BIG DATA INVESTMENT, SKILLS, AND FIRM VALUE

Prasanna Tambe
NYU Stern School of Business
ptambe@stern.nyu.edu

Forthcoming, *Management Science*

Last revised on January 4, 2014

Abstract

This paper considers how labor market factors have shaped early returns to investment in big data technologies. It tests the hypothesis that returns to early investments in Hadoop—a key big data infrastructure technology—have been concentrated in select labor markets due to the importance of aggregate corporate investment levels within a labor market for producing a supply of complementary technical skills during the early stages of technology diffusion. The analysis uses a new data source—the LinkedIn skills database—enabling direct measurement of firms' investments into emerging technical skills such as Hadoop, Map/Reduce, and Apache Pig. Productivity estimates indicate that from 2006 to 2011, firms' Hadoop investments were associated with 3% faster productivity growth, but only for firms a) with significant existing data assets and b) in labor networks characterized by significant aggregate Hadoop investment. Evidence for the importance of labor market concentration disappears for investments in mature data technologies, such as SQL-based databases, for which the skills are diffused and readily available through universities and other channels. These findings underscore the importance of geography, corporate investment, and channels for technical skill acquisition for explaining differences in productivity growth rates across labor markets during the spread of new IT innovations.

This research has been generously supported by the Alfred P. Sloan Foundation. I am indebted to Monica Rogati, Daniel Tunkelang, Pete Skomoroch and other members of the Data Science Team at LinkedIn for valuable assistance with data collection and analysis. I have also benefited greatly from assistance or comments from Erik Brynjolfsson, Avi Goldfarb, Lorin Hitt, Heekyung Kim, Frank Nagle, Andrea Meyer, Foster Provost, Adam Saunders, Wen Wen, and seminar participants at LinkedIn, New York University, University of Maryland, Michigan State University, University of British Columbia, Boston College, The Conference Board, Academy of Management, and the Workshop for Information Systems and Economics.

1.0 Introduction

US businesses appear to be on the cusp of a data-driven revolution in management. Firms capture enormous amounts of fine-grained data on social media activity, RFID tags, web browsing patterns, consumer sentiment, and mobile phone usage, and the analysis of these data promises to produce insights that will revolutionize managerial decision-making. Because this type of data analysis has, in many cases, outpaced firms' existing technological capabilities, there has been growing interest in the potential economic impact of investment in "big data" technologies, which enable data analysis at a scale exceeding the capabilities of existing database systems, and which many academic and industry observers argue will drive a new wave of innovation (McKinsey 2011; Brynjolfsson and McAfee 2011). Early big data adopters, however, face significant challenges. One in particular -- difficulties acquiring the technical skills required to support big data tools -- has attracted significant media attention (Rooney 2012 is an example).[1] A recent article describing Sears' implementation experiences with Hadoop, a technology that is central to the early wave of big data investment, captures the tradeoffs managers face: "Enter Hadoop, an open source data processing platform gaining adoption on the strength of two promises: ultra-high scalability and low cost compared with conventional relational databases … The downside of Hadoop is that it's an immature platform, perplexing to many IT shops, and Hadoop talent is scarce. Sears learned Hadoop the hard way, by trial and error. It had few outside experts available to guide its work when it embraced the platform in early 2010" (Henschen 2012). Amidst rapidly growing demand for big data technologies, Sears' experience reflects broader concerns that difficulties acquiring big data skills will limit the rate at which these technologies lead to productivity growth (McKinsey 2011).

These observations reflect a gap in the academic literature on IT-enabled growth. Prior research focuses on the effects of organizational factors in explaining variation in IT returns (Melville, Gurbaxani, and Kraemer 2004 provide a review). However, recent work finds evidence of systematic differences in growth rates across labor markets during large waves of investment in new IT innovations (Forman, Goldfarb, and Greenstein 2012), although most inputs required for implementing new IT practices are available at common prices throughout the US. The speed at which knowledge barriers fall, however, can impact the rate of diffusion of IT innovations and differ across labor markets (Attewell 1992; Fichman and Kemerer 1997). In particular, differences in the supply of workers with the skills complementary to the new information technologies, especially during the early diffusion period when there are few

[1] The use of big data technologies has been associated with the emergence of new technical skills such as Hadoop, Map/Reduce, Apache Pig, Hive, and HBase.

channels through which to acquire these skills, may explain differences in the rates at which firms in different labor markets are able to unlock value from new IT innovations.

This paper examines how labor markets have shaped early returns to investment in a key big data technology—Hadoop-based systems.[2] It tests the hypothesis that returns to Hadoop investments have been concentrated in select labor markets due to the importance of aggregate corporate investment within a labor market as a determinant of the early stock of technical human capital required to support firms' own Hadoop investments. Because technical know-how is embodied in IT labor, external corporate investment within the same labor market improves the skill content of the firm's own labor market by pooling demand for emerging skills and facilitating on-the-job learning for workers who can subsequently be hired. This argument is related to the literature on how external R&D investment impacts the success of a firm's own R&D efforts (Jaffe 1986; Cassiman and Veugelers 2006) and a related literature on the knowledge-based micro-foundations of agglomeration (Saxenian 1996; Porter and Stern 2001). As with R&D, firms that invest in new information technologies should derive significant benefits from the related investments of other firms while the complementary know-how is scarce. During this period, hiring employees from other early adopters may be an especially important channel through which to acquire technical expertise. As the technologies mature and alternative channels emerge through which workers can acquire the complementary skills (e.g. university degree programs), differences in labor market thickness, "spillovers" from the investments of nearby firms, and the performance advantages of being located in specific labor markets should decline, which is consistent with a literature on how the geographic concentration of production changes as spillovers weaken (Desmet and Rossi-Hansberg 2009). For data technologies, this leads to three testable predictions, a) that investment in emerging data technologies should be concentrated in select labor markets, b) that investments in these technologies should yield higher returns in these labor markets, and c) that the advantages of labor market concentration should disappear for investments in mature data technologies.

Testing these hypotheses requires data that can distinguish the investments firms make in emerging data technologies from investments in mature data technologies. The primary innovation in this paper is analysis of a new data source describing technical skills for a large fraction of the US-based IT workforce, collected from LinkedIn, a popular online professional network[3] on which participants post employers and occupations, professional technical skills such

[2] Section 2 provides a brief technological overview of big data technologies and Hadoop-based systems.
[3] See http://www.linkedin.com. These data were analyzed while the author was visiting LinkedIn.

as SQL, as well as emerging skills such as Hadoop, HBase, and Apache Pig. This data source is used to measure firms' investments in human capital complementary to specific technologies.

The construction of measures using data from online labor market intermediaries raises several concerns—such as those related to sampling—that merit a longer discussion and are explicitly addressed later in the analysis. On the other hand, the granularity of this data source provides advantages over measurement approaches used in the prior IT value literature. Earlier work measures the returns to specific technologies using data collected from firms or software vendors (e.g. Hitt, Wu, and Zhou 2002). However, the analysis of labor markets for skills complementary to specific IT innovations has been largely absent from the empirical literature. This is surprising because value from technological investment is determined in part by the supply of professionals who can translate technologies into business outcomes, and the economic importance of these professionals is reflected in wide-ranging policy discussions on the importance of IT labor supply for national competitiveness. Most existing IT workforce studies have been occupation-level analyses, but how supply adjusts to demand in markets for skills complementary to specific technologies, such as big data technologies, is likely to be important for understanding temporal and regional dynamics in the growth resulting from new IT innovations. Therefore, data on the fine-grained structure of skills within the IT labor force are well suited for understanding how labor markets impact returns to new technological innovations.

The empirical evidence is broadly supportive of the three hypotheses stated above. At the time of data collection, over 30% of workers with Hadoop skills were employed in Silicon Valley, compared with 4% of total US IT employment in that region. Mature technical skills were much less geographically concentrated. Estimates from short-run demand equations are consistent with complementarities between a firm's own Hadoop investments and the investments of other firms in its labor market. Direct complementarities tests indicate that firms' Hadoop investments yield higher returns in Hadoop-intensive labor markets. The most robust productivity estimates indicate that the output elasticity of firms' Hadoop investments is about 3%, and that these returns are principally captured by firms that are in data-intensive industries and are located in Hadoop intensive labor markets. On the other hand, the estimates indicate no measurable returns to Hadoop investments made outside of Hadoop intensive labor markets. By comparison, the evidence for labor market complementarities disappears for investments in mature data technologies, such as SQL-driven databases, for which the technical skills are widely available—the returns to investments in mature data technologies appear to be unaffected by the labor markets in which the investments are made. These findings are robust to several specifications as well as tests that place bounds on the effects of various sources of estimation bias as well as measurement error in the skills data.

These findings are closely related to several academic literatures. First, they contribute to an emerging literature on the value of modern data analytic technologies (Brynjolfsson, Hitt, and Kim 2011; Barua, Mani, and Mukherjee 2012). Empirical evidence of the benefits of the new data technologies has been primarily restricted to case evidence (discussed below). There is still an active debate about whether and under what conditions big data technologies have driven generalized economic gains (e.g. see Glanz 2013 and Harris 2013 for contrasting viewpoints). There is a need, therefore, for large-sample evidence of the impact of these investments on firm performance, as well as analysis of firm-level factors that can impact the magnitude of these returns. The few existing empirical studies on data analytics do not distinguish returns to emerging data technologies from returns to traditional database systems, at least in part due to measurement limitations. In fact, it may be uniquely difficult to assess the impact of new data technologies using archival data on hardware or software expenses due to the reliance of these technologies on open source software and commodity hardware.[4] Instead, investments in complementary human capital may command a larger share of expenditures for big data technologies than for earlier information technologies, and data on skills may therefore provide benefits for empirically distinguishing firms' investments in specific data technologies.

Second, this paper extends the broader IT value literature by providing evidence that labor market adjustments can explain why higher IT returns concentrate in select labor markets during the emergence of new IT innovations. Explaining firm-level variation in IT returns has been a topic of long-standing interest in the IT value literature (e.g., see Brynjolfsson and Hitt 2000), and recent empirical work demonstrates that IT returns are unevenly distributed across geographic regions (Dewan and Kraemer 2000; Bloom, Sadun, and Van Reenen 2012). By leveraging new data sources on the distribution of fine-grained technical skills among workers in different firms and labor markets, this paper provides evidence that the importance of labor market spillovers—a potentially important source of variation in IT returns across labor markets—vary according to the maturity of technical skills. In doing so, it connects an emerging literature on IT spillovers (Cheng and Nault 2007, 2011; Chang and Gurbaxani 2012; Tambe and Hitt, forthcoming) to a recent literature that documents geographic divisions in IT returns during periods of rapid IT innovation and argues that technological change in general—and the spread of big data technologies in particular—has the potential to foster "digital divides" and inequality

[4] See Greenstein and Nagle (2012) for a detailed discussion of the difficulties associated with measuring open source software use and value.

across regions (Freeland 2010; Dewan, Ganley, and Kraemer 2010; Forman, Goldfarb, and Greenstein 2012). Finally, by examining how labor networks change in importance over the lifecycle of IT innovations, the paper contributes to a literature on IT human resource management and firm performance (Agarwal and Ferratt 2001; Ang, Slaughter, and Ng 2002; Levina and Xin 2007; Bapna et al. 2013). Implications for managerial and policy decisions related to big data technology investment are discussed at the end of the paper.

2.0 Technology Background

The term "big data" is used to describe technologies enabling the collection, management, and analysis of datasets that are too large for conventional database systems (Dumbill 2012), and recent work discusses how big data tools are enabling new decision-making capabilities for firms (Provost and Fawcett 2013). To address the limitations of existing database systems, big data technologies use massively parallel computing approaches. Although distributed data processing has a long history (e.g., see Provost and Kolluri 1999 for a survey of the literature over a decade ago), the scale and rate of data collection in recent years has raised the returns to innovation in data processing technologies. The origins of the most recent wave of new data technologies can be traced to employees at Google who, in 2004, began using big data algorithms to support distributed processing. Apache Hadoop, the most widely used software platform for big data analytics, is derived from the Map/Reduce framework, implemented in the Java programming language, and freely distributed under an open source license. This open source project has a number of subprojects such as Cassandra, Pig, Hive, and HDFS, that handle different parts of the Hadoop cluster interface, communication, and processing flow. Big data infrastructure requires the implementation of this software and data environment on computer clusters. Because both the hardware and software required to support big data processing are readily available to firms, one of the primary expenses that firms face when implementing big data systems is the acquisition of expertise required to install, maintain, and facilitate these clusters to support data analysis.

Although there is debate about the potential economic impact of these technologies, some case-level evidence has begun to emerge from the business press about how the use of big data technologies generates value for specific firms in different industries. Big data technologies allow firms to extract business intelligence from petabyte-scale data in nearly real-time, a data processing task that requires managers using older data technologies to make compromises on either data size or processing time. For instance, Sears used Hadoop clusters to lower marketing analysis time for loyalty club members from six weeks to weekly, and even daily for online and mobile scenarios, while improving the granularity of its targeting (Henschen 2012). Orbitz Worldwide uses Hadoop to parse unstructured data on users' trip planning activities to develop portraits of user preferences that can be used for delivering personalized information (Schaal 2011). Netflix uses a Hadoop based infrastructure to analyze customers' viewing habits and deliver viewing recommendations (Harris 2012). Finally, Morgan Stanley has used Hadoop to determine, in real-time, how financial market events affect site activity by examining web logs, a process that in the past took months (Groenfeldt 2012). These examples illustrate how big data technologies enable firms to derive intelligence from Internet scale data in nearly real-time, improving the speed and the accuracy of managerial decision-making.

3.0 Data and Key Measures

3.1 Primary Data Source

The primary data source used for this analysis is the LinkedIn database. LinkedIn is a professional networking website that had over 175 million users worldwide at the time of the analysis.[5] Website participants report professional information on their profiles, including employment histories, education, geographic locations, accomplishments, and interest groups. LinkedIn also invites participants to list skills such as C++, Java, and Hadoop.

Among emerging data technologies, Hadoop investments have been identified by industry observers as some of the most closely associated with the recent wave of investment in data technologies (e.g., see Dumbill 2012, Bertolucci 2012). This analysis focuses specifically on firms' Hadoop investments, which are measured using the employment of technical workers who report having or using Hadoop skills. Similar measurement approaches, based on human capital investments, have been used in prior work on IT value, due to the large share of IT investment commanded by technical labor (Lichtenberg 1995; Brynjolfsson and Hitt 1996; Tambe and Hitt 2012). Due to the open source nature of the infrastructure software for Hadoop and its reliance on commodity hardware, human capital investments are likely to comprise an especially large share of investment into big data technologies. Therefore, data on human capital investment is likely to be highly correlated with overall Hadoop investment, and in fact may be one of the few available markers that can distinguish investment in emerging data technologies from investment in older generations of data technologies.

One caveat of developing and using a measure of Hadoop investment is that it will be correlated with the use of related data technologies within the firm[6], so the coefficient estimates produced using this measure are likely to reflect the returns to Hadoop investments as well as related investments in emerging data technologies. For instance, Hadoop investments are associated with a variety of new technical skills and technologies as well as increased demand for

[5] See http://www.linkedin.com. These data were collected and analyzed while the author was visiting LinkedIn.

[6] This is likely to be true when using measures of any specific information technology.

existing technical skills such as machine learning. Figure 1 uses the LinkedIn skills database to compare the technical skill mix of firms with Hadoop investments with that of other firms. Firms with Hadoop investments have disproportionately more workers with data skills such as "apache pig" and "map/reduce", in addition to skills such as "recommender systems" and "text classification" that have experienced increased demand from investments in new data technologies. The coefficient estimates produced by using Hadoop as a marker of firms' data technology investments are likely to reflect these broader underlying differences in technology and human capital across firms, and therefore, must be interpreted accordingly.

Because LinkedIn profiles include geographic data, Hadoop investment can also be measured at the firm-region levels. This observational unit is not as precise as the establishment level comparisons conducted in prior IT adoption research (e.g. Forman 2005), but provides some within-firm variation in how the labor pool impacts IT returns, and much of the economics literature treats the metropolitan region as the appropriate observational unit for labor market analysis (e.g., Card 1990; Borjas et al. 1996). Similar methods are used to create firm-region measures of other technical skills. Firm-level IT measures are created using the number of US-based IT workers in the database who report working for an employer in a given year. This approach follows prior work that uses employment history databases (Tambe and Hitt 2012), and due to the large fraction of the US technical workforce represented in the LinkedIn database, requires few sampling corrections. The firm's labor market is circumscribed using the firm-to-firm flows of technical workers as reported on workers' employment histories. This approach also directly follows prior work (Tambe and Hitt, forthcoming), and in addition to being a more precise measure of the labor market than geographic region, has the advantage that it allows comparisons between firms in the same regions that are embedded in different labor networks.

3.2 Measurement Error

The most important caveat to this measurement approach is uneven sampling across firms, skills, and regions, affecting which users choose to post which skills information into the database. Measurement error has been a common problem in most data sets used in firm-level IT research, and the error variance for even the most commonly-used IT measures has been estimated to be as high as 30-50% of the variance of the IT measure (Brynjolfsson and Hitt 2003). It is important, therefore, to understand the potential direction and magnitude of biases produced by these errors.

Estimation bias arises when omitted variables affect firm output as well as a) the propensity for a firm's IT workers participate on LinkedIn or b) the probability that a firm's employees accurately report skills. In general, the large sample size mitigates concerns related to website participation. LinkedIn includes much of the white-collar workforce, and within the US-based IT workforce, the size of the LinkedIn sample appears to be over 80% of the size of the total US IT workforce as reported by the Bureau of Labor Statistics. Correlations with external data sources indicate that the IT employment measures created using LinkedIn data accurately measure the total size of the firm's IT labor force. In logs, correlations between the IT employment measures generated using LinkedIn data and a) the IT employment measures developed using similar methods in recent work is 0.61 (Tambe and Hitt 2012), b) with survey data analyzed by Brynjolfsson and colleagues is 0.70, and c) with total employment in the packaged software industry (SIC 7372), in which a very large fraction of employees are IT employees, is 0.81.

The more important measurement concern relates to the likelihood that IT workers report their technical skills in a representative way.[7] The likelihood of reporting a particular skill online is dependent upon employer, skill, and worker attributes. For instance, there is a potential bias in online platform participation towards younger workers who use emerging technologies. Older IT workers using mature information technologies may have less incentive and lower proclivity to post their technical skills on LinkedIn. Moreover, workers at some firms may have more or less incentive to report their skills than workers at other firms.

However, firm-level estimates produced using these measures are robust to several of these sources of error. Systematically higher reporting rates for Hadoop relative to other skills will not impact the estimates if they do not affect the distribution of Hadoop skills across employers. Consistently higher reporting rates for skills at more productive firms can be addressed by normalizing firm-level skill measures by the rate at which other skills are posted at the firm. Factors that raise the rate of posting only Hadoop skills for workers only at more productive firms are more difficult to address, because this is one of the key sources of variation generating the estimates, and there are no administrative data at the skill level that can be used for direct comparisons. However, Spearman rank correlations reject the hypothesis of systematic differences between the distributions of IT employment across labor markets and the distribution of IT employees across labor markets who report skills (ρ=0.998), as well as between the distribution of IT employment across firms and the distribution of IT employees across firms who report skills (ρ=0.983). These correlations are inconsistent with large differences in skill reporting rates across firms.

[7] There has been recent concern about the possibility of fake profiles on social networks (Thier 2012). However, fake profiles will not bias the coefficient estimates unless big data skills are over- or under-represented in fake profiles, so they should not directly impact the estimates. Moreover, falsely reporting skills on one's profile will also not tend to exert an upward bias on the estimates unless individuals are more likely to falsely report these skills at high-performing firms.

However, because some measurement error bias is likely to affect the estimates, results are reported at the end of the analysis from sensitivity tests of the key estimates to measurement error. These are conducted by constructing alternative measures that differ in their sensitivity to error—for instance, results are reported from regressions using binary measures of firms' Hadoop investments, which are less sensitive to how many workers within each firm report skills. Similarly, the use of four-digit industry controls removes the effects of systematic differences in reporting rates across four-digit industries, which limits the impact of measurement problems to differences in skill reporting among firms within a four-digit industry. Fixed-effects estimators based on within-firm changes can remove some of the impact of these sources of bias, as long as they are time-invariant in the short panel. Although these tests do not remove the effects of measurement error from key estimates, they place bounds on the size of a bias term produced this type of measurement error.

A final source of measurement error is that workers only report current skills. The timing of the data capture, therefore, during a period in which there is significant cross-sectional variation across labor markets in big data investment, is important. However, one caveat is that using these data to impute skill distributions at prior employers is noisy. It produces an upward bias on key estimates if more productive firms are more likely to attract the types of workers who will eventually learn Hadoop, although this is mitigated by the short duration of the panel used in the analysis and is addressed by some of the tests reported below that account for timing, labor expenses, and the lagged productivity of firms.

3.3 Supplementary Data Sources

The Compustat database was used to create measures of capital, non-IT employment, labor expenses, and value added (sales minus materials), and to construct dummy variables for industry and year. Different analyses described below use industry variables constructed at either the two-digit and four-digit SIC levels. Value-added was chosen as a dependent variable to maintain consistency with prior IT productivity research and has the benefit that it is less subject than measures of total output to bias introduced by unobserved variables that affect demand and employment choices. Measures of capital and value-added were adjusted using methods common in the micro-productivity literature and were deflated to a base year using industry-level deflators posted at the Bureau of Economic Analysis.

Supplementary measures were created using firms' investments in SQL skills, as well as survey-based measures of firms' investments in data-driven decision-making practices, measured through surveys administered in 2008 and used in prior research on the impact of data-driven decision-making on firm performance (Brynjolfsson, Hitt, and Kim 2011).[8] The data-driven decision-making measure used in this analysis is based on questions about the extent to which data are used by the firm to make decisions about new products or services.[9] Rather than use these data at the firm level, which would significantly restrict the sample size, these data are used to construct industry level measures of data-driven decision-making, computed as the mean value of all firms in the four-digit industry for which survey responses are available.

4.0 Methods

4.1 Complementarities Theory

This paper formalizes the notion that returns to firms' Hadoop investments are increasing in the investments of other firms in the labor market by testing for (Edgeworth) complementarities between the investments of firms in the same labor pool. This definition of complementarities, formalized by Milgrom and Roberts (1990, 1994), has been the basis of an influential literature on organizational complementarities (Arora and Gambardella 1990; Huselid 1993; Milgrom and Roberts 1994; Ichniowski, Prennushi, and Shaw 1997; Athey and Stern 1998; Bresnahan, Brynjolfsson, and Hitt 2002; Bloom, Sadun, and Van Reenen 2012) as well as a literature on the relationship between firms' own investments and external investments in other knowledge-bearing assets such as R&D (Laursen and Foss 2003; Mohnen and Roller 2005; Cassiman and Vuegelers 2006). This definition of complementarities requires that the contribution of elements of an organizational choice vector to a payoff function (Π), such as productivity or innovation, is higher in the presence of the complementary elements.[10]

(1) $\Pi(1,1) - \Pi(0,1) \geq \Pi(1,0) - \Pi(0,0)$

This framework enables complementarities tests between choice variables for which the assumptions required for the application of differential techniques, such as the divisibility of inputs and non-convexity, are violated. Empirical studies in this literature commonly examine complementarities using the following two tests:

a) Examining correlations in inputs (*indirect tests*)

b) Testing how a "payoff" function is affected when complementary inputs are used in combination and independently (*direct tests*).

[8] I am very grateful to the authors for providing access to this data.

[9] This measure departs from the BHK measure of data-driven decision making because the other measures used in their construct reflect both firms' use of data assets as well as successful use of data technologies. The measure used in this analysis is based on a question that primarily focuses on practices because it used to test the benefits of modern data technologies, conditional on firms' practices.

[10] This definition of complements differs somewhat from the notion of complements used in factor demand theory (Amir 2005). The latter approach has been used by a number of studies in the IT value literature to estimate how firms' IT usage affects other factors, such as labor and capital (notable studies in this stream include Brynjolfsson and Hitt 1995 and Dewan and Min 1997). The two definitions of complements are equivalent under certain conditions described in Milgrom and Roberts (1990). Thanks are due to an anonymous reviewer for these points.

4.2 Complementarities Tests

Indirect correlation tests are conducted by testing how firms' own Hadoop investments are associated with firm, industry, and labor market factors. Tests for complementarities between inputs are implemented using short-run demand equations for different technical investments, conducted at the end of the sample period (similar to the approach used in Bresnahan et al. 2002). For evidence of complementarities, Hadoop investment should be conditionally correlated with the investment levels of other firms in the labor market, after controlling for other factors.

Direct complementarities tests are implemented using productivity as an outcome variable. The impact of IT investments on productivity has been measured using a variety of functional forms including the Cobb-Douglas (Brynjolfsson and Hitt 1996) as well as more flexible forms such as the translog and CES-translog (Dewan and Min 1997). However, direct complementarities tests using productivity as a dependent variable have most often been implemented using the Cobb-Douglas framework, due in part to its ease of augmentation (Brynjolfsson and Milgrom 2012 review this literature). In logs, this model has the following form, where *Y* is a measure of output such as value added and *X* are the firm's inputs and *i* and *t* index firm and year:

(2) $\log Y_{it} = \mathrm{C} + \Sigma\log \mathrm{X}_{it} + \textit{controls} + e_{it}$

Brynjolfsson and Milgrom (2012) discuss how direct complementarities tests between inputs can be implemented using the Cobb-Douglas model with the complementary inputs entered independently and in pairs to test the inequality restrictions in equation (1).

This paper uses the framework in (2) to implement direct complementarity tests between own and labor market Hadoop investment. Prior work in the IT productivity literature extends the Cobb-Douglas using investments in specific technologies, such as data practices or ERP adoption (Brynjolfsson, Hitt, and Kim 2011; Hitt, Wu, and Zhou 2002), and prior work in the micro-productivity literature has augmented this model to include external IT and R&D investment as a factor of production. The most common method of measuring external investment in knowledge-bearing inputs, based in an empirical literature on R&D spillovers (Griliches 1992 reviews this literature), uses data on other firms' investments combined with weighting measures reflecting the strength of the knowledge transmission path between firms. This method has been adapted for an emerging literature on IT spillovers (Cheng and Nault 2007, 2011; Chang and Gurbaxani 2012) and the method used in this paper follows prior work measuring labor market spillovers from IT investment (Tambe and Hitt, forthcoming). LinkedIn employment histories are used to directly model investments within the firm's labor pool (*S*) as:

(3) $\mathrm{S}_{\mathrm{T_i}} = \sum_{\mathrm{j}} \mathrm{w}_{\mathrm{ij}} T_{\mathrm{j}}$

where w_{jij} is the share of incoming IT labor that firm *i* acquires from firm *j* in each year and T is the relevant investment, measured as the number of technical workers with skills complementary to the relevant technology [e.g., overall employment of IT workers or employment of workers with Hadoop skills] of firm *j*, respectively, in that year. An alternative measure used in some regressions substitutes w_{jr}, the share of IT labor hired from firm *j* in a particular metropolitan area *r*, for w_j and substitutes T_{jr}, the investment levels for firm *j* in that metropolitan area, for T_j.

Direct complementarities tests can then be implemented by augmenting the production function in (2) with own and labor market Hadoop investment or other technologies independently and in pairs to test the hypothesis that the relationship in (1) is satisfied after controlling for other factors affecting productivity levels. The complementarities hypothesis is supported by positive coefficients on the interaction terms. Brynjolfsson and Milgrom discuss additional test statistics (2012).

5.0 Descriptive Statistics

5.1 Preliminary Evidence for Labor Market Complementarities

Table 1 reports industries with the largest Hadoop investments, measured using the skills data. Most are IT industries, but over 30% of Hadoop investment is in non-IT industries, including finance, transportation, utilities, and retail. Figure 2 illustrates the geographic distribution of this investment. Measures in Figure 2 are normalized by the IT labor force size in each region and represent the intensity of investment into Hadoop skills within the IT workforce, which is greatest in the San Francisco Bay area. As discussed above, these geographic imbalances in Hadoop skills reflect broader differences in underlying changes to the technical skills in these labor markets. Figure 3 compares the distribution of technical skills in the San Francisco Bay area, the most Hadoop intensive region, to the distribution of skills in the rest of the US IT labor force. The vertical axis is the fraction of each skill in the Bay area. The figure indicates a disproportionately high concentration of skills required to support large-scale data analytics, such as "Apache Pig," "Hadoop," "distributed algorithms," "recommender systems," and "HBase."

Figure 4 plots the age of some other major technical skills against their geographic concentration, where the ages of technical skills were collected using Internet data sources,[11] and where the geographic concentration of skills is computed by summing the squared share of the skill in each metro area across all metro areas, such that a value of one corresponds to all employees with a particular technical skill being located in a single metropolitan region. Some of the most concentrated skills are associated with emerging data technologies, such as Hadoop and Map/Reduce, and the least concentrated are older technical skills such as Cobol and Fortran,

[11] For example, see http://en.wikipedia.org/wiki/History_of_programming_languages.

which is consistent with the hypothesized importance of labor market concentration for emerging technical skills.

These comparisons suggest that the complementary human capital is concentrated for emerging IT innovations but diffuses as labor markets adjust. Figures 5 and 6 provide evidence that this labor market variation is consistent with patterns of returns to firms' Hadoop investments. For each firm in the sample, Figure 5 plots the Hadoop investment of firms in its labor market against changes to the firm's value-added relative to its industry from 2005 to 2011, where firms are divided into adopters and non-adopters of Hadoop. For firms that have made Hadoop investments, performance changes relative to the industry are increasing in levels of labor market Hadoop investment, but the performance of firms without Hadoop investments does not appear to be correlated with labor market investment levels.

Figure 6 uses SQL instead of Hadoop as the focal investment variable. Unlike Hadoop investment, there is no apparent benefit to being embedded in SQL-intensive labor markets when investing in SQL-based technologies, which is consistent with the short-run nature of the complementarities hypothesized in this study. Labor market concentration appears to matter for emerging technologies for which it is important to acquire technical workers from other companies, but the skills required to support mature technology investments can be acquired through other channels. Section 6 explores these relationships in a regression framework.

5.2 Summary Statistics

Table 2 reports means and standard deviations for the key measures used in the regression analysis along with statistical tests for the presence of significant differences in means between firms making Hadoop investments and other firms in the sample. Firms with Hadoop investments are characterized by higher employment (t=10.21) and greater IT-intensity (t=23.52), in part due to the higher fraction of these firms in IT industries. Measures of total IT employment and Hadoop-intensity within a firm's labor market are also significantly larger for firms with Hadoop investments. Table 3 reports simple correlations using 2011 values for the key regression measures.

6.0 Regression Analyses

6.1 Indirect Complementarities Tests

Table 4 implements correlation tests (indirect complementarity tests) on the key measures. The estimates in column (1) indicate greater Hadoop investment in industries characterized by higher levels of data-driven decision-making, which is consistent with the higher potential benefits from investing in new data technologies in these industries (t=1.95). The direct effects of labor market Hadoop investment are also significant (t=2.09), but the coefficient estimate on labor market SQL investment is not significant. By contrast, the coefficient estimates on the labor market investment measures are insignificant when SQL investment is used as the dependent variable in column (2), providing indirect evidence favoring strategic complementarities for firms' Hadoop investments, but not for investment in more mature data technologies. Column (3) adds SQL investments on the right-hand side of the equation. Hadoop investment is negatively associated with firms' SQL investment (t=2.79), perhaps indicating slower adoption when firms face higher replacement costs for existing technologies.

Because workers often acquire skills through hands-on interaction with new technologies, these correlations should be strongest for measures that capture whether firms hire IT labor from the specific establishments of other firms in which Hadoop investments are being made. Indeed, correlations with the broader firm-level measures of Hadoop investment disappear after including the more precise firm-region level measures of Hadoop investment in (4) (t=3.77). In general, all of the correlations reported in Table 4 are consistent with the hypothesis that being in the same labor networks as firms making similar investments is complementary to investment in emerging data technologies, but not for mature data technologies.

6.2 Baseline Productivity Estimates

Before presenting the results from the direct complementarities tests, baseline results are reported from embedding Hadoop investment measures into a productivity equation. As discussed earlier in the paper, the magnitude of the impact of the new data technologies on firm productivity remains a question of empirical interest.

The OLS estimates in column (1) of Table 5 indicate an IT output elasticity that is comparable to studies that use similar specifications (t=15.5) (e.g., see Lichtenberg 1995). The larger coefficient estimate when using IT employment instead of IT capital stock as an IT investment measure can be attributed to the use of employment, rather than labor expense, as the labor input measure. In the absence of direct labor expenses, higher wages paid to more educated workers in IT-intensive firms[12] are partially reflected in the IT coefficient rather than the labor coefficient. The estimated output elasticity on Hadoop investment is positive and significant (t=4.93), but interpretation of this estimate (as well as the estimates on Hadoop investment reported throughout this paper) is subject to the caveat described above—that it reflects not only firm-level differences in Hadoop investment, but also differences in related data technology and human capital investments. F-tests reject the equality of the coefficient estimates on Hadoop and IT investment (F=10.68, $p<0.001$). Marginal products are difficult to compute because the

[12] Bresnahan, Brynjolfsson, and Hitt (2002) provide evidence that IT use is associated with greater demand for skilled workers.

Hadoop investment measures mark larger underlying investments in new data technologies—however, the coefficient estimates on IT and Hadoop suggest a higher marginal product for Hadoop investment under reasonable assumptions about the share of investment commanded by Hadoop technologies relative to overall IT spending.

Columns (2) and (3) add SQL investment measures and data-driven decision-making (DDD) variables into the regression to control for confounding effects on the Hadoop investment measure caused by returns to correlated data practices which provide value to the firm. The DDD variable is statistically significant and similar in magnitude to estimates from prior work using measures based on these survey data (Brynjolfsson, Hitt, and Kim 2011). The Hadoop estimate remains significant after including these data variables (t=5.07), and is of the same magnitude as in the baseline regressions, suggesting that the estimated output elasticity on the Hadoop measure is not substantially biased upwards by omitted variables related to the use of other data technologies and practices.

However, the OLS estimate on the Hadoop measure indicates an output elasticity of 10%, which is a large value, some of which may be attributable to other omitted variable bias. Columns (4) through (6) add firm fixed-effects, which remove biases attributable to firm-level factors that are not time varying. The magnitude of the coefficient estimates on Hadoop investment falls considerably when firm fixed-effects are added in (4), indicating significant unobserved heterogeneity between firms making Hadoop investments and other firms. The coefficient estimate on Hadoop investment produced by the fixed-effects estimator indicates an output elasticity of only 1% to 2%, and is statistically significant at only the 15% level. Part of the explanation for this reduced coefficient estimate, however, may be that firm-fixed effects also remove the effects of assets that impact the returns to Hadoop investment. Columns (5) and (6) include firm fixed-effects but are separated by whether firms are above or below the median value of the DDD variable. With firm fixed-effects, the coefficient estimate on Hadoop investment indicates an output elasticity of slightly over 3% for firms with data assets (t=2.22), but the estimate is insignificant for firms in other industries. These estimates indicate that heterogeneity in the complete panel masks significantly higher productivity levels in firms with Hadoop investments and substantial data assets.

6.3 Direct Complementarity Tests

Table 6 introduces labor market measures and implements direct complementarity tests using productivity regressions. Labor market measures of IT, Hadoop, and SQL investment are standardized with means removed, so the main effects of investment in each of these technologies can be interpreted as the contribution to value-added from these investments for firms in labor pools with average investment levels.[13] Column (1) includes measures of own Hadoop investment as well as a measure of aggregate labor pool IT investment. The estimate on labor pool IT investment is statistically significant and the magnitude of the estimate indicates that hiring technical workers from labor markets with IT investment levels that are one standard deviation above the mean is associated with an output elasticity of 2% (t=2.4), which is very close to estimates from prior work that use similar methods but different data sources to quantify the aggregate impact of IT labor market spillovers on productivity growth (Tambe and Hitt, forthcoming). After including labor pool measures of Hadoop investment in (2), the estimate on the IT investment pool is no longer statistically significant, which is consistent with the argument that the "spillover" effect from labor market investment is generated by investments in new information technologies. The magnitude of the output elasticity on the Hadoop pool measure is larger than the estimate on the IT pool in (1) (t=3.25).

Column (3) adds interaction terms between labor pool investment and firms' own investments. The interaction term for Hadoop investment is positive (t=1.75), and the coefficient estimate on firms' own Hadoop investments is no longer significantly different than zero after including the interaction term, which is consistent with complementarities between internal and external Hadoop investment. The main effect on the labor market Hadoop investment measure is positive and statistically significant, perhaps due to factors allowing firms to acquire labor from other firms with Hadoop investments, but it becomes insignificant after including firm fixed-effects in (4) and (5). The interaction of own Hadoop investment and labor market Hadoop investment remains positive and significant after including firm fixed-effects. The estimates in columns (1) through (5) are consistent with the presence of complementarities between own and labor market investment in emerging data technologies, but not for general technological investments. For firms with Hadoop investments, OLS regressions indicate that being in a labor pool with investment levels that are one standard deviation higher than the mean is associated with an output elasticity of 8%, and panel estimates suggest an elasticity of 3% to 4%. However, in a labor pool with average Hadoop investment levels, the coefficient estimates on firms' own Hadoop investments are not significantly different than zero.

Columns (6) and (7) report results from OLS regressions using data on the firm-region combinations in which workers are acquired. The labor pool measures constructed using these data are more precise than the earlier measure but only available for 2011, so only cross-sectional regressions are presented. After including these measures in (6), the region-based measures are

[13] Labor market measures were computed as in equation (3) and then were logged before removing the mean and standardizing the variables.

significant but the estimate on the original Hadoop pool that does not account for region is no longer significant and in (7) the interaction term using the original Hadoop pool measure is no longer significantly different than zero after including the region-based interaction measure. These regression results imply that spillovers occur when hiring workers from the specific establishments of firms that are making Hadoop investments. These findings are more consistent with human capital explanations than labor network homophily if this type of homophily is more likely to arise along firm-level variables than along firm and region specific variables. For comparison with SQL investment, column (8) tests a specification similar to that in column (2) but using SQL measures. Unlike Hadoop investment, labor pool investment in SQL does not exhibit a statistically significant association with the firm's value-added.

Table 7a implements another form of direct complementarities tests proposed by Brynjolfsson and Milgrom (2012) that contrasts the productivity levels of firms with varying combinations of own and labor market investment. Each of the variables is dichotomized, where the firm's Hadoop investment variable is coded such that 1 represents employing at least one worker with Hadoop skills, and all other firms are coded 0, and the labor pool investment variable is coded as 1 for firms in labor pools in the top quartile of investment. Complementarities imply that after controlling for other inputs, value-added for firms with Hadoop investments are higher when these firms are located in labor pools in the top quartile of Hadoop investment.

The highest productivity group is where firms have higher levels of both factors (1, 1), where values are expressed as productivity levels relative to the omitted (0, 0) group. F-tests of productivity differences between the (1, 1) group and groups with any other combination of factors are just short of significant at the 10% level (F(1,1691) = 2.55; p = 0.111), but chi-squared tests reject the hypothesis that observations are randomly distributed across the four cells ($X^2(1)=129.2$, $p<0.01$), which is consistent with complementarities between own and labor market investment. The results from these tests are consistent with definitions of complementarities based on increasing differences (Brynjolfsson and Milgrom 2012). Table 7b reports the results of tests using SQL rather than Hadoop as the focal data technology investment. Evidence for complementarities disappears (F(1,1691) = 0.05; p = 0.823), which is consistent with the argument that labor market spillovers lose importance for explaining differences in returns to investments in mature technologies. Chi-squared tests reject the hypothesis that firms are independently distributed across the quadrants ($X^2(1)=7.32$, $p<0.01$), but these effects are not as strong as they are for Hadoop investments.

6.4 Endogeneity Tests

Like most estimates from large-scale empirical studies in the IT value literature, the estimates reported above are subject to endogeneity concerns, including omitted-variable bias and reverse causality. These concerns reflect well-known limitations with obtaining unbiased estimates of returns to IT investment in large-sample studies. Recent papers use econometric approaches to eliminate sources of endogeneity (Aral, Brynjolfsson, and Wu 2006; Tambe and Hitt 2012), but biased IT productivity estimates, especially due to a scarcity of effective instruments for IT investment, remain a persistent problem. Fixed-effects specifications address some issues, but time-varying factors affecting output and information technology investment can still impose an upward bias on the estimates.

The most significant concern is that unobserved firm-level factors, such as the quality of a firm's management or an anticipated increase in the demand for a firm's output, can exert an upward bias on the coefficient estimate on Hadoop investment. Although such sources of bias are difficult to completely remove, this section of the paper argues for a causal interpretation of the estimates presented above based on three types of evidence: 1) the pattern of correlations observed in the complementarities tests, 2) the timing of the observed productivity effects, and 3) evidence from additional robustness tests.

First, an explanation favoring reverse causality or simultaneity would be consistent with a pattern of estimates in which higher productivity firms systematically make Hadoop investments. However, the estimates in Table 6 indicate that firms' Hadoop investments, in the absence of complementary investments by other firms in the labor market, exhibit no statistical associations with higher productivity levels, which is somewhat inconsistent with an explanation in which higher productivity firms make Hadoop investments unless this investment behavior only occurs in specific labor markets. Similarly, biases on the labor market investment measures may occur if these measures are correlated with firm-level factors lowering the costs of attracting employees from Hadoop intensive firms (e.g. employment reputation), but this bias term should appear on the main effect of the labor pool measure rather than the complementarity term. In general, the emphasis on complementarities tests minimizes the importance of sources of bias that could be expected to impact the main-effect estimates on firms' own investment or labor market investment, rather than acting at their confluence.

Along similar lines, in Table 6, labor market position is not correlated with productivity unless firms hire from the specific establishments in which other firms are making Hadoop investments. Correlations between Hadoop investment and productivity levels disappear for firms that hire technical workers from the establishments of firms that have not made Hadoop investments, even if these establishments' parent firms have made Hadoop investments. This reduces the likelihood of estimation bias on the labor market measures that is not associated with hiring within specific regions.

Second, the timing of the observed productivity changes favors a causal relationship between Hadoop investment and firm performance. Figure 7 indicates that labor productivity levels (value added per employee) at firms making Hadoop investments diverge from that of other firms after 2009.[14] Firms making these investments were more productive on average (consistent with the drop in the magnitude of the coefficient estimate after including firm fixed-effects in Table 5), but the more recent *divergence* in performance between firms is less easily explained by the argument that more productive firms tend to invest in superior data capabilities. Interestingly, in the absence of Hadoop investments, there is no apparent difference in labor productivity levels for firms in Hadoop intensive labor markets and other markets, indicating that the recent productivity lift experienced by firms in these markets is principally due to complementarities with investment in these emerging technologies. This suggests that the IT-enabled labor market wage divergence documented in prior work (Forman, Goldfarb, and Greenstein 2012) may reflect the performance of users of emerging technologies in these markets, rather than being shared by all labor market participants.

Third, estimates from additional robustness tests are presented in Table 8 that attempt to minimize bias due to residual sources of firm-level heterogeneity, including time-varying firm-level heterogeneity. Columns (1) and (2) separate the sample into firms that are above and below median labor productivity levels in 2006 in order to remove some firm-level heterogeneity. The coefficient estimate on the Hadoop measure is statistically significant in both samples, which implies that the estimates are robust to restricting the sample to low-performing firms. Including lagged productivity measures in (3) provides a control for unobserved and time-varying differences in past-productivity levels. Including this measure reduces the estimate on Hadoop investment to half its prior magnitude, but it remains significant. It is not significant after including fixed-effects and lagged productivity levels in (4), but the point estimate is close to its value in prior fixed-effects tests. Finally, columns (5) and (6) report results from sub-samples of firms that are growing and shrinking their headcounts. Firms with shrinking headcounts are less likely to be making investments in new technical capabilities in the absence of productivity improvements. The estimate on Hadoop investment, however, remains significant and similar in magnitude in both samples of firms.

Another concern, due to the use of employment instead of labor expense as a workforce measure, is that Hadoop investment reflects unmeasured heterogeneity in workforce quality

[14] I thank an anonymous editor for suggesting this test.

across firms. Columns (7) and (8) include labor expenses rather than employment to measure human capital differences, which substantially reduces the sample size due to the limited availability of labor expense data in the Compustat database. OLS regressions in (7) using labor expenses produce similar results to using employment, but including fixed-effects on the reduced sample in (8) eliminates the statistical significance of all coefficients except labor.

6.5 Measurement Error Tests

As described earlier in the paper, another class of concerns with the estimates presented above is that unobserved differences in the propensity of firms' employees to report Hadoop skills on a platform such as LinkedIn may be correlated with firm-level performance measures. It is difficult to remove the effects of biases produced by this form of measurement error, but robustness tests can bound the impact of this source of bias. To test the sensitivity of the key estimates to error in the skills measures, measures of Hadoop investment with different error characteristics are used in the baseline regressions.

Column (1) of Table 9 uses a binary adoption variable that takes the value one when at least one of a firm's employees reports Hadoop as a technical skill. Converting Hadoop investment into a binary measure mitigates the degree to which factors causing employees at productive firms to report skills at significantly higher rates could bias the key estimates, because the estimates from the binary measure are generated only from variation between productivity at firms with some Hadoop investment and those with none. The use of this measure has relatively little impact on the estimated returns to Hadoop investment. The use of the binary measure along with a fixed-effects estimator in column (2) produces an output elasticity of about 3%, which is similar in magnitude to the key estimates presented earlier in the paper, although the estimate is insignificant.

Column (3) uses a measure of Hadoop investment that is normalized by the number of employees at the firm reporting SQL skills, which removes the effects of factors that increase overall skill reporting rates for personnel at specific firms, unless these factors increase the rate of reporting Hadoop but not other skills. The estimate produced by the normalized measure is consistent with estimates generated earlier in the paper (t=2.93). Column (4) is a similar exercise that Java skills in the denominator, rather than SQL skills, which may be a better comparison if the distribution of Hadoop skills shares more in common with software development than with data management, but the results are again similar. Finally, column (5) presents regression results when limiting the sample to firms making investments in Hadoop skills, so that variation in the sample is produced from differences in quantities of workers with Hadoop skills in firms who have at least one employee listing these skills. Estimates from this regression indicate that correlations between Hadoop investment and productivity observed in earlier regressions reflect not only productivity differences between Hadoop-using firms and other firms, but also differences in productivity between firms with different intensities of Hadoop investment (t=3.98). These tests rule out most measurement factors except for those that cause employees at more productive firms to report Hadoop but not other skills. However, some of the tests described above, such as the binary measure in (1) and the timing results reported in Figure 7, minimize the likelihood that this type of measurement issue is driving the estimates.

7.0 Summary and Conclusions

Hadoop investment appears to be associated with higher productivity levels in data-intensive industries. However, the analysis underscores the tradeoffs described earlier in the context of Sears: managers of data-intensive firms must balance the benefits of extracting greater value from their data using big data technologies against the higher costs of acquiring the required expertise in a tight labor market. Outside of labor markets characterized by high levels of Hadoop investment, the estimated returns to firms' own Hadoop investments were not statistically significant. For managers who choose not to incur the expense required to attract the necessary expertise in a tight labor market, investments in traditional database systems—for which the skills are widely available—may remain more effective. Alternatively, managers can wait. Big data technologies are maturing and the channels through which to acquire the complementary skills, such as university programs, are expanding to new markets (Thibodeau 2012). Managers, therefore, should weigh the competitive benefits offered by big data technologies against the costs of acquiring the skills, both of which should fall over time.

For high-tech labor policy, these findings underscore the importance of skill acquisition channels for understanding why IT-enabled growth differs across labor markets during large waves of new IT investment. The findings suggest that access to complementary skills is associated with performance advantages for early adopters of big data technologies, but that the diffusion of complementary know-how erodes the productivity advantages experienced by firms located in these labor markets. Therefore, the channels through which these skills diffuse merit greater attention because the rate of this process has implications for the duration of the growth differences that result from the spread of big data technologies. Policies accelerating the diffusion of big data know-how to other labor markets, such as those that accelerate the establishment of courses in business analytics by institutions providing education or training, can narrow inequality in the stock of complementary skills across labor markets, but if there are significant lags in this process, firms in big data intensive labor markets will continue to experience faster productivity growth than other firms.

There are many related areas for future research related to the diffusion of big data technologies. A question of policy interest, as the use of big data technologies becomes widespread, is the extent to which large-scale data driven decision-making will complement or substitute other types of human capital in the labor market (such as statistical proficiency). Furthermore, acquiring complementary skills is not the only obstacle to successful big data use. Effective big data use may require changes to existing data assets, management practices, and data governance. Prior work provides insight into how management practices provide superior performance by enabling firms to analyze interactions with customers, competitors, and suppliers (Mendelson 2000; Tambe et al 2012); the use of big data technologies can raise the returns to these practices by improving the depth of insight that firms derive from these interactions as well as the speed at which they respond. Installing these capabilities often requires organization-wide changes to complement data-driven practices.

References

Agarwal, R., T. Ferratt. (2001) Crafting an HR Strategy to Meet the Need for IT Workers. *Communications of the ACM*. 44(7):58-64.

Ang, S., S. Slaughter, and K. Ng. (2002) Human Capital and Institutional Determinants of Information Technology Compensation: Modeling Multilevel and Cross-Level Interactions. *Management Science*, 48(11):1427-1445.

Aral, S., Brynjolfsson, E., & Wu, D. J. (2006) Which came first, IT or productivity? Virtuous cycle of investment and use in enterprise systems. *Virtuous Cycle of Investment and Use in Enterprise Systems*.

Arora, A., & Gambardella, A. (1990) Complementarity and external linkages: the strategies of the large firms in biotechnology. *The Journal of Industrial Economics*, 361-379.

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

Attewell, P. (1992) Technology diffusion and organizational learning: The case of business computing. *Organization Science*, 3(1):1-19.

Audretsch, D. and M. Feldman. (1996) R&D Spillovers and the Geography of Innovation and Production. *American Economic Review.* 86(3):630-640.

Bertolucci, J. (2012) Big Data's Wild West Period Stars Hadoop. *InformationWeek.* (Accessed online at http://www.informationweek.com/big-data/news/big-data-analytics/240006652/big-datas-wild-west-period-stars-hadoop)

Bapna, R., Langer, N., Mehra, A., Gopal, R., & Gupta, A. (2013) Human Capital Investments and Employee Performance: An Analysis of IT Services Industry. *Management Science*, 59(3):641-658.

Barua, A., D. Mani, and R. Mukherjee. (2012) Measuring the Business Impacts of Effective Data. Report accessed at http://www.sybase.com/files/White_Papers on Sep 15, 2012.

Bloom, N., Sadun, R., & Van Reenen, J. (2012) *Americans do IT better: US multinationals and the productivity miracle* (No. w13085). National Bureau of Economic Research.

Borjas, G., R. Freeman, and L. Katz. (1996) Searching for the Effect of Immigration on the Labor Market. *American Economic Review,* 246-251.

Bresnahan, T. F., Brynjolfsson, E., & Hitt, L. M. (2002) Information Technology, Workplace Organization, and Labor Demand: Firm-Level Evidence. *Quarterly Journal of Economics*, 117(1):339-376.

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

Brynjolfsson, E. and L. Hitt. (1996) Paradox Lost? Firm-Level Evidence on the Returns to Information Systems Spending. *Management Science*. 42:4:541-558.

Brynjolfsson, E., & Hitt, L. M. (2000) Beyond computation: Information technology, organizational transformation and business performance. *The Journal of Economic Perspectives*, 23-48.

Brynjolfsson, E., L. Hitt, and H. Kim. (2011) Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance? Working Paper.

Brynjolfsson, E. and A. McAfee. (2011) The Big Data Boom is the Innovation Story of Our Time. *The Atlantic*. Nov 21, 2011. Accessed at http://www.theatlantic.com/business/archive/2011/11/the-big-data-boom-is-the-innovation-story-of-our-time/248215/ on Sept 15, 2012.

Brynjolfsson, E., & Milgrom, P. (2012) Complementarity in organizations. *The Handbook of Organizational Economics*, 11.

Card, D. (1990) The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial and Labor Relations Review* 43(2):245-257.

Cassiman, B., & Veugelers, R. (2006) In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition. *Management science*, 52(1):68-82.

Chang, Y. B., & Gurbaxani, V. (2012) The Impact of IT-Related Spillovers on Long-Run Productivity: An Empirical Analysis. *Information Systems Research*, 23(3-Part-2):868-886.

Cheng, Z. and B. Nault. (2007) Industry Level Supplier-Driven IT Spillovers. *Management Science* 53(8):1199-1216.

Cheng, Z. and Nault, B. R. (2012) Relative Industry Concentration and Customer-Driven IT Spillovers. *Information Systems Research*, 23(2):340-355.

Desmet, K. and Rossi-Hansberg, E. (2009) Spatial Growth and Industry Age, *Journal of Economic Theory.* 144:2477-2502.

Dedrick, J., V. Gurbaxani, and K. Kraemer. (2003) Information Technology and Economic Performance: A Critical Review of the Empirical Evidence. *ACM Computing Surveys*, 35(1):1-28.

Dewan, S. and K. Kraemer. (2000) Information Technology and Productivity: Evidence from Country-Level Data, *Management Science* 46(4):548-562.

Dewan, S., Ganley, D., & Kraemer, K. L. (2010) Complementarities in the diffusion of personal computers and the Internet: Implications for the global digital divide. *Information Systems Research*, 21(4), 925-940.

Dewan, S. and C. Min. (1997) The Substitution of Information Technology for Other Factors of Production: A Firm-Level Analysis, *Management Science* 43(12):1660-1675.

Dumbill, E. (2012) What is Apache Hadoop? (Accessed online at http://strata.oreilly.com/2012/02/what-is-apache-hadoop.html on September 10th, 2012)

Duranton, G., & Puga, D. (2004) Micro-foundations of urban agglomeration economies. *Handbook of regional and urban economics*, 4:2063-2117.

Fichman, R. G., & Kemerer, C. F. (1997) The assimilation of software process innovations: an organizational learning perspective. *Management Science*,43(10):1345-1363.

Forman, C. (2005) The Corporate Digital Divide: Determinants of Internet Adoption. *Management Science.* 51(4):641-654.

Forman, C., Goldfarb, A., and Greenstein, S. (2012) The Internet and Local Wages: Convergence or Divergence? *American Economic Review,* 102:556-575.

Freeland, C. (2012) In Big Data, Potential for Big Division. *New York Times*. January 12, 2012. Accessed at http://www.nytimes.com/2012/01/13/us/13iht-letter13.html on September 14, 2012.

Glanz, J. (2013). Is Big Data an Economic Big Dud? *New York Times*. August 17, 2013. Accessed at http://www.nytimes.com/2013/08/18/sunday-review/is-big-data-an-economic-big-dud.html?pagewanted=all&_r=1& on January 1st, 2014.

Greenstein, S. and F. Nagle (2012) Digital Dark Matter and the Economics of Apache. Working Paper.

Griliches, Z. (1992) The Search for R&D Spillovers, *Scandinavian Journal of Economics* (94):29-47.

Groenfeldt, T. (2012) Morgan Stanley Takes on Big Data With Hadoop, *Forbes*, May 30, 2012. Accessed online at http://www.forbes.com/sites/tomgroenfeldt/2012/05/30/morgan-stanley-takes-on-big-data-with-hadoop/ on March 12, 2013.

Harris, D. 2012. (2012) Netflix analyzes a lot of data about your viewing habits. *Gigaom*. June 14, 2012. Accessed at http://gigaom.com/2012/06/14/netflix-analyzes-a-lot-of-data-about-your-viewing-habits/ on March 12, 2013.

Henschen, D. (2012) Why Sears is Going All-In On Hadoop. *Informationweek*. Accessed online at http://www.informationweek.com/global-cio/interviews/why-sears-is-going-all-in-on-hadoop/240009717 on March 8, 2013.

Hitt, L. M., Wu, D. J., & Zhou, X. (2002) Investment in enterprise resource planning: Business impact and productivity measures. *Journal of Management Information Systems*, 19(1):71-98.

Hitt, L. M., & Snir, E. M. (1999). The role of information technology in modern production: complement or substitute to other inputs. In *Workshop on Information Systems and Economics*.

Huselid, M. (1995) The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of management journal*, 38(3):635-672.

Ichniowski, C., Shaw, K., & Prennushi, G. (1995). *The effects of human resource management practices on productivity* (No. w5333). National Bureau of Economic Research.

Jaffe, A. (1986) Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value", *American Economic Review* 76(5):984-1001.

Laursen, K., & Foss, N. J. (2003). New human resource management practices, complementarities and the impact on innovation performance.*Cambridge Journal of economics*, 27(2):243-263.

Levina, N. and M. Xin. (2007) Comparing IT Workers' Compensation Across Country Contexts: Demographic and Institutional Factors, *Information Systems Research* 18(2):193-210.

Lichtenberg, F. (1995) The Output Contributions of Computer Equipment and Personnel. A firm-level analysis, *Economics of Innovation and New Technology* 3(3-4):201-218.

Mckinsey Global Institute. (2011) Big Data: The Next Frontier for innovation, competition, and productivity.

Melville, N., K. Kraemer, and V. Gurbaxani. (2004) Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value. *MIS Quarterly*. 28(2):283-322.

Mendelson, H. (2000) Organizational architecture and success in the information technology industry. *Management science*, 46(4):513-529.

Milgrom, P., & Roberts, J. (1990) The economics of modern manufacturing: Technology, strategy, and organization. *The American Economic Review*, 511-528.

Milgrom, P., & Roberts, J. (1994) Complementarities and systems: Understanding Japanese economic organization. *Estudios Economicos*, 3-42.

Mohnen, P., & Röller, L. H. (2005) Complementarities in innovation policy. *European Economic Review*, 49(6):1431-1450.

Provost, F., & Fawcett, T. (2013) Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1):51-59.

Provost, F. and V. Kolluri. (1999) A Survey of Methods for Scaling Up Inductive Algorithms. *Data Mining and Knowledge Discovery*, 3(2):131-169.

Porter, M., & Stern, S. (2001) Location matters. *Sloan Management Review*, 42(4):28-36.

Rooney, B. (2012) Big Data's Big Problem: Little Talent. *Wall Street Journal.* April 29, 2012.

Saxenian, A. (1996) *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*, Harvard University Press, Cambridge, USA.

Schaal, D. (2011) Orbitz CEO plans to revolutionize hotel sort – holy Hadoop. Nov 18, 2011. Accessed at http://www.tnooz.com/2011/11/18/news/orbitz-ceo-plans-to-revolutionize-hotel-sort-holy-hadoop/ on March 12, 2013.

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

Tambe, P. and L. Hitt. (forthcoming) Job hopping, information technology spillovers, and productivity growth. *Management Science*.

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

Thibodeau, P. (2012) Grad schools add big-data degrees. *Computerworld*. Oct 8th, 2012. Accessed at http://www.computerworld.com/s/article/9232106/Grad_schools_add_big_data_degrees on January 17, 2013.

Thier, D. (2012) An Estimated 83 Million Facebook Profiles are Fake. Accessed at http://www.forbes.com/sites/davidthier/2012/08/02/83-million-estimated-facebook-profiles-are-fake/ on December 3, 2012.

Figure 1: Comparison of Skill Distributions in Firms with Hadoop Investment and Other Firms

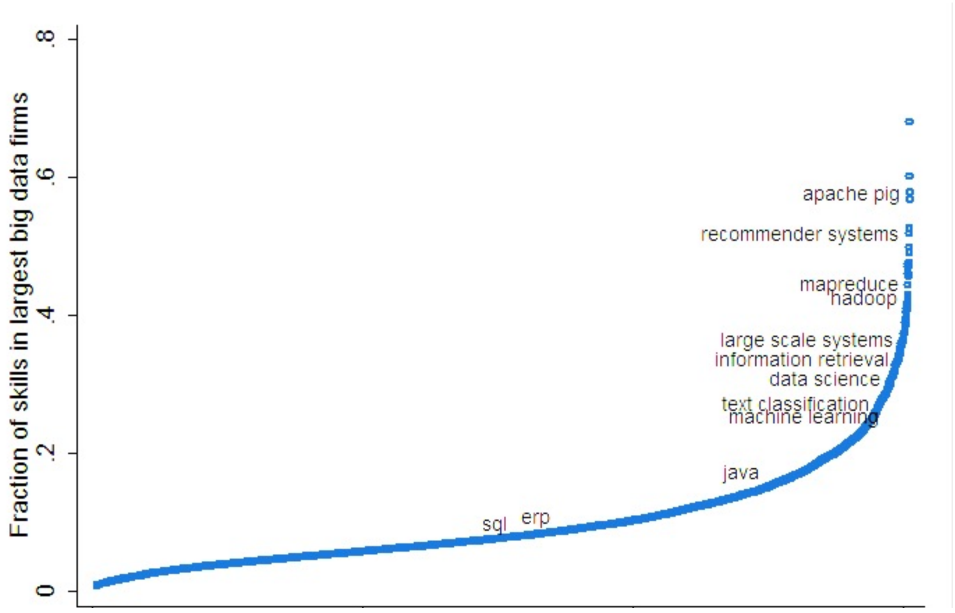

Figure notes: Y-axis is the fraction of workers with each skills employed at firms with large Hadoop investments. SQL and ERP in the middle of the graph are in proportion to the fraction of IT employment at these firms. Technical skills to the right are disproportionately higher values for these firms.

Table 1: Top Ten Industries by Hadoop Investment*

6-Digit NAICS Industry	*% of Hadoop Engineers Employed in Industry*
Software Publishers	20.4
Internet Publishing and Broadcasting	13.0
Computer Systems Design	5.2
Radio and Television Broadcasting	5.0
Internet Shopping	4.4
Computer Peripheral Manufacturing	4.3
Computer Services	3.9
Commercial Banking	3.1
Computer Storage Manufacturing	2.5
Wired Telecommunication	2.2
All other sectors	36.0
Total	100.0

*Industries based on 6 digit NAICS codes. Table only includes 6-digit NAICS industries with at least ten firms and is based on publicly traded firms only.

Figure 2: Top Metropolitan Regions by Hadoop Investment

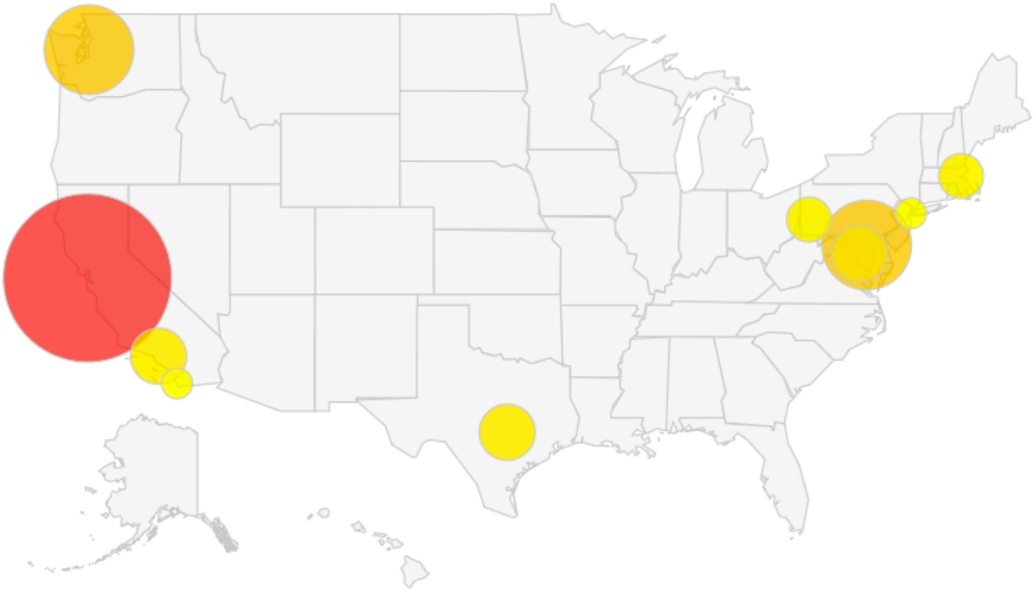

Figure notes: The size of each circle represents the number of technical workers with Hadoop skills in each region normalized by total IT labor force size for each region.

Figure 3: Skill Distribution in San Francisco Bay Area Compared with Other Skills

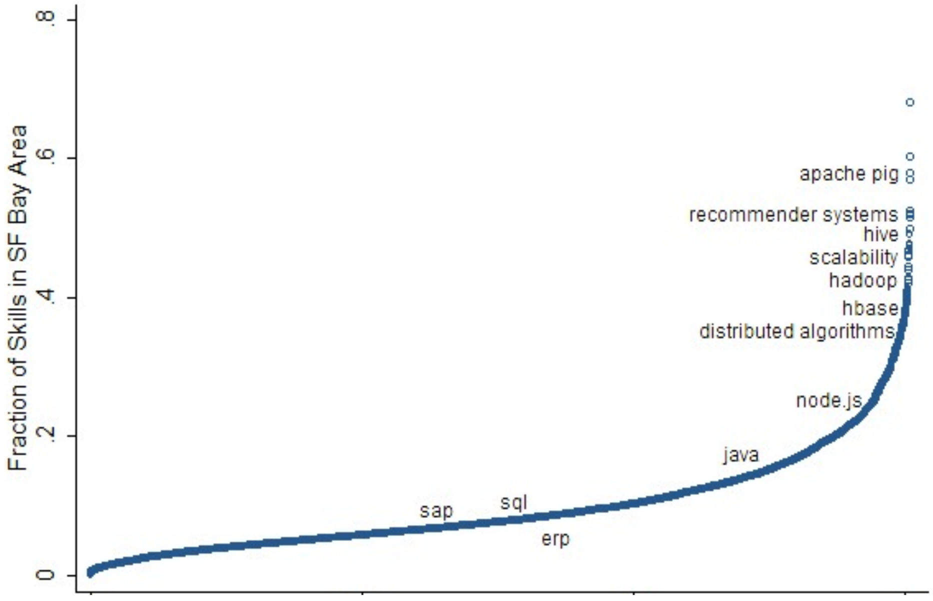

Figure notes: The y-axis is the fraction of workers with each skills employed in the San Francisco Bay Metropolitan Area. SQL and ERP are close to in proportion to total IT employment in the region. Skills to the right are disproportionately found in this metropolitan region.

Figure 4: Technical Skill Age and Geographic Concentration

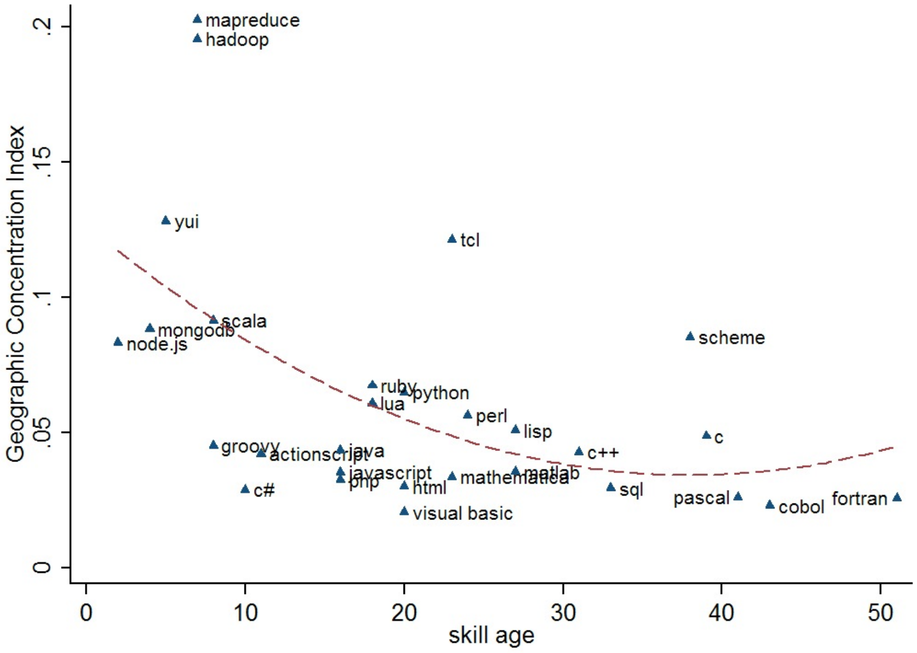

Figure 5: Sales Growth Plotted Against Labor Market Hadoop Investment

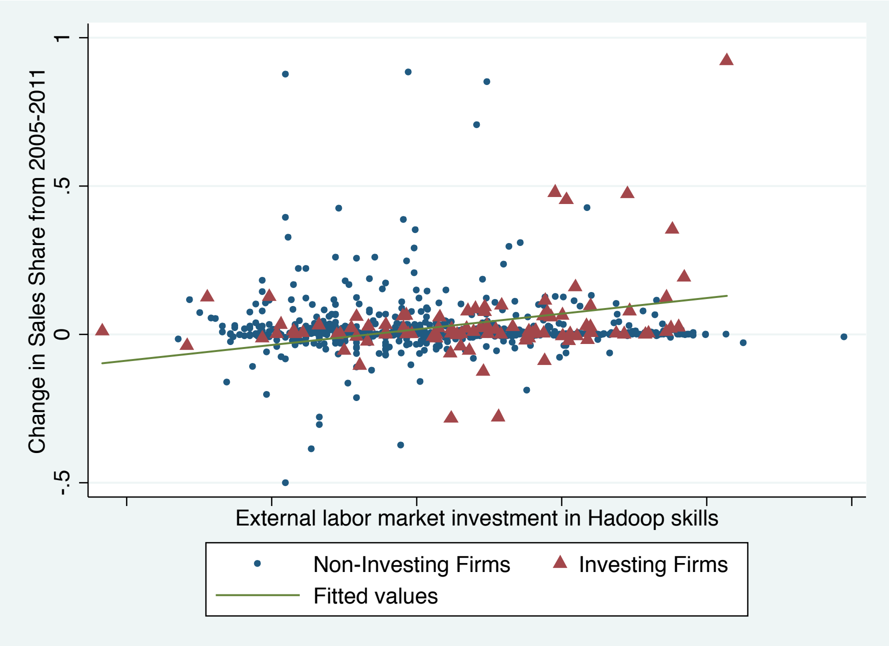

Figure notes: The y-axis is change in sales share relative to other firms in industry. Red triangles are firms with Hadoop investment. Blue circles are all other firms.

Figure 6: Sales Growth Plotted Against Labor Market SQL Investment

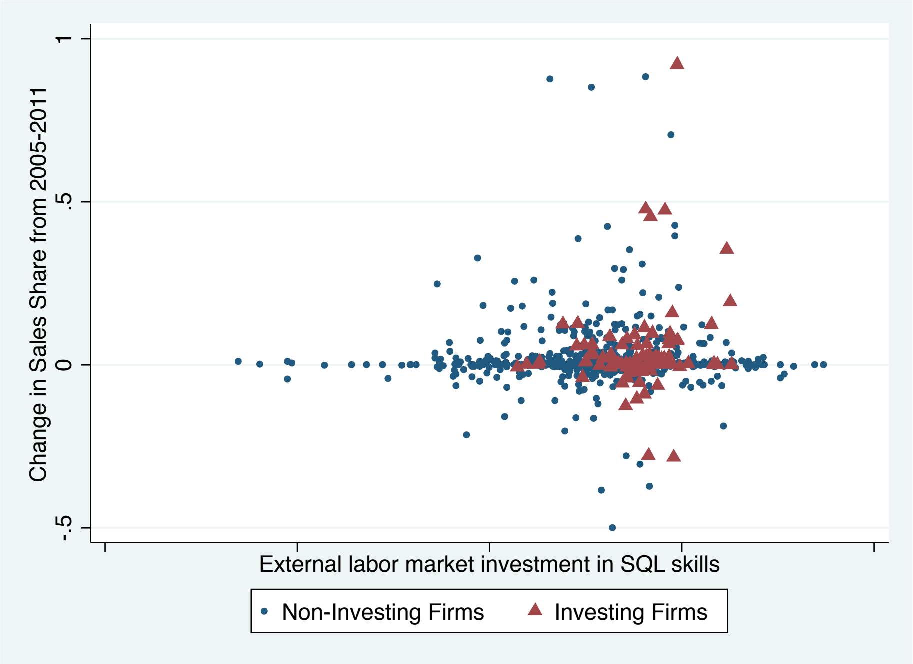

Figure notes: The y-axis is change in sales share relative to other firms in industry. Red triangles are firms with SQL database investment. Blue circles are all other firms.

Table 2: Summary Statistics and Mean Comparisons for Key Regression Variables

	(1)	(2)	(3)	(4)	(5)
	Mean	*Std. Dev.*	*Hadoop Using Firms*	*All Other Firms*	*t-test*
Log(Value added)	6.89	1.04	7.54	6.12	11.52**
Log(Capital)	5.51	2.39	6.65	5.72	5.61**
Log(Non-IT employment)	8.28	1.91	9.46	8.31	8.83**
Log(Labor expenses)[a]	6.66	1.85	7.86	6.68	4.15**
Log(IT employment)	4.06	1.63	5.63	3.31	22.53**
Log(Hadoop)	.132	.463	1.27	0	63.01**
Log(IT pool)	0	1	.248	-.719	10.05**
Log(Hadoop pool)	0	1	2.49	.605	12.16**
Log(SQL)[+]	1.45	2.60	4.08	1.09	17.34**
N	12,677		211	1,484	

Economic figures are from 2011 Compustat data. Value-added, labor expenses, and capital are reported in millions of dollars and are deflated to 2006 values. Non-IT employment, IT employment, Hadoop, and SQL are reported in number of employees. $^{**}p<.05$. A significant value in column (5) rejects the hypothesis that the means in (3) and (4) are equal. [+]2011 values only. Means and standard deviations in Columns (1) and (2) are reported for all observations in panel. Mean comparison statistics in Columns (3) through (5) are only for 2011 values. [a]Labor expense means and standard deviations are reported for 1,693 available observations for full sample and for 200 available observations for 2011 values. IT pool is the IT employment of other firms weighted by incoming IT labor share and Hadoop pool is the Hadoop employment of other firms weighted by incoming IT labor share (see equation (3)). Both pools are standardized with means removed.

Table 3: Correlations for Key Regression Variables

	1	**2**	**3**	**4**	**5**
1. Log(Capital)	1.00				
2. Log(Non-IT employment)	0.76	1.00			
3. Log(Hadoop employment)	0.16	0.22	1.00		
4. Log(IT employment)	0.42	0.59	0.48	1.00	
5. Log(IT pool)	0.11	0.20	0.21	0.44	1.00
6. Log(Hadoop pool)	0.11	0.22	0.26	0.49	0.84
Simple correlations are shown for 2011 values of each variable. N=1,692.					

Table 4: Demand for Skills as a Function of Firm, Industry, and Labor Market Variables

	(1)	(2)	(3)	(4)
	Log(Hadoop)	*Log(SQL)*	*Log(Hadoop)*	*Log(Hadoop)*
VARIABLES	OLS	OLS	OLS	OLS
DDD [Industry]	0.034*	-0.097	0.033*	0.034*
	(0.018)	(0.063)	(0.017)	(0.017)
Log(IT employment)	0.114***	1.094***	0.131***	0.126***
	(0.016)	(0.065)	(0.018)	(0.018)
Log(Non IT employment)	-0.047**	-0.017	-0.047**	-0.041**
	(0.018)	(0.066)	(0.018)	(0.018)
Log(Capital)	0.011	-0.053	0.010	0.012
	(0.014)	(0.059)	(0.014)	(0.014)
Log(Value added)	0.069***	0.200***	0.072***	0.066***
	(0.021)	(0.075)	(0.021)	(0.021)
Log(IT pool)	-0.002	-0.457	-0.010	-0.008
	(0.034)	(0.279)	(0.034)	(0.034)
Log(Hadoop pool)	0.040**	0.169	0.043**	0.005
	(0.019)	(0.153)	(0.019)	(0.021)
Log(SQL pool)	-0.022	0.168	-0.020	-0.019
	(0.019)	(0.154)	(0.019)	(0.019)
Log(SQL)			-0.016***	-0.017***
			(0.006)	(0.006)
Log(Hadoop pool-region)				0.054***
				(0.014)
Controls	Industry	Industry	Industry	Industry
Observations	902	902	902	902
R-squared	0.309	0.675	0.312	0.319
All regressions are short-run demand equations, estimating how firm's investments in Hadoop and SQL skills are associated with other factor inputs, industry variables, and labor pool variables. OLS estimates are reported using 2011 values. Robust standard errors are shown in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Industry controls are included at the two-digit SIC level.				

Table 5: Baseline Productivity Equations

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	*All*	*With SQL measures*	*With SQL and DDD measures*	*All*	*DDD above median*	*DDD below median*
	OLS	OLS	OLS	FE	FE	FE
Log(Capital)	0.304***	0.315***	0.314***	0.117***	0.145***	0.173***
	(0.016)	(0.015)	(0.019)	(0.010)	(0.018)	(0.022)
Log(Non IT emp)	0.477***	0.430***	0.433***	0.646***	0.551***	0.618***
	(0.021)	(0.020)	(0.024)	(0.015)	(0.024)	(0.032)
Log(IT emp)	0.155***	0.194***	0.201***	0.081***	0.087***	0.045*
	(0.012)	(0.012)	(0.015)	(0.011)	(0.018)	(0.024)
Log(Hadoop)	0.095***	0.132***	0.144***	0.013	0.031**	-0.019
	(0.019)	(0.027)	(0.028)	(0.009)	(0.014)	(0.028)
Log(SQL)		0.009*	0.010*			
		(0.005)	(0.006)			
DDD [Industry]			0.038**			
			(0.018)			
Controls	Industry Year	Industry Year	Industry Year	Year	Year	Year
Observations	12,677	12,677	7,594	12,677	3,699	3,895
R-squared	0.891	0.863	0.863	0.358	0.464	0.260

All regressions are from Cobb-Douglas production functions using logged value added as the dependent variable. Standard errors are clustered on firms and shown in parentheses; *** p<0.01, ** p<0.05, * p<0.1. DDD is the data-driven decision-making variable that uses the survey instrument described in Brynjolfsson, Hitt, and Kim (2011). Log(Hadoop) is the log of the number of employees reporting Hadoop skills. Industry controls are included at the two digit SIC level.

Column (1) is an OLS productivity regression using the Hadoop investment measure.

Column (2) adds a measure of SQL investment to the regression in (1).

Column (3) adds the industry level DDD variable to the regression in (2).

Column (4) adds firm fixed-effects to the regression in (1).

Column (5) is the fixed-effects regression in (4) restricted to the subsample of firms with the DDD variable above the mean.

Column (6) is the fixed-effects regression in (4) restricted to the subsample of firms with the DDD variable below the mean.

Table 6: Direct Complementarity Tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample Period	*All years*	*All years*	*All years*	*All years*	*All years*	*2011*	*2011*	*2011*
VARIABLES	OLS	OLS	OLS	FE	FE	OLS	OLS	OLS
Log(Capital)	0.308***	0.308***	0.307***	0.195***	0.117***	0.382***	0.381***	0.405***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.010)	(0.021)	(0.021)	(0.032)
Log(Non IT emp)	0.478***	0.478***	0.478***	0.573***	0.645***	0.371***	0.368***	0.327***
	(0.021)	(0.021)	(0.021)	(0.020)	(0.015)	(0.025)	(0.025)	(0.034)
Log(IT emp)	0.164***	0.163***	0.163***	0.124***	0.081***	0.167***	0.179***	0.138***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.018)	(0.020)	(0.032)
Log(Hadoop)	0.115***	0.110***	0.000	0.024***	-0.030	0.113***	0.127	0.177***
	(0.023)	(0.023)	(0.063)	(0.010)	(0.028)	(0.036)	(0.114)	(0.044)
Log(IT pool)	0.021***	0.005	-0.027	0.002	-0.026**	-0.027	-0.072	-0.093
	(0.008)	(0.010)	(0.023)	(0.006)	(0.011)	(0.036)	(0.048)	(0.061)
Log(Hadoop pool)		0.033***	0.032***	0.012*	0.009	0.046	0.043	0.042**
		(0.012)	(0.012)	(0.007)	(0.006)	(0.040)	(0.041)	(0.020)
Hadoop x Had pool			0.085*		0.037*		-0.219	
			(0.049)		(0.022)		(0.155)	
IT x IT pool			0.010*		0.007**		0.022	
			(0.006)		(0.003)		(0.015)	
Hadoop region						0.087***	0.073***	
						(0.021)	(0.022)	
Hadoop x region							0.152**	
							(0.074)	
Log(SQL)								0.039***
								(0.013)
Log(SQL pool)								0.055
								(0.050)
Controls	Industry Year	Industry Year	Industry Year	Year	Year	Industry	Industry	Industry
Observations	12,677	12,677	12,677	12,677	12,677	1,692	1,692	902
R-squared	0.891	0.891	0.891	0.352	0.359	0.870	0.870	0.864

All regressions are from Cobb-Douglas production functions using log(value added) as the dependent variable. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Standard errors are clustered on firm. All pool variables are constructed as the skill-based investments of all other firms weighted by the share of IT labor coming from those firms. Columns (6)-(8) only include 2011 observations only.

Table 7a: Productivity Matched and Mismatched on Complements (*Hadoop*)

Hadoop \\ Labor pool	1	0
1	0.296** (0.065) N=118	0.056 (0.045) N=298
0	0.081 (0.077) N=92	0 (N/A) N=1,184

Huber-White robust standard errors are shown in parentheses and clustered on firm. Pearson Chi-Sq(1)=129.2, p<0.01. Year 2011 observations only. The Hadoop variable takes the value 1 for firms with Hadoop investments and 0 otherwise. The labor pool variable takes the value 1 if firms are in labor markets in the top quartile of Hadoop investment and 0 otherwise. The coefficient estimate in each quadrant indicates differences in mean logged value added for firms in each quadrant relative to firms in the omitted group (0,0), after controlling for levels of other inputs.

Test Statistic: F(1,1) + F(0,0) - F(0,1) - F(1,0)
F(1, 1691) = 2.55; p = 0.111

Table 7b: Productivity Matched and Mismatched on Complements (*SQL*)

SQL \\ Labor pool	1	0
1	0.105 (0.080) N=392	0.071 (0.118) N=31
0	0.061 (0.073) N=1,116	0 (N/A) N=153

Huber-White robust standard errors are shown in parentheses and clustered on firm. Pearson Chi-Sq(1)=7.32, p<0.01. Year 2011 observations only. SQL takes the value 1 for firms with SQL investments. The labor pool variable takes the value 1 if firms are in labor markets in the top quartile of SQL investment and 0 otherwise.

Test Statistic: F(1,1) + F(0,0) - F(0,1) - F(1,0)
F(1, 1691) = 0.05; p = 0.823

Figure 7: Labor Productivity Separated by Hadoop/Labor Pool quadrants (see Table 7a)

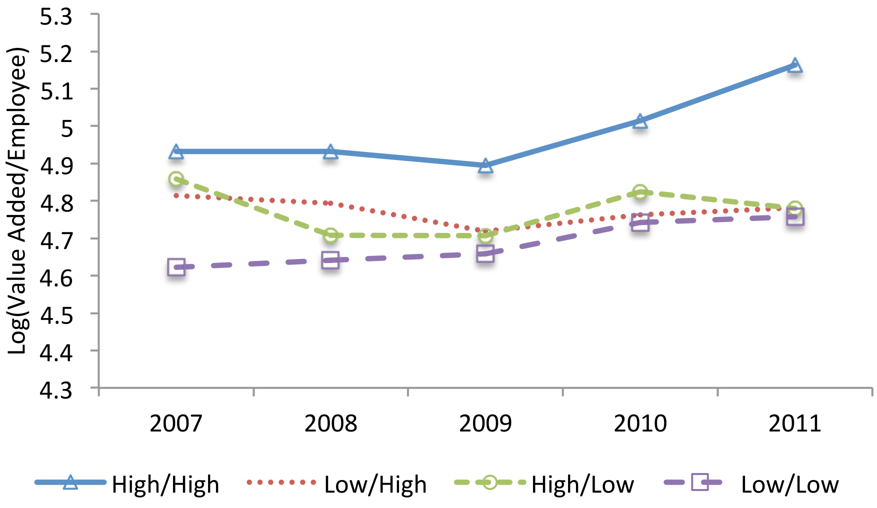

Figure notes: The y-axis is the log of value added divided by total employment. For each of the high/high, low/high, high/low, and low/log groups, the first value corresponds to own Hadoop investment levels and the second value corresponds to labor market Hadoop investment levels.

Table 8: Additional Robustness Tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	High Prod	*Low Prod*	*Lagged VA*	*Lagged VA*	*Growing Firms*	*Shrinking Firms*	*Labor Expenses*	*Labor Expenses*
VARIABLES	OLS	OLS	OLS	FE	OLS	OLS	OLS	FE
Log(Capital)	0.115***	0.327***	0.074***	0.040***	0.309***	0.337***	0.191***	-0.011
	(0.013)	(0.014)	(0.009)	(0.012)	(0.026)	(0.021)	(0.050)	(0.033)
Log(Non IT emp)	0.797***	0.410***	0.101***	0.518***	0.456***	0.394***		
	(0.018)	(0.019)	(0.011)	(0.018)	(0.033)	(0.027)		
Log(IT emp)	0.095***	0.121***	0.020***	0.018	0.190***	0.202***	0.120**	-0.070*
	(0.014)	(0.012)	(0.007)	(0.014)	(0.023)	(0.018)	(0.049)	(0.040)
Log(Hadoop)	0.083***	0.113***	0.046***	0.016	0.137***	0.133***	0.139***	-0.022
	(0.025)	(0.039)	(0.009)	(0.011)	(0.042)	(0.031)	(0.048)	(0.031)
Log(Labor)							0.697***	0.777***
							(0.099)	(0.050)
Log(VA) – Lag 1			0.778***	0.235***				
			(0.021)	(0.011)				
Observations	5,013	7,664	8,968	8,968	2,648	4,424	1,693	1,693
R-squared	0.908	0.823	0.965	0.349	0.918	0.905	0.904	0.240

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered on firm. All regressions are Cobb Douglas using logged value added as the dependent variable. Dollar figures are deflated to 2006 values.

Columns (1) and (2) are OLS regressions separated into subsamples with labor productivity above and below the mean.

Columns (3) and (4) are OLS and FE estimates with lagged value added directly included into the productivity regression.

Columns (5) and (6) are OLS regressions separated into subsamples of firms that are growing and shrinking employment.

Columns (7) and (8) are OLS and FE estimates using labor expenses instead of total employment.

Table 9: Robustness Tests for Measurement Error in the Skills Variables

	(1)	(2)	(3)	(4)	(5)
	Binary Hadoop Measure	*Binary Hadoop Measure*	*Hadoop Compared with SQL*	*Hadoop Compared with Java*	*Firms with Hadoop Investment*
VARIABLES	OLS	FE	OLS	OLS	OLS
Log(Capital)	0.308***	0.118***	0.384***	0.305***	0.346***
	(0.016)	(0.010)	(0.025)	(0.017)	(0.043)
Log(Non-IT employment)	0.479***	0.646***	0.410***	0.484***	0.351***
	(0.021)	(0.015)	(0.030)	(0.022)	(0.055)
Log(IT employment)	0.171***	0.081***	0.157***	0.179***	0.221***
	(0.011)	(0.011)	(0.019)	(0.012)	(0.050)
Hadoop y/n	0.183***	0.030			
	(0.054)	(0.020)			
Log(Hadoop / SQL)			0.044***		
			(0.015)		
Log(Hadoop / Java)				0.030**	
				(0.012)	
Log(Hadoop)					0.167***
					(0.042)
Observations	12,677	12,677	1,690	1,690	1,335
R-squared	0.891	0.358	0.906	0.890	0.929

The dependent variable in all regressions is logged value added. Robust standard errors are reported in parentheses; *** p<0.01, ** p<0.05, * p<0.1.

Column (1) reports OLS results when using a binary measure of investment in Hadoop skills.

Column (2) reports fixed effects results when using a binary measure of investment in Hadoop skills.

Column (3) normalizes investment in Hadoop skills by investment in SQL skills.

Column (4) normalizes investment in Hadoop skills by investment in Java skills.

Column (5) limits the sample to firms that have invested in Hadoop skills.