OECD Economics Department Working Papers No. 1731

Measuring data as an asset: Framework, methods and preliminary estimates

Carol Corrado,
Jonathan Haskel,
Massimiliano Iommi,
Cecilia Jona-Lasinio

https://dx.doi.org/10.1787/b840fb01-en

OECD

Organisation for Economic Co-operation and Development

ECO/WKP(2022)32

Unclassified **English - Or. English**

15 November 2022

ECONOMICS DEPARTMENT

MEASURING DATA AS AN ASSET: FRAMEWORK, METHODS AND PRELIMINARY ESTIMATES

ECONOMICS DEPARTMENT WORKING PAPERS No. 1731

By Carol Corrado, Jonathan Haskel, Massimiliano Iommi and Cecilia Jona-Lasinio

Authorised for publication by Luiz de Mello, Director, Policy Studies Branch, Economics Department.

All Economics Department Working Papers are available at www.oecd.org/eco/workingpapers.

JT03507686

This document was produced with the financial assistance of the German Ministry of Justice and Consumer Protection. The views expressed herein do not necessarily represent the official views of the German Ministry of Justice and Consumer Protection, or of the Organisation for Economic Co-operation and Development and its member countries.

ABSTRACT / RESUME

Measuring data as an asset: Framework, methods and preliminary estimates

Data are shown to generate efficiency gains but to have been unevenly shared across firms and households, and the subpar economic performance of most advanced economies (prior to the pandemic) has been attributed to increased market power originating, at least in part, from the increased use of data. To sharpen our understanding of these divergent perceptions of the modern digital age, this paper puts the recent increase in use of digitized information, i.e., data, into an economic framework amenable to measurement and analysis.
Data is conceptualized as an intangible asset: a storable factor input that is only partially captured in existing macroeconomic and financial statistics. Our proposed framework treats data as an intangible asset that contributes to final production in an economy.
This paper provides the conceptual groundwork that is needed for defining and measuring data investments. We also provide a review of methods that are used to measure data, and we offer an experimental implementation of our framework.

We also develop preliminary estimates of data assets intended to fully encompass the “intelligence” or “knowledge” generated by the use of data that are coherent with national accounts data at the industry-level of analysis as well as with measures of intangibles developed by EUKLEMS-INTANProd.[1]

JEL classification codes: O47, E22, E01
Keywords: intangible capital, data, innovation, productivity growth

****

Mesurer les données en tant qu'actif : cadre, méthodes et estimations préliminaires

Il a été démontré que les données génèrent des gains d'efficacité mais que ces derniers sont inégalement réparties entre les entreprises et les ménages, et que la performance économique médiocre de la plupart des économies avancées (avant la pandémie) pourrait être attribuée à un pouvoir de marché accru provenant, au moins en partie, de l'utilisation accrue des données. Afin d’améliorer notre compréhension de ces effets associés à l'ère numérique moderne, cet article place la récente augmentation de l'utilisation de l'information numérisée, c'est-à-dire des données, dans un cadre économique se prêtant à la mesure et à l'analyse.

Les données sont conceptualisées comme un actif intangible : un facteur de production stockable qui n'est que partiellement pris en compte dans les statistiques macroéconomiques et financières existantes. Le cadre que nous proposons traite les données comme un actif immatériel qui contribue à la production finale dans une économie. Cet article fournit ainsi les bases conceptuelles nécessaires pour définir et mesurer les investissements en données. Nous passons également en revue les méthodes utilisées pour mesurer les données et nous proposons une mise en œuvre expérimentale de notre cadre. Nous développons également des estimations préliminaires des données en tant qu’actifs qui englobent pleinement "l’intelligence" ou la "connaissance" générée par l'utilisation de données de façon cohérentes avec les données des comptes nationaux au niveau de l'industrie ainsi qu'avec les mesures des biens incorporels développées par EUKLEMS-INTANProd.

JEL classification codes: O47, E22, E01
Keywords: intangible capital, data, innovation, productivity growth

[1] https://euklems-intanprod-llee.luiss.it/

Table of contents

Tables

Figures

Boxes

Measuring data as an asset: Framework, methods and preliminary estimates

By Carol Corrado, Jonathan Haskel, Massimiliano Iommi and Cecilia Jona-Lasinio[2]

Introduction

1. A defining aspect of the digital age is the use of data, specifically large stores of digitized information referred to as "bigdata". Much popular work on bigdata appears in the business strategy literature. By a very long way, the best-selling book on the subject is "Big Data: A Revolution That Will Transform How We Live, Work, and Think" by Mayer-Schönberger and Cukier (2013). The book makes several statistical claims, suggesting that data will be used to disprove some casually held causal intuition and reduce many measurement problems. The Economist Magazine (2017) has asserted that data is the new oil, and Google's Eric Schmidt is quoted as stating that as much data/information is being created every two days as was created from the dawn of civilization to 2003 (Wong 2012). Such statements suggest that data have significant impacts on economic activity.

2. Business strategists draw upon what technologists refer to as the "data stack," or "information value chain," which depicts the transformation of raw records to usable information and knowledge via modern digital technologies (e.g., see Varian 2018). Practitioners make strong claims about the ease with which business users can deploy such technologies, in line with observers (e.g., Brynjolfsson and McAfee 2014) who describe the transformational promise of data-intensive digital tools, e.g., deep learning algorithms that adjust themselves to perform better as they are exposed to more data.[3] The gains from increased data use are also viewed to have been unevenly shared across firms and households, and the subpar economic performance of most advanced economies (prior to the pandemic) has been attributed to increased market power originating, at least in part, from the increased use of data (e.g., see Chapter 4

[2] Corresponding authors : Carol Corrado (Carol.Corrado@conference-board.org) from LUISS Lab of European Economics (LLEE) and The Conference Board , Jonathan Haskel (j.haskel@imperial.ac.uk) from the Imperial College Business School and the Bank of England, Massimiliano Iommi (miommi@luiss.it) from LUISS Lab of European Economics and the Italian Institute of Statistics (ISTAT) and Cecilia Jona-Lasinio (cjonalasinio@luiss.it) from LUISS University and the Italian Institute of Statistics (ISTAT). The authors thank Giuseppe Nicoletti, Dan Andrews, Alain de Serres and Luiz de Mello (Economics Department) for fruitful discussions and insights, and John Mitchell (Statistics and Data Directorate), Mauro Pisu (Economics Department) and Simon Lange and Vincenzo Spiezia (Directorate for Science, Technology and Innovation), for helpful comments on preliminary versions of the paper. They also thank delegates and participants to the OECD meetings on the "Value of Data" (May 2020) and "Data Shaping Firms and Markets" (June 2021) for helpful comments and suggestions. The authors are indebted to Ines Gómez-Palacio for editorial assistance. All remaining errors and omissions are authors' responsibility.

[3] Artificial intelligence (AI), machine learning, neural networks, and deep learning are terms often used interchangeably to describe software programs that behave intelligently. In a nutshell, each is a subset of the prior term, i.e., deep learning has been called "scalable machine learning." For definitions, examples, and further information see IBM's Cloud Blog (2020).

of the 2019 IMF *World Economic Outlook* report). To sharpen our understanding of these divergent perceptions of the modern digital age, this paper puts the recent increase in use of digitized information, i.e., data, into an economic framework amenable to measurement and analysis.

3. Data is conceptualized as an intangible asset in this paper: a storable factor input that is only partially captured in existing macroeconomic and financial statistics. We address the measurement challenge of valuing data assets from a resource cost perspective and develop estimates of data capital based on asset creation in each "layer" of the technologists' data stack. This covers the formation of all data processed, transformed, and used in an economy. The resulting estimates of data capital assets may be used in productivity analysis along with industry-level (mostly tangible) capital estimates currently available in national accounts. They are also compatible with the industry-level estimates of intangible assets available via the INTAN-Invest online database, though an open question—addressed in the paper—is the conceptual and practical overlap between data capital and intangible capital.[4]

4. Available statistics are not up to the demands of analysing the role of data and intangibles in generating value in modern economics. Official macroeconomic data do currently include investments in some intangible assets, e.g., notably R&D and computer software (which may include databases developed within firms) but their recognition in national accounts is relatively recent, and many intangible assets, including significant data assets remain excluded (see Box 1 below). Though the move to recognize intangible assets in macroeconomic data seems slow, the pace of their recognition in company-level financial accounts is tectonic and, worse, the treatment incoherent from an economic perspective (see Box 2 below). Some of the largest and fastest growing internet companies (Google, Facebook, Twitter etc.) are built mainly on the economics of transforming personal information into business and marketing intelligence, and the valuation of data is viewed with keen interest by policymakers and macroeconomic analysts. Still, neither accounting standards nor government surveys of "production" ask businesses to record their use of intangible or data assets, much less to record their investments and purchases of such assets on balance sheets.

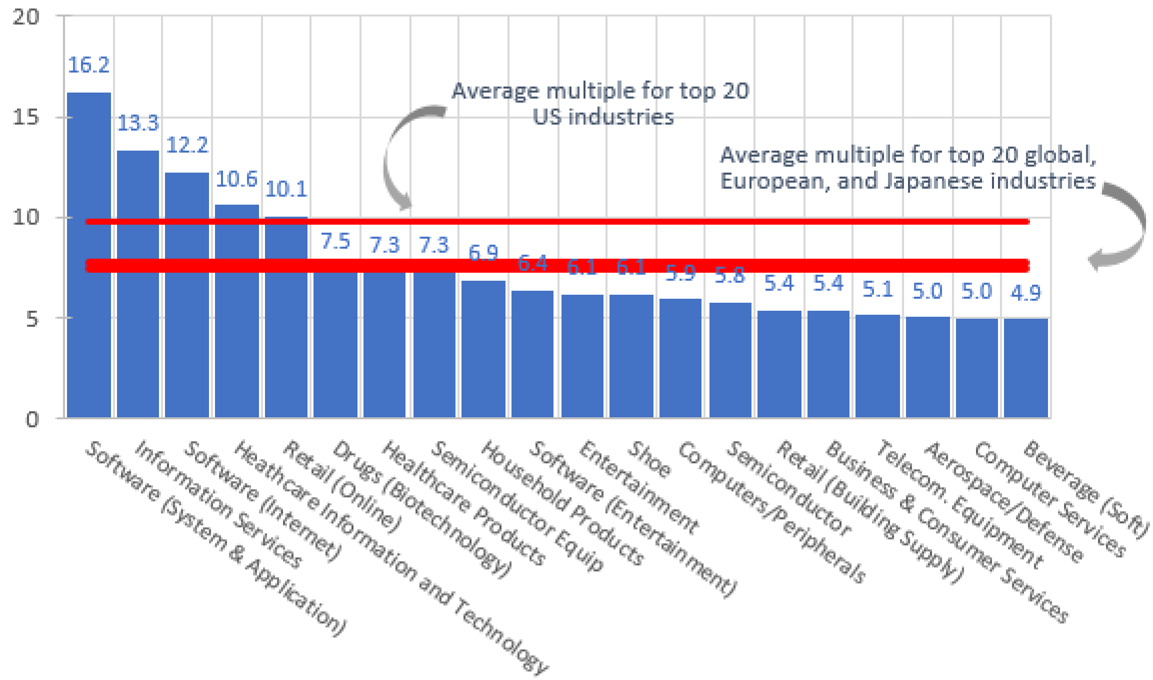
5. Though it is typical for the market value of nonfinancial companies to exceed the value recorded in financial statements (or national accounts) for technical valuation reasons and/or the impacts of brand recognition, such reasons do not reconcile *growing* differences. Indeed, analysis of the widening gap between equity market and accounting valuation during the 1990s led to the detection of software systems and business practices as intangible assets driving many of the most innovative companies of the time (Lev 2001); see also Hulten (2010), who used an intangible capital framework and company financial statistics to analyse the hidden value driving the astonishing growth and performance of Microsoft from 1988 to 2006.

6. Figure 1 sheds light on the hidden value in today's global industries by examining market-to-book multiples for individual nonfinancial industries. This figure highlights three phenomena. First and perhaps unsurprisingly, the software systems and applications industry, which includes Microsoft, Oracle, and Dropbox along with many lesser-known AI-powered software companies post the largest global market-to-book multiple. Second, companies driven mainly by their transformation and use of personal data, e.g., Google/Alphabet, Facebook, Baidu, and Tencent are grouped in the software entertainment industry, which has a multiple of nearly 6-1/2. While this sits a bit below the average global multiple for the top 20

[4] INTAN-Invest refers to the cross-country harmonized industry-level estimates of intangible investment available at www.intaninvest.net. The coverage of these estimates is undergoing improvement with the release of the updated EUKLEMS production accounts in December 2021. The update extends the usual production accounts provided by the project to include productivity estimates that incorporate cross-country harmonized estimates of intangible capital (INTANPROD). The EUKLEMS update and INTANPROD expansion is funded by a grant from the European Commission's Directorate-General for Economic and Financial affairs (DG ECFIN). Further information is available at either www.intaninvest.net or www.euklems.net.

industries (7.6), it is well above that for all nonfinancial industries (2.5, not shown).[5] Finally, besides these software and data producing industries, there is substantial "hidden value" in industries ranging from online retailing to the design of new drugs and specialized equipment—industries whose engineering and production processes are driven increasingly by data.

Figure 1. Industries with "Hidden Value": Market-to-book multiples for the top 20 global Industries (ranked by size of multiple) in 2020



Note*:* Series plotted are ratios of industry level enterprise value (market value of equity plus debt) to invested capital (book value of equity and debt less cash). The top 20 ranking is roughly the top quartile of multiples calculated for 82 detailed global nonfinancial industries, which excludes the finance, insurance, and real estate industries. Global data are derived from more than 41,623 global public companies; figures for the average top 20 multiple in United States, Europe and Japan are based on 6,253, 6,035 and 3,706 public companies, respectively. Source data pertain the latest full year available as of December 2020. Companies included in each industry are listed here: http://www.stern.nyu.edu/~adamodar/pc/datasets/indname.xls

Source: Authors' calculations based on data accessed 10/15/21 from Damodaran online. https://pages.stern.nyu.edu/~adamodar/New_Home_Page/home.htm

7. This paper sets out to unearth the role of data in the "hidden value" of modern corporations using an approach that puts data into a framework amenable to comprehensive measurement. As previously mentioned, the framework treats data as an intangible asset that contributes to final production in an economy. Like frameworks for measuring intangible capital geared toward usage with national accounts e.g., Corrado, Hulten, and Sichel (2005, 2009), our framework for data capital is designed to augment existing macroeconomic statistics to include investments in the full complement of data assets and digital tools.

8. A first step in designing a comprehensive framework to measure data as an asset is of course to define data assets and determine the scope of the economic activities that need to be captured. Though data capital is not explicitly considered a fixed asset in both national statistics and available measures for intangible capital, there are important overlaps between data capital and the existing assets in both

[5] The global top 20 average is a tad below Europe's top 20 industry average (7.7) and above that for Japan (7.4).

systems, e,g., software, and the design of a data measurement framework requires assessing the scope of its overlap with assets in existing systems (Corrado, 2021). The next steps—to determine the most appropriate method and statistical sources to use to estimate data capital, including feasibility via experimental implementation—flow in large part from these scoping assessments.

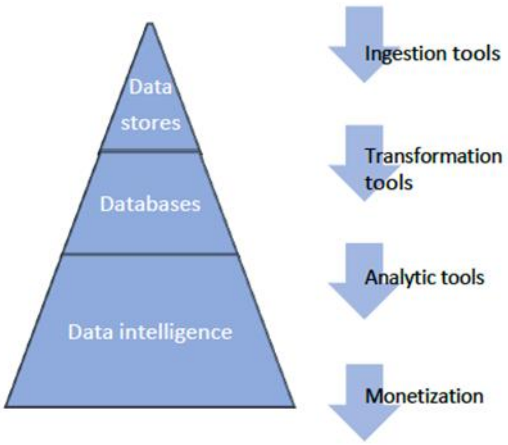
9. This paper provides, we believe, much of the conceptual groundwork that is needed for defining and measuring data investments, i.e., targeting activities that need to be captured and the extent to which they are found in existing systems. We also provide a review of methods that are used to measure data, and we offer an experimental implementation of our framework. The measures of data investment developed in this paper are preliminary, but they are comprehensive measures that are intended to fully encompass the "intelligence" or "knowledge" generated by the use of data; related works (Statistics Canada 2019a, 2019b; Goodridge, Haskel, and Edquist 2021) seem to fall short in this regard, missing major aspects of business technical, computer and engineering design intelligence. The estimates reported in this paper are also notable in that they are coherent with national accounts data at the industry-level of analysis.

10. In the next section, we set out the data-as-an asset framework and discuss its alignment with the intangible assets framework; we also provide information on coverage of data, explicitly or implicitly, in national and company accounts. We then review existing methods to measure the value of data, and in section four, we provide our experimental estimates for investments in data capital. These cover six major European economies: Germany, France, Italy, Spain, Sweden, and the United Kingdom from 2012 to 2018.[6] A final section concludes and discusses directions for future research.

Data Value Creation and the Intangibles Framework

11. Data does not provide a flow of services to production simply because records of economic transactions accumulate at an astonishing pace at little to no cost. The accumulation of digital by-product data within an economy has the potential to boost real output only when it also invests in transforming such records, possibly along with other available economic or social information, into analytical insights and actionable intelligence.

Figure 2. Data value creation: Data assets, digital tools, and monetization



[6] The investment estimates reported in this paper are under further development, including expansion to additional countries, for use in comparative productivity analysis.

Note: The stack to the left depicts the stages of data asset value creation based on applying layers of tools shown on the right.

Data value creation

12. As explained in greater context in Corrado, Haskel, Iommi, and Jona-Lasinio (2021), we identify three types of data assets: raw records, databases, and data intelligence. These are based on three layers of value identified in the data stack set out by technologists or the information value chain discussed by management strategists. These frameworks for data value creation are summarized via the depiction in Figure 2.

13. Data value creation involves the application of successive "layers" of data technologies to generate data assets of productive value. These assets are shown on the left as three major types of data that are embedded in social and economic systems. The corresponding sequencing of tools/technologies used in the creation of each layer of value in the "data stack" are shown on the right.

14. As may be seen, the data asset stack has three layers of value—data stores, databases, and data intelligence—each corresponding to a set of tools used in its creation. The three types of value are defined more precisely as follows:

- *Data stores* are raw records that have been stored but not yet cleaned, formatted, or transformed for analysis, e.g., data scraped from the web or sensor and economic data captured from production or transactions activity. Raw records also cover the raw data collected from experiments, statistical surveys, or administrative records.
- *Databases* consist of transformed raw data, records that have been cleaned, formatted, and structured such that they are suitable for using with visualization tools or some form of data analytics.
- *Data intelligence* reflects the further integration of data with advanced analytic tools (e.g., machine learning algorithms, or deeper; see again footnote 1); data intelligence is a set of quantitative inputs that provide actionable guidance for decision-makers, including solutions to scientific problems.

15. Figure 2 depicts *monetization* as an application stage of data value creation. This stage features prominently in the business literature and refers to an organization's capability for implementing actions guided by data intelligence. These actions are extremely relevant for understanding the impacts of data capital. Data monetization mainly reflects observed (though not explicitly identified) adjustments to existing factor inputs—labor, capital (tangible or intangible), and intermediates—in response to the increased use of data capital.

16. The measurement of the value chain concepts depicted in Figure 2 is thus approached by developing estimates of the largely "missing" activities. These are set out explicitly in table 1. As previously indicated, there are three data asset types included in data capital, and the task is to estimate the *value of investment* in these assets, including the costs of tools used to create them.

Table 1. Data as an asset: investment activities and assets produced

	Activity	Asset Type	Comments
	(1)	(2)	(3)
1	Generation/collection	Data stores ("raw" records)	Experiments and platforms that create and/or collect data
2	Aggregation	Databases (query-ready)	Processes and platforms for combining and curating data from multiple sources
3	Analysis/Analytics	Data intelligence (actionable)	Gleaning of insights from data that can be acted upon

Intangible capital

17. Data assets are largely subsumed—though not explicitly identified—in the intangible capital framework set out by Corrado, Hulten, and Sichel (2005, 2009). Intangible investment covers a wide class of investments, from databases to business processes, engineering design, and market research, that would appear to be relevant for analysing the consequences of the increased use of data in economies. In this section we explore this approach to measuring and analysing data by considering the definitional/conceptual overlap between the data assets enumerated in table 1 and activities covered by existing measures of intangible assets.

18. The identified asset types in the Corrado, Hulten and Sichel framework are set out in Table 2. Column 1 of the table shows that there are three major categories of intangible assets: digitized information, innovative property, and economic competencies. Column 2 reports specific assets used to populate each major category, and column 3 reports the intangible assets covered in national accounts. As may be seen, only a subset is included (lines 1 through 5).

19. At first blush one might infer from column 1 of table 2 that the digitized information grouping of intangible assets includes the data stores and databases assets listed in table 1, but as may be seen in the itemized list in column 2 of table 2, only databases appear. National accounts estimates of the value of investment in databases will thus exclude the cost of acquiring or ingesting the data stores they contain, and outright purchases of data stores and databases are only included to the extent they are embedded or sold as software products (see Box 1 for elaboration). Furthermore, the national accounts of most countries do not publish databases as a unique asset category. A feature of the national accounts combined “software & databases” measure, however, is its coverage of investments in digital tools used to create data assets (see again Figure 2).

Table 2. Intangible investment: major categories and asset types

Categories (1)	Investment by Asset Type (2)	NA (3)	Examples of Assets and Property (4)
Digitized Information	1. Software	Yes	Digital capabilities, tools
	2. Databases	Yes	Trade secrets (data)
Innovative Property	3. Research and development (R&D)	Yes	Patents, licenses
	4. Mineral exploration	Yes	Mineral rights
	5. Artistic, entertainment, and literary originals	Yes	Copyrights, licenses
	6. Attributed designs (industrial)	No	Patents, trademarks
	7. Financial product development	No	Trademarks, software patents
Economic Competencies	8. Brand and market research	No	Brand equity, customer lists, market insights
	9. Business process and organizational practices	No	Operating models and platforms, supply chains and distribution networks, and management and employee practices
	10. Employer-provided training	No	Firm-specific human capital

Note. Column 3 indicates whether the asset type is currently included as investment in national accounts (NA).
Source: Updated version Corrado, Hulten and Sichel (2005) as set out in Corrado (2021).

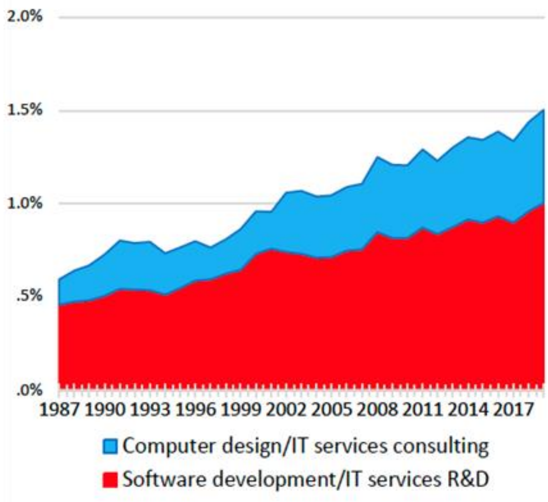
20. Consider now data intelligence, the most valuable, final stage of the data value chain per Figure 2. This is where the utility of the intangibles framework becomes apparent. The knowledge created from data encompasses all modern, data-driven business, financial, marketing, engineering, and scientific intelligence. The inclusion of business operations, financial products, and general marketing intelligence in intangible capital is seen via lines 7, 8 and 9 of Table 2. An increase in the use of data capital in R&D activities (line 3), will cover novel forms of data-derived scientific intelligence (including the development of new AI techniques); it will exclude, however, the application of modern engineering data analysis that produce systems used in future production but whose solutions are not sufficiently novel to be included in R&D. Such investments are covered in the intangible framework via line 6, architectural and engineering design, and line 9, which includes certain investments in computer design, e.g., investments to re-engineer computer systems and computer network platforms to make use of cloud infrastructure services, data analytic services, and data.

21. The intangible capital perspective sheds key insights on the economic activities that generate data assets. While Figure 2 and table 1 depict overall data asset-generating activities, the intangibles framework captures many forms of data intelligence—virtually all assets in the framework are potentially data driven. Seen from another perspective, prior works seeking to implement the data value chain seem to miss some key application areas of modern data science (Statistics Canada 2019a, 2019b; Goodridge, Haskel, and Edquist 2021) that are covered in the intangible investment framework. While financial "engineering" and marketing forms of intelligence are explicitly included in these prior works, architectural/industrial engineering and computer design forms of data intelligence are not, at least not comprehensively.

22. Partial evidence for the notion that intangible capital is picking up data intelligence not covered by R&D is provided in Figure 3. The figure shows two series that Byrne and Corrado (2017) argued captured, at least in part, the data driven demand for cloud services. The two series are business R&D in IT services and software development and purchases of computer and network design consulting services; these are underlying components of the R&D and business process investment intangible investment categories listed on Table 2 (lines 3 and 9).

Figure 3. Intangible investment components related to data, cloud services, and AI

Investment as percent of GVA, private industries, 1987 to 2019



Source: Authors' update of Byrne and Corrado (2017)

23. As may be seen, these data driven components of intangible investment are very dynamic, having nearly tripled relative to private sector GDP over the period shown. Their share relative to total GDP is 1.2 percent in 2018, which would not include *public* funding for AI research, suggesting that the true contribution of AI software research to total GDP is higher.

24. In summary, beyond the main message of this section that data capital is largely subsumed within intangible capital, key findings regarding the measurement of data capital are as follows:

- Data value creation involves the generation of data assets--data stores, databases, and data intelligence—including the tools used to create them
- Data stores and most forms of data intelligence are not being captured in official statistics.
- Data intelligence has many forms—business, marketing, engineering, and scientific—and not all forms have been comprehensively measured in previous works.
- Data intelligence is the most valuable, and final, stage of the data value chain as it pertains to investments in modern digital business practices and engineering-based systems.

Box 1. Data and intangibles in national accounts

In the System of National Accounts (SNA, United Nations 2008), fixed assets include only a subset of the intangible assets identified in the intangible capital framework set out by Corrado, Hulten, and Sichel (2005, 2009) (referred to as intellectual property products in the SNA): digitized information (software and database) and innovative properties excluding design and new financial products. Expenditures for purchasing other intangibles not included in the SNA fixed asset boundary (such as brand or attributed design) are considered intermediate consumption, thus not classified as an investment (gross fixed capital formation in national accounts) and not included in GDP. Production of other intangibles made in-house goes completely unrecorded, as only the own-account output of final products (i.e., household final consumption expenditure or fixed investment) is included in the SNA production boundary. Finally, the other intangibles are not included in capital stocks and do not generate any depreciation (consumption of fixed capital in national accounts terminology).

The SNA defines databases as “files of data organized in such a way as to permit resource-effective access and use of the data and recognizes that they may be developed exclusively for own use or sale as an entity or sale through a license to access the information contained.” The value of databases, either produced for own use or sale, should not include the cost of the database management system (DBMS), which should be treated as a computer software asset.

The 2008 SNA recommends an inconsistent treatment for data depending on whether a database is developed for own use or for sale. The value of databases produced for own final use should be estimated by a sum-of-costs approach and should only include the cost of preparing data in the appropriate format but not the cost of acquiring or producing the data. On the other hand, databases for sale should be valued at their market price, which includes the value of the information content.

The national accounts definition of databases is consistent with the data definition proposed in this paper, which considers three different data assets: data stores, databases, and data intelligence. On the other hand, the SNA asset boundary does not explicitly include data stores (except for its potential inclusion on the value of databases for sale) and data intelligence (except for its possible overlap with software and R&D).

An expansion of the asset boundary to include data in the next edition of the SNA is under discussion as part of the process of updating the 2008 SNA. At the time of writing, the task team on digitalization recommends updating the 2008 SNA to include data in the production and asset boundaries (ISWGNA 2020; see also the discussion in Goodridge et al 2021).

The ISWGNA[7] proposes the following statistical definition for data: "data is information content that is produced by collecting, recording, organizing, and storing observable phenomena in a digital format, which can be accessed electronically for reference or processing. Data from which its owner(s) derive economic benefits by using it in production for at least one year is an asset."

Two crucial points emerge from the proposed definition of data. First, the revised asset boundary will explicitly include data stores and databases but not data intelligence. Second, only long-lived data qualify as fixed assets and should be capitalized.

[7] ISWGNA stands for Intersecretariat Working Group on National Accounts. It is one interagency body set up by the United Nations Statistical Commission (UNSC) to enhance cooperation among international organisations and comprises Eurostat, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, and the World Bank. The ISWGNA has received its mandate from the United Nations Statistical Commission (UNSC) to, among other tasks, revise and update the SNA.

Box 2. Data in financial reports

Financial statements are vital to informing economic policy, whether used directly to conduct economic analyses or as the primary source to produce business surveys (as in the case of EU Structural Business Statistics in almost all European countries). In addition, financial accounts are a relevant data source for compiling annual national accounts.

In the business community, there is growing consensus that intangibles and data assets are the main drivers of total value and growth for modern organizations, and an increasing number of firms conduct data valuation initiatives. However, while the results of data valuations are widely used to inform business management, current accounting standards rarely recognize data and other intangible investments as capital assets (Zéghal and Maaloul 2011). In brief, the accounting rules allow to capitalize intangibles that are purchased or obtained when mergers or acquisitions occur (in the latter case, usually as part of goodwill). International Financial Reporting Standards, IFRS (but not the US Generally Accepted Accounting Principles, GAAP) allow to capitalize at cost some of internally-generated intangibles (such as development costs, the "D" of R&D), but only if they meet six different criteria (such as there is robust evidence of future economic benefits, and costs can be reliably determined).

Lev (2019) shows that data comparability across firms might be somewhat difficult because of the current treatment of intangibles. First, not recognizing intangibles as assets means that they are expensed (i.e., treated as costs in the income statement), thereby understating the earnings and assets of intangibles-growing firms and overstating the earnings and assets of intangibles-declining enterprises. Second, the fundamental inconsistency between the accounting treatment of internally generated (expensed) and acquired intangibles (capitalized) precludes a meaningful comparison of performances between companies with different innovation strategies (internal generation vs. acquisition).

Including acquired data assets in goodwill is a standard practice in business accounting. For example, Reinsdorf and Ribarsky (2019) reported that in 2017, Nielsen's[8] revenue was USD 6.6 billion while its balance sheet included only USD 168 million of data assets. These figures were recorded when Nielsen acquired Gracenote in 2017 for USD 585 million. One of the primary motivations for Nielsen to acquire Gracenote was to access its global content database. However, a large portion of the cost of acquiring Gracenote was allocated to goodwill and amortizable intangible assets.

Mazzi et al. (2019) investigates financial statements of listed companies from more than 20 countries (20,475 firm-year observations) adopting International Financial Reporting Standards in 2005 or later. They find that for the period 2006 to 2015, in conforming to the requirements and conditions set out in international accounting standards, most companies either fully or partly expense expenditures referred to R&D. Therefore, R&D expenditure, as included in the balance sheets, underestimates the real amount of R&D expenditure and thus of intangible assets. In addition, there are country- and industry-level differences in R&D capitalization, further hindering cross-industry and cross-country comparisons.

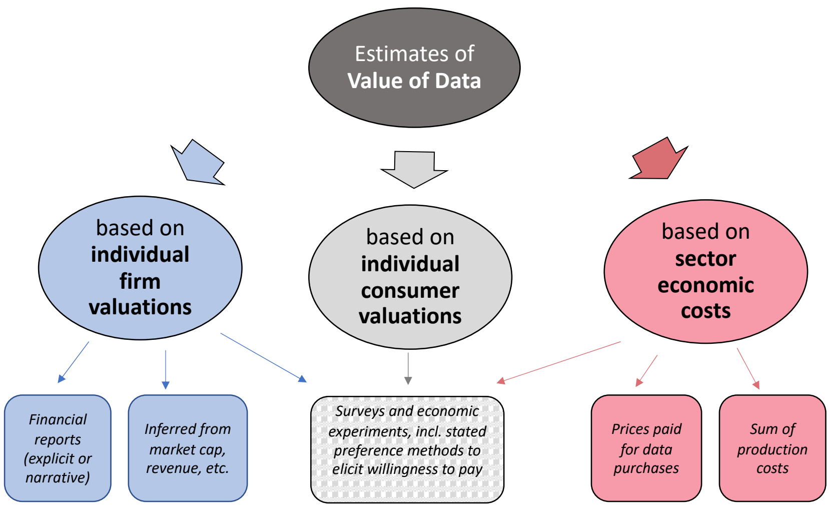
Current accounting treatment of data and intangibles hampers the quality of firm-level data based on financial statements. First, the value of data and intangible assets in firm-level data is underestimated. Second, much of the observed variability across firms is likely to reflect different accounting practices or business strategies (internally generated vs. purchased vs. mergers and acquisitions). Finally, data are not fully comparable across industries and countries.

Measuring data: approaches and methods

25. How much value do firms derive from data? And how is this related to the value of personal information or to costs sustained by firms to obtain the data that are used and/or created via the data stack?

26. In addressing these questions, one encounters different perspectives and different measurement approaches to the valuation of data. The economics literature has taken three main directions to develop estimates of the value of data. As depicted in the middle panel of figure 4, these include approaches based on individual firm valuations, approaches based on consumers' valuations, and approaches based on sector economic costs. The bottom panel of the figure indicates methods used under each approach. Surveys and economic experiments (the middle box in the bottom panel) are of course methods that are not unique to a given approach, as the figure indicates.

Figure 4. Approaches to the valuation of data



[8] Nielsen is a world leader in market research and ratings, whose business is based on an extensive foundation of proprietary data assets.

27. Approaches aimed at valuing consumers' personal information will not encompass the full data value chain of Figure 2, which covers all digitized information in an economy.[9] Our review of methods is targeted at those that can yield comprehensive coverage of data use in market activities in economies, and we thus proceed as follows: We first discuss methods that have been used to estimate the value of data for individual firms and/or based on individual firm-level data. That is followed by a discussion of stated preference methods applied to measure the value of data used in business—a little appreciated niche in this literature, though long used by market researchers to study consumer preferences and recently ingeniously employed by Brynjolfsson, Collis, and Eggars (2019) to estimate the value of free digital goods.

28. We then briefly summarize the gist of the sector cost approach as deployed by national accountants. This approach, also called the sum-of-costs approach, is used to develop the experimental estimates of data investment reported in section 4.

Methods Based on Firm-level Estimates

29. Below we review data valuation approaches used and/or emerging in financial reporting, followed by a review of methods used in key studies. These studies provide essential insights on measuring the value of data, even if their methods cannot be readily adapted to compile macroeconomic statistics sufficiently comprehensive to inform economic policy analysis.

Business reporting methods

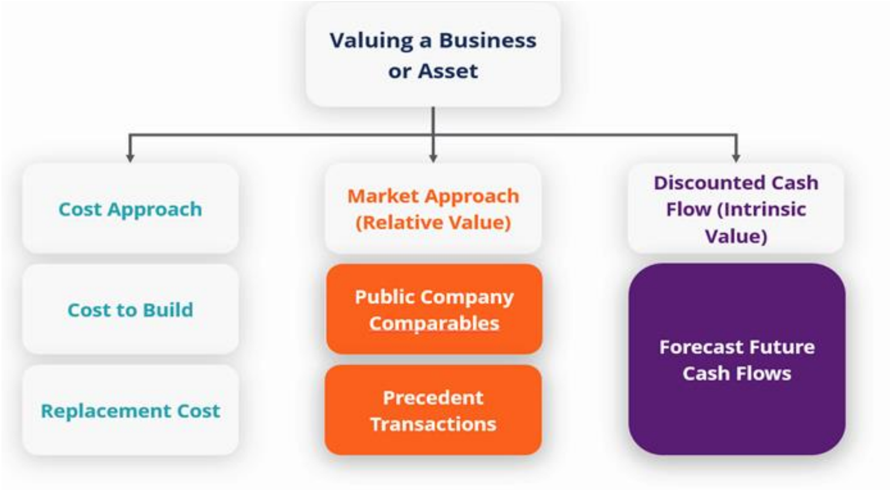
30. Most of the approaches adopted for valuing data in the business context consist of an implementation of the three traditional valuation methods used to value any asset type: income, market, and cost approaches (Figure 5). The income methodology measures the incremental cash flows (increased revenues and/or reduced costs) that the data are expected to generate in the future. The market approach captures the value of a given data asset using the information about the value of a comparable data asset whose value is observable in an active market or transaction. The cost approach estimates the value as the cost for recreating a replica of the data or replicating the data's utility.

31. The growing importance of intangibles in corporate activity and the evidence that they do not fit very well in the current financial reporting has generated a debate among the accounting community about the opportunity to deliver more information on intangibles promoting its disclosure of financial reporting (see, for instance, UK Financial Reporting Council 2019) or by capitalizing intangibles as assets in balance sheets (ACCA 2019, Lev 2019) The UK Financial Reporting Council (2019) proposes two ways to get more information on intangibles in financial reporting. One is to revise the statement of profit or loss to provide information on expenditure on future-oriented intangibles, analysed by nature. The other is the provision of more details on intangibles in the narrative sections of financial reporting.[10]

[9] That said, the "personal data value chain" is a conceptual construct that sits within the overall data value chain in which public open data and business-specific information also reside and contribute to value creation; see Corrado, Haskel, Iommi, and Jona-Lasinio (2021, section 2.4) for further elaboration.

[10] The term 'narrative reporting' includes reports with titles such as 'Management Commentary' or 'Strategic Report' generally forming part of the annual report, and also other information (such as Preliminary Earnings Announcements) that an entity provides primarily for the information of investors.

Figure 5. Valuation Techniques in financial accounting



Source*: Corporate Finance Institute. Image accessed 12/12/2021.*

32. The first option is more beneficial for compiling business statistics and for economic analysis based on firm-level data. First, it would facilitate gathering information via business surveys. Based on current financial reporting standards, respondents to business surveys would typically be unable to identify expenditures for data and several other intangibles separately. Second, improved and more comprehensive disclosure of spending on intangibles (in addition to the value of existing stocks) would be consistent with the needs of national accounts compilers of collecting information on outlays (not on the value of the assets). Finally, more precise information on expenditure for intangibles and data would be available to firm-level data users.

33. It is worth adding that there are advocates for a "stakeholder value approach" that aims at measuring the economic value created by the data for each stakeholder (including customers, employees, suppliers, communities, and the environment). For example, a stakeholder data valuation performed for Highways England[11] shows that the total value created through data is £39 billion, around 30 percent of Highways England's physical asset value (the strategic road network) (https://anmut.co.uk/why-you-should-be-treating-your-data-as-an-asset/). These and related data valuation approaches are possibly biased by subjective evaluations thus very difficult to generalize.

Revenue-based approaches

34. Another interesting approach suggested by Nguyen and Paczos (2020) aims at capturing the value of data on the basis of the revenue shares driven by data monetization across different types of firms (e.g., manufacturers, utility providers, banks, or online platforms). Nguyen and Paczos (2020) adopt a stylized taxonomy of business models distinguishing two main categories: data-enhanced or data-enabled. Their assumption is that by looking at the business models adopted in different productive sectors it is possible to identify specific characteristics from which to infer a general measure of the value of data at the industry level. This approach can be easily implemented even if it requires additional efforts from national statistical institutes to conduct ad-hoc economic surveys and coordinate internationally to guarantee comparable results across countries.

[11] The UK government-funded company that maintains England's motorways.

Depreciation-based approach

35. Coyle and Li (2021) develop a demand-side methodology for estimating the size of data markets using the recent finding that an online platform's entry can disrupt incumbent firms' organizational capital by affecting its depreciation rate (Li and Chi 2021). They calculate the stocks of organizational capital based on before-entry and after-entry depreciation rates. This difference captures the loss due to the failure of using data to cope with changes in competition due to the entry of an online platform. Thus, it can be used to measure the potential size of the demand for data by incumbent firms in the industry sectors disrupted by online platforms. In other words, they use the loss of the value incumbent firms' organizational capital to measure firms' maximum willingness to pay for the access to data.

36. Coyle and Li (2021) apply their model to study the impact of the entry of Airbnb on existing firms in the hospitality industry. They find that the market size for data in the global hospitality sector is USD 43 billion in 2018 and that this data market has also grown rapidly at an average growth rate of 35%, meaning that its size has been doubling in less than three years.

37. Consistent with the existing literature on measuring intangible capital from firm-level data, Coyle and Li (2021) use the selling, general, and administrative (SG&A) expenses as a proxy for a firm's investment in organizational capital. Firms report these expenses in their annual income statements. They include expenditures for employee training costs, brand enhancement activities, consulting fees, and supply chains' installation and management costs, thus covering the economic competencies category in the list of intangibles proposed by Corrado et al. (2005). On this basis, they estimate the value of data considering the extent to which online platform entry can disrupt incumbent firms' economic competencies assets.

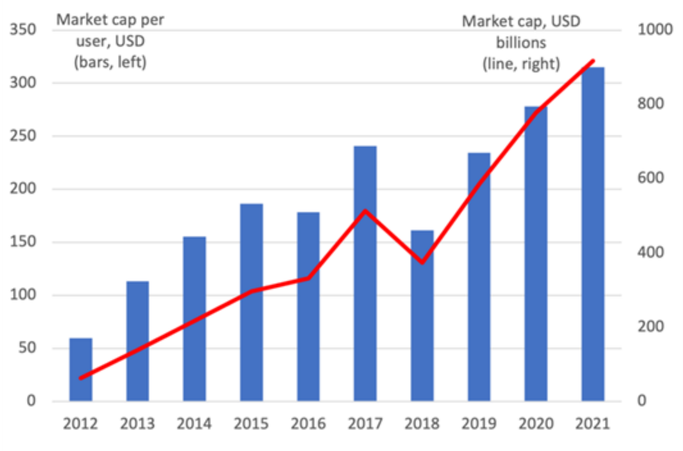
Market prices

38. Market prices paid and received in actual transactions are the best proxy for quantifying the value of data. However, the adoption of this approach might be complex. First, there is no well-defined market for many types of data, and, when available, transaction-based valuations may often rely on obsolete information. Second, as the value of data is highly context-dependent, the same dataset might be valued differently across different data suppliers, users, and regulators (Nguyen and Paczos 2020). Finally, market transactions in unprocessed data would only capture the input data and not the entire transformation chain necessary to generate digitized information (Reinsdorf and Ribarsky 2019).

39. Large-scale market transactions typically exist primarily for third-party data[12] produced by data brokerage or data aggregator companies. These companies usually collect information from publicly available personal records and then aggregate, store and sell it to different customers through licensing subscriptions or contractual arrangements. As third-party data is widely accessible, they are less valued than first and, to a lesser extent, second-party data (Reinsdorf and Ribarsky 2019).

[12] Input data can be classified based on how it is obtained: First-party are data collected by the business itself about its users or customers (e.g., cookie-based data on browsing activity or data on past purchases); Second-party data are essentially someone else's first-party data who can be obtained either by purchasing them from the company that owns it or by working out arrangements with trusted partners who are willing to share their customer data (e.g., a grocery store selling its customer loyalty data to a credit card company); third-party data is data that a firm can buy from large data aggregators that pull it from various other platforms and websites.

Figure 6. Facebook/Meta Market Cap per Active User, 2012-2021



Sources: Number of FB monthly active users from Statistica, Q4 of each year (Q3 for 2021). Market capitalization for Facebook/Meta is Nasdaq value for December 11 of each year.

40. It is also illustrative to examine financial indicators per record from companies that derive nearly all of their income from advertising linked to personal data, e.g., Facebook/Meta. Figure 6 shows that the value of an individual (active) record currently is more than 300 USD. The firm's valuation is approaching 1 trillion USD.

41. Ahmad, Ribarsky, and Reinsdorf (2017) calculate a value equivalent to around 0.02 percent of global GDP for the user data collected by five major digital services (Facebook, Twitter, Instagram, LinkedIn, and Gmail) based on the number of active users and assumed prices of a user profile. The estimate was based on the maximum user profile price obtainable from a calculator available from the Financial Times that used industry pricing data from various sources in the United States of USD.

Stated Preference Methods

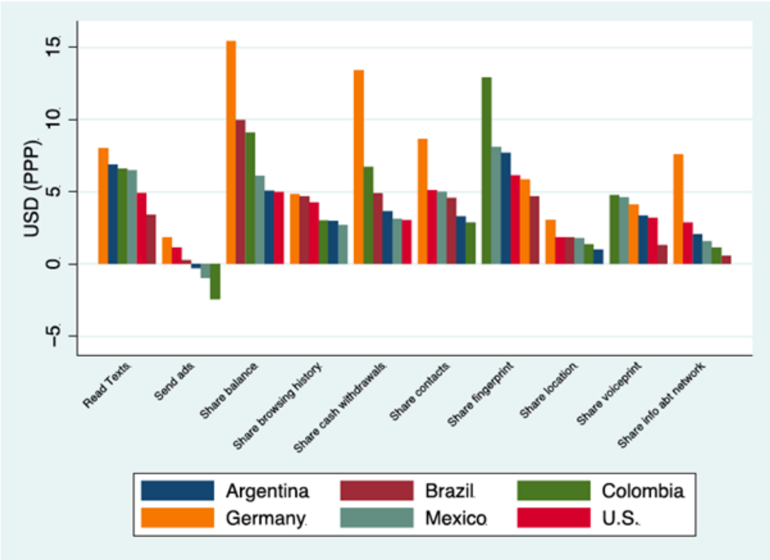
42. Some studies have provided estimates of data value using stated preference methods (including contingent valuation, conjoint analysis, and discrete choice analysis). This approach asks surveys participants to directly report their willingness to pay (WTP) to obtain a specified good or willingness to accept (WTA) to give up a good. The value of a non-market good or service is the amount that users are "willing to pay" for it, or "willing to accept" in return for not having it. Contingent valuation methods are widely used to understand consumer valuations and preferences in contexts with no monetary prices, such as environmental or cultural goods (see, e.g., Carson, Flores & Meade, 2001 and McFadden & Train, 2017 for surveys).

43. An excellent example to illustrate the use of stated preference methods for business valuation of data is the case of Landsat. The Landsat program consists of a series of Earth-observing satellite missions jointly managed by NASA and the US Geological Survey. Landsat data products are processed and made available for download to all users at no cost. Based on surveys of data users, Miller et al. (2013) estimated the economic benefit of Landsat data for the year 2011 to be $1.79 billion for US users and $400 million for international users. The annual benefit to US users is two times greater than the cost of building and launching Landsat-8, the Land-sat satellite launched in 2013 and still operating.

44. From a different perspective, a growing literature relies on stated preferences methods to study the monetary valuation of privacy. Prince and Wallsten (2020) conducted a discrete choice survey across six countries: the United States, Mexico, Brazil, Colombia, Argentina, and Germany. They find that

Germany places the highest value on privacy compared to the US and Latin American countries. Across countries, people place the highest value on keeping financial and biometric information private—balance and fingerprint data in particular. Germany's status as the country with the highest value for privacy is driven mainly by strong preferences for keeping financial data private. German respondents were willing to share bank balance information in exchange for monthly payments of $15.43 and cash withdrawal information for $13.42/month.

Figure 7. Average payment consumers would demand for permission to share data across countries by feature



Source: Prince and Wallsten (2020), Figure 2, page 6.

45. Stated preference methods are also used to assess the value of public information assets, e.g., official statistics. The United Nations Economic Commission for Europe (UNECE 2018) has called on national statistical agencies to develop approaches to calculate the monetary value of official statistics, which cannot be measured using market prices as many official statistics datasets are accessible under public license with no monetary price. UNECE (2018) recommends various possible valuation methods, including using the stated preference method and reports that it was used to explore the economic value of the UK Economic and Social Data Service (ESDS). ESDS is a distributed service that aims to promote the broader and more informed use of data for research and teaching in social sciences. In the ESDS study, respondents were asked to express their willingness to pay in terms of an annual (subscription) fee and on a pay-per-access basis. This resulted in an estimated willingness to pay of around £25 million per annum among the survey population.

Sector economic costs

46. National accounts estimate investment by asset type based on economic costs. Though the approach differs substantially in context and application from the cost-based valuation method used in financial accounting (and depicted in Figure 5), the concepts do overlap. National accounts aim at consistently recording investment flows and capital stocks for each industry sector, and doing so involves estimating values for all sources of supply for each asset and deriving the asset valuations and quantities using information on price change in newly produced assets and information on the rate at which an asset's

value declines as it ages. The focus here, and in the application set out in section 4, is on the estimation of investment flows.

47. If firms purchased all or most data from market transactions, as they do with tangible assets, measuring the cost of data would be conceptually similar to measuring expenditures for a construction firm's purchase of excavators and concrete mixers. Instead, most digitized information used by businesses (and other intangibles such as software and R&D as well) is not transacted on markets but produced in-house. Thus, national accounts compilers must come up with two components, own-account investment (when data are produced and used in-house) and purchased investment (when data are bought and sold in market transactions) to measure nominal investment flows in data assets. Let us take each component in turn, beginning with how in-house production of data assets might be estimated.

A factory within a factory

48. Imagine a firm having a "software factory" or "R&D factory" inside it—and your task is to estimate the gross output of this hypothetical factory based on the market value of the payments made to factors employed by it (labor, capital, and intermediates). The key to accomplishing this task is to identify the occupations of workers employed in the factory and to estimate their compensation. Based on knowledge of the compensation paid to these workers, the total payments made to all factors involved in the in-house production can be estimated. As a practical matter, the identified workers may not be involved in producing new assets their entire workday; for example, the conventional approach to measuring in-house software production in national accounts is to assume that software developers spend just 50 percent of their time working in their firm's "software factory" to produce original code. In-house production of data assets can be estimated in a similar fashion.

49. The SNA explicitly recommends that national statistical offices use the sum-of-costs approach to estimate software and databases (unless produced for sale) and R&D (unless the market value of the R&D is observed directly) and the own-account component of any product for which it is not possible to find the price of a similar product. The INTAN-Invest database uses a sum-of-costs approach to estimate the own-account component of non-national accounts intangibles.

Purchased data assets

50. Purchased data should be valued at the transaction price. Although conceptually simple, measuring the purchased component of data investment is challenging because comprehensive data sources are scant. All told, information about the expenditures on data usually is missing in surveys of production or capital spending, and the national accountant's total supply approach is difficult to implement. Ker and Mazzini (2020) considered business statistics sources and looked at the revenues generated by firms that create explicit value from data (those collecting, compiling, and selling databases). But they found that focusing mainly on industry classifications is likely to generate an inexact identification of these activities. For example, Zillow sells its data on home real estate valuations, Nielsen sells it survey data, as do credit agencies such as Experian, but these firms are in widely different industries, and monetizing databases is not necessarily the primary line of business for many firms who charge for purchased databases or data intelligence.

This study vs prior studies using sum-of-costs approach

51. Statistics Canada (2019a, 2019b) prepared experimental estimates of in-house investments in data based on a sum-of-costs approach, counting in effect all production as in-house production. Occupational groups were selected from among those generally associated with converting observations into digital format (the process of digitization). Their estimated values for investments in all three data types ranged from 1-3/4 to 2-1/4 percent of the country's GDP in 2018.

52. Goodridge at al. (2021) took the same approach and estimated the combined value of software and databases (from national accounts) and other data capital investments for 16 EU countries using essentially the same implementation in terms of occupations covered. Their results suggest that including the Statistics Canada grouping of occupations engaged in producing data stores and data intelligence (which they refer to as data transformation and knowledge creation) raises own-account GFCF by around 60 percent compared to own-account investment in software and databases measured in EU official national accounts.

53. In the next section, we implement a sum-of-costs approach to estimate in-house production of three types of data assets. Our identification of workers engaged in producing data intelligence yields a somewhat broader list of occupations than used in previous works. In line with the intangible capital framework, our estimates of data intelligence include data-driven engineering design. The joint evolution of engineering design (ED) and data science is discussed and subjected to meta analysis in Chiarello, Belingheri, and Fantoni (2021). As they report, though ED is recognized as a key element of the innovation process at-large, only in recent years has data-driven design become more prominent due to developments in AI. With increased competition, increased digitization of manufacturing, coupled with new methodologies to collect data on product characteristics, product performance and customer requirements, ED is often accompanied by big data.

Value of data investment in selected European countries

54. As explained above, we implement a cost-based approach coherent with the intangible framework illustrated in section 2 to generate experimental estimates of investment in data stores and data intelligence for the market sector of six European economies (Germany, France, Spain, Italy, Sweden and UK) over the years 2012-2018.[13] To the best of our knowledge, these are the first harmonized and internationally comparable measures of data investment produced so far for these countries. Notice that although experimental and preliminary, our approach is a new attempt for defining a measurement framework consistent with national accounts for computing time series of investments in data assets, including estimates in volume terms (i.e., adjusted to consider price changes) as well as data capital stock measures.

55. Measures of data asset have been generated using the information on total in-house costs incurred for transforming the data (to generate data stores) and for data analytics (to develop data intelligence) considering the occupation types engaged in producing data-related assets. Consequently, our measures of data assets capture the amount of data value produced in the market sector regardless of whether the produced output is intended for own final use or final sale. In what follows, we consider the produced value of data as a good proxy for data investment in the market sector assuming that data transactions between the government and the market sector are rather small. [14]

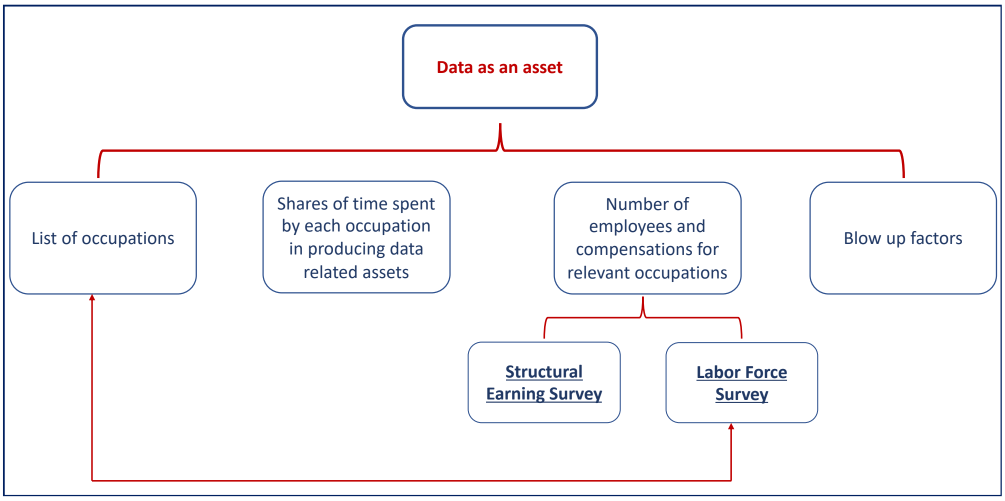
56. To develop the estimates of data investment it is necessary to collect the information illustrated in Figure 8. The main data sources are the EU Structure of Earnings Survey (SES) for 2014 and the EU Labor Force Survey (LFS) for 2012-2018. The SES provides information on the number of employees by

[13] We define market sector as all industries excluding NACE sections O (public administration and defense; compulsory social security), P (education), Q (human health and social work activities), T (activities of households as employers; undifferentiated goods - and services - producing activities of households for own use), and the imputed rents of owner-occupied dwellings component section L (real estate activities).

[14] Strictly speaking, estimating output based market sector investment would require adjusting the output for data transactions between government and the market sector as well as for imports and exports data flows. But information on international data trade is currently not available from official statistics.

occupation (at the three-digit level of the 2008 International Standard Classification of Occupations, ISCO) and economic activity and their annual earnings, while the LFS provides the number of persons employed.

Figure 8. Information used to measure investment in value data assets



57. These sources are used to identify a detailed list of occupations as well as the number of employees for the relevant occupations and their compensations. Then, making some assumptions about specific data-related occupations and industry it is possible to identify the shares of time spent by each specific occupation in producing data-related assets. Finally, we compute blow-up factors to account for other costs (intermediate consumption and gross operating surplus) to be used to get a measure of output consistent with national accounts definitions. All the steps of the calculation are described in appendix 1.

58. Compared to existing estimates from Goodridge et al. (2021), which derive estimates of average wages for workers with high, medium and low educational attainment in each country based on the EUKLEMS 2019 database,[15] we get a somewhat more precise result on the basis of the country-occupation wages from the SES.

59. Also, the list of occupations and assumptions to derive time-use factors has been defined merging the information from Statistics Canada (2019) with Goodridge et al. (2021) and making three main adjustments:

- First, as discussed in section 3, we select a wider list of occupations to account for the different types of data intelligence beyond data science, particularly science and marketing data intelligence. In addition to professions like mathematicians, statisticians, financial analysts, and economists, we have considered life, physical, and earth science professionals, engineers, and sales and marketing professionals.
- Second, our time-use factors are relatively more prudent as we aim at avoiding double-counting with national accounts estimates of computer software and databases, and to capture the output of long-lived data and data intelligence (as short-lived ones do not qualify as assets) only.
- Finally, the SES provides data on occupations only at the three-digit level of the ISCO classification but some occupational groups relevant for data-related asset production are only identifiable at the

[15] More precisely, they use the average wage of workers with high and medium (combined) educational attainment for software and database professionals, other ICT professionals and analytical occupations, and the average wage for low educational attainment for data entry clerks.

four-digit level. Hence, we need adjust the time use factors to account for those occupational groups including workers not engaged in data assets production.

60. Overall, our measurement approach has the advantage of being fully consistent with national accounts thus allowing robust comparisons and integration of our findings with official data on computer software and database investment. However, the consistency with national accounts entails some limitations as it is not simple to quantify exactly the overlapping between our estimates of data stores and data intelligence and national accounts estimates of software and databases. Finally, a possible bias can be produced by the lack of harmonization in sources and methods used by national statistical institutes to estimate investment in software and databases.[16]

61. The sum-of-costs approach also has some well-known limitations. Ideally, the selection of the relevant occupations, time-use assumptions and blow-up factors should be based on empirical evidence (e.g., employees' time-use and tasks surveys and surveys on cost structure of intangibles and data production). As empirical information is very scant, they are based on assumptions in practice, which may hamper the cross-country comparability of national accounts if different countries use very different assumptions.

62. As the sum-of-cost approach based on labor costs for relevant occupations is used for different intangible assets, there is also the risk of overlapping between the estimates of data investment and other intangibles (e.g., some costs for data stores may already be included in the valuation of the databases or some costs for software in R&D). Ideally, double-counting would be avoided by a consistent allocation of the time-use of each occupation across different assets.

63. The next sub section illustrates our experimental findings and provides a first attempt of testing possible overlapping between non National Account intangible investment and investment in data assets.

Results

64. As mentioned above, our estimates cover the two components of data investments (data stores and data intelligence) likely not captured by national accounts investment in software and databases. Results are shown in Table 3 reporting GDP shares of the value of fixed investment in data assets for the market sector of six European countries in 2018. Columns (a) and (b) show our measures of investment in data stores and data intelligence and suggest that investment in data intelligence is systematically larger than investment in data stores in all sample economies. GDP shares of data intelligence varies from 1.9 percent in Spain to 3.2 percent in Sweden, while investment in data stores from 1.6 percent in Spain to 2.5 percent in Sweden. Summing up the value of data stores (a) and data intelligence (b) we get an estimate of the value of data investment (c) estimated with our cost-based approach and not currently captured by national accounts. Estimates in column (c) indicate that Sweden, UK and Germany are relatively data intensive (5 percent of GDP on average) and that data intelligence accounts for more than a half of the estimated value of data asset. Italy, France and Spain are below the total sample average (4.6 percent of GDP).

65. Including national account estimates of software and databases (column d) into the picture provides more mixed evidence across countries. Germany shows a rather small GDP share of software and database (0.7 percent of GDP) followed by the Mediterranean economies and UK (1.5 percent of GDP on average). The GDP shares are instead more than double for France and Sweden (3-3.5 percent).

66. Then adding national account software and database (d) to data capital (c) we get the total value of data capital (e). Estimated values suggest a different picture for data intensities compared to column (c)

[16] National statistical offices may use different methods to estimate the purchased component of software investment (expenditure approach or supply approach) and include different occupational groups and/or different time-use assumptions to estimate the own-account component of software and databases.

as the category software and database accounts for a large share of total data in Sweden and France making them the most data intensive countries. While UK and Italy are confirmed to be somewhat data intensive but below the sample average (6.6 percent of GDP). Germany and Spain lag a bit behind.

Table 3. Market sector investment in data and intangible assets, 2018

Share in GDP

	Data Stores	Data Intelligence	Data investment	NA SW & DB	Total data investment	Total Intangible Assets	Total Intangible Assets excl SW & DB	Total Intangible Assets excl SW & DB & R&D
	(a)	(b)	(c=a+b)	(d)	(e=c+d)	(f)	(g=f-d)	(h)
Germany	2.3%	2.8%	5.0%	0.7%	5.7%	8.1%	7.5%	5.0%
Spain	1.6%	1.9%	3.5%	1.6%	5.1%	6.0%	4.4%	3.5%
France	1.7%	2.2%	4.0%	3.0%	7.0%	11.1%	8.1%	6.1%
Italy	2.0%	2.7%	4.7%	1.4%	6.1%	8.5%	7.1%	5.7%
Sweden	2.5%	3.2%	5.7%	3.4%	9.0%	13.7%	10.3%	7.1%
United Kingdom	1.9%	3.0%	5.0%	1.6%	6.5%	9.7%	8.1%	7.0%

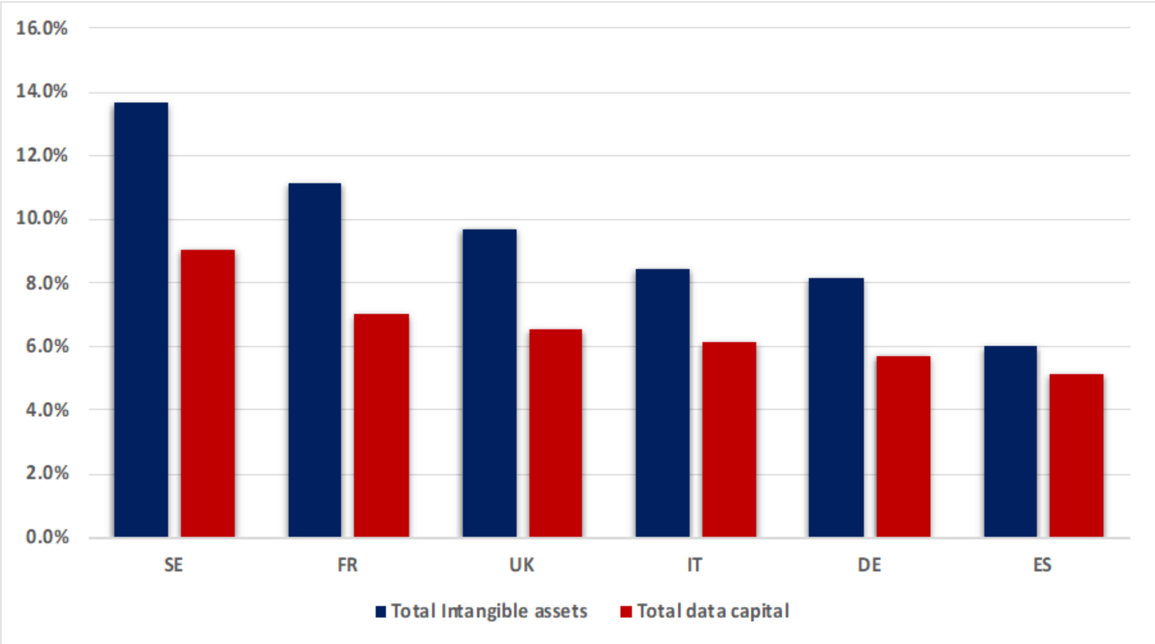
Note: Total intangible assets include all the asset types listed in table 2.

67. In this paper data is considered as an intangible asset but it is likely too early to assess the extent INTAN-Invest estimates of total intangibles of conceptual and empirical overlap between our estimated data investment and intangible investment. However, here we provide some comparisons between existing INTAN-Invest estimates of total intangibles (column f) also disaggregated by their main asset categories (columns g and h), and total investment in data assets (column e) to get the sense of possible overlaps. For the sake of clarity, we illustrate the comparisons in Figures 9 and 10 below.

68. Our experimental estimates of total data investment support the assumption that available measures of total intangible investment likely capture a large part of the value of data assets. Chart 9 shows, as expected, that total intangibles account for a larger share of GDP compared to total data investment in all sample economies.

Figure 9. Market sector total intangible assets vs total data investment, 2018

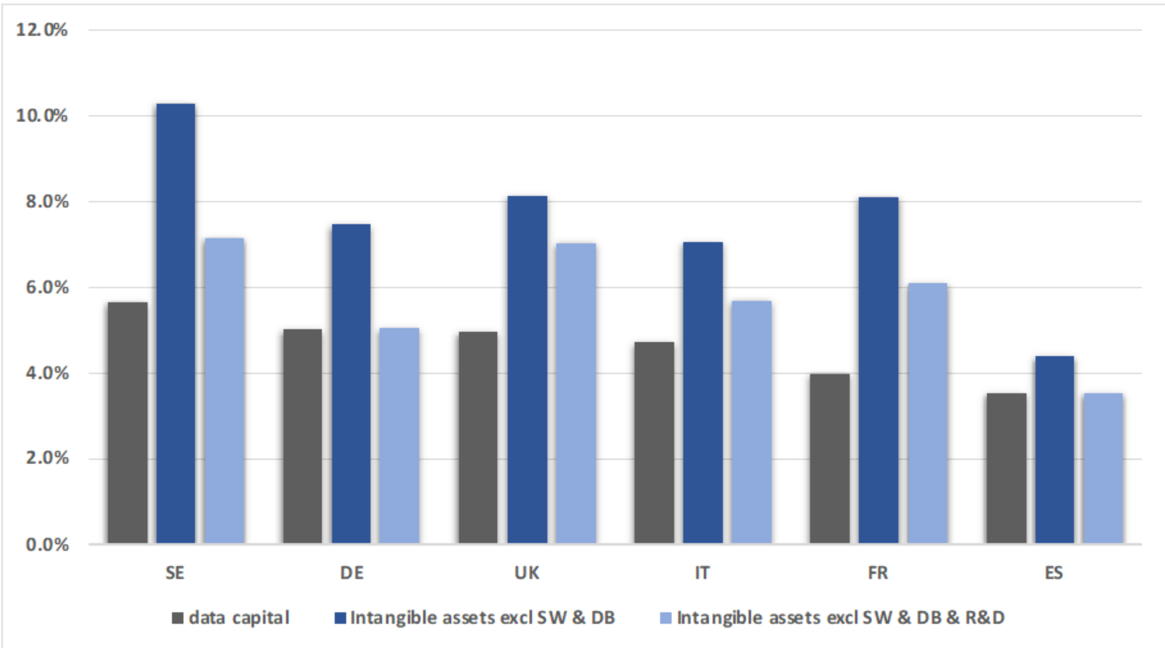
Share in GDP



Note: Total intangible assets include all the asset types listed in table 2.

Figure 10. Market sector data investment vs categories of intangibles, 2018

Share in GDP



69. Then to get some insights about the possible drivers of this difference, chart 10 compares our estimated value of data assets (data stores and data intelligence) with different categories of intangibles: total intangibles excluding national account software and databases and then excluding also R&D. When software and databases is subtracted from total intangibles, investment in intangible assets remain larger than data assets.

70. Further, also excluding R&D we get non national account intangibles, mainly economic competencies and design. The comparison with data investment suggests that intangibles still account for a larger share of GDP compared to data assets especially for the most intangible intensive economies (SE, FR and UK) while this is not the case for Germany and Spain.

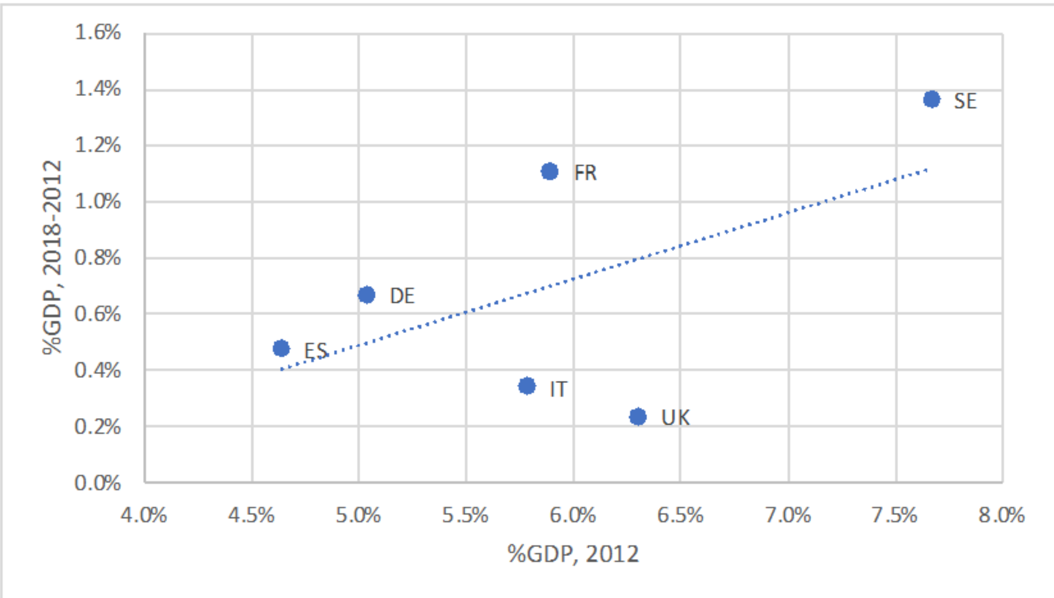
71. A careful look at charts 9 and 10 reveals that more intangible intensive economies are not necessarily high data capital intensive and that the gap between intangibles and data might be sizable. In digital intensive economies typically, economic competencies, provides an important contribution to shape competitive advantages, thus requiring large investment in organizational capital.

72. Figure 11 shows GDP shares of data investment for the sample economies in 2012 versus the average difference of the shares between 2012 and 2018. There is no clear pattern either of convergence or divergence in data investment intensity.

73. Sweden was already the most data intensive country in 2012 accounting for 7.7 percent of GDP and keeping its position of fast adopter of data asset until 2018 (the GDP share increased on average by 1.4 percentage in six years). Also, in France data investment kept growing fast since 2012 (1.1 percent on average) but starting from a smaller GDP share (5.8 percent). Notably, in UK and Italy the growth rates of data investment shares were below 0.5 percent. In 2012, Germany and Spain had a much lower data intensity than the other four countries and did not speed up.

74. Overall, data investment intensity has an heterogenous dynamics across the major EU economies making difficult to identify a common sample pattern at this stage.

Figure 11. Market sector investment intensity in data and software in 2012 vs. difference 2018-2012



Source: Authors' elaborations on EU Structure of Earnings Survey, Labour Force Survey and official national accounts.

75. To better understand the results illustrated above we look also at the composition of the data related occupations assumed to be the main producers of data stores and data intelligence. For this purpose, tables 4 and 5 show the structure of market sector output in data stores and data intelligence by type of occupation for the last available year (2014).[17] We have selected the list of occupations identified at three-digits ISCO minor groups reported in Table A1 in the appendix. Here, however, we illustrate more synthetic figures at the level of two-digits ISCO sub-major groups.

76. In the EU economies, the activity for generating data stores and data intelligence requires the contribution of various types of occupations besides science and engineering professionals. France is partially an exception, with science and engineering professionals accounting for 61 percent of output of data intelligence. On the other hand, about 50 percent of data intelligence can be attributed to business and administration associate professionals in Italy. The relatively high value for Business and administration associate professionals in Italy depends on the high share of Financial and mathematical associate professionals" (ISCO 331) on total occupations in the Italian Financial sector.

77. Additionally, notice that for some occupations, such as Social Professionals, there is no information available from the SES and LFS. Therefore, for these cases we have made some assumptions following the results in Statistics Canada (2019b) at the cost of no variability of the shares across countries.

78. The figures reported in Tables 4 and 5 have been computed making harmonized time-use assumptions and using coherent data sources. Therefore, the differences between the composition of occupations across countries likely reflect a different structure of employment by occupational groups[18].

Table 4. Composition of market sector output in data stores by occupation, 2014

	Germany	Spain	France	Italy	Sweden	United Kingdom
Science and engineering professionals	23%	23%	34%	14%	26%	23%
Business and administration professionals	14%	20%	7%	10%	14%	29%
Information and communications technology professionals	10%	9%	23%	11%	21%	19%
Social Professionals	7%	7%	7%	7%	7%	7%
Business and administration associate professionals	10%	6%	9%	28%	10%	9%
Information and communications technicians	4%	8%	4%	8%	6%	4%
Data-Related Clerical Support Workers	33%	27%	15%	23%	17%	8%

Source: Authors' elaborations on EU Structure of Earnings Survey, Labour Force Survey and official national accounts.
Note: The share of Social Professional is based on the results of Statistics Canada (2019)

[17] Composition of data investment by detailed type of occupation is only available for 2014, which is the year for which we have the information from the SES and that we have used as a benchmark year for our calculations.

[18] Further analysis of the SES 2014 is needed to determine the extent to which the result for France and Italy is due to high employment for these occupational groups being highly concentrated in one or few industries or being widespread in the whole market sector.

Table 5. Composition of market sector output in data intelligence by occupation, 2014

	Germany	Spain	France	Italy	Sweden	United Kingdom
Science and engineering professionals	41%	39%	61%	23%	43%	34%
Business and administration professionals	25%	36%	9%	16%	26%	38%
Social Professionals	13%	13%	13%	13%	13%	13%
Business and administration associate professionals	21%	12%	17%	48%	18%	15%

Source: Authors' elaborations on EU Structure of Earnings Survey, Labour Force Survey and official national accounts.
Note: The share of Social Professional is based on the results of Statistics Canada (2019).

A focus on the information and communication industry

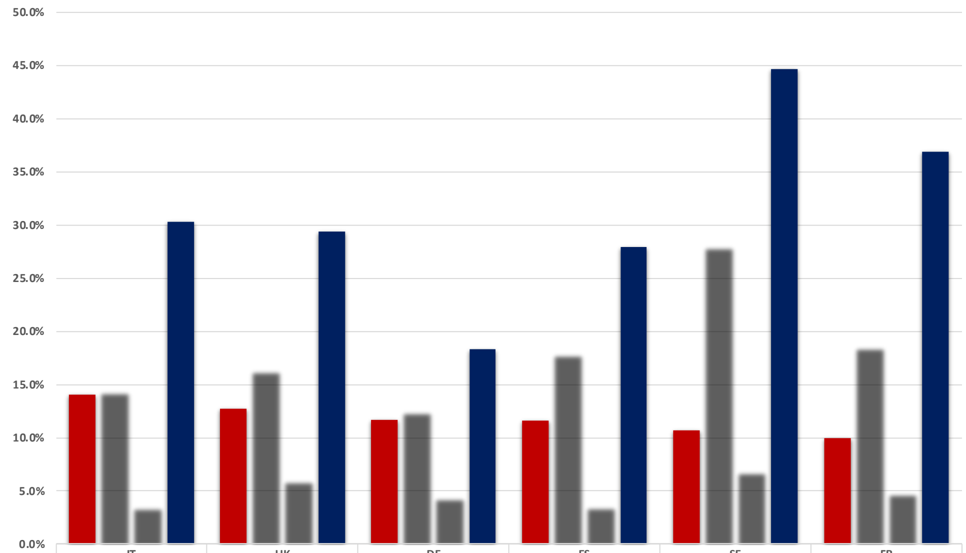
79. The information and communication industry (industry J) is an interesting case study for testing our estimates of data capital as it is at the centre of the digital transformation of modern economies. The capability of a country of leveraging the full productive capacity of its digital transformation relies substantially on the structure and characteristics of the information and communication industry producing a wide range of products such as: software and services as well as data processing and related activities. Traditionally, industry J provides significant contributions to labor productivity growth thus being also a target for innovation and digital policies.

80. In our sample, the value added of industry J accounts for around 6 percent of GDP in Sweden and UK and around 3 percent of GDP in Italy and Spain (Figure 12). However, data capital intensity seems to be uncorrelated with the size of the sector, as in Italy and Spain the information and technology sectors look rather relatively more data investment intensive than in Sweden and UK despite their smaller size.

81. Data intensity within the market sector[19] indicates that industry J accounts for a larger share in more intangible intensive economies (Sweden and France) corroborating the evidence that the big data companies operating in this sector play a main role in the digital transformation, as also shown in Figure 1.

[19] As captured by the share of data capital in industry J over data capital of the market sector in Figure 12.

Figure 12. Information and communication industry, investment in data and intangible assets, 2018



	IT	UK	DE	ES	SE	FR
Data capital	14.1%	12.7%	11.7%	11.6%	10.7%	9.9%
(Data capital)I / (Data capital) MKT	14.3%	16.2%	12.4%	17.8%	27.9%	18.4%
(VA)I / (GDP)	3.4%	5.9%	4.3%	3.4%	6.7%	4.7%
Total Intangible Assets	30.3%	29.4%	18.3%	27.9%	44.6%	36.9%

Note: *Data capital and total intangible is expressed as a percent of value-added in the sector. Total intangible assets include all the asset types listed in table 2.*

82. The main takeaway from our analysis is that existing estimates of intangible investments not included in National Accounts capture a large amount of data investment. More specifically, investment in data intelligence is largely captured by non National Account Intangibles and R&D as suggested by the high correlation between investment in data intelligence and total intangibles excluding software and databases in Table 6. On the other hand, the relatively lower correlation between data stores and intangibles excluding software and databases indicates that only a small fraction of investment in data stores is captured by non National Account intangibles. Further, the even smaller correlation between data stores and National Account measures of Software and DB suggests that they are currently not included in National Accounts[20].

83. Finally, these correlations and the empirical results illustrated above corroborate the assumption of strong complementarity between data intelligence, a key element for artificial intelligence, and non NA intangibles, such as such as design, training, and business process re-engineering (Corrado, et al. 2021).

[20] To test the significance of these results we have tested the same correlations on industry level data. Industry level findings support more aggregate correlation results, especially the lack of a significant relationship between data stores and national account software.

Table 6. Correlations between data assets and intangibles

	Data Stores	Data Intelligence	Data investment	National Account SW & DB	Total data investment	Total Intangible assets	Total Intangible Assets excl SW & DB	Total Intangible Assets excl SW & DB & R&D
Data Stores	1.00	0.84	0.92	0.12	0.60	0.56	0.72	0.48
Data Intelligence	0.84	1.00	0.99	0.10	0.62	0.61	0.81	0.78
Data investment	0.92	0.99	1.00	0.16	0.68	0.66	0.84	0.74
NA SW DB	0.12	0.10	0.16	1.00	0.84	0.82	0.59	0.52
Total data investment	0.60	0.62	0.68	0.84	1.00	0.98	0.91	0.80
Total Intangible assets	0.56	0.61	0.66	0.82	0.98	1.00	0.95	0.86
Total Intangible Assets excl SW & DB	0.72	0.81	0.84	0.59	0.91	0.95	1.00	0.91
Total Intangible Assets excl SW & DB & R&D	0.48	0.78	0.74	0.52	0.80	0.86	0.91	1.00

Note: Total intangible assets include all the asset types listed in table 2.

Directions for Future Research and Conclusions

84. Data are everywhere but in macroeconomic data and financial statements. National accounts do currently include investments in databases developed within firms, but they are valued excluding the cost of acquiring or producing the data they contain. In financial statements, data are mostly recognized as assets only when mergers or acquisitions occur and are included in goodwill most of the time.

85. This paper contributes to increasing data on data in two ways. First, we show that data assets are largely subsumed—though not explicitly identified—within the intangible capital framework adopted in the INTAN-Invest database, which extends the coverage of national accounts and includes investments in data-driven business and marketing intelligence. Second, applying the data-as-an-asset framework, we produce experimental - and new - explicit estimates of market sector nominal investment in data stores and data intelligence (the components of data not currently included in official national accounts) for Germany, France, Italy, Spain, Sweden, and UK in 2012-2018.

86. Future research should focus on refining the estimates using more detailed data on occupations, completing the accounts with investment in volume terms, capital stocks, and industry level estimates, expanding the country coverage to all or most EU27 countries and the US. Furthermore, separate estimates of investment in computer software and in databases should be produced. Doing this will also help to understand whether the striking differences in the intensity of investment in software and databases across countries as measured in official national accounts are real or reveal measurement issues.

References

ACCA - Association of Chartered Certified Accountants (2019), ""Business Reporting of Intangibles: Realistic proposals – Comments from ACCA", Accessed on 15 September 2021 at: https://www.accaglobal.com/gb/en/technical-activities/technical-resources-search/2019/may/frc-intangibles-response.html.

Brynjolfsson, Erik, Avinash Collis and Felix Eggers (2019), "Using Massive Online Choice Experiments to Measure Changes in Well-being," *Proceedings of the National Academy of Sciences* 116(15): 7250-7255.

Byrne, David and Carol Corrado (2017), "ICT Prices and ICT Services: What do they tell us about Productivity and Technology?",*I nternational Productivity Monitor*, Number 33 (Fall), p. 150-181.

Chiarello, Filippo, Paola Belingheri, and Gualtiero Fantoni (2021), "Data Science for Engineering Design: State of the Art and Future Directions," *Computers in Industry* 129 (103447).

Coyle, D. and Li WCY (2021), "The Data Economy: Market Size and Global Trade", ESCoE Discussion Paper No. 2021-09

Deloitte (2020), "Data Valuation: Understanding the Value of your Data Assets," Accessed on 15 March 2021 at: https://www2.deloitte.com/global/en/pages/finance/solutions/datavalue.html.

The Economist (2017), "Leader: The world's most valuable resource is no longer oil, but data," The Economist Group Limited (6 May).

Goodridge, P., J. Haskel, and H. Edquist (2021), "We See Data Everywhere Except in the Productivity Statistics," Review of Income and Wealth. https://doi.org/10.1111/roiw.12542

IBM Cloud Blog (2020), AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?, https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks (Accessed November 6, 2021).

ISWGNA (2020), "Recording and Valuation of Data in National Accounts", issue note prepared by the ISWGNA subgroup on digitalization for the 14th Meeting of the Advisory Expert Group on National Accounts, October 2020.

Lev, B. (2019), "Ending the Accounting-for-Intangibles Status Quo", European Accounting Review, 28:4, 713-736, DOI: 10.1080/09638180.2018.1521614

Li, WCY and P.J. Chi (2021), "Online Platforms' Creative "Disruption"" in Organizational Capital – The Accumulated Information of the Firm, Moon Economics Institute Discussion Paper, No. 1, August.

Loomis, John, Steve Koontz, Holly Miller, and Leslie Richardson (2015), "Valuing Geospatial Information: Using the Contingent Valuation Method to Estimate the Economic Benefits of Landsat Satellite Imagery," Photogrammetric Engineering & Remote Sensing 81 (8): 647–56. https://doi.org/10.14358/PERS.81.8.647.

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

Mazzi, F., R. Slack, I. Tsalavoutas and F. Tsoligkas (2019), "The capitalisation debate: R&D expenditure, disclosure content and quantity, and stakeholder views." ACCA research reports. Accessed on 15 march 2021 at: https://www.accaglobal.com/content/dam/ACCA_Global/professional-insights/Intangibles/pi-intangibles-R%26D.pdf.

Miller, Holly H., Leslie A. Richardson, Stephen R. Koontz, John Loomis, and Lynne Koontz (2013), "Users, Uses, and Value of Landsat Satellite Imagery: Results from the 2012 Survey of Users." https://doi.org/10.3133/ofr20131269.

OECD (2010), "Handbook on Deriving Capital Measures of Intellectual Property Products", OECD Publishing.

Prince J. and S. Wallsten (2020), "How Much is Privacy Worth Around the World and Across Platforms?", TPRC48: The 48th Research Conference on Communication, Information and Internet Policy. Accessed o 0 october 2021 at https://techpolicyinstitute.org/publications/privacy-and-security/how-much-is-privacy-worth-around-the-world-and-across-platforms-2/.

PwC (2019), "Putting a value on data." Accessed on 15 March 2021 at: https://www.pwc.co.uk/data-analytics/documents/putting-value-on-data.pdf.

Reinsdorf, M. and J. Ribarsky (2019), "Measuring the Digital Economy in Macroeconomic Statistics: The Role of Data", paper presented at the ASSA 2020 Annual Meeting

Reinsdorf, M. and P. Schreyer (2019), "Measuring consumer inflation in a digital economy", OECD Statistics Working Papers, No. 2019/01, OECD Publishing, Paris, https://doi.org/10.1787/1d002364-en.

Skatova, Anya, Rebecca L. McDonald, Sinong Ma, and Carsten Maple. (2019), "Unpacking Privacy: Willingness to Pay to Protect Personal Data," PsyArXiv. 12 May. doi:10.31234/osf.io/ahwe4.

Statistics Canada (2019), "The Value of Data in Canada: Experimental Estimates." (July). Accessed on 10 September 2021 at: https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm

UNECE (2018), "Recommendations for Promoting, Measuring and Communicating the Value of Official Statistics"

United Nations (2008). "System of National Accounts 2008"

UK Financial Reporting Council (2019), "Business Reporting of Intangibles: Realistic proposals". Accessed on 15 September 2021 at https://www.frc.org.uk/consultation-list/2019/discussion-paper-business-reporting-of-intangibl.

Varian, H. (2018), "Artificial Intelligence, Economics, and Industrial Organization." NBER working paper #24839

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

Yamaguchi, Shinichi & Oshima, Hidetaka & Saso, Hiroaki & Aoki, Shihoko (2020), How do people value data utilization? An empirical analysis using CVM in Japan, Technology in Society, 62. 101285. 10.1016/j.techsoc.2020.101285.

Zéghal D. and A. Maaloul (2011), "The accounting treatment of intangibles – A critical review of the literature." *Accounting Forum*, 35 262– 274.

Annex A. Detailed Description of Data Sources and Estimation Method for Data Stores and Data Intelligence

87. In this Annex, we describe the cost-based approach and the data sources we have used to estimate market sector investment in data stores and data intelligence (the components of data not currently included in official national accounts) for Germany, France, Italy, Spain, Sweden, and the UK in 2012-2018.

In broad terms, cost-based estimates of output can be derived as follows:

Estimated value of data stores /data intelligence output at basic prices

equals

Labour costs of relevant personnel (compensation of employees)

plus

Intermediate costs used in these activities

plus

Cost of capital services used in these activities (gross operating surplus)

plus

Net taxes on production related to these activities

The standard way to implement the above calculation is to first estimate the labor cost component as:

Labour costs of relevant personnel

equals

Total number of employees working on producing data stores/data intelligence

times

Average remuneration

times

Proportion of time spent on these activities

Then, gross output at basic prices is obtained as:

Gross output at basic prices

equals

Labour costs of relevant personnel

times

Blow-up factor

88. The calculation requires i) a detailed list of occupations; ii) occupation-specific (and industry-specific, if relevant) assumptions on the share of time spent in producing data-related assets; iii) data on the number of employees for the relevant occupations and their compensations; iv) blow-up factors to account for other cost components (intermediate consumption and gross operating surplus) to derive an output measure consistent with national accounts definitions.

89. Occupations identified from the ISCO-08 as engaged in capital formation are presented in Table A1 along with time-use assumptions.

90. The selection of relevant occupations is constrained by the level of detail of the available data sources. For this paper, we use micro-data of the EU Structure of Earnings Survey (SES) for 2014 and the EU Labor Force Survey (LFS) for 2012-2018. The SES provides information on the number of employees by occupation (at the three-digit level of the 2008 International Standard Classification of Occupations, ISCO) and economic activity and their annual earning, while the LFS only provides data on the number of persons employed with no information on their wages. In the LFS, occupations are available at the three-digit level of ISCO for all the six countries, while SES data for Germany, Spain, and Sweden are available at two-digit. We have disaggregated two-digit ISCO into three digits for these three countries based on the share of each relevant three-digit occupation from the LFS.

91. As some occupational groups relevant for data-related asset production are only identifiable at the four-digit level, we have tweaked our assumptions on the time-use factors accordingly to consider that the occupational groups include workers not engaged in data assets production.

92. The calculation for 2014 (the year for which we have the SES) is a follows:

1. Calculate total employment for each relevant occupational group involved in producing data stores/data intelligence (identified at three-digit ISCO)
2. Apply occupation-specific time-use assumptions to each occupation's employment
3. Calculate total wages for each (time-use adjusted) relevant occupation.
4. Calculate the total share of all occupations involved in producing data stores/data intelligence in total wages from the SES
5. Calculate labor cost component consistent with national accounts by applying the share calculated at step 4 to national accounts' compensation of employees.

93. Gross output is then obtained applying country-specific blow-factors to the labor cost component derived at step 5. We have derived blow-up factors based on the cost structure of the industry J62_63 (Computer programming, consultancy industry, and information service activities) as the ratio of gross output to compensation of employees. We have removed the component of intermediate consumption that is likely to be due to subcontracting activity of the industry J62_63, which we deem not relevant to the cost structure of data production. We assume that subcontracting costs equal 50% of computer programming, consultancy and related services, and information services (CPA_62_63) bought by industry J62_63. For each country, we have used the 2010-2017 average.[21]

94. The value of output in 2014 is then extrapolated to 2018 and retropolated to 2012 based on information from the LFS. For each country, we have calculated the share of (time-use adjusted) relevant occupations in total employment for 2012-2018 from LFS. We have then used the employment share as an indicator to extrapolate/retropolate the 2014 wage shares obtained from the SES.

95. We have made the calculations by industry, at the level of Nace sections, and then aggregated the result to the the market sector, defined as all industries excluding Nace sections O (public

[21] 2017 is the last year for which the USE tables are available for all the six countries. The blow-up factors are 2.3 for Germany, 2.2 for France, 2.1 for UK, 3.4 for Italy, 2.4 for Spain and 2.8 for Sweden.

administration and defence; compulsory social security), P (education), Q (human health and social work activities), T (activities of households as employers; undifferentiated goods - and services - producing activities of households for own use), and the imputed rents of owner-occupied dwellings component section L (real estate activities).

Table A.1. Relevant occupations in measurement of investment in data stores and data intelligence and time-use assumptions

ISCO-08 sub-major group	ISCO-08 minor group	Occupation description	Data Stores		Data Intelligence	
			Time-use (%)	Statcan/Goodridge et al.	Time-use (%)	Statcan/Goodridge et al.
21 - Science and engineering professionals	211	Physical and earth science professionals	0.1	No	0.25	no
	212	Mathematicians, actuaries and statisticians	0.1	yes	0.25	yes
	213	Life science professionals	0.1	no	0.25	no
	214	Engineering professionals (excluding electrotechnology)	0.1	no	0.25	no
	215	Electrotechnology engineers	0.1	no	0.25	no
	216	Architects, planners, surveyors and designers	0.1	yes	0.1	yes
24 - Business and administration professionals	241	Finance professionals	0.1	yes	0.25	yes
	242	Administration professionals	0	-	0.1	no

	243	Sales, marketing and public relations professionals	0.1	yes	0.1	no
25 - Information and communications technology professionals	251	Software and applications developers and analysts	0.1	yes	0	-
	252	Database and network professionals	0.1	yes	0	-
33 - Business and administration associate professionals	331	Financial and mathematical associate professionals	0.1	yes	0.25	yes
35 - Information and communications technicians	351	Information and communications technology operations and user support technicians	0.1	yes	0	-
41 - General and keyboard clerks	413	Keyboard Operators	0.05	yes	0	-
42 - Customer services clerks	422	Client information workers	0.05	yes	0	-
43 - Numerical and material recording clerks	431	Numerical clerks	0.05	yes	0	-
	432	Material-recording and transport clerks	0.05	yes	0	-

Note: 413 includes data entry clerks (4132); 422 includes survey and market research interviewers (4227); 431 includes statistical, finance and insurance Clerks (4312).