Graphical Presentation of Longitudinal Data

Introduction

Let us begin with a few kind words about the bubonic plague. In 1538, Thomas Cromwell, the Earl of Essex (1485–1540), issued an injunction (one of 17) in the name of Henry VIII that required the registration of all christenings and burials in every English Parish. The London Company of Parish Clerks compiled weekly *Bills of Mortality* from such registers. This record of burials provided a way to monitor the incidence of plague within the city. Initially, these *Bills* were circulated only to government officials; principal among them, the Lord Mayor and members of the King's Council.

They were first made available to the public in 1594, but were discontinued a year later with the abatement of the plague. However, in 1603, when the plague again struck London, their publication resumed on a regular basis.

The first serious analysis of the *London Bills* was done by John Graunt in 1662, and in 1710, **Dr. John Arbuthnot**, a physician to Queen Anne, published an article that used the christening data to support an argument (possibly tongue in cheek) for the existence of God. These data also provide supporting evidence for the lack of existence of statistical graphs at that time.

Figure 1 is a simple plot of the annual number of christenings in London from 1630 until 1710. As we will see in a moment, it is quite informative. The preparation of such a plot is straightforward, certainly requiring no more complex apparatus than was available to Dr. Arbuthnot in 1710. Yet, it is highly unlikely that Arbuthnot, or any of his contemporaries, ever made such a plot.

The overall pattern we see in Figure 1 is a trend over 80 years of an increasing number of christenings, almost doubling from 1630 to 1710. A number of fits and starts manifest themselves in substantial jiggles. Yet, each jiggle, save one, can be explained. Some of these explanations are written on the plot. The big dip that began in 1642 can only partially be explained by the onset of the English Civil War. Surely the chaos common to civil war can explain the initial drop, but the war ended in 1649 with the beheading of Charles I at Whitehall, whereas the christenings did not return to their earlier levels until 1660 (1660 marks the end of the protectorate of Oliver and Richard Cromwell and the beginning of the restoration). Graunt offered a more complex explanation that involved the distinction between births and christenings, and the likelihood that Anglican ministers would not enter children born to Catholics or Protestant dissenters into the register.

Many of the other irregularities observed are explained in Figure 1, but what about the mysterious drop in 1704? That year has about 4000 fewer christenings than one might expect from observing the adjacent data points. What happened? There was no sudden outbreak of a war or pestilence, no great civil uprising, nothing that could explain this enormous drop.

The plot not only reveals the anomaly, it also presents a credible explanation. In Figure 2, we have duplicated the christening data and drawn a horizontal line across the plot through the 1704 data point. In doing so, we immediately see that the line goes through exactly one other point −1674. If we went back to Arbuthnot's table, we would see that in 1674 the number of christenings of boys and girls were 6113 and 5738, exactly the same number as he had for 1704. Thus, the 1704 anomaly is likely to be a copying error! In fact, the correct figure for that year is 15 895 (8153 boys and 7742 girls), which lies comfortably between the christenings of 1703 and 1705 as expected.

It seems reasonable to assume that if Arbuthnot had noticed such an unusual data point, he would have investigated, and finding a clerical error, would have corrected it. Yet he did not. He did not, despite the fact that when graphed the error stands out, literally, like a sore thumb. Thus, we must conclude that he never graphed his data. Why not? The answer, very simply, is that graphs were not yet part of the statistician's toolbox. (There were a very small number of graphical applications prior to 1710, but they were not widely circulated and Arbuthnot, a very clever and knowledgeable scientist, had likely not been familiar with them.)

Reproduced from the *Encyclopedia of Statistics in Behavioral Science*. © John Wiley & Sons, Ltd.
ISBN: 0-470-86080-4.

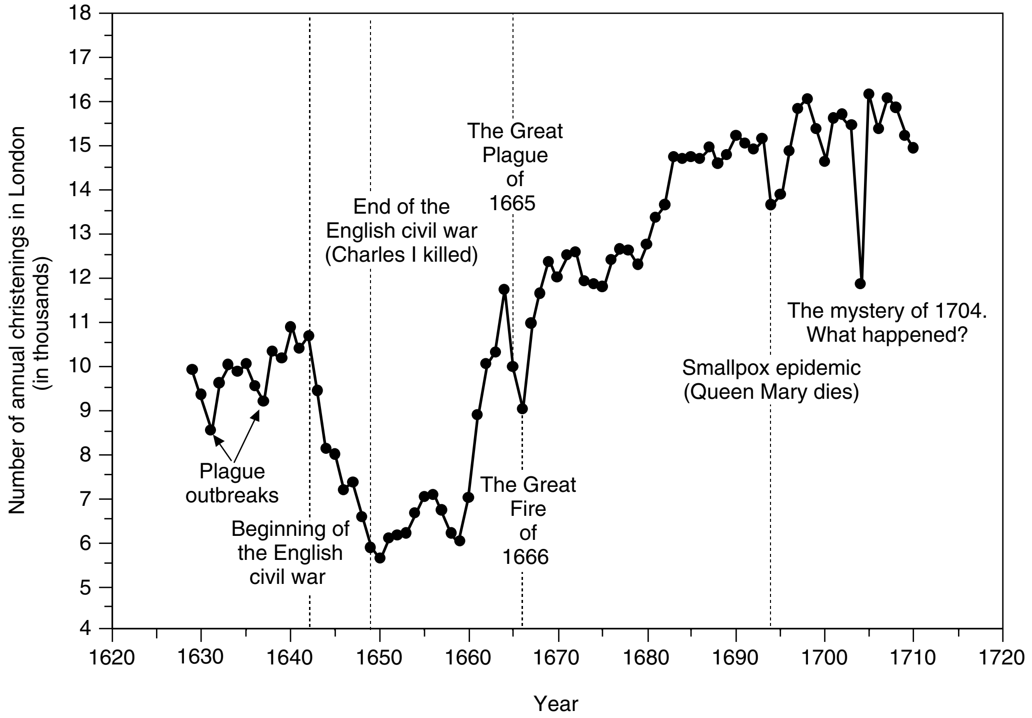


Figure 1 A plot of the annual christenings in London between 1630 and 1710 from the London Bills of Mortality. These data were taken from a table published by John Arbuthnot in 1710

The Beginnings of Graphs

Graphs are the most important tool for examining longitudinal data because they convey comparative information in ways that no table or description ever could. Trends, differences, and associations are effortlessly seen in the blink of an eye. The eye perceives immediately what the brain would take much longer to deduce from a table of numbers. This is what makes graphs so appealing – they give numbers a voice, allowing them to speak clearly. Graphs and charts not only show what numbers tell, they also help scientists tease out the critical clues from their data, much as a detective gathers clues at the scene of a crime. Graphs are truly international – a German can read the same graph that an Australian draws. There is no other form of communication that more appropriately deserves the description 'universal language.'

Who invented this versatile device? Have graphs been around for thousands of years, the work of inventors unknown? The truth is that statistical graphs were not invented in the remote past; they were not at all obvious and their creator lived only two centuries ago. He was a man of such unusual skills and experience that had he not devised and published his charts during the Age of Enlightenment we might have waited for another hundred years before the appearance of statistical graphs.

The Scottish engineer and political economist, William Playfair (1759–1823) is the principal inventor of statistical graphs. Although one may point to solitary instances of simple line graphs that precede Playfair's work (see Wainer & Velleman, [10]), such examples generally lack refinement and, without exception, failed to inspire others. In contrast, Playfair's graphs were detailed and well drawn; they appeared regularly over a period of more than 30 years; and they introduced a surprising variety of practices that are still in use today. He invented three of the four basic forms: the statistical line graph, the **bar chart**, and the **pie chart**. The other important basic form – the **scatterplot** – did not appear until atleast a half century later (some credit Herschel [4] with its first use, others believe that

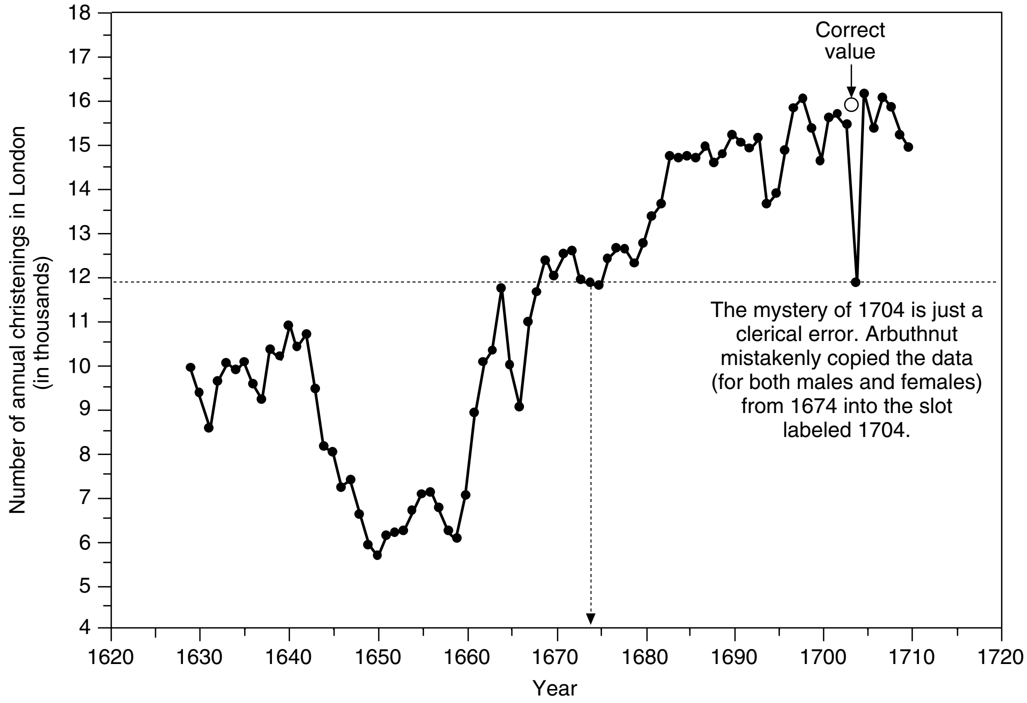


Figure 2 The solution to the mystery of 1704 is suggested by noting that only one other point (1674) had exactly the same values as the 1704 outlier. This coincidence provided the hint that allowed Zabell [11] to trace down Arbuthnot's clerical error. (Data source: Arbuthnot 1710)

Herschel's plot was a time-series plot, no different than Playfair's). Playfair also invented other graphical elements, for example, the circle diagram and statistical Venn diagram; but these innovations are less widely used.

Two Time-series Line Graphs

In 1786, Playfair [5] published his *Commercial and Political Atlas*, which contained 44 charts, but no maps; all of the charts, save one, were variants of the statistical time-series line graph. Playfair acknowledged the influence of the work of Joseph Priestley (1733–1804), who had also conceived of representing time geometrically [6, 7]. The use of a grid with time on the horizontal axis was a revolutionary idea, and the representation of the lengths of reigns of monarchs by bars of different lengths allowed immediate visual comparisons that would otherwise have required significant mental arithmetic. An interesting sidelight to Priestley's plot is that he accompanied the original (1765) version with extensive explanations, which were entirely omitted in the 1769 elaboration when he realized how naturally his audience could comprehend it (Figure 3).

At about the same time that Priestley was drafting his time lines, the French physician Jacques Barbeu-Dubourg (1709–1779) and the Scottish philosopher Adam Ferguson (1723–1816) produced plots that followed a similar principle. In 1753, Dubourg published a scroll that was a complex timeline spanning the 6480 years from The Creation until Dubourg's time. This is demarked as a long thin line at the top of the scroll with the years marked off vertically in small, equal, one-year increments. Below the timeline, Dubourg laid out his record of world history. He includes the names of kings, queens, assassins, sages, and many others, as well as short phrases summarizing events of consequence. These are fixed in their proper place in time horizontally and grouped vertically either by their country of origin or in Dubourg's catch-all category at the bottom of the chart 'événements mémorables.' In 1780, Ferguson

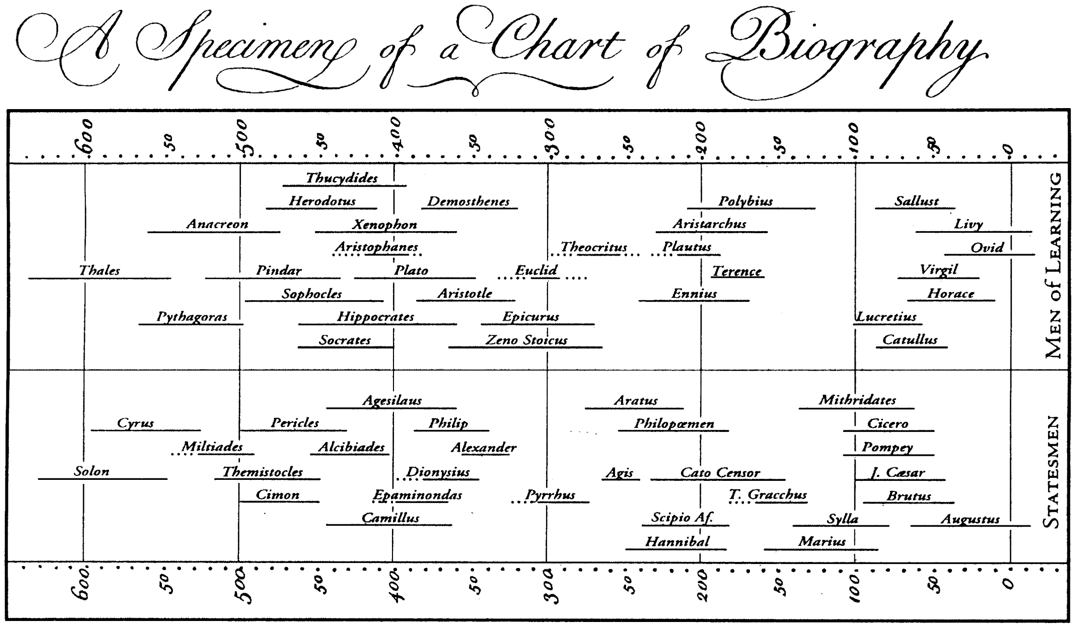


Figure 3 Lifespans of 59 famous people in the six centuries before Christ (Priestley, [6]). Its principal innovation is the use of the horizontal axis to depict time. It also uses dots to show the lack of precise information on the birth and/or death of the individual shown

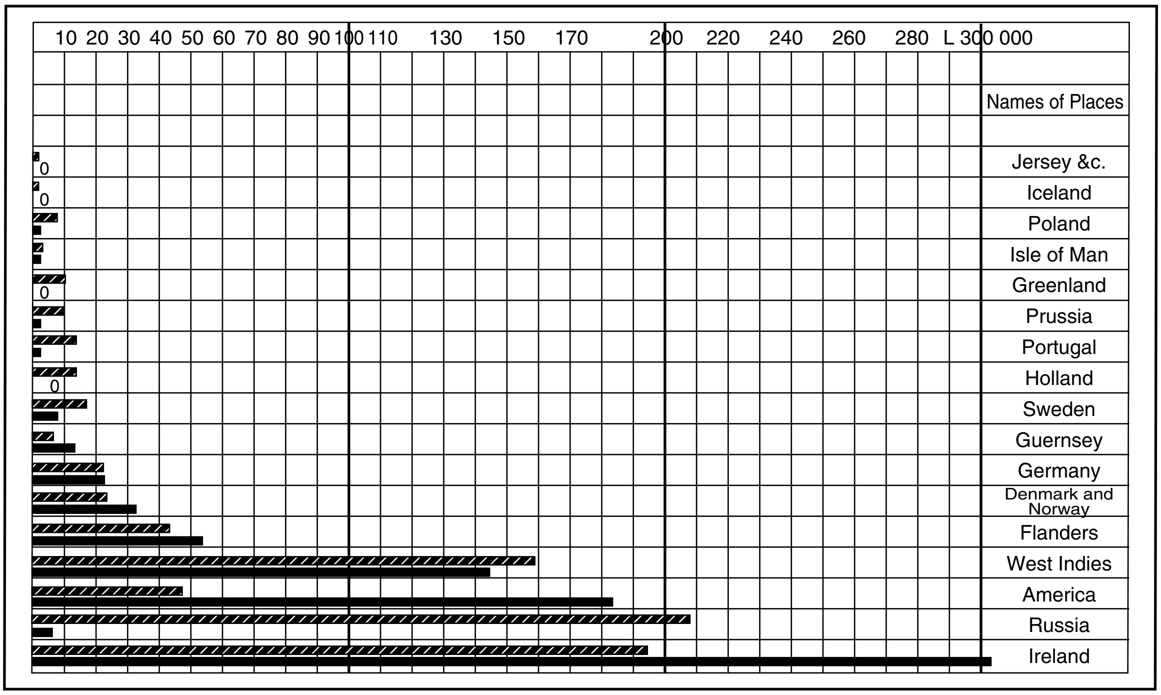
published a timeline of the birth and death of civilizations that begins at the time of the Great Flood (2344 BC – indicating clearly, though, that this was 1656 years after The Creation) and continued until 1780. And in 1782, the Scottish minister James Playfair (unrelated to William), published *A System of Chronology*, in the style of Priestley.

The motivation behind the drafting of graphical representations of longitudinal data remains the same today as it was in eighteenth-century France. Dubourg declared that history has two ancillary fields: geography and chronology. Of the two he believed that geography was the more developed as a means for studying history, calling it ‘lively, convenient, attractive.’ By comparison, he characterizes chronology as ‘dry, laborious, unprofitable, offering the spirit a welter of repulsive dates, a prodigious multitude of numbers which burden the memory.’ He believed that by wedding the methods of geography to the data of chronology he could make the latter as accessible as the former. Dubourg’s name for his invention *chronographie* tells a great deal about what he intended, derived as it is from the Greek chronos (time) and grapheikos (writing). Dubourg intended to provide the means for chronology to be a science that, like geography, speaks to the eyes and the imagination, ‘a picture moving and animated.’

Joseph Priestley used his line chart to depict the life spans of famous figures from antiquity; Pythagoras, Socrates, Pericles, Livy, Ovid, and Augustus, all found their way onto Priestley’s plot. Priestley’s use of this new tool was clearly in the classical tradition.

Twenty-one years later, William Playfair used a variant on the same form (See Figure 4) to show the extent of imports and exports of Scotland to 17 other places. Playfair, as has been amply documented (Spence & Wainer, [8]), was an iconoclast and a versatile borrower of ideas who could readily adapt the chronological diagram to show economic data; in doing so, he invented the bar chart. Such unconventional usage did not occur to his more conservative peers in Great Britain, or on the

Exports and imports of Scotland to and from different parts for one year from Christmas 1780 to Christmas 1781



The Upright Divisions are ten thousand pounds each. The black lines are Exports, the ribbed lines, imports.

Published as the Act directs June 7th1786 by W$^{m.}$ Playfair *Neele sculp^{t} 352 strand London*

Figure 4 Imports from and exports to Scotland for 17 different places (after Playfair, [5], plate 23)

Continent. He had previously done something equally novel when he adapted the line graph, which was becoming popular in the natural sciences, to display economic time series. However, Playfair did not choose to adapt Priestley's chronological diagram because of any special affection for it, but rather of necessity, since he lacked the time-series data he needed to show what he wanted. He would have preferred a line chart similar to the others in his *Atlas*. In his own words,

> 'The limits of this work do not admit of representing the trade of Scotland for a series of years, which, in order to understand the affairs of that country, would be necessary to do. Yet, though they cannot be represented at full length, it would be highly blameable entirely to omit the concerns of so considerable a portion of this kingdom.'

Playfair's practical subject matter provides a sharp contrast to the classical content chosen by Priestley to illustrate his invention.

In 1787, shortly after publishing the *Atlas*, Playfair moved to Paris. Thomas Jefferson spent five years as ambassador to France (from 1784 until 1789). During that time, he was introduced to Playfair personally Donnant [2], and he was certainly familiar with his graphical inventions. One of the most important influences on Jefferson at William and Mary College in Virginia was his tutor, Dr. William Small, a Scots teacher of mathematics and natural philosophy – Small was Jefferson's only teacher during most of his time as a student. From Small, Jefferson received both friendship and an abiding love of science. Coincidentally, through his friendships with James Watt and John Playfair, Small was responsible for introducing the 17-year-old William Playfair to James Watt, with the former serving for three years as Watt's assistant and draftsman in Watt's steam engine business in Birmingham, England.

Although Jefferson was a philosopher whose vision of democracy helped shape the political structure of the emerging American nation, he was also a farmer, a scientist, and a revolutionary whose feet were firmly planted in the American ethos. So it is not surprising that Jefferson would find uses for graphical displays that were considerably more down to earth than the life spans of heroes from classical antiquity. What is surprising is that he found time, while President of the United States, to keep a keen eye on the availability of 37 varieties of vegetables in the Washington market and compile a chart of his findings (a detail of which is shown in Figure 5).

When Playfair had longitudinal data, he made good use of them, producing some of the most beautiful and informative graphs of such data ever made. Figure 6 is one remarkable example of these. Not only is it the first 'skyrocketing government debt'

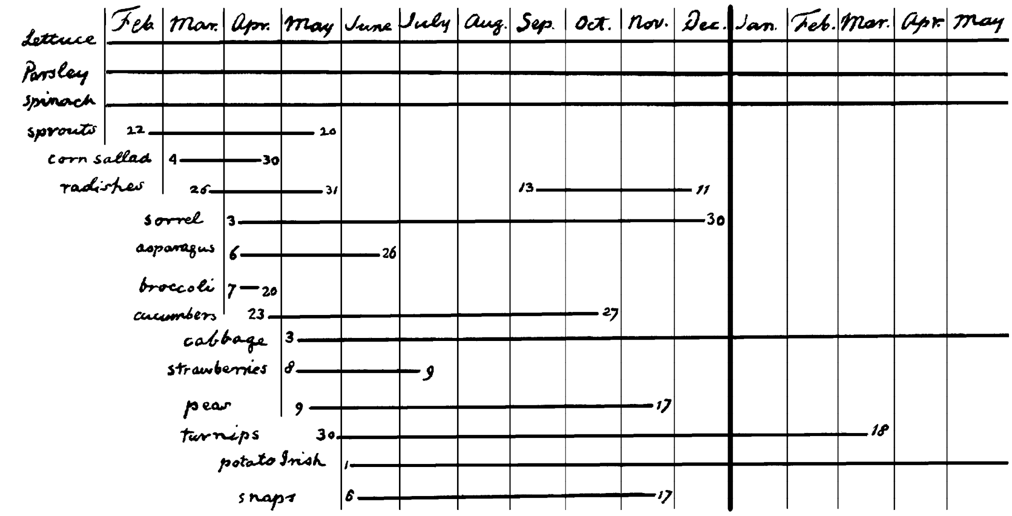


Figure 5 An excerpt from a plot by Thomas Jefferson showing the availability of 16 vegetables in the Washington market during 1802. This figure is reproduced, with permission, from Froncek ([3], p. 101)

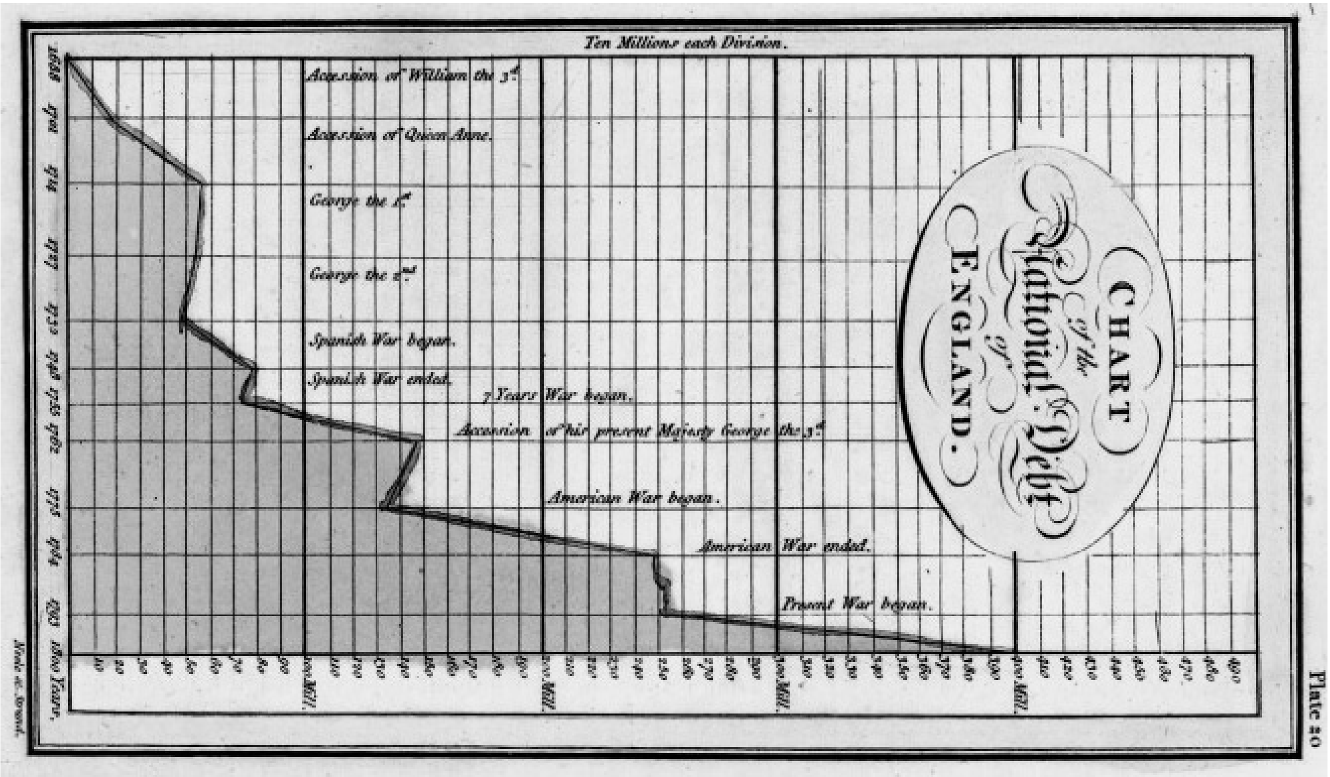


Figure 6 This remarkable ‘Chart of the National Debt of England’ appeared as plate 20, opposite page 83 in the third edition of Playfair’s *Commercial and Political Atlas* in 1801

chart but it also uses the innovation of an irregularly spaced grid along the time axis to demark events of important economic consequence.

Modern Developments

Recent developments in displaying longitudinal data show remarkably few modifications to what was developed more than 200 years ago, fundamentally because Playfair got it right. Modern high-speed computing allows us to make more graphs faster, but they are not, in any important way, different from those Playfair produced. One particularly useful modern example (Figure 7) is taken from Diggle, Heagerty, Liang & Zeger ([1], p. 37–38), which is a hybrid plot combining a scatterplot with a line drawing. The data plotted are the number of CD4+ cells found in HIV positive individuals over time. (CD4+ cells orchestrate the body's immunoresponse to infectious agents. HIV attacks this cell and so keeping track of the number of CD4+ cells allows us to monitor the progress of the disease.) Figure 7 contains the longitudinal data (*see* **Longitudinal Data Analysis**) from 100 HIV positive individuals over a period that begins about two years before HIV was detectable (seroconversion) and continues for four more years. If the data were to be displayed as a **scatterplot**, the time trend would not be visible because we have no idea of which points go with which. But (Figure 7(a)) if we connect all the dots together appropriately, the graph is so busy that no pattern is discernable. Diggle et al. [1] propose a compromise solution in which the data from a small, randomly chosen, subset of subjects are connected (Figure 7(b)). This provides a guide to the eye of the general shape of the longitudinal trends. Other similar schemes are obviously possible: for example, fitting a function to the aggregate data and connecting the points for some of the residuals to look for idiosyncratic trends.

A major challenge of data display is how to represent multidimensional data on a two-dimensional surface (*see* **Multidimensional Scaling**; **Principal**

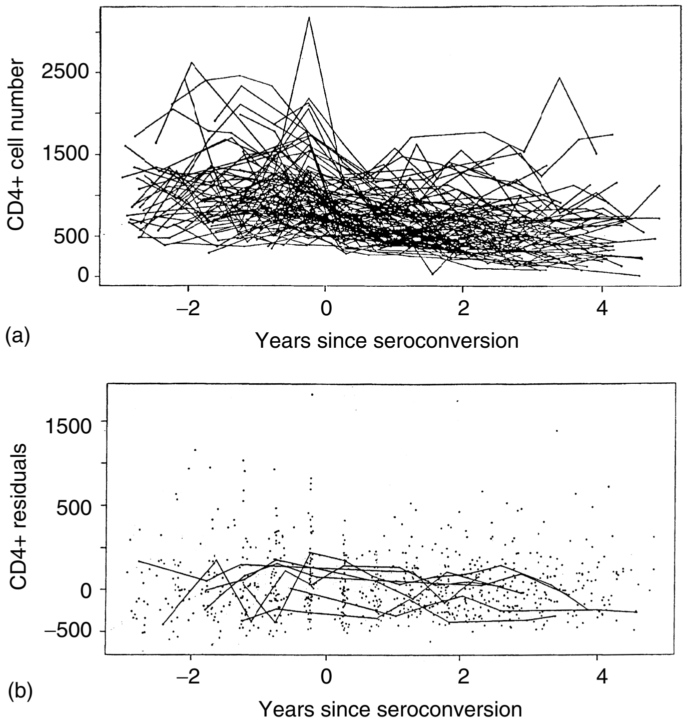


Figure 7 Figures 3.4 and 3.5 from Diggle et al., [1] – reprinted with permission (p. 37–38), showing CD4+ counts against time since seroconversion, with sequences of data on each subject connected (a) or connecting only a randomly selected subset of subjects (b)

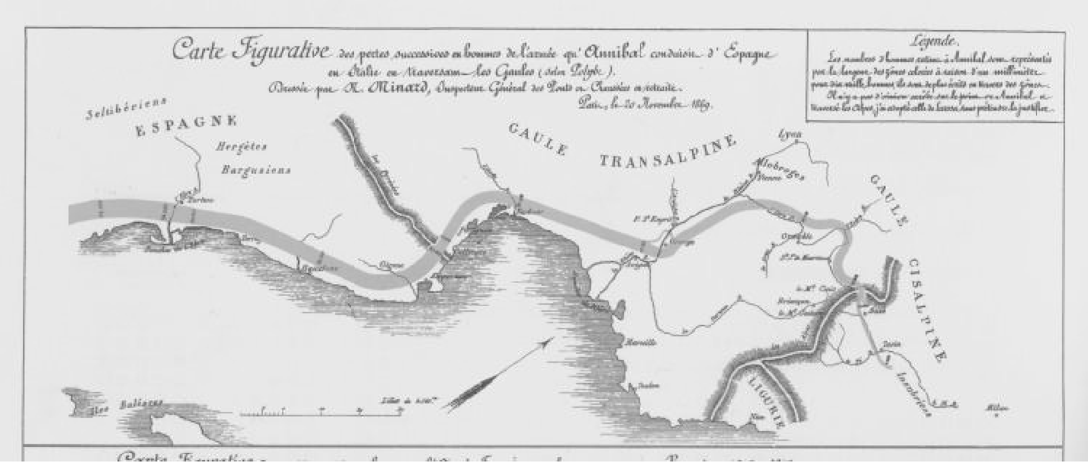

Carte Figurative des pertes successives en hommes de l'armée qu'Annibal conduisit d'Espagne en Italie en traversant les Gaules (selon Polybe).
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Légende.
Seltibériens
ESPAGNE
Hergètes
Barguasiens
GAULE TRANSALPINE
Lyon
Allobroges
GAULE CISALPINE
LIGURIE
Marseille
Iles Baléares

Figure 8 An 1869 plot by Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard, 1845–1869* depicting the size of Hannibal's Army as it crossed from Spain to Italy in his ill-fated campaign in the Second Punic War (218–202 BC). A portfolio of Minard's work is held by the Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris. This figure was reproduced from Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut © 1983, 2001), p. 176. with permission

Component Analysis). When longitudinal data are themselves multivariate (*see* **Multivariate Analysis: Overview**), this is a problem that has few completely satisfying solutions. Interestingly, we must look back more than a century for the best of these. In 1846, the French civil engineer Charles Joseph Minard (1781–1870) developed a format to show longitudinal data on a geographic background. He used a metaphorical data river flowing across the landscape tied to a timescale. The river's width was proportional to the amount of materials being depicted (e.g., freight, immigrants), flowing from one geographic region to another. He used this almost exclusively to portray the transport of goods by water or land. This metaphor was employed to perfection in his 1869 graphic (Figure 8), in which, through the substitution of soldiers for merchandise, he was able to show the catastrophic loss of life in Napoleon's ill-fated Russian campaign. The rushing river of 4 22 000 men that crossed into Russia when compared with the returning trickle of 10 000 'seemed to defy the pen of the historian by its brutal eloquence.' This now-famous display has been called (Tufte, [9]) 'the best graph ever produced.' Minard paired his Napoleon plot with a parallel one depicting the loss of life in the Carthaginian general Hannibal's ill-fated crossing of the Alps in the Second Punic War. He began his campaign in 218 BC in Spain with more than 97 000 men. His bold plan was to traverse the Alps with elephants and surprise the Romans with an attack from the north, but the rigors of the voyage reduced his army to only 6000 men. Minard's beautiful depiction shows the Carthaginian river that flowed across Gaul being reduced to a trickle by the time they crossed the Alps. This chart has been less often reproduced than Napoleon's march and so we prefer to include it here.

Note

1. This exposition is heavily indebted to the scholarly work of Sandy Zabell, to whose writings the interested reader is referred for a much fuller description (Zabell, [11, 12]). It was Zabell who first uncovered Arbuthnot's clerical error.

References

[1] Diggle, P.J., Heagerty, P.J., Liang, K.Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd Edition, Clarendon Press, Oxford, pp. 37–38.

[2] Donnant, D.F. (1805). *Statistical account of the United States of America.* Messrs Greenland and Norris, London.

[3] Froncek, T. (1985). *An Illustrated History of the City of Washington*, Knopf, New York.

[4] Herschel, J.F.W. (1833). On the investigation of the orbits of revolving double stars, *Memoirs of the Royal Astronomical Society* **5**, 171–222.

[5] Playfair, W. (1786). *The Commercial and Political Atlas*, Corry, London.

[6] Priestley, J. (1765). *A Chart of History*, London.

[7] Priestley, J. (1769). *A New Chart of History*, London. Reprinted: 1792, Amos Doolittle, New Haven.

[8] Spence, I. & Wainer, H. (1997). William Playfair: a daring worthless fellow, *Chance* **10**(1), 31–34.

[9] Tufte, E.R. (2001). *The Visual Display of Quantitative Information*, 2nd Edition, Graphics Press, Cheshire.

[10] Wainer, H. & Velleman, P. (2001). Statistical graphics: mapping the pathways of science, *The Annual Review of Psychology* **52**, 305–335.

[11] Zabell, S. (1976). Arbuthnot, Heberden and the Bills of Mortality, Technical Report #40, Department of Statistics, The University of Chicago, Chicago, Illinois.

[12] Zabell, S. & Wainer, H. (2002). A small hurrah for the black death, *Chance* **15**(4), 58–60.

Further Reading

Ferguson, S. (1991). The 1753 *carte chronographique* of Jacques Barbeu-Dubourg, *Princeton University Library Chronicle* **52**(2), 190–230.

Playfair, W. (1801). *The Commercial and Political Atlas*, 3rd Edition, John Stockdale, London.

Spence, I. & Wainer, H. (2004). William Playfair and the invention of statistical graphs, in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, ed., Academic Press, San Diego.

Wainer, H. (2005). *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*, University Press, Princeton.

HOWARD WAINER and IAN SPENCE