Journal of Big Data
a SpringerOpen Journal

RESEARCH Open Access

Big data, Big bang?

Jacques Bughin[1,2,3]

Correspondence:
jacques_bughin@mckinsey.com
[1]Director McKinsey & Company, Brussels, Belgium
[2]Fellow of the University of Brussels, Brussels, Belgium
Full list of author information is available at the end of the article

Abstract

Using a random sample consisting of hundreds of companies worldwide, we are testing the impact on company performance of investing in big data projects targeted on three major business domains (namely, customer interface, company supply chain and competitors). The performance test relies on a so-called trans-logarithmic production function, allowing for a more direct test of the complementarity between big data capital and big data labour investments; further, we have used a Heckman correction to adjust for the fact that companies investing in big data are generally more productive than their peers.
We confirm and extend early results of a productivity impact from big data. We find that for the average of our sample, more productive firms are also faster adopters of big data than their industry peers (this explains 2.5% of productivity difference). Big data investments in labour and IT architecture are complements, with a total productivity growth effect of about 5.9%. Big data projects targeting customers and competitive intelligence domains bring slightly more performance than big data projects aimed at supply chain improvements.

Keywords: Information systems; Big data; Data analytics; Competitive performance; Organization assets

> "Exploiting [big data] can improve (…) performance. But first you will have to change your decision-making culture"
>
> Erik Brynjolfsson and Andrew McAfee, Harvard Business Review, Oct. 2012 [1]

Introduction

Big data, or the handling of vast amounts of data through a parallel IT architecture, is a buzz nowadays. Multiple reports suggest that data creation will continue to grow at a rate between 40 and 60% a year [2], while a quick look at Google Trends, a Google analytic tool aggregating search queries, reveals that big data queries have grown tenfold in a matter of some 2.5 years.[1]

Internet companies have been pioneering successful big data investment projects, due to the massive amount of almost real time data that they are handling. While Google was indexing a million pages for a few million searches in 1998, it was indexing more than a trillion pages ten years later, for more than 3.5 billion search queries performed every day, or 1.2 trillion searches a year, according to the tracking website, Internetlivestats.com.[2] Likewise, Facebook is handling about a billion content information queries every day, and Netflix has accumulated billions of viewer ratings, with members searching and adding millions of items every day [3].

Netflix has used big data to improve its content recommendation engine, first via a crowd-sourced algorithm using customer rating, then, through machine-learning-based algorithms which are able to develop new insights from the mash ups of a wide range of data (show features, social data from other Netflix users, or box-office). Using big data generated recommendations, Netflix movie and TV series consumption has been boosted by a factor of four [3]. Google is running a vast amount of experiments in order to induce faster search query clicks on its domain, with a few micro-seconds translating into additional millions of dollars being spent.[3]

More traditional companies have also adopted big data programmes, however with mixed success, if the results emerging from public case studies are analysed. High profile companies such as Harrah's or Tesco, have been early adopters and successful in investing in big data Hadoop-like infrastructures [4]. Gartner, an IT consultancy, also provides other case studies of companies, such as Macy's in retail, Infinity in Insurance, or American Express in credit card payments, which have leveraged big data for the benefit of their bottom-lines .[4] Yet, some scepticism remains as to the real value created by the average non-Internet based company from launching big data projects; as a case in point, Information Week has recently reported that the typical US company was only generating 55 US dollars cents for every dollar invested in big data projects.[5]

Clearly, this calls for looking beyond case studies and for performing a more systematic, larger-scale and statistical-driven study as to whether or not big data investment has led to an improvement of a company's performance trajectory.

Background and literature review

Works of this nature have recently been undertaken through the lens of big data investment impact on company productivity, e.g., Brynjolfsson et al. [5], Tambe [6], or still, Bakhshi et al. [7]. The seminal study by Brynjolfsson, et al. [5], leveraged survey collection at corporate level by McKinsey & Company, to document a 5–6% increase in global productivity from leveraging data-driven analytics, over the non-big data-friendly company. Using similar approaches, Bakhshi et al. [7] confirmed a productivity effect of 8% for UK firms.

Strictly speaking, though, the above studies may have scope and bias issues, leading to an overly optimistic effect of big data on productivity measures. Regarding scope, most of the quoted studies concentrate on the effect from "data" analytics, rather than on the more focused domain of big data. Bakhshi et al. [7] mention that big data techniques (in their study, text- or data-mining techniques) are used for only half of the companies in their sample.

Another issue is bias. One bias may be due to the data collection timing, leading to a big data effect featuring early adopters more than mainstream companies. For example, Brynjolfsson et al. [5], rely on a set of data collected before 2010, in the early stage of big data project adoption. Finally, the specification of a productivity equation, with big data adopters presenting a drift in productivity, may be subject to an endogeneity bias, e.g., high productivity firms tend to early adopters of technology [8].

The recent work by Tambe [6] on US firms goes a long way towards limiting such biases. Correcting for them as diligently as possibly, Tambe [6] shows that US companies that have hired IT labour with specific Hadoop/no SQL skills have achieved labour productivity that is 3% higher than the average. This effect is lower, but likely to be more robust than in other studies. Further, Tambe [6] shows that the big data effect is not universal: it is statistically significant only for firms established in data-intensive industries, such as IT-related or financial services, where data is a key strategic production input, and for the various firms that were geographically located in a Hadoop-intensive labour market, so as to secure a large enough pool of complementary talents for investment in big data projects.

A final constraint of all the recent, and still scarce, big data research is that the productivity equation is relatively simple, and does not explicitly breakdown big data's many input factors, e.g., the labour skills needed to maintain big data-related architectures and to run the associated applications, as well as the capital investments in new flexible IT architectures.

This paper follows this route, and offers three major changes to the existing literature. First, we are testing big data on profit changes, not only on productivity for more flexible forms of production function. In particular, we are using a translog production function. Further, in order to test formally for some bias, we have developed a Heckman procedure [9], that tests for a *selection* bias arising from big data adopters being more productive firms. Third, the effects are detailed for various business domains that should be the most prone to big data impact (e.g., customer-centric domains such as customer care or marketing; and competitive domains, such as strategy, or business intelligence).

The results confirm that big data investments lead to higher performance, with estimates in the high range of other studies found, that is, 6% higher profitability effects. This effect is however to be understood as the combination of direct as well as complementarity effects of joint investment in capital architecture and talents within big data. This effect is also purged as far as possible, by a selection bias, from the tendency of high performing firms to be early adopters of big data projects. Finally, we find that the effect of big data is slightly higher for application domains such as business intelligence and customer interfaces, than for supply chain applications.

The next section presents the data, then the statistical model (Research design and methodology). Results are provided and discussed in Results and discussion. A concluding section provides avenues for research.

Research design and methodology

Big data definition

We denote by the binary variable, BDi, the fact that the i-th company invests in big data. Big data decisions can be assessed around many dimensions, namely, the related IT architecture, the data model used, the business domains in which the big data is applied, and the type of labour skills needed for running big data projects.

There are many data models, from visualization techniques to regression techniques, fuzzy clustering, discriminant analysis, and machine learning, for deriving powerful insights. Our focus is on any type of data model, provided it relates to big data architecture. Regarding the IT architecture, we define the variable KD_i, for the i-th company, as the stock of capital invested in architecture, servers and applications such as Hadoop or MarkLogic that allow the handling of massive flows of (mostly unstructured) data. Big data also requires big data-specific IT labour skills such as data scientists, data architects, data analysts, etc. both to maintain the systems and architecture as well as to operate the applications. We define the variable LD_i, as the human stock associated with big data at the i-th firm.

Finally, we are concerned with fields/business domains in which big data techniques are applied. Currently, many case studies focus on whether big data investments improve sales and marketing functions, e.g., development of next product to buy, enhanced on-line recommendation tools. Amatriain [3] describes how Netflix uses big data techniques to improve video watching. Evidently, big data can be applied in other fields, such as corporate supply chains (e.g., leveraging RFID data to optimize stock replenishment, [10]), or for business intelligence (e.g., anticipating likely competitor actions through web nowcasting, [8]). We define three business domains, CUST, SUPPLY, COMP, which are worth 1, 0 otherwise, if the company has launched big data projects in the related domains. These domains are not only chosen for reference to the existing big data literature, but mainly because companies collecting large amounts of information in these three domains are known to require major advanced technologies such as big data, and new organizational assets (see [11, 12] [6]).

Performance definition

The typical metric used in recent works (e.g., [6]) has been to measure the effect of big data on the i-th company's added value expansion, (call it, Y_i), that is, recent work measuring the big data effect as a productivity effect. We extend the analysis on whether big data provides a differentiated advantage, on the i-th company's profitability, Π_i, with $\Pi_i = \text{margin}_i{}^*Y_i$. That is, we measure the effect of big data on *both* productivity *and* margin development.

Noting the industry profitability by Π^*, our dependent variable is the yearly change in relative profit Π_i/Π^* for the i-th company. This metric can be positive if the margin and/or output expands faster than the industry average, even if we do not have a true split between the margin and output components; in particular, company margin information is not readily available. We do however ask for the relative development of margin and output versus the competition in our survey. The trend is being measured by a binary variable which takes the value of 0 (if reduction), 1 (stabilization), or 2 if increase versus the previous year. Combining the value assumed by each of the two variables, a variable is being constructed as equivalent to a Likert scale, with ordinary values from 0 to 4. A value of 0 means a decrease in both market share and profit margin in the last year, while a value of 4 means an increase in both performance metrics. On top of Π_i/Π^*, we build $PERF_i$, which stands for the ordinary variable construct of relative profit trend, after standardizing with an average of 0 and a standard deviation of 1, so the variable takes a value bounded into the interval (0,1).

The big data performance model

As mentioned in the introduction, typical models of big data impact consider a production model, linking Y_i to labour input, L_i, and capital, K_i. Those models further assume a simple Cobb-Douglas function between those inputs, and add a drift variable, measuring big data adoption. In mathematical form:

$$\text{Log}\,(Y_i) = a + \beta.\,\text{Log}\,(K_i) + \delta.\,\text{Log}\,(L_i) + \chi.\,\text{Log}\,(BD_i) + u_i \tag{1}$$

Where u is an error term; α, β, χ, δ are parameters to be estimated; β and δ are the technical coefficients from the Cobb-Douglas function, and χ is the drift parameter, measuring the effect of big data on productivity increases.

Equation (1) presents some challenges. First, it does not purge L and K from the portion of capital and labour linked to big data. Second, the Cobb-Douglas assumes complete substitution between K and L, but this may be rather restrictive, and we may assume more general technical combinations of inputs for generating a company's added value.

We define K = KD+ KND, and L = LD + LND where KND (respectively LND) is the stock of all types of capital, including machinery and IT, but outside big data capital investment (respectively, is the stock of labour outside of the big data workforce). We also consider a generalization of Cobb-Douglas to a translog function for KD and LD. Limiting the interaction terms to the ones of interest, we thus have [7]:

$$\begin{aligned}\mathrm{Log}\,(\mathrm{Y_i}) = \mathrm{a} + \beta_0.\,\mathrm{Log}\,(\mathrm{KND_i}) + \beta_1.\,\mathrm{Log}\,(\mathrm{KD_i}) + \mathscr{S}.\,\beta_2.\,(\mathrm{Log}\,(\mathrm{KD_i}))^2 + \delta_0.\,\mathrm{Log}\,(\mathrm{LND_i}) + \\ \delta_1.\,\mathrm{Log}\,(\mathrm{LD_i}) + \mathscr{S}.\,\delta_2.\,(\mathrm{Log}\,(\mathrm{LD_i}))^2 + \gamma.\,\mathrm{Log}\,(\mathrm{KD_i}.\,\mathrm{LD_i}) + \mathrm{u_i}\end{aligned} \tag{2}$$

Where:

- At the mean of sample, $\beta_{1+}\beta_{2.}$ Log(KD), as well $\delta_{1+}\delta_{2.}$ Log(LD) measure big data capital and labour elasticities;
- The sign of β_2 as well of δ_2 measures growing (if positive) or declining (if negative) returns in new big data investments in capital and labour input;
- The sign of γ, measures the extent of substitution, (if negative) or of complementary (if positive) between big data labour and capital.

Further, we note Log (Π/Π*) = Log (margin$_i$)- Log(Π*) + Log (Y$_i$),for which we approximate the first two terms by:

$$\mathrm{Log}\,(\mathrm{margin_i})\text{-}\,\mathrm{Log}\left(\prod *\right) = \phi + \mathrm{k.\,PERF_i} + \mathrm{v_i} \tag{3}$$

where v is an error term and k (k > 0) is the parameter tackling the common result that margins developments in oligopoly markets are often correlated with market share development (see [13]). Integrating (3) into (2), we now have a profit equation of the form (4):

$$\begin{aligned}\mathrm{Log}\left(\prod\nolimits_{i}/\prod\nolimits_{*}\right) = \tau + \kappa.\,\mathrm{PERF_i} + \beta_0.\,\mathrm{Log}\,(\mathrm{KND_i}) + \beta_1.\,\mathrm{Log}\,(\mathrm{KD_i}) + \mathscr{S}.\,\beta_2.\,(\mathrm{Log}\,(\mathrm{KD_i}))^2 \\ + \delta_0.\,\mathrm{Log}\,(\mathrm{LND_i}) + \delta_1.\,\mathrm{Log}\,(\mathrm{LD_i}) + \mathscr{S}.\,\delta_2.\,(\mathrm{Log}\,(\mathrm{LD_i}))^2 + \gamma.\,\mathrm{Log}\,(\mathrm{KD_i.LD_i}) + \mathrm{w_i}\end{aligned} \tag{4}$$

where PERF has been defined earlier, and we posit k > 0, $\tau = \alpha + \phi$; $w_i = v_i + u_i$.

We are estimating a model (4') equivalent to equation (4) above, making still two adjustments. First, our sample (see description later on) includes a set of heterogeneous firms. We are thus in need of incorporating a set of corporate control effects, FIRM$_i$, for each i-th company. Consistent with Riemer et al. [14], and based on available data, the FIRM vector includes information for continent location (North America is the reference), company size (revenue below 1 billion sales is the reference, for 3 categories:

< 1 billion, 1–5 billion, >5 billion annual sales), and company sectors (B2B is the reference, three categories, B2B, B2C services and B2B goods). We also acknowledge that company performance may not be random and can exhibit some forms of persistence [15]. We thus include the lag of our performance variable as an additional regressor.

Second, we may not assume that the disturbance term w is randomly distributed. We in fact posit that companies early in the big data investment cycle may have already been performing better than their peers. This is a common trend in other technologies adopted by companies, e.g., ERP or Enterprise 2.0 (see [8] and references there-in).[8] In other words, there may be a risk of selection bias in (4). We are correcting for this possible bias through a Heckman correction procedure [9].[9]

Specifically, for each i-th firm, we add another regressor into Equation (4), CBD_i. CBD_i is the inverse of the Mills ratio measured from regressing BD_i on the following variables, from equation (5):

$$\mathrm{BDi} = \theta + \rho'.\mathrm{FIRMi} + \upsilon.\,\mathrm{SPILLOVER}_i + r_i \tag{5}$$

where r is a disturbance term assumed to be randomly distributed and SPILLOVER is defined below. The significance of a CBD effect on performance is thus equivalent to a test of a self-selection variable, i.e., companies that are quick to adopt big data are on average performing better. Technically, Equation (5) must include regressors excluded from equation (4). The SPILLOVER variable measures the extent to which other companies in the same sector have already invested in a big data projects. Sector is defined here among a list of two-digit NACE/SIC industry codes. Imitation strategy is typically visible in consumer purchase behaviour, but companies are also engaged in imitation, in particular, companies resort to such tactics when they look into investing in new technologies [16, 17], and when they must build organizational learning [18].[10] Another argument would be that companies investing in big data projects, are tapping into common external factor markets. Tambe [6] shows convincing evidence of this, as well as [19].

Sampling

We use a data panel for the year 2013, aimed at assessing the adoption of technologies, and constructed jointly by McKinsey and a major global research firm, TNS. The latter company not only maintains the panel but has also trained C-suite respondents to complete the questionnaires submitted to the panel appropriately. Other articles using the panel are by Bughin et al. [8], Bughin and Manyika, [20]. Brynjolfsson et al. [5] also leverage the panel for their seminal study on big data.

The data originates from more than 60 countries. Technically, the survey was based on a random sample of 11,000 companies, delivering a response rate of 14% for 2013, or an actual sample of about 1500 companies. Telecom, high-tech and financial services companies are the most represented in the sample. North America and Europe account for 70% of the companies, while about 30% of them achieve annual sales of more than 5 billion. We have weighted the questionnaire results by the relative size of the countries. We have then reweighted them per GDP size in US dollars in 2013, as published by the IMF.

The questionnaire we have implemented on the data panel is a follow-up questionnaire on the adoption of collaboration technologies for the years 2012–2013. Given the constraints, questions on big data investments within this questionnaire were relatively limited, and confined to the share of big data investment in total investment, and of big- data workforce in total workforce, as well as domains of big data use. Only 55% of companies answered the questions relating to big data investment and domains of use. Likewise, only 65% of companies reported their relative performance. Further, we were able to analyse firms with published financial accounts in order to collect data on total workforce and capital investments, for about 69% of cases. Our final sample, with full data completion, consists of 714 firms.

The summary statistics are presented in Table 1. We note that companies have achieved a financial performance that is quite similar to performance shown in many global publications [8]. EBIT margin and return on assets are around 10%, investment rate is in the range of 6% of revenue, with a total workforce of about 420 Full-Time Equivalents.[11]

More specifically regarding big data, 36% of the C-suite respondents claim that their companies have invested in big data architecture, within which, 72% claim to leverage big data for customer domains (CUST), 45% for supply chain domains (SUPPLY) and 35% for competition (COMP) domains. This big data adoption level is roughly in line with other available statistics; for instance, a recent report by IDG Enterprise claims that 16% of large US enterprises have fully invested in their big data infrastructures, and 20% are in the process of investing.[12] Likewise, companies reported over-investment in areas relating to their customer markets. This also tallies with anecdotal evidence. Taking a sample of about 50 McKinsey-supported implementations of big data projects in the last two years, in sectors as varied as chemicals, grocery retail, telecom, or oil and gas, roughly two-thirds of them seem to be biased towards leveraging customer data.

Regarding the workforce, we include not only the IT workforce providing all the technical big data ability, but also, the amount of workforce devoted to leveraging the technical inputs for business insights, e.g., big data analysts, etc. Typically, the pure IT technical provider side is roughly just less or the same size, in terms of employees, as the big data user side.[13] Taking this global definition, companies investing in big data are also building an appropriate big data workforce, already comprising about 1.7% of their total workforce, or 2.2% of the total value share. This percentage of companies

Table 1 Big data sample statistics

Variables	Average	St Deviation*
EBIT margin	11.3%	10.79%
Return on assets	7.4%	11.20%
Employees (total workforce)	420	745
Capex to revenue ratio	6.34%	7.28%
Big data adoption rate	36%	24%
Big data capex share	7.2%	8.44%
Big data labour share	2.2%	3.81%
Big data domain: customers	72%	
Big data domain: supply chain	45%	
Big data domain: competition	35%	

Note:* standard deviation across regions, size and segment clusters.

hiring data talent tallies with other sources. The IDG enterprise study, referred to above, reports that many companies are investing in the skill sets necessary for big data deployment, including Data Scientists (27%), Data Architects (24%), Data Analysts (24%), Data Visualizers (23%), Research Analysts (21%), and Business Analysts (21%).

Finally, we observe that 38% of companies have reported an improvement (35% a reduction) of their margin versus competition, while 35% (45%) claim to have won (lost) market share (the balance is companies reporting no effect). 47% of companies claimed to have improved on at least one performance-related dimension. Dissecting the data between companies investing in big data and others, 58% of companies with big data investments have reported improvement in their performance relative to their competitors, versus 40% for those not investing in big data, or a net effect of PERF = 18%. Furthermore, 76% of companies which have invested both in big data architecture and in additional big data labour skills have reported improvement in both performance measurements, for a net effect of PERF = 52%.

The above clearly adds weight to the hypothesis that big data may lead to an improved performance, and especially when big data investments are complemented by an increase of the big data workforce. We formally test this hypothesis econometrically in the following results section.

Results and discussion

Reference model

The final model is composed of two sequential equations. We first estimate Equation (5), then derive the inverse of the Mills ratio for big data project adoption, and include it as regressor in Equation (4a), – an amended version of Equation (4) with extra investment control variables[14]:

$$\begin{aligned}\mathrm{Log}\left(\prod_{\mathrm{i}//}\prod_{*}\right) = {} & \tau + \kappa.\ \mathrm{PERF_i} + \beta_0.\ \mathrm{Log}\ (\mathrm{KND_i}) + \beta_1.\ \mathrm{Log}\ (\mathrm{KD_i}) + \$.\ \beta_2.\ (\mathrm{Log}\ (\mathrm{KD_i}))^2 \\ & + \delta_0.\ \mathrm{Log}\ (\mathrm{LND_i}) + \delta_1.\ \mathrm{Log}\ (\mathrm{LD_i}) + \$.\ \delta_2.\ (\mathrm{Log}\ (\mathrm{LD_i}))^2 + \\ & \gamma.\ \mathrm{Log}\ (\mathrm{KD_i}.\ \mathrm{LD_i}) + \omega_1.\ \mathrm{CBD_i} + \omega_2.\ \log\left(\prod \mathrm{i}_{/}\prod_{*}\right)_{-1} + \sigma.\ \mathrm{FIRM'_i} + \mathrm{w_i}\end{aligned} \tag{4a}$$

where w is a random term assumed normally distributed; σ is a vector of parameter linked to our set of company control variables.

The final model is estimated by traditional linear regression techniques. Given large amounts of unobserved data, we estimate a model with variables computed in difference versus their average, which is similar to a fixed effect model. We test the following hypotheses: $\kappa > 0$ (oligopoly effect), $\omega_1 > 0$ (selection bias effect), $0 < \omega_2 < 1$ (partial adjustment effect in long-term performance), and mostly: $\beta_{1,}\delta_1 > 0$, (productivity effects of big data) $\beta_{2,}\delta_2 > 0$, (increasing returns of big data) as well as $\gamma > 0$ (complementarity effect between big data investment in capital and labour.

Decision to invest in big data projects

Table 2 presents the estimated results for the adoption equation (5), with associated *P*-values. Regarding control, we find that larger companies, especially those which generate more than USD 5 billion in sales, are more prone than smaller ones to invest in big data projects. Likewise, North America is quicker to invest, as are companies in the B2C service sector.

Regarding our hypothesis of either a significant imitation propensity, or the prevalence of large common externalities among companies (k > 0) [19, 21], we find a significant marginal effect of k = 0.38. If this was only a measure of imitation effect, this effect is smaller than in other recent technological investments - e.g., Leroux et al. [22] regarding Internet and ERP technologies. If this was a measure of a common externality effect, this estimate is rather close to the externality effect of 20–30% found in [19].

Effects of big data on corporate performance

We use Table 2 results to build the Mills ratio, and estimate the performance equation, as represented in Equation (4a). Results are displayed in Table 3. The last column reports P-values computed from robust standard errors at company level. As acknowledged by Tambe [6] among others, empirical studies in the IT value literature are subject to concerns of causality bias, omitted variables biases, etc. We have addressed some of the biases through company control effect, a Heckman correction, etc. The cross-sectional nature of our data, measuring big data investment in 2013, prevents a lot more correction. We tested our results, clustering by industry NACE code, and by sub-sample of higher, lower, or neutral PERF companies. The results did not significantly change.

We first comment on parameters outside of the production function. We notice a statistically significant hysteresis effect in performance, with ω = 0.33. Regarding the company control variables, a set of variables exhibits no significance on a standalone basis, but a F-test passes the test of their joint relevance at 10% (F = 0.071): companies of intermediate size (1–5 billion sales) seem to perform marginally better, contrarily to companies in the B2C arena; the US companies in our sample exhibit better performance than their peers in Europe and South America. Regarding the effect of PERF, we find κ = 0.6%. This positive effect on performance implies that the average big data company (with PERF = 18%- see Table 1) has been generating 0.6%/18% = 3% better margin performance.

Table 2 Big data adoption (Equation (5))

Explanatory variables	Coefficients	*P*-value
SPILLOVER	0.38	0.006
CONTROL		
More than 5 billion sales	0.11	0.002
More than 1, less than 5 billion sales	0.07	0.054
Europe	−0.08	−0.004
South America	−0.11	−0.011
Asia, Pacific	−0.09	−0.012
B2C goods	0.06	0.120
B2C services	0.12	0.047

Notes:
1. Default size is sales of less than 1 billion, default continent is North America, and default industry is B2B.
2. Adjusted R-square is 0.472; F = 0.012.

Table 3 Performance equation (Equation (4a))

Explanatory variable (Parameter)		Coefficient	*P*-value
PERF	(κ)	0.006	0.067
Log [KND]	(β_0)	0.251	0.015
Log [KD]	(β_1)	0.021	0.001
$[\text{Log[KD]}]^2$	(½.β_2)	0.015	0.097
Log [LND]	(δ_0)	0.541	0.000
Log [LD]	(δ_1)	0.012	0.045
$[\text{Log[LD]}]^2$	(½. δ_2)	0.027	0.018
[Log[KD]. [Log[LD]]	(γ)	0.105	0.031
[CBD]	(ω_1)	0.025	0.040
Log [Π] $_{-1}$	(ω_2)	0.326	0.002
FIRM CONTROL:			
More than 5 billion sales		−0.026	0.223
More than 1, less than 5 billion sales		0.013	0.072
Europe		−0.021	0.062
South America		−0.037	0.093
Asia, Pacific		−0.002	0.401
B2C goods		−0.027	0.037
B2C service		0.024	0.567

Notes
1. Default size is sales of less than 1 billion, default continent is North America, and default industry is B2B.
2. Adjusted R-square is 0.762; F = 0000.

We also find a statistically significant effect of the inverse Mills ratio on performance, with $\omega_1 = 2.5\%$. As anticipated, this means that higher productive companies are also quicker to adopt big data.

We now turn to coefficients of the production function. We note that the labour and capital elasticities lie in the typical range of elasticities estimates for added-value production functions at company level with Cobb-Douglas specification (see for example [5]).

What is more interesting is that the elasticities of big data capital and big data labour stock are significant, for nearly all coefficients at 5%. Their economic impact on performance is also rather large – as the additive effect amounts to up to 5.9% of the effect on total productivity growth for the companies in our sample. While the size of these effects converges with the early literature, its nature is relatively different.

First, we have already corrected for the self–selection bias that higher productive companies tend to be early adopters of big data. Second, the big data effect is mostly driven by complementarity between investment in big data labour skills with big data IT investment.

The total big data effect is computed as follows from our Table 3. We observe that the growth in KD (=log(KD)) is 11%, and in LD (=log (LD)) is 10%. Including in computation only the statistically significant production parameters at 5%, the complete big data capital elasticity, $\varepsilon_{kd} = \beta_1 + \beta_2.\text{Log(KD)} + y.\text{log(LD)}$, amounts to: 2.1% + (10.5%*10%) = 3.15%. Likewise, the complete big data labour elasticity $\varepsilon_{ld} = \delta_1 + \delta_2.\text{Log(LD)} + y.\text{Log(KD)}$ is worth 1.2% + (2.7%*15%) + (10.5%*11%) = 2.75%. The sum of both labour and capital elasticities is 2.75% + 3.15%, or 5.9%.

Remark that δ_2 is positive, meaning increasing returns on big data labour investment. Further, the complementarity effect in labour and capital boosts the big data capital elasticity by 50% (1.05%/2.1% = 50%), as well as the big data labour elasticity by more than 70% (1.15%/1.6% = 72%). Complementarity between labour and capital is an essential driver of the big data effect on corporate performance.

Big data effects by big data domain

We have also measured big data investment domains, e.g., for customer interface, supply chain and intelligence. We are interested in testing whether these domains (versus all others) are producing an even greater productivity effect from big data. To test this hypothesis, we can interact all the big data productivity terms with the binary variables, CUST, SUPPLY and COMP, as defined above.

As the number of regressors in Equation (4a) is already quite large, we run into some multi-collinearity problems if we directly include interaction terms in (4'). We thus rather resort to another estimation strategy, whereby we compute the big data capital and labour elasticities ε_{kd} and ε_{ld} for each company that has invested in big data, which we directly correlate with each whether the i-th firm has or not invested in either big data domain.

Formally, we have:

$$\varepsilon_{kd} = a_0 + a_1.\,\mathrm{CUST} + a_2.\,\mathrm{SUPPLY} + a_3.\,\mathrm{COMP} + e_{kd} \tag{6}$$

$$\varepsilon_{ld} = b_0 + b_1\,\mathrm{CUST} + b_2.\,\mathrm{SUPPLY} + b_3.\,\mathrm{COMP} + el_d \tag{7}$$

where ε is computed from the estimation of (4'), $a_{i's}$ as well as $b_{i's}$ (i = 1,2,3) are marginal effects on big data elasticities, and the last terms are error terms. The results are shown in Table 4.

As seen from Table 4, big data domains such as CUST and COMP are statistically significant at traditional risk levels. The positive sign implies that big data productivity capital and labour elasticities are higher for companies invested in CUST and COMP domains. A test of the average further suggests that b1 > a1 that is the effect of big data projects targeted on customers has a greater impact on capital productivity impact than on labour productivity.

Table 4 Big data adoption (Equations (6–7))

Explanatory variables parameters		Coefficients	*P*-value
CUST	a_1	0.424	0.046
	b_1	0.667	0.092
SUPPLY	a_2	−0.142	0.127
	b_2	−0.617	0.521
COMP	a_3	0.772	0.044
	b_3	0.802	0.036

Notes:
1. Default big data domain is all domains outside sales/marketing, supply chain and competitive intelligence.
2. Adjusted R-square for equation (6) is 0.566; F = 0.023.
3. Adjusted R-square for equation (7) is 0.392; F = 0.047.

At the sample average, the complete effect on productivity domains is however clustering at 0.3–0.5% effect versus 3% effect for big data input – thus it adds a 10% improvement to the average big data project.

Clearly domains play a role but what really matters is to invest in complementary input in big data projects in order to see a major shift in companies' overall performance trajectory.

Conclusions

This article extends the recent work on the effects of big data on corporate performance. Its main innovation is to develop a more general approach to corporate performance and to the production function, that allows us to answer more directly questions such as the (importance of) complementarity of big data capital and labour.

After adjusting for higher performing firms being early adopters of big data via a Heckman procedure (1979), we find that the major performance impact of big data resides in the close complementarity between big data IT investment and labour skills. The performance effect is also slightly higher for application domains such as business intelligence and customer interface domains.

Avenues for research are clear. The first avenue, ironically, relates to better data itself. We have only cross-sectional information collected and we would like to see how big data evolves over the years, including a productivity development time. The second avenue is to investigate larger interactions between traditional forms of capital and labour and big data-in our translog: we assume relative independence between these types of inputs, as otherwise, a typical regression estimation becomes quickly too complex.

Endnotes

[1] The analytic tool, Google trends (www. Googletrends.com) describes the intensity of search query terms among total searches. It reveals that the search intensity of the term "big data" grew 10 times from Sept 2011 to March 2015. Meanwhile, total searches had grown by a compound rate of 10% a year (see www.internetlivestats.com; accessed on March 19 2015).

[2] Accessed March 19, 2015.

[3] A Marine software research conducted at the time of Google Instant launch, reported 9.4% boosts in search performed per day; not all searches were monetized at the same rate as per traditional search, leading to about a 2.5% revenue upside for Google, see www.prweb.com.

[4] See http://searchcio.techtarget.com/opinion/Ten-big-data-case-studies-in-a-nutshell

[5] As reported in http://www.informationweek.com/big-data/big-data-analytics/3-roadblocks-to-big-data-roi/

[6] We thank a referee for suggesting this point and for providing related references.

[7] Given the Taylor-like expansion, is well-known that a complete model of translog will lead to a non-tractable number of parameters, let alone a major risk of multicollinearity. We are thus concentrating only here on big data inputs. As a referee suggests, this omission may lead to some biases in the big data effect on performance. We ran extra regression tests by incorporating one extra cross-effect at the time. Big data

elasticities were affected in the range of −12%, +12% at maximum, with an average among simulations of −3%;+5%.

[8] Tambe [6] also finds evidence of more productive firms adopting new technologies faster.

[9] Obviously, the Heckman procedure removes only biases linked to selection bias; other biases can still prevail, i.e., omitted variables, time-dynamics of big data on productivity, etc.

[10] Competitive intensity is another critical variable, but we lack firm data on the matter. The fixed effect model picks up part of this effect.

[11] Statement is made with respect to publicly quoted companies, using statistics from Reuters.

[12] Summary details made public are available at http://www.idgenterprise.com/report/big-data

[13] Statistics on the split are provided in a recent seminar by the European Commission, on big data, workshop on big data skills for Europe, Nov 2014.

[14] We designed two models to estimate the full model composed of equations (4a)-(5). The first estimates (5) on total sample, then (4') on the restricted sample of the 257 firms investing in big data. The second one is based on a transformation LD' = 1 + LD, and KD' = 1 + KD to accommodate the full sample, including companies with no big data investment (LD = KD = 0). In such a case, the coefficients of (4') are to be interpreted in terms of semi-elasticities given the approximation, say for LD, that Log(1 + LD) = LD for small level of LD. The reported results in the paper are regarding the first methodology. The second methodology estimates are plagged by multicollinearity between CBD and transformed variables, but coefficient signs and mean levels are very close to first model.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This paper has been prepared for the "Strategic Management Society Meeting" in Madrid on September 21, 2014 for Session 463 re "Big Data: Revolutionizing Innovation and Competition". The author acknowledges that the paper has benefitted from significant suggestions by two referees. The content of this article only engages his author.

Author details

[1]Director McKinsey & Company, Brussels, Belgium. [2]Fellow of the University of Brussels, Brussels, Belgium. [3]Fellow of the Catholic University of Leuven, Leuven, Belgium.

Received: 9 October 2014 Accepted: 24 March 2015
Published online: 07 January 2016

References

1. Brynjolfsson E, McAfee A (2012) Big data: the management revolution. Harvard Business Review
2. OECD (2013) New sources of growth- knowledge based capital. OECD, Paris. http://www.oecd.org/sti/inno/knowledge-based-capital-synthesis.pdf
3. Amatriain X (2013) Beyond Data: from user information to business value through personalized recommendations and consumer science, CIKM'13. San Francisco, CA, USA
4. Davenport TH (2014) Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press (ed)
5. Brynjolfsson E, Hitt L, Kim HH (2011) Strength in numbers: how does data-driven decision making affect firm performance? MIT - Sloan School of Management
6. Tambe P (2014) Big data investment, skills and firm value. Manag Sci 60/6:1452–1469
7. Bakhshi H, Bravo-Biosca A, Mateos-Garcia J (2014) Inside the datavores: how data and online analytics affect business performance, Nesta
8. Bughin J, Hung Byers A, Chui M (2011) How social technologies are extending the organization. McKinsey Quart. Available at http://www.mckinsey.com/insights/high_tech_telecoms_internet/how_social_technologies_are_extending_the_organization
9. Heckman J (1979) Sample selection bias as a specification error. Econometrica 47(1):153–161
10. Zaslavsky A, Ch P, Georgakopoulos D (2013) Sensing as a service and big data, arXiv 1301–1059

11. Mendelson H (2000) Organizational architecture and success in the information technology industry. Manag Sci 46(4):513–529
12. Tambe P, Hitt LM, Brynjolfsson E (2012) The extroverted firm: how external information practices affect innovation and productivity. Manag Sci 58(5):843–859
13. Clarke R, Davies SW (1982) Market structure and price–cost margins. Economica 49(195):277–287
14. Riemer K, Steinfeld CH, Vougel D (2009) e-collaboration: on the nature and emergence of communication and collaboration technologies. Electron Markets Int J 19:23
15. Villalonga B (2004) Intangible resources, Tobin's q, and sustainability of performance differences. J Econ Behav Organ 54:205–230
16. Hollenstein H (2002) Determinants of the adoption of ICT: an empirical analysis based on firm level data for the Swiss business sector. Druid summer Conference on Industrial dynamics of the new and old economy, Copenhagen
17. Lee SG, Trimi S, Kim C (2013) Innovation and imitation effects' dynamics in technology adoption. Industrial Manage Data Syst 113(6):772–799
18. Fichman RG, Kemerer C (1997) The assimilation of software process innovations: an organizational learning perspective. Manag Sci 43(10):1345–1363
19. Tambe P, Hitt LM (2013) Job hopping, information technology spillovers and productivity growth. Manag Sci 60(2):338–355
20. Bughin J, Manyika J (2008) Bubble or paradigm change? Assessing the global diffusion of enterprise 2.0. In: Koohang A, Harman K, Britz J (eds) Knowledge Management-research and applications, Information Science Press
21. Hannan T, McDowell J (1987) Rival precedence and the dynamics of technology adoption: an empirical analysis'. Economica 54:155–171
22. Leroux E, Pupion P-C, Sahut J-M (2011) ERP diffusion and mimetic behaviors. Int J Bus 16(2):1–27
23. Bughin J. Online fuzz and searches as telecoms brand performance indicators, available at SSRN, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1904328. 2011