Data Analytics Supports Decentralized Innovation

Abstract

Modern analytics technology can accelerate the inventive process by increasing the rate at which existing knowledge can be identified, accessed, combined and deployed to 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 how the benefits of analytics on 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 analytics can have a direct influence on innovative output, especially those that are especially likely to be influenced by analytics. Furthermore, organizational structures, such as decentralization, have a substantial influence on the ability to firms to capture innovation benefits from analytics, primarily because analytics can facilitate the linking of distant ideas coming from different decentralized groups.

Introduction

Innovation is critical to the growth of advanced economics. Technology, especially information technology, has always been closely linked to innovation as both an enabler and an output 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 stock of existing knowledge (Joshi et al. 2010, Majchrzak and Malhotra 2016). Analytics capabilities, advanced by artificial intelligence and enabled by digitization, have substantially improved the ability to detect hidden patterns in large-scale data. By accessing a wide range of knowledge, both within and external to a firm, 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 having the necessary data (Bradstreet 2017). One possible explanation is that not all firms have strategies or other organizational structures that are complementary 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 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 is likely to hold true for investments in analytics capabilities, although the complements critical to leveraging analytics appear to differ from those for leveraging general IT.

Prior work has also suggested that IT has 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 enhances or

inhibits these relationships. As a first step, we focus specifically on one important form of investment in analytics through hiring employees with data skills (Tambe 2014, Wu et al. 2017) and one structural characteristic of firms, the choice of centralization versus decentralization, that is known to have important influence on the nature and outcome of innovation (Argyres and Silverman 2004, Siggelkow and Rivkin 2006). Our goal is to provide empirical support for what we believe is the most plausible theoretical 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 is not inconsistent with presence of additional complementary practices we are unable to observe.[1] As analytics 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 can 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 firm. In such a structure where information is often hard to transfer across organizational boundaries (Von Hippel 1994), individual groups directly involved in firm operations may be better at understanding the nature of operational problems and at identifying and implementing solutions. By contrast, at Apple, much of the innovation output is centralized in a few tightly-

[1] Any set of complementary practices can be studied by examining pairwise relationships between any two practices (see e.g., (Brynjolfsson and Milgrom 2013, Milgrom and Roberts 1990)).

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

linked clusters whose connections outside their groups 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. Accordingly, 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 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. Apple is known for the creation of novel, breakthrough products, while Google has been most successful in generating a steady stream of improvements to their search and targeted advertising technology. 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 through firms with different ex-ante innovation structures, which change much more slowly than analytics, provides the opportunity to identify the relationship between analytics and innovation.

We hypothesize that, although analytics may provide benefits to all firms in their innovation processes, decentralized structures, such as those found at Google are complementary to analytics capability. Therefore, firms that combine analytics with decentralized innovation are likely to receive greater benefits than firms that combine analytics with centralized innovation. 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 the likelihood of knowledge sharing outside each group and limiting opportunities to link innovative ideas across areas (Henderson and Clark 1990). Analytics can mitigate this disadvantage

to some extent by automating the search of patterns in innovations from communities outside of the immediate group, and thereby facilitate the transfer and use of knowledge across organizational boundaries.

This effort is accelerated by recent advances in data analytics such as deep learning and machine learning that have become increasingly effective in 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), a feat that can potentially generate 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, BenevolentAI, a British AI firm was able identify five potential hypotheses for the treatment of ALS (Amyotrophic lateral sclerosis) in less than a week, one of which was proven to prevent the death of motor neurons (Smalley 2017), overcoming a critical difficulty in treating ALS that has no known cure. Like the IBM Watson example, BenevolentAI found the 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 existing 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, put cheap sensors to collect vast amount of data on how the chassis respond to stresses, strains, temperature, displacement and all other factors that might affect performance. Using the sensor data as well as chassis design knowledge from a variety of sources, analytics helped to create a substantially different chassis designs from traditional ones. For example, unlike the traditional designs, the new chassis is asymmetric, which mirrors the fact that a race car turn more often in one direction than the other and thus subject the chassis to different forces. Designing a deeply asymmetric chassis would not have been possible without data analytics to link newly captured sensor data and traditional chassis

principles together and to uncover hidden patterns within the data (McAfee and Brynjolfsson 2017). As decentralized innovation structures often face difficulty in sharing information from different departments and recombining the information into new inventions, analytics can be particularly helpful for decentralized innovation structure to integrate and uncover hidden common patterns in data across many sources. Thus, we expect that 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 novel innovation (Arora et al. 2014, Siggelkow and Rivkin 2006), which has less to gain from the current state of analytics technology since there is often little or no available data for novel or breakthrough products because they do not have existing producers or customers prior to their initial deployment. Despite these critical knowledge sharing constraints in novel innovation, it has been noted that much of the relevant information is tacit and shared through informal communications channels (Avery and Norton 2014), thus making it less amenable to digitization or automated analysis. Firms who 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 no (or even perhaps a slightly negative) 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 innovations communities are dispersed for the firm (Appendix A). Our metric has the advantages that it measures the actual structure of innovation regardless of the formal hierarchy (e.g. org chart), is entirely data driven (requiring no arbitrary definitions of organizational boundaries), and 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 they can be consistently measured[4] and there is a large and robust literature on patent-related measurement approaches. We also measure novelty in patent: whether an innovation involves the creation of a new technology class (or a subclass in patent classification).This can be measured at a firm (technology class new to the firm) or global level (a new subclass of patent that no one else created). For our analysis, it is important to distinguish reuse of existing combinations from the creation of new combinations (Akcigit et al. 2013). We measure this difference by segmenting innovations into existing combinations and new combinations. These can also be measured at both the firm level (combination is new to the firm, but other firms have done similar combinations) or at a global level (an original combination that no one else has

[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.

[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.

created). Thus, we have a total of 6 ways to classifying innovation along two dimensions: (1) new technology, reuse and new recombination, and (2) local (firm) versus global (see Appendix B).

Analytics. For the purpose of our analysis we define analytics to be 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 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 a relevant firm outcome (performance), and the effects are most powerful statistically when 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 of using a complementary system diminishes as its adoption spreads.

For econometric identification, we take the view that data skills are rapidly changing (and potentially endogenous) while organization of innovation is quasi-fixed and (Hannan and Freeman 1984, Lam 2005)

and therefore exogeneous. 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 so we standardize on the Multilevel algorithm for our main results)

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 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 innovation (data analytics skills, number of inventors, and patent stock) (Columns 2-4). In general, we find data analytics is positively associated with productivity while the direct effect of patents and inventors is indistinguishable from zero. Our key result (Column 3) shows that while having decentralized innovation structures is negatively correlated with productivity (albeit not significantly so; it is perhaps due to the impact of fixed effects with a slow-changing quasi-fixed factor), the interaction with data analytics is positive. On average, a firm that is more decentralized than the average (one standard deviation above in our decentralization metric) receives a 2.2% increase in productivity for every standard deviation increase in data skills above the mean.

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 data analytics 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 skills to be influenced by prior performance (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 add additional instruments derived from information within the panel. The results on the key coefficients continue to be consistent, suggesting that potential endogeneity of data analytics is not leading to an upward bias of our complementarities measure.

To explore the complementarities between different types of innovation, we use the same variables from our previous models (data analytics, decentralization, their interactions, and other controls) to determine their relationship with the production of different types of patents (see Table 4). First, we find that none of our metrics have 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 novel 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 ideas. First, we do not distinguish new combinations from refinements of existing combinations, and we find that the combination of data analytics and decentralized innovation facilitate recombination in innovation (Column 3). After distinguishing new combinations from the reuse of existing combinations, we find that our data-decentralization complementarity facilitates both combinations that are new to the firm (Column 5) and those new to the overall market (Column 4). On average, the interaction term implies that a one standard deviation increase in data analytics increase the number of new recombination patents by 3.96% in firms that are one standard-deviation above the mean in innovation dispersion. Given the average new recombination patents in a firm is about 100, a 3.96% increase is about 4 additional patents per year. 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 particular benefit of data analytics (direct or complementary) to this type of innovation (Columns 6 & 7).

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 ideas into a firm's patent portfolio. Following Lin et. al. (2016), we use the portion of citations to a firms' own prior art in their new patents, which captures the amount of external information 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 innovations based on in-house innovation, and the complementarity strengthens as the proportion of external information increases (60% internal/40% external). However, the complementarity begins to drop and even turns negative (although not significant) (see Columns 6-10) although there is a significant difference in the coefficient between the 60% cutoff and the 10% cutoff ($p<.05$). 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 knowledge, 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 skills of named inventors. While named inventors are only a subset of all employees with data skills and matching limits the number that we can identify, we can still conduct an exploratory analysis where we divide analytics skills between inventors and other employees who also have data skills. 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 higher level of data skills in innovators (Table 6). The productivity test results are also similar. Furthermore, the interaction to data skills for non-inventors show 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, complementarities effect is stronger for employees who provide infrastructure and extensive support to the inventors than the inventors who have analytics skills themselves. Replication of our other results using this inventor-based measure are similar in magnitude but do not show consistently significant results, likely due to 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 disadvantages of decentralized structures. However, data analytics provide limited help in improving completely novel innovation or known combinations of existing technologies. This finding also appears reasonable because these situations are where 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 to alternative measurement of innovation, decentralization, and whether we consider data 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, especially from the advances in artificial intelligence and new ways of gathering data (e.g. using robotics to capture optimal machine toolpaths from 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 or quickly adjust to acquire such capabilities would be likely to benefit substantially from using data analytics.

Our results cover two potentially important complements 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 these two complements. Future work could productively identify additional complements (specific technologies, additional organizational practices) or extend the investigation to 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, and help explain and perhaps mitigate the recent observed decline in innovative output that has been observed in advanced economies (see e.g., Bloom et al. (2017)).

Table 1. Summary Statistics

VARIABLES	(1) # obs.	(2) Mean	(3) Std. dev.	(4) Min	(5) Max
ln(Sales ($ million))	15,001	5.590	2.279	-6.830	13.00
ln(Materials ($ million))	15,001	5.245	2.025	-2.315	12.74
ln(Capital ($ million))	15,001	4.918	2.338	-1.427	12.41
ln(IT labor)	15,001	3.849	1.687	0	11.68
ln(Other labor)	15,001	6.810	2.406	0	14.60
ln(Dispersion)	15,001	0.514	0.163	0	0.690
ln(Data)	15,001	3.706	3.517	0	12.23
ln(R&D ($ million))	14,761	2.559	1.958	-4.542	9.513
ln(Recombination)	15,001	0.323	0.118	0	0.693
ln(New-tech Global)	15,001	0.0197	0.144	0	2.862
ln(New-combination Global)	15,001	1.429	1.364	0	7.514
ln(Reuse Global)	15,001	0.952	1.172	0	6.804
ln(New-tech Local)	15,001	0.926	0.951	0	5.150
ln(New-combination Local)	15,001	1.202	1.373	0	7.535
ln(Reuse Local)	15,001	0.736	1.089	0	6.615

Correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. ln(Sales)	1														
2. ln(Materials)	0.946	1													
3. ln(Capital)	0.900	0.910	1												
4. ln(IT labor)	0.429	0.441	0.436	1											
5. ln(Other labor)	0.815	0.783	0.796	0.609	1										
6. ln(Dispersion)	0.0923	0.107	0.104	0.0226	0.0727	1									
7. ln(Data)	0.434	0.435	0.420	0.352	0.477	0.0369	1								
8. ln(R&D)	0.375	0.455	0.399	0.259	0.307	0.213	0.260	1							
9. ln(Recombination)	-0.0617	-0.0451	-0.0689	-0.0393	-0.0653	0.164	0.0537	0.113	1						
10. ln(New-tech Global)	0.170	0.184	0.174	0.125	0.157	0.00862	0.144	0.218	0.0323	1					
11. ln(New-combination Global)	0.401	0.438	0.426	0.264	0.370	0.284	0.285	0.544	0.202	0.331	1				
12. ln(Reuse Global)	0.398	0.429	0.407	0.248	0.330	0.238	0.259	0.564	0.115	0.314	0.837	1			
13. ln(New-tech Local)	0.365	0.403	0.393	0.245	0.346	0.261	0.253	0.463	0.195	0.305	0.907	0.744	1		
14. ln(New-combination Local)	0.407	0.441	0.425	0.261	0.363	0.258	0.287	0.567	0.159	0.340	0.954	0.886	0.793	1	
15. ln(Reuse Local)	0.405	0.431	0.412	0.259	0.337	0.197	0.263	0.547	0.0849	0.328	0.814	0.963	0.700	0.847	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)	0.405***	0.404***	0.403***	0.403***	0.460***
	(0.0601)	(0.0602)	(0.0600)	(0.0601)	(0.0749)
std(Patents)	0.158	0.152	0.159	0.158	0.120
	(0.139)	(0.141)	(0.138)	(0.138)	(0.140)
ln(R&D)	0.193***	0.196***	0.192***	0.193***	0.194***
	(0.0543)	(0.0543)	(0.0543)	(0.0543)	(0.0575)
ln(Dispersion: multilevel)	0.863**				0.823**
	(0.405)				(0.407)
ln(Dispersion: infomap)		0.565*			
		(0.338)			
ln(Dispersion: leading eigenvector)			0.963**		
			(0.393)		
ln(Dispersion: fastgreedy)				0.909**	
				(0.397)	
ln(Total Neighbor Data)					-0.0396
					(0.218)
ln(Total Neighbor ERP)					0.987**
					(0.439)
ln(Total Neighbor HCM)					0.185
					(0.188)
Lagged ln(Total Neighbor Data)					0.206
					(0.216)
Lagged ln(Total Neighbor ERP)					-0.542
					(0.362)
Lagged ln(Total Neighbor HCM)					-0.146
					(0.160)
Observations	14,761	14,761	14,761	14,761	11,049
R-squared	0.657	0.657	0.657	0.657	0.685
# of firms	1,856	1,856	1,856	1,856	1,594
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 4 of Table 3.
(3) Data source 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 3. The Productivity Test: Data Analytics Skills and Innovation Community Structure

DV: ln(Sales)	(1)	(2)	(3)	(4)	(5)
Model	FE	FE	FE	2SLS	GMM/IV
ln(Materials)	0.642***	0.642***	0.643***	0.588***	0.552***
	(0.0224)	(0.0224)	(0.0226)	(0.0248)	(0.0559)
ln(Capital)	0.122***	0.122***	0.125***	0.0808***	0.209***
	(0.0209)	(0.0209)	(0.0206)	(0.0219)	(0.0578)
ln(IT labor)	0.0190***	0.0190***	0.0198***	0.0451	0.00312
	(0.00557)	(0.00555)	(0.00562)	(0.0569)	(0.0315)
ln(Other labor)	0.213***	0.213***	0.213***	0.261***	0.258***
	(0.0251)	(0.0251)	(0.0251)	(0.0280)	(0.0694)
ln(Emp w/ college+)	0.00596**	0.00596**	0.00599**	-0.00331	0.0401
	(0.00256)	(0.00261)	(0.00262)	(0.00660)	(0.0249)
ln(Data)		2.16e-05	-0.00147	0.0320	-0.00460
		(0.00136)	(0.00135)	(0.0446)	(0.00516)
ln(Dispersion)			-0.0744	-0.0636	-0.0463
			(0.0506)	(0.0687)	(0.155)
Data X Dispersion			0.0223***	0.0273**	0.0255*
			(0.00400)	(0.0134)	(0.0136)
Observations	15,001	15,001	15,001	11,049	11,049
R-squared	0.984	0.984	0.984	0.732	
# of firms	1,864	1,864	1,864	1,594	1,594
Industry	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES

Note:

(1) The Multilevel algorithm is used to measure the dispersion of innovation communities. *Data X Dispersion* has been standardized on the interaction between *ln(Data)* and *ln(Dispersion)*. We also replicate Column 3 of Table 3 with other measurements of decentralization. They show similar results for the estimation on interaction term (0.0211 for leading eigenvector, 0.0148 for infomap, 0.0207 for fastgreedy). 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 ($F(12, 9405) = 139.62$) 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=0.324). Neither the Hansen test of over-identification (p= 0.373) nor the difference-in-Hansen test of the system GMM instruments (p= 0.676) 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 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)	-9.24e-05	-0.00132	-0.00251	-0.00351	-0.00106	0.00180
	(0.000397)	(0.00326)	(0.00393)	(0.00380)	(0.00352)	(0.00331)
ln(Dispersion)	0.0124	0.938***	1.284***	0.944***	0.749***	0.466***
	(0.0104)	(0.0900)	(0.123)	(0.116)	(0.0966)	(0.0881)
Data X Dispersion	-0.000184	0.0287**	0.0396***	0.0339**	0.0108	0.0158
	(0.00180)	(0.0112)	(0.0143)	(0.0137)	(0.0120)	(0.0112)
ln(Sales)	0.000452	0.0358*	0.0497**	0.0421*	0.0513**	0.0575***
	(0.00214)	(0.0198)	(0.0240)	(0.0224)	(0.0215)	(0.0195)
ln(Employees)	0.00853**	0.180***	0.258***	0.266***	0.241***	0.199***
	(0.00419)	(0.0304)	(0.0406)	(0.0384)	(0.0339)	(0.0313)
ln(Emp w/ college+)	0.000552	0.00600	0.00667	0.00141	0.00195	0.000729
	(0.000494)	(0.00566)	(0.00728)	(0.00681)	(0.00593)	(0.00563)
Observations	15,001	15,001	15,001	15,001	15,001	15,001
R-squared	0.473	0.695	0.811	0.829	0.792	0.797
# 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. *Data X Dispersion* has been standardized on the interaction between *ln(Data)* and *ln(Dispersion)*.
(2) Data source 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. Innovation Quality Compared to Firm's Own Patent Stock from the previous 5 years

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
DV	100	Post 90	Post 80	Post 70	Post 60	Prior 40	Prior 30	Prior 20	Prior 10	0
Model	FE	FE	FE	FE	FE	FE	FE	FE	FE	FE
ln(Data)	0.000911	0.000735	-0.000530	-0.000805	-0.000375	-0.00296	-0.00219	-0.00126	-0.00185	-0.00148
	(0.00258)	(0.00263)	(0.00295)	(0.00319)	(0.00352)	(0.00390)	(0.00379)	(0.00358)	(0.00333)	(0.00299)
ln(Dispersion)	0.173**	0.179**	0.275***	0.377***	0.552***	1.215***	1.073***	0.873***	0.653***	0.499***
	(0.0853)	(0.0870)	(0.0961)	(0.0988)	(0.108)	(0.112)	(0.107)	(0.101)	(0.0869)	(0.0782)
Data X	0.0205**	0.0243***	0.0338***	0.0391***	0.0466***	0.0116	0.00364	-0.00107	-0.00767	-0.0111
Dispersion	(0.00872)	(0.00899)	(0.0103)	(0.0111)	(0.0123)	(0.0135)	(0.0132)	(0.0124)	(0.0112)	(0.00982)
ln(Sales)	0.0229	0.0190	0.0132	0.0171	0.0130	0.0643***	0.0496**	0.0432*	0.0339*	0.0365*
	(0.0148)	(0.0151)	(0.0168)	(0.0182)	(0.0203)	(0.0240)	(0.0238)	(0.0223)	(0.0201)	(0.0190)
ln(Employees)	0.116***	0.123***	0.151***	0.162***	0.190***	0.278***	0.270***	0.249***	0.210***	0.160***
	(0.0245)	(0.0255)	(0.0297)	(0.0320)	(0.0354)	(0.0379)	(0.0372)	(0.0341)	(0.0301)	(0.0285)
ln(Emp w/	-0.00497	-0.00511	-0.00484	-0.00549	-0.00514	0.00737	0.00641	0.00746	0.00588	0.00387
college+)	(0.00558)	(0.00567)	(0.00633)	(0.00666)	(0.00695)	(0.00628)	(0.00606)	(0.00606)	(0.00561)	(0.00538)
Observations	15,001	15,001	15,001	15,001	15,001	15,001	15,001	15,001	15,001	15,001
R-squared	0.774	0.773	0.775	0.779	0.786	0.802	0.798	0.795	0.785	0.781
# 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. *Data X Dispersion* has been standardized on the interaction between *ln(Data)* and *ln(Dispersion)*.
(2) Data source 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 6. The Correlation Test using Data Analytics Skills of Inventors

DV: ln(Data & Inventor)	(1)
Model	FE
ln(Dispersion)	0.279***
	(0.0909)
ln(Sales)	0.0302*
	(0.0173)
std(Patents)	0.146*
	(0.0846)
ln(R&D)	0.0402***
	(0.0139)
Observations	13,033
R-squared	0.387
# of firms	1,833
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)	0.601***	0.600***
	(0.0236)	(0.0236)
ln(Capital)	0.108***	0.110***
	(0.0195)	(0.0195)
ln(IT labor)	0.00856**	0.00849**
	(0.00410)	(0.00411)
ln(Other labor)	0.262***	0.261***
	(0.0185)	(0.0184)
ln(Emp w/ college+)	0.00342	0.00339
	(0.00248)	(0.00249)
ln(Data & Inventor)	-0.0225	-0.0229
	(0.0151)	(0.0151)
ln(Data & Non-Inventor)		-0.000900
		(0.00127)
ln(Dispersion)	0.00111	0.00190
	(0.0508)	(0.0504)
(Data & Inventor) X Dispersion	0.0181*	0.0178
	(0.0109)	(0.0109)
(Data & Non-Inventor) X Dispersion		0.0177***
		(0.00362)
Observations	13,033	13,033
R-squared	0.986	0.986
# of firms	1,833	1,833
Industry	YES	YES
Year	YES	YES

Note: The Multilevel algorithm is implemented to measure the dispersion of innovation community. Similar standardization has been performed for (*Data & Inventor) X Dispersion* and (*Data & Non-Inventor) X Dispersion* as for *Data X Dispersion*. 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

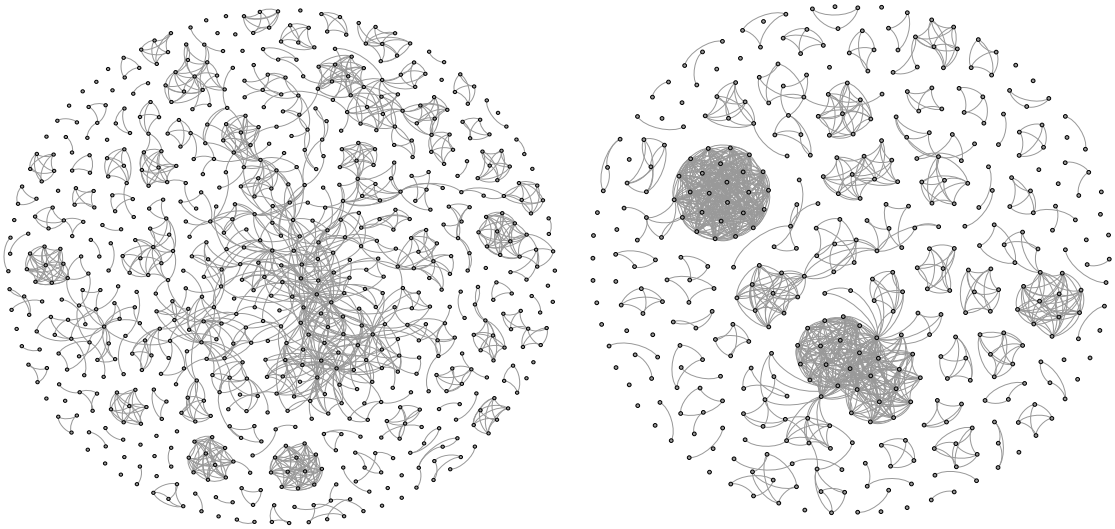

Figure 1. Comparing innovation structures between Google (left) and Apple (right). Each node on the graph is a particular inventor, while the edges are inventors appearing on the same patent. This graph is generated from the patent data in our sample.

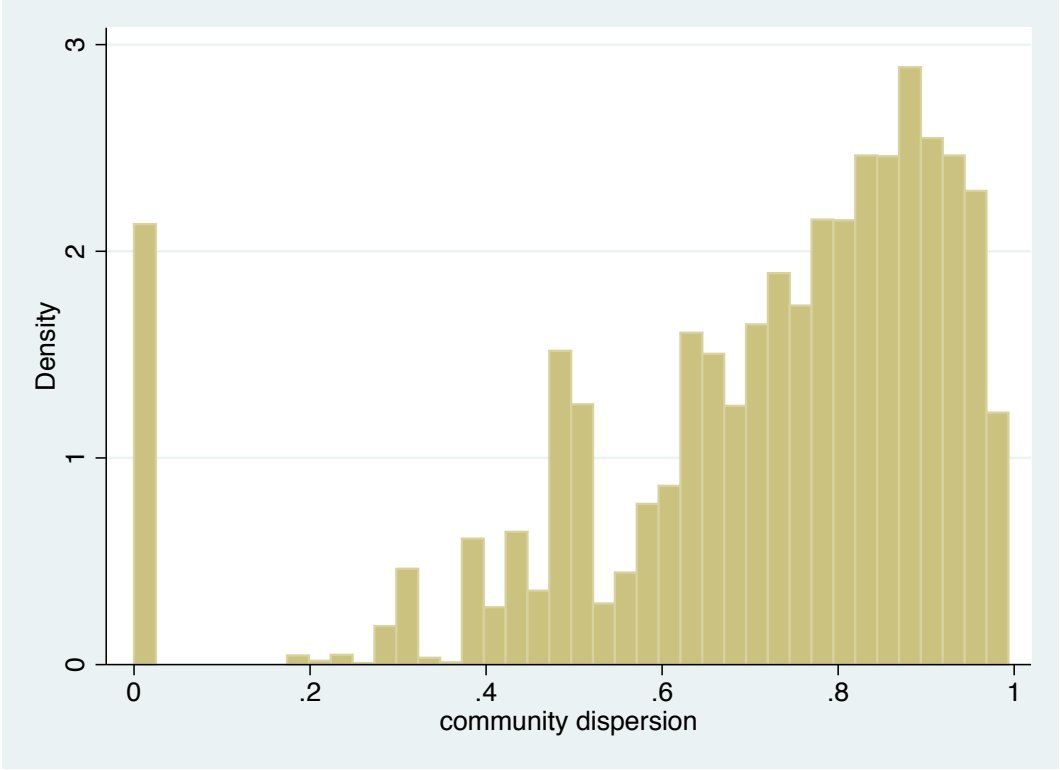

Figure 2. Community dispersion distribution.

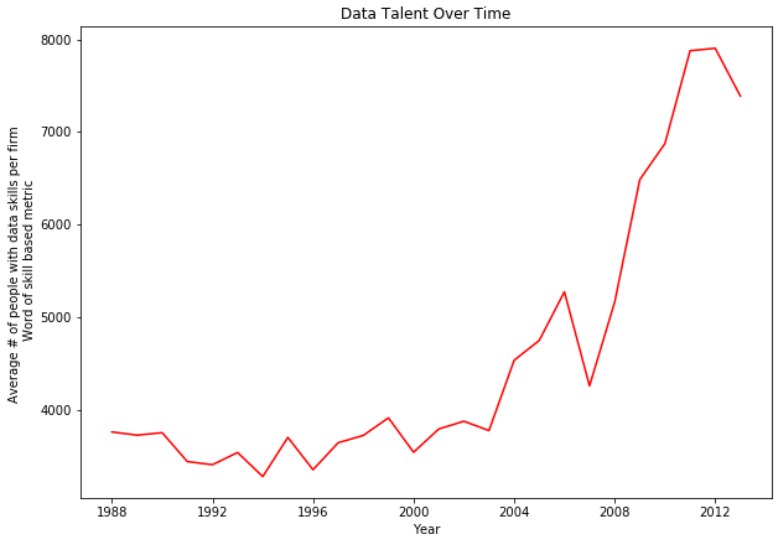

Figure 3: Average data talent over time

References:

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

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.

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.

Brynjolfsson E, Milgrom P (2013). Complementarity in organizations. *The handbook of organizational economics* 11-55.

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.
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.
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 L, 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.