Niebel, Thomas; Rasel, Fabienne; Viete, Steffen

Working Paper

BIG data - BIG gains? Empirical evidence on the link between big data analytics and innovation

ZEW Discussion Papers, No. 17-053

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Niebel, Thomas; Rasel, Fabienne; Viete, Steffen (2017) : BIG data - BIG gains? Empirical evidence on the link between big data analytics and innovation, ZEW Discussion Papers, No. 17-053, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, https://nbn-resolving.de/urn:nbn:de:bsz:180-madoc-437068

This Version is available at:
https://hdl.handle.net/10419/171456

Discussion Paper No. 17-053

BIG Data – BIG Gains? Empirical Evidence on the Link Between Big Data Analytics and Innovation

Thomas Niebel, Fabienne Rasel,
and Steffen Viete

ZEW

Zentrum für Europäische
Wirtschaftsforschung GmbH

Centre for European
Economic Research

Discussion Paper No. 17-053

BIG Data – BIG Gains? Empirical Evidence on the Link Between Big Data Analytics and Innovation

Thomas Niebel, Fabienne Rasel,
and Steffen Viete

BIG Data - BIG Gains? Empirical Evidence on the Link Between Big Data Analytics and Innovation

Thomas Niebel* Fabienne Rasel † Steffen Viete‡

Abstract

This paper analyzes the relationship between firms' use of big data analytics and their innovative performance in terms of product innovations. Since big data technologies provide new data information practices, they create novel decision-making possibilities, which are widely believed to support firms' innovation process. Applying German firm-level data within a knowledge production function framework we find suggestive evidence that big data analytics is a relevant determinant for the likelihood of a firm becoming a product innovator as well as for the market success of product innovations. These results hold for the manufacturing as well as for the service sector but are contingent on firms' investment in IT-specific skills. Subsequent analyses suggest that firms in the manufacturing and service sector rely on different data sources and data-related firm practices in order to reap the benefits of big data. Overall, the results support the view that big data analytics have the potential to enable innovation.

Keywords: Big data, data-driven decision-making, product innovation, firm-level data.

JEL Classification Numbers: D22, L20, O33.

*Corresponding author: ZEW Mannheim, *email*: niebel@zew.de, Centre for European Economic Research (ZEW) Mannheim, Digital Economy Research Department, P.O. Box 103443, 68034 Mannheim, Germany.

†ZEW Mannheim, rasel@zew.de

‡ZEW Mannheim, viete@zew.de.

For further information on the author's other projects see www.zew.de/staff_tni, www.zew.de/staff_frl and www.zew.de/staff_sve as well as the ZEW annual report on www.zew.de/en. We would like to thank Irene Bertschek, Chris Forman and participants at the ZEW ICT seminar, the 28th European Regional Conference of the International Telecommunication Society (ITS) in Passau, Germany, and the 6th European Conference on Corporate R&D and Innovation (CONCORDi) in Seville, Spain, for their helpful comments and suggestions. Kevin Krzyzok provided excellent research assistance. All remaining errors are ours alone.

1 Introduction

The latest technological trends like connected devices and machines, wearables, and the universal application of sensors as well as (user-generated) online content are drivers of a vast and constantly increasing amount of data. In reference to the large volumes of diverse data and associated new data information practices that have become available to firms, *big data analytics* has become an important topic among practitioners, policy makers and scientists. Broadly speaking, the concept of big data encompasses the amount and complexity of newly available data and the technical challenges of processing them (Dumbill, 2013). A narrower definition of the term, which is commonly used in the literature, highlights the following three characteristics: (1) an enormous amount of data (volume), (2) a wide variety of data coming from highly diverse sources (variety), and (3) the pace of data processing (velocity). Enormous progress in computing power, storage capacity, and software have been necessary for the surge of big data technologies.

Much of the debate and research has centered around possible implications of big data for firms and businesses. As big data alters the sources and types of information available to decision-makers in the firm, it is expected to impact on established ways of decision- and strategy-making which have traditionally relied on predefined data collected for specific needs (Constantiou and Kallinikos, 2015). In particular, data which has become available to firms is often not collected intentionally, but in a heterogenous and unstructured way (Varian, 2010; Anderson, 2008). The ability to analyze such data, extract insights and appropriate value from it represents a key challenge to firms. One problem big data poses to decision-making is that correlations identified from the raw data are erroneously interpreted as causal relationships or that misleading patterns are found in the data (Lazer et al. 2014; McAfee and Brynjolfsson 2012). Starting from such data patterns found with big data analytics, decisions without potential for improvement or even unwise decisions can be made. That is why the use of big data analytics may not guarantee sustainable, positive effects on firm performance (‚Big Gains'). The grey areas with respect to privacy, data protection, the regulatory environment, or an insufficient internet connection are viewed as the other main barriers to the diffusion of big data.

Despite these challenges associated with big data, a widely shared expectation is that the ongoing changes in how data is being generated and made relevant for firms can help to increase business value through profitable use of data, that previously had even been used to be produced as ‚waste' product of business activity before the surge of big data technologies. New data information practices and better informed decision-making can be particularly advantageous for firms' innovation processes, which often involve high uncertainty and risk. In this vein, mining of consumption patterns or social network and consumer sentiment analysis, for instance, might improve the adoption and market success of new products. Data obtained from sensors can facilitate the detection of product defects and the subsequent improvement of existing products. Insights obtained from big data can furthermore reduce the duration and costs of the innovation process. Besides improving the R&D process, big data can also be at the core of the innovation

itself. Monitoring transactions and combining different information facilitates the development of new personalized services (Varian, 2010) and other data-intensive innovations. This applies to highly digitized services as well as to more traditional manufacturing industries. For instance, by exploiting real time data on the geospatial position of users, mapping apps now provide drivers with real time information about potential road congestion (Kshetri, 2014). A successful example in traditional manufacturing can be found at the Ford Motor Company that started capturing consumer data from vehicles through sensors and remote app-management software. Based on analysis of data from the cars' voice recognition system the company found that surrounding noise affected the performance of the software. This led to an improvement of the system by means of noise reduction technology and the repositioning of microphones (Erevelles et al., 2016). Overall, big data is widely expected to enable firms from all industries to create new products and services, improve existing ones, and to develop new business models (e.g. Manyika et al., 2011; Gobble, 2013).

High potentials to foster innovation, productivity, and growth are also ascribed to big data by policymakers. For instance, the European Commission (EC) stressed the importance of data for growth and innovation in a knowledge-based economy in their policy report on the strategy for a digital single market. Furthermore, the EC has already taken measures to promote the data-driven economy, e.g. through public-private-partnerships for projects on big data or by supporting the development of standards and interoperability in data usage (European Commission, 2014).

Despite the high expectations associated with big data and the prominent position it has gained as a current key technological trend, there is a paucity of empirical evidence on its effect on firm performance overall, and firms' innovation performance in particular. Against this background, we analyze the relation of firms' use of big data and innovation performance using large-scale firm-survey data from German manufacturing and services industries. Extending classical knowledge production functions by firms' use of big data, we find that big data information practices are associated with a higher propensity to innovate, as well as a higher innovation intensity.

Our paper contributes to the literature in various respects: (i) we provide first large-scale empirical evidence based on representative firm-level data on the role of big data for firm performance in terms of the product innovation activities of manufacturing and service firms. (ii) The paper further contributes to a better understanding of the relationship between data analysis and innovation output across industries and helps to assess the potential benefits of big data analytics.

The remainder of the article is structured as follows. Section 2 reviews the empirical literature on the potential effects of big data analytics on firm performance. Section 3 lays out our empirical framework. Section 4 describes the data and measures. Section 5 and 6 discuss the descriptive and econometric results. Section 7 then analyzes which types of digital information and data-related practices are pursued by firms who indicate to make use of big data technologies. Finally, Section 8 concludes.

2 Related Empirical Literature

The reports of Manyika et al. (2011) and the OECD (2015) provide a general overview of the definition and application scope of big data analytics and the potential economic benefits of the use of big data technologies and of data-driven innovation.[1] Up to now, empirical evidence on the potential effects of big data analytics on firm performance has been scarce. There exist only a few empirical studies based on selective U.S. datasets for specific sectors or limited to listed companies (e.g. Brynjolfsson et al., 2011; Tambe, 2014; Brynjolfsson and McElheran, 2016*a*). The common finding from these studies is that firms with more intensive data usage are more productive. Furthermore, some studies show complementarities between big data usage and employment of highly qualified workers (e.g. Tambe, 2014; Brynjolfsson and McElheran, 2016*a*).

Concerning the diffusion process of data-related activities, Saunders and Tambe (2015) demonstrate an increasing trend toward the use of data-related activities in U.S. firms within the IT industry in the period from 1996 to 2012. Likewise, Brynjolfsson and McElheran (2016*a*) find that the use of data-driven decision-making almost tripled in the U.S. during the period from 2005 to 2010, where the adoption was particularly high in larger firms and in firms with more skilled workers and a higher IT capital stock.

With respect to the role of data-driven decision-making for productivity, Brynjolfsson et al. (2011) find that such practices are associated with a 5 to 6 percent increase in productivity and output among publicly traded U.S. firms. Similarly, Brynjolfsson and McElheran (2016*a*) show that data-related management practices caused a productivity increase of 3 percent for firms in the U.S. manufacturing sector. However, the authors highlight heterogeneity in the productivity returns of data-related practices with respect to firm characteristics, with the productivity return of data-related management practices appearing to be lower for larger, older and capital-intensive multi-unit firms. In addition, they find evidence for complementarity between data-driven decision-making and a high IT capital stock prior to the adoption of data-related practices as well as complementarity between data practices and the presence of more highly educated workers.

Tambe (2014) shows evidence for labor market complementarities between investments in and productivity returns from a particular big data technology, namely Hadoop, and the availability of employees with the skills for using this big data technology. The hypotheses for labor market complementarities between technology and human capital are supported by findings that indicate that U.S. firms' Hadoop investments yield higher productivity returns in geographic labor markets with high availability of workers with Hadoop skills. Wu and Hitt (2016) find evidence for complementarity between data analysis skills and process-related decisions, which is suggested by positive productivity returns for firms in which employees have a higher level of data skills and

[1] Goodridge and Haskel (2015) develop an economic framework to determine the importance of big data for GDP and for GDP growth. Applying their framework to the UK, they find that big data in the form of transformed data and data-based knowledge accounted for 0.02 percent of growth in market sector value added from 2005 to 2012.

the use of practices that aim at improving business processes is more intensive.

Overall, the findings on the role of big data analytics in firm performance are compatible with prior evidence on the complementarity and performance effects of information and communication technologies (ICT). There is a large literature on the productivity effects of ICT investment as well as on complementarities between ICT and human capital.[2] Generally, ICT is viewed as an enabler for innovation (e.g. Brynjolfsson and Saunders, 2010; Spiezia, 2011). In terms of the role of data use for realizing innovation, Bertschek and Kesler (2017) find that the adoption of a Facebook page and user activity on this page are significant determinants for the realization of product innovations by firms.

To the best of our knowledge, there is no study yet that explicitly examines the role of big data analytics for innovation performance at the firm level across industries. Based on the findings from the literature on the role of big data in firm performance and generally the contribution of ICT to innovation, we expect a positive relationship between big data analytics and product innovation - however, possibly not uniformly for all firms but rather contingent on potential complementary factors.

3 Empirical Framework

We analyze the contribution of big data to firms' innovation performance within the widely used knowledge production function framework introduced by Griliches (1979). This framework postulates a transformation process which links various inputs associated with knowledge accumulation, such as investments in R&D or human capital, to the firms' innovative output. Knowledge production functions have been the workhorse model in understanding the importance of various knowledge sources besides formal R&D. In the present work, we explicitly account for big data in the firms' knowledge production processes in order to provide initial insights into the relevance of big data for firms' innovation activities.

The following section outlines our empirical model of the knowledge production function. We denote y_{1i}^* the latent propensity of firm i to achieve product innovations, given the firm's use of big data analytics, $bigdata_i$, as well as the firm's R&D intensity and other firm- and market-specific characteristics denoted by the vector $\mathbf{c}_{1i}$. For simplicity of the formal exposition of the analysis, let us further collect the variable on the firm's big data use and further control variables in the vector $\mathbf{x}_1 \equiv (bigdata, \mathbf{c}_1)$. The first step of the empirical model of the knowledge production function assumes a linear additive relationship and amounts to

$$y_{1i}^* = \beta_1 bigdata_i + \gamma_1' \mathbf{c}_{1i} + \epsilon_{1i} = \delta_1' \mathbf{x}_{1i} + \epsilon_{1i} \tag{1}$$

where β denotes the parameter of interest, capturing the effect of the firm's engagement in big data analytics on the propensity to innovate. ϵ_{1i} denotes an idiosyncratic error term, which captures

[2]For an overview see e.g., Draca et al. (2007), Van Reenen et al. (2010), Cardona et al. (2013).

unobserved variables affecting y_{1i}^* and is assumed to be identically and independently normally distributed, $\epsilon_{1i} \sim NID(0, \sigma_1^2)$. The observed variable is the innovation success, i.e. the event of introducing a new product to the market, y_{1i}, which is defined by the following observation rule:

$$y_{1i} = \mathbf{1}[y_{1i}^* > 0] \tag{2}$$

where $\mathbf{1}[\,]$ is the indicator function taking the value 1 if the condition is satisfied and 0 otherwise. Equations (1) and (2) describe the first part of our analysis, in which we estimate the relationship between the use of big data and firms' innovation propensity via a simple Probit model.[3]

Beyond the relationship between big data and the propensity to innovate, we want to assess the relationship with the firms' innovation intensities. Thus, let y_{2i}^* denote the firms' potential innovation intensities given the firms' use of big data, R&D intensity and further firm- and market-specific characteristics, such that

$$y_{2i}^* = \beta_2 bigdata_i + \gamma_2' \mathbf{c}_{2i} + \epsilon_{2i} = \delta_2' \mathbf{x}_{2i} + \epsilon_{2i} \tag{3}$$

where, again, $\epsilon_{2i} \sim NID(0, \sigma_2^2)$ denotes the normally distributed idiosyncratic error term and $\mathbf{x}_2 \equiv (bigdata, \mathbf{c}_2)$. In line with much of the empirical literature investigating innovation intensities, the observed innovation intensity, which is typically measured by the sales ratio of innovative products and services, is assumed to be defined by the following observation rule:

$$y_{2i} = \mathbf{1}[y_{2i}^* > 0] y_{2i}^*. \tag{4}$$

Equations (3) and (4) together result in the standard Tobit model (Tobin, 1958), which takes account of the nonlinear nature of the conditional expectation function $E(y_{2i}|\mathbf{x}_{2i})$ due to the nontrivial fraction of firms which do not generate sales with newly introduced products.[4]

The conditional expectation for the model made up of Equations (3) and (4) is given by

$$E(y_{2i}|\mathbf{x}_{2i}) = \Phi(\delta_2' \mathbf{x}_{2i}/\sigma)\delta_2' \mathbf{x}_{2i} + \sigma\phi(\delta_2' \mathbf{x}_{2i}/\sigma) \tag{5}$$

where $\Phi_i(\cdot)$ and $\phi_i(\cdot)$ denote the standard normal cumulative distribution function and density function, respectively.[5]

A potential problem in estimating the Tobit model arises due to its strong and restrictive distributional assumptions. Unlike Ordinary Least Squares estimation, in cases of heteroskedasticity or non-normality, Tobit estimates will generally be inconsistent.[6] Due to the limitations of the

[3] Given the distributional assumption in Equation (1), we have $P(y_{1i} = 1|\mathbf{x}_{1i}) = P(y_{1i}^* > 0|\mathbf{x}_{1i}) = P(\epsilon_{1i} \leq \mathbf{x}_{1i}'\beta) = \Phi(\mathbf{x}_{1i}'\beta)$ under the normalization restriction $\sigma_1^2 = 1$, which we estimate by Maximum Likelihood.

[4] Note that, in line with the general literature, in the Tobit model with zero lower limit we ignore the upper limit of the innovation intensity. However, as the share of observations at the upper limit (of 1) is well below 1 percent, we regard the effect of upper limiting cases on the estimates to be negligible.

[5] For a more detailed description of Tobit type models see for instance Amemiya (1984) or Maddala (1986).

[6] Note that the assumption of normality and constant variance of ϵ_{2i} is crucial in deriving the conditional expectation

standard Tobit model, we check our results against the fractional logit model proposed by Papke and Wooldridge (1996). This model builds on the logistic distribution function to model the conditional expectation of a fractional dependent variable

$$E(y_{2i}|\mathbf{x}_{2i}) = \frac{exp(\delta_2'\mathbf{x}_{2i})}{1 + exp(\delta_2'\mathbf{x}_{2i})}. \tag{6}$$

Using a Bernoulli link function the model is estimated by Maximum Likelihood. Crucially for our application, the fractional logit model allows for y_{2i} to take on the boundaries 0 and 1 with positive probability, as opposed to other common solutions to model proportions, such as using the logit transformation of y_{2i}.

The standard Tobit and the fractional logit model discussed above assume that the observed innovation intensity is the result of a single process influenced by the same set of determinants. As the innovation intensity is a fractional variable with a lot of observations clustering at zero, one possible concern is that a single model fitted to all data might be insufficient. In particular, while big data might be related to the propensity to innovate, it could at the same time be unrelated to the innovation intensity, i.e. the market success of the firms' innovations, conditional on being an innovator. In that case, the simple Tobit model in Equations (3) and (4) is too restrictive. Alternatively, we can consider a framework in which the models for the propensity to innovate and for the innovation intensity conditional on being an innovator differ. Overall, there is no consensus in the empirical innovation literature whether a one stage model, such as the simple Tobit model described above, or an alternative two stage model is more appropriate to model firms' innovation intensities.[7] We therefore also estimate an alternative two stage model. In particular, we consider that, alternative to Equation (4), the observed innovation intensity is defined by the observation rule

$$y_{2i} = \mathbf{1}[y_{1i}^* > 0]y_{2i}^* \tag{7}$$

such that the sales ratio of innovations is observed if the firm's propensity to innovate is sufficiently large (e.g. Raymond et al., 2015). In addition, let the unobserved errors $(\epsilon_{1i}, \epsilon_{2i})$ be jointly normally distributed with covariance σ_{12}. Equations (3) and (7) together with the distributional assumptions on the error terms yield the Tobit Type II or Heckman Selection model, in which the conditional expectations of interest are given by:

$$E(y_{1i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Phi(\delta_1'\mathbf{x}_{1i}) \tag{8}$$

$$E(y_{2i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1) = \delta_2'\mathbf{x}_{2i} + \sigma_{12}\frac{\phi(\delta_1'\mathbf{x}_{1i})}{\Phi(\delta_1'\mathbf{x}_{1i})} \tag{9}$$

Given both models, the simple Tobit as well as the Heckman Selection model, are being used

in Equation (5).

[7] See for instance Cassiman and Veugelers (2006), Andries and Czarnitzki (2014) or Hottenrott and Lopes-Bento (2016) for other studies applying both types of models to model innovation shares.

in the empirical innovation literature, we estimate both to check the robustness of our findings to the common modeling assumptions.

The main caveat here is that our study is subject to common endogeneity concerns in the empirical literature on the value of ICT. Omitted variables might confound the relation between the use of big data and firms' innovation performance. The main advantage of our data is the wide variety of background characteristics we can account for. In particular, our data contain rich information on firms' use of alternative digital technologies, which help to disentangle the quality and features of big data analytics activities from the firms' general ICT intensities as well as the use of legacy systems. Since the empirical literature on ICT performance generally suffers from a lack of good instrumental variables, reverse causation is another common endogeneity concern. We note that our study runs the risk of being confounded by reverse causation since we are only able to provide controlled correlation applying a new cross-sectional dataset. Nevertheless, we believe that our analysis is an important first step in understanding how firms make use of big data analytics and in shedding light on the often discussed role of big data technologies in the innovation process of firms.

4 Data and Measures

Our analysis is based on the ZEW ICT survey which is a survey of manufacturing and services firms located in Germany with five or more employees.[8] In total, six waves were collected in 2000, 2002, 2004, 2007, 2010 and 2015. We exploit the 2015 wave, which is the first to contain information on the firms' use of big data. About 4400 firms were interviewed about their characteristics and particularly about their ICT usage. The data were collected via computer-aided telephone interviews (CATI) based on a sample stratified with respect to industry and firm size. The respondent is usually from the board of management or the head of the IT department.[9]

4.1 Big Data Analytics

Our main variable of interest is the dummy variable for big data analytics that is equal to one if the firm is using big data technologies. More precisely, the following question was asked in our survey:

> *"Up next a question about so-called big data, i.e. the processing of large amounts of data. Does your company systematically analyze large amounts of data to support business operations?"*

As we aim at measuring firms' engagement with big data across different industries and firm sizes, our measure of big data use leaves room for the subjective assessment of the interviewee. This

[8]The data are available at the ZEW Research Data Centre - http://kooperationen.zew.de/en/zew-fdz.
[9]For more information about the survey see Bertschek et al. (2017).

was done deliberately, because despite the public recognition of big data as one of the current key technological trends, the term lacks a generic definition and does not constitute a unified concept. Furthermore, the technology for big data has been advancing quickly over time. As the size of datasets is continuously growing and increasingly sophisticated tools arise to analyze them, big data has always been an evolving concept. The most commonly accepted definition is based on the "3 Vs" formulated by Laney (2001). They are the enormous amount of data (volume), (2) the variety of data coming from highly diverse sources (variety), and (3) the pace of data processing (velocity). Focusing on the novelty of big data technologies and architectures Apache Hadoop defines big data as data that *"could not be captured, managed, and processed by general computers within an acceptable scope"* (Chen et al., 2014, p. 173). An insightful delineation of big data can be found in Chen et al. (2012). The authors describe big data as digitized information and analytical technologies which have not been incorporated into standard commercial business intelligence platforms and enterprise software systems. In this vein, the authors highlight new web-based, mobile and sensor-generated data as well as techniques such as opinion mining, social network analysis or machine learning techniques.[10]

The definition of big data might also be contingent on the industrial context and depend on the specific software used and the common size of datasets in a particular industry (Manyika et al., 2011). Product innovations based on big data analytics will also vary a lot between industries. For instance, Luckow et al. (2015) describe potential innovations in the automotive industry. Based on the steadily growing number of sensors per vehicle, new innovative services like traffic prediction, safety warnings, vehicle diagnostics, and location-based services are based on big data analytics. High potentials are also ascribed to big data technologies in health care, where big data can help to identify drug interactions and design improved drug therapies (Kshetri, 2014). Another often cited example is that of insurance companies making use of different data sources and big data technologies to design improved premium policies and new forms of contracts (Varian, 2010).

4.2 Innovation Outcomes

Our data include items on innovation and R&D activities following the Community Innovation Survey (CIS) and the guidelines of the Oslo Manual by the OECD and Eurostat (Mortensen et al., 2005). In particular, we consider the event of introducing a product innovation to the market as the first outcome of the knowledge production process. The relevant measure is a binary indicator, which takes the value one if the firm has introduced a new or substantially improved product or service to the market over the past three years (*Product Innovation*). The product can be new to the market overall or new to the firm. In addition to the propensity to innovate, we investigate the intensity of innovation, which we measure by the share in total sales resulting from new products in the year 2013 (*% of Sales New Product*). In contrast to a mere innovation count, the sales share of innovations weights each innovation by its success in total turnover. In this way, our

[10]For an extensive review of definitions of the big data phenomenon see for instance Wamba et al. (2015).

innovation intensity measure captures the market success of product innovations (Mairesse and Mohnen, 2002; Laursen and Salter, 2006).

4.3 Control Variables

Following the empirical innovation literature, we control for an extensive set of firm characteristics which have been shown to affect innovation performance. We measure R&D intensity, the potentially single most important input factor to knowledge production, as R&D expenditures over total sales (*% of R&D Expenses*). The firms' R&D intensities affect the propensity to innovate as well as the firms' innovation successes (Pakes and Griliches, 1980) and reflect the relative importance of innovation activities for the firm. Firms which are making use of big data analytics are in general likely to be more intensive ICT users. Similarly, ICT intensity can be expected to positively affect firms' innovation performance (Hempell and Zwick, 2008). Therefore, we control for firms' ICT intensities by the share of employees who mainly work with personal computers (*% of Emp. Predom. Using PC*) as well as the share of employees with access to the internet in the workplace (*% of Emp. Using Internet*). Furthermore, as the use of enterprise software systems has been shown to be related to firms' innovation activities (Engelstätter, 2012), we include a binary variable into the model indicating whether or not the firm has an enterprise software system implemented (*Enterprise Software*). We note that our additional measures on the firms' ICT use capture the effect of mature software systems and data technologies, which lack the quality of large-scale data analytics, such as structured data collected through standard Enterprise Resource Planning Systems and stored in conventional relational database management systems. Furthermore, firms' innovative capabilities are affected by the employees' human capital, their knowledge, abilities and creativity (Vinding, 2006). Thus, we control for the share of highly skilled employees, i.e. workers with degrees from universities and technical colleges (*% Highly Qualified Employees*), as well as the share of employees with vocational training (*% Medium Qualified Employees*). In order to account for the firm's investment in IT-specific knowledge, we control for the share of employees who participated in IT-specific training over the past year (*% of Emp. IT-Training*). We furthermore account for the age structure of the workforce by controlling for the share of employees below 30 years of age (*% of Employees < Age 30*) and above 50 years of age (*% of Employees > Age 50*). As the maturity of the firm might affect both, the use of cutting-edge technology as well as their innovative capabilities (Huergo and Jaumandreu, 2004), we control for the years since the founding year of the firm (*Age*). Younger firms might also achieve higher sales shares with new products merely because they have fewer established products in their portfolio. Firm size has been found to be an important determinant of technology adoption (Haller and Siedschlag, 2011). Likewise, potential relations between firm size and innovation have already been found by Schumpeter (1942). Overall, larger firms can be expected to have better internal financial resources and enjoy economies of scale and scope, which benefits both, technology adoption as well as innovative capabilities. We thus control for firm size measured by the log of

the number of employees (*Employees*). As the likelihood of innovating has been shown by some studies to increase with physical capital intensity (e.g. Lööf and Heshmati, 2006), we control for the log of gross investments (*Investment*). The exposure to international product markets affects the potential market size for new products as well as the competitive pressure to innovate (Hottenrott and Lopes-Bento, 2016). We thus include an indicator for whether the firm exports to foreign markets (*Exporter*) and whether it is part of a multinational enterprise (*Multinational*). As Brynjolfsson and McElheran (2016*b*) show that multi-unit firms are more likely to adopt data driven decision-making, we additionally account for the firms' ownership structure by a binary variable indicating whether the firm is part of a national enterprise group (*Group*). Finally, we account for structural regional differences between the two former German states by a binary indicator for firms' location in former Eastern Germany (*East Germany*) as well as structural differences between industries by including a set of 16 industry dummies constructed from 3-digit NACE industry codes.[11]

5 Descriptive Statistics

Table 5.1 provides summary statistics on the variables used in the analysis. The share of firms that introduced new products or services amounts to 48 percent and the average share of sales due to new products and services is 8.4 percent. In our estimation sample, 22 percent of firms rely on big data to support their decision-making. With a share of 56 percent, considerably more firms have implemented an enterprise software system. About 45 percent of employees predominately work with computers. The average number of employees in the sample is 89, so the sample mainly consists of small and medium-sized enterprises. We apply the data to shed light on the incidence of data driven decision-making and to discover which firms exploit data strategically for their decision-making. Figure A.1 provides the in-sample share of firms which are using big data analytics by industry. Overall, the use of data analytics is higher in the service sector. As noted by other authors as well (e.g. Chen et al., 2014), data driven decision-making has proliferated in the financial sector, where over half of the firms in the sample indicated that they systematically apply data as a form of strategic support for their business operations. Firms in the retail and wholesale trade sectors also make intensive use of data in their decision-making process with a diffusion of around 30 percent. Amongst the manufacturing industries, big data is used most intensively in the chemicals and motor vehicles sectors, by around 23 percent of the firms in each sector. The sector in which the least firms rely on data for their decision-making is manufacturing of consumer goods with a diffusion rate of only 13 percent. Figure A.1 also depicts the share of firms innovating by industry. Among manufacturers of chemicals, electronics and machinery as well as in the ICT service sector over 70 percent of firms introduced new products or services within the previous three years. The share of innovating firms is lowest in the transport service sector with

[11] Table A.1 provides an overview of the industries and their distribution in the estimation sample.

only 23 percent. Overall, the variation over industries depicted in Figure A.1 does not provide a clear picture on the relation between the use of big data and innovation performance. While some sectors with a high diffusion of big data also exhibit high shares of innovating firms, this is certainly not true for all industries. For example, while in the manufacturing of machinery industry around 71 percent of the firms innovate, only 16 percent rely on big data for their decision-making.

Table 5.1: Summary Statistics: Estimation Sample

	N	Mean	SD	Min	Max
Product Innovation	2706	0.48	0.50	0	1
% of Sales New Product	2706	0.084	0.15	0	1
Big Data	2706	0.22	0.41	0	1
% of Emp. Predom. Using PC	2706	0.45	0.34	0	1
% of Emp. Using Internet	2706	0.57	0.37	0	1
Enterprise Software	2706	0.56	0.50	0	1
% of R&D Expenses	2706	0.050	0.11	0	1
Employees	2706	89.3	243.5	5	4100
Employees (in logs)	2706	3.43	1.30	1.61	8.32
Investment in Mill. Euro	2706	0.88	4.61	0.00050	130
Investment (in logs)	2706	-2.03	1.83	-7.60	4.87
Exporter	2706	0.45	0.50	0	1
% Highly Qualified Employees	2706	0.19	0.24	0	1
% Medium Qualified Employees	2706	0.63	0.27	0	1
% of Employees < Age 30	2706	0.24	0.17	0	1
% of Employees > Age 50	2706	0.27	0.19	0	1
East Germany	2706	0.25	0.43	0	1
% of Emp. IT-Training	2706	0.092	0.19	0	1
Age (in logs)	2706	3.17	0.92	0	6.39
Group	2706	0.29	0.46	0	1
Multinational	2706	0.093	0.29	0	1

SOURCE: ZEW ICT-Survey 2015.

To further investigate which firms exploit data strategically for their decision-making, Table A.2 provides summary statistics of firm characteristics conditional on the firms' use of big data. In general, firms which have introduced big data technologies are using ICT more intensively overall, are larger in terms of employees and investments, have higher R&D expenditures, more likely to belong to a multi plant or multinational firm and are more likely to export their goods and services. Importantly, firms using big data analytics are on average more innovative, both at the extensive and intensive margin. Still, a thorough investigation of the relation betweeen big data and firms' innovation performance calls for a multivariate analysis as outlined above.

6 Econometric Results

The following section provides the main estimation results. Table 6.1 presents the estimation results of the Probit models analyzing the relation between big data utilization and the firms' innovation propensity for the full sample as well as for the estimation sample split into the manufacturing and service sector, respectively. The estimate of the coefficient on the big data indicator is positive and statistically significant in all three estimations. Moreover, the estimated relation between a firm's use of big data and the likelihood of that same firm introducing a new product or service to the market is economically meaningful. Looking at the results for the full sample in column (1), the firms' application of big data analytics is associated with a 6.7 percentage point increase in the propensity to innovate. Interestingly, the results are of comparable magnitude when differentiating between manufacturing and service firms in columns (2) and (3). The respective results show that firms using big data analytics are 6.5 percentage points more likely to innovate in the manufacturing sector and 6.8 percentage points more likely to innovate in the service sector. Looking at the estimated coefficients on other control variables, in particular those for other measures of ICT use by the firm, we find that the firms' general ICT intensity measured by the share of employees working predominantly with PCs is not significantly related to innovation propensity. Our estimation results furthermore confirm existing research on the positive relation between enterprise software and innovation (e.g. Engelstätter, 2012). ERP Systems typically serve for the planning and controlling of business processes across different sections of the value chain. They moreover constitute a platform to integrate more specific applications, such as Supply Chain Management or Customer Relationship Management Software. While firms using ERP Systems are typically integrating information across different business processes and engage in data driven decision-making, the features of classical ERP Software systems lack the quality of big data analytics in terms of the amount of data that is being processed and the software tools which are used to analyze the data. Furthermore, ERP systems are used to process data that has been purposefully generated by the firm through business transactions while big data often stems from heterogenous sources outside of the firm. Importantly, our measure for big data use explains the firms' innovation propensity beyond the effect of these legacy software systems. Further strong predictors for how likely a firm is to innovate over all three models are the firm's R&D intensity and export status as well as whether or not the firm belongs to a multinational enterprise.

Table 6.1: Dependent Variable: Dummy for Product Innovation - Probit Regression - Average Marginal Effects

	(1) Full Sample	(2) Manufacturing	(3) Services
Big Data	0.067***	0.065*	0.068**
	(0.023)	(0.035)	(0.029)
% of Emp. Predom. Using PC	-0.012	-0.104	0.050
	(0.042)	(0.075)	(0.051)
% of Emp. Using Internet	0.074**	0.076	0.067
	(0.036)	(0.050)	(0.052)
Enterprise Software	0.081***	0.112***	0.059**
	(0.020)	(0.030)	(0.026)
% of R&D Expenses	0.905***	1.104***	0.774***
	(0.158)	(0.267)	(0.176)
Employees (in logs)	0.011	0.015	0.010
	(0.012)	(0.017)	(0.015)
Investment (in logs)	0.024***	0.018*	0.029***
	(0.007)	(0.011)	(0.010)
Exporter	0.164***	0.144***	0.183***
	(0.021)	(0.029)	(0.032)
% Highly Qualified Employees	0.154**	0.353***	0.041
	(0.062)	(0.124)	(0.080)
% Medium Qualified Employees	-0.043	-0.020	-0.099
	(0.043)	(0.056)	(0.069)
% of Employees < Age 30	-0.028	-0.069	-0.006
	(0.052)	(0.077)	(0.071)
% of Employees > Age 50	-0.015	-0.054	0.022
	(0.049)	(0.070)	(0.070)
East Germany	0.005	0.032	-0.036
	(0.021)	(0.028)	(0.030)
% of Emp. IT-Training	0.137***	0.165	0.126**
	(0.052)	(0.125)	(0.055)
Age (in logs)	-0.010	0.004	-0.025*
	(0.010)	(0.014)	(0.014)
Group	0.036*	0.056*	0.014
	(0.020)	(0.030)	(0.028)
Multinational	0.135***	0.133***	0.123**
	(0.036)	(0.047)	(0.055)
Industry Dummies	Yes	Yes	Yes
Pseudo R^2	0.209	0.182	0.216
Observations	2706	1404	1302
Log likelihood	-1481.155	-788.318	-683.426

Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept.
Source: ZEW ICT-Survey 2015.

Table 6.2 reports the results from the Tobit and the Fractional Logit estimations modelling the sales share of new products, i.e. the market success of firms' innovations. The table reports average marginal effects on the conditional expectations in Equations (5) and (6). Overall, results show that the use of big data is not only related to firms' innovation status, but also to their innovation intensity. Over both empirical models in all three samples, big data is positively and statistically significantly associated with the sales share of innovations. Again the estimates are economically meaningful and of equal magnitude for the full sample and within the manufacturing and the service sector. In particular, for the full sample (columns (1) and (2)) the use of big data is associated with a 2.5 to 2.9 percentage point increase in the sales share from innovations. All other coefficients are in line with prior expectations. R&D intensity is a strong predictor of the sales share of innovations. Over most specifications, a firms' age is negatively associated with innovation intensity. Thus, younger firms achieve a larger share of their sales with newly introduced products or services.

Table 6.2: Dependent Variable: % Share of New Products in Turnover- Tobit/FracReg Regressions

	Full Sample		Manufacturing		Services	
	(1)	(2)	(3)	(4)	(5)	(6)
	Tobit	FracReg	Tobit	FracReg	Tobit	FracReg
Big Data	0.025***	0.029***	0.026***	0.032***	0.026***	0.029***
	(0.006)	(0.008)	(0.009)	(0.011)	(0.008)	(0.010)
% of Emp. Predom. Using PC	0.004	0.007	-0.009	-0.003	0.015	0.019
	(0.012)	(0.013)	(0.019)	(0.022)	(0.014)	(0.015)
% of Emp. Using Internet	0.016*	0.014	0.021	0.022	0.014	0.007
	(0.010)	(0.011)	(0.013)	(0.015)	(0.014)	(0.018)
Enterprise Software	0.020***	0.018***	0.032***	0.030***	0.012*	0.011
	(0.005)	(0.006)	(0.007)	(0.008)	(0.007)	(0.008)
% of R&D Expenses	0.253***	0.196***	0.321***	0.243***	0.199***	0.158***
	(0.020)	(0.024)	(0.035)	(0.050)	(0.023)	(0.024)
Employees (in logs)	-0.007**	-0.014***	-0.004	-0.009*	-0.009**	-0.019***
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
Investment (in logs)	0.007***	0.008***	0.003	0.002	0.011***	0.014***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)
Exporter	0.037***	0.030***	0.034***	0.028***	0.038***	0.029***
	(0.005)	(0.007)	(0.007)	(0.009)	(0.008)	(0.010)
% Highly Qualified Employees	0.034**	0.026	0.055**	0.027	0.013	0.026
	(0.016)	(0.020)	(0.028)	(0.032)	(0.022)	(0.023)
% Medium Qualified Employees	-0.015	-0.018	-0.019	-0.032	-0.020	-0.001
	(0.012)	(0.015)	(0.015)	(0.020)	(0.019)	(0.021)
% of Employees < Age 30	0.001	0.013	0.015	0.036	-0.011	-0.003
	(0.014)	(0.017)	(0.021)	(0.023)	(0.018)	(0.023)
% of Employees > Age 50	-0.001	0.001	-0.008	-0.000	0.007	0.005
	(0.013)	(0.015)	(0.019)	(0.023)	(0.018)	(0.020)
East Germany	0.002	0.002	0.012	0.013	-0.009	-0.009
	(0.006)	(0.006)	(0.008)	(0.009)	(0.008)	(0.008)
% of Emp. IT-Training	0.027**	0.019	0.007	-0.002	0.031**	0.027*
	(0.012)	(0.015)	(0.025)	(0.023)	(0.014)	(0.016)
Age (in logs)	-0.008***	-0.012***	-0.004	-0.008*	-0.011***	-0.015***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)
Group	0.008	0.007	0.015*	0.016	0.001	-0.002
	(0.006)	(0.007)	(0.008)	(0.010)	(0.007)	(0.009)
Multinational	0.024***	0.023**	0.015	0.011	0.035**	0.039**
	(0.009)	(0.009)	(0.011)	(0.011)	(0.014)	(0.016)
Industry Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.363	0.092	0.398	0.069	0.335	0.127
Observations	2706	2706	1404	1404	1302	1302
Censored	1432		633		799	
Uncensored	1274		771		503	
Log likelihood	-659.328	-709.616	-257.934	-410.628	-376.715	-295.430

Standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors in columns 2, 4, and 6. All models include an intercept.

Source: ZEW ICT-Survey 2015.

Finally, we turn to the estimation results of the Heckman Selection Model. Theoretically, the model is identified by the functional form assumptions. That is, even if the set of regressors in both equations of the model is identical ($\mathbf{x}_1 = \mathbf{x}_2$), the model is identified due to the nonlinearity of the inverse Mills ratio in the second equation.[12] However, in practice it is desirable to have at least one exclusion restriction, i.e. a variable that enters the selection equation but not the second equation, for more reliable identification of the model parameters (e.g. Wooldridge, 2010, p.805ff). Ideally, the exclusion restriction is selected on theoretical grounds. However, there is no variable available which would theoretically affect the firms' likelihood of innovating while leaving the firms' innovation intensity unaffected. We thus follow, for instance, Andries and Czarnitzki (2014) or Peters and Schmiele (2010) and search empirically for an exclusion restriction in order to ensure that identification of the model parameters does not merely rest on functional form assumptions. When including the full set of variables in both equations of the model, the firms' export status is strongly and significantly related to the firms' propensity to innovate, whereas the respective parameter estimate in the second equation is very small and statistically insignificant (see Table A.3 in the appendix for the respective estimation results). We thus rely on the firms' export status as an exclusion restriction.[13] We note, however, that the validity of our exclusion restriction cannot be tested.

Table 6.3 reports the average marginal effects of the Heckman model estimation. For each of the three samples, the first column reports the partial effects on the propensity to innovate while the second column reports the expected innovation intensity, conditional on being an innovator, according to Equation (9). Overall, the previous results are confirmed by the estimation of the selection model. The application of big data analytics is associated with a 6.5 to 6.7 percentage point higher innovation propensity over all samples. The estimated partial effect on the innovation intensity conditional on being an innovator ranges between 2.3 percentage points in the full sample and 2.5 percentage points in the manufacturing and service sector samples. Note that, in contrast, the use of enterprise software is only positively and statistically significantly related to the propensity to innovate, while the estimated partial effect on the conditional innovation intensity is negative, small and statistically insignificant.

Finally, it should be noted that over all three samples we cannot reject independence between the two equations comprising the model. Consequently, we can re-estimate the equation modeling the firms' innovation intensity on the subsample of innovating companies only. In fact, all the above results were confirmed and detailed regression results are thus omitted for the sake of brevity.

[12] The inverse Mills ratio corresponds to the term $\frac{\phi(\delta_1' \mathbf{x}_{1i})}{\Phi(\delta_1' \mathbf{x}_{1i})}$ in Equation (9).

[13] As a robustness check, we used enterprise software as well as firms' export status together with enterprise software as exclusion restrictions with substantially similar results (results not reported; available upon request).

Table 6.3: Heckman Selection Model with Exclusion Restriction (Export Status), Marginal Effects

	Full Sample		Manufacturing		Services	
	(1)	(2)	(3)	(4)	(5)	(6)
	1st	2nd	1st	2nd	1st	2nd
Big Data	0.066***	0.023***	0.065*	0.025**	0.067**	0.025**
	(0.022)	(0.008)	(0.035)	(0.010)	(0.030)	(0.013)
% of Emp. Predom. Using PC	-0.014	0.001	-0.108	0.015	0.048	-0.011
	(0.043)	(0.017)	(0.072)	(0.022)	(0.054)	(0.027)
% of Emp. Using Internet	0.074**	-0.005	0.075	0.010	0.069	-0.024
	(0.036)	(0.014)	(0.050)	(0.016)	(0.053)	(0.028)
Enterprise Software	0.082***	-0.007	0.112***	-0.002	0.059**	-0.012
	(0.020)	(0.008)	(0.030)	(0.010)	(0.026)	(0.013)
% of R&D Expenses	0.940***	0.223***	1.210***	0.285***	0.789***	0.182***
	(0.112)	(0.025)	(0.203)	(0.039)	(0.128)	(0.036)
Employees (in logs)	0.011	-0.020***	0.015	-0.011*	0.010	-0.031***
	(0.012)	(0.004)	(0.017)	(0.006)	(0.015)	(0.007)
Investment (in logs)	0.024***	0.004	0.018	-0.005	0.029***	0.014***
	(0.008)	(0.003)	(0.011)	(0.004)	(0.010)	(0.005)
Exporter	0.162***		0.138***		0.181***	
	(0.021)		(0.029)		(0.031)	
% Highly Qualified Employees	0.155**	0.003	0.358***	-0.037	0.040	0.043
	(0.062)	(0.023)	(0.114)	(0.034)	(0.081)	(0.040)
% Medium Qualified Employees	-0.040	-0.010	-0.018	-0.026	-0.099	0.032
	(0.043)	(0.017)	(0.057)	(0.019)	(0.069)	(0.036)
% of Employees < Age 30	-0.027	0.028	-0.069	0.069***	-0.004	-0.007
	(0.051)	(0.020)	(0.076)	(0.026)	(0.069)	(0.032)
% of Employees > Age 50	-0.015	0.007	-0.050	0.005	0.018	0.018
	(0.048)	(0.019)	(0.068)	(0.023)	(0.069)	(0.034)
East Germany	0.005	0.001	0.032	0.005	-0.036	-0.001
	(0.021)	(0.008)	(0.028)	(0.009)	(0.030)	(0.015)
% of Emp. IT-Training	0.133***	0.005	0.166	-0.016	0.119**	0.018
	(0.051)	(0.017)	(0.113)	(0.028)	(0.056)	(0.022)
Age (in logs)	-0.010	-0.013***	0.004	-0.009*	-0.025*	-0.019***
	(0.010)	(0.004)	(0.013)	(0.005)	(0.014)	(0.007)
Group	0.034*	-0.000	0.054*	0.007	0.014	-0.009
	(0.021)	(0.007)	(0.031)	(0.010)	(0.028)	(0.012)
Multinational	0.134***	0.012	0.130***	-0.004	0.123**	0.033*
	(0.036)	(0.010)	(0.047)	(0.012)	(0.055)	(0.019)
Industry Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2706	2706	1404	1404	1302	1302
$\hat{\sigma}_{12}$	-0.232		-0.283		-0.250	
LR-Test $H_0 : \sigma_{12} = 0$ $[\chi^2(1)]$, p-Val	0.166		0.113		0.391	
Log Likelihood	-969.745		-412.029		-524.323	

Standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept.
Source: ZEW ICT-Survey 2015.

As outlined in Section 2, existing empirical evidence has thus far highlighted the notion that the returns to employing big data analytics is contingent on human capital and the skills of the workforce (e.g. Brynjolfsson and McElheran, 2016*a*). In particular, Tambe (2014) provides empirical evidence that positive returns to Hadoop investments depend on the firm operating in labor markets with a sufficient supply of relevant technical skills.

Exploring these previous findings in the context of innovation, we conduct a further split sample analysis differentiating between firms with low vs. high general human capital and firms with low vs. high investment in the IT skills of their employees. Specifically, we define a firm as a low (high) human capital firm if the share of employees with degrees from universities and technical colleges is below (above) the industry specific median. Similarly, firms are defined as having low (high) investment in IT-specific skills if the share of employees who participated in IT-specific training in the previous year is below (above) the industry specific median.

Table 6.4 shows the regression results for Probit models analyzing the relation between big data utilization and the firms' innovation propensity. Columns 1 and 2 show the result for firms with low and high general human capital and columns 3 and 4 the respective results for firms with low and high investment in IT-specific skills. Interestingly, while the relation of big data analytics and the propensity to innovate is not contingent on general human capital in our data, it appears to be, in fact, contingent on the firm's investment in specific IT skills. For firms with low investment in IT skills, the parameter estimate reduces in magnitude and we cannot reject the null hypothesis of no association between big data analytics and the propensity to innovate. For firms with high investment in IT-specific skills the point estimate is now more than twice as large as for firms with low investment. We note that this finding does not carry over to the intensity of innovation, where results are similar to our previous findings, irrespective of the modeling assumptions.[14] Overall, the estimation results support findings on the importance of the acquisition of technical skills for the successful use of big data analytics and show them to be of particular relevance in the context of firms' innovative performance.

[14]Estimation tables for the innovation intensity equations using split samples by human capital and investment in IT-specific skills are excluded for brevity.

Table 6.4: Dependent Variable: Dummy for Product Innovation - Probit Regression by Skill Group - Average Marginal Effects

	General Human Capital		IT-specific skills	
	(1)	(2)	(3)	(4)
	low	high	low	high
Big Data	0.075**	0.066**	0.044	0.096***
	(0.033)	(0.030)	(0.034)	(0.029)
% of Emp. Predom. Using PC	-0.128**	0.089	-0.006	-0.033
	(0.064)	(0.058)	(0.059)	(0.061)
% of Emp. Using Internet	0.092*	0.010	0.110**	0.041
	(0.049)	(0.050)	(0.046)	(0.056)
Enterprise Software	0.112***	0.044	0.099***	0.046
	(0.028)	(0.028)	(0.026)	(0.030)
% of R&D Expenses	0.684***	0.995***	0.805***	0.960***
	(0.204)	(0.220)	(0.190)	(0.270)
Employees (in logs)	0.006	0.011	0.017	-0.009
	(0.017)	(0.016)	(0.016)	(0.017)
Investment (in logs)	0.031***	0.021**	0.017*	0.035***
	(0.010)	(0.011)	(0.009)	(0.012)
Exporter	0.167***	0.160***	0.151***	0.181***
	(0.030)	(0.030)	(0.028)	(0.032)
% Highly Qualified Employees	0.334*	0.167	0.175**	0.080
	(0.176)	(0.106)	(0.083)	(0.095)
% Medium Qualified Employees	-0.015	-0.118*	0.000	-0.147**
	(0.055)	(0.072)	(0.055)	(0.074)
% of Employees < Age 30	-0.078	0.017	-0.133**	0.119
	(0.076)	(0.072)	(0.067)	(0.082)
% of Employees > Age 50	0.033	-0.067	-0.055	0.007
	(0.067)	(0.072)	(0.062)	(0.081)
East Germany	0.015	-0.005	0.015	-0.017
	(0.031)	(0.028)	(0.027)	(0.031)
% of Emp. IT-Training	0.219***	0.101	0.192	0.084
	(0.083)	(0.064)	(0.589)	(0.065)
Age (in logs)	-0.014	-0.004	-0.018	-0.003
	(0.014)	(0.014)	(0.013)	(0.015)
Group	0.047	0.024	0.035	0.031
	(0.029)	(0.029)	(0.029)	(0.029)
Multinational	0.089	0.158***	0.092*	0.163***
	(0.057)	(0.043)	(0.056)	(0.043)
Industry Dummies	Yes	Yes	Yes	Yes
Pseudo R^2	0.192	0.239	0.186	0.215
Observations	1394	1312	1491	1215
Log likelihood	-765.502	-688.597	-813.189	-647.612

Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept.
Source: ZEW ICT-Survey 2015.

7 Big Data in the Context of Firms' Digital Transformation

As outlined above, big data is not a unified concept and its exact definition along with associated technologies and methodologies might differ between industries. Our main finding above suggests that knowledge reaped from digitized data by means of big data analytics can be a relevant intangible asset in the innovation process. We now want to further investigate how to best describe the availability and use of data within firms which claim to rely on big data technologies to support business operations. In particular, we investigate which data-related firm practices and digital technologies are used by firms who indicate to make use of big data. This enables a more precise interpretation of the empirical measure of big data use in our analysis and furthermore sheds light on the sources of digital information firms are currently exploiting to reap benefits of big data technologies. Hence, in a subsequent survey conducted in 2015/16, we obtained more detailed information on data-related firm practices for a subsample of our data consisting of around 2,000 observations.

We identify several data-related firm practices and digital technologies which the literature mainly discusses as driver of big data analytics. Generally, big data is generated either in an unstructured manner from heterogenous sources including internet clicks, mobile transactions and user-generated content, or from purposefully generated content through business transactions or sensors (George et al., 2014).

The former is largely generated as online content and many authors stress the relevance of social media as a main driver of big data (e.g. Wamba et al., 2015). Social media sites, user-generated content and the diffusion of mobile devices among individuals have become sources of data related to consumer behavior. The appropriate exploitation of such data can enable firms to better predict consumer needs and sentiments, which might be particularly advantageous for the innovation process (e.g. Erevelles et al., 2016). Previous empirical research has already shown that social media adoption and use are complementary to data skills within the firm (Hitt et al., 2016). We thus investigate whether the use of big data is associated with activities related to social media and online content. In particular, we can account for whether the firm offers customers the opportunity to evaluate products or services online (*s_feedback*), whether the firm systematically searches for user-generated content about its own products, services or the company (*s_content*) and whether the firm engages in online advertising (*s_ads*).

Data purposefully generated by the firm are created during the production process, or by digital technologies embedded in the final products and services. In this vein, data-driven activities in firms are captured for instance by production data, inventory data, or financial data. Many scholars particularly stress the relevance of data coming from sensors in the production of tangibles (George et al., 2014; Kitchin, 2017). By exploiting such digital information, big data is deemed valuable in the way it enables the optimization of daily operations in the production process and

value chain. Big data technologies can enable a highly variable production process and, in the form of predictive maintenance, reduce errors and down times. Furthermore, deploying big data analytics in logistic processes is discussed as a means to reduce inefficiencies, such as delayed shipments or inconsistent supplies (Wang et al., 2016; Zhong et al., 2015). To capture firms' data-driven activities in the production process of goods and services firms were asked whether they use automated data recording, processing and transmission to make internal processes more efficient (*p_efficiency*), provide employees with digital assistance systems (*p_assistance*), exchange information with suppliers and customers (*p_edi*), customize products or services (*p_customize*), or adapt internal processes flexibly (*p_adapt*). We furthermore measure whether the firm has been integrating IT systems between different business processes or divisions during the previous two years (*p_network*).

Finally, we account for digital technologies embedded in final products and services, which can be a source of digitized information used in the innovation process. In particular, we measure whether the firm produces products with embedded sensors (*g_sensor*), whether the firm offers mobile applications (*g_apps*), whether the firm offers product-support services (*g_service*), or whether partner firms offer product-support services (*g_partner_service*).

To uncover which of the data-related practices are associated with the use of big data analytics we look at controlled correlations between the indicator of big data use and indicators for the above digital technologies and firm strategies in the context of social media and online content, digital data in the production of goods and services and digital technology embedded in goods and services. Descriptive statistics of the new measures are provided in Tables A.5-A.7 and the controlled correlations of interest are provided in Tables A.8-A.10, respectively. In all correlation tables we control for the share of employees working predominantly with PCs and the use of ERP software as measures for general ICT-intensity, employees (in logs), industry and regional indicators. Furthermore, we control for the share of employees who have participated in IT-training. As Brynjolfsson and McElheran (2016*b*) show that multi-unit firms and firms with high human capital are more likely to adopt data-driven decision-making, we additionally control for whether the firm is part of a national enterprise group, as well as for the share of high and medium-qualified employees.

Looking at the controlled correlations in Table A.8 we see that a firm's use of big data is associated with a higher likelihood of the firm exploiting social media and online content in both the manufacturing and the service sector. However, the results indicate differences in the relevant types of data-related activities. Manufacturing firms which rely on big data have a statistically significant higher likelihood of allowing customers to evaluate products and services online and systematically searching for user-generated content about their own products or services or about the company. Interestingly, this is not the case in the service sector, where firms using big data analytics instead have a higher likelihood of engaging in online advertising.

Further differences between manufacturing and service firms are revealed with regard to data-related activities in the production process of goods and services (Table A.9). In the manufacturing

sector, the use of big data is statistically significantly associated with a higher likelihood of engaging in all data-driven activities except for the use of automated data processing for the sake of efficiency gains. On the contrary, while automated data exchange with suppliers and customers as well as the integration of IT systems between business processes and divisions are on average more common in the service sector, service firms relying on big data analytics are not significantly more likely to engage in these data-related activities.

Interestingly, while serialized product identification enabled by sensors, RFID, barcodes or radio tags is often regarded as an important driver of big data technologies (e.g. Wamba et al., 2015; Chen et al., 2012), we only find a significant association between big data use and products with embedded sensors in the services sector (Table A.10). This is mostly driven by media and ICT services. Furthermore, in the manufacturing sector, we do not find a significant relation between big data and the provision of mobile apps or the provision of product support services supplied directly by the firm. However, manufacturing firms who indicated that they used big data technologies exhibit a statistically significant higher likelihood of relying on partner firms to offer product support services. On the contrary, in the service sector the use of big data technologies is statistically significantly associated with all variables on digital technologies embedded in products and services, which we account for in the analysis.

Overall, while our main analysis suggests that both manufacturing and service firms can equally benefit from big data analytics in their innovation processes, our in-depth analysis suggests that firms in the two sectors exploit different sources of digital information in order to reap the benefits from big data technologies.

8 Conclusions

This paper investigates the relationship between the use of big data analytics and firms' propensity to innovate, as well as firms' innovation intensity, which we measure by the sales share resulting from new products or services and which constitutes a measure of the market success of the firms' innovations. Our results show that the use of big data analytics is associated with a higher propensity to innovate, as well as a higher innovation intensity. Importantly, this relation holds when we control for the use of mature software systems and data technologies, such as Enterprise Resource Planning Software, which lack more sophisticated features encompassed by big data analytics. These results are robust with respect to various alternative specifications and econometric methods. As the knowledge production process and innovative output likely differ between manufacturing and service firms, we investigate potential effect heterogeneity with regard to the two sectors. Interestingly, the associations we measure are of similar magnitude among firms in the manufacturing and service industries. However, subsequent analyses suggest that firms in the manufacturing and service sectors that apply big data rely on different sources of digital information and different data-related firm practices to reap the benefits of big data analytics.

Furthermore, while the relation between a firm's use of big data and the likelihood of the firm innovating is not contingent on general human capital, it is contingent on firms' investment in IT-specific knowledge and skills. Overall, our results are consistent with positive returns of big data analytics in terms of product innovations at the extensive and intensive margin. They support the view that knowledge reaped from digitized data by means of big data analytics can be a relevant intangible asset in the innovation process.

References

Amemiya, T. (1984), 'Tobit Models: A Survey', *Journal of Econometrics* **24**(1-2), 3–61.

Anderson, C. (2008), 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired magazine* **16**(7), 16–07.

Andries, P. and Czarnitzki, D. (2014), 'Small Firm Innovation Performance and Employee Involvement', *Small Business Economics* **43**(1), 21–38.

Bertschek, I. and Kesler, R. (2017), Let the User Speak: Is Feedback on Facebook a Source of Firms' Innovation?, ZEW Discussion Paper 17-015, Centre for European Economic Research.

Bertschek, I., Ohnemus, J. and Viete, S. (2017), 'The ZEW ICT Survey 2002 to 2015: Measuring the Digital Transformation in German Firms', *Jahrbücher für Nationalökonomie und Statistik* . Available at: https://doi.org/10.1515/jbnst-2016-1005.

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

Brynjolfsson, E. and McElheran, K. (2016*a*), Data in Action: Data-Driven Decision Making in U.S. Manufacturing, Working Papers 16-06, Center for Economic Studies, U.S. Census Bureau.

Brynjolfsson, E. and McElheran, K. (2016*b*), 'Digitization and Innovation: The Rapid Adoption of Data-Driven Decision-Making', *The American Economic Review* **106**(5), 133–139.

Brynjolfsson, E. and Saunders, A. (2010), *Wired for Innovation: How IT is Reshaping the Economy*, The MIT Press.

Cardona, M., Kretschmer, T. and Strobel, T. (2013), 'ICT and Productivity: Conclusions from the Empirical Literature', *Information Economics and Policy* **25**(3), 109–125.

Cassiman, B. and Veugelers, R. (2006), 'In Search of Complementarity in Innovation Strategy: Internal R&D and External Knowledge Acquisition', *Management Science* **52**(1), 68–82.

Chen, H., Chiang, R. H. and Storey, V. C. (2012), 'Business Intelligence and Analytics: From Big Data to Big Impact.', *MIS Quarterly* **36**(4).

Chen, M., Mao, S. and Liu, Y. (2014), 'Big Data: A Survey', *Mobile Networks and Applications* **19**(2), 171–209.

Constantiou, I. D. and Kallinikos, J. (2015), 'New Games, New Rules: Big Data and the Changing Context of Strategy', *Journal of Information Technology* **30**(1), 44–57.

Draca, M., Sadun, R. and Van Reenen, J. (2007), Productivity and ICT: A Review of the Evidence, *in* C. Avgerou, R. Mansell, D. Quah and R. Silverstone, eds, 'The Oxford Handbook of Information and Communication Technologies', Oxford University Press, pp. 100–147.

Dumbill, E. (2013), 'Making Sense of Big Data', *Big Data* **1**(1), 1–2.

Engelstätter, B. (2012), 'It's not all About Performance Gains – Enterprise Software and Innovations', *Economics of Innovation and New Technology* **21**(3), 223–245.

Erevelles, S., Fukawa, N. and Swayne, L. (2016), 'Big Data Consumer Analytics and the Transformation of Marketing', *Journal of Business Research* **69**(2), 897–904.

European Commission (2014), 'Towards a Thriving Data-Driven Economy'. COM(2014) 442 final. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1404888011738&uri=CELEX:52014DC0442. Last Accessed: 26 September 2017.

George, G., Haas, M. R. and Pentland, A. (2014), 'Big Data and Management', *Academy of Management Journal* **57**(2), 321–326.

Gobble, M. M. (2013), 'Big Data: The Next Big Thing in Innovation', *Research-Technology Management* **56**(1), 64–66.

Goodridge, P. and Haskel, J. (2015), How Does Big Data Affect GDP? Theory and Evidence for the UK, Discussion Paper 2015/06, Imperial College Business School.

Griliches, Z. (1979), 'Issues in Assessing the Contribution of R&D to Productivity Growth', *Bell Journal of Economics* **10**(1), 92–116.

Haller, S. A. and Siedschlag, I. (2011), 'Determinants of ICT Adoption: Evidence from Firm-Level Data', *Applied Economics* **43**(26), 3775–3788.

Hempell, T. and Zwick, T. (2008), 'New Technology, Work Organisation, And Innovation', *Economics of Innovation and New Technology* **17**(4), 331–354.

Hitt, L. M., Jin, F. and Wu, L. (2016), Data Analytics Skills and the Corporate Value of Social Media, Research Paper No. 16-61, Kelley School of Business.

Hottenrott, H. and Lopes-Bento, C. (2016), 'R&D Partnerships and Innovation Performance: Can There be too Much of a Good Thing?', *Journal of Product Innovation Management* **33**(6), 773–794.

Huergo, E. and Jaumandreu, J. (2004), 'How Does Probability of Innovation Change with Firm Age?', *Small Business Economics* **22**(3-4), 193–207.

Kitchin, R. (2017), 'Big Data-Hype or Revolution', *The SAGE Handbook of Social Media Research Methods* pp. 27–39.

Kshetri, N. (2014), 'Big Data's Impact on Privacy, Security and Consumer Welfare', *Telecommunications Policy* **38**(11), 1134–1145.

Laney, D. (2001), 3D Data Management: Controlling Data Volume, Velocity and Variety, Application Delivery Strategies, 6 February 2001, META Group.

Laursen, K. and Salter, A. (2006), 'Open for Innovation: The Role of Openness in Explaining Innovation Performance among UK Manufacturing Firms', *Strategic Management Journal* **27**(2), 131–150.

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014), 'The Parable of Google Flu: Traps in Big Data Analysis', *Science* **343**(6176), 1203–1205.

Lööf, H. and Heshmati, A. (2006), 'On the Relationship Between Innovation and Performance: A Sensitivity Analysis', *Economics of Innovation and New Technology* **15**(4-5), 317–344.

Luckow, A., Kennedy, K., Manhardt, F., Djerekarov, E., Vorster, B. and Apon, A. (2015), Automotive Big Data: Applications, Workloads and Infrastructures, *in* '2015 IEEE International Conference on Big Data (Big Data)', IEEE, pp. 1201–1210.

Maddala, G. S. (1986), *Limited-Dependent and Qualitative Variables in Econometrics*, number 3 *in* 'Econometric Society Monographs', Cambridge University Press.

Mairesse, J. and Mohnen, P. (2002), 'Accounting for Innovation and Measuring Innovativeness: An Illustrative Framework and an Application', *The American Economic Review* **92**(2), 226–230.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011), *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute. Available at http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation. Last Accessed: 26 September 2017.

McAfee, A. and Brynjolfsson, E. (2012), 'Big Data: The Management Revolution', *Harvard Business Review* **90**(10), 60–68.

Mortensen, P. S., Bloch, C. W. et al. (2005), *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data*, Organisation for Economic Cooperation and Development, OECD, Paris.

OECD (2015), *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris, Available at http://dx.doi.org/10.1787/9789264229358-en.

Pakes, A. and Griliches, Z. (1980), 'Patents and R&D at the Firm Level: A First Report', *Economics Letters* **5**(4), 377–381.

Papke, L. E. and Wooldridge, J. M. (1996), 'Econometric Methods for Fractional Response Variables with an Application to 401 (k) Plan Participation Rates', *Journal of Applied Econometrics* **11**(6), 619–632.

Peters, B. and Schmiele, A. (2010), The Influence of International Dispersed vs. Home-Based R&D on Innovation Performance, ZEW Discussion Paper 10-102, Centre for European Economic Research.

Raymond, W., Mairesse, J., Mohnen, P. and Palm, F. (2015), 'Dynamic Models of R&D, Innovation and Productivity: Panel Data Evidence for Dutch and French Manufacturing', *European Economic Review* **78**, 285–306.

Saunders, A. and Tambe, P. (2015), 'Data Assets and Industry Competition: Evidence from 10-K Filings'. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2537089.

Schumpeter, J. A. (1942), *Capitalism, Socialism and Democracy*, Harper & Brothers, London.

Spiezia, V. (2011), 'Are ICT Users More Innovative?: An Analysis of ICT-Enabled Innovation in OECD Firms', *OECD Journal: Economic Studies* **2011**(1), 1–21.

Tambe, P. (2014), 'Big Data Investment, Skills, and Firm Value', *Management Science* **60**(6), 1452–1469.

Tobin, J. (1958), 'Estimation of Relationships for Limited Dependent Variables', *Econometrica: Journal of the Econometric Society* **26**(1), 24–36.

Van Reenen, J., Bloom, N., Draca, M., Kretschmer, T. and Sadun, R. (2010), *The Economic Impact of ICT*, Centre for Economic Performance, London School of Economics. Available at http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=669. Last Accessed: 26 September 2017.

Varian, H. R. (2010), 'Computer Mediated Transactions', *The American Economic Review* **100**(2), 1–10.

Vinding, A. L. (2006), 'Absorptive Capacity and Innovative Performance: A Human Capital Approach', *Economics of Innovation and New Technology* **15**(4-5), 507–517.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015), 'How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study', *International Journal of Production Economics* **165**, 234–246.

Wang, G., Gunasekaran, A., Ngai, E. W. and Papadopoulos, T. (2016), 'Big Data Analytics in Logistics and Supply Chain Management: Certain Investigations for Research and Applications', *International Journal of Production Economics* **176**, 98–110.

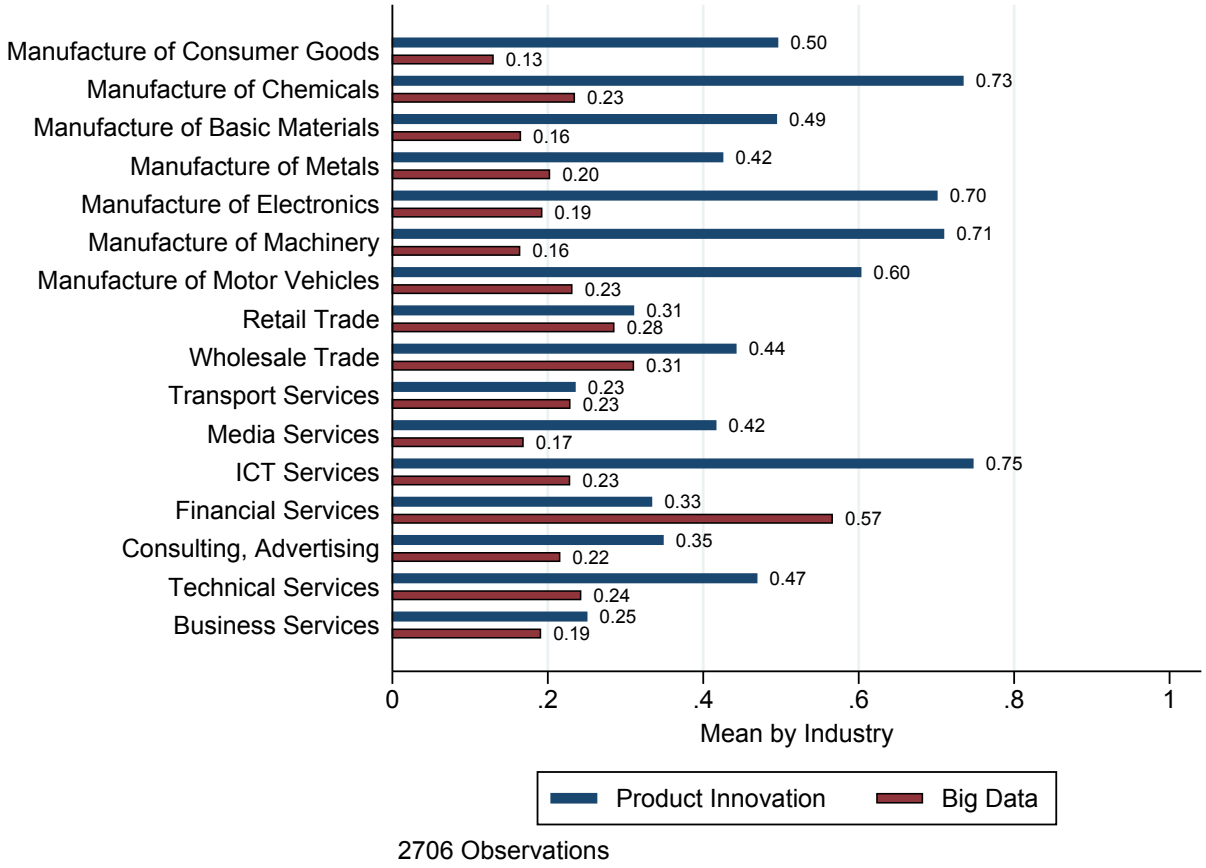
Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, Mass.

Wu, L. and Hitt, L. M. (2016), How Do Data Skills Affect Firm Productivity: Evidence from Process-Driven vs. Innovation-Driven Practices. Working Paper.

Zhong, R. Y., Huang, G. Q., Lan, S., Dai, Q., Chen, X. and Zhang, T. (2015), 'A Big Data Approach for Logistics Trajectory Discovery from RFID-Enabled Production Data', *International Journal of Production Economics* **165**, 260–272.

A Appendix

Figure A.1: Industry Means of Product Innovation and Big Data: Estimation Sample



Source: ZEW ICT survey 2015.

Table A.1: Distribution of Firms across Industries: Estimation Sample

	N	Percentage
Manufacture of Consumer Goods	448	16.56
Manufacture of Chemicals	94	3.47
Manufacture of Basic Materials	249	9.20
Manufacture of Metals	193	7.13
Manufacture of Electronics	177	6.54
Manufacture of Machinery	165	6.10
Manufacture of Motor Vehicles	78	2.88
Retail Trade	158	5.84
Wholesale Trade	129	4.77
Transport Services	149	5.51
Media Services	125	4.62
ICT Services	158	5.84
Financial Services	129	4.77
Consulting, Advertising	158	5.84
Technical Services	128	4.73
Business Services	168	6.21
Total	2706	100.00

SOURCE: ZEW ICT-Survey 2015.

Table A.2: Summary Statistics by Big Data Use of Firms: Estimation Sample

	No Big Data		Use Big Data		Total	
	N	Mean	N	Mean	N	Mean
Product Innovation	2121	0.44	585	0.60	2706	0.48
% of Sales New Product	2121	0.07	585	0.12	2706	0.08
Big Data	2121	0.00	585	1.00	2706	0.22
% of Emp. Predom. Using PC	2121	0.42	585	0.55	2706	0.45
% of Emp. Using Internet	2121	0.55	585	0.65	2706	0.57
Enterprise Software	2121	0.50	585	0.77	2706	0.56
% of R&D Expenses	2121	0.04	585	0.07	2706	0.05
Employees	2121	62.22	585	187.53	2706	89.31
Employees (in logs)	2121	3.23	585	4.17	2706	3.43
Investment in Mill. Euro	2121	0.53	585	2.12	2706	0.88
Investment (in logs)	2121	-2.29	585	-1.09	2706	-2.03
Exporter	2121	0.44	585	0.49	2706	0.45
% Highly Qualified Employees	2121	0.19	585	0.21	2706	0.19
% Medium Qualified Employees	2121	0.63	585	0.61	2706	0.63
% of Employees < Age 30	2121	0.23	585	0.26	2706	0.24
% of Employees > Age 50	2121	0.28	585	0.26	2706	0.27
East Germany	2121	0.25	585	0.22	2706	0.25
% of Emp. IT-Training	2121	0.08	585	0.13	2706	0.09
Age (in logs)	2121	3.13	585	3.31	2706	3.17
Group	2121	0.26	585	0.43	2706	0.29
Multinational	2121	0.08	585	0.15	2706	0.09

Source: ZEW ICT-Survey 2015.

Table A.3: Heckman Selection Model (Without Exclusion Restriction), Marginal Effects

	Full Sample		Manufacturing		Services	
	(1)	(2)	(3)	(4)	(5)	(6)
	1st	2nd	1st	2nd	1st	2nd
Big Data	0.066***	0.022***	0.065*	0.024**	0.067**	0.025*
	(0.022)	(0.008)	(0.034)	(0.011)	(0.030)	(0.013)
% of Emp. Predom. Using PC	-0.015	0.001	-0.110	0.017	0.048	-0.013
	(0.043)	(0.017)	(0.072)	(0.023)	(0.054)	(0.028)
% of Emp. Using Internet	0.074**	-0.005	0.075	0.010	0.069	-0.025
	(0.036)	(0.014)	(0.050)	(0.016)	(0.053)	(0.028)
Enterprise Software	0.082***	-0.007	0.112***	-0.002	0.059**	-0.014
	(0.020)	(0.008)	(0.030)	(0.010)	(0.026)	(0.013)
% of R&D Expenses	0.947***	0.223***	1.224***	0.286***	0.792***	0.180***
	(0.111)	(0.026)	(0.202)	(0.040)	(0.127)	(0.037)
Employees (in logs)	0.011	-0.020***	0.015	-0.010*	0.010	-0.031***
	(0.012)	(0.005)	(0.017)	(0.006)	(0.015)	(0.008)
Investment (in logs)	0.024***	0.004	0.018	-0.005	0.029***	0.014***
	(0.008)	(0.003)	(0.011)	(0.004)	(0.010)	(0.005)
Exporter	0.164***	-0.007	0.141***	-0.010	0.184***	-0.010
	(0.021)	(0.008)	(0.029)	(0.011)	(0.031)	(0.014)
% Highly Qualified Employees	0.154**	0.002	0.358***	-0.039	0.040	0.045
	(0.062)	(0.023)	(0.113)	(0.035)	(0.081)	(0.041)
% Medium Qualified Employees	-0.040	-0.010	-0.018	-0.027	-0.099	0.034
	(0.043)	(0.017)	(0.057)	(0.019)	(0.068)	(0.037)
% of Employees < Age 30	-0.026	0.027	-0.068	0.068***	-0.004	-0.008
	(0.051)	(0.020)	(0.076)	(0.026)	(0.069)	(0.033)
% of Employees > Age 50	-0.015	0.007	-0.049	0.006	0.017	0.017
	(0.048)	(0.020)	(0.067)	(0.024)	(0.069)	(0.035)
East Germany	0.005	0.000	0.032	0.004	-0.036	-0.001
	(0.021)	(0.008)	(0.028)	(0.010)	(0.030)	(0.016)
% of Emp. IT-Training	0.133***	0.004	0.166	-0.017	0.118**	0.017
	(0.051)	(0.017)	(0.113)	(0.028)	(0.055)	(0.023)
Age (in logs)	-0.010	-0.013***	0.004	-0.009*	-0.025*	-0.019***
	(0.010)	(0.004)	(0.013)	(0.005)	(0.014)	(0.007)
Group	0.034*	-0.001	0.054*	0.007	0.014	-0.010
	(0.021)	(0.008)	(0.031)	(0.010)	(0.028)	(0.013)
Multinational	0.133***	0.012	0.129***	-0.004	0.122**	0.036*
	(0.036)	(0.010)	(0.047)	(0.012)	(0.055)	(0.020)
Industry Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2706	2706	1404	1404	1302	1302
$\hat{\sigma}_{12}$	-0.270		-0.316		-0.311	
LR-Test $H_0 : \sigma_{12} = 0$ $[\chi^2(1)]$, p-Val	0.113		0.068		0.262	
Log Likelihood	-969.413		-411.619		-524.052	

Standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

All models include an intercept.

Source: ZEW ICT-Survey 2015.

Table A.4: Variable Descriptions for Supplementary Survey

Variable	Description/Question
Social Media and Online Content	
s_feedback	firm allows customers to evaluate products or services online
s_content	firm systematically searches for user-generated content about own products or services or about the company
s_ads	firm engages in online advertising
Data in Production of Goods and Services	
We use automated data recording, processing and transmission in order to...	
p_efficiency	...make internal processes more efficient, e.g. reduce material or energy consumption.
p_assistance	...provide our employees with digital assistance systems, e.g. in logistics, production, maintenance.
p_edi	...exchange information with suppliers and customers.
p_customize	...customize products/services to individual customer needs.
p_adapt	...adapt internal processes flexibly or fix errors.
p_network	firm introduced integration of IT between different business processes or divisions during the previous two years
Digital Technology in Goods and Services	
g_sensor	firm manufactures products with embedded RFID-Chips, QR-codes, sensors
g_apps	firm offers mobile apps for products or services
g_service	firm offers product support services, e.g. online platforms, software
g_partner_service	partner firms offer product support services

Source: ZEW ICT-Survey 2015 and supplementary survey 2015/2016.

Table A.5: Summary Statistics - Social Media and Online Content

	Manufacturing		Services	
	Mean	SD	Mean	SD
s_feedback	0.18	0.39	0.20	0.40
s_content	0.22	0.42	0.24	0.43
s_ads	0.22	0.42	0.26	0.44
Observations	872		726	

Source: ZEW ICT-Survey 2015 and supplementary survey 2015/2016.

Table A.6: Summary Statistics - Data in Production of Goods and Services

	Manufacturing		Services	
	Mean	SD	Mean	SD
p_efficiency	0.38	0.49	0.33	0.47
p_assistance	0.53	0.50	0.52	0.50
p_edi	0.48	0.50	0.56	0.50
p_network	0.21	0.41	0.29	0.46
p_customize	0.29	0.45	0.41	0.49
p_adapt	0.39	0.49	0.49	0.50
Observations	872		726	

Source: ZEW ICT-Survey 2015 and supplementary survey 2015/2016.

Table A.7: Summary Statistics - Digital Technology in Goods and Services

	Manufacturing		Services	
	Mean	SD	Mean	SD
g_sensor	0.12	0.33	0.07	0.26
g_apps	0.06	0.23	0.19	0.39
g_service	0.22	0.42	0.38	0.49
g_partner_service	0.15	0.36	0.34	0.47
Observations	872		726	

Source: ZEW ICT-Survey 2015 and supplementary survey 2015/2016.

Table A.8: Controlled Correlations: Big Data and Social Media and Online Content

	s_feedback	s_content	s_ads
Manufacturing	0.077**	0.093**	0.054
	(0.039)	(0.041)	(0.040)
Services	0.049	0.030	0.107**
	(0.037)	(0.039)	(0.041)

NOTES: This Table shows partial correlations between big data use and firms' utilization of various aspects of social media and online content. Parameter estimates of OLS regression analysis are shown. Included control variables are the share of employees working predominantly with PCs, an indicator for use of ERP software, employees (in logs), the share of highly and medium qualified employees, the share of employees who received IT-training, industry and regional indicators. Results are based on 872 obs. in manufacturing and 726 obs. in services. Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

Table A.9: Controlled Correlations: Big Data and Data in Production of Goods and Services

	p_efficiency	p_assistance	p_edi	p_network	p_customize	p_adapt
Manufacturing	0.069	0.139***	0.088*	0.133***	0.106**	0.146***
	(0.042)	(0.039)	(0.045)	(0.043)	(0.044)	(0.043)
Services	0.020	0.070	0.049	0.012	0.094**	0.087*
	(0.043)	(0.043)	(0.043)	(0.042)	(0.045)	(0.044)

NOTES: This Table shows partial correlations between big data use and various forms of firms' utilization of data in the production process. Parameter estimates of OLS regression analysis are shown. Included control variables are the share of employees working predominantly with PCs, an indicator for use of ERP software, employees (in logs), the share of highly and medium qualified employees, the share of employees who received IT-training, industry and regional indicators. Results are based on 872 obs. in manufacturing and 726 obs. in services. Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

Table A.10: Controlled Correlations: Big Data and Digital Technology in Goods and Services

	g_sensor	g_apps	g_service	g_partner_service
Manufacturing	0.011	0.038	0.012	0.079**
	(0.033)	(0.026)	(0.041)	(0.037)
Services	0.056**	0.100***	0.080*	0.071*
	(0.026)	(0.037)	(0.042)	(0.043)

NOTES: This Table shows partial correlations between big data use and various forms of digital technology embedded in final goods and services. Parameter estimates of OLS regression analysis are shown. Included control variables are the share of employees working predominantly with PCs, an indicator for use of ERP software, employees (in logs), the share of highly and medium qualified employees, the share of employees who received IT-training, industry and regional indicators. Results are based on 872 obs. in manufacturing and 726 obs. in services. Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.