

# A Measure of Firms' Information Practices Based on Textual Analysis of 10-K Filings

DRAFT

August 2013

Adam Saunders  
Sauder School of Business  
University of British Columbia  
Vancouver, BC, Canada V6T 1Z2  
adam.saunders@sauder.ubc.ca

Prasanna Tambe  
Stern School of Business  
New York University  
New York, NY 10012  
ptambe@stern.nyu.edu

## Abstract

*There is a scarcity of consistent data on the firm-level adoption of technological innovations and IT-enabled business practices. Yet this type of data is necessary in order to understand how IT-related innovations have created business value in the last decade. In an effort to bridge this information gap, we perform textual analysis on 10-K filings from publicly traded U.S. firms to identify technologies and practices related to the collection, storage, and analysis of data. We then develop a panel dataset of close to 40,000 observations from 1996-2010, which allows us to estimate how these technologies and practices are affecting firm value and performance. We find that before data-related keywords appear in a firm's 10-K, its market value is below the industry average (at the six-digit NAICS level). Yet after the appearance of the keywords, the firm's market value is significantly greater than the industry average. We also find that operating profits relative to the industry are significantly higher after a period of three to five years. These findings are robust to a battery of controls, as well as long-difference and fixed-effects estimators that use within-firm variation over time. Moreover, for within-firm long differences, there is a persistence of excess value for up to 4 years after the first appearance of these keywords. In addition to quantifying the economic effects of the spread of data-intensive practices, this analysis demonstrates that textual analysis can be used on readily available public data to measure otherwise "invisible" economic activity.*

We thank Ashish Agarwal, Anitesh Barua, Erik Brynjolfsson, V.C. Choudhary, Chris Forman, Vijay Gurbaxani, Shawndra Hill, Lorin Hitt, Steve Kimbrough, Kristina McElheran, Shivendu Shivendu, the participants of the 2011 Workshop on Information Systems and Economics (WISE) and the 2012 Economics of Digitization Workshop at the NBER Summer Institute for helpful comments and suggestions on this work. We also thank Wonguen (Johnny) Lee, Chi-Ted Edward Li, Stefanus Soegiarto, Christine Tedijanto, and Chris Wang for outstanding research assistance.

## 1. Introduction

Massive investments in information technologies over the last two decades have vastly improved firms' abilities to collect, store, and analyze data from customers, suppliers, and partners.<sup>1</sup> These investments are leading to a “data revolution,” where some IT-intensive firms are developing the capabilities to analyze immense stores of data to inform decisions about their major activities, including procurement, manufacturing, pricing, and marketing. There are a number of prominent examples of firms—such as Amazon.com or Caesars Entertainment<sup>2</sup>—that have used these new data capabilities to achieve competitive advantage and superior performance (Loveman 2003; Hopkins 2010). Furthermore, there has been a recent wave of interest in the growth of this data collection and analysis, a phenomenon that has been termed “big data” (McKinsey 2011). For example, in 2011, venture capital in the high-tech sector increased by 33%, and Internet-related venture capital funding went up by 68%.<sup>3</sup> Yet, our understanding of the economic implications of data-intensive technologies and practices, such as “data-driven decision making” (Brynjolfsson, Hitt, and Kim 2011), is still in its infancy.

One reason for our relatively poor understanding of this phenomenon is that spending on these new technological innovations and IT-enabled business practices remains invisible on corporate balance sheets and income statements (Brynjolfsson and Saunders 2010, 2011; Corrado and Hulten 2010). In general, there exists a severe deficit of publicly available data about IT spending and practices. The last freely available dataset about IT spending was

---

<sup>1</sup> For instance, starting in 1987, more than 40% of U.S. business investment in equipment has been in information processing equipment (hardware, software and communications), and since 2000, it averaged 50% annually. Source: Bureau of Economic Analysis, National Income and Product Accounts Table 5.3.5.: “Private Fixed Investment by Type,” available at <http://www.bea.gov>.

<sup>2</sup> Caesars Entertainment was formerly known as Harrah's Entertainment before 2010.

<sup>3</sup> Source: “Moving Ahead,” a report by PricewaterhouseCoopers and the National Venture Capital Association (NVCA), based on data from Thomson Reuters, available at <http://www.pwcmoneytree.com>.

produced by *InformationWeek* magazine in 1997. Fee-based datasets that measure IT spending (the most notable being the Harte-Hanks CITDB database) and other datasets that require special deputized access (such as micro-level data from the U.S. Census Bureau) do not adequately cover investments in data centers, cloud computing, and data analytics software. Furthermore, these datasets may be fundamentally inadequate for tracking the spread of innovations because they are focused on firms' spending on IT hardware and software. In particular, the development of big data infrastructure, may not be easily tracked through IT spending numbers alone—it is often rooted in combining commodity hardware with “free” Apache software from sources such as the Hadoop project. While managerial surveys have recently been used to capture differences in the data intensity of firms (Barua, Mani, and Mukherjee 2011; Brynjolfsson, Hitt, and Kim 2011), such sources have been generally limited to one-time data collections of 150-200 firms each. The limited number of firms surveyed is due to the cost and difficulty of collecting responses from hundreds of firms, as well as the even greater problem of gathering repeated observations from the same firms over time.

Fortunately, we observe that the explosion of online data about firms' activities presents new, alternative channels for economic measurement. For example, text from public financial documents, product announcements, news stories, analyst reports, blogs, and even Twitter data can be utilized to observe firms' actions and how they are perceived by customers and market participants. A recent report prepared by authors from the Conference Board and the National Science Foundation (NSF) identified the collection and interpretation of cyber-data as the most promising frontier in the measurement of organizations and innovation (Corrado and Lane 2009). Moreover, innovation in text analytics has improved our ability to extract meaning from unstructured text into structured data that can be investigated using statistical analysis.

In this study, we develop a new dataset of close to 40,000 observations over the period 1996-2010 through a textual analysis of 10-K documents, annual reports filed by publicly traded firms to the U.S. Securities and Exchange Commission (SEC). Each 10-K filing includes a description of the business, risk factors, major business initiatives, operations, and financial statements. In the 10-Ks, we search for a set of keywords drawn from *Big Data: The Next Frontier for Innovation, Competition and Productivity*, a 2011 industry report by the McKinsey Global Institute, to identify technologies and practices related to the collection, storage, and analysis of data.<sup>4</sup> For each keyword, we draw a random sample of fifty 10-K documents in which that keyword appears to examine it in context.<sup>5</sup> We examine whether the keyword primarily describes the *usage* of a technology or data practice, or whether the keyword primarily appears in a *non-usage* context (such as in product descriptions or industry trends). We then match the appearance of these keywords with corresponding firm identifiers to generate a 15-year panel database, focusing on the relationship between *usage*-related keywords, market value, and performance.<sup>6</sup>

We demonstrate that firms with two or more distinct data usage-related keywords (e.g., “automated decision,” “data-driven,” or “data warehouse”) in their 10-Ks are worth significantly more than their peers. Significant economic returns persist after removing firm effects by estimating long-difference and fixed effects specifications. Furthermore, the gains in market value are greatest in the two-year difference estimates, where we estimate almost a 20% increase in market value relative to the 6-digit NAICS industry. However, we find significant excess

---

<sup>4</sup> The report is available at: [http://www.mckinsey.com/mgi/publications/big\\_data/index.asp](http://www.mckinsey.com/mgi/publications/big_data/index.asp).

<sup>5</sup> For any keyword that appeared in less than fifty 10-K documents during the sample period, we examine all 10-Ks during the sample period with that keyword. However, any keyword that appeared in less than five unique firms in the sample was not classified.

<sup>6</sup> Results based on the non-usage keywords might only be capturing the present value of growth or excess returns in technology markets.

value (relative to the industry) remaining up to three to four years after the first appearance of two or more distinct usage keywords in a 10-K. We also find significant increases in operating profits that do not occur right away, but rather, that grow over time through a five-year period after the keywords appear. One explanation consistent with these findings is that complementary organizational or managerial changes are required while the firm implements data-related initiatives. We should note that usage keywords are not merely a proxy for highly valued firms, as we find no significant increases in market value *preceding* the appearance of usage-related keywords. We estimate that shortly before adopting these data practices, such firms are worth less than their peers, whereas afterwards, they are worth significantly more than their peers.

Creating this database allows us to analyze how data-intensive practices have diffused among firms and industries in the United States, as well as to estimate the returns to adopting these practices. The sample size in this study is dramatically greater than what has been used in prior work, as we create a panel of 2,000-3,000 publicly traded firms in the United States from 1996-2010. The longitudinal dimension of our data is particularly useful for enabling the application of panel data estimators that remove the effects of time-invariant unobservables—such as management quality—that can potentially bias estimates produced using cross-sectional data.

The contribution of our work is to apply textual analysis of 10-Ks to measure the use of data-intensive practices and then quantify how these practices affect firm value and performance. This analysis is of current interest given the wave of investment in data analytics, cloud computing, and social media. Due to the spending and managerial attention focused on these technologies, it is of fundamental interest to quantify how the diffusion of this technology can be expected to impact business value in the coming years. We also demonstrate that textual

analysis can be used to convert unstructured textual information into a new and useful source of data about firms' innovation patterns. A relatively recent stream of academic papers uses textual analysis to extract consumer demand information from the unstructured text in online product and service reviews (Archak, Ghose, and Ipeiritis 2011; Ghose, Ipeiritis, and Li 2012), as well as investor sentiment from news stories and 10-Ks (Tetlock 2007; Tetlock, Saar-Tsechansky, and Macskassy 2008; Loughran and McDonald 2011). While the feasibility of measuring firms' economic assets through textual analysis has been discussed in prior work (Zadrozny 2006), to the best of our knowledge, no prior research has implemented it. Our dataset of close to 40,000 observations over a 15-year period is one of the largest panel datasets ever constructed to measure data-intensive practices and technologies. Moreover, our results are remarkably robust considering that we rely on only one document per firm, per year—the 10-K. Indeed the use of 10-Ks as the underlying source of textual information suggests that our approach is conservative: Firms are not required to disclose technologies or business practices in their 10-Ks, making it all the more noteworthy how well our keyword data correlates with excess market value and returns.

In the next section, we describe how we generate the data to produce the sample. In Section 3, we present our theoretical model. In Section 4, we produce descriptive statistics related to the adoption and diffusion of data practices, and report results of how the adoption of these technologies affects firms' market value and performance. We conclude in Section 5 and discuss future research directions.

## **2. Data Generation**

For our textual analysis, we collected 10-K filings from Westlaw's LIVEDGAR system covering fiscal years 1996-2010.<sup>7</sup> Each 10-K filing includes a description of the business, risk

---

<sup>7</sup> 1996 was the first year we could obtain a reliably large sample of 10-K documents electronically. Westlaw is a subscription-based service used to facilitate ease of search and download, but the 10-K text is identical to that which

factors, major business initiatives, operations, and financial statements. We use the original 10-K filing where possible, if had at least 2,000 words (Loughran and McDonald, 2011).<sup>8</sup> We use an externally generated set of keywords to identify technologies and practices related to collection, storage, and analysis of data. The source for our keywords is the 2011 industry report by the McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition and Productivity*.<sup>9</sup> To create the set of keywords, we use all possible combinations of words and phrases (up to 4 words) that could describe usage of data, or data-related practices. We then developed a software program that searched the 10-Ks for the keywords. The program ignores punctuation, is not case sensitive, and accounts for word stemming. Thus, a search for the term “real time” will count *real time*, *real-time* and *Real Time*, and a search for “predictive model” will also count *predictive modeling*. In Table 1, we list a number of examples of our keywords that appeared in 10-Ks from our sample firms.

For each keyword, we draw a random sample of fifty 10-K documents in which that keyword appears to examine it in context. We determine whether the keyword primarily describes the *usage* of a technology or data practice, or, if the keyword primarily appears in a *non-usage* context (such as in product descriptions or in industry trends). If a keyword was classified in a *usage* context for at least three-quarters of the unique 10-Ks in the random sample,

---

can be found on the SEC website for free at <http://www.sec.gov/edgar.shtml>. We used the body of the 10-K, and did not include exhibits.

<sup>8</sup> If the original 10-K filing did not have at least 2,000 words, we use the first amended filing (if there was one) in which at least 2,000 words appeared for that fiscal year. This screen removes only about 0.5% of the sample, and is designed to remove 10-Ks with so few words that comparisons with longer 10-Ks are less meaningful. The word count is based on the body of the 10-K, and does not include additional exhibits, such as the Annual Report to Shareholders. Thus, firms with 10-K documents of less than 2,000 words each may be including their business information in exhibits.

<sup>9</sup> Available at: [http://www.mckinsey.com/mgi/publications/big\\_data/index.asp](http://www.mckinsey.com/mgi/publications/big_data/index.asp).

we consider the keyword a *usage* keyword for the population as a whole.<sup>10</sup> Our randomly drawn selection for each keyword comes from the final screened sample, and if a keyword appeared in less than fifty 10-Ks from unique firms, we examine all 10-Ks from unique firms during the sample period with that keyword. However, any keyword that appeared in less than five unique firms from 1996-2010 was not classified.<sup>11</sup> We list the results of our keyword classifications based on the randomly generated samples of 10-Ks in Table 2.

Using the Central Index Key (CIK) number associated with each filing entity, we matched the keyword counts from 10-Ks to financial data from *Compustat's* North America Fundamental Annual Database for the period 1996-2010. Our full, unscreened sample is a dataset of 104,070 observations that each have CIK identifiers, year, industry code, and a 10-K with 2,000 or more words. After the screening procedure described in detail in Appendix A, our final sample contains 37,918 observations from 1996-2010.

### 3. Model

We begin with the principle that the market value of the firm should be equal to the sum of its tangible and intangible assets (Hall 2000, 2001; Brynjolfsson, Hitt, and Yang 2002):

$$MV = A + I \tag{1}$$

where  $MV$  is the market value of the firm,  $A$  is the replacement value of the firm's tangible assets, and  $I$  is the value of the firm's intangible assets. If we divide both sides by  $A$ , and use the

---

<sup>10</sup> For each 10-K, if a keyword appeared exactly once, then the keyword was classified by the context of its single appearance. If a keyword appeared multiple times in a 10-K, we classified the keyword by the majority of *usage* or *non-usage* instances. It is worth noting that when a keyword appeared multiple times in a 10-K, the context of that keyword in terms of *usage* or *non-usage* was almost always the same throughout that 10-K. In a very rare case of a tie between *usage* and *non-usage* in the same 10-K, we scored the keyword as *usage* for that 10-K, since the technology, is in fact, being used by the firm, our primary consideration in our econometric analysis.

<sup>11</sup> As the sample used to classify the keyword would be too small to be considered reliable. We also do not classify the keyword "R", as the use of R language (the open source statistical programming language) by firms is clearly not the predominant context for "R" in 10-Ks.



identity that Tobin's  $q$  is the market value of the firm divided by the replacement value of the firm's tangible assets, we have

$$q = \frac{MV}{A} = 1 + \frac{I}{A} \quad (2)$$

In a simple case of a cross-section of firms, this would imply an estimating equation of the following form, with the addition of our keyword data to proxy for data-related intangible assets:

$$q_i = \alpha + \beta_1 \frac{I_{1i}}{A_i} + \beta_2 \frac{I_{2i}}{A_i} + \dots + \beta_n \frac{I_{ni}}{A_i} + \delta keywords_i + controls_i + u_i \quad (3)$$

where  $I_{1i}$ ,  $I_{2i}$ , through  $I_{ni}$  are the  $n$  intangible assets of the firm, such as R&D or brand, and  $keywords_i$  is a variable capturing intensity of data, with subscript  $i$  denoting firm  $i$ .

To account for industry and year effects, we subtract the average  $q$  for the industry in that year.<sup>12</sup> We use the most detailed industry code available with at least five firms in that year (starting with the 6-digit NAICS, and successively narrowing to 5, 4, 3, and then 2-digit NAICS until there are at least five firms). We use a variety of control variables, which take the form of ratios: Leverage (long-term debt to assets), capital expenditures to sales, earnings before interest and taxes (EBIT) to sales, and cash to balance sheet assets. These are designed to control for previous profitability, growth opportunities, and free cash flow that may affect excess value. We also include controls for firm size and age. We use long-difference specifications of varying lengths (from 1 to 5 years) as well as fixed-effects specifications, which remove time-invariant omitted variables.

In addition to a Tobin's  $q$  regression of excess value, we also examine firm performance using the same specifications and controls:<sup>13</sup>

---

<sup>12</sup> Which is equivalent to using an industry-year fixed effect, based on Lang and Stulz (1994) and Villalonga (2004a).

<sup>13</sup> Except the ratio of EBIT to sales, since profits are included in ROA.

$$ROA_i = \alpha + \delta keywords_i + controls_i + u_i \quad (4)$$

Our left-hand side performance variable is return on assets (ROA), which is the ratio of earnings before interest, taxes, depreciation and amortization (EBITDA) to balance sheet assets.

## 4. Analysis

### 4.1. Descriptive Statistics

Figure 1 illustrates the growing use of data-related keywords in 10-Ks from 1996-2010. By 2010, more than 40% of the firms in the sample were using at least one of the data-related keywords from Table 2 in their 10-Ks (whether or not it was in a usage-related context). Almost 30% of firms had a usage-related keyword in their 10-K. In Table 3, we summarize the variety of different usage keywords by year (regardless of how many times the keyword appears in the 10-K). In 1996, about 6% of the 10-Ks in the sample had one usage keyword, which grew to almost a quarter of the sample by 2010. The number of 10-Ks with two or more distinct usage keywords grew from almost zero in 1996 to more than 5% in 2010.

We then examine the industry patterns of keyword appearances in Table 4. We use the full, unscreened sample of 104,070 observations consisting of all *Compustat* firms with CIK identifiers, year, industry code, and 2,000 or more words in the 10-K. Not surprisingly, the usage keywords appear the most frequently in the Information sector (which includes software firms) as well as the Professional, Scientific, and Technical Services sector (which includes the IT services firms). By 2010, over 60% of firms in the Information industries used data-intensive keywords in their 10-Ks. However, operationally intensive industries, such as transportation, warehousing, utilities, and retail, also appear to be heavy users of data-intensive practices. Furthermore, in these data-intensive industries, the adoption rate for these practices appears to

have accelerated in the five-year period from 2005 to 2010. In the aggregate, approximately 35% of firms had at least one data usage keyword in their 10-Ks in 2010.

Although our dataset begins with 1996, our regression analysis focuses on the period 2001-2010 for two reasons. First, equity markets were in a technology-related bubble in the late 1990s. While we control for industry-year effects, it is questionable as to how accurately markets valued the use of technologies at that time. Second, the use of our selected keywords was sparse prior to 2001, making market value estimation less reliable in that period. In Table 5, we report descriptive statistics about the subsample from 2001-2010. The median firm in our sample had a market value of \$549 million, sales of about \$433 million, Tobin's  $q$  of 1.04, and a return on assets of 11.3% during that time. A detailed description of our variable construction can be found in Appendix A.

#### ***4.2. Long-Difference Specifications***

In Table 6, we present long-difference specifications of relative Tobin's  $q$  as a function of data usage keywords from 2001-2010. The independent variable being differenced is a dummy variable equal to 1 if 2+ distinct usage keywords ever appeared together in the same 10-K for that firm from 1996 until that year, and 0 otherwise. Therefore, the differenced independent variable takes the value of 1 if 2+ distinct usage keywords appeared in a firm's 10-K for the first time sometime during the difference period. Using a subsample of firms with five or more continuous years of data from 2001-2010, we find the greatest changes in relative Tobin's  $q$  occurs in the two-year difference estimates. In other words, our estimates indicate that the change in firms'  $q$  value if 2+ usage keywords appeared for the first time during the middle year or the end-year of the two-year period is almost 19% larger than that of their 6-digit industry

rivals ( $t=3.04$ ). This estimate grows smaller over time, but the effect is still practically and statistically significant through the end of 4 years.<sup>14</sup>

In Table 7, we perform long-difference specifications with relative profits (ROA minus the industry average) as a function of data usage keywords from 2001-2010. There is little change to relative ROA over short time periods, such as one or two years. However, we find a remarkable pattern of increasing returns over time, with statistically significant returns in years four and five. This observed pattern of returns is similar to an analysis of total factor productivity (TFP) conducted by Brynjolfsson and Hitt (2003). Using long difference specifications, they found the returns to IT were five times greater over longer time periods (5-7 years) than in shorter time periods (such as one year).

It is possible that firms that discuss data-related practices and technologies are fundamentally different than firms that do not, and this is not accounted for adequately through controls or through long-difference specifications that remove time-invariant factors. We therefore create a restricted sample, keeping firms that 1) had at least one keyword from Table 2 appear in its 10-K by the beginning of the difference period, and 2) after the appearance of such a keyword, had five or more continuous years of data from 2001-2010. This single keyword could be *usage*, *non-usage*, or even *unclassified*. Thus, the control group in Tables 8 and 9 consists solely of firms that had discussed some data-related practice or technology, but not intensively (since only one keyword had ever appeared in a previous 10-K, not two). Even on this restricted sample, we find significant increases in value and profits over time (although the effect on  $q$  is not as strong as before.)

---

<sup>14</sup> We perform the same analysis on the appearance of 1+ data usage keywords appearing rather than 2+ words. We find no excess value accruing to 1+ data words appearing over time. These findings suggest that a 2+ word variable captures intensity of data usage and is much less subject to measurement error than a 1+ keyword variable.

### 4.3. *Endogeneity*

In light of the correlation between the inclusion of data usage keywords and excess value and profits relative to the 6-digit NAICS industry, we need to consider the question of causality: Are data practices and technologies causing higher excess value, or, are more valuable firms the first to adopt new technologies? If this alternative scenario were the case, there would be a spurious statistical association between data usage and financial performance, and we would then expect value to rise before, not after, the appearance of these words in the 10-K. To provide evidence consistent with the causal direction of our argument, in Table 10 we more closely examine the timing of changes to market value that are associated with the adoption of data practices.<sup>15</sup>

In Table 10, we examine excess  $q$  just before and after the first appearance of two or more usage keywords in a firm's 10-K filing. We estimate three different specifications of excess  $q$ , based on three different dummy variables: 1) At one year before the first appearance of 2+ usage keywords in a firm's 10-K; 2) At the close of the fiscal year, approximately 60-90 days before the first appearance; and 3) Approximately 9-10 months after the first appearance, which is the end of the first fiscal year after the 10-K filing.<sup>16</sup>

We do not find a corresponding increase in market value *before* the first appearance of two or more distinct usage keywords. In fact, our point estimate is slightly negative one year beforehand, and almost zero 60-90 days before the 10-K filing is released. After the first appearance of the usage keywords however, such firms are worth 15.3% more than their

---

<sup>15</sup> We do not examine changes in relative ROA in Table 10, as we would not expect immediate changes in ROA based on keyword usage, whereas we would for market value.

<sup>16</sup> Historically, the deadline for 10-K filings was 90 days after the close of the fiscal year, although a shorter deadline for large companies was phased in to 60 days after the close of the fiscal year. The 10-Ks in our sample are always filed *after* the close of the fiscal year, with the earliest being approximately one month after the end of fiscal year.

competition, which is both practically and statistically significant (Column 4,  $t=1.86$ ). This is consistent with undervalued firms adopting data practices, or firms undertaking a set of costly organizational changes before adopting data practices, rather than an alternative story of the usage keywords merely being a proxy for highly-valued firms. We perform the same analysis using fixed effects estimation. One year before the first appearance of the keywords, these firms are worth about 6% less than their own average excess value during the sample period, whereas in the year immediately afterwards, they are worth about 17% more than their own average excess value ( $t=2.85$  in Column 8).

In Table 10, Columns 9-16, we run the same analyses as Columns 1-8 although we only retain firms that had at least one keyword from anywhere in Table 2 ever appear in its 10-K, and 2) on or after the appearance of such a keyword, had five or more continuous years of data from 2001-2010. We still find that excess value increases after, but not before the appearance of 2+ distinct usage keywords in a single 10-K. However, while in fixed effects the increase in market value after the appearance of usage keywords remains statistically significant (Column 16), in OLS, it does not (Column 12).

## **5. Conclusion**

In this study, we created a dataset of close to 40,000 observations based on a textual analysis of 10-K filings to measure the usage of data practices within firms from 1996-2010. We also analyzed how this shift towards the greater use of data in recent years has affected the market value and performance of firms. The results indicate that firms that have invested in data are rewarded with significantly higher market value, of approximately 10-20%, compared to their 6-digit NAICS competitors. We also find that these firms see significant increases in profitability (as measured by ROA), but not right away. Rather, it takes from three to five years

after the appearance of usage-related keywords for significant increases in profits to materialize. These results persist after including a number of control variables, when using long-difference and fixed-effects estimates that remove the effects of unobserved firm characteristics, and when dropping observations that have never used a single keyword before. Increases in value do not occur before the first appearance of keywords in the 10-K. However, there are significant and persistent gains to market value afterwards.

Our findings have important implications for the study of technology diffusion across firms and industries. As with previous generations of IT-related technological innovation, current investments in data are heavily concentrated within a few sectors where the skills and technologies required for handling data at such a massive scale are selected, refined, and standardized. However, as with prior technologies, these data-related skills will eventually diffuse across sectors, facilitated by a number of mechanisms such as consulting companies that specialize in big data practices or the mobility of new classes of workers (e.g., “data scientists”) who can introduce the new technologies and practices to firms in other sectors. The size and scope of the sample provide a unique opportunity to examine the diffusion of these technologies and analyze the performance effects of adoption even within fine-grained industry classifications.

Our main results—that Tobin’s  $q$  relative to the industry increases after the appearance of data usage keywords, but not before, and that ROA increases between 4-5 years after the appearance of such keywords, but not immediately—is based on a single document, per firm, per year (the 10-K). This is even more remarkable when one considers that disclosure of data-related technologies and practices is voluntary. Thus, even with the potential measurement error from this approach, our results are robust to a number of different specifications.

We see this paper as the first step in a much larger stream of research. We plan to extend this work by widening the scope of our text to include more SEC filings (such as 10-Qs and 8-Ks), as well as press releases, analyst reports, news stories, and other related sources. We will also examine a broader set of intangibles, including innovation, management quality, leadership, and brand. Furthermore, we plan to further refine our method for representing the intensity of data usage inside firms—currently based on whether certain keywords were mentioned in a 10-K—to better account for the context in which words appear. Such refinement is needed to establish elasticities of investment in new data technologies, which would enable us to make more precise statements about the economic contribution of these inputs to productivity and growth. Over time, we hope that this method will become an important resource for researchers interested in constructing firm-level datasets for IT and other intangible assets.



## Table 1. Examples of Data-Related Keywords in 10-Ks

Array BioPharma Inc., 2002 10-K, page 16. Filed September 30, 2002.

Predicting drug characteristics, such as potency, dosing frequency and potential side effects, requires powerful **data mining** and management tools. Our informatics teams comprise computational chemistry, scientific computing and medicinal chemistry experts who work together to increase the probability of creating a successful drug. Our proprietary software enables our scientists to search databases of existing drugs, to generate novel predictive databases and to create modeling programs designed to better forecast drug characteristics.

Cincinnati Financial Corporation. 2010 10-K, page 10. Filed February 25, 2011.

In the second half of 2009, we began to use **predictive modeling** tools that align individual insurance policy pricing to risk attributes for our workers' compensation line of business.

Coach, Inc. 2008 10-K, page 8. Filed August 21, 2008.

Complementing its ERP system are several other system solutions, each of which Coach believes is well suited for its needs. The **data warehouse** system summarizes the transaction information and provides a single platform for all management reporting.

Harrah's Entertainment, Inc. 2007 10-K, page 22. Filed February 29, 2008.

We believe that the customer-relationship marketing and **business-intelligence** capabilities fueled by Total Rewards, our customer loyalty program, are constantly bringing us closer to our customers so we better understand their preferences, and from that understanding, we are able to improve entertainment experiences we offer accordingly.

Owens-Illinois, Inc. 2010 10-K, page 1. Filed February 29, 2008.

The Company is pursuing a strategy aimed at leveraging its global capabilities, broadening its market base and focusing on modern management technologies and fundamentals including incentive compensation linked to cash flows and fact-based, **data-driven** decision making.

The Talbots, Inc. 2009 10-K, page 37. Filed April 16, 2009.

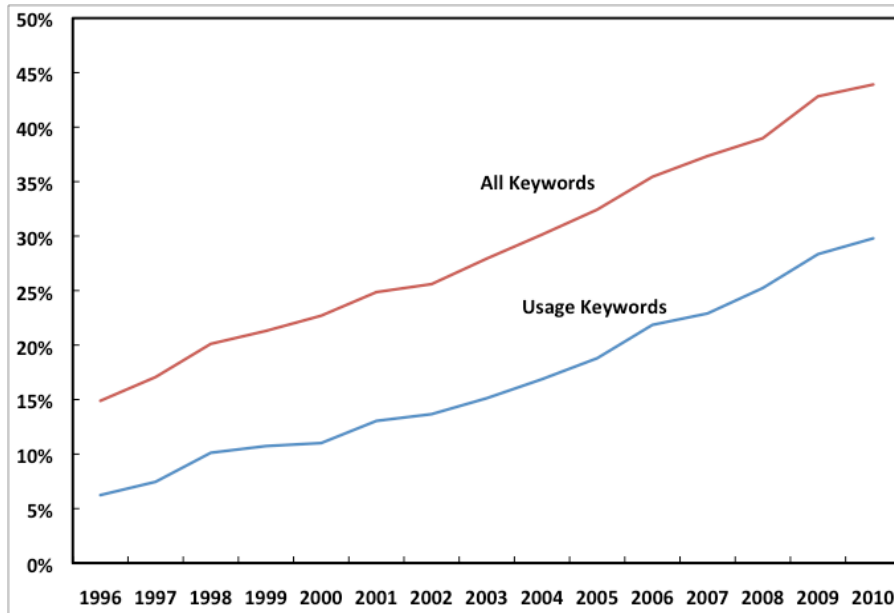
During 2008 and 2009, we took the following actions in an effort to seek to improve our gross margins: changed our promotional cadence to monthly markdowns rather than our historical four clearance sales events per year, held a leaner inventory position, concentrated on better product flow and content, and adopted a new **price optimization** tool.

**Table 2. Classification of Data-Related Keywords in the Screened Sample, 1996-2010**

Keyword	Percentage of 10-Ks from Unique Firms Classified as:			Word in Less Than 5 Unique Firms
	Usage	Selling	Other	
<b>Usage Keywords</b>				
analyze/analyzing customer	98	0	2	A/B/N test
automated decision	92	0	8	analytical talent
customer data	86	2	12	association rule learning
data analysis	76	13	11	big data
data center	91	0	9	bucket testing
data driven	87	5	8	cluster analysis
data mining	78	0	22	clustergram
data warehouse	94	0	6	crowdsourcing
neural network	80	10	10	ensemble learning
price/pricing optimization	75	13	13	extract, transform, and load
regression	80	0	20	genetic algorithm
simulation	76	10	14	Google File System
<b>Non-Usage Keywords</b>				
A/B Test	20	0	80	hadoop
advanced algorithms	50	25	25	Hbase
analytics	60	10	30	highly scalable computing resource
analyze/analyzing data	70	9	21	history flow
business intelligence	49	15	37	Internet of Things
Cassandra	0	0	100	large datasets
cloud computing	26	44	30	machine learning
collect data	64	10	26	MapReduce
combine data	63	11	26	market basket analysis
data fusion	17	83	0	mashup
data integration	26	44	30	mass collaboration
data mart	73	18	9	natural language processing
data model	67	13	21	non-relational database
digital data	32	28	40	predictive analytics
digital form	8	2	90	real-time management dashboard
digitization	42	34	24	real-time sales data
distributed system	33	67	0	semi-structured data
forecast sales	64	0	36	sentiment analysis
fraud detection	50	43	7	spatial analysis
network analysis	29	71	0	spatial information flow
pattern recognition	33	33	33	split testing
predictive model	73	11	16	supervised learning
real time	55	29	16	tag cloud
RFID	43	20	37	time series analysis
signal processing	49	22	30	time series forecast
Web 2.0	33	25	42	unstructured data
				unsupervised learning
				visualization

For each keyword, a random sample of fifty 10-Ks from unique firms from the screened *Compustat* sample of 37,918 observations was selected. Keywords were classified as *usage* terms if they appeared in that context in 75% or more of the random sample. If there were less than 50 filings from unique firms with the keyword, then all filings from unique firms with that keyword were used. Keywords were classified if they appeared in 5 or more unique firms in the sample.

**Figure 1. Data-Intensive Keywords in Screened Sample of 10-Ks, 1996-2010**



The sample is all *Compustat* firms based on the screening procedure described in the text, for a total of 37,918 observations.

**Table 3. Year-by-Year Totals of Data Usage Keywords in Sample**

Percentage of 10-Ks by Distinct Usage Keyword Occurrences							
Year	0	1	2	3	4	5	Observations
1996	93.73	5.79	0.48	---	---	---	2,917
1997	92.52	6.79	0.60	0.06	0.03	---	3,180
1998	89.90	8.99	0.92	0.16	0.03	---	3,060
1999	89.43	9.04	1.14	0.28	0.10	---	2,886
2000	89.63	8.58	1.40	0.23	0.12	0.04	2,565
2001	87.11	11.10	1.47	0.20	0.08	0.04	2,451
2002	86.38	11.92	1.43	0.20	0.04	0.04	2,525
2003	84.91	13.09	1.50	0.42	0.04	0.04	2,406
2004	83.40	13.96	2.05	0.42	0.17	---	2,386
2005	81.16	15.98	2.23	0.55	0.08	---	2,372
2006	78.28	17.66	3.52	0.46	0.04	0.04	2,389
2007	76.92	18.72	3.97	0.30	0.04	0.04	2,318
2008	74.74	20.88	3.75	0.59	0.05	---	2,189
2009	71.63	23.60	4.04	0.69	0.05	---	2,178
2010	70.66	23.43	5.10	0.62	0.19	---	2,096
<b>1996-2010 Total</b>							<b>37,918</b>
<b>Addenda: 2001-2010 Total</b>							<b>23,310</b>

The sample is all *Compustat* firms based on the screening procedure described in the text. Usage keywords are listed by *variety* in each 10-K, regardless of the total occurrences of that keyword in the 10-K.

**Table 4. Variety of Data Usage Keywords in 10-Ks by NAICS sector and Year**

NAICS	Sector Name	1996			2000			2005			2010		
		0	1	2+	0	1	2+	0	1	2+	0	1	2+
11	Agriculture, Forestry, Fishing and Hunting	100	0	0	100	0	0	94.44	5.56	0	88.89	11.11	0
21	Mining, Quarrying, and Oil and Gas Extraction	95.88	3.74	0.37	95.09	4.53	0.38	89.64	10.00	0.36	72.44	25.09	2.47
22	Utilities	98.94	1.06	0	96.51	2.91	0.58	84.30	14.53	1.16	57.45	35.46	7.09
23	Construction	100	0	0	95.65	4.35	0	90.00	10.00	0	85.00	13.33	1.67
31-33	Manufacturing	92.31	7.26	0.43	89.15	9.96	0.89	84.45	13.57	1.98	75.14	21.00	3.86
42	Wholesale Trade	95.87	3.49	0.63	92.81	5.76	1.44	81.87	17.62	0.52	78.32	18.18	3.50
44-45	Retail Trade	94.40	5.60	0	91.12	6.18	2.70	83.04	14.78	2.17	65.80	28.50	5.70
48-49	Transportation and Warehousing	92.16	7.84	0	87.31	11.19	1.49	81.01	15.82	3.16	66.41	25.78	7.81
51	Information	76.30	20.78	2.92	65.98	27.84	6.19	60.29	32.00	7.71	39.59	41.42	18.99
52	Finance and Insurance	85.18	14.16	0.66	75.72	22.35	1.93	58.57	37.15	4.28	46.42	43.33	10.25
53	Real Estate and Rental and Leasing	95.00	5.00	0	94.22	4.62	1.16	87.10	12.10	0.81	76.44	19.56	4.00
54	Professional, Scientific, and Technical Services	77.05	20.33	2.62	65.79	27.63	6.59	59.75	30.96	9.29	46.15	39.82	14.03
55	Management of Companies and Enterprises	---,--	---,--	---,--	---,--	---,--	---,--	---,--	---,--	---,--	---,--	---,--	---,--
56	Administrative, Support, Waste Mgmt and Remediation Svcs.	92.00	6.86	1.14	80.58	17.48	1.94	78.57	17.14	4.29	63.16	30.53	6.32
61	Educational Services	85.71	14.29	0	78.13	21.88	0	79.17	16.67	4.17	42.86	53.57	3.57
62	Health Care and Social Assistance	96.32	3.68	0	88.89	10.42	0.69	81.97	17.21	0.82	75.79	17.89	6.32
71	Arts, Entertainment and Recreation	95.45	4.54	0	94.29	5.71	0	91.67	8.33	0	89.74	2.56	7.69
72	Accommodation and Food Services	94.35	4.62	1.03	95.00	5.00	0	92.11	7.89	0	80.95	16.67	2.38
81	Other Services (except Public Administration)	94.59	5.41	0	92.31	5.13	2.56	88.89	11.11	0	76.47	23.53	0
99	Unclassified	100	0	0	96.77	3.23	1.10	96.67	2.50	0.83	89.29	10.71	0
<b>Average</b>		<b>90.02</b>	<b>9.23</b>	<b>0.74</b>	<b>83.09</b>	<b>14.74</b>	<b>2.17</b>	<b>76.01</b>	<b>20.78</b>	<b>3.21</b>	<b>64.68</b>	<b>28.38</b>	<b>6.93</b>

N=104,070 observations. The sample is all firms in *Compustat* with CIK identifiers, year, industry code, and 10-K data (2,000 or more words in the 10-K). The totals in each cell represent the percentage of 10-Ks in that industry and year that contain either zero, one, or two or more *different* usage keywords.

**Table 5. Descriptive Statistics of 2001-2010 Subsample**

Variable	Mean	Median	Minimum	25th Percentile	75th Percentile	Maximum	Std. Deviation
Market Value (\$mm)	3,987.9	549.1	2.36	131.2	2,111.9	513,805.6	15,802.4
Tobin's <i>q</i>	1.35	1.04	.074	.748	1.58	24.64	1.02
Return on Assets (EBITDA/Assets)	.103	.113	-.701	.063	.167	.589	.123
Balance sheet Assets (\$mm)	3,106.8	432.3	10.0	111.1	1,746.7	304,594	11,808.4
Sales (\$mm)	2,922.2	433.0	10.0	99.1	1,650.1	425,071	12,936.8
R&D/Assets	.103	0	0	0	.065	9.66	.314
Advertising Stock/Assets	.025	0	0	0	.016	3.50	.083
EBIT/Sales	.015	.068	-32.79	.017	.133	.975	.545
CAPX/Sales	.103	.034	0	.016	.075	13.48	.313
Leverage (Long-Term Debt/Assets)	.211	.170	0	.017	.321	5.75	.223
Cash/Assets	.102	.056	0	.018	.137	.997	.126

N=23,310 observations from 2001-2010. The subsample is from the 1996-2010 *Compustat* sample of 37,918 observations.

**Table 6. Long Difference OLS Regressions of Relative Tobin's  $q$  as a Function of Data Usage Keywords, 2001-2010**

	(1)	(2)	(3)	(4)	(5)
Number of Years Differenced	1	2	3	4	5
<b>2+ different usage keywords ever appearing (D=1)</b>	.104* (.056)	.187*** (.061)	.149*** (.057)	.112* (.060)	.034 (.067)
<b>Number of Observations</b>	13,560	11,836	10,094	8,335	6,553
<b>R-squared</b>	.029	.033	.038	.039	.048

The sample is all *Compustat* firms based on the screening procedure described in the text, and five or more continuous years of data from 2001-2010. The dependent variable being differenced is Tobin's  $q$  minus the average Tobin's  $q$  for the most detailed NAICS industry with at least 5 firms in that year (starting at the 6-digit NAICS level). Controls (that are differenced) include R&D Assets/Total Assets, Advertising Assets/Total Assets, No R&D, No Advertising, Leverage, Size, CAPX/Sales, EBIT/Sales, and Cash/Total Assets. Firm Age is included and not differenced. Robust standard errors clustered by firm are in parentheses.

\*significant at the 10% level, \*\*significant at the 5% level, \*\*\*significant at the 1% level.

**Table 7. Long Difference OLS Regressions of Relative Return on Assets (ROA) as a Function of Data Usage Keywords, 2001-2010**

	(1)	(2)	(3)	(4)	(5)
Number of Years Differenced	1	2	3	4	5
<b>2+ different usage keywords ever appearing (D=1)</b>	.001 (.005)	.005 (.005)	.010 (.006)	.014** (.007)	.014* (.008)
<b>Number of Observations</b>	13,560	11,836	10,094	8,335	6,553
<b>R-squared</b>	.033	.036	.040	.047	.044

The sample is all *Compustat* firms based on the screening procedure described in the text, and five or more continuous years of data from 2001-2010. The dependent variable being differenced is Return on Assets (ROA) minus the average ROA for the most detailed NAICS industry with at least 5 firms in that year (starting at the 6-digit NAICS level). Controls (that are differenced) include R&D Assets/Total Assets, Advertising Assets/Total Assets, No R&D, No Advertising, Leverage, Size, CAPX/Sales, and Cash/Total Assets. Firm Age is included and not differenced. Robust standard errors clustered by firm are in parentheses.

\*significant at the 10% level, \*\*significant at the 5% level, \*\*\*significant at the 1% level.

**Table 8. Long Difference OLS Regressions of Relative Tobin's  $q$   
as a Function of Data Usage Keywords,  
Firms with Previous Keyword Usage before Difference Period, 2001-2010**

	(1)	(2)	(3)	(4)	(5)
Number of Years Differenced	1	2	3	4	5
<b>2+ different usage keywords ever appearing (D=1)</b>	.036 (.048)	.091* (.053)	.070 (.060)	.043 (.066)	-.029 (.071)
<b>At least one keyword from Table 2 appeared in 10-K by the beginning of difference period</b>	✓	✓	✓	✓	✓
<b>Number of Observations</b>	5,471	4,880	4,240	3,556	2,824
<b>R-squared</b>	.028	.038	.049	.050	.068

The sample is all *Compustat* firms based on the screening procedure described in the text, and further restricted to firms that had at least one keyword from Table 2 appear in its 10-K by the beginning of the difference period. Any firm with five or more continuous years of data from 2001-2010 after the appearance of a single keyword are retained. The dependent variable being differenced is Tobin's  $q$  minus the average Tobin's  $q$  for the most detailed NAICS industry with at least 5 firms in that year (starting at the 6-digit NAICS level). Controls (that are differenced) include R&D Assets/Total Assets, Advertising Assets/Total Assets, No R&D, No Advertising, Leverage, Size, CAPX/Sales, EBIT/Sales, and Cash/Total Assets. Firm Age is included and not differenced. Robust standard errors clustered by firm are in parentheses.

\*significant at the 10% level, \*\*significant at the 5% level, \*\*\*significant at the 1% level.

**Table 9. Long Difference OLS Regressions of Relative Return on Assets (ROA)  
as a Function of Data Usage Keywords,  
Firms with Previous Keyword Usage before Difference Period, 2001-2010**

	(1)	(2)	(3)	(4)	(5)
Number of Years Differenced	1	2	3	4	5
<b>2+ different usage keywords ever appearing (D=1)</b>	-.002 (.006)	.004 (.005)	.012* (.006)	.015** (.007)	.015* (.009)
<b>At least one keyword from Table 2 appeared in 10-K by the beginning of difference period</b>	✓	✓	✓	✓	✓
<b>Number of Observations</b>	5,471	4,880	4,240	3,556	2,824
<b>R-squared</b>	.038	.048	.055	.065	.053

The sample is all *Compustat* firms based on the screening procedure described in the text, and further restricted to firms that had at least one keyword from Table 2 appear in its 10-K by the beginning of the difference period. Any firm with five or more continuous years of data from 2001-2010 after the appearance of a single keyword are retained. The dependent variable being differenced is Return on Assets (ROA) minus the average ROA for the most detailed NAICS industry with at least 5 firms in that year (starting at 6-digit NAICS). Controls (that are differenced) include R&D Assets/Total Assets, Advertising Assets/Total Assets, No R&D, No Advertising, Leverage, Size, CAPX/Sales, and Cash/Total Assets. Firm Age is included and not differenced. Robust standard errors clustered by firm are in parentheses.

\*significant at the 10% level, \*\*significant at the 5% level, \*\*\*significant at the 1% level.

**Table 10. Relative Tobin's  $q$  and the Timing of the First Appearance of 2+ Usage Keywords, 2001-2010**

Specification	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) FE	(6) FE	(7) FE	(8) FE
<b>Variable</b>								
One year before first appearance of 2+ distinct usage keywords in 10-K (D=1)	-.066 (.064)	-.042 (.064)			-.035 (.053)	-.062 (.053)		
End of fiscal year, approx. 60-90 days prior to first appearance of 2+ distinct usage keywords in 10-K (D=1)			.006 (.068)				.023 (.049)	
Approx. 9-10 months after first appearance of 2+ distinct usage keywords in 10-K (D=1)				.153* (.083)				.174*** (.061)
Had at least five continuous years of data from 2001-2010		✓	✓	✓		✓	✓	✓
<b>Number of Observations</b>	15,226	12,377	14,159	13,893	15,226	12,377	14,159	13,893
<b>R-squared</b>	.042	.040	.038	.039	.046	.042	.045	.042
Specification	(9) OLS	(10) OLS	(11) OLS	(12) OLS	(13) FE	(14) FE	(15) FE	(16) FE
<b>Variable</b>								
One year before first appearance of 2+ distinct usage keywords in 10-K (D=1)	-.054 (.069)	-.043 (.077)			.004 (.061)	-.015 (.064)		
End of fiscal year, approx. 60-90 days prior to first appearance of 2+ distinct usage keywords in 10-K (D=1)			-.076 (.062)				.007 (.045)	
Approx. 9-10 months after first appearance of 2+ distinct usage keywords in 10-K (D=1)				.083 (.080)				.128** (.059)
Had at least five continuous years of data from 2001-2010		✓	✓	✓		✓	✓	✓
Used at Least One Keyword from Table 2 Previously	✓	✓	✓	✓	✓	✓	✓	✓
<b>Number of Observations</b>	6,748	4,915	5,647	5,544	6,748	4,915	5,647	5,544
<b>R-squared</b>	.044	.038	.042	.043	.041	.044	.051	.052

The sample is all *Compustat* firms based on the screening procedure described in the text. The dependent variable is Tobin's  $q$  minus the average Tobin's  $q$  for the most detailed NAICS industry with at least 5 firms in that year (starting at 6-digit NAICS). Controls include R&D Assets/Total Assets, Advertising Assets/Total Assets, No R&D, No Advertising, Leverage, Age, Size, CAPX/Sales (current and lagged to 3 years), EBIT/Sales (current and lagged to 3 years), and Cash/Total Assets (current and lagged to 3 years). Robust standard errors clustered by firm are in parentheses in Columns 1-4 and 9-12, robust standard errors are in parentheses in Columns 5-8 and 13-16.

\*significant at the 10% level, \*\*significant at the 5% level, \*\*\*significant at the 1% level.



## Appendix A

### A.1. Detailed Variable Construction

#### A.1.1. Tobin's $q$ Relative to Industry

The method we use to construct Tobin's  $q$ —the market value of the firm divided by the replacement cost of its tangible assets—is based on the Lindenberg and Ross (1981) algorithm, modified by Smirlock, Gilligan, and Marshall (1984), and subsequently used by Lang and Stulz (1994), McGahan (1999), and Villalonga (2004b).

The numerator, market value, is constructed as the price of common stock at the end of the fiscal year times the number of common shares outstanding, plus the value of preferred stock, plus long-term debt and short-term debt. (Using *Compustat* mnemonics, this is equivalent to:  $PRCC\_F * CSHO + PSTK + DLTT + DLC$ ).

The denominator of Tobin's  $q$  is the sum of three parts. The first is the replacement cost of property, plant, and equipment (PP&E). As described in Villalonga (2004b), the replacement cost of a firm's PP&E is assumed to be equal to its book value in 1970 (net PP&E, *Compustat* mnemonic PPENT) or the first year the firm appears in *Compustat*, whichever is later. The replacement value from the previous year is reduced by a 5 percent rate of depreciation, and then adjusted for price changes according to the GDP deflator for nonresidential fixed assets.<sup>17</sup> The change in the book value of gross PP&E (*Compustat* mnemonic PPEGT) during the year is then added to last year's replacement-cost PP&E.<sup>18</sup> Finally, the technological change parameter is

---

<sup>17</sup> See Bureau of Economic Analysis, National Income and Product Accounts Table 5.3.4, "Price Indexes for Private Fixed Investment by Type," line 2, available at <http://www.bea.gov>.

<sup>18</sup> We also account for the possibility that major reductions to gross PP&E (for example, if a large amount of capital is retired and thus taken out of service) could result in negative imputed values for the firm's replacement cost of capital (since the change in gross stock of PP&E in the current year is added to last year's adjusted replacement cost PP&E). In such cases, we reset the value of replacement cost of PP&E to be equal to net PP&E as reported by the firm, and continue the algorithm from there for future years.

assumed to be zero (Smirlock, Gilligan, and Marshall 1984; Lang and Stulz 1994; McGahan 1999; Villalonga 2004b).

The second part of the denominator is the replacement cost of inventories. We adjust the book value of inventories (INVT in *Compustat*) according to the method outlined in Lindenberg and Ross (1981), McGahan (1999) and Villalonga (2004b). We assume that the replacement cost of inventory is the same as the book value of inventory as reported by the firm in 1970 or the first year in which the firm appears in *Compustat*, whichever is later. We then account for annual inflation depending on the inventory valuation method used by the firm (different methods apply whether the firm uses FIFO, LIFO, average cost, or retail cost accounting), and the price deflator for inventories.<sup>19</sup> In the absence of data about the method of inventory valuation, we assume the replacement cost of inventory is equal to its book value.

Finally, the third part of the denominator is all other assets of the firm on the balance sheet, which we assume are listed at replacement cost. Thus, Tobin's  $q$  is the market value of the firm divided by the following: balance sheet assets reported by the firm (mnemonic AT), minus net PP&E as reported by the firm (PPENT) and the book value of inventories (INVT), plus the replacement cost of PP&E and the replacement cost of inventories.

Tobin's  $q$  relative to the industry is the firm's Tobin's  $q$  minus the average Tobin's  $q$  in the 6-digit NAICS industry in that year, if there are at least five firms to compute this average. If there aren't at least five firms at the 6-digit NAICS level, we move successively to the 5-digit, 4-digit, 3-digit, and 2-digit NAICS level until there are at least five firms in that year to compute the average industry  $q$ .<sup>20</sup>

---

<sup>19</sup> See Bureau of Economic Analysis, National Income and Product Accounts Tables 5.7.9A and 5.7.9B, "Implicit Price Deflators for Private Inventories by Industry," available at <http://www.bea.gov>.

<sup>20</sup> We drop a small handful of observations in which there were less than 5 firms at the 2-digit NAICS level in a given year.

### **A.1.2. Return on Assets (ROA) Relative to Industry**

Return on Assets (ROA) is calculated as the ratio of EBITDA (earnings before interest, taxes, depreciation and amortization, or EBITDA in *Compustat*) divided by total balance sheet assets (AT). ROA relative to the industry is the firm's ROA minus the average ROA for the 6-digit NAICS industry in that year, if there are at least five firms to compute this average. If there aren't at least five firms at the 6-digit NAICS level in that year, we move successively to the 5-digit, 4-digit, 3-digit, and 2-digit NAICS level until there are at least five firms in that year to compute the average.

### **A.1.3. Keywords**

We construct a set of dummy variables based on data usage keywords as identified in Table 2. Because 10-K filings will have repeated instances of sentences, or in some cases, entire paragraphs, measuring intensity by the overall number of appearances of the keyword in the document could be fraught with measurement error. Thus, as a measure of intensity, we rely on the variety of keywords that appear in the 10-K, regardless of the number of occurrences of those keywords in the 10-K.<sup>21</sup>

For the long-difference specifications, we begin by creating a dummy variable equal to 1 if two or more distinct usage keywords ever appeared in a single 10-K for the firm from 1996 up to and including that year. This assumes that once a data usage keyword appears in a 10-K, the firm continues to use that technology or practice henceforth, even if it does not discuss it in the 10-K every year.<sup>22</sup> Thus, if an  $n$ -year difference of this variable is equal to 1, then 2+ keywords

---

<sup>21</sup> When we encounter a 10-K with more than 20 repetitions of a keyword, we drop that observation and all future observations from that firm.

<sup>22</sup> Some firms include data-related keywords in multiple years, while others use them only once. In future work, we plan to further examine the differences between those firms.

appeared in its 10-K for the very first time at some point between the beginning and end of that difference period.

We also generate a number of dummy variables in order to examine the timing of the appearance of keywords and excess  $q$  in Table 10. We set a dummy variable equal to 1 the first time that 2+ distinct usage keywords appeared in a firm's 10-K, and zero otherwise. We create similar dummies for the year before, and the year after the first appearance of two or more distinct usage keywords in a firm's 10-K.

#### **A.1.4. Research and Development (R&D) Assets**

We construct R&D assets for each firm by converting nominal R&D spending as reported by the firm (XRD in *Compustat*) into constant-dollar flows. For a price deflator, we use the input price series available from the Bureau of Economic Analysis.<sup>23</sup> We use a 15% depreciation rate (Hall, Jaffe, and Trajtenberg, 2005), and, as an initial value, we begin with the first available year of R&D spending, or 1970, whichever is later. Firms are required to report R&D spending if it exceeds 1% of sales (Zhao 2002), and approximately half of our sample reports R&D spending. Thus, we assume a value of 0 for R&D spending if the firm does not report R&D. We create a dummy variable (No R&D) equal to 1 if R&D assets are equal to 0.

#### **A.1.5. Advertising Assets**

Similar to R&D, we convert nominal advertising spending as reported by the firm (XAD in *Compustat*) into constant-dollar flows and then aggregate those flows together into an advertising asset. We begin with the first available year that the firm reports advertising

---

<sup>23</sup> See Table 4.1, "Aggregate Input Price Indexes for R&D Investment, 1959-2007," available at: <http://www.bea.gov/national/newinnovation.htm>. We extrapolate this series through 2009 by assuming that its growth rate in 2008 and 2009 is the same as the rate of growth in the gross output deflator of NAICS 5417, Scientific Research and Development Services, taken from "Gross Output by Industry in Current Dollars, Quantity Indexes by Industry, Price Indexes by Industry," available at [http://www.bea.gov/industry/gdpbyind\\_data.htm](http://www.bea.gov/industry/gdpbyind_data.htm). We then assume the growth rate of this series in 2010 is the average of the growth rates for 2008 and 2009.

spending, or 1970, whichever is later. We blend three price series together to form a price deflator for advertising<sup>24</sup> and assume an annual 45% rate of depreciation (Villalonga 2004b). Firms are required to report advertising spending if it is material, and approximately one-third of our sample firms reports advertising spending. We assume a value of 0 for advertising spending if the firm does not report it. Finally, we add a dummy variable (No Advertising) equal to 1 if advertising assets are equal to 0.

#### **A.1.6. Other Controls**

We use a number of other variables that are designed to control for previous profitability, growth opportunities, or free cash flow that may affect excess value (Berger and Ofek 1995; Campa and Kedia 2002). We use leverage, which is the ratio of long-term debt (*Compustat* mnemonic DLTT) to total balance sheet assets (AT). We also use the ratios of capital expenditure (CAPX) to sales (SALE), earnings before interest and taxes (EBIT) to sales, and cash (CH) to balance sheet assets. We also include controls for firm size (the logarithm of total balance sheet assets) and age (proxied as the logarithm of one plus the number of years the firm has been listed in *Compustat*, a series that dates back to 1950).

#### **A.2. Detailed Screening Procedure**

Our full, unscreened sample is a dataset of 104,070 observations that each have CIK identifiers, year, industry code, and a 10-K with 2,000 or more words. Our screening procedure

---

<sup>24</sup> For the period 1970-1976, we use the BEA Output Price Index for Professional, Scientific and Technical Services. For the period 1977-1994, we use the BEA Output Price Index for Miscellaneous Technical Services. These price deflators are available from “Value Added by Industry in Current Dollars, Quantity Indexes by Industry, and Price Indexes by Industry, 1947-1997; Value Added by Industry, Gross Output by Industry, Intermediate Inputs by Industry, and the Components of Value Added by Industry, 1987-1997” at [http://www.bea.gov/industry/gdpbyind\\_data.htm](http://www.bea.gov/industry/gdpbyind_data.htm). For the period 1995-2010, we use the Producer Price Index (PPI) for Advertising Agencies, series PCU541810541810 from the Bureau of Labor Statistics, available at <http://www.bls.gov/ppi/>.

is as follows: We drop observations with negative or missing values of market value,<sup>25</sup> balance sheet assets, sales, cash, capital expenditures, property, plant and equipment, inventories, or employment,<sup>26</sup> as well as observations with missing values of industry code or profits.<sup>27</sup> We also drop observations with less than \$10 million of assets or sales to eliminate potential financial anomalies, as well as to ensure thick enough trading volume for reasonable estimates of market value.<sup>28</sup> We drop a handful of firms that either have unclassifiable industry codes (NAICS 99), or belong to 2-digit NAICS sectors with less than 5 observations in that year. In keeping with the prevailing practice of the literature, we drop financial firms (NAICS 52), outliers with extreme Tobin's  $q$  of less than 25% or greater than 400% of the industry average  $q$ , and observations with ROA relative to industry below the 1<sup>st</sup> percentile or greater than the 99<sup>th</sup> percentile (Berger and Ofek 1995; Campa and Kedia 2002; Villalonga 2004a; Giroud et al. 2012).

Since the focus of this work is on how the *usage* of technologies and data practices is associated with market value and performance, we also drop firms that are IT producers. Keywords that appear in an IT-producing firm's 10-K are much more likely to be in sales or product-related contexts. Therefore, we drop firms with primary NAICS codes in the hardware, software, telecommunications, and IT services industries (NAICS 334, 5112, 5133, 5415, 517-519). Since some IT-producing firms are not primarily classified in the IT sector but nonetheless

---

<sup>25</sup> Constructed as the sum of the market value of equity, preferred stock, and debt.

<sup>26</sup> Although we do not use employment in our analysis, we drop firms with zero employment as they may be holding companies that do not produce goods or services.

<sup>27</sup> Profits defined either as earnings before interest and taxes (EBIT) or earnings before interest, taxes, and depreciation (EBITDA).

<sup>28</sup> We adopt this criteria in the spirit of McGahan and Porter (1999, 2003) who apply the \$10 million asset and sales threshold to business segments instead of firms. Lang and Stulz (1994) eliminate firms that average less than \$100 million in assets, McGahan (1999) eliminates firms with less than \$50 million in assets, and Villalonga (2004b) eliminates firms that never reach \$50 million in assets in the sample period.

sell IT products or services, we also drop observations with more than 20 instances of any keyword as well as all future observations from that firm.<sup>29</sup> After the entire screening procedure, our final sample contains 37,918 observations from 1996-2010.

---

<sup>29</sup> Since information about the firm's products and services is an important part of the 10-K, keywords that appear more than 20 times in a single 10-K are very likely describing the firm's products, services, or markets instead of the firm's usage of data. Including these firms (which, in one case in the unscreened data, have as many as 318 repetitions of a keyword in a single 10-K) could introduce significant measurement error. About 0.5% of the observations were removed because a current or previous 10-K had more than 20 instances of a data-related keyword.

## References

- Aral, S., Brynjolfsson, E., Wu, D.J. 2006. Which Came First, IT or Productivity? The Virtuous Cycle of Investment and Use in Enterprise Systems. *ICIS 2006 Proceedings, Paper 110*. Available at: <http://aisel.aisnet.org/icis2006/110>.
- Aral, S., Weill, P. 2007. IT Assets, Organizational Capabilities, and Firm Performance: How Resource Allocations and Organizational Differences Explain Performance Variation. *Organization Science*. 18(5) 763-780.
- Archak, N., Ghose A., Ipeirotis, P.G. 2011. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*. 57(8) 1485-1509.
- Arrow, K.J. 1974. *The Limits of Organization*. New York: Norton.
- Barua, A., Mani, D. Mukherjee, R. 2010. Measuring the Business Impacts of Effective Data. Available at <http://www.sybase.com>.
- Bloom, N., Garicano, L., Sadun, R., Van Reenen, J. 2011. The Distinct Effects of Information Technology and Communication Technology on Firm Organization. Working paper.
- Bresnahan, T. F., Brynjolfsson E., Hitt, L. M. 2002. Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence. *Quarterly Journal of Economics*. 117(1) 339-376.
- Brynjolfsson, E., Hitt., L.M. 1996. Paradox Lost? Firm-level Evidence on the Returns to Information Systems Spending. *Management Science*. 42(4) 541-558.
- Brynjolfsson, E., Hitt, L.M. 2003. Computing Productivity: Firm-Level Evidence. *Review of Economics and Statistics*. 85(4) 793-808.
- Brynjolfsson, E., Hitt, L., Kim, H. 2011. Strength in Numbers: How does Data Driven Decisionmaking Affect Firm Performance? Available at SSRN: <http://ssrn.com/abstract=1819486>.
- Brynjolfsson, E., Hitt, L., Yang, S. 2002. Intangible Assets: Computers and Organizational Capital. *Brookings Papers on Economic Activity*. 2002(1) 137-181.
- Brynjolfsson, E., McAfee, A. 2011. The Big Data Boom is the Innovation Story of our Time. *The Atlantic Monthly*. November. Available at <http://www.theatlantic.com/business/archive/2011/11/the-big-data-boom-is-the-innovation-story-of-our-time/248215/>.
- Brynjolfsson, E., Saunders, A. 2010. *Wired for Innovation: How Information Technology is Reshaping the Economy*. MIT Press.



- Chwelos, P., Ramirez, R., Kraemer, K., Melville, N. 2009. Research Note—Does Technological Progress Alter the Nature of Information Technology as a Production Input? New Evidence and New Results. *Information Systems Research*. 21(2) 392-408.
- Corrado, C., Hulten, C. 2010. How do you Measure a Technological Revolution? *American Economic Review*. 100(2) 99-104.
- Corrado, C., Lane, J. 2009. Using Cyber-Enabled Transaction Data to Study Productivity and Innovation in Organizations. Global COE Hi-Stat Discussion Paper Series.
- Dewan, S., Min, C. 1997. The Substitution of Information Technology for Other Factors of Production: A Firm Level Analysis. *Management Science*. 43(12) 1660-1675.
- Forman, C., Goldfarb, A., Greenstein, S. 2002. Digital Dispersion: An Industrial and Geographic Census of Commercial Internet Use. NBER Working Paper #9287.
- Forman, C., Goldfarb, A., Greenstein, S. 2005. How Did Location Affect Adoption of the Commercial Internet: Global Village vs. Urban Leadership. *Journal of Urban Economics*. 58(3) 389-420.
- Forman, C., Goldfarb, A., Greenstein, S. 2008. Understanding the Inputs into Innovation: Do Cities Substitute for Internal Firm Resources? *Journal of Economics and Management Strategy*. 17(2) 295-317.
- Forman, C., Goldfarb, A., Greenstein, S. 2012. The Internet and Local Wages: A Puzzle. *American Economic Review*. 102(1) 556-75.
- Galbraith, J.R. 1974. Organization Design: An Information Processing View. *Interfaces*. 4(3) 28-36.
- Ghose, A., Ipeirotis, P.G., Li, B. 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content. *Marketing Science*. 31(3) 493-520.
- Giroud, X., Mueller, H.M., Stomper, A., Westerkamp, A. Snow and Leverage. *The Review of Financial Studies*. 25(3) 680-710.
- Hall, B.H., Jaffe, A., Trajtenberg, M. 2005. Market Value and Patent Citations. *The RAND Journal of Economics*. 36(1) 16-38.
- Hall, R. E. 2000. E-Capital: The Link between the Stock Market and the Labor Market in the 1990s. *Brookings Papers on Economic Activity*. 2000(2) 73-118.
- Hall, R. E. 2001. The Stock Market and Capital Accumulation. *The American Economic Review*. 91(5) 1185-1202.

- Hitt, L.M., Brynjolfsson, E. 1996. Productivity, Business Profitability, and Consumer Surplus: Three Different Measures of Information Technology Value. *MIS Quarterly*. 20(2) 121-142.
- 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.
- Hopkins, M. 2010. Interview with Erik Brynjolfsson: The Four Ways IT is Revolutionizing Innovation. *MIT Sloan Management Review*. 51(3) 51-56.
- Lang, L.H.P., Stulz, R.M. 1994. Tobin's  $q$ , Corporate Diversification, and Firm Performance. *The Journal of Political Economy*. 102(6) 1248-1280.
- Lee, S., Song, J., Baker, J., Kim, Y., Wetherbe, J.C. 2011. The Commoditization of IT: Evidence from a Longitudinal Text Mining Study. *Communications of the Association for Information Systems*. 29(12) 221-242.
- Lichtenberg, F.R. 1995. The Output Contributions of Computer Equipment and Personnel: A Firm-Level Analysis. *Economics of Innovation and New Technology*. 3(3-4) 201-218.
- Lindenberg, E.B., Ross, S.A. 1981. Tobin's  $q$  Ratio and Industrial Organization. *Journal of Business*. 54(1) 1-32.
- Loughran, T., McDonald, B. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*. 66(1) 35-65.
- Loveman, G. 2003. Diamonds in the Data Mine. *Harvard Business Review*. 81(5) 109-113.
- McElheran, K.S. 2011. Do Market Leaders Lead in Business Process Innovation? The Case(s) of E-Business Adoption. *Harvard Business School Working Paper* 10-104.
- McGahan, A.M. 1999. The Performance of U.S. Corporations: 1981-1994. *The Journal of Industrial Economics*. 47(4) 373-398.
- McGahan, A.M., Porter, M.E. 1999. The Persistence of Shocks to Profitability. *The Review of Economics and Statistics*. 81(1) 143-153.
- McGahan, A.M., Porter M.E. 2003. The Emergence and Sustainability of Abnormal Profits. *Strategic Organization*. 1(1) 79-108.
- McKinsey Global Institute. 2011. Big Data: The Next Frontier for Innovation, Competition and Productivity. Available at: [http://www.mckinsey.com/mgi/publications/big\\_data/index.asp](http://www.mckinsey.com/mgi/publications/big_data/index.asp).

- Mendelson, H. 2000. Organizational Architecture and Success in the Information Technology Industry. *Management Science*. 46(4) 513-529.
- Radner, R. 1993. The Organization of Decentralized Information Processing. *Econometrica*. 61(5) 1109-1146.
- Saunders, A., Brynjolfsson, E. 2012. Valuing IT-Related Intangible Assets. Working paper.
- Smirlock, M., Gilligan, T., Marshall, W. 1984. Tobin's  $q$  and the Structure-Performance Relationship. *The American Economic Review*. 74(5) 1051-1060.
- Tambe, P., Hitt, L. 2012. The Productivity of Information Technology Investments: New Evidence from IT Labor Data. *Information Systems Research*. 23(3) 599-617.
- Tambe, P., Hitt, L., Brynjolfsson, E. 2012. The Extroverted Firm: How External Information Practices Affect Innovation and Productivity. *Management Science*. 58(5) 843-859.
- Tetlock, P. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*. 62(3) 1139-1168.
- Tetlock, P., Saar-Tsechansky, M., Macskassy, S. 2008. More than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*. 63(3) 1437-1467.
- Van Zandt, T. 1999. Decentralized Information Processing in the Theory of Organizations, in *Contemporary Economic Issues, Vol. 4: Economic Behavior and Design*, pp. 125-160. M.R. Sertel (ed.). London: Macmillan Press.
- Villalonga, B. 2004a. Diversification Discount or Premium? New Evidence from the Business Information Tracking Series. *The Journal of Finance*. 59(2) 479-506.
- Villalonga, B. 2004b. Intangible Resources, Tobin's  $q$ , and Sustainability of Performance Differences. *Journal of Economic Behavior & Organization*. 54(4) 205-230.
- Zadrozny, W. 2006. Leveraging the Power of Intangible Assets. *MIT Sloan Management Review* 48(1) 85-89.