Old Dominion University

CS 791/891
Visualization Seminar
Fall 2015

Color, Design Principles, and Data

Dr. Michele C. Weigle

http://www.cs.odu.edu/~mweigle/CS891-F15/

Outline

- How -- color
- Design Principles
- Data Cleaning

Outline

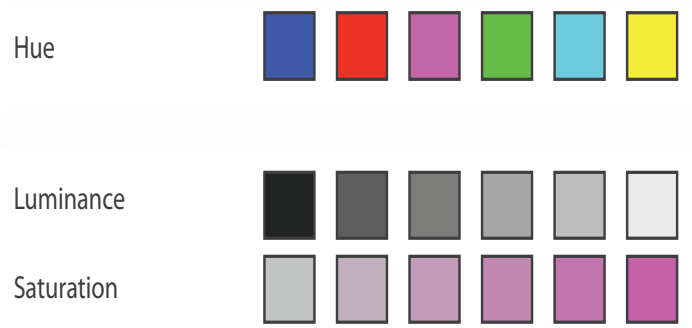
- Visualization Analysis Framework
  - Introduction: Definitions - Ch 1
  - Analysis: What, Why, How - Chs 2, 3
  - Marks and Channels - Ch 5
- Idiom Design Choices
  - Arrange Tables - Ch 7
  - Map Color - Ch 10
  - Manipulate, Facet, Reduce - Chs 11, 12, 14

Main reference:
Visualization Analysis and Design, Munzner. AK Peters / CRC Press, Oct 2014.
http://www.cs.ubc.ca/~tmm/courses/533/book/

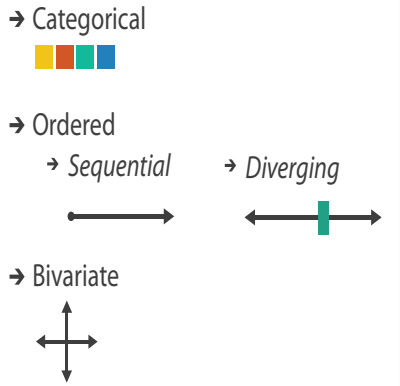
Slides in this style are shamelessly borrowed from Tamara Munzner's VAD minicourse, June 2014 (http://www.cs.ubc.ca/~tmm/talks.html#minicourse14)

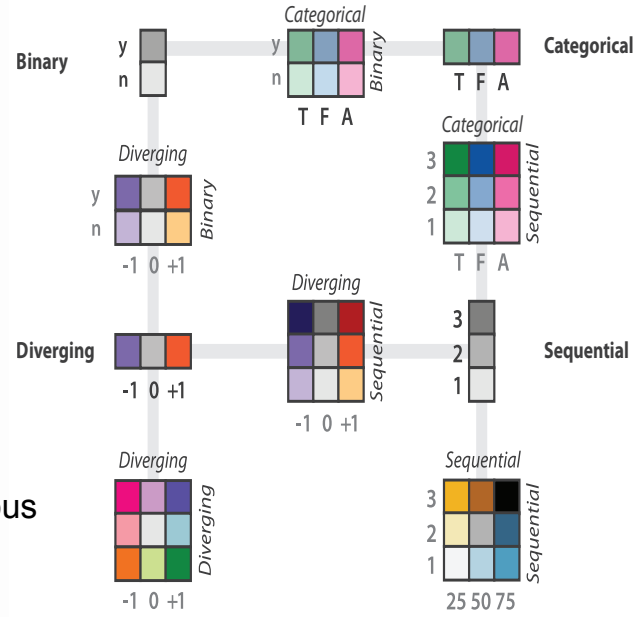
Color: Luminance, saturation, hue

- 3 channels
  - what/where for categorical
    - hue
  - how-much for ordered
    - luminance
    - saturation



Colormaps



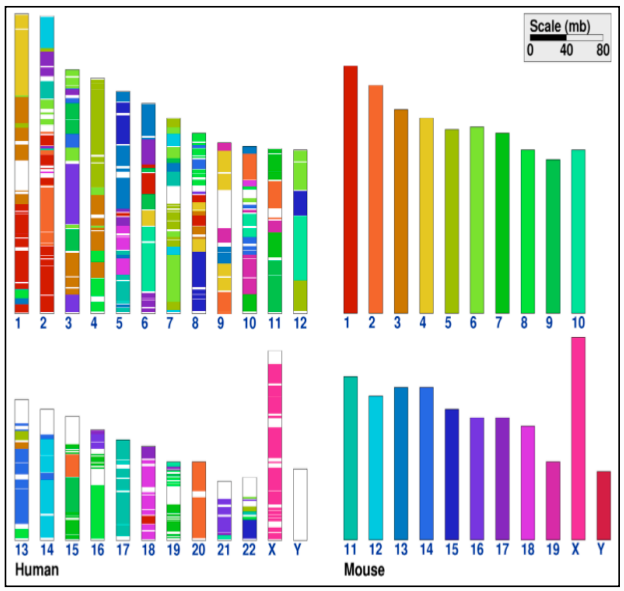


- categorical limits: noncontiguous
  - 6-12 bins hue/color
    - far fewer if colorblind
  - 3-4 bins luminance, saturation
  - size heavily affects salience
    - use high saturation for small regions, low saturation for large

after [Color Use Guidelines for Mapping and Visualization. Brewer, 1994. http://www.personal.psu.edu/faculty/c/a/cab38/ColorSch/Schemes.html]

Categorical color: Discriminability constraints

- noncontiguous small regions of color: only 6-12 bins



[Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. Sinha and Meller. BMC Bioinformatics, 8:82, 2007.]

Stone, *IEEE Computer Graphics and Applications*, 2012

Stone's Advice on Color

- Modesty! Less is more
- Use blue in large regions, not in thin lines
- Use red and green in the center of the field of view (edges of retina not sensitive to these)
- Use black, white, yellow in periphery
- Use adjacent colors that vary in hue and value

Stone, *IEEE Computer Graphics and Applications*, 2012

Stone's Advice on Color

- For large regions, don't use highly saturated colors (pastels a good choice)
- Do not use adjacent colors that vary in amount of blue
- Don't use high saturation, spectrally extreme colors together (causes after images)
- Use color for grouping and search
- Beware effects from adjacent color regions

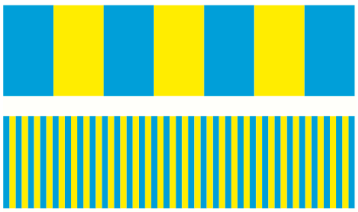


Figure 3. An example of spreading, demonstrating that the human eye blends small areas of color that are close together. The alternating cyan and yellow colors are the same in both bars.

Stone, *IEEE Computer Graphics and Applications*, 2012

Color and Size

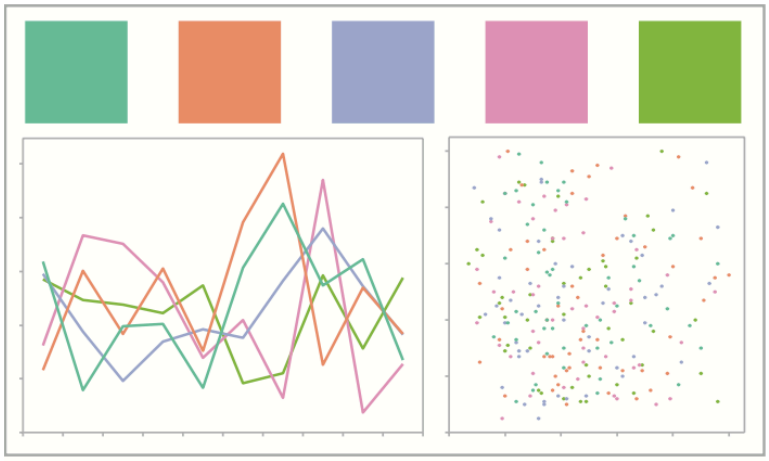


Figure 1. A set of colors with different hues but equal luminance (L* = 72). In the squares, these colors are distinct and visually balanced. However, in a line graph or scatter plot, they are harder to distinguish. The colors in this figure are adapted from one of the ColorBrewer qualitative color sets (www.colorbrewer.org).

Stone, *IEEE Computer Graphics and Applications*, 2012

Color and Size

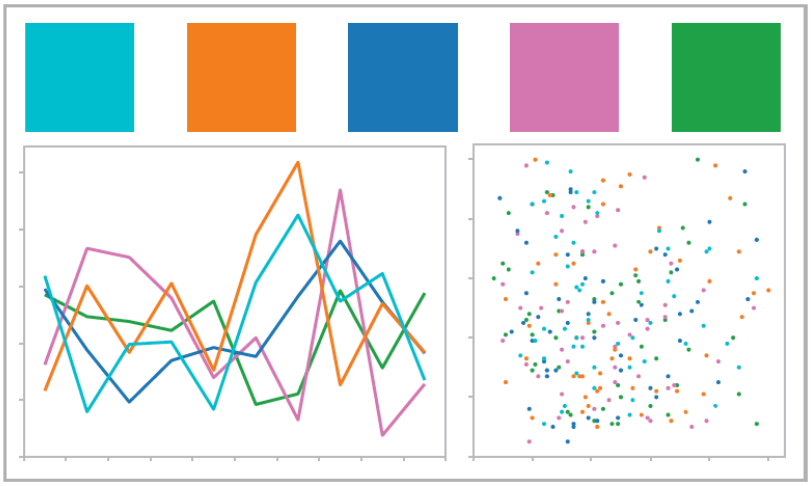


Figure 4. These colors vary in luminance, which makes them more distinguishable when rendered small. From left to right, L* = 70, 68, 47, 65, and 58.

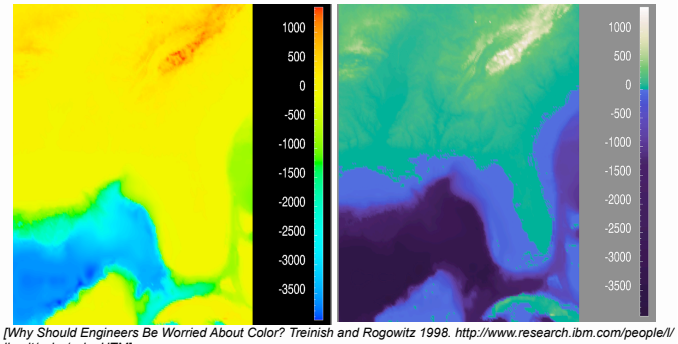
Do No Harm

- "The often scant benefits derived from coloring data indicate that even putting a good color in a good place is a complex matter. Indeed, so difficult and subtle that avoiding catastrophe becomes the first principle in bringing color to information: *Above all, do no harm*."

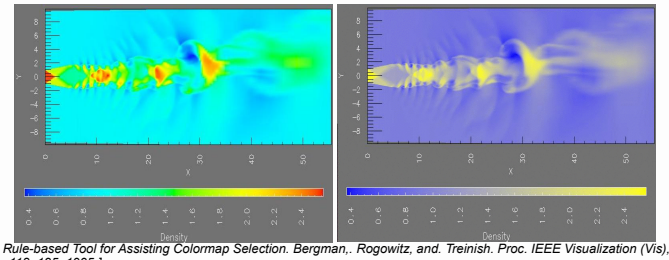
Tufte, *The Visual Display of Quantitative Information*

Ordered color: Rainbow is poor default

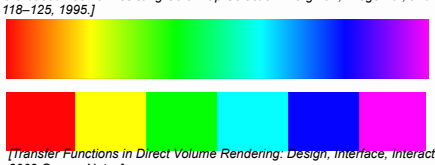
- problems
  - perceptually unordered
  - perceptually nonlinear
- benefits
  - fine-grained structure visible and nameable
- alternatives
  - fewer hues for large-scale structure
  - multiple hues with monotonically increasing luminance for fine-grained
  - segmented rainbows good for categorical, ok for binned



[Why Should Engineers Be Worried About Color? Treinish and Rogowitz 1998. http://www.research.ibm.com/people/l/lloydt/color/color.HTM]

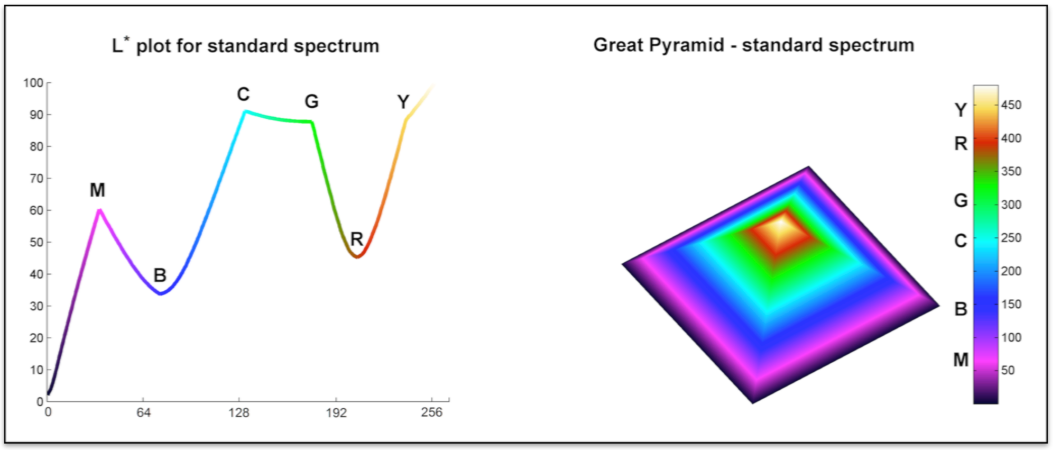


[A Rule-based Tool for Assisting Colormap Selection. Bergman,. Rogowitz, and. Treinish. Proc. IEEE Visualization (Vis), pp. 118–125, 1995.]



[Transfer Functions in Direct Volume Rendering: Design, Interface, Interaction. Kindlmann. SIGGRAPH 2002 Course Notes]

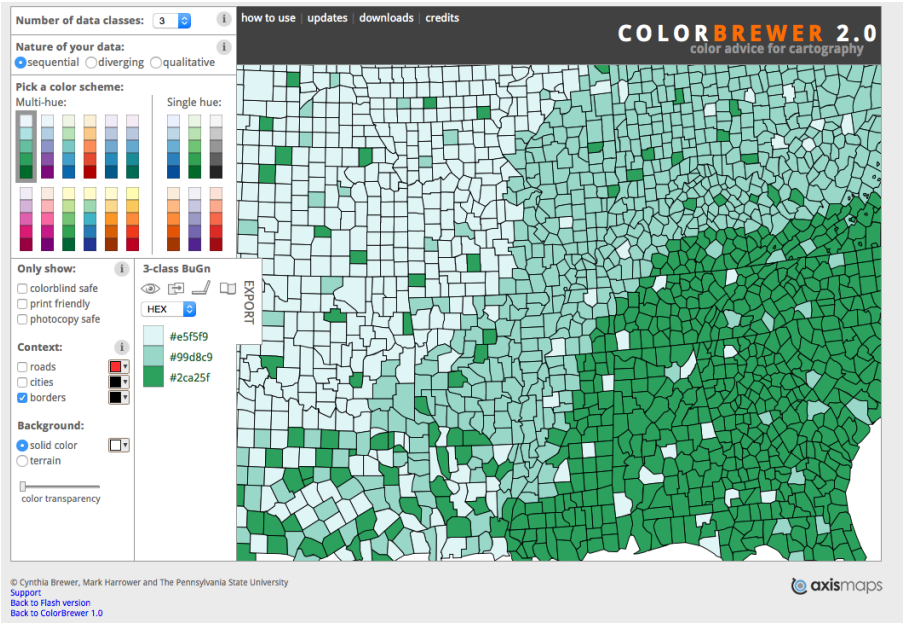
Here's why rainbow is bad



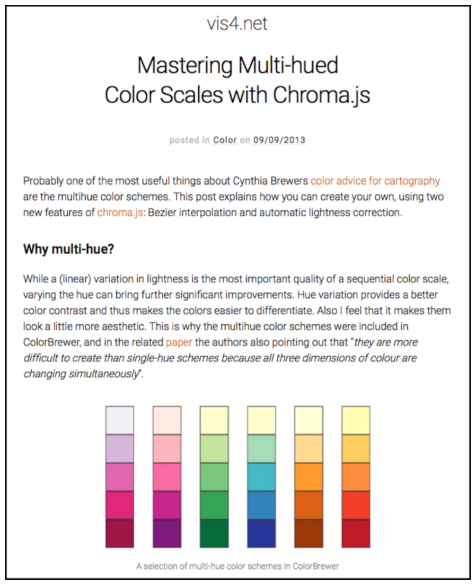
https://mycarta.files.wordpress.com/2013/01/spectrum-vs-cubeyf.png

ColorBrewer

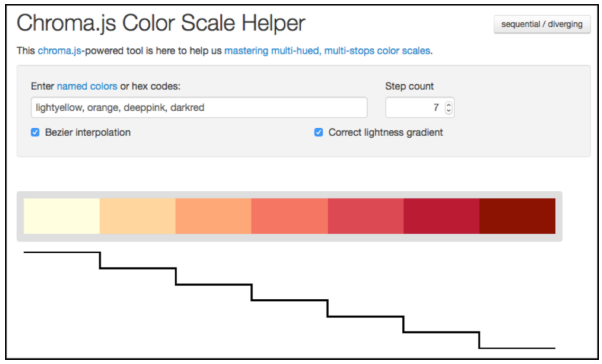
http://colorbrewer2.org



Multi-Hued Color Scale Help



https://vis4.net/blog/posts/mastering-multi-hued-color-scales/



http://gka.github.io/palettes/

Outline

- How -- color
- Design Principles
- Data Cleaning

Design Principles
Basic Principles

- Highlight the data
  - "Above else, show the data" -Tufte
- Organize the data
- Reduce the non-data ink
  - Would the data suffer loss of meaning or impact if this was eliminated?
- Enhance the data ink
- Revise and edit

 Few, *Show Me the Numbers*

Maximize Data Density
p. 166

"Data-rich designs give a *context* and *credibility* to statistical evidence.

Low-information designs are suspect: what is left out, what is hidden, why are we shown so little?

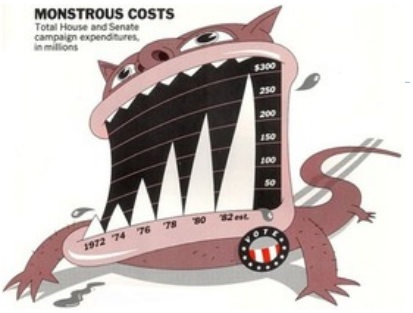
High-density graphics help us to compare parts of the data by displaying much information within the view of the eye: we look at one page at a time and the more on the page, the more effective and comparative our eye can be.

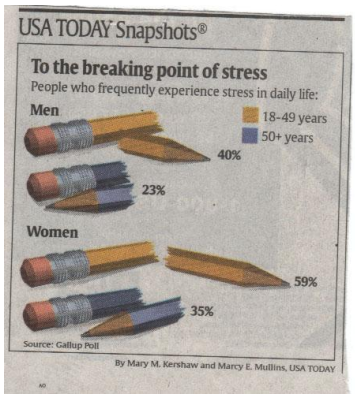
The principle, then, is: *Maximize data density and the size of the data matrix, within reason*."

 Tufte, *The Visual Display of Quantitative Information*

Chartjunk
USA Today

http://www.usatoday.com/news/snapshot.htm



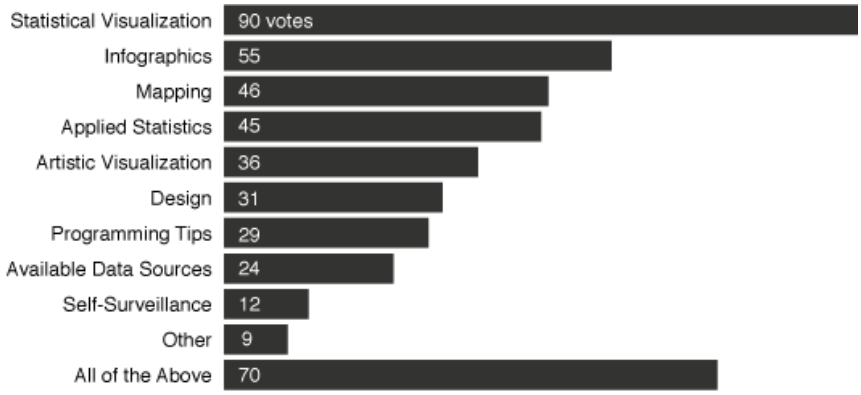


Design Principles
Few's Selection and Design Process

- Where to display each variable?
  - vertical vs. horizontal bars
  - multiple bars
  - small multiples

Where to Display Values
Vertical vs. Horizontal Bars

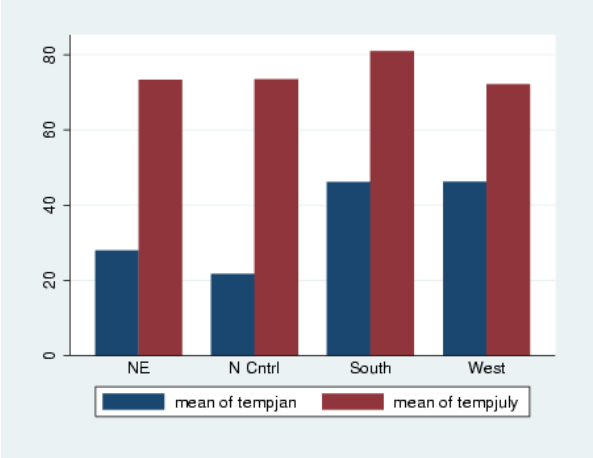
- Horizontal can be good if long text labels or many items



http://flowingdata.com/2008/08/19/3-worthwhile-alternatives-to-the-pie-chart/

Where to Display Values
Multiple Bars

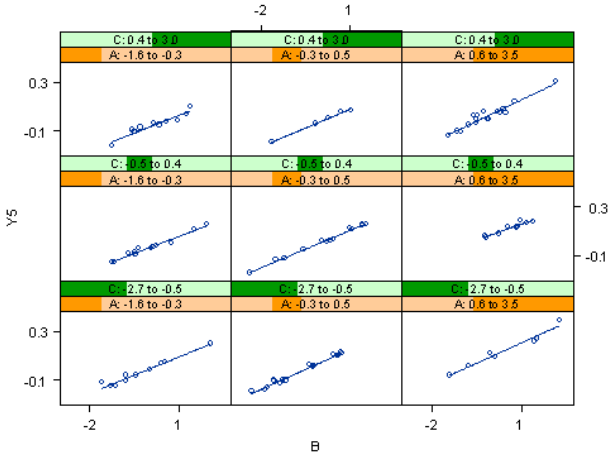
- Can be used to encode another variable



http://www.stata.com/support/faqs/graphics/gph/graphdocs/bar3.html

Where to Display Values
Small Multiples

- Can distribute a variable across graphs
- Sometimes called a *trellis display* or *small multiples*



http://www.amstat.org/publications/jse/02-046/02-046_blind.html

Few's Selection and Design Process
Best Design for the Remaining Objects

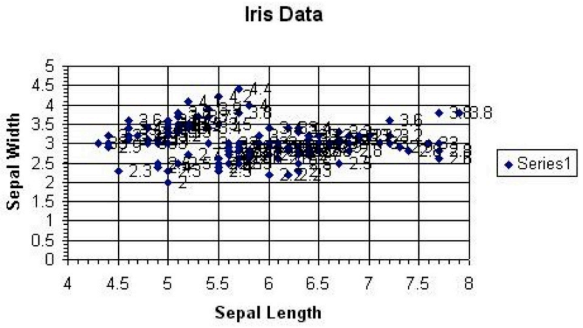
- Range of the quantitative scale?
  - bars - start from 0
  - lines, points - adjust scale so that it extends a little below the lowest data value and a little above the highest
  - bars - if you want to narrow the scale, switch from bars to points
- If a legend is required, where to place it?
  - the more directly you label the data, the better
  - bar - arrange labels to match the arrangement of bars
- Tick marks required for axis?
  - only necessary on quantitative scales
  - only major tick marks are necessary (5-10 usually enough)

Few's Selection and Design Process
Best Design for the Remaining Objects

- Best location for the quantitative scale?
  - avoid placing scale on right side unless necessary
  - if quantitative scale ranges between positive and negative, place axis at 0 (but don't let labels interfere with data)
- Grid lines required?
  - values cannot be interpreted with the necessary degree of accuracy
  - subset of points in multiple related scatter plots must be compared
  - if used, make them barely visible
- What descriptive text is needed?
  - title
  - axes titles

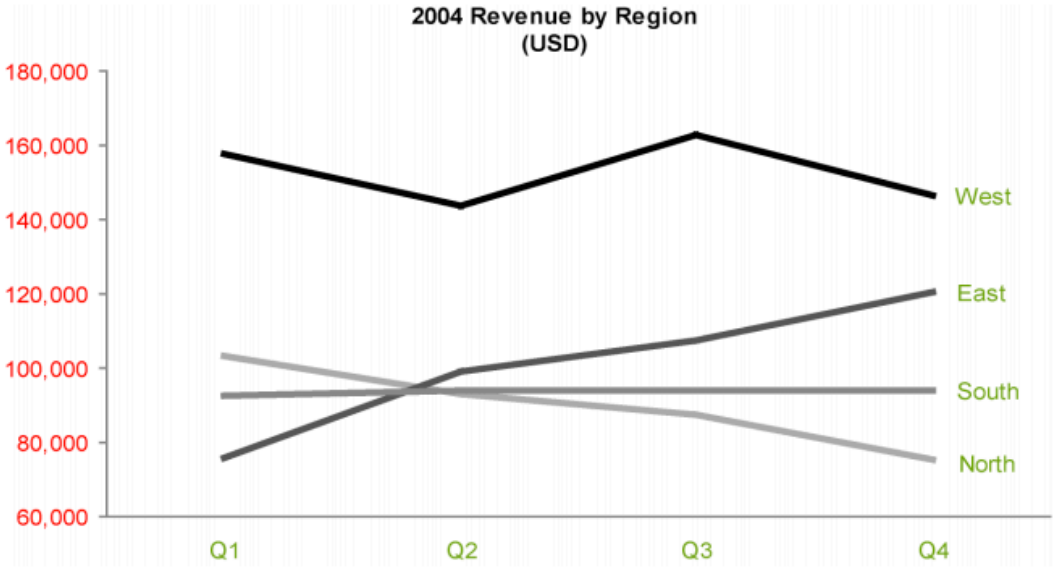
Keys, Labels, and Legends

- Caption
  - include data fields displayed and mappings used
- Grid/Tick Marks
  - but not too much
- Label Axes
  - *include units!*



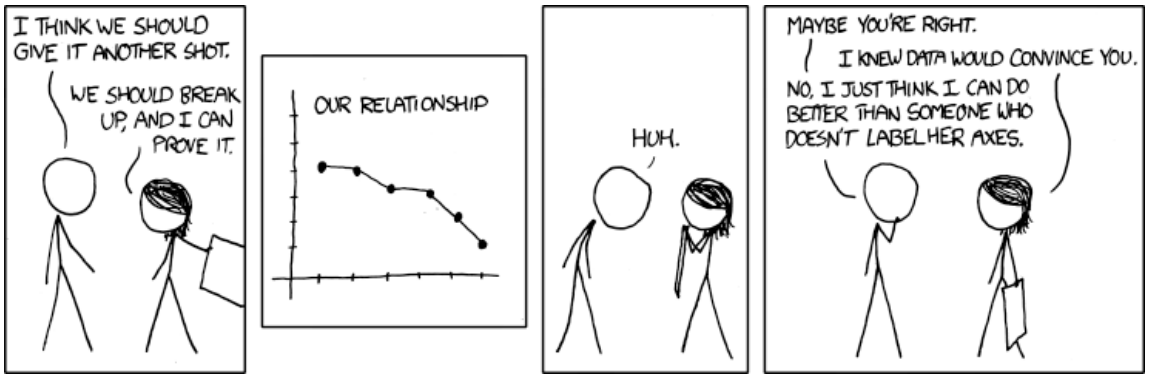
- If using multiple views, be consistent in labeling and use of grid/tick marks

Labeling Data



 Few, "Effectively Communicating Numbers"

Axis Labels
My Pet Peeve



http://xkcd.com/833/

Design Principles
Few's Selection and Design Process

- Should particular data be featured, and if so, how?
  - use bright or dark colors (when using soft colors for everything else)
  - bars - place borders around only those bars that should be highlighted
  - lines - make lines to be highlighted thicker
  - points - make featured points larger or include fill color in them alone

Design Principles
Overplotting Solutions

- Reducing size of data objects
- Removing all fill color from data objects
- Changing the shape of data objects
- Jittering data objects
- Making data objects transparent
- Encoding the density of values
- Reducing the number of values
  - Aggregating the data
  - Filtering the data
  - Breaking the data into a series of separate graphs
  - Statistically sampling the data

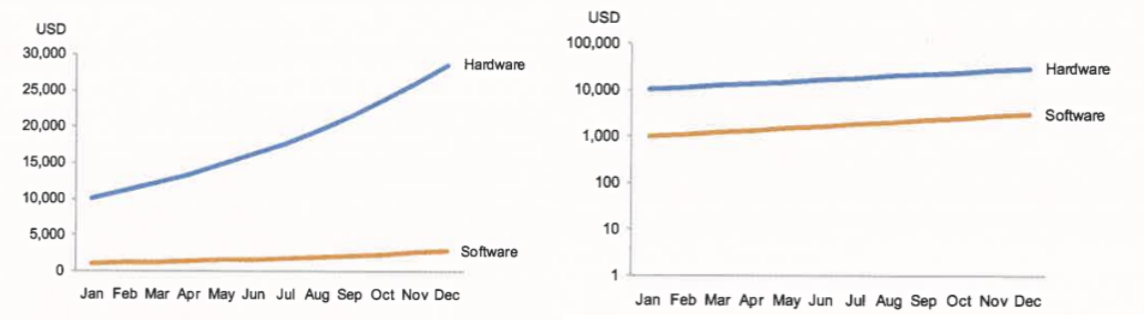
Design Principles
Content Is King

- Quality, relevance and integrity of the content is fundamental
- What's the analysis task? Make the visual design reflect that
- Integrate text, chart, graphic, map into a coherent narrative

 Stasko, GaTech 2011

Figs 7.42-7.43, p. 172, Few, *Now You See It*

Importance of Scale

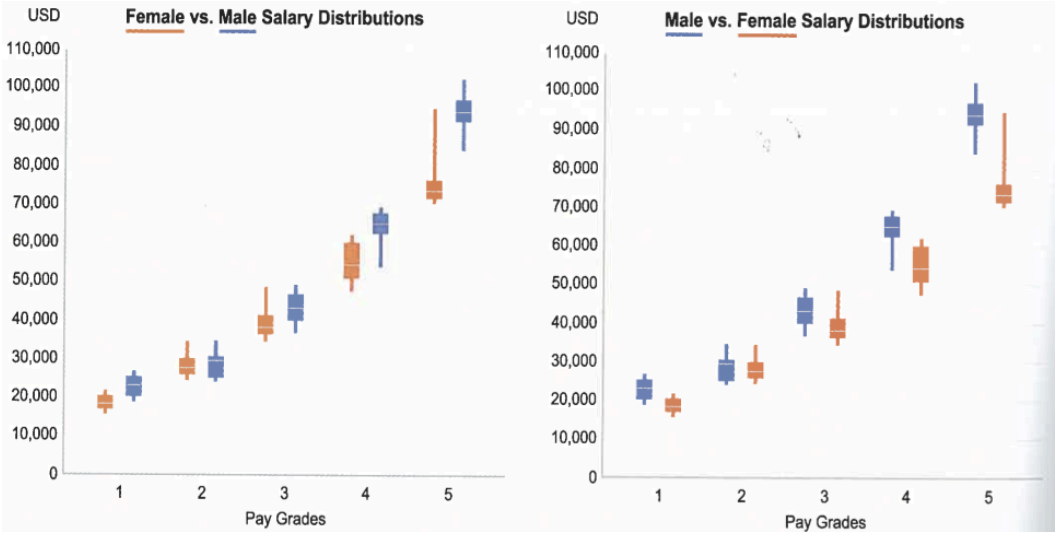


- Which is increasing at a faster rate, hardware sales or software sales?
  - Both at same rate, 10%
  - Log scale shows this

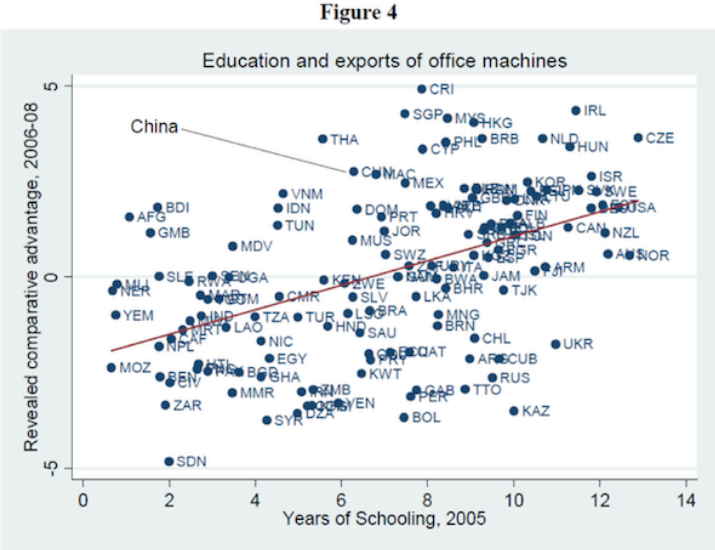
Fig 10.36, p. 234, Few, *Now You See It*

Comparisons

- Note how first one's curve is smooth (not such a noticeable difference). Second one is more noticeable. Same data.

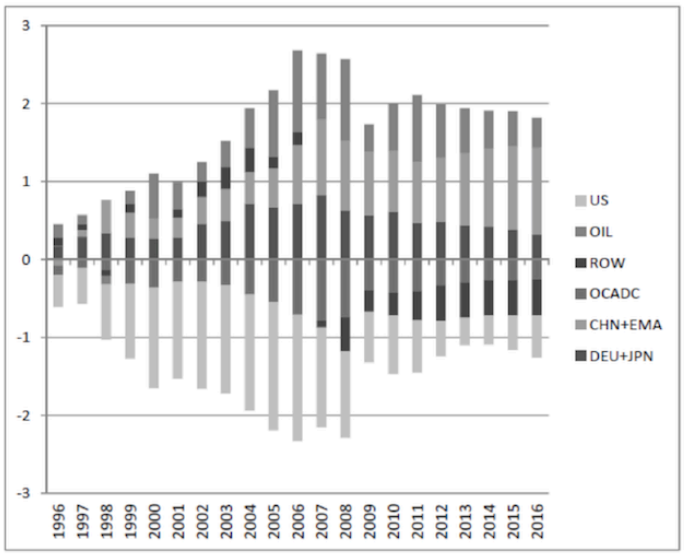


What's the problem here?



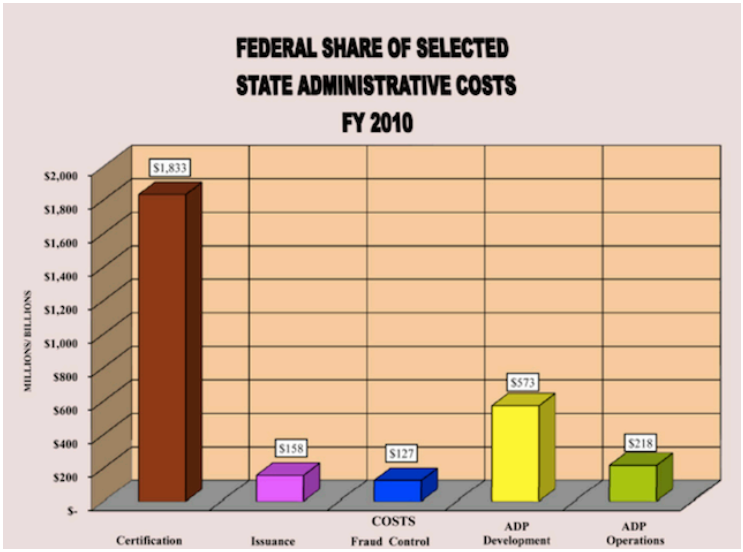
http://www.visualisingdata.com/index.php/2012/08/guest-post-how-governments-can-better-use-data-visualization/

What's the problem here?



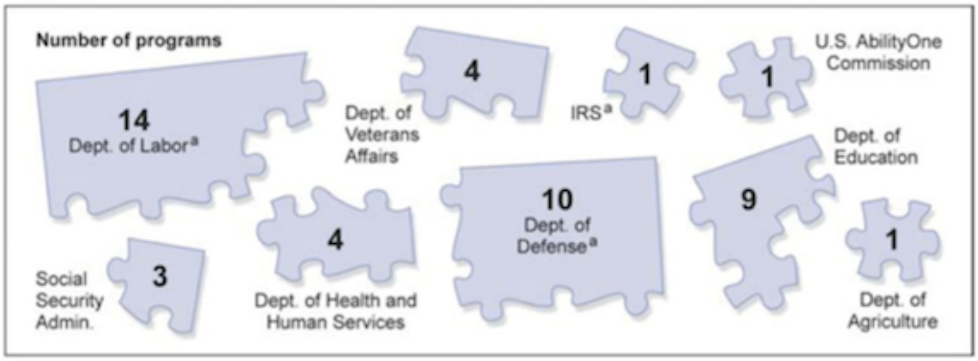
http://www.visualisingdata.com/index.php/2012/08/guest-post-how-governments-can-better-use-data-visualization/

What's the problem here?



http://www.visualisingdata.com/index.php/2012/08/guest-post-how-governments-can-better-use-data-visualization/

What's the problem here?



http://www.visualisingdata.com/index.php/2012/08/guest-post-how-governments-can-better-use-data-visualization/

Outline

- How -- color
- Design Principles
- Data Cleaning

Data Formats

- Delimited Text
  - tabbed delimited
  - comma delimited (CSV)
- Extensible Markup Language (XML)
  - looks a bit like HTML
  - user-defined tags to identify data
- JavaScript Object Notation (JSON)
  - collection of name/value pairs
  - smaller than XML
  - easier to parse

Example of Data Formats

CSV

```
20090101,26
20090102,34
20090103,27
```

JSON

```
{"observations": [
   {"date":"20090101", "max_temp":26},
   {"date":"20090102", "max_temp":34},
   {"date":"20090103", "max_temp":27}
 ]
}
```

XML

```
<weather_data>
  <observation>
        <date>20090101</date>
        <max_temp>26</max_temp>
  </observation>
  <observation>
        <date>20090102</date>
        <max_temp>34</max_temp>
  </observation>
  <observation>
        <date>20090103</date>
        <max_temp>27</max_temp>
  </observation>
</weather_data>
```

Yau, *Visualize This*, Ch 2

How to convert between data formats?

- Write a program to convert from one format to another
  - awk (my favorite, but I'm old school), Python, Perl, PHP, ...
- Other tools
  - search Google for "csv to json", "csv to xml", "xml to json"
- Mr. Data Converter
  - http://shancarter.github.io/mr-data-converter/
  - developed by a graphics editor at *The New York Times*
  - input: CSV or tab-delimited data
  - output: HTML table, JSON, MySQL, Python, PHP, Ruby, XML, ...

Data in the Real World

- Data can be missing, have typos, be inconsistent, spread over multiple tables, etc.
- Two big issues:
  - format
  - accuracy

World Disasters - Inconsistent

	A	B	C	D	E	F	G	H	I	J
1	Start	End	Country	Location	Type	Sub_Type	Name	Killed	Cost	Affected
2	102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...	Drought	Drought			280000	
3	72006	2006	Afghanistan		Drought	Drought			1900000	
4	52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...	Drought	Drought		37	2580000	0.05
5	81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	Drought				
6	11969	1969	Afgh						48000	0.2
7	29072006	29072006	Afgh					1	935	
8	13122005	13122005	Afgh					5	501	
9	8102005	8102005	Afgh					1		0.05
10	18072004	18072004	Afgh					2	1040	
11	10042003	10042003	Afgh					1	1001	
12	12042002	12042002	Afgh					50	6150	
13	25032002	25032002	Afgh					1000	91228	
14	3032002	3032002	Afgh					150	3513	
15	1062001	1062001	Afgh					4	270	
16	25022001	25022001	Afgh							

	A	B	C	D
1	Start	End	Country	Location
2	102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...
3	72006	2006	Afghanistan	
4	52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...
5	81971	1973	Afghanistan	Central, North-West, Nort ...
6	11969	1969	Afghanistan	Paktia province
7	29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...
8	13122005	13122005	Afghanistan	Hindu Kush
9	8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...
10	18072004	18072004	Afghanistan	Paktia province
11	10042003	10042003	Afghanistan	Yakabagh (Takhar province ...
12	12042002	12042002	Afghanistan	Dawabi, Khojakeder (Nahri ...
13	25032002	25032002	Afghanistan	Nahrin (Baghlan province) ...
14	3032002	3032002	Afghanistan	Dakhli-Ezeu (Hindu Kush m ...
15	1062001	1062001	Afghanistan	Jabul Saraj, Gumbahar, Pa ...
16	25022001	25022001	Afghanistan	Faizabad region (Badakhsh ...

http://www.infochimps.com/datasets/disasters-worldwide-from-1900-2008

World Disasters - Missing

	A	B	C	D	E	F	G	H	I	J
1	Start	End	Country	Location	Type	Sub_Type	Name	Killed	Cost	Affected
2	102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...	Drought	Drought			280000	
3	72006	2006	Afghanistan		Drought	Drought			1900000	
4	52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...	Drought	Drought		37	2580000	0.05
5	81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	Drought				
6	11969	1969	Afghanistan						48000	0.2
7	29072006	29072006	Afghanistan					1	935	
8	13122005	13122005	Afghanistan					5	501	
9	8102005	8102005	Afghanistan					1		0.05
10	18072004	18072004	Afghanistan					2	1040	
11	10042003	10042003	Afghanistan					1	1001	
12	12042002	12042002	Afghanistan					50	6150	
13	25032002	25032002	Afghanistan					1000	91228	
14	3032002	3032002	Afghanistan					150	3513	
15	1062001	1062001	Afghanistan					4	270	
16	25022001	25022001	Afghanistan							

	A	B	C	H	I	J
1	Start	End	Country	Killed	Cost	Affected
2	102008	102008	Afghanistan		280000	
3	72006	2006	Afghanistan		1900000	
4	52000	2002	Afghanistan	37	2580000	0.05
5	81971	1973	Afghanistan			
6	11969	1969	Afghanistan		48000	0.2
7	29072006	29072006	Afghanistan	1	935	
8	13122005	13122005	Afghanistan	5	501	
9	8102005	8102005	Afghanistan	1		0.05
10	18072004	18072004	Afghanistan	2	1040	
11	10042003	10042003	Afghanistan	1	1001	
12	12042002	12042002	Afghanistan	50	6150	
13	25032002	25032002	Afghanistan	1000	91228	
14	3032002	3032002	Afghanistan	150	3513	
15	1062001	1062001	Afghanistan	4	270	
16	25022001	25022001	Afghanistan			

http://www.infochimps.com/datasets/disasters-worldwide-from-1900-2008

What to do with dirty data?

	A	B	C	D	H	I	J
1	Start	End	Country	Location	Killed	Cost	Affected
2	102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...		280000	
3	72006	2006	Afghanistan			1900000	
4	52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...	37	2580000	0.05
5	81971	1973	Afghanistan	Central, North-West, Nort ...			
6	11969	1969	Afghanistan	Paktia province		48000	0.2
7	29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...	1	935	
8	13122005	13122005	Afghanistan	Hindu Kush	5	501	
9	8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	1		0.05
10	18072004	18072004	Afghanistan	Paktia province	2	1040	
11	10042003	10042003	Afghanistan	Yakabagh (Takhar province ...	1	1001	
12	12042002	12042002	Afghanistan	Dawabi, Khojakeder (Nahri ...	50	6150	
13	25032002	25032002	Afghanistan	Nahrin (Baghlan province) ...	1000	91228	
14	3032002	3032002	Afghanistan	Dakhli-Ezeu (Hindu Kush m ...	150	3513	
15	1062001	1062001	Afghanistan	Jabul Saraj, Gumbahar, Pa ...	4	270	
16	25022001	25022001	Afghanistan	Faizabad region (Badakhsh ...			

Data Cleaning Tools

Quick Tools

- Data Science Toolkit
  - http://www.datasciencetoolkit.org/
  - lots of quick conversion tools
- Mr. People
  - http://people.ericson.net/
  - formats lists of names
- Mr. Data Converter
  - http://shancarter.github.io/mr-data-converter/

Full Apps

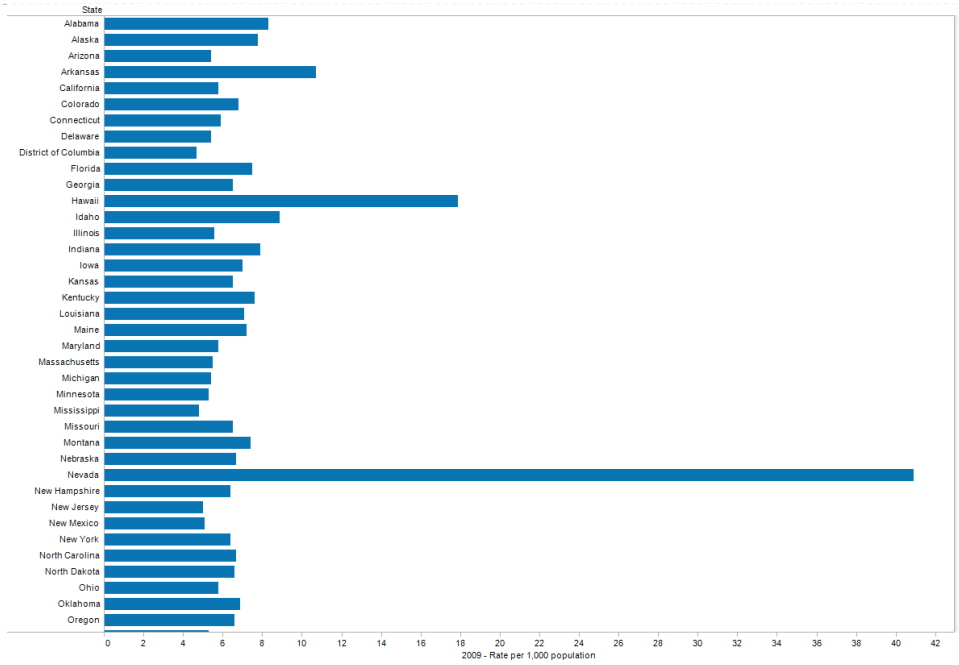
- Data Wrangler
  - http://vis.stanford.edu/wrangler/app/
  - video: http://vimeo.com/19185801
- Open Refine (was Google Refine)
  - http://openrefine.org/
  - video: http://www.youtube.com/watch?v=B70J_H_zAWM
  - more info: http://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning

What about accuracy?

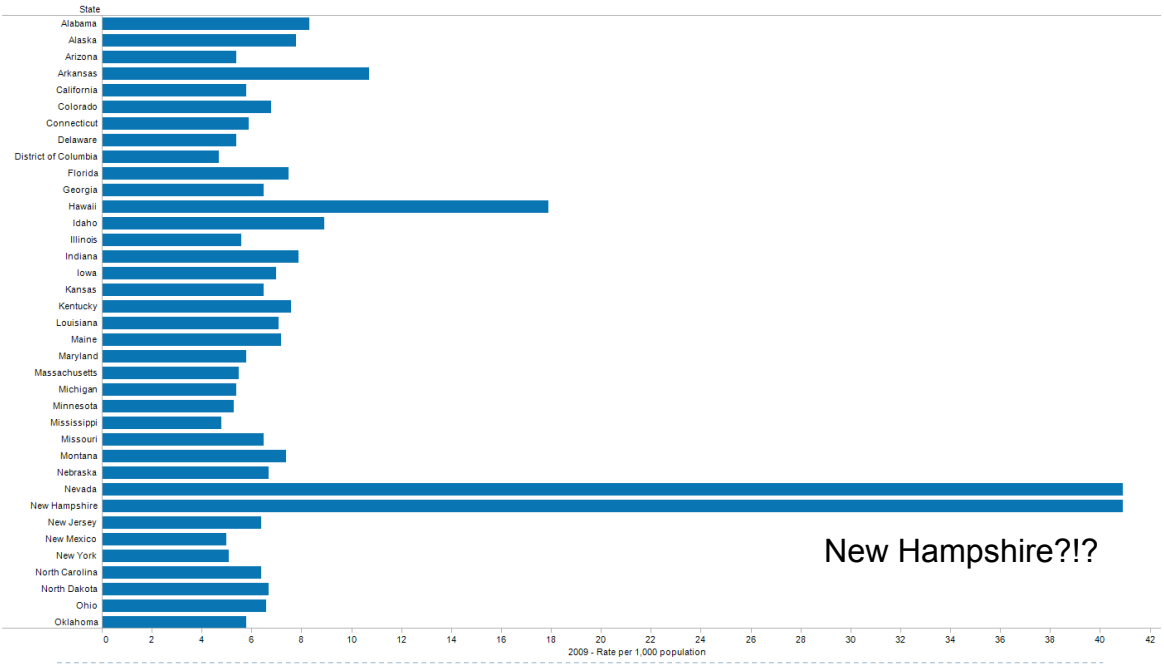
- Nathan Yau (*Visualize This, Data Points*) was intern at *The New York Times*
- One day, his entire goal was to verify 3 numbers in a dataset
- *Must have accurate data before can trust the visualization*

Yau, *Visualize This*, Ch 1

Yes, this is actually a problem



Graphing the raw data



Let's look at the Excel file

	Rate 1,000 population \2						
	1990	2000	2005	2006	2007	2008	2009
Maryland	9.7	5.8	6.9	6.6	6.5	6.0	5.8
Massachusetts	7.9	6.7	6.2	5.9	5.9	5.7	5.5
Michigan	8.2	6.8	6.0	5.9	5.9	5.5	5.4
Minnesota	7.7	6.9	5.9	6.0	6.0	5.5	5.3
Mississippi	9.4	7.8	5.8	5.7	5.7	5.2	4.8
Missouri	9.6	7.3	7.0	6.9	6.9	6.8	6.5
Montana	8.6	7.6	7.4	7.5	7.5	7.7	7.4
Nebraska	8.0	72.2	7.0	6.8	6.8	6.9	6.7
Nevada	99.0	9.4	57.8	52.6	52.6	43.1	40.9
New Hampshire	9.5	6.0	7.2	7.1	7.1	6.8	40.9
New Jersey	7.6	6.0	5.7	5.5	5.5	5.4	6.4
New Mexico \5	8.8	8.0	6.6	6.9	6.9	4.0	5.0
New York \5	8.6	7.1	6.8	6.8	6.8	6.5	5.1
North Carolina	7.8	8.2	7.3	7.3	7.3	7.0	6.4
North Dakota	7.5	7.2	6.9	6.8	6.8	6.7	6.7
Ohio	9.0	7.8	6.5	6.3	6.3	6.0	6.6
Oklahoma	10.6	(NA)	7.3	7.3	7.3	7.1	5.8
Oregon	8.9	7.6	7.3	7.2	7.2	6.9	6.9
Pennsylvania	7.1	6.0	5.8	5.7	5.7	5.6	5.3
Rhode Island	8.1	7.6	7.0	6.5	6.5	6.2	5.9

http://www.census.gov/compendia/statab/2012/tables/12s0133.xls

Now, let's look at the PDF

	Number (1,000)			Rate per 1,000 population [2]		
	1990	2000	2009	1990	2000	2009
Maryland	46.3	40.0	32.4	9.7	7.5	5.8
Massachusetts	47.7	37.0	36.7	7.9	5.8	5.5
Michigan	76.1	66.4	53.1	8.2	6.7	5.4
Minnesota	33.7	33.4	28.4	7.7	6.8	5.3
Mississippi	24.3	19.7	14.5	9.4	6.9	4.8
Missouri	49.1	43.7	39.8	9.6	7.8	6.5
Montana	6.9	6.6	7.1	8.6	7.3	7.4
Nebraska	12.6	13.0	12.5	8.0	7.6	6.7
Nevada	120.6	144.3	108.2	99.0	72.2	40.9
New Hampshire	10.5	11.6	8.5	9.5	9.4	6.4
New Jersey	58.7	50.4	46.3	7.6	6.0	5.0
New Mexico [5]	13.3	14.5	10.2	8.8	8.0	5.1
New York [5]	154.8	162.0	120.1	8.6	7.1	6.4
North Carolina	51.9	65.6	65.8	7.8	8.2	6.7
North Dakota	4.8	4.6	4.3	7.5	7.2	6.6
Ohio	98.1	88.5	64.8	9.0	7.8	5.8
Oklahoma	33.2	15.6	23.5	10.6	(NA)	6.9
Oregon	25.3	26.0	23.5	8.9	7.6	6.6
Pennsylvania	84.9	73.2	64.2	7.1	6.0	5.3
Rhode Island	8.1	8.0	6.5	8.1	7.6	5.9

http://www.census.gov/compendia/statab/2012/tables/12s0133.pdf

But wait, there's more funny stuff

	PDF	Excel
Maryland	5.8	5.8
Massachusetts	5.5	5.5
Michigan	5.4	5.4
Minnesota	5.3	5.3
Mississippi	4.8	4.8
Missouri	6.5	6.5
Montana	7.4	7.4
Nebraska	6.7	6.7
Nevada	40.9	40.9
New Hampshire	6.4	40.9
New Jersey	5.0	6.4
New Mexico [5]	5.1	5.0
New York [5]	6.4	5.1
North Carolina	6.7	6.4
North Dakota	6.6	6.7
Ohio	5.8	6.6
Oklahoma	6.9	5.8
Oregon	6.6	6.9
Pennsylvania	5.3	5.3
Rhode Island	5.9	5.9

Bottom Line

- If you see something weird in your graph that you can't explain, go back and double-check your data
- Even if you didn't make an error, maybe someone else did

Data Sources

- Some notable ones
  - Data.gov
  - Google Public Data Explorer
    - http://www.google.com/publicdata/directory
  - Census Bureau
    - Census data visualization gallery - http://www.census.gov/dataviz/
  - Federal Reserve
    - FRASER - http://fraser.stlouisfed.org/
    - FRED - http://research.stlouisfed.org/fred2/