
GTK CYBER
GET ON THE FAST TRACK

Module 4: Data Visualization

Our Lawyers Make Us Say This

Exploring Data Through Visualization

Exploratory Visualizations

Explanatory Visualizations

Why Visualize Data?

Visualizing data can inspire you to ask new and more refined questions of your data and ultimately lead to better analysis.

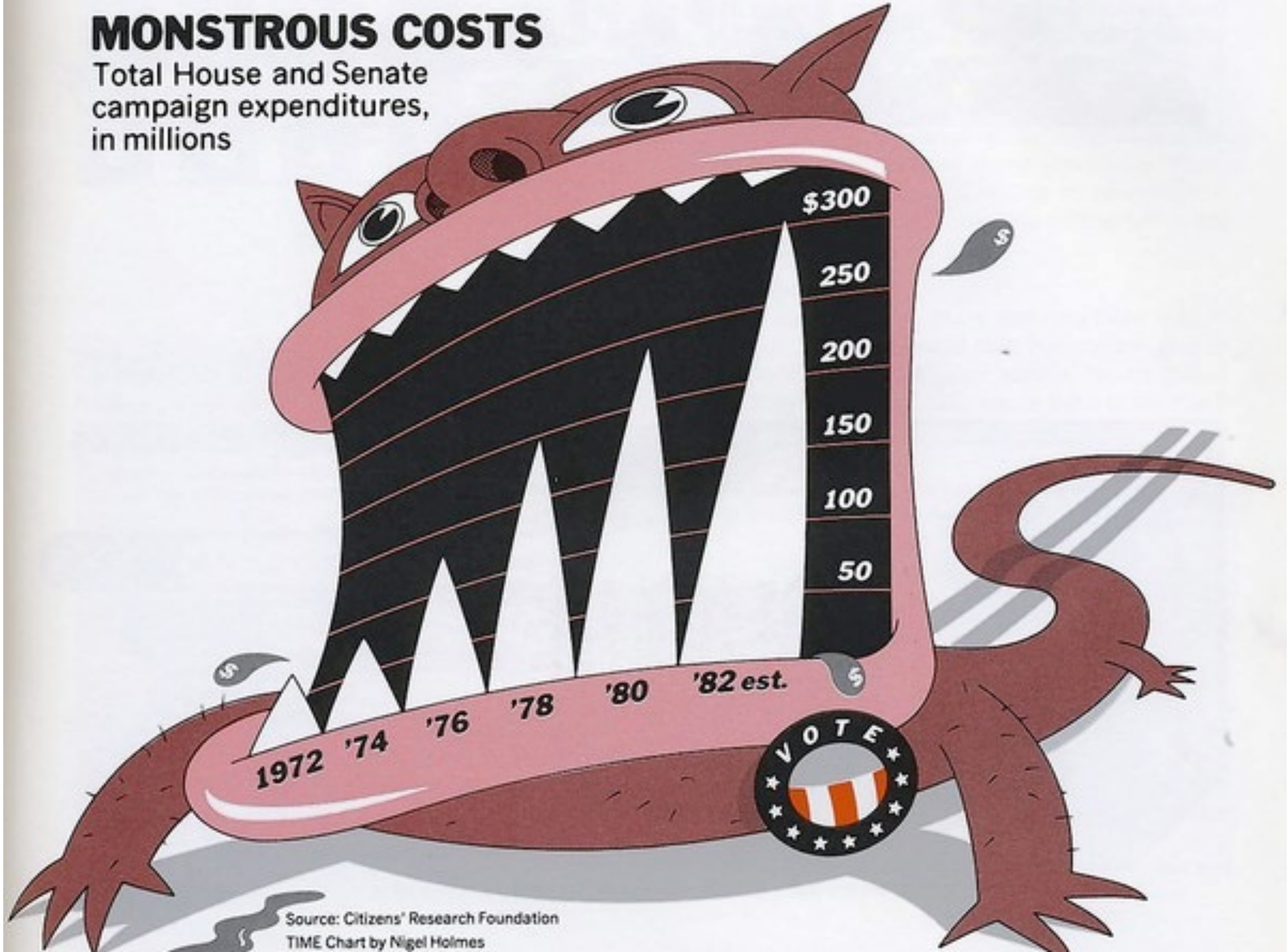
"The greatest value of a picture is when it forces us to notice what we never expected to see."

–John Tukey, 1977

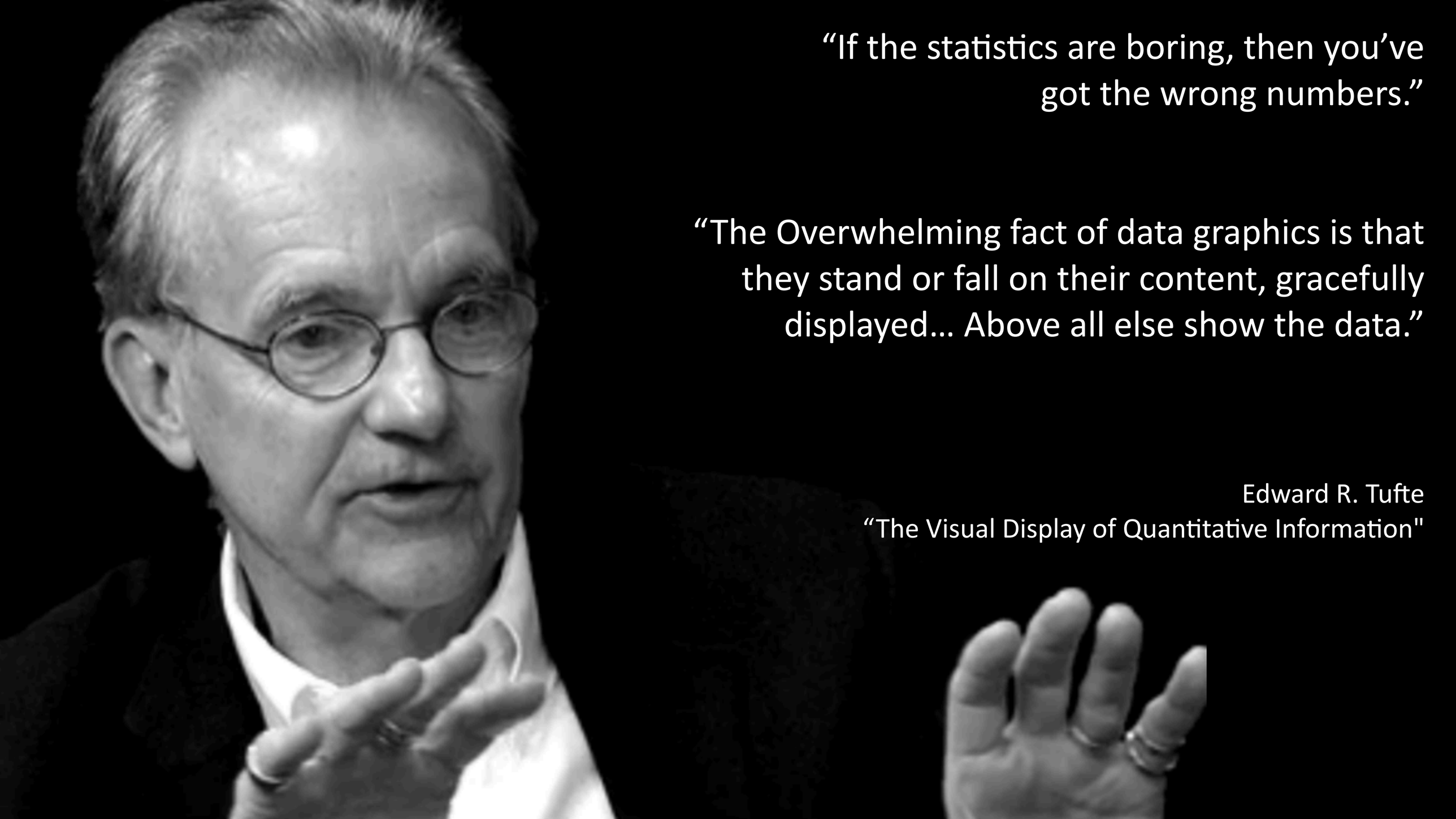
The power of visualization comes from **illustrating relationships**, contrasts and comparisons between many different dimensions of data.

MONSTROUS COSTS

Total House and Senate campaign expenditures, in millions



Source: Citizens' Research Foundation
TIME Chart by Nigel Holmes



“If the statistics are boring, then you’ve got the wrong numbers.”

“The Overwhelming fact of data graphics is that they stand or fall on their content, gracefully displayed… Above all else show the data.”

Edward R. Tufte
“The Visual Display of Quantitative Information"

Remove

to improve

(the **data-ink** ratio)

Show Comparisons, Contrasts and Differences

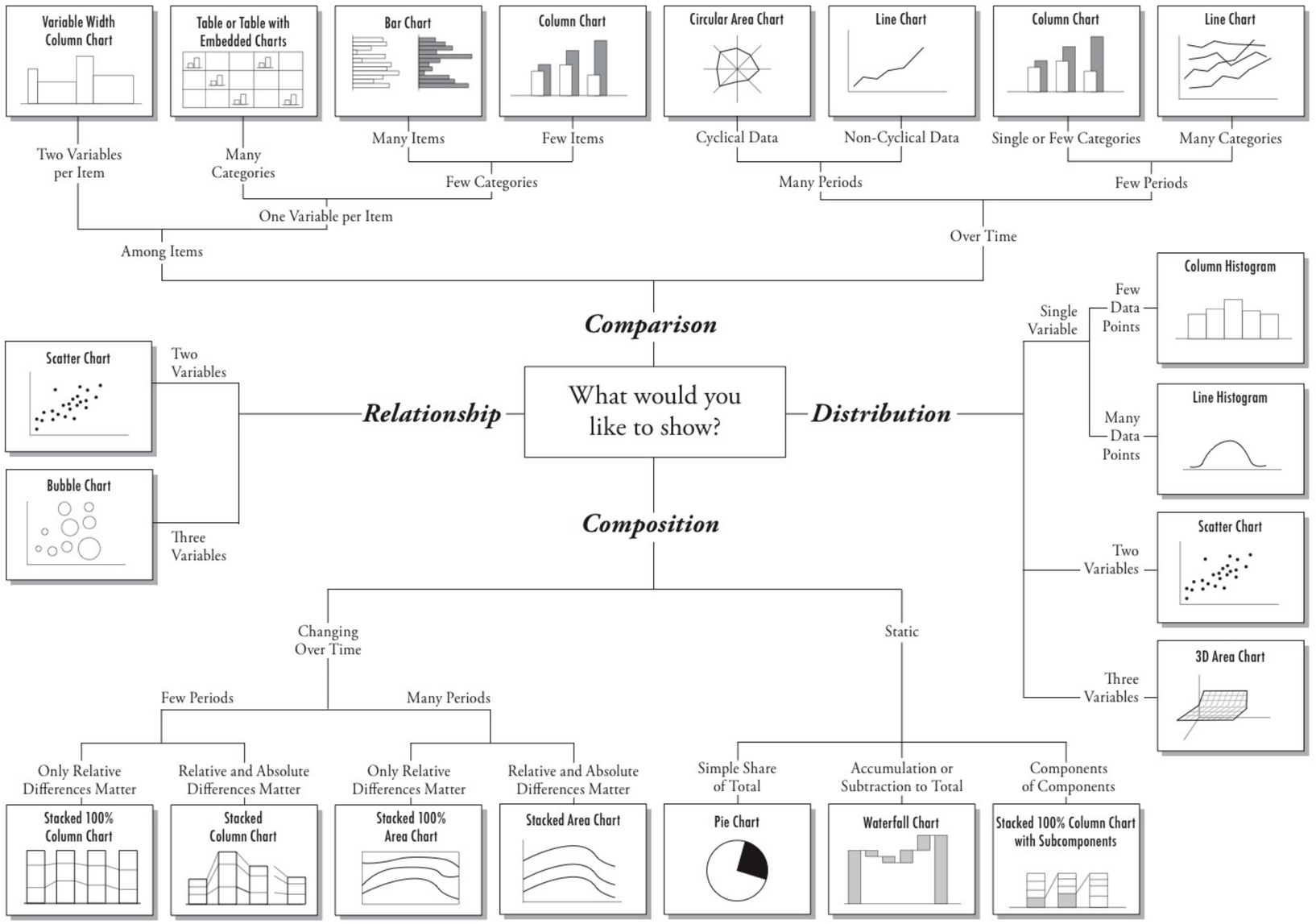
Show Multivariate Data

Integrate words, numbers, images and diagrams

Document your Evidence

Ultimately, the quality, relevance and integrity of the content is most important.

Chart Suggestions—A Thought-Starter



Visualization Goals

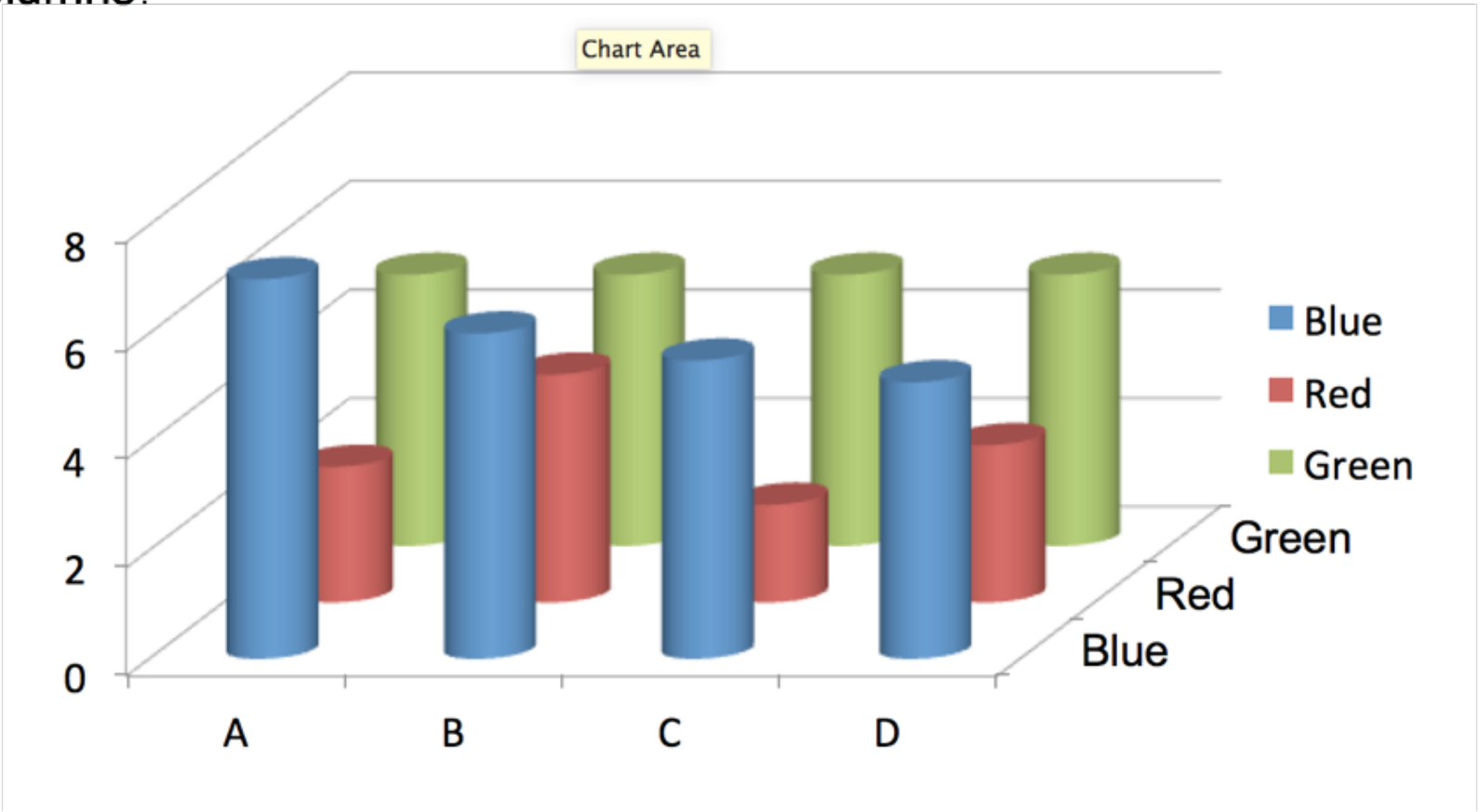
- Analyze
- Explore
- Assess
- Determine
- Decide
- Communicate
- Explain
- Present
- Prove
- Persuade

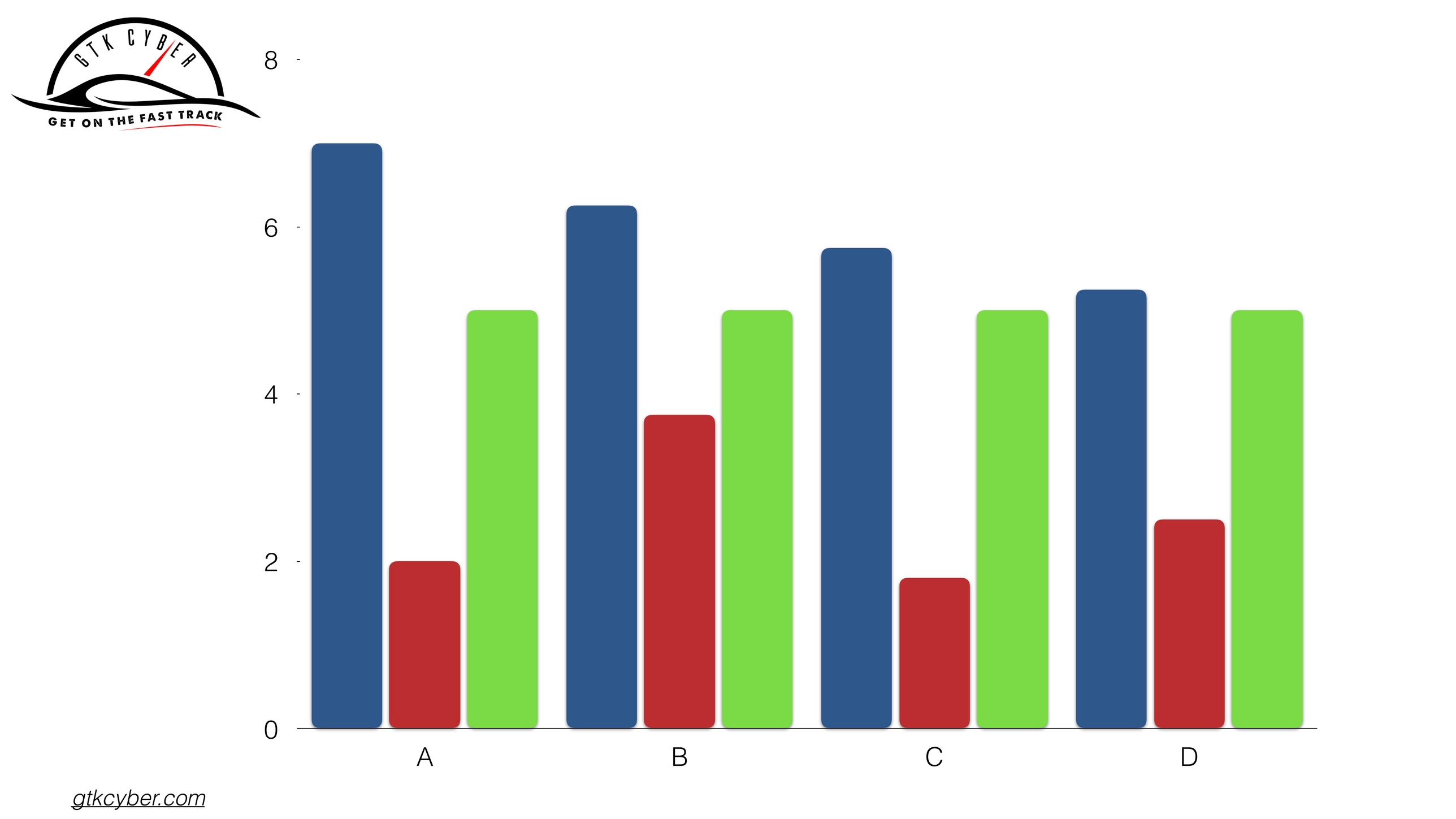
Elements of Good Visualizations

1. Graphical Integrity
2. Simple
3. Proper Display
4. Proper Color
5. Tells a story

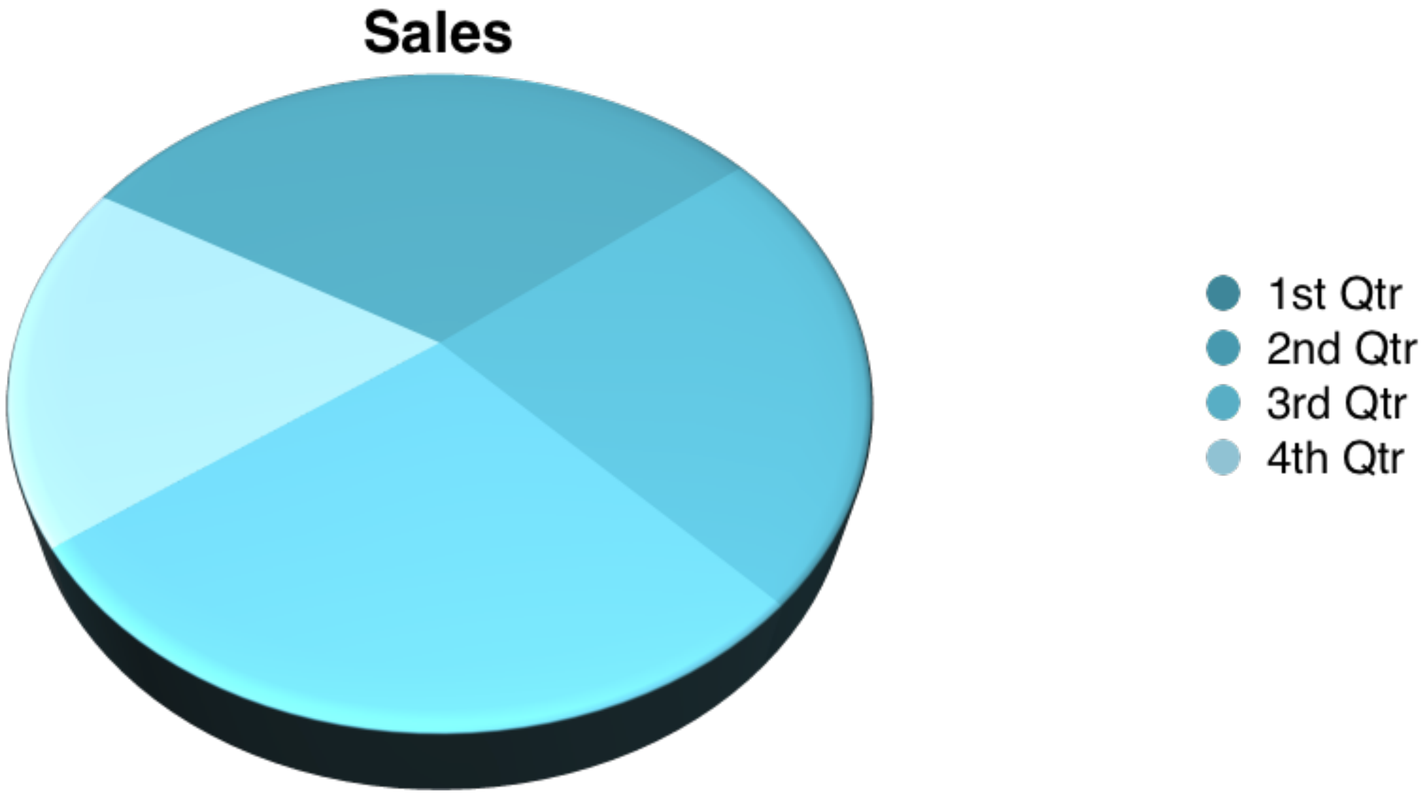
Proper Display

Questions: What is the height of the green columns? For which categories (A,B,C,D) are the blue columns taller than the green columns?

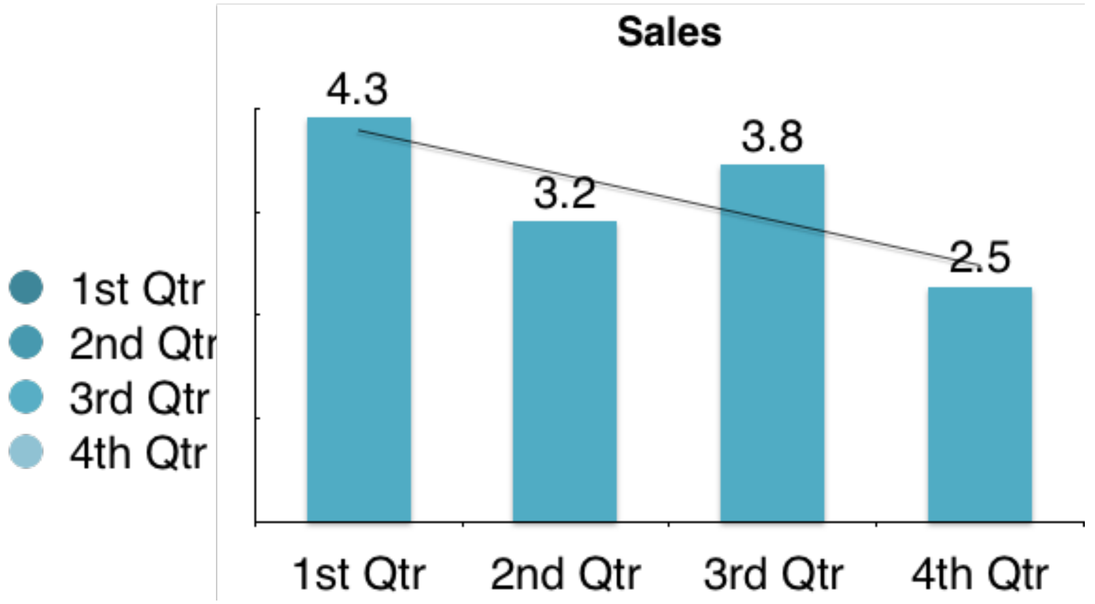



GTK CYBER
GET ON THE FAST TRACK
8
6
4
2
0
A
B
C
D

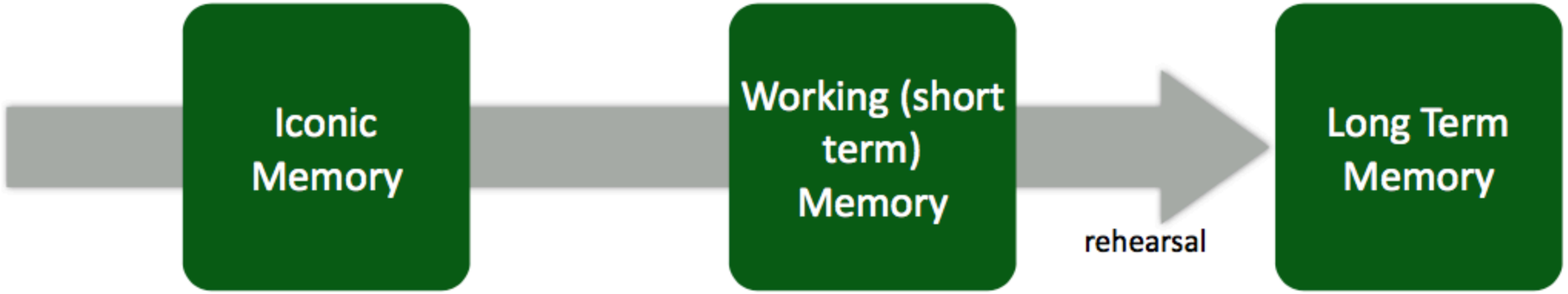
Proper Display



Proper Display



Visual Processing System



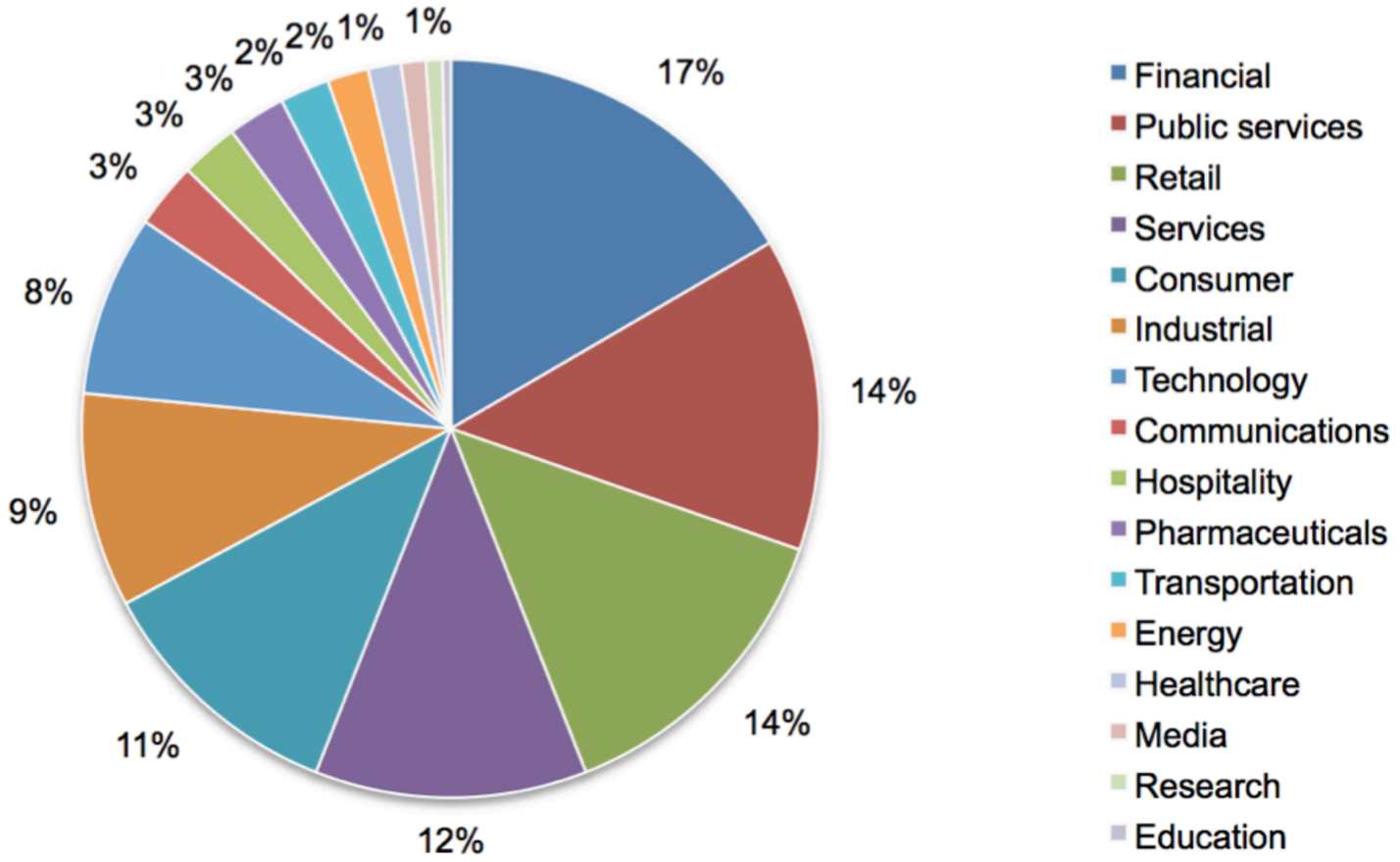
- < 1 second (or until replaced)
- Preattentive processing
- Saccadic movement
- ~20 objects

- Visual groupings
- Flexible, conscious
- 3-5 objects

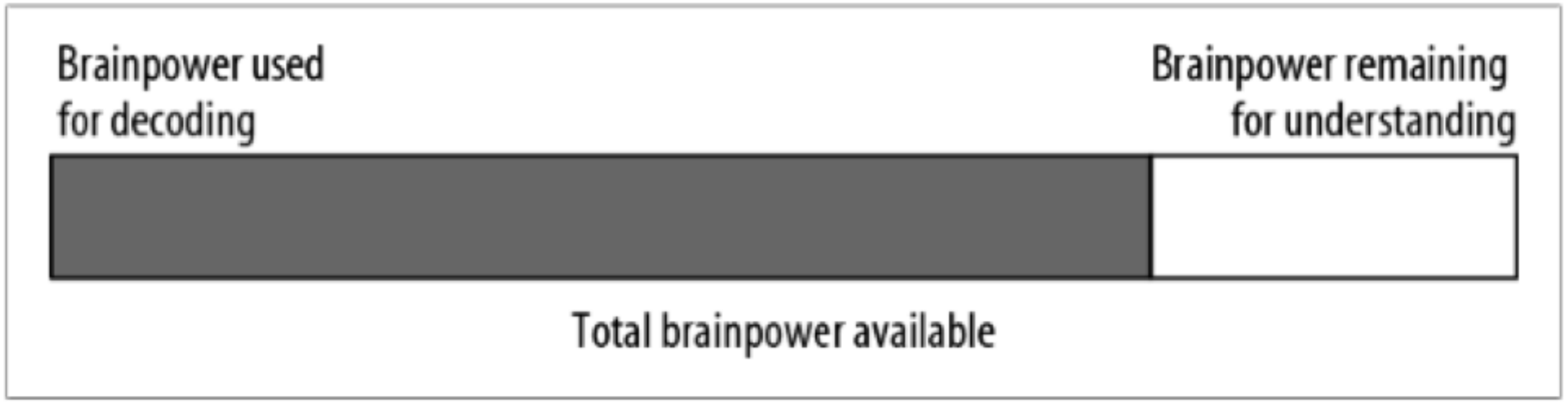
- Affects dataviz through expectations

Overworking Visual Memory

Figure 20. Distribution of the benchmark sample by industry segment
Consolidated (n = 277 organizations)



https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf


Brainpower used
for decoding
Brainpower remaining
for understanding
Total brainpower available

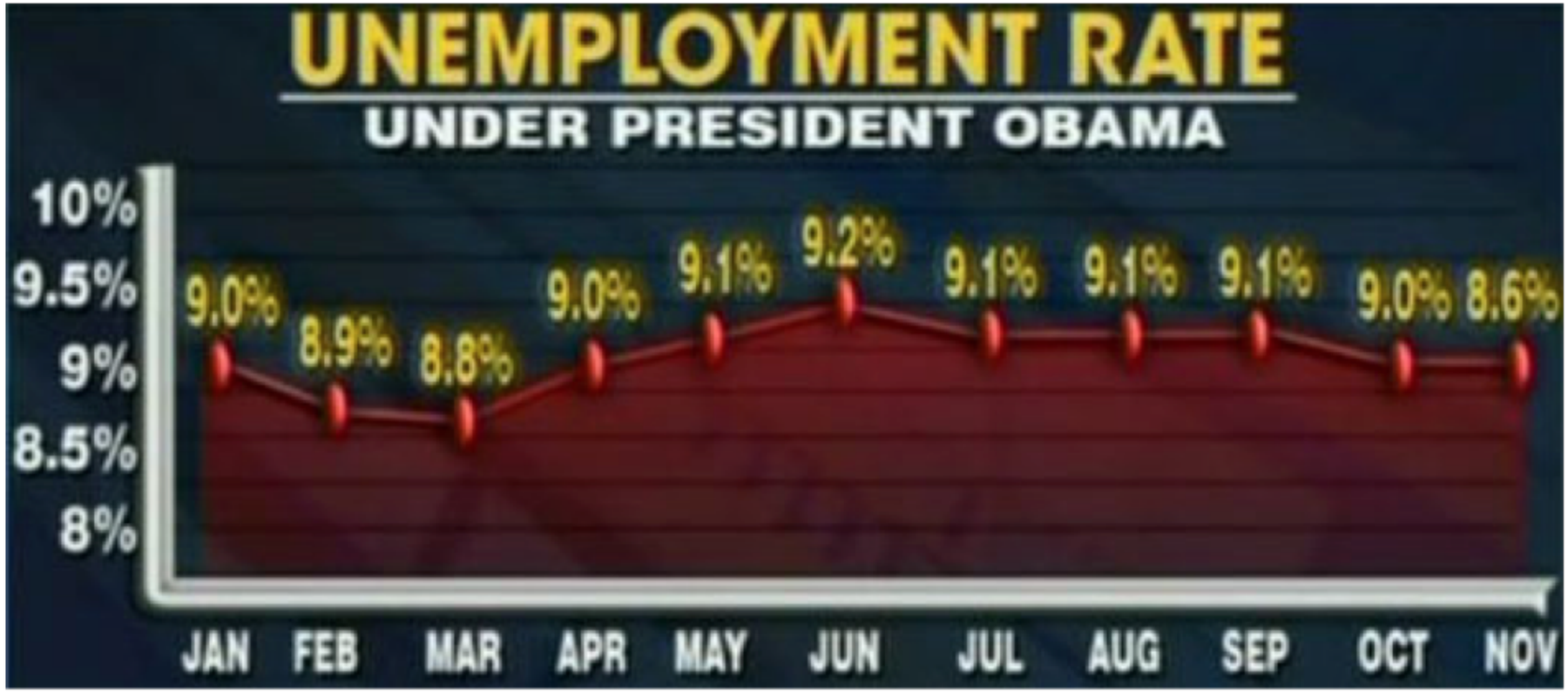
Encode

Noise

Decode

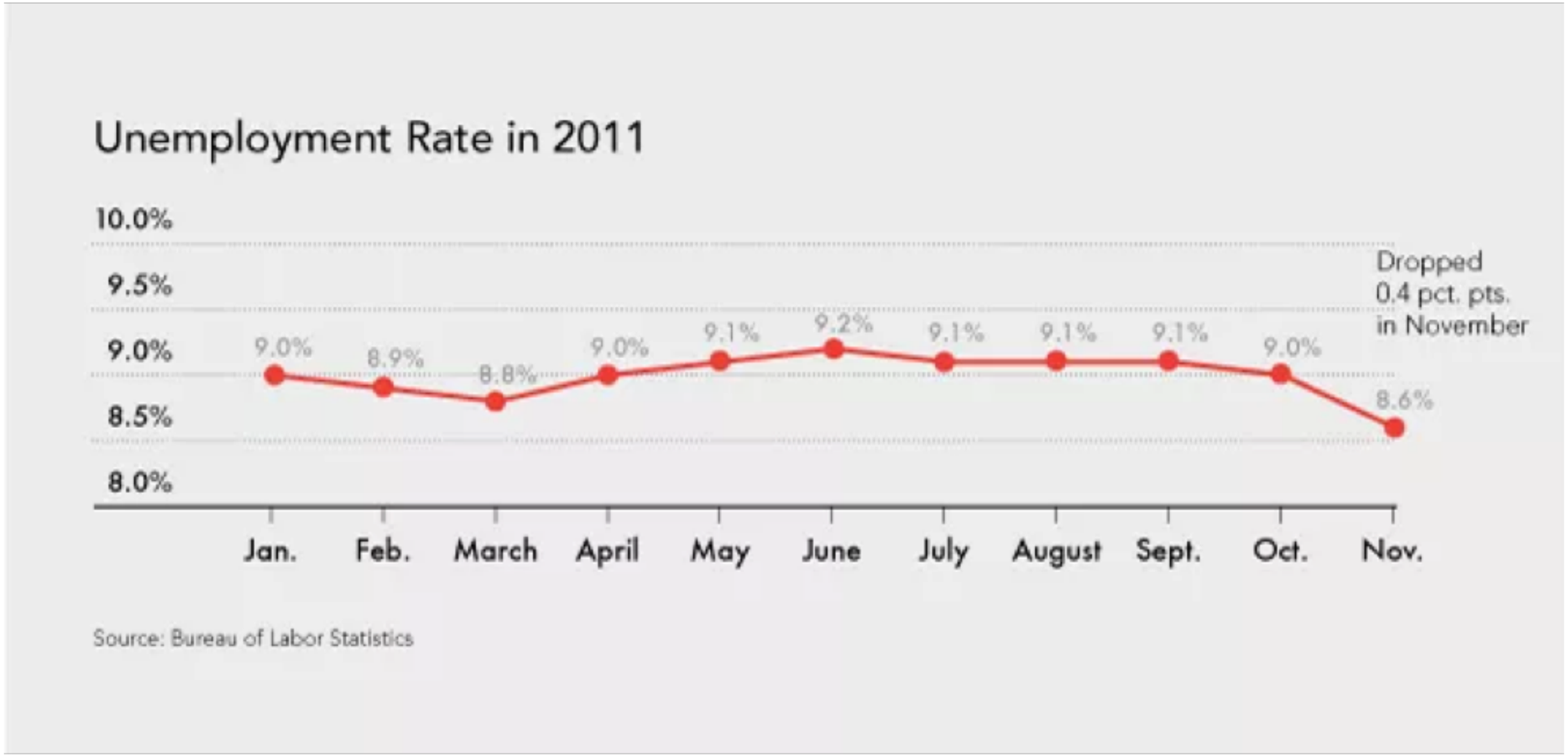
The Bad...

Graphical Integrity



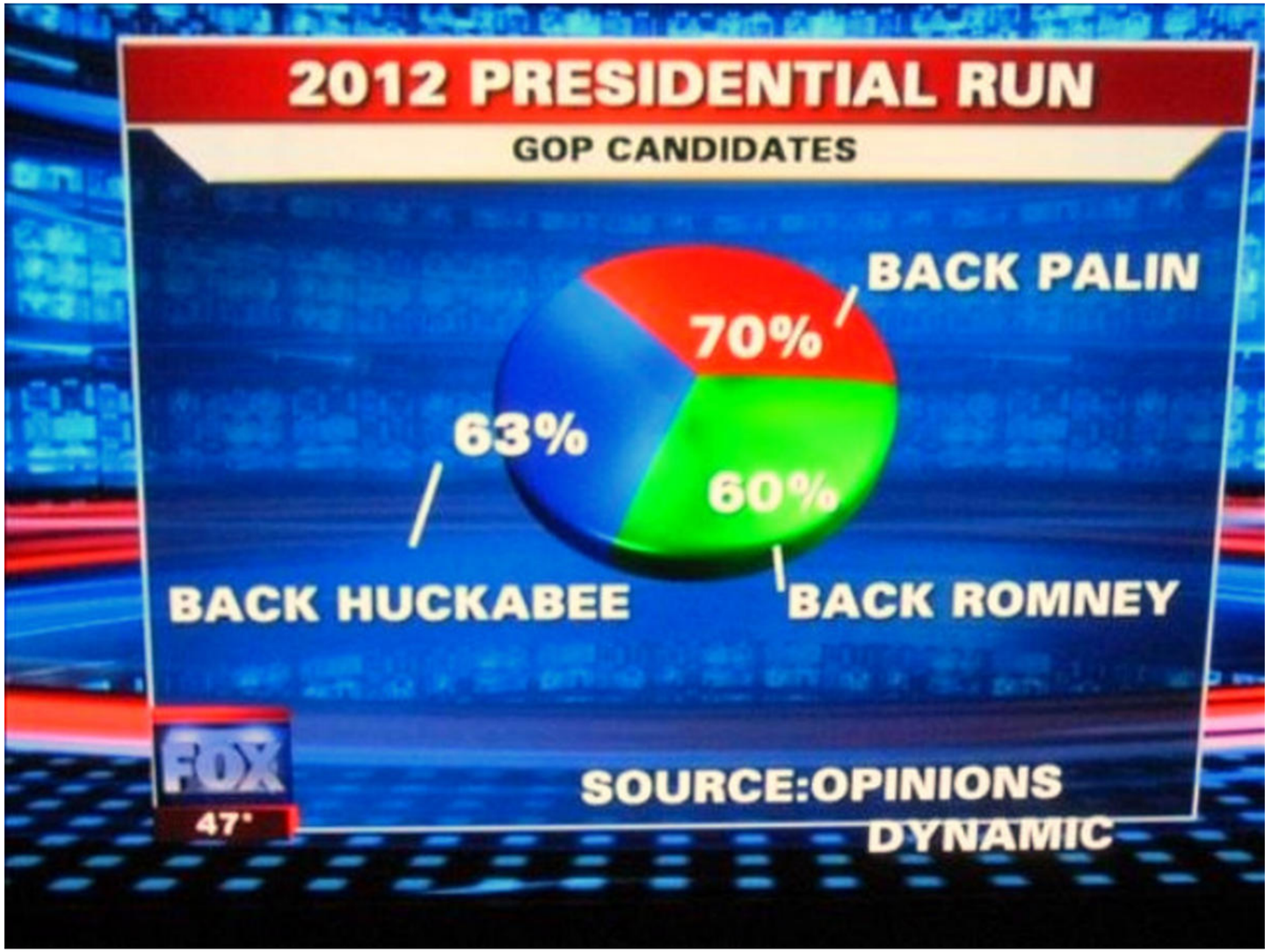
flowingdata.com

Graphical Integrity

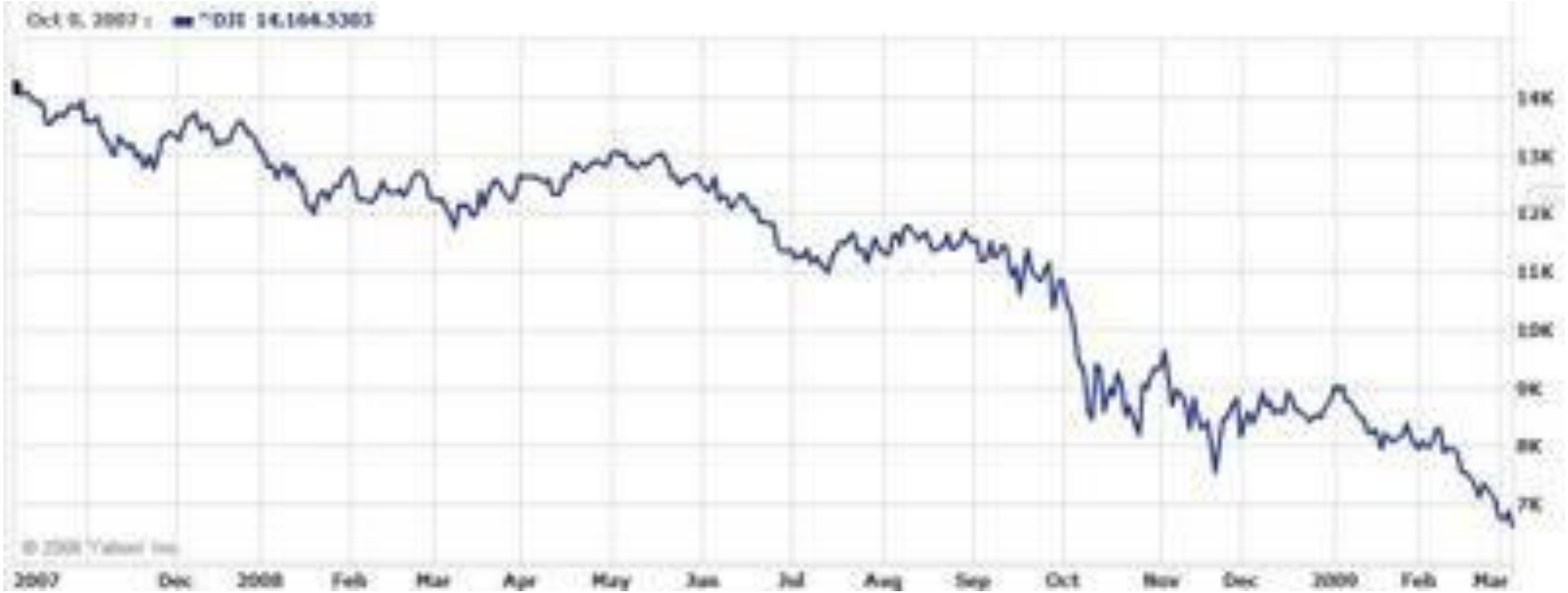


flowingdata.com

Graphical Integrity?

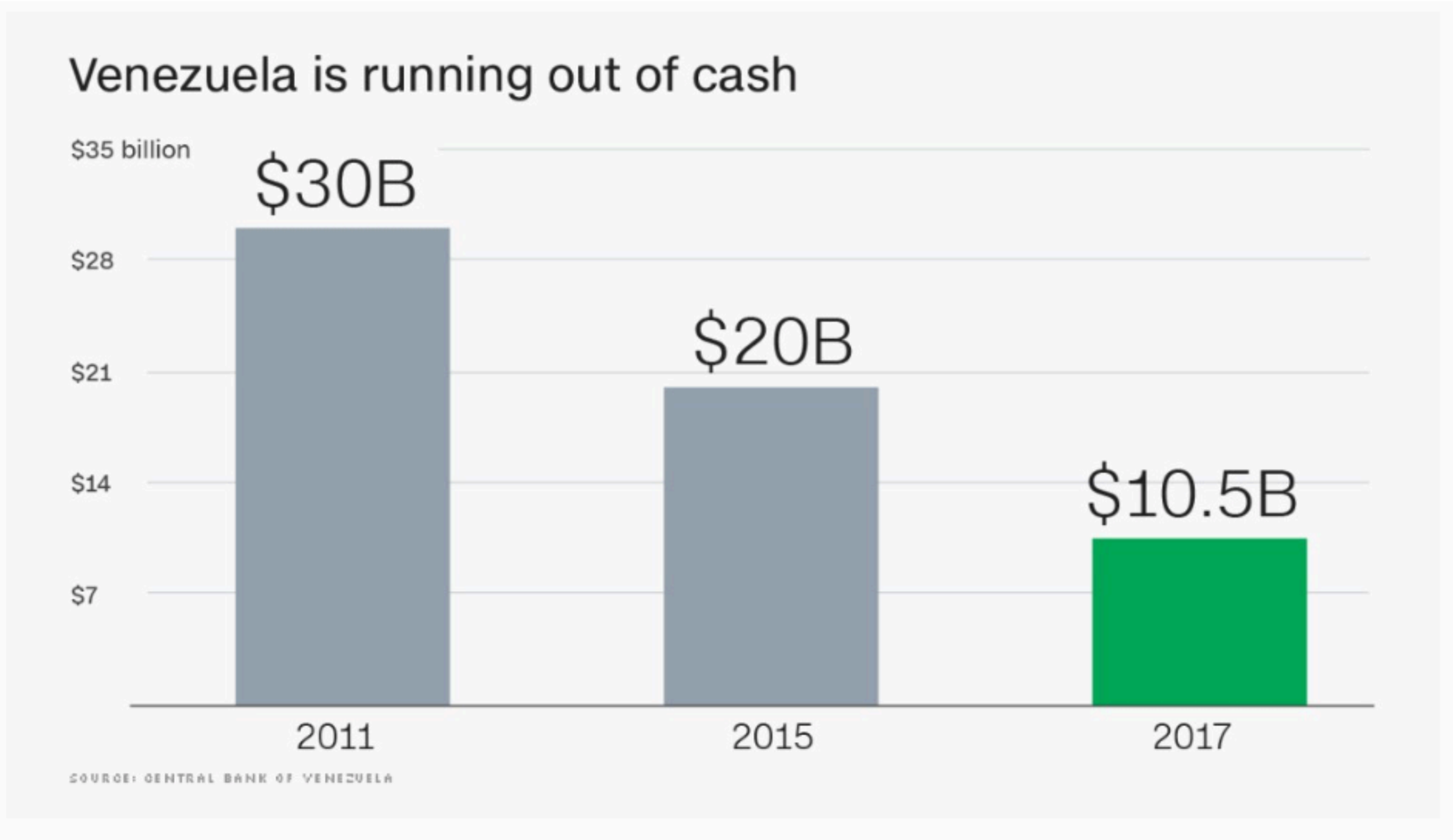



STOCK MARKET SLIDE
NOV. 4, 2008
ELECTION DAY
9625.30
JAN. 20, 2009
INAUGURATION DAY
7949.10
FRIDAY
MAR. 5, 2009
6626.94
SOURCE: BUSINESSWEEK
6:31 CT
MSNBC

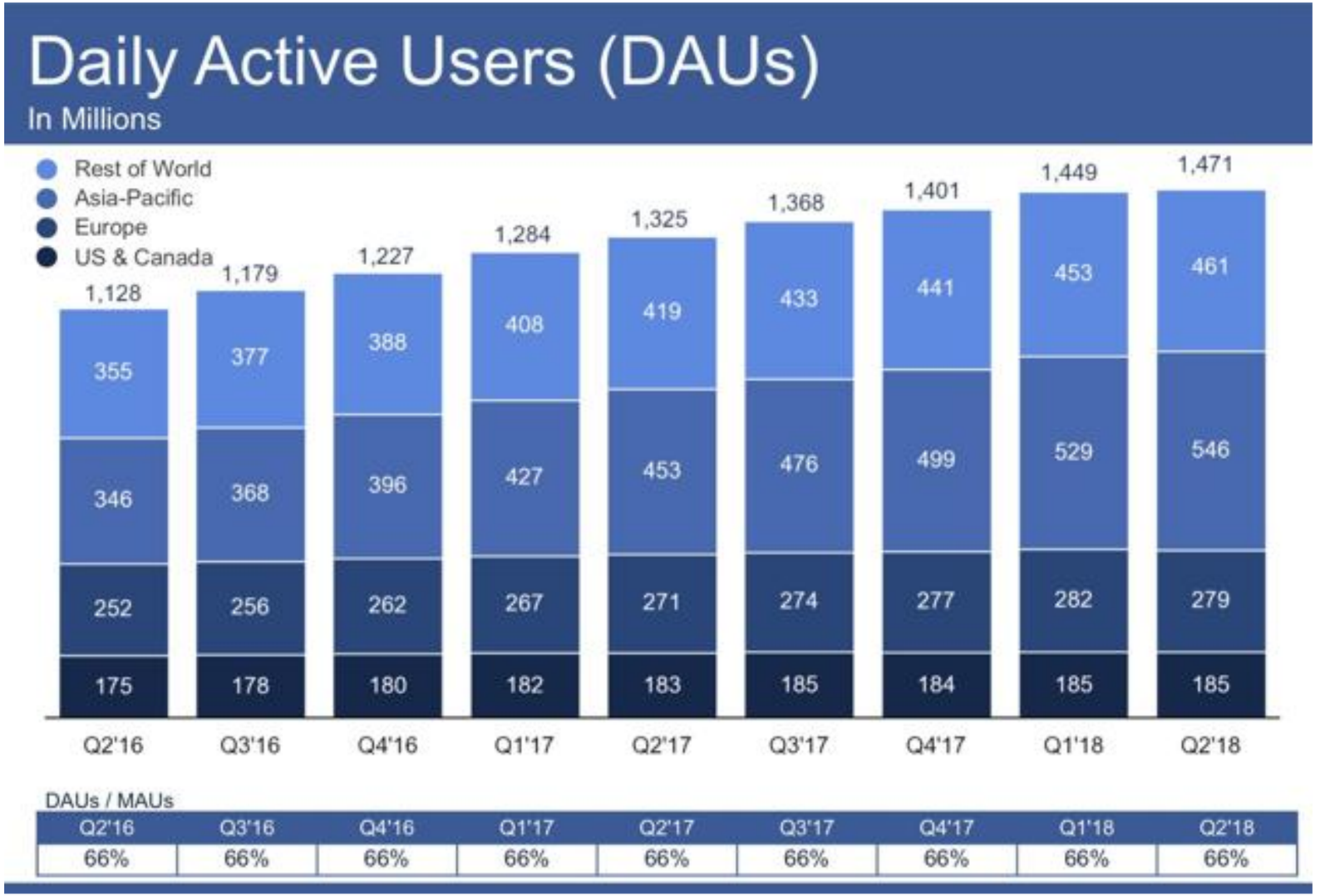

Oct 9, 2007 :
^DJI 14,164.5303
14K
13K
12K
11K
10K
9K
8K
7K
2007
Dec
2008
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2009
Feb
Mar


GEORGIA
PRESIDENT
FEELINGS ABOUT THE FEDERAL GOVERNMENT
SATISFIED
D CLINTON 81%
R TRUMP 14%
L JOHNSON 4%
CNN EXIT POLL
26:23
NEXT POLLS CLOSE
LIVE
CNN
CNN ELECTORAL MAP
19 TRUMP
3 CLINTON
IN PRESIDENT
R TRUMP 126,791 65.4%
D CLINTON 59,284 30.6%

Graphical Integrity

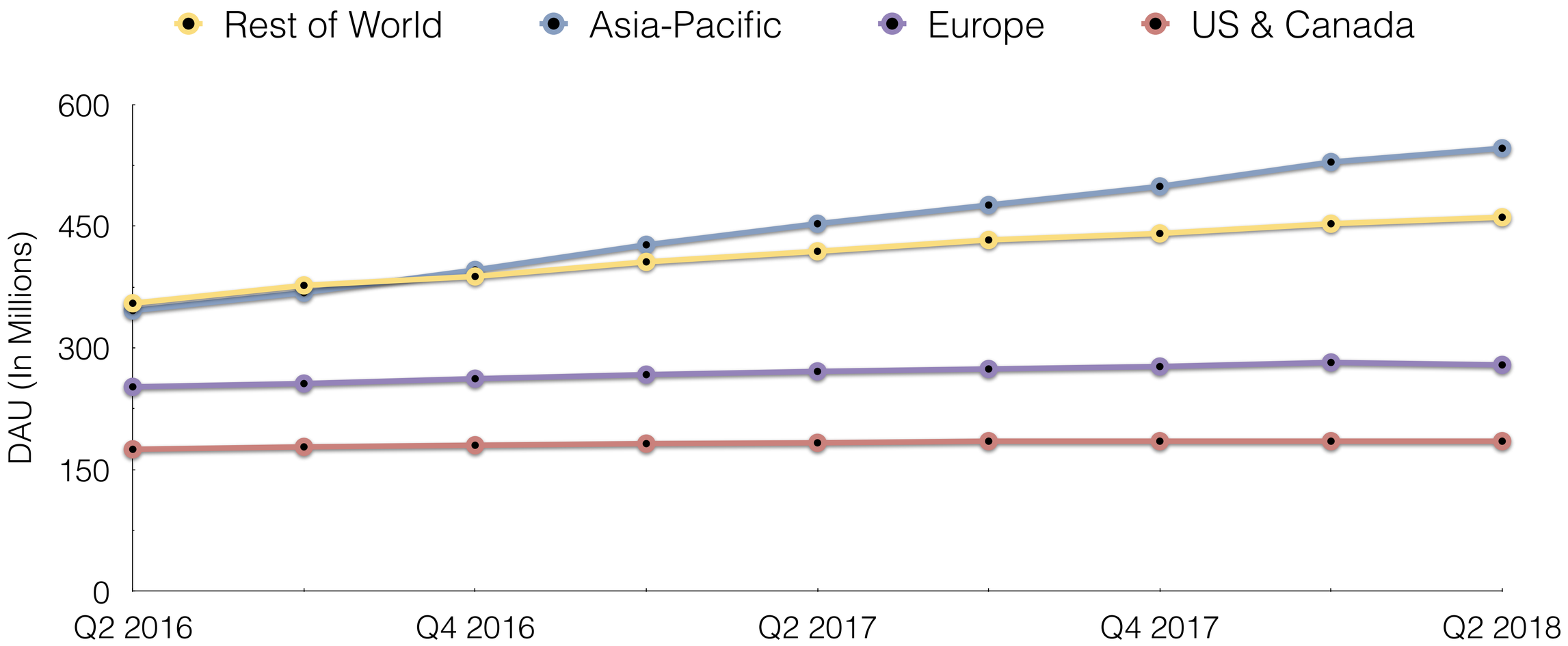


Proper Display

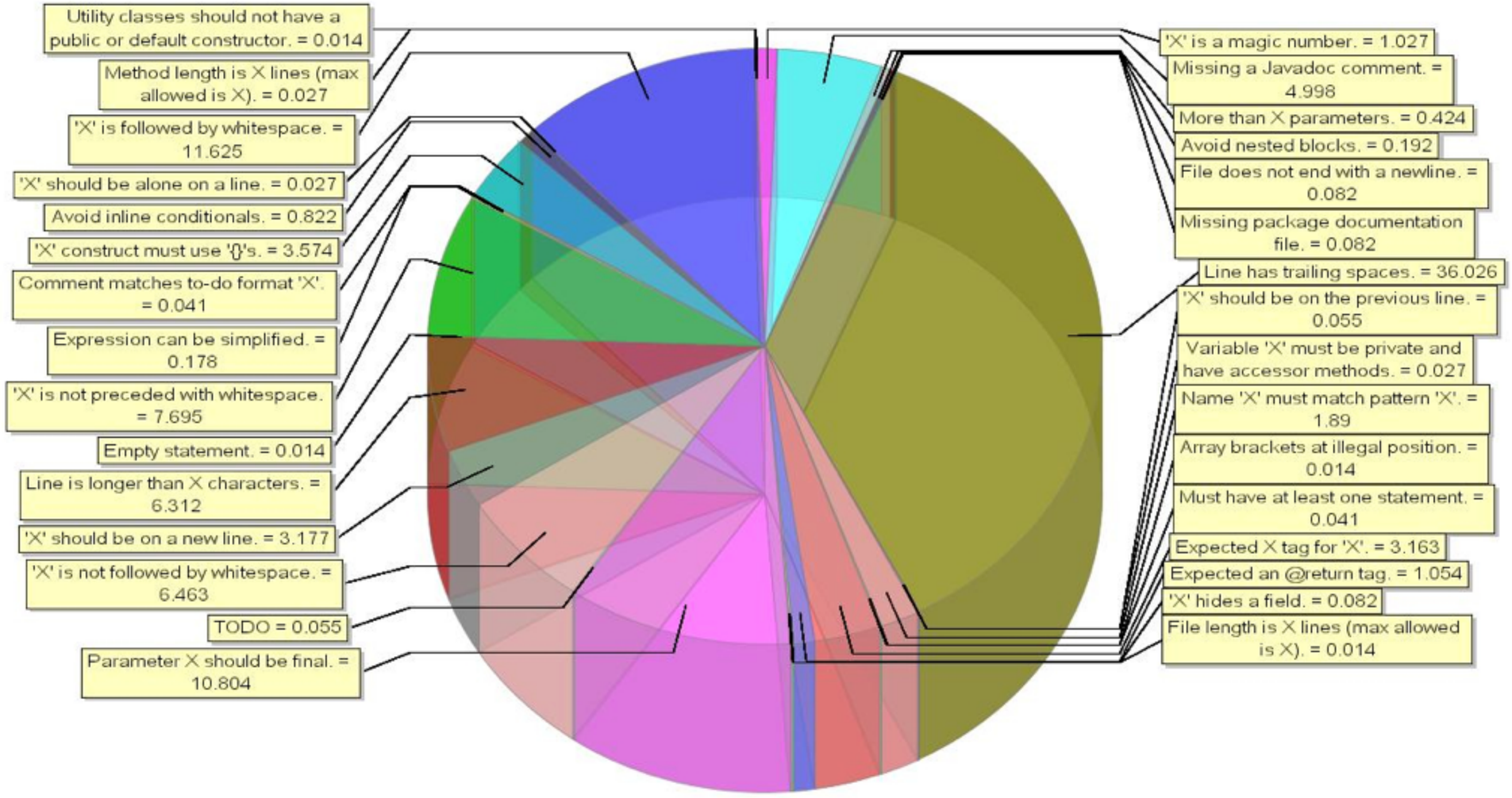


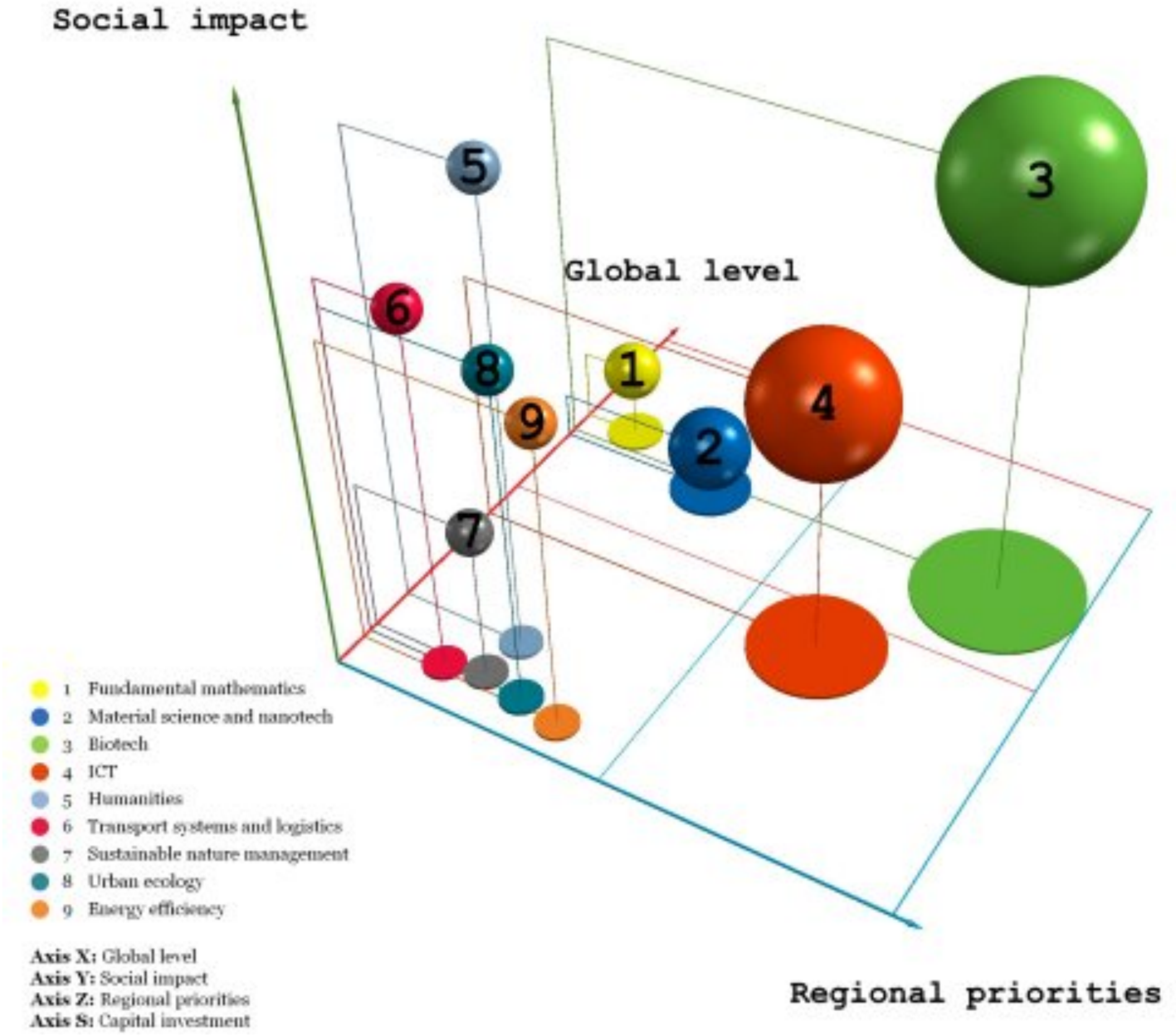
https://s21.q4cdn.com/399680738/files/doc_financials/2018/Q2/Q2-2018-Earnings-Presentation.pdf

Proper Display

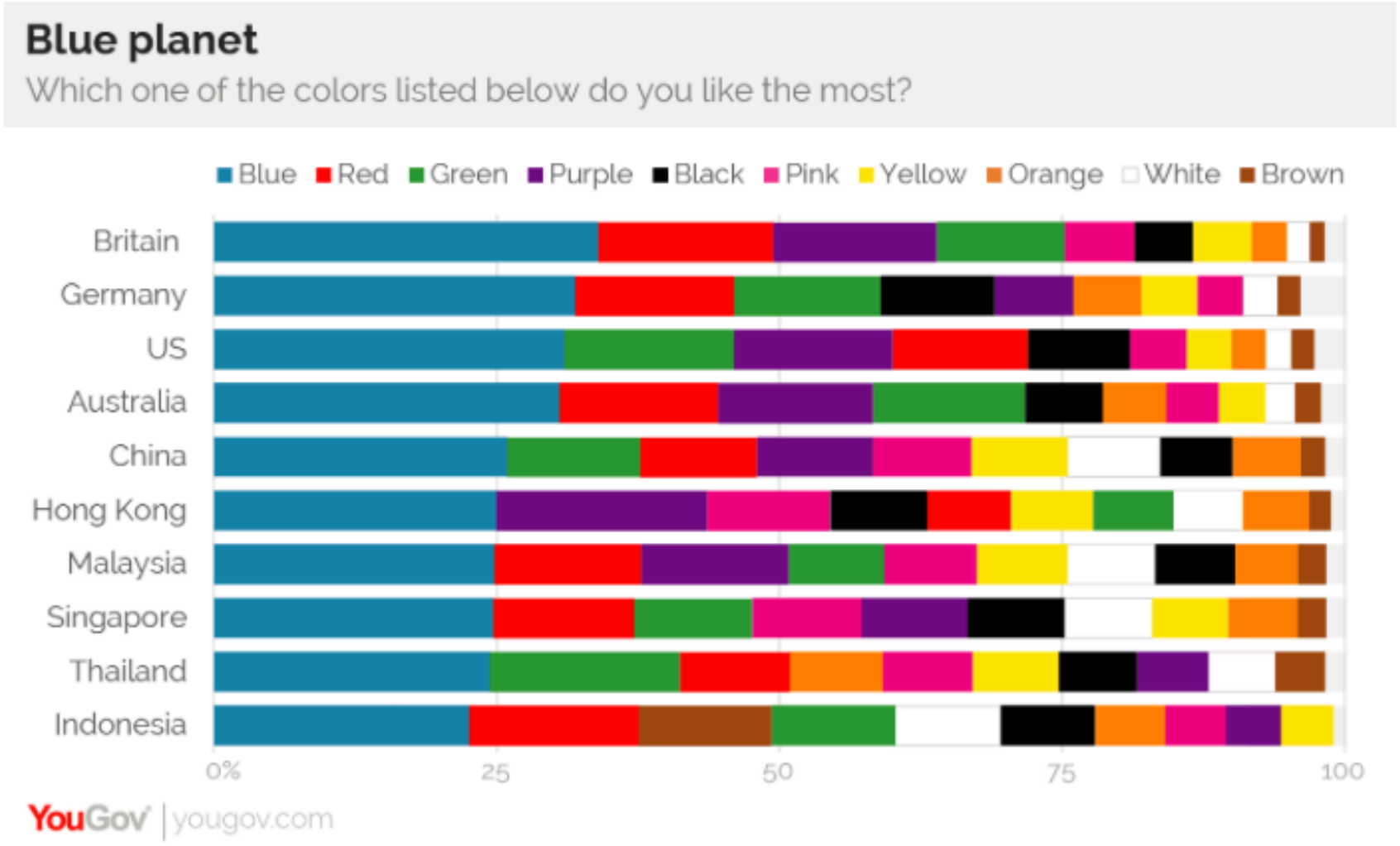


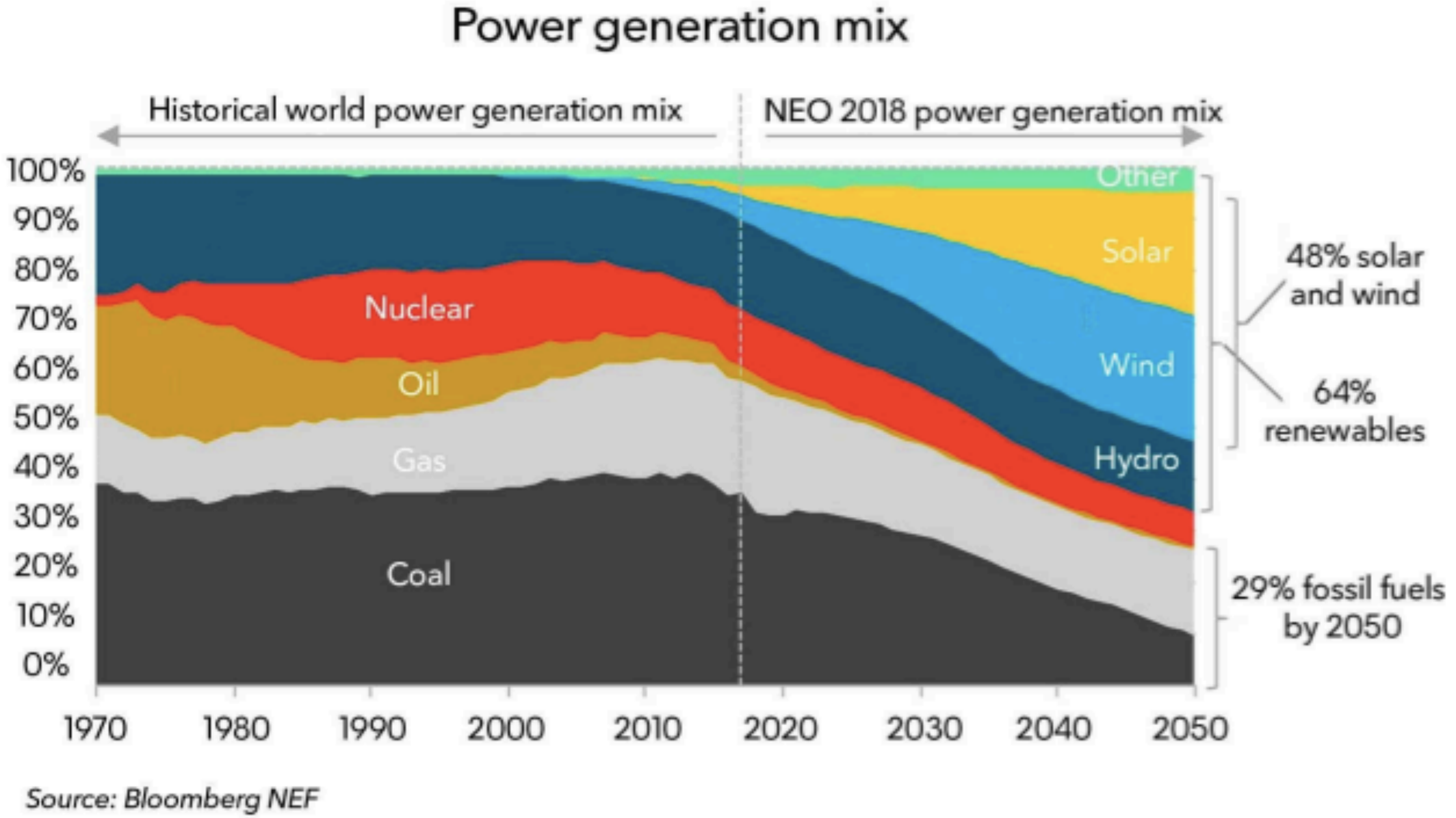
Simple


Utility classes should not have a public or default constructor. = 0.014
Method length is X lines (max allowed is X). = 0.027
'X' is followed by whitespace. = 11.625
'X' should be alone on a line. = 0.027
Avoid inline conditionals. = 0.822
'X' construct must use '{}'s. = 3.574
Comment matches to-do format 'X'. = 0.041
Expression can be simplified. = 0.178
'X' is not preceded with whitespace. = 7.695
Empty statement. = 0.014
Line is longer than X characters. = 6.312
'X' should be on a new line. = 3.177
'X' is not followed by whitespace. = 6.463
TODO = 0.055
Parameter X should be final. = 10.804
'X' is a magic number. = 1.027
Missing a Javadoc comment. = 4.998
More than X parameters. = 0.424
Avoid nested blocks. = 0.192
File does not end with a newline. = 0.082
Missing package documentation file. = 0.082
Line has trailing spaces. = 36.026
'X' should be on the previous line. = 0.055
Variable 'X' must be private and have accessor methods. = 0.027
Name 'X' must match pattern 'X'. = 1.89
Array brackets at illegal position. = 0.014
Must have at least one statement. = 0.041
Expected X tag for 'X'. = 3.163
Expected an @return tag. = 1.054
'X' hides a field. = 0.082
File length is X lines (max allowed is X). = 0.014

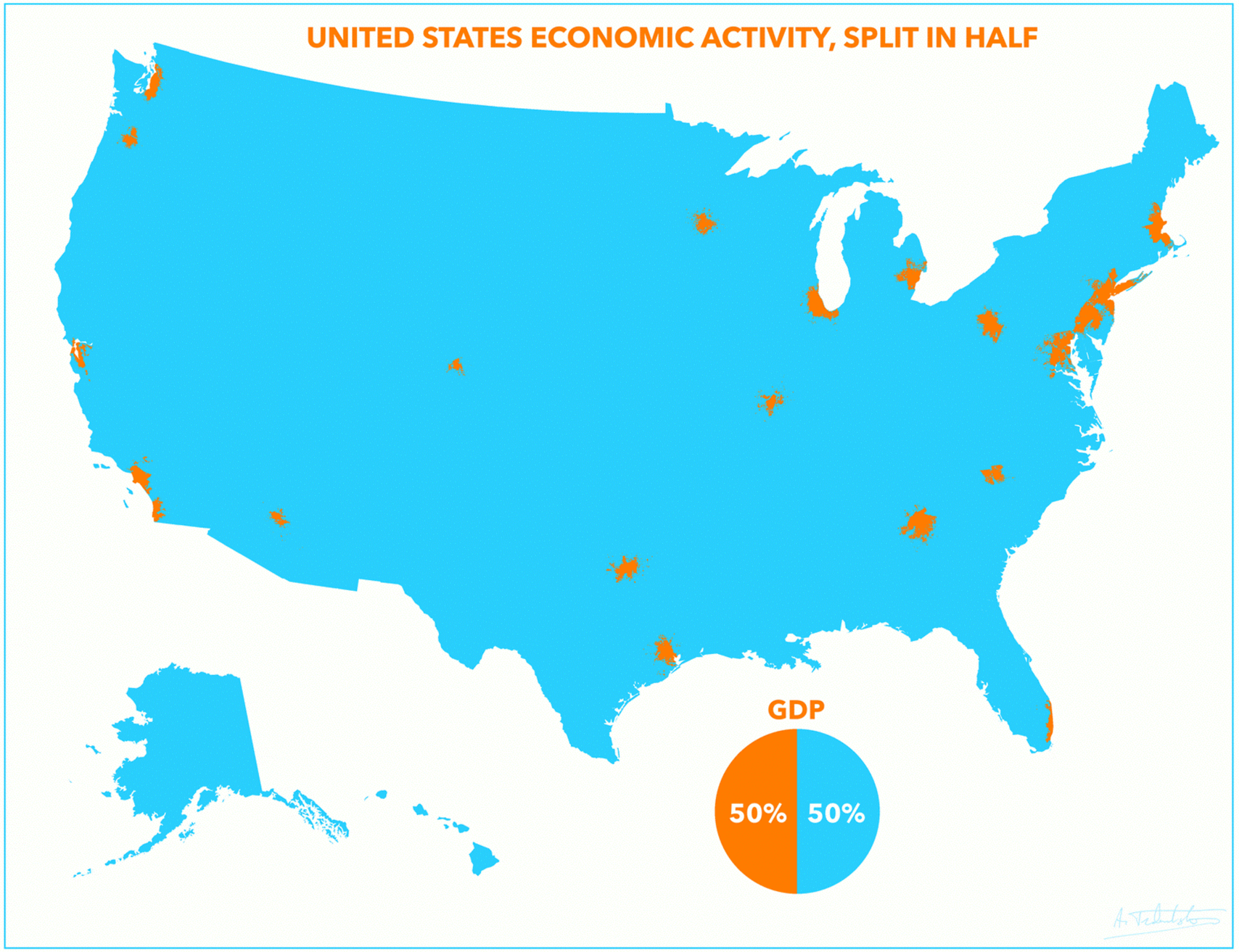


http://livingqlikview.com/the-9-worst-data-visualizations-ever-created/


Blue planet
Which one of the colors listed below do you like the most?
Blue
Red
Green
Purple
Black
Pink
Yellow
Orange
White
Brown
Britain
Germany
US
Australia
China
Hong Kong
Malaysia
Singapore
Thailand
Indonesia
0%
25
50
75
100
YouGov | yougov.com

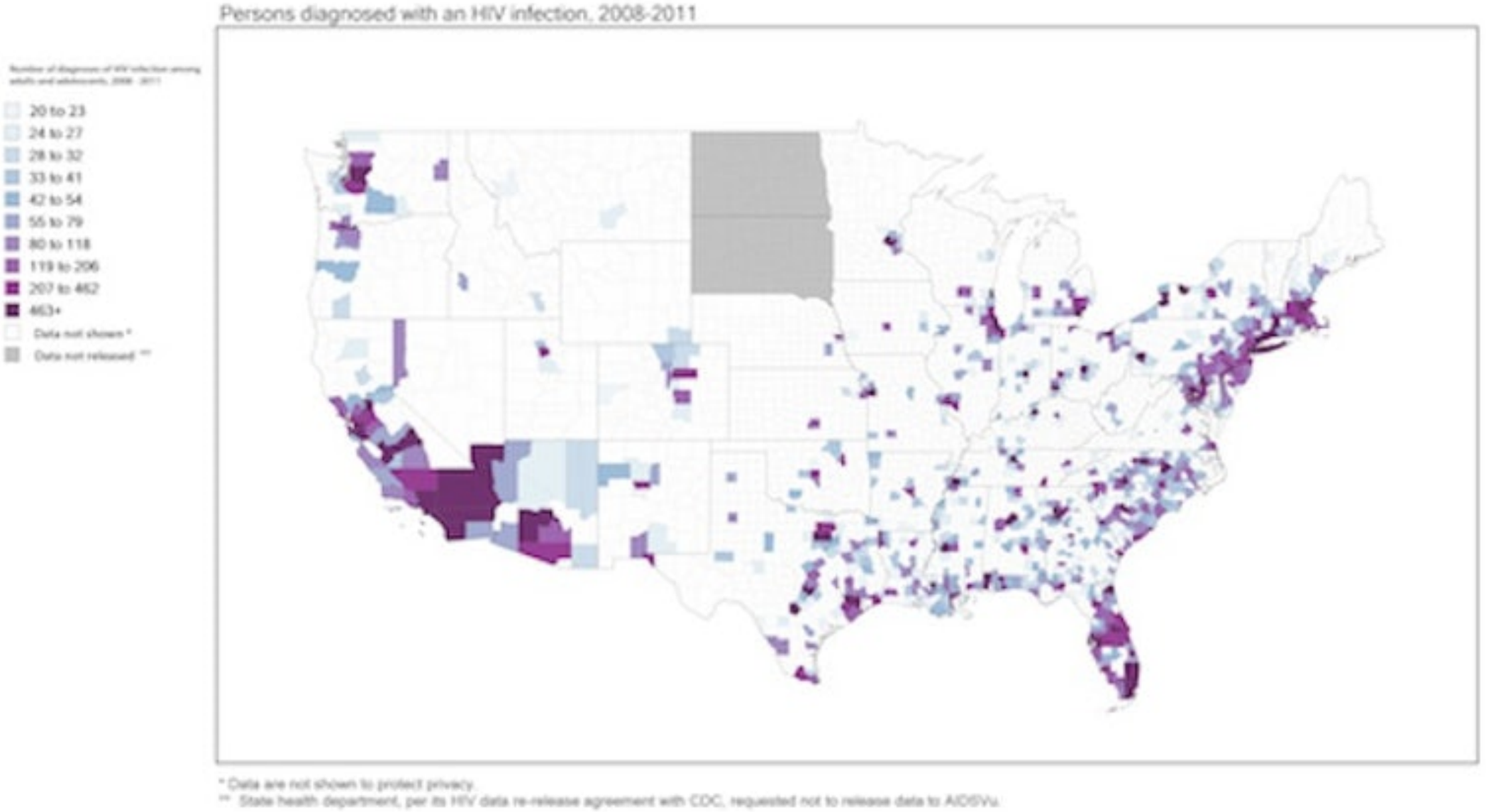

Power generation mix
Historical world power generation mix
NEO 2018 power generation mix
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Other
Solar
Wind
Hydro
Nuclear
Oil
Gas
Coal
48% solar and wind
64% renewables
29% fossil fuels by 2050
1970
1980
1990
2000
2010
2020
2030
2040
2050
Source: Bloomberg NEF

GTK CYBER
GET ON THE FAST TRACK



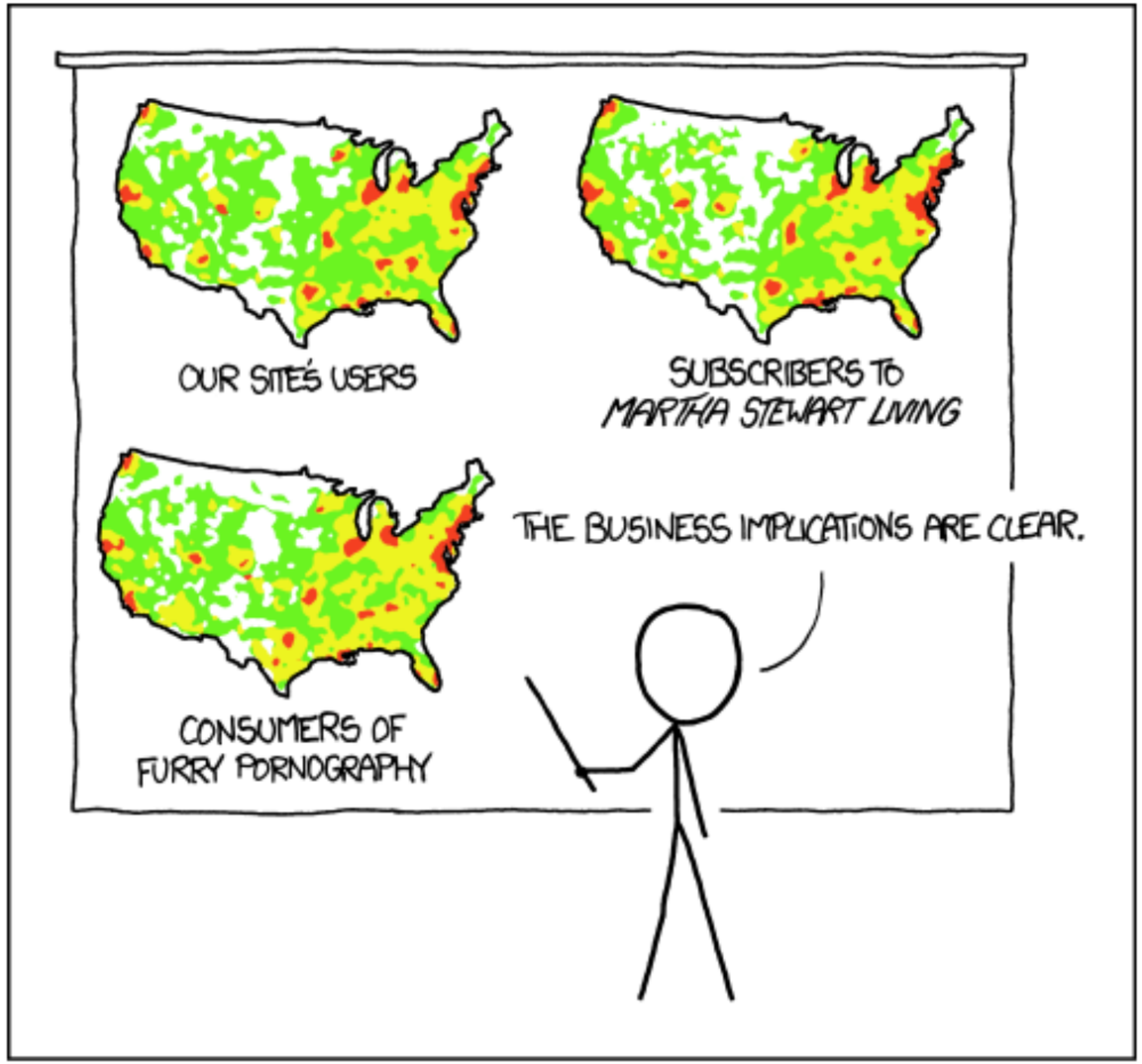
http://www.thefunctionalart.com/2014/02/the-incredible-gdp-map-that-shows-that.html

Startling New Map: 92 Percent of New HIV Cases are in 25 Percent of Counties



gtkcybe

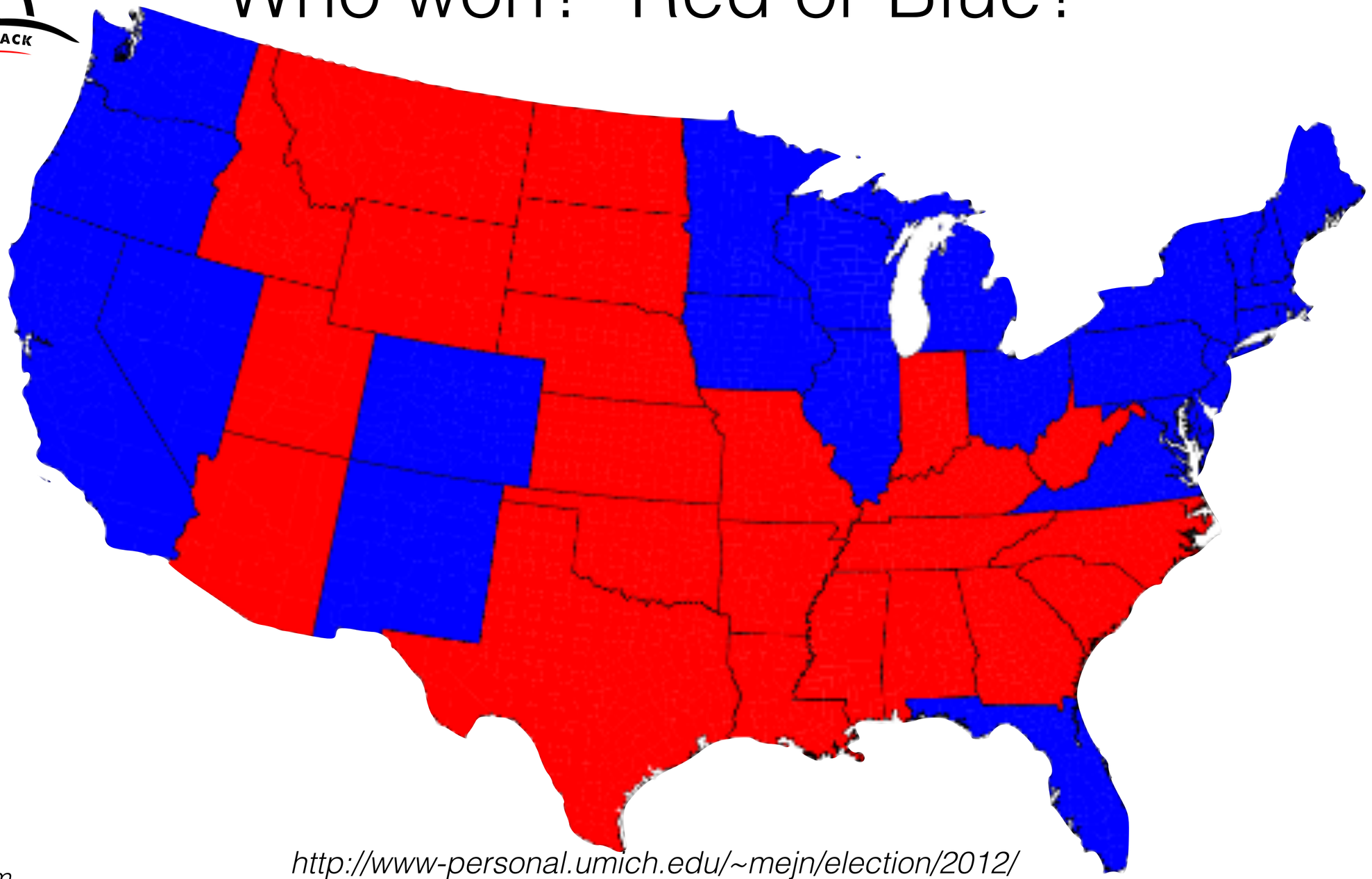
https://newrepublic.com/article/114822/maps-92-percent-new-hiv-cases-are-25-percent-counties



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

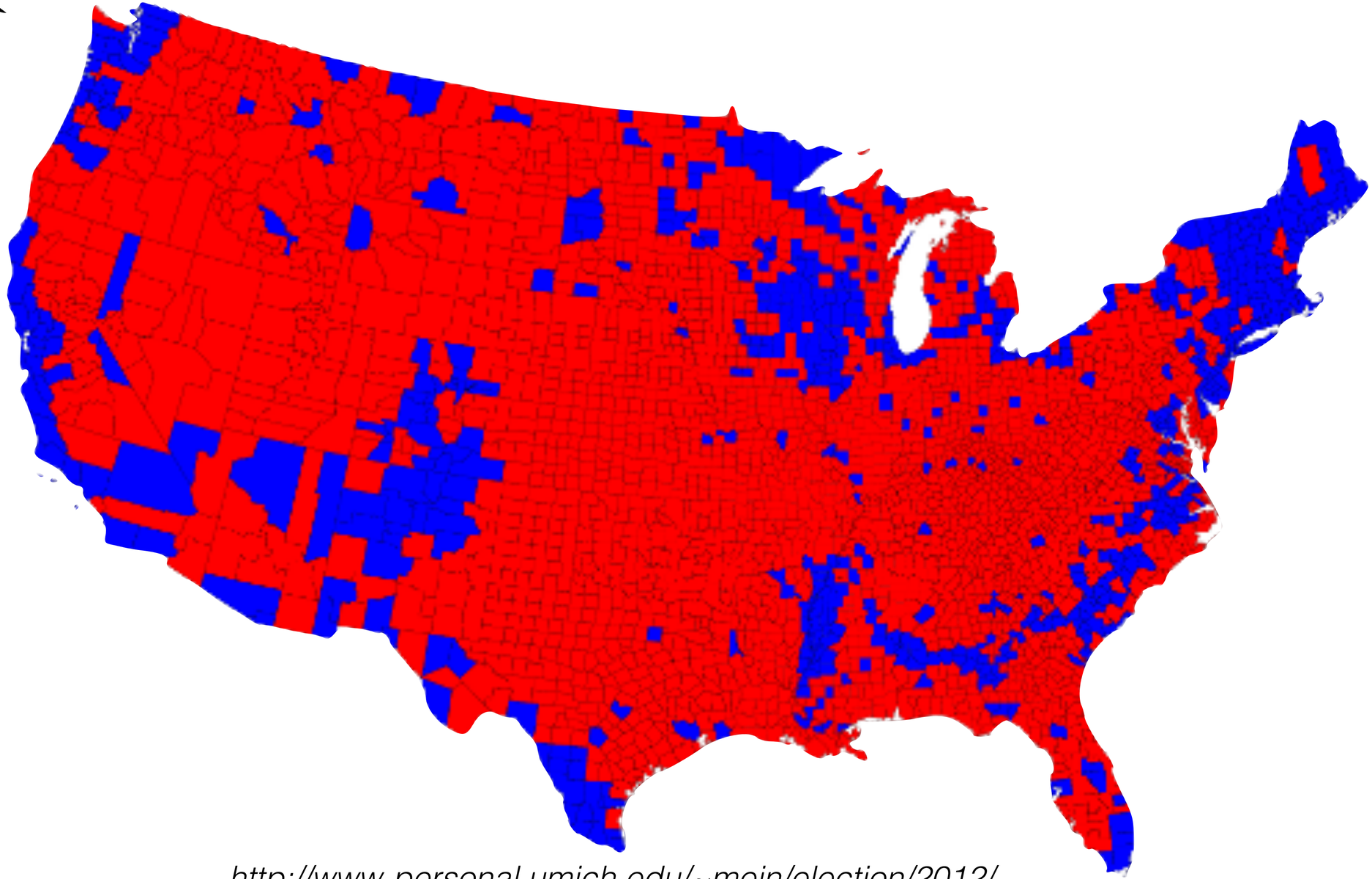
https://xkcd.com/1138/

Who won? Red or Blue?



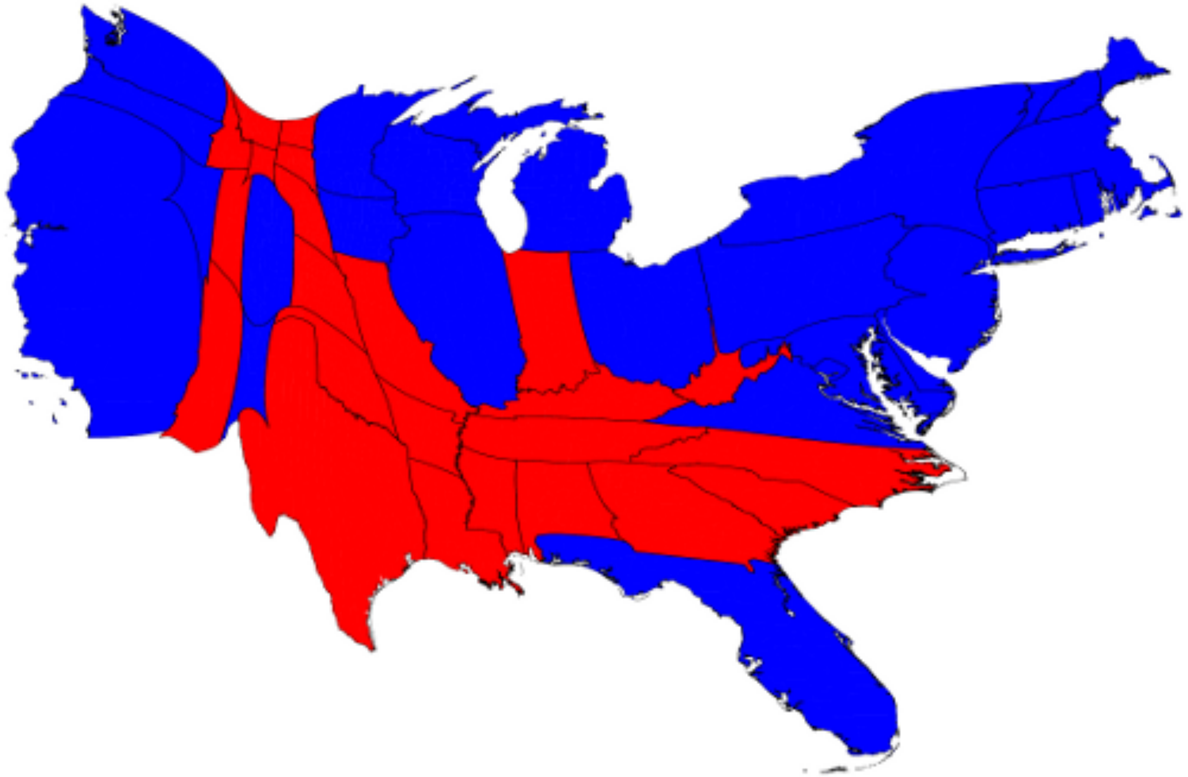
http://www-personal.umich.edu/~mejn/election/2012/

Who won? Red or Blue?

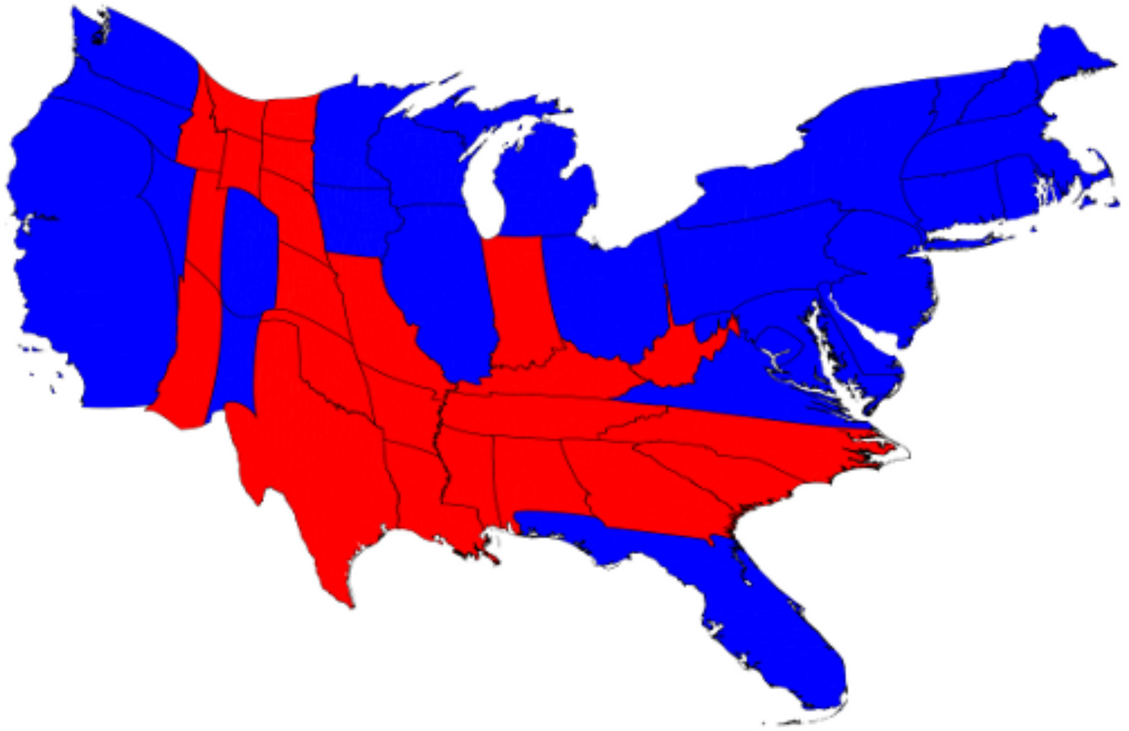


http://www-personal.umich.edu/~mejn/election/2012/

Who won? Red or Blue?



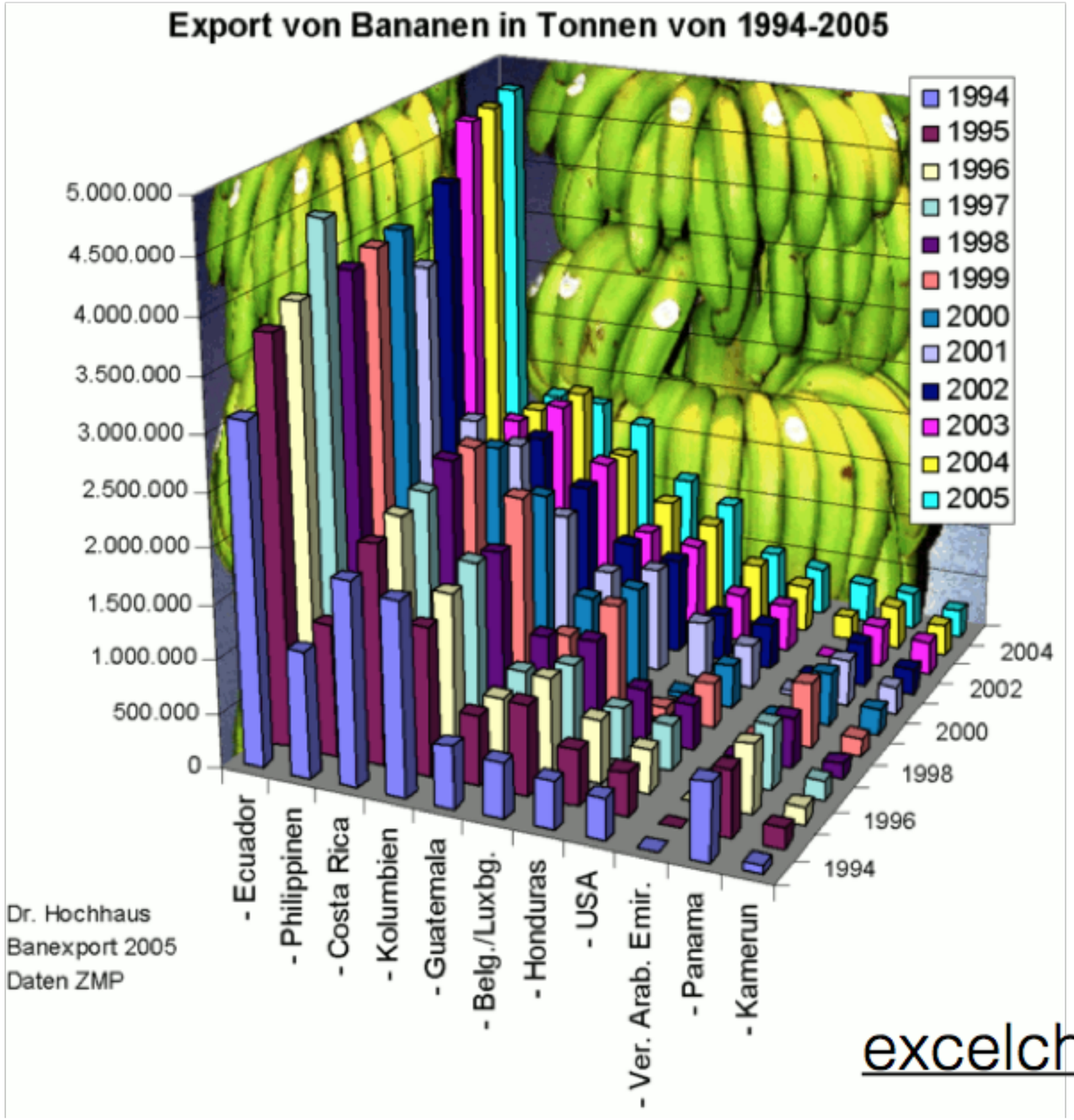
Scaled by Population



Scaled by Electoral Votes

http://www-personal.umich.edu/~mejn/election/2012/

The ugly

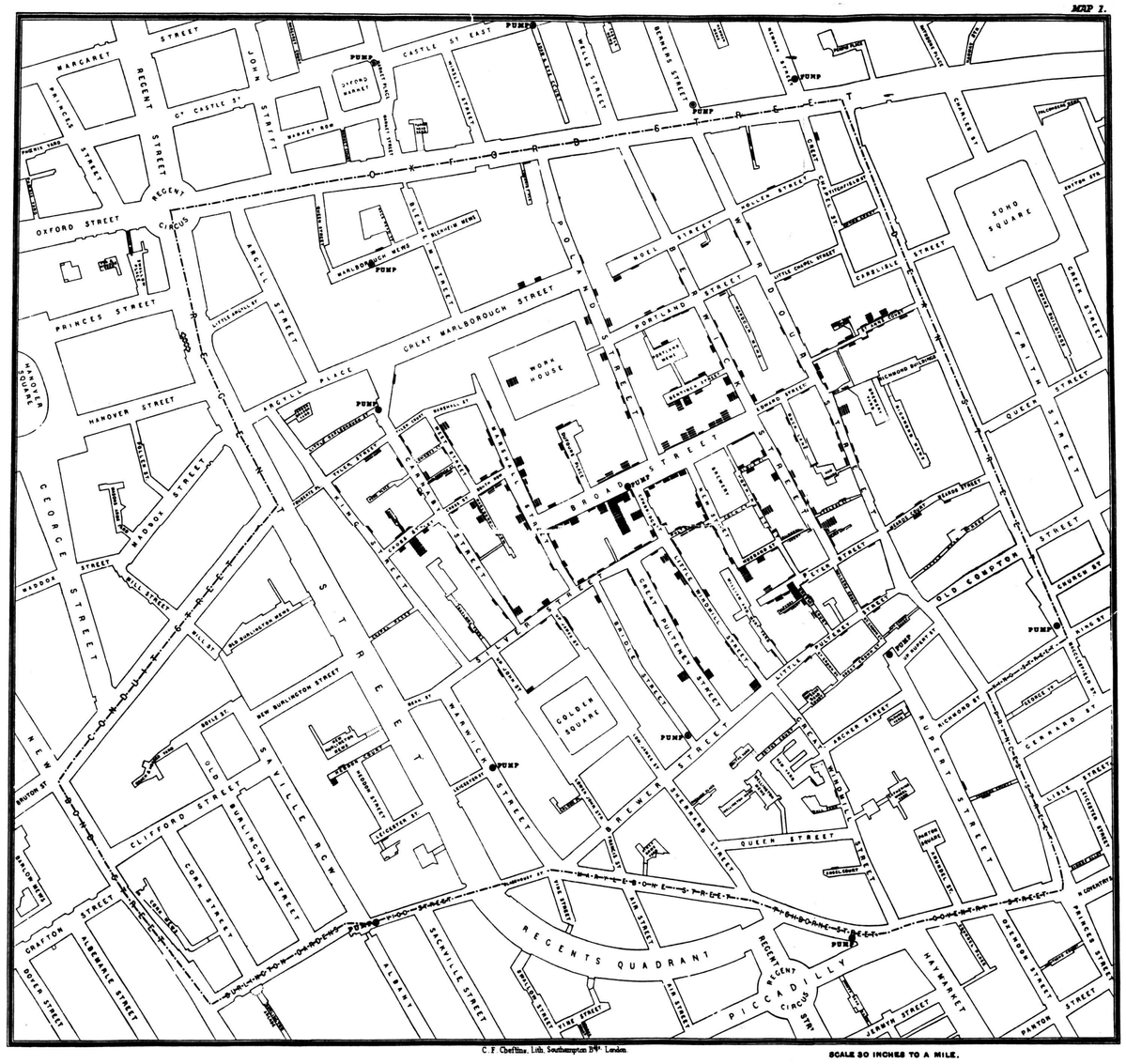


excelcharts.com

The good

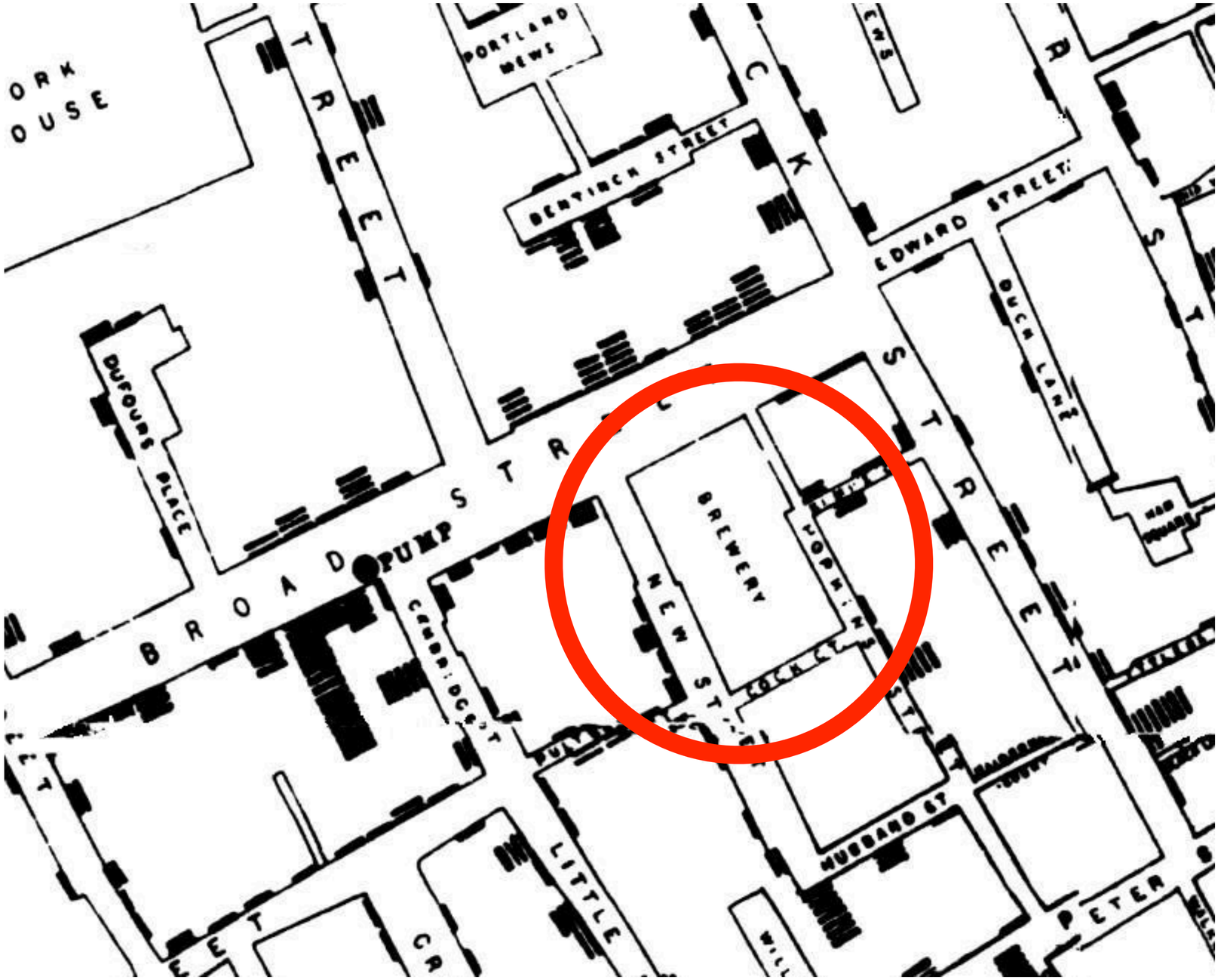
Proper Display

Google Finance



- Created by John Snow in 1854
- This visualization proved that cholera was water-borne
- Snow identified the source of the outbreak as a water pump on Broad Street
- Spurred city to create a true sewage system

GTK CYBER
GET ON THE FAST TRACK

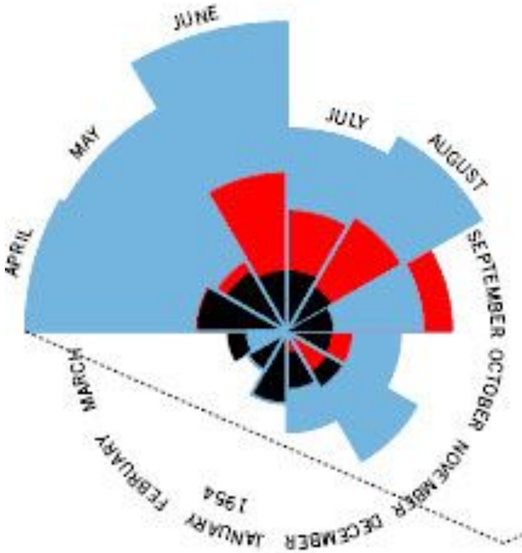


gtkcyber.com

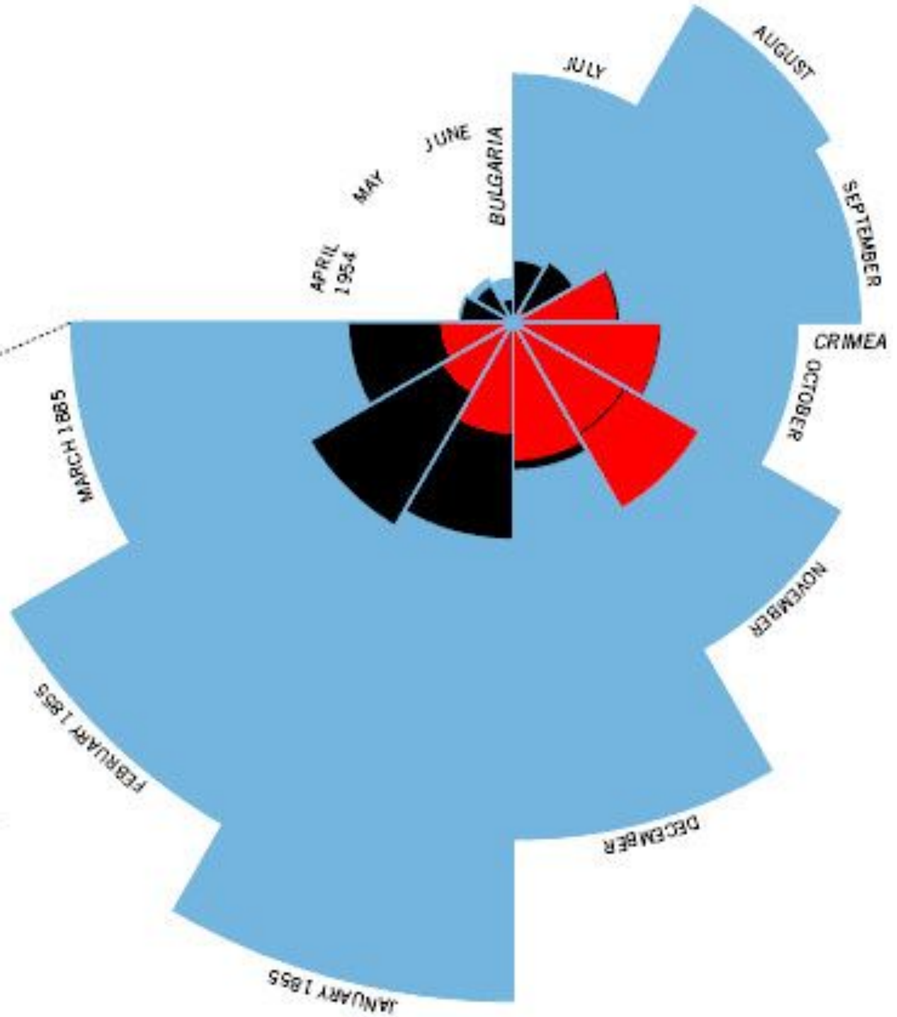
https://www.youtube.com/watch?v=YpKbO6O3O3M

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST

2.
APRIL 1855 TO MARCH 1856

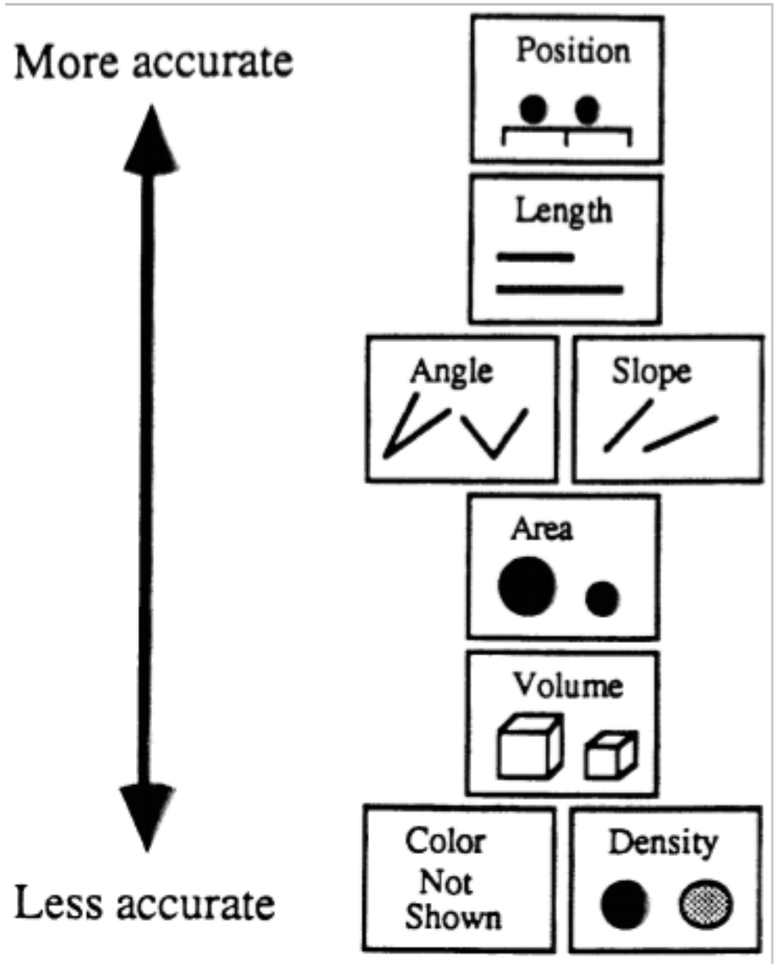


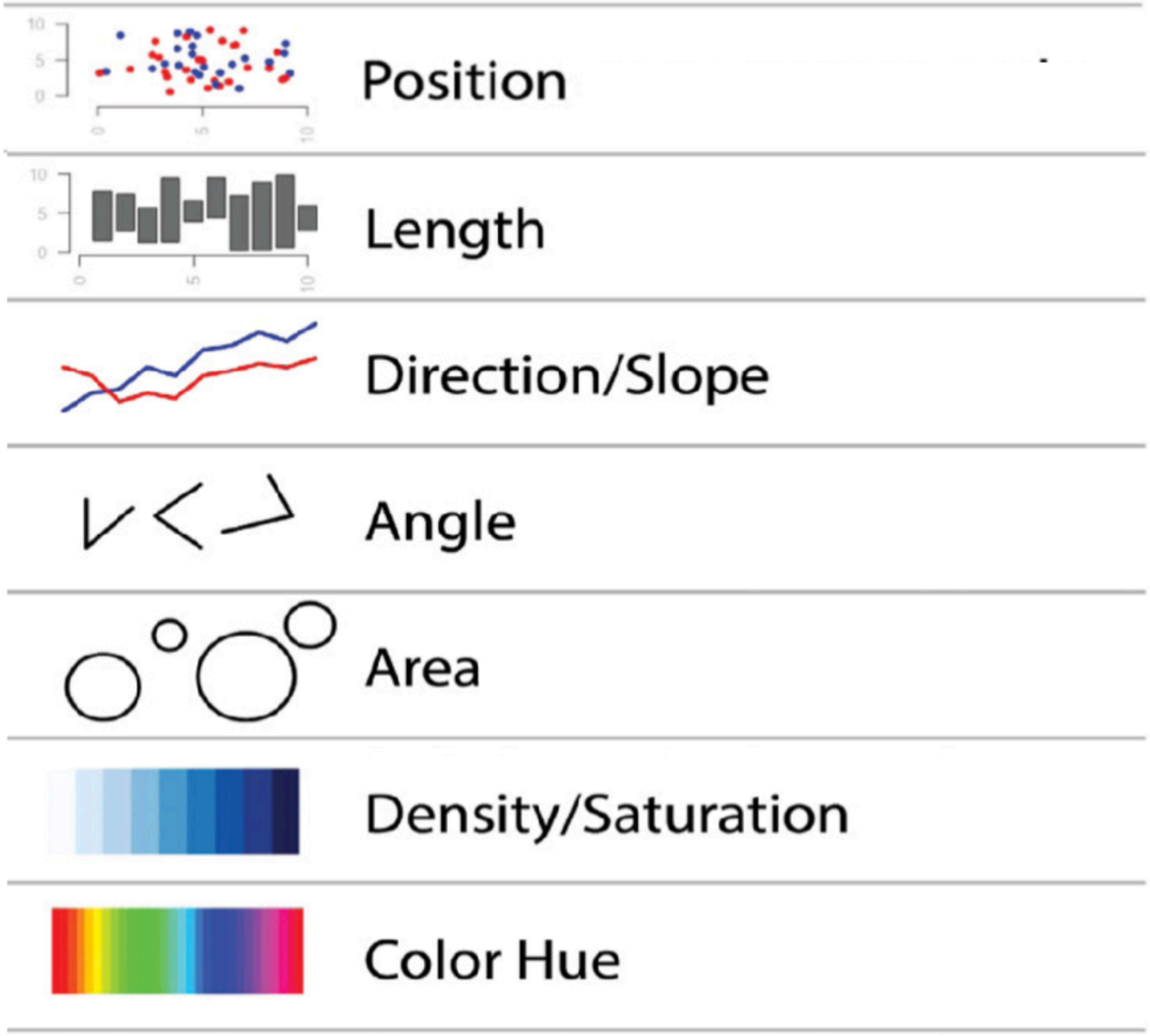
1.
APRIL 1854 TO MARCH 1855



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes
The black line across the red triangle in Nov' 1854 marks the boundary of the deaths from all other causes during the month
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

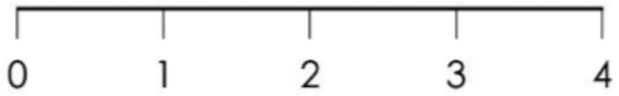
Visual Encoding


More accurate
Position
Length
Angle
Slope
Area
Volume
Color Not Shown
Density
Less accurate


Position
Length
Direction/Slope
Angle
Area
Density/Saturation
Color Hue

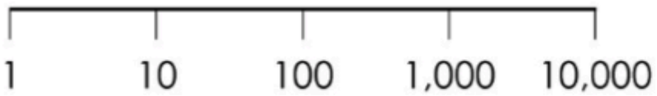
Linear

Values are evenly spaced



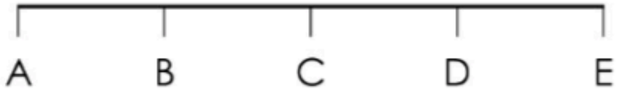
Logarithmic

Focus on percent change



Categorical

Discrete placement in bins

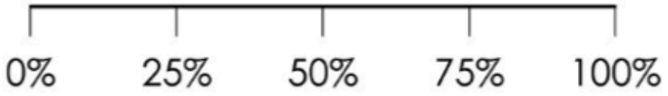


Ordinal

Categories where order matters

Percent

Representing parts of a whole

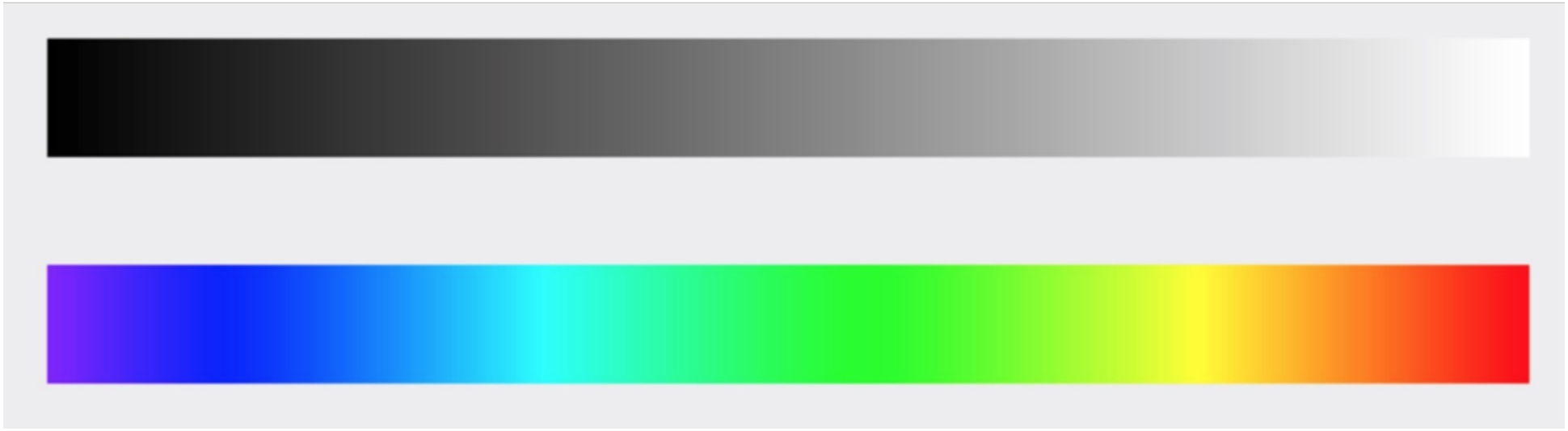


Time

Units of months, days, or hours

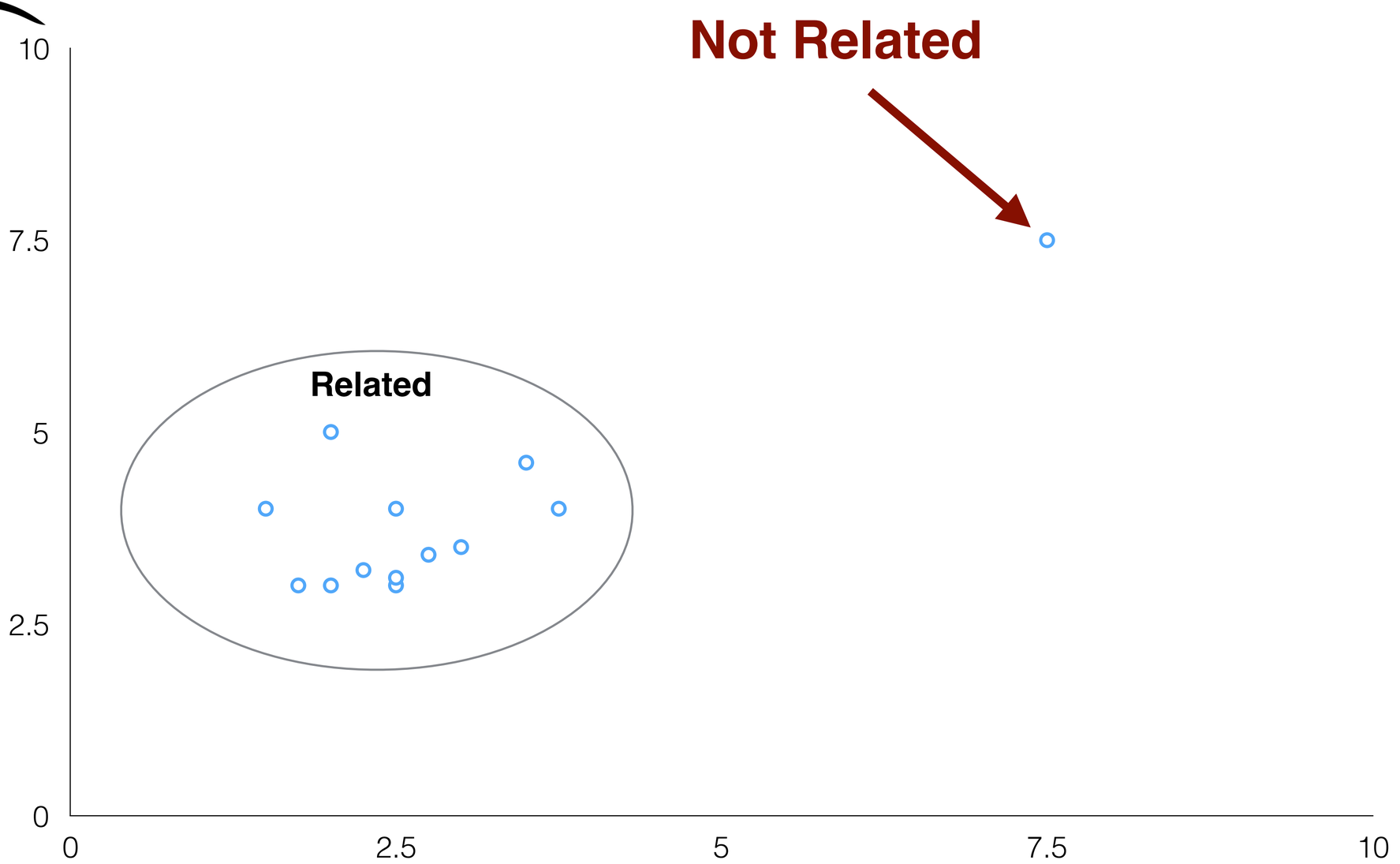
Source: Nathan Yau, Data Points

Color

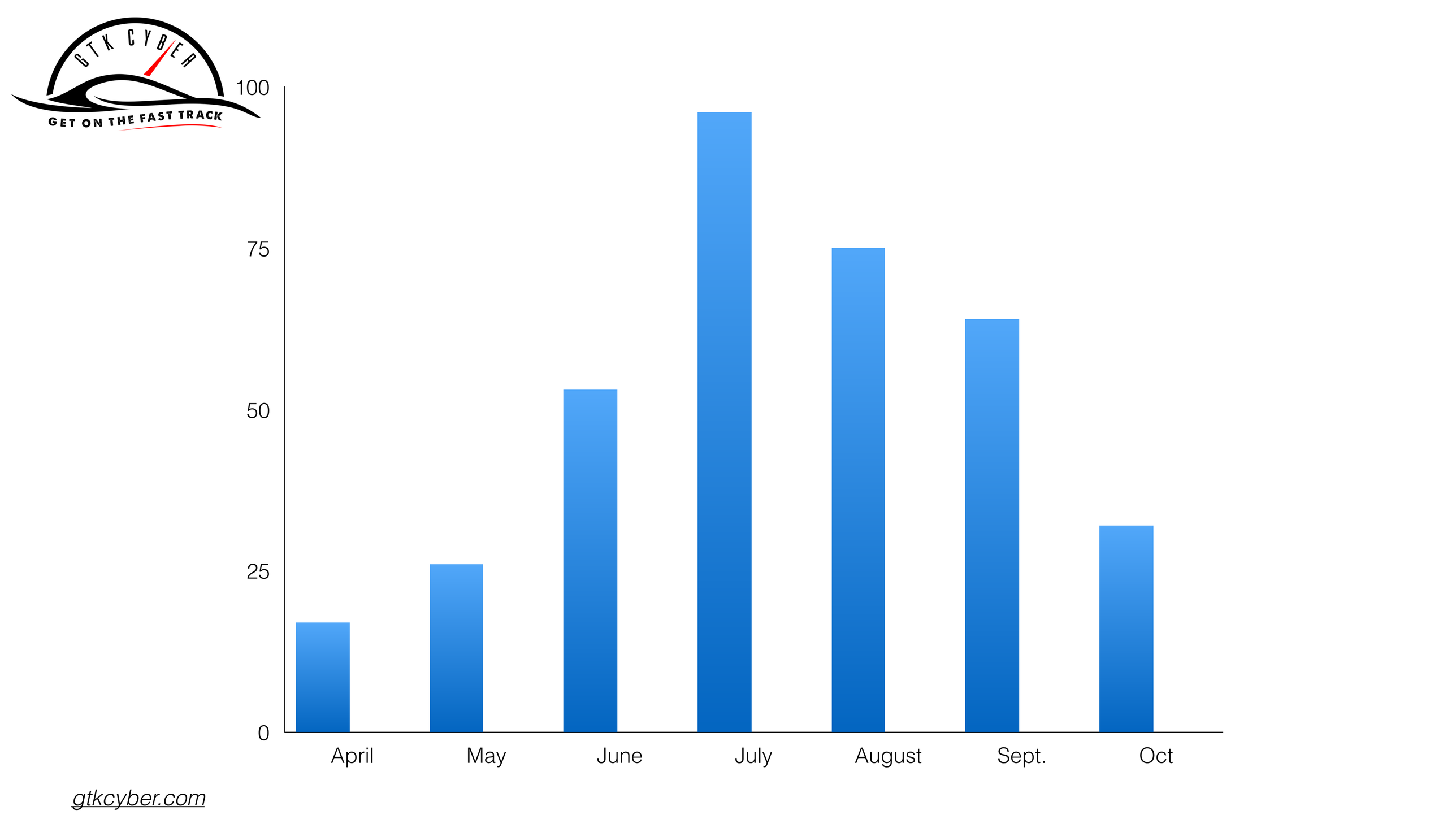


Visual Encodings

Visual Encodings: **Position**

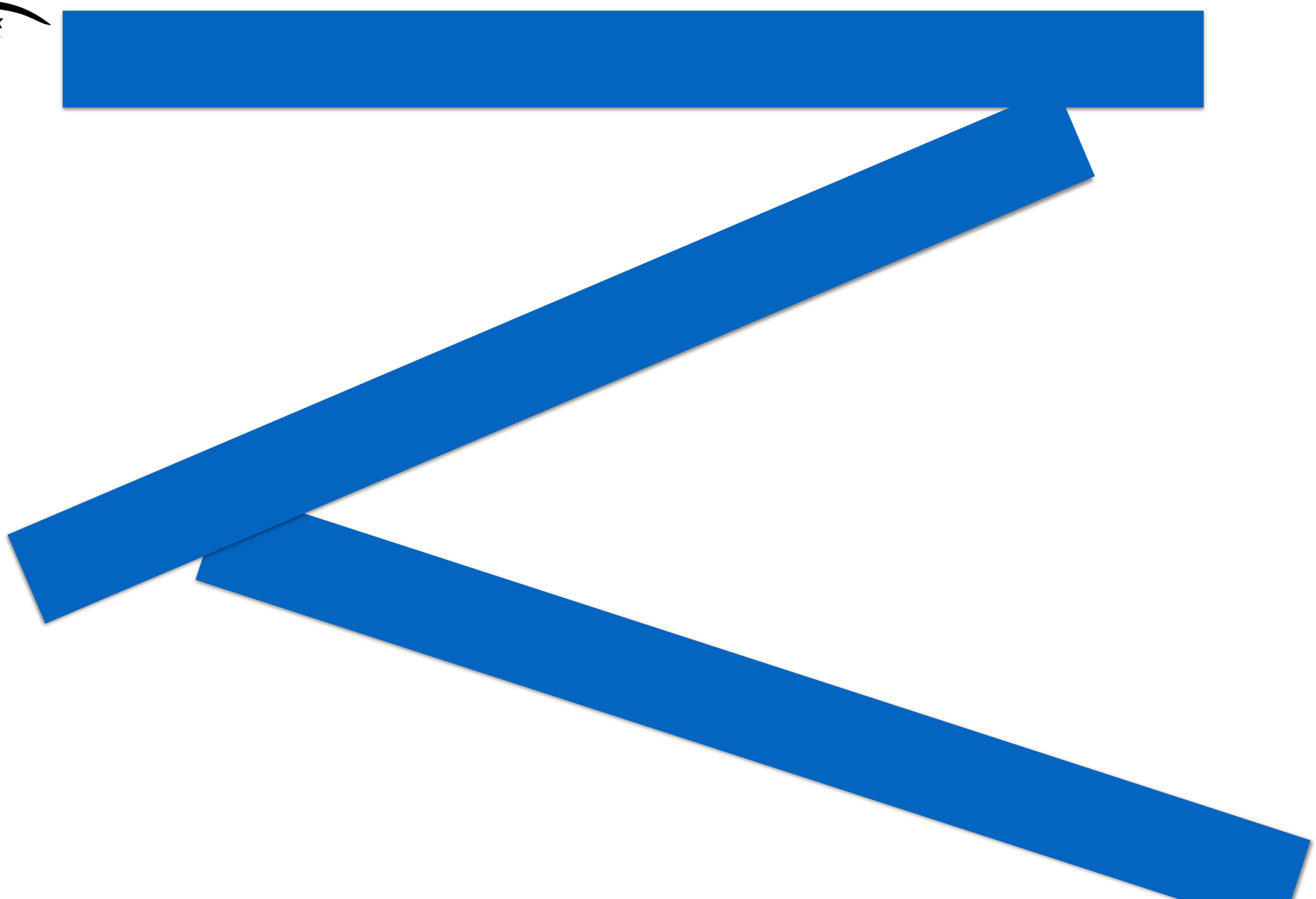

Not Related
Related
10
7.5
5
2.5
0
0
2.5
5
7.5
10

Visual Encodings: **Length**



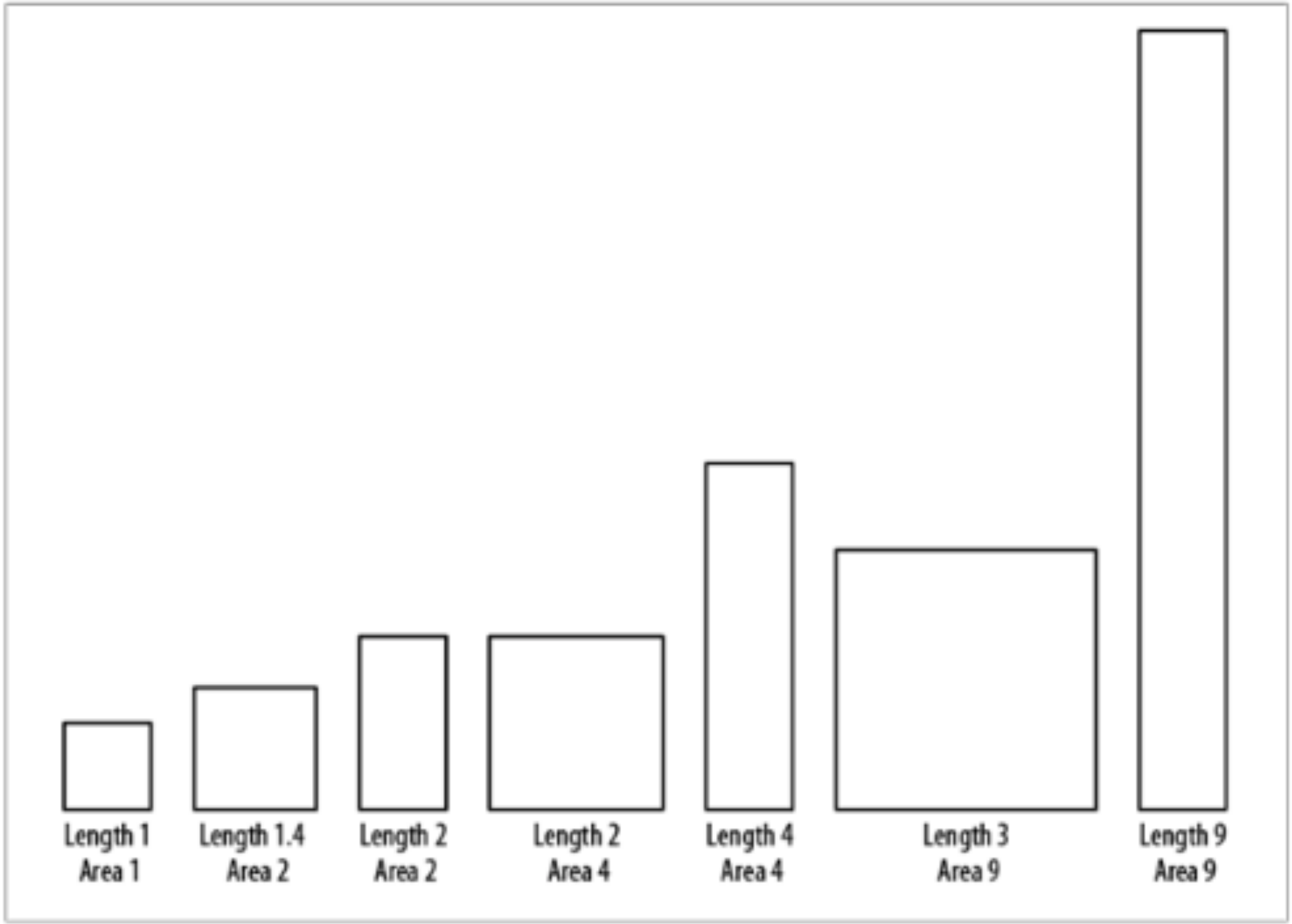
gtkcyber.com

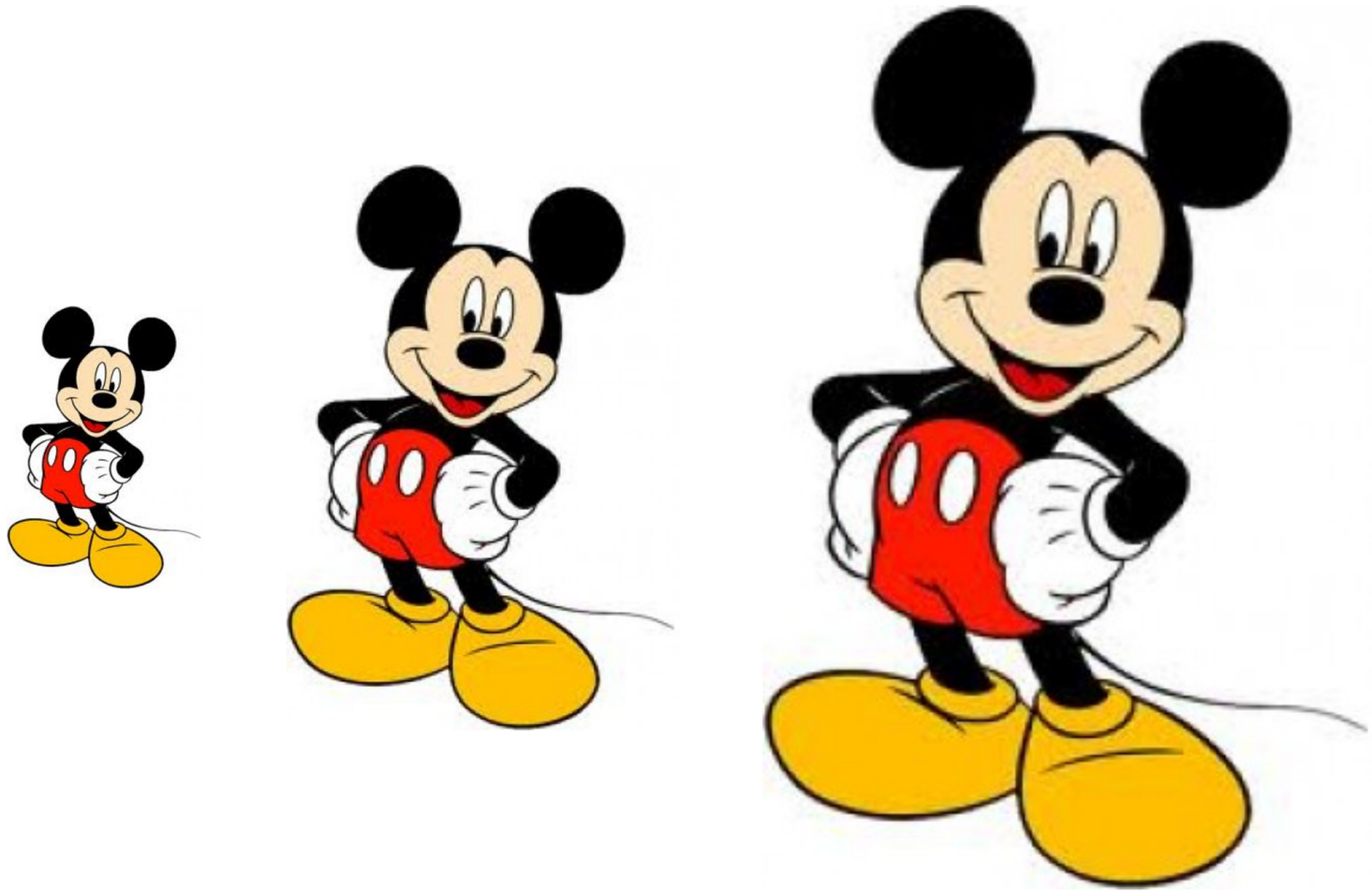

GTK CYBER
GET ON THE FAST TRACK



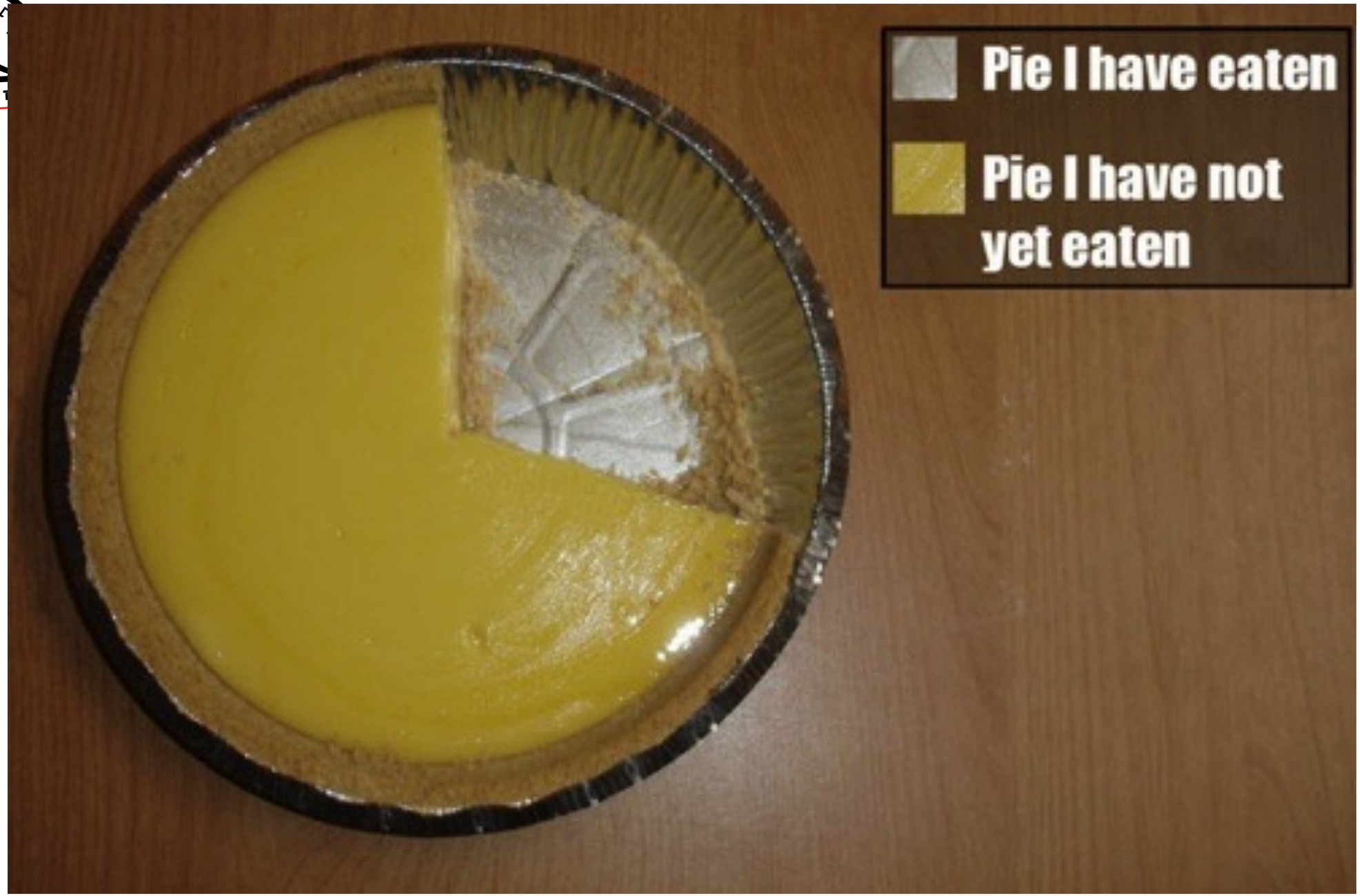
gtkcyber.com

Visual Encodings:
Size


Length 1
Area 1
Length 1.4
Area 2
Length 2
Area 2
Length 2
Area 4
Length 4
Area 4
Length 3
Area 9
Length 9
Area 9

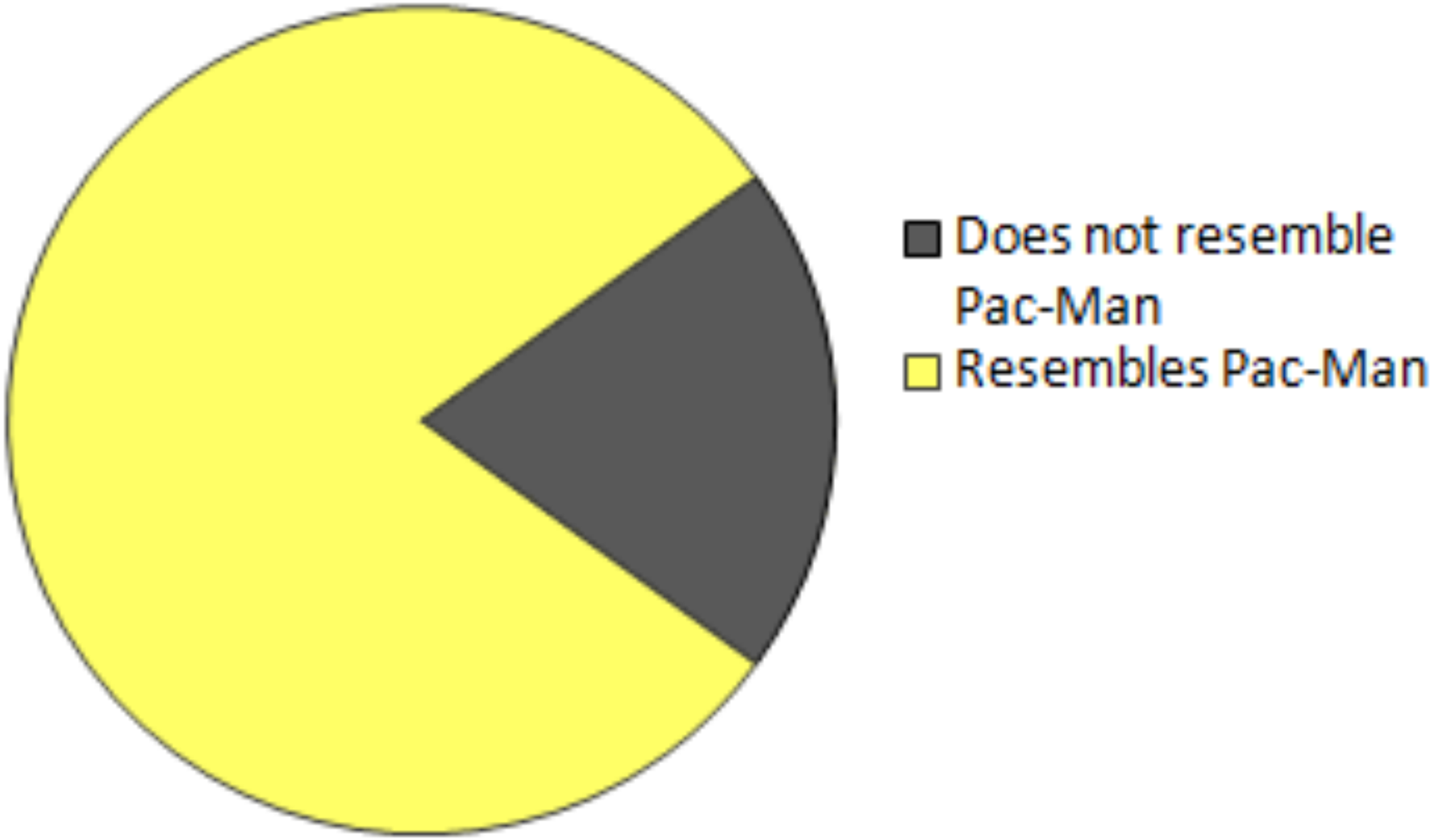


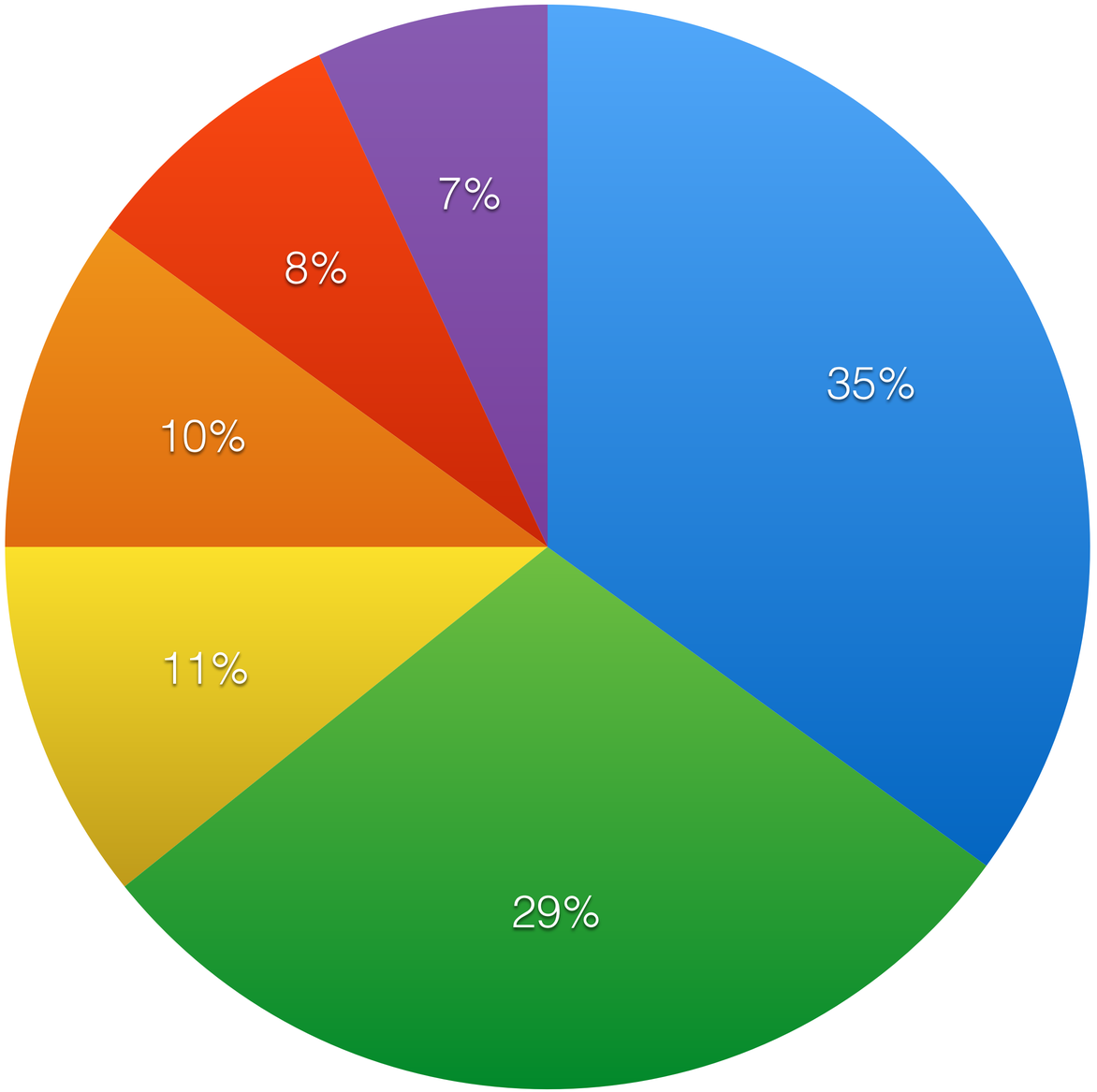
A brief interlude: Why never to use a Pie Chart


Pie I have eaten
Pie I have not yet eaten

GTK CYBER
GET ON THE FAST TRACK

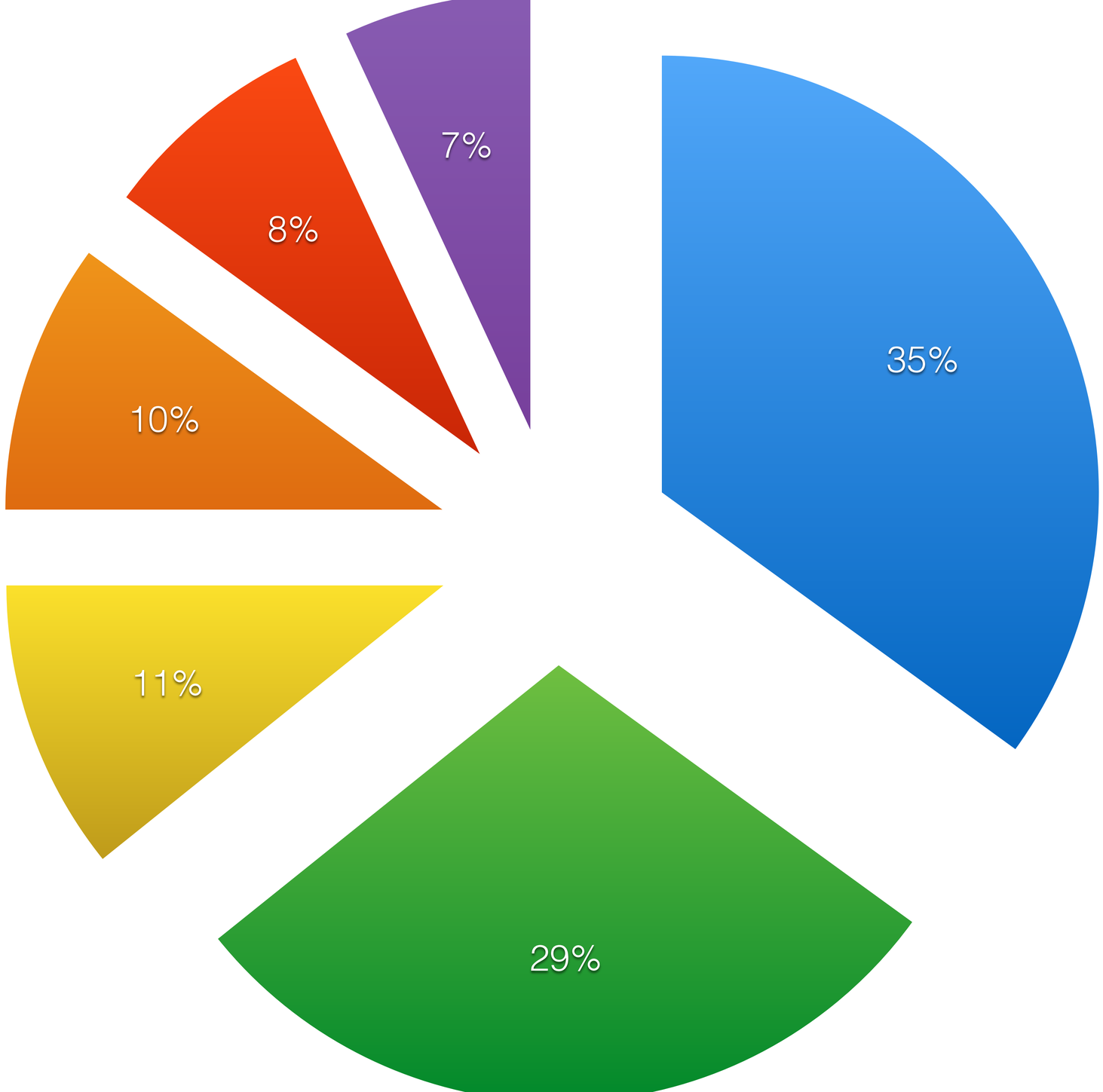
Percentage of Chart Which Resembles Pac-Man



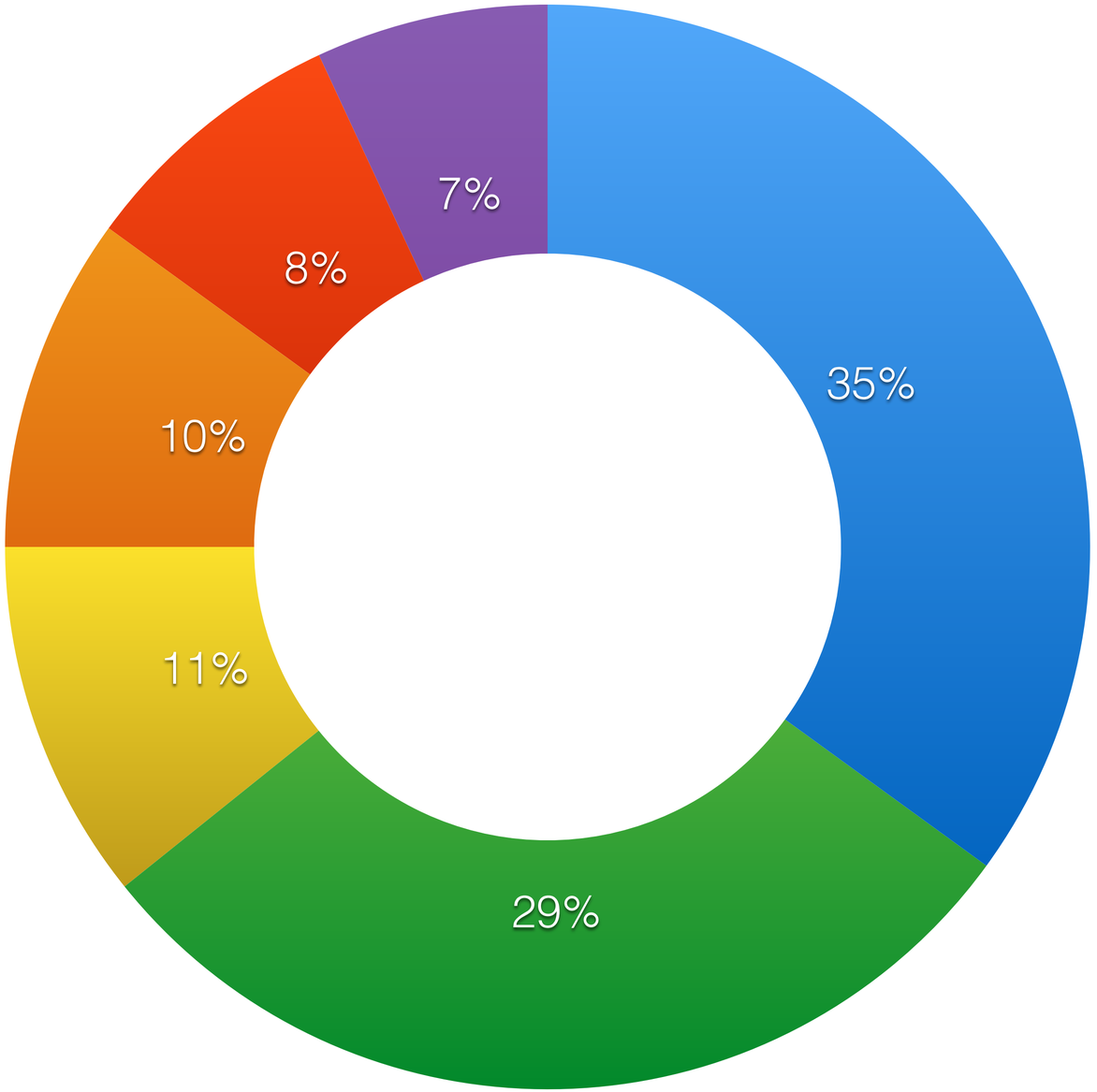

7%
8%
35%
10%
11%
29%

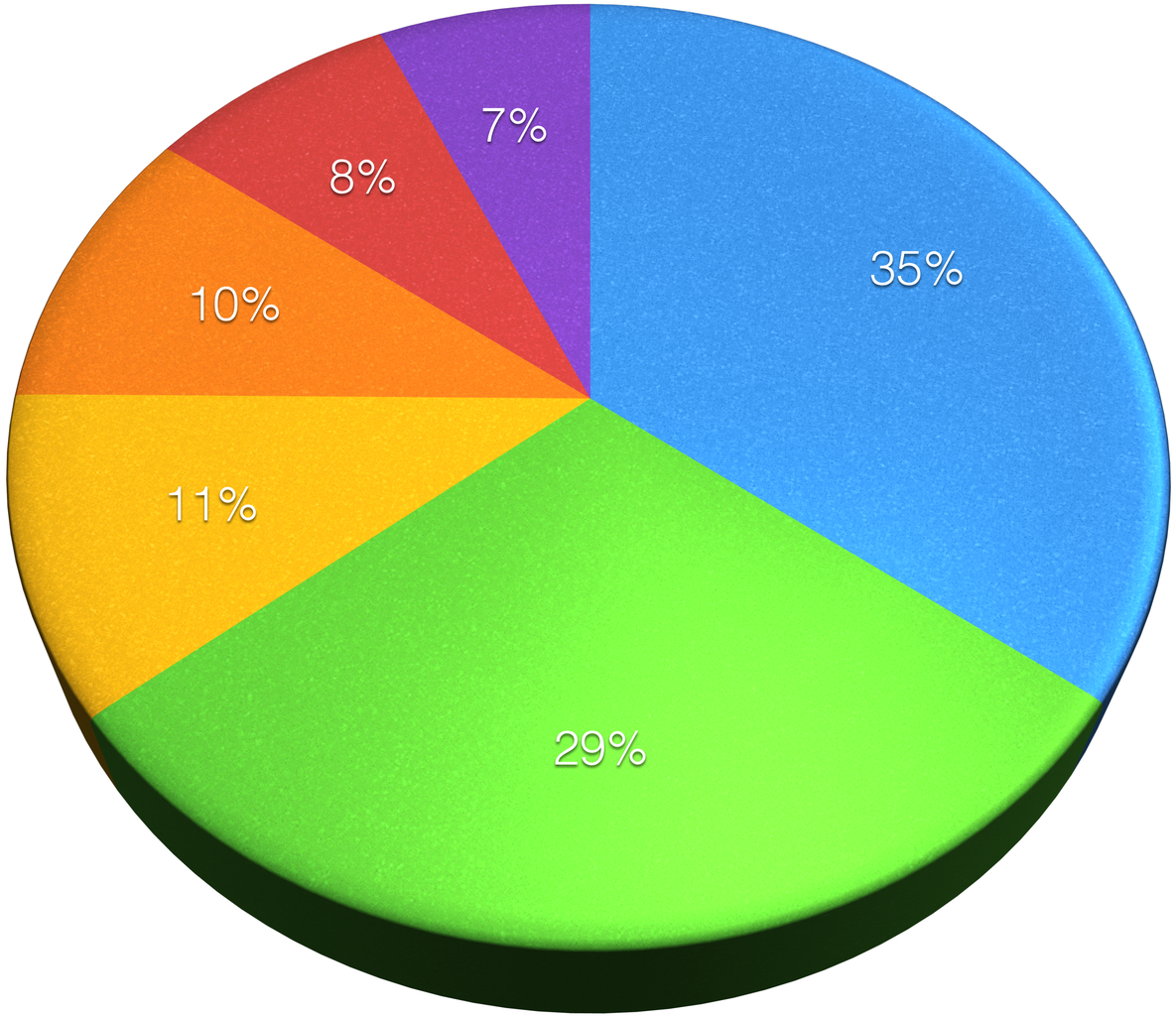
GTK CYBER

GET ON THE FAST TRACK



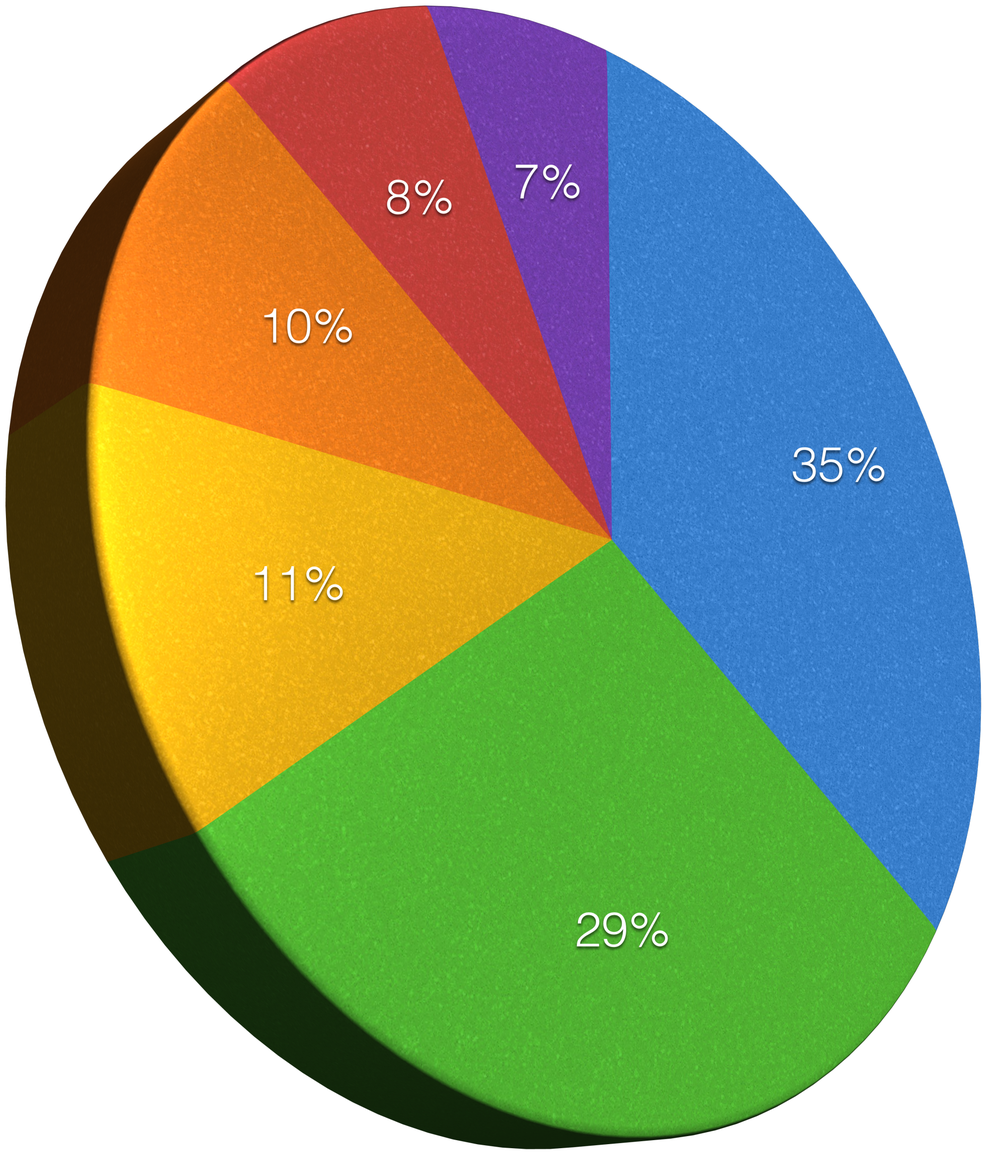
gtkcyber.com

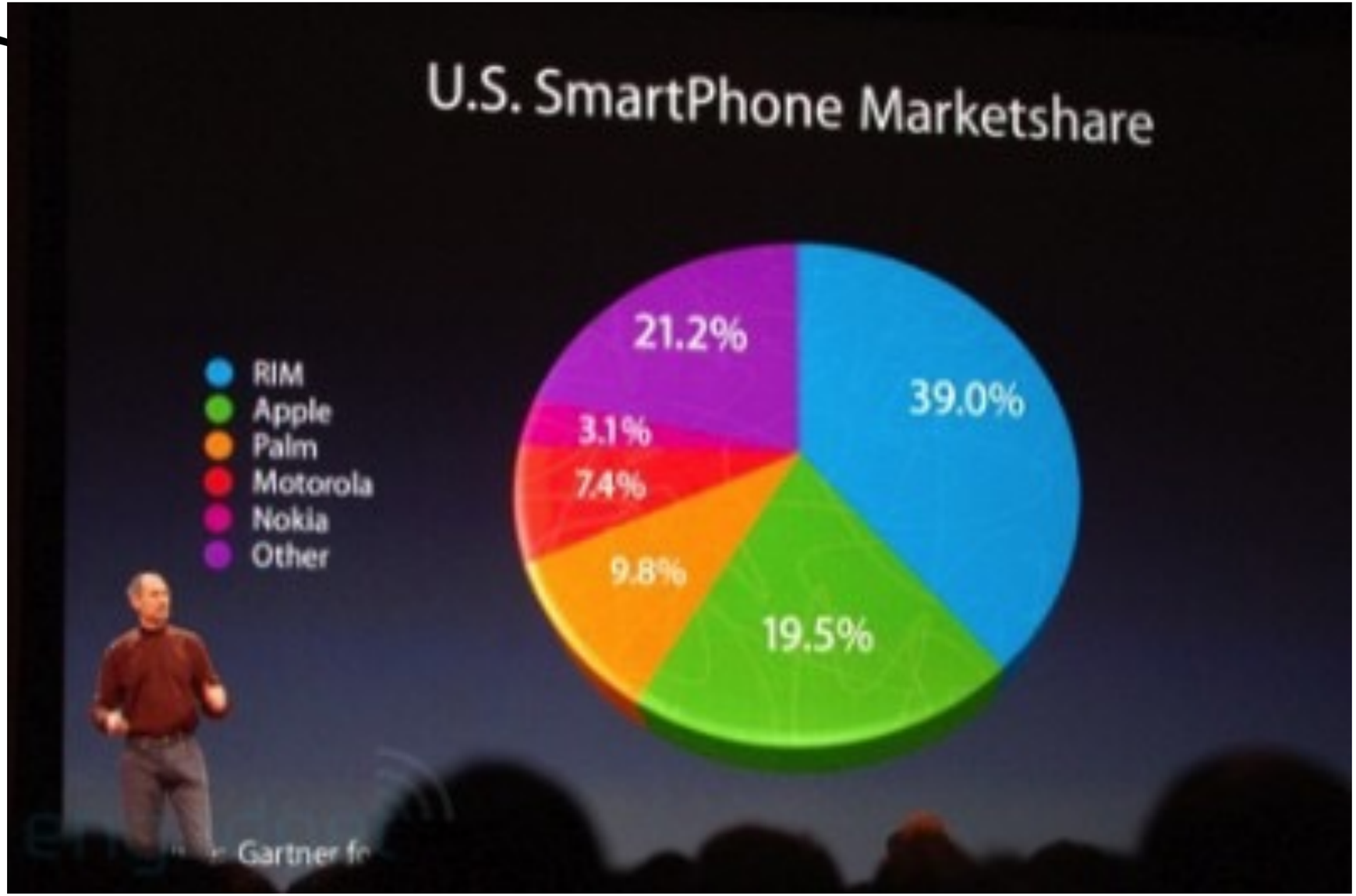

35%
29%
11%
10%
8%
7%


7%
8%
10%
11%
29%
35%

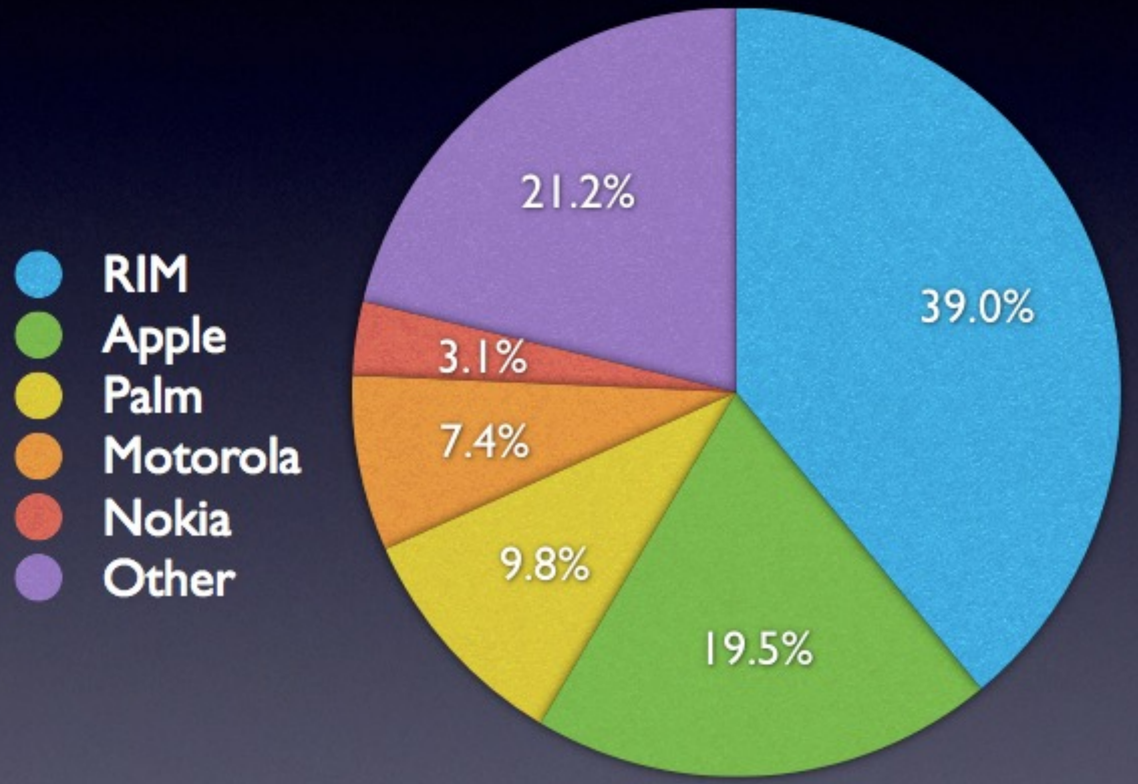
GTK CYBER

GET ON THE FAST TRACK

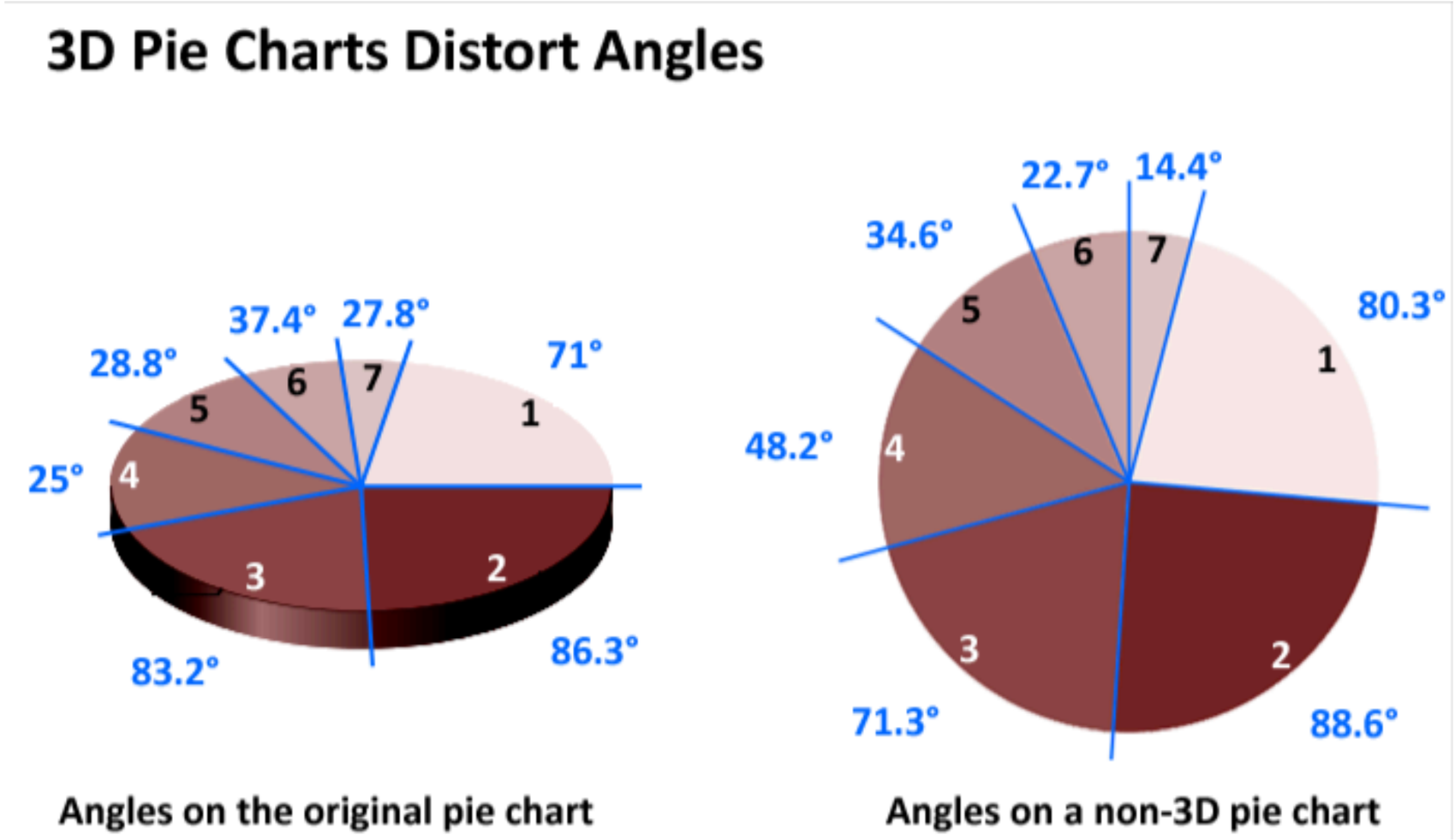



U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
39.0%
19.5%
9.8%
7.4%
3.1%
21.2%
Gartner

U.S. SmartPhone Marketshare



3D Pie Charts are BAD!



h8p://themendozaline.org/post/113828947026/3d-pie-charts-are-lie-charts

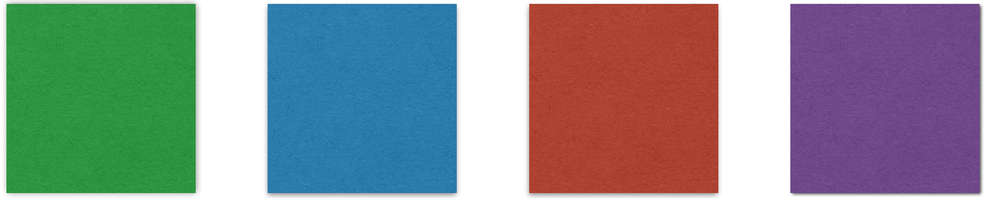
Visual Encodings: **Color**

Visual Encodings: **Color**

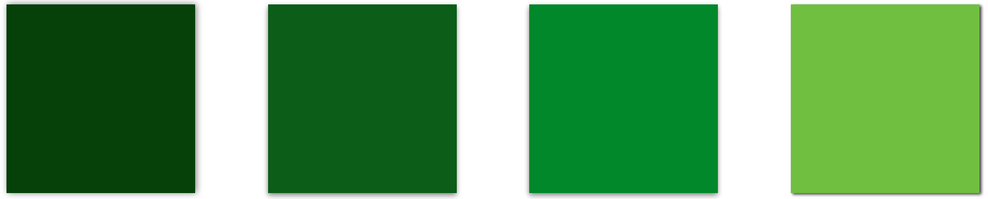
- **Hue** should be used to encode categorical data
- **Saturation** should be used to encode intensity or a continuous value

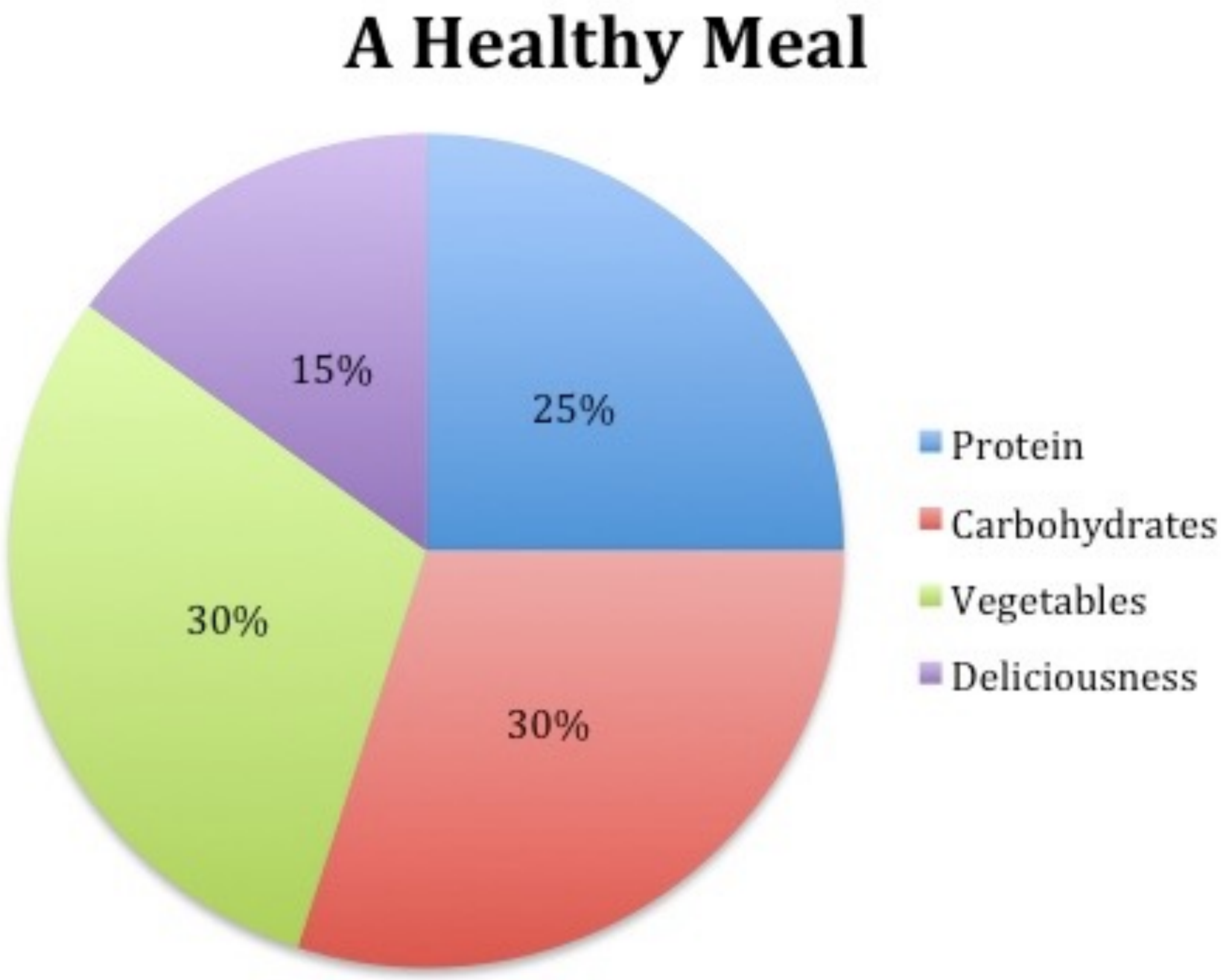
GTK CYBER
GET ON THE FAST TRACK

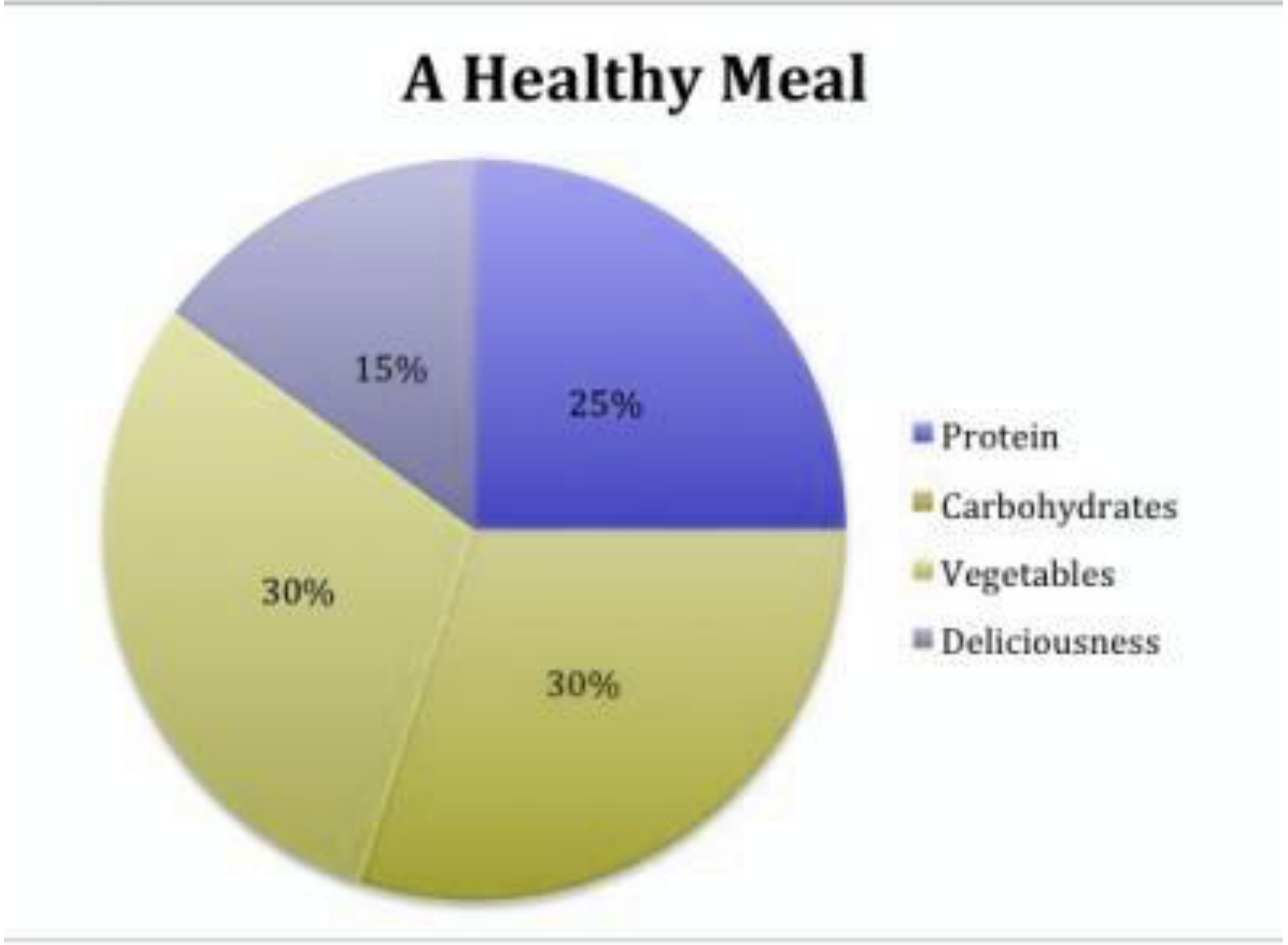
Hue



Saturation




A Healthy Meal
25%
30%
30%
15%
Protein
Carbohydrates
Vegetables
Deliciousness


A Healthy Meal
15%
25%
30%
30%
Protein
Carbohydrates
Vegetables
Deliciousness


elastica
Dashboard
Users
Total (192)
28
High Risk
0
Med Risk
4
Blocked
Policies
Total (231)
3
Blocking
142
Alerting
Policy Alerts
Alerting
443
Blocked
13
Rest
0
Threat Alerts
High Risk
2
Med Risk
2
Low Risk
17
Audited Services
by Users
High Risk (736)
Medium Risk (3k)
Low Risk (3k)
3k
494.3 GB
963k
243

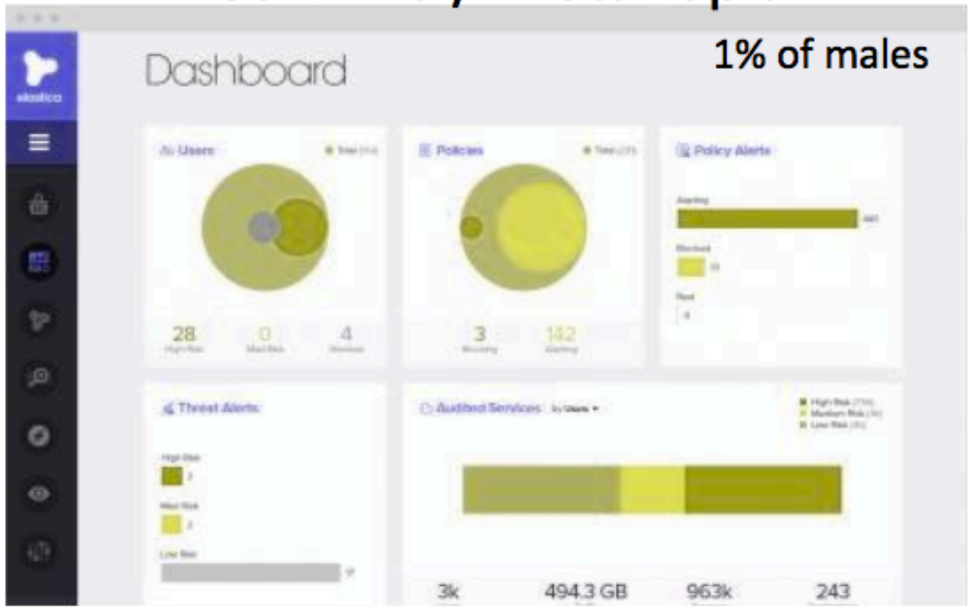
Original

Green-Blind / Deuteranopia

8% of male
0.5% female

Red-Blind / Protanopia

1% of males

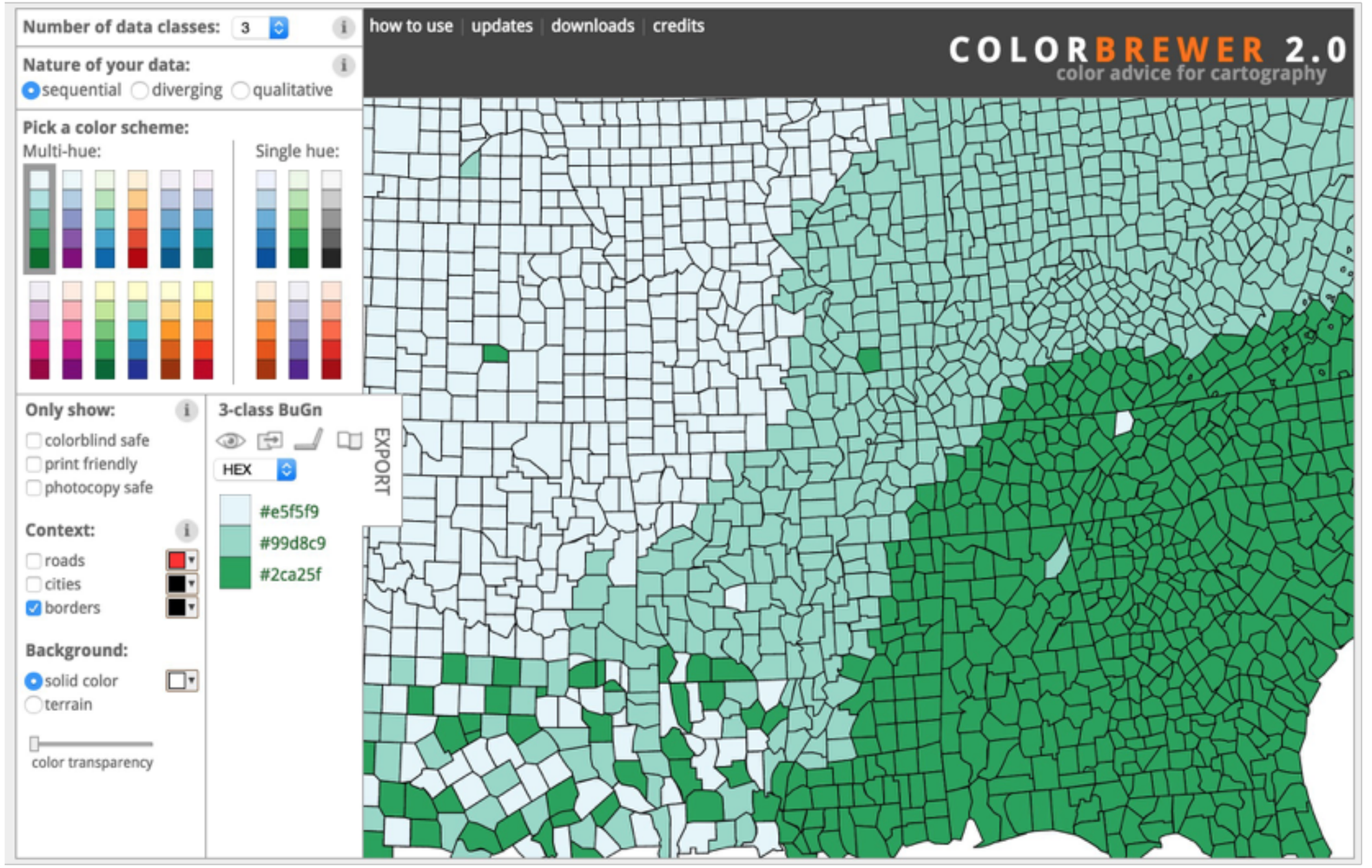


Blue-Blind / Tritanopia

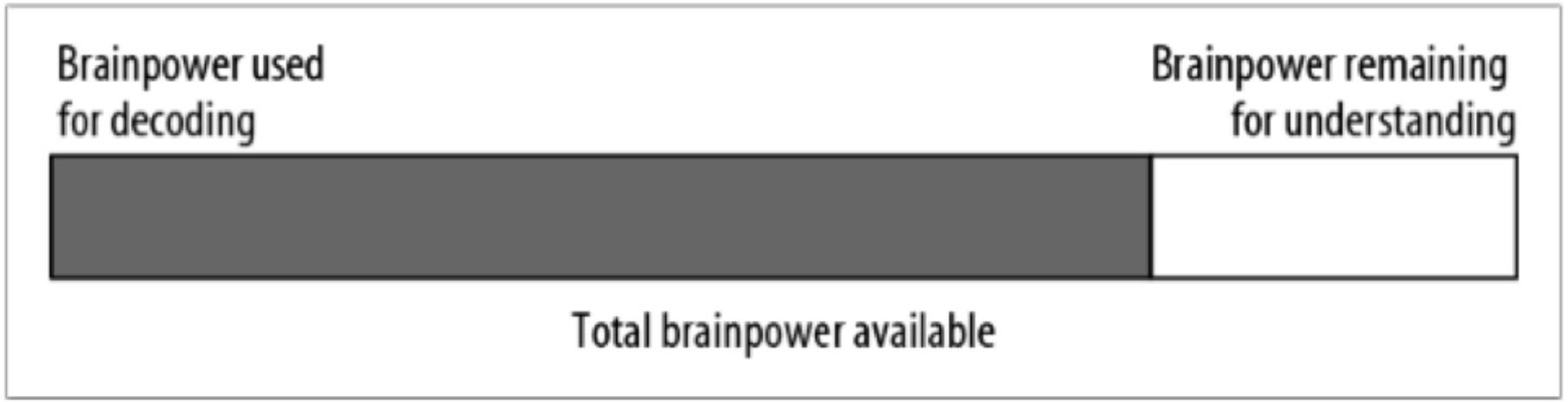
rare

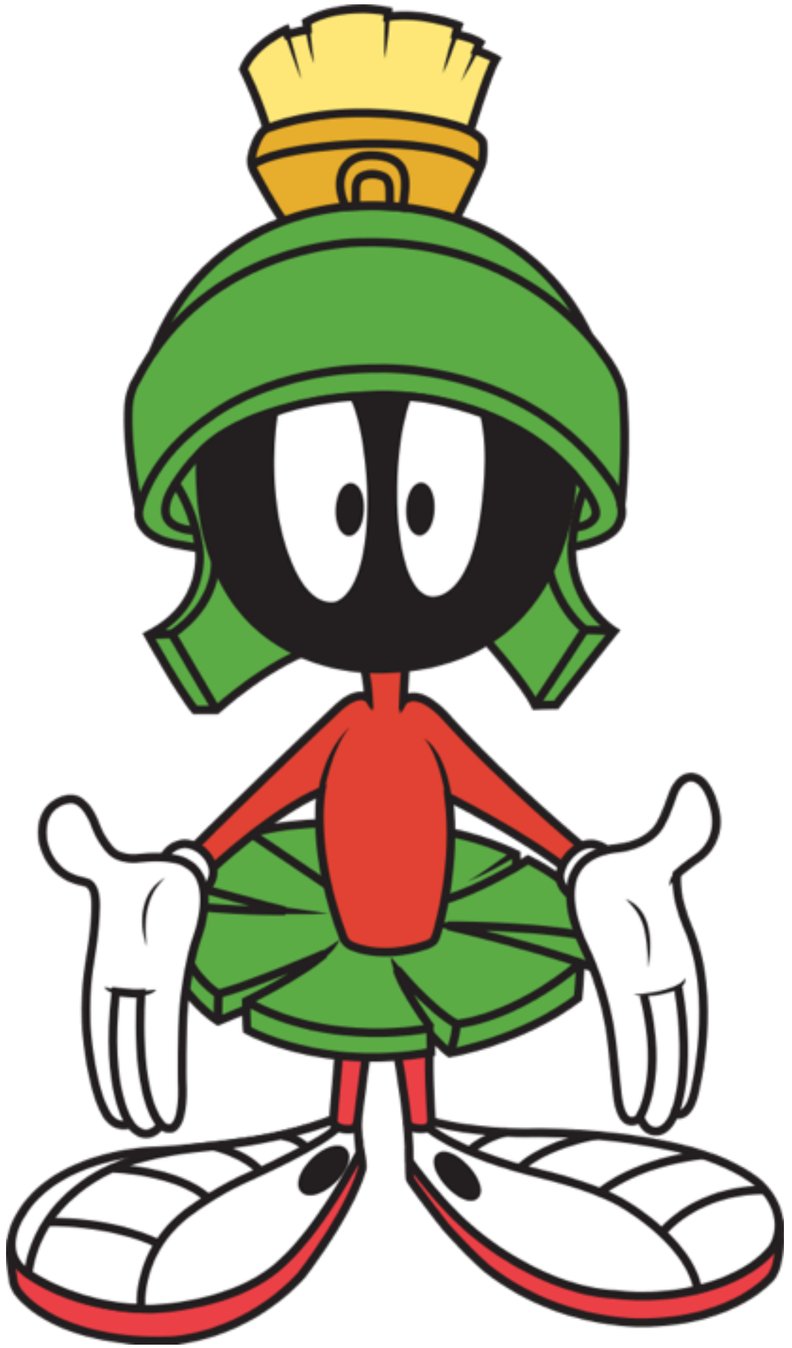
http://www.color-blindness.com/coblis-color-blindness-simulator/

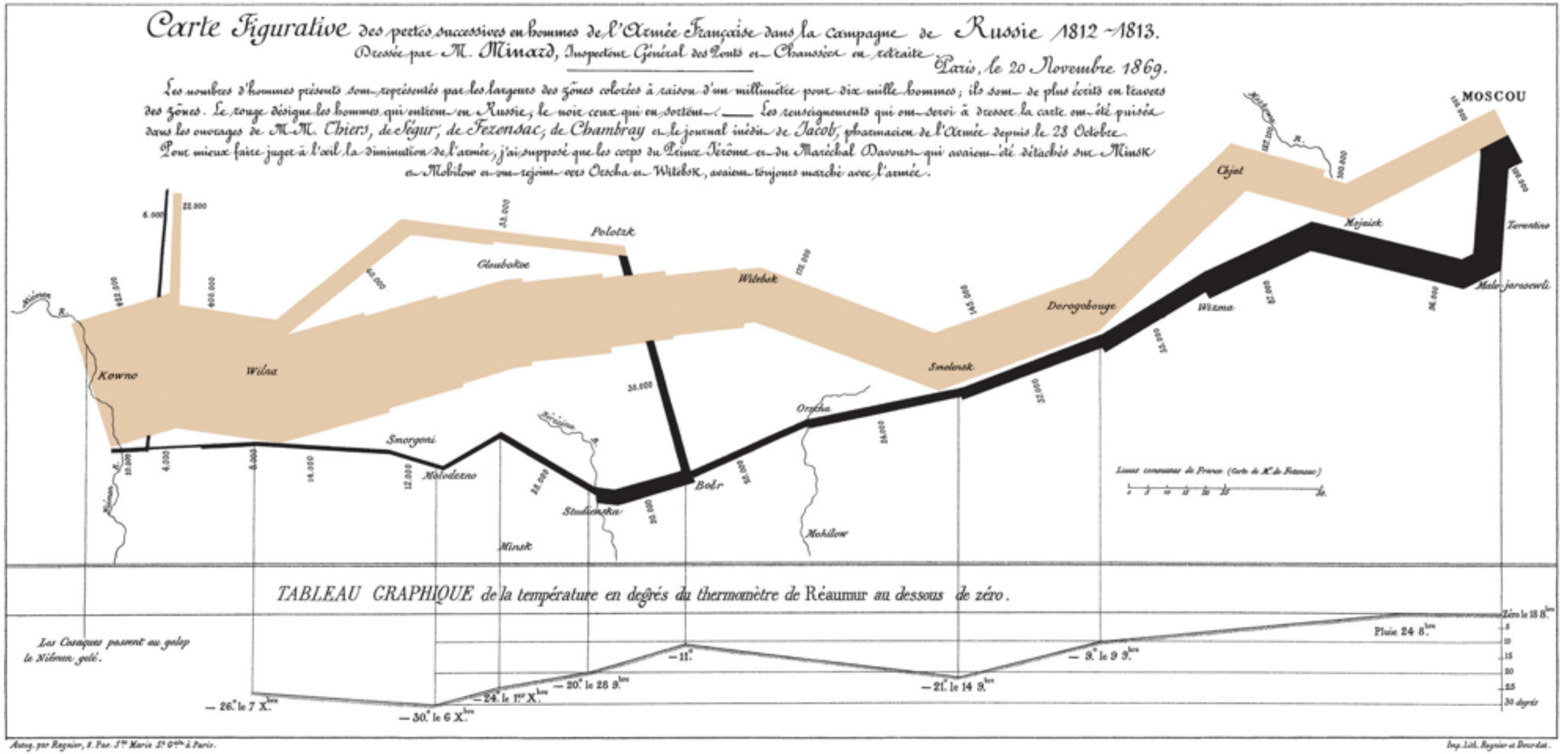
Color



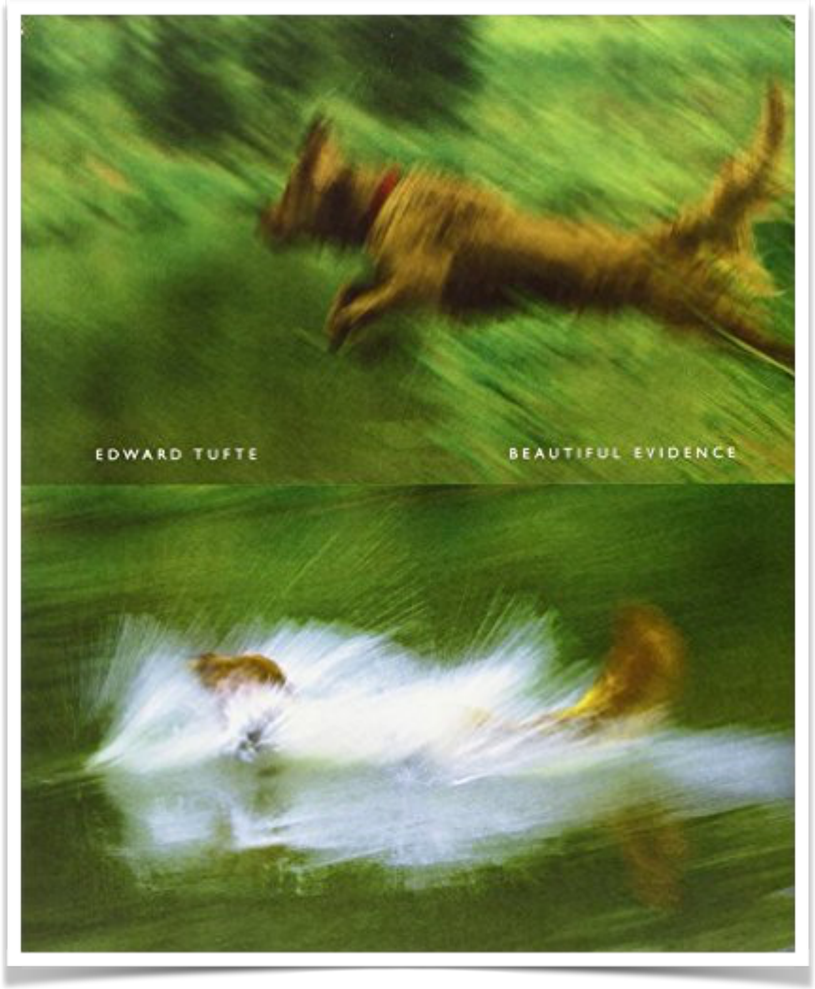
colorbrewer2.org


Brainpower used
for decoding
Brainpower remaining
for understanding
Total brainpower available

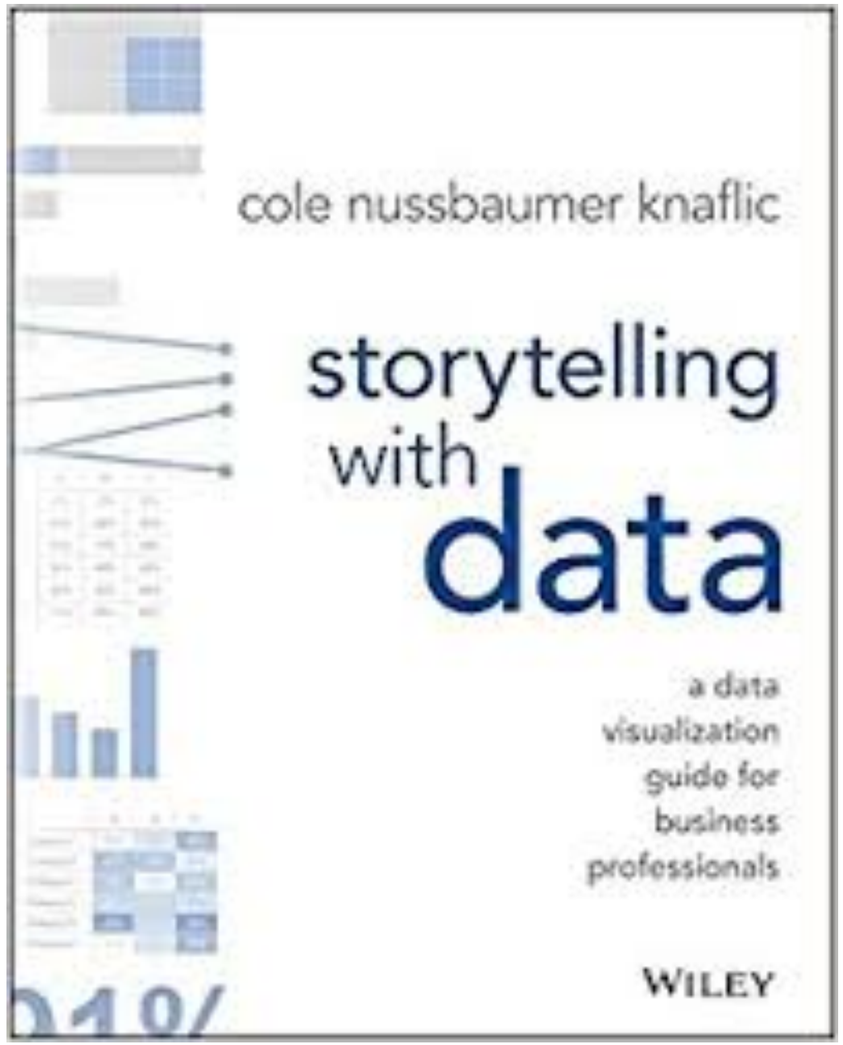




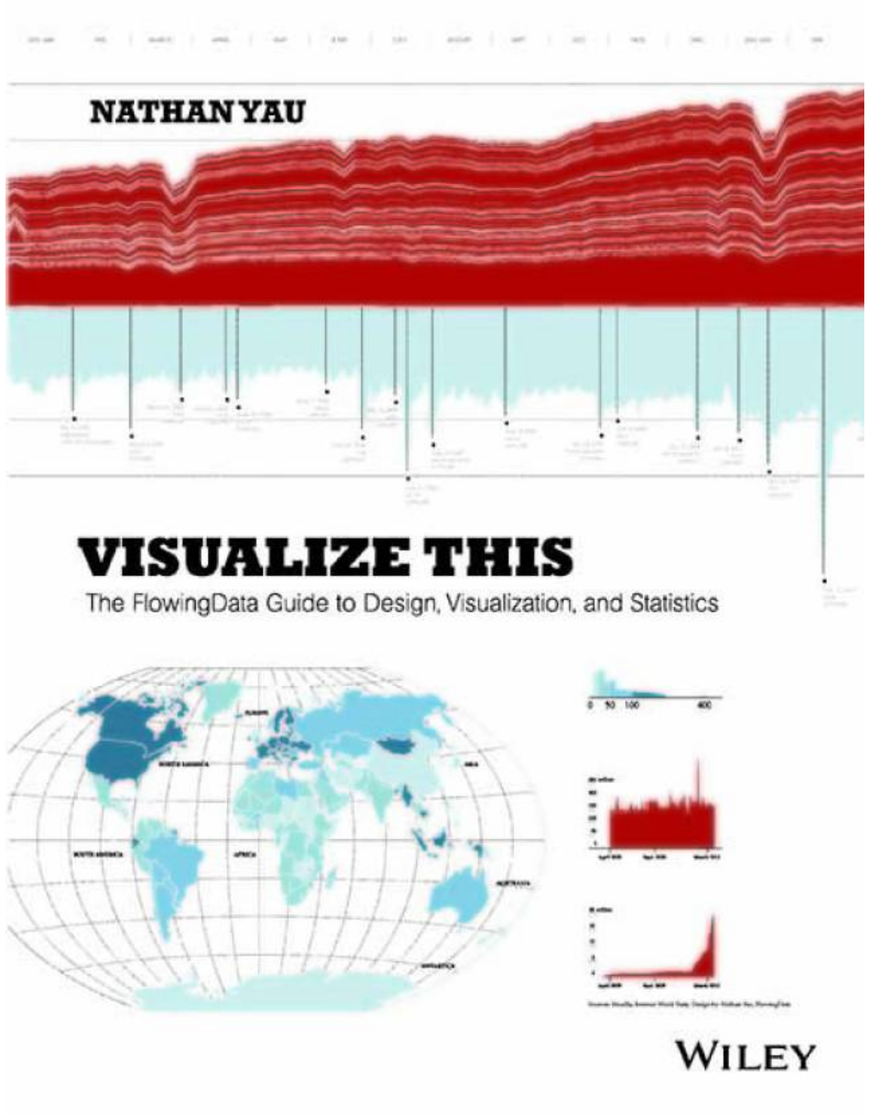
Recommended Reading



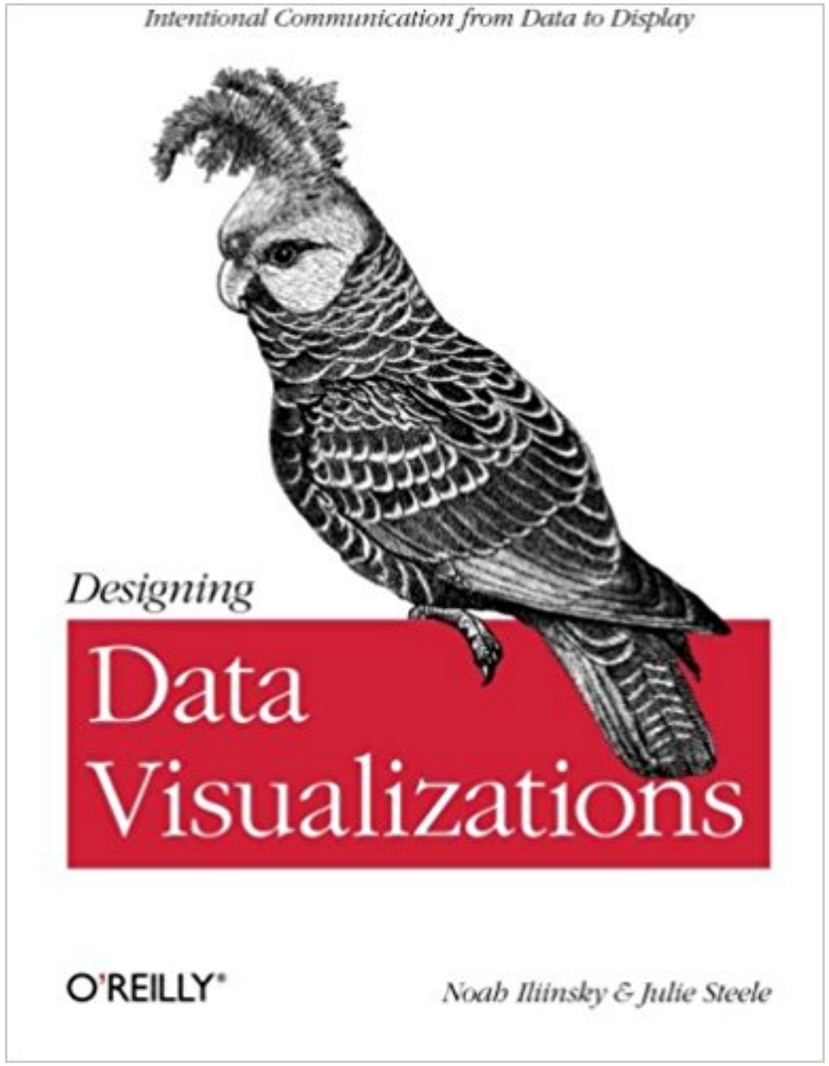
Recommended Reading



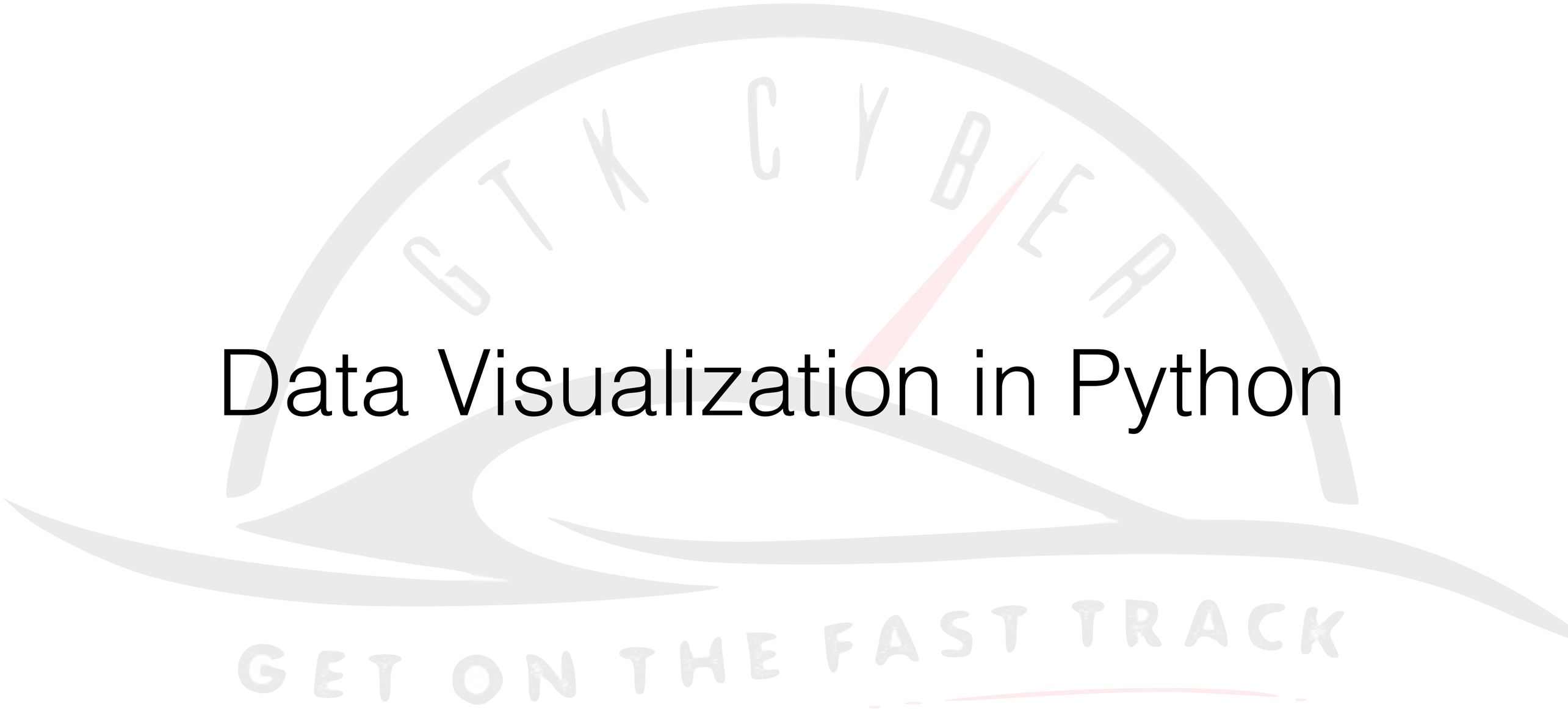
Recommended Reading

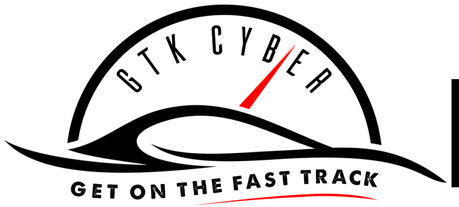


Recommended Reading



Data Visualization in Python





Python Visualization Libraries

- **Matplotlib**: “Lowest” level visualization library. Very powerful, but little abstraction
- **Pandas/Pyplot**: Easier, but requires transformation for complex visualizations
- **Seaborn:** Good for advanced statistical visualizations
- **Altair:** New library: declarative statistical visualization library
- **Plotly / Cufflinks:** Make it interactive!

Other Visualization Libraries

Panel

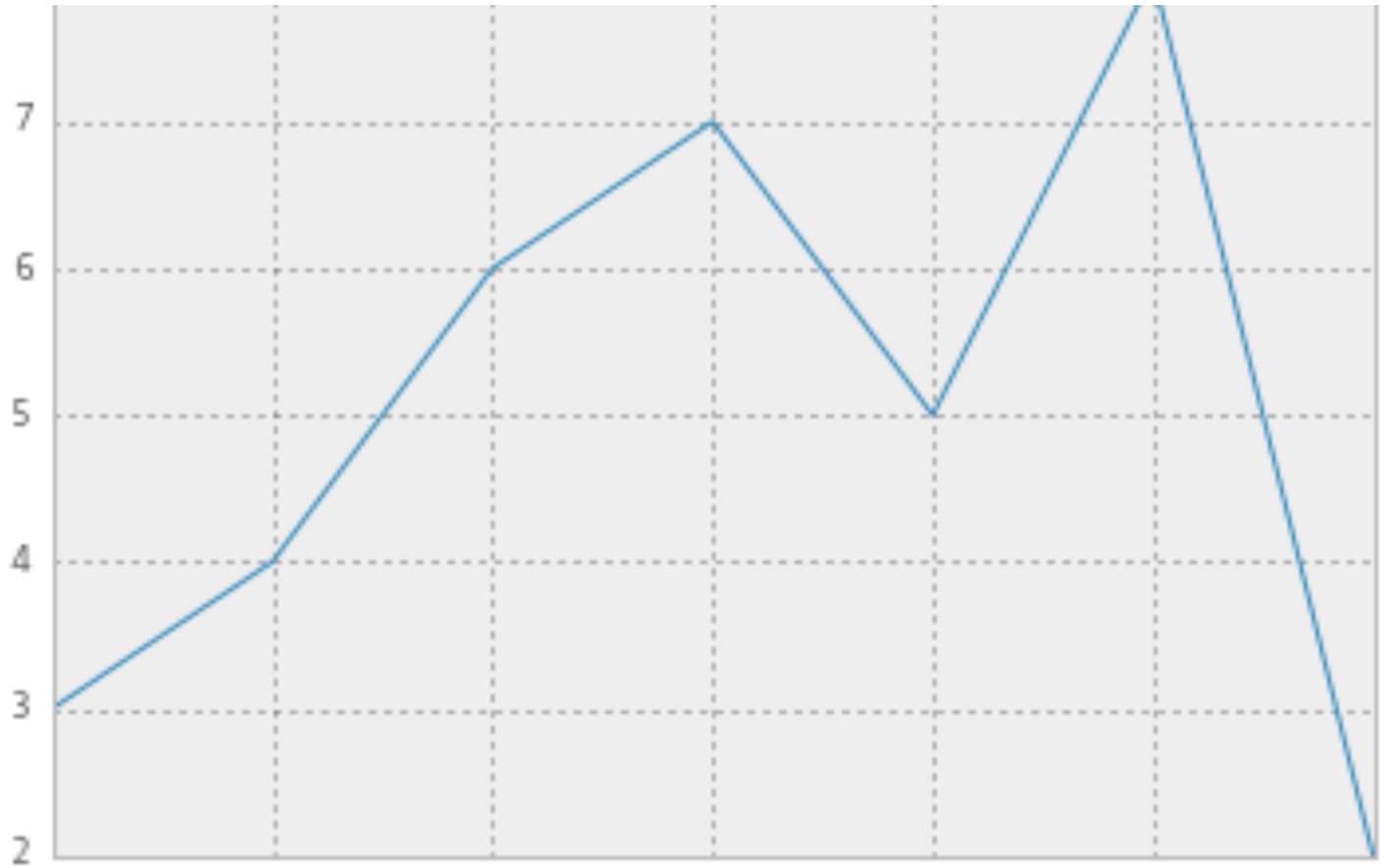
hvPlot

HoloViews

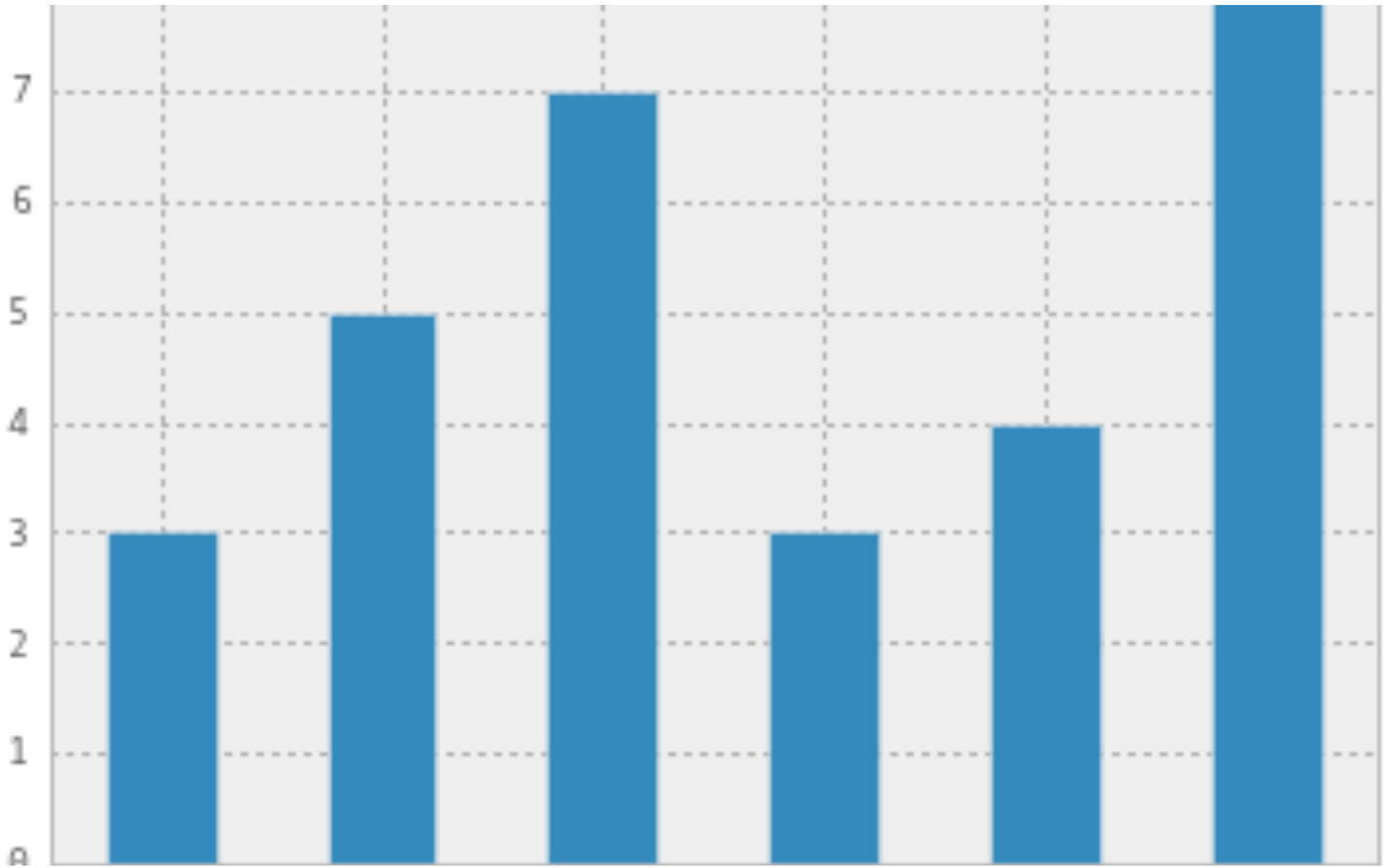
GeoViews

- **Panel**: Powerful library for creating apps and dashboards
- **hvPlot**: Quickly generate interactive plots from data
- **HoloViews:** Helps you make all data instantly visualizable
- **GeoViews:** HoloViews for geo-spatial data
- **ggplot2/PlotNine:** This is Python's implementation of R's ggplot2
- For more useful info about Python visualization, check out **pyviz.org**

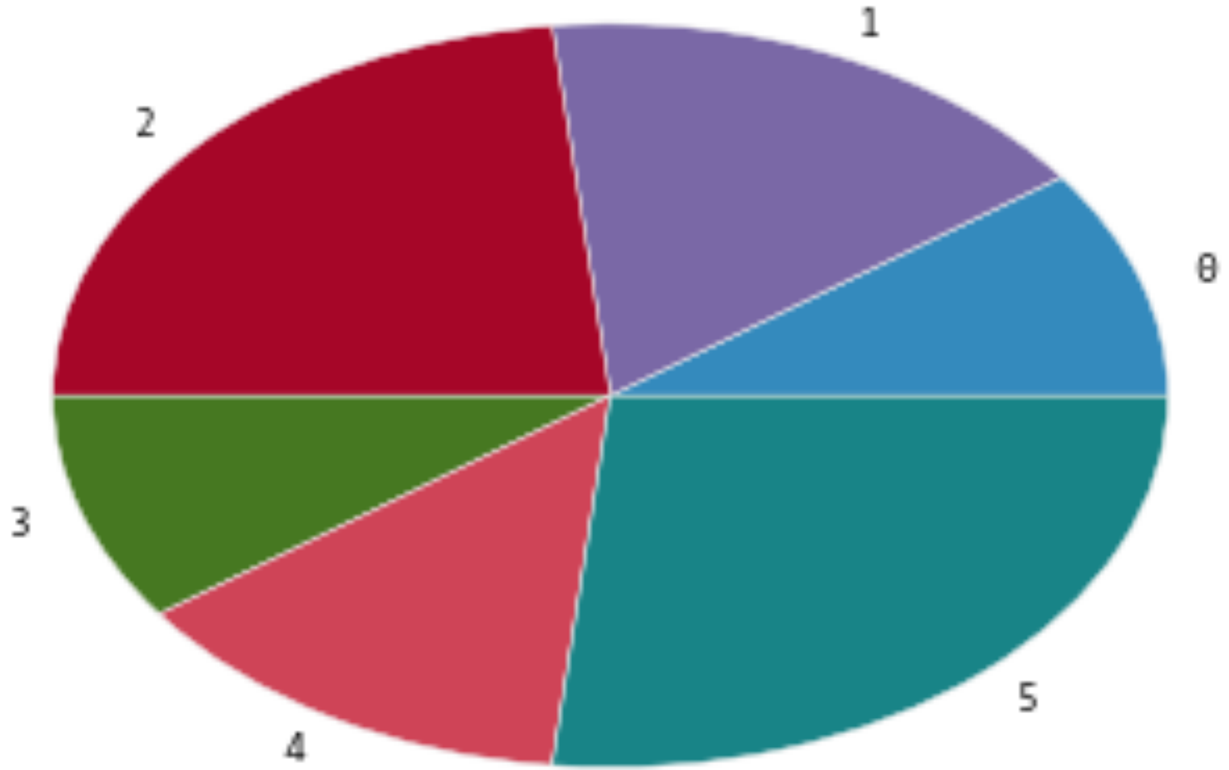
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
%matplotlib inline
```



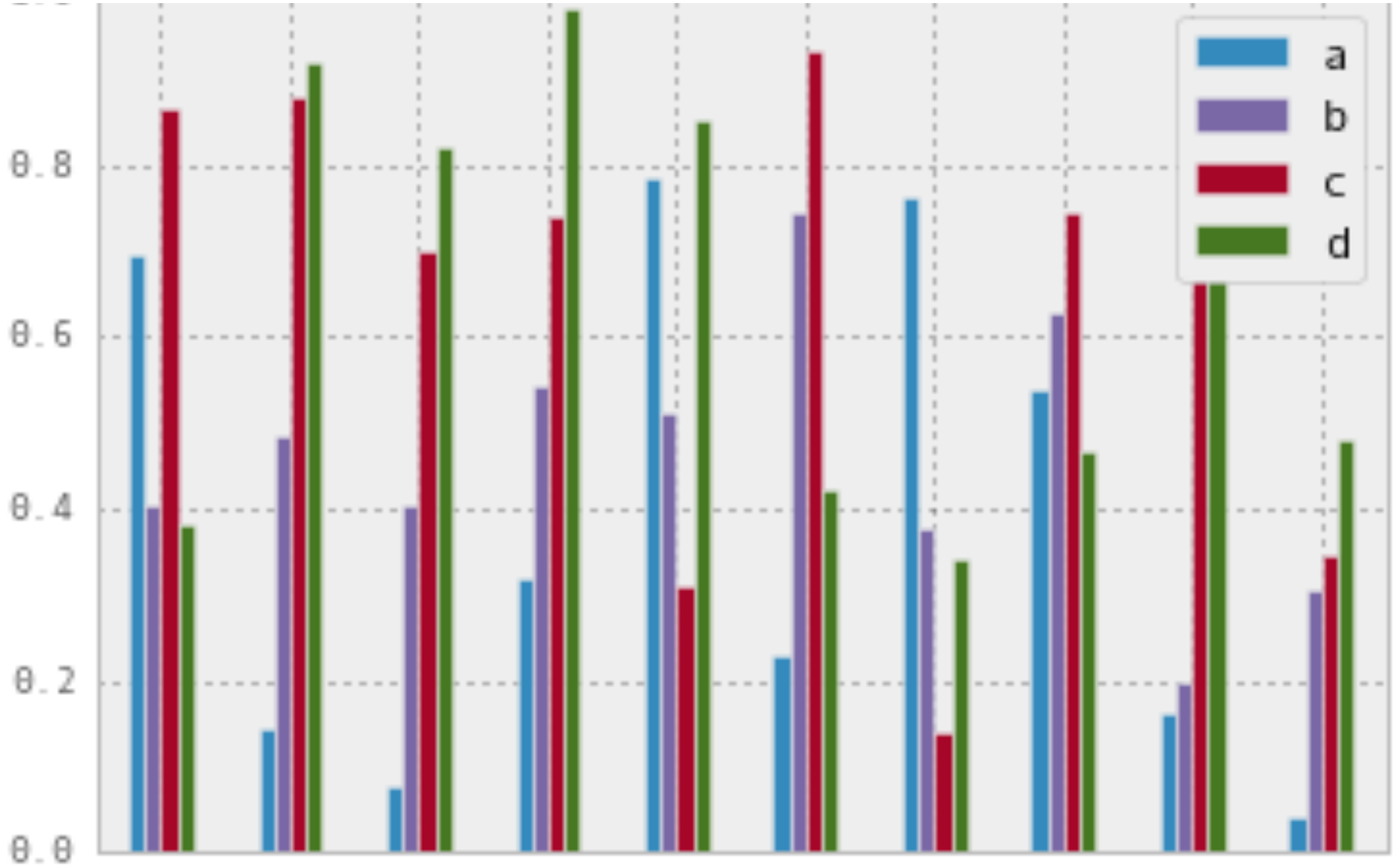

```
data = pd.Series( [3,4,6,7,5,8,2 ])
graph = data.plot()
```



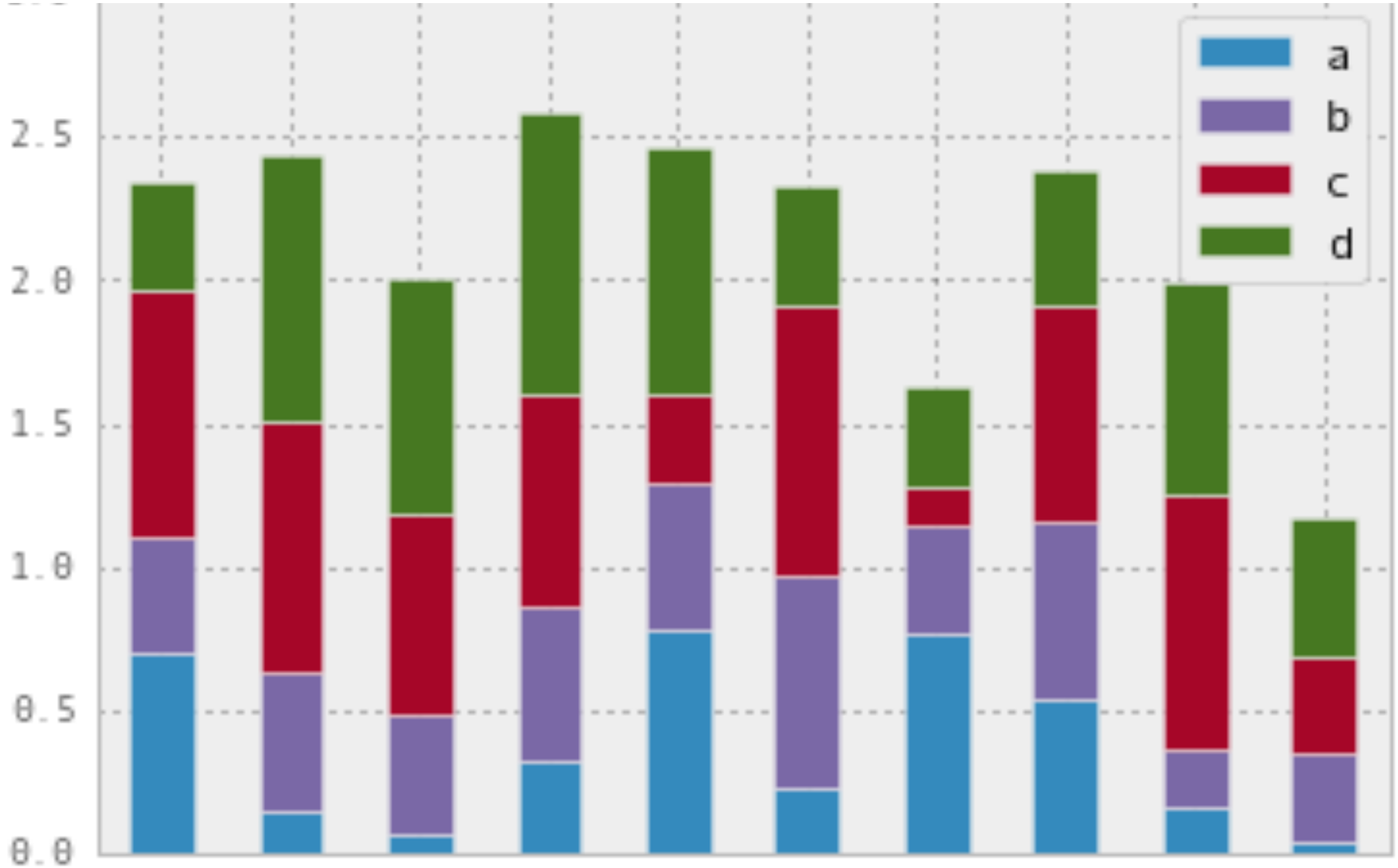

```
barchart = data.plot( kind="bar" )
```



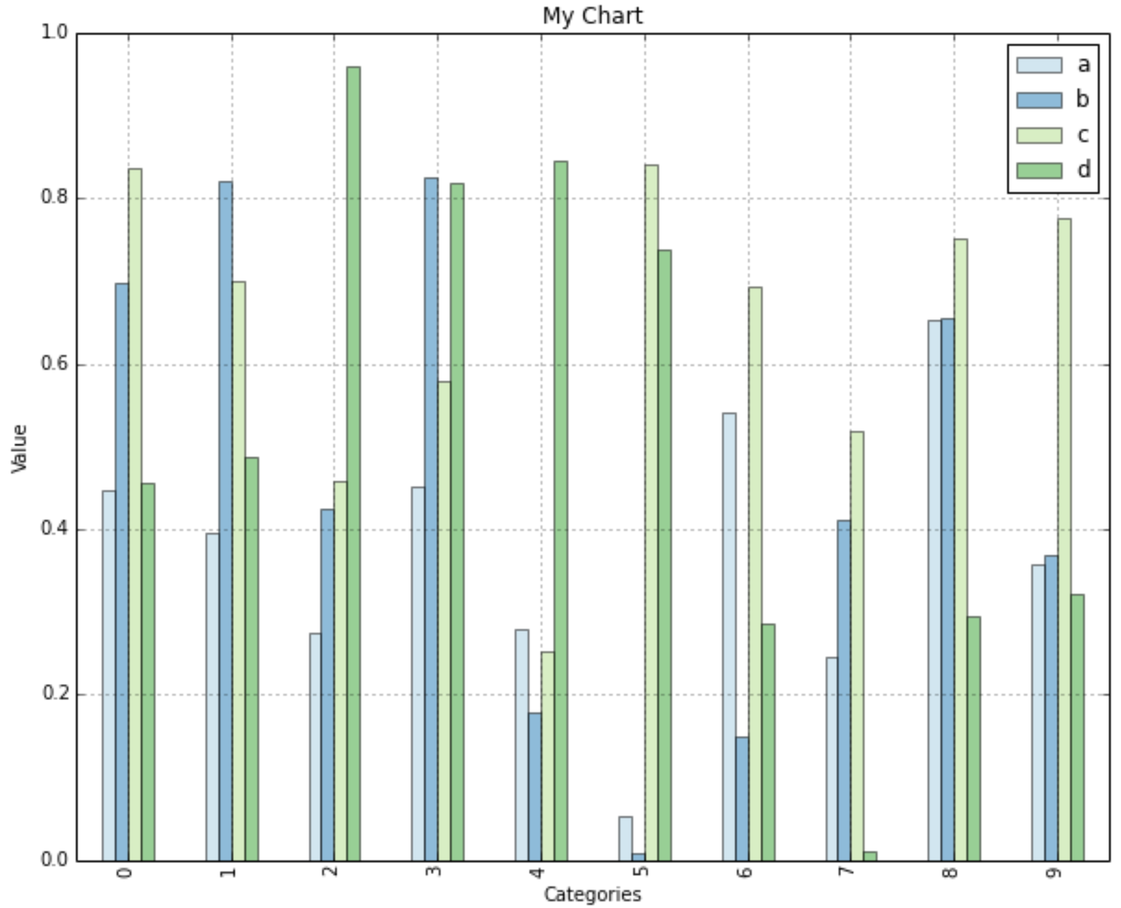

```
piechart = data.plot( kind="pie" )
```



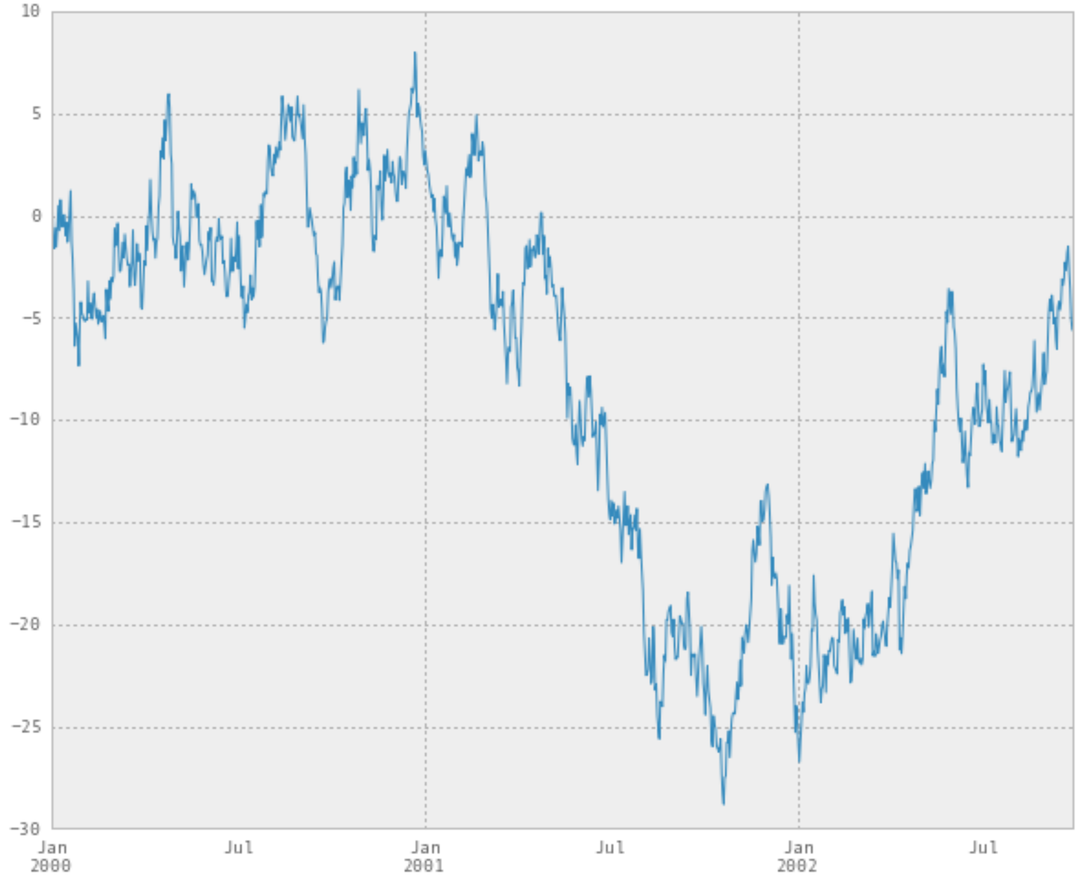

```
df2 = pd.DataFrame(np.random.rand(10, 4), \
    columns=['a','b','c','d'])

df2.plot( kind='bar' )
```



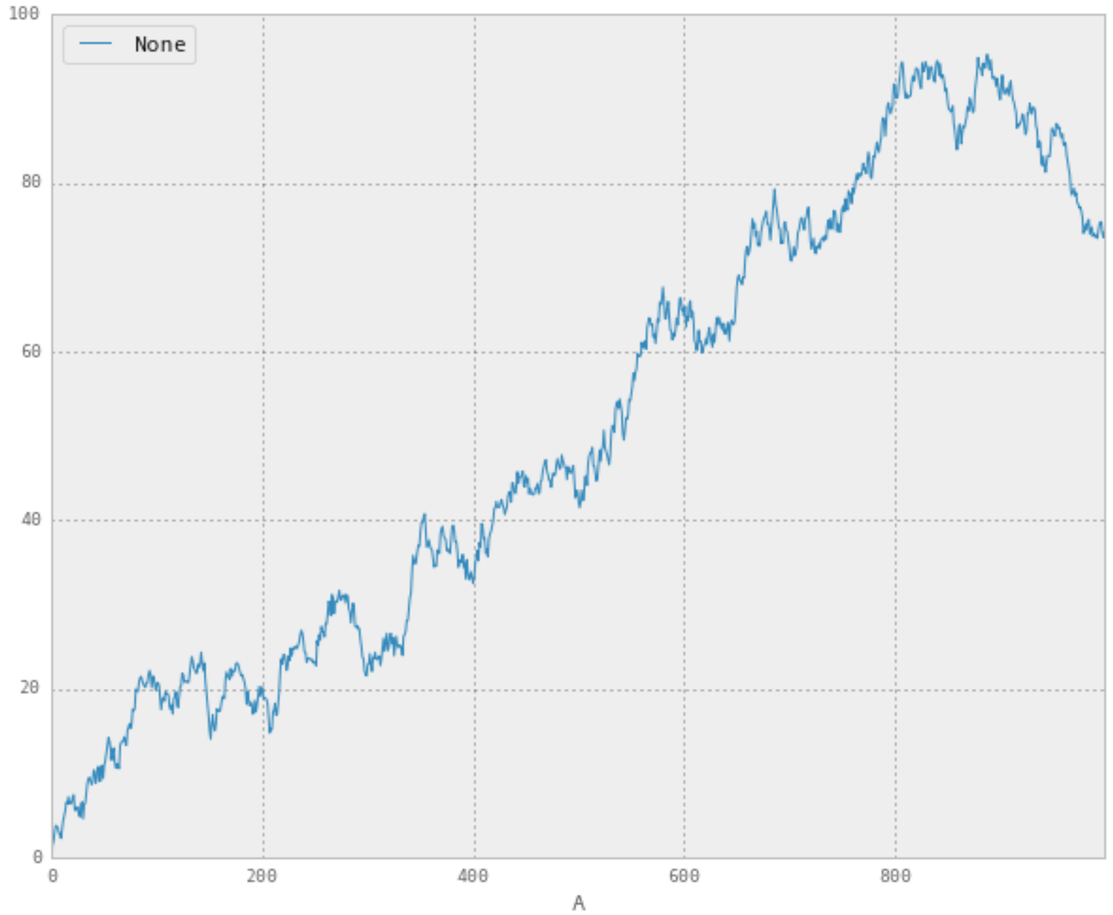

```
df2.plot( kind='bar', stacked=True )
```



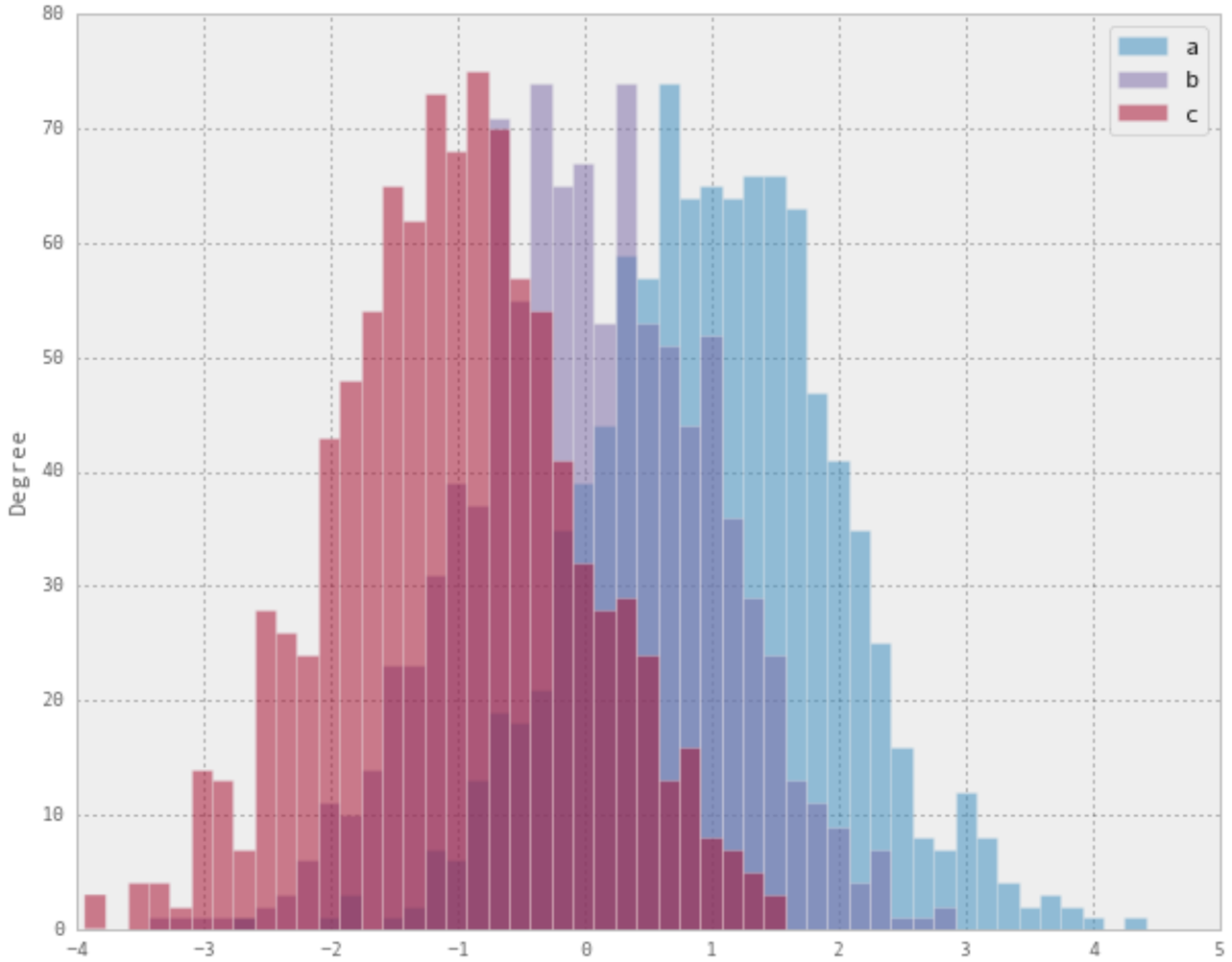

```
df2.plot( kind='bar',
          color=('#a6cee3','#1f78b4','#b2df8a','#33a02c'),
          alpha=0.5,
          width=0.5,
          figsize=(10,8))
plt.title( "My Chart" )
plt.xlabel( "Categories" )
plt.ylabel( "Value" )
```




```
ts = pd.Series(np.random.randn( 1000),
index=pd.date_range('1/1/2000', periods=1000))
ts = ts.cumsum()
timeseriesChart = ts.plot( figsize=(10, 8) )
```

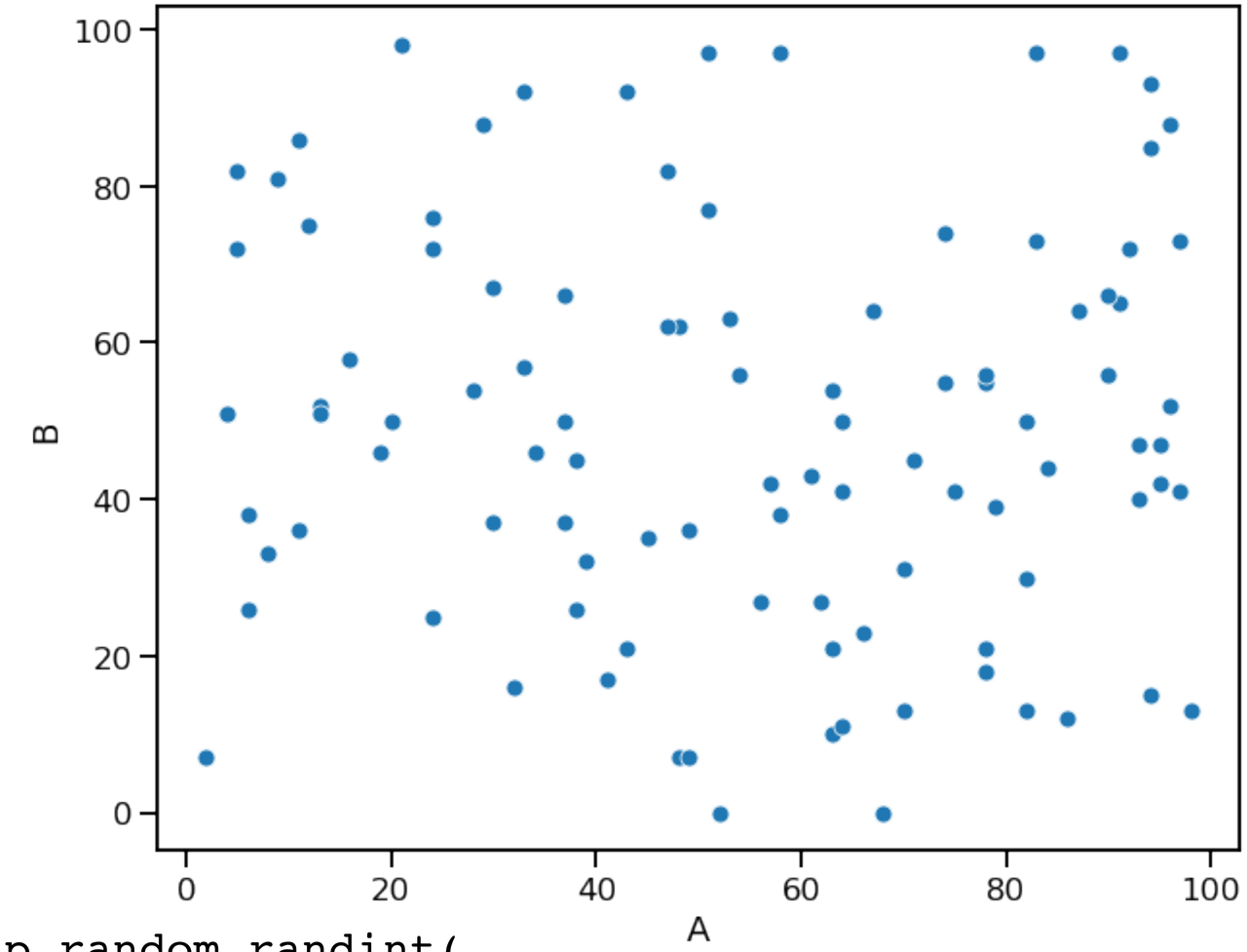



```
df3 = pd.DataFrame(np.random.randn(1000, 2),
     columns=['B', 'C']).cumsum()
df3['A'] = pd.Series(list(range(len(df3))))

df3.plot( x='A', y='B' )
```

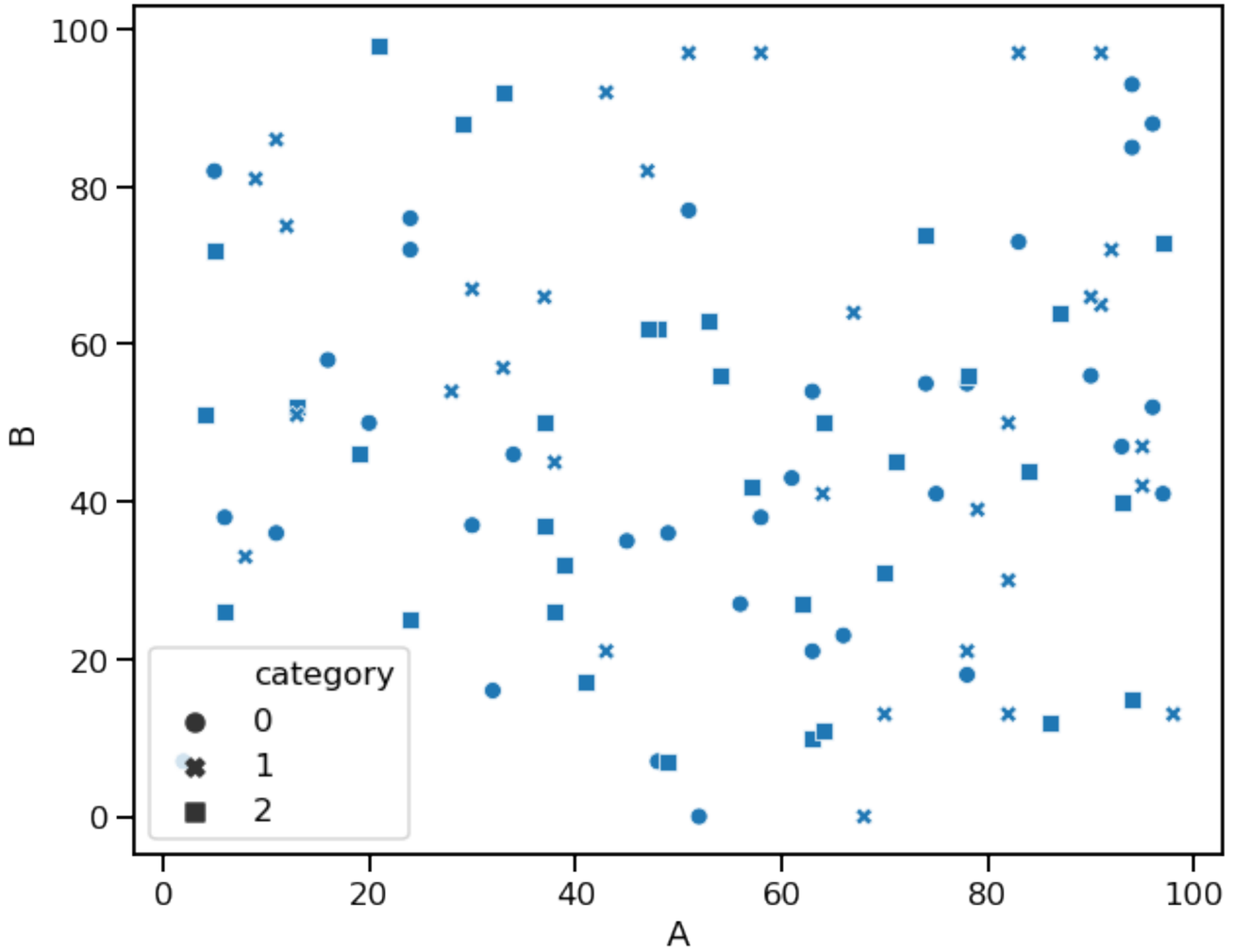



```
df4.plot(kind='hist',
         alpha=0.5,
         bins=50 )
```

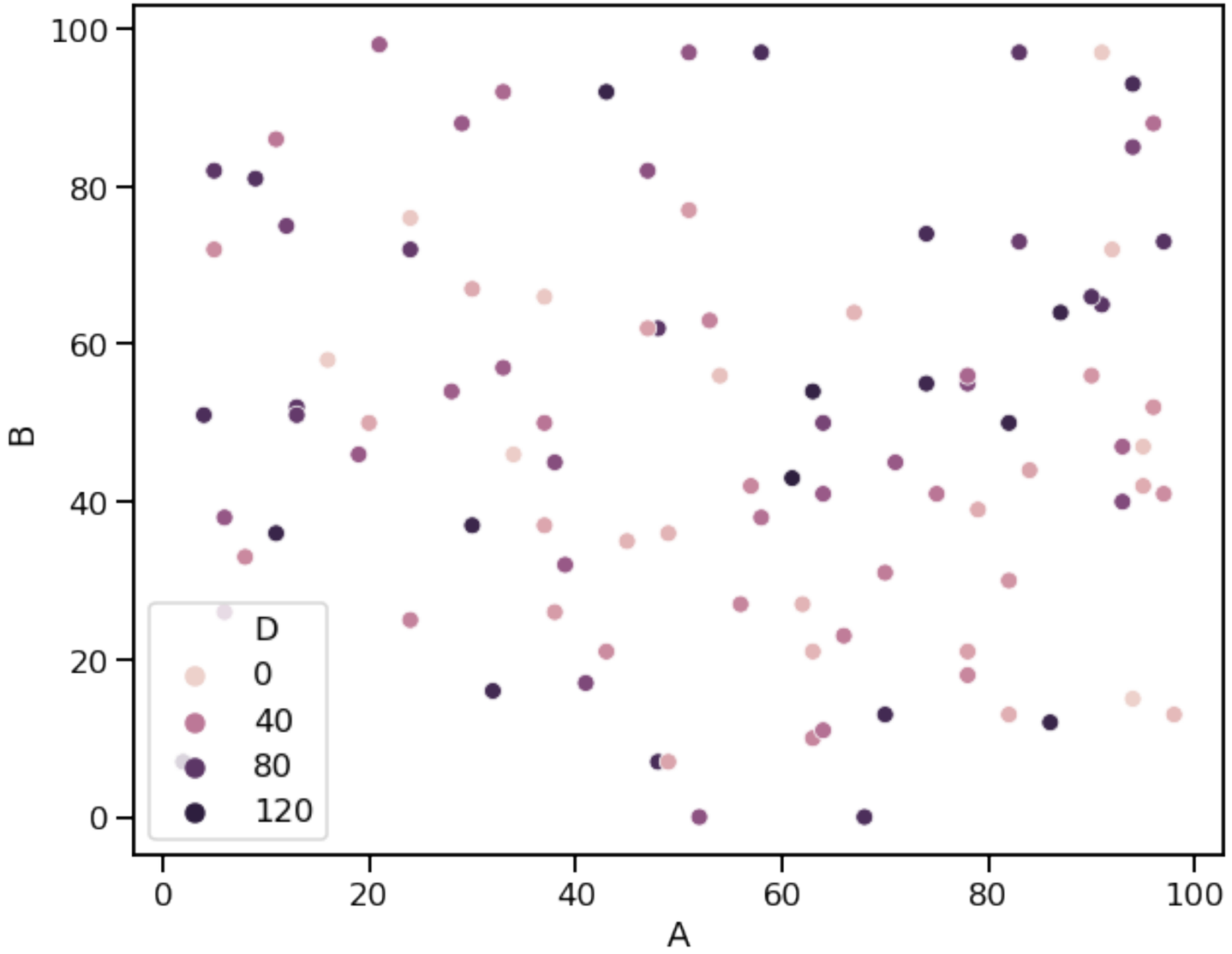

```
import seaborn as sns
```



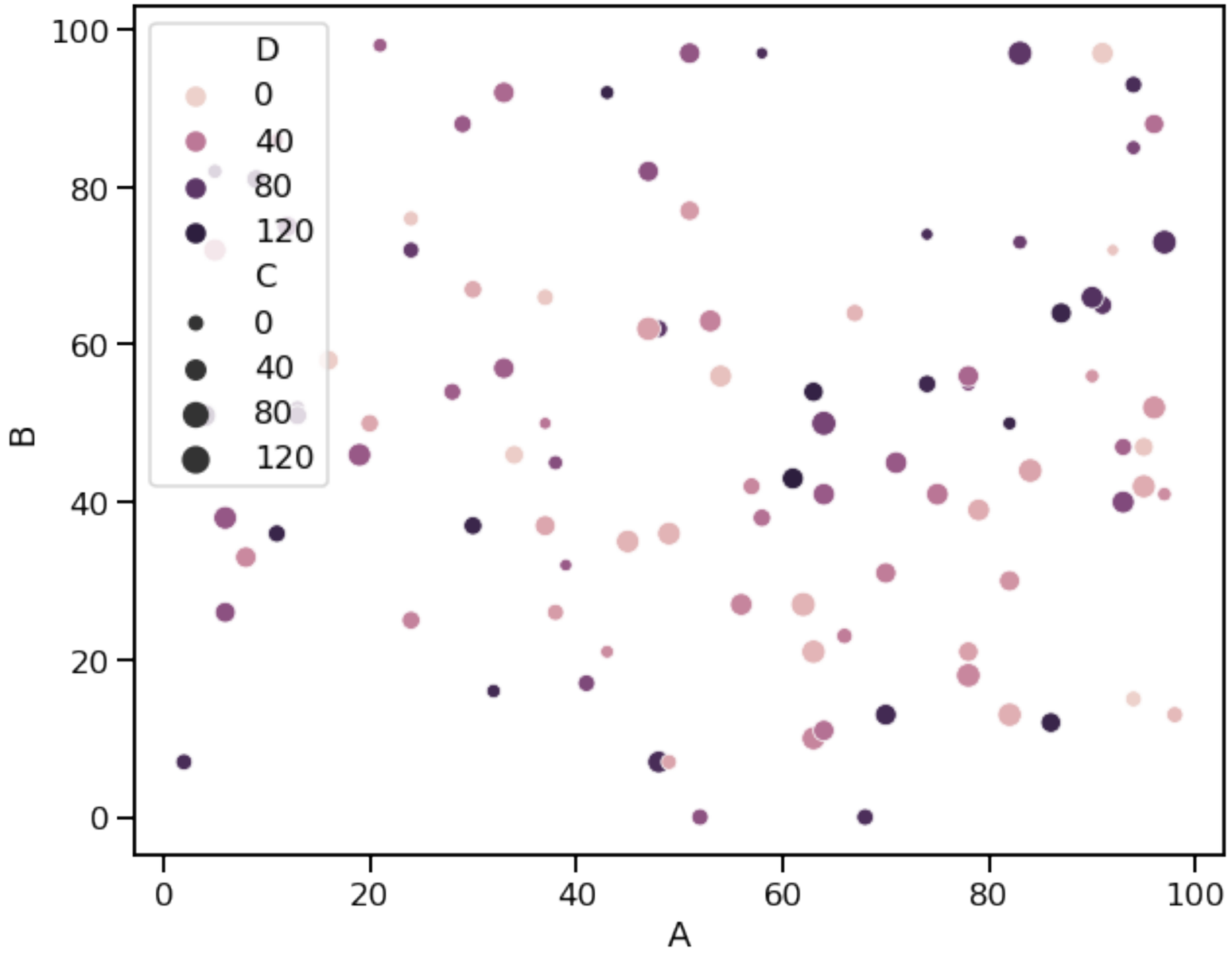

```
df = pd.DataFrame(np.random.randint(
    0,100,size=(100, 4)),
    columns=list('ABCD'))
df['category'] = df['C'] % 3

sns.scatterplot(x=df['A'], y=df['B'], data=df)
```




```
sns.scatterplot(x=df['A'], y=df['B'], style=df['category'])
```

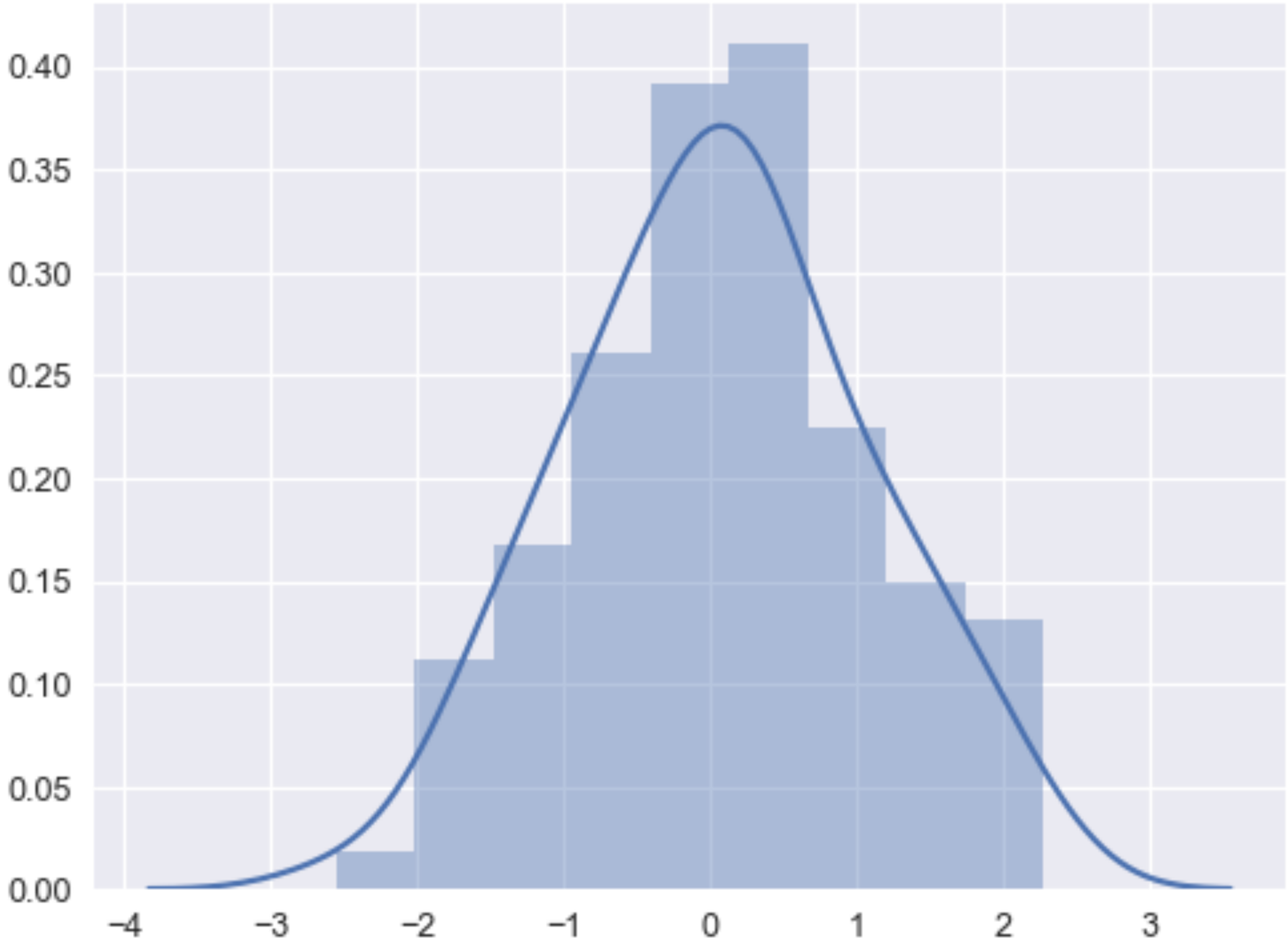



```
sns.scatterplot(x=df['A'], y=df['B'], hue=df['D'])
```

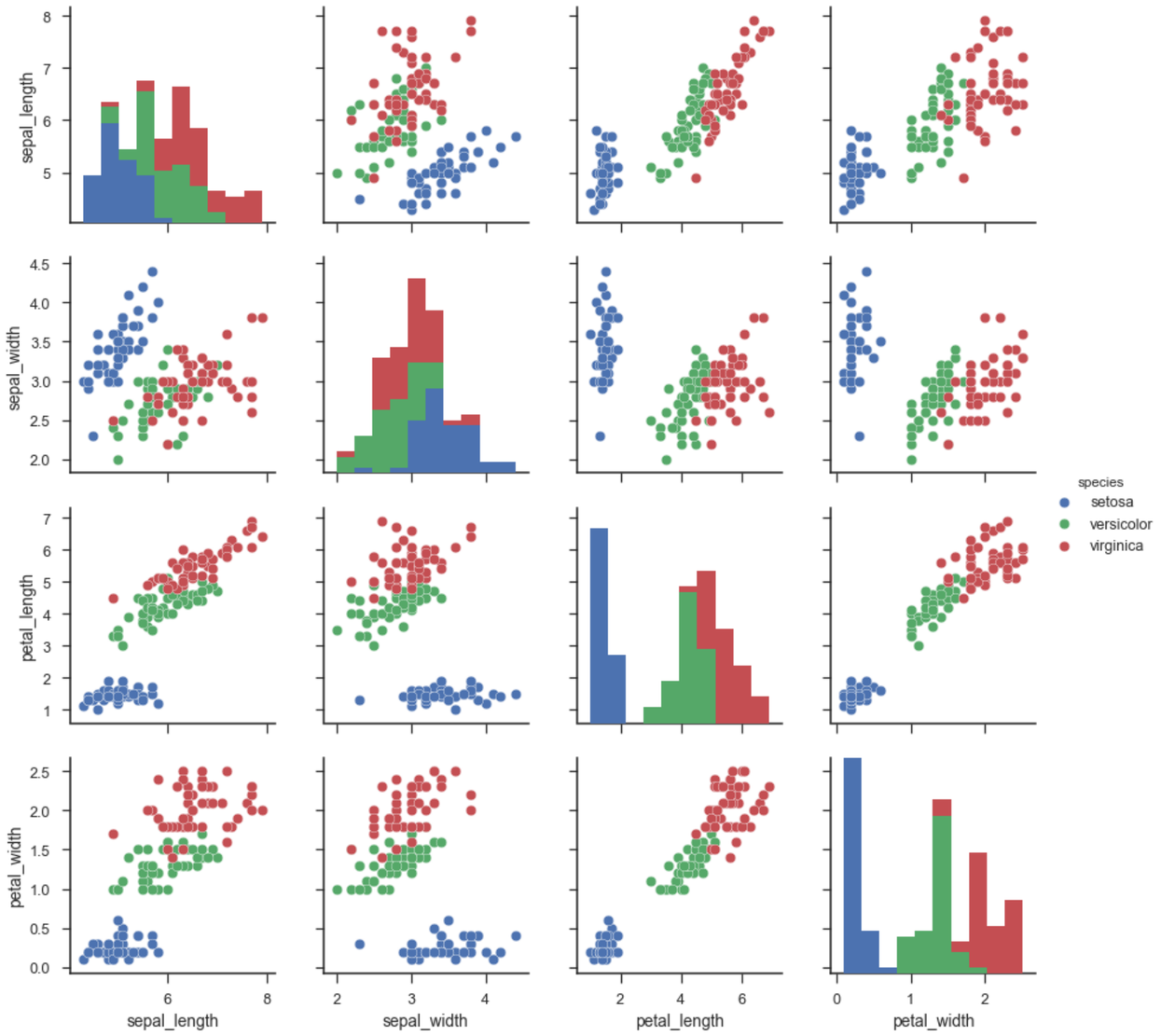



```
sns.scatterplot(x=df['A'], y=df['B'], size=df['C'],
                hue=df['D'])
```

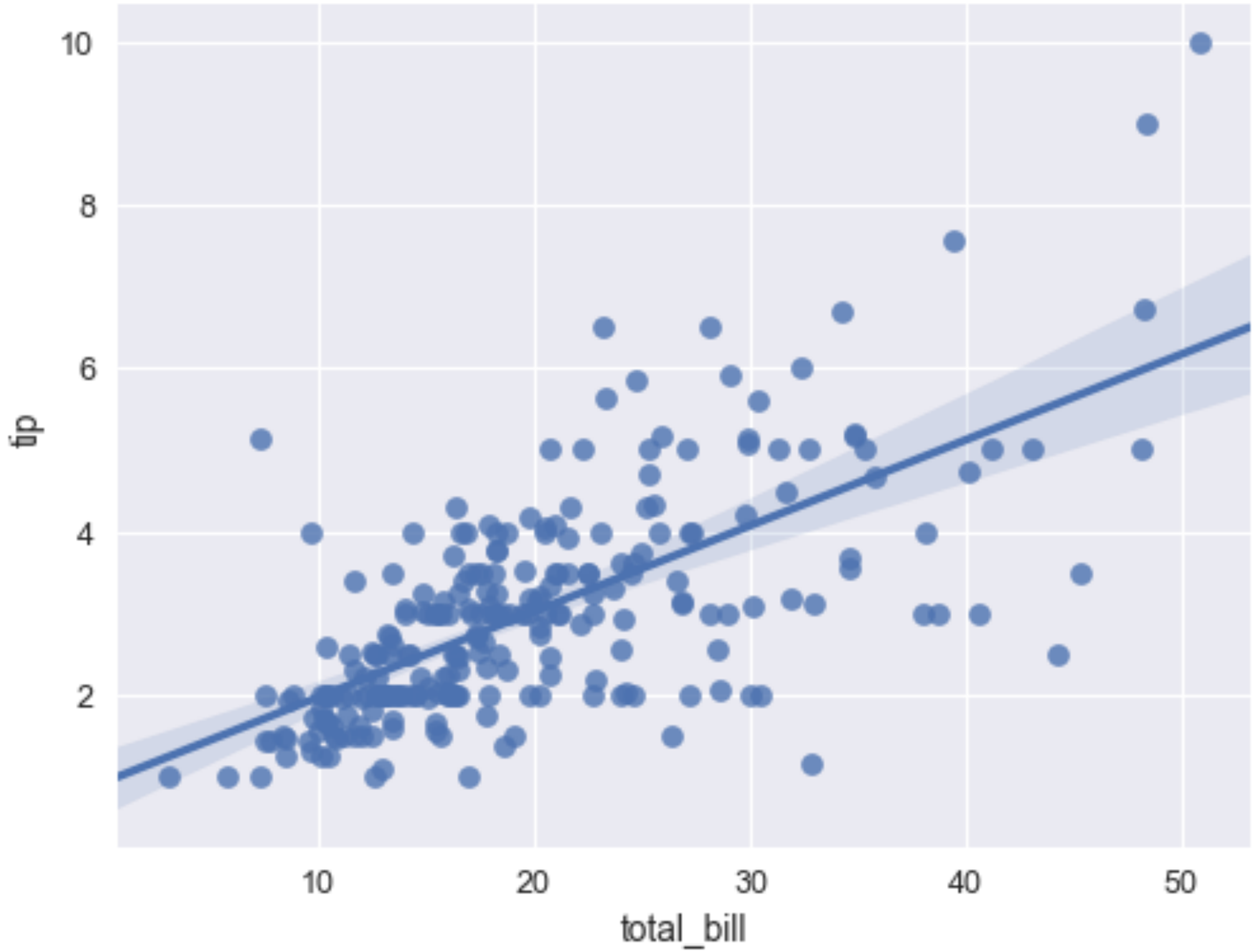

```
ax = sns.distplot(<data>)
```



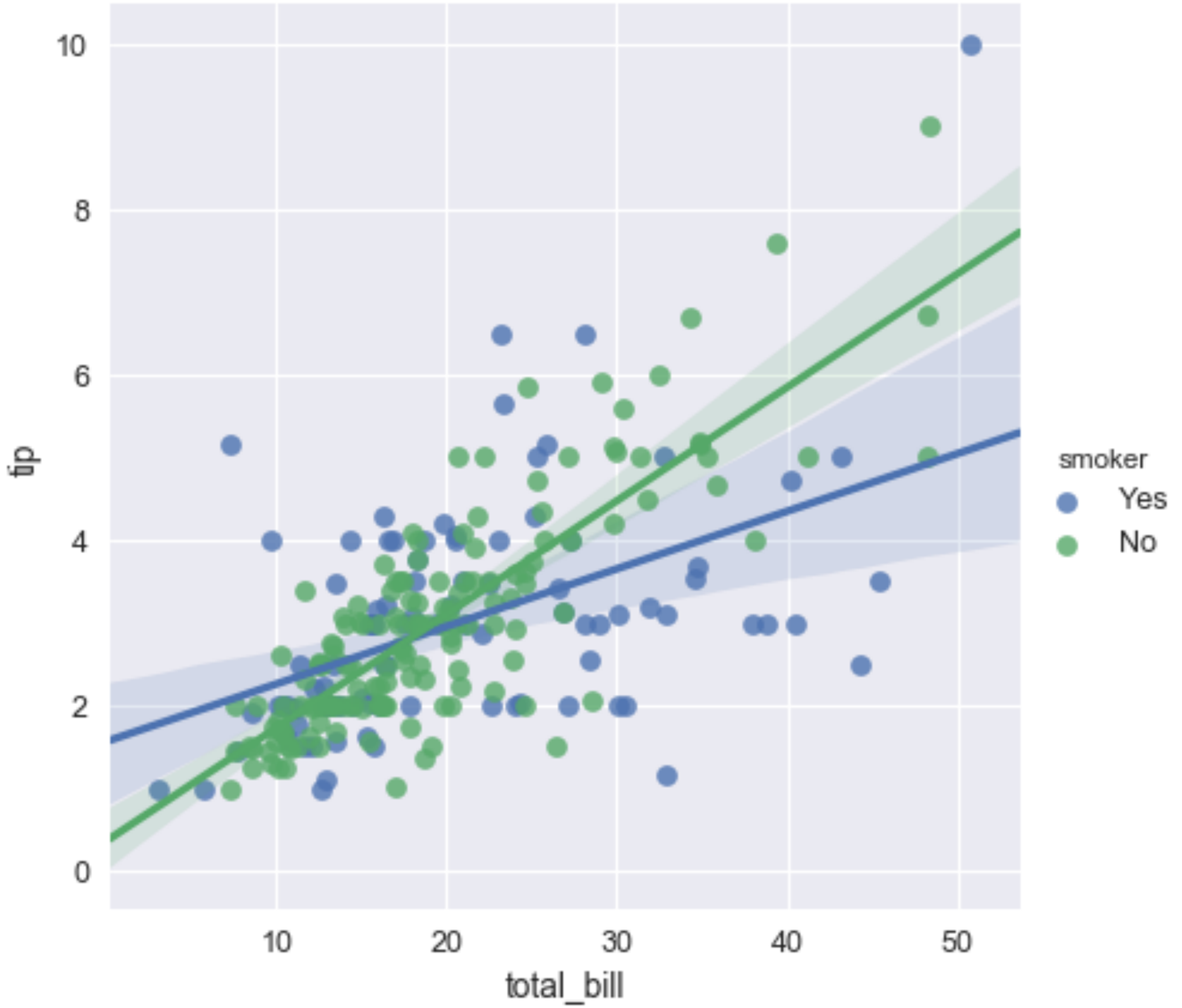
```
g = sns.pairplot(<data>, hue="<target>")
```



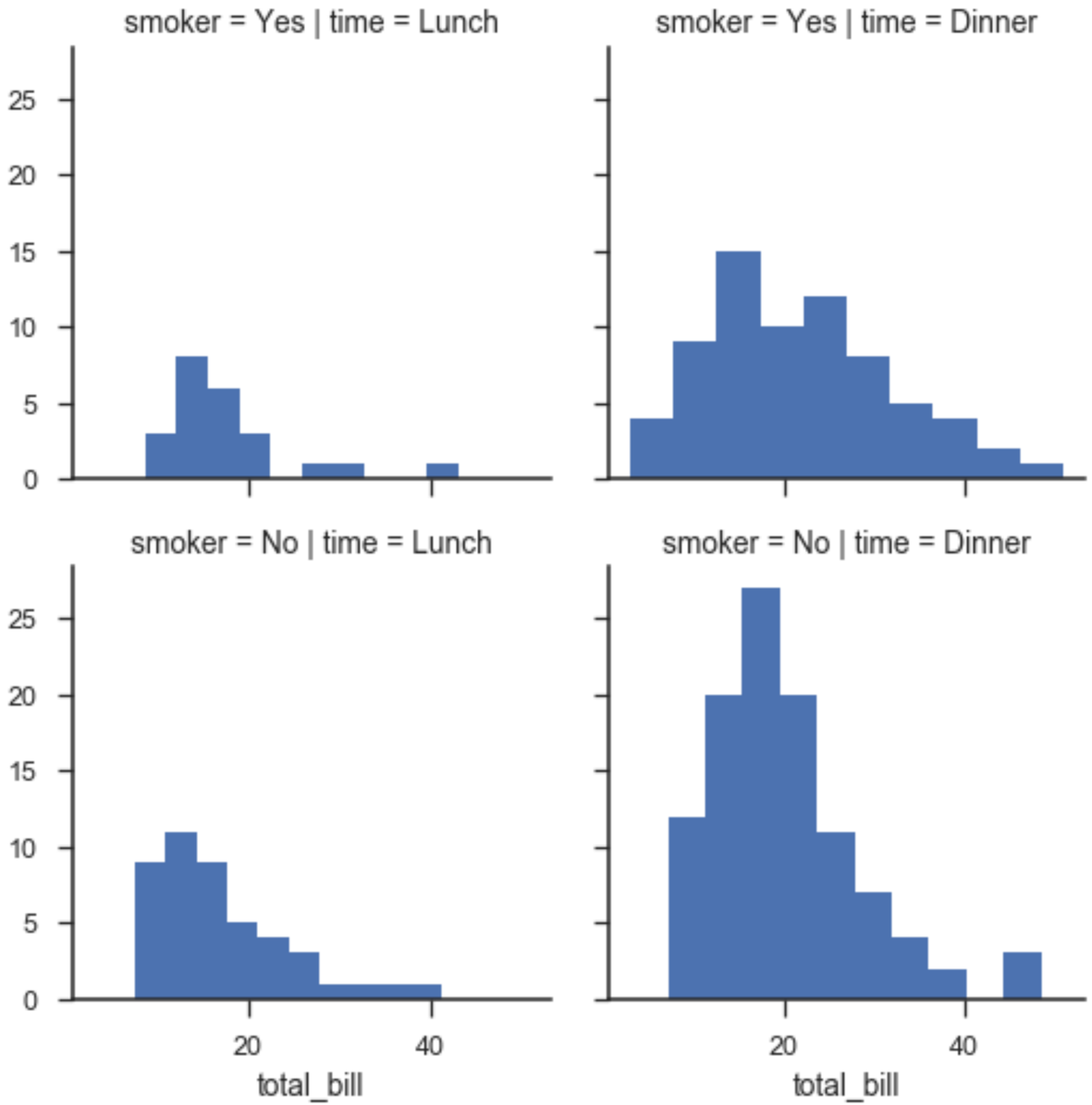

```
ax = sns.regplot(x="total_bill",
                 y="tip", data=tips)
```




```
g = sns.lmplot(x="total_bill",
y="tip", hue="smoker", data=tips)
```



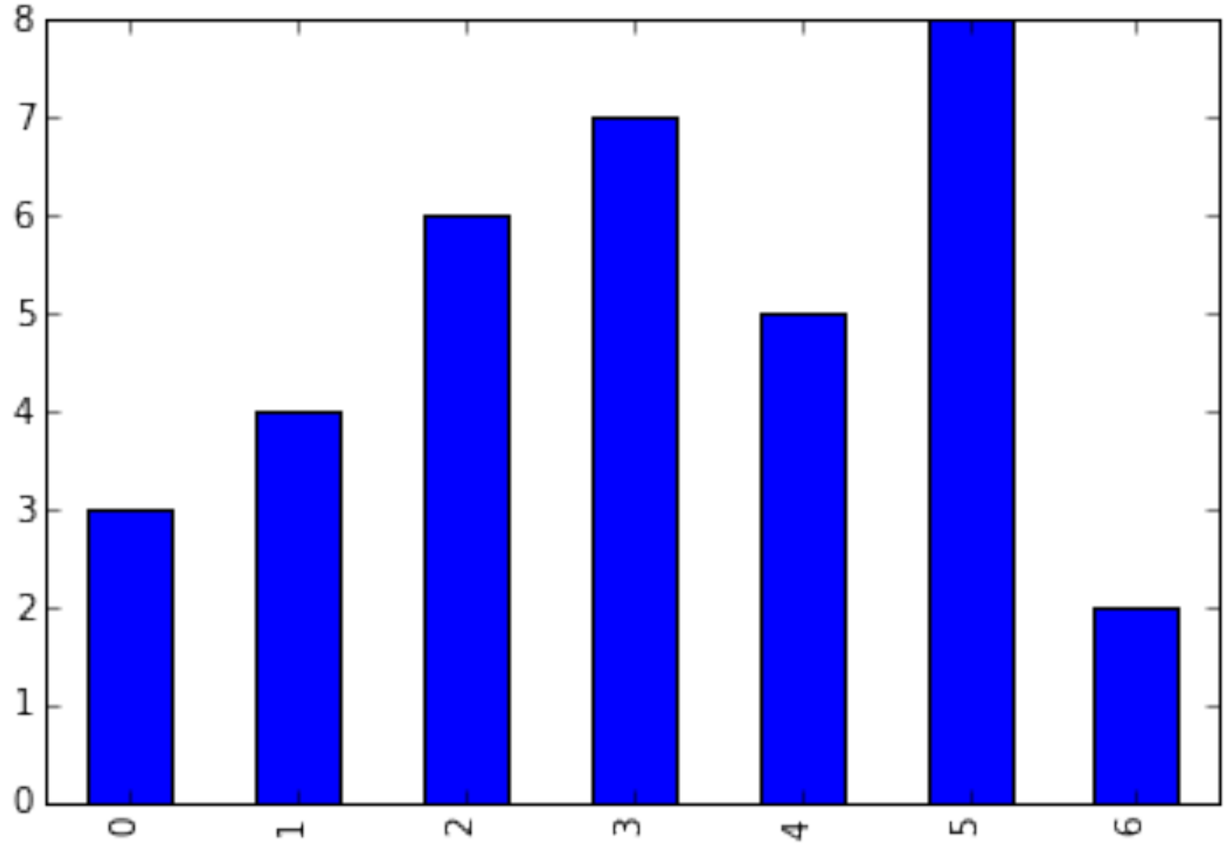

```
g = sns.FacetGrid(tips, col="time",  row="smoker")
>>> g = g.map(plt.hist, "total_bill")
```



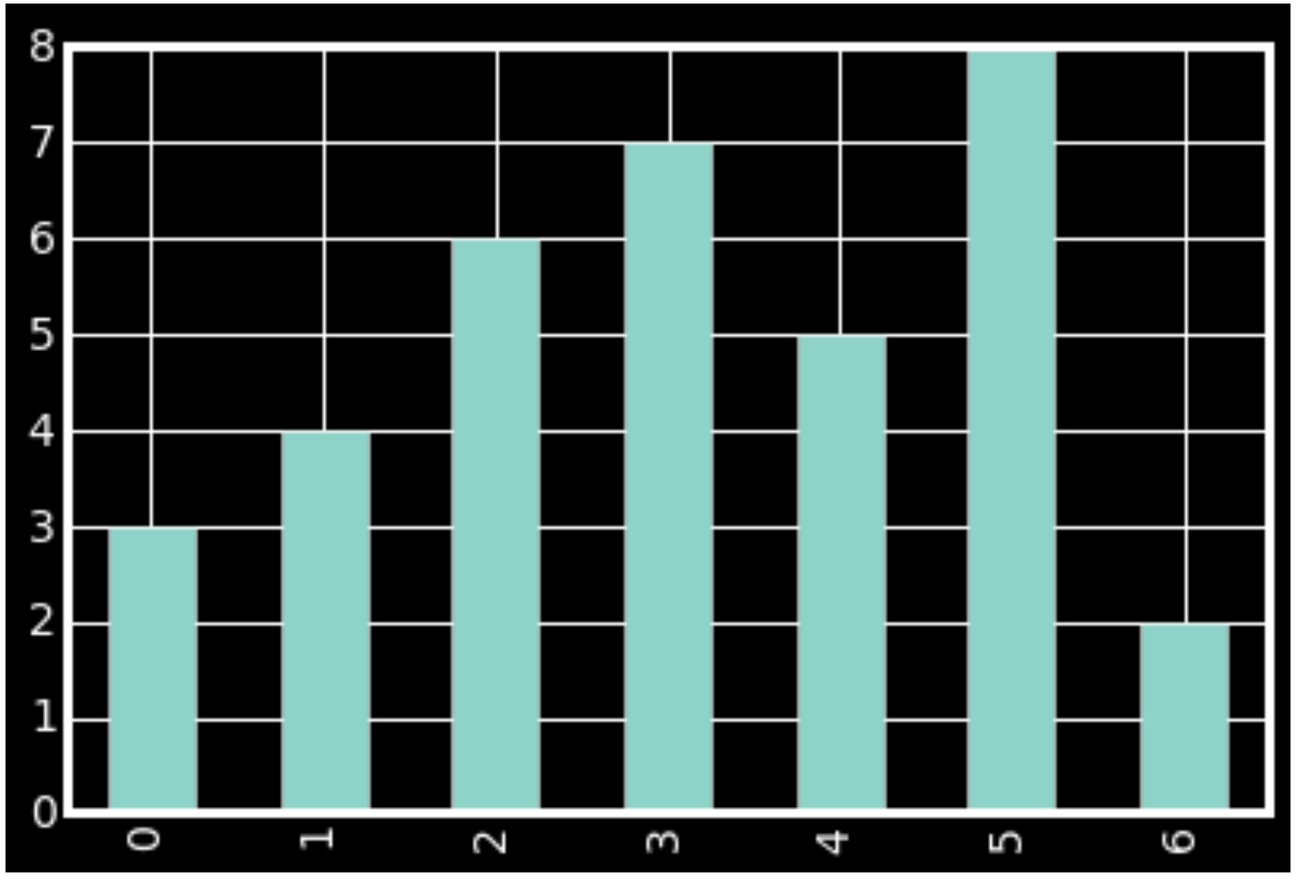

```
scatterPlot.get_figure().savefig( "scatterPlot.png" )
```


```
print(plt.style.available)

['dark_background', 'grayscale', 'ggplot',
'bmh', 'fivethirtyeight']
```

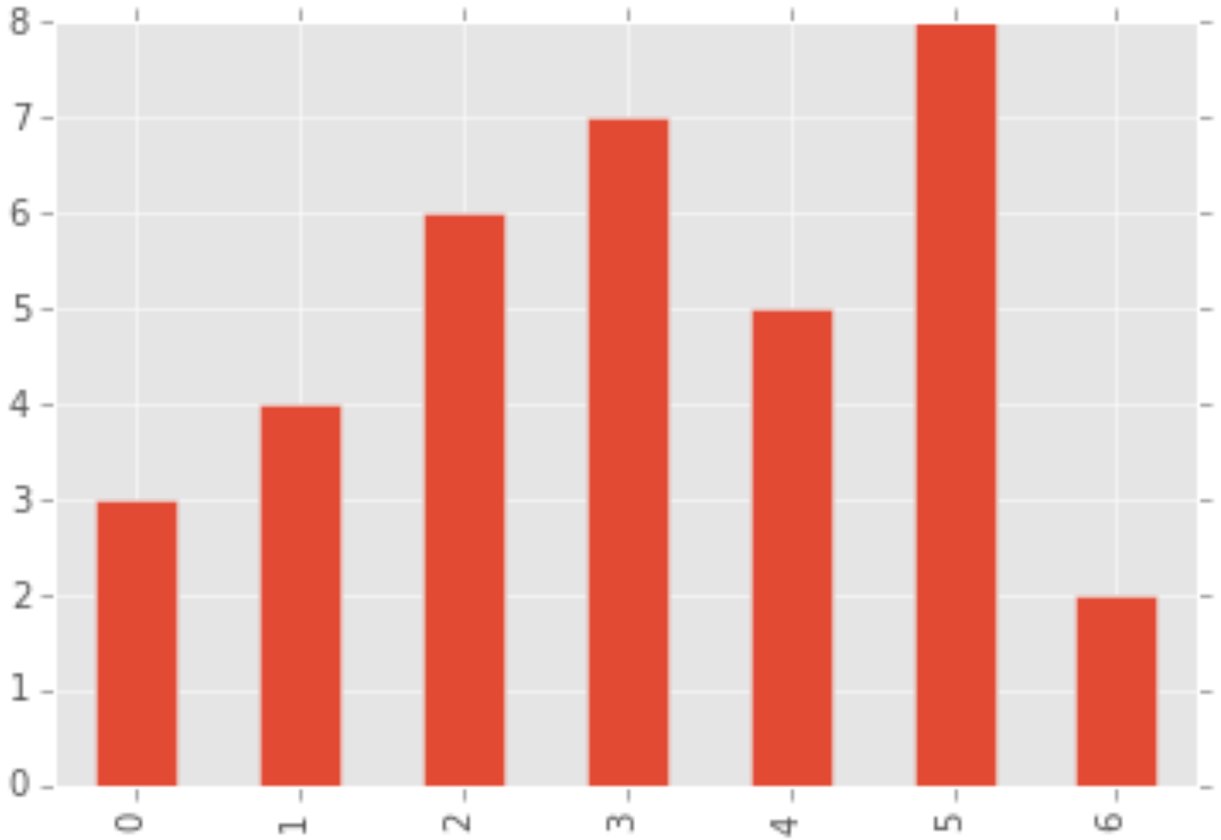
UGLY



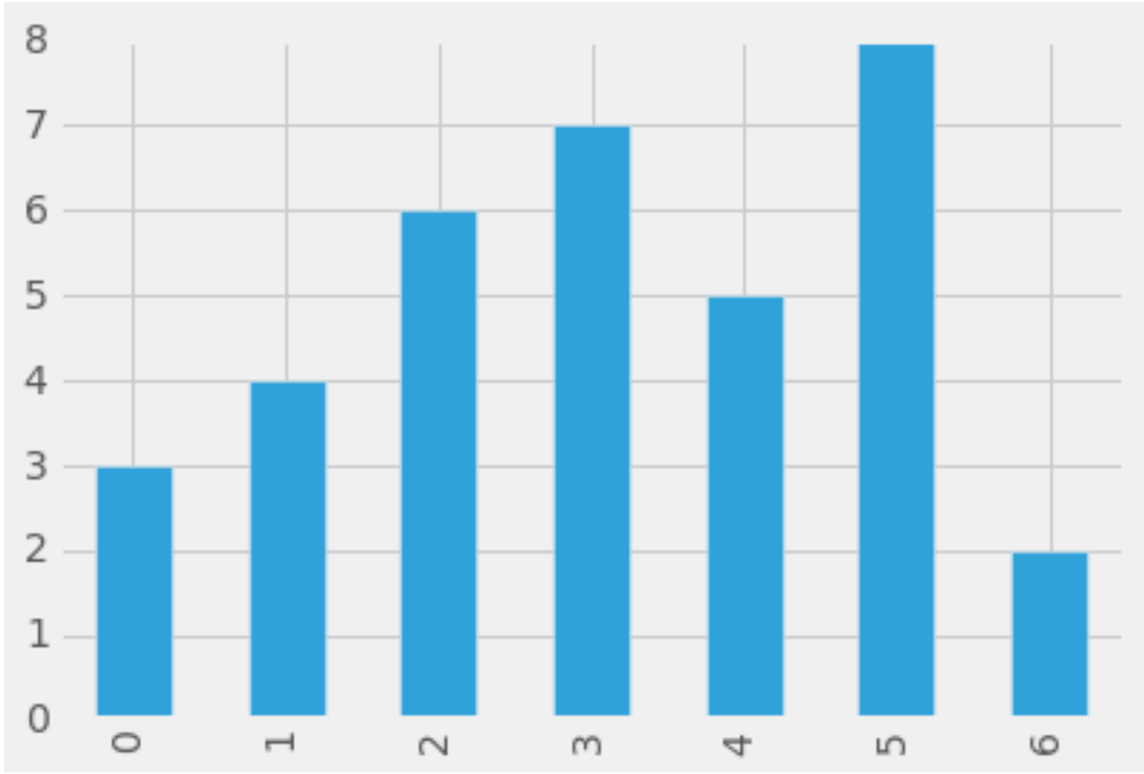

```
data = pd.Series( [3,4,6,7,5,8,2 ])
barchart = data.plot( kind="bar" )
```



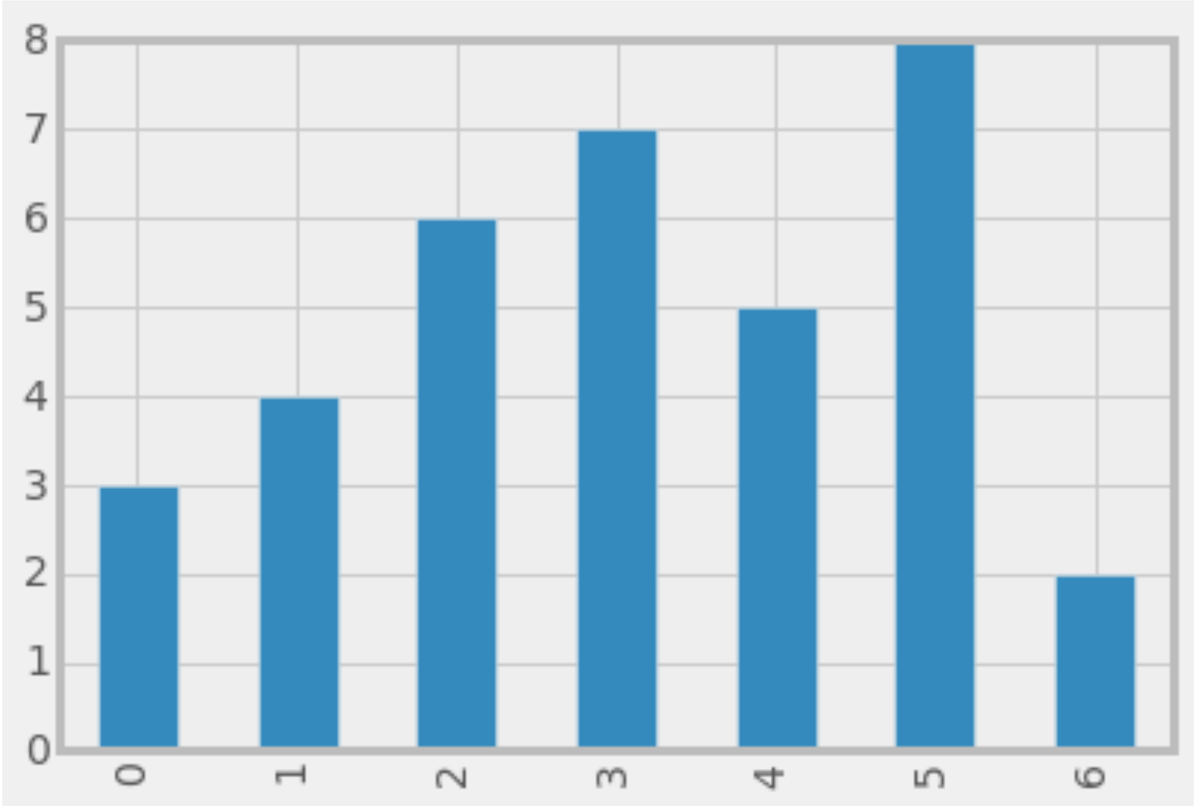

```
plt.style.use('dark_background')
barchart = data.plot( kind="bar" )
```




```
plt.style.use('ggplot')
barchart = data.plot( kind="bar" )
```



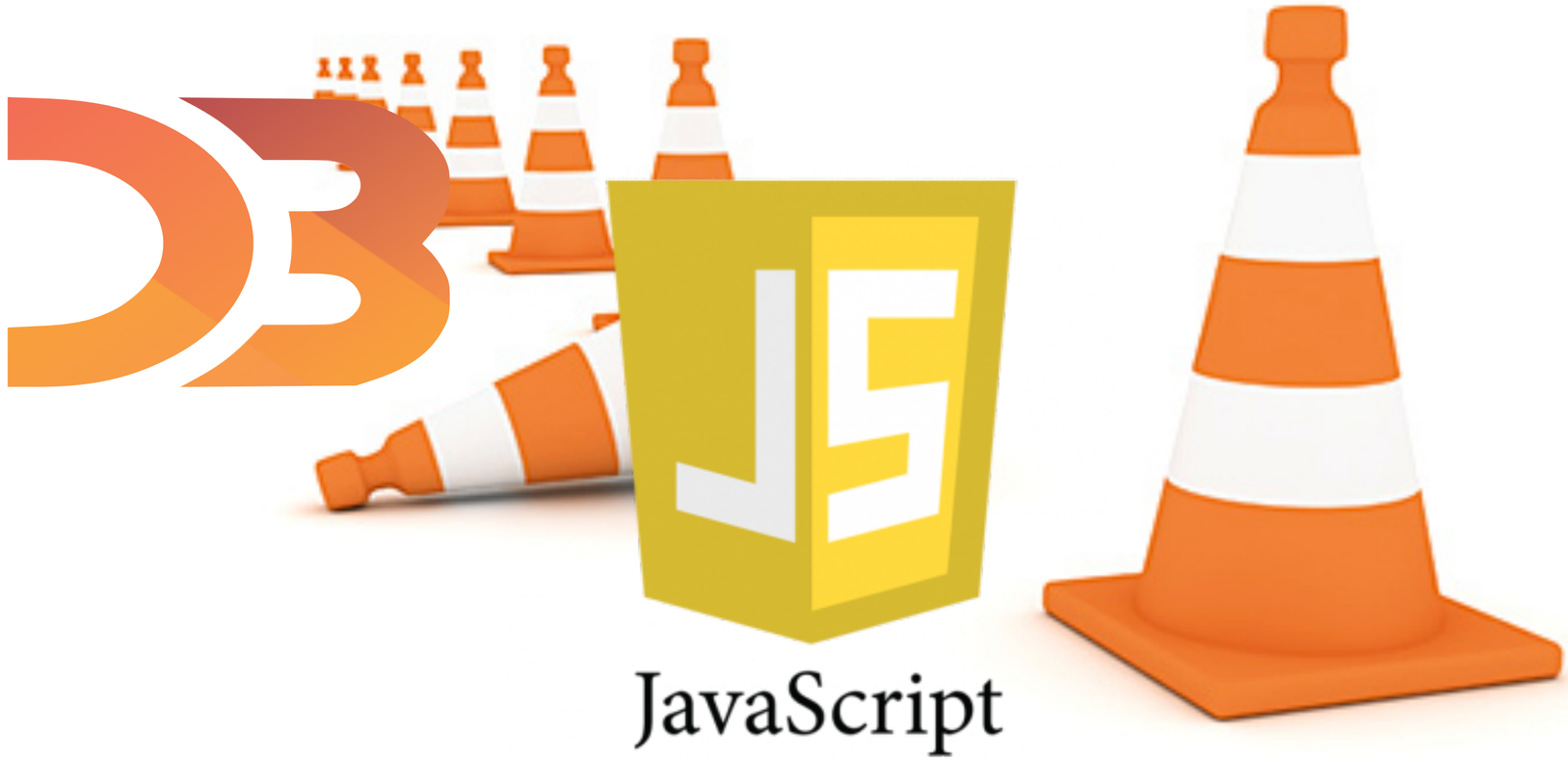

```
plt.style.use('fivethirtyeight')
barchart = data.plot( kind="bar" )
```




```
plt.style.use('bmh')
barchart = data.plot( kind="bar" )
```

Interactive Visualizations

QlikView

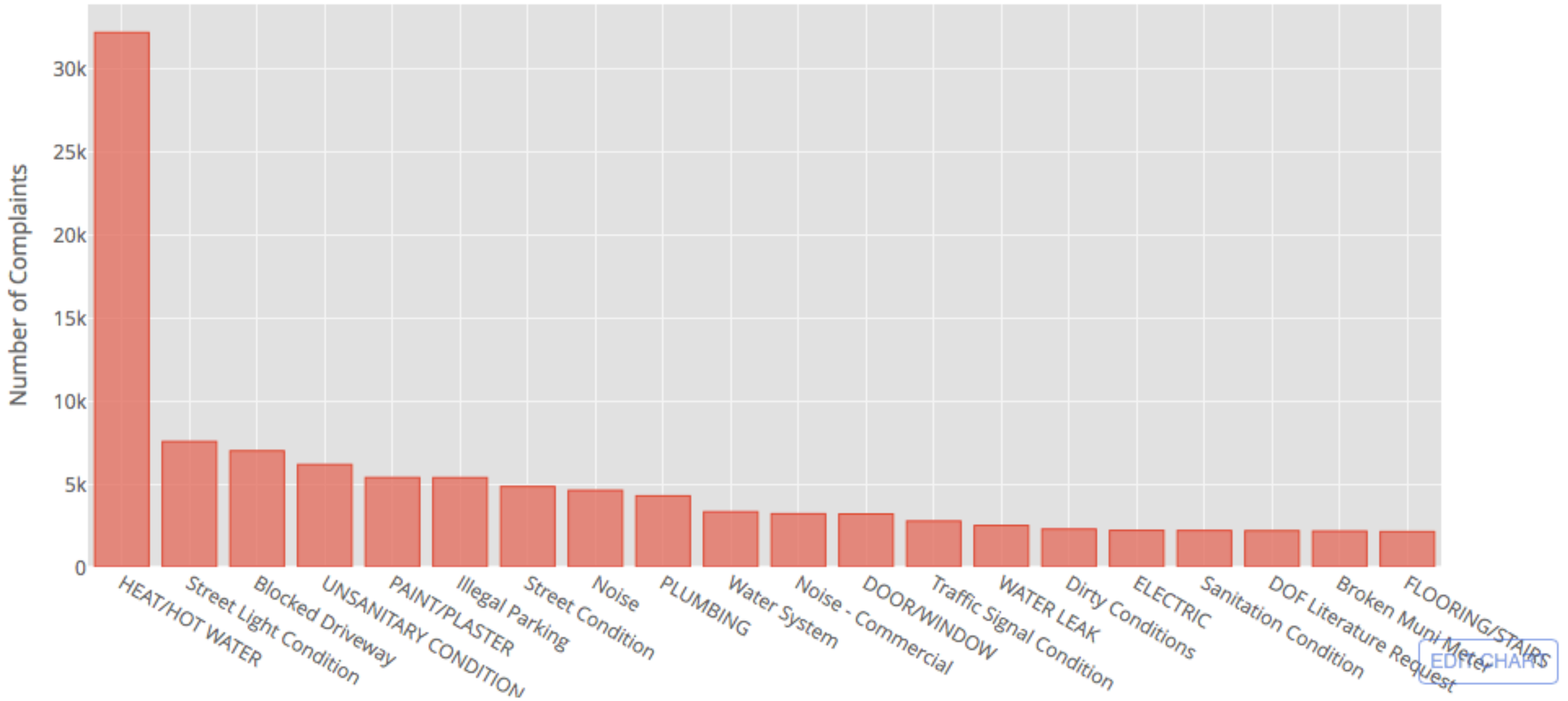

JavaScript

Easy to use… if you know R

GTK CYBER
GET ON THE FAST TRACK

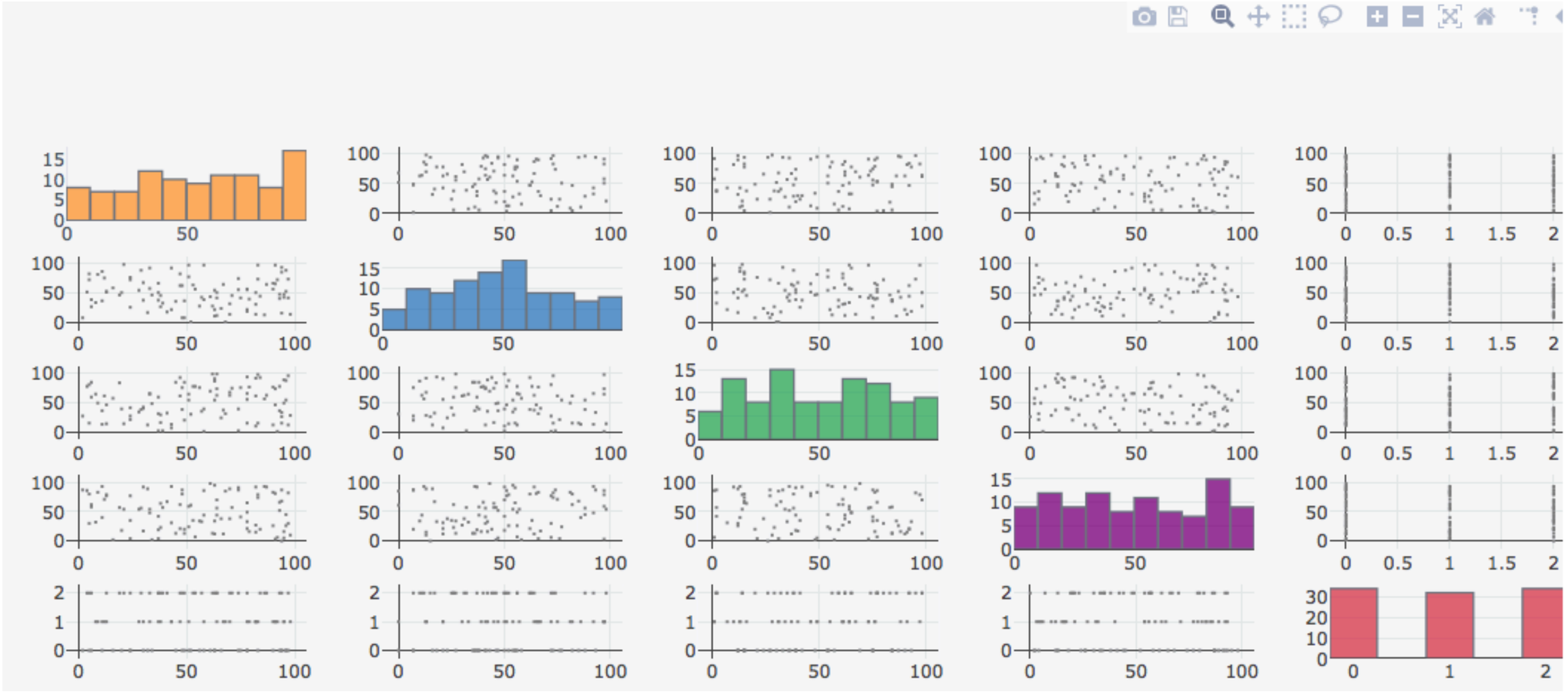
Introducing Cufflinks

gtkcyber.com




```
series.iplot(kind='bar', yTitle='Number of Complaints', title='NYC 311
Complaints')
```


https://plot.ly/ipython-notebooks/cufflinks/



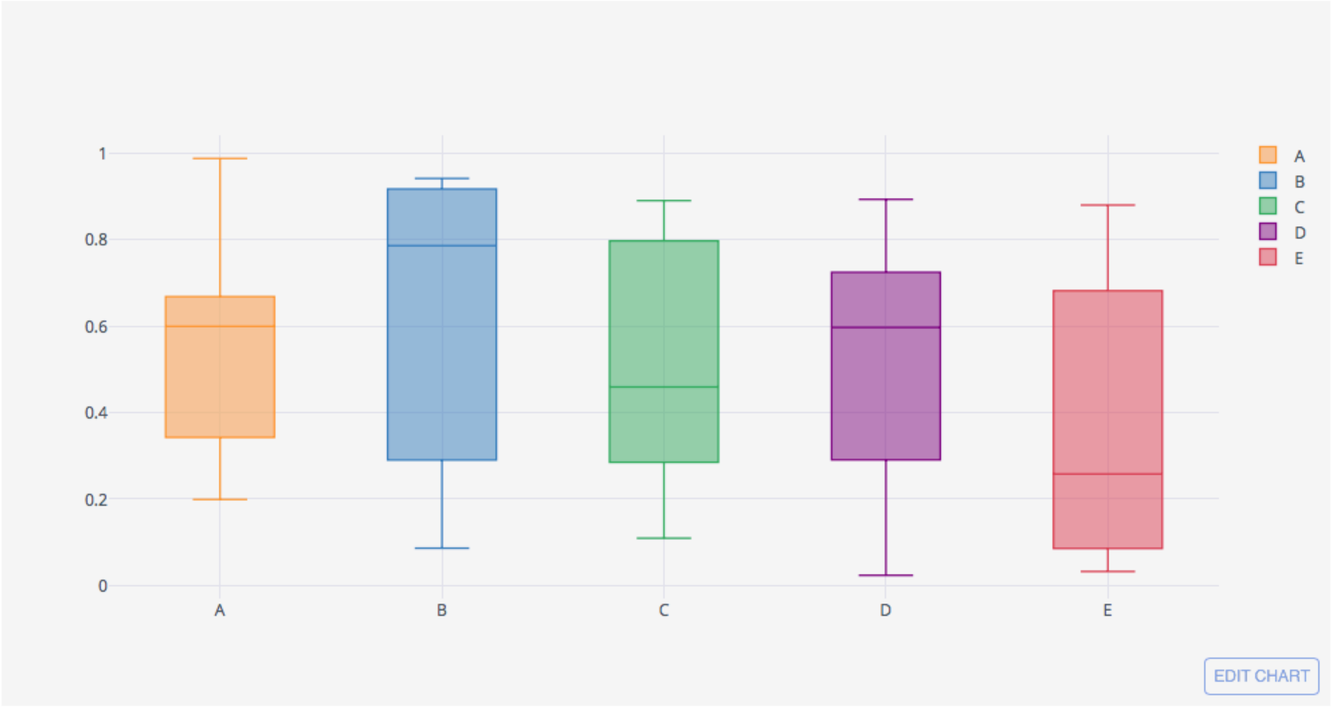
```
df.scatter_matrix()
```

https://plot.ly/ipython-notebooks/cufflinks/

Supported Chart Types

```
df.iplot(kind='<type>')
```

- line
- bar
- histogram
- box (boxplot)
- area
- scatter
- bubble
- heat map



https://plot.ly/ipython-notebooks/cufflinks/

In Class Exercise

Please complete Worksheet 4: Data Visualization

Please run the following command before running the worksheet.

```
conda install -c conda-forge panel
```

Questions?