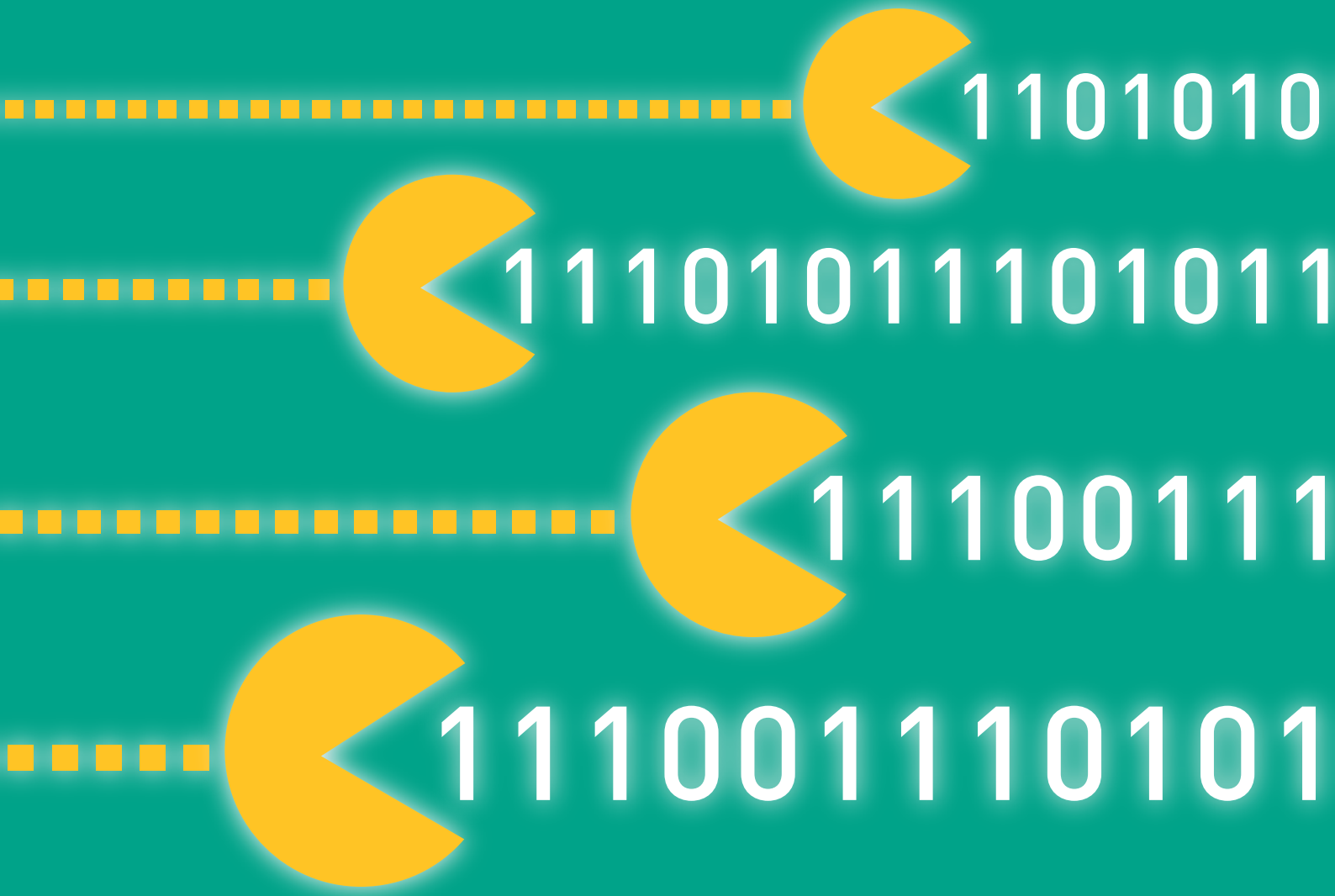
Nesta...

INSIDE THE DATAVORES

ESTIMATING THE EFFECT OF DATA AND ONLINE ANALYTICS ON FIRM PERFORMANCE

Hasan Bakhshi, Albert Bravo-Biosca and Juan Mateos-Garcia.

March 2014



Nesta...

Nesta is the UK's innovation foundation.

An independent charity, we help people and organisations bring great ideas to life. We do this by providing investments and grants and mobilising research, networks and skills.

Nesta is a registered charity in England and Wales with company number 7706036 and charity number 1144091.
Registered as a charity in Scotland number SCO42833. Registered office: 1 Plough Place, London, EC4A 1DE.

www.nesta.org.uk

SUMMARY

Despite numerous claims that data is a critical source of competitive advantage for firms, there is little empirical analysis of its link with firm performance. This paper uses a survey of data activity for 500 UK firms, which are commercially active online, to quantify the contribution that online data use – that is, the collection, analysis and deployment of online customer data – makes to business productivity.

We find that a one–standard deviation greater use of online data is associated with an 8 per cent higher level of productivity (TFP): firms in the top quartile of online data use are, other things being equal, 13 per cent more productive than those in the bottom quartile. When we distinguish between the different data–related activities that firms undertake, we find that greater data analysis and reporting of data insights have the strongest link with productivity, whereas amassing data has little or no effect on its own. Consistent with this, we report significant links between online data analysis and reporting and profitability measures.

We also study the complementarities between online data activity and other organisational attributes and behaviours. We find that the impact of online data use is stronger for firms with higher levels of employee autonomy, and for firms willing to disrupt their business processes. An implication for managers is that their data investments stand to generate more benefits when they are accompanied by other organisational changes.

Acknowledgements

We are grateful to Christian Fons Rossell, Josh Siepel and Andrew Whitby for valuable comments and discussion.

INSIDE THE DATAVORES

ESTIMATING THE EFFECT OF DATA AND ONLINE ANALYTICS ON FIRM PERFORMANCE

CONTENTS

1. INTRODUCTION 5
   - Inside the Datavores 6

2. MODELS 8
   - a) Online data activity and total factor productivity 8
   - b) Complementarities between online data activities and other firm characteristics and behaviours: employee autonomy 9
   - c) Complementarities between online data activities and other firms characteristics and behaviours: process innovation 10

3. DATA AND MEASURES 11
   - a) Measures 11
   - b) Estimation issues 12

4. RESULTS 13
   - a) Simple correlations 13
   - b) The contribution of online data activities to firm productivity 13
   - c) Complementarities between online data activities and employee autonomy 14
   - d) Complementarities between online data activities and process innovation 14
   - e) Robustness 14
   - f) Discussion and implications 16

5. CONCLUSIONS 17

BIBLIOGRAPHY 18

ENDNOTES 20

FIGURES AND TABLES 21

1. INTRODUCTION

It is hard to ignore the current excitement about the commercial potential of the 'data revolution' (MGI, 2011), with data analytics, big data and allied concepts holding out the promise of big returns in seemingly every sector of the economy. 'Big data' is now competing with topics like 'cloud computing' and '3D printing' – two other 'hot' technology areas – as popular terms on Google Search (Figure 1).

While definitions of this data revolution vary, most accounts focus on three dimensions of data: its unprecedented volume, velocity and variety.

As regards volume, IBM has famously estimated that in just two years, mankind generated as much data as it had done in all of its history up until that point (IBM, 2012). The OECD projects that global data creation will grow by 40 per cent yearly, compared with 5 per cent growth in IT expenditure overall (OECD, 2013).

Data is also being created, analysed and acted upon with increasing velocity. Brands routinely monitor discussions in real time on social media platforms (Divol, Edelman, and Sarrazin, 2012). Google has used real-time search data to monitor the spread of flu more rapidly (although not always more accurately) than traditional epidemiological surveillance networks (Butler, 2013).

Data is also gaining in variety, with businesses increasingly having to deal with different types of data, often in unstructured formats. This creates challenges for the management, integration and analysis of data across many sources, such as text, images, videos, sound, and GPS and sensor data.

The abundance of data has made attention and analysis a scarce resource, raising concerns about 'information overload' (Shapiro and Varian, 1998; Simon, 1996). However, improvements in IT hardware as well as software innovations such as Hadoop, a framework for the distributed processing of large amounts of data, and NoSQL, a flexible type of database, not only have reduced the cost of storing and managing large data sets, but also of extracting valuable insights from them. New developments in analytical techniques and methods have also helped.

As a result, data has passed from being a modest and oft-discarded by-product of firms' operations to become an active resource with the potential to increase firm performance and economic growth through 'data-driven decision making', and data-driven goods and services. McKinsey estimates that big data will contribute up to $325 billion to US GDP by 2020 (MGI, 2013), while the OECD has included 'big data' in its group of 'knowledge-based' capital assets that will act as new sources of growth in advanced economies (OECD, 2013). There are also examples of well-known companies across a range of industries that have adopted 'big data' to surge ahead of their competitors, such as Google, Wal-Mart, Marriott Hotels, Amazon and Netflix to name just a few (Davenport and Harris, 2007).

The existing evidence appears broadly supportive of this optimistic assessment. In addition to case studies, executive surveys by consultants, analysts, industry observers and technology vendors have linked the adoption of 'big data' and data analytics to self-reported improvements in business performance. (Economist Intellience Unit, 2011; Kiron, Shockley, Kruschwitz, Finch, and Haydock, 2011; LaValle, Lesser, Shockley, Hopkins and Kruschwitz, 2011). For instance, Bakhshi and Mateos-Garcia (2012), using the same firm survey dataset that we draw on in this paper, reports that UK firms which rely more on data and analysis to make decisions – the 'datavores' – are twice as likely as the average to report significant benefits from their online customer data. The box below sets out the connection between this earlier report and the current one.

Inside the Datavores

Inside the Datavores uses the same survey data as *Rise of the Datavores,* but instead of adopting a summary measure to determine if a firm acts in a data-driven way (*"whether a firm uses data and analysis over experience and intuition when making decisions on how to grow its sales"*), it utilises more granular metrics of a firm's data activity, like comprehensiveness in its collection of different types of data, the number of techniques it uses for data analysis and communication, and the importance of data for decisions in different parts of the business (see Table 1 on page 22).

Arguably, these more detailed measures help us address a problem we identified in the *Rise of the Datavores* – namely, that in some instances companies that identify themselves as data-driven are in truth doing little data collection, analysis or deployment. By including in our models the variables that capture these data behaviours directly, we hope to reduce measurement error and reporting biases. It should in any case be noted that the average scores of the 'datavores' identified in our 2012 report are for all our measures of data activity significantly higher than for the rest of our sample.

Inside the Datavores also differs from the earlier report by making use of a follow-up survey of IT employment in firms and by using company accounts measures of firm performance.

As so often happens with new technology areas, the academic literature, however, lags behind the 'grey' literature, and there are very few quantitative studies that examine the impacts of 'big data', data analytics and data-driven decision making. Brynjolfsson et al., (2011) uses a survey of HR and IT managers to measure the adoption of data-driven decision making in a sample of 179 US publicly listed companies, and finds that a one-standard deviation increase in adoption has a positive and significant effect on productivity levels – in the region of 5 per cent to 7 per cent – as well as on business profitability. Tambe (2013) uses LinkedIn data to study the complementarities between 'big data' adoption and skills in the US, and finds a strong relationship between productivity and firms' investment in 'big data', proxied by their employment of individuals mentioning Hadoop skills in their personal LinkedIn profile.

This paper contributes to this emerging literature by examining the impact of online data – specifically, the collection, analysis and deployment of online customer data – on business productivity in a sample of 500 UK, mostly medium-sized, privately-owned firms. We adopt a production function approach and show that, other things equal, online data use is associated with stronger business performance. Specifically, we estimate that a one-standard deviation greater online data use is associated with an over 8 per cent higher level of productivity (TFP).

We also exploit several detailed survey questions about specific data activities in the responding firms and find that greater analysis of online data and the reporting of its findings have the strongest link with productivity, in contrast with the comprehensiveness of a firm's online data collection which on its own appears to have no effect. This is consistent with received industry wisdom (and an equivalent finding in the ICT adoption literature) that amassing larger amounts of online data will do little for firm performance if the data is not analysed and acted upon (LaValle et al., 2011). Data analysis and reporting is also associated with higher profitability in some of the metrics that we consider.

Another finding in the literature on ICT adoption is that organisations need to make complementary investments to fully benefit from their technology investments (Bloom, Garicano, Sadun and Van Reenen, 2009; Bresnahan, Brynjolfsson and Hitt, 2002; Crespi, Criscuolo and Haskel, 2007). Similar claims have been made about data. In this paper, we investigate whether there are complementarities between online data activity and employee autonomy and process innovation. This allows us to explore the managerial implications of widespread data access, as the way a firm is organised may impact on its ability to act efficiently and quickly upon its data-driven insights (Aghion, Bloom and Van Reenen, 2013; Bloom et al., 2009; Garicano and Wu, 2012).

The findings support the idea of complementarities between online data activities and employee autonomy: those firms in our sample that are more intensive in their online data use and grant their employees more autonomy enjoy a boost in their productivity as much as four times larger than those firms which are similarly intensive in their online data use, but centralise decision making. The complementarities that we estimate are again strongest with respect to our measure of online data analysis and reporting.

One interpretation is that firms will be able to best reap the benefits of their data if they allow their employees to act upon its insights without necessarily first having to clear their actions with their managers.

We also look at self–reported measures of process innovation in the survey, which can serve as a proxy for firms' willingness to adapt their workflows and practices in order to benefit from their data. We detect some evidence of complementarities between data use and process innovation, though the findings are less statistically robust than is the case with employee autonomy. Nonetheless, we interpret this result as suggestive that those firms which are more willing to reconfigure – and perhaps even disrupt – their production processes in response to the opportunities created by the increasing availability of online data enjoy higher productivity gains.

Our paper is most closely related to Brynjolfsson et al., (2011), although with three key differences. First, our sample includes both medium–sized and large firms, mostly privately held, rather than just large public companies. Second, we only consider firms that are commercially active on the internet.[1] Third, we focus only on a subset of all data in firms – online customer data – and consider the totality of the 'value chain' for this data, including its collection, analysis and reporting, and deployment, using a survey instrument designed for this purpose.

Lastly, it is important to note that this paper is about data – and online data specifically – rather than exclusively 'big data', and that it considers a variety of data analysis methodologies in addition to advanced 'data science' techniques.[2] This means that our examination involves looking at the performance of firms attempting to harness data sets that may be getting 'bigger' relative to what they are accustomed to, without necessarily fulfilling volume–based definitions of 'big data'. Arguably, this is where the 'data revolution' may have its most substantial impact, by enabling innovation and productivity growth in a swathe of firms beyond the (currently) small elite who are in a position to use 'big data' sets.

The structure for the paper is as follows: we first set out our hypotheses, their links to the existing literature and the econometric models we use to address them. We then describe our data sources and measures. Following this, we present our empirical findings, and test for their robustness. We conclude with a brief discussion of the implications of the paper for business managers, and signpost avenues for further research.

2. MODELS

a) Online data activity and total factor productivity

A growing body of evidence documents how intangible assets account for an increasing share of business investment across countries (Corrado, Hulten and Sichel, 2009; Goodridge, Haskel and Wallis, 2012). In addition to research and development (R&D), these include investments in design and branding, databases and market research, and in management capabilities. In their recent analysis of new sources of growth, the OECD (2012) conceptualises data as a foundation for the development of some these assets – by enhancing R&D, helping develop new products and services, supporting the optimisation of processes, improving marketing, and informing decision making more generally. Economists and management scholars have also highlighted the importance of data and analytical capabilities for production by characterising firms as 'information processors' (Radner, 1993), and proposing the 'knowledge–based' view of the firm (Grant, 1996).

However, raw data – impressions about an organisation's internal and external environment captured by multiple sensors – is in itself insufficient to generate value. In order to have an economic impact, data needs to be processed and structured into information (that is, into meaningful statements about the state of the world) and knowledge (models of the relationship between different variables, such as behaviour and outcomes) that can be used to inform action.[3] The sequential nature of this process is captured by the idea of a data value chain (Bakhshi and Mateos–Garcia, 2012; OECD, 2013).[4]

The survey instrument we use in this study attempts to capture a firm's activities across this data value chain, including: (1) its collection of data from online sources; (2) its analysis and reporting using various analytics methods and dissemination formats, and (3) its deployment in making decisions across the business. Based on the responses to these questions we create an indicator that measures a firm's joint engagement with those three activities – we refer to this data score as 'online data use'.

The first question that we seek to address with our data is therefore:

Q1: What is the link between online data use (and its constituent data activities) and firm productivity?

We do so by estimating a value–added production function equation akin to that used in Brynjolfsson et al., (2011):

$$\ln (Y)_{it} = \beta_0 + \beta_1 \ln(K)_{it} + \beta_2 \ln(L)_{it} + \beta_{3x} Data_{xit} + controls + \varepsilon_{it} \qquad (1)$$

where Y is value added, K is the stock of tangible capital, L is employment, and $Data_x$ is adoption of an online data activity x (i.e. data collection, analysis, or deployment, or their combination in a single summary score of online data use). When estimating this production function, we allow for heterogeneity in production functions across industries by interacting the production factors K and L with industry dummy variables.[5]

In the regression model, we control for a firm's IT intensity (its IT employment as a share of its overall employment), the average education level of its workforce (proxied by average wages), the extent to which it uses the web to generate revenues, and its levels of product and process innovations: intuitively all of these may be associated with both online data activities and productivity.[6] We also include industry and year dummy variables in our model.

β_{3x} the coefficient of online data activity $Data_x$ is our measure of its contribution to total factor productivity (TFP), the increase in value added resulting not from an increase in production inputs but from a more efficient use of them.

b) Complementarities between online data activities and other firm characteristics and behaviours: employee autonomy

Previous research has provided substantial evidence of complementarities between ICT investments and certain organisational characteristics and behaviours. These complementarities capture the fact that a firm may need to adopt particular practices, or invest in certain capabilities (e.g. workforce skills), at the same time as it invests in ICT in order to reap the full benefits in terms of increased productivity (Autor, Levy and Murnane, 2003; Bresnahan et al., 2002).

The relationship between the internal organisation of firms and ICT has received special attention. The idea is that ICT impacts on the cost of transmitting information within the firm, and that this has implications for its organisational structure and the optimal allocation of decision–making rights between managers and workers (Garicano and Wu, 2012). Bloom et al., (2009) find evidence that different types of ICT capital have a divergent effect on decentralisation. On the one hand, the adoption of information technologies such as Enterprise Resource Planning (ERP) systems or Computer Aided Design/Manufacturing (CAD/CAM) makes it easier for workers to access the information they need to make decisions without consulting their managers, enabling greater decentralisation. On the other hand, communication technologies that decrease the costs of transmitting information to managers, such as intranets, lead to increases in centralisation. Bresnahan et al., (2002) examines 'skills–biased technical change' and finds complementarities between ICT investment in aggregate and higher levels of employee empowerment, which it links to the increased flexibility in production processes afforded by the adoption of ICTs.

When we consider online data within this organisational complementarities framework, it is not clear whether their adoption should be linked to more employee autonomy, or the other way around. While the knowledge derived from online data can be distributed to employees, potentially increasing their ability to make decisions independently from managers – as, say, with the LexisNexis case database in the legal sector – it can also lower the costs of codifying local and personal knowledge, and as a result, reduce employee autonomy (Aghion et al., 2013).

The idea here is that in situations where knowledge is fragmented across an organisation, managers may prefer to allow their employees to use local and difficult–to–transmit knowledge to inform actions. If knowledge is codified, managers are less reliant on the tacit expertise of their employees, and can centralise decision making – an extreme example of this is where the information needed to undertake a certain task is perfectly codified, so that it can be undertaken by an automaton or algorithm (i.e. employee autonomy is minimised). Past management studies of the oil and gas sectors do in fact show higher levels of centralisation in those business areas where decisions are based on quantitative information (e.g. treasury and financial risk management) compared with other parts of the business such as strategic planning or investment appraisal, where idiosyncratic and tacit knowledge is more important (Grant and Cibin, 1996). If online data analytics lead to an increased 'quantification' of knowledge across firms, it could conceivably result in higher levels of centralisation in decision making.

The question that stems from this discussion is therefore:

Q2: Is online data use (and its constituent data activities) complementary with employee autonomy?

We address it by estimating this model:

$$\ln Y_{it} = \beta_0 + \beta_1 \ln(K)_{it} + \beta_2 \ln(L)_{it} + \beta_{3x}\, Data_{xit} + \beta_4 Autonomy_{it} + \beta_5 Data_{xit} * Autonomy \quad (2)$$

$$+ controls + \varepsilon_{it}$$

In this case, *Autonomy* is a measure of employee empowerment (and relatedly, organisational decentralisation) derived from the survey. A positive sign on the coefficient of the interaction between online data activity $Data_x$ and employee autonomy on productivity indicates that, other things equal, the benefits in terms of higher productivity that a firm derives from using online data are higher the more decentralised the firm is.

c) Complementarities between online data activities and other firms' characteristics and behaviours: process innovation

One implication of the existence of complementarities is that organisations may need to change their processes and practices to benefit from ICT. For instance, by restructuring their organisation in line with the discussion above, by modifying their HR processes to identify, recruit and incentivise workers with the right set of skills, or by building new channels for communication with customers (Brynjolfsson and Saunders, 2010). Therefore, we might expect to see a positive complementarity between ICT investment and process innovation (i.e. the willingness to implement those changes).

An analogous point has been made regarding the importance of modifying production processes in response to the advent of 'big data' – for example, access to social media data now allows TV companies to measure with more precision their audiences' engagement with programmes, and this is transforming their commissioning processes (Vanderbilt, 2013). In this case we should find a positive complementarity between online data use and process innovation.

We therefore examine the following question:

Q3: Is online data use (and its constituent activities) complementary with higher levels of process innovation?

We use a similar equation to (2) to address explore this hypothesis:

$$\ln(Y)_{it} = \beta_0 + \beta_1 \ln(K)_{it} + \beta_2 \ln(L)_{it} + \beta_{3x} Data_{xit} + \beta_4 Process\ Innovation_{it} \quad (3)$$

$$+ \beta_5 Data_{it} * Process\ Innovation_{it} + controls_{it} + \varepsilon_{it}$$

Where *process innovation* is a self–reported measure derived from the survey. The interpretation of the coefficient of the interaction between any given online data activity $Data_x$ and *Process Innovation* is similar to the one on *Autonomy*. In short, a positive coefficient suggests that the benefits of intensifying a data activity are higher when firms innovate in their processes to take advantage of this new technology.

3. DATA AND MEASURES

We test our three hypotheses using a dataset that links responses to a telephone survey of online data practices carried out for Nesta by the survey company Ipsos–MORI in spring 2012 and reported in Bakhshi and Mateos–Garcia (2012), a follow–up survey of IT employment in those same firms undertaken in autumn 2012, and financial performance data from Bureau Van Dijk's Financial Analysis Made Easy (FAME) database covering the period 2006–2012.[7]

The purpose of the telephone survey was to measure the adoption of online data practices in a sample of UK firms (at the establishment level) which were active online, as well as a range of other practices, investments and organisational behaviours. We included eight sectors in our sampling frame that according to the Office for National Statistics' E–Commerce Survey displayed a higher than average propensity to transact online, and added to that list financial services (which is not covered by the E–Commerce Survey) (ONS, 2011).[8] We drew a random sample of firms in those sectors from FAME, with the additional condition that they had more than 50 employees in 2010. FAME obtains the data from the mandatory annual accounts filings that UK firms, both public and private, need to provide to Companies House, but the financial coverage for firms with fewer than 50 employees is significantly worse due to weaker requirements (implying that such firms for which FAME does have financial data may not be representative).

The survey was targeted at Chief Marketing Officers or people with an equivalent role within the firm.[9] We also excluded from the survey any firms that did not use the internet to generate revenues, either through e–commerce sales, by selling advertising space on their websites or by advertising on other websites. This had the goal of surveying firms where online data and its use were more likely to be relevant.[10] Five hundred firms participated in the survey.

The telephone survey included questions on IT budgets and employment in the responding firms, but the response rate to these questions was low (just over 25 per cent). The importance of these questions as conditioning variables in our analysis led us to carry out a follow–on online survey of IT managers in the same sample of firms to plug the gaps; 174 firms responded to this follow–on survey, giving us a total of 300 firms with data on IT employment.[11]

a) Measures

Table 1 presents a list of the variables generated with data from the online data practices and IT surveys, together with their descriptive statistics. First, it includes our composite measures of online data activities – data collection, data analysis and reporting, data deployment, and the overall 'data score' (that is, the average of the three first data activities).[12] It also describes all the control variables derived from the survey data and the variables we use in our analysis of complementarities. All our composite indicators are standardised.[13]

We have tested for the reliability of our composite indicators by computing their Cronbach alpha scores. The Cronbach alphas are all within the acceptable range (around 0.6 or higher), which supports the idea that composite indicators are a valid summary of their underlying components.[14]

It is worth noting several things about our measures of online data activity from this table.

First, two of them – data collection and data deployment – are based on subjective scales relating to the 'comprehensiveness' in the collection of various types of online data, and the 'importance' of online data in making business decisions, so measurement error is a potential concern.

Our second measure of online data activity – data analysis and reporting – is based on the sum of scores which take either the value of 1 or 0 (i.e. whether any given tool or visualisation output is used or not by the firm). It therefore indicates how comprehensive a firm is in the methods it deploys to analyse its online customer data and to communicate the insights thus generated, although not necessarily the intensity with which they are used.

Table 2 presents the definitions and descriptive statistics for a cross-section of our financial indicators and other firm characteristics using the year 2010 (the last year for which we have an almost complete set of financial data). The average firm in our sample has 455 employees and is 23 years old (even if some are start-ups and others century-old companies). The average value added, constructed as turnover minus costs of goods sold, is £20 million, while the average remuneration per employee is £32,000. We use average remuneration per employee as a proxy for the average levels of human capital in each firm, but also test whether our main results are robust to excluding this variable. We have deflated all financial variables using producer prices and implied investment deflators from the Office for National Statistics.[15] Where possible, we have done this at the major group level (2-digit SIC code). In cases where 2-digit price indices or investment deflators are not available we have had to use lower resolution deflators at the level of services or manufacturing. We have also winsorised these variables at the 1 per cent level to deal with outliers (i.e. we replace the scores of the 1 per cent most extreme observations with those immediately below them).[16]

b) Estimation issues

Whilst we have financial performance data for the firms for the period 2006–2012, our survey data on online data practices and IT spend refer only to the period when the survey data were collected (2012). This is a common challenge in econometric studies of new technologies at an early stage of their adoption as longitudinal data are lacking. (Bresnahan et al., 2002; Brynjolfsson et al., 2011).

Because of this, we are not able to use panel techniques to establish whether there is a causal relationship between online data activity and productivity: the best we can do is test whether or not there is a correlation, controlling for other plausible determinants of firm performance. In other words, if we do detect a statistically significant relationship we cannot rule out the possibility that business performance and online data engagement are jointly caused by a third, unobserved factor, such as the disposition of the management. We try to reduce potential omitted variable biases by including controls for innovation (product and process) and other firm characteristics which we might conceivably expect to correlate with unobservable drivers of productivity and technology adoption, but we cannot rule them out.

An additional problem caused by the cross-sectional nature of our survey data is that our models implicitly assume that the online data activities we are studying have remained constant over the period under consideration. Although there is a rich body of literature showing that organisational routines and behaviours tend to be quite stable over short periods of time (Nelson and Winter, 1982), there are obvious tensions with the alleged novelty of the technologies and behaviours that are the subject our analysis. We try to address this issue to a degree by estimating our models with a sample restricted to the later years of our sample for which we have full data (2010 and 2011), over which we would expect the assumptions about stability in practices to be less problematic; we report the results in the robustness analysis.

Since we have pooled observations for the same firms over different periods, we adjust our standard errors by clustering responses at the individual firm level.

4. RESULTS

a) Simple correlations

Table 3 displays the pairwise correlations between our online data activities and the control variables. The first thing worth noting is that the correlation coefficients between our indicators of online data activity are high: those firms that are more comprehensive in their data collection also tend to use more tools for its analysis and reporting, and also to deploy data for decision making in different parts of the business. In other words, firms tend to carry out different online data activities at the same time – consistent with the idea of a 'data value chain'.

We also expect online data activities to correlate with other organisational investments, capabilities and behaviours that are also linked to productivity, including firm age, the levels of education in the workforce, reliance on the internet to generate revenues, IT investment and levels of innovation. The correlation matrix in Table 3 shows that this is generally the case: data-intensive businesses are more reliant on the internet to generate business (online business share), and are more likely to report high levels of product innovation. We also find a significant (although small) correlation between data analysis and reporting and our proxy for human capital in the workforce (logarithm of average remuneration). We detect no correlation between measures of online data use and firm age, however.

Table 3 also shows that there is a positive correlation between all online data activities and our measures of employee autonomy and process innovation, which is as we would expect if our conjectures regarding their mutual complementarities are valid.

b) The contribution of online data activities to firm productivity

Table 4 presents the findings of our regression models of productivity on online data activities. For each of our predictors, we estimate three models with different sets of control variables.

In the first model for each of our online data activities, we estimate a baseline production function, with standard production factors K and L (with industry-specific factor income shares), including online data activity, industry and year fixed effects, and no other control variables.[17] We see a significant association between productivity and several measures of online data activity – including data collection, data analysis and reporting, and the overall data score. Surprisingly, this is not the case for online data deployment, even though this variable appears to capture data-driven decision making behaviours previously shown to have had a positive effect on firm productivity by Brynjolfsson et al., (2011).

When we look at our coefficients, we find that data analysis and reporting has the strongest effect on performance – according to this model, firms that are one standard deviation above the mean in their levels of online data analysis and reporting (16 per cent of firms in the sample) are almost 15 per cent more productive.

In subsequent models, we estimate the contribution of each of our online data activities to productivity after including other firm controls. We do this in two steps: first, we add controls for firm-level characteristics (firm age, logarithm of average remuneration, online business share and IT employment share), and, second, we add measures of innovation (product innovation, and process innovation). In general, once we add these controls to our model, the size of the coefficients of online data activity variables on productivity are reduced in magnitude and, in the case of data collection, become insignificant. There is one exception to this – online data deployment – which actually becomes positive and significant after we include our second group of controls (model 8). Once we consider our innovation measures, this variable loses its significance again (model 9).

Our coefficient estimate for data analysis and reporting with all controls (model 6) shows that firms that are one standard deviation above the mean in that data activity are almost 11 per cent more productive. Our overall data score is also positive and significant. Using this measure, higher overall levels online data use are associated with over 8 per cent higher productivity (model 12).[18]

c) Complementarities between online data activities and employee autonomy

Table 5 presents the results of our tests of complementarity between online data activities and employee autonomy. There are two models for each of our online data activity measures. The first is a baseline with the measure and all control variables. The second adds the interaction between that online data activity and employee autonomy as well as the interaction between employee autonomy and IT share of employment as an extra control. This last interaction variable is included because we want to test for complementarities between online data activities and employee autonomy over and above those that past studies have shown exist between IT investments (which we proxy by IT employment) and autonomy (bearing in mind too, the importance of IT as a covariate for online data activities).

Our results suggest that there are significant complementarities between employee autonomy and online data activity (specifically, for data analysis and reporting, and the overall data score). Firms that intensify their data analysis and reporting while granting their employees autonomy experience a boost in productivity almost four times as high as those that are similarly intense in their data analysis and reporting but who have centralised decision making (18.6 per cent compared to 4.7 per cent).[19]

d) Complementarities between online data activities and process innovation

Table 6 presents the results of the complementarity tests between online data activities and process innovation, using the same structure as in Table 5, and conditioning on the interactions between process innovation and share of IT employment for similar reasons. Our results support the idea that process innovations and online data activities are complements – but there are some interesting variations across the data value chain.

In particular, data collection, which was insignificant in all previous models, presents a positive interaction with process innovation (even if only significant at the 10 per cent level). A literal interpretation is that there is a positive association between comprehensive data collection and productivity only in firms that innovate in their processes. We also find evidence of a positive complementarity between data deployment and process innovation. We interpret this finding as suggestive that involvement in process innovation influences whether firms are able to benefit from deploying data to make decisions in different areas of their business. In contrast with our previous tests of employee autonomy, however, we find no evidence of complementarities between process innovation and data analysis and reporting.

e) Robustness

Table 7 examines the robustness of our results for data collection (Panel A), data analysis and reporting (Panel B), data deployment (Panel C) and the overall data score (Panel D).

Using average remuneration per employee as a proxy for human capital in the firm can be problematic since it not only captures the skill level of employees but also the employees' ability to extract better wages from their employers, which in turn is associated with the profitability of the firm. Because of this, columns 1–4 re–estimate the main models excluding average remuneration from the regression, with very similar results with regards to the impact of data on firm performance for all our variables, but much weaker complementarities.

Another question with our results is whether treating our measures of online data activity as continuous is the best choice, in particular given the existence of measurement error (this is a particular concern with data collection and data deployment, both of which were based on self–reported five–point Likert scales). Columns 5–8 use a 1–0 dummy variable instead of a continuous measure, which takes the value 1 if the underlying indicator is above the median and 0 otherwise. These dichotomous variables reveal a stronger association between online data activity and productivity than is the case with the continuous measures. In particular, we note that when we use a dummy measure of data deployment instead of the continuous measure, the coefficient for this variable turns positive and significant: one interpretation of this is that measurement error in our original metric for data deployment masks what is in fact a positive link between this online data activity and productivity, in line with previous findings (Brynjolfsson et al., 2011). Finally, the coefficients capturing complementarities between process innovation and online data activity lose their significance when a dummy measure of data activity is used.

We have also tested for the sensitivity of the results when restricting the sample to the later years for which we have good coverage in terms of financial data (2010 and 2011), when arguably the assumption that firms' online analytics behaviours and IT employment shares are constant is more plausible. The downside is that the number of observations is lower, so less information is available to estimate the production function parameters. Reassuringly, the estimation produces very similar results (columns 9–12).

In addition, we have carried out some other unreported tests to establish whether the results are robust to how we have constructed our online data activities variables and how we have specified the regression models. These include considering different treatments of missing values and carrying out our estimations without first winsorising the measures of financial performance, both of which lead to very similar results. We have also considered what happens when considering multiple variables of data online activities simultaneously in one regression, in which case data analysis and reporting dominates the others (which typically lose their significance), and when exploring the complementarities of online data activities with autonomy and process innovation together, which leads to weaker estimates for both of them.

Finally, in Table 8 we explore how robust our main findings are to the use of alternative measures of firm performance, and in particular profitability. We consider three alternative accounting measures of profitability: EBITDA per employee, return on assets (ROA) and return on equity (ROE).[20] Each of the four panels shows the results of estimating the impact of our online data activities (and its interactions with employee autonomy and process innovation) on the three measures of profitability. We use the same model as before but include capital intensity as an additional control variable.[21]

Although the main effects of data use on these profitability metrics are positive, they are statistically insignificant with one main exception: data analysis and reporting is positive and statistically significant for two of our profitability metrics – EBITDA per employee and return on equity. For example, column 2 in Panel B suggests that firms that are one standard deviation above the average in their data analysis and reporting measure generate an additional operating profit of £3,180 per employee. Looking at column 10 in that same panel, firms that are more intensive (i.e. one standard deviation above the mean) in their data analysis and reporting generate a return on equity 4.3pp higher than the average.

The results relating to complementarities between data and employee autonomy and process innovation are less consistent when looking at profitability. Not only are the coefficients of the interactions typically insignificant, but they often also have a negative sign. In fact, when looking at the return on equity (column 12 in Panel B), the interaction of data with process innovation is negative and significant at the 10 per cent level. In contrast, we find a positive and significant interaction of data analysis and reporting with employee autonomy when looking at return on assets.

f) Discussion and implications

Our results strongly support the idea that firms that engage more deeply in online data activities are more productive, even after controlling for a host of covariates which we expect to affect firm productivity. Specifically, firms whose levels of online data use are one standard deviation above the average have around 8 per cent higher productivity. Our results are particularly robust for the data analysis and reporting stage of the data value chain. This variable is also significantly linked to two of the profitability metrics that we have considered in our robustness tests – EBITDA per employee and return on equity.

Our results are consistent with the argument, though given the cross-sectional nature of the survey data not conclusive proof, that firms can enhance their business performance by using their data more intensely.[22]

We find substantial differences in the effects of specific online data activities on performance when we consider them independently from each other, however. For example, our results suggest that more comprehensive data collection does not on its own contribute to business performance, echoing both a finding from the literature on ICT adoption and an idea often discussed in case studies and management magazines, namely that collecting data alone does not yield benefits unless the data is analysed and the resulting insights used to inform action (LaValle et al., 2011). In contrast, online data analysis and reporting – that is, how many techniques a firm deploys to analyse its data (ranging from basic descriptive analysis and customer segmentation to controlled experiments or data and text mining), and how they report the insights (through reporting, dashboard and visualisations, reporting of trends, etc.) – is very strongly associated with higher productivity. This result highlights the importance of extracting reliable insights from online data using a variety of techniques, and communicating them effectively to their users. In other words, the benefits of becoming an 'analytical firm'. To the extent to which adopting these analytical techniques may require specialist and 'deep' analytical skills, our finding lends weight to Hal Varian's famous quip that *"the sexy job in the next ten years will be statisticians."* (McKinsey and Company, 2009).

The lack of significance in the association between data deployment and productivity is somewhat puzzling in the light of previous findings in the literature (Brynjolfsson et al., 2011), and the growing importance of 'data-driven decision making' inside companies (Davenport and Harris, 2007; MGI, 2011). This discrepancy may, however, be explained by measurement error in our continuous measure of data deployment. The finding that when we instead consider a dichotomous measure of data deployment in our robustness tests the association between data deployment and productivity becomes positive and significant (if only at the 10 per cent level) is consistent with this explanation.

Our first set of complementarity tests – between online data activities and the extent of employee autonomy – indicate that using online data is particularly beneficial for those firms whose organisational structures are decentralised, and where employees are empowered to make decisions informed by the increasing amounts of data that are available to them. As Steve Ballmer, Microsoft's CEO put it recently in a memo to employees: *"As a company, we need to make the right decisions, and make them more quickly, balancing all the customer and business imperatives. Each employee must be able to solve problems more quickly and with more real-time data than in the past."* Consistent with Ballmer's steer to his staff, our results shows that there are particularly strong complementarities between employee autonomy and data analysis and reporting (Thusoo, 2009). In other words, firms seeking to boost their performance by prioritising analysis should also empower their employees to act on the insights with some autonomy.

Our second set of complementarity tests – between online data activities and process innovation – is less clear-cut, with some of the results losing significance in our robustness analysis. Nonetheless, they are broadly in line with the idea that the benefits of online data will be more likely realised by those firms that adapt their business processes. For instance, manufacturing firms that are integrating 'demand sensing' techniques based on web data into their logistics chains to manage inventories more efficiently (Wheatley, 2013), or 'lean start-ups' where entrepreneurs rapidly iterate their services in 'live' data-rich environments, (Croll and Yoskovitz, 2013; Ries, 2011).

5. CONCLUSIONS

In this study we have examined the economic realities behind the hype about the 'data revolution' in a sample of 500 UK firms with 50 employees and above which are commercially active online.

Our findings suggest that online data is making a substantial contribution to the productivity of firms. Activities related to the analysis and the reporting of online data appear to play a critical role, underscoring organisational psychologist Herbert Simon's remark about the importance of deploying attention more effectively as it becomes scarcer in a data-rich world (Shapiro and Varian, 1998; Simon, 1996).

The findings raise an obvious question: Why, given these apparent economic benefits, are only a minority of the firms in our sample heavily involved in data collection, analysis and reporting, and deployment? For example, only around a quarter of survey respondents collect online customer transaction data, and only 20 per cent say that online data plays a very important role in the formulation of their business strategy (Bakhshi and Mateos-Garcia, 2012). Only a small minority of 18 per cent – the datavores – say that they primarily rely on data and analysis when making decisions aimed at growing their sales; 43 per cent say that they prefer to use intuition and experience when making these decisions.

This disconnect between the levels of online data activity and the benefits that we estimate may in part be explained by our other finding that firms need to introduce complementary changes in order to reap the full returns from their online data activity. This may include disruptive – and therefore possibly controversial – changes to their organisational structures and business processes.

The link between data benefits and employee autonomy is particularly interesting: historically, the incorporation of new types of knowledge in the firm has gone hand in hand with changes in the organisation of work, the skills content of the workforce, and the emergence of new corporate functions. Are we, perhaps, seeing something similar as a consequence of the bigger volumes of data that are becoming available for firms to analyse and deploy? Ongoing debates about the need for more 'data scientists' and 'Chief Data Officers' suggest this might well be the case, as do the results of our complementarity tests, which lend weight to the idea that the data boom may be putting a premium on employee creativity, with potentially substantial implications for educational policy and management practice.

We think it is especially important to reach a more precise understanding of the mechanisms through which the joint presence of autonomy and data are linked to business performance. Is it by reducing intra-organisational communication costs and increasing flexibility, or by allowing the more effective use of workers' knowledge in decision making? Or is it because data-driven organisations allow their employees to take the initiative and (sometimes) fail, and are therefore more innovative as a result? We aim to explore these issues in our future research.

BIBLIOGRAPHY

Aghion, P., Bloom, N. and Van Reenen, J. (2013) 'Incomplete Contracts and the Internal Organization of Firms.' Cambridge MA: NBER.

Autor, D. H., Levy, F. and Murnane, R. J. (2003) The Skill Content of Recent Technological Change: An Empirical Exploration. 'The Quarterly Journal of Economics.' 118(4), 1279–1333. doi:10.1162/003355303322552801.

Bakhshi, H. and Mateos-Garcia, J. (2012) 'Rise of the Datavores.' London: Nesta.

Bloom, N., Garicano, L., Sadun, R. and Van Reenen, J. (2009) 'The distinct effects of Information Technology and Communication Technology on firm organization.' Cambridge MA: NBER.

Bresnahan, T. F., Brynjolfsson, E. and Hitt, L. M. (2002) Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence. 'The Quarterly Journal of Economics.' 117(1), 339–376. doi:10.1162/003355302753399526.

Brynjolfsson, E. and McAfee, A. (2012) 'Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity and Irreversibly Transforming Employment and the Economy.' (p. 100). Digital Frontier Press.

Brynjolfsson, E., Hitt, L. M. and Kim, H. H. (2011) Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? 'SSRN Electronic Journal.' doi:10.2139/ssrn.1819486.

Brynjolfsson, E. and Saunders, A. (2010) 'Wired for Innovation.' Cambridge MA: MIT Press.

Butler, D. (2013) When Google got flu wrong. 'Nature.' 494(7436), 155–6. doi:10.1038/494155a.

Corrado, C., Hulten, C. and Sichel, D. (2009) Intangible Capital and US Economic Growth. 'Review of Income and Wealth.' 55(3), 661–685. doi:10.1111/j.1475-4991.2009.00343.x.

Crespi, G., Criscuolo, C. and Haskel, J. (2007) 'Information technology, organisational change and productivity growth: evidence from UK firms.' London: Centre for Economic Performance, London School of Economics and Political Science.

Croll, A. and Yoskovitz, B. (2013) 'Lean Analytics: Use Data to Build a Better Startup Faster.' (p. 409). Sebastopol CA: O'Reilly Media, Inc.

Cukier, K. and Mayer-Schonberger, V. (2013) 'Big Data: A Revolution That Will Transform How We Live, Work and Think.' (p. 256). London: Hachette UK.

Davenport, T. H. and Harris, J. G. (2007) 'Competing on Analytics: The New Science of Winning.' (p. 218). New York: Harvard Business Press.

Divol, R., Edelman, D. and Sarrazin, H. (2012) 'Demystifying social media.' Retrieved from http://www.mckinsey.com/insights/marketing_sales/demystifying_social_media

Economist Intelligence Unit (2011) 'Big data Harnessing a game-changing asset.' London: Economist Intelligence Unit.

Garicano, L. and Wu, Y. (2012). Knowledge, communication, and organizational capabilities. 'Organization science.' 23(5), 1382–1397.

Goodridge, P., Haskel, J. and Wallis, G. (2012) UK Innovation Index: Productivity and Growth in UK Industries. 'SSRN Electronic Journal.'

Grant, R. M. (1996) Toward a knowledge-based theory of the firm. 'Strategic Management Journal.' 17, 109–122.

Grant, R. M. and Cibin, R. (1996) Strategy, structure and market turbulence: The international oil majors, 1970–1991. 'Scandinavian Journal of Management.' 12(2), 165–188.

Hamilton, B. (2003) EBITDA: Still Crucial to Credit Analysis. 'Commercial Lending Review.' September 2003.

IBM (2012) IBM Analytics – IT Business Intelligence – United Kingdom. IBM Corporation – Smarter Planet. Retrieved 29 October, 2013, from http://www.ibm.com/smarterplanet/uk/en/business_analytics/article/it_business_intelligence.html

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G. and Haydock, M. (2011) Analytics: The widening divide. 'MIT Sloan Management Review.' 1–20.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S. and Kruschwitz, N. (2011) Big data, analytics and the path from insights to value. 'MIT Sloan Management Review.' 52(2), 21–31.

McKinsey and Company (2009) 'Hal Varian on how the Web challenges managers.' Retrieved from: http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers

MGI (2011) 'Big data: The next frontier for innovation, competition, and productivity.' McKinsey Global Institute.

MGI (2013) 'Disruptive technologies: Advances that will transform life, business, and the global economy.' McKinsey Global Institute.

Nelson, R. R. and Winter, S. (1982) 'An evolutionary theory of economic change.' Cambridge, MA: Harvard University Press.

OECD (2013) 'New Sources of Growth: Knowledge–Based Capital.' Paris: OECD.

ONS (2011) 'E–commerce and ICT activity.' 2010 Edition. Retrieved from http://www.ons.gov.uk/ons/dcp171778_245829.pdf

Provost, F. and Fawcett, T. (2013) 'Data Science for Business.' Sebastopol CA: O'Reilly Media.

Radner, R. (1993) The organization of decentralized information processing. 'Econometrica: Journal of the Econometric Society.' 1109–1146.

Ries, E. (2011) 'The Lean Startup: How Constant Innovation Creates Radically Successful Businesses.' (p. 336). London: Penguin UK.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. 'Journal of Information Science.' 33(2), 163–180. doi:10.1177/0165551506070706.

Shapiro, C. and Varian, H. (1998) 'Information Rules.' (p. 352). New York: Harvard Business Press.

Simon, H. A. (1996) Designing organizations for an information–rich world. 'International Library of Critical Writings in Economics.' 70, 187–202.

Tambe, P. (2013) Big Data Investment, Skills, and Firm Value. 'SSRN Electronic Journal.' doi:10.2139/ssrn.2294077

Thusoo, A. (2009) 'Hive – A Petabyte Scale Data Warehouse using Hadoop.'

Vanderbilt, T. (2013) The Nielsen family is dead. 'Wired.' abr. 19 March.

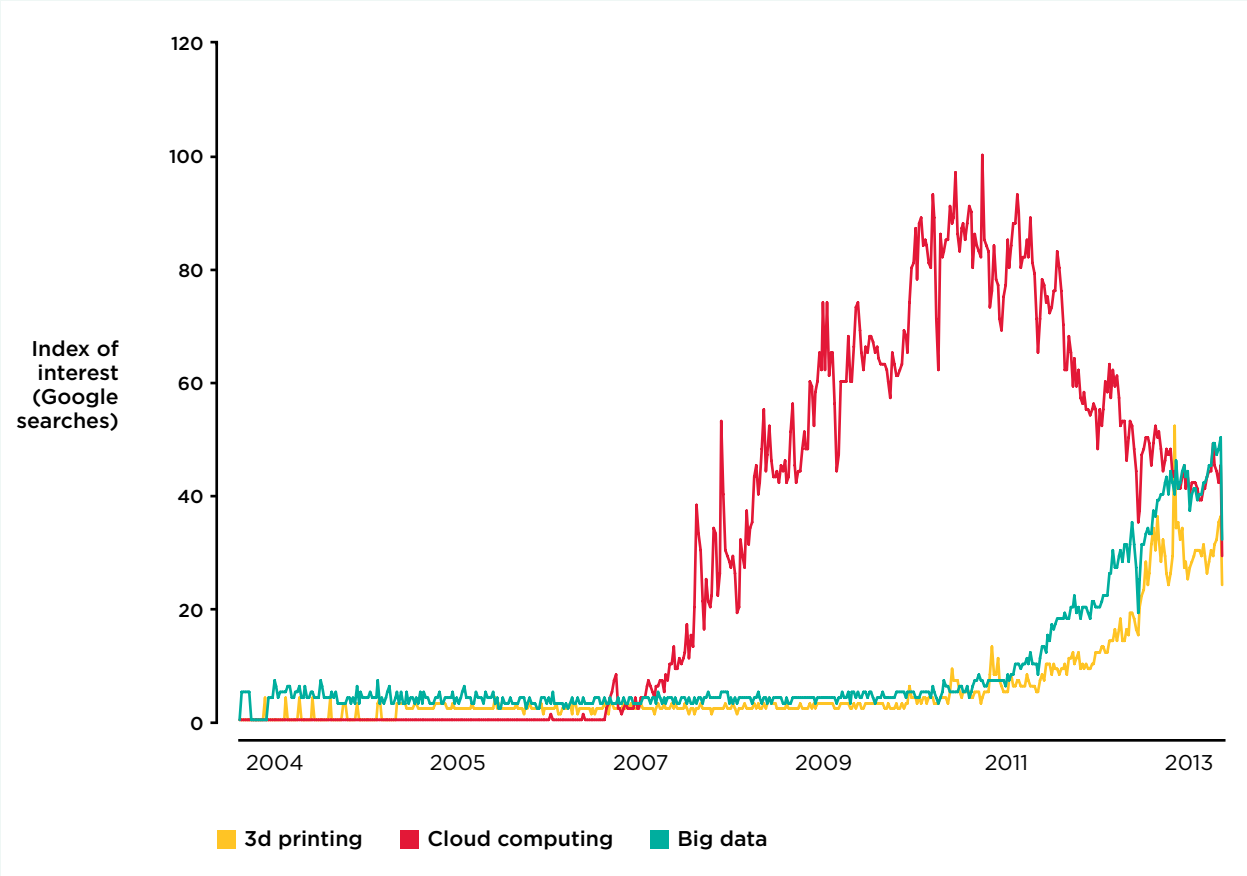
Wheatley, M. (2013) The big three for big data. 'The Manufacturer.' 7 May.

ENDNOTES

1. We define this as firms that are either involved in commercial transactions online, generate revenue through adverts in their websites, or pay for advertisements online. We discuss this further in Section 3, and in Bakhshi and Mateos-Garcia (2012).
2. See Provost and Fawcett (2013) for a discussion of how these concepts relate to each other in a business context.
3. It is worth noting that these models might involve an understanding of the causal mechanisms that link the relevant variables, or instead be based on the existence of mere statistical correlations between these variables. Cukier and Mayer-Schonberger (2013) claim that 'big data' increases the potential of correlations as a guide for making decisions. For example, it argues, Amazon does not need a theory of its users' preferences to recommend to them new products, relying instead on their past choices and those of similar users. Furthermore, 'data-based actions' may be taken by algorithms (i.e. an expert system which flags up a suspect transaction, or a high-frequency trading program) as well as by humans. It is increasingly widely believed that 'big data' will be a driver in the automation of physical 'non-routine' and knowledge work (Brynjolfsson and McAfee 2012; MGI, 2013).
4. The idea of the 'data value chain' is related to the 'data information knowledge wisdom' hierarchy (Rowley, 2007).
5. Dummy variables take the value of 1 in the case that the firm is in a particular industry and 0 otherwise.
6. We acknowledge potential problems with our proxy for human capital (i.e. if it simply captures a firms' ability to pay higher wages to its personnel), and in some of our robustness tests we have excluded it from our models.
7. For detailed information on the design and testing of the questionnaire, see Bakhshi and Mateos-Garcia (2012). The final version of the questionnaire can be downloaded from: http://www.nesta.org.uk/library/documents/Datavores_Questionaire.pdf
8. The final list of sectors includes Wholesale, Retail, Business Support Services, Information and Communications, Knowledge Intensive Business Services, Manufacturing, Professional Activities and Other.
9. A screener question sought to ensure that the respondent was able to provide information about online data analytics in their firm. If not, they were given the option of nominating another individual.
10. Four in ten respondents were excluded at that stage of the survey (ibid).
11. Those firms that provided IT data were significantly smaller than those that did not ($p=0.0000$). A chi2 test finds that there are significant differences ($p=0.075$) between the sectoral distribution of those firms which provided IT data and those which did not. In particular, Manufacturing and KIBS firms were overrepresented in the group of firms that provided IT data, and business services, financial services and professional services were underrepresented. In our econometric models we test whether these differences have significant implications for the estimated relationships between analytics and firm productivity and we find that they do not.
12. We have replaced 'don't know' scores in these measures of online data activity with the mean for all other observations for which we had data. The purpose of this was to avoid situations where we had to drop a firm from our analysis as a consequence of a 'don't know' score for a single item within one indicator. We acknowledge the potential measurement errors introduced by this imputation approach. As a robustness test, we have estimated all our models with two alternative treatments of 'don't know' scores – making all 'don't know' scores missing values, and replacing them with the lowest possible score for that item. All our findings are robust to these changes, although our estimators naturally become less precise when we treat 'don't knows' as missing values, because of the reduction in sample size.
13. For each component, we subtract from a firm's score the sample mean, and divide by the standard deviation. We add them into the composite measure and normalise again. As a consequence, their mean is zero, and their standard deviation is one.
14. Alternatively, we carried out a factor analysis of our individual components to extract summary indicators. Each of our data measures is correlated with its factor analysis version with $p>0.99$.
15. These data are available from the authors on request.
16. Our results are robust when we repeat our analysis without winsorisation.
17. We do not report the coefficients for production factors, industry and year fixed effects and interactions between production factors and industry fixed effects for clarity of presentation, but all these results are available on request.
18. The three measures of online data activities are highly collinear so the results of 'horse-race' regressions that include the three of them simultaneously would need to be interpreted with care. In unreported regressions where we do so we find that data analysis and reporting is always statistically significant, while the other two metrics are not.
19. In this illustration, we are comparing firms with one standard deviation above the mean in their data analysis and reporting, and autonomy variables, with firms one standard deviation above the mean in their data analysis and reporting, and one standard deviation below the mean in their autonomy (i.e. centralised).
20. EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) captures a firm's ability to generate healthy net profits from operations – it is generally considered a good measure of 'how well a company is managing revenues and costs' over time (Hamilton, 2003). Return on assets (profits and losses before taxes over total assets) measures the rate of return on a firm's invested capital, while Return on equity (profits and losses before taxes over shareholder funds) measures the rate of return on the capital invested by shareholders.
21. This controls for the fact that EBITDA per employee is typically higher for firms with higher capital intensity. However, we get similar results if we do not control for capital intensity.
22. Some potentially fruitful ways of establishing causality include collecting panel data through longitudinal firm surveys, by drawing on web data sources which capture relevant dimension of a firm's online data activity, and running a controlled experiment involving random assignment to firms of an 'online data analytics' intervention. We are exploring all of these options in our future research.

FIGURES AND TABLES

Figure 1: **Search trends for 'big data' compared with other technology and business areas**



Source: **Google Insights for Search.UK.**

Table 1: **Survey–based measures description**

Variable	Definition	N	Mean	Sd	Min	Max	Alpha
Data collection	**Online data collection indicator (standardised)** *Based on 1–5 scores for 5 items: Comprehensiveness in collection of online transaction data, service and support data, user online activity data, marketing data and lifestage data.*	500	0	1	-2.06	2.02	0.93
Data analysis & reporting	**Online data analysis and communication indicator (standardised)** *Based on binary scores (1 or 0) for 7 items: Adoption of A/B tests, trend analysis and reporting, forecasting, dashboard and visualisations, segmentation, regression and propensity score modelling, and data and text mining.*	500	0	1	-1.46	2.07	0.93
Data deployment	**Online data use indicator (standardised)** *Based on 1–5 scores for 9 items: Importance of online data for making decisions regarding customer segmentation, tailoring of marketing and sales, developing products and services suited to customers, improving the website, predicting customer behaviour, reporting on performance, informing business strategy, optimising pricing, and designing and evaluating social media strategy.*	500	0	1	-2.05	1.74	0.98
Data score	**Combined online data use indicator (standardised)** *Based on the scores for Data collection, Data analysis and reporting, and Data deployment above.*	500	0	1	-2.19	2.08	0.8
Product innovation	**Respondent launches goods and services ahead of competitors (standardised)**	484	0	1	-1.86	1.38	
Process innovation	**Respondent is willing to disrupt its business processes (standardised)**	484	0	1	-2.07	1.50	
Online business share	**Proportion of revenues generated through the website**	427	0.15	0.20	0.00	1.00	
IT employment share	**IT employees as a proportion of the workforce in 2010**	300	0.04	0.08	0.00	0.74	
Autonomy	**Decentralisation indicator (standardised)** *Based on 1–5 scores for three items: workers set the pace of work, workers decide how tasks should be performed, people are free to try new things.*	480	0	1	-2.63	2.17	0.59

Table 2: **FAME–based descriptive statistics (2010)**

Variable	N	Mean	Sd	Min	Max
Number of employees	496	455	1046	50	8781
Firm age (years)	500	23	21	0	115
Value added (£000s)	497	20376	46244	338	348817
Tangible assets (£000s)	496	12052	44060	18	337879
Capital intensity (K/L) (£000s)	496	24	56	0	448
Average remuneration (£000s)	496	32	15	1	86
EBITDA per employee (£000s)	496	15	27	-54	175
Return on assets (%)	496	7	13	-48	53
Return on equity (%)	440	23	51	-188	270

Table 3: **Correlation matrix**

	Data collection reporting	Data analysis and	Data deployment	Data score	Firm age	Log (Average remuneration)	Online business share	IT employment share	Product innovation	Process innovation	Autonomy
Data collection	1 500										
Data analysis and reporting	0.5613* 500	1 500									
Data deployment	0.6342* 500	0.5230* 500	1 500								
Data score	0.8653* 500	0.8215* 500	0.8503* 500	1 500							
Firm age	-0.0097 500	0.0202 500	-0.0095 500	0.0004 500	1 500						
Log (Average remuneration)	0.0018 496	0.1284* 496	-0.0472 496	0.0327 496	-0.0236 496	1 496					
Online business share	0.3428* 427	0.3176* 427	0.3054* 427	0.3792* 427	0.0021 427	0.0291 423	1 427				
IT employment share	0.0191 300	0.0535 300	0.0517 300	0.0487 300	-0.0651 300	0.1766* 300	0.0209 259	1 300			
Product innovation	0.2714* 484	0.2458* 484	0.2305* 484	0.2940* 484	0.015 484	0.0867 480	0.1405* 421	0.0045 293	1 484		
Process innovation	0.2073* 484	0.2718* 484	0.2228* 484	0.2765* 484	-0.0744 484	-0.0026 480	0.1740* 418	0.1479* 293	0.2175* 474	1 484	
Autonomy	0.1852* 480	0.1820* 480	0.1955* 480	0.2216* 480	0.0032 480	0.0686 476	0.1589* 416	0.0903 292	0.1866* 472	0.2964* 473	1 480

* indicates significant at the 5 per cent level. Average remuneration data for 2010.

Table 4: **Online data activities and firm productivity**

This table estimates a standard production function, allowing the coefficients for capital (K) and labour (L) to vary across industries. The dependent variable is Log (Value added). Data score corresponds to the average of the Data collection, Data analysis and reporting, and Data deployment indicators. All regressions include industry and year fixed effects, as well as production factors, Log (K) and Log (L), interacted with industry. The table reports coefficients estimated with OLS, with robust standard errors clustered at firm level in parentheses. ***, **, * indicate significance levels of 1 per cent, 5 per cent and 10 per cent respectively.

	Log (Value added)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data collection	0.0775** (0.0326)	0.0598 (0.0392)	0.0423 (0.0428)									
Data analysis and reporting				0.147*** (0.0352)	0.110*** (0.0409)	0.109** (0.0427)						
Data deployment							0.0433 (0.0326)	0.0789** (0.0361)	0.0647 (0.0393)			
Data score										0.105*** (0.0340)	0.0979** (0.0384)	0.0877** (0.0423)
Firm age		-0.00342 (0.00213)	-0.00372* (0.00210)		-0.00354* (0.00206)	-0.00390* (0.00204)		-0.00355* (0.00209)	-0.00383* (0.00207)		-0.00348* (0.00210)	-0.00382* (0.00206)
Log (Average remuneration)		1.009*** (0.0896)	1.038*** (0.0935)		0.987*** (0.0868)	1.017*** (0.0912)		1.015*** (0.0888)	1.043*** (0.0928)		1.005*** (0.0881)	1.037*** (0.0929)
Online business share		0.143 (0.298)	0.182 (0.289)		0.0554 (0.276)	0.0898 (0.272)		0.130 (0.291)	0.163 (0.285)		0.0622 (0.288)	0.105 (0.281)
IT employment share		0.780 (0.615)	0.166 (0.450)		0.714 (0.588)	0.130 (0.426)		0.723 (0.600)	0.146 (0.441)		0.733 (0.589)	0.165 (0.442)
Product innovation			0.000295 (0.0363)			-0.00691 (0.0368)			0.000657 (0.0357)			-0.00981 (0.0363)
Process innovation			0.0398 (0.0325)			0.0313 (0.0320)			0.0354 (0.0323)			0.0324 (0.0324)
Observations	2,119	1,090	1,059	2,119	1,090	1,059	2,119	1,090	1,059	2,119	1,090	1,059
R-squared	0.644	0.749	0.760	0.654	0.754	0.766	0.642	0.751	0.762	0.647	0.753	0.763

Table 5: **Complementarities between data activities and employee autonomy**

This table expands the baseline model to include a measure of employee autonomy interacted with indicators of online data activity. Autonomy is a standardised average of three survey–based indicators: workers set the pace of work, workers decide how tasks should be performed, and people are free to try new things. All regressions include industry and year fixed effects, as well as production factors Log (K) and Log (L) interacted with industry. Unreported control variables are Firm age, Log (Average remuneration), Online business share, IT employment share, Product innovation and Process innovation. The table reports coefficients estimated with OLS, with robust standard errors clustered at firm level in parentheses. ***, **, * indicate significance levels of 1 per cent, 5 per cent and 10 per cent respectively.

	Log (Value added)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Data collection	0.0423 (0.0428)	0.0519 (0.0432)						
Data collection x Autonomy		0.0439 (0.0296)						
Data analysis and reporting			0.109** (0.0427)	0.117*** (0.0431)				
Data analysis and reporting x Autonomy				0.0697** (0.0293)				
Data deployment					0.0647 (0.0393)	0.0727* (0.0402)		
Data deployment x Autonomy						0.0428 (0.0330)		
Data score							0.0877** (0.0423)	0.0972** (0.0434)
Data score x Autonomy								0.0592* (0.0317)
Autonomy		-0.0470 (0.0403)		-0.0384 (0.0394)		-0.0421 (0.0411)		-0.0461 (0.0405)
IT employment share x Autonomy		0.270 (0.723)		0.198 (0.713)		0.227 (0.728)		0.270 (0.737)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,059	1,051	1,059	1,051	1,059	1,051	1,059	1,051
R-squared	0.760	0.759	0.766	0.767	0.762	0.760	0.763	0.764

Table 6: **Complementarities between data activities and process innovation**

This table expands the baseline model to include a measure of organizational restructuring interacted with indicators of online data activity. Process innovation is a survey–based standardised indicator of a firm's willingness to disrupt its business processes. All regressions include industry and year fixed effects, as well as production factors Log (K) and Log (L) interacted with industry. Unreported control variables are Firm age, Log (Average remuneration), Online business share, IT employment share, Product innovation and Process innovation. The table reports coefficients estimated with OLS, with robust standard errors clustered at firm level in parentheses. ***, **, * indicate significance levels of 1 per cent, 5 per cent and 10 per cent respectively.

	Log (Value added)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Data collection	0.0423 (0.0428)	0.0536 (0.0425)						
Data collection x Process innovation		0.0630* (0.0338)						
Data analysis and reporting			0.109** (0.0427)	0.112*** (0.0422)				
Data analysis and reporting x Process innovation				0.0297 (0.0332)				
Data deployment					0.0647 (0.0393)	0.0636* (0.0377)		
Data deployment x Process innovation						0.0720** (0.0337)		
Data score							0.0877** (0.0423)	0.0899** (0.0408)
Data score x Process innovation								0.0640* (0.0328)
IT employment share x Process innovation		0.500 (0.525)		0.551 (0.517)		0.406 (0.534)		0.509 (0.526)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,059	1,059	1,059	1,059	1,059	1,059	1,059	1,059
R-squared	0.760	0.763	0.766	0.767	0.762	0.765	0.763	0.767

Table 7: **Robustness regressions**

Columns 1–4 estimate the baseline model without including Log(Average remuneration) as a control. Columns 5–8 estimate the baseline model using a dummy instead of a continuous score for Data collection (top panel) and Analysis & reporting (bottom panel), which takes the value 1 if the underlying indicator is above the median and 0 otherwise. Columns 9–12 estimate the baseline model restricting the sample to observations for the years 2010 and 2011. All regressions include industry and year fixed effects, as well as production factors Log(K) and Log(L) interacted with industry. Control variables are Firm age, Log(Average remuneration), Online business share, IT employment share, Product innovation and Process innovation. Regressions that include Autonomy interacted with data usage also include Autonomy and its interaction with IT employment share. Similarly, when an interaction with Process Innovation is included, its interaction with IT employment share is as well. The table reports coefficients estimated with OLS, with robust standard errors clustered at firm level in parentheses. ***, **, * indicate significance levels of 1 per cent, 5 per cent and 10 per cent respectively.

Panel A: **Data collection**												
	Excluding Log (Average remuneration)				Data collection dummy				2010–2011			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data collection	0.0775** (0.0326)	0.0199 (0.0480)	0.0239 (0.0490)	0.0252 (0.0485)	0.131* (0.0680)	0.0953 (0.0943)	0.112 (0.0956)	0.102 (0.0945)	0.0802** (0.0328)	0.0436 (0.0434)	0.0605 (0.0436)	0.0573 (0.0423)
Data collection x Autonomy			0.0119 (0.0311)				0.0812 (0.0683)				0.0735** (0.0319)	
Data collection x Process innovation				0.0368 (0.0366)				0.0775 (0.0744)				0.0909** (0.0355)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,119	1,059	1,051	1,059	2,119	1,059	1,051	1,059	924	471	467	471
R-squared	0.644	0.644	0.639	0.645	0.643	0.760	0.759	0.762	0.651	0.775	0.777	0.781

Panel B: **Data analysis and reporting**												
	Excluding Log (Average remuneration)				Analysis and reporting dummy				2010–2011			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data analysis and reporting	0.147*** (0.0352)	0.139*** (0.0504)	0.146*** (0.0517)	0.140*** (0.0500)	0.292*** (0.0699)	0.238*** (0.0771)	0.255*** (0.0815)	0.244*** (0.0768)	0.146*** (0.0360)	0.0992** (0.0427)	0.111** (0.0437)	0.101** (0.0432)
Data analysis and reporting x Autonomy			0.0645* (0.0362)				0.162*** (0.0610)				0.0691** (0.0306)	
Data analysis and reporting x Process innovation				0.0165 (0.0395)				0.0869 (0.0633)				0.0148 (0.0350)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,119	1,059	1,051	1,059	2,053	1,031	1,023	1,031	924	471	467	471
R-squared	0.654	0.654	0.653	0.654	0.658	0.766	0.768	0.768	0.660	0.779	0.781	0.780

Table 7: **Robustness regressions (continued)**

Columns 1–4 estimate the baseline model without including Log(Average remuneration) as a control. Columns 5–8 estimate the baseline model using a dummy instead of a continuous score for Data collection (top panel) and Analysis & reporting (bottom panel), which takes the value 1 if the underlying indicator is above the median and 0 otherwise. Columns 9–12 estimate the baseline model restricting the sample to observations for the years 2010 and 2011. All regressions include industry and year fixed effects, as well as production factors Log(K) and Log(L) interacted with industry. Control variables are Firm age, Log(Average remuneration), Online business share, IT employment share, Product innovation and Process innovation. Regressions that include Autonomy interacted with data usage also include Autonomy and its interaction with IT employment share. Similarly, when an interaction with Process Innovation is included, its interaction with IT employment share is as well. The table reports coefficients estimated with OLS, with robust standard errors clustered at firm level in parentheses. ***, **, * indicate significance levels of 1 per cent, 5 per cent and 10 per cent respectively.

Panel C: **Data deployment**												
	Excluding Log (Average remuneration)				Data deployment dummy				2010–2011			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data deployment	0.0433 (0.0326)	0.0327 (0.0450)	0.0358 (0.0468)	0.0319 (0.0443)	0.105 (0.0696)	0.141* (0.0760)	0.150* (0.0808)	0.140* (0.0758)	0.0403 (0.0333)	0.0472 (0.0391)	0.0569 (0.0393)	0.0444 (0.0370)
Data deployment x Autonomy			0.0113 (0.0376)				0.0892 (0.0679)				0.0586* (0.0349)	
Data deployment x Process innovation				0.0409 (0.0373)				0.0899 (0.0663)				0.0735** (0.0364)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,119	1,059	1,051	1,059	2,119	1,059	1,051	1,059	924	471	467	471
R-squared	0.642	0.644	0.640	0.645	0.645	0.764	0.762	0.765	0.654	0.777	0.780	0.781

Panel D: **Overall data score**												
	Excluding Log (Average remuneration)				Data dummy				2010–2011			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data score	0.105*** (0.0340)	0.0784 (0.0477)	0.0844* (0.0500)	0.0786* (0.0468)	0.172** (0.0671)	0.178** (0.0786)	0.186** (0.0827)	0.180** (0.0786)	0.104*** (0.0347)	0.0773* (0.0423)	0.0919** (0.0428)	0.0765* (0.0404)
Data score x Autonomy			0.0321 (0.0344)				0.0690 (0.0682)				0.0764** (0.0334)	
Data score x Process innovation				0.0379 (0.0364)				0.0709 (0.0660)				0.0695** (0.0349)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,119	1,059	1,051	1,059	2,119	1,059	1,051	1,059	924	471	467	471
R-squared	0.647	0.647	0.643	0.648	0.645	0.764	0.762	0.765	0.654	0.777	0.780	0.781

Table 8: **Profitability regressions**

Columns 1–4 estimate the baseline model using EBITDA per employee (£000s) as dependent variable, while columns 5–8 consider Return on assets and columns 9–12 Return on equity. All regressions include industry and year fixed effects, as well as production factors Log(K) and Log(L) interacted with industry. All regressions include Capital intensity as an additional control. Baseline control variables are Firm age, Log (Average remuneration), Online business share, IT employment share, Product innovation and Process innovation. Regressions that include Autonomy interacted with data usage also include Autonomy and its interaction with IT employment share. Similarly, when an interaction with Process Innovation is included, its interaction with IT employment share is as well. The table reports coefficients estimated with OLS, with robust standard errors clustered at firm level in parentheses. ***, **, * indicate significance levels of 1 per cent, 5 per cent and 10 per cent respectively.

Panel A: **Data collection**												
	EBITDA per employee				Return on assets				Return on equity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data collection	0.498 (0.680)	-1.091 (1.062)	-0.995 (1.045)	-0.999 (1.072)	0.591 (0.495)	1.040 (0.664)	0.997 (0.691)	1.034 (0.659)	2.921* (1.755)	3.576 (2.381)	3.589 (2.289)	3.201 (2.248)
Data collection x Autonomy			0.136 (0.763)				0.317 (0.656)				-2.920 (2.107)	
Data collection x Process innovation				0.0578 (0.883)				-0.0433 (0.639)				-3.247 (2.452)
Capital intensity (K/L)	0.241*** (0.0382)	0.225*** (0.0465)	0.219*** (0.0454)	0.221*** (0.0440)	0.00786 (0.0113)	-0.00348 (0.0173)	0.000510 (0.0191)	-0.00341 (0.0169)	0.0308 (0.0455)	-0.0328 (0.0823)	-0.0234 (0.0849)	-0.0313 (0.0749)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,159	1,077	1,069	1,077	2,159	1,076	1,068	1,076	1,930	950	943	950
R-squared	0.295	0.399	0.401	0.403	0.041	0.068	0.082	0.069	0.059	0.102	0.104	0.107

Panel B: **Data analysis and reporting**												
	EBITDA per employee				Return on assets				Return on equity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data analysis and reporting	2.332** (0.907)	3.180*** (1.136)	3.339*** (1.218)	3.462*** (1.142)	0.0974 (0.480)	0.998 (0.709)	1.014 (0.728)	1.028 (0.699)	0.431 (1.579)	4.353* (2.360)	4.391* (2.415)	4.733** (2.365)
Data analysis and reporting x Autonomy			0.180 (0.969)				1.441** (0.639)				-1.078 (2.177)	
Data analysis and reporting x Process innovation				-1.096 ((1.137)				-0.382 (0.611)				-3.591* (2.019)
Capital intensity (K/L)	0.240*** (0.0378)	0.227*** (0.0463)	0.218*** (0.0449)	0.224*** (0.0445)	0.00768 (0.0114)	-0.00386 (0.0177)	-4.95e-05 (0.0203)	-0.00309 (0.0173)	0.0290 (0.0457)	-0.0368 (0.0835)	-0.0299 (0.0889)	-0.0332 (0.0780)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,159	1,077	1,069	1,077	2,159	1,076	1,068	1,076	1,930	950	943	950
R-squared	0.295	0.399	0.401	0.403	0.041	0.068	0.082	0.069	0.059	0.102	0.104	0.107

Table 8: **Profitability regressions (continued)**

Columns 1–4 estimate the baseline model using EBITDA per employee (£000s) as dependent variable, while columns 5–8 consider Return on assets and columns 9–12 Return on equity. All regressions include industry and year fixed effects, as well as production factors Log(K) and Log(L) interacted with industry. All regressions include Capital intensity as an additional control. Baseline control variables are Firm age, Log (Average remuneration), Online business share, IT employment share, Product innovation and Process innovation. Regressions that include Autonomy interacted with data usage also include Autonomy and its interaction with IT employment share. Similarly, when an interaction with Process Innovation is included, its interaction with IT employment share is as well. The table reports coefficients estimated with OLS, with robust standard errors clustered at firm level in parentheses. ***, **, * indicate significance levels of 1 per cent, 5 per cent and 10 per cent respectively.

Panel C: **Data deployment**												
	EBITDA per employee				Return on assets				Return on equity			
	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
Data deployment	-0.321 (0.844)	-1.361 (1.232)	-1.299 (1.322)	-1.321 (1.210)	-0.0108 (0.484)	-0.243 (0.674)	-0.326 (0.718)	-0.238 (0.680)	-0.120 (1.830)	-0.852 (2.304)	-0.688 (2.408)	-0.839 (2.276)
Data deployment x Autonomy			-1.559 (1.221)				0.477 (0.734)				-1.009 (2.474)	
Data deployment x Process innovation				-1.472 (0.997)				-0.194 (0.706)				-3.532 (2.587)
Capital intensity (K/L)	0.241*** (0.0381)	0.226*** (0.0459)	0.218*** (0.0447)	0.225*** (0.0432)	0.00770 (0.0115)	-0.00410 (0.0181)	0.00113 (0.0206)	-0.00352 (0.0175)	0.0291 -0.0456	-0.0354 -0.0828	-0.0278 -0.0869	-0.0326 -0.0732
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,159	1,077	1,069	1,077	2,159	1,076	1,068	1,076	1,930	950	943	950
R-squared	0.289	0.390	0.394	0.395	0.041	0.064	0.069	0.064	0.059	0.097	0.098	0.102

Panel D: **Overall data score**												
	EBITDA per employee				Return on assets				Return on equity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Data score	0.995 (0.792)	0.318 (1.024)	0.459 (1.115)	0.432 (1.007)	0.264 (0.499)	0.712 (0.704)	0.646 (0.748)	0.714 (0.696)	1.237 (1.767)	2.816 (2.416)	3.098 (2.517)	2.941 (2.373)
Data score x Autonomy			-0.457 (0.859)				0.823 (0.674)				-2.232 (2.355)	
Data score x Process innovation				-0.850 (0.892)				-0.274 (0.678)				-4.237* (2.513)
Capital intensity (K/L)	0.240*** (0.0383)	0.226*** (0.0470)	0.219*** (0.0456)	0.224*** (0.0449)	0.00764 (0.0114)	-0.00397 (0.0175)	0.000496 (0.0201)	-0.00324 (0.0167)	0.0290 (0.0456)	-0.0353 (0.0828)	-0.0284 (0.0869)	-0.0291 (0.0732)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	2,159	1,077	1,069	1,077	2,159	1,076	1,068	1,076	1,930	950	943	950
R-squared	0.290	0.388	0.390	0.391	0.041	0.066	0.073	0.066	0.059	0.099	0.102	0.106

Nesta...

Nesta

1 Plough Place
London EC4A 1DE

research@nesta.org.uk
www.twitter.com/nesta_uk
www.facebook.com/nesta.uk

www.nesta.org.uk

March 2014