the review of income and wealth

Review of Income and Wealth
Series 68, Number 4, December 2022
DOI: 10.1111/roiw.12542

WE SEE DATA EVERYWHERE EXCEPT IN THE PRODUCTIVITY STATISTICS

BY PETER GOODRIDGE*

The Productivity Institute, The University of Manchester

JONATHAN HASKEL

Bank of England, Imperial College Business School, CEPR and IZA

AND

HARALD EDQUIST

Ericsson Research and Research Institute of Industrial Economics

This paper uses Labor Force Survey data for European countries to estimate national investment in data assets, where the asset boundary is extended beyond that for software and databases as currently defined in the System of National Accounts. We find that: (a) in 2011–2018, 1.4 percent of EU-28 employment was engaged in the formation of (software and) data assets, with a mean growth rate of 5 percent per annum (pa); (b) on average in 2011–2016, expanding the asset boundary raises the level of own-account GFCF in software and databases in the EU-16 by 61 percent, and mean growth in real investment in own-account software and data assets to 6.9 percent pa, compared to 2.7 percent pa in national accounts; (c) in 2011–2016, expansion of the asset boundary raises labor productivity growth in the EU-13 from 0.79 percent to 0.83 percent pa, and the contribution of software and data capital deepening over three-fold, from 0.03 percent to 0.10 percent pa.

JEL Codes: O47, O57, O32

Keywords: data, information, software, investment, productivity

1. INTRODUCTION

Critical aspects of the latest stage in the ICT revolution include the Internet of Things (IoT) and artificial intelligence (AI). Related is rapid growth in a vast

Note: Funding from: Ericsson AB (556056-6258); EPSRC (Releasing the Value of Big Data, EP/K039504/1, EP/K503733/1 (Jonathan Haskel was a co-investigator), Accelerator Grant for investigation into Effect of Telecoms prices on Economic Growth ES/M500562/1); and ESRC (Programme on Innovation and Diffusion, ES/V009478/1 and The Productivity Institute, ES/V002740/1). We thank Josh Martin of the ONS and Melker Loberg of Statistics Sweden for sharing data and methodological detail on estimates of investment in software and databases in the UK and Sweden respectively. We also thank Eurostat for supplying data extracted from the European Labor Force Survey (LFS) and representatives of various national statistics offices for sharing information on the proportion of GFCF in software and databases produced for own final use in their respective countries. Finally, we thank an editor and three anonymous referees for very helpful comments which improved the paper. Errors and opinions are our own and not those of our affiliated institutions.

*Correspondence to: Peter Goodridge, The Productivity Institute, Alliance Manchester Business School, The University of Manchester, Booth ST W, Manchester. M15 6PB, United Kingdom (peter.goodridge@manchester.ac.uk).

collection of data/information available for analysis and the advancement of knowledge.[1] Eric Schmidt is quoted as saying that as much data is created every two days as was from the dawn of civilization to 2003 (Wong, 2012). IDC (Gantz and Reinsel, 2010) project that between 2009 and 2020 the volume (stock) of data created will have grown over forty-fold.

Analysis of data is not a new activity. Firms, other agents and humankind have long sought to acquire knowledge from information. What has changed is the scale and ubiquity of that activity. Developments in ICT mean that far more data is being created (captured), stored, combined and aggregated in more systematic ways; with greater opportunities for richer, more complex, analysis.

Acquisition of knowledge from data generates an economic return in the form of higher revenues or lower costs. Therefore, provided they meet asset criteria of repeated contribution to production over more than one year, the case for considering data and the knowledge acquired from data as assets is clear. Databases have been classified as an asset in the System of National Accounts (SNA) (United Nations, 2008) since 1993.

Goodridge *et al*. (2015) and Goodridge and Haskel (2015a, 2015b) document fast growth in UK employment engaged in the transformation and analysis of data, and thus in UK investment in data assets and their contribution to growth. We seek to build on that work. Our two primary contributions are that we: (a) first, present harmonized estimates of investment in data assets based on an expanded asset boundary across EU-28 countries[2]; and (b) second, we estimate the contribution of data capital deepening to growth in productivity.

In general we find that over half (57 percent) of employment engaged in software and data capital formation is already accounted for in national accounts measurement of own-account investment in software and databases. The remainder is largely outside the asset boundary as currently defined in the SNA. That element outside the current asset boundary is growing faster than the national accounts measure in a number of European countries.

Our more detailed findings are as follows. First, in 2011–2018, 1.4 percent of EU-28 employment was engaged in the formation of own-account software and data assets. Second, in the EU-28 in 2011–2018, mean growth in employment engaged in own-account software and data capital formation was 5 percent per annum (pa). Third, combined with estimates of wage and non-wage costs, our estimates imply that extending the asset boundary raises EU-16 investment in own-account software and data assets by around 61 percent on average in 2011–2016. Fourth, in the EU-16 in 2011–2016, mean growth in newly defined real investment in own-account software and data assets was 6.9 percent pa, compared to 2.7 percent pa in national accounts. Fifth, in the context of growth-accounting, incorporating our expanded definition of investment changes both output and input. In the EU-13 in 2011–2016: (i) labor productivity growth is raised from 0.79 percent

[1] Connections between items of capital equipment and the wider internet (IoT) use data and create ("exhaust") data as a by-product. AI technologies provide new and powerful methods to enhance the analysis of data.

[2] This paper was written before the UK exited the European Union. References to the EU-28, EU-16 and EU-13 therefore include the UK.

pa to 0.83 percent pa, which translates to €6.7bn pa of additional output growth in 2016 if applied to the EU-28 aggregate[3] and (ii) the contribution of capital deepening in software and data assets is raised over three-fold, from 0.03 percent pa to 0.1 percent pa, which translates to €9.4bn pa in 2016 if applied to the EU-28 aggregate.

In the next section we review the relevant literature. In section three we review the treatment of data and databases in the SNA and implications for measurement. In section four we define our framework to show how data relates to information and knowledge and discuss it in the light of the SNA. In section five we document employment engaged in (own-account software and) data capital formation in the EU-28 and construct new estimates of investment in the EU-16. In section six, we estimate the economic impact of expanding the asset boundary for software and data on growth-accounting measures for the EU-13[4] and provide a range of sensitivity analyses to the underlying assumptions. Section seven concludes.

2. Existing Literature

The terms data, information and knowledge; are often used interchangeably. The framework in Goodridge *et al.* (2015) and Goodridge and Haskel (2015a, 2015b) seeks to give them a more precise context, drawing on definitions from the information science and economics literature, also summarized in Rassier *et al.* (2019).

Ackoff (1989) introduced the data-information-knowledge-wisdom (DIKW) hierarchy defining: data as symbols representing properties; information as processed data containing descriptions, where processing serves to increase usefulness; knowledge as conveyed by instructions; and wisdom as ability to increase effectiveness. The transmission of information to knowledge is described as analytic thinking.

Dictionary definitions refer to data as: (i) quantities and symbols; or (ii) information (Statistics Canada, 2019a). One of the descriptions in the online Merriam-Webster dictionary defines data as something “that must be processed to be meaningful” (Statistics Canada, 2019a). Statistics Canada (2019a) define data as “observations that have been converted into a digital form that can be stored, transmitted or processed and from which knowledge can be drawn.” In their framework, observations are an intangible raw material and data are something that is digitized, stored and can be analyzed.

Shapiro and Varian (1998) define information as digitized data. Boisot and Canals (2004) distinguish between data and information and argue that information

[3] Estimated as 0.05 percent (0.05 percent rather than 0.04 percent, due to rounding) multiplied by EU-28 nominal value-added in 2016 (€13,397 bn), where EU-28 value-added is the sum of PPP(GDP)-adjusted measured nominal value-added in EU-28 countries.

[4] Note, our estimates of employment are for the EU-28, estimates of investment are for the EU-16, and growth accounting analyses are for the EU-13. The reason is that we have estimates of employment for all EU-28 countries. However, we are only able to construct estimates of national accounts own-account investment in software and databases for EU-16 countries. Additionally, growth-accounting data in EUKLEMS is incomplete for some countries so we are only able to conduct that analysis on EU-13 countries.

is regularities in data which agents attempt to extract (at a cost). Following Arrow (1984) they define knowledge as a set of expectations modified by new information. Bakhshi *et al.* (2014) argue that to generate value, raw data must be processed and structured into: (a) information, defined as meaningful statements about the state of the world; and (b) knowledge, defined as models of relationships between variables, which can be used to inform action.

In the economics literature, definitions of knowledge use terms including instructions, ideas, recipes or blueprints (Romer, 1990, 1992; Jones, 2005). Fransman (1998) notes different knowledge can be formed from the same information suggesting information can be used repeatedly in the formation of knowledge.

Mokyr (2003) argues that knowledge exists in the human mind implying that knowledge constitutes an understanding of information. Mokyr (2003) also introduces a distinction between propositional and prescriptive knowledge. The former catalogues natural phenomena and regularities and includes knowledge of nature, properties (i.e. science) and geography. Prescriptive knowledge has some base in propositional knowledge but prescribes actions for the purposes of production and so can be thought of in terms such as recipes, blueprints, techniques and instructions. Romer (1990) also distinguishes between basic and commercial knowledge.

The R&D literature distinguishes been basic and applied R&D according to features described by Romer (1990) and Mokyr (2003) including the property of excludability. Basic knowledge is freely available to all agents (calculus or economic theory for example) but applied (commercial) knowledge, produced or acquired by firms, is not. Mokyr (2003) notes the two are linked. Commercial knowledge can derive from freely available knowledge and in turn can feed back and enhance the epistemic base, creating a positive feedback between science and innovation.

The justification for treating data, and the knowledge acquired from data, as capital is inherent in the intangibles literature, which follows on from the seminal work of Corrado *et al.* (2005, 2009) and applied by Marrano *et al.* (2009) and Fukao *et al.* (2009). Similarly, Jones (2005) refers to a "stock of knowledge or ideas." It is also present in earlier work. Machlup (1962) makes the case for considering knowledge accumulation as capital formation, correctly noting that the defining feature of investment is the devotion of current resources to future productivity gain.

The treatment of data and databases in the SNA is summarized in Ahmad and Van De Ven (2018) and Ahmad (2004, 2005a, 2005b).[5] Van De Ven (2017) notes that the SNA has changed dramatically over the past decades and will continue to change as a consequence of changes in the environment of producing statistics. Increased creation and use of data is one such change. SNA recommendations and their implications are discussed further in the next section.

Estimation of national accounts gross fixed capital formation (GFCF) is the domain of national statistics authorities. Statistics Canada (2019a) discuss the conceptual case for capitalization of data, databases and data science, using a wider asset boundary than currently applied to databases in the SNA, and present estimates for Canada (Statistics Canada, 2019b).[6] They present an information value

[5] Available at: https://unstats.un.org/unsd/nationalaccount/aegm.asp.
[6] Discussed in more detail in Appendix A.

chain similar to the data value chain presented in Goodridge and Haskel (2015a, 2015b) and this paper. The US Bureau of Economic Analysis (BEA) have conducted preliminary work on extending the asset boundary for data assets beyond SNA recommendations in US national accounts (Rassier *et al*., 2019). In the UK, the Office for National Statistics (ONS) have developed estimates of own-account GFCF in software and databases to better capture the output of employees engaged in database investment (McCrae and Roberts, 2019), based on occupations identified in Goodridge *et al*. (2015) and Goodridge and Haskel (2015a, 2015b).

3. Data and Databases in the System of National Accounts (SNA)

In the SNA (United Nations, 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."* The SNA distinguishes between produced and non-produced assets. Produced assets are those generated as output from production processes that fall within the production boundary and include intellectual property products (IPPs): *"the result of research, development, investigation or innovation leading to knowledge that the developers can market or use to their own benefit in production."*

The 1993 SNA update recommended capitalization of software and "large" databases, with databases considered a special case of software, defined as: *"files of data organized in such a way as to permit resource-effective access and use of the data."* OECD (2010) recommend 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."* The 2008 SNA update clarified capitalization criteria and dropped the restriction to large databases. Those clarifications are described in Ahmad and Van De Ven (2018) and Ahmad (2004, 2005a, 2005b), with the key point being that the SNA distinguishes sharply between data and databases.

Like other assets, particularly IPPs, databases may be developed: (i) exclusively for own final use; or (ii) for sale as an entity or by means of a license.[7] In the case of databases developed for own final use (own-account capital formation), the SNA recommends the sum of costs method of measurement, as used for own-account software and R&D. Given that databases are part-software and the difficulty in accurately distinguishing between employees working on software and database capital formation, the asset class in national accounts is "software and databases" and it is not separated into its two components.

According to the SNA, databases consist of two components: (a) the supporting software or database management system (DBMS), which provides or facilitates access to the data; and (b) embodied data. Databases and the underlying DBMS are an IPP and produced asset. However, in the SNA, data is a non-produced asset created outside the production boundary. Clearly there are examples where data is created as part of a production process. The reasoning is therefore partly pragmatic.

[7] With standard conditions applying for whether expenditures constitute capital formation i.e. economic benefits to owners and repeated use in production for more than one year.

It is argued that that if data (or information/knowledge in the form of data) were treated as a produced asset, that would open the door to capitalization of all forms of information/knowledge, including in written form or embedded in humans, swamping the accounts and reducing their meaning (Ahmad and Van De Ven, 2018).

As a consequence, measured fixed capital formation in databases is recommended to include just two components: (i) the cost of the DBMS, already recorded in software; and (ii) costs associated with preparation and transfer (including digitization) of data to the format/structure required by the database. As a result, national statistics agencies generally assume that database investment is largely captured by methods for measuring software investment.[8]

The SNA and OECD (2010) recommend that the value of embodied data and the costs of data acquisition be excluded from measured capital formation in databases. On embodied data, Ahmad (2005a) notes that attempts at valuation would be impractical as it would arguably be necessary to include all costs incurred in all business processes that generate data, but that the cost of transferring data to database format is at least meaningfully measurable. That is consistent with methods for measuring other own-account knowledge capital formation i.e. R&D GFCF does not include a valuation of knowledge discovered or embedded in R&D output.[9] However, the exclusion of costs of data acquisition seems at odds with the standard sum of costs approach taken in valuing other own-account capital formation (Rassier *et al*., 2019).

As a result of all the above, the recommended sum of costs approach (Ahmad, 2005a) reduces to:

Total number of employees working on converting data with an expected working life of more than one year from one medium/format onto that required by the database and on the DBMS application *

Average remuneration *

Proportion of time spent on these activities +

Other intermediate costs used in these activities (excluding any costs associated with data acquisition) +

Notional operating surplus related to these activities (costs of capital services, for example capital services of scanning machines and computers)

[8] Details on national methods to estimate capital formation in software and databases are provided in Appendix A for a number of European countries. We also include details of experimental estimates from Statistics Canada (2019a, 2019b), which go beyond SNA recommendations and extend the asset boundary to include capital formation in data and data science, as well as databases.

[9] Although we note that it does create an anomaly in the treatments of own-account and purchased capital formation. Database *purchase*s are valued at market prices, where prices will include a valuation of information content. The same anomaly exists for purchased R&D where prices will include some valuation of knowledge discovered. This is somewhat pragmatic as attempting to remove the value of embodied information (or knowledge in the case of R&D) from market prices would be impractical. However, Ahmad and Van De Ven (2018) state that the value of embodied information in database purchases ought to be recorded in the accounts under transactions in goodwill, a non-produced asset that is not a productive asset for the purposes of fixed capital formation, implying that database purchases ought to be excluded from fixed capital formation entirely.

where costs should include an estimate of the net rate of return to capital in database production and costs incurred in updating the database.[10]

A key aspect of data capital formation not explicitly addressed in the SNA and outside the SNA asset boundary for databases is activity in data analytics or data science, which is the process of creating or extracting knowledge from data. Goodridge and Haskel (2015a, 2015b) and Statistics Canada (2019a, 2019b) argue that data science clearly meets the SNA and Frascati Manual definition of R&D: *"the value of expenditures on creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and use of this stock of knowledge to devise new applications."* Ahmad and Van De Ven (2018) note that extending the sum of costs approach to include data science activity would result in a clear improvement to estimated investment.

Although data science is consistent with the definition of R&D, in practice it does not appear to be included in measured R&D capital formation. As far as we are aware, in most countries measurement of R&D is undertaken using surveys sent to R&D performers known to the statistics authority. This is the case in the UK and in Canada (Statistics Canada, 2019a). As a result, measurement is concentrated on known performers of traditional scientific R&D.

In the UK, survey respondents are explicitly asked about R&D activity in software development, but not in database development or data science/analytics. Since software expenditures are already capitalized, in the UK reported R&D activity in software development is excluded from measured R&D output (GFCF) to avoid double-counting. Firms are not provided with further guidance or definitions. If they consider their data science activities to be a form of R&D they may include them in their response, but if they classify those expenditures under software development, they will subsequently be excluded from R&D GFCF.[11]

Therefore, whilst data science meets the definition of R&D in principle it is, to the best of our knowledge, not included in measured R&D output in practice. Some data science activity will inevitably be unintentionally included in estimates of own-account investment in software and databases because occupations identified as engaged in the writing of software are also engaged in the transformation and analysis of data (Goodridge *et al*., 2015; Goodridge and Haskel, 2015a, 2015b). In some cases, the occupational classification is not sufficiently granular to distinguish between them, and in other cases the same employees undertake each activity. For this reason, we argue that investments in the extraction of knowledge from data are best considered in the context of software and data(bases), rather than R&D or a new asset category.

[10] Provided updates have a working life greater than one year.

[11] For there to be double-counting (i.e. for data science activity to be recorded as capital formation in both: (i) software and databases and (ii) R&D; occupations identified in measurement of software and database capital formation would need to include those engaged in data science, including mathematical, statistical, economic and other analytical occupations, as well as those typically associated with software creation. In reviewing the methods of European national statistics agencies (Appendix A), we show that such occupations are not typically included in measurement of own-account software and database capital formation. Exceptions are the UK and Germany, which do include some of these occupations. Additionally, the activity would need to be recorded as R&D capital formation. Our understanding of UK practice is that it is not, in that country at least.

4. Our Framework

In Figure 1 we present a simplified exposition of the data value chain, which uses concepts similar to those described in Statistics Canada (2019a, 2019b), Nguyen and Paczos (2020) and Mayer-Schonberger and Cukier (2013). The value chain is presented in three stages to summarize the production process for information and knowledge. Investment is defined as total costs incurred in the creation of information, defined as computerized analyzable data, and knowledge, defined as an understanding of (or insights from) information to be used in the production of goods and services.[12] Although presented linearly, feedbacks exist between stages. The three stages can exist in-house within the same vertically integrated firm or in distinct specialist firms.

4.1. *Data-Building/Transformation Stage*

The top of Figure 1 highlights the data-building or transformation process, which transforms observations (unprocessed data, which may or may not have a cost)[13] into computerized data of a usable and analytical format (information). Once transformed, data are a (produced) asset (information) and the production of information is capital formation. Data building (transformation) may involve digitizing, structuring, formatting, cleaning, aggregating or matching, and is sometimes referred to as: data management; data acquisition; data warehousing; or ETL (Extract, Transform, Load). The costs of transformation may be low or (close to) zero where the process is automated.[14]

4.2. *Knowledge Creation*

The second stage is the knowledge creation process, usually referred to as data analytics or data science. Other terms include: data/text mining; knowledge recovery; business intelligence; and machine learning; with the latter referring to the use of AI to discover correlations. This stage takes the information output of the data-building stage and uses it to conduct analysis. The output is knowledge created from analysis of information to be applied in final production. Different knowledge can be formed from the same information (Fransman, 1998) and information can be used repeatedly in the formation of knowledge. Again, direct costs may be low if the process is automated (e.g. machine learning).[15]

[12] Our framework for estimating GFCF in data (transformation and knowledge creation) is based on total costs incurred, which is consistent with methods for other intangibles in national accounts and those outside the SNA framework (Corrado et al., 2005, Goodridge et al., 2014). As with other intangibles, there may be a divergence (or mark-up) between the value of (unique) knowledge assets (and the capital services they generate) and the costs incurred in their creation. Ideally we would be able to observe their value in market prices. However, as these assets are typically created and used in-house, we instead estimate investment by observing costs of production.

[13] Where there is a purchase, the acquisition of data is an incurred cost, including where already transformed data are purchased for further transformation. They are an intermediate cost if used within the accounting period and a capital expenditure if used repeatedly over more than one accounting period (although we note that it may be difficult to distinguish for the practical purposes of measurement). Where observations are created as a by-product (e.g. exhaust data) of some other process, there is no incurred cost.

[14] Where automated, measured costs should include the nominal capital services of hardware and software used in automation.

[15] But measured costs should again include the nominal capital services of hardware and software (AI) used in extracting knowledge from information.

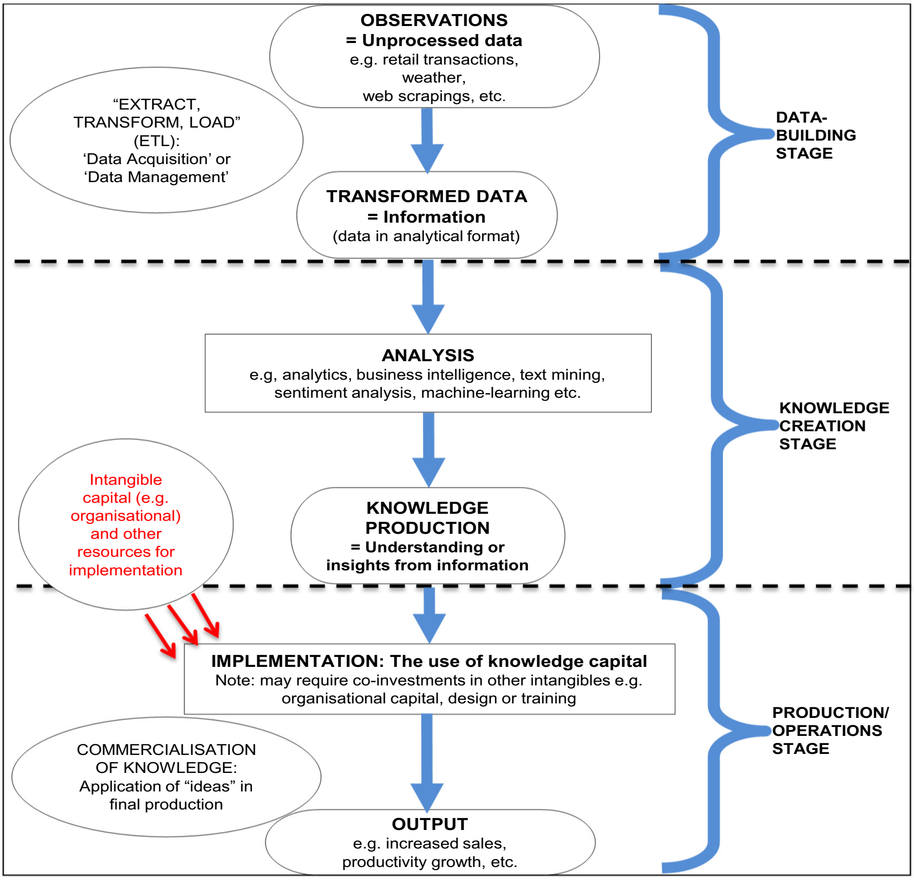
OBSERVATIONS
= Unprocessed data
e.g. retail transactions, weather, web scrapings, etc.
"EXTRACT, TRANSFORM, LOAD" (ETL):
'Data Acquisition' or 'Data Management'
TRANSFORMED DATA
= Information
(data in analytical format)
DATA-BUILDING STAGE
ANALYSIS
e.g, analytics, business intelligence, text mining, sentiment analysis, machine-learning etc.
KNOWLEDGE CREATION STAGE
Intangible capital (e.g. organisational) and other resources for implementation
KNOWLEDGE PRODUCTION
= Understanding or insights from information
IMPLEMENTATION: The use of knowledge capital
Note: may require co-investments in other intangibles e.g. organisational capital, design or training
PRODUCTION/ OPERATIONS STAGE
COMMERCIALISATION OF KNOWLEDGE:
Application of "ideas" in final production
OUTPUT
e.g. increased sales, productivity growth, etc.

Figure 1. Data Value Chain

Notes: Investment defined as costs incurred in the creation of information (computerized analyzable data) and knowledge (acquired from information). Commercialization is the embodiment of knowledge in final output. We use the term commercialization as our focus is on the market sector, but note that the framework can be applied to the non-market sector.

Source: Authors' representation (Goodridge and Haskel, 2015a, 2015b). "Raw records" are labeled here as "observations." [Colour figure can be viewed at wileyonlinelibrary.com]

4.3. *Downstream Production of Final Goods and Services*

The final stage incorporates the application or implementation of (produced) data-based insights in downstream production of final goods and services. Knowledge could be used to generate additional revenue or reduce costs. For instance, it could be a re-optimisation of processes based on knowledge derived from information formed from observations emitted by sensors embedded in tangible capital (Internet of Things). The downstream is an operations stage that does not undertake capital formation but employs/rents labor and (tangible and knowledge) capital to produce final goods and services. Implementation may require complementary applications or co-investments of other forms of knowledge capital, such as organizational (business process change), design or reputational (brand) capital. Information and data-based knowledge assets may also be used in the upstream creation of other types of knowledge capital.

4.4. *Implications for Measurement in Context of the SNA*

The framework suggests measuring two types of capital formation: the creation of information; and knowledge. We argue that the framework is consistent with the SNA in the following ways. First, measuring capital formation in information (i.e. the transformation of data) is an extension but consistent with SNA and OECD recommendations to measure the costs of data preparation, transfer and digitization. However, occupations identified as engaged in information capital formation ought to go beyond the software and database professionals typically observed in national accounts estimation, to include all occupations engaged in data transformation activity. Second, data-based knowledge creation (data science) meets the SNA and Frascati Manual definition of R&D, but, for reasons described above, is unlikely recorded in measured national accounts R&D GFCF in practice. We believe our framework adds clarity, particularly in distinguishing between observations (unprocessed data) and transformed data (information),[16] which are both simply "(embodied) data" in the SNA.

Like data in the SNA, in this framework observations do not constitute a fixed asset and so for practical purposes can be regarded as a naturally occurring non-produced good, typically generated as a by-product of some other process. Alternatively, if observations are considered produced, they may be regarded as an intermediate. Whether or not observations are produced is a difficult question, conceptually. However, our key argument is that data does not constitute an asset until it is transformed into a usable, analyzable, long-lived form that can repeatedly be used in the extraction of knowledge and contribute to production. Therefore, in this framework, generation of observations is not investment activity.

In contrast to observations (unprocessed data), information is digitized analyzable data and is a produced asset. Statistics Canada (2019a) note that the production process that creates analyzable information (data in their nomenclature) from observations (also observations in their nomenclature and also regarded as non-produced) implies that it is a produced asset. This overcomes the objection that capitalization opens the door to capitalization of all other forms of information, since this framework requires that information be produced (computerized, transformed, analyzable and long-lived). A paper-based telephone directory does not constitute information unless digitized and transformed. The fact that information assets are owned is another signal that they are produced (Statistics Canada, 2019a, 2019b). In this framework, knowledge extracted from data is also a produced asset.

Since data assets (information and knowledge) are typically created and used in-house, the appropriate method for measuring capital formation is costs incurred in creation. This can be done by adapting the SNA and OECD (2010) recommended sum of costs approach to identify all workers engaged in the production of information and knowledge. In the data transformation sector, this would encompass job titles including: data administrator; data manager; data engineer; data entry

[16] Statistics Canada (2019a, 2019b) make the same distinction.

and data control. Relevant job titles in the knowledge creation sector include: data scientist; data engineer; business intelligence; analyst; statistician; and economist. In practice, the occupational classifications used by statistics authorities are not granular enough to identify many of these job titles and many will be recorded as "software and database professionals" and other broad occupations (Goodridge *et al*., 2015; Goodridge and Haskel, 2015a, 2015b). Their roles could include aspects of both data transformation and knowledge creation, as well as software creation.

The non-rival nature of information and knowledge assets does present challenges related to multiple counting. Some assets will be created for own final use, others for sale and some for both purposes. One practical way of dealing with this, and the approach taken in this paper, is to simply measure all (data transformation and analytics) activity at the macro level using the sum of costs method, regardless of whether output is intended for own use or sale. While firms may purchase information assets from other firms, and conduct further transformation, provided all transformation activity is measured, then purchases need not be measured separately. Costs of data acquisition can be incorporated into cost adjustment factors.[17] Similarly, firms may purchase analytics services from specialist firms but the macro approach will capture the analytics activity. However, mark-ups present in the price of traded information and knowledge assets will not be included using this approach. An alternative approach would be to measure all purchases separately. Then measurement of own-account production would require adjustment to avoid multiple counting (i.e. to deduct production of assets intended for final sale).[18]

5. Data Investment Activity in EU Countries

5.1. *Employment Engaged in Data Capital Formation*

In this section we present estimates of the volume of (software and) data investment activity in EU-28 countries, using occupation-based estimates of employment engaged in capital formation from the EU Labor Force Survey (LFS)[19]. Our approach is to identify all employment engaged in capital formation regardless of whether assets are destined for sale or own final use.[20] The method allows us to compare the volume of activity where: (a) the method is harmonized across countries; and (b) the asset boundary is expanded to fully include both data transformation and knowledge creation.[21]

To identify workers engaged in capital formation, we inspect the International Standard Classification of Occupations (ISCO-08) in light of: methods used in

[17] If treated as an intermediate. If it is capital expenditure then it is purchased GFCF and the acquired data generates capital services in the production of information.

[18] An overview of national accounts methods to measure GFCF in software and databases is provided in Appendix A, including information on how statistics authorities deal with potential double counting. Most countries make an adjustment to observed own-account activity to remove production of software and databases destined for final sale, but the methods focus on software.

[19] Data received on request from Eurostat.

[20] In forming estimates of investment, in the next section, we apply an adjustment to exclude capital formation activity in software that is destined for final sale.

[21] Current national accounts methods mean these activities are partially included in measured GFCF, mainly because occupations identified in national accounts measurement (largely software and database professionals) spend some of their time working on data transformation and knowledge creation.

experimental estimates for Canada and national accounts in the UK, Sweden and other EU countries;[22] and work in Goodridge *et al.* (2015) and Goodridge and Haskel (2015a, 2015b).[23] Inspection of ISCO-08 suggests the list of occupations in Table 1.

Some workers in these occupations will be engaged in either or both data transformation and knowledge creation, and some will also be engaged in software creation. We stress that occupational classifications are not sufficiently granular to disentangle entirely workers engaged in software creation from those engaged in data transformation and analysis. Further, even if classifications were sufficiently fine, in practice both activities are sometimes undertaken by the same workers (Goodridge *et al.*, 2015; Goodridge and Haskel, 2015a, 2015b). As a result, our estimates of employment engaged in capital formation are indicators of both software and data investment activity and we are unable to separate our estimates into their three respective components (software, data transformation and knowledge creation). While we would expect data entry occupations to be engaged in data transformation, analytical occupations will typically be engaged in both transformation and analytics, while ICT professionals may be engaged in one, two or all three activities.

Column 1 shows that we categorize our identified occupations in four groups: (1) ICT (software and database) professionals; (2) Data Entry; (3) Other ICT; and (4) Analytical. Columns 2 and 3 are ISCO codes and titles in each group. The identified occupations go beyond the ICT (software and database) professionals identified in Sweden and most other EU countries[24] and include analytical occupations similar to those identified by Statistics Canada (2019b) and, to a lesser extent, the UK ONS. In their advice for estimating capital formation in software and databases, OECD-Eurostat (OECD, 2020) recommend focusing on occupations in ISCO 25, in particular codes 251 (software professionals) and 2521 (Database designers and administrators).

Column 4 is our assumed time-use factor for each occupation group, to adjust for the amount of time spent creating assets, informed by those used in the UK, Sweden and Canada.[25] There is inevitably some subjectivity in time-use assumptions that are not based on formal time-use surveys. Our reasoning for each is as follows. On software and database professionals (group 1), the UK method assumes that software professionals (SOC 2136) spend 50 percent of their time engaged in software and database capital formation. That estimate is based on an informal survey of the trade association, Intellect UK, which reported that software professionals spend 70 percent of time on capital formation. The ONS chose to apply a lower factor of 50 percent in line with OECD recommendations (Chamberlin *et al.*, 2006, 2007). Sweden assumes a range of 20–76 percent for ICT professionals (ISCO 25) depending on their industry. We have chosen a factor of 50 percent although it could

[22] See Appendix A.

[23] Who identify relevant workers and occupations by mapping keywords to member profiles on an employment-based social media network.

[24] See Appendix A.

[25] Outlined in Tables A1, A2 and A4, respectively, in Appendix A. We note that OECD-Eurostat (OECD, 2020) recommend a time-use factor of 50 percent where no other information is available and that time-use factors do not exceed 50 percent on average (across all firms/industries). However, this seems a conservative assumption in cases where occupations would typically be expected to spend most of their time creating assets.

TABLE 1
Relevant Occupations in ISCO-08

Broad Occupation Groups	ISCO-08	Occupation Title	Assumed Time-use Factor	Mean % Engaged in Capital Formation (Time-use Adjusted, EU-28, 2011–2018)
1. Software and database	25	Information and Communications Technology Professionals (251: Software; 252: Database)	50%	57%
2. Data entry	4132	Data Entry Clerks	90%	9%
3: Other ICT	133	Information and Communications Technology Services Managers	25%	14%
	351	Information and Communications Technology Operations and User Support Technicians		
4: Analytical	212	Mathematicians, Actuaries and Statisticians	66%	20%
	2413	Financial Analysts		
	2631	Economists		
	3314	Statistical, Mathematical and Related Associate Professionals		

Notes: Column 1 are broad occupation groups. Columns 2 and 3 are ISCO-08 codes and titles that we consider engaged in (software and) data capital formation, within each broad group, based on inspection of the ISCO and methods used by national statistics agencies detailed in the Appendix. Column 4 is our assumed time-use factor for each broad group, partly based on those in Appendix Tables A1 and A4. Column 5 is the average (2011–2018) percentage of (time-use adjusted) employment in each group in the EU-28 total e.g. data entry clerks account for 9 percent of (time-use adjusted) employment engaged in capital formation across all occupations in Table 1. ISCO code 25 (Information and Communication Technology Professionals) is a two-digit aggregate of three-digit components ISCO 251 (Software and Applications Developers and Analysts) and ISCO 252 (Database and Network Professionals), which respectively include: 2511 (Systems Analysts); 2512 (Software Developers); 2513 (Web and Multimedia Developers); 2514 (Applications Programmers); 2519 (Software and Applications Developers and Analysts n.e.c.); and 2521 (Database Designers and Administrators); 2522 (Systems Administrators); 2523 (Computer Network Professionals); and 2529 (Database and Network Professionals n.e.c.). ISCO codes 133 and 212 include no other four-digit components. ISCO code 351 (ICT Operations and User Support Technicians) includes: 3511 (ICT Operations Technicians); 3512 (ICT User Support Technicians); 3513 (Computer Network and Systems Technicians); and 3514 (Web Technicians). A table of all identified ISCO codes and their sub-components is provided in Appendix D.

be argued that a larger factor is more appropriate. The chosen factor recognizes that: (i) the input of software and database professionals in the formation of software and database assets is already measured in national accounts and most countries apply a recommended time-use factor of 50 percent; (ii) the occupational group is broad with some workers engaged in software activity, others in data activity, and others in both; and (iii) workers will not spend all of their time on capital formation activity.

On data entry (group 2), it could be argued that a factor of 100 percent is appropriate, as in Statistics Canada (2019b), as these workers are likely fully engaged in data transformation. However, some may spend at least some time performing other administrative or clerical duties. We therefore choose a factor of 90 percent to account for most of their time.

On other ICT occupations (group 3), Table A1 shows that the ONS apply time-use factors of 5 percent–35 percent for similar occupations. Statistics Canada (2019b) apply a factor of 30 percent to Computer and information systems managers. We choose a factor of 25 percent that is consistent with this range and is similar to the factor used for IT managers and technicians in the UK method.

On analytical occupations (group 4) we note that, for similar occupations included in the UK method, the factor chosen is just 10 percent, which is low but likely reflects the intent to capture database production rather than data science. From Table A4, Statistics Canada apply much larger factors to similar occupations, in the range of 70 percent–90 percent when summed across activity in both data (i.e. transformation) and data science. While many such occupations likely work on data transformation and knowledge creation for much of their time, some, including those more senior, do not. We therefore, somewhat arbitrarily, assume these occupations on average spend two-thirds of their time engaged in data capital formation.

Column 5 presents the mean (2011–2018) percentage of employment from each group in the total for the EU-28, after applying time-use factors. Over half (57 percent) of activity is from software and database professionals, followed by analytical occupations (20 percent), other ICT occupations (14 percent) and data entry (9 percent). This accords with Goodridge *et al*. (2015) and Goodridge and Haskel (2015a, 2015b) who find that of UK workers engaged in data capital formation, 65 percent are already accounted for in official UK measurement of own-account investment in software and databases.

Charts detailing country-level information on levels and growth rates in employment engaged in (software and) data capital formation are presented in Appendix B. In brief, we find that: (a) 1.4 percent of EU-28 employment is engaged in the formation of (software and) data assets, ranging from 3.5 percent in Luxembourg (LU) to 0.5 percent in Greece (GR); (b) for the EU-28 as a whole, growth in the volume of employment engaged in capital formation activity was 5 percent pa over the period 2011–2018, ranging from 12.9 percent pa in Portugal (PT) to −2.4 percent pa in Latvia (LV).

5.2. *Estimating Data Investment and Comparison with National Accounts*

In this section we use our data on occupation-employment to form estimates of investment in (software and) data assets, where estimates are: (a) harmonized across countries; and (b) based on an expanded asset boundary that incorporates data transformation and knowledge creation. We compare our results with estimates of own-account capital formation in software and databases in national accounts to gain some understanding of how extending the asset boundary changes estimates of investment. As we only have national accounts own-account information for sixteen of the countries in our dataset, we carry out the analysis for the EU-16. We convert our employment values to nominal investment using the sum of costs method, as also used to estimate own-account investment in software and databases in national accounts.

First, to form credible estimates that maintain consistency with national accounting methods in EU countries, we adjust our estimates of employment to exclude labor input on *software* destined for final sale. This is necessary to avoid double-counting with purchased software, which is already included in measured national accounts GFCF. As detailed in information for the UK, Sweden and other countries in Appendix A this is a standard adjustment typically carried out by excluding the input of identified occupations in the software industry (NACE 62, Computer programing, consultancy and related activities).

With ideal data we could simply subtract ISCO 251 (software professionals) employment in NACE 62 from our ISCO 25 whole economy estimates. However, EU LFS sample sizes are not sufficient for reliable estimates at both detailed occupation- and industry-level. We do however have country-year estimates of total employment in NACE 62, for all occupations. We therefore subtract NACE 62 employment from ISCO 25 employment. We note however that this implicitly assumes that all workers in NACE 62 are in ISCO 251 and engaged in software capital formation, which is an over-adjustment but we lack more precise information. Thus we avoid potentially double-counting with purchased software and ensure that the adjustment is largest in countries with a larger software industry.

Second, after excluding employees in NACE 62, we apply our time-use factors (see Table 1) to form estimates of time-use adjusted labor input to software and data capital formation. Third, we must convert time-use adjusted employment to wage costs in capital formation. Ideally we would do this using country-occupation salaries but unfortunately this information is not available from our EU LFS data. We do however have labor composition data from EUKLEMS (Stehrer *et al.*, 2019), which we can use to derive estimates of average wages for workers with high, medium and low educational attainment in each country.[26] We therefore multiply time-use adjusted employment by a derived wage for each occupational group in each country-year. For software and database professionals, other ICT professionals and analytical occupations, we use the average annual wage of workers with high and medium (combined) educational attainment. For data entry we use the average annual wage for low educational attainment. After summing the wagebills

[26] The EUKLEMS Labor file includes shares of employment and labor compensation by labor composition group for characteristics of educational attainment (high, medium, low), age and gender. Summing across other characteristics (gender and age) gives shares for each educational attainment group. Applying the shares to total employment and total labor compensation, respectively, gives estimates of employment and labor compensation in each education group. We then divide the group wagebill (high, medium or low) by group employment to derive an average salary for each attainment group. We carry out the same procedure for an aggregate of the high and medium groups, effectively giving us a weighted average salary for high and medium attainment workers.

for each occupational group, we have estimates of time-use adjusted wage costs in (software and) data capital formation.

Fourth, to account for non-wage costs (non-wage labor costs, overheads, intermediates and capital services including the net rate of return to capital) we multiply wage costs by two, which is in line with methods in the UK, Sweden and other EU countries.[27] This gives us estimates of (own-account software and) data investment according to an expanded asset boundary, based on total costs incurred in capital formation.

To estimate the scale of how much estimates of investment in software and data change after extending the asset boundary, we compare our estimates with those in national accounts. We have estimates of national accounts GFCF in software and databases from EUKLEMS for EU-26 countries[28] but those estimates include expenditures on purchased software.[29] The correct comparator for our estimates is national accounts investment with the purchased element removed (i.e. own-account investment).

National accounts GFCF in software and databases separated into purchased and own-account components are not available for most EU countries. We have however found some country-specific estimates of the proportion of GFCF that is own-account for EU-16 countries.[30] Full details on the proportion of GFCF that is own-account in each EU-16 country and the source of the information is provided in Appendix F. Table A16 shows that the proportion varies widely across countries but is fairly stable within countries and across time.

Figure 2 presents the mean (2011–2016)[31] ratio of our newly constructed measure of (own-account software and) data investment ($P_N N^{GHE}$) to our estimates of own-account GFCF in software and databases in national accounts ($P_N N'_{oa}$) in EU-16 countries. A ratio close to one suggests that coverage of (occupations in) national accounts own-account GFCF already well captures data capital formation activity according to the asset boundary used in this paper. A ratio considerably greater than one implies that extending the asset boundary to include the labor input of all occupations in Table 1 substantially raises the estimate of capital formation. The method used means that identified additional investment (over and above that in national accounts) is largely due to the input of analytical

[27] See Appendix A. The adjustment factor used for experimental estimates in Canada is 1.55. The factor used in estimating own-account GFCF in software and databases is: 2.2 in Germany (DE); 2.4 in Denmark (DK, 2.2 in the non-market sector); 1 in Estonia (EE); 2.7 in Sweden (SE); and approximately 2 in the UK. We choose a factor of 2, which lies close to the middle of this range (an unweighted average of these seven values gives a factor of 2.01) and allows us to harmonize our method across countries.

[28] Estimates for Belgium (BE) and Croatia (HR) are not available from EUKLEMS.

[29] Based on information for the UK and Sweden (see Appendix A) we conjecture that the majority of purchased expenditure is on pre-packaged and custom software, rather than databases.

[30] The UK, Czechia (CZ) and (in a sense) Slovenia (SI) (see Appendix F) publish annual estimates. Belgium (BE), Denmark (DK), Estonia (EE) and Sweden (SE) publish a point estimate for an individual year in their GNI inventories. Austria (AT), Cyprus (CY), Ireland (IE), Luxembourg (LU), Malta (MT) and Slovakia (SK) provided annual estimates after we contacted the national statistics agency, and Germany (DE), the Netherlands (NL), Portugal (PT) and Romania (RO) provided us with average values. We were unable to gather any estimates for Bulgaria (BG), Spain (ES), Finland (FI), France (FR), Greece (GR), Croatia (HR), Hungary (HU), Italy (IT), Lithuania (LT), Latvia (LV) and Poland (PL).

[31] Estimates of GFCF in software and databases in EUKLEMS are missing for 2017 for some countries. We therefore only show data up to 2016.

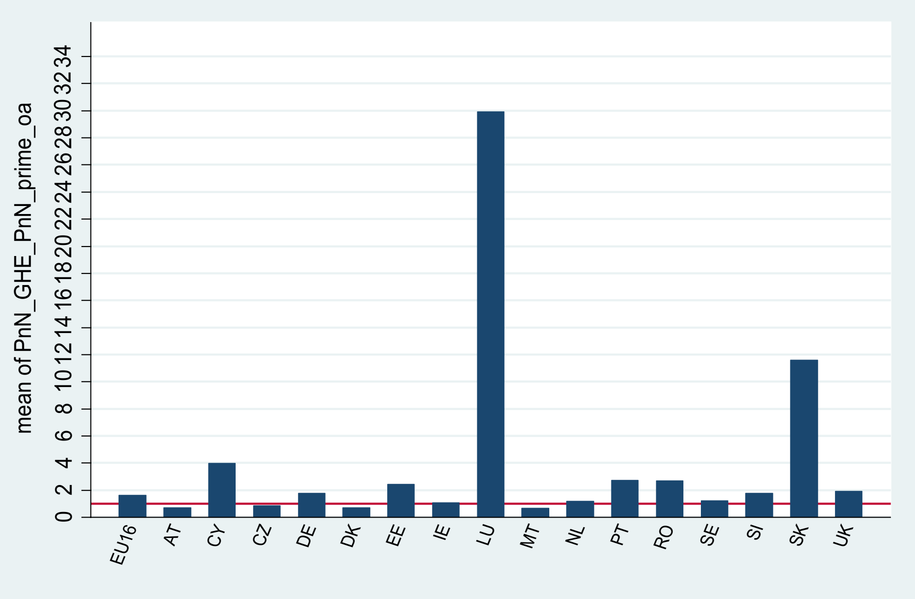

Figure 2. Ratio of GHE Investment (this Paper) to National Accounts Own-account GFCF in Software and Databases ($P_N N^{GHE}$: $P_N N'_{oa}$), Mean (2011–16), by Country (EU-16)

Notes: Y-axis is mean ratio of nominal investment in (own-account software and) data estimated in this paper ($P_N N^{GHE}$) to our estimate of nominal national accounts own-account investment in software and databases ($P_N N'_{oa}$). Solid horizontal line is 1. Data are averages for 2011–2016.

Source: Authors' estimates. $P_N N^{GHE}$ are estimates constructed for this paper, based on employment values from the EU LFS and wage rates for labor composition groups from EUKLEMS (Stehrer *et al.*, 2019). National accounts own-account GFCF in software and databases ($P_N N'_{oa}$) are authors' estimates derived using country-specific estimates of the proportion of GFCF that is own-account (see Appendix F) and estimates of GFCF in software and databases from EUKLEMS. Some EU LFS employment values are imputed for missing and unpublished observations. For details, see Table A12, Appendix C. Some EU LFS employment values are flagged by Eurostat for low reliability. For details, see Table A11, Appendix C. [Colour figure can be viewed at wileyonlinelibrary.com]

occupations (group 4) and to a lesser extent, other ICT (group 3) and data entry (group 2) occupations.[32]

The mean ratio of 1.61 for the EU-16 aggregate[33] suggests that identifying a wider range of occupations engaged in capital formation raises own-account GFCF by around 60 percent. The ratio is less than one in: Malta (MT), Austria (AT), Denmark (DK) and Czechia (CZ); suggesting that our adjustment to avoid double-counting with purchased software is, for these countries at least, an over-adjustment. The ratio is greater than one but lower than for the EU-16 aggregate in: Ireland (IE); the Netherlands (NL); and Sweden (SE). It is higher than the ratio

[32] Information in Appendix A confirms that countries largely base own-account estimates on the input of software and database professionals (ISCO 25), as recommended by OECD-Eurostat (OECD, 2020). Some countries also include other ICT and data entry occupations (e.g. Austria (AT) and Germany (DE)). With the exceptions of the UK (economists and other analytical occupations) and Germany (DE) (data scientists), input from analytical occupations are not typically included in estimation of GFCF in European countries.

[33] Constructed using PPP(GDP)-adjusted values of investment. PPP data downloaded from Eurostat. https://ec.europa.eu/eurostat/web/purchasing-power-parities/data/database.

for the EU-16 aggregate for all other countries. The particularly high value for Luxembourg (LU) reflects very large estimates of employment in analytical occupations. The ratio is also relatively high in the Slovak Republic (SK), Cyprus (CY), Portugal (PT) and Romania (RO). Of these countries, LU, SK, PT, and, to a lesser extent CY, all report low shares of own-account GFCF in total GFCF in software and databases, which may partly explain these results. The results suggest that extending the asset boundary would substantially increase measured own-account investment in these four countries in particular.

The above data are averages. Figure 3 presents annual estimates of our measures of nominal investment ($P_N N^{GHE}$) and national accounts own-account GFCF ($P_N N'_{oa}$).

In Appendix H, we study the correlation between our measure of investment and other forms of (tangible and intangible) capital formation. We find: a positive correlation with measured investment in software and databases, as expected due to consistency in methods of measurement; a negative correlation with R&D, providing some support for our view that R&D as measured in national accounts does not typically include activity in data analytics; and a positive correlation with mineral exploration. We consider the possibility of double counting with R&D and mineral exploration in sensitivity analyses presented in Table 3.

While comparisons of nominal investment are interesting, it is changes in real volumes that determine productivity growth and the contribution of capital. We study implications for growth in the next section but here we first compare annual changes in real investment in EU-16 countries. Real investment is derived using (country-specific) investment price indices for software and databases from EUKLEMS.[34] Figure 4 compares growth in the two measures of real investment ($\Delta \ln N^{GHE}$and $\Delta \ln N'_{oa}$). We estimate growth in newly expanded real investment ($\Delta \ln N^{GHE}$) of 6.7 percent pa for the EU-16, compared to 2.7 percent pa in national accounts ($\Delta \ln N'_{oa}$). The difference is particularly large in Germany (DE) and Romania (RO).

6. Economic Impact of Data Capital Formation on Growth

6.1. *Framework*

In this section we use our new estimates of investment to quantify the economic impact of expanding the scope of (software and data) capital formation in a

[34] With ideal data we would construct our own price index for our measure of investment, following the method typically used for own-account capital formation in national accounts. That is, construct an own-account price index based on the wages of relevant occupations and prices for other inputs to capital formation, and assuming no productivity gain in production. However these input price data were not available. Instead we use the EUKLEMS price index for software and databases in national accounts. We note however that the EUKLEMS investment price index for software and databases is implicitly a weighted average of price indices for purchased and own-account investment, and it is the latter which we measure. Although of course measured own-account prices are based on the input of occupations identified by national statistics agencies (mainly software and database professionals) and do not include the wages of other occupations, such as analysts, that we identify in this paper. As measured purchased software prices do not typically rise as fast as measured own-account prices, the likely impact is that we under-estimate price increases and therefore over-estimate real GFCF growth. As a result, in the next section, we likely over-estimate adjusted growth in labor productivity and the contribution of software and data capital deepening. However, as the EUKLEMS price index includes an own-account component, these errors are likely small.

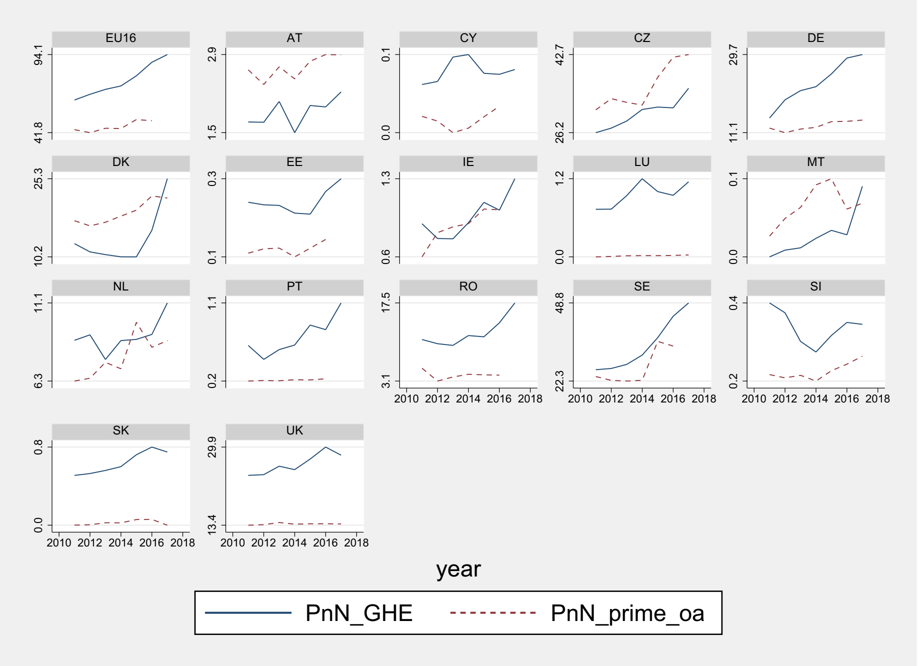

Figure 3. Nominal Investment: GHE ($P_N N^{GHE}$, this Paper) vs National Accounts Own Account GFCF in Software and Databases ($P_N N'_{oa}$), 2011–2017, National Currency (bns) by Country (EU-16)

Notes: Estimates in billions of national currency units. EU-16 aggregate is € billions, and is the sum of PPP(GDP)-adjusted values for underlying countries. Estimates of own-account GFCF in 2017 are missing in EUKLEMS for CY, EE, IE, PT, RO and SE. Therefore estimates for these countries and the EU-16 are up to 2016.

Source: Authors' estimates. $P_N N^{GHE}$ (this paper, blue line) based on employment values from the EU LFS and wage rates for labor composition groups from EUKLEMS (Stehrer *et al.*, 2019). National accounts own-account GFCF in software and databases ($P_N N'_{oa,}$ red line) are authors' estimates derived using country-specific estimates of the proportion of GFCF that is own-account (see Appendix F) and estimates of GFCF in software and databases from EUKLEMS. Some EU LFS employment values are imputed for missing and unpublished observations. For details, see Table A12, Appendix C. Some EU LFS employment values are flagged by Eurostat for low reliability. For details, see Table A11, Appendix C.

growth-accounting context when estimation is harmonized across countries, by comparing new estimates of the growth decomposition with those in the EUKLEMS database (Stehrer *et al.*, 2019). We carry out the analysis for EU-13 countries.[35]

We first set out a simple model that is typical in the intangibles literature (e.g. Corrado *et al.*, 2005), where part of investment is already included in measured national accounts GFCF and another part is newly identified investment as a result of expansion of the asset boundary.

First, we set out how measured investment in software and databases ($P_N N'$) relates to estimates in this paper ($\mathrm{P_N N^{GHE}}$). In the notation below, N is real investment in software and data and P_N is its price. Superscripts $^{\mathrm{GHE}}$ refers to estimates in this paper, $'$ to measured national accounts estimates and * to newly identified additional investment (i.e. investment over and above that recorded in national accounts). Subscripts $_{\mathrm{purch}}$ and $_{\mathrm{oa}}$ refer to purchased and own-account investments respectively:

[35] Of the EU-16 countries for which we estimate investment, growth-accounting data for Cyprus (CY), Malta (MT) and Romania (RO) are incomplete in the EUKLEMS database.

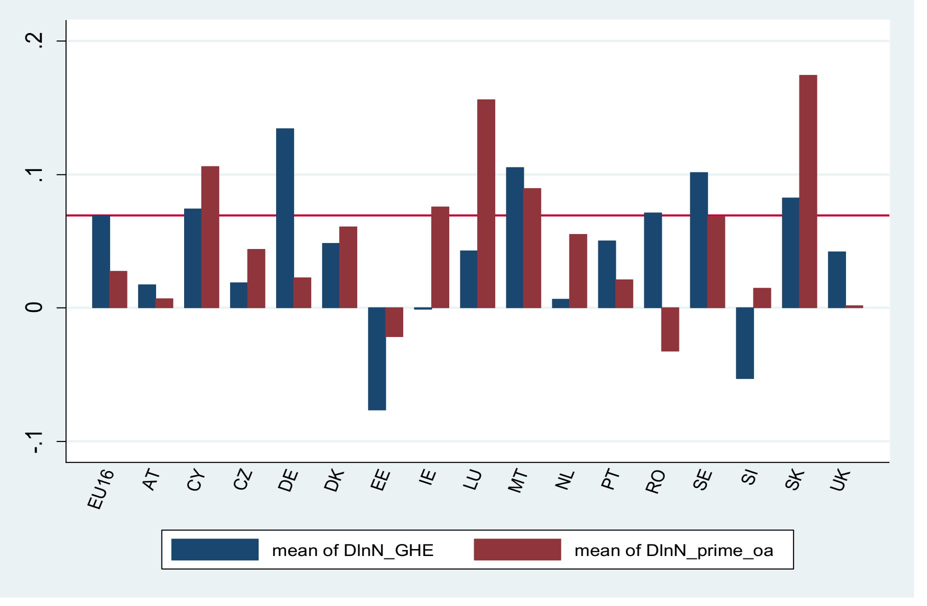

Figure 4. Mean Growth Rate in Real GHE Investment ($\Delta \ln N^{GHE}$, this Paper) vs Mean Growth in Real (National Accounts) GFCF in Own-Account Software and Databases ($\Delta \ln N'_{oa}$), Mean (2011–2016), by Country (EU-16)

Notes: Each category of nominal investment deflated using software and database investment price index from EUKLEMS. *Y*-axis is mean growth rate. Growth rates calculated as changes in the natural log. Estimates for the EU-16 aggregate constructed as a weighted average of country growth rates using PPP(GDP)-adjusted nominal investment as weights. Solid horizontal line highlights mean growth rate in real GHE investment ($\Delta \ln N^{GHE}$) for the EU-16 (6.9 percent pa). Data are averages for 2011–2016.

Source: Authors' estimates. GHE data investment based on employment values from the EU LFS and wage rates for labor composition groups from EUKLEMS (Stehrer *et al.*, 2019). National accounts own-account GFCF derived using country-specific estimates of the proportion of GFCF that is own-account (see Appendix F) and estimates of GFCF in software and databases from EUKLEMS. Some EU LFS employment values are imputed for missing and unpublished observations. For details, see Table A12, Appendix C. Some EU LFS employment values are flagged by Eurostat for low reliability. For details, see Table A11, Appendix C. [Colour figure can be viewed at wileyonlinelibrary.com]

$$
\begin{aligned}
P_N N' &= P_N N'_{purch} + P_N N'_{oa} \\
P_N N &= P_N N'_{purch} + P_N N^{GHE} \\
&= P_N N' + P_N N^{*} \\
P_N N^{*} &= P_N N^{GHE} - P_N N'_{oa}
\end{aligned}
\tag{1}
$$

Equation (1) says that total measured national accounts GFCF in software and databases ($P_N N'$) is the sum of purchased ($P_N N'_{purch}$) and own-account ($P_N N'_{oa}$) components. An expanded definition of total investment ($P_N N$) equals the sum of national accounts purchased software investment[36] ($P_N N'_{purch}$) and the

[36] From discussions with national statistics agencies and reviews of their methods, to the best of our knowledge national accounts purchased investment largely, if not wholly, consists of purchases of software rather than databases. See Appendix for further information.

measure of investment constructed in this paper ($P_N N^{GHE}$). $P_N N$ is also equivalent to total measured national accounts GFCF in software and databases ($P_N N'$) plus newly identified additional investment due to expansion of the asset boundary ($P_N N^*$). That additional investment can also be derived as the measure constructed in this paper ($P_N N^{GHE}$) minus measured own-account GFCF ($P_N N'_{oa}$).

How does this relate to national accounts output (GDP)? Ignoring government output and net trade, from the expenditure side, measured (V) and adjusted (Q) output, where adjusted output includes newly identified investment, are:

$$\begin{aligned} P_V V &= P_C C + P_I I + P_N N' \\ P_Q Q &= P_C C + P_I I + P_N N \\ &= P_V V + P_N N^* \end{aligned} \tag{2}$$

where V is real value-added, Q is adjusted real value-added, C is real consumption, I is real investment in all (tangible and intangible) assets except software and data, and P are their prices. Equation (2) simply states that adjusted nominal output ($P_Q Q$) is equal to measured nominal output ($P_V V$) plus newly identified nominal investment ($P_N N^*$).[37]

The relation between measured (V) and adjusted (Q) real output growth can be written as:

$$\begin{aligned} \Delta \ln Q_t &= s_Q^V \Delta \ln V_t + s_Q^{N^*} \Delta \ln N_t^* \\ &= \left(1 - s_Q^{N^*}\right) \Delta \ln V_t + s_Q^{N^*} \Delta \ln N_t^* \\ &= \Delta \ln V_t + s_Q^{N^*} \left(\Delta \ln N_t^* - \Delta \ln V_t\right) \end{aligned} \tag{3}$$

where N* is newly identified real investment and s_Q are shares in adjusted nominal value-added which sum to one.[38] From (3) it is clear that, provided $s_Q^{N^*} > 0$, faster growth in N^* relative to V will result in $\Delta \ln Q > \Delta \ln V$, showing that capitalization of N^* changes both input and output. The sources of growth decomposition for measured and adjusted labor productivity growth can be written as:

$$\begin{aligned} \Delta \ln (V/H)_t &\equiv s_V^{X'} \Delta \ln (X/H)_t + s_V^{R'} \Delta \ln \left(R'/H\right)_t + \Delta \ln TFP'_t \\ \Delta \ln (Q/H)_t &\equiv s_Q^{X} \Delta \ln (X/H)_t + s_Q^{R} \Delta \ln (R/H)_t + \Delta \ln TFP_t \end{aligned} \tag{4}$$

[37] This is true for the market sector where expansion of the asset boundary results in the identification and capitalization of additional (gross) output, with the value of that output estimated as costs incurred in production. The adjustment to output is different in the non-market sector. In general, for most countries, government output is estimated as the sum of costs incurred in generating output. Since we also estimate investment as the sum of costs incurred, capitalization of data in the non-market sector means that the correct adjustment involves adding a measure of consumption of fixed capital (CFC i.e. depreciation) to output rather than a measure of GFCF. As our EU LFS occupation data are at whole economy level, we are unable to produce separate estimates for the market and non-market sectors. Our analysis therefore implicitly assumes that market GFCF and non-market CFC are growing at similar rates, which may not be correct. We note this inaccuracy.

[38] Estimated as averages of the share in the current and previous period to form a superlative index.

where total factor productivity (TFP) is defined as a residual, H are annual person-hours and X is an index of labor (L) and (tangible and intangible) capital inputs (K) except software and databases. R is the stock of software and data capital. s is an income share for each factor input $Z(=X, R)$, estimated as an average over two periods (we omit the usual overbar to ease notation):[39]

$$s_Q^Z \equiv \frac{1}{2}\left[\left(\frac{P_Z Z}{P_Q Q}\right)_t + \left(\frac{P_Z Z}{P_Q Q}\right)_{t-1}\right] \tag{5}$$

Capital (K and R; K only shown here for simplicity) and labor (L) services are translog aggregations over heterogeneous capital types a and labor types b, respectively:

$$\Delta \ln K = \sum s_K^{Ka} \Delta \ln K_{a,t} \tag{6}$$

$$\Delta \ln L = \sum s_L^{Lb} \Delta \ln H_{b,t} \tag{7}$$

where shares (s) are of total capital and labor payments for each type, again averaged over the current and previous period in order to form a superlative index:

$$s_L^{Lb} \equiv \frac{1}{2}\left[\left(\frac{P_{Lb} L_b}{P_L L}\right)_t + \left(\frac{P_{Lb} L_b}{P_L L}\right)_{t-1}\right] \tag{8}$$

$$s_K^{Ka} \equiv \frac{1}{2}\left[\left(\frac{P_{Ka} K_a}{P_K K}\right)_t + \left(\frac{P_{Ka} K_a}{P_K K}\right)_{t-1}\right] \tag{9}$$

Labor is in natural units, hours. For capital (K, R), stocks for each type (a) are constructed using nominal investment and a price index for capital goods of each type in a perpetual inventory model (PIM) so that:

$$\begin{aligned} K_{a,t} &= \frac{P_{Ia} I_a}{P_{Ia}} + \left(1 - \delta^{Ka}\right) K_{a,t-1} \\ R_t &= \frac{P_N N}{P_N} + \left(1 - \delta^{R}\right) R_{t-1} \end{aligned} \tag{10}$$

where δ is an asset-specific depreciation rate. The inputs side of the model is completed by the user-cost relation between P_{Ia} (or P_N) and P_{Ka} (or P_R):

$$P_R = P_N \left(\rho + \delta^R - \left(\Delta P_N / P_N\right)\right) \tag{11}$$

[39] Shares of measured value-added (V) are estimated in the same way with measured value-added as the denominator.

where $\Delta P/P$ is the capital gain/loss from holding the asset and ρ is an economy-wide nominal net rate of return estimated such that gross operating surplus is exhausted. As set out in Corrado *et al.* (2020), when the asset boundary is expanded to incorporate newly identified investment, the wedge between measured and adjusted growth consists of four terms, each of which can be seen in the equations above. They are as follows.

6.1.1. Wedge in Output Growth

First, on the output side, there is the change to growth in output or labor productivity, as set out in equation (3), which can be written as:

$$\Delta \ln Q_t - \Delta \ln V_t = s_Q^{N^*} \left(\Delta \ln N_t^* - \Delta \ln V_t\right) \tag{12}$$

6.1.2. Wedge in Contribution of Software and Data Capital Deepening

Second, on the input side, from equation (4) there is a wedge between the measured and adjusted contribution of capital deepening in software and data(bases):

$$s_V^{R'} \Delta \ln\left(R'/H\right)_t - s_Q^{R} \Delta (R/H)_t \tag{13}$$

where $'$ denotes measured. Equation (13) shows that the difference is due to a change to the income share and also a change to growth in software and data capital services (R).

6.1.3. Wedge in Contribution of Other Factor Inputs

Third, from equation (4), there is an additional effect on the input side due to a change to the contribution of other factor inputs (X):

$$s_V^{X'} \Delta \ln (X/H)_t - s_Q^{X} \Delta (X/H)_t \tag{14}$$

The difference is due to a change in the income shares for other factor inputs (L and K), summarized in s^X.

6.1.4. Wedge in Total Factor Productivity (TFP) Growth

Finally, equation (4) shows that there is a difference between measured and adjusted growth in total factor productivity, which is the cumulative effect of the other three terms.

Below we estimate each of these effects using data from the EUKLEMS (Stehrer *et al.*, 2019) database.

6.2. *Capitalization of Newly Identified Investment in Data*

In the notation that follows, $'$ refers to measured (i.e. national accounts) data.

6.2.1. Output

First, to estimate the change to output, we estimate $P_N N^*$ as in the final line of equation (1). We derive N* using the measured price index for software and databases (P'_N) from EUKLEMS. $s_Q^{N^*}$ is estimated using newly identified investment over and above national accounts estimates ($P_N N^*$) and adjusted value-added ($P_Q Q$), as in equations (1) and (2).

6.2.2. Contribution of Capital Deepening in Software and Databases

Second, to estimate the adjustment to the contribution of capital deepening in software and databases, we first divide the measured contribution ($s_V^{R'} \Delta \ln R'$) by measured growth in capital services ($\Delta \ln R'$) to derive the measured income share ($s_V^{R'}$):

$$s_{V,t}^{R'} = \frac{s_V^{R'} \Delta \ln R'_t}{\Delta \ln R'_t} \tag{15}$$

We then multiply the measured income share ($s_V^{R'}$) by measured value-added ($P_V V$) to recover implied measured rental payments to software and databases ($P_R R'$). We divide measured rental payments by the measured real stock (R') to back out the implied measured rental price (P'_R):

$$P_R R'_t = s_{V,t}^{R'} * P_V V_t \tag{16}$$

$$P'_{R,t} = \frac{P_R R'_t}{R'_t} \tag{17}$$

To estimate our new measure of adjusted capital services based on an expanded definition of data capital, we estimate $P_N N$ as in (the second equation in) equation (1), and derive (real) N using the EUKLEMS price index (P_N). We then re-build the capital stock in a PIM from 2011, as in equation (10), using the depreciation rate (0.315) for software and databases in EUKLEMS. This gives us estimates of adjusted growth in capital services, $\Delta \ln R$, and growth in capital deepening, $\Delta \ln (R/H)$.

To derive the new adjusted contribution of software and data capital deepening to growth, we require an estimate of the new adjusted income share. We therefore re-estimate rental payments as the measured rental price[40] (P'_R) times the (level of the) new adjusted real stock, which incorporates our expanded definition of investment (R). The new income share (s_Q^R) is estimated as new adjusted rental payments ($P_R R$) divided by adjusted value-added ($P_Q Q$).

$$P_R R_t = P'_{R,t} * R_t \tag{18}$$

[40] We therefore implicitly assume the same net rate of return as that found in the EUKLEMS growth-accounting exercise. Strictly, that rate would change if re-estimated using the ex-post method in a new growth-accounting exercise with an expanded definition of data capital, but the effect on the rental price and estimated contributions would be small.

$$s^{R}_{Q,t} = \frac{P_R R_t}{P_Q Q_t} \tag{19}$$

Finally, we re-estimate the contribution of capital deepening for our expanded definition of capital ($s^R_Q \Delta\ln(R/H)$ using the new income share (s^R_Q) and new estimates of capital deepening ($\Delta\ln(R/H)$. We are now able to compare the new adjusted contribution with the measured contribution in EUKLEMS.

$$s^R_Q \Delta\ln(R/H)_t = s^R_{Q,t} * \Delta\ln(R/H)_t \tag{20}$$

6.2.3. Contribution of Other Factor Inputs

Third, to estimate the wedge in the contribution of other (capital and labor) inputs, we estimate the measured and adjusted income share for all other factor inputs (X) as one minus the measured and adjusted share for software and data, respectively:

$$\begin{gathered} s^{X\prime}_{Vt} = 1 - s^{R\prime}_{t} \\ s^{X}_{Qt} = 1 - s^{R}_{t} \end{gathered} \tag{21}$$

We derive measured growth in non-R factor inputs per hour ($\mathit{\Delta ln(X/H))}$ using the sum of measured factor contributions ($s^{X'}_V \Delta\ln(X/H)$ and income shares ($s^{X'}_{V,t}$) from EUKLEMS (excluding those for software and databases):

$$\Delta\ln(X/H)_t = \frac{s^{X'}_V \Delta\ln(X/H)_t}{s^{X'}_{V,t}} \tag{22}$$

where:

$$s^{X'}_V \Delta\ln(X/H)_t = s^{L'}_V \Delta\ln(L/H)_t + s^{K,NICT\prime}_V \Delta\ln\left(K^{NICT}/H\right)_t + s^{K,ICT\prime}_V \Delta\ln\left(K^{ICT}/H\right)_t +$$

$$s^{K,RD\prime}_V \Delta\ln\left(\frac{K^{RD}}{H}\right)_t + s^{K,OIPP\prime}_V \Delta\ln\left(\frac{K^{OIPP}}{H}\right)_t$$

and:

$$s^{X'}_{V,t} = s^{L\prime}_{V,t} + s^{K,NICT\prime}_{V,t} + s^{K,ICT\prime}_{V,t} + s^{K,RD\prime}_{V,t} + s^{K,OIPP\prime}_{V,t}$$

where NICT is non-ICT tangible capital and OIPP is other IPPs (artistic originals and mineral exploration).

We estimate the new contribution of non-software and database factor inputs ($s^X_Q \Delta\ln(X/H)$) using the newly estimated share (s^X_Q) (from equation (21)) and measured $\Delta\ln\mathit{(X/H)'}$ (from equation (22)).

$$s_Q^X \Delta\ln(X/H)_t = s_{Q,t}^X * \Delta\ln(X/H)_t \quad (23)$$

6.2.4. Total Factor Productivity Growth

Finally, using our new estimates of adjusted growth in output ($\Delta\ln(Q/H)$), the contribution of software and data capital deepening based on an expanded definition of data capital formation ($s_Q^R \Delta\ln(R/H)$), and the contribution of non-software and database factor inputs ($s_Q^X \Delta\ln(X/H)$), we can re-estimate $\Delta\ln TFP$ as a residual.

6.3. *Results: Economic Impact*

Table 2 presents our results. We compare measured estimates from EUKLEMS with our new adjusted estimates and show the scale of adjustment due to an expanded definition of data investment. Growth-accounting data for Cyprus (CY), Malta (MT) and Romania (RO) are incomplete in the EUKLEMS database. Therefore, we present estimates for EU-13 countries and a weighted average for the aggregate. Measured data are in plain font, new adjusted data are in *italics* and the wedge between the two is in bold.

Column 1 is measured labor productivity growth, column 2 is adjusted labor productivity growth and column 3 is the wedge. Column 4 is the measured contribution of capital deepening in software and databases, column 5 is the adjusted contribution and column 6 is the wedge. Column 7 is the measured contribution of other non-R factor inputs (X), column 8 is the adjusted contribution and column 9 is the wedge. Column 10 is measured growth in TFP, column 11 is adjusted growth in TFP and column 12 is the wedge. The memo item in column 13 is the share of country adjusted value-added in the EU-13 aggregate.

Estimates in the final row are weighted averages for the EU-13. Estimates are weighted using PPP(GDP)-adjusted shares of measured (V) and adjusted (Q) value-added. The latter are shown as a memo item in column 13.[41] Column 13 shows that EU-13 estimates are dominated by data for Germany (DE) and the UK.

The estimated wedge is zero for Austria (AT), Czechia (CZ) and Denmark (DK) as for these countries we estimate $P_N N^* = 0$ in all years (see Figures 2 and 3). That is, for these countries, expanding the definition of data capital does not raise estimated investment in software and databases.[42]

On the wedge in labor productivity growth (LPG, columns 1 to 3), a positive term means that adjusted labor productivity growth is greater than measured growth due to expansion of the production boundary. The wedge term is positive and largest in Germany (DE) and the UK. A negative value means that expansion

[41] Estimated as an average of the share in the current (t) and previous period ($t-1$) to form a superlative index.

[42] This finding reflects less activity in data capital formation in these countries. It may also reflect inaccuracy in our estimation of national accounts own-account investment in software and databases due to an over-adjustment in our effort to remove software investment purchases.

TABLE 2
Decomposition of Labor Productivity Growth, Averages for 2011–2016

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	***(13)***
			=(2)-(1)			=(5)-(4)			=(8)-(7)	=(1)-(4)-(7)	=(2)-(5)-(8)	=(11)-(10)	***Memo:***
	Dln(V/H)	*Dln(Q/H)*	**Wedge**	sR′*	*sR**	**Wedge**	sX′*	*sX**	**Wedge**	DlnTFP′	*DlnTFP*	**Wedge**	***sQ***
				Dln(R′/H)	***Dln(R/H)***		Dln(X/H)	***Dln(X/H)***					
AT	0.68%	*0.68%*	**0.00%**	0.05%	*0.05%*	**0.00%**	0.48%	*0.48%*	**0.00%**	0.14%	*0.14%*	**0.00%**	*0.04*
CZ	1.24%	*1.24%*	**0.00%**	0.04%	*0.04%*	**0.00%**	0.57%	*0.57%*	**0.00%**	0.64%	*0.64%*	**0.00%**	*0.03*
DE	0.86%	*0.95%*	**0.09%**	0.02%	*0.09%*	**0.07%**	0.08%	*0.08%*	**0.00%**	0.77%	*0.79%*	**0.02%**	*0.39*
DK	1.17%	*1.17%*	**0.00%**	0.02%	*0.02%*	**0.00%**	0.37%	*0.37%*	**0.00%**	0.78%	*0.78%*	**0.00%**	*0.03*
EE	1.17%	*1.08%*	**−0.09%**	0.02%	*0.09%*	**0.07%**	0.72%	*0.71%*	**0.00%**	0.44%	*0.28%*	**−0.16%**	*0.00*
IE	4.74%	*4.73%*	**−0.02%**	0.03%	*0.07%*	**0.03%**	1.83%	*1.82%*	**0.00%**	2.88%	*2.84%*	**−0.05%**	*0.03*
LU	0.18%	*0.21%*	**0.02%**	0.08%	*0.35%*	**0.27%**	0.79%	*0.78%*	**−0.01%**	−0.69%	*−0.92%*	**−0.24%**	*0.01*
NL	0.55%	*0.52%*	**−0.03%**	0.08%	*0.14%*	**0.06%**	0.36%	*0.35%*	**0.00%**	0.11%	*0.02%*	**−0.08%**	*0.09*
PT	0.34%	*0.36%*	**0.02%**	0.01%	*0.12%*	**0.11%**	0.91%	*0.91%*	**0.00%**	−0.59%	*−0.67%*	**−0.09%**	*0.03*
SE	0.96%	*0.99%*	**0.04%**	0.26%	*0.17%*	**−0.09%**	0.45%	*0.44%*	**0.00%**	0.25%	*0.38%*	**0.13%**	*0.05*
SI	0.98%	*0.91%*	**−0.07%**	−0.01%	*0.04%*	**0.04%**	0.55%	*0.55%*	**0.00%**	0.44%	*0.32%*	**−0.12%**	*0.01*
SK	2.19%	*2.23%*	**0.04%**	0.08%	*0.11%*	**0.02%**	0.87%	*0.87%*	**0.00%**	1.23%	*1.26%*	**0.02%**	*0.02*
UK	0.18%	*0.23%*	**0.05%**	−0.02%	*0.11%*	**0.13%**	−0.14%	*−0.14%*	**0.00%**	0.34%	*0.26%*	**−0.08%**	*0.28*
EU13	0.79%	*0.83%*	**0.05%**	0.03%	*0.10%*	**0.07%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.54%*	**−0.02%**	*1.00*

Notes: Decomposition of labor productivity growth constructed using: (a) measured estimates from EUKLEMS; and (b) new adjusted estimates that incorporate estimates of investment in data after expanding the asset boundary. Growth rates estimated as changes in the natural log. May not sum exactly due to rounding. The first panel from the left-hand side is growth in value-added per hour worked. The second panel is the contribution of software and data capital deepening. The third panel is the contribution of other factor inputs. The fourth panel is growth in total factor productivity. The memo item in the final column is the share in adjusted output for that country in the EU-13, estimated as an average of the share in the current and previous period. Measured estimates in plain font and new adjusted estimates in *italics*. The difference between them is the wedge, presented in bold. A positive sign for the wedge means that the adjusted estimate is greater than the measured estimate, where the former incorporates newly identified investment. Growth-accounting data for 2017 is incomplete in the EUKLEMS database. Therefore, we present estimates up to 2016.

TABLE 3

Sensitivity Analyses: Decomposition of Labor Productivity Growth for the EU-13, Averages for 2011–2016

	(1)	***(2)***	(3)	(4)	***(5)***	(6)	(7)	***(8)***	(9)	(10)	***(11)***	(12)
			=(2)-(1)			=(5)-(4)			=(8)-(7)	=(1)-(4)-(7)	=(2)-(5)-(8)	=(11)-(10)
	Dln(V/H)	***Dln(Q/H)***	Wedge	sR*	***sR****	Wedge	sX′*	***sX****	Wedge	DlnTFP′	***DlnTFP***	Wedge
				Dln(R′/H)	***Dln(R/H)***		Dln(X/H)	***Dln(X/H)***				
	(1) Baseline											
EU13	0.79%	*0.83%*	**0.05%**	0.03%	*0.10%*	**0.07%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.54%*	**-0.02%**
	(2) Potential double counting with R&D: exclude all analytical occupations											
EU13	0.79%	*0.81%*	**0.03%**	0.03%	*0.07%*	**0.04%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.55%*	**-0.01%**
	(3) Potential double counting with R&D: reduce time-use factor for analytical occupations by 50% (to 33%of time)											
EU13	0.79%	*0.82%*	**0.04%**	0.03%	*0.08%*	**0.06%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.54%*	**-0.02%**
	(4) Potential double counting with mineral exploration: subtract 50% of employment in mining and quarrying sector											
EU13	0.79%	*0.82%*	**0.04%**	0.03%	*0.08%*	**0.05%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.55%*	**-0.01%**
	(5) Potential double counting with mineral exploration: subtract 20% of GFCF in other IPPs (artistic originals and mineral exploration)											
EU13	0.79%	*0.83%*	**0.05%**	0.03%	*0.09%*	**0.07%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.54%*	**-0.02%**
	(6) Sensitivity: set all *time-use factors = 50% (consistent with OECD recommendation for software professionals)*											
EU13	0.79%	*0.84%*	**0.05%**	0.03%	*0.11%*	**0.08%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.53%*	**-0.03%**
	(7) Sensitivity: set all *time-use factors = 100% (i.e. all time spent on capital formation)*											
EU13	0.79%	*0.91%*	**0.13%**	0.03%	*0.36%*	**0.34%**	0.20%	*0.20%*	**0.00%**	0.56%	*0.36%*	**-0.20%**

Notes: Decomposition of labor productivity growth constructed using: (a) measured estimates from EUKLEMS; and (b) new adjusted estimates that incorporate estimates of investment in data after expanding the asset boundary. Growth rates estimated as changes in the natural log. May not sum exactly due to rounding. The first panel from the left-hand side is growth in value-added per hour worked. The second panel is the contribution of software and data capital deepening. The third panel is the contribution of other factor inputs. The fourth panel is growth in total factor productivity. Measured estimates in plain font and new adjusted estimates in *italics*. The difference between them is the wedge, presented in bold. A positive sign for the wedge means that the adjusted estimate is greater than the measured estimate, where the former incorporates newly identified investment. Growth-accounting data for 2017 is incomplete in the EUKLEMS database. Therefore, we present estimates up to 2016. In row 1 we present our baseline estimates, as in the final row of Table 2. In remaining rows we carry out a series of sensitivity checks. In row 2, we exclude all analytical occupations in case of double-counting with R&D. In row 3, we reduce the time-use factor for analytical occupations by 50 percent (to 33 percent) for the same reason. In row 4, we subtract 50 percent of employment in NACE 05 to 09 (mining and quarrying) from our estimate of (time-use adjusted) employment engaged in capital formation, in case of double-counting with mineral exploration. In row 5, we subtract 20 percent of GFCF in Other IPPs (an aggregate of GFCF in artistic originals and mineral exploration) from our estimate of capital formation. The assumption that 20 percent represents mineral exploration is based on data observed for the UK in 2014 (Goodridge *et al.*, 2016). In row 6, we set time-use factors for all occupations to 50 percent, consistent with the OECD recommendation for software and database professionals. Finally, in row 7, we use a 100 percent time use factor for all occupations, thus assuming they spend all their time on capital formation.

reduces growth in labor productivity, as newly identified investment is growing more slowly than measured value-added. The wedge term is negative in: Estonia (EE); Slovenia (SI); the Netherlands (NL); and Ireland (IE). For the EU-13, the wedge is 0.05 percent pa meaning that expansion of the asset boundary adds 0.05 percent pa to LPG in 2011–2016.

The wedge in the contribution of software and database capital deepening (columns 4 to 6) is positive for all countries except Sweden (SE), where it is negative. For the EU-13, expansion of the asset boundary raises the contribution of capital deepening over three-fold, from 0.03 percent pa to 0.10 percent pa in 2011–2016.[43] The wedge in the contribution of other factor inputs (X) is very small in all countries.

For the EU-13 total, we find that, after rounding, the input and output adjustments explain around 0.02 percent pa (4 percent) of measured TFP growth.

6.3.1. Sensitivity Analyses

To test the robustness of our results, in Table 3 we conduct a series of sensitivity checks. Table 3 is set out in the same format as Table 2, with all results for the EU-13 weighted average.

In row 1 we repeat our baseline estimates from Table 2. In row 2, in case of double counting with R&D,[44] we test the sensitivity of our results to the exclusion of all analytical occupations (group 4). Relative to the baseline, the wedge in labor productivity growth and the contribution of software and data capital deepening are reduced by approximately 50 percent.

In row 3, we undertake a similar check, this time reducing the time-use factor for analytical occupations by 50 percent (to 33 percent of time). Relative to the baseline, the wedge in labor productivity growth and the contribution of software and data capital deepening are reduced, but by less than in row 2.

In row 4, we consider potential double counting with another knowledge asset already recorded as GFCF in national accounts. GFCF in mineral exploration could include costs incurred in data-building and analytics.[45] Therefore, in case of any double counting, we subtract 50 percent of employment in the mining and quarrying sector[46] (NACE 05 to 09) from the employment values and wage costs

[43] Our new estimates of the contribution of software and data capital deepening are artificially raised in the earlier years of our estimation by our introduction of newly identified investment in 2011. Of course, investments in data transformation and knowledge creation were occurring before 2011. This therefore creates a discontinuity in capital services with growth higher than it would have been had our estimates of $P_N N^*$ extended further back. One option to remove the discontinuity would be to backcast our new estimate of investment with the existing measured series, but that would implicitly assume that the share of data investment in total software and data investment was the same in earlier years as in later years, which does not seem correct for a growing activity. However, the fast depreciation rate (0.315) used means that this effect is removed or depreciated away in later years. For comparison, Table A17 in Appendix G presents estimates for just 2014–2016, which minimizes this initial years effect.

[44] Although we think potential for double-counting with R&D is limited for reasons outlined above in Section 3 and the negative correlation presented in Appendix H.

[45] We thank an anonymous referee for pointing out the potential of double counting with mineral exploration.

[46] Which can be interpreted either as: 50 percent of employment in mining and quarrying; or 100 percent of employment with an assumed time-use factor of 50 percent. We use our estimate of the high-medium attainment wage rate from EUKLEMS to estimate.

that feed into our estimates of investment. Relative to the baseline, the wedge in labor productivity growth and the contribution of software and data capital deepening are reduced by amounts similar to row 3.

The method in row 4 probably over-estimates any potential double counting with mineral exploration as not all employees in that industry will be engaged in capital formation. Therefore, in row 5, we conduct an alternative check. In the EUKLEMS data, investment in mineral exploration is aggregated with investment in artistic originals to form GFCF in "Other Intellectual Property Products (IPPs)". Separate data on GFCF in mineral exploration is not available for most European countries. We do however have the observation from Goodridge *et al.* (2016) that, in the UK in 2014, 17.4 percent of GFCF in Other IPPs was in mineral exploration.[47] We therefore subtract 20 percent of GFCF in Other IPPs from our estimate of investment in software and data. Relative to the baseline, estimates of the wedge to labor productivity growth and capital deepening are unchanged.

The final two rows test sensitivity to our assumed time-use factors. In row 6 we set all time-use factors to 50 percent, which is consistent with OECD-Eurostat (2020) recommendation for estimating own-account capital formation in software and databases based on the input of software and database professionals. Relative to the baseline, growth in labor productivity and the contribution of software and data capital deepening increase slightly. In row 7 we set all time-use factors to 100 percent, which is consistent with all observed occupations spending all of their time on capital formation. Relative to the baseline, the wedge in labor productivity growth is more than doubled and that in software and data capital deepening increases more than four-fold.

7. Conclusions

Investments in the transformation and analysis of data are a key aspect of the latest wave of the ICT revolution and related to developments in artificial intelligence (AI) and the Internet of Things (IoT). Databases have been recognized as productive capital assets in the System of National Accounts (SNA) since 1993.

However, SNA and OECD recommendations for capitalization are limited to the cost of the database management system (DBMS, which is software) and the cost of transferring the data to the format required by the DBMS.

In this paper we extend the definition of the asset boundary to incorporate capital formation activity in data transformation and data analytics, where both processes create produced (information and knowledge) assets.

Applying our framework to EU-28 countries, we find that over half (57 percent) of employment engaged in an expanded definition of (software and) data capital formation is already accounted for in the measurement of own-account investment in software and databases. The remaining 43 percent is generally not incorporated in national accounts measurement, where the asset boundary is narrower than that used in this paper. Our analysis shows that activity in data capital

[47] We recognise however that the composition of GFCF in Other IPPs is likely to vary widely between countries.

formation currently outside the national accounts asset boundary is growing faster than national accounts measures in a number of countries.

Our main findings are as follows. First, we find that, in 2011–2018, 1.4 percent of EU-28 employment was engaged in the formation of (software and) data assets, ranging from 3.5 percent in Luxembourg to 0.5 percent in Greece. Second, in 2011–2018, mean growth in employment engaged in (software and) data capital formation in the EU-28 was 5 percent pa, ranging from 12.9 percent pa in Portugal to −2.4 percent pa in Latvia. Third, expanding the definition of investment in software and data raises own-account GFCF in the EU-16 by 61 percent in 2011–2016. Fourth, in 2011–2016, mean growth in real expanded investment in own-account software and data assets in the EU-16 was 6.9 percent pa, compared to 2.7 percent pa in the narrower definition used in national accounts. Fifth, in the context of growth-accounting, incorporating a wider definition of data capital changes both output and input. In the EU-13, in 2011–2016: (i) labor productivity growth is raised from 0.79 percent pa to 0.83 percent pa, which translates to €6.7bn pa of additional output growth in 2016 if applied to the EU-28 aggregate; and (ii) the contribution of capital deepening in software and data assets is raised over three-fold, from 0.03 percent pa to 0.1 percent pa, which translates to €9.4bn pa in 2016 if applied to the EU-28 aggregate.

References

Ackoff, R., “From Data to Wisdom,” *Journal of Applied Systems Analysis*, 16, 3–9, 1989.

Ahmad, N. T., *Measurement of Databases in the National Accounts: An Issue Paper Prepared for the December 2004 Meeting of the Advisory Expert Group on National Accounts*, Technical Report, OECD, 2004.

Ahmad, N., *Results of the AEG e-Discussion on Measurement of Database in National Accounts*. Technical Report, 2005a. https://unstats.un.org/unsd/nationalaccount/aeg/papers/m3reportDatabases.PDF.

______, *Issues Paper for the AEG July 2005 – SNA Update Issue 12 Follow-up to the Measurement of Databases in the National Accounts*, Technical Report, OECD, 2005b.

Ahmad, N. and P. Van De Ven, *Recording and Measuring Data in the System of National Accounts Meeting of the Informal Advisory Group on Measuring GDP in a Digitalised Economy*, Working Party on National Accounts, OECD, 2018.

Arrow, K. J., *The Economics of Information*, Harvard University Press, Cambridge, MA, 1984.

Bakhshi, H., A. Bravo-Biosca, and J. Mateos–Garcia, *Inside the Datavores: Estimating the Effect of Data and Online Analytics on Firm Performance*, Technical report, Nesta, 2014.

Boisot, M. and A. Canals, “Data, Information and Knowledge: Have We Got it Right?” *Journal of Evolutionary Economics*, 14, 43–67, 2004.

Chamberlin, G., A. Chesson, T. Clayton, and S. Farooqui, “Survey Based Measures of Software Investment in the UK,” *Economic Trends*, 627, 61–72, 2006.

Chamberlin, G., T. Clayton, and S. Farooqui, “New Measures of UK Private Sector Software Investment,” *Economic and Labour Market Review*, 1, 17–28, 2007.

Corrado, C., J. Haskel, M. Iommi, and C. Jona-Lasinio, “Intangible capital, Innovation, and Productivity à la Jorgenson: Evidence from Europe and the United States,” in B. M. Fraumeni (ed), *Measuring Economic Growth and Productivity: Foundations, KLEMS Production Models, and Extensions*, Academic Press, Cambridge, MA, 363–85, 2020.

Corrado, C., C. Hulten, and D. Sichel, “Measuring Capital and Technology: An Expanded Framework,” in C. Corrado, J. Haltiwanger, and D. Sichel (eds), *Measuring Capital in the New Economy, Studies in Income and Wealth No. 65*, University of Chicago Press, Chicago, IL, 11–46, 2005.

______, “Intangible Capital and U.S. Economic Growth,” *Review of Income and Wealth*, 55, 661–85, 2009.

Danmarks Statistik, *Danish GDP and GNI Sources and Methods 2012*, Technical report, Danmarks Statistik, 2016.

Fransman, M., "Information, Knowledge, Vision and Theories of the Firm," in G. Dosi, D. J. Teece, and J. Chytry (eds), *Technology, Organization, and Competitiveness: Perspectives on Industrial and Corporate Change*, Oxford University Press, Oxford, 147–91, 1998.

Fukao, K., T. Miyagawa, K. Mukai, Y. Shinoda, and K. Tonogi, "Intangible Investment in Japan: Measurement and Contribution to Economic Growth," *Review of Income and Wealth*, 55, 717–36, 2009.

Gantz, J. and D. Reinsel, *The Digital Universe Decade: Are You Ready?*, Technical report, IDC, 2010.

Goodridge, P., O. Chebli, and J. Haskel, *Measuring Activity in Big Data: New Estimates of Big Data Employment in the UK Market Sector*, Working Papers 25158, Imperial College London, 2015.

Goodridge, P. and J. Haskel, *How Does Big Data Affect GDP? Theory and Evidence for the UK*, Discussion Paper 2015/06, Imperial College London, 2015a.

______, *How Much Is UK Business Investing in Big Data?*, Discussion Paper 2015/05, Imperial College London, 2015b.

Goodridge, P., J. Haskel, and G. Wallis, *UK Innovation Index 2014*, Nesta Working Paper No. 14/07, 2014.

______, *UK Intangible Investment and Growth: New Measures of UK Investment in Knowledge Assets and Intellectual Property Rights*, Discussion Paper 2016/08, Imperial College London, 2016.

Jones, C. I., "Growth and Ideas," in P. Aghion and S. N. Durlauf (eds), *Handbook of Economic Growth*, North-Holland, Amsterdam, 1063–111, 2005.

Machlup, F., *The Production and Distribution of Knowledge in the United States*, Princeton University Press, Princeton, NJ, 1962.

Marrano, M. G., J. Haskel, and G. Wallis, "What Happened to the Knowledge Economy? ICT, Intangible Investment and Britain's Productivity Record Revisited," *Review of Income and Wealth*, 55, 686–716, 2009.

Mayer-Schönberger, V. and K. Cukier, *Big Data: A Revolution that will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt, Boston, MA, 2013.

McCrae, A. and D. Roberts, *National Accounts Articles: Impact of Blue Book 2019 Changes on Gross Fixed Capital Formation and Business Investment*, Technical report, ONS, 2019.

Mokyr, J., *The Knowledge Society: Theoretical and Historical Underpinnings*, Paper presented to the Ad Hoc Group on Knowledge Systems, United Nation, New York, 2003.

Nguyen, D. and M. Paczos, *Measuring the Economic Value of Data and Cross-border Data Flows: A Business Perspective*, OECD Digital Economy Papers, No. 297, OECD Publishing, Paris, 2020. https://doi.org/10.1787/6345995e-en.

OECD, *Handbook on Deriving Capital Measures of Intellectual Property Products*, OECD Publishing, 2010.

______, *Joint Eurostat—OECD Task Force on Land and other Non-Financial Assets. Report on Intellectual Property Products*, Technical report, OECD Statistics and Data Directorate, Committee on Statistics and Statistical Policy, 2020.

Rassier, D. G., R. J. Kornfeld, and E. H. Strassner, *Treatment of Data in National Accounts*, Technical report, Paper prepared for the BEA Advisory Committee, 2019.

Romer, P. M., "Endogenous Technological Change," *Journal of Political Economy*, 98, 71–102, 1990.

______, "Two Strategies for Economic Development: Using Ideas and Producing Ideas," *World Bank Economic Review*, 6, 63–91, 1992.

Shapiro, C. and H. R. Varian, *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Cambridge, MA, 1998.

Statistics Canada, *Measuring Investment in Data, Databases and Data Science: Conceptual Framework*, Technical report, Statistics Canada, 2019a.

______, *The Value of Data in Canada: Experimental Estimates*, Technical report, Statistics Canada, 2019b.

Stehrer, R., Bykova, A., Jäger, K., Reiter, O., and Schwarzhappel, M., *Industry Level Growth and Productivity Data with Special Focus on Intangible Assets: Report on Methodologies and Data Construction for the EU KLEMS Release 2019*, Technical Report, The Vienna Institute for International Economic Studies, 2019.

United Nations, *System of National Accounts 2008*, 2008.

Van De Ven, P., "Present and Future Challenges to the System of National Accounts: Linking Micro and Macro," *Review of Income and Wealth*, 63, S266–86, 2017.

Wong, D., *Data is the Next Frontier, Analytics the New Tool*, Technical Report, Big Innovation Centre, London, 2012.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site:

Supplementary Material

Appendix A. Measurement of investment in software and databases in practice

Appendix B. Detailed description of dataset based on occupations

Appendix C. EU LFS data

Appendix D. ISCO-08 catagories

Appendix E. Occupation shares, by country, no time-use adjustment

Appendix F. % of GFCF in software and databases that is own-account

Appendix G. Decomposition of growth (2014-16)

Appendix H. Correlation between software and data investment (PNNGHE, this paper) and investment in other assets, 2011-17