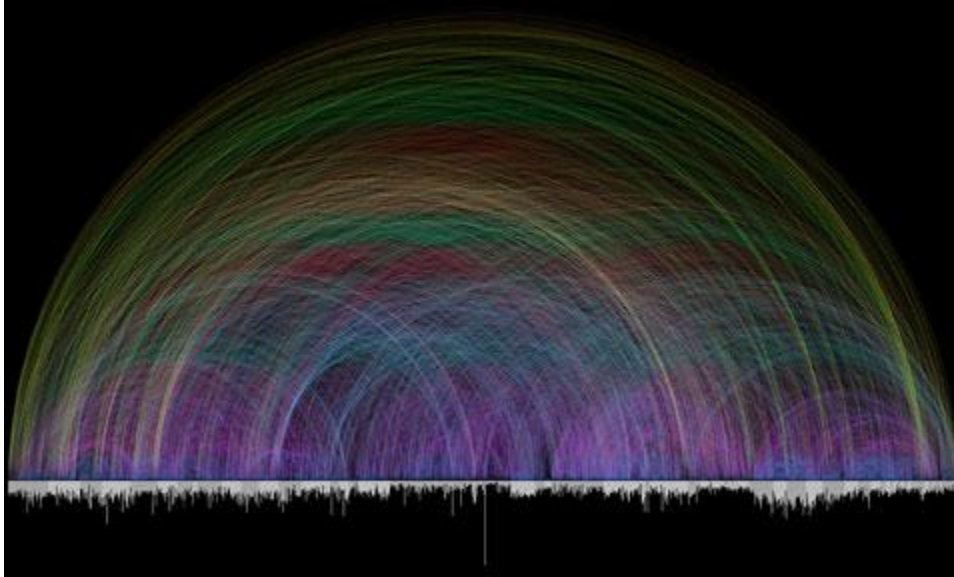


# Tooling Up for the Digital Humanities, from Stanford

## Data Visualization



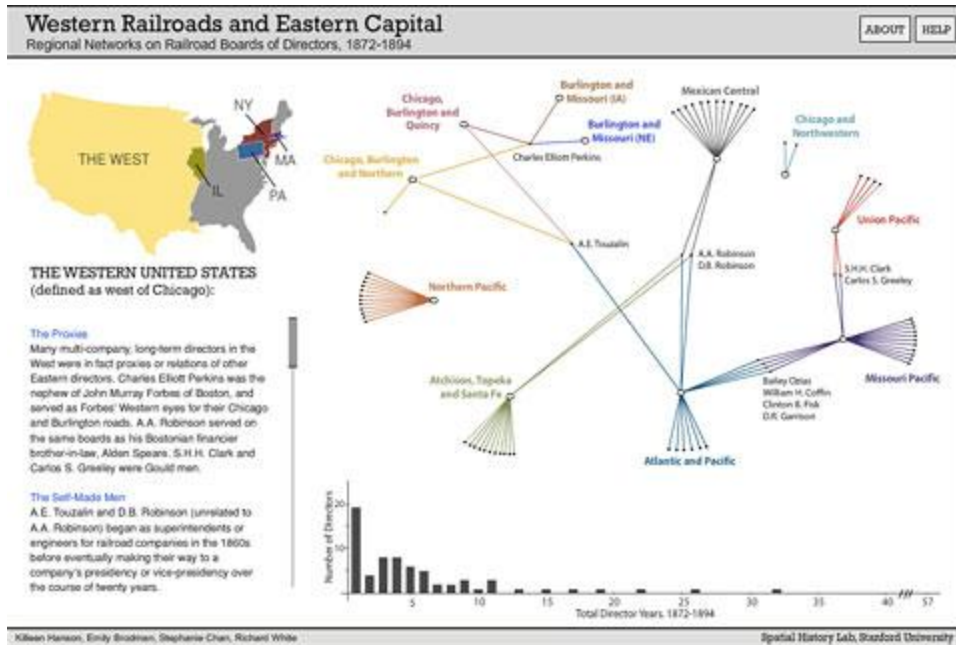
New computer programs offer innovative ways to present and convey information visually—a function particularly important in representing large quantities of data. While much of this information can, conceivably, be conveyed through text, many new visualization tools are appealing precisely because they can communicate certain ideas better than verbal language alone. Often they communicate something that escapes words. Why should scholars accustomed to written forms of communication limit themselves to text when a visualization might convey certain information more effectively?

## 1: Introduction

A visualization can take any number of forms. While some will be geographic and map-based, others will involve various ways of displaying quantitative data, such as a simple graph, a word cloud, or a visualized network.

The Spatial History Lab at Stanford has experimented with a number of visualization technologies. While most are geography-based representations, others, like this display of human social networks (rendered in Pajek and published with Adobe Flash) show the value of other visualization tools.

The text is copied here for printing ease; this is the dynamic version:  
[http://toolingup.stanford.edu/?page\\_id=1247](http://toolingup.stanford.edu/?page_id=1247)



Although many visualization technologies are often used merely as illustrations, the real power of visualization comes in its ability to make powerful arguments, and show data in a way that raises new questions. We may tend to think of visualization as a finished product, as part of presentation, but it may be more useful to think of visualization as part of the research process. Visualizing data, argues Richard White in “[What is Spatial History](#),” is “a means of doing research; it generates questions that might otherwise go unasked, it reveals historical relations that might otherwise go unnoticed, and it undermines, or substantiates, stories upon which we build our own versions of the past.”

But conveying essential data in an elegant and clear way is harder than one might expect. There are plenty of visualizations that are appealing in some aesthetic way, but which might contain distractions and fail to convey the intended message. At worst, a poorly executed visualization can distort data and leave a reader confused.

The pursuit of clarity in visual representations of data has been one of the pursuits of renowned visualization scholar Edward Tufte. Tufte famously railed against the widespread use of “[chartjunk](#),” which he argued was a distraction in conveying information as efficiently as possible. Others have [defended chart junk](#) as a way of engaging readers. An effective visualization requires a careful balance of design and function.



Visualizations offer new challenges in communication for another simple reason: it seems people are more willing to accept certain visual representations as objective or scientific. In a field in which scholars work with great amounts of uncertainty and ambiguity, what is the humanities scholar's obligation in visualizing a data set? How must we account for visualizing data sets that may be incomplete or skewed?

The Republic of Letters project at Stanford, which is visualizing epistolary connections in the eighteenth century, has had to grapple with this problems. See linked [video](#) and [interactive visualization](#).



In fact, visualizations are themselves fundamentally human creations, each one representing many small choices made by an author to show certain things and hide others; to emphasize certain features and not others. In that sense, visual arguments are like the forms of communication that

The text is copied here for printing ease; this is the dynamic version:  
[http://toolingup.stanford.edu/?page\\_id=1247](http://toolingup.stanford.edu/?page_id=1247)

humanities scholars are more accustomed to using, but it may be able to convey new information and raise new questions in a number of scholarly fields.

## 2: Getting Started

At its core, visualization starts with data. Scholarship in the digital age will increasingly call on researchers to reach a comfort level with databases and database management systems. Fortunately, new tools and techniques are emerging that make dealing with data easier and more interoperable than ever.

For one thing, if you rely on public data sources for information - say, historical economic statistics by country- you might find the tabular information in the pdf's you download, or the html you copy and paste to be a terrible mess. Columns may be merged together, entities like place names may be written in different formats (Nev., Nevada or NV, for example). Tools like Stanford's Wrangler and Google Refine let you interactively track down those glitches, fix them and output clean, usable data quite easily.

The screenshot shows the DataWrangler interface. On the left, there is a 'Transform Script' panel with the following options:

- Split data repeatedly on newline into rows
- Split data repeatedly on ;

Below the script panel, there are tabs for 'Text', 'Columns', 'Rows', 'Table', and 'Clear'. A 'Split' script is active, showing the following rules:

- Split split between 'Unempl' and 'oyment'
- Split split between 'Unempl' and ' any word '
- Split split between 'Unempl' and ' any lowercase word '

The main table has the following columns: 'split', 'split17', and 's'. The data is as follows:

	split	split17	s
1	Unemployment rates	Unempl	oyment rate
2	As a percentage of total labour force	As a percentage of total labour force	
3			
4			
5			
6	Australia	Australia	
7	Austria	Austria	
8	Belgium	Belgium	
9	Canada	Canada	
10	Czech Republic	Czech Republic	
11	Denmark	Denmark	
12	Finland	Finland	
13	France	France	
14	Germany	Germany	
15	Greece	Greece	
16	Hungary	Hungary	
17	Ireland	Ireland	
18	Italy	Italy	

For getting your information into a database, Google also offers not only online spreadsheets but the increasingly powerful Google Fusion Tables, which lets you upload, edit, visualize and share data files of up to 100 megabytes in size. Whereas a desktop application like Microsoft Excel might become unresponsive after opening a spreadsheet of ten thousand rows, Fusion Tables lets you manage enormous files relatively painlessly and, moreover, lets you join them with related data quite easily. Want to create a heatmap of occurrences of what you're studying on a world map? With a few clicks you're there. Plus you can output your data set in any of the predominant file formats that you might need to create visualizations with other tools.

The text is copied here for printing ease; this is the dynamic version:  
[http://toolingup.stanford.edu/?page\\_id=1247](http://toolingup.stanford.edu/?page_id=1247)

One of the keys to good visualization is understanding what your immediate goals are. Are you visualizing data to understand what's in it, or are you trying to communicate meaning to others?

### 3: For Analysis and Understanding

Let's look at the first category, or what some people call exploratory data analysis. In this case, you're simply throwing your data set at a wall to see what's in it. If it's numeric data, what are the highs and lows? The median, mean, and standard deviation (a measure of the statistical distribution of your values)? Are there spatial patterns in geographic data, or trends over time? A number in isolation doesn't tell you anything; visualization can help you understand the context for your data. For data analysis, free online tools can be the very useful; while the final interface of the visualization might not be the most beautiful, and the interactivity might be limited, free tools can provide great visual sandboxes for you to play around with your data. Among the most accessible tools for quantitative information are **Tableau Public**, a free PC application that lets you apply quite powerful graphical analysis to your data, and **ManyEyes**, a web application that IBM developed to help people share data sets and look for patterns.

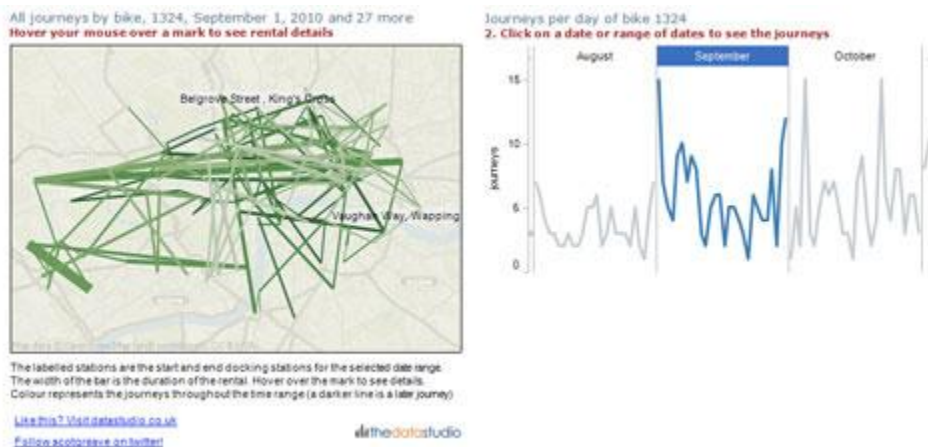
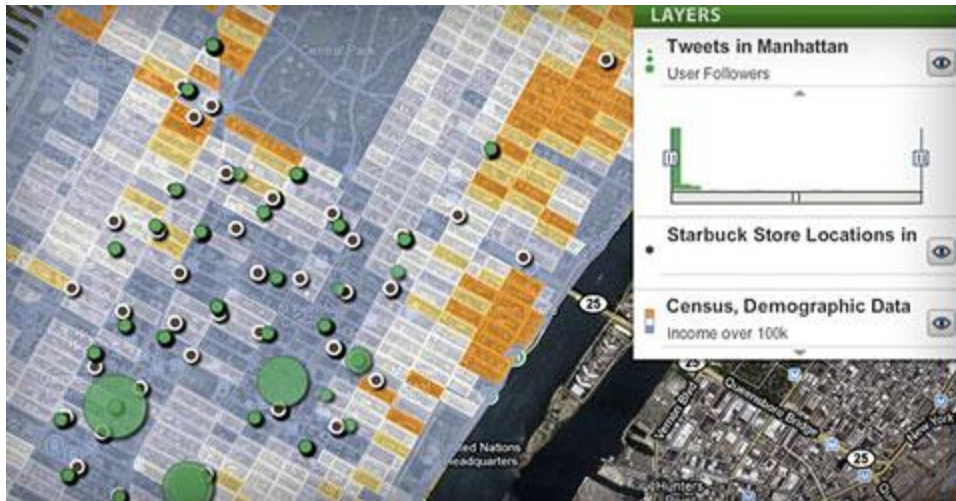


Tableau has quite good mapping abilities for data that contain spatial information, but there are other map-focused tools that can help you make more sophisticated data maps. One of these, called **Geocommons**, enables you to upload data sets that reference specific place names or geographic coordinates (in the form of latitude and longitude); it can recognize spatial references in your data set and plot it on a map or match it with administrative boundaries like counties, census tracts, states etc. You can then output a static map or an embeddable interactive map.

The text is copied here for printing ease; this is the dynamic version:  
[http://toolingup.stanford.edu/?page\\_id=1247](http://toolingup.stanford.edu/?page_id=1247)



There are also tools that support plotting your information over time, including the free timeline creation tools [Dipity](#) and MIT's [Simile Timeline](#), and the more sophisticated free desktop application called [TimeFlow](#), originally developed for investigative journalists.

For plotting the non-geographic relationships between people, organizations and ideas, there are also a number of network analysis tools available. (See "[Network Analysis Software and Services](#)" postscript for more information.)

Finally, for building something more interactive and application-like, [Adobe Flash](#) and [Flex](#) remain powerful and popular tools. For more than a decade, they have served as the standard for building online infographics and multimedia, particularly on news and information sites ([Geocommons](#), shown above, is one of several visualization tools built with Flash). While the lack of support for Flash on the iPhone and iPad have hurt its cross-platform compatibility, Flash is still a far more full featured tool for authoring interactive data visualizations than Javascript and HTML, although the latter have made great advances in recent years. Moreover, Adobe now offers a tool called "[Flash Packager for iPhone](#)" that can convert Flash documents into iOS applications, and they have a [similar tool in development](#) for translating Flash to HTML.

The key to using Flash lies in mastering Actionscript, its internal language for creating interactions and processing data for visualization. Syntactically, it is very similar to Javascript, but because it is "compiled" by the development tool into a self-contained application, it can run much faster.

In addition to the Spatial History Project and the Republic of Letters, probably the best showcase of the type of work possible with Flash is the online work done by The New York Times' [graphics desk](#).

## 4: For Communication and Storytelling

But what about visualizations that communicate to others? This is perhaps the most challenging area of visualization. Charts and other forms of information graphics have been around for centuries, but the computerized process of taking data and using it to generate graphics is relatively new and rapidly changing as we speak. While you can sometimes create a very clear and compelling chart at the touch of a button now, visualized data sets often demand that the viewer be given a fair amount of hand-holding. Many viewers are not necessarily used to reading visualizations critically.



For static charts published in print or posted as images, annotative overlays such as labels, arrows and descriptive text are highly important. For animated or interactive charts created with Flash, Java, Processing or other software, labels are also important. But more than that, they present the opportunity to construct a narrative that can lead the viewer through the main points of your visualization. Is it a time series showing how a particular issue took shape? Is it a general view that zooms into the specific, say birth rates across the U.S. before you zoom in on the state of Utah? There is a growing body of research and dialogue among journalists about the best practices of narrative storytelling with interactive graphics: A good survey of methods used in journalism can be found in research conducted at Stanford by Jeffrey Heer and Edward Segal: <http://vis.stanford.edu/papers/narrative>. There is also a documentary video about visualization in journalism at [datajournalism.stanford.edu](http://datajournalism.stanford.edu).

The text is copied here for printing ease; this is the dynamic version:  
[http://toolingup.stanford.edu/?page\\_id=1247](http://toolingup.stanford.edu/?page_id=1247)

**Journalism in the Age of Data** a data report on data visualization as a storytelling medium. Archived during a 2014-2015 copyright, creative commons license. Hosting time, as always, with digital information and tools.

**Ben Fry**  
Design and Software Consultant

**CHAPTERS**

- I. Introduction
- II. Data Viz in Journalism
- III. Telling "Data Stories"
- IV. A New Site for Infographics
- V. Life as a Data Stream
- VI. Exploring Data
- VII. Technologies and Tools
- VIII. Final Steps

**RELATED INFORMATION FOR THIS CHAPTER**

Ben Fry, Fathom Consulting

Fry's work spans the worlds of science, technology, art and communication. At the MIT Media Lab, he co-developed the open-source programming tool Processing with Casey Reas. Processing provides a simple, accessible interface for nonprogrammers to harness the graphics generation capabilities of the Java programming language. He also wrote the book "Visualizing Data" (2007), which offers detailed step-by-step instructions for creating some of his most well-known visualizations. Fry recently started a new consulting firm called *fathom* that is producing interactive visualizations for such clients as US Healthcare. His work has been shown in major museums around the world.

**RELATED LINKS**

- Fathom Consulting
- Personal Site
- Visualizing Data
- Interviews
- Twitter: Ben Fry

**ABOUT ME** **ACKNOWLEDGEMENTS**

LEOFF MCGREE is an online journalist specializing in multimedia and information graphics. Over the past decade he has worked at The New York Times and ABCNews.com, and in France at Le Monde Interactif. In 2009-2010, he spent a year as Knight Journalism Fellow at Stanford University studying data visualization. In July 2010, McGree began a new job developing visualizations and interactive content for the Bill Lane Center for the American West at Stanford.

Creative Commons Attribution-NonCommercial-ShareAlike 3.0 licensed license

Other interesting examples of narrative visualization include the Swedish public health researcher Hans Rosling's BBC program "[The Joy of Stats](#)," the Economist Magazine's online "[video graphics](#)," and Pedro Cruz's visual rendering of [the decline of empires](#).

## 5: Visualizations and Accountability

Because visualization is also about transparency, about letting other people explore your source data, there is another important attribute of interactive visualizations: helping people find the "state" of a visualization that highlights the insight you want to share. Where is the center point of the display? How far is it zoomed in? Is the data sorted, filtered or focused on a particular item? This ability to "track back" makes it possible not only for you to highlight the pertinent information - the smoking gun or the needle in the haystack - but it also allows other users to potentially do the same. Did somebody using your visualization spot something interesting? Then let them save that state of the visualization and [share it with others](#).





And beyond the visualization itself, there is a growing trend toward exposing the underlying dataset for your users to download. Within the bounds of protecting information that might be truly proprietary or contractually limited, letting other people download your data and find their own ways to visualize it goes hand-in-hand with the open spirit of data visualization. One need look no further than the Guardian newspaper in the UK and their “[data store](#),” where you can download any number of public and private data sets to visualize yourself, or the [contest sponsored by New York public radio](#), where users entered the history of their moves around the United States, and where those same users could download the collected data of other users and take a shot at visualizing it. Just recently Google and Eyebeam launched the “[Data Vis Challenge](#),” inviting users to create interactive visualizations showing how U.S. taxpayers’ money is spent by the government.

The text is copied here for printing ease; this is the dynamic version:  
[http://toolingup.stanford.edu/?page\\_id=1247](http://toolingup.stanford.edu/?page_id=1247)



Smashing Magazine has produced some good roundups of visualization genres and techniques:

[Modern approaches to data visualization](#)

[Inspiring data visualizations and infographics](#)

Web Designer Depot has [more examples here](#).

## Additional Info on Network Analysis Software and Services

International Network for Social Network Analysis: <http://www.insna.org/> ([list of software](#))

R: (<http://www.r-project.org/>) For advanced users, R is free, open source, and it is a programming language, which makes it more pliant but also increases the steepness of the learning curve.

[UCINET](#) is a low cost (~\$30) social network analysis program that works in tandem with freeware program called NETDRAW for visualizing networks. It is relatively easy to use and has good user manual data.

[NodeXL](#) is a useful tool for visualizing network graphs quickly and easily. It is an Excel plug-in developed by Microsoft engineers.

[Pajek](#) is another tool for visualizing network data. It offers more functionality than some programs, but support and documentation suffer from translation issues (the product was developed in Slovenia).

### 7: Discussion

You can download the slides from Geoff McGhee's presentation "[Getting Started with Data Visualization](#)" [here](#).

Discuss the material for the chapter on Data Visualization on the Stanford site:

[http://toolingup.stanford.edu/?page\\_id=1295](http://toolingup.stanford.edu/?page_id=1295)

The text is copied here for printing ease; this is the dynamic version:

[http://toolingup.stanford.edu/?page\\_id=1247](http://toolingup.stanford.edu/?page_id=1247)