CHAPTER 5

Inputs and Outputs: Data Science and the Role of Media

Abstract In order to be able to make use of statistics in measuring, promoting, and guiding wellbeing policies, people must know how to access and interpret statistics, especially official statistics. “Interpretation” includes being able to assess their quality and trustworthiness. Data accessibility and use is not uniform around the world, and this impacts UN plans for sustainability. However, new sources and new kinds of data, such as social media data, administrative data, telecomms data, and unstructured data hold considerable promise—though they come with challenges of their own. A basic level of statistical literacy is necessary to be able to assess the reliability of statistics and to interpret them appropriately.

Keywords Data · Trustworthiness · Quality · Reliability · Data science

5.1 Introduction

This book examines how effectively politicians, policy-makers, and others understand and use statistics in measuring and promoting sustainable wellbeing. To meet that broad aim we need to ask more nuanced questions, including to examine how measures are used and what measures should be used. Beyond that, lies the question of what can be done to facilitate the use of appropriate measures.
An obvious starting point is that, to use official statistics as evidence to support change, people need to know that the numbers are there and be able to access them. A case in point is the website of the UK’s Office for National Statistics which, some years ago, came under criticism as being difficult to navigate, so that even experts could not find the data and statistics they needed. Responding to the criticism, extensive redesign work has vastly improved it, so that facts can be more readily found to steer policy and other decisions.

However, knowing that the numbers are there is but the first step. Numbers can only be expected to guide policy decisions (and to call people to account) if the intrinsic truth in them is acknowledged, and such an acknowledgement should only be expected if the numbers are accurate, valid, timely, and relevant. People might decline to use numbers in the form of official statistics because they deny the validity or relevance of the numbers (arguing that they are incorrect or irrelevant in some way) or, perhaps worse, because they simply ignore the numbers. While lack of correctness or relevance can be remedied, probably in discussion with those who point out the shortcomings, it is more difficult to cope with someone who simply turns a blind eye. (We are reminded of the apparent growth of acceptance of the notion that the Earth is flat, https://theflatearthsociety.org/home/index.php).

The United Nations has long been aware of the need for accurate data and the opportunities provided by the data revolution for supporting sustainable development. UN Global Pulse (Global Pulse 2020) is the UN Secretary General’s initiative on big data and artificial intelligence for development, humanitarian action, and peace. Established in 2010, Global Pulse is a network of laboratories partnering with the private sector and academia to access new data sources and explore how to use them for measuring wellbeing. Global Pulse works in three interconnected areas: discovery, policy, and scale. Discovery explores the application of big data and AI for sustainable development, humanitarian action, and peace. Policy contributes to global efforts to establish trusted frameworks for ethical and privacy-protective data practices and for digital cooperation. Scale provides the UN System and public sector organisations with tools and technical assistance needed for mainstream adoption of data and AI practices.

A World That Counts, a report published in November 2014 by the United Nations Independent Expert Advisory Group on a Data Revolution for Sustainable Development (Data Revolution Group 2014,
INPUTS AND OUTPUTS: DATA SCIENCE AND THE ROLE OF MEDIA

This report, made five key recommendations to the UN Secretary General:

1. **Develop a global consensus on principles and standards**, bringing together public, private, and civil society data providers.
2. **Share data technology and innovation for the common good**.
3. **New resources for capacity development**: provide new resources for supporting the data revolution for sustainable development.
4. **Leadership for coordination and mobilisation**: a global partnership to mobilise and coordinate organisations to make the data revolution serve sustainable development.
5. **Exploit some quick wins on SDG data**: establishing an SDG data lab.

The report commented: “Data needs improving. Despite considerable progress in recent years, whole groups of people are not being counted and important aspects of people’s lives and environmental conditions are still not measured. For people, this can lead to the denial of basic rights, and for the planet, to continued environmental degradation. Too often, existing data remain unused because they are released too late or not at all, not well documented and harmonized, or not available at the level of detail needed for decision-making”.

That was 2014, and while progress has been made, much remains undone. The UN website on *Big Data for Sustainable Development* (UN Big Data 2020), containing useful links to UN reports in this area, comments “Critical data for global, regional and national development policymaking is still lacking. Many governments still do not have access to adequate data on their entire populations. This is particularly true for the poorest and most marginalized, the very people that leaders will need to focus on if they are to achieve zero extreme poverty and zero emissions by 2030, and to ‘leave no one behind’ in the process”. It is clear that there is still a long journey ahead.

From within the data science community, one of the exciting changes is that new types of data, discussed below, are arising all the time: new ways of capturing data, data with new kinds of features, entirely new (unstructured) kinds of data, and so on. From outside, however, these exciting new sources come with associated challenges, not least their short-term nature. Classical official statistics often have time series stretching back
years, or even decades, whereas (for example) social media data might stretch back only a short time (and is vulnerable to changing data capture methods—not least companies going out of business). On the other hand, the different characteristics might be complementary. This has certainly proved to be the case in population studies, where, as we describe below, conventional survey sampling complements modern administrative data, so that the combination yields a powerful synergy.

Other organisations (for example, the Commonwealth Partnership for Technology Management [CPTM 2019] and the Asia Global Institute [Spence 2019]) are also exploring how to take advantage of modern data technology to move beyond conventional economic indicators and develop sound evidence-based (i.e. data-based) tools for assessing well-being and sustainable progress.

5.2 How Can Data Science Help?

Data science is the discipline of extracting value from data. As you might expect from that potted definition, data science is primarily statistics, with significant aspects of computer science (e.g. for manipulating and organising data) along with features which are domain-specific (e.g. different tools might be emphasised in healthcare than in finance, though there is very substantial overlap). “Value” in the definition might be in terms of enhanced understanding of something or improved ability to make decisions or predictions in some domain. For us in this book, primary interest lies in the power of data science for evaluating and assessing the effectiveness of public policies, especially societal progress where the current reported position is that “well-being has, in some respects, improved … Yet progress has been slow, or has even deteriorated in other areas” (OECD 2020, p. 19).

As readers of this book will almost certainly be aware, publicity about the potential of data science has reached new heights in recent years. Couple this with concomitant publicity about associated areas (which overlap to a very great extent with data science, and might legitimately be regarded as subsumed within it), such as machine learning, data mining, and artificial intelligence, and an obvious question is to what extent can these developments assist in monitoring and enhancing progress in sustainable wellbeing. In this section we explore these questions, looking at some of the new kinds of data which have arisen and exploring their properties, and the extent to which they can facilitate our aims.
Discussions of modern data science often focus on the sizes of the data sets and the speed with which new data are acquired. Data sets consisting of billions of points are now common, and data often accumulate in a non-stop stream (hence “streaming data”). Examples are data arising from transactions, such as plastic card financial transactions, mobile phone calls, internet accesses, and so on. These examples illustrate a novel feature which we believe is even more important than the size or speed of acquisition of the data sets. This is the fact that so much data acquisition is automatic. Whereas surveys require an effort on behalf of those being surveyed, credit card, phone call, and web search data are an automatic spin-off of the transactions themselves. No extra effort is required to capture the data. This is what underlies the sizes of many modern data sets—the cost of acquisition is minimal.

Data generated as an intrinsic and necessary part of a transaction (e.g. the amount spent in a credit card purchase, or the destination address of an email) have been supplemented by data deliberately captured by automatic measuring devices. Examples are wrist fitness monitors and odometers in vehicles. Telemetry in general is a rich new source of data—for example, monitoring the ongoing state of an engine, or the way a vehicle is being driven (which can then lead to reduced insurance premiums). But, again, the immediate operational use of such data (is the engine about to break down? Is someone driving erratically?), while perhaps the initial motivation for the measurement, is then supplemented by storing the data in large databases, where retrospective analysis can be undertaken to great effect (e.g. is there a common pattern to engine failures? What type of person is the safest driver?).

In general, assurance must be provided that personal data will be used only for certain purposes—like measuring wellbeing or the effectiveness of policies. Although Article 5, Clause 1(b) of the General Data Protection Regulation says that personal data shall be “collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes” it then goes on to say “further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall … not be considered to be incompatible with the initial purpose”. (Our italics)

Likewise, in the UK “The Statistics and Registration Service Act 2007, as amended by the Digital Economy Act 2017, creates a legal framework providing the Authority with access to data held by Crown bodies, other public authorities and private undertakings (including charities) to
support the Authority’s statistical functions. As amended, the 2007 Act
requires these data suppliers to consult the Authority before changes to
data collection are made in order to protect the continuity of data supply,
as well as the accuracy and reliability of statistics and statistical research
derived from these data sources” (Digital Economy Act 2018, para. 2.2).

Data held by “private undertakings” is potentially a very useful supple-
ment to data captured by official organisations, provided suitable agree-
ments can be reached. So-called “data trusts” are one way in which this
might be done. Precursors to data trusts have long been used in the
retail financial services industry, where data are pooled to yield improved
models for default prediction, with the models being fed back to the
diverse organisations providing the data, and without detailed data being
divulged to competitor organisations. The result is that the industry as
a whole, along with its customers and shareholders, benefits from the
more accurate and reliable predictions. Such strategies are also used in
fraud detection, in recognition of the fact that fraud is an industry-
wide challenge, and that everyone benefits if fraud is minimised (except,
presumably, the fraudsters).

Structured data are still the most common form of data. “Struc-
tured” describes data which come in a clearly specified format—with,
for example, each record containing well-defined slots for data entries
(which does not mean they are all necessarily filled). Increasingly,
however, “unstructured” or “heterogeneous” data are being analysed.
These include things like speech signals, images, video clips, free text,
and so on. It will be obvious that this has the potential to dramatically
increase the sources from which data can be captured, and the perspec-
tives on wellbeing that they imply. Examples which spring to mind are
voice assistants such as Alexa and Siri, surveillance cameras, footfall moni-
toring in retail outlets, as well as web scraping, social media, data from
drones and airships, and so on.

The tremendous richness of the modern data ecosystem will be
apparent from the examples above, so a natural first reaction must be
that these innovations will help the sustainable wellbeing and alternative
to GDP initiative. However, it is important to be aware that new kinds of
data have novel kinds of properties and are not without challenges of their
own. Typically, these challenges are different from, and often comple-
mentary to, the challenges of more conventional data. To illustrate, we
shall compare conventional survey data with modern administrative data.
The importance of administrative data is indicated by the efforts made by the UK Office for National Statistics to move towards a census based on administrative data after 2021 (ONS 2020).

Administrative data, data initially collected for some, typically operational, purpose, and then stored in a database, appear to have various extremely valuable characteristics. For example (and see Hand 2018a, for a detailed discussion):

1. Minimal additional expense is incurred in collecting them; all that is required is to store them in the database.
2. In many situations we might assume that all the data are available: a credit card company will have records of all the transactions undertaken using its card, and a mobile phone company will have records of all calls made. Such data must be generated if the companies are to function.
3. Survival of an organisation might depend on high quality data: incorrect customer charges will not encourage extra custom, and will lead to mistaken company revenue.
4. In many cases the data are as up-to-the-minute as they can be. As soon as a transaction or call is made, it can be recorded in the database.
5. Administrative data tell us what people actually do, reflecting social reality better than survey data saying what people claim to do.

However, closer examination shows that these ideals might not be met. Using the list above:

1. The data generated during an operation might not be in a format suitable for storage and later large-scale analysis, so cleaning and polishing will be required. This will incur a cost. If, further, the data are to be linked to other data from different sources before value can be realised, costs might be incurred in obtaining the additional data.
2. Even if all of an organisation’s transactions are stored in a database, the aim might be to make a statement about a wider population. For example, in the credit card example, we might want to make a statement about all retail purchasers, not merely those who use plastic cards or one particular company’s plastic card, and while tax
records show the details of those who pay, or might pay, tax, they do not reveal details of the black economy.

3. While sampling variability, which is an intrinsic characteristic of survey data, will not apply, administrative data will have other sources of error. Often these induce a systematic bias, which is more difficult to cope with than random variability. An additional complication is that the underlying definitions will (necessarily) be those required for operational purposes (e.g. to run a company, tax office, hospital, school, etc.) which may not match those required for the purposes of building a wellbeing indicator. This can induce distortion.

4. While the organisation collecting the data might have instant access to them, other organisations, such as a government department measuring wellbeing, might have to wait. Unless there are well-established and smoothly running procedures for transferring data, the wait can be a long one.

5. Discovering what people do rather than what they say they do is all very well, but sometimes it is the latter that is needed. This can clearly apply in a wellbeing context, where subjective feelings are relevant.

This example of administrative data serves to illustrate that new data sources (often generally referred to as “alternative data” in some contexts), often have properties complementary to existing sources. That means that using two sources together can yield a very useful synergy. A simple example is the combined use of administrative data, collected on everyone in a population, and survey data, collected on just a small subsample, to infer characteristics of small segments of the population which contain only a relatively few survey cases: tools like regression estimation and Bayesian methods can “steal power” from neighbouring or related segments. Likewise, estimates from entirely different data sources can be used to triangulate and check values: if very different estimates are obtained it suggests that something is wrong somewhere—perhaps in data collection (selection biases creeping in?), or in recording (inaccuracies?), or in calculations (coding errors?), and so on.

A now classic example of the power of alternative data source was the “Billion Prices Project” (Cavallo 2013; Cavallo and Rigobon 2016). This collected the online prices of goods displayed on the web and used them to construct a daily price index—an inflation index—for Brazil, Chile,
Colombia, Venezuela, and Argentina. The effort involved in collecting such data is trivial compared with the effort involved in traditional inflation indexes data collection methods. The well-known result was that while the online price indexes for the first four countries roughly followed the official estimates, that for Argentina was nearly three times the official estimate: the official estimate was misrepresenting the true state of the economy.

Since then, official statistics offices around the world have begun to use data from other sources to estimate inflation and other economic and social indicators (see, for example, Data Science Campus 2019).

Something else which will be apparent from the outline of alternative data sources above is that often data acquisition is more than a merely passive exercise. Often data will be collected dynamically, with its characteristics changing over time or in response to data which have already been collected. At the least this allows effort to focus where it is most needed. Ricciato et al. (2019) describe this sort of change in detail, describing how the term “trusted smart statistics” has been coined to describe the “shift of focus from data sources to data systems and a change of perspective about innovation in official statistics from incremental augmentation towards a systemic paradigm change. The concept of data system is meant to signify an augmentation of the capabilities and role of data source beyond the mere generation of raw input data” (their italics). The concept of trusted smart statistics has been further examined in Vichi and Hand (2019).

Complications can arise when international comparisons are concerned, since different countries have different approaches to different kinds of data, different attitudes to privacy and confidentiality, and different data capture capabilities. We also commented above about challenges arising from the short-term nature of many alternative data streams. Worse still, although a data stream might superficially appear to be internally consistent over an extended period, often this is an illusion. Whereas traditional data capture methods work to strict definitions and collection strategies, many modern data streams, especially those collected as a side-effect of some other activity (like administrative data) are subject to relative arbitrary definitional changes. Consider web scraping, for example. As discussed in Hand (2020, p. 301), Google’s search engine undergoes regular updates to make it more effective. According to Moz (2020, p. 1), “Each year, Google makes hundreds of changes to search. In 2018, they reported an incredible 3234 updates—an average of almost 9 per day, and more than 8 times the number
of updates in 2009. While most of these changes are minor, Google occasionally rolls out a major algorithmic update (such as Panda and Penguin) that affects search results in significant ways”. The challenge of constructing a consistent time series using data collected from such changing methods will be apparent, especially given that the details of what changes have been made may not be known.

In summary, exciting new ways of collecting data are leading to new kinds of data with greater granularity, comprehensiveness, timeliness, relevance, and bulk, with the potential to benefit the sustainable wellbeing initiative. However, these merits come with caveats: new types of data have new kinds of problems. It is still necessary to keep alert for distortions and misrepresentations in the data. Worse, these distortions might be of a kind we are relatively unfamiliar with. Decades of statistical theory means we can comfortably handle sampling variability, but systematic biases induced by hitherto unsuspected shortcomings in the way the data are collected are not so easy to spot. The context is that traditional official statistics have a long history of rigorous development: usually they can be relied upon.

However, there have been dramatic examples where this has gone wrong. We mentioned the case of measuring Argentina’s inflation above. Daniel and Lanata Briones have examined higher-level debate about Argentina’s consumer price index, the INDEC CPI, as played out over 2007–2015, concluding that “The strong criticism of the official index derived from the fact that a matter regarded as purely technical appeared polluted by political interests. This ‘contamination’ undermined the INDEC CPI’s attributed objectivity. The tensions around this index were also translated into the public sphere as a moral problem related to the lack of honesty and transparency from those in charge of elaborating the index” (Daniel and Lanata Briones 2019, p. 136). They also remark that “Inflation can be thought, expressed, defined and quantified in multiple ways. Differences between these forms are not simple, technical details, but have a historical, political and sociological significance. In Argentina, the battle over the CPI was a missed opportunity to discuss the techniques of measuring inflation, as well as its purpose, beyond the walls of a technical agency. Unfortunately, it was also a missed opportunity to deliberate openly about the ways of conceiving, understanding and managing inflation in a country that still suffers from this problem” (p. 145).
The opportunities missed in Argentina are especially pertinent to our study here of new measures of sustainable wellbeing. Without full attention being given to the key characteristics of official statistics—that is to trustworthiness, value, and quality, as described in the UK Statistics Authority’s Code of Practice (UKSA 2018)—there is an increased likelihood that the world will continue to suffer as SDGs set for 2030 will fail to be achieved.

5.3 The Role of the Media

The unique selling point of national statistical institutes (NSIs) is (or at least should be) the quality, reliability, and trustworthiness of their output, created using the best available raw data. However, this output is of little value if it cannot be accessed by those who need it—by governments, commercial organisations, and citizens. In particular for us, of course, measures of wellbeing and sustainability are no use if people cannot see them or if they cannot understand them. Communication is a key step in whether or not such numbers are useful.

In the past, this communication was achieved by (paper) publications from the NSIs, which were then picked up and propagated by various news media such as newspapers, radio, and television. Nowadays, however, we have in addition the web, and all the different channels which ride on that. And therein lies a risk, because not all these channels are equally trustworthy. Indeed, they sometimes serve as a breeding ground for conspiracy theories and misinformation epidemics. That last phrase seems to have been recently coined in the context of anxiety about the coronavirus pandemic, COVID-19. Bizarre, misleading, and potentially harmful stories about the source of the virus and how to protect oneself have spread, including crackpot notions that the virus was a deliberately engineered bioweapon, that one should avoid spicy food, that consuming large amounts of vitamin C is protective, that blowing a hairdryer up one’s nose will kill the virus, that it is caused by 5G telephone masts, and that one should avoid cold food and drinks such as ice cream (see, for example, BBC 2020).

There are many other examples of misinformation spread across the web, including false claims about the dangers of vaccinations, conspiracy theories surrounding the 9/11 terrorist atrocity, and misleading information during election campaigns. It goes without saying that there are
many potential adverse consequences from such misinformation, not least on policy-makers, who might be misled into adopting harmful policies or misunderstanding the real consequences of policies.

The fundamental problem is that anyone can publish statistics, on the web or elsewhere. The web, certainly, has no overall quality control or veracity control. This leads us to the question of how to decide who to trust. It is encouraging that, as we reported in Chapter 4, national statistical institutes (NSIs) can gain broad trust. However, an earlier, cross-Europe study showed large variability in the trust in specific economic indicators between countries.

Given the effort made to ensure that the statistics produced by an NSI are trustworthy, valid, and accurate, this trust seems well placed. It also suggests that there would be fewer misunderstandings and misrepresentations if more people were to go directly to their NSI for the information that they need. However, as already mentioned, the truth is that they typically go to other sources, which may involve partially digested statistics (even if originally from the NSI), or even deliberately distorted material. Some outlets—some newspapers, television channels, and blogs—are well-known for strategic choice of material to promote a position, rather than attempting to give a properly balanced perspective. Others have the same effect accidentally. Hand (2020, pp. 217–219) discusses these distortions in detail.

A step towards a solution to the problem might be to require that assertions are accompanied by a source (Hand, 2018b). This is a standard practice in scientific publications, but there is no reason why it should not be more widely adopted, and it is simplicity itself on the web. Providing a source citation means that people can follow it up to check things if they want to, not that they must. The possibility that an assertion might be checked should serve as a moderator on the more absurd claims. Of course, such a practice alone is not a complete solution—that is probably impossible—but it will help.

There are also higher-level issues. Numbers produced by an NSI, even with the best of intentions, might be in error. They might be subtly misleading because they use a definition which is slightly different from the one you are using. There can be no guarantee that they are right, but we can expect them to be produced by a trustworthy procedure, so that we might legitimately have confidence in them.

We have said little about the deliberate promulgation of false ideas for nefarious purposes—fraud, as one might call it. Such things do sometimes
happen. The case of the official Argentinian price index introduced above would appear to have elements of this, and other cases are described in Oreskes and Conway (2010).

5.4 Statistical Literacy

The question of trust in statistical conclusions, and how this arises from the trustworthiness of the raw input data (whether it is subject to sampling distortions, for example) and the trustworthiness of the statistical tools used to analyse the data (whether they are appropriate to answer the question of interest, for example) is the central one. It has become the focus of considerable interest within the statistical community in the context of the so-called “reproducibility crisis” in science mentioned in Chapter 1. Of course, as far as science is concerned, we must remember that science is necessarily contingent: a scientific assertion may be falsified by further data which contravene it. That is the fundamental nature of science.

Lack of statistical literacy is one of the primary enablers of the misunderstandings consequent on misinformation. An ability to critically assess numbers would reduce the prevalence of such mistakes. However, it is absurdly unrealistic to expect everyone to acquire a sufficient level of statistical expertise to be able to do this. Moreover, that assumes a rationality that is not supported by our increasing understanding of how human cognitive processes operate, particularly when seeking and responding to evidence. Endorsing the seminal work of Kahneman (2011) in this field, Wolf (2019, p. 13) has pointed out situations in which “We do not notice that a number simply cannot be right. And we certainly do not notice (or care) that we are getting only part of the story”.

The best that can be hoped for is to raise awareness of the issues, to encourage people to undertake sense checks and reality checks, and to ask about the provenance of the numbers and the statistics they are presented with, as in the call in Hand (2018b) described above. The case for enhanced education in numeracy is clear. This should explicitly cover official statistics. As Nagaraja (2019, p. 6) has argued, “the public need a greater understanding of how these [official] statistics are put together, what they count, and why”.

An abundance of measures, nominally of the same thing, but in fact using subtly different definitions and underlying ways of being calculated,
can also cause confusion—and can lead to disagreements as different protagonists adopt different statistics. Tackling this sort of issue inevitably requires digging down into complex technical matters. Again, since this is often beyond the expertise of those engaged in the discussion, it poses a real challenge. (The issue is not one unique to measuring wellbeing. The 2008 financial crash was in large part due to a lack of understanding of the limitations of the inevitably complex statistical models underlying risk estimation).

The subtitle of this book is “changing statistics or changing lives?” Statistics are indicators of current states, and change in their values indicates progress, or regress. But, of course, this assumes that the right statistics are being used. There are clear analogies elsewhere. Measurements of height would be of little use in a study of weight-loss diets. Self-reported measurements of calories consumed would be a little better, but would still miss the essential idea. (And there is evidence that that particular measure lacks quality, and is unreliable, [Harper and Hallsworth 2016]). Measurements of BMI or fat fold thickness would be nearer to the point. In the same way, and as we have repeatedly emphasised, if one is interested in sustainable wellbeing, while measures of GDP tap into part of what is needed, they gloss over other key parts and so have the potential to be seriously misleading.

In short, lives will only be changed if those who should make use of the statistics do in fact make use of them, which requires the conditions discussed above to be satisfied: the numbers are accepted as representing the state of society and people are aware that the numbers exist.

References

BBC. (2020). https://www.bbc.co.uk/news/blogs-trending-51271037. Accessed 12 March 2020.

Cavallo, A. (2013). Online and Official Price Indexes: Measuring Argentina’s Inflation. *Journal of Monetary Economics, 60*, 152–165.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives, 30*, 151–178.

CPTM. (2019). http://cptm.org/documents/Draft%20Outline%20Agenda.pdf. Accessed 11 March 2020.

Daniel, C. J., & Lanata Briones, C. T. (2019). Battles Over Numbers: The Case of the Argentine Consumer Price Index (2007–2015). *Economy and Society, 48*(1), 127–151. https://doi.org/10.1080/03085147.2019.1579438.
Data Revolution Group. (2014). *A World That Counts*. https://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf. Accessed 12 March 2020.

Data Science Campus. (2019). https://datasciencecampus.ons.gov.uk/faster-indicators-of-uk-economic-activity/. Accessed 9 April 2020.

Digital Economy Act. (2018). https://www.gov.uk/government/consultations/digital-economy-act-part-5-data-sharing-codes-and-regulations/statistics-statement-of-principles-and-code-of-practice-on-changes-to-data-systems. Accessed 9 April 2020.

Global Pulse. (2020). https://www.unglobalpulse.org/what-we-do/. Accessed 11 March 2020.

Hand, D. J. (2018a). Statistical Challenges of Administrative and Transaction Data. *Journal of the Royal Statistical Society Series A*, 181, 555–605.

Hand, D. J. (2018b, August). Who Told You That?: Data Provenance, False Facts, and Separating the Liars From the Truth-Tellers. *Significance*, 8–9.

Hand, D. J. (2020). *Dark Data: Why What You Don’t Know Matters*. Princeton: Princeton University Press.

Harper, H., & Hallsworth, M. (2016). *Counting Calories: How Under-Reporting Can Explain the Apparent Fall in Calorie Intake*. https://www.bi.team/wp-content/uploads/2016/08/16-07-12-Counting-Calories-Final.pdf. Accessed 12 March 2020.

Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Allen Lane.

Moz. (2020). https://moz.com/google-algorithm-change. Accessed 12 March 2020.

Nagaraja, C. H. (2019, December). Measuring society. *Significance*, 6–7.

OECD. (2020). *How’s Life? 2020: Measuring Well-Being*. OECD Publishing, Paris. https://doi.org/10.1787/9870c393-en Accessed 28 March 2020.

ONS. (2020). https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject. Accessed 12 March 2020.

Oreskes, N., & Conway, E. M. (2010). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York: Bloomsbury Press.

Ricciato, F., Wirthmann, A., Giannakouris, K., Fernando, R., & Skaliotis, M. (2019). Trusted Smart Statistics: Motivations and Principles. *Statistical Journal of the IAOS*, 35, 589–603.

Spence, M. (2019). The “Digital Revolution” of Wellbeing. https://www.project-syndicate.org/commentary/digital-revolution-impact-on-wellbeing-by-michael-spence-2019-06. Accessed 12 March 2020.

UKSA. (2018). *Code of Practice*. https://www.statisticsauthority.gov.uk/wp-content/uploads/2018/02/Code-of-Practice-for-Statistics.pdf. Accessed 17 February 2020.
UN Big Data. (2020). https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html. Accessed 12 March 2020.

Vichi, M., & Hand, D. J. (2019). Trusted Smart Statistics: The Challenge of Extracting Usable Aggregate Information From New Data Sources. *Statistical Journal of the IAOS, 35*, 605–613.

Wolf, A. (2019). Science and Statistical Understanding in the Media. In A. M. Herzberg (Ed.), *Statistics, Science and Public Policy XXI, Environment, Education and the Global Economy* (pp. 9–14). Kingston: Queen’s University.