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

Michael Friendly Matthew Sigal Derek Harnanansingh

August 22, 2012

Abstract

Approaches to modern data visualization have evolved substantially from the first thematic maps of the 1600s and the first bar charts and line graphs in the early 1800s to the dynamic and interactive graphics of today. Over the course of this history, scholars have taken a variety of approaches to show how particular factors of interest have changed over time.

The purpose of this chapter is threefold: first, to introduce the reader to an online resource called the Milestones Project. This website highlights 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 visual examples that deal with conveying aspects of history over time, drawn from this resource and discussed in terms of their goals, similarities, and differences. Finally, the Milestones database itself will be used to showcase how such a resource can serve as an interesting source of information 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.

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.[2] When this work

[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] 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) and thematic mapping (Friendly and Palsky, 2007, Palsky, 1996, Robinson, 1982), and so forth.

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, and containing 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 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.

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 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.

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 honor. 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 turned out to involve astronomy, archival research, patronage in the 17^{th} century and even an unsolved problem of cryptography, but also serves as one example of statistical historiography. For the present purposes we note simply that the Milestones Project provided the infrastructure for this research— a time-based, cross-referenced catalog of images, references and links to related work.

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. The next major invention, and the first true bivariate display is scatterplot whose use by Galton (1886) led to the discovery of correlation and regression, and ultimately to much of present multivariate statistics. So, it is perhaps surprising that there is no one widely credited with the invention of this idea.

In Friendly and Denis (2005) we trace the early origins of ideas related to the scatterplot, why, in Playfair's time, it was nearly impossible to think about and visualize bivariate relations, and how Galton's visual insight from a scatterplot contributed to the rise of modern statistics and graphics. But, the resources available in the Milestones Project allowed us to 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 this history of data visualization into epochs, each of which turned out to be describable by coherent themes. For reasons we describe 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 ($\pm$10) the Golden Age of Statistical graphics (Friendly, 2008b).

Figure 2 shows the time distribution of 260 miletone events listed in the Milestones Project in 2007 together with the labels we used for epochs. In Friendly (2008b) we trace the origin of this period in terms of the infrastructure required to produce this explosive growth of contributions to data visualization: systematic data collection by state agencies, the rise of statistical and visual thinking, and enabling developments of technology.

2 The Milestones Project

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 in documenting the history of data visualization in a convenient form for browsing, searching and analysis.

2.1 Origin, structure and evolution

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

This began as 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 using a unix command pipe like

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

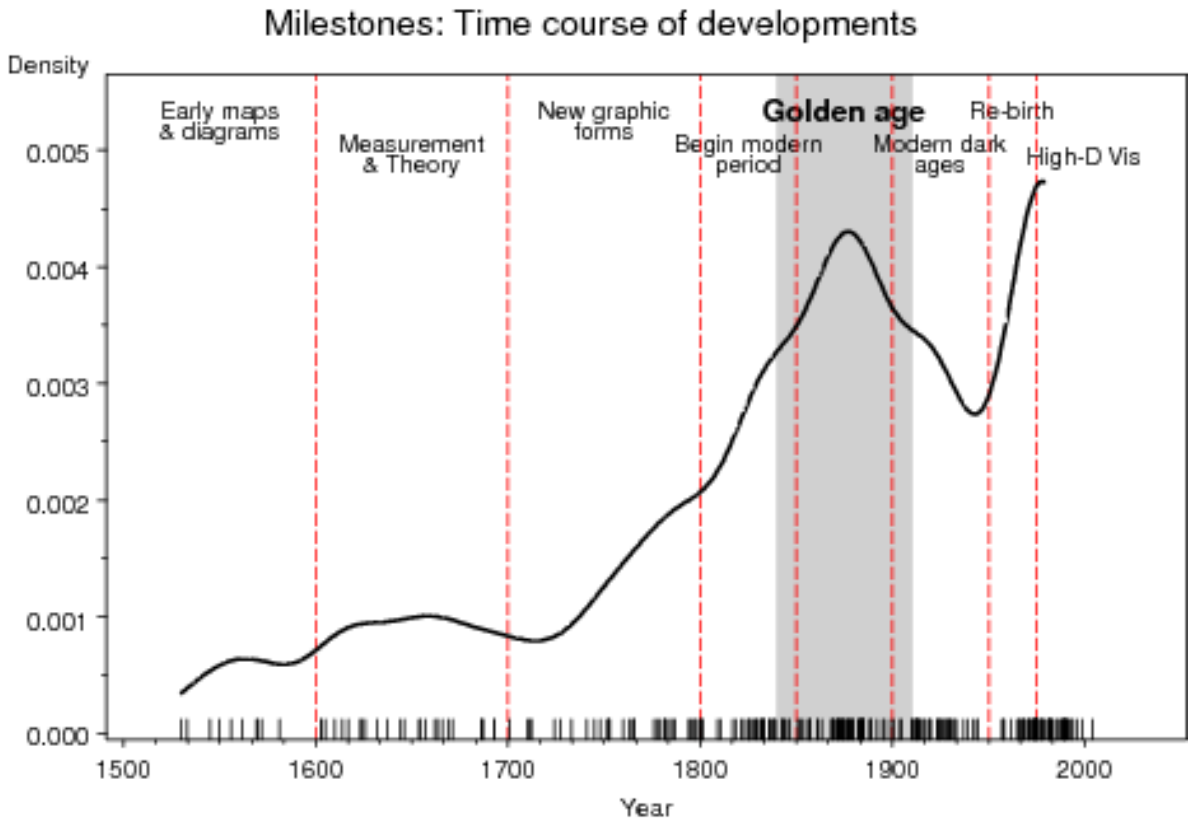

Figure 2: The time distribution of events considered milestones in the history of data visualization, shown by a rug plot and density estimate. The density estimate is based on $n = 260$ significant events in the history of data of data visualization from 1500–present. The developments in the highlighted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

It soon became apparent that such a text-based representation was inadequate. Around 2005, we began to convert this to a true database and completely redesigned the Milestones web site. ...

At present, the Milestones Project lists 288 contributions to this history, with nearly 350 references, information on 336 authors and 774 "media items", comprising 371 images appearing online on the `http://datavis.ca` site and 403 links to images and documents at other sites. In addition, we maintain an offline image database comprising over 1100 images collected from various sources, ...

Figure 3 shows the timeline view of the miletsones items displayed on the landing page. ...

3 Visualizing Time: Historical Precedents

- Visions of history from the past
- Correlated pasts
- Non-linear scales
- Dynamic/interactive timelines

4 Using the Milestones Project for Statistical Historiography

- comparisons of graphical innovations by field (e.g. physical sciences vs. social sciences), in terms of content and form

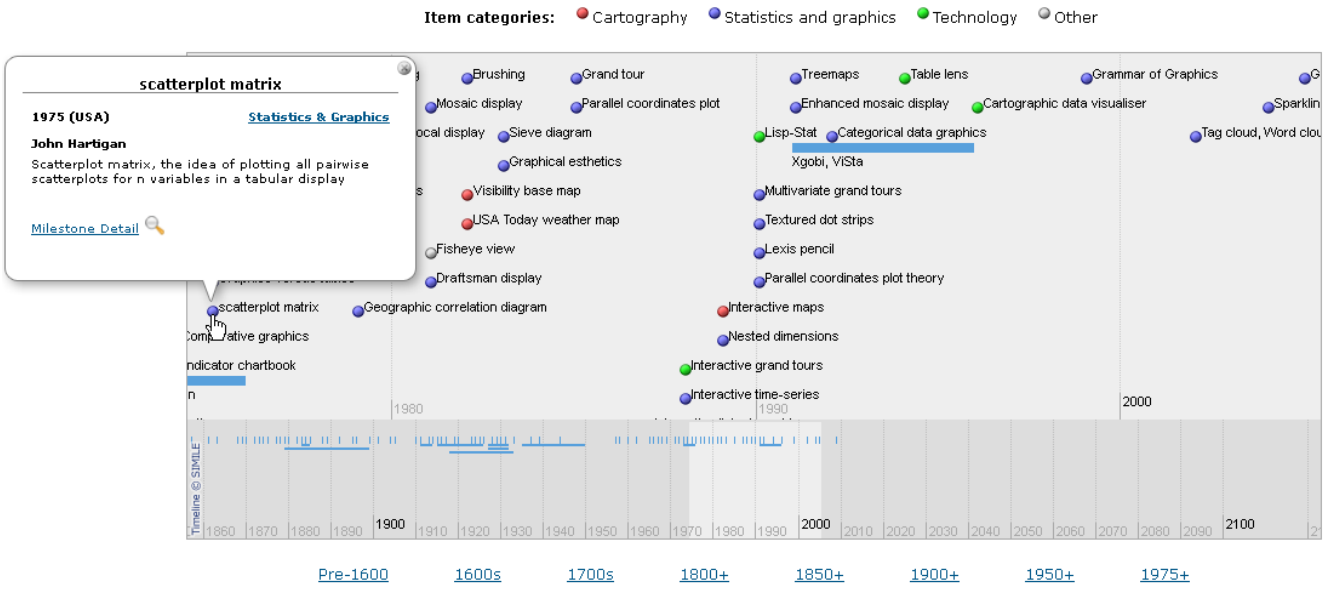

Figure 3: Timeline view of the Milestones Project on the site `http://datavis.ca`. In this view, the top panel shows a detailed view of the segment of history highlighted in the bottom panel, both of which can be separately scrolled. Items in the top panel show a brief tag, color-coded in coarse categories. Clicking on an item in this panel brings up small description, linked to the details of the milestone item.

- extract themes and relationships between content areas

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.[3] 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.

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.[4] On more than one occasion, we wondered whether there wasn't something lost with the advent of modern software: We can now analyze massive data sets and generate many graphs with simple mouse clicks or software commands, but designing a truly effective graphic display requires much thought and a lot of manual intervention.

There is, of course, one principal requirement for statistical historiography: **data**. The milestones database is the repository of all the information we have so far recorded, and modern database tools allow the possibility of simple or complex queries, limited only by the available information.

In related work, we have also collected and disseminated data sets of historical interest on a variety of topics in statistics and data visualization, for example in the R packages HistData (Friendly, 2011) and Guerry (Friendly and Dray, 2010).

[3] 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*.

[4] Some examples are: Charles Joseph Minard's famous depiction of Napoleon's March on Moscow (Friendly, 2002), 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) ...

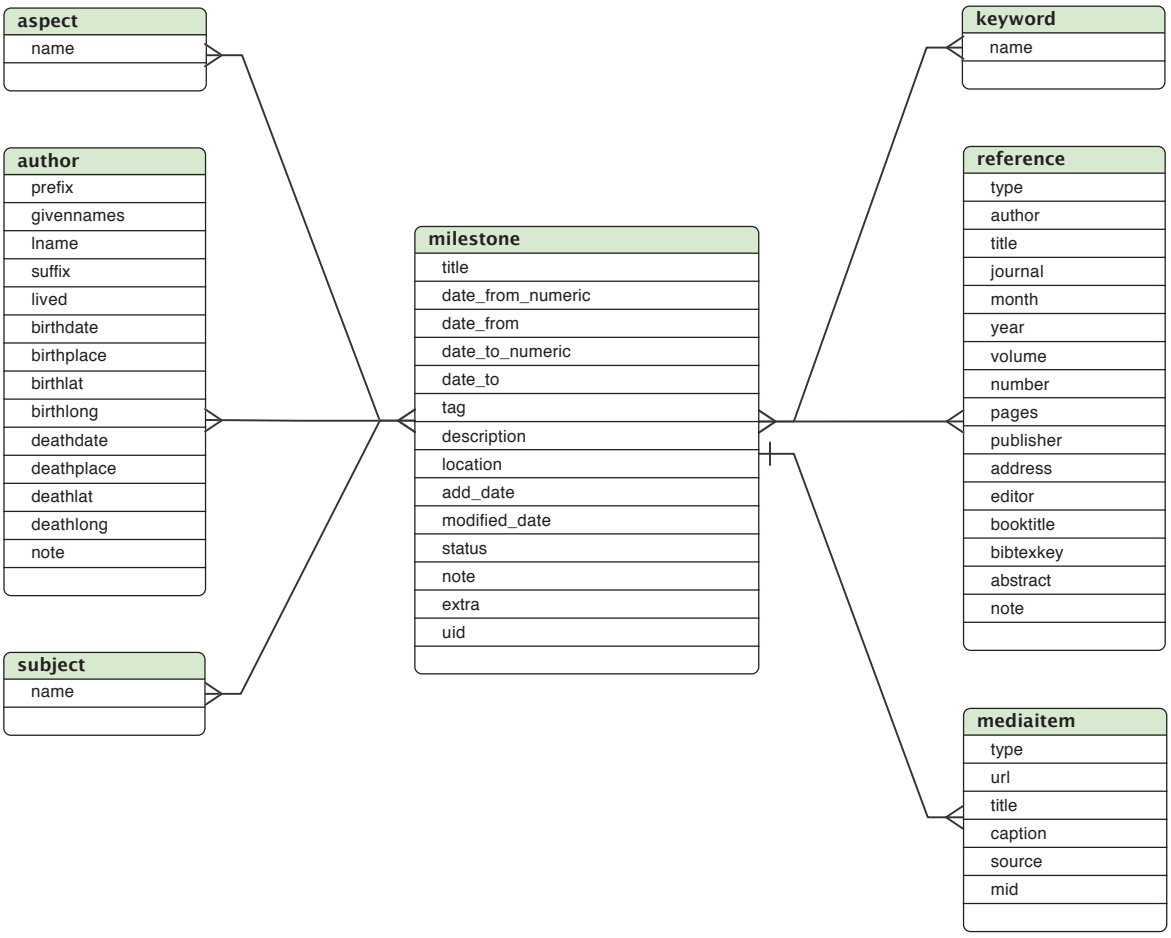

Figure 4: Simplified schema for the MySQL database for the Milestones Project. The main table (`milestone`) contains information regarding each of the items considered a milestone in the history of data visualization, linked to other tables (e.g., `reference`, `mediaitem`) by unique keys. Other supporting tables, not shown here, provide for convenient lookup of descriptors of these milestones items (`subject`, `aspect`, `keyword`).

In the subsections below, we describe a few applications of these ideas using the milestones 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 1836 extensive 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. Internet and biographical searches for these persons allowed us to determine the dates and places of their birth and death in a large number of cases.

One simple question is how long did these contributors live? As illustrated earlier (figref:) Joseph Priestley was the first to develop the idea of using a graphic representation to show the lifespan of famous

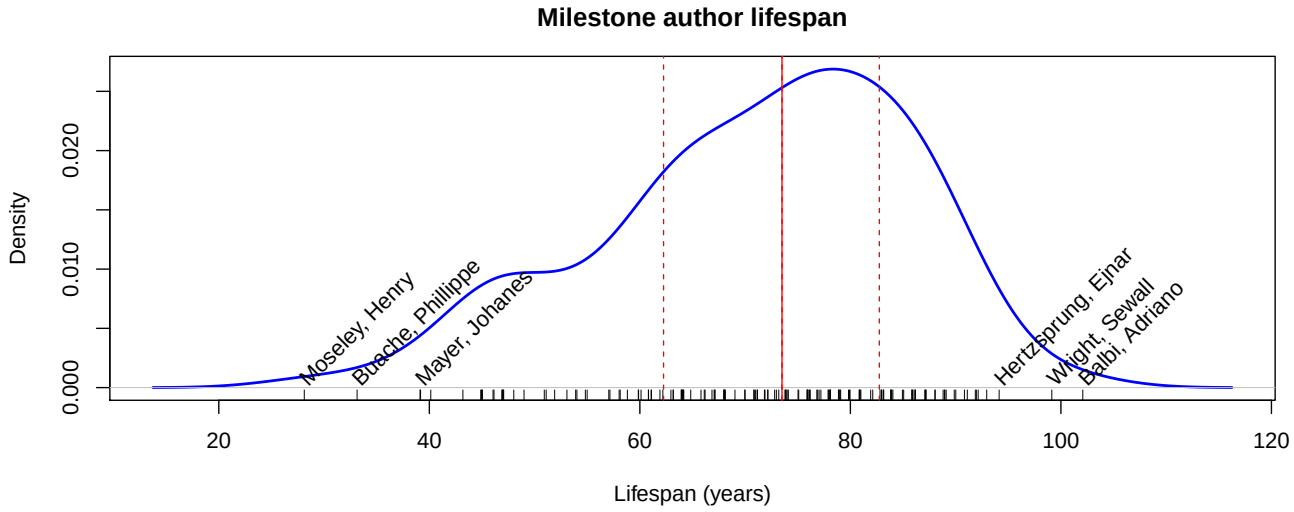

Figure 5: Density plot of the lifespan of 172 authors in the milestones database born after 1500 and for whom lifespan can be determined. Individual observations are shown by a (jittered) rug plot, and the three extremes on each end are identified by name. The dashed lines show the quartiles of the distribution.

men. His "charts of biography" did this in a particularly evocative form, showing each person by a line segment identified by name, and grouped into occupational categories.

However, such "lifeline" charts don't provide any answers to this question. With the author table, it is a simple matter to calculate lifespan, and give a direct answer with the help of software. Figure 5 shows one display of this information, using a combined density plot and rug plot, as we used in Figure 2.

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 by David de la Croix and Omar Licandro (`http://www.fcs.edu.uy/archivos/BCU_clebrities.pdf`) of famous people from 2400 BC to 1880 AD, fluctuating around a mean of 61 years for 4 millennia, and only reaching a mean of 69 years by the end of their sample. To take this analysis further would require more data, for example a classification of authors by occupational groups.

Second, there is a noticeable bump in the distribution around 45 years. This occurrence also 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 5 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 in 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 2-3. 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.

4.3 Milestone authors: geography

The Milestones Project web site provides an initial page showing an interactive timeline of the events 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.

The geography of these developments so far is only represented in the birth place and death place information we recorded in the author table. For example, we know that 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 place names listed in the author table. Using this and the R package googleVis (Gesmann and de Castillo, 2011) we easily created the interactive map shown in Figure 6

Like other Google maps, this can be panned and zoomed using mouse controls. The place markers display tool tips when hovered and, when clicked, link to a search page showing milestone items related to that author. This tool seems sufficiently useful that we are adding it as another visual overview page to the milestones site.

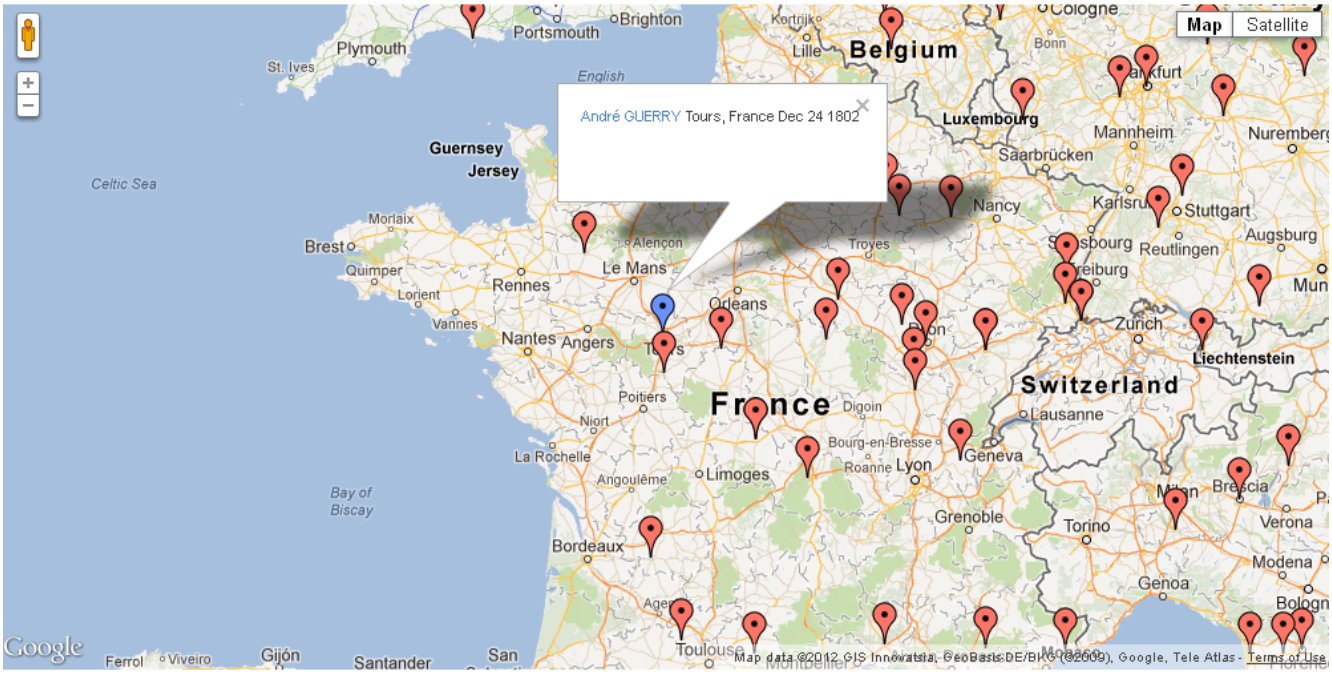

Figure 6: Birth places of 188 milestones authors, shown on an interactive Google map, zoomed to show locations in part of Europe centered on France. Each geographic marker is linked to a query on the `datavis.ca` web site listing the contributions by that author.

5 Conclusion

-Summarize and conclude

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.

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

Friendly, M. (2002). 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.

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.

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

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.

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.

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.