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

Marshall Reinsdorf and Jennifer Ribarsky (International Monetary Fund)[1]

DRAFT, December 28, 2019

Abstract

The strategic focus of businesses has long been to generate and control traditional intellectual property (IP) assets, such as patents and copyrights, but the strategic focus is now becoming the collection, control and deployment of data. An economy that is knowledge-based and data-driven poses significant measurement challenges for national income accountants. Business accounting either ignores the value of data or co-mingles it with assets like goodwill, so businesses' financial reports and survey responses are of little help for measuring data assets for national accounts purposes. Methods that national accountants could turn to for valuing digitized information assets are market prices for data, the costs of the data collection, processing and analysis involved in creating the assets, and the income attributable to the digitized information assets. Of these three, the cost approach is the most generally applicable.

In the framework for identifying and classifying assets of the *System of National Accounts 2008* (United Nations *et al*., 2009) raw data – the events and conditions that are observed – can be regarded as a non-produced asset, while information assets are produced through digitizing and processing the data and analysis of the processed data to extract meaning. In practice much of the value of digitized information assets may already be captured as part of software and databases or R&D (and the value of short-lived data used as an intermediate input, such as web browsing behavior used for targeted advertising, is captured as part of the value of the downstream outputs).

[1] The views expressed are those of the authors and do not necessarily represent the views of the International Monetary Fund (IMF), its Executive Board, or IMF management. This paper was prepared under the supervision of Gabriel Quirós-Romero.

I. Introduction

Data flows so abundantly throughout the modern economy that it has been called "the new 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 percent of all IP traffic will be video by 2022, up from 75 percent in 2017.[4] Setting aside the differences in value between different types of data flows are considered, the main physical use of data flows is *the delivery of content* (e.g. movies, books, music). In fact, according to Sandvine, Netflix is 15 percent of the total downstream[5] volume of traffic across the *entire Internet*.[6]

The distribution of content via data rather than physical media such as paper creates opportunities for flows in the other direction in which the content distributors to collect data on consumers. These data raise questions about *how businesses use data in a productive sense*. For example, how does Netflix use data on customers' browsing and viewing history (e.g., whether you binge watch, abandon a show or complete an entire season, etc.) and ratings to predict the popularity of movies and television that might be acquired, to guide the creation of content, and to make personalized viewing suggestions with a precision that no other platform will be able to match? In this respect, Netflix is about 3 percent of the total upstream volume of traffic (the traffic uploaded by users such as browsing the Netflix library). Businesses use this upstream traffic and other means of acquiring data *to generate information and knowledge*.

This paper considers how the value of data and assets derived from data, including digitized information and the knowledge, can be incorporated in national accounts. The SNA criteria for identifying and classifying asset types are applied to assets linked to data, and methods for estimating the asset values are discussed. The value transformation chain, in which raw data is processed and analyzed, and the resulting information and knowledge used to produce other assets and items for current consumption, means that in many cases the value of the data is already included in GDP as part of software and databases or R&D assets.

[2] According to Wikipedia, Internet Protocol (IP) is the principal communications protocol for relaying datagrams across network boundaries. Its routing function enables inter-networking, and essentially establishes the Internet. IP has the task of delivering packets to their destination based solely 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] A 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 from downloading items such as a video or music stream, a file, or a smartphone app.

[6] https://www.sandvine.com/2018-internet-phenomena-report. The report omits some data from China and India.

II. THE DATA TRANSFORMATION CHAIN

According to dictionary.com data is defined as

1. facts and statistics collected together for reference or analysis.

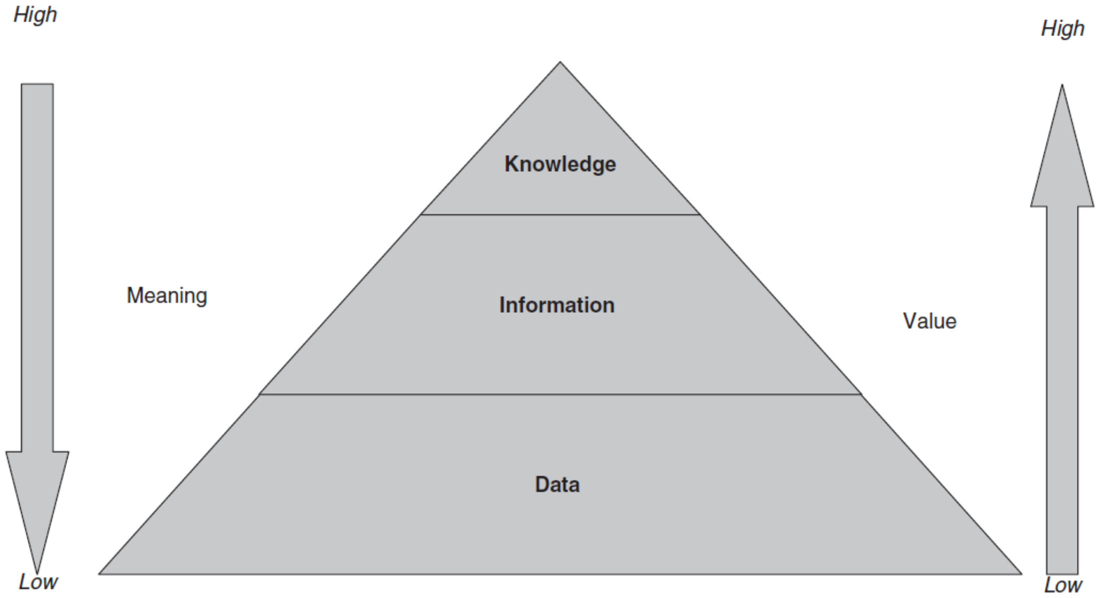
2. the quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

While data can be in analog form (e.g., stored in paper books), it is the ability to transmit, store and manipulate digital data in electronic form (bits and bytes) that has transformed the usefulness of data. Businesses use data to improve products and processes, to plan both short-run moves and long-run strategies, and to create artificial intelligence (AI) software via machine learning. They may also monetize the data immediately by selling it to a data broker, using it to generate in-app purchases or to advertise to targeted audiences, or renting access for purposes such as targeted advertising and fraud prevention. Such monetization of customer data is often what is meant when analysts refer to the "data-driven" economy.

While it is widely agreed that data drives the modern economy, its value is more contentious. Some say that data is very valuable. Others say that data has no value in and of itself because the value is found in the *insights derived from data*, and the *products created from data.*

The information and knowledge management literature discusses the nuances of the distinctions between data, information, and knowledge. As Rowley (2017) explains, data is not knowledge. The facts or observations of raw data are of no use until they are organized and stored in a suitable form. Data only acquires value once it has been transformed into knowledge, 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 embodying the transformation of information into instructions. Knowledge can be further differentiated into explicit knowledge (recorded in information systems) and tacit knowledge that has become part of the human psyche.

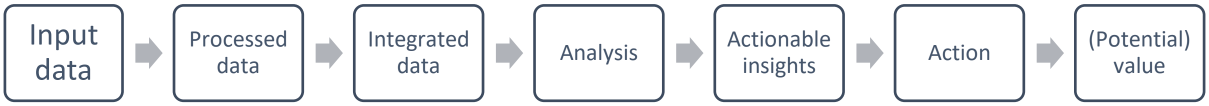
Figure. 1 Data, information, and knowledge



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

To become useful, data must undergo multiple processing steps involving collection, recording, organizing, structuring, storage, combination and integration with other data sources or information. 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 mark a transformation of 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. Factors Influencing the Value of Data

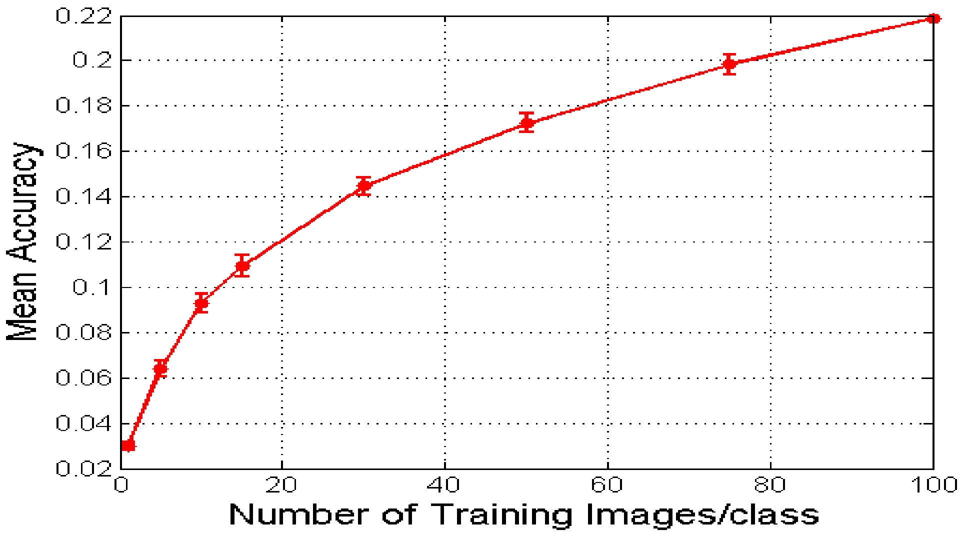
Value increases as data moves through the data transformation chain. The input data has much less value than the information (integrated data) and know-how (actionable insights) derived by processing it.

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

While an individual piece of data may be useless on its own, the value starts to increase when it is combined with other data to become a dataset. Bringing the data together creates an entity whose value is greater than the sum of its parts. A certain minimum amount of data is needed to be able to discern patterns and trends that improve information and knowledge, and with more data comes more certainty and more precision.

However, past a certain point, data generally ceases to exhibit increasing returns to scale. Varian (2018) uses the Stanford Dogs Dataset to illustrate decreasing returns to scale in machine learning. As seen in figure 3, accuracy improves as the number of training images increases, but at a decreasing rate. Whether data exhibits increasing or decreasing returns to scale may also depend on whether the increase in data involves simply adding another record of the same type (one more dog image) or a combination with complementary data that makes new applications possible.

Figure 3. Mean accuracy of Training Images



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

Input data can be classified based on how it is obtained:

- **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.

Data from the first two sources is usually not associated with a market transaction. However, there are exceptions, such as the reported payments by Facebook to certain users of up to USD 20 per month plus referral fees for access to their data after they installed 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).

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.

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.

What accounts for such divergent estimates of what a user profile is worth?

- **Quality of the 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.

- **Who else has access to the data**. Third party data is usually widely accessible, so a business is not necessarily gaining unique audience intelligence that is not also

[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/

available to their competitors. In the Facebook Research app example cited above, the access to the user’s data is most likely restricted to the company itself and is not widely accessible.

- **The identity of the user of data**. Input data has a number of possible uses that depend on the user of the data and the context – in other words how the business will use the information provided.

IV. DATA AS A FACTOR OF PRODUCTION

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. Representing data in electronic form has allowed its analysis 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 has allowed for new information/knowledge creation that could not have been done if the data were not in digital form.

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, the data used to “train” AI algorithms replaces some of the labor inputs for software creation, and AI algorithms then enable knowledge extraction and decision-making from diverse types of data that might otherwise have been hard to analyze. Finally, digital platforms and connected devices (e.g., Internet of Things – IoT) collect vast amounts of data. Consequently, digital data has become another factor of production, and Bean (2016) puts it on a par with physical and intangible capital.

Many new goods and novel features of new models have been enabled by technologies that use data as a factor of production. How much of the revenue of a data-centric enterprise goes to costs of data inputs may be hard to determine. Also, the consumer and producer surplus generated by new technologies for leveraging data inputs are part of the story of the value of data, but they may be more attributable to productivity gains and the value created during the production process than to the value of the data inputs.

V. IS DIGITAL DATA AN ASSET?

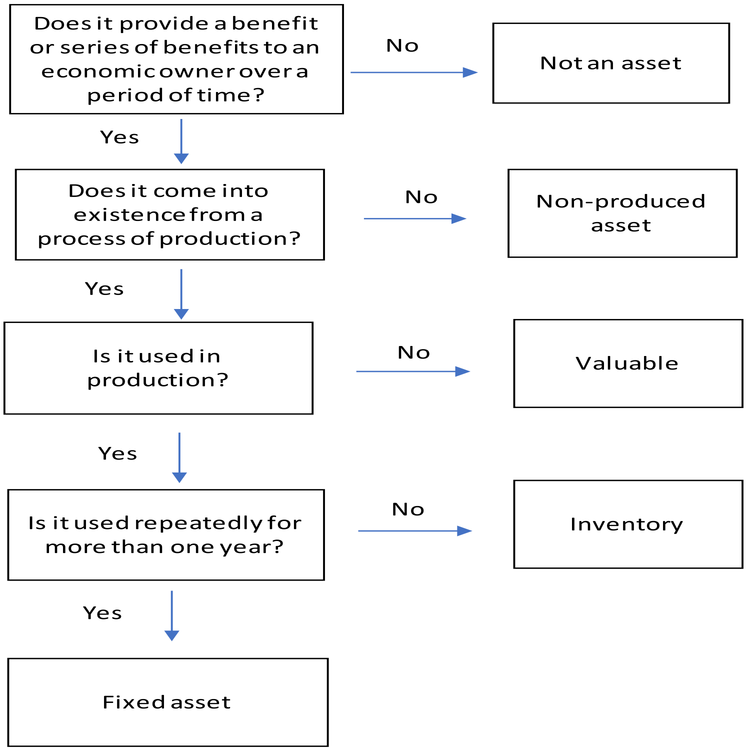
A. Digital data

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” (SNA 2008, 10.8). Non-financial assets fall into two broad categories: produced (coming into existence as outputs from production processes) and non-produced (e.g. land). Produced assets have three main types: fixed assets, inventories, and valuables. Fixed assets and inventories are held only by producers. Valuables may be held by any institutional unit and are primarily held as a store of value (SNA 2008, 10.10). To qualify as a **fixed** asset, a good or service must be used repeatedly or continuously in a production process **for more than one year** (SNA 2008, 10.11). Inventories consist of goods and services that 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 (SNA 10.12).

Criteria for classifying data as asset

Figure 4 shows a decision tree for determining if data is an asset and, if so, what type of asset. Data clearly provides economic benefits, but the first question also specifies that an asset must have an economic owner. Determining the economic ownership of data can be tricky because data is non-rival – the same data can be used simultaneously by multiple parties. To be an economic owner, the party in possession of the data must be able to prevent unauthorized copying (either legally or by restricting physical access) and must have general rights to use the data. On the other hand, a license to use the data for a particular purpose for a period of more than one year may represent an asset of the licensee. Distinguishing a license to use the data for a limited time or purpose as a different kind of asset from ownership of the data itself can help to avoid confusion.

Figure 4. Decision tree to determine if something is an asset and, if so, 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

"Period of time" is another potentially tricky part of the first question. Many kinds of data are not used repeatedly for more than one year and hence will not qualify as a fixed asset at the bottom of the decision tree. In particular, data collected on browsing activities for use in targeted advertising typically have a short shelf life – if the pair of shoes that you clicked on while shopping for shoes follows you around, you can be pretty sure that it will be gone within a month. Yet a framework for establishing inventories of short-lived data such as browsing behavior data would present some complications that might be best to avoid.

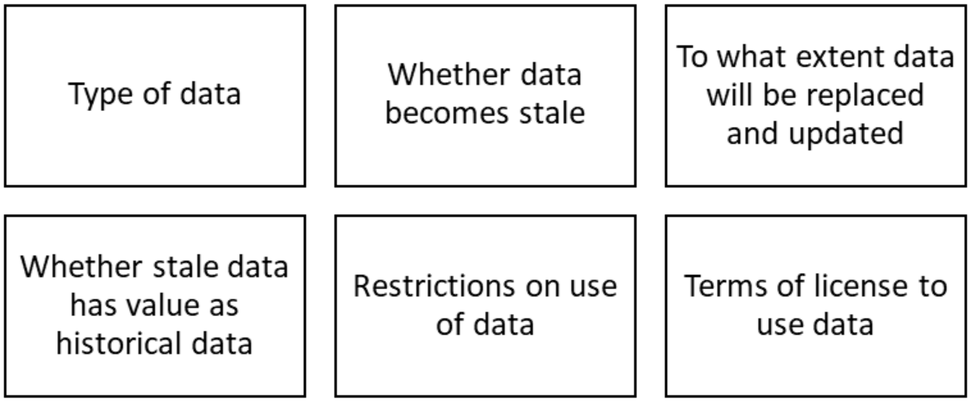
The SNA stipulates that "inventories consist of stocks of outputs that are still held by units that produced them prior to being further processed, sold, delivered to other units or used in other ways and stocks of products acquired from other units that are intended to be used for intermediate consumption or for resale without further processing" (SNA 2008, 10.12). Inventory items exit the stock by being used up in production, not via a depreciation allowance. For example, in the special case of a data set created solely to train a particular AI algorithm, the data set would be a work-in-progress inventory whose value will eventually be subsumed in the finished software.

Nevertheless, in principle, data can always be used again and again, and the inapplicability of the assumption that items are extinguished when used is one of the distinctive features of data.

This creates a dilemma, because dropping the assumption that the inventories of short-lived data are used up in production would have unacceptable consequences. Counting both the gross production of inventories of short-lived data and the full price of the products embodying the data will cause GDP to double-count output, just as if the flour and the bread had been both been included with no adjustment for the flour that went into the bread. Rather than subtracting depreciation of short-lived data assets from GDP—unlike every other kind of depreciation—it may be preferable to simply omit these assets. After all, the value of short-lived data assets should be well-captured in the estimates of the output of the products that these data help to create.

Data with a useful life of more than one year is conceptually more straightforward. Axciom, the data broker, capitalizes costs related to the acquisition or licensing of data used over two or more years in providing data products and other services. **These costs are amortized over the useful life of the data, which ranges from two to seven years**. To estimate the useful life of any acquired data, Axciom considers several factors, shown in figure 5.[11]

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 Annual Report for 2017

Data assets as a type of goodwill

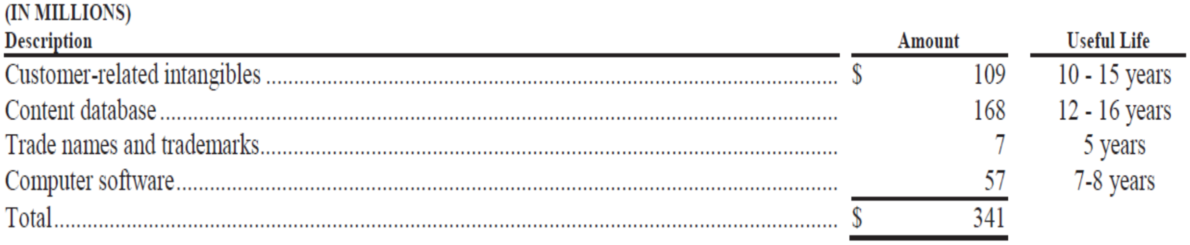
The decision tree of Figure 4 also has a question to distinguish between produced assets and non-produced assets. Unlike fixed capital formation and net additions to inventories, acquisitions of non-produced assets do not count in GDP. Ahmad and van de Ven (2018) identify two drawbacks of calling data in the abstract a produced asset. First, treating data as produced would seem to imply that knowledge in general is a produced asset, with unworkable implications. Second, as a practical matter, this approach would pose risks of double-counting,

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

as data would first be directly added to GDP and later included in GDP again as part of the value of the products derived from the data. Ahmad and van de Ven (2018) therefore recommend recording transactions related to data only when a monetary transaction occurs and to include them as a sub-item of goodwill.

Including acquired data assets in goodwill is a standard practice in business accounting. For example, Nielsen, a company founded in 1923, is a world leader in market research and ratings. In 2017, Nielsen's revenue was USD 6.6 billion, and its financial report[12] states that the 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." Yet, Nielsen's balance sheet includes a relatively tiny amount of data assets (USD 168 million, figure 6), all of which were recorded when Nielsen acquired Gracenote in 2017 for USD 585 million. Most of the purchase price for Gracenote was allocated to goodwill (USD 316 million) and amortizable intangible assets (USD 341 million). One of the reasons that Nielsen acquired Gracenote was to acquire Gracenote's global content database, which spans across platforms including multi-channel video programming distributors (MVPD's), smart television, streaming music services, connected devices, media players and in-car infotainment systems.

Figure 6. Nielsen's intangible asset 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	

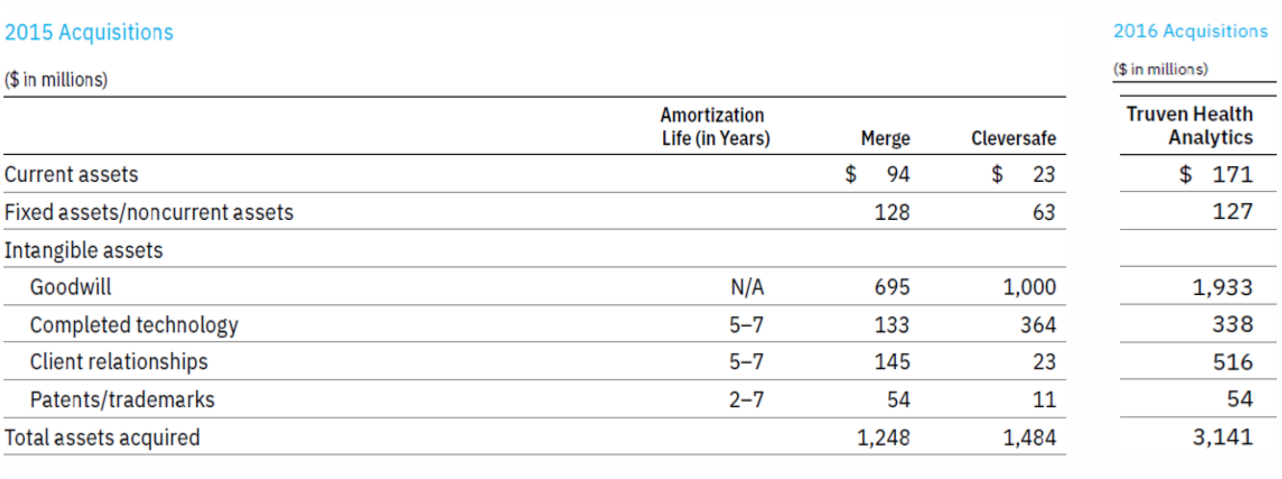
Source: Nielsen Annual Report for 2017.

Although purchased data assets are occasionally shown in financial reports of businesses (Annex 1), they are more likely to end up in goodwill. For example, when IBM acquired Merge in 2015 and Truven Health Analytics in 2016 to gain access to data for training its AI software, Watson, their data assets were all included in goodwill (Figure 7). Direct collection of values of businesses' data assets in official statistics will probably require a change in business accounting standards; at present, respondents to business surveys would be unable to separately identify data assets. Methods discussed below may allow estimation of investment in digitized information assets derived from data for national accounts purposes. However, the challenges for external sector statistics may be more difficult. Going beyond standard business accounting

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

to measure foreign direct investment positions and reinvested earnings could be impossible, and methods to estimate cross-border transactions in data assets have not yet been developed.

Figure 7. IBM's acquisitions of Data and Data Management Software
Millions of U.S. Dollars



2015 Acquisitions

($ in millions)

2016 Acquisitions

($ in millions)

	Amortization Life (in Years)	Merge	Cleversafe	Truven Health Analytics
Current assets		$ 94	$ 23	$ 171
Fixed assets/noncurrent assets		128	63	127
Intangible assets				
Goodwill	N/A	695	1,000	1,933
Completed technology	5–7	133	364	338
Client relationships	5–7	145	23	516
Patents/trademarks	2–7	54	11	54
Total assets acquired		1,248	1,484	3,141

Data as a non-produced asset

Business accounting's treatment of data assets as goodwill (or even as out of consideration) is not a good solution for national accounts. However, regarding raw data as non-produced would have advantages.

The states of nature and events that are the subject of observation are not produced but they are the basis of the value of the raw data. For example, a person's data is worth more if that person has a health condition, and the value that comes from presence of a health condition does not represent production. Production is characterized by control and management of costly inputs of labor, capital and materials to create an output (SNA 2008, 6.2), but large amounts of valuable data are generated as the costless by-product of running a business. The fact that raw data may be "found" through serendipity is also characteristic of a non-produced asset.

Even if present, a small element of production would not prevent a treatment of raw data as non-produced. Two kinds of non-financial transactions are recognized in the SNA: purchases of goods and services, and distributions of income (e.g. wages, interest and dividends).[13] Payments for the use of a produced asset, known as rentals, are purchases of services, while payments for the use of non-produced assets, known as rents, are distributions of income (SNA 2008, 7.153). Rent on a field used for growing crops is treated as a payment for access to a

[13] Payments that change the composition of net worth, such as repaying a debt, are financial transactions.

non-produced asset even though part of the value of the field comes from improvements like clearing, grading, and drainage

The production involved in transforming raw data into usable data sets by collecting, assembling, standardizing, organizing and storing in a format appropriate for analytical purposes is important enough to bring a dataset that is ready for use inside the production boundary. Note, however, that increases in the value of a dataset caused events such as the discovery of a new application for data, better technology for processing the data, or a change in data prices are "other changes in value", not production. Advances in hardware and software have unlocked the value of diverse kinds of data that were previously worthless.

If fixed capital formation in data assets is measured either from the projected market value of the data or the present value of the future cash flows that the data is expected to generate, the estimate will have a high risk of including a non-produced component. In contrast, the costs of the labor, capital and materials involved in collecting, organizing, assembling, cleaning and storing the (long-lived) data **can** be used to measure fixed capital formation in data assets. This represents the produced part of the value of data assets coming into existence in a given time period.

B. Information Derived from Digital Data

Digital data is an input for producing information and knowledge assets. The value of the data is embedded in the value of these assets, and potential problems of double-counting can be avoided by focusing on them. The view that information and knowledge are the relevant assets and not the data itself is consistent with Varian's (2018) perspective that it is the organization and analysis of the data that creates the value.

The phenomena that are observed and captured in the data (e.g. a person's characteristics and behavior) do not arise from a production process. Production occurs with the steps of gathering, organizing, cleaning, and formatting that transform the data into useful digitized information. Information and knowledge derived from data come into existence through a process of production: the input data is put into a digital form, organized in a database, and analyzed using inputs of labor (e.g. data scientists and software developers), and capital (computers, software and database structures).

User profiles that are developed by analyzing data to determine patterns of behavior are an example of a digitized information asset. Companies such as Axciom use this type of information repeatedly 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.

VI. Estimating the Value of Digitized Information Assets

Despite the potential for double counting output if the both the value of the data used as an intermediate input and the products embodying the value of that data are added to GDP, estimates of the value of short-lived data could be included in supply and use tables for digital economy and estimates of the value of long-lived data may be relevant to include in GFCF. As the information and knowledge derived from input data are typically not subject to market transactions, 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. The market-based approach

The 2008 SNA (3.119) states that transactions should be valued at market prices – prices paid by willing buyers to acquire something from willing sellers. If market prices are unavailable, market-price-equivalents can provide an approximation (SNA 2.123).

On a *conceptual basis* the market-based approach is the preferred concept of the SNA. The problem with applying this approach to data is that, except for commercial third-party databases, a truly comparable product sold on the market generally does not exist. One might attempt to estimate the value of unprocessed consumer data using the market price of user profiles sold by data brokers, but this is not an exact equivalent to third-party user profile data has undergone processing (e.g. organizing). Organizing data, cleaning it, and making it fit for use require significant resources. In addition, as discussed in above, market transactions in unprocessed data would only provide an estimate of the input data and not an estimate for the entire transformation chain needed to arrive at digitized information.

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 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 a range of sources in the United States of USD 5.1512.[14] This figure – which assumes that the subject has a rare health condition – is much greater than the USD 0.005 per profile quoted in the *Atlantic Monthly* article but less than the USD 20 per user that Facebook paid to acquire slightly more extensive data on users through their Facebook Research App. Reported prices of user data vary widely depending on the source and, more importantly, the intended use of the data.

Could market transactions also be used to value the vast amounts of data created by the IoT, primarily comes from sensors? Sensor data are often used as a current input into a production or control process, not problems requiring data histories. Also, IoT data tend to have value primarily to owner of the IoT device. These factors limit the potential for data markets to develop. A market price equivalent for IoT data will likely be possible to estimate in only a few special cases.

However, marketplaces have begun to appear for some kinds of IoT data. A company called Terbine has created a system to enable IoT data trading alliances. Terbine makes available data on the "physical world", ranging 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. It 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. 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.

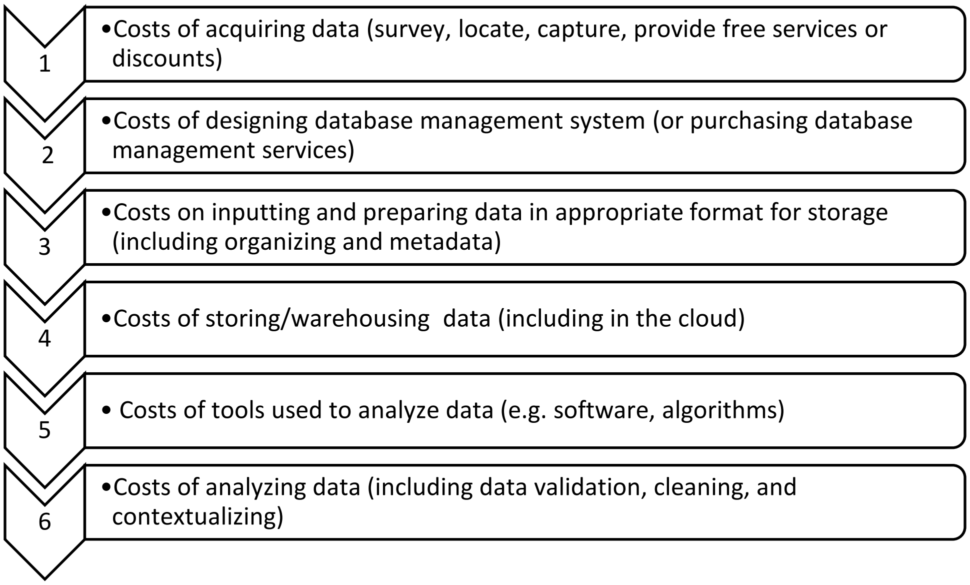
B. The cost approach

In the absence of a market price or market-price-equivalent, the NA recommends two alternatives for valuing an asset. These are valuation by the cost of producing the asset and valuation by income that the asset is expected to generate. Of these, the SNA gives preference to production costs. Indeed, many statistical agencies use the "sum-of-costs" approach to measure own-account gross fixed capital formation (GFCF) in software and databases and in research and development. Note that for market producers, the SNA includes a mark-up in the measure of costs (SNA 2008, 3.135). The SNA does not explain the details, but a mark-up consistent with a risk-adjusted competitive rate of return on the capital assets involved in the production would capture the opportunity cost of the capital deployed.

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

Figure 8 lists the potential direct costs of the steps for production of digitized information:[15]

Figure 8. Direct costs for creating digitized information assets



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)

Costs 1-4 are associated with database creation as a step on the way to digitized information/knowledge via costs 5 and 6. The first step in creating digitized information is acquisition of the input data. As noted above, some data are acquired through purchases (e.g. the rights to use the data) from third-party data providers, or, occasionally, explicit payments to households (e.g. the Facebook Research App example discussed above). Firms also hire market research firms to conduct surveys, focus groups, and interviews to collect data they need. To capture such transactions, national statistical offices could add questions on costs of acquiring data to their business surveys.

However, the data with the most distinctive roles in the new digital economy are acquired with no explicit payment. Large amounts of data are effectively generated for free as automatic by-products of the firms' daily operations – e.g. Amazon's data on customer order histories. The cost method is appropriate for measuring the contribution of these data to GFCF (which is zero), but for balance sheet purposes, an approach that would imply a positive value could be considered. More controversial are the data that digital platforms quietly collect on the activity of their users – their browsing, movements, communications and content creation. The direct cost of the software, storage and computation required to capture these data is small. However, a more comprehensive implementation of the cost approach could also include as indirect expenses the cost of producing the services that attract and engage the users so that their data can be collected.

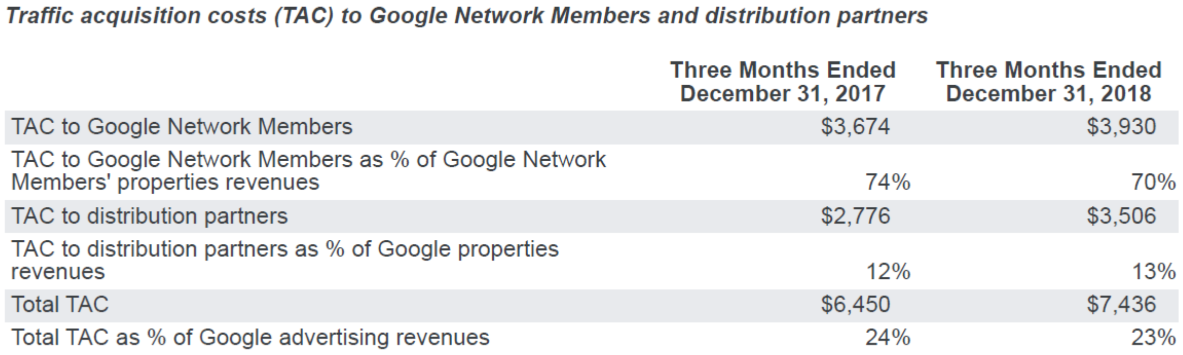
[15] A full accounting of costs would also allocate a portion of administrative and overhead costs.

A problem with including the costs to attract and engage users is that they generally have multiple objectives, just one of which is data collection. The same expenditures that create opportunities for the platform operator to collect users' data also create opportunities to show advertisements and sell one's own products. In addition, they allow the platform to benefit from the network effects and economies of scale that come with keeping existing users attached to the platform and attracting new users. Although the business value of customer relationships and network externalities is clear, adding customer relationships and network externalities to the fixed assets of the SNA would cause many conceptual and practical problems.

Traffic acquisition costs are an example of expenses whose purposes include creating data collection opportunities among other things. Traffic acquisition costs are payments made by one digital platform (e.g., Google, Yahoo, Baidu, Facebook) to another for directing consumer and business traffic to their websites. They are a major cost of revenue for many digital platforms. For example, Google reports that they are 24 percent its advertising revenues (Figure 9), and Goldman Sachs estimated that Google would pay Apple USD 12 billion in 2019 to make it the default search engine.[16] Expenses to create opportunities for data acquisition from consumers can also take the form of subsidies to prices of smart devices. For example, a smart TV's price is subsidized because the manufacturer uses the TV for data collection, advertising and content delivery.[17]

Figure 9. Google traffic acquisition costs[18]

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

The vast amount of data collection done by governments would also have to be considered if data is to be added to types of assets measured in national accounts. Government data is a key

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

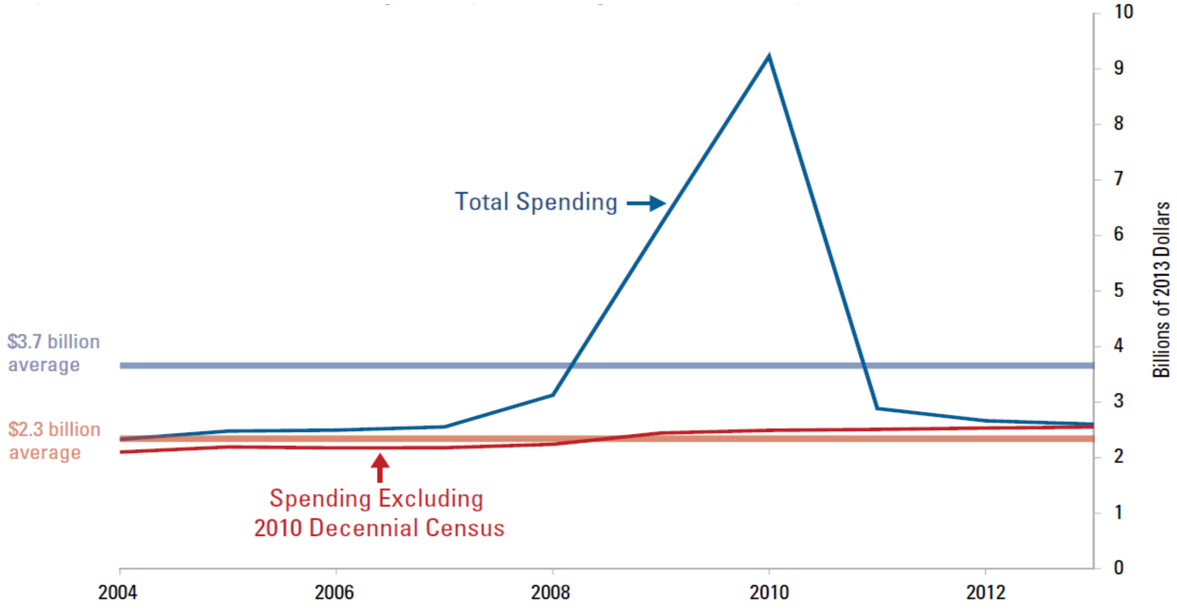
[17] https://www.theverge.com/2019/1/7/18172397/airplay-2-homekit-vizio-tv-bill-baxter-interview-vergecast-ces-2019

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

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)[19] reported that the United States spent USD 3.7 billion annually, adjusted for inflation, on data collection and dissemination by the Principal Statistical Agencies[20]. The Decennial Census is associated with a surge in expense. Excluding the Decennial Census, the average is about $2.3 billion (Figure 10). A consistent approach is critical in an integrated system like the SNA, and if data acquisition is to be capitalized, the costs of government acquisition of data cannot be ignored.

Figure 10. 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)

Once the data is collected, it must be put into a format suitable for use. This involves steps to organize, index, correct, and contextualize the data through metadata. Businesses that gather data often devote significant resources to these activities. The costs may include compensation

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

[20] These agencies' 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.

of staff such as data entry personnel, database architects, software architects and engineers, and the subject matter experts needed to help contextualize the data.

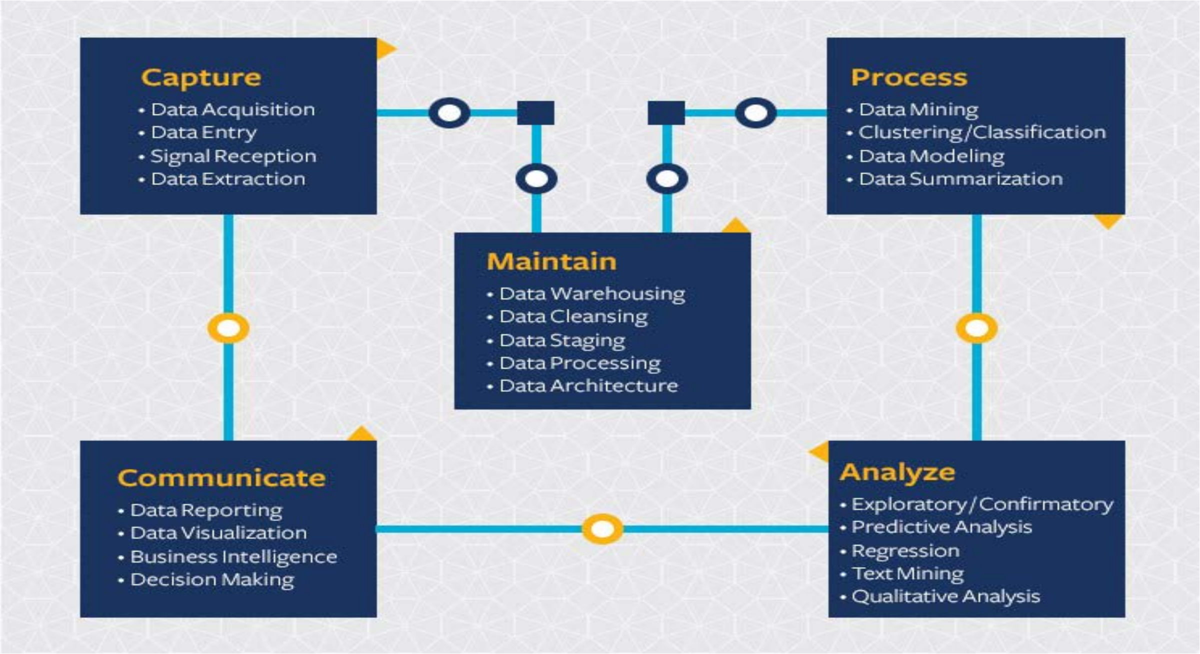
Data can be stored in company-owned servers or the cloud. Cloud computing encompasses several types of services delivered over the Internet, including data processing, storage, database management, and software. Even large data-driven businesses make use of cloud computing services. For example, Netflix uses Amazon Web Service (AWS) for its computing and storage needs, from customer data to recommendation algorithms.[21]

Once the data are in an appropriate format, statisticians[22] and data scientists – in the past we might have also called these people statisticians – use software tools (developed in-house, purchased, or even open source such as R and Python) to mine and analyze the data, or the data may be used to train an AI algorithm via machine learning. The analysis of the data may include additional data validation, cleaning, contextualizing for a given use and combining with other kinds of data. Figure 11 shows the tasks of the five stages of the data science life cycle.

[21] To be clear, Netflix designed the software (database architecture and algorithms) that run on AWS data processing and storage hardware. See https://www.computerworlduk.com/cloud-computing/how-netflix-moved-cloud-become-global-internet-tv-network-3683479/ .

[22] 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 11. The Data Science Life Cycle



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

C. Special Case of Collection of Data on Users of Digital Platforms

Collection of data on users of digital platforms is a distinctive feature of the so-called surveillance economy and a central part of the discussion of the need to measure data assets in national accounts. To facilitate data collection on users, a platform may offer services that attract and engage a large user base, or the maker of a "smart" device such as a TV may subsidize the purchase price and update its software for free to ensure that it stays in service. This raises a question of whether to include the cost of producing the services that attract platform users in the cost of acquiring data.

Much of the time that households' spend on platforms is for entertainment, so for simplicity we will call the free services to platform users "entertainment". Platform users are often said to barter their data for entertainment services. A depiction of the users as producing a service of data provision would misrepresent their role in the process. The platform users spend time on the platform being entertained and furthering their own objectives, not assembling their data and transmitting it to the platform in order to receive a payment. Nevertheless, platform users who know their data is being collected and consent to it they can still be regarded as providing something of value to the platform. They could be said to barter a license to collect data on what they do and say for access to the platform’s services.[23] As events that can be observed are not a produced asset, the imputed payment for permission to track the platform users would be classified as a rent, not a purchase of services. The users’ imputed income from the license to collect their data would then fund imputed purchases of entertainment services from the platform.

The platform’s saving would not change, as the imputed revenue and expenses would cancel out. Nevertheless, imputing consumption of entertainment services funded by users’ income from licensing access to their data would cause a double counting problem. Assume that the platform uses the data to sell targeted advertising, with the prices of the advertised products including mark-ups that fund the ads. If the output of the platform is recorded as its actual sales of advertising services plus the entertainment services that generate the data used to produce the advertising services, a portion of the platform’s output will be counted twice. The mistake would be analogous to measuring the output of a baker that mills its own flour by gross additions to flour inventories plus sales of bread.

Another potential problem with treating the platform users as bartering their data for entertainment is the failure to hold of the assumption of informed consent to collect data that is owned by the platform users. Assuming that the users are aware of collection of their data and consent to it may be unrealistic,[24] and assuming that platform users have ownership of their data might prevent the national accounts from later reflecting the change in their data ownership should data privacy laws be enacted. The SNA would include an acquisition of data assets by stealth in “other changes in volume of assets”.

D. The income-based approach

The present discounted value of the expected income flow from the data, or digitized information, may imply a different value for the asset from the cost-based approach depending on the uses of the information. The 2008 SNA recommends caution in using income to value the asset: “*If none of the methods mentioned above [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 of certainty, and given that assumptions are also needed about the asset’s life length and the discount factor …, the other*

[23] Nakamura, Samuels, and Soloveichick (2017) impute a related kind of barter transaction in which the platform users exchange the service of viewing advertising and marketing material for entertainment. The entertainment services are valued by their production cost.

[24] A journalist’s surprise at the findings of an audit of the hidden trackers his phone, which were found to number more than 5000, illustrates this point. See "It’s the middle of the night. Do you know who your iPhone is talking to?", Washington Post, May 28, 2019. https://www.washingtonpost.com/technology/2019/05/28/its-middle-night-do-you-know-who-your-iphone-is-talking/

sources of valuation…should be exhausted before resorting to this method.” (SNA 2008, 3.137).

The income approach is most straightforward to apply when the asset has a single, easily identifiable use, and is recommended for musical, literary, and photographic works – items for which there is an established system of royalty flows (OECD, 2010). However, a common difficulty with the income approach is the impossibility of isolating the cash flows (net of associated costs) uniquely related to the asset being valued.[25] Although crediting the entire operating income of a digital platform that could not exist without its data assets to those assets might provide an interesting perspective for a philosophical discussion of the importance of data, the necessary assumptions would be unacceptable for national accounts purposes. In most cases, the platform's income equally well be attributed to the other tangible and intangible assets, including technical and business competencies, customer relationships and network externalities, and innovations. Another potential drawback of the income approach is that the implied estimates of the capital stock will be unsuitable for analysis of total factor productivity (TFP) if the income generated by the asset embodies the increase in TFP under analysis.

Many digital platforms that collect data get all (or nearly all) of their revenue from targeted advertising. For purposes of illustration, assume that the useful life span of the data used for targeting is long enough to treat it as a fixed asset. Attributing the entire cash flow after direct expenses from targeted advertising to the data will overstate the value of the data – some of what the advertisers are paying for is general access to "eyeballs". The European advertising revenue of the *New York Times* did not decline when it eliminated targeting in response to the General Data Protection Regulation (GDPR)[26] but an income-approach estimate of zero for the value of the data would be too low. Johnson et al. (2018) find that ads shown to users who opt-out from being behaviorally targeted garner 52 percent less revenue on an ad exchange than do comparable ads for users who allow behavioral targeting. This suggests that just over half of online advertising revenue could be attributed to data used for targeting.

In the U.S., the Internet Publishing and Broadcasting and Web Search Portals industry earned the bulk of its revenue from advertising, USD 105.2 billion in 2017 or 62 percent of total revenue (Table 1). If just over half of the revenue and associated costs[27] from online advertising space can be attributed to digitized information, assuming a 3 year life and a real discount rate of 8 percent gives an estimate of the value for the stock of digitized information for the U.S. Internet Publishing industry of USD 85.5 billion (Table 2, column 3, line 1). If the industry's entire revenue from online advertising is attributed to the digitized information, then the estimate rises to USD 164.4 billion (Table 2, column 3, line 2). To put these amounts in

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

[26] https://digiday.com/media/new-york-times-gdpr-cut-off-ad-exchanges-europe-ad-revenue/

[27] We assume costs are 55% of revenue based on Facebook's average costs-to-revenue in 2018-2014.

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 choice of 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

Table 2. Value of digitized 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		
Assumption	**7 years**	**3 years**	**7 years**	**3 years**	**Memo: Software stock**
Digitized information responsible for part of advertising revenue	148.3	85.5	162.2	89.0	…
Digitized information responsible for *all* advertising revenue	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 digitized information asset for U.S. Internet Publishing and Broadcasting and Web Search Portals industry; Software figures are from the Bureau of Economic Analysis, current-cost net stock data from NIPA Table 2.1, accessed 15 March 2019.

VII. What is Already Recorded in National Accounts?

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*
> “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**.”
>
> “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.”
>
> “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*
> 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*

> 10.199 **The value of goodwill[28] 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.

The intangible assets of data-driven businesses include data, software and databases (storage for the data) and R&D, and these items can be hard to separate out. For example, the financial filings of Facebook show substantial R&D expenditures (USD 10.3 billion in 2018, a little over one-third of total costs and expenses), consisting primarily of compensation for software engineers and other technical employees who are responsible for building new products as well as improving existing products.[29] The employees engaged in R&D, software development, and development of data assets are hard to distinguish from each other. The overlap between R&D and software is well-known (as the example of Facebook states that their R&D includes work of software engineers), and so national statistical offices–if they estimate databases at all–estimate a pooled "software and databases" category. Estimating the value of a digitized information asset (e.g. the information and knowledge derived from digital data) in isolation may be impossible, but not a combined estimate of data assets, software and database assets, and R&D assets.

Overlap with other intangibles

The above discussion shows that digitized information has considerable overlap with other intangible assets already capitalized within the 2008 SNA. On the one hand, the digitized information assets that are already being measured as part of the value of other kinds of intangible assets help to improve the accuracy of the totals of intangible asset stocks and investment. On the other hand, the similarity of the technologies used to produce the various types of inter-mingled intangible assets makes distinguishing the digitized information that remains uncaptured a challenge. Furthermore, in actual practice, the procedures used to estimate the other kinds of intangible assets may have gaps. In particular, the proliferation of data and the low cost of processing and analyzing data have "democratized" the performance of R&D, moving it outside the realm of specialists, and existing R&D surveys may not include the new performers of R&D.

[28] 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/

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

The sum-of-costs approach is generally used to estimate investment in own-account software, databases, and R&D assets, and the same approach will likely be applied to digitized information assets. The question is how to identify the additional costs that will need to be included.

Data acquisition costs

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 value of the database asset. 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. If the treatment of digital data is changed, the capitalization of data will then depend on how it is stored. If embodied in a database, it will be capitalized. Otherwise, (e.g. if stored on paper), it will not be capitalized. If one accepts that digital data differs fundamentally from non-digital data (perhaps because 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 a concern: digital data is a different product.

Software and databases

Even though the SNA defines software and databases as conceptually separate, it recognizes that a digital database cannot be developed independently of a database management system (DBMS), which is software (SNA 2008, 10.109). Because of the practical impossibility of separately estimating the two, the usual practice is to compile a combined value of investment in own-account software and databases based on labor costs for relevant occupations plus a markup for other expenses (including costs of capital used). Occupations that are included are: 251 “Software and applications developers and analysts”, and 2521 “Database designers and administrators”, where the code are defined in the International Standard Classification of Occupations – ISCO-08 (Ribarsky, Konijn, Nijmeijr, and Zwijnenburg, 2018).

Therefore, the in-house software development part of the production of digitized information assets is already be captured in software GFCF (though any free open source software tools used in this development would not be reflected).[30] 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 purchased inputs associated with database creation) should be included in database assets. These “other costs” include the costs of software licenses. In practice, however, some of the other costs involved in software and database creation may be missed. The cost of the data used to train AI algorithms, and the costs of cloud computing and cloud storage could be overlooked when measuring investment in software and databases using the cost approach.

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

Another element of costs that could potentially be missed is the data analysis. The place of data scientists in the current ISCO-08 classification system is unclear, as this occupation has not yet been added. Data scientists might be included in the "Mathematicians, actuaries, and statisticians" group 2120 or possibly spread across multiple occupational categories: "Database analysts" are included in 2521 and "Data miner" are included in 2529. If the data scientists can be identified, including their labor in the compilation of investment in software and databases could capture this missing element of investment in information assets. (However, some of this labor cost may already be included in R&D.)

Own-account R&D

Many countries use the R&D surveys based on the OECD's Frascati Manual (FM) as a data source for deriving estimates of R&D (Ribarsky, Konijn, Nijmeijr, and Zwijnenburg, 2018). These R&D estimates are derived from reported expenses for R&D performance (activity), and so occupations are not used to include or exclude what is included in R&D costs.

This approach creates a potential for overlap with what would be included in a new digitized information asset. In the United States, the starting point for the data on R&D expenditure of businesses collected in the FM survey is the R&D spending figure in the financial report.[31] Facebook, for example, reports large expenditures on R&D. Much of this may be related to advancements in mining videos, pictures, and text for information,[32] things that might be included in a new digitized information asset. If the R&D expenditure figure includes work by software engineers to develop innovative algorithms, there is even a potential for triple-counting (in R&D, software, and digitized information) if national accountants are not careful about removing overlaps.

VIII. SUMMARY AND CONCLUSIONS

Digital data is a key factor of production in today's data-driven economy, but the value of data inputs and the information derived from the data can be hard to determine. This paper explores the issues for national accounts in measuring, classifying and recording assets derived from data, referred to here as digitized information. Digitized information assets clearly provide economic benefits and are used repeatedly in production, but further research will be needed on how national accountants can identify the long-lived digitized information that would

[31] 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.")

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

qualify as a fixed asset. Attempting to record inventories of short-lived digital information (defined as having a useful service life of a year or less) would pose a dilemma of choosing between a distortion picture of output in which the production of the digitized information would be double-counted or a distorted picture of production processes involving information in which information can only be used once.

Of the three possible approaches to valuing digitized information, the market-based approach is least likely to be both feasible to implement and reliable. Most digitized information is not transacted on markets, and even when a market for data does exist, the prices may misrepresent the value of non-traded data because the value of data depends heavily on the way it is used. The income-based approach is a promising method for valuing digitized information that tied solely to a particular use, such as targeted advertising. However, this is a special case. In general, the cost-based approach will be the most feasible option for national statistical offices to implement.

A considerable amount of digitized information is already capitalized as part of other kinds of assets. The own-account investment in related intangible assets (software, databases, and R&D) is already estimated using the cost-based approach. Use of the same approach for digitized information assets may be helpful for identifying overlaps in order to avoid double counting. In contrast, with mixed estimation approaches, national accountants will need to take extra care not to double (or even triple) count.

If the cost-based approach is accepted as an appropriate method to calculate GFCF of digitized information then both the acquisition costs of data used to populate databases and the costs associated with analyzing the data (e.g., data scientists) could be included with adjustments for items already counted as part of R&D or software. The impact on the GDP of adopting this conceptual change may be turn out to be small if most of the costs are already capitalized. Yet it is also possible that the search for data-related assets will uncover investment in R&D and software that was hitherto unnoticed, and this could noticeably affect estimated GDP.

Regardless of the appropriate treatment of acquisitions of data in GDP, supplementary measures of data-driven businesses are in high demand. 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 on inclusion of data acquisition costs when deriving the value of databases using the cost approach (including government databases).

- Explore the possibility of surveying businesses on costs associated with developing digitized information assets. Care would be necessary to identify any overlap with the other intellectual property products already captured in the national accounts. Therefore, any data collection should delineate costs related to software and R&D.

- Further research into overlaps between R&D and digitized information assets and the expansion of R&D surveys to account for digitized information assets currently not being captured. In addition, research on uses of data for R&D should consider whether existing R&D surveys cover all the industries performing R&D. The abundance of data and the low cost of processing and analyzing these data may have made performance of R&D nearly ubiquitous.

- Explore the possibility of using an income-based approach to value digitized information assets of platforms that collect user data, particularly those that use the data for targeted advertising.

- Update the classification systems to better identify data-related activities, products, and occupations, including adding “data scientist” as an occupation.

- Further explore the links between data and digital trade, including the evidence on the importance of cross-border data flows and the treatment of digitized information assets in Balance of Payments and International Investment Position statistics.

IX. 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/

Nakamura, Leonard, Jon Samuels, and Rachel Soloveichick, 2017, "Measuring the 'Free' Digital Economy within the GDP and Productivity Accounts", ESCoE Discussion Paper 2017-03.

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 .

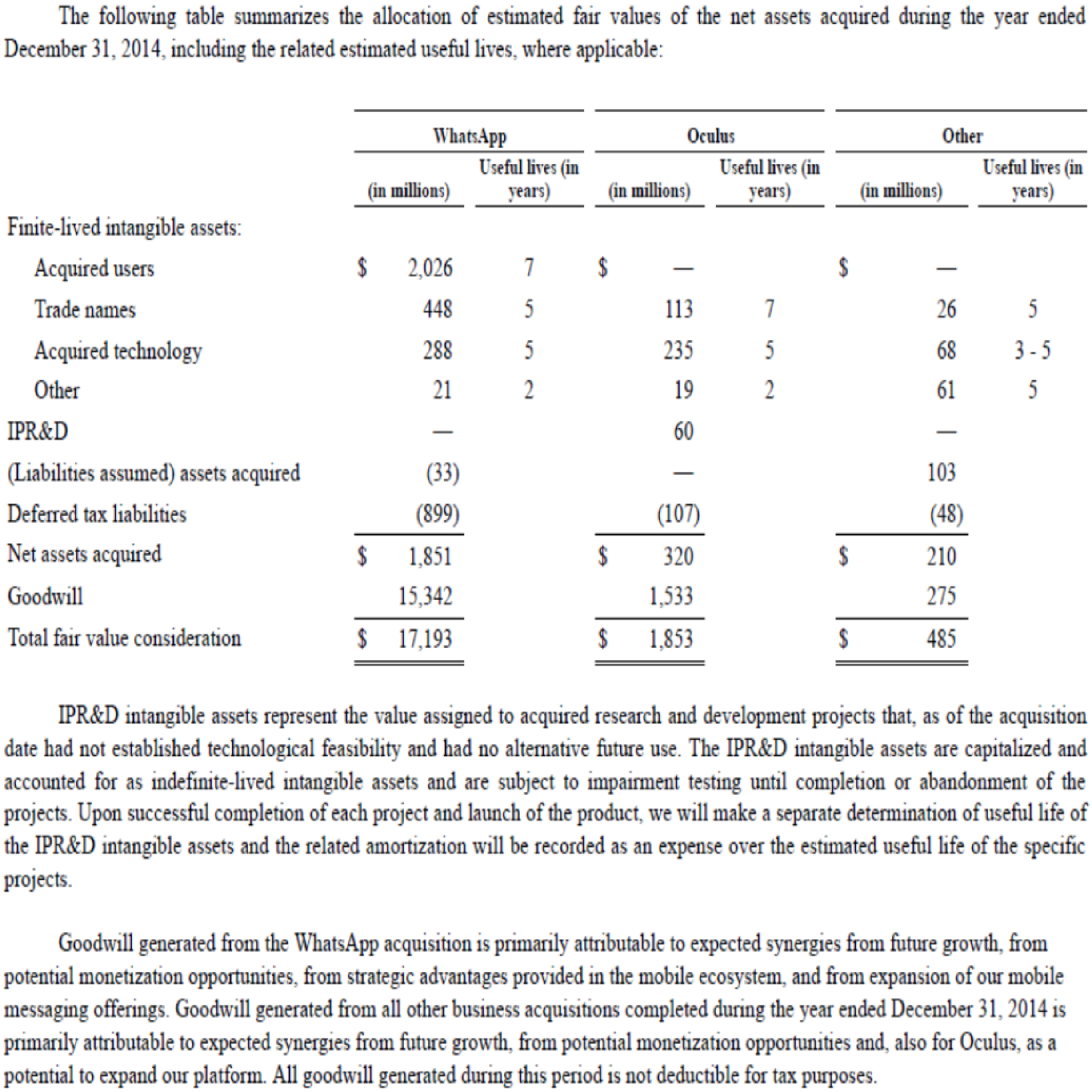
United States Department of Commerce. Economics and Statistics Administration. 2014. “Fostering innovation, creating jobs, driving better decisions: The value of government data.” Available at https://www.bea.gov/sites/default/files/2018-02/fostering-innovation-creating-jobs-driving-better-decisionsthe-value-of-government-data714.pdf

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

Varian, H., (undated) “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.