United Nations ECE/CES/GE.20/2019/14

Economic and Social Council

Distr.: General
29 March 2019

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on National Accounts

Eighteenth session
Geneva, 10-12 April 2019
Item 7 of the provisional agenda
Current research related to digitalization

Measuring the Digital Economy in Macroeconomic Statistics: The Role of Data

Prepared by the International Monetary Fund[1]

Summary

The strategic focus of businesses in the modern knowledge-based economy has been to generate and control traditional intellectual property (IP) assets, such as patents and copyrights. Nowadays, the strategic focus of businesses is increasingly to generate and control data. As the economy becomes more knowledge-based and data-driven it is increasingly difficult to measure. National statistical compilers often rely on business accounting records or surveys of businesses to derive estimates, yet business accounting methods have not evolved to measure the value of data. This creates difficulties for national statistical compilers.

This paper explores what data is, what's its value, what's its role in the modern economy and whether it is an asset in the national accounts– the perimeter of what could be capitalized while considering what is already included in R&D, software and databases– and potential estimation methods.

[1] Prepared by Jennifer Ribarsky, International Monetary Fund

I. Introduction

1. Data flows through the modern economy and it is often said it is the “oil”, the fuel of the future. According to Cisco, the annual global Internet Protocol (IP)[2] traffic was 1.5 zettabytes[3] in 2017 and is projected to increase threefold over the next 5 years reaching 4.8 zettabytes by 2022. Yet, not all of these packets of data flows are the same, 82% of all IP traffic will be video by 2022, up from 75 percent in 2017.[4] This is perhaps an important point to distinguish: *data flows can be a means of delivery of content* (e.g. movies, books, music). In fact, according to Sandvine, Netflix is 15% of the total downstream[5] volume of traffic across the *entire Internet*.[6] While this might be of interest to understand the migration of content from physical and analog forms (e.g. books, broadcast television) to digitized forms, *it is of most interest to understand how businesses use data in a productive sense*. For example, how does Netflix use data from their customer’s Netflix browsing history, viewing history (e.g., whether you binge watch, abandon a show or complete an entire season, etc.) and ratings to select movies and television shows, create content, and make decisions? In this respect, Netflix is about 3% of the total upstream volume of traffic. This is the traffic volume uploaded to the Internet (e.g. browsing the Netflix library). Businesses use this upstream traffic and other means of acquiring data *to generate information and knowledge*.

II. What is data?

2. It may be useful to first discuss what is meant by data. According to dictionary.com data is defined as

(i) facts and statistics collected together for reference or analysis.

(ii) the quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

3. While data can be in analog form (e.g., stored in paper books), businesses leveraging data in electronic form (e.g., stored in bits and bytes) to improve their products and processes or to monetize the data itself, either directly (e.g. data brokers) or indirectly (e.g. targeted advertising), is often what is thought of when talking about the “data-driven” economy.

4. While it is often said that data is the raw material that drives the modern economy, its value is hard to obtain. Some say that data is very valuable, while others say that it is not valuable at all because it is the *insights derived from data*, *how data is used* is what creates the value.

[2] According to Wikipedia, Internet Protocol (IP) is the principal communications protocol in the Internet protocol suite for relaying datagrams across network boundaries. Its routing function enables internetworking, and essentially establishes the Internet. IP has the task of delivering packets from the source host to the destination host solely based on the IP addresses in the packet headers. For this purpose, IP defines packet structures that encapsulate the data to be delivered. It also defines addressing methods that are used to label the datagram with source and destination information.

[3] One zettabyte is 1 followed by 21 zeroes.

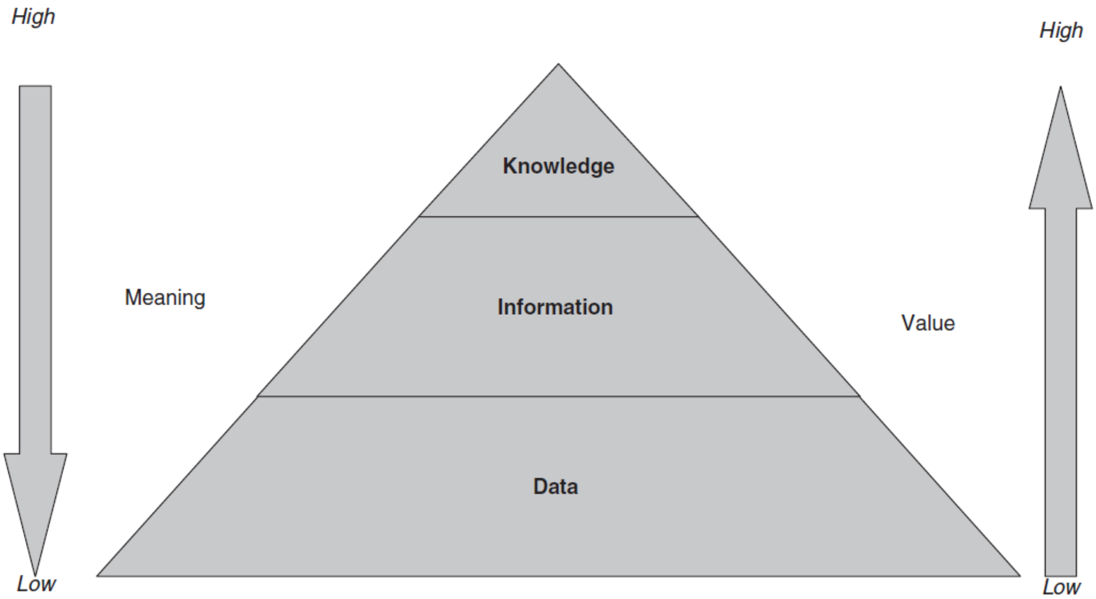
[4] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html#_Toc529314187

[5] This is the traffic volume downloaded from the internet. Examples would be a video stream, a file download, or an app download from iTunes.

[6] The report however does not include significant data from China or India. https://www.sandvine.com/2018-internet-phenomena-report

5. So, how can such divergent views exist? First, it depends on the perimeter of what you are talking about. It may, therefore, be useful to look at the information and knowledge management literature to help us dissect the nuances between data, information, and knowledge. As Rowley (2017) explains, data is not knowledge. Data are the products of facts or observations and are of no use until they are in a useable (i.e. relevant) form. In other words, data itself has no or little intrinsic value as shown in figure 1. Organizing and processing data lends the data relevance for a specific purpose or context, and thereby makes it meaningful, valuable, and useful. Information is inferred from data. Information is then used to create know-how, the knowledge that makes possible the transformation of information into instructions. Knowledge can be further differentiated into explicit knowledge (recorded in information systems) and tacit knowledge (cannot be recorded since it is part of the human mind).

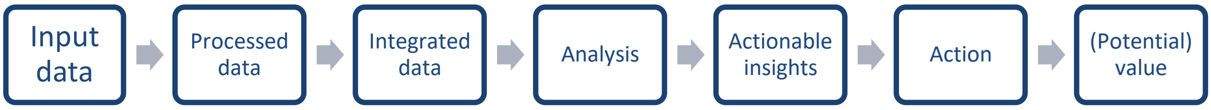
Figure 1
Data, information, and knowledge



Source: Rowley (2017). Data, information, and knowledge according to Chaffey and Wood

6. Before input data becomes useful it must undergo several processing steps, such as collection, recording, organizing, structuring, storage, combination and integration (potentially with other data sources). Mawer (2015), figure 2, shows the progression of products that are created as input data is processed, integrated and then analyzed with *context*, the “information” layer in figure 1, to produce actionable insights. The actionable insights transform this information into know-how or knowledge (figure 1, layer 3), which can lead to action and, potentially, value.

Figure 2
Data transformation chain



Source: Mawer (2015), https://www.svds.com/valuing-data-is-hard/

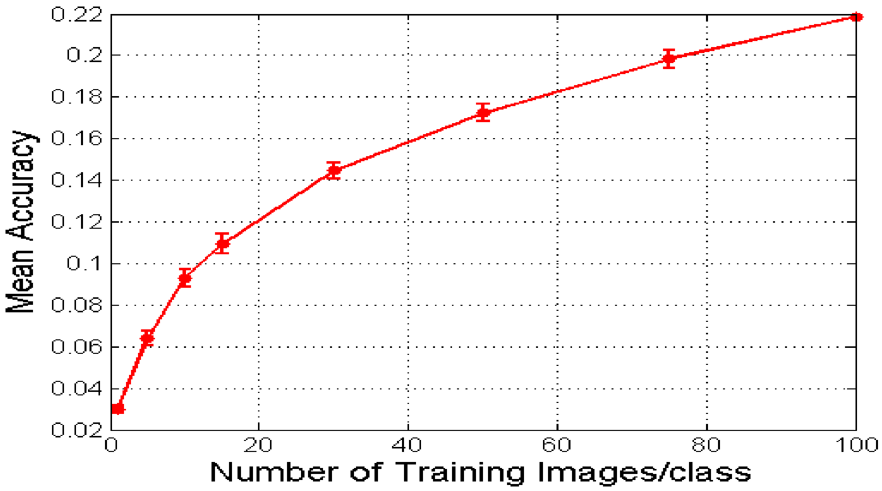
III. What is the value of data?

7. The potential value of data depends on where in the chain it lies, and the value increases as it moves through the transformation chain. The input dataset has much less value than the value of the information and know-how that is provided once the analysis is done.

What's the value of the input data(set)?

8. While each piece of data may have no or little intrinsic value, the value can increase when it is combined with other data to become a dataset. This value can be greater than the sum of its parts. In other words, the more data you have, the more likely you will be able to identify patterns and trends that improve your information and knowledge. However, that is not to say that data always exhibits increasing returns to scale. As Varian (2018) describes, data can exhibit decreasing returns to scale as shown by using the Stanford Dogs Dataset. As seen in figure 3, the accuracy of the machine learning improves as the number of training images increases, but at a decreasing rate. Whether data exhibits increasing or decreasing returns to scale may be due to whether the data is simply adding another record (one more dog image and hence not much more in terms of information can be gleaned) versus whether the data is combined with complementary data which may expand the possibilities for data use.

Figure 3
Mean accuracy of Training Images



Source: http://vision.stanford.edu/aditya86/ImageNetDogs/

9. Input data can be obtained by businesses in various ways.

- **First-party 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**- essentially someone else's first-party data. Second-party data is not usually bought and sold. Businesses work out arrangements with trusted partners who are willing to share their customer data with them (and vice versa). For instance, a high-end watch company might partner with a yacht blog to find new customers, based on demographic overlap.
- **Third-party data**- any data collected by an entity that does not have a direct relationship with the user the data is being collected on.
- **Public data-** open or freely available without payment, e.g. data produced by the government and made freely available for anyone to use.

10. The first two sources are usually not associated with a market transaction, in other words they are not bought and sold. However, there are sometimes exceptions, as recent reports indicate that since 2016 Facebook has been paying certain users up to USD 20 per month plus referral fees to sell their data by installing the "Facebook Research" app.[7] This app essentially lets Facebook acquire *all data* on a user's phone and web activity, not just the activity done on Facebook's products (Facebook, Instagram, WhatsApp).

11. Exchange of third-party data–often obtained from data brokers or data aggregators such as Axciom– usually involves a market transaction or a partnership to use the data in exchange for profit-sharing. Third party input data are often obtained by businesses through licensing, subscription, or contractual arrangements. Axciom's financial reports note that many of the licensing arrangements are in the form of recurring monthly billings, as well as transactional revenue based on volume or one-time usage.

12. These data brokers sell consumer profiles in large chunks, e.g., 10,000 in a batch. One source reported that the price for a list of a thousand people with health conditions like anorexia, substance abuse, or depression was USD 79 or USD 0.079 per user profile.[8] Data on health conditions are worth the most, as shown in the Financial Times calculator[9], so using the value of health data overestimates the typical value of data to advertisers. According to an *Atlantic Monthly*[10] article, user profile data go for USD 0.005 per profile based on advertising-industry sources.

13. What accounts for such diverging estimates of what a user profile is worth?

- **Quality of data**. The price differential could be accounted for by the quality of the data. Data quality can be measured along a number of dimensions such as accuracy, completeness, breadth, latency, and granularity.
- **Access to data**. Third party data is usually widely accessible, so a business is not necessarily gaining unique audience intelligence that is not also available to their competitors. In the Facebook Research app example cited above, the

[7] https://techcrunch.com/2019/01/29/facebook-project-atlas/

[8] https://www.webfx.com/blog/general/what-are-data-brokers-and-what-is-your-data-worth-infographic/

[9] https://ig.ft.com/how-much-is-your-personal-data-worth/?ft_site=falcon#axzz2z2agBB6R

[10] https://www.theatlantic.com/technology/archive/2012/03/how-much-is-your-data-worth-mmm-somewhere-between-half-a-cent-and-1-200/254730/

access to the user's data is most likely restricted to the company itself and is not widely accessible.

- **Use of data**. Input data has a number of possible uses that depend on the user of the data and the context, in other words what information it may provide and how the business will use that information.

IV. Data as a factor of production

14. Data have always had a central role in business decision making. Businesses strive to gather data on customers, to improve products and processes to enhance productivity, improve performance, and increase profitability. As storage and acquisition costs decreased and processing capacity (software, IT hardware) increased this led to an explosion in data accumulation. The simple fact that the data is in electronic form allows it to be analyzed for insights and decision-making at an unprecedented scope and scale. In some sense data itself has been transformed: it has become *digital data*. This digital data has allowed for new information/knowledge creation that could not have been done if the data were not in digital form.

15. The modern data-driven economy has moved beyond databases of "structured" data (e.g. lists of names and well-defined personal characteristics) into "unstructured" data. Yet, even "unstructured" data have standardized structures and meaning can be extracted by using data and text mining and data analytics. For example, aggregated GPS data could be used to help a retailer choose the location of its next store, while a city government could use GPS data (even the same input data since data are non-rivalrous) to better plan its roads. Both these uses would require different analysis and possible integration with different data sources. Furthermore, vast amounts of data are collected by digital platforms and Internet connected devices (e.g., Internet of Things (IoT)). Consequently, digital data is becoming another factor of production and Bean (2016) states that it is analogous to physical and intangible capital.

V. Is digital data an asset?

A. Digital data

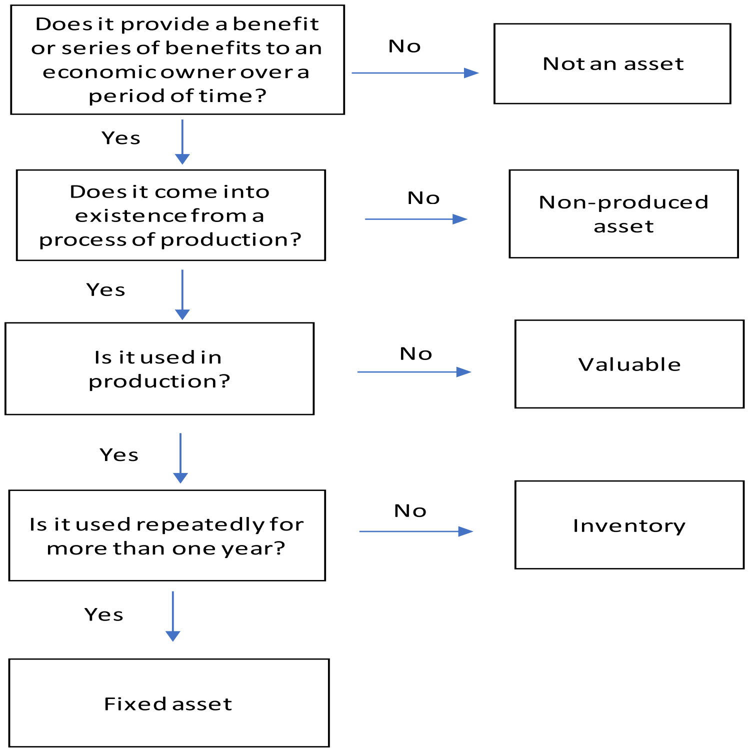
16. Data is a source of value creation for a business, but does it satisfy the System of National Accounts (SNA) criteria of an asset? According to the 2008 SNA, an asset is a "store of value representing a benefit or series of benefits accruing to the economic owner by holding or using the entity over a period of time. It is a means of carrying forward value from one accounting period to another".[11] Non-financial assets are grouped into two broad categories: produced (coming into existence as outputs from production processes that fall within the production boundary, e.g. equipment, software, R&D) and non-produced (coming into existence in ways other than through the process of production, e.g. land). There are three main types of produced assets: fixed assets, inventories and valuables. Both fixed assets and inventories are assets that are held only by producers for purposes of production. Valuables may be held by any institutional unit and are primarily held as a store of value.[12] To qualify as a **fixed** asset, a good or service must be used repeatedly or continuously in the

[11] 2008 SNA paragraph 10.8

[12] SNA paragraph 10.10.

production process **for more than one year**.[13] Inventories are produced assets that consist of goods and services, which came into existence in the current period or in an earlier period, and that are held for sale, use in production or other use at a later date.[14] Figure 4 shows a decision tree that can be used to determine if data is an asset.

Figure 4
Decision tree to determine if asset and what type


Does it provide a benefit or series of benefits to an economic owner over a period of time?
No
Not an asset
Yes
Does it come into existence from a process of production?
No
Non-produced asset
Yes
Is it used in production?
No
Valuable
Yes
Is it used repeatedly for more than one year?
No
Inventory
Yes
Fixed asset

17. Data provides economic benefits to its users and it can be used repeatedly. Data is non-rivalrous because the same data source can be used for multiple analysis (multiple uses) and by multiple users. The trickier part in answering the first question "Does it provide a benefit or series of benefits to an economic owner over a **period of time**" is how a period of time is defined.

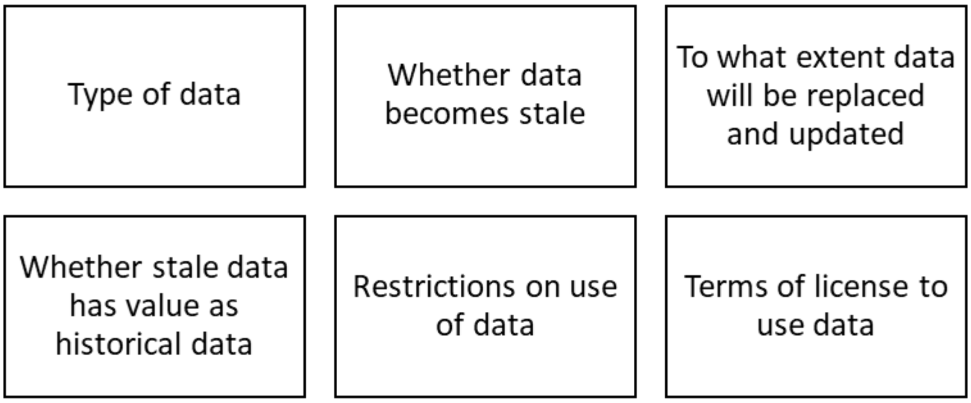
18. In principle, data in digital form can be stored forever but many kinds of data are not likely to be used repeatedly for more than one year. For example, data collected on browsing activities that are then used for targeted advertising of specific products do not have a long life for that particular use. When I shop for shoes online, the pair of shoes that I clicked on only "follow" me through targeted advertising for about a month. Therefore, much of this data could be considered as intermediate inputs to targeted advertising. As such, the value is already appropriately included in output.

[13] SNA paragraph 10.11.
[14] SNA paragraph 10.12

19. While data could be viewed as an intermediate input another alternative is to consider it as a new kind of inventory. The one-year criterion does not apply to inventories. The SNA stipulates that "inventories consist of stocks of outputs that are still held by units that produced them prior to being further processed, sold, delivered to other units or *used in other ways* and stocks of products acquired from other units that are intended to be used for intermediate consumption or for resale without further processing". [15] If digital data is recorded as a type of inventory then gross capital formation would be recorded, thus impacting GDP.

20. On the other hand, some data may have a useful life of more than one year. Axciom, the data broker, capitalizes costs related to the acquisition or licensing of data for their proprietary databases which are used in providing data products and services. **These costs are amortized over the useful life of the data, which is from two to seven years**. To estimate the useful life of any acquired data, Axciom considers several factors, shown in figure 5.[16] If digital data is recorded as a type of fixed asset then gross *fixed* capital formation would be recorded, thus impacting GDP.

Figure 5
Axciom's criteria to determine if data acquisition costs should be capitalized



Type of data

Whether data becomes stale

To what extent data will be replaced and updated

Whether stale data has value as historical data

Restrictions on use of data

Terms of license to use data

Source: Axciom's 2017 Financial Report.

21. National statistical compilers also need to determine if data comes into existence because of a production process. Ahmad and van de Ven (2018) propose to treat data as a non-produced asset. The recommendation is to record transactions related to data only when a monetary transaction occurs and to include them as a sub-item of goodwill. It should be noted that no gross capital formation is recorded and therefore, it has no impact on GDP. This may be the most prudent approach, as it appears there could be some information in business accounts that could help identify such transactions, although this is not always the case (see Annex 1).

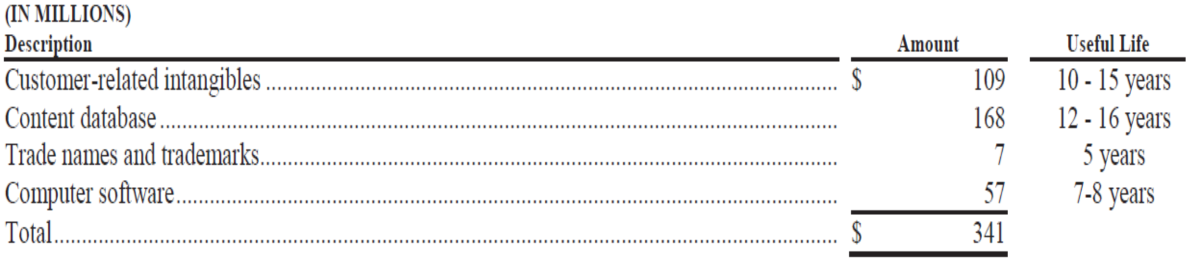
[15] SNA paragraph 10.12.

[16] See Axciom's financial report at https://www.sec.gov/Archives/edgar/data/733269/000073326917000039/acxm-20170331x10k.htm

22. Nielsen[17], a company founded in 1923, is a leader in market research and ratings. Nielsen states in their financial report that their business is based on “an extensive foundation of proprietary data assets designed to yield essential insights for our clients to successfully measure, analyze and grow their businesses and manage their performance.” In 2017, Nielsen’s total revenue was USD 6.6 billion. However, their balance sheet only includes a small amount of data assets (USD 168 million, figure 6) that were recorded when Nielsen acquired Gracenote in 2017 for USD 585 million. Most of the purchase price was allocated to goodwill (USD 316 million) and amortizable intangible assets (USD 341 million). One of the reasons that Nielsen acquired Gracenote was to expand their footprint with major clients by acquiring Gracenote’s global content database which spans across platforms including multichannel video programming distributors (MVPD’s), smart television, streaming music services, connected devices, media players and in-car infotainment systems.

Figure 6
Nielsen’s intangible acquisitions from Gracenote

Millions of U.S. Dollars



(IN MILLIONS)

Description		Amount	Useful Life
Customer-related intangibles	$	109	10 - 15 years
Content database		168	12 - 16 years
Trade names and trademarks		7	5 years
Computer software		57	7-8 years
Total	$	341	

Source: Nielsen’s 2017 Financial Report.

B. Digital data-based information

23. Instead of determining if data itself is an asset this paper considers whether the information and knowledge derived from the data is. This is consistent with Varian’s perspective that it is the organizing and analysis of data that creates value and not data itself. It is the analytics on the data and the creation of algorithms that create value. It can be argued that the information and knowledge derived from data come into existence through a process of production: the input data is put into a relevant (electronic) form that can be processed using resources (labor, e.g. data scientist, and capital, e.g. computers, software and databases). One may argue that the facts (e.g. a person’s gender or age) do not come about because of a production process, but the act of gathering the data and digitizing it requires resources; and therefore, the digital data-based information is produced.

24. User profiles that are developed by analyzing data to determine patterns of behavior may be an example of data-based information that qualifies as a fixed asset. Companies such as Axciom repeatedly use this type of information to generate licensing revenue. The information and know-how created from data is the result of a production process and in deciding whether it qualifies as an asset, the entire data transformation chain should be considered.

[17] See Nielsen’s financial report at https://s1.q4cdn.com/199638165/files/doc_financials/Annual/2018/04/2017-Annual-Report.pdf

VI. How to value digital data-based information?

25. As the information and knowledge derived from input data are typically not evidenced by a market transaction various estimation approaches have been proposed. These approaches, summarized by Li et al (2018) are market-based, cost-based, and income-based.

- **Market-based**: value is determined based on the market price of comparable products on the market.
- **Cost-based**: value is determined by how much it costs to produce the information/know-how derived from data.
- **Income-based**: value is determined by estimating the future cash flows that can be derived from the data.

A. Market-based approach

26. The 2008 SNA states that transactions should be valued at market prices, defined as amounts of money that willing buyers pay to acquire something from willing sellers,[18] and that if market prices for transactions are not observable, valuation according to market-price-equivalents can provide an approximation[19]. So, on a *conceptual basis* the market-based approach is the preferred concept of the SNA. The problem is that in most cases, except for commercial third-party databases, a comparable product sold on the market does not exist. One, therefore, may be able to estimate the value of, for example, unprocessed consumer data using the market price of user profiles sold by data brokers, but this is not an exact equivalent as the third-party user profile data has undergone processing (e.g. organizing). Organizing data, cleaning it, and making it fit for use, may require significant resources. In addition, as discussed in detail above, this would only provide an estimate of the input data and not an estimate for the entire transformation chain needed to derive digital data-based information.

27. Ahmad, Ribarsky, and Reinsdorf (2017) derive a value for user input data (e.g., number of active users * value of user profile) for five major digital services (Facebook, Twitter, Instagram, LinkedIn, and Gmail), equivalent to around 0.02% of global GDP. The maximum user profile price was USD 5.1512[20], based on a *Financial Times* calculator that used industry pricing data from a range of sources in the United States. This figure – which assumes that the subject has a rare health condition – is much more than the USD 0.005 per profile quoted in the *Atlantic Monthly* article but much less than the USD 20 per user that Facebook paid to acquire data through their Facebook Research App. This illustrates that the price assumption used may vary significantly depending on the source of the price and, more importantly, what the intended use of the data will be.

28. Could this approach be used to value the vast amounts of data created by the IoT (e.g. sensor data)? Most sensor data are valuable as an input into the production process (and most often only for short-term use) of the data collector but not very useful to others. However, there may be a market developing for this type of data. A company called Terbine has created a system to enable IoT data trading alliances. Terbine makes available data on the "physical world" from solar farms' power output to tracking drones and airplanes in the sky. Terbine cleans, indexes, and grades the sensor data so that it can be priced dynamically.

[18] 2008 SNA paragraph 3.119.

[19] 2008 SNA paragraph 3.123.

[20] The estimate was broken down into a valuation of digital identities (USD 0.5296) and a valuation of digital footprints (USD 4.6217).

29. Terbine has three offerings: public data subscriptions, branded data exchanges, and a market place. For the public data subscriptions, the company has a team of "Data Searchers" that seek, locate, characterize and index machine-generated data feeds from public agencies around the world.

30. Terbine's branded data exchange allows organizations (e.g., supplier-buyer groups, trade associations, government agencies and research institutions) to exchange data. An example is the Intelligent Transportation Society of America (ITSA) that has over 300 member-organizations ranging from major automakers to federal, state and municipal highway authorities. The ITSA Data Exchange offers an environment for participants to share, and in some cases *monetize*, their IoT data feeds.

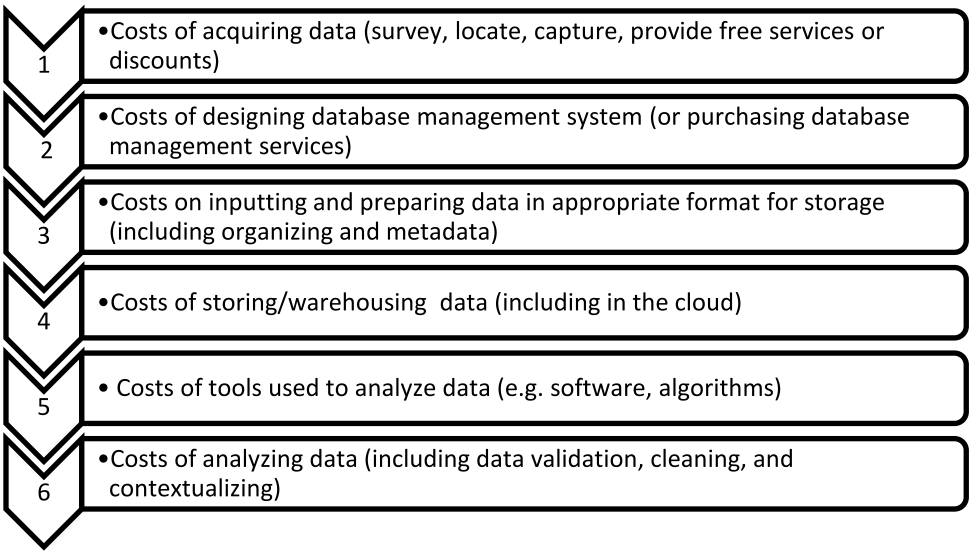
31. While IoT data markets may develop for certain types of sensor data it will most likely be hard to determine a true market price equivalent for data that are not exchanged by a market transaction. In addition, as with consumer/user profile data most of the value of the IoT data will be the information derived from the sensor data.

B. Cost-based approach

32. If there is no appropriate market price, or market-price-equivalent, from which the value can be derived the 2008 SNA notes that valuation by costs or valuation using the income-based approach can be used. In general, the SNA gives preference to valuation by costs; indeed, own-account gross fixed capital formation (GFCF) in software and databases and research and development are measured using the "sum-of-costs" approach. In addition, for market producers, the SNA recommends including a mark-up that reflects the operating surplus or mixed income attributable to the producer.[21]

33. Figure 7 considers potential direct costs for creating digital data-based information:[22]

Figure 7
Direct costs for creating digital data-based information



1 •Costs of acquiring data (survey, locate, capture, provide free services or discounts)

2 •Costs of designing database management system (or purchasing database management services)

3 •Costs on inputting and preparing data in appropriate format for storage (including organizing and metadata)

4 •Costs of storing/warehousing data (including in the cloud)

5 • Costs of tools used to analyze data (e.g. software, algorithms)

6 •Costs of analyzing data (including data validation, cleaning, and contextualizing)

[21] 2008 SNA paragraph 3.135.

[22] Full accounting of costs would also include a portion of administrative and overhead costs.

34. One could consider costs 1-4 to be associated with database creation and costs 5 and 6 to be associated with digital data-based information/knowledge creation.

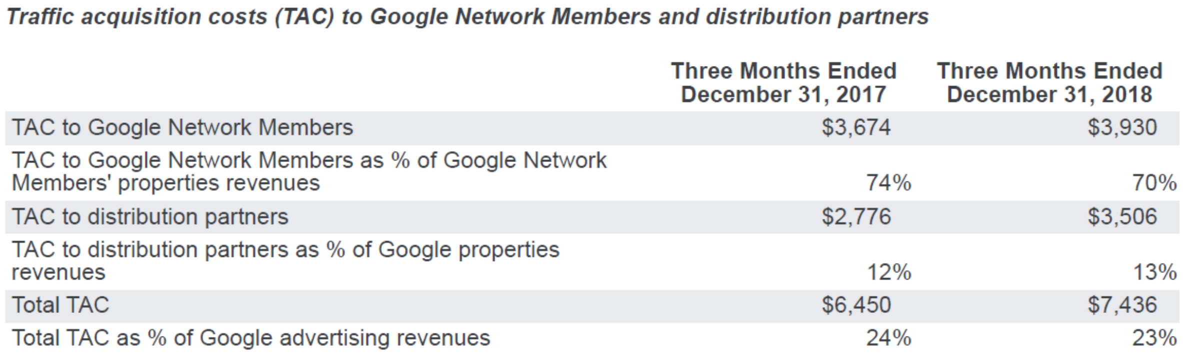
35. The first step in creating digital data-based information is acquiring the input data. As discussed above some data are acquired through purchases (e.g. the rights to use the data) from third-party data providers or occasionally there are explicit payments to households (e.g. the Facebook Research App example discussed above). Firms can also hire market research firms to conduct surveys, focus groups, and interviews to collect the required data. Thus, national statistical offices' business surveys could be expanded to explicitly ask for the costs of acquiring data.

36. However, as discussed above, much data is acquired by businesses without explicit payments (e.g., data as byproducts of production, data acquisition in exchange for free services or through pricing discounts). There has been discussion on whether an imputation should be made for data that is acquired through "barter" in exchange for free services. One could consider imputing a cost for data acquired through barter that could be based on a user profile price as discussed under the market-based approach above. An alternative approach would be to review business accounts to see if some of the expenditures that a business makes to obtain users could be considered as data acquisition costs. Although expenditures to attract or acquire users have multiple objectives of which collection of their data is just one. The expenditures also enable the platform operator to show users advertisements, sell them products, and attract other users to the platform. In effect, the expenditures can be viewed as own-account capital formation to create a bundle of intangible assets.

37. An example of such expenses are traffic acquisition costs. Traffic acquisition costs are payments made by some digital platforms (e.g., Google, Yahoo, Baidu, Facebook) to others for directing consumer and business traffic to their websites and are a critical cost of revenue– 24% of Google advertising revenues as shown in figure 8. Goldman Sachs estimates that Google will pay Apple USD 12 billion in 2019 to make it the default search engine.[23]

Figure 8
Google traffic acquisition costs[24]

Millions of U.S. Dollars



Traffic acquisition costs (TAC) to Google Network Members and distribution partners

	Three Months Ended December 31, 2017	Three Months Ended December 31, 2018
TAC to Google Network Members	$3,674	$3,930
TAC to Google Network Members as % of Google Network Members' properties revenues	74%	70%
TAC to distribution partners	$2,776	$3,506
TAC to distribution partners as % of Google properties revenues	12%	13%
Total TAC	$6,450	$7,436
Total TAC as % of Google advertising revenues	24%	23%

Source: Alphabet's Fourth Quarter and Fiscal Year 2018 Results

[23] https://www.mobilemarketer.com/news/goldman-apple-will-charge-google-12b-to-be-default-search-engine-in-2019/538469/

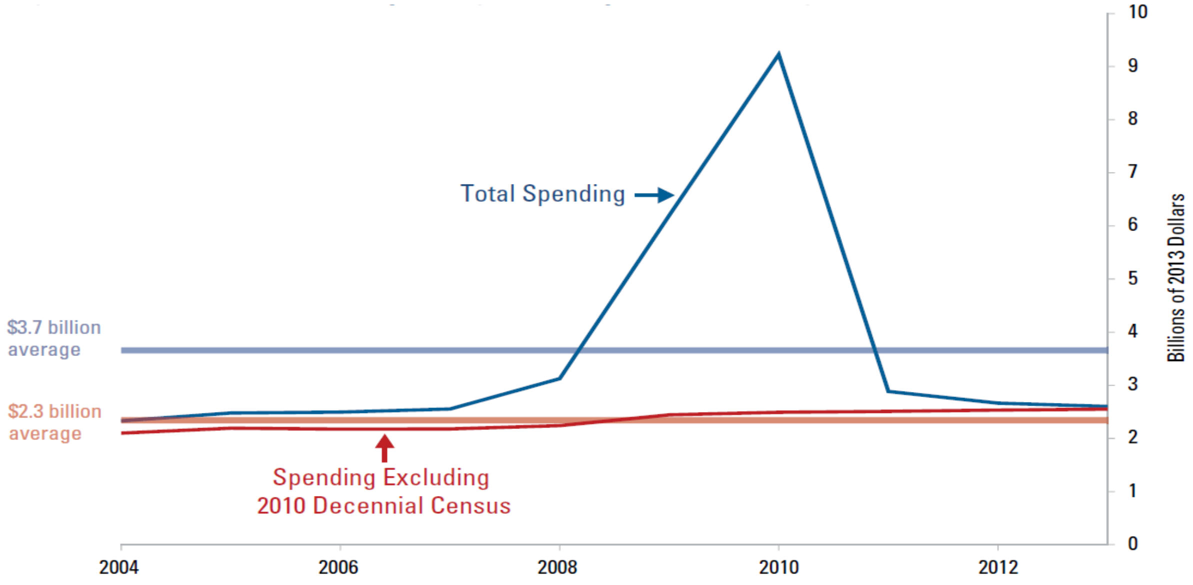
[24] https://abc.xyz/investor/static/pdf/2018Q4_alphabet_earnings_release.pdf?cache=adc3b38

38. Yet, this would not fully account for all unpriced data acquisition from consumers. For example, a smart TV's price is subsidized because the manufacturer uses the TV for data collection, advertising and content delivery.[25]

39. Considerations should also be given to the vast amount of data collection that is done by the government. Government data is a key input to a wide variety of commercial goods and services in the economy (as well as a key input to government policy making). The U.S. Department of Commerce (2014)[26] reported that the United States spent USD 3.7 billion annually, adjusted for inflation, on data collection and dissemination by the Principal Statistical Agencies[27]. The Decennial Census is associated with a surge in expense. Excluding the Decennial Census, the average is about $2.3 billion (Figure 9). If one were to approach this holistically, one could not ignore the costs of government acquisition of data.

Figure 9
U.S. Federal Government Spending on the Principal Statistical Agencies

Billions of U.S. 2013 Dollars



Sources: Budget information compiled from *Analytical Perspectives, President's Budget*; *Statistical Programs of the U.S. Government Supplement to President's Budget*; actual agency budgets; *Principles and Practices for a Federal Statistical Agency*
Note: Budget amounts converted to real 2013 dollars using Government Consumption Expenditures deflator.

Source: U.S. Department of Commerce (2014)

[25] https://www.theverge.com/2019/1/7/18172397/airplay-2-homekit-vizio-tv-bill-baxter-interview-vergecast-ces-2019

[26] https://www.bea.gov/sites/default/files/2018-02/fostering-innovation-creating-jobs-driving-better-decisionsthe-value-of-government-data714.pdf

[27] These agencies' core missions are to collect, compile, process, analyze, and disseminate information for statistical purposes. They are: Bureau of Economic Analysis, Bureau of the Justice Statistics, Bureau of Labor Statistics, Bureau of Transportation Statistics, Bureau of the Census, Economic Research Service, Energy Information Administration, National Agricultural Statistics Service, National Center for Education Statistics, National Center for Health Statistics, National Center for Science and Engineering Statistics, Office of Research and Statistics, and Statistics of Income Division.

40. Once data is collected it must be put into the appropriate format for use. It must be organized, indexed, and contextualized through metadata. Businesses that gather data from many different sources may devote significant resources to this activity. Cost to include would be staff involved in this work such as data entry personnel, database architects, software architects and engineers, and subject matter experts needed to help contextualize the data.

41. Data can be stored in company-owned data centers or businesses can leverage the cloud. Cloud computing encompasses several types of services delivered over the Internet, including data processing, storage, database management, and software. Even large data-driven businesses make use of cloud computing services. For example, Netflix does not maintain any data centers and uses Amazon Web Service (AWS) for its computing and storage needs, from customer data to recommendation algorithms. To be clear, Netflix still designed the software (database architecture and algorithms) that utilize AWS data processing and storage.[28]

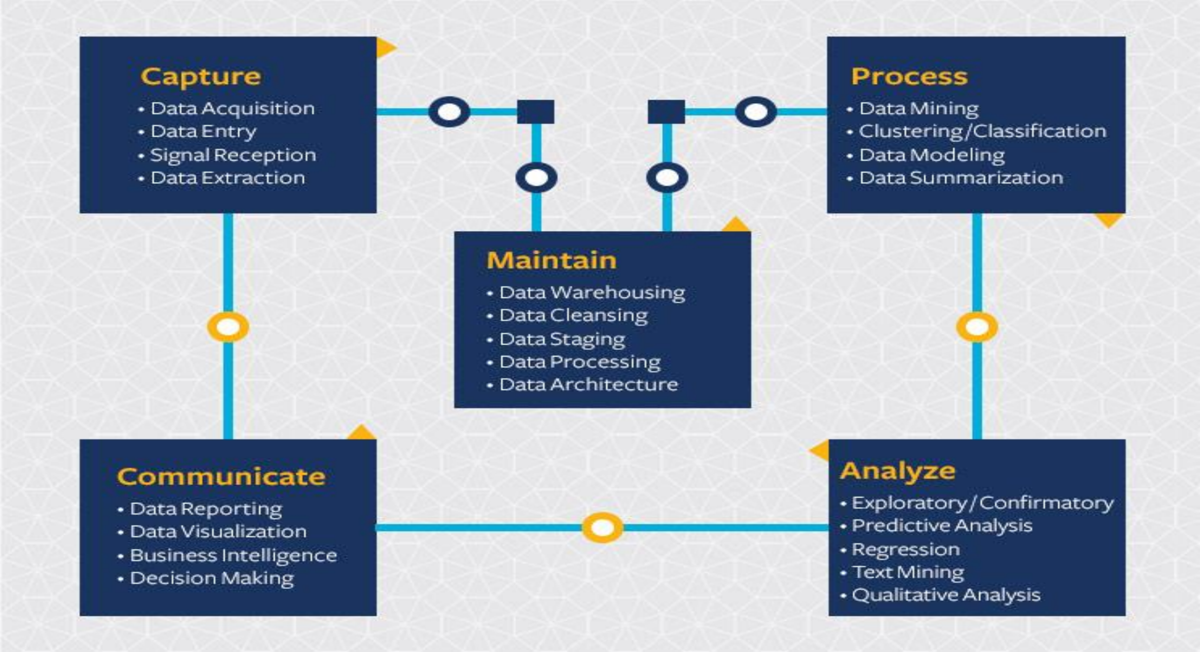
42. Once the data are in an appropriate format, data processing can begin by using software tools– developed in-house, custom, or packaged software, or even using open source software such as R and Python– that can mine and model the data making it ready for analysis. Data scientists and/or software engineers may even design new algorithms. The analysis of the data– nowadays often done by data scientists but in previous years we may have called these people statisticians[29]– may include data validation, cleaning, and contextualizing for a given use, possibly combining the data with other data to gain additional insight.

Figure 10 illustrates the tasks involved in the five stages of the data science life cycle which includes database creation.

[28] https://www.computerworlduk.com/cloud-computing/how-netflix-moved-cloud-become-global-internet-tv-network-3683479/

[29] In 2009, Google's Chief Economist Hal Varian said the following on data and statistics "I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?" https://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/

Figure 10
The Data Science Life Cycle



Source: https://datascience.berkeley.edu/about/what-is-data-science/

C. Income-based approach

43. Since the value of data is context-dependent, it may be necessary to value the digital data-based information using an income-based approach. While the income-based valuation approach is an acceptable method, the 2008 SNA advises caution in its use. The 2008 SNA recommends the income-based valuation approach "*If none of the methods mentioned above [e.g. market-price-equivalent, valuation at cost] can be applied, stocks, or flows arising from the use of assets, may be recorded at the discounted present value of expected future returns….However, because it may be difficult to determine the future earnings with the appropriate degree about the asset's life length and the discount factor applied, the other sources of valuation…should be exhausted before resorting to this method.*"[30] One of the primary difficulties with this approach is distinguishing the cash flows (net of associated costs) uniquely related to the asset from the cash flows related to the whole company.[31] However, this approach is recommended in valuing musical, literary, and photographic works– industries where there is an established system of royalty flows.[32]

44. This approach may be a viable option for data that can easily be tied to a particular use, such as targeted advertising. Many digital platforms fund their operations through

[30] 2008 SNA paragraph 3.137.

[31] https://www.cgma.org/content/dam/cgma/resources/tools/downloadabledocuments/valuing-intangible-assets.pdf

[32] See chapter V in OECD, 2010, Handbook on Deriving Capital Measures of Intellectual Property Products. http://www.oecd.org/sdd/na/44312350.pdf

advertising. Yet, attributing the entire cash flow from advertising revenue to the data alone would be an overstatement. Advertisers not only pay for behaviorally-targeted ads, but also for distribution, and general access to the users.

45. Johnson et al (2018) find that ads shown to users who opt-out from being behaviorally targeted fetch 52% less revenue on an ad exchange than do comparable ads for users who allow behavioral targeting. [33] This suggests that around half of online advertising revenue is generated because the advertiser has access to “eyeballs”.

46. The U.S. Internet Publishing and Broadcasting and Web Search Portals industry earns much of its revenue selling online advertising space, USD 105.2 billion in 2017 or 62% of total revenue (Table 1). If just over half of the revenue (and associated costs)[34] from online advertising space can be attributed to data-based information and using 3 years for the life-length[35] and a real discount rate of 8%, I derived an estimated value for the stock of data-based information for the U.S. Internet Publishing industry of USD 85.5 billion (Table 2, column 3, line 1). If one assumes that the entire revenue from online advertising space can be attributed to data-based information, then the estimated value of data would be USD 164.4 billion (Table 2, column 3, line 2). To put these amounts in perspective, in 2017 the current-cost net capital stock of private fixed assets for prepackaged and own-account software are USD 176.4 billion and 146.3 billion, respectively (Table 2, column 6). Table 2 also shows the calculation with an alternative discount rate and service life, showing the sensitivity to the chosen parameters.

Table 1
U.S. Internet Publishing and Broadcasting and Web Search Portals

Millions of US Dollars

Source of revenue	2017	2016	2015	2014	2013
Revenue	170,781	148,039	125,868	109,414	96,951
Publishing and broadcasting of content on the Internet	42,806	37,948	33,763	34,079	30,765
Licensing of rights to use intellectual property	4,317	4,125	3,590	4,133	3,782
Online advertising space	105,190	90,288	75,266	54,670	49,805
All other operating revenue	18,468	15,678	13,249	16,532	12,599

Source: Table 4. Estimated Sources of Revenue for Employer Firms. U.S. Census Bureau Services Annual Survey

[33] Johnson G., Shriver S., and Du S (2018) “Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry? ”. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3020503

[34] I assume costs are around 55% of revenue using an average of Facebook’s cost to revenue share from 2018-2014.

[35] The assumption of 3 years is made because much of the data used for targeted-advertising may only have a service life of a few months. Therefore, a shorter service life may be representative of the data-based information stock as a whole. Table 2, also shows an alternative with a longer service life.

Table 2
Value of data-based information for U.S. Internet Publishing and Broadcasting and Web Search Portals industry, 2017

Billions of US Dollars

	Discount factor: 8% Service life assumption		Discount factor: 5% Service life assumption		NIPA
Net stock, 2017	**7 years**	**3 years**	**7 years**	**3 years**	**stock**
Data-based information, portion of AD space	148.3	85.5	162.2	89.0	…
Data-based information, all AD space	285.2	164.4	311.8	171.1	…
Software, NIPA current-cost net stock of private fixed assets	…	…			644.4
Prepackaged software, NIPA current-cost net stock of private fixed assets	…	…			176.4
Custom software, NIPA current-cost net stock of private fixed assets	…	…			321.6
Own-account software, NIPA current-cost net stock of private fixed assets	…	…			146.3

Source: Author's calculations using U.S. Census Bureau data for data-based information asset for U.S. Internet Publishing and Broadcasting and Web Search Portals industry; NIPA current-cost net stock data from Table 2.1, accessed 15 March 2019.

VII. What is recorded in national accounts?

47. The 2008 SNA currently recognizes several types of intellectual property products that may be linked to data: software and databases, R&D, and goodwill and marketing assets. Software and databases and R&D are considered produced non-financial fixed assets and goodwill and marketing assets are non-produced non-financial assets.

Databases

48. “10.112 Databases consist of files of data organized in such a way as to permit resource-effective access and use of the data. **Databases may be developed exclusively for own use or for sale as an entity or for sale by means of a license to access the information contained. The standard conditions apply for when an own-use database, a purchased database or the license to access a database constitutes an asset**.”

49. “10.113 **The creation of a database will generally have to be estimated by a sum-of-costs approach.** The cost of the data base management system (DBMS) used should not be included in the costs but be treated as a computer software asset unless it is used under an operating lease. **The cost of preparing data in the appropriate format is included in the cost of the database but not the cost of acquiring or producing the data**. Other costs will include staff time estimated based on the amount of time spent in developing the database, an estimate of the capital services of the assets used in developing the database and costs of items used as intermediate consumption.”

50. “10.114 **Databases for sale should be valued at their market price, which includes the value of the information content.** If the value of a software component is available separately, it should be recorded as the sale of software.”

Research and development

51. 10.103 **Research and [experimental] development consists of the value of expenditures on creative work undertaken on a systematic basis in order to increase the stock of knowledge**, including knowledge of man, culture and society, and **use of this stock of knowledge to devise new applications**. This does not extend to including human capital as assets within the SNA. …Unless the market value of the R&D is observed directly, it may, by convention, be valued at the sum of costs, including the cost of unsuccessful R&D.

Goodwill and marketing assets

52. 10.199 **The value of goodwill[36] and marketing assets is defined as the difference between the value paid for an enterprise as a going concern and the sum of its assets less the sum of its liabilities, each item of which has been separately identified and valued.** Although goodwill is likely to be present in most corporations, for reasons of reliability of measurement it is only recorded in the SNA when its value is evidenced by a market transaction, usually the sale of the whole corporation. Exceptionally, identified marketing assets may be sold individually and separately from the whole corporation in which case their sale should also be recorded under this item.

53. Data combined with software and databases (storage for the data) and R&D are all likely intangible assets of data-driven businesses. For example, the financial filings of Facebook– a data-driven business– show significant R&D expenditures (USD 10.3 billion in 2018, a little over one-third of total costs and expenses) which consist primarily of compensation for software engineers and other technical employees who are responsible for building new products as well as improving existing products.[37] In other words, it may be hard in practice to distinguish between data and these other types of intangible assets. It is well-recognized that R&D and software overlap (as the example of Facebook states that their R&D includes work of software engineers) and that national statistical offices–if they estimate databases at all– estimate a combined “software and databases” category instead of trying to separately identify databases from software. Therefore, any estimation methods to value a digital data-based information asset (e.g. the information and knowledge derived from digital data) must consider these overlaps.

Overlap with other intangibles

54. Clearly the above discussion shows that valuing data-based information has considerable overlap with other intangible assets already capitalized within the 2008 SNA. So, what may need to be updated if one wanted to include the full value of the digital data-based information asset without double counting?

55. Since own-account software, database, and R&D production/GFCF are all estimated using the sum-of-costs approach a prudent approach to valuing a data-based information asset could be to review what possible additional costs are needed to fully account for the value of the data-based information asset.

[36] While conceptually goodwill and marketing assets are included in the SNA framework, in practice, very few countries publish estimates. Data are only available for the Czech Republic and France in the OECD database: Table 9B Balance sheets for non-financial assets. https://stats.oecd.org/

[37] https://investor.fb.com/financials/default.aspx

Data acquisition costs

56. What first stands out about the 2008 SNA paragraph 10.113 is that the cost of acquiring or producing the data is not included in the cost of the database. Bean (2016) notes that this recommendation is meant to avoid capitalizing the value of the data as a form of ‘knowledge’ in the national accounts and that data’s capitalization would depend on how it was stored. If the data were stored and embodied in a database, it would be capitalized. However, if data were stored elsewhere, e.g. on paper files, it would not be capitalized. If one agrees that digital data is fundamentally different than non-digital data as discussed above (i.e., digital data has allowed for new information/knowledge creation that could not have been done if the data were not in digital form) then this “inconsistency” is not of much concern: digital data is a different product.

Software and databases

57. Even though Software and databases are defined separately, the 2008 SNA recognizes the need to group the two assets together into one category because a computerized database cannot be developed independently of a database management system (DBMS), which is itself computer software.[38] Many countries use a macro approach for estimating own-account software and databases. This approach is based on labor costs for relevant occupations plus a markup for other expenses (including the costs of the capital used). Occupations (according to International Standard Classification of Occupations (ISCO)-08) that are included are 251 “Software and applications developers and analysts” and 2521 “Database designers and administrators”.[39]

58. Therefore, most likely work related to own-account software development that is done in the creation of the data-based information asset would already be included in software GFCF.[40] However, if the software development is done using open source software tools this would not be accounted for in GFCF. The OECD *Handbook on Measuring Intellectual Property Products* (OECD 2010) emphasizes that besides the labor costs, other costs such as overheads associated with employing staff engaged in database creation and updating as well as intermediate consumption associated with database creation should be included in database assets. These other costs should include the costs of software not recognized as a fixed asset. In this respect, it should be noted that to provide an exhaustive estimate of database creation one should also include payments for the use of cloud services for storage.

59. What potentially is missing are the costs incurred in analyzing the data. It is unclear where data scientists fit in the current ISCO-08 classification as the term does not appear when the classification document is searched. Data scientist could be included in the “Mathematicians, actuaries, and statisticians” group 2120 or possibly spread across multiple occupational categories because “Database analysts” are included in 2521 and “Data miner” are included in 2529. Including the labor costs of data scientists in the calculation of software and databases could be a way to derive more complete estimates of GFCF. However, some of these costs could be included in R&D.

60.

[38] SNA paragraph 10.109.

[39] Ribarsky, Konijn, Nijmeijr, and Zwijnenburg (2018) “Measuring the Stocks and Flows of Intellectual Property Products” http://www.iariw.org/copenhagen/konijn.pdf

[40] If countries use the “supply-side” approach in estimating own-account software production, then usually only software occupations such as software engineers and architects are included.

Own-account R&D

61. Many countries use the OECD's Frascati Manual (FM) based R&D surveys as a data source for deriving estimates of R&D.[41] As a starting point, the FM-based R&D estimates are derived from what is considered R&D performance (activity) and therefore, occupations are not used to include or exclude what is included in R&D costs.

62. In the United States, the starting point for the data collected on R&D expenditure in the FM survey is the item on the financial report.[42] As discussed above, Facebook has a significant amount of R&D expenditure. Much of this expenditure may be related to advancements in mining videos, pictures, and text for information[43] which could also be considered as expenditure for creating a data-based information asset. In addition, if the R&D expenditure includes work by software engineers in the creation of highly intelligent algorithms then there is the potential of triple counting (in R&D, software, and data-based information) if national accountants are not careful to remove the overlap in the estimation process.

63. Therefore, there appears to be the potential for overlap in what would be included in a data-based information asset and what is already included in R&D.

VIII. Summary and Conclusions

64. Digital data is a key factor of production in the modern data-driven economy, but the value of data inputs and the information derived from the data is hard to determine. This paper explores the perimeter of what national accountants may want to consider when determining the value of data-based information as well as whether it qualifies as a (fixed) asset. Clearly data-based information provides economic benefits and it can be used repeatedly in production. Further research needs to be done to determine how national accountants can identify long-lived data-based information (e.g., more than one year) versus short-lived.

65. While there are three approaches to valuing data-based information, the market-based approach is the least feasible to implement because most data-based information is not transacted on markets. For data that is transacted on the market, it is hard to determine if it would be an appropriate price for unpriced data since the value of data is heavily dependent on the use of the data.

66. The income-based approach could be a viable method for data-based information that can easily be tied to a particular use, such as targeted advertising. However, this method may be hard to use for all types of data-based information. The cost-based approach appears to be a feasible approach for national statistical offices to implement. In this respect, there seems to be a fair amount of data-based information already capitalized within the SNA. The related own-account intangible assets (software, databases, and R&D) are estimated using the cost-based approach, thus, using the same approach may help identify and avoid overlap. While

[41] Ribarsky, Konijn, Nijmeijr, and Zwijnenburg (2018) "Measuring the Stocks and Flows of Intellectual Property Products" http://www.iariw.org/copenhagen/konijn.pdf

[42] See question 2-1 of the 2016 Business Research and Development and Innovation Survey https://www.nsf.gov/statistics/srvyindustry/#qs : What was the total worldwide R&D expense for your company in 2016? If your company is publicly traded, this amount is equivalent to that disclosed on SEC Form 10-K as defined in FASB ASC Topic 730, Research and Development (FASB Statement No. 2, "Accounting for Research and Development Costs.")

[43] https://venturebeat.com/2015/04/22/why-facebooks-rd-spend-is-huge-right-now/

mixed estimation approaches can be used, national accountants would need to take care not to double (or triple) count.

67. If the cost-based approach is determined to be an appropriate method in calculating GFCF of data-based information then– in a future update to the international standards and/or guidance– one may want to consider including (1) acquisition costs of data in the calculation of databases and (2) costs associated with analyzing the data (e.g., data scientists) if they are not already accounted for in R&D or software. If this conceptual change is adopted statistical agencies may not find much of an impact if most of the costs are already capitalized.

68. Whether the SNA (fixed) asset boundary is extended to include these additional costs, data on data-driven businesses is in demand by policy makers. In this respect, the following could be explored:

- Further research into business accounts of firms that provide free services in exchange for data to see if certain costs can be identified as being related to data acquisition.
- Research the impact of including the acquisitions costs of data in the calculation of databases (including government).
- Explore the possibility of asking businesses the costs associated with developing digital data-based information. Care must be taken to understand the overlap with the other intellectual property products currently captured within the national accounts. Therefore, any data collection should delineate costs related to software and R&D (e.g. split between R&D and non-R&D work).
- Further research into potential overlap between R&D and digital data-based information. What is considered R&D by data-driven businesses and what is captured in surveys. Explore the possibility of expanding R&D surveys to account for data-based information creation if not already captured. In addition, ensure that R&D surveys are exhaustive and cover all industries that may perform R&D and not just selected ones.
- Identify advertising driven digital platforms and explore the possibility of using an income-based approach to valuing data-based information. Determine if the income-based approach is a viable method for other data-driven businesses.
- Refining the classification systems to better identify data-related activities, products, and occupations.
- Further explore the link with digital trade: cross-border data flows; Further research into recording in Balance of Payments and International Investment Position statistics.

References

Ahmad N., Ribarsky J., and Reinsdorf M., No. 2017/09, "Can potential mismeasurement of the digital economy explain the post-crisis slowdown in GDP and productivity growth?", OECD Publishing, Paris. Available at https://doi.org/10.1787/a8e751b7-en.

Ahmad N., van de Ven P. (2018); "Recording and measuring data in the System of National Accounts". Available at https://unstats.un.org/unsd/nationalaccount/aeg/2018/M12_3c1_Data_SNA_asset_boundary.pdf

Bean, C. (2016); Independent Review of UK Economic Statistics. Available at

https://www.gov.uk/government/publications/independent-review-of-uk-economic-statistics-final-report

Johnson G., Shriver S., and Du S (2018) "Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry? ". Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3020503

Li W., Nirei M., and Yamana K., "Value of Data: There's no Such Thing as a Free Lunch in the Digital Economy". Available at https://www.imf.org/en/News/Seminars/Conferences/2018/04/06/6th-statistics-forum

Mawer, C. (2015) "Valuing data is hard". Available at https://www.svds.com/valuing-data-is-hard/

Organisation for Economic Co-operation and Development, *Handbook on Deriving Capital Measures of Intellectual Property Products*, 2010. Available at http://www.oecd.org/sdd/na/44312350.pdf .

Organisation for Economic Co-operation and Development, *Frascati Manual 2015*, 2015. Available at http://www.oecd.org/publications/frascati-manual-2015-9789264239012-en.htm

Ribarsky J., Konijn P., Nijmeijr H., and Zwijnenburg J. (2018) "Measuring the Stocks and Flows of Intellectual Property Products". Available at http://www.iariw.org/copenhagen/konijn.pdf

Rowley, J. (2007) "The wisdom hierarchy: representations of the DIKW hierarchy"

Available at https://unigis.at/schnuppermodul/modul_gisintro/html/lektion5/media/rowley-2007.pdf

United Nations, European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development, World Bank, *System of National Accounts (SNA) 2008*, New York, 2009. Available at https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf .

United States Department of Commerce. Economics and Statistics Administration. 2014. "Fostering innovation, creating jobs, driving better decisions: The value of government data." Available at https://www.bea.gov/sites/default/files/2018-02/fostering-innovation-creating-jobs-driving-better-decisionsthe-value-of-government-data714.pdf

Varian, H. (2018) "Artificial Intelligence, Economics, and Industrial Organization". Available at https://www.nber.org/chapters/c14017.pdf

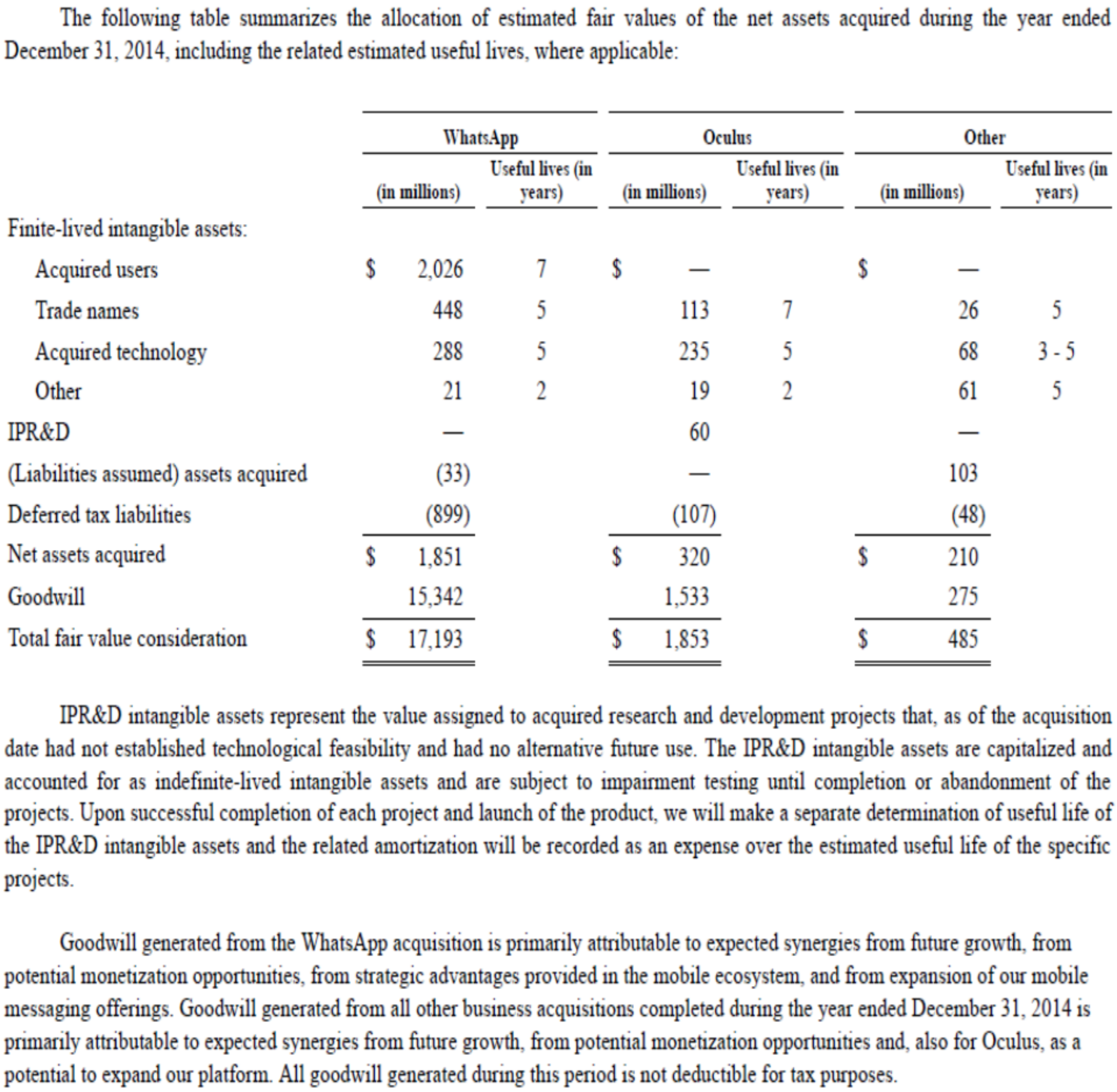
Varian, H., "In Conversation: Hal Varian on the Economics of Data"

https://www.lowyinstitute.org/news-and-media/multimedia/audio/conversation-hal-varian-economics-data

Annex 1

Merger and acquisitions

Many refer to the large amounts spent by digital businesses to acquire other companies as evidence that the value of data is large. In looking at Facebook’s acquisition of WhatsApp in 2014, they valued the acquisition of WhatsApp users at USD 2 billion, around 10% of the total fair value of USD 17.2 billion. Unlike the Nielsen acquisition of Gracenote, Facebook did not capitalize any acquisition of “content database”. Could the capitalization of “acquired users” be considered as the amount that is attributable to data? In the case of Nielsen, they made separate estimates of “customer-related” intangibles and “content database” intangibles.



The following table summarizes the allocation of estimated fair values of the net assets acquired during the year ended December 31, 2014, including the related estimated useful lives, where applicable:

	WhatsApp		Oculus		Other	
	(in millions)	Useful lives (in years)	(in millions)	Useful lives (in years)	(in millions)	Useful lives (in years)
Finite-lived intangible assets:						
Acquired users	$ 2,026	7	$ —		$ —	
Trade names	448	5	113	7	26	5
Acquired technology	288	5	235	5	68	3 - 5
Other	21	2	19	2	61	5
IPR&D	—		60		—	
(Liabilities assumed) assets acquired	(33)		—		103	
Deferred tax liabilities	(899)		(107)		(48)	
Net assets acquired	$ 1,851		$ 320		$ 210	
Goodwill	15,342		1,533		275	
Total fair value consideration	$ 17,193		$ 1,853		$ 485	

IPR&D intangible assets represent the value assigned to acquired research and development projects that, as of the acquisition date had not established technological feasibility and had no alternative future use. The IPR&D intangible assets are capitalized and accounted for as indefinite-lived intangible assets and are subject to impairment testing until completion or abandonment of the projects. Upon successful completion of each project and launch of the product, we will make a separate determination of useful life of the IPR&D intangible assets and the related amortization will be recorded as an expense over the estimated useful life of the specific projects.

Goodwill generated from the WhatsApp acquisition is primarily attributable to expected synergies from future growth, from potential monetization opportunities, from strategic advantages provided in the mobile ecosystem, and from expansion of our mobile messaging offerings. Goodwill generated from all other business acquisitions completed during the year ended December 31, 2014 is primarily attributable to expected synergies from future growth, from potential monetization opportunities and, also for Oculus, as a potential to expand our platform. All goodwill generated during this period is not deductible for tax purposes.