United Nations ECE/CES/GE.20/2020/4

Economic and Social Council

Distr.: General
14 February 2020

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on National Accounts

Nineteenth session
Geneva, 27-30 April 2020
Item 5 of the provisional agenda
Digitalization

The value of data, databases and data science in Canada

Prepared by Statistics Canada

Summary

The paper provides an overview of the statistical framework used to develop an experimental set of estimates of the amounts invested in Canadian data, databases and data science in the recent years. Measurement of intangibles within national accounts has focused on traditional forms of intangibles, including investment in software, research and development, and mineral exploration and evaluation. However, as data take on a far more prominent role in Canada and, indeed, all over the world, data, databases and data science have become a staple of modern life. The increasing use of and investment in data is driving economic growth, changing the employment landscape and reshaping how and from where we buy and sell goods. Yet, data is not well measured in the existing statistical system. Given the 'lack of data on data', Statistics Canada has initiated new research to produce a first set of estimates of the value of data, databases and data science. The estimated value of the stock of data, databases and data science in 2018 was roughly equivalent to two third of the estimated value of established crude bitumen reserves in Canada.

GE.20-02320(E)

2002320

I. Introduction

1. All over the world the use of data has increased exponentially largely due to the ease with which information is captured, converted to digital format, stored and analyzed for the extraction of knowledge. In the 1930s and 1940s, the first computers were rudimentary, slow, expensive and cumbersome with little memory or storage capacity. Today, after many decades of innovation, they are fast, cheap and miniaturized with enormous memory and storage capabilities and capable of executing complex algorithms. These developments have both enabled and encouraged a rapid growth in the collection, digital storage and usage of a wide variety of types of data.

2. Yet despite these indisputable trends, data still only have a small explicit role to play and little visibility in the modern national accounting framework. This is because data usage, to a large extent (though certainly not always), is unpriced in the modern economy while the economic indicators released by statistical agencies are mostly about market-determined values. Some data are produced by businesses and governments for their own use but not sold in the marketplace, for example by internal corporate accounting departments. Other data are supplied by households to businesses and governments as payment-in-kind in exchange for other services, as for example in the case of Facebook, Google and many other online services. In these and other situations data flows are a crucial part of the economic landscape, but they are not readily apparent in the economic indicators.

3. To address this situation, an expansion of the current national accounting concepts and statistical methods is proposed to measure data activities which will shed light on these highly consequential changes in society that are related to the rising usage of data. Given the 'lack of data on data', Statistics Canada has initiated new research to produce a first set of estimates on the value of data, databases and data science.

4. In 2018, Canadian investment in data, databases and data science is estimated to be as high as $40 billion. This is larger than the annual investment in industrial machinery, transportation equipment and research and development and represents approximately 12% of total non-residential investment in 2018. Average annual growth in data investment between 2015 and 2018 was 6.2%, much higher than the average annual growth in machinery and equipment (+2.2%), non-residential buildings (+2.0%), engineering structures (-4.7%) and research and development (+0.5%).

II. An 'information value chain'

5. 'Data' is a common word, but what does it mean exactly? What should it be defined as for economic analysis purposes? The word data has evolved to an extent where it is synonymous with information that either is or can be stored, transmitted and processed in digital form.

6. For the purposes of this study the 'data' will be defined as: "observations that have been converted into a digital form that can be stored, transmitted or processed and from which knowledge can be drawn". This definition does not imply that everything digitized is therefore data. For example, a song that has been converted into digital format (or even recorded in digital format) is still a song – it will not be redefined as data just because there is a digital representation of the song. The definition proposed in this paper is limiting the definition of data to those observations (such as the weather, or the number of 'likes' on my latest post, or the number of goals my favourite hockey player scored in her last game) that someone or something has converted into a digital form and can therefore be stored, retrieved, manipulated and investigated at a point in time.

7. Having narrowed the definition of data this must now be put into a broader context. One way to think about data, as defined above, is that it is part of a larger information chain. This information value chain can be envisioned as having four unique and separable states. At the base of the chain is simply observations. Observations can be anything—from the

temperature, to the fact that someone bikes to work or eats lunch at a specific time. Individuals, objects and the environment emit observations continuously. Observations are often fleeting and intangible. Observations do not necessarily need to be perceived by humans. In other words, objects and the environment can 'emit observations' even if there is no human observing them. While many observations are irrelevant and will never be recorded they can be seen to represent the sum total of all activity—human or otherwise.

8. Often, for various reasons, someone may choose to record observations. In the past, prior to the advent of digital technologies, these observations were often recorded in books and ledgers. This was mainly to keep an historical record of activities either because some regulation required it or the observations would be needed at a later point in time to execute a task. In today's digital world, pencil and paper have been replaced by the keyboard, sensors and electronic storage devices. This second layer in the value chain, where observations are converted into digital form, will be referred to as data.

9. Data is the digital representation of observations or activities. In order for data to come to be, someone has to decide that something needs to be recorded and has to set up the capture system so the observations can be both taken and stored. This recording implies that something is being done by someone. In layman's terms, when something is done for economic reasons, or an economic purpose, the System of National Accounts recommends recording this as production. In other words, there is a strong argument, in this case, that data are produced.

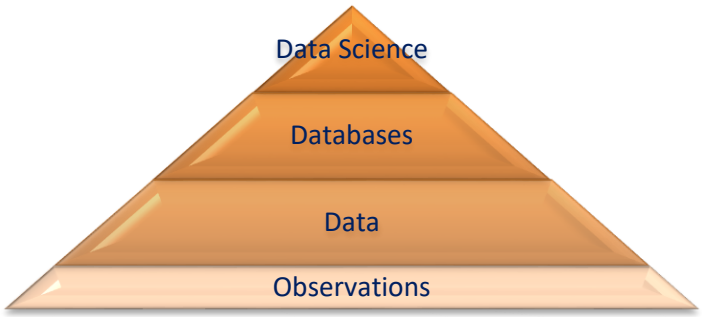
10. Additional value can be added to this chain by organizing and structuring the captured bytes of data. The 2008 SNA defines the product 'databases'. It states (paragraph 10.112): "*Databases consist of files of data organized in such a way as to permit resource-effective access and use of the data. 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*."

11. It is important to distinguish between data and databases. They are not the same. Data are observations that have been converted into a digital form that are stored. They can be thought of as raw material. They are the bytes of information that have not yet been structured and are not easily interpretable. A database is an organized store of data that can be readily retrieved and manipulated. Databases or structured data can then be considered the third tier in the information chain. The boundary between data and databases can be blurry. The main distinguishing feature between the two is that there generally is a normalization process that occurs between data and databases. Data or digitized observations can be seen as singular and separate, a database brings these observations together in a structured way. For example, a small business may record the IP addresses that visit their website. Each visit is a data point. The small business may decide to load all of these observations or data points into a database for retrieval or further analysis. The task (or production) of bringing the data together into a single database is separate from the task (or production) of digitizing the observation of someone visiting the website.

12. The fourth tier, and possibly the most valuable one, is when individuals are able to glean insights or new knowledge from the observations that were digitized and became data and then were organized in a database to facilitate retrieval and analysis. It is true that each data point or datum embodies some knowledge. This fourth tier goes beyond measuring the knowledge contained in each datum. It involves the collective knowledge that can only be gleaned when a volume of data is looked at as a whole. This new knowledge includes patterns and relationships that are not evident when looking at each datum in isolation. The definition for this activity is embedded in the 2008 SNA definition of research and development where it states that research and development are undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and use this stock of knowledge to devise new applications. This part of the information chain does

not signify a deviation from the 2008 SNA standard. For the purposes of this paper this activity is referred to as data science.

Figure 1
Information value



13. The data science activity is different and separable from the databases that support it, the raw data and the underlying observations contained in each datum. There is a circular flow to the information value chain and it may be characterized alternatively as an information cycle. In many ways, observations become data, data are stored into databases, new knowledge is drawn out of the databases through systematic investigation and this new knowledge becomes observations.

III. Nature of data, databases and data science

14. An essential question about observations, data, databases and data science—or the information value chain—is: what part of the chain is 'produced' and what part is 'non-produced'? The answer to this question determines what gets included in GDP and what is excluded.

15. 2008 SNA already answers this question for databases and data science. Databases are recognized as assets and are produced. Given it is difficult to distinguish databases from database management software, the value of a database is often grouped with its supporting software. Similarly, 2008 SNA defines research and development (which includes the definition of data science) as "*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*." (2008 SNA 10.103).

16. Research and development, including data science, are recognized as produced assets within the Canadian System of National Accounts (CSNA) under the asset category intellectual property products (IPP). While the CSNA conceptually includes these assets, the data system employed to measure data science activities needs to be re-examined.

17. While 2008 SNA clearly argues that research and development and databases are produced assets it says little about the other parts of the information value chain. As a result, countries do not record observations nor data as defined in this paper. The 2008 SNA argues that since there is no production process leading to their existence, observations and data fall outside the economic production boundary. Increases in observations or data will therefore have no impact on measures of economic activity such as gross domestic product or national wealth. Given the different ways observations and data are being used, it is important to re-examine this guidance. Are observations and/or data, in fact, produced and should they enter the production boundary?

18. Some observations can be thought of as a natural resource. Much like fresh air exists, or trees exist or minerals exist, observations in their purest form simply exist. They are a

consequence of the actions of humans and the environment. In some cases, you could argue some observations are produced, such as someone observing someone else riding their bike—it takes a lot of work to produce that observation. In some sense observations are everything we do. We go to work, we have dinner with our family, we exercise, the wind blows, it is cold, it is sunny—all of these are observations. We use these observations each day to manage our activities and make decisions. We exchange observations every day: whenever you ask someone how they are doing and they respond, you are the recipient of an observation they are providing to you. While most observations are related to 'doing', most are not being done for economic purposes.

19. Given these examples and assumptions, it is difficult to argue that observations are produced assets. For the purposes of this paper, observations will therefore be treated as non-produced. This is not to say that observations do not have value. Observations can have significant, often life-saving value. This is only stating they are not produced and therefore fall outside the economic accounting production boundary.

20. What about data? As previously-defined, data are "observations that have been converted into a digital form that can be stored, transmitted or processed and from which knowledge can be drawn". In the preceding paragraphs it is argued that observations are non-produced. Does this extend to data? Are data different from the observations they embody? Are data produced? There are a few attributes of data that provide a clue as to the answers to these questions.

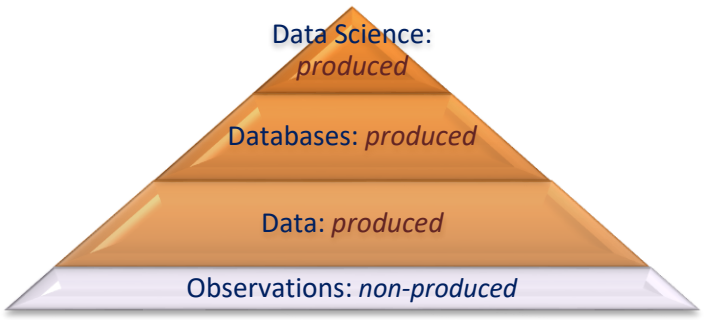
21. First, there is a process that needs to take place to convert observations into bytes of data. Sometimes this process can be costless or have low marginal cost such as when data are generated using a sensor. Often, these processes do not require a 'labour input', such as in the case of a sensor reading the quality of air. Regardless of the cost there is some sort of transformation that occurs by which an observation changes states from non-digital to digital.

22. Businesses and households spend significant resources protecting data. This is a strong indication that data are owned or at least someone is acting as a custodian of the data. The fact that data appear to be owned is another signal that, by extension, they are produced.

23. Finally, more and more businesses are selling data as either a primary or secondary output. Data are a product and in order for data to be sold they must first be produced. An analogy could be drawn with air. Most of the air life on earth breathes is not produced and has no market price. But scuba divers need air and there are companies that compress the air into tanks and sell it to them. So even though most air is not produced, some air is produced. The same applies to observations. Observations are not produced but when they are digitized and sold something has been produced. The cost of producing the data may be very low or zero at the margin but there appears to be a market for data—independent of the software used to store and retrieve the data.

24. Given the presence of these 'clues', for the purposes of this paper it is considered that data are produced. They do not simply 'appear'. There must be an action that brings the data into existence.

Figure 2
Refined information value



25. Given the previous arguments Figure 1 can be updated by delineating that portion of the information value chain that is produced (and therefore should be valued) and that portion that is non-produced. See Figure 2.

IV. Valuing data, databases and data science

26. Data, databases and data science can be produced and either used by firms on 'own-account' or sold on the market. Those sold on the market are in theory valued at market price (the value of the transaction). Ideally, Statistics Canada would survey Canadian firms and obtain information related to their market sales of data, databases and data science. At this time, Statistics Canada has very little information on the market sales of data, databases and data science. Data, databases, and data science that are used on 'own-account' are valued at the cost of producing the product including an estimate for return on capital. Since Statistics Canada does not have information on the market sales of data, databases and data science, the production of all data, databases and data science (whether for market sale or for use on own account) have been valued at the cost of producing the product.

27. The estimates are calculated from employment and wage information collected by the quinquennial Census of Population and the monthly Labour Force Survey, combined with a number of important, but as yet largely untested, assumptions. Occupational groups are selected from among those in the National Occupational Classification (NOC) that are generally associated with data activities (converting observations into data, organizing the data into databases, analysing those databases to gain knowledge).

28. Employees working in these NOC categories are unlikely to spend all of their work time producing data. They may also be involved in several other kinds of activity. Information on the share of their work time applied to data activities is presently unavailable, so subjective assumptions were made. In view of the uncertainty associated with these assumptions, two alternatives were considered (upper and lower ranges). Additional work is required in the future to collect factual information about both the specific occupational groups that engage in data production activities and the shares of their labour inputs associated with the activity.

V. Data

29. Data are produced and therefore included and valued within the System of National Accounts (SNA) production boundary. In some cases, data are bought and sold in market transactions. In these situations the value is simply the market price. In other perhaps more common cases, data are produced and used within a business, a government or a non-profit

institution. In these instances, since an arm's-length market-determined value is unavailable, the associated value must be estimated.

30. So, if a business purchases data from another business, the value is the transaction price. For example, if Statistics Canada purchases financial information from Bloomberg Canada, the data will be valued at the price negotiated between the two parties.

31. Traditionally, the method used to value own-account products (created and used in-house) has been to add up the costs to produce them, 'marked up' by a normal return to capital. As noted, often the cost of digitizing observations on own-account (at the margin) is close to zero since it may not require labour input. Examples of the types of activities involved in producing data range from the labour costs associated with capturing information from paper in a machine-readable form to the costs associated with operating a drone to acquire digital images for a geographic location. Further, advances in artificial intelligence and machine learning make it possible for complex natural language algorithms to be constructed which take digital unstructured information (such as a photo) and turn it into coded and highly structured information from which databases can be built and knowledge can be acquired.

VI. Databases

32. Recommended methods to value databases are outlined in the 2008 SNA. It notes that the value of a database will generally have to be estimated by a sum-of-costs approach (para.10.113). The costs include the cost of preparing data in the appropriate format for the database; the time spent by staff in developing the database; the capital services of the assets used in developing the database; and the costs of items used as intermediate consumption.

33. Databases purchased on the market should be valued at purchasers' prices, while those developed in-house should be valued at their estimated basic price or at their costs of production (including a return to capital for market producers) if it is not possible to estimate the basic price (para. A3.60).

34. In most cases, the challenge for national accountants is not a conceptual one but rather a 'lack of information' problem. The lines between software, databases and services (such as client support services) are often blurred. As a result, in many cases statistical agencies assume that database investments are captured in the estimates of gross fixed capital formation in software. Statistics Canada does this as well, although there is evidence that databases are not fully captured under the current methodology.

35. Statistics Canada's existing methodology to estimate own-account software (including databases) investment involves identifying a number of occupational groups related to the development of software and databases and making assumptions about the amount of time these groups of employees spend developing software and databases for own final use within the enterprise. In addition to the labour input cost, Statistics Canada also includes non-labour costs associated with the development of the software such as electricity, building rental and other types of overhead.

VII. Data science

36. Similar to databases, the 2008 SNA manual provides national accountants with a standard method to estimate the value of investment in research and development. When research and development results are sold on the market, the market price is used for valuation. When research and development is undertaken for own final use, a sum-of-costs approach is used. In the case of Canada, while data analytics are included in research and development in principle and the conceptual framework and methods for measuring them exist, the increasing range of businesses engaged in data analytics means there is a potential statistical under-estimation. The problem is that the current collection instruments are geared

to gather information from a relatively small set of businesses that are known to be research-intensive.

37. Canada's estimates of research and development activities are derived from two main sources of information. The Research and Development in Canadian Industry (RDCI) survey is used to measure the research and development activities of firms in the non-financial and financial corporate sectors. It is a cross-economy survey of approximately 8,000 firms. A number of federal and provincial government surveys are used to measure research and development activities in the government sector.

38. While the RDCI sample and survey strategy are appropriate for traditional forms of research and development such as pharmaceutical research and development or software engineering, they are not as well designed from either an instrument or sampling perspective to capture the growing research using big data—what has been referred to as data science. For example, retailers and banks are using insights gleaned from their massive stores of personal data to help drive sales. These kinds of insights fit the 2008 SNA definition of research and development. The problem, at least in the case of Canada, is that current statistical methods and tools do not fully capture this research and development investment activity.

39. In order to develop an order-of-magnitude estimate of the value of data science investment the same approach described above for data and databases is adopted. A share of production activities is assumed for each of occupational groups and as with data and databases, an assumed markup for non-direct labour and other costs is applied to the direct labour costs. The estimates (presented as ranges) of the value of investment in data, databases and data science are presented in Table 1.

Table 1
Estimates of the value of investment in data, databases and data science

	2005	*2010*	*2015*	*2018*
	millions of dollars			
Total of all data-related categories				
lower range	14,693	17,788	26,029	29,455
upper range	19,995	24,125	35,192	40,025
'Data'				
lower range	6,777	7,559	8,916	9,418
upper range	9,742	10,840	13,448	14,216
'Databases'				
lower range	3,087	4,143	5,945	8,046
upper range	4,564	6,104	8,599	11,625
'Data science'				
lower range	4,829	6,085	11,168	11,991
upper range	5,689	7,181	13,145	14,184
		percent		
Annual growth rate				
lower range	...	3.90	7.91	4.21
upper range	...	3.83	7.84	4.38

... not applicable

Source: Statistics Canada, special tabulation.

VIII. Stock of data related assets

40. This paper proposes that a significant amount of 'information-related' activity creates stores of value, from which firms draw in subsequent periods to produce goods and services. Given that data, databases and data science investments are being made by businesses, governments and non-profit institutions each day, a stock of these assets is also being accumulated. This stock needs to be included on the balance sheet of the sector that owns it, at its market value.

41. Valuing the stock of data, databases and data science assets poses a number of interesting challenges. The first relates to the depreciation profile. These assets do not physically depreciate, so a naturally observable profile cannot be drawn upon. In some cases firms store massive amounts of information indefinitely, although the perceived utility of these assets may decline. At the same time, in other cases the value of information is fleeting and should not be capitalized if it is not used for a period extending beyond one year (the knowledge that it would rain yesterday was most useful before and during yesterday).

42. For pragmatic reasons it is assumed that data have a useful life of 25 years, databases have the same useful life as software which is 5 years, and data-driven research and development have a useful life of 6 years, the same assumption that is made for other forms of research and development.

43. The reason for the assumption of a 25-year useful life for data is based on how long it is expected that a firm will store data or at least draw upon stored data to gain insight. Since much of the data that are currently used are behavioural, it can be assumed such data will only retain their value for a 'generation'. A generation is often defined as the period it takes for children to be born, grow, become adults and start to have children of their own. Of course there are many other types of data as well. The 25-year assumption should be regarded as quite tentative and more research is required on this matter. For all three data-related asset types a net capital stock with a geometric depreciation profile was estimated.

44. The second challenge associated with measuring the stock of data is in establishing a 'current market price'. In the previous sections an approach to data valuation is outlined, but that approach only applies to the initial value of data. The market value of data can change substantially from one period to another. Since it is believed that most of the data, databases and data science as outlined in this paper are internally produced and used by businesses, governments and non-profit organizations, the price of these assets will be a function of the input costs related to direct labour compensation and non-direct labour compensation and non-labour costs such as utilities, employee support services and capital services. For the purposes of this paper only the direct labour input costs were considered when estimating the price of data, databases and data science.

45. Table 2 shows the price indexes of data, databases and data science, based on the input cost of each and an assumed 3% capital service charge. It also shows the estimated net capital stocks at the end of the four years 2005, 2010, 2015 and 2018, at current prices, and the average annual rates of growth of those capital stock estimates.

Table 2
Price indexes of data, databases and data science

	2005	*2010*	*2015*	*2018*
		2005=100		
Total, price indexes for data-related categories	100.0	109.5	119.1	126.3
'Data'	100.0	112.6	122.0	130.7
'Databases'	100.0	103.6	113.3	121.7
'Data science'	100.0	108.4	116.9	121.2
		millions of dollars		
Total, net capital stock for data-related categories				
lower range	74,058	100,512	131,950	157,067
upper range	97,855	136,055	181,098	217,659
'Data'				
lower range	53,549	74,181	92,133	104,824
upper range	71,571	102,231	130,569	150,993
'Databases'				
lower range	6,926	9,302	13,015	18,692
upper range	10,290	13,740	18,954	27,050
'Data science'				
lower range	13,582	17,029	26,801	33,551
upper range	15,993	20,084	31,576	39,616
		percent		
Annual growth rate for total net capital stock				
lower range	...	6.3	5.6	6.0
upper range	...	6.8	5.9	6.3

... not applicable

IX. Conclusion and next steps

46. Data science and its antecedents, data and databases, are becoming more and more central in the modern world. So much of what we do nowadays is digitally chronicled as data, loaded into databases and exploited analytically for a wide variety of purposes. During the day our purchases, travel, reading, listening and media viewing activities, physical activities, likes and dislikes and so much more are stored for use toward various ends. Even our physical states while sleeping are increasingly being recorded. Yet while the growing importance of the information chain is evident, the existing framework for economic measurement does not reveal much about it.

47. These estimates and the previous releases on this topic, are an effort to build upon the established economic statistics framework in a way that makes the roles of and temporal changes in data, databases and data science more evident. They entail a number of assumptions that need to be tested, so the numerical estimates are tentative and presented as

ranges rather than point valuations. Nevertheless, the estimates indicate significant and growing investment expenditure and capital stock in data, databases and data science.

48. Statistics Canada will continue to build upon these initial estimates in order to provide a comprehensive picture of data activities in Canada. Plans are underway to engage with the private sector to refine the estimates of time spent on these activities as well as the types of occupations involved. Due to their fluidity, it is difficult to determine economic ownership of these assets as well as the estimation of cross-border transfers. Close collaboration with input data suppliers will be required to modernize the methodology and ensure correct measurement.

49. Statistics Canada is working with the international community to seek general acceptance of the framework, additionally, discussions are underway as part of the planned revision to international standards for national accounting in the digitalization stream. The date of incorporation within the CSNA is not yet determined.

Appendix

Measuring data, databases and data science: conceptual framework
https://www150.statcan.gc.ca/n1/daily-quotidien/190624/dq190624a-eng.htm

The value of data: experimental estimates
https://www150.statcan.gc.ca/n1/daily-quotidien/190710/dq190710a-eng.htm