The Milestones Project: A Database for the History of Data Visualization

Michael Friendly* Matthew Sigal Derek Harnanansingh

November 25, 2012

Abstract

Methods of data visualization have evolved substantially over their history. Some landmarks in this story were the first thematic maps in the 1600s, the invention of the bar chart and line graph in the early 1800s, and the dynamic and interactive graphics of today. While these developments have been previously detailed in various written micro-histories, there has never been an attempt to collect a complete, macro-history in a single place for study, search or query, and even data analysis or graphics based on this history.

The purpose of this chapter is threefold: first, to introduce the reader to our solution: an online resource called the Milestones Project. This web site details important events in the history of data visualization, and enables users to interactively travel through time to see and explore the context that surrounded their developments. Secondly, we present some striking visual examples that deal with conveying aspects of history over time, drawn from this resource.

Finally, the Milestones database will be used to showcase how such a resource can serve as "data" for *statistical historiography*, which entails the use of statistical and graphical methods for the analysis and understanding of historical innovations, developments, and trends.

1 Introduction

If you would understand anything, observe its beginning and its development. —Aristotle

Questions regarding the history of data visualization are (or at least should be) of great importance to historians of science, to current developers of graphical methods for statistical analysis and the related info-vis community, as well those just interested in the history of ideas. In the history of science, diagrams, graphs, maps and other visualizations have often played important roles in discoveries that arguably might not have been achieved otherwise.[1] At the same time, in the fields of statistical graphics and information visualization, developers often create "new" methods without any appreciation that they have deep roots in the past.[2]

*This work was supported by Grant OGP0138748 from the National Sciences and Engineering Research Council of Canada to Michael Friendly. We are grateful to Dan Denis, Antoine de Falguerolles, Stephen Stigler, Ben Shneiderman and Howard Wainer for constructive comments on this chapter.

[1]Some salient examples are: Francis Galton's 1861 discovery of anti-cyclonic movement of wind around low-pressure areas from contour maps; Edward Maunder's "butterfly diagram" of the variation of sunspots over time leading to the discovery of the "Maunder minimum," from 1645–1715; and Henry Moselely's 1913 discovery of the concept of atomic number, based largely on graphical analysis (a plot of serial numbers of the elements vs. square root of frequencies from their X-ray spectra).

[2]For example, mosaic displays for frequency tables were thought to have been invented by Hartigan and Kleiner (1981) and extended to show the pattern of residuals in loglinear models by Friendly (1994). But it turns out that the essential idea behind this area-based display goes back to Georg von Mayr in 1877 (Friendly, 2002a).

These two perspectives provided the motivation for the development of the Milestones Project. This stemmed from the fact that historical accounts of events, ideas and techniques that relate *inter alia* to modern data visualization were fragmented and scattered across a wide number of fields.[3]

When this work began in the mid-1990s, there were no accounts or resources that spanned the entire development of visual thinking and the visual representation of data across different disciplines and perspectives. The Milestones Project began simply as an attempt to collate these diverse contributions into a single, comprehensive listing, organized chronologically, that contained representative images, references to original sources, and links to further discussion— a source for "one-stop shopping" on the history of data visualization.

In Section 2, we describe the evolution and structure of the Milestones Project. Section 3 presents some historical and modern approaches to one self-referential question: how can data visualization be applied to its own history? Section 4 introduces another self-referential topic we call *statistical historiography*, which entails the use of statistical and graphical methods for the analysis and understanding of historical innovations, developments, and trends. But first we give some brief vignettes of historical topics and questions for which the Milestones Project has proved invaluable in our own research.

1.1 The first statistical graph

In the history of statistical graphics (Friendly, 2008a), as in other artful sciences, there are a number of inventions and developments that can be considered "firsts" in these fields. The catalog of the Milestones Project (Friendly and Denis, 2001) lists 70 events that can be considered to be the initial use or statement of an idea, method or technique that is now commonplace, but there is probably no question more fundamental than that of the first visual representation of statistical data.

In Friendly *et al.* (2010) we argue that the 1-dimensional line graph shown in Figure 1 by Michael Florent van Langen (van Langren, 1644) should be accorded this honour. The graph shows 12 estimates of the distance in longitude between Toledo and Rome, overlaid on a modern map. Van Langren used this to demonstrate that these estimates were all subject to large errors and to propose to King Phillip of Spain that only *he* had a sufficiently precise method for the determination of longitude for navigation at sea.

The telling of van Langren's story not only turned out to involve astronomy, archival research, the history of patronage in the 17^{th} century, and even an unsolved problem of cryptography, but also serves as an example of statistical historiography. The Milestones Project provided the infrastructure for this research – through the use of a time-based, cross-referenced catalog of images, references and links to related work, van Langren's tale was able to be studied and reported upon.

1.2 Who invented the scatterplot?

Although there are earlier precursors, the main graphical methods used today— pie charts, line graphs and bar charts— are generally attributed to William Playfair in works around the beginning of the 19^{th} century (Playfair, 1786, 1801). All of these are essentially univariate displays of some aspect of a single variable.

A logical next step would be to invent a method to reveal the relationship between two variables— what we now know as the scatterplot. By 1886, Francis Galton had utilized this truly bivariate display,

[3] Among these are general histories in the fields of probability (Hald, 1990), statistics (Pearson, 1978, Porter, 1986, Stigler, 1986), astronomy (Riddell, 1980), cartography (Wallis and Robinson, 1987). More specialized accounts focus on the early history of graphic recording (Hoff and Geddes, 1959, 1962), statistical graphs (Funkhouser, 1936, 1937, Royston, 1970, Tilling, 1975), fitting equations to empirical data (Farebrother, 1999), cartography (Friis, 1974, Kruskal, 1977), thematic mapping (Friendly and Palsky, 2007, Palsky, 1996, Robinson, 1982), and so forth.

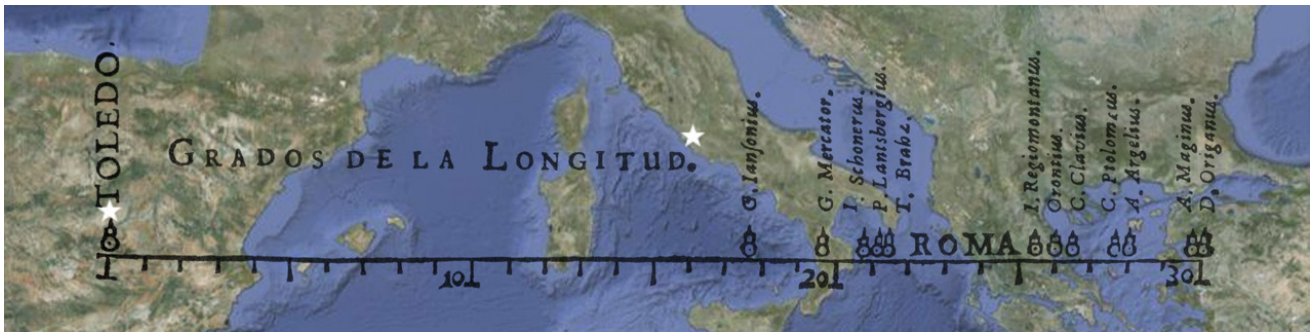

Figure 1: van Langren's 1644 graph, re-scaled and overlaid on a modern map of Europe. Toledo is located at lat/long (+39.86°N, −4.03°W), Rome is located at (+41.89°N, +12.5°W), both shown by markers (stars) on the map. This image makes clear what van Langren wished to communicate: the wide variability of the estimates, but also shows how far the estimates were biased.

which led to the discovery of correlation and regression, and ultimately to much of present multivariate statistics. However, he was not the first to use this graphical technique, and it is surprising that no one is widely credited with its invention.

In Friendly and Denis (2005), we delved into this mystery. This involved tracing the early origins of ideas related to the scatterplot, which led to two compelling narratives: how, in Playfair's time, it was nearly impossible to think about and visualize bivariate relationships; and, later, how the scatterplot was essential for Galton's visual insights that would lead to the rise of modern statistics and graphics. It was the resources available in the Milestones Project that allowed us to focus upon the events in this period and attribute the essential ideas of the scatterplot to J. F. W. Herschel in two 1832 papers.

1.3 The Golden Age of statistical graphics

In our initial web presentation of the Milestones Project, it proved convenient to sub-divide the history of data visualization into epochs, each of which turned out to be describable by coherent themes. As we illustrate later, one period turned out to be particularly noteworthy, both for the sheer number of contributions, and for the beauty and elegance of their execution. We call this period, from roughly 1850 to 1900 (±10), the Golden Age of statistical graphics (Friendly, 2008b).

Figure 2 shows the time distribution of the 260 significant events that had been included in the Milestones Project database by 2007, demarcated by the labels we used for epochs. In Friendly (2008b), we traced the origin of this period in terms of the infrastructure required to produce such an explosive growth of contributions to data visualization, and found three primary sources: the systematic data collection by state agencies, the rise in popularity of statistical and visual thinking, and the enabling developments of technological innovations.

2 The Milestones Project

Direction is more important than speed. We are so busy looking at our speedometers that we forget the milestone. —Anonymous

An early overview of the content and aims of the Milestones Project appeared in Friendly (2005). Here we update that description and provide a few technical details on some problems that were encountered in attempting to make the history of data visualization convenient for collecting, browsing, searching, and analysis.

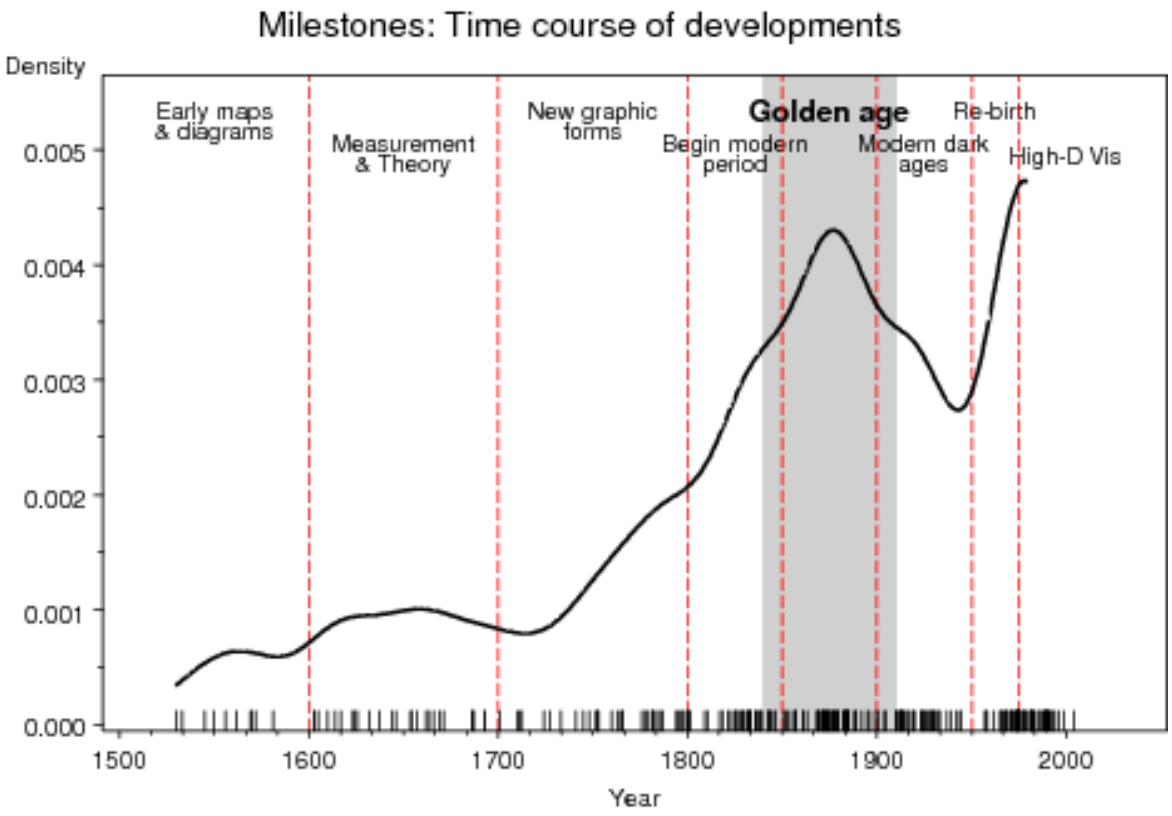

Figure 2: The time distribution of events considered milestones in the history of data visualization, shown by a rug plot and density estimate. The density estimate is based on $n = 260$ significant events in the history of data of data visualization from 1500–present. The developments in the highlighted period, from roughly 1840–1910 comprise the Golden Age of statistical graphics.

2.1 Origin, structure and evolution

The initial step in portraying the history of data visualization was to create a simple chronological listing of milestone items with capsule descriptions, bibliographic references, markers for date, person, place, and links to portraits, images, related sources and more detailed commentaries. The initial database contained the 105 developments listed by Beniger and Robyn (1978), and incorporated additional records from Hankins (1999), Tufte (1983, 1990, 1997), Heiser (2000), among others.

This began as a single LaTeX file (with markup tags for all relevant bits of information), used to produce a hyper-linked PDF document. A variety of software tools (perl scripts, Unix utilities) allowed us to turn this single source *directly* into the web version originally shown at `http://www.math.yorku.ca/SCS/Gallery/milestone`. Other custom software tools allowed us to add new milestones items from text files using a template of tags (DATE:, AUTHOR:, WHAT:, REF:, IMG:, etc.) and extract the information about milestones items, authors, images, etc. in a variety of forms (CSV, XML, JSON) that could be used as input for analyses and graphic displays. For example, Figure 2 was produced in SAS software by piping the output of a latex to csv translator:

```
itemdb -o milestones.csv < milestones.tex | sas -i milestones.csv mileyears.sas
```

It soon became apparent that such a text-based representation was inadequate. Updating the milestones data required that the LaTeX file be shared among several collaborators; milestone assets, such as images, web links and references, were not easily accessible by others, which made collaboration

cumbersome. Further, each update to the web site required an inefficient number of steps of verification, re-building, and synchronization with the server, meaning the website was often out of date.

Around 2005, we began to make the process into a more dynamic one; to convert the flat file into a relational database; create a Milestones administrative system and completely redesign the Milestones web site. Specifically, we wanted to facilitate contributions by any number of trusted collaborators via an easy-to-use web administration area, and allow for the dissemination of milestones data via an easy-to-browse public user interface.

Migrating the data to this format provided some challenges. First, the existing milestones data needed to be restructured logically and have redundancy minimized. To do this, we partitioned the data into its relevant entities: namely the milestone itself, and its descriptors, such as its aspect, author, subject, keywords, reference, and linked media items (such as images). The aspect, author, subject, keyword, and reference descriptors exist as a many-to-many relationship between it and the milestone. For example, an aspect can belong to one or more milestones, and the milestone can belong to one or more aspects. Media items, on the other hand, can only belong to one milestone at a time, with multiple media items possible for a single milestone. Figure 3 illustrates these relationships.

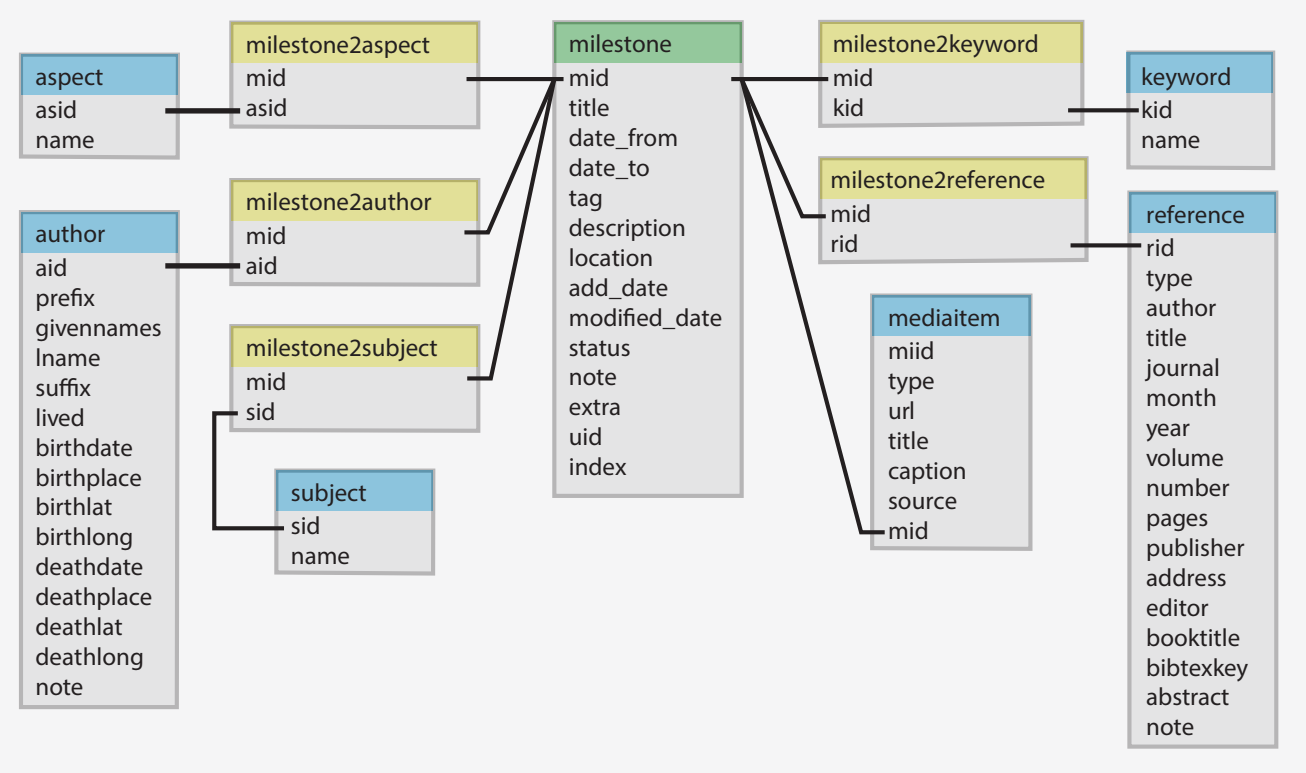

Figure 3: Simplified schema for the MySQL database for the Milestones Project. The main table (`milestone`) contains information regarding each of the items considered a milestone in the history of data visualization, linked to other tables (e.g., `reference`, `mediaitem`) by unique (primary) keys. Other supporting tables (e.g., `milestone2aspect`) provide for convenient lookups of descriptors of these milestones items (`subject`, `aspect`, `keyword`).

Normalizing the data in this way enabled us to free the database of modification anomalies; ensured that the database structure was scalable, and could be extended with minimum modifications. Most importantly, it allows for future growth, and provides a query-neutral database model (Codd, 1971) that could be used to power web presentation, and customized indexed search. The last major benefit, which will be demonstrated in Section 3, is that this schema allows for any type of analysis of the Milestones

data itself.

At present, the Milestones Project documents 288 contributions, with nearly 350 references, information on 336 authors, and 774 media items, made up of 371 images appearing online on the `http://datavis.ca/milestone` site, and 403 hyperlinks to images and documents that are externally hosted. In addition, we maintain an offline image database comprising over 1,100 images collected from various sources. Over time, these too will be incorporated into the database.

2.2 User interface

The second challenge related to how to display such a large amount of information in an easy-to-use interface that would provide overview, search, and details about these events in the history of data visualization. We decided to retain the time-based grouping of the milestones content by epochs (Pre-1600, 1600s, 1700s, etc.), each with a theme (e.g., 1600–1699: Measurement and Theory) and descriptive text. The visual design of the interface adopts Ben Shneiderman's mantra: "Overview first, zoom and filter, then details on demand" (Shneiderman, 1996). To do this, we added a timeline view (Figure 4) of the milestones items displayed on the overview landing page.

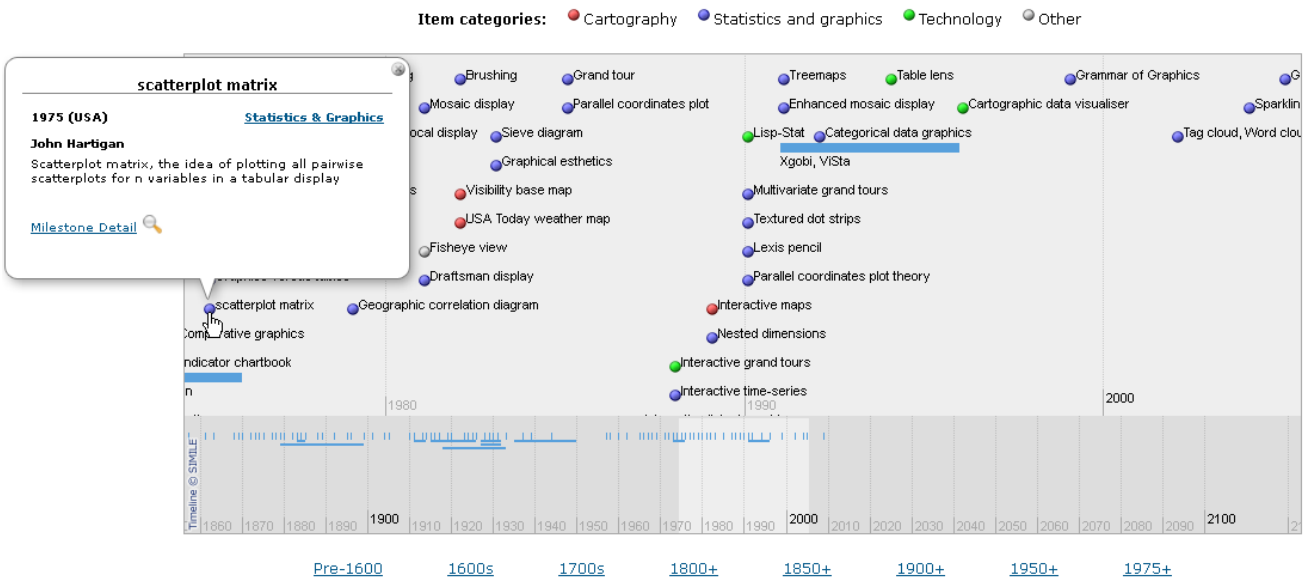

Figure 4: Timeline view of the Milestones Project on the site `http://datavis.ca/milestone`. In this view, the top panel shows a detailed view of the segment of history highlighted in the bottom panel, both of which can be separately scrolled. Items in the top panel show a brief text tag, colour-coded by category. Clicking on an item in this panel brings up a small description, which is further linked to the details of the milestone item.

This timeline, based on the SIMILE Timeline Widget (`http://www.simile-widgets.org/timeline`) allows multiple connected time bands, showing events at different resolutions. Each band can be separately panned by dragging left or right with the mouse pointer, scroll wheel, or keyboard arrow keys. The timeline view, although most obvious, is just one of several possibilities for a visual overview or interaction with the display of the milestones database. The software design of the site, using open-source tool kits, makes it relatively simple to add new ones. For example, the database can also be navigated via a list view (with drop down quick links), and in Section 4.3 we will illustrate how it can be explored using a map-based display.

3 Visualizing Time and History

What does history look like? How do you draw time? —Rosenberg and Grafton (2010, p. 10)

The questions in this quotation introduce an important topic in the history of data visualization: how can such a history be visualized? What methods might be called upon to detail the richness of its past?[4] Time provides an obvious dimension, but what else could be included in a static display that might reveal a story previously hidden? What kinds of dynamic or interactive displays might fascinate and intrigue viewers?

An annotated visual gallery of some timeline designs and visual histories can be found in our Data Visualization Gallery at `datavis.ca/gallery/timelines.php`. The topics covered include early visual histories, encyclopedic charts, special purpose charts, correlated histories showing events in one domain in the context of events in other areas, non-linear scales for time and space, as well as dynamic, interactive timelines. Here we present a few inventive selections from this scholarship.

3.1 The first timelines, reconsidered

Although there are earlier precursors, the first timelines of modern design— featuring a horizontal, linear axis for time, and vertical positions for place, theme or category of events— were produced in the mid 1700s. Most notable of these prototypes were Jacques Barbeau-Doubourg's 1753 *Carte Chronologique*, and Joseph Priestley's 1765 *Chart of Biography*.

Priestley first published a small "Specimen" of this chart as a proof-of-concept, showing the lifespan of famous men in the years 600 BC to 0 AD, classified as "statesmen" (from Solon to Augustus) and "men of learning" (from Thales to Ovid). Later that year, Priestly published a detailed version 1765 that quickly became the most popular and influential timeline of the 19^{th} century. The full graphic details the lifespans of more than 2,000 people from 1200 BC to 1750 AD, classified by their areas of achievement (statesmen & warriors, mathematicians & physicians, artists & poets, and so on).

Priestly's timeline charts can be seen on our Data Visualization Gallery, and we don't reproduce them here. Instead we show (see Figure 5) a re-design, in his style, of the lifespans of 79 authors from the Milestones database who were born in France or the United Kingdom between 1500 and 2000.

Rosenberg and Grafton (2010, p. 117) called Priestley's charts "masterpieces of visual economy." Indeed, they were at the time. However, in his charts, the famous people were arranged haphazardly within category groups, so it is difficult to find specific individuals, and nearly impossible to uncover any trends, either over time or across categories.

In our version, authors are sorted by birth year within each country and the names are printed alternately at the year of birth and death. The result, which resembles a cumulative distribution plot: (a) allows easier visual lookup of names, (b) provides an overall "lifespan envelope," and, (c) highlights a few individuals who lived conspicuously shorter or longer than their contemporaries (e.g., shorter: Willam Jevons, James Maxwell, John Snow, and Phillipe Buache). Of course, to display lifespan *directly* requires a different kind of plot, but one that would not have been even thinkable by Priestley in 1765. We return to this question in Section 4.2 (see Figure 10).

3.2 Universal histories

In addition to unrivalled thematic maps and statistical diagrams, the Golden Age of statistical graphics also gave rise to a variety of novel attempts to visualize history in a comprehensive manner, combining

[4] Another recent book, *Visualizing Time* (Wills, 2012), discusses a range of modern graphical methods for visualizing time-based data.

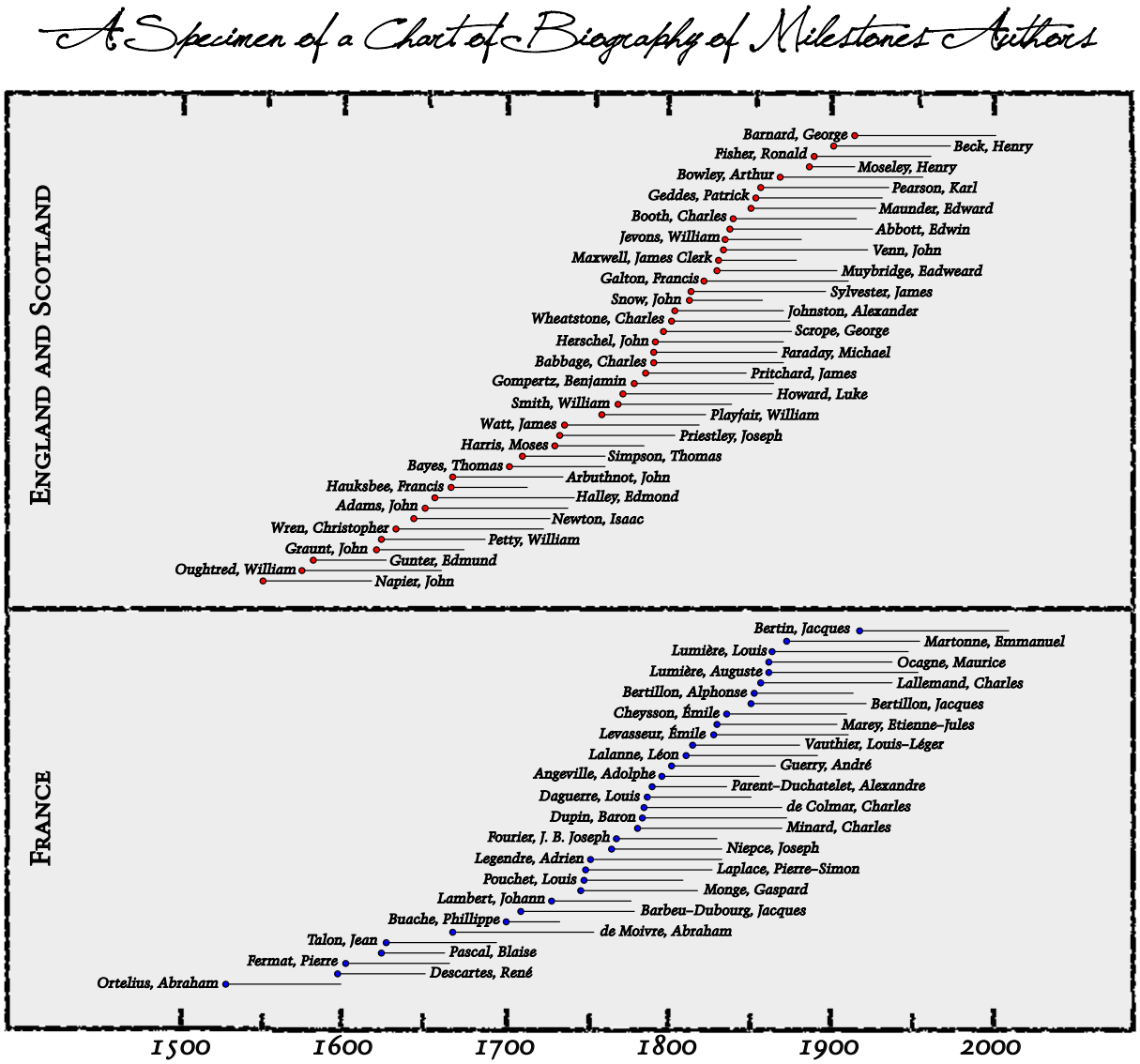

Figure 5: A modern re-design of Priestley's 1765 *Chart of Biography*, using information on authors in the Milestones database born in France or the United Kingdom. Authors are sorted within country by year of birth and labeled alternately at birth and death years, allowing better lookup and visual comparison.

parallel, intertwined time-flows, text, illustrations, maps, and other visual forms. Among the most impressive is the series of Synchronological Charts of Universal History produced by Sebastian Adams between 1871–1885. The 1881 version is 23 feet long and captures 5,885 years of history, from 4004 B.C. to 1881 A.D. Rosenberg and Grafton (2010, p. 172) call it "nineteenth-century America's surpassing achievement in complexity and synthetic power."

Figure 6 shows the entire chart at the top (note the increasing visual density towards the right) and a small portion below; The entire chart can be viewed in high-resolution at `http://www.davidrumsey.com/blog/2012/3/28/timeline-maps`. Adams used a linear scale for time, and so it is understandable why it took 23 linear feet to include all of recorded history.

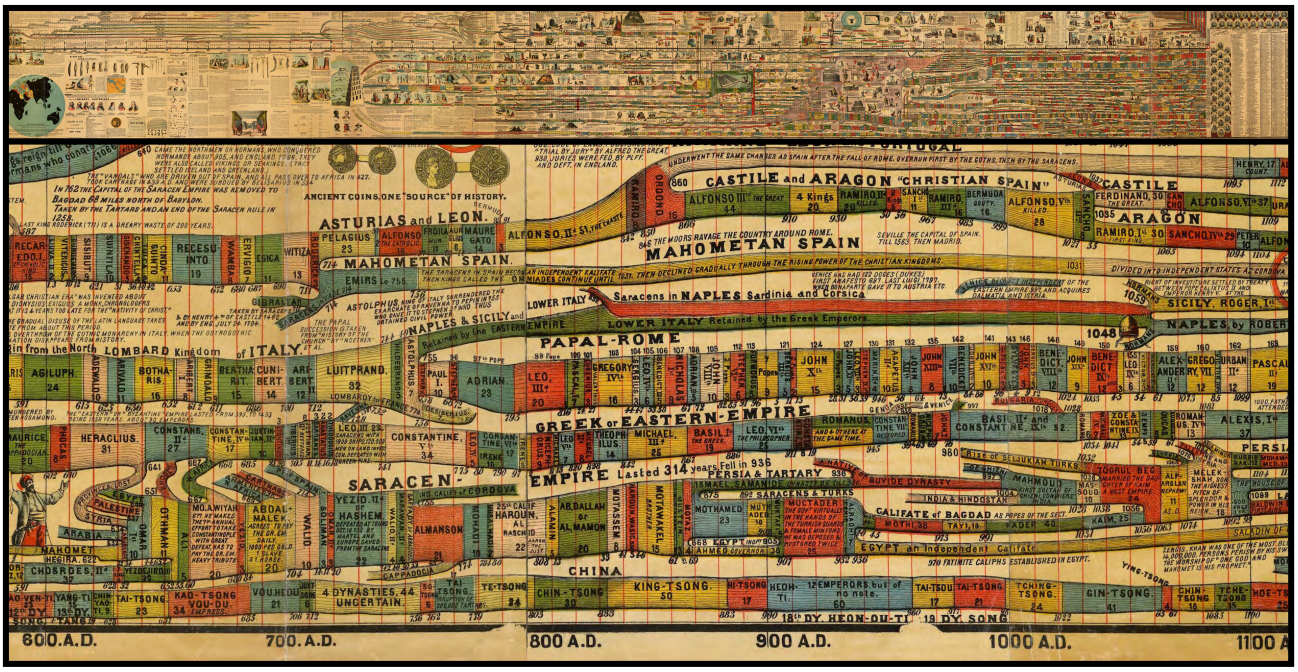

Figure 6: Top: The entirety of Sebastian Adams' *Synchronological Chart of Universal History*, 1881. Bottom: An excerpt detailing the 600–1100AD period. Horizontal bands trace developments in different countries, with detailed text describing significant events, and which break up or merge according to political factors.

3.3 Categorization and non-linear scales

Linear time scales have the advantage that they provide uniform resolution and detail across the entire time span, but events in time, or our interest in them are rarely uniformly distributed. As exemplified by the Milestones Project, most visual histories are rather sparse at their beginning and very crowded at their end. Utilizing non-linear scales can allow resolution to vary smoothly across the range, providing greater detail in regions of interest, which are most often the recent past.[5]

Figure 7 is a proof-of-concept sketch for something that a graphic artist could use as a starting point for a chart of the history of data visualization. It uses the events from the Milestones Project, categorized by two correlated factors: Subject area, in which the content has been categorized as dealing with human populations, physical properties of the world, or mathematics and statistics; and, the milestone's aspect or form, which has been categorized as dealing with cartography, graphs and diagrams, or technology.

To provide greater resolution for more recent events, we have used a reverse square-root scale going backward from the year 2000. Specifically, 'Year' on the horizontal time axis is actually plotted according to the formula $\text{Year}^\star = 2*(25-\sqrt{(2000-\text{Year})})$, giving the more pleasing result that the modern period 1800–2000 occupies about 60% of the scale, despite only comprising 40% of the range. This is conveyed visually by the spacing between tick marks on the X-axis.

[5]Of course, interactive graphics offer the possibity to vary resolution dynamically, by moving a "lens" across the display, as in a hyperbolic viewer.

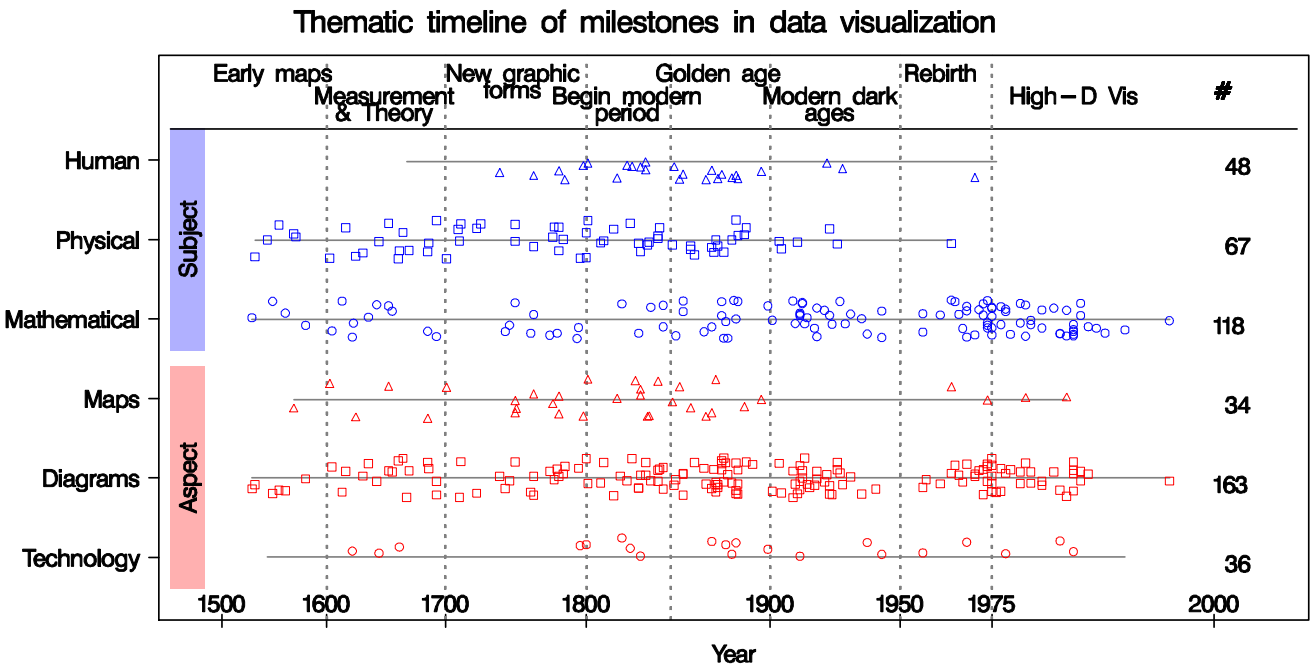

Figure 7: Sketch for a thematic timeline of milestones items, 1500–present, categorized by both the Subject (content) and Aspect (form) of the milestone item. To provide greater resolution for more recent events, time (Year) is shown on a square-root scale, going backward from the year 2000.

4 Using the Milestones Project for Statistical Historiography

Vision is the art of seeing things invisible. —Johnathan Swift, 1711

4.1 Statistical historiography

We use the term "statistical historiography," to refer to the use of statistical and graphical methods to explore, study and describe historical problems and questions.[6] This topic has a delightful self-referential quality when applied to the history of data visualization itself, since we have often found ourselves using modern methods of statistical analysis and graphics to study the development of ideas in this area. As in the quotation from Swift above, one goal is to make previously hidden aspects of this history visible.

At the same time, our examination of some of the most impressive graphic works of the past sometimes left us awe-struck by their exquisite beauty and visual design.[7] On more than one occasion when looking at these elegant presentations, we wondered whether there wasn't something lost with the advent of modern software. While we can now analyze massive data sets, and generate a multitude of graphics

[6] As far as we know, the initial expression of this idea appeared in a paper by Rubin (1943) discussing various ways in which statistical methods could be applied to historical topics. These included: the use of sampling methods to test historical theories; statistical distributions applied to historical data; and, the use of time series graphs with smoothed curves to study historical trends. More recently, many examples of the application of these ideas to statistical topics can be found in Stigler (1986, 1999), as well as our own papers on the history of data visualization, cited *inter alia*.

[7] Some examples are: Charles Joseph Minard's famous depiction of Napoleon's March on Moscow (Friendly, 2002b), Francis Galton's detailed study of weather patterns in Europe (see: Friendly, 2008b), and André-Michel Guerry's (Guerry, 1864, Plate 17) semi-graphic table depicting the relations of occurrence of crimes to a wide variety of social and demographic factors (see: Friendly, 2007).

with a simple mouse click, we still feel that designing a truly effective visual display of information requires thought and manual intervention.

For this reason, it is often quite instructive to attempt to re-create or even re-vision a graphic work from the past (Friendly, 2002b). We can learn from this undertaking an appreciation for the insight and hard labor of our graphic heros, and can sometimes better understand or improve on their designs by a process we call "understanding through reproduction," another facet of statistical historiography.

We illustrate this approach with an analysis of a graph from Playfair (1821), shown in Figure 8, in some ways a tour-de-force of early graphic presentation. In this graph, Playfair used three parallel time series in different forms to show the price of wheat (bar chart), weekly wages (line graph), and reigning monarch (intervals) over a ∼250 year span from 1565 to 1820. His graphic goal was rhetorical: he wished to argue that workers had become better off in the most recent years. Surely this must be counted among the best early data graphics.

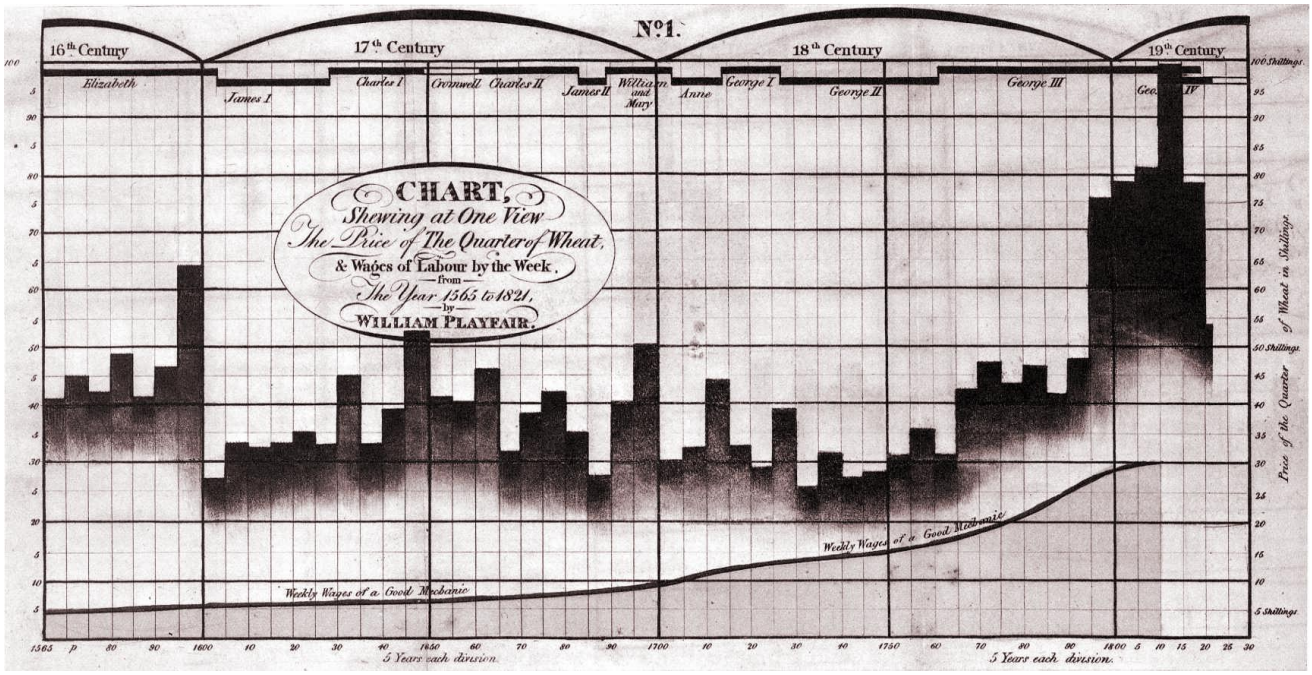

Figure 8: William Playfair's 1821 time series graph of prices, wages, and ruling monarch over a 250 year period. *Source*: Playfair (1821), image from Tufte (1983, p. 34)

Yet, as we have argued elsewhere (Friendly and Denis, 2005), this graph is both sinful and a communication failure for Playfair's purpose. It is sinful because the use of separate y axes for wages (left axis, range: 0–100) and prices (right axis, range: 0–30) on different scales provides the opportunity to tell very different stories simply by re-scaling one or both axes.

It is also a graphic failure because the visual impression that wages increased relative to prices toward the right end is at best indirect and is obscured by the large fluctuations in prices of wheat. What Playfair might have done to show the relation directly, is to plot the *ratio* of price to wages, representing the labor cost of a unit of wheat, as we have done in Figure 9. Adding a non-parametric (loess) smoothed curve to the line plot makes Playfair's message directly apparent, and also shows that the reduction in the amount of labor required to purchase one unit of week in fact levels off in the last 40 years.

However, in order to conduct such statistical historiography, there is one principal requirement: **data**. The Milestones Project database is the repository of all the information we have so far recorded,

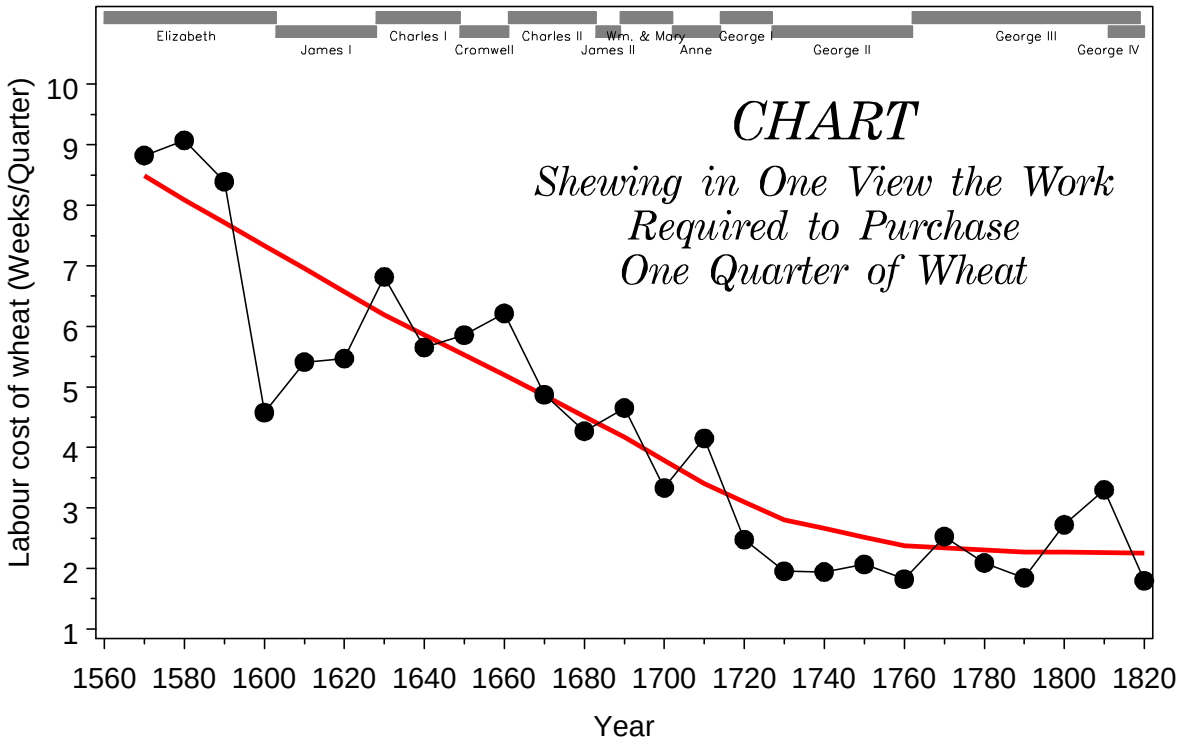

Figure 9: Redrawn version of Playfair's time series graph showing the ratio of price of wheat to wages, together with a loess smoothed curve.

and modern database tools allow the possibility of simple or complex queries, limited only by the available information.[8] In related work, we have collected and disseminated data sets of historical interest on a variety of topics in statistics and data visualization, for instance via the `R` packages HistData (Friendly, 2011) and Guerry (Friendly and Dray, 2010). These can be considered another source for data, pictures, and stories related to statistical historiography, and understanding through reproduction. This is the essence of the motto on the `datavis.ca` web site: *Looking back, going forward.*

In the subsections below, we describe a few applications of these ideas using the Milestones Project database and case studies that arose from this work. There is an interesting interplay between such historical analyses and these data collections. Some studies called for us to find and incorporate new data sources, such as our paper (Friendly, 2007) on Guerry's *Moral statistics of France* and the Guerry package, to which we added Angeville's extensive 1836 data on social and economic characteristics of France. In other cases, our analyses suggested new or different ways to visualize historical data.

4.2 Milestone authors: lifespan

As noted earlier, we record information relevant to the contributors of milestones events in an author table in the database. For most of these individuals, internet and biographical searches allowed us to determine the dates and places of their birth and death.

One simple question that can be posed using this information is how long did these contributors live? As illustrated earlier (see Figure 5), Joseph Priestley was the first to develop the idea of using

[8]It should be noted that, beyond the basics of recording milestones items, images and references, inputting the other meta-data (content and form categories, keywords, etc.) is highly labor-intensive. Thanks are due to many research assistants and graduate students who have and continue to work on the Milestones Project, including Dan Denis, Matt Dubins, Yvonne Lai, Avi Lipton, and Carolina Patryluk.

a graphic representation to show the lifespan of famous men. His "charts of biography" did this in a particularly evocative form, representing each person by a line segment whose length was defined by the individual's lifespan, and then grouped by occupational category.

These "timespan" charts tell an interesting story, but they do not provide an answer to the question of how long, in general, these individuals lived. However, with the author table from the Milestones Project, it is a simple matter to calculate lifespan, and obtain a direct answer to this query. Figure 10 shows one display of this information, using a combined density plot and rug plot, similar to the one used in Figure 2.

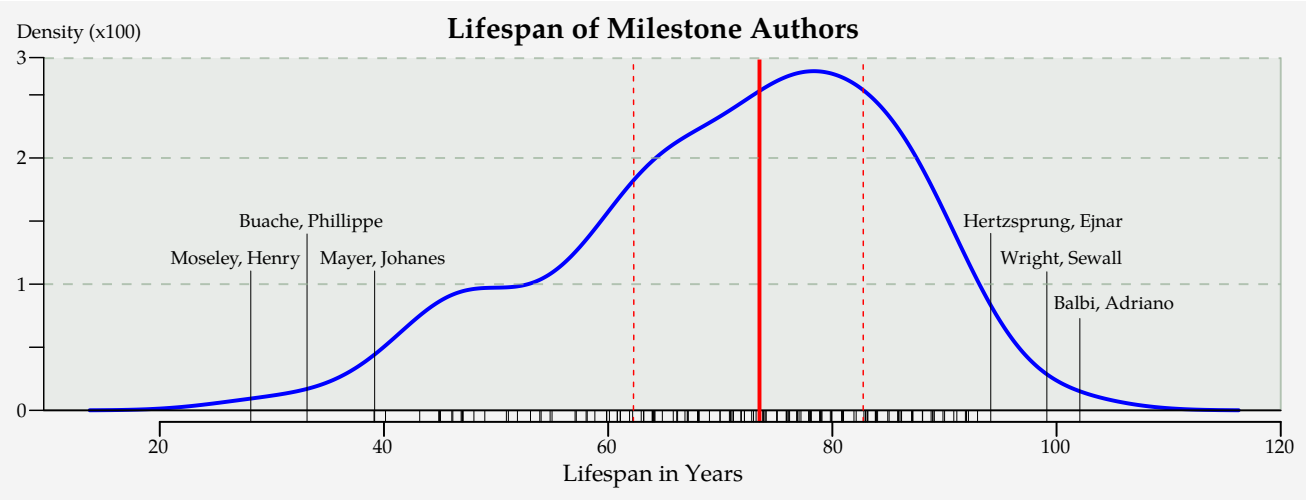

Figure 10: Density plot of the lifespan of the 172 authors in the Milestones Project database who were born after 1500 and for whom lifespan can be determined. Individual observations are shown by a (jittered) rug plot, and the three extremes on each end are identified by name. The red vertical lines show the quartiles of the distribution.

Several features of this plot deserve comment, and also invite further inquiry: Most notable is that, by and large, milestones authors generally lived to a ripe old age— the median lifespan is 73.0, but the density plot peaks at around 79. This contrasts with a detailed study on famous people between 2400 BC to 1880 AD by David de la Croix and Omar Licandro (`http://www.fcs.edu.uy/archivos/BCU_clebrities.pdf`). In their research, it was found that the typical lifespan fluctuated around a mean of 61 years for four millennia, and only gradually reached 69 toward the end of their sample. Such a discrepancy between the two studies might warrant further investigation; for instance, by classifying the individuals into occupational, locational, or otherwise more delineated groups, and looking for trends.

Another interesting feature that becomes apparent in this graphic is the noticeable bump in the distribution around 45 years. This occurrence calls for some attempt at further explanation. We don't pursue this here, but again note that such graphs often suggest further analyses (breakdowns by region or time period), or cry out for the collection of more data.

Finally, although Figure 10 is just a summary graph, we have labeled a few extreme observations on each end, which may relate to telling parts of the story of the history of data visualization. Among these, Henry Moselely, who is known for the discovery of atomic number from a graphical display, died the youngest, as a consequence of serving in the British Army during World War I. But, we were surprised to see the noted and prolific French cartographer Phillippe Buache, and the German physicist and astronomer Johann Tobias Mayer, show up in positions two and three. On the other end, we were delighted to see that Adriano Balbi, a Venetian geographer and early collaborator of André-Michel Guerry (Balbi and Guerry, 1829) had the longest lifespan, just exceeding the population geneticist,

Sewall Wright, who invented path analysis and the path diagram around 1920. By incorporating these details, the visualization is able to reveal narratives that otherwise would have been concealed.

4.3 Milestone authors: geography

The Milestones Project web site provides an initial page showing an interactive timeline of the events in this history as a visual overview (Figure 2). A long-term goal has been to provide other views of this history and other tools for searching and exploring the database. With recent technological developments, it became evident to us that one such method would be through the use of geographical data.

So far, the primary geographic information we have encoded in the database refers to the birth and death place of the milestone authors. This is an imperfect representation, as these locations may not accurately represent the author's primary residence. For instance, Charles Joseph Minard was born in Dijon, and died in Bordeaux, but all of his work was done in Paris while he worked at the École Nationale des Ponts et Chaussées. Nevertheless, a geographic view of the available information is potentially useful. In this regard, we used the Google geocoding tools to provide latitude and longitude for the locations listed in the author table. Using this and the R package googleVis (Gesmann and de Castillo, 2011), we created the interactive map shown in Figure 11.

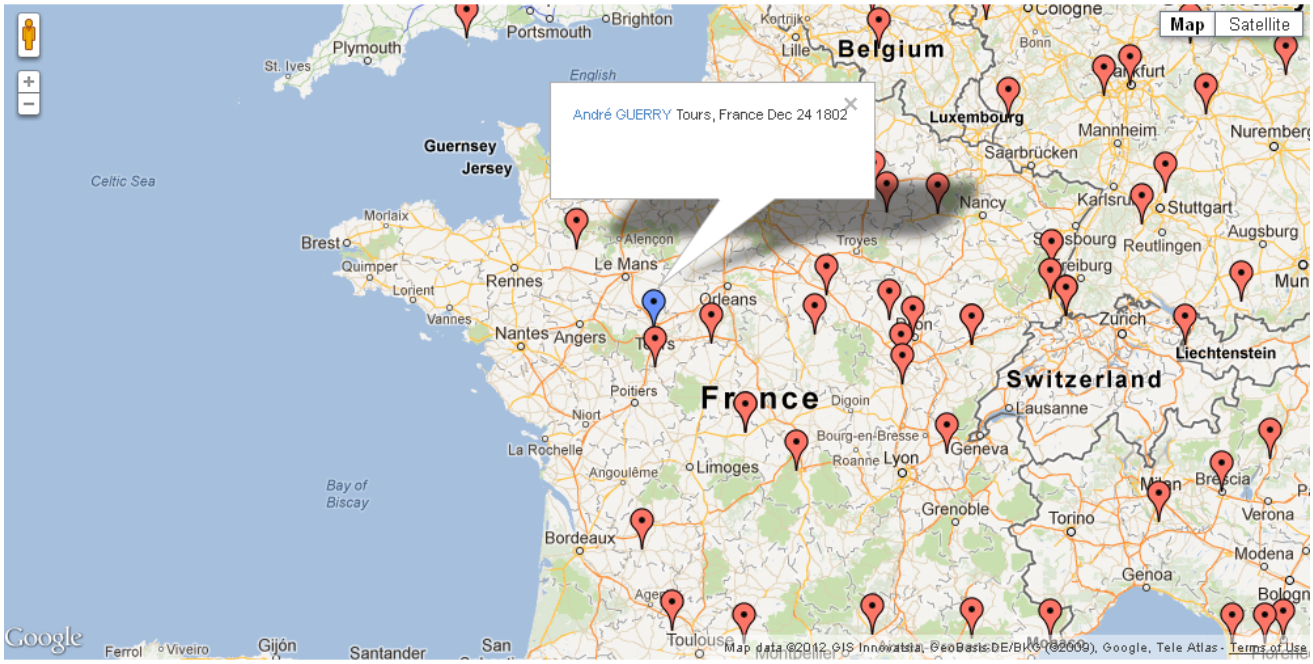

Figure 11: Birth places of 188 milestone authors, shown on an interactive Google map, centred on France. Each geographic marker is linked to an author query on the `datavis.ca` web site that lists the contributions by that individual.

Like other instances of Google Maps, this graphic can be panned and zoomed using mouse controls. The place markers display tool tips when hovered over and, when clicked, link to a search page that details all of the Milestone items that are related to that author. This interesting visualization will soon be revealed on the Milestones Project website, with future work planned to incorporate other types of data in addition to the birth and death locations.

4.4 Milestones: themes and trends

The records in the Milestones Project database also feature various text fields for each logged event. These include a brief item tag, a full description of the event, and relevant keywords, as well as categorical codes for the content (Subject), and form (Aspect) of the item. Treating this information as "data" allows us and others to study themes and trends in these developments. Modern methods of text mining and data visualization can provide insights into this history not available through other means.

As one simple illustration of this approach, Figure 12 shows two mosaic displays[9] that explore the relationships among Epoch, Subject, and Aspect. The left panel shows changes in the distributions of milestone events by Subject over time. It can readily be seen that while most of the milestone innovations up to the end of the 18^{th} century were about the physical world (astronomy, geodetic measurement, weather, etc.), this trend changed in the 19^{th} century, where there was a large shift toward problems that related to human populations (e.g., pertaining to mortality, births, disease, crime). Beginning in the early 1900s, the pattern changes again, with advances in mathematics and statistics becoming the dominating force.

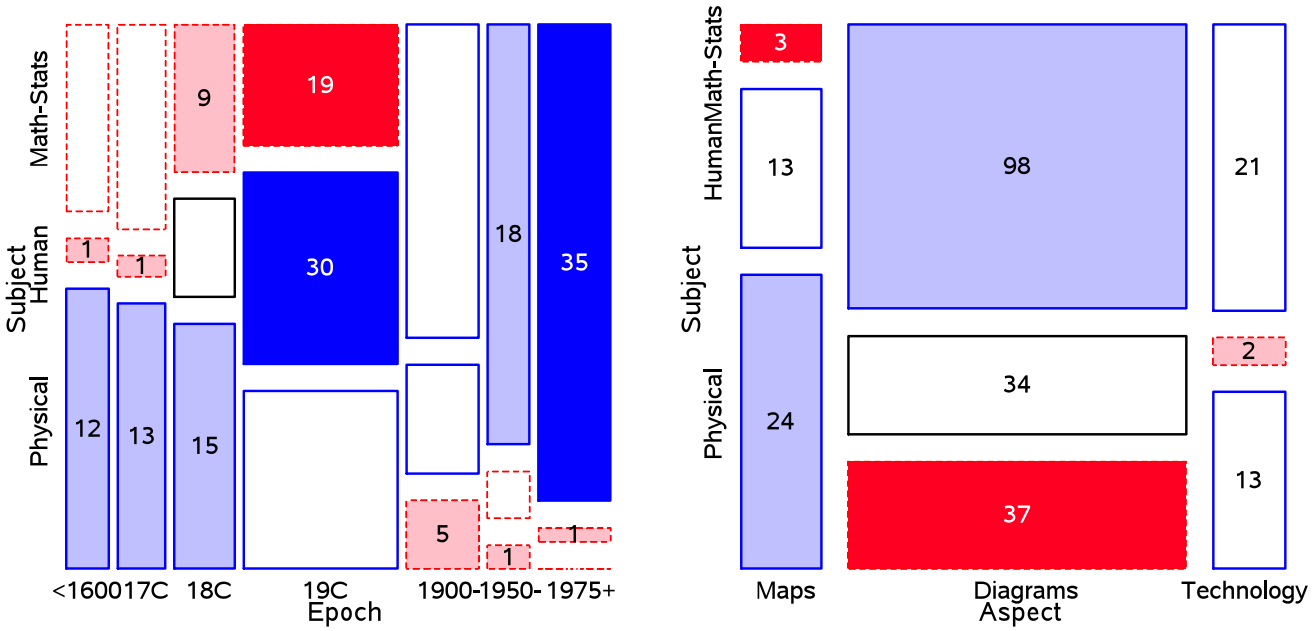

Figure 12: Mosaic displays for milestone items, classified by Epoch, Subject and Aspect. Left: mosaic for the marginal table showing differences in Subject across Epochs; Right: mosaic for the marginal table showing differences in Subject across Aspect. Numbers in the tiles give the number of milestone items.

The right panel shows the association between Subject and Aspect, pooled over Epoch. As is not surprising, maps and other cartographical representations were most often used to show data of the physical world, while graphs and diagrams were most often associated with mathematical and statistical subjects.

Other statistical graphs and analyses could be used to explore these and other relationships in more detail. The key to this is of course the existence and availability of data— in this case reflected by the coding of graphical milestones in our database.

[9]Mosaic displays show the frequencies in cells of a cross-classified table by the area of each tile. The tiles are shaded according to departure from a null model of no-association, using blue for cells with frequencies substantially greater than chance, and red for cells with frequencies that are lower than expected.

5 Conclusion and Future Directions

The Milestones Project began as a simple attempt to collect a comprehensive history of innovations and developments in data visualization in a single, "one-stop shopping" location. Like Topsy, it "just grew" over time, with images, historical papers and references, suggestions, and other contributions graciously provided by friends and collaborators, most notably from the members of *Les Chevaliers des Albums de Statistique Graphique*.

In this chapter, our primary goal was to introduce the second and latest iteration of this project. The redesign was undertaken to make this history more accessible for browsing and searching, and to attempt to make the database more amenable to additions, edits, and extensions among collaborators. However, we find that the most exciting aspect of the new structure is its flexibility in terms of data retrieval, and our newfound ability to use and manipulate the data for graphic-based statistical historiography.

One goal for the future, as we suggested earlier (Section 4.3), is to extend the user interface to provide multiple views and advanced text search and filtering capabilities. One convenient path for this development is provided by the SIMILE Exhibit framework (`www.simile-widgets.org/exhibit3/`). This provides web software libraries (Ajax, javascript, css) for timelines, interactive maps, tabular displays, image "tiles" and other visualizations. Various views can be composed for browsing as tabbed, alternatives or as faceted displays, showing, for example an interactive timeline and a map.

Equally important, the Exhibit framework allows us to present some of the milestones tables to be used as filters for the items displayed in these views. Tables for subject, aspect, keywords, location, epoch, etc. would allow the user to select select milestone events based on some or all of these criteria, providing a way to ask such questions as "what milestones events between 1700-1900 involving social science occurred in Europe?"

Finally, we would like to make the milestones database more publicly accessible for use by others on the history of data visualization. For the examples we have shown here, we connect to the milestones database directly via MySQL or ODBC interfaces to SAS and R, but this presents security risks. Happily, the Exhibit framework also provides methods for data export from various views, using JSON or CSV formats. In addition, we contemplate adding facilities for users in the data visualization community to add comments, notes, references and links to milestones items. These extensions will comprise the Milestones Project 3.0.

References

Balbi, A. and Guerry, A.-M. (1829). Statistique comparée de l'état de l'instruction et du nombre des crimes dans les divers arrondissements des académies et des cours royales de France. Jules Renouard, Paris. BL:Tab.597.b.(38); BNF: Ge C 9014 .

Beniger, J. R. and Robyn, D. L. (1978). Quantitative graphics in statistics: A brief history. *The American Statistician*, 32, 1–11.

Codd, E. F. (1971). Further normalization of the data base relational model. *IBM Research Report, San Jose, California*, RJ909.

Farebrother, R. W. (1999). *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. New York: Springer.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89, 190–200.

Friendly, M. (2002a). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 11(1), 89–107.

Friendly, M. (2002b). Visions and Re-Visions of Charles Joseph Minard. *Journal of Educational and Behavioral Statistics*, 27(1), 31–51.

Friendly, M. (2005). Milestones in the history of data visualization: A case study in statistical historiography. In C. Weihs and W. Gaul, eds., *Classification: The Ubiquitous Challenge*, (pp. 34–52). New York: Springer.

Friendly, M. (2007). A.-M. Guerry's Moral Statistics of France: Challenges for multivariable spatial analysis. *Statistical Science*, 22(3), 368–399.

Friendly, M. (2008a). A brief history of data visualization. In C. Chen, W. Härdle, and A. Unwin, eds., *Handbook of Computational Statistics: Data Visualization*, vol. III, chap. 1, (pp. 1–34). Heidelberg: Springer-Verlag.

Friendly, M. (2008b). The Golden Age of statistical graphics. *Statistical Science*, 23(4), 502–535.

Friendly, M. (2011). *HistData: Data sets from the history of statistics and data visualization.* R package version 0.6-12.

Friendly, M. and Denis, D. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. Web document. `http://www.math.yorku.ca/SCS/Gallery/milestone/`.

Friendly, M. and Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2), 103–130.

Friendly, M. and Dray, S. (2010). *Guerry: Guerry: maps, data and methods related to Guerry (1833) "Moral Statistics of France".* R package version 1.4.

Friendly, M. and Palsky, G. (2007). Visualizing nature and society. In J. R. Ackerman and R. W. Karrow, eds., *Maps: Finding Our Place in the World*, (pp. 205–251). Chicago, IL: University of Chicago Press.

Friendly, M., Valero-Mora, P., and Ulargui, J. I. (2010). The first (known) statistical graph: Michael Florent van Langren and the "Secret" of Longitude. *The American Statistician*, 64(2), 185–191.

Friis, H. R. (1974). Statistical cartography in the United States prior to 1870 and the role of Joseph C. G. Kennedy and the U.S. Census Office. *American Cartographer*, 1, 131–157.

Funkhouser, H. G. (1936). A note on a tenth century graph. *Osiris*, 1, 260–262.

Funkhouser, H. G. (1937). Historical development of the graphical representation of statistical data. *Osiris*, 3(1), 269–405. Reprinted Brugge, Belgium: St. Catherine Press, 1937.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.

Gesmann, M. and de Castillo, D. (2011). *googleVis: Interface between R and the Google Visualisation API.* R package version 0.2.12.

Guerry, A.-M. (1864). *Statistique morale de l'Angleterre comparée avec la statistique morale de la France, d'après les comptes de l'administration de la justice criminelle en Angleterre et en France, etc.* Paris: J.-B. Baillière et fils. BNF: GR FOL-N-319; SG D/4330; BL: Maps 32.e.34; SBB: Fe 8586; LC: 11005911.

Hald, A. (1990). *A History of Probability and Statistics and their Application before 1750*. New York: John Wiley and Sons.

Hankins, T. L. (1999). Blood, dirt, and nomograms: A particular history of graphs. *Isis*, 90, 50–80.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy, ed., *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, (pp. 268–273). New York, NY: Springer-Verlag.

Heiser, W. J. (2000). Early roots of statistical modelling. In J. Blasius, J. Hox, E. de Leeuw, and P. Schmidt, eds., *Social Science Methodology in the New Millenium: Proceedings of the Fifth International Conference on Logic and Methodology*. Amsterdam: TT-Publikaties.

Hoff, H. E. and Geddes, L. A. (1959). Graphic recording before Carl Ludwig: An historical summary. *Archives Internationales d'Histoire des Sciences*, 12, 3–25.

Hoff, H. E. and Geddes, L. A. (1962). The beginnings of graphic recording. *Isis*, 53, 287–324. Pt. 3.

Kruskal, W. (1977). Visions of maps and graphs. In *Proceedings of the International Symposium on Computer- Assisted Cartography, Auto-Carto II*, (pp. 27–36). 1975.

Palsky, G. (1996). *Des Chiffres et des Cartes: Naissance et développement de la cartographie quantitative française au XIX^{e} siècle*. Paris: Comité des Travaux Historiques et Scientifiques (CTHS).

Pearson, E. S., ed. (1978). *The History of Statistics in the 17th and 18th Centuries Against the Changing Background of Intellectual, Scientific and Religeous Thought*. London: Griffin & Co. Ltd. Lectures by Karl Pearson given at University College London during the academic sessions 1921–1933.

Playfair, W. (1786). *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century*. London: Debrett; Robinson; and Sewell. Re-published in Wainer, H. and Spence, I. (eds.), *The Commercial and Political Atlas and Statistical Breviary*, 2005, Cambridge University Press, ISBN 0-521-85554-3.

Playfair, W. (1801). *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. London: Wallis. Re-published in Wainer, H. and Spence, I. (eds.), *The Commercial and Political Atlas and Statistical Breviary*, 2005, CAmbridge, UK: Cambridge University Press, ISBN 0-521-85554-3.

Playfair, W. (1821). Letter on our agricultural distresses, their causes and remedies; accompanied with tables and copperplate charts shewing and comparing the prices of wheat, bread and labour, from 1565 to 1821. BL: 8275.c.64.

Porter, T. M. (1986). *The Rise of Statistical Thinking 1820–1900*. Princeton, NJ: Princeton University Press.

Priestley, J. (1765). *A Chart of Biography*. London: (n.p.). BL: 611.l.19.

Riddell, R. C. (1980). Parameter disposition in pre-Newtonain planetary theories. *Archives Hist. Exact Sci.*, 23, 87–157.

Robinson, A. H. (1982). *Early Thematic Mapping in the History of Cartography*. Chicago: University of Chicago Press.

Rosenberg, D. and Grafton, A. (2010). *Cartographies of Time: A History of the Timeline*. New York: Princeton Architectural Press.

Royston, E. (1970). Studies in the history of probability and statistics, III. a note on the history of the graphical presentation of data. *Biometrika*, 43, 241–247. Pts. 3 and 4 (December 1956); reprinted In *Studies in the History Of Statistics and Probability Theory*, eds. E. S. Pearson and M. G. Kendall, London: Griffin.

Rubin, E. (1943). The place of statistical methods in modern historiography. *American Journal of Economics and Sociology*, 2(2), 193–210.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, (pp. 336–343). Washington, DC, USA: IEEE Computer Society.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard University Press.

Tilling, L. (1975). Early experimental graphs. *British Journal for the History of Science*, 8, 193–213.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual Explanations*. Cheshire, CT: Graphics Press.

van Langren, M. F. (1644). *La Verdadera Longitud por Mar y Tierra*. Antwerp: (n.p.). Ii + 14 pp., folio; BL: 716.i.6.(2.); BeNL: VB 5.275 C LP.

Wallis, H. M. and Robinson, A. H. (1987). *Cartographical Innovations: An International Handbook of Mapping Terms to 1900*. Tring, Herts: Map Collector Publications.

Wills, G. (2012). *Visualizing Time: Designing Graphical Representations for Statistical Data*. Statistics and computing. New York: Springer.