How much is UK business investing in big data?

Peter Goodridge, Jonathan Haskel

Discussion Paper 2015/05
July 2015
How much is UK business investing in Big Data?*

Peter Goodridge
Imperial College Business School

Jonathan Haskel
Imperial College Business School; CEPR and IZA

July 2015

Abstract

We present an economic approach to measuring investment in Big Data. We present a conceptual framework to understand and measure the production of transformed data and data-based knowledge. We use this framework to understand how current official datasets and concepts used by Statistics Offices might already measure Big Data investment in GDP, or might miss it. We also set out how unofficial data sources might be used to measure investment and estimate that in 2010, UK market sector investment in data-based assets was $5.7bn. Of that, we estimate that £4.3bn is already counted in official measurement of investment in software and databases, and that £1.4bn is previously unidentified investment.

* 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 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. This work uses research datasets which may not exactly reproduce National Statistics aggregates. All errors are of course our own.
1 Introduction

This paper sets out a conceptual framework for measuring investment in and around ‘Big Data’, or more broadly, data and data analytics. In it we set out the case for treating data, and the knowledge gleaned from data analytics, as long-lived assets that contribute to final production. As part of that, we present a framework for measuring investment in data-building and data analytics, and show how those estimates and concepts can be integrated into wider national accounting and productivity frameworks, in a consistent way that avoids double counting. In doing so, we rely heavily on estimates of ‘big data employment’ as estimated and presented in Chebli, Goodridge et al. (2015). Together these papers form a first step in estimating the contribution of data to the UK economy.¹

Since there has been an explosion of work on ‘Big Data’ one might wonder what, if any, this current paper can contribute. First, it is perhaps still unclear what constitutes Big Data exactly, with discussion over Big versus Wide Data, raw data versus analytics etc. Hanging over this is the seemingly interchangeable use of the words “data”, “information” and “knowledge”. Official UK definitions for example, state that the Data Protection Act “controls how your personal information is used”, that the UK Information Commissioner “promotes data privacy for individuals”, and the Freedom of Information Act allows citizens to request publically-held datasets (all our italics).². In addition, Romer (1991), Romer (1993) and Jones (2005) use terms like “ideas”, “blueprints” and “instructions”.

Second, much of what has been published on ‘Big Data’ (BD) and data analytics has focused on the sheer volume, or growth in volume, of data available to firms. On volume, Google’s Eric Schmidt is commonly quoted as stating that as much data/information is being created every two days as was created from the dawn of civilisation to 2003 (Wong 2012). Without diminishing the technological wonders that lie behind this statement, the implications of this for investment and GDP are unclear, but clearly our framework must take this into account.

Third, currently, much of the work on BD has focussed on the business strategy implications and speculated on the information value chain. Thus for example, Mayer-Schönberger and Cukier (2013) classify the properties of BD according to the three Vs (volume, variety and velocity) and, using a rich variety of case studies, document how companies have managed to profit from BD in various ways.

¹ The work is therefore consistent with, and complementary to, a broader work programme that seeks to measure investment in intangible (or knowledge) assets that make long-lived contributions to growth, see for example Corrado, Hulten et al. (2005) or Goodridge, Haskel et al. (2012) for an application to the UK.
Helbing (2014) and Schwab, Marcus et al. (2011) assert that Big Data is the new input to the economy in the 21st Century in the way that oil was in the 20th Century. Once again, our purpose here is not to question these cases, but to encompass them in a consistent framework.

Fourth, a stream of work has documented the dizzying changes in the industrial and occupational structure around BD (e.g. OECD (2013)) for example: the entry of data analytics companies into the market, as well as “data lockers”, “data brokers” and “data exchanges”; or the employment of data scientists and other occupational titles unheard of even a few years ago. From a conventional GDP measurement standpoint, this creates obvious challenges for statistical authorities who work with industrial and occupational classifications that do not distinctly identify many of these types of companies or workers. Furthermore, much work has documented the internal transformation of many companies who have started to use such techniques (e.g. Manyika, Chui et al. (2011)), retailers for example, and to the extent that this work occurs in house, it is again effectively invisible to statistical authorities using conventional survey techniques and standard industrial/occupational classifications.

Finally, there is a stream of work that is closer to the objectives of this paper and wider work programme, which seeks to quantify activity around BD and understand how its use might affect productivity, unit costs and growth. Thus for example, some work has tried to measure those employed in BD, see for example Mandel (2012; 2013; 2014), e-skills UK (2013a; 2013b) and Chebli, Goodridge et al. (2015). Attempts to measure the macroeconomic benefit of BD have been made by CEBR (2012) and Manyika, Chui et al. (2011). There are also micro studies that study the link between the use of data and productivity, see for example Brynjolfsson, Hitt et al. (2011), Bakhshi, Bravo-Biosca et al. (2014) and Tambe (2013).

Our approach is therefore macro-based. That is, we are interested in asking how BD appears in current estimates of GDP (if at all) and how it might affect GDP growth. In particular, we are interested in measuring the investments made in the acquisition or management of transformed data and data-based knowledge, and the contribution those investments make to growth. ³ Our agenda therefore fits the measurement agenda outlined by OECD (2014), which specifically encourages business, statistical and research communities to “measure and value digitised data as an intangible asset, and analyse its contribution to productivity and business performance”. There is of course a growing set of micro-estimates of the impact of BD in various firms and industries, to which this current paper does not do justice. Rather our attempted contribution here is to try to better understand how BD might fit into current and future GDP measures. We believe this to be of interest for the

³ In this paper we focus on the measurement of investment. In Goodridge and Haskel (2015b) we use the estimates in this paper to conduct a sources of growth decomposition and so estimate the contribution of data and data-based knowledge to growth in the UK market sector.
following reasons. First, there are a number of forecasts of how BD will contribute to growth and prosperity in the future and we think that an explicit framework of how BD affects GDP will help better inform such estimates. Second, as we shall argue, some parts of (investment in) BD are in fact already counted in GDP, a point which we do not think is sufficiently appreciated. Third, our work should provide a road map for investigators and statistical agencies for what we need to measure to understand BD’s effect on the macro economy. Fourth, we try to set out how BD fits into the extensive research programme on the information/knowledge economy, since data must be related in some way to information and knowledge.

We propose to treat transformed data (or information) and commercial knowledge acquired from data\(^4\) as assets in a national accounting framework. It is therefore also necessary to consider how data is currently treated in the official System of National Accounts (SNA) (United Nations 2008). Currently, the SNA recommends the capitalisation of expenditures on ‘computerised information’, which is comprised of software and databases, but is mostly considered to consist of software. We therefore review the current UK methodology for estimating investment in this asset category, to help gauge just how much investment activity in data and data analytics is already counted in the official UK national accounts. This is a natural place to focus since, although it is mathematics and statistics that are the foundations of data analytics, both data-building and data analytics require software programming skills, so that investments in these assets are very much related.

The rest of this paper is set out as follows. Section two sets out definitions to be applied in the rest of the paper and our conceptual framework. Section three discusses the justification for treating transformed data and data-based knowledge as assets, including a discussion of SNA criteria, and how activities in and around data and data analytics fit into those, as well as detail on official practice. Section four considers measurement of investment in data from the perspective of purchases and in-house activity. Section five presents results and section six concludes.

## 2 Framework

### 2.1 Definitions: data, information and knowledge

Current literature on the subject of data and information, and other literature on data and data analytics, uses terms in a variety of ways. It will therefore be useful to set out some definitions, as we do below, and which we endeavour to stick to throughout the rest of this paper and future work.

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

\(^4\) We provide formal definitions of these terms in a following section.
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.\(^5\) 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.\(^6\)

Commercial knowledge is therefore that knowledge that is invested in by firms and applied in the process of production. The economics literature has long considered private expenditures on R&D as constituting investment (e.g. Abramovitz (1956)). In this paper, we shall consider expenditures on the transformation and analysis of data in a similar vein, and in future work estimate the contribution those investments make to economic growth.\(^7\)

---

\(^5\) Boisot and Canals (2004) distinguish between data and information, arguing that information is regularities in data which agents attempt to extract, and that this extraction comes with a cost. Regularities in data for us constitute knowledge. In turn they define knowledge as an agent’s set of expectations that are modified by new information (Arrow 1984). Using that definition, information is extracted from raw data and used to build knowledge, which is in line with the schematic we present below.

\(^6\) Goodridge and Haskel (2015b) provide a fuller discussion of these definitions and how they align with similar concepts in the wider literature.

\(^7\) The idea underlying this is that growth can come from two sources. The first is duplication i.e. increasing output via increases in factor inputs, capital and labour. The second is technical progress, which can be thought of as increased efficiency in combining inputs to generate output. Technical progress can derive from freely available knowledge, or from knowledge that is acquired privately, by firms, using costly resources. Since there are limits to the quantities of (tangible) capital and labour that can be employed, it is widely recognised that it is knowledge and innovation which ultimately drives long-run economic growth. We attempt to measure this process, treating the expenditures made in acquiring commercial knowledge as investments in national accounting and growth accounting frameworks, thus allowing us to estimate the economic contribution of private investments in knowledge, and separate them from the TFP residual, so that the latter is more reflective of growth derived from free sources of knowledge (see Goodridge and Haskel (2015b)).
2.2 The big data “supply chain”

With these definitions in mind, we now summarise the big data supply chain, that is, the process of producing transformed data (information) and commercial data knowledge, and the use of that knowledge in final production. 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. We also assume for the moment that the firm employs no other intangible capital (such as R&D), again this assumption is relaxed below.

2.2.1 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, a process sometimes referred to as “data management”, “data acquisition” or “data warehousing”. The literature 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 and of an analytical format; 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.

---

8 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.

9 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 transformed data or 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.
2.2.2. 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.

To preview our analysis below, in trying to understand investment by firms, we need to understand the costs they pay at every stage of production. We have in mind that the knowledge creation stage does not get transformed/readable data for free; rather, it must pay a market price for it, which we assume is a rental price like the licence fee to use a patent. Details are set out below.
2.2.3. **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, please see Goodridge, Haskel et al. (2014).\(^\text{10}\)

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.\(^\text{11}\)

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 primary focus here is on knowledge creation in the

\(^{10}\) The case studies reviewed below also emphasise how data could be employed in those other upstreams e.g. in the production of R&D, market research or design. We document and comment on these complementarities in our results section.

\(^{11}\) The treatment is therefore perfectly symmetrical with purchased tangible capital (i.e. buildings, machinery etc.), for which firms pay an implicit but unobserved annual rental for use of the asset.
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.

2.2.4. Industry case studies
To fix ideas, we consider whether this framework accommodates the examples of applications, or potential applications, of data and data analytics in market production, as reported in Manyika, Chui et al. (2011) and OECD (2013). In these examples, we think of data being used to create data-based knowledge which in turn is applied in downstream production, increasing productivity by either: a) improving efficiency and reducing costs; or b) by adding to the quantity/quality of goods and services produced, thereby increasing output.

a) Retail
First, consider retail, a sector which generates large volumes of data via its transactions with customers and suppliers and in its routine operations. To emphasise the volume of data present in the industry, it has been said that Tesco generate 1.5 billion new items of data per month (Manyika, Chui et al. 2011). Other retailers such as Amazon and Wal-Mart are also known for their extensive use of data; for instance, it is estimated that Walmart collects more than 2.5 petabytes\(^{12}\) of data per hour from its customer transactions (McAfee and Brynjolfsson 2012). Such data is used to improve processes in and around marketing, the management of supply chains (better inventory management), pricing (modelling of demand elasticities) and in the provision of new services. Indeed the retail sector has made use of data in its operations for decades particularly in the optimisation of distribution and logistics, and more recently, growing e-commerce activity means that retailers have access to click-stream data to improve their marketing functions.

In terms of specific applications, consider the customer loyalty schemes operated by numerous retailers, particularly supermarkets,\(^{13}\) and Amazon’s recommendation feature, with each used to improve marketing campaigns and better target consumers. In these cases, firms have developed aspects of their service that generate raw records from which they can ultimately derive knowledge and improve the productivity of their downstream processes. In our framework, the resources they devote to managing and transforming those data, and extracting knowledge, are investments in capital that ultimately contribute to downstream production.

More recently, the retail industry is developing the ability to track its products, as well as its customers (using web-bugs and cookies), and gauge online customer behaviour and sentiment. Wal-Mart, for example, mine data on consumer preferences and behaviour to gain concessions from

---

12 A petabyte is one quadrillion bytes.
13 For instance, ASDA use data provided by Experian to offer personalised promotions to its customers.
suppliers (Manyika, Chui et al. 2011; OECD 2013). Real-time location data can also be used to target potential customers close to, or inside, stores, and track their movements in-store, with the resulting knowledge used to improve store layout for example. Further, new retail services include data-based knowledge services, such as the provision of price comparison services and information on features/characteristics gathered from online feedback and customer reviews.

b) Manufacturing

Second, more so than any other sector, the manufacturing sector generates and stores large quantities of data in its routine operations (Manyika, Chui et al. 2011) and uses that data to either improve efficiency or increase the quality of output. On efficiency, knowledge acquired from data is used to improve forecasts of demand and better manage supply chains, with data sometimes combined with that from other parts of the supply chain, such as the retail sector. Machine-embedded sensors also generate real-time data that can be used to either optimise production or monitor the performance of goods sold. On improving the quality (and therefore increasing the real quantity) of output, computer-aided processes around design, engineering and R&D generate data which, when brought together, can enable increased experimentation and provide platforms for co-creation in partnership with suppliers or customers. Manufacturers can also implicitly acquire customer input in the design process, through the mining of comments on social media, or the mining of ideas submitted through open idea submission processes. In terms of our framework it seems that in manufacturing in particular but also in other sectors, data and data-based knowledge are used as inputs in the creation of other forms of knowledge-based capital, such as new product and process designs, as well as in downstream production.

c) Telecommunications

A third sector that generates large amounts of data is telecommunications, which in particular acquires great volumes of personal location data via GPS on mobile phones or cell-tower triangulation data, some of which is used within the industry including in the provision of new data-based services, and some used to generate knowledge in outside industries.

Major applications of telecommunications data include the provision of location-based applications and services e.g.: routing information which can be based on traffic/weather/road conditions and has the potential to improve productivity in transport and logistics; safety-related applications that track people or locate stolen property; or the provision of information services on local points of interest and services available to consumers. Personal location data can also be used in geo-targeted advertising with personalised promotions to individuals close to particular locations, electronic toll collections or the monitoring of individual behaviour by insurance companies (Manyika, Chui et al. 2011). As an example of the latter, in the UK the insurance company Norwich Union install a black
box in customers cars to monitor driving patterns, and then offer a price/rate based on the knowledge they acquire from that data (Bollier and Firestone 2010). Other more macro applications of telecommunications data can include urban planning, making use of data on traffic, transit etc..

2.2.5. Statistical evidence
Above we have provided some descriptive cases for how data can function as a durable asset in firms/industries, and thus contribute to growth in final output and productivity. There is also a growing body of statistical evidence on the contribution of data and data-based knowledge to productivity, thus emphasising its status as an asset. Using firm-level data on business practices and controlling for traditional capital including ICT, Brynjolfsson, Hitt et al. (2011) find that the use of data-driven decisionmaking (DDD) can explain a 5 to 6% increase in firm output and productivity and is also associated with significantly higher firm profitability and market value, with potential issues around reverse causality addressed using instrumental variable techniques. Similarly, using data from a NESTA survey on data activity, Bakhshi, Bravo-Biosca et al. (2014) find that data active firms are on average 8% more productive than their counterparts. The same authors also report strong links between data analytics and firm productivity, with firms that empower employees to implement insights gleaned from data found to be 16% more productive. Further support for the treatment of data as an asset can be found in: Economist Intelligence Unit (2012), which reports results of a survey of managers who on average stated that (big) data had improved their organisations performance by 26% over the past three years; Davenport and Harris (2007) who make the link between the use of data analytics and acquiring a competitive advantage; and LaValle, Hopkins et al. (2010) who show that firms that employ data analytics are twice as likely to be among the industry top performers.

3 Economics and national accounting conventions on treating information and data-based knowledge as assets
The above framework modelled information and data-based knowledge as assets that make long-lived contributions to production. However, if we are to count them as assets in a national accounting context, it is worth saying a little more on the justification for that, including a discussion of official capitalisation criteria as set out in the SNA.

3.1 Do information and commercial data knowledge function as assets?
To assess whether or not (transformed) data and data-based knowledge ought to be counted as assets, and whether the expenditures towards their creation ought to be counted as investments, it is worth reminding ourselves of the definitions of capital and investment.
As pointed out in Jorgenson and Griliches (1967) and Hulten (1979), savings and investment are a means of sacrificing current consumption in order to increase future consumption, making the appropriate definition of economic investment the devotion of current resources to the pursuit of future returns (Weitzman 1976; Hulten 1979). Consistent application of that definition immediately makes clear that whether expenditure is on a factory or a virtual data centre for long-term use does not matter to the question of what ought to be classified as investment. What matters for the purposes of capitalisation is whether data (information) and data-based knowledge function as assets that generate future returns and make long-lived contributions to production. As noted in CEBR (2013), data that enables firms to derive future economic benefits ought to be regarded as assets. The next section considers whether these assets meet the capitalisation criteria as set out in the System of National Accounts (SNA) and OECD recommendations (OECD 2010), and whether investments in data are already measured in the official national accounts data on gross fixed capital formation (GFCF).

3.2 System of National Accounts (SNA)

SNA investment criteria have the same interpretation as those from the economic literature above. If an input contributes to production over more than one accounting period, its acquisition ought to be counted as investment. However, some features of data and databases, namely that data is, in national accounts nomenclature, a non-produced asset, have implications for measurement which we expand on below.

Generally, according to the SNA (2008), assets are “entities that must be owned by some unit..., from which economic benefits are derived by their owner(s) by holding them or using them over a period of time” (United Nations 2008). Intermediate consumption is the consumption of goods or services in production, such that those goods are used up in the course of the accounting period (one year). Gross fixed capital formation (GFCF) is investment in produced assets that are used repeatedly in production over more than one accounting period. The distinction between GFCF and intermediate consumption therefore depends on whether or not the good in question is used up in the course of one year, termed the “asset boundary” in the SNA, with the key feature of an asset being its repeated use in production over a period longer than one year.

Further, the SNA describes intellectual property products (IPPs) as assets that are “the result of research, development, investigation or innovation leading to knowledge that the developers can market or use to their own benefit in production”, and states that such knowledge remains an asset until it is ether no longer protected or becomes obsolete. We note that provided they are repeatedly

---

14 Where acquisition can include the purchase of an asset in a market transaction, or own-account (in-house) asset production.

15 Intermediates not used up in the accounting period form inventories, which are part of Gross Capital Formation (GCF) but not Gross Fixed Capital Formation (GFCF) since they do not meet asset criteria.
used over more than one accounting period, transformed data (information) and commercial data knowledge meet the SNA definitions for both assets and, more specifically, IPPs.

Current national accounting convention capitalises the following types of produced intangible assets (or IPPs): computerised information (software and databases); entertainment, literary and artistic originals; mineral exploration; and, most recently, R&D. The latest revision of the SNA (2008) places increased emphasis on databases as assets, with databases defined as “files of data organized in such a way as to permit resource-effective access and use of the data”. The latest revision to the SNA also mandates that databases are to be explicitly included as a separate sub-component of ‘software and databases’, and recommends that estimates of GFCF in databases be estimated separately. In the 1993 revision to the SNA, only “large” databases were considered assets. The 2008 revision correctly recognises that all databases, regardless of size, that provide an economic benefit to their owner and with a useful service life greater than one year, should be treated as fixed assets. To help indicate whether this is so, the OECD Handbook on Deriving Capital Measures of Intellectual Property Products (OECD 2010) recommends that “a database should be recorded as a fixed asset if a typical datum is expected to be stored on the database, or archived on a secondary database, for more than one year.”

However, measurement of investment in databases is complicated by the fact that, like land, data (at least in its raw form, raw records in our nomenclature) is a non-produced asset. Therefore, according to the SNA, while expenditures on what we term data-building or data transformation ought to be recorded as investments, expenditures on acquiring that data ought not. We note that this fits with our measurement framework in the sense that we aim to count the investments made in the data-building and data analytics processes, with raw records modelled as non-produced assets that are either generated for no cost, or paid for in the same way as raw materials.

Conceptually, and in common with R&D, mineral exploration and indeed all other assets, all investment in transformed data or data-based knowledge should be recorded as GFCF, regardless of whether or not it is successful i.e. whether or not it generates some useful knowledge to be used in production. Failed investments can also generate the knowledge to make subsequent investments a success. It is also expected that investors/owners consider the chance of failure in their investment decision, and that successful investments provide benefits that compensate for those that are unsuccessful. Only counting those investments that are successful would result in over-estimation of the returns to those investments. Failure should be accounted for in the applied rate of depreciation, which accounts for the rate of obsolescence and discard/retirement.
3.3 Measurement of UK investment in ‘computerised information’: purchased and own-account

As we stress above, investments in data-based information and knowledge can either be purchased/outsourced or alternatively, conducted in-house. In the UK national accounts, estimates of investment in computerised information (software and databases) are comprised of two components: a) purchased, which includes purchases of either pre-packaged or custom-made software, and in theory ought to include the purchase of data(bases); and b) own-account, which consists of software written in-house by firms for their own use and again ought to include investments in data(bases) developed in-house. Note that the SNA does not explicitly recommend that expenditures on data analytics be recorded as investments.16

3.3.1. Purchased GFCF: Computerised information (software and databases)

For purchases of databases, or copies of databases, the SNA and OECD (2010) recommendations are consistent with those for software. That is, if a licence to a copy is purchased with annual payments over multiple years, or a one-off payment to cover multiple years, and if the licensee assumes all risks/rewards associated with economic ownership of the copy, then that purchase is GFCF. If annual payments are made for a licence to use a copy without a long-term (greater than one year) contract, then those are payments for database services. In cases where users regularly pay for access to a database that is frequently updated then they are consuming database services rather than undertaking GFCF.17 If the original database is sold and the owner divests itself of all responsibilities to issue and service copies, then this is the sale of part, or all, of the asset represented by the original, making it positive GFCF (i.e. an acquisition) for the purchaser and negative GFCF (i.e. a disposal) for the seller.

Purchases of databases are collected via the capital expenditure (CAPEX) survey which we discuss more in Appendix 1, but we conclude that purchases of data are likely not well captured in the official measures. Data purchases are however considered to be rare and most investment in databases is thought to occur on own-account (OECD 2010) suggesting that is where measurement ought to be focused.

3.3.2. Own-account GFCF: Computerised information (software and databases)

The ONS also produce estimates of own-account investment in computerised information, although the methodology was developed in order to better estimate investment in software (Chamberlin,

---

16 However, we note that expenditures on data analytics do meet the SNA definition for IPPs above, as well as the definition of R&D from the Frascati Manual. We note therefore the potential for such expenditures to be recorded in business investment in R&D, as noted in Chebli, Goodridge et al. (2015) and expanded on below.

17 Note that in the case of a firm consuming database services, then the supplying firm must have invested in creating or purchasing a data asset that they are using to provide those services. In which case, we need to measure the investment of the database service provider.
Chesson et al. 2006; Chamberlin, Clayton et al. 2007). This is estimated separately, with the CAPEX survey not thought to well capture own-account investments.

Measures of own-account investment capture an implicit transaction between the producer (for own final use) and user of the asset, in the case where they are both the same entity. Since no market transaction actually takes place, the SNA recommends valuation based on the sum of costs incurred in producing the asset. Therefore, for market producers, own-account output should be valued as the sum of: intermediate consumption, compensation of employees, consumption of fixed capital, a net return to fixed capital, and taxes less subsidies on production.\(^\text{18}\) That output is a direct component of value-added in the wider market sector. Additionally the SNA states that, in the case of IPPs, all own-account asset production is to be treated as GFCF as work takes place, regardless of whether it is work-in-progress or complete, unlike say construction projects which are recorded as inventories whilst still in progress and GFCF when complete. The SNA also specifically recommends that when valuing the own-account output of databases, the costs of developing the database management system be excluded (since that is software), the cost of preparing data in the appropriate format be included, and the costs of acquiring or producing the data be excluded. The last point relates to the treatment of data as a non-produced asset, as noted above. OECD (2010) further states that the costs of updating a database should also be recorded as GFCF.

How do these principles fit in with our framework? In short, all are consistent with the framework for measurement presented in this paper. We note that the SNA and OECD recommend that the “cost of preparing data in the appropriate format” and “updating”, be included. In our framework, these activities form part of the data-building (D) process. The OECD also recommend that the “costs of acquiring or producing the data be excluded”. As we will show below, in our framework we model raw records as freely available and thus acquired for no cost, for example, it could be data scraped from the web or generated automatically via transactions or some other business process, with a marginal cost of (almost) zero. Finally, the OECD recommend that the “costs of developing the database management system” be excluded since they are already in software. This could include for example the in-house modifications made to open-source software packages typically used in data-building (D) and knowledge creation (N) processes, such as Hadoop, Python or R. To minimise the

\(^{18}\) Where: intermediate consumption includes all relevant materials and intermediates; compensation of employees refers to wage and non-wage labour payments (therefore including social security, pension contributions and any income-in-kind) of those engaged in own-account production, including if necessary some assumption/estimate on the amount of time spent engaged in asset production; consumption of fixed capital accounts for the depreciation of fixed assets used in own-account production; the net return to fixed capital accounts for the returns to those assets (together with consumption of fixed capital this accounts for the gross return on capital employed); and taxes and subsidies on production account for all taxes/subsidies paid/received on the costs of asset production (OECD 2010). For non-market producers, the sum of costs should exclude the net return to fixed capital used in production, since, by national accounting convention, non-market producers do not generate any operating surplus over and above the consumption of fixed capital.
risk of double-counting here, we will employ an approach that is fully consistent with the methodology for estimating investment in own-account software. The additional element in our framework, not explicitly considered in the official recommendations, is the data analytics conducted in the knowledge creation (N) stage, which we also count as an investment activity.

An important issue that needs to be considered in the measurement of own-account investment is the potential for double-counting. For example, if we were to estimate the own-account production of an original database, and the ownership rights and the right to reproduce, were subsequently sold and purchased by another firm, it would be double-counting to measure both the purchase and the original own-account production as GFCF. In that case, the own-account production should be recorded as positive GFCF for the producer, and the sale as negative GFCF (disposal) for the producer and positive GFCF (acquisition) for the purchaser. In such a situation it would be appropriate to either exclude that item of own-account production, or if that is not possible, only record the own-account production and not add on the subsequent purchase. If, on the other hand, the owner of the original were to sell a copy of the data but retain the original asset, then according to SNA and OECD recommendations, the purchase of the copy should be counted in GFCF as an ‘investment in use’ (OECD 2010), and the original own-account production should still be recorded as positive GFCF. In practice, distinguishing between these two types of transaction is likely to be problematic.

As with purchased investments, official estimates of own-account investment in computerised information cannot be separated out into their two respective components, software and databases, with part of the reason being that a database “cannot be developed independently of a database management system (DBMS) which is itself computer software” (United Nations 2008).

3.3.3. Computerised information GFCF: Summary

Therefore in current UK practice, estimates of investment in databases are in principle included in estimates for GFCF in computerised information. Unfortunately, at this moment in time, those estimates are not separated out into their respective components for a) software and b) databases, although they are to be in future following the latest guidance from the SNA. It is unclear however just how well current methods capture UK investments in (transformed) data and commercial data knowledge. Before considering a revised model, it is worth noting though that one aspect of ‘data investment’, namely expenditures on ‘data management software’, whether purchased or undertaken in-house, will already be included in the measured investment data.

19 See Appendix 1 for discussion of why purchased investments in data are currently likely not well captured in official measurement. Note the SNA does not capitalise investments in the knowledge creation/data analytics process.

Consider an economy that produces long-lasting assets (investment goods) and consumption goods, and is composed of four distinct sectors: a data-building sector \( D_t \) that manages and transforms data, a knowledge creation \( N \) sector, a separate software \( S \) sector and a final goods sector \( Y \) that uses these outputs in final production.

The structure of the production chain is as set out in Figure 1. The upstream data-building sector \( D_t \) manages data and transforms raw records into data (information) of a format ready for analysis, using labour \( L_t^D \), tangible capital \( K_t^D \), software \( R_t^{D,S} \) and intermediate goods/services \( Z_t^D \). Nominal sector output \( P_t^D \) is equal to the sum of factor payments, where factor payments are at competitive prices for capital and labour, multiplied by a factor, \( \mu_t^D \). 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,\(^{20}\) or paid for in the same way as other material/intermediate inputs.

The factor \( \mu_t^D \) is incorporated because the D sector might be able to mark up prices over competitive costs. First, it might 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 in the sector (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 market competition, the scarcity of that information, and its commercial value to the ultimate users. Of course the acquisition and maintenance of market power provides a further incentive for the sector to exist in-house.

The output of the data-building sector (information assets, \( D_t \)) forms a stock of (bytes of) information \( B_t \) which is used as an input in the knowledge creation sector. That sector uses transformed data to create commercial data knowledge \( N_t \), employing tangible capital \( K_t^N \), software \( R_t^{N,S} \) and labour. Factor payments thus include those paid for the use of bytes of information \( P_t^B B_t^N \) and again the value of sector output (commercial data knowledge assets) incorporates a product mark-up, \( \mu_t^N \), to account for the market power acquired by the owners of unique data-based knowledge assets.

\(^{20}\) Where exhaust data is data generated as a by-product of some other online or digital process.
How are we to measure investment in Big Data assets in such an economy? Consider a statistical ideal. In a non-vertically integrated economy all these activities would be carried out in different firms. Each firm would be classified to one of the four industries. We could then measure investment using either (a) the output of industries that produce investment goods, the D, N and S industries or (b) we could ask the firms in the Y industry how much they are spending on investment in D, N and S goods i.e. buying the assets outright rather than renting them.

In practice however, some of the production of D, N and S goods occurs in-house and some is conducted by stand-alone firms for purchase by other firms. Thus instead of industries, we consider activity in terms of sectors, where sectors can either exist in-house or in stand-alone industries. To measure investment we must identify both forms of activity, thus we have to add to the sales of the D, N and S industries (adjusting for an open economy) the in-house (own-account) spending. Thus we may write for the D, N and S sectors that GFCF is:

\[
\hat{\text{GFCF}}(P^D D) = P^D D^{\text{PUR}} + \mu^D \left( P^L L^D + P^K K^D + P^R R^D + P^Z Z^D \right)
\]

\[
\text{GFCF}(P^N N) = P^N N^{\text{PUR}} + \mu^N \left( P^L L^N + P^K K^N + P^R R^N + P^B B^N + P^Z Z^N \right) 
\]

\[
\text{GFCF}(P^S S) = P^S S^{\text{PUR}} + \mu^S \left( P^L L^S + P^K K^S + P^R R^S + P^Z Z^S \right)
\]

Where GFCF in each asset (D, N, S) incorporates assets purchased from specialist industries (denoted PUR) and those produced in-house. In-house spend on production of D assets includes that on labour (L), tangible capital (K), software (R) and intermediates (Z) multiplied by a mark-up (\(\mu\)). The equations for the N and S sectors are similar, but inputs in the N sector also include (implicit or explicit) payments for the use of the stock of transformed data or information (B).

4.1. Actual data

In considering (1) there are a number of issues. Consider first the PUR terms, being the sales of the D, N and S industries. The SIC is not fine enough to distinguish the D and N industries; rather, ONS groups D and N activities within the two industries SIC 62012, “Business and domestic software and development” and SIC 63110, “Data processing, hosting and related activities” (see Appendix 2 for more detail). Let us suppose true D and N sales to be a fraction (\(\gamma\)) of total sales in those two industries (\(P^D Q\)).

\[\text{21} \] In the S sector, inputs include software assets, with some aspects of own-account investment including in-house modifications to previously acquired software.

18
Second, when constructing own-account investment, the ONS does not in practice have data on capital and intermediate usage (except in unusual cases) and so uses the wage bill of software workers. It then applies two adjustments. First, it assumes that software workers spend some of their time on non-long lived asset creation (e.g. day to day activities, administration, maintenance etc.) and so only a fraction, denote this $\tau$, is on long-lived GFCF. Second, one must adjust wage costs for other factors e.g. overheads etc., call this $\lambda$. Thus we have (2), where in the final terms we are summing over D, N and S occupations, and we exclude those workers in the D, N and S industries, since their output is included in the first term.

\[
GFCF(P^D) = \gamma^D (P^D Q)^{\text{I=SIC}62012,63110} + \mu^D \sum_{OCC\in D, I=SIC62012,63110} \lambda^{D,\text{Cost}} \tau^D (P^L L^D)^{OCC,I}
\]

\[
GFCF(P^N) = \gamma^N (P^N Q)^{\text{I=SIC}62012,63110} + \mu^N \sum_{OCC\in N, I=SIC62012,63110} \lambda^{N,\text{Cost}} \tau^N (P^L L^N)^{OCC,I}
\]

\[
GFCF(P^S) = \gamma^S (P^S Q)^{\text{I=SIC}62} + \mu^S \sum_{OCC\in S, I=SIC62} \lambda^{S,\text{Cost}} \tau^S (P^L L^S)^{OCC,I}
\]

4.2. Practical measurement

There are two practical issues in the implementation of this. First, as noted in Appendix 2, whilst the UK Standard Industrial Classification (SIC) does encompass the D and N industries, we have no information on $\gamma^D$ or $\gamma^N$. In addition, to the extent that such services are bought internationally e.g. rented from a Cloud provider abroad, they will not show up in the output of UK companies. Also, if such services are also provided direct to households that would not be investment by firms.

However, we are able to circumvent this issue by directly measuring all own-account production by counting the wage bill in D and N occupations, but including I=SIC62012, 63110. Thus we measure D and N activity, regardless of whether it takes place in specialist D and N industries or in outside industries.

Second, statistical authorities do measure software, including own-account investment as in the second term, but in practice some of the D and N workers might be included in that measurement as follows. In the estimation of investment in computerised information, the ONS does not distinguish between D, N and S workers. Rather, it takes a list of occupations thought likely to be involved in software creation (S). As shown in Chebli, Goodridge et al. (2015), some of those occupations

\[22\] Although the asset as defined by the SNA is “computerised information”, the methodology was primarily designed to estimate investment in software (Chamberlin, Chesson et al. 2006; Chamberlin, Clayton et al. 2007).
include D and N workers, but some D and N workers are outside such a list. Second, when surveying firms on their purchases, ONS asks for purchased software and databases, but it seems that firms likely do not reply to the latter very well (see Appendix 1). In light of this we can write measured $P^S_S$, denoted as $P^S_S^{MEAS}$, as:

$$GFCF(P^S_S^{MEAS}) = (P^S_S^{PUR} + \zeta^{D,PUR} P^D D^{PUR}) +$$

$$\mu^S \left( \sum_{OCCj \in S, SIC \neq software(62)} \lambda^S (P^L L^S)^{OCCj} + \sum_{OCCj \in D, SIC \neq software(62)} \zeta^{D,OCC} \lambda^D (P^L L^D)^{OCCj} + \sum_{OCCj \in N, SIC \neq software(62)} \zeta^{N,OCC} \lambda^N (P^L L^N)^{OCCj} \right)$$

(3)

Where $\zeta^{D,PUR}$ highlights that purchases of data are likely only partially captured in $P^S_S^{MEAS}$, if at all. Similarly, $\zeta^{OCC}$ highlights that while some D and N workers are included in the estimation of own-account software, some are excluded. Note that workers in the (broad definition of the) software industry (thus including a lot of what we term the D and N industries) are excluded from the final term since they are in the purchased term. Note also that in the ONS method, $\mu^S = 1$.

To summarise, in the measured data some part of measured purchases may include expenditures on databases, although likely not all in practice. The method for measuring own-account investment means that a significant amount of D and N activity is measured there too. Therefore this must be accounted for in any measurement of D and N activity, so that estimates of investment in data (and data-based knowledge) are the sum of that part already recorded in the measurement of software, plus any additional investment not recorded.

4.3. Summary of measurement methods

Now that we have set out what we want to measure, as well as what is already measured, our steps are as follows. First, we can go to the occupational data and measure $L^D$ and $L^N$ and then add in assumptions about $P^L$, $\tau$, $\lambda$, and $\mu$. Second, to do this, we must be aware that at least some of these will already be counted in $P^S_S^{MEAS}$ and so we will have to take that out.

5. Direct measurement of investment in D and N activity

5.1. Own account spending on D and N services

Our starting point is the wage bill of D and N workers. Thus we use estimates of big data employment as set out in Chebli, Goodridge et al. (2015), and apply the ONS method for estimating own-account investment in computerised information which we review briefly below.
UK estimates of investment in own-account computerised information are constructed using data on the labour input of occupations considered engaged in asset production (collected via the Annual Survey of Hours and Earnings (ASHE)). Specifically, the ONS estimate own-account output as set out in (3), by first observing the wagebill of relevant workers with some adjustment for the fraction of their time spent building assets (see Table 1 for detail on occupations). Relevant workers are those with an occupation, such as “software professional” that the ONS judges makes them likely to be engaged in the creation of long-lived software. Then, to account for non-wage labour costs (income-in-kind, pensions, social security etc.), that estimate is multiplied by a factor of 1.16, giving a final estimate of $PL$. Then to account for materials and the depreciation of fixed capital, the ONS multiply $PL$ by a factor of 1.84, based on the ratio of non-employment costs (including depreciation of capital) to employment costs in the software industry itself. Finally they multiply that by a factor of 1.15 to account for the net return to fixed capital in upstream production, thus deriving a final estimate of own-account output. The ONS implicitly assume that $\mu=1$.

We apply the same method in Table 1 which reads as follows. It is centred around occupations identified in Chebli, Goodridge et al. (2015) as engaged in creating D/N assets, set out in columns 1 and 2. The upper panel shows the occupations that are identified by ONS as creating software and the lower panel those not considered by ONS as creating software. Column 3 shows market sector employment for each occupation, constructed from AHSE microdata (Office for National Statistics). As the total shows, there are just over 2m workers in these occupations, of whom 748,769 are in software occupations. Column 4 sets out estimates of Big Data employment, drawn from Chebli, Goodridge et al. (2015). Using a dataset constructed from the publically available profiles of members of an employment based social media network, Chebli, Goodridge and Haskel identify the fractions of people in various occupational groups with skills or job descriptions in big data using a list of keywords (skills in Hadoop, Python, sentiment analysis or predictive analytics for example). As the total shows, they estimate that in 2010 the UK market sector employed 190,000 workers in D and N activity. Of those, they estimate that 123,000 (65\%) are already measured in own-account investment in computerised information, with 67,000 additional workers (35\%) not already counted in official measurement.

Thus, as column 5 shows, 65\% of big data employment is already measured in the employment of workers in computerised information (Panel 1), with as much as 40\% in the single occupational code of “software professionals” (2132), and with 35\% in outside occupations (Panel 2) and so not

---

23 In the notation of (3), $\lambda=1.84*1.15=2.12$

24 For example, Chebli, Goodridge et al. (2015) show that in 2010, UK market sector employment of “software professionals” according to the ASHE was 289,823. They estimate that 26.2\% of workers in this occupational group are engaged in D and N activity, hence the estimate of (0.262*289,823=)75,873 D and N workers for SOC 2132 in Table 1.
CEBR (2013) also argue that current national accounting methods do not well account for the growing importance of software and information services, implying that some aspects of investment are missing from official data. In work on defining the information economy using dynamic mapping techniques, Spilsbury (2015) also identifies emerging job roles in big data as partly present in typical IT occupations, and partly present elsewhere in other occupations, such as “Business and related associate professionals” (3539), which is one of the additional occupations included in Panel 2 of Table 1.

Looking at columns 5 and 6 of Table 1, 40,000 (or 21% of the total) D and N workers are in occupations classified as research and development managers (1137), science professionals (211), engineering professionals (212) and research professionals (232). We note therefore the potential for these workers to already be counted in the measurement of R&D. The working definition of R&D in the context of the national accounts is taken direct from the Frascati Manual (OECD 2002), and is defined as: “creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications”. In our discussions with the ONS they have told us that they would expect data analytics activity to be included in R&D provided it meets the above definition.

We also note the 12,000 (7%) of workers with occupations classified as marketing and sales managers (1132), advertising and public relations managers (1134) and design associate professionals (342). In the industry case studies presented earlier we discussed how data and data analytics can be employed in the creation of other forms of knowledge-based capital, including in advertising, marketing and design functions. There are two possibilities here. First, these workers may simply be “users” of big data insights (i.e. implementation), and we may be overcounting by including them. Alternatively, these workers may truly be involved in D/N production. For now we assume the latter but note that a more conservative estimate could remove these occupations.

The remaining columns convert the employment numbers into investment. Column 6 shows an estimated salary, calculated as a weighted average of the salaries for each occupational group using ASHE microdata (Office for National Statistics). Columns 8 and 9 show the assumed $\tau$ (based on our discussions with industry experts) and $\lambda$, and column 10 the investment estimates by occupation.

---

25 The factor used by the ONS ($\lambda=2.12$) is based on employment and non-employment costs recorded in a broad definition of the software industry (SIC 62), thus including most of what we term the D and N industries. It has however been suggested to us that the appropriate factor for D/N may be higher, given the data security costs associated with data centres. For now we conservatively choose a factor consistent with software in the national accounts.
Table 1: Investment in data-building and data-based knowledge creation, UK market sector, 2010

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1136</td>
<td>Information and communication technology managers</td>
<td>159,974</td>
<td>6,328</td>
<td>0.03</td>
<td>52,571</td>
<td>0.39</td>
<td>0.15</td>
<td>2.12</td>
<td>0.12</td>
</tr>
<tr>
<td>2131</td>
<td>IT strategy and planning professionals</td>
<td>99,387</td>
<td>5,487</td>
<td>0.03</td>
<td>45,418</td>
<td>0.29</td>
<td>0.35</td>
<td>2.12</td>
<td>0.21</td>
</tr>
<tr>
<td>2132</td>
<td>Software professionals</td>
<td>289,823</td>
<td>75,873</td>
<td>0.40</td>
<td>37,333</td>
<td>3.29</td>
<td>0.50</td>
<td>2.12</td>
<td>3.48</td>
</tr>
<tr>
<td>3131</td>
<td>IT operations technicians</td>
<td>93,430</td>
<td>26,067</td>
<td>0.14</td>
<td>31,596</td>
<td>0.96</td>
<td>0.20</td>
<td>2.12</td>
<td>0.41</td>
</tr>
<tr>
<td>3132</td>
<td>IT user support technicians</td>
<td>61,860</td>
<td>6,820</td>
<td>0.04</td>
<td>24,394</td>
<td>0.19</td>
<td>0.15</td>
<td>2.12</td>
<td>0.06</td>
</tr>
<tr>
<td>4136</td>
<td>Database assistants/clerks</td>
<td>30,796</td>
<td>1,242</td>
<td>0.01</td>
<td>16,044</td>
<td>0.02</td>
<td>0.05</td>
<td>2.12</td>
<td>0.00</td>
</tr>
<tr>
<td>5245</td>
<td>Computer engineers, installation and maintenance</td>
<td>13,499</td>
<td>1,038</td>
<td>0.01</td>
<td>27,069</td>
<td>0.03</td>
<td>0.05</td>
<td>2.12</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Panel 1: ONS software occupations (Chamberlin, Clayton and Farooqui (2007))

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1132</td>
<td>Marketing and sales managers</td>
<td>514,849</td>
<td>8,459</td>
<td>0.04</td>
<td>56,367</td>
<td>0.55</td>
<td>0.10</td>
<td>2.12</td>
<td>0.12</td>
</tr>
<tr>
<td>1134</td>
<td>Advertising and public relations managers</td>
<td>30,252</td>
<td>718</td>
<td>0.00</td>
<td>45,881</td>
<td>0.04</td>
<td>0.10</td>
<td>2.12</td>
<td>0.01</td>
</tr>
<tr>
<td>1137</td>
<td>Research and development managers</td>
<td>40,848</td>
<td>1,469</td>
<td>0.01</td>
<td>54,005</td>
<td>0.09</td>
<td>0.10</td>
<td>2.12</td>
<td>0.02</td>
</tr>
<tr>
<td>211</td>
<td>Science professionals</td>
<td>51,703</td>
<td>5,323</td>
<td>0.03</td>
<td>37,503</td>
<td>0.23</td>
<td>0.20</td>
<td>2.12</td>
<td>0.10</td>
</tr>
<tr>
<td>212</td>
<td>Engineering professionals</td>
<td>396,375</td>
<td>22,086</td>
<td>0.12</td>
<td>36,632</td>
<td>0.94</td>
<td>0.20</td>
<td>2.12</td>
<td>0.40</td>
</tr>
<tr>
<td>232</td>
<td>Research professionals</td>
<td>103,455</td>
<td>11,590</td>
<td>0.06</td>
<td>30,003</td>
<td>0.40</td>
<td>0.20</td>
<td>2.12</td>
<td>0.17</td>
</tr>
<tr>
<td>2423</td>
<td>Management consultants, actuaries economists and statisticians</td>
<td>120,311</td>
<td>8,051</td>
<td>0.04</td>
<td>46,928</td>
<td>0.44</td>
<td>0.40</td>
<td>2.12</td>
<td>0.37</td>
</tr>
<tr>
<td>342</td>
<td>Design Associate Professionals</td>
<td>59,625</td>
<td>3,233</td>
<td>0.02</td>
<td>25,040</td>
<td>0.09</td>
<td>0.10</td>
<td>2.12</td>
<td>0.02</td>
</tr>
<tr>
<td>3539</td>
<td>Business and related associate professionals n.e.c.</td>
<td>67,338</td>
<td>6,421</td>
<td>0.03</td>
<td>27,028</td>
<td>0.20</td>
<td>0.40</td>
<td>2.12</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Panel 2: Other occupations that include D and N workers (Chebli, Goodridge and Haskel (2015))

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1132</td>
<td>Marketing and sales managers</td>
<td>514,849</td>
<td>8,459</td>
<td>0.04</td>
<td>56,367</td>
<td>0.55</td>
<td>0.10</td>
<td>2.12</td>
<td>0.12</td>
</tr>
<tr>
<td>1134</td>
<td>Advertising and public relations managers</td>
<td>30,252</td>
<td>718</td>
<td>0.00</td>
<td>45,881</td>
<td>0.04</td>
<td>0.10</td>
<td>2.12</td>
<td>0.01</td>
</tr>
<tr>
<td>1137</td>
<td>Research and development managers</td>
<td>40,848</td>
<td>1,469</td>
<td>0.01</td>
<td>54,005</td>
<td>0.09</td>
<td>0.10</td>
<td>2.12</td>
<td>0.02</td>
</tr>
<tr>
<td>211</td>
<td>Science professionals</td>
<td>51,703</td>
<td>5,323</td>
<td>0.03</td>
<td>37,503</td>
<td>0.23</td>
<td>0.20</td>
<td>2.12</td>
<td>0.10</td>
</tr>
<tr>
<td>212</td>
<td>Engineering professionals</td>
<td>396,375</td>
<td>22,086</td>
<td>0.12</td>
<td>36,632</td>
<td>0.94</td>
<td>0.20</td>
<td>2.12</td>
<td>0.40</td>
</tr>
<tr>
<td>232</td>
<td>Research professionals</td>
<td>103,455</td>
<td>11,590</td>
<td>0.06</td>
<td>30,003</td>
<td>0.40</td>
<td>0.20</td>
<td>2.12</td>
<td>0.17</td>
</tr>
<tr>
<td>2423</td>
<td>Management consultants, actuaries economists and statisticians</td>
<td>120,311</td>
<td>8,051</td>
<td>0.04</td>
<td>46,928</td>
<td>0.44</td>
<td>0.40</td>
<td>2.12</td>
<td>0.37</td>
</tr>
<tr>
<td>342</td>
<td>Design Associate Professionals</td>
<td>59,625</td>
<td>3,233</td>
<td>0.02</td>
<td>25,040</td>
<td>0.09</td>
<td>0.10</td>
<td>2.12</td>
<td>0.02</td>
</tr>
<tr>
<td>3539</td>
<td>Business and related associate professionals n.e.c.</td>
<td>67,338</td>
<td>6,421</td>
<td>0.03</td>
<td>27,028</td>
<td>0.20</td>
<td>0.40</td>
<td>2.12</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Notes to table: Columns 1 and 2 are occupations that include D and N workers, as estimated in Chebli, Goodridge et al. (2015). Column 3 is UK market sector employment in each occupation estimated using ASHE microdata held in the Secure Data Service at the UK Data Archive (Office for National Statistics). Column 4 is our estimate of big data employment, which is a subset of employment in that particular occupational code (see Chebli, Goodridge et al. (2015)). Column 5 is the share of big data employment for that occupational code. Column 6 is the average market sector annual wage for that occupation, also estimated using ASHE microdata. Column 7 are estimated labour payments to D and N workers, that is column 4 times column 6. Column 8 is the time-use assumption (i.e. proportion of time assumed spent building long-lived assets) based on our discussions with industry experts. Column 9 is the ONS estimate of λ to capture upstream use of materials and fixed capital. Finally column 10 presents estimates of total (data-building and knowledge creation) investment derived from the occupational data. In the final three rows we present the total aswell as subtotals to capture that part already measured in own-account software, plus the additional part from occupations outside the official software method.
Thus by combining our estimate of big data employment with occupation specific parameters for wages and time-use, and a general parameter for upstream use of materials and capital, we estimate that in 2010 the UK market sector invested £5.7bn in D and N activity. Note that in our estimates we have implicitly assumed that the product mark-ups, $\mu^B$ and $\mu^N$, are equal to one, as is standard in official measurement practice. As we have also noted, some of that £5.7bn is already counted in official estimates of investment in computerised information, as some of the occupations that write own-account software are the same as those involved in data transformation and data-based knowledge production. Using data from Table 1, we estimate that £4.3bn of D/N investment is already counted as an implicit part of own-account software, and £1.4bn can be regarded as “missing” or previously unidentified investment, not currently recorded in the official data. To provide some context on our estimates of investment, the following table summarises UK market sector investment in various other (tangible and intangible) assets in 2010.

Table 2: Nominal UK market sector investment, comparison by asset

<table>
<thead>
<tr>
<th>UK Market Sector investment in:</th>
<th>£bns (2010, nominal)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tangibles</strong></td>
<td>94.4</td>
</tr>
<tr>
<td>of which:</td>
<td></td>
</tr>
<tr>
<td>Buildings</td>
<td>44.9</td>
</tr>
<tr>
<td>Plant &amp; Machinery (excl. IT hardware)</td>
<td>30.4</td>
</tr>
<tr>
<td>Vehicles</td>
<td>13.6</td>
</tr>
<tr>
<td>IT hardware</td>
<td>5.4</td>
</tr>
<tr>
<td><strong>Intangibles (measured, national accounts)</strong></td>
<td>44.5</td>
</tr>
<tr>
<td>of which:</td>
<td></td>
</tr>
<tr>
<td>R&amp;D</td>
<td>14.8</td>
</tr>
<tr>
<td>Mineral Exploration</td>
<td>0.6</td>
</tr>
<tr>
<td>Artistic Originals</td>
<td>5.7</td>
</tr>
<tr>
<td>Computerised information</td>
<td>23.4</td>
</tr>
<tr>
<td>Software (SOFT\neq DN)</td>
<td>19.1</td>
</tr>
<tr>
<td>Data (SOFT=DN)</td>
<td>4.3</td>
</tr>
<tr>
<td><strong>Additional investment in data (DN+)</strong></td>
<td>1.4</td>
</tr>
<tr>
<td><strong>Total Investment in data (DN)</strong></td>
<td>5.7</td>
</tr>
<tr>
<td><strong>Memo items:</strong></td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>12.8</td>
</tr>
<tr>
<td>Advertising &amp; Market Research</td>
<td>13.5</td>
</tr>
<tr>
<td>Training</td>
<td>27.4</td>
</tr>
<tr>
<td>Organisational Capital</td>
<td>27.0</td>
</tr>
<tr>
<td>Financial Product Innovation</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Note to table: UK market sector investment in 2010 in current prices, £bns. Data for tangibles and national accounts intangibles based on ONS GFCF for selected industries, except R&D which are authors own estimates constructed from BERD. Memo items include estimated investment in intangibles not currently capitalised in the national accounts, taken from Goodridge, Haskel et al. (2014).

---

26 In estimating investment in own-account computerised information, in-house R&D, and investments in film, TV & Radio and artistic works, the ONS implicitly assume that $\mu=1$. 
Our estimate for investment in data in 2010 is £5.7bn. Therefore, in that year, investments in data transformation and data-based knowledge creation were around the same as those in artistic (Film, TV & Radio, Literary, Musical and Art) originals, 38% of investments in R&D and 44% of investments in design. Comparing to tangible assets, we estimate that investments in data were slightly higher than those in ICT hardware (£5.4bn). We also estimate that around (4.3/23.4=)18% of measured investments in ‘software and databases’ (P²S²MEAS) are actually investments in data, with an additional £1.4bn of investment previously uncounted. Thus data accounted for (5.7/(23.4+1.4)=)23% of ‘total’ or ‘adjusted’ investment in computerised information in 2010. However, the growth rate in data investments means that proportion may be growing. It is anticipated that investments in data transformation and data-based knowledge will grow rapidly in the coming years. However we note that at this point in time, our estimates of investment are relatively low in comparison to some other knowledge-based assets. Theory tells us that, at the margin, firms will invest up to the point where marginal benefits equate to marginal costs. We believe this is worth bearing in mind when considering the statements that are being made around the contribution of (big) data to growth in output and productivity.

For future work we require a time-series of investment. The dataset we use to estimate employment only provides a snapshot of data in 2010. The data is a snapshot in the sense that whilst we identify workers with Big Data skills, we have no information on when they acquired those skills. Other estimates of Big Data employment (e-skills UK 2013b; Mandel and Scherer 2014) also provide snapshots for 2013 and 2014 respectively. However, whilst we do not have a time-series for Big Data employment, e-skills UK (2013a; 2014) do provide a time-series for Big Data vacancies. Whilst in general we may expect the relationship between employment and vacancies to be relatively stable, it is unlikely to be so in an emerging and growing field such as Big Data. In particular we would expect there to be a lag such that it takes employment a number of periods to adjust to any growth in vacancies. We therefore use cyclical data on vacancies and employment for the wider economy to estimate the following regression:

$$\Delta \ln N_t = 0.135 \Delta \ln V_t - 0.34 \ln N_{t-1} + 0.0052 \ln V_{t-1}$$

(4)

Which implies that the growth rate in employment can be estimated as 13.5% of the growth rate in vacancies. Combining this parameter with data on the growth rate in vacancies and growth rates derived from other e-skills forecasts of Big Data employment, we extend our employment estimates to cover the period 2008 to 2013. We also have a time series for the (weighted) wage of Big Data

---

workers. Applying that, and the factors summarised in Table 1, enables us to generate a short time-series of investment (2008-14). We extend that estimate back further using growth in measured investment in own-account computerised information. Figure 2 presents a time-series for investment in data, from 1997 to 2013, with total investment estimated as having risen from £2.7bn in 1997 to £7.1bn in 2013.

**Figure 2: Nominal UK investment in D and N assets, £bns, 1997-2013**

![Time-series for investment in data](image)

Source: authors calculations

Note to figure: Time-series for investment for each source estimated using growth rates implied by growth rates in Big Data employment (e-skills UK 2013b), Big Data vacancies (e-skills UK 2013a) and own-account software investment.

We noted above that investments in data-based information and knowledge (D and N) are made up of components for purchases and in-house production. We attempt to circumvent the lack of data on purchases by directly observing all asset production regardless of whether it is produced for sale or own final use. This is fine provided we do not count sales/purchases as well as in-house production which would be double counting. We therefore make the implicit assumption, we think reasonably, that the current design of the ONS CAPEX survey means that very few purchased investments in data (D) are currently recorded in the official measures for purchased investments in computerised information. However, of the £5.7bn of investment identified in 2010, some proportion will be produced within the D and N industries, for final sale to the wider market sector, and some will be produced in outside industries for in-house use. Although we cannot identify purchases separately, we do have the additional information from e-skills UK (2013b) that of big data activities, 89% are conducted in-house, and 11% are purchased. Using these proportions to split estimates of investment,

Note the ONS CAPEX survey is not designed to include purchases of N (data analytics) services.

See Appendix 1 for more detail on the ONS CAPEX survey. In light of the latest revision to the SNA, the survey is to be re-designed in order to better estimate investment in databases, which are to be reported separately from investments in software.
we can form estimates of the amount that takes place in the D and N industries, and the amount that
takes place in-house in outside industries. The estimates are summarised below in Table 3.

Table 3: Estimated investment in the D and N industries and in outside industries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5.70</td>
<td>0.63</td>
<td>5.07</td>
</tr>
</tbody>
</table>

Note to table: Estimates of investment in production of data-based assets that occurs in the D and N industries for sale to the wider economy, and in in-house investment in outside industries, based on the information that 11% of big data activities are outsourced/purchased. Thus column 1 is total UK market sector investment in 2010, column 2 is 11% of estimated investment which we allocate to the D and N industries and estimated as sold to the wider market sector. Column 3 is the remainder of investment, corresponding to in-house/own-account activity in outside industries, and is column 1 minus column 2.

How do these estimates compare to the turnover numbers in industries 62012 and 63110 which we consider to include the D and N industries (see Appendix 2)? In Appendix Table A2.2 we show that in 2010, total sales in these two industries were £18.5bn. We therefore estimate that £0.6bn, or 3.5%, of these sales were of D and N assets. The remainder of sales were for a variety of other activities, as reported in Table A2.1.

5.2. Mark-ups

From (1) it is clear that the product mark-up is a key parameter in estimating the value of upstream output, in either the data-building or knowledge creation sector. However, little is known about actual values of \( \mu \), either in the context of data and data analytics, or indeed other intangible assets such as scientific R&D, design, artistic originals or brand equity. The reason is that, with some exceptions, most intangibles are created on a firms’ own-account (i.e. the upstream intangibles sector is situated in-house). Therefore little data on actual market transactions exists, making comparisons between the value of output and its cost of production difficult. Further, firms are understandably reticent to share any information on the mark-ups they expect to earn on data and other knowledge-based investment activities.

What work has been done on estimating mark-ups for intangible assets has mostly been in the context of R&D. Hulten and Hao (2008) estimate a mark-up for the additional profit earned by R&D assets using the share of R&D in current expenditure, with the share used to allocate a proportion of operating surplus to R&D. Using data for six pharmaceutical companies, they estimate a mark-up of 1.5 for the year 2006.

\[ \text{Note, these estimates are based on data for the supply-side. If sales of D and N incorporate mark-ups, then estimates from the demand-side (i.e. based on sales or purchases) would be higher.} \]
In the US R&D satellite account, the costs of R&D exchanged between R&D establishments classified in a different industry to the parent/owner firm are also marked up (Robbins and Moylan 2007). The mark-up is estimated using the ratio of net operating surplus to gross output for miscellaneous professional, scientific, and technical services, which for the US averages about 0.15, implying an average mark-up of 1.15 (Corrado, Goodridge et al. 2011).

In work on estimating UK investment in artistic originals, Goodridge (2014) includes estimates of investment in music originals based on the revenues earned by those assets. Invoking assumptions of steady-state conditions, aggregate revenues earned by assets equate to the value of investment. In that work, UK investment in music originals in 2008 was estimated as £1,331m. For the same year, a cost-based approach yielded an estimate of investment of £224m, implying an innovator mark-up of \( \frac{1,331}{224}=5.9 \).

A similar approach can be taken to estimating a mark-up for television broadcasting originals. ITV is a UK commercial broadcaster that earns revenues from the sale of advertising carried on its television broadcasts. An approximate mark-up for ITV originals can be estimated using data on ITV costs of television production and the revenues generated through the sale of advertising. Data from OFCOM (2013) show that in 2012 ITV costs of production were £814m. Data from the ITV Annual Report show that net advertising revenues were £1,510m (ITV 2013), implying a mark-up of \( \frac{1,510}{814}=1.86 \).

In the context of data-building and data analytics, we are not aware of any work that has sought to explicitly estimate the value of mark-ups. We do note however that some companies explicitly value their data assets and in some cases include that value on their balance sheets (SAS). SAS report that 20% of large UK companies assign a financial value to data assets on their balance sheets. This is a nominal value of the stock of transformed data (in terms of the notation above, \( \Sigma P^D_B \)). This value, set against estimates of the accumulated costs of production of all transformed data assets, would provide an implicit estimate of the mark-up earned by information assets (\( \mu^D \)). Unfortunately we have not been able to acquire data that would allow us to estimate D and N sector mark-ups. We also note that the UK national accounting convention for intangible assets (i.e. software, R&D, most artistic originals) is to implicitly assume \( \mu=1 \). We therefore take a conservative approach in our empirical work and similarly assume that product mark-ups in the D and N sector equal one.
5.3. Overlap with measured R&D employment and investment

So far we have focused on the links between investments in data and official measured investments in software. A further comparison that we have not yet made is that between estimated employment/investment in D/N activity, and that in other knowledge creation activities aside from software. The ONS Business Enterprise Research and Development (BERD) release\(^{31}\) states that in 2010, UK R&D employment was 154,000 (see Table 4), consisting of 87,000 scientists and engineers, 41,000 technicians and 27,000 administrative staff. If we take the sum of the first two of those occupations, then our estimates suggest that big data employment is \((190,000/128,000=)148\%\) of UK R&D employment. If we also incorporate clerical staff, it is 123\%. Considering the attention devoted to R&D, these are clearly significant estimates. We do note however that R&D is largely concentrated in manufacturing\(^{32}\) whilst data activities are likely to be more dispersed across industries, potentially being a feature of any firm/industry that generates, or has access to, raw records or information.

It is also worth considering another aspect of the BERD data. BERD expenditure data is broken down into various product groups (pharma, aerospace etc.). One of those product groups is “Computer programming and information service activities”. Of £17.1bn of UK R&D activity in 2012, £1.9bn (or 11\%) occurred in this product group. BERD also provides figures for R&D employment in this product group (see Table 4), which stood at 22,000 in 2010 (note this is comprised of 11,000 scientists and engineers, 8,000 technicians and 3,000 clerical staff). BERD data for all business R&D and for R&D in computer programming and information and service activities are presented below.

Accessed on 18th September 2014.

\(^{32}\) Table 27 of the BERD release shows that, in 2012, of the £12.4bn of R&D that occurred outside of the R&D industry, £6.7bn (54\%) took place in manufacturing.
Table 4: UK Business R&D employment

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>Of which:</th>
<th>Of which:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Scientists and engineers</td>
<td>Technicians, laboratory assistants and draughtsmen</td>
</tr>
<tr>
<td>2000</td>
<td>145</td>
<td>86</td>
<td>30</td>
</tr>
<tr>
<td>2001</td>
<td>152</td>
<td>93</td>
<td>28</td>
</tr>
<tr>
<td>2002</td>
<td>158</td>
<td>96</td>
<td>31</td>
</tr>
<tr>
<td>2003</td>
<td>155</td>
<td>99</td>
<td>28</td>
</tr>
<tr>
<td>2004</td>
<td>150</td>
<td>94</td>
<td>27</td>
</tr>
<tr>
<td>2005</td>
<td>146</td>
<td>94</td>
<td>25</td>
</tr>
<tr>
<td>2006</td>
<td>147</td>
<td>92</td>
<td>27</td>
</tr>
<tr>
<td>2007</td>
<td>158</td>
<td>90</td>
<td>35</td>
</tr>
<tr>
<td>2008</td>
<td>151</td>
<td>86</td>
<td>37</td>
</tr>
<tr>
<td>2009</td>
<td>151</td>
<td>85</td>
<td>40</td>
</tr>
<tr>
<td>2010</td>
<td>154</td>
<td>87</td>
<td>41</td>
</tr>
<tr>
<td>2011</td>
<td>159</td>
<td>90</td>
<td>42</td>
</tr>
<tr>
<td>2012</td>
<td>161</td>
<td>91</td>
<td>44</td>
</tr>
<tr>
<td>2013</td>
<td>178</td>
<td>98</td>
<td>52</td>
</tr>
</tbody>
</table>

Source UK BERD, ONS

We noted above that the analysis of data and the creation of data-based knowledge meets the definition of R&D in the context of the national accounts and the Frascati Manual (OECD 2002). In particular we conjecture that some part of R&D in “computer programming and information service activities” may include investments in data analytics. Unfortunately however we currently have no information on how much of such activity may be included in the BERD data. The guidance notes to the BERD survey do state that “consumer surveys, advertising and market research” and “general purpose or routine data collection” are to be excluded from R&D figures provided, but the potential for some data analytics activity to be included does remain. We note that our occupation-based estimates of investment include contributions from research and development managers (1137), science professionals (211), engineering professionals (212) and research professionals (232), all of which have the potential to be included in the R&D data as measured in BERD. Excluding those occupations from estimation would result in an estimate of investment of £5bn, compared to £5.67bn with them included. The reason is that we apply a time-use assumption of between 0.1 and 0.2 for these occupations, whereas the implicit time-use parameter in R&D estimation is equal to one. The impact of excluding these occupations would therefore be small.

Additionally, official estimates of R&D investment are already adjusted to avoid double-counting with investments in software/computerised information (Ker 2014). In particular, 100% of expenditures on R&D in the field of “computer programming, consultancy and related activities” made in the industry of the same name (SIC 62) are excluded, and 37.5% of expenditures in that field but outside the industry are also excluded from R&D investment.
Therefore, in estimating investment in D and N activity, one option is to subtract (some proportion of) measured R&D in “Computer programming and information service activities” from investments in data, to eliminate any risk of double counting. Alternatively we may consider that the time-use assumption used, and the adjustments made by ONS, already largely eliminate the risk of double counting and so make no additional adjustment. We take this latter approach.

5.4. **Split between D and N activity**

The framework we presented earlier considered investment as composed of data-building (D) and knowledge creation (N) activity. However, the approach we have taken to measurement, and the fact that some workers in this field (particularly software professionals and the like) typically work on both D and N activities, means that we cannot separate investment in data into its respective components. We do however have some additional information from our discussions with industry experts. Initially one might expect the costs in the D stage to be relatively low, particularly in cases where raw records are generated automatically for free (or almost free). However, firms devote a lot of resources to the management of their data, in particular to integrate it with other data sources and also to improve its quality, for instance, in terms of consistency or the removal of duplicates. In our discussions it has been put to us that the “acquisition” or data-building process actually represents around 60-80% of the total costs of producing data-based knowledge. If we take a central estimate of 70%, we can split our estimate such that approximately: \( P^D = £4bn \) and \( P^N = £1.7bn \) in 2010.

6. **Conclusions**

This paper has set out a framework for the measurement of investment in big data and data analytics. In it we define data, information and knowledge. As part of our measurement, we show how some investments in data are already measured in GDP, but some are missed. We then use the framework and data from various sources to derive estimates of investment in data (transformation and analytics). We estimate that in 2010, UK market sector investment in data was around £5.7bn. We estimate that three quarters of that (£4.3bn) is already recorded in official estimates of investment in computerised information (software and databases). The remaining quarter represents previously unidentified investments not currently recorded in the national accounts, although we note the potential for some of it to be included in R&D, but any overlap is considered to be small. In future work we shall use our estimates of investment in a growth-accounting decomposition, to form a measure of the contribution of data to UK growth.
References

Bollier, D. and C. M. Firestone (2010). The promise and peril of big data, Aspen Institute, Communications and Society Program Washington, DC, USA.


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


SAS "SAS High-Performance Analytics: Transforming Big Data into Corporate Gold."


Appendix 1: Purchases of databases in official capital expenditure survey (CAPEX)

The most direct method for measurement of purchases of data (D) or data-based knowledge (N) would be to ask companies for their purchases of D and N via capital expenditure surveys (CAPEX). However, in official measurement of purchases of ‘computerised information’, the actual question on CAPEX only asks firms to report the expenditures incurred on software (see Figure A2.1).

Figure A2.1: ONS CAPEX Survey: Question

Source: Excerpt from ONS capital expenditure survey (CAPEX)

The guidance notes at the back of the survey state: “Include all expenditure on computer software to be used for more than one year. This includes the purchase or development of large databases and licence fees for the use of software.” Therefore, in practice, with the only reference to databases being in the guidance notes, and with the restriction to large databases, it is unclear how many purchased investments in data actually make it into official estimates of GFCF. In light of the modified treatment to databases in the 2008 revision to the SNA, the CAPEX survey is to be revised, and firms asked to report expenditure on software and databases separately. Until then however it is not possible to split official estimates of purchases of ‘software and databases’ into their two respective components.

We also note that to the extent that purchases of databases are captured in the official data, in our framework they refer to the purchase of D, (transformed) data (or information). They will not include the purchases of knowledge (N) gleaned from data analytics, by specialist data analytics providers. As data analytics becomes more widespread, we would expect there to be growth in specialist firms which sell either: knowledge acquired from data; or data itself; to firms in other industries. On the latter, these firms could include those situated in positions to capture the large flows of data generated in transactions, for example, firms that interface with large volumes of consumers; are involved in the processing of transactions; or who enable global supply chains (Manyika, Chui et al. 2011).

To summarise, it appears that in the official data based on CAPEX survey responses, purchases of data (information) are likely not well captured, and purchases of data-based knowledge not included at all.
Appendix 2: D and N industries in the Standard Industrial Classification (SIC)
We have noted that investments in data/information (D) and data-based knowledge (N) may be in the form of direct purchases or own-account (in-house) asset production. Ideally we would wish to measure purchases of D and N directly. As noted in OECD (2013), a number of specialist firms and entities are entering this arena, with new developments including: “data lockers” that enable users to allow access to their data in exchange for some portion of revenues earned from its use; the emergence of “data brokers” (e.g. Experian, LexisNexis, Intelius, Locate Plus) who process data and sell it on, and could include either specialist data providers or firms who collect personal data and sell that data to other firms (for instance, airlines, retailers, mobile phone networks and internet service providers (ISPs) are examples of organisations that collect, and are in a position to sell, data on their customers); specialist data analytics firms whose services include the sale of refined customer profiles aggregated from a variety of sources e.g. public data, proprietary data and data from institutional research; and “data exchanges” where firms can bid for data, for example, the data exchange “BlueKai” offers advertisers the opportunity to bid on data on web surfing activity (OECD 2013).

A2.1 Data from the size of the D and N producing sector
There are two types of purchase we need to consider:
1) The D sector: purchases of transformed data (information) assets
2) The N sector: purchases of commercial data-based knowledge assets

Since these are sales of assets to the rest of the market sector, they must be produced by specialist firms located somewhere in the Standard Industrial Classification (SIC). One option therefore is look at the sales data for the D and N industries. Inspection of the 2007 SIC reveals the following industries, shown in Table A2.1, whose activities are potentially relevant to data-building (transformation) or data analytics (knowledge creation).

Table A2.1: SIC07 Industries whose activities might include data-building and/or data analytics

<table>
<thead>
<tr>
<th>SIC (2007)</th>
<th>Industry Description</th>
<th>Activity Description</th>
<th>Where in our framework?</th>
</tr>
</thead>
<tbody>
<tr>
<td>62012</td>
<td>Business and domestic software development</td>
<td>Business and domestic software development</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Custom software development</td>
<td>Data-building (D) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data analysis consultancy services</td>
<td>Data-building (D) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Designing of structure and content of business and domestic software database</td>
<td>Data-building (D) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Made-to-order software</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Programming services</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Software house</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Software systems maintenance services</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>System maintenance and support services</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Systems analysis (computer)</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Web page design</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td>63110</td>
<td>Data processing, hosting and related activities</td>
<td>Batch processing</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data conversion</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data preparation services</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data processing</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data storage services</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Database running activities</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tabulating service</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time sharing services (computer)</td>
<td>Knowledge Creation (N) sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Web hosting</td>
<td>Knowledge Creation (N) sector</td>
</tr>
</tbody>
</table>

Note to table: Excerpt from the 2007 Standard Industrial Classification

However, we must also be careful not to double-count. It would be incorrect to count both the in-house production and the sale as GFCF.
Of particular interest are the two industries: Business and domestic software development (62012) and Data processing, hosting and related activities (63110). The third column lists the activities included in each industry. Highlighted in red are the activities we consider potentially part of either data-building (D) or knowledge creation (N) (indicated in final column).

Unfortunately, data are not available for each activity, column 3, in Table A2.1. Instead, the five-digit level of the SIC (as in Column 1) is the lowest aggregation available. From Table A2.1, we can assume that some part of the sales of industry 63110 relate to data-building. We can also assume that some part of industry 62012 relates to knowledge production, and another part to data building. Industry data for SICs 62012 and 63110 are presented in Table A2.2.

### Table A2.2: Annual Business Survey (ABS) data

<table>
<thead>
<tr>
<th>Standard Industrial Classification (Revised 2007)</th>
<th>Description</th>
<th>Year</th>
<th>Number of enterprises</th>
<th>Total Turnover</th>
<th>Approximate gross value added at basic prices</th>
<th>Total purchases</th>
<th>Total employment - average during the year (1)</th>
<th>Total employment costs</th>
<th>Total net capital expenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.01/2 Business and domestic software development</td>
<td>2008</td>
<td>18,323</td>
<td>13,681</td>
<td>6,712</td>
<td>6,976</td>
<td>107</td>
<td>4,614</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>11,197</td>
<td>12,899</td>
<td>6,928</td>
<td>6,126</td>
<td>82</td>
<td>4,128</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>15,653</td>
<td>12,859</td>
<td>7,355</td>
<td>5,556</td>
<td>102</td>
<td>3,888</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2011</td>
<td>22,085</td>
<td>14,889</td>
<td>8,077</td>
<td>6,017</td>
<td>108</td>
<td>4,390</td>
<td>299</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>27,147</td>
<td>15,562</td>
<td>9,319</td>
<td>6,329</td>
<td>109</td>
<td>4,771</td>
<td>273</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{\textsuperscript{*}} 63.110 \quad \text{Data processing, hosting and related activities}
\]

<table>
<thead>
<tr>
<th>Year</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>2,850</td>
<td>3,447</td>
<td>1,789</td>
<td>38</td>
<td>1,633</td>
</tr>
<tr>
<td>2009</td>
<td>2,850</td>
<td>3,447</td>
<td>1,343</td>
<td>*</td>
<td>1,409</td>
</tr>
<tr>
<td>2010</td>
<td>2,783</td>
<td>3,711</td>
<td>1,882</td>
<td>*</td>
<td>1,618</td>
</tr>
<tr>
<td>2011</td>
<td>2,996</td>
<td>4,220</td>
<td>2,168</td>
<td>*</td>
<td>1,621</td>
</tr>
<tr>
<td>2012</td>
<td>3,038</td>
<td>4,221</td>
<td>2,438</td>
<td>*</td>
<td>1,643</td>
</tr>
</tbody>
</table>

Let us start with industry 62012, whose turnover is about £15.5bn in 2012. As Table A2.2 shows, some of that industry output would be of data related activity, but some will be the sale of either pre-packaged or custom software to domestic and non-domestic firms and households. Figure A2.1 shows industry sales of 62012 and also UK GFCF in purchased software. By subtracting the latter from the former, we can derive a potential upper bound for the sales of data-building and data analytics services from industry 62012, of around £2.2bn in 2011 and £0.55bn in 2012, and averaging £1.5bn over the years 2008 to 2012.\footnote{We note that this is very much an upper bound since it makes no allowance for household purchases of software, which would reduce the total. Adjusting for international trade may also reduce this estimate if UK net trade in data is positive (i.e. if exports are greater than imports).}
Figure A2.1: Domestic software sales and UK purchased software investment

Note to figure: Red line is sales of industry 62012, taken from the Annual Business Survey. Green line is whole economy UK GFCF in purchased software. Purple line is sales of 62012 less UK purchased software GFCF which can be viewed as an upper bound of D and N asset sales from industry 62012.

Returning to Table A2.1, industry 63110 predominantly includes activities that we model as D sector production, as well as some additional services: as the table shows, turnover is around £6.6bn in 2012: Adding this number to that derived for industry 62012, as in Figure A2.2, therefore provides an upper bound on the purchases of D and N activities of around £8.7bn in 2011 and £7.2bn in 2012, and averaging £7.2bn over the years 2008 to 2012.

Figure A2.2: Approximate upper-bound for purchases of data (D) and data analytics (N), current prices (£mns)

Note to figure: Blue line is sales of 62012 less purchased investments in software. Red line is turnover of 63110. Green line is the sum of these two series’.

While these figures provide some indication of the upper limit to purchases of data-based assets, the estimates also include a range of other activity that we do not wish to include e.g. web page design, web hosting etc. Unfortunately lower-level industry data is not available. Note however that there is an additional complication: if we were to estimate purchased investments in D/N using data on sales of the D/N industry, then we would need to ensure that own-account investment in industry D/N is
excluded to avoid double counting. Therefore by estimating all own-account investment, including that within the “data industry”, we can circumvent the problem of a lack of data on purchases. Further, by directly estimating data transformation and knowledge creation activity, we can ensure that our method meets OECD recommendations around the non-produced nature of data and the types of expenditures that ought to be recorded as investment (in particular the requirement that purchases of raw records are excluded as they are payments for a non-produced asset). We do note however that were estimates of purchases available then they would incorporate the product mark-ups, $\mu^D$ and $\mu^N$, which the own-account estimates will not without some additional information.

---

35 For example, in the measurement of investment in software, own-account investment in the software industry is excluded, as it is assumed that this investment forms assets that are then sold to firms making purchased investments in either pre-packaged or custom software.