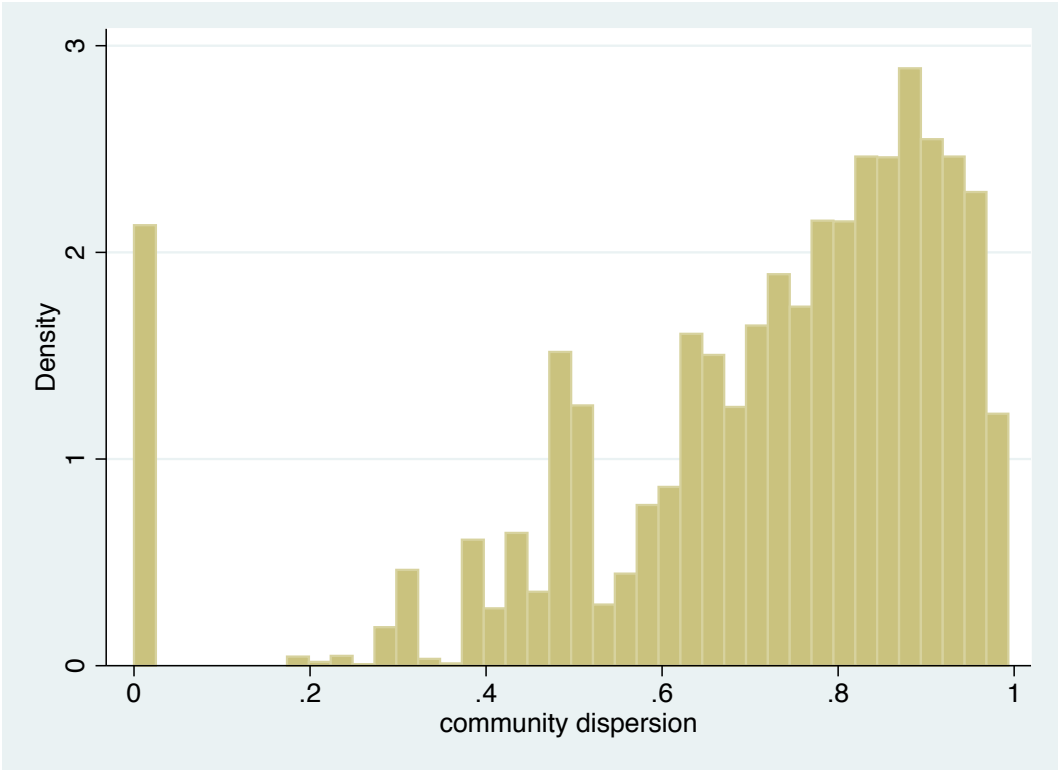

Figure 2. Community dispersion distribution.

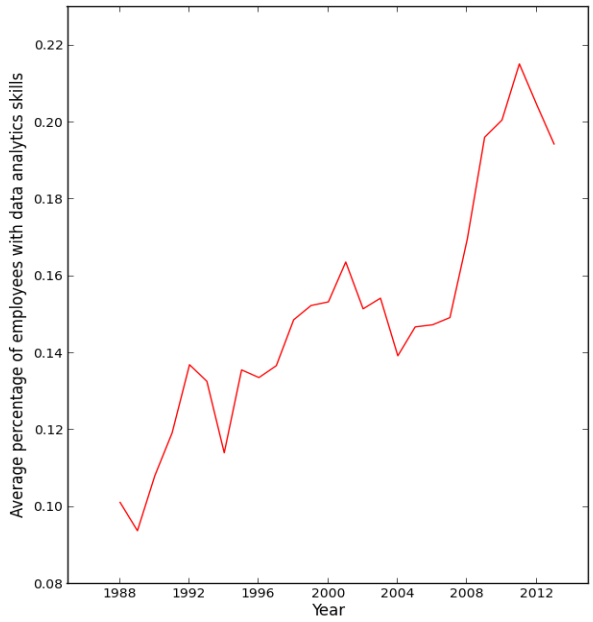

Figure 3: Average data talent over time

References:

Akcigit U, Kerr WR, Nicholas T (2013). *The Mechanics of Endogenous Innovation and Growth: Evidence from Historical US Patents*. Technical report, Working Paper.

Amabile TM, Pillemer J (2012). Perspectives on the social psychology of creativity. *The Journal of Creative Behavior*. 46(1) 3-15.

Argyres NS, Silverman BS (2004). R&D, organization structure, and the development of corporate technological knowledge. *Strategic Management Journal*. 25(8-9) 929-958.

Arora A, Belenzon S, Rios LA (2014). Make, buy, organize: The interplay between research, external knowledge, and firm structure. *Strategic Management Journal*. 35(3) 317-337.

Audia PG, Goncalo JA (2007). Past success and creativity over time: A study of inventors in the hard disk drive industry. *Management Science*. 53(1) 1-15.

Autor DH, Levy F, Murnane RJ (2003). The Skill Content of Recent Technological Change: An Empirical Exploration*. *The Quarterly Journal of Economics*. 118(4) 1279-1333.

Avery J, Norton M (2014). Learning From Extreme Consumers.

Bharadwaj S, Bharadwaj A, Bendoly E (2007). The performance effects of complementarities between information systems, marketing, manufacturing, and supply chain processes. *Information systems research*. 18(4) 437-453.

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*. 2008(10) P10008.

Bloom N, Jones CI, Van Reenen J, Webb M (2017). *Are ideas getting harder to find?* National Bureau of Economic Research.

Bradstreet D (2017). *Analytics Accelerates Into the Mainstream*. Dun & Bradstreet.

Bresnahan TF, Brynjolfsson E, Hitt LM (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The Quarterly Journal of Economics*. 117(1) 339-376.

Brynjolfsson E, Hitt L (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management science*. 42(4) 541-558.

Brynjolfsson E, McElheran K (2016). The rapid adoption of data-driven decision-making. *American Economic Review*. 106(5) 133-139.
Chen Y, Elenee Argentinis JD, Weber G (2016). IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clinical Therapeutics*. 38(4) 688-701.
Cross R, Parker A, Prusak L, Borgatti SP (2001). Knowing what we know:: Supporting knowledge creation and sharing in social networks. *Organizational dynamics*. 30(2) 100-120.
Gao G, Hitt LM (2012). Information technology and trademarks: Implications for product variety. *Management Science*. 58(6) 1211-1226.
Griliches Z (1990). *Patent statistics as economic indicators: a survey*. National Bureau of Economic Research.
Hall BH, Jaffe AB, Trajtenberg M (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*. National Bureau of Economic Research.
Hannan MT, Freeman J (1984). Structural inertia and organizational change. *American sociological review* 149-164.
Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*. 47 98-115.
Henderson RM, Clark KB (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly* 9-30.
Joshi KD, Chi L, Datta A, Han S (2010). Changing the competitive landscape: Continuous innovation through IT-enabled knowledge capabilities. *Information Systems Research*. 21(3) 472-495.
Kleis L, Chwelos P, Ramirez RV, Cockburn I (2012). Information technology and intangible output: The impact of IT investment on innovation productivity. *Information Systems Research*. 23(1) 42-59.
Lam A (2005). *Organizational innovation*. Oxford University Press.
Lin C, Liu S, Manso G (2016). Shareholder litigation and corporate innovation. *University of Hong Kong and University of California at Berkeley Working Paper*.
Majchrzak A, Malhotra A (2016). Effect of knowledge-sharing trajectories on innovative outcomes in temporary online crowds. *Information Systems Research*. 27(4) 685-703.
McAfee A, Brynjolfsson E. (2017). *Machine, platform, crowd: Harnessing our digital future*. WW Norton & Company.
McElheran K, Brynjolfsson E (2015). Data in Action: Data-Driven Decision Making in US Manufacturing.
Milgrom P, Roberts J (1990). The economics of modern manufacturing: Technology, strategy, and organization. *The American Economic Review* 511-528.
Nagle F (2017). Open Source Software and Firm Productivity. *Management Science*. Forthcoming.
Pang M-S (2016). Politics and Information Technology Investments in the US Federal Government in 2003–2016. *Information Systems Research*. 28(1) 33-45.
Siggelkow N, Rivkin JW (2006). When exploration backfires: Unintended consequences of multilevel organizational search. *Academy of Management Journal*. 49(4) 779-795.
Smalley E (2017). *AI-powered drug discovery captures pharma interest*. Nature Publishing Group.
Tafti A, Mithas S, Krishnan MS (2013). The effect of information technology-enabled flexibility on formation and market value of alliances. *Management Science*. 59(1) 207-225.
Tambe P (2014). Big data investment, skills, and firm value. *Management Science*. 60(6) 1452-1469.
Von Hippel E (1994). "Sticky information" and the locus of problem solving: implications for innovation. *Management science*. 40(4) 429-439 %@ 0025-1909.
Wu L, Hitt LM, Lou B (2017). Data Analytics Skills, Innovation and Firm Productivity.
Zafar SY, Peppercorn JM, Schrag D, Taylor DH, Goetzinger AM, Zhong X, Abernethy AP (2013). The financial toxicity of cancer treatment: a pilot study assessing out-of-pocket expenses and the insured cancer patient's experience. *The oncologist*. 18(4) 381-390.

Appendix

A. Patent Data and Networks

Our measures of organization of innovation are based on network analysis of patent co-authorship relationships. We begin by constructing a list of patents attributable to each firm from 1988 to 2013. We use the latest version of NBER Patent Citation Data File (Hall et al. 2001), which links patents to publicly-traded firms through 2006. To capture the recent change from the growth of data analytics from years 2007 to 2013, we obtain the original patent data from the USPTO and use the same methods as Hall et al. (2001) to match patents to firms. These methods correct for matching related issues, such as misspellings, mergers and acquisitions, or patents assigned to subsidiaries. Following the convention in the R&D literature (Griliches et al. 1986, Hall et al. 2005, Saunders 2011), we use the application year (as opposed to issue year) because this is likely closer to the time the technology is developed and deployed.

Using the authors listed for each patent, we construct an intra-firm patent network for each firm annually between 1988 and 2013, using a sliding window of 5 years.[1] A node in this network represents an employee inventor and an edge (link) indicates the presence of one or more coauthored patents between two inventors in the same firm. The sliding window approach allows the patent network to be more accurately reflected than one that uses only a single year of patents (Wu 2013), although our results using one year of patents for each network do not qualitatively change. We then apply machine learning-based community detection algorithms to these networks to identify distinct innovation communities within each firm: a community of inventors is identified when they work closely with each other but have limited collaboration outside the community. Our primary analyses rely on the Multilevel community detection algorithm to identify clusters of innovative activity (Blondel et al. 2008)[2]. As suggested by Yang et al.

[1] Each intra-firm network for a particular year includes all inventors listed on the patents that were applied in the current year, plus two years prior and two years after the current year. Thus for the 2013 year, we use patents applied in year 2014 and 2015 as well.

[2] We use a greedy clustering algorithm that begins by treating every node (inventor) as a separate community and then moves each node to adjacent communities to see if it improves the modularity score: a high score indicates dense connections within communities but sparse connections across communities, while a low score indicates the communities are not clearly delineated. The first stage of the algorithm terminates when the movement of a single node no longer leads to a material improvement in modularity. For the next stage, the nodes in each community are

(2016), we confirm our main results using other popular community detection algorithms (Fastgreedy, Leading eigenvector, and Infomap), all of which generated results (both in the metrics and the results of analysis of these metrics) similar to the Multilevel algorithm. These results suggest limited potential biases arising from the algorithm choice (Table A1).

Table A1. The Productivity Test: Using Other Community Detection Algorithms

DV: ln(Sales)	(1)	(2)	(3)	(4)
Model	FE	FE	FE	FE
ln(Materials)	0.641***	0.642***	0.642***	0.642***
	(0.0223)	(0.0224)	(0.0224)	(0.0224)
ln(Capital)	0.121***	0.124***	0.125***	0.124***
	(0.0208)	(0.0205)	(0.0205)	(0.0205)
ln(IT labor)	0.0191***	0.0193***	0.0195***	0.0195***
	(0.00552)	(0.00554)	(0.00555)	(0.00556)
ln(Other labor)	0.212***	0.212***	0.212***	0.212***
	(0.0251)	(0.0252)	(0.0251)	(0.0251)
ln(Emp w/ college+)	0.00627**	0.00618**	0.00615**	0.00615**
	(0.00251)	(0.00255)	(0.00256)	(0.00256)
ln(Data)		-2.44e-05	-0.000321	-0.000288
		(0.00136)	(0.00135)	(0.00135)
ln(Dispersion: infomap)		-0.00336		
		(0.0510)		
Data X Dispersion: infomap		0.0149***		
		(0.00436)		
ln(Dispersion: leading eigenvector)			-0.0470	
			(0.0481)	
Data X Dispersion: leading eigenvector			0.0212***	
			(0.00392)	
ln(Dispersion: fastgreedy)				-0.0391
				(0.0493)
Data X Dispersion: fastgreedy				0.0207***
				(0.00396)
Observations	14,960	14,960	14,960	14,960
R-squared	0.985	0.985	0.985	0.985
# of firms	1,864	1,864	1,864	1,864
Industry	YES	YES	YES	YES
Year	YES	YES	YES	YES

Note:
(1) All interactions are appropriately centered.
(2) Data source dummy for measuring data analytics skills, and patent variables including firm's patent stock and

aggregated into a single node which will represent a community as opposed to a single inventor. The link between two communities in this new network is the sum of all links between inventors in one community to another (before the aggregation occurred). The greedy algorithm is applied again by assigning nodes to adjacent communities until modularity cannot be improved anymore. The final set of clusters identified is the distinct communities identified in the co-authorship network.

number of inventors are also controlled.
(3) Robust (clustered by firm) standard errors are reported in parentheses.
(4) *** p<0.01, ** p<0.05, * p<0.1

Using the communities identified, we create a dispersion metric which is one minus the Herfindahl concentration index of the fraction of inventors in each innovation community. The value is always between 0 and 1. A low value indicates most inventors are concentrated in a few large communities, and a high value indicates that inventors are dispersed across many different communities. The higher the value on this metric, the more decentralized is the structure of innovation.

Our measure of decentralization differs from previous measurements in two important ways. Earlier approaches measure the degree of decentralization through formal hierarchies (e.g., organizational charts) within the firm or formal firm boundaries between the parent firm and the affiliates (Arora et al. 2014). This approach is difficult to extend to large scale analyses because internal R&D and innovation structures are extremely difficult to observe limiting much of this work to small sample studies (see e.g., Argyres and Silverman 2004; Lerner and Wulf 2007) and divisional-level analysis may be too coarse for large firms and may not be applicable at all to firms that do not operate as multiple legal entities.

Our approach has advantages over other metrics based on organizational relationships because it is entirely data driven – the organization is identified by the interaction of inventors rather than (potentially) idiosyncratically-defined organizational units. Moreover, our metric is not subject to the criticism that the informal and formal structures of organizations are different, which has been an important observation in the knowledge management literature (Cross et al. 2002, Cross et al. 2004, Cross et al. 2001). Finally, to obtain a valid patent, all inventors must be disclosed on the application, which further ensures that we have likely identified the relevant inventors accurately.

While there are many advantages of using networks and community detection algorithms to measure decentralization, it is still a relatively new method. A main drawback is that there lacks “ground-truth” of the communities underlying the network which makes model test and comparison difficult. Recent comparative analysis has used simulated graphs (Fortunato 2010; Leskovec et al. 2010; Yang et al. 2016)

to evaluate the model performance based on properties of networks that are observed. We rely on these recommendations to determine what algorithms to use based on the graph characteristics[3]. To ensure that our measure is not idiosyncratic to a single measurement, we also implemented four popular community detection algorisms and the results from these algorithms are very similar, as shown in Table 2 in the main body of the paper and Table A1. However, field in community detection algorithms is rapidly advancing and future work should incorporate new methods especially if they can consider more detailed information about the nodes and multiplex tie relationships.

We graph the distribution of the community dispersion metrics using the Multilevel method (Figure 2 in the paper). The chart shows a mass at zero (maximally centralized). This is likely due to a limited number of inventors and patents in some firms. To ensure that our results are not driven by firms with a single cluster, we include a dummy variable for the mass at zero and its interaction with analytics skills in our analysis. As shown in Table A2, including the dummy and the interaction does not change our main result.

Table A2. The Productivity Test: Controlling for Firms with a Single Community

DV: ln(Sales)	(1)	(2)	(3)
Model	FE	FE	FE
ln(Materials)	0.641***	0.642***	0.642***
	(0.0223)	(0.0224)	(0.0224)
ln(Capital)	0.121***	0.124***	0.125***
	(0.0208)	(0.0205)	(0.0204)
ln(IT labor)	0.0191***	0.0194***	0.0197***
	(0.00552)	(0.00556)	(0.00556)
ln(Other labor)	0.212***	0.212***	0.212***
	(0.0251)	(0.0251)	(0.0251)
ln(Emp w/ college+)	0.00627**	0.00611**	0.00611**
	(0.00251)	(0.00255)	(0.00256)

[3] We rely on results from general comparison in these studies that test models on benchmark networks, which are synthetic graphs constructed by a priori standard of grouping nodes. The test framework is first proposed in (Girvan et al. 2002), called GN benchmark, where the artificial random network has 128 nodes, partitioned into four equivalently sized communities with 32 nodes each. The nodes in each group have approximately the same degree, and the average degree of the network is 16. When the expected number of links joining each node to those nodes in different groups is less than 8, each node has more links with the other nodes in its own group than with the rest of the network, and thus four groups are well defined communities. LFR benchmark suggested in (Lancichinetti et al. 2008) is more scalable and introduces heterogeneity in the distributions of node degrees and of community sizes to extend GN benchmark. Further research could examine the accuracy of each partition if the detailed hierarchical structure of network is available (Blondel et al. 2008).

ln(Data)		0.00203	-5.22e-05
		(0.00139)	(0.00145)
Dispersion==0		0.0448**	0.0258
		(0.0227)	(0.0356)
Data X Dispersion==0		-0.0668***	-0.0118
		(0.0135)	(0.0208)
ln(Dispersion)			-0.0395
			(0.0796)
Data X Dispersion			0.0200***
			(0.00611)
Observations	14,960	14,960	14,960
R-squared	0.985	0.985	0.985
# of firms	1,864	1,864	1,864
Industry	YES	YES	YES
Year	YES	YES	YES

Note:
(1) The Multilevel algorithm is used to measure the dispersion of innovation communities. All interactions are appropriately centered. *Dispersion==0* is a binary variable that indicates whether dispersion score is 0.
(2) Data source dummy for measuring data analytics skills, and patent variables including firm's patent stock and number of inventors are also controlled.
(3) Robust (clustered by firm) standard errors are reported in parentheses.
(4) *** p<0.01, ** p<0.05, * p<0.1

B: Innovative Output

Our measures of innovative output are also based on patent data consistent with prior work on R&D productivity (Griliches 1990, Griliches et al. 1986, Hall et al. 2001). Like these prior studies we exploited the fact that these data are public, consistent (defined by law and subject to a largely stable regulatory environment), and rich, including information on inventors and their affiliations, detailed records of citations that can be used to measure impact (e.g., citation count), general area of technology described in the patent (patent class and subclass) that can be used to identify the extent to which the innovation is the result of combining existing technologies or creating new ones, and an extensive textual description of the innovation that can be used to rank the level of novelty. We recognize that we cannot measure unpatented innovation and thus the associated effects of data analytics and decentralization on such innovations (Nagle 2014, Saunders and Brynjolfsson 2013). However, R&D expenditure, which can approximate a firm's overall effort on innovation, is often found to be positively correlated with patents (Kleis et al. 2012). Thus,

the patent-based metrics could potentially underestimate the effect of data analytics and innovation structure on innovation output[4].

To measure recombination, we begin with a measure created by Hall et al. (2001) [also described as "technological diversity" (Kaplan and Vakili 2015)] that is a Herfindahl concentration index, calculated based on the distribution of different three-digit main classes of prior art that are cited by the focal patent. The exact formulation is $1 - \sum_j^{n_i} s_{i,j}^2$, where $s_{i,j}$ represents the percentage of citations that patent *i* cited that belong to patent class *j*, out of n_i different patent classes. Patents that cite a wider variety of classes will have a measure closer to 1, whereas patents citing a single technology class will have a score of 0. Thus, the technology diversity metric represents the level of recombination because a patent that cites many technology classes is more likely to represent a synthesis of existing technologies than a patent that cites few or a single area of prior art. To account for potential bias in this metric for patents with small numbers of citations (that is, patents with fewer citations necessarily cite fewer different categories of prior art) we adjust this measure by citations as proposed by Hall et al, 2001, pp 44-46):

$$Recombination = \left(\frac{Cites}{Cites-1}\right)\left(1 - \sum_j^{n_i} s_{i,j}^2\right). \quad (1)$$

We can aggregate this measure at the firm level by using a weighted average on citations. For our analysis, it is important to distinguish reuse of existing combinations from the creation of new combinations (Akcigit et al. 2013). Our approach is based on the idea that reuse is defined as recombination of technologies that have been previously used; more specifically, if the firm creates a patent with technology areas that were combined in the past, we consider that patent reuse. We also classify the patent with only one single technology area as the reuse type if this technology area has been used in other prior patents. If a firm combines technology classes for a patent in a new way, we consider this to be a new combination of existing technology classes. When a patent creates a new patent

[4] Nagle (2014) and Saunders and Brynjolfsson (2013) have detailed discussions on valuing unpatented innovation. While our work use patent output to approximate the overall innovation output, future work should examine these effects for innovations that are not patentable as well as other intangibles that are difficult to measure.

technology class that did not exist before, we consider it to be new technology. To obtain a firm level measure, we aggregate all the patents (m) at the firm level in a single year and count how many patents are reuse or new recombination or new technology.

$$reuse = \sum_{j=1}^{m} 1(existing\ technology\ or\ combination_j) \quad (2)$$

$$new\ recombination = \sum_{j=1}^{m} 1(new\ combination_j) \quad (3)$$

$$\text{new } technology = \sum_{j=1}^{m} 1(new\ technology_j) \quad (4)$$

This distinction for each type of innovation can be made at the firm level (e.g. a new combination for the firm, even if other firms have done similar combinations) or on a global level (e.g. a combination is new only if no other firm has ever combined that set of patent classes to create a new patent). This distinction between firm-specific prior knowledge (local) and global (all available knowledge) is useful because it enables discrimination between innovations that bring in external knowledge versus innovations based on existing firm knowledge. Thus, we have a total of 6 ways to classifying innovation along two dimensions: (1) new technology, reuse and new combination, and (2) local vs. global.

Lastly, we experiment with another alternative method to measure novelty. We note that patents contain a considerable amount of free-form text data describing the innovation, which can be used to represent the underlying inventive ideas. Prior research on patent innovation, including in information system research, used bibliometric methods and citation based analysis to study innovation impact and evolution (Hall et al. 2001, Kleis et al. 2012, Saunders 2011). However, the use of raw text from the patent has been less frequent with a few notable exceptions (Azoulay et al. 2007, Kaplan and Vakili 2015, Upham et al. 2010, Wu 2013). Our measure is based on the idea that novel ideas can often be detected through vocabulary shifts. In our case, we examine the appearance of new words or topics in the patent abstract, which provides a summary of the innovation. We apply a "bag of words" model to identify patent terms and then calculate the age of each word by when it first appears in the prior art. Each (non-stop) word in the patent is assigned an age. A word that has not appeared previously is assigned an age of zero. If the

word has appeared in another patent before, each word in each patent is assigned an age that is the difference between the application date of the focal patent and the time that word first appeared in any patent. The overall novelty score is the average age of the words in a patent:

$$\text{Novelty} = \frac{1}{N}\sum_{w=1}^{N}\frac{1}{\text{Age}_w+1}. \qquad (5)$$

The novelty of a firm's patent portfolio for a particular year is equal to weighted average of the novelty score of firm's own patents.[5]

We chose to keyword-based metrics because it is likely to be more sensitive to shift in vocabulary than topic modeling methods such as Latent Dirichlet Allocation (LDA), which relies on identifying a stable set of topics for classification (Blei et al. 2003, Shaparenko and Joachims 2009); in fact, the shifting vocabulary likely makes it difficult to find a stable ranking of topics, which is essential for accurate classification of LDA (Greene et al. 2014). In addition, the large topic space (essentially the language of all scientific endeavor over 20 years) would make it difficult to establish a reasonable topic space without expert input, which defeats the benefits of automated classification. In contrast, our keyword-based analyses are relatively simple, and have long been established and extended to detect novelty in sentences and documents (Voorhees 2003). They do not require auxiliary assumptions about the number of topics or degree of association between words needed to define a topic. As a check, we also ran an LDA model to identify topics in patents and measured the novelty of each topic in patents according to when it was first detected. The LDA results are directionally consistent. Using the vocabulary-based metric to measure novel technologies produced qualitative similar results to the novelty metric based on the technology class.

C: Data Analytics

We use two data sources—resumes and job reviews—to measure a firm's data analytics capabilities. First, we use resume data of employees at the firm to calculate the total number of employees possessing the relevant data analytics skills. The total universe of resumes used across all of the firms in this study is six million, collected in 2007. Since a resume provides a time stamp of employment, this provides a panel

[5] We also used the sum of the novelty score and the results are very similar to using the average.

data structure with measures from 1988 through 2007. Prior work suggests that resumes are representative of employment generally due to their large size and the fact that many employees in information-intensive occupations engage in passive job seeking – posting their resumes publicly, making them available to potential employers even when they are not actively searching for a job – and thus reduce the concerns of selection bias (see further discussion of the advantages and disadvantages of resume data in Tambe and Hitt, 2014).[6] To capture the level of investment in data analytics skills for more recent time periods (2008-2013), we also gathered data on employee job titles (a subset of what is available in the resume data) from a large online job-posting site where employees can post employer reviews.

To classify data analytics skills from the resume data, we apply natural language processing techniques on free-form text and job titles to identify the skills of each employee from 1988 to 2007. We include direct matches on keywords like "analytics" as well as inferred matches on phrases such as "regression modeling," "Hadoop," or "Machine Learning." Table A3 shows the list of words used. For each employee in our data set, we identify whether a data-related skill is present in his or her resume and approximate the timing of when the skill was acquired from the job history in the resume. We then measure firm-level data analytics capabilities by aggregating all employees with data analytics skills for each firm in each year and adjust it by a sampling rate[7] (Tambe and Hitt 2013). This approach assumes that the observed employees are sampled from the underlying population of all workers. We also assume that firm and occupation-specific factors with respect to the likelihood of posting career trajectory or job reviews are uncorrelated.[8] Thus, our firm-level sample rate could be estimated by $\theta_j' = \frac{x_j}{L_j}$, where x_j is the number of employees in all occupations in our sample for firm j, and L_j is the total number of employees at firm j obtained from Compustat.

[6] Our use of resumes is similar to other studies that have used resumes to study network relationships among individuals and firms or used them to identify the presence of particular skills within a firm. For instance, Tambe and Hitt, (2013b) used such an approach to measure information technology investment by counting the number of IT employees in the firm and Tambe (2014) used resume data to analyze the diffusion of specific IT skills, such as Hadoop, across firms and locations.

[7] The sampling rate is calculated annually for each firm.

[8] To the extent such matching varies systematically across firms, we can address this problem by using fixed effect models to alleviate the bias from time-invariant firm characteristics.

Table A3: Taxonomy used to classify data analytics skills	
General:	Business intelligence, Data analysis, Data center, Data driven, Data fusion, Data integration, Data mining, Data warehouse
System oriented:	Cassandra, Cloud computing, Distributed system, Hadoop, HBase, Google File System, MapReduce
Algorithms & Methodology oriented:	A/B test, Ensemble learning, Genetic algorithm, Machine learning, Natural language processing, Neural network, Network analysis, Optimization, Pattern recognition, Predictive model, Regression, Signal processing, Simulation, Supervised learning, Unsupervised learning, Visualization

Note: The Sovren resume parser (http://www.sovren.com) provides us with the text mining and skills-based semantic matching algorithms that map free-form text responses to a taxonomy of skills. It is also used by a number of online job sites to facilitate skills-based resume searches (among other activities). We then classify data analytics skills by the taxonomy.

To extend the data analytics metric from 2008 to 2013,[9] we use a second data source containing information from a large online job review platform that has 3.7 million job reviews. Each review has information on job title, employer name and length of employment. While these data are not as detailed as the resume data, they do provide job titles and employer names which can then be used to identify skills available at a firm at any point in time. Specifically, we use the job title classification from O*Net to identify analytics related positions, similar to how IT labor is distinguished from other employees in earlier work (Tambe and Hitt 2013, Tambe and Hitt 2013). If any of the skills listed under the job title is related to data analytics, we assume anyone with such a job title to possess data analytics skills. We aggregate these individual level skills for each firm-year observation, tracing back to the year when the employment started and adjusted the raw aggregation by the underlying sampling rate as we did with the resume data. We complement our title-based classification by identifying three relevant phrases about data analytics ("data," "statistic," and "business analytics"). The productivity test results by job-title based measure of data

[9] We search the word "data" in the Occupation Search box provided by https://www.onetonline.org/, and obtain a list of job titles ranked according to how well these job titles matched to the keyword "data," according to the matching process in https://www.onetonline.org/help/online/search#keyword. A total of 576 primary data-related job titles are found, with detailed reports on the requisite technology skills and general skills. An additional 788 alternate occupation titles are found from the primary job titles. For example, the occupation "Database Administrator" has several alternate job titles, such as data engineer, data miner and data center manager. We count a person as having data analytics skills if any of the skills in that job title is related to data.

analytics skills are consistent as shown in Table A4. Column 3 and 4 suggest that the complementarity is driven by employees with data analytics skills rather than IT skills.

Table A4. The Productivity Test: Using Title-based Data Analytics Skills

DV: ln(Sales)	(1)	(2)	(3)	(4)
Model	FE	FE	FE	FE
ln(Materials)	0.641***	0.642***	0.642***	0.642***
	(0.0223)	(0.0225)	(0.0224)	(0.0224)
ln(Capital)	0.121***	0.124***	0.124***	0.124***
	(0.0208)	(0.0205)	(0.0204)	(0.0204)
ln(IT labor)	0.0191***	0.0194***	0.0189***	0.0192***
	(0.00552)	(0.00560)	(0.00573)	(0.00573)
ln(Other labor)	0.212***	0.213***	0.212***	0.212***
	(0.0251)	(0.0251)	(0.0249)	(0.0248)
ln(Emp w/ college+)	0.00627**	0.00637**	0.00640**	0.00638**
	(0.00251)	(0.00250)	(0.00251)	(0.00251)
ln(Data)		-0.000756	-0.000108	-0.000716
		(0.00151)	(0.00150)	(0.00152)
ln(Dispersion)		-0.0662	-0.0629	-0.0657
		(0.0501)	(0.0498)	(0.0501)
Data X Dispersion		0.00665**		0.00635*
		(0.00307)		(0.00338)
IT labor X Dispersion			0.00301	0.00150
			(0.00580)	(0.00606)
Observations	14,960	14,960	14,960	14,960
R-squared	0.985	0.985	0.985	0.985
# of firms	1,864	1,864	1,864	1,864
Industry	YES	YES	YES	YES
Year	YES	YES	YES	YES

Note:
(1) The Multilevel algorithm is used to measure the dispersion of innovation communities. All interactions are appropriately centered.
(2) Data source dummy for measuring data analytics skills, and patent variables including firm's patent stock and number of inventors are also controlled.
(3) Robust (clustered by firm) standard errors are reported in parentheses.
(4) *** p<0.01, ** p<0.05, * p<0.1

We also appropriately scale the two different data sources to ensure that the metrics are comparable between the sources. This involves scaling the job-title-based classification used in job reviews to the full-text-based classifications used in resumes. We first translate job title-based metrics from job reviews to job title-based metrics in resumes. Then we translate this title-based metrics from resumes to the corresponding full-text-based metrics. When a firm does not exist in the matching data, we use the industry average of scaling ratios to obtain and approximate the appropriate scale. Thus, our data cover the critical period for

the rising demand in data analytics ranging across 26 years from 1988 to 2013. Although we combine two different data sources to measure data analytics skills, the data source and year dummies help to control for any idiosyncratic characteristics associated with each data source that may bias our results. Furthermore, we use the logarithm of employees with data analytics skills which should alleviate the potential scaling differences in the underlying data. We also perform a Chow test on our main results that cannot reject the hypothesis that the analytics coefficients using the combined data are equal[10] and thus we can reasonably combine the two data sources. To ensure our results are not driven by the merging of two different data sources, we estimate our models in each data source separately and the results are consistent with the combined sample and observe the complementarities to grow stronger in later years using the job reviews (Column 5 and 6 in Table 3 in the main body of the paper).

D: Method and Identification

Our primary analysis tests for complementarities between the deployment of data analytics and innovation structure. Two types of statistical tests have been developed to assess the existence of such complementarities (Arora and Gambardella 1994, Brynjolfsson and Milgrom 2013, Milgrom and Roberts 1990): correlations (adoption or demand equations) and performance differences (productivity equations). The correlation test, which can be estimated in the form of conditional correlations or reduced-form demand equations, determines whether a cluster of practices is more likely to be adopted jointly rather than separately. Although the demand equations have the advantage that they are relatively simple and provide the greatest power if firms are optimally matching complementary practices, they have the disadvantage that the simplicity makes them vulnerable to unobserved heterogeneity. In addition, such an analysis will tend to understate the strength of complements if not all firms are endowed with or able to change to the optimal match between complements. An alternative approach to examining complementarities is to

[10] For instance, the Chow test for pooling our data in our key regression (see Column 2 of Table 3 in the main body of the paper) on the complementarities between data analytics and dispersion of innovation community is $F(2, 1863) = 1.91$, $p<0.15$).

measure whether there are performance differences between firms that adopt complementary practices as a group and those that do not. These metrics have the advantage of a direct tie to a relevant firm outcome (performance), and the effects are most powerful statistically if not all firms have found the optimal match, which may be likely for a relatively new change in business practices. Over time, as the complementarities system diffuses to other firms, the correlation would increase but the productivity premium would decrease because the relative advantage for using the system diminishes as its adoption spreads.

We take the view that firm capabilities to pursue specific innovative practices are quasi-fixed because it is difficult to change them due to organizational inertia, path dependency, culture, norms, and prior strategic choices (Hannan and Freeman 1984, Lam 2005). The rapid diffusion of analytics technology and a relatively recent, exogenous shift in the amount of data available to firms enable us to observe the effects of data analytics on different firm practices related to innovation, even those that are not optimally configured to benefit from analytics capabilities. A similar assumption has been made previously in IT research for other types of IT-related complements (Autor et al. 2003, Bresnahan et al. 2002, Brynjolfsson and Hitt 1996, Milgrom and Roberts 1990).

We first explore the correlation between data analytics and the dispersion of innovation communities. According to our theory, we expect the two to be positively correlated. Thus, we relate data analytics capabilities to innovation dispersion after controlling for sales, innovation inputs, and firm (fixed effects) (γ_i) and year (y_t). This can be viewed as a reduced-form demand equation to the extent that the "price" of data analytics skills can be viewed as varying by firm and time, and we are treating innovation structure as quasi-fixed relative to data analytics skills.[11]

$$ln(data)_{it} = \beta_0 + \beta_1\, ln(sales)_{it} + \beta_2\, ln(R\&D)_{it} + \beta_3 patent_{it} + \beta_4 ln(dispersion)_{it} + y_t + \gamma_i + \epsilon_{it} \quad (6)$$

The productivity test examines whether the hypothesized system of complements is more productive if adopted together or separately; it is typically implemented by including the direct effects as well as the

[11] A typical factor demand equation relates the quantity of a factor to the prices of that factor and the levels of quasi-fixed factors. Sales can also appear in this equation if the demand system is assumed to arise from cost minimization (since this model treats sales as exogenous). In our specification, sales also serve as a control for firm heterogeneity in size.

interaction effect in a productivity or performance model. A complementarities relationship is consistent with a significant interaction term but can also be estimated by sample comparisons that would also imply increasing benefits of one factor in the presence of increasing levels of another complement (see e.g., Brynjolfsson and Milgrom (2013)), a technique that has been applied empirically in a number of recent papers (see e.g., (Aral et al. 2012, Tambe et al. 2012)).

Although there are several approaches to measuring the marginal effect of information technology on firm performance, we choose the multi-factor productivity framework implemented with the Cobb-Douglas production function, which is the most commonly used framework in estimating IT value. Using this framework, we relate firm output such as sales or value-added to various firm inputs such as labor, capital and materials after controlling for temporal and firm-specific variations. The residual of this equation can be interpreted as the multi-factor productivity, and we are interested in exploring whether data analytics skills and their interaction with innovation structure are related to this measure of productivity.

Equation 7 below represents a standard function that has been used in the IT-productivity literature augmented with the additional terms for data analytics skills and innovation (and their interactions). Specifically, we relate sales to production inputs such as materials expense (m), physical capital stock (k) and labor (*IT labor and other labor*, measured as the number of employees with IT skills and without IT or data analytics skills) as well as controls for firm fixed effects (i) and year (t) dummy variables that address firm-specific unobserved heterogeneity, temporal productivity shocks and measurement error in price deflators. We also control for educational level of employees and firm's innovation activity including patent stock and number of inventors to resolve time-variant unobserved heterogeneity in labor quality and innovation in general. Here and throughout we will use italics to designate specific regression measures. Our panel analysis includes the performance and labor metrics on a 26-year panel from 1988 to 2013. To this base specification, we add data analytics skills (*data)*, the innovation structure of the firm (*dispersion*) as well as their interactions (*data X dispersion)*. Of particular interest in this equation is β_7, since a positive value of this coefficient suggests the presence of complementarities.

$$\ln(sales)_{it} = \beta_0 + \beta_1 \ln(m)_{it} + \beta_2 \ln(k)_{it} + \beta_3 \ln(other\ labor)_{it} + \beta_4 \ln(IT\ labor)_{it} +$$

$$\beta_5 \ln(data)_{it} + \beta_6 \ln(dispersion)_{it} + \beta_7 data_{it} X\ dispersion_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (7)$$

Next, we explore the nature of innovation as enabled by these complementarities. We first investigate whether the interaction of data analytics skills and innovation structure could help produce a new technology element that can serve as a building block for future combination in innovation. We then examine the extent to which data analytics can facilitate the generation of a new combination innovation or reuse of existing combination and technology.

$$\ln(new\ technology)_{it} = \beta_0 + \beta_1 \ln(data)_{it} + \beta_2 dispersion_{it} + \beta_3 data_{it} X\ dispersion_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (8)$$

$$\ln(new\ combination)_{it} = \beta_0 + \beta_1 \ln(data)_{it} + \beta_2 dispersion_{it} + \beta_3 data_{it} X\ dispersion_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (9)$$

$$\ln(reuse)_{it} = \beta_0 + \beta_1 \ln(data)_{it} + \beta_2 dispersion_{it} + \beta_3 data_{it} X\ dispersion_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (10)$$

Complementarity relies on the benefits of matching two or more investment decisions rather than the effectiveness of the decisions themselves. As a result, it is naturally robust for some types of endogeneity and reverse causality problems. While there are many explanations for why a particular investment might be correlated with unobserved shocks (e.g. firms spend unexpected free cash flow on new investment), it is difficult to explain why such a relationship would be present only when two investments are made together or not at all. In addition, the correlation test and the productivity test for complementarities are subject to different sources of bias which provide some ability to cross-check results. Nonetheless, we estimate that there are at least two endogenous factors that could lead to a positive bias. First, high-performing firms can have resources available that can be redeployed to increase investment in data talent acquisition, which could lead to a reverse causality between performance and data analytics. By itself this is not a problem in interpreting our complementarities result, but it could introduce bias if this tendency depends on the organizational form. (This is not implausible given that human capital practices and organizational practices are often closely linked). Second, there could be omitted variables, such as higher management quality

attracting higher quality labor, which could lead to an upward bias. We address this problem in two ways. First, selection bias from omitted variables is less likely in the complementarities framework if we find the use of data analytics is only complementary to some firm practices but not others. Second, we treat data analytics skills and IT skills as endogenous and use three sets of instruments to correct potential biases: 1) data analytics skills in a firm's existing hiring network; 2) a neighbor's adoption of an enterprise resource planning (ERP) system and the total number of neighbors that started to use ERP systems; 3) a neighbor's adoption event and the total number of neighbors that started to use Human Capital Management (HCM) systems. All these instruments are derived from a firm's immediate hiring network, which can be used as a proxy to measure the ease of access to existing data analytics talent and therefore the "cost" of acquiring analytics skills. This hiring network is constructed by examining employee flows across firms using techniques similar to those in Wu et al. (2016).

Firms are represented by nodes and firm-to-firm employee movements are represented by edges in the network. A directed edge is identified between Firm A and Firm B if A hired someone from B in the previous five years[12] and the weight of that edge is expressed as the number of employees hired. The first instrumental variable is a Hausman-type instrumental variable, based on the neighboring firm's existing data talent. It measures the potential labor pool accessible to the firm through employee mobility, which reflects the cost of acquiring additional analytics talent. At the same time, the data analytics skills pool from the neighboring firms is only the potential access a firm could have, and thus it should not affect the firm's own performance. The only point at which the talent pool could affect performance of the focal firm is when the firm hires from a neighboring firm, which the firm fixed-effect model can address because it can capture the effect from the change in data analytics skills between years. Accordingly, the potential for productivity spillovers through labor movement would not invalidate our instruments.

The second and third instrumental variables are derived from measures of the adoption of ERP systems broadly and the HCM module by firms in the same labor pool. Because these types of systems are likely

[12] This type of graph has been previously utilized to identify the position of a firm in the network of labor flows (Wu et al. 2016).

complementary to the use of data analytics talent,[13] this can change the overall access to data analytics skill as firms acquire employees with these skills during implementation and employees transfer these skills through job-hopping. The ERP adoption event may have an especially large effect of increasing the availability of employees because of a transition from an implementation team to operations may free up some types of staff. We focus specifically on the point at which a firm implements these systems (rather than simply licenses, which could be several years earlier) to prevent common industry-level shocks from potentially generating an increased demand for these systems in nearby firms as well as a direct shock to productivity in the focal firm (see related arguments in (Hitt and Brynjolfsson 1996, Hitt et al. 2002)). Furthermore, our instrumental variables derived from hiring neighbors are inherently more protected against common industry shocks because a firm's hiring network is composed of a much more diverse set of firms than firms in the same industry. Thus, while a neighboring firm's adoption of ERP should not directly affect the focal firm's productivity, it should have an impact on the focal firm's access to data analytics-related talent. To illustrate this, we show in Figure A1 the impact on the access to data talent of the focal firm (Firm A) when its neighbor (Firm B), from which A hires people, adopted an ERP system. B now has a need to hire employees with ERP experience as well as experience working with the associated process change that often accompanies the ERP implementation, and starts to hire from new firms, such as C. Not only has B's hiring pool expanded, at the same time, this network change also increases to A's access to data analytics-related talent. As shown in Figure A1, A still hired from the same three firms but its access to data talent has expanded because of B's newly hired data talent that it acquired from C. This example shows that B's ERP implementation induces a change in A's access to talent even though A has not changed its hiring practices. Thus, we can use a neighbor's ERP adoption event as an instrument for the focal firm's access to data analytics talent. The first stage results are shown in Column 5 of Table 2 in the main body of the paper.

[13] ERP systems are off-the-shelf software packages that offer a variety of functions including finance, human resources, supply-chain management, consumer relationship management and business intelligence (Hitt and Brynjolfsson 1996; Hitt et al. 2002). These types of enterprise information systems often produce a large amount of data about firms' business processes, which can affect the demand for data analytics skills.

We have two available measures of ERP adoption. First, we can infer ERP adoption by examining whether employees in a firm have job titles or job descriptions in their resumes that suggest the use of ERP (e.g., employees stating in their past job description that they "used an ERP system to conduct inventory control"). We can infer the adoption of HCM systems using data from a major ERP vendor that provided the dates when clients purchased and implemented the HCM module. This is used to construct a binary variable to represent the firm's HCM module implementation status, which is 0 in the years before the firm started using the system and 1 afterward.

Innovation structure could also be endogenous. Here, we rely on the assumption that organizational structure is often quasi-fixed, an assumption which has been used in prior work on IT-organizational complementarities (Bresnahan 2001, Brynjolfsson and Hitt 1996, Levy and Murnane 1996, Milgrom and Roberts 1990). The quasi-fixity of our innovation structure measures is likely to be correct given the literature that has documented the difficulty firms face in changing their innovation practices and portfolios, because changing them requires firms to substantially alter practices in many areas including finance, talent acquisition, monitoring policy, incentive pay and reporting relationships (Nagji and Tuff 2012, Wessel and Christensen 2012). We verify that the quasi-fixed assumption by examining the extent to which the structure of the innovation process varies across time within the same firm. Overall, the standard deviations of the dispersion metric within firms are very small, suggesting that it is difficult for firms to change their innovation structure. Thus, our regression results can be interpreted as assessing whether pre-existing firm differences in innovation structure influence the benefits of acquiring data analytics capabilities. Even though it is quasi-fixed, there is some temporal variation in this metric across so the coefficient of the direct term can be identified, but this variable should be interpreted with caution due to the likelihood of a low signal to noise ratio in our fixed effects regressions on the decentralization variable alone; this concern does not apply to the interaction term which is the focus of our analysis. We also fixed the firm's dispersion score by taking the average of the yearly scores for each firm. In this regression, the dispersion variable is dropped out of the regression in the fixed-effect model. As shown in Table A5, results are consistent with the time varying measure of firm dispersion.

Table A5. The Productivity Test: Time Invariant Innovation Community Structure

DV: ln(Sales)	(1)	(2)	(3)	(4)
Model	FE	FE	2SLS	GMM/IV
ln(Materials)	0.641***	0.642***	0.582***	0.594***
	(0.0223)	(0.0224)	(0.0244)	(0.0397)
ln(Capital)	0.121***	0.124***	0.0819***	0.132***
	(0.0208)	(0.0205)	(0.0216)	(0.0414)
ln(IT labor)	0.0191***	0.0192***	0.0456	0.00896
	(0.00552)	(0.00556)	(0.0549)	(0.00887)
ln(Other labor)	0.212***	0.212***	0.260***	0.310***
	(0.0251)	(0.0252)	(0.0280)	(0.0545)
ln(Emp w/ college+)	0.00627**	0.00609**	-0.00332	0.00871
	(0.00251)	(0.00255)	(0.00654)	(0.0120)
ln(Data)		0.000471	0.0322	0.00148
		(0.00133)	(0.0424)	(0.00222)
Data X Dispersion		0.0153***	0.0824**	0.0165**
		(0.00459)	(0.0419)	(0.00746)
Observations	14,960	14,960	11,012	11,012
R-squared	0.985	0.985	0.733	
# of firms	1,864	1,864	1,592	1,592
Industry	YES	YES	YES	YES
Year	YES	YES	YES	YES

Note:
(1) The Multilevel algorithm is used to measure the dispersion of innovation communities. All interactions are appropriately centered.
(2) Six instrumental variables are used: 1) total number of employees with data analytics skills in neighboring firms; 2) one-period lagged values of 1); 3-4) total number of neighbors that adopt an enterprise resource planning (ERP) Human Capital Management (HCM) systems; 5-6) one-period lagged values for ERP or HCM variables in what are calculated for 3-4). The first-stage F-statistics passes the weak instrument test.
(3) We use the Blundell and Bond (1998) SYS-GMM estimator which derives additional instruments using the lagged level and differences from production inputs. This method was originally developed specifically for micro-productivity applications, especially for measuring the productivity of R&D. We use three-period lags of the endogenous variables as instruments in addition to the external instruments used in 2SLS. The Arellano-Bond test for AR(2) autocorrelation in first differences suggests that serial correlation is not a concern in the first-differencing equations. Neither the Hansen test of over-identification nor the difference-in-Hansen test of the system GMM instruments rejects the null that the instruments are uncorrelated with the error term, ensuring the validity of the instruments used in the GMM estimation. We also test other lags and find they don't qualitatively change our results.
(4) Data source dummy for measuring data analytics skills, and patent variables including firm's patent stock and number of inventors are also controlled.
(5) Robust (clustered by firm) standard errors are reported in parentheses.
(6) *** p<0.01, ** p<0.05, * p<0.1

Finally, we address general endogeneity of production inputs with respect to output using standard panel-based instrumental variables methods (the SYS-GMM estimator of Blundell and Bond (1998)), that

were developed to specifically address problems of endogeneity in micro-level productivity analysis. The combination of instrumental variables, SYS-GMM, and labor quality controls should substantially reduce the potential for bias in our estimates.

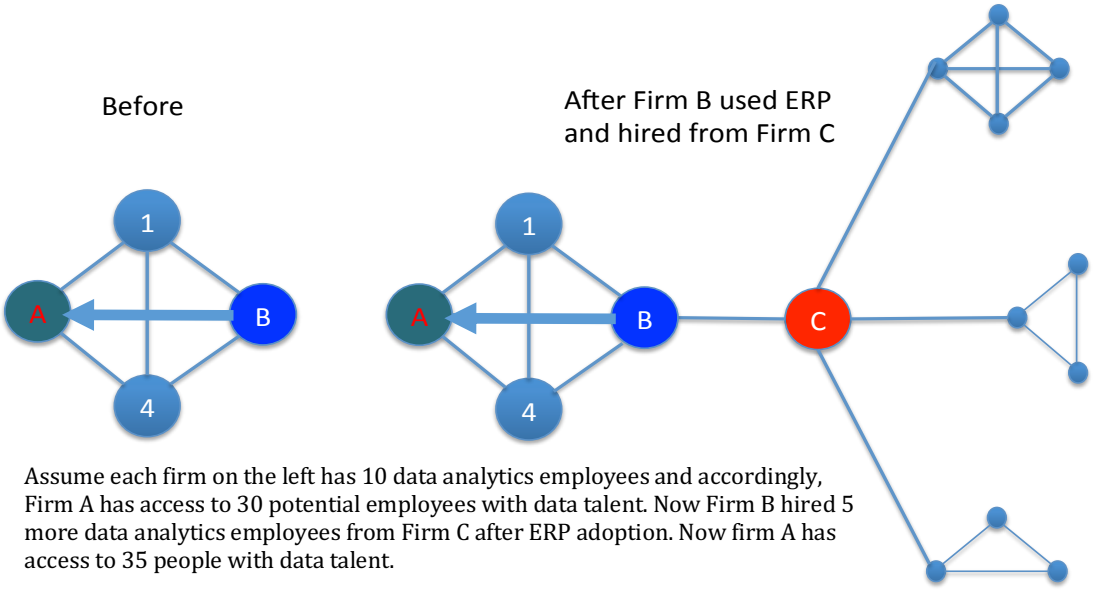

Figure A1: Instrumental Variables