Cool Statistics!
Econ 29000
Kevin R Foster, CCNY
Fall 2011

> *Learning Objective: Students will understand that statistics is important, exciting, and worth hard work.*

Stats pervade our everyday life, particularly online. Google is dominant in search because they've figured out how to give people what they want, using sophisticated statistics-based algorithms to do so. After you type even a single letter the site is making suggestions, based on your past browsing.

Smart uses of stats drive many top companies, from Google's advertisements to Facebook's friend suggestions and advertisements to Netflix with its movie suggestions to Amazon's "people who bought … also bought …" Even game designers use stats on attention spans – Angry Birds didn't just happen! Every company is always looking to improve, to get a competitive edge.

Hal Varian, chief economist at Google and previous Dean of the business school at Berkeley, notes "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." (*New York Times*, August 9, 2009, "For Today's Graduate, Just One Word: Statistics.")

Of course economics is built on statistics, from GDP to unemployment to interest rates or consumer confidence.
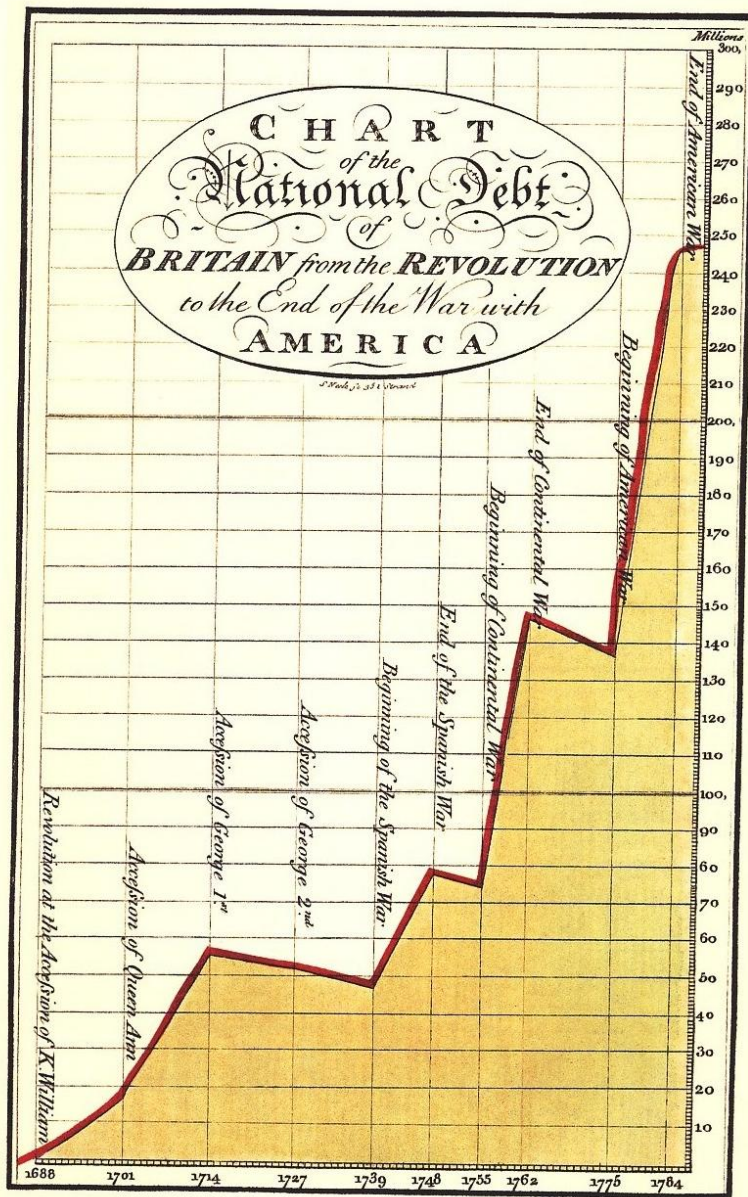
Companies battle to best use available data to forecast trends. Walmart is dominant because they built an inventory management system that used real-time data. In finance, quants rule: every trader is looking for new algorithms to figure out market moves, even (with high-frequency trading) to calculate the answer just a split second faster than their competitors. Risk managers then construct models to protect against those traders. In sports, data-driven managers are dominating the old guard who made decisions based on their gut instinct. Politicians depend on polls for nearly every utterance. Market research informs smart business decisions. Manufacturer's quality control is based on stats. Of course statistics is a bedrock of science and was developed in parallel to the scientific revolution.

We can use statistics to learn about the world and try to overcome our misconceptions, as in this presentation from the TED talks by Hans Rosling, 2006,
http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html
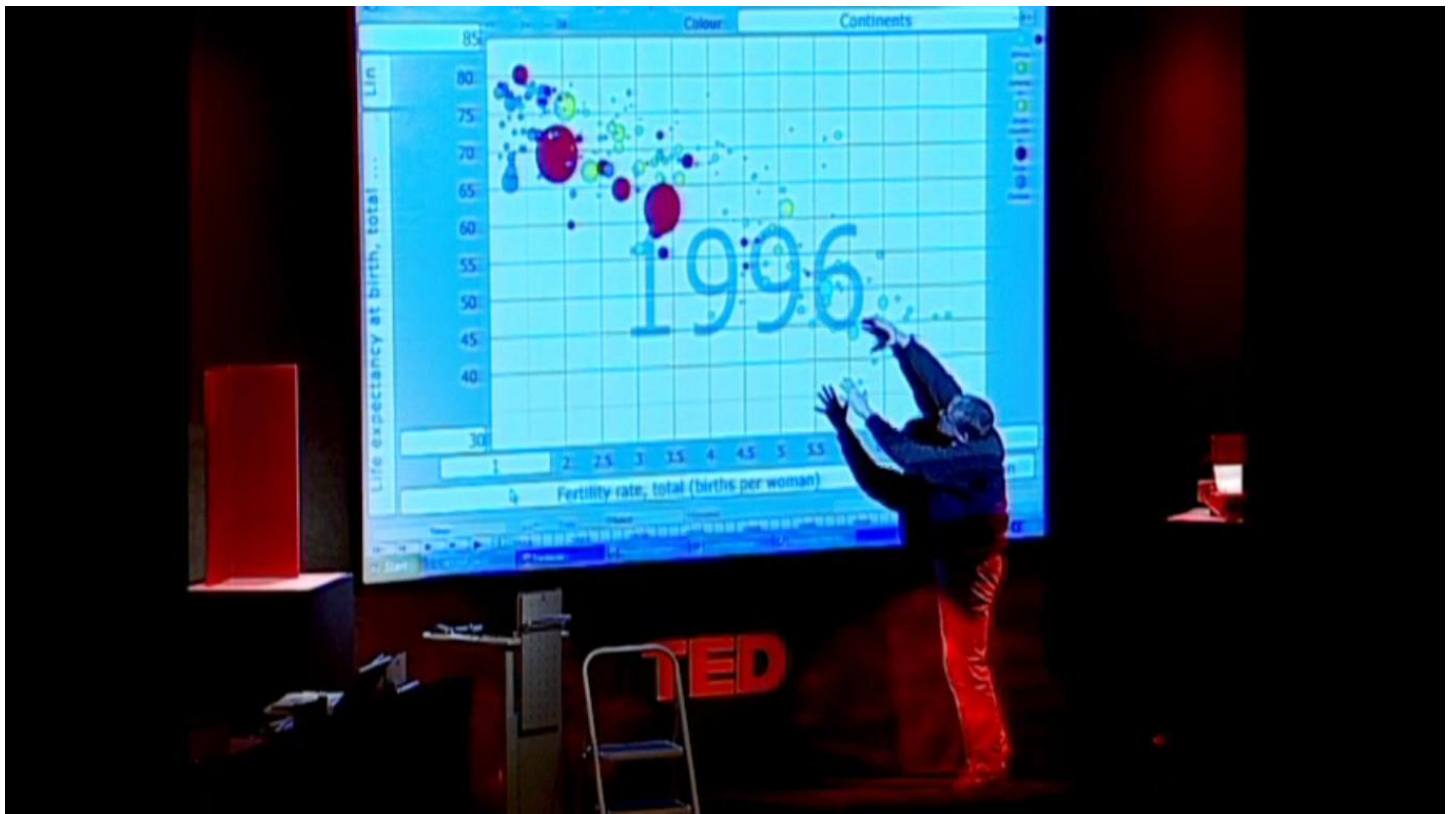
Where he explains why his students are dumber than chimps.

One of the earliest time-series charts, showing the effect of war on Great Britain's national debt, is due to William Playfair (credited as being the inventor of the pie chart, bar chart, and other displays still in use today):
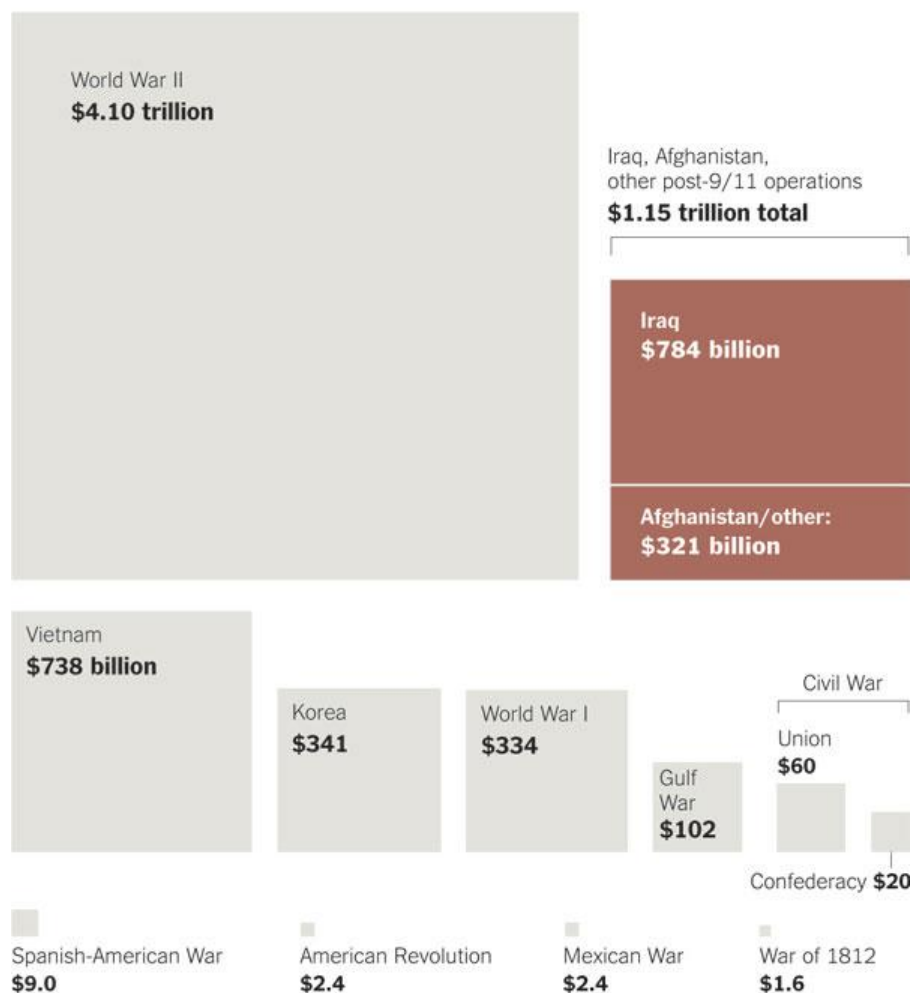
CHART of the **National Debt** of **BRITAIN** from the **REVOLUTION** to the End of the War with **AMERICA**

Millions

300
290
280
270
260
250
240
230
220
210
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10

End of American War

Beginning of American War

End of Continental War

Beginning of Continental War

End of the Spanish War

Beginning of the Spanish War

Accession of George 1st

Accession of George 2nd

Accession of Queen Ann

Revolution at the Accession of K. William

1688 1701 1714 1727 1739 1748 1755 1762 1775 1784

The Divisions at the Bottom are Years, & those on the Right hand Money.

This chart shows the British debt from "the Revolution" (i.e. the Glorious Revolution when William and Mary ascended) to the end of the American Revolution.

Some charts from the NYTimes looked at the costs of American wars (Week In Review July 25, 2010):
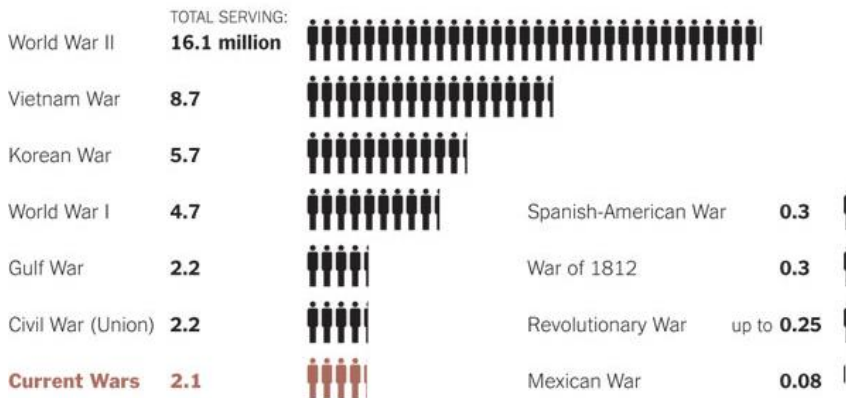
The Current Conflicts Are the Second-Most Expensive…

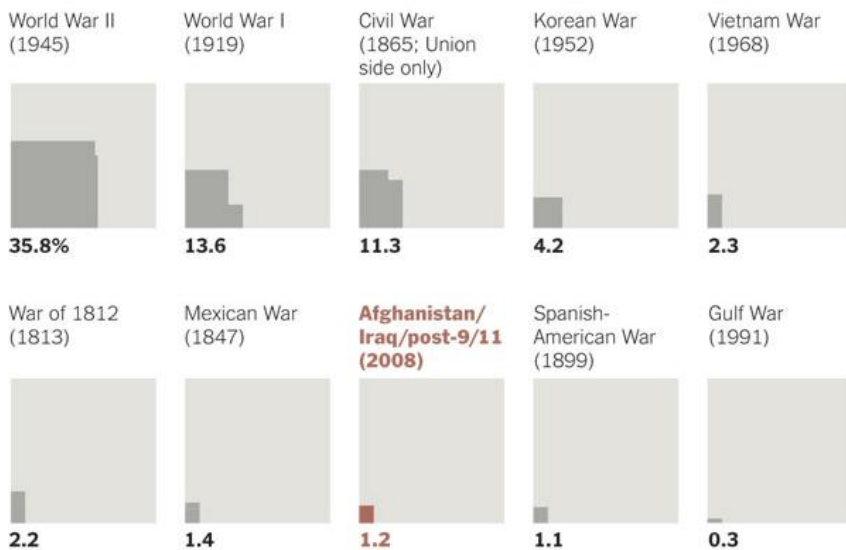The combined costs of Iraq, Afghanistan and other post-9/11 operations are exceeded only by those of World War II.

World War II
$4.10 trillion

Iraq, Afghanistan,
other post-9/11 operations
$1.15 trillion total

Iraq
$784 billion

Afghanistan/other:
$321 billion

Vietnam
$738 billion

Korea
$341

World War I
$334

Gulf
War
$102

Civil War

Union
$60

Confederacy **$20**

Spanish-American War
$9.0

American Revolution
$2.4

Mexican War
$2.4

War of 1812
$1.6

But They Have Involved Far Fewer People...

Each figure represents 500,000 people who served during the course of each conflict (includes those fighting in theaters of war and those serving in support areas).
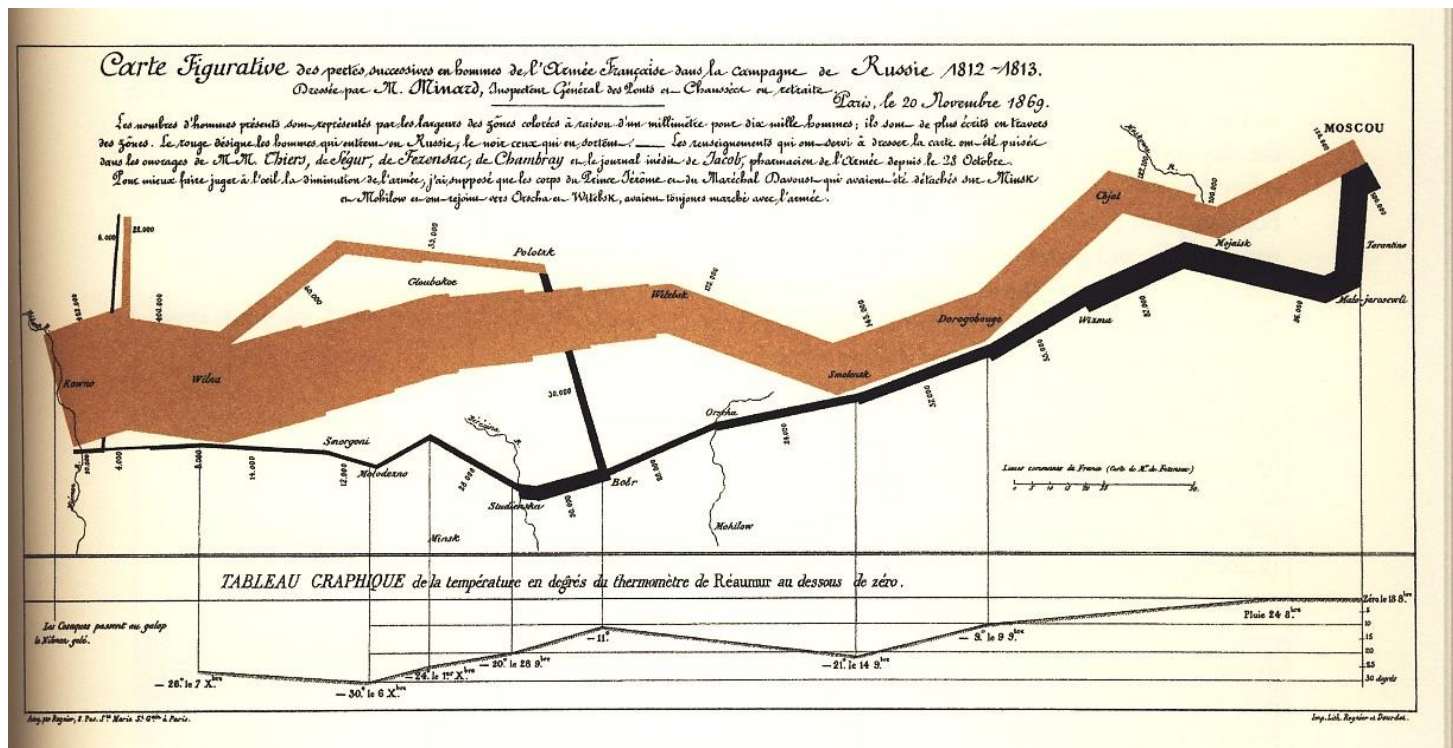
	TOTAL SERVING:	
World War II	16.1 million	
Vietnam War	8.7	
Korean War	5.7	
World War I	4.7	
Gulf War	2.2	
Civil War (Union)	2.2	
Current Wars	2.1	

Spanish-American War	0.3	
War of 1812	0.3	
Revolutionary War	up to 0.25	
Mexican War	0.08	

At a Smaller Cost to the Economy.

War costs as a percentage of United States gross domestic product in the peak year of each conflict. Estimate not available for Revolutionary War-era G.D.P.

World War II (1945)	World War I (1919)	Civil War (1865; Union side only)	Korean War (1952)	Vietnam War (1968)
35.8%	13.6	11.3	4.2	2.3

War of 1812 (1813)	Mexican War (1847)	Afghanistan/Iraq/post-9/11 (2008)	Spanish-American War (1899)	Gulf War (1991)
2.2	1.4	1.2	1.1	0.3

The modern guru of graphical display is Edward Tufte. This illustration, which he commonly uses and indentifies as "probably the best statistical graphic ever drawn", is by Charles Joseph Minard:

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
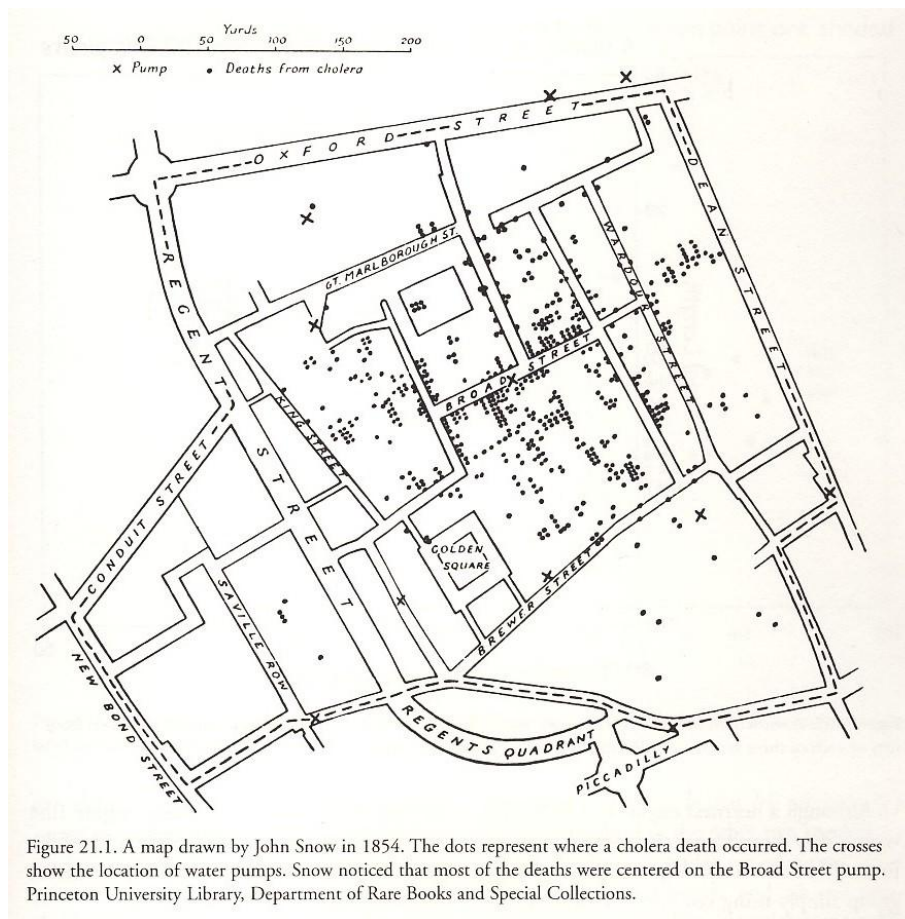
This shows Napoleon's invasion of Russia. The width of the line indicates the number of soldiers as well as how they moved to Moscow. The bottom graph shows the cold on the retreat (in degrees Reaumur; multiply by 4/5 to get Celsius so it bottomed at -24°C).

Wainer (among others) also pick out another important innovation, the early graphical display which had the most immediate improving effect on human welfare, by John Snow.

Snow's map shows deaths from cholera (the dots) as well as the location of well pumps (the x's) in London.

Figure 21.1. A map drawn by John Snow in 1854. The dots represent where a cholera death occurred. The crosses show the location of water pumps. Snow noticed that most of the deaths were centered on the Broad Street pump. Princeton University Library, Department of Rare Books and Special Collections.

The story tells that Snow went to the Broad Street pump, removed the handle, and shortly cholera deaths began to diminish. Prevailing public health theory had told that cholera was transmitted through the air; this chart and the associated experiment provide strong evidence that the problem was water quality.

Sometimes statistics reveal surprising things about the world; my colleague Prof. Bengoa noted a recent economic study showing that "size matters": there is a surprising correlation between economic growth and the average size of what the paper coyly refers to as "the male organ."

Figure 1: GDP in 1985 and the size of male organ in 76 countries, ORGAN in linear and quadratic form, $\bar{R}^2 = 0.15$
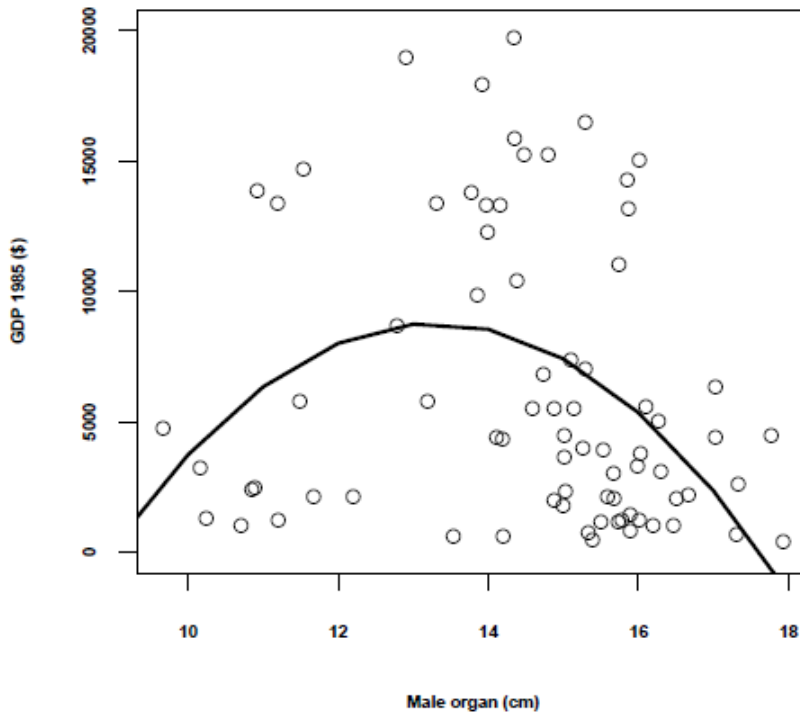
from "Male Organ and Economic Growth: Does Size Matter?" by Tatu Westling, July 2011.

There are many modern uses and surveys. The NY Times had this graphic showing the different uses of time during the day, gathered from the American Time Use Survey (which we'll use in class). Here http://www.nytimes.com/interactive/2009/07/31/business/20080801-metrics-graphic.html is the full interactive chart where you can compare the time use patterns of men and women, employed and unemployed, and other groups . (The article is http://www.nytimes.com/2009/08/02/business/02metrics.html?_r=2)
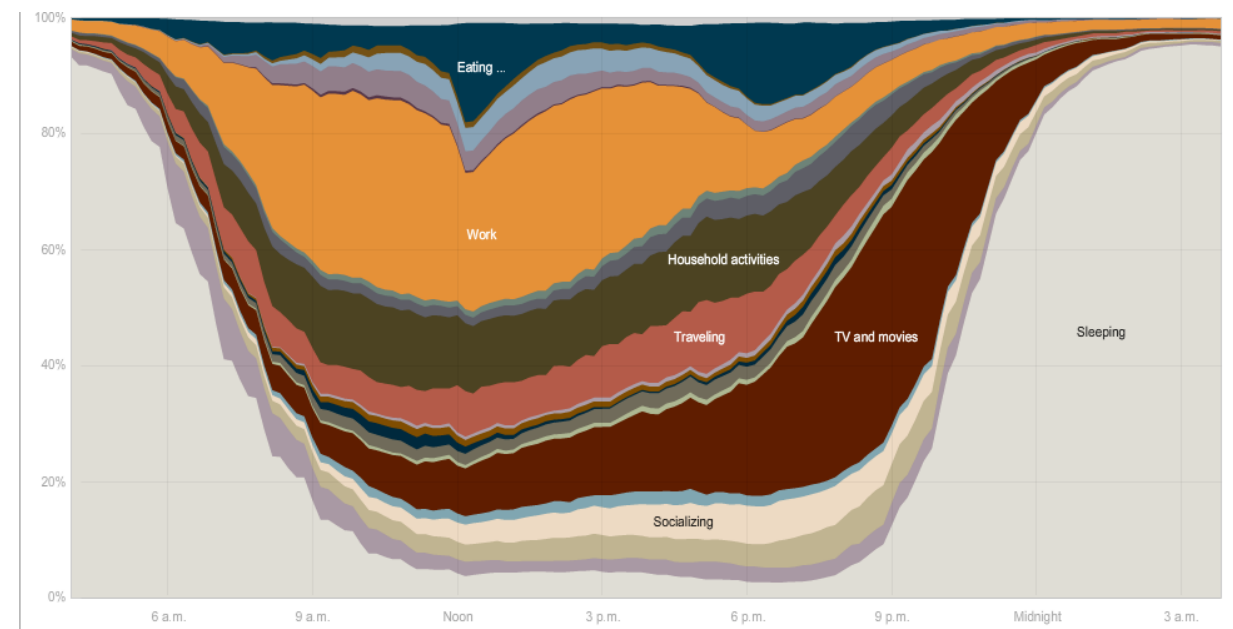
You can just browse online to find many beautiful examples of data presentation. Some are more effective as art; some are better at presenting the data.

The comic, xkcd, offers this recent graphic,
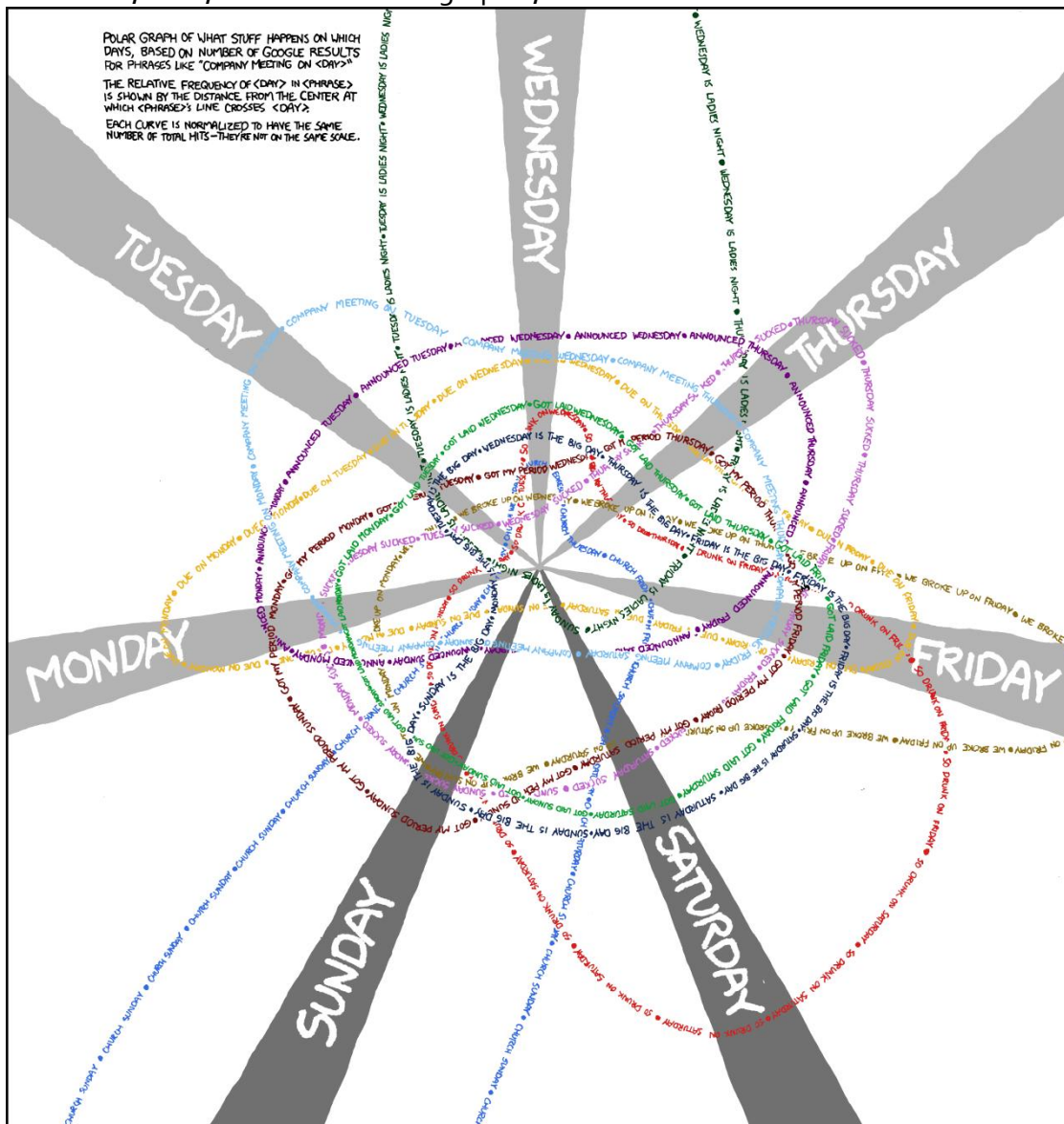
from xkcd.com/930/.

Many Eyes, from IBM, http://manyeyes.alphaworks.ibm.com/manyeyes/page/Visualization_Options.html, gives an overview of both the classics and more recent innovations such as word clouds.

Here is a word-cloud image of Shakespeare's most-used words,

Visual Complexity has some beautiful images, http://www.visualcomplexity.com/vc/.

Such as this one on NYC Subway Ridership 1905-2006, http://diametunim.com/shashi/nyc_subways/
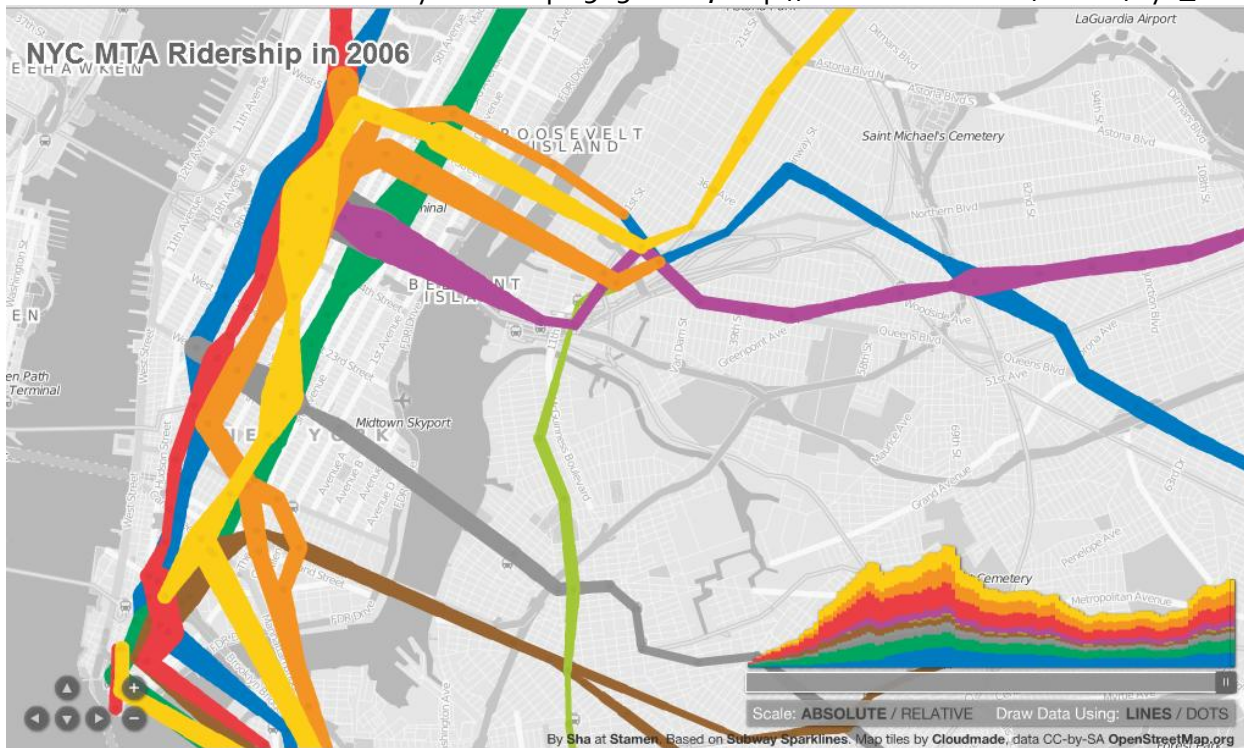
Note the slider on the right to change the time back to 1905.

Or here is a map of the moods of different places in the US, based on an analysis of Twitter feeds, http://www.visualisingdata.com/index.php/2010/07/twitter-visualisation-of-happiness/

state in question:
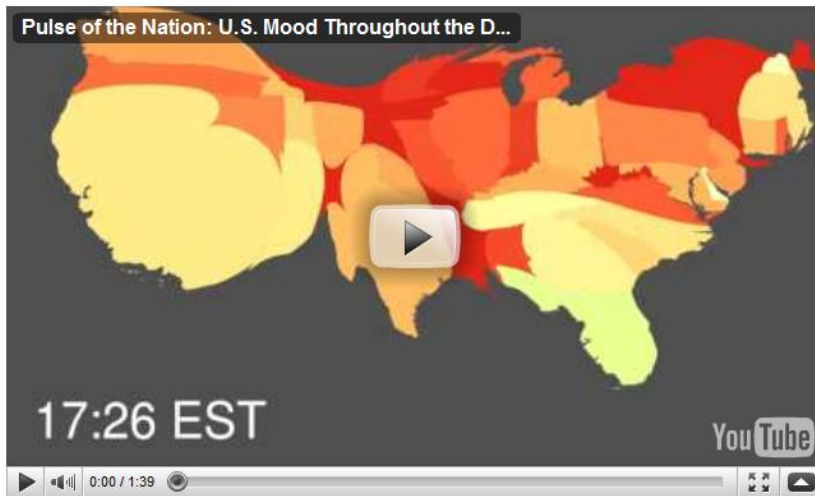
This is [an interesting site](#) that uses Amazon's Mechanical Turk (which blasts simple calculations to many human workers) to do things like rate pictures, to determine how attractiveness changes with age for men and women:
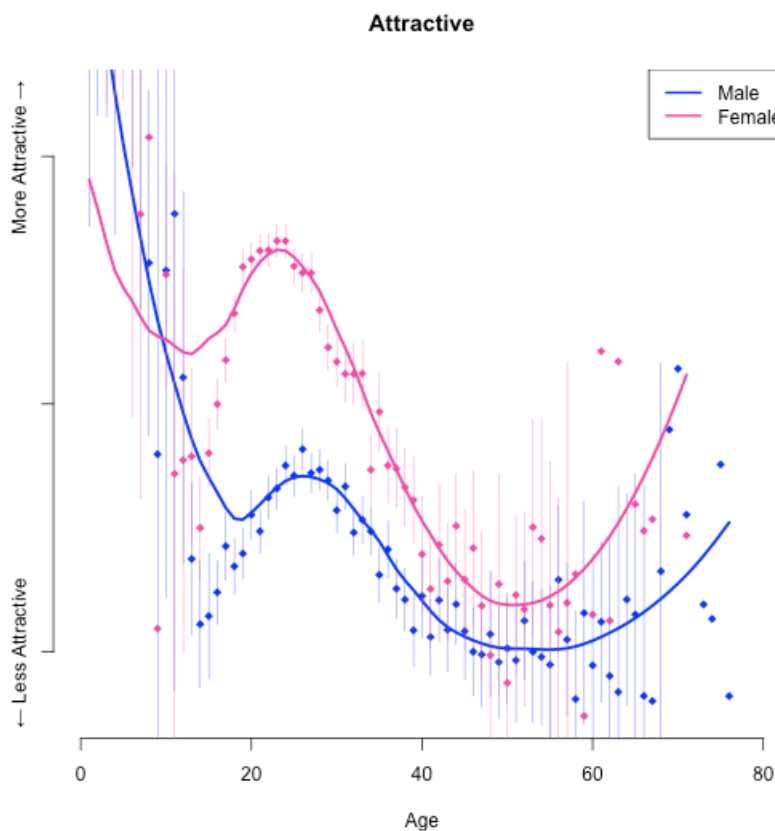
Where, except for babies, women are consistently rated as more attractive although the variation across age is interesting. The [whole paper](#) about perceptions of people based on thumbnail pictures is here.

Then there are the awful statistics. From *Damned Lies and Statistics*, Joel Best nominates the single worst,

> When I read the quotation, I assumed the student had made an error in copying it. I went to the library and looked up the article the student had cited. There, in the journal's 1995 volume, was exactly the same sentence: "Every year since 1950, the number of American children gunned down has doubled."

(Note that 2^{45} is more than 35 trillion.)

How?

Hopefully you're convinced that you want to learn stats. How can you best do it? Previous students have provided advice on a survey: "Don't take the class lightly. Make sure you take your time and make sure that you understand everything," "Study every day," "do every homework," "form a study group."

Bottom line: it takes work. If you work hard, you'll succeed. If you don't, you'll fail. It's your choice.

The class syllabus explains how we'll get there. Below I provide some specifics on the skills that you will be expected to have by the end of the term.

Want to learn more, about how to do good and avoid bad?

If you begin a love affair with Statistics and want to read more, here are some suggestions of books:
- Leonard Mlodinow, *Drunkard's Walk*
- Edward R. Tufte *The Visual Display of Quantitative Information*, *Visual Explanations: Images and Quantities, Evidence and Narrative* (in library)
- Howard Wainer, *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*
- David Salsburg, *Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*
- James Stock & Mark Watson, *Introduction to Econometrics* and Peter Kennedy, *A Guide to Econometrics*
- Jane E. Miller, *The Chicago Guide to Writing about Numbers* (in library)
- John W. Tukey, *Exploratory Data Analysis* (in library)
- Stephen Stigler, *Statistics on the Table* (in library) and *The History of Statistics: The Measurement of Uncertainty before 1900* (in library)
- Dierdre McCloskey , *Economical Writing* and *The Rhetoric of Economics* (in library)

Visualization/Stats/Econometrics blogs

Visualization
- http://www.visualisingdata.com/
- http://infosthetics.com/
- http://www.informationisbeautiful.net/
- http://fellinlovewithdata.com/
- http://flowingdata.com/
- http://smartdatacollective.com
- http://www.b-eye-network.com
- http://www.information-management.com/
- http://www.kdnuggets.com
- http://www.analyticsbridge.com

- http://journal.drawar.com/d/tag/lawofsimplicity/

Stats
- To learn basics, Khan Academy is good – see this article from Wired
- http://junkcharts.typepad.com/numbersruleyourworld/
- http://statisticsforum.wordpress.com/
- http://andrewgelman.com/

Econometrics
- http://davegiles.blogspot.com/
- http://freakonometrics.blog.free.fr (I admit, I run the French through Google Translate if it seems interesting enough)

Here's one on the Dark Side,
- http://wiki.darkpatterns.org/Home

Economics
Of course just about every economics blog is intensely focused on data, trying to make sense of how the economy works, such as
- http://www.calculatedriskblog.com
- http://economistsview.typepad.com/
- http://www.econbrowser.com.

Many of these are from a list called "Great web sites for Analytic people"
http://www.analyticbridge.com/profiles/blogs/great-web-sites-for-analytic

For a longer list (hundreds), see http://www.datashaping.com/analytics_links.shtml.