MIT IDE RESEARCH BRIEF VOL. 2016.09

BIG DATA INVESTMENT, SKILLS, AND FIRM VALUE

Prasanna Tambe

U.S. businesses are in the midst of a data-driven management revolution. 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 this data promises to produce insights that will revolutionize managerial decision-making.

Because this type of data analysis has, in many cases, outpaced existing technological capabilities, interest is growing in the potential economic impact of big data technology investments—as well as challenges that may prevent implementation. One particular challenge for enterprises is the difficulty of acquiring the technical skills required to support big-data tools. A 2012 report[1] about Sears' early adoption of Hadoop, for instance, noted that as demand for big data technologies grows, so do the problems of finding sufficient skills. Questions have been raised about whether talent shortages could limit the rate of productivity growth.

My research found a gap in the academic literature on IT-enabled growth. Prior research[2] focused on organizational factors to explain variation in IT returns. However, recent work[3] finds evidence of systematic differences in growth rates across labor markets during large waves of investment in new IT innovations. The speed at which knowledge barriers fall can impact the rate of IT innovation diffusion; and there are wide differences across labor markets. In particular, differences exist in the supply of workers and the skills complementary to the new information technologies--especially during the early period when there are few channels for acquiring these skills.[4] This variation may explain why some firms unlock value from new IT innovations faster than others.

IN THIS RESEARCH BRIEF

- Labor-market factors have shaped early returns on investment (ROI) in big data technologies such as Hadoop.
- Aggregate corporate investment levels produce a supply of complementary technical skills during the early stages of technology diffusion.
- From 2006 to 2011, Hadoop investments were associated with 3% faster productivity growth for firms with significant existing data assets and in labor networks with significant aggregate Hadoop investment.
- For mature data technologies, such as SQL-based databases, the importance of labor-market concentration disappears because skills are diffused and readily available.
- Geography, corporate investment and channels for technical-skill acquisition are important factors in productivity growth rates during the spread of new IT innovations. Hadoop investment appears to be associated with higher productivity levels in data-intensive industries.

1 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-onhadoop/240009717 on March 8, 2013.

2 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.

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

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

MIT INITIATIVE ON THE DIGITAL ECONOMY

LABOR MARKETS AND ROI

I also examined how labor markets have shaped early returns on investment (ROI) in a specific big-data technology—Hadoop-based systems. I tested the hypothesis that returns on Hadoop investments have been concentrated in select labor markets, such as Silicon Valley, due to aggregate corporate investment in the early stock of technical human capital required to support Hadoop. In other words, the more Hadoop expertise there is in the labor market, the easier it is to hire skilled employees.

As with R&D, firms that invest in new IT should derive significant benefits from the related investments of other firms when the complementary know-how is scarce. During this period, hiring employees from other early adopters may be an especially important means of acquiring technical expertise. As technologies mature and complementary-skill channels emerge (e.g. university degree programs), the importance of differences in labor market "spillovers" from the investments of nearby firms should decline. The same is true for the performance advantages of being located in specific labor markets. This leads to three predictions:

A) Investment in emerging data technologies should be concentrated in select labor markets.

B) Investments in these technologies should yield higher returns in these labor markets.

C) The advantages of labor market concentration should disappear for investments in mature data technologies.

To test these hypotheses I used LinkedIn data to distinguish the investments firms make in emerging data technologies versus investments in mature data technologies and to measure investments in human capital complementary to specific technologies.[5]

Such a study is important because the value of technological investment is determined, in part, by the supply of professionals who can translate technologies into business outcomes. 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 I examined how supply adjusts to the demand for skills complementary to specific technologies--such as big data technologies--to better understand temporal and regional dynamics resulting from new IT innovations. Data on the fine-grained structure of skills within the IT labor force will help in understanding how labor markets impact returns on new technological innovations.

MEASURING THE VALUE OF HADOOP SKILLS

At the time of data collection, over 30% of workers with Hadoop skills were employed in Silicon Valley, compared with 4% of total U.S. IT employment in that region. Mature technical skills were much less geographically concentrated. Direct complementarities tests indicate that a firm's Hadoop investments yield higher productivity returns in Hadoop-intensive labor markets.

The most robust productivity estimates indicate that the output elasticity of 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 ROI in mature data technologies appears to be unaffected by the labor markets in which the investments are made.

These findings are closely related to several academic papers[6] on the value of modern data analytic technologies, but there is still active debate about whether, and under what conditions, big data technologies have driven generalized economic gains. Because investments in complementary human capital may command a larger share of expenditures for big-data technologies than for earlier IT, data on skills may provide benefits for empirically distinguishing firms' investments in specific data technologies.

5 Fichman, R. G., & Kemerer, C. F. (1997) The assimilation of software process innovations: an organizational learning perspective. Management Science, 43(10):1345-1363.
Note that Microsoft has bid to buy LinkedIn in June http://www.wsj.com/articles/microsoft-to-acquire-linkedin-in-deal-valued-at-26-2-billion-1465821523

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

Additionally, this paper provides evidence that labor-market adjustments can explain geographic variations in IT returns at the firm level -- a topic of long-standing interest.[7]The research also indicates that the importance of labor market spillovers—a potentially important source of variation in IT returns across labor markets—varies according to the maturity of technical skills.

Apache Hadoop, among 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 free or low-cost, and readily available, one of the primary expenses that firms face is the acquisition of expertise required to install, maintain, and facilitate these clusters to support data analysis.

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 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.

Netflix uses a Hadoop-based infrastructure to analyze customers' viewing habits and deliver viewing recommendations. And Morgan Stanley has used Hadoop to determine, in real-time, how financial market events affect site activity by examining web logs, a process that previously took months. 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.

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

DATA AND KEY MEASURES

The primary data source used for this analysis is the LinkedIn database. At the time of this analysis, LinkedIn had over 175 million users worldwide. 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. (Some of the potential measurement errors of using this data are addressed in the full research paper which can be found here.)

This analysis focuses specifically on Hadoop investments, which are measured using the employment of workers with Hadoop skills. Due to Hadoop's open-source nature 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. Hadoop investments also are associated with a variety of new technical skills and technologies as well as increased demand for existing technical skills such as machine learning.

ONE OF THE PRIMARY EXPENSES THAT FIRMS FACE IS THE ACQUISITION OF EXPERTISE REQUIRED TO INSTALL, MAINTAIN, AND FACILITATE THESE CLUSTERS TO SUPPORT DATA ANALYSIS.

BIG DATA INVESTMENT, SKILLS, AND FIRM VALUE

Prasanna Tambe

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

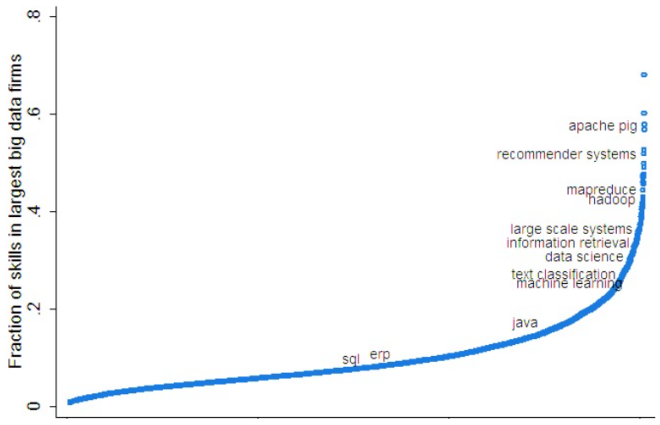

Figure notes: Y-axis is the fraction of workers with each skills employed at firms with large Hadoop investments. SQL and ERP in the middle of the graph are in proportion to the fraction of IT employment at these firms. Technical skills to the right are disproportionately higher values for these firms.

Figure 1 uses the LinkedIn skills database to compare the technical skill mix of firms with Hadoop investments versus 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.

Because LinkedIn profiles include geographic data, Hadoop investment can also be measured at the firm-region levels. This observational unit provides some within-firm variation in how the labor pool impacts IT returns. Similar methods are used to create firm-region measures of other technical skills. Firm-level IT measures are created using the number of U.S.-based IT workers in the database who report working for an employer in a given year.

COMPLEMENTARITIES THEORY

This paper formalizes the notion that returns to firms' Hadoop investments are increasing the investments of other firms in the labor market by testing for complementarities between the investments of firms in the same labor pool.

I found that industries with the largest Hadoop investments, based on skills data, are mostly IT industries. More than 30% of Hadoop investment, however, is in non-IT industries, including finance, transportation, utilities and retail.

I also found that the intensity of investment into Hadoop skills within the IT workforce is greatest in the San Francisco Bay area. The geographic imbalance in Hadoop skills reflect broader differences in underlying changes to the technical skills in these labor markets.

These comparisons suggest that the complementary human capital is concentrated for emerging IT innovations but diffuses as labor markets adjust. For firms that have made Hadoop investments, for example, 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.

HADOOP INVESTMENT APPEARS TO BE ASSOCIATED WITH HIGHER PRODUCTIVITY LEVELS IN DATA-INTENSIVE INDUSTRIES.

TRADEOFFS TO CONSIDER

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.

BIG DATA INVESTMENT, SKILLS, AND FIRM VALUE

Prasanna Tambe

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 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 business analytics courses, 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 others.

Acquiring complementary skills is not the only obstacle to successful big-data use: changes to existing data assets, management practices and data governance may also be needed. Prior work[8] provides insight into how management practices provide superior performance by enabling firms to analyze interactions with customers, competitors and suppliers; the use of big data technologies can raise the returns to these practices by improving the depth of insight derived, as well as the speed at which they respond. As with many innovative practices, installing these capabilities often requires organization-wide changes to complement data-driven technologies.

8 Mendelson, H. (2000) Organizational architecture and success in the information technology industry. Management science, 46(4):513-529. Tambe, P., and Hitt, L. M. (2012) The productivity of info rmation technology investments: New evidence from IT labor data. Information Systems Research, 23 (3-Part-1):599-617.

MIT INITIATIVE ON THE DIGITAL ECONOMY

The MIT Initiative on the Digital Economy brings together internationally recognized researchers seeking solutions to how people can – and will – thrive in a digital world. Drawing on MIT's strengths in technology and innovation, IDE explores the profound impact of a rapidly advancing digital economy, and how it's changing the ways we live and work.

SUPPORT THE MIT IDE

Foundations, private donors and corporate members are critical to the success of the IDE. Their support fuels cutting-edge research by MIT faculty and graduate students, and enables new faculty hiring, curriculum development, events, and fellowships. Contact Christie Ko (cko@mit.edu) to learn how you or your organization can support the IDE.

TO LEARN MORE ABOUT THE IDE, INCLUDING UPCOMING EVENTS, VISIT MITSLOAN.MIT.EDU/IDE

MIT INITIATIVE ON THE DIGITAL ECONOMY

Prasanna Tambe is an Associate Professor of Information, Operations and Management Sciences at the New York University Stern School of Business, and a Fellow at the MIT IDE.