Important Dates

- Assignment #1: Intro to Python – Opens 9/13
- Research Proposal Instructions will be handed out. (9/13)
- Assignment #2: Applying Analysis Techniques and Statistical Inference to Data: Opens 10/4
- Assignment #1: Due 10/4
- Students Research Proposals Due: (10/4)

Agenda

- **Guest Lecture: Ruth Agada from BSU**
- **Lecture 4: Data Visualization**
- **Lab:**
  - **Web Scraping**
  - **EDA**

Why is Graphics a topic in this course?

- Visualization belongs in every stage of the data life cycle
- Plots can uncover structure in data that can't be detected from numerical summaries
- Visualization is an important communication skill

Goals of this lecture

- Guidelines and general philosophy
  - Reveal the data
  - Facilitate Comparisons
  - Add information
  - Iterate
- Techniques for following guidelines
  - Scale
  - Conditioning
  - Perception
  - Transformations
  - Adding context
  - Smoothing & other large data considerations

Good Starting Place – Know your data type

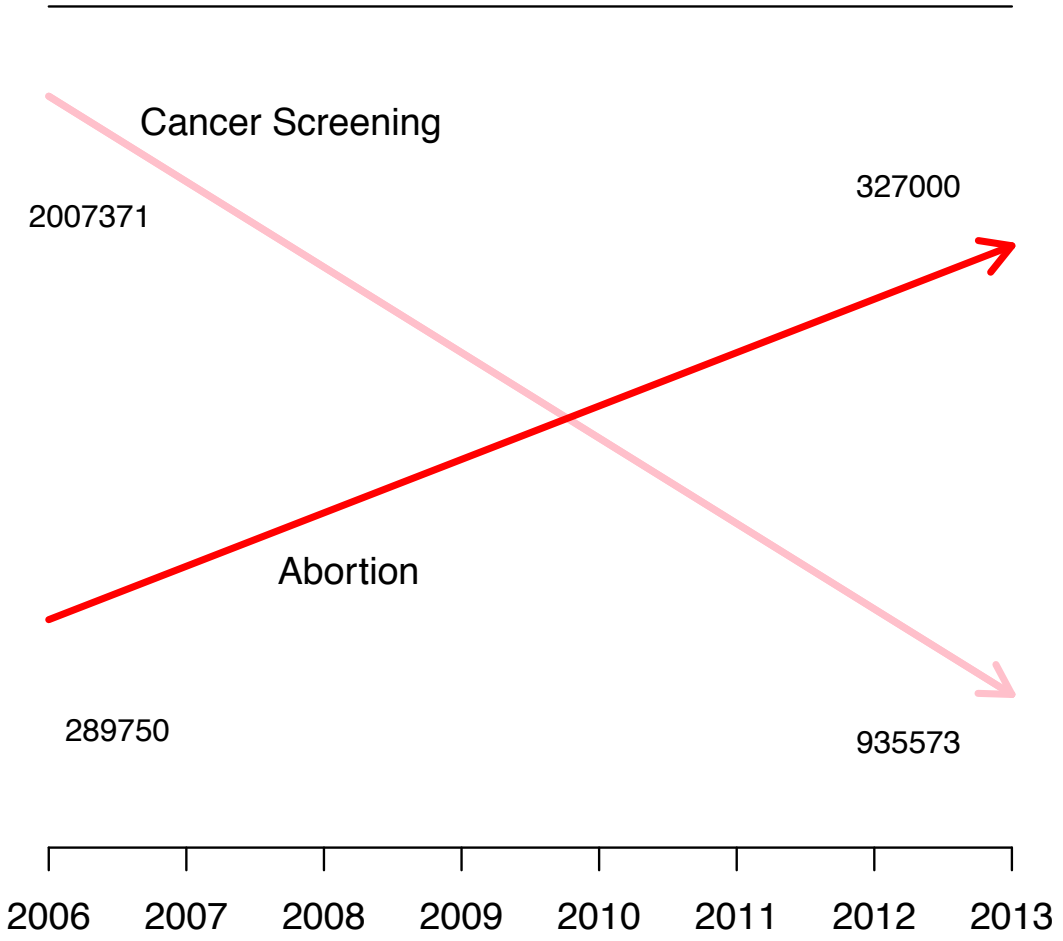
- Quantitative (Numeric)
  - Continuous (e.g., health care expenditure)
  - Discrete (e.g., number of siblings)
- Qualitative (Categorical)
  - Nominal (e.g., lane of traffic, country)
  - Ordinal (e.g., Yelp rating, education level)
- See table that maps data types to plot types at end of slides

Examples

2015 Congressional Hearing: Planned Parenthood

- Congressman Chaffetz (R-UT), chair US House Oversight Committee
- Investigation of federal funding of Planned Parenthood
- Chaffetz showed a plot which originally appeared in a report by Americans United for Life (http://www.aul.org/).
- Report available at https:// oversight.house.gov/ interactivepage/plannedparenthood/

Planned Parenthood Procedures



- Procedures:
  - Cancer screenings
  - Abortion
- Time: 2006 to 2013
- How many data points are in this plot?
- What's suspicious about this plot?

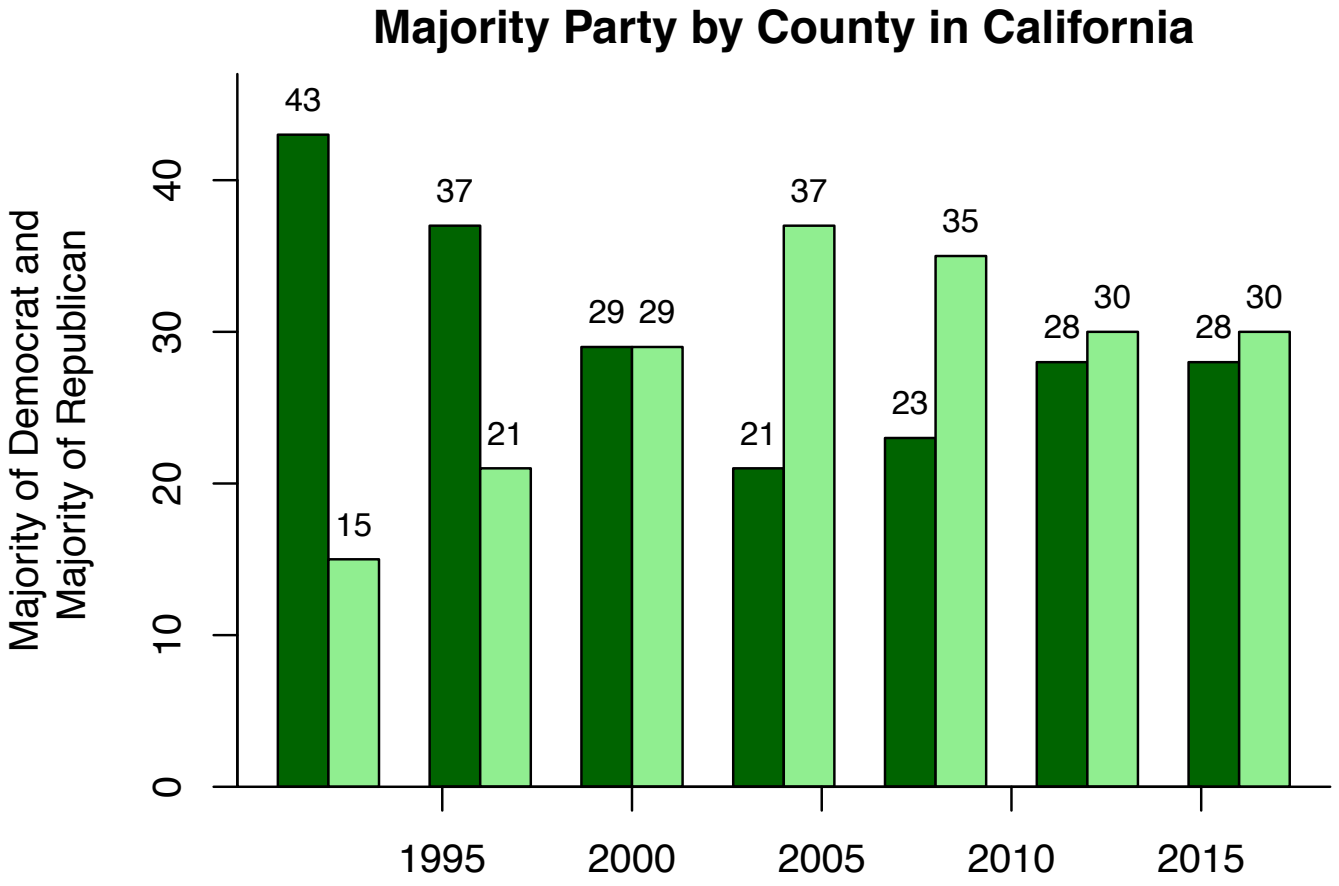
Voter Registration Trends in California

- State of California publishes voter registration summaries

http://www.sos.ca.gov/elections/ror/60day presprim/hist reg stats.pdf

- Historical registration counts available for presidential election years

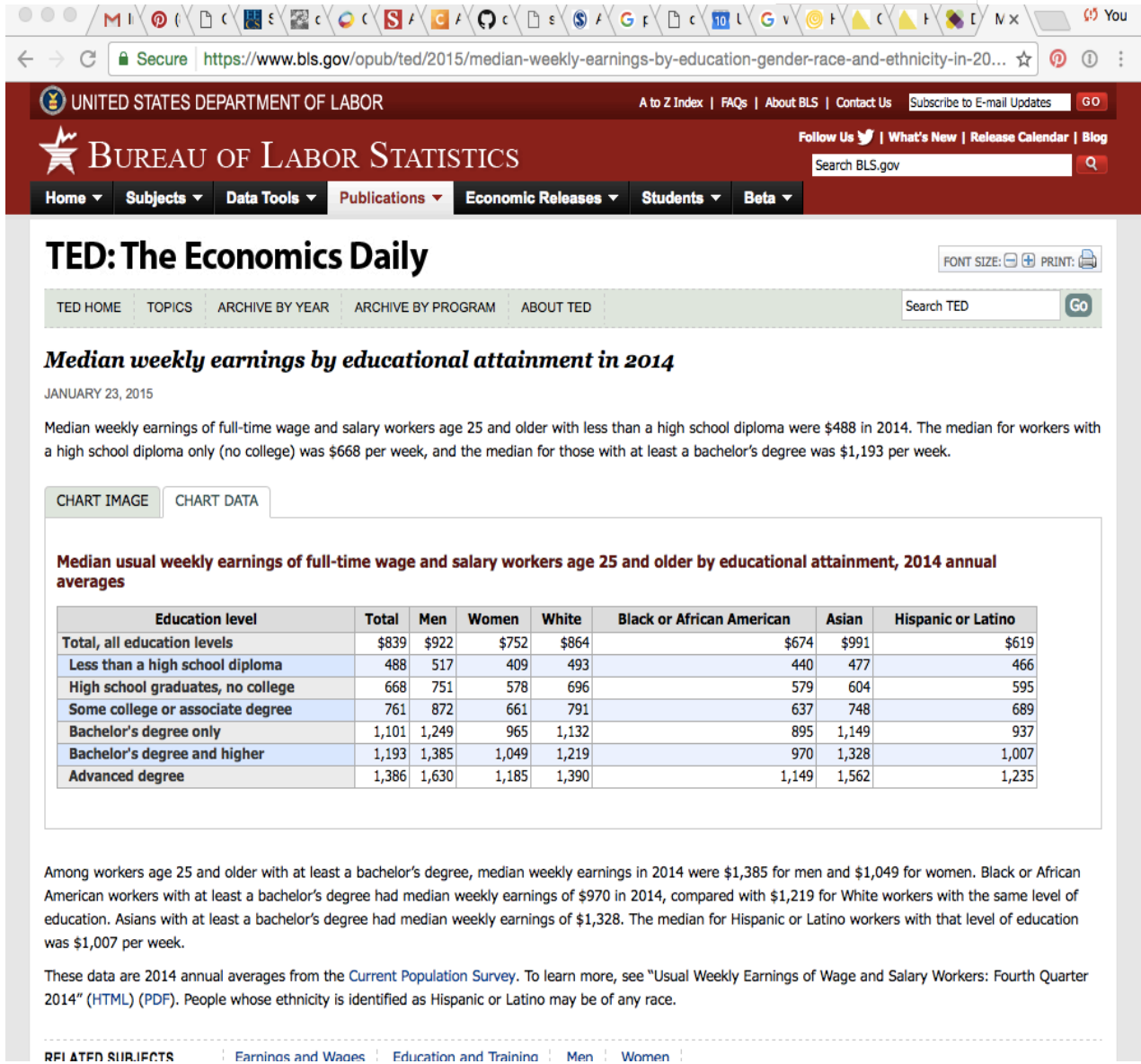
Voter Registration Trends in California



- What's confusing or annoying about this plot?

Earnings

- Bureau of Labor Statistics
  - Oversees scientific surveys related to economic health of the country
- Current Population Survey
  - Collects data on the earnings
  - www.bls.gov - Web interface to a report generating app



Secure https://www.bls.gov/opub/ted/2015/median-weekly-earnings-by-education-gender-race-and-ethnicity-in-20...

UNITED STATES DEPARTMENT OF LABOR

A to Z Index | FAQs | About BLS | Contact Us | Subscribe to E-mail Updates GO

BUREAU OF LABOR STATISTICS

Follow Us | What's New | Release Calendar | Blog

Search BLS.gov

Home | Subjects | Data Tools | Publications | Economic Releases | Students | Beta

TED: The Economics Daily

FONT SIZE: PRINT:

TED HOME | TOPICS | ARCHIVE BY YEAR | ARCHIVE BY PROGRAM | ABOUT TED | Search TED Go

Median weekly earnings by educational attainment in 2014

JANUARY 23, 2015

Median weekly earnings of full-time wage and salary workers age 25 and older with less than a high school diploma were $488 in 2014. The median for workers with a high school diploma only (no college) was $668 per week, and the median for those with at least a bachelor's degree was $1,193 per week.

CHART IMAGE | CHART DATA

Median usual weekly earnings of full-time wage and salary workers age 25 and older by educational attainment, 2014 annual averages

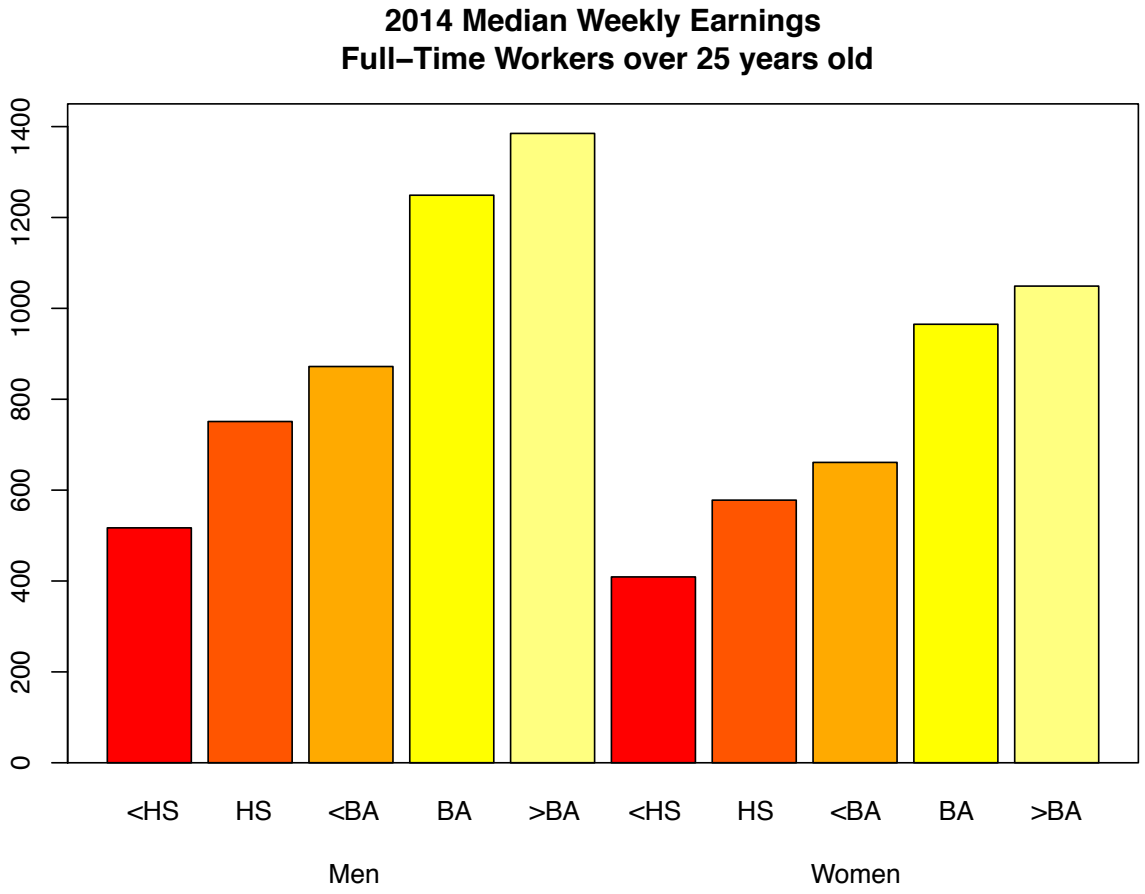
Education level	Total	Men	Women	White	Black or African American	Asian	Hispanic or Latino
Total, all education levels	$839	$922	$752	$864	$674	$991	$619
Less than a high school diploma	488	517	409	493	440	477	466
High school graduates, no college	668	751	578	696	579	604	595
Some college or associate degree	761	872	661	791	637	748	689
Bachelor's degree only	1,101	1,249	965	1,132	895	1,149	937
Bachelor's degree and higher	1,193	1,385	1,049	1,219	970	1,328	1,007
Advanced degree	1,386	1,630	1,185	1,390	1,149	1,562	1,235

Among workers age 25 and older with at least a bachelor's degree, median weekly earnings in 2014 were $1,385 for men and $1,049 for women. Black or African American workers with at least a bachelor's degree had median weekly earnings of $970 in 2014, compared with $1,219 for White workers with the same level of education. Asians with at least a bachelor's degree had median weekly earnings of $1,328. The median for Hispanic or Latino workers with that level of education was $1,007 per week.

These data are 2014 annual averages from the Current Population Survey. To learn more, see "Usual Weekly Earnings of Wage and Salary Workers: Fourth Quarter 2014" (HTML) (PDF). People whose ethnicity is identified as Hispanic or Latino may be of any race.

RELATED SUBJECTS | Earnings and Wages | Education and Training | Men | Women

Earnings

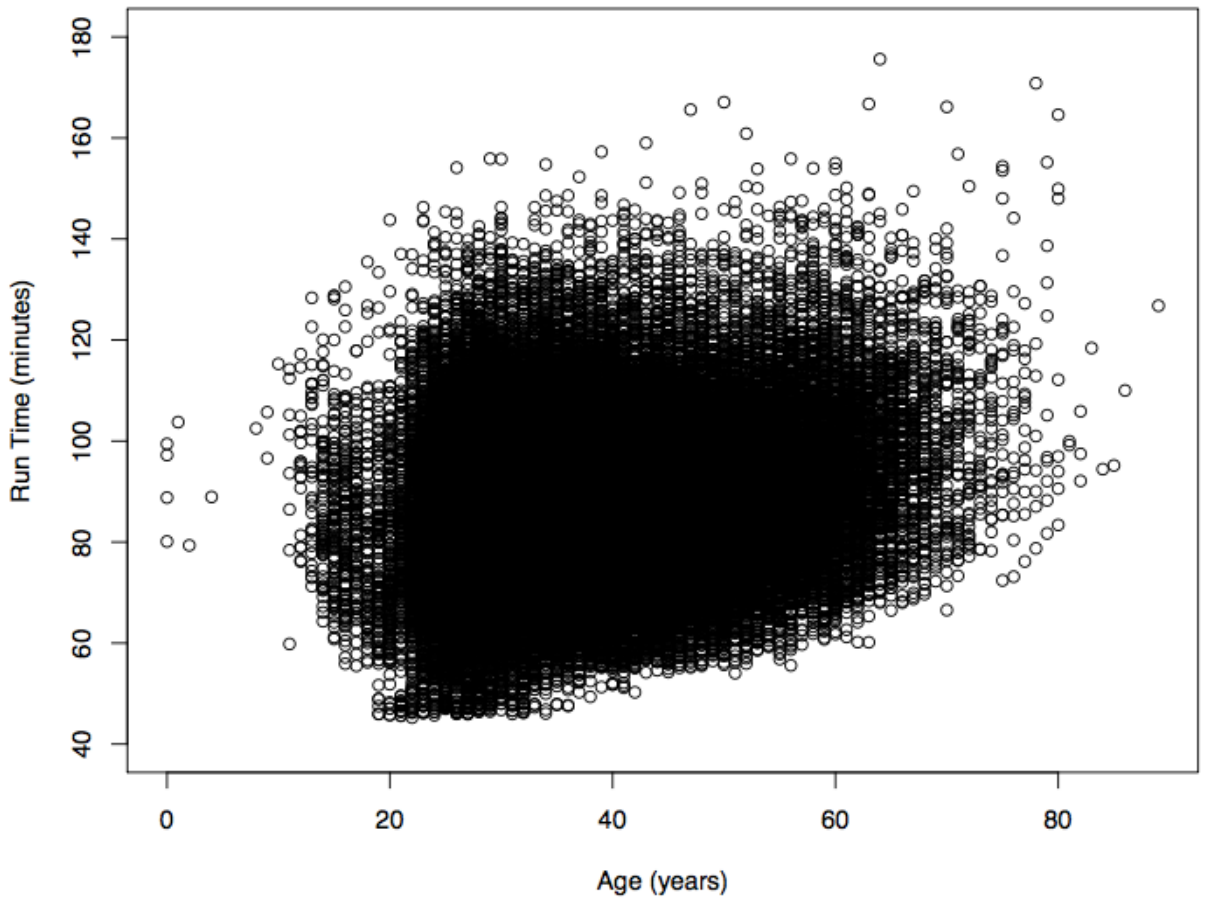


- Which comparisons can be easily made with this plot?

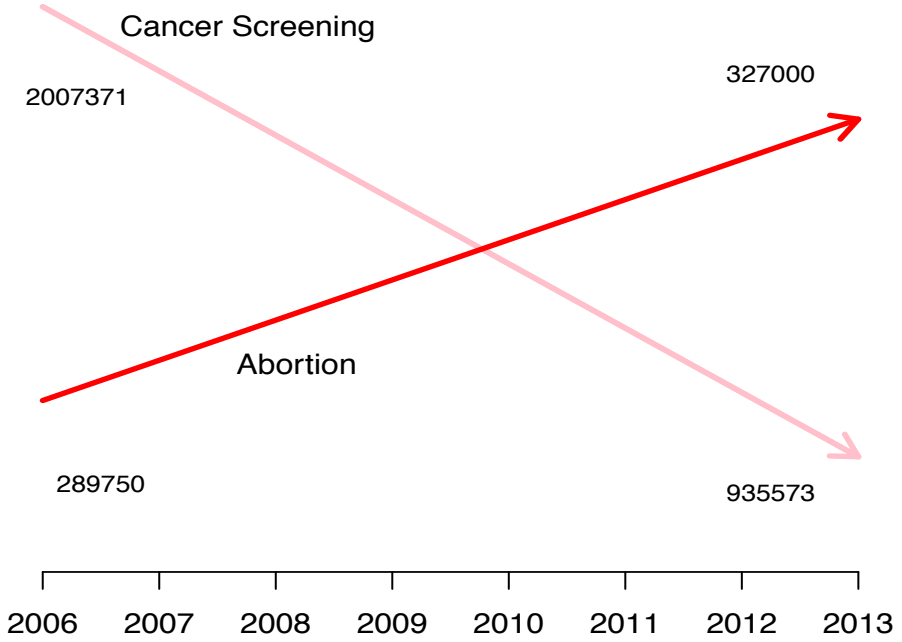
Cherry Blossom Run

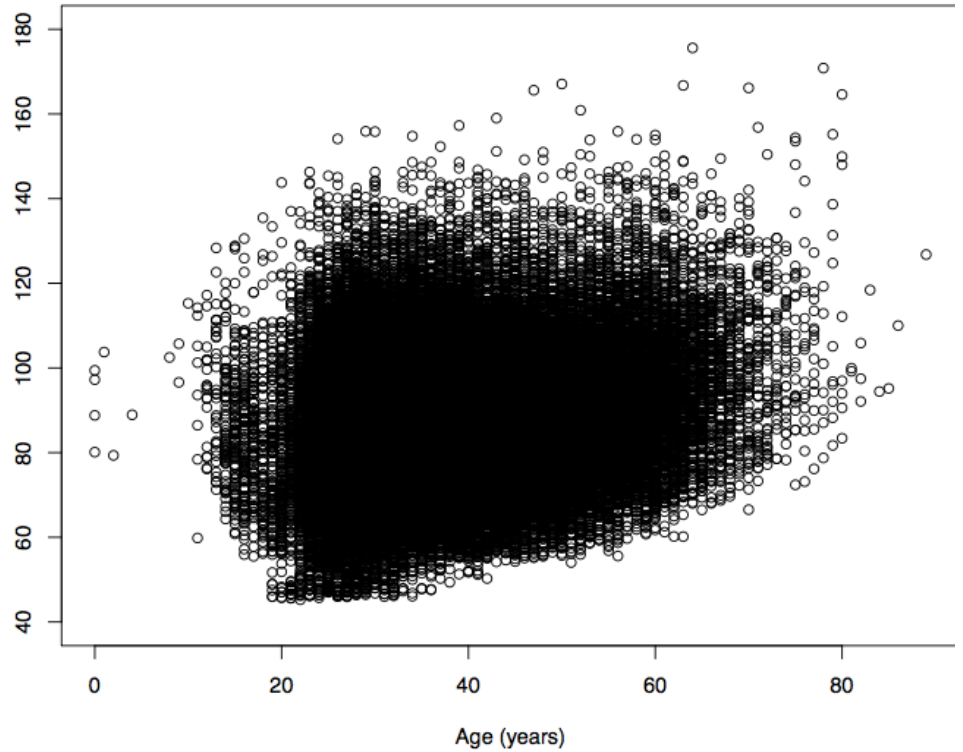
- 10 mile run in Washington DC each April
- Race organizers make results available on Web
  - Runner name, age, gender, address, hometown, time
  - Race results from 1999 to 2016
  - In 2012 nearly 17,000 runners ranging in age from 9 to 89 participated
  - http://www.cherryblossom.org/

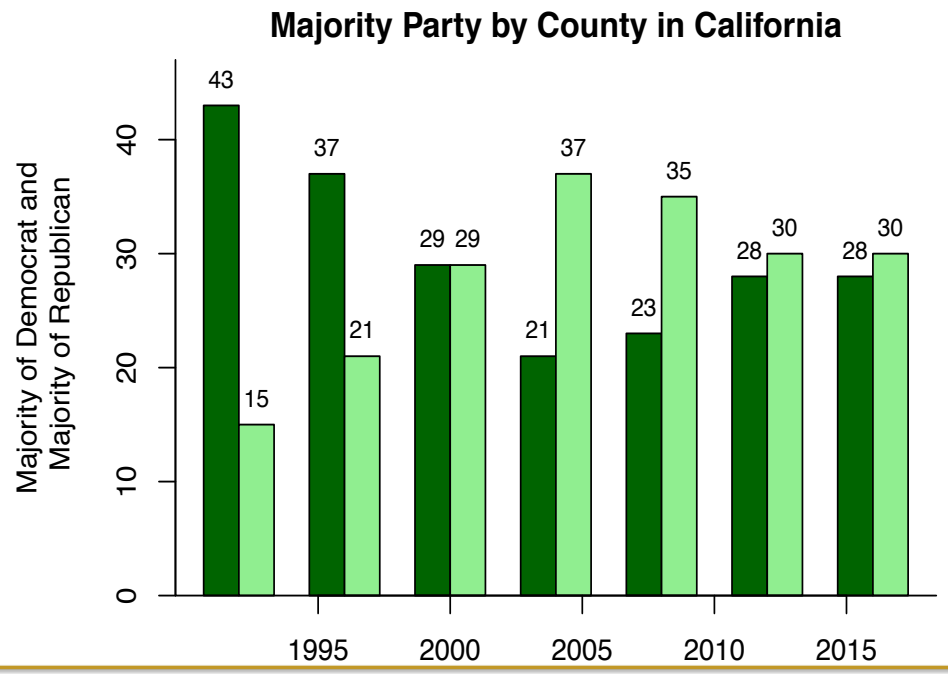
Cherry Blossom Run

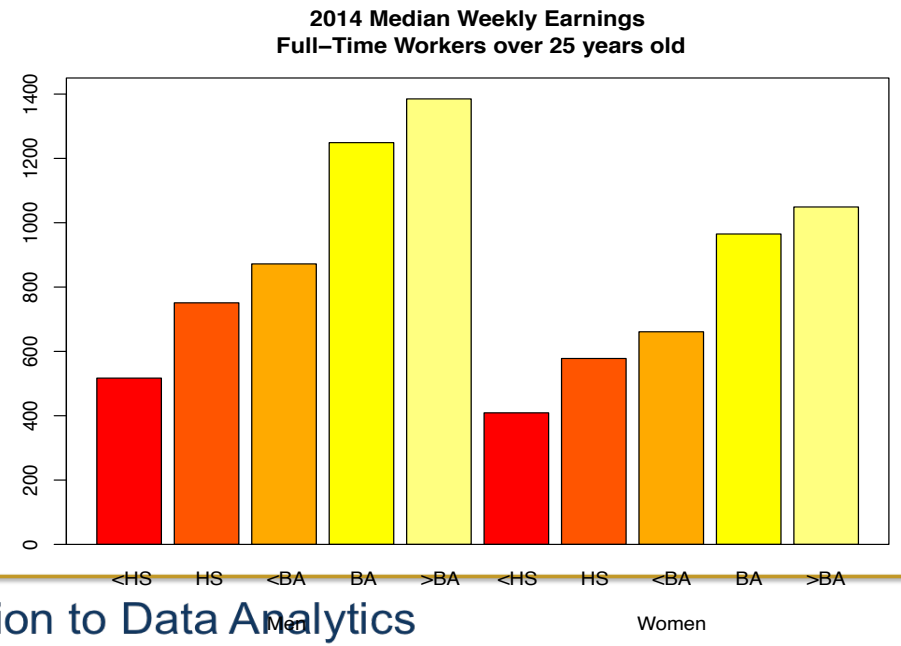


- Scatter plot of run time (minutes) by age (years)
- 70,000 points in this plot.
- What's the relationship between run time and age?


Cancer Screening
2007371
327000
Abortion
289750
935573
2006
2007
2008
2009
2010
2011
2012
2013

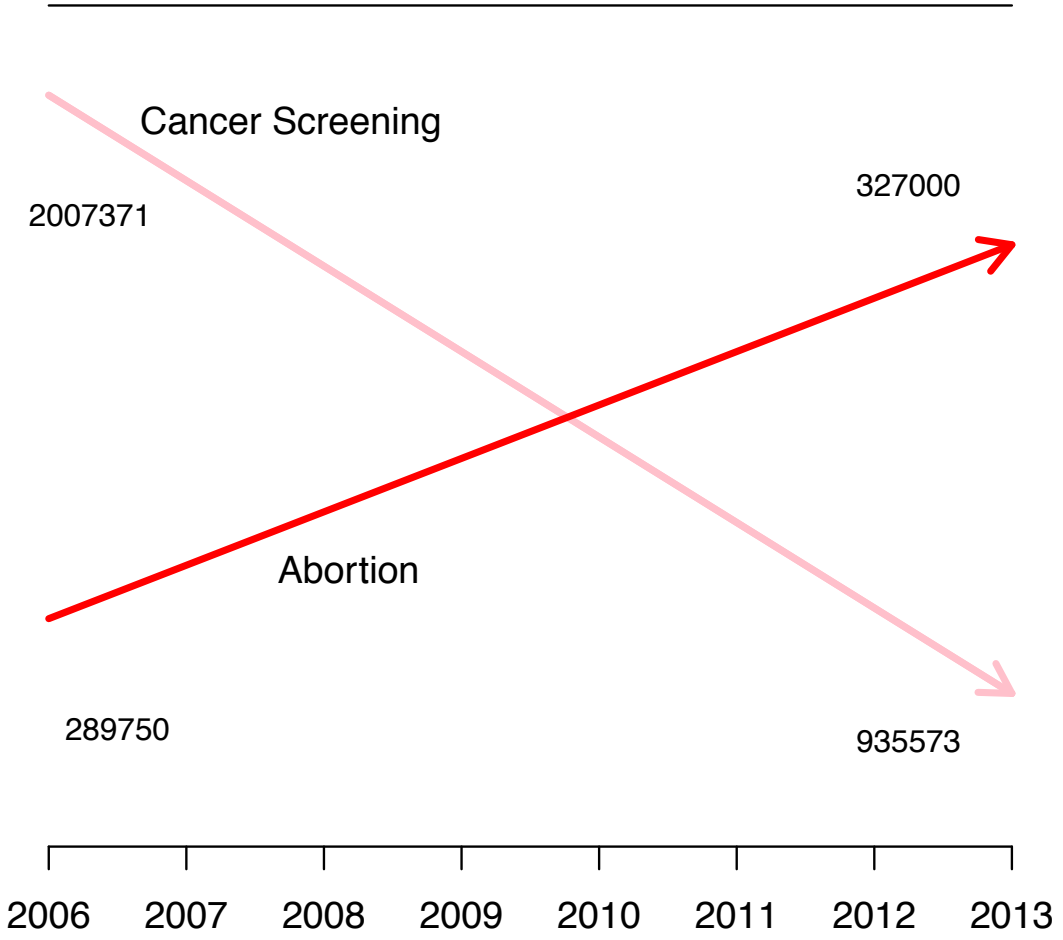

40
60
80
100
120
140
160
180
0
20
40
60
80
Age (years)

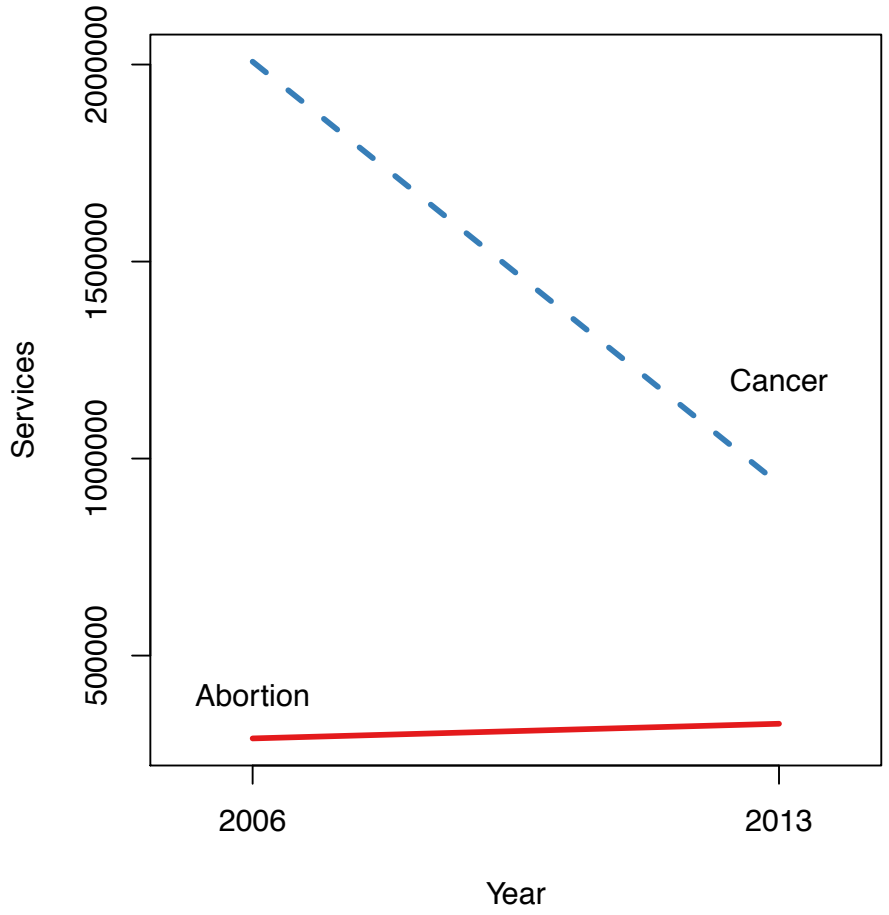

Majority Party by County in California
Majority of Democrat and Majority of Republican
0
10
20
30
40
43
15
37
21
29
29
21
37
23
35
28
30
28
30
1995
2000
2005
2010
2015


2014 Median Weekly Earnings
Full–Time Workers over 25 years old
0
200
400
600
800
1000
1200
1400
<HS
HS
<BA
BA
>BA
<HS
HS
<BA
BA
>BA
Men
Women

Techniques

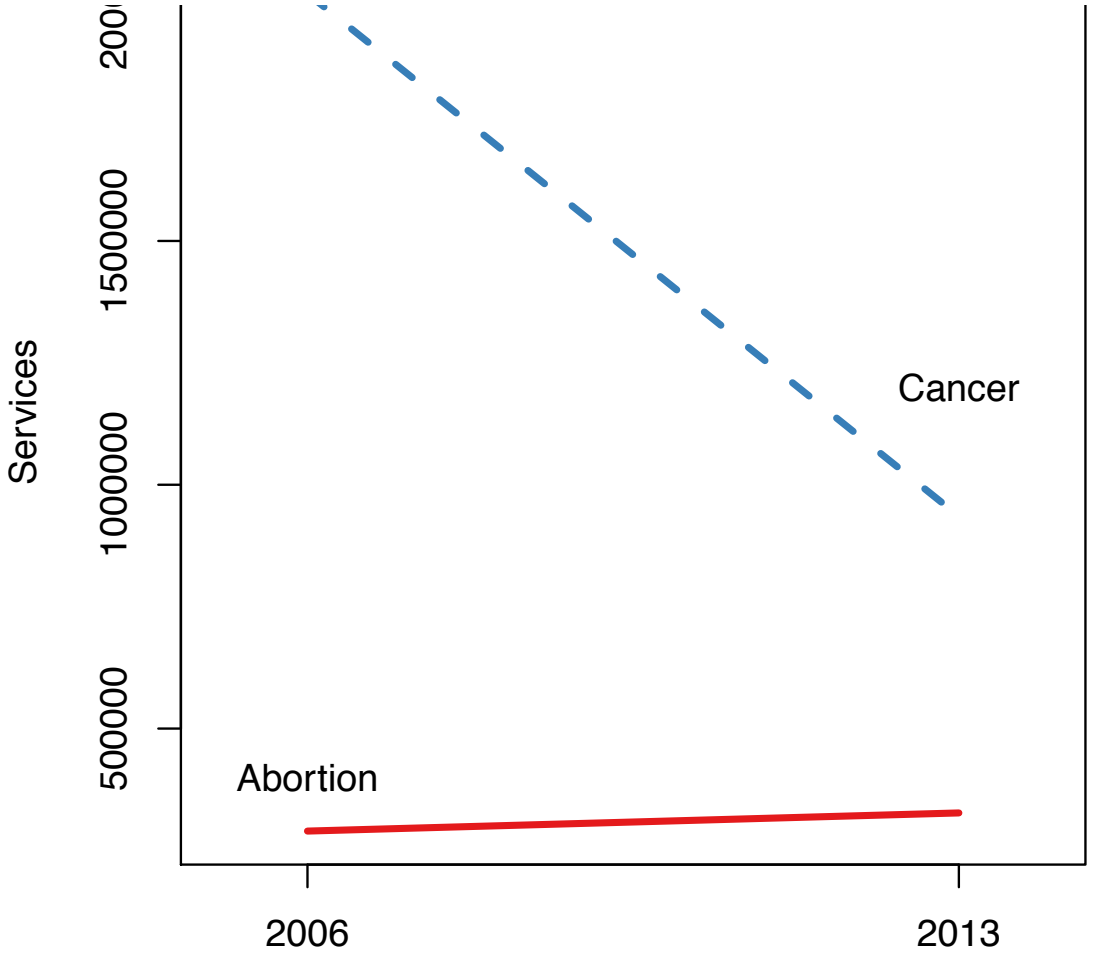
Planned Parenthood Procedures





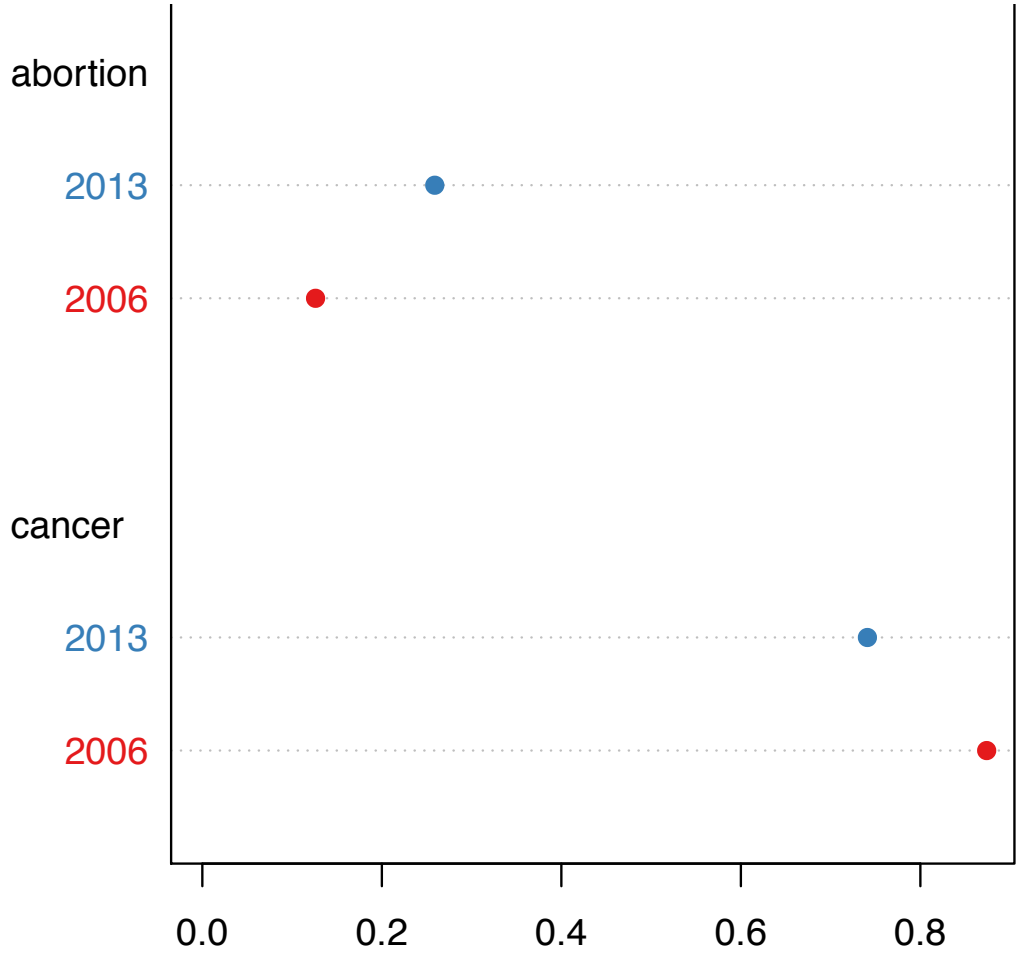
Scale

Planned Parenthood Procedures



- All points and lines are on the same scale
- How does this plot change the perception of the information?
- There has been a dramatic decrease in cancer screenings which dominates this plot
- The scales of the two procedures are very different – consider representing as percentages instead

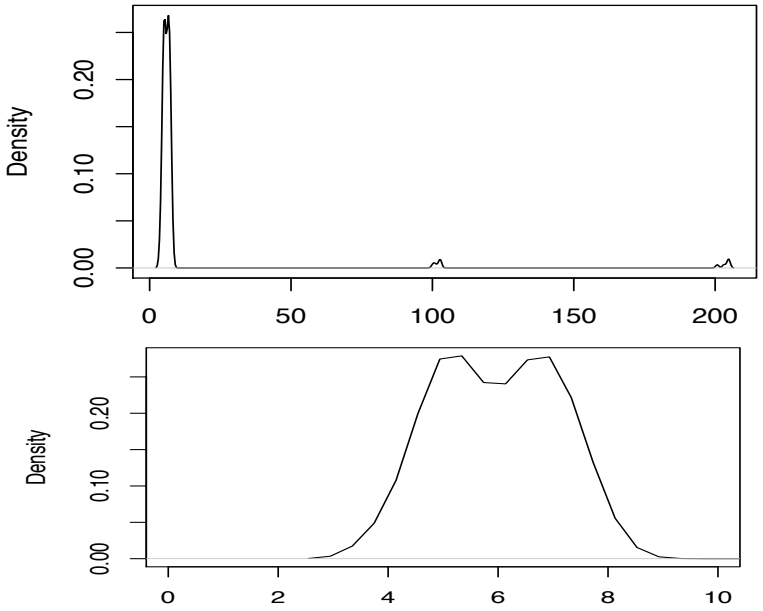
Planned Parenthood Procedures



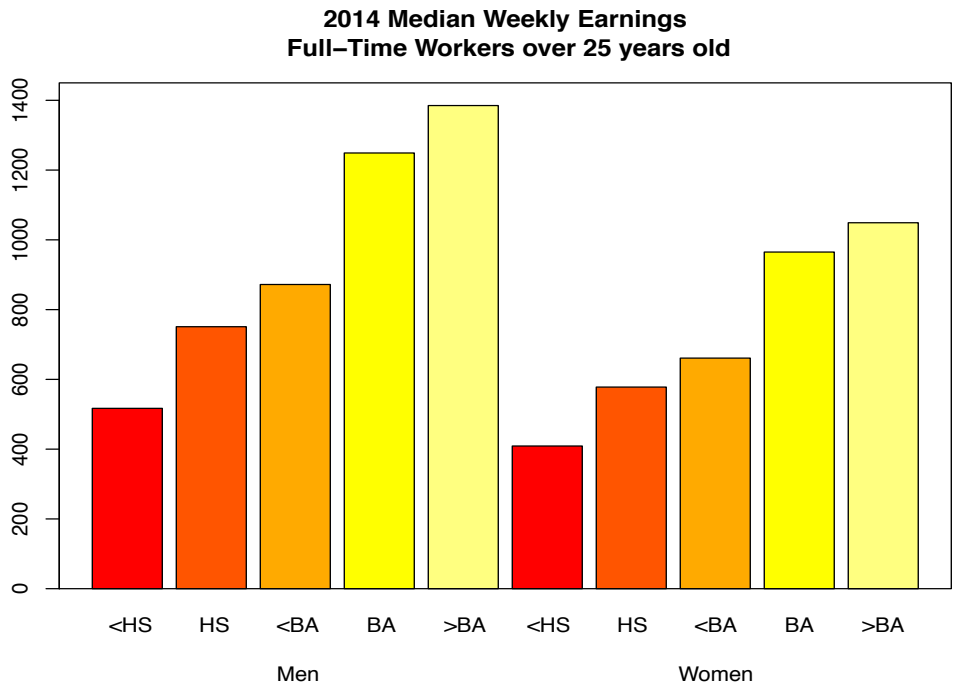
- Procedures in 2006 and 2013 as a percentage
- Abortions increased from 13% to 26% of total procedures
- May want to plot the percent change, screenings fell 50%

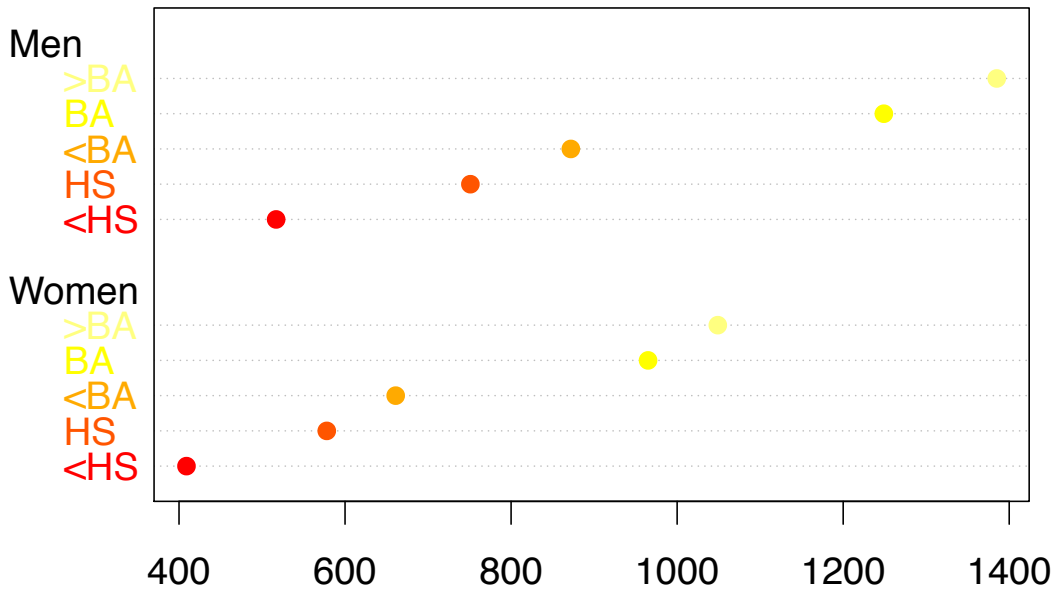
Choosing the Scale

- Choose axis limit to fill the plotting region
- If necessary,
  - Zoom in to focus on region with bulk of data
  - Make multiple plots of different regions
  - Transform data to improve resolution (TBC)
- Don’t change scale mid-axis
- Don’t use two different scales for the same axis



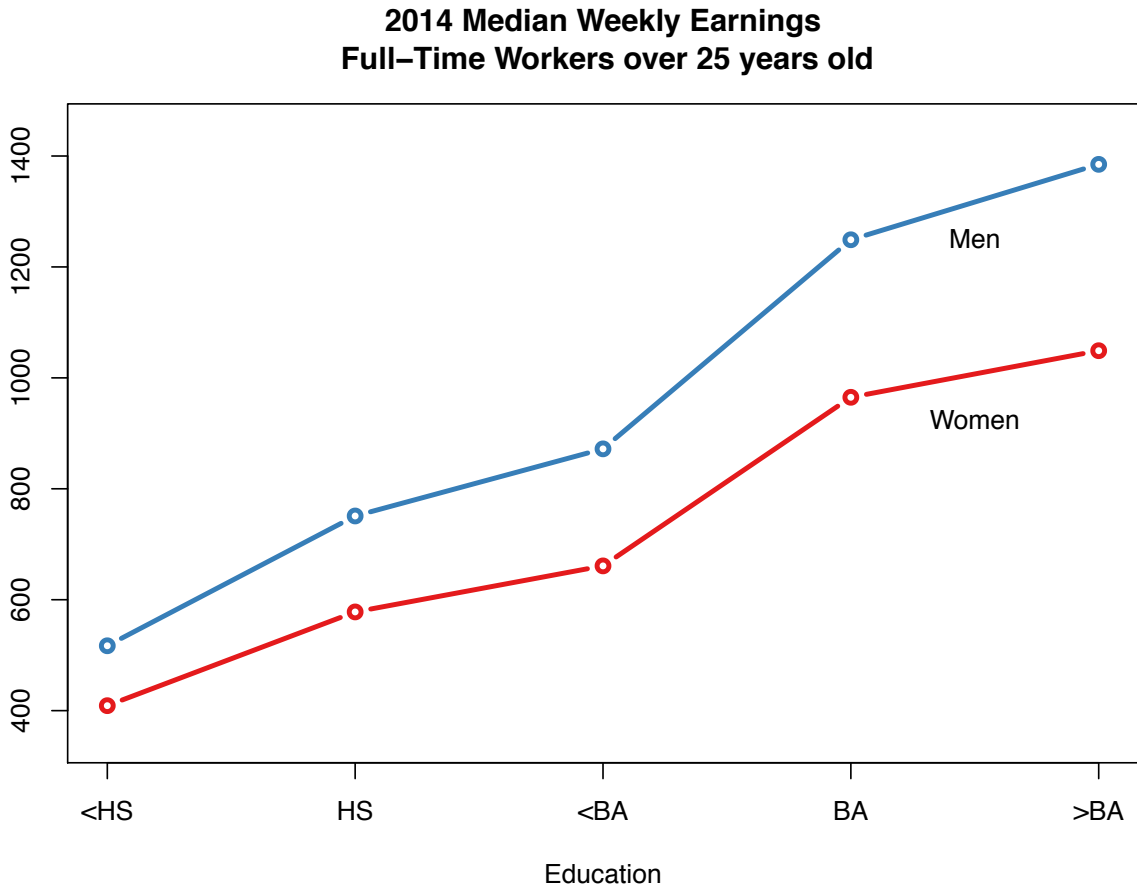
Earnings


2014 Median Weekly Earnings
Full–Time Workers over 25 years old
0
200
400
600
800
1000
1200
1400
<HS
HS
<BA
BA
>BA
<HS
HS
<BA
BA
>BA
Men
Women


Men
>BA
BA
<BA
HS
<HS
Women
>BA
BA
<BA
HS
<HS
400
600
800
1000
1200
1400

Conditioning

Earnings



- Emphasize the important difference –
- Lines make it easier to see growth in gap
- Placement of one point above the other makes it easier to compare males & females

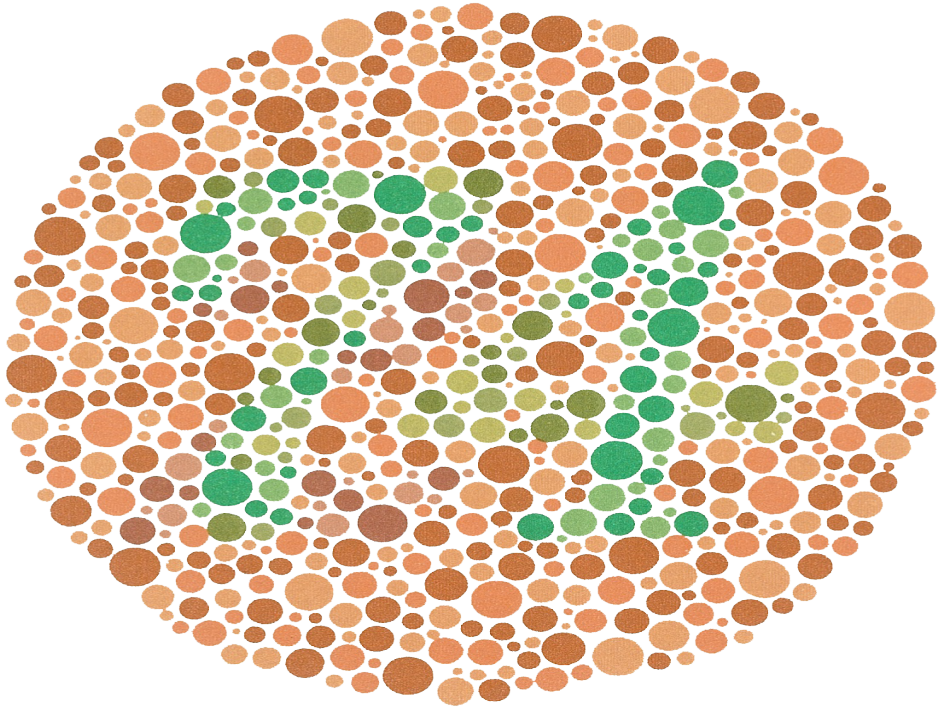
Conditioning – Distributions & Relationships in subgroups

- Superpose density curves, fitted curves and lines from different subgroups
- Juxtapose scatter plots, histograms & keep x and y scales the same across plots to facilitate comparison
- Use color and plotting symbols to represent additional variables

Perception - Color

Color Guidelines

- Choosing a set of colors which work well together is a challenging task for anyone who does not have an intuitive gift for color
- 7-10% of males are red-green color blind.

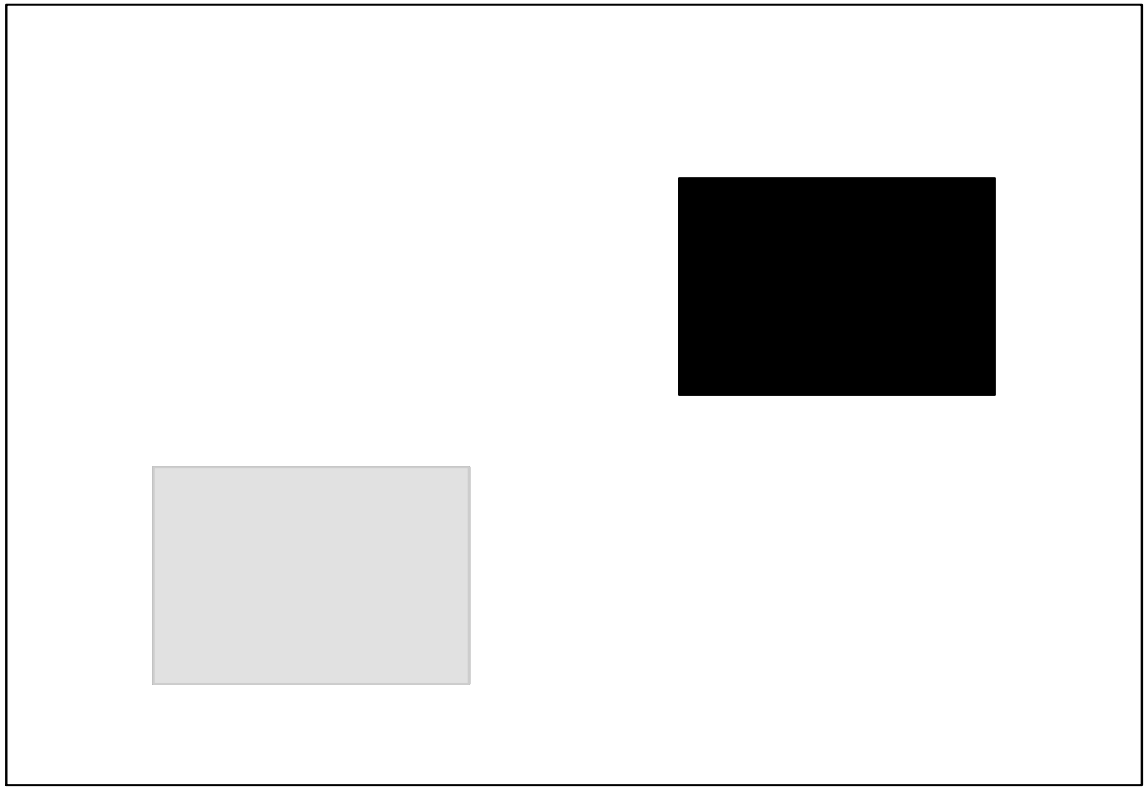


Colorfulness

- Saturated/colorful colors are hard to look at for a long time.
- They tend to produce an after-image effect which can be distracting.

Luminance

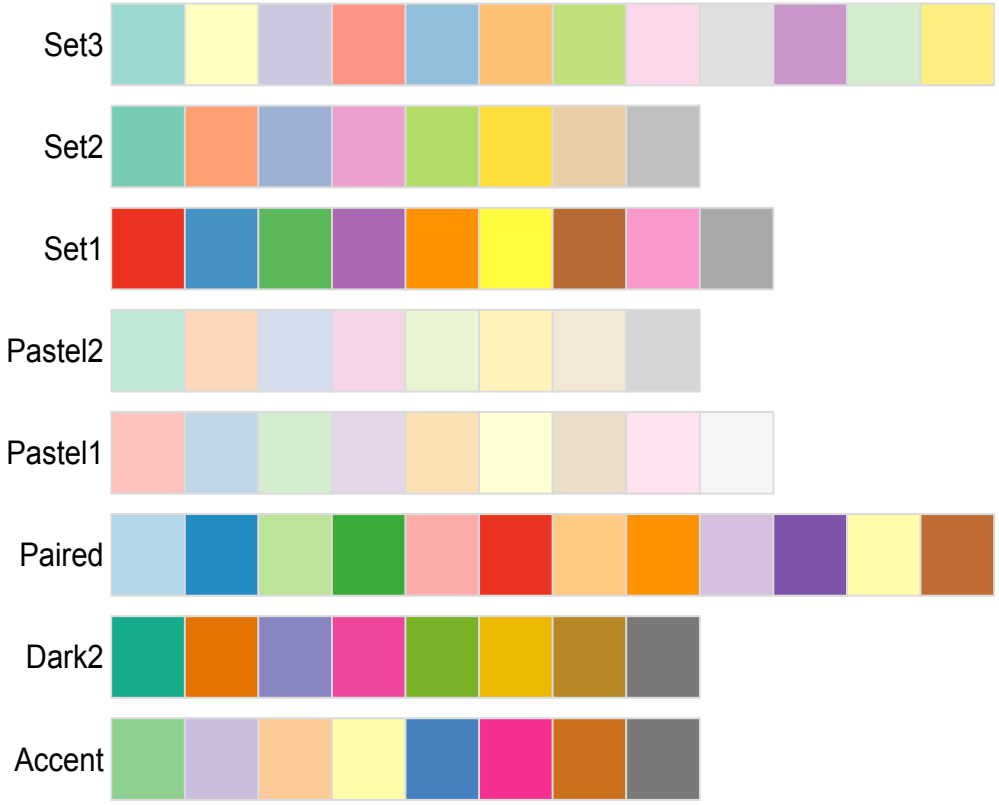
- Areas should be rendered with colors of similar luminance (brightness).
- Lighter colors tend to make areas look larger than darker colors



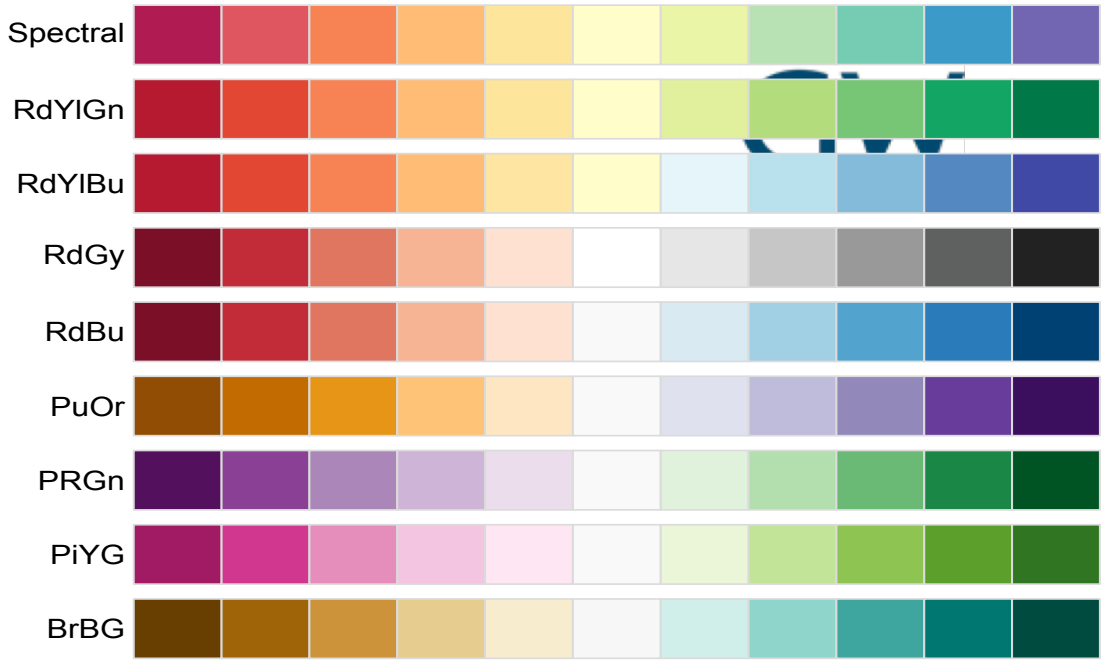
Data Type and Color

- Qualitative – Choose a **qualitative** scheme that makes it easy to distinguish between categories
- Quantitative – Choose a color scheme that implies magnitude.
  - Does the data progress from low to high? Use a **sequential** scheme where light colors are for low values
  - Do both low and high value deserve equal emphasis? Use a **diverging** scheme where light colors represent middle values

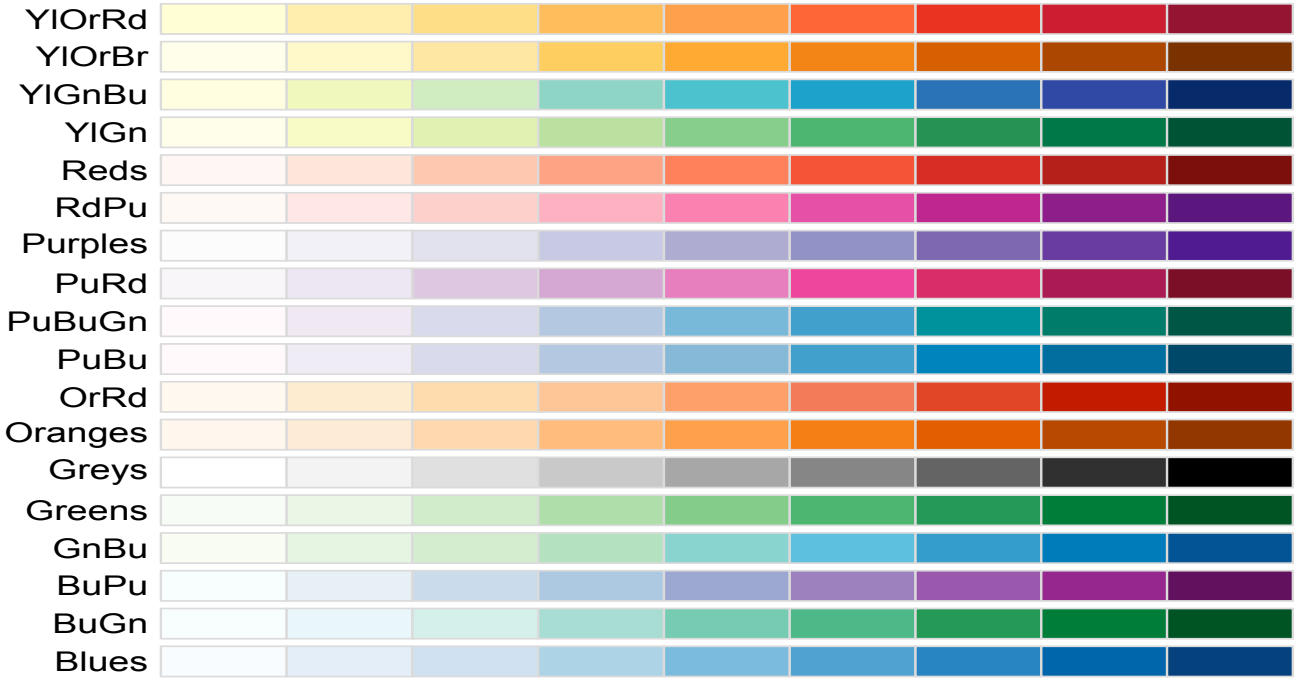
Examples of Palettes



Qualitative



Diverging



Sequential

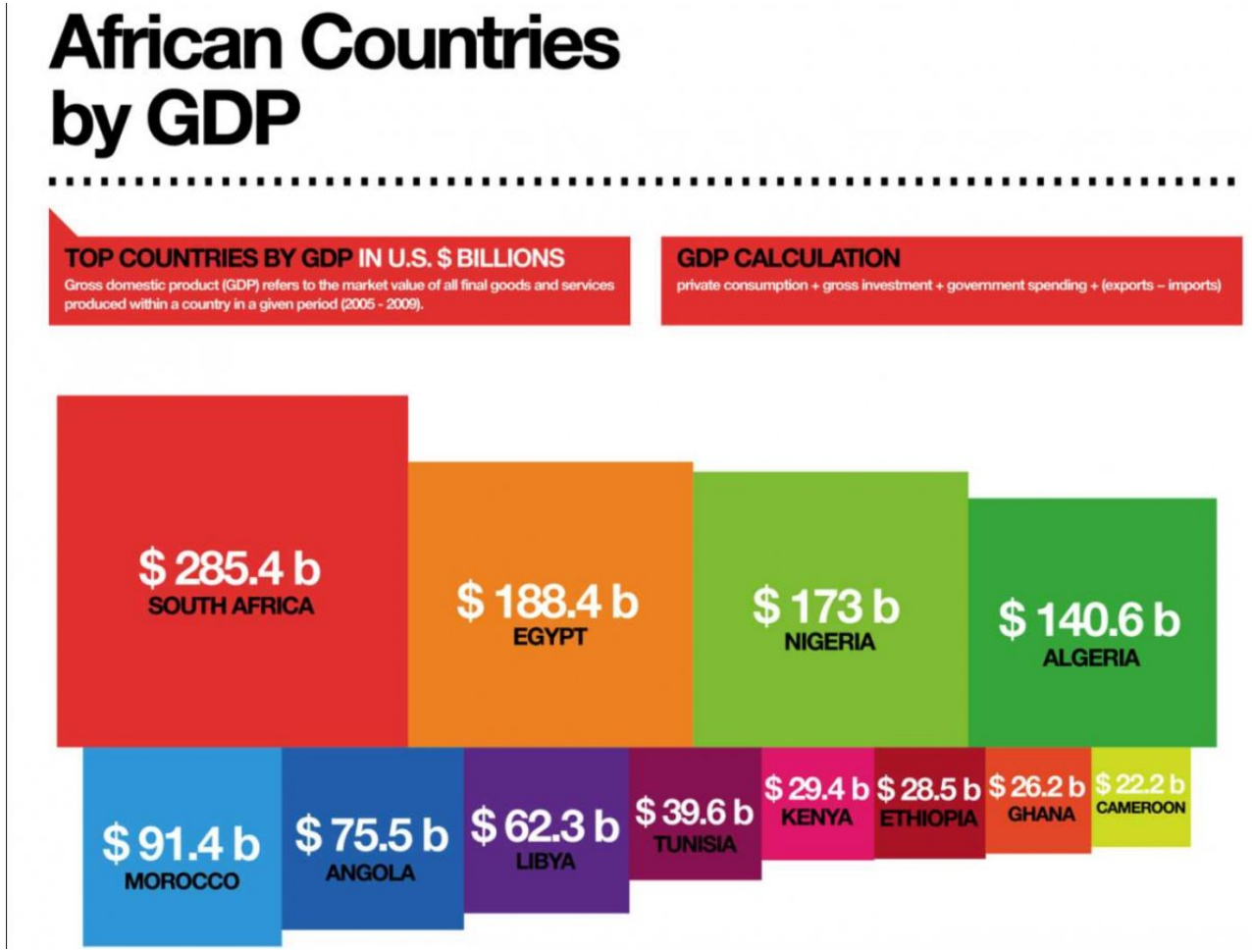
Scale	Conditioning	Perception	Transformation	Context	Smoothing	Philosophy

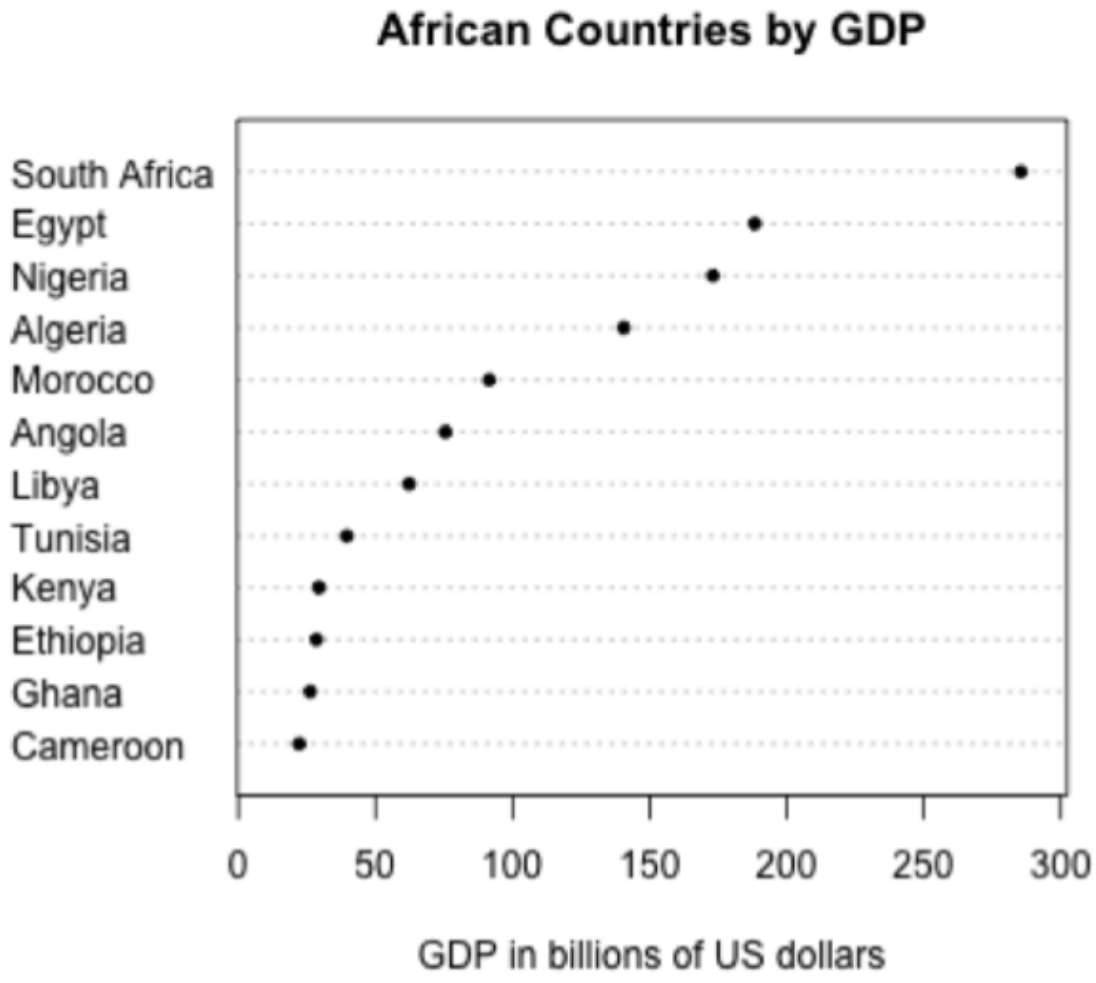
Perception - Length

Bar plot, Pie chart, Dot chart

- Experiments found that angle judgments based on pie charts are less accurate than length judgments from bar charts
- Length is easier to compare than area or volume
- Lengths that fall on a line are easier to compare than lengths on parallel lines, i.e., judgments based on dot charts are easier to make than judgments based on bar plots

Perception


African Countries by GDP
TOP COUNTRIES BY GDP IN U.S. $ BILLIONS
Gross domestic product (GDP) refers to the market value of all final goods and services produced within a country in a given period (2005 - 2009).
GDP CALCULATION
private consumption + gross investment + government spending + (exports – imports)
$ 285.4 b
SOUTH AFRICA
$ 188.4 b
EGYPT
$ 173 b
NIGERIA
$ 140.6 b
ALGERIA
$ 91.4 b
MOROCCO
$ 75.5 b
ANGOLA
$ 62.3 b
LIBYA
$ 39.6 b
TUNISIA
$ 29.4 b
KENYA
$ 28.5 b
ETHIOPIA
$ 26.2 b
GHANA
$ 22.2 b
CAMEROON


African Countries by GDP
South Africa
Egypt
Nigeria
Algeria
Morocco
Angola
Libya
Tunisia
Kenya
Ethiopia
Ghana
Cameroon
0
50
100
150
200
250
300
GDP in billions of US dollars

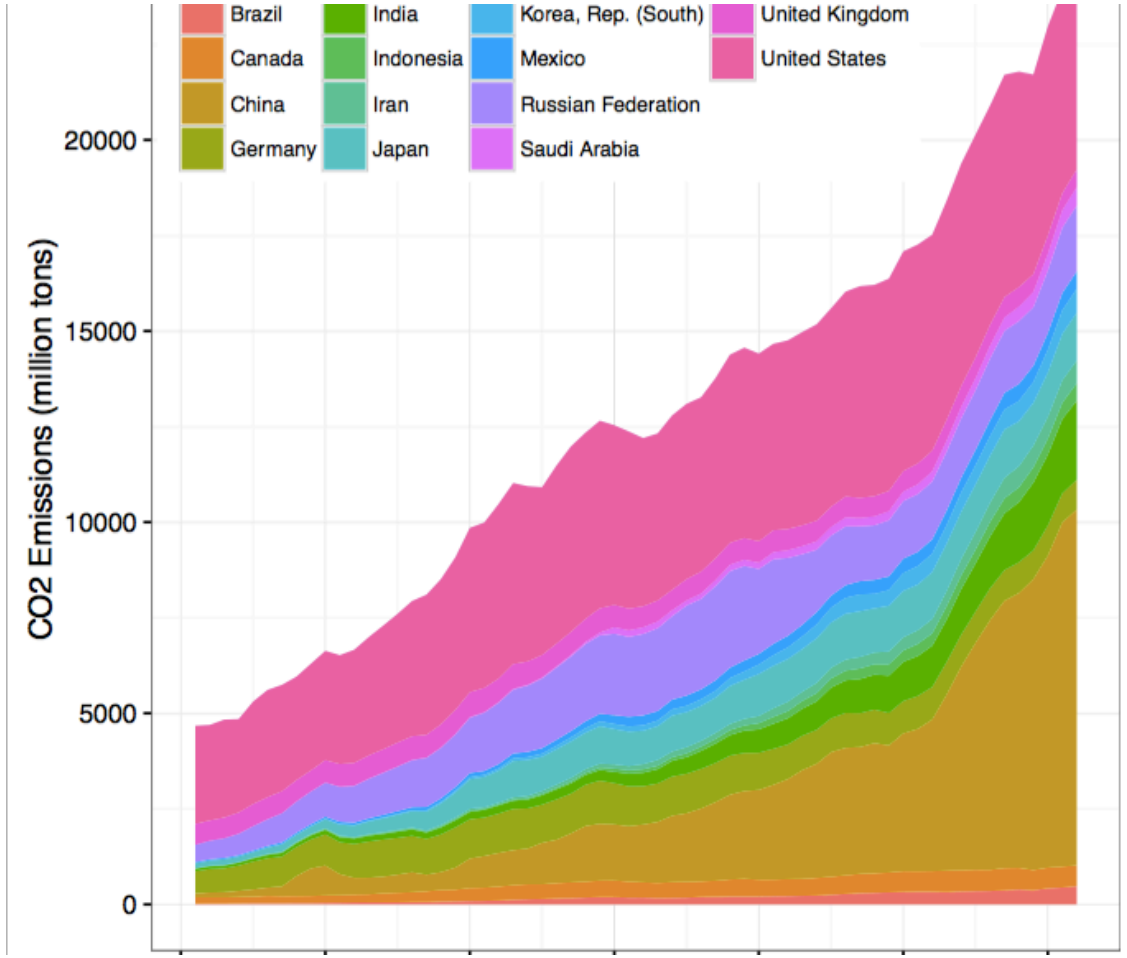
Stacking and Jiggling

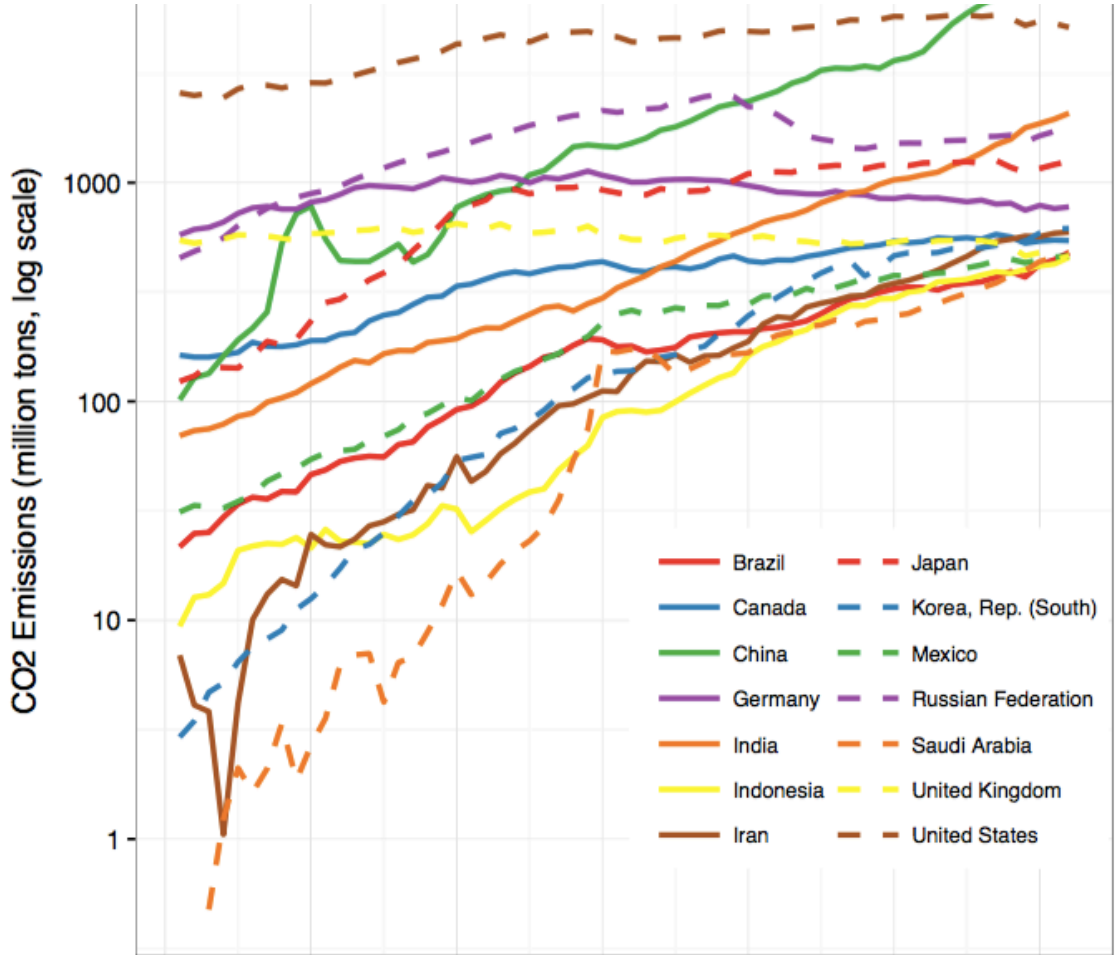
- Stacked bar plots and histograms are difficult to read because the base line moves from one bar to the next
- Line plots where the area between successive lines represent the measurement are very difficult to read because the base line jiggles up and down.

CO2 Emissions from Fuel Consumption

- Data on historical carbon dioxide (CO2) emissions from fuel combustion (http://cait.wri.org)
- Country annual CO2 emissions date back to 1850
- Typical report on trends since 1950 for the 14 countries that emitted the greatest amount of CO2 in 2012
- World Resources Institute (http://www.wri.org/)

CO2 Emissions





Transformations

Why Transform Variables?

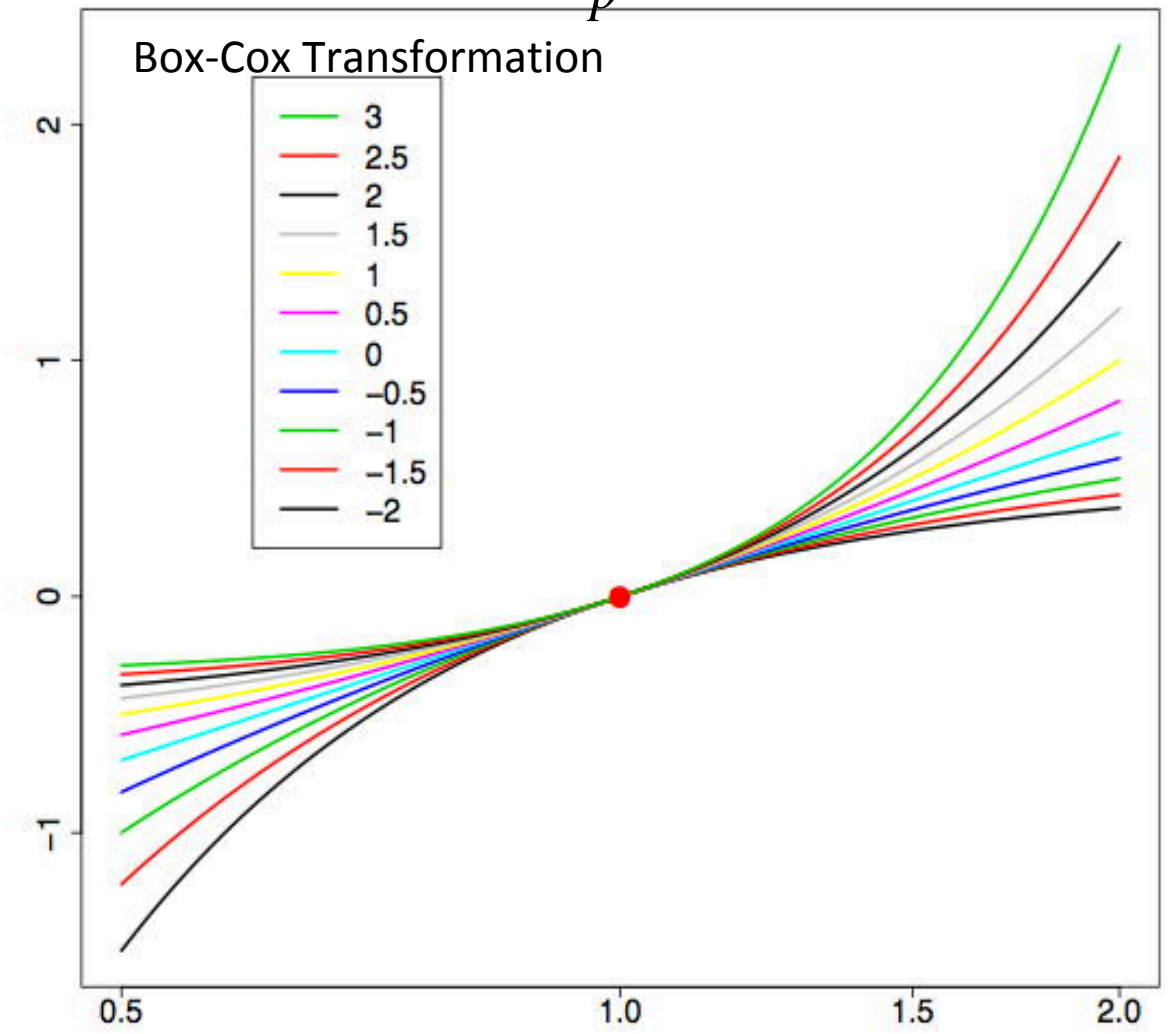
- Reveal distribution of most of the observations (otherwise much of the data is squashed in a small region)
- Numerical summaries of a transformed data are better summaries of a symmetric distribution
- Choose a transformation that's simple and easily interpreted in the context of the problem, e.g., a power of 2, 3, ½, 0 (log), -1

Power Transformation

$$x^{(p)} = \frac{x^p - 1}{p}$$

- Preserve order of values
- Effective when max / min > 5
- Sometimes add a shift before transform
- Ratio of hinges can help select a transformation

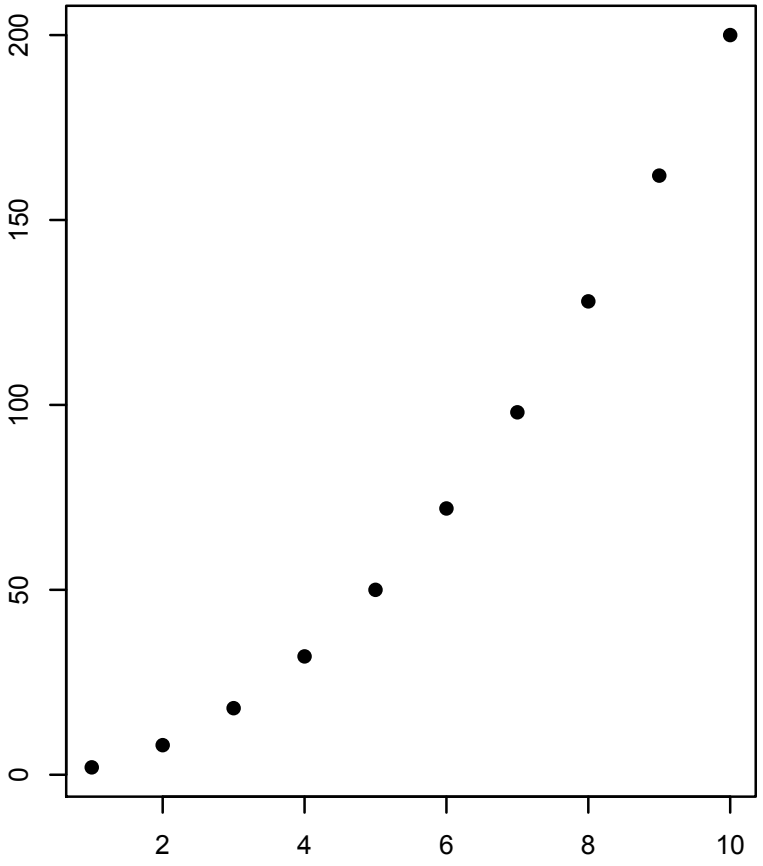
$$\frac{Upper\ Quartile - Median}{Median - Lower\ Quartile} \approx 1$$

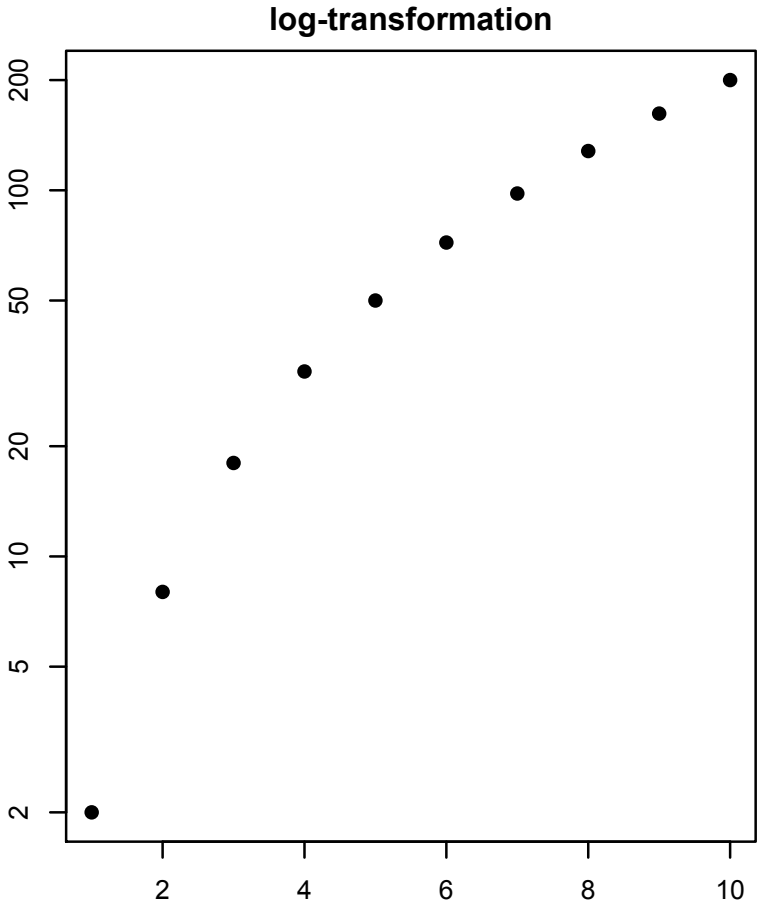


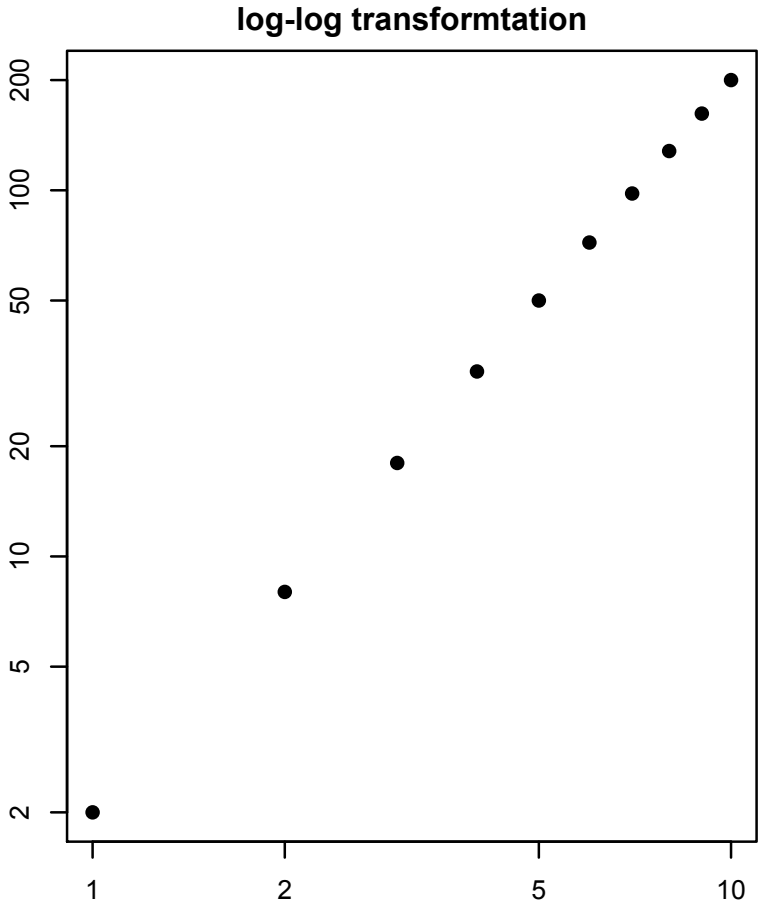
Why Straighten Relationships?

- Easier to uncover the form of the relationship if we can transform it to linear relationship; we see what transformation used to make it linear
- Linear relationships are particularly simple to interpret & fit
- Choose a transformation that's simple and easily interpreted in the context of the problem, e.g., a power of 2, 3, ½, 0 (log), -1

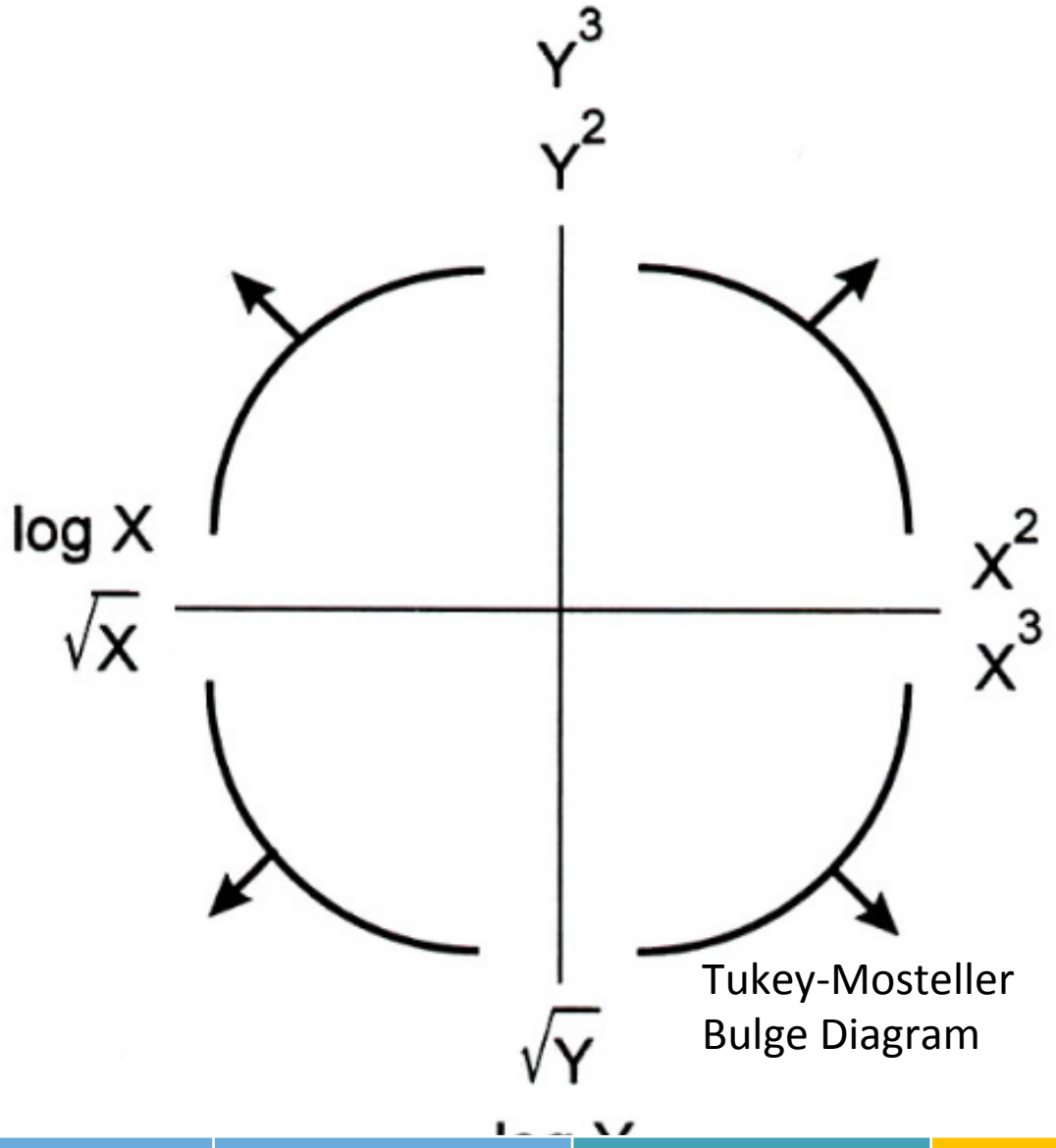
Straighten Relationships with Transformations



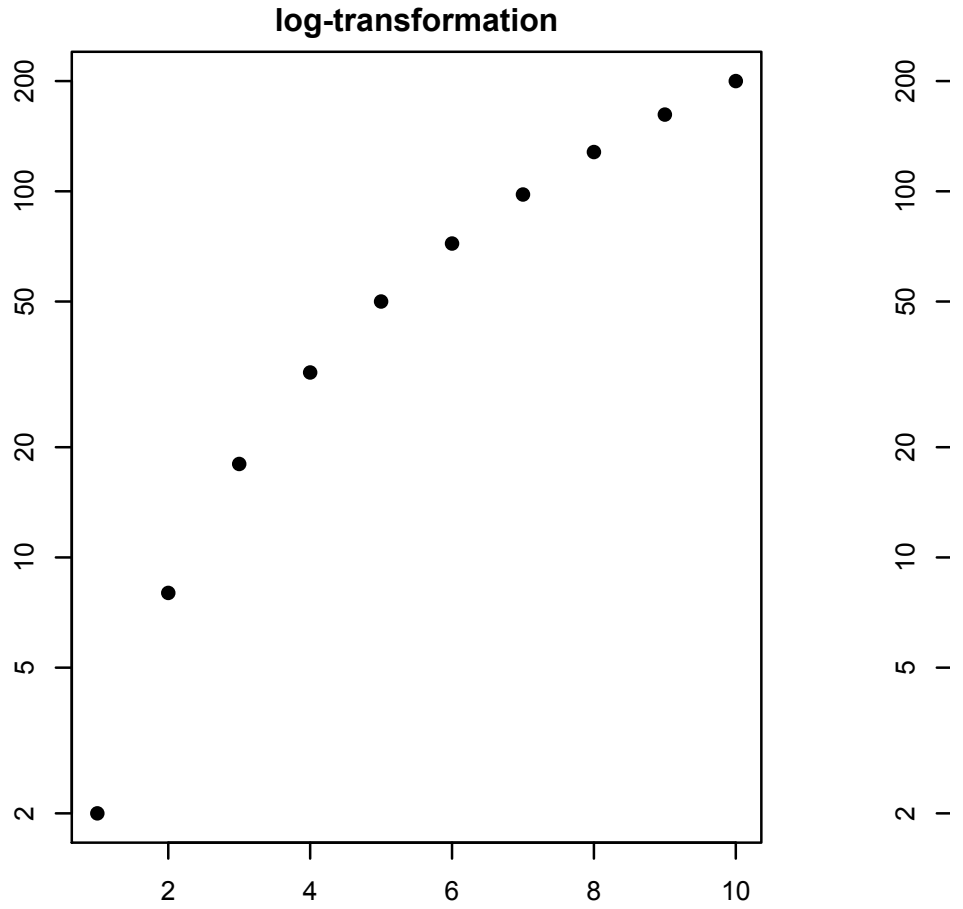




Straighten Relationships with Transformations



Tukey-Mosteller Bulge Diagram

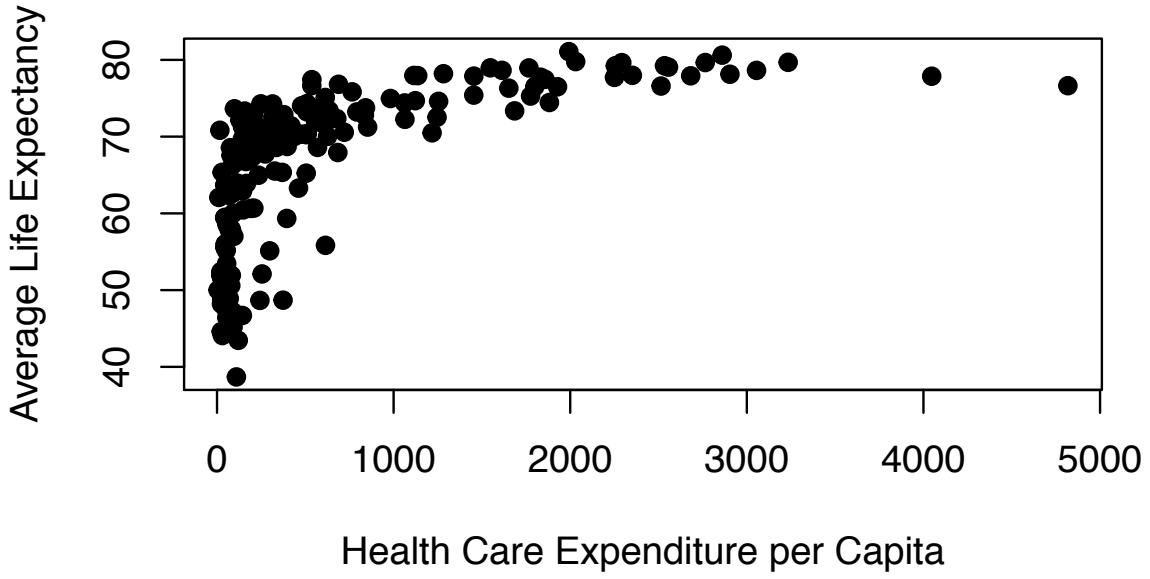


World Bank Country Statistics

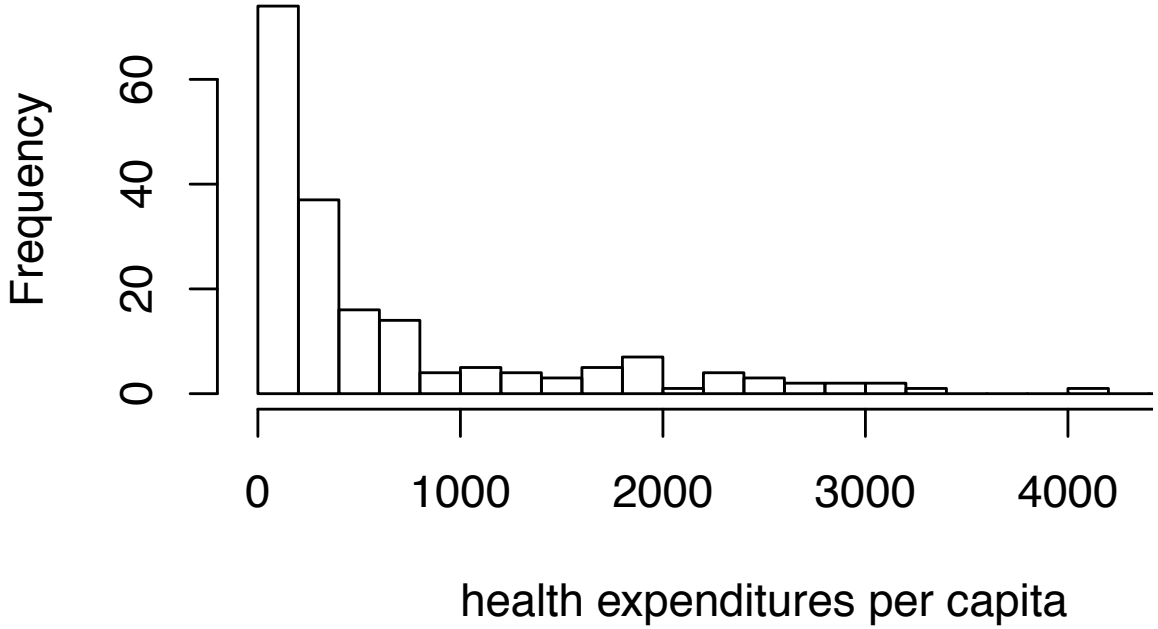
- World Bank provides financial and technical assistance to developing countries.
- In 2010, the World Bank launched an Open Data Website that provides access to data from their reports on topics such as GDP, education, health, and the environment.
- We are interested in the relationship between life expectancy and health expenditures
- These variables are measured at the country level

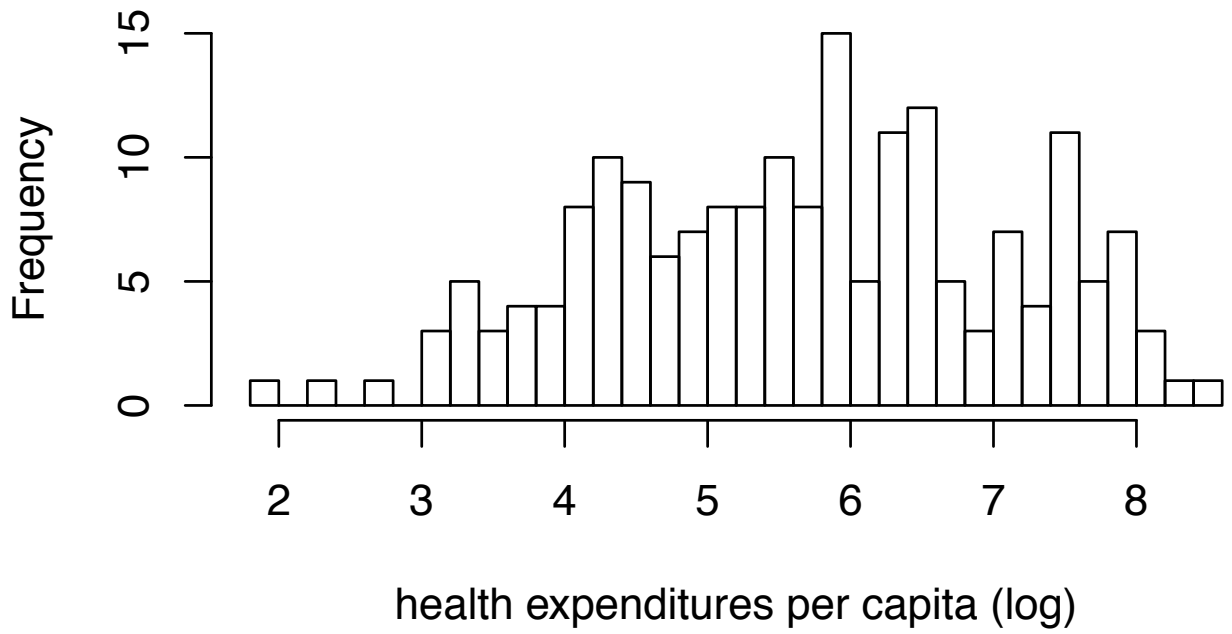
Healthcare Costs & Life Expectancy

- What does the bulge diagram tell us?

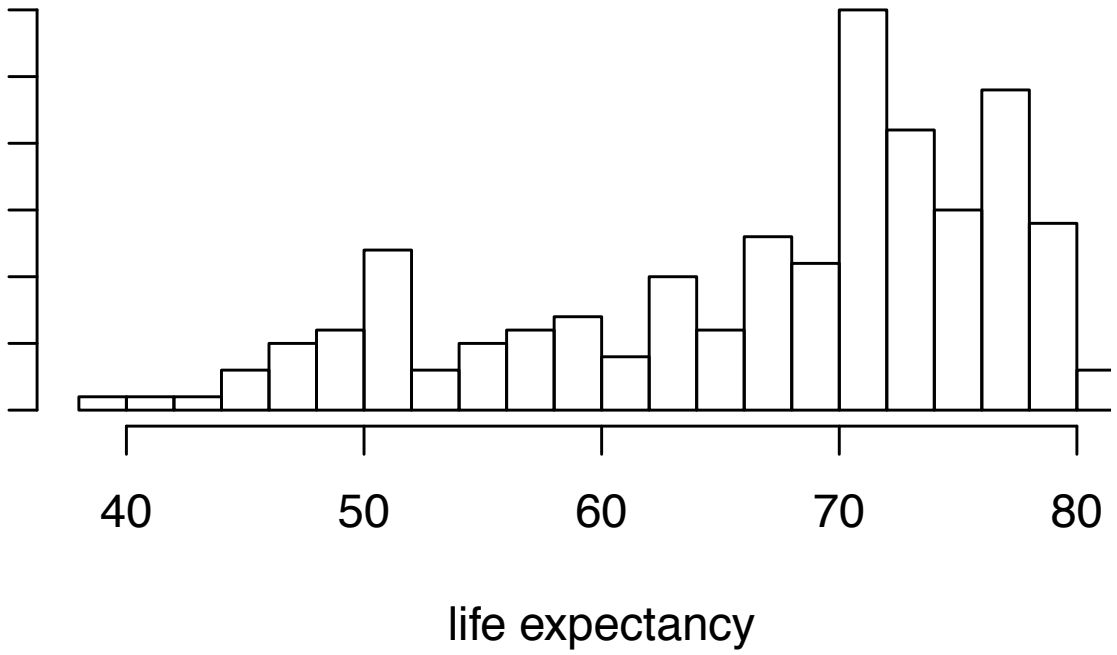


Healthcare Costs



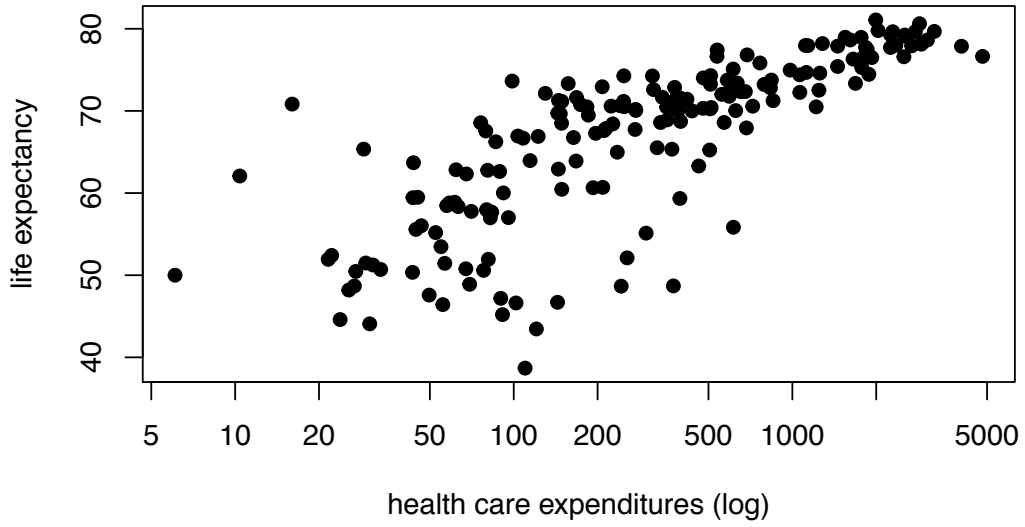


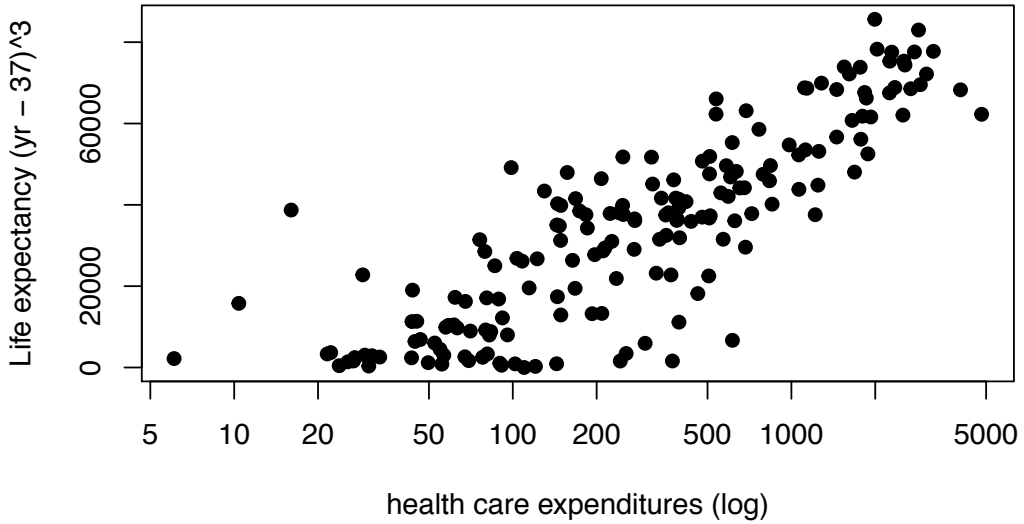
Life Expectancy



- Unusual feature – left skew. Upper bound on life expectancy
- What transformation?
  - Range is factor of 2 so shift first
  - Pull up the low end with a square or cube transformation

Healthcare Costs & Life Expectancy





Issues remaining:

3 unusual countries

Complexity of cube model

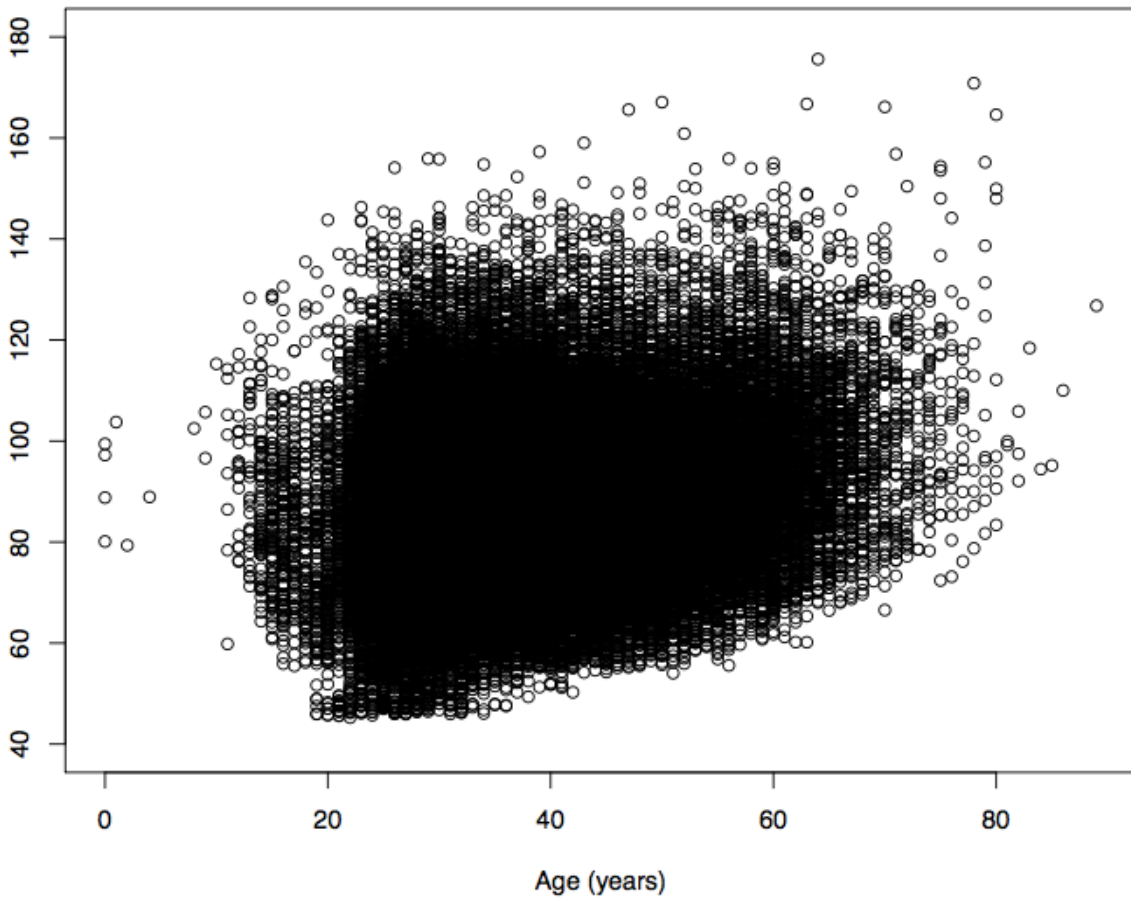
Add Context

Add Context

- Label axes, including units
- Add Reference lines and markers for important values
- Label points of unusual/interesting observations
- Include captions that describe data, how plotted, and describe important features

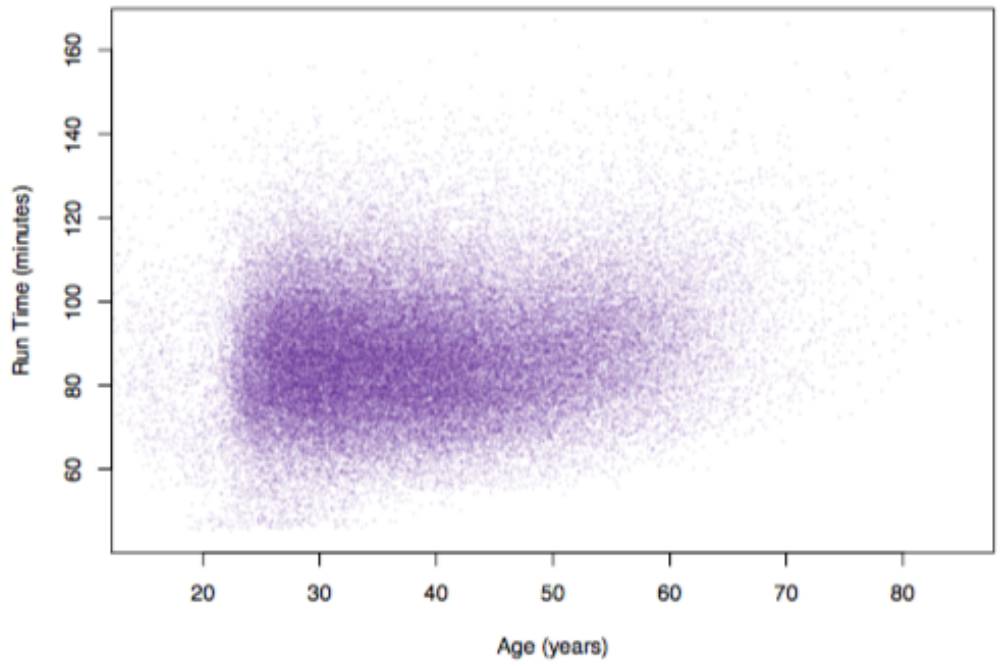
Large n (records)

Cherry Blossom Run



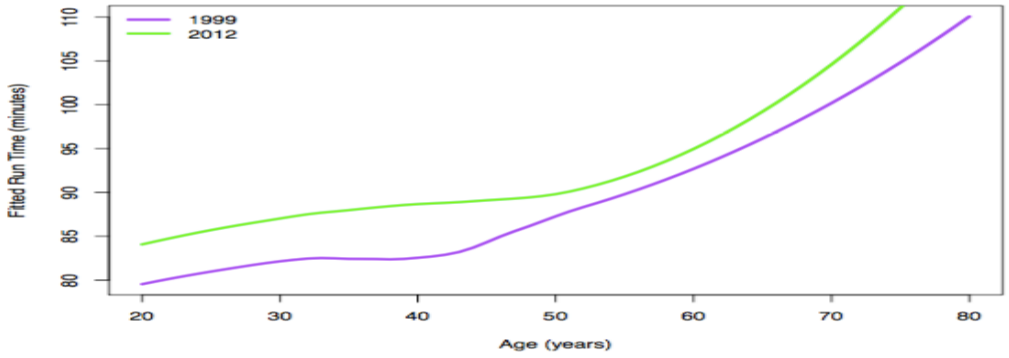
- 3-dimensional histogram is needed, but hard to come by
- Use heat map or hexbin plot or transparency
- Add smooth curve that takes local averages to see the conditional center, i.e., average y in a neighborhood of x

Cherry Blossom Run



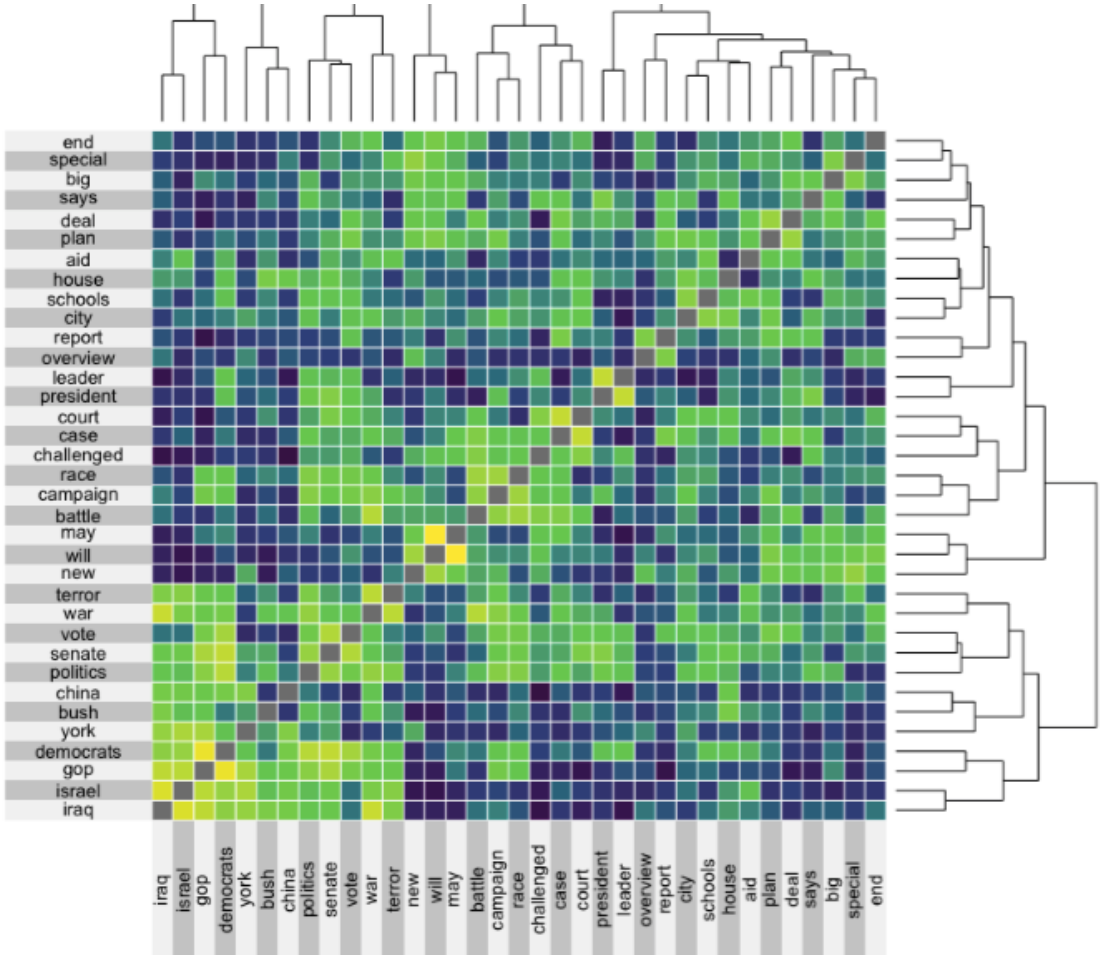
Slightly transparent points

- Local Smoothing helps us see the center
- Control for year – race popularity
- Observational data – snapshot in time
- Not Longitudinal – follow same people in time



Large p (variables)

Heat map



- Documents – records
- Words counts – variables
- Hierarchical clustering groups documents that have similar distributions of words
- Heat Map - Color is used to denote closeness

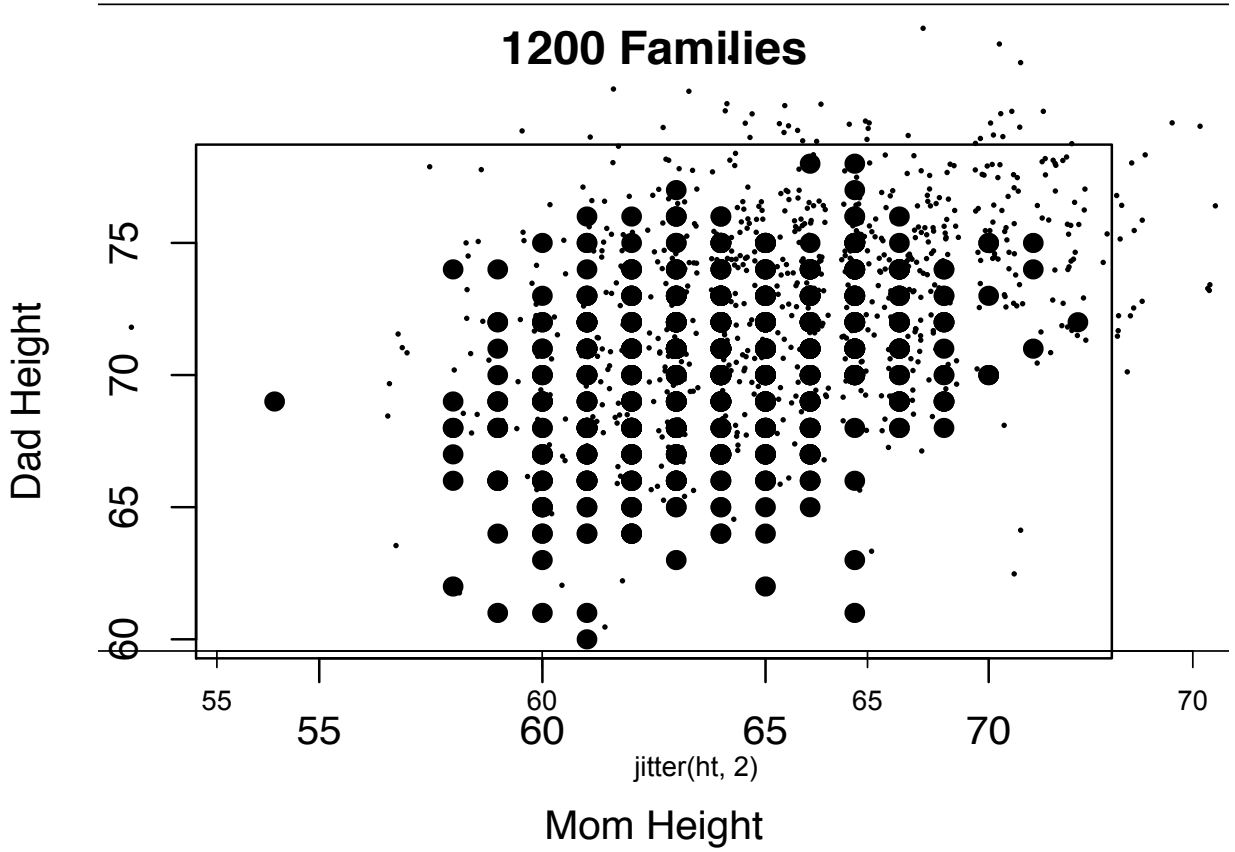
Philosophy

Reveal the Data

- Choose scale appropriately
- Avoid having other graph elements interfere with data
- Use visually prominent symbols
- Eliminate superfluous material, aka chart junk
- Avoid over-plotting

Avoid over-plotting

Why are there so few data points?



Jitter: Add random noise so the values aren't plotted on top of each other

Shrink the plotting symbol so they don't plot on top of each other

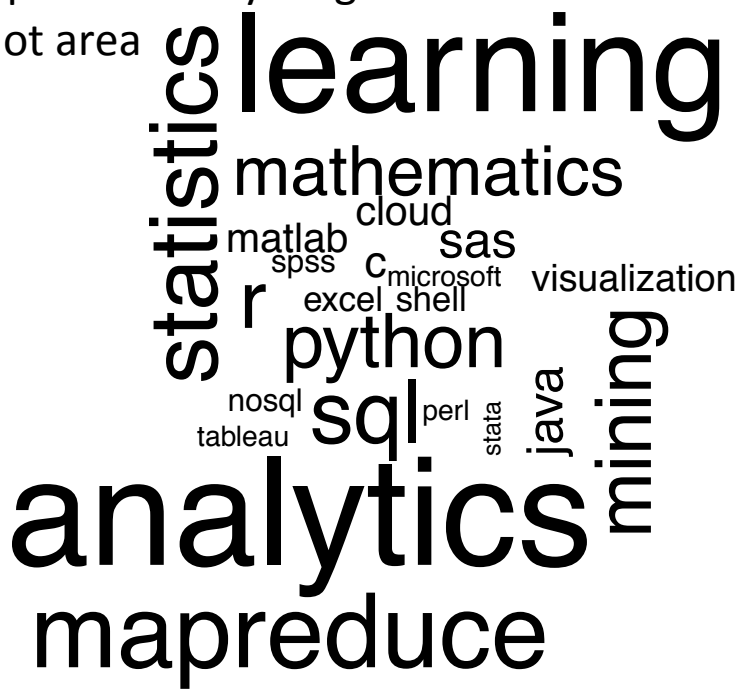
Facilitate Comparisons

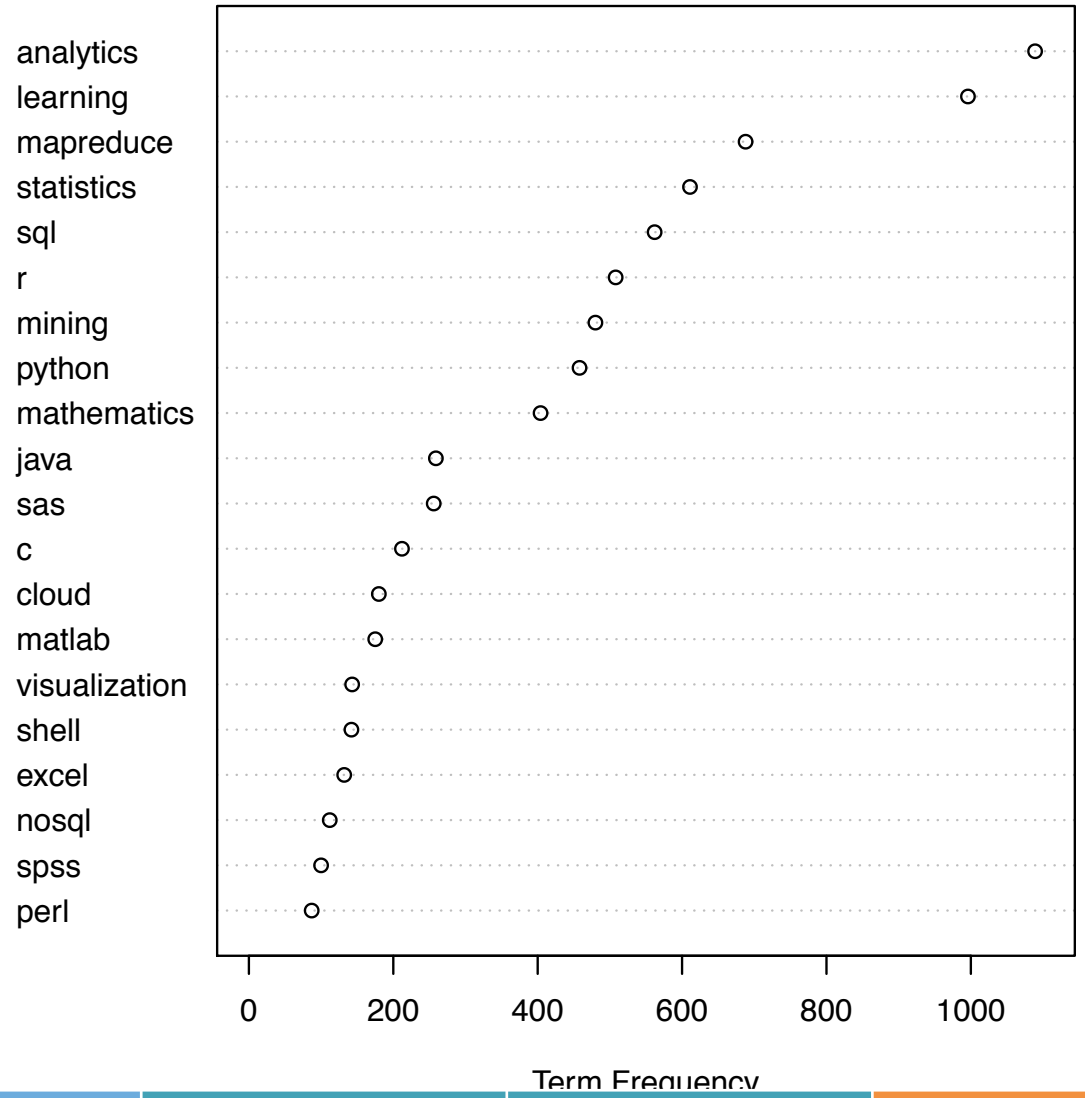
- Put Juxtaposed plots on same scale
- Make it easy to distinguish elements of *superposed* plots, e.g. color, line type
- Avoid Stacking and Jiggling the baseline
- Avoid angles, extra dimensions (e.g., areas rather than lines)
- Don’t break the visual metaphor, i.e., if use rectangles, then area should correspond to value

Comparison: area vs Length

Order of words/counts is random – makes it difficult to compare

Broken Visual metaphor –
count is represented by height
of word, not area





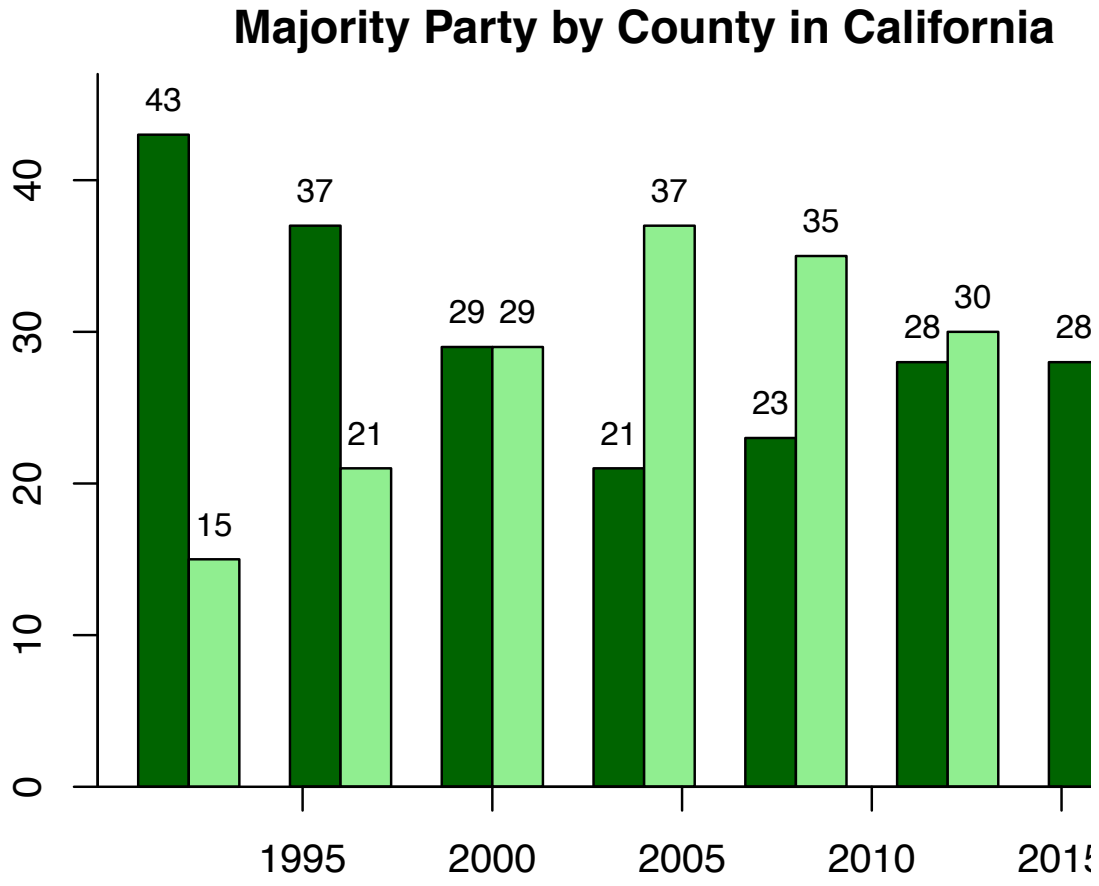
Make a plot information rich

- Describe what you see in the Caption
- Add context with Reference Markers (lines and points) including text
- Add Legends and Labels
- Use color and plotting symbols to add more information
- Plot the same thing more than once in different ways/scales
- Reduce clutter

Captions

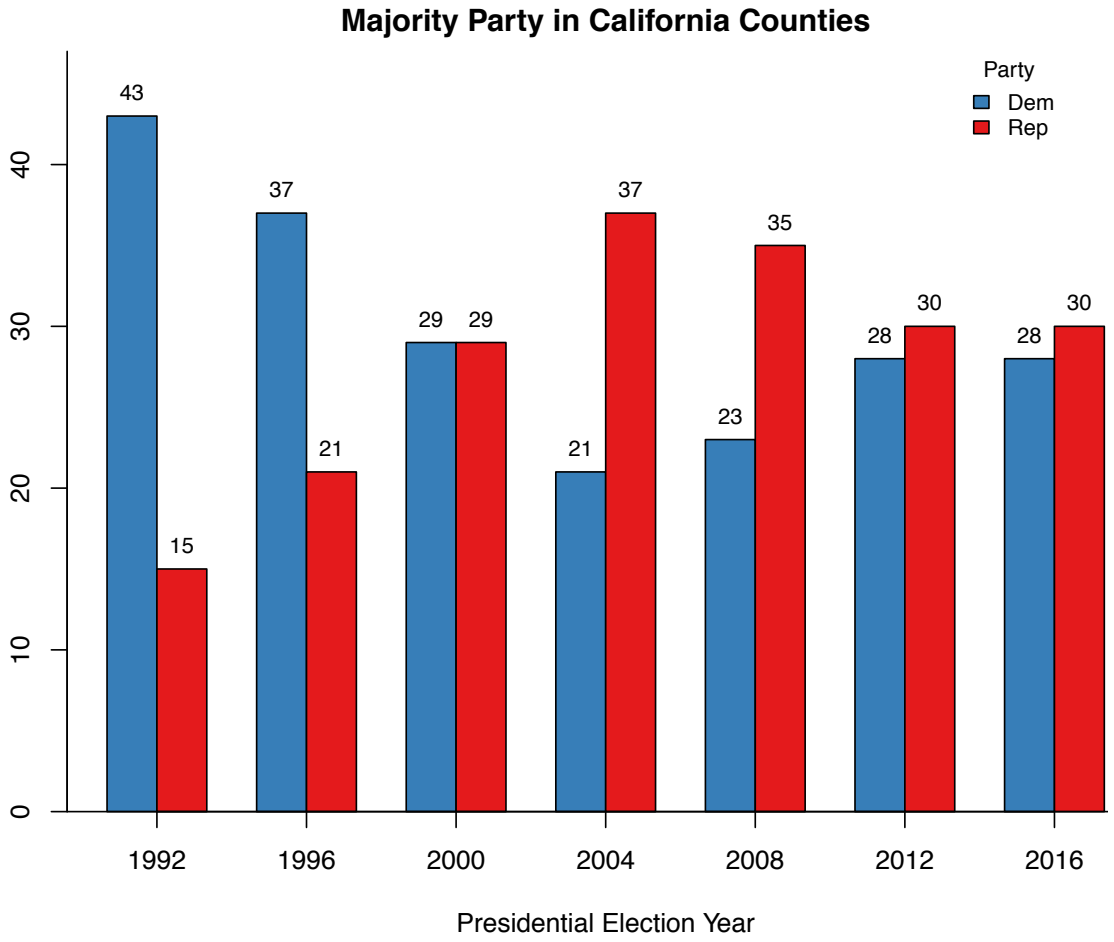
- Captions should be comprehensive
- Self-contained
- Captions should:
  - Describe what has been graphed
  - Draw attention to important features
  - Describe conclusions drawn from graph

Iterate – Example Voter Registration



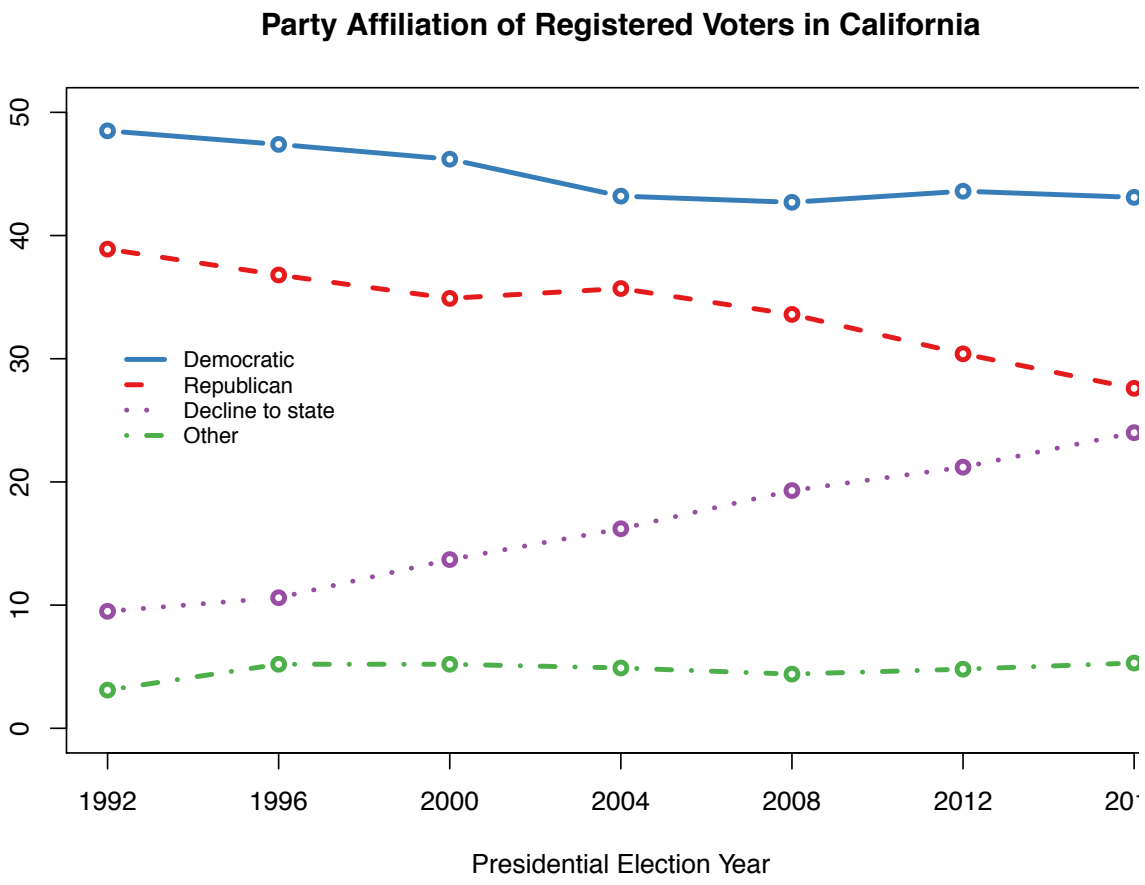
- Location of tick marks under bars
- Color of bars – indicate party
- Title confusing
- Y-axis label confusing
- X-axis label missing

Voter Registration Trends



- Check data for understanding of how plot is made
- Observation? People vote, not counties
- Lurking variable? County size - small counties tend to be rural and conservative

Voter Registration Trends



- What is the message?
- Can we improve it?
- Collect better/more data
  - Decline to state and other parties are missing
  - Voter registration totals may be more useful

Good Plot Making Practice

- Put major conclusions in graphical form
- Provide reference information
- Proof read for clarity and consistency
- Graphing is an iterative process
- Multiplicity is OK, i.e., two plots of the same variable may provide different messages
- Make plots data rich

Univariate Displays

Type	Plot
Numeric –	few observations Histogram, Density curve Box plot, Violin plot Normal quantile plot Few Observations - Rug plot, Dot plot Caution if discrete: density curves and box plots may be misleading
Categorical – Counts of categories	Dot chart Bar chart Pie chart (avoid!) Caution if ordinal –order of bars, dots, etc. should reflect category order

Bivariate Displays

	Numeric	Categorical
Numeric	Scatter plot Smooth scatter Smooth lines and curves	Multiple histograms, density curves, Avoid jiggling!
Categorical		Side-by-side bar plot Overlaid Lines plot Side-by-side dot chart Mosaic plot Avoid stacking!

Chart Design: Simplifying

- Example from Tim Bray

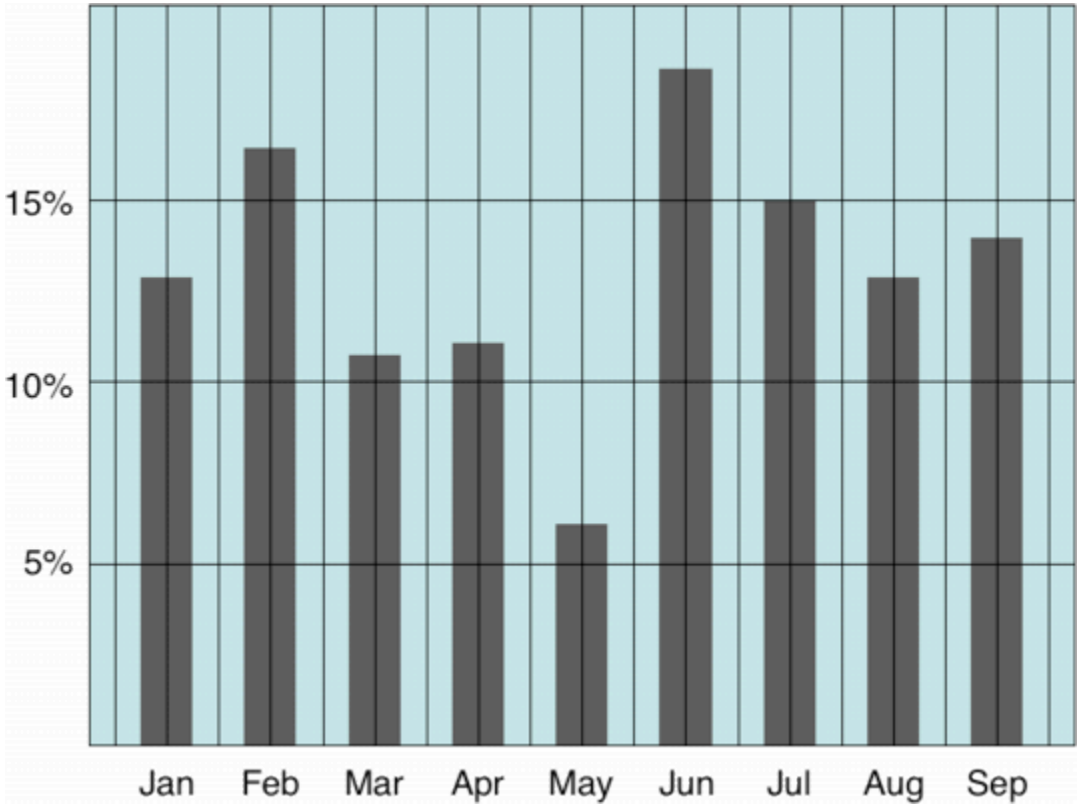


Chart Design: Simplifying

- Example from Tim Bray

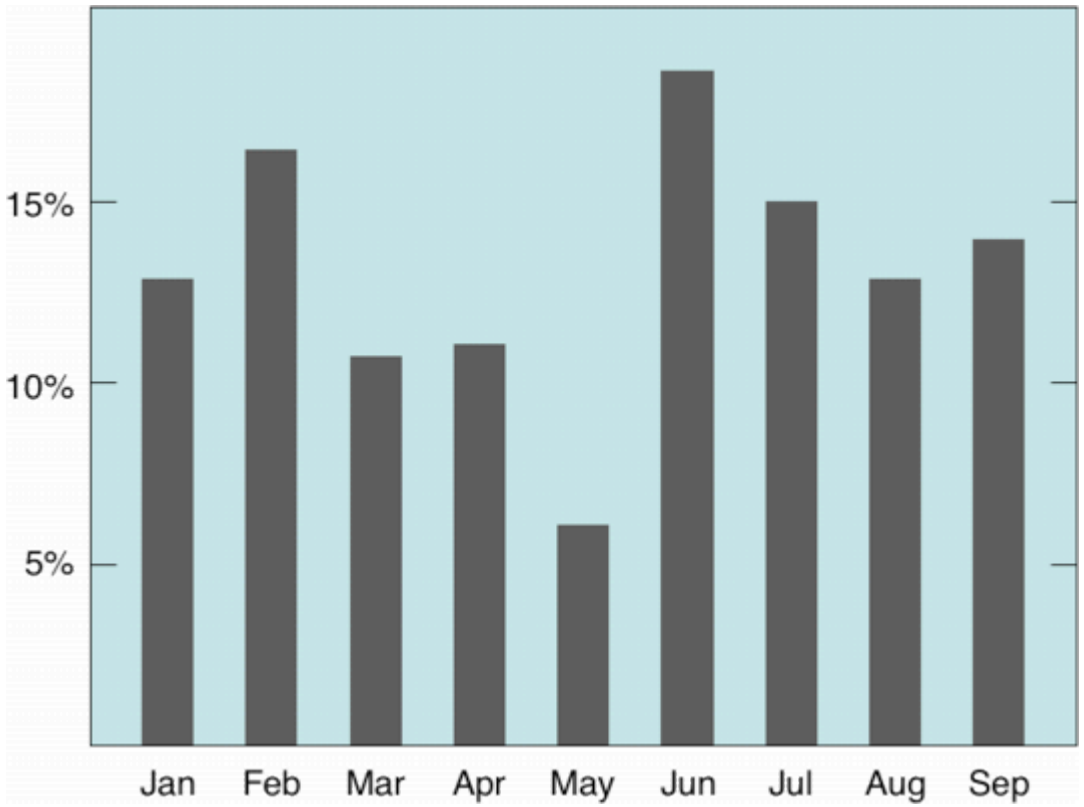


Chart Design: Simplifying

- Example from Tim Bray

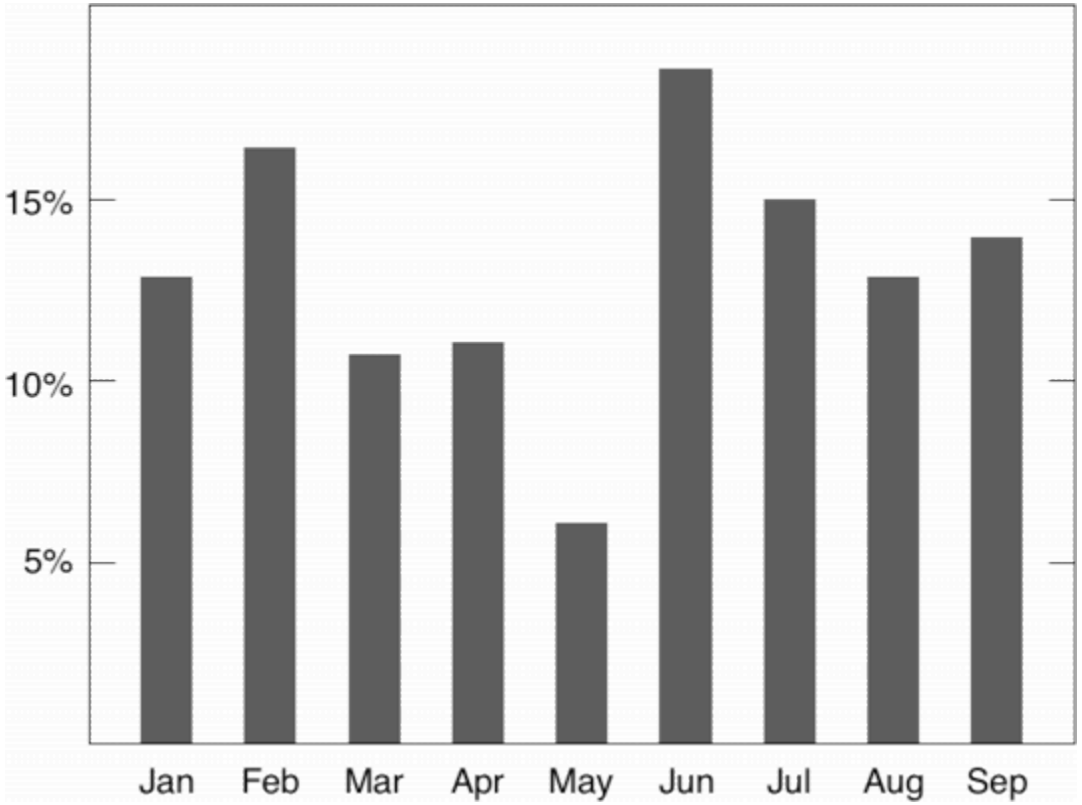


Chart Design: Simplifying

- Example from Tim Bray

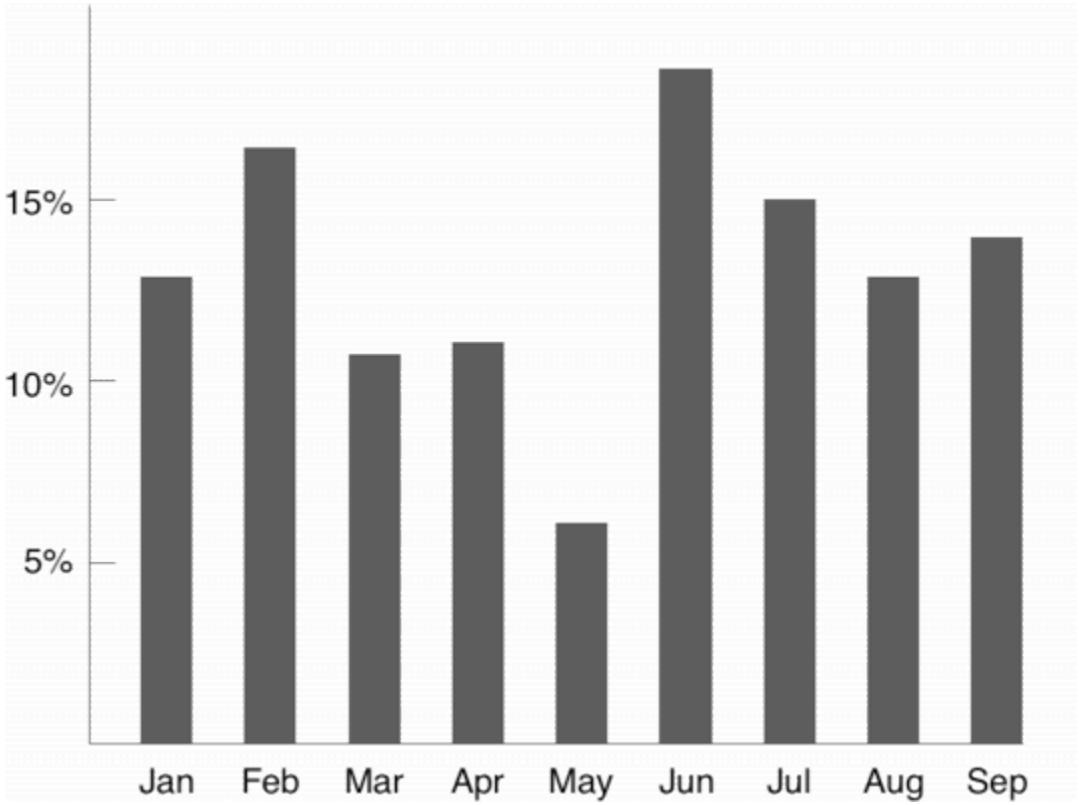


Chart Design: Simplifying

- Example from Tim Bray

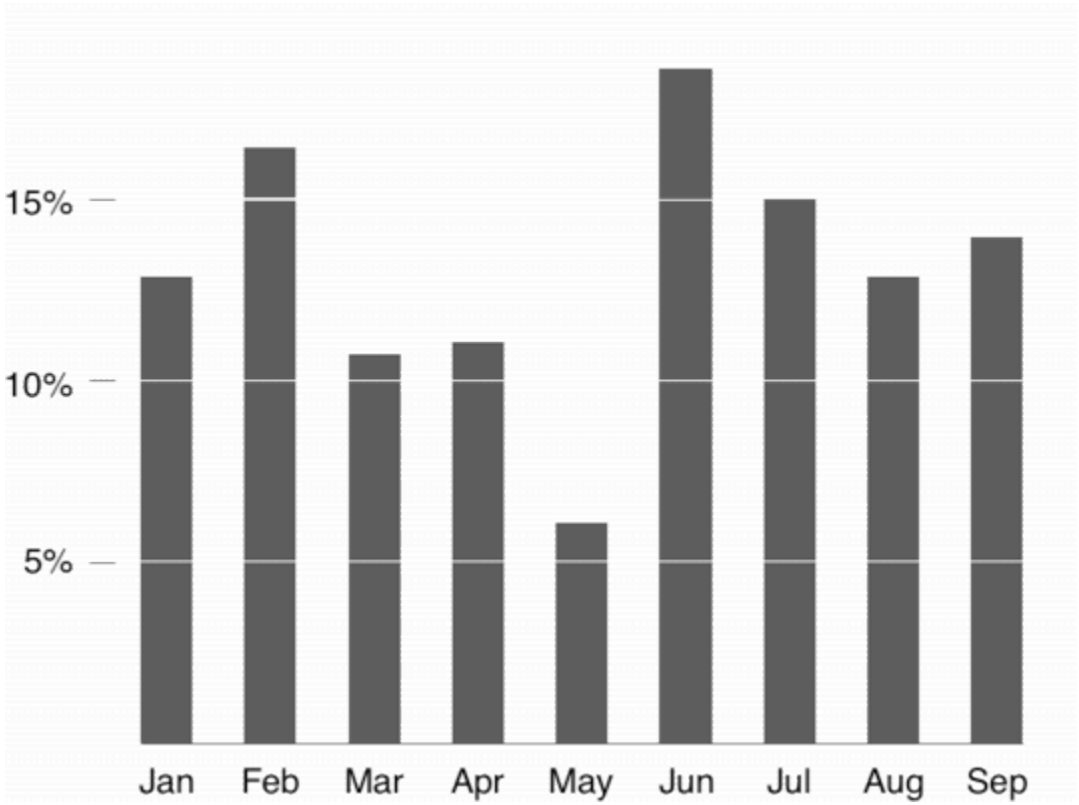
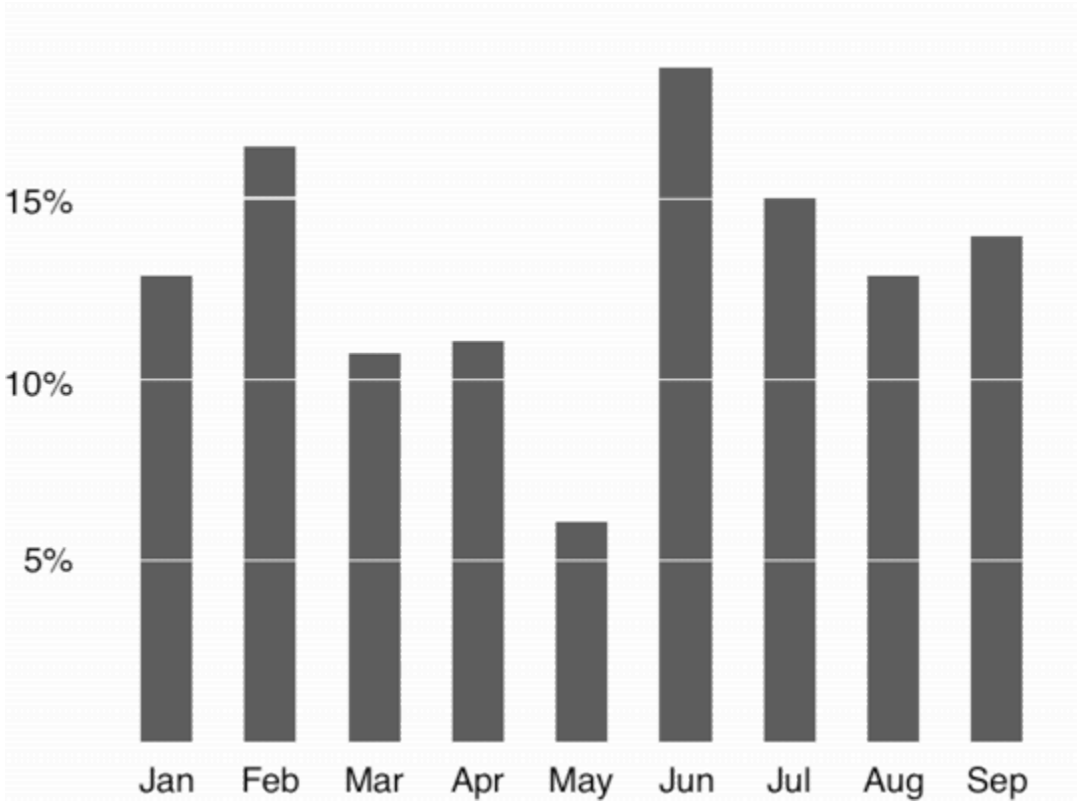


Chart Design: Simplifying

- Example from Tim Bray



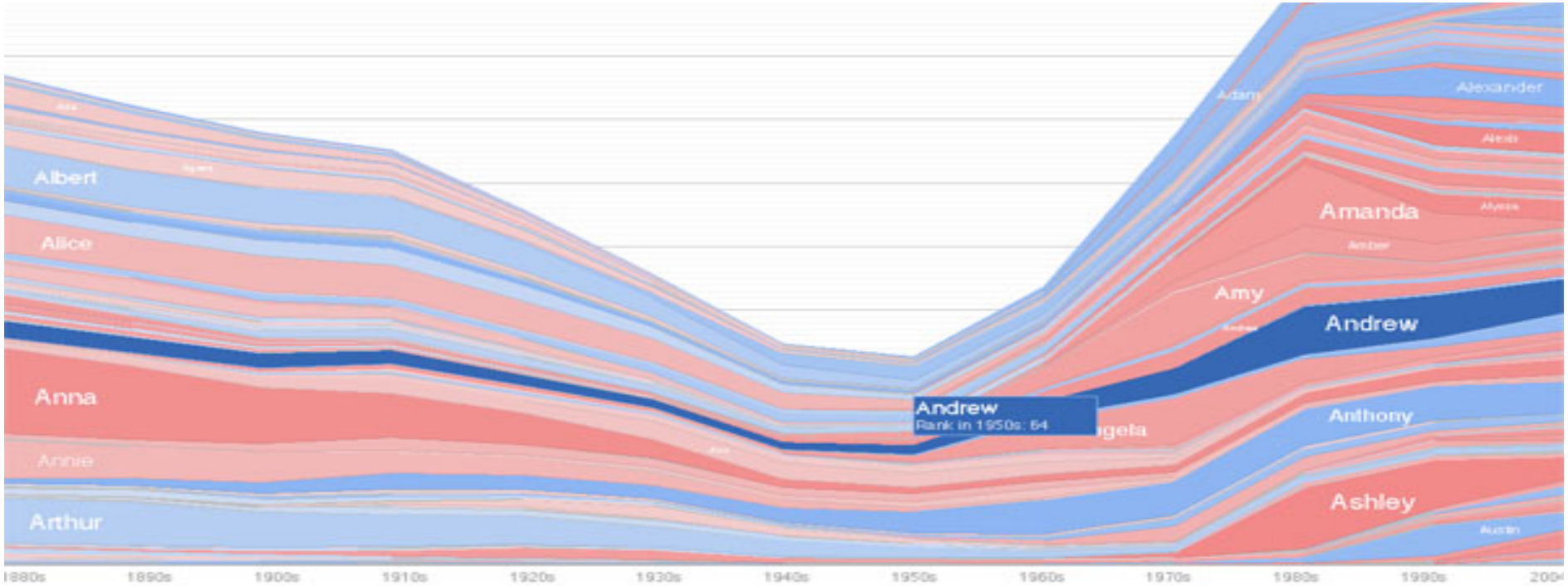
Principle 1: Simplify

- Tables and charts
  - Reduce chartjunk/tablejunk; increase data-ink ratio
  - Lessons from perception: Limit the number of objects displayed at once

- Beware:
  - Gratuitous 3D
  - Shadows
  - Gratuitous animation
- How do you tell if a feature is gratuitous?
  Ask whether using it reveals more information.

Interactive Chart Design: Simplifying

- With interactive charts you can keep things very simple by **hiding** and **dynamically revealing** important structure.
- On an interactive chart, you reveal the information most useful for **navigating** the chart.



Principle 2: Understand Magnitudes

Which is brighter?

Principle 2: Understand Magnitudes

(128, 128, 128)

(144, 144, 144)

Which is brighter?

Just Noticeable Difference

- JND (Weber's Law)

$$\Delta S = k \frac{\Delta I}{I}$$

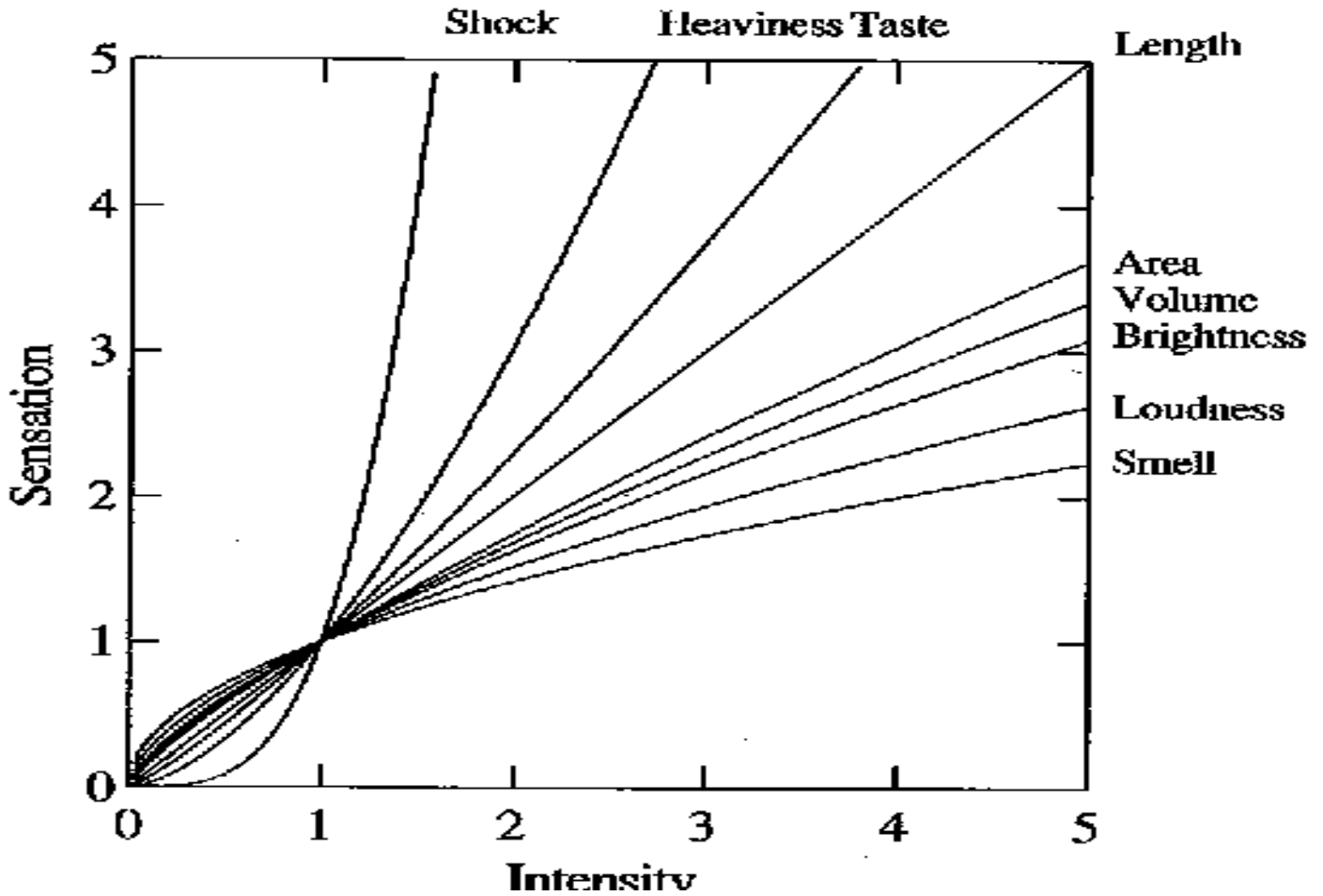
- Ratios more important than magnitude
- Most continuous variations in stimuli are perceived in discrete steps

Steven's Power law

$$S = I^p$$

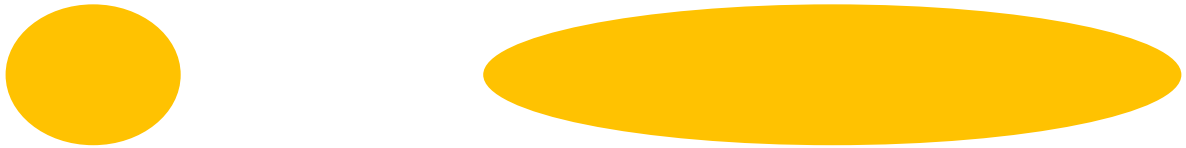
S = sensation
I = intensity

p < 1 : underestimate
p > 1 : overestimate

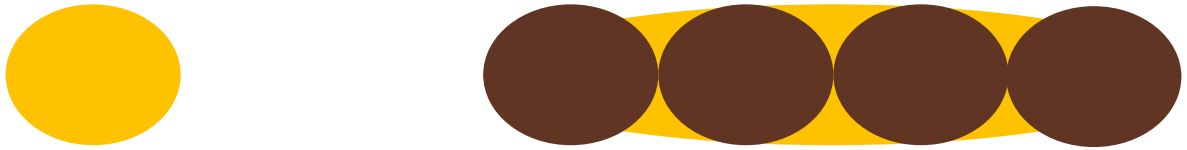


[graph from Wilkinson 99, based on Stevens 61]

[alternate graph : http://www.undergrad.ahs.uwaterloo.ca/~wchedder/stevenspowerlaw.htm]



Compare area of circles

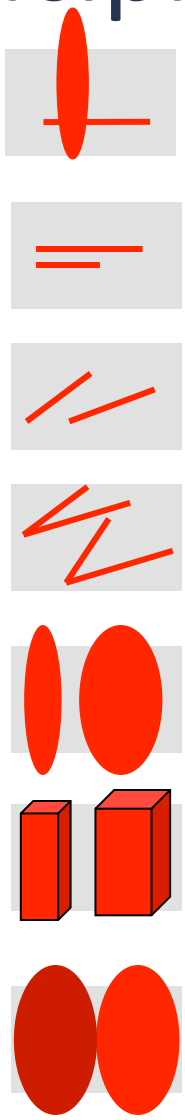


Compare area of circles

Principle 2: Understand Magnitudes

Most accurate

Least accurate



Position (common) scale

Position (non-aligned) scale

Length

Slope

Angle

Area

Volume

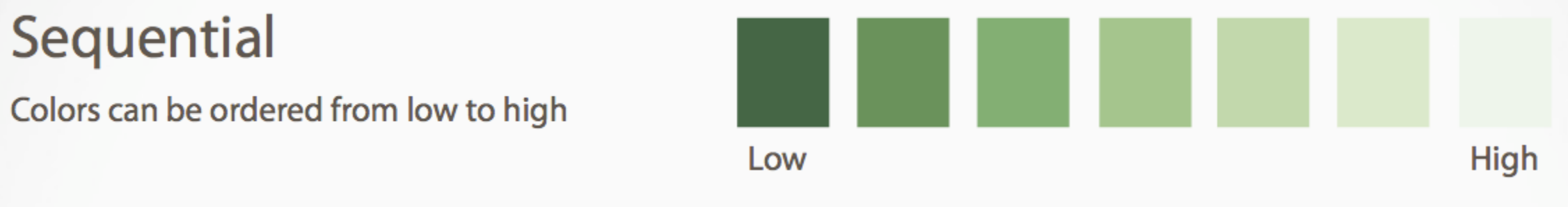
Color hue-saturation-density

Principle 3: Use Color

- Color
  - Choose colors based on the information you want to convey
    - Sequential
    - Diverging
    - Categorical
  - Use online resources to discover and record your color schemes
    - Color Brewer
    - Kuler
    - Colour Lovers
  - Where possible, use your organization's palette

Principle 3: Use Color

- Color

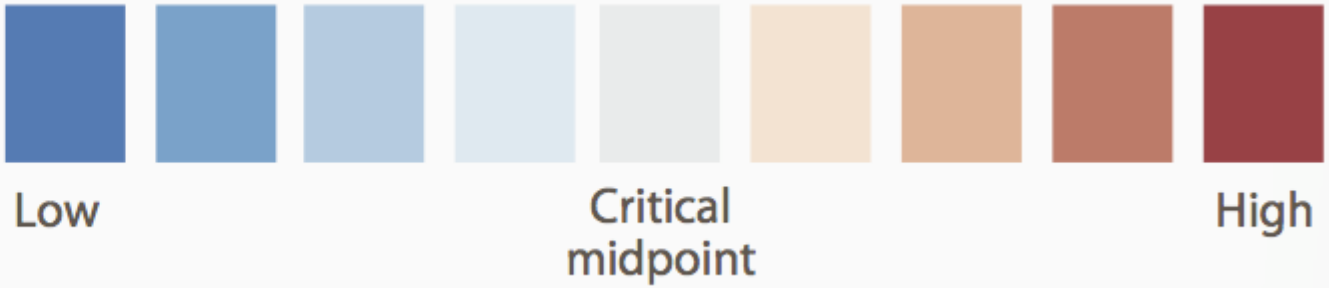


Principle 3: Use Color

- Color

Diverging

Two sequential schemes extended out from a critical midpoint value

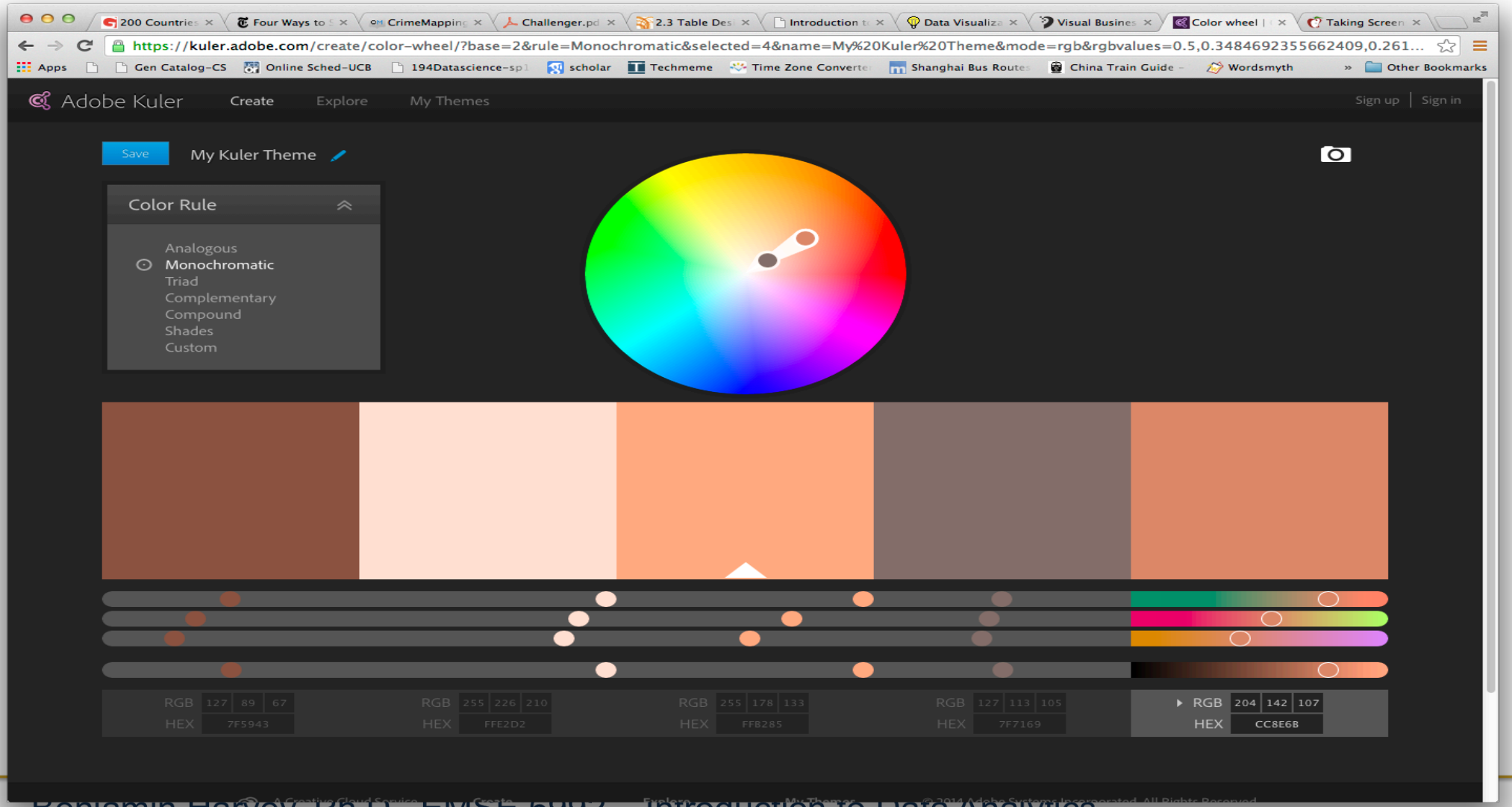


Principle 3: Use Color

- Color

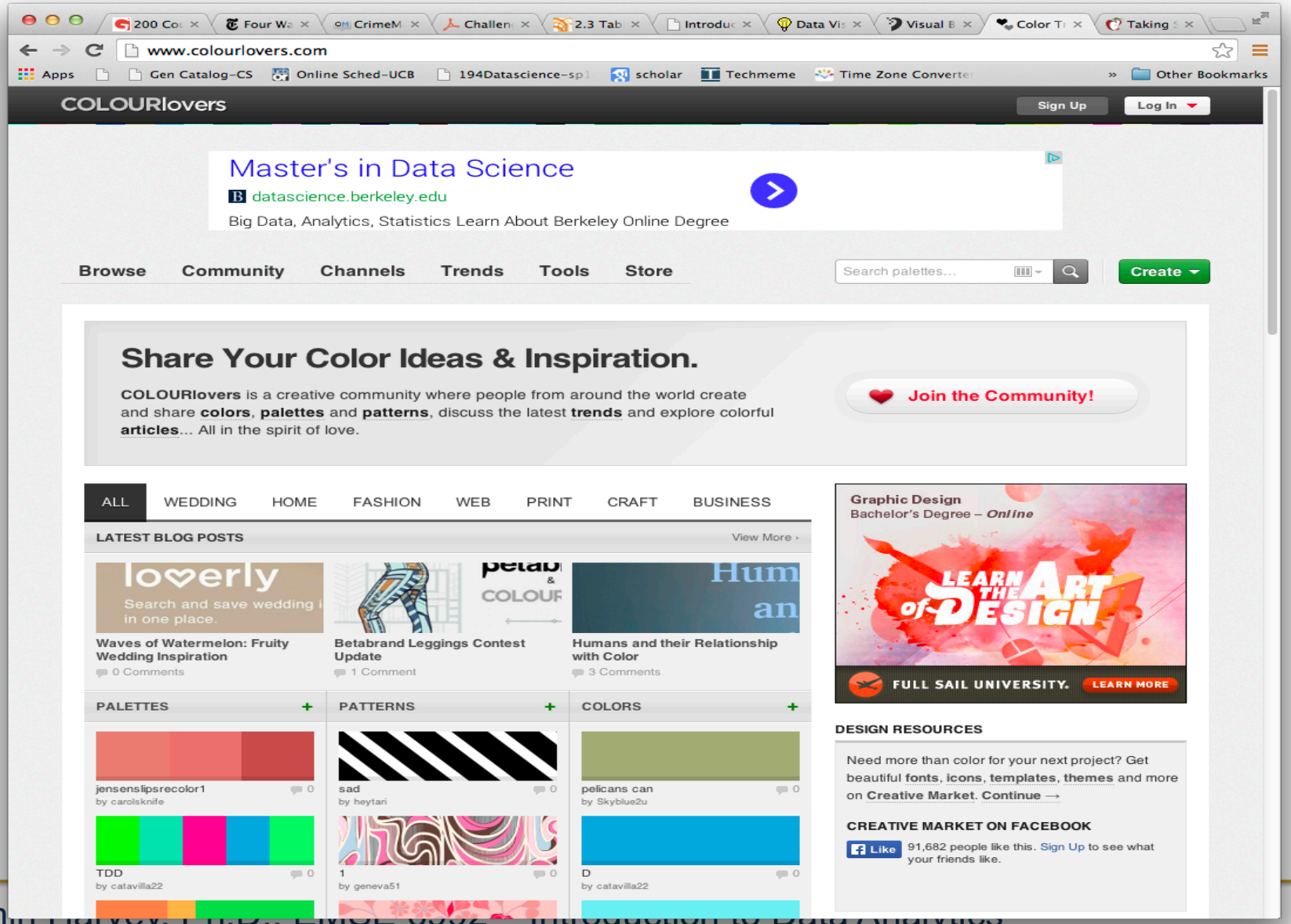
Principle 3: Use Color

- Color



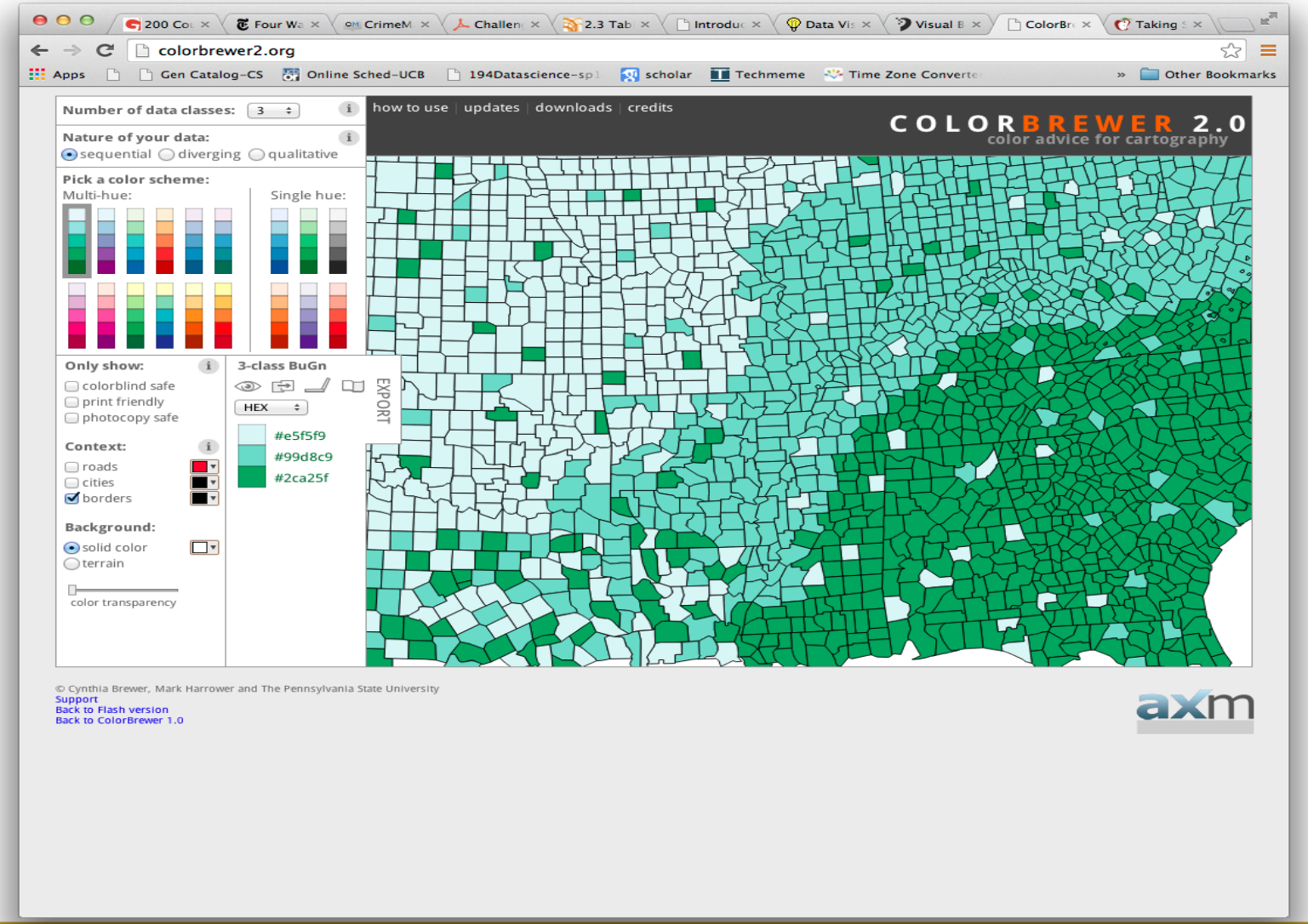
Principle 3: Use Color

- Color



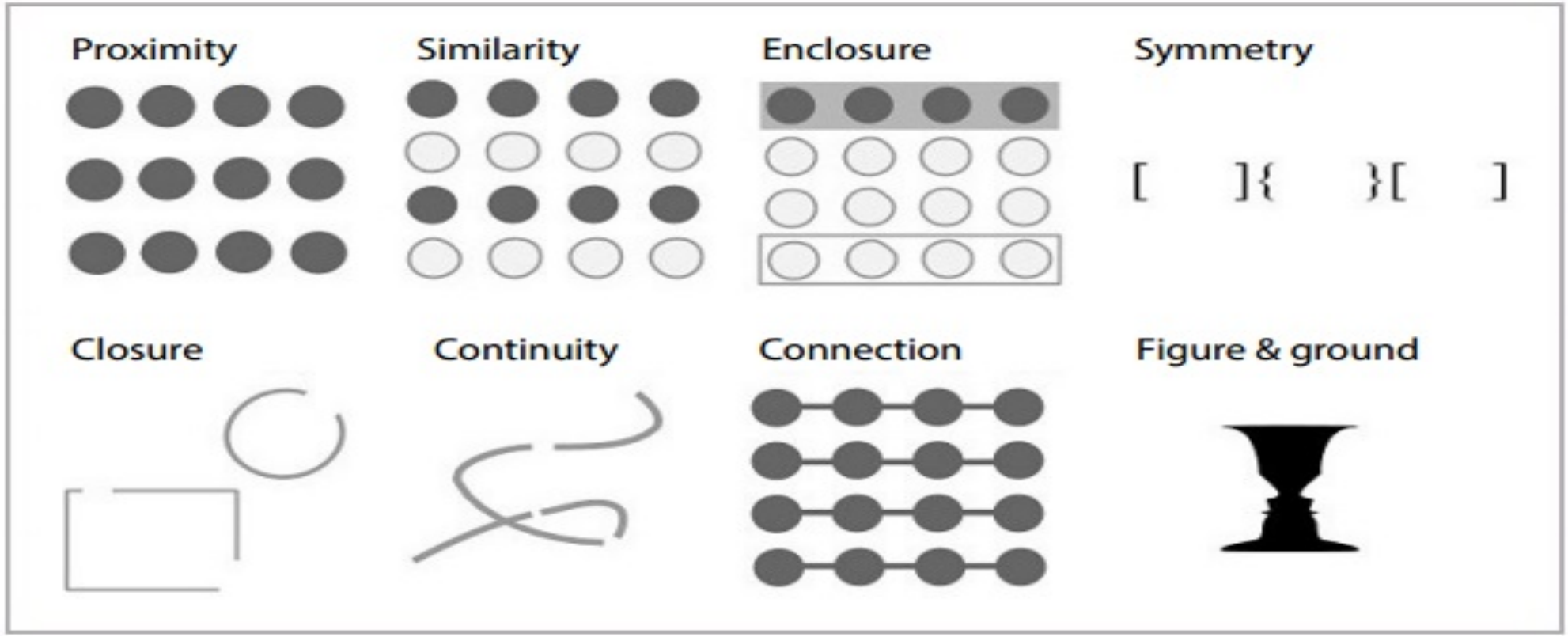
Principle 3: Use Color

- Color



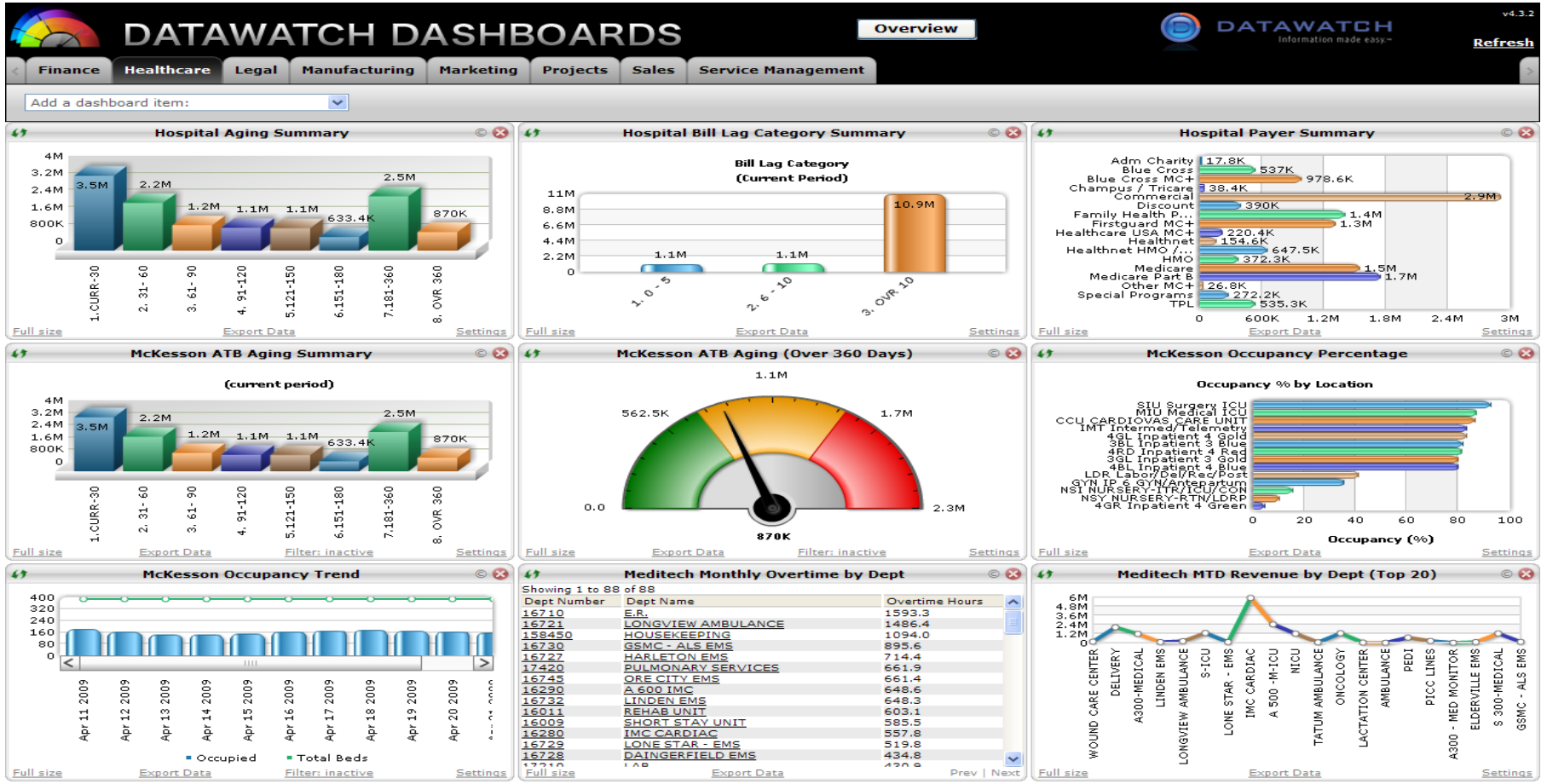
Principle 4: Use Structure

- Gestalt Psychology principles (1912):



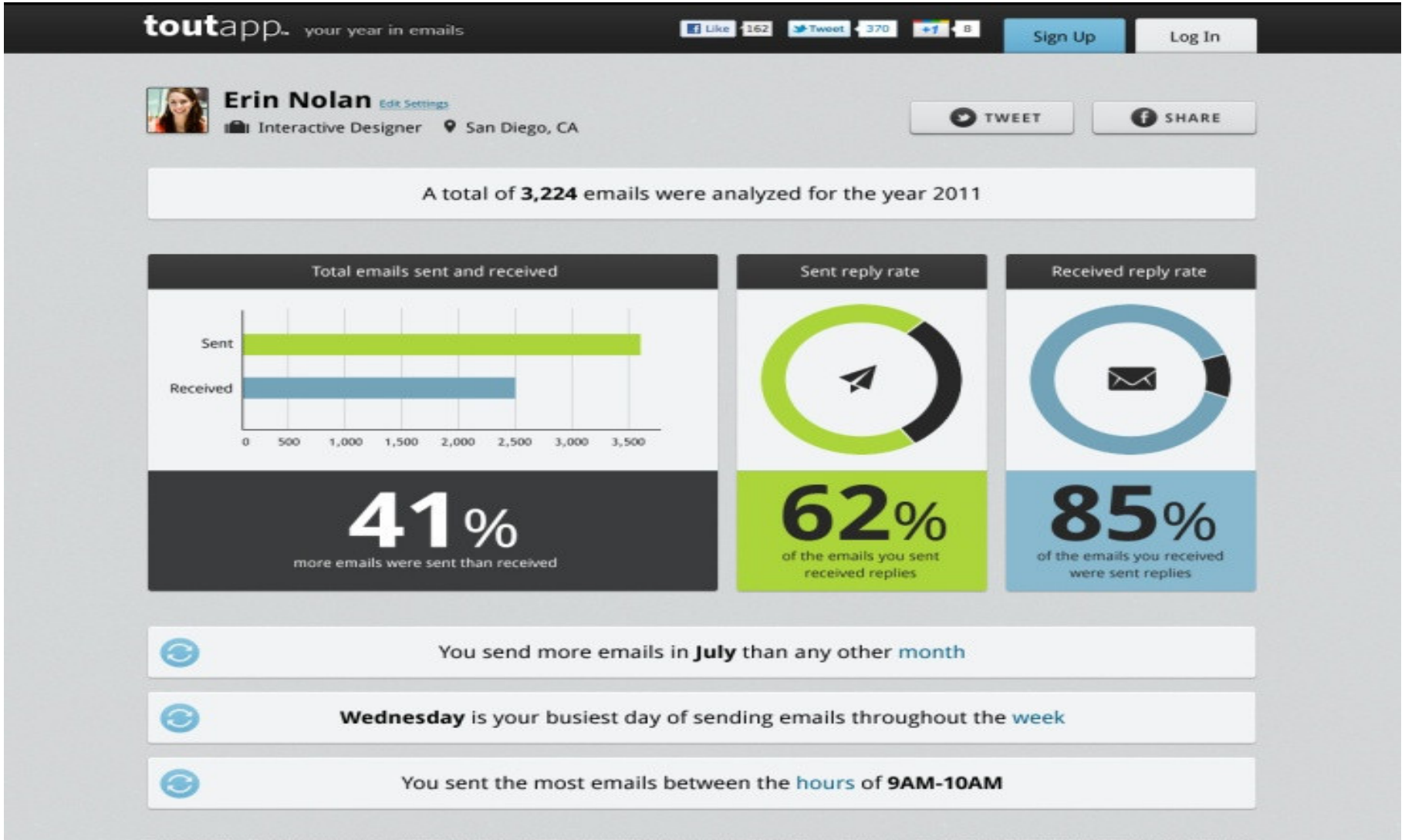
Source http://blog.fusioncharts.com/2014/03/how-to-use-the-gestalt-principles-for-visual-storytelling-podv/

Principle 4: Use Structure (but not like this)



Source https://www.vocalabs.com/blog/my-dashboard-pet-peeve

Principle 4: Use Structure



Source https://www.vocalabs.com/blog/my-dashboard-pet-peeve

Chart Selection – Andrew Abela

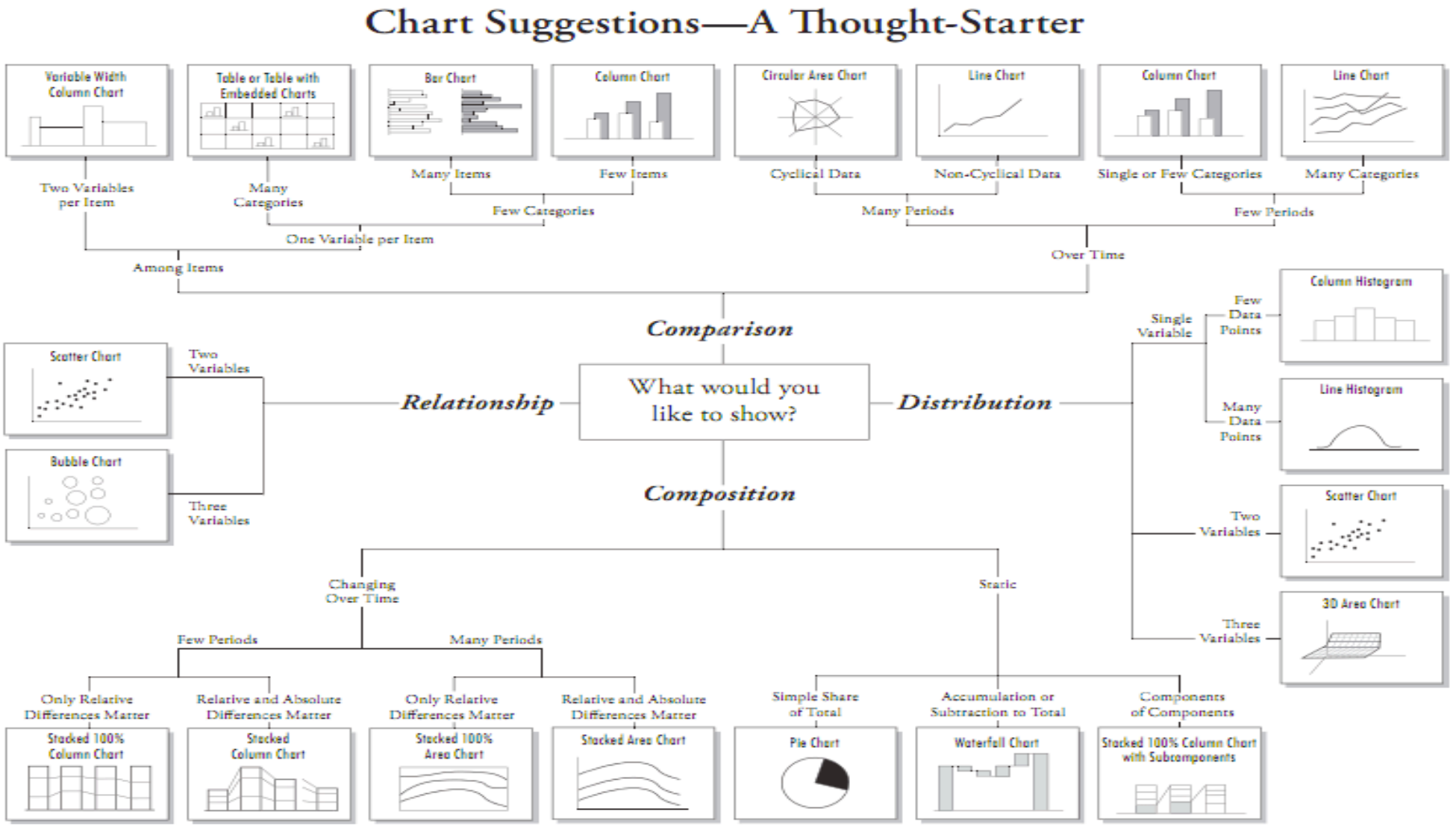
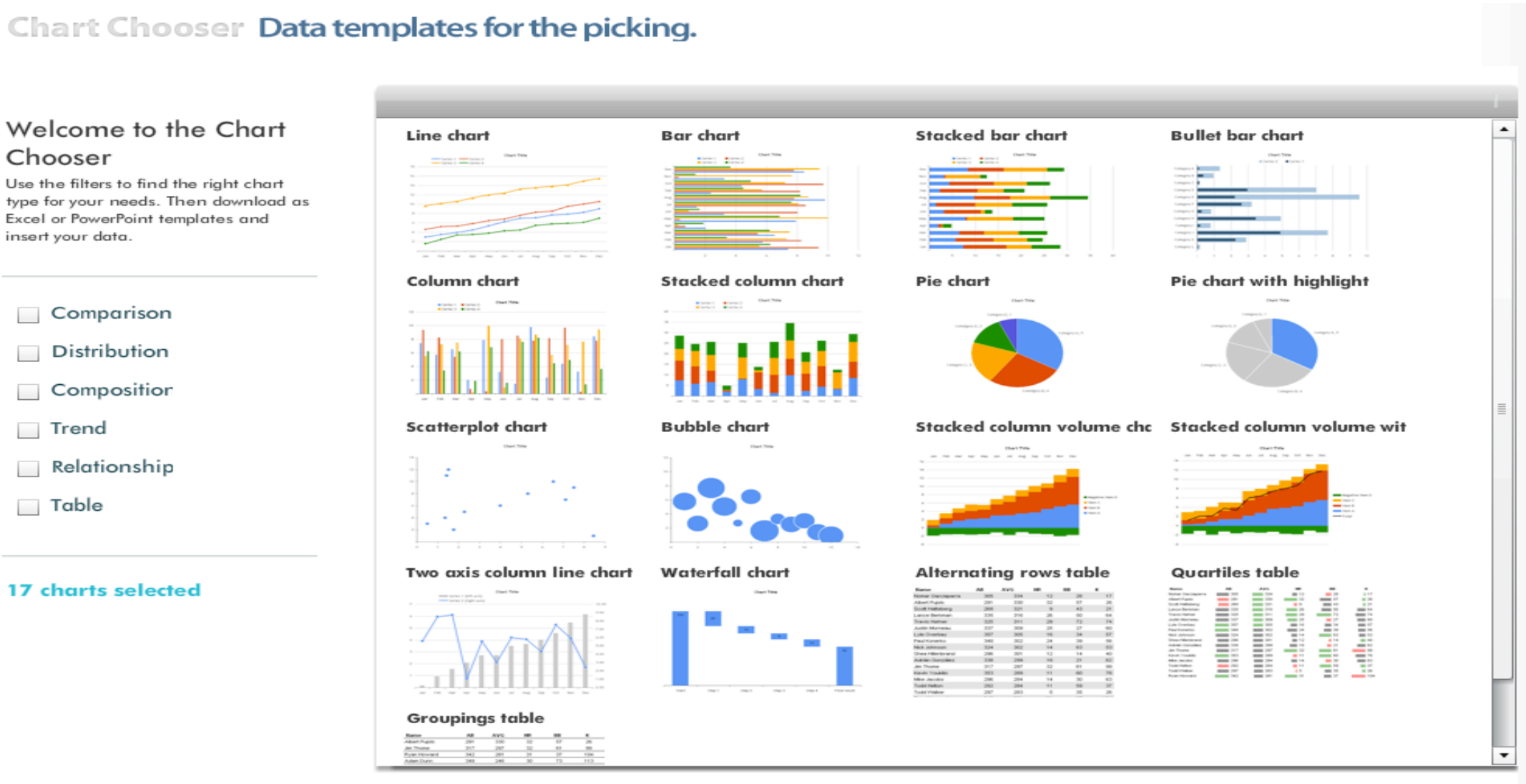
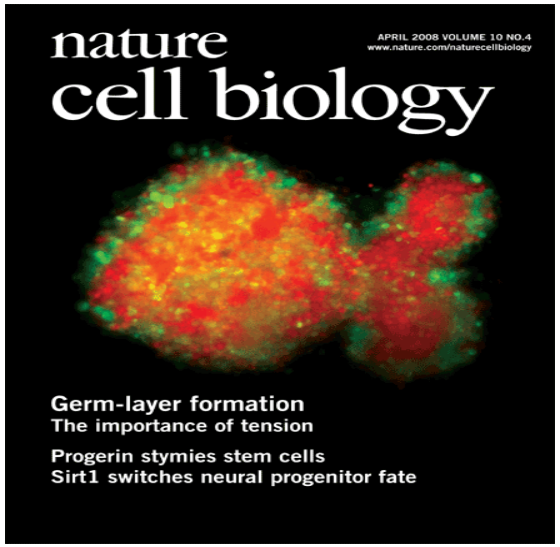
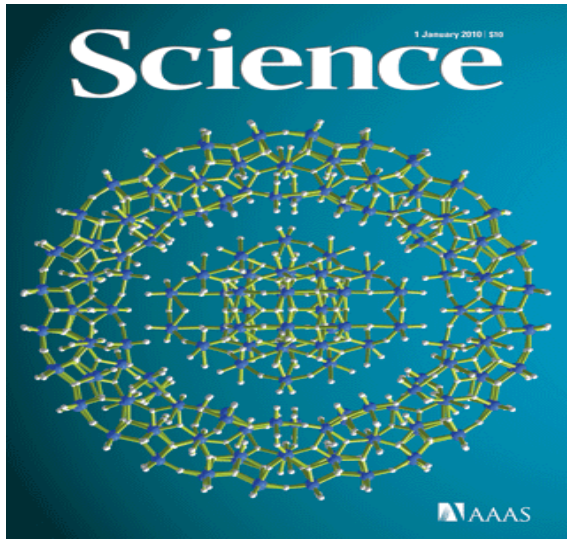


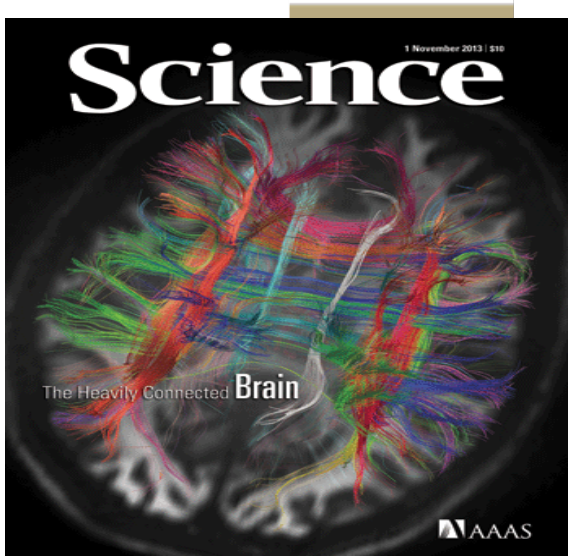
Chart Selection – Juice Analytics

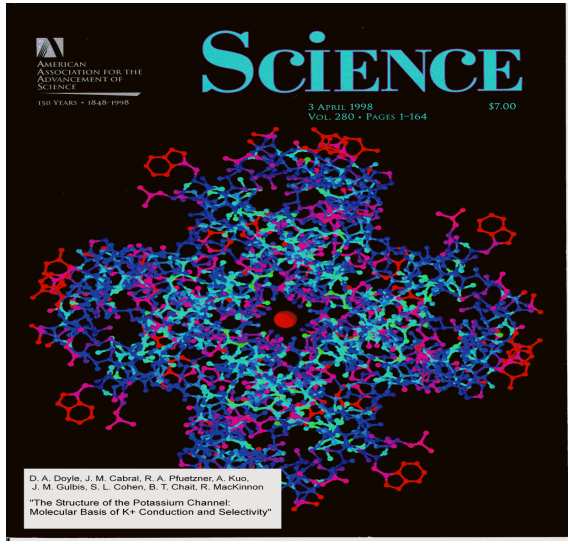


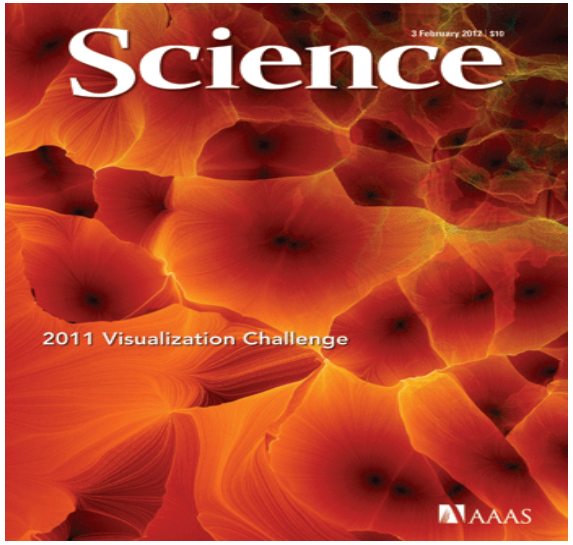
Data Viz in the Sciences











National Science Foundation
WHERE DISCOVERIES BEGIN

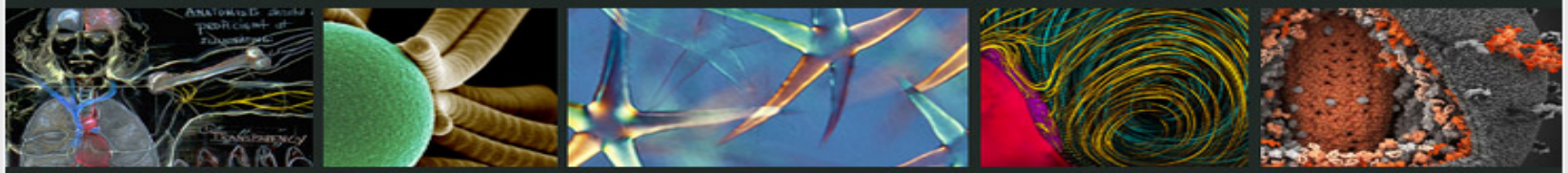
QUICK LINKS

SEARCH

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT NSF | FASTLANE

THE VIZZIES
VISUALIZATION CHALLENGE

The most beautiful visualizations from the worlds of science and engineering



Home | About | Timeline | Categories | Guidelines | Prizes | Winners

Important Dates

- The Visualization Challenge competition closes September 30, 2014.
- The deadline for all entries is 11:59 p.m. PST on Sept. 30, 2014.
- Competition judging rounds take place in October 2014.
- The 2014 winning entries will be announced in March 2015.
- Contest results will be publicly announced in *Popular Science* and on popsci.com in March 2015, and *Popular Photography* will recognize the winning photo. NSF will also publish the names of the winners on its website.

FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT NSF | FASTLANE

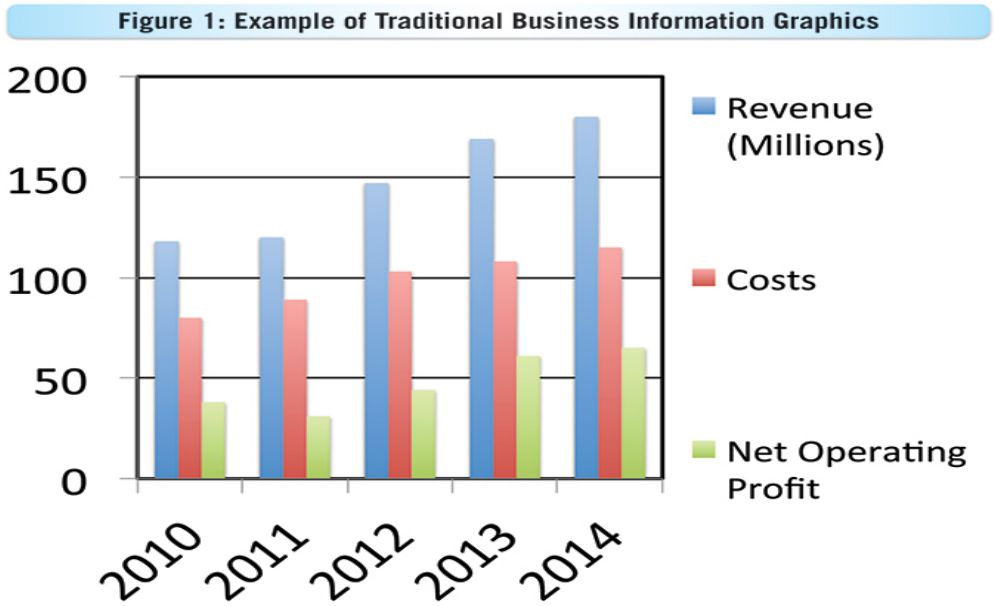
A case for Ugly visualizations

People instinctively gravitate to attractive visualizations, and they have a better chance of getting on the cover of a journal.

But does this conflict with the goals of visualization?:

- Rapid exploration
- Focus on most important details
- Easy and fast to develop and customize

e.g. Powerpoint vs Keynote



Interactive Toolkits: D3

Without Doubt, the most widely used interactive visualization framework is ***D3***, developed around 2011 by Jeff Heer, Mike Bostock and Vadim Ogievetsky.

Note from the authors: *D3 is intentionally a low-level system. During the early design of D3, we even referred to it as a "visualization kernel" rather than a "toolkit" or "framework"*