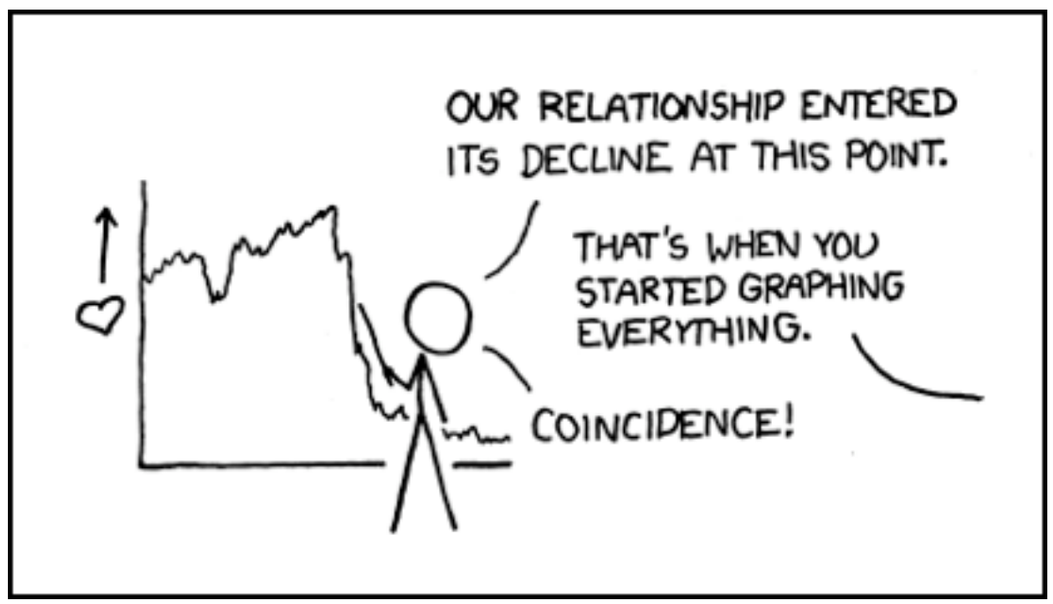
CS109 Data Science
Statistical Graphs

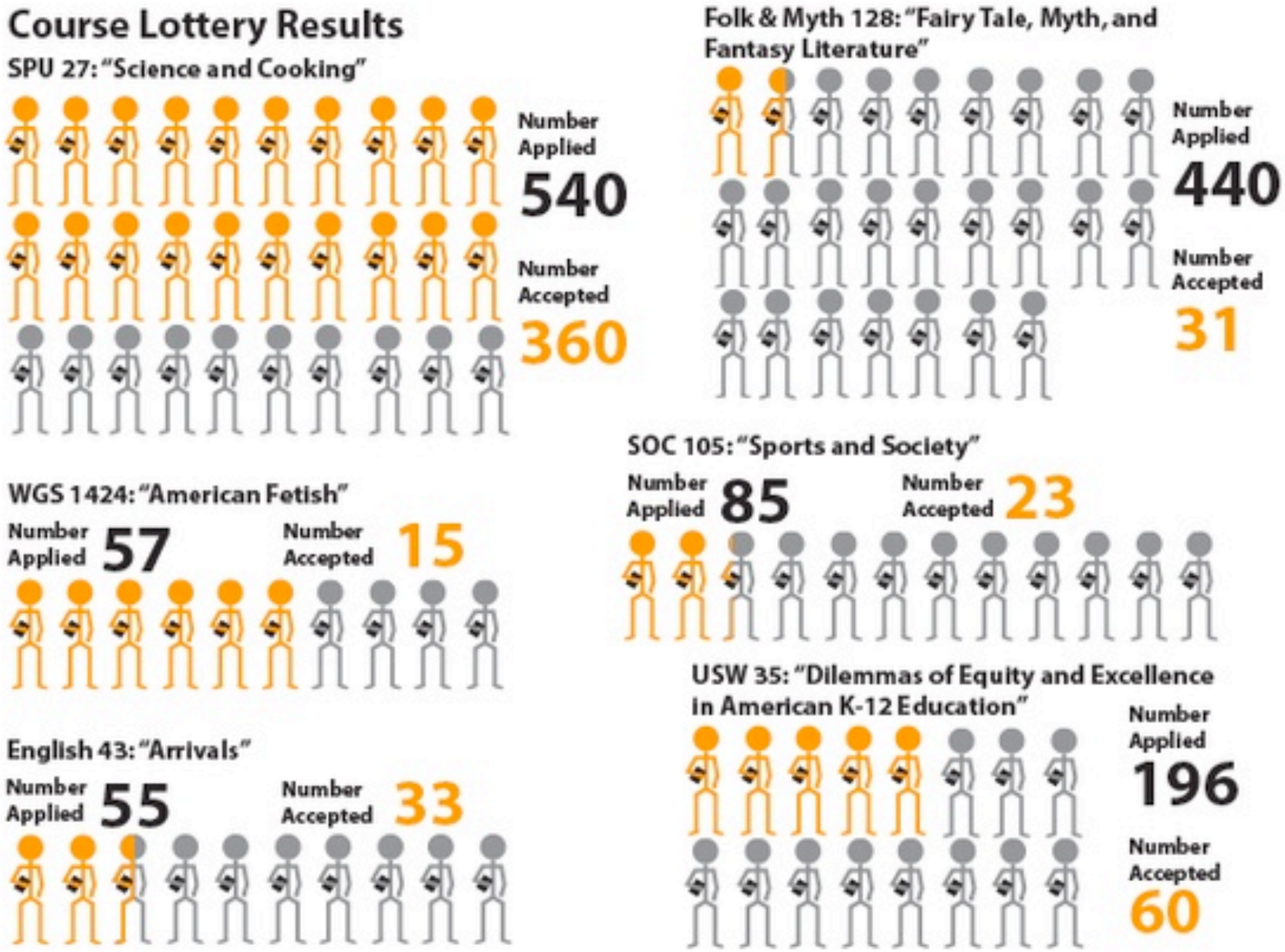
Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



xkcd

This Week

- HW0 - due today Tuesday (not graded)
- HW1 - due Thursday, Sept 19 - start now!
- Friday lab **10-11:30 am** in MD G115
    - *Data Scraping with Python* with Ray and Johanna
- Thank you for being patient, flexible, and constructive while we work through startup issues



With Demand for Popular Courses High, Course Lotteries See Low Admissions Rates, Harvard Crimson, Sept 10, 2013

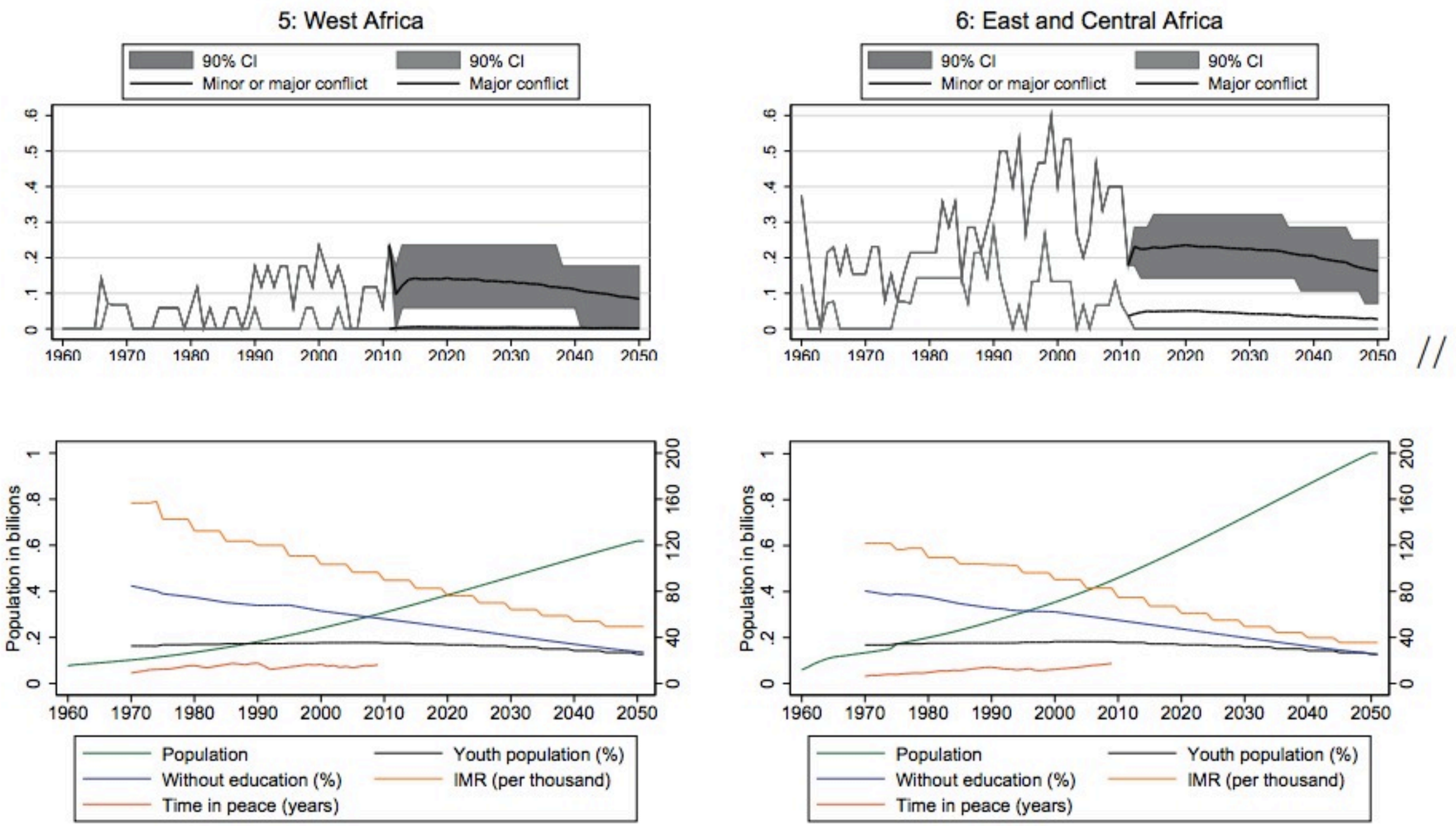


Figure 5. Predicted share of countries in conflict and average predictor values, West Africa (left) and East and Central Africa (right), 1960–2050

Peace on Earth?: The future of internal armed conflict,
Håvard Hegre, February 2013

significance
statistics making sense
ASA AMERICAN STATISTICAL ASSOCIATION
ROYAL STATISTICAL SOCIETY

Peace on Earth?: The future of internal armed conflict

FREE

Håvard Hegre

Article first published online: 15 FEB 2013
DOI: 10.1111/j.1740-9713.2013.00628.x

Issue

Significance
Volume 10, Issue 1, pages 4–8, February 2013

Additional Information (Show All)

How to Cite | Author Information | Publication History

Abstract | References | Cited By

Warfare seems endemic to mankind. Nations around the world are riven by conflict. But is the impetus to war decreasing?**Håvard Hegre** finds statistical grounds for hope.

What we need here is some peer review for statistical graphics

Posted by Andrew on 8 September 2013, 9:49 am

Under the heading, "Bad graph candidate," Kevin Wright points to this article [link fixed], writing:

> Some of the figures use the same line type for two different series.
>
> More egregious are the confidence intervals that are constant width instead of increasing in width into the future.

Indeed. What's even more embarrassing is that these graphs appeared in an article in the magazine Significance, sponsored by the American Statistical Association and the Royal Statistical Society.

Perhaps every scientific journal could have a graphics editor whose job is to point out really horrible problems and require authors to make improvements.

The difficulty, as always, is that scientists write these articles for free and as a public service (publishing in Significance doesn't pay, nor does it count as a publication in an academic record), so it might be difficult to get authors to fix their graphs. On the other hand, if an article is worth writing at all, it's worth trying to convey conclusions clearly.

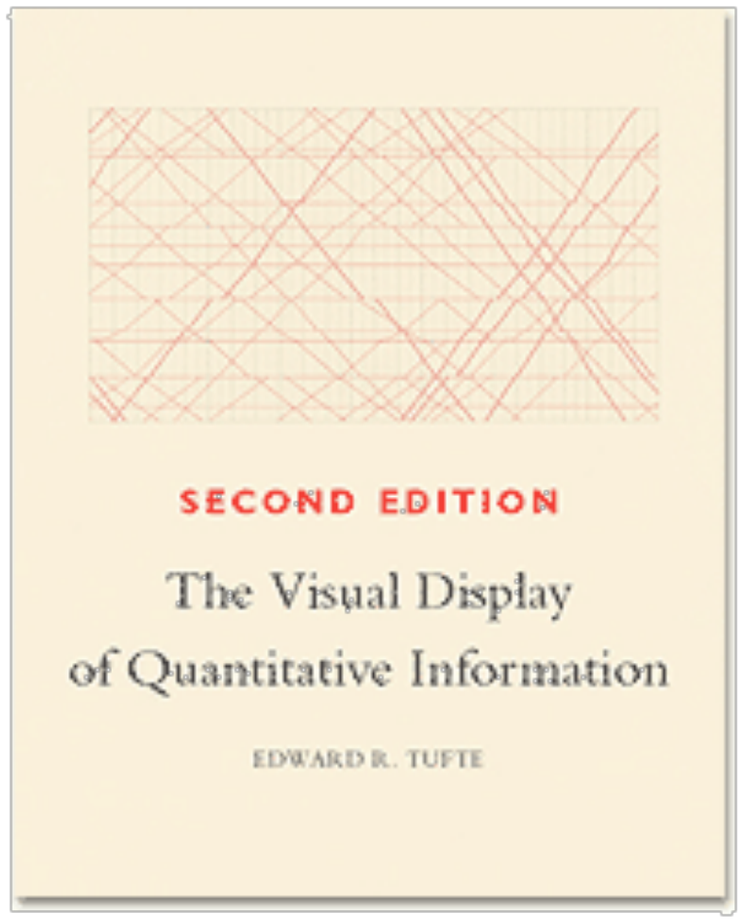
I'm not angry at the authors for publishing bad graphs—scientists typically don't get training in how to construct or evaluate graphical displays, indeed I've seen stuff just as bad in JASA and other top statistics journals—but it would be good to catch this stuff before it gets out for public consumption.

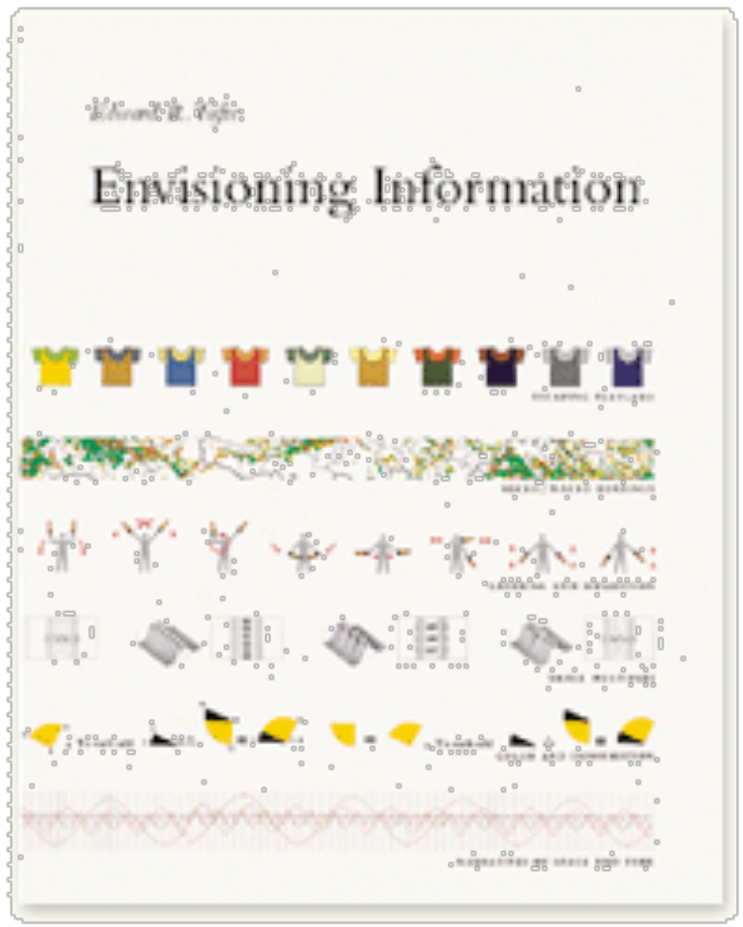
Filed under Sociology, Statistical graphics

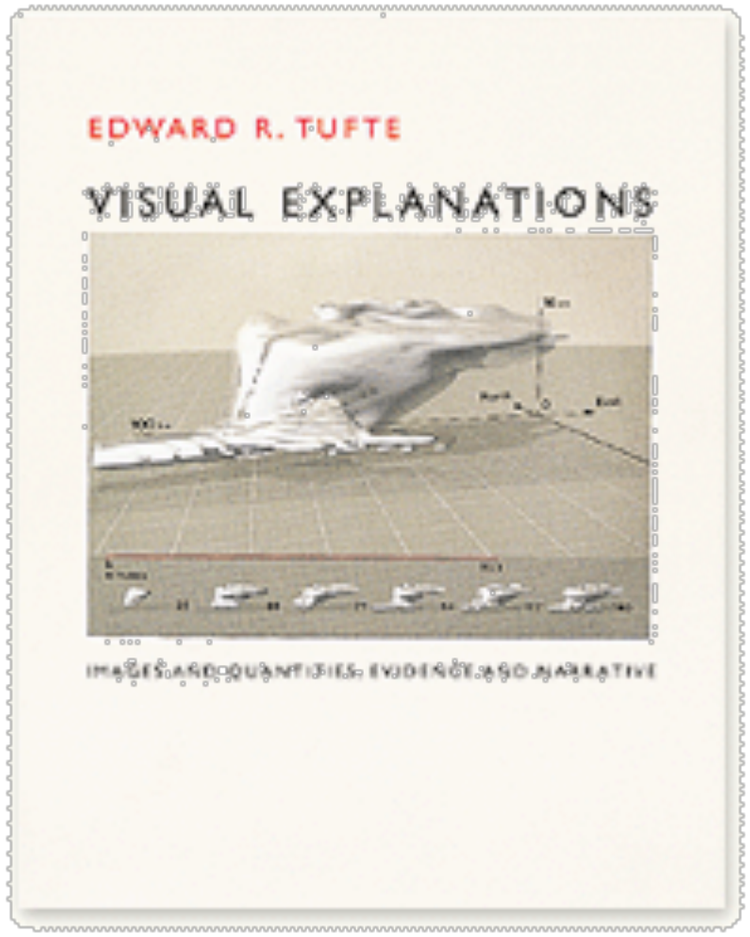
Comment (RSS) | Trackback | Permalink

Design Principles

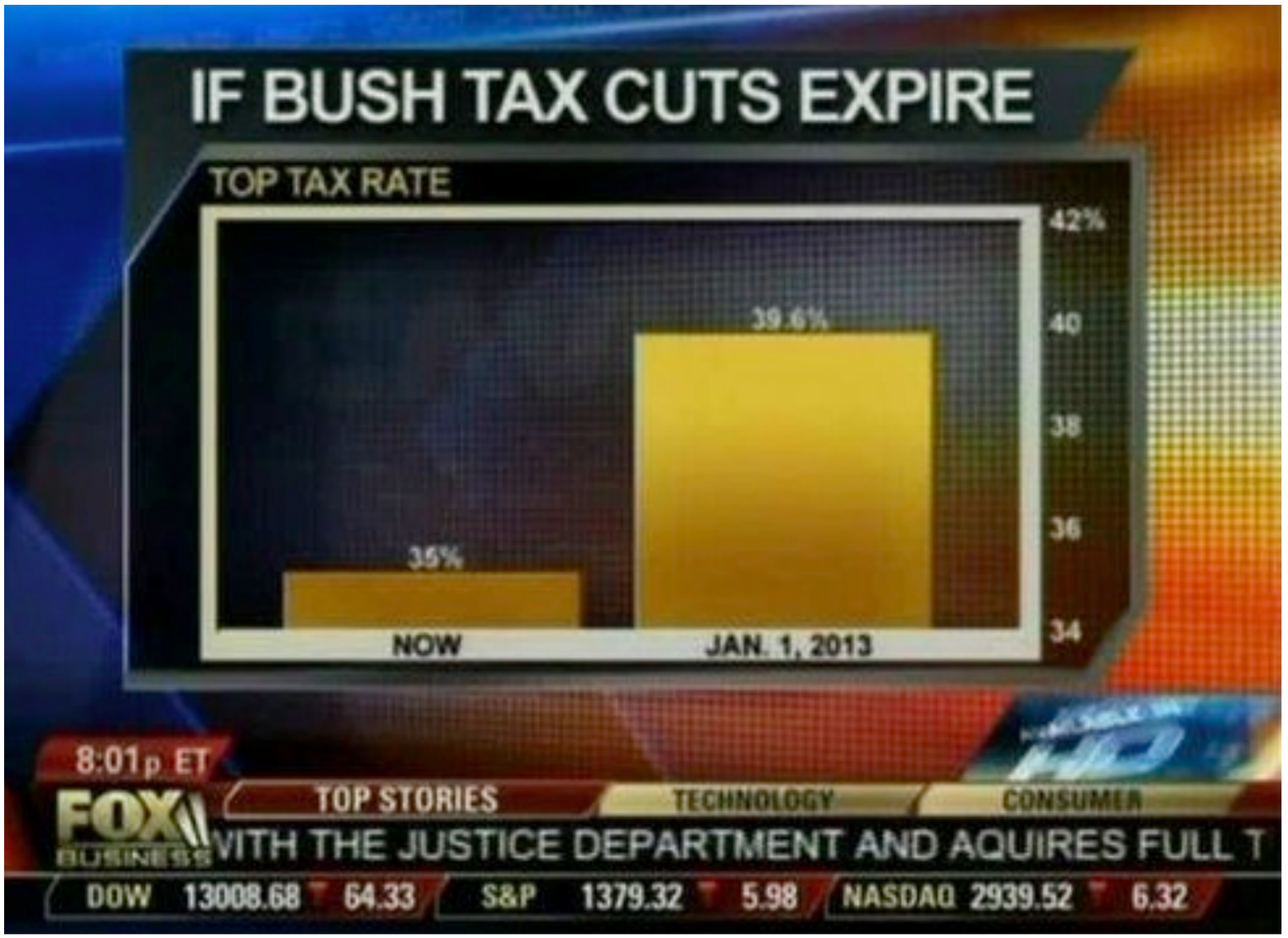
Edward Tufte





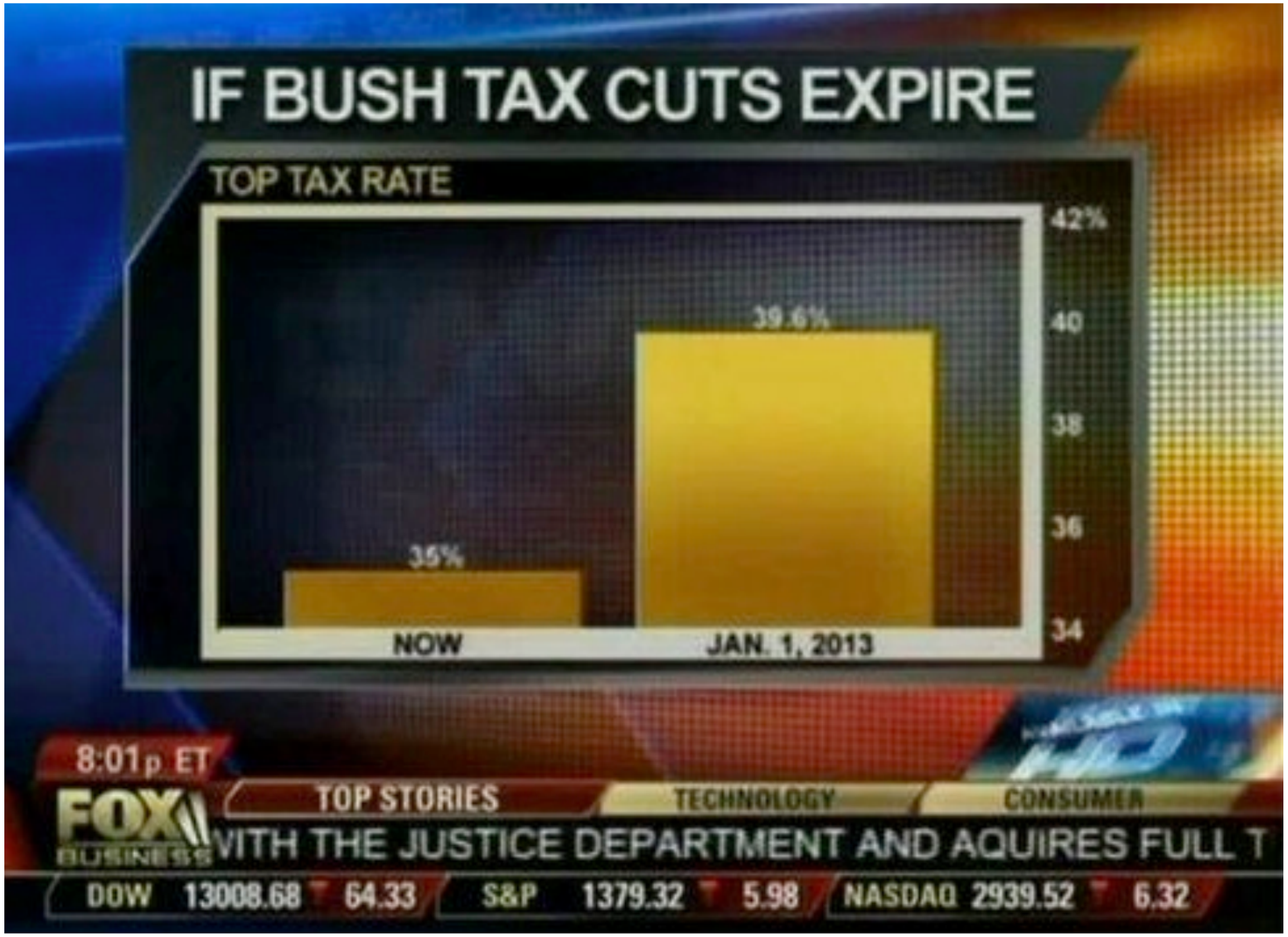


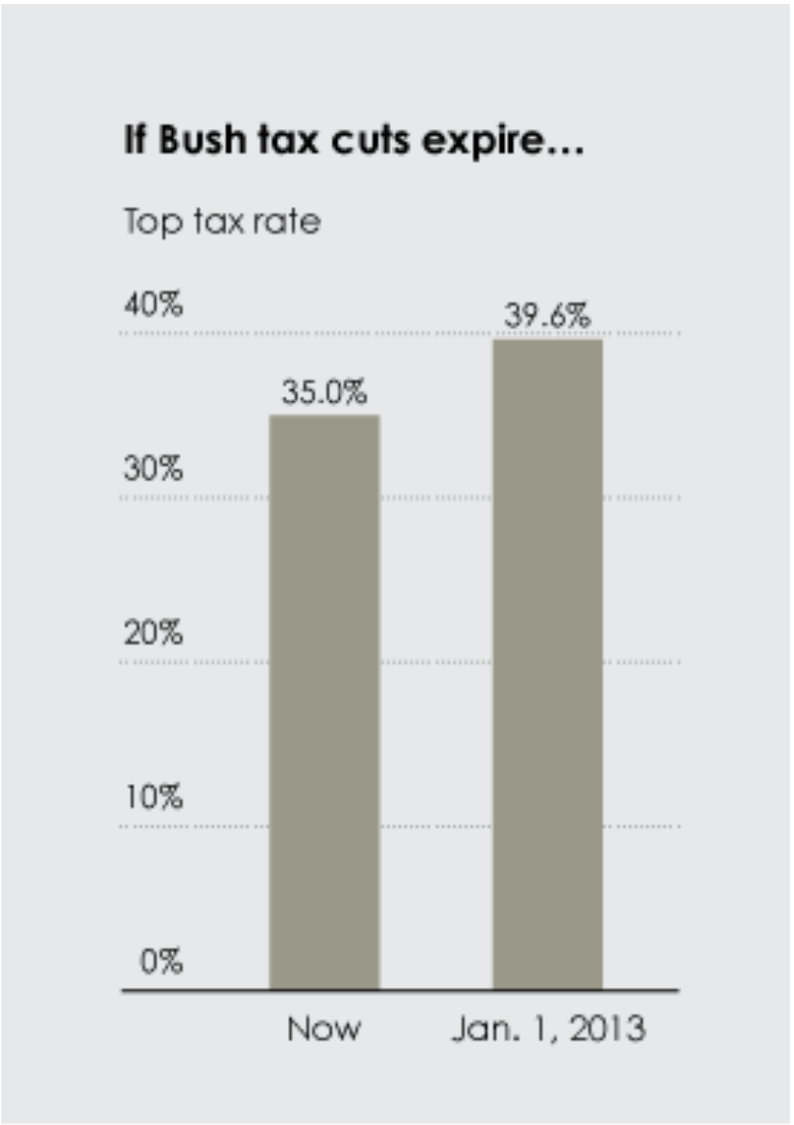
Graphical Integrity



Scale Distortions

BAR GRAPHS SHOULD ALWAYS START AT 0

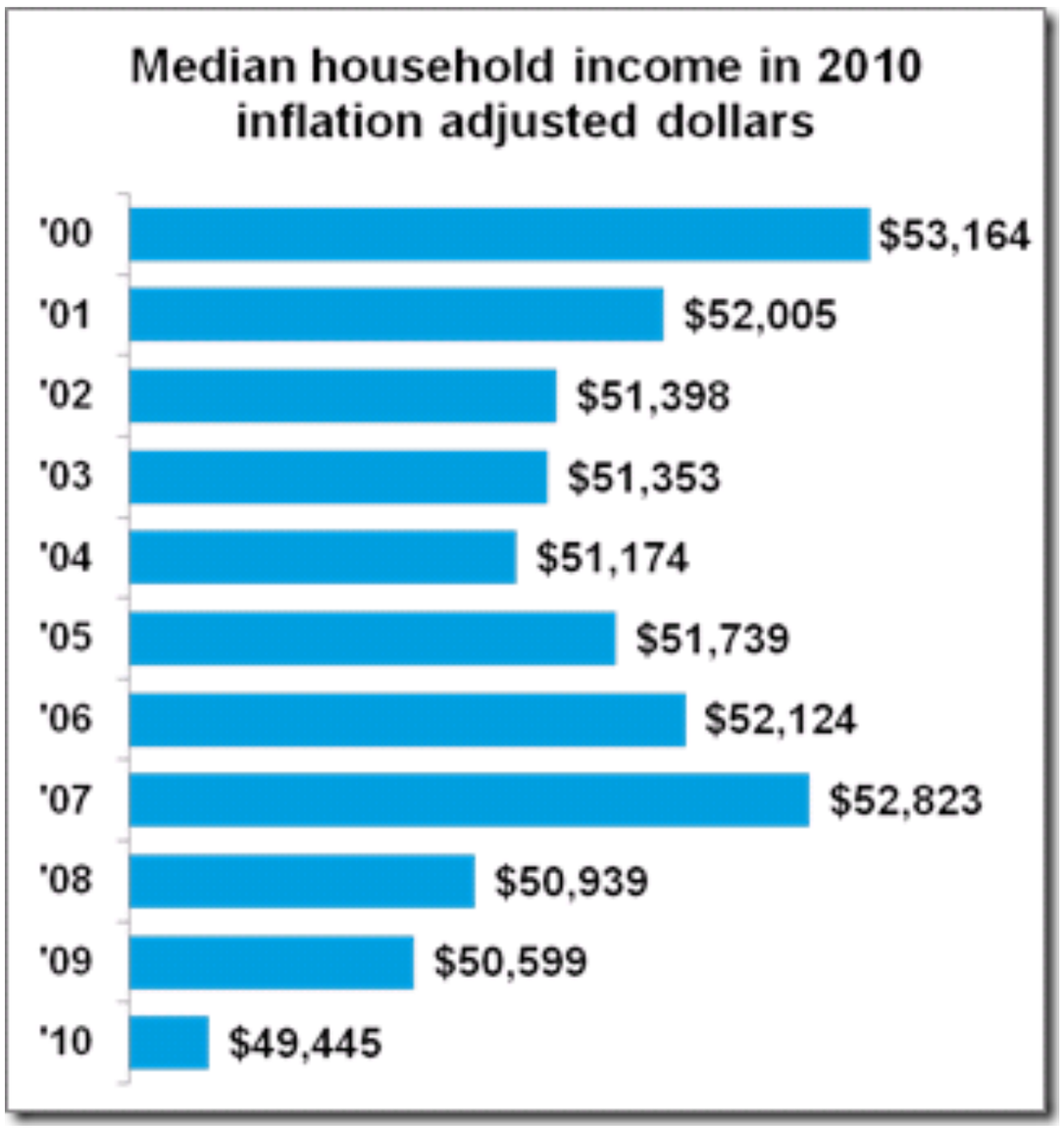




Scale Distortions

Scale Distortions

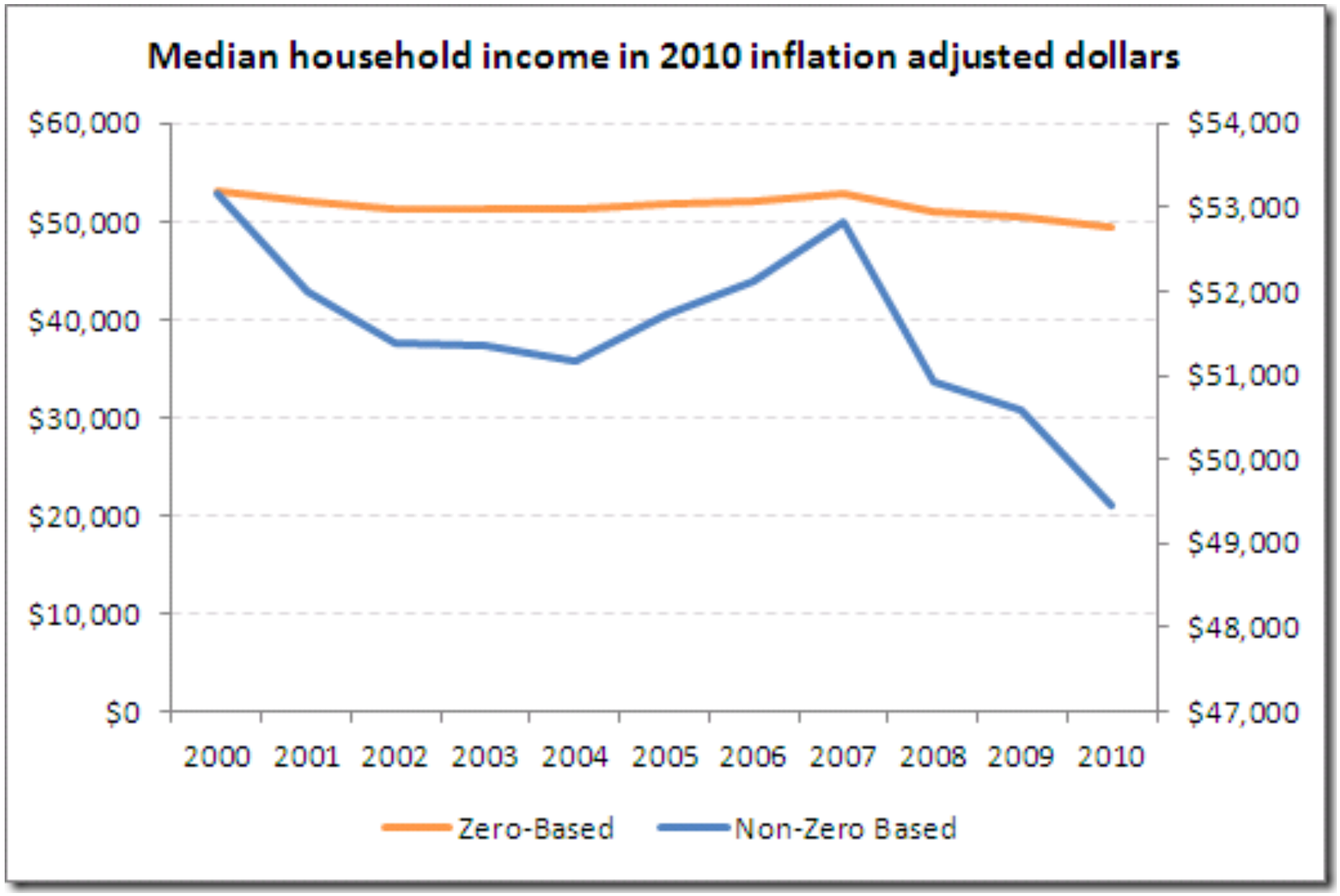
Always start your bar graphs at zero!



Nielson

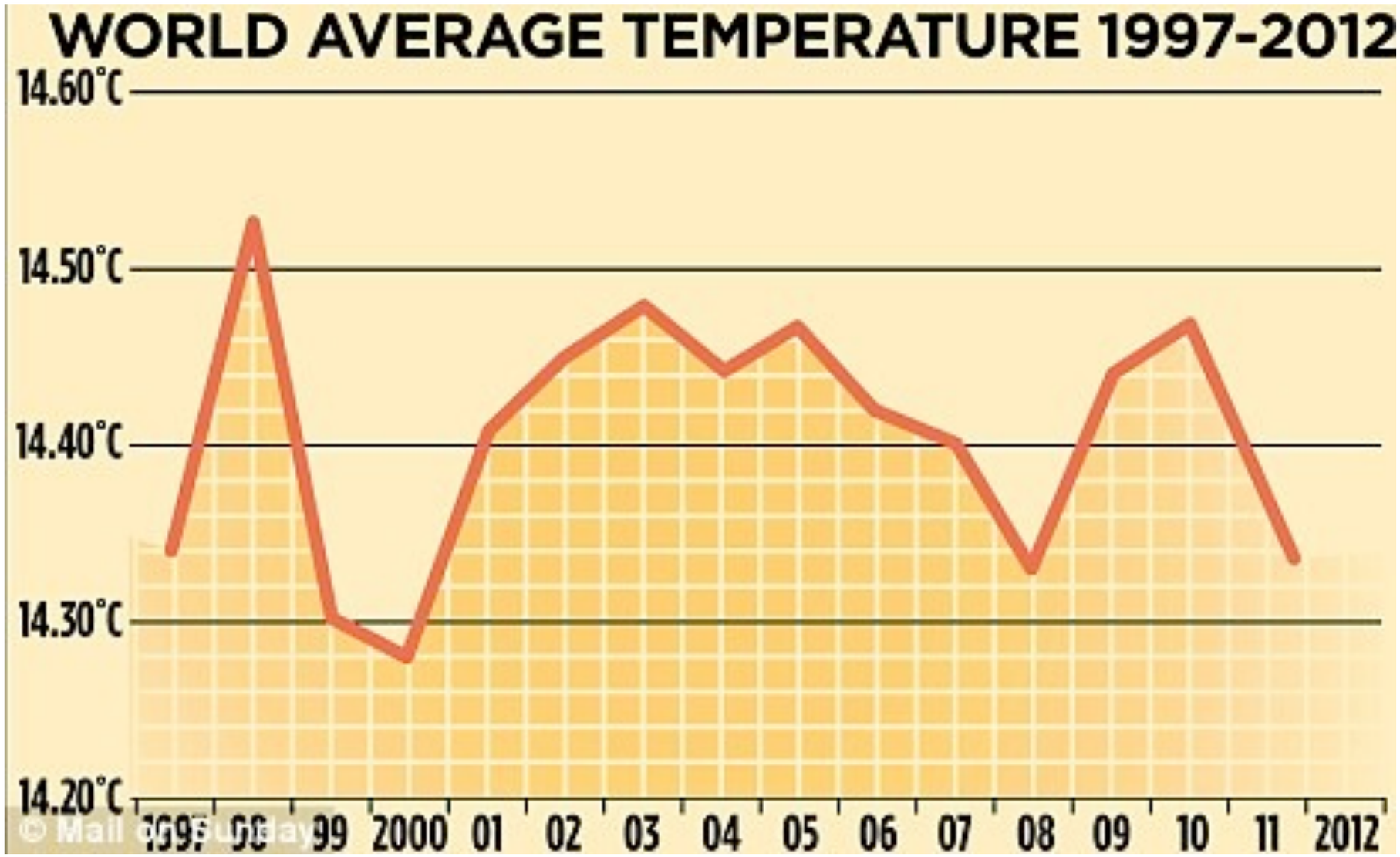
Scale Distortions

Line Graphs do not need to start at 0, but there should be an acknowledgment for the image below that the scales were changed



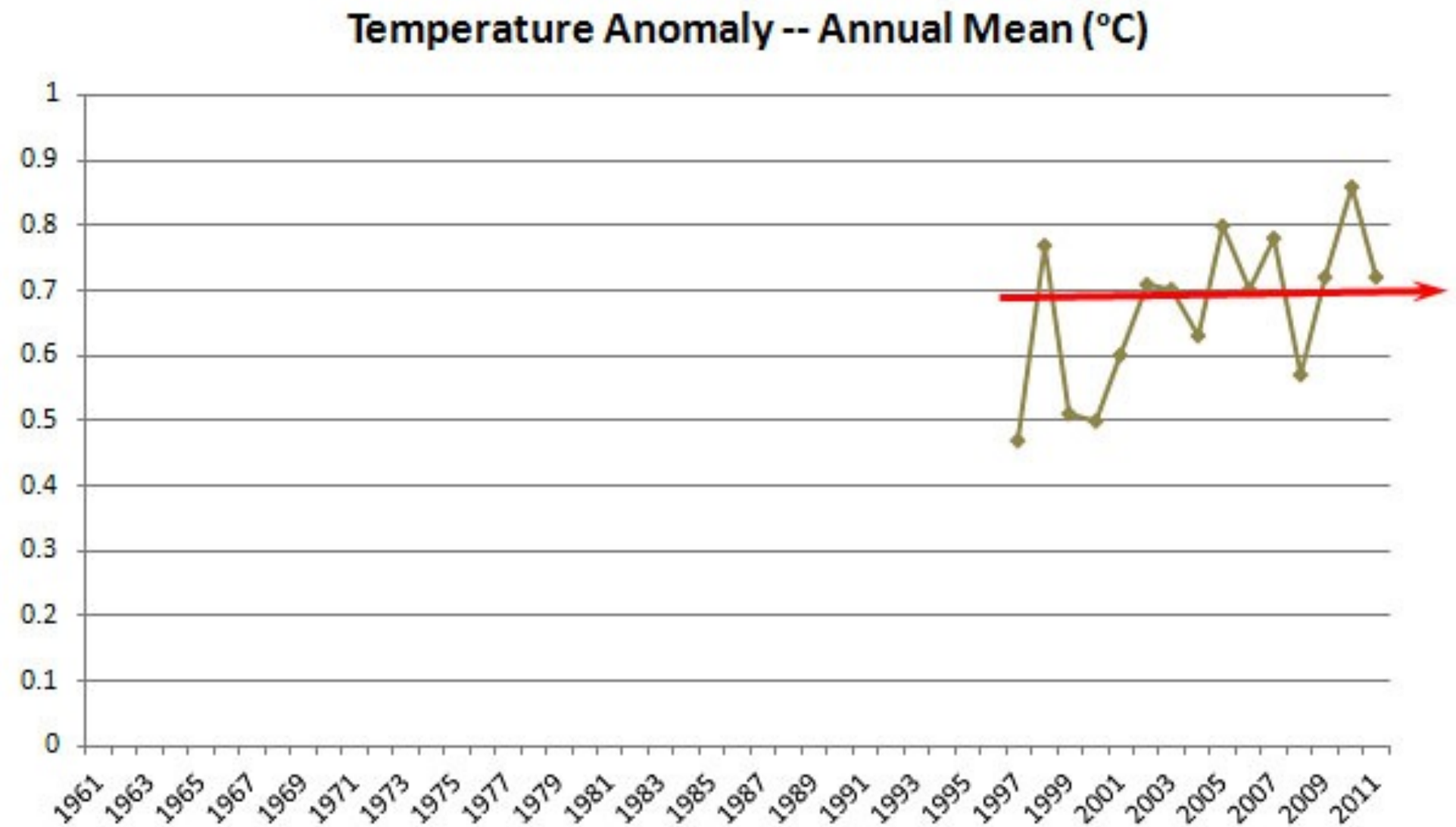
A. Kriebel, VizWiz

Global Warming?



The Daily Mail, UK, Jan 2012

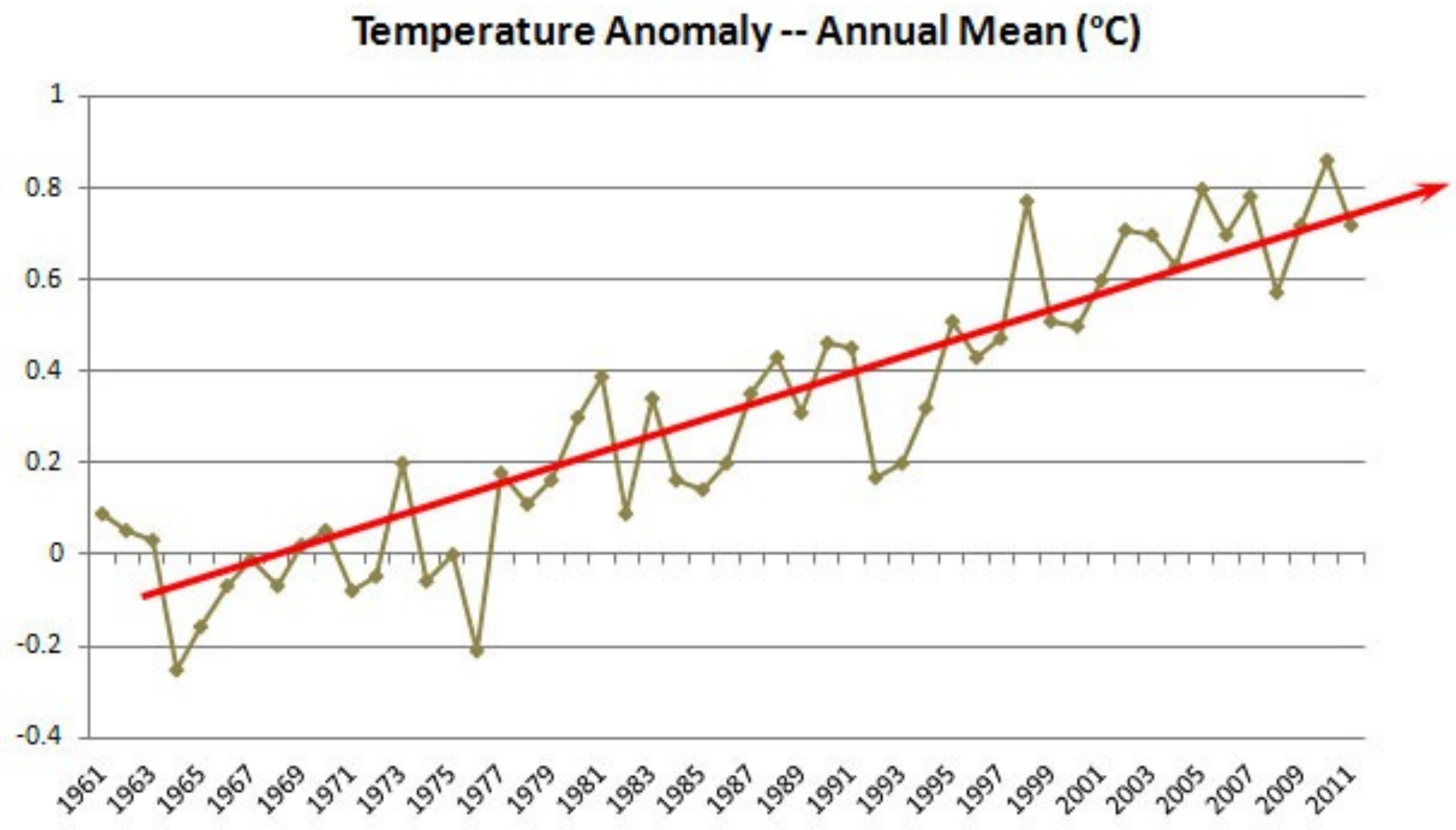
Global Warming?



Mother Jones

Global Warming!

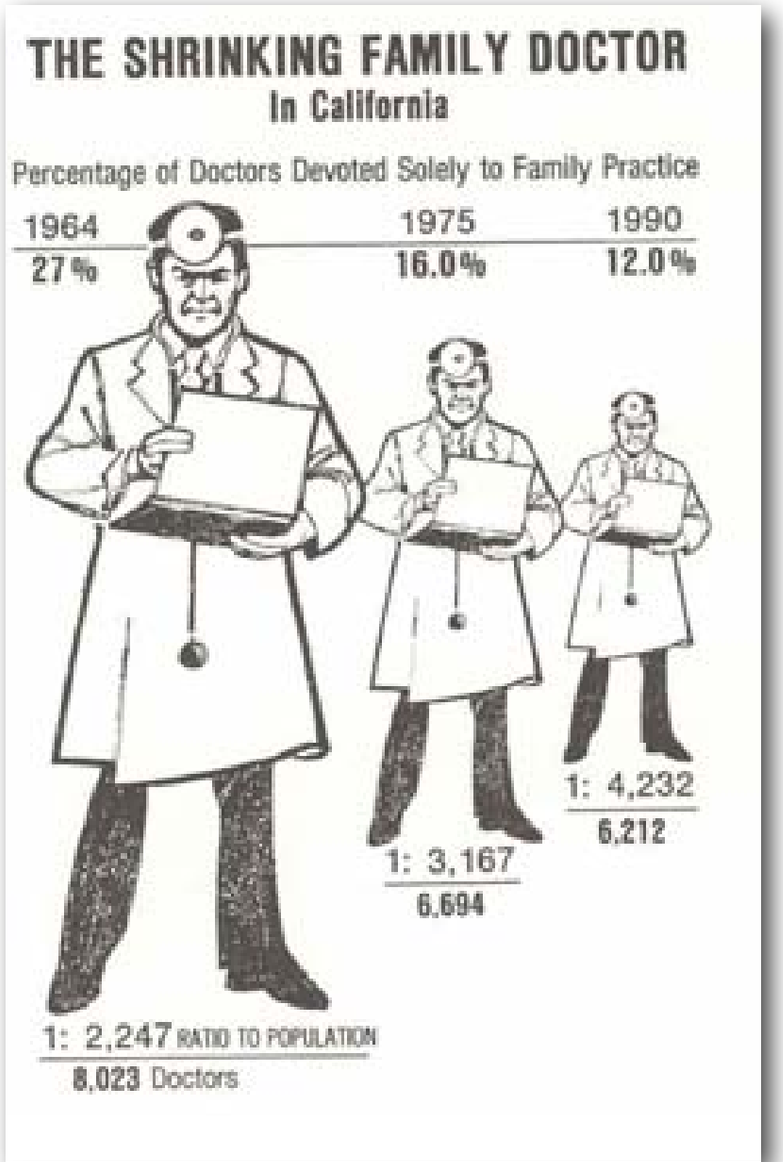
Make sure that you show enough “context”

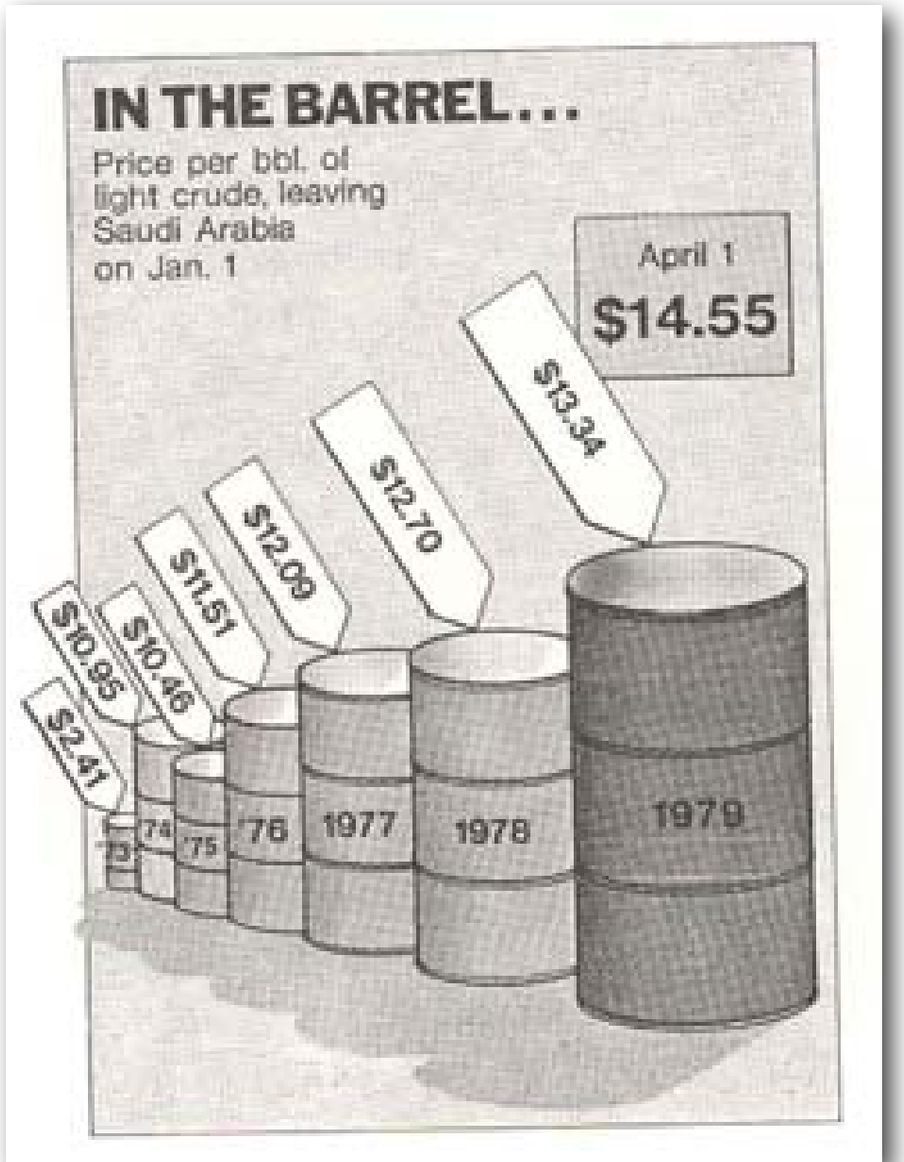


Mother Jones

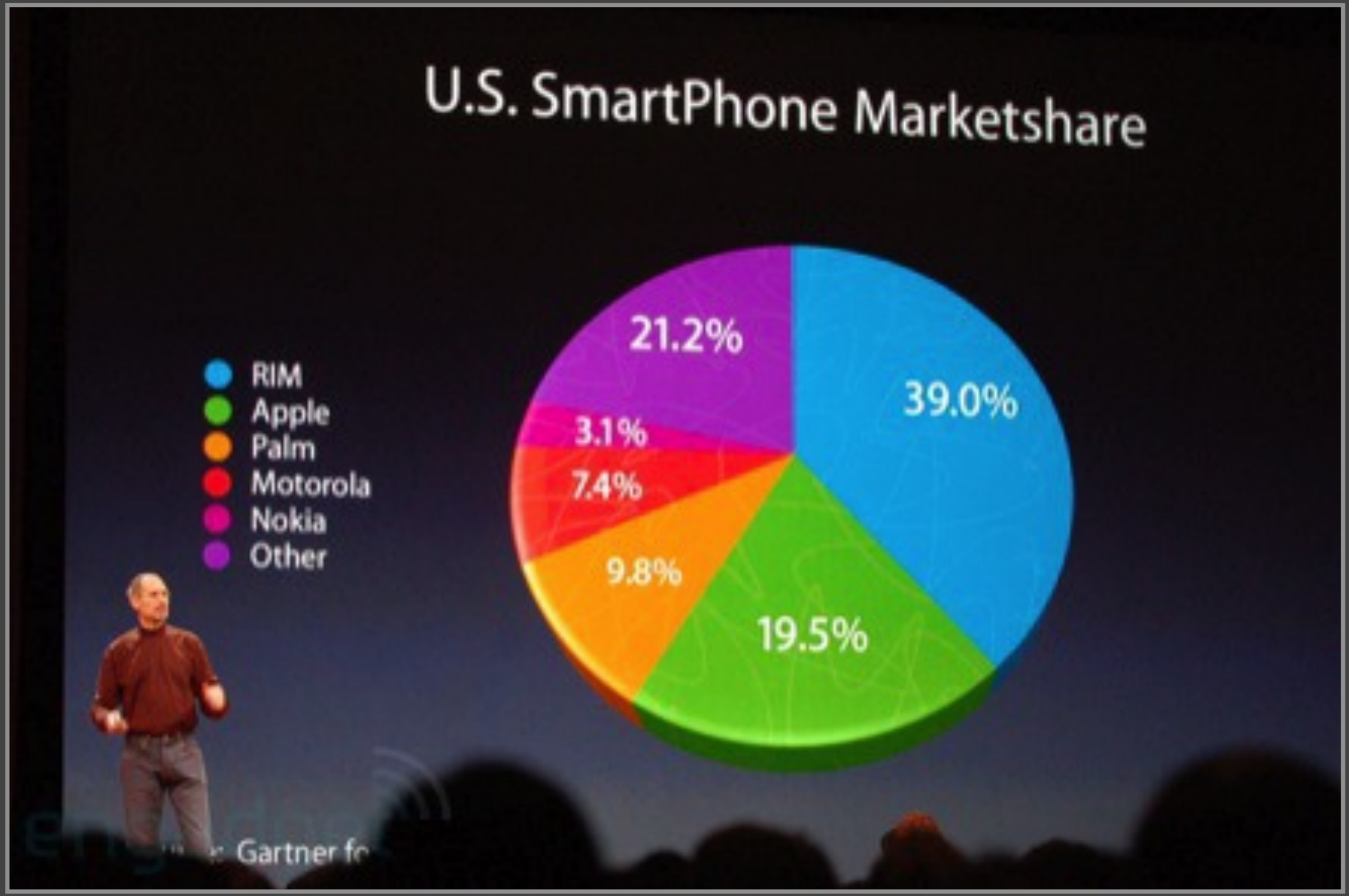
The Lie Factor

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

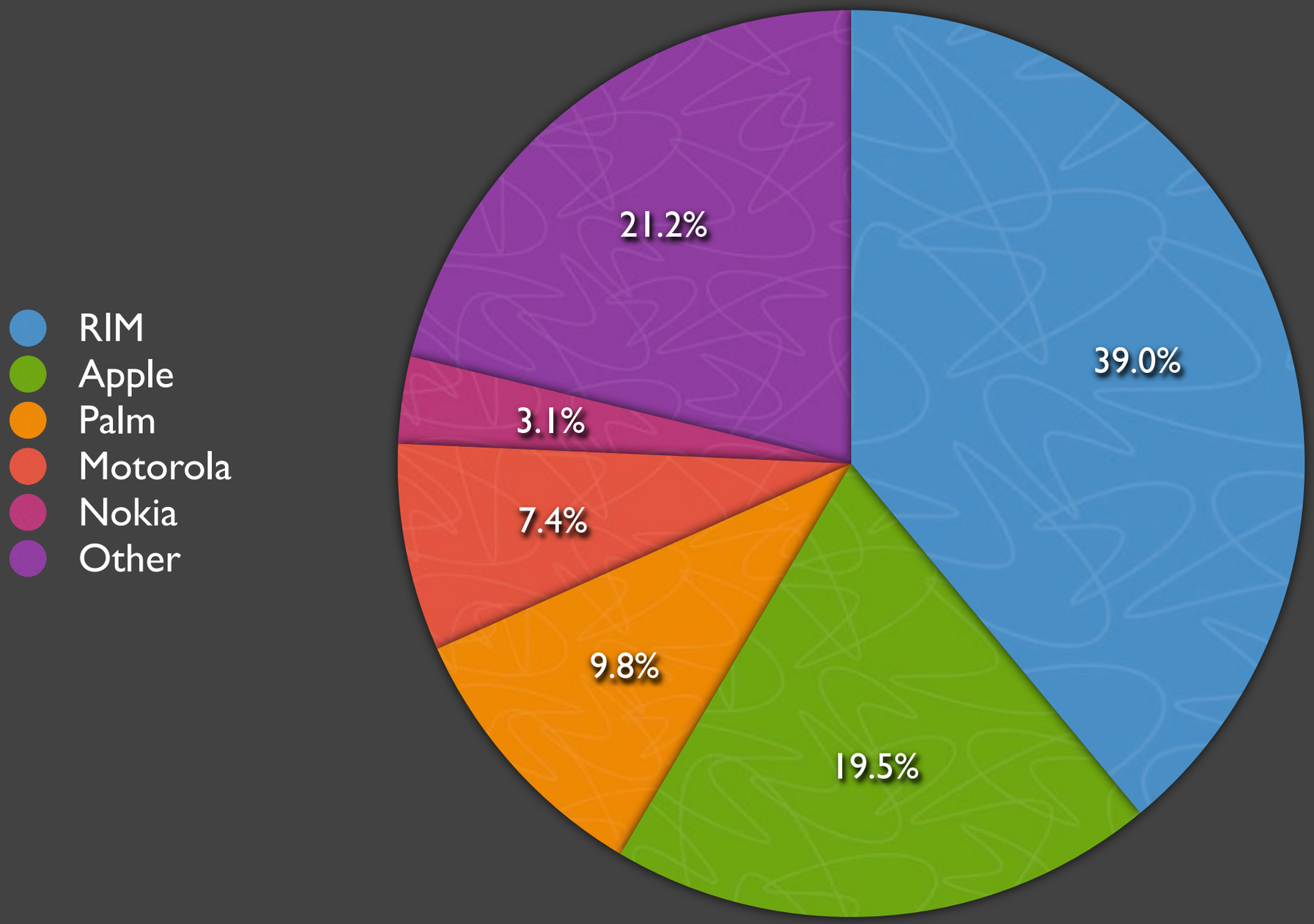




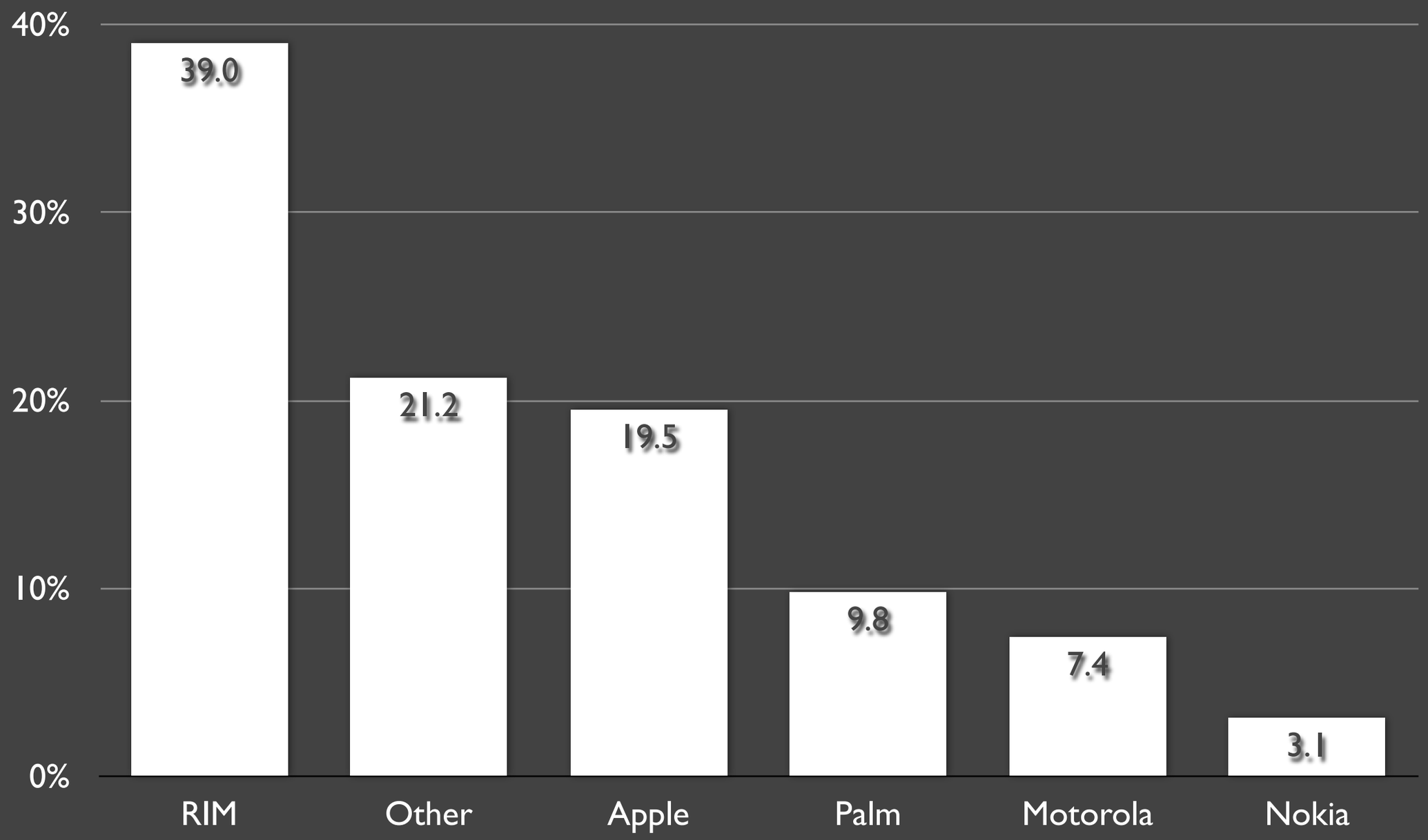
Tufte, VDQI


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

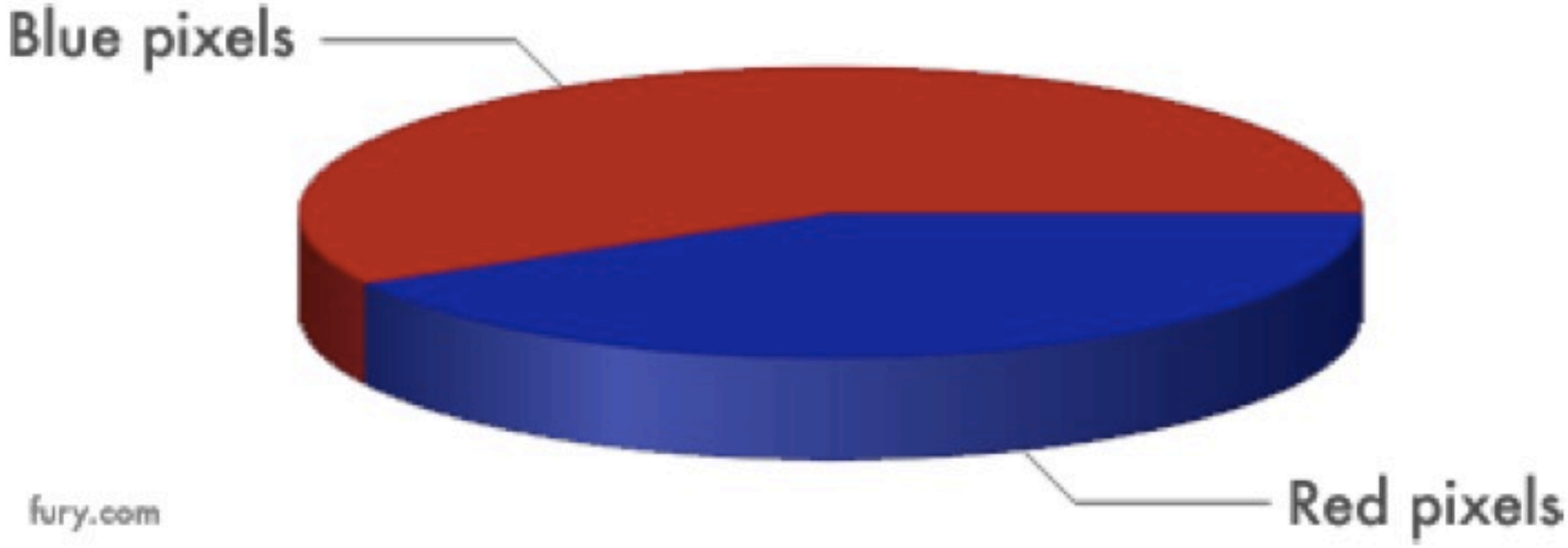
U.S. SmartPhone Marketshare



U.S. SmartPhone Marketshare



Why 3D Pie Charts are Bad

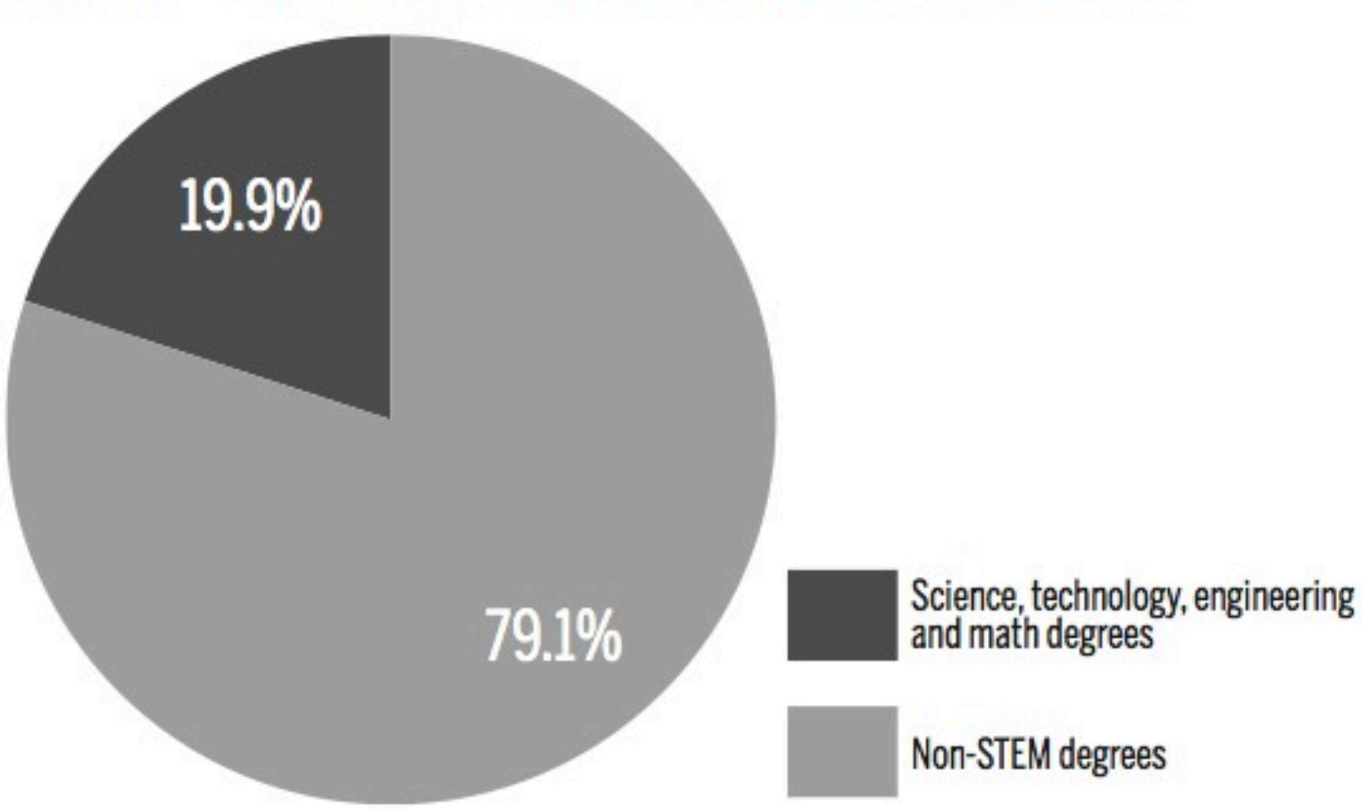


Kevin Fox

Yale Daily News

THE OLDEST COLLEGE DAILY
ESTABLISHED 1878

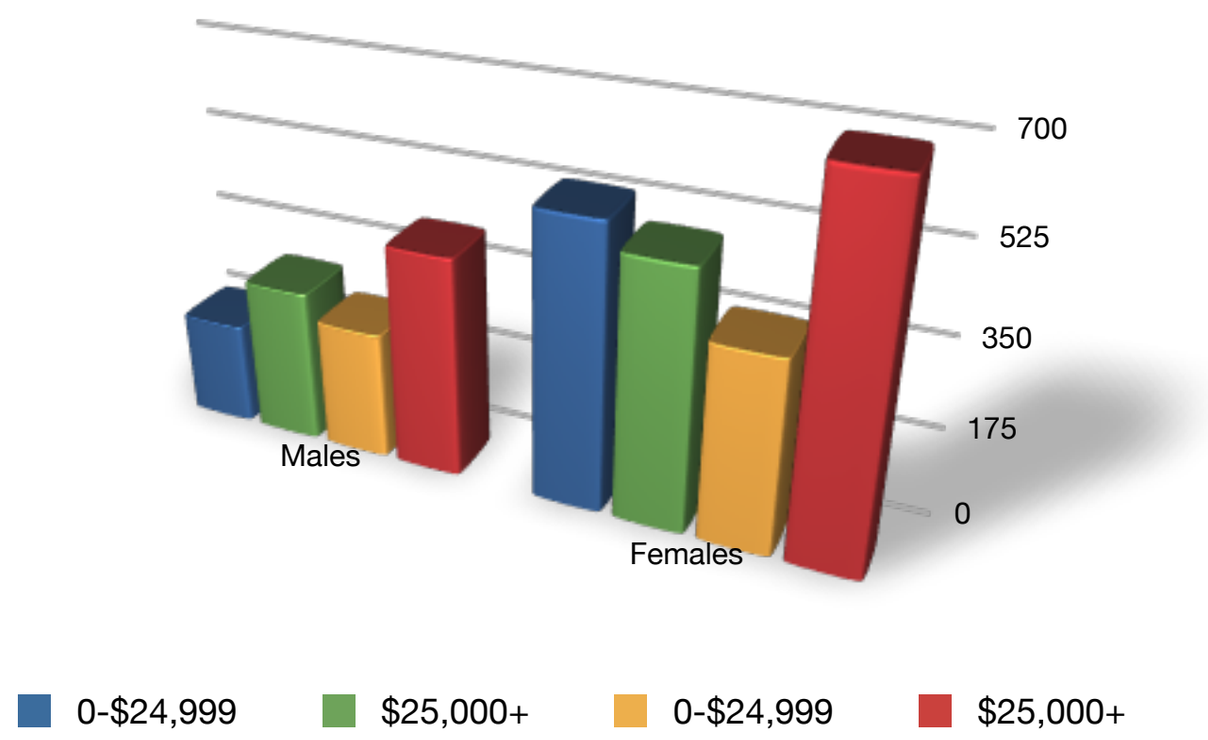
CHART YALE GRADUATES' MAJORS, CLASS OF 2011



September 13, 2011

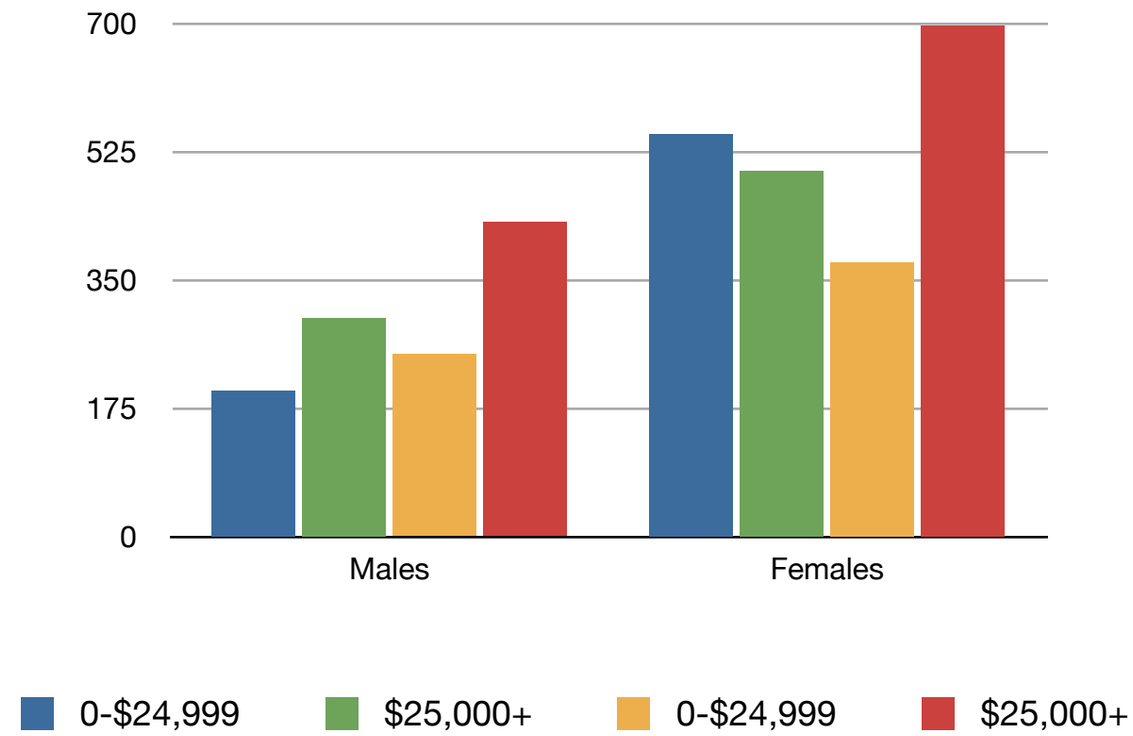
Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

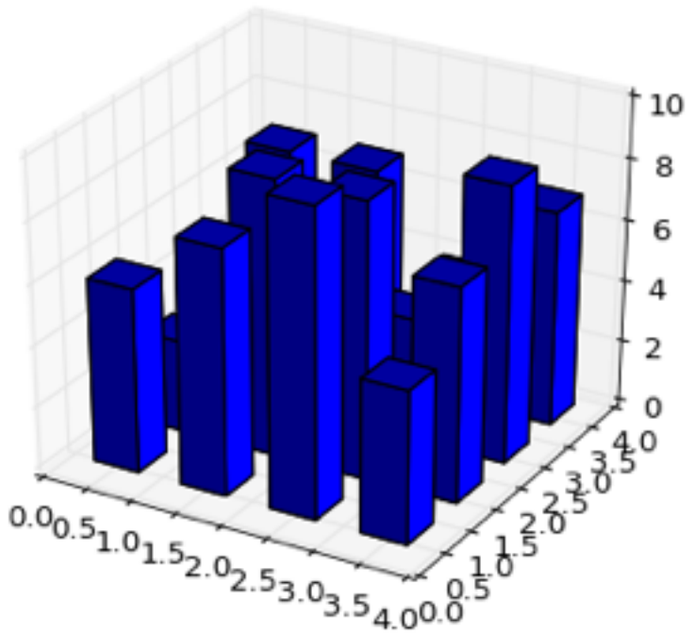


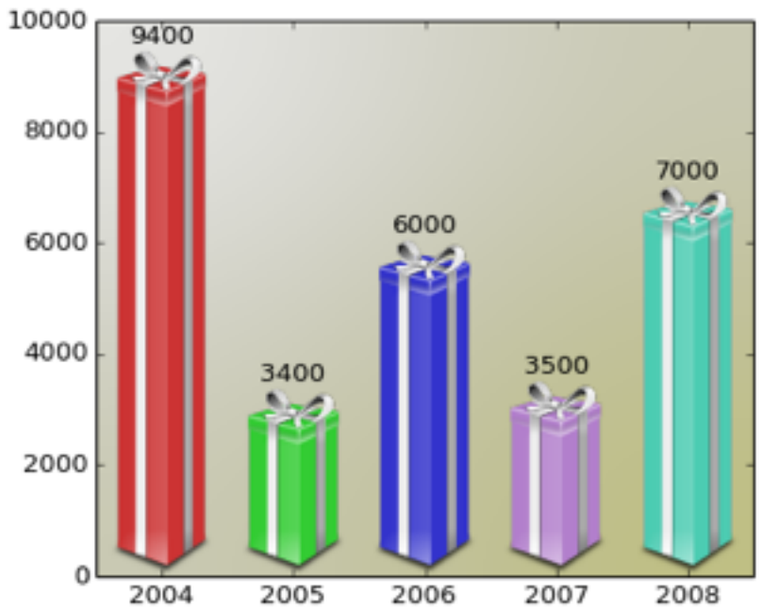
Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

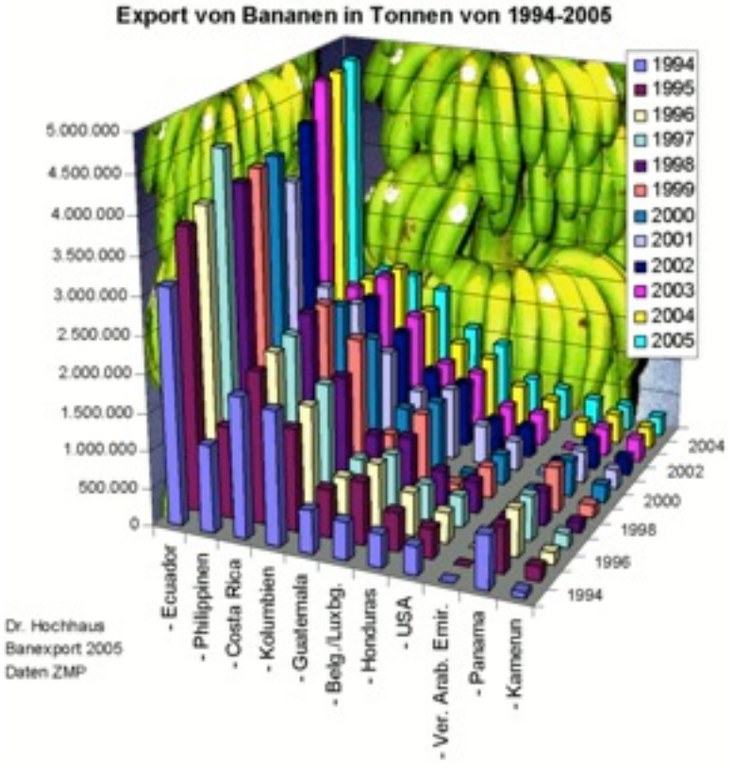


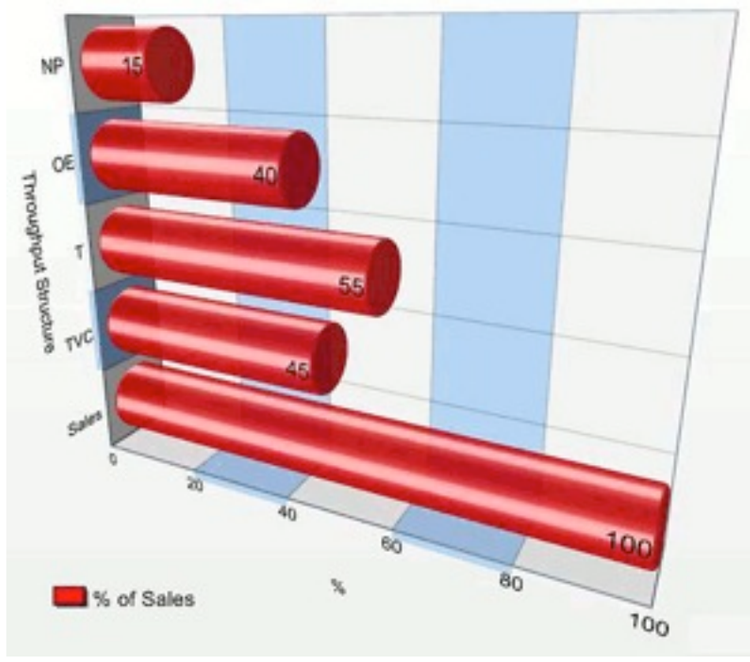
Don’t





matplotlib gallery

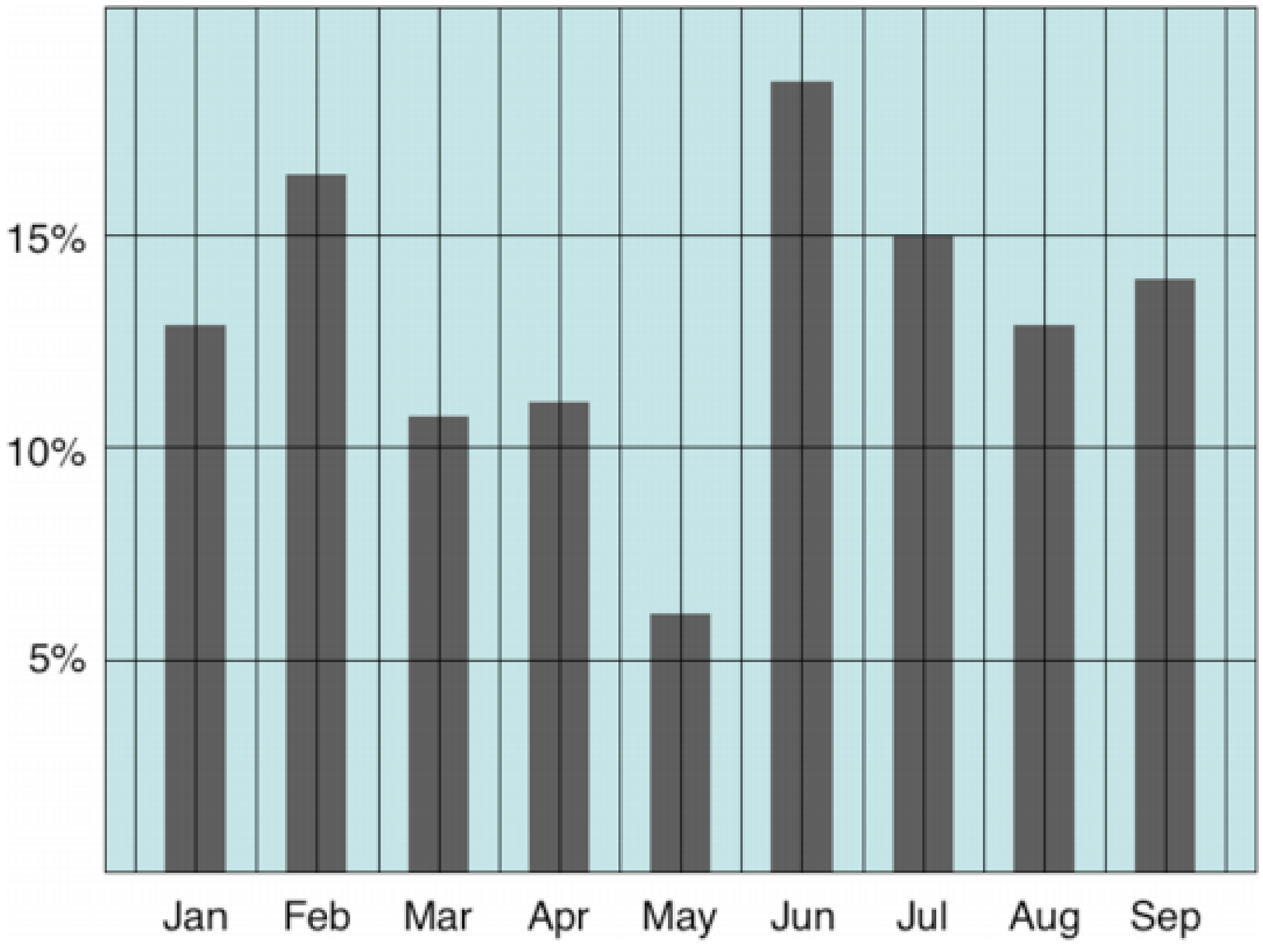




Excel Charts Blog

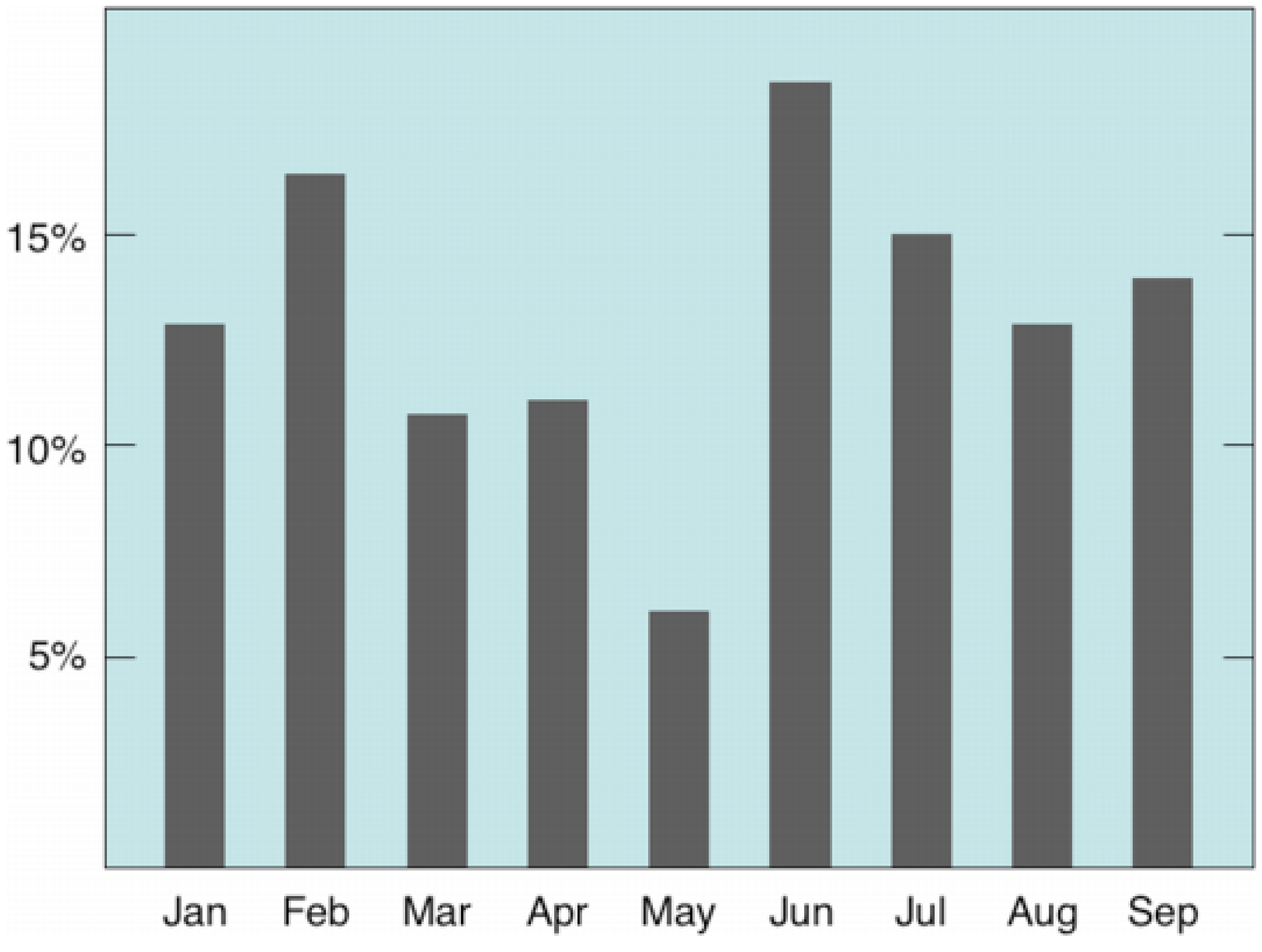
Avoid Chartjunk

Extraneous visual elements that distract from the message



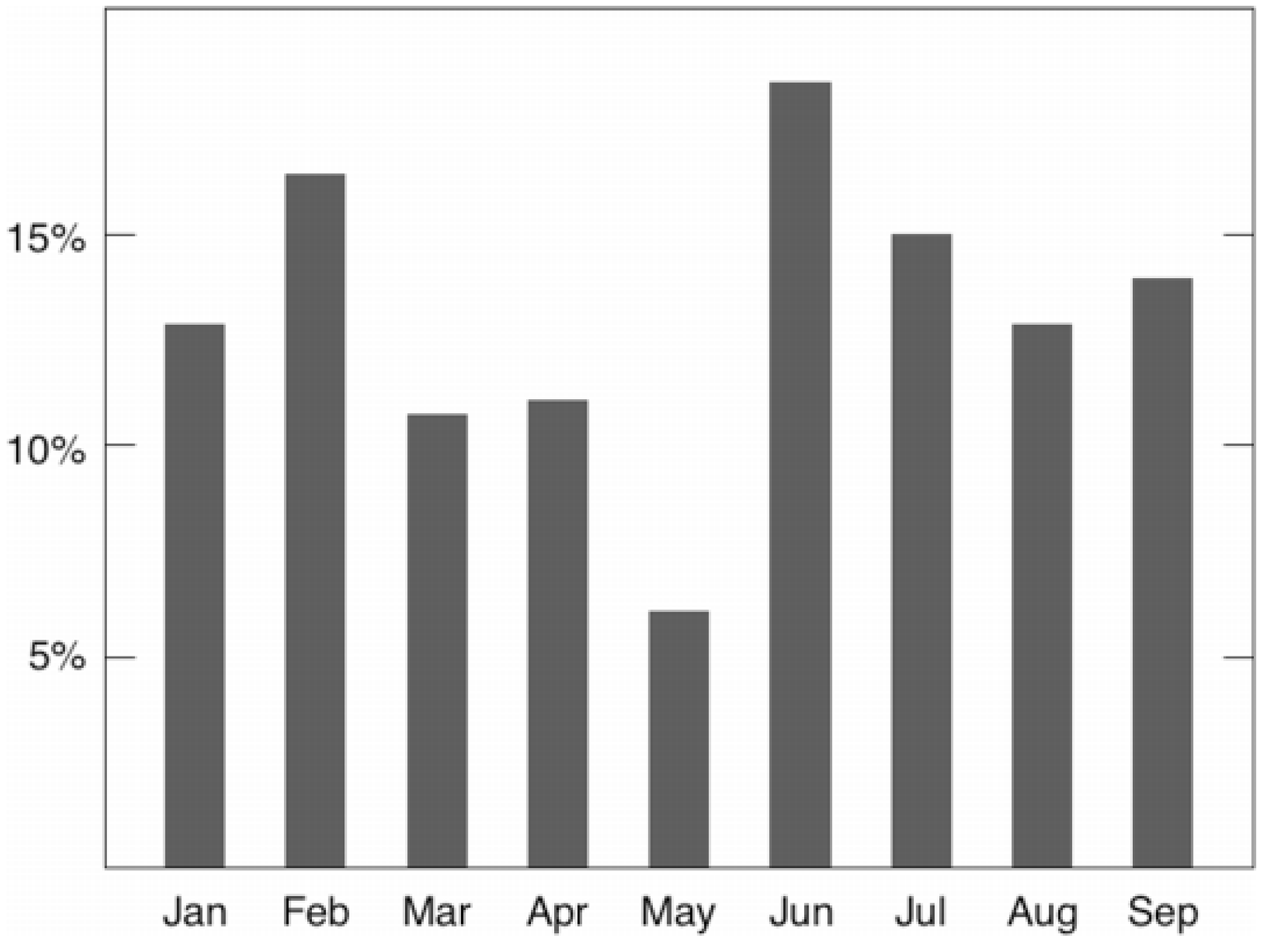
ongoing, Tim Brey

Avoid Chartjunk



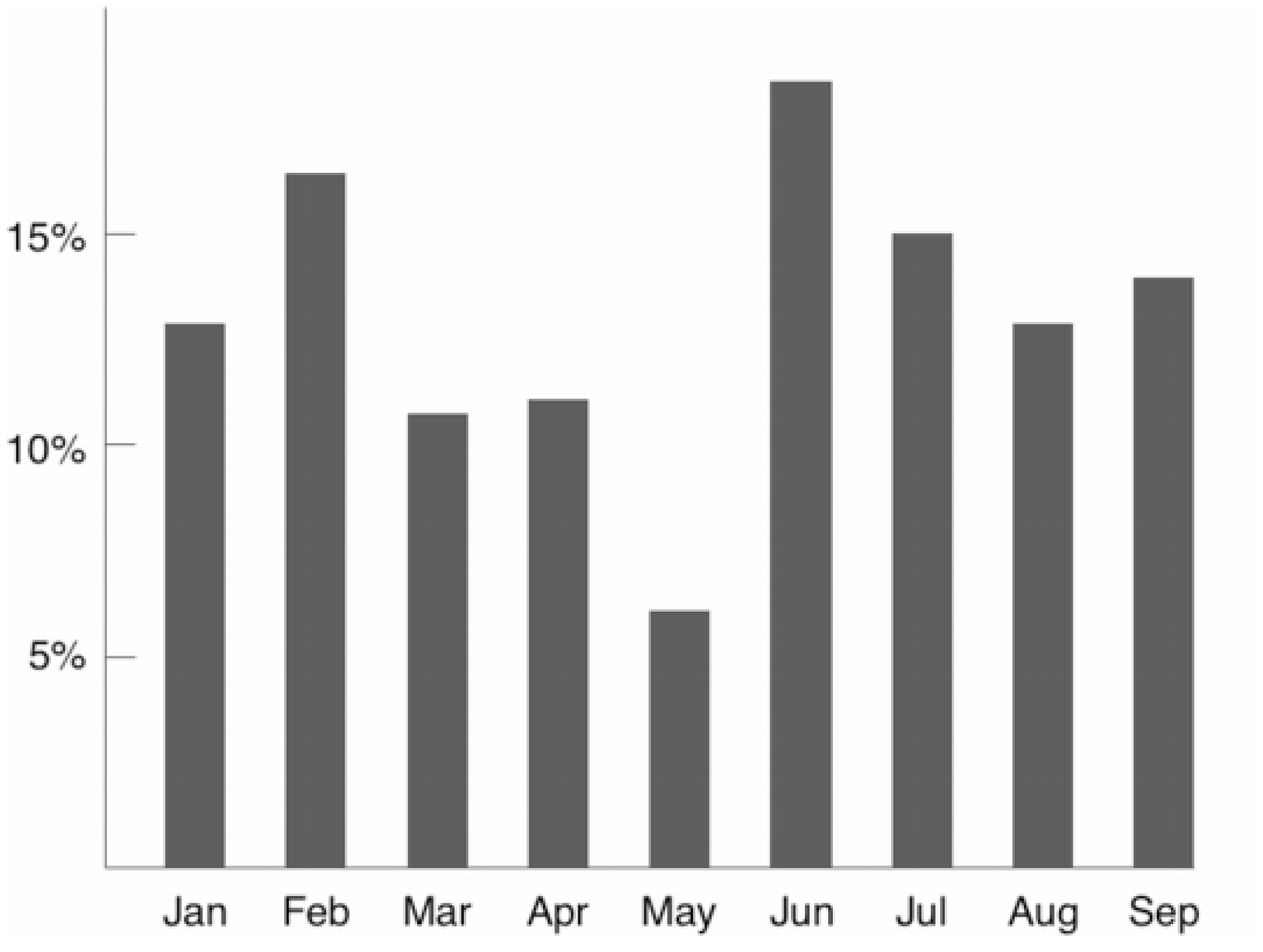
ongoing, Tim Brey

Avoid Chartjunk



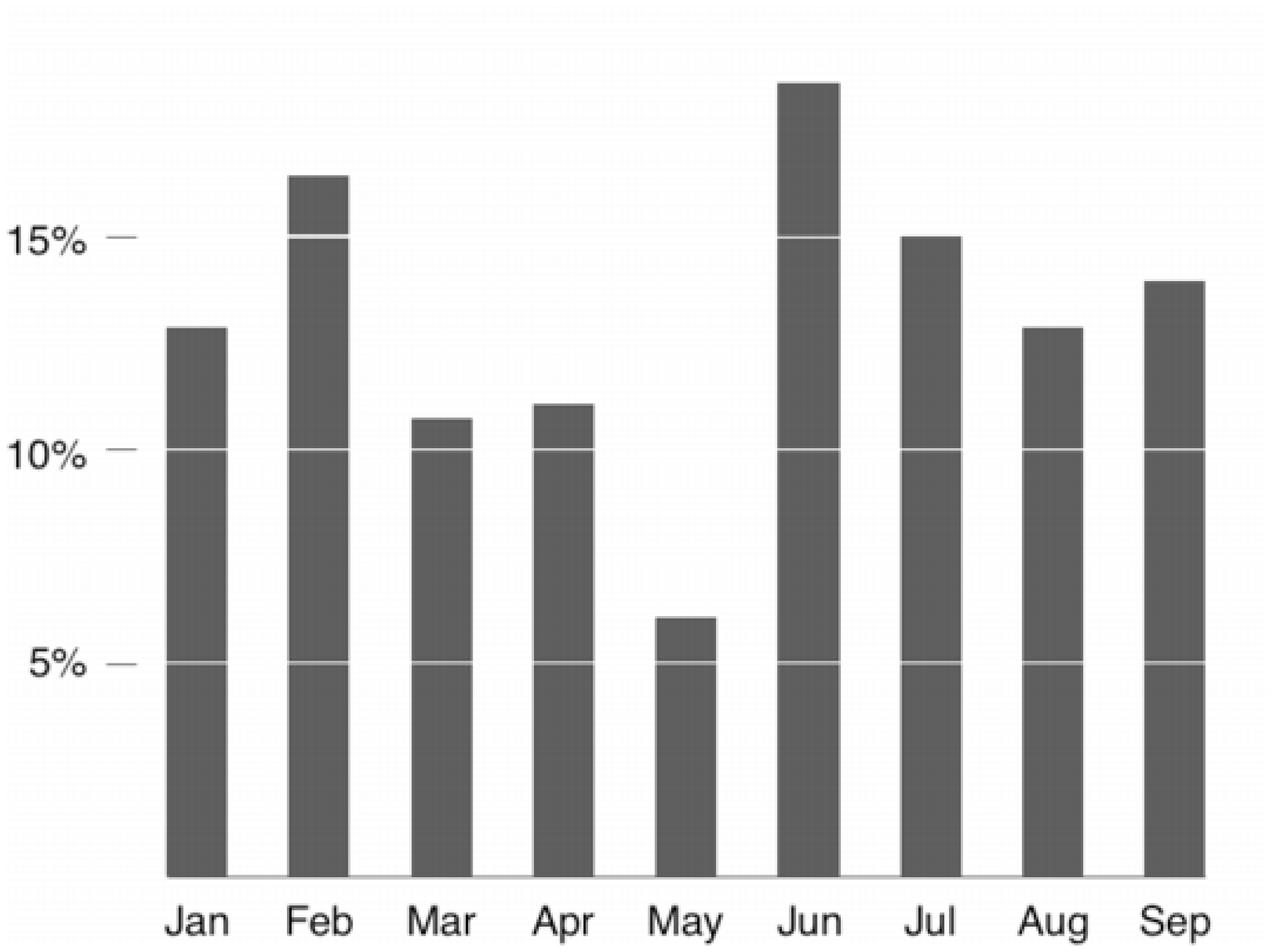
ongoing, Tim Brey

Avoid Chartjunk



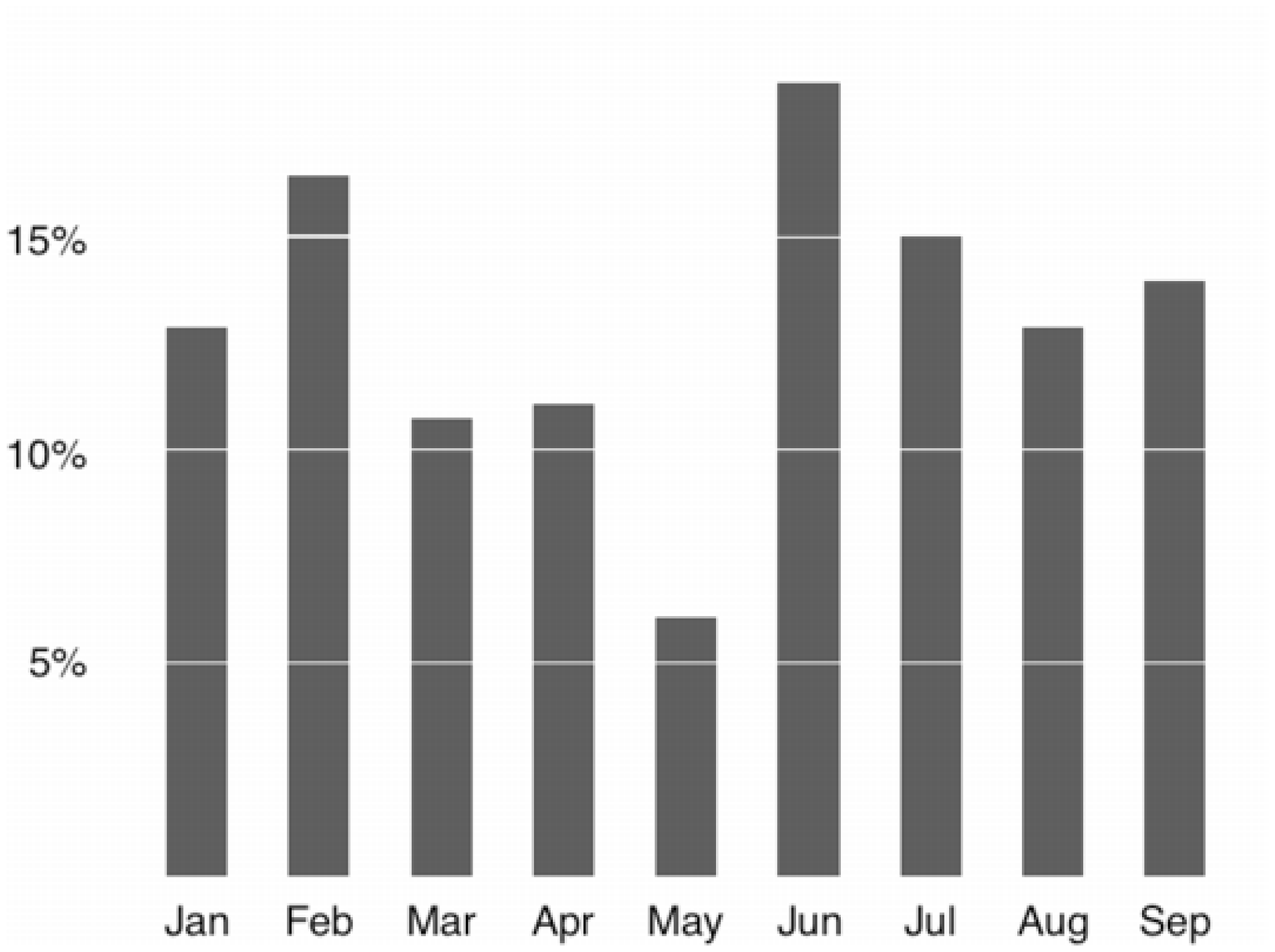
ongoing, Tim Brey

Avoid Chartjunk



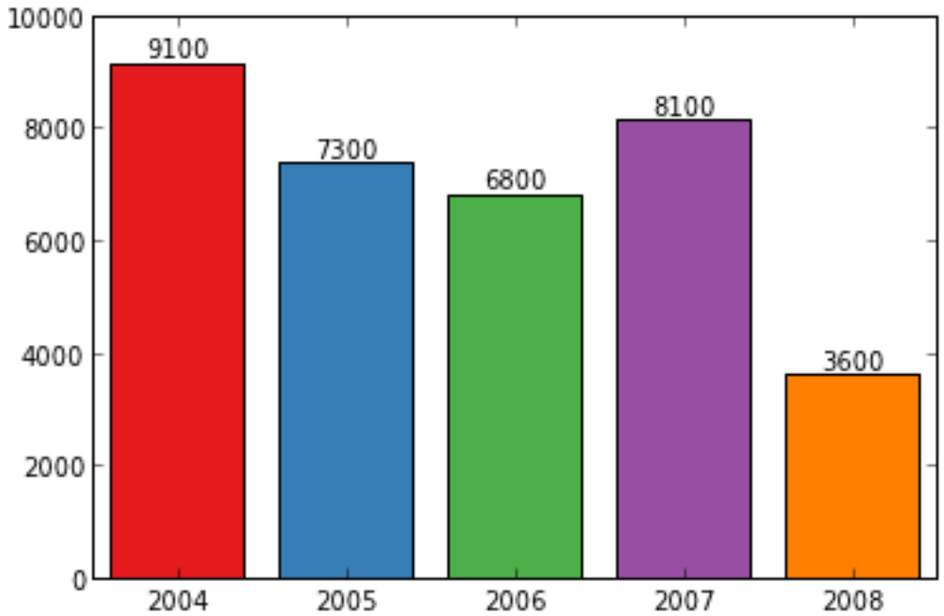
ongoing, Tim Brey

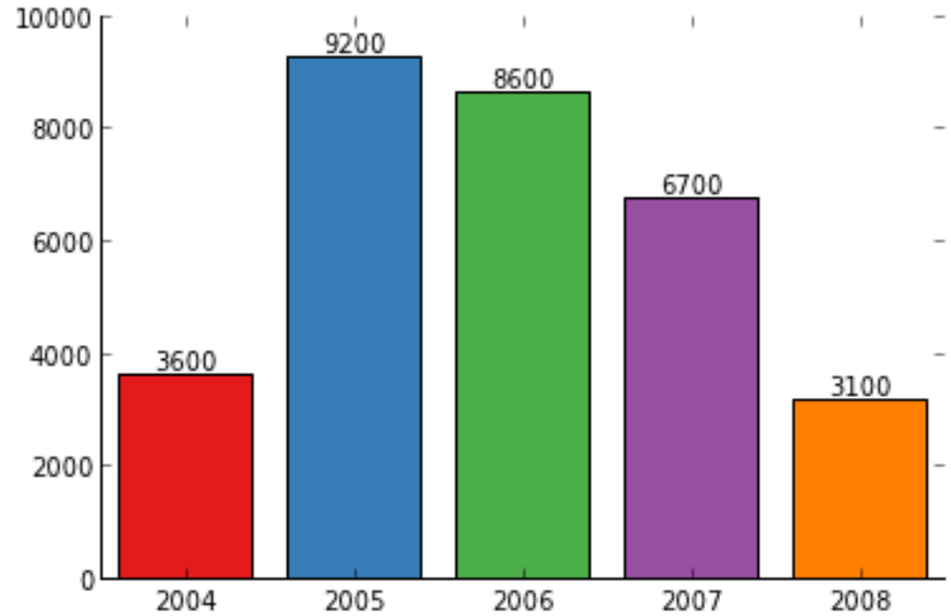
Avoid Chartjunk

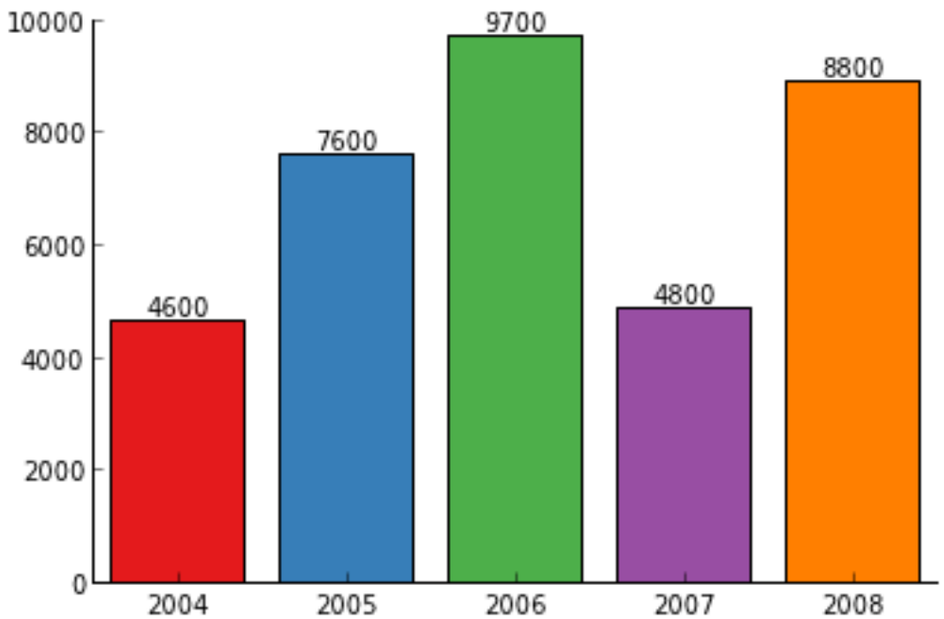


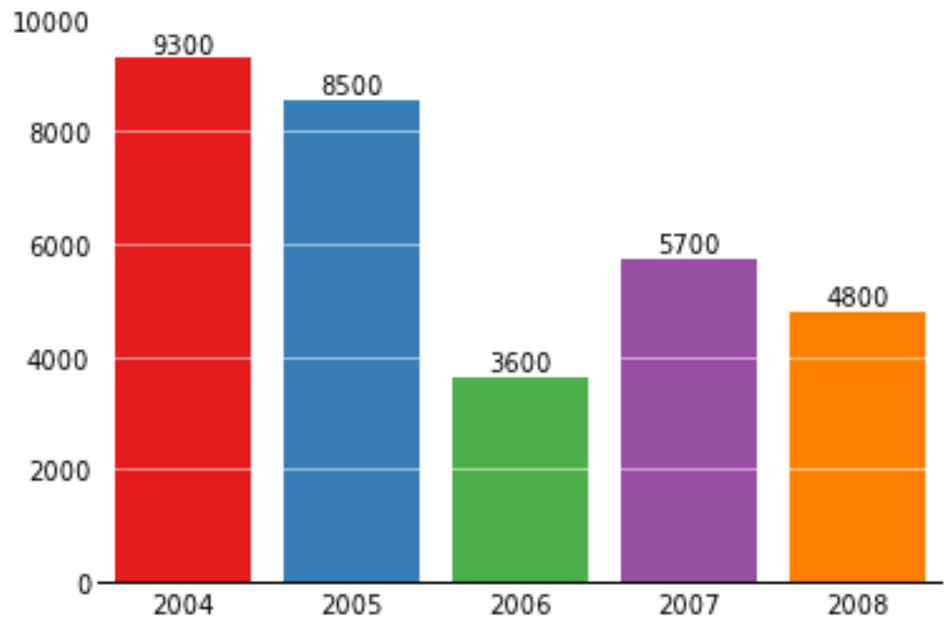
ongoing, Tim Brey

Matplotlib Example









http://nbviewer.ipython.org/5357268

Matplotlib Notebooks

- Links under "Readings" on the course web site
- Demonstrate how to make nice graphs in MPL by Chris Beaumont

A Gallery of Statistical Graphs in Matplotlib

Inspiration taken from http://nbviewer.ipython.org/5357268

```
In [78]: #brewer2mpl makes it easier to use color tables from colorbrewer2.org in matplotlib
         !pip install brewer2mpl

Requirement already satisfied (use --upgrade to upgrade): brewer2mpl in /Users/beaumont/anaconda/lib/python2.7/site-packages
Cleaning up...

In [79]: %matplotlib inline
         from urllib import urlopen

         import brewer2mpl
         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import pandas as pd

In [80]: # Set up some better defaults for matplotlib
         from matplotlib import rcParams

         #colorbrewer2 Dark2 qualitative color table
         dark2_colors = brewer2mpl.get_map('Dark2', 'Qualitative', 7).mpl_colors

         rcParams['figure.figsize'] = (10, 6)
         rcParams['figure.dpi'] = 150
         rcParams['axes.color_cycle'] = dark2_colors
         rcParams['lines.linewidth'] = 2
         rcParams['axes.facecolor'] = 'white'
         rcParams['font.size'] = 14
         rcParams['patch.edgecolor'] = 'white'
         rcParams['patch.facecolor'] = dark2_colors[0]
         rcParams['font.family'] = 'StixGeneral'


         def remove_border(axes=None, top=False, right=False, left=True, bottom=True):
             """
             Minimize chartjunk by stripping out unnecesasry plot borders and axis ticks

             The top/right/left/bottom keywords toggle whether the corresponding plot border is draw
             """
             ax = axes or plt.gca()
             ax.spines['top'].set_visible(top)
             ax.spines['right'].set_visible(right)
             ax.spines['left'].set_visible(left)
             ax.spines['bottom'].set_visible(bottom)

             #turn off all ticks
```

Home FAQ IPython Bookmarklet Download Notebook

A Gallery of Statistical Graphs in Matplotlib (Matplotlib Defaults)

```
In [24]: %matplotlib inline
         from urllib import urlopen

         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import pandas as pd

         from matplotlib import rcParams
         rcParams['figure.figsize'] = (10, 6)
         rcParams['figure.dpi'] = 150
```

Example Data

```
In [25]: file = urlopen('https://raw.github.com/cpcloud/pydatasets/master/datasets/ggplot2/diamonds.csv')
         diamonds = pd.read_csv(file)

         file = urlopen('http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/resources/R/titanic.csv')
```

IPython creator

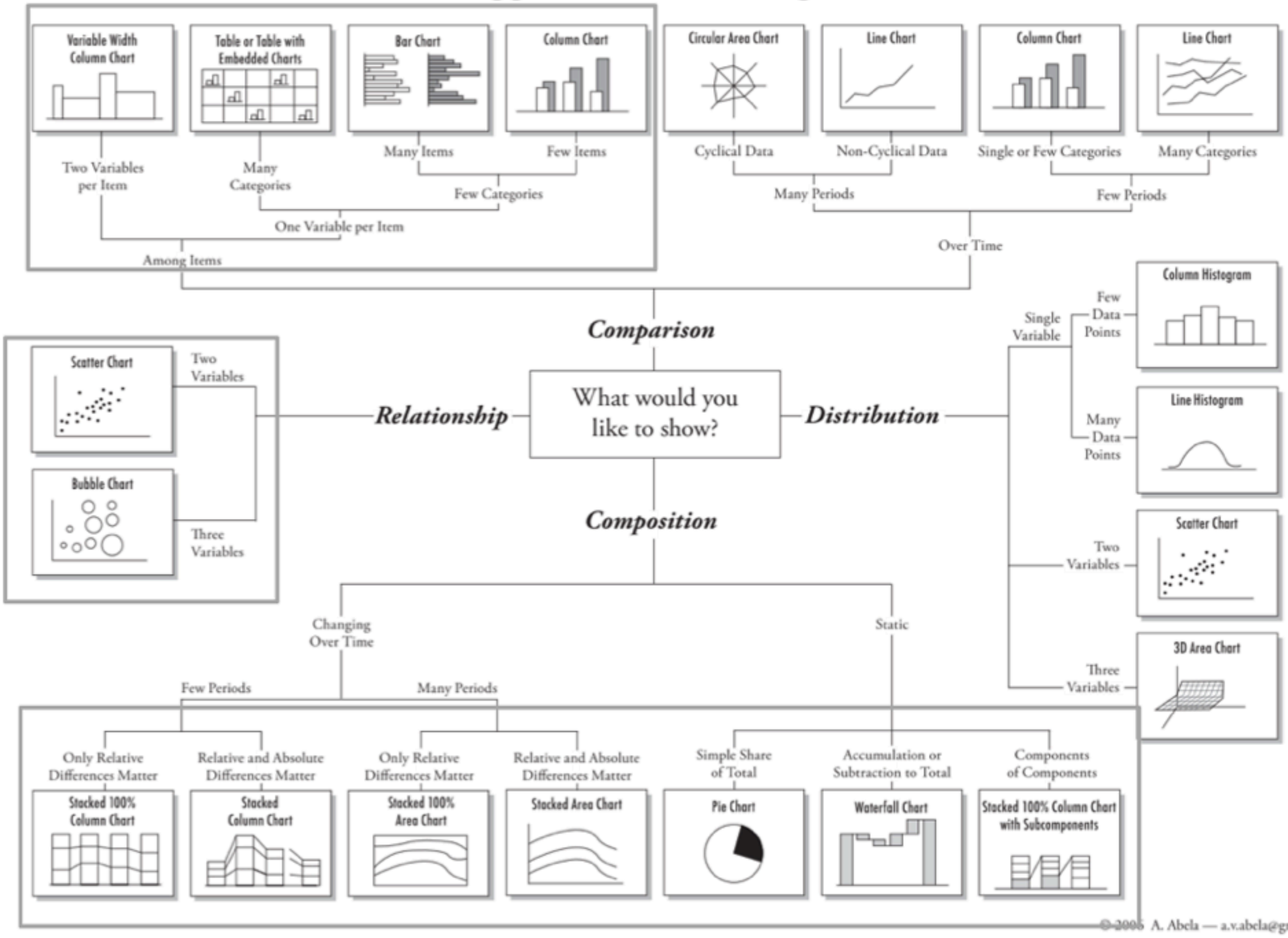
ggplot2 (R) creator

Tufte's Design Principles

- Clear, detailed, and thorough labeling and appropriate scales
- Size of the graphic effect should be directly proportional to the numerical quantities ("lie factor")
- Maximize data-ink ratio
- Avoid chart junk

Graph Types

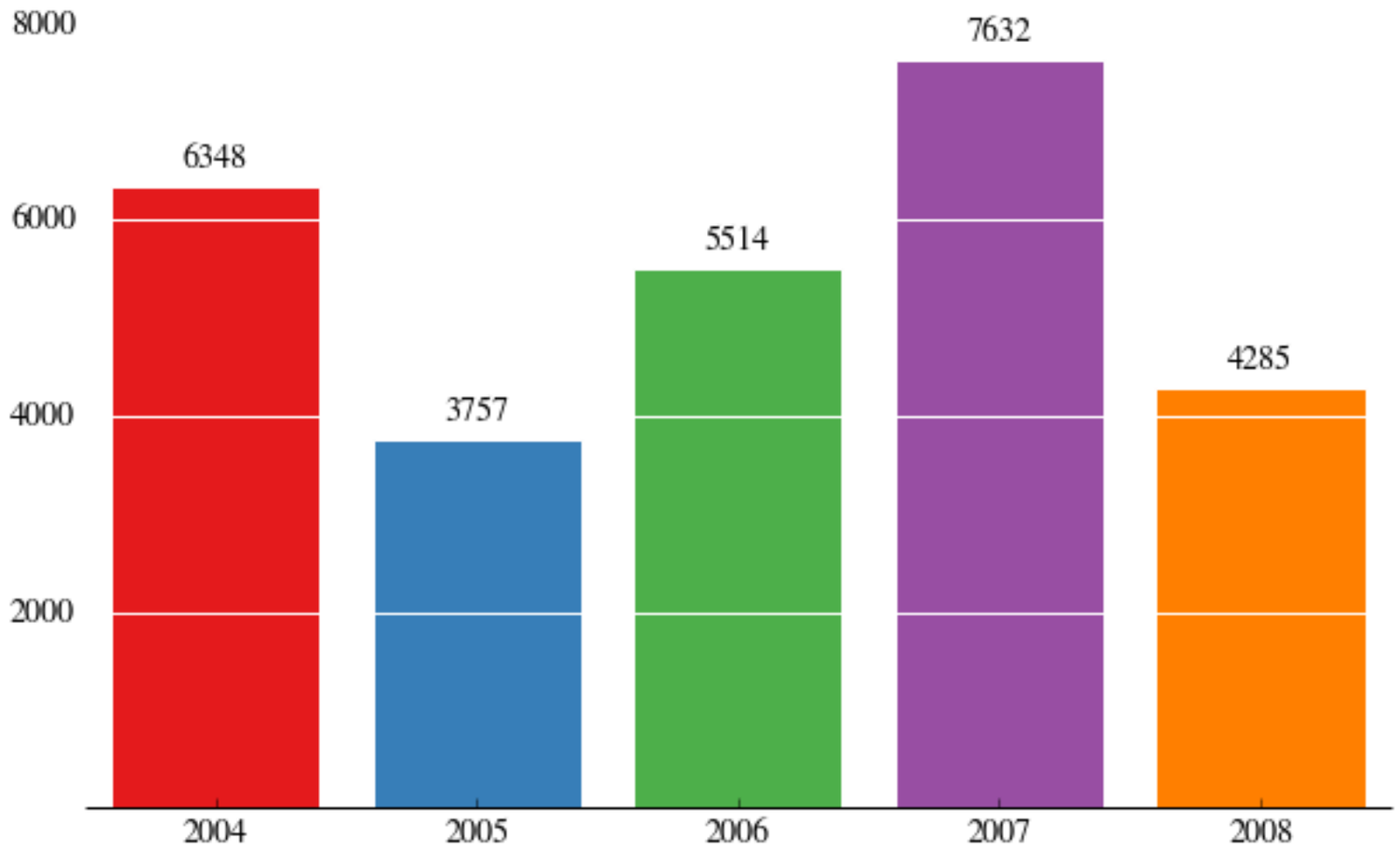
Chart Suggestions—A Thought-Starter



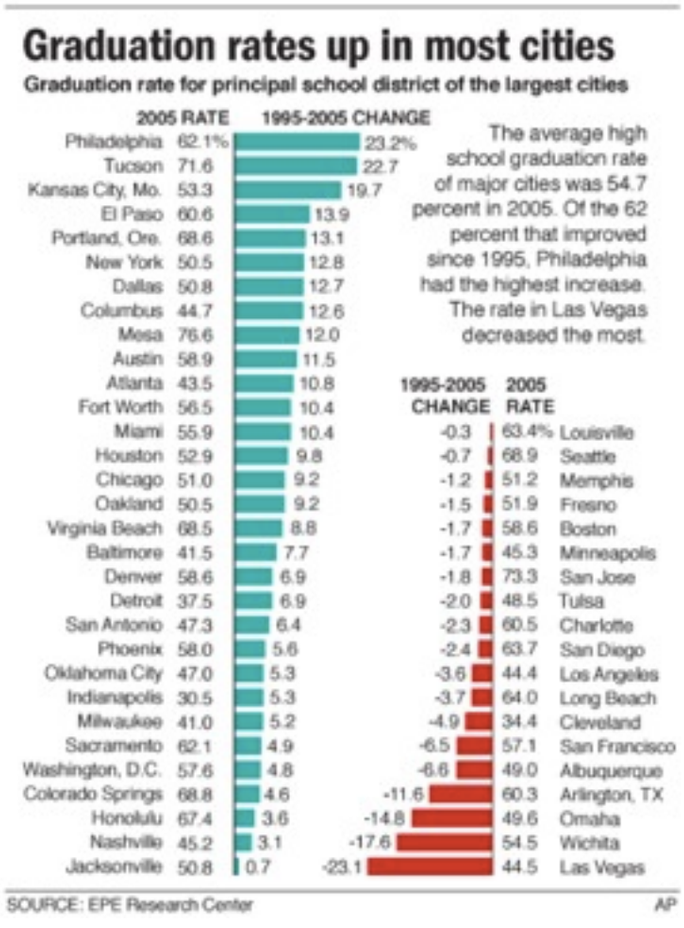
http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

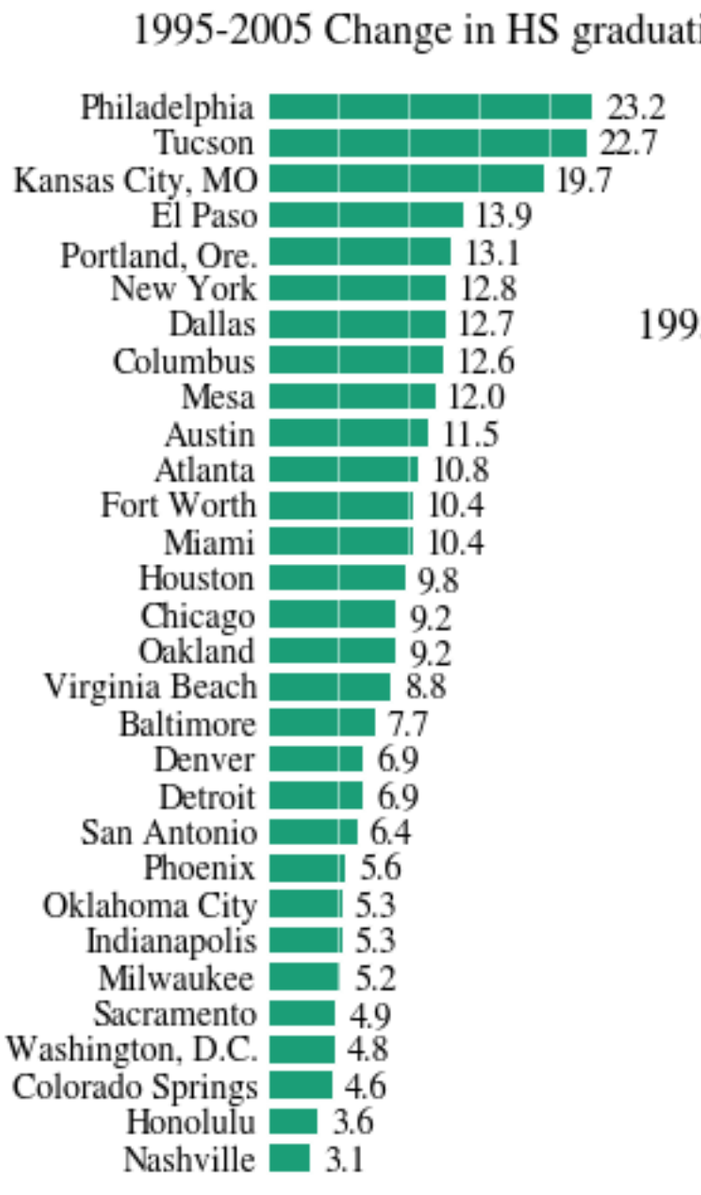
Comparisons

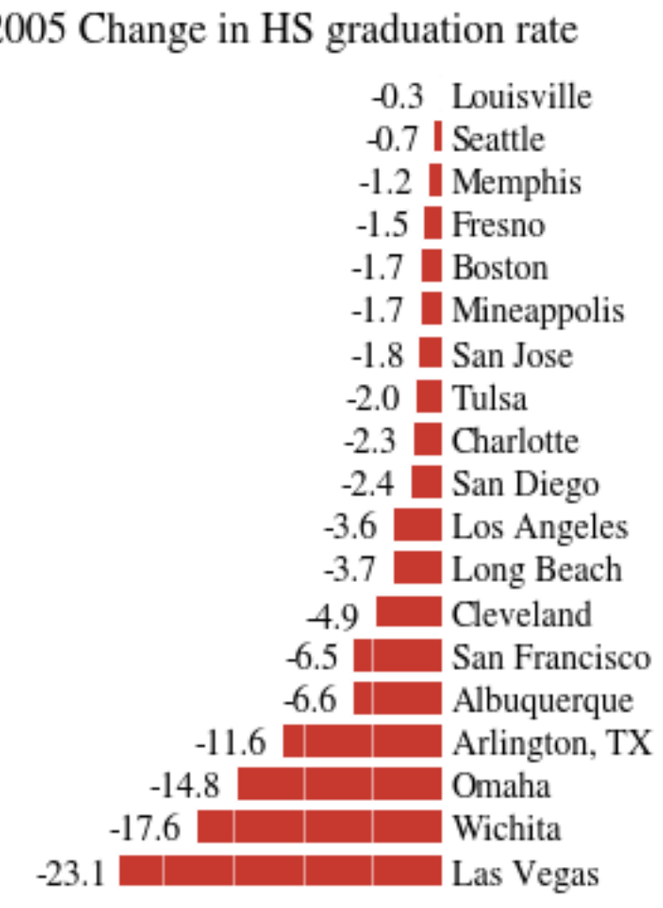
Bar Chart


8000
6000
4000
2000
6348
3757
5514
7632
4285
2004
2005
2006
2007
2008

Direction



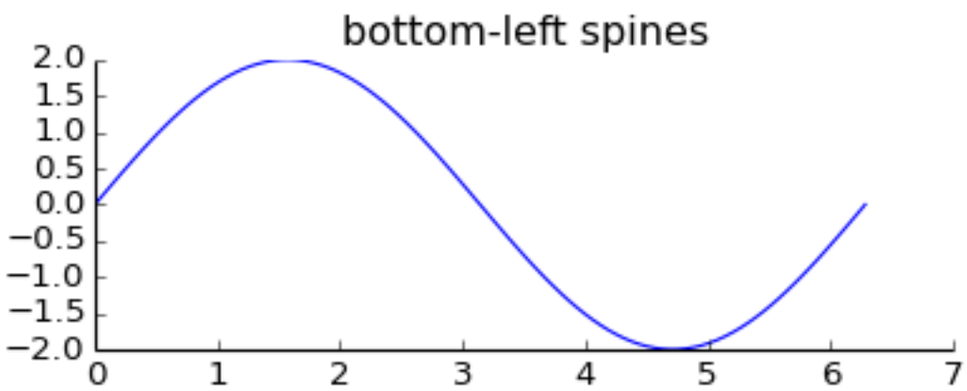


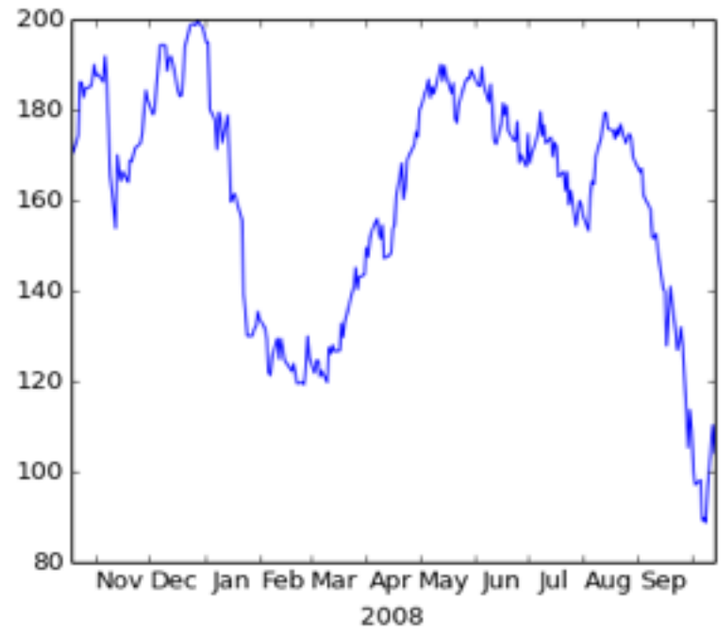


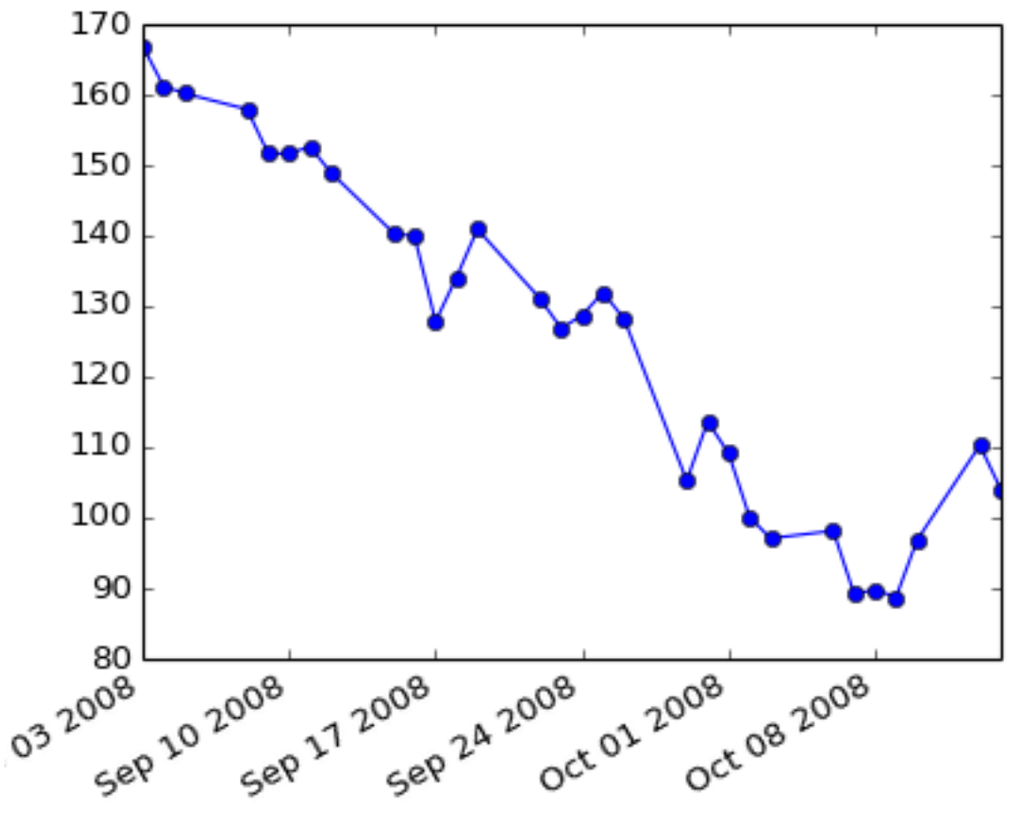
Nicolas Rapp

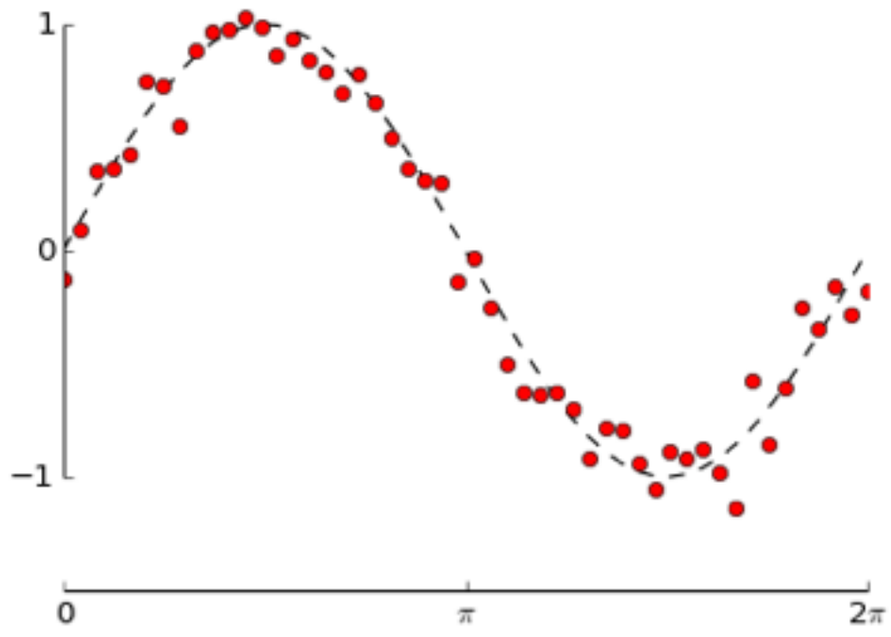
Trends Over Time

Line Charts





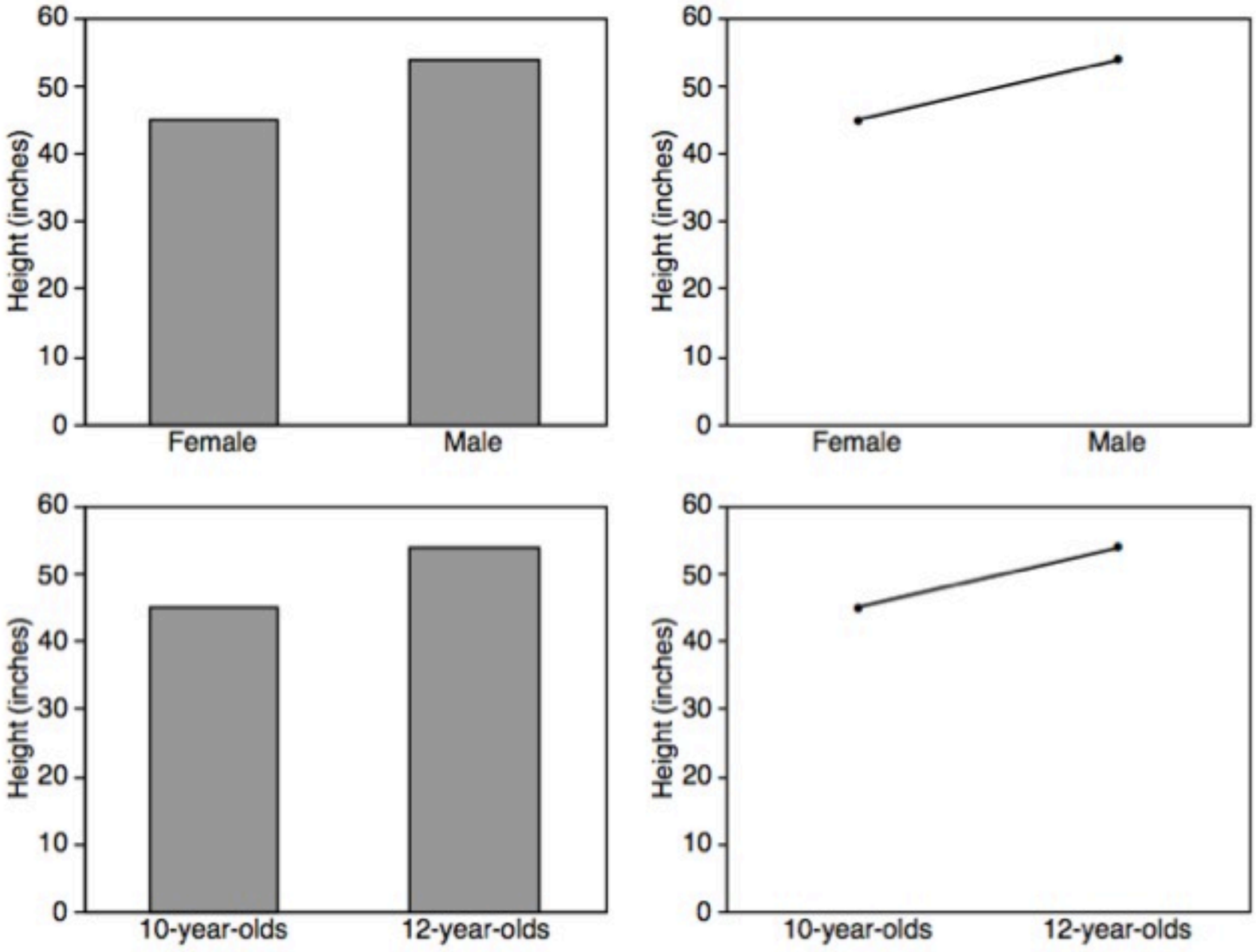




matplotlib gallery

Bars vs. Lines

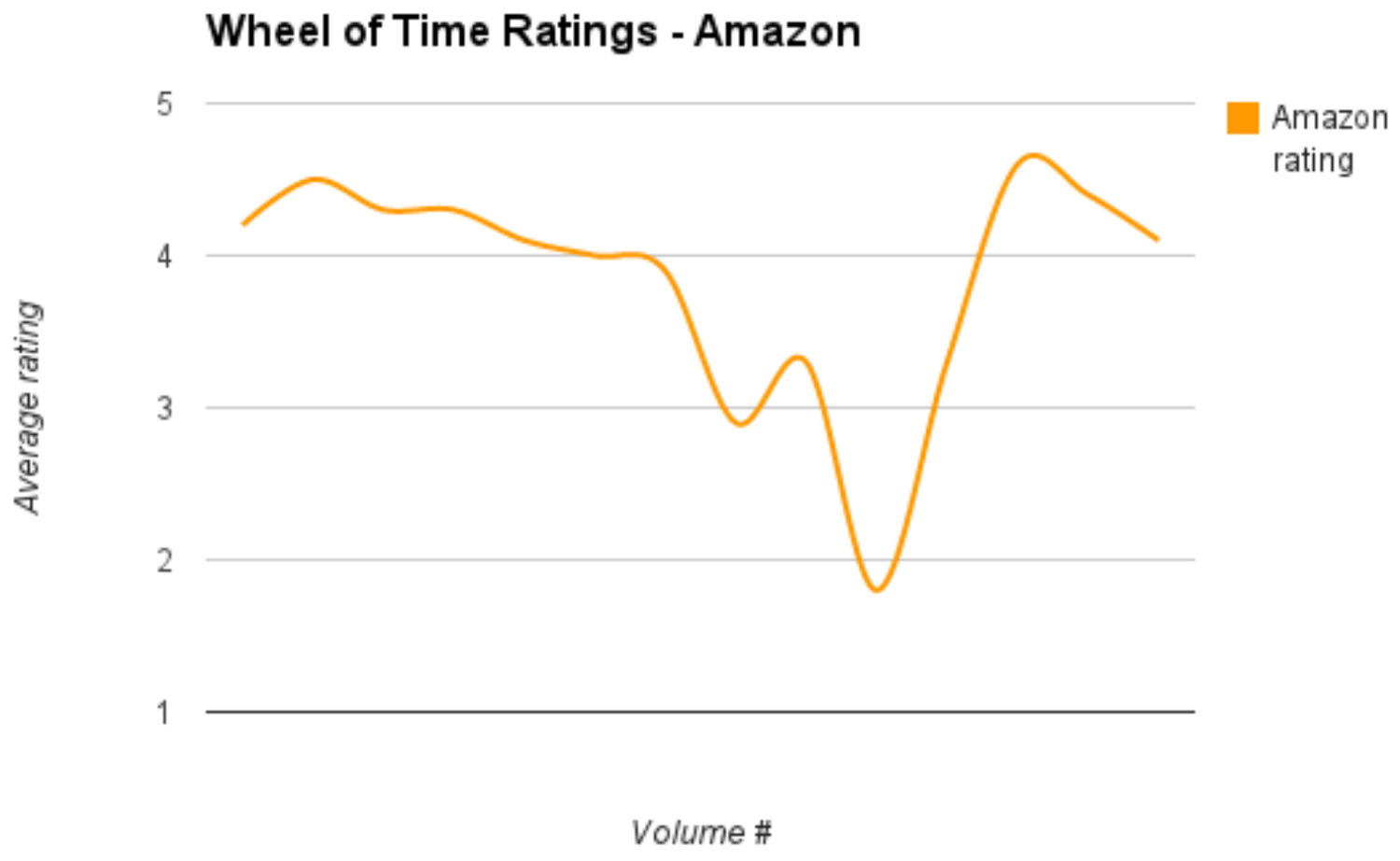
Lines imply connections - do not use for categorical data



Zacks 1999

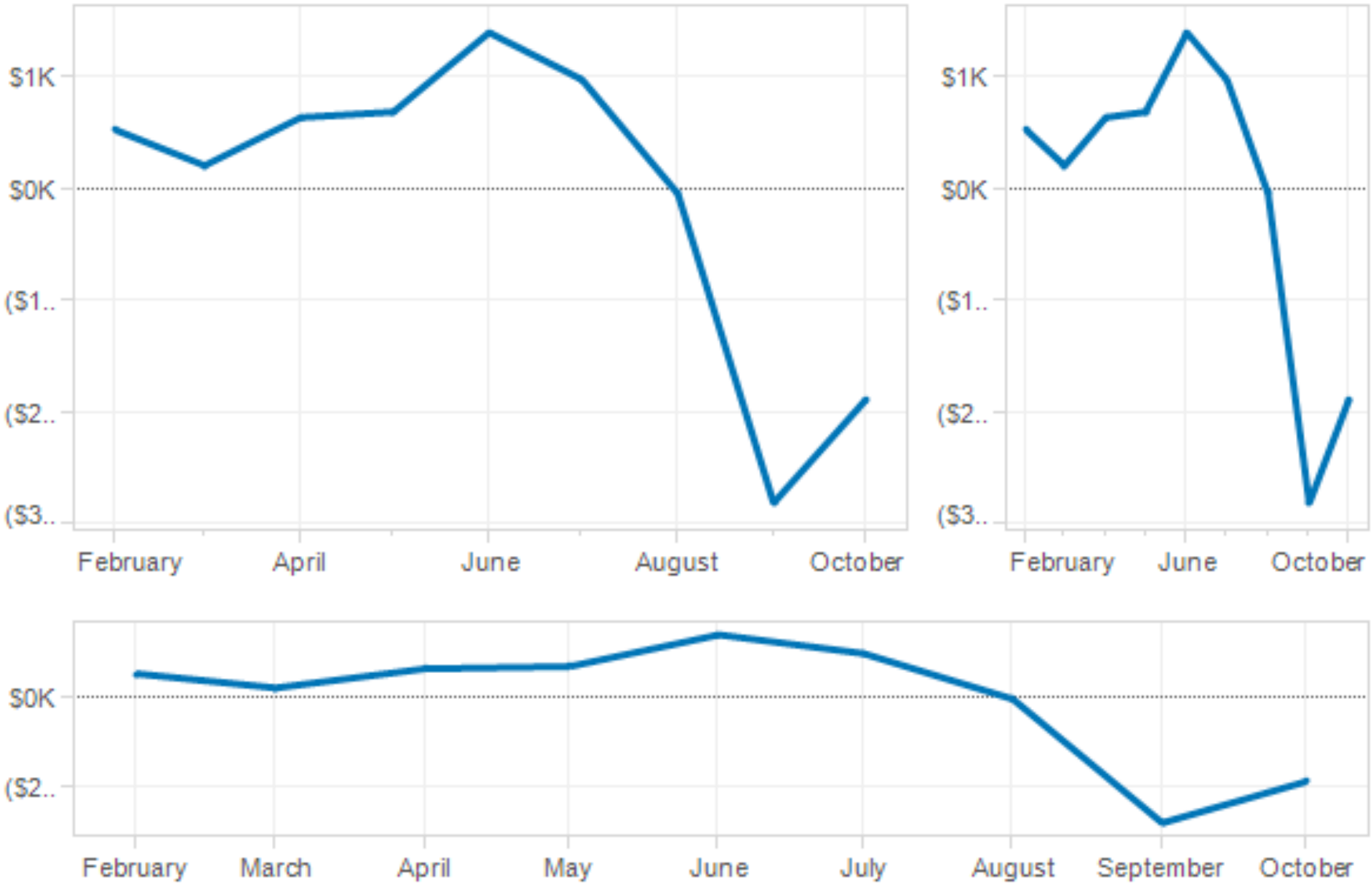
Don’t

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

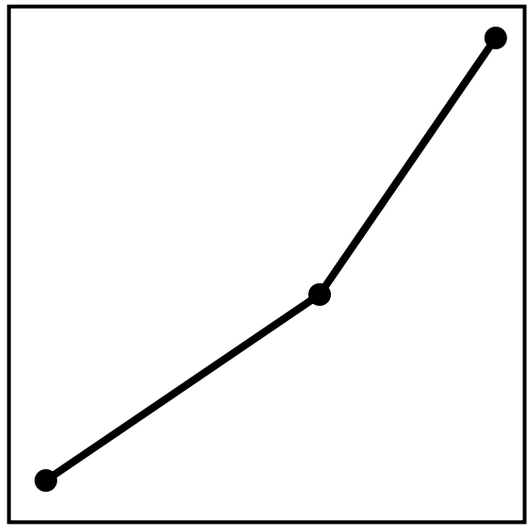


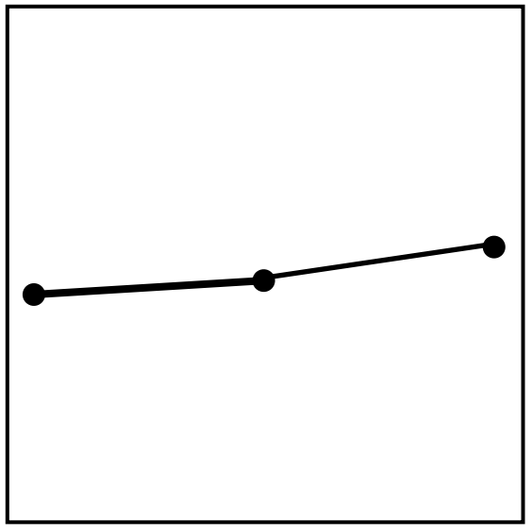
eagereyes.org

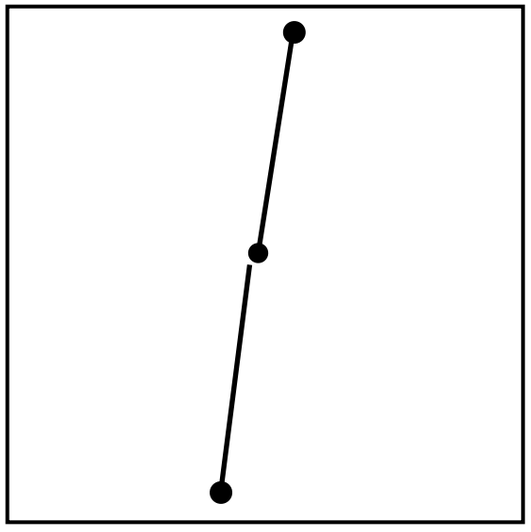
Banking to 45°

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

W. Cleveland

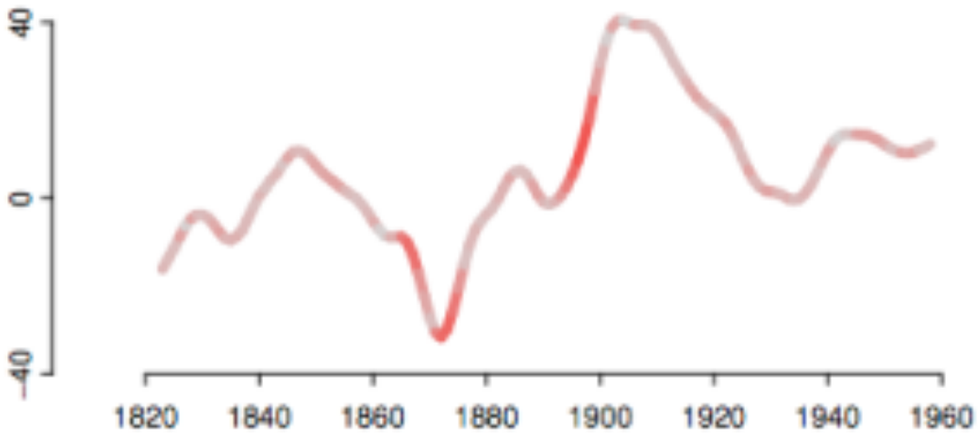




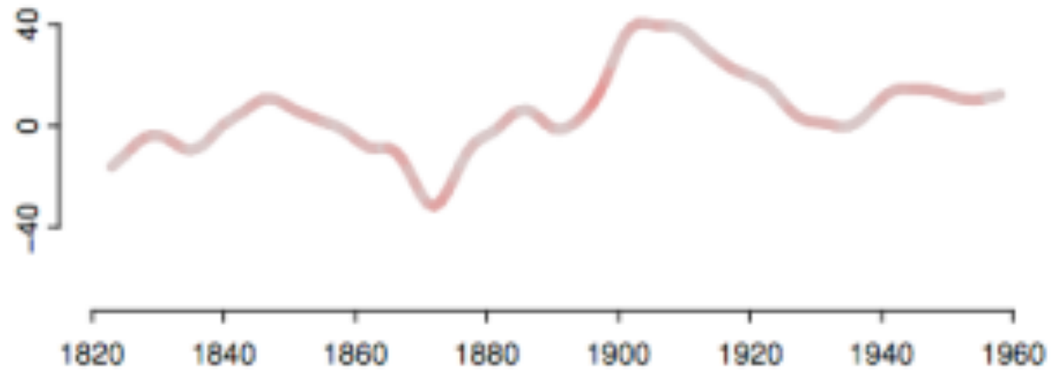


Slide based on J. Heer / M. Agrawala

Banking to 45°



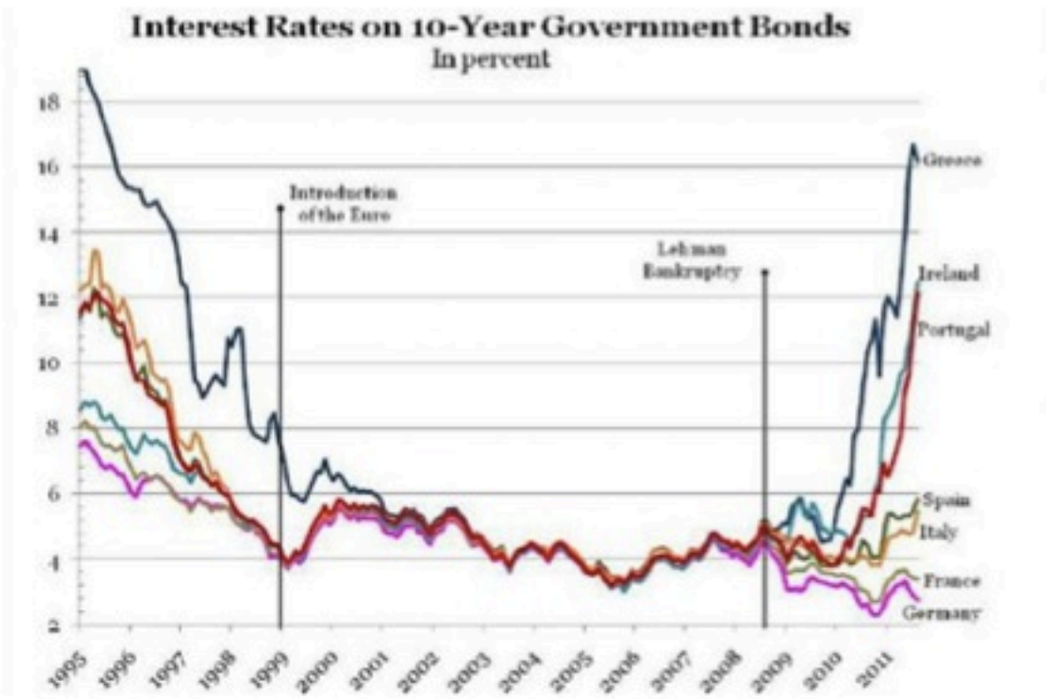
Error Prone

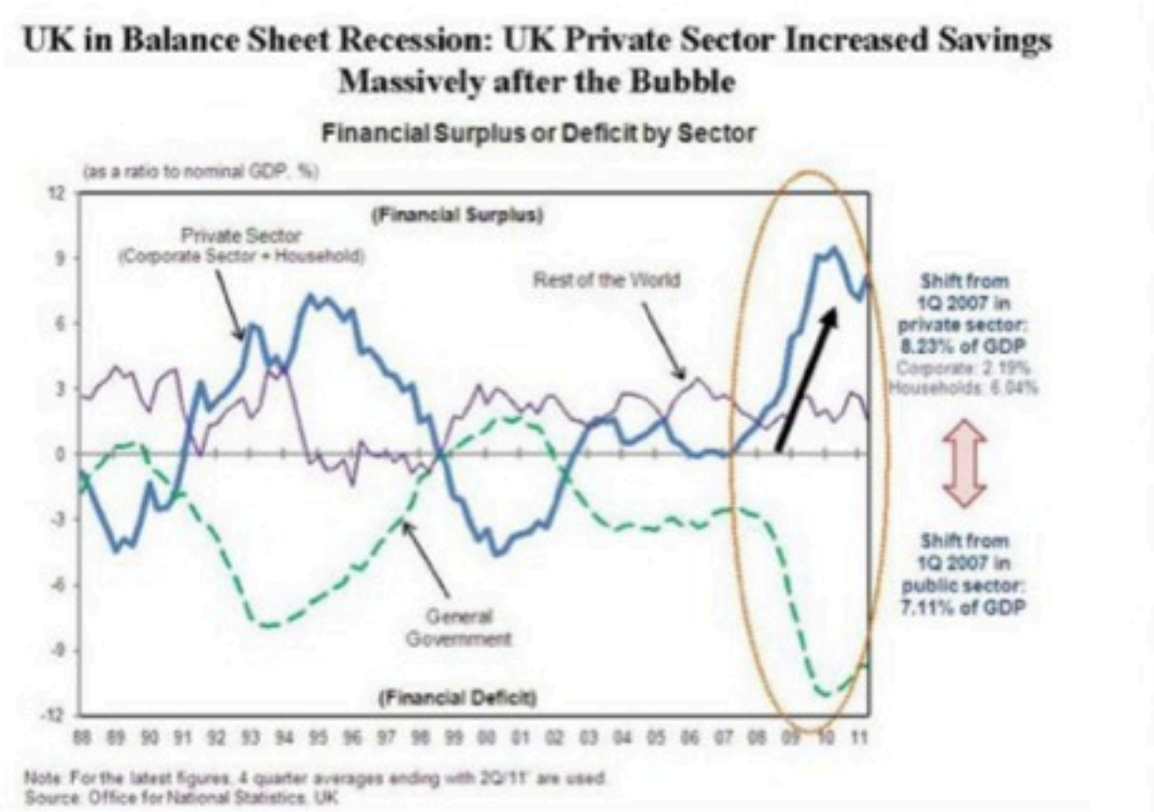


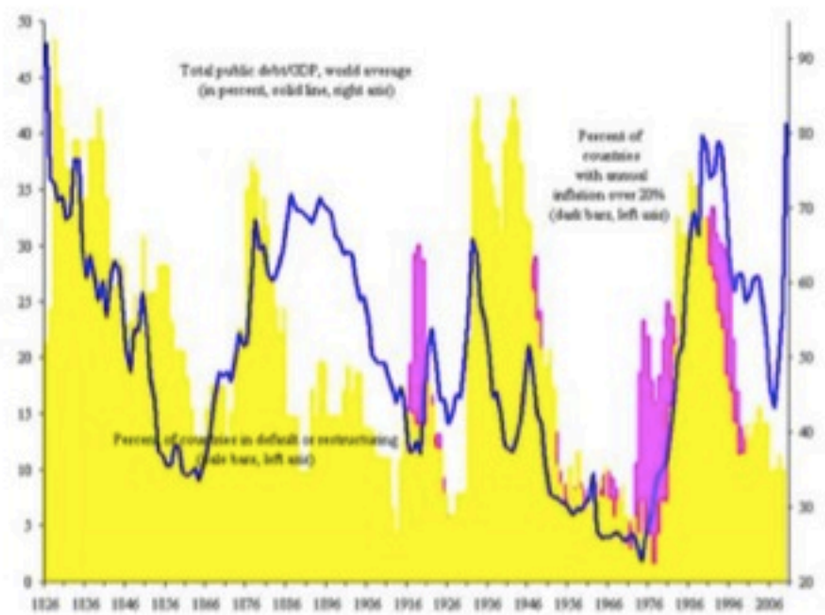
Optimal Aspect Ratio

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

Don't

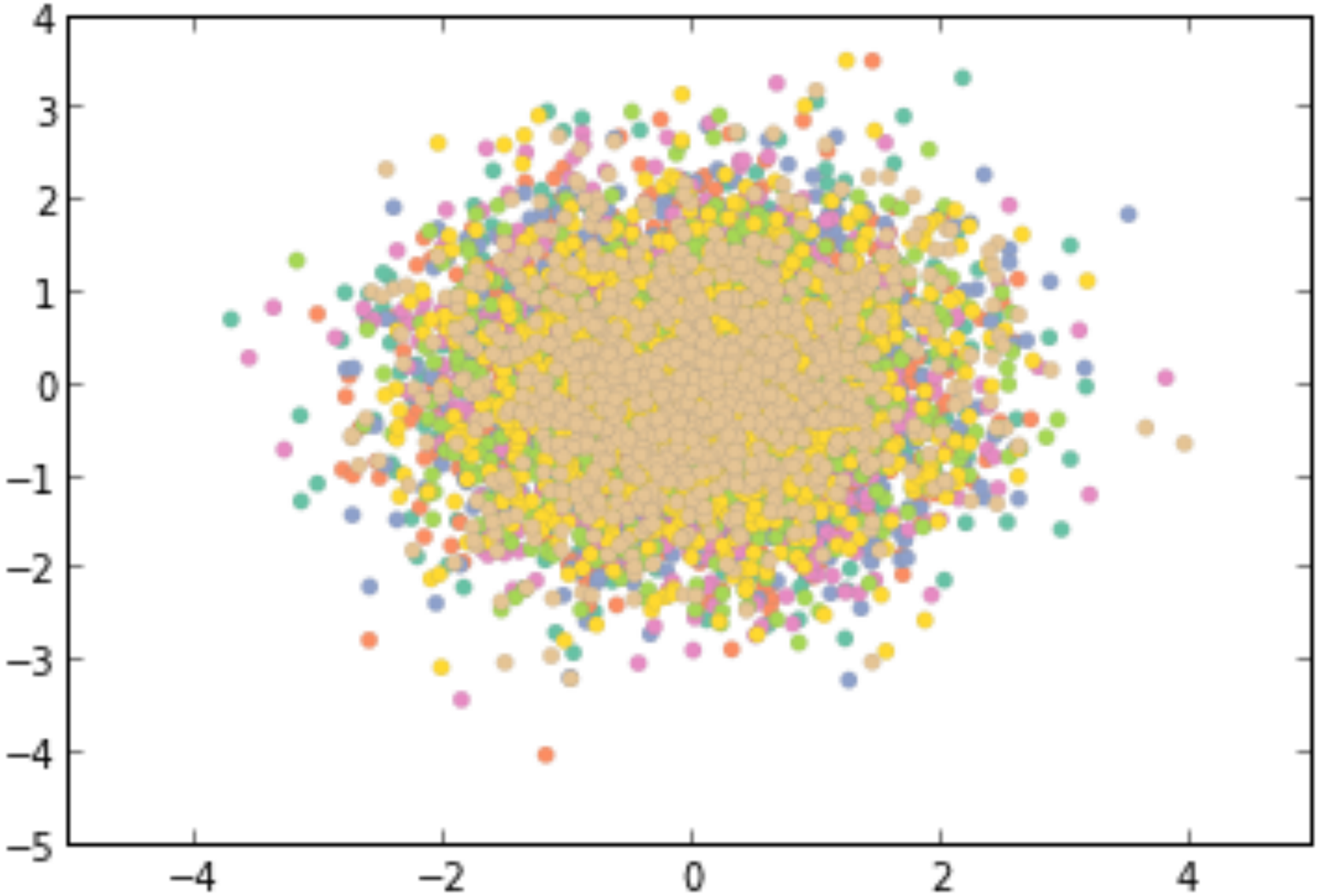

Interest Rates on 10-Year Government Bonds
In percent
Introduction of the Euro
Lehman Bankruptcy
Greece
Ireland
Portugal
Spain
Italy
France
Germany
18
16
14
12
10
8
6
4
2
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011


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
12
9
6
3
0
-3
-6
-9
-12
88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11
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)
1826 1836 1846 1856 1866 1876 1886 1896 1906 1916 1926 1936 1946 1956 1966 1976 1986 1996 2006

Correlations

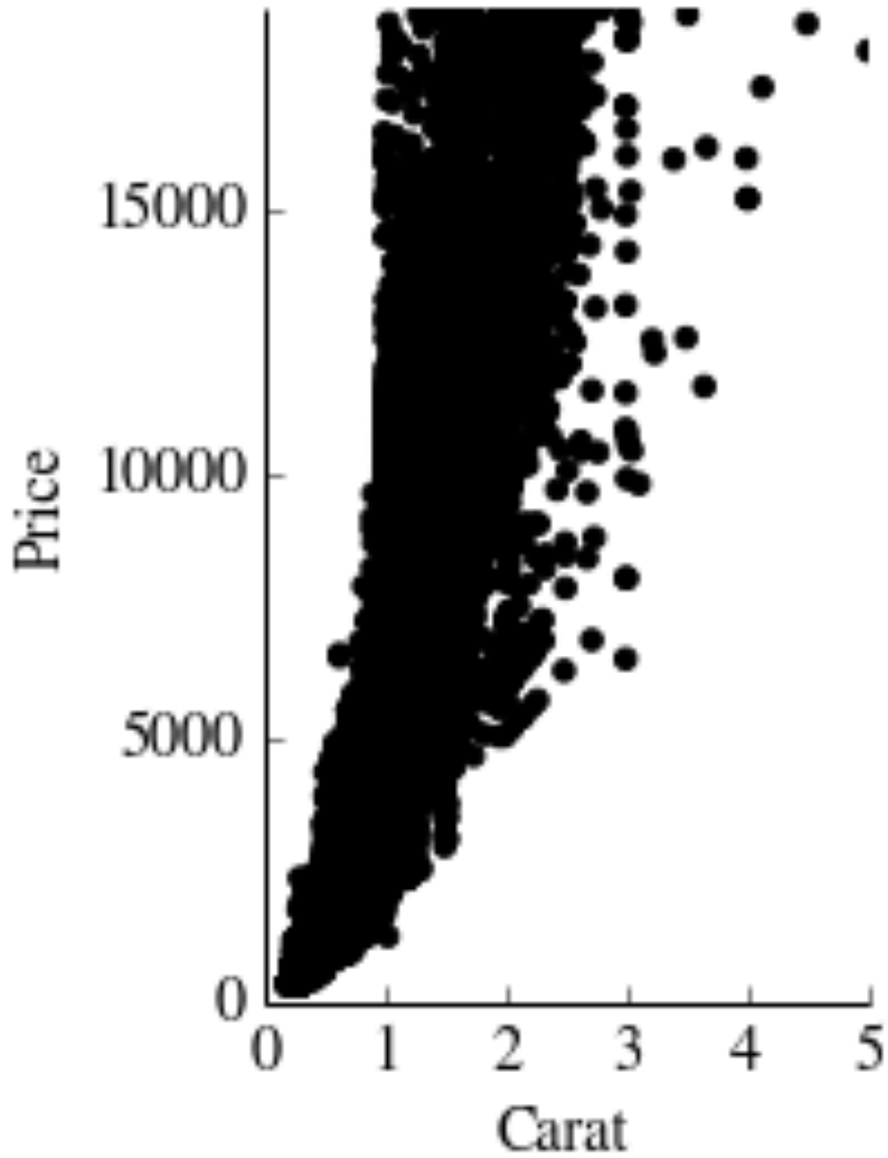
Scatterplots



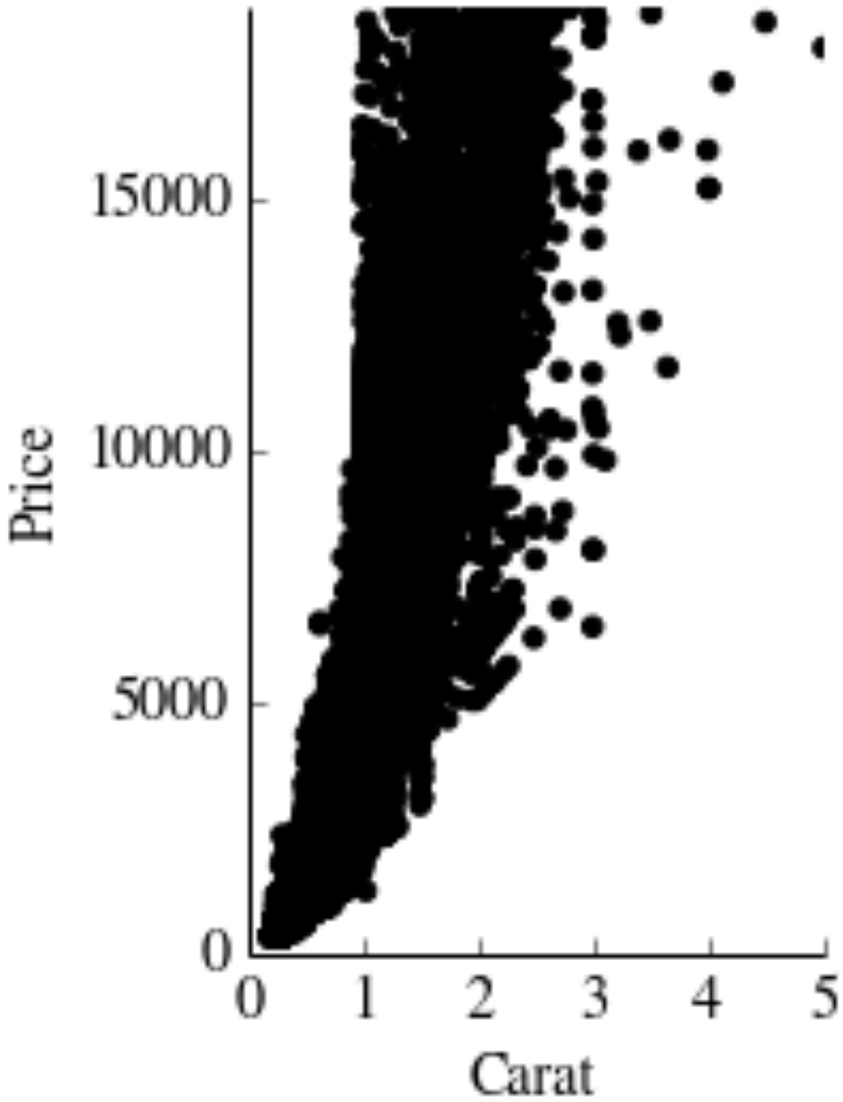
Light Grey Border

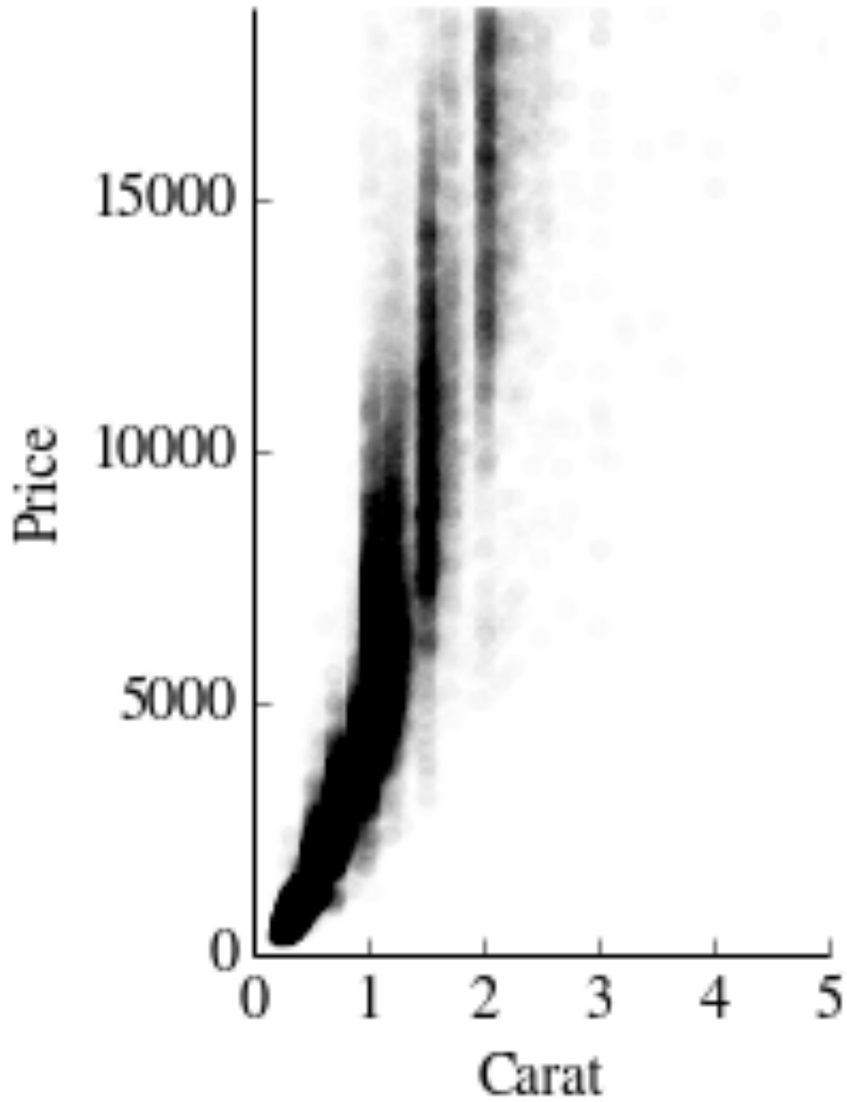
http://nbviewer.ipython.org/5357268

Overplotting



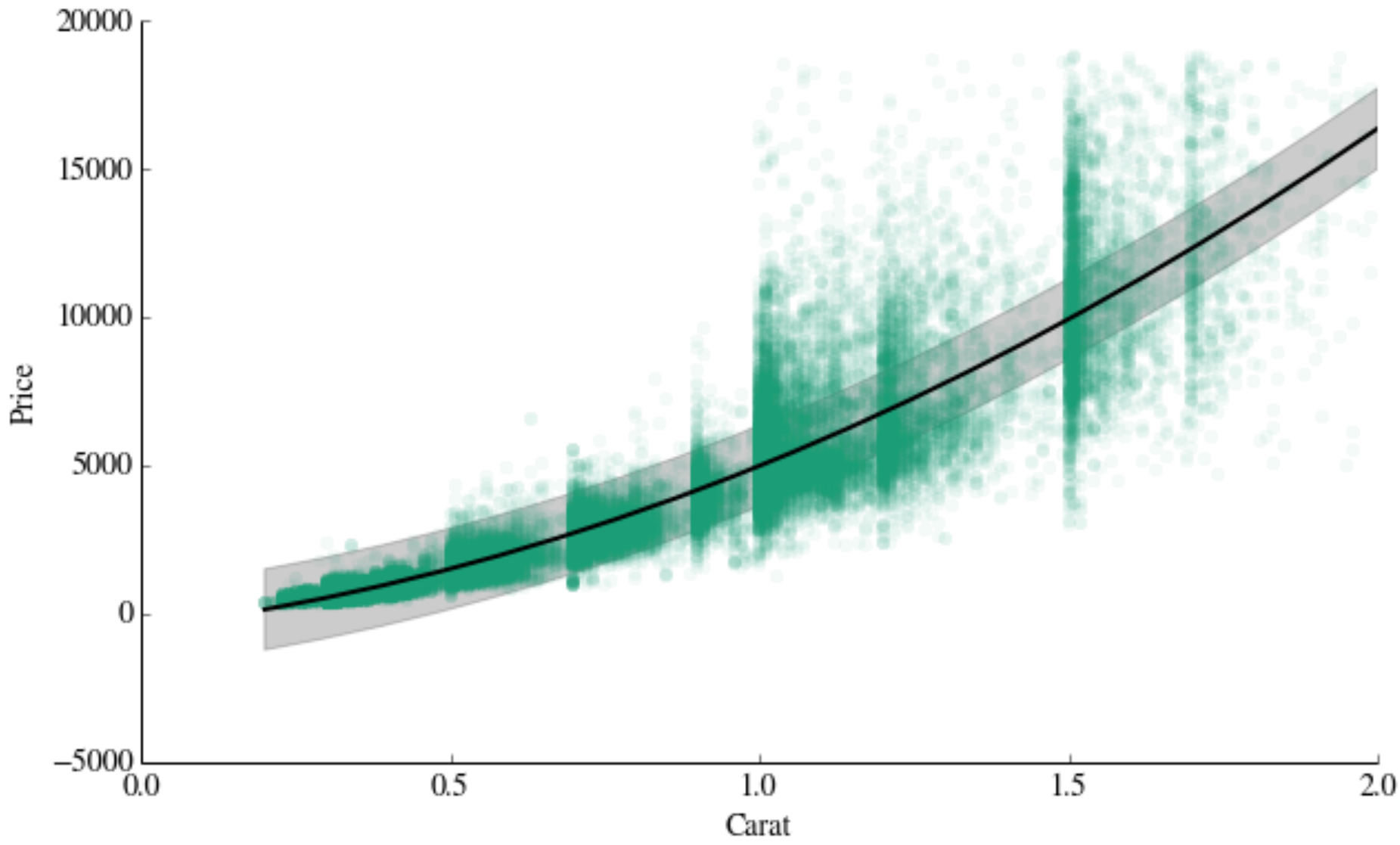
Overplotting



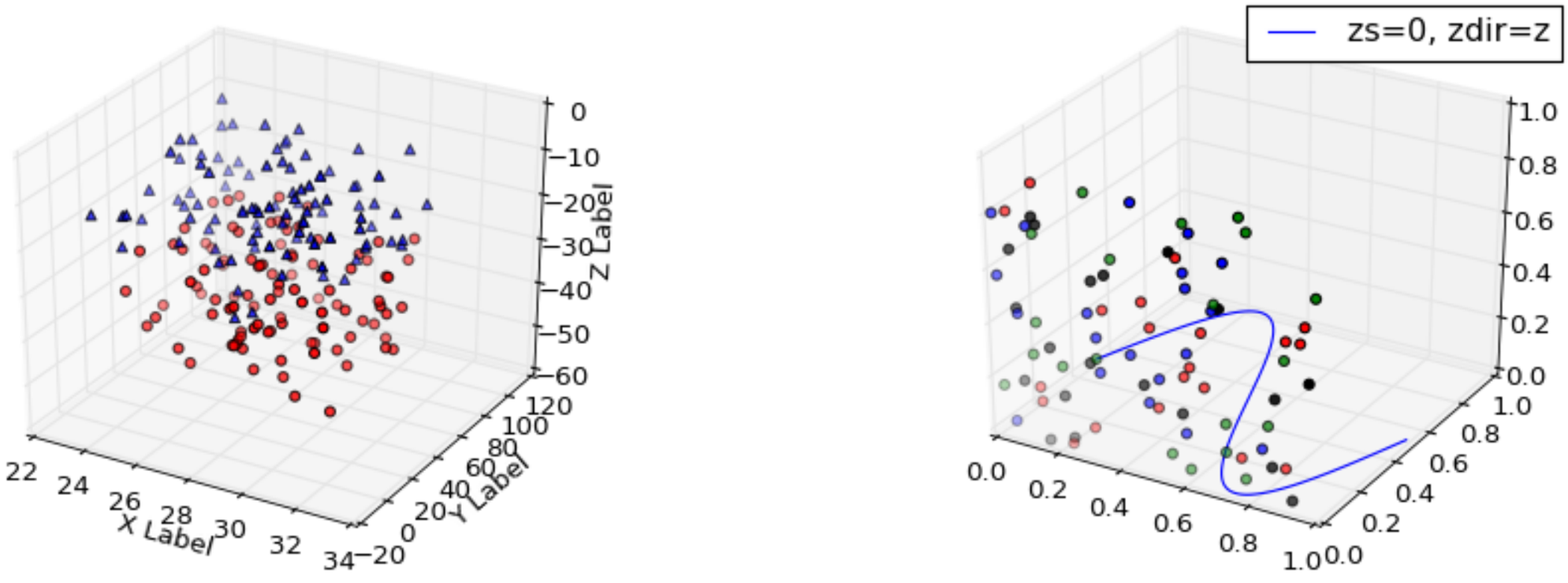


alpha = 1/100

Trend Lines

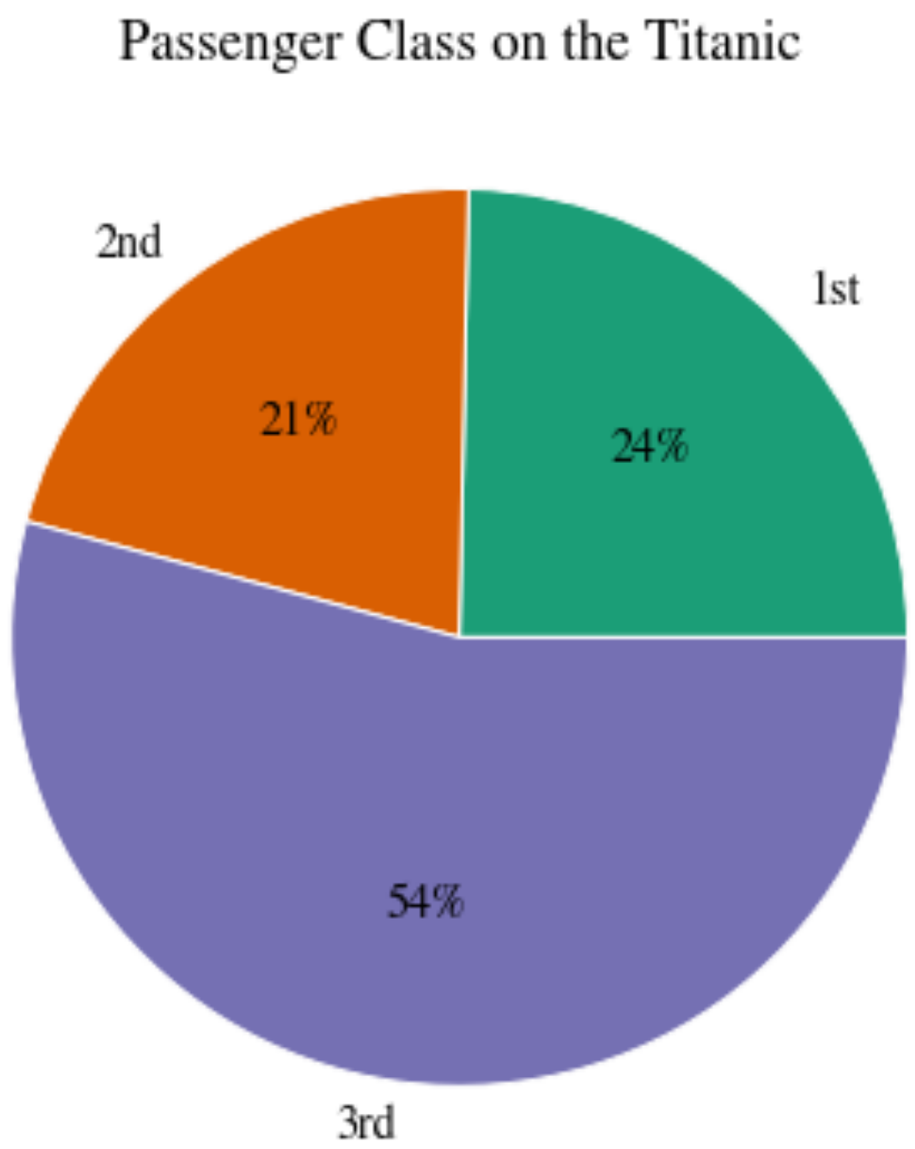


Don‘t



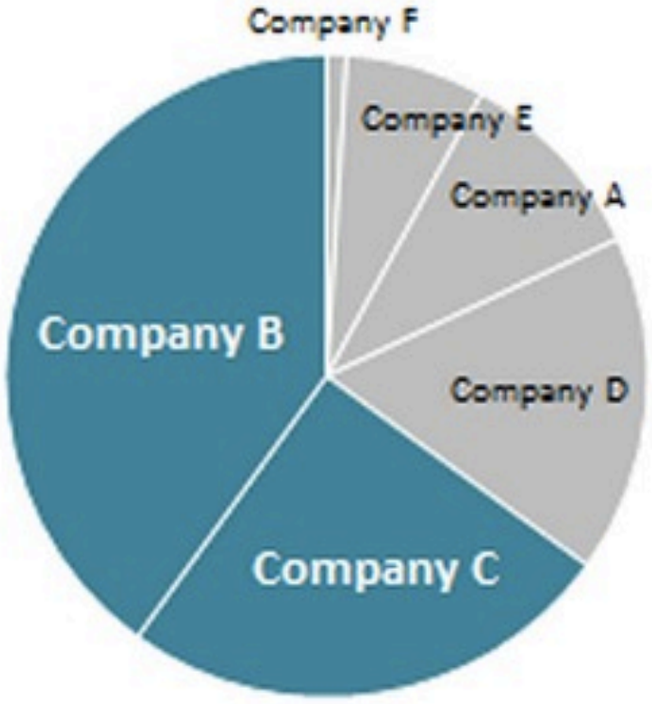
Compositions

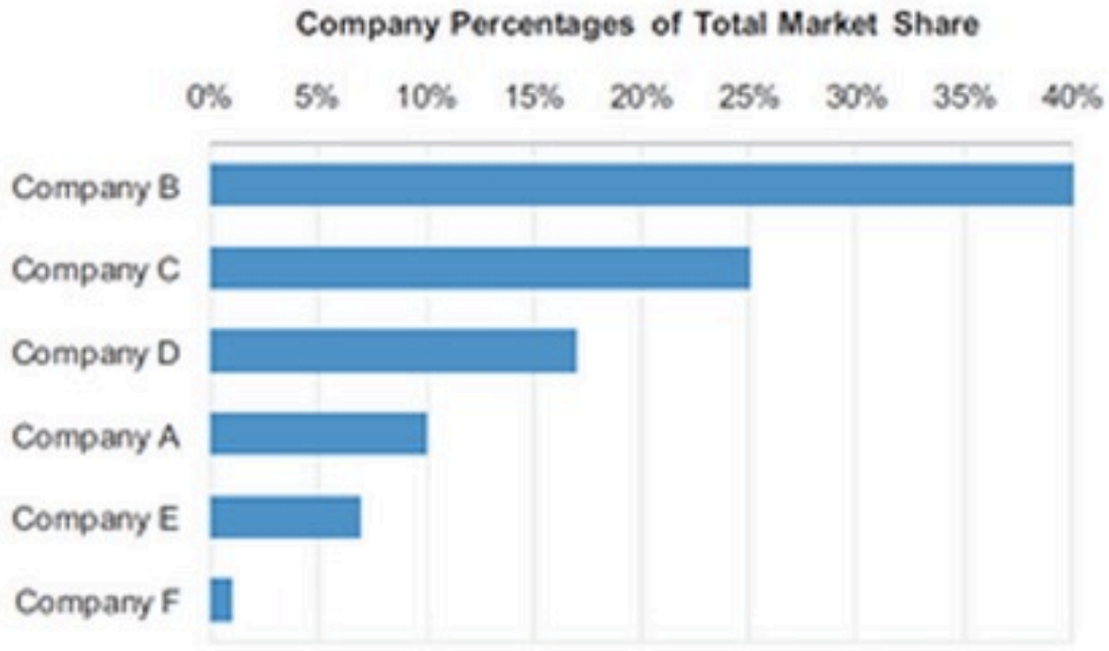
Pie Charts



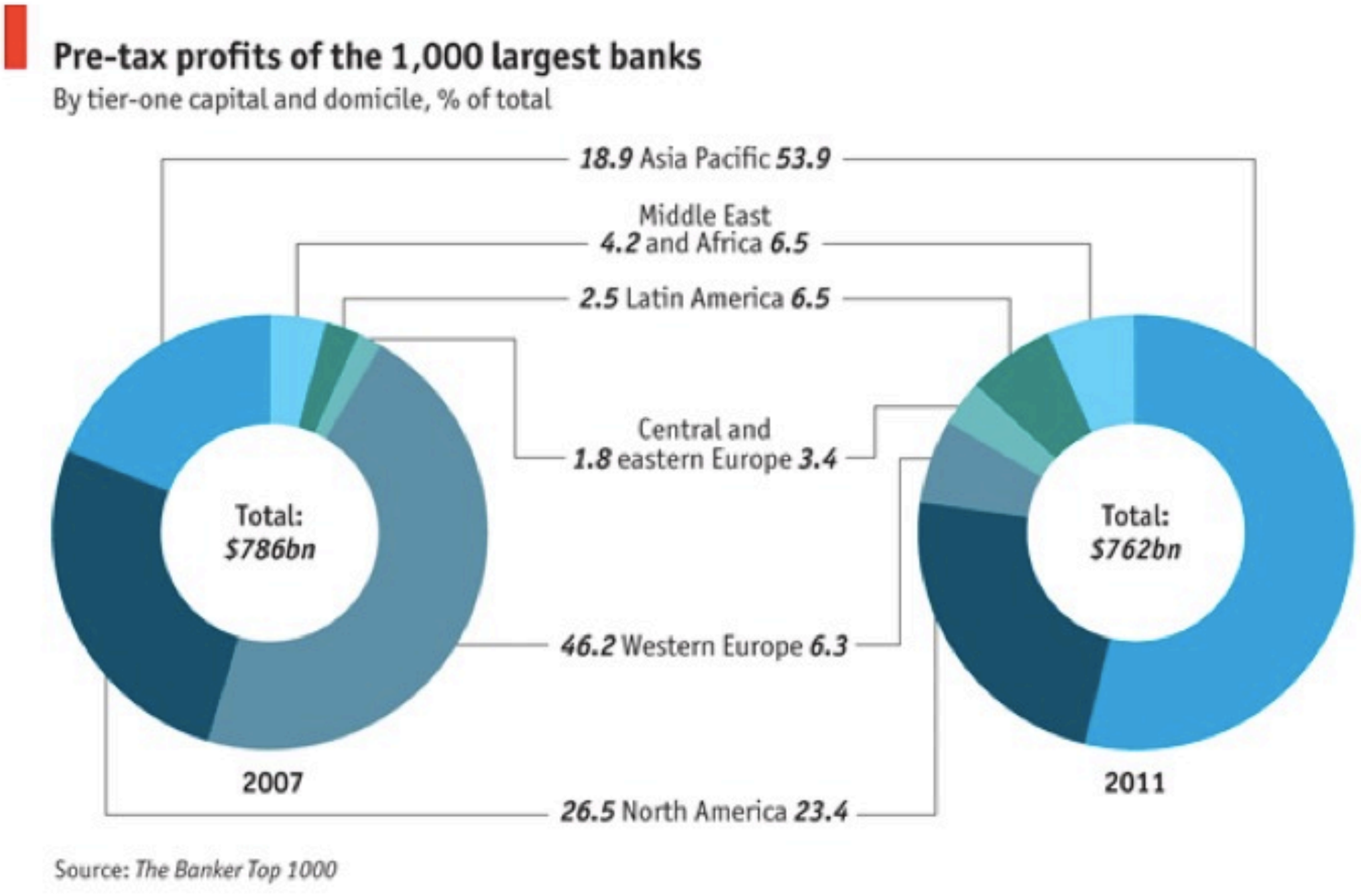
Pie vs. Bar Charts

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



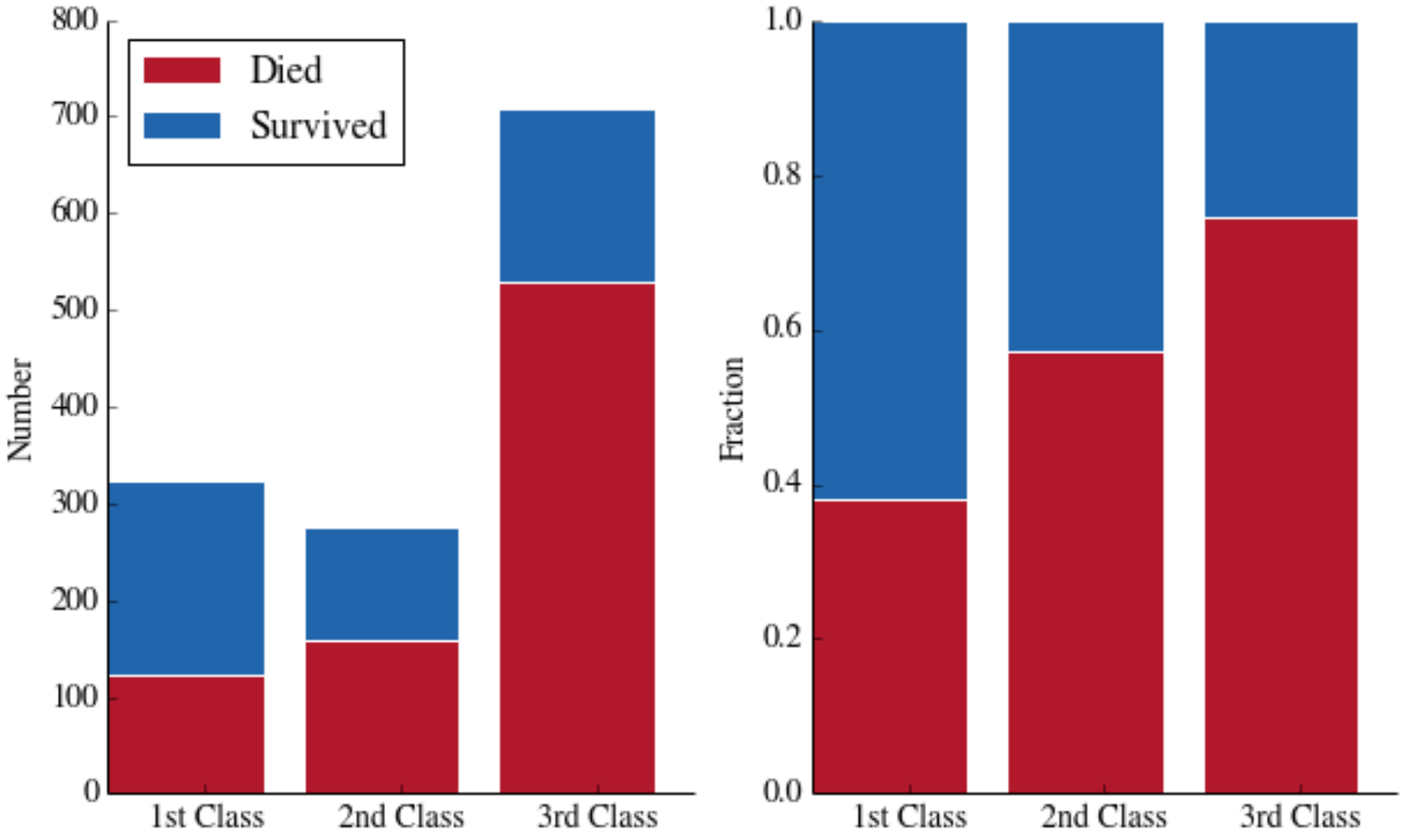


Donut Chart

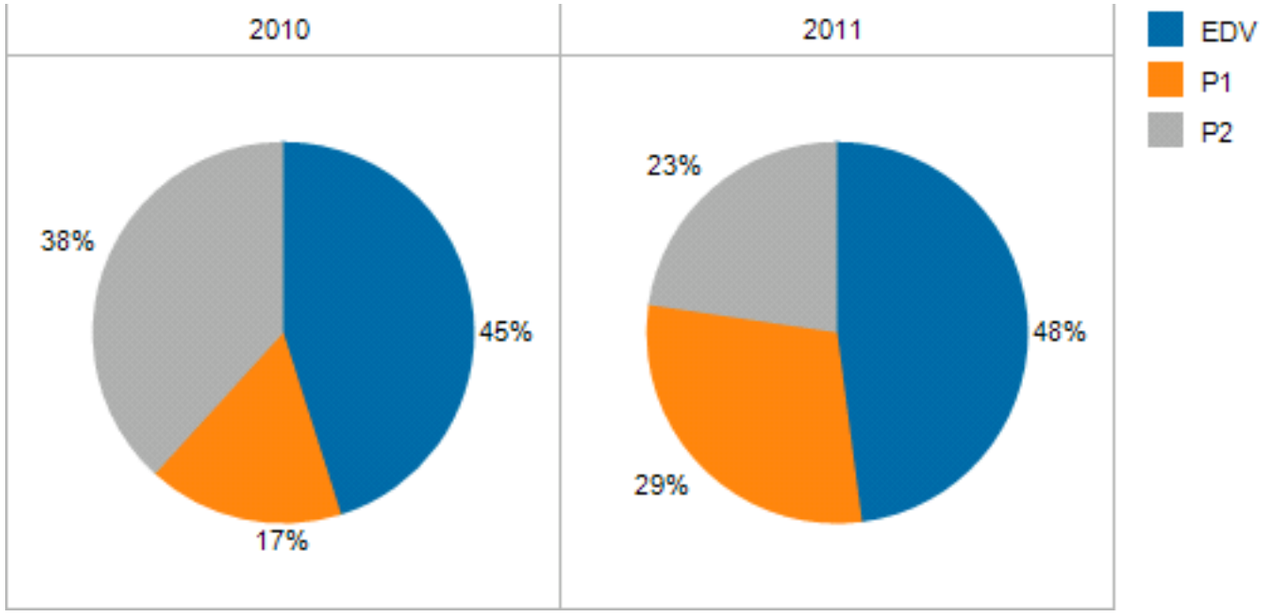


The Economist Daily Chart

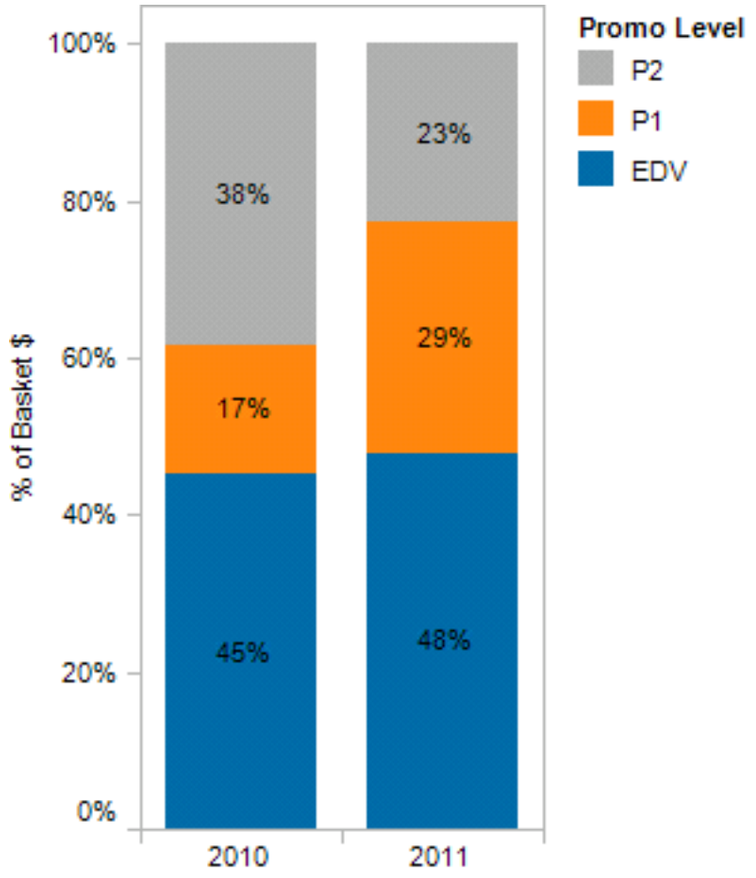
Stacked Bar Chart



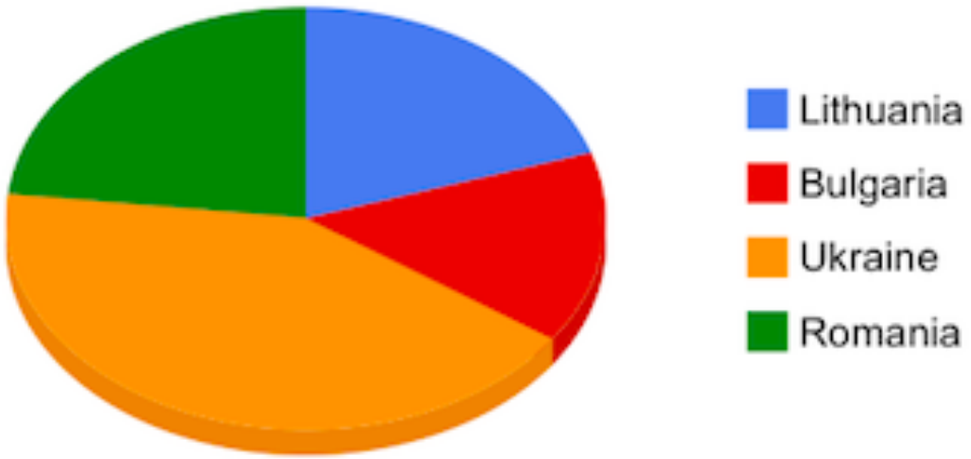
Stacked Bar Chart



V.S.



Don’t



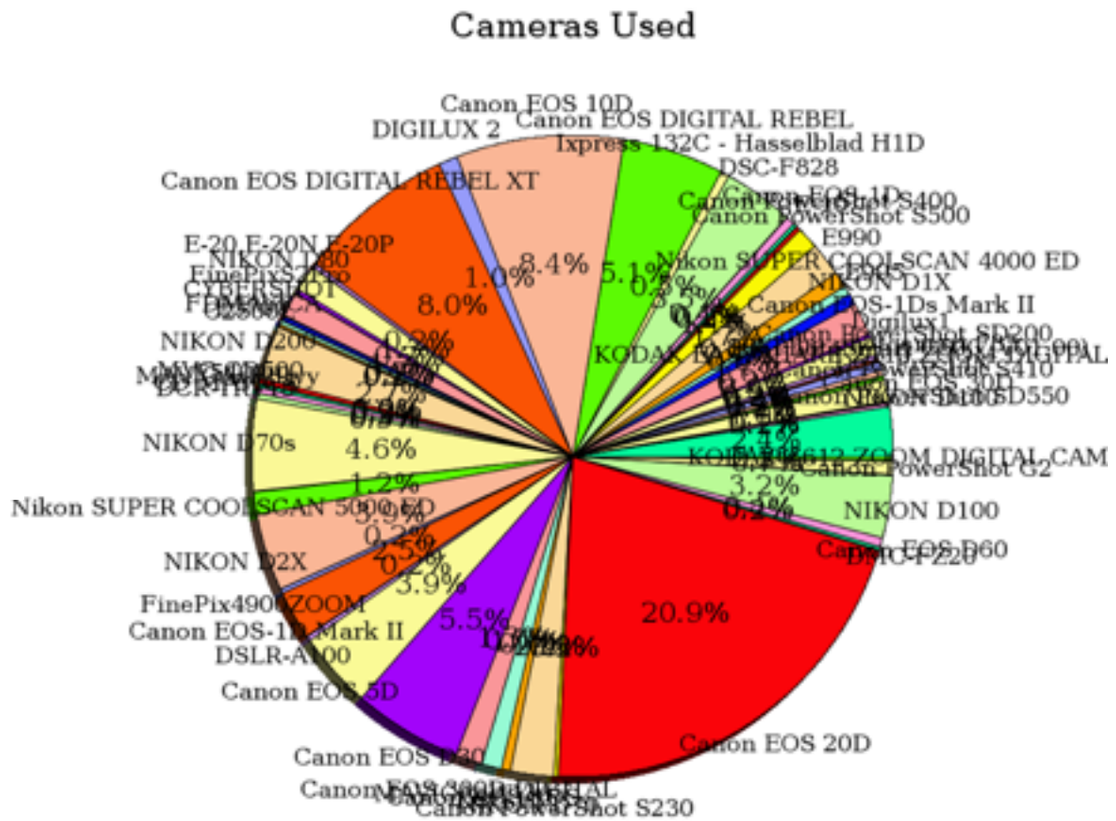
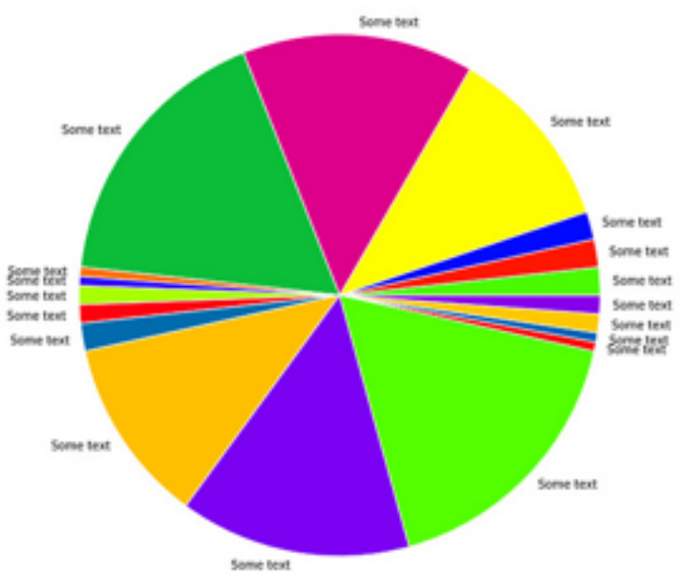
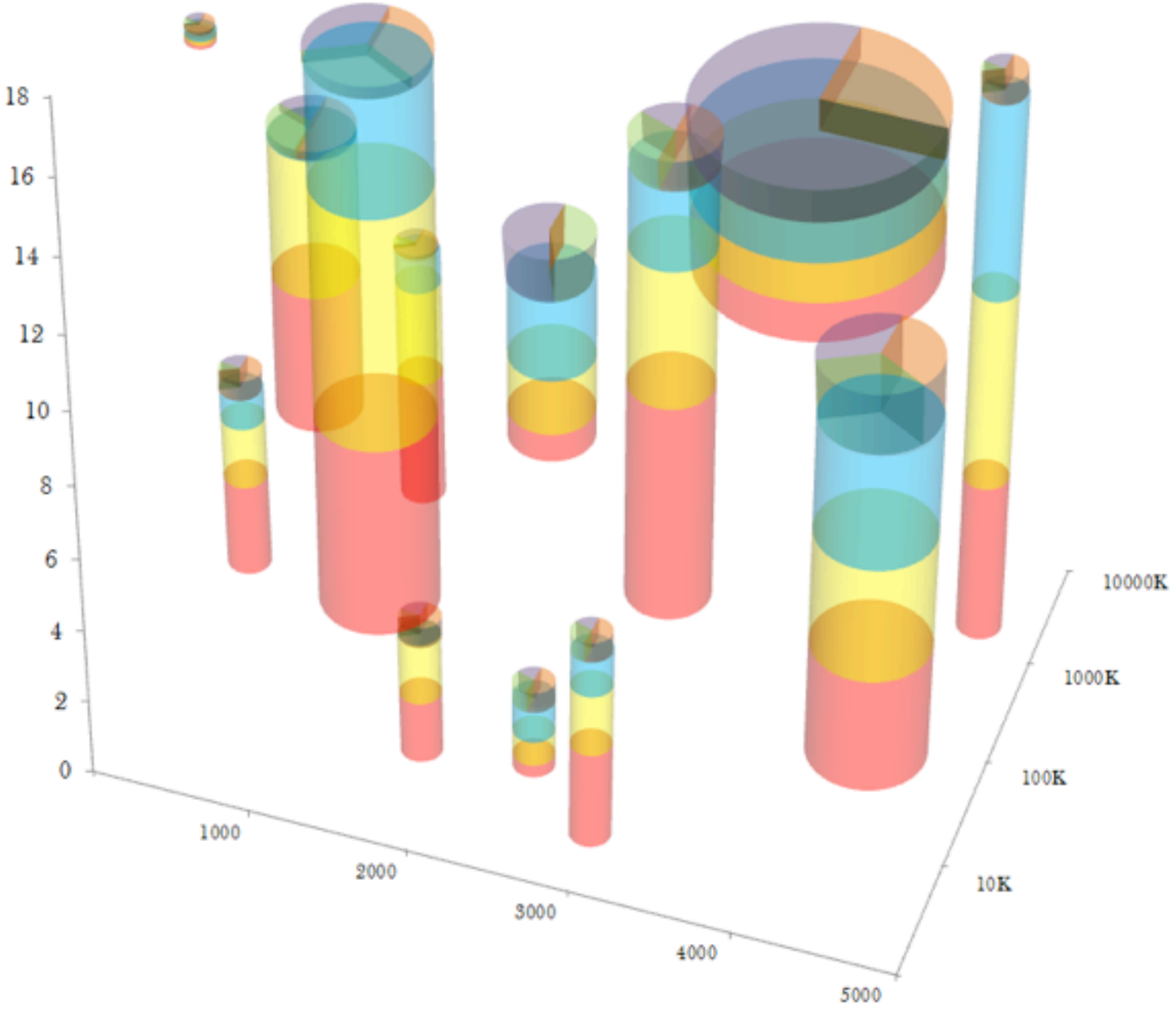


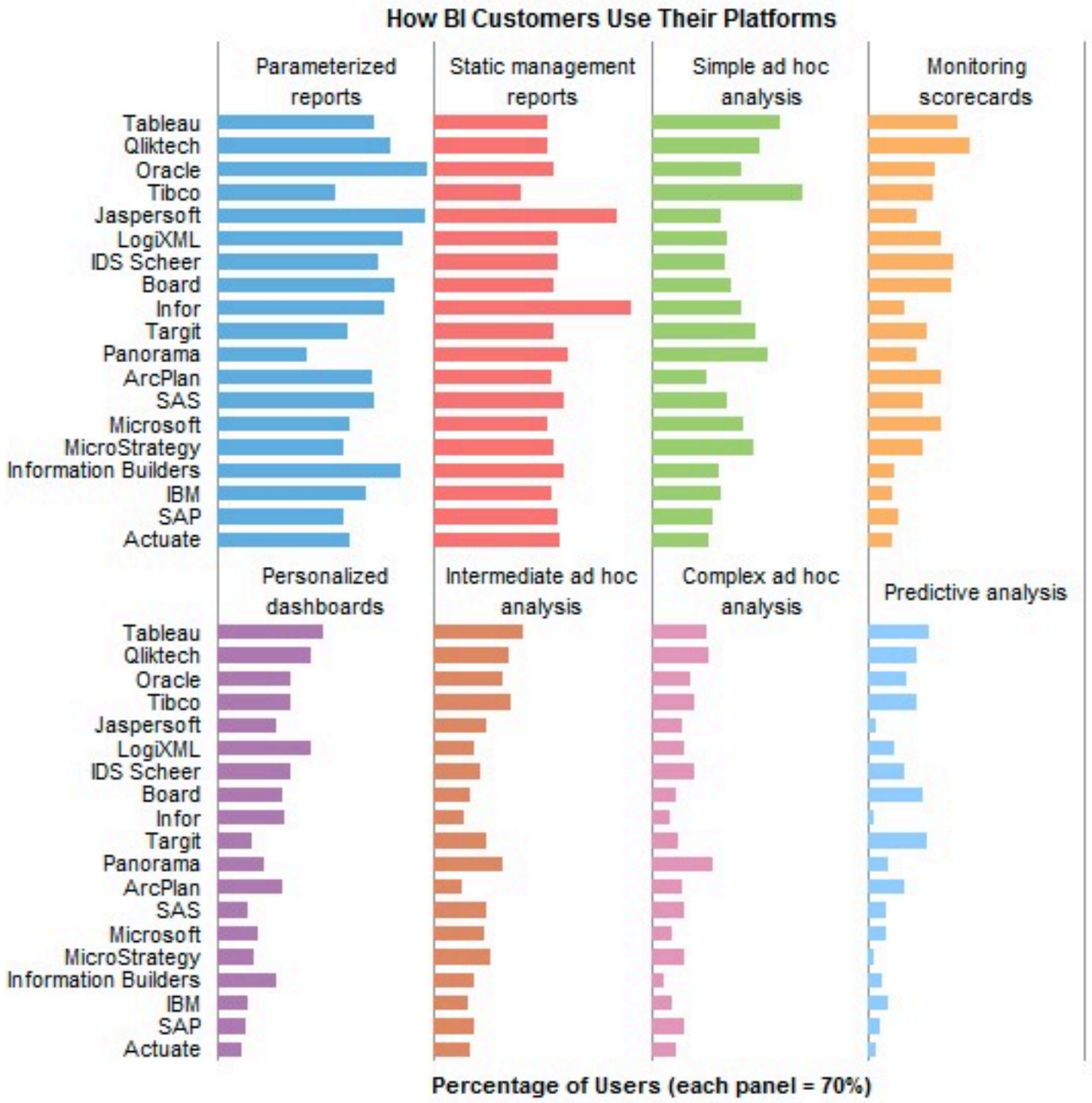
Figure 5



eagerpies.com

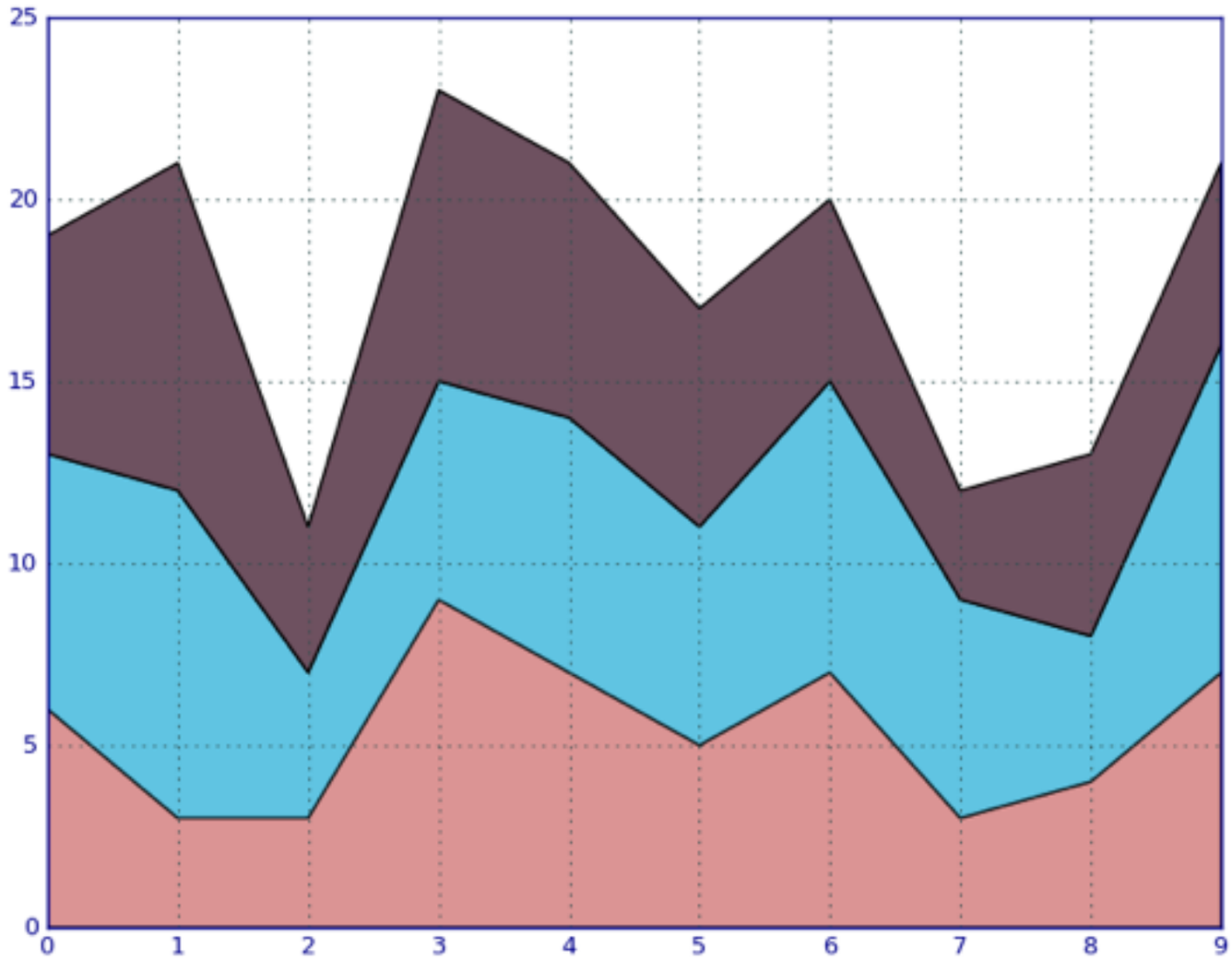


Small Multiples



Peltier Tech Blog

Stacked Area Chart

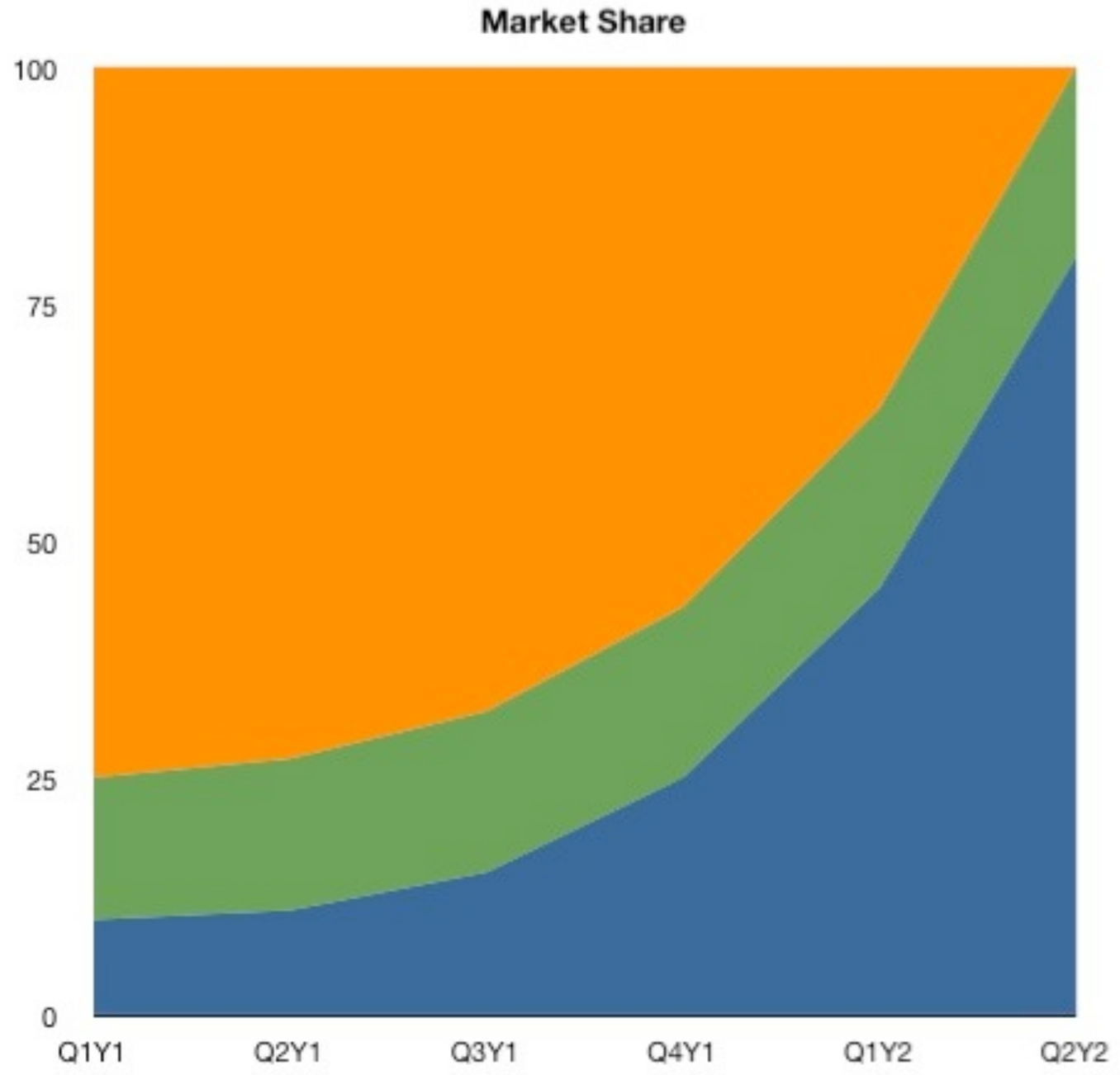


http://stackoverflow.com/questions/2225995/how-can-i-create-stacked-line-graph-with-matplotlib

100% Stacked Area Chart

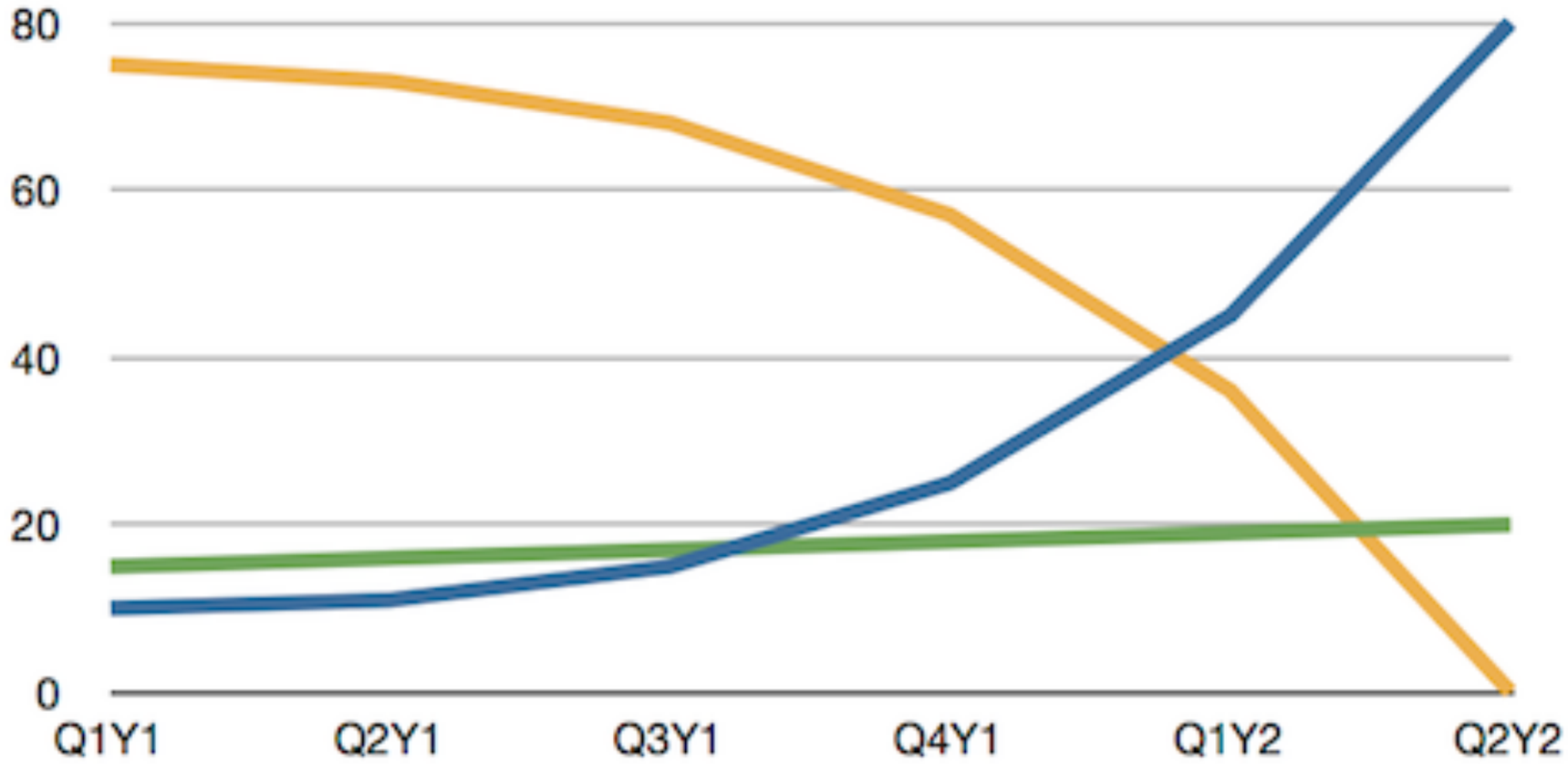
http://stackoverflow.com/questions/16875546/create-a-100-stacked-area-chart-with-matplotlib

Stacked Area vs. Line Graphs

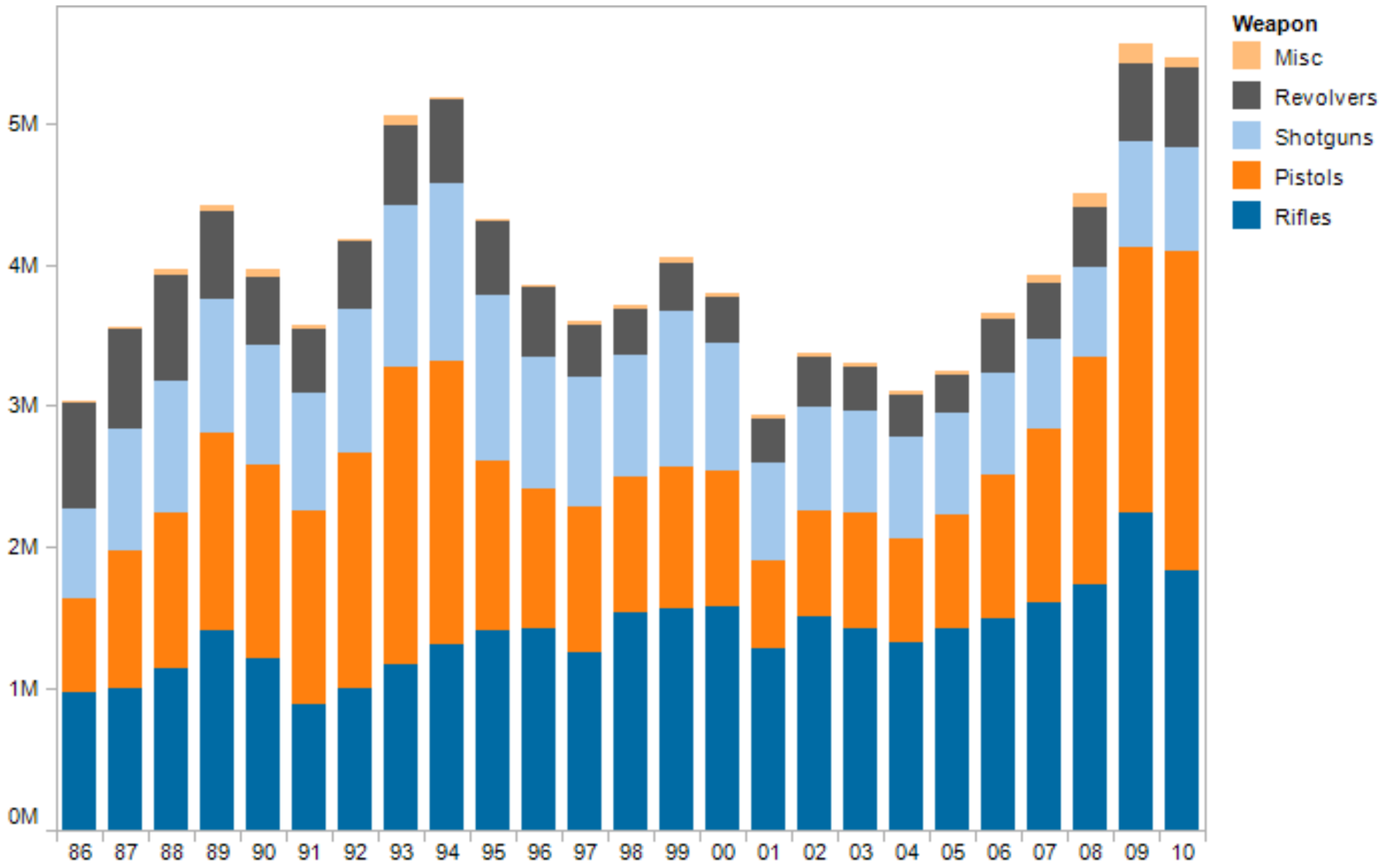


leancrew.com &
Practically Efficient

Stacked Area vs. Line Graphs

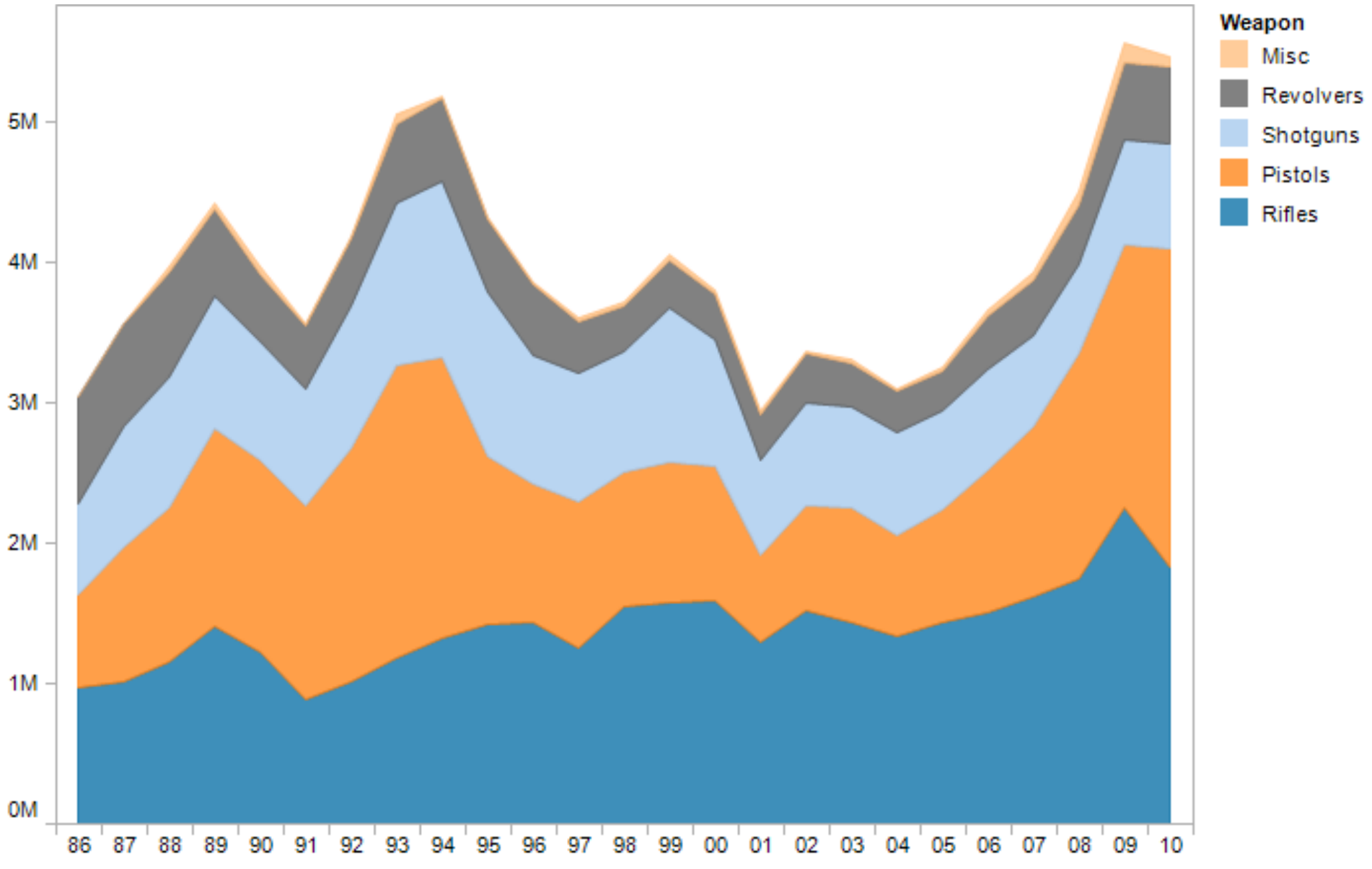


Stacked Area vs. Line Graphs



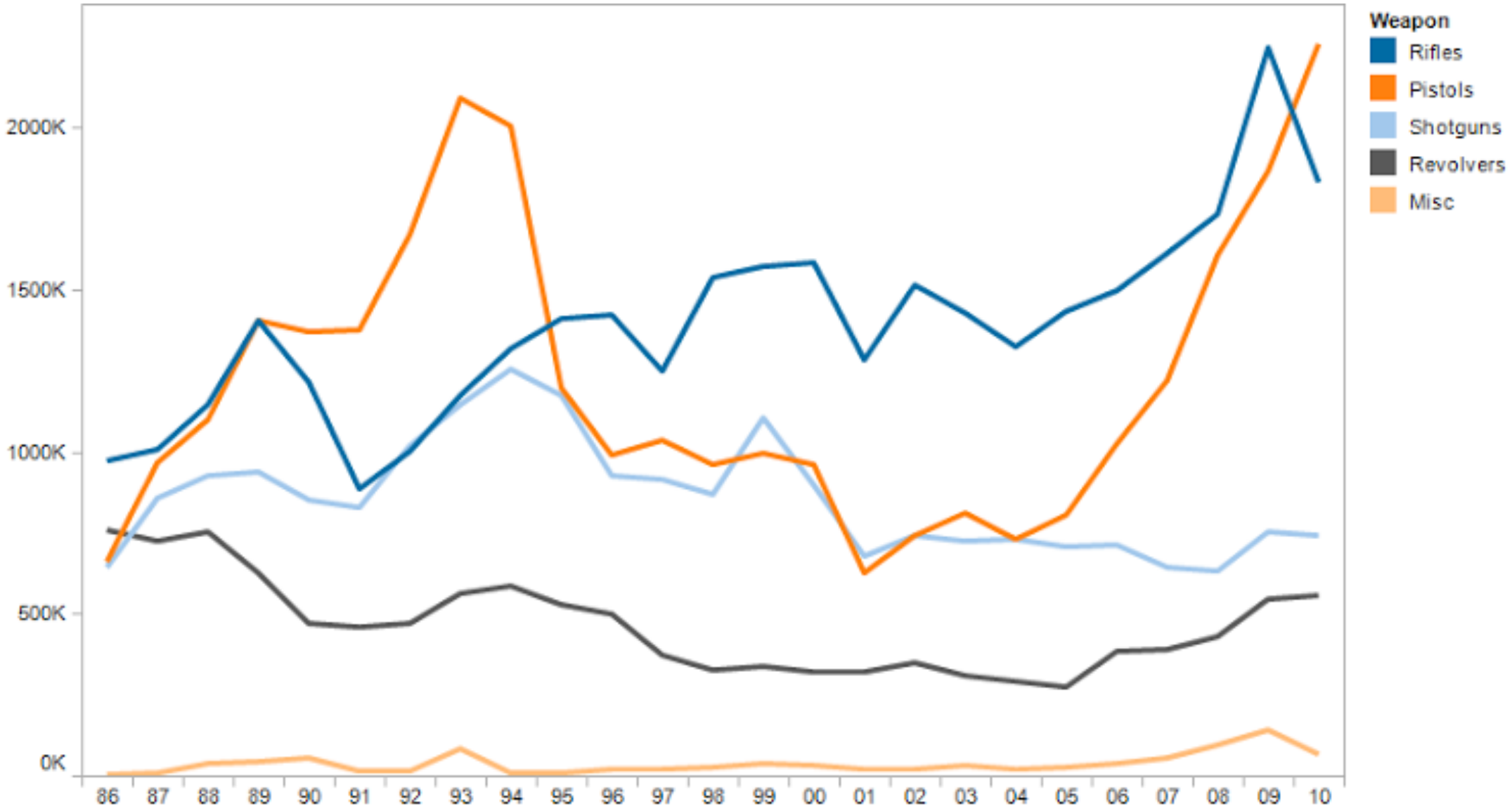
VizWiz, A. Kriebel

Stacked Area vs. Line Graphs



VizWiz, A. Kriebel

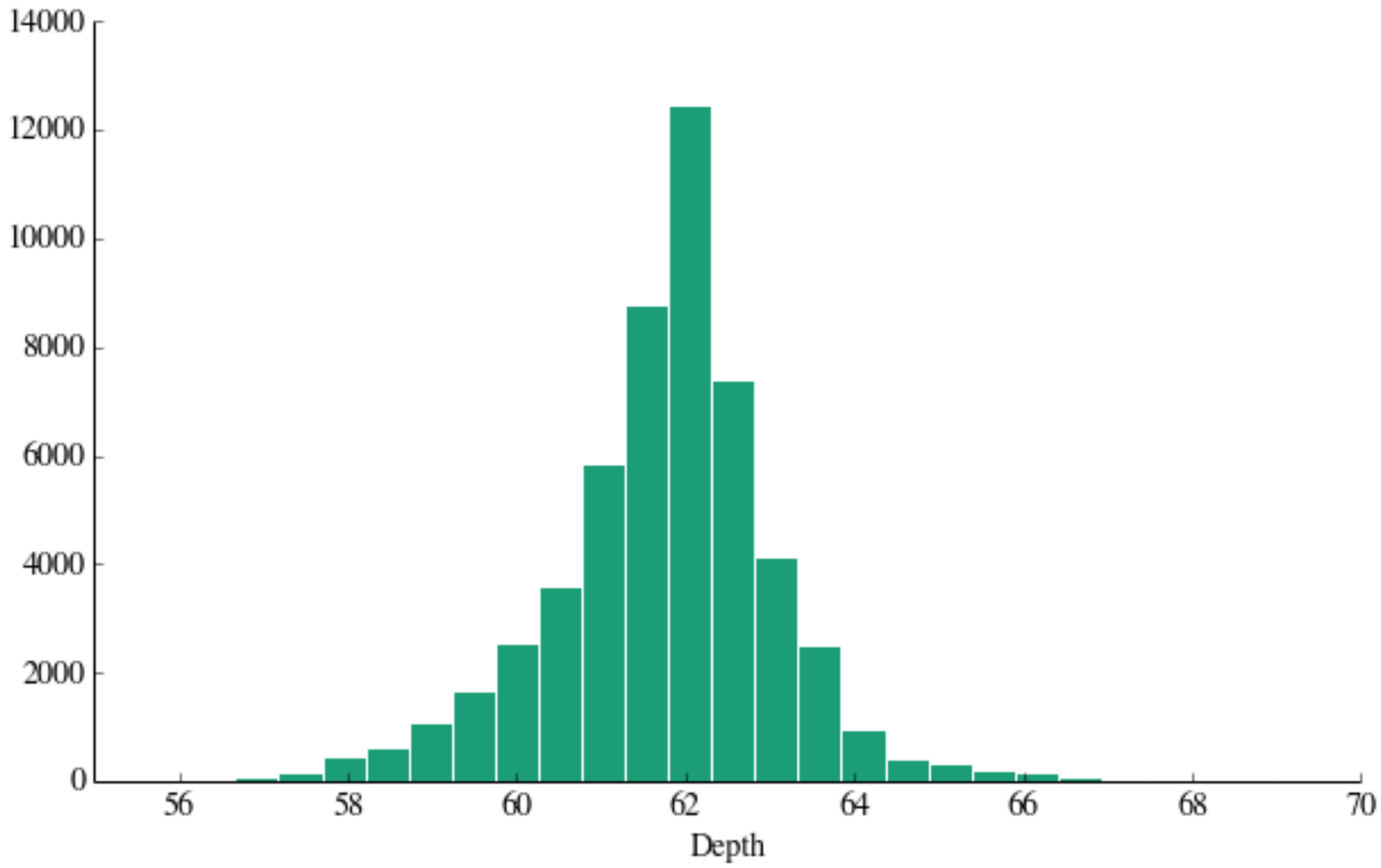
Stacked Area vs. Line Graphs



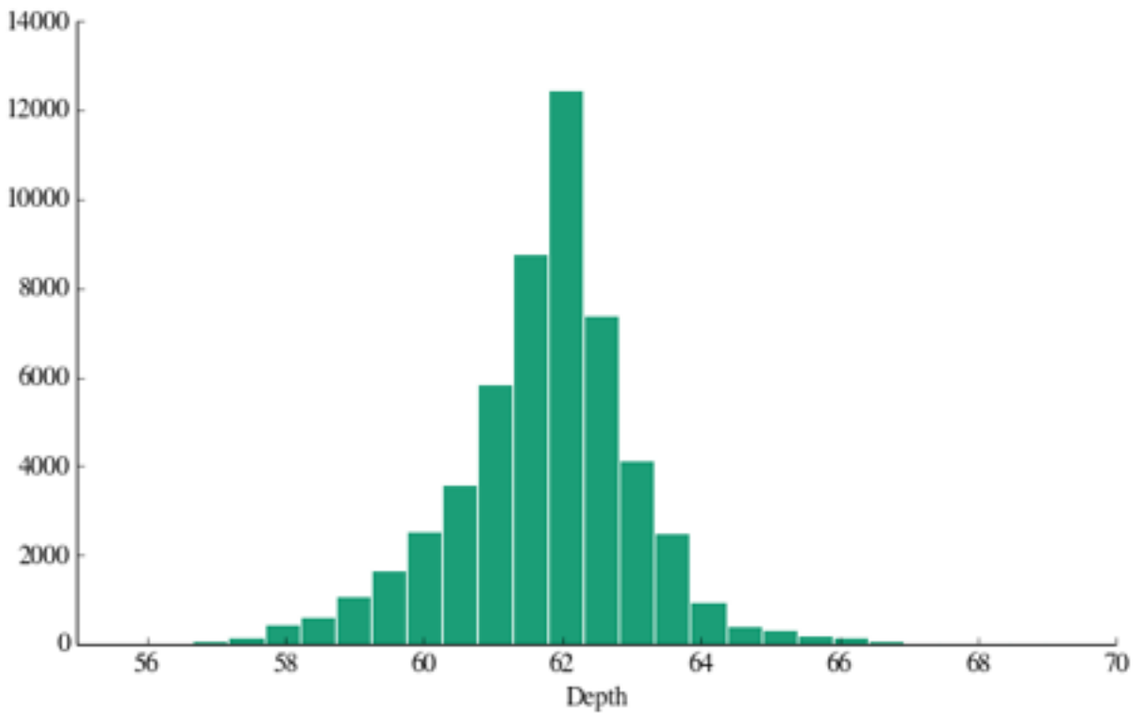
VizWiz, A. Kriebel

Distributions

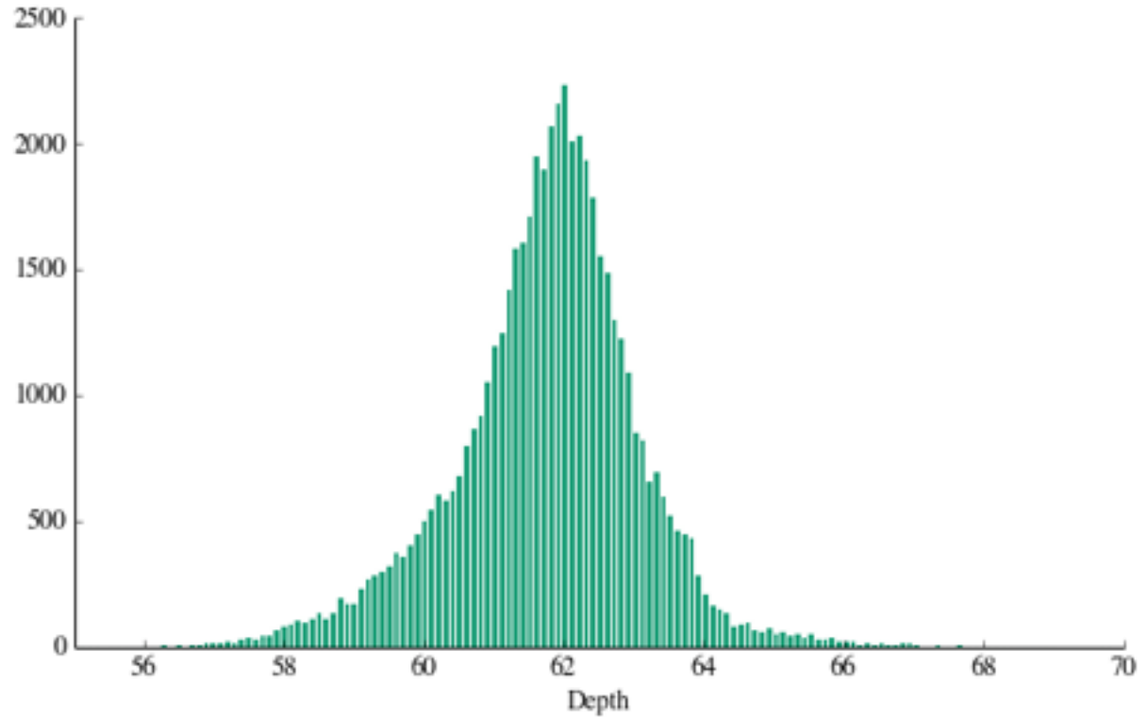
Histogram


14000
12000
10000
8000
6000
4000
2000
0
56
58
60
62
64
66
68
70
Depth

Histogram

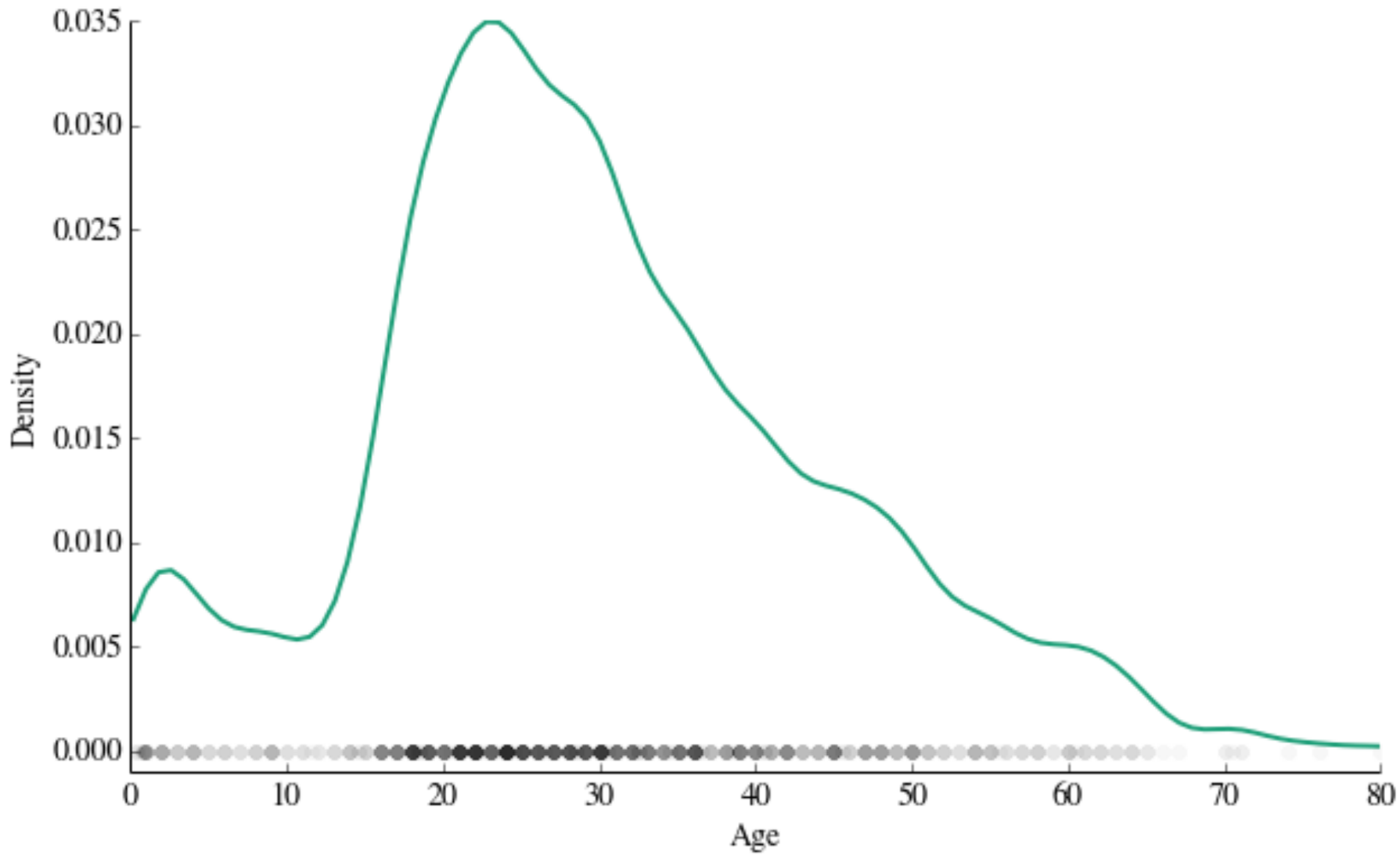


40 bins

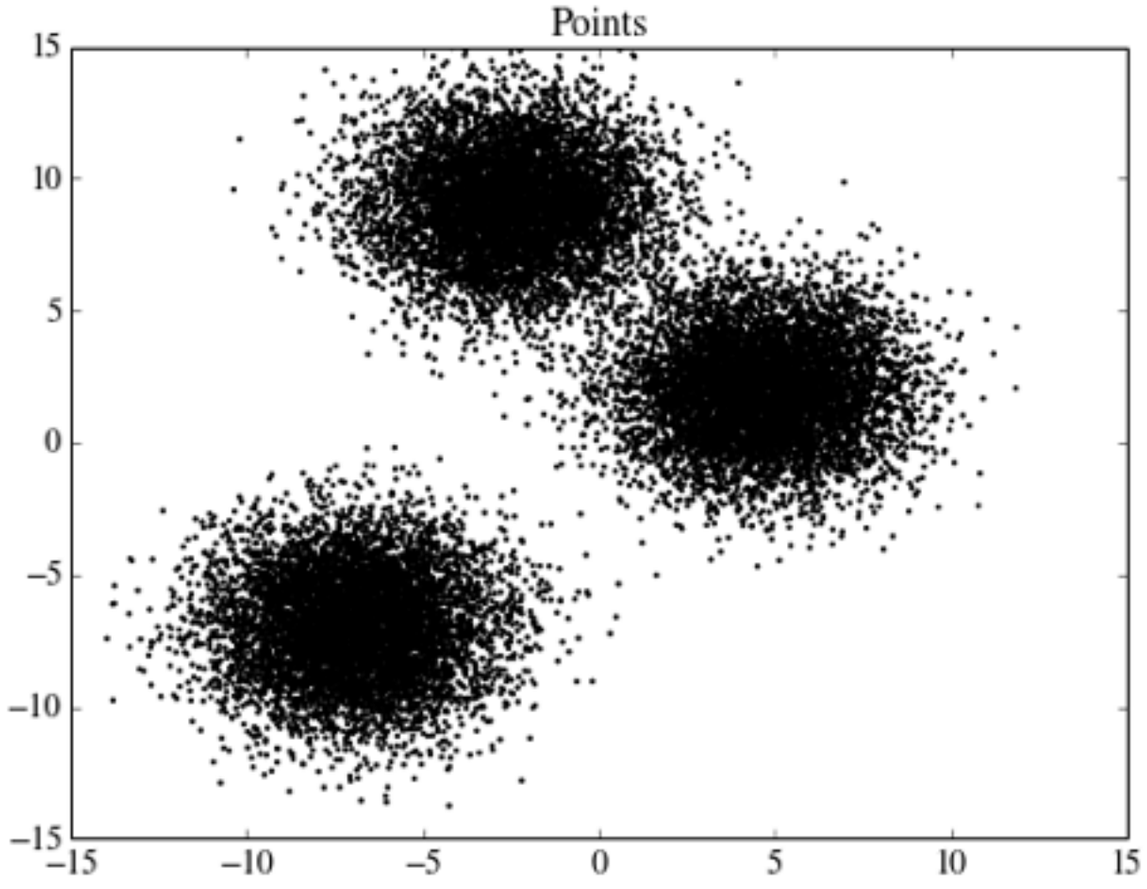


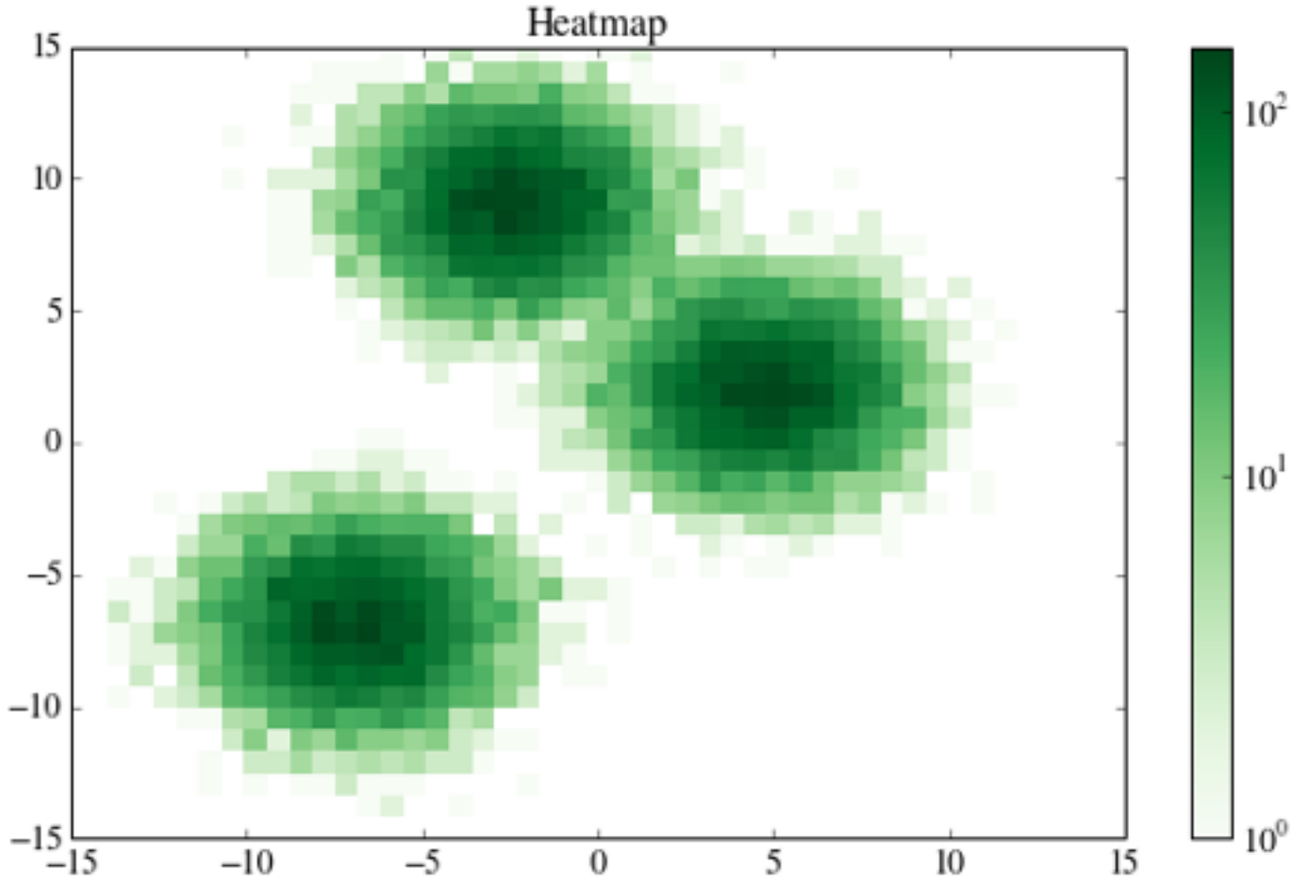
200 bins

Density Plots



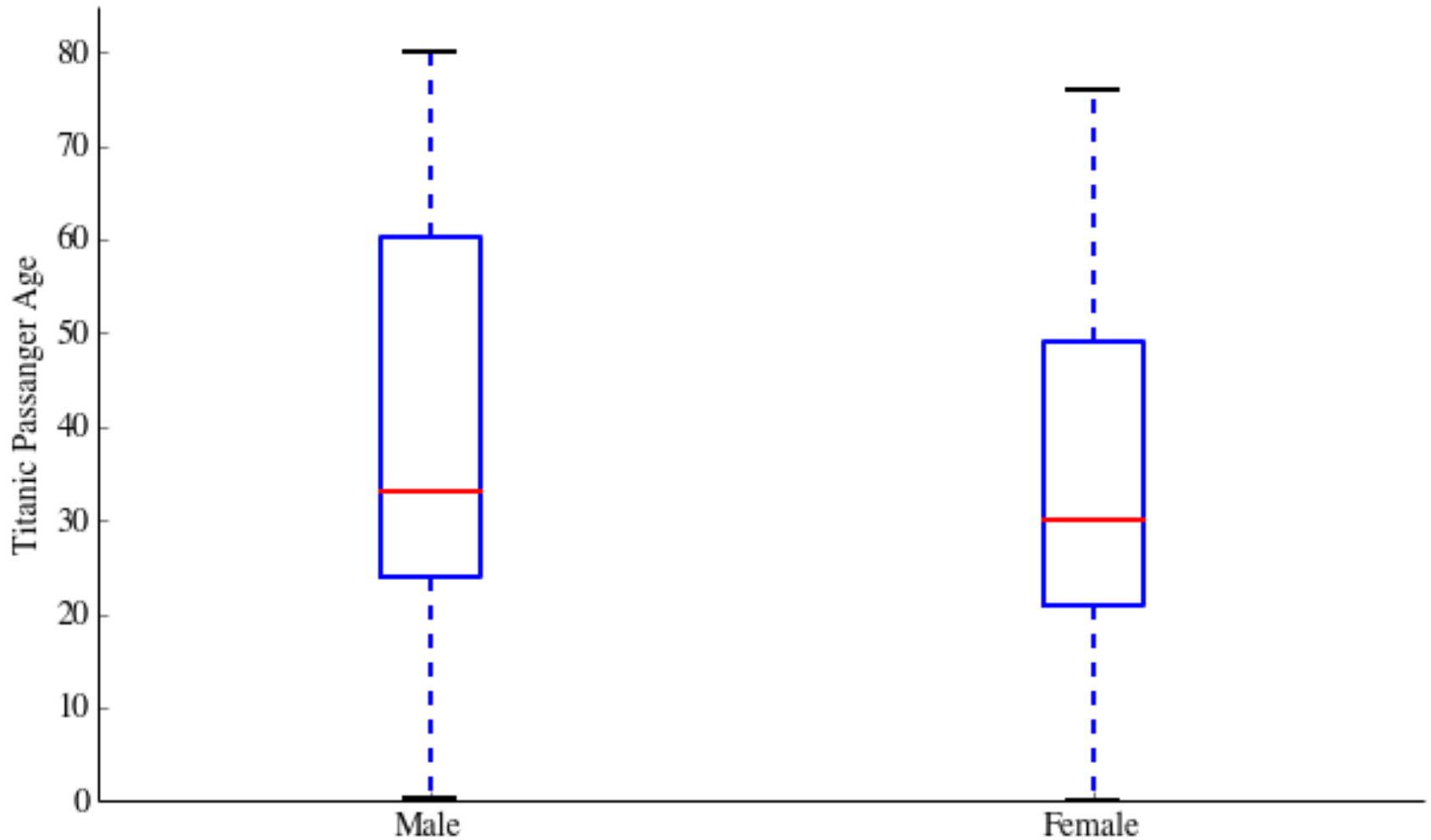
Heat Maps



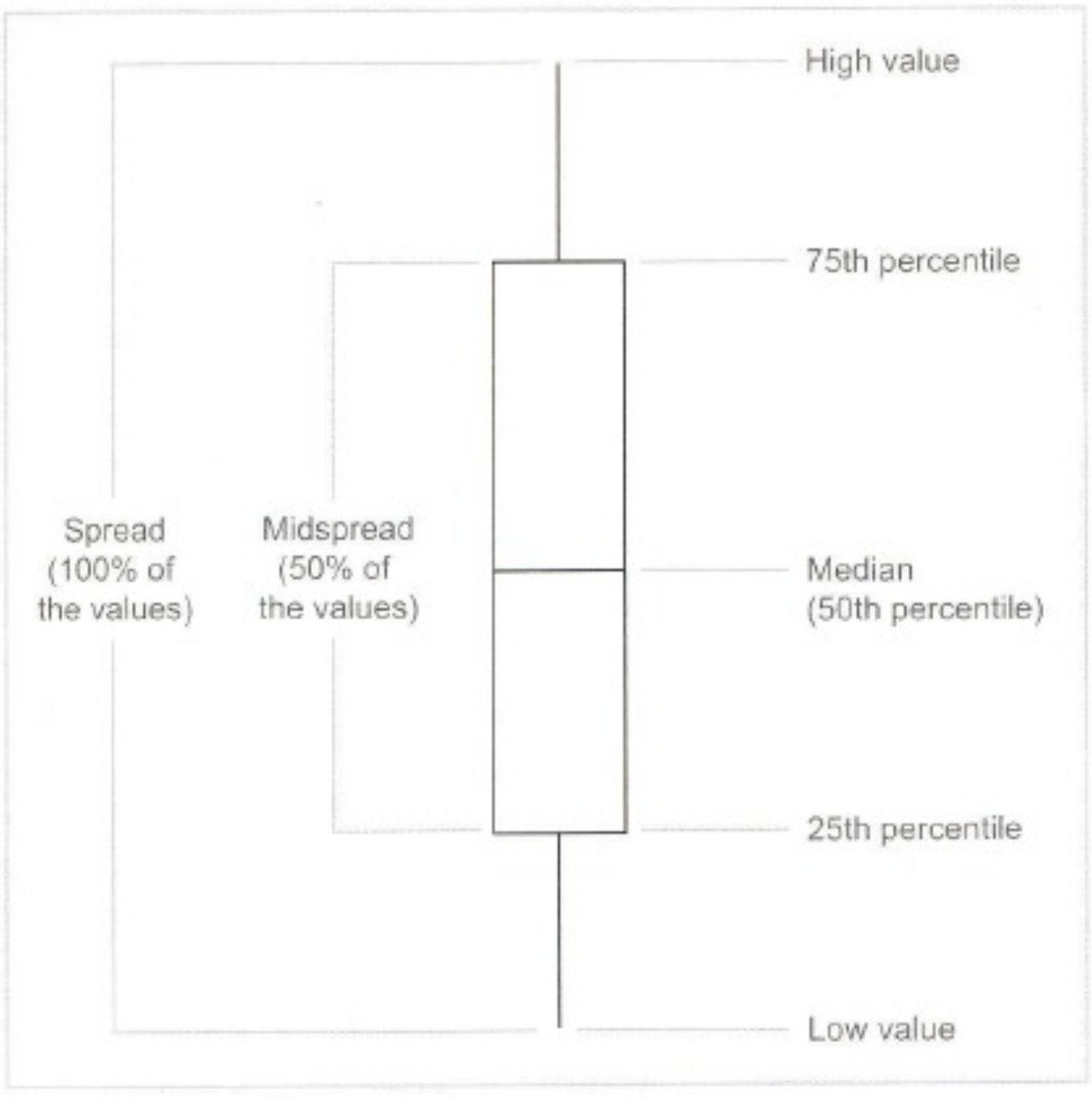


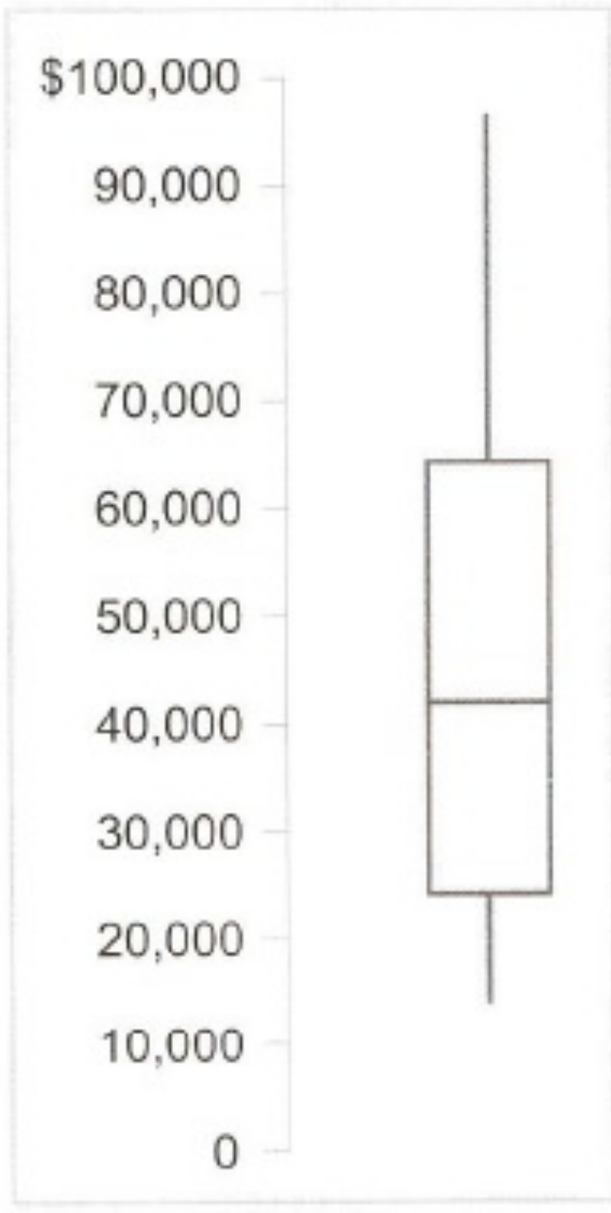
2D Density Plots

Box & Whisker Plots



Box & Whisker Plots

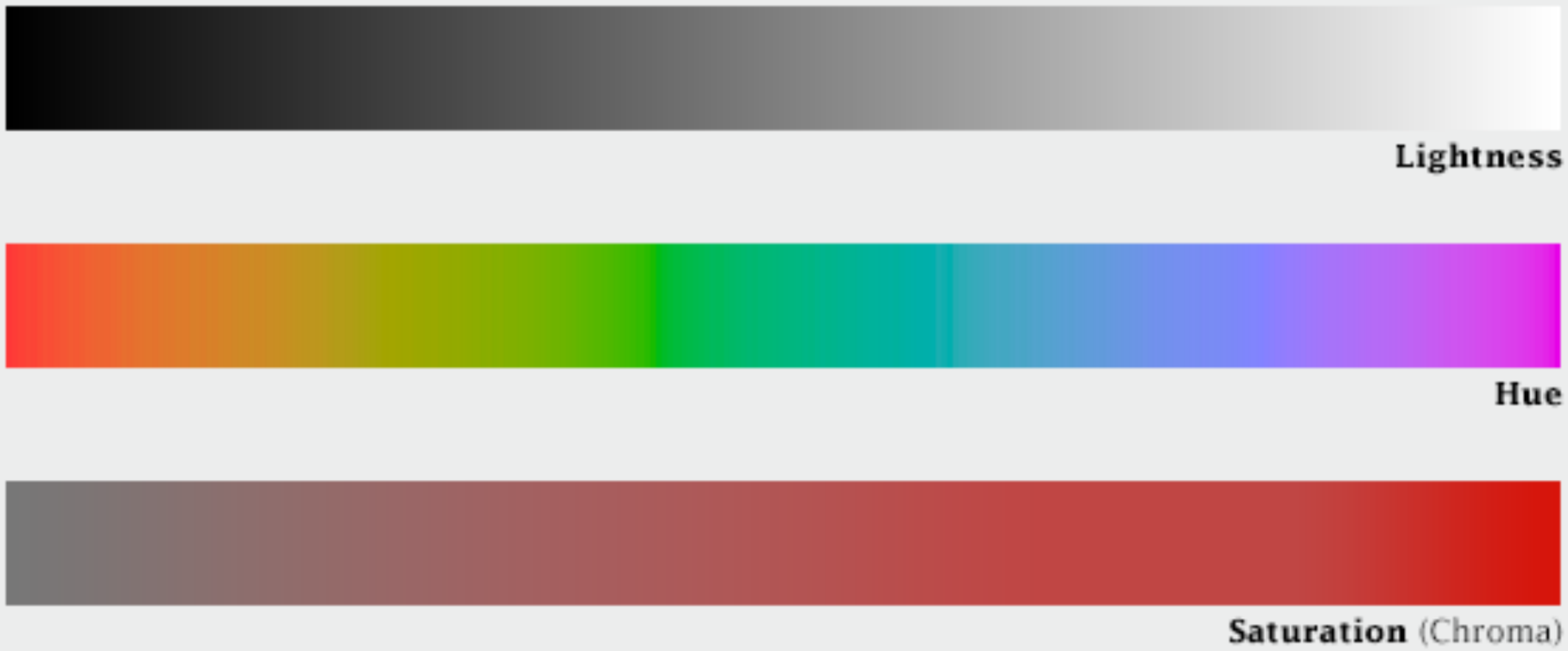




Few, "Show me the Numbers"

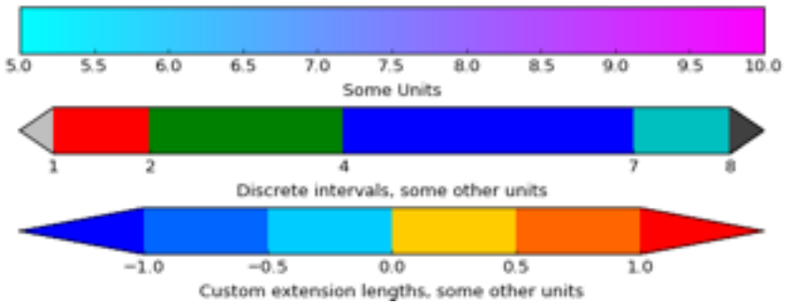
Color

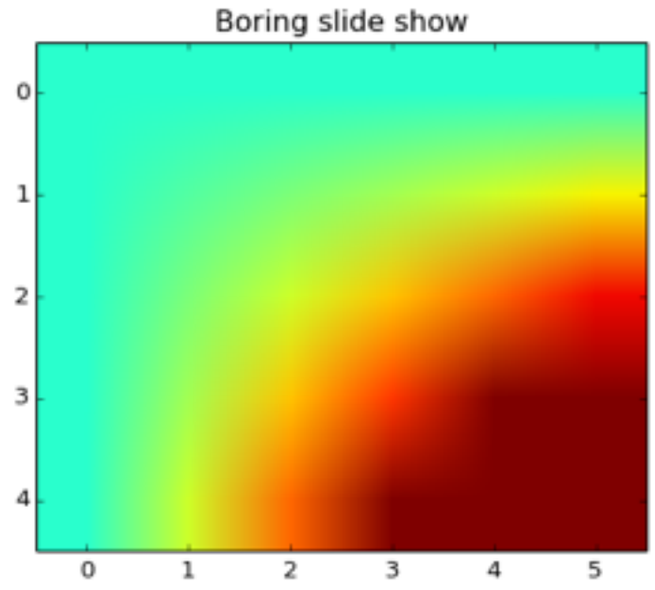
Color Space

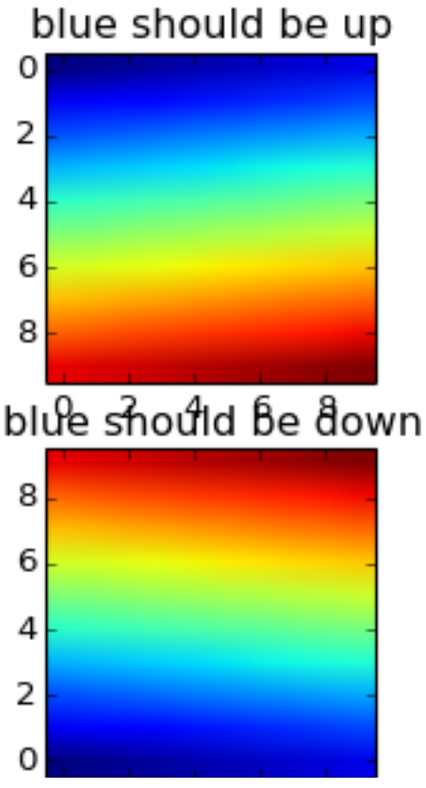


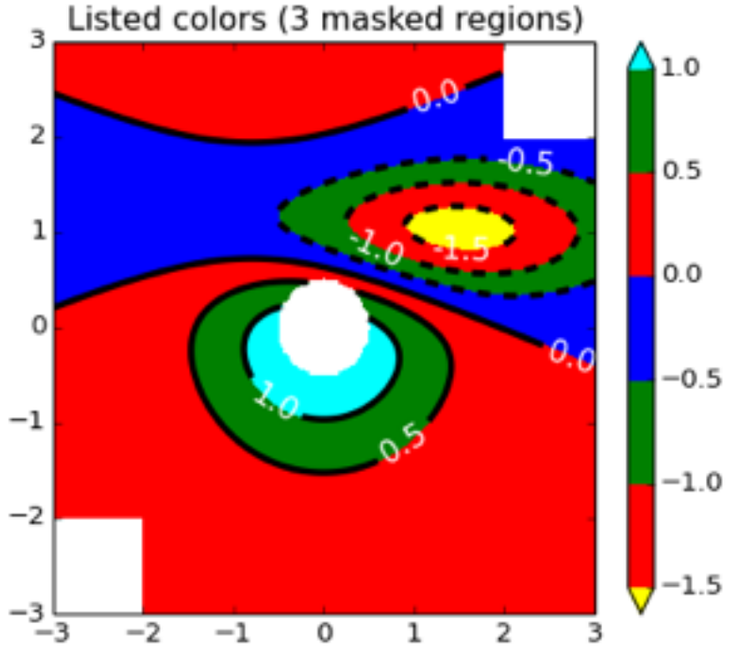
R. Simmon

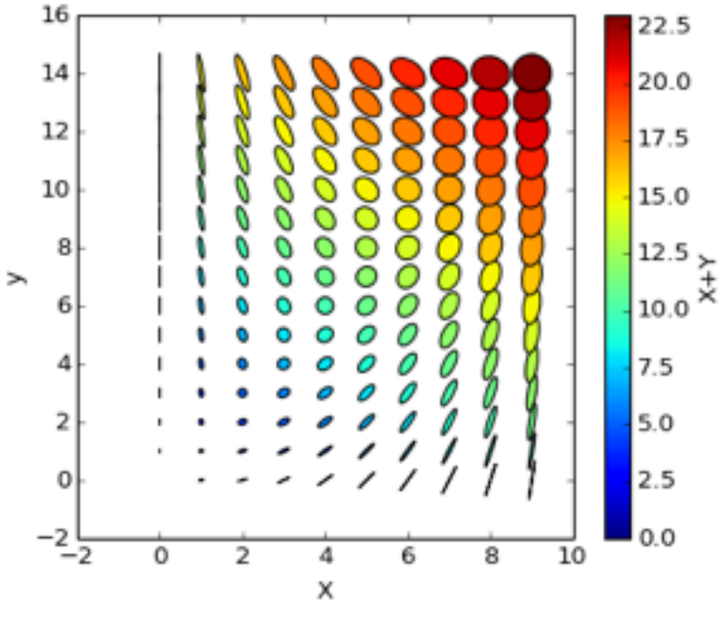
Rainbow Colors

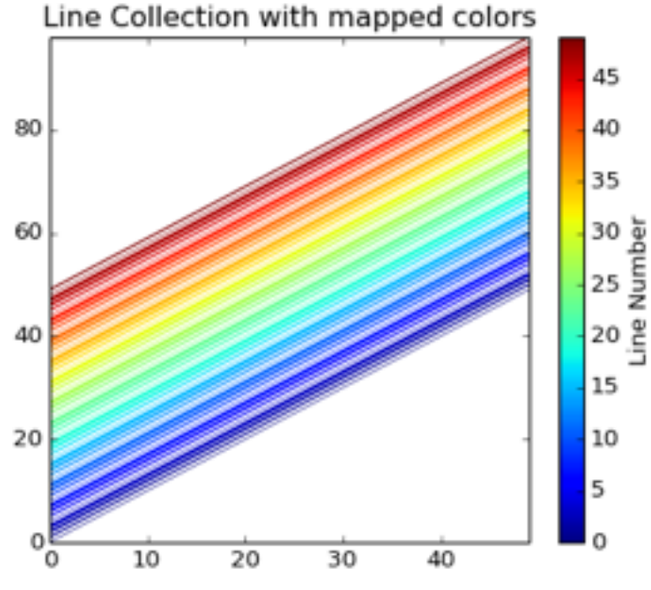








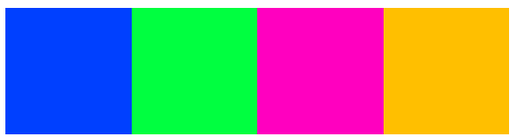


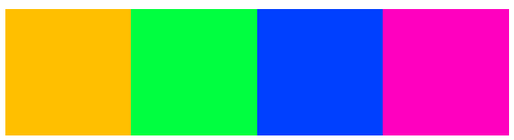


matplotlib gallery

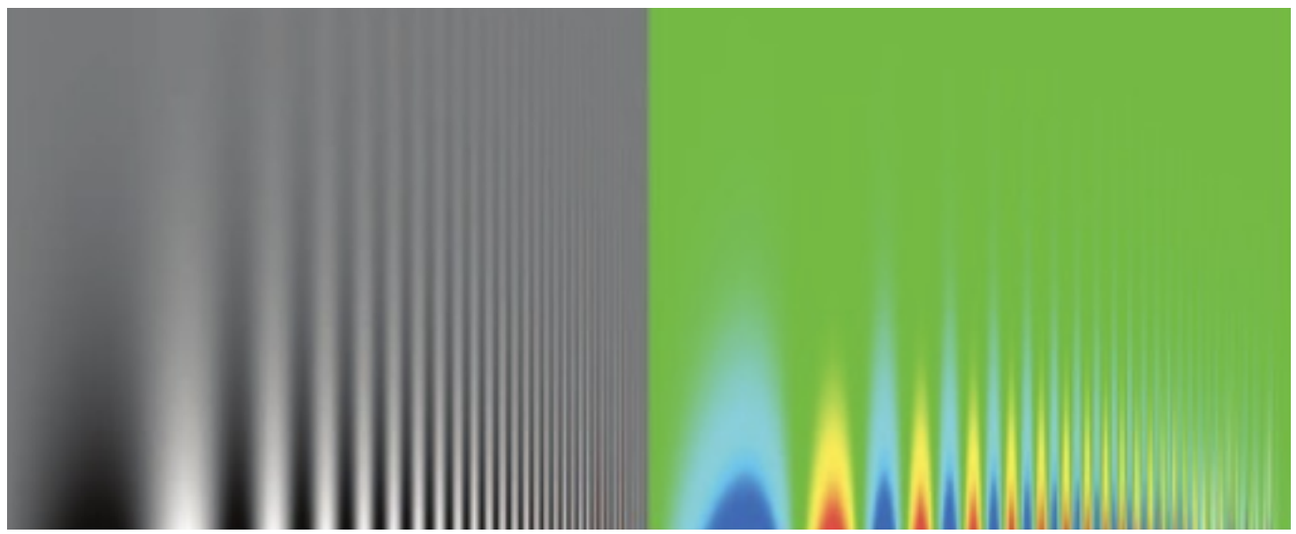
Rainbow Colormap

hard to order



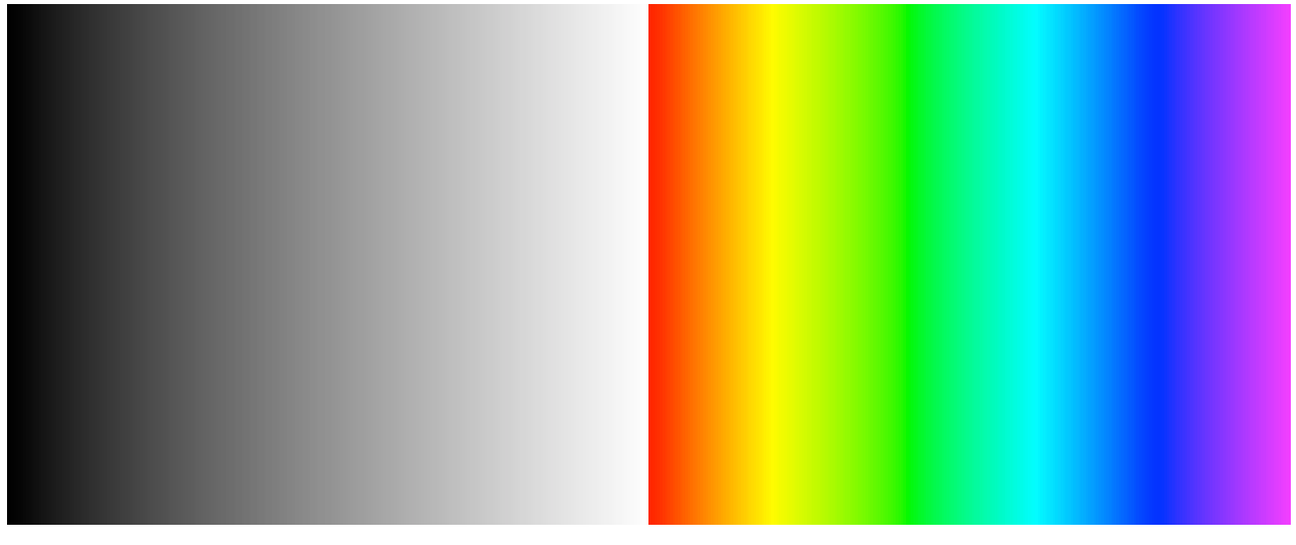


lower resolution



easy to order

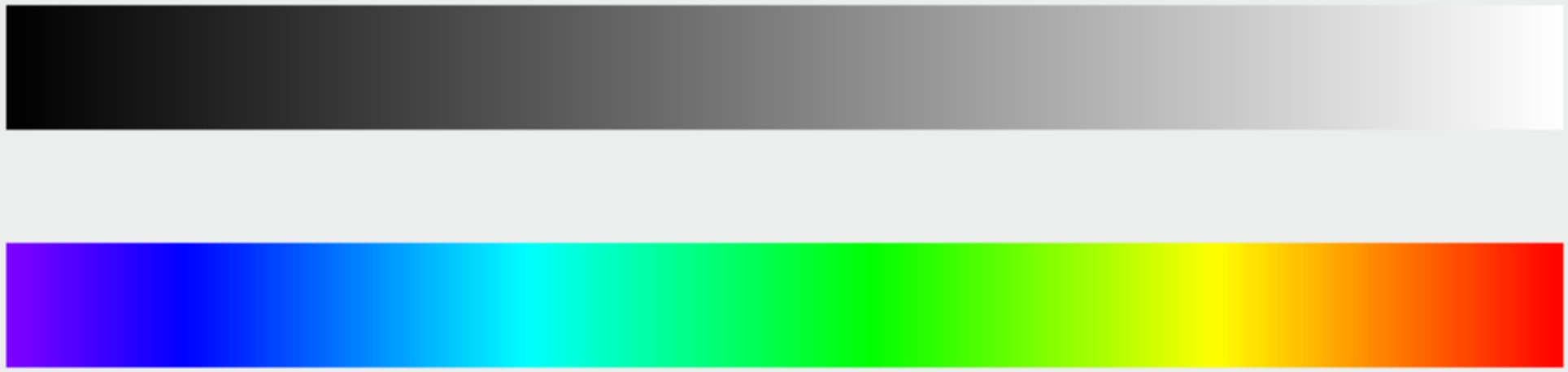
creates artifacts



Borland 2007

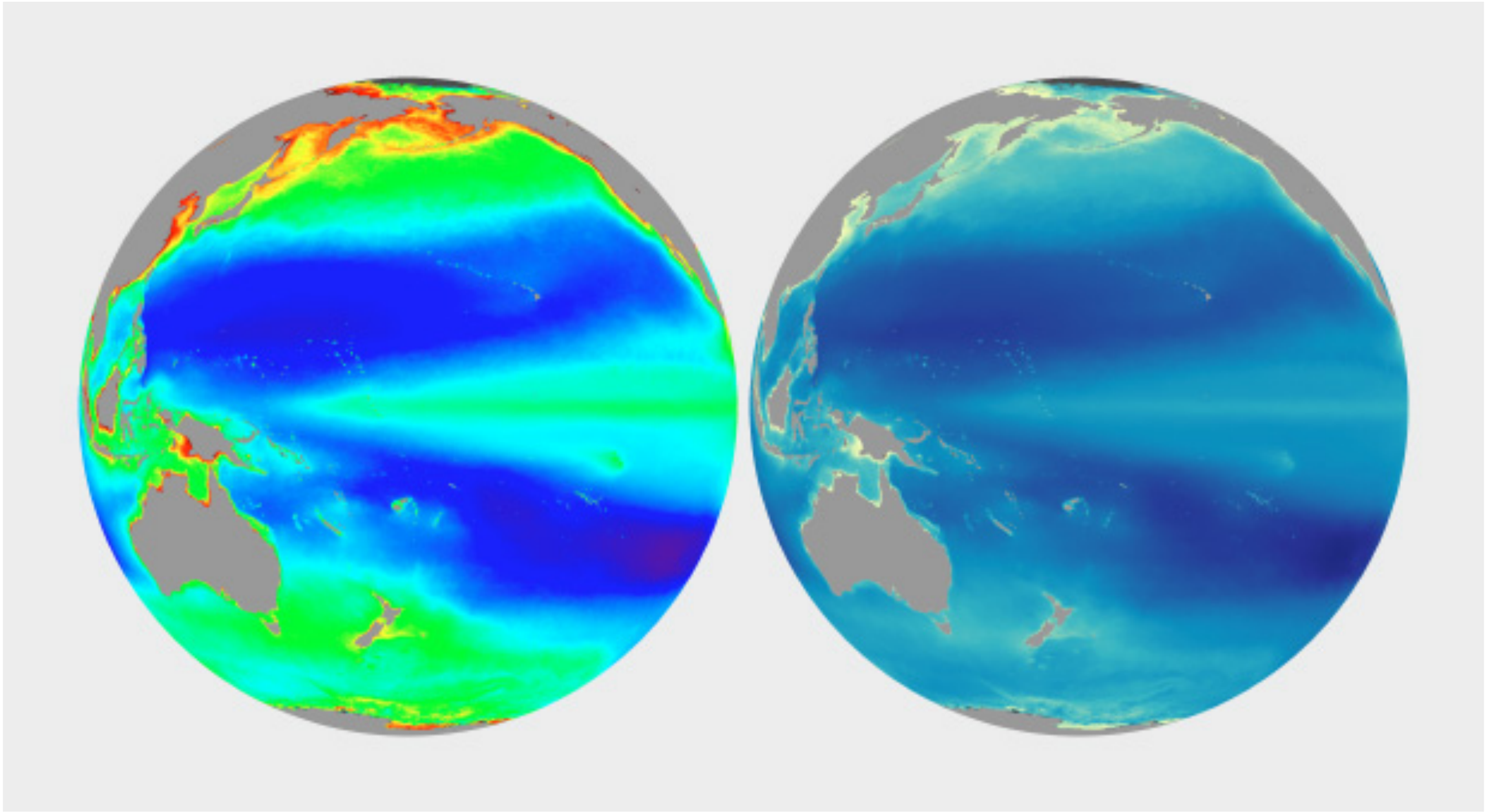
Rainbow Colormap

Rainbow colormap is perceptually nonlinear



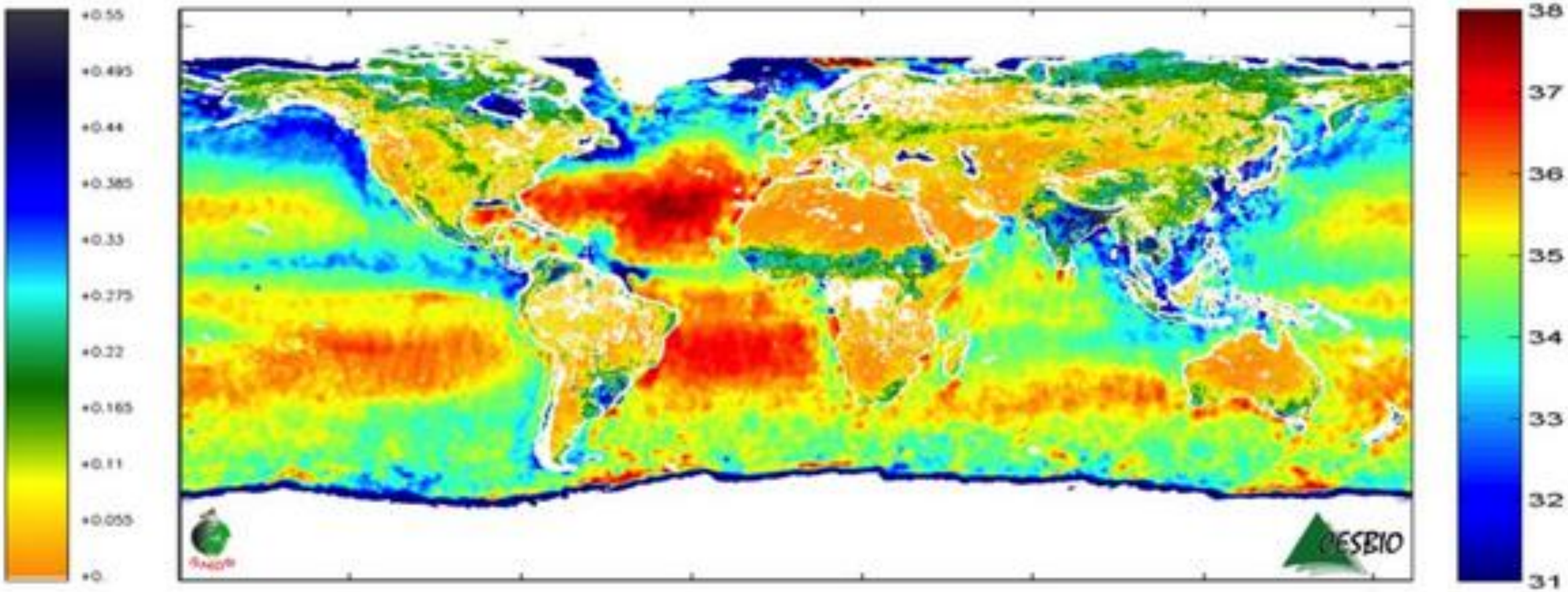
R. Simmon

Rainbow Colormap



R. Simmon

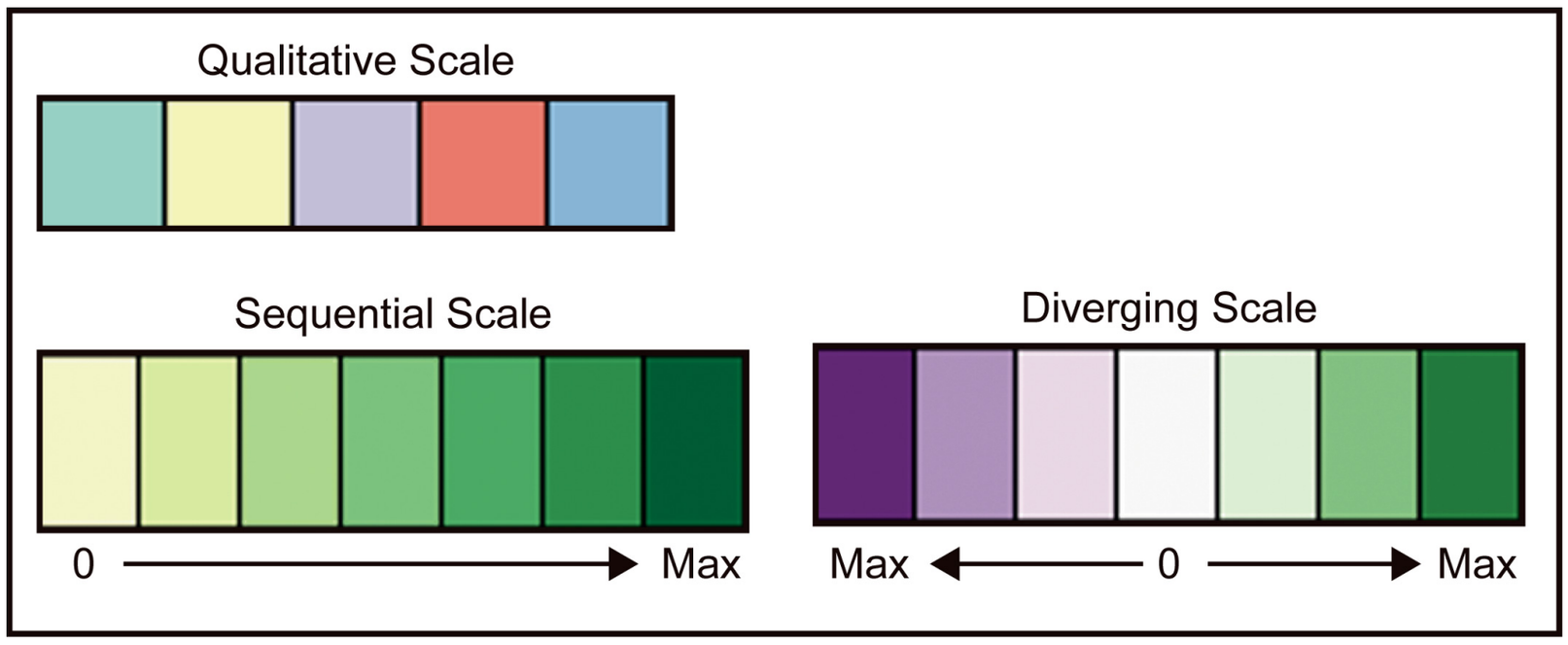
Rainbow Colormap



Brewer Scales

Nominal

Ordinal



Cynthia Brewer, Color Use Guidelines for Data Representation

how to use | updates | credits

COLORBREWER 2.0
color advice for cartography

number of data classes on your map

3 | learn more >

the nature of your data

sequential | learn more >

pick a color scheme: GnBu

multihue | single hue

(optional) only show schemes that are:

- [x] colorblind safe
- [ ] print friendly
- [ ] photocopy-able

learn more >

pick a color system

224, 243, 219
168, 221, 181
67, 162, 202

(•) RGB ( ) CMYK ( ) HEX

adjust map context

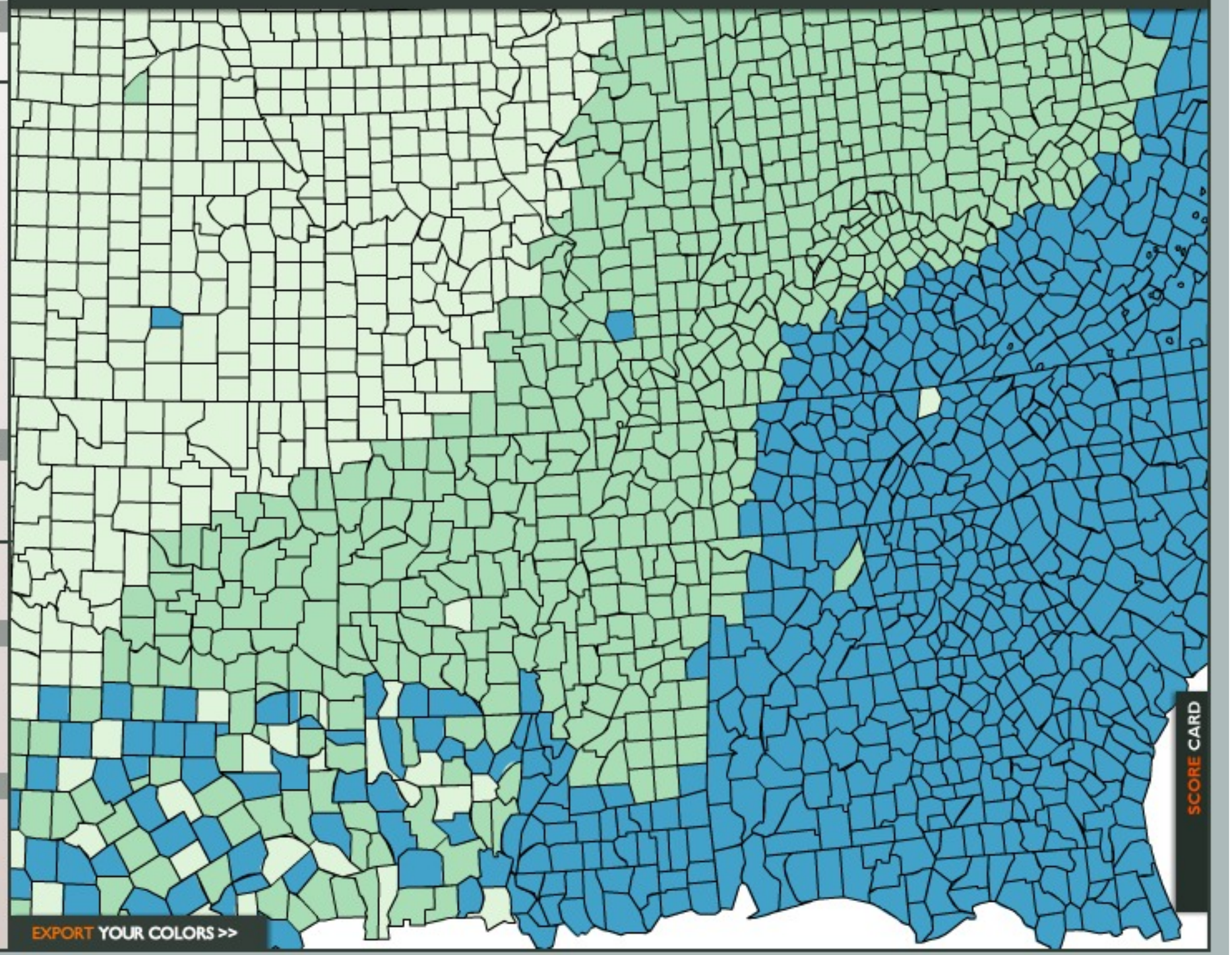
- [ ] roads
- [ ] cities
- [x] borders

select a background

(•) solid color
( ) terrain

color transparency

learn more >

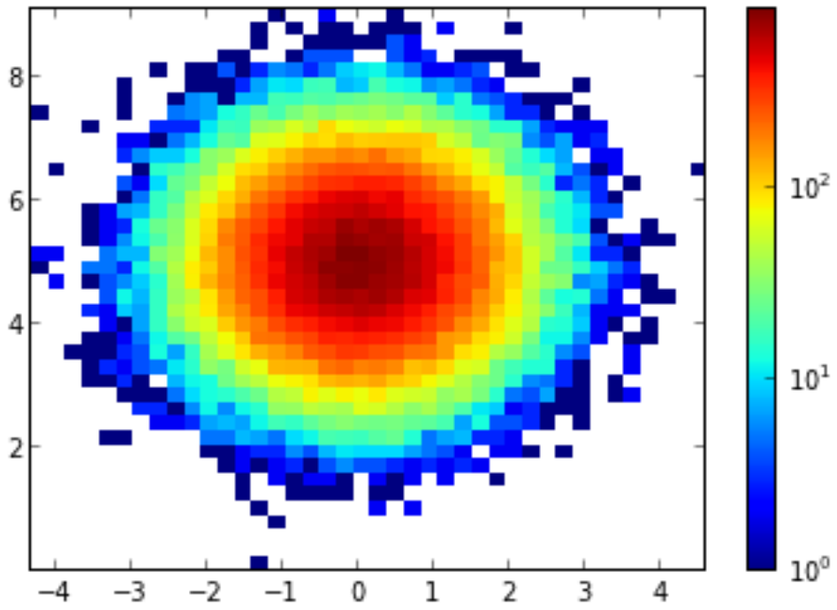


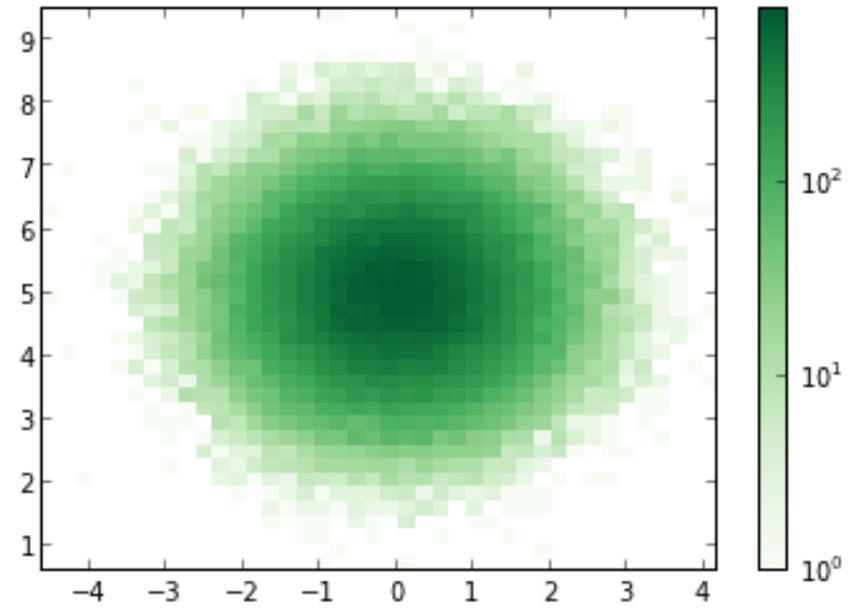
Support
Back to ColorBrewer 1.0

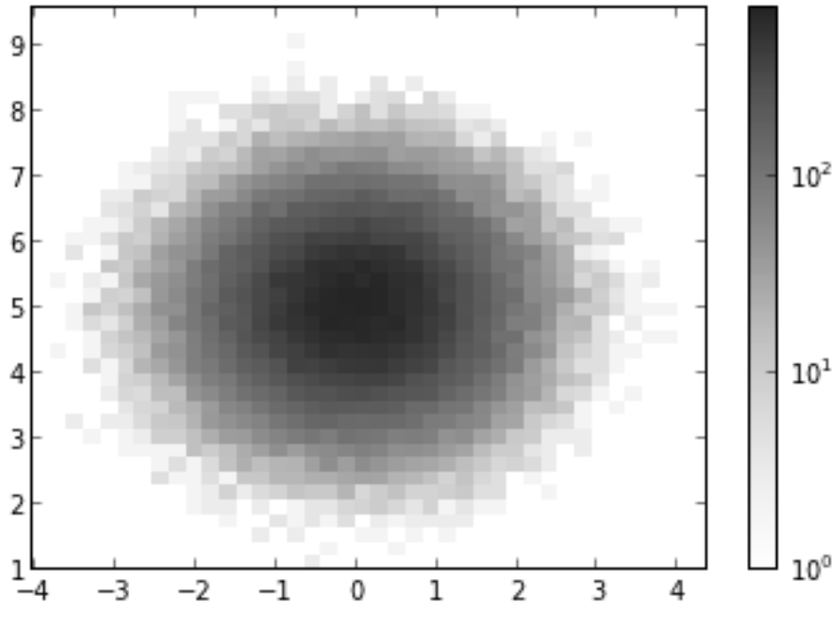
axm

Sequential Brewer Scales

No!



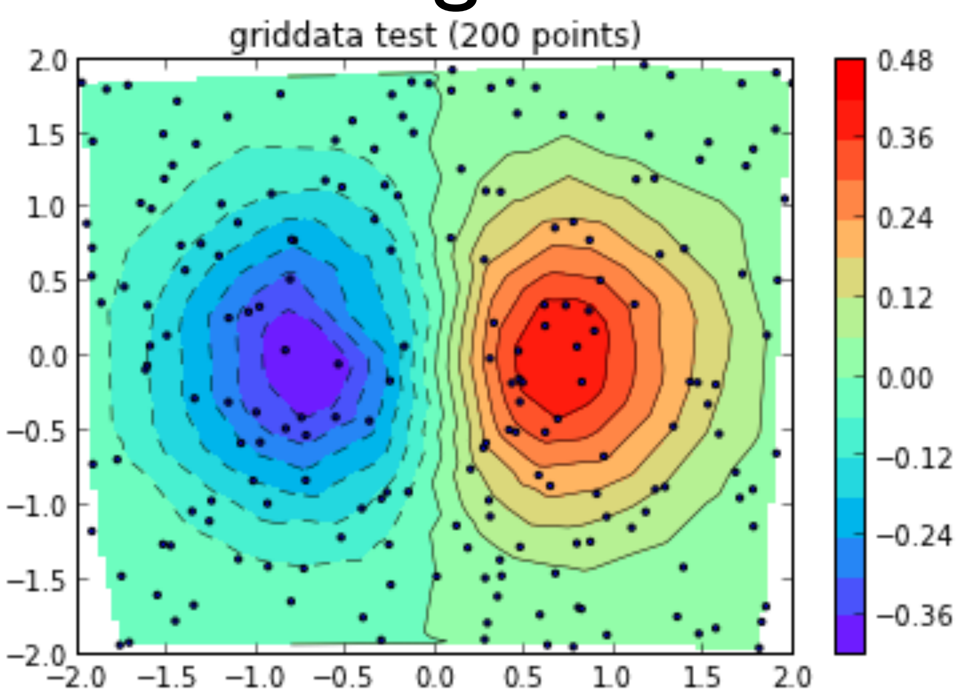


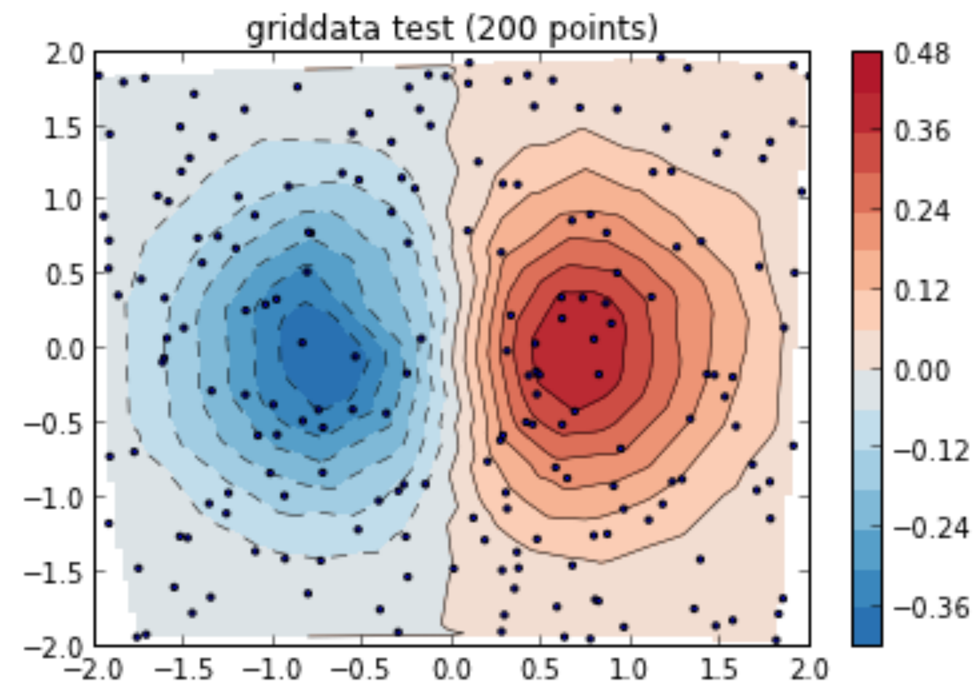


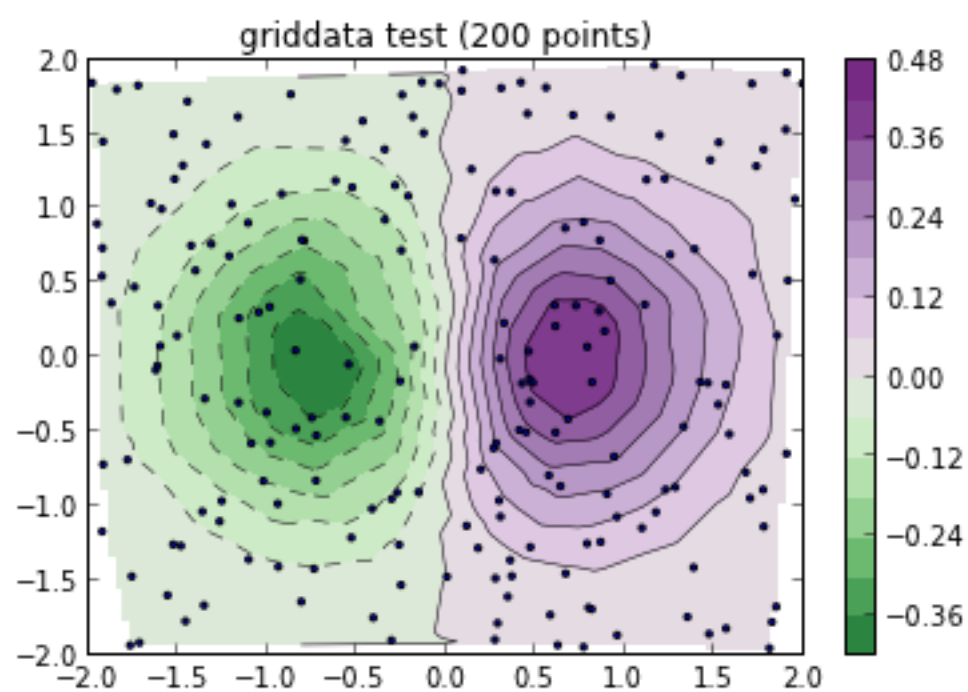
http://nbviewer.ipython.org/5357268

Divergent Brewer Scales

Not great



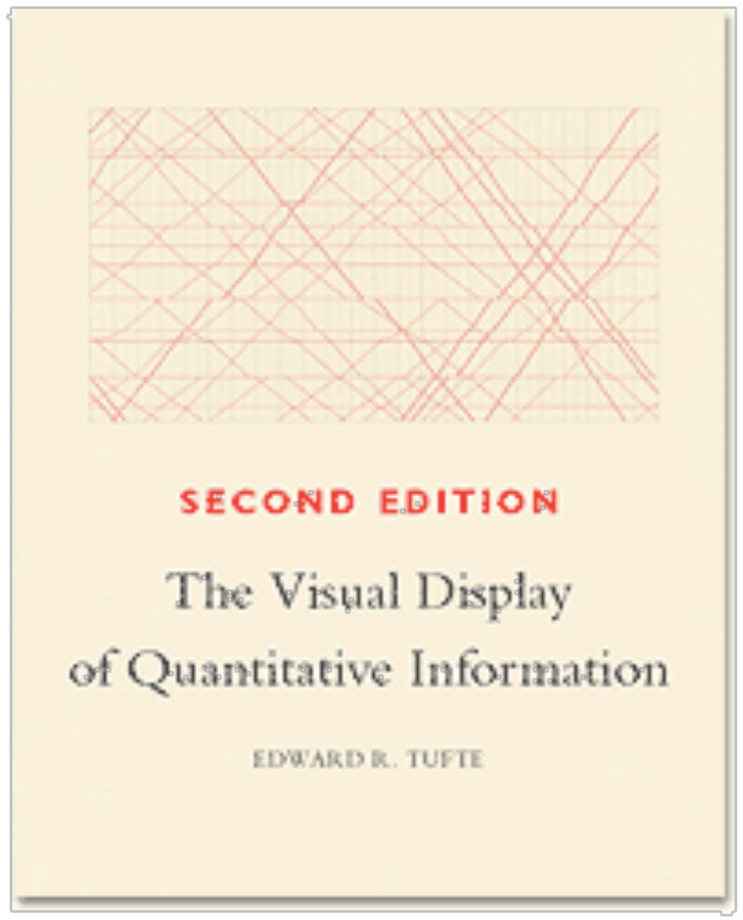


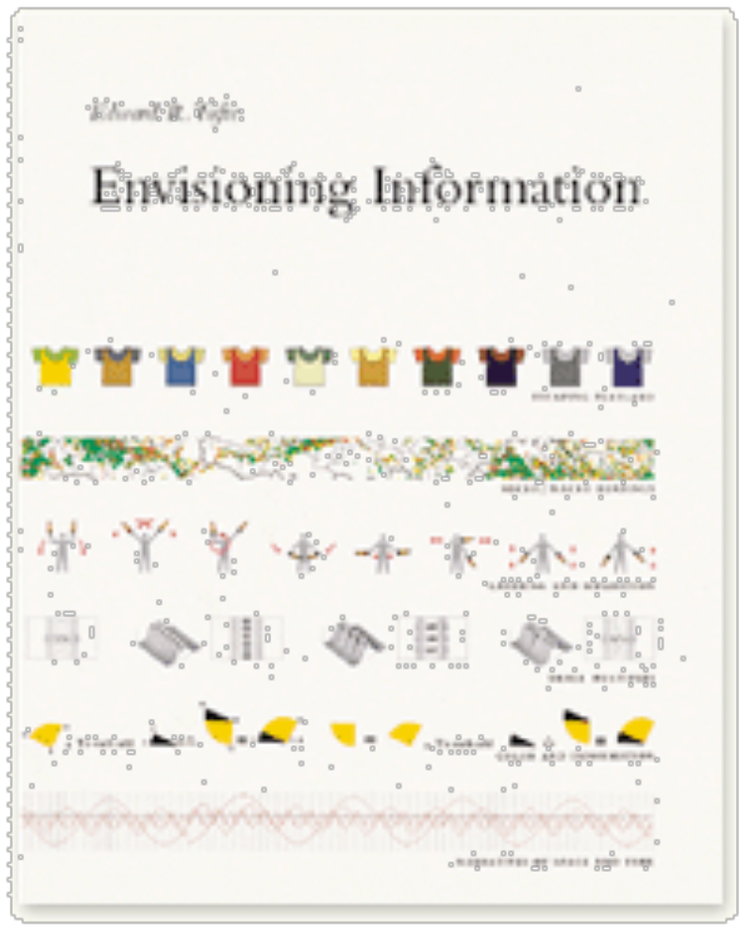


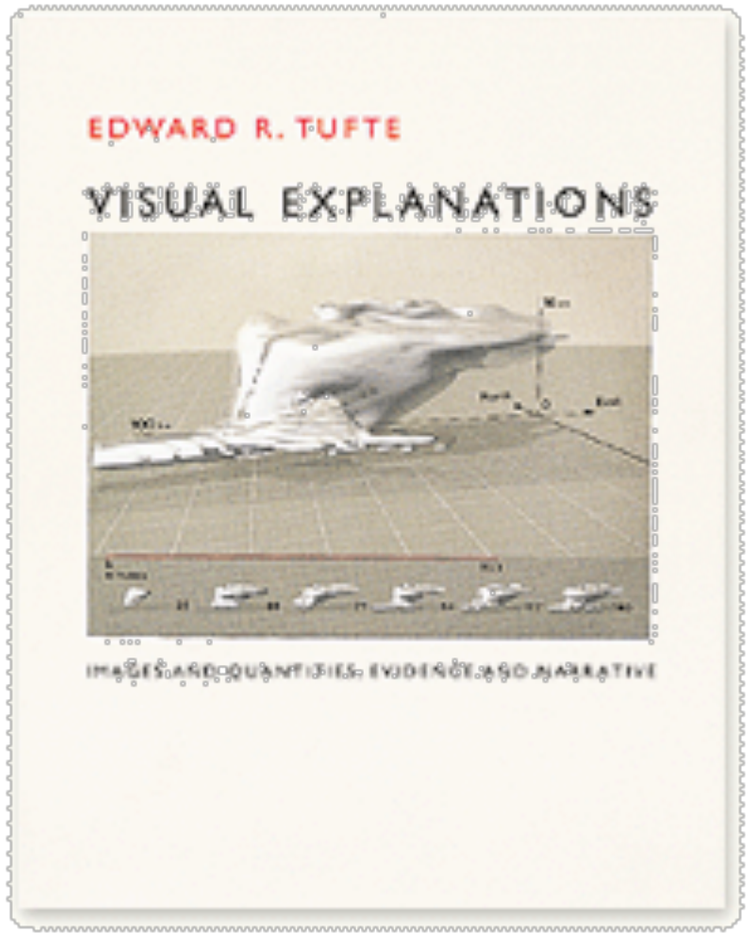
http://nbviewer.ipython.org/5357268

Further Reading

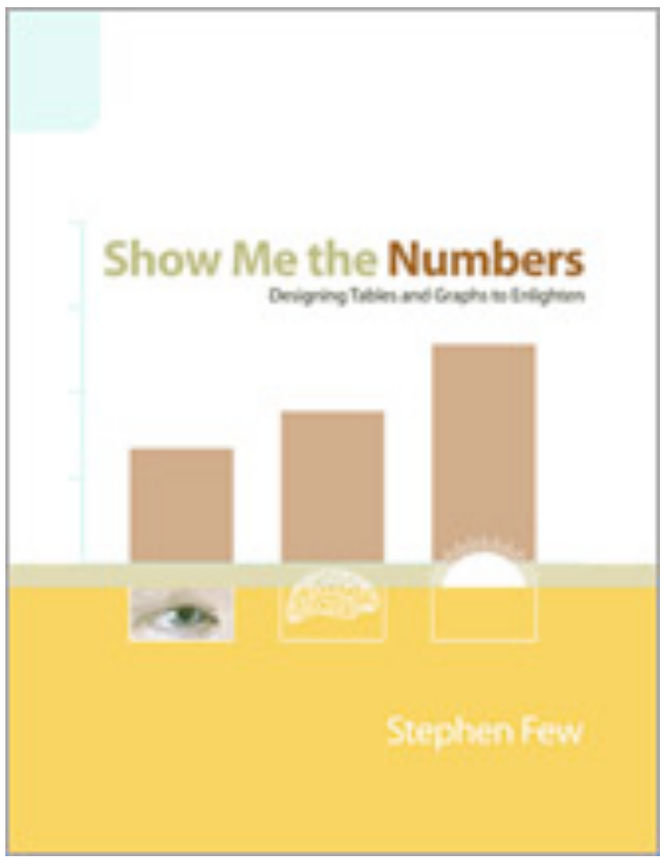
Edward Tufte

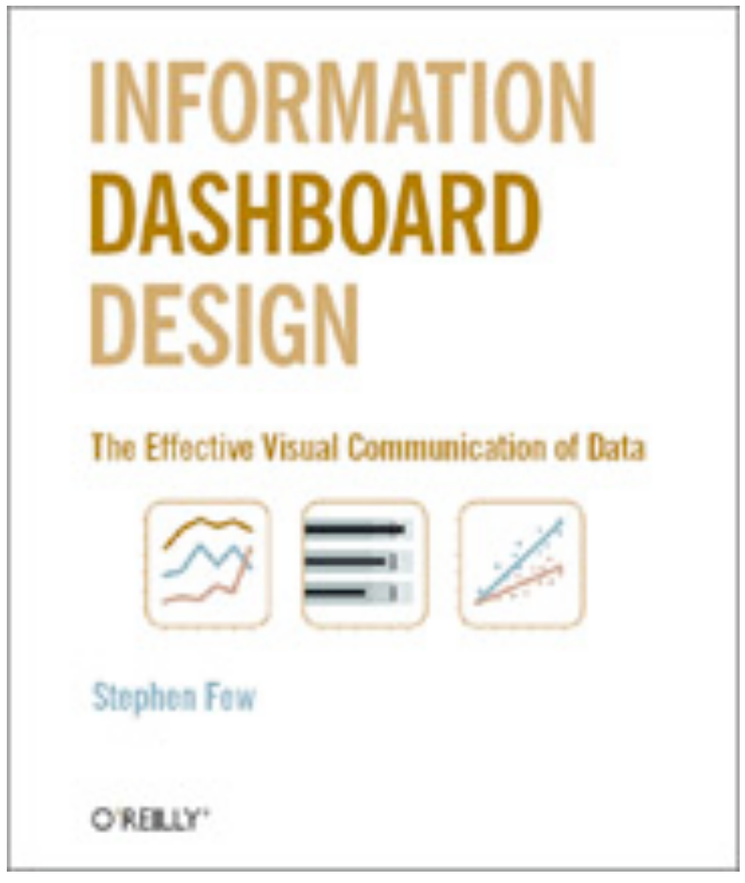






Stephen Few


Show Me the Numbers
Designing Tables and Graphs to Enlighten
Stephen Few


INFORMATION
DASHBOARD
DESIGN
The Effective Visual Communication of Data
Stephen Few
O'REILLY


Now you see it
Stephen Few