Lecture 2: Introduction To Data Data Visualization

Instructor: Saravanan Thirumuruganathan

Outline

1. Data Mining Terminology
2. Basics of Visualization
   - Graph integrity
   - 2D visualization
   - Basics of higher dimensional visualization

Piazza

- Enrollment done for registered students
- Auditing students - send me your email id
- "Search for Teammates" enabled

Socrative

- Instant Student Feedback
- Accessible via Smart Phone, Tablet, Laptop
- No login needed from Student's end
- Use a consistent name throughout the semester

In-Class Quizzes

- **URL:** `http://m.socrative.com/`
- **Room Name:** **4f2bb99e**

Misc Announcements

- Slides for Lecture 1 updated
- Change Office hour timings?
- Installation of Scientific Python

Other Relevant Online Classes

- Machine Learning, Stanford: https://www.coursera.org/course/ml
- Mining of Massive Datasets, Stanford: https://www.coursera.org/course/mmds
- Statistical Learning, Stanford: https://class.stanford.edu/courses/HumanitiesandScience/StatLearning/Winter2015/about

Data Mining Terminology

Data Matrix

Table 1.1. Extract from the Iris dataset

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Data Matrix

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

Data Matrix

- n rows and d columns
- Row $\Rightarrow$ Tuple/Entities
- Column $\Rightarrow$ attribute/feature
- Special column called **Class**
- x_i: i-th row, X_j: j-th column
- Row $\Rightarrow$ entities, instances, examples, records, transactions, objects, points, feature-vectors, tuples
- Column $\Rightarrow$ attributes, properties, features, dimensions, variables, fields
- $n \Rightarrow$ size, $d \Rightarrow$ dimensionality of data

Geometric View

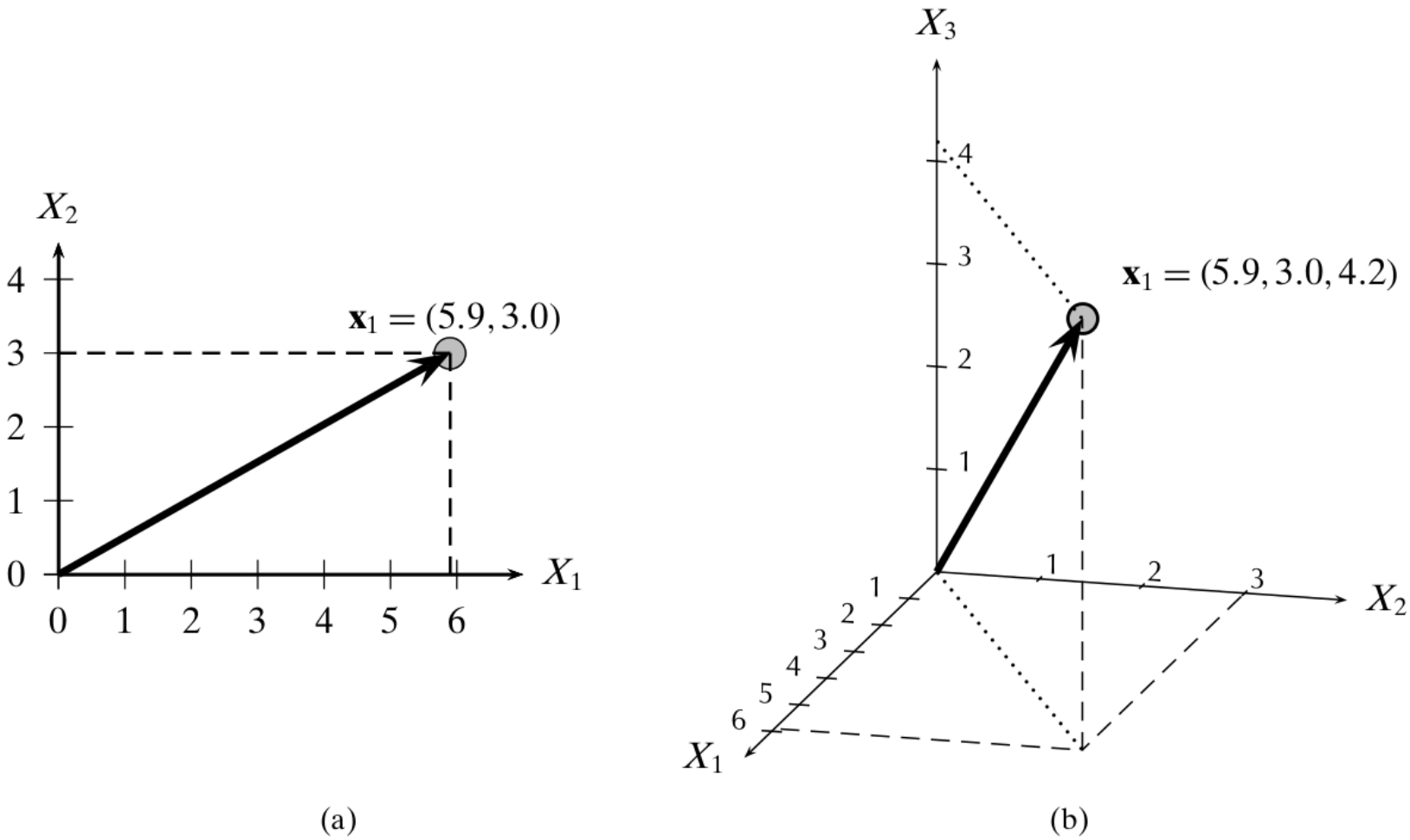
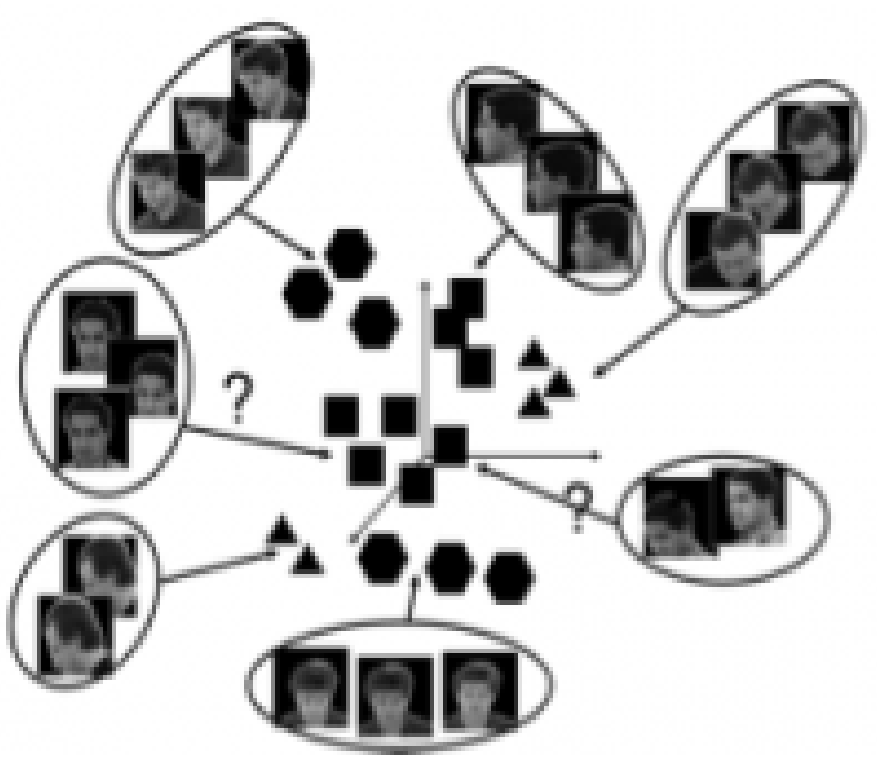


Figure 1.1. Row $\mathbf{x}_1$ as a point and vector in (a) $\mathbb{R}^2$ and (b) $\mathbb{R}^3$.

Implications

- Each photo in the universe is some point in high dimension
- Each book (written or in future) are some point in high dimension



Data Types

Ben Shneiderman, 1996:[1]

- 1D (sequences)
- Temporal
- 2D (maps)
- 3D (shaped)
- nD (relational)
- Trees (hierarchical)
- Networks (graphs)
- Others (text)

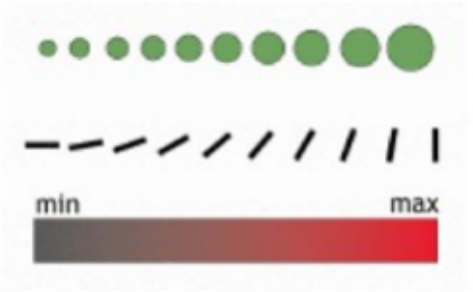
[1] The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization [Shneiderman, 96]

Semantics vs. Types

- Data Semantics: real-world meaning
  - e.g., company name, day of the month, person height, etc.
- Data Type: Interpretation in terms of scales of measurements
  - e.g., quantity or category, sensible mathematical operations etc.

Data Types

- Nominal (Categorical) (N)
  - Are = or ≠ to other values
  - Apples, Oranges, Bananas,...
- Ordinal (O)
  - Obey a < relationship
  - Small, medium, large
- Quantitative (Q)
  - Can do arithmetic on them
  - 10 inches, 23 inches, etc.



On the theory of scales and measurements [S. Stevens, 46]

Data Types

- Q - Interval (location of zero arbitrary)

  Dates: Jan 19; Location: (Lat, Long)

  Like a geometric point. Cannot compare directly.

  Only differences (i.e., intervals) can be compared

- Q - Ratio (zero fixed)

  Measurements: Length, Mass, Temp, ...

  Origin is meaningful, can measure ratios & proportions

  Like a geometric vector, origin is meaningful

Data Types

- N - Nominal (labels)
  - Operations: $=, \neq$
- O - Ordinal (ordered)
  - Operations: $=, \neq, >, <$
- Q - Interval (location of zero arbitrary)
  - Operations: $=, \neq, >, <, +, -$
- Q - Ratio (zero fixed)
  - Operations: $=, \neq, >, <, +, -, \times, \div$

Quiz!

What is the data type of:

- Gender:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age: Ordinal
- Height:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age: Ordinal
- Height: Quantitative - Ratio
- Date:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age: Ordinal
- Height: Quantitative - Ratio
- Date: Quantitative - Interval

Data Dimensions

- Univariate (1D)
- Bivariate (2D)
- Trivariate (3D)
- Multivariate (nD)

Introduction To Data Visualization

Visualization Goals

- **Presentation**
  - Known facts about data
  - Task: Communicate results
- **Exploration**
  - Data without hypothesis
  - Task: Generate hypothesis
- **Confirmation**
  - Hypothesis is given
  - Task: Verify / falsify hypothesis

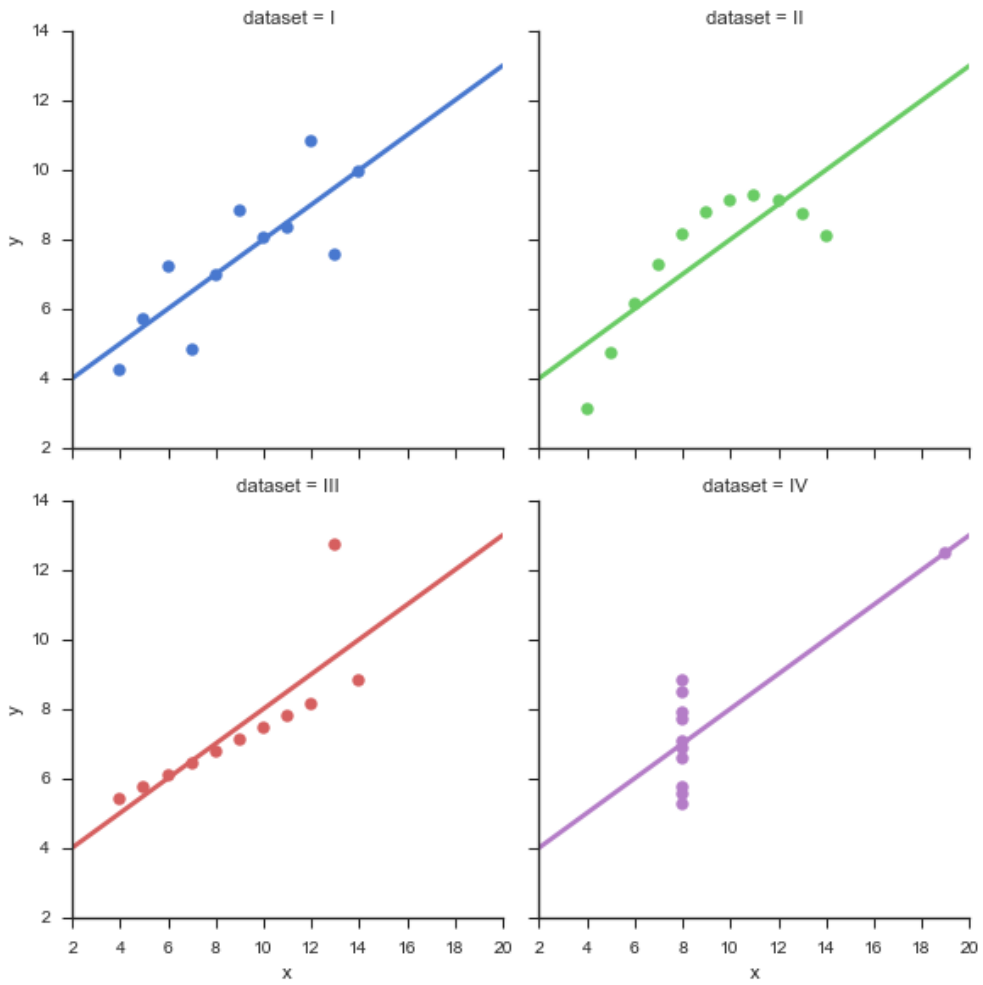
Visualization Goals

"The greatest value of a picture is when it forces us to notice what we never expected to see."

-John Tukey (1915 - 2000)

Anscombe's Quartet

Same mean, variance, correlation, and linear regression line

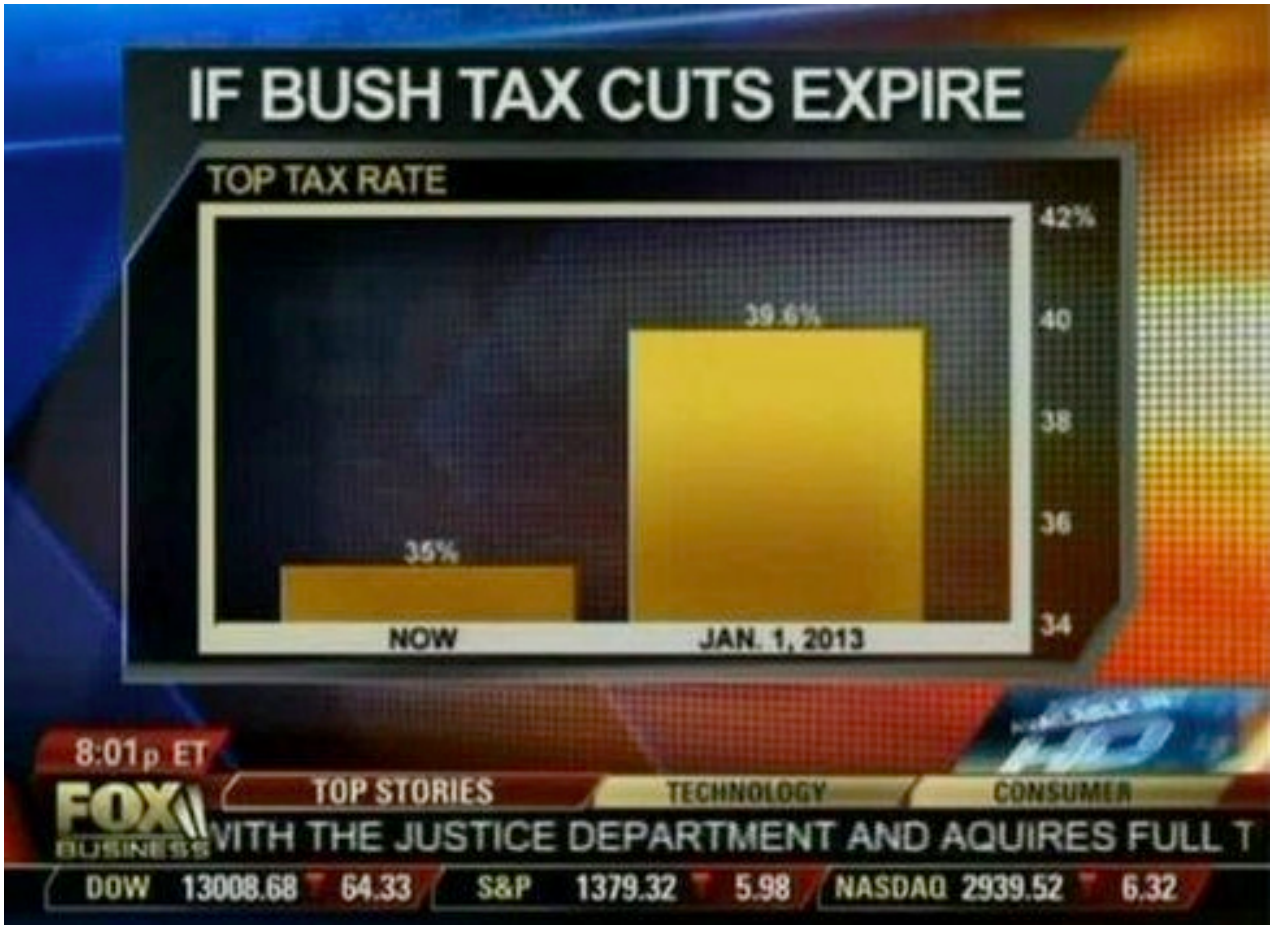


Graphical Integrity

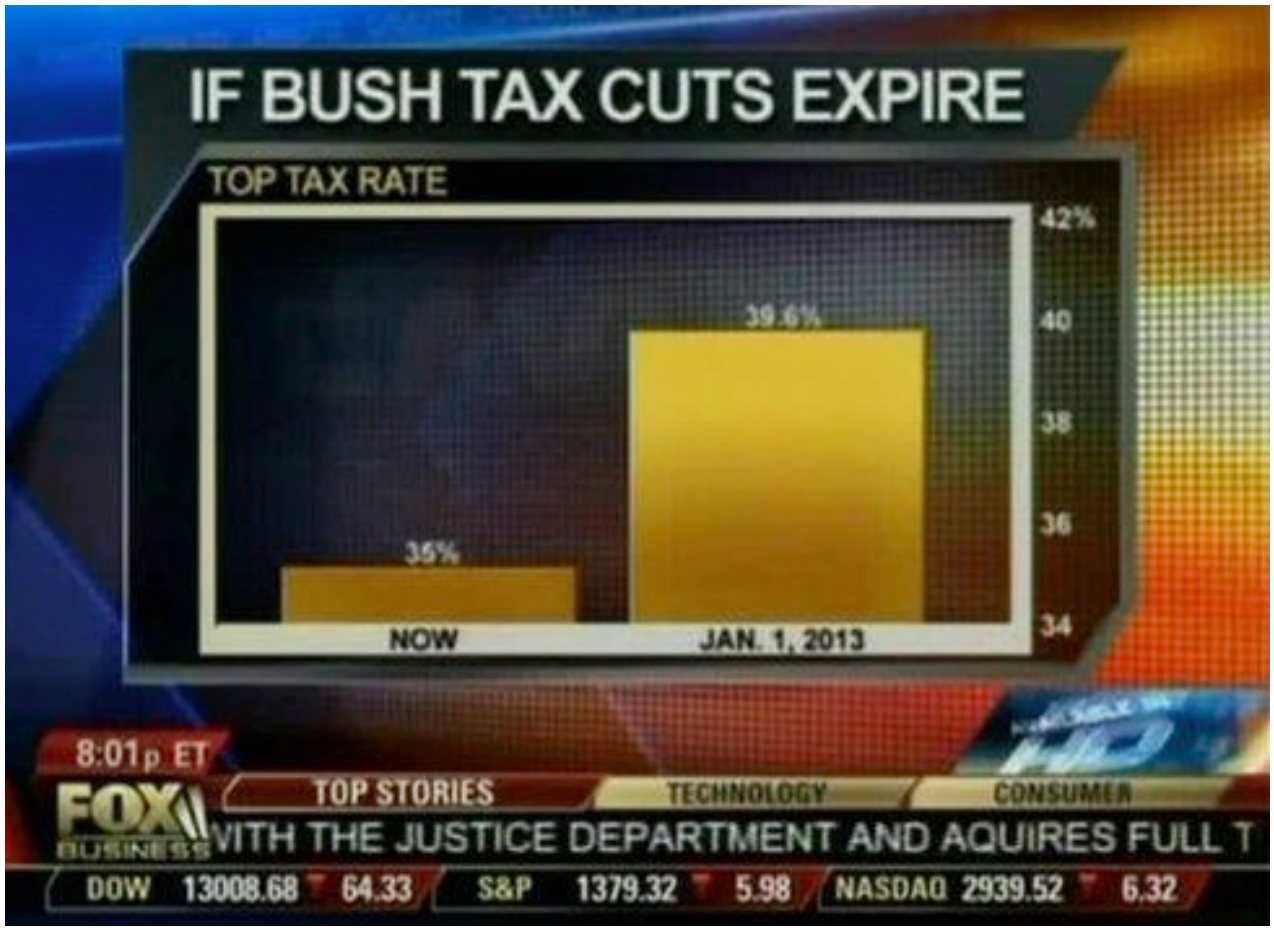
"There are three kinds of lies: lies, damned lies, and statistics."

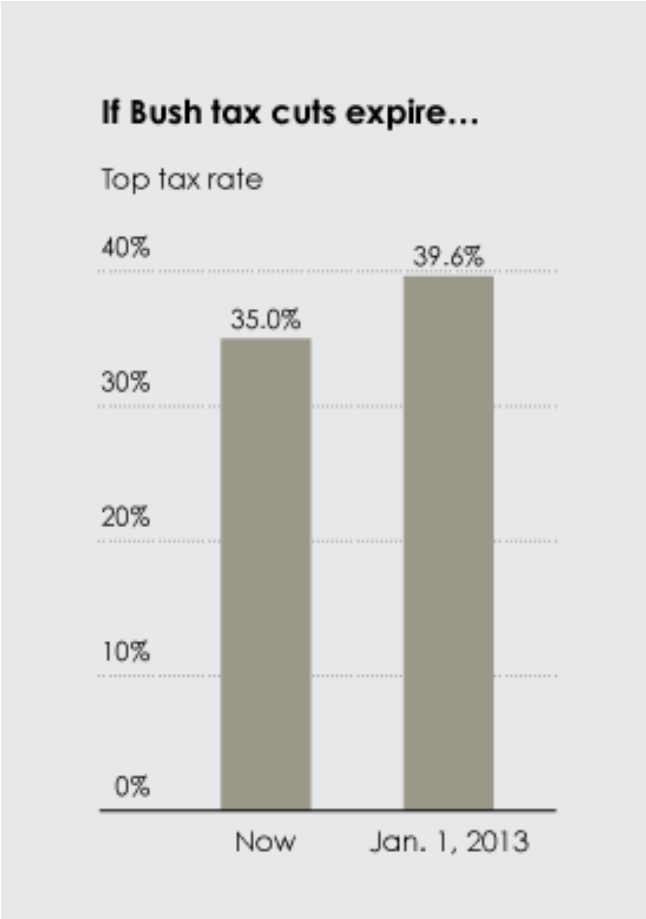
- attributed to Benjamin Disraeli in 19th Century

Graphical Integrity

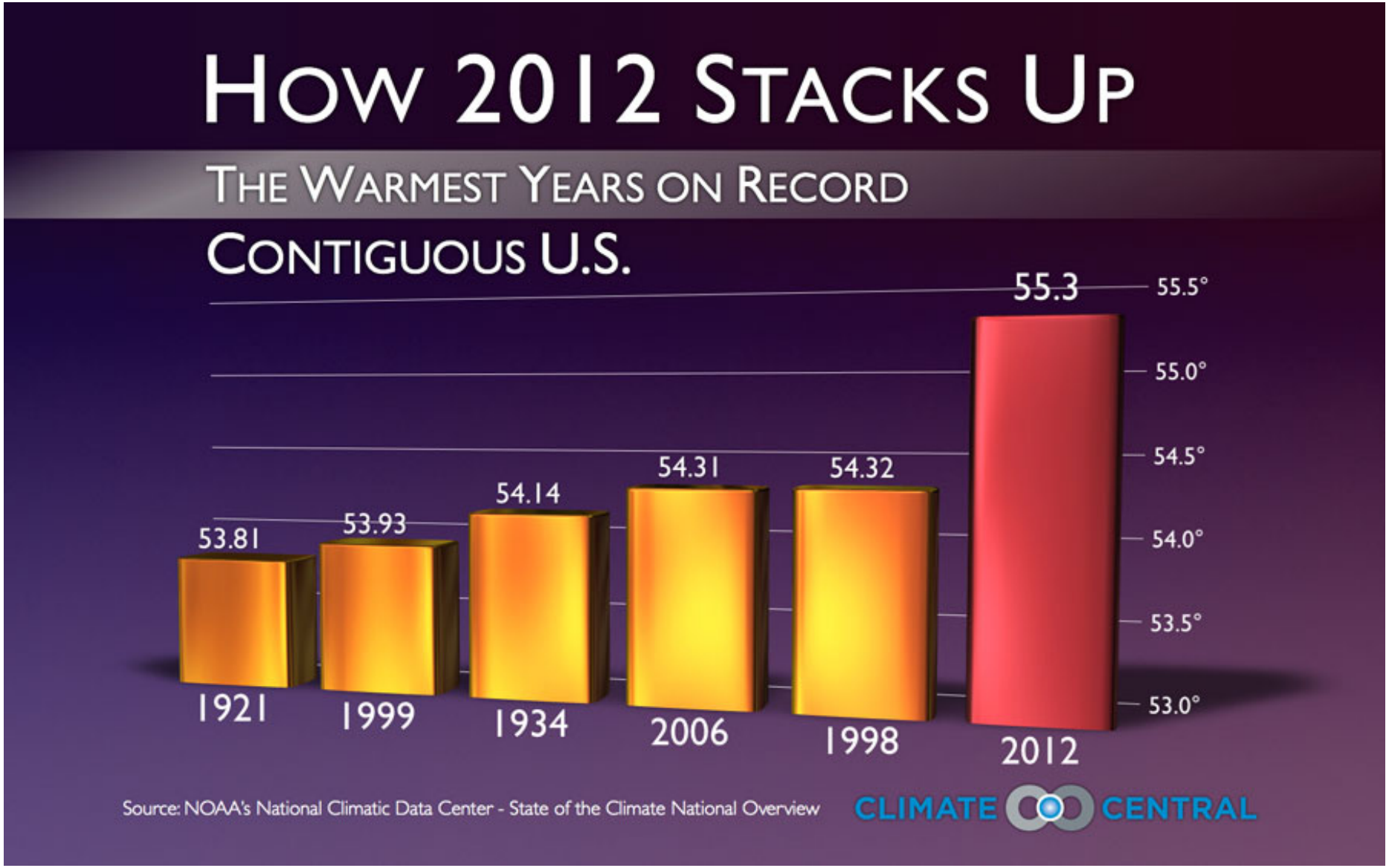


Scale Distortions



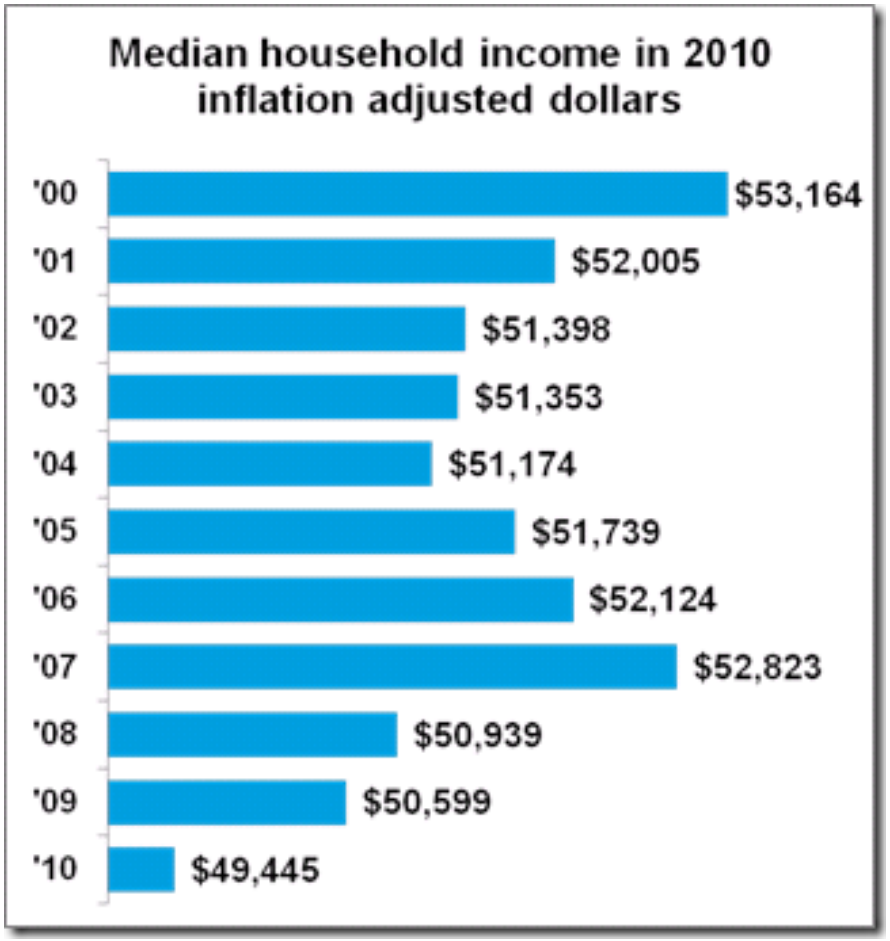


Scale Distortions



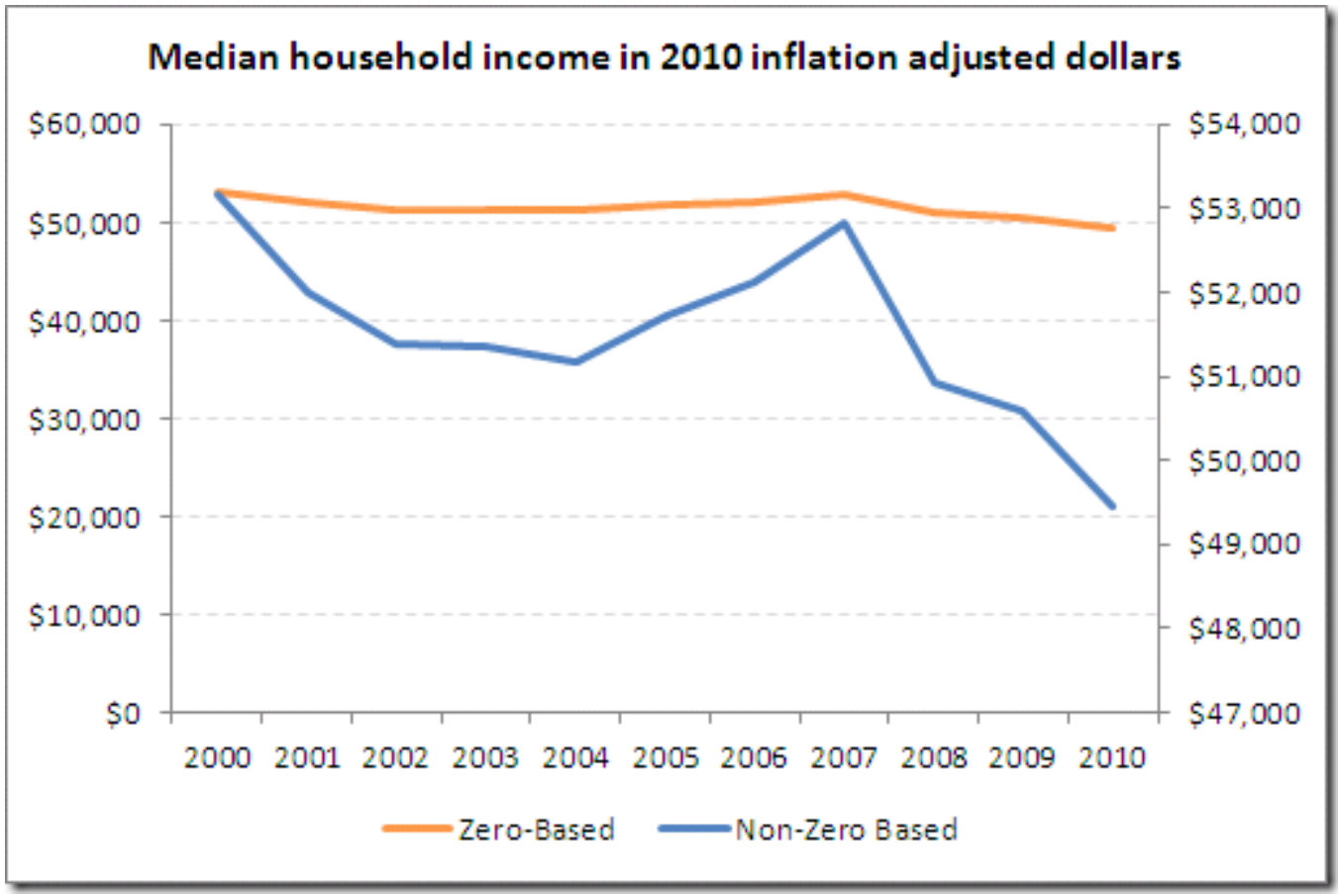
Scale Distortions

Always start your bar graphs at zero!



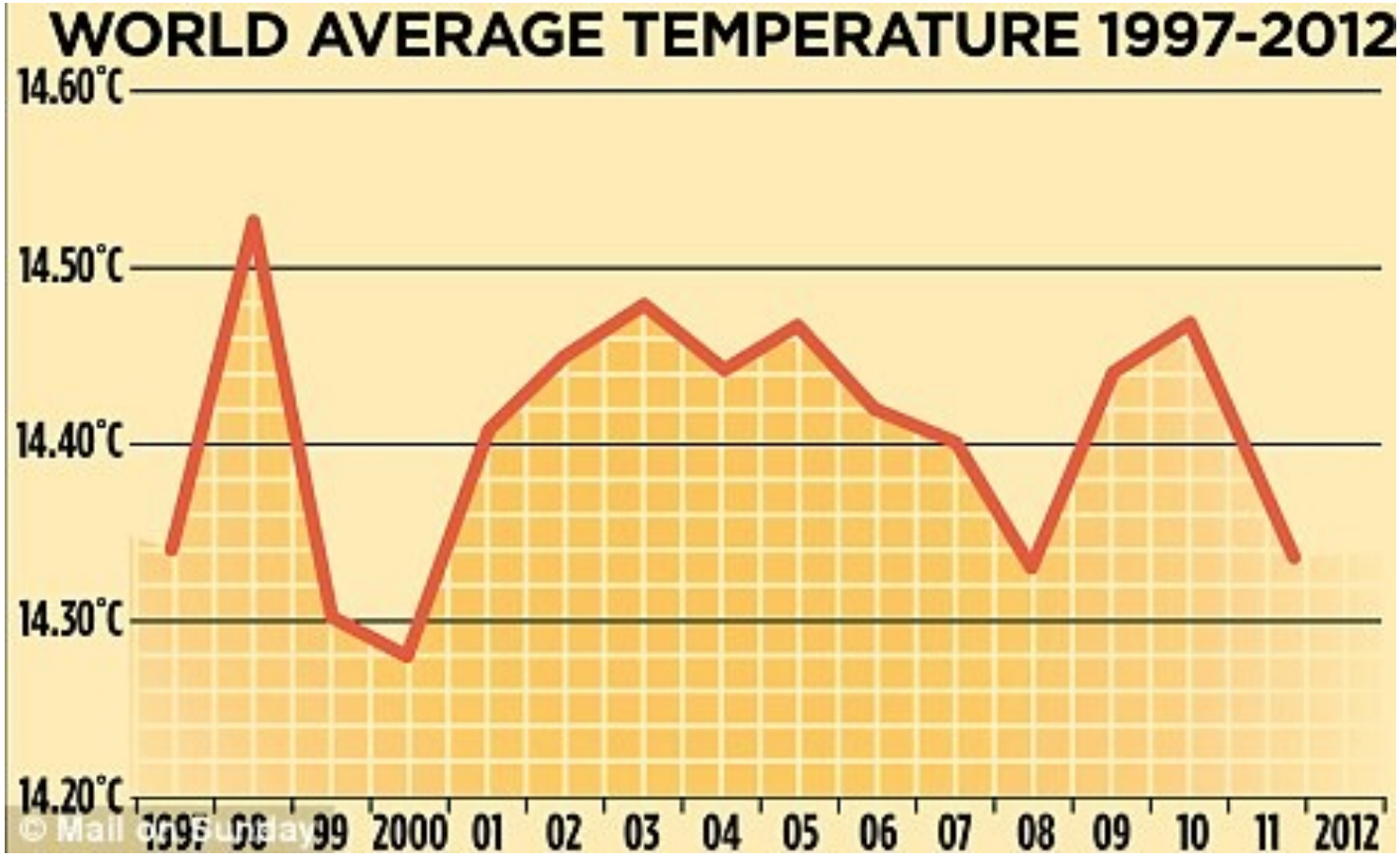
Nielson

Scale Distortions



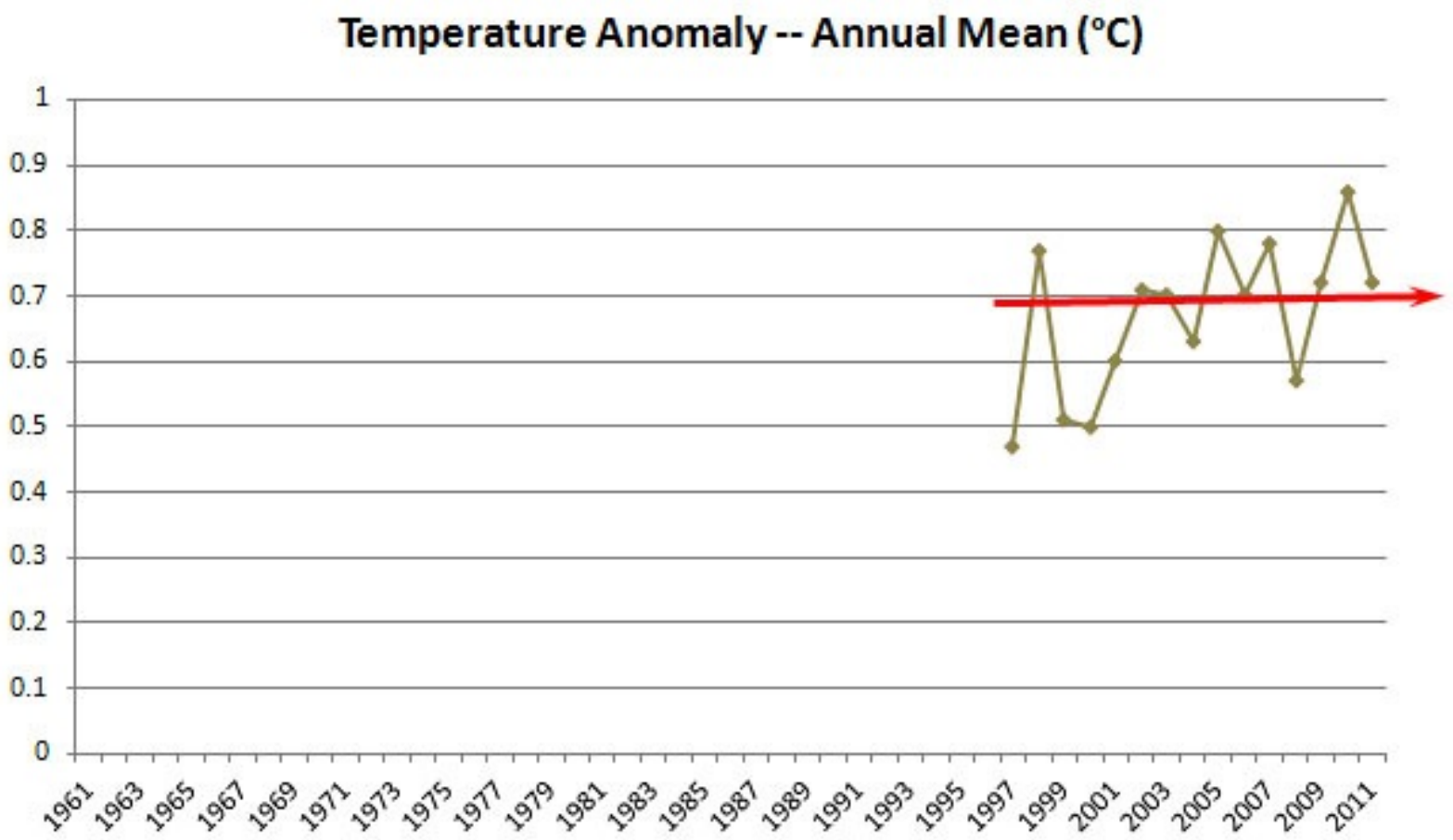
A. Kriebel, VizWiz

Global Warming?



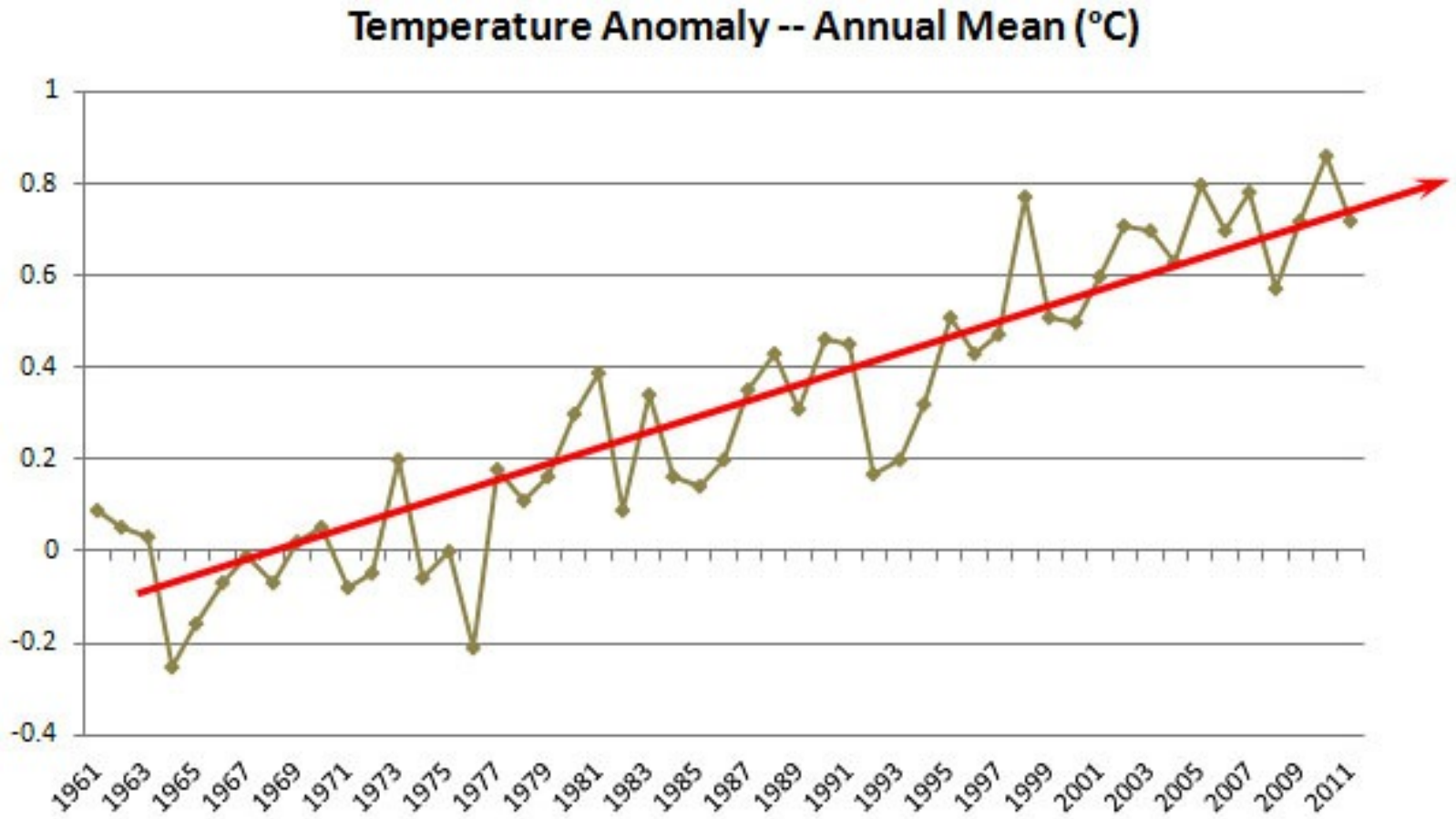
The Daily Mail, UK, Jan 2012

Global Warming?



Mother Jones

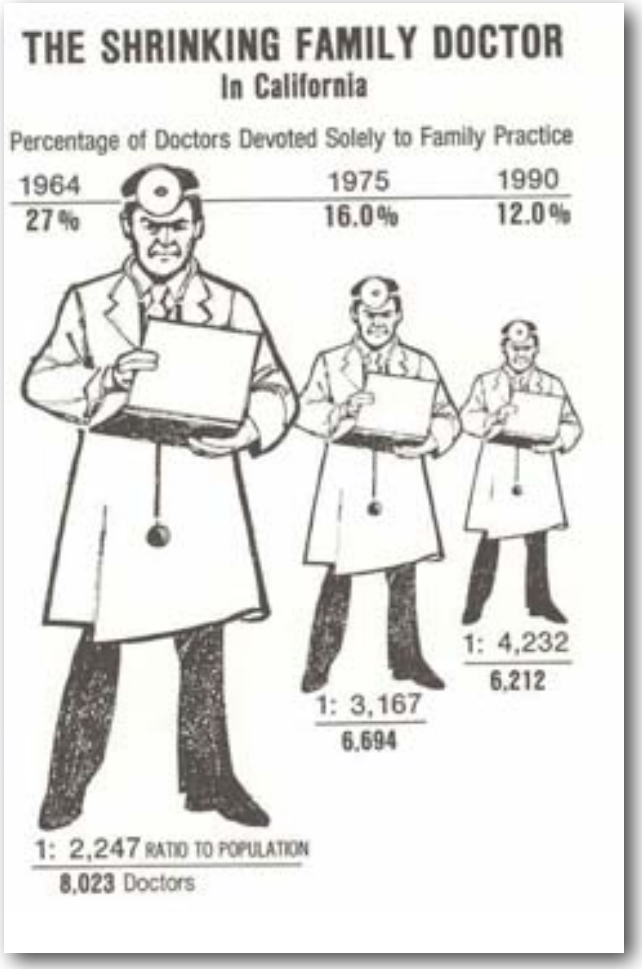
Global Warming!

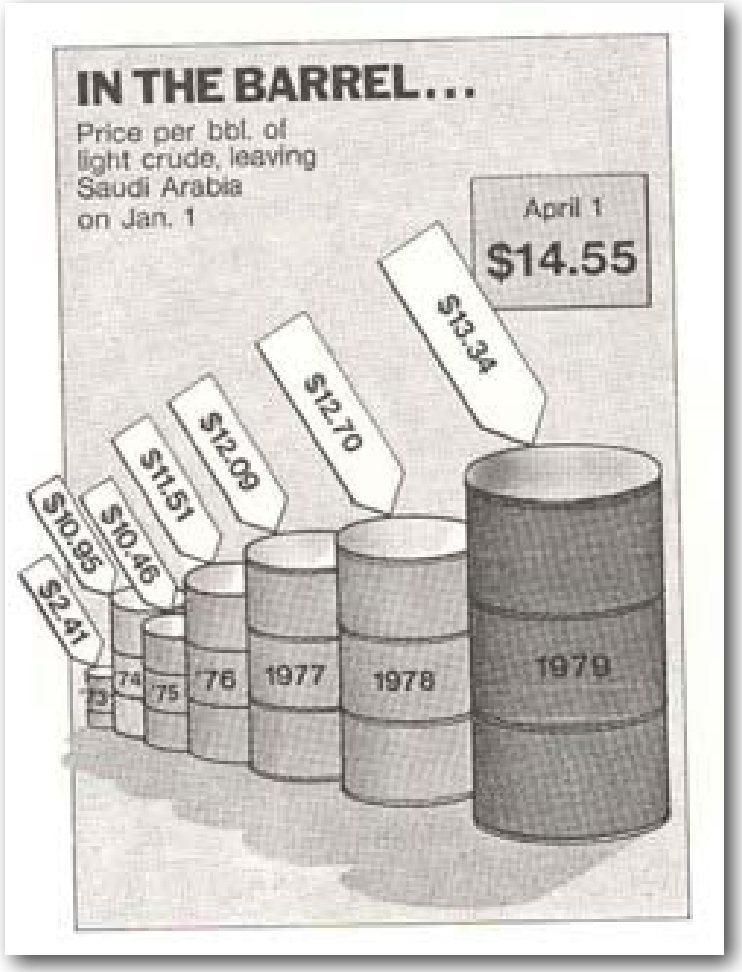


Mother Jones

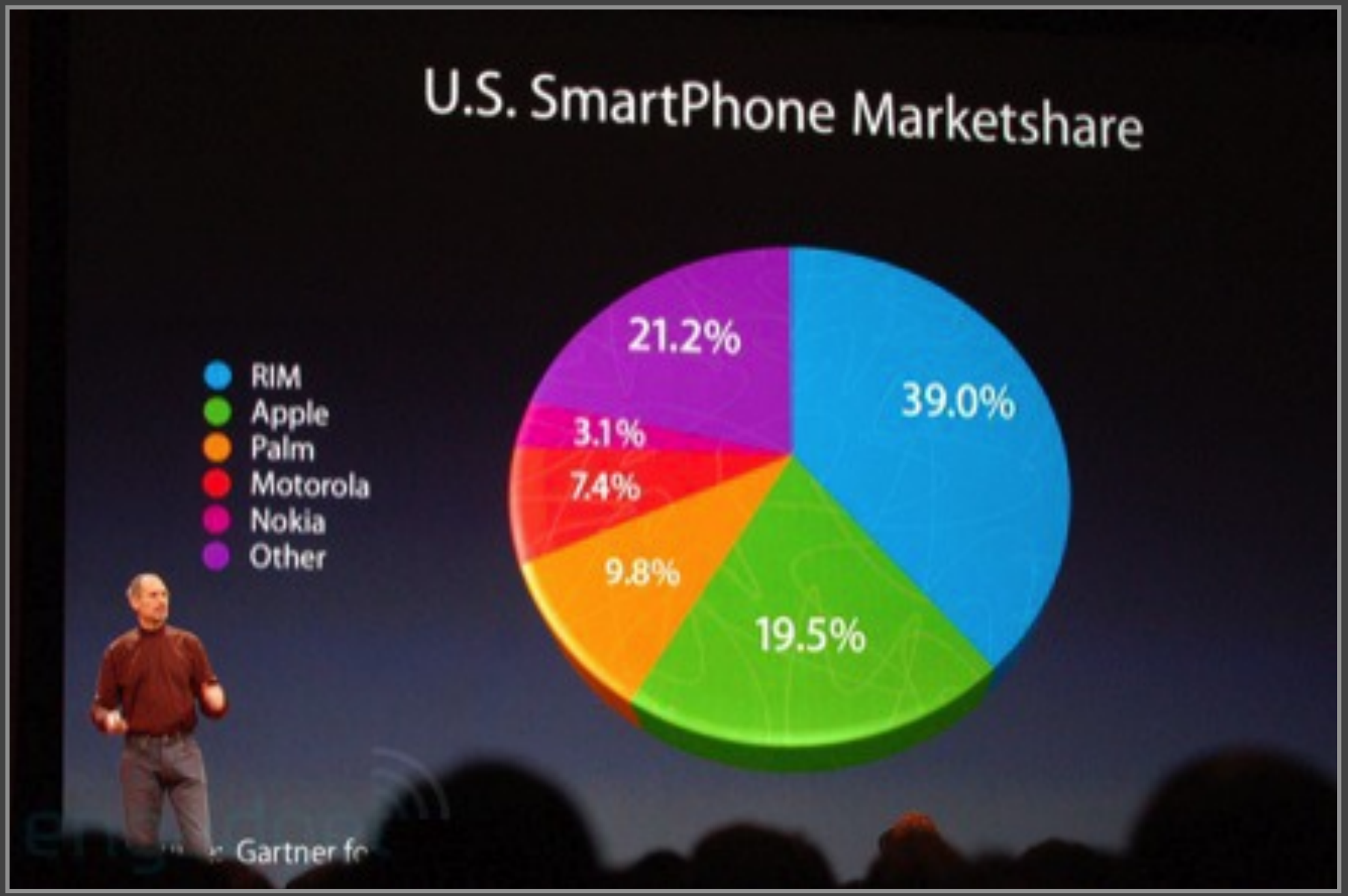
The Lie Factor

$$\frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$

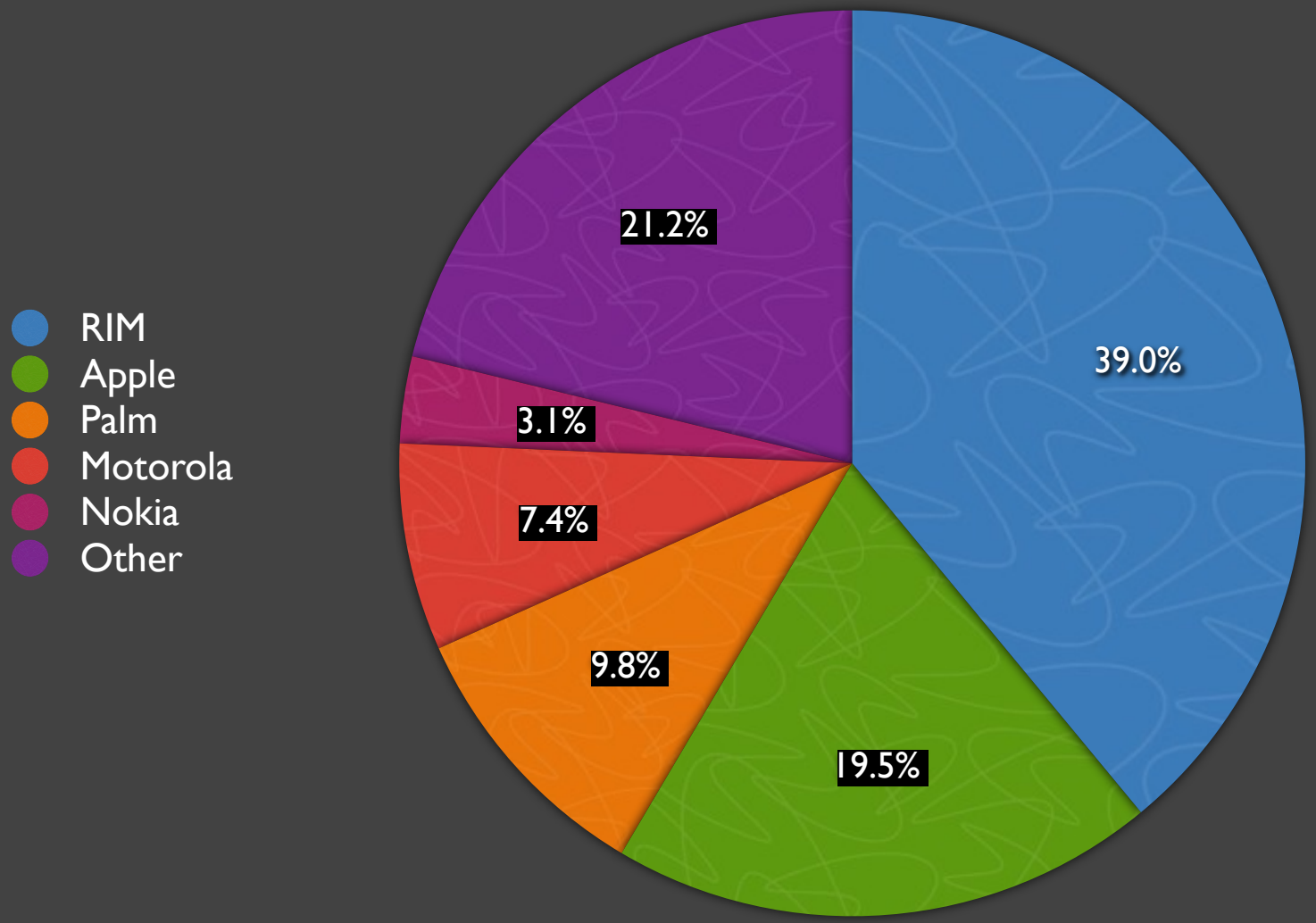




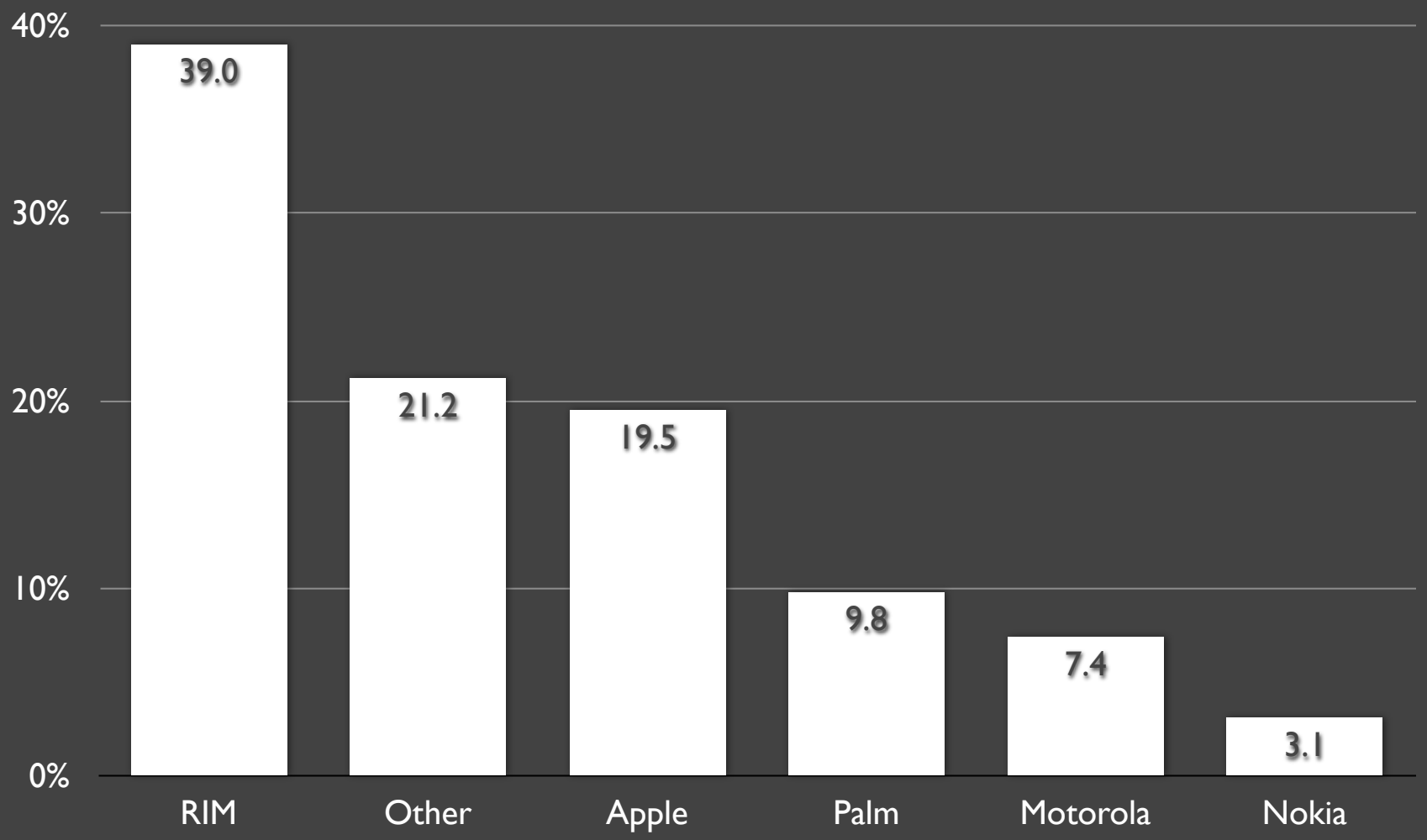
Tufte, VDQI


U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
39.0%
19.5%
9.8%
7.4%
3.1%
21.2%
Gartner fo

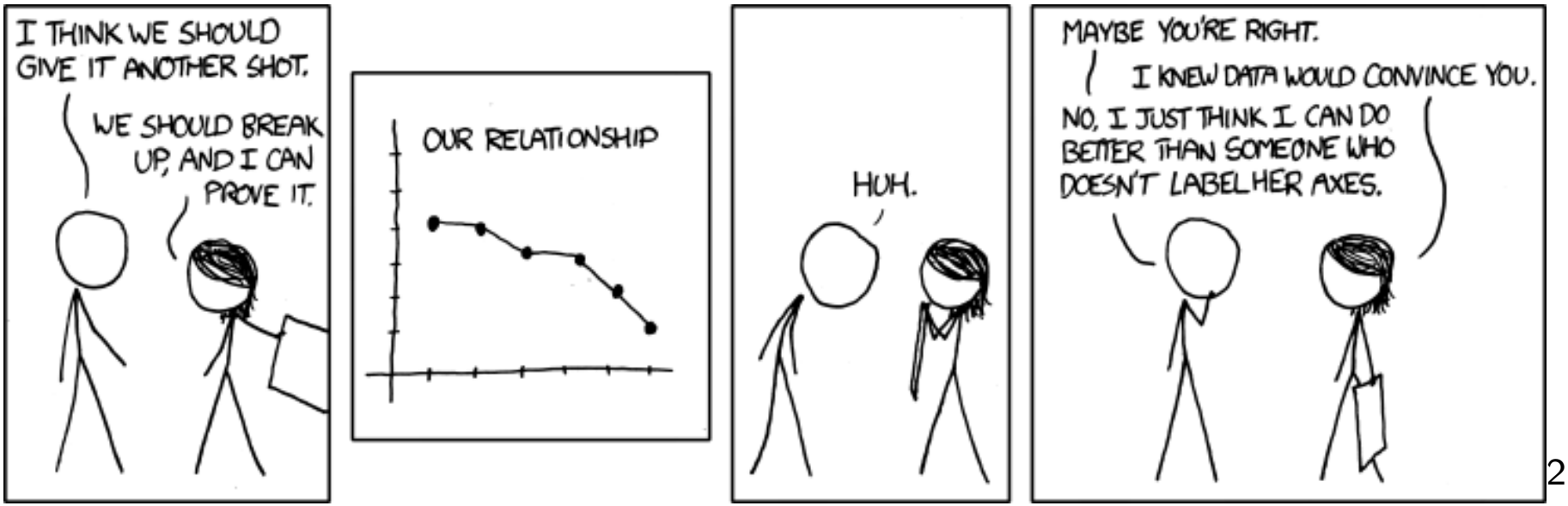
U.S. SmartPhone Marketshare



U.S. SmartPhone Marketshare



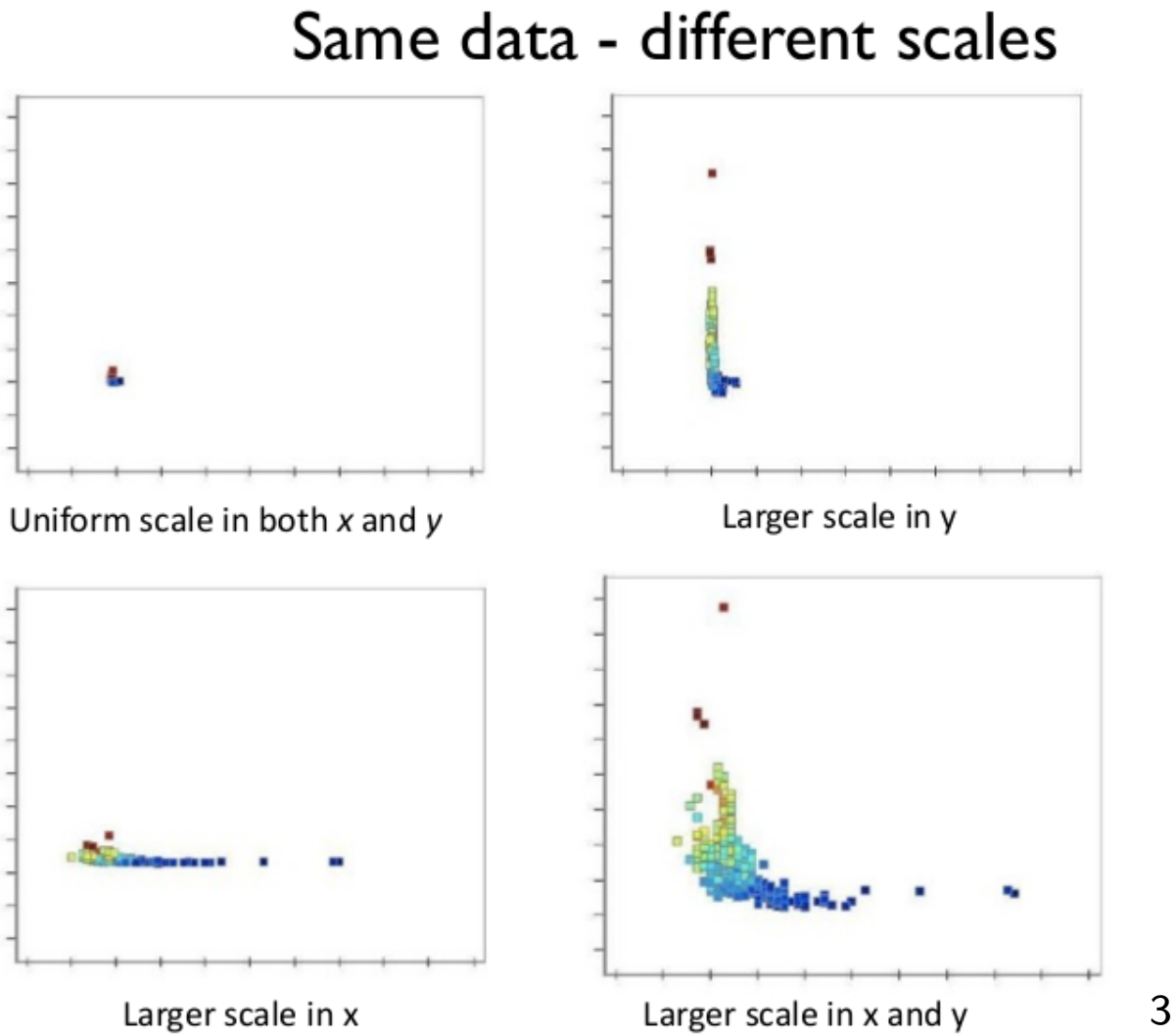
Labelling Chart Axes



[2]

[2] http://xkcd.com/833/

Lying with Scales



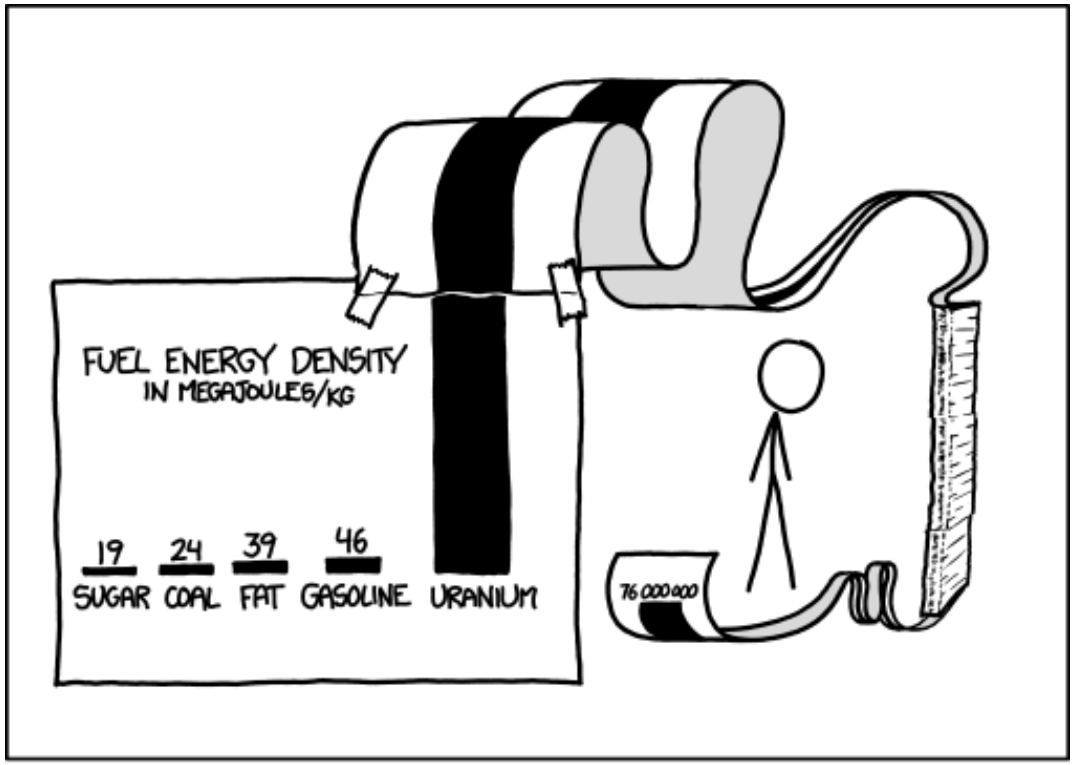
[3]

[3] Ward, Grinstein, Keim, 2011

Scales are Critical!

- What are your bounds upper and lower?
- What scale works? - Linear? Log? Clipping? Breaks?
- Relative or absolute values?
- How can you make things comparable?

Log Scale



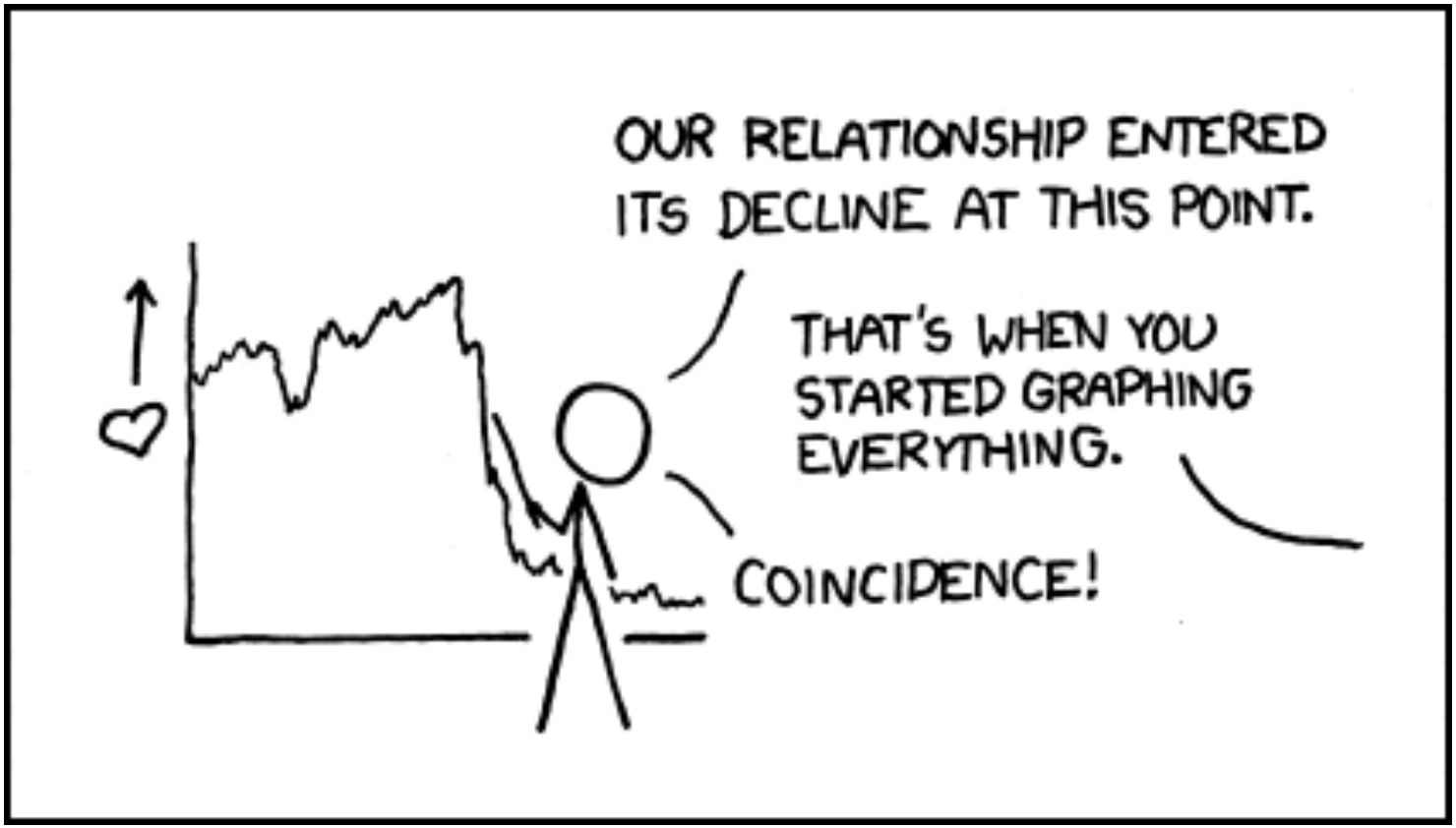
SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

4

[4] http://xkcd.com/1162/

Graph Types (2D and nD)

Statistical Graph Types

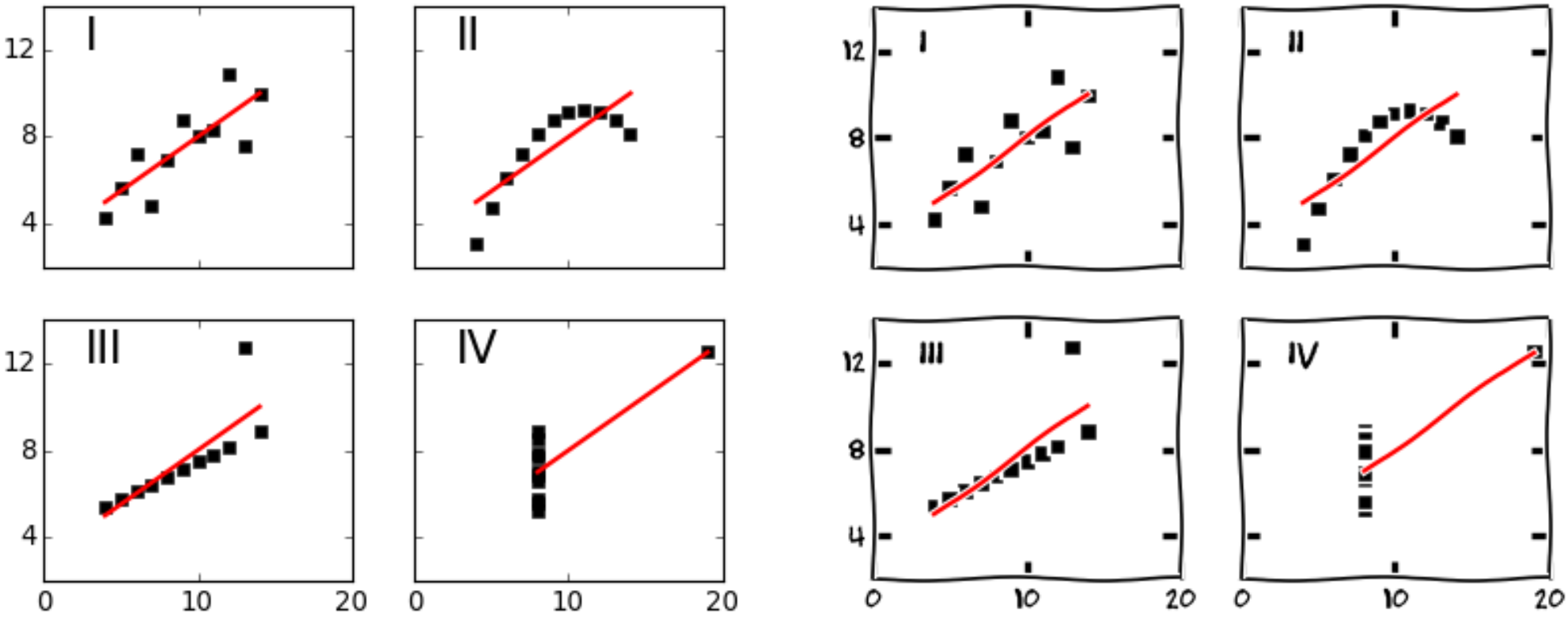


http://xkcd.com/523/

Side Note

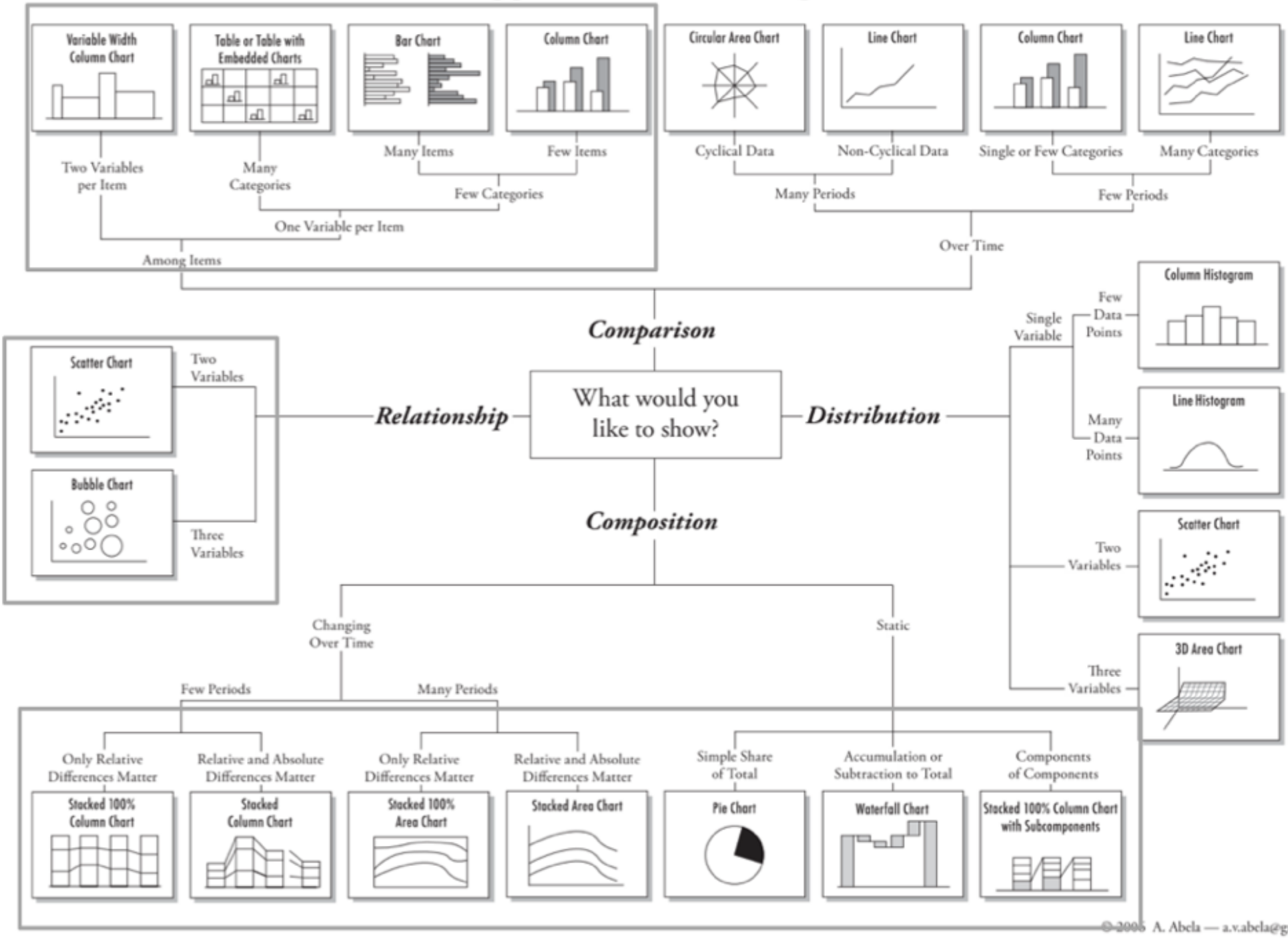
MATPLOTLIB

XKCD-ify your plot



http://matplotlib.org/xkcd

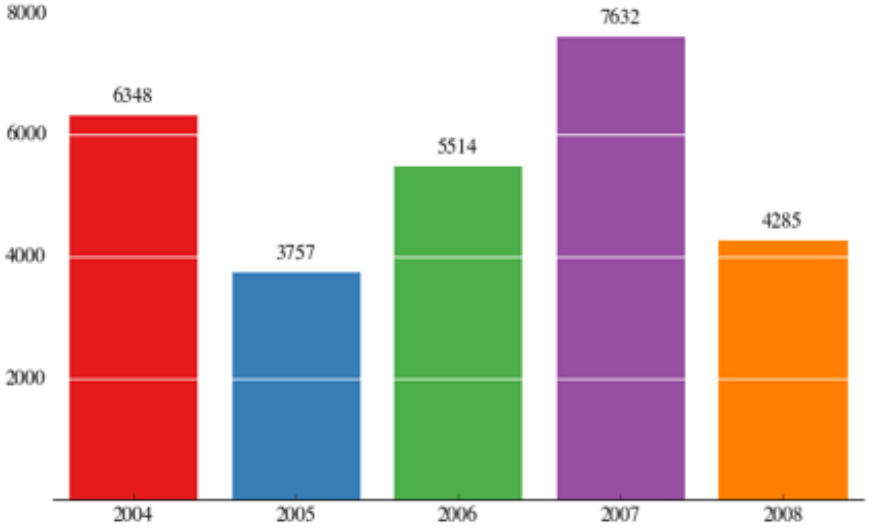
Chart Suggestions—A Thought-Starter



http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

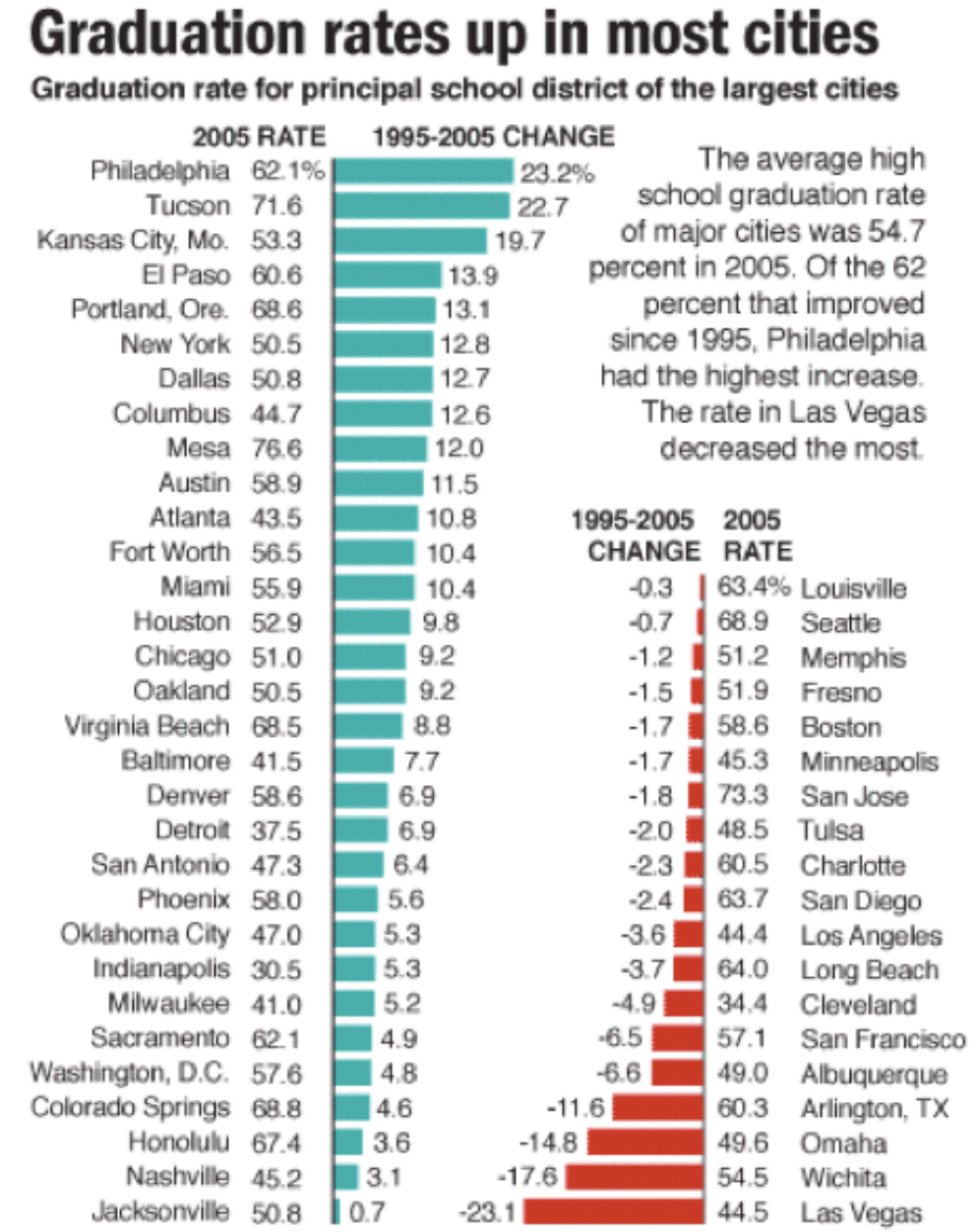
Comparisons

Bar Chart



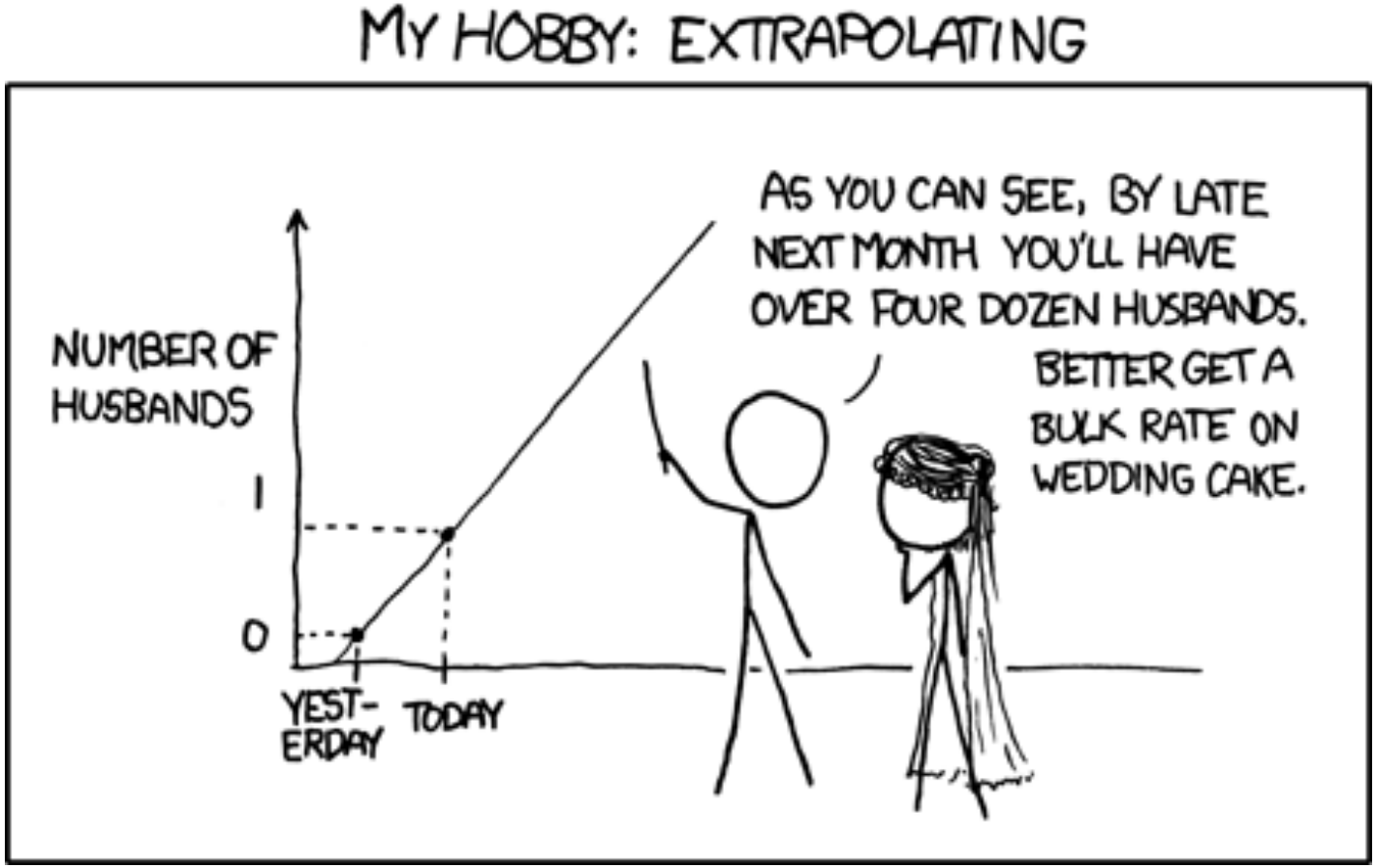
http://nbviewer.ipython.org/gist/olgabot/5357268

Direction



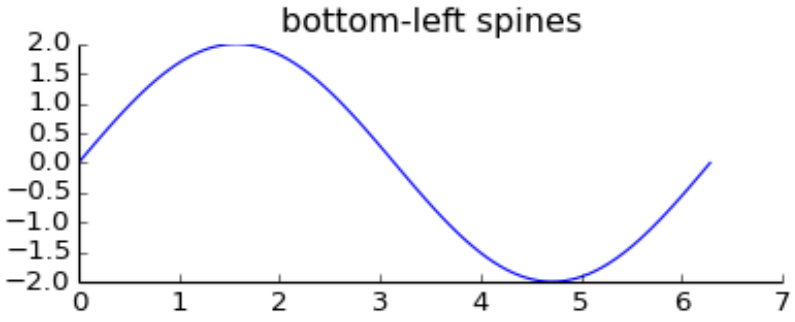
Nicolas Rapp

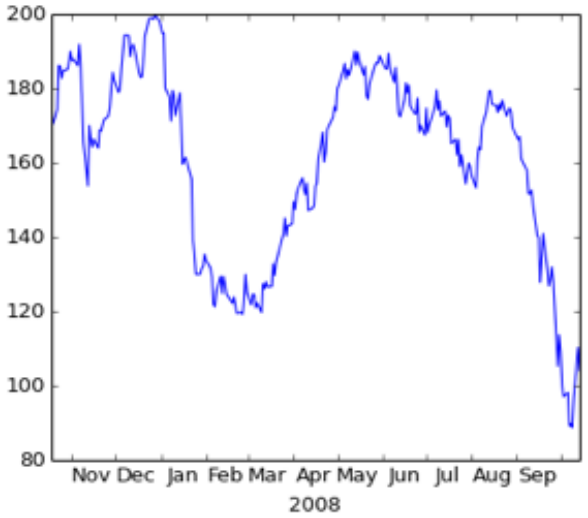
Trends Over Time

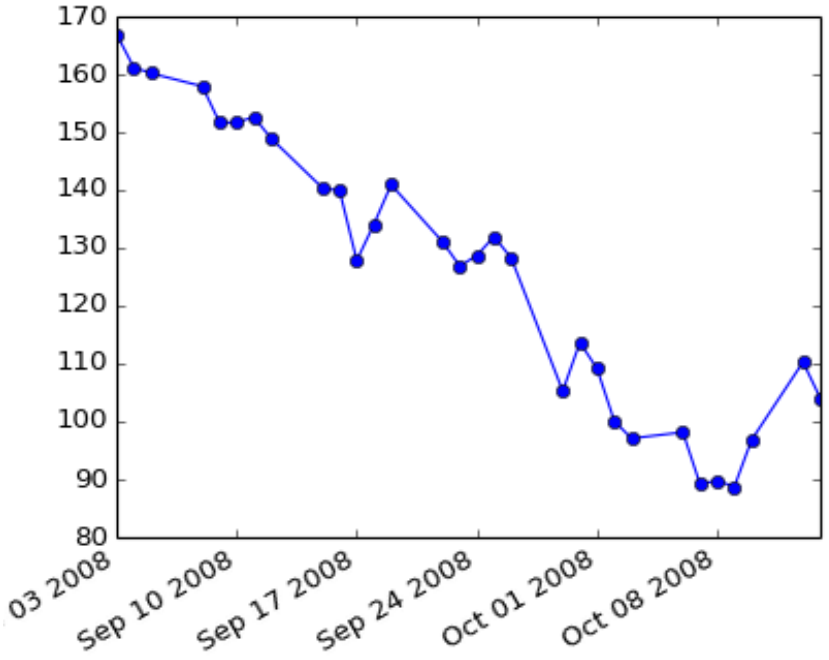


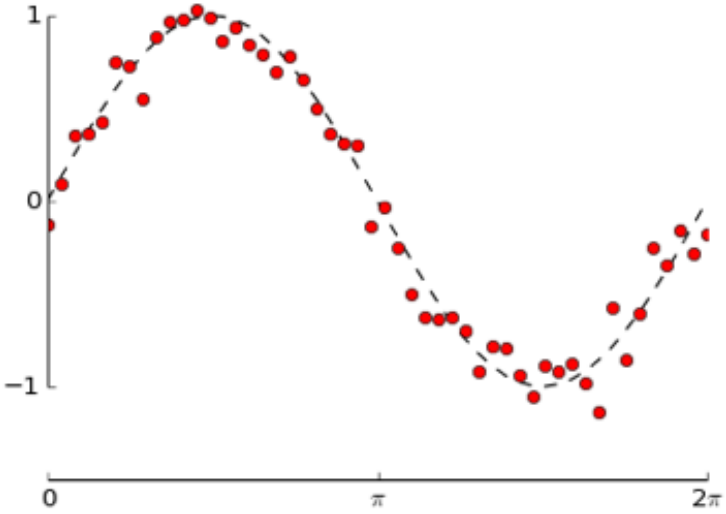
http://xkcd.com/605/

Line Charts



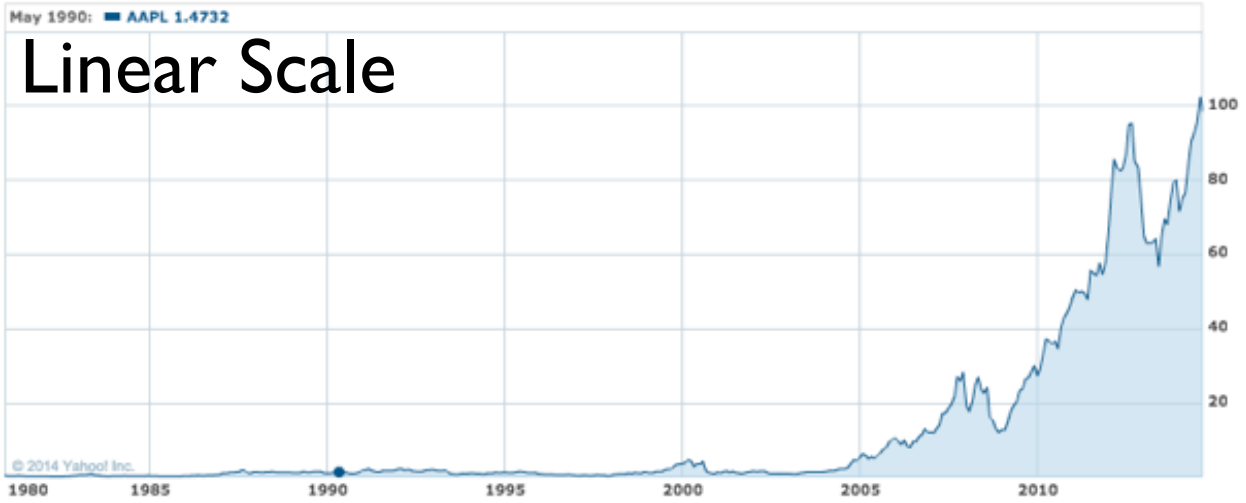


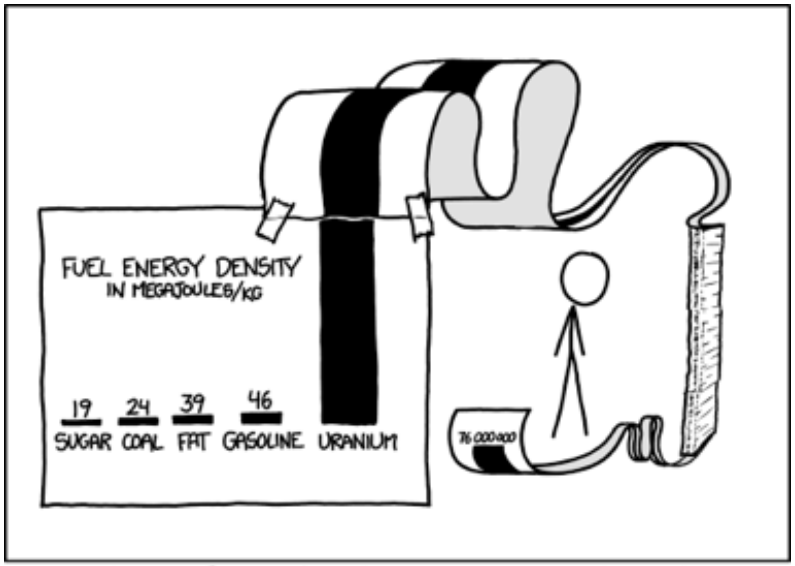




matplotlib gallery

Linear vs. Logarithmic Scale





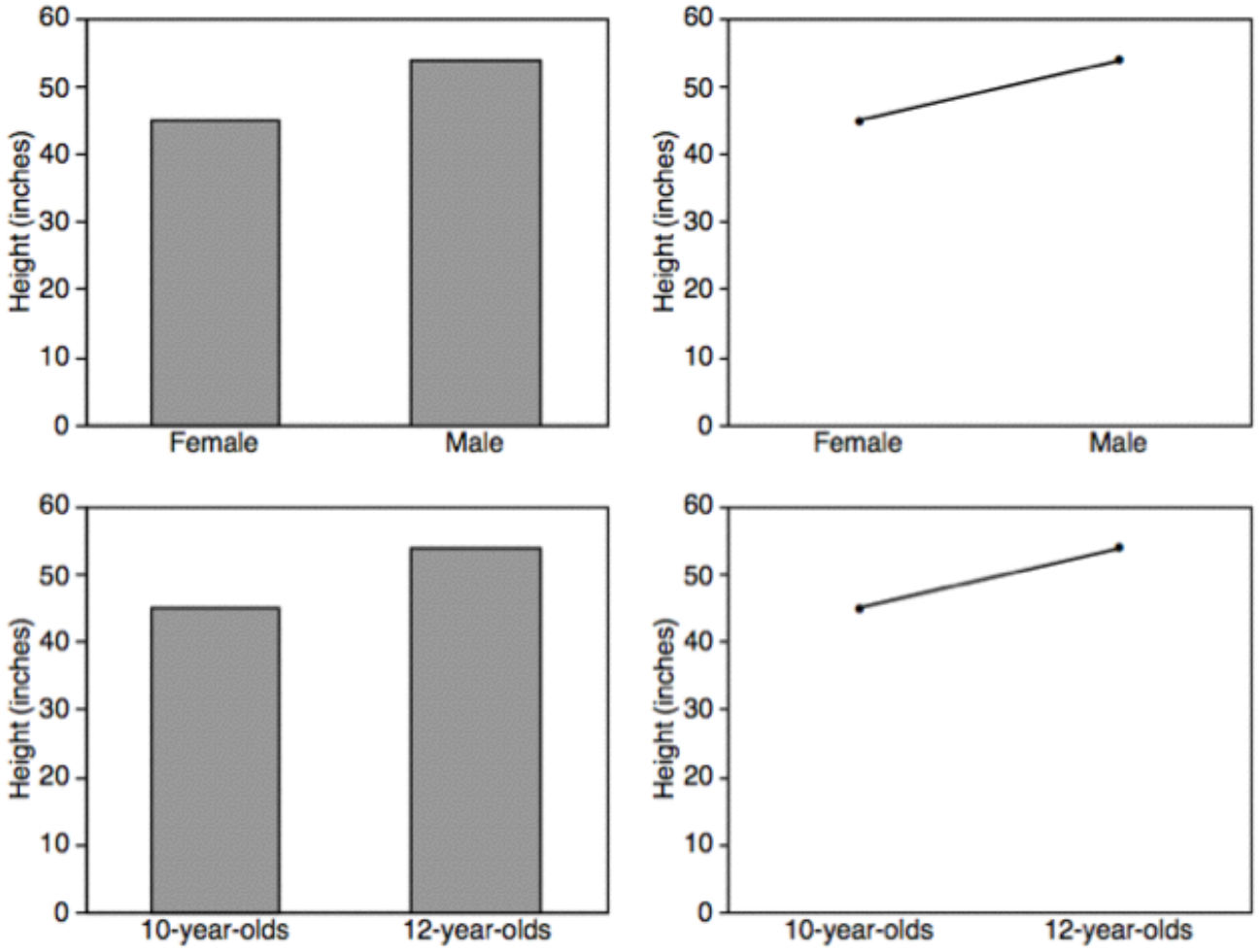
SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

http://xkcd.com/1162/

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

Bars vs. Lines

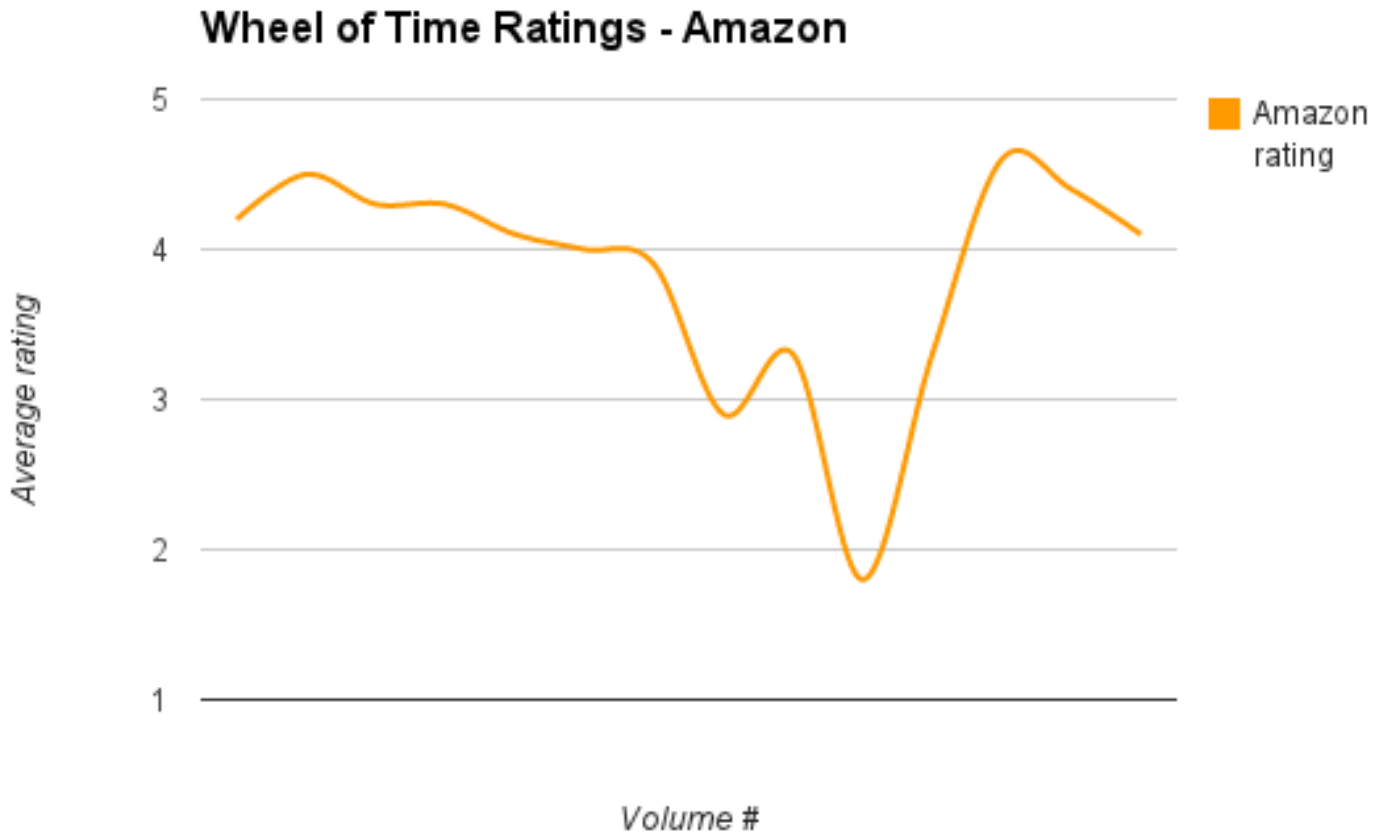
Lines imply connections - do not use for categorical data



Zacks 1999

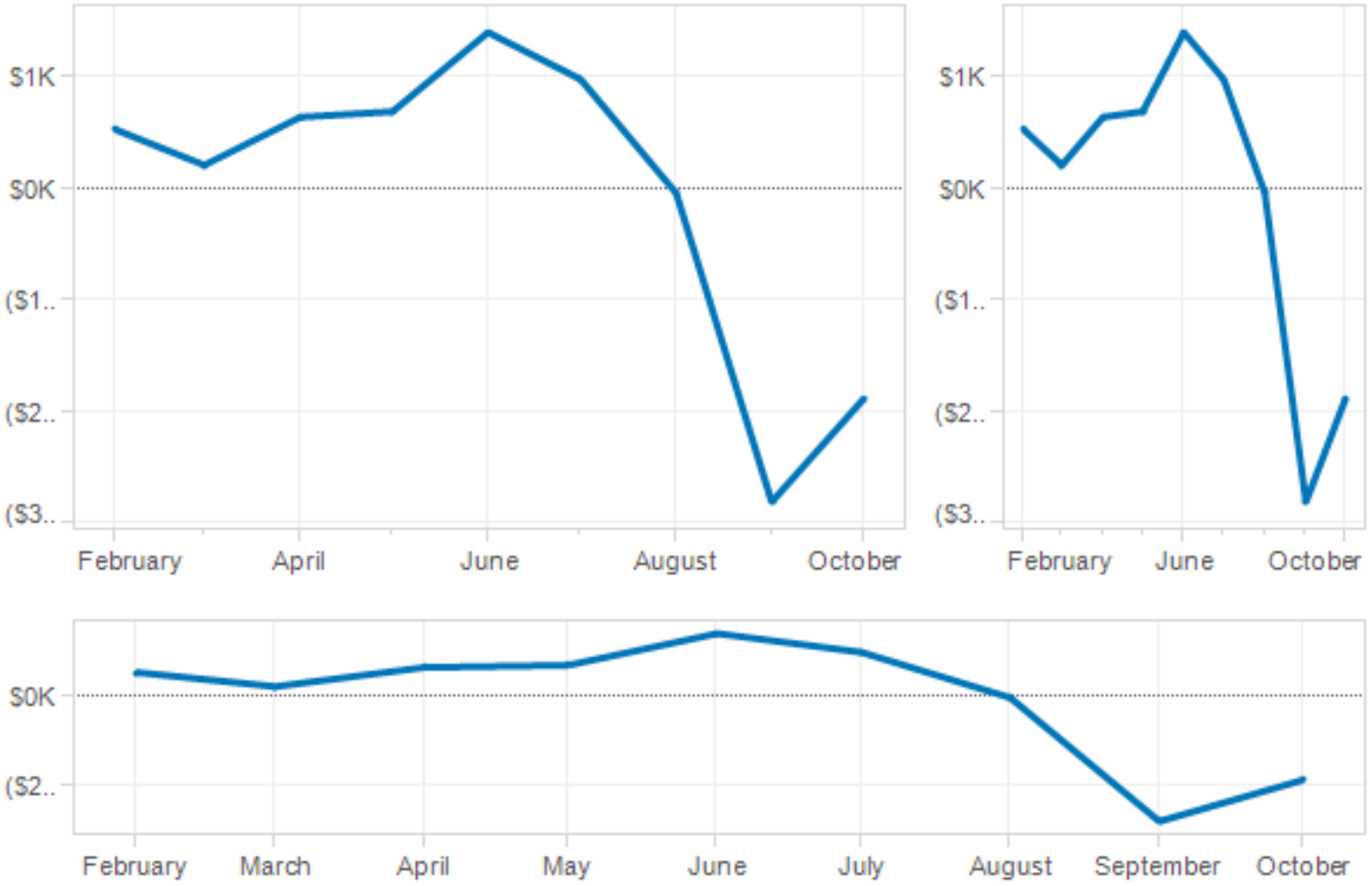
Don’t

Use bar charts to compare book ratings



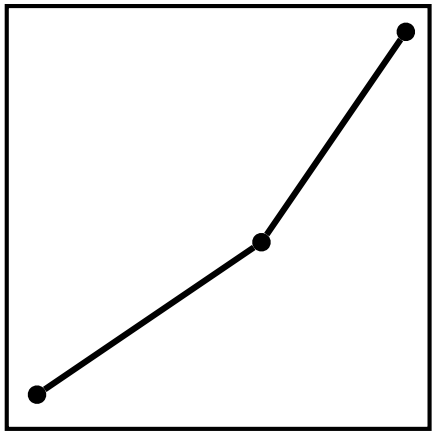
“Visualizing The Wheel of Time: Reader Sentiment for an Epic Fantasy Series”, J. Siddle, Sept 2013

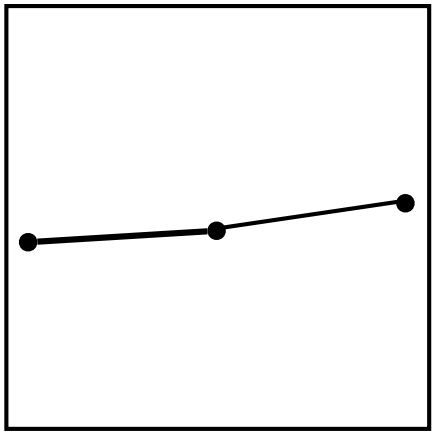
Aspect Ratios

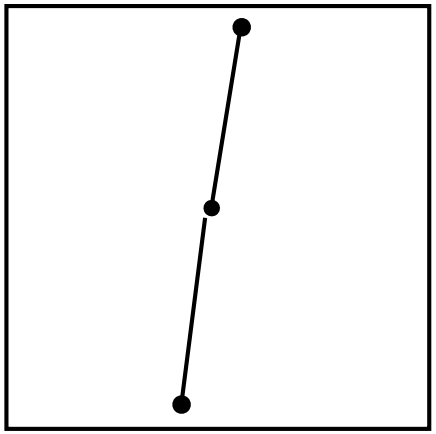


Banking to 45°

Two line segments are maximally discriminable when their average absolute angle is 45° W. Cleveland

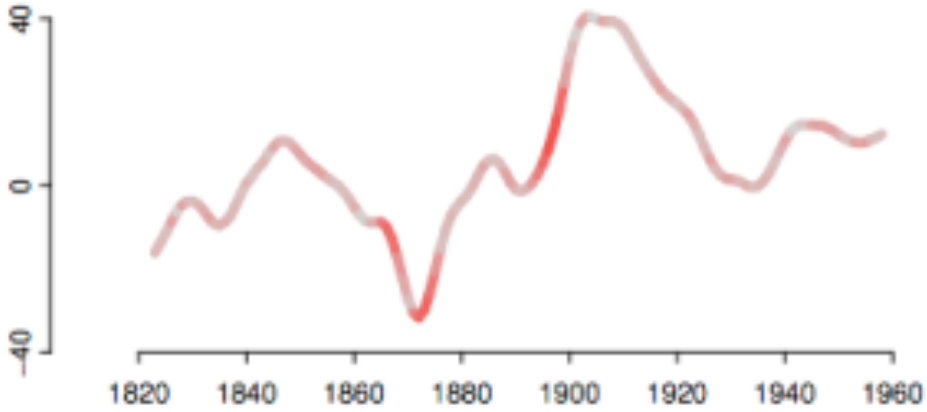




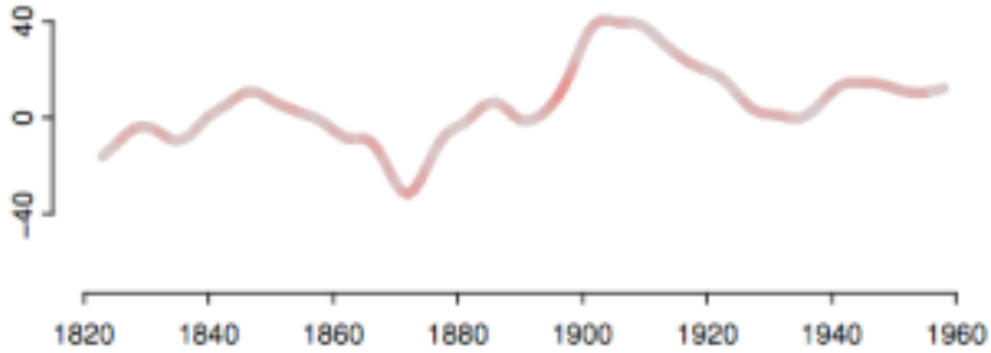


Slide based on J. Heer / M. Agrawala

Banking to 45°



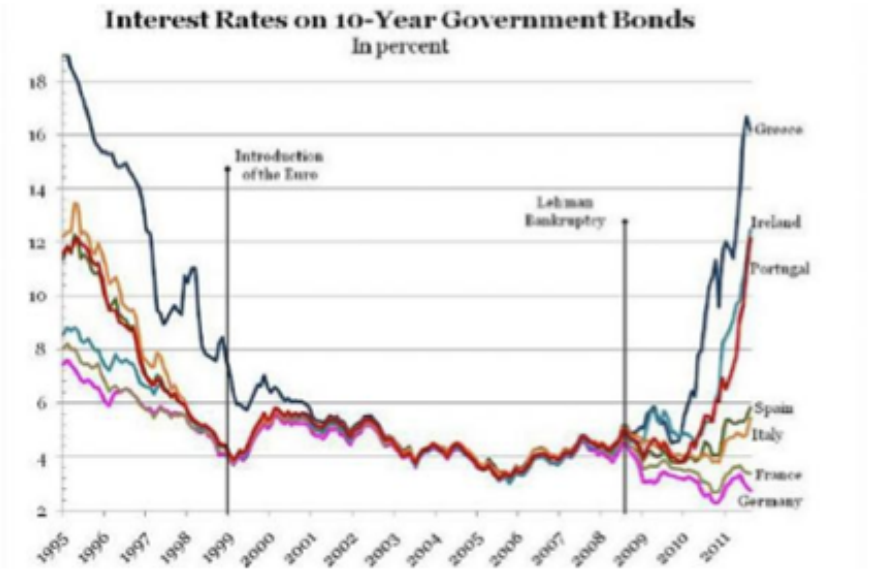
Error Prone

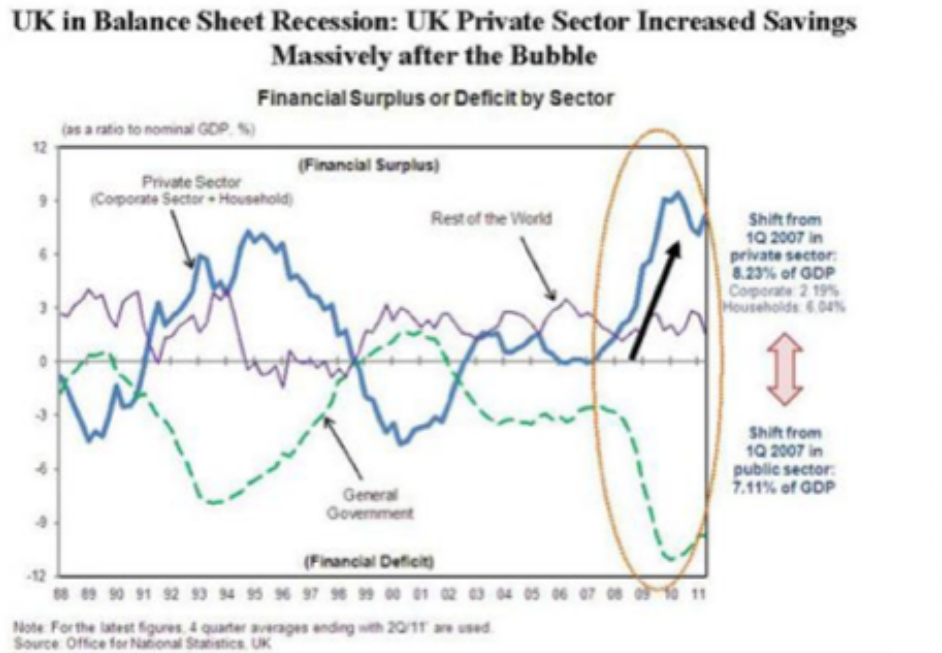


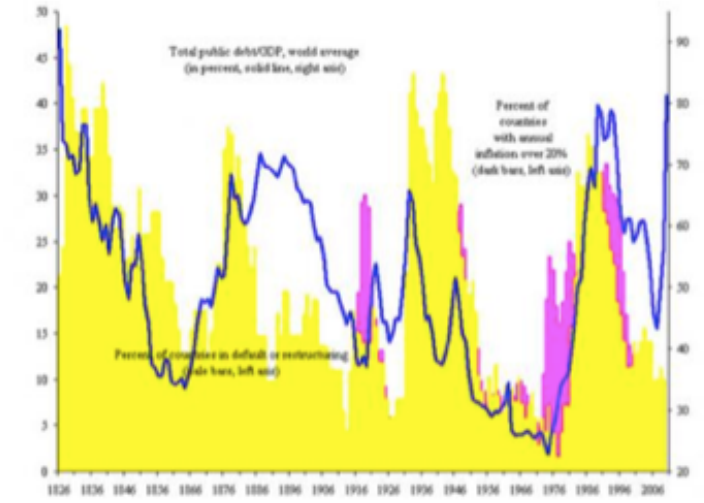
Optimal Aspect Ratio

An Empirical Model of Slope Ratio Comparisons,
Justin Talbot, John Gerth, Pat Hanrahan

Don’t

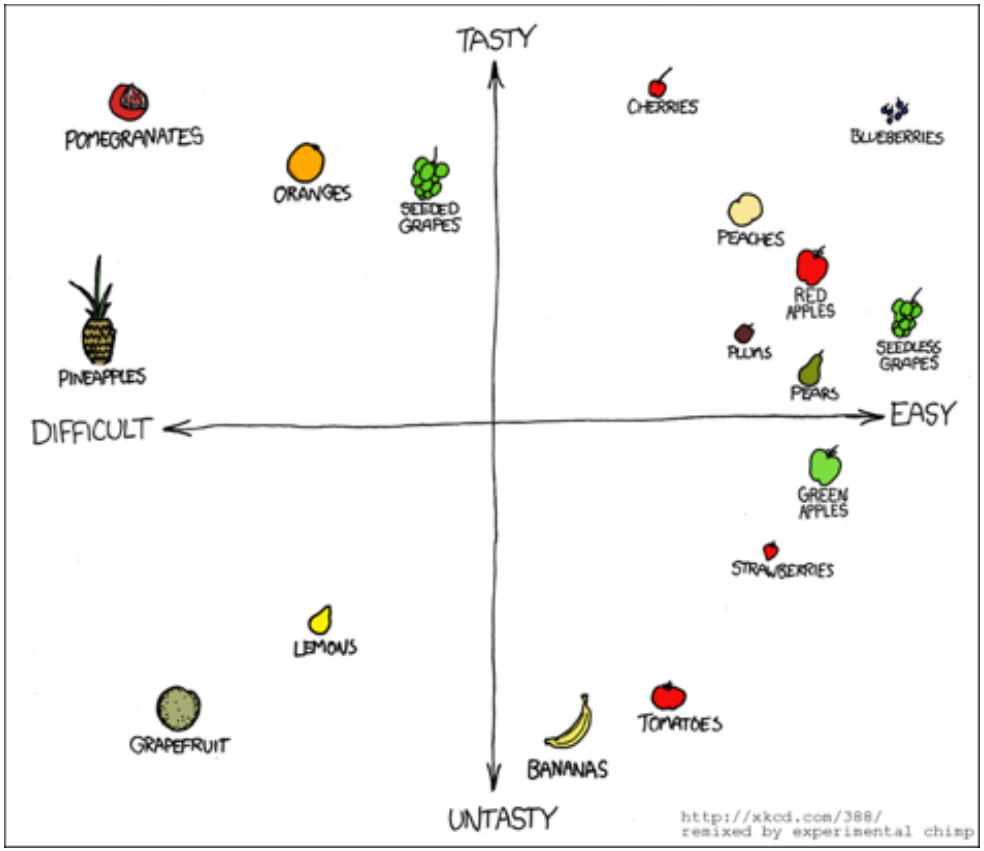

Interest Rates on 10-Year Government Bonds
In percent
Introduction of the Euro
Lehman Bankruptcy
Greece
Ireland
Portugal
Spain
Italy
France
Germany

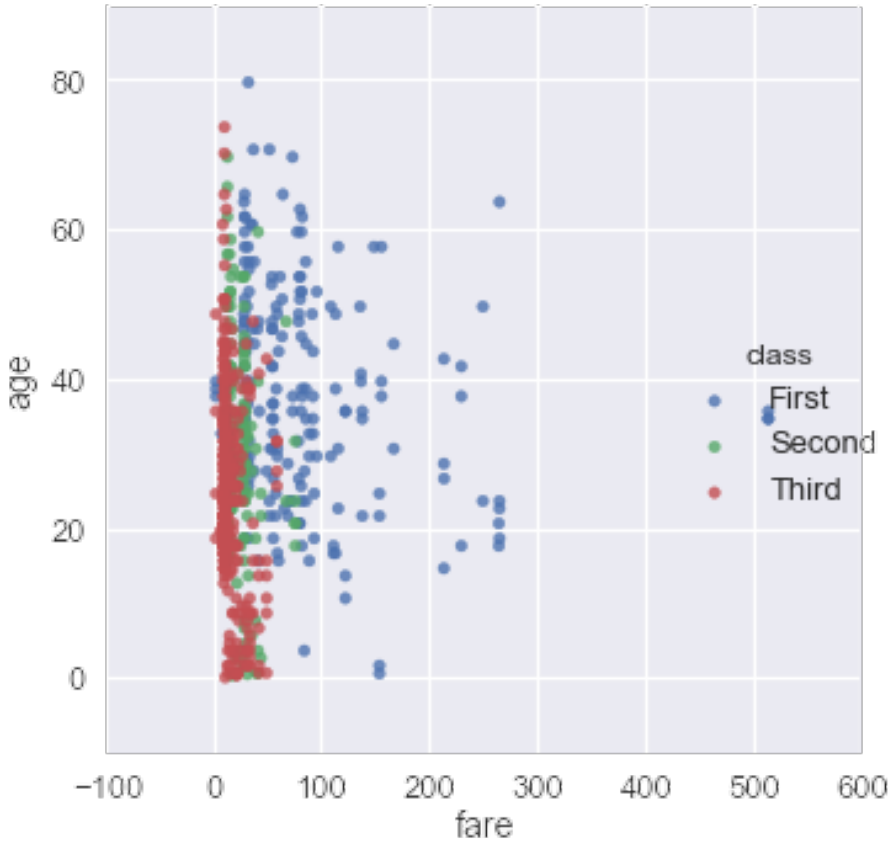

UK in Balance Sheet Recession: UK Private Sector Increased Savings Massively after the Bubble
Financial Surplus or Deficit by Sector
(as a ratio to nominal GDP, %)
(Financial Surplus)
Private Sector (Corporate Sector + Household)
Rest of the World
General Government
(Financial Deficit)
Shift from 1Q 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
Percent of countries with annual inflation over 20%

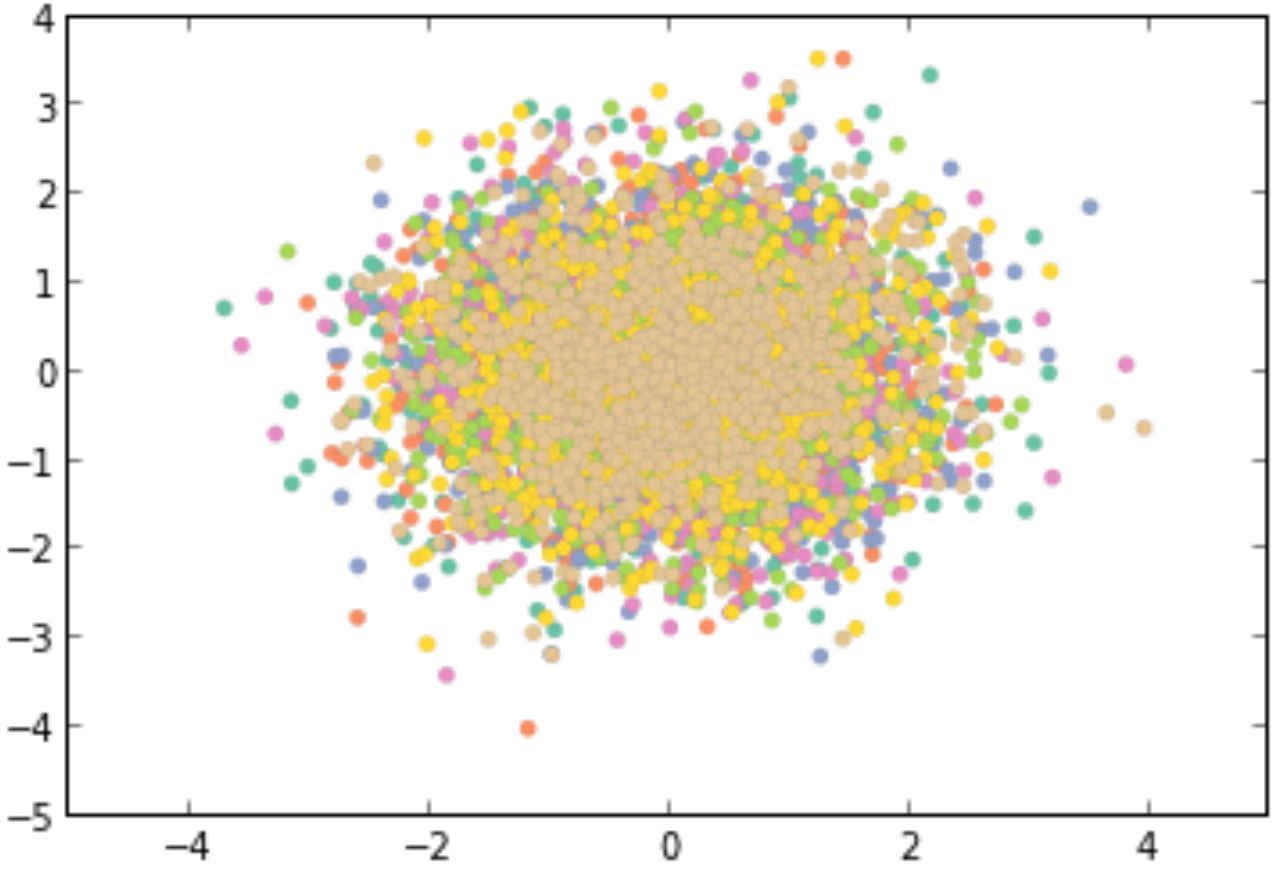
Correlations

Scatterplots





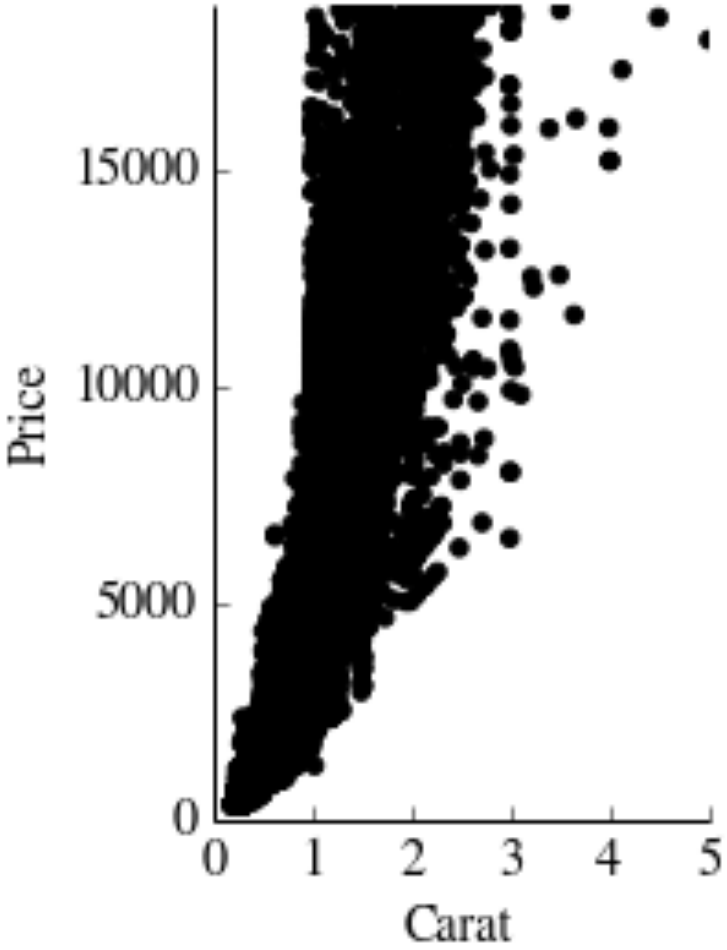
Scatterplots

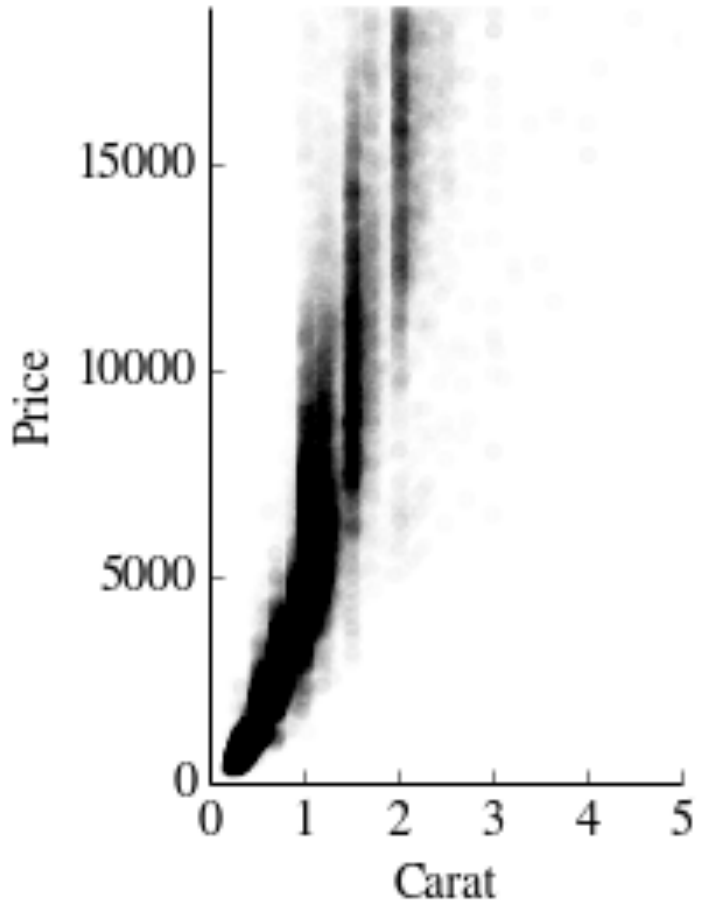


Light Grey Border

http://nbviewer.ipython.org/5357268

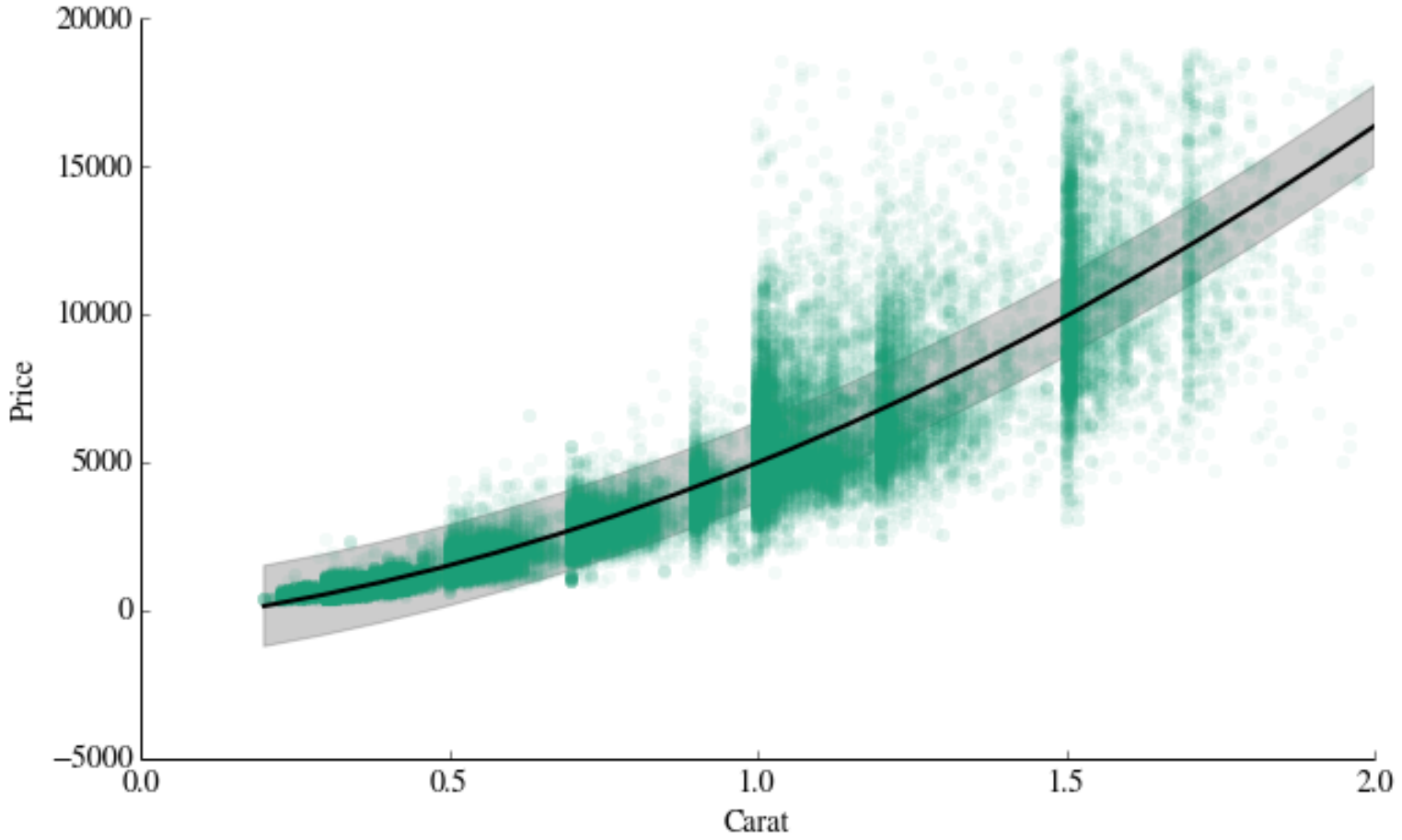
Overplotting



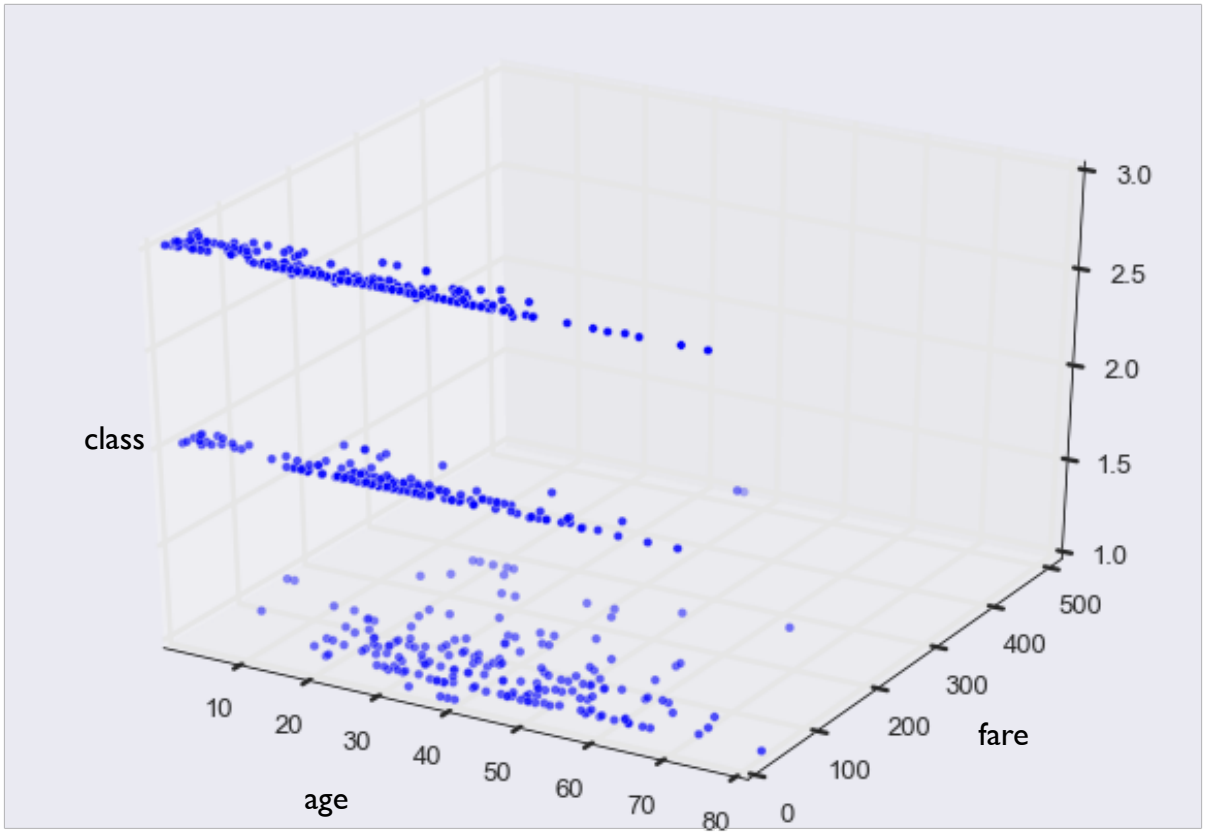


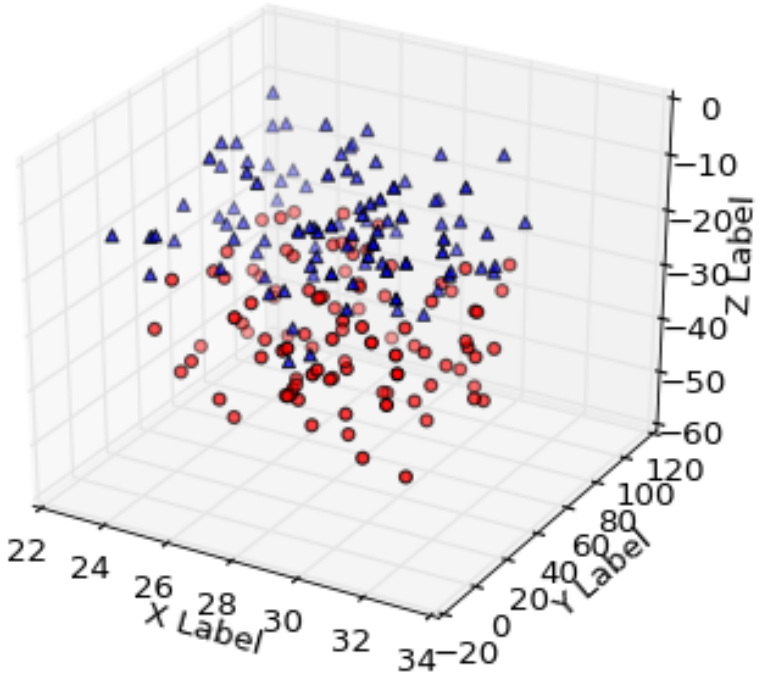
alpha = 1/100

Trend Lines



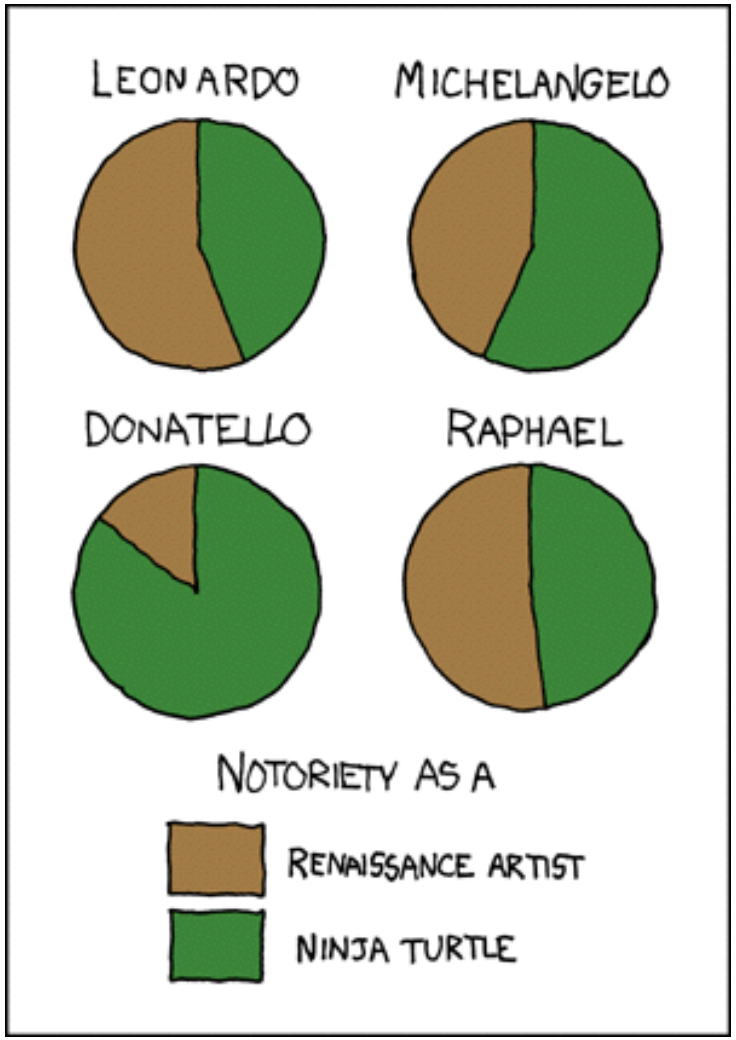
Don‘t


class
3.0
2.5
2.0
1.5
1.0
10
20
30
40
50
60
70
80
age
0
100
200
300
400
500
fare

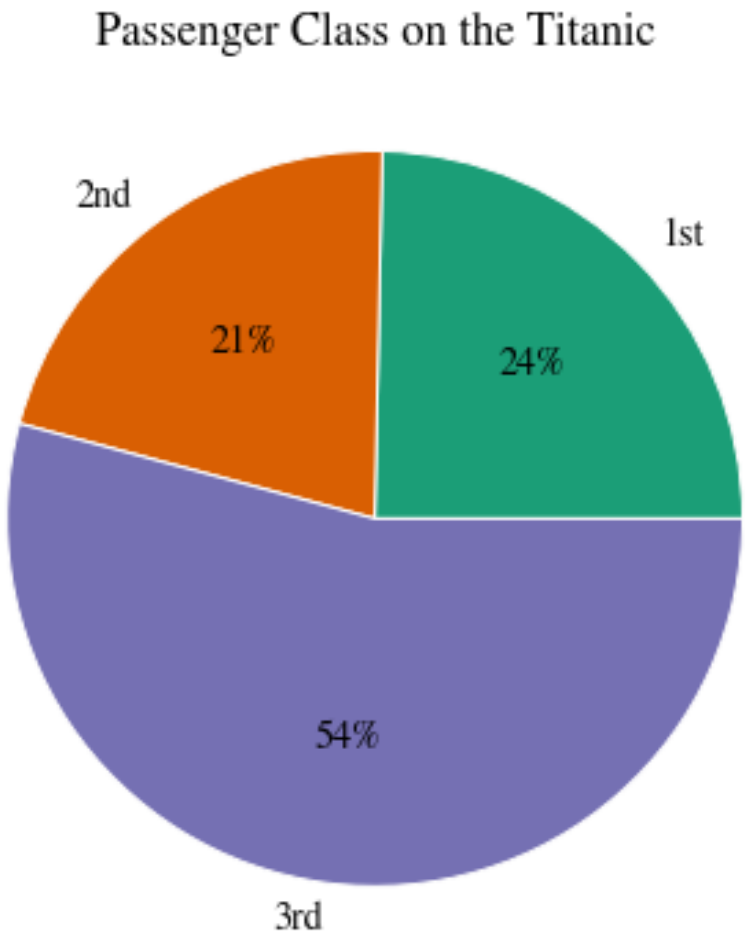

22
24
26
28
30
32
34
X Label
0
20
40
60
80
100
120
Y Label
0
−10
−20
−30
−40
−50
−60
Z Label

Compositions

Pie Charts

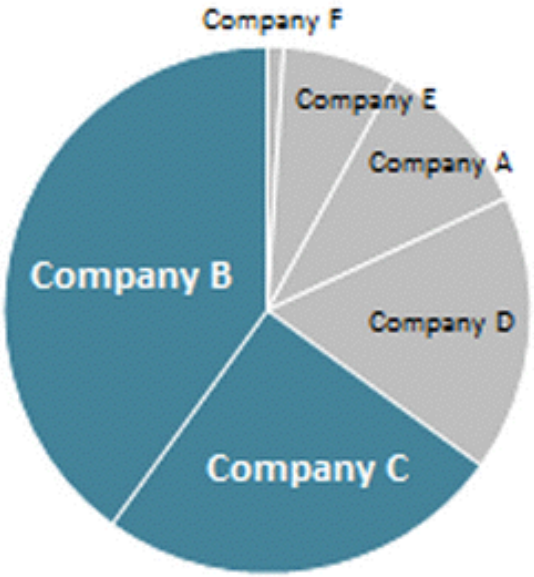


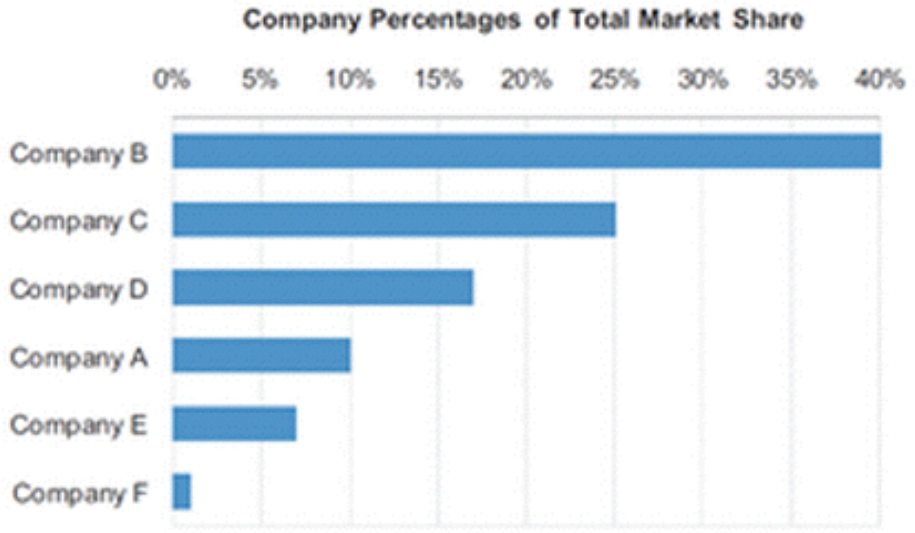
http://xkcd.com/197/



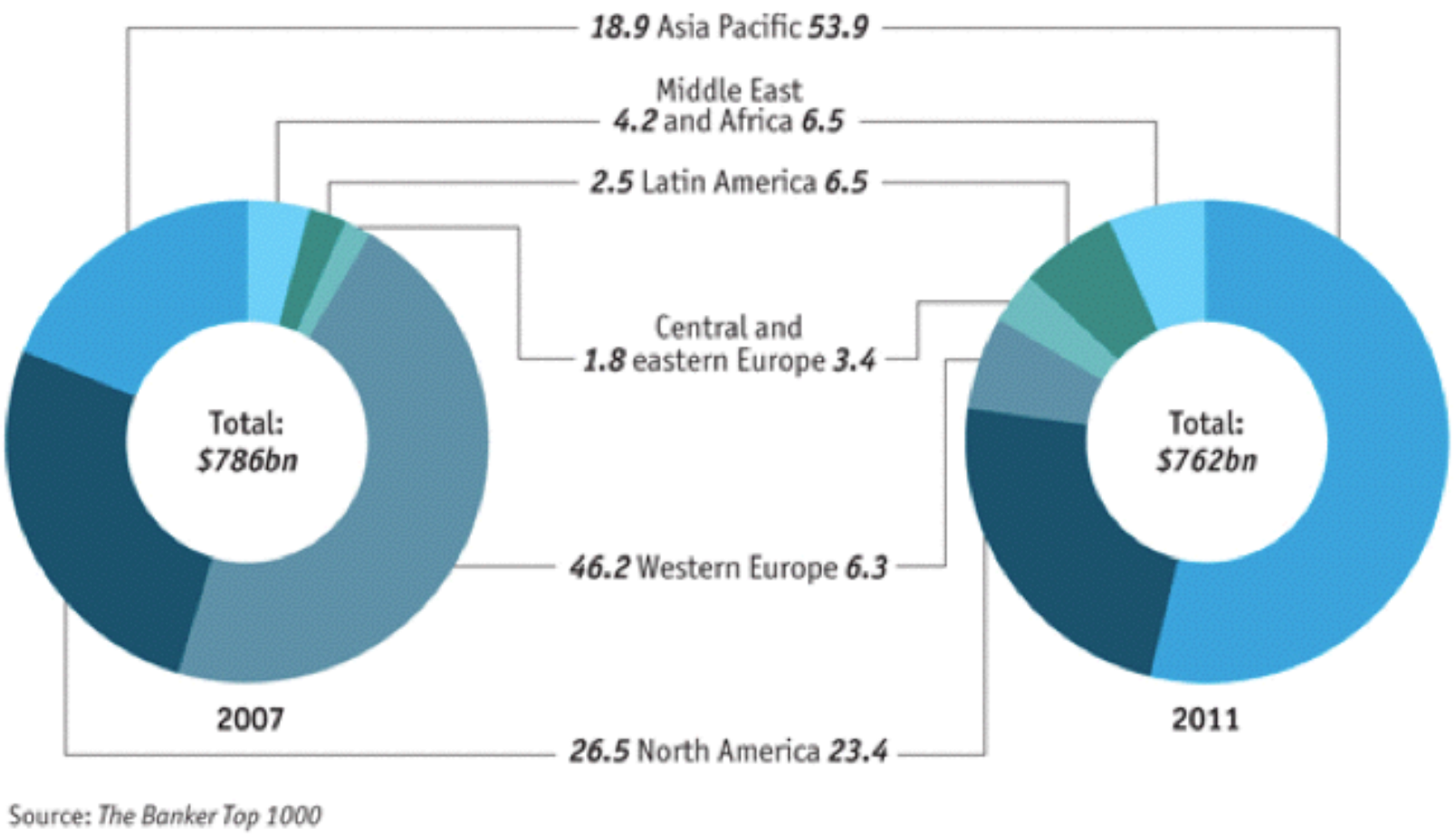
Pie vs. Bar Charts

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

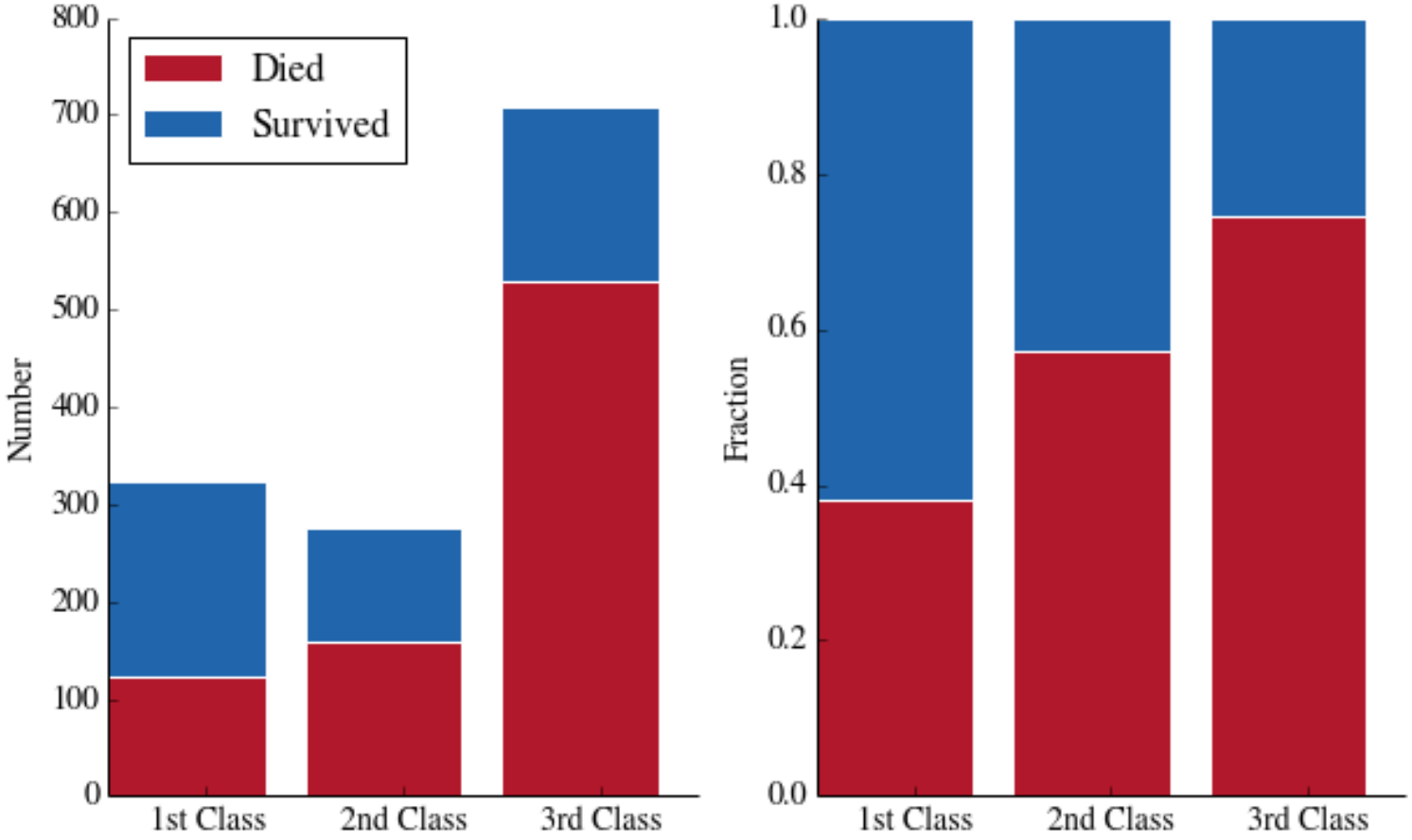




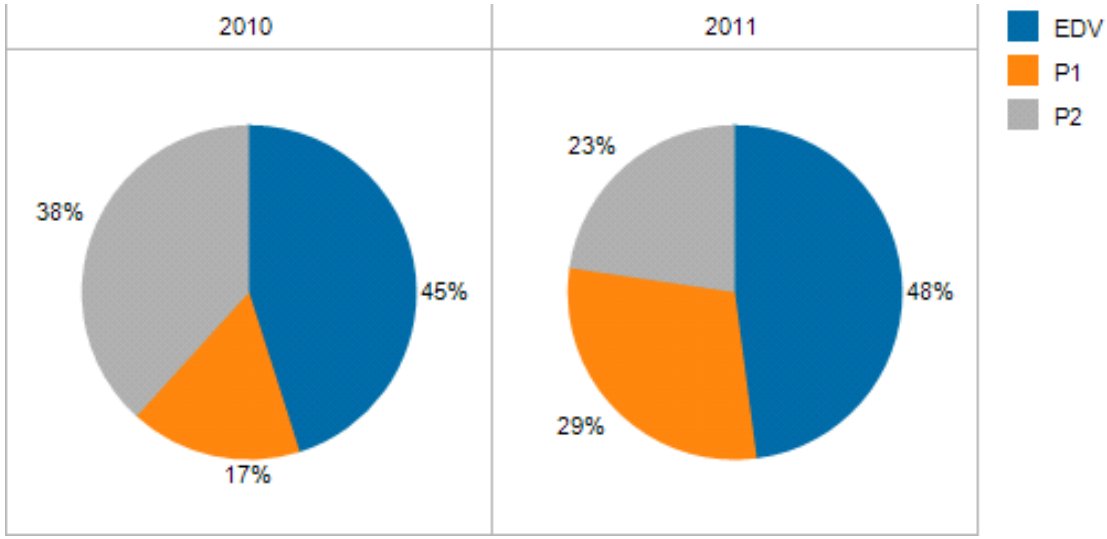
Donut Chart


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

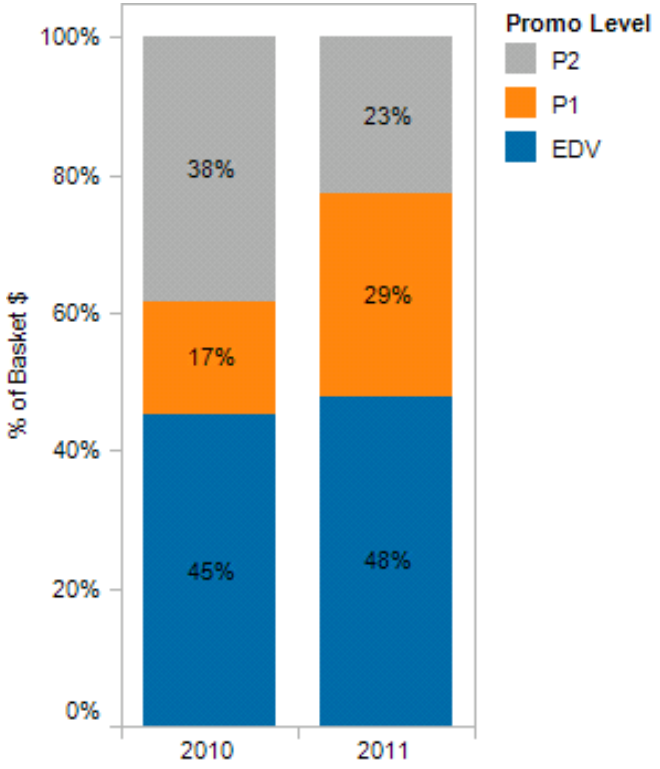
Stacked Bar Chart



Stacked Bar Chart

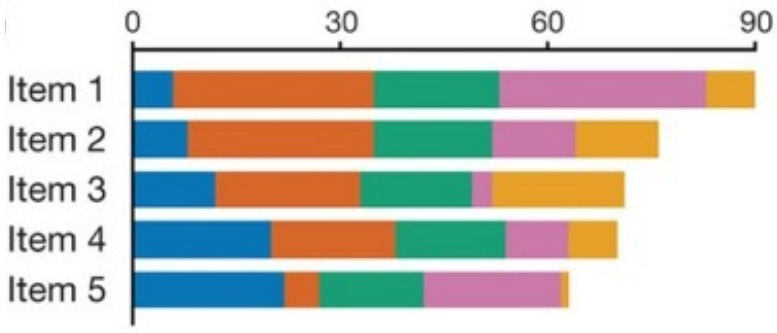


vs.



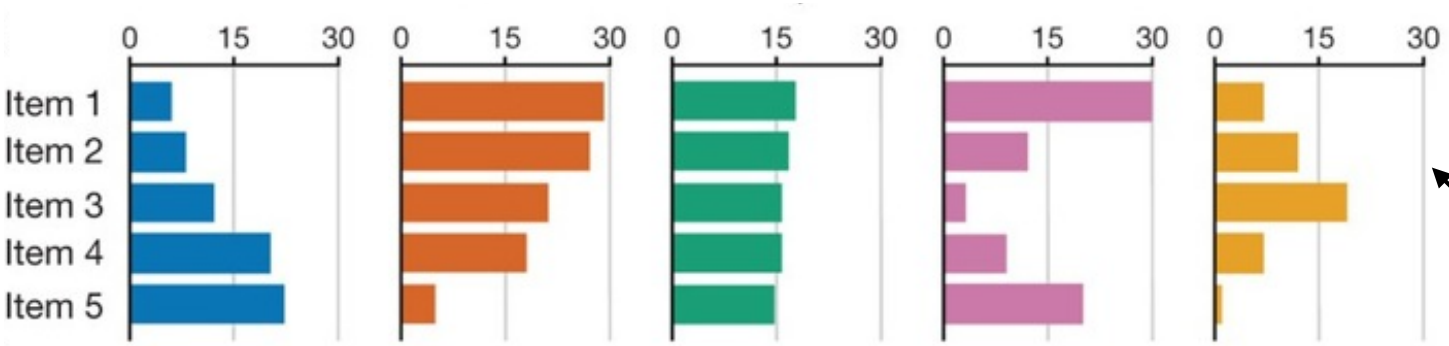
Comparison of bar chart types

Pie Chart



Stacked bar chart

Layered Bar Chart

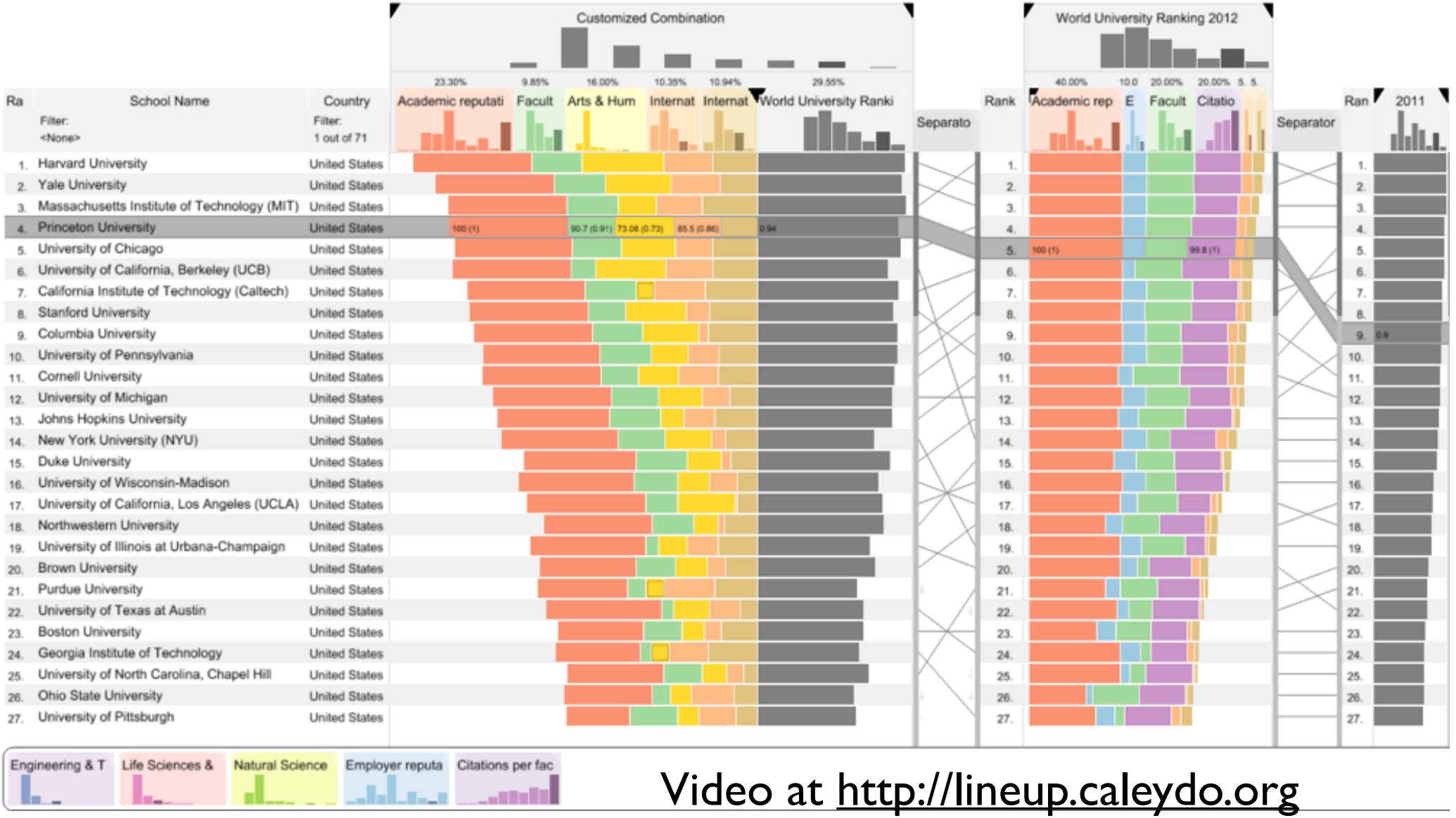


Grouped Bar Chart

Small Multiples

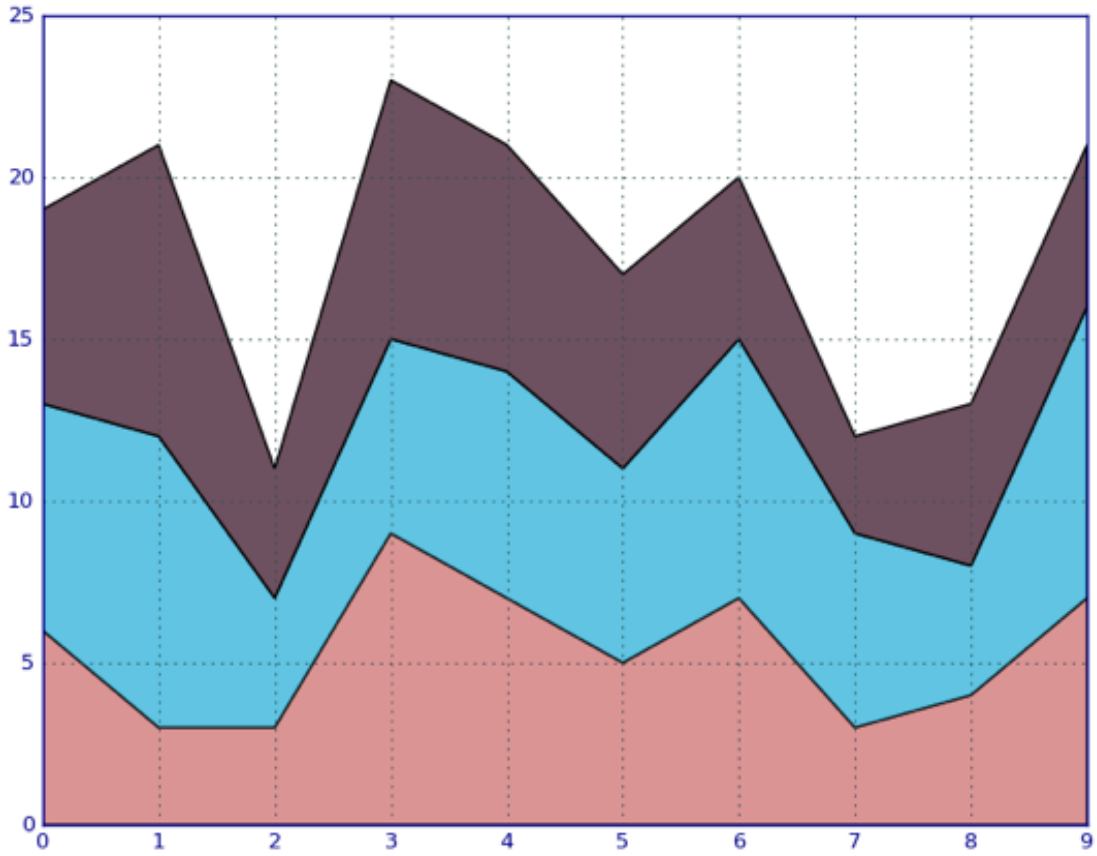
Streit & Gehlenborg, PoV, Nature Methods, 2014

LineUp



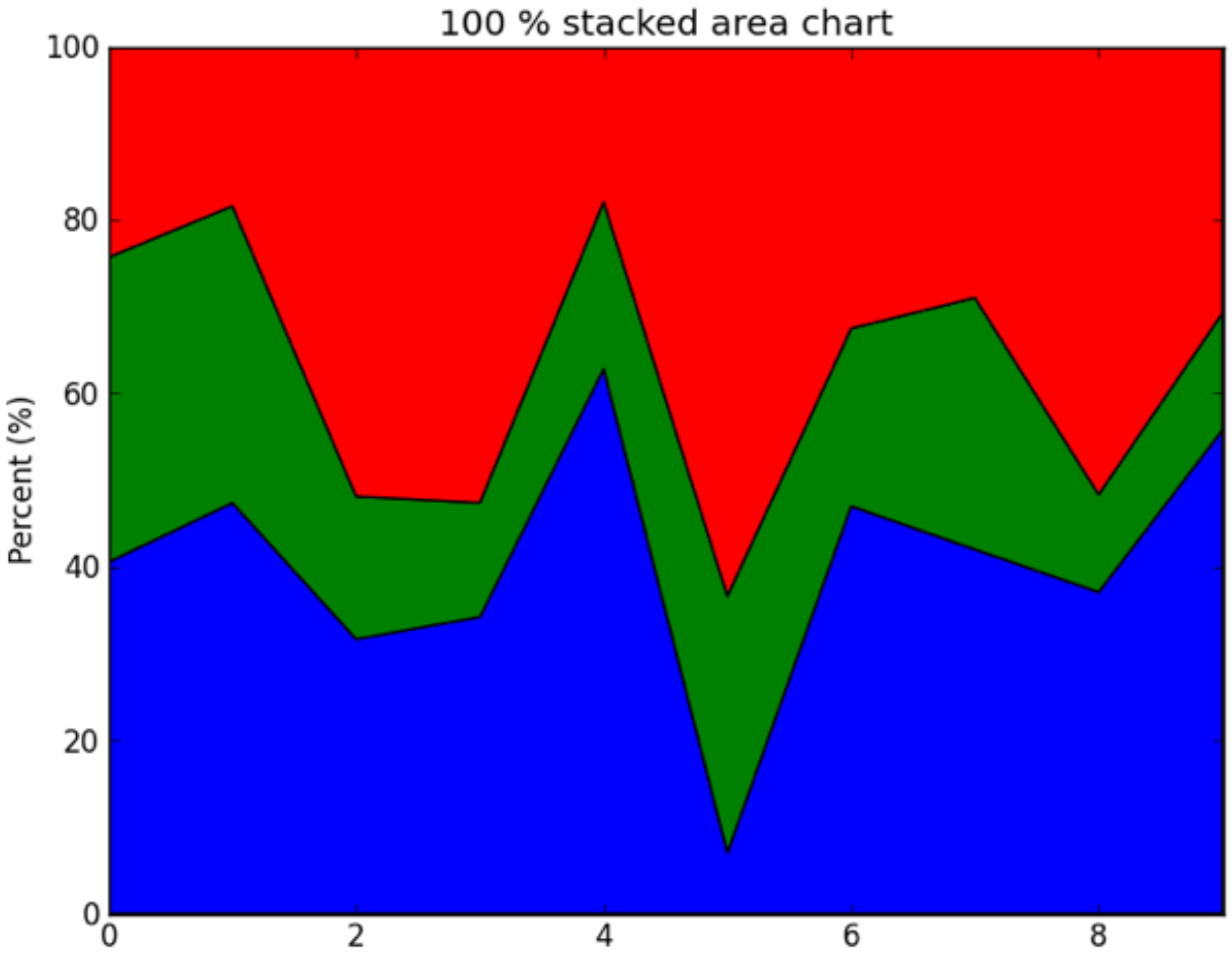
Video at http://lineup.caleydo.org

Stacked Area Chart



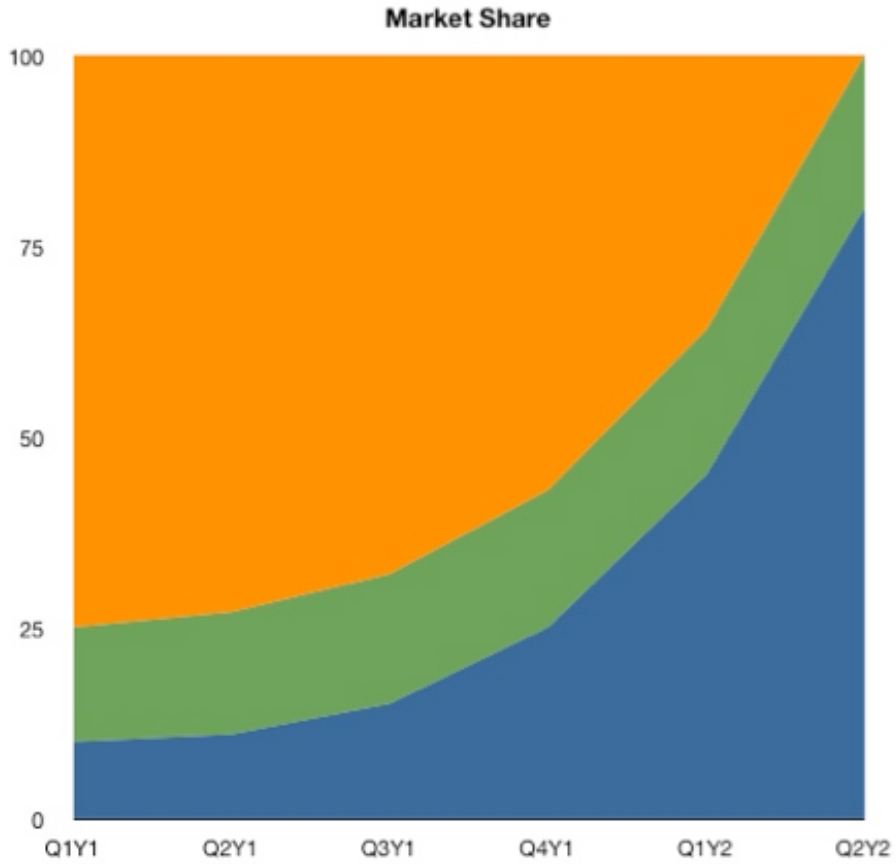
http://stackoverflow.com/questions/2225995/how-can-i-create-stacked-line-graph-with-matplotlib

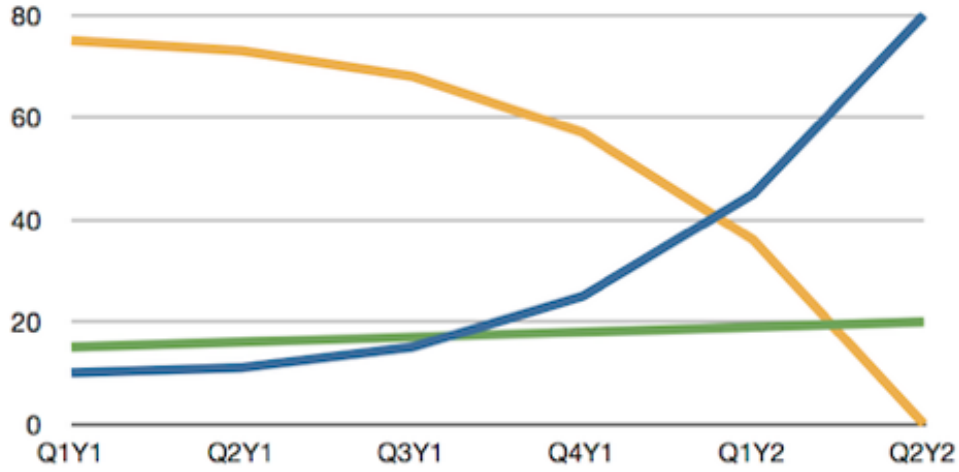
100% Stacked Area Chart



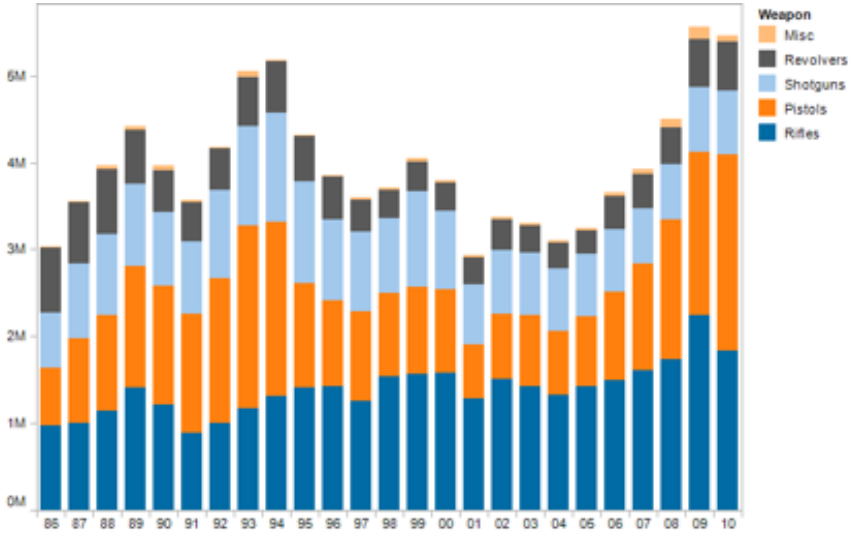
http://stackoverflow.com/questions/16875546/create-a-100-stacked-area-chart-with-matplotlib

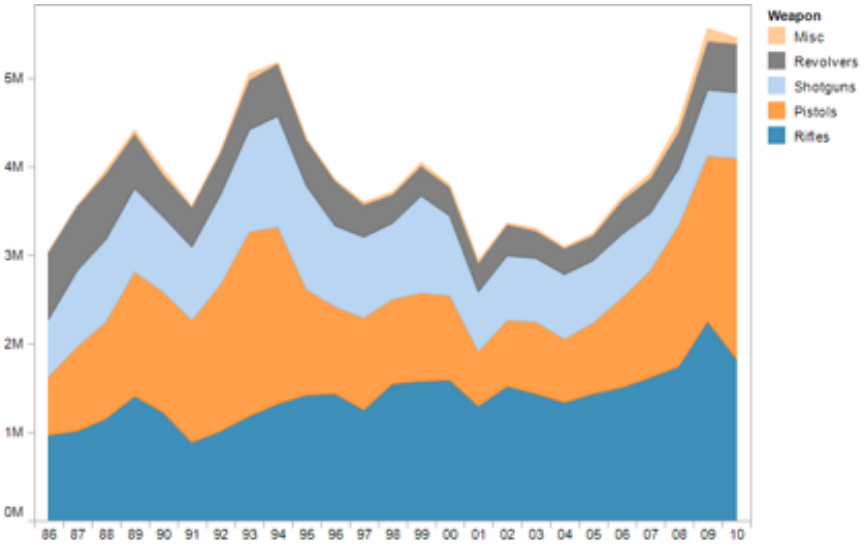
Stacked Area vs. Line Graphs

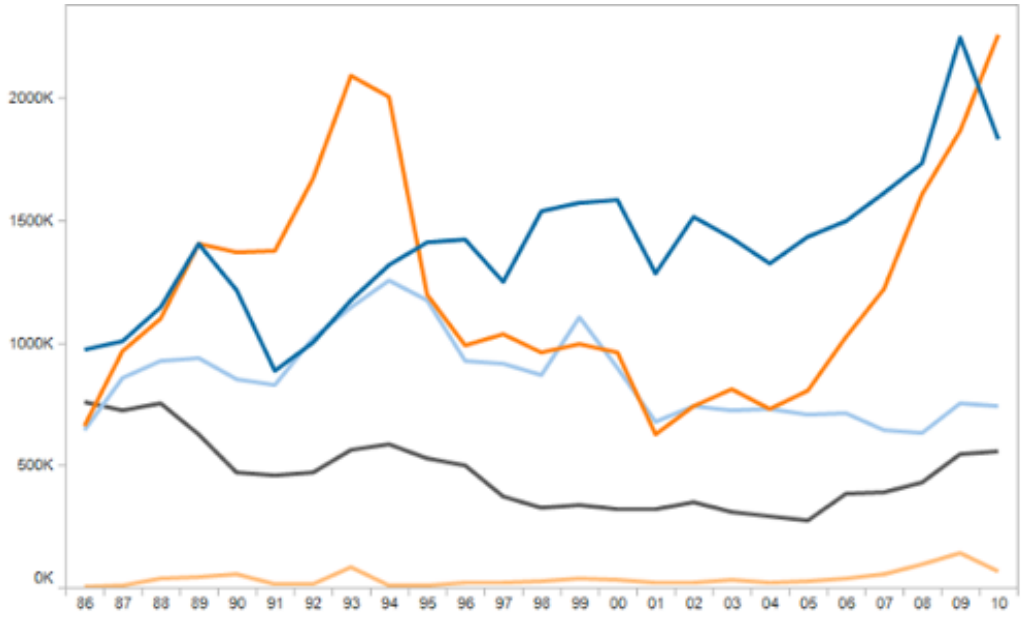




leancrew.com &
Practically Efficient



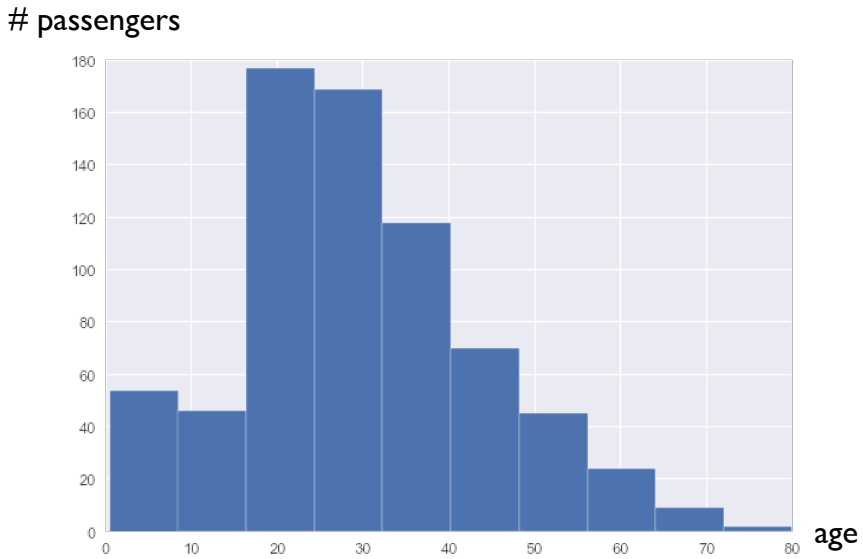




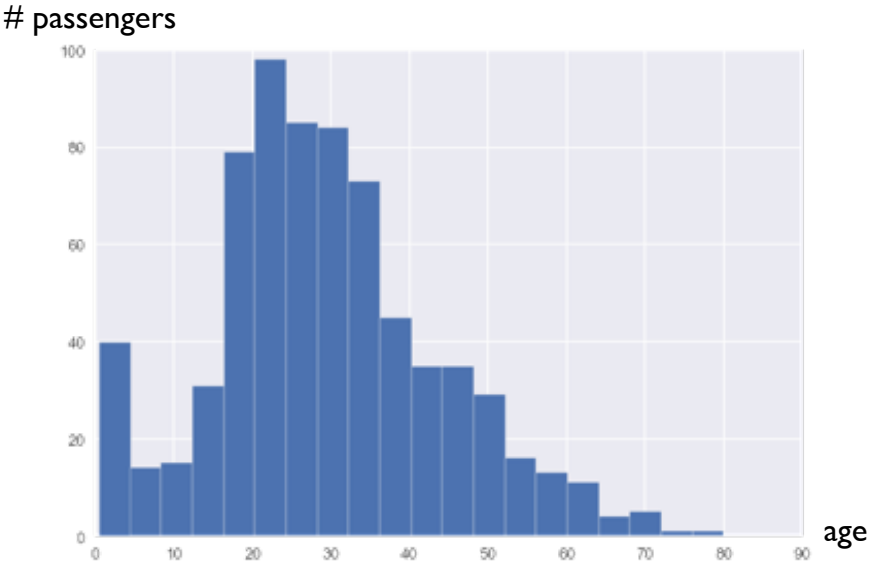
VizWiz, A. Kriebel

Distributions

Histogram

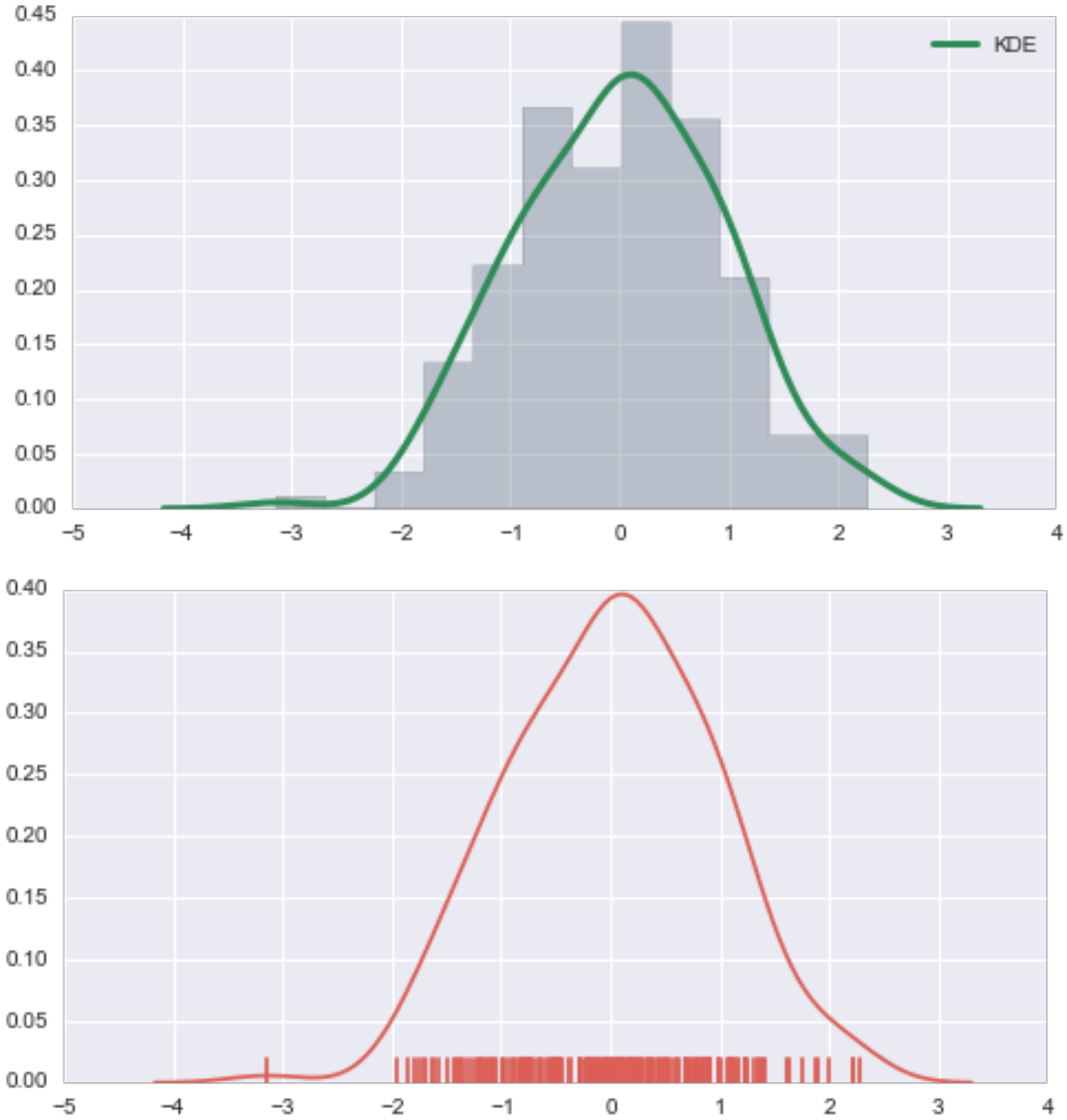


10 Bins



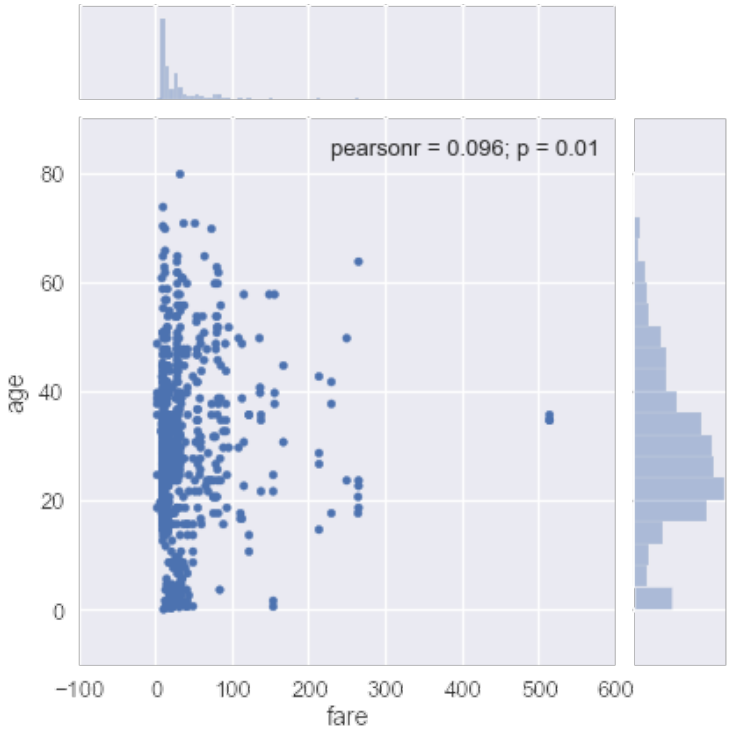
20 Bins

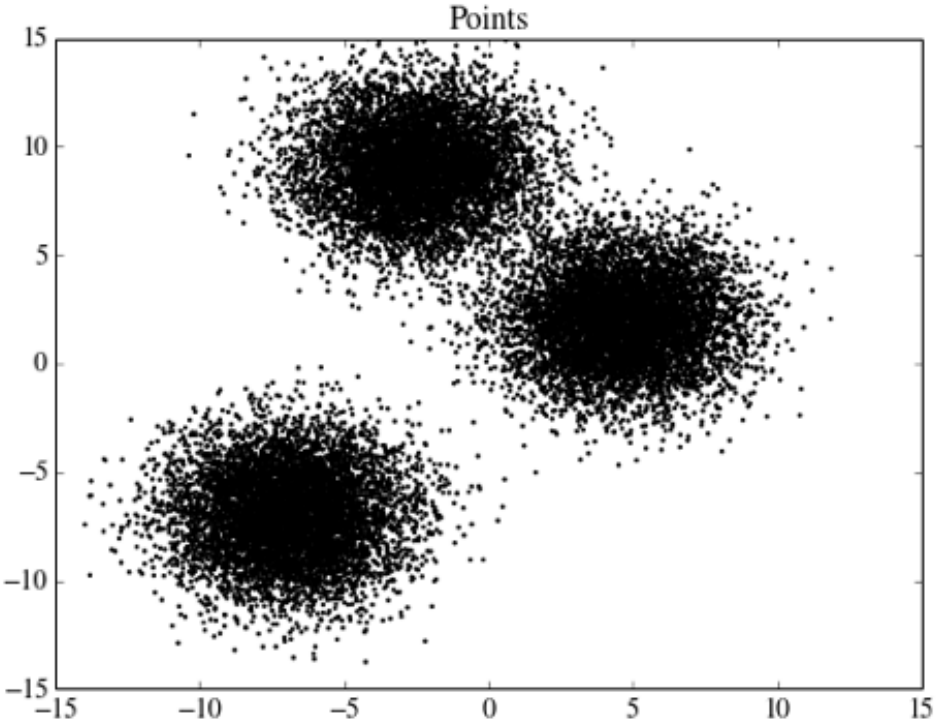
Density Plots

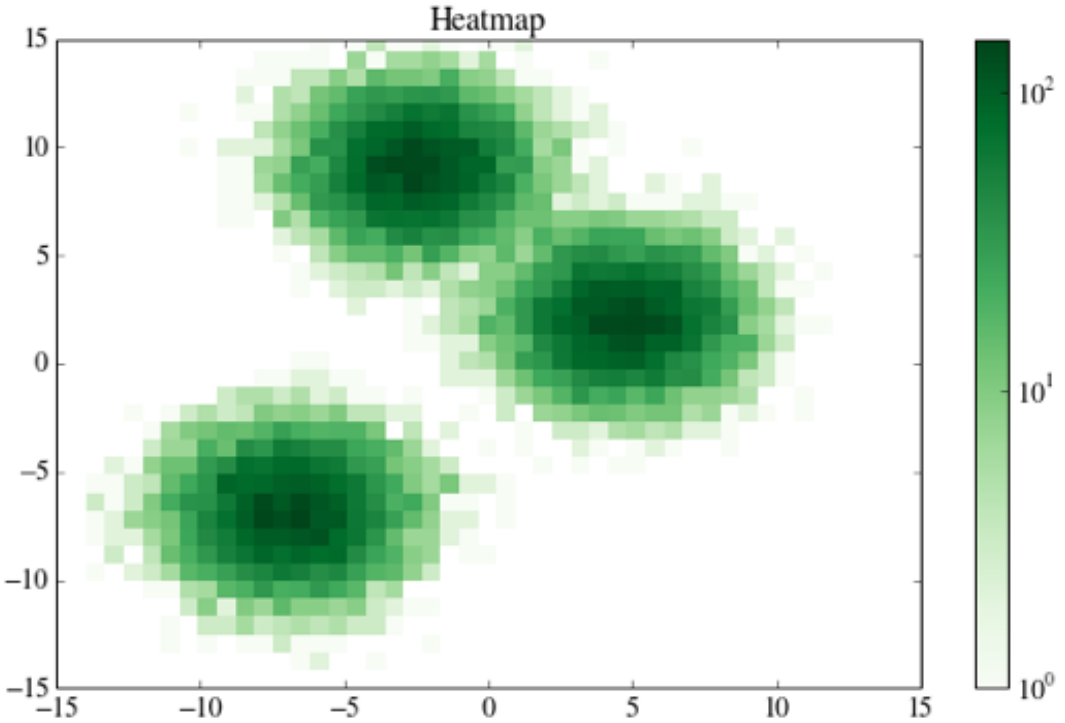


http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Heat Maps



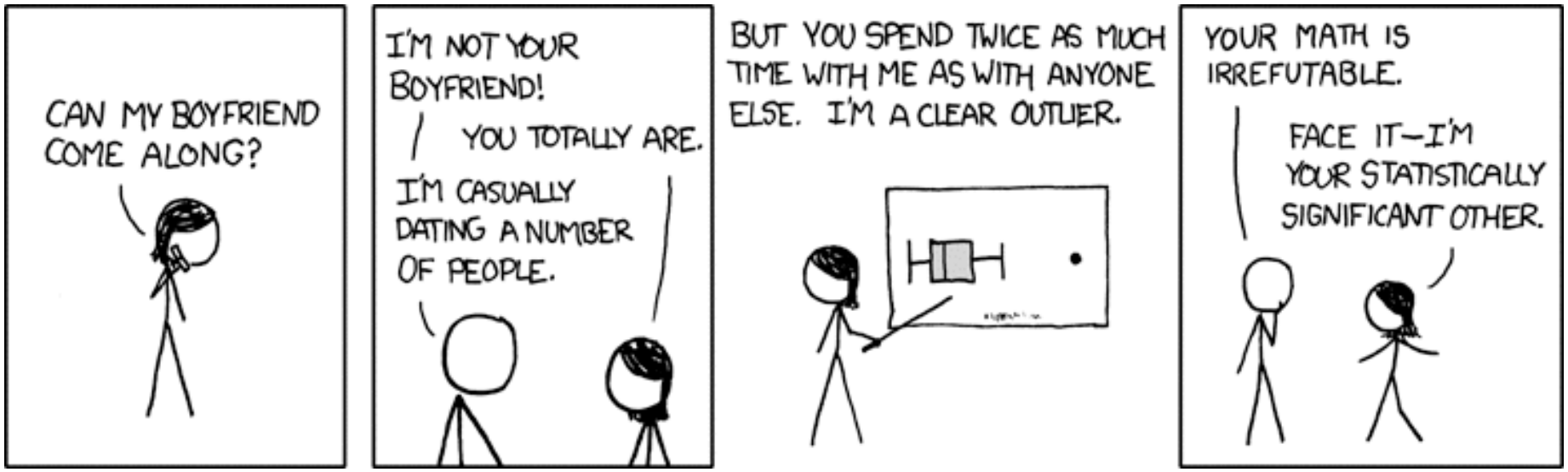




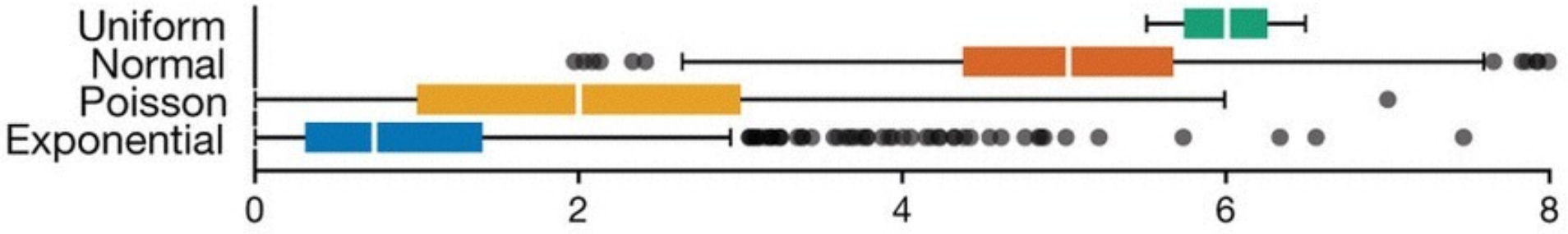
2D Density Plots

Box Plots

aka Box-and-Whisker Plot

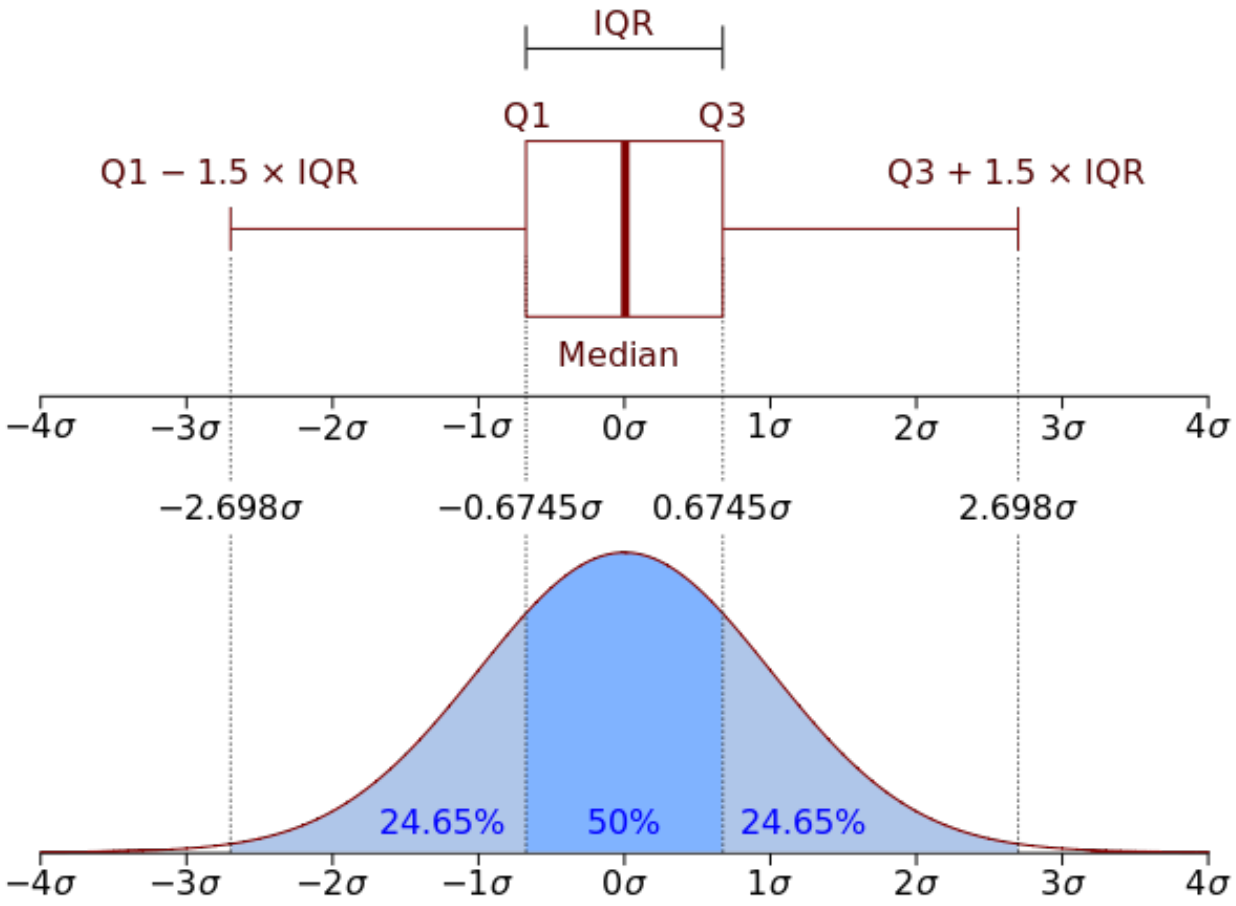


http://xkcd.com/539/



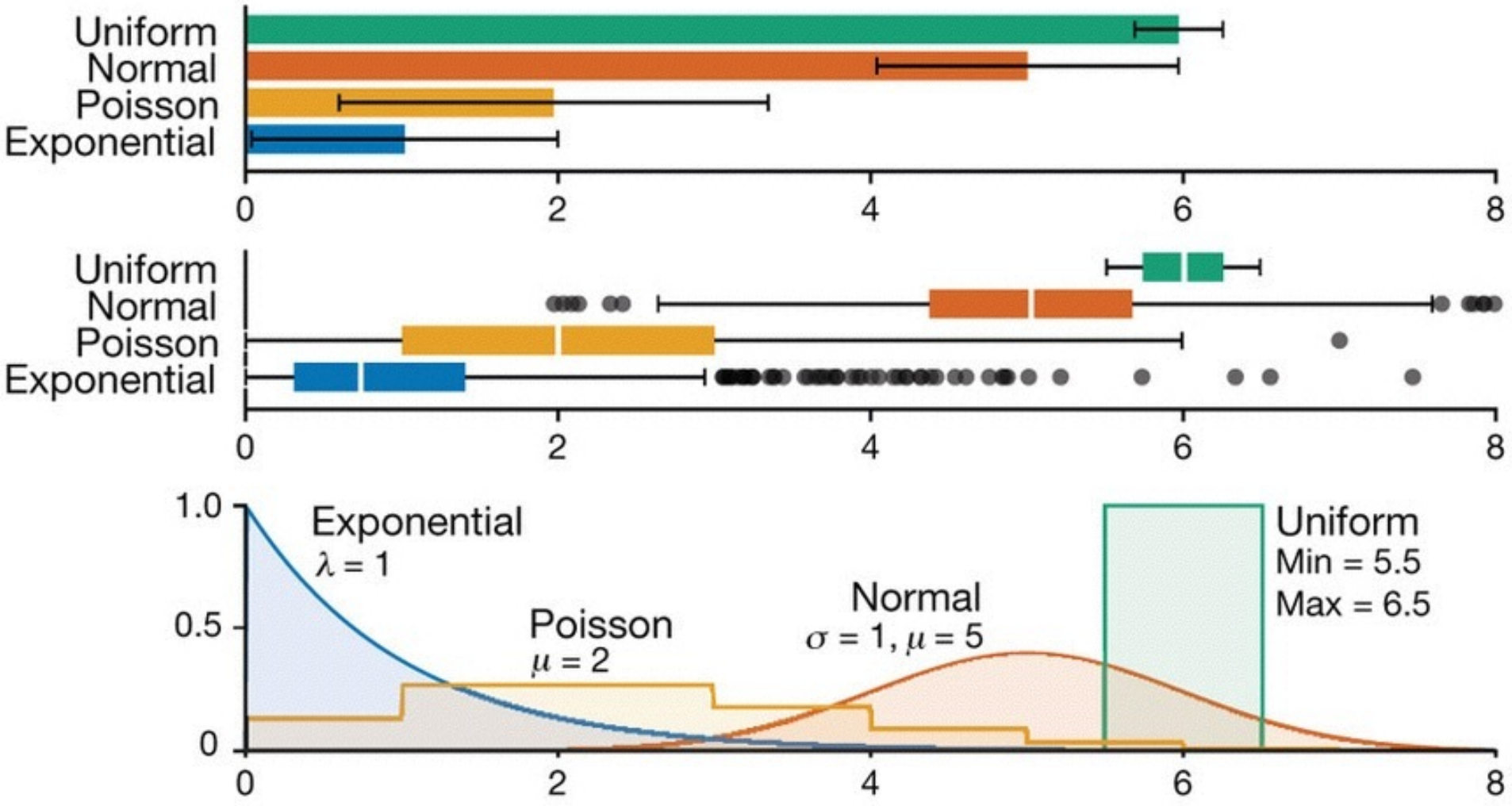
Box Plots

aka Box-and-Whisker Plot



Wikipedia

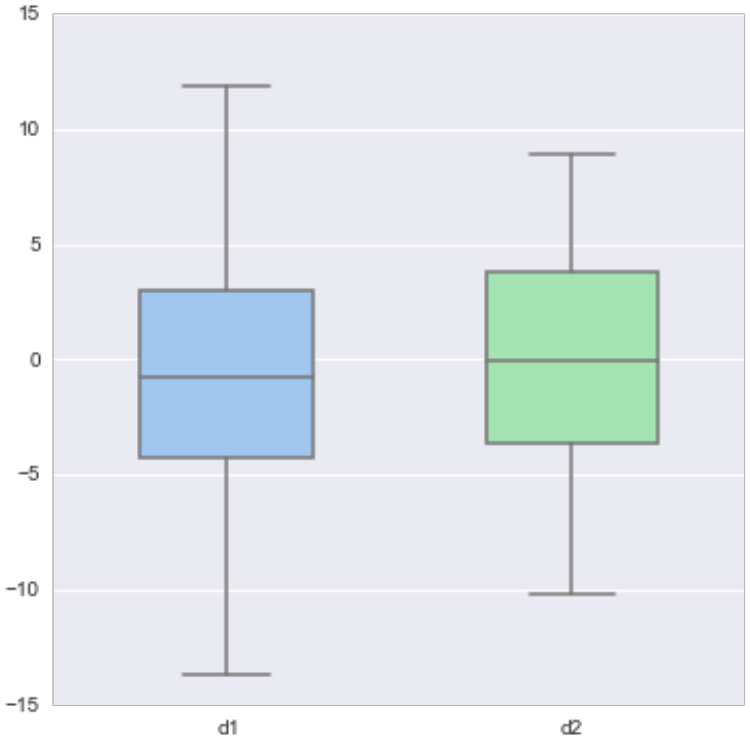
Comparison

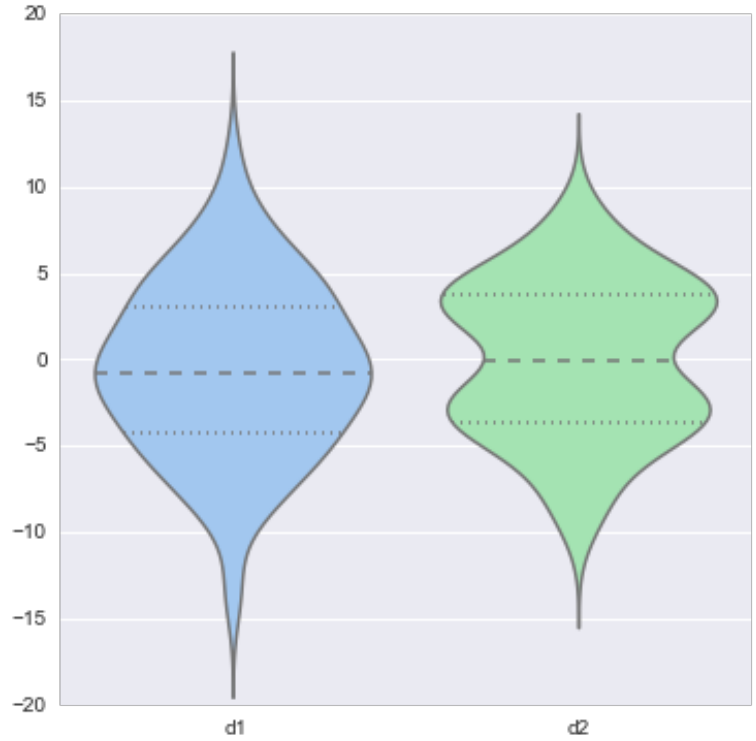


Streit & Gehlenborg, PoV, Nature Methods, 2014

Violin Plot

= Box Plot + Probability Density Function





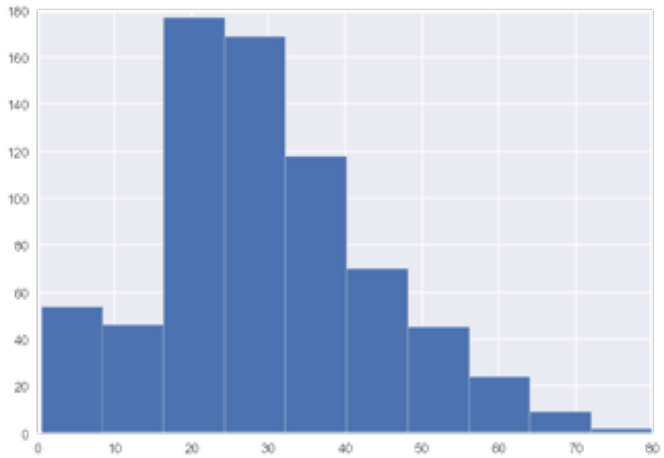
http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Multi-Dimensional Data Visualization

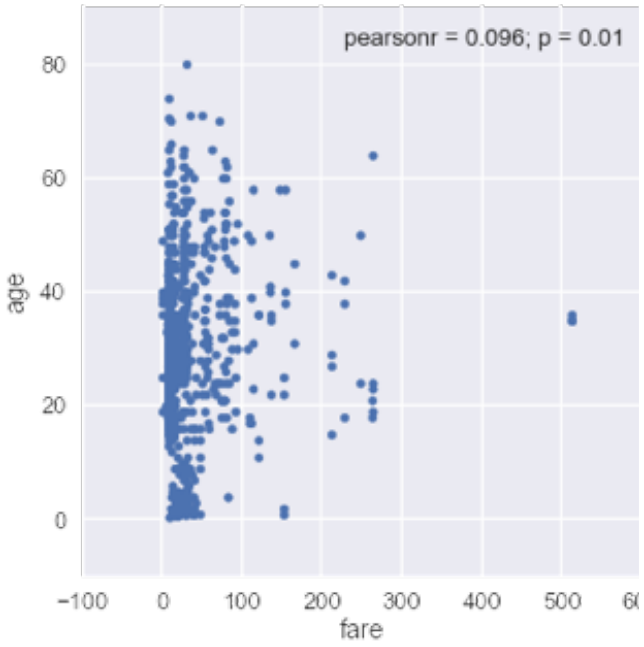
survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0	30.0708	C	Second	child	False		Cherbourg	yes	False
1	3	female	4.0	1	1	16.7	S	Third	child	False	G	Southampton	yes	False
1	1	female	58.0	0	0	26.55	S	First	woman	False	C	Southampton	yes	True
0	3	male	20.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male	39.0	1	5	31.275	S	Third	man	True		Southampton	no	False
0	3	female	14.0	0	0	7.8542	S	Third	child	False		Southampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

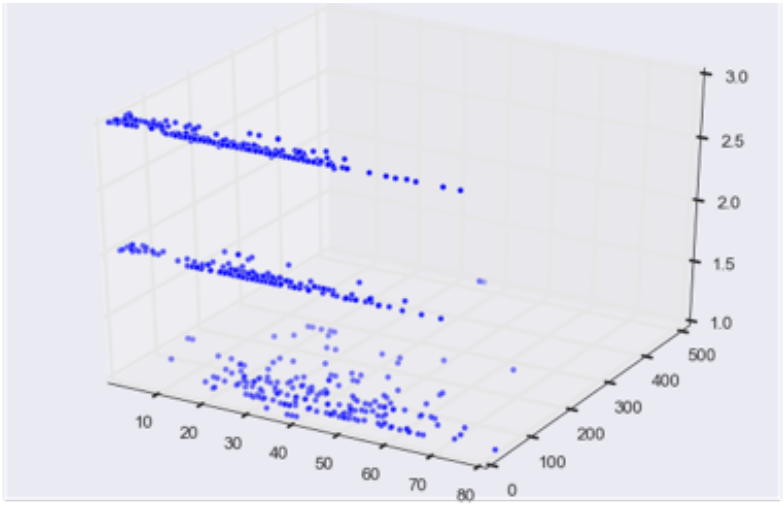
Example: Titanic Dataset

1D



2D





3D

4D?

What is “high” dimensional?

How many dimensions (attributes)?

- ~50 – tractable with “just” vis
- ~1,000 – need analytical methods

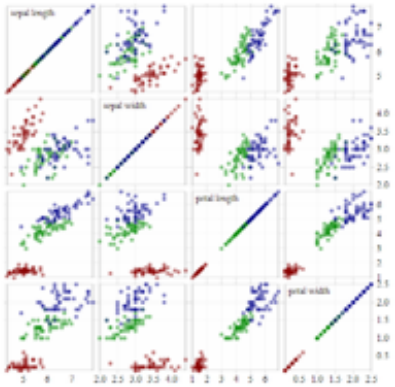
How many items?

- ~ 1,000 – “just” vis is fine
- >> 10,000 – need analytical methods

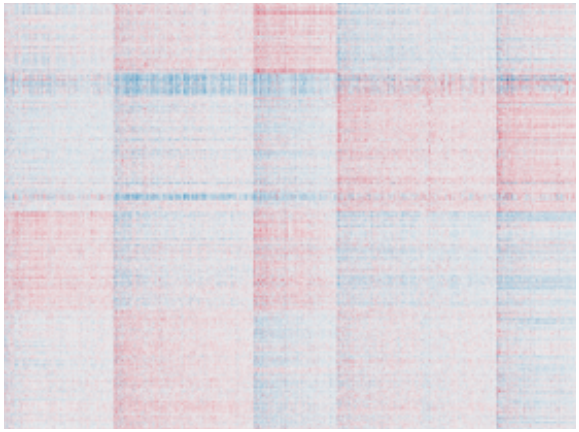
Homogeneity

- Same data type?
- Same scales?

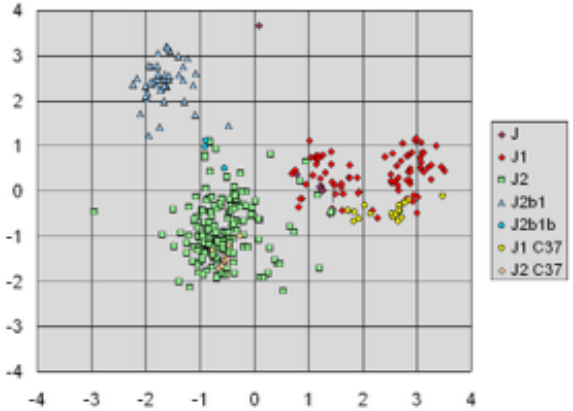
Analytic Component



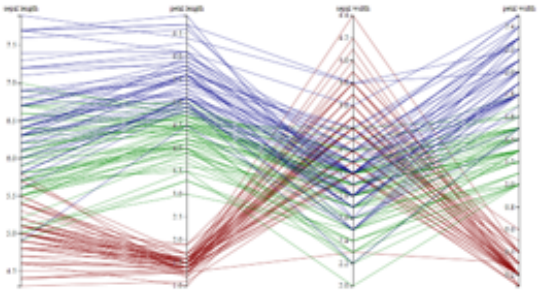
Scatterplot Matrices



Pixel-based Visualizations / Heat Maps



Dimensionality Reduction (e.g., PCA)



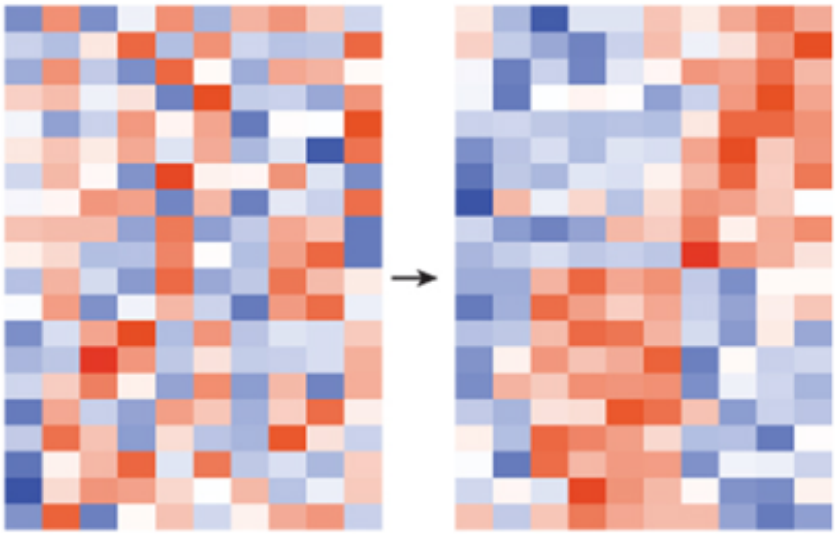
Parallel Coordinates

no / little analytics

strong analytics component

Heat Map

Each cell is a “pixel”, value encoded using color

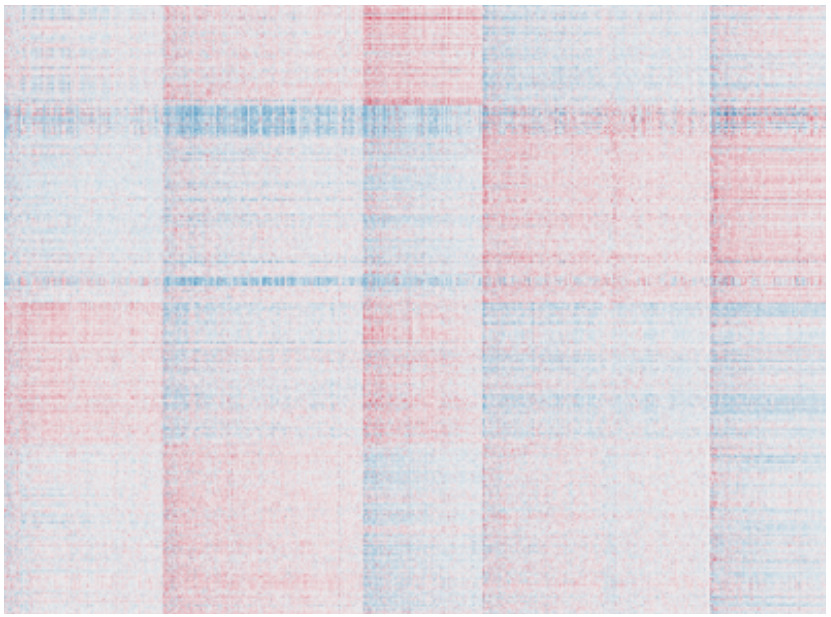


Meaning derived from ordering

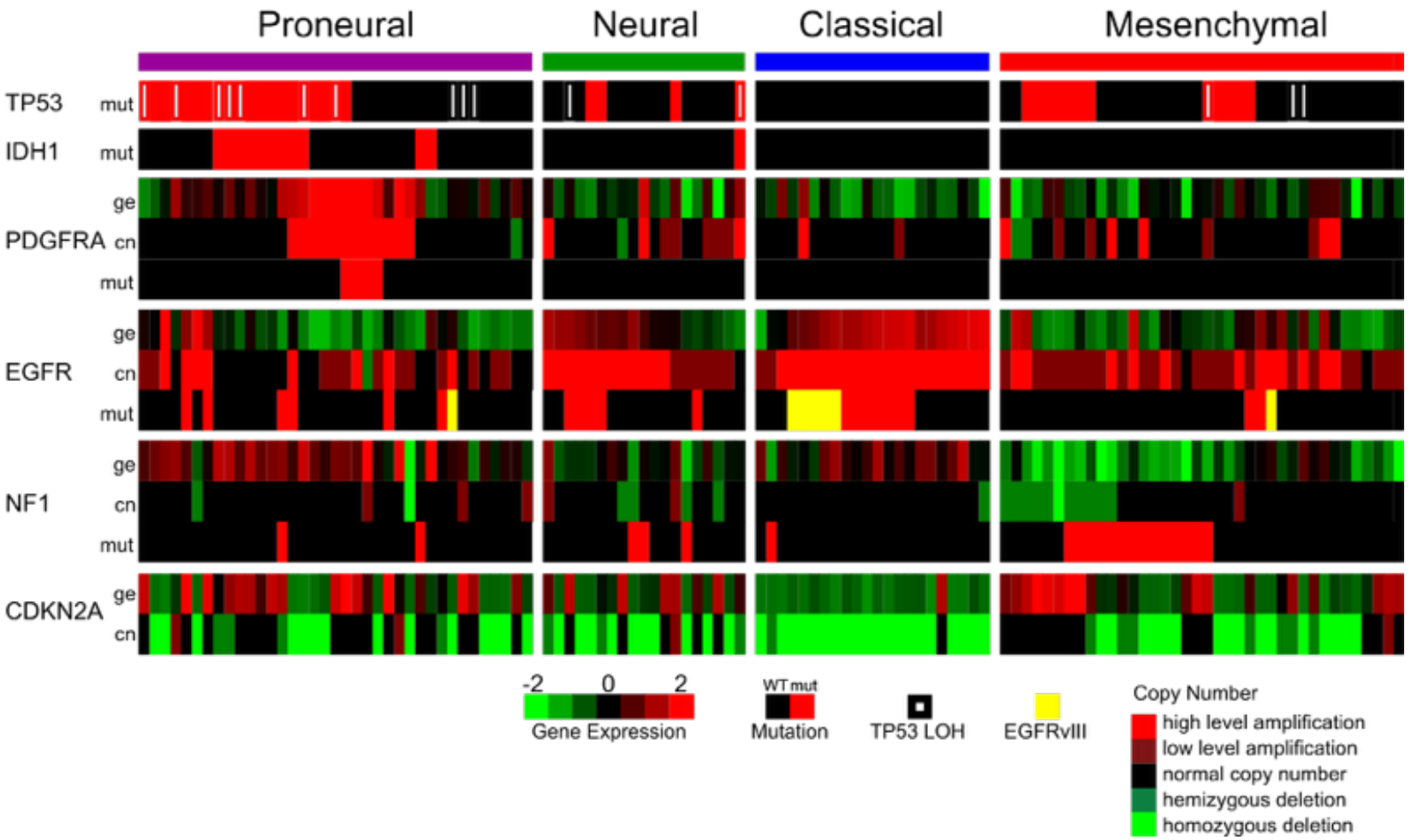
If no ordering inherent, clustering is used

Scalable – 1 px per item

Good for homogeneous data



Heterogeneous Data?



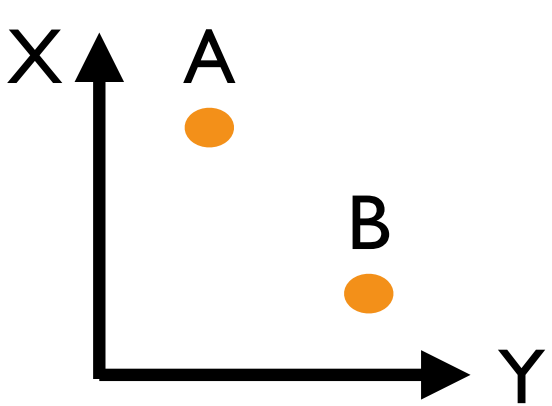
Verhaak 2012

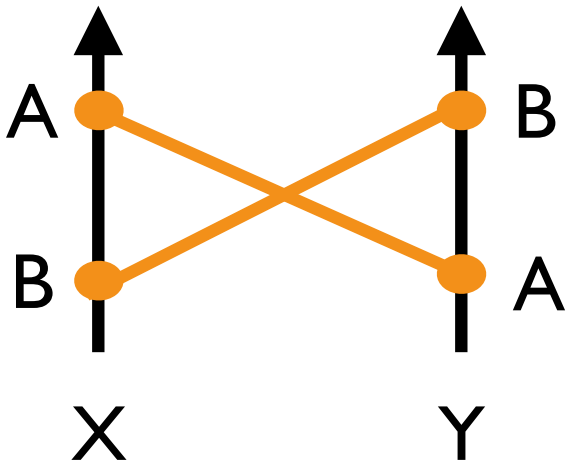
Parallel Coordinates (PC)

Inselberg 1985

Axes represent attributes

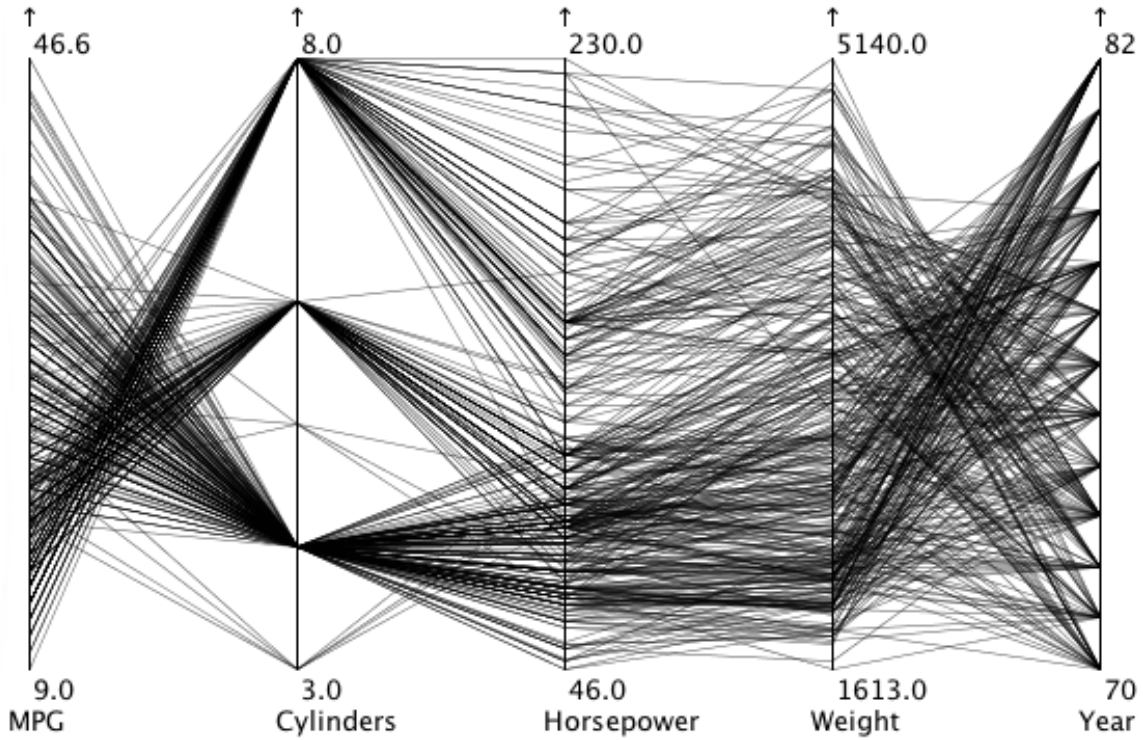
Lines connecting axes represent items





Example: Cars Dataset

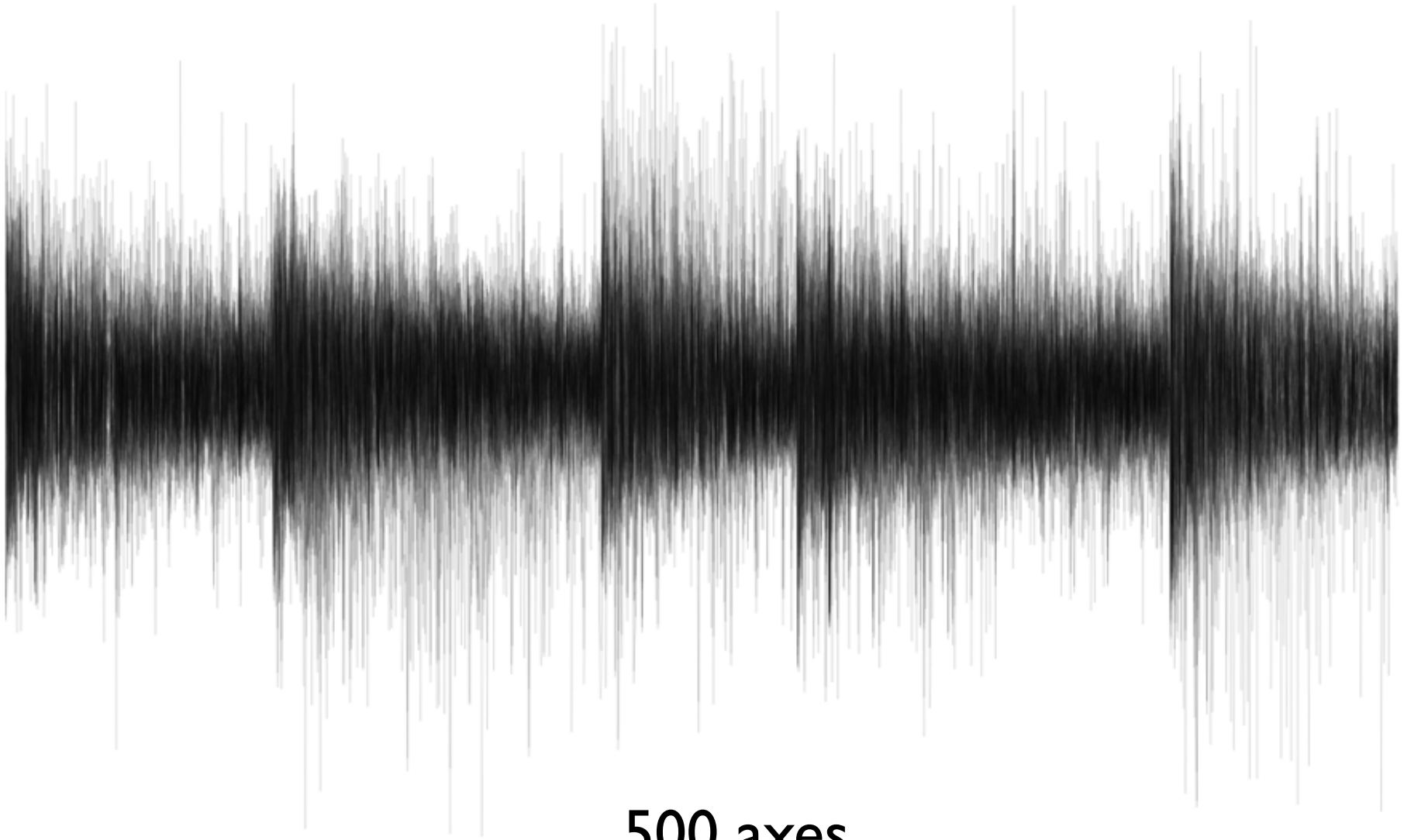
1	MPG	Cylinders	Horsepower	Weight	Acceleration	Year	Origin
2	18	8	130	3504	12	70	USA
3	15	8	165	3693	11.5	70	USA
4	18	8	150	3436	11	70	USA
5	16	8	150	3433	12	70	USA
6	17	8	140	3449	10.5	70	USA
7	15	8	198	4341	10	70	USA
8	14	8	220	4354	9	70	USA
9	14	8	215	4312	8.5	70	USA
10	14	8	225	4425	10	70	USA
11	15	8	190	3850	8.5	70	USA
12	15	8	170	3563	10	70	USA
13	14	8	160	3609	8	70	USA
14	15	8	150	3761	9.5	70	USA
15	14	8	225	3086	10	70	USA
16	24	4	95	2372	15	70	Europe
17	22	6	95	2833	15.5	70	USA
18	18	6	97	2774	15.5	70	USA
19	21	6	85	2587	16	70	USA



Limitations of PC?

PC Limitations

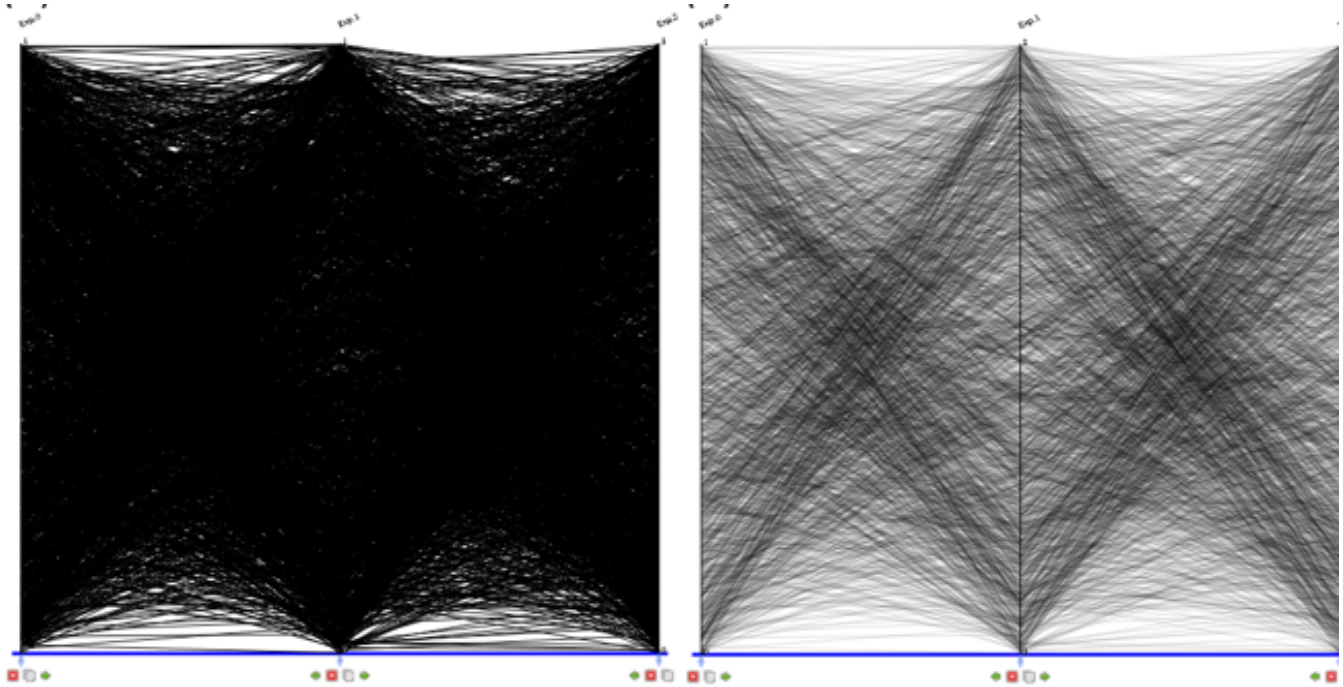
Scalability to Many Dimensions



500 axes

PC Limitations

Scalability to Many Items

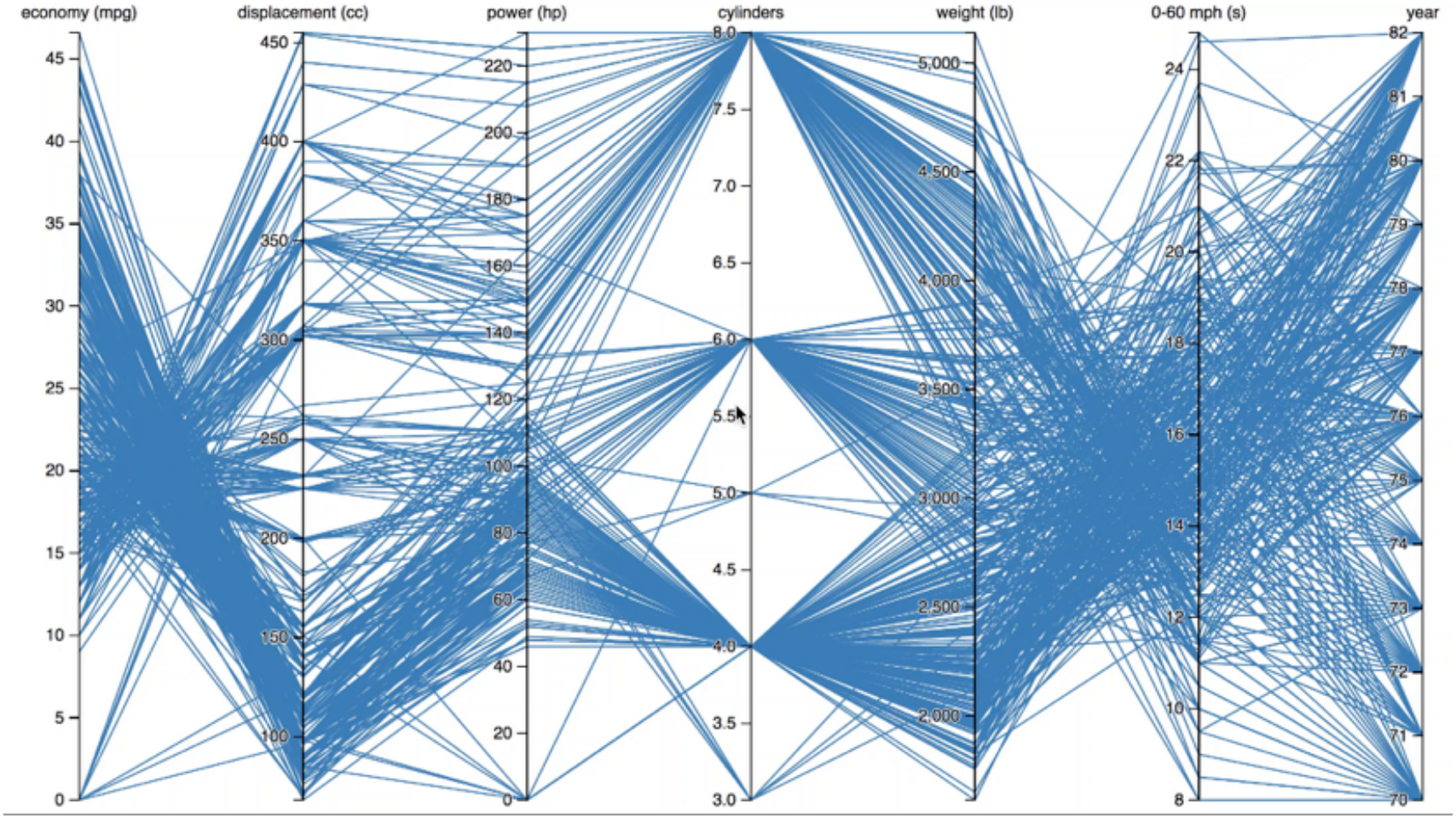


Solutions:

- Transparency
- Bundling, Clustering
- Sampling

PC Limitations

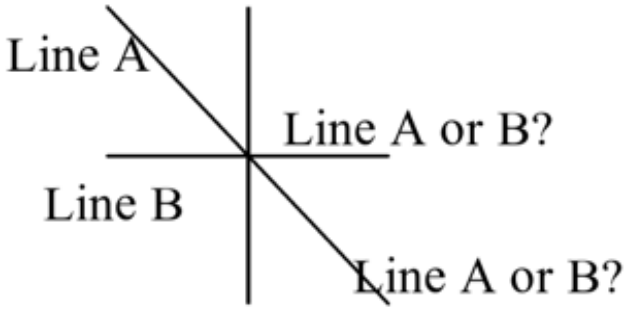
Correlations only between adjacent axes

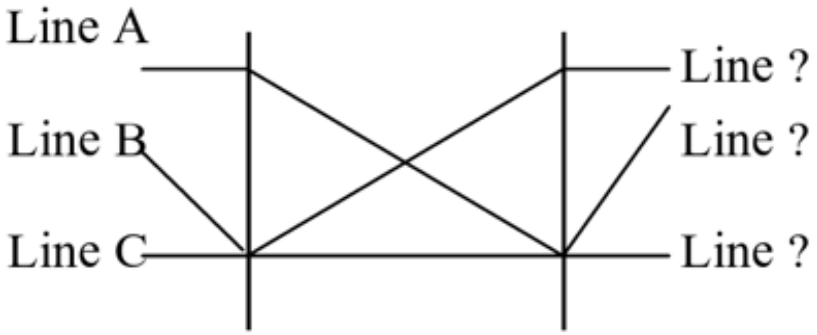


Solution: Let user change order

PC Limitations

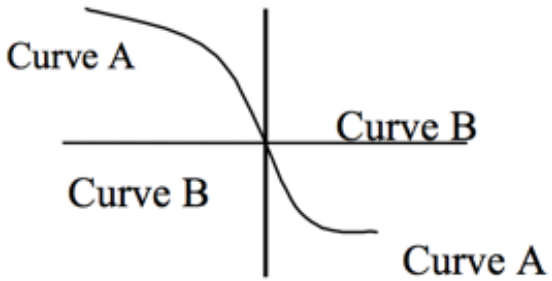
Ambiguity

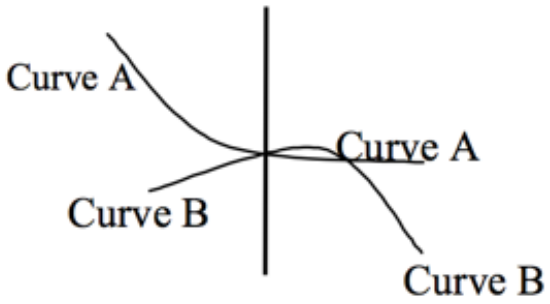


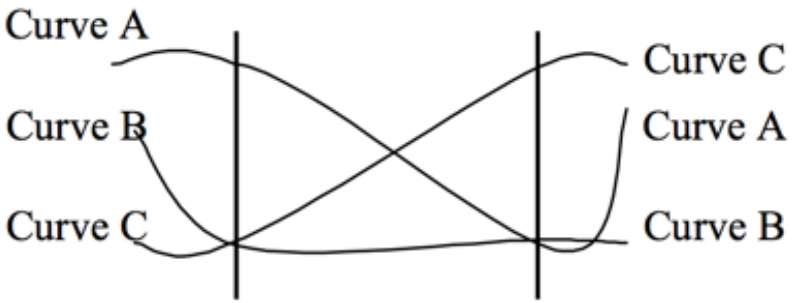


Solutions:

- Interactive highlighting
- Curves







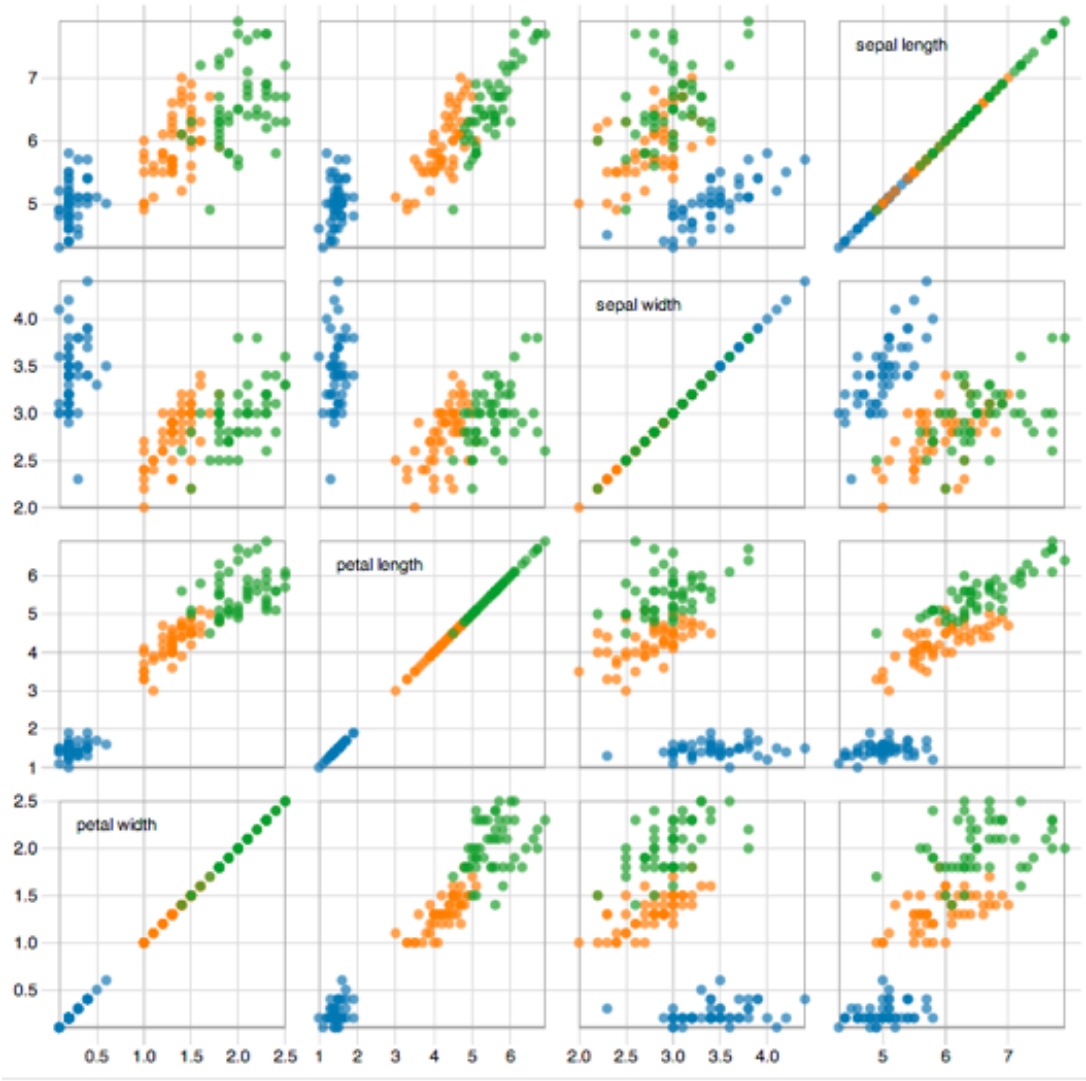
Graham and Kennedy 2003

Scatterplot Matrix (SPLOM)

N dimensions

N^2 scatterplots

Limited scalability
(~20 dims,
~500-1k items)

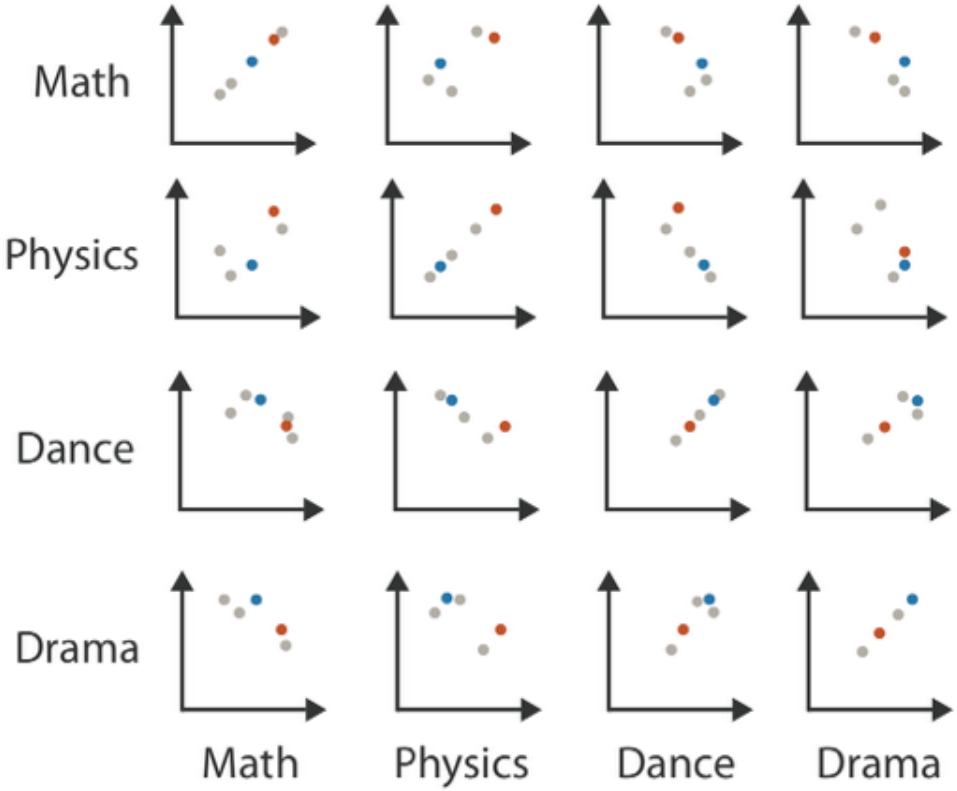


http://bl.ocks.org/mbostock/4063663

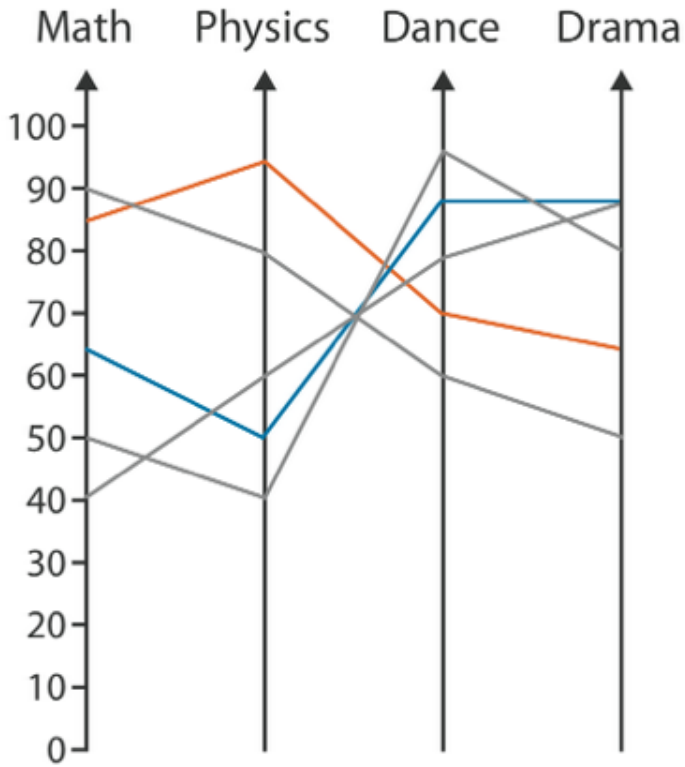
Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatterplot Matrix



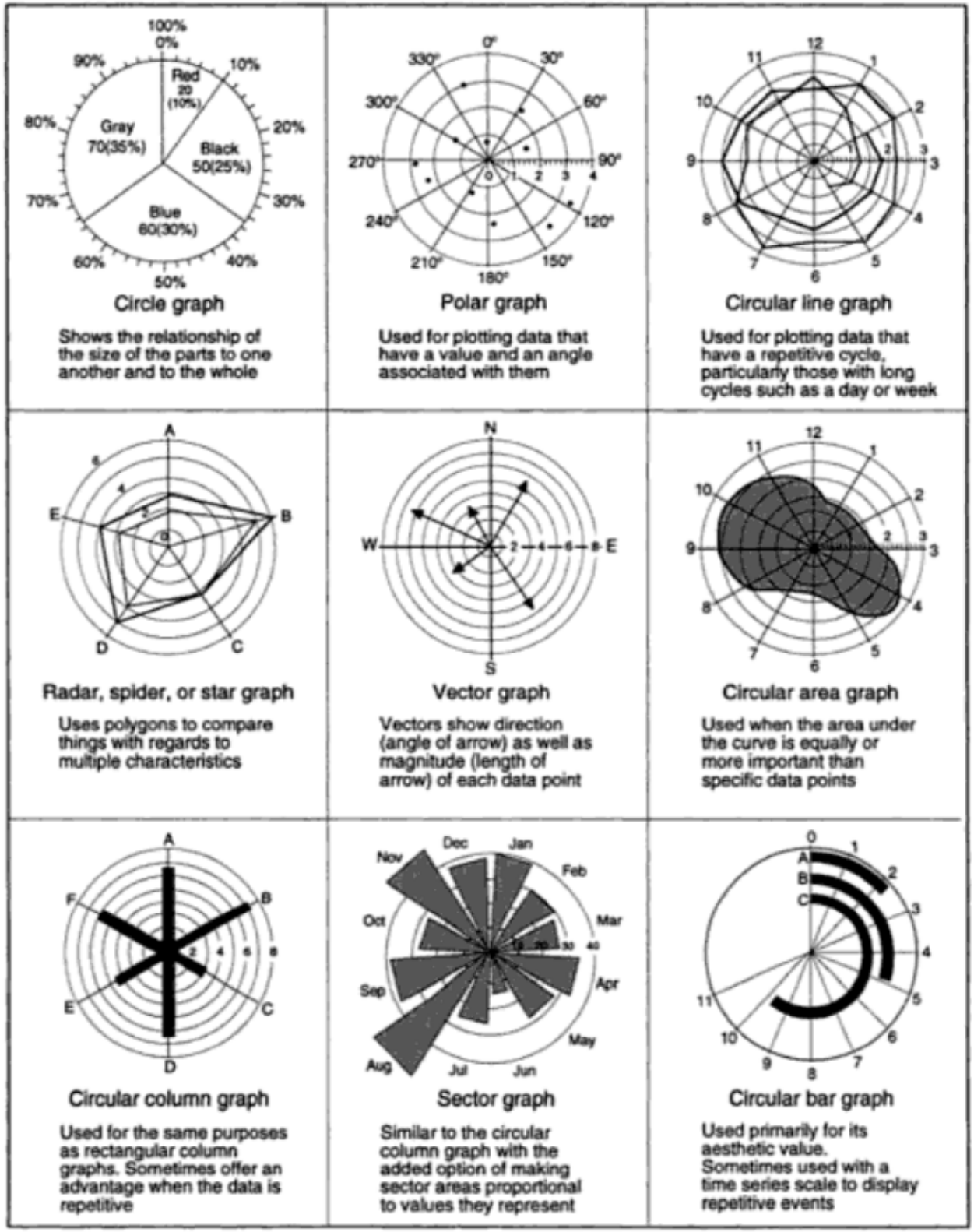
Parallel Coordinates



Radial Axis Techniques

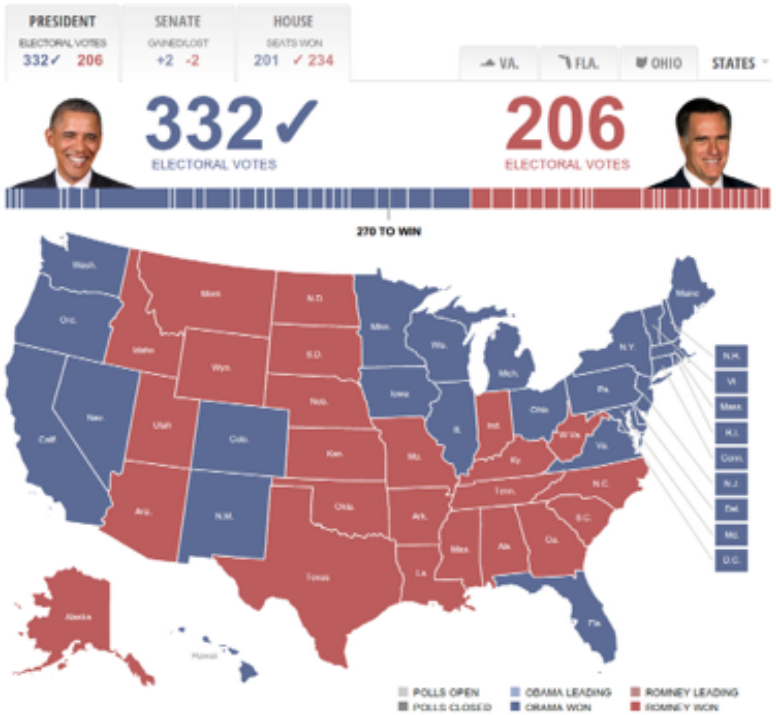
Similar to parallel coordinates

Axes radiate from a common origin

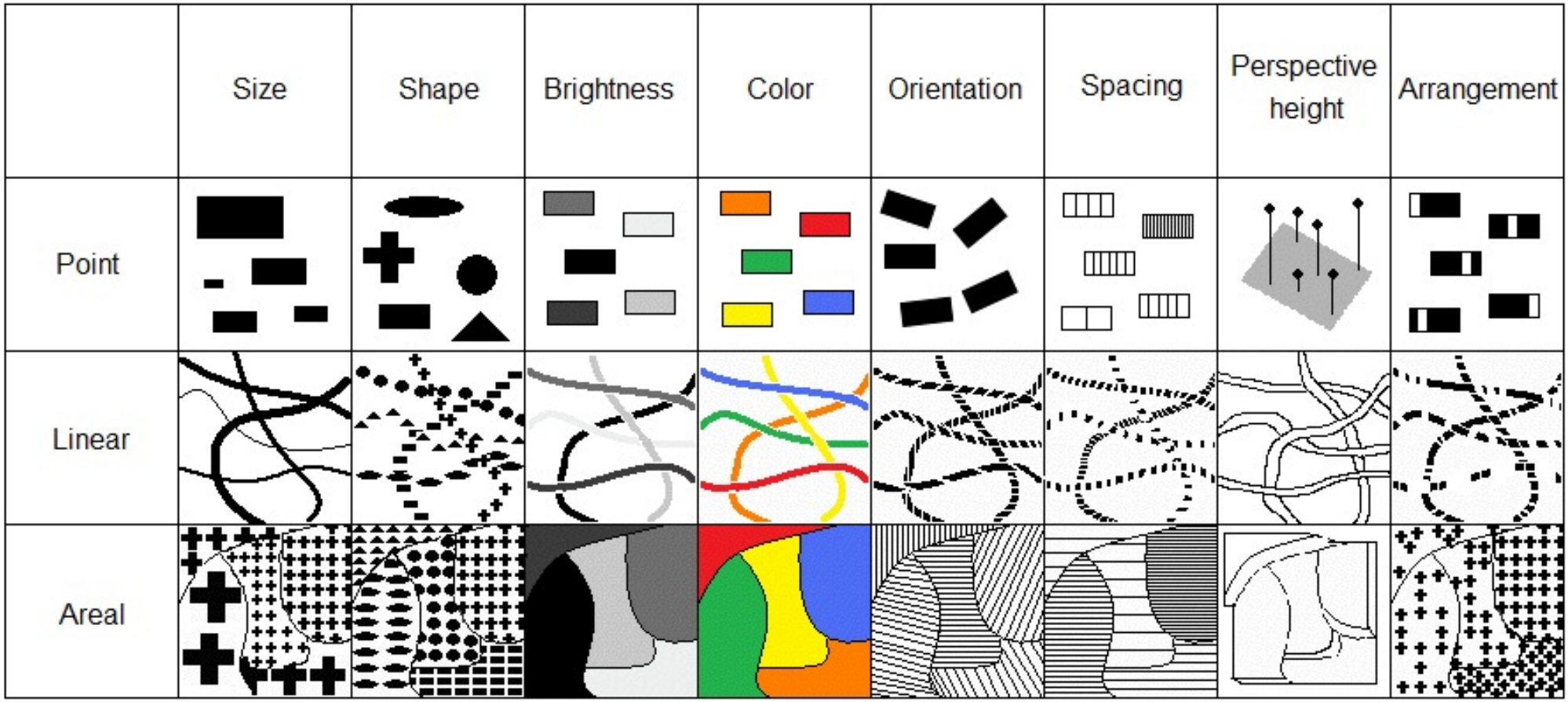


Harris 1999

Map Visualization



Visual Variables for Spatial Data

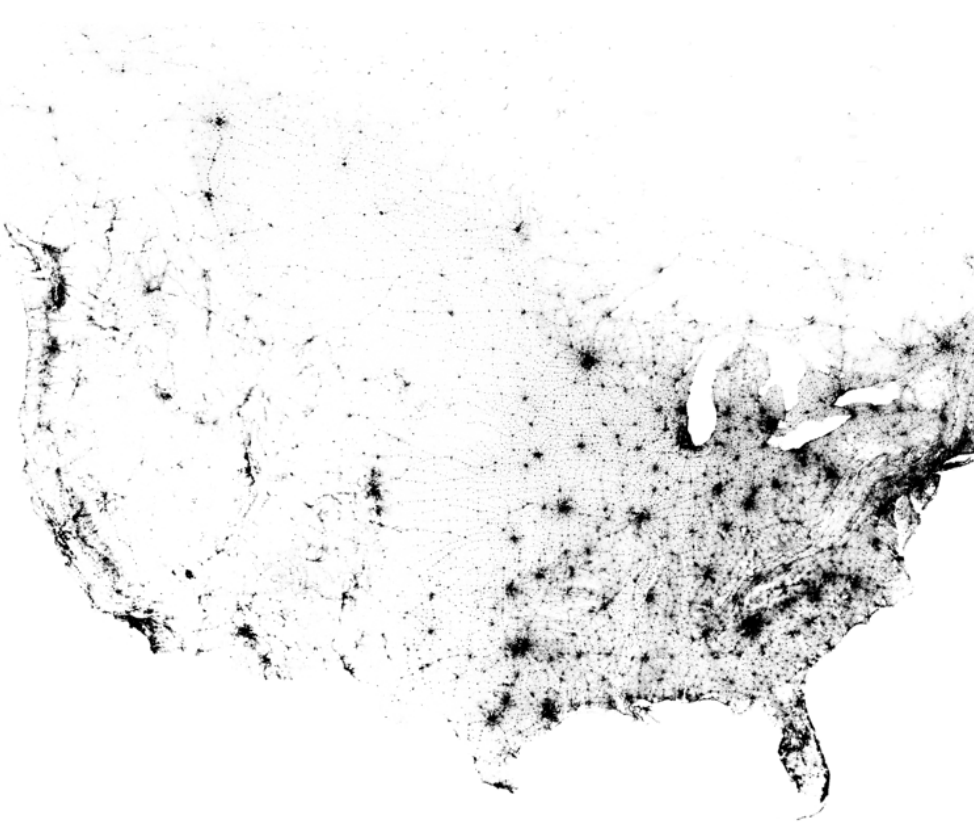


Slocum 1999

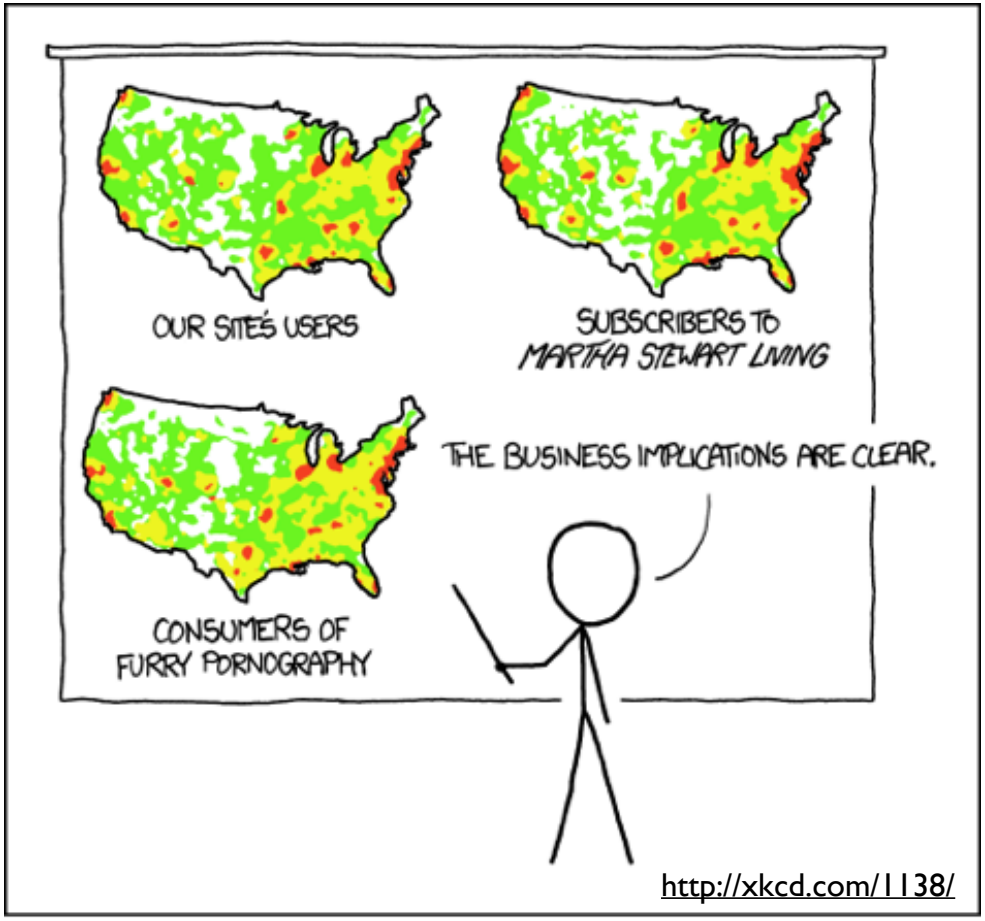
Dot Map

A Dot for Every Data Item

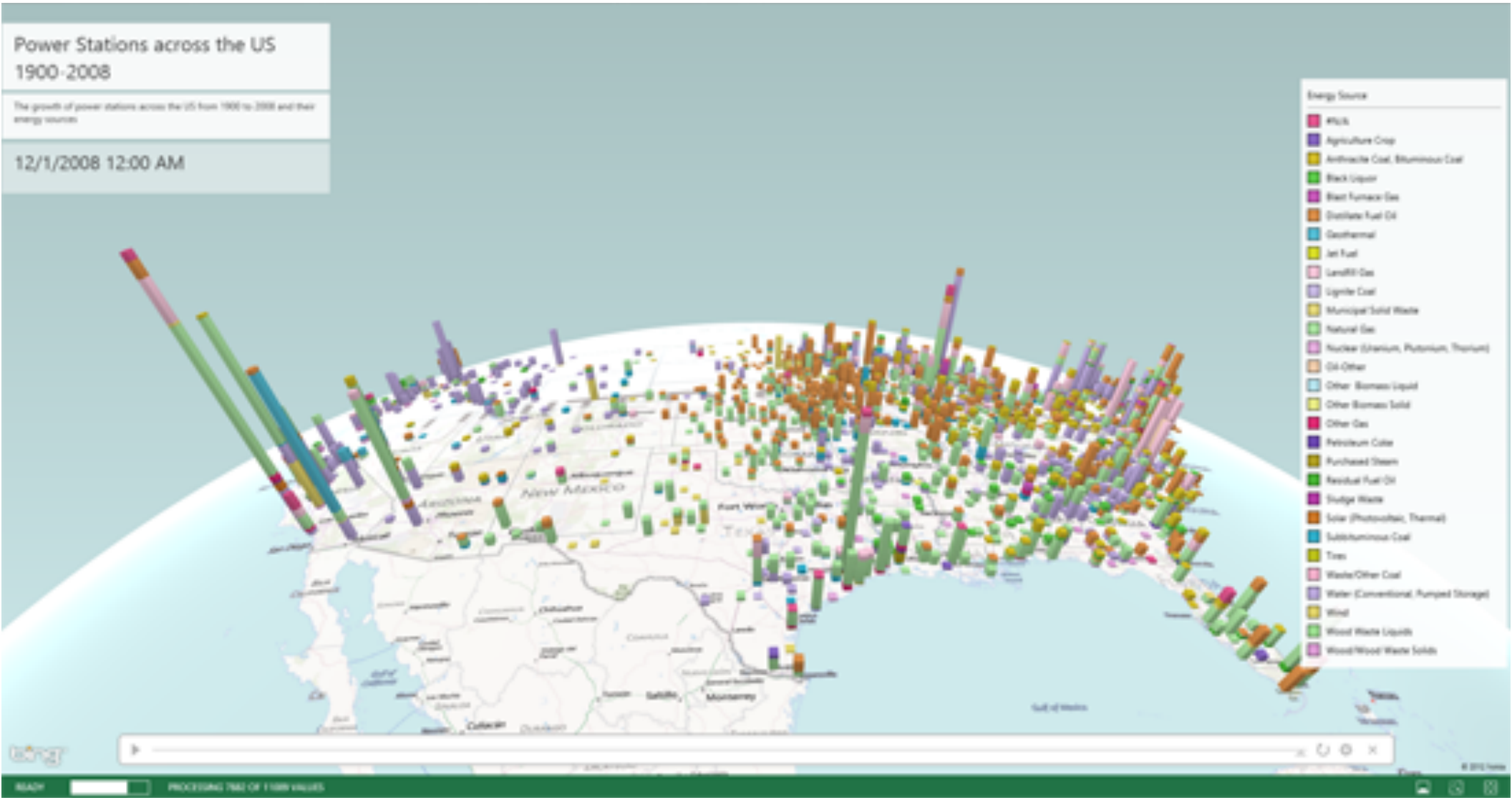
Maps can lie, too!



http://www.cairco.org/news/those-scary-dots-population-america



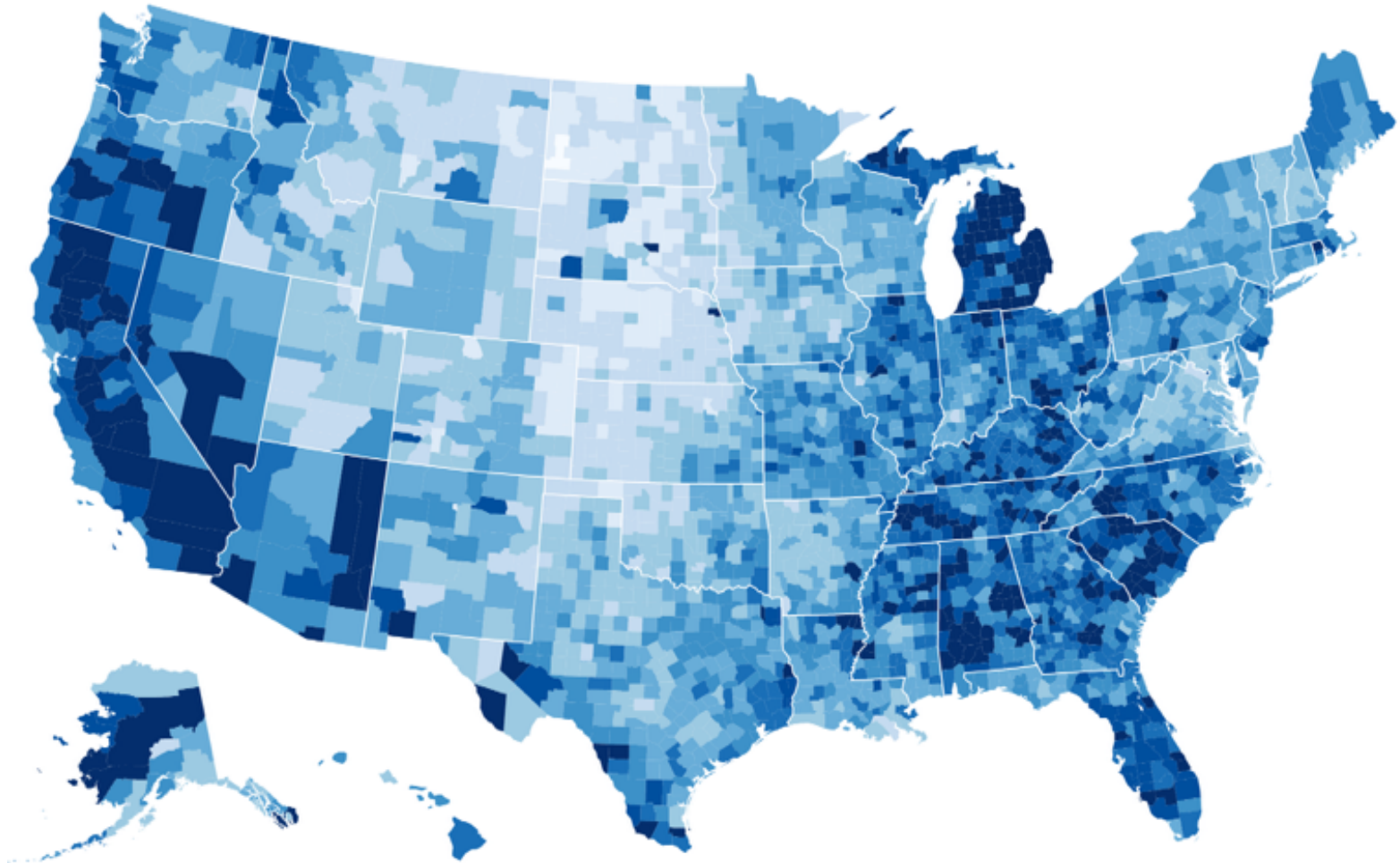
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS



Microsoft GeoFlow – Part of Excel 2013

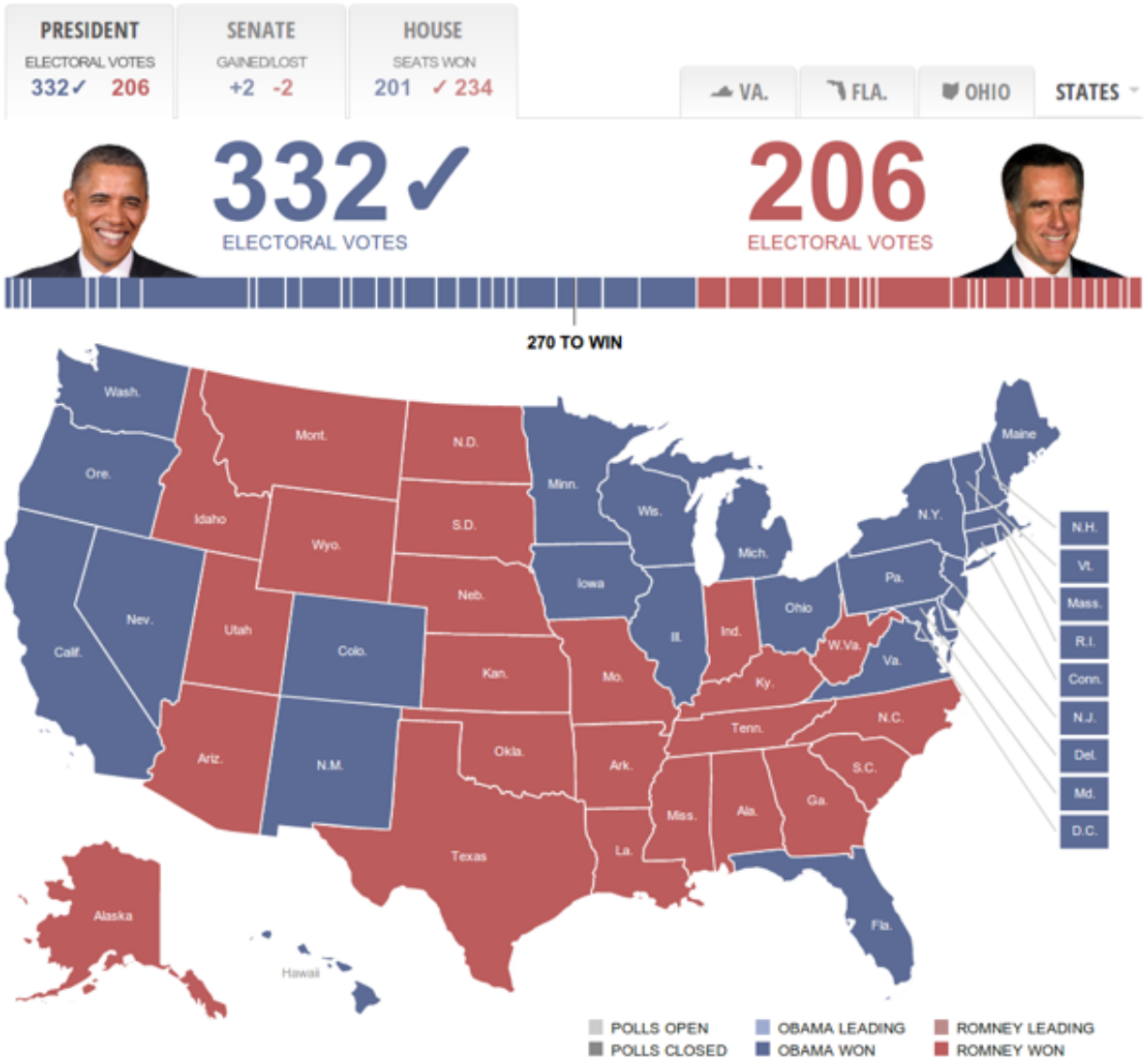
Choropleth Map

Attribute uniformly distributed in region



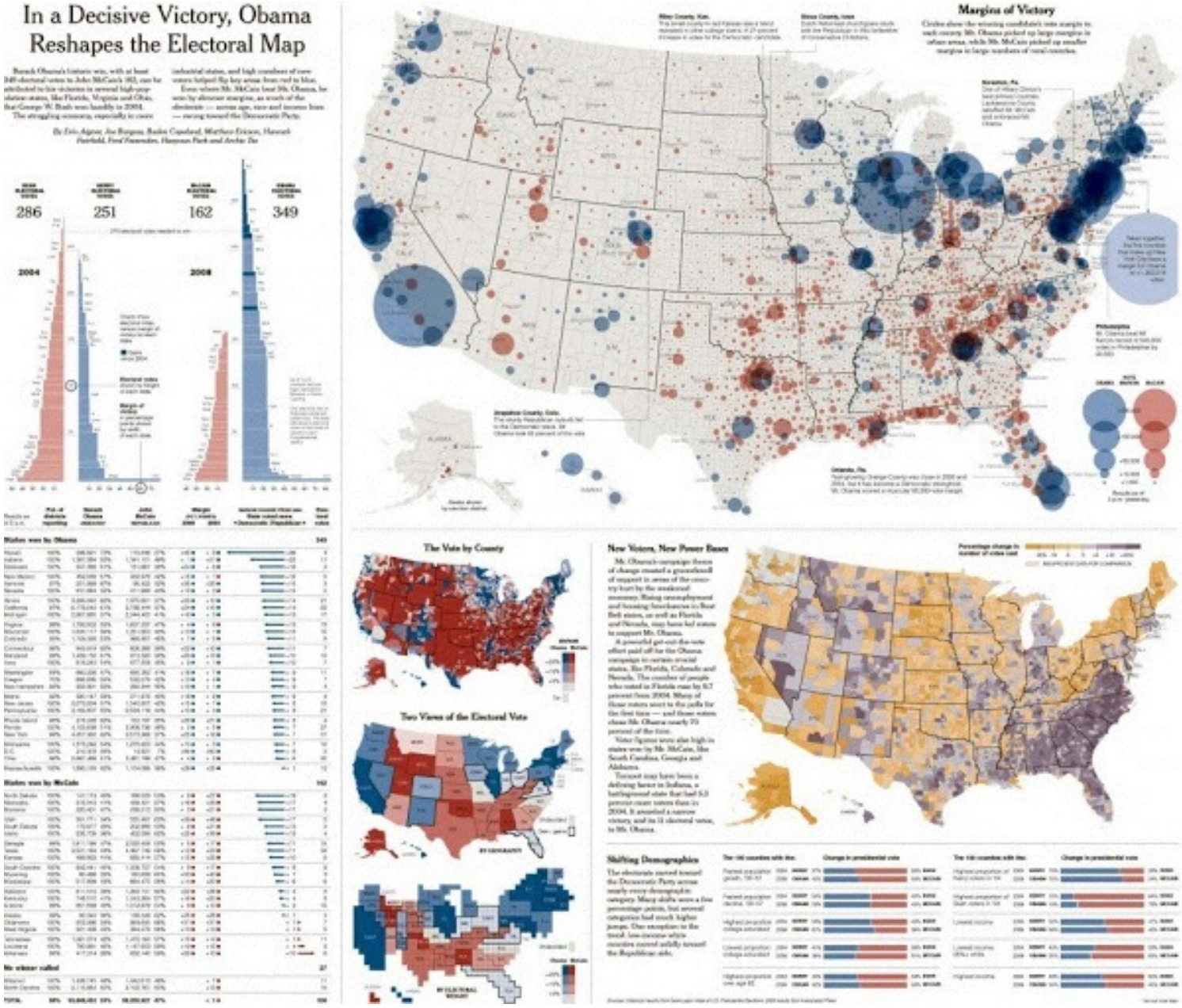
http://bl.ocks.org/mbostock/4060606

Misleading Coropleth Map



Huffington Post, http://elections.huffingtonpost.com/2012/results

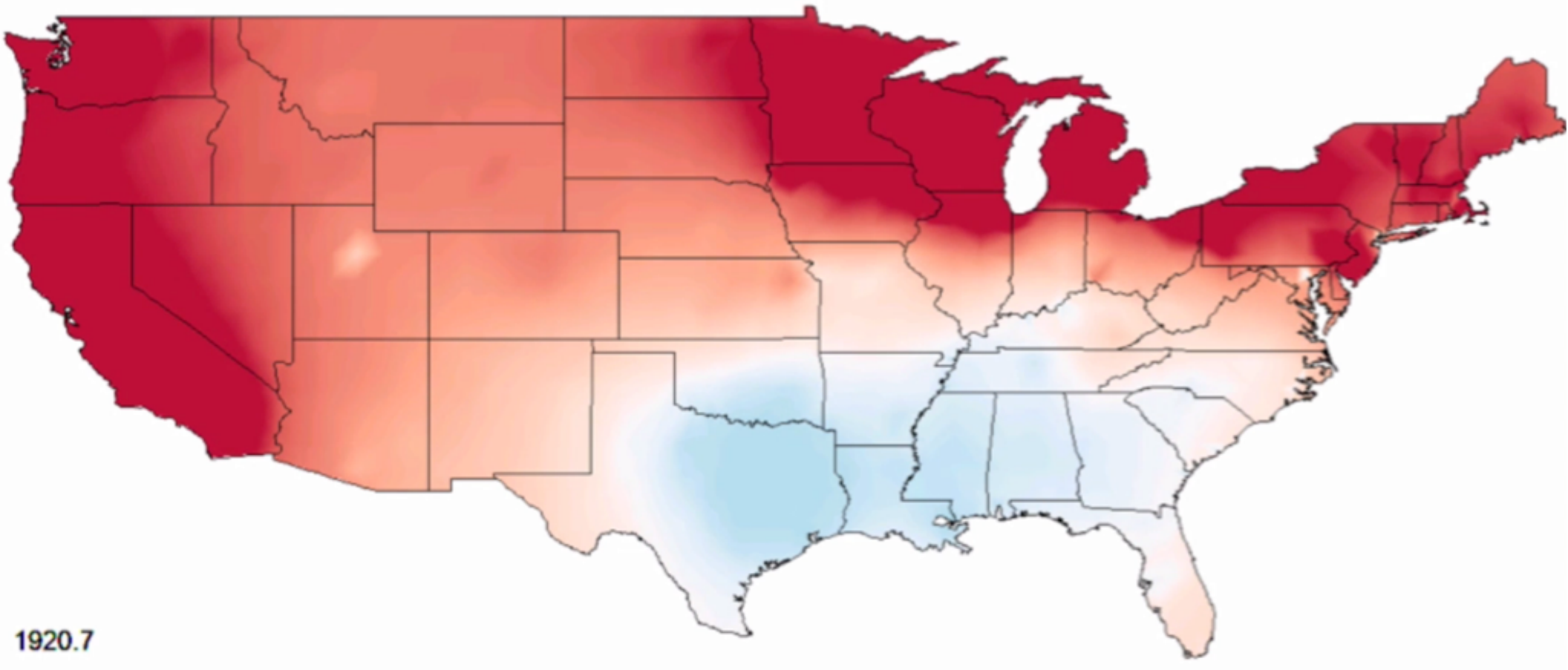
Better Version by NYT



http://www.thefunctionalart.com/2012/11/when-choropleth-maps-deceive.html

Isarithmic Map

Color coding continuous phenomena

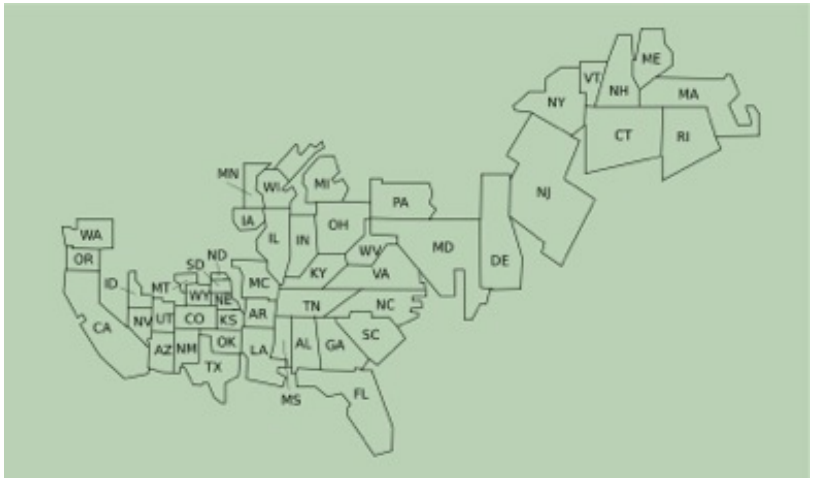


http://ecopolitology.org/2010/11/23/stunning-animated-time-lapse-map-of-red-blue-america-1920-2008/

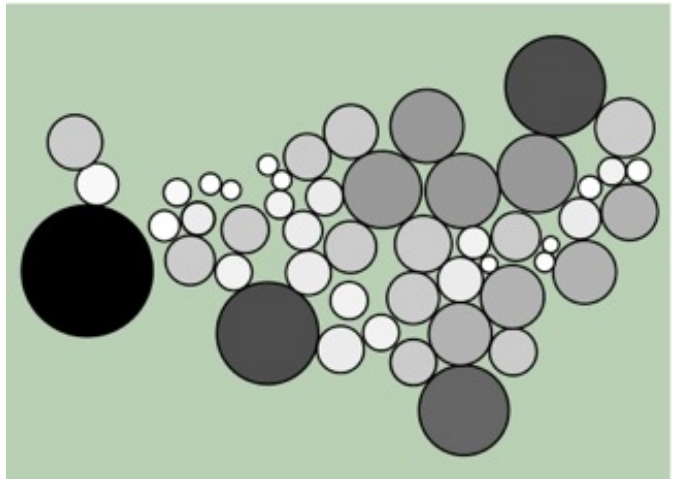
Cartograms

Size of region scaled to attribute value

Noncontinuous cartogram



Noncontiguous cartogram

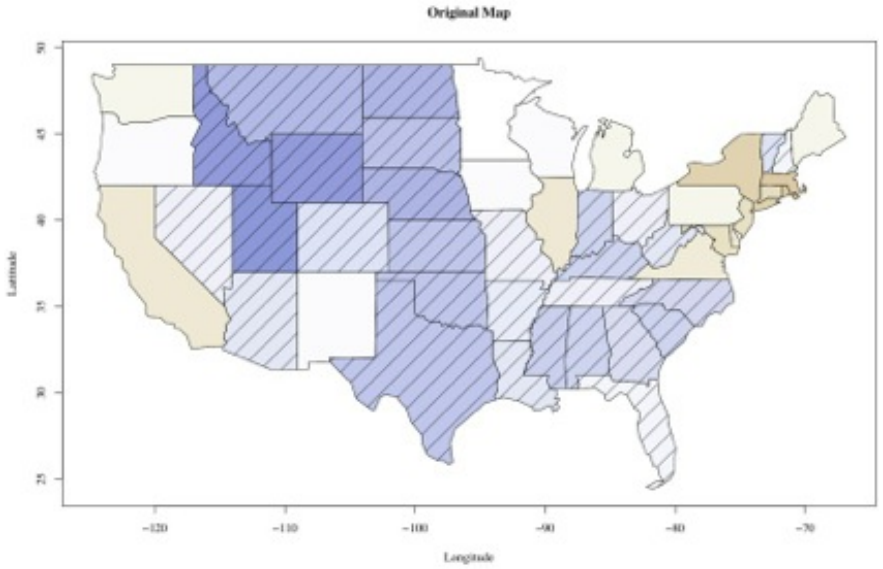


Circular cartogram

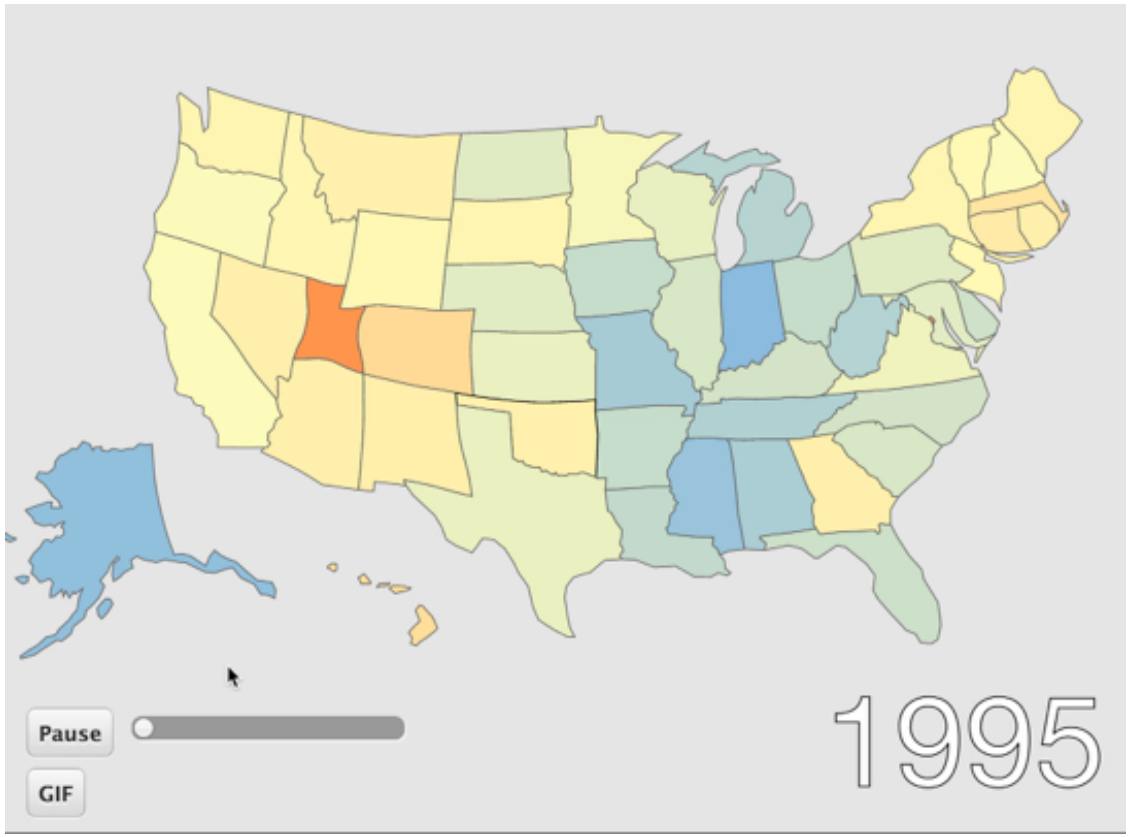
Continuous cartogram

Keim 2005

Continuous Cartogram Example

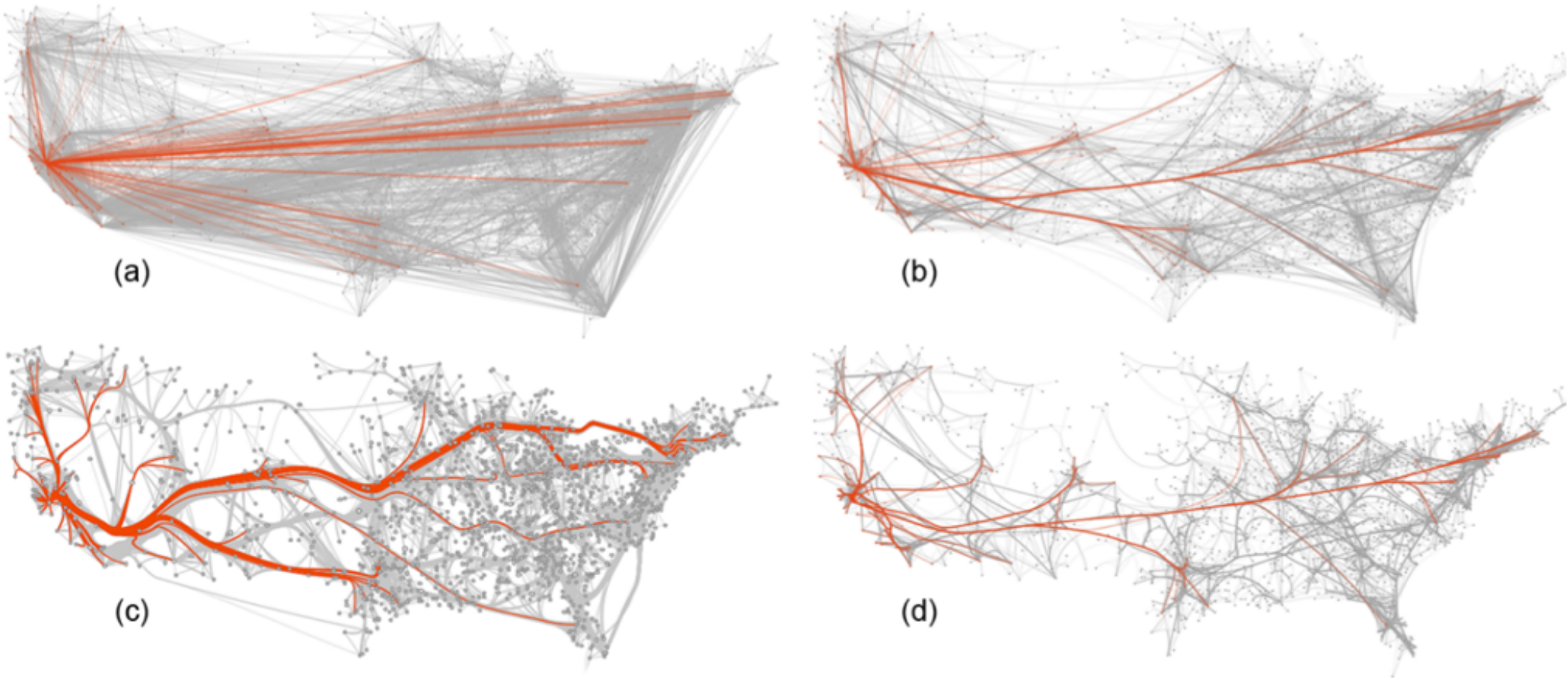


Keim 2003



http://tributary.io/inlet/5651806

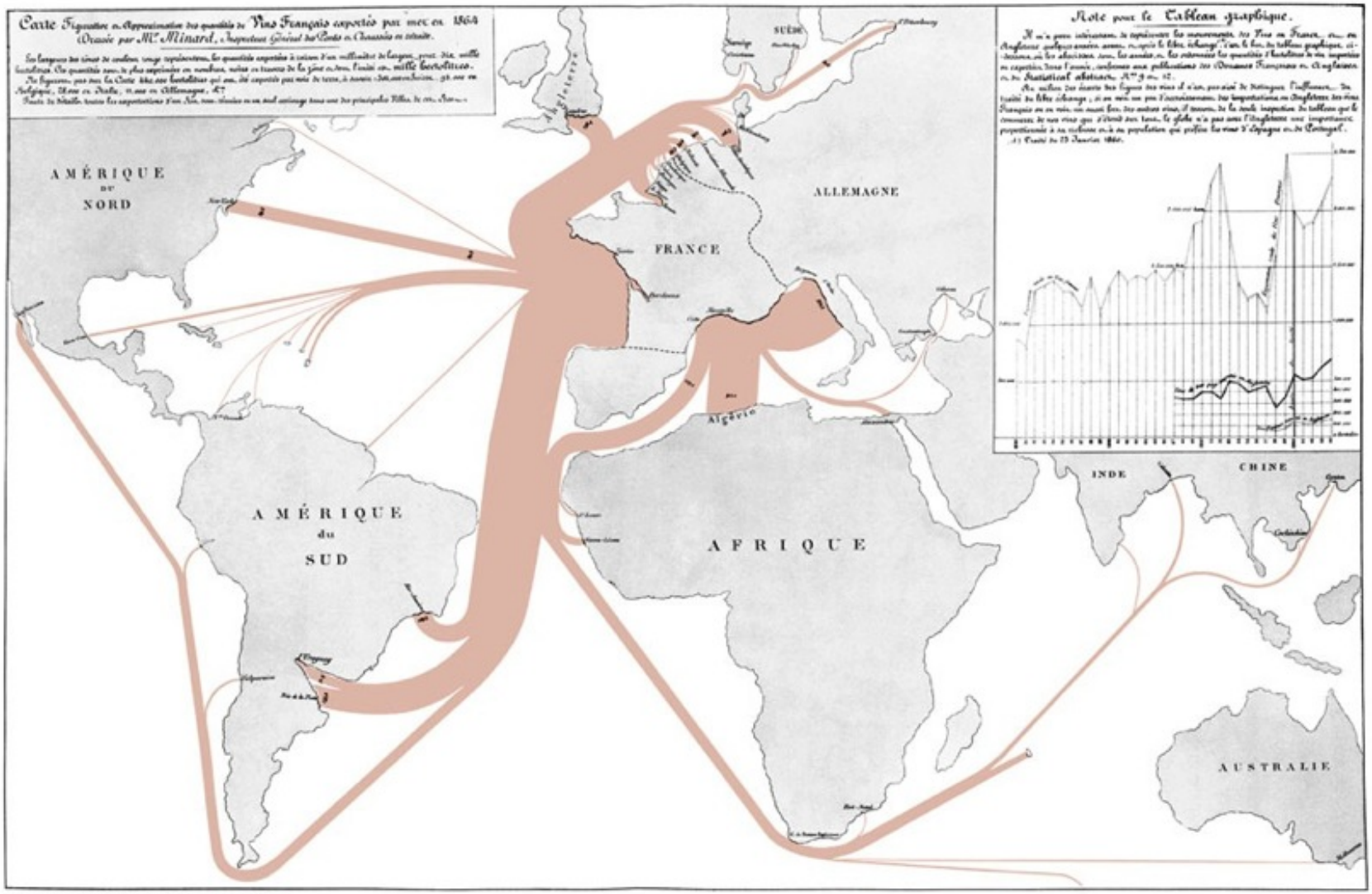
Edges Augmented Onto Map



Holten 2009

Force Directed Edge Bundles
US migration graph (1715 nodes, 9780 edges)

Flow Maps



Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard*, 1845–1869, a portfolio of his work held by the Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris.

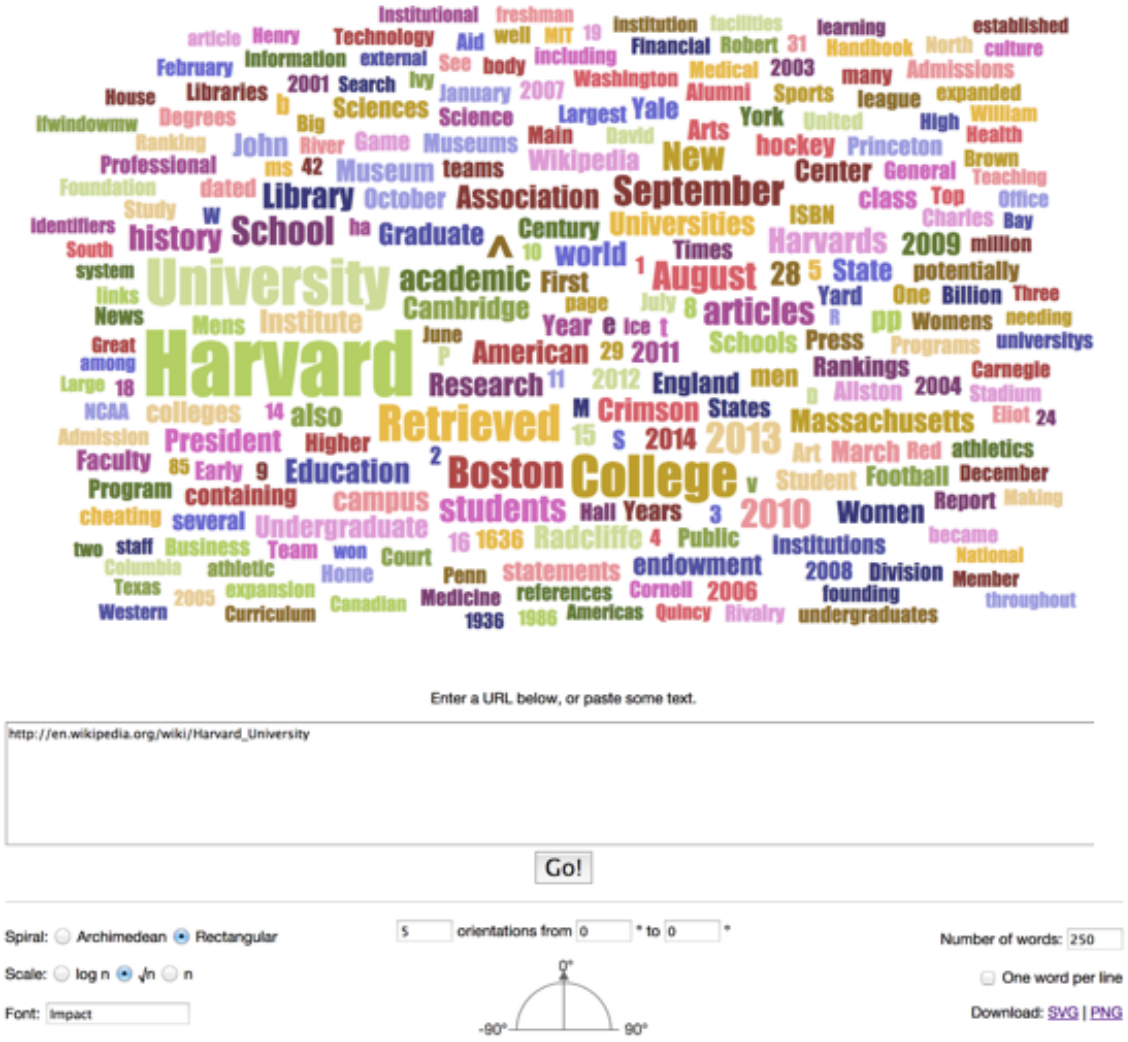
Text Visualization

$59,413,405,476,974

The outstanding United States public debt as of Sept 15, 2014.

Tag Cloud / Word Cloud / Wordle

Change word size/color by frequency

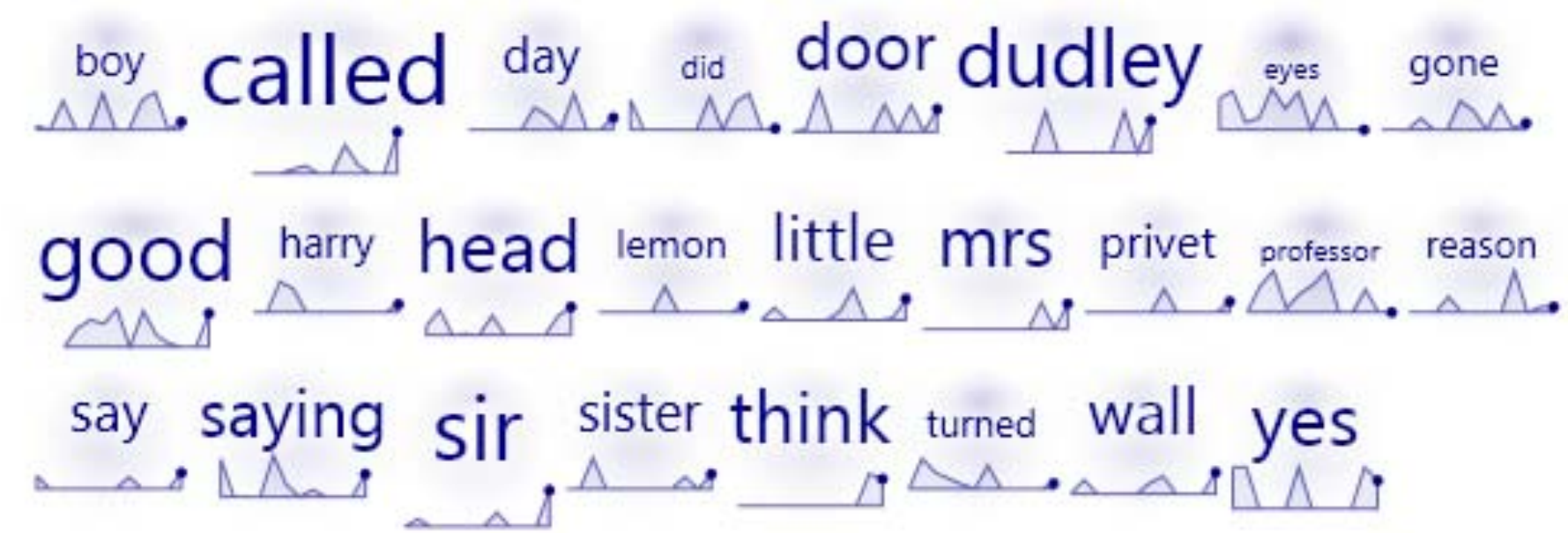


http://www.jasondavies.com/wordcloud/

Spark Clouds

Lee 2010

Convey trends between multiple tag clouds over time



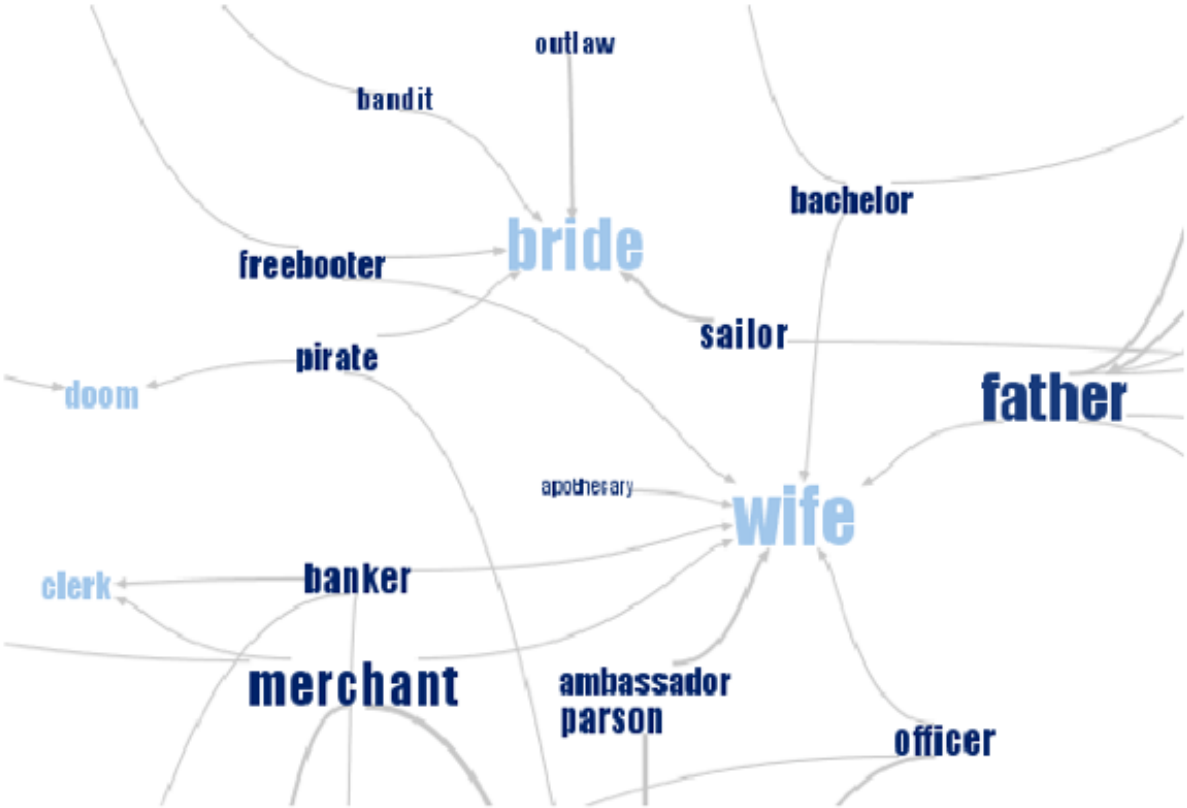
http://research.microsoft.com/en-us/um/redmond/groups/cue/infovis/

Example: Phrase Net

van Ham et al. 2009

Visual overviews of unstructured text

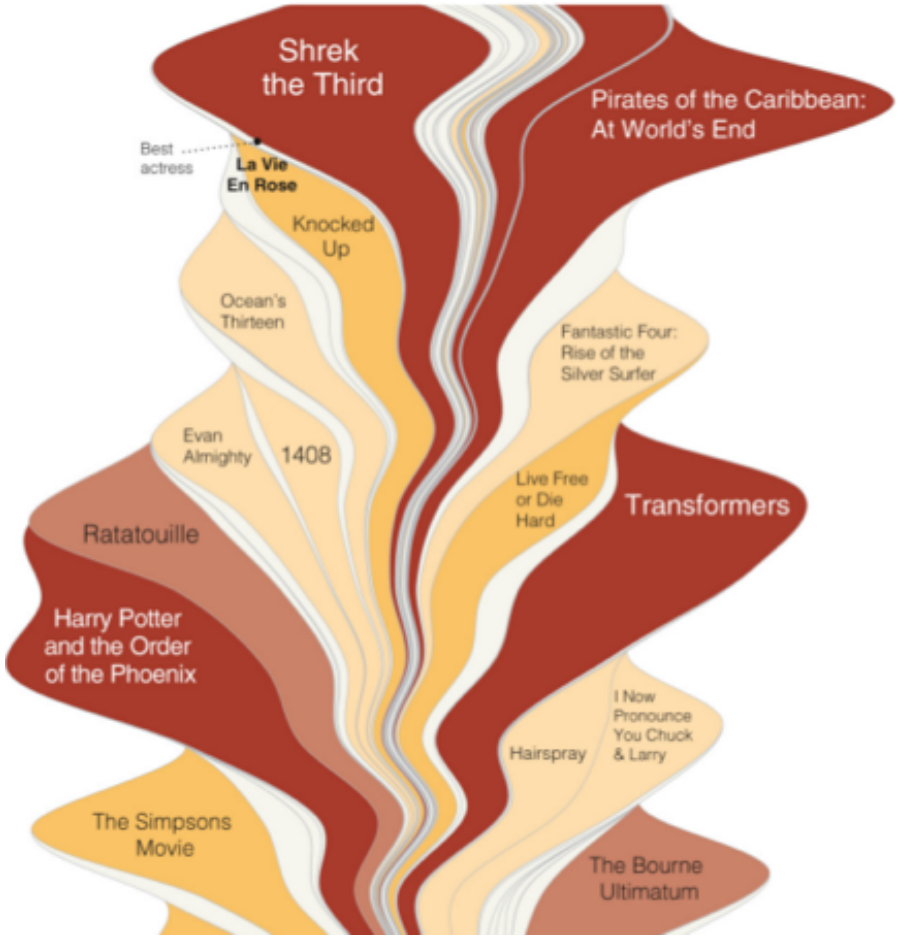
Graph; nodes = words; edges = linked by user specified relation



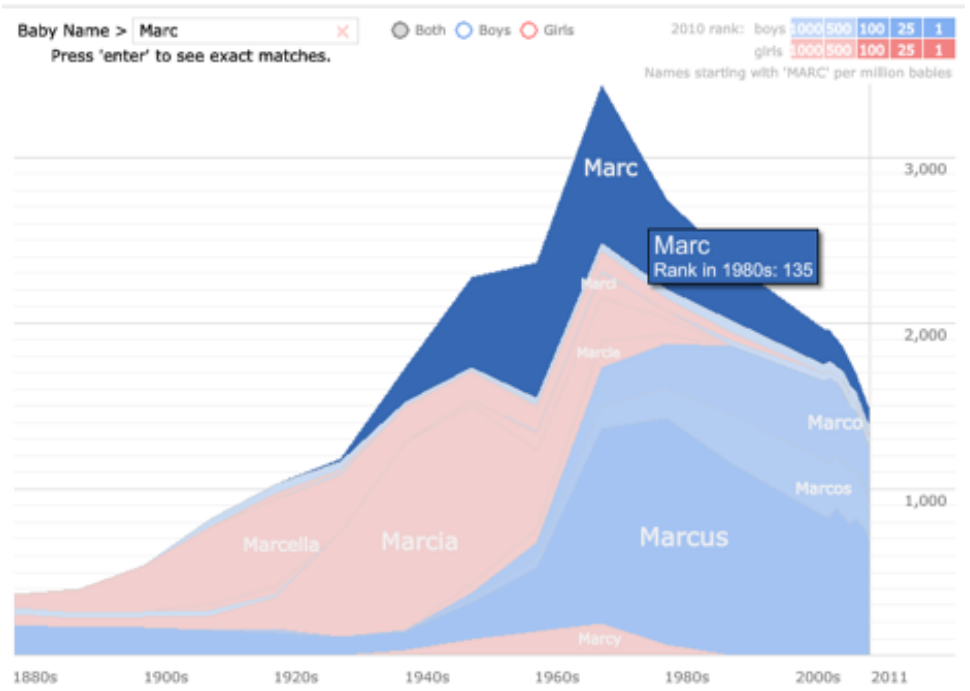
Relations in 18th and 19th century novels

Theme River / Stream Graph

Thematic changes over time. Height = frequency

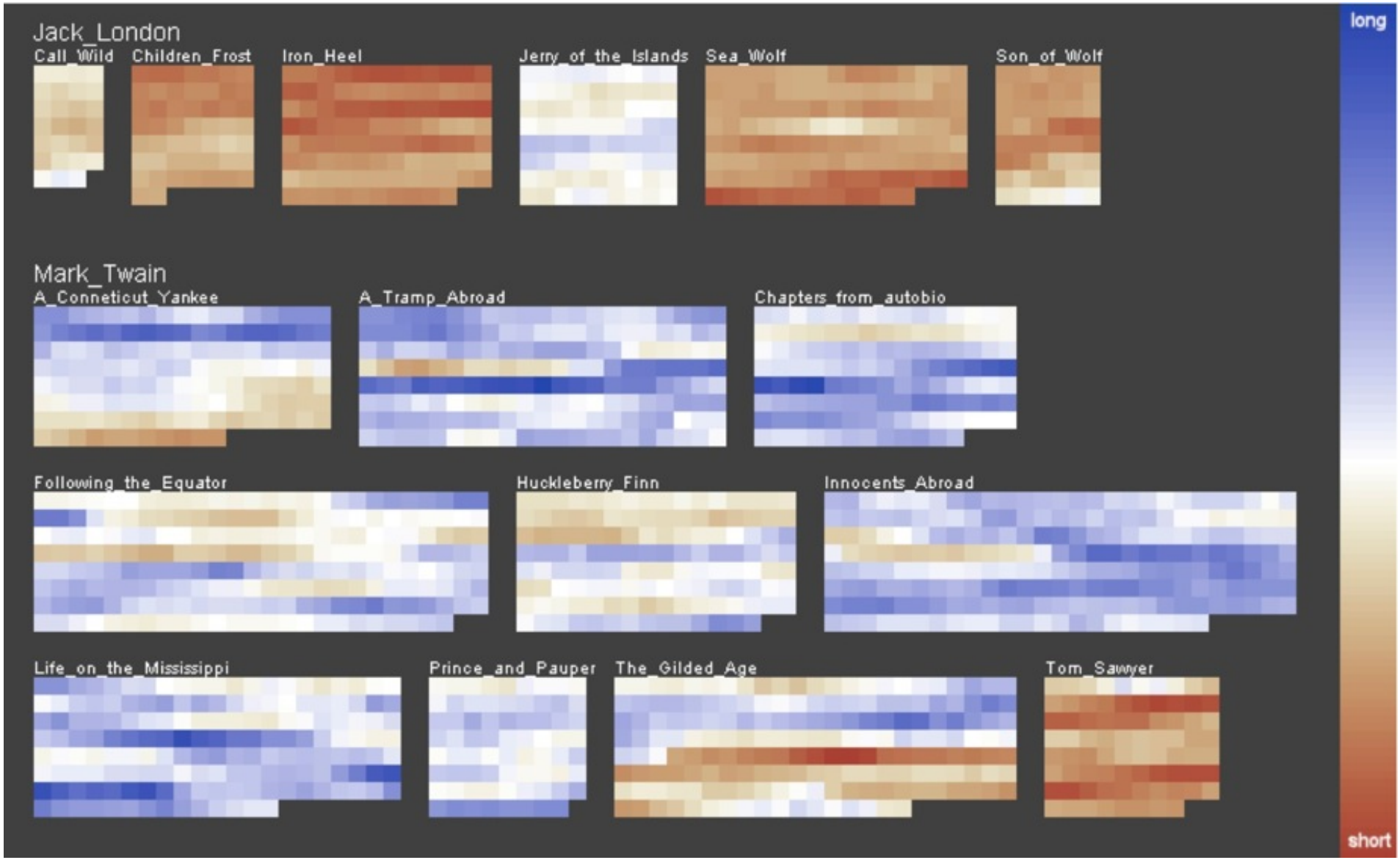


StreamGraph: Films from summer 2007
[Byron & Wattenberg 2008]



Baby Name Voyager
http://www.babynamewizard.com/
Wattenberg 2005

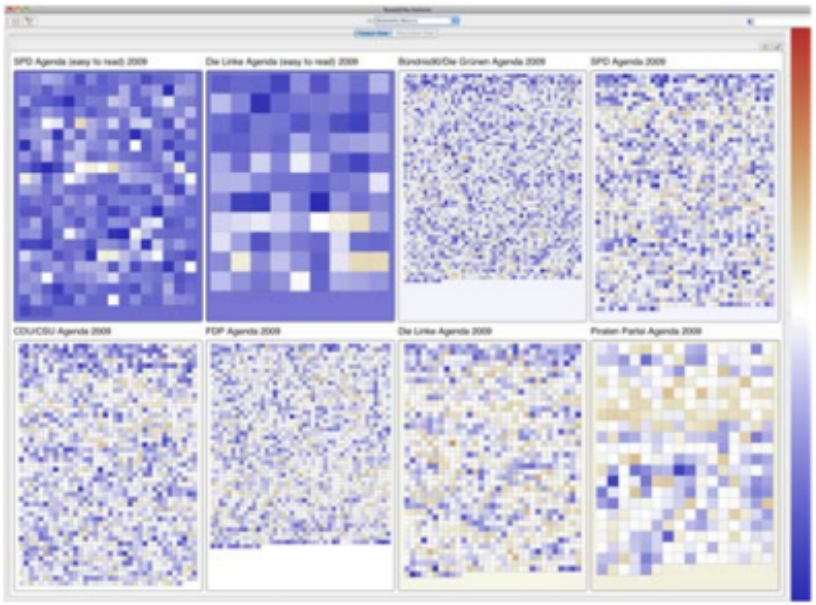
Example: Literature Fingerprinting



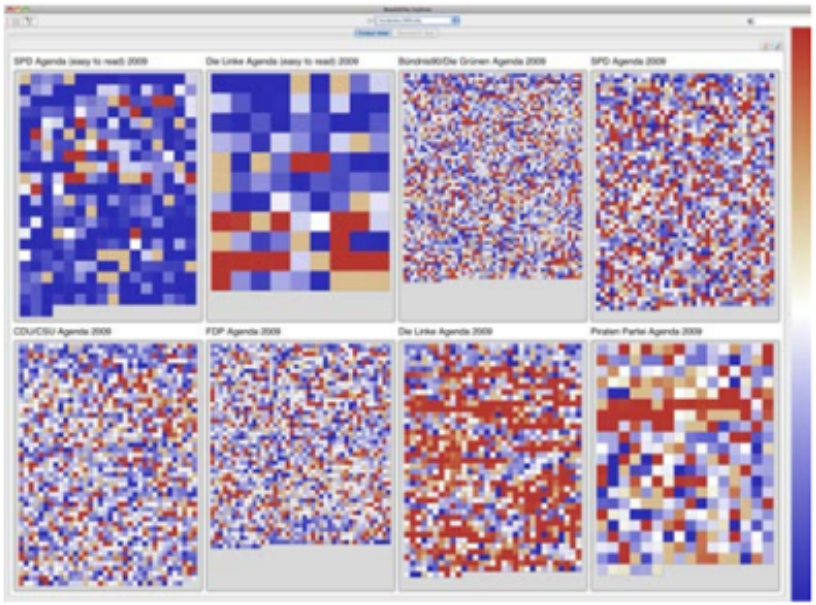
Visualize several text measures to discriminate between authors.
Pixel = text block, Group = book, color = average sentence length

Keim and Oelke 2007

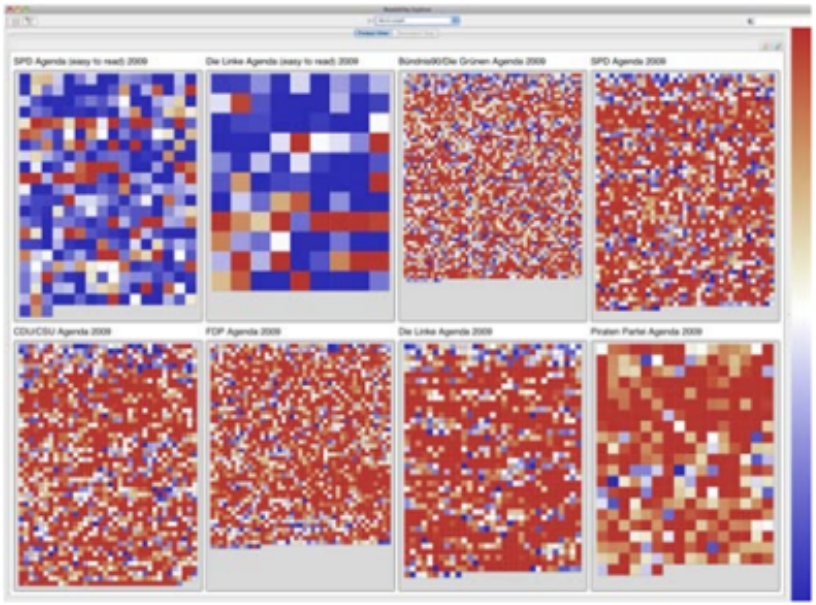
Example: Readability Analysis



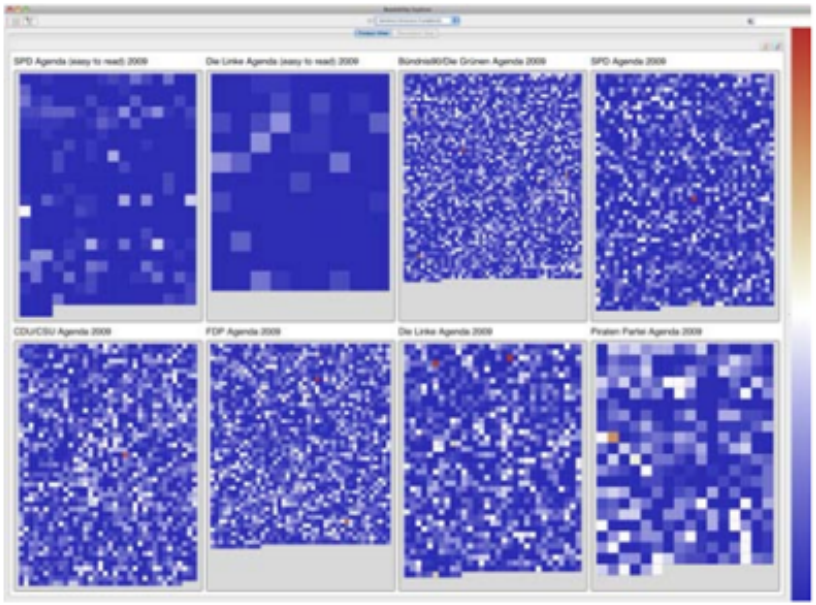
(a) Average Readability Score



(b) Feature: Vocabulary Difficulty



(c) Feature: Word Length



(d) Feature: Sentence Structure Complexity

Oelke 2007

Summary

Major Concepts:

- Data mining Terminology
- Visualization basics
- Graphical Integrity
- Graph Types for 2D and nD

Slide Material References

- Slides from Harvard CS 109 (2013 and 2014)