Data: A Love Story for the Ages

Dr Victoria Baines, IT Livery Company Professor of IT

25 February 2025

Data is first of all a given. The word *datum* is the past participle of the Latin *dare*, meaning “to give”. As a noun it means “a given thing”, “a gift”. In later scientific thought it takes on a secondary meaning of “a given”, as in an accepted piece of information, premise, or principle. The idea of data as a gift can be a helpful lens through which to view its supply, use, and interpretation across different time periods and diverse geographical contexts.

Data is…or data are? Since *datum* is a neuter noun, *data* is the correct plural form, meaning “given things”. You will still hear some (particularly Classicists like me) pluralising sentences as a result – “the data are”, “the data show”, etc. – but increasingly, the dominant usage is to treat *data* as a singular. As is often the case, what is technically correct in a language gives way to what is more natural. If it’s any consolation, scientists were already arguing about this in the 18th century, John Wesley and the linguist John Elphinston among those debating whether *datum* or *data* was the more correct – or at least the less heinous - form.

Data’s appearance in English-language science comes about partly through translation from Greek into Latin of Euclid’s work on geometry, *Dedomena* (Δεδομένα) in the 12th century CE. When in 1661 John Leeke and George Serle published the first English translation of what was then known as the *Data*, they retained the Latin title rather than calling it *Givens*.

Early Modern Data

If we are looking for specific tipping points when the world (or at least parts of it) became data-based societies, the middle of the 17th century seems to be a good candidate. Whether one categorises this period as the tail end of the Scientific Revolution, the Scientific Renaissance, or the earliest stirrings of the Age of Enlightenment, we certainly see a greater interest in data as premises and data as what we now know as statistics.

One of its most mystifying expressions and its first appearance in an English language work of geometry is in the 1645 *Trissotetras*. Written by the Scottish aristocrat Sir Thomas Urquhart, it is, I am told, a sound treatise on trigonometry. Its prose style, however, is unconventional, to say the least. In the ‘Lexicidion’ (glossary) to this work, he lists the following data-related terms:

> *Data, is said of the parts of a Triangle, which are given us, whether they be sides or Angles, or both, of do, datum, dare.*
>
> *Datimista, are those Datas, which are neither Angles onely, nor sides onely, but Angles, and sides intermixedly: of data, and mista, from misceo.*
>
> *Datangulary, is said of the Concordances of those Moods, for the obtaining of whose Praenoscendas, we have no other Datas, but Angles, unto the foresaid Moods common.*
>
> *Datapurall, comes from datapura, which be those Datas, that are either meerly Angles, or meerly sides.*
>
> *Datolaterall, is said of the Concordances of those Moods, for the obtaining of whose Praenoscendas, the same sides serve for Datas.*
>
> *Datoquaere, is the very Problem it selfe, wherein two or three things are given, and a third or fourth*

required, as by the composition of the word appears.

Datisterurgetick, is said of those Moods which agree in the Datas of the last work: of data, ὕστερον, postremum, and ἔργον,opus.

Confused? Don't worry, so was everyone else. One imagines that Urquhart's use of the double plural 'datas' horrified his contemporaries. The other terms listed above were his own coinages. It may help to know that in his prolific publishing career he also produced a work in which he introduced the world's first universal language, anticipating Esperanto by more than two hundred years. In characteristically florid, classicising style, he named this work the *Logopandecteision* ('all telling/showing speech', 1653). Now considered something of a mathematical and linguistic maverick, he is said to have died in a fit of laughter.

Urquhart's invented terms didn't catch on with his fellow scientists, something about which we may now be rather thankful. Nevertheless, the work of one contemporary demonstrated the burgeoning of interest in analysis of data as a way of understanding society. Recognised as one of the founders of demography and epidemiology, in 1662 John Graunt published *Natural and Political Observations Made upon the Bills of Mortality*, a comparative analysis of weekly data on the numbers and causes of deaths recorded by London parishes. It's a striking example of how data can be turned into usable information or intelligence. His analysis showed, for instance, that 36 out of every 100 children died in infancy, and it identified peaks in deaths from preventable diseases. It earned him election to the Royal Society, and by a twist of fate the Great Fire of 1666 saw to it that it is the only surviving source for the parish records that were destroyed. Just as in the case of the Library at Hellenistic Alexandria, physical destruction can mean the disappearance of data. It is by no means as permanent as our current digital age would lead us to believe.

Who counts? Who is counted?

Despite his legacy, Graunt did not make it into a visual 'Hall of Fame' produced a century later by the philosopher and theologian Joseph Priestley, although a few of his fellow Royal Society members – and some Gresham professors – did. Priestley's 1765 *Chart of Biography* is the ultimate Enlightenment infographic. Priestley was elected to the Royal Society following presentation of this work, just as Graunt had on presentation of his analysis of the Bills of Mortality. It shows the lifespans of more than 2000 eminent humans, described as "a distinct, and a comprehensive view of the succession of great men: of every kind." As well as being an extraordinary undertaking designed to be used as a teaching aid, it speaks to us about data selection, data permanence, and the often-blurred distinction between description and prescription.

Effective infographics prompt as many questions as they answer and can be viewed at multiple levels. Priestley's chart is no exception. When one zooms out to take in the chart in its entirety, one immediately sees a concentration of names to the far right, closer to the author's own time. In the book accompanying the chart, he explains this as evidence of acceleration due to the "continual propagation and extension of knowledge." While it's undoubtedly demonstrable that scientific progress can proliferate further discoveries, Priestley doesn't appear to acknowledge that we have significant gaps in the data for much of the time covered by the left half of his chart. We know a lot less about these people simply because there isn't as much documentary evidence available to us.

Priestley's data selection is at once highly subjective and draws on received notions of what makes a great personage. It is a visual illustration of those deemed worthy of commemoration. Given their lack of access to education, exclusion from office holding and from careers in fields here identified as distinguished, coupled with comparative obscurity in written records, we should be entirely unsurprised that there are only 25 women on this chart – fewer than 1 per cent of those listed. Among them are the Ancient Greek poet Sappho, Cleopatra, "Boadicia" (Boudicca), and the female English monarchs.

Even at the time, the near absence of women in such compilations did not go unnoticed and drew attempts to restore the balance, among them de la Croix's 1769 *Dictionnaire portatif des femmes célebres* and Mary Matilda Betham's 1804 *A Biographical Dictionary of the Celebrated Women of Every Age and Country*. In addition to its noticeable gaps, Priestley's chart contains data that we would now consider to be anomalous: the life of legendary craftsman Daedalus of Greco-Roman myth is plotted in the 11th century BCE, even though there is no historical evidence for his existence. Two vertical dashes on the timeline at 1483 and 1635 denote the lifespan of Old Tom Parr, reported to have reached the grand age of 152. There is evidence also that some data was added later, among them the listing for poet and physician Mark Akenside, whose death

was recorded in 1770 – five years after the original publication date - and whose name does not appear in the text index accompanying the chart.

Priestley's chosen classifications also determine how and for what a figure is remembered, which can prove problematic for those who have distinguished themselves in multiple disciplines. Isaac Newton is situated among the "Mathematicians &c. Physicians", but so, too, are Gresham professor Robert Hooke and friend of Gresham John Wilkins. Gresham colleague Christopher Wren is categorised along with the likes of Hogarth as an artist, even though he was Professor of Astronomy at the College and Hooke and Wren were equally active as architects. The treatment of polymaths here shows how even in the pre-digital era, complex humans can be reduced to just one dimension in data. Moreover, presentation of the data in this format is instructional, directive, and prescriptive. It shows us who to value and how to remember them. It tells us about who counts in both senses: the people selecting the data and the people worthy of being commemorated.

Human invention, human error, human perspectives

Anomalies in data can give us intriguing insights into how analysis and graphical representations are produced. As well as the untimely inclusion of a still living poet in Priestley's *Chart of Biography*, data can survive on maps long after the physical landscape has altered. This appears to have been the case for the extraordinary *Peutinger Table* (*Tabula Peutingeriana*), which measures nearly 7 metres long and takes the form of an *itinerarium*, an ancient map of public routes on which all roads quite literally lead to Rome. While the surviving copy has been identified as dating from the 13th century, Glen Bowersock and others have made a strong case that it takes much of its content from a map commissioned by Marcus Vipsanius Agrippa in the 1st century BCE and put on display in the centre of Rome.[1] This would account for the inclusion of out-of-date information about Arabia reflecting its status under Roman rule, and also that of Pompeii, which was obliterated by the eruption of Vesuvius in 79 CE. At the same time, it shows the city of Constantinople, which wasn't founded until 328 CE, and this suggests that some changes in the landscape triggered amendments and additions – perhaps those deemed of sufficient significance to the rulers of Western Europe in a subsequent period. While not displaying the same type of technical sophistication as the real time updates we now see in digital mapping apps, the Peutinger Table was nevertheless a living document that bears witness to the changeability of data.

It is also testament to what can happen when data is lost, and why such losses may come about. The Western portion of the map containing the British Isles, Iberia, and Mauretania does not survive: the parchment has been cut at a point that allows us to view Colchester (Camuloduno) but not London, and it has been suggested that this was done so that the reverse side could be reused. If true, this is perhaps one of the earliest instances of which we know of structured data being deleted as part of a cost-cutting exercise. For whomever was in possession of the map at the time, the price of parchment was more pressing than the value of the data contained in that section. Given its discovery in the German city of Worms in 1494, it is tempting to speculate that the furthest reaches of the West were considered of little relevance to someone in Central Europe. But we don't know where the map was before this time, and nor do we know when the sections were removed.

Just as it is tempting to speculate about the why and the when of its transmission, gaps in data can prompt creative attempts to compensate based on informed assumption. In 1887, German art historian Konrad Miller published a facsimile of the extant sections of the Peutinger Table. In a subsequent work in 1898, he published a reconstruction of the lost British and Iberian sections. This is not entirely a work of imagination: in the accompanying text, Miller demonstrates how he took the data from other historical and cartographic sources. But neither can it claim to be an authentic part of the map. And yet, as a cursory web search reveals, it is often accepted as such, even though it has since been argued that what we are missing is not one but three sections.[2] Barring rediscovery, we can't possibly know for certain what was really in the lost section(s). Equally, we can't say that Miller has done a bad job – not least because he has clearly made an effort to include features with mythical associations, among them the Pillars of Hercules and island of Thule. Something that was an interesting thought experiment has come to be viewed as closer to hard data or knowledge. In some cases, virtual history is now difficult to distinguish from accepted fact.

The knowledge producers of the past also show us that data has never been infallible, especially where there

[1] Glen Bowersock. 1983. *Roman Arabia*. Harvard University Press.
[2] Richard J. A. Talbert. 2010. *Rome's World: The Peutinger Map Reconsidered*. Cambridge University Press.

is room for human intervention. The Mappa Mundi kept at Hereford Cathedral is one of the most comprehensive medieval maps constructed in the 'T and O' (*orbis terrarum*) format described by Isidore of Seville in the 7th century CE. In this configuration, East is at the top, West at the bottom, and Europe occupies the bottom left quadrant:

Fig. 1 Diagram of the 'T and O' configuration from a printed edition of Isidore of Seville's Etymologiae

Spare a thought, then, for the limner (illuminator) who, perhaps in a late stage of its production, was tasked with labelling the continents on the Hereford Mappa Mundi. For some reason which we cannot know for certain, he labelled Europe and Africa the wrong way around. The word 'AFFRICA' in gold lettering is visible in the bottom left quadrant populated by European cities and topographical features, and 'EUROPA' is visible in the bottom right quadrant, which includes Alexandria and the Nile delta. How might this have happened? We might imagine a poor, bleary-eyed soul struggling to focus in the dim candlelight. Perhaps he had been brought in solely to add these two labels and had insufficient knowledge of cartographic convention? Could he really have not seen that he was placing the 'R' and 'I' of 'AFFRICA' just above Paris, and in the same section as Hereford itself? Was his mind elsewhere and the mistake too costly and too damaging to rectify?[3] Might this have been a deliberate act of mischief or rebellion? Was it career-ending? These two data points may have had life-changing consequences for someone.

There are similarities between the Hereford map and the Peutinger Table, even while they have visibly different formats. Both are to some degree *cosmographia*, representations not only of places but peoples, myths, and narratives, and above all the ordering of the world according to a particular belief system. Where the one reflects the medieval inheritance of an imperial or perhaps Late Antique view of the world centred on Rome, the other (and kindred *mappae mundi* such as the Ebstorf map) situates Jerusalem at the centre with Christ at the top, presiding over all. Both maps feature a variety of exotic animals, drawing on the prevailing knowledge of the time as related in bestiaries. Both maps show camels: in the Peutinger Table one can be found in the centre of Jerusalem. The lynx shown on the Black Sea coast on the Hereford map is accompanied by the following description: "The Lynx sees through walls and urinates a black stone." Here taking the Greek form *monoceros*, a unicorn stands in North Africa: while we may view them as entirely mythical, these also featured in contemporary bestiaries. Contemporary travel narratives provide source material for articulating difference between peoples on the Hereford and Ebstorf maps, with several figures reminiscent of descriptions in the unreliable *Travels of Sir John Mandeville*, which we encountered in my

[3] https://mortimerhistorysociety.org.uk/wp-content/uploads/Mortimers/Articles/MappaMundi.pdf

previous lecture, "Who's Afraid of Robots?"[4] Blemmyes with heads in their chests, sciapods with one giant foot each, troglodites and dog-headed cynocephali are literally outlandish, inhabiting the furthest reaches of the known world at the time from this point of view – Africa, India, and the Arctic Circle.

This kind of map is an encyclopedia as much as it is a navigational aid. Notably absent are lines denoting borders between territories. Rulers of course exercised jurisdiction over their subjects and control of their lands. But it was not until the 17th century that lines were clearly drawn. In my very first Gresham lecture, "Who Owns the Internet?",[5] we explored how states apply to cyberspace a concept of national sovereignty that is most closely associated with the Peace of Westphalia (1648). This group of treaties ended the Thirty Years War and the Eighty Years War between Catholic and Protestant powers in mainland Europe and is viewed by International Relations experts as a defining moment in Western politics. Several territories changed hands as part of the settlement, the Netherlands and the Swiss Confederation were formally confirmed as states, and the Peace marked the end of the ability of the Holy Roman Empire to interfere in states' affairs.

Spatially accurate maps were therefore required to enforce sovereignty. Maps with hard-line borders and states shaded in distinct colours became more prevalent, and spatial data came to dominate, while demographic, natural history, mythical, and even wayfinding data began to fade. As political historian Steve Pickering has observed, just three years after the Peace of Westphalia was concluded, Nicolas Sanson's "*L'Europe*" delineated countries' extents with brightly coloured borders.[6] Within two hundred years, colonial powers would boast of the amount of territory "coloured in" with their chosen hue. In this way, maps came to display the extent of ownership and exploitation. The stories of the people on these lands came to be told in alternative formats.

The oldest census? The impact of wishful thinking on A.I.

In my research for this lecture I wanted to find out when the earliest census that we know of was taken. Numerous sources online, including several national statistics authorities, state that the Babylonians conducted a census around 3800 BCE, and that it enumerated population, livestock, and crops in the interest of food security. One refers to a clay tablet in the British Museum as evidence. This evidence has, however, proved rather elusive. In the catalogue of the Cuneiform Digital Library Initiative (CDLI), I found an administrative clay tablet bearing a Sumerian inscription which enumerates dairy production, including the following entries:[7]

from Lugal-šunire;

3 barig 2 ban2 5 sila3 butter oil,

41 milk cows × 5 = 205

1 gur 7 ½ sila3 kašk cheese,

41 milk cows × 7.5 = 307.5

from Šeškala;

1 barig 5 sila3 butter oil,

13 milk cows × 5 = 65

1 barig 3 ban2 7 1/2 sila3 kašk cheese,

13 milk cows × 7.5 = 97.5

from Ur-Suda;

1 barig 1 ban2 butter oil,

[4] https://www.youtube.com/watch?v=7vIKkmK5lMU&t=2486s
[5] https://www.youtube.com/watch?v=tJS_IEU-7oU
[6] https://isprs-archives.copernicus.org/articles/XL-4-W3/111/2013/isprsarchives-XL-4-W3-111-2013.pdf
[7] https://cdli.mpiwg-berlin.mpg.de/artifacts/118388

14 milk cows × 5 = 70

1 barig 4 ban2 5 sila3 kašk cheese,

14 milk cows × 7.5 = 105

It was found on the site of the ancient city of Umma (modern Jokha in Iraq) and it is believed to date to around 2000 BCE. It is held not in the British Museum but in a college in the US. It is signed by "Atu, Chief Cattle Manager" and dated to "the year in which the silver-chair of Enlil was fashioned." Several debits are also listed, including king's destruction of Urbilum (modern Erbil in Iraqi Kurdistan) and his attendance at a beer festival. On the reverse of the tablet, Atu's calculations are accompanied by data on the receipt of silver:

1/3 mana 1 shekel silver

the first time,

2/3 mana 8 1/2 shekels silver

the second time,

via Lukala,

9 shekels silver,

via Ur-Šara the chief accountant,

9 2/3 shekels 15 grains silver

via Lu-Zabala,

its [scales] stone overhead: 1 1/3 shekel 2/3 grains silver,

This clearly isn't the famed Babylonian census of 3800 BCE – in fact I've since been relieved to have my suspicion confirmed by data scientist Andrew Whitby's assessment (see Further Reading) that the "oft repeated claim, dating the 'first census' to 3800 BCE in Babylon, seems to have resulted from a misinterpretation by an overzealous early twentieth-century statistician." According to Whitby, both this and the assertion of a census in Egypt around 2500 BCE originate from George Knibbs, Australia's first national statistician. If I may add my own conjecture to Whitby's excellent and reassuring analysis, 2500 BCE is around the time the pyramids at Giza are thought to have been constructed. The inference that a census was taken could well have been based on deduction that such a sizeable workforce would need to be enumerated in order to establish that it was sufficient. It would also have aligned with the general appreciation – evidenced in hieroglyphic inscriptions and funerary monuments – that Ancient Egypt had a mature civil service.

In the age of AI, the fact that several trusted sources have regurgitated erroneous claims as accepted fact results in their being promoted not only in web search results, but also in text summaries produced by large language models, including the one that operates at the top of Google's results page. Inaccurate data is being pushed to users on the basis that its sources are perceived to be authoritative. This is not a case of generative AI hallucinating: it is an example of what happens when it is fed poor quality data published by humans.

Putting us in buckets – data as an expression of power

Gathering data about people has proved to be an efficient way of identifying numbers of those eligible for a range of activities including military service, political participation, and payment of tax. The Greek historian Herodotus reports (*Histories* 2.177.2) that Amasis II, pharaoh in the 6th century BCE, "made the law that every Egyptian declare his means of livelihood to the ruler of his district annually, and that omitting to do so or to prove that one had a legitimate livelihood be punishable with death" and that "Solon the Athenian got this law from Egypt and established it among his people." Aristotle describes Solon's assessment of property as determining a man's class status (*Athenian Constitution* 7.3). Both he and the 1st century CE biographer Plutarch highlight that men in the lowest class (*thetes*) were barred from holding office. Humans therefore

have a long history of excluding each other and withholding opportunity using data.

Just before Plutarch's time, at the turn of the Common Era we find evidence of two censuses that make different but equally important contributions to our understanding of how people are datafied. The census taken by the Han dynasty of China in 1/2 CE is the earliest for which we still have the numbers. It recorded 59,594,978 people, and is considered by modern estimates to be reasonably accurate:

Year (A.D.)	Persons	Households	Persons per Household
2	a59,594,978	a12,233,062	4·9
57	21,007,820	4,279,634	4·9
75	34,125,021	5,860,572	5·8
88	43,356,367	7,456,784	5·8
105	53,256,229	9,237,112	5·8
125	b49,690,789	9,647,838	5·2
140	c49,150,220	c9,698,630	5·1
144	49,730,550	9,946,919	5·0
145	49,524,183	9,937,680	5·0
146	d47,566,772	9,348,227	5·1
156	e56,486,856	e10,677,960	5·3

Fig.2 China: Recorded Population Statistics, 2-156 CE, taken from John D. Durand (1960). "The population statistics of China, AD 2-1953". Population Studies 13:3.

What Andrew Whitby describes as "China's numerical supremacy" lies for me in the ability since at least 2 CE (and probably earlier), to compare people data over time. But where the capability remained constant, the land mass changed. John Durand in his analysis notes that the totals for 2 and 140 CE included some areas in Korea, Mongolia, Turkestan, and Vietnam. Census data is therefore an expression of expansion and ownership, and it continues to this day: according to the non-profit Population Reference Bureau, when in 2010 China recorded the largest population in the world (1.3 billion people), this included 23 million Taiwanese people, thereby claiming the population of the Republic of China for the People's Republic of China.[8] The people who are the subjects of these data do not appear to have the power to object to being co-opted in this way.

Around the time of the Western Han census of 2 CE, Europe was engaged in its own large scale population surveys. The census of Quirinus (6 CE) is famously described in the Gospel of Luke (2.1-5) as follows

> *And it came to pass in those days, that there went out a decree from Caesar Augustus, that all the world should be taxed. And this taxing was first made when Cyrenius was governor of Syria. And all went to be taxed, every one into his own city. And Joseph also went up from Galilee, out of the city of Nazareth, into Judaea, unto the city of David, which is called Bethlehem; (because he was of the house and lineage of David:) To be taxed with Mary his espoused wife, being great with child.*

The context for the Nativity – and the reason for its location in Bethlehem - is taxation. As the text above suggests, wholesale taxation of residents in the provinces was an innovation of the first emperor Augustus (27 BCE – 14 CE), and by this time there had been two censuses in Gaul (27 BCE and 12 BCE). As a new acquisition in 6 CE, Judaea presented an unprecedented opportunity to raise funds. Roman censuses assessed citizens and non-citizens alike with respect to the extent of their land (*tributum soli*) and, especially in the case of non-citizens in the provinces, their eligibility for poll tax (*tributum capitis*). Later that century, the First Jewish Revolt (66-74 CE) resulted not only in the destruction of Jerusalem and its Temple, but also in the imposition of the *fiscus judaicus*, a supplementary tax of 2 denarii per person for all Jews in the Roman Empire. It symbolically redirected to the reconstruction of the Temple of Jupiter in Rome monies that had previously funded the maintenance of the Temple in Jerusalem. The Roman biographer Suetonius recalls first-hand how 'Jewishness' was defined and determined (*Life of Domitian* 12):

> *Besides the exactions from others, the poll-tax on the Jews was levied with extreme rigour, both on those who lived after the manner of Jews in the city, without publicly professing themselves to be such, and on those who, by concealing their origin, avoided paying the tribute imposed upon that people. I remember, when I was a youth, to have been present, when an old man, ninety years of age, had his person exposed to view in a very crowded court, in order that, on inspection, the*

[8] https://www.prb.org/resources/milestones-and-moments-in-global-census-history/

procurator might satisfy himself whether he was circumcised.

Enforced examination of physical characteristics with the aim of categorising individuals presages some of the darkest applications of eugenics in modern history. It speaks to the dual meanings of value, and of a person's worth. We may collect data on resources to know how many of them we have at our disposal. But the more we collect data on people, the more we know about *what kind* of people they are, and the greater the opportunity to profile them – whether for exploitation, to assess risk, or to control the influence of those we deem undesirable. Considered together, the census of Quirinus and the *fiscus judaicus* are an early example of personal data collection as an explicit instrument of control.

In Roman society, these statistical and moral functions were formally blurred. Our modern word 'censorship' has its roots in the role of the censors, Roman magistrates who oversaw both the taking of the census and the morals of the community. Censors' written comments on the conduct of citizens in the official list could result in their being removed from their tribes and no longer eligible to vote – although still eligible to pay tax. Membership of the senatorial and equestrian upper classes was also in the power of the censors, with expulsion a real threat for those considered to no longer make the grade. Previously elected by senate committee, in 22 BCE, two censors were appointed directly by Augustus, marking the point at which the moral and administrative oversight of the role fell under the authority of the emperor – the government as opposed to the governing body.

This kind of data collection takes us far away from the notion of it being volunteered or freely given. Indeed, it is tempting to support Johanna Drucker's suggestion (see Further Reading) that 'data' be renamed *capta* ('things taken'), to reflect the reality of the practice of its collection. Data collection against the will or without the knowledge of the subject is not solely the preserve of the digital age. What changes with the introduction of digital technology are the speed and scale at which data can be collected and analysed, the sheer number of opportunities for people to volunteer ever more detailed and personal information about themselves, and the increased use of algorithms to categorise and make predictions about them.

Baby, remember my name

Just as data can be created without our explicit consent, so too can it be removed. The upper classes in Ancient Rome were preoccupied with legacy and posterity. One of the best-known extant manifestations of this is the *Res Gestae Divi Augusti*, the Emperor Augustus' account of his own achievements, which was inscribed in full on stone monuments around the empire. By the same token, among the penalties for those judged to be enemies of the state was condemnation of their memory (*damnatio memoriae*), which typically included destroying their images and erasing their name from public records and inscriptions. The decrees of condemned emperors were also reversed. Suetonius details the condemnation of the 'bad' emperor Domitian in suitably destructive terms: his images were "dashed in pieces upon the floor of the senate-house"; his titles were "obliterated" (*eradendos*), and all memory of him abolished (*abolendam*, *Life of Domitian* 23). Indeed, numerous stone artefacts survive in which his name or, indeed, the entire text of the inscription has been chiselled out. Classical scholars have noted how, then as now, creating a void in history by 'cancelling' a public figure can serve to draw attention to their negative example as much as it seeks to consign them to oblivion.[9]

Data can also persist by accident. The earliest surviving document found in Britain with a date on it is legible to us only because someone pressed too hard with their stylus, breaching the wax of the tablet and scratching the wood beneath. It's an 'IOU' note, dated to 8th January 57 CE.[10] The site of the ancient town of Oxyrynchus in Egypt has yielded thousands of everyday documents and fragments of lost Greek and Roman authors because its rubbish dumps lay undisturbed until the 19th century. They survived *because* not despite of becoming the ancient equivalent of chip paper. Part of the text on which I wrote my PhD thesis only came to light when in 1899 an Oxford undergraduate discovered it in an 11th century manuscript in the Bodleian Library. And data can survive against the creator's will: according to Suetonius, we have the text of Virgil's *Aeneid* only because the Emperor Augustus overruled the poet's wish for it to be burned if he died before it was finished.

Data can be structured to justify power and reverence. Organigrams tell you your place in a corporate hierarchy. Royal family trees have for centuries served as visualisations of lineage but are by no means

[9] https://web.sas.upenn.edu/discentes/2020/08/21/damnatio-memoriae-on-facing-not-forgetting-our-past/
[10] https://www.bbc.co.uk/news/uk-england-london-36415563

standardised in their representation: for example, in the cases of the House of Saud and the Tree of Jesse, earliest ancestors are the roots at the bottom, with the more recent descendants on the higher branches. If our ancestors were not public figures or otherwise noteworthy/notorious (the latter often commemorated in orally transmitted folk songs), their posterity was for millennia largely confined to the living memory of those who knew them personally. When those people themselves died, that information faded from memory. The fact that they existed and ceased to exist might persist in the records of a place of worship, along with a possible date of marriage. With the introduction of modern, regular censuses, data began to be collected systematically and stored centrally about people's demographics, membership of households, locations and professions. That data was not intended to be used by descendants seeking to understand their ancestry, but it has become a treasure trove for amateur genealogists and the genealogy industry, powered by digital technology and scientific advances.

Apps that cater to those who want to build a family tree crawl vast databases of births, marriages, deaths, censuses, and military service records. My 8th great grandfather, William Ardern, clearly couldn't have known that I would be able to access his life story quite as easily and systematically as I can now. But then neither would his descendant, my great-grandfather, the cotton mill worker whom you 'met' in my first lecture of this year, "The Ancient History of Computers and Code"[11]. My understanding of my heritage is undoubtedly richer as a result of having access to this data. It would be anachronistic to apply modern concepts of consent to its collection. Nevertheless, consideration of its re-use beyond its original purpose chimes with the guiding principles of the regulations we have now, to make people aware of the reasons we collect their data so that they can give *informed* consent to its processing, and to minimise its use for other than the agreed purposes. Time moves on, technology evolves, and regimes change. Cold case crimes can now be solved with retrieval and analysis of DNA data; equally, the data you shared about your sexuality with one service provider could be requisitioned or intercepted by an intolerant government. The application of new technologies to old artefacts and the repurposing of datasets brings new benefits and new risks.

Seeing the future = needing more data

Communications technology is both fertile ground and a catalyst for predictive analysis. Weather forecasting is a case in point. Until weather stations could transmit readings more quickly than the weather moved, there was limited predictive ability beyond seasonal and cyclical expectations based on historical observations. These could be impressive enough in themselves: Luke Howard's 1818 work, *The Climate of London*, compared air temperatures from three rural sites and one in the centre of the city. His data visualisations are things of beauty, and have much in common with modern infographics. As Howard himself acknowledged (p.xvi), to improve the accuracy of prediction what was needed above all was more data:

> *There is no subject on which the learned and the unlearned are more ready to converse, and to hazard an opinion, than on the Weather — and none on which they are more frequently mistaken! This, alone, may serve to show that we are in want of more* ***data*** *[original emphasis], of a greater store of facts, on which to found a Theory that might guide us to more certain conclusions; and facts will certainly multiply together with observers.*

Permit me to make a brief linguistic detour here. You will often hear weather forecasters say that a specific day/month/year is the hottest/coldest/wettest "since records began." When we interrogate this more closely, we realise that what they really mean is not "since anyone started recording weather data" but "since the organisation that provides the data for our forecasts started collecting and storing data systematically." Diarists, community authorities, and others have all recorded information about the weather. What they hadn't necessarily done until the 19th century was set out to take regular readings according to agreed and consistent parameters. And, equally important, they lacked the ability to share these in a timely fashion.

Widespread adoption of the electric telegraph enabled fast time sharing of data, both from stations taking the readings and central offices issuing weather warnings; in contrast, the optical telegraph (semaphore) tended to fall foul of the weather. As data could be transmitted nationwide and internationally more quickly than before, weather patterns could be anticipated in a way that made sense to include them in newspapers. Radio, television, and the Internet have all come to be used as media for transmission. So, simply collecting data is very rarely enough. It's what you do with it that counts, and sometimes developments in other branches of technology need to give a helping hand.

---

[11] https://www.youtube.com/watch?v=Ass6vZt2lAE

We are now used to the idea that data drives and can even change public policy, but this was not always the case. The science of statistics has its linguistic roots in German *Statistik*, coined by economist Gottfried Achenwall in 1749 to describe analysis of data on the state. Since then, the presentation of data as evidence has become increasingly prominent in decision-making – for instance, when seeking to introduce new laws to tackle a particular type of crime, or to introduce new public health measures. Two superb Gresham lectures by Lynn McDonald and Sarah Hart explore how Florence Nightingale sought to improve conditions in military hospitals.[12] Her polar area charts illustrate very effectively the scale and proportion of deaths of soldiers from preventable or mitigable diseases in contrast to deaths from wounds and indeed any other cause. As Eileen Magnello notes, "The War Office showed no interest, but public outrage, informed by her graphs, forced it to take action."[13] In 1859 Nightingale and Harriet Martineau, a leader-writer for the *Daily News*, published *England and her Soldiers*, in which Nightingale's data visualisations were paired with Martineau's text. It was part of a very public campaign for better healthcare for the military, communicated to a general audience.[14]

Data and hindsight

Data can question our gut feelings, test our assumptions, and challenge our belief systems, and it is often in need of revision. 1859 also saw the publication of Charles Darwin's *The Origin of Species*. Also intended for a general audience, Darwin's theory of evolution was a clear challenge to the Christian creation narrative, popularising doubts about the assertion made in 1650 by James Ussher, Archbishop of Armagh, that the world was created on 22nd October, 4004 BC. Ussher's chronology was based on genealogies set down in the Old Testament and became the prevailing knowledge in Christendom at the time due to its inclusion in the King James Bible, in the margin of the Book of Genesis. While it's easy to mock what in hindsight appears as comparative ignorance, we do this from a privileged vantage point which benefits from subsequent geological discoveries. And this case helps us understand another key point about knowledge and evidence. Major religions claim different dates for the creation of the Earth, some of them closer than others to what we now know from radiometric dating. Using different datasets can result in different answers to the same question.

Data can also defy our expectations by helping us put our fears in perspective. This regularly happens in relation to crime, where perception routinely outstrips its reported incidence. In England and Wales, recorded crime has been falling for thirty years, but members of the public – 79% of those surveyed by the Office for National Statistics – believe that it is rising. There are several possible factors playing into this, including media representations and political rhetoric. But it may also bear out physician and statistician Hans Rosling's theory that people tend to believe things are getting worse, even when the data shows otherwise – what he called the 'Negativity Instinct.' In his book, *Factfulness*, and in the resources produced by his Gapminder project, Rosling repeatedly demonstrated that things were better than we thought, and that many of the world's ills – including legal slavery, infant mortality, HIV infections, and extreme poverty – were in fact continuing to decline.[15]

According to my reading of his work, he wasn't suggesting that there was cause for complacency: one child death is too many for anyone, let alone a former paediatric doctor such as Rosling. His theory helps us put a name to and give context to the distrust of science and scientists that we see now, especially in relation to priority issues such as public health and climate change. As a researcher who has planned collection of data for both qualitative and quantitative analysis, I know full well that data doesn't exist independently 'out there' as an unassailable, absolute truth. It is selected and interpreted by humans (or tools built and programmed by humans), and very often presented by humans as evidence. This doesn't mean that data is inherently untrustworthy, nor that we should suspect everyone who uses it. As with all information presented to us, it is up to us to use our critical thinking to question and stress-test it; to understand why it might have been collected, interpreted, and presented in a certain way. And it is up to those who present it to qualify it with that all important contextual information, just as mapmakers provide a legend to explain cartographical features, and Florence Nightingale included a text rubric on the different colours in her diagram.

This would seem to be particularly important in an age when more and more sensitive data is being collected

---

[12] https://www.gresham.ac.uk/watch-now/florence-nightingale-and-her-crimean-war-statistics-lessons-hospital-safety; https://www.gresham.ac.uk/watch-now/maths-nightingale
[13] https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2013.00706.x
[14] https://journal.sciencemuseum.ac.uk/article/nightingale-and-martineau/#abstract
[15] https://www.gapminder.org/facts/improvements/

about every single one of us. Our love affair with data has lasted a very long time, and one might say it is only getting more intense. We are 'humans in data' more than ever before. Every aspect of our lives is liable to datafication. We need to be able to exercise the right to be forgotten, as in European data protection legislation, to choose what data is collected, what data is kept, and what is consigned to oblivion. At the same time, we need to resist our reduction as humans to data points, and to remind data processors and each other that we are so much more.

Our current concerns about data quality, representation, and algorithmic justice – fair outcomes from analysis of the data about us – are understandable, legitimate, and by no means new. The lack of female representation in Joseph Priestley's *Chart of Biography*, Luke Howard's demand for more data in the interest of greater accuracy, and the historic profiling of people according to their attributes foreshadow some of the dilemmas we now face regarding Artificial Intelligence. Digital tools' documented inability to recognise darker skinned faces has to some extent been explained by their insufficient representation in training data sets.[16] Equally, where training data does not include specific markers for characteristics such as ethnicity, researchers have warned that this invisibility could exacerbate health inequalities.[17] The logical solution is to collect more data on underrepresented groups to redress the balance. But as we have seen, this results in further datafication of individuals, with further potential for discriminatory profiling.

As soon as we start enumerating humans, we must ask ourselves, "Who counts?". Who is doing the counting, and who has value? The answer to the former will tell us why we are being counted. The answer to the latter must surely be "everyone."

Further Reading and Resources

Johanna Drucker. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5.1. https://dhq-static.digitalhumanities.org/pdf/000091.pdf

Luciano Floridi. 2008. "Data", in The *International Encyclopedia of the Social Sciences*, ed. William A. Darity. Macmillan. https://philarchive.org/rec/FLOD-2

Daniel Rosenberg. 2013. "Data before the Fact", in *"Raw Data" is an Oxymoron*, ed. Lisa Gitelman. MIT Press. https://projects.iq.harvard.edu/files/eswg/files/rosenburg_-_rawdata.pdf

Hans Rosling. 2018. *Factfulness: Ten Reasons We're Wrong About the World – and Why Things Are Better Than You Think*. Hodder & Stoughton.

Andrew Whitby. 2020. *The Sum of the People: How the Census has Shaped Nations, from the Ancient World to the Modern Age*. Basic Books.

Chris Wiggins & Matthew L. Jones.2023. *How Data Happened: A History from the Age of Reason to the Age of Algorithms*. W. W. Norton.

[16] https://www.turing.ac.uk/sites/default/files/2020-10/understanding_bias_in_facial_recognition_technology.pdf
[17] https://www.rsm.ac.uk/media-releases/2023/ai-must-not-worsen-health-inequalities-for-ethnic-minority-populations/