National Data Strategy: Review of commonly quoted statistics

The National Data Strategy's aim: 'The NDS aims to build a world leading data economy unlocking the power of data across government and the wider economy, while building public trust and confidence in its use'

The government's National Data Strategy needs to be grounded in a firm understanding of the available evidence on the data economy. The National Data Strategy (NDS) team in the Department for Digital, Culture, Media & Sport (DCMS) has heard a range of facts and statistics. In January 2019, Policy Lab facilitated an evidence kick-off workshop for the team to examine statistics quoted by several governmental and non-governmental bodies. The following document summarizes the sources and methodological foundation behind the statistics identified. If any statistic lacks an identifiable source or defensible research methodology, it has been flagged below with a short description in **orange**.

This Policy Lab evidence review highlights that many of the most commonly quoted statistics around the use of data are not always as robust as we might think them to be - reinforcing the need for DCMS to launch a 'call for evidence' to underpin the National Data Strategy process. Improving, via this process, our collective knowledge and understanding of the use and value of data will be beneficial to organisations across sectors and across the economy.

Contents

Executive summary

‘Data-driven companies are 10% more productive’

Statistic 1

This statistic is based on a relatively small sample of verbal interviews with HR professionals. It may be true, but requires more robust evidentiary support.

‘The value of the U.K. data economy is £73.3 billion’

Statistic 2

This statistic is backed by both primary research (market surveys) and publicly available statistical data, but its data is outdated and sensitive to variations in how the ‘data economy’ is defined. The actual value of the U.K. data economy in 2016 may range somewhere between £61.3 billion and £73.3 billion, and this value may be considerably different in 2019.

‘The value of the U.K. data economy was £216 billion over the years 2012-2017’

Statistic 3

This statistic is a projection using 2011 economic estimates and projected data analytics adoption rates through 2017; it does not accurately reflect actual valuations after 2011. Further, it sources its data from an undisclosed body of academic and industry literature.

‘The total amount of global data is expected to grow by 40% year on year for the next decade’

Statistic 4

The data and research methodology behind this statistic can be sourced but not verified. Additionally, the statistic only reflects the projected rate of global data expansion through 2020.

‘The value of removing fees from all currently available data for re-use is estimated to be £404 million over 10 years’

Statistic 5

No source can be verified for this statistic. Various sources with similar information have been provided in its section below.

‘Effective use of data can create £66 billion of new business and innovation opportunities in the U.K.’

Statistic 6

This statistic erroneously applies projected cumulative economic benefits of the adoption of big data analytics to data sharing and ‘data liquidity’. The underlying data represents the cumulative benefits to all public and private U.K. businesses for the years 2012 through 2017; it is not an annual estimate. Further, this statistic is derived from estimates and projections made in 2011 based on an undisclosed body of academic and industry literature.

‘90% of data was created in the last 2 years’

Statistic 7

While the statistic represents a trend of exponentially increasing data production corroborated by several sources, its underlying data and methodology cannot be verified.

Statistic 1

'Data driven companies are 10% more productive'

***Source*:** 2015 Nesta report

Direct Quote

'Companies in the Datavores and Data Builders group are, on average, over 10 per cent more productive than Dataphobes.'

Takeaway

Companies who make heavy use of data to drive decision-making and innovate have increased their financial strength faster than those companies who don't. While this effect persists when both sector and factor controls are added, it is correlational, not causational, and **based on a relatively small sample of verbal interviews with HR professionals.** The statistic may be true (or even more extreme, given that companies without plans to use data were excluded), but would **need to be verified by further research.**

Summary

This report builds on earlier research by Nesta on how companies use data. They found four subgroups, three of which are active data users (Datavores, Data Builders, and Data Mixers) and one of which does not incorporate data into decision-making (Dataphobes). **Datavores** are data-driven companies that build their businesses around data analytics. **Data Builders** are companies that do not rely on data processes for decision-making, but still collect and work with large volumes of data. **Data Mixers** are similar to Data Builders, but combine data from different sources without creating large, new datasets themselves. **Dataphobes** 'work with few, small datasets, and rarely use analysis to make decisions'.

When controlling for extraneous variables, Datavores and Data Builders are significantly more productive (measured as longitudinal financial strength) than Dataphobes, but Data Mixers are not. Nesta suggest that 'firms that are less advanced with data use it incrementally, with a focus on efficiency, whereas companies further along their data journey make it part of their innovation and strategic processes.' Data Mixers may not be using data to a sufficient level to significantly increase productivity over the Dataphobe benchmark.

Methodology

The Nesta report draws its analysis from a random sample of 404 medium-sized and large-sized businesses in Bureau Van Dijk's FAME Database[1]. The sectors analysed were Manufacturing, Retail, ICT, Creative Media, Financial Services, and Pharmaceuticals.

Analysis was conducted on verbal responses to a phone questionnaire based on a previous Nesta survey[2]. The survey includes questions regarding:

- The **data value chain** (data collection practices, data processing and analysis, and data applications within businesses)
- **Data talent sources** (where are companies getting their data talent?)
- **Labour market experiences** (recent recruitment experiences and skills shortages)
- **Skills strategies** (investments in tools and training to improve employee skills)
- **Data talent management practices** (practices used to create value from data talent, including team make-up, organisation, communication and compensation structures)

Respondents were screened to only include those companies where data plays a non-negligible role and where there is an intention to build up an analytical capability. (Dataphobes still fall within these inclusion criteria.) In most cases, data was collected from those serving HR functions.

[1] Bureau Van Dijk's FAME (Financial Analysis Made Easy) is a database of about nine million companies in the UK and Ireland. For more information, visit: http://www.bvdinfo.com/Products/Company-Information/National/FAME.aspx

[2] Livingstone, I. and Hope, A. (2011) 'Next Gen. Transforming the UK into the World's Leading Talent Hub for the Video Games and Visual Effects Industries. London: NESTA; Bakhshi, H. and Mateos–Garcia, J. (2012) 'Rise of the Datavores.' London: Nesta; Winterbotham, M. 'UK Commission's Employer Skills Survey 2013: UK Results, January 2014.' See: https://media.nesta.org.uk/documents/next_gen_wv.pdf

The methodology is prone to bias in non-anonymous verbal responses by HR representatives at companies willing to take the call (the response rate was 19%). Accordingly, Nesta supplemented the questionnaire responses with econometric data from the FAME database, controlled for **factor controls** (e.g. production input), **sector controls**, and **general controls** (company age, innovativeness, etc.) . The statistic is taken from this econometric analysis.

Statistic 2

'The value of the U.K. data economy is £73.3 billion'

***Source*:** Digital Realty Data Economy Report 2018 (see also the European Data Market Monitoring Tool)

Direct Quote

'The U.K. Data Economy is estimated to have generated economic output (GVA) worth £73.3 billion in 2016' - Data Economy Report, p. 35.

Takeaway

The statistic is backed by both primary research (market surveys) and publicly available statistical data, but its data, from 2016, is **outdated** and **its validity depends wholly on the definition of 'Data Economy' used**. The actual value of the U.K. data economy in 2016 may range somewhere between £61.3 billion and £73.3 billion (a **£12 billion difference**), and this value may be considerably different in 2019.

Summary

Data Realty defines the Data Economy as the 'financial and economic value created by the storage, retrieval, and analysis - via a sophisticated software and other tools - of large volumes of highly detailed business and organisational data at very high speeds'. Also included in this definition is the Internet of Things (IoT) ecosystem.

Aggregating GVA data from the Office of National Statistics (ONS) for all of the sectors included in Data Realty's broad definition combined to produce a total GVA of £73.3 billion in 2016. The ICT, financial services, and professional services sectors - the largest contributors in the U.K. - accounted for 64% of this total.

Using the same yearly £:€ conversion ratio used by Data Realty[3], IDC and Open Evidence's EU Data Market study (via their European Data Market Monitoring Tool) values the 2016 value of the U.K. Data Economy at £61.3 billion. Here, the Data Economy is defined as 'the overall impacts of the data market on the economy as a whole', which 'involves the generation, collection, storage, processing, distribution, analysis, elaboration, delivery, and exploitation of data enabled by digital technologies' and 'includes the direct, indirect, and induced effects of the data market on the economy.'

Methodology

The Data Realty report based its findings on GVA metrics taken from the 2016 statistics reported by the ONS (up-to-date statistics available here) as well as a 'detailed review of relevant economic and business literature'. The exact SIC07 descriptions within the sectoral scope of the Data Economy are not explained, but include a large variety of sectors, including ICT, financial services, professional services, and manufacturing. (The difference between the Data Realty estimate and that of the European Data Market Monitoring Tool may be caused by this broader definition.)

The EU Data Market study, in contrast, conducted data collection through two ad hoc surveys carried out in Spring 2015, which 'targeted potential data companies in the ICT and professional service industries (235 interviews) and data users in eleven industries (1,184 interviews)'. This new data was combined with previous data sources (including Eurostat business demography statistics in the European Union, IDC's market forecast estimates for IT hardware, software, and IT services from 2014 to 2016, the IDC Worldwide Black Book, etc.) to estimate the value of the data economy within several E.U. and non-E.U. countries. However, IDC found it impossible to estimate the forward indirect impacts of the data economy and thus excluded it from analysis despite it being the 'most relevant in terms of quantitative impacts' of the data economy within its broader context.

[3] 1:1.225; see https://www.ofx.com/en-gb/forex-news/historical-exchange-rates/yearly-average-rates/.

Statistic 3

'The value of the U.K. data economy was £216 billion over the years 2012-2017'

***Source*:** 'Data equity, Unlocking the value of big data', CEBR Report for SAS, April 2012

Direct Quote

'Our study has identified £216 billion worth of potential benefits through efficiency, innovation and creation gains, driven by insights unlocked from big data. These economic benefits are dependent on businesses adopting big data analytics solutions to gain a competitive advantage in their industry.'

Takeaway

The £216 billion amount was a **cumulative projection using 2011 economic estimates and projected data analytics adoption rates and does not accurately reflect actual valuations after 2011**. The report assumed that the value of the data economy would come predominantly from the improvements to business efficiency, innovation, and creation that stemmed from more widespread use of data analytics. Given how the data economy has evolved and expanded since 2011, the U.K.'s data economy may very well be valued higher than the projected £216 billion. However, this statistic only reflects outdated and limited projections from 2011 based on a **undisclosed body of academic and industry literature**.

Summary

The CEBR report used 2011 estimates of the business efficiency, innovation, and creation benefits of data analytics, as well as projected adoption rates of big data analytics from 2011 to 2017, to estimate cumulative economic benefits from 2012 through 2017 totaling roughly £216 billion. This amount stems from a simple calculation: the estimated economic benefits of big data analytics in 2011 (totaling

£25.09 billion) divided by the estimated 2011 adoption rate of big data analytics (34%, or 0.34) then multiplied by the sum of the projected adoption rates for years 2012 through 2017 (totaling 293%, or 2.93).

The projected adoption rates for each year are provided below:

- 38% in 2012
- 45% in 2013
- 50% in 2014
- 52% in 2015
- 54% in 2016
- 54% in 2017

Methodology

CEBR developed an economic model based on a review of academic and industry literature to quantify the impact of data equity on the broader economy via three sources: enterprise-level business **efficiency gains** from big data, enterprise-level business **innovation gains** from big data, and enterprise-level business **creation gains** from big data. Estimates of these gains were combined with estimates of the 2011 through 2017 levels of industry-specific adoption rates of big data analytics (based on a review of industry reports from technology experts and an assessment of industry 'potentiality' based on five factors: data intensity, earnings volatility, product differentiation, supply chain complexity, and IT intensity).

CEBR did not disclose the literature reviewed, which informed both the model design and the weighting of various factors used to project future adoption rates. However, U.K.-wide estimates of the big data analytics adoption rates were included, as well as potential drivers of and barriers to the growth of adoption rates.

Statistic 4

'The total amount of global data is expected to grow by 40% year over year for the next decade'

Source: The 2014 IDC Digital Universe study for EMC

Direct Quote

'The digital universe… is growing 40% a year into the next decade, expanding to include not only the increasing number of people and enterprises doing everything online, but also all the "things" - smart devices - connected to the Internet, unleashing a new wave of opportunities for businesses and people around the world.

Takeaway

Without more clarity on the data used and research methodology for the IDC study, **this statistic cannot be verified**. Additionally, the report from which it stems only projects the rate of global data expansion **until 2020**, not through the next decade. Therefore, **the statistic is outdated**.

Summary

IDC's most recent Digital Universe report, authored in 2014, provided an estimate of trends in global data production. It predicted that the 'digital universe', which encompasses all data produced, would (1) double in size every two years through 2020, (2) shift so that emerging markets produced more data than mature markets, (3) be composed increasingly by Internet of Things devices (7% in 2013 to 15% in 2020).

Notably, the IDC report projected that data produced by surveillance technologies would 'wind down' through 2020 as the transition from analog to digital systems was completed. More recent trends in data collection from both governmental and nongovernmental bodies suggest that this projection was false.

Methodology

IDC does not provide a comprehensive overview of their data source(s) and methodology. They do explain (here) that their approach to modeling the 'Digital Universe' involves four steps:

1. 'Develop a forecast for the installed base of any of 40 or so classes of device or application that could capture or create digital information'
2. 'Estimate how many units of information - files, images, songs, minutes of video, calls per capita, packets of information - were created in a year'
3. 'Convert the units of information to megabytes using assumptions about resolutions, compression, and usage.'
4. 'Estimate the number of times a unit of information might be replicated, either to share or store.'

There is no explanation provided for how these estimates are made or what data sources are used.

Statistic 5

'The value of removing fees from all currently available data for re-use is estimated to be £404 million over 10 years'

Source

No source can be verified for this statistic. See below for various sources with similar statistics.

Additional Evidence

1. 'Emerging research indicates that open data provides return on investment of up to 50 times the sums invested in terms of customer benefit. In this specific case, London has gained around £100m direct value by technological investment in London and elsewhere off the back of its data. Through saved customer time saved by journey planning, and more efficient utilisation of their resources, TfL's open data ecosystem has enabled an estimation of 1000 jobs. 175K people are now employed in the digital technology industry in London, in 45K companies with a £30bn annual turnover.' - Transport for London case study, 2018 European Commission Impact Assessment (p. 109)
2. 'Existing macroeconomic studies provide ambitious forecasts about the potential value of Open Data, estimating large gains as a result of Open Data. Estimations vary from EUR 27 bn., to EUR 40 bn. per year, to EUR 59.7 bn. per year[4] in the EU, to even an amount of USD 900 bn. in the EU[5]. The 2015 Creating Value through Open Data report is the most recent and comprehensive

[4] European Commission (2017), The European Data Market Study: Final Report. Available at http://www.datalandscape.eu/study-reports

[5] McKinsey (2013). Open Data: Unlocking innovation and performance with liquid information, McKinsey Global Institute: 116. Available at https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information

study that considers the market size of Open Data at the European level. The study estimated a total market size between EUR 193 bn. and EUR 209 bn. for 2016, with an estimated projection of EUR 265-286 bn. for 2020, including inflation corrections.' - 'Analytical Report 9: The Economic Benefits of Open Data', 2017 European Data Portal Report

3. 'For 2016, the direct market size of Open Data is expected to be 55.3 bn EUR for the EU 28+. Between 2016 and 2020, the market size is expected to increase by 36.9%, to a value of 75.7 bn EUR in 2020. The total market value of Open Data [i.e. both direct and indirect market value] is estimated between 193 bn EUR and 209 bn EUR for 2016 with an estimated projection of 265- 286 bn EUR for 2020, including inflation corrections./ - 'Creating Value Through Open Data, 2015 European Data Portal study
4. 'Deloitte analysis quantifies the direct value of [U.K. public sector information (PSI), i.e. data made available for re-use at low or no cost] at around £1.8bn with wider social and economic benefits taking that up to around £6.8bn' - 'An Independent Review of Public Sector Information', 2013 Shakespeare Review report (p. 21)
5. 'Numerous studies in the last decade have attempted to quantify the economic value generated by the re-use of public sector information. All the studies converge in concluding that the re-use of open data leads to substantial economic and societal benefits. Depending on the methodology used (e.g. the definition of 'open data', the focus on the direct impact on the data economy or wider socio-economic impacts, etc.) various figures have been put forward. The support study for this IA assesses the current direct economic value of open data to be 52 billion EUR a year for the 28 EU Member States[6].' - 'Impact Assessment Accompanying the document Proposal for a Directive of the European Parliament and of the Council on the re-use of public sector information', European Commission (2018)

[6] Impact Assessment Support Study, Deloitte, SMART 2017/0061.

Statistic 6

'Effective use of data can create £66 billion of new business and innovation opportunities in the U.K.'

Source: 'The Digital Catapult and Productivity', 2015 RAND Report

Direct Quote

'Building on evidence that improvements to the more effective sharing of data within and between organisations (i.e. 'data liquidity') can unlock £149bn of organisational efficiencies and £66bn of new business and innovation opportunities in the UK economy, the focus of the Digital Catapult lies in enabling the effective sharing of proprietary data in better and more trusted ways.' (p. 3)

Takeaway

This statistic **erroneously applies projected economic benefits of the adoption of big data analytics to data sharing and 'data liquidity'.** While the numbers provided in the statistic - £149bn and £66bn - can be sourced back to a 2012 CEBR report, they represent **cumulative projections of economic benefits to all UK public and private businesses for the years 2012 through 2017 using 2011 economic estimates and projected data analytics adoption rates. The values here do not accurately reflect actual valuations after 2011.**

As with Statistic 3, this statistic only reflects outdated and limited projections from 2011 based on an **undisclosed body of academic and industry literature**.

Summary

This statistic, used to support the basis for the Digital Catapult report, was derived from the same section of the 2012 CEBR report entitled 'Data equity: Unlocking the

value of big data' from which Statistic 3 is sourced. The CEBR report separated the total economic benefits of big data into benefits derived from increases to **business efficiency**, **business innovation**, and **business creation**. From 2012 through 2017, the cumulative projected economic benefits from improvements to business efficiency were estimated to be £149bn; the cumulative projected benefots from increases in business innovation and business creation were estimated to be £66bn. Importantly, these benefits were the projected results of increased adoption of big data analytics within U.K. private and public sector businesses. They were not the result of increased 'data liquidity'.

Methodology

See methodology for Statistic 3.

Statistic 7

'90% of data were created in the last 2 years'

Source

Several possible sources:

1. 'Big Data - for better or worse', 2013 SINTEF Report
2. 2013 IBM Big Data Overview

Direct Quote

'A full 90% of all the data in the world has been generated over the last two years.'

Takeaway

Without a robust and accessible methodology underpinning this statistic (and with both SINTEF and IBM claiming ownership), it remains **unverified and outdated** (most likely - see below).

One caveat: the statistic *may* be true, both in 2013 and today, due to the exponential growth of data production in recent years. IBM has continued to use this statistic as if it was current as recently as 2017, with one expert quoted as saying, 'by 2016, 90 percent of the world's data had been created in the last 12 months' (p. 11). Despite the likely trend and continued usage, more research will be necessary to verify its accuracy.

Summary

The SINTEC report, drawn from large amounts of social media data, suggests that an increasing amount of data is being produced due to social media and surveillance: either our physical movements are being recorded digitally or our online behaviors and preferences are being tracked. The report came at the start of the 'Big Data' analytic craze and used the statistic to showcase the value of Big Data analytics.

Methodology

This statistic is taken from an analysis of aggregated social media data from the Wisdom Professional consumer analytics tool by MicroStrategy, which was applied to the DelTA-project research study by SINTEF. The data set encompasses aggregated user data from 20 million people as of 22 May 2013, but the MicroStrategy dataset is proprietary and thus its source, quality, and representativeness cannot be independently verified.

The source suggests that its analysis of the Wisdom Professional dataset uncovered the fact that 90% of data had been produced within the two years prior to the report, but offers no methodological foundation for the claim.