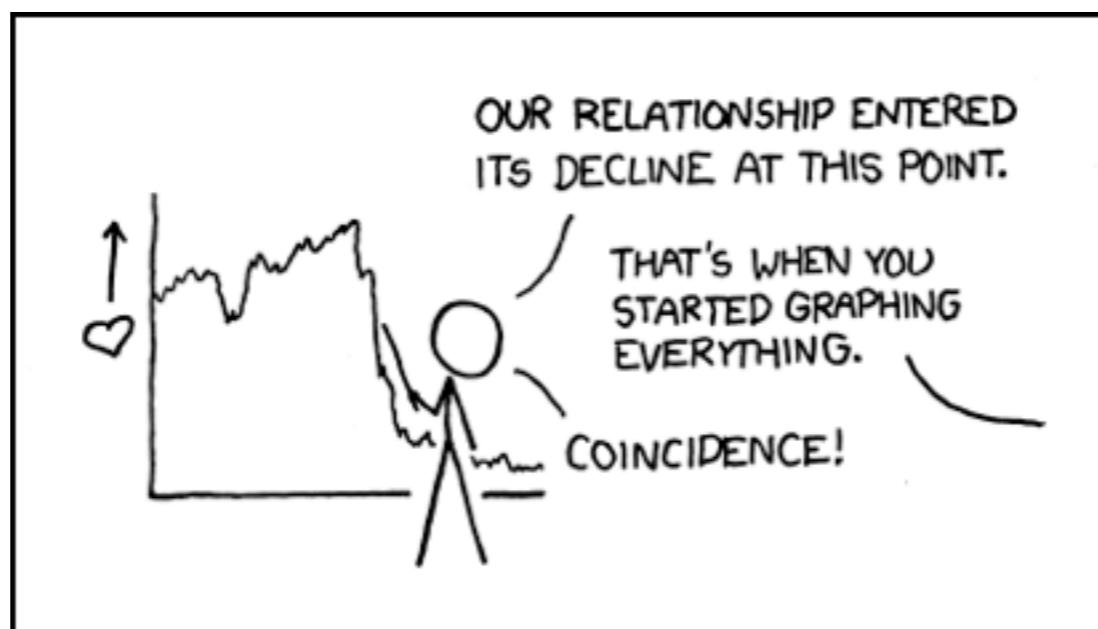


# CS I 09 Data Science

## Statistical Graphs

Hanspeter Pfister & Joe Blitzstein

[pfister@seas.harvard.edu](mailto:pfister@seas.harvard.edu) / [blitzstein@stat.harvard.edu](mailto:blitzstein@stat.harvard.edu)

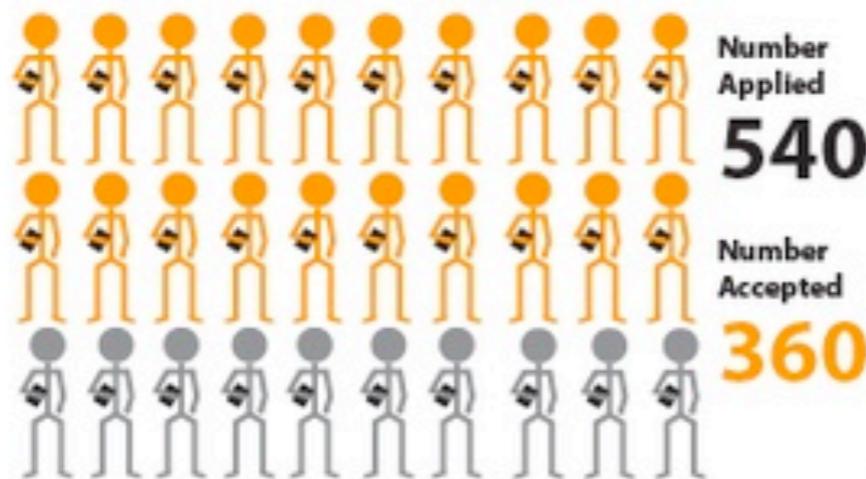


# This Week

- 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

## Course Lottery Results

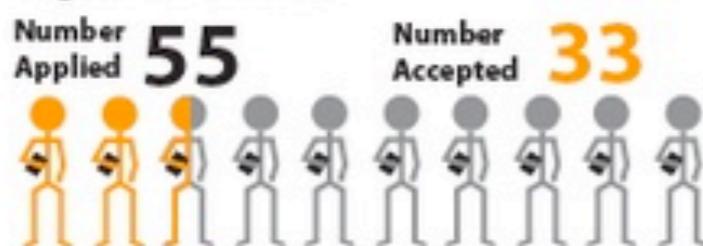
SPU 27: "Science and Cooking"



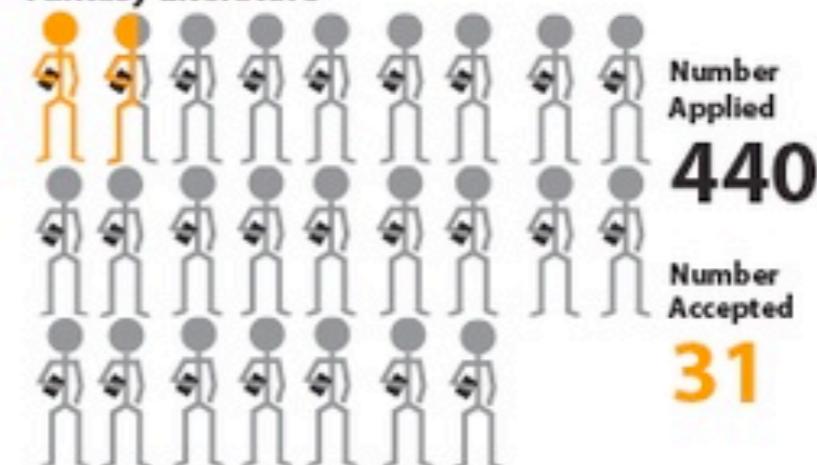
WGS 1424: "American Fetish"



English 43: "Arrivals"



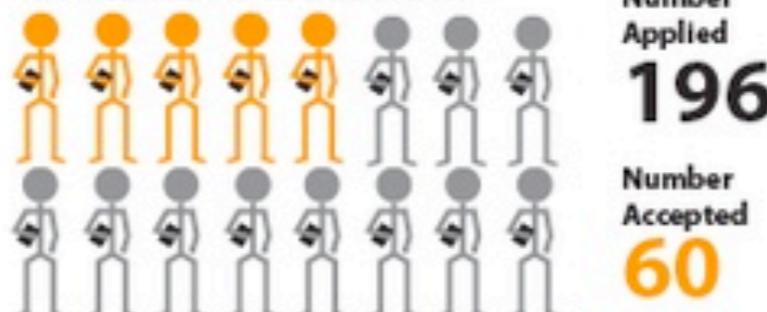
Folk & Myth 128: "Fairy Tale, Myth, and Fantasy Literature"



SOC 105: "Sports and Society"



USW 35: "Dilemmas of Equity and Excellence in American K-12 Education"



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

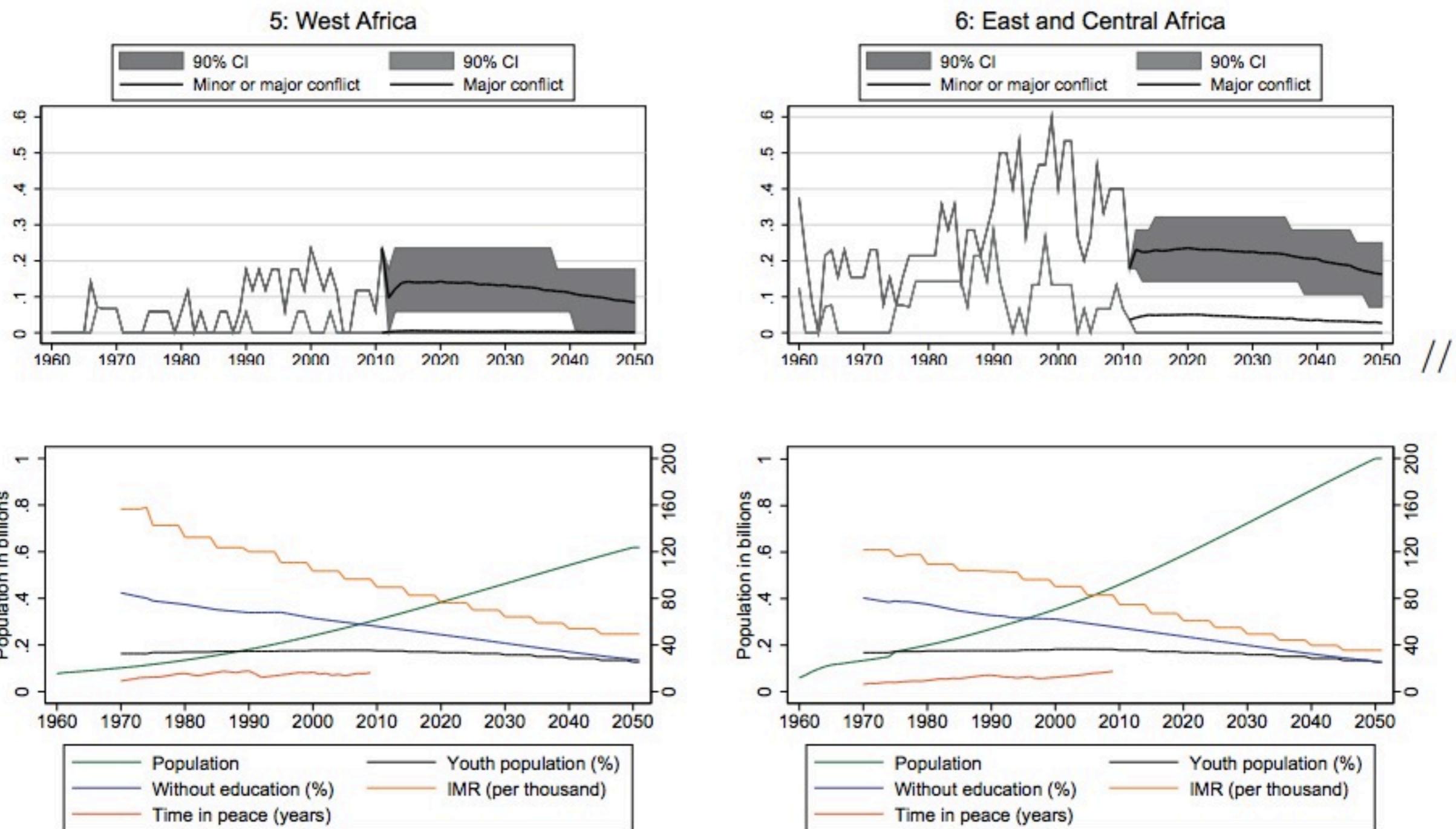


Figure 5. Predicted share of countries in conflict and average predictor values, West Africa (left) and East and Central Africa (right), 1960–2050


**JOURNAL TOOLS**

- [Get New Content Alerts](#)
- [Get RSS feed](#)
- [Save to My Profile](#)
- [Get Sample Copy](#)
- [Recommend to Your Librarian](#)

**JOURNAL MENU**
[Journal Home](#)
**FIND ISSUES**

- [Current Issue](#)
- [All Issues](#)
- [Virtual Issues](#)

**FIND ARTICLES**

- [Most Accessed](#)
- [Most Cited](#)

**GET ACCESS**
[Subscribe / Renew](#)
**FOR CONTRIBUTORS**

- [OnlineOpen](#)
- [Author Guidelines](#)

**ABOUT THIS JOURNAL**

- [Society Information](#)
- [News](#)
- [Overview](#)
- [Editorial Board](#)
- [Permissions](#)
- [Advertise](#)

# significance

statistics making sense



## Peace on Earth?: The future of internal armed conflict



Håvard Hegre

Issue



Article first published online: 15 FEB 2013

DOI: [10.1111/j.1740-9713.2013.00628.x](https://doi.org/10.1111/j.1740-9713.2013.00628.x)

© 2013 The Royal Statistical Society

**Significance**
**Volume 10, Issue 1, pages 4–8, February 2013**
**SEARCH**
[In this issue](#)

[Advanced >](#) [Saved Searches >](#)
**ARTICLE TOOLS**

- [Get PDF \(752K\)](#)
- [Save to My Profile](#)
- [E-mail Link to this Article](#)
- [Export Citation for this Article](#)
- [Get Citation Alerts](#)
- [Request Permissions](#)

Share

Get PDF (752K)

**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.

Get PDF (752K)

[More content like this](#)

« [Job openings at American University](#)

[False memories and statistical analysis](#) »

## 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.

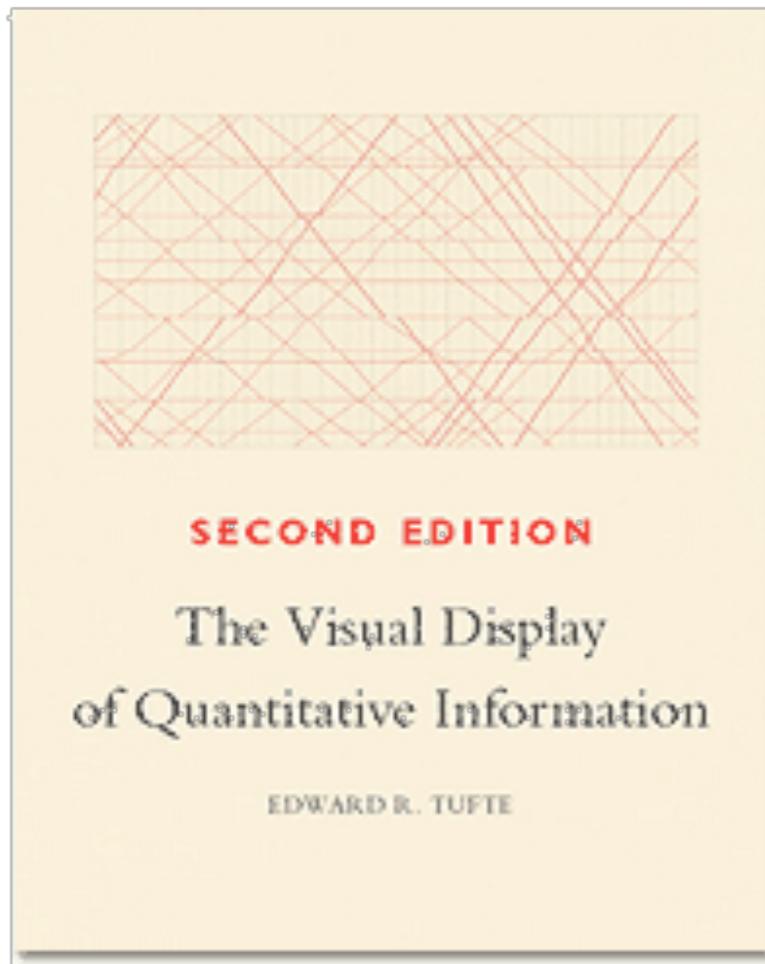


Filed under [Sociology](#), [Statistical graphics](#)

[Comment \(RSS\)](#) | [Trackback](#) | [Permalink](#)

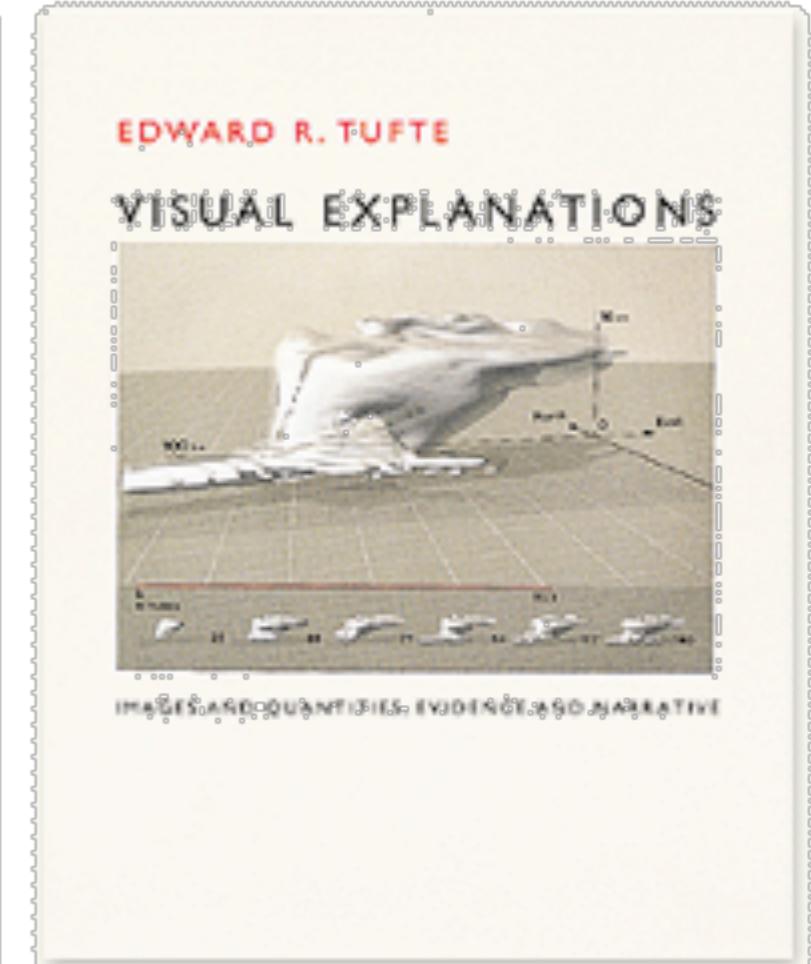
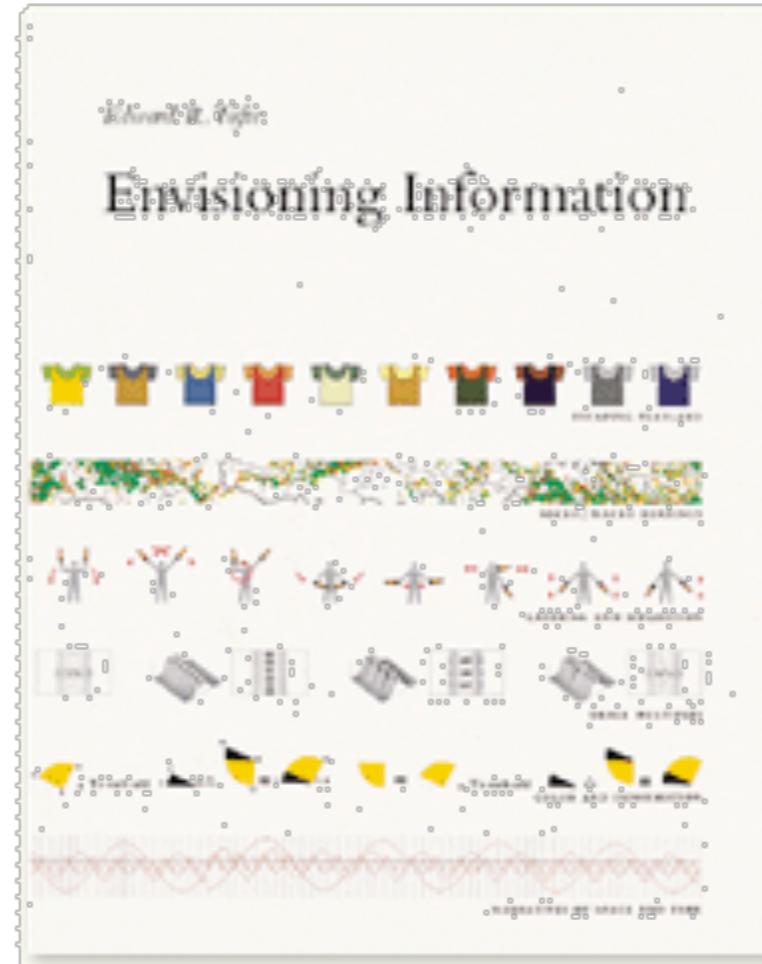
# Design Principles

# Edward Tufte



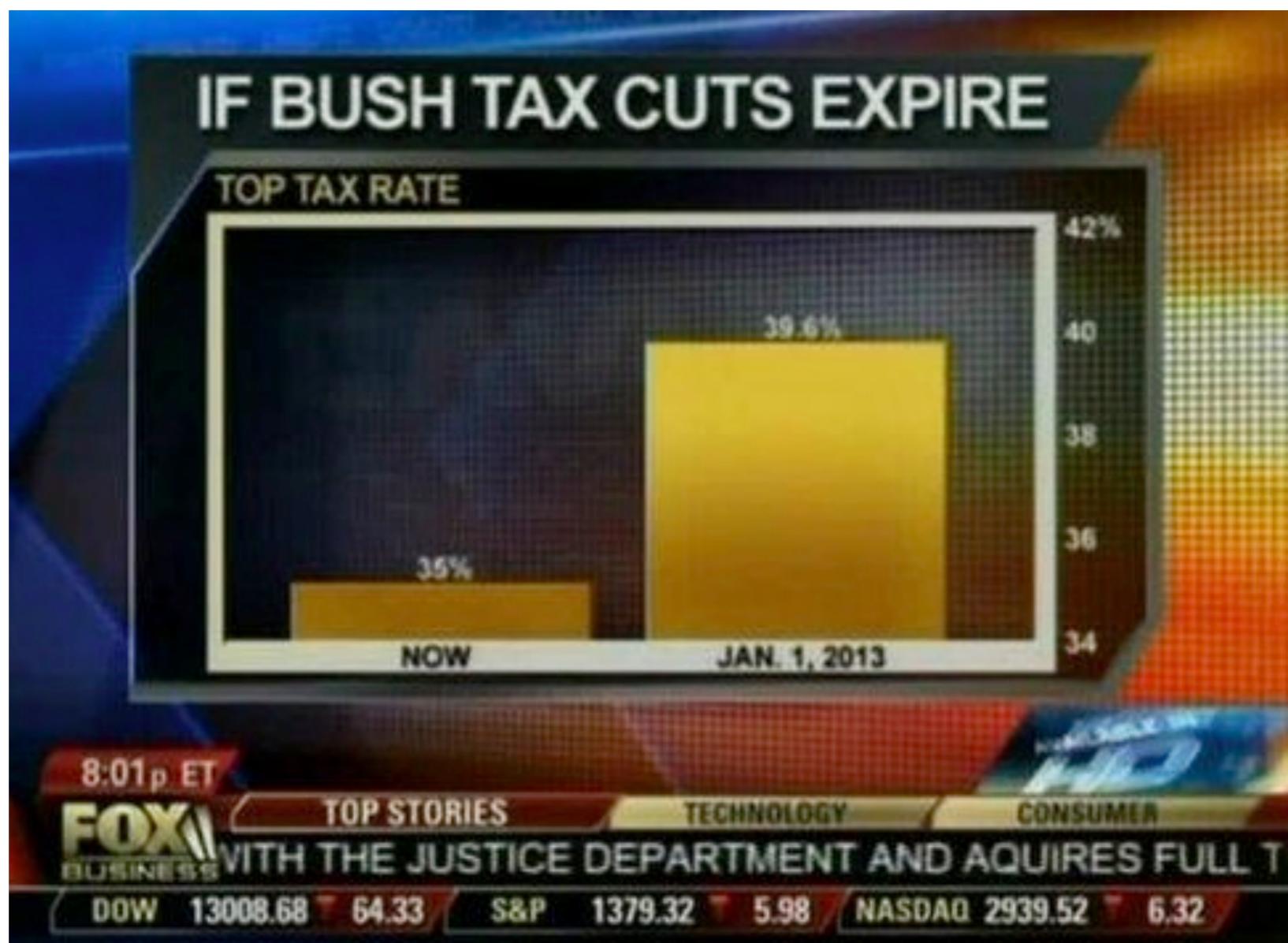
**SECOND EDITION**  
The Visual Display  
of Quantitative Information

EDWARD R. TUFTE

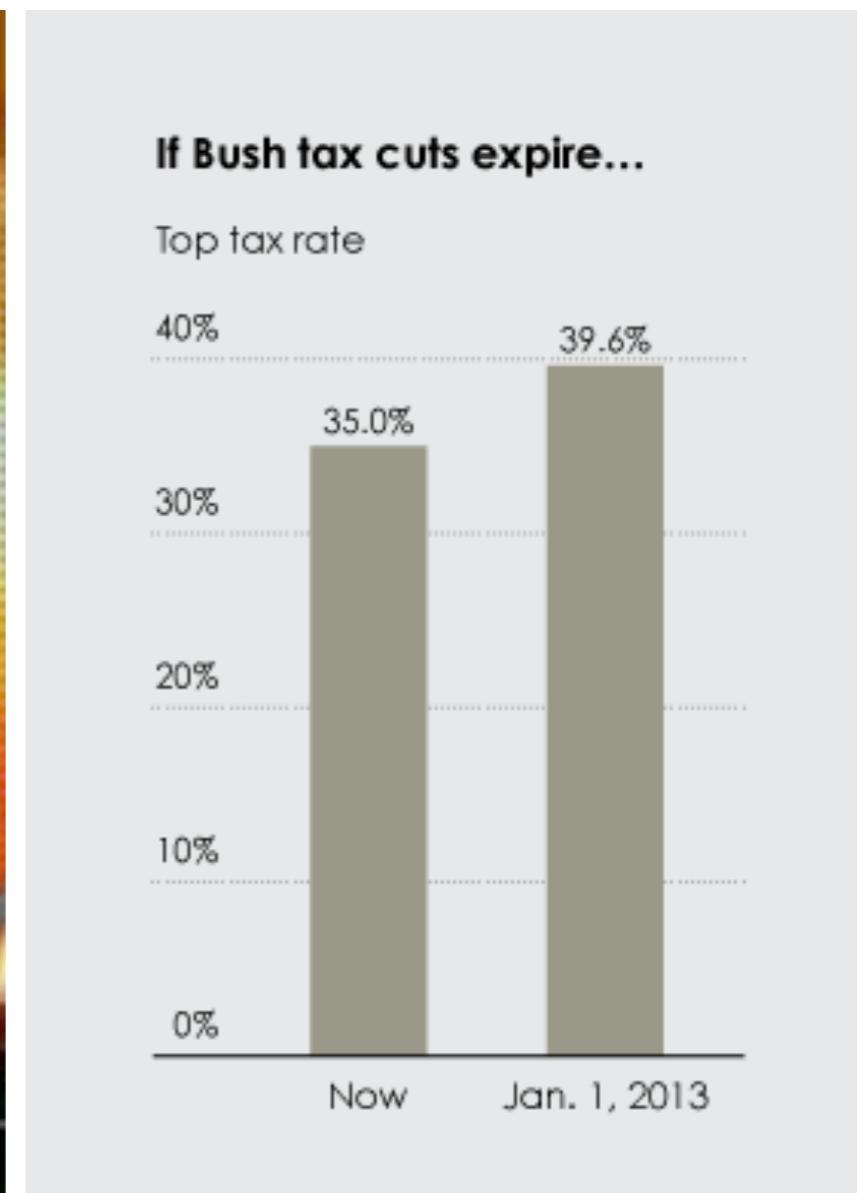
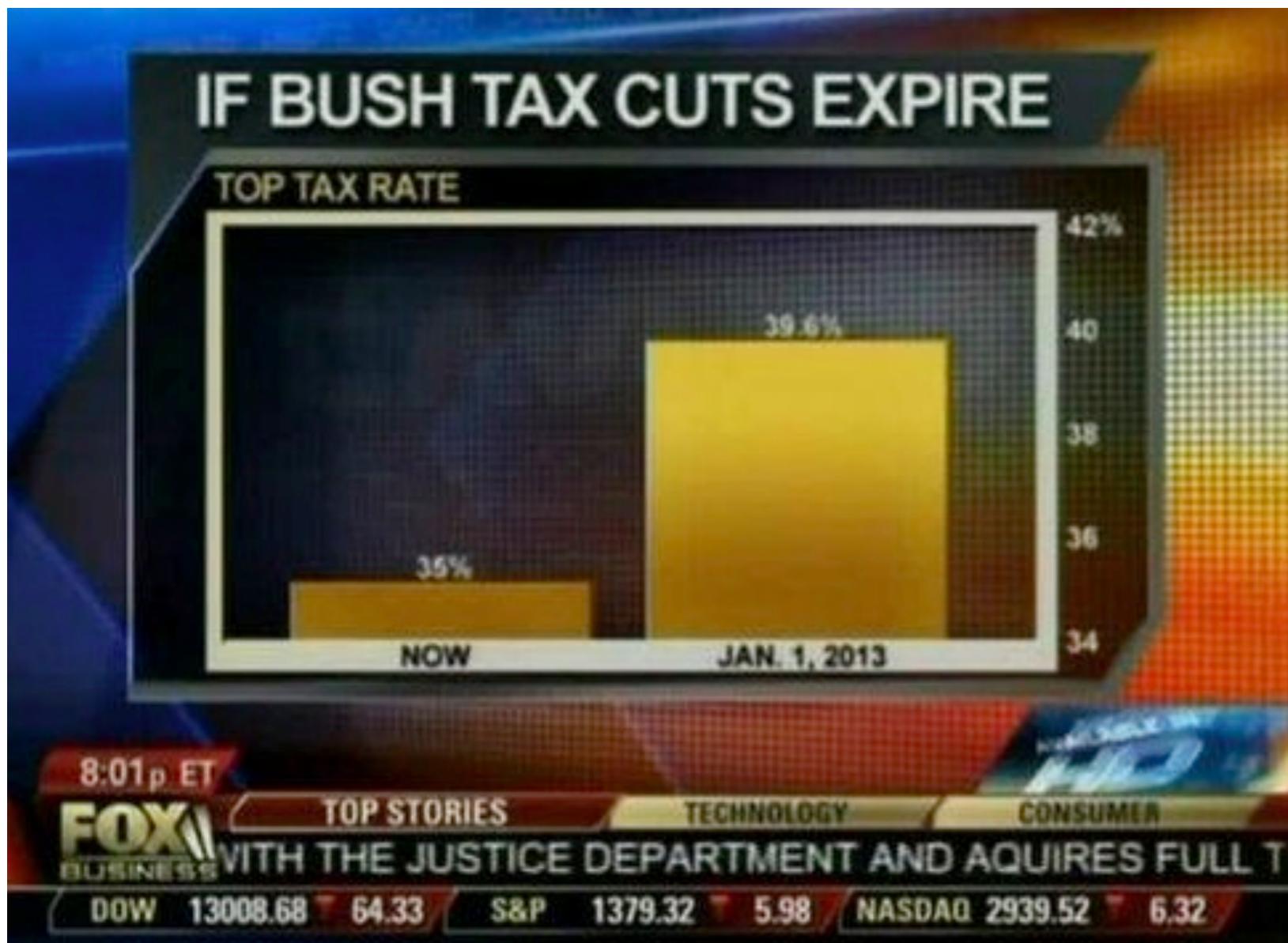


EDWARD R. TUFTE  
**VISUAL EXPLANATIONS**

# Graphical Integrity



# Scale Distortions



# Scale Distortions

## How 2012 STACKS UP

THE WARMEST YEARS ON RECORD  
CONTIGUOUS U.S.

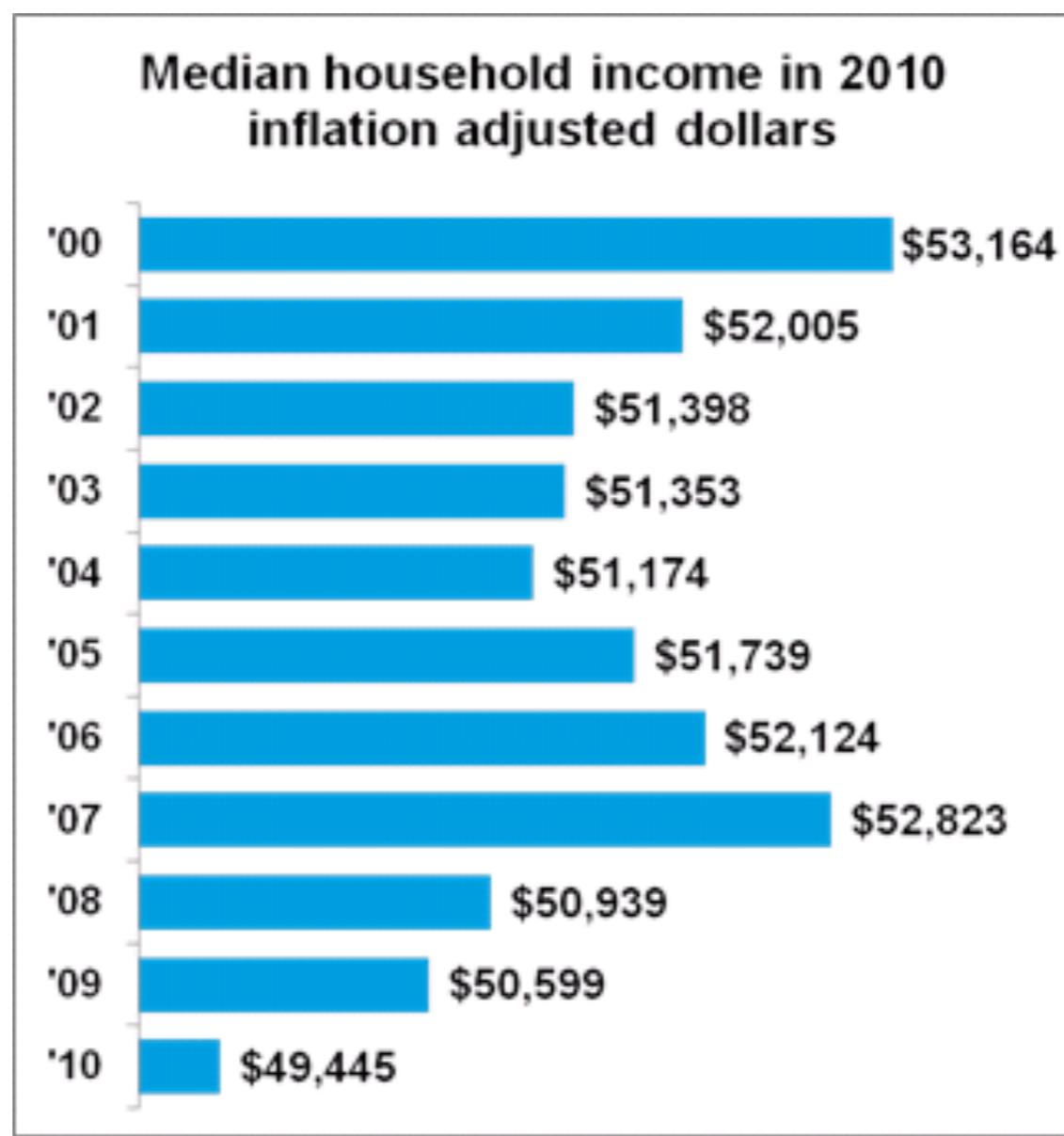


Source: NOAA's National Climatic Data Center - State of the Climate National Overview

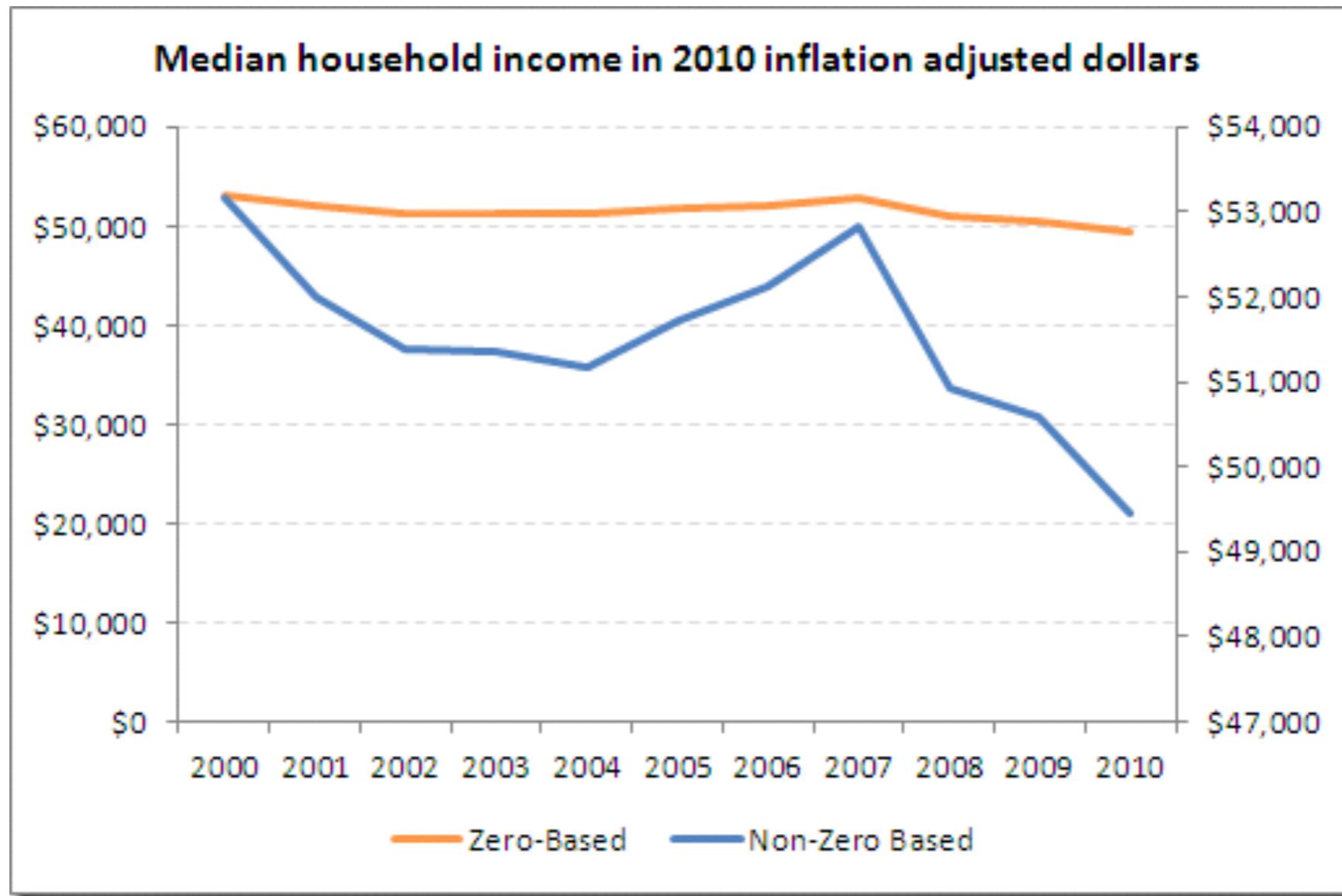
CLIMATE  CENTRAL

# Scale Distortions

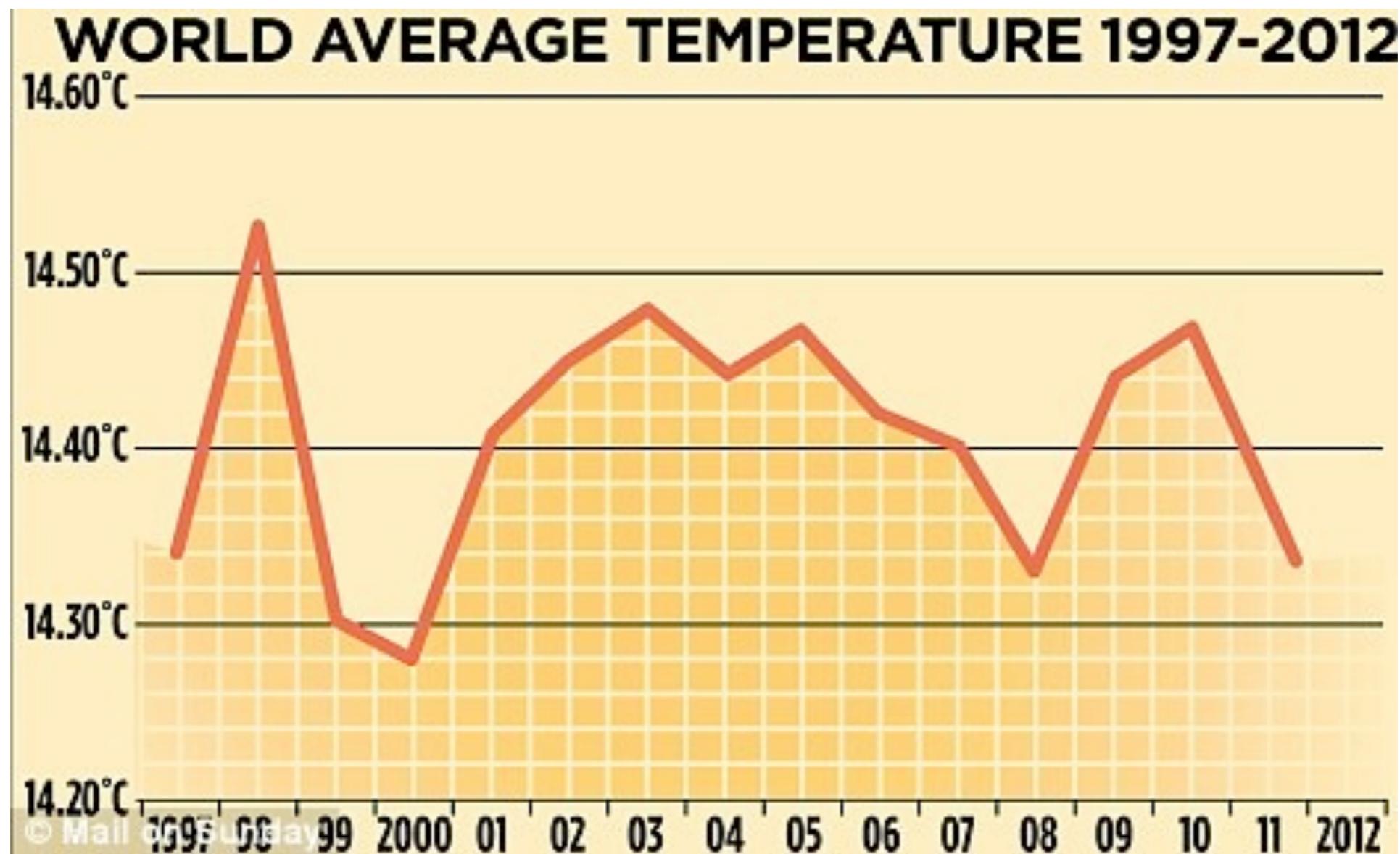
*Always start your bar graphs at zero!*



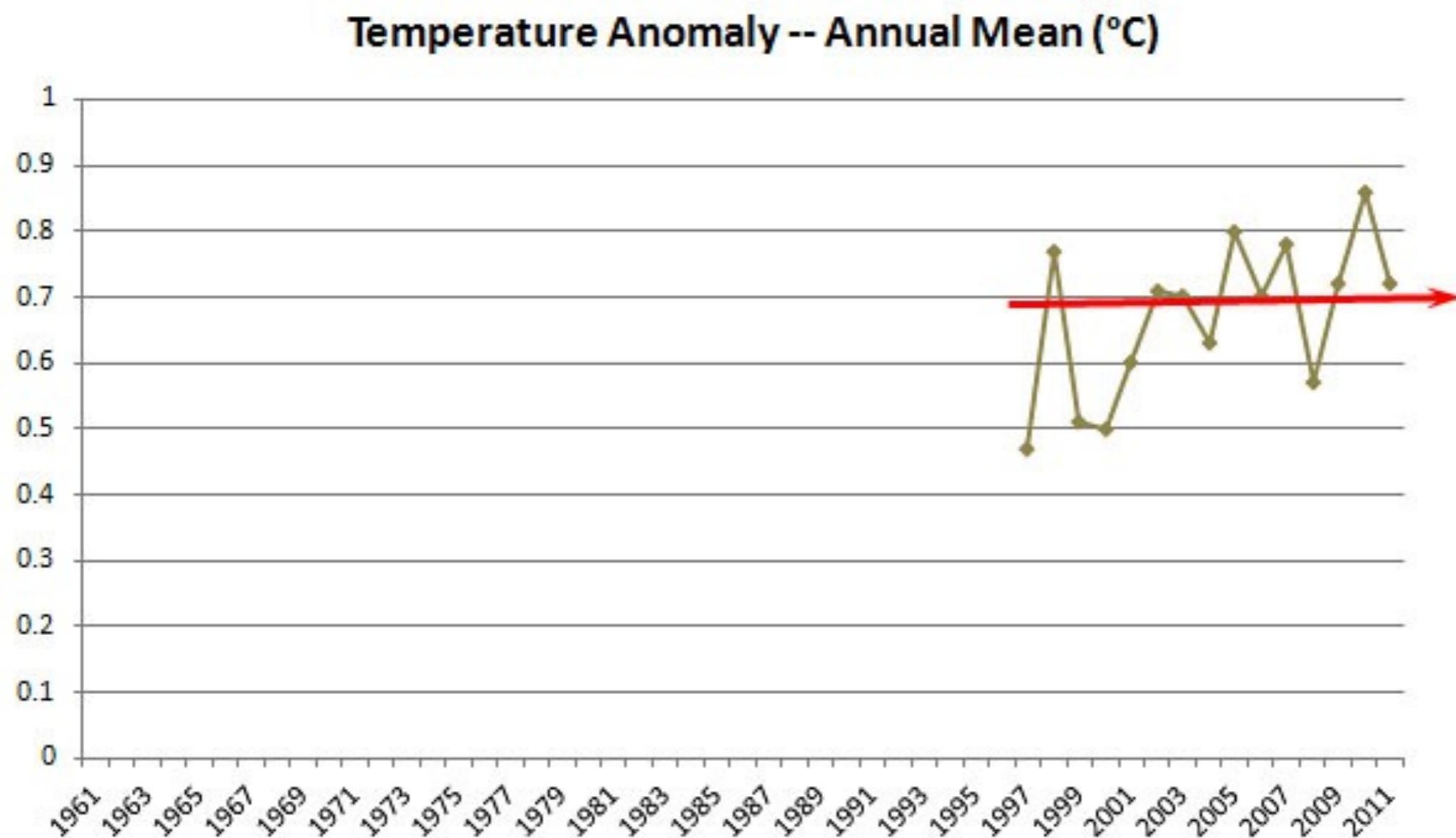
# Scale Distortions



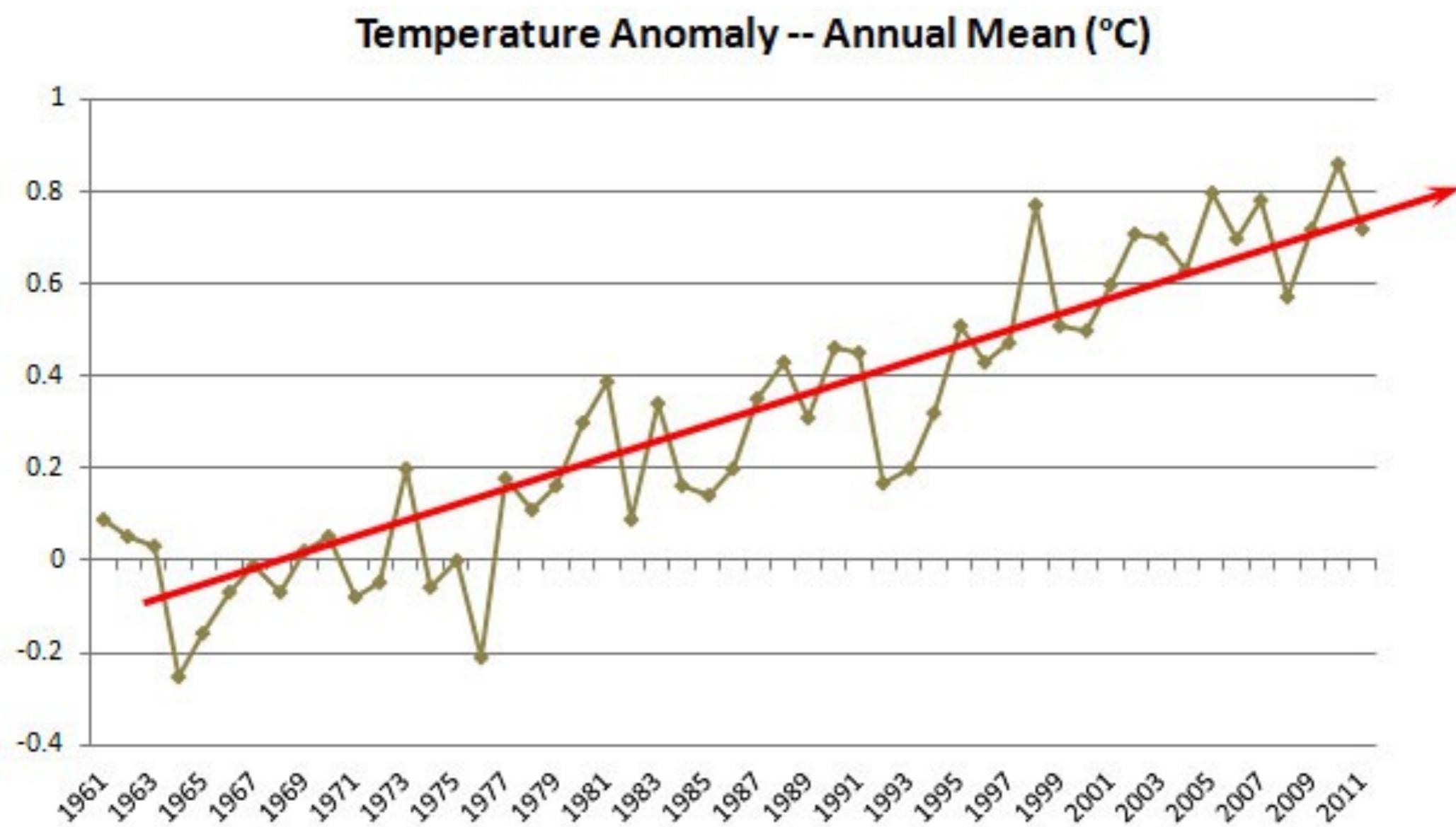
# Global Warming?



# Global Warming?



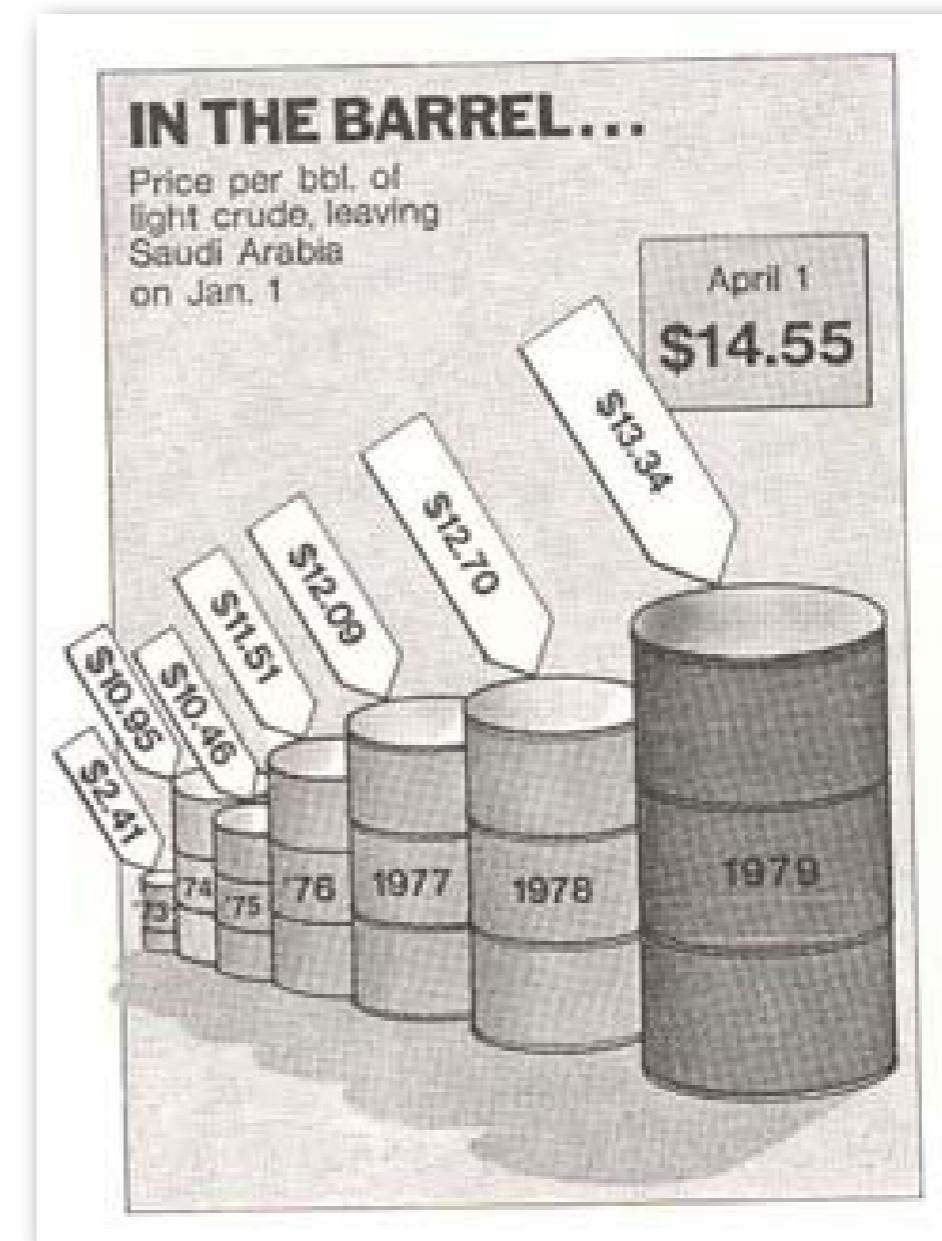
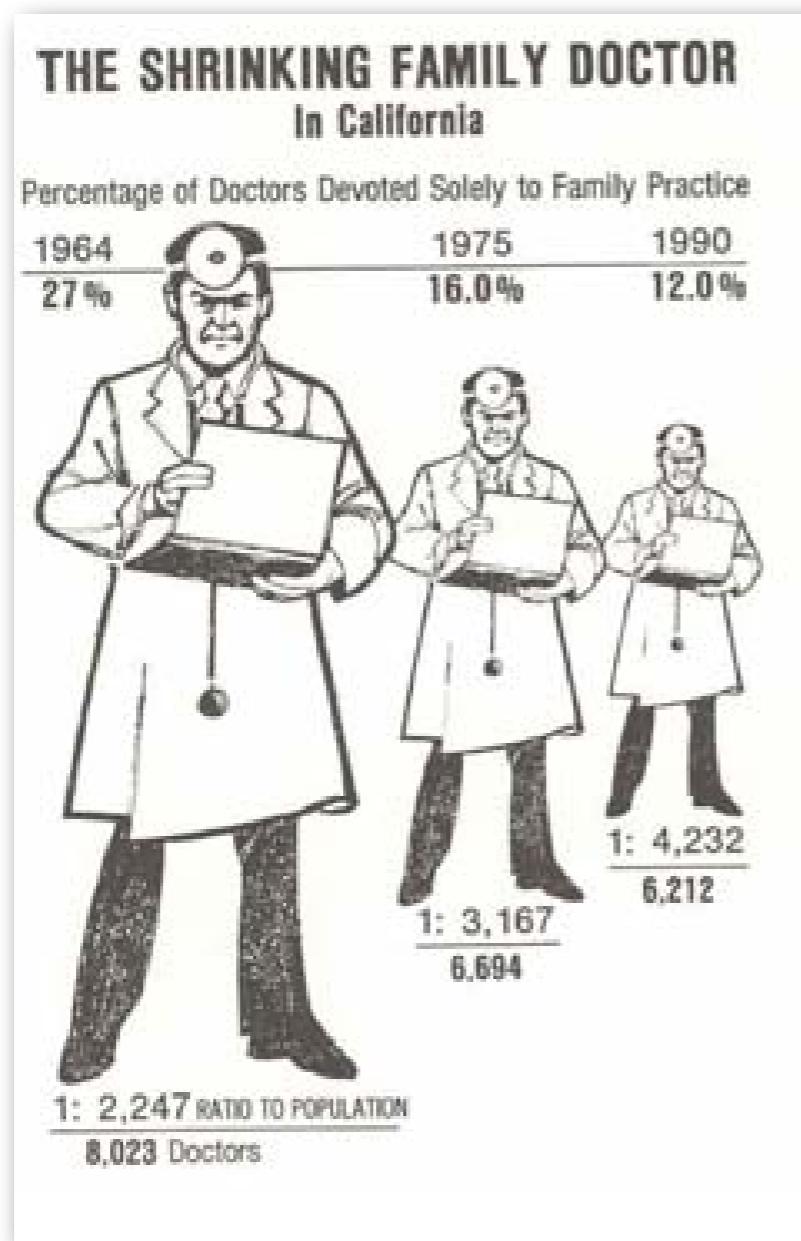
# Global Warming!



# The Lie Factor

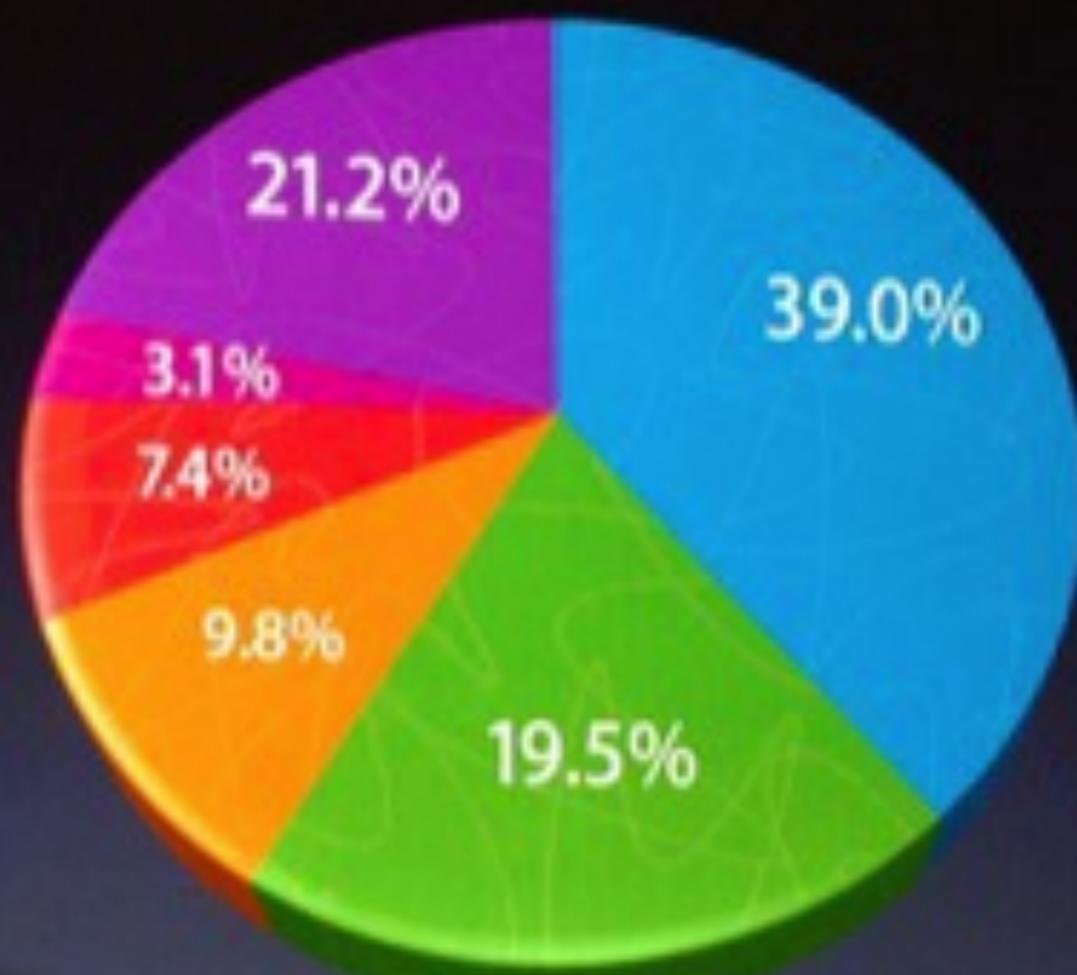
Size of effect shown in graphic

Size of effect in data



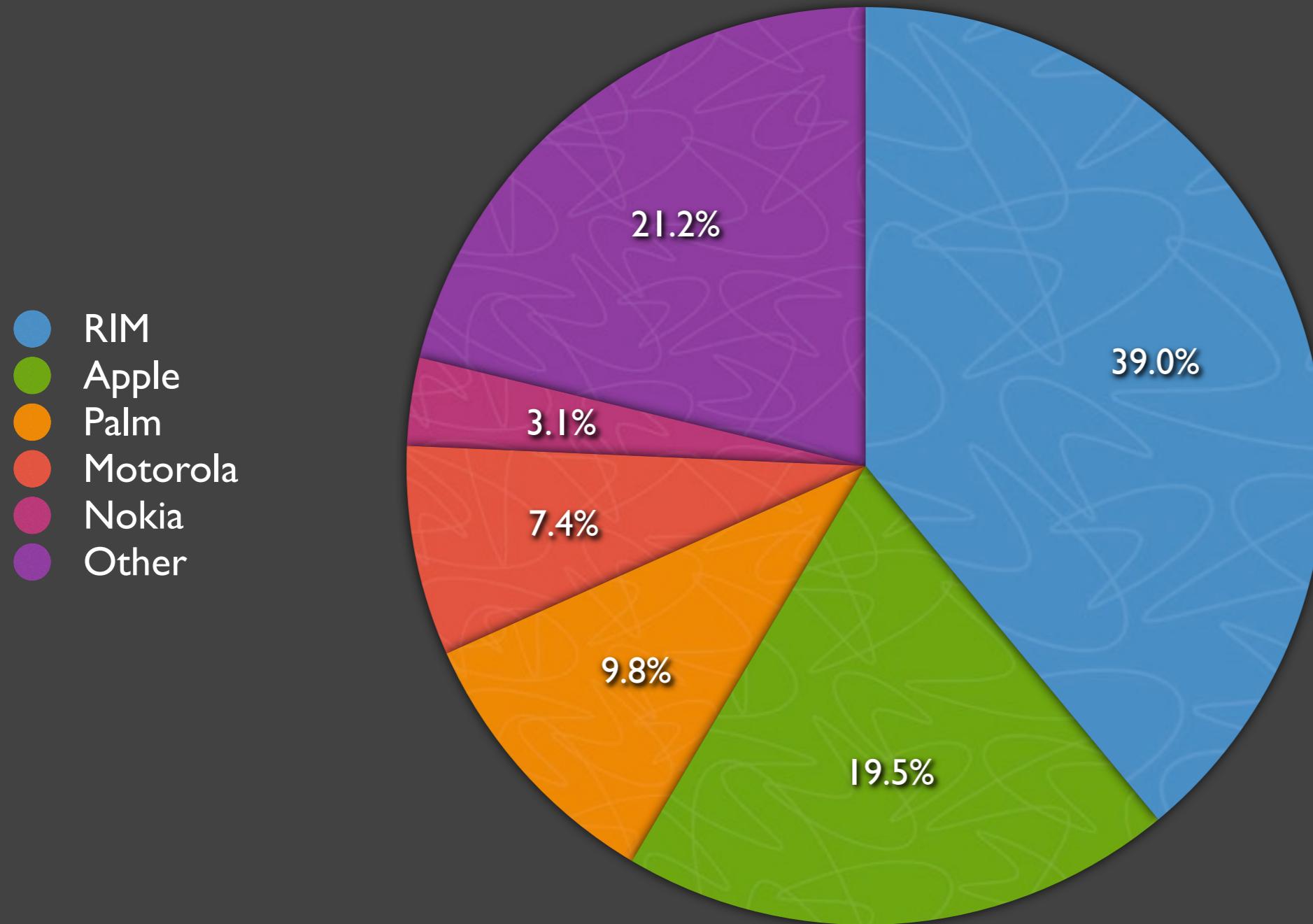
# U.S. SmartPhone Marketshare

- RIM
- Apple
- Palm
- Motorola
- Nokia
- Other

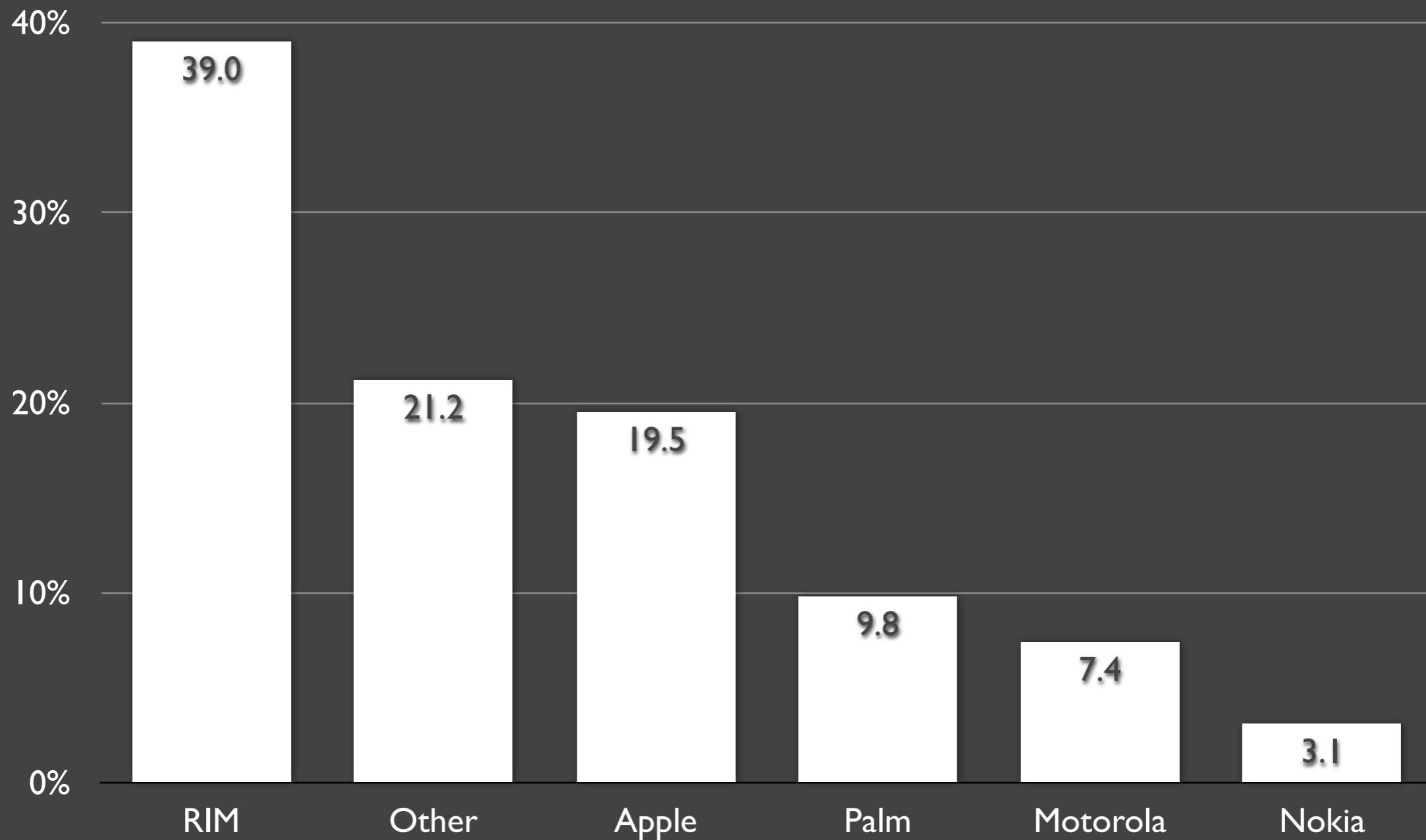


e... Gartner fo

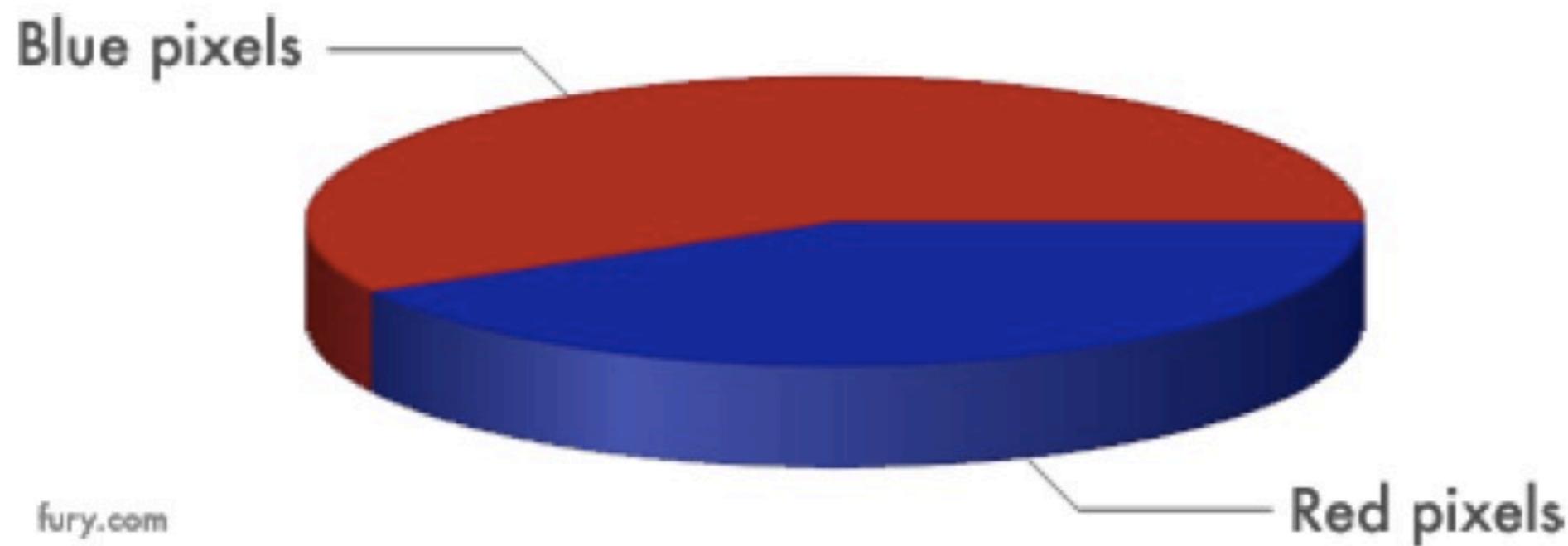
# U.S. SmartPhone Marketshare



# U.S. SmartPhone Marketshare



# Why 3D Pie Charts are Bad





Same Veritas. More Lux.

## Yale Summer Session

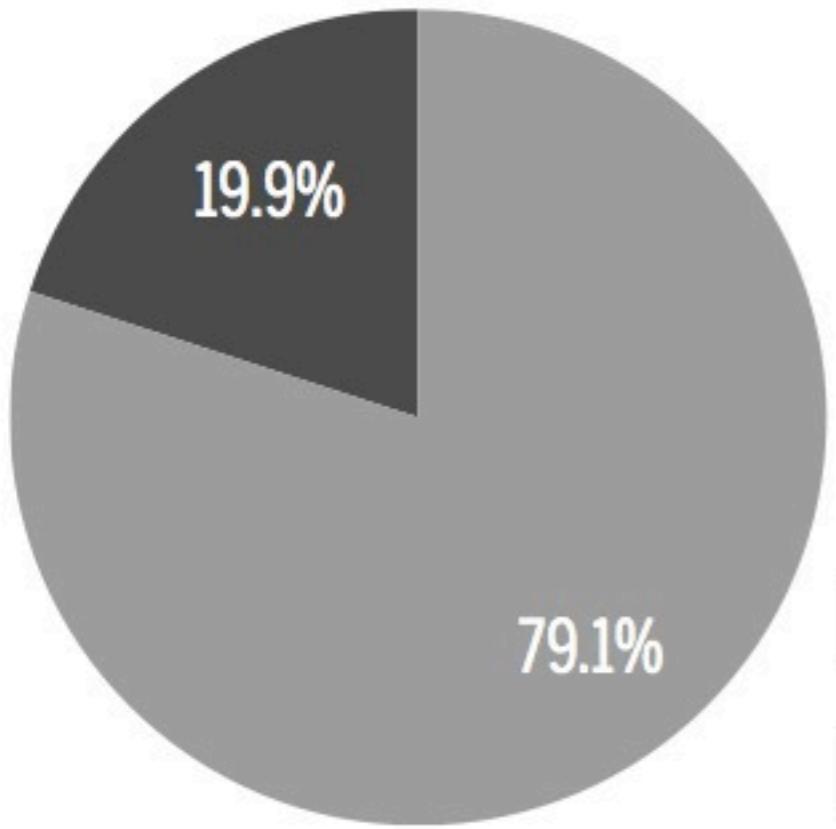
Over 200 full-credit courses.

June 4 – July 6 , July 9 – Aug 10

2012 experience Yale



### CHART YALE GRADUATES' MAJORS, CLASS OF 2011

Science, technology, engineering  
and math degrees

Non-STEM degrees

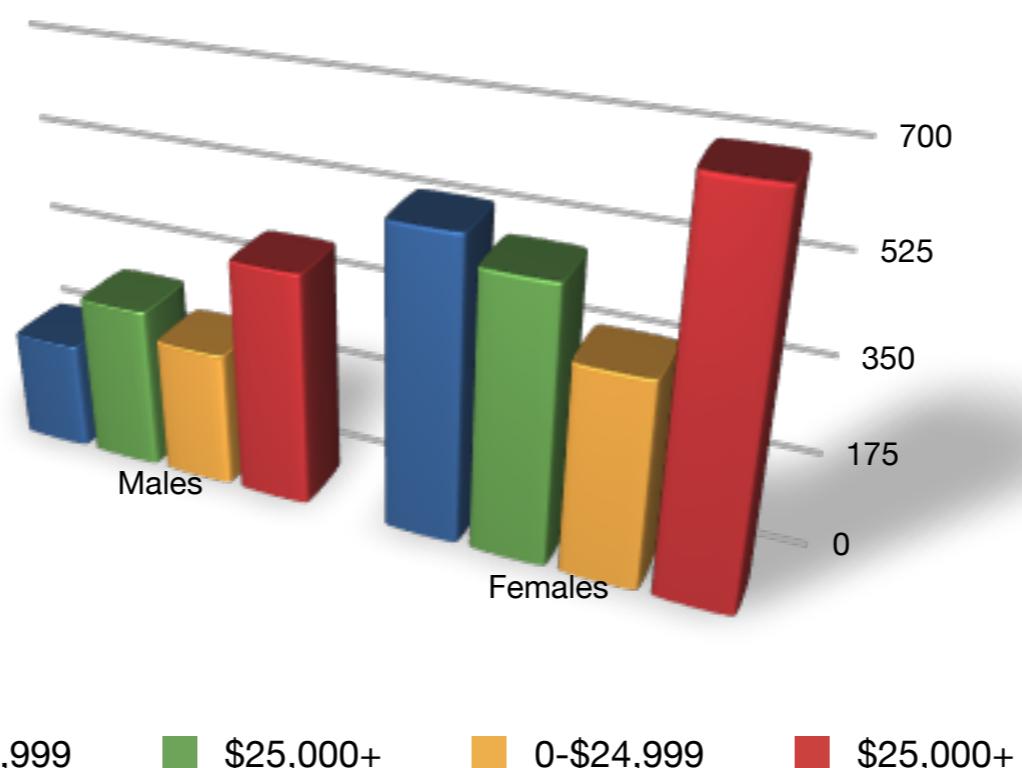
### Facebook Recommendations

[Shake Shack to open in New Haven](#)  
277 people recommend this.[Popular anti-religion creates false dichotomy](#)  
15 people recommend this.[Friends remember Foucher LAW '14](#)  
10 people recommend this.[AIDS activist speaks about documentary film](#)  
8 people recommend this.[Panel outlines changes in hip-hop](#)  
30 people recommend this.

Facebook social plugin

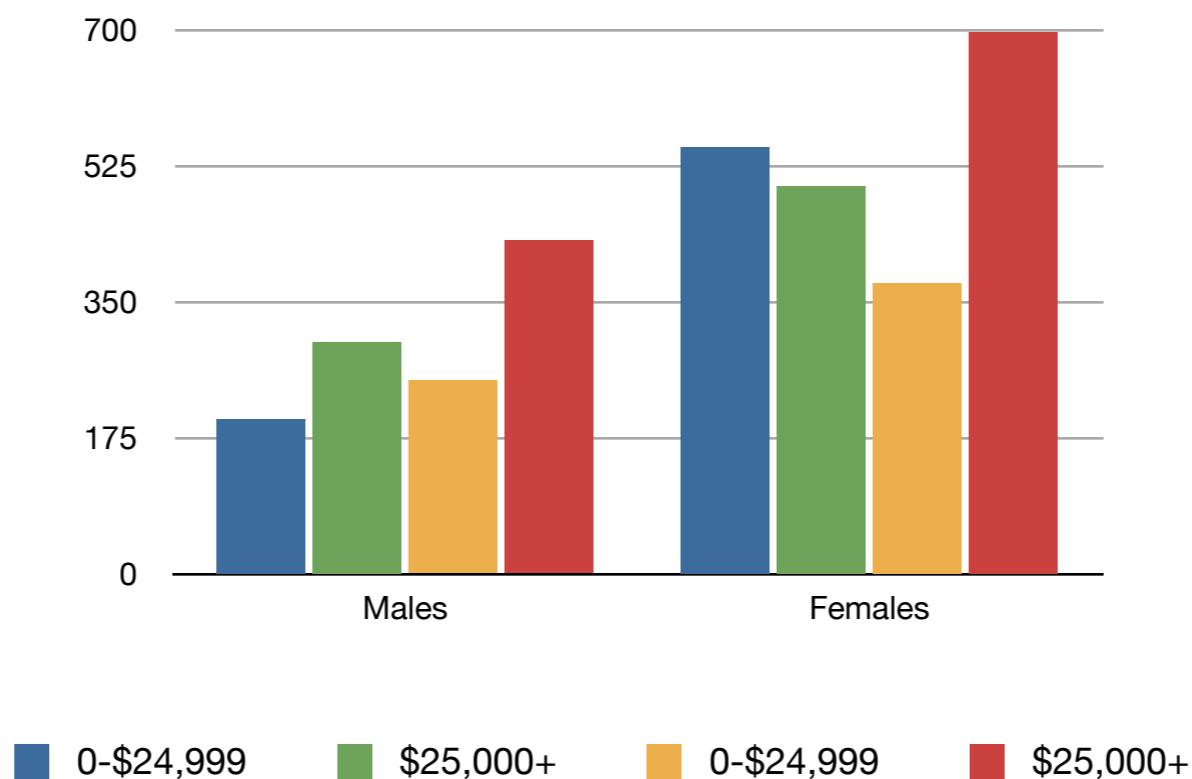
# Maximize Data-Ink Ratio

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

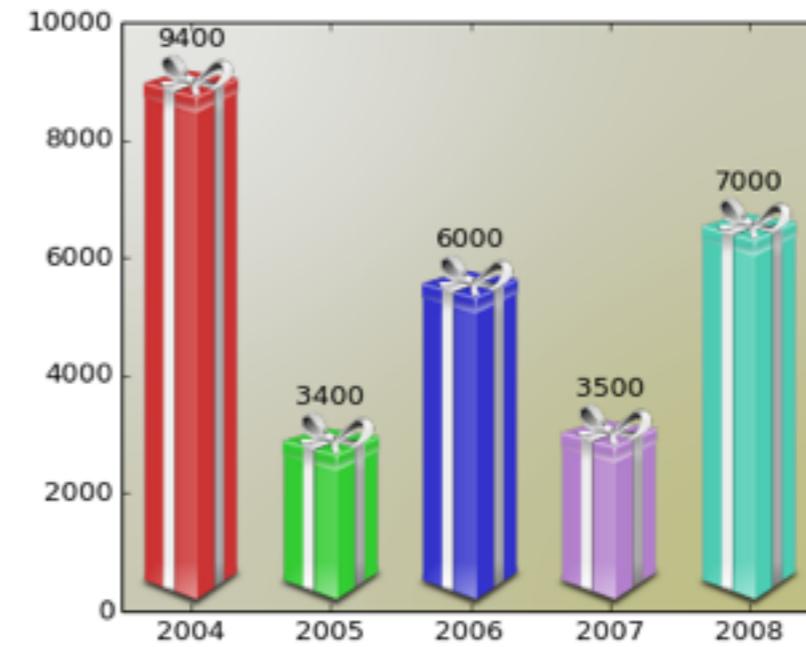
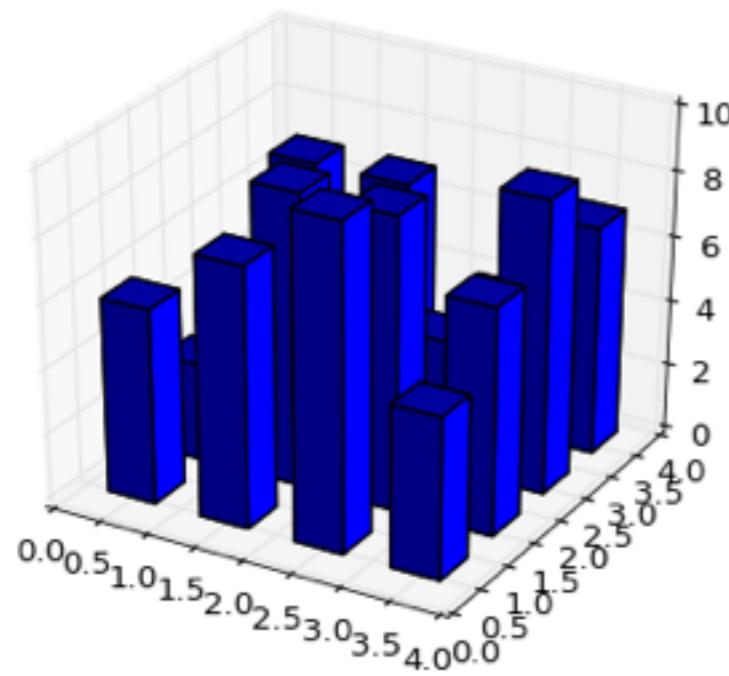


# Maximize Data-Ink Ratio

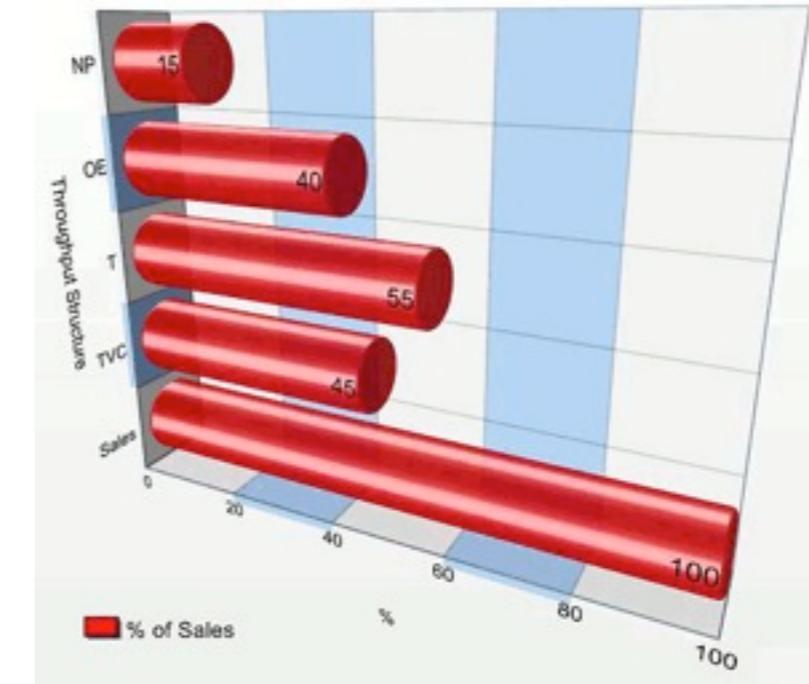
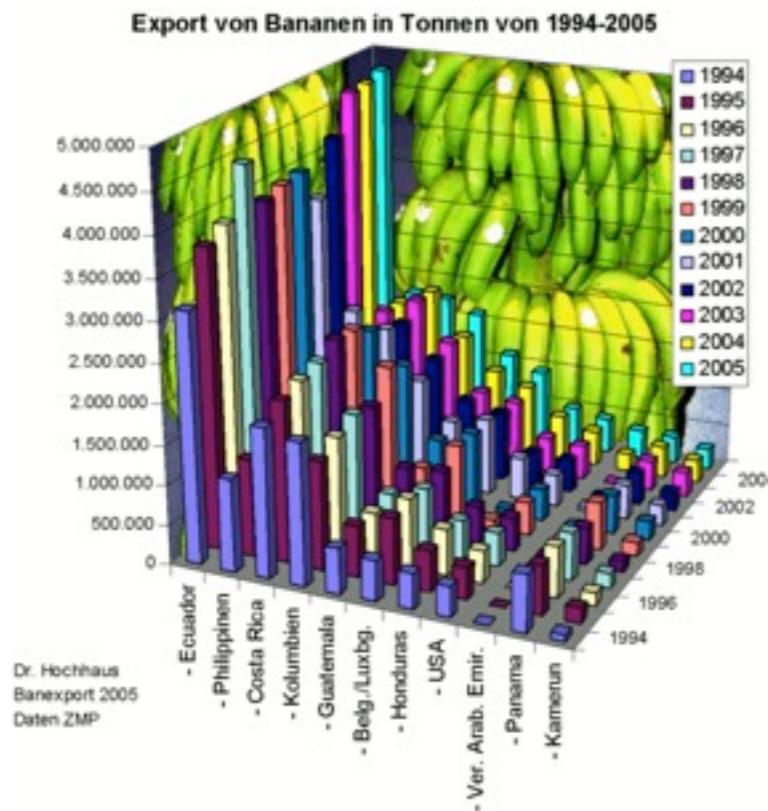
Data-Ink Ratio =  $\frac{\text{Data ink}}{\text{Total ink used in graphic}}$



# Don't



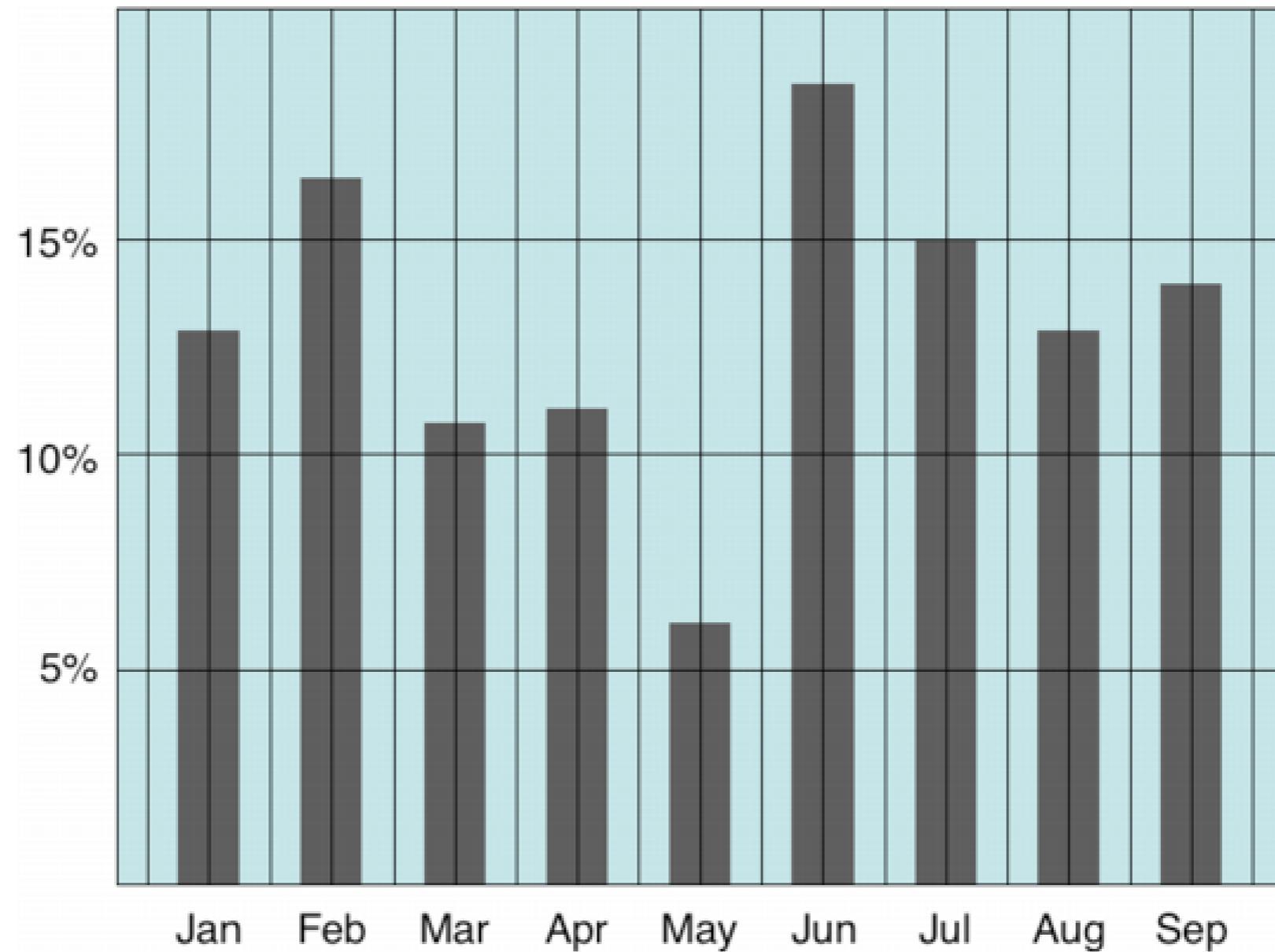
matplotlib gallery



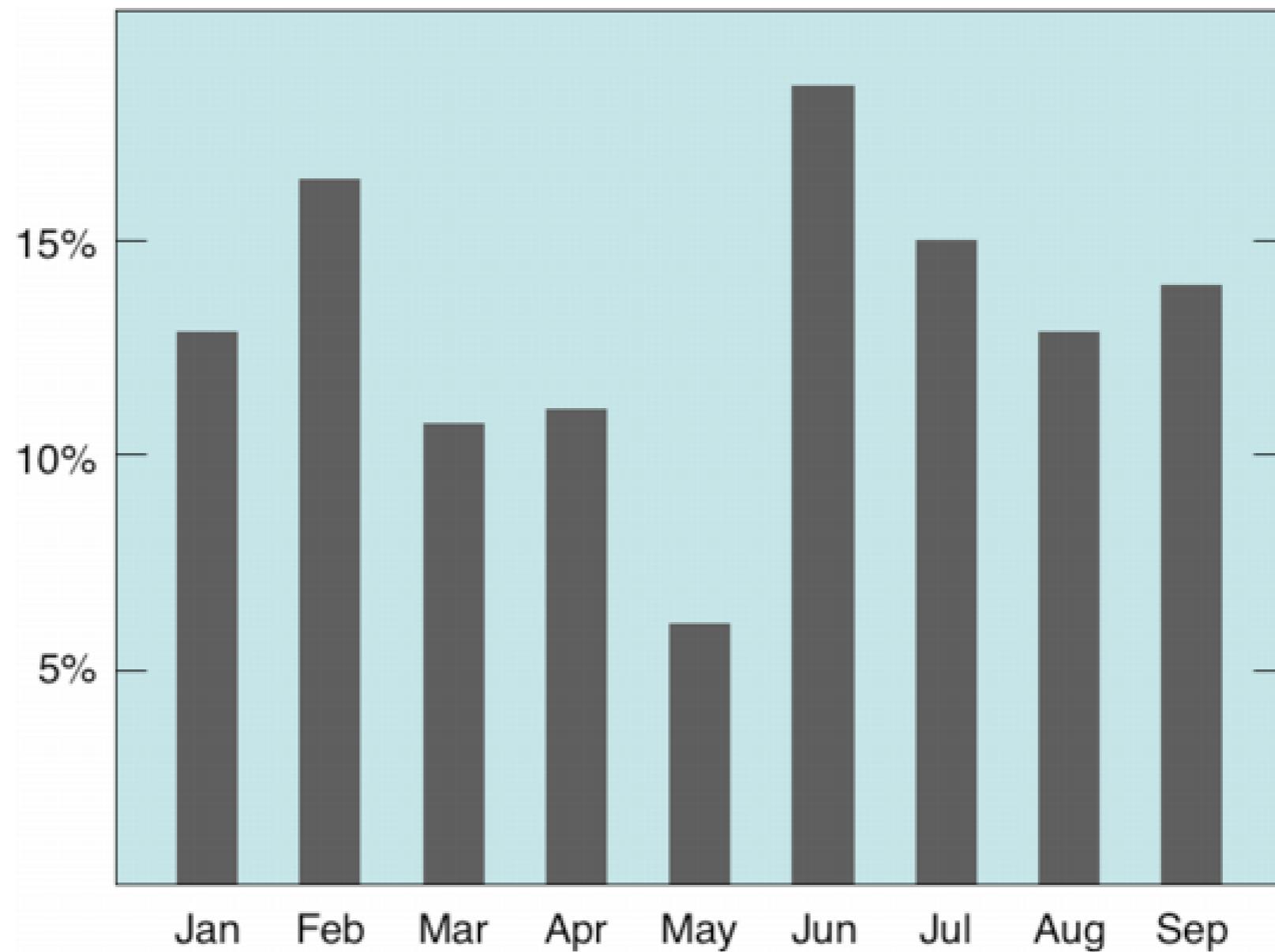
Excel Charts Blog

# Avoid Chartjunk

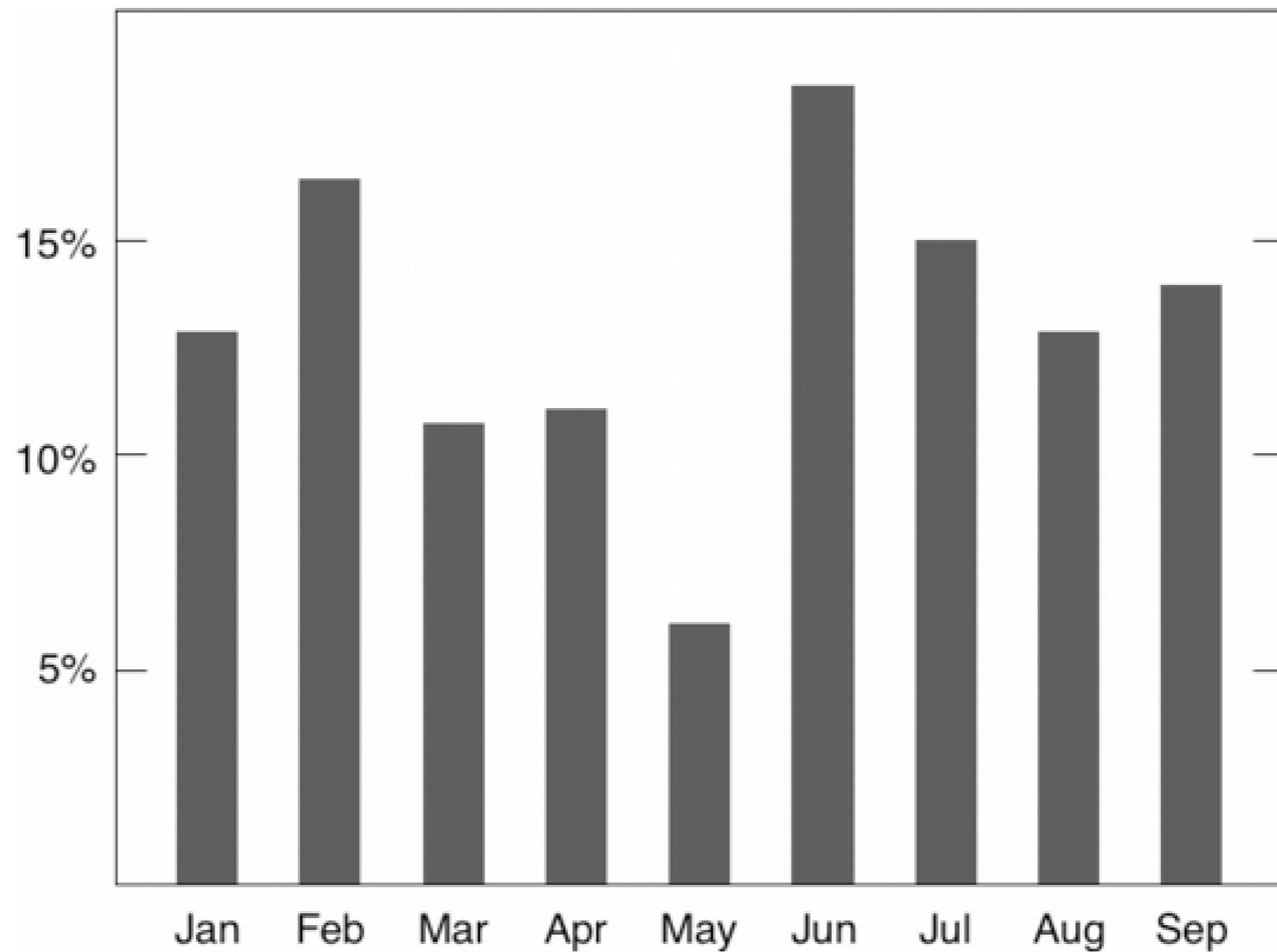
Extraneous visual elements that distract from the message



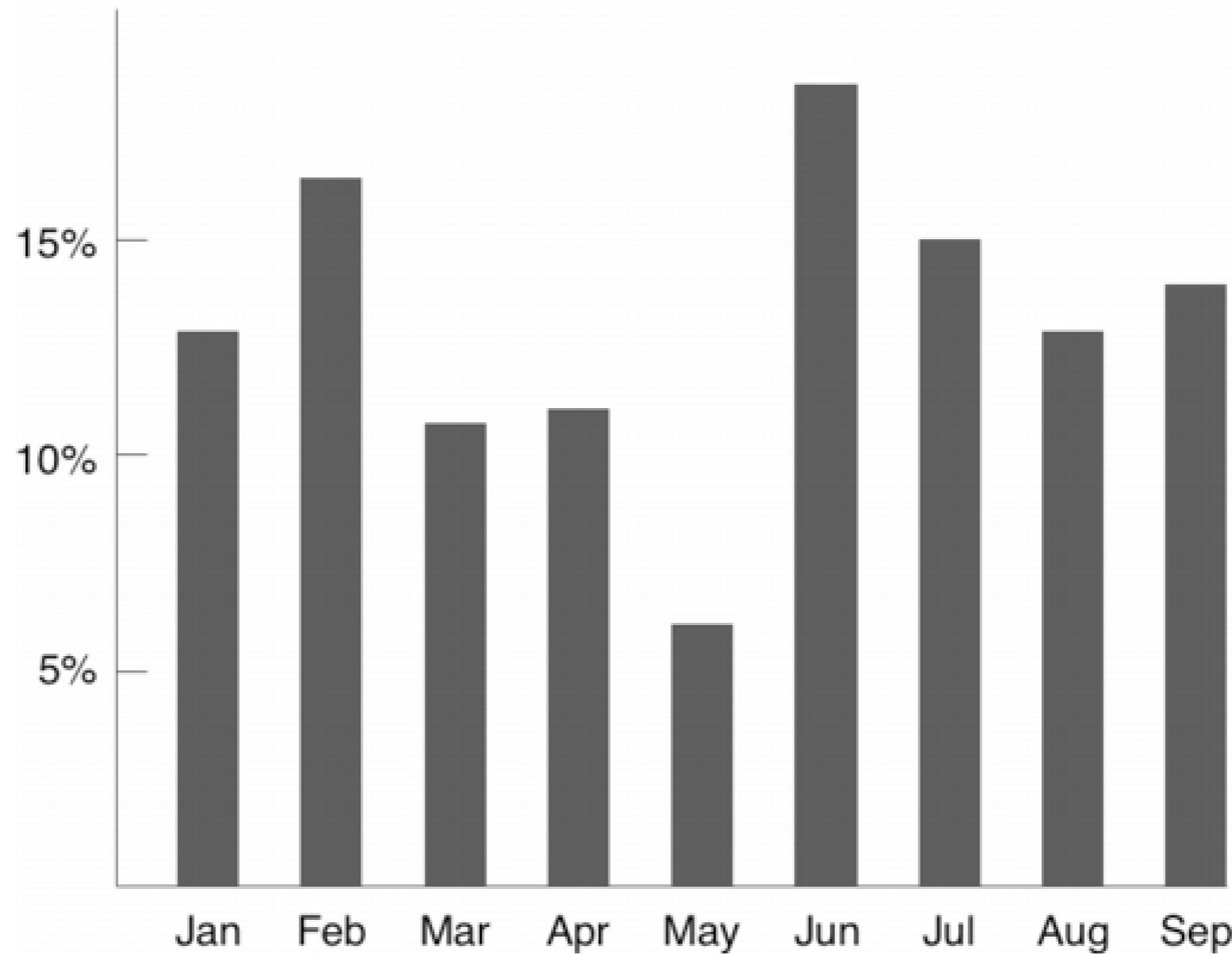
# Avoid Chartjunk



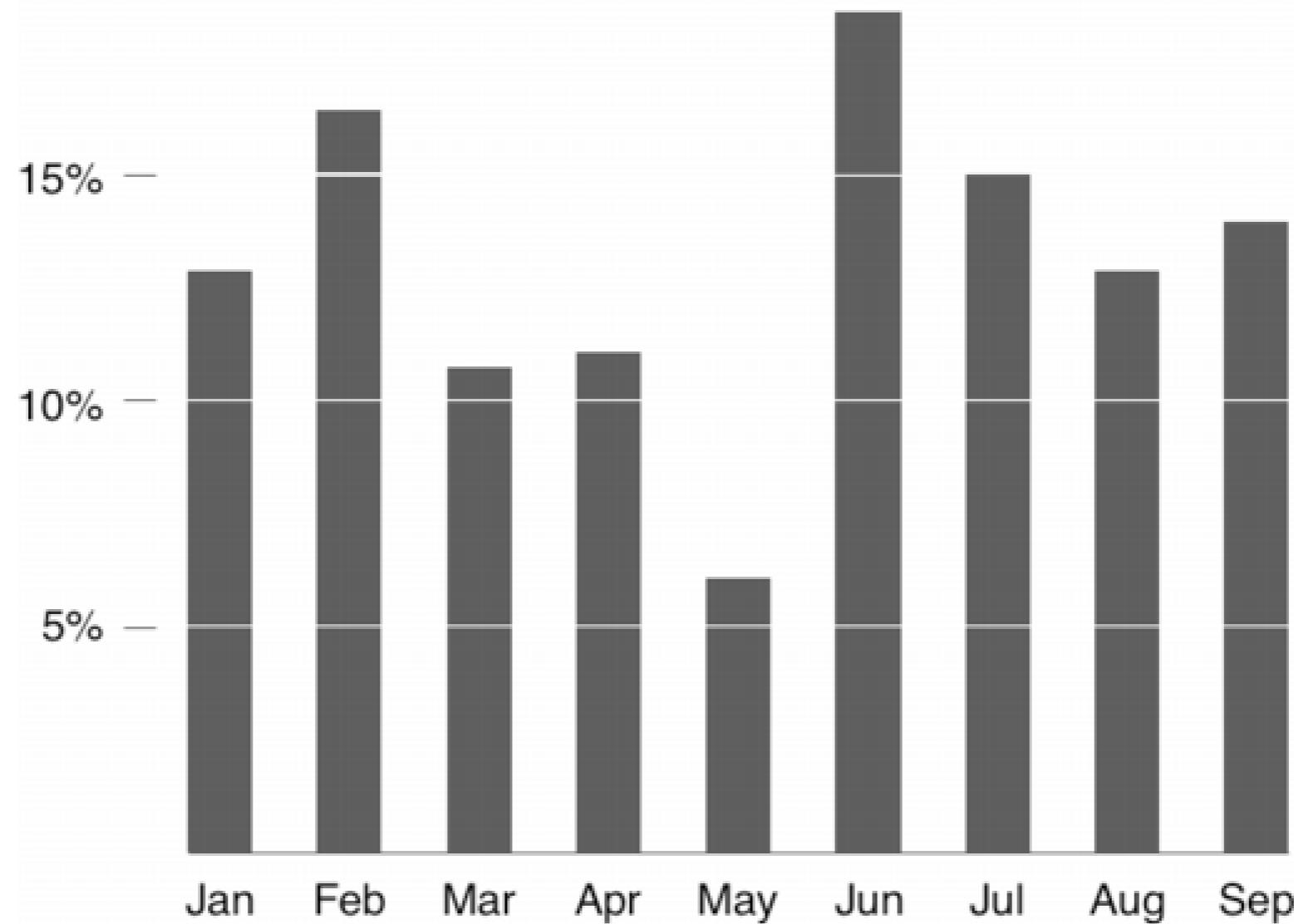
# Avoid Chartjunk



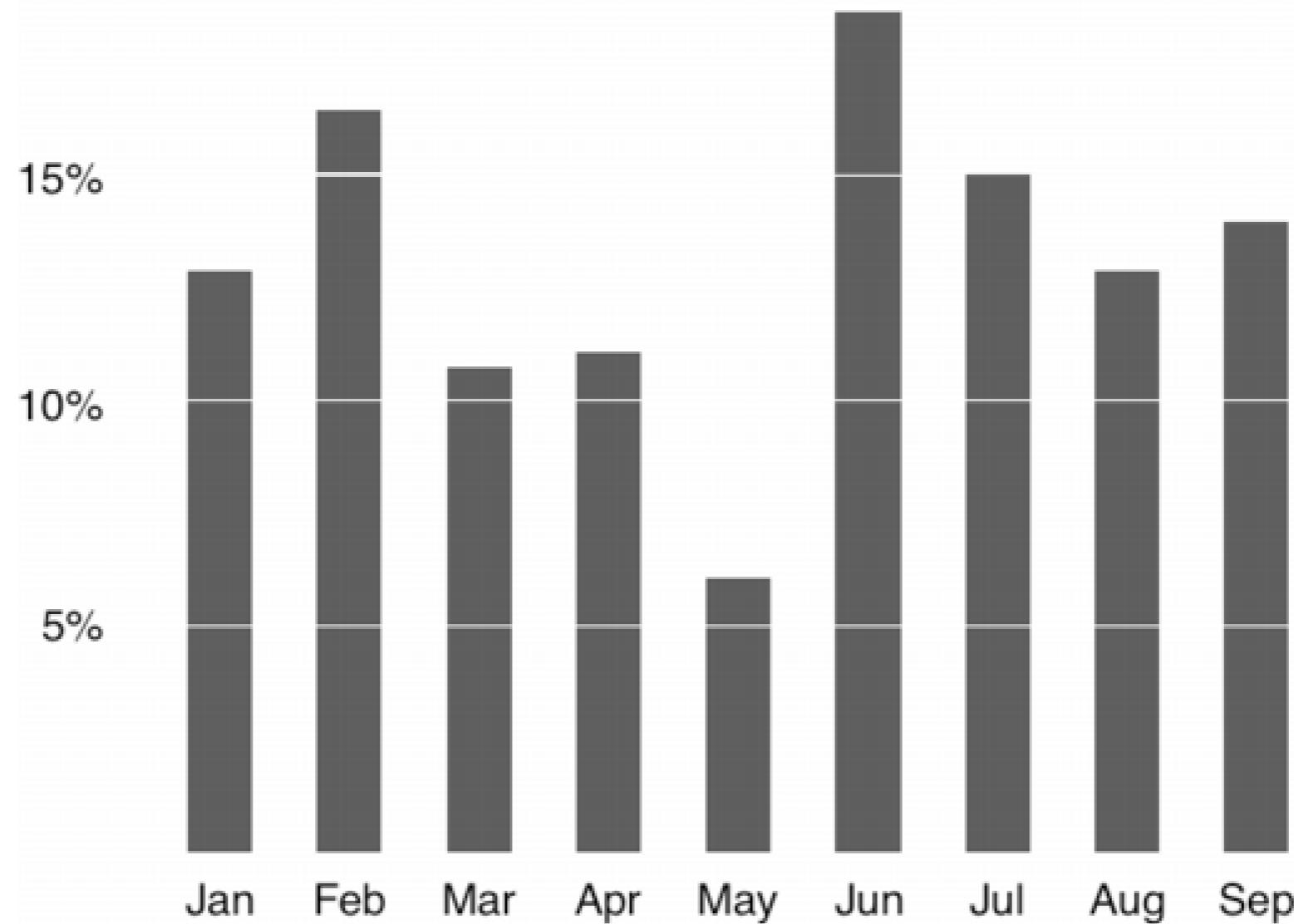
# Avoid Chartjunk



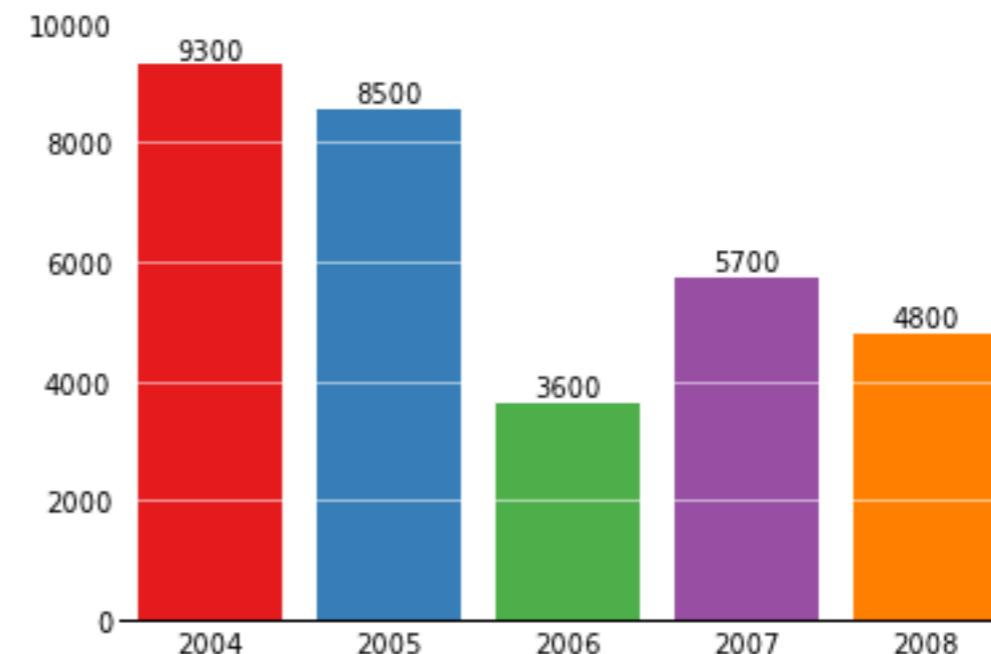
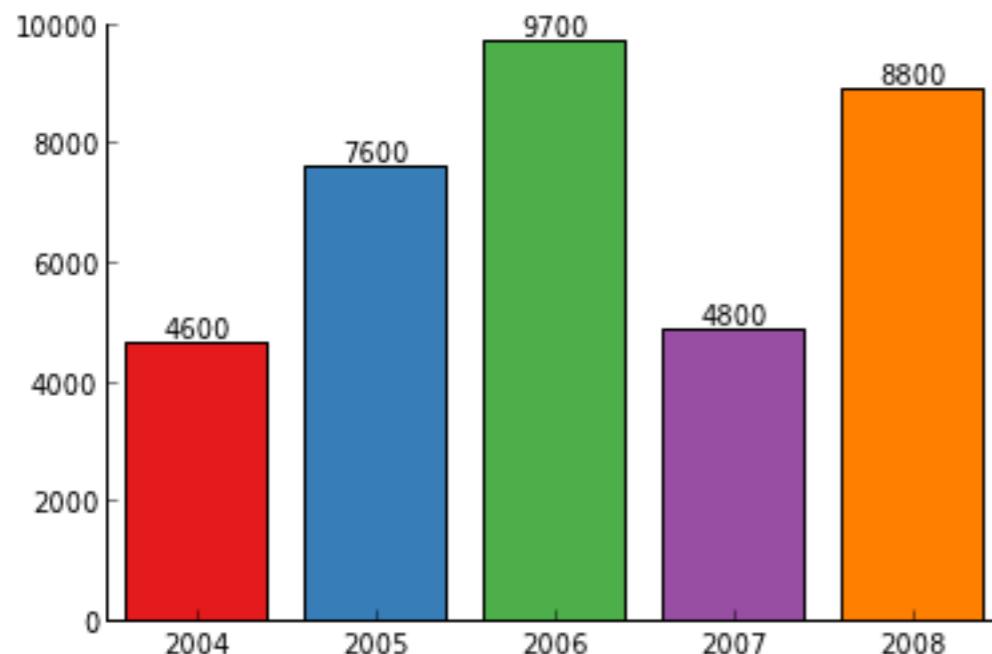
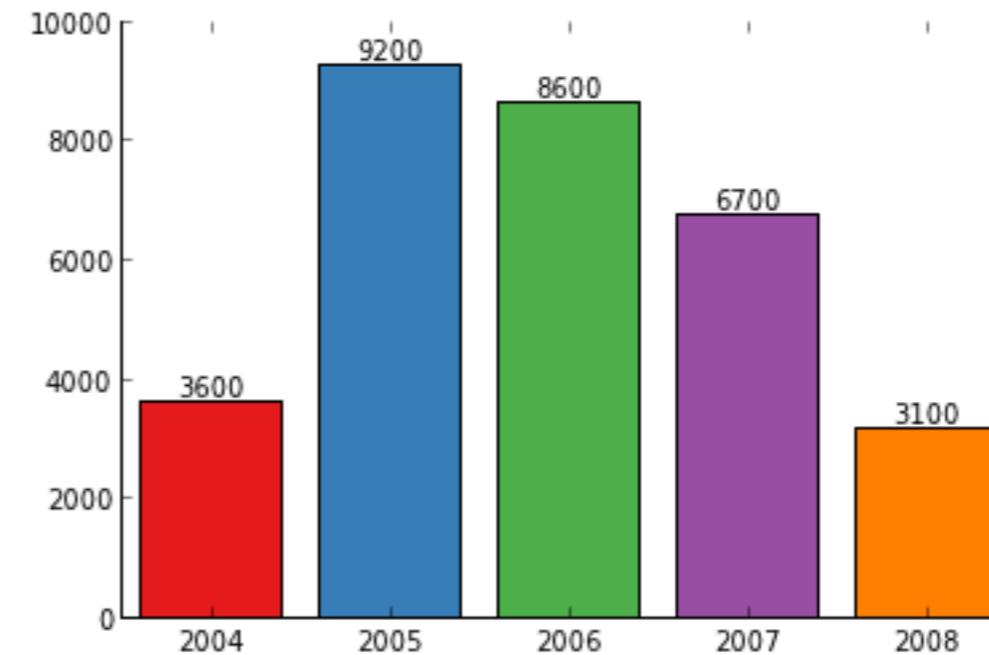
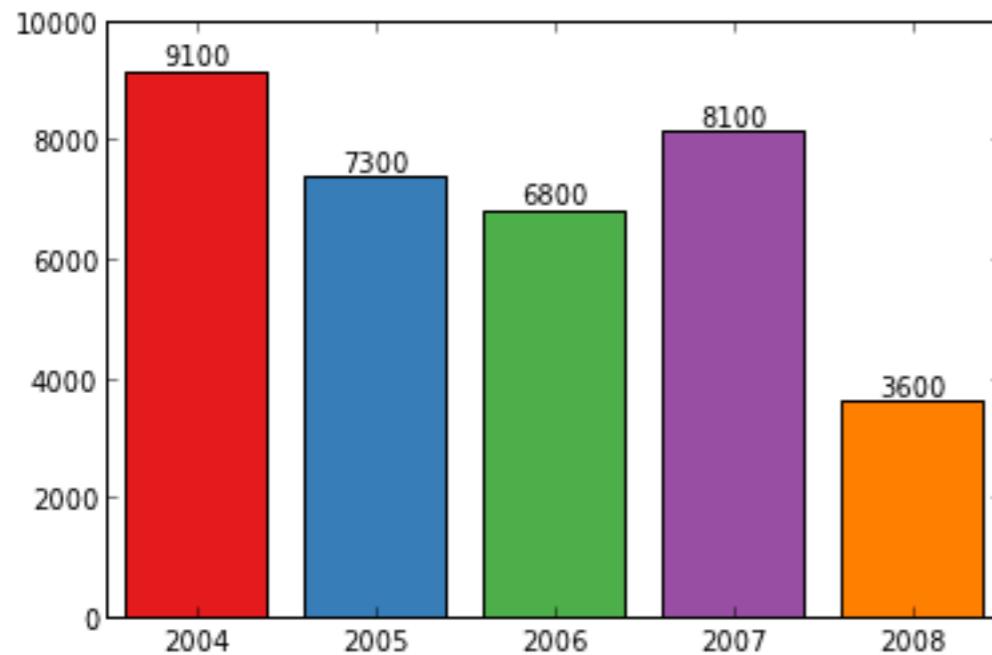
# Avoid Chartjunk



# Avoid Chartjunk



# Matplotlib Example



# Matplotlib Notebooks

- Links under “Readings” on the course web site
- Demonstrate how to make nice graphs in MPL

The screenshot shows a Jupyter Notebook interface. On the left, there is a code cell (In [78]) containing Python code to install the brewer2mpl package. On the right, the notebook title 'A Gallery of Statistical Graphs in Matplotlib' is displayed along with the author's name, Chris Beaumont. Below the title, there is a section titled 'A Gallery of Statistical Graphs in Matplotlib (Matplotlib Defaults)' which contains another code cell (In [24]). This cell contains the same setup code as the first one. At the bottom of the notebook, there is a section titled 'Example Data' with a code cell (In [25]) that reads data from a CSV file.

A Gallery of Statistical Graphs in Matplotlib

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 unnecessary plot borders and axis ticks
    The top/right/left/bottom keywords toggle whether the corresponding plot border is drawn
    """
    ox = axes or plt.gca()
    ox.spines['top'].set_visible(top)
    ox.spines['right'].set_visible(right)
    ox.spines['left'].set_visible(left)
    ox.spines['bottom'].set_visible(bottom)

    #turn off all ticks
    ox.xaxis.set_ticks_position('none')
    ox.yaxis.set_ticks_position('none')
```

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
```

A Gallery of Statistical Graphs in Matplotlib (Matplotlib Defaults)

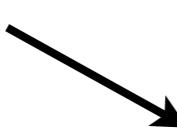
In [25]:

```
file = urlopen('https://raw.githubusercontent.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')
```

Example Data

I  
Python  
creator



**Tweets**

**Fernando Perez** @fperez\_org 8h  
A Gallery of Statistical Graphs in Matplotlib: nicer plots than the default MPL produces, with all code:  
[nbviewer.ipython.org/urls/raw.github...](http://nbviewer.ipython.org/urls/raw.github...)

[Collapse](#)

**24 RETWEETS** **37 FAVORITES**

8:50 PM - 9 Sep 13 · Details

**Hadley Wickham** @hadleywickham 1h  
@fperez\_org it's still embarrassing how much code you need to produce a nice graph in matplotlib. why is python visualisation so far behind?

[Expand](#)

ggplot2 (R)  
creator

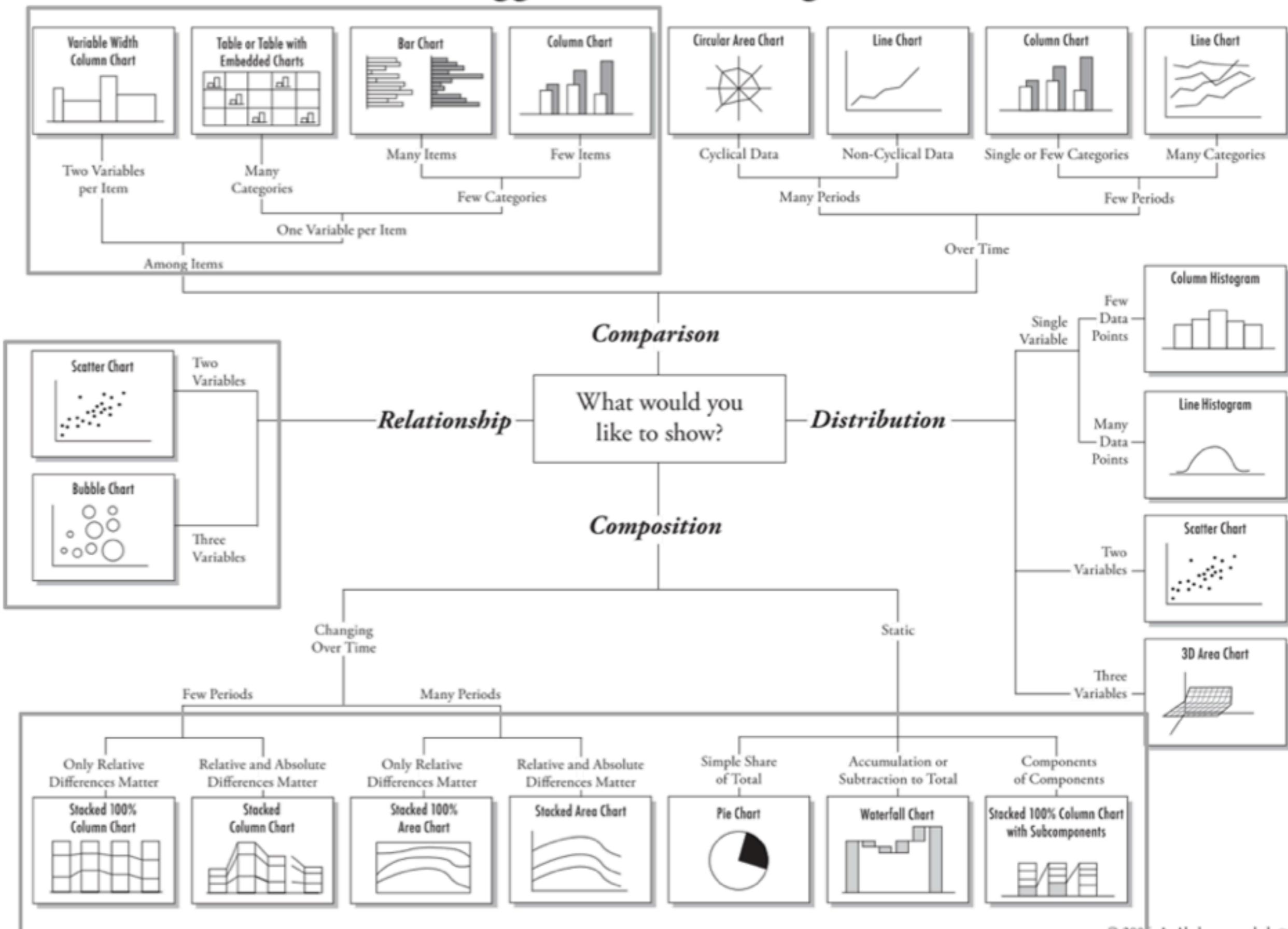


# Tufte's Design Principles

- 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

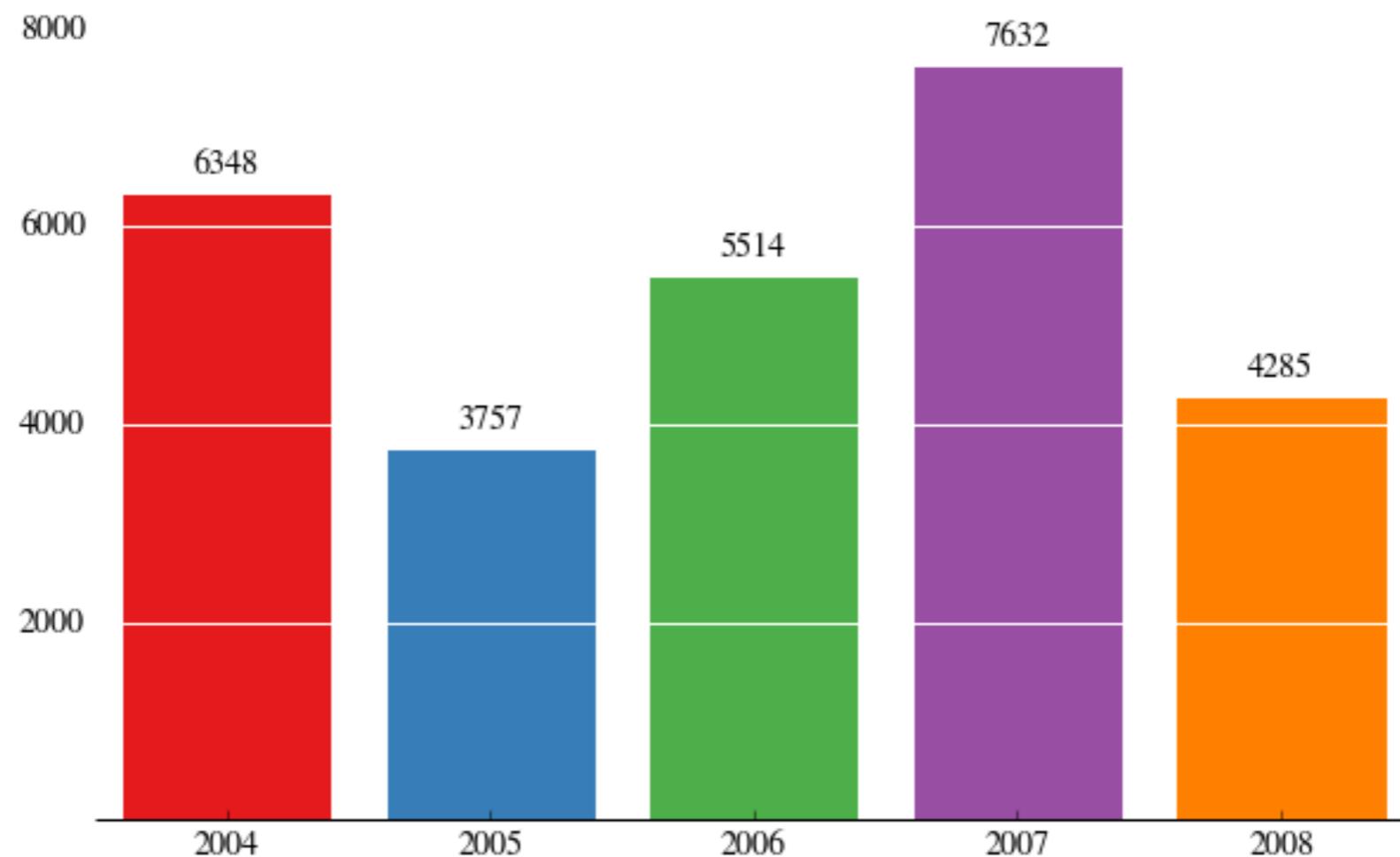
# Graph Types

# Chart Suggestions—A Thought-Starter



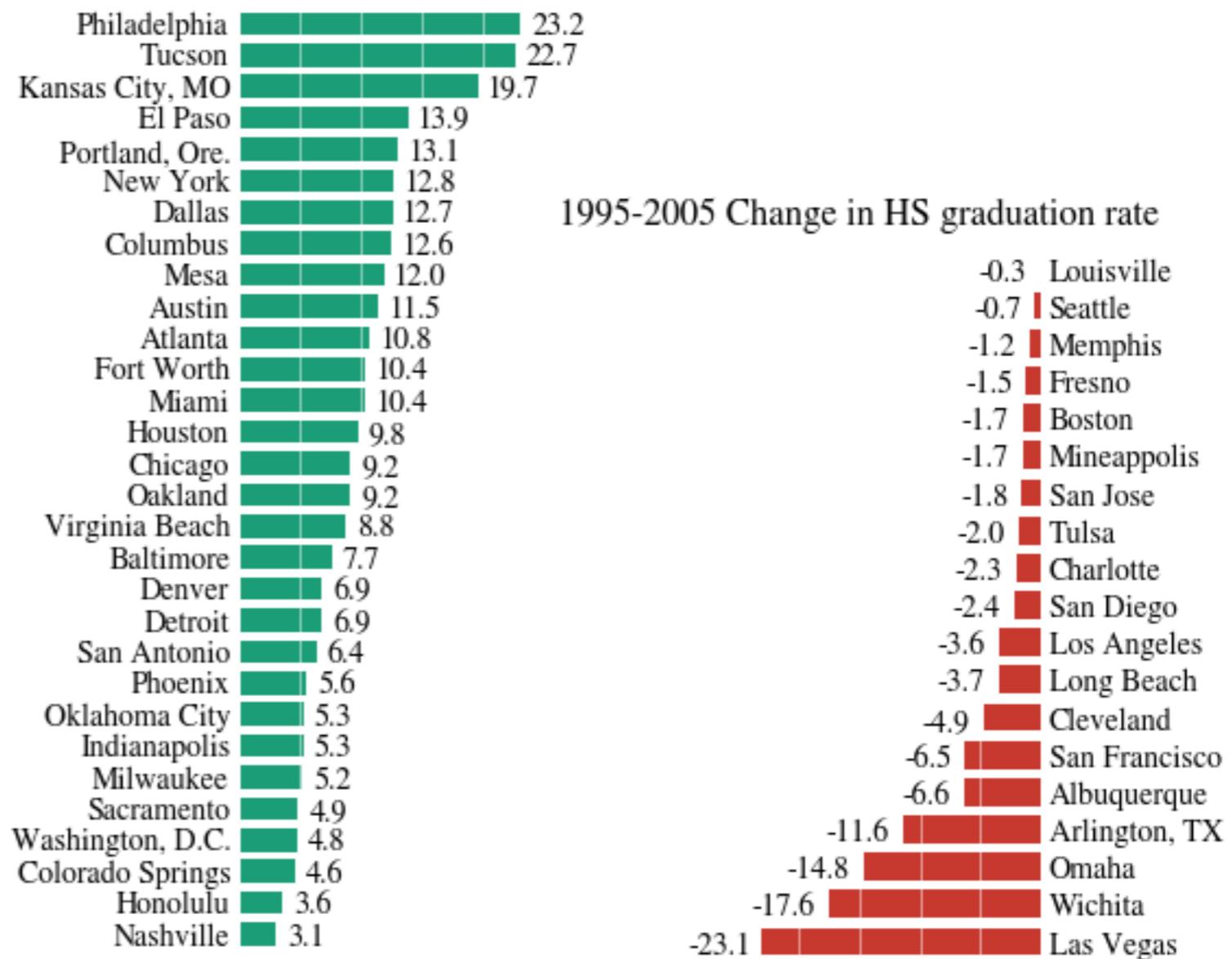
# Comparisons

# Bar Chart



# Direction

## 1995-2005 Change in HS graduation rate



## Graduation rates up in most cities

Graduation rate for principal school district of the largest cities

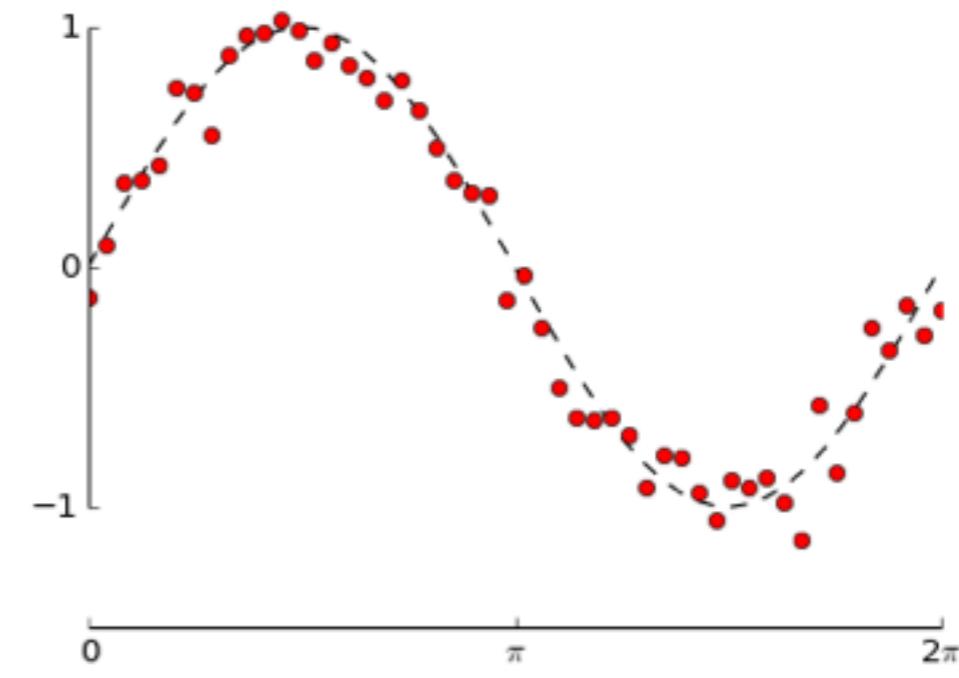
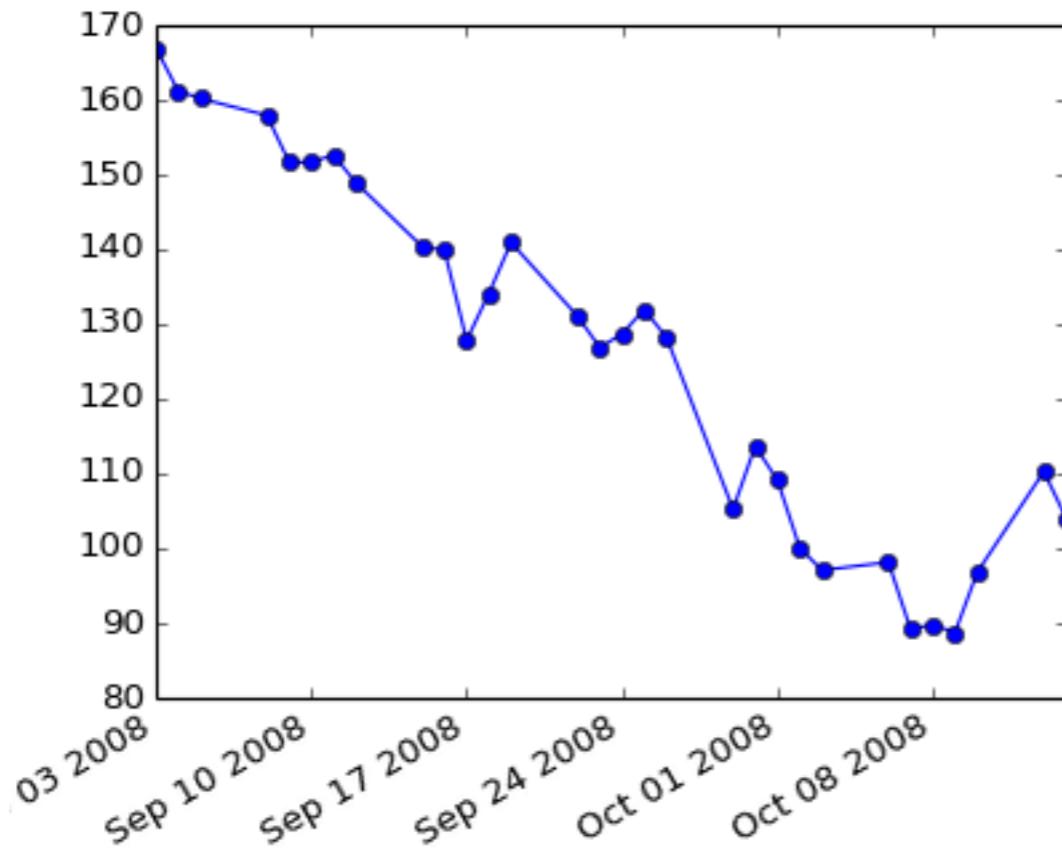
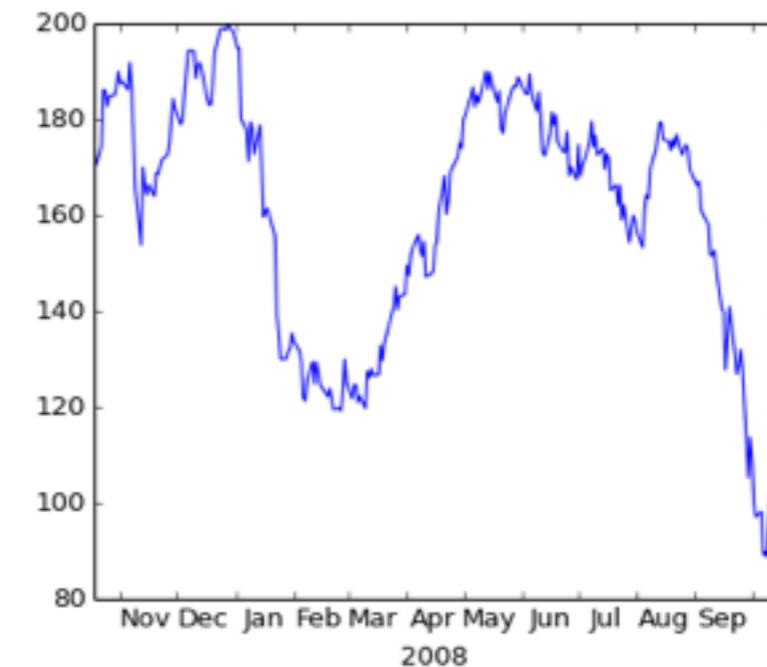
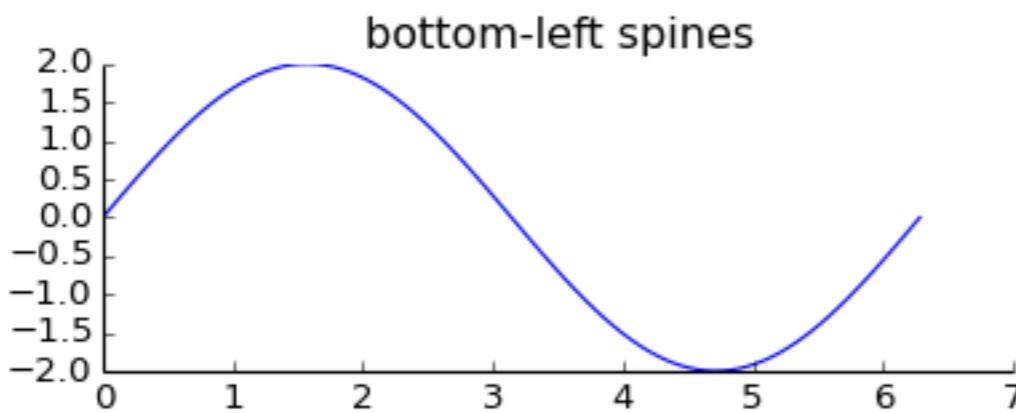
|                  | 2005 RATE | 1995-2005 CHANGE |  |
|------------------|-----------|------------------|--|
| Philadelphia     | 62.1%     | 23.2%            | The average high school graduation rate of major cities was 54.7 percent in 2005. Of the 62 percent that improved since 1995, Philadelphia had the highest increase. |
| Tucson           | 71.6      | 22.7             |  |
| Kansas City, Mo. | 53.3      | 19.7             |  |
| El Paso          | 60.6      | 13.9             |  |
| Portland, Ore.   | 68.6      | 13.1             |  |
| New York         | 50.5      | 12.8             |  |
| Dallas           | 50.8      | 12.7             |  |
| Columbus         | 44.7      | 12.6             |  |
| Mesa             | 76.6      | 12.0             | The rate in Las Vegas decreased the most.  |
| Austin           | 58.9      | 11.5             |  |
| Atlanta          | 43.5      | 10.8             |  |
| Fort Worth       | 56.5      | 10.4             |  |
| Miami            | 55.9      | 10.4             |  |
| Houston          | 52.9      | 9.8              |  |
| Chicago          | 51.0      | 9.2              |  |
| Oakland          | 50.5      | 9.2              |  |
| Virginia Beach   | 68.5      | 8.8              |  |
| Baltimore        | 41.5      | 7.7              |  |
| Denver           | 58.6      | 6.9              |  |
| Detroit          | 37.5      | 6.9              |  |
| San Antonio      | 47.3      | 6.4              |  |
| Phoenix          | 58.0      | 5.6              |  |
| Oklahoma City    | 47.0      | 5.3              |  |
| Indianapolis     | 30.5      | 5.3              |  |
| Milwaukee        | 41.0      | 5.2              |  |
| Sacramento       | 62.1      | 4.9              |  |
| Washington, D.C. | 57.6      | 4.8              |  |
| Colorado Springs | 68.8      | 4.6              |  |
| Honolulu         | 67.4      | 3.6              |  |
| Nashville        | 45.2      | 3.1              |  |
| Jacksonville     | 50.8      | 0.7              |  |

SOURCE: EPE Research Center

AP

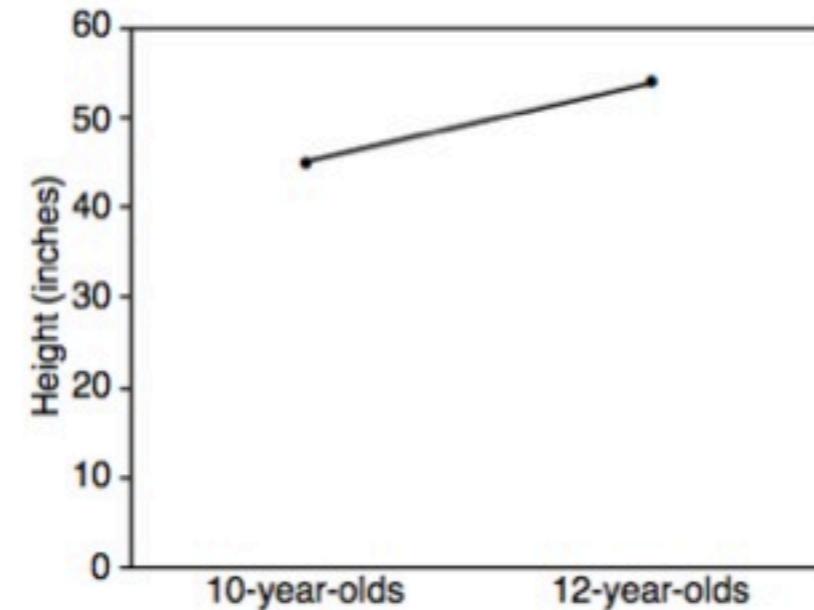
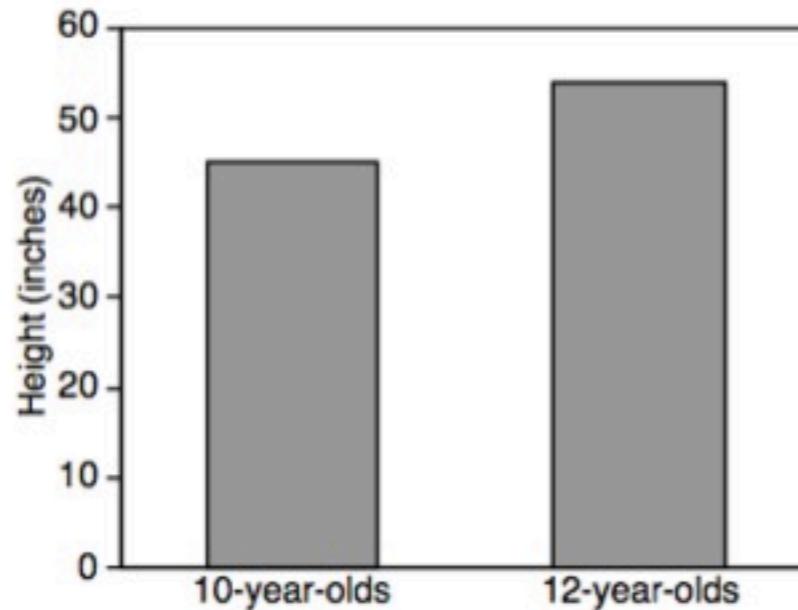
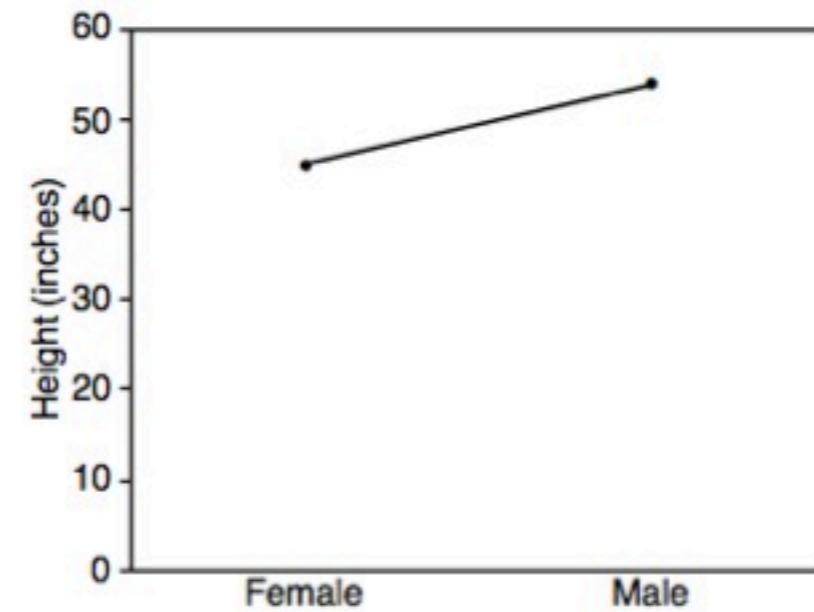
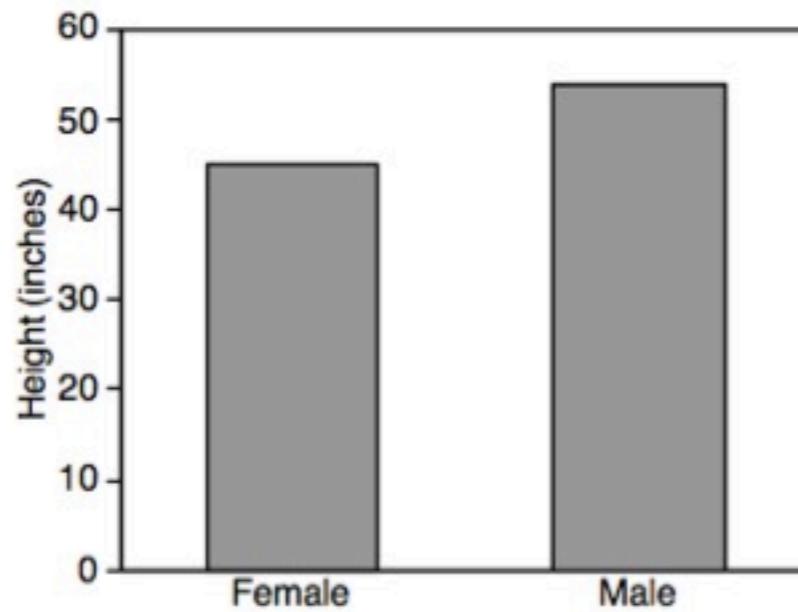
# Trends Over Time

# Line Charts



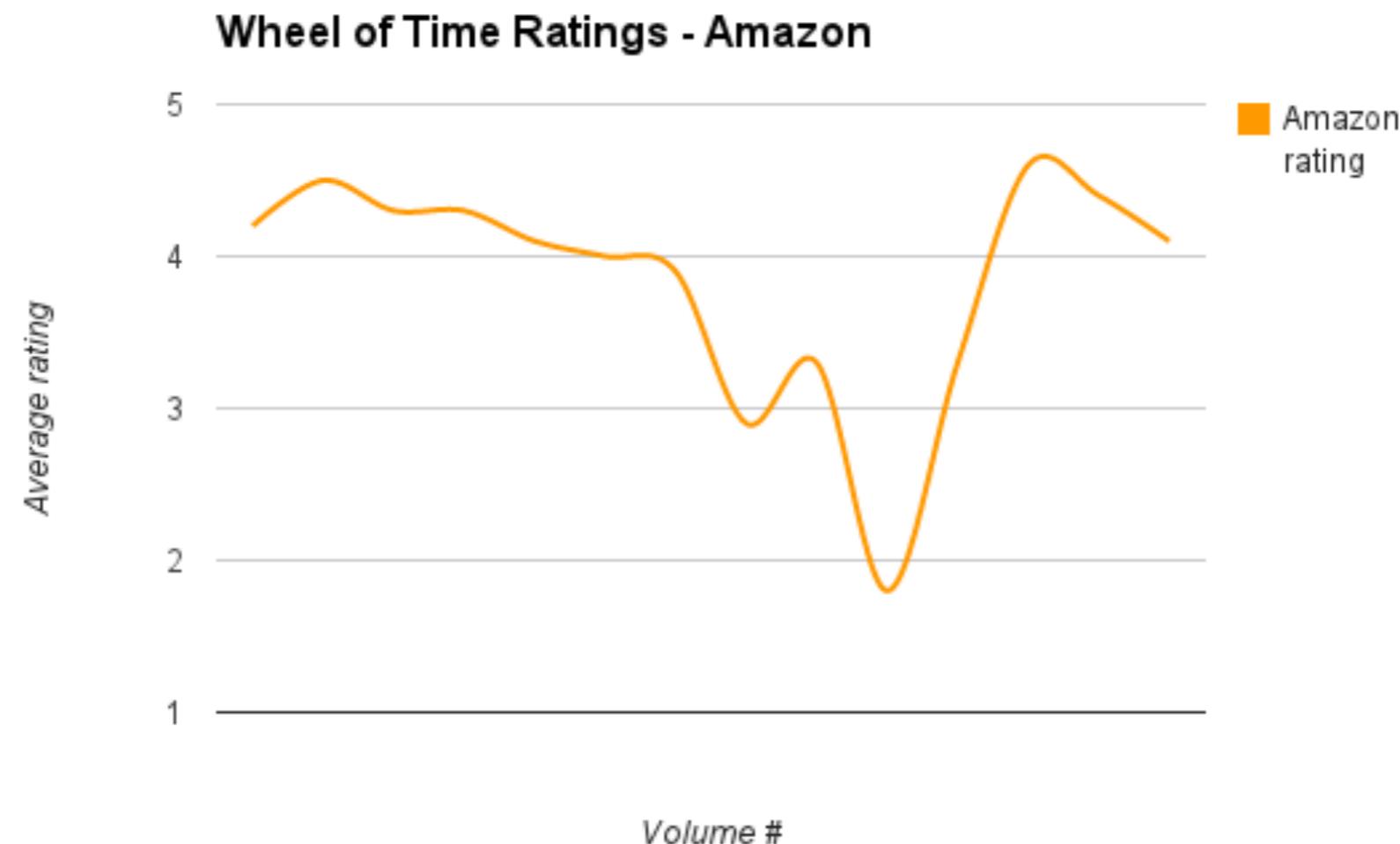
# Bars vs. Lines

Lines imply connections - do not use for categorical data



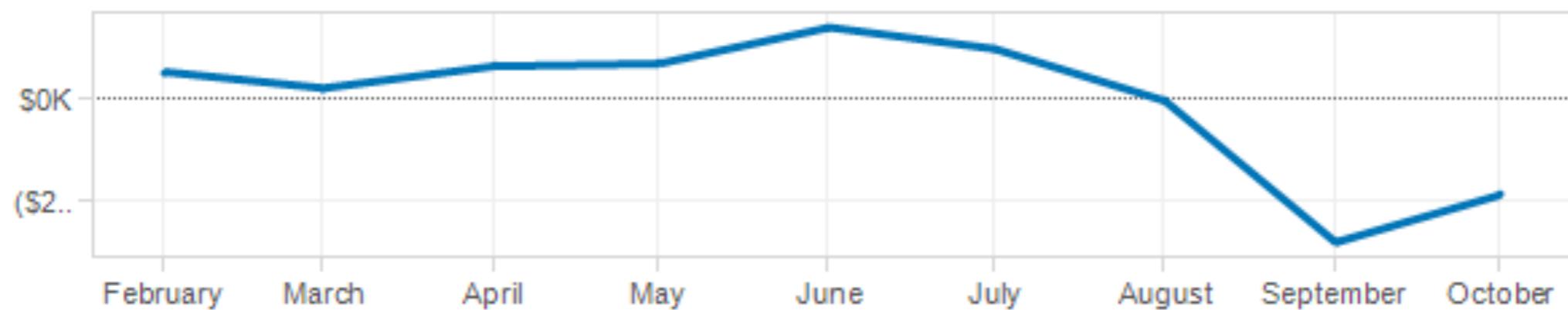
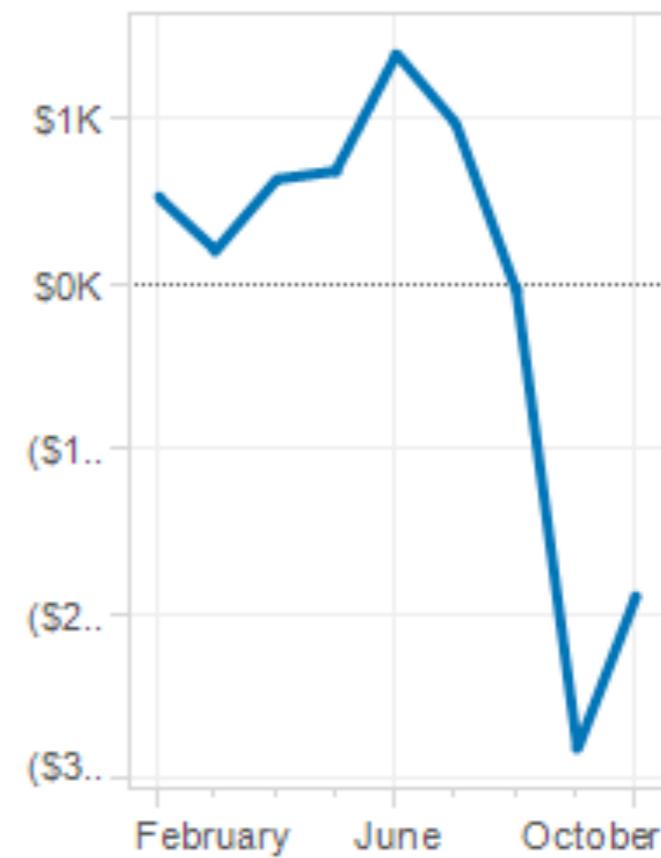
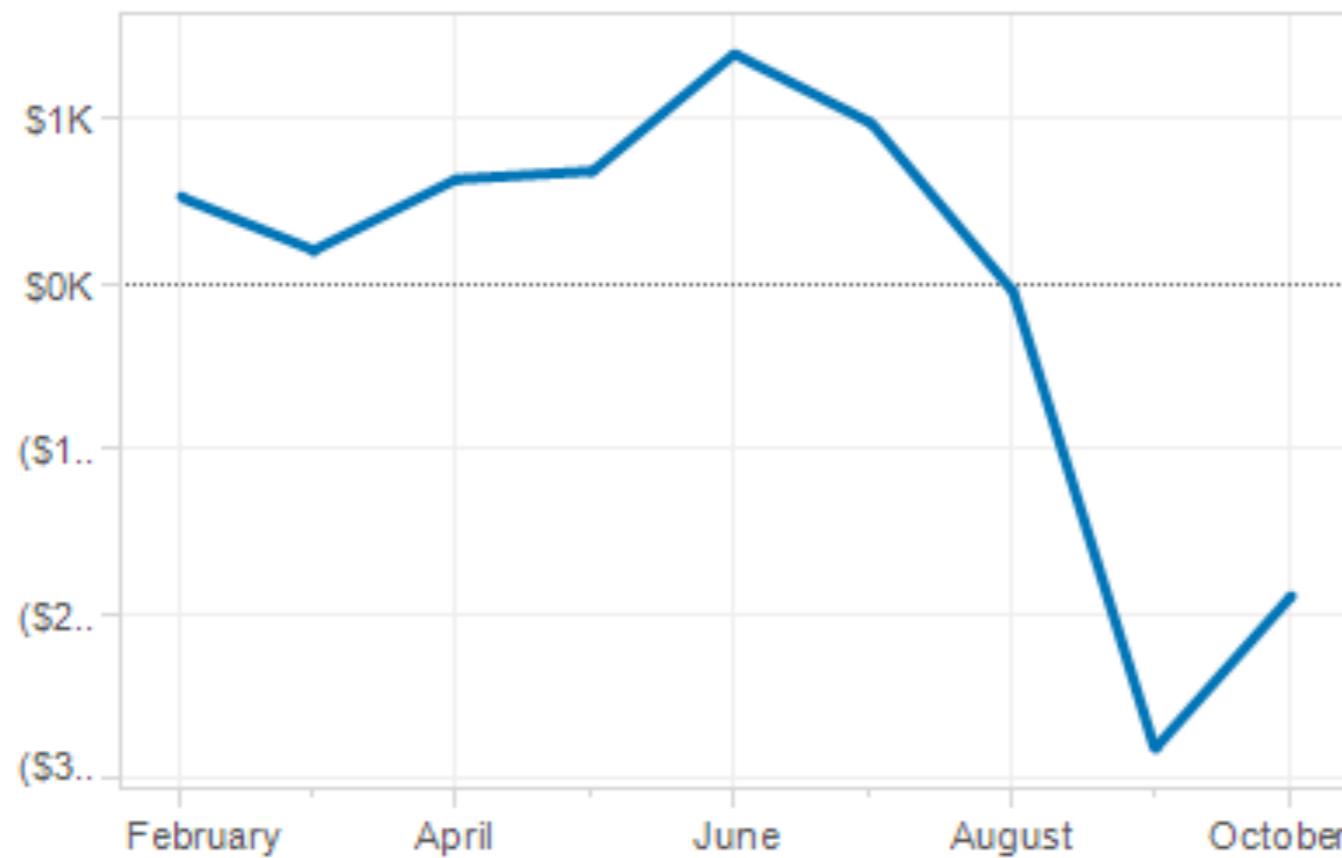
# Don't

## Use bar charts to compare book ratings



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

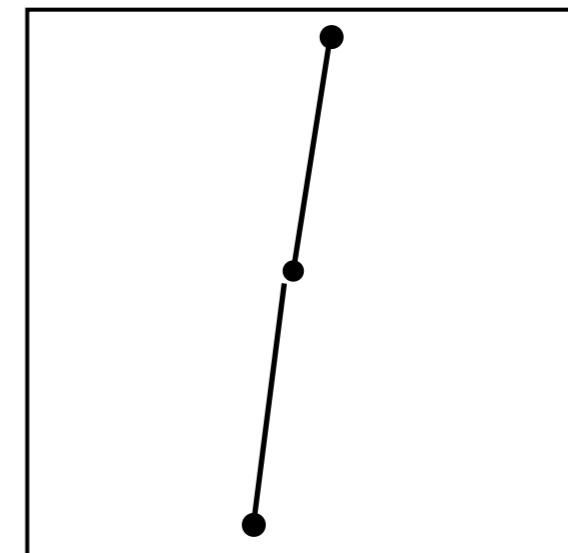
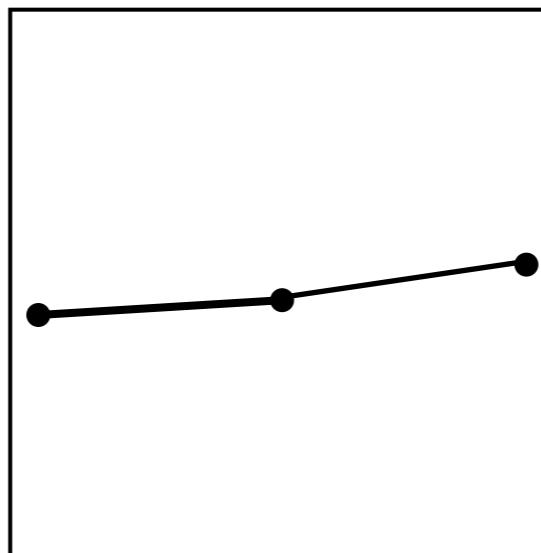
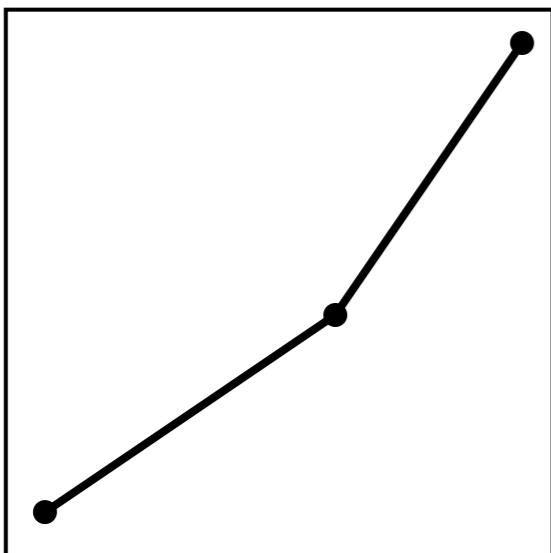
# Aspect Ratios



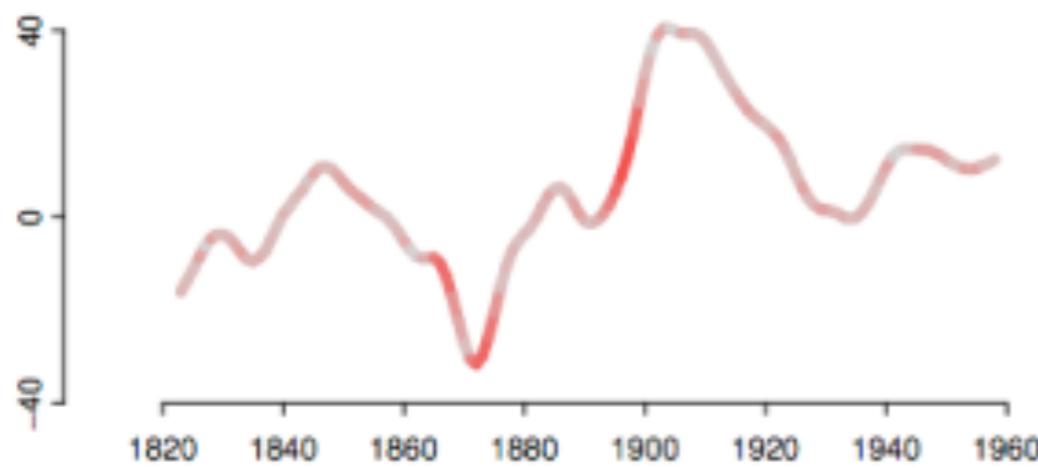
# Banking to $45^\circ$

Two line segments are maximally discriminable when their average absolute angle is  $45^\circ$

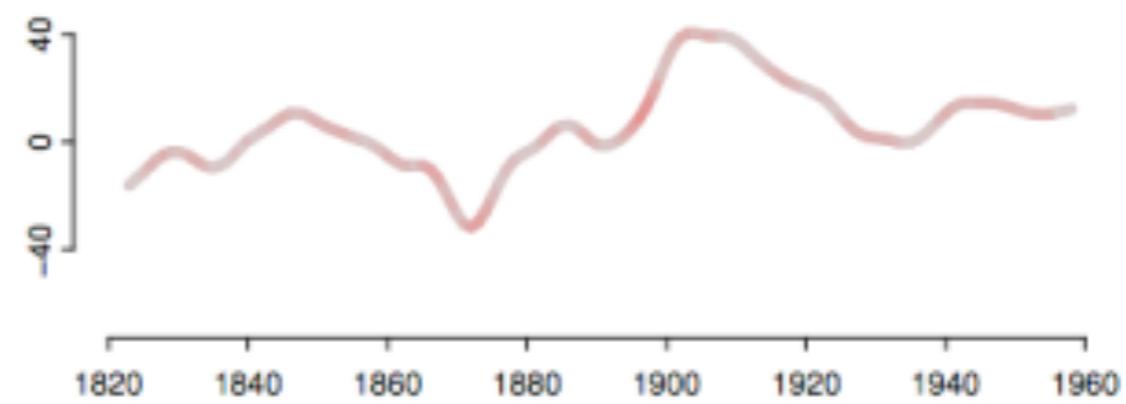
W. Cleveland



# Banking to 45°



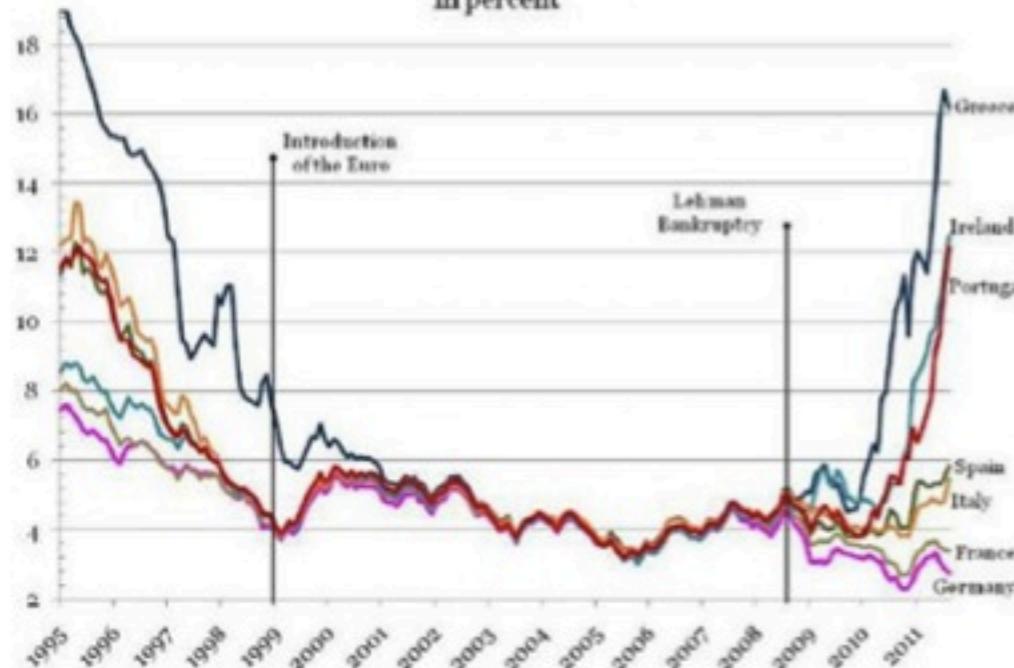
Error Prone



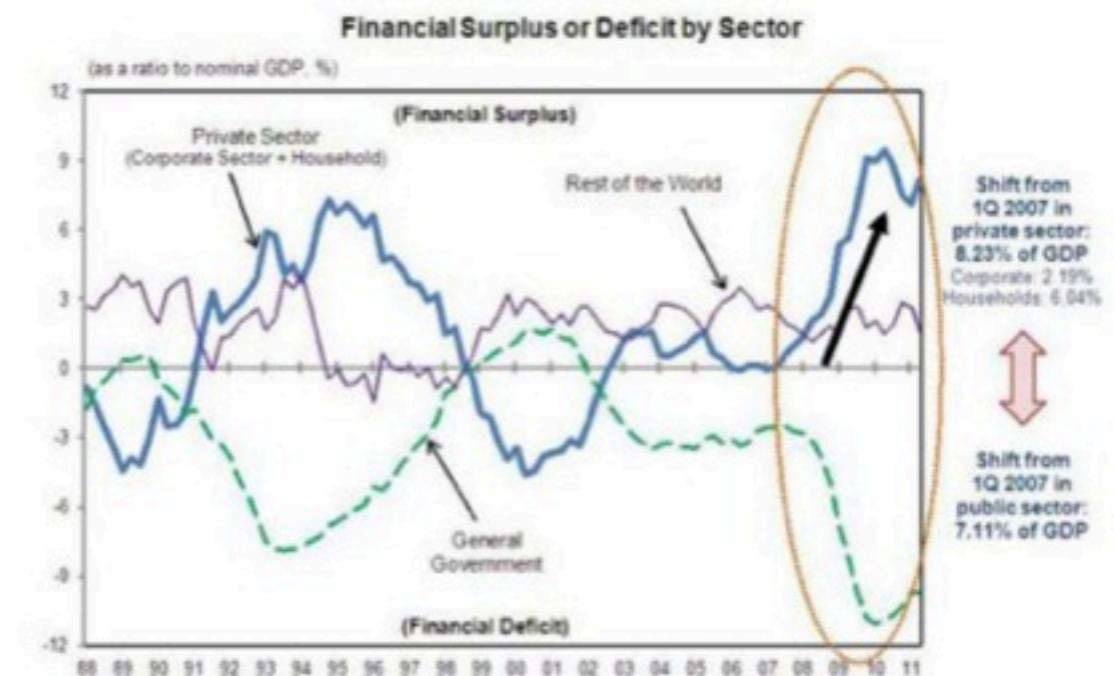
Optimal Aspect Ratio

# Don't

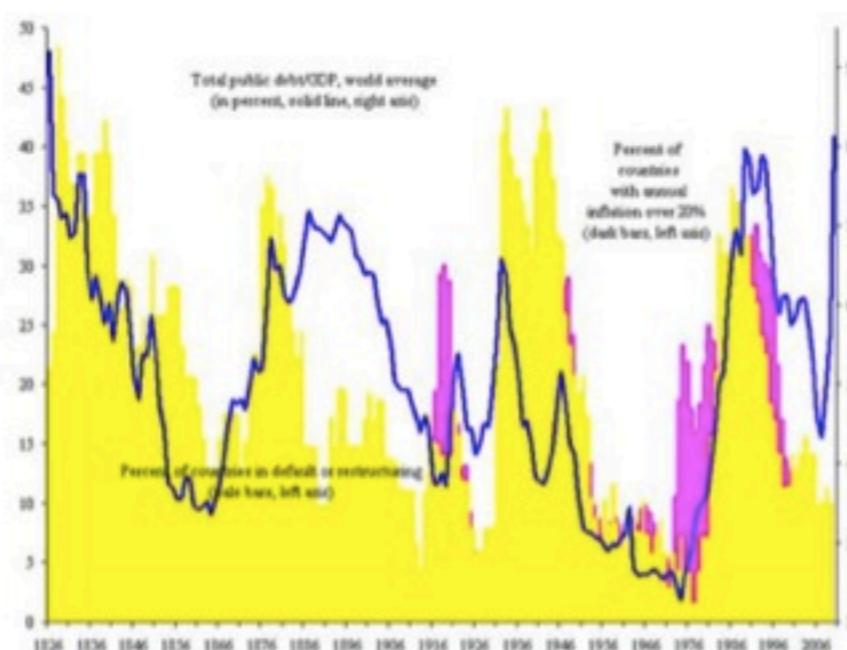
**Interest Rates on 10-Year Government Bonds**  
In percent



**UK in Balance Sheet Recession: UK Private Sector Increased Savings  
Massively after the Bubble**

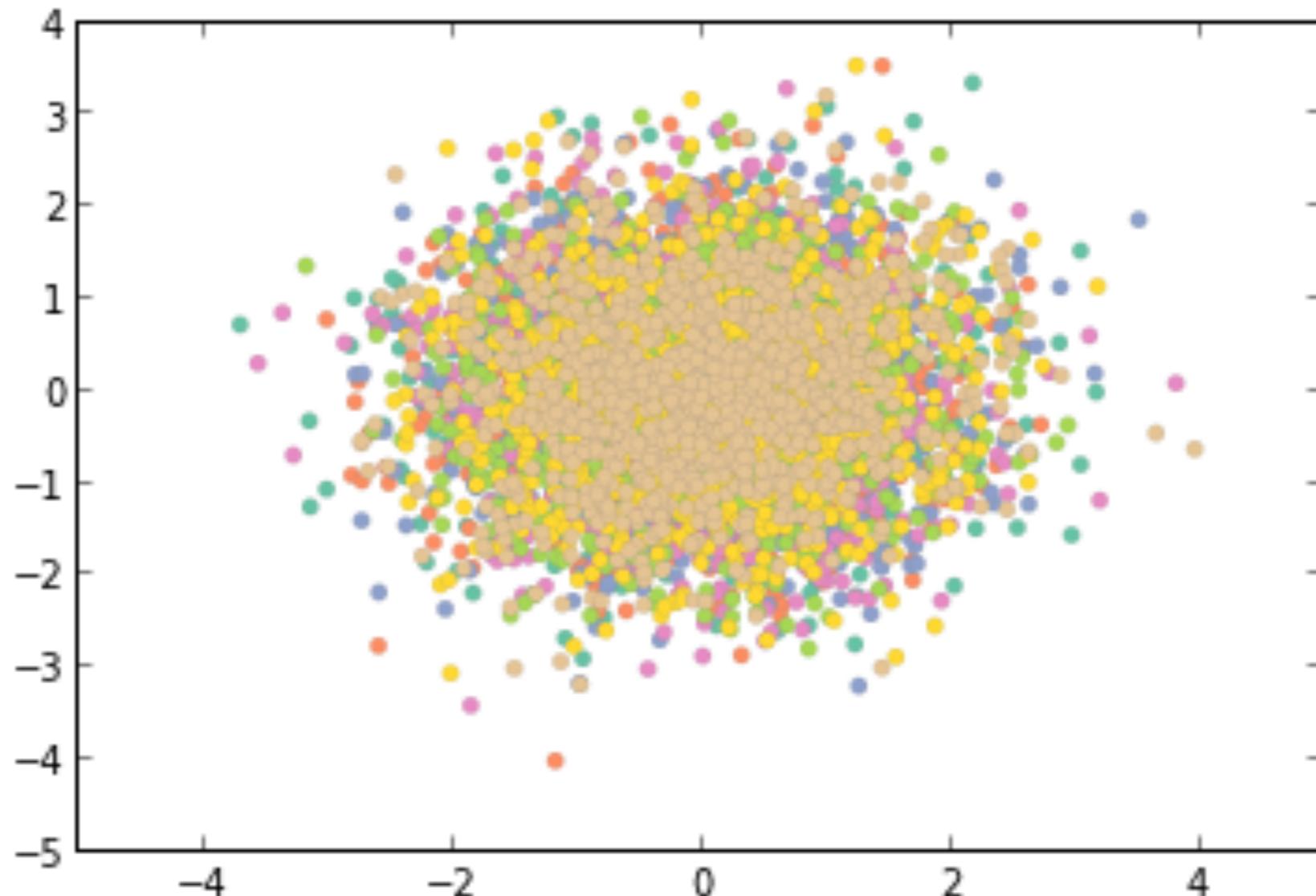


Note: For the latest figures, 4 quarter averages ending with 2Q/11 are used.  
Source: Office for National Statistics, UK



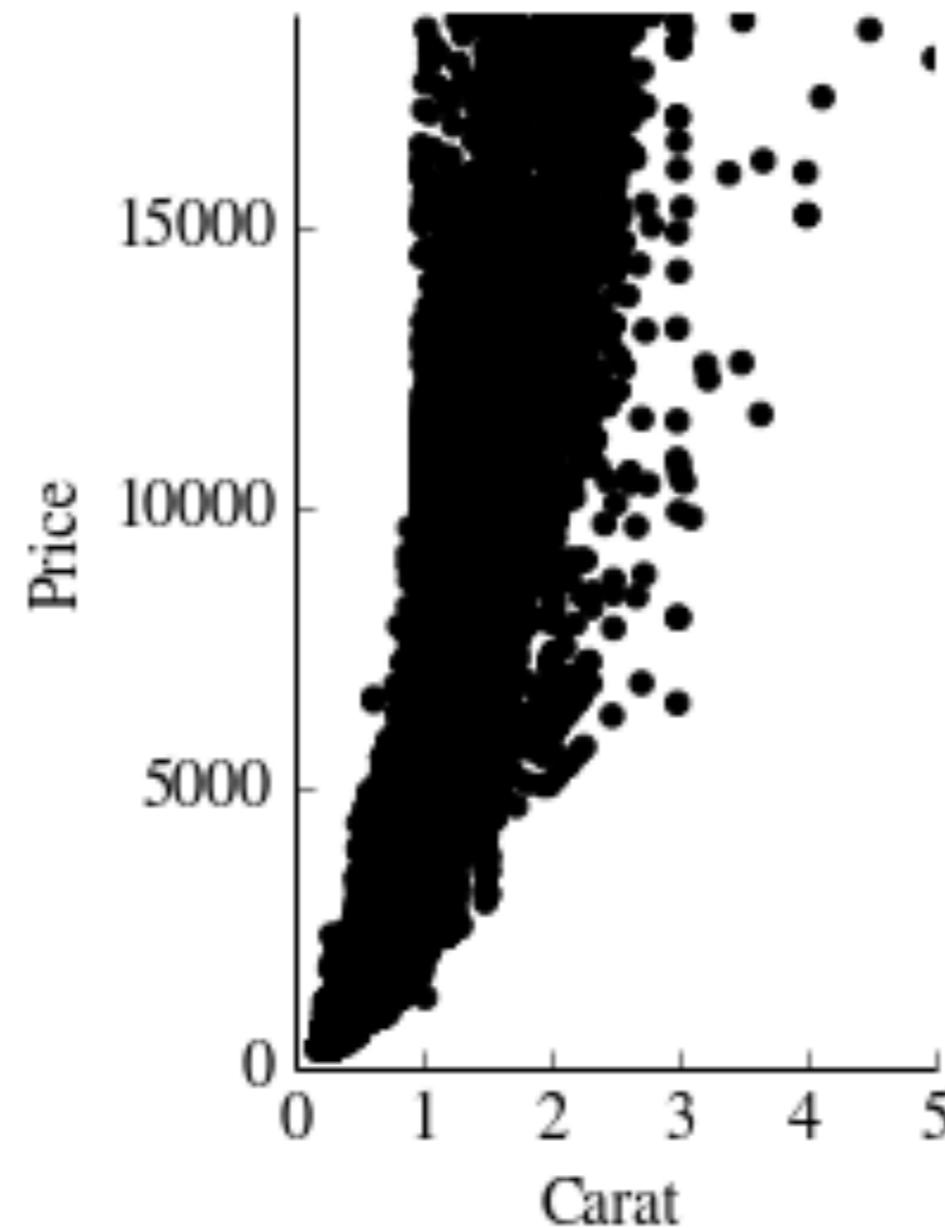
# Correlations

# Scatterplots

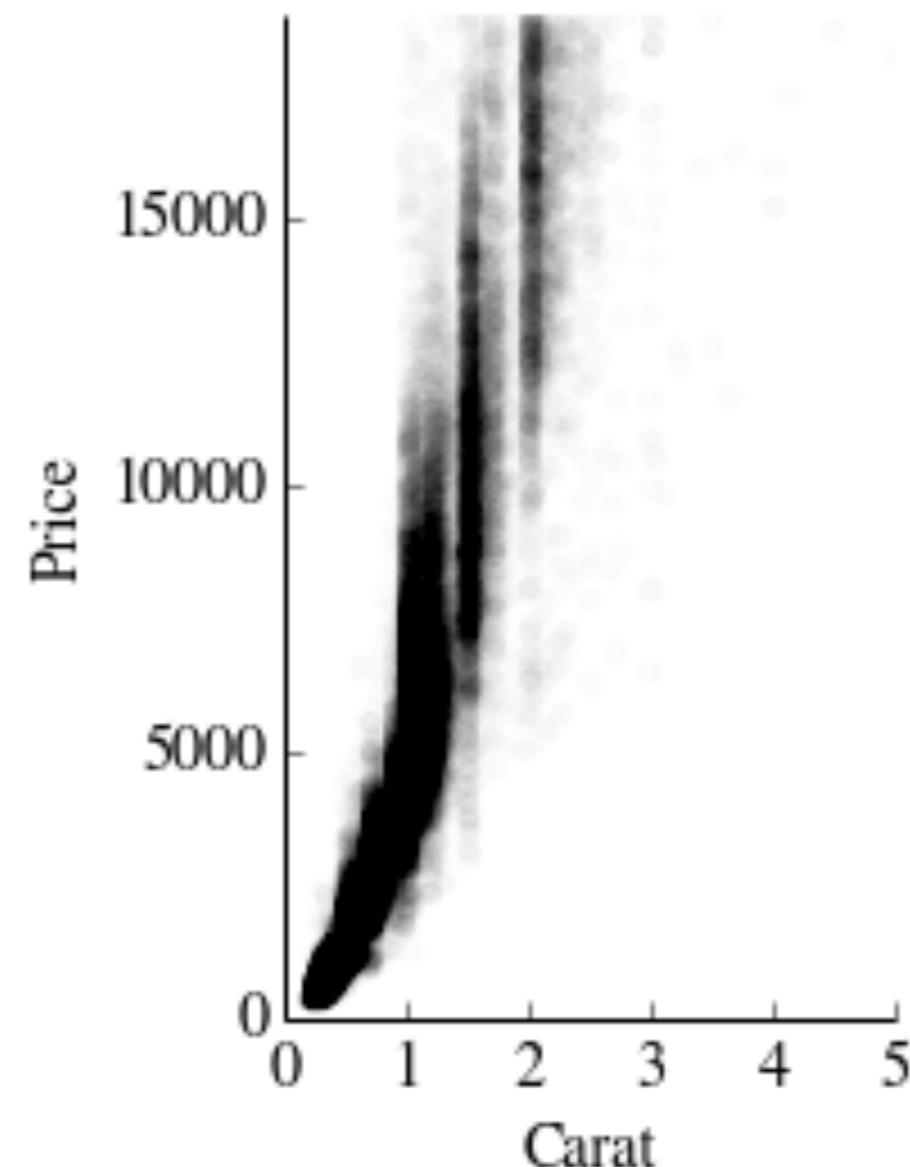
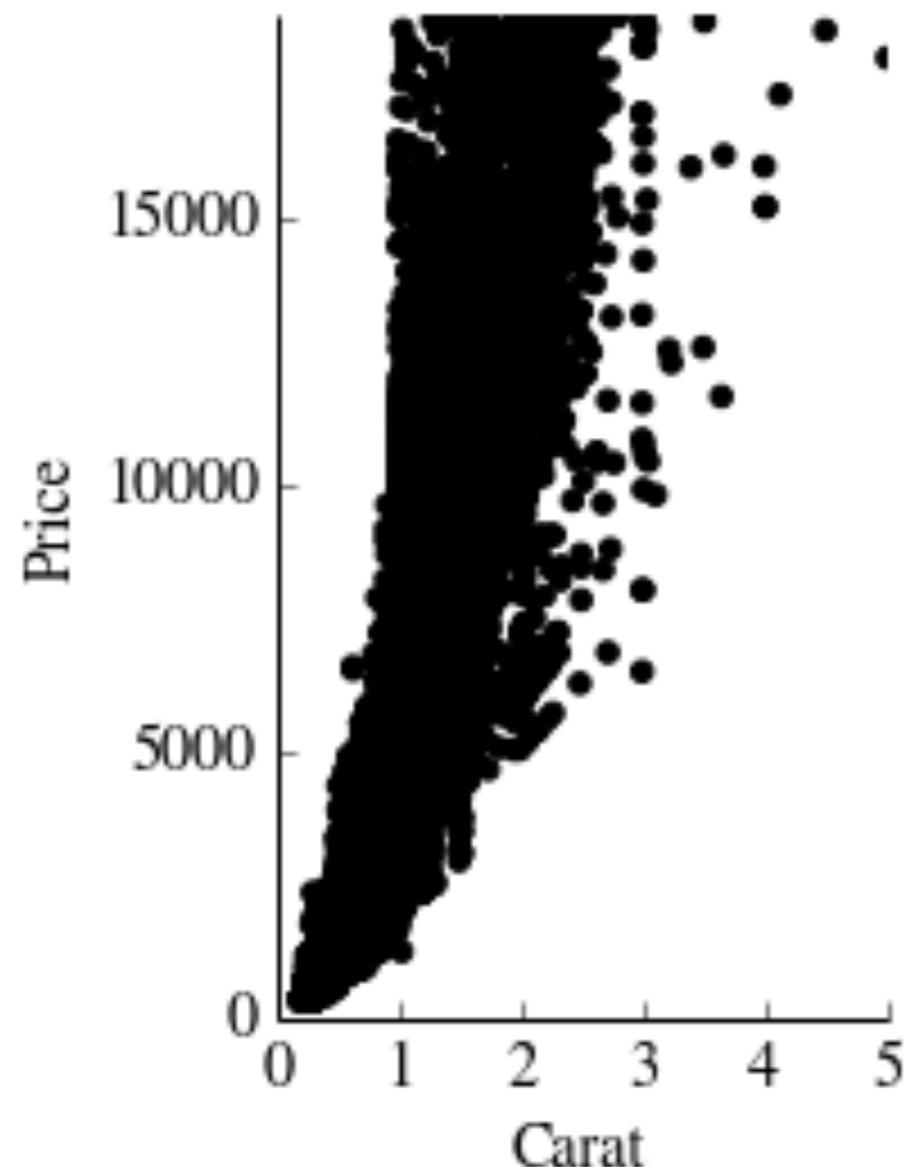


Light Grey Border

# Overplotting

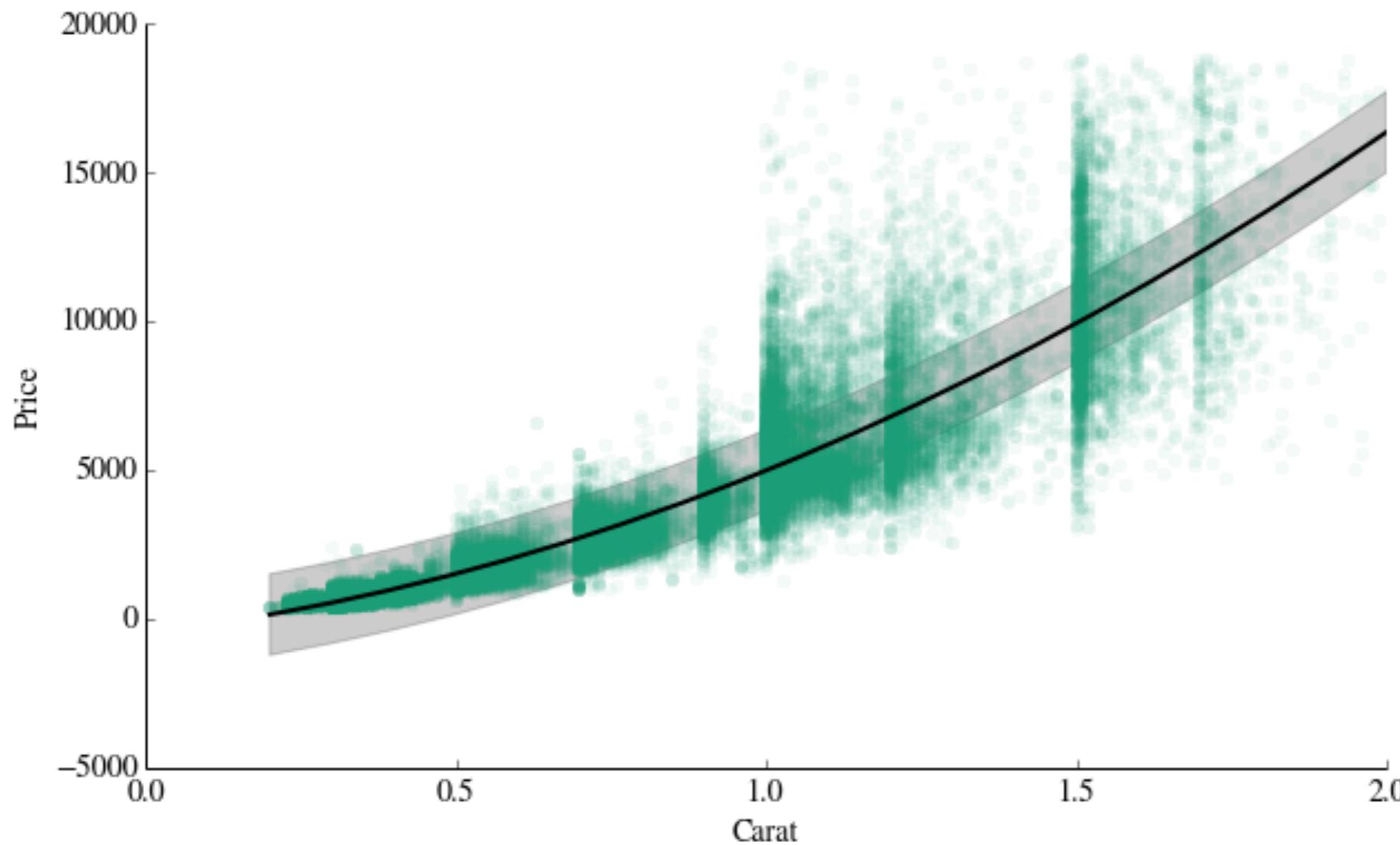


# Overplotting

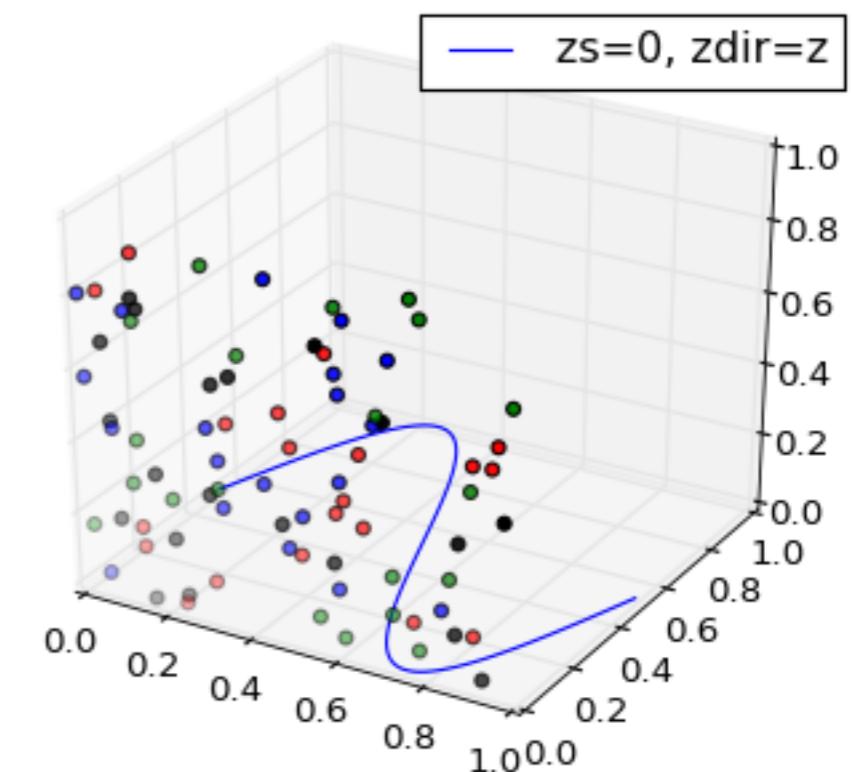
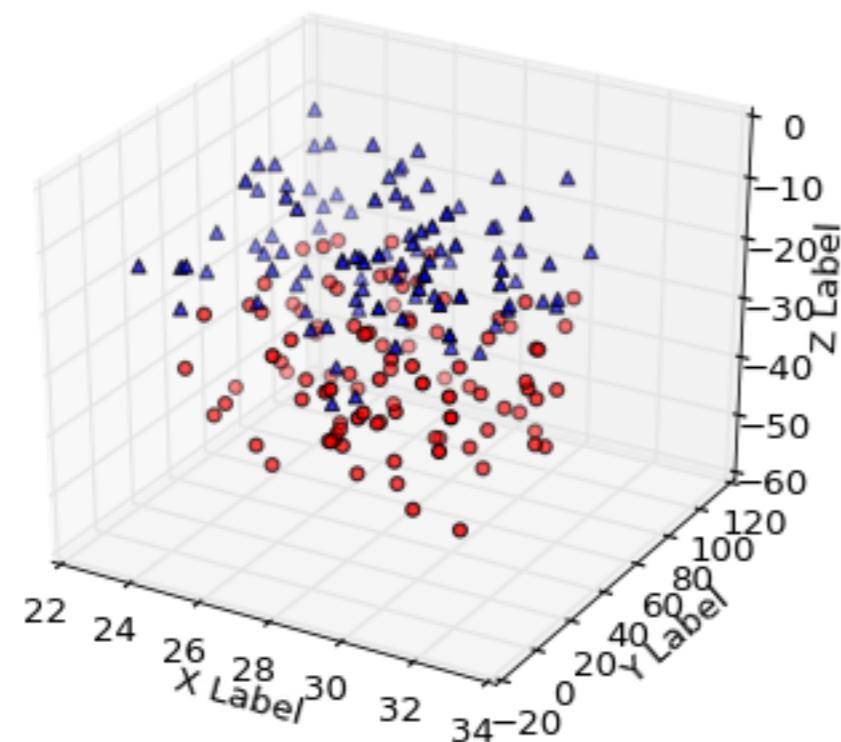


$\text{alpha} = 1/100$

# Trend Lines



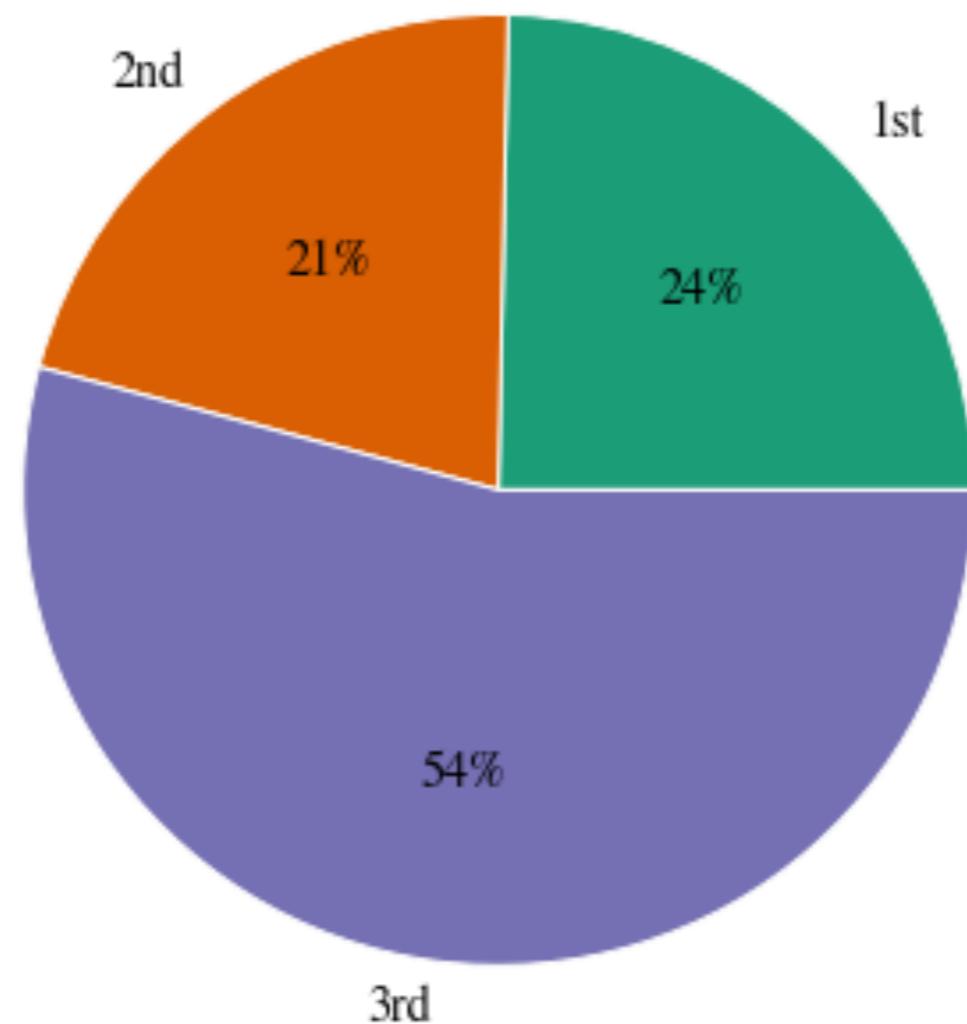
# Don't



# Compositions

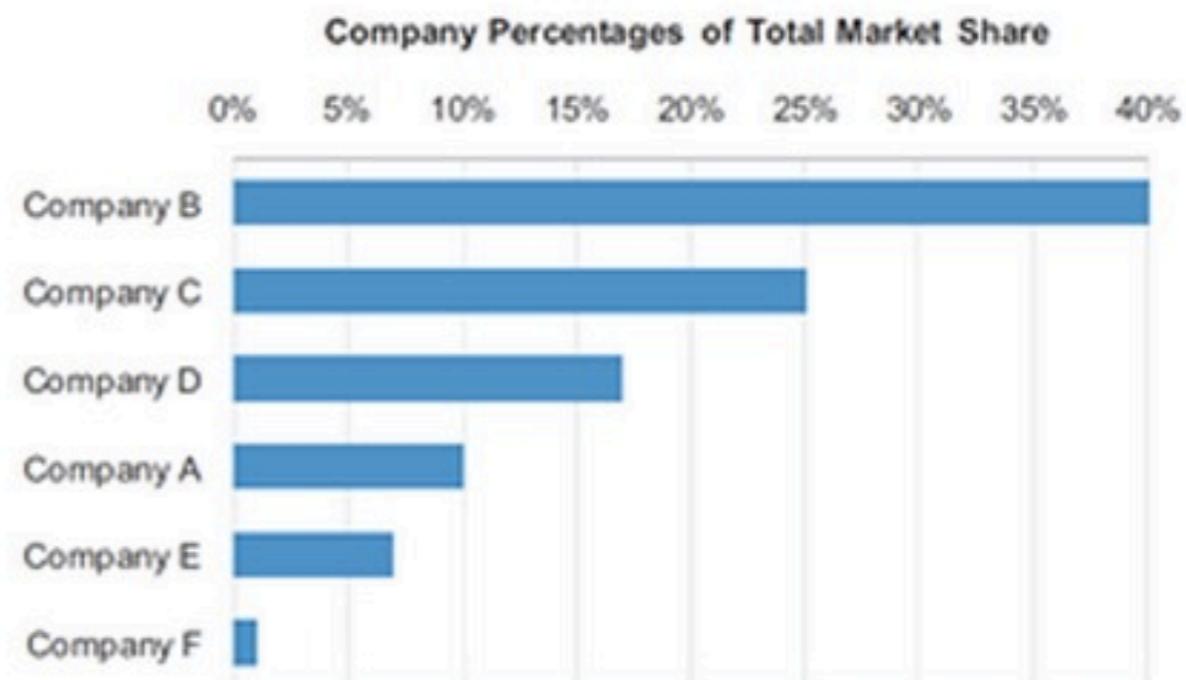
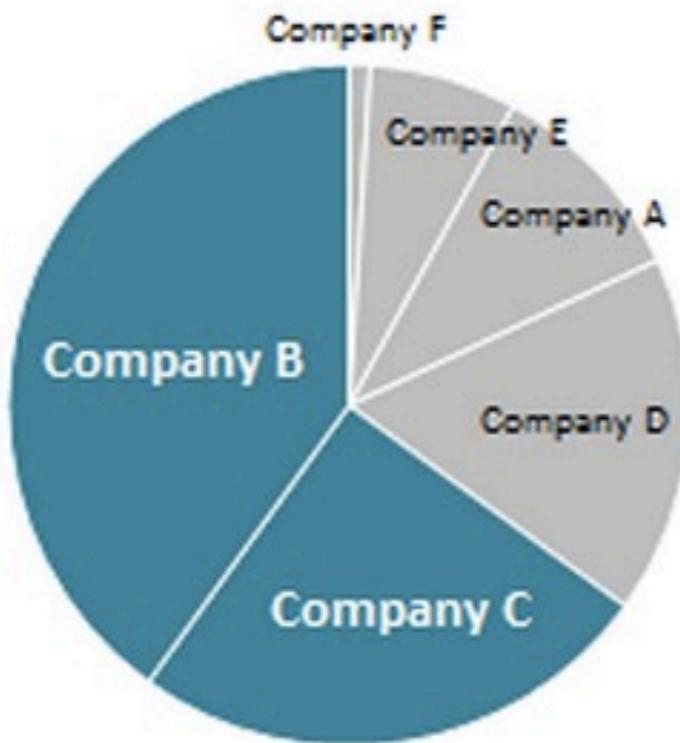
# Pie Charts

Passenger Class on the Titanic



# Pie vs. Bar Charts

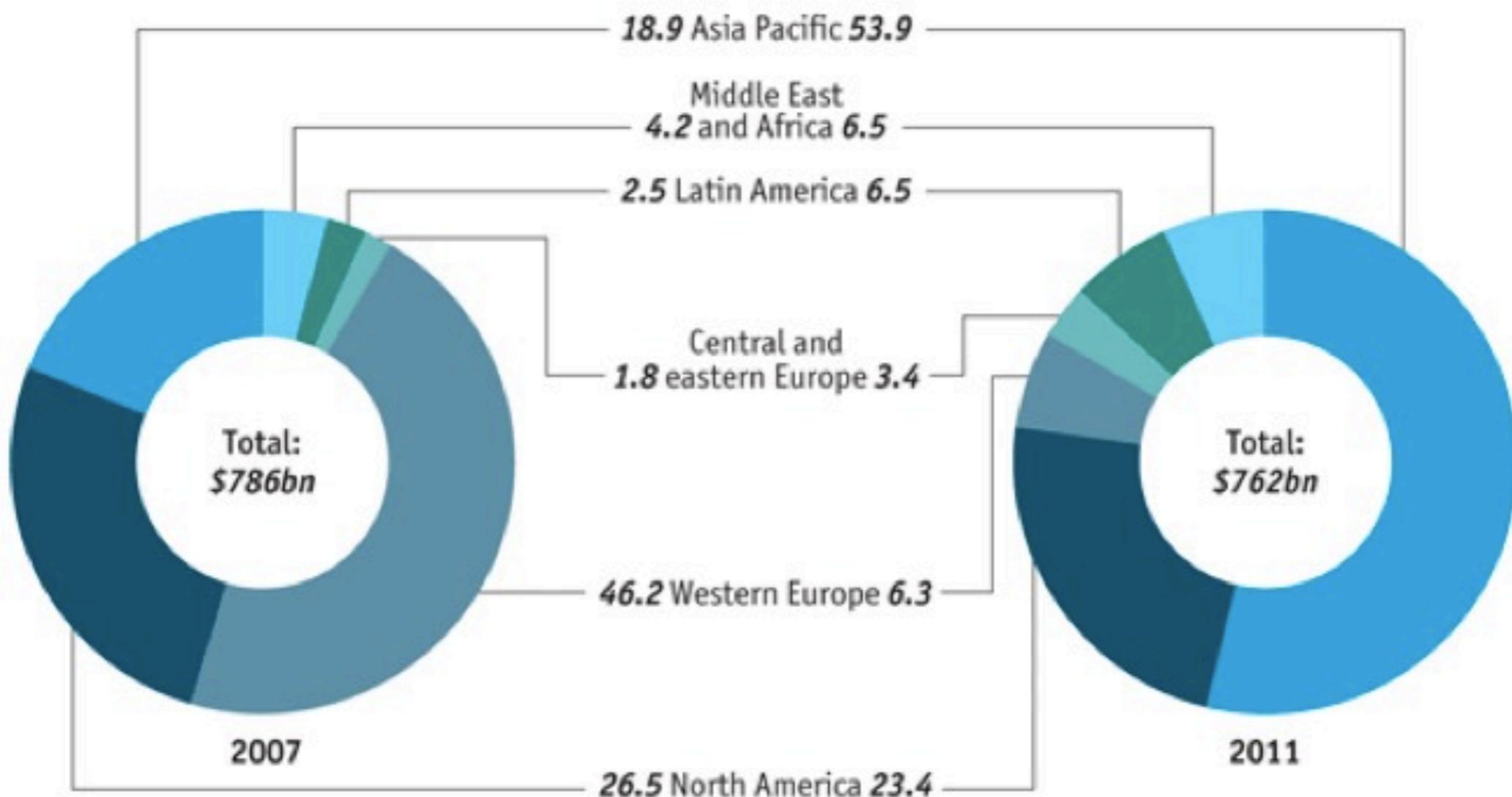
65% of the market is controlled by companies B and C



# Donut Chart

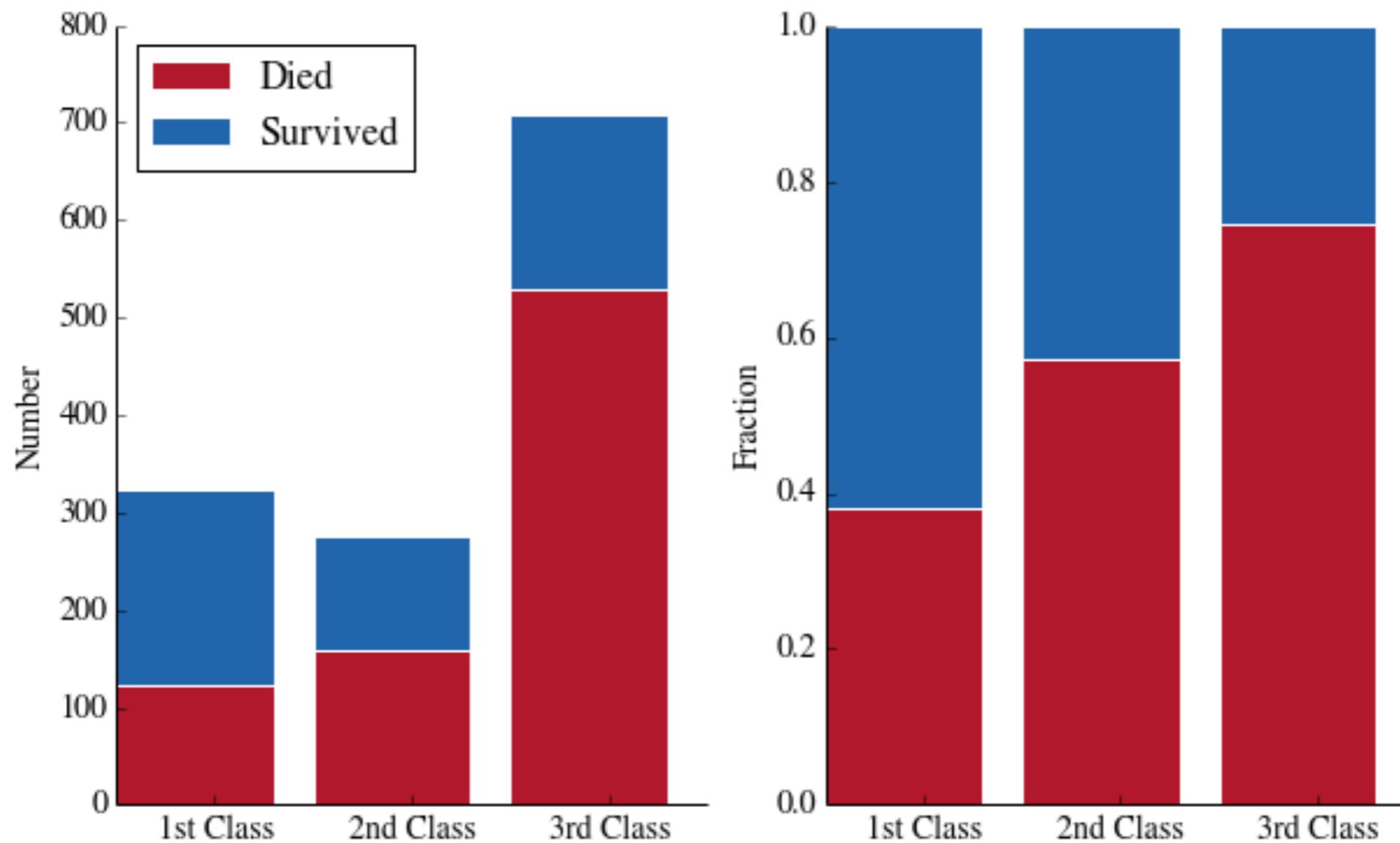
## Pre-tax profits of the 1,000 largest banks

By tier-one capital and domicile, % of total

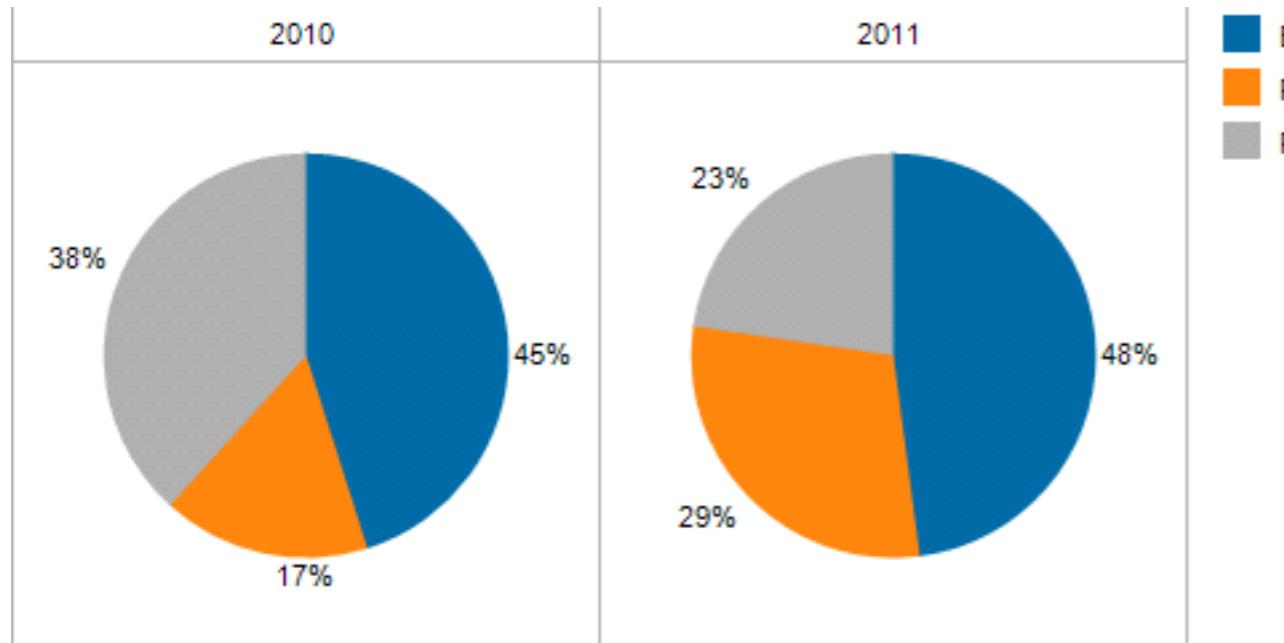


Source: *The Banker Top 1000*

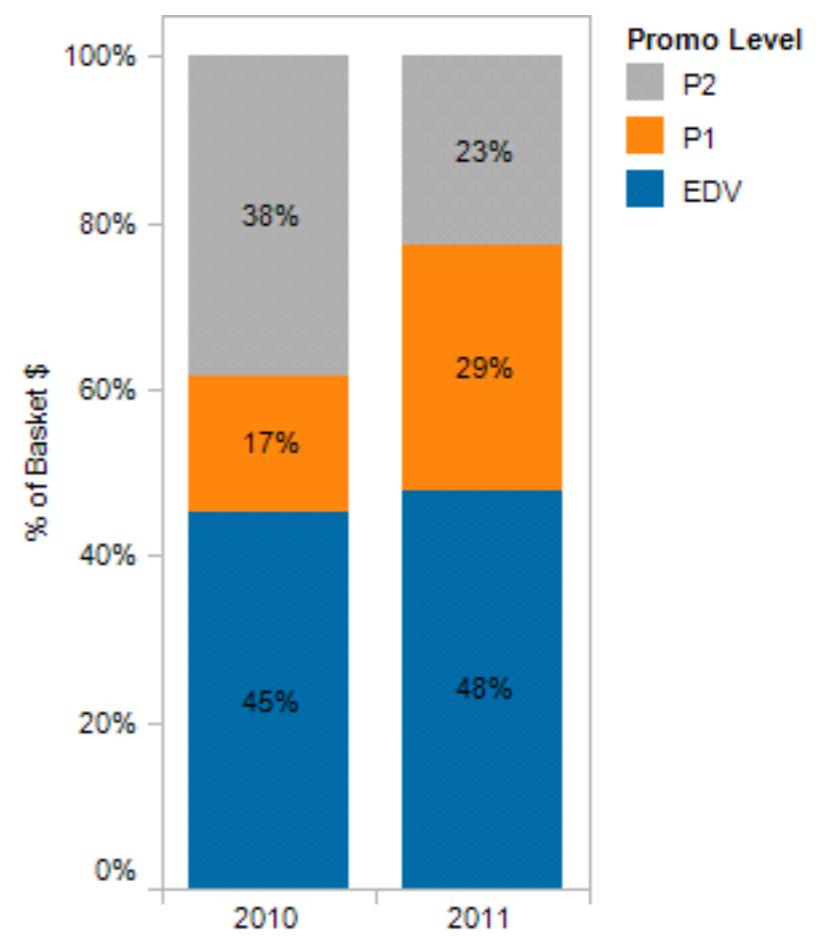
# Stacked Bar Chart



# Stacked Bar Chart



V.S.



# Don't

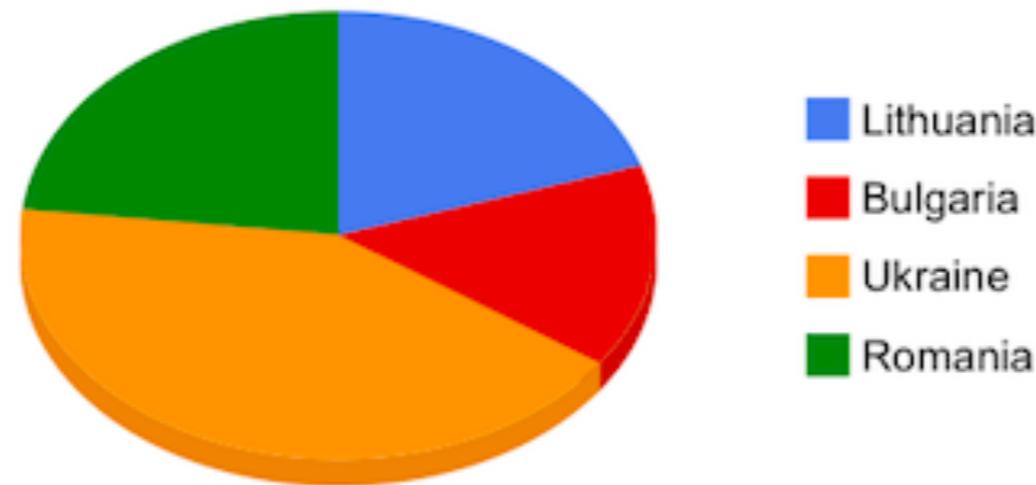
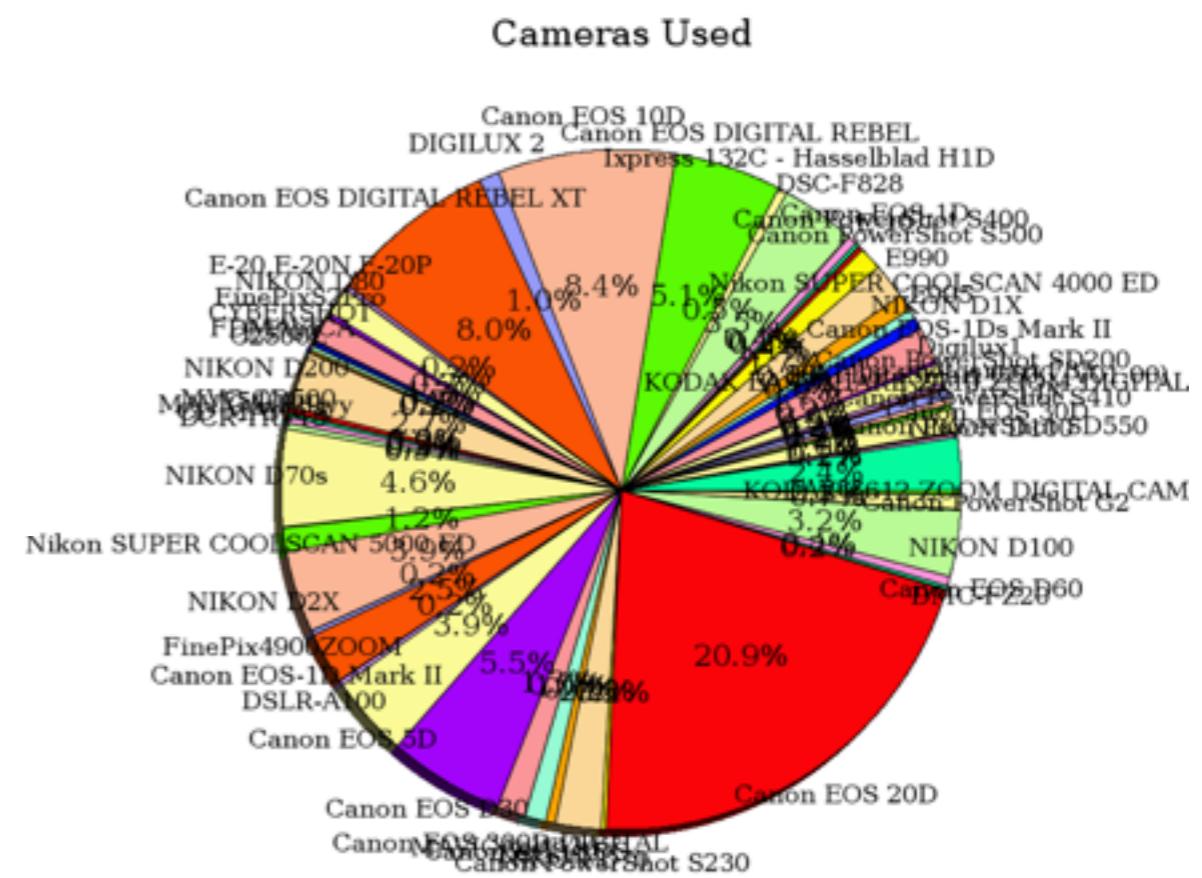
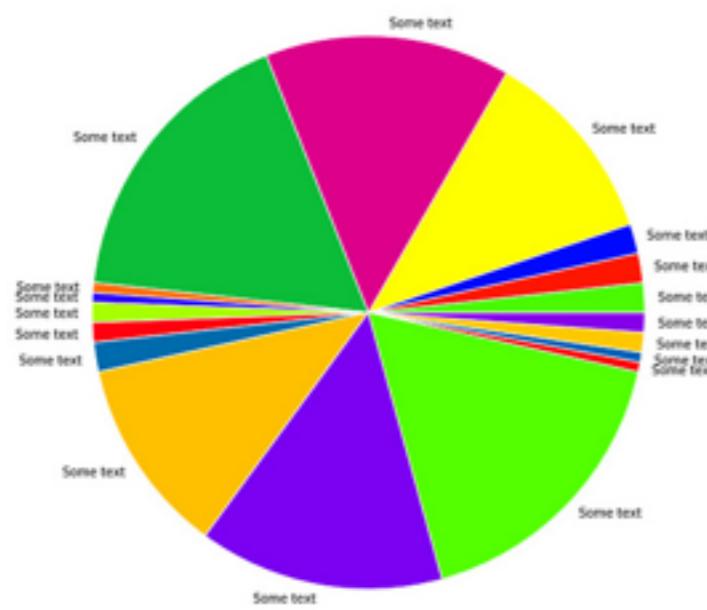
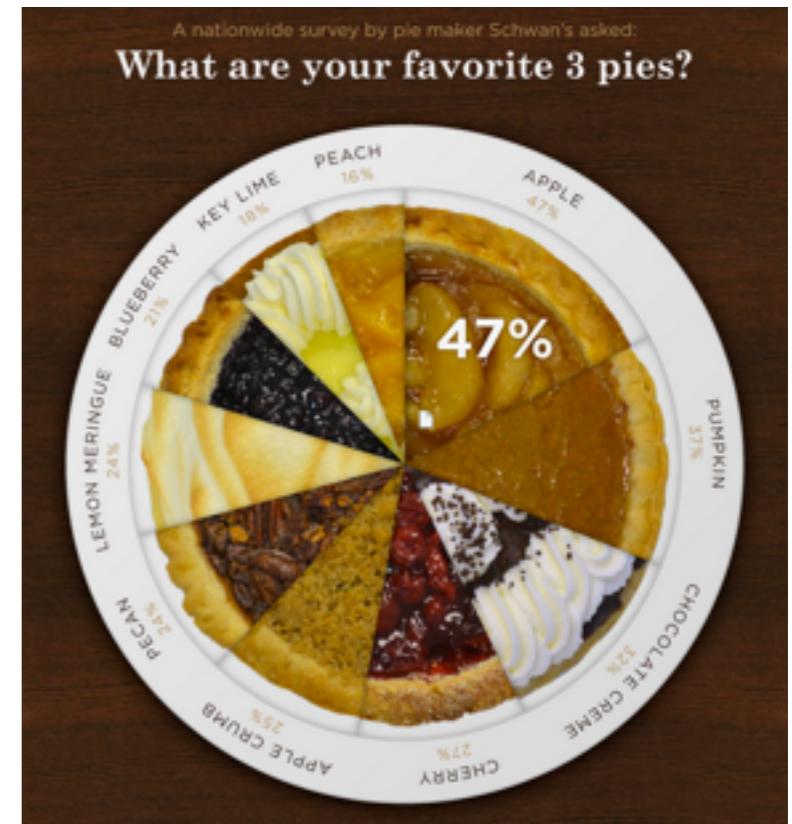
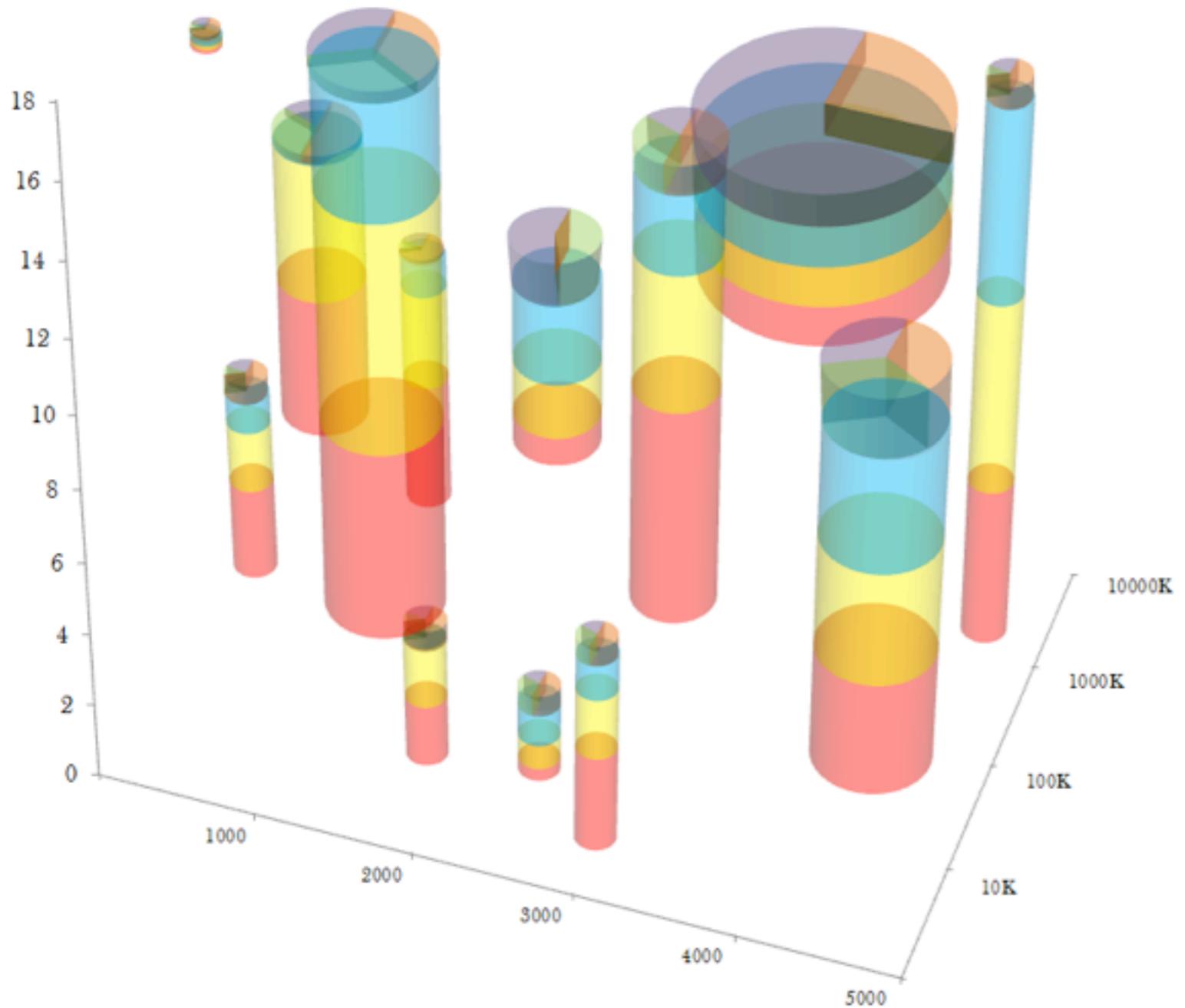


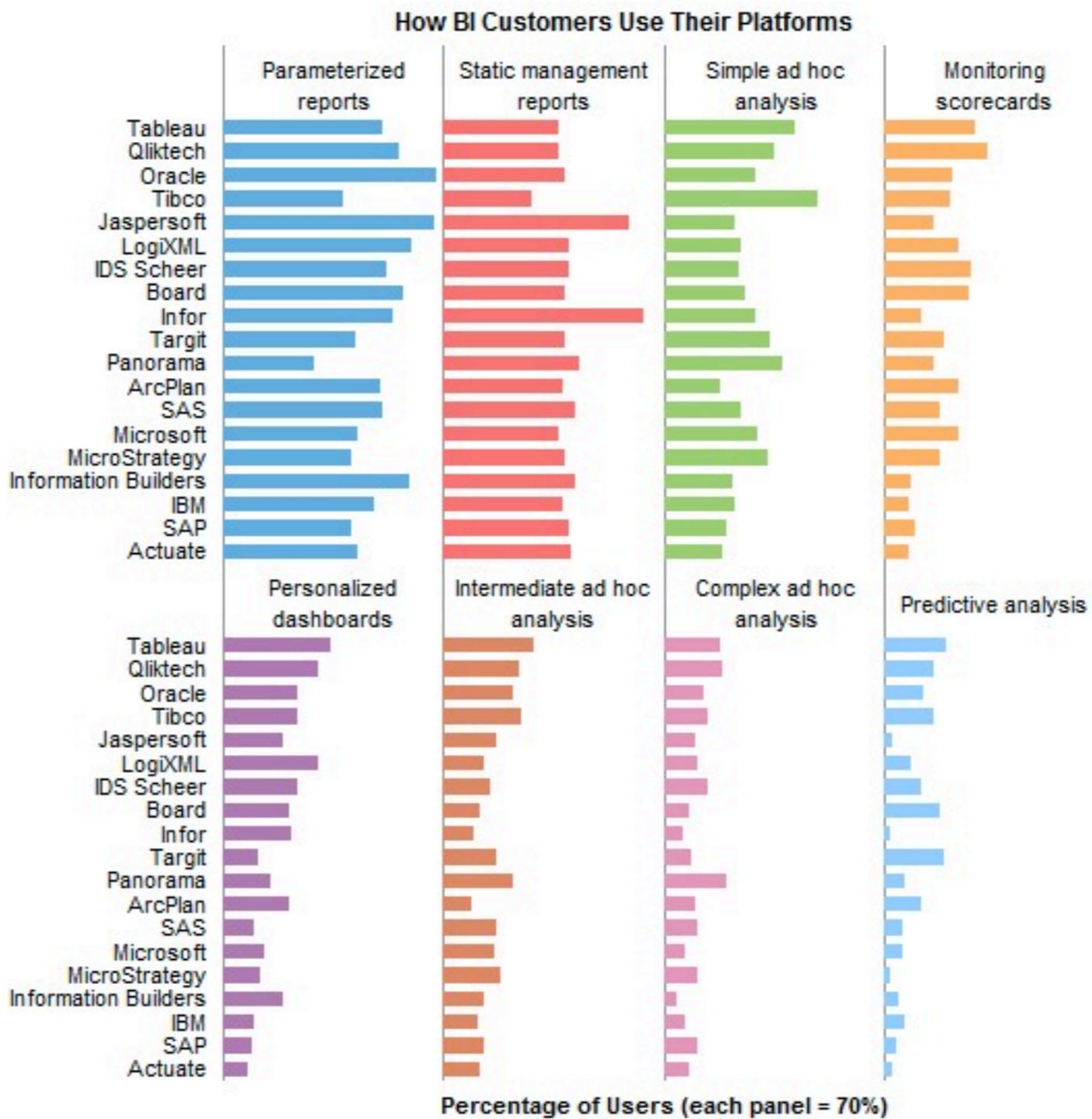
Figure 5



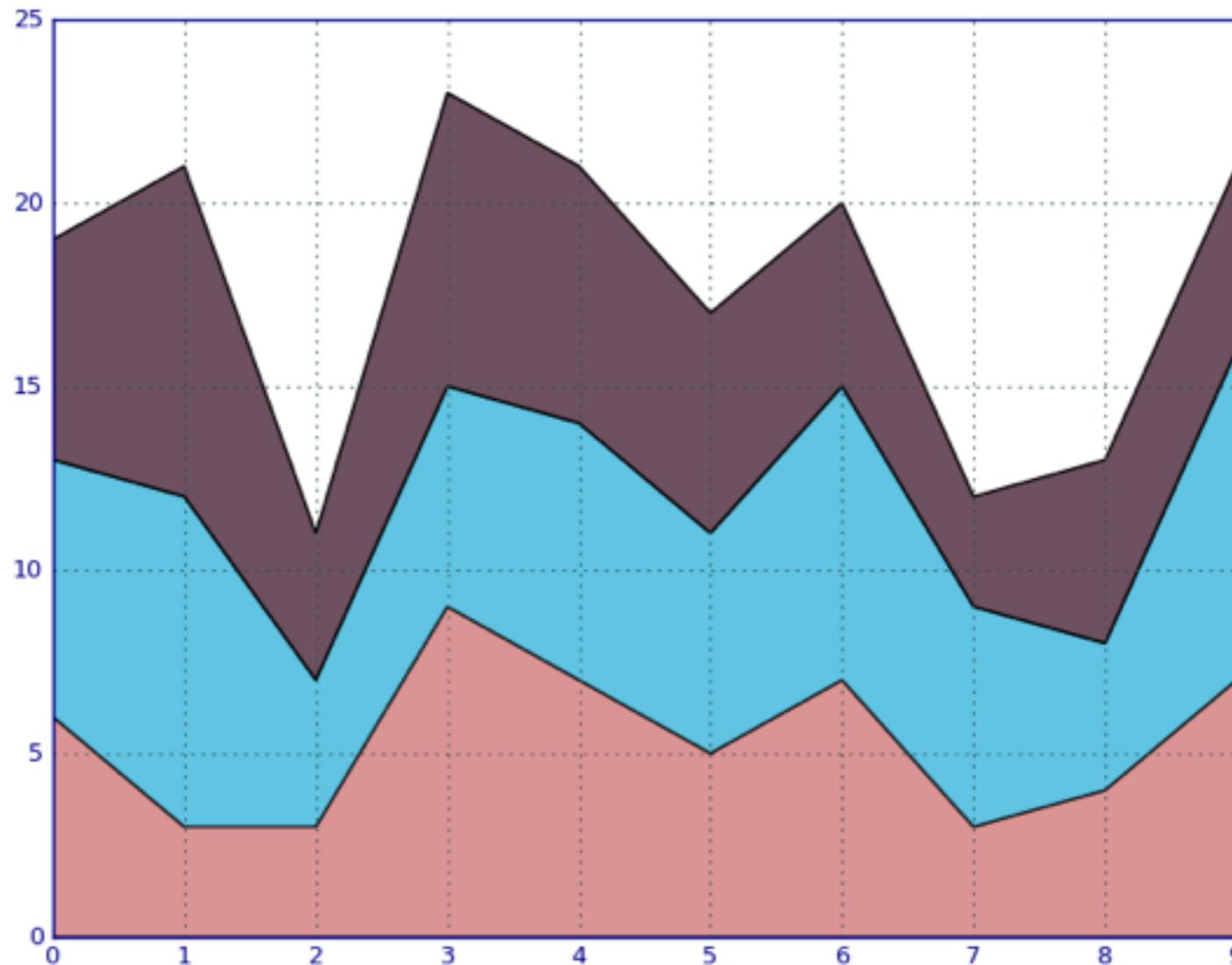
# eagerpies.com



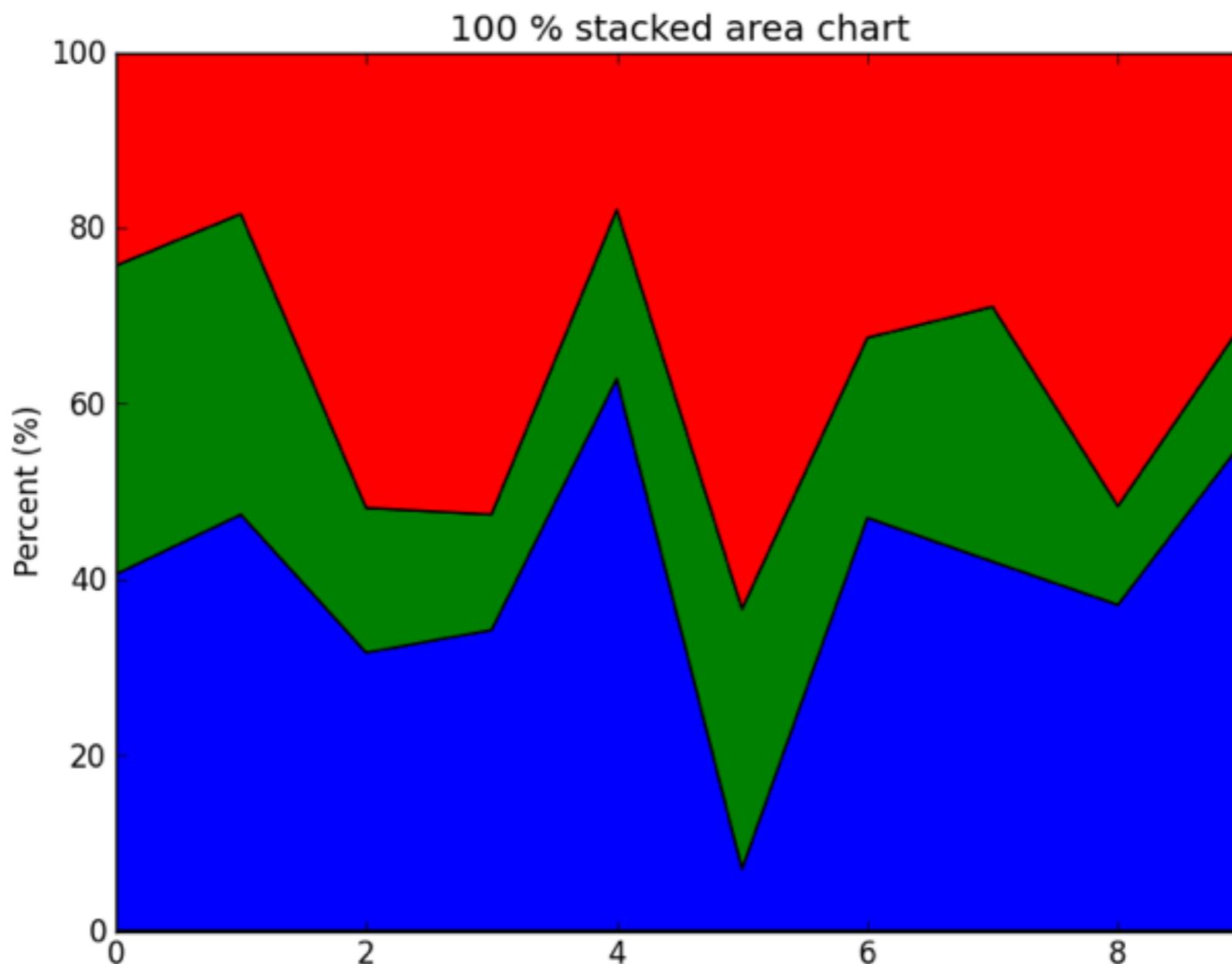
# Small Multiples



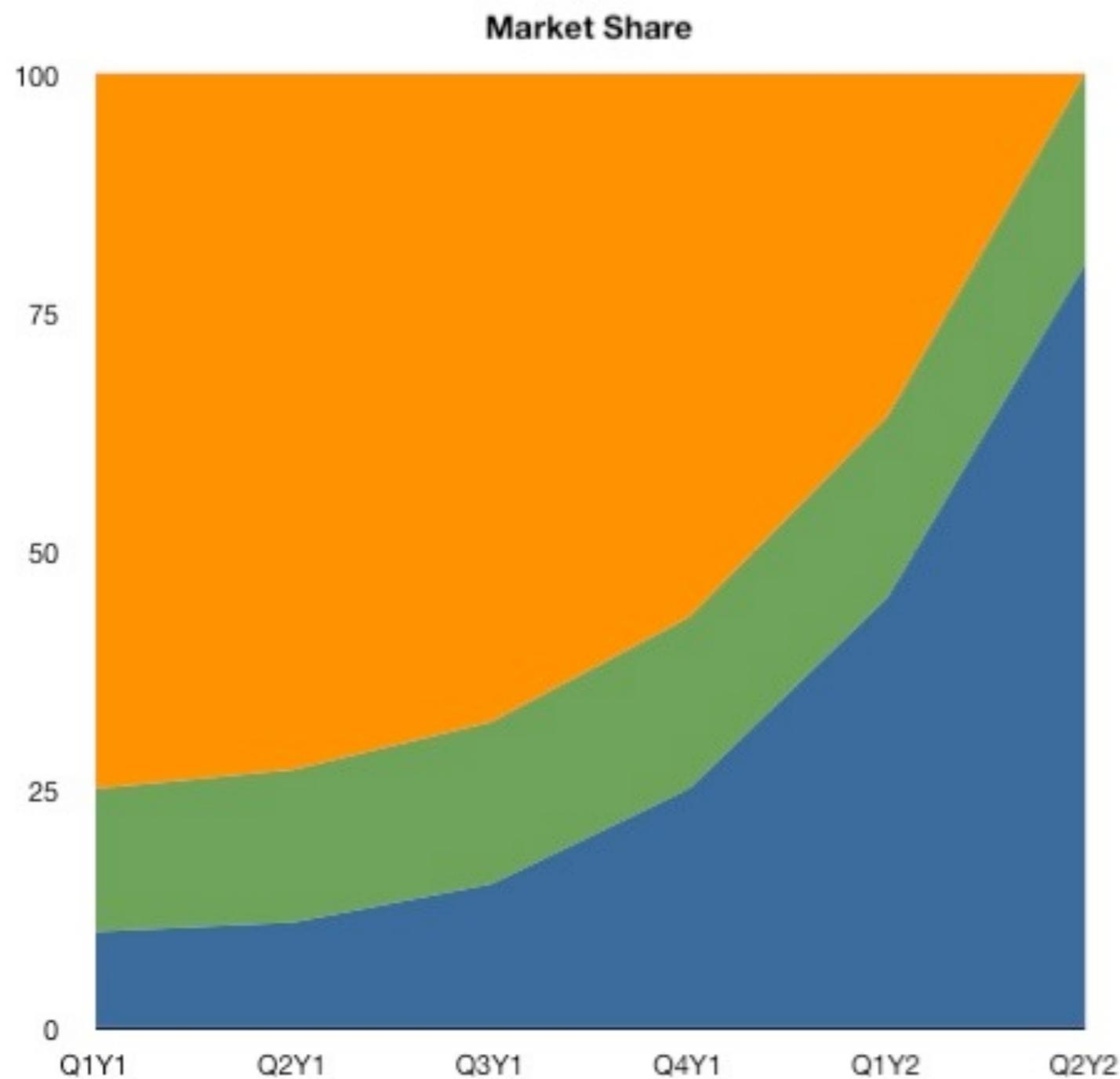
# Stacked Area Chart



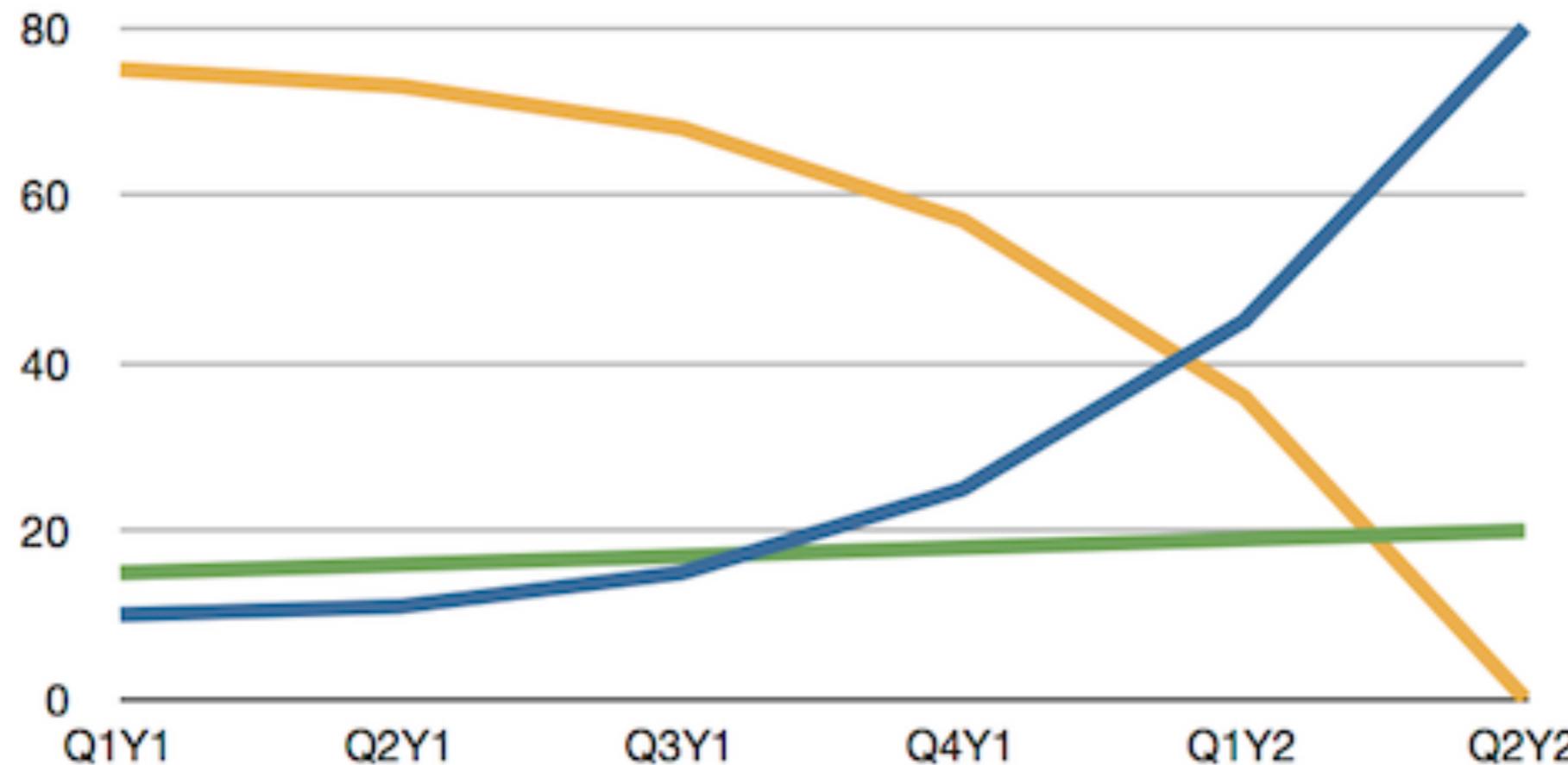
# 100% Stacked Area Chart



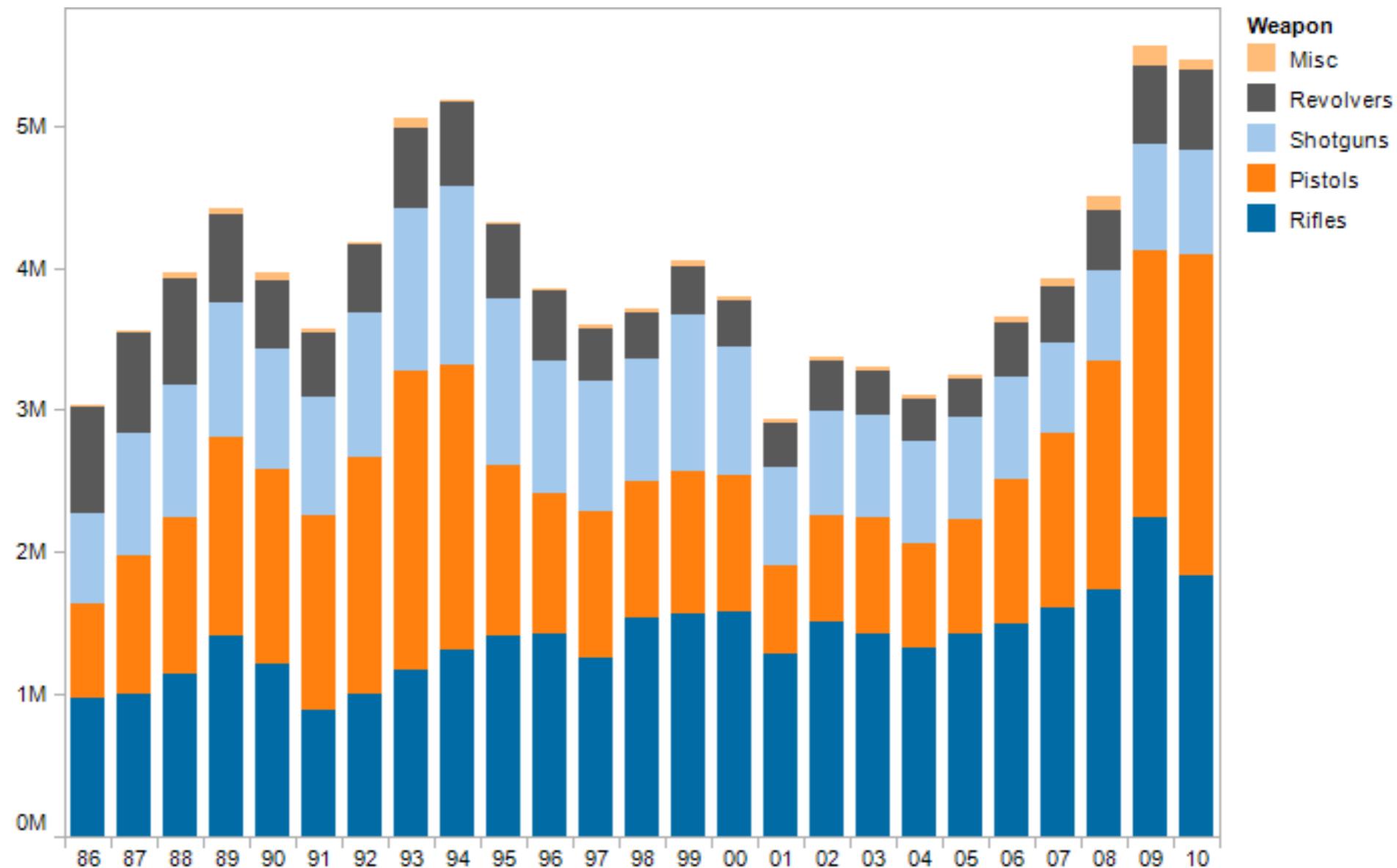
# Stacked Area vs. Line Graphs



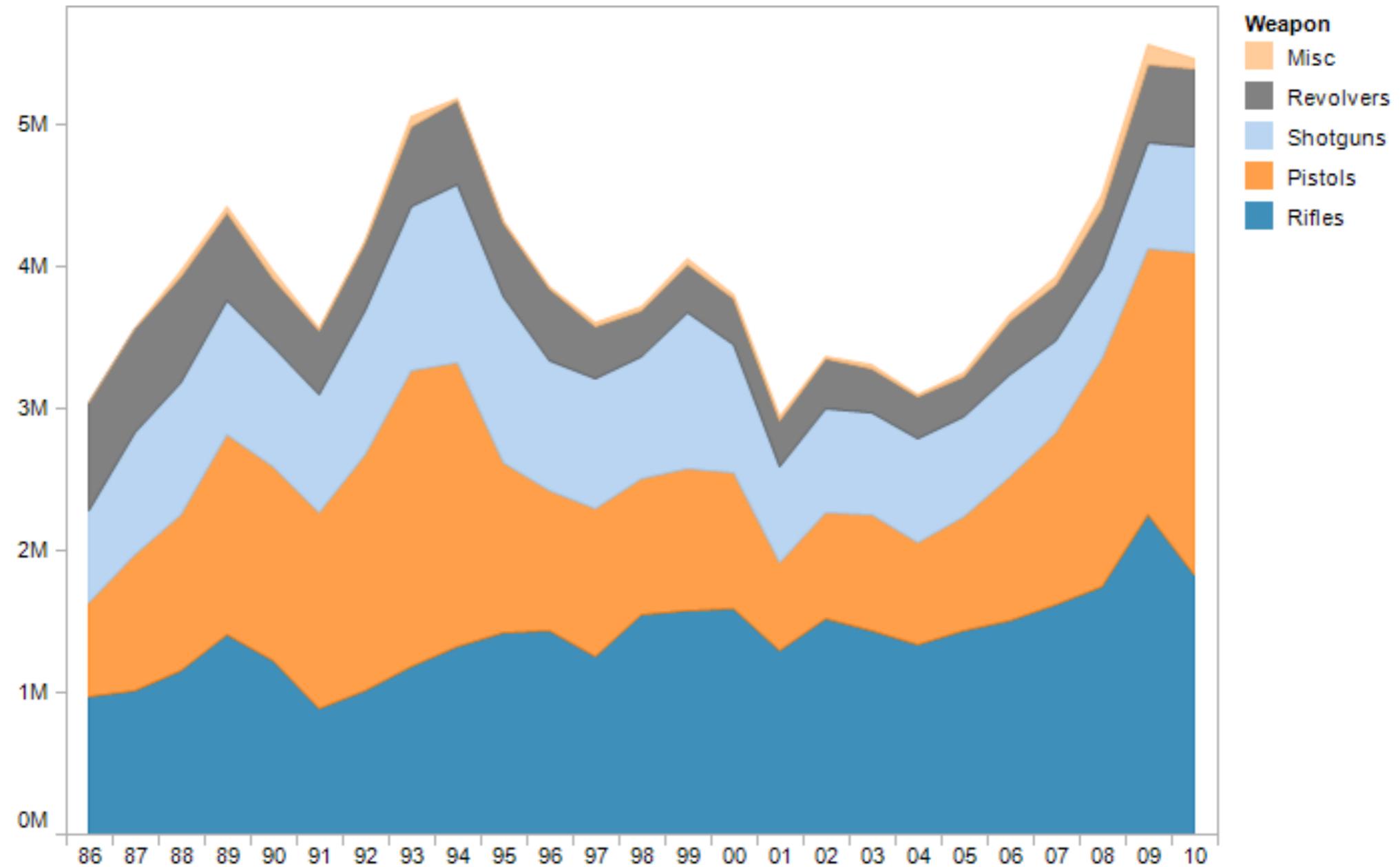
# Stacked Area vs. Line Graphs



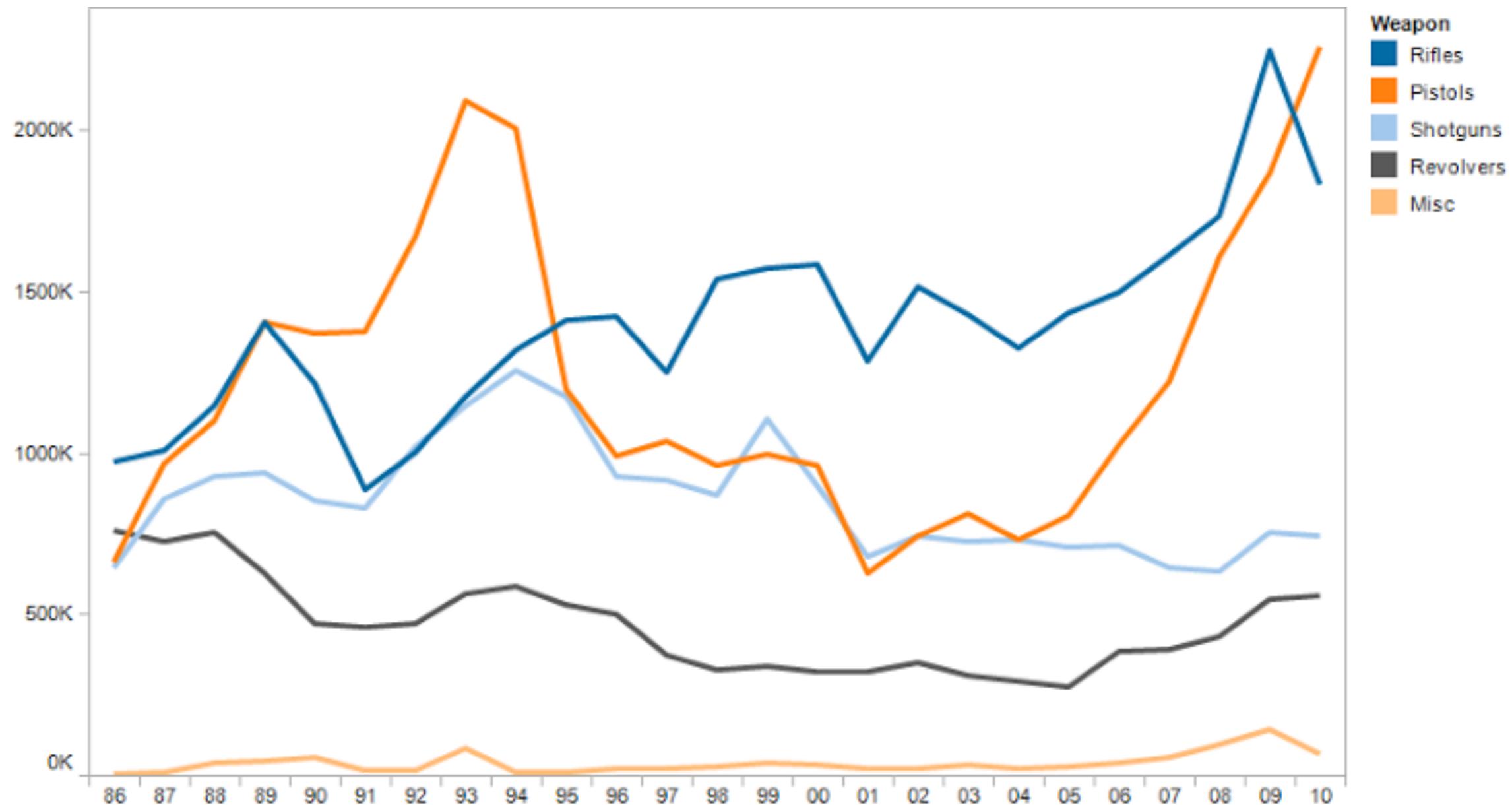
# Stacked Area vs. Line Graphs



# Stacked Area vs. Line Graphs

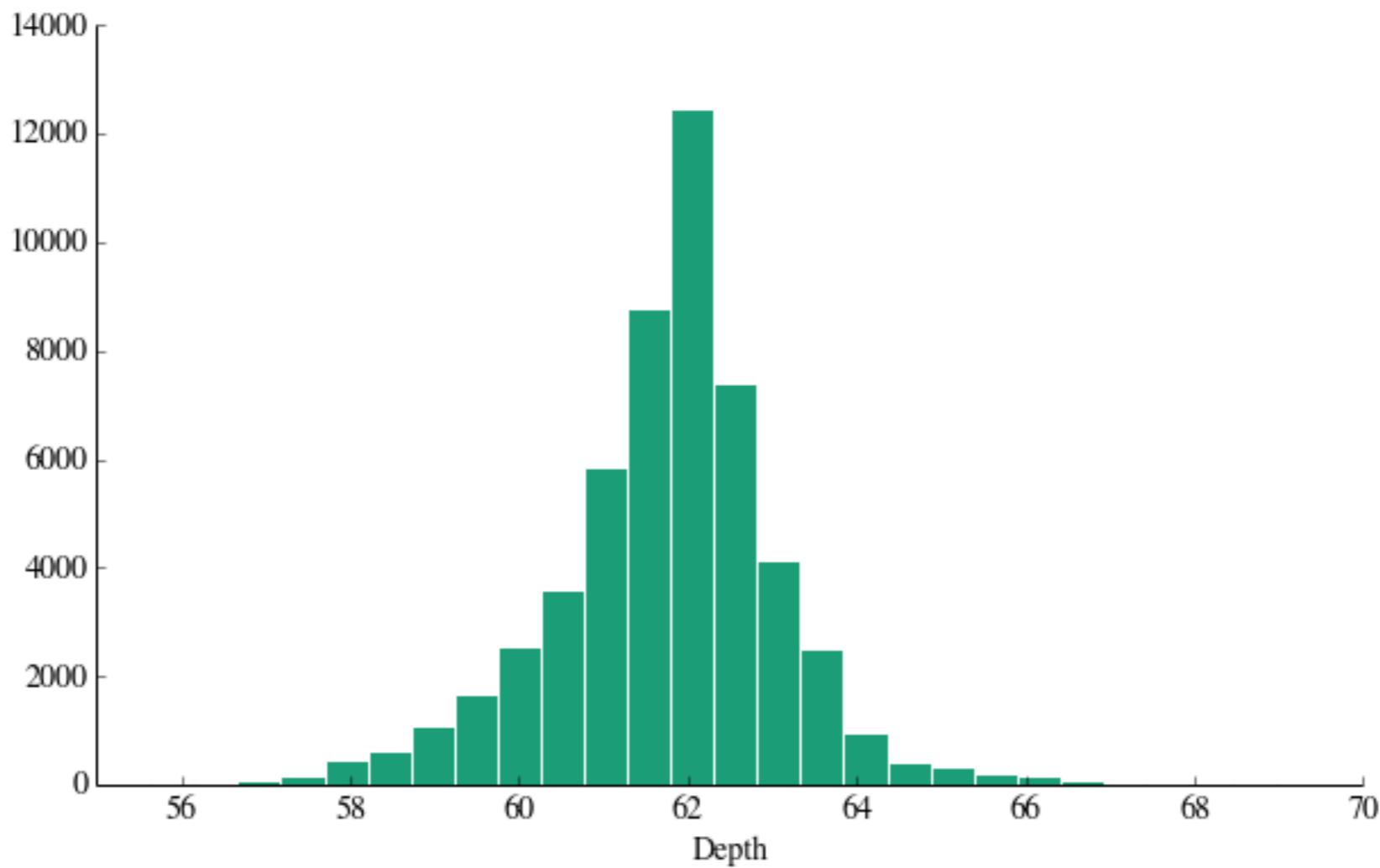


# Stacked Area vs. Line Graphs

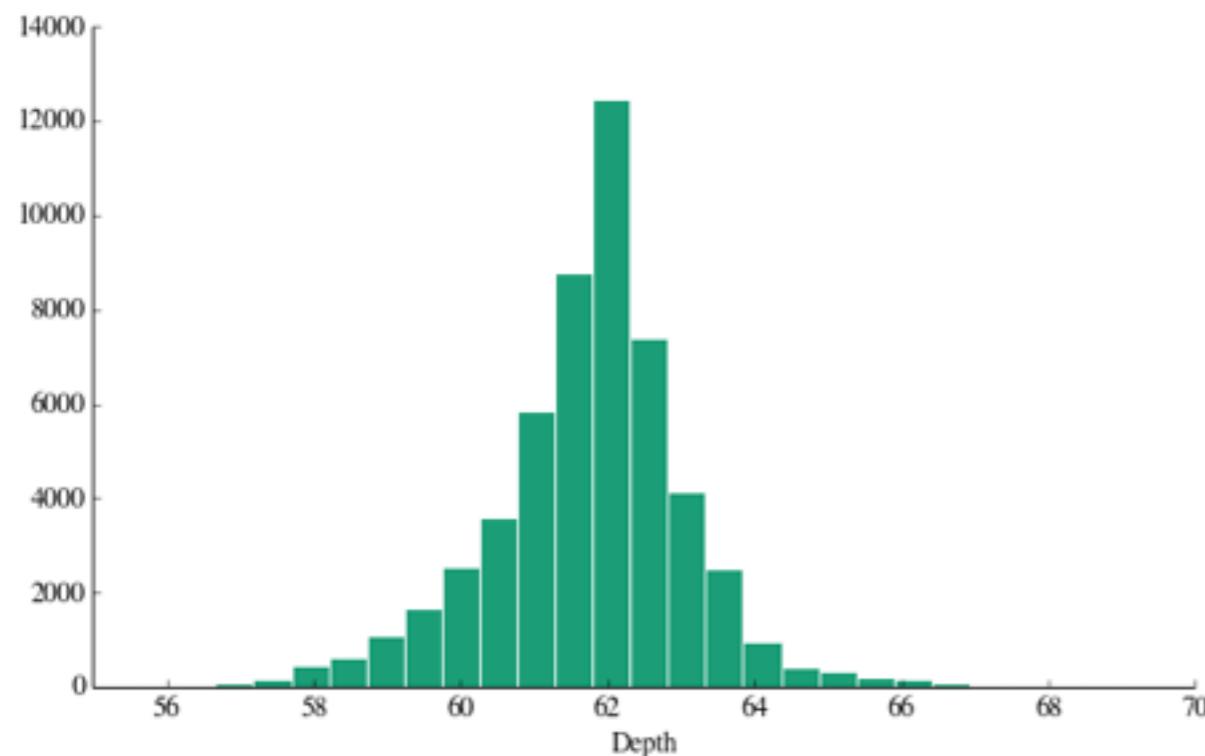


# Distributions

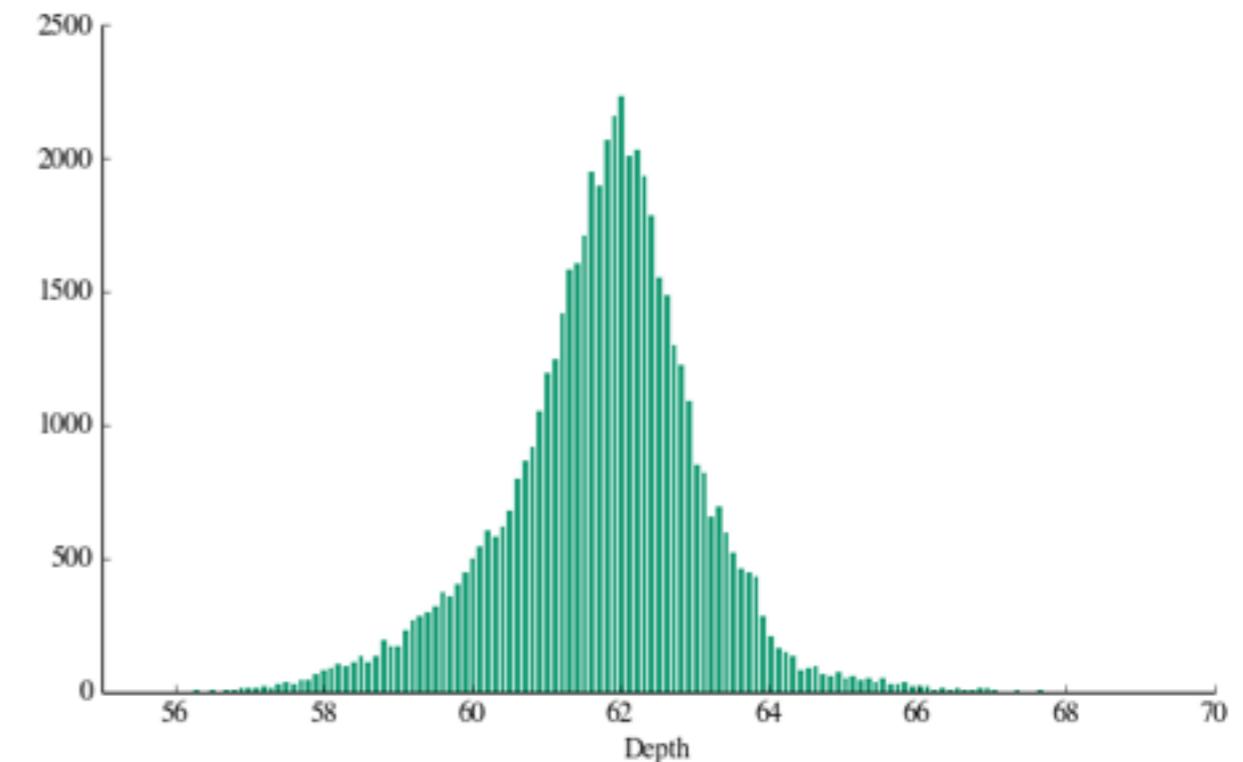
# Histogram



# Histogram

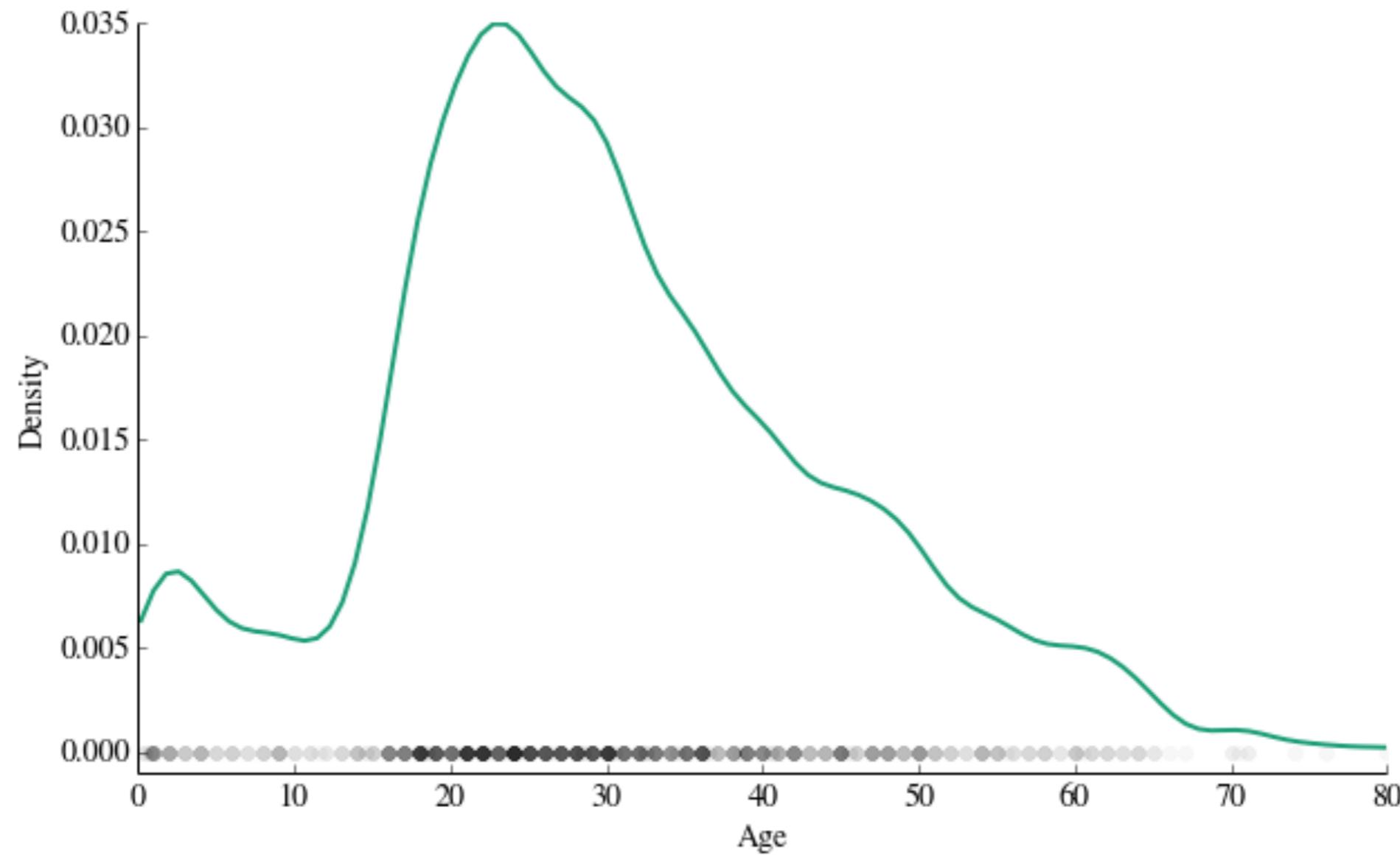


40 bins

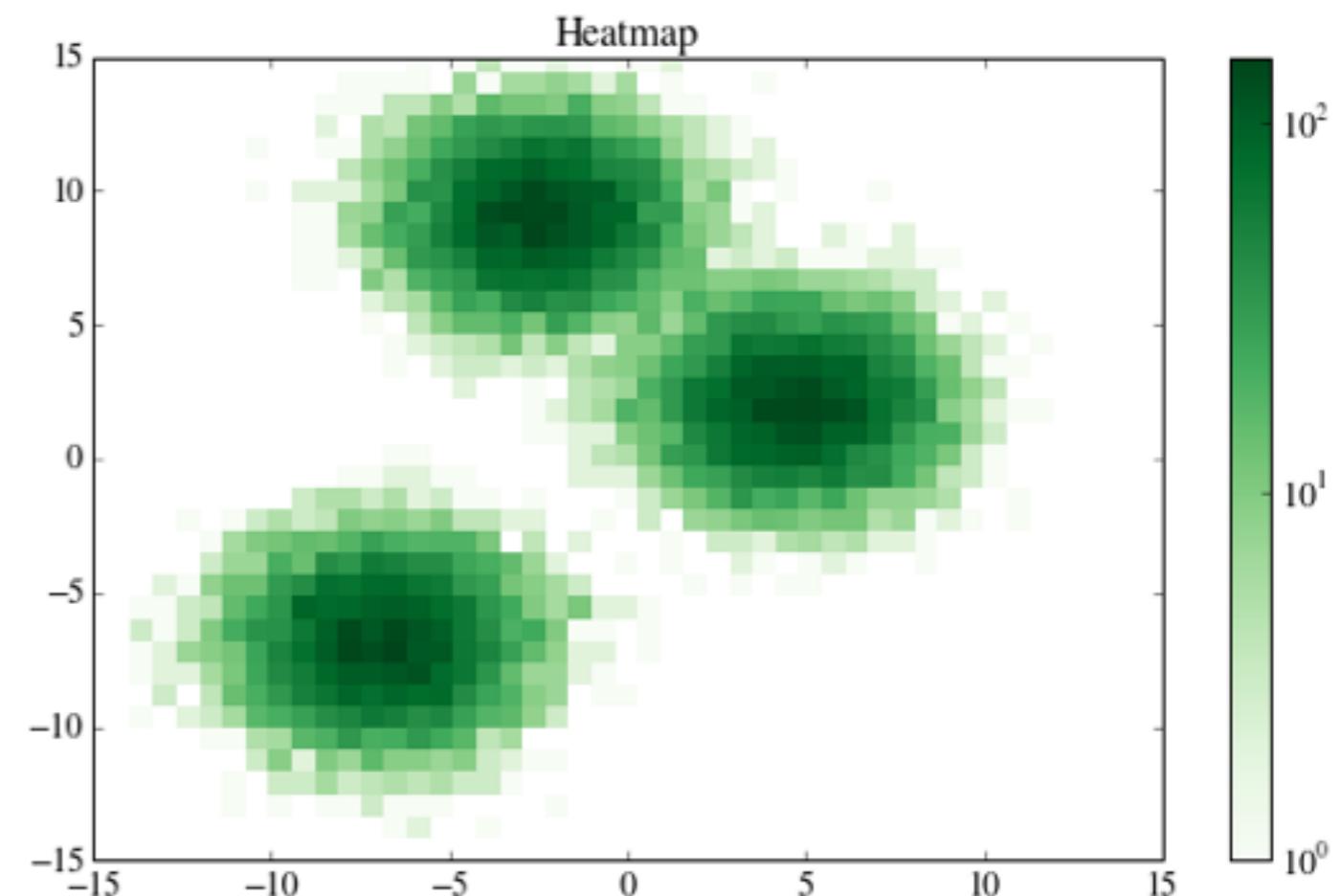
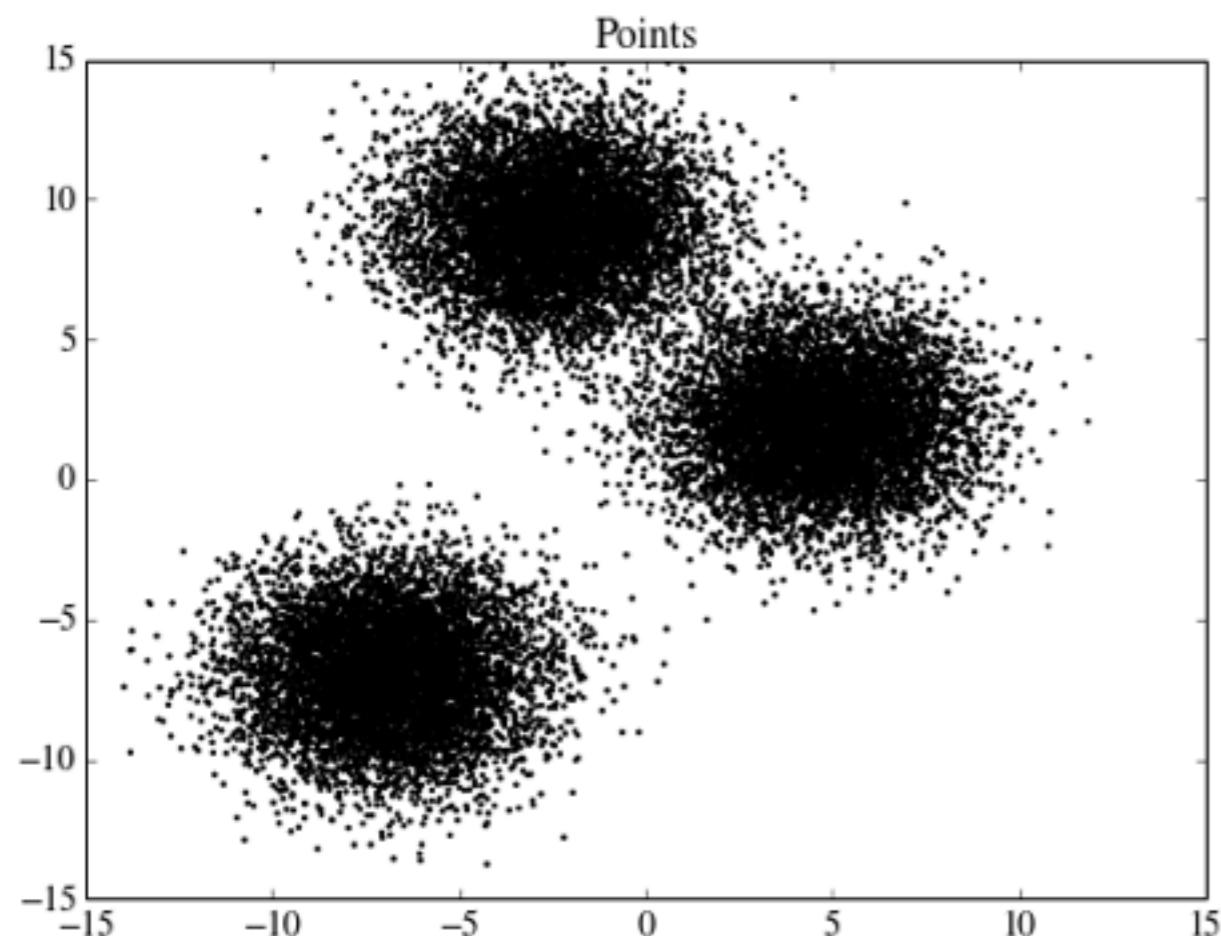


200 bins

# Density Plots

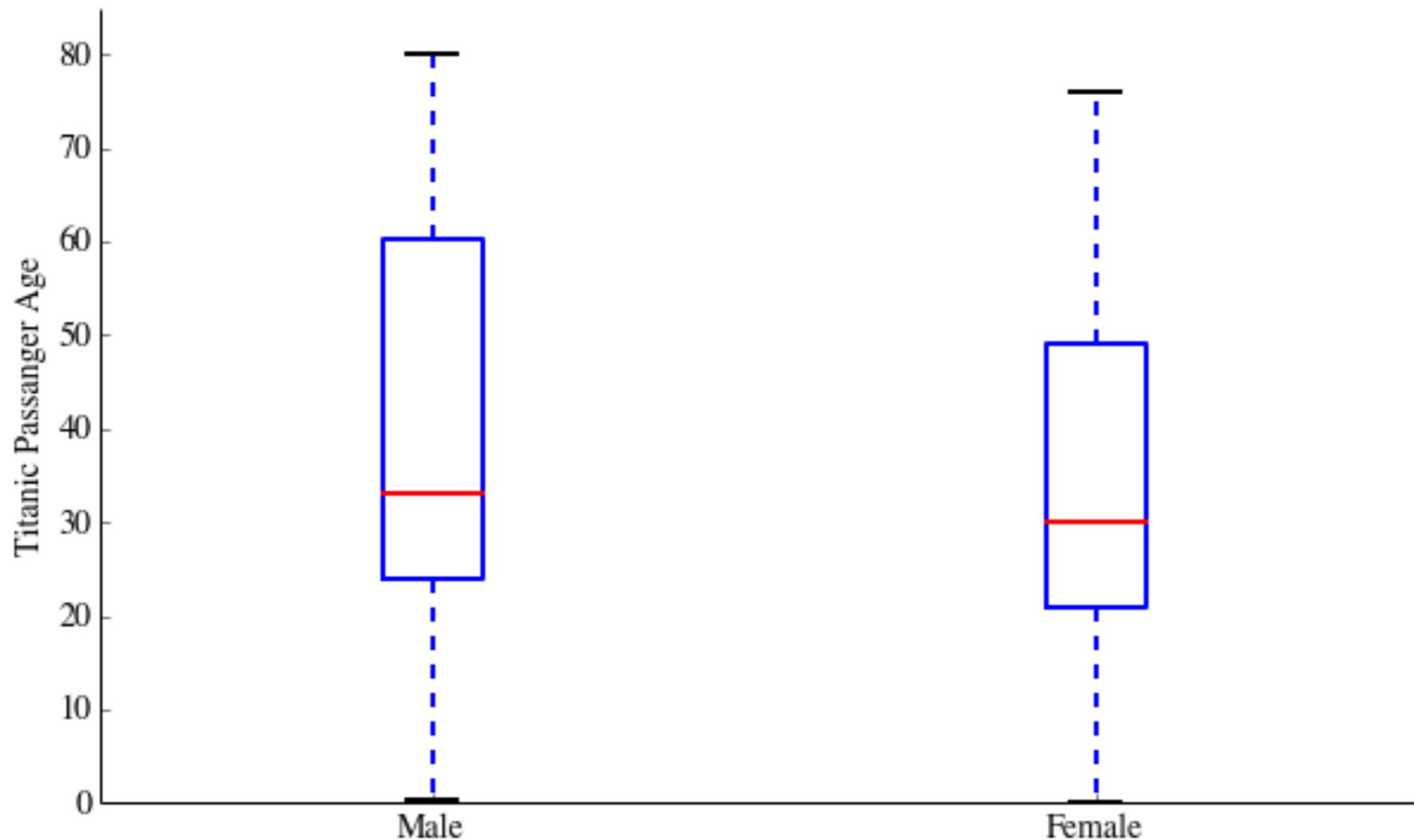


# Heat Maps

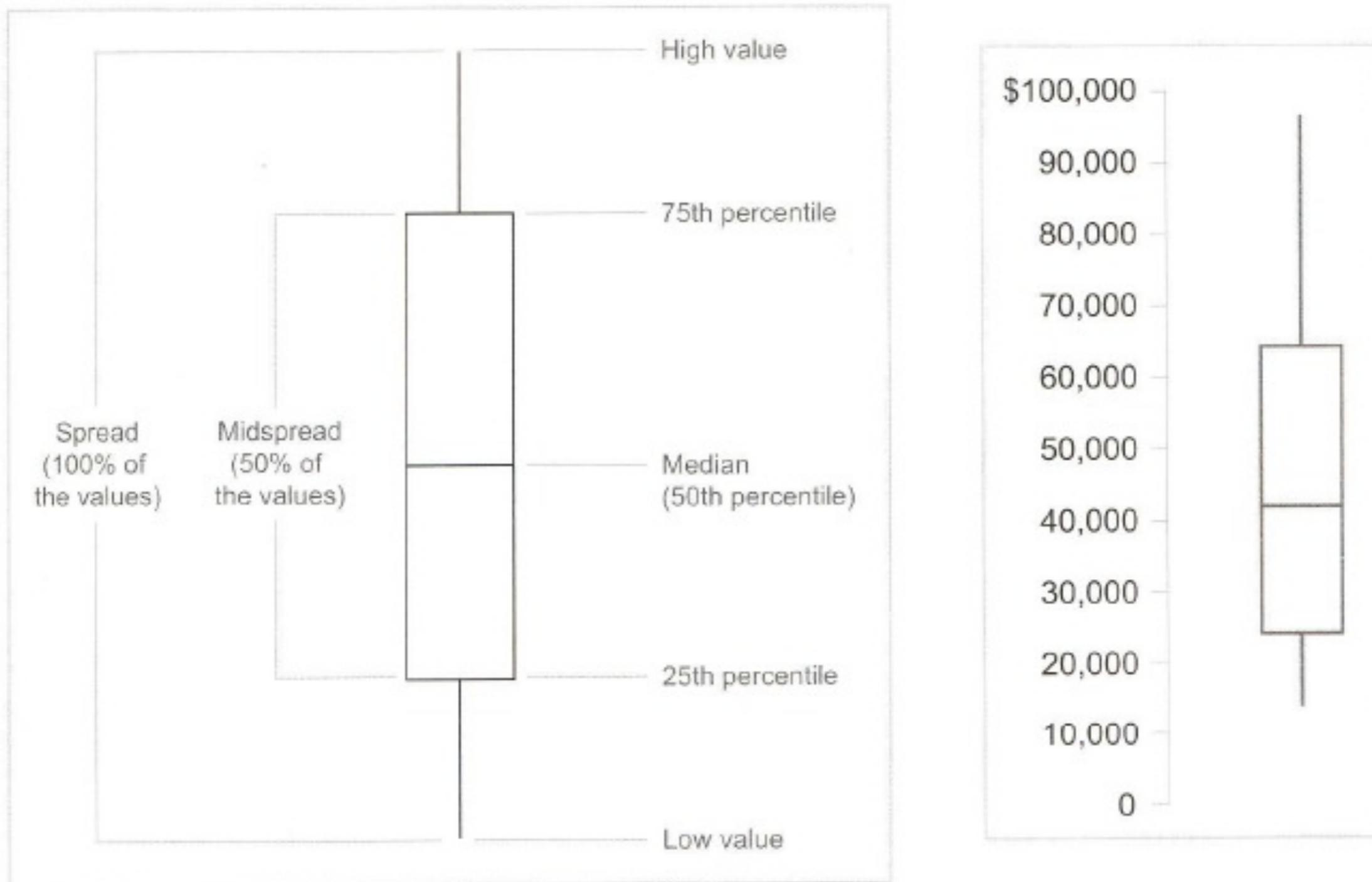


2D Density Plots

# Box & Whisker Plots

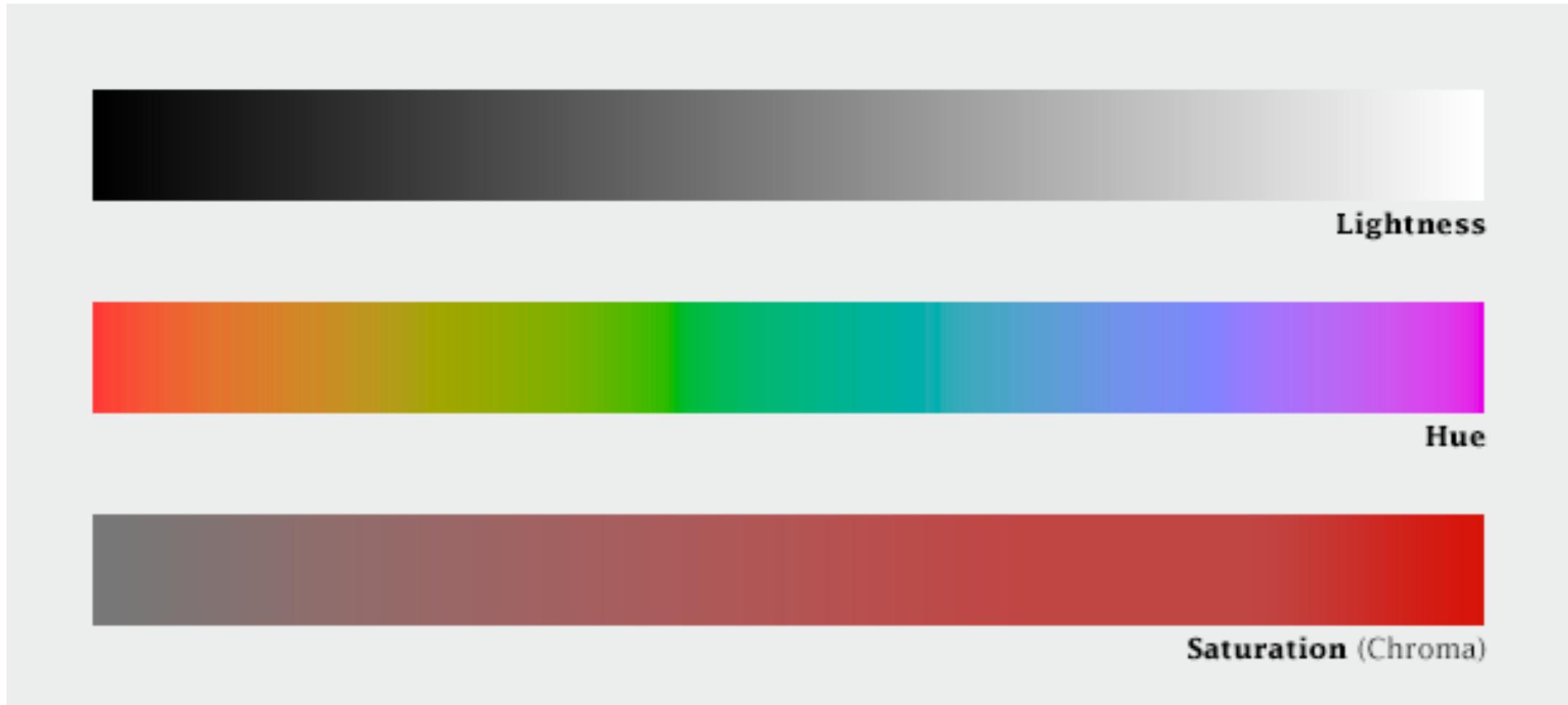


# Box & Whisker Plots

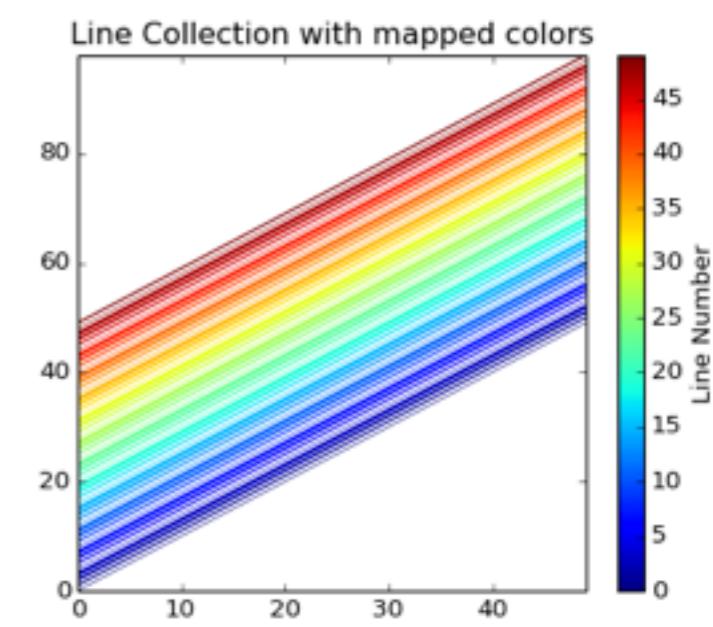
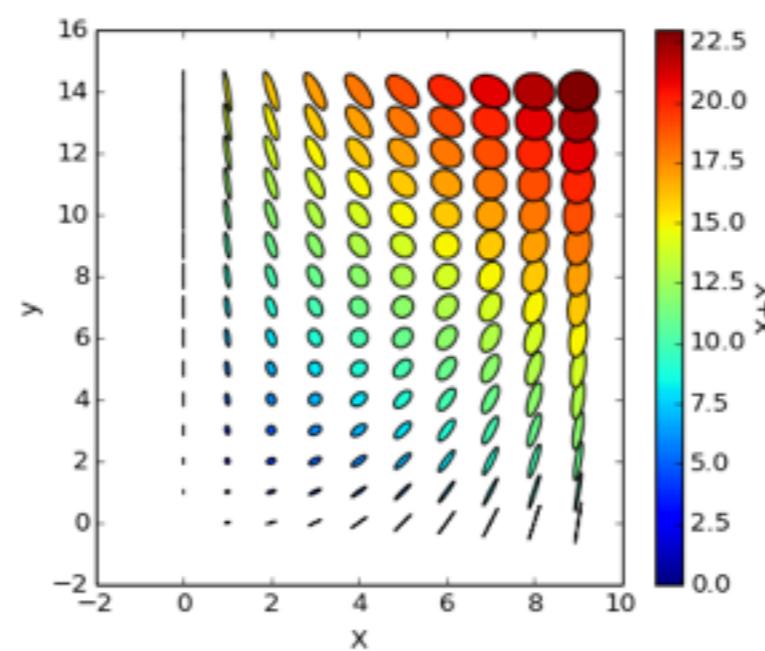
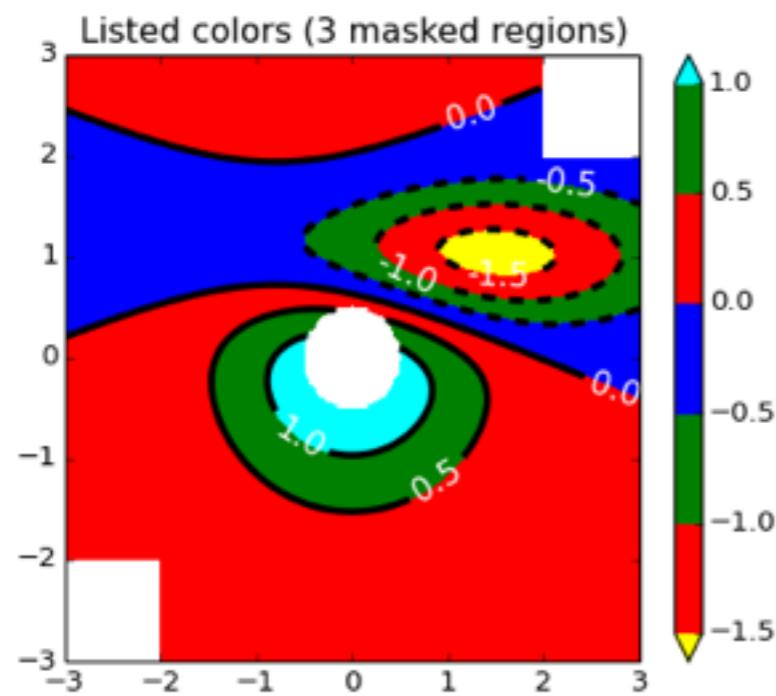
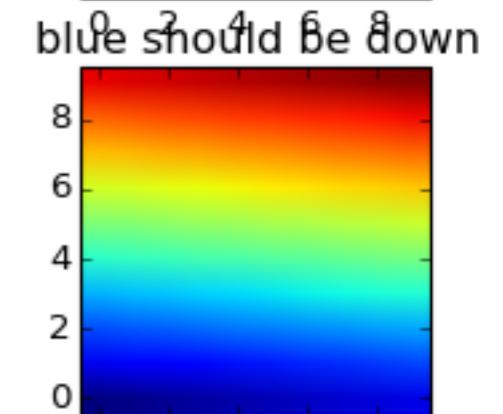
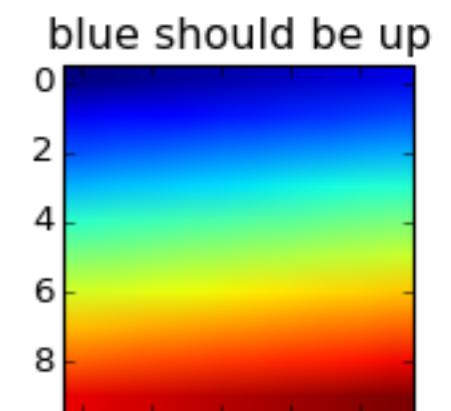
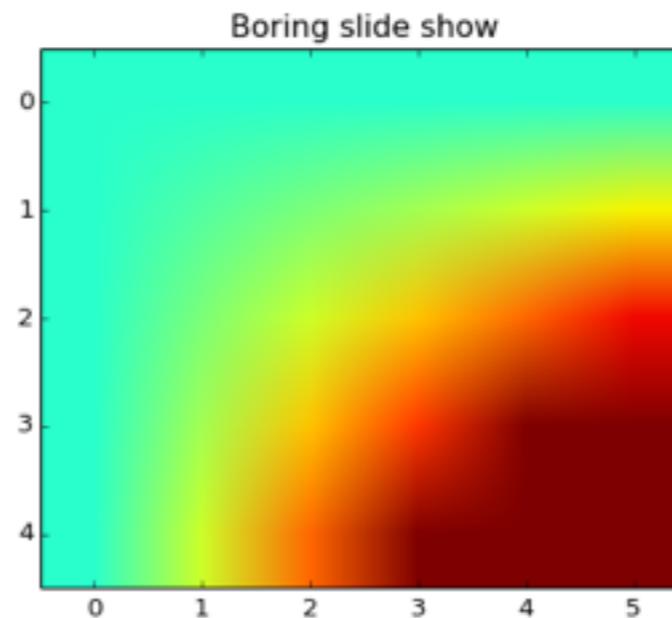
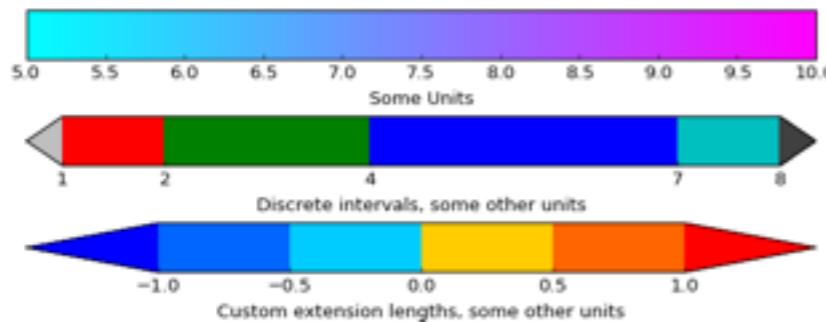


# Color

# Color Space

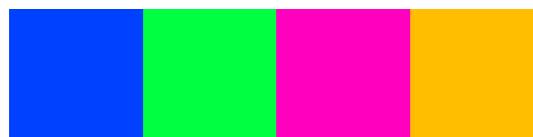


# Rainbow Colors

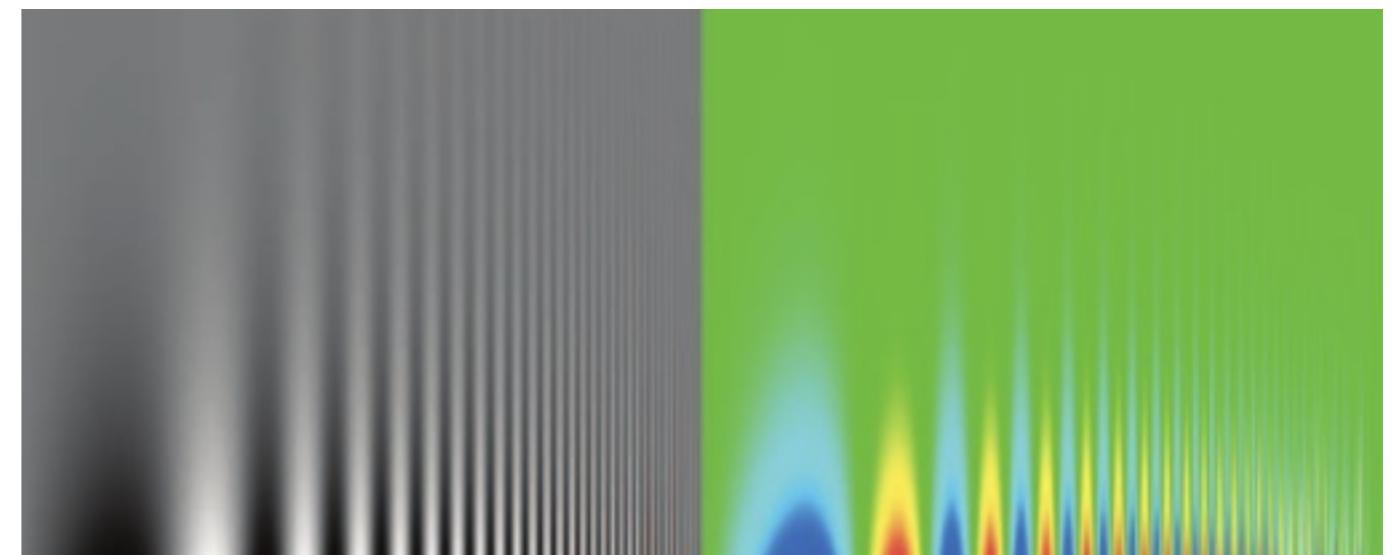


# Rainbow Colormap

hard to order



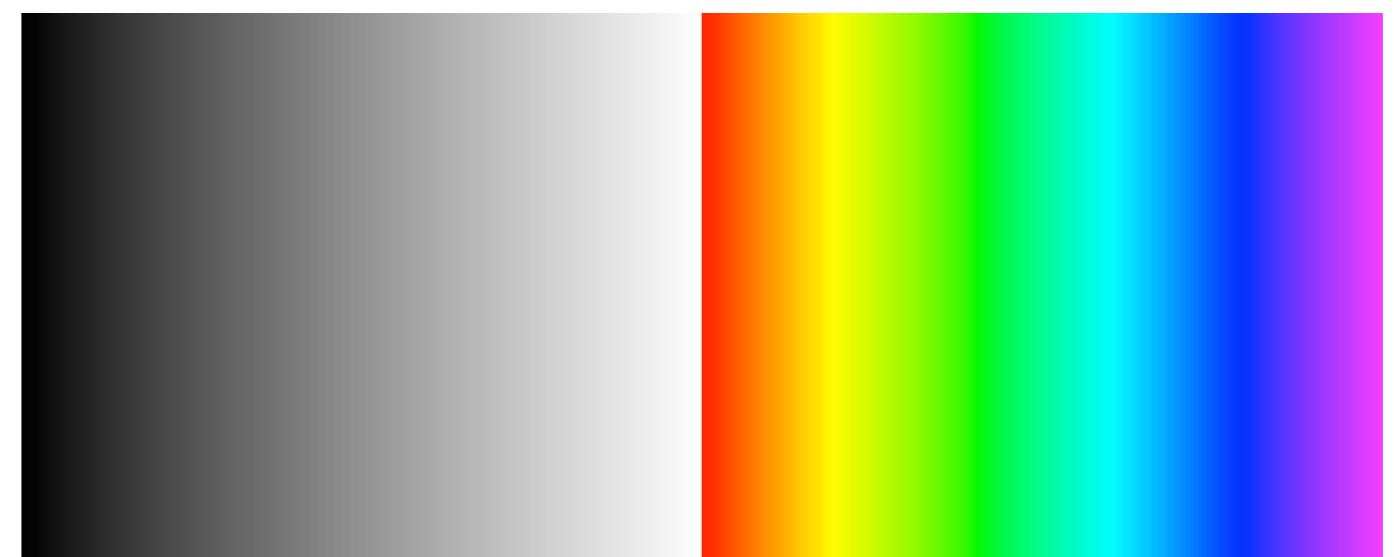
lower resolution



easy to order

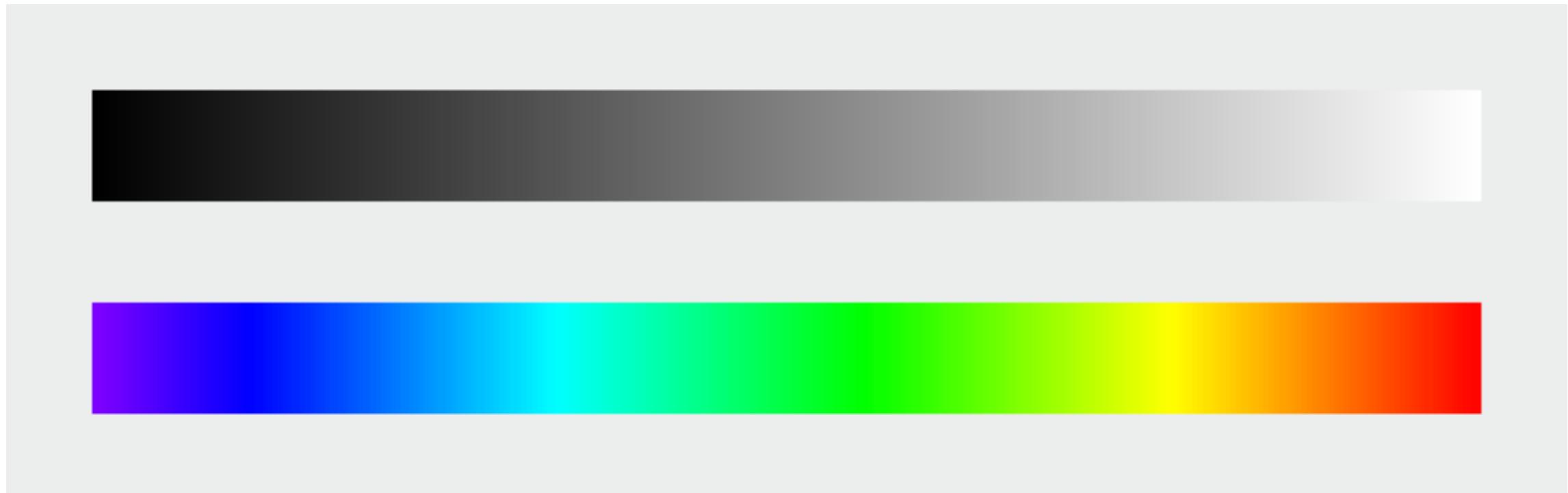


creates artifacts

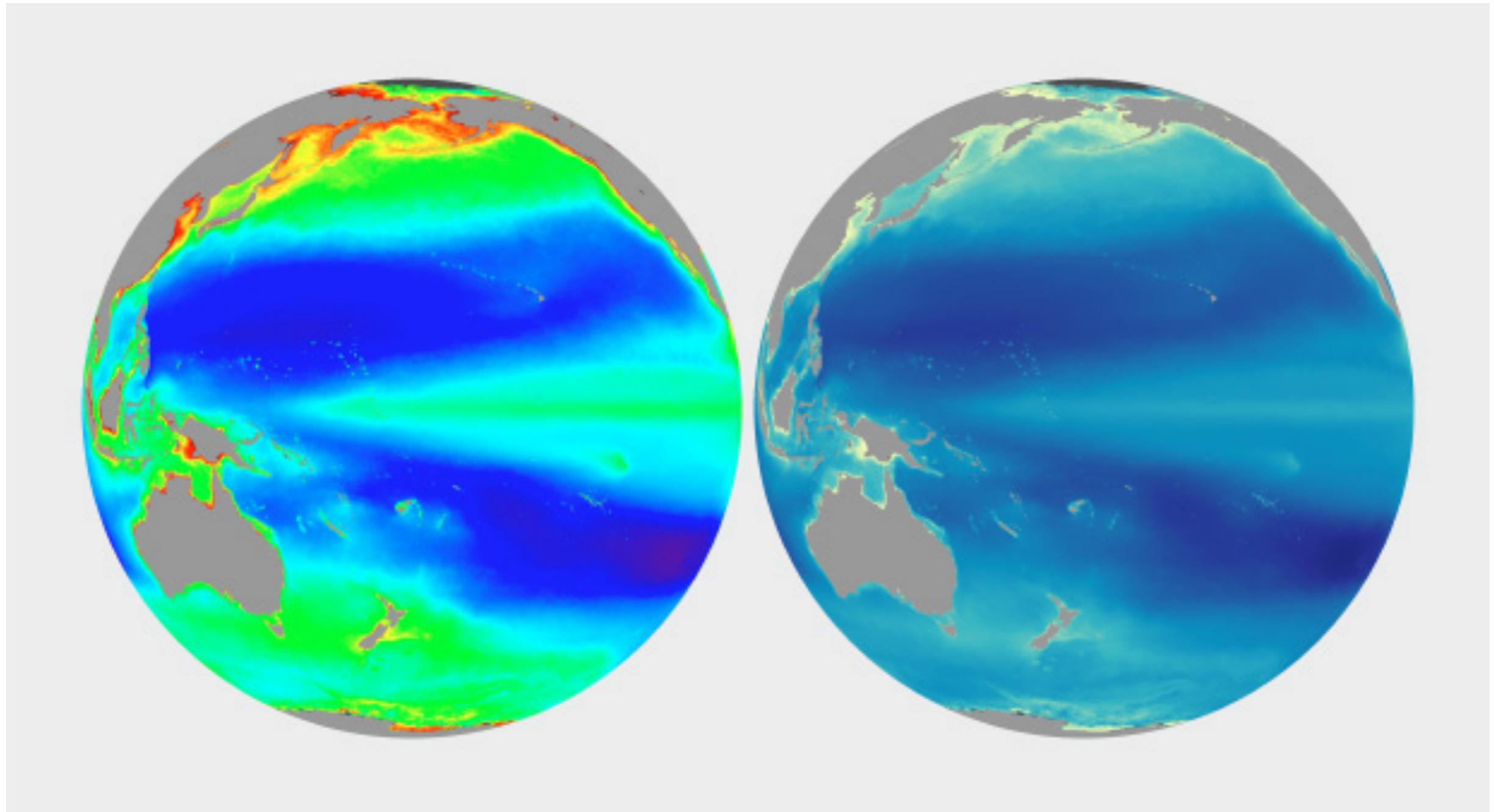


# Rainbow Colormap

Rainbow colormap is perceptually nonlinear

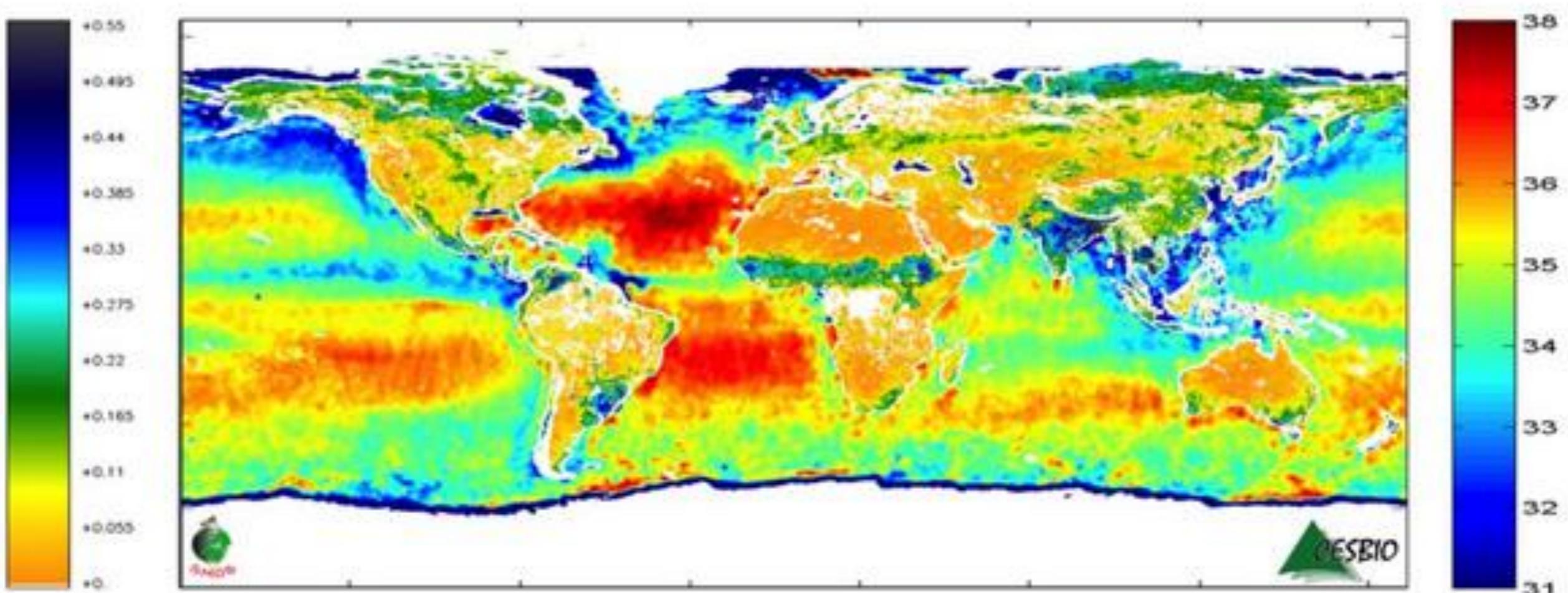


# Rainbow Colormap



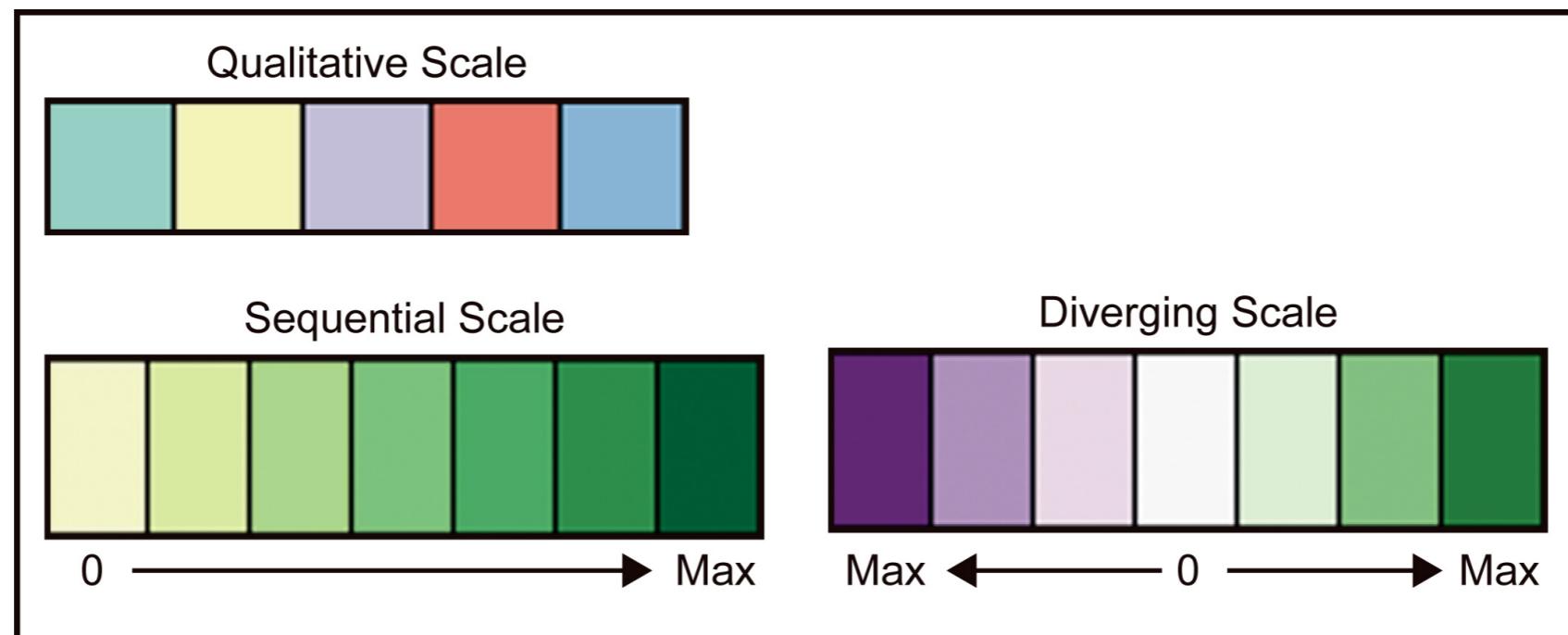
R. Simmon

# Rainbow Colormap



# Brewer Scales

Nominal



Ordinal

number of data classes on your map

3 | [learn more >](#)

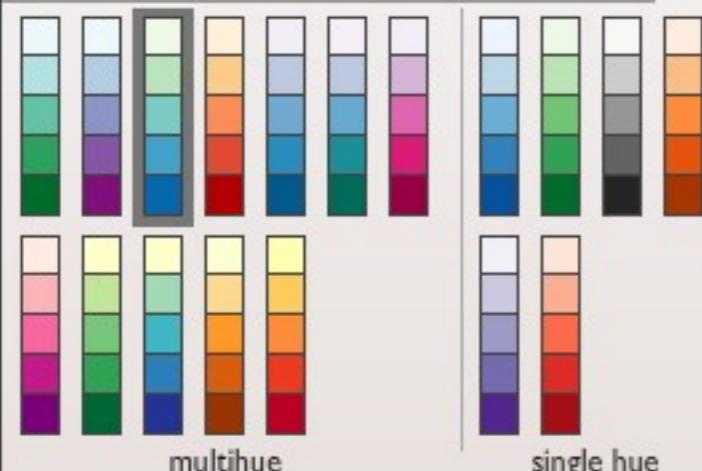
[how to use](#) | [updates](#) | [credits](#)

**COLORBREWER 2.0**  
color advice for cartography

the nature of your data

sequential | [learn more >](#)

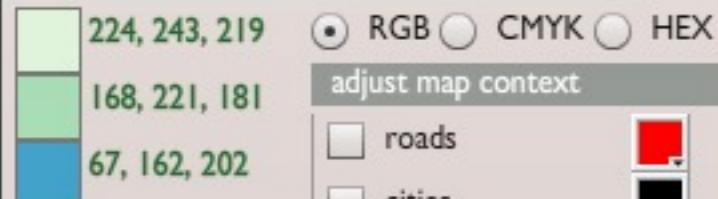
pick a color scheme: GnBu



(optional) only show schemes that are:

- colorblind safe  print friendly  
 photocopy-able [learn more >](#)

pick a color system



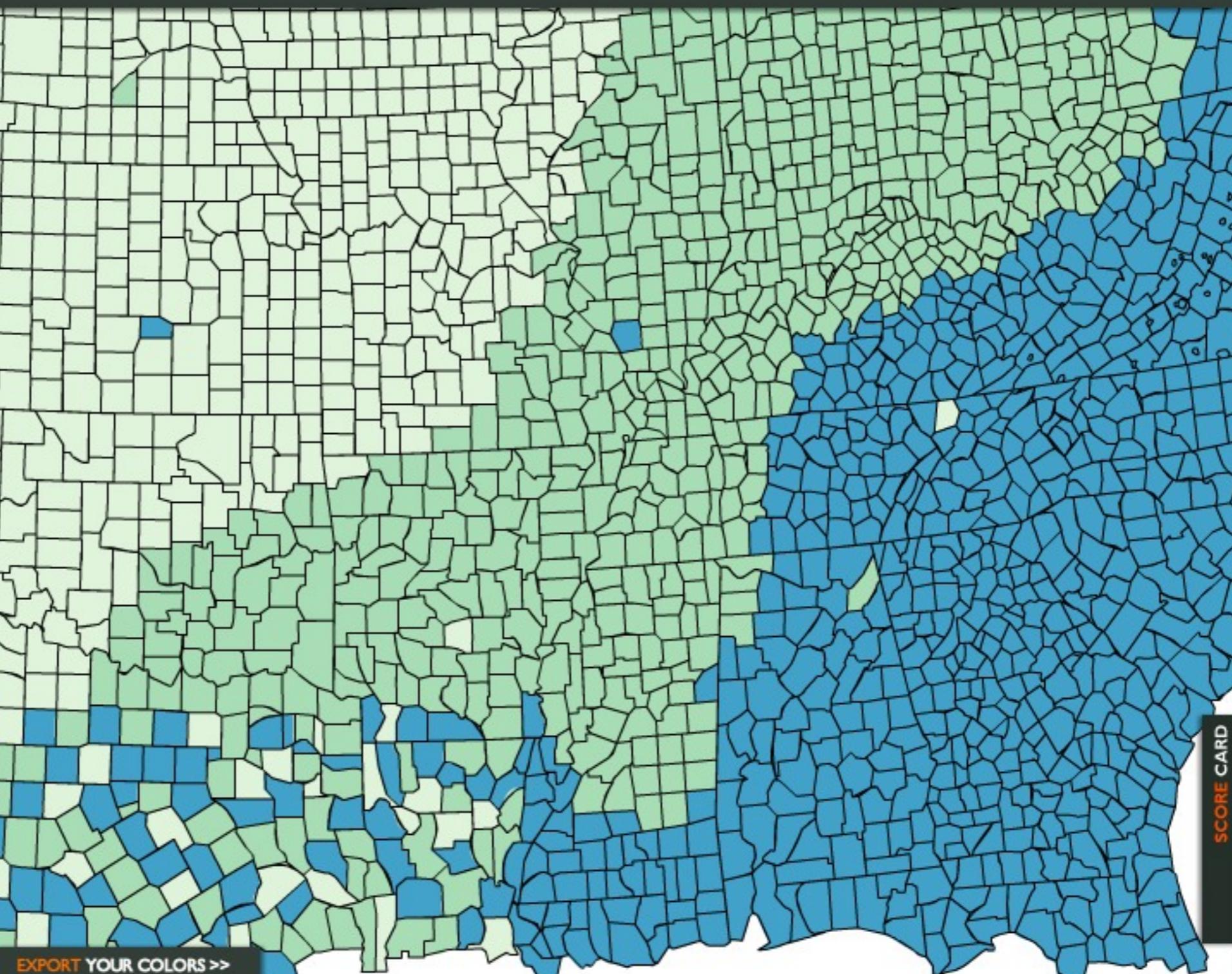
[adjust map context](#)

- roads   
 cities   
 borders

[select a background](#)

- solid color   
 terrain

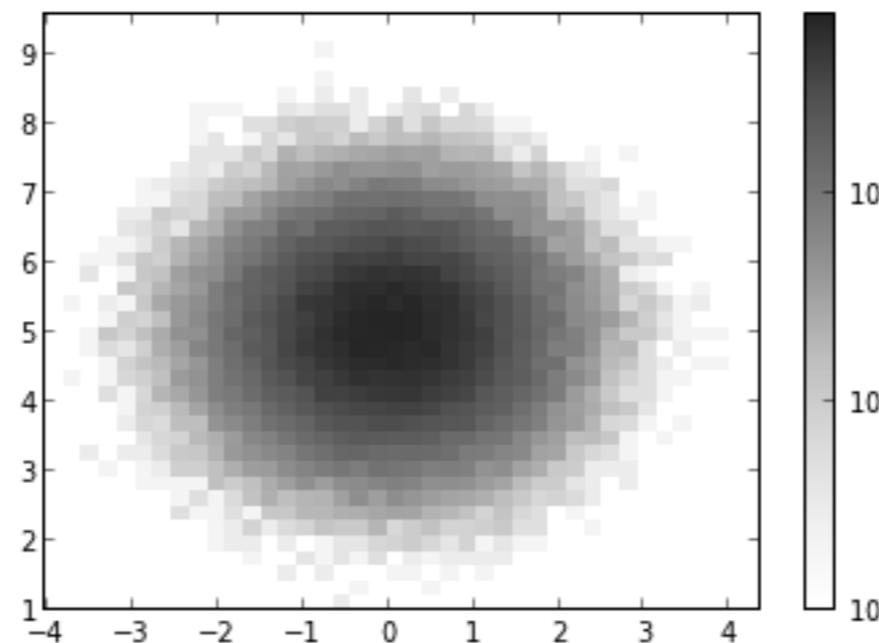
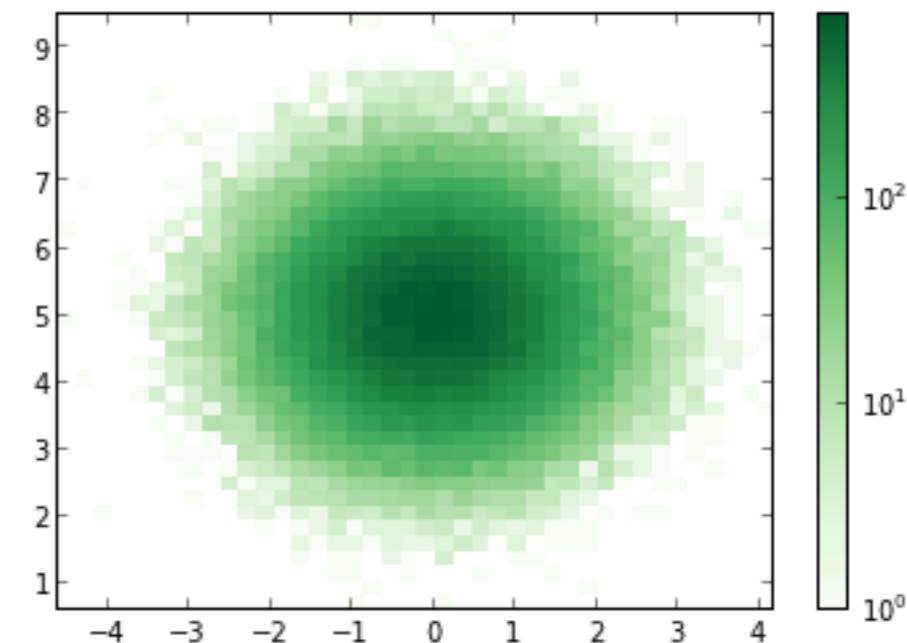
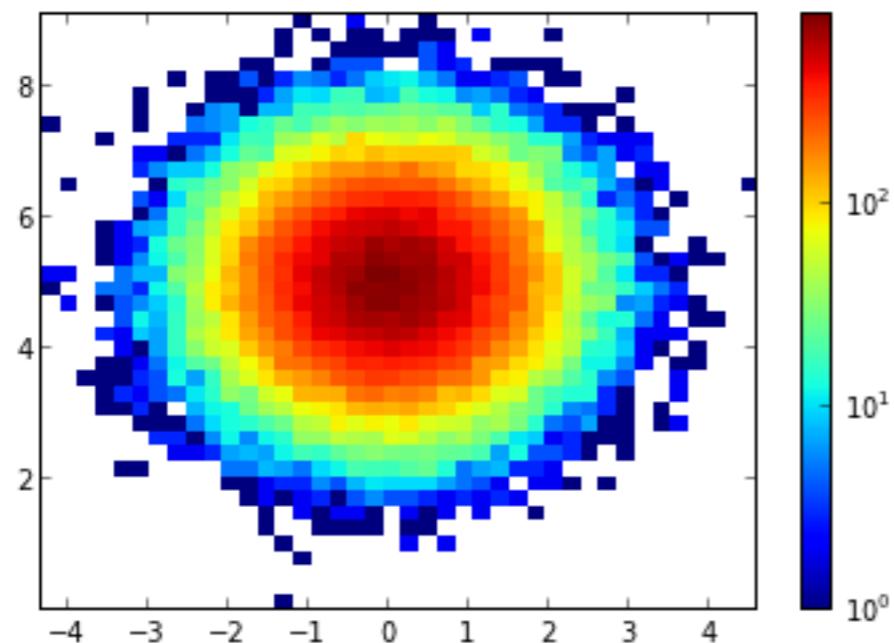
[color transparency](#)



[EXPORT YOUR COLORS >>](#)

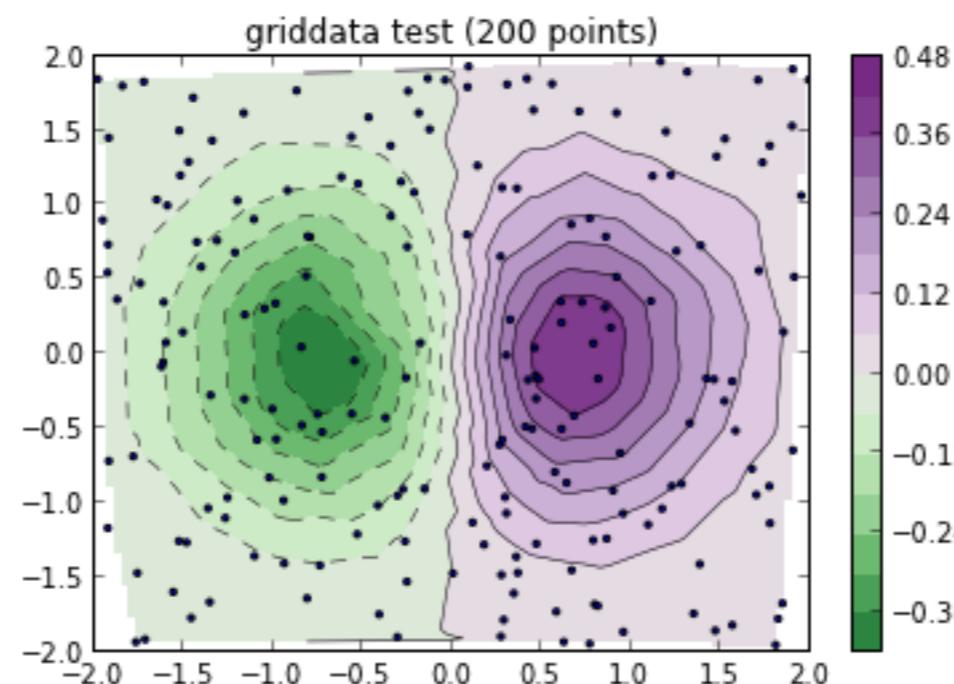
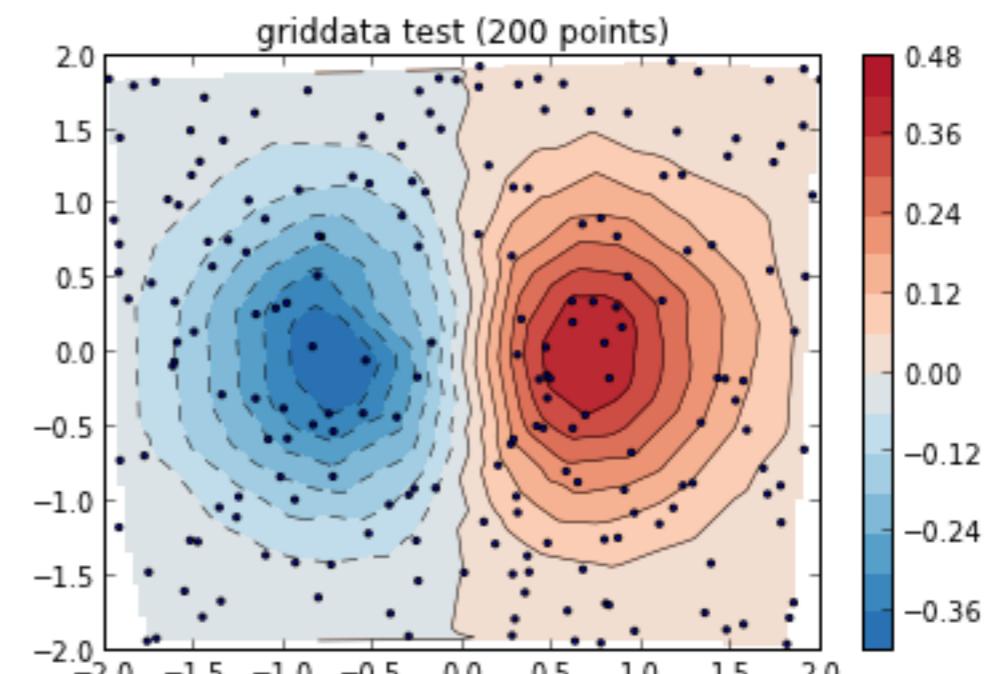
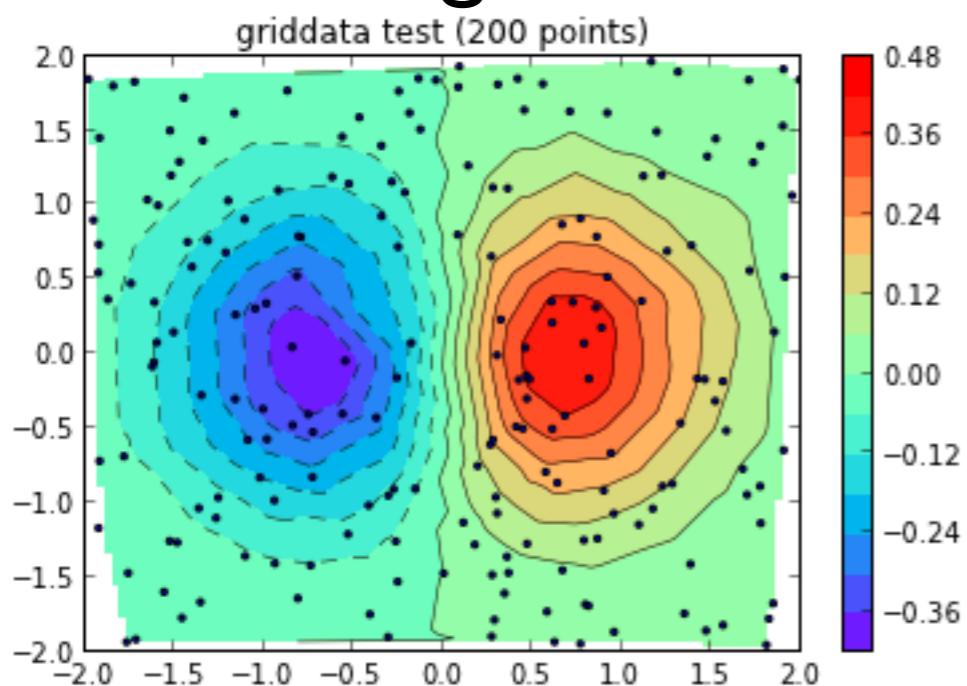
# Sequential Brewer Scales

No!



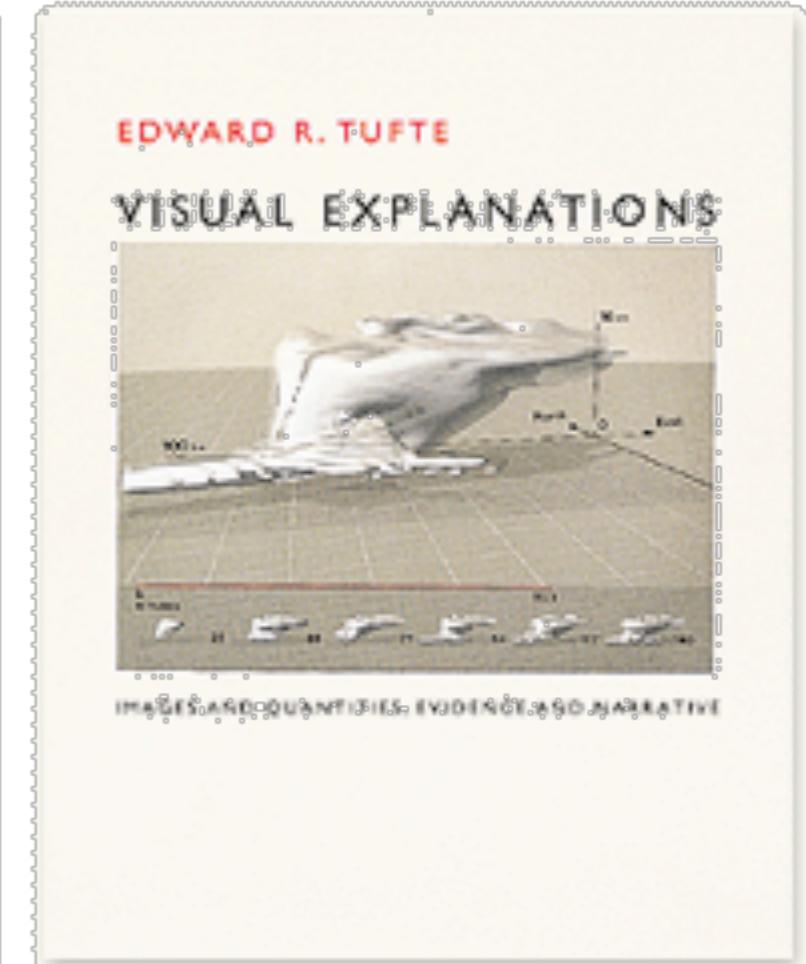
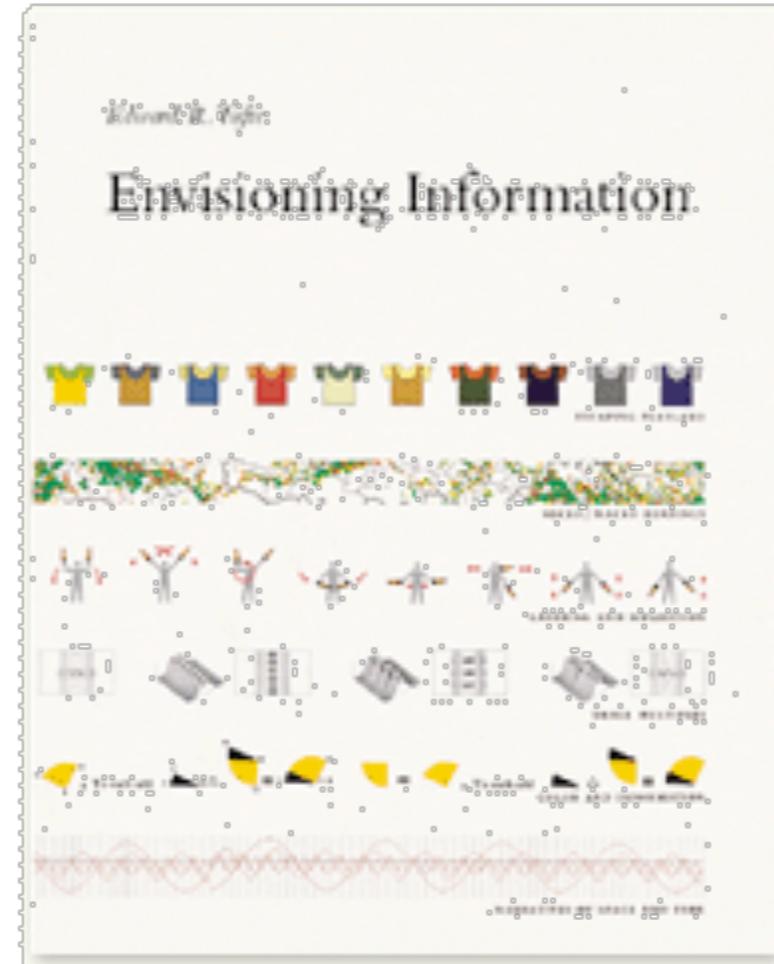
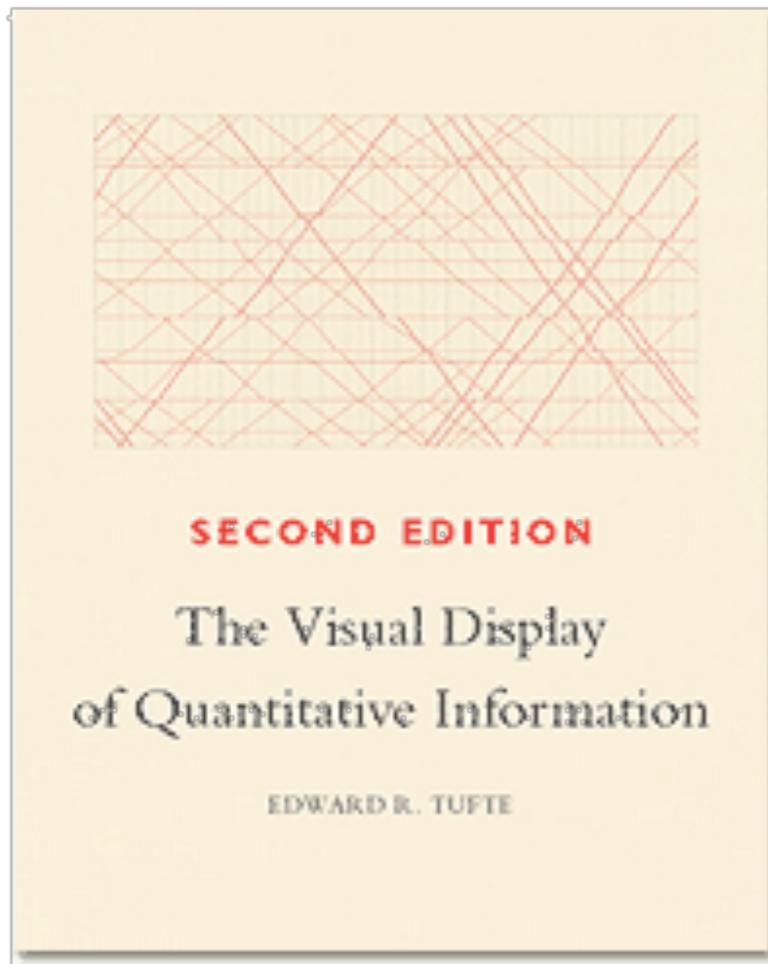
# Divergent Brewer Scales

Not great



# Further Reading

# Edward Tufte



# Stephen Few

