Giuseppe Sollazzo

Data in public communications

Battistini Lecture
Institute of Advanced Studies, University of Bologna
1 December 2020

Acknowledgments

Cover photo by Jason Coudriet on Unsplash
https://unsplash.com/photos/eQux_nmDew0

The latest version of this lecture is at www.puntofisso.net/battistini.

Contents

Introduction

Let me explain the context of this lecture. The development of my background is somewhat unconventional: I studied computer science, then worked with large data in the healthcare system and setting up computational research services, but by doing so I came close to the issues of data standards, openness, policy, transparency, and how they impact the way Governments and the media use data. The technical nature of my background together with the subsequent multidisciplinary set of experiences in the public sector has two consequences in the way I interpret the topic of this lecture.

First, I view the concept of data in the broadest sense possible: data, in this lecture, is everything that can be stored in a digital format and that can be processed on a computer; text is data, images are data, sound is data, maps are data.

Second, when I speak of *public communications* I'm also looking at a very extensive and inclusive concept. *Public* in this lecture refers to the idea that the intended recipient of the communication channel is as broad as possible and that the producer of the communication is someone working with a broad public in mind.

In this lecture we will see examples that involve both public authorities and the media. The aim is to reflect on the concerns and opportunities for those who work in the public sphere when they opt to choose data. In the course of this lecture, we will discuss some of the milestones in the long march of data into public consciousness; we will visit a few snapshots of data use in public communications, especially with the advent of official national statistics ; we will see a few issues emerged with the pervasive use of data during the recent pandemic; and, hopefully, this will offer a reflection on a few common problems in the way we use data in the world of policymaking and media.

We are in the age of data and we have been for some time. If we chart the frequency over time of the occurrence of the word *data* in literature and politics we get similar results. For example, using the Google Books[1] corpus, or the Hansard record of debates at the House of Commons in the UK Parliament[2] , we see in both cases (Figures 1 and 2) an upward trend during the 20th century. Members of the UK Parliament have been increasingly referring to

[1] Google Books Ngrams Viewer, https://books.google.com/ngrams/graph?content=data&year_start=1800&year_end=2019&corpus=26&smoothing=3&direct_url=t1%3B%2Cdata%3B%2Cc0#t1%3B%2Cdata%3B%2Cc0

[2] Giuseppe Sollazzo, Parli-N-Grams, http://parli-n-grams.puntofisso.net/index.php?ngrams=commons%3Adata

data (although politicians seem to be picking up matters with a 10-year delay, but that's a point for a different lecture). This is just a snapshot, but it represents well the increasing importance of data in the public discourse.

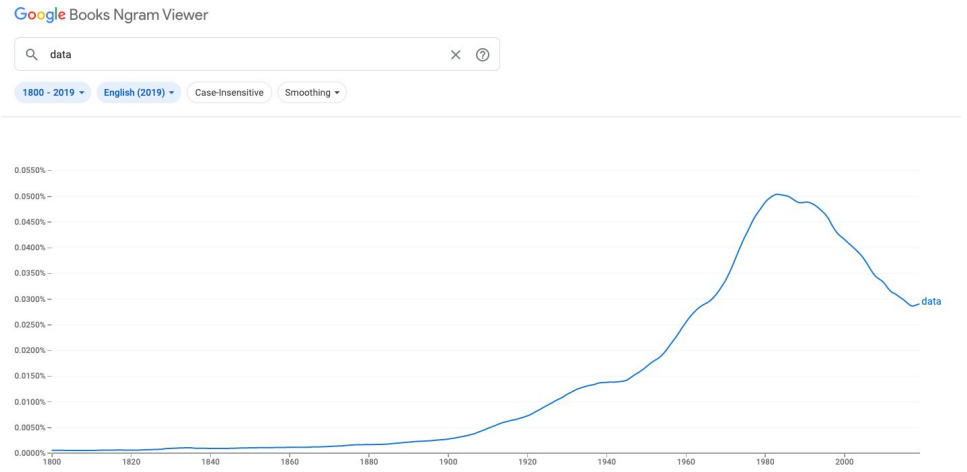

Figure 1

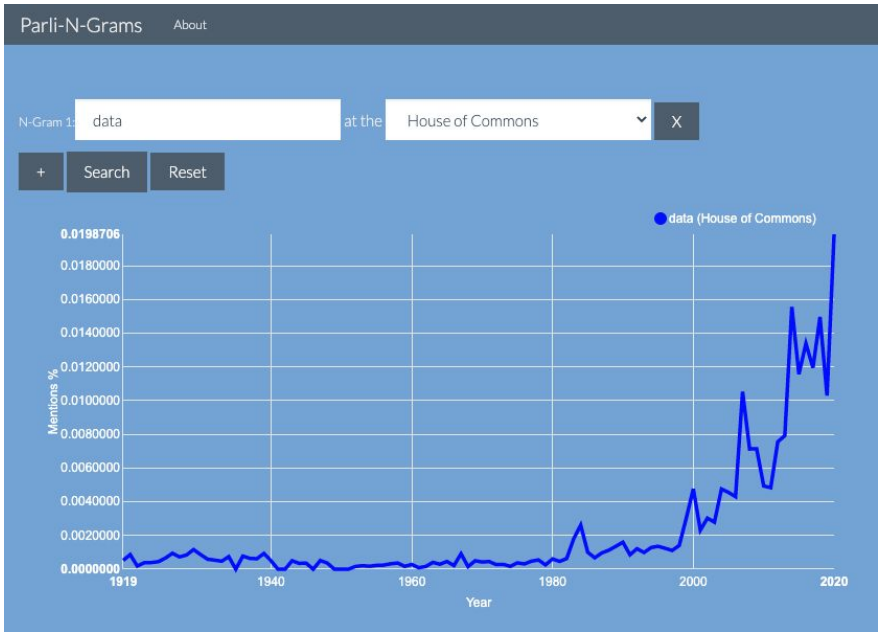

Figure 2

From Cesare Battisti to data definitions

While I was researching some of the main concepts of using data in public communications, I came across the writings of Cesare Battisti[3] (Figure 3). Italians should have heard of Battisti before: countless roads and squares are named after him. Cesare Battisti was an Italian nationalist, citizen of the Austro-Hungarian Empire, who was involved in the Italian annexation of the Trentino region. But Cesare Battisti was also a geographer whose graduation thesis, published in 1898, researched the geography and anthropology of the region[4], which interestingly offers some elements that are of interest to the topic of this lecture.

Firstly, we register the fact that Battisti has a few complaints on the difficulty of finding data for his thesis. He was researching data and national statistics (and let's not forget that *national*, for Battisti, means those of the Austrian Empire). In the preface, he writes: *"I could have done more, especially in the statistical field, if there was not in our country, and in private individuals and moral bodies, such a reluctance to confide in the public domain data, facts and news."* This is an interesting phrase, which captures a few issues we'll touch upon in the course of this lecture: the emergence of national statistics, the idea of Government transparency, and of course the politics of data collection and use.

Figure 3[5]

[3] Maurizio Napolitano, Mappa di Trento 1915 – da un libro di Cesare Battisti, https://medium.com/@napo/mappa-di-trento-1915-da-un-libro-di-cesare-battisti-84935794b1ed

[4] Cesare Battisti, Il Trentino, Giovanni Zippel Editore, 1898. Available on WikiSource: https://it.wikisource.org/wiki/Il_Trentino

[5] Cesare Battisti, Source: Wikipedia, https://en.wikipedia.org/wiki/Cesare_Battisti_(politician)#/media/File:Cesare_Battisti,_Milano,_1915_(portrait).jpg

Secondly, the political nature of Battisti's thesis is particularly fascinating: Battisti uses data and statistics, but not in a neutral way. Battisti's thesis is fundamentally an argument for the annexation of Trentino to Italy, which uses data and statistics to back this political argument. To highlight the political nature of this work, look no further than the first chapter, which opens with the following words: *"Only part of the region within the geographic borders of Italy coincides with the borders of linguistic Italy"*.
Thirdly, the work of Battisti – his thesis, but also his future writing – is full of maps (Figure 4), and maps are intimately related to some of the most intriguing (and, to some, technical) aspects of data. Maps are interesting because they often are, alongside charts, the first way that many people come into contact with data. Maps are often thought of as something entirely graphical, a way to find a route to place, sometimes as beautiful art pieces. But maps are, most fundamentally, graphical representations of data and, as such, they are a good starter to explore data concepts.

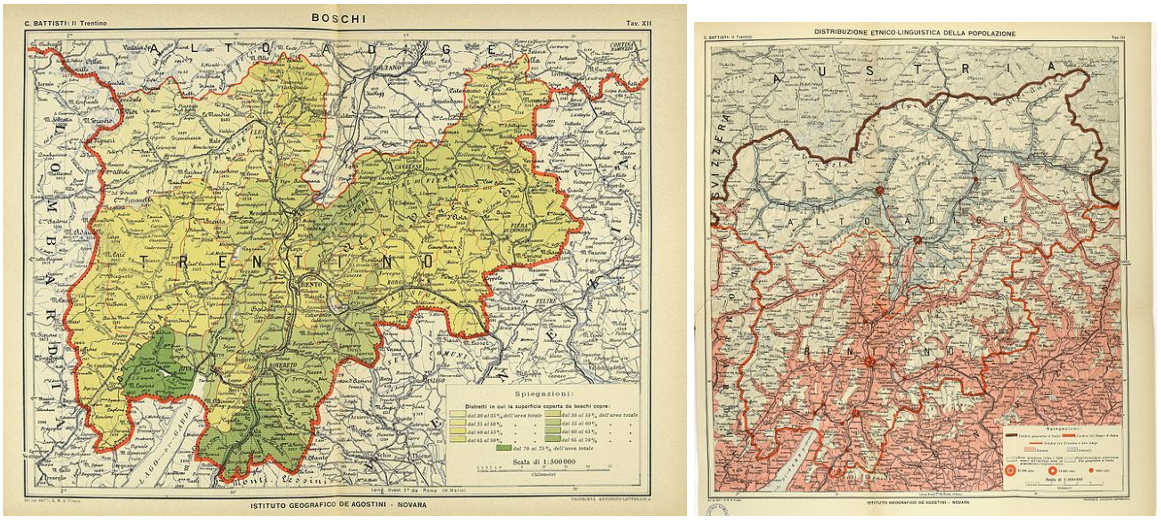

Figure 4[6]

What are these data concepts? The most important concept is that of the entities on the maps, that is the object that are mapped: buildings, streets, points of interest, lakes, rivers, seas. The key message to remember is that some of these entities will change over the course of time. For example, rivers can have their course deviated by a dam. If you look at maps of the city of Trento from before the annexation in comparison with a more recent map (Figure 5), you might note that some of the streets of the 1800s Trento are still in the same place, but they've had their name changed. Therefore, if we want

[6] Illustration from Cesare Battisti, Il Trentino, 1915, https://commons.wikimedia.org/wiki/Category:Illustrations_from_Battisti_-_Il_Trentino,_cenni_geografici,_storici,_economici,_1915

to make a comparison, say, between the 1800s and 2020s Trento, we might ask ourselves how, in data, we can keep track of changing names, boundaries, and paths; how we link in data two entities that we know are related in real life.

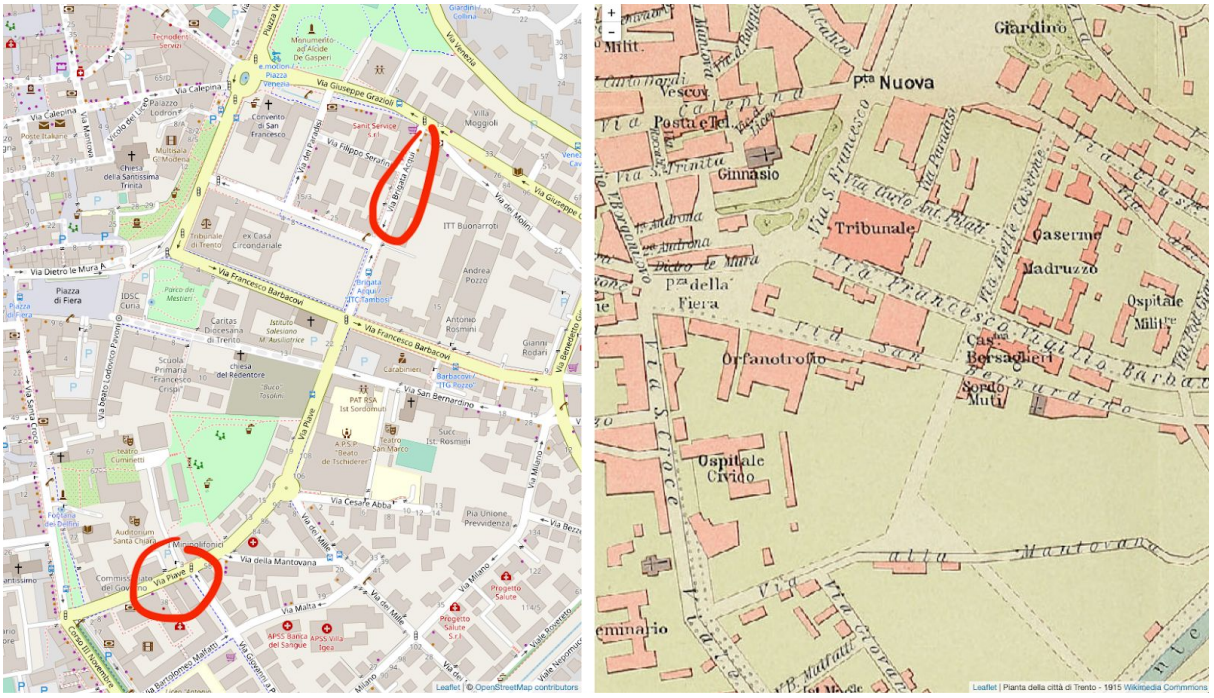

Figure 5[7]

The concepts of entities, shapes, and boundaries, are important if we want to count the evolution of figures pertaining to a certain area over time. For example, we might want to know how each province of Italy has contributed to the national GDP over time. If you look at two maps of the provinces of Italy, one from today and one from 1942, or simply look at the number of provinces in existence in Italy over time (Figure 6), you will see that this is not simple at all, because provinces change over time. Therefore, in order to track how provinces contribute to the national GDP over time, we need to keep track of the changing boundaries of the provinces, to decide how to deal with provinces that were created by carving out territory from larger provinces, provinces that were merged into one, and provinces that ceased to exist. Maurizio Napolitano[8], a researcher at the Bruno Kessler Foundation who works a lot in the space of data advocacy, points out that entities are not always attributed to their "natural" owner. For example, there are towns in

[7] Labmod, Trento nel 1915, http://labmod.org/maps/trento1915/

[8] Maurizio Napolitano, Fondazione Bruno Kessler, https://ict.fbk.eu/people/detail/maurizio-napolitano/

Veneto whose cadastral land register is run by the Province of Trento for historical reasons[9].

Number of provinces [hide]

Year	Provinces
1861	59
1866	68
1870	69
1923	75
1924	76
1927	92
1934	93
1935	94
1941	95
1944	94
1945	93
1947	91
1954	92
1968	93
1970	94
1974	95
1992	103
2001	107
2004	110
2016	107

Figure 6[10]

This brings up a question of ownership: who owns the definition of such entities and has the final word on it? That's a problematic question, if you start to think about the ownership of the definition of borders. Border disputes are more common than one might think. The Mont Blanc Summit is still disputed between France and Italy, and there is a three-way dispute between Germany, Austria, and Switzerland over the exact path of the border that falls inside of Lake Constance. Regional disputes are also known, for example the Marmolada[11] mountain has swapped hands a few times between the Italian regions of Trentino and Veneto, and the current settlement is still debated. Maps need to take borders into account. They are often drawn by the state to reify its territorial claims. Hence, it would not be uncommon to see maps of Mont Blanc that show different borders according to where they were printed. In this respect, maps (and the underlying data), are tools of governance that do not necessarily reflect the reality of the spaces they represent.
The issue of ownership does not pertain only to state governments. OpenStreetMap is a community-led website that allows people to create a map of the world collaboratively, by adding layers of data. It's commonly referred to

[9] Il passaggio all'Italia, Sistema catastale tavolare, Wikipedia, https://it.wikipedia.org/wiki/Sistema_catastale_tavolare#Il_passaggio_all'Italia
[10] Source: Wikipedia 1942, https://en.wikipedia.org/wiki/Provinces_of_Italy#/media/File:Kingdom_of_Italy_1942_with_provinces.svg and 2020, https://it.wikipedia.org/wiki/Province_d%27Italia#/media/File:Italian_regions_provinces.svg
[11] Di chi è la Marmolada?, Il Post, https://www.ilpost.it/2018/09/25/marmolada-veneto/

as “the Wikipedia for maps”. There have been numerous examples of conflicts in this community over the definitions of borders and the naming of places. The most famous of these is the conflict over the official naming of Jerusalem[12] [13]. OpenStreetMap allows multiple names in multiple languages to co-exist. However, there is also one overall “master name”. In Jerusalem’s case, two competing groups fought over the right to assign that master name. The community took a Solomonic approach and opted for not assigning the master name at all until the two groups can reach an agreement. As Mikel Maron, former Head of the OpenStreetMap’s Data Working Group, remarks[14] on their community forum, “Jerusalem is an edge case of everything”. This, of course, is sadly true, but it is helpful to understand that the issue of definitions is not a pure academic exercise. At the time of writing, as you probably expect, the naming of Jerusalem hasn’t had a resolution and it’s not likely to have one soon, but it sheds some light on the concept of the ownership of definitions.
We’ll see that agreed definitions, their authoritative ownership, and the process of collecting data – in two words, data standards – are a very important element in the use of data in public communications and their credibility.

[12] Open Street Map, Forum, Edit War over Jerusalem, https://forum.openstreetmap.org/viewtopic.php?id=13178
[13] Open Street Map, Data Working Group, Disputes, https://wiki.openstreetmap.org/wiki/Data_working_group/Disputes
[14] Mikel Maron, Open Street Map, Forum, Edit War over Jerusalem, https://forum.openstreetmap.org/viewtopic.php?pid=181435#p181435

Data uses in history

Historically, we have some great examples of data and data visualization used for political and social purposes. William Du Bois (Figure 7), an American professor of Sociology, a black man born only a few years after the abolition of slavery, was the primary organizer of a display for the 1900 World's Fair, called "*The Exhibit of American Negroes*"[15]. The Exhibit was a varied collection, widely remembered for its inclusion of a large set of what today we would call *'infographics'* (Figure 8). These infographics had, of course, an agenda. In this case it was a very positive agenda: to demonstrate the progress that had been made socially and economically by African Americans only a few years after the abolition of slavery.

Figure 7[16]

[15] W.E.B du Bois, African American Photographs Assembled for the 1900 Paris Exposition, Library of Congress, http://www.loc.gov/pictures/search/?st=grid&co=anedub

[16] W.E.B. Du Bois in 1918, Source: Wikipedia, https://en.wikipedia.org/wiki/W._E._B._Du_Bois#/media/File:WEB_DuBois_1918.jpg

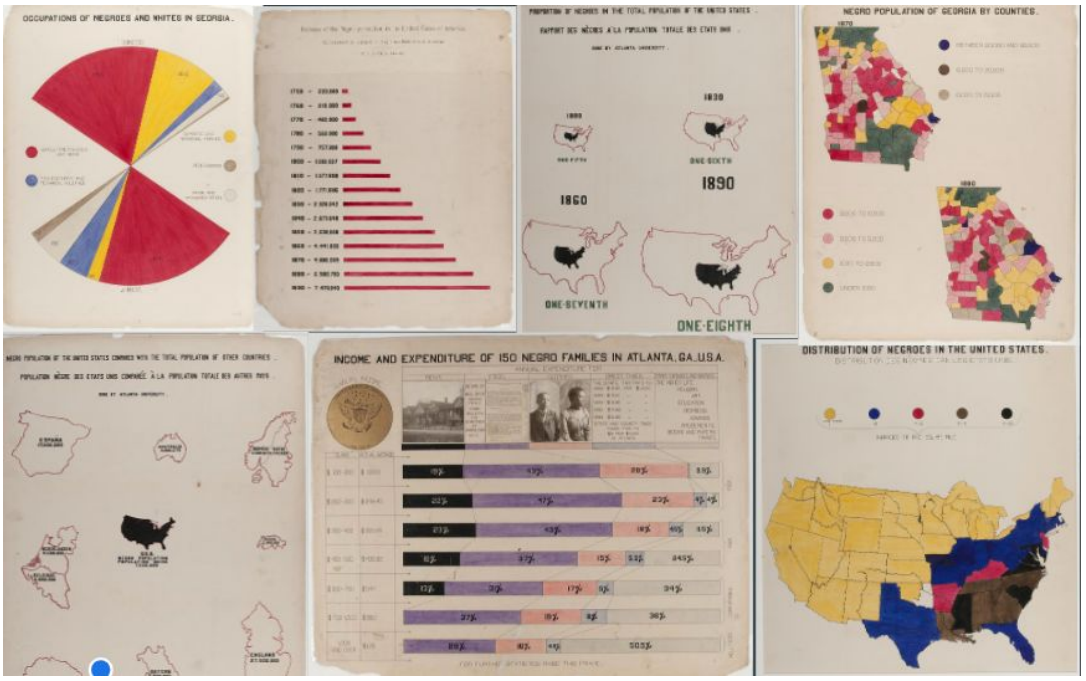

Figure 8[17]

In relation to this exhibition, an interesting point to make is that the US had in fact been running a national census every 10 years since 1790, but this didn't collect a lot of data about African-Americans. The undercounting of African Americans is, sadly, still an existing problem[18], albeit to a lesser extent than in Du Bois' time. As a consequence, du Bois reportedly *employed [...] clerks, students and [... others ...] to assemble and run 'the great machinery of a special census'.* This is clearly another important issue: the existence or lack of data collection can itself be a political action, in this case an act of discrimination against a specific group; and on the other hand, this is an example of data used against discrimination.

The most famous historical example of data collection and analysis is John Snow's map (Figure 9) of cholera-related deaths, which is often hailed as the founding event in the history of epidemiology. In 1854, there was a deadly outbreak of cholera around an area of Central London. John Snow, at the time a physician researching the disease, took a data-driven approach. He interviewed local residents asking details about the deaths in the area, he analysed the patterns of those deaths, and produced this now famous map that suggests a very localised problem, which he found in a dirty water pump. The local council was persuaded to disable the water pump which massively reduced the numbers of deaths.

---

[17] From W.E.B du Bois, African American Photographs Assembled for the 1900 Paris Exposition, Library of Congress, http://www.loc.gov/pictures/search/?st=grid&co=anedub

[18] William O'Hare, Differential Undercounts in the U.S. Census, https://www.springer.com/gp/book/9783030109721

There is a lot of myth around this story. However, the map became a staple for every public health specialist because it just captures how good data, and good data visualization, can be actioned. Snow is remembered for his scientific contribution to asserting the germ theory, but the actionability of data analysis and visualization is also a strong legacy of his.

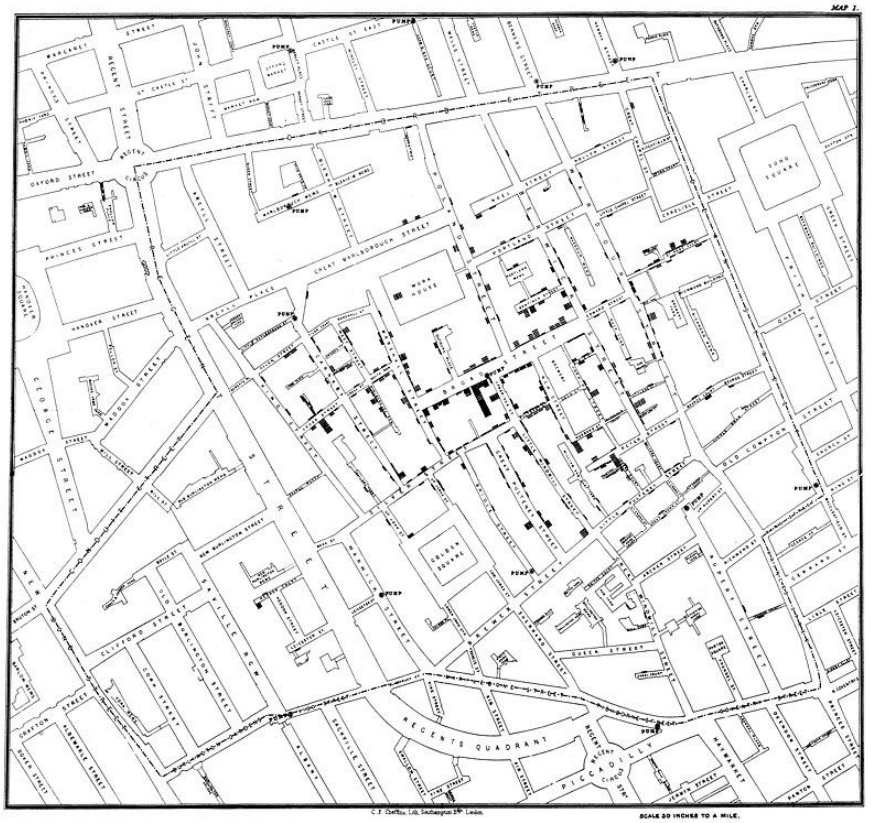

Figure 9[19]

We will discuss the recent COVID-19 pandemic in more detail later. Let me just note that, in comparison with COVID-19, cholera is relatively easy to diagnose with its well defined symptoms. However, there is an interesting similarity to COVID-19: in John Snow's times there wasn't any readily available data source on cholera cases, nor any existing data collection in place, therefore Snow had to create his own data source. He did this by speaking to people, who are, to an extent, an unreliable source of data. Keep these themes in mind as they will be useful later.

[19] John Snow's map, Source: Wikipedia, https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak#/media/File:Snow-cholera-map-1.jpg

The age of official statistics

The age of official national statistics starts in earnest in the 20th century, although we know of censuses dating back thousands of years[20]. Those of us who were brought up in the Western world will be familiar with the events told by the Gospel of Luke. The holy family was travelling from Galilee to Judeah in order to respond to the census run by the Romans, the rulers of that region. Historians suggest this is probably an inaccurate account, but the Romans did indeed run censuses around 2,000 years ago. The ancient Babylonians and Egyptians did too, thousands of years before the Romans. The Han Dynasty in China in 2 AD recorded data about over 57 million people living in over 12 million households. A census in the ancient era was often a way to determine how many taxes or how much property a citizen owned. In the modern era, the US has been running a census every 10 years since 1790, the UK since 1801.
Censuses were a prelude to the establishment of national statistics, which are a characteristic of the 20th century. The Italian ISTAT was created in 1926. The British Central Statistics Office, a precursor of the current Office for National Statistics, was officially set up in 1941. Why were these institutions created? There was an increasing request to answer questions with figures, a consequence of social transformations, the expansion of the electoral franchise, and new economic activities. Policy-making started to have a thirst of evaluation that only data could quench. The birth of official statistics puts data firmly in the public communications arena and, with this, we start to see the first issues in the way data is used in communications.
Let's see for example employment statistics. Employment statistics were among the first ever collected. They are also highly politically sensitive. Employment statistics are connected to the two major points I'd like to make: process and definition. By *process*, I mean the way the data is collected and processed; by *definition*, I mean the way data becomes a translation of a common sense concept.
In most countries, official employment statistics are an estimate. They don't come from a magic box that records every employed and unemployed person; employment statistics come from surveys: there is a panel of people who are repeatedly and frequently interviewed about their employment status. Surveys are based on the idea of sampling, and therefore they have the characteristic of

[20] Census-taking in the ancient world, Office for National Statistics, https://www.ons.gov.uk/census/2011census/howourcensusworks/aboutcensuses/censushistory/censustakingintheancientworld

being representative of a larger population within a certain degree of accuracy. Communicating properly the concepts of accuracy and error in surveys is vitally important.

David Spiegelhalter's recent book "The Art of Statistics"[21] is a great resource, offering examples of bad communication of statistics. I strongly recommend it if you want to explore the issues of communicating data with the public. One of these examples puts the BBC in the spotlight (Figure 10). In 2018, the BBC reported[22] the fall of UK unemployment by 3,000 people in the previous trimester. Except, the official figures come from a survey; and that survey, had a reported margin of error of ± 77,000[23] – which means that in reality there was a range of potential values that ranged from a fall in employment of 80,000 to a rise of 74,000. The BBC journalists interpreted a survey as a firm figure rather than an estimate, producing a potentially misleading headline.

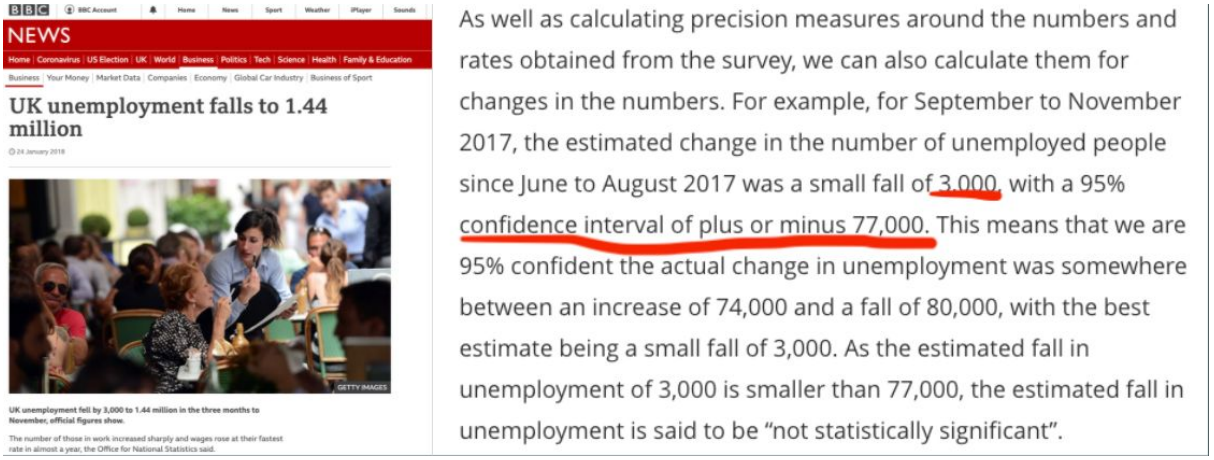

Figure 10

The issue of definitions is my favourite, though. It's a great party conversation starter (well, if you attend the parties I attend...) If you take a random person and ask them if they work, they will certainly have a clear answer. They know if they work or not, whether this is full time, part time, casual work, retirement, or anything else in that spectrum.

However, when we're looking at employment statistics what we are seeing is data about a specific definition of employment, which might be incredibly technical, so technical that it might be hard to relate to it, for people who are not statistically trained. Additionally, that definition can be different between

[21] David Spiegelhalter, The Art of Statistics – Learning with Data, Penguin, 2020, https://www.penguin.co.uk/books/294/294857/learning-from-data/9780241258767.html

[22] BBC News, UK unemployment falls to 1.44 million, https://www.bbc.co.uk/news/business-42802526

[23] ONS, UK labour market: January 2018, https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/uklabourmarket/january2018

different countries, and it changes also within the same country over time. In the UK, the definition of employment comes from the Labour Force Survey[24], which defines an employed person as "*anyone aged 16, or over, who has completed at least one hour of work per week*". This is a technical definition which might not relate to the experience of what meaningful employment looks like.

This definition is not set in stone. For example[25], in December 2019 the ONS and the House of Commons made a change to the way they report the unemployment rate. The rate is usually calculated as follows: it's the total number of people who are unemployed divided by the total population. What total population count was most appropriate? The ONS and the Commons Library decided to change it from *the economically active population aged 16-64* to *the total population aged 16-64*. This is a subtle and relatively minor change, but it shows that definitions can be confusing for those who are not trained in statistics and economics.

To give you an additional angle[26] on the issue of definitions, Figure 11 presents the questions that the US Bureau of Labor Statistics asks its employment panel to determine whether someone is employed. It's a complex set of 11 questions that then get combined into an employment rate. In fact, the outcome of the survey is to produce not one but six alternative unemployment rates[27] with different uses (Figure 12). In Europe, Germany famously used to exclude anyone aged 58 or over from the count on the grounds that they were too close to retirement to be significant[2829]. These differences don't make comparisons between years or countries impossible, but it becomes really important to be aware of a growing number of caveats in order to make a honest, accurate comparison.

[24] ONS, A guide to labour market statistics, https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/aguidetolabourmarketstatistics#unemployment

[25] House of Commons Library, "Understanding statistics on employment, unemployment and earnings" Research Briefing, Published Tuesday, 17 December, 2019, https://commonslibrary.parliament.uk/research-briefings/sn07119/

[26] How the Government Measures Unemployment, Labor Force Statistics from the Current Population Survey, U.S Bureau of Labor Statistics,

[27] Alternative Measures of Labor Underutilization (U-1 through U-6), Labor force characteristics, Labor Force Statistics from the Current Population Survey, U.S. Bureau of Labor Statistics, https://www.bls.gov/cps/lfcharacteristics.htm#altmeasures

[28] Andreas Becker, Debunking the myth of low German unemployment, Deutsche Welle, 2018, https://www.dw.com/en/debunking-the-myth-of-low-german-unemployment/

[29] Viktor Steiner, The labor market for older workers in Germany, Der Arbeitsmarkt für ältere Arbeitnehmer in Deutschland, Journal for Labour Market Research 50, 1–14, Springer Open, https://doi.org/10.1007/s12651-017-0221-9

1. Does anyone in this household have a business or a farm?

2. **Last week**, did you do **any** work for (either) pay (or profit)?
   If the answer to question 1 is "yes" and the answer to question 2 is "no," the next question is:

3. **Last week**, did you do any unpaid work in the family business or farm?
   For those who reply "no" to both questions 2 and 3, the next key questions used to determine employment status are:

4. **Last week**, (in addition to the business) did you have a job, either full or part time? Include any job from which you were temporarily absent.

5. **Last week**, were you on layoff from a job?

6. What was the main reason you were absent from work **last week**?
   For those who respond "yes" to question 5 about being on layoff, the following questions are asked:

7. Has your employer given you a date to return to work?
   If "no," the next question is:

8. Have you been given any indication that you will be recalled to work within the next 6 months?
   If the responses to either question 7 or 8 indicate that the person expects to be recalled from layoff, he or she is counted as unemployed. For those who were reported as having no job or business from which they were absent or on layoff, the next question is:

9. Have you been doing anything to find work during the last 4 weeks?
   For those who say "yes," the next question is:

10. What are all of the things you have done to find work during the last 4 weeks?
    If an active method of looking for work, such as those listed at the beginning of this section, is mentioned, the following question is asked:

11. **Last week**, could you have started a job if one had been offered?
    If there is no reason, except temporary illness, that the person could not take a job, he or she is considered to be not only looking but also available for work and is counted as unemployed.

Figure 11

HOUSEHOLD DATA
Table A-15. Alternative measures of labor underutilization
[Percent]

Measure	Not seasonally adjusted			Seasonally adjusted					
	Oct. 2019	Sept. 2020	Oct. 2020	Oct. 2019	June 2020	July 2020	Aug. 2020	Sept. 2020	Oct. 2020
U-1 Persons unemployed 15 weeks or longer, as a percent of the civilian labor force	1.3	4.5	3.8	1.3	2.1	5.0	5.1	4.6	3.8
U-2 Job losers and persons who completed temporary jobs, as a percent of the civilian labor force	1.4	5.5	4.5	1.6	8.9	8.1	6.4	5.7	4.8
U-3 Total unemployed, as a percent of the civilian labor force (official unemployment rate)	3.3	7.7	6.6	3.6	11.1	10.2	8.4	7.9	6.9
U-4 Total unemployed plus discouraged workers, as a percent of the civilian labor force plus discouraged workers	3.5	8.0	6.9	3.8	11.5	10.6	8.7	8.2	7.2
U-5 Total unemployed, plus discouraged workers, plus all other persons marginally attached to the labor force, as a percent of the civilian labor force plus all persons marginally attached to the labor force	4.1	8.8	7.7	4.3	12.5	11.3	9.6	8.9	8.0
U-6 Total unemployed, plus all persons marginally attached to the labor force, plus total employed part time for economic reasons, as a percent of the civilian labor force plus all persons marginally attached to the labor force	6.5	12.4	11.6	6.9	18.0	16.5	14.2	12.8	12.1

NOTE: Persons marginally attached to the labor force are those who currently are neither working nor looking for work but indicate that they want and are available for a job and have looked for work sometime in the past 12 months. Discouraged workers, a subset of the marginally attached, have given a job-market related reason for not currently looking for work. Persons employed part time for economic reasons are those who want and are available for full-time work but have had to settle for a part-time schedule. Updated population controls are introduced annually with the release of January data.

Figure 12

A general point that we can make here is that the simplification of real life into data points is troubling. Most categorisations elide useful details, and policy making and communication based purely on end point statistics is problematic. There is a lot of nuance in communicating insight based on data that is simply very difficult to provide in all cases (and this obviously conflicts with the urge to make it actionable).

Averages are another example of something with a common sense meaning that differs from its technical definition. The term *average* in mathematics is not a single, unique concept. Already in primary school we learn that there are 3 averages: the mean, the median, and the mode. Pythagoras defined three

ways to calculate the mean: the arithmetic, geometric, and harmonic means, and there are many more. When people speak about an average they commonly refer to the arithmetic mean and that's how they interpret the word *average* whenever they hear it in the media. Except it's not always what it means.

For example, the Office for National Statistics uses the geometric[30] mean for its average house prices series; it uses both the arithmetic mean and the median for the average household income[31]. Rarely, if ever, these concepts are fully explained by the media reporting on the publications of these statistical series. This lack of full disclosure is particularly hard to address, and it makes it difficult for the average listener to discern whether we are talking about an average income (which is a mean) as opposed to the income of the average person (which is a median); or if we're talking about the price of the average house (a median), rather than the average price of a house (a mean, but not an arithmetic mean).

Speaking more broadly, it can be easy to misuse or misunderstand mathematical concepts, especially as the media and politicians themselves sometimes struggle to understand them, or to use them properly. A famous case involved Michael Gove, when he was Secretary of State for Education, speaking about the rating of school. He said that all schools should be rated *good*. Broadly speaking, a school is rated *good* if it is above the national average [32]. Taken mathematically, these two statements are incompatible. Michael Gove appeared before a Parliamentary Select Committee where he was attacked and ridiculed by the Chair of the committee for this statement[33]. The following conversation is on the record:

Chair: [...] if "good" requires pupil performance to exceed the national average, and if all schools must be good, how is this mathematically possible?
Michael Gove: By getting better all the time.
Chair: So it is possible, is it?
Michael Gove: It is possible to get better all the time.
Chair: Were you better at literacy than numeracy, Secretary of State?
Michael Gove: I cannot remember.

---

[30] Office for National Statistics, UK House Price Index: March 2020, https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/housepriceindex/march2020

[31] Office for National Statistics, Average household income, UK: financial year ending 2019, https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/householddisposableincomeandinequality/financialyearending2019

[32] This is not entirely accurate as multiple factors are at stake during an Ofsted inspection. The *result* of an Ofsted evaluation is that a good school will be broadly above average, but it's not a case – nor was a case – of a straight calculation. Once again, this is an example in itself of misuse of data and mathematical definitions.

[33] House of Commons, Oral Evidence taken before the Education Committee, The Responsibilities of the Secretary of State for Education, Tuesday 31 January 2012, https://publications.parliament.uk/pa/cm201012/cmselect/cmeduc/uc1786-i/uc178601.htm

Of course, Michael Gove was not referring to a mathematical definition of average. He was thinking of a quality threshold, set at one point in time and challenging schools to go beyond it over time. But by using a term with a mathematical connotation, he exposed himself to criticism. However, it is even more remarkable that no one on the panel asked him to actually specify his definition of average, which might have clarified his position. On the opposite side of the UK political spectrum, Diane Abbott[34] was heavily criticised when she gave a set of incoherent statistics about police funding, which seemed to show a failure to grasp basic arithmetic operations.

Episodes like these have encouraged two equally damaging behaviours: the media, in the context of rapid news churning created by the social media age, have increasingly started to chase politicians' mathematical failures for their comedic value, rather than focus on rational, pondered, and slower fact-checking[35]. The other damaging behaviour has been the introduction of evasive maneuvers by politicians in order to avoid answering such questions[36] entirely.

The more we look at this subject, the more worrying it gets for those of us working with data to produce public communications. In 2012, the Royal Statistical Society ran a survey among Members of Parliament. 97 MPs out of 650 were asked: "*if you spin a coin twice, what is the probability of getting two heads?*". Only 40% of MPs gave the right answer. The survey also asked if they generally felt confident when dealing with numbers, and over 70% expressed confidence. Probability might not be the easiest topic for those who aren't trained in its quirks, but this is one of the most basic probability problems. How can we expect the world of politics and media to report and use figures accurately if such a simple question has such a poor ratio of correct answers?

One element of reflection about this use and misuse of statistics and other mathematical concepts is a piece of research run by the National Centre for Social Research. They survey[37] the general public on their perceptions and attitudes towards National Statistics. What's remarkable is that although the public has strong confidence in National Statistics, with figures around 90%

---

[34] For example, John Crace, The Guardian, "Diane Abbott has several numbers on police costs – sadly they are all wrong", https://www.theguardian.com/politics/2017/may/02/diane-abbott-has-several-numbers-on-police-costs-sadly-they-are-all-wrong

[35] There are some excellent exceptions, such as the fact-checking efforts started by a number of media outlets, like the BBC Reality Check service, https://www.bbc.co.uk/news/reality_check

[36] Eleanor Busby, The Independent, Schools minister Nick Gibb refuses to answer maths question on TV as he launches times tables tests, https://www.independent.co.uk/news/education/education-news/nick-gibb-maths-questions-tv-gmb-refuse-answer-times-tables-test-primary-schools-launch-a8210281.html

[37] NatCen, Public Public Confidence in Official Statistics 2018, https://natcen.ac.uk/media/1714099/Public-confidence-in-official-statistics-%E2%80%93-technical-report-2018.pdf

percent, only a minority thinks that the Government presents official figures honestly, 45%, and a smaller share, 17%, thinks that newspapers present official figures honestly.

The attitude of politics and media towards the use of numbers in the political debate is not helpful. Over the past 20 years an expectation has developed that leaders give yes/no answers. They are expected to quote from memory data to back those binary answers. There are two issues with this. The first issue is that of *data-washing*, which is the increasing use of data in a pseudoscientific way to back claims that cannot really be backed by data; the second is that often there is much uncertainty to be conveyed, and the political discourse hasn't yet learned to cope well with uncertainty. Reality is complex.

The drive in the media for clear answers from politicians is a reaction to years in which politicians gave elusive statements and evaded scrutiny; but as all reactions, it has produced a generation of politicians and journalists that doesn't cope well with doubt. Politicians are now intimately convinced that answering "I don't know" is a bad thing, and journalists won't allow any kind of doubtful answer without ridiculing their guest.

This problem goes beyond data. We all have fun when Jeremy Paxman or Andrew Neil push politicians to answer difficult questions about complex problems; the problem with difficult questions and complex problems is that they need equally complex answers. The introduction of data in the public sphere has made this problem worse because there is now a way to get out of those difficult questions by misquoting data. We hear that *"The data shows"*, "*the data says*". But data is not a crystal ball and should not be used that way.

Journalism brings data into the mainstream

An important development in recent years has been that of data journalism, which is the use of data to tell journalistic stories, sometimes through the adoption of interactive applications. Data journalism has brought data into the media mainstream. Today's broadsheet newspapers present data-driven stories on their front pages. Data journalism is broadly a very positive development. It is not entirely without controversy. Journalistic stories usually need an angle. Remember what I said at the beginning: data is rarely neutral.

For example, in 2018 the Financial Times[38] represented the results of the Italian Elections by making a map (Figure 13) that shows the party or coalition that got the most votes in each constituency. The result shows a country divided in three, which was the line taken by the journalist (the article's title was "*Italian election shines harsh light on economic divide*"). This is obviously a captivating story and it does somehow reflect on other cultural, social, and economic aspects of Italy.

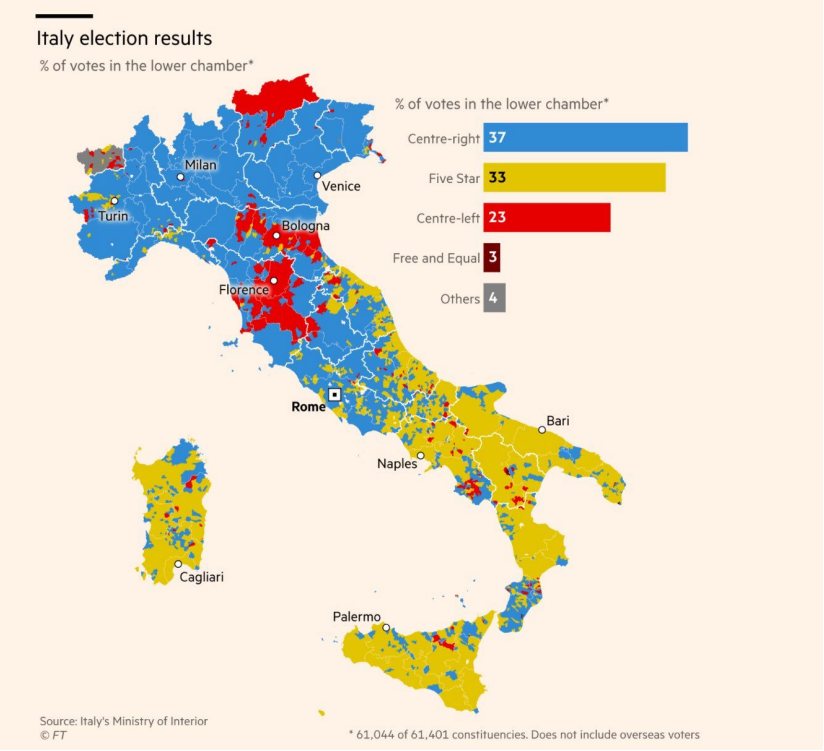

Figure 13

[38] Valentina Romei, Financial Times, Italian election shines harsh light on economic divide, https://www.ft.com/content/d11902f6-2062-11e8-a895-1ba1f72c2c11?segmentId=6132a895-e068-7ddc-4cec-a1abfa5c8378

But let's not forget: telling this story is a choice. I produced a couple of dot maps to make a counter-argument[39] to this story (Figure 14). Dot maps don't show the winner in a region; they simply show a dot for every 10,000 votes each party or coalition gets. These are obviously rather ugly maps, entirely driven by my lack of aesthetic taste. However, what becomes apparent is that I'm telling a different story while using the same data. The story I'm choosing to tell is that Italy wasn't as divided as it was confused.
Neither of these two stories is the ultimate truth. Maps and data don't lie: they tell a story the way we choose to present by asserting an editorial choice. Note that I'm not criticising data journalists for doing this: making editorial choices is their job. What we need to remember is to be aware of this phenomenon; and be aware that "*a map is not the territory*"[40].

Figure 14

On a positive note, many in data journalism have taken up the idea of replicable journalism, taken from academic research. There are some great examples of it. Recently, journalists at the Washington Post[41], created a

[39] Giuseppe Sollazzo, 8 March 2018, Le vere mappe delle elezioni italiane, https://medium.com/@puntofisso/le-vere-mappe-delle-elezione-italiane-a0cb89d27d9e
[40] Map-territory relation, Wikipedia, https://en.wikipedia.org/wiki/Map%E2%80%93territory_relation#%22A_map_is_not_the_territory%22
[41] How The Washington Post Estimates Outstanding Votes for the 2020 Presidential Election, Washington Post Engineering, October 2020,

narrative based on data models, and they released their source code; they offered a technical write-up with citations, credits, and release notes, exactly as if this was an academic research paper. This degree of honesty and transparency can help build trust, in the assumption that the reader base is educated to a level that enables them to understand it. This means that we are at a crossroads: public communications of data must be able to balance the need for transparency and honesty with the need to be accessible to a public who's not (yet) fully able to appreciate very difficult concepts that cannot be simplified.

The other big shift in journalism has been the increase in predictive journalism. We have seen it in many shapes, ranging from sports to electoral predictions – a very hot topic with the recent US Election – and we've seen models about hospital beds during the COVID-19 pandemic. Predictions are always dangerous because they... can be wrong, it is in their nature. This move towards predictive data journalism is intriguing and exciting, although it brings us a philosophical reflection about the use of data in journalism: is the job of journalism to tell stories that describe the world as it is, based on data, or is it to use that data to tell what the future holds? In fact, it can be a bit of both; the balance is a philosophical quest as much as it is a market-driven choice.

Mona Chalabi, who's the Data Editor for the US edition of the Guardian, has written and spoken extensively on this topic. In a recent interview[42] in which she spoke about the allegedly wrong predictions of the US Election 2016, she asks a straight question: *"Since when is it our job to predict?"*. That's a very good question, although the boundary between data-driven storytelling and data-driven prediction is full of grey areas.

Moving slightly sideways, I believe that data journalism has been broadly positive in addressing one of the major problems in the use of data in public communication: how to convey the concept of *uncertainty*.

For example, in the 2020 US Presidential Election, The Guardian[43] built an interactive article that allowed readers to play with the different routes that Joe Biden and Donald Trump could take to winning the election; the Financial Times[44] had something similar; the New Statesman allowed readers to re-run

---

https://washpost.engineering/2020/10/22/how-the-washington-post-estimates-outstanding-votes-for-the-2020-presidential-election/

[42] Today In Focus, Guardian Podcast, US election 2020: can we trust the polls?, https://www.theguardian.com/news/audio/2020/oct/22/us-election-2020-can-we-trust-the-polls-podcast

[43] Helena Robertson, Ashley Kirk and Frank Hulley-Jones, The Guardian, Build your own US election result: plot a Biden or Trump win, https://www.theguardian.com/us-news/ng-interactive/2020/oct/30/build-your-own-us-election-result-plot-a-win-for-biden-or-trump

[44] Biden vs Trump: live results 2020, Financial Times, https://ig.ft.com/us-election-2020/#calculator

their model[45] and get new results; FiveThirtyEight, the blog of statistician Nate Silver who rose to fame for correctly predicting 49 out of 50 states in the 2008 election, offered a way to see the prediction being recalculated[46] in real-time whenever the reader assigned a state to one of the two candidates.

Another clever way to approach uncertainty in data-driven journalism is to encourage the reader to reflect on their own lack of knowledge. The German newspaper Berliner Morgenpost has repeatedly taken this approach to invite their readers to reflect on the consequences of the reunification of the two German states. In one example[47], the Morgenpost invited readers to draw the path of the Berlin Wall on a map, then revealed what other readers thought together with the real path. In another occasion they allowed[48] the reader to cut Germany in two, revealing, using official statistics, how the two sides of the cut compare.

These are powerful examples: by using data-driven interactions, they allow the reader to explore the uncertainty that is present in the data and to understand their lack of knowledge about a topic. The reader can learn something new, or have a Eureka moment about what they don't know. Let's not forget, however, that *visualization is not always the most appropriate tool*[49] and that sometimes there is a degree of uncertainty that is impossible to properly capture in a graphic. The ultimate necessity for public data communicators in our age is to make the public appreciate that data can carry uncertainty and trigger questions, instead of seeing it as the crystal ball that provides certain answers.

[45] Ben Walker, New Statesman, US 2020 presidential election forecast model: will Donald Trump or Joe Biden win? https://www.newstatesman.com/international/2020/11/us-2020-presidential-election-forecast-model-will-donald-trump-or-joe-biden

[46] FiveThirtyEight, https://projects.fivethirtyeight.com/trump-biden-election-map/

[47] Wissen Sie noch, wo die Mauer Berlin teilte?, Berliner Morgenpost, https://interaktiv.morgenpost.de/berliner-mauer/

[48] Ost-West? Nord-Süd? Oder ganz anders?, https://interaktiv.morgenpost.de/deutschland-teilen-deutsche-einheit-wiedervereinigung/

[49] Levontin, P., Walton, J.L., Kleineberg, J., Barons, M., French, S., Aufegger, L., McBride, M., Smith, J.Q., Barons, E., and Houssineau, J., Visualising Uncertainty: A Short Introduction (London, UK: AU4DM, 2020), https://spiral.imperial.ac.uk/bitstream/10044/1/80424/2/VUI_221219.pdf

Data in the COVID-19 pandemic

Some of the themes we have explored so far have become evident, and in some cases amplified, during the COVID-19 pandemic. The pandemic saw the emergence of data as a matter of daily discussion by politicians and journalists. The most important question emerging from the pandemic has always been *what kind data is most useful to monitor and manage it*. A characteristic of the debate on data was a ceaseless series of arguments as to what available data was the most useful to understand and deal with the pandemic. Let's name a few datasets and the caveats attached to them. The number of positive patients was one of the first to be used; however, the number of positives depends on the number of tests being administered. The number of hospital admissions is important to make sure that the health system isn't overwhelmed; this, however, carries a lot of uncertainty because patients might be admitted with a variety of other conditions, so its usefulness depends on *how* the data is collected and reported. The ratio of available/occupied beds is also a rather uncertain variable, because when under pressure hospitals will commonly squeeze in a few more beds: in Italy, for example, there is a habit of having a set of so-called *sub-intensive* therapy beds that can be transitioned into full intensive therapy upon need[50]. Think also about the controversy in Italy that happened around counting beds in intensive therapy units[51] in Calabria: the local authority decided to exclude assisted ventilation patients from the total count of intensive therapy patients, allegedly as a way to be classified as a lower-risk area. In addition, bed availability data is hardly readily available in any easy-to-consume format. Finally, reporting death figures is useful to understand how serious the pandemic is affecting patients, especially in comparison with longer-term death rates; but we will see in a moment how fuzzy death can be. All these different datasets can be helpful. The real area of enquiry is understanding what they mean, how accurate they are, and how actionable they are.

Death offers us another dig into the problem of definitions. You can probably see this coming now: how we define death has an impact on how we account for it. Let's start with a local example. In August 2020, Public Health England

[50] Isaia Invernizzi, Come leggere i dati sulle terapie intensive, https://www.ilpost.it/2020/11/03/dati-terapie-intensive-coronavirus/

[51] Alessia Candito, Il trucco con cui la Calabria è passata da 26 a 10 ricoveri in terapia intensiva, La Repubblica, 5 November 2020, https://www.repubblica.it/cronaca/2020/11/05/news/cosi_la_calabria_e_passata_in_poche_ore_da_26_a_10_letti_occupati_in_terapia_intensiva-273201335/

changed its definition of COVID-19-related death[52] (Figure 15). They decided to extend the period between a positive test and a recorded death. They basically started to include deaths that happened a bit later on than what they had been doing before. This matters because there might not be a way to clearly attribute a death to COVID. There's no smoking gun suggesting that "COVID did it". This expanded definition might capture more COVID-related deaths but it might also capture deaths of patients diagnosed with COVID whose death was not quite a major consequence of COVID. In this process, there is no right or wrong: once again, it is all a matter of understanding how definitions of common sense concepts inform our understanding of the pandemic. The press famously started to use the phrase "dying *with* COVID" as opposed to "dying *from* COVID", as a way to capture the uncertainty of the cause of death. But other problems might be impacting death counts; for example, at the beginning of the pandemic the counts missed a lot of deaths happening at home or in care homes. It is important to be aware that processes are set up for a reason, and that sometimes what we might perceive as a straightforward question, such as *was this death caused by COVID-19?*, might not have yes/no answer.

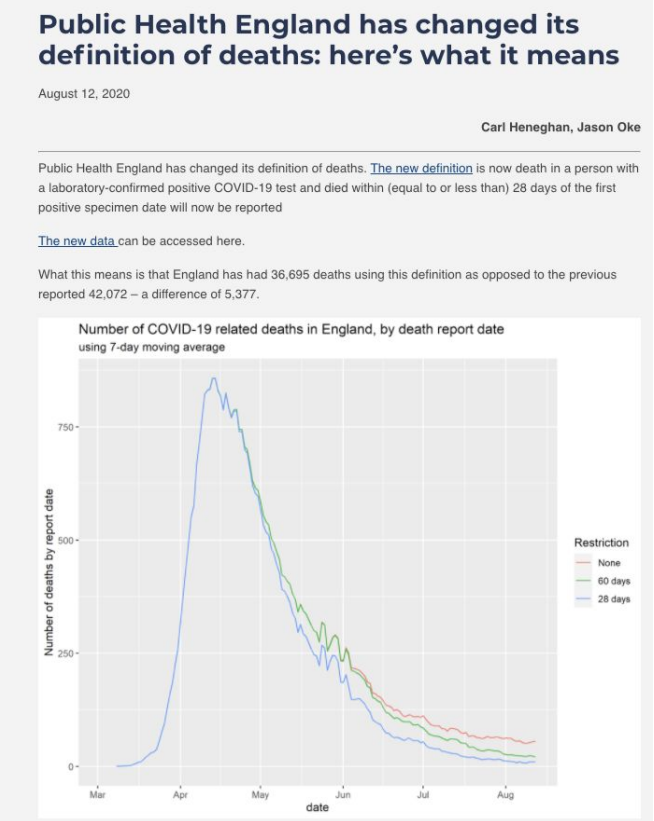

Figure 15

[52] Public Health England has changed its definition of deaths: here's what it means, https://www.cebm.net/covid-19/public-health-england-death-data-revised/

The problem of defining death, though, is particularly intriguing. I'm sorry if this sounds morbid or overly philosophical. As you can imagine, the issue of defining death is not limited to COVID-19. Another book I recommend is "*How to Make the World Add Up*" by the Financial Times' economist Tim Harford[53], which mentions a rather sobering example from a few years ago. In the UK there was a case where mortality rates for newborn babies varied massively between different parts of the country. Subsequent research also showed significant differences between different countries of Europe[54][55]. This would of course be cause for a lot of concern. But it turned out that there wasn't necessarily a difference in mortality between different areas: there was, instead, a level of uncertainty in the definitions that caused a recording of such events that followed different informal standards. Let's see why. Normally, when a pregnancy ends early, let's say long before the 23rd week, this event is recorded as a miscarriage. On a more cheerful note, when a baby is born prematurely, from the 24th week onwards regulations mandate that doctors record this event as a live birth. It turns out that when pregnancies ended somewhat early but close enough to that 24th week mark, the situation was fuzzy enough that the data recording would be largely discretionary. Some doctors would record these events as two separate events: a live birth and a subsequent death; others would mark them as a single miscarriage event. This difference was caused by a variety of largely cultural considerations.

What matters to us is that an event that we feel as definite as death is not as clear when we are translating it into data; therefore, we create definitions to represent it. These definitions might appear entirely arbitrary. Of course, definitions can change in response to a number of developments, and we should always explore if the definitions we use are culturally appropriate, and if they work in their normative context. Some definitions might have ethical implications; culture and ethics are not immutable, therefore we need to approach data collection with care.

Let me add a personal example. A few years ago I helped a group of clinical statisticians carry out a differential geographic analysis[56] of survival rates for a highly specialised elective surgical procedure (*elective open supra-renal aneurysm repair*, if you're curious). We looked at how mortality for this type of

[53] Tim Harford, How to make the world add up, https://timharford.com/books/worldaddup/

[54] Jon Sharman, Tragedy of stillbirths in Europe underestimated due to statistical guidelines, study suggests, The Independent, https://www.independent.co.uk/news/health/stillbirth-rates-europe-underestimated-official-statistics-new-study-who-a8558286.html

[55] Lucy Smith et al., Quantifying the burden of stillbirths before 28 weeks of completed gestational age in high-income countries: a population-based study of 19 European countries, The Lancet, https://doi.org/10.1016/S0140-6736(18)31651-9

[56] Alan Karthikesalingam et al., Elective Open Suprarenal Aneurysm Repair in England from 2000 to 2010 an Observational Study of Hospital Episode Statistics, PLOS One, 2013, https://doi.org/10.1371/journal.pone.0064163

operation changed across the country producing a map (Figure 16) representing survival rate at Strategic Health Authority level, which is a regional-level of health administration in the UK. At a basic level this map tells us a simple story: people die to a different extent when this highly specialised type of procedure is carried out in different parts of the country. But if we start taking a closer look at the map, we might want to probe further: why are the regions defined as they are? How exactly did people die after the surgery? How closely linked was their death to the procedure? How long after the operation did the death occur? How many operations are run in each area, are these rates significant? We did address some of these questions in the paper, but you can see that answering them requires discussing definitions and an in-depth conversation that is hard to capture on a map.

In order to take action with data, there is a need to understand the process of data collection: how it operates, for what purpose it was established, for how long it stores data, and so on. It is important to keep an eye open to wrong assumptions. It's very easy to make wrong assumptions.

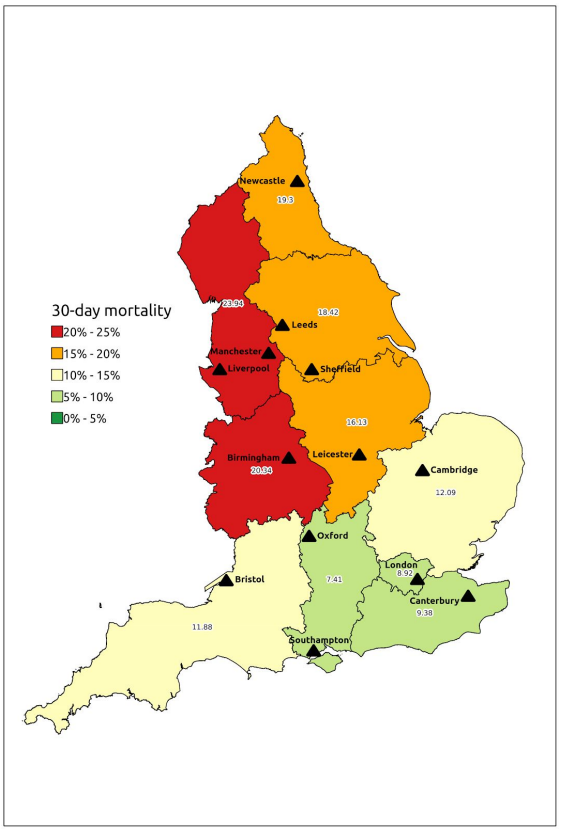

Figure 16

Take the issue of *where* to attribute the record of a COVID infection. This became an apparent problem in the UK when the figures of students who had

been tested positive were incorrectly attributed[57] to their parents' address rather than to the location where they would normally live during term time (in many cases, their hall of residence). The data collection asked for a home address rather than the current or ordinary location of residence. This little error resulted in inflated infection rates in the borough of Richmond, a relatively wealthy suburb of London, where families are likely to have university-age children who study away from home. "Fixing" this problem requires an understanding of the context in which data collection and reporting happens. To communicate data effectively, analysts should always work together with situational experts.

The pandemic also put data journalists' data representation choices under the spotlight. John Burn-Murdoch, a journalist at the Financial Times, and a widely followed authority on the journalistic use of data, spent a lot of time on video and social media to engage with the public in order to explain the newspaper's choices, defend them, and sometimes change the way the article represented data. I think this is something really positive in terms of building trust. Some of these discussions were about themes that are now familiar: should we show absolute counts of cases or figures per capita adjusted for population size[58]? Should we display the chart on a linear or a logarithmic scale[59]? Is a rolling average a better indicator of trends as opposed to a spline[60]? John and other journalists who took this path of engaging with the public should be praised, because they tried to make it clear that what we get from data is not set in stone, and that sometimes at different stages of the pandemic different choices in data visualization are necessary in order to convey the real message, for example around the seriousness of the infection.

[57] Nicholas Cecil, London's coronavirus figures 'skewed by university students in other cities' amid fears capital is 'two weeks behind the North West', Evening Standard, 9 October 2020, https://www.standard.co.uk/news/health/london-covid-stats-skewed-university-students-a4567441.html

[58] John Burn-Murdoch, Twitter, https://twitter.com/jburnmurdoch/status/1248731769798623235

[59] John Burn-Murdoch, Financial Times, https://www.ft.com/video/9a72a9d4-8db1-4615-8333-4b73ae3ddff8

[60] John Burn-Murdoch, Twitter, https://twitter.com/jburnmurdoch/status/1247669865982427136

Trust and engagement in data communications

Data collection can be controversial and political. We all appreciate the need for accurate data reporting about a pandemic, because data enables us to monitor the spread, to take action, and to mitigate the risks and the impact of the virus. Despite the many issues we've discussed, without data we wouldn't be able to do anything about the pandemic and we would be in the same situation as our ancestors dealing with the bubonic plague.

Data collection can be relatively uncontentious in the context of a global pandemic. In the UK the COPI regulations[61] enable data collection at a time of a health crisis. But data collection is not without problems if seen more generally, beyond the protecting context of a crisis. Quickly setting up data collection for tracking and tracing patients in the context of an emergency enjoys relatively good support. We should make two reflections. The first is that emergency situations aren't the black swan events we sometimes make them out to be, and therefore we should be ready and prepared to initiate new data collections when an emergency strikes. The second is that, in reality, in a very fast-paced situation there might be little time to set up new data processes. Therefore many decisions related to data collection must be taken "at peacetime", when most people don't see the immediate benefit of data collection (or might even actively oppose it).

A few years ago, the then Health and Social Care Information Centre, announced a data collection programme called care.data. The idea behind care.data was brilliant: getting family doctors to automatically extract data about all their patients, and transfer them into a central facility, in order to power medical research. Two acts of parliament provided legal backing to this programme, which was set up to allow patients to opt out of it. care.data[62] turned out to be so controversial that it was shut down just three years later.

What happened? The programme had somehow mishandled the question of trust. Patients' rights organisations objected that there simply hadn't been enough effort to raise patient awareness, despite the distribution of information leaflets. Doctors started to complain that they didn't know the process to let patients opt out. Privacy activists raised the alarm that the programme wasn't clear about data sharing to third parties, and media stories started to emerge on

[61] Control of Patient Information, https://www.legislation.gov.uk/uksi/2002/1438/contents/made

[62] Care.data, NHS Board Updates, July 2014, https://www.england.nhs.uk/wp-content/uploads/2019/07/04-care-data.pdf

the potential misuse by insurance companies. The criticism escalated and the programme was ultimately abandoned.

Care.data was the kind of data collection programme that could be helpful to prevent, monitor, understand, and take action in a future health crisis. But we cannot do that without really addressing the issue of trust, and proper engagement to communicate the benefits. The public needs to travel the journey into data along public authorities and the media.

There is a lot that could be said about the issues of engagement and trust and we don't have time here, but I wanted to touch upon them as they will be increasingly important in the next few years. Let me just say that the concepts of ethics and trust are now increasingly present in policy-making. The UK Government, for example, has set up, a few years ago, a Centre for Data Ethics and Innovation with the mandate to research and adopt a more solid ethical approach for the use of data across government.

The age of open data

We've seen data-driven metrics in the political discourse used to back policies and to monitor the delivery of Government services. The final chapter of this lecture is about data graduating and becoming a first-class citizen and a policy objective in its own right. This happened with the advent of the Open Data agenda.

What is Open Data? There are many competing definitions. According to a commonly recognised definition by the Open Knowledge Foundation, *Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike*[63]. In general, this definition is interpreted as to guarantee that the data is somehow in a format that can be automatically processed by a computer rather than, for example, an image or a PDF file. The definition suggests that there should not be any imposition of limitations on the way data can be used, although in practice this is interpreted in many different ways. *Openness* refers to the fact that data should be interoperable, a technical word that means that it should be usable on different computer platforms and linkable to different datasets.

How do we go from something this technical into something that has to do with public communications? We must thank a natural disaster. In 2005, Hurricane Katrina killed almost 2,000 people and caused $125B of damage in the US, particularly in Louisiana. The Government struggled to react and, as a response to this lack of intervention, isolated citizens started to organise in order to help their neighbours. To better direct their help, they put together datasets of population statistics and other useful information, by copying the information off Government websites.

Governments ultimately loved this idea. It makes a good story: data can help people, it can even save lives. Governments can become really popular, that was the thinking, if they enable all of this by making data available.

How strongly data entered the Zeitgeist is highlighted by the "armchair auditors" statement by David Cameron, then the Leader of the Opposition. In 2009, there was an expenses scandal at the House of Commons, and the Government was considering the release of the MPs' expense claims. Cameron connected these events to the growing interest in open data and said "*Just imagine the effect that an army of armchair auditors is going to have on those*

[63] Open Data Definition, https://opendefinition.org/

expense claims."[64] Armchair auditors would be ordinary people taking the data made available by the Government, analysing it, holding the powerful to account, and making public authorities more efficient.

Governments around the world created public data portals based on this premise. The world of consultancy jumped into it, as it often does, suggesting that Open Data would unleash massive economic development and business opportunities. A McKinsey report[65] stated in 2013 "*Open data—public information and shared data from private sources—can help create $3 trillion a year of value in seven areas of the global economy.*" In those years, the $3T figure was repeated like a mantra in all kinds of communications about open data. And so it was that open data became a tool in the political arsenal: party election manifestos produced for the General Election in 2010 and 2015 devoted several pages to tell what their parties, once in power, will achieve with Open Data.

How to measure the success of Open Data initiatives was always somehow fuzzy: if the data is released without control or imposition, it is very hard to make any direct quantitative measurement of its use. The easy metric, then, became the number of releases; sadly, this is also the wrong metric. Sometimes the drive for more data to be released turned into a competition of how many datasets were added to data.gov.uk; very little effort was put into measuring the usefulness and impact of the data. In 2015, a statement by the UK Environment Secretary famously set a target of releasing at least[66] 8,000 datasets in one year. The Department got very busy trying to meet this target, so much so that there were accusations[67] of "salami-slicing" datasets (splitting datasets in multiple chunks, in order to increase the count of single data releases).

Years after Katrina, what have been the outcomes of the Open Data movement? I'm yet to meet armchair auditors who really exposed anything useful; McKinsey's assessment didn't measurably materialise; there's a lot of data on Government portals but it's often hard to find and unusable. Was Open Data just a PR exercise? I don't entirely think so. Open data has injected two important ideas into public administrations: first, the awareness that data can

---

[64] Cameron in 'people power' pledge, BBC News, http://news.bbc.co.uk/1/hi/uk_politics/8067505.stm

[65] McKinsey Global Institute, Open data: Unlocking innovation and performance with liquid information, October 2013, https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information#

[66] Environment Secretary unveils vision for open data to transform food and farming, Gov.Uk News Story, https://www.gov.uk/government/news/environment-secretary-unveils-vision-for-open-data-to-transform-food-and-farming

[67] Owen Boswarva, OpenDefra: 8,000+ open datasets released in one year, https://mapgubbins.tumblr.com/post/146514977190/opendefra-8000-open-data-releases-in-a-year

be *used* in operations; second, that transparency and engagement are a positive value that all public authorities should aspire to.

Clare Moriarty, the former permanent secretary at the Department of the Environment at the time of their thousands of data releases, was a strong supporter of open data. She has strongly suggested that open data has created something positive, almost by side effect. She said[68]: *"Opening our data was at most half the issue. More important was the effect it had on open working among colleagues [...]. That led to interesting discussions about how open government relates to accountability [...] It took us into how we can be more open as individuals [...] building respect and trust. Open has all sorts of dimensions.[...] being open to challenge, open to new experiences, open to new ideas and to difference".*

Open data marked a very important step in the relationship between administration and the public. I've been an advocate of Open Data for years and my career developed in part by trying to make open data an effective movement. I'm still an advocate, but these days I'm also very aware that we need to be careful about overstating what data can do. An interesting concept is that of *careless openness*, described by Terence Eden, a Government Technology Advisor[69]. While analysing the gas meter dataset, a dataset that is openly *accessible* although not openly *licensed*, he found that it was very simple to discover his neighbour's gas supplier. The Land Registry has a public and open record of all residential property transactions. This means that anyone can easily find out how much a neighbour paid for their property, which can give away information about their wealth. Linking up this dataset to others, for example the equally open Register of Companies, may reveal more and more about individuals. We need to be vigilant for two reasons: the first is that the privacy issue is real and there is a genuine threat to privacy and a risk of ID fraud; the second is that the privacy issue may be exploited for the wrong reasons by those who have a vested interest in Governments being less transparent. These are antipodean problems, but are intrinsically related.

[68] Clare Moriarty, Redefining success, FDA, https://www.fda.org.uk/home/Newsandmedia/Features/clare-moriarty-women-into-leadership-speech.aspx

[69] Terence Eden, Open Data - but not *too* open, https://shkspr.mobi/blog/2020/11/open-data-but-not-too-open/

Conclusions

I hope you've followed me so far. It's time for me to wrap up. I've presented you with a combination of horror stories and good uses of data in the public sphere. My major message is simple: using data, numeric or otherwise, in public communications is difficult and full of hidden caveats. Communicating this difficulty should be our major concern whether we are operating as public officials, journalists, or scientists.

In order to communicate data well, we need to know the data that we're talking about from the inside out. We shouldn't make any assumptions about it. We need to know the process that generates that data. We need to know and communicate the context in which that data is collected, and the context in which it is used. We need to know enough about the problems we're trying to address with data, and we need data to be good enough to be actionable.

I've mentioned definitions extensively. Definitions are the most important aspect of effective data communications, because the way we translate a concept into its data equivalent tends to restrict the concept. Sometimes this restriction creates a data definition that is rather far from the common sense idea that we have about the same concept. It's also important to have in mind whether there is an authoritative source of the data and who owns it.

Communicating data is entirely a multidisciplinarity endeavour. Communicating data surely requires the technical capability of analysing it, the ability to tell stories with it, sometimes the skills to summarise its key lessons by visualizing it. But none of these will work in isolation. And these days, there's more: good data communication requires an appreciation of ethical and legal issues.

Most of the problems I've mentioned in this lecture are not unique to data. Data communication suffers from the same issues that affect science communication. These are generally centred on the issue of uncertainty, and the risk for scientific findings of being twisted willingly or unwillingly by the media. There is also an obvious risk in politics: data-driven evidence might clash with policy. In the UK a decade ago the Government sacked its chief drug adviser. Professor David Nutt, the chairman of the Advisory Council on the Misuse of Drugs, and a well respected academic, was guilty of making a claim in a research paper[70] that ecstasy was less dangerous than alcohol. This was

[70] David J Nutt, Estimating Drug Harms: A Risky Business?, January 2009, https://www.researchgate.net/publication/242626407_Estimating_Drug_Harms_A_Risky_Business

part of an evidence-based statement in a peer-reviewed publication, but it was against that Government's policy.

The relationship between science, politics, policy, and the judiciary is troublesome. A few years ago the debate on the Investigatory Powers Act included calls for Internet Providers to weaken encryption in ways that are mathematically impossible. Science should not be interpreted politically, but it often is, and data suffers from the same fate.

The key ingredients to fight back against these issues are effective data stewardship and curation curation, and inspirational leadership in the multidisciplinary world of data. On the one hand, we need to be really careful not to fall into the trap of impartiality. Data is not impartial because there is no way to "just use the data" to explain a phenomenon. The use of data can always be driven by an agenda, even unconsciously. Using data in public communications requires explanations and context, which goes somehow against the fast and furious news cycle of today. On the other hand, we need a bit of healthy data scepticism. Data alone cannot explain everything. Data, very often, has questions to offer; data cannot give enlightening answers without a lot of work around it. If we are tempted to say that "data speaks for itself", we need to stop and rethink. Because data doesn't speak.