Imperial College London
BUSINESS SCHOOL

Big Data in UK industries: An intangible investment approach

Peter Goodridge, Jonathan Haskel



Discussion Paper 2016/01
May 2016

Big Data in UK industries: An intangible investment approach*

Peter Goodridge
Imperial College Business School

Jonathan Haskel
Imperial College Business School; CEPR and IZA

September 2015

Abstract

In Goodridge and Haskel (2015b) we present an economic framework for measuring the contribution of Big Data to UK growth, building on measures of investment and employment presented in Goodridge and Haskel (2015a) and Chebli, Goodridge et al. (2015) respectively. In this paper we extend that framework to present industry estimates of UK investment and employment in producing data-based information and knowledge assets, and their contribution to industry-level growth. In doing so, we focus on industries that are considered to be intensive users of knowledge or intangible capital, including for instance Financial services and Manufacturing. We find that the four industries: Information and Communication, Professional and Administrative Services, Financial Services and Manufacturing account for 88% of big data employment and 89% of investments in data-based assets in the UK market sector. Similarly, of the total contribution of data-based assets to UK growth, we find that it is fully accounted for by these four industries.

*Contacts: Peter Goodridge, Jonathan Haskel, Imperial College Business School, Imperial College, London. SW7 2AZ. p.goodridge10@ic.ac.uk j.haskel@ic.ac.uk. We thank Omar Chebli for the dataset used. We are very grateful for financial support for this research from EPSRC (EP/K039504/1 and EP/I038837/1). We also thank Rashik Parmar (IBM) and attendees of a forum hosted by TechUK for useful thoughts and insights. This work contains statistical data from ONS which is Crown copyright and reproduced with the permission of the controller of HMSO and Queen's Printer for Scotland. The use of these data does not imply the endorsement of the data owner or the UK Data Service at the UK Data Archive in relation to the interpretation or analysis of the data. This work uses research datasets which may not exactly reproduce National Statistics aggregates. All errors are of course our own.

1. Introduction

There has been much speculation that activity in and around big data will have dramatic implications for business models, growth and productivity in particular industries (e.g. Mayer-Schönberger and Cukier (2013) and OECD (2013)). For instance, Manyika, Chui et al. (2011) look in particular at potential applications of big data and data analytics in the retail, manufacturing, telecommunications, healthcare and public sector industries and provide estimates of its potential impact on future industry revenues and productivity growth. In this paper we combine intangible investment measurement and growth-accounting methods to produce empirical estimates of the contribution of data-based knowledge capital to growth in UK market sector industries.

Goodridge and Haskel (2015b) set out an economic framework for estimating the contribution of (big) data to UK market sector growth, building on measures of investment and employment in data activity, estimated in Goodridge and Haskel (2015a) and Chebli, Goodridge et al. (2015) respectively. In this paper we disaggregate those data to a nine-industry market sector breakdown, and look in more detail at the intensity of data-based investment activity in each industry and the contribution of data-based knowledge capital to industry growth in value-added. In doing so we use an occupational-based approach which allows us to also consider how the nature of big data activity may differ between industries.

Our main findings are as follows. First, we find that the composition of big data employment and investment differs by industry. In Manufacturing and other production industries, big data activity is typically conducted by workers in scientific, engineering and research occupations as well as those in occupations typically considered engaged in the writing of software. In services, big data workers are predominantly among software and ICT occupations, but also include significant numbers of consultants, economists and statisticians, and business associate professionals. Second, we find that just four industries: Information and Communication, Professional and Administrative Services, Financial Services and Manufacturing account for 88% of big data employment. Third, these same four industries carry out 89% of UK market sector investment in data-based assets. Note, we estimate that total UK market sector investment in data-based assets in 2010 was £5.7bn, growing to £6.6bn in 2012. In the same years, UK market sector investment in R&D is estimated at £14.8bn in 2010, and £15.5bn in 2012 (Goodridge and Haskel 2015a). Thus we estimate UK investments in data-based assets as around 40% of those in R&D, and growing. Fourth, of the total contribution of data-based assets to UK growth, we find that it is fully accounted for by just these four industries.

The rest of this paper is set out as follows. Section two sets out our framework, first using an informal exposition of the big data process and then a more a formal model. Section three describes our data and section four our results. Section five concludes.

2. Our framework

2.1. An informal exposition: The big data supply chain

Before introducing our approach to the big data production process, it will be useful to first summarise the definitions of some terms, further reasoning can be found in Goodridge and Haskel (2015b).

On *data*, we define two kinds of data: raw records and transformed data. Raw records are raw data not yet cleaned, formatted or transformed ready for analysis. They can include, for instance, data scraped from the web, data generated by transactions between agents, data generated by sensors embedded in machines, or data generated as a by-product of some other business operation or process. Transformed data are those that have been cleaned, formatted, combined and/or structured such that they are suitable for some form of data analytics.

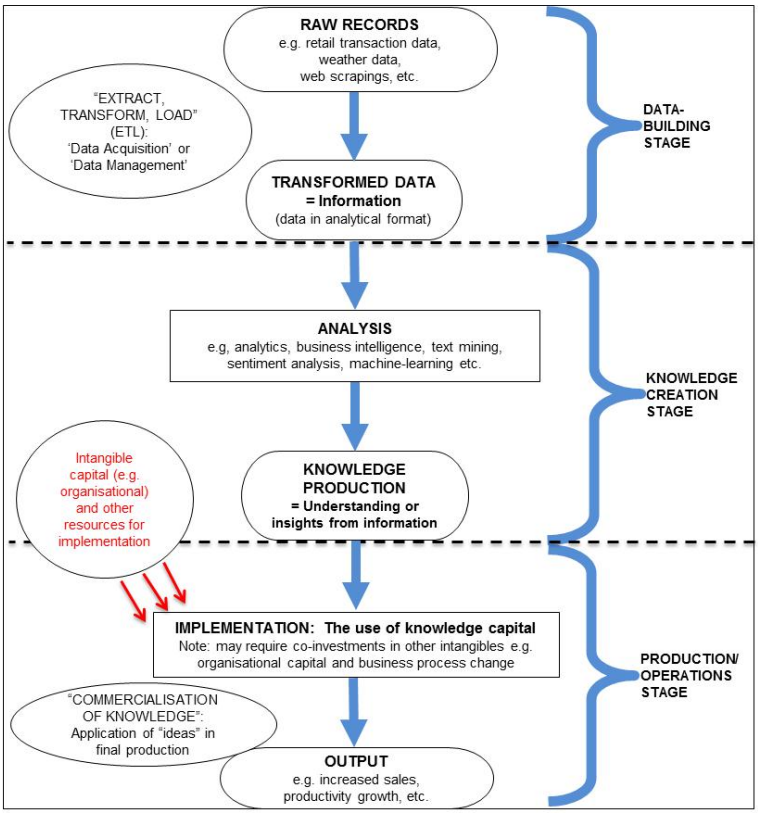
Turning to *information*, Shapiro and Varian (1998) take information to mean anything that can be digitised, thereby implicitly defining information as digitised data. We consider information in a similar vein and treat it as synonymous with transformed data. For example, analysable data on two variables, such as the prices and quantities of goods sold, constitutes information.

We define *knowledge* as the connections made between pieces of information, supported by evidence, to form a coherent understanding. Knowledge cannot exist without information, and knowledge is required to fully understand and interpret information. Knowledge can therefore include theories, hypotheses, correlations, or causal relationships observed in data. To continue with the same example, the observed correlation between the price of a good and the quantity sold constitutes knowledge. Note that different pieces of knowledge can be formed from the same piece of information (Fransman 1998), suggesting that information can be used repeatedly in the formation of new knowledge, as is explicit in the framework we present below.

With these definitions in mind, we now summarise the supply chain around big data, that is, the process of producing transformed data (information) and commercial data knowledge, and the use of that knowledge in final production. The following diagram illustrates how these concepts fit into that process. [1]

[1] We thank Rashik Parmar (IBM), Christopher Royles (Oracle), Harvey Lewis (Deloitte), Chris Francis (SAP) and Jon Steinberg (Google) for insights around the process of data transformation and data-based knowledge creation, and discussion around the value chain presented in Figure 1.

Figure 1: Data, Information and Knowledge: Data-building, knowledge creation and the commercialisation of knowledge



Note to figure: Commercialisation is the embodiment of knowledge into the output of goods and services, which may be sold for profit or made freely available. We therefore use the term commercialisation as our focus is on the market sector, but note that the framework can also be applied to the non-market sector.

Figure 1 shows that we consider the process of creating, and using, data-based knowledge, as consisting of three "stages". We use "stages" and not "industries" or "firms" since a stage might exist in-house, that is within the same firm, or within distinct specialist firms: this is discussed below.[2] We also assume for the moment that the firm employs no other intangible capital (such as R&D), again this assumption is relaxed below.

[2] Currently it is expected that the three stages predominantly exist in-house. However, as the field develops, it is likely that more companies will specialise at different points in the chain/process (i.e. provision of raw records, producers of information, producers of data-based knowledge, etc.). As an example, Google are a case where all three stages exist in-house. As a by-product of providing search services, Google automatically generate raw records on the search histories of users. They then employ labour and capital to manage, clean and transform those data into an analytical format, producing information. Google then use that transformed data (i.e. it rents from the Google stock of transformed data) to produce commercial knowledge. As a trivial example, this may be the knowledge that users that search for product X (say, flights) also consume product Z (say, hotel accommodation). In the downstream, Google sell advertising services to other firms. In doing so Google rents from its stock of commercial knowledge to sell advertising that can be targeted at specific consumers e.g. in this example, hotels in a region advertise to those searching for aeroplane flights to that area. Alternatively, consider a firm such as Experian. They operate in the knowledge creation stage, buying or acquiring transformed data from numerous sources, and using that information to produce data-based knowledge which they sell to other firms. The credit scores they sell to banks are just one example of the data-based knowledge services they provide.

2.1.1. Data-Building (Transformation)

Starting at the top of the diagram, we first consider the *data-building* or *transformation* (D) process, which transforms raw records into information of a format ready for analysis. Thus data building may involve digitising, structuring, aggregating, formatting, and/or cleaning. This process is sometimes referred to as "data management", "data acquisition" or "data warehousing". The literature on data warehousing and data analytics commonly describes this as the ETL process, an acronym for 'Extract, Transform, Load'. Using the above definitions: 'Extract' is the extraction of raw records; 'Transform' refers to the transformation of raw records into data, often of improved quality, of a format ready for analysis; and 'Load' to the loading of the data into the database or data warehouse. The linking, matching and aggregation of datasets may take place in this stage, or later in the knowledge creation stage.

2.1.2. Knowledge creation

The next stage is the *knowledge creation* (N) process, more commonly referred to as 'data analytics'. This sector takes the output of the data-building stage, and uses that data/information to conduct analysis. That analysis could take a number of forms. It will include activities commonly referred to in the literature as 'data science', 'data/text mining', 'knowledge recovery', 'business intelligence' and 'machine learning', with the latter referring to the use of artificial intelligence to discover correlations in data. Whatever the method, the output of the analytics process is a piece of commercial knowledge formed from the analysis of information, and used to construct advice to be implemented in the final production of goods and services.

2.1.3. Downstream production of final goods and services

The final stage incorporates the application of knowledge in the downstream production of final goods and services (Y). We emphasise that the downstream is a pure operations stage, that does not create any form of capital, but just employs labour and (tangible and intangible) capital to deliver final goods and services. Therefore, use of data-based knowledge in the downstream does not equate to investment in the downstream. The downstream is a pure using sector, with all investment occurring in the upstream.

However, we are aware that the implementation of data-based knowledge in downstream production may require co-investments in other forms of intangible capital such as organisational (business process change) or reputational (brand) capital. There are of course other upstreams that create such other forms of intangible capital that are also used in downstream production, but we do not seek to measure those here. Rather, our focus here is on the measurement of the data-building and data-based

knowledge creation upstreams. For estimates of a fuller range of intangible investment by industry, see Goodridge, Haskel et al. (2014).

As noted above, the upstream stages may either be situated in-house or in specialist firms operating along the value chain presented in Figure 1. In the case where these stages exist in distinct firms, the knowledge (or advice) could be sold to the downstream firm for an explicit fee, just as plant and machinery is typically sold for an observed price. In the case where these stages exist in-house, the downstream operations unit will receive advice from the upstream knowledge creation unit, located in the same firm, for which it must pay an implicit but unobserved rental, just as the knowledge creation stage must pay a rental for the use of transformed data.[3]

The downstream therefore receives advice formed on the basis of knowledge and takes action to implement that knowledge in final production. For instance, it could be the knowledge that the cross-promotion of goods results in increased sales, or it could be a re-optimisation of downstream processes to improve productivity, based on say knowledge acquired from data emitted from machine-embedded sensors (the "internet of things"). We refer to this implementation as the commercialisation of knowledge. The term commercialisation obviously has connotations with the market and a profit motive. That is because our focus here is on knowledge creation in the market sector. We emphasise however that the framework can be applied more generally to the application of knowledge in non-market production, such as in the delivery of public services.

2.1.4. Summary

We therefore propose to treat data-based information and knowledge as capital assets in an industry growth-accounting framework. For a full justification of the treatment of these goods as capital assets according to economic and SNA investment criteria, see Goodridge and Haskel (2015a) and Goodridge and Haskel (2015b).

2.2. A formal model

The previous section defined the terms that underlie our framework and presented an informal exposition of the processes of data transformation, knowledge creation and commercialisation. In this section we present a model of the industry payments and productivity underlying that process. The model explicitly treats transformed data (information) and data-based knowledge as capital assets in national accounting and growth-accounting frameworks.

[3] The treatment is therefore perfectly symmetrical with purchased tangible capital (i.e. buildings, machinery etc.), for which a firm pays an implicit but unobserved annual rental for use of the asset.

The framework is analogous to the market sector model in Goodridge and Haskel (2015b) but with an added industry dimension. In the model each industry consists of three sectors. First, there are the two upstream sectors: the data-building (transformation) sector ("D") and the knowledge creation sector ("N"). We emphasise that the upstream sectors can of course exist in-house. Finally, there is the downstream sector, which is engaged in the production of final goods and services. In equation (1) we present the value-added production function and income accounting identity for the three sectors in each industry, i.

$$
\begin{aligned}
D_{i,t} &= A_{i,t}^{D} F_{i}^{D}(L_{i,t}^{D}, K_{i,t}^{D}, R_{i,t}^{D}); \qquad & P_{i,t}^{D} D_{i,t} &= \mu_{i,t}^{D}\left(P_{i,t}^{L,D} L_{i,t}^{D} + P_{i,t}^{K,D} K_{i,t}^{D}\right) \\
N_{i,t} &= A_{i,t}^{N} F_{i}^{N}(L_{i,t}^{N}, K_{i,t}^{N}, B_{i,t}^{N}); \qquad & P_{i,t}^{N} N_{i,t} &= \mu_{i,t}^{N}\left(P_{i,t}^{L,N} L_{i,t}^{N} + P_{i,t}^{K,N} K_{i,t}^{N} + P_{i,t}^{B,N} B_{i,t}^{N}\right) \\
Y_{i,t} &= A_{i,t}^{Y} F_{i,t}^{Y}(L_{i,t}^{Y}, K_{i,t}^{Y}, R_{i,t}^{Y}); \qquad & P_{i,t}^{Y} Y_{i,t} &= P_{i,t}^{L,Y} L_{i,t}^{Y} + P_{i,t}^{K,Y} K_{i,t}^{Y} + P_{i,t}^{R,Y} R_{i,t}^{Y}
\end{aligned}
\tag{1}
$$

Taking each industry-sector in turn. In each industry, the upstream *data-building* sector ($D_{i,t}$) manages data and transforms industry raw records ($R_{i,t}^{D}$) into data (information) of a format ready for analysis. Real transformed data output is thus a function of primary factors, labour ($L_{i,t}^{D}$), capital ($K_{i,t}^{D}$) and raw records ($R_{i,t}^{D}$). Sectoral technical progress is represented by $A_{i,t}^{D}$, which incorporates among other things, disembodied technical change and the contribution of freely available knowledge, diffusion of knowledge from other industry-sectors and factor utilisation. The income identity shows that nominal industry-sector output ($P_{i,t}^{D} D_{i,t}$) is equal to the sum of factor payments multiplied by a factor, $\mu_{i,t}^{D}$. Note that there are no factor payments for raw records. This is because we do not model raw records as an asset, but rather as a raw material that may either be generated for free or almost free, where data comes as exhaust data, or paid for in the same way as other material/intermediate inputs.

The factor $\mu_{i,t}^{D}$ enters the output identity because the data-building sector might be able to mark up prices over competitive costs. First, the firm/industry might either have access to a unique type of raw records or be in a position to generate unique information assets. Second, it might be able to patent its information asset. Third, there might be increasing returns,[4] for example, if data is non-rival and can be shared in the production of goods (e.g. mistakes from Google searches are also used for Google's spellchecker): this is the mechanism in Romer (1991) for example. In practice, the value of the mark-up will differ for each individual information asset produced, dependent on the degree of product

[4] As shown in a series of papers from Basu and Fernald e.g. Basu, Fernald et al. (2001), mark-ups and increasing returns to scale are linked: mark-ups greater than one imply that factor elasticities sum to more than one, which is the definition of increasing returns to scale.

market competition, the scarcity of that information, and its commercial value to the ultimate users. Of course the acquisition and maintenance of this market power provides a further incentive for the upstream to exist in-house.[5]

From (1), in each industry the data-building sector produces information assets ($D_{i,t}$) which evolve into an industry stock of (bytes of) information according to the perpetual inventory method (PIM):

$$B_{i,t} = D_{i,t} + (1-\delta_i^B)B_{i,t-1} \quad (2)$$

Where $B_{i,t}$ are accumulated bytes of information and δ_i^B is a geometric rate of depreciation which in theory may be firm- or industry-specific. One might think that provided it is stored, information does not physically depreciate since it is not subject to wear and tear. But, as noted in Mayer-Schönberger and Cukier (2013), data/information assets do depreciate economically. For example, a retailer may retain the transaction histories of their customers. However, as consumers age, their tastes and preferences typically change. As a result, firms actively test their data to separate the useful data from that which has become less useful, with the latter culled. CEBR (2013) also emphasise that data does not have an infinite life, and some data can quickly become outdated, for instance social media and financial trading data. Some might have to be deleted as well due to regulation. Note some of these factors may be industry-specific. Therefore information (like other intangible) assets, do depreciate, not due to wear and tear, but rather due to obsolescence and decay in the profile of revenues they earn. This concept of depreciation applicable to intangible assets was first introduced in Pakes and Schankerman (1984).

Consider now the *knowledge creation* ($N_{i,t}$) sector. That sector uses transformed data to create commercial data knowledge, also employing capital ($K_{i,t}^N$) and labour ($L_{i,t}^N$), and freely available knowledge (captured within $A_{i,t}^N$). Factor payments include those paid for the use of bytes of information ($P_{i,t}^{B,N}B_{i,t}^N$). Note, just as with tangible capital, these payments could be explicit rental payments (i.e. licence fees), or they could be implicit in the case where the information asset is owned in-house. Again the cost of resources devoted to producing knowledge consists of the payments to each factor, and the value of industry-sector output (commercial data knowledge assets) incorporates a product mark-up, $\mu_{i,t}^N$, to account for the market power acquired by the owners of unique data-based

[5] Of course, if there is a mark-up, the P^K in the D and N sectors are to be understood as competitive returns to capital.

knowledge assets. Again, that potential to appropriate a mark-up provides an incentive for the *knowledge creation* upstream to exist in-house.

Just as with information, the industry stock of commercial data-based knowledge (R_t) can be modelled as evolving according to the PIM:

$$R_{i,t} = N_{i,t} + (1-\delta_i^R)R_{i,t-1} \quad (3)$$

The knowledge creation sector therefore gleans knowledge and insights from information. Note that multiple pieces of knowledge can be generated from the same stock (or even piece) of information. Therefore information can be used repeatedly in the production of knowledge, and knowledge can be used repeatedly in the production of downstream output.

Finally, data-based knowledge is ultimately employed in final production in the industry *downstream* sector ($Y_{i,t}$). The downstream is a pure operations sector that produces the final goods and services typically associated with that industry, employing labour ($L_{i,t}^Y$), tangible capital ($K_{i,t}^Y$), and commercial data-based knowledge capital ($R_{i,t}^Y$). Nominal industry downstream output ($P_{i,t}^Y Y_{i,t}$) is thus the sum of industry downstream factor payments, where payments include implicit or explicit payments for data-based knowledge. Note that there is no mark-up in the downstream since we assume it is competitive, a reasonable assumption if we consider market power to derive from features associated with intangible assets such as unique knowledge, designs, superior technology or branding.[6]

In this model, all production of knowledge-based capital (KBC) takes place in the upstream sectors. We note that there are of course other upstream sectors that create various other knowledge assets. For instance, an upstream marketing sector that builds reputational capital, an upstream design sector, an upstream software sector, an upstream scientific research sector etc.. At the industry-level, the R&D upstream is a significant component of manufacturing output and similarly (own-account) software in financial services (Borgo, Goodridge et al. 2013). Here, for simplicity of exposition, we just model the data-building and knowledge creation upstreams, and the downstream.

[6] Therefore what look like mark-ups in the downstream are actually returns to intangible capital which we can explicitly account for in this framework. Returns to intangible capital therefore flow back to the relevant knowledge upstream e.g. to the producers of information or knowledge, which may or may not be located in the same firm/industry.

It was mentioned above that data-based information/knowledge assets can either be created in-house or purchased via market transactions. However those assets are acquired has no impact on the model. Consider the case of say a manufacturer that employs tangible capital with embedded sensors that monitor the performance and output of those machines. Those sensors generate raw records which are then transformed, in-house, into information assets. In turn those information assets are used to glean commercial knowledge and insights, which in turn are used to optimise downstream operational (manufacturing) business processes. The output of this hypothetical firm is the sum of output(s) from each sector: $P^D_{i,t}D_{i,t}$, $P^N_{i,t}N_{i,t}$ and $P^Y_{i,t}Y_{i,t}$. $P^N_{i,t}N_{i,t}$ includes implicit annual payments for the use of transformed data ($P^{B,N}_{i,t}B^N_{i,t}$), and $P^Y_{i,t}Y_{i,t}$ includes implicit annual payments for the use of data-based knowledge ($P^{R,Y}_{i,t}R^Y_{i,t}$). Thus we model the firm as a producer of information and knowledge assets as well as a producer of manufactured final goods. The output of assets (here $P^D_{i,t}D_{i,t}$ and $P^N_{i,t}N_{i,t}$) are related to the factor payments for their use via the Hall-Jorgenson user costs relation (Hall and Jorgenson 1967):

$$P^B_{i,t} = P^D_{i,t}(r + \delta^B_i - \pi^D_{i,t}); P^R_{i,t} = P^N_{i,t}(r + \delta^R_i - \pi^N_{i,t}) \tag{4}$$

Where r is the economy-wide nominal net rate of return to capital (i.e. a net profit rate) and π accounts for capital (holding) gains/losses from changes in the asset price. Asset-level factor payments (or capital compensation) therefore consist of a net return to capital plus depreciation, minus any holding gain, with all these components directly proportional to the nominal value of the stock ($P^D_{i,t}B^N_{i,t}$ or $P^N_{i,t}R^Y_{i,t}$).

2.2.1. Relation with industry-value-added and market sector GDP

Industry value-added is the sum of value added earned in each industry-sector. With no intermediates, this is the sum of industry-sector output i.e. the output of the industry D, N and Y sectors. Or equivalently, industry output is the sum of industry factor incomes, that is, the sum of all industry payments to labour and (tangible and intangible) capital.

$$\begin{aligned} P^V_{i,t}V_{i,t} &= P^D_{i,t}D_{i,t} + P^N_{i,t}N_{i,t} + P^Y_{i,t}Y_{i,t} \\ &= P^L_{i,t}L_{i,t} + P^K_{i,t}K_{i,t} + P^B_{i,t}B_{i,t} + P^R_{i,t}R \end{aligned} \tag{5}$$

Where:

$$P_{i,t}^{L} L_{i,t} = P_{i,t}^{L,D} L_{i,t}^{D} + P_{i,t}^{L,N} L_{i,t}^{N} + P_{i,t}^{L,Y} L_{i,t}^{Y}$$
$$P_{i,t}^{K} K_{i,t} = P_{i,t}^{K,D} K_{i,t}^{D} + P_{i,t}^{K,N} K_{i,t}^{N} + P_{i,t}^{K,Y} K_{i,t}^{Y} \quad (6)$$

Market sector value-added is the sum of industry value-added, where D and N goods are capitalised in each industry.[7]

$$\begin{aligned} P_t^V V_t &= \sum_{i=1}^{I} P_{i,t}^V V_{i,t} \\ &= \sum_{i=1}^{I} P_{i,t}^D D_{i,t} + \sum_{i=1}^{I} P_{i,t}^N N_{i,t} + \sum_{i=1}^{I} P_{i,t}^Y Y_{i,t} \\ &= \sum_{i=1}^{I} P_{i,t}^L L_{i,t} + \sum_{i=1}^{I} P_{i,t}^K K_{i,t} + \sum_{i=1}^{I} P_{i,t}^B B_{i,t} + \sum_{i=1}^{I} P_{i,t}^R R_{i,t} \end{aligned} \quad (7)$$

Thus for each industry we have the following value-added based TFP:

$$\Delta \ln TFP_{i,t} \equiv \Delta \ln V_{i,t} - \overline{v}_{i,t}^{K} \Delta \ln K_{i,t} - \overline{v}_{i,t}^{L} \Delta \ln L_{i,t} - \overline{v}_{i,t}^{R} \Delta \ln R_{i,t} \quad (8)$$

Where B is subsumed into R due to an inability to separate DN into its respective components (see Chebli, Goodridge et al. (2015) and Goodridge and Haskel (2015a) for more detail on estimation of D and N), and the terms in "v" are shares of factors costs in industry value-added averaged over two periods.

In Goodridge and Haskel (2015b) we estimate the contribution of data-based information and knowledge to growth in the UK market sector. In this paper we work at the industry level, with the relation between industries and the market sector aggregate defined as follows. First, the definition of market sector TFP is:

$$\Delta \ln TFP_t \equiv \Delta \ln V_t - \overline{v}_t^{K} \Delta \ln K_t - \overline{v}_t^{L} \Delta \ln L_t - \overline{v}_t^{R} \Delta \ln R_t \quad (9)$$

Where the "v" terms here, that are not subscripted by "i", are shares of payments in market sector wide nominal value added. Now we define the relation between industry value-added and market

[7] Note, there is a slight complication here. In the above framework (e.g equation (1)) the P^K term represents the competitive cost of capital were it observed. Since, in practice, capital compensation is estimated residually, then (5) holds but the cost of capital incorporates mark-ups which manifest as above-competitive returns to capital.

sector value-added, which is that changes in aggregate real value added are a weighted sum of changes in industry real value added:

$$\Delta \ln V_t \equiv \sum_i \overline{w}_{i,t} \Delta \ln V_{i,t}, \quad w_{i,t} = P_{i,t}^V V_{i,t} \Big/ \sum_i (P_{i,t}^V V_{i,t}), \overline{w}_{i,t} = 0.5(w_{i,t} + w_{i,t-1}) \tag{10}$$

We are now in a position to write down our desired relationship, that is the relation between economy-wide real value added growth and its industry contributions:

$$\Delta \ln V_t = \left(\sum_i \overline{w}_{i,t} \overline{v}_{i,t}^K \Delta \ln K_{i,t} \right) + \left(\sum_i \overline{w}_{i,t} \overline{v}_{i,t}^L \Delta \ln L_{i,t} \right) + \left(\sum_i \overline{w}_{i,t} \overline{v}_{i,t}^R \Delta \ln R_{i,t} \right) + \sum_i \overline{w}_{i,t} \Delta \ln TFP_{i,t} \tag{11}$$

Which says that the contributions of K_i, L_i and R_i to market sector value added growth depend upon the share of V_i in total V (w_i) and the shares of K, L and R in industry value-added. Which is equivalent to saying that the contributions of K_i, L_i and R_i depend upon their income share in aggregate value-added. The contribution of $\Delta \ln TFP_i$ also depends on the share of V_i in total V (w_i).

Finally, in reality we do not of course have one capital and labour unit, but many. These are then aggregated across different types: for labour, see below, we use, education, age (experience), and gender; for capital, different types of both tangible and intangible assets. Denoting the capital and labour types *k*, *r* and *l* we have the following industry and aggregate variables for each type where industry is defined as industry *i* and the aggregate variables are unsubscripted:

$$\begin{aligned}
&\Delta \ln K = \sum_k \overline{w}_k \Delta \ln K_k, \quad \textit{capital type k} \\
&\Delta \ln L = \sum_l \overline{w}_l \Delta \ln L_l, \quad \textit{labour type l} \\
&\Delta \ln R = \sum_r \overline{w}_r \Delta \ln R_r, \quad \textit{capital type r} \\
&\overline{w}_k = P_k^K K_k / \sum_k P_k^K K_k, \quad \overline{w}_l = P_l^L L_l / \sum_l P_l^L L_l, \quad \overline{w}_r = P_r^R R_r / \sum_r P_r^R R_r, \\
&K_i = \sum_i K_{k,i} \forall k, \quad L_i = \sum_i L_{l,i} \forall l, \quad R_i = \sum_i R_{r,i} \forall r, \\
&\overline{w}_t = 0.5(w_t + w_{t-1})
\end{aligned} \tag{12}$$

Before proceeding to the data, some further theory remarks on the measurement of capital. As pointed out by e.g. Jorgenson and Griliches (1967), the conceptually correct measure of capital in this productivity context is the flow of capital services. This raises a number of measurement problems

set out, for example, in the OECD productivity handbook (2001b). We estimate the now standard measure as follows. First, we build a real capital stock via the perpetual inventory method whereby for any capital asset k or r, the stock of assets evolves according to:

$$K_{k,t} = I_{k,t} + (1-\delta_{k,t})K_{k,t-1}$$
$$R_{r,t} = DN_{dn,t} + (1-\delta_{r,t})R_{r,t-1} \quad (13)$$

Where I and DN are real investments over the relevant period and δ the geometric rate of depreciation. Real investment comes from nominal investment deflated by an investment price index. Second, that investment price is converted into a rental price using the Hall-Jorgenson relation, where we assume an economy-wide net rate of return such that the capital rental price times the capital stock equals the total economy-wide operating surplus (on all of this, see for example, Oulton and Wallis (2014) and Oulton and Srinivasan (2003)).

3. Data

3.1. Time period

For the industry analysis, since we work with value-added we use the official ONS data from 1997 to 2011. For data-based information and knowledge, we work with the estimates described in Goodridge and Haskel (2015a), but here disaggregated by industry.

3.2. Industries

We work with the nine broad market sector industries based on SIC07 described in Table 1, thus excluding real estate, public administration & defence, education and health.

Table 1: Definition of nine industries

#	Sectors	SIC(2007) code		NACE1 sections
1	Agriculture, Mining and Utilities (AgMinUtil)	1-9 & 35-39	A	Agriculture, forestry and fishing
			B	Mining and quarrying
			D	Electricity, Gas, Steam and Air Conditioning Supply
			E	Water Supply, Sewerage, Waste Management and Remediation Activities
2	Manufacturing (Mfr)	10-33	C	Manufacturing
3	Construction (Constr)	41-43	F	Construction
4	Wholesale and Retail Trade, Accomodation and Food (RtAcc)	45-47 & 55-56	G	Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
			I	Accomodation and Food Service Activities
5	Transportation and Storage (Tran)	49-53	H	Transportation and Storage
6	Information and Communication (InfoCom)	58-63	J	Information and Communication
7	Financial Services (FinSvc)	64-66	K	Financial and Insurance Activities
8	Professional and Administrative Services (ProfAdmin)	69-82	M	Professional, Scientific and Technical Activities
			N	Administrative and Support Service Activities
9	Recreational and Personal Services (PersSvc)	90-98	R	Arts, Entertainment and Recreation
			S	Other Service Activities
			T	Activities of Households as Employers; Undifferentiated Goods and Services Producing Activities of Households for Own Use

Note to table: We break the market sector down into 9 broad industries based on SIC07, as reported above.

3.3. Outputs, inputs and factor shares

For labour composition and hours worked we use the ONS Quality-adjusted labour input (QALI) data, based on ONS person-hours by industry (Franklin and Mistry 2013). The ONS use these data along with LFS microdata to estimate composition-adjusted person hours, where the adjustment uses wage bill shares for composition groups for age, education and gender to account for services from different labour types. Person hours are annual person-hours, with persons including the employed, self-employed and those with two jobs. Data are grossed up using population weights.

We also use ONS data for industry Gross Value Added at current basic prices and the corresponding price and volume indices. Data on labour income, that is compensation of employees plus a proportion of mixed (self-employed) income, are from the ONS. Capital compensation is estimated residually as nominal gross value-added less total labour compensation. The Compensation of Employees (COE) data are consistent with the labour services data. Mixed income is allocated to labour and capital according to the ratio of labour payments to MGVA excluding mixed income, as used in the ONS publication of QALI. Gross operating surplus (GOS) is always computed as MGVA less COE so that GOS+COE=MGVA by construction. We shall of course amend capital compensation to incorporate compensation for data-based information and knowledge capital assets.

The tangible capital variables are based on Oulton and Wallis (2014). Their estimates combine the latest ONS investment series and price deflators, which only go back to 1997, with historic series to

estimate UK capital stock and capital services growth since the 1950s. The tangible capital data distinguishes four asset types, which are: buildings, computer hardware, (non-computer) plant & machinery and vehicles. We exclude dwellings (they are not capital for firm productivity analysis). We also incorporate appropriate tax adjustment factors for all assets, based on Wallis (2012b).

3.4. Details of measurement of intangible assets

In this paper we work with a national accounting definition of capital, thus only including the following intangibles: computerised information (software and databases); mineral exploration; artistic originals and R&D.[8] Investments in data-based information and knowledge (DN) fall under the category of computerised information. The methodology for estimating DN investments, described in full in Chebli, Goodridge et al. (2015) and Goodridge and Haskel (2015a), is consistent with the official methods for estimating investment in computerised information, allowing us to separate that category out into software excluding data-based information and knowledge (SOFT≠DN), data-based information and knowledge already accounted for in official measurement (SOFT=DN), plus additional investments in DN not accounted for in the official measured data (DN^{+}). In this paper we disaggregate those estimates by industry in order to present an industry breakdown of big data employment and investment.

(1) Computerised information: Software and databases

(a) National Accounts measure

Computerised information comprises computer software, both purchased (pre-packaged and custom) and own-account, and computerised databases. This category is already capitalised, as described by Chamberlin, Clayton et al. (2007), and thus we use these data as our starting point. Purchased software data are based on company investment surveys and own-account based on the wage bill of employees in computer software occupations, adjusted downwards for the fraction of time spent on creating new software (as opposed to, say routine maintenance) and then upwards for associated overhead costs.

(b) Data-based information and knowledge (DN)

In Chebli, Goodridge et al. (2015) and Goodridge and Haskel (2015a) we use publically available social media data to identify the occupations where workers in big data reside. We show that of the identified 190,000 big data workers in UK firms, 65% are already counted in the own-account

[8] For work that incorporates a fuller range of intangible capital at the industry-level, see Borgo, Goodridge et al. (2013) or Goodridge, Haskel et al. (2014).

computer software occupations described above, with 35% in other occupations. We therefore use that occupational data to estimate the part of data investment already recorded in the national accounts and also additional investments in data not already recorded. In 2010, of total investment of £5.7bn, we estimate that £4.3bn is already counted within official measures, with £1.4bn of additional investment currently uncounted. Thus we adjust the measured data and effectively separate data for computerised information into components for software and data respectively. See Goodridge and Haskel (2015a) for full details.

In this paper, we break down our estimates of big data employment by occupation and industry, using the self-reported industries of members of our social media dataset, and form industry-level estimates of investment in D and N activity. The data show that the largest investments in DN goods are being made in the "Information and Communication" and "Professional and Administrative (business) services" industries (see Table 3).

(2) R&D, mineral exploration and artistic originals

For business *R&D* we use industry expenditure data derived from the Business Expenditure on R&D survey (BERD). To avoid double counting of R&D and software investment, we subtract R&D spending in "computer and related activities" (SIC 62) since this is already included in the software data.[9] Since BERD also includes physical capital investments we convert those investments into a capital compensation term, using the resulting physical capital stocks for the R&D sector and the user cost relation.[10] The BERD breakdown also includes R&D performed in the R&D services industry. We allocate that spend to purchasing industries using information from the IO tables.

Mineral exploration and production of *artistic originals* (copyright for short) are also already capitalised in the National Accounts. Data for mineral exploration here are simply data for Gross Fixed Capital Formation (GFCF) from the ONS, valued at cost (ONS National Accounts, 2008) and explicitly not included in R&D. Data for copyright are new estimates recently included in the national accounts, based on our own estimates produced with the co-operation of ONS and the Intellectual Property Office (Goodridge 2014). The production of artistic originals covers, "original

[9]BERD gives data on own-account R&D spending. Spending is allocated to the industry within which the product upon which firms are spending belongs. That is, we assume that R&D on say, pharmaceutical products takes place in the pharmaceutical industry. General R&D spending is allocated to professional, scientific and technical services and then allocated to purchasing industries using data from the supply use tables. Thus the BERD data differs from that in the supply use tables, which estimates between-unit transactions of R&D where units can be within the same firm.

[10] Ignoring capital gains, $P^K=P^I(\rho+\delta)$, where P^K is the rental price of physical capital; P^I is the asset price, ρ is the net rate of return to capital and δ is the depreciation rate.

films, sound recordings, manuscripts, tapes etc., on which musical and drama performances, TV and radio programmes, and literary and artistic output are recorded."

3.5. Prices and depreciation

Rates of depreciation and the prices of intangible assets are less well established. This is particularly true of assets in data and the knowledge acquired from data. The R&D literature appears to have settled on a depreciation rate of around 15-20%, and OECD recommend 33% for software. Given the close links between software and data, in terms of both measurement and actual investment activity, as emphasised in Chebli, Goodridge et al. (2015) and Goodridge and Haskel (2015a), we use the same rate of depreciation for data as we do for software (33%). However, as shown in Goodridge and Haskel (2015b), estimates of the contribution of data-based assets to market sector growth are not very sensitive to deprecation rates, unless one assumes much slower rates, in which case intangibles are even more important than suggested here.

On prices, in past work we have made extensive use of the implied GDP deflator. The price of intangibles is an area where very little is known, aside from some very exploratory work by the BEA (e.g. Copeland, Medeiros et al. (2007)) and Corrado, Goodridge et al. (2011)). These papers attempt to derive price deflators for knowledge from the price behaviour of knowledge intensive industries and the productivity of knowledge producing industries. Two observations suggest that using the GDP deflator overstates the price for knowledge, and so understates the impact of knowledge on the economy. First, many knowledge-intensive prices have been falling relative to GDP. Second, the advent of the internet and computers would seem to be a potential large rise in the capability of innovators to innovate, which would again suggest a lowering of the price of knowledge, in contrast to the rise in prices implied by the GDP deflator. Thus use of the GDP deflator could understate the importance of intangible assets.

A number of alternatives are available so we test the robustness of our results using alternative deflators. First, for our baseline we use a wage index built from the reported salaries of the occupations that make up big data employment ($\Delta \ln P_{DN}^{L}$) as estimated in Chebli, Goodridge et al. (2015) and Goodridge and Haskel (2015a). The wage index is a share-weighted index based on the reported salaries of sixteen occupations and their shares in estimated big data employment.

Second, given the strong links between investments in software and data, particularly own-account investments, we use the deflator for own-account software, as used in the ONS Volume Index of Capital Services (VICS). In practice, this is similar to our baseline wage index due to the overlap in

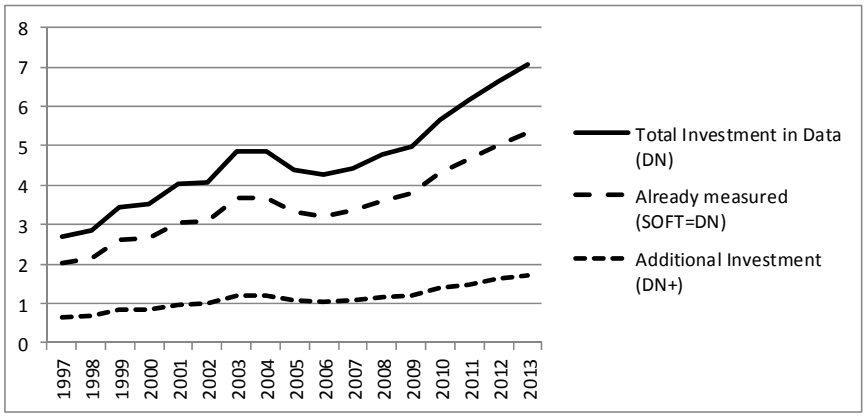
software and data occupations, with an additional productivity adjustment to at least partially account for productivity in the DN sector. Third, as a more aggressive option, we experiment with the fast-falling US price index for pre-packaged software, on the basis that productivity in data investment activities may have been growing similarly fast to that in production of pre-packaged software.

4. Results

4.1. Industry-level employment and investment

This section sets out our results. First, in Figure 2, we present time-series' for total market sector investments in DN, separated into the part already measured within software in the national accounts (SOFT=DN) and the additional component previously unidentified (DN^+), as estimated in Goodridge and Haskel (2015a). In that paper they show that some of the occupations engaged in the building of DN assets are the same as those identified in the measurement of own-account software (SOFT=DN), whilst additional estimates (DN^+) are based on estimates of DN workers in occupations not traditionally considered engaged in the writing of software and the building of databases (e.g. business professionals, economists, statisticians etc.). As outlined in Goodridge and Haskel (2015a), this estimate of market sector employment in 2010 translates to a total estimate of market sector investment in data-based information and knowledge of £5.67bn. Of that, we estimate that £4.29bn (76%) is already measured in the official recording of investment in software and databases. The remaining £1.37bn consists of additional investment not currently recorded in the national accounts.

Figure 2: Time-series of nominal UK market sector investment in data, £bns current prices, 1997-2013



Note to figure: Solid line is total measured investment in data (DN). Dashed line is the component of investment that we estimate as already measured in official gross fixed capital formation (GFCF) (SOFT=DN). Dotted line is additional investment not already recorded in official GFCF (DN^+). Data for the UK market sector, defined as SIC07 sections: A-K, MN, R-U.
Source: Goodridge and Haskel (2015a)

In this paper, we use the social media dataset described in Chebli, Goodridge et al. (2015) to identify the industry of DN workers and thus estimate DN investment, and the intensity of investment in DN goods, in each industry. The dataset also allows us to explore whether the composition of the DN

workforce differs by industry. For instance, in some industries, big data and analytics activities may largely be conducted by ICT occupations typically considered to be engaged in the writing of software. In other industries, big data activities may be more the realm of other types of technical and analytical occupations e.g. economists and statisticians, or marketing and advertising professionals.

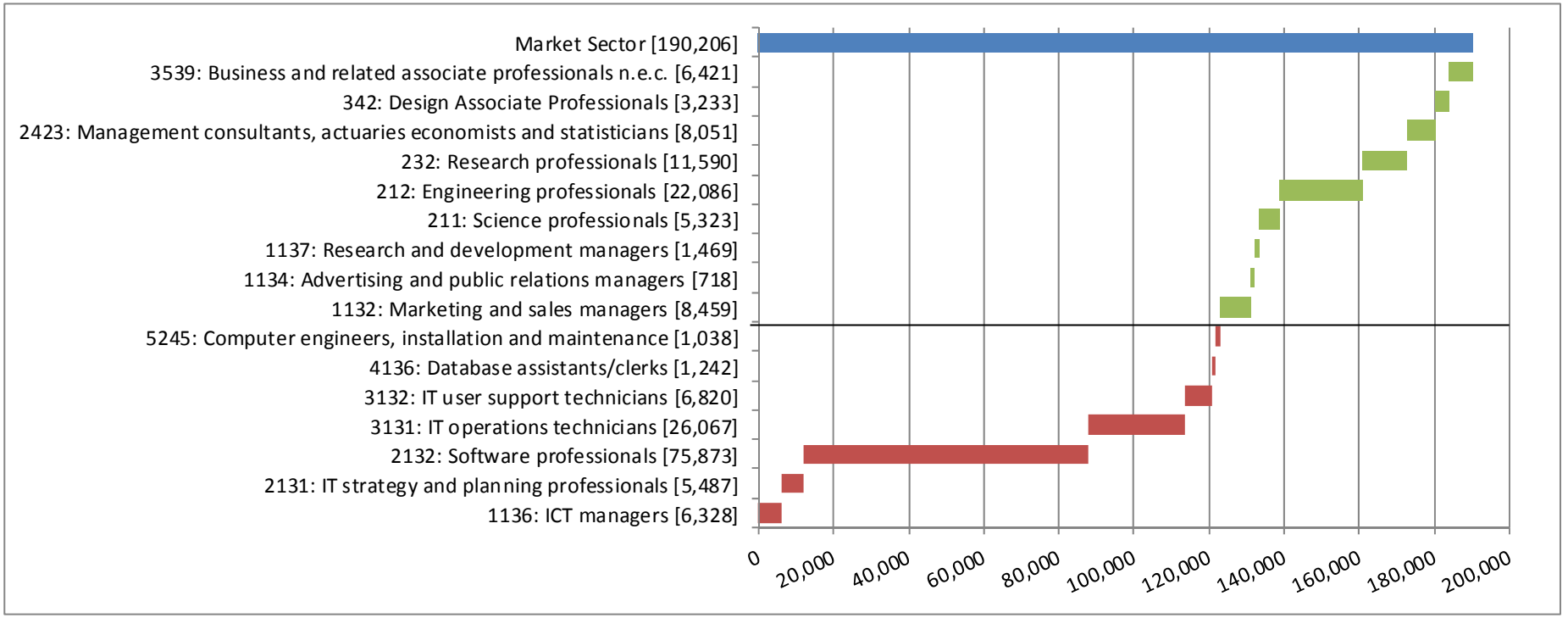
As set out in Chebli, Goodridge et al. (2015), we estimate that in 2010, UK market sector employment of DN workers stood at 190,000. Figure 3 presents how that 190,000 is allocated across various occupational codes in the Standard Occupational Classification (SOC2000). The chart is broken down into two panels. The lower panel shows DN employment in occupations used in the official measurement of own-account software (and databases) which sums to 122,855 workers. The upper panel shows DN employment in other occupations not so measured in the official data, summing to 67,350 workers. As Figure 3 shows, (75,973/190,206=)40% of UK DN employment is accounted for by the single occupation "software professionals" (SOC 2132), and in total, occupations used in the measurement of own-account software account for (122,855/190,206=)65% of DN workers. Of the other occupations in the top panel, we note they are largely composed of: "engineering professionals" (22,086 or 12%); "research professionals" (11,590 or 6%); "marketing and sales managers" (8,459 or 5%); and "management consultants, actuaries, economists and statisticians" (8,051 or 4%).

In which industries is this £5.67bn of investment activity being carried out? That is answered in Table 2 and Table 3. We form estimates of industry investment in DN as follows. First, Goodridge and Haskel (2015a) present estimates of the amount of investment accounted for by each occupation in 2010. Second, using the dataset described in Chebli, Goodridge et al. (2015), we use the self-reported industries of identified DN workers to form industry estimates of DN employment for each occupation, and re-aggregate DN investment to form estimates by industry.[11] Table 2 presents grossed-up estimates[12] of DN employment by industry.

[11] This implicitly assumes that industry shares of investment in 2010 have been constant over time. This is slightly unsatisfactory but since our underlying data are a snapshot for 2010, we have insufficient information to observe changes in industry composition across time.

[12] Data are grossed up using estimates of employment by Standard Occupational Classification (SOC) from the Annual Survey of Hours and Earnings (ASHE) (Office for National Statistics). See Goodridge and Haskel (2015a) for more details.

Figure 3: UK market sector employment of DN workers, by occupation (2010)



Note to table: UK market sector employment of DN workers by occupational code as defined in the 2000 Standard Occupational Classification (SOC 2000). Headings on the left are occupations as defined by the SOC, and the number of DN workers within that occupation, as estimated in Chebli, Goodridge et al. (2015). Top blue bar is total DN employment in the UK market sector. In the lower panel, red bars are DN employment in occupations used to measure own-account investment in computerised information (software and databases). In the upper panel, green bars are DN employment in other occupations outside official measurement of investment in computerised information.

Table 2: UK market sector big data employment, by industry and occupation

		SIC07:	ABDE	C	F	GI	H	J	K	MN	RSTU	TOTAL
	Standard Occupational Classification (2000):		Agriculture, Mining, Utiliies	Manufacturing	Construction	Wholesale and Retail Trade, Accomodation and Food	Transportation and Storage	Information and Communication	Financial Services	Professional and Administrative Services	Recreational and Personal Services	UK Market Sector
Panel 1: ONS software occupations	1136	Information and communication technology managers	80	253	23	264	57	3,273	402	1,769	207	6,328
	2131	IT strategy and planning professionals	154	184	0	307	46	3,166	507	953	169	5,487
	2132	Software professionals	1,750	3,463	214	2,321	1,928	48,844	7,462	8,641	1,250	75,873
	3131	IT operations technicians	357	693	84	1,302	378	14,724	1,764	5,818	945	26,067
	3132	IT user support technicians	196	294	49	589	0	3,876	834	736	245	6,820
	4136	Database assistants/clerks	248	0	0	248	0	497	0	248	0	1,242
	5245	Computer engineers,installation and maintenance	0	234	0	67	0	569	0	134	33	1,038
Panel 2: Other occupations	1132	Marketing and sales managers	43	470	43	684	171	4,785	684	1,367	214	8,459
	1134	Advertising and public relations managers	33	100	0	50	0	184	33	317	0	718
	1137	Research and development managers	0	0	0	122	61	612	122	490	61	1,469
	211	Science professionals	184	1,927	0	275	184	1,560	0	1,010	184	5,323
	212	Engineering professionals	821	3,286	183	548	639	13,598	639	1,643	730	22,086
	232	Research professionals	740	1,480	0	0	740	4,685	493	2,959	493	11,590
	2423	Management consultants, actuaries economists and statisticians	249	274	12	348	162	3,509	1,319	1,991	187	8,051
	342	Design Associate Professionals	0	323	0	162	0	1,778	0	647	323	3,233
	3539	Business and related associate professionals n.e.c.	262	393	0	262	131	1,835	786	2,097	655	6,421
	TOTAL:		5,117	13,375	608	7,550	4,497	107,497	15,047	30,818	5,696	190,205
	Industry % of total:		*2.7%*	*7.0%*	*0.3%*	*4.0%*	*2.4%*	*56.5%*	*7.9%*	*16.2%*	*3.0%*	*100.0%*
Memos:	**Subtotal: "In software"**		2,785	5,122	370	5,099	2,410	74,950	10,970	18,299	2,849	122,855
	Subtotal: Additional "outside software"		2,332	8,253	238	2,451	2,087	32,546	4,077	12,520	2,847	67,350

Note to table: Columns 1 and 2 are occupations that include D and N workers, as estimated in Chebli, Goodridge et al. (2015). Columns 3 to 11 are employment in each occupation/industry estimated using industry/occupation shares as reported from our social media dataset, scaled to market sector data from Chebli, Goodridge et al. (2015). Column 12 is total UK market sector employment in each occupation. Data in rows are by SOC classification. The upper panel shows the SOC classifications used in official measurement of investment in software and databases. The lower panel shows data for other SOCs not used in official measurement of software and databases. Data in the final row are totals and the industry share (%) in big data employment. Data in final memo items are subtotals for the occupations used in official measurement of investment in software and databases ("In software") and those in other occupations (Additional "outside software").

Table 2 (row beneath total) shows that the majority of big data employment is in two industries: Information and Communication (J) and Professional and Administrative Services (MN), where J accounts for 57% of DN employment and MN for 16%. Financial Services and Manufacturing account for 8% and 7% respectively. Thus together just these four industries account for 88% of big data employment. Thus our results are quite consistent with those reported in OECD (2013), who cite that in the market sector, data intensity (average amount of data per organisation) is highest in financial services, communication and media, utilities and discrete manufacturing (Manyika, Chui et al. 2011); and that the estimated number of data management and analytics professionals per 1000 employees produces a similar ranking. Using US data from population surveys, they also report that the five industries with the largest share of database administrators per 10,000 employees in 2012 were: financial services; professional and business services; wholesale and retail trade; manufacturing; and information (OECD 2013).

Before considering the composition of big data employment in each industry, it is worth making a point on the industry shares in Table 2. As noted above, 40% of total DN employment is accounted for by the single occupational code "software professionals" (2132). According to Table 2, (48,844/75,873=)64% of those software professionals engaged in DN activity are employed in the Information and Communication industry. While we would expect a high degree of software professionals in this industry, these estimates of industry DN employment and the industry share are very large. The allocated industry for DN workers is based on self-reported industry information from members of our social media dataset. Thus we note these self-reported industries may differ from the Standard Industrial Classification (SIC) of the firms these employees work for. For instance, a software or ICT professional may identify themselves as working in the "ICT sector" (i.e. Information and Communication) when in practice the firm they are working for is actually allocated to a different SIC. We note therefore the potential for some bias to our estimates, in particular some upward bias to the estimates for industry J.

What about the differing composition of DN employment by industry?[13] In Agriculture, Mining and Utilities (ABDE), DN employment is approximately equally split between the software occupations (mainly "software professionals") (54%) and other occupations (46%), mainly "engineering professionals" (16%) and "research professionals" (14%). In Manufacturing (C) the split is different, whilst 38% of DN employment is in software occupations (26% are "software professionals"), 62% are in other occupations, mainly "engineering professionals" (25%), "science professionals" (14%) and "research professionals" (11%). In Construction (F) and the distributive trades (GI), DN employment is predominantly composed of the software occupations. In Construction (F), the

[13] Appendix Table A1 presents shares of DN employment by occupation/industry.

software occupations account for 61% of DN employment (35% are "software professionals"), whilst other occupations account for 39% (30% are "engineering professionals). In Wholesale and Retail; Accommodation and Food (GI), 68% of DN workers are among the software occupations (31% are "software professionals") and 32% are in other occupations (10% are "marketing/advertising/public relations managers" and 7% are "engineering professionals"). In Transportation and Storage (H) the split is more even: 54% of DN employment is in the software occupations (43% "software professionals") and 46% in other occupations (16% are "research professionals" and 14% are "engineering professionals"). In Information and Communication (J) DN employment is predominantly made up from the software occupations which account for 70% ("software professionals" alone account for 45%), other occupations account for the other 30%, including "engineering professionals" which account for 13%.[14] In Financial Services (K) 73% of DN employment is among the software occupations (as much as 50% are "software professionals"), with 27% in other occupations (9% are "management consultants, economists and statisticians", 5% are "marketing/sales managers" and 5% are "business and related associate professionals n.e.c."). In Professional and Administrative Services (MN) the software occupations again account for the majority of DN employment: 59% are in software occupations and 41% in outside occupations. Of the former, 28% are "software professionals" and, of the latter, 10% are "research professionals" and 7% are "business and related associate professionals n.e.c.". Finally, in Recreational and Personal Services (RSTU) the split is even. Of those recorded under software occupations, 22% are "software professionals". Of those recorded under other occupations, 13% are "engineering professionals", 12% are "business and related associate professionals n.e.c.", 9% are "research professionals" and 6% are "design associate professionals".

Thus the composition of big data employment differs considerably by industry. In manufacturing, and the production industries more generally, many big data workers can be found among the scientific, engineering and research professions as well as occupations considered engaged in the writing of in-house software. In distribution, 10% of workers engaged in DN activity include marketing and advertising managers. In other services, big data workers are more predominantly among the software occupations, but also include significant numbers of consultants, economists and statisticians, and business associate professionals.

What do these industry-level estimates of employment mean for industry investment? This is answered in Table 3, which provides a snapshot of UK industry investment in 2010. The table reads as follows. Each cell provides an estimate of investment for each industry/occupation. Thus reading

[14] In our data, industry and occupation are self-reported. Therefore there may be a classification issue if some of the workers categorised as "engineers" are actually "software engineers" and should actually be allocated to the top panel, in which case the software occupations make up an even larger part of big data employment than estimated here.

across the rows gives an estimate of investment by occupation. Reading down the columns gives an estimate of investment for each industry. Looking at the final row we see that the largest industry investors in DN goods are: first, Information and Communication (J) which accounts for 59% of market sector investment; second, Professional and Administrative Services (MN) which accounts for 15%; third, Financial Services (K), which accounts for 9%; and fourth, Manufacturing (C), which accounts for 6%. Thus together, just these four industries account for 89% of UK market sector investment in DN activity.

Memo items in the final two rows show the part of DN industry investment already measured in the official GFCF data for software and databases (SOFT=DN) and the part not currently captured by official methods (DN^{+}). The data show that the current method of measuring the labour input of occupations in panel 1 is fairly comprehensive for some industries but less so for others. For instance, we find that official methods account for 76% of DN investment in the UK market sector. In Information and Communication the proportion accounted for by official methods is 81% and in Financial Services it is 78%. In contrast, in Manufacturing, official methods account for only 55% of DN investment, 60% in Recreational and Personal Services and 66% in Agriculture, Mining and Utilities.

Table 3: UK big data investment by occupation and industry (2010), £bns, nominal

			ABDE	C	F	GI	H	J	K	MN	RSTU	A-K, MN, R-U
Investment in DN activity, by industry/occupation (2010): £bns			Agriculture, Mining, Utiliies	Manufacturing	Construction	Wholesale and Retail Trade, Accomodation and Food	Transportation and Storage	Information and Communication	Financial Services	Professional and Administrative Services	Recreational and Personal Services	UK Market Sector
Panel 1: ONS software occupations	1136	Information and communication technology managers	0.00	0.00	0.00	0.01	0.00	0.06	0.01	0.03	0.00	0.12
	2131	IT strategy and planning professionals	0.01	0.01	0.00	0.01	0.00	0.12	0.02	0.04	0.01	0.21
	2132	Software professionals	0.08	0.16	0.01	0.11	0.09	2.24	0.34	0.40	0.06	3.48
	3131	IT operations technicians	0.01	0.01	0.00	0.02	0.01	0.23	0.03	0.09	0.01	0.41
	3132	IT user support technicians	0.00	0.00	0.00	0.01	0.00	0.03	0.01	0.01	0.00	0.06
	4136	Database assistants/clerks	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5245	Computer engineers,installation and maintenance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Panel 2: Other occupations	1132	Marketing and sales managers	0.00	0.01	0.00	0.01	0.00	0.07	0.01	0.02	0.00	0.12
	1134	Advertising and public relations managers	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	1137	Research and development managers	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.02
	211	Science professionals	0.00	0.04	0.00	0.01	0.00	0.03	0.00	0.02	0.00	0.10
	212	Engineering professionals	0.01	0.06	0.00	0.01	0.01	0.25	0.01	0.03	0.01	0.40
	232	Research professionals	0.01	0.02	0.00	0.00	0.01	0.07	0.01	0.04	0.01	0.17
	2423	Management consultants, actuaries economists and statisticians	0.01	0.01	0.00	0.02	0.01	0.16	0.06	0.09	0.01	0.37
	342	Design Associate Professionals	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02
	3539	Business and related associate professionals n.e.c.	0.01	0.01	0.00	0.01	0.00	0.05	0.02	0.06	0.02	0.17
	Total Industry Investment (DN):		0.14	0.33	0.02	0.20	0.14	3.34	0.52	0.84	0.14	5.67
Memos:	**Subtotal: "In software" (SOFT=DN)**		0.10	0.19	0.01	0.15	0.10	2.70	0.41	0.57	0.08	4.29
	Subtotal: Additional "outside software" (DN+)		0.05	0.15	0.00	0.05	0.04	0.64	0.11	0.27	0.06	1.37

Note to table: Estimated UK investment in current prices (£bns) by occupation and industry. Data in rows are for occupations and data in columns for industries. Final column is data for aggregate UK market sector. Data in final row is industry investment in DN activity. Memo items show the part already measured in official estimates of investment in software and databases (SOFT=DN) and the part not currently measured (DN^{+}).

Above we have reported absolute values for DN investment by industry. To understand more about the relative intensity of DN investment by industry, Table 4 presents: a) the shares of intangible investment accounted for by each intangible asset category capitalised in the SNA; and b) the share of industry value-added accounted for by each of those intangibles. That is, we compare industry shares of intangible investment for software,[15] data-based assets (DN), R&D, mineral exploration and the creation of artistic originals, and also the intensity of investment in each asset in value-added.

The top panel of Table 4 presents shares of total industry intangible investment by asset in 2011, giving some idea of the relative importance of different intangible assets in each industry. First, looking at R&D, row 1, as we would expect this is most important in Manufacturing, where it accounts for 77% of industry intangible investment. R&D is also important in Agriculture, Mining and Utilities, where it accounts for 18% of industry intangible investment, Professional and Administrative Services, where it accounts for 13%, and Construction where it accounts for 12%. Results for mineral exploration (row 2) and the creation of artistic originals (row 3) are as expected. That is, mineral exploration is important in Agriculture, Mining and Utilities, where it accounts for 28% of industry intangible investment, and artistic originals are important in Information and Communication (i.e. publishers) where they account for 45% of industry intangible investment, and Recreational and Personal Services (i.e. artists) where they account for 35% of industry intangible investment.

Next we consider the SNA asset category of computerised information (software and databases). Looking at software (SOFT≠DN), row 4, we see that this asset accounts for the majority of intangible investment in most industries. Software is particularly important in the distributive trades: in Wholesale and Retail, Accommodation and Food it accounts for 89%, and in Transportation and Storage, for 83%. In Financial Services and also Construction, software accounts for 84% of industry intangible investment.

On data-based assets (DN), row 5, in Information and Communication they account for 31% of industry intangible investment. Other large shares occur in Professional and Administrative Services where investments in DN account for 19% of intangible investment, Financial Services where they account for 14%, and Transportation and Storage where they account for 12%. The memo item in the final row presents shares for the combined SNA asset category, computerised information (software and databases), which accounts for as much as 99% of industry intangible investment in Financial Services, 94% in Transportation and Storage, and 93% in Wholesale and Retail; Accommodation and Food.

[15] Estimates for software are adjusted in the sense that the part that is estimated to represent investments in data is subtracted from software so that software is SOFT≠DN, and added to DN, so that DN=(SOFT=DN)+(DN^{+}).

Thus, in general, the shares of intangible investment accounted for by DN are larger in services, with the production industries tending to invest in more "traditional" forms of intangible activity, such as investments in R&D or Mineral Exploration.

The bottom panel of Table 4 presents investment as a share of value-added for each asset. Again we note the importance of R&D in Manufacturing, where it accounts for 8.6% of industry value-added, but R&D is much less important outside Manufacturing. One exception is Information and Communication where R&D is 1.2% of value-added. On software (SOFT≠DN), this accounts for around 2 to 3% of value-added in most industries, except Construction where it is just 0.4% of industry value-added. This is despite software accounting for 21% of industry intangible investment (top panel), with total intangible investment in this industry being small.

On DN assets, we find that by far the most intensive investor in DN is the Information and Communication industry, where DN investments are as much as 4% of industry value-added. In general, investment intensity in DN assets tends to be higher in services, at 0.6% in Professional and Administrative Services, 0.5% in Financial Services, and 0.3% in both Recreational and Personal Services and Transportation and Storage. Considering the combined asset category of computerised information (memo item), investment intensity is 6% in Information and Communication, 3.3% in Financial Services, and in the range of 2 to 3% in all other industries with the exception of Construction where it is just 0.5% of industry value-added.

Table 4: Industry intangible investment shares and intensities

	ABDE	C	F	GI	H	J	K	MN	RSTU
Shares of industry intangible investment (2011):	**Agriculture, Mining, Utiliies**	**Manufacturing**	**Construction**	**Wholesale and Retail Trade, Accomodation and Food**	**Transportation and Storage**	**Information and Communication**	**Financial Services**	**Professional and Administrative Services**	**Recreational and Personal Services**
1. R&D	18%	77%	12%	7%	6%	9%	1%	13%	1%
2. Mineral Exploration	28%	0%	0%	0%	0%	0%	0%	0%	0%
3. Artistic Originals	0%	0%	0%	0%	0%	45%	0%	0%	35%
4. Software (SOFT≠DN)	48%	21%	84%	89%	83%	15%	84%	68%	55%
5. Data-based assets (DN)	6%	2%	4%	5%	12%	31%	14%	19%	9%
Memo: "software and databases" (=4+5)	*54%*	*23%*	*88%*	*93%*	*94%*	*46%*	*99%*	*87%*	*64%*
Shares of industry value-added (2011):									
1. R&D	0.6%	8.6%	0.1%	0.2%	0.1%	1.2%	0.0%	0.4%	0.0%
2. Mineral Exploration	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
3. Artistic Originals	0.0%	0.0%	0.0%	0.0%	0.0%	5.9%	0.0%	0.0%	1.3%
4. Software (SOFT≠DN)	1.8%	2.3%	0.4%	2.1%	1.7%	2.0%	2.8%	2.1%	2.0%
5. Data-based assets (DN)	0.2%	0.2%	0.0%	0.1%	0.3%	4.0%	0.5%	0.6%	0.3%
Memo: "software and databases" (=4+5)	*2.0%*	*2.5%*	*0.5%*	*2.2%*	*2.0%*	*6.0%*	*3.3%*	*2.6%*	*2.4%*

Notes to table: Top panel are shares of intangible investment for each intangible asset capitalised in the SNA, including additional investments in data-based information and knowledge assets (DN) as estimated in this paper. Bottom panel are investment shares of industry value-added for each intangible asset capitalised in the SNA, again including additional investments in DN. Note therefore that investments in DN are the sum of those already measured in the official data (SOFT=DN) plus additional investments (DN^{+}). Memo items combine estimates for software and DN to re-create estimates for the asset category "computerised information" (software and databases) as used in the national accounts.

4.2. DN investment: issues around own-account .vs purchased and performance .vs. ownership

Our estimation of DN investment is a cost-based approach that is conceptually and practically analogous to official methods of measuring investment in own-account software and R&D. That is, we estimate investment in DN assets by observing or estimating upstream factor payments using a sum of costs approach, as in equation (1).

However, the official methods for measuring own-account software and R&D include an important adjustment that we are unable to implement in our data due to a lack of information. In measuring investment in software, the ONS estimate two components. The first is purchased, with data on investment in purchased software gathered via capital expenditure (CAPEX) surveys. The second is own-account, estimated by observing the labour input of occupations engaged in the writing of own-account software and adjusting for time-use and other factors (use of capital and materials), the same method as used in this paper. However, in the measurement of own-account software, the ONS exclude workers in those occupations working in the software industry (SIC 62) itself, assuming that the software they produce is produced for general sale and so already captured in the purchased data.

A similar adjustment is made in the measurement of R&D. Official estimates of investment in R&D are also produced using a sum of costs approach, with the data collected via official R&D expenditure surveys. Those surveys include questions which ask R&D performers on the ownership and destination of the knowledge produced. If the knowledge is intended for own final use, then the performer is considered the owner of the R&D knowledge asset, and investment is allocated to the performing industry (as with own-account software and the own-account estimation of DN conducted in this paper). However, if the knowledge is to be transferred and owned by some other domestic or non-domestic organisation, then the R&D is allocated to "R&D market output", with investment allocated to the acquiring domestic industry, or to exports if the destination of the knowledge is outside the UK.

However, as explained in Goodridge and Haskel (2015a; 2015b), we do not have available data on purchased investments in DN, or the ownership/destination of DN assets, to make a similar adjustment. We have noted that our estimates show that (3.34/5.67=)59% of DN production occurs in the Information and Communication (SIC J) industry. However, as with software and R&D, some of this production will be market output produced for sale to either the wider market sector or for export. Thus there will be an upward bias to our estimates for that industry, with investment allocated to the Information and Communication industry including that market output produced for general sale. Similarly, our estimates for industries outside J will incorporate a downward bias, since they do not

include the purchases of DN from the market output of DN assets produced in the Information and Communication (J) industry. Some DN assets will also be imported, which our data also do not capture.

Therefore, our method implicitly allocates ownership on the basis of performance and does not allow for the purchase/sale of DN assets between industries. Therefore, whilst we feel we adequately the scale of DN performance at the aggregate level, there is likely some bias to our allocation at the industry level due to our inability to identify final ownership/destination of DN assets. To get some idea of the magnitude of the potential bias, we have the following information. Of R&D performed by UK market producers,74% is produced for own final use (i.e. own-account investment) and the remaining 26% is market output produced for sale (i.e. purchased by other industries).[16] Specifically on DN, we do also have the estimate from e-skills UK (2013b) that of big data (DN in our nomenclature) activities, 89% are conducted in-house, and 11% are outsourced/purchased. This implies that of our total estimate of investment of £5.7bn, approximately (0.89*£5.7bn=)£5.1bn is own-account investment and £0.6bn is "market output" sold to, and purchased by, other industries. If all DN assets produced for sale are produced in the Information and Communication (J) industry, this suggests our estimate for investment in this industry (£3.3bn, see Table 3) may more accurately be in the order of £2.7bn, with the £0.6bn re-allocated to outside (purchasing) industries on a pro rata basis. Table 5 presents a re-allocation on this basis.

Thus if the estimate from e-skills UK, that only 11% of DN activity is outsourced, is accurate, the bias to our industry breakdown is small. In the growth-accounting estimates that follow we work with the unadjusted investment estimates presented in Table 3 and Table 5, but note that inability to identify purchases of DN assets from official or unofficial data has implications for measurement of investment at the industry-level. However the scale of the adjustment implied by the e-skills estimate means that the impact on our growth-accounting results is minimal.

[16] We are grateful to David Lewis and Foyzunnesa Khatun of the ONS for providing this information.

Table 5: Industry investment re-estimated using assumption on percentage destined for final sale (2010)

	ABDE	C	F	GI	H	J	K	MN	RSTU	A-K, MN, R-U
	Agriculture, Mining, Utiliies	Manufacturing	Construction	Wholesale and Retail Trade, Accomodation and Food	Transportation and Storage	Information and Communication	Financial Services	Professional and Administrative Services	Recreational and Personal Services	UK Market Sector
Total Industry Investment (DN), unadjusted:	0.14	0.33	0.02	0.20	0.14	3.34	0.52	0.84	0.14	5.67
Total Industry Investment (DN), adjusted for estimated purchases:	0.18	0.42	0.02	0.25	0.17	2.71	0.66	1.06	0.18	5.67

Note to table: Estimates of current price investment in DN assets by industry in 2010. Top row are unadjusted estimates, as presented in Table 3. Bottom row adjusts these estimates based on information from e-skills UK (2013b) that of big data (DN in our nomenclature) activities, 89% are conducted in-house, and 11% are outsourced/purchased. Thus in the bottom row we assume that all purchases of DN assets are purchased from industry J, so that investment in that industry is reduced by (0.11*£5.67)=£0.63bn, with that £0.63bn re-allocated to other industries on a pro rata basis.

4.3.Industry-level growth-accounting

In this section we use our industry-level estimates of investment to conduct an industry-level growth-accounting analysis and estimate the contribution of DN capital deepening to productivity growth in each industry. Doing so of course requires estimates of the price of data-based capital. Therefore as our deflator we use a wage index for DN workers, as outlined in Goodridge and Haskel (2015a) and Goodridge and Haskel (2015b), although we do test robustness to alternative deflators. We therefore implicitly assume that the price of data-based assets is the same across industries.

Table 6 sets out the results of our growth-accounting exercise for 2000-11. The table reads as follows. Column 1 is industry growth in value-added per hour. Column 2 is the contribution of labour services per hour (i.e. labour composition), namely growth in labour services per hour times the labour share in GVA. Column 3 is the contribution of computer capital services per hour, namely growth in computer capital services per hour times the income share in GVA. Column 4 is the contribution of other tangible capital services, buildings, plant & machinery and vehicles, per hour. Column 5 is the contribution of R&D capital services per hour. Column 6 is the contribution of other intangible capital services, mineral exploration and artistic originals, per hour. Column 7 is the contribution of software capital services per hour.[17] Column 8 is the contribution from data-based information and knowledge (DN) capital services per hour. Finally, column 9 is ΔlnTFP which is column 1 less the sum of columns 2 to 8.

Before looking at the contribution of DN capital deepening, a few points to note from Table 6. From column 1, the strongest performers in terms of labour productivity growth are: Manufacturing (3.41% pa); Information and Communication (3.06% pa); Financial Services (2.45% pa); and Professional and Administrative Services (2.01% pa). From column 2 we see that these same four industries also have the four largest contributions from labour composition, which can be interpreted as the contribution of labour "quality" or "skills". The contribution of labour composition is particularly high in Financial Services (0.82% pa) and Information and Communication (0.69% pa).

These four industries also have something else in common. From column 5 we see that R&D makes a strong contribution to labour productivity growth in Manufacturing (0.42% pa). Column 6 shows that the creation of artistic originals makes a large contribution to growth in Information and Communication (0.21% pa). Column 7 shows that the largest contributions from software are in:

[17] Data for software are adjusted with the part estimated as representing investments in data (SOFT=DN) excluded and included in column 8 (DN).

Table 6: Industry-level growth accounting, 2000-11 (all data in mean % pa)

Decomposition of growth in labour productivity (2000-11)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Memo
Industry	DlnV/H	sDln(L/H)	sDln(K/H) comp	sDln(K/H) othtan	sDln(R/H) rd	sDln(R/H) mincop	sDln(R/H) soft	sDln(R/H) DN	DlnTFP	VA share: VAi/ΣVA
Agriculture, Mining, Utilities	-3.57	0.19	0.11	0.53	-0.02	-0.15	0.10	0.00	-4.34	0.07
Manufacturing	3.41	0.51	0.08	0.69	0.42	0.00	0.13	0.01	1.57	0.18
Construction	1.18	0.06	0.06	1.72	0.00	0.00	0.03	-0.01	-0.68	0.09
Wholesale and Retail Trade, Accommodation and Food	1.64	0.41	0.08	1.22	0.01	0.00	0.10	0.00	-0.17	0.19
Transportation and Storage	-0.06	0.22	0.08	1.00	0.00	0.00	0.08	0.00	-1.45	0.06
Information and Communication	3.06	0.69	0.05	0.39	0.06	0.21	0.17	0.09	1.41	0.09
Financial Services	2.45	0.82	0.24	0.88	0.00	0.00	0.29	0.02	0.21	0.11
Professional and Administrative Services	2.01	0.51	0.03	0.86	0.02	0.00	0.13	0.01	0.45	0.16
Recreational and Personal Services	0.01	0.43	0.21	2.00	0.00	0.05	0.12	0.00	-2.79	0.04

Note to table: Decomposition of industry labour productivity growth, based on an industry value-added production function. Growth rates calculated as changes in natural logs. Data are averages for the period 2000-11. Column 1 is industry growth in value-added per hour. Column 2 is the contribution of labour services per hour (i.e. labour composition), namely growth in labour services per hour times the labour share in GVA. Column 3 is the contribution of computer capital services capital deepening, namely growth in computer capital services per hour times the income share in GVA. Column 4 is the contribution of other tangible capital deepening: buildings, plant & machinery and vehicles. Column 5 is the contribution of R&D capital deepening. Column 6 is the contribution of other intangible capital deepening: mineral exploration and artistic originals. Column 7 is the contribution of software capital deepening. Data for software are adjusted with the part estimated as representing investments in data excluded and included in column 8 (DN). Column 8 presents the contribution from data-based information and knowledge (DN) capital deepening. Finally, column 9 is ΔlnTFP which is column 1 less the sum of columns 2 to 8.

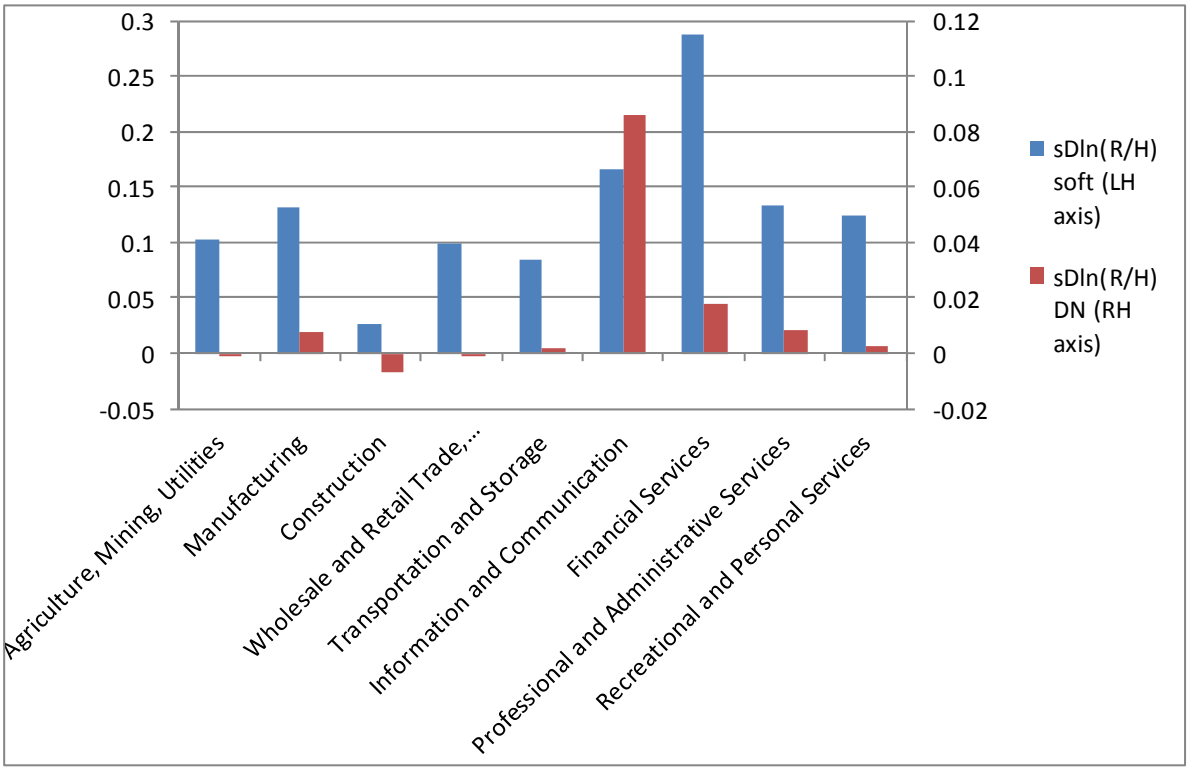
Financial Services (0.29% pa), Information and Communication (0.17% pa); Professional and Administrative Services (0.13% pa); and Manufacturing (0.13% pa). Similarly column 9 shows that the contribution of capital deepening in data-based information and knowledge (DN) is largest in: Information and Communication (0.09% pa); Financial Services (0.02% pa); Professional and Administrative Services (0.01% pa); and Manufacturing (0.01% pa). These patterns and correlations between the noted contributions above are not surprising – it is widely accepted that it is human and knowledge capital that drives growth in modern, advanced economies.

The memo item presented in Table 6 is the industry share of nominal value-added. As set out in a previous section, we can use this share with the industry contributions to estimate the contribution of each industry to the total. Thus we find, in 2000-11, that DN capital deepening contributed 0.01% pa of UK market sector labour productivity growth, and of that, 67% was contributed by the Information and Communication industry, 16% from Financial Services, 11% from Manufacturing and 11% from Professional and Administrative Services.[18]

Therefore we find that the contribution of 'big data' is largest in the intangible- and skill-intensive sectors of the UK market economy. It is currently estimated as negligible in other industries, although that may change in coming years, with the contribution of data-based assets projected to grow over the next decade (see Goodridge and Haskel (2015b)). The following chart emphasises that the contribution of data-based assets is highest in industries that also generate large contributions from software.

[18] Contributions sum to more than 100% as the contribution of DN is estimated as negative in some industries due to negative growth rates in DN capital services in some industries.

Figure 4: Industry contributions of software and DN capital deepening, 2000-11



4.4.Industry-level growth-accounting: Robustness checks

It might be asked, how robust are these estimates to different assumptions on the price of DN assets? Appendix Table A2 provides growth-accounting estimates constructed using various different deflators for DN investments. Baseline estimates presented above were based on an estimated wage index for DN workers (based on the occupations in Table 3), constructed from ASHE microdata (Office for National Statistics). Given the links between investments in DN and software, in both practice and measurement, we test the impact of first, using the UK own-account software deflator, and second using the US pre-packaged software deflators. The former is also a wage index based on the occupations in panel 1 of Table 3 (the wage element of the two deflators is very similar) but with an important conceptual difference in that it adjusts for assumed productivity growth in own-account software creation.[19] Thus real DN investment and capital services typically grow slightly faster when using this deflator.

The second price index tested is the US pre-packaged software deflator, which is a hedonic price index produced by the BEA. Thus we are testing here the contribution of DN if we assume productivity growth in DN production is as rapid as that in the production of pre-packaged software.

[19] Using data for UK labour productivity growth in the wider service sector.

Results using each of these deflators are presented in panel 1 and 2 respectively in Appendix Table A2. Using the own-account software deflator raises the mean contribution of DN capital deepening in the Information and Communication (J) industry from 0.09% pa in the baseline estimates to 0.16% pa using the software deflator, almost double. The total market sector contribution is raised from 0.01% pa to 0.03% pa. The pattern of industry contributions to the total is similar, with Information and Communication contributing 59% of the 0.03% pa, Financial Services contributes 15%, Professional and Administrative Services 14% and Manufacturing, 8%.

In panel 2 we test the much more aggressive US pre-packaged software deflator, implicitly assuming very fast TFP in the upstream D and N sectors. Now the total market sector contribution is raised to 0.05% pa, of which, Information and Communication contributes 55%, Professional and Administrative Services 15%, Financial Services 14% and Manufacturing 7%.

5. Conclusions

This paper presents estimates industry-level estimates of employment and investment in big data activity, building on previous work reported in Chebli, Goodridge et al. (2015) and Goodridge and Haskel (2015a). We also conduct growth-accounting to estimate the contribution of data-based information and knowledge to growth in industry value-added. Our main findings are as follows. First, we find that the composition of big data employment and investment differs by industry. In Manufacturing and other production industries, big data activity is conducted by workers in scientific, engineering and research occupations as well as among occupations typically considered engaged in the writing of software. In services, big data workers are predominantly among software and ICT occupations, but also include significant numbers of consultants, economists and statisticians, and business associate professionals. Second, we find that the four industries: Information and Communication, Professional and Administrative Services, Financial Services and Manufacturing account for 88% of big data employment. Third, these same four industries carry out 89% of UK market sector investment in data-based assets. Note, we estimate that total UK market sector investment in data-based assets in 2010 was £5.7bn, growing to £6.6bn in 2012. In the same years, UK market sector investment in R&D is estimated at £14.8bn in 2010, and £15.5bn in 2012. Thus we estimate UK investments in data-based assets as around 40% of those in R&D, and growing. Fourth, of the total contribution of data-based assets to UK growth, we find that it is fully accounted for by these four industries.

References

Arrow, K. J. (1984). The economics of information, Harvard University Press.

Basu, S., J. G. Fernald, et al. (2001). Productivity growth in the 1990s: technology, utilization, or adjustment? Carnegie-Rochester conference series on public policy, Elsevier.

Boisot, M. and A. Canals (2004). "Data, information and knowledge: have we got it right?" Journal of Evolutionary Economics **14**(1): 43-67.

Borgo, M. D., P. Goodridge, et al. (2013). "Productivity and Growth in UK Industries: An Intangible Investment Approach*." Oxford Bulletin of Economics and Statistics **75**(6): 806-834.

CEBR (2013). "Data on the Balance Sheet: Report for SAS, June 2013."

Chamberlin, G., T. Clayton, et al. (2007). "New measures of UK private sector software investment." Economic and Labour Market Review **1**(5): 17-28.

Chebli, O., P. Goodridge, et al. (2015). "Measuring activity in Big Data: New estimates of Big Data employment in the UK market sector."

Copeland, A. M., G. W. Medeiros, et al. (2007). "Estimating Prices for R&D Investment in the 2007 R&D Satellite Account." BEA Papers.

Corrado, C., P. Goodridge, et al. (2011). "Constructing a Price Deflator for R&D: Calculating the Price of Knowledge Investments as a Residual."

e-skills UK (2013b). "Big Data Analytics: Adoption and Employment Trends, 2012-2017."

Franklin, M. and P. Mistry (2013). "Quality-adjusted labour input: estimates to 2011 and First Estimates to 2012." ONS, available at http://www. ons. gov. uk/ons/dcp171766_317119. pdf.

Fransman, M. (1998). "Information, knowledge, vision and theories of the firm." Technology, organization, and competitiveness: perspectives on industrial and corporate change: 147-191.

Goodridge, P. (2014). "Film, Television & Radio, Books, Music and Art: Estimating UK Investment in Artistic Originals." Imperial College Discussion Paper **2014/2**(March 2014).

Goodridge, P. and J. Haskel (2015a). "How much is UK business investing in Big Data?" Imperial College Discussion Paper.

Goodridge, P. and J. Haskel (2015b). "How does big data affect GDP? Theory and evidence for the UK." Imperial College Discussion Paper.

Goodridge, P., J. Haskel, et al. (2014). "UK Innovation Index 2014."

Hall, R. E. and D. W. Jorgenson (1967). "Tax policy and investment behavior." The American economic review **57**(3): 391-414.

Jorgenson, D. W. and Z. Griliches (1967). "The explanation of productivity change." The Review of Economic Studies **34**(3): 249-283.

Manyika, J., M. Chui, et al. (2011). "Big data: The next frontier for innovation, competition, and productivity."

Mayer-Schönberger, V. and K. Cukier (2013). Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt.

OECD (2001b). "Measuring productivity: measurement of aggregate and industry-level productivity growth." Statistics Portal: 1-156.

OECD (2013). Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data". OECD Digital Economy Papers, OECD Publishing.

OECD (2013). Supporting Investment in Knowledge Capital, Growth and Innovation. O. Publishing.

Office for National Statistics "Annual Survey of Hours and Earnings, 1997-2011: Secure Data Service Access [computer file]. Colchester, Essex: UK Data Archive [distributor], April 2013. SN:6689."

Oulton, N. and S. Srinivasan (2003). "Capital stocks, capital services, and depreciation: an integrated framework." Bank of England Working Paper No. 192.

Oulton, N. and G. Wallis (2014). "An Integrated Set of Estimates of Capital Stocks and Services for the United Kingdom: 1950-2013." Paper for the 33rd General Conference of the IARIW, Rotterdam, August 24-30, 2014.

Pakes, A. and M. Schankerman (1984). The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources, University of Chicago Press.

Romer, P. (1991). Endogenous technological change, National Bureau of Economic Research.

Shapiro, C. and H. Varian (1998). Information rules, Harvard Business Press.

Wallis, G. (2012b). "Tax Incentives and Investment in the UK." Forthcoming and available on request.

Appendix 1

Appendix Table A1: Shares of big data employment by occupation/industry

		SIC07:	ABDE	C	F	GI	H	J	K	MN	RSTU	TOTAL
		Standard Occupational Classification (2000):	Agriculture, Mining, Utiliies	Manufacturing	Construction	Wholesale and Retail Trade, Accomodation and Food	Transportation and Storage	Information and Communication	Financial Services	Professional and Administrative Services	Recreational and Personal Services	UK Market Sector
Panel 1: ONS software occupations	1136	Information and communication technology managers	2%	2%	4%	3%	1%	3%	3%	6%	4%	3%
	2131	IT strategy and planning professionals	3%	1%	0%	4%	1%	3%	3%	3%	3%	3%
	2132	Software professionals	34%	26%	35%	31%	43%	45%	50%	28%	22%	40%
	3131	IT operations technicians	7%	5%	14%	17%	8%	14%	12%	19%	17%	14%
	3132	IT user support technicians	4%	2%	8%	8%	0%	4%	6%	2%	4%	4%
	4136	Database assistants/clerks	5%	0%	0%	3%	0%	0%	0%	1%	0%	1%
	5245	Computer engineers,installation and maintenance	0%	2%	0%	1%	0%	1%	0%	0%	1%	1%
Panel 2: Other occupations	1132	Marketing and sales managers	1%	4%	7%	9%	4%	4%	5%	4%	4%	4%
	1134	Advertising and public relations managers	1%	1%	0%	1%	0%	0%	0%	1%	0%	0%
	1137	Research and development managers	0%	0%	0%	2%	1%	1%	1%	2%	1%	1%
	211	Science professionals	4%	14%	0%	4%	4%	1%	0%	3%	3%	3%
	212	Engineering professionals	16%	25%	30%	7%	14%	13%	4%	5%	13%	12%
	232	Research professionals	14%	11%	0%	0%	16%	4%	3%	10%	9%	6%
	2423	Management consultants, actuaries economists and statisticians	5%	2%	2%	5%	4%	3%	9%	6%	3%	4%
	342	Design Associate Professionals	0%	2%	0%	2%	0%	2%	0%	2%	6%	2%
	3539	Business and related associate professionals n.e.c.	5%	3%	0%	3%	3%	2%	5%	7%	12%	3%
	TOTAL:		100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Memos:	**Subtotal: "In software"**		54%	38%	61%	68%	54%	70%	73%	59%	50%	65%
	Subtotal: Additional "outside software"		46%	62%	39%	32%	46%	30%	27%	41%	50%	35%

Note to table: Shares of estimated big data employment in 2010 by occupation/industry. Shares sum vertically to 100% for each industry. Final column shows shares for the aggregate UK market sector.

Appendix Table A2: Industry growth-accounting, 2000-11, using alternative deflators for DN assets

Panel 1: UK own-account software deflator (incl prod adj) for DN	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Memo
Industry	DlnV/H	sDln(L/H)	sDln(K/H) comp	sDln(K/H) othtan	sDln(R/H) rd	sDln(R/H) mincop	sDln(R/H) soft	sDln(R/H) DN	DlnTFP	VA share: VAi/ΣVA
Agriculture, Mining, Utilities	-3.57	0.19	0.11	0.53	-0.02	-0.15	0.10	0.00	-4.34	0.07
Manufacturing	3.42	0.51	0.08	0.69	0.42	0.00	0.13	0.01	1.57	0.18
Construction	1.18	0.06	0.06	1.72	0.00	0.00	0.03	-0.01	-0.68	0.09
Wholesale and Retail Trade, Accommodation and Food	1.64	0.41	0.08	1.22	0.01	0.00	0.10	0.00	-0.17	0.19
Transportation and Storage	-0.06	0.22	0.08	1.00	0.00	0.00	0.08	0.01	-1.46	0.06
Information and Communication	3.07	0.69	0.05	0.39	0.06	0.21	0.17	0.16	1.35	0.09
Financial Services	2.46	0.82	0.24	0.88	0.00	0.00	0.29	0.03	0.20	0.11
Professional and Administrative Services	2.02	0.51	0.03	0.86	0.02	0.00	0.13	0.02	0.44	0.16
Recreational and Personal Services	0.01	0.43	0.21	2.00	0.00	0.05	0.12	0.01	-2.80	0.04
Panel 2: US pre-packaged software deflator for DN										
Agriculture, Mining, Utilities	-3.57	0.19	0.11	0.53	-0.02	-0.15	0.10	0.01	-4.35	0.07
Manufacturing	3.42	0.51	0.08	0.69	0.42	0.00	0.13	0.02	1.56	0.18
Construction	1.18	0.06	0.06	1.72	0.00	0.00	0.03	0.00	-0.69	0.09
Wholesale and Retail Trade, Accommodation and Food	1.64	0.41	0.08	1.22	0.01	0.00	0.10	0.01	-0.18	0.19
Transportation and Storage	-0.06	0.22	0.08	1.00	0.00	0.00	0.08	0.01	-1.46	0.06
Information and Communication	3.10	0.69	0.05	0.39	0.06	0.21	0.17	0.28	1.26	0.09
Financial Services	2.46	0.82	0.24	0.88	0.00	0.00	0.29	0.06	0.17	0.11
Professional and Administrative Services	2.02	0.51	0.03	0.86	0.02	0.00	0.13	0.04	0.42	0.16
Recreational and Personal Services	0.02	0.43	0.21	2.00	0.00	0.05	0.12	0.02	-2.80	0.04

Note to table: Decomposition of industry labour productivity growth, based on an industry value-added production function. Growth rates calculated as changes in natural logs. Data are averages for the period 2000-11. Panel 1 deflates DN investment with the own-account software deflator (including a productivity adjustment) and panel 2 deflates DN investment using the US pre-packaged software deflator. Column 1 is industry growth in value-added per hour. Column 2 is the contribution of labour services per hour (i.e. labour composition), namely growth in labour services per hour times the labour share in GVA. Column 3 is the contribution of computer capital services capital deepening, namely growth in computer capital services per hour times the income share in GVA. Column 4 is the contribution of other tangible capital deepening: buildings, plant & machinery and vehicles. Column 5 is the contribution of R&D capital deepening. Column 6 is the contribution of other intangible capital deepening: mineral exploration and artistic originals. Column 7 is the contribution of software capital deepening. Data for software are adjusted with the part estimated as representing investments in data excluded and included in column 8 (DN). Column 8 presents the contribution from data-based information and knowledge (DN) capital deepening. Finally, column 9 is ΔlnTFP which is column 1 less the sum of columns 2 to 8.

This paper has been produced by the Department of Management at Imperial College Business School

Copyright © the authors 2014
All rights reserved

ISSN: 1744-6783

Imperial College Business School
Tanaka Building
South Kensington Campus
London SW7 2AZ
United Kingdom

T: +44 (0)20 7589 5111
F: +44 (0)20 7594 9184

www.imperial.ac.uk/business-school