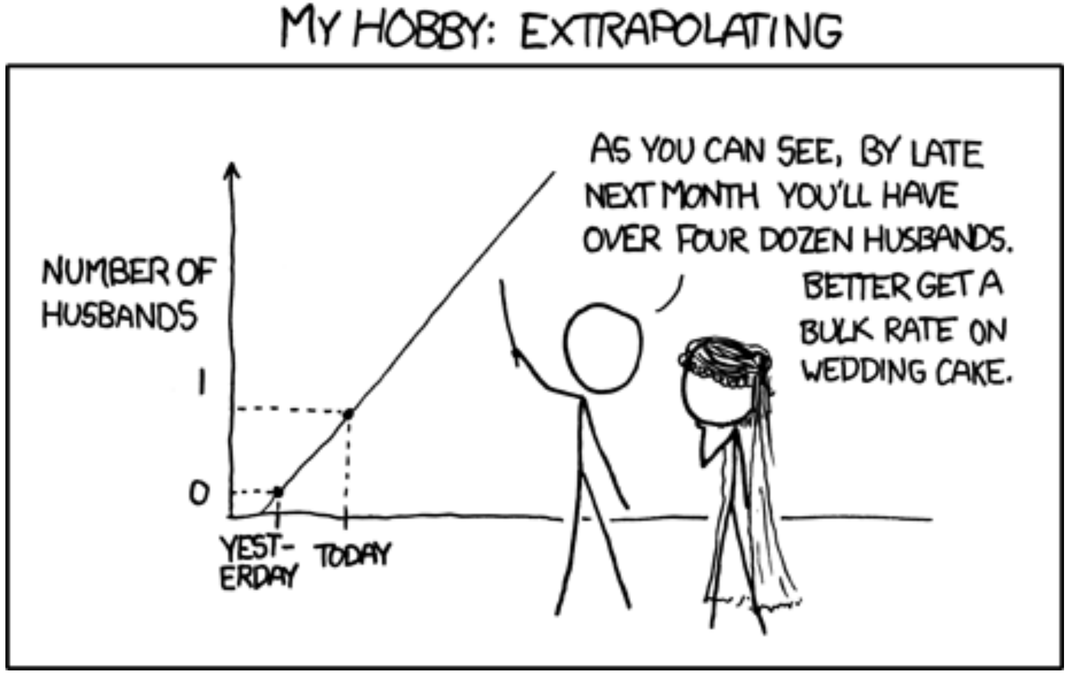
CS171 Visualization

Alexander Lex
alex@seas.harvard.edu

Tables



[xkcd]

HARVARD
School of Engineering and Applied Sciences

This Week

Reading: VAD, Chapters 6 & 7

Lecture 9: Tables

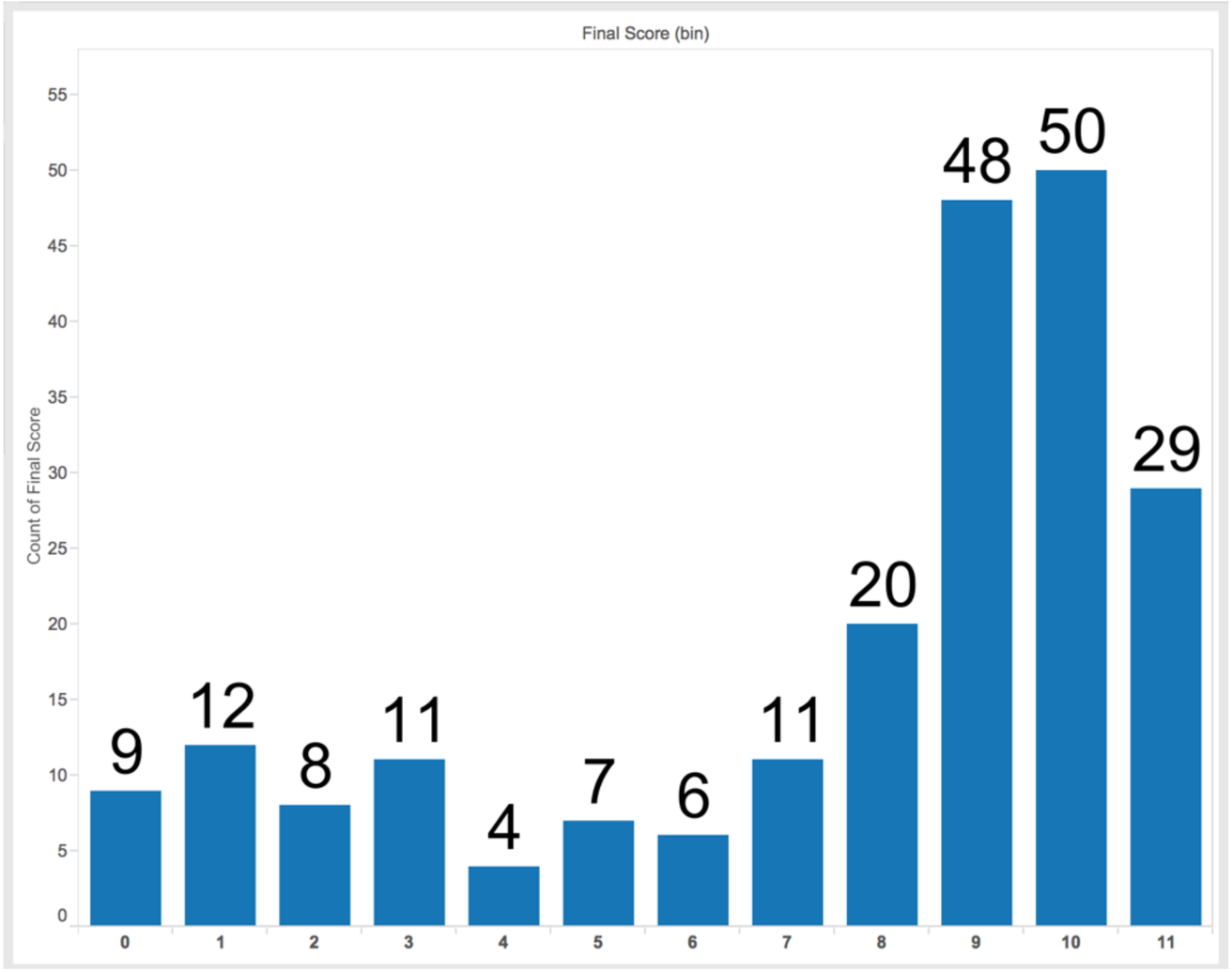
Lecture 10: Graphs

Sections: Designing your Visualization

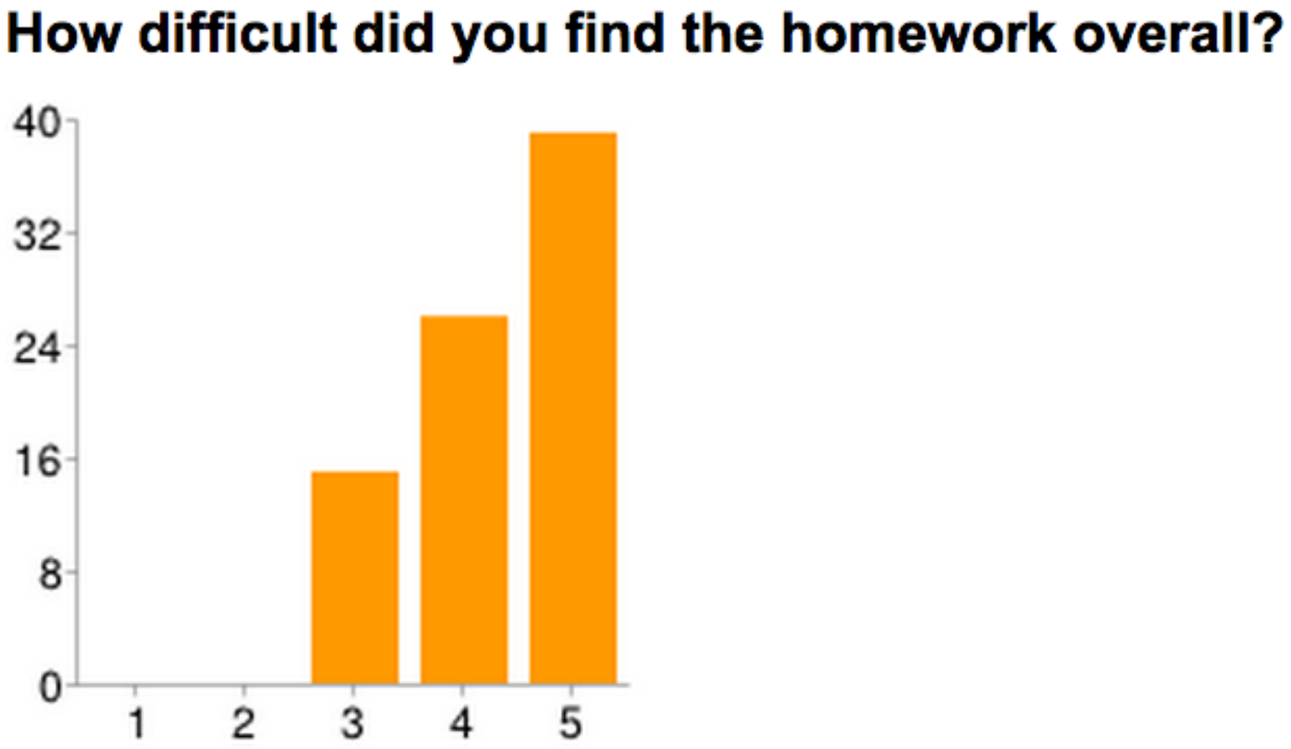
Homework 1 Review

Score Distribution

Average: 7.8



How Difficult?



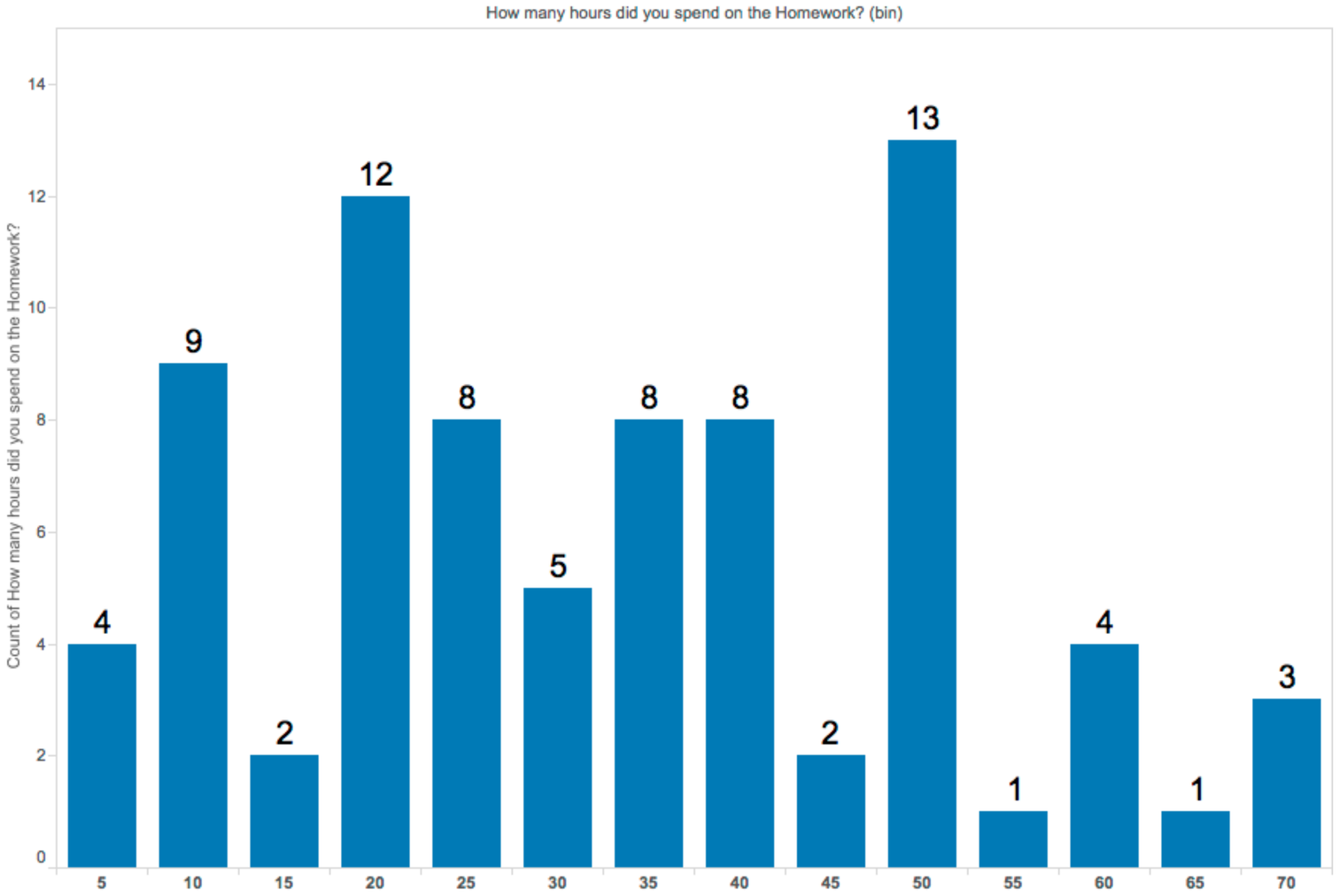
1	**0**	0%
2	**0**	0%
3	**15**	19%
4	**26**	33%
5	**39**	49%

How Long?

N=81

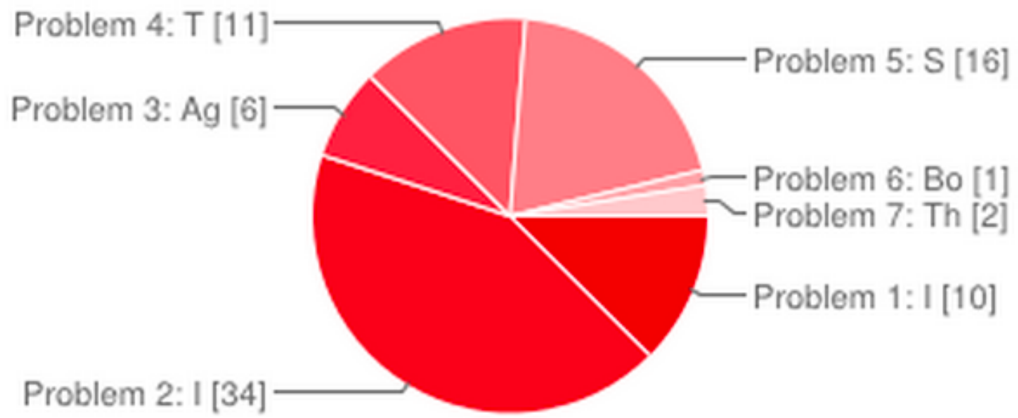
Average: 33.85

Goal: 20



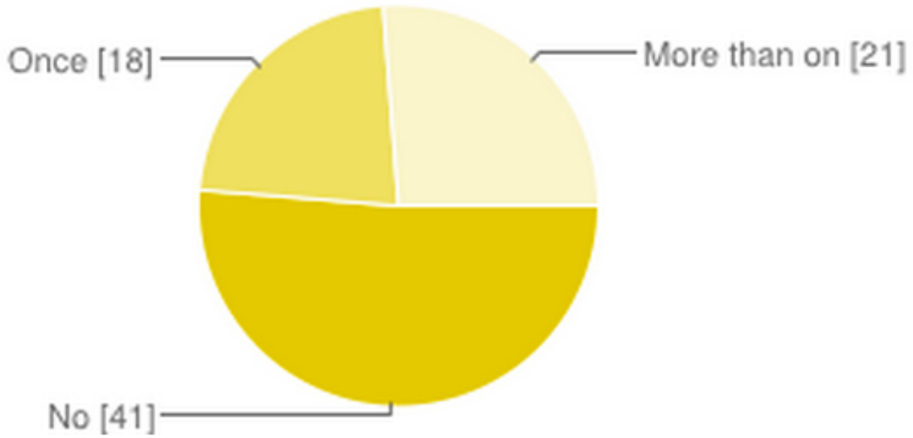
Which part took longest?

What part of HW1 did you spend the most time on?



Problem 1: Improving the visual table design	**10**	13%
Problem 2: Interactive filtering	**34**	43%
Problem 3: Aggregating continents	**6**	8%
Problem 4: Time-dependent visualization	**11**	14%
Problem 5: SVG bar chart	**16**	20%
Problem 6: Bonus	**1**	1%
Problem 7: Theory	**2**	3%

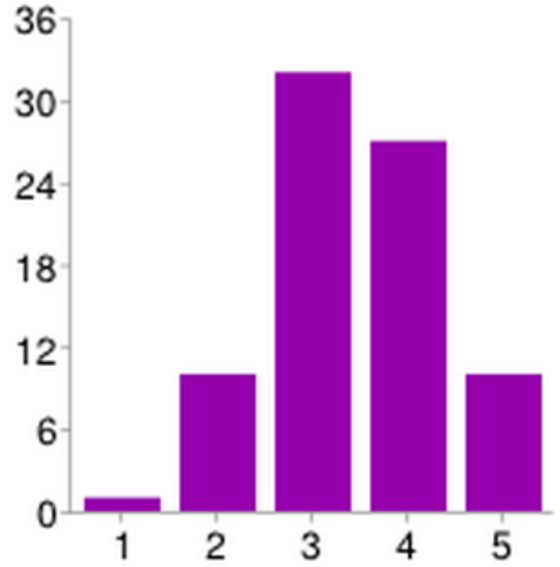
Office Hours Attendance



No	**41**	51%
Once	**18**	23%
More than once	**21**	26%

Are Sections Helpful?

How helpful do you find the sections for the homework?



1	**1**	1%
2	**10**	13%
3	**32**	40%
4	**27**	34%
5	**10**	13%

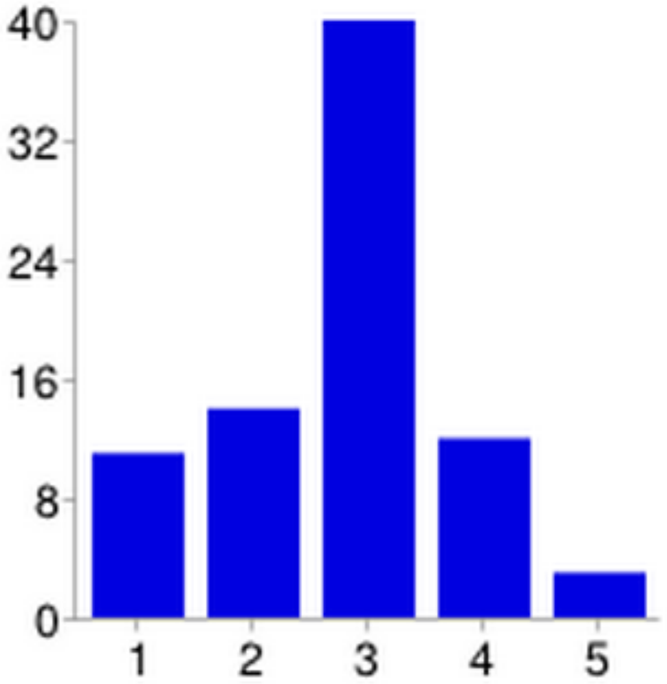
Section Comments

“Pertinent and just enough momentum to get you thinking in the right direction. Section presenter delivered an enthusiastic and polished lesson.”

“Topics covered were too easy! Homework problems were way harder.”

Design Studio

How helpful did you find the design studio (already for HW2)?



1	**11**	14%
2	**14**	18%
3	**40**	50%
4	**12**	15%
5	**3**	4%

Design Studio Comments

“I felt it was a huge waste of time because I'm still struggling with d3 let alone attempting a creative design. Also, we didn't really do anything in class.”

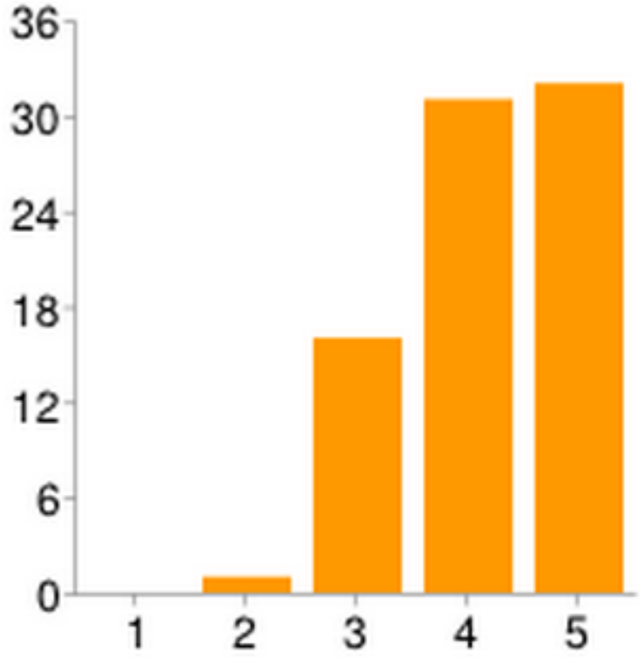
“DESIGN STUDIOS ARE HARD. Wow, it was cool to see our group trying to think of all of the complex things we could draw and just how quickly it all got overly complex. Might be nice to see an example DS after HW2 is submitted.”

“A lot of fun!”

“nice chance to interact with more people while working”

General Difficulty

In general, how difficult are you finding the course?



1	**0**	0%
2	**1**	1%
3	**16**	20%
4	**31**	39%
5	**32**	40%

General Comments

“The learning curve is quite steep for someone who does not do programming regularly”

“I think there is a large discrepancy between the contents of lecture and the problem sets that we are given. Generally, I don't understand why most of the lectures focus on visualization theory and do not discuss actual coding itself.”

“Theory might need to be a little bit harder. Some of the code, I think is too hard. Really freaking good course though.”

“Please teach us some real code and design problems in lecture. It's a disaster for people who learn Javascript first time.”

What you need to know

Theory

Lecture
Reading
Discussion

Design Lecture
Design Studios

Sections
D3 reading
Self-study
Office hours

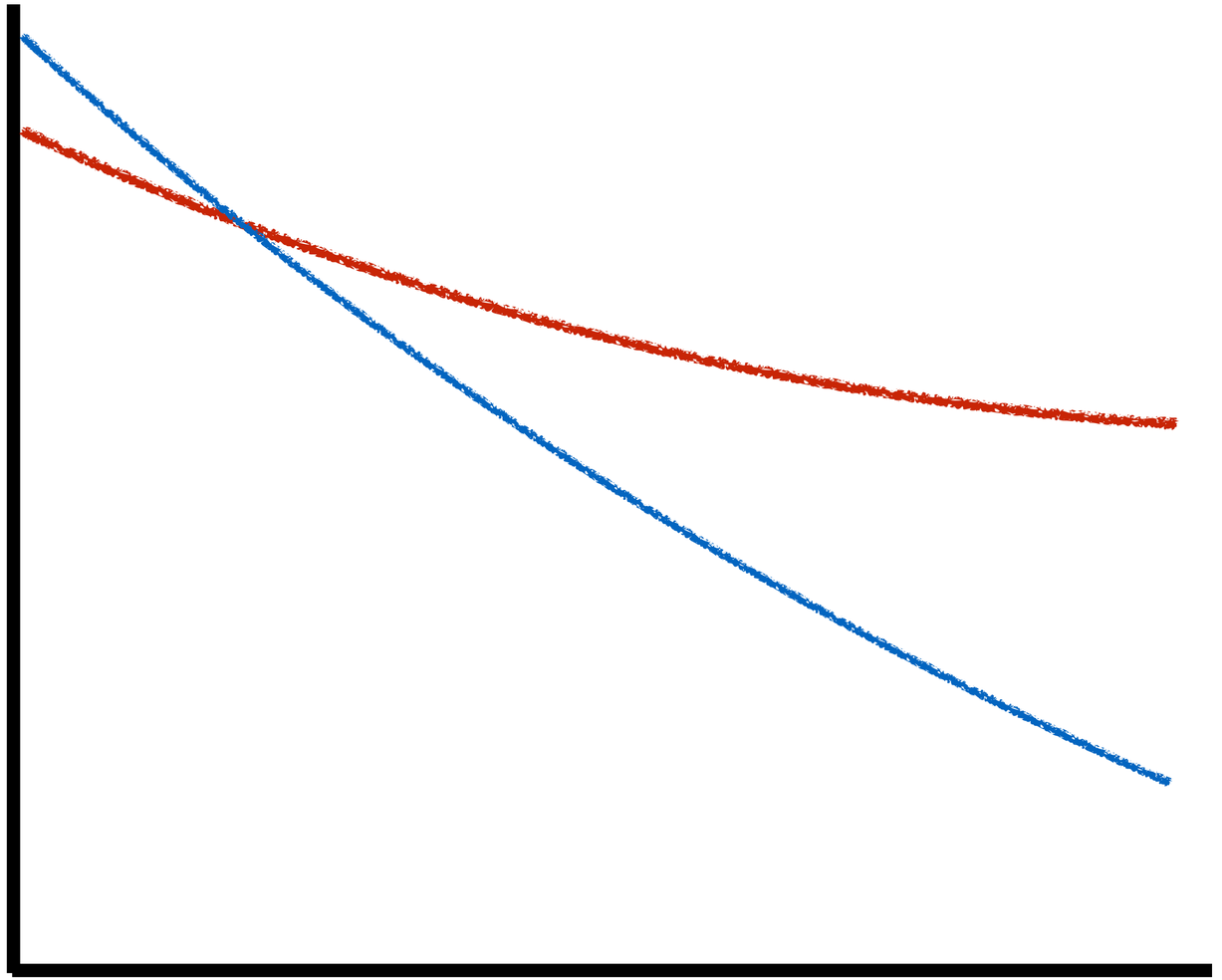
Design Skills ⟷ Coding Skills


```
<!DOCTYPE html>
<meta charset="utf-8">
<style>

text {
  font: 10px sans-serif;
}

</style>
<body>
<script src="http://d3js.org/d3.v3.min.js"></script>
<script>
```

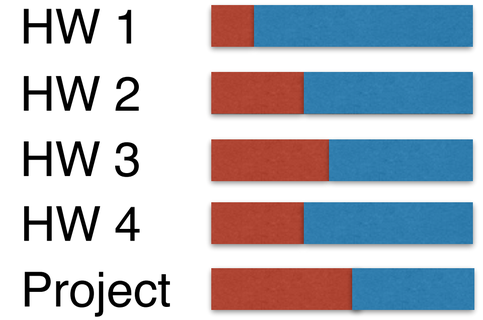
Half-Life of Knowledge

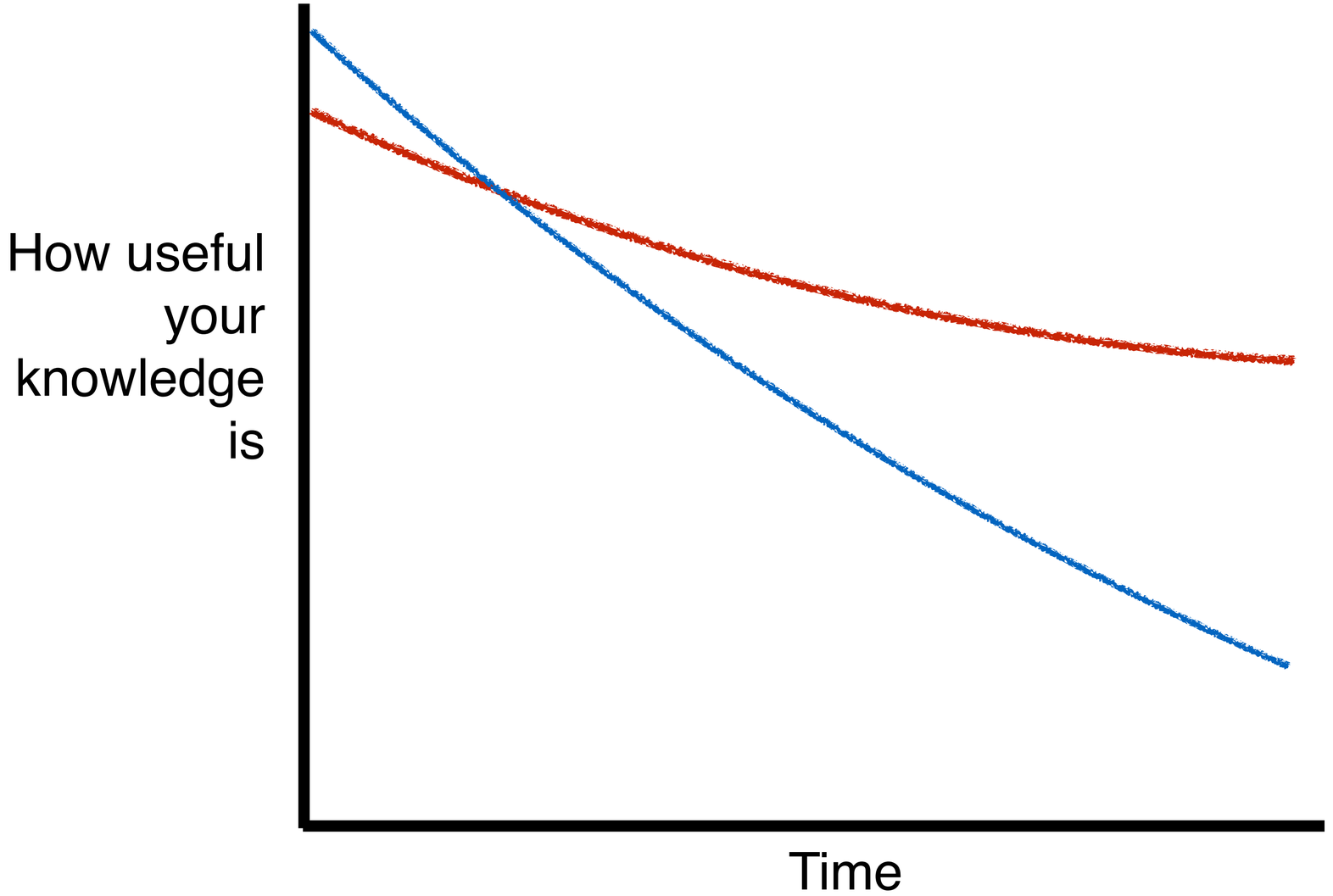


Fundamentals & problem solving skills (University Education)

Knowledge about a specific technology (Tutorials, etc.)

Half-Life of Knowledge





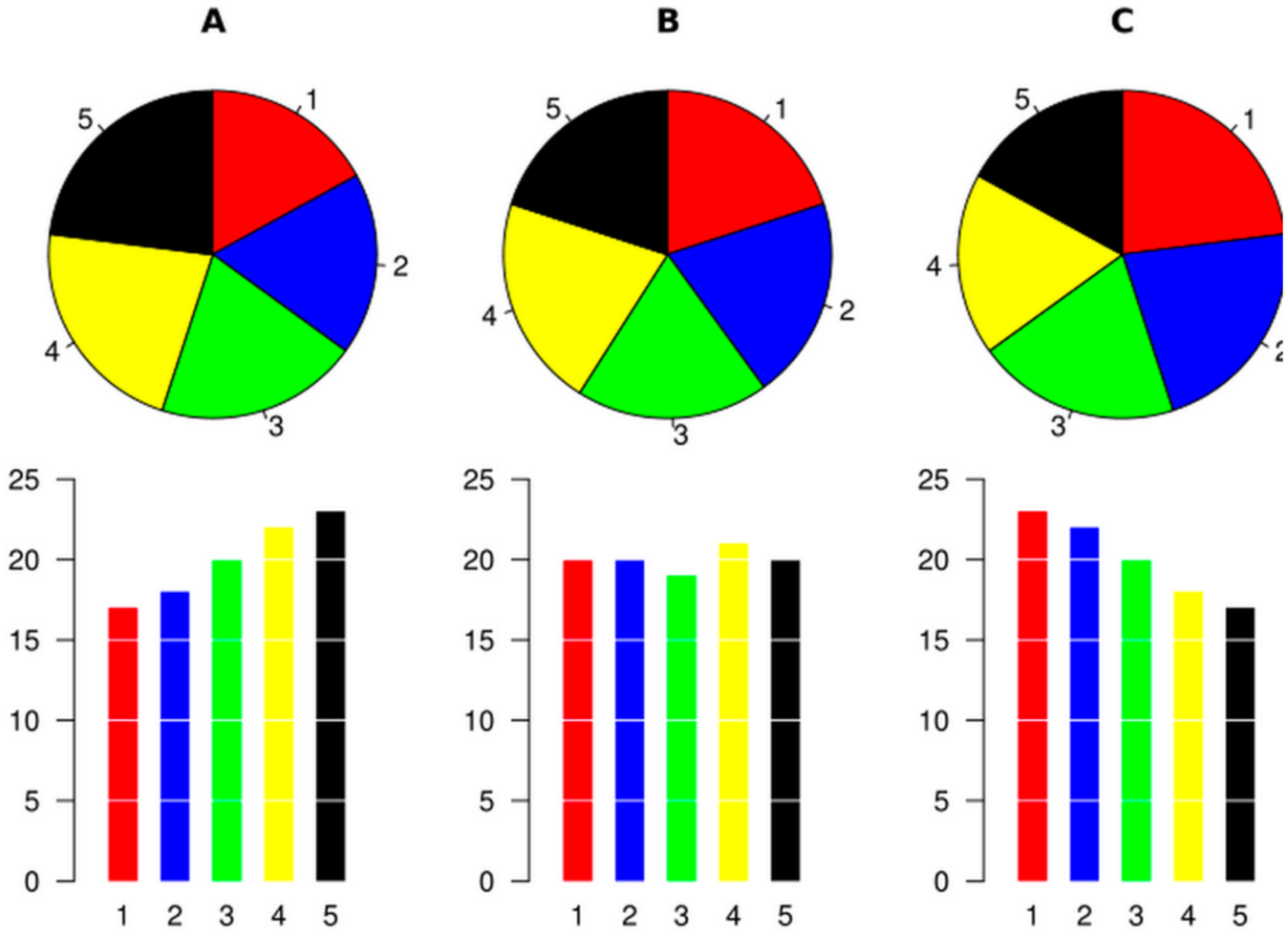
Visualization Principles and Theory

Your D3/JavaScript Ninja Skills

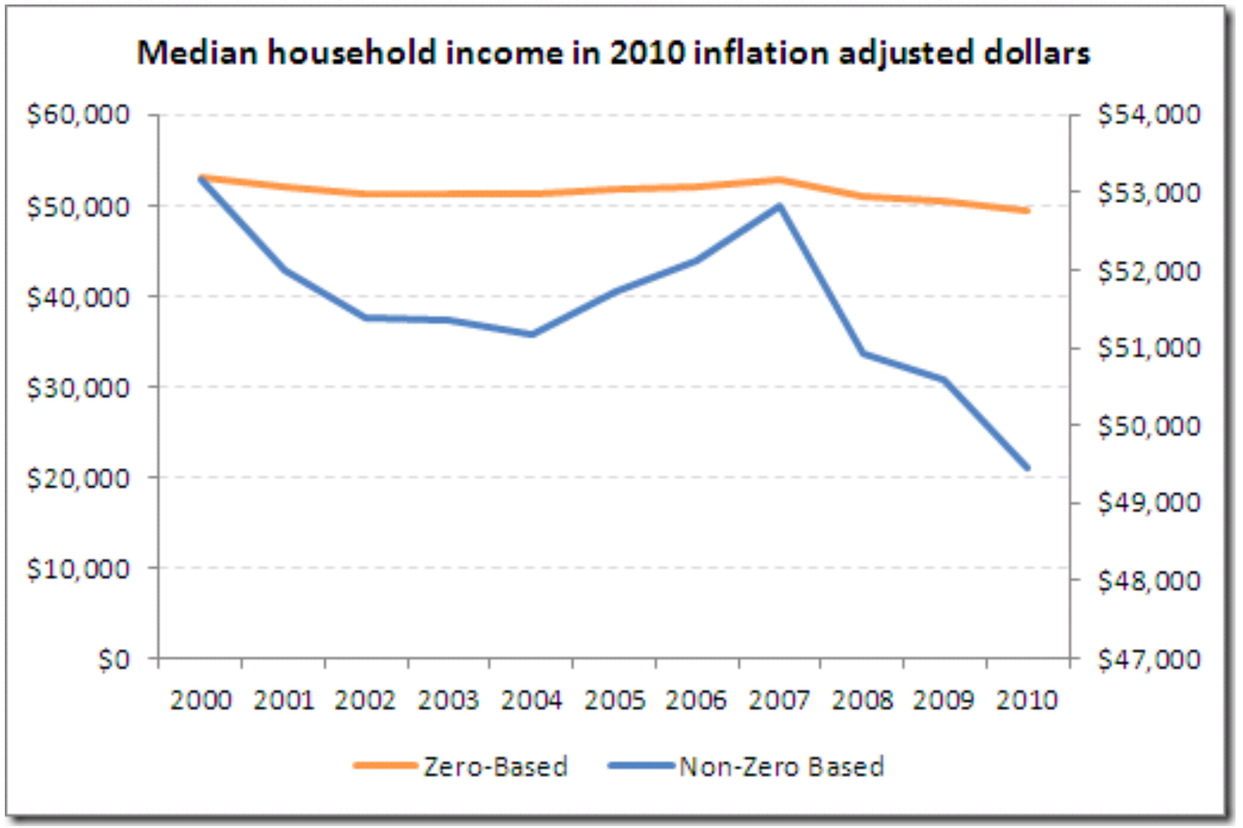
Two Weeks Ago

Vis Guidelinies
Tasks

Can you spot the differences?

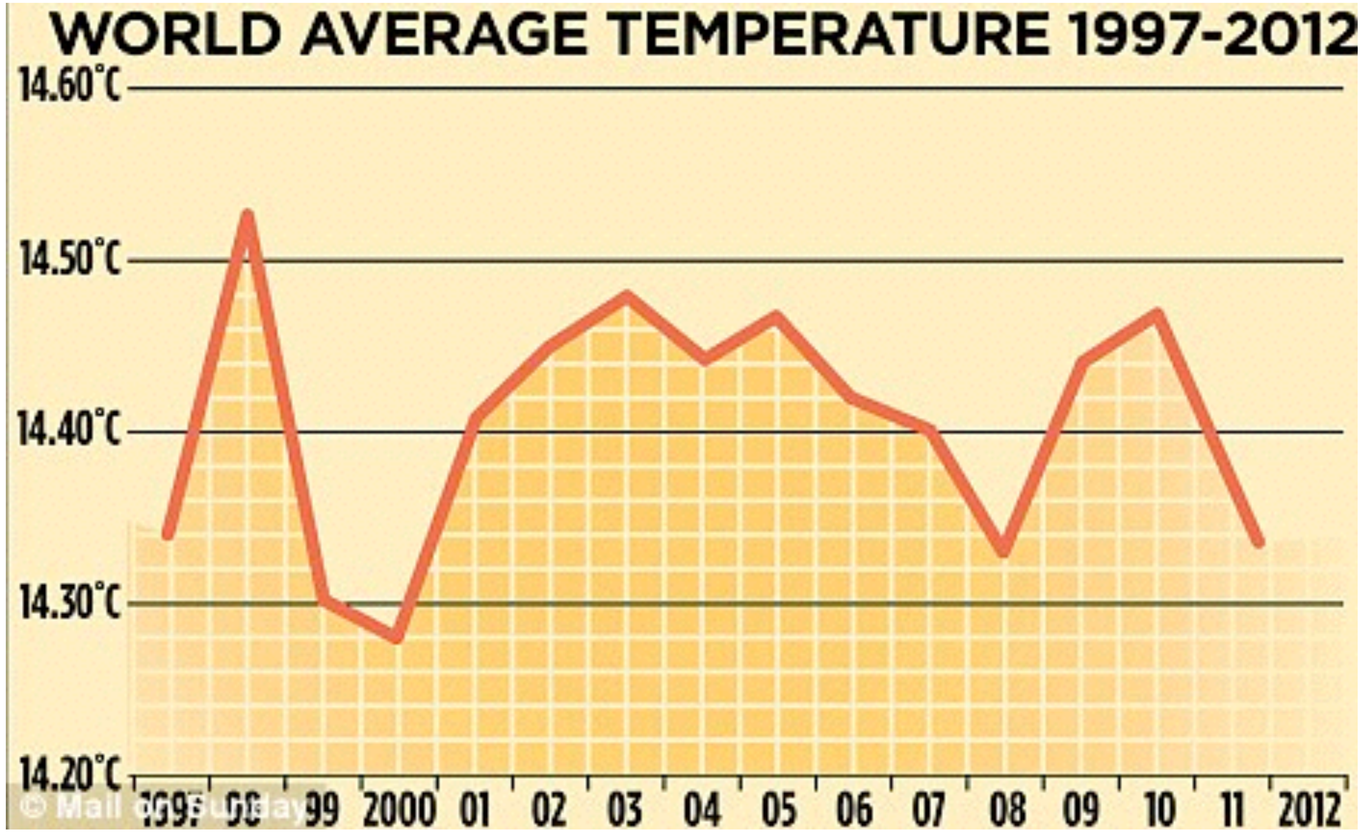


Start Scales at 0?



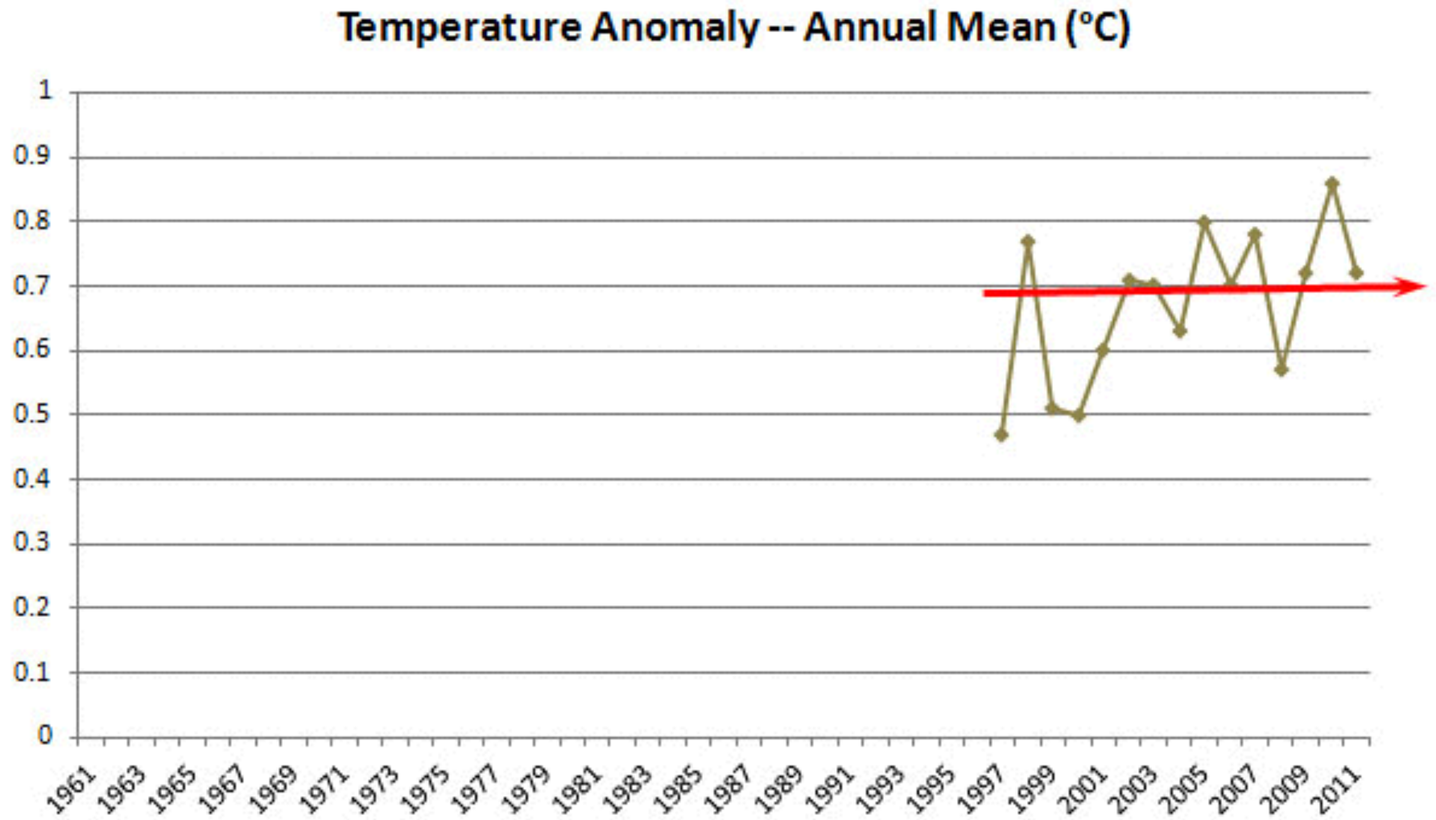
A. Kriebel, VizWiz

Global Warming?



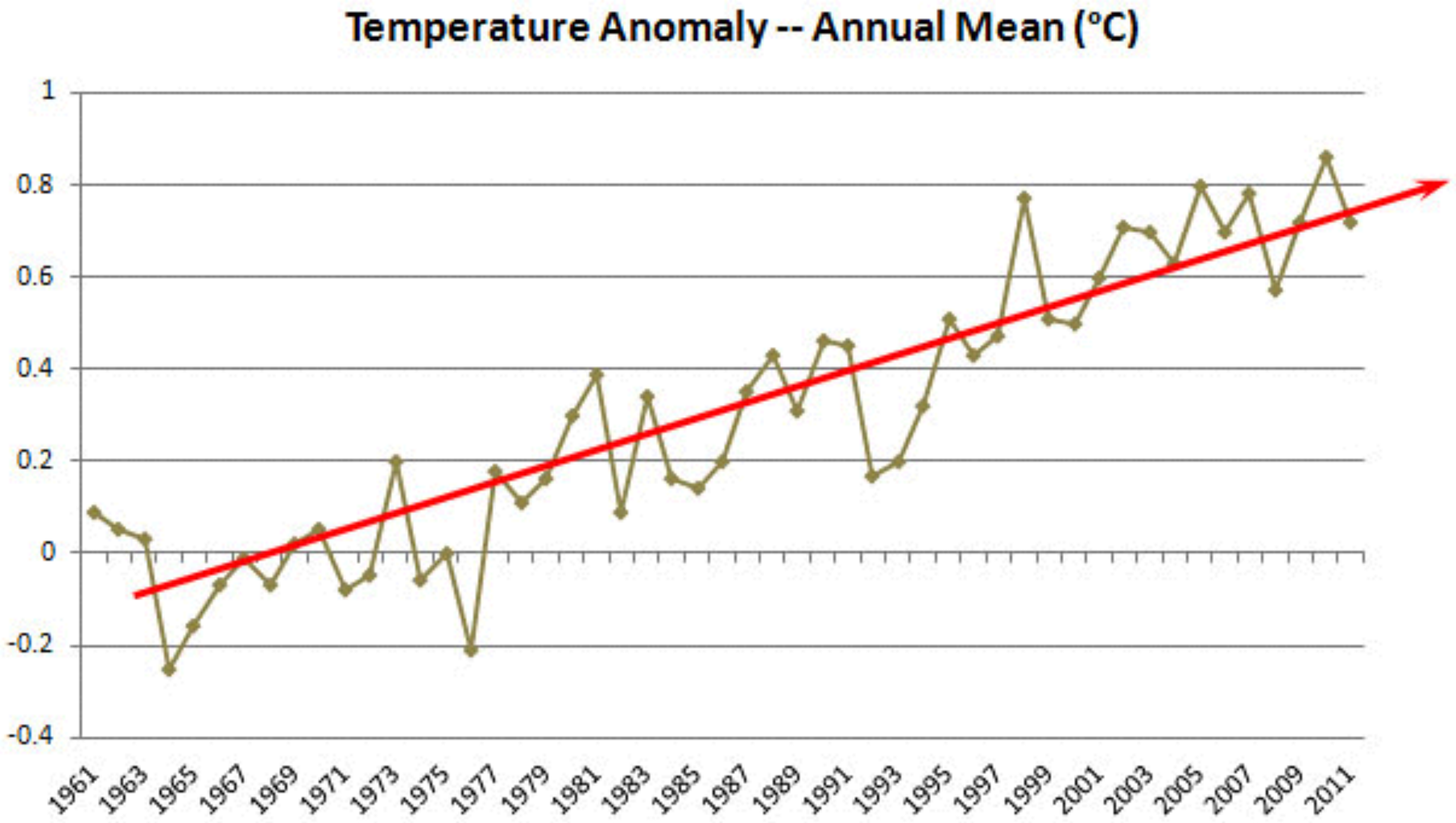
The Daily Mail, UK, Jan 2012

Global Warming?



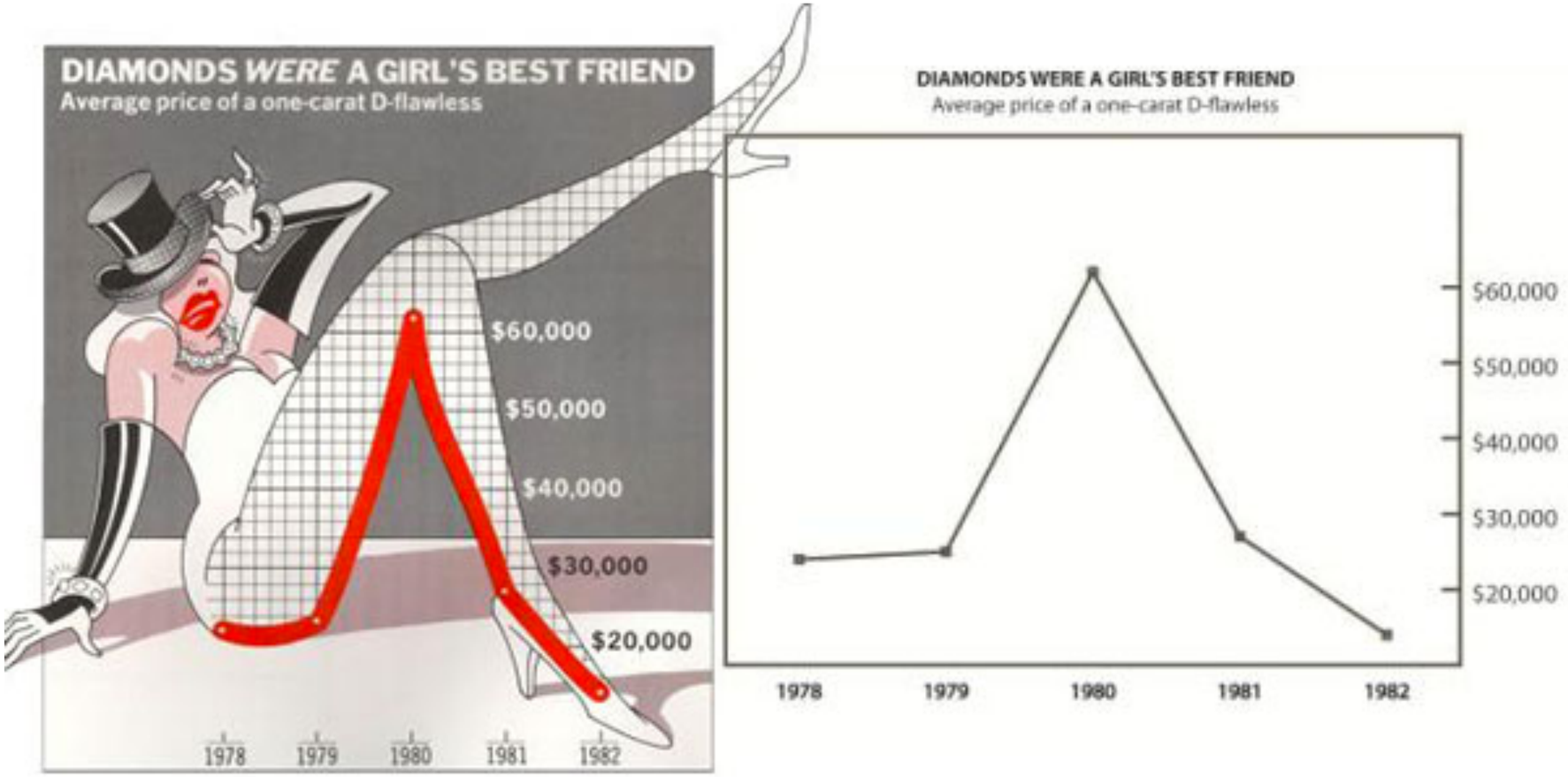
Mother Jones

Global Warming - Frame the Data



Mother Jones

Which is better?



[Bateman et al. 2010]

Tasks

Why are we using Visualization?

Domain and Abstract Tasks

Infinite numbers of domain tasks

Can be broken down into simpler abstract tasks

We know how to address the abstract tasks!

Identify task - data combination: solutions probably exist

High-level actions: Analyze

Consume

discover vs present

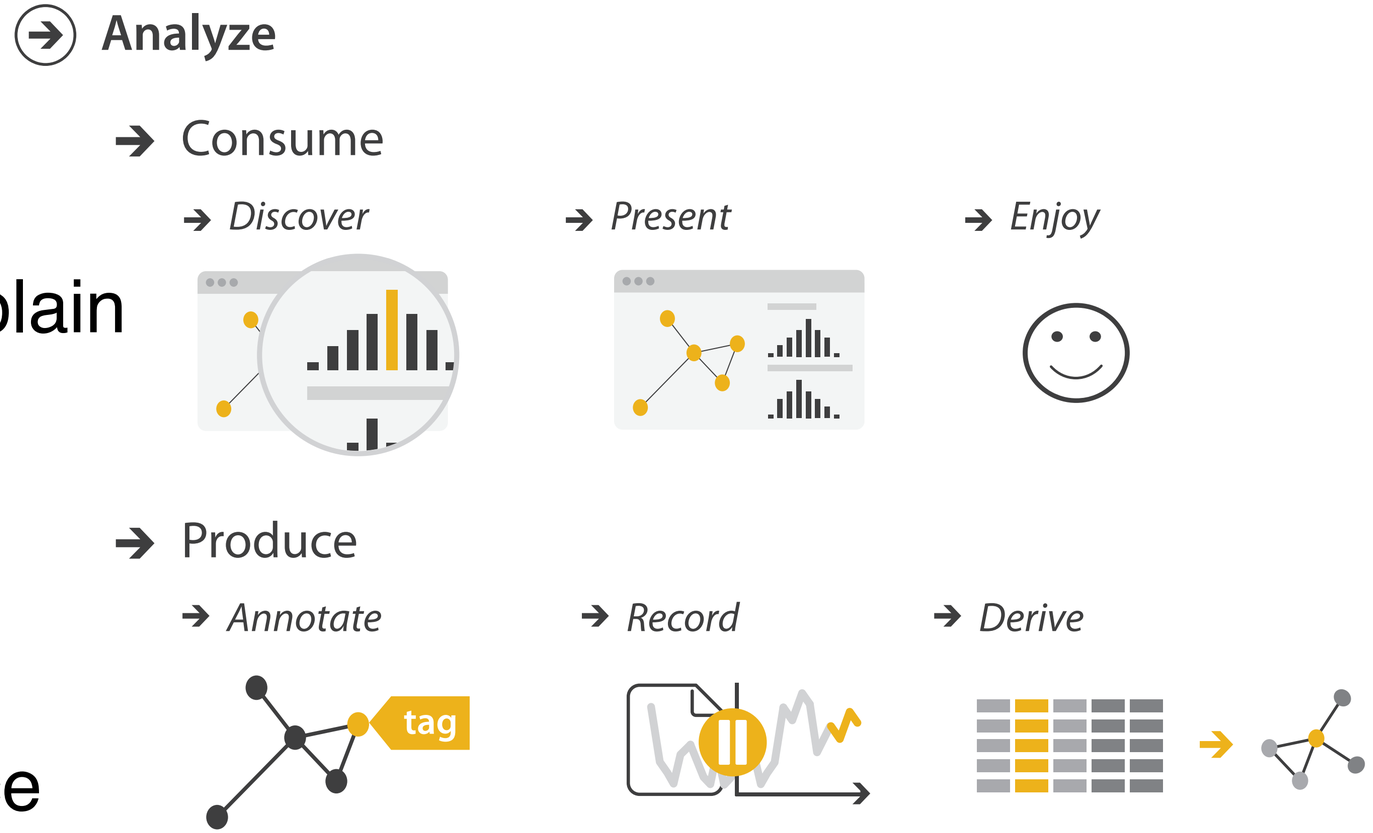
classic split: explore vs explain

enjoy: casual, social

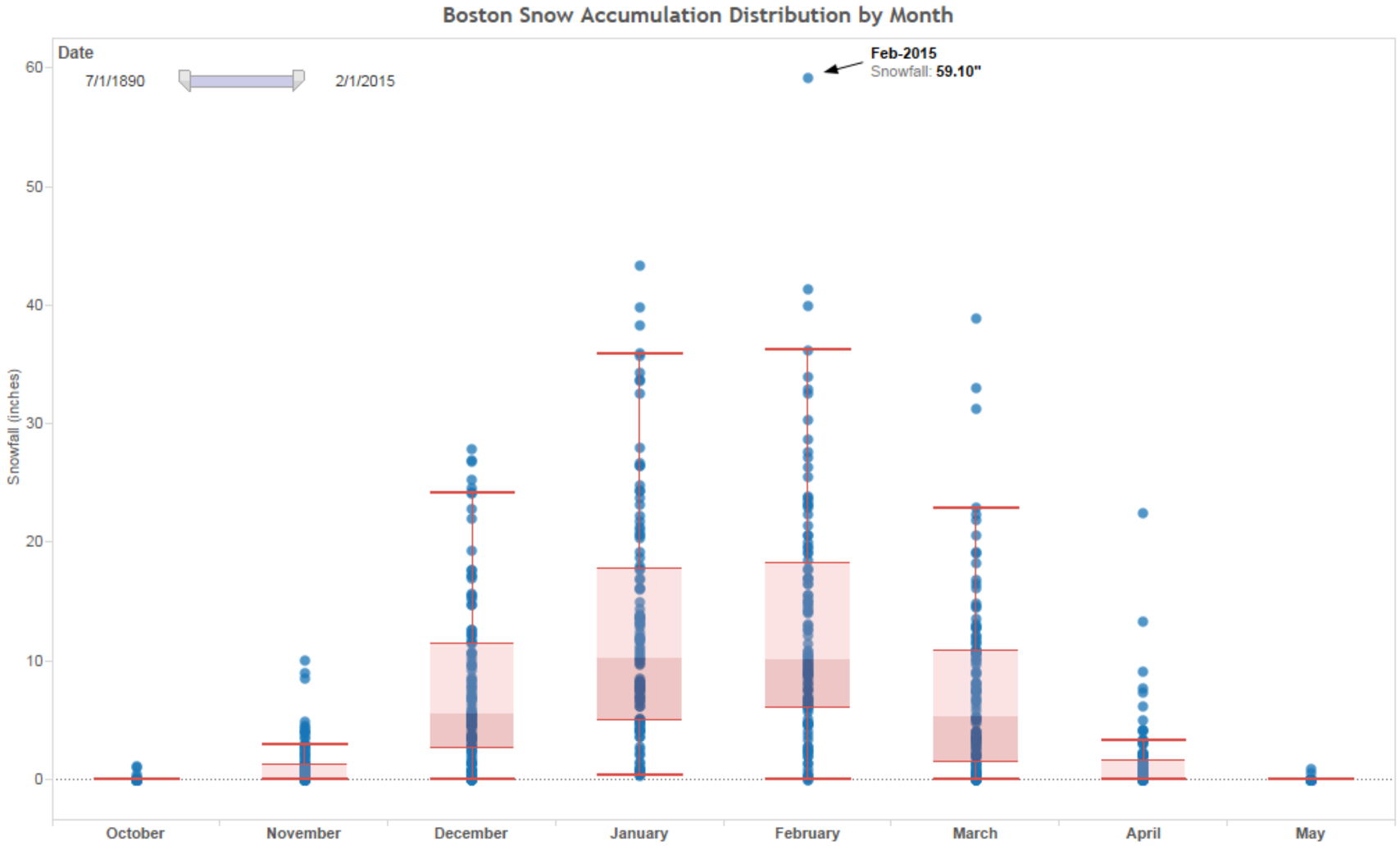
Produce

Annotate, record

Derive: crucial design choice



Example: Derive



Actions: Mid-level search, low-level query

what does user know?

target, location

how much of the data matters?

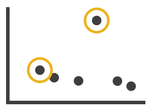
one, some, all

Search

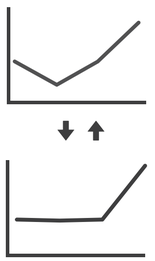
	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Query

→ Identify

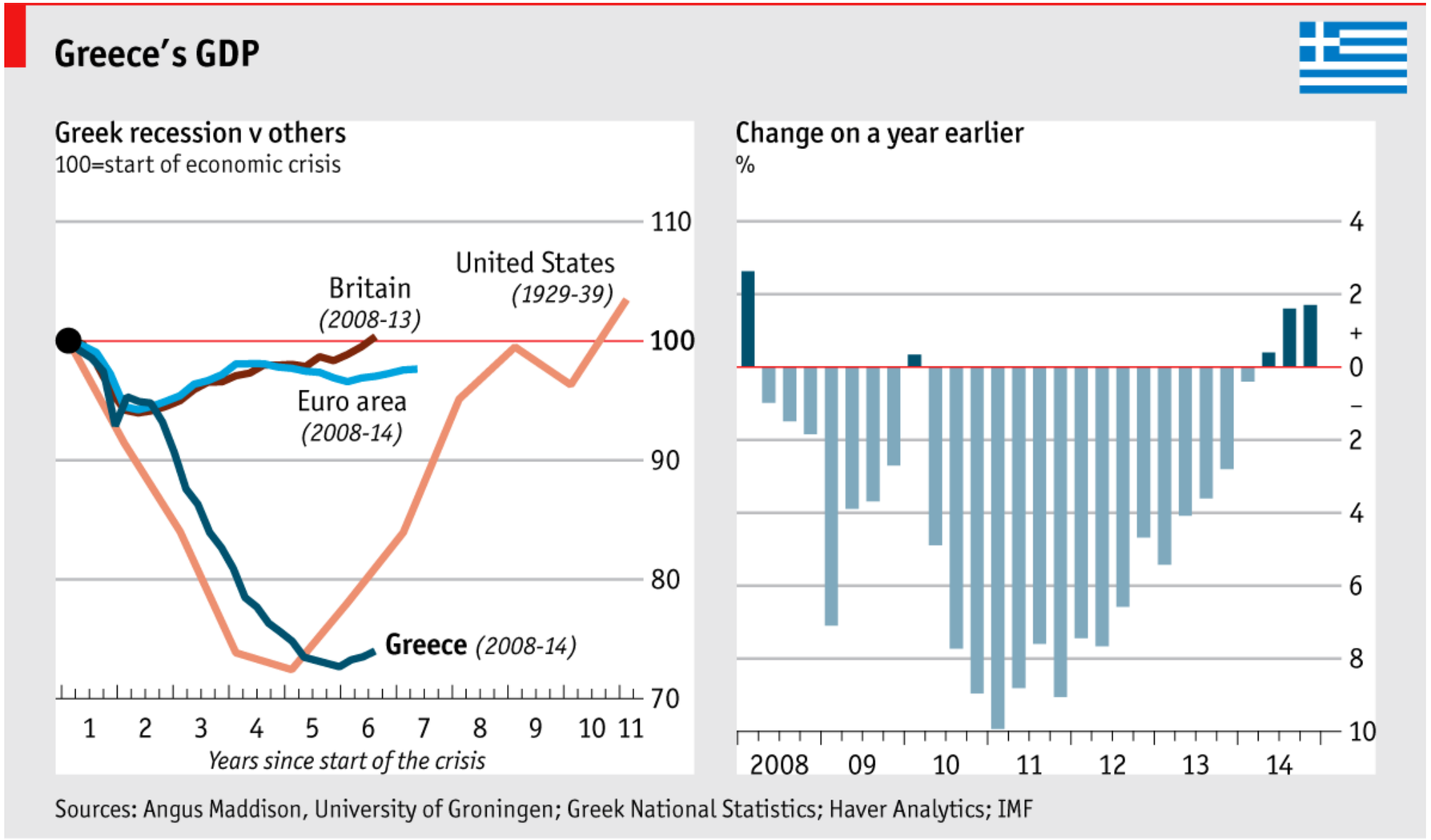


→ Compare

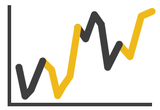
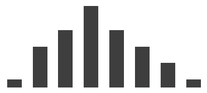
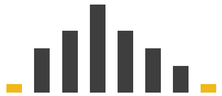
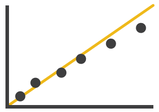
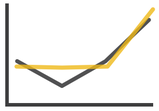
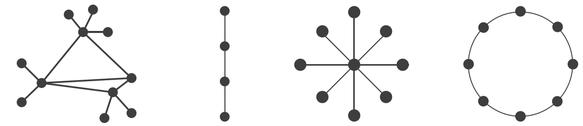
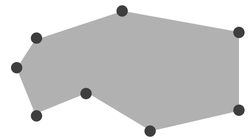


→ Summarize

Example Compare (& Derive)



Why: Targets

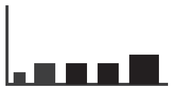
- ALL DATA
  - Trends
    
  - Outliers
    
  - Features
    
- ATTRIBUTES
  - One
    - *Distribution*
      
    - *Extremes*
      
  - Many
    - *Dependency*
      
    - *Correlation*
      
    - *Similarity*
      
- NETWORK DATA
  - Topology
    
  - *Paths*
    
- SPATIAL DATA
  - Shape
    

How? A Preview

Encode

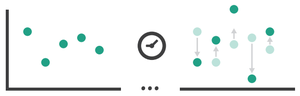
- **Arrange**
  - Express
  - Separate
  - Order
  - Align
  - Use



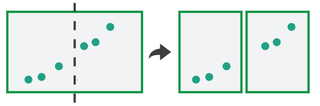

Manipulate

- **Change**
- **Select**
- **Navigate**



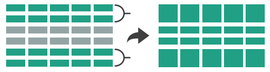
Facet

- **Juxtapose**
- **Partition**
- **Superimpose**

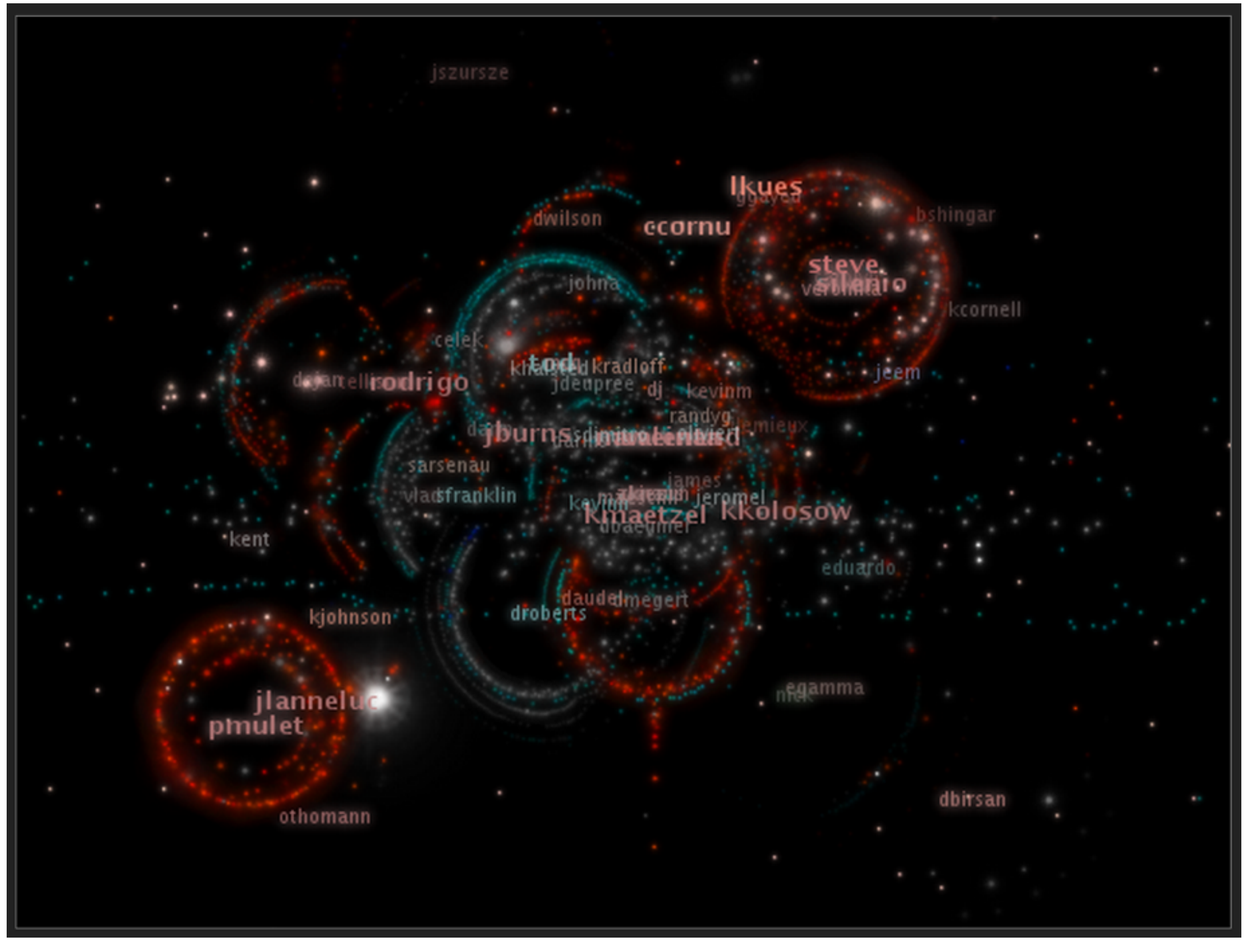

Reduce

- **Filter**
- **Aggregate**
- **Embed**


Design Critique

CodeSwarm: http://goo.gl/9exsZH



http://vis.cs.ucdavis.edu/~ogawa/codeswarm/

Tables & Multi-Dimensional Data

Basic Plots for Basic Tasks

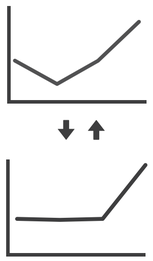
Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Query

- Identify

- Compare



- Summarize

ALL DATA

- Trends

- Outliers

- Features

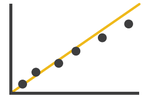
ATTRIBUTES

- One
  - *Distribution*

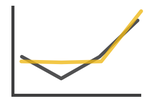
  - *Extremes*

- Many
  - *Dependency*

  - *Correlation*

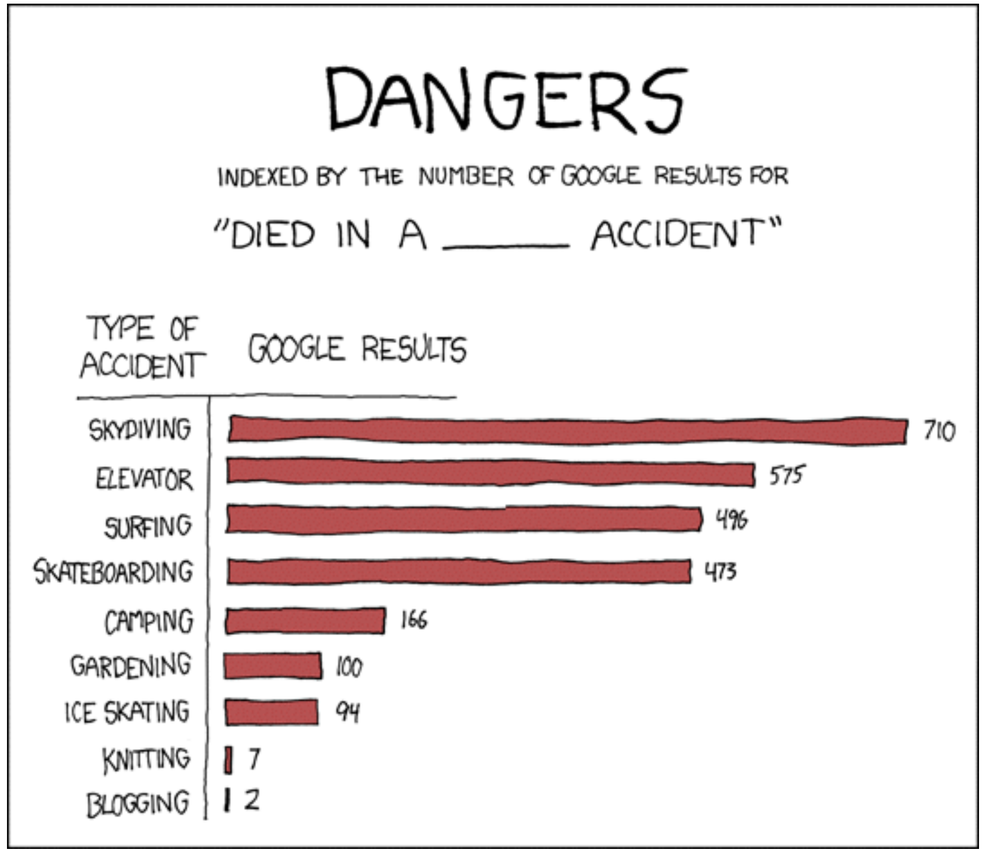


  - *Similarity*

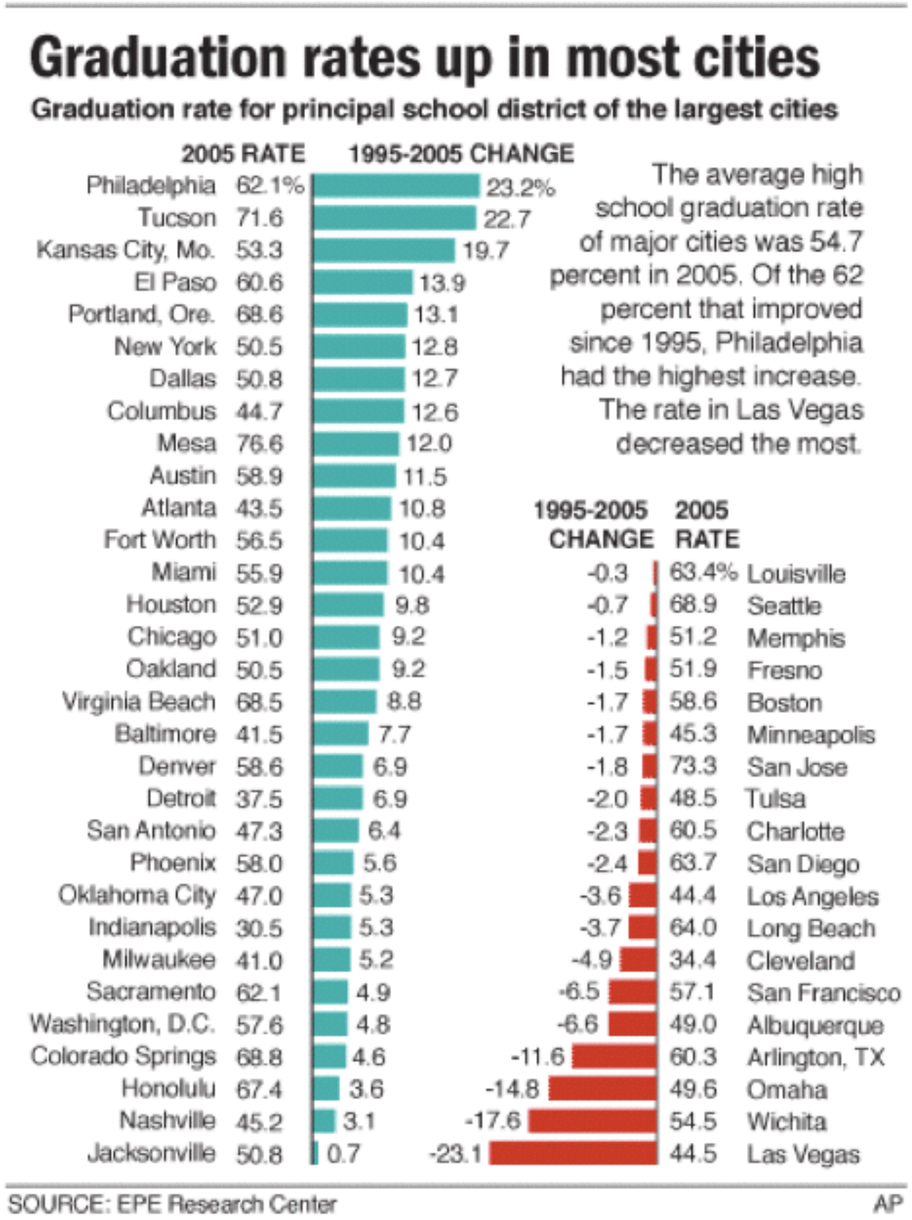


Comparisons

Bar Chart

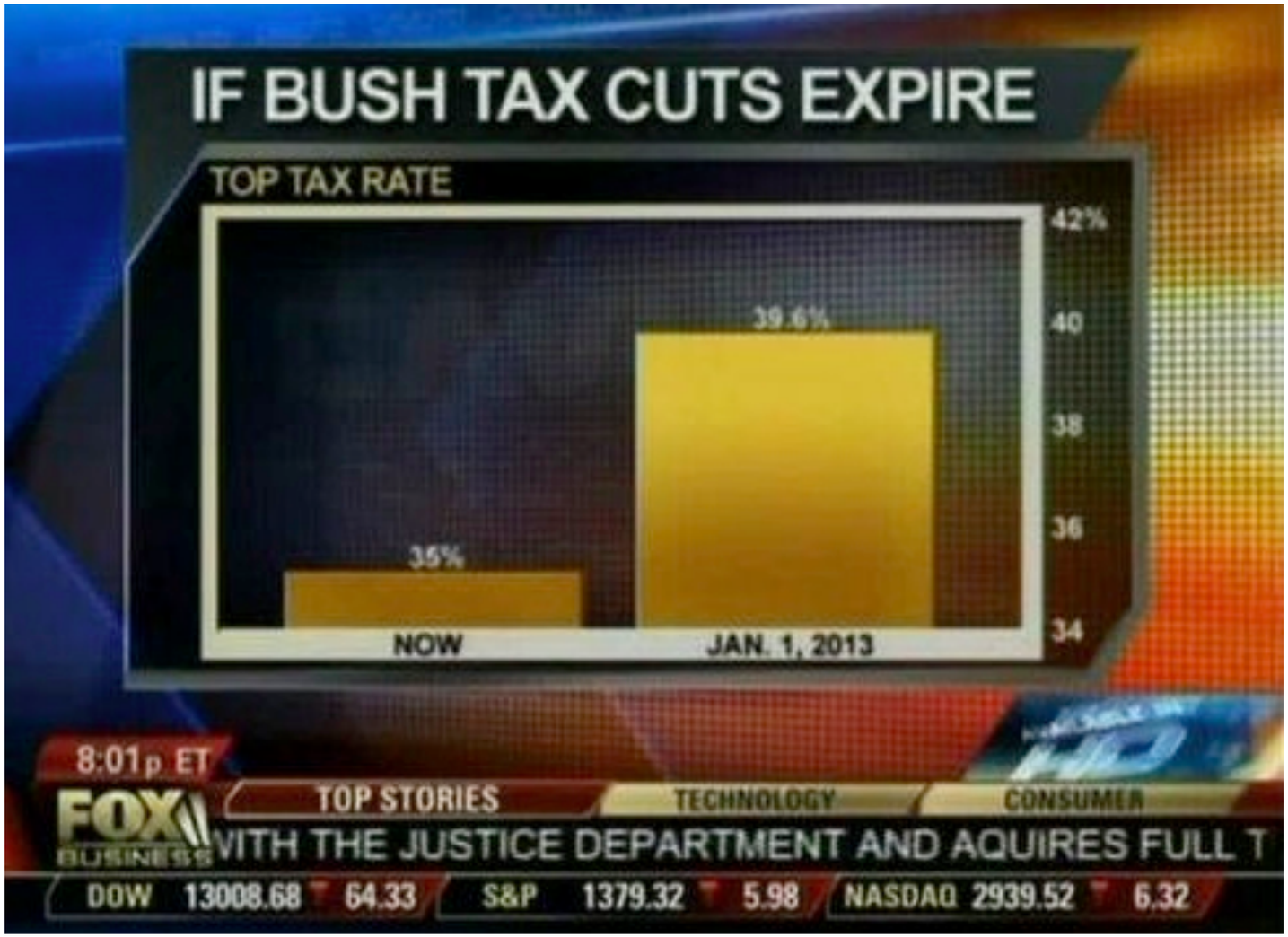

DANGERS
INDEXED BY THE NUMBER OF GOOGLE RESULTS FOR
"DIED IN A _____ ACCIDENT"
TYPE OF ACCIDENT
GOOGLE RESULTS
SKYDIVING
710
ELEVATOR
575
SURFING
496
SKATEBOARDING
473
CAMPING
166
GARDENING
100
ICE SKATING
94
KNITTING
7
BLOGGING
2

Direction



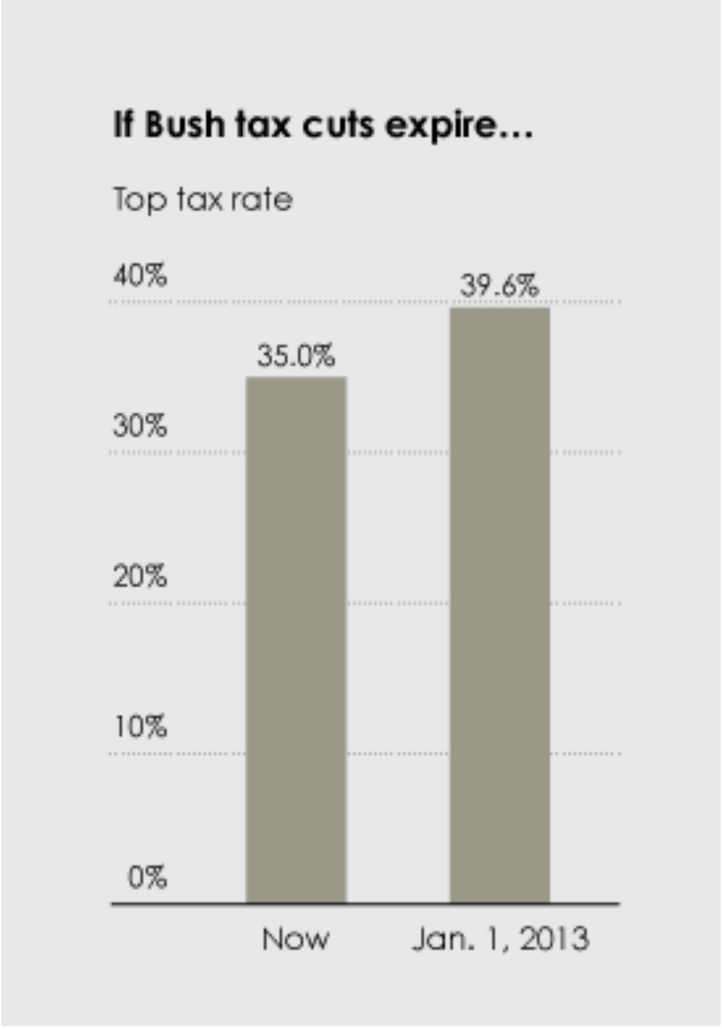
Nicolas Rapp

Baseline Problem



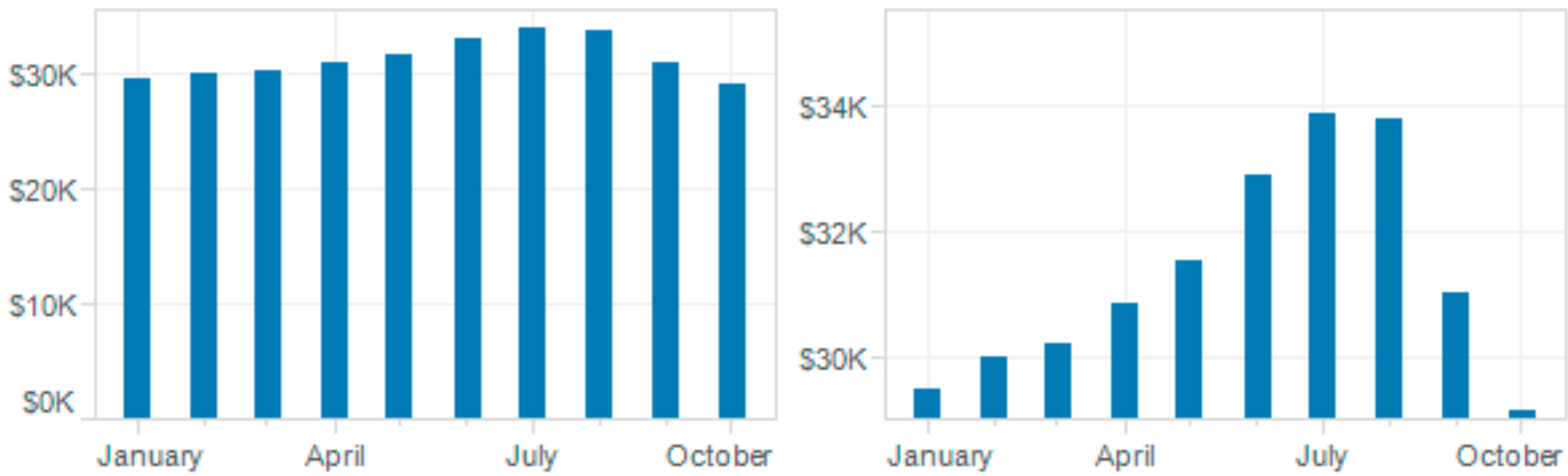
Flowing Data

Baseline Problem



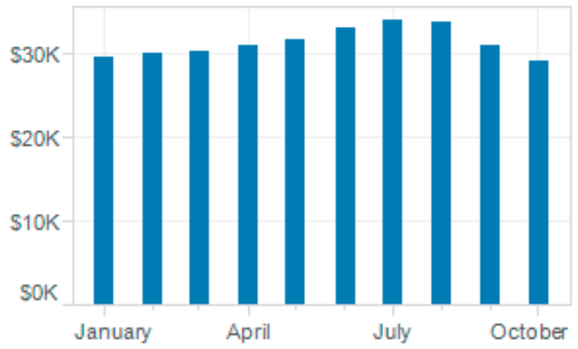
Flowing Data

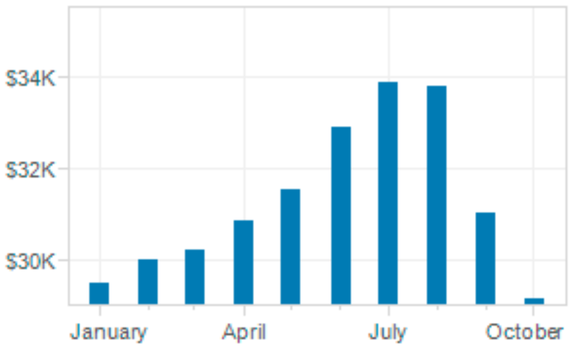
Different Baselines

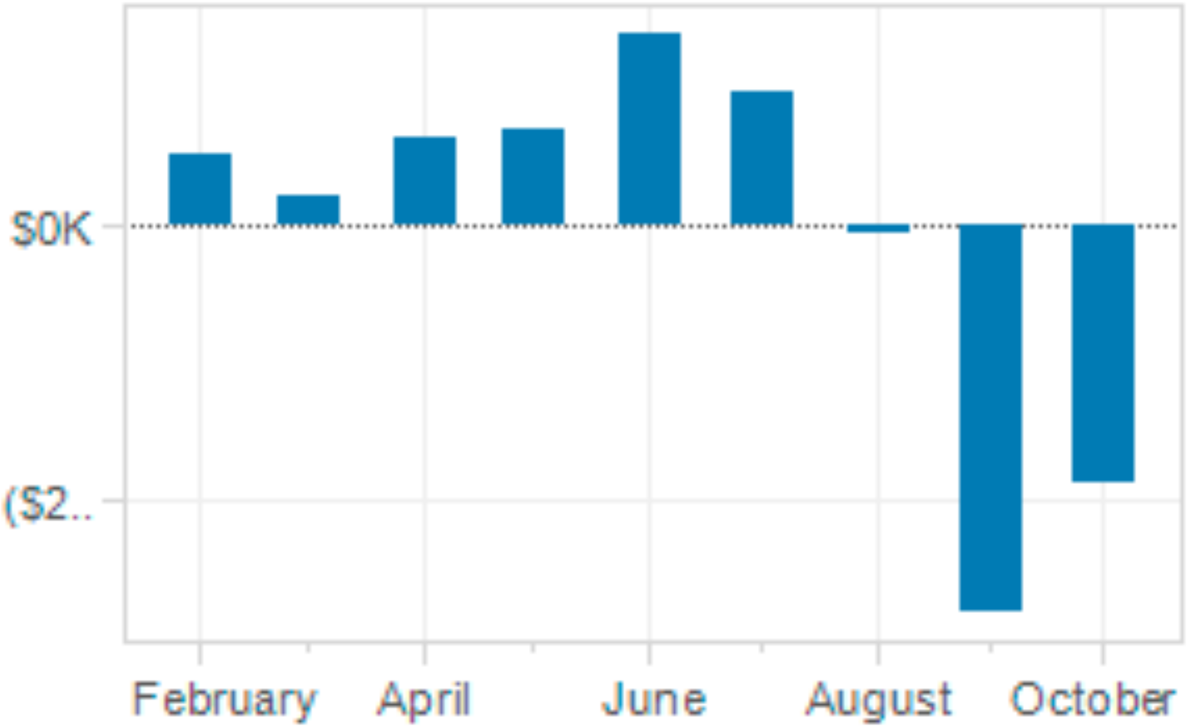


https://eagereyes.org/basics/baselines

Plot Change Instead

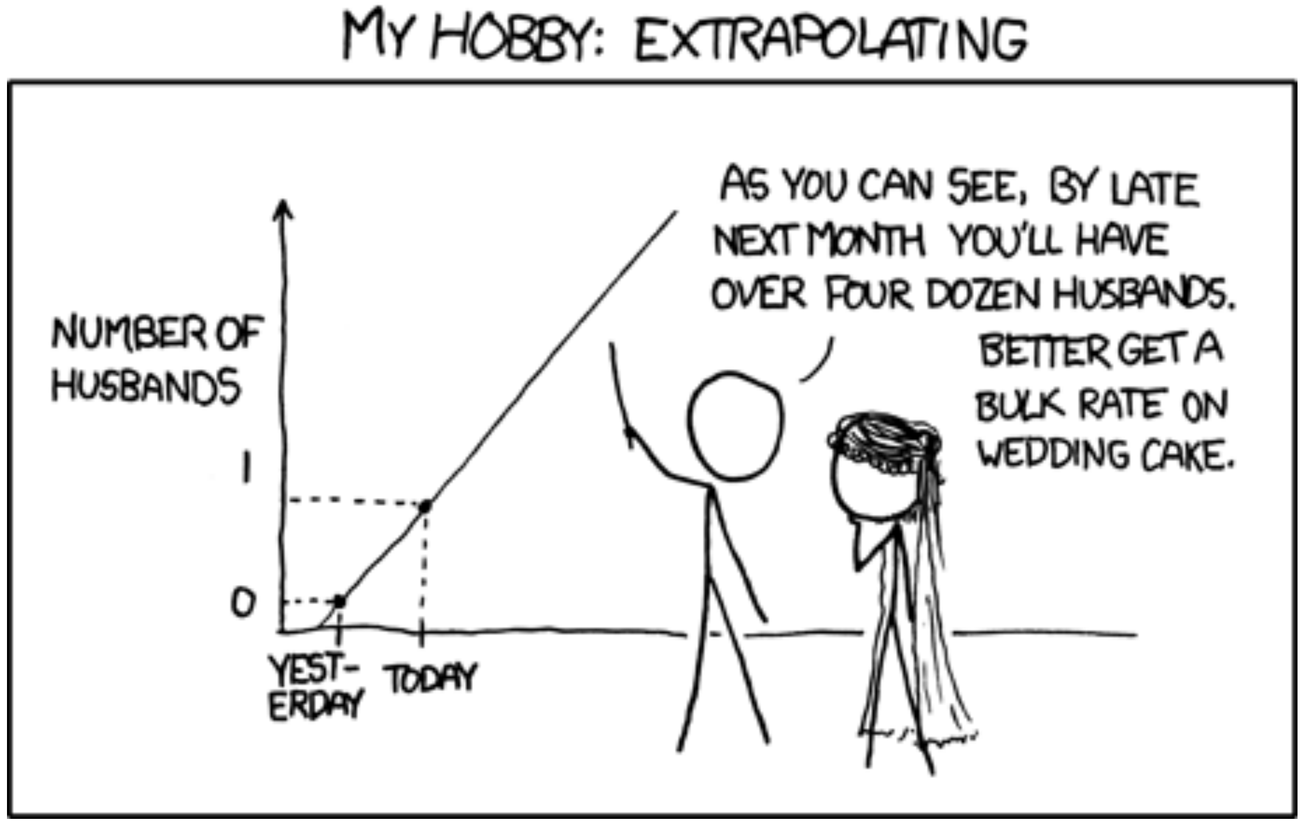






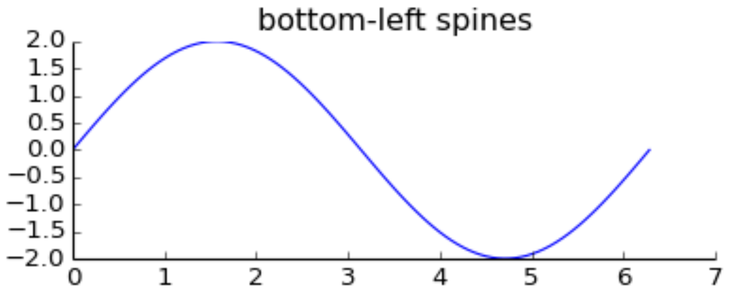
https://eagereyes.org/basics/baselines

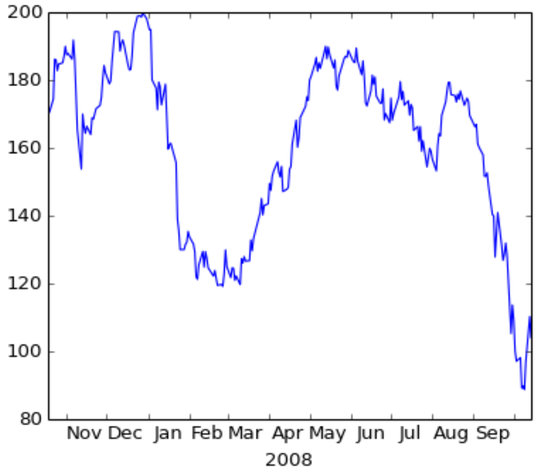
Trends Over Time

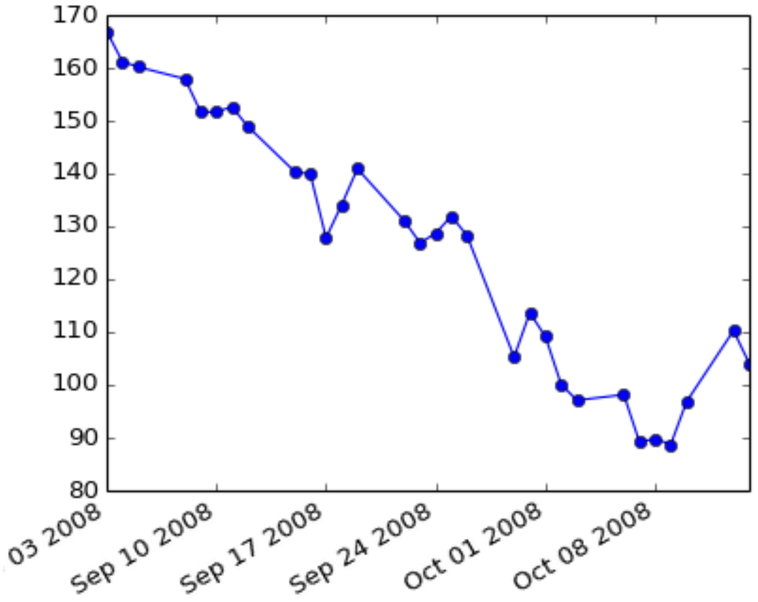


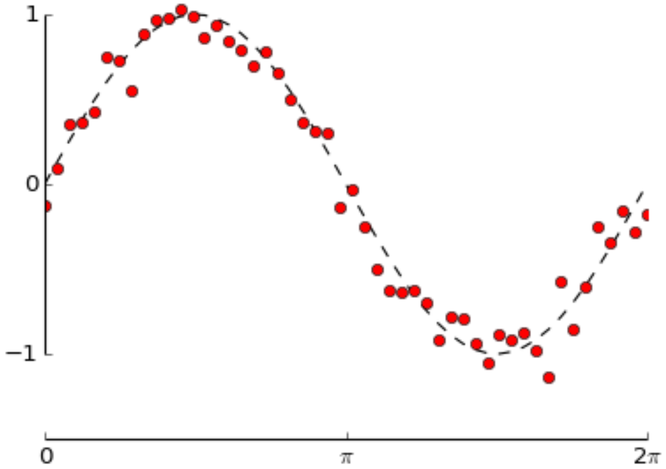
http://xkcd.com/605/

Line Charts





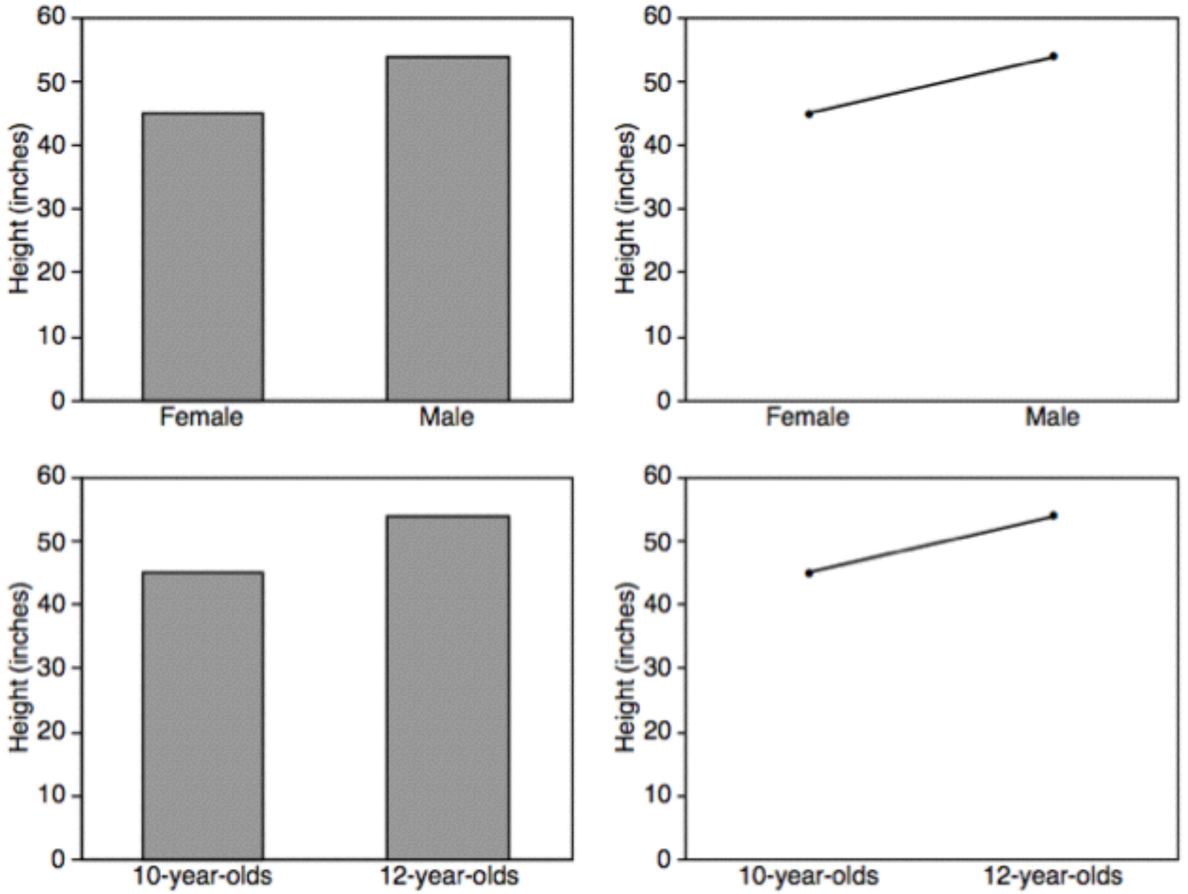




matplotlib gallery

Bars vs. Lines

Lines imply connections & sampling from continuous data.
Do not use for categorical data.

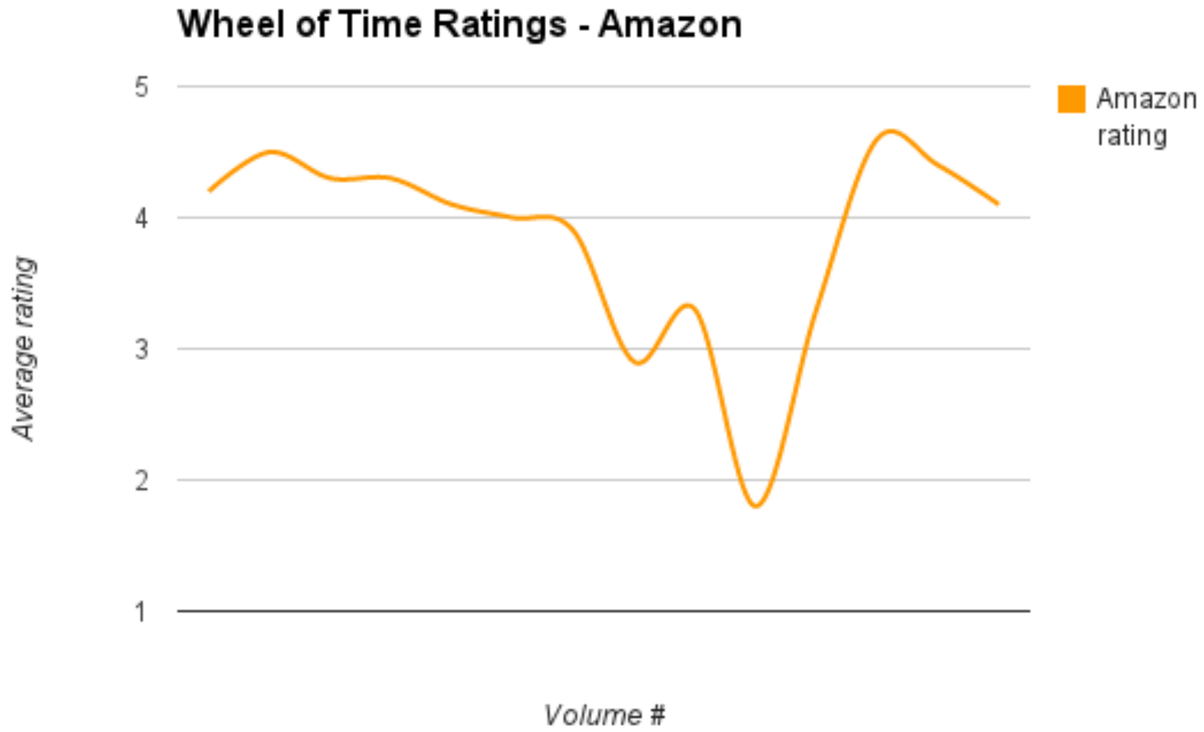


Zacks 1999

Don’t

Use bar charts to

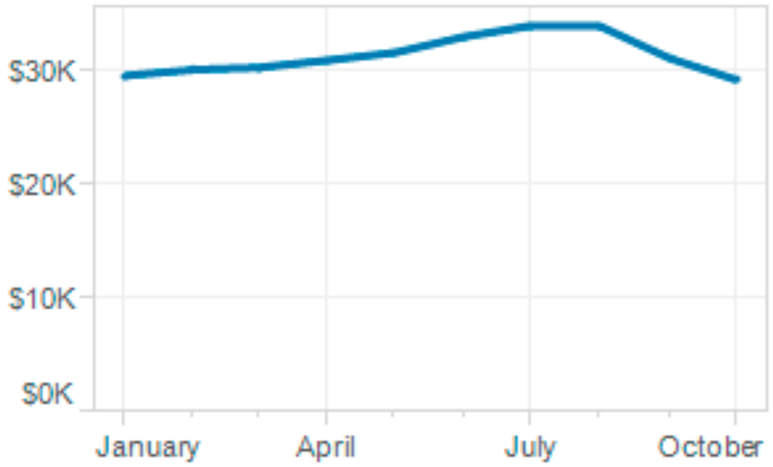
compare ratings of books…



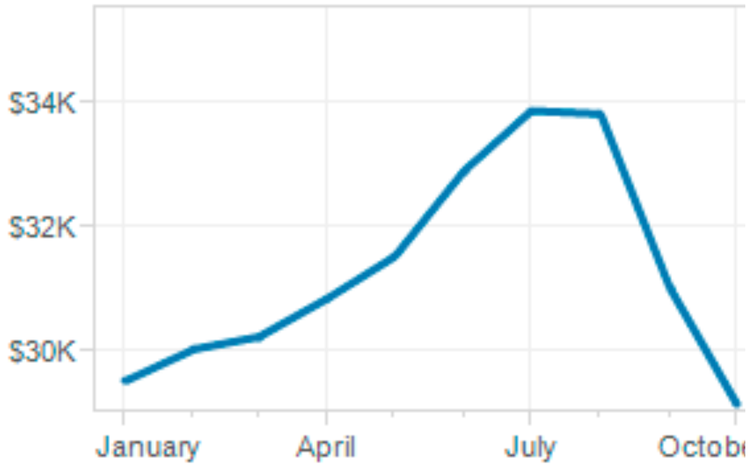
“Visualizing The Wheel of Time: Reader Sentiment for an Epic Fantasy Series”, J. Siddle, Sept 2013

Baseline Problem (again)

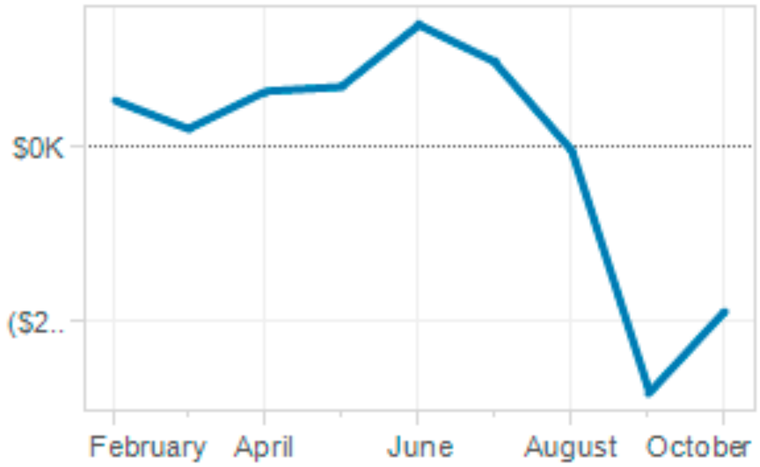
True Baseline



Clipped Baseline

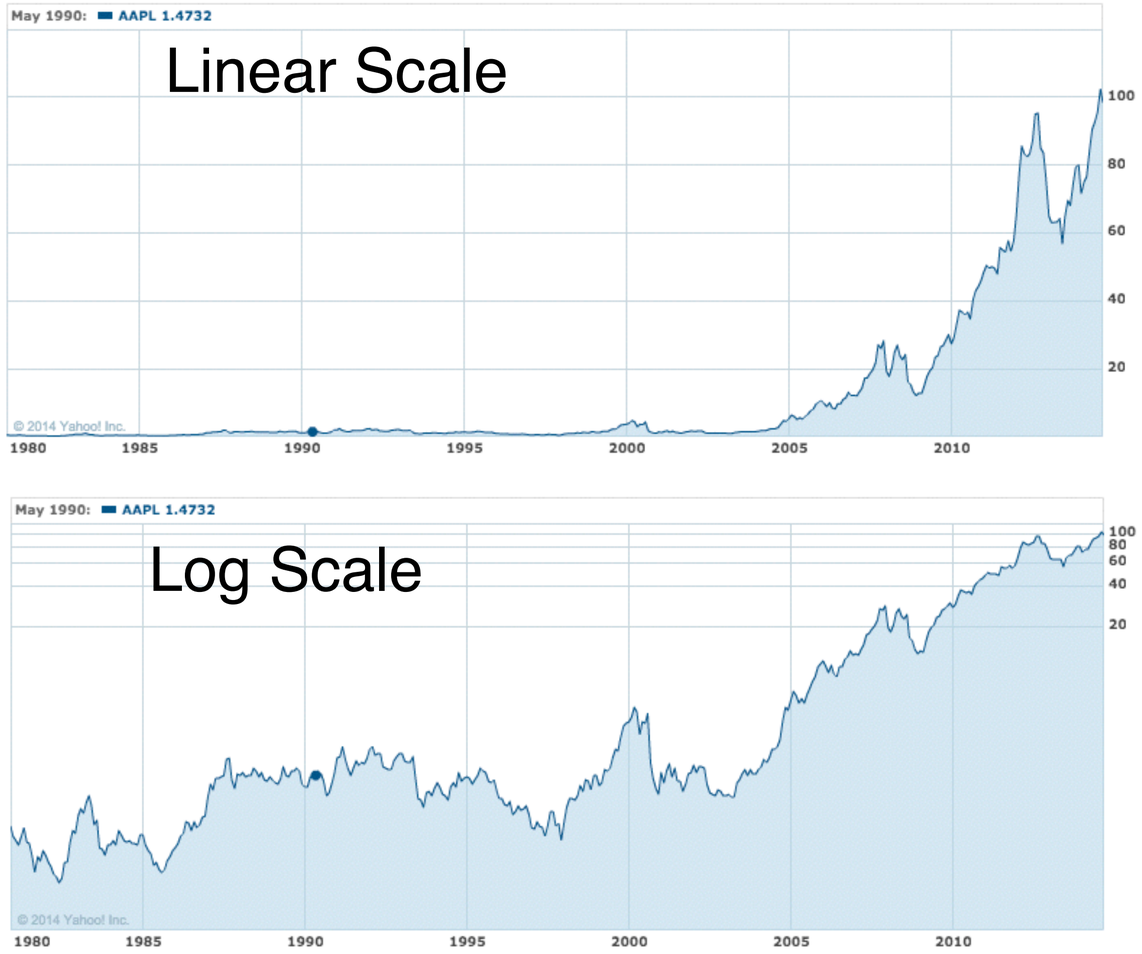


Plotting Change



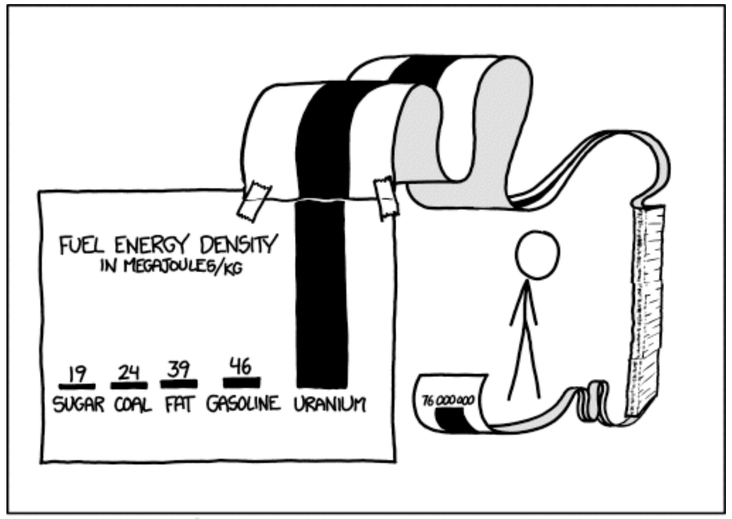
https://eagereyes.org/basics/baselines

Linear vs. Logarithmic Scale



Apple Stock Price

http://finance.yahoo.com/echarts?s=AAPL



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

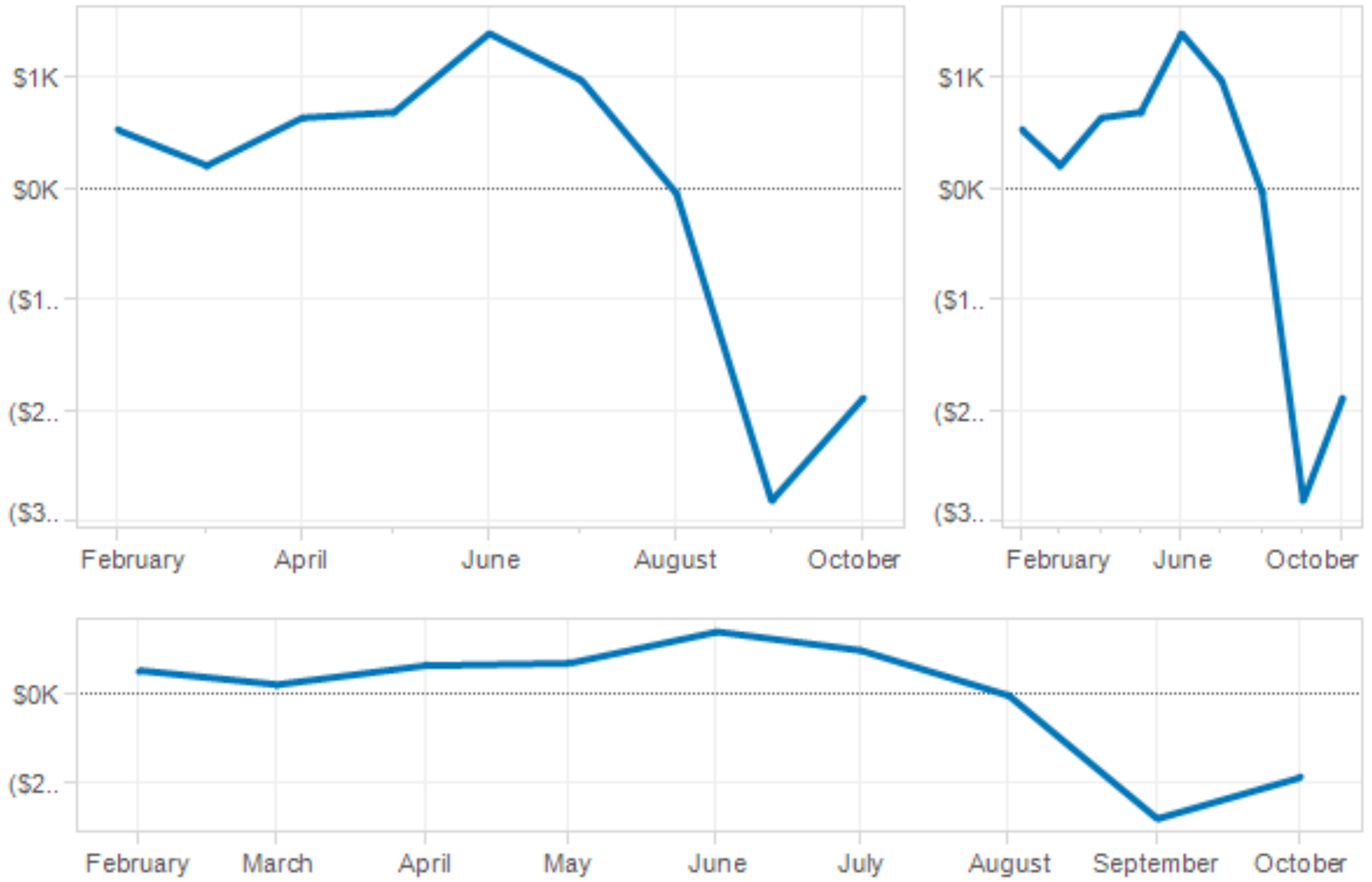
http://xkcd.com/1162/

Aspect Ratios

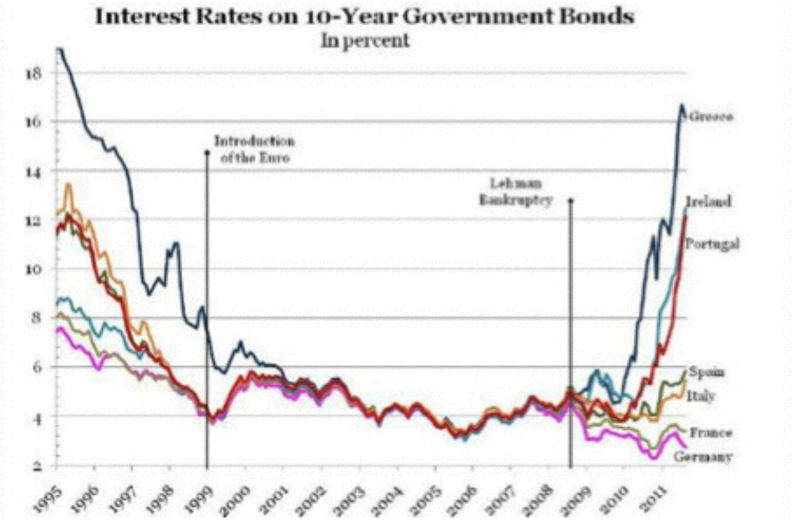
Rule of Thumb:

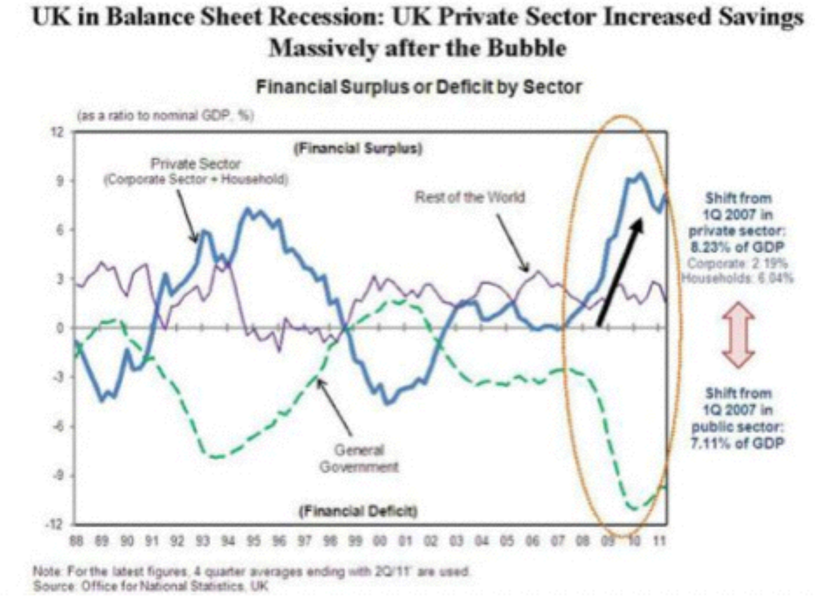
Banking to 45°

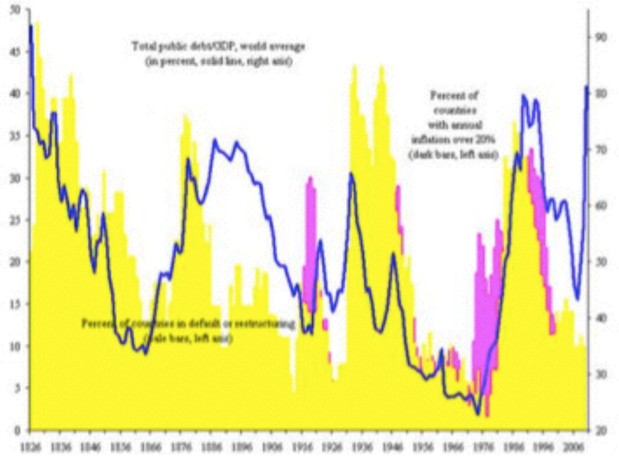
(average line slope: 45°)



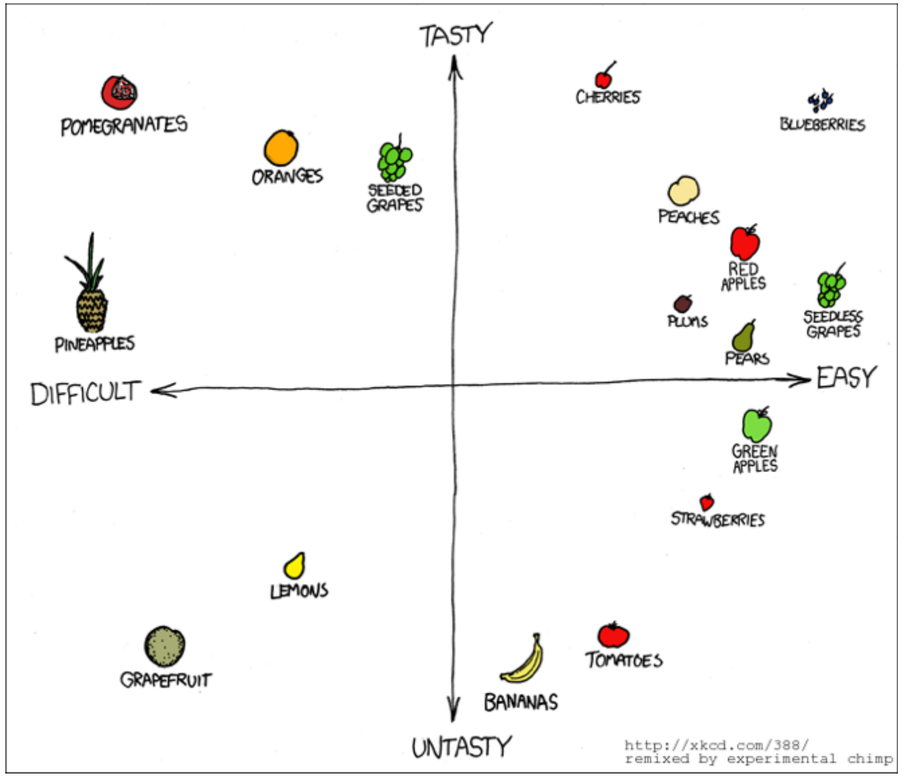
Don’t


Interest Rates on 10-Year Government Bonds
In percent
Introduction of the Euro
Lehman Bankruptcy
Greece
Ireland
Portugal
Spain
Italy
France
Germany


UK in Balance Sheet Recession: UK Private Sector Increased Savings Massively after the Bubble
Financial Surplus or Deficit by Sector
(as a ratio to nominal GDP, %)
(Financial Surplus)
Private Sector
(Corporate Sector + Household)
Rest of the World
General Government
(Financial Deficit)
Shift from 1Q 2007 in private sector: 8.23% of GDP
Corporate: 2.19%
Households: 6.04%
Shift from 1Q 2007 in public sector: 7.11% of GDP
Note: For the latest figures, 4 quarter averages ending with 2Q'11' are used.
Source: Office for National Statistics, UK

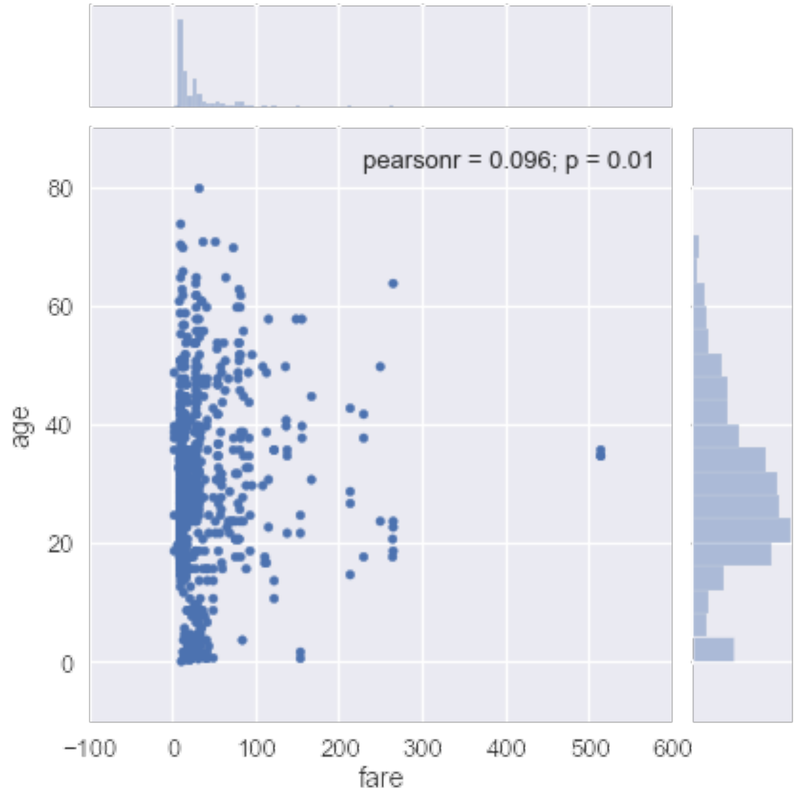

Total public debt/GDP, world average (in percent, solid line, right axis)
Percent of countries with annual inflation over 20% (dark bars, left axis)

Correlations



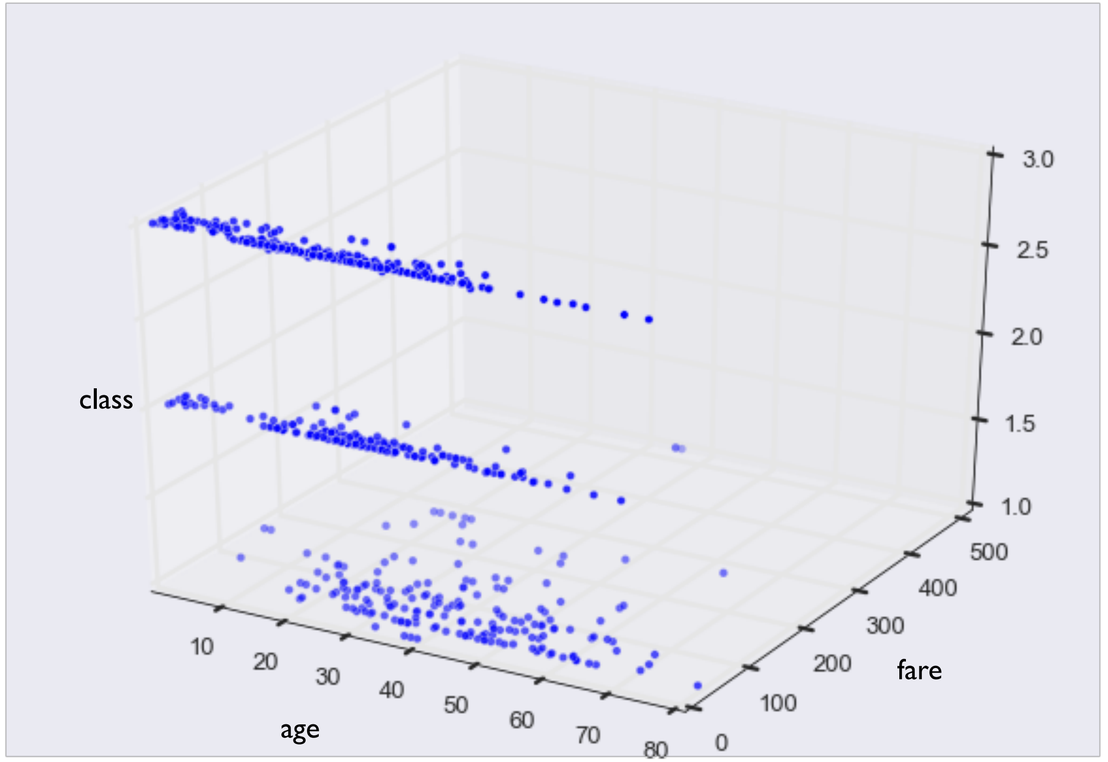
age	fare
22.0	7.25
38.0	71.2833
26.0	7.925
35.0	53.1
35.0	8.05
	8.4583
54.0	51.8625
2.0	21.075
27.0	11.1333
14.0	30.0708
4.0	16.7
58.0	26.55
20.0	8.05
39.0	31.275
14.0	7.8542
55.0	16.0
2.0	29.125
	13.0
31.0	18.0
	7.225
35.0	26.0
34.0	13.0
15.0	8.0292

Scatterplots



age	fare	class
22.0	7.25	Third
38.0	71.2833	First
26.0	7.925	Third
35.0	53.1	First
35.0	8.05	Third
	8.4583	Third
54.0	51.8625	First
2.0	21.075	Third
27.0	11.1333	Third
14.0	30.0708	Second
4.0	16.7	Third
58.0	26.55	First
20.0	8.05	Third
39.0	31.275	Third
14.0	7.8542	Third
55.0	16.0	Second
2.0	29.125	Third
	13.0	Second
31.0	18.0	Third
	7.225	Third
35.0	26.0	Second
34.0	13.0	Second
15.0	8.0292	Third

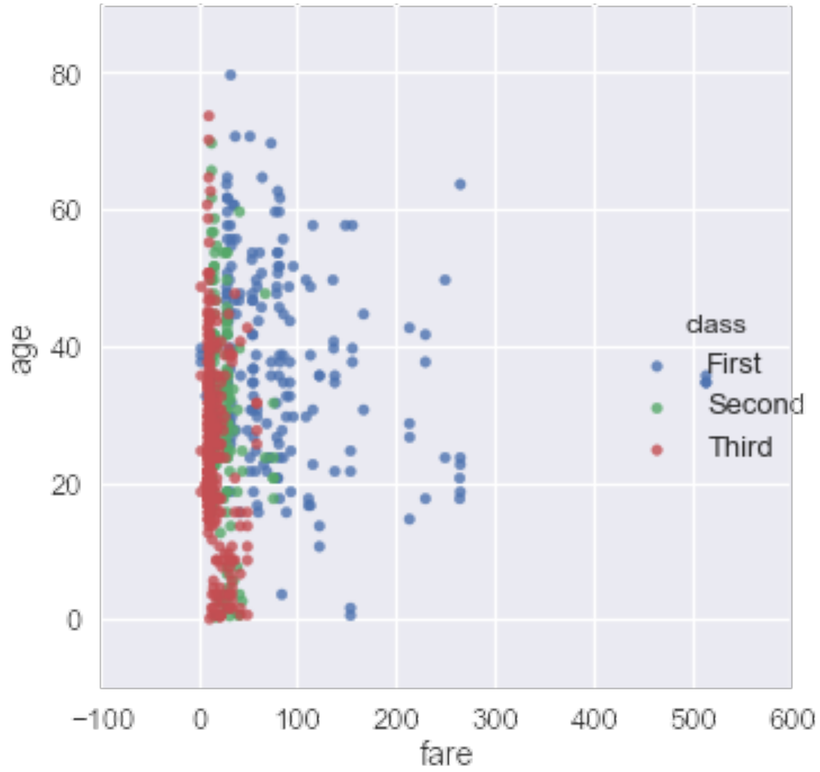
Trivariate Data



Do NOT use 3D scatterplots!

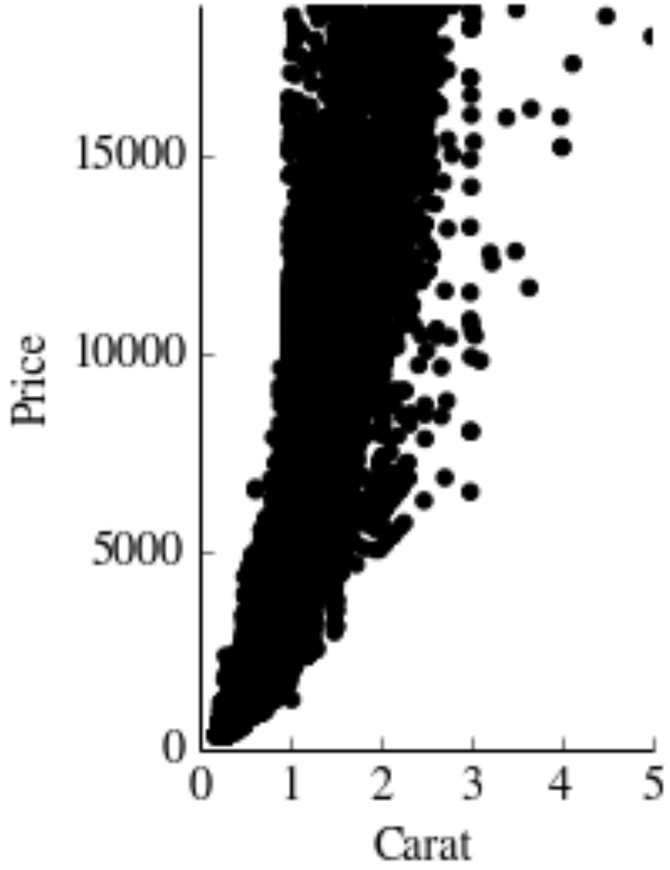
age	fare	class
22.0	7.25	Third
38.0	71.2833	First
26.0	7.925	Third
35.0	53.1	First
35.0	8.05	Third
	8.4583	Third
54.0	51.8625	First
2.0	21.075	Third
27.0	11.1333	Third
14.0	30.0708	Second
4.0	16.7	Third
58.0	26.55	First
20.0	8.05	Third
39.0	31.275	Third
14.0	7.8542	Third
55.0	16.0	Second
2.0	29.125	Third
	13.0	Second
31.0	18.0	Third
	7.225	Third
35.0	26.0	Second
34.0	13.0	Second
15.0	8.0292	Third

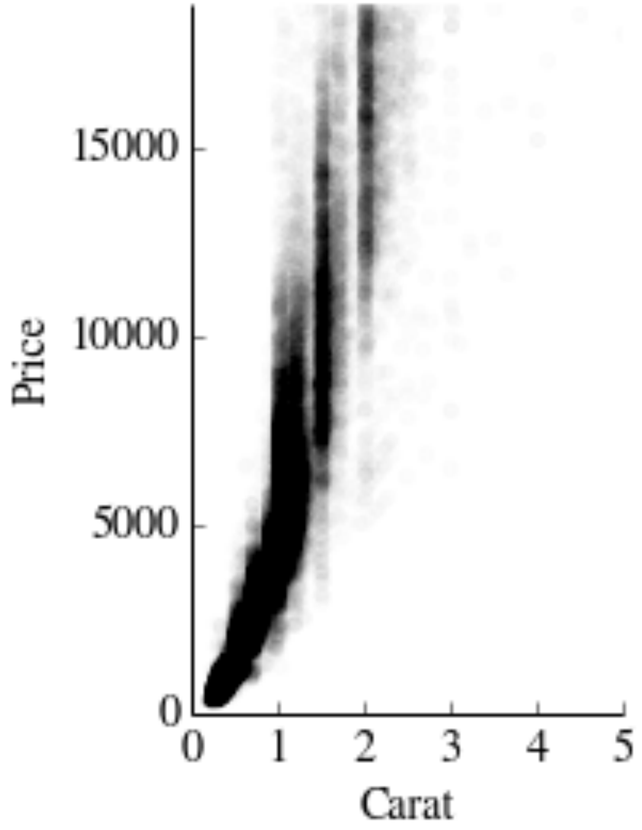
Trivariate Data



Map the third dimension to some other visual attribute

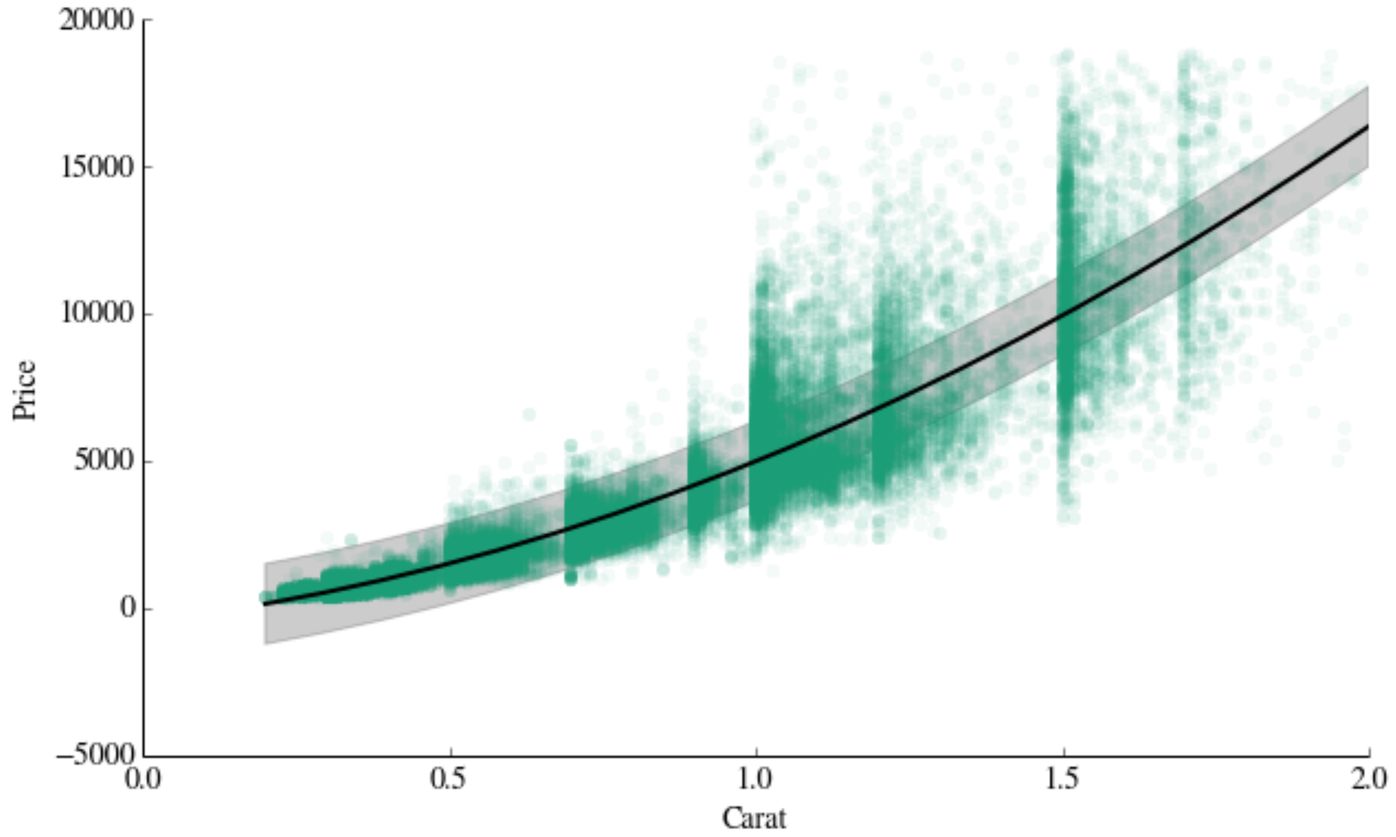
Overplotting





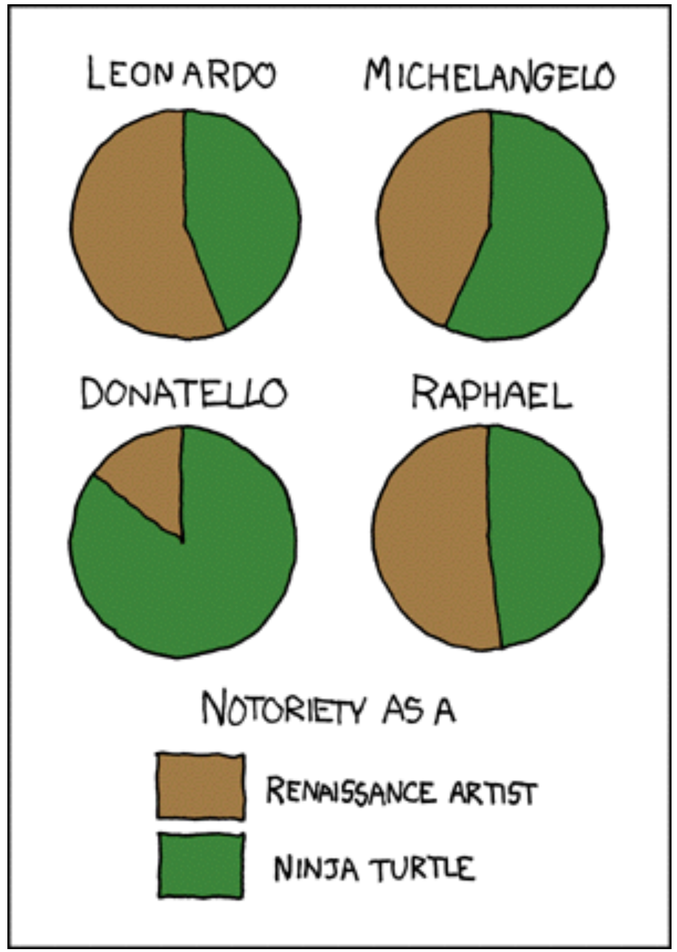
alpha = 1/100

Trend Lines

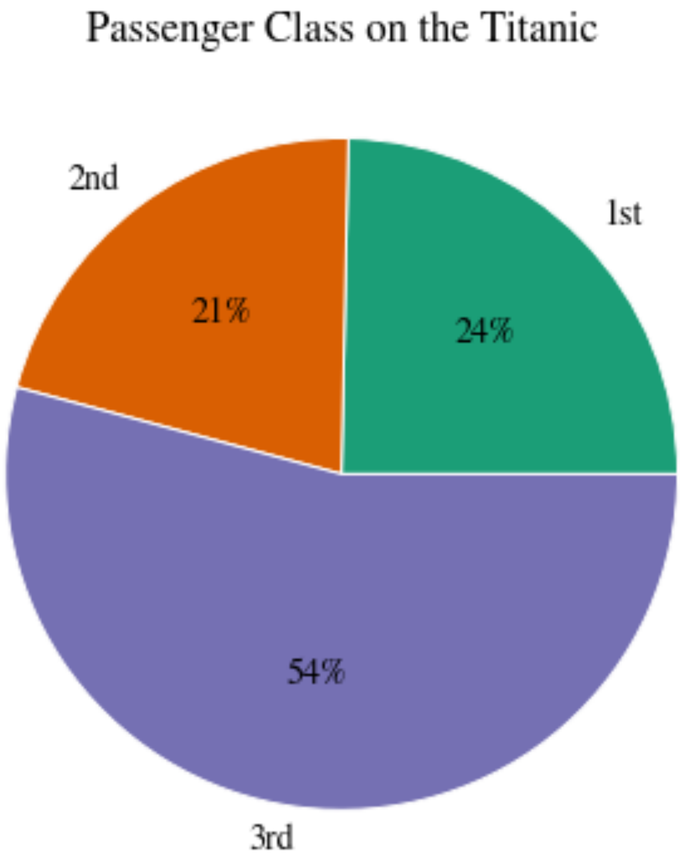


Compositions

Pie Charts

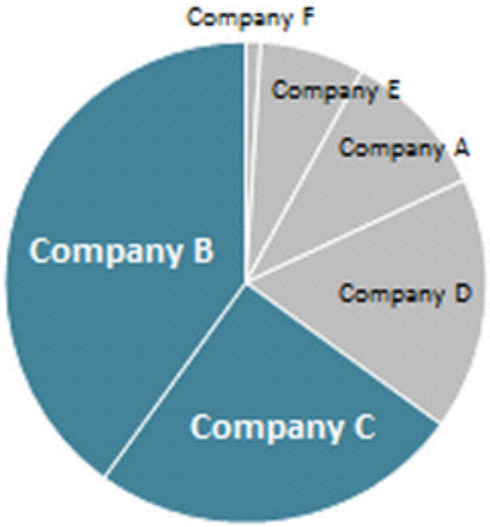


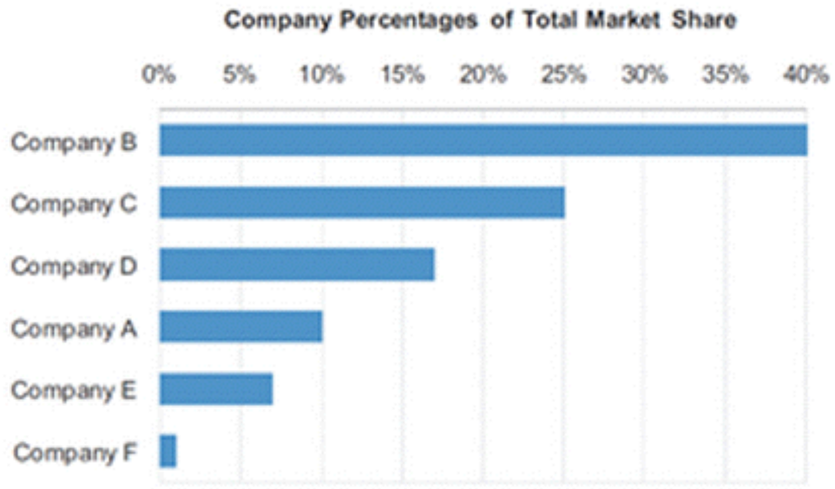
http://xkcd.com/197/



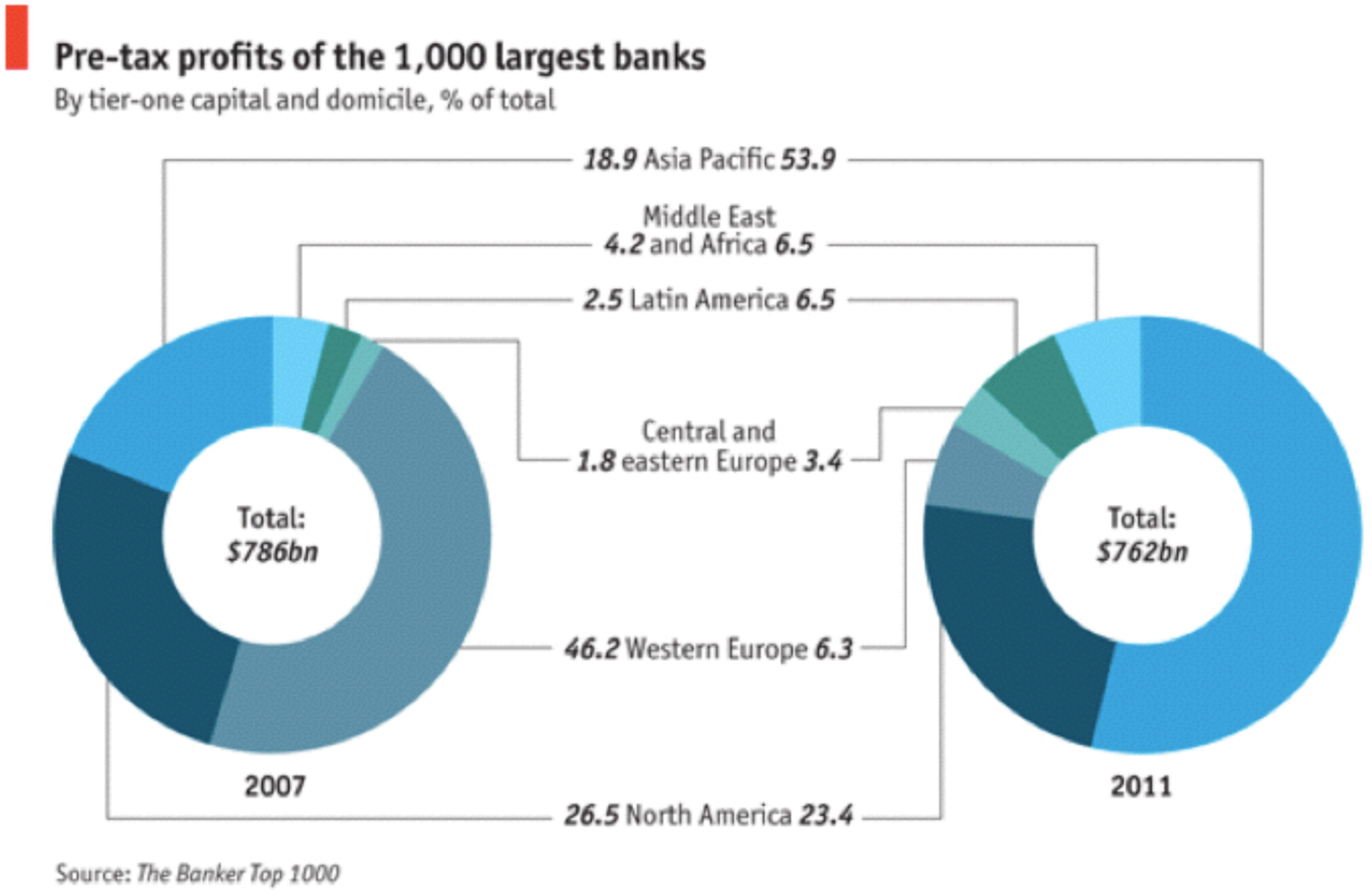
Pie vs. Bar Charts

65% of the market is controlled by companies B and C

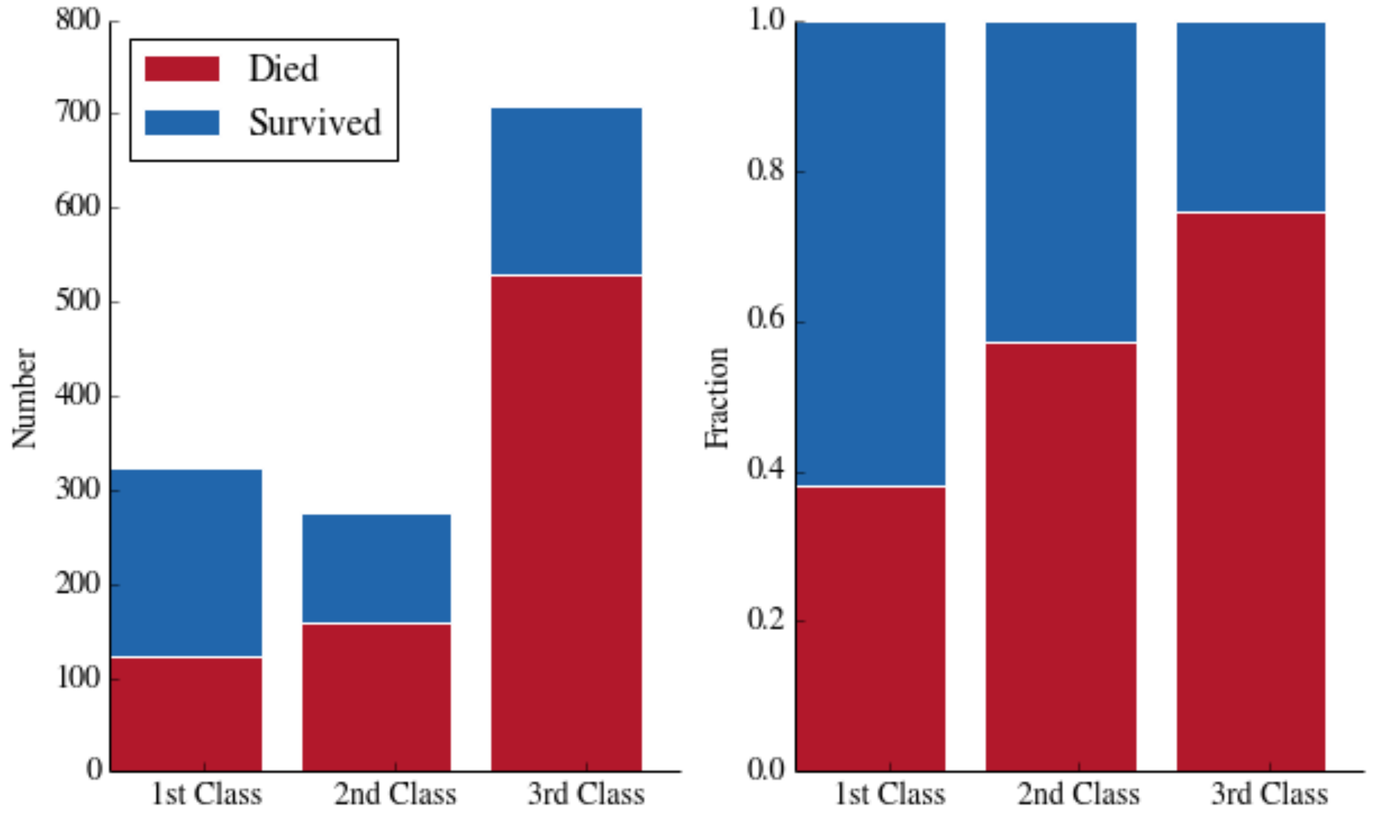




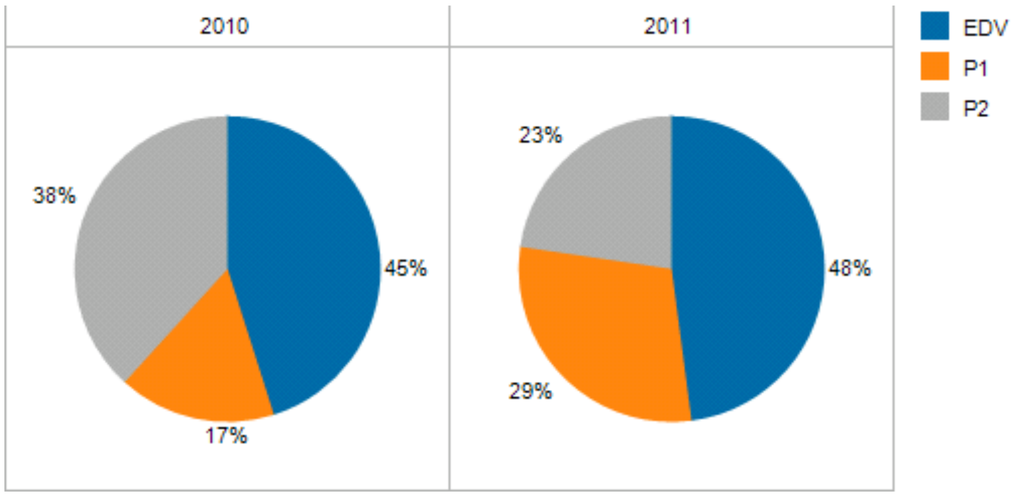
Donut Chart


Pre-tax profits of the 1,000 largest banks
By tier-one capital and domicile, % of total
18.9 Asia Pacific 53.9
Middle East
4.2 and Africa 6.5
2.5 Latin America 6.5
Central and
1.8 eastern Europe 3.4
46.2 Western Europe 6.3
26.5 North America 23.4
Total:
$786bn
2007
Total:
$762bn
2011
Source: The Banker Top 1000

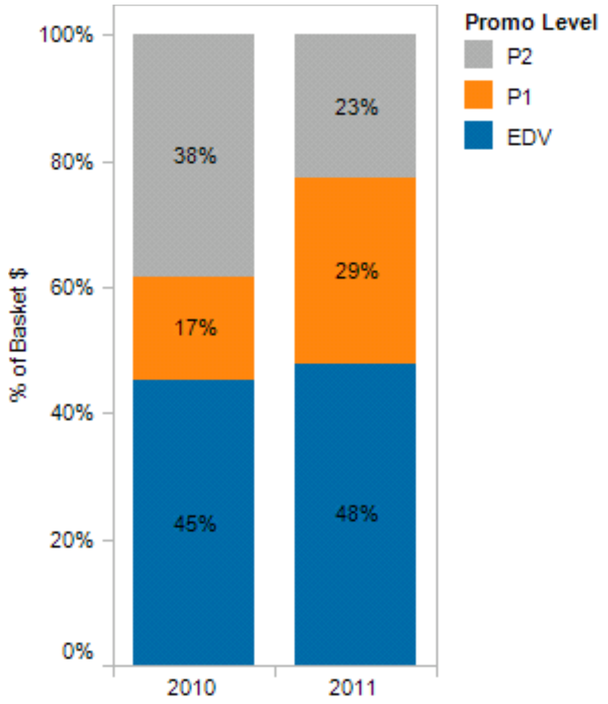
Stacked Bar Chart



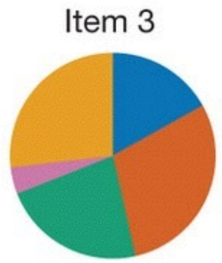
Stacked Bar Chart



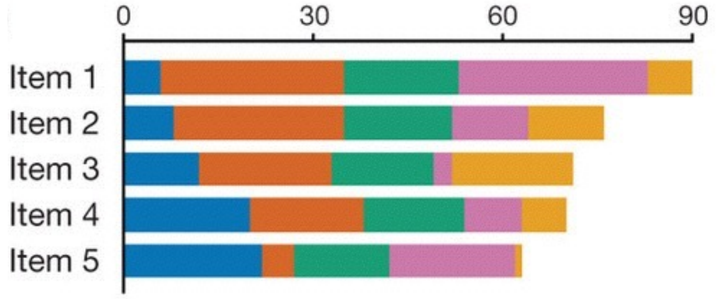
vs.



Comparison of bar chart types

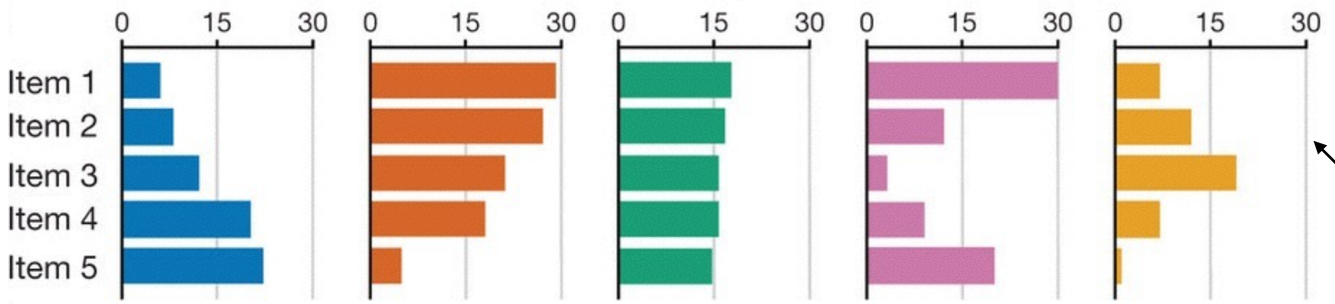


Pie Chart

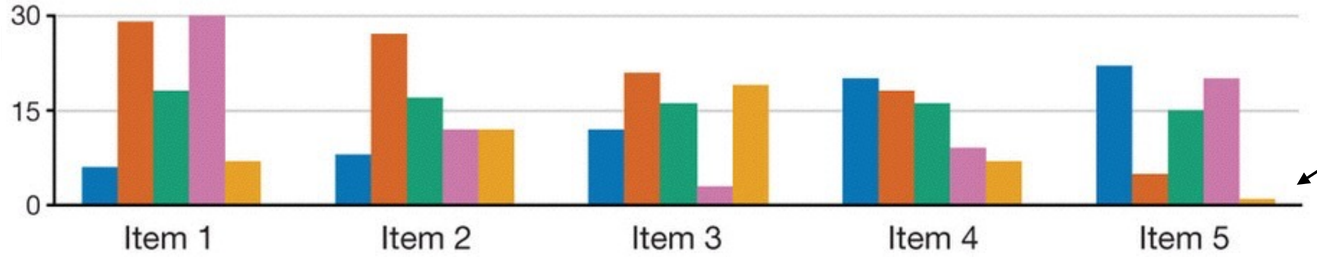


Stacked bar chart

Layered Bar Chart



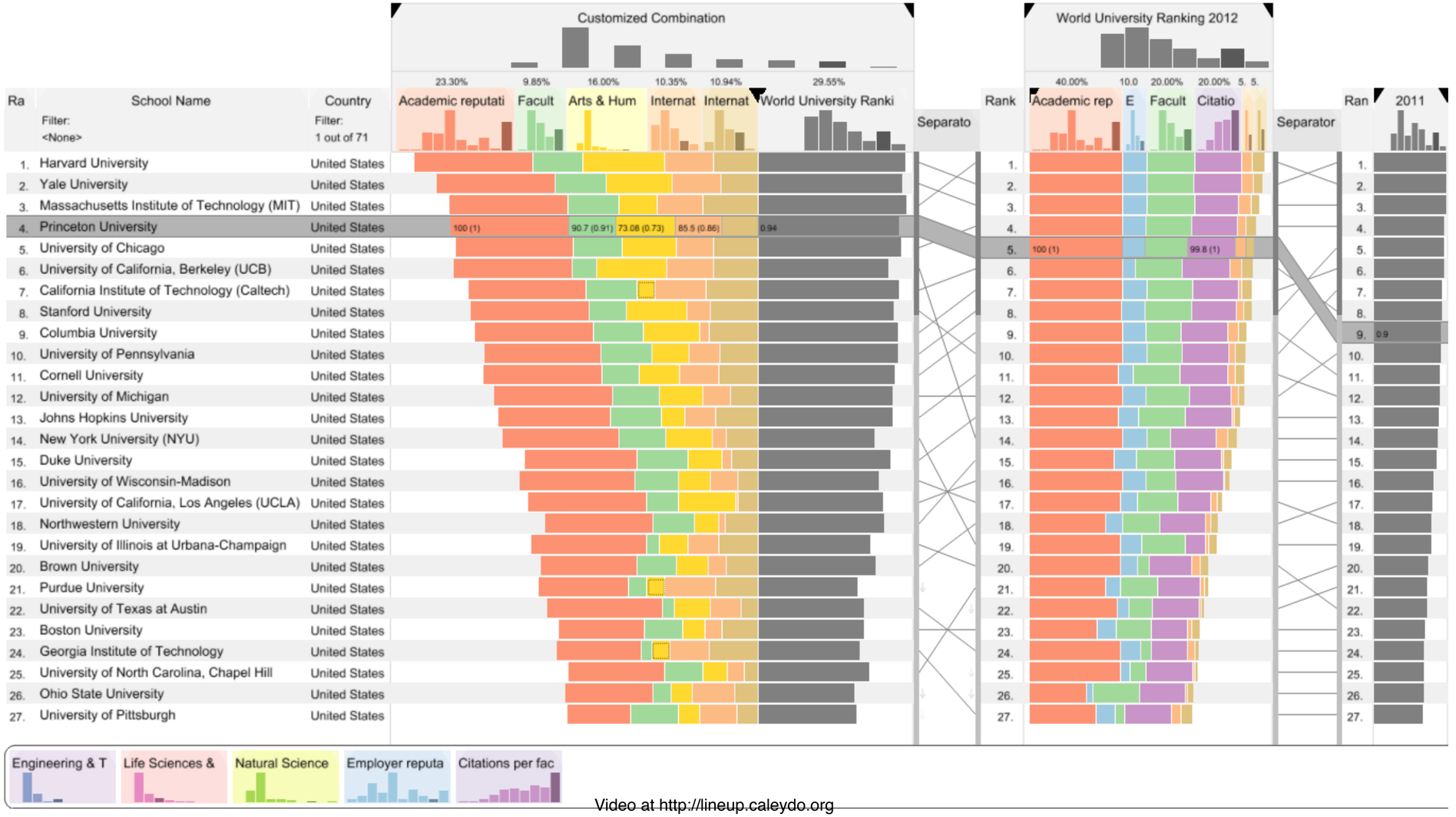
Grouped Bar Chart



Small Multiples

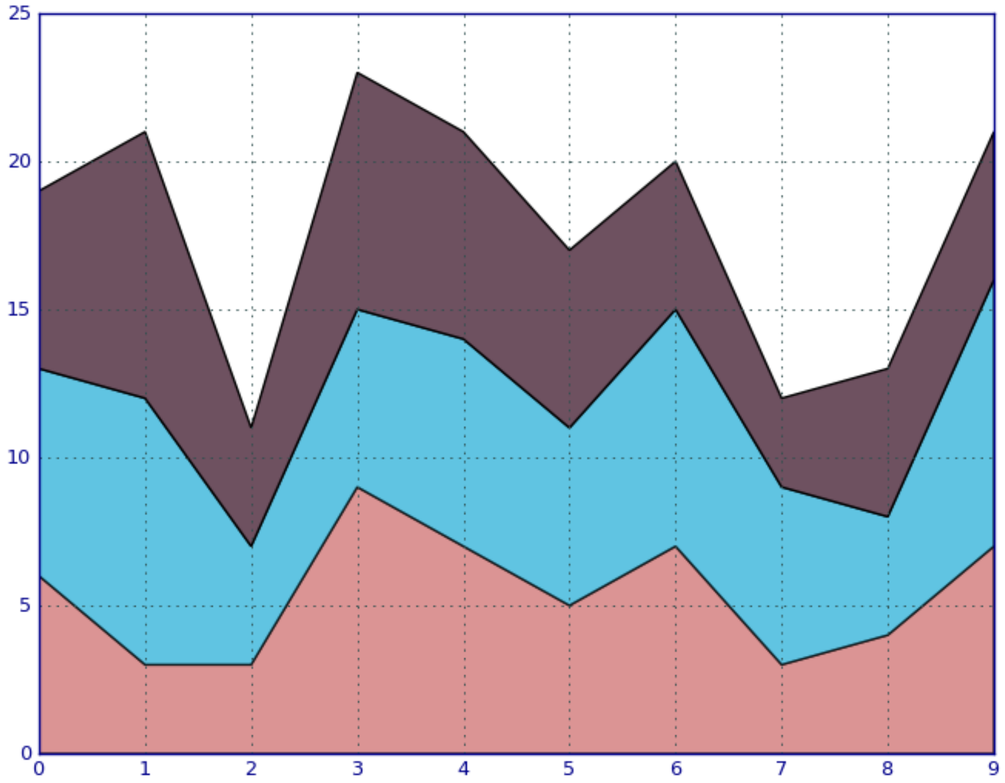
Streit & Gehlenborg, PoV, Nature Methods, 2014

LineUp



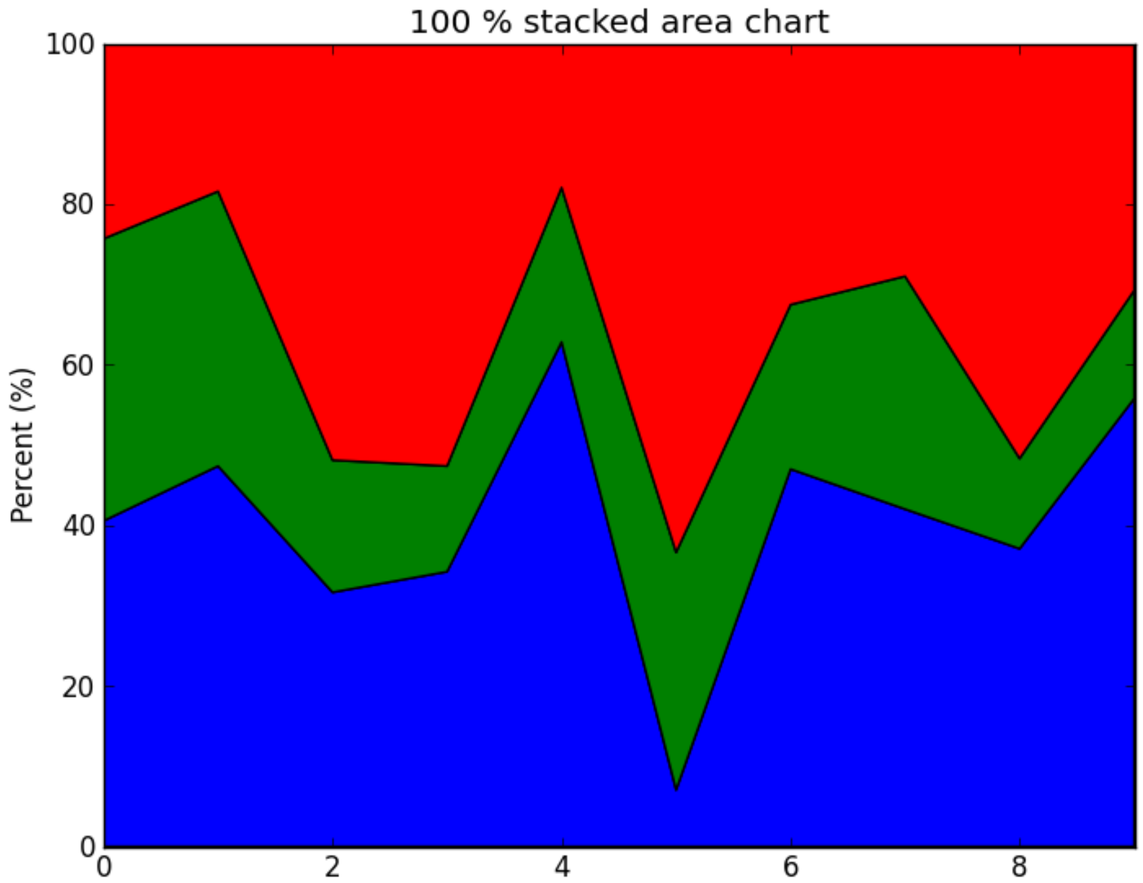
Video at http://lineup.caleydo.org

Stacked Area Chart



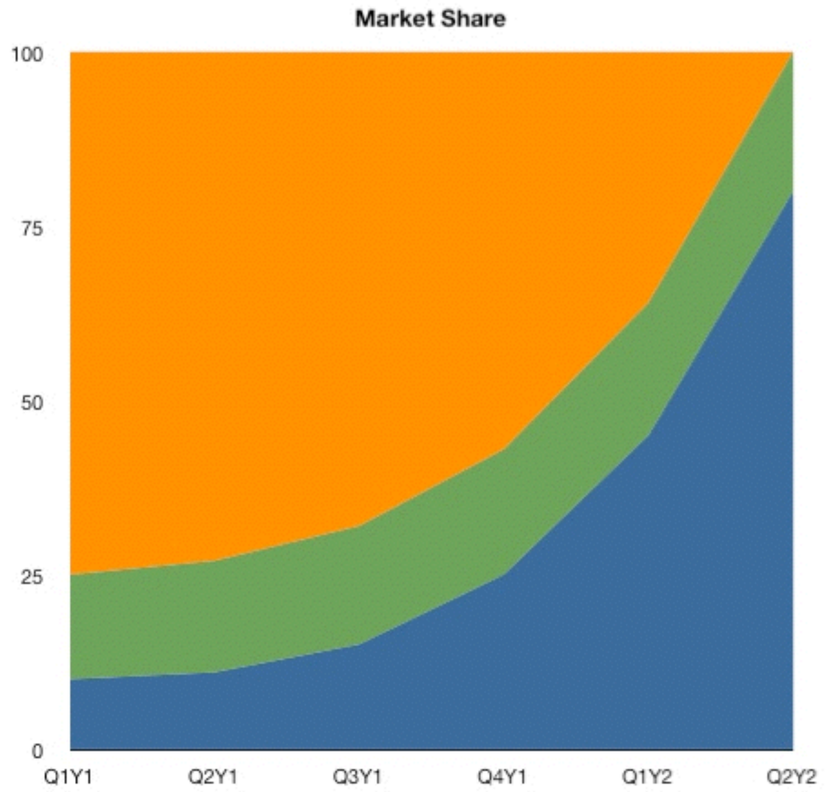
http://stackoverflow.com/questions/2225995/how-can-i-create-stacked-line-graph-with-matplotlib

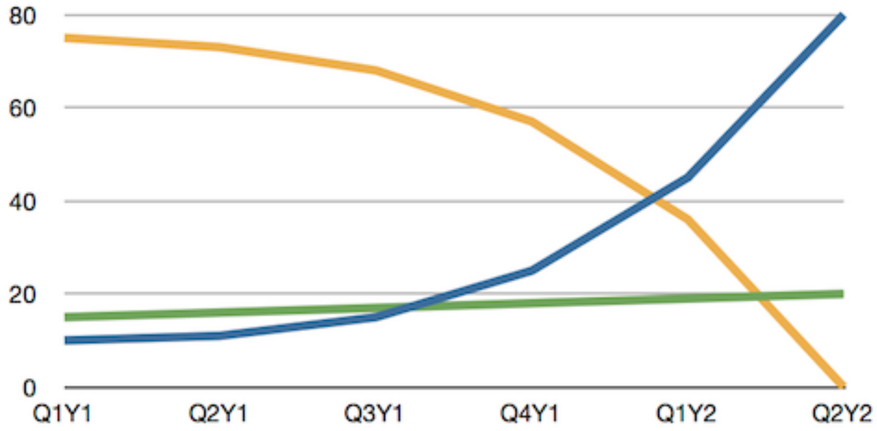
100% Stacked Area Chart



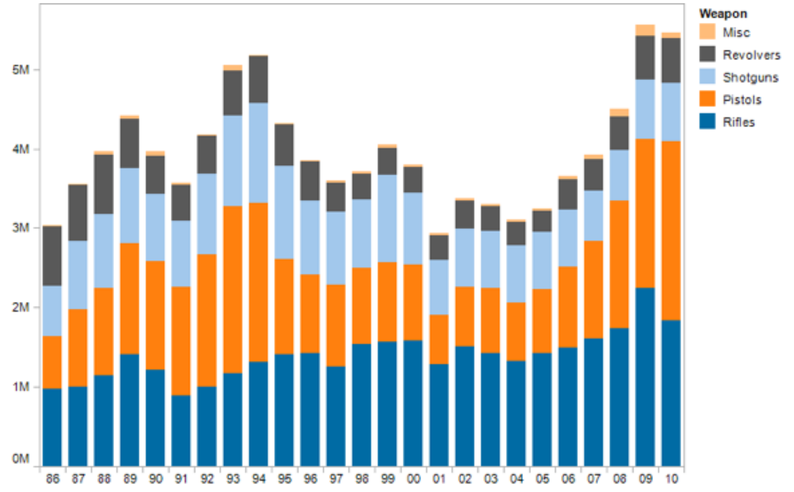
http://stackoverflow.com/questions/16875546/create-a-100-stacked-area-chart-with-matplotlib

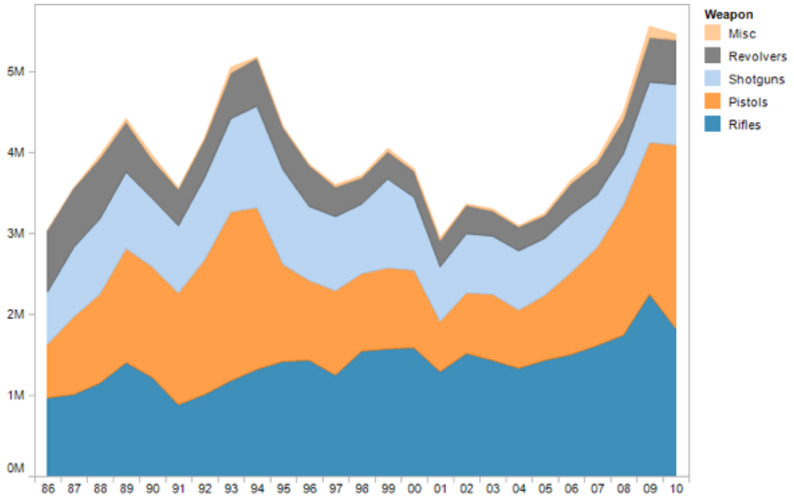
Stacked Area vs. Line Graphs

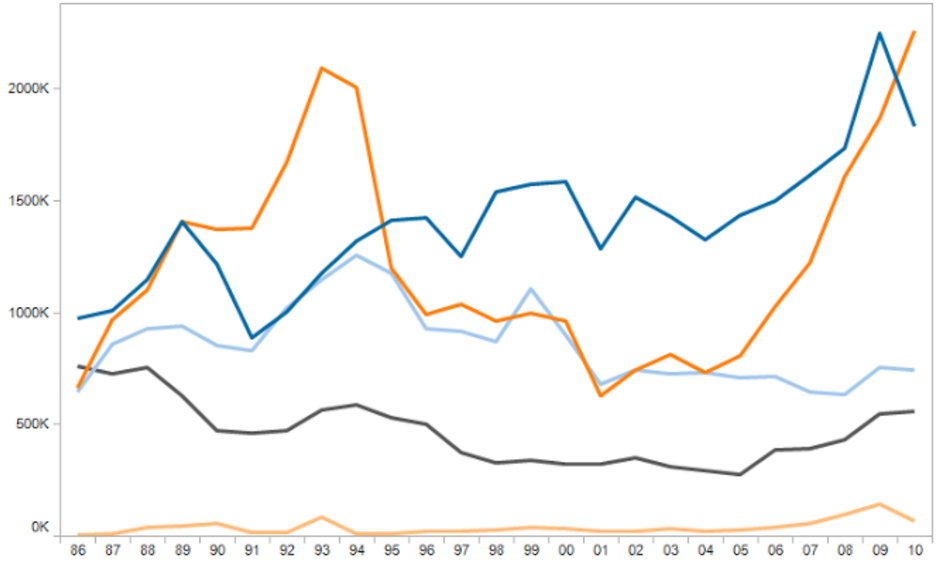




leancrew.com &
Practically Efficient







VizWiz, A. Kriebel

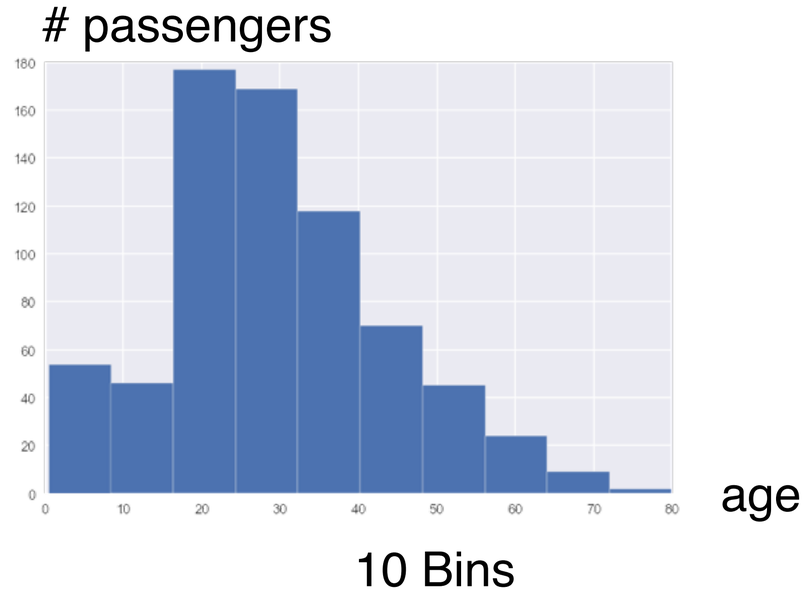
Distributions

Histogram

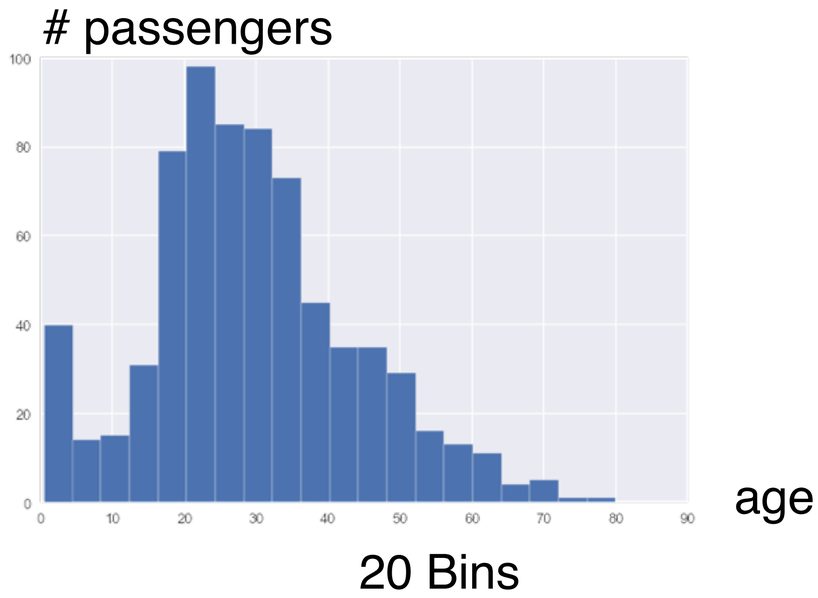
#bins hard to predict

make interactive!

rule of thumb: #bins = sqrt(n)

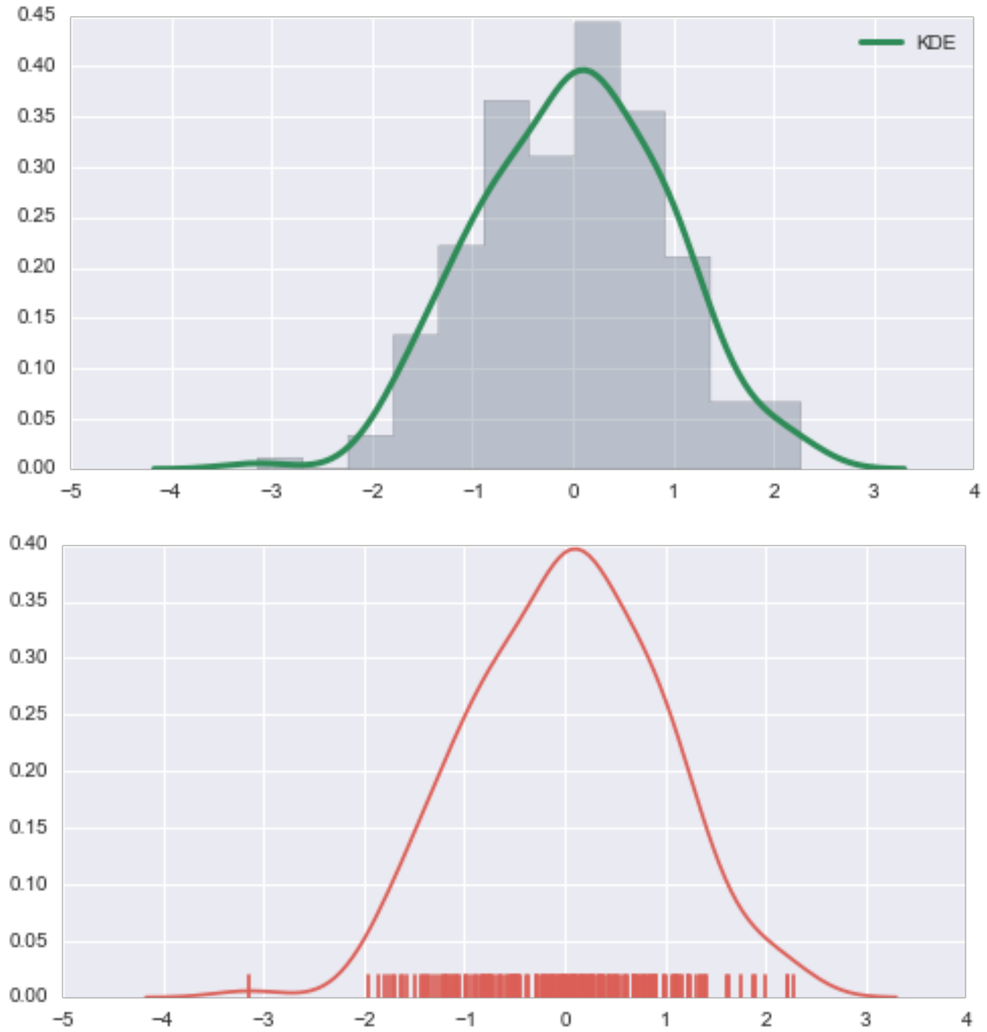


10 Bins



20 Bins

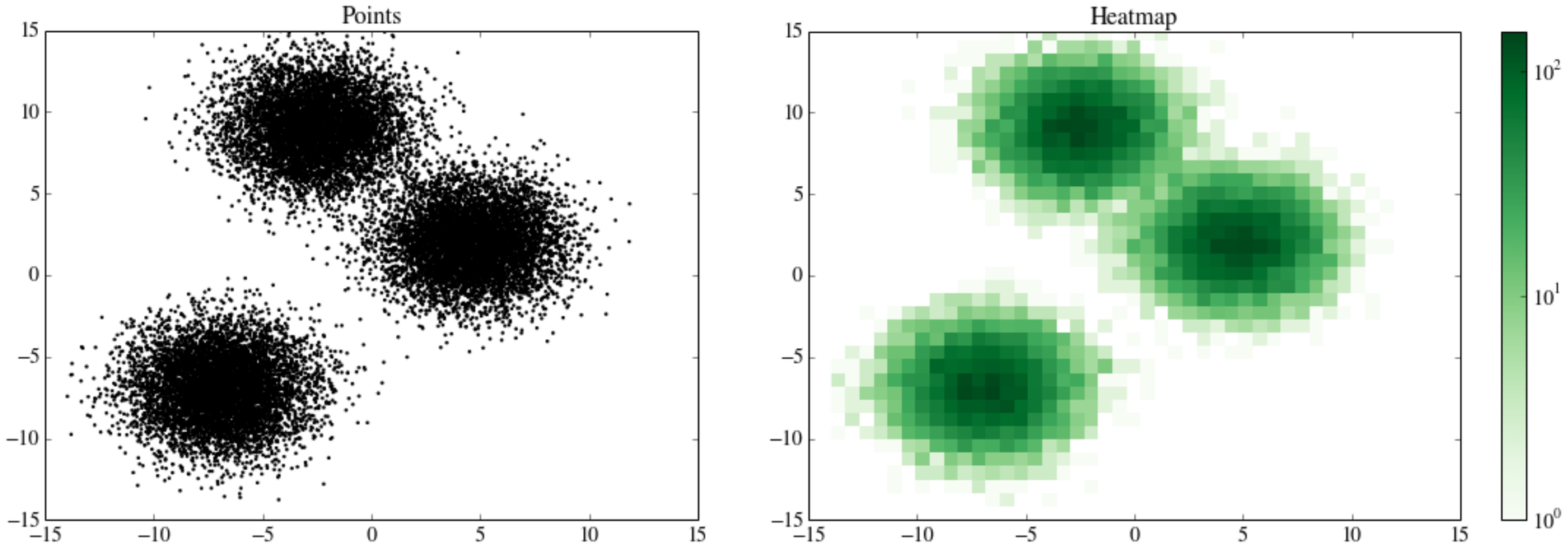
Density Plots



http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

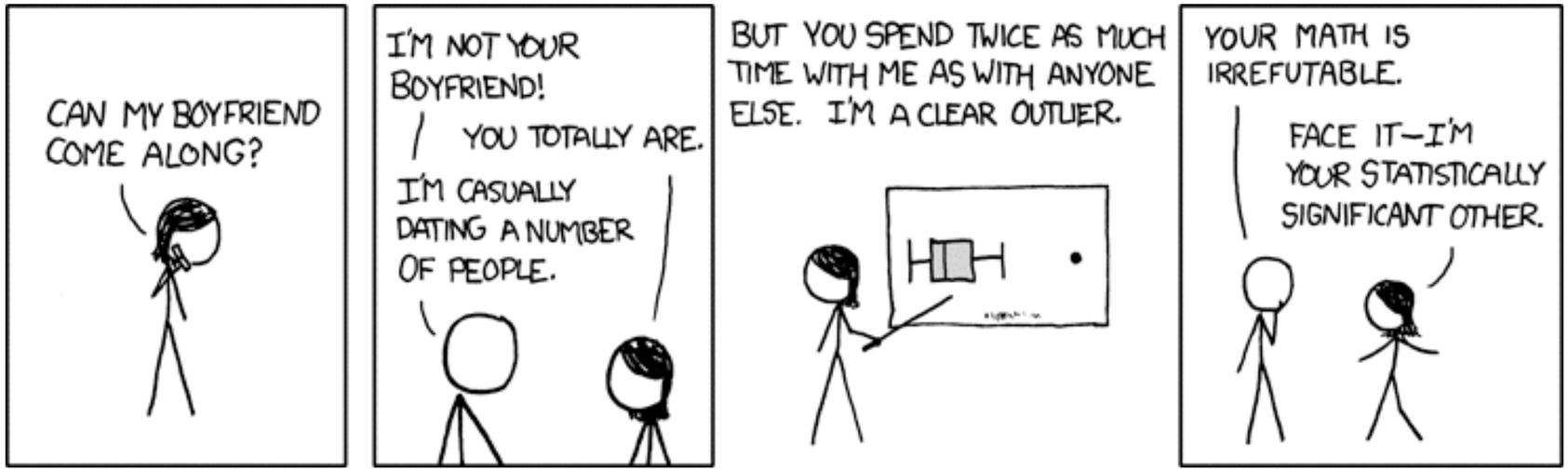
Heat Maps

binning of scatterplots

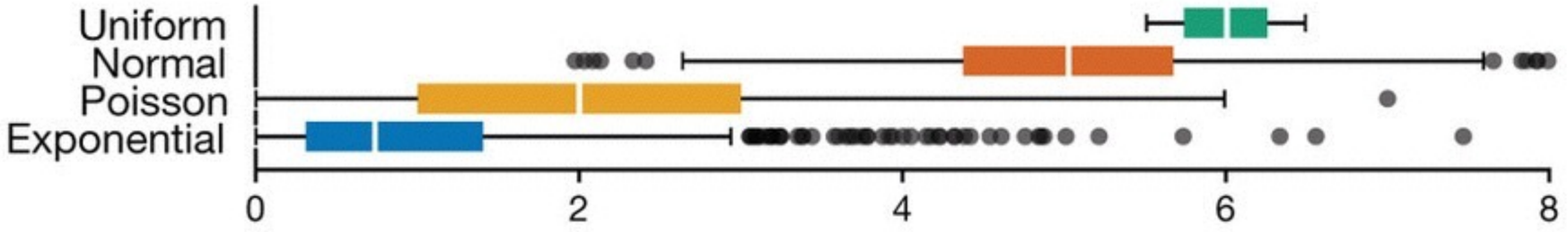


2D Density Plots

Box(and Whisker) Plots

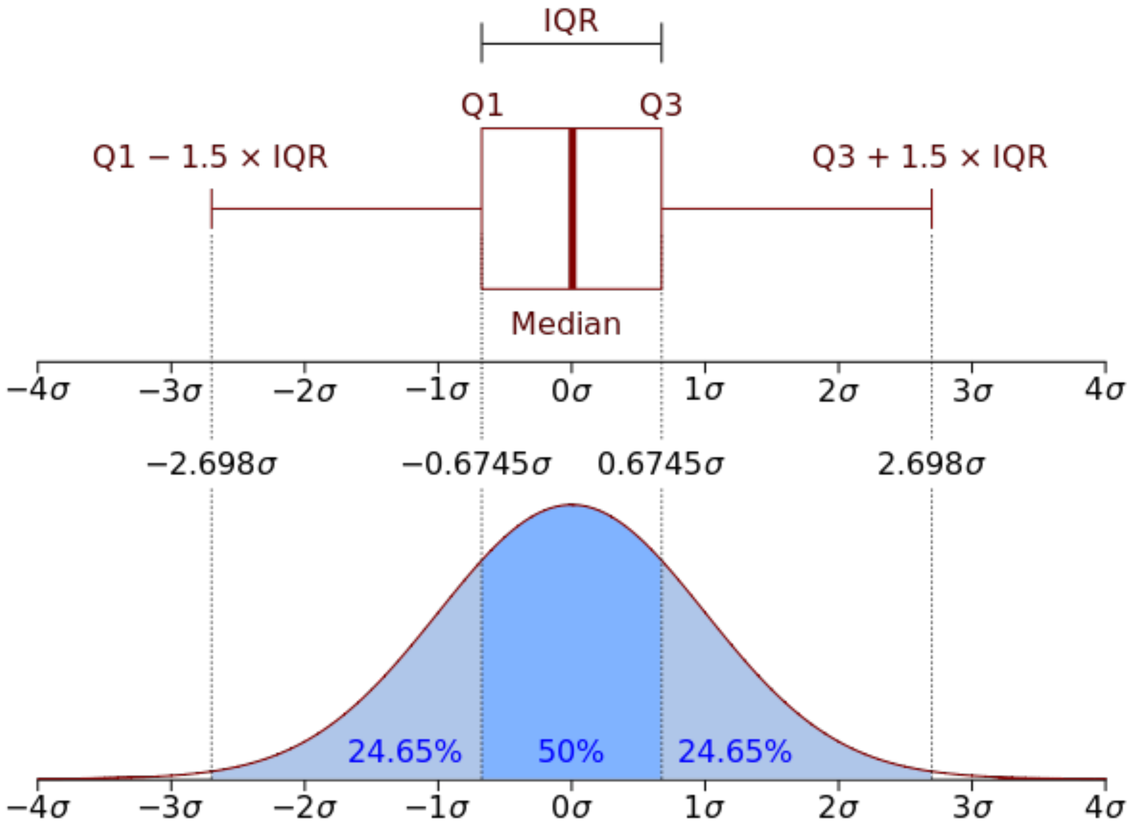


http://xkcd.com/539/



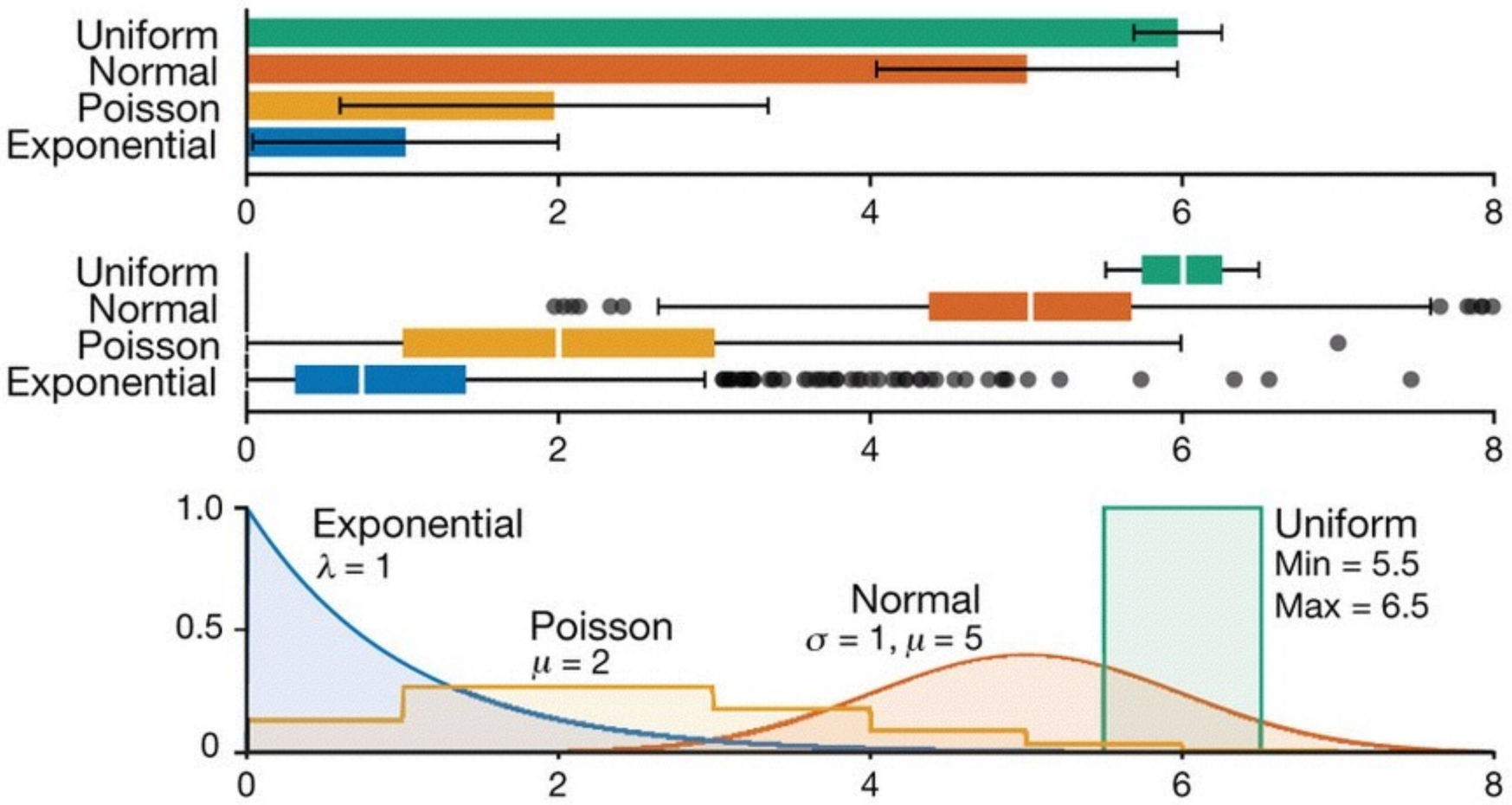
Box Plots

aka Box-and-Whisker Plot



Wikipedia

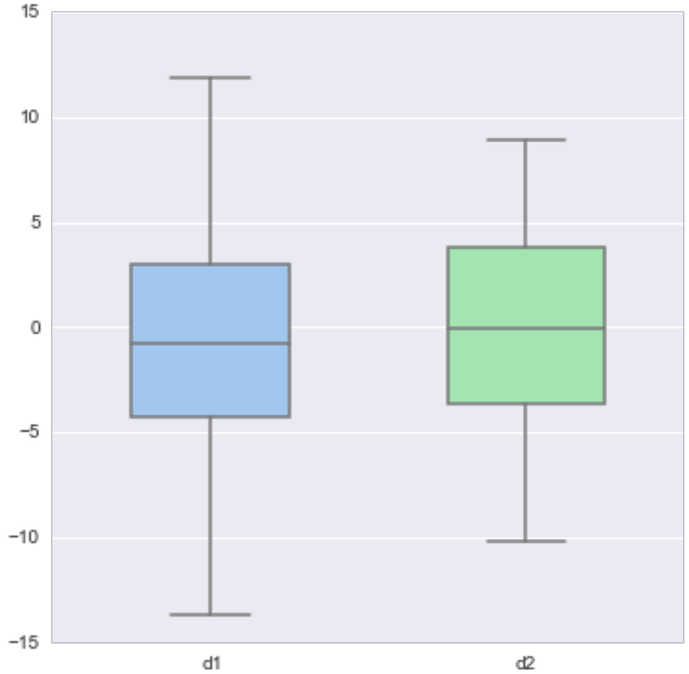
Comparison

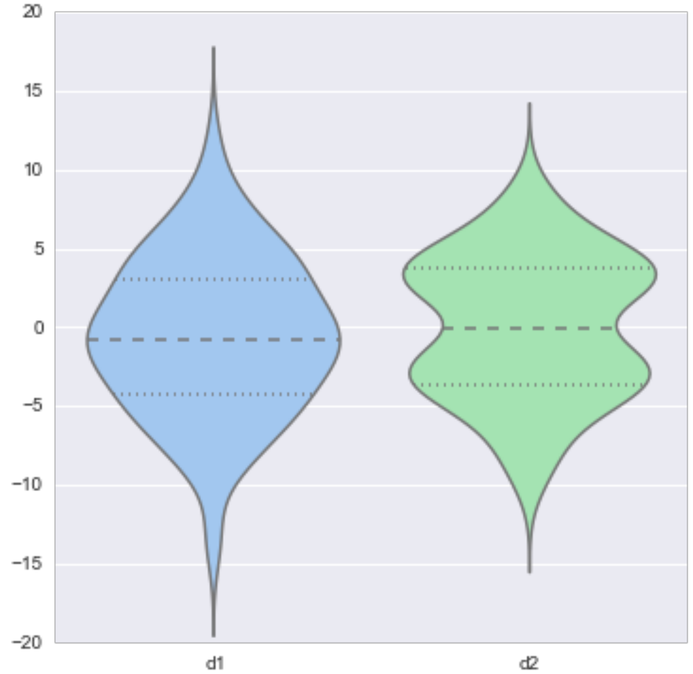


Streit & Gehlenborg, PoV, Nature Methods, 2014

Violin Plot

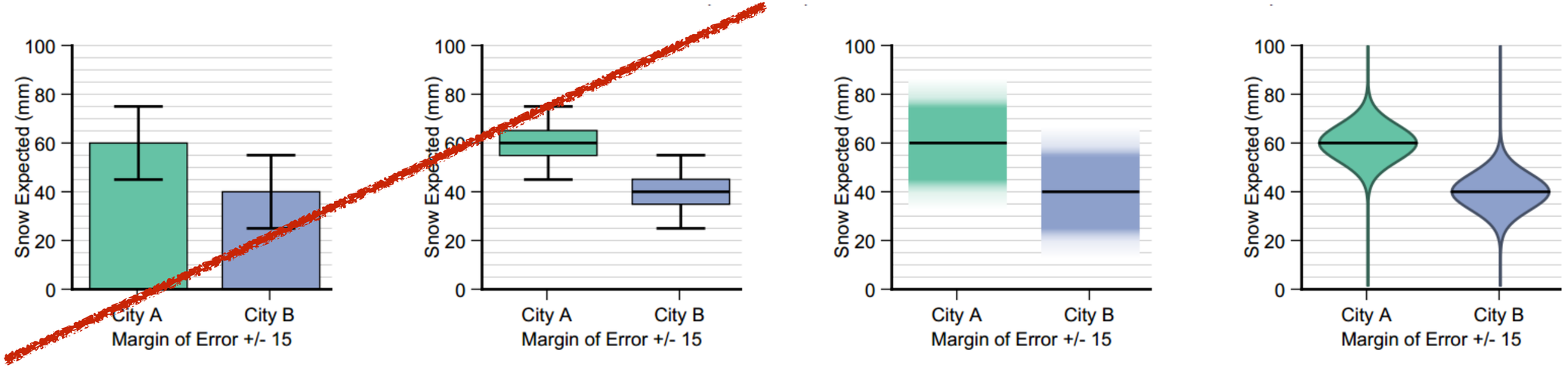
= Box Plot + Probability Density Function





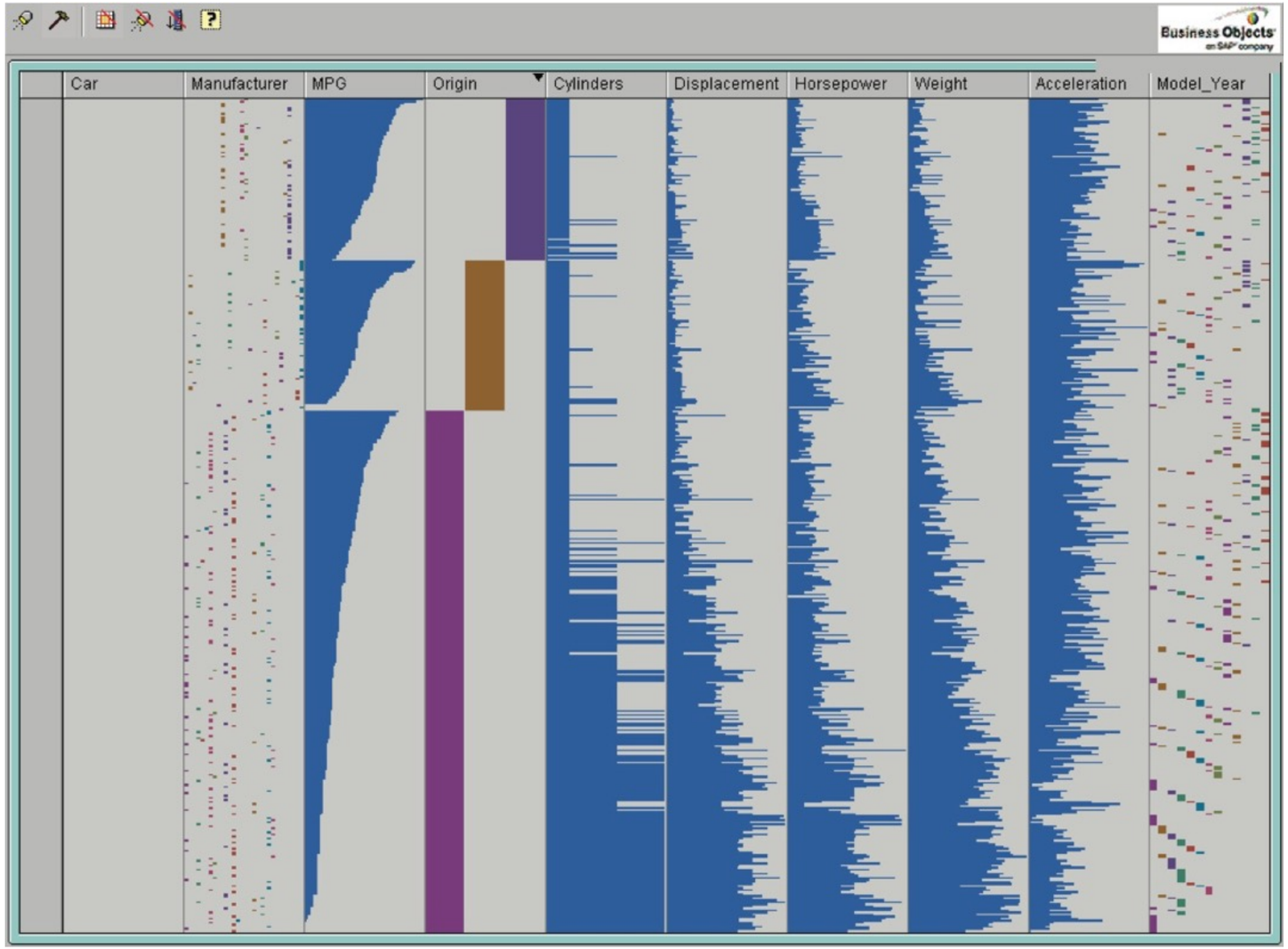
http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Showing Expected Values & Uncertainty



Error Bars Considered Harmful:
Exploring Alternate Encodings for Mean and Error
Michael Correll, and Michael Gleicher

Table Lens

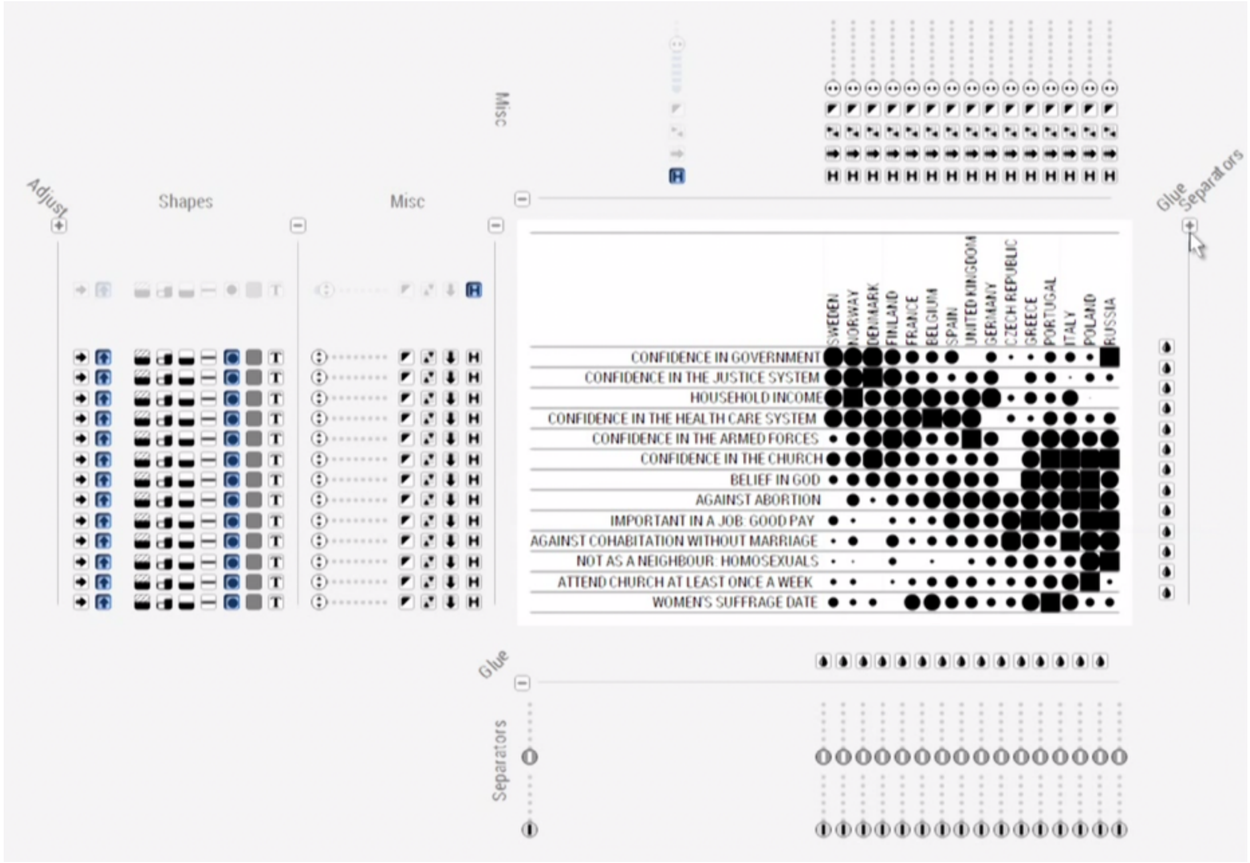


Rao & Card 1994

Bertifier

Matrix/Table representation

Authoring Interface



http://www.aviz.fr/bertifier

Charles Perin, Pierre Dragicevic and Jean-Daniel Fekete

Highdimensional Data

What is High-dimensional Data?

Tabular data, containing

- rows (items)
- columns (attributes or items)
- rows >> columns

	Age	Gender	Height
Bob	***25***	***M***	***181***
Alice	***22***	***F***	***185***
Chris	***19***	***M***	***175***

High-Dimensional Data Visualization

How many dimensions?

~50 – tractable with “just” vis

~1000 – need analytical methods

How many records?

~ 1000 – “just” vis is fine

>> 10,000 – need analytical methods

Homogeneity

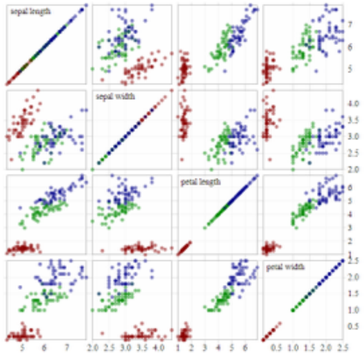
Same data type?

Same scales?

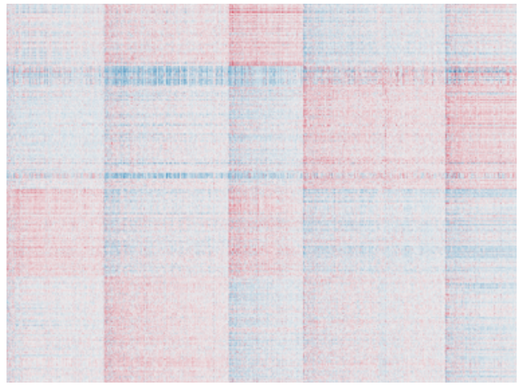
	Age	Gender	Height
Bob	*25*	*M*	*181*
Alice	*22*	*F*	*185*
Chris	*19*	*M*	*175*

	BPM 1	BPM 2	BPM 3
Bob	*65*	*120*	*145*
Alice	*80*	*135*	*185*
Chris	*45*	*115*	*135*

Analytic Component

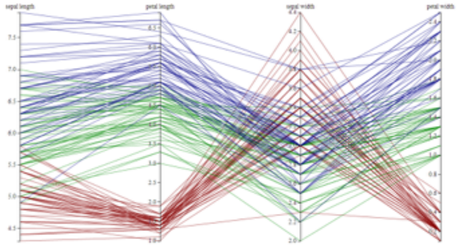


Scatterplot Matrices
[Bostock]

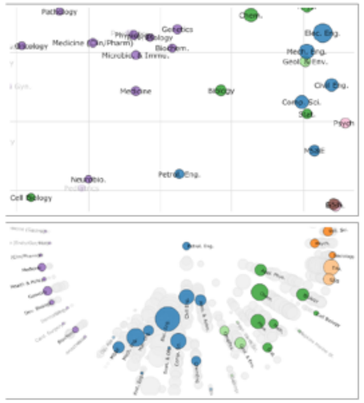


Pixel-based visualizations / heat maps

Multidimensional Scaling
[Doerk 2011]



Parallel Coordinates
[Bostock]



[Chuang 2012]

no / little analytics

strong analytics component

More next time…