Michael Grossberg

Data Visualization

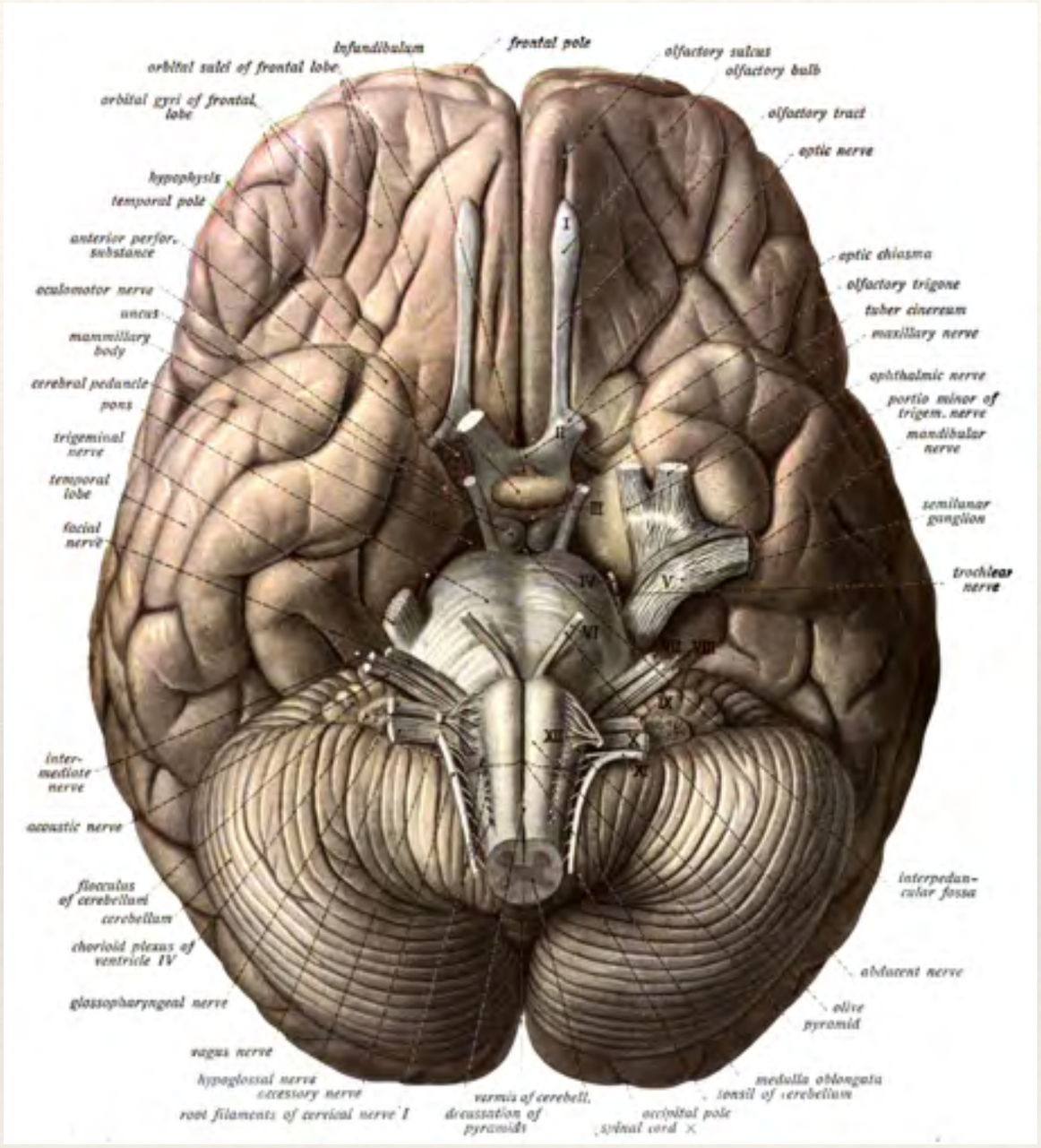
Basics

Tools, Principles and Pitfalls

Visualization as Tool

Whats the problem?

Information

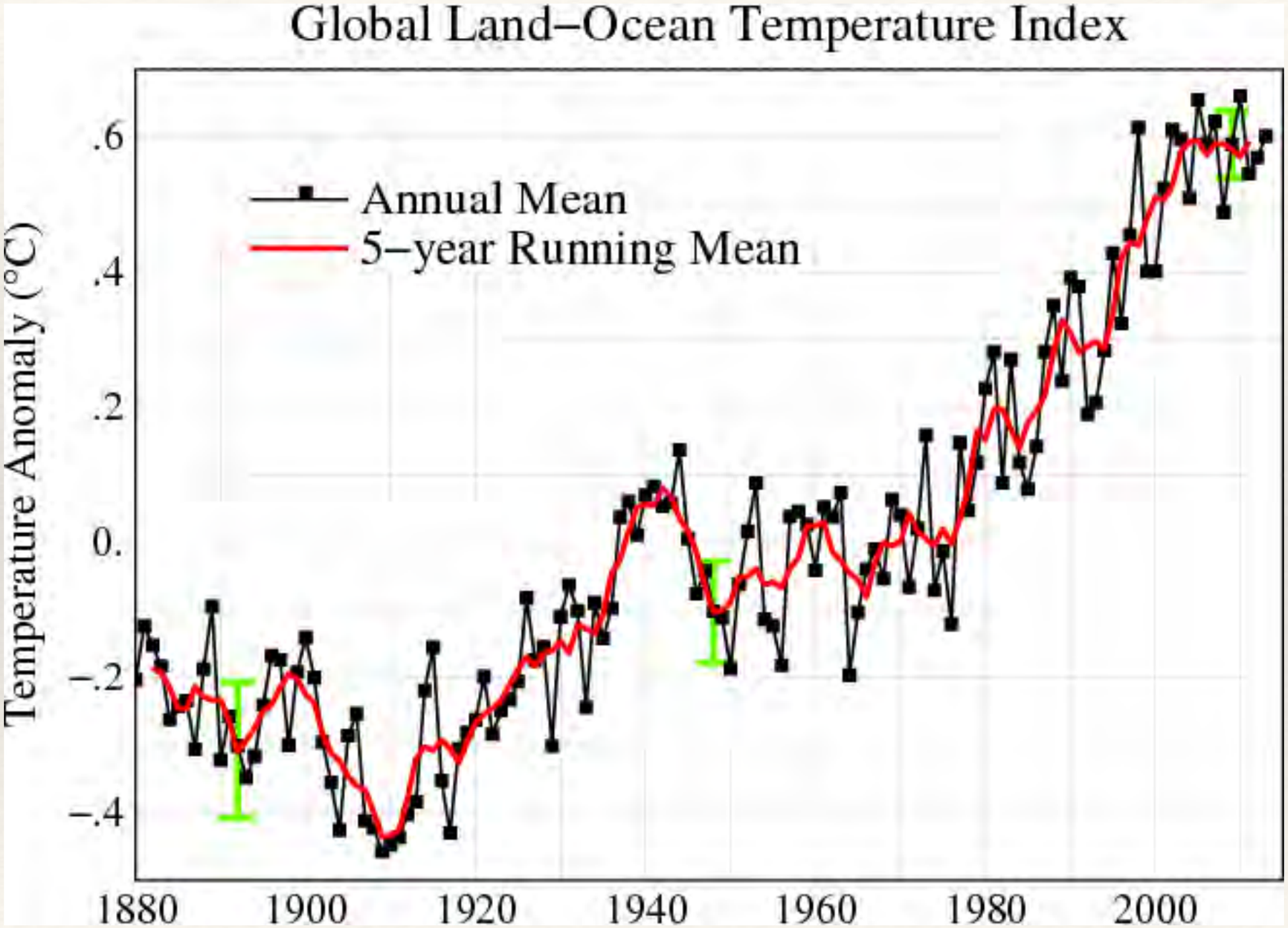


Understanding

Data Visualization Global Temp

Year	Annual_Mean	5-year_Mean
1880	-0.20	*
1881	-0.12	*
1882	-0.15	-0.19
1883	-0.18	-0.19
1884	-0.26	-0.22
1885	-0.24	-0.25
1886	-0.23	-0.25
1887	-0.31	-0.21
1888	-0.19	-0.23
1889	-0.09	-0.23
1890	-0.32	-0.23
1891	-0.26	-0.26
1892	-0.30	-0.31
1893	-0.35	-0.29
1894	-0.32	-0.27
1895	-0.24	-0.25
1896	-0.17	-0.24
1897	-0.17	-0.21
1898	-0.30	-0.19
1899	-0.19	-0.20
1900	-0.14	-0.23
1901	-0.20	-0.24
1902	-0.30	-0.28
1903	-0.36	-0.31
1904	-0.43	-0.32
1905	-0.29	-0.35
1906	-0.25	-0.36
1907	-0.41	-0.37
1908	-0.42	-0.40
1909	-0.46	-0.44
1910	-0.45	-0.43
1911	-0.44	-0.43
1912	-0.40	-0.38
1913	-0.38	-0.32
1914	-0.22	-0.30
1915	-0.16	-0.31
1916	-0.35	-0.29
1917	-0.43	-0.31
1918	-0.31	-0.33
1919	-0.28	-0.30
1920	-0.26	-0.27
1921	-0.20	-0.26
1922	-0.28	-0.25
1923	-0.25	-0.23
1924	-0.23	-0.21
1925	-0.21	-0.19
1926	-0.08	-0.17
1927	-0.17	-0.18
1928	-0.16	-0.16
1929	-0.30	-0.16
1930	-0.11	-0.15
1931	-0.06	-0.16
1932	-0.10	-0.12
1933	-0.24	-0.13
1934	-0.09	-0.14
1935	-0.14	-0.11
1936	-0.10	-0.05
1937	0.04	-0.03
1938	0.06	0.02
1939	0.01	0.05
1940	0.07	0.06
1941	0.08	0.06
1942	0.05	0.08
1943	0.06	0.07
1944	0.14	0.04
1945	0.01	0.02
1946	-0.08	-0.02
1947	-0.04	-0.07
1948	-0.10	-0.10
1949	-0.11	-0.10
1950	-0.19	-0.09
1951	-0.06	-0.05
1952	0.02	-0.05
1953	0.09	-0.04
1954	-0.11	-0.06
1955	-0.12	-0.06
1956	-0.18	-0.07
1957	0.04	-0.04
1958	0.05	-0.02
1959	0.03	0.02
1960	-0.04	0.02
1961	0.05	0.03
1962	0.04	-0.01
1963	0.07	-0.03
1964	-0.20	-0.05
1965	-0.10	-0.06
1966	-0.04	-0.08
1967	-0.01	-0.03
1968	-0.05	-0.00
1969	0.06	-0.01
1970	0.04	0.00
1971	-0.07	0.04
1972	0.02	0.02
1973	0.16	0.01
1974	-0.07	-0.01
1975	-0.01	0.02
1976	-0.12	-0.00
1977	0.15	0.04
1978	0.05	0.08
1979	0.12	0.16
1980	0.23	0.15
1981	0.28	0.20
1982	0.09	0.20
1983	0.27	0.17
1984	0.12	0.14
1985	0.08	0.18
1986	0.14	0.19
1987	0.28	0.22
1988	0.35	0.28
1989	0.24	0.33
1990	0.39	0.31
1991	0.38	0.28
1992	0.19	0.29
1993	0.21	0.30
1994	0.28	0.29
1995	0.43	0.34
1996	0.32	0.42
1997	0.45	0.44
1998	0.61	0.44
1999	0.40	0.48
2000	0.40	0.51
2001	0.52	0.51
2002	0.61	0.53
2003	0.60	0.58
2004	0.51	0.59
2005	0.65	0.59
2006	0.59	0.57
2007	0.62	0.59
2008	0.49	0.59
2009	0.59	0.58
2010	0.66	0.57
2011	0.55	0.59
2012	0.57	*
2013	0.60	*
2014	*	*

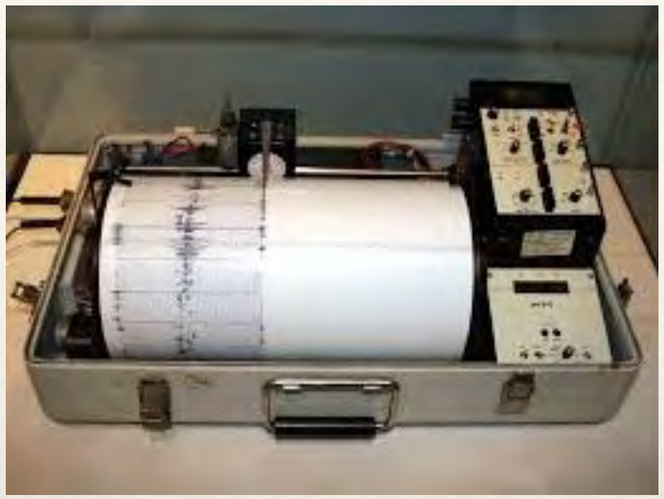
Global Means Temp as Graph



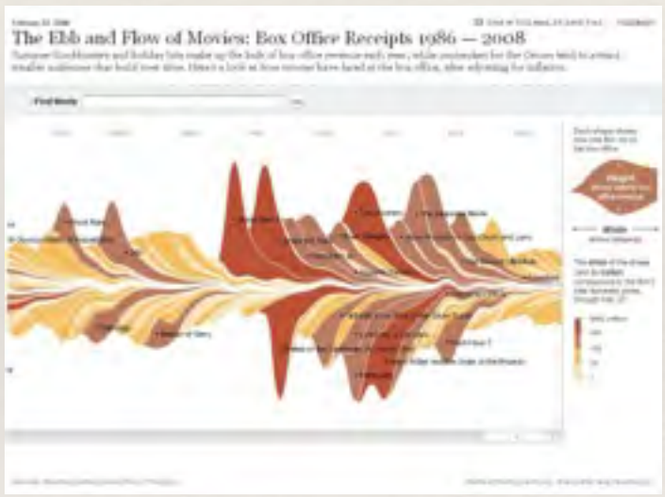
Hansen et al. (2006), NASA GISS

Goals of Visualization

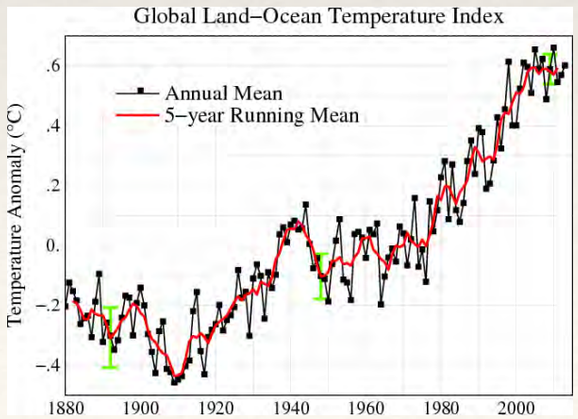
- Record



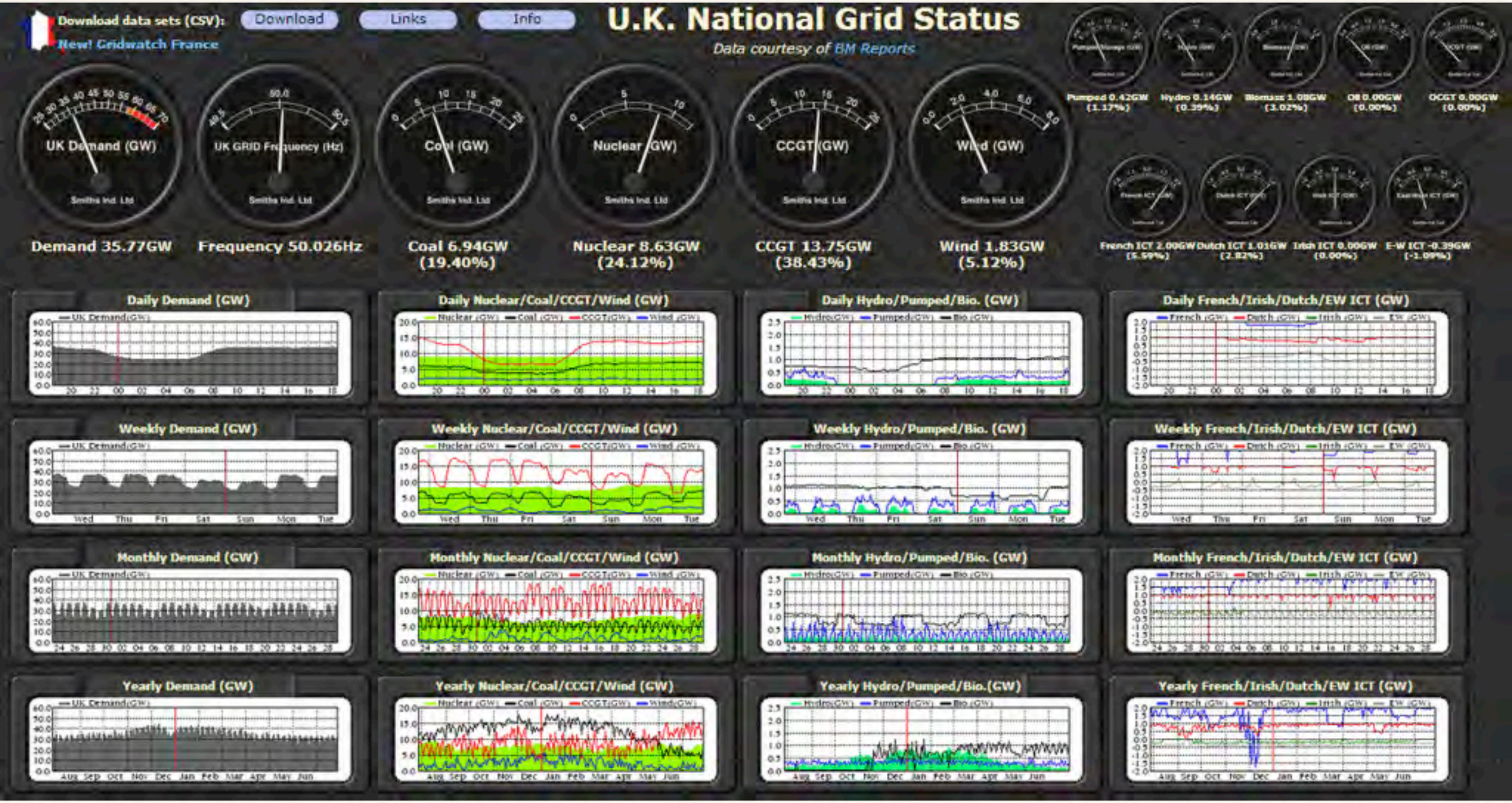
- Analyze



- Communicate

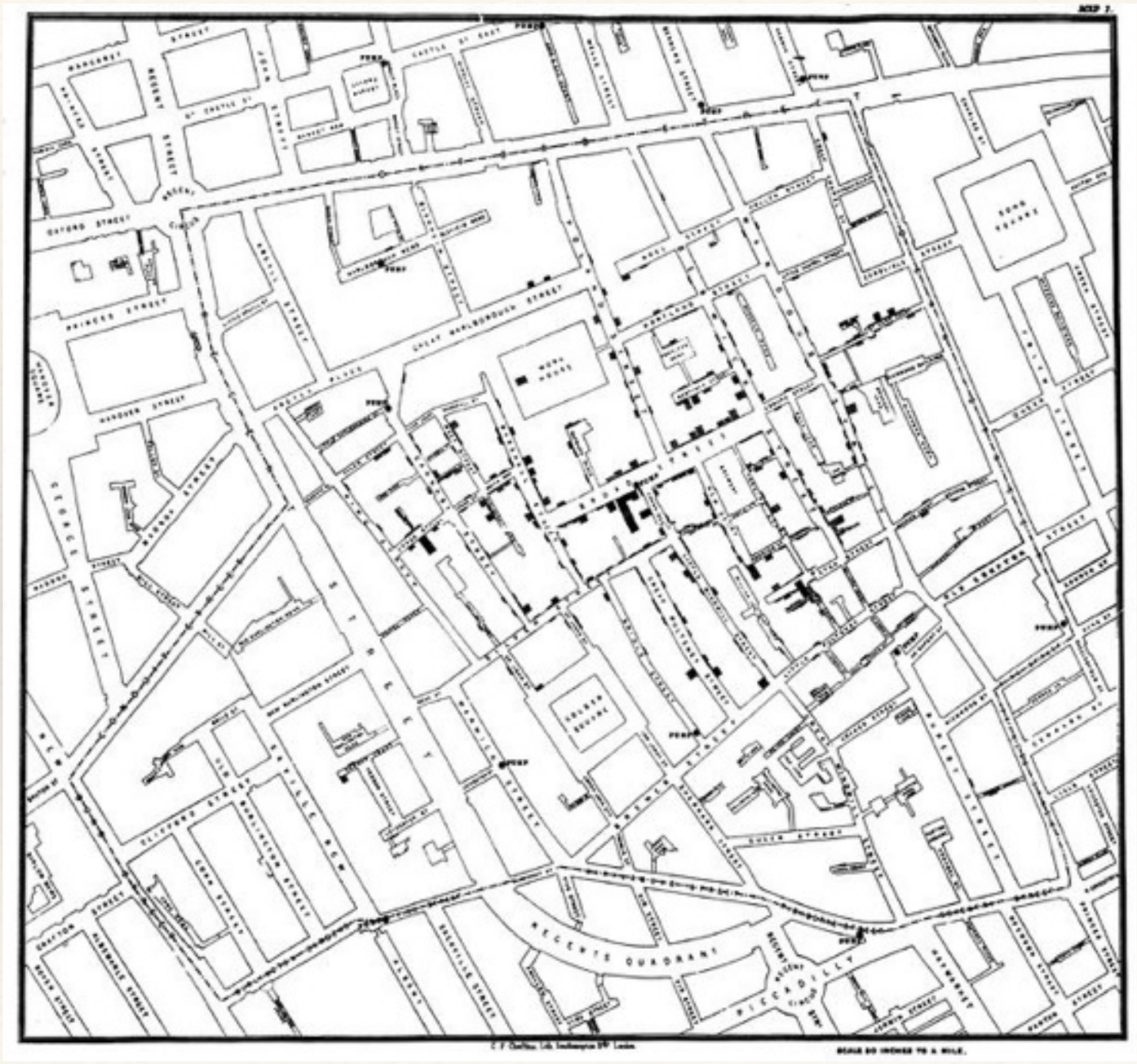


Analyze/Monitor



Analyze

Exploratory Data Analysis (EDA)

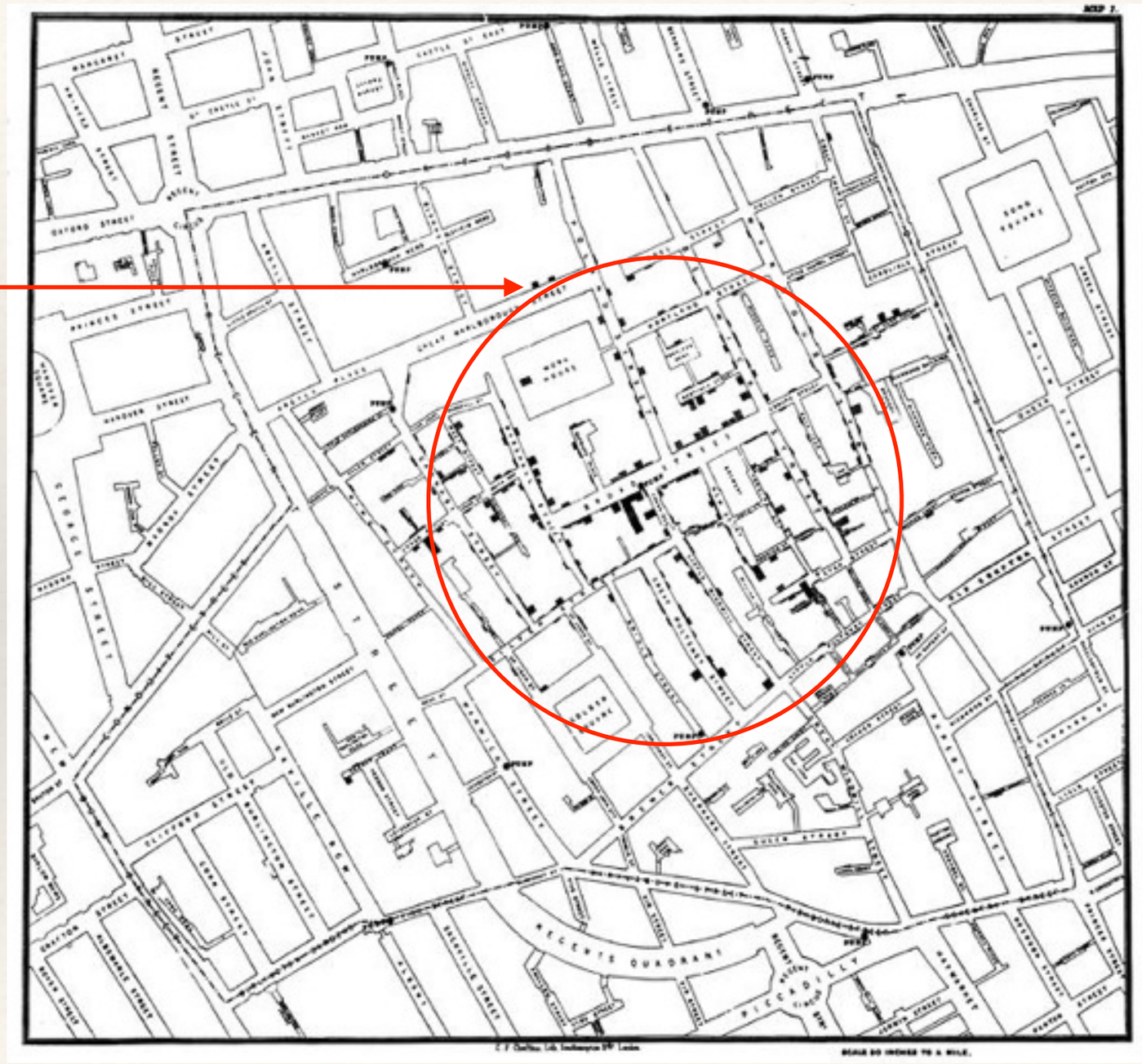


John Snow, 1854

Analyze

Cluster Region

Exploratory Data Analysis (EDA)



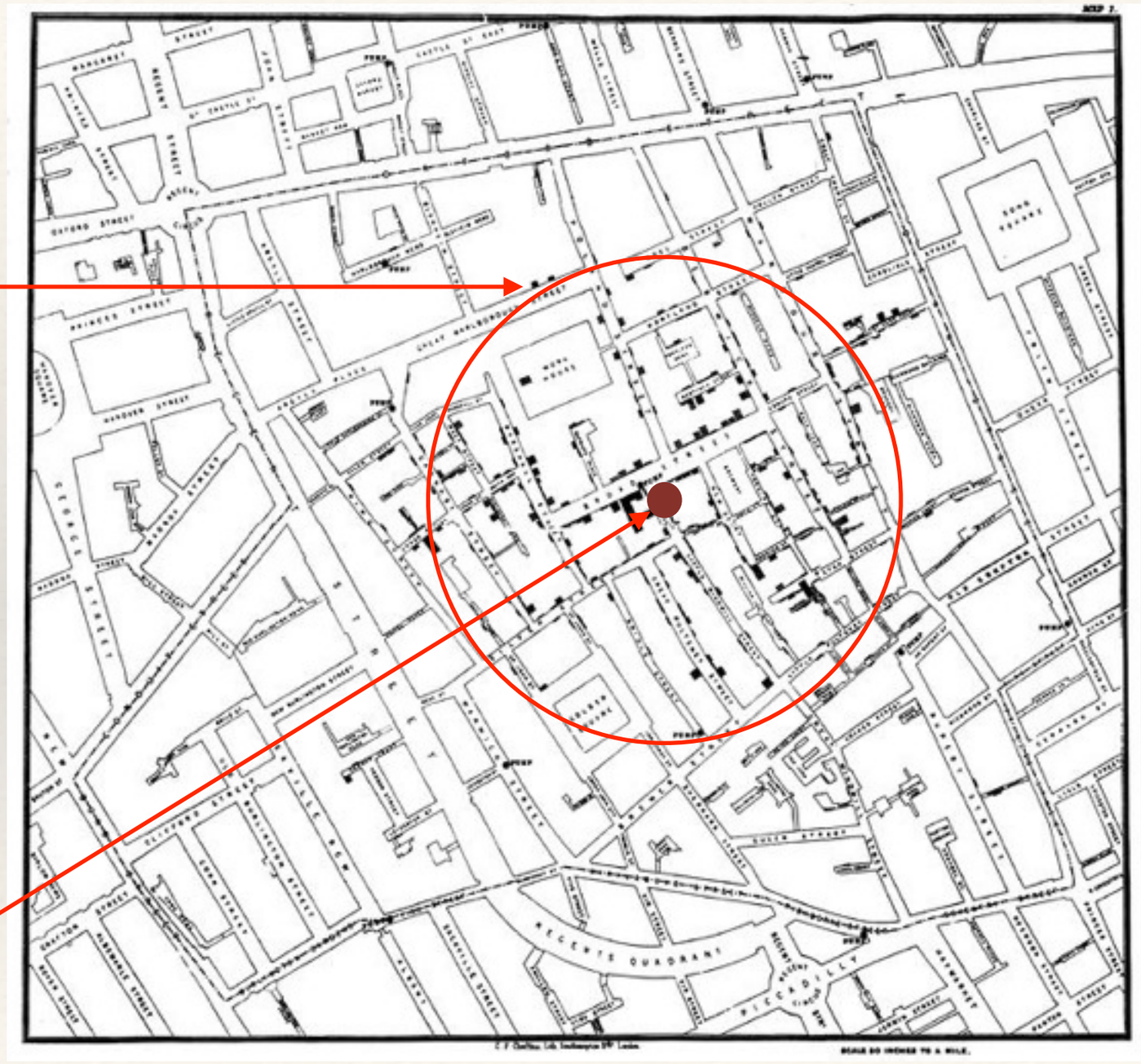
John Snow, 1854

Analyze

Exploratory Data Analysis (EDA)

Cluster Region

Cluster Center



John Snow, 1854

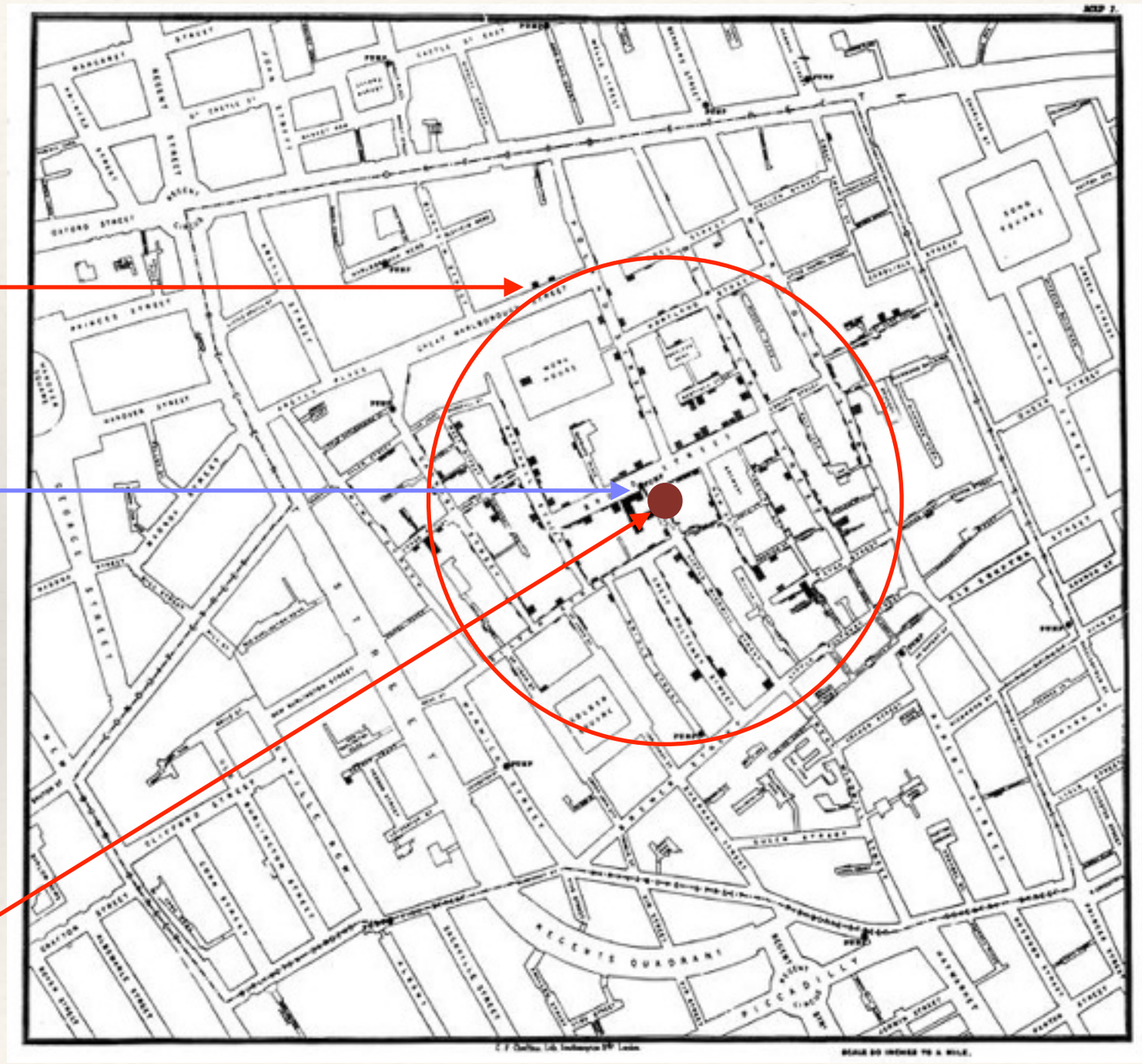
Analyze

Exploratory Data Analysis (EDA)

Cluster Region

Pump

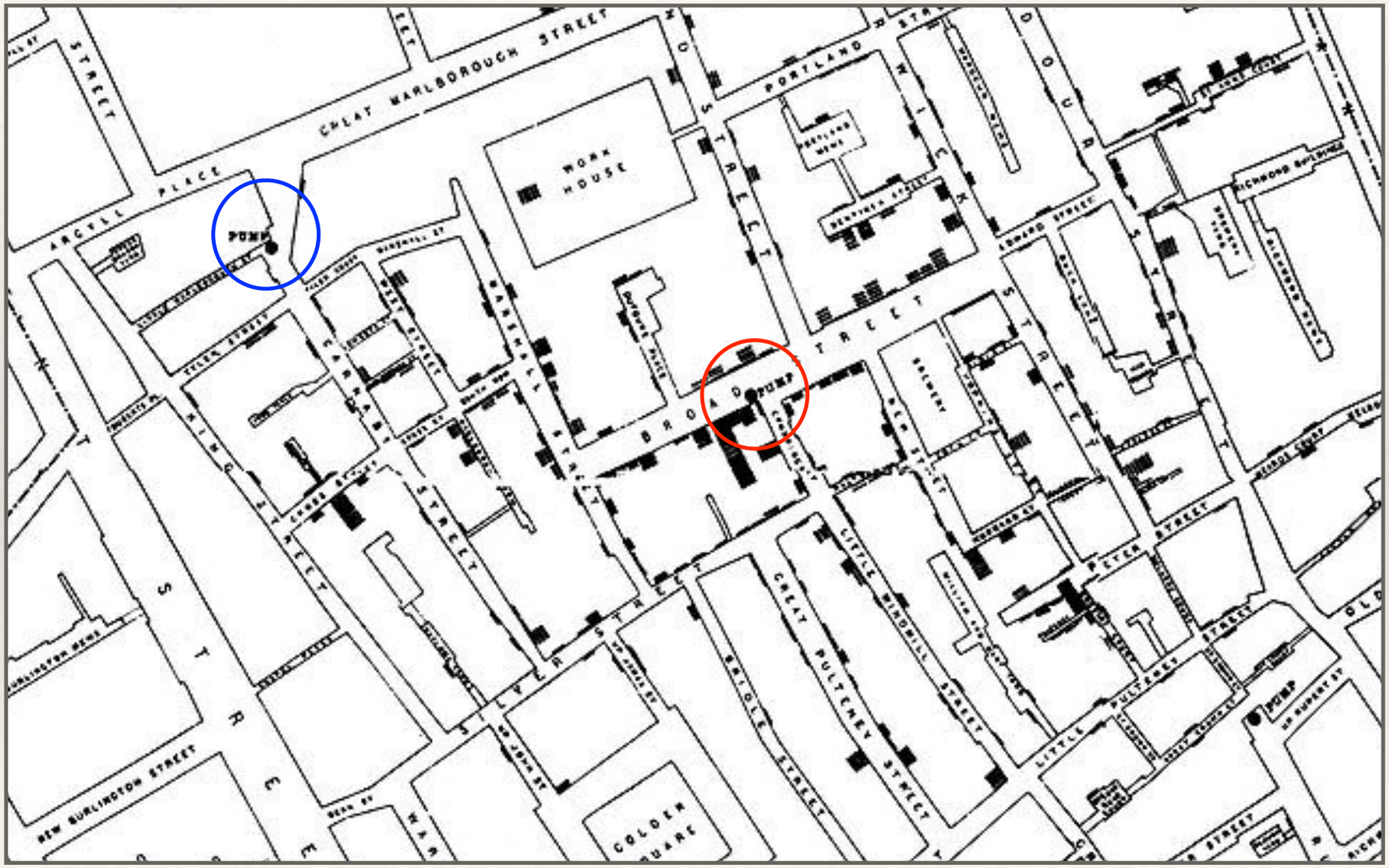
Cluster Center



John Snow, 1854

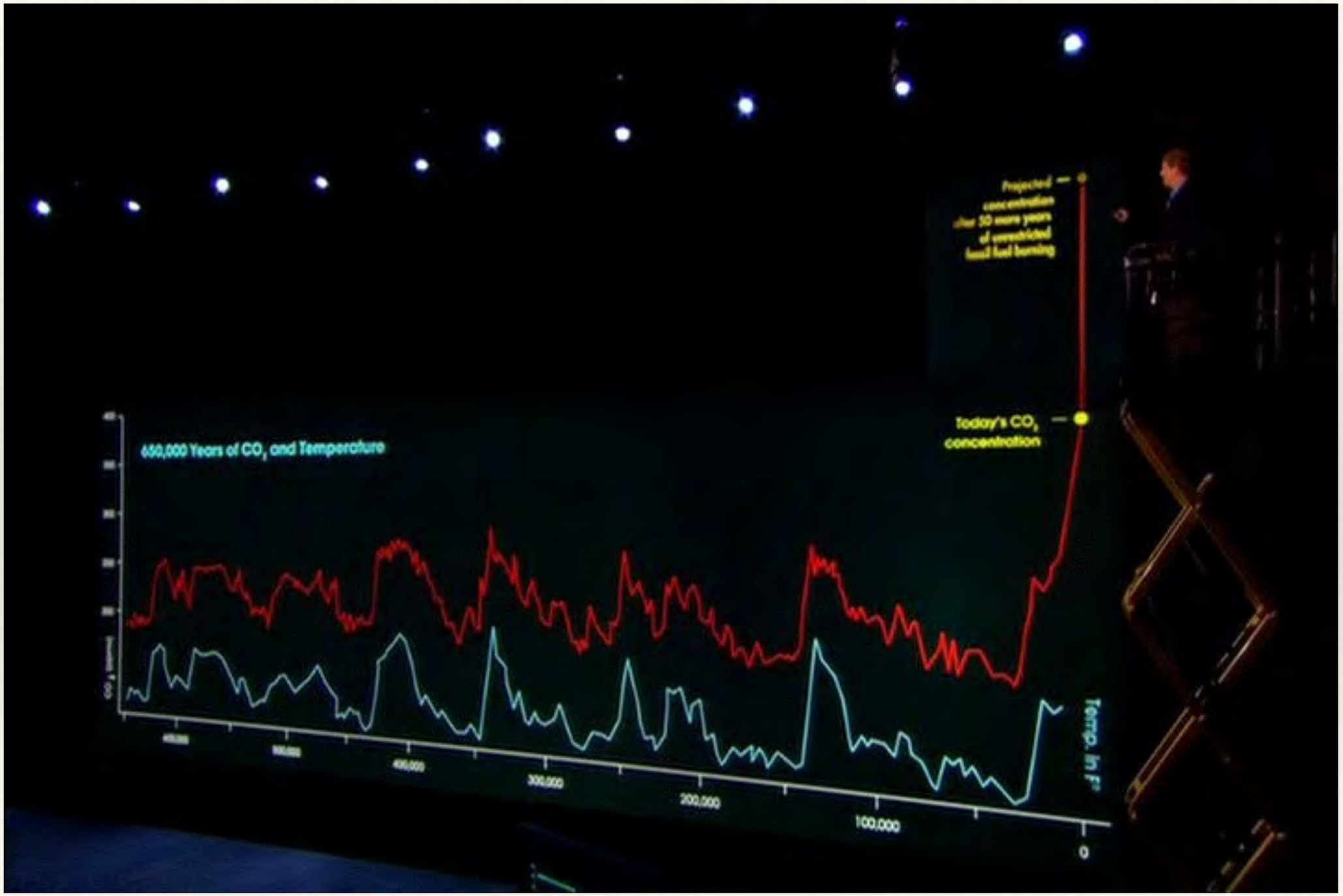
Analyze/Communicate

Confirmatory Data Analysis (CDA)



John Snow, 1854

Communicate/Convince



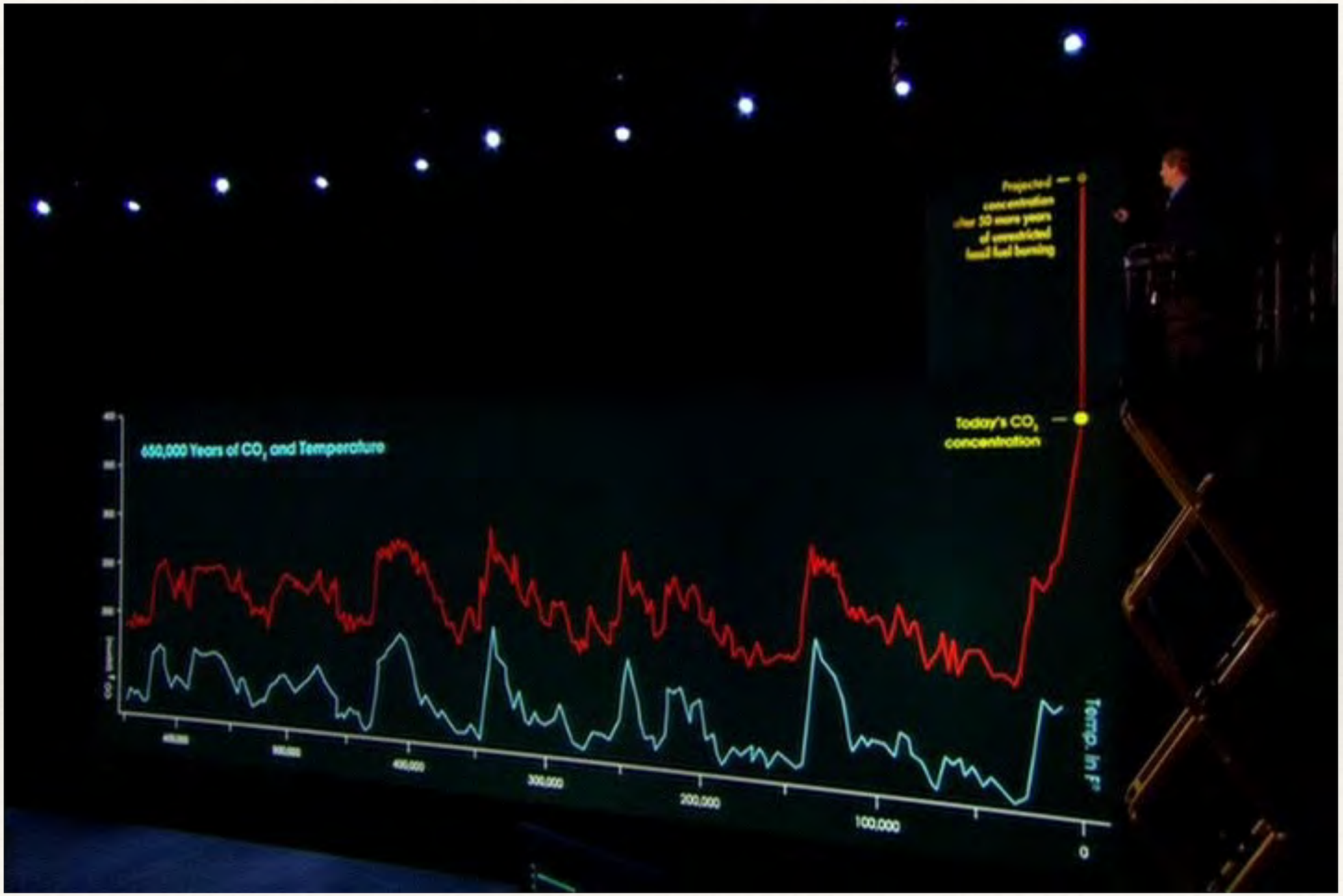
Al Gore, An Inconvenient Truth 2006

Communicate

http://www.gapminder.org/

What do you want to accomplish?

Don’t Build to Convince



Al Gore, An Inconvenient Truth 2006

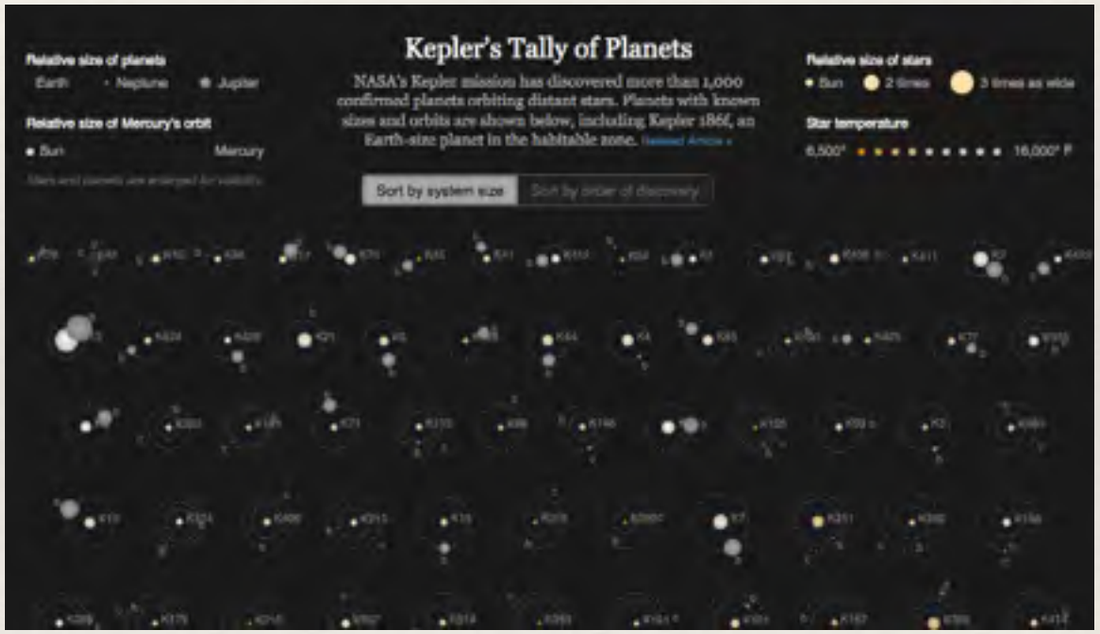
If the goal is Monitoring

Most of your visualizations

Convince

Explore

http://nyti.ms/17AErgX



http://nyti.ms/1dRTdxQ

What visual queries do you support?

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Are These Data Sets The Same?

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

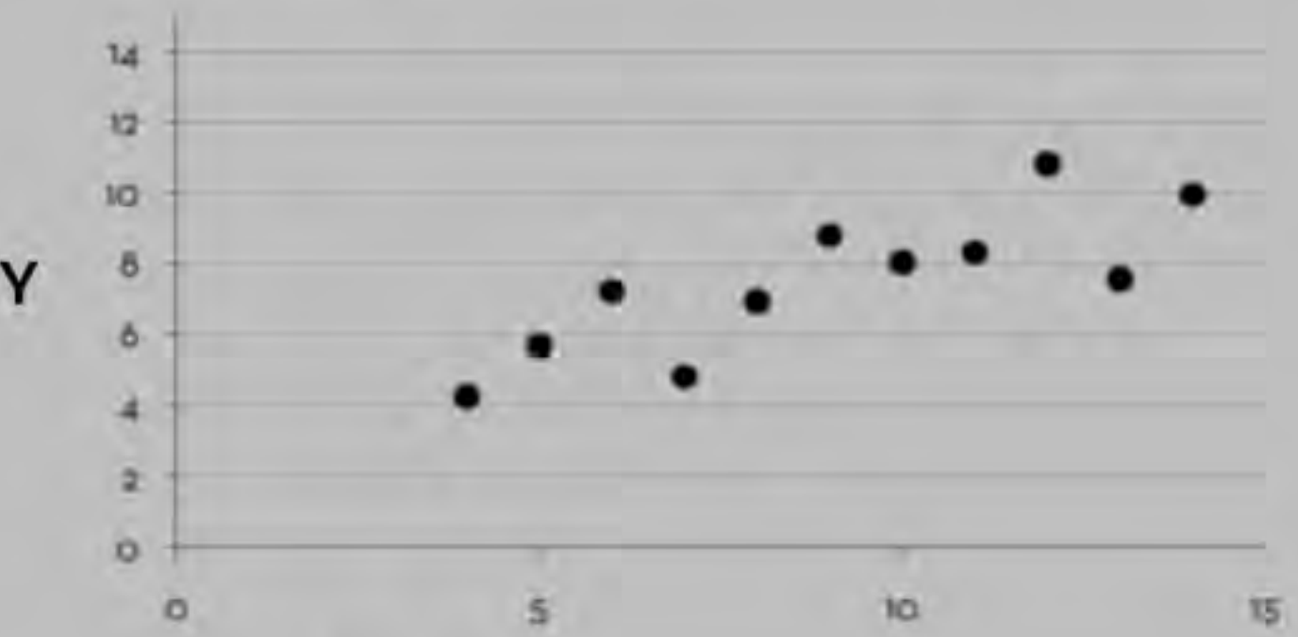
Linear Regression

$Y = 3 + 0.5\ X$

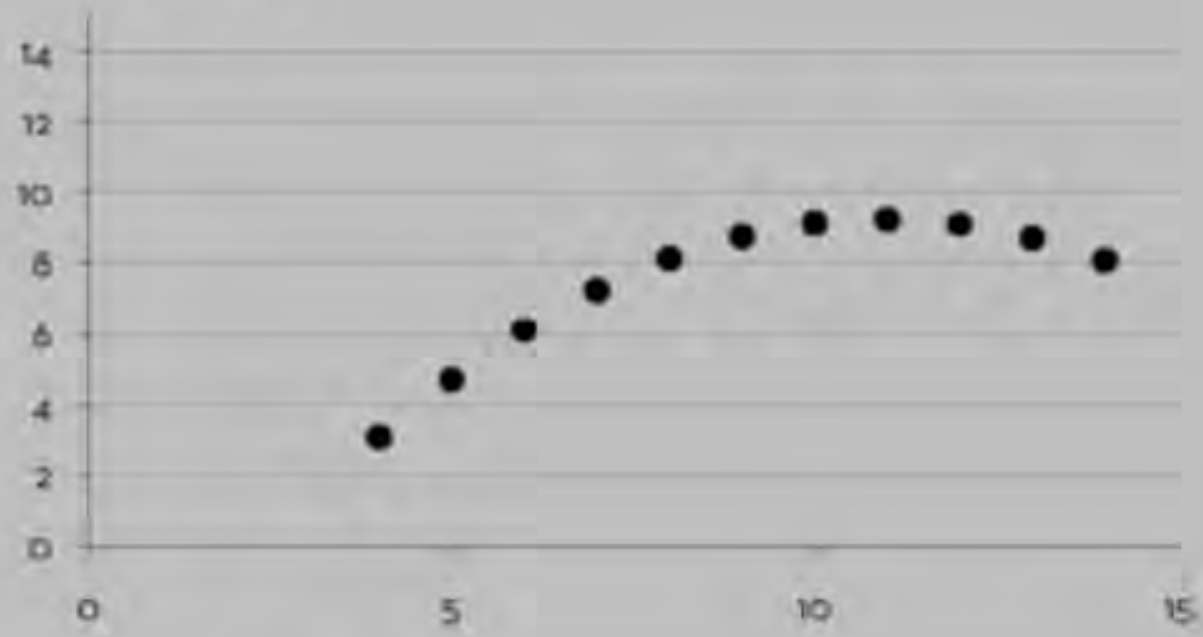
$R^2 = 0.67$

[Anscombe 73]

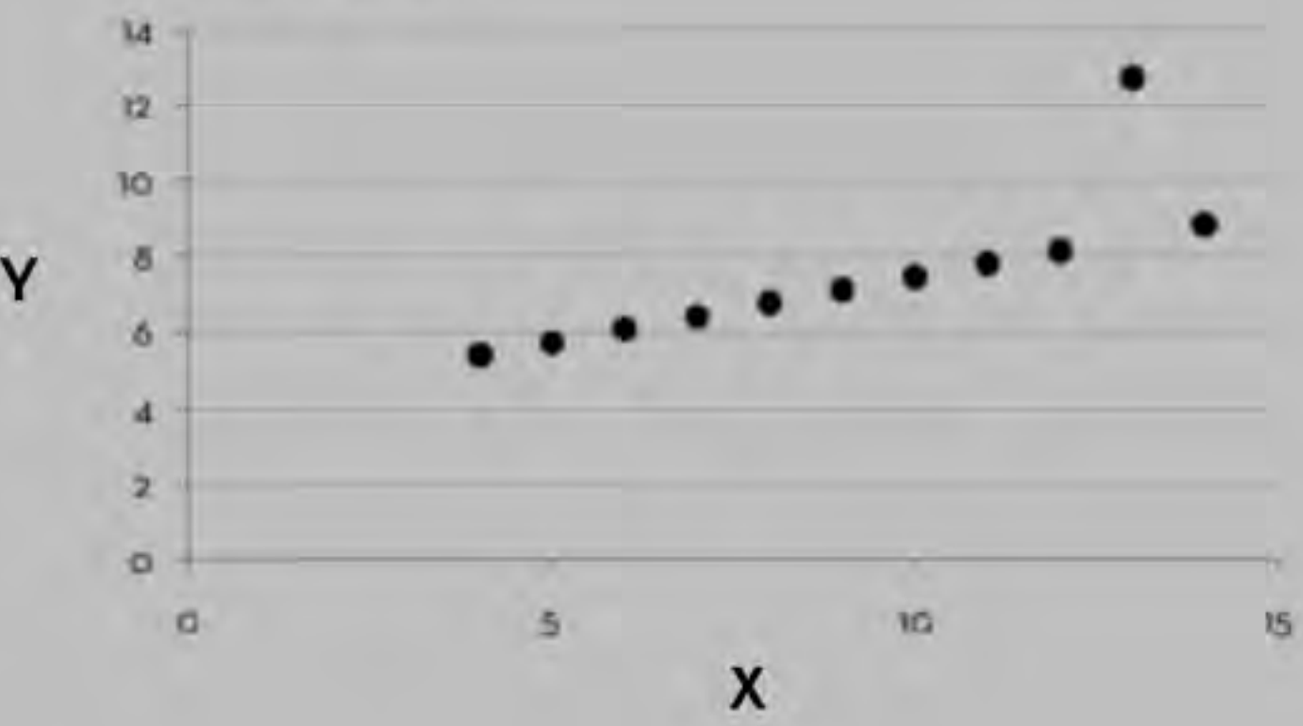
Set A



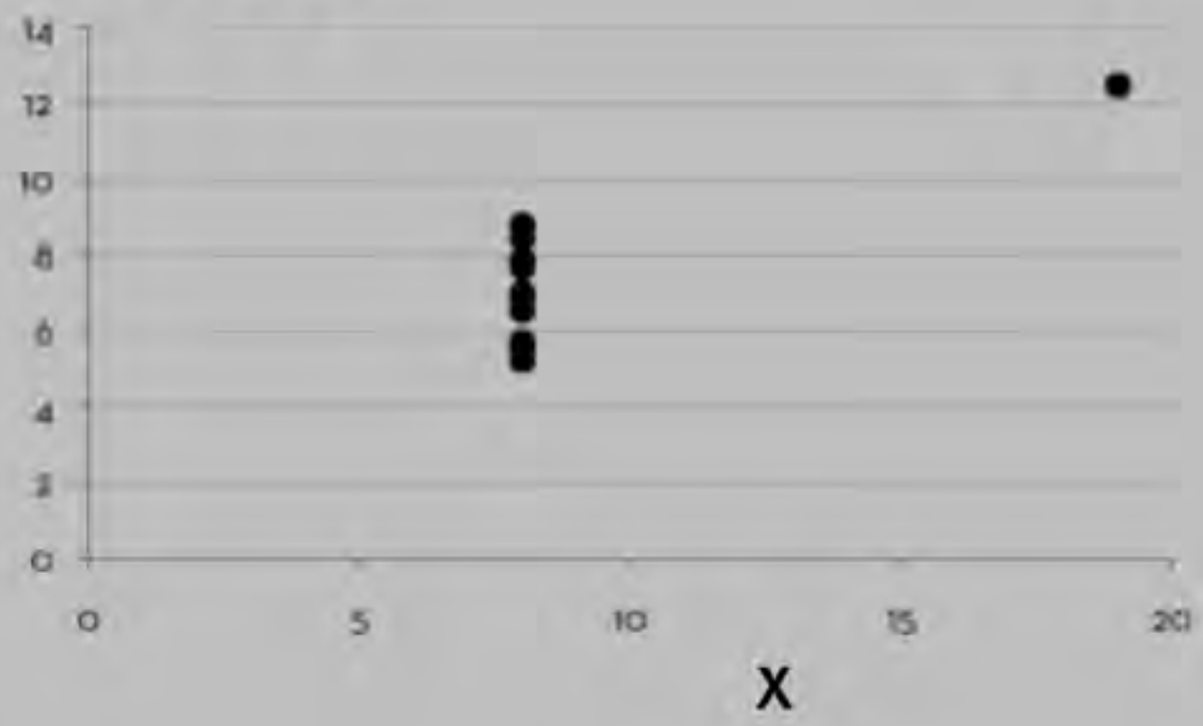
Set B



Set C



Set D

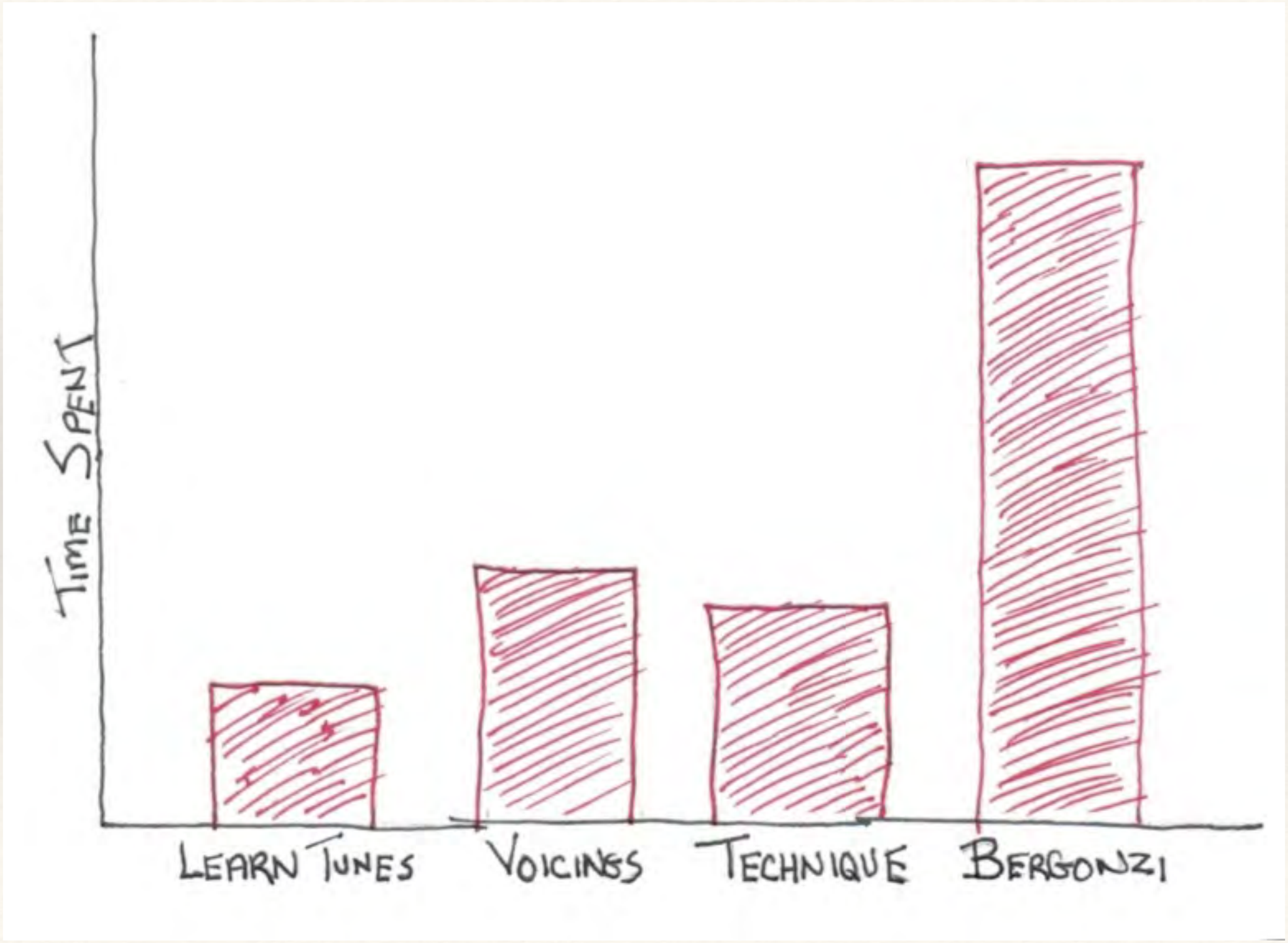


Iterate

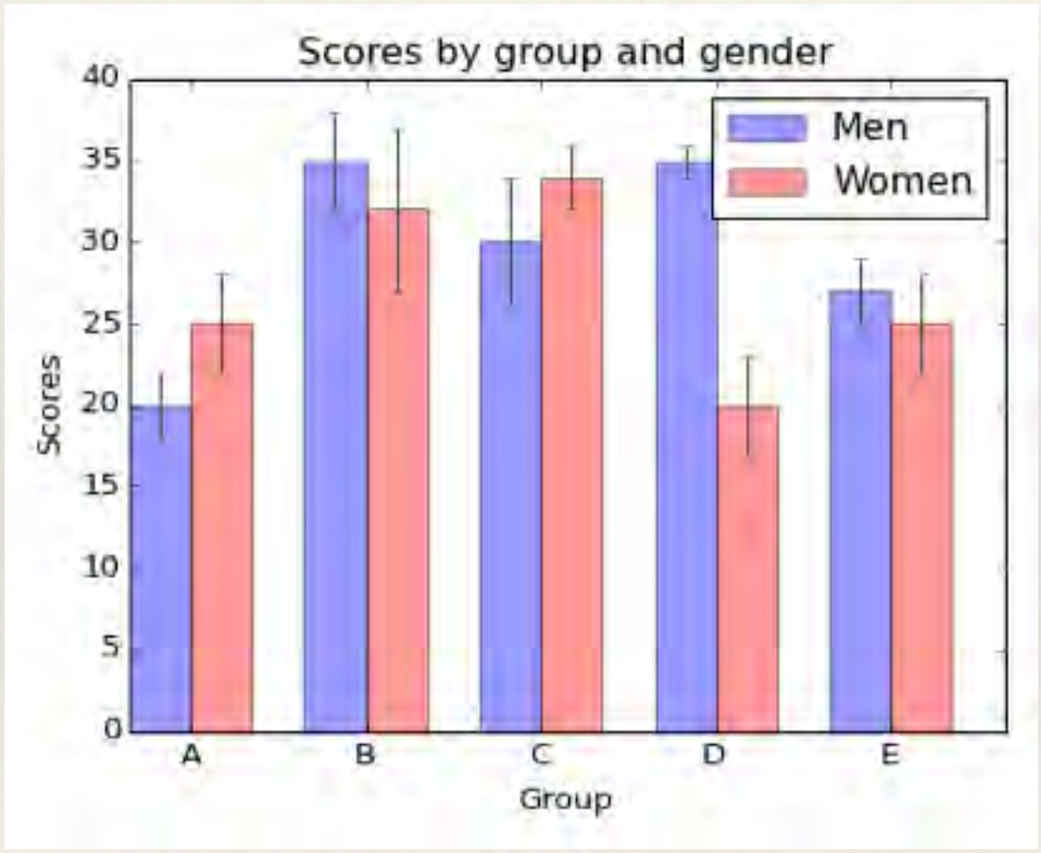
- Build many simple graphs first
  - Use Ipython/Excel/OpenOffice/Tabeau

Fully Explore Your Data First

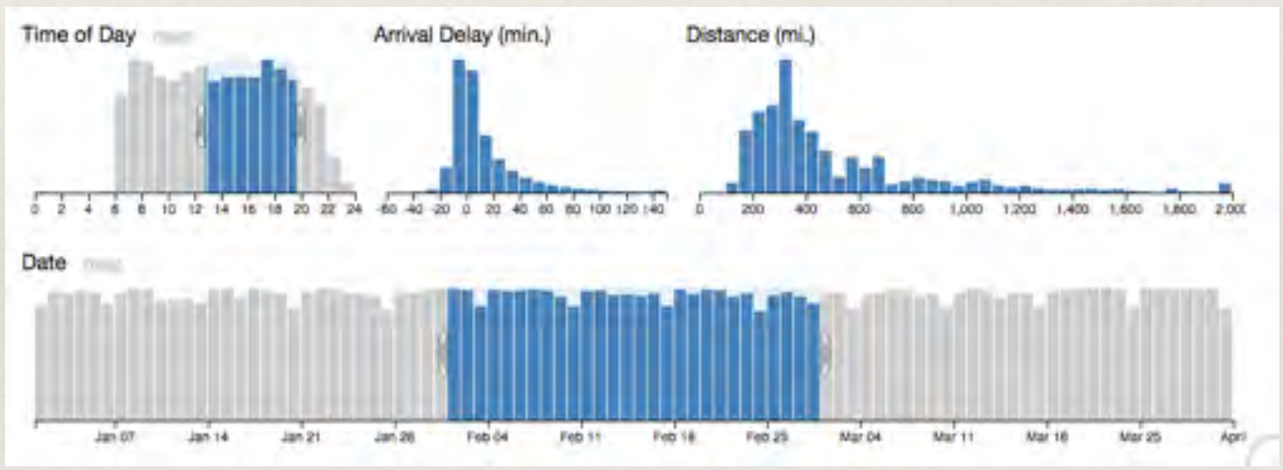
Start Design with paper and pencil/pen



Build Static BEFORE Interactive



Build these (Matplotlib)



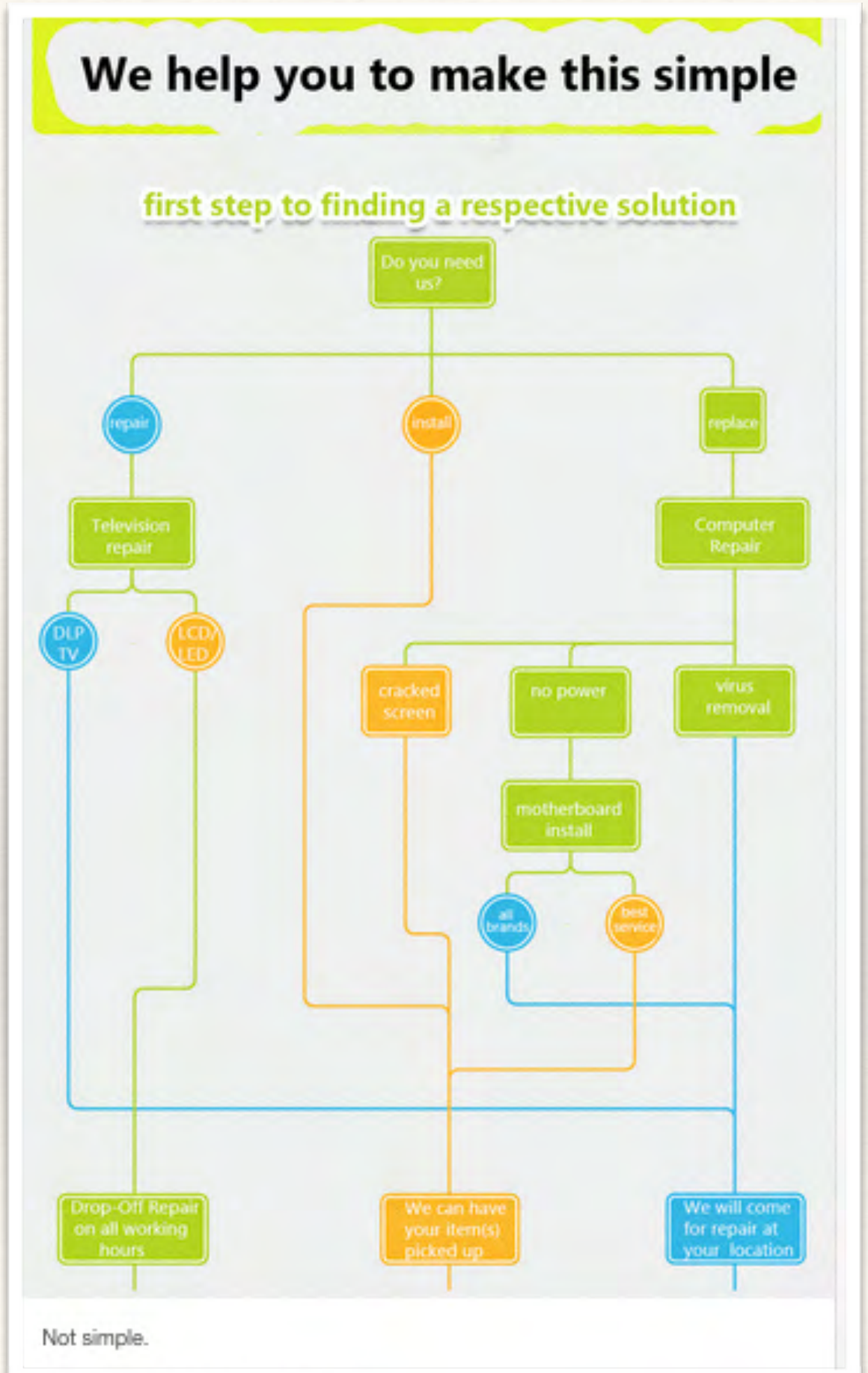
Before these (D3)

KISS principle:

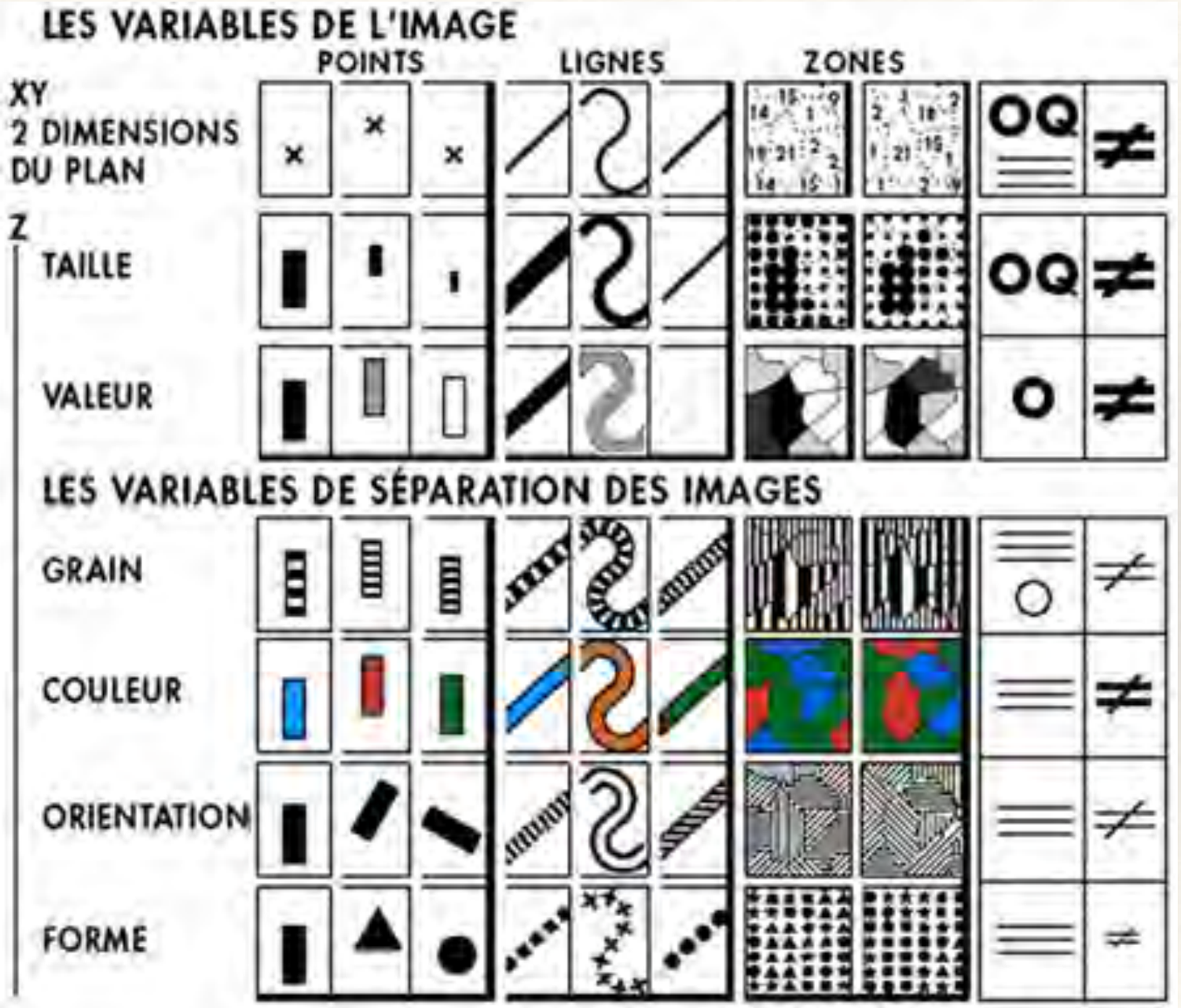
Keep It Simple Stupid

http://viz.wtf/

http://visual.ly/mad-mad-world-we-live

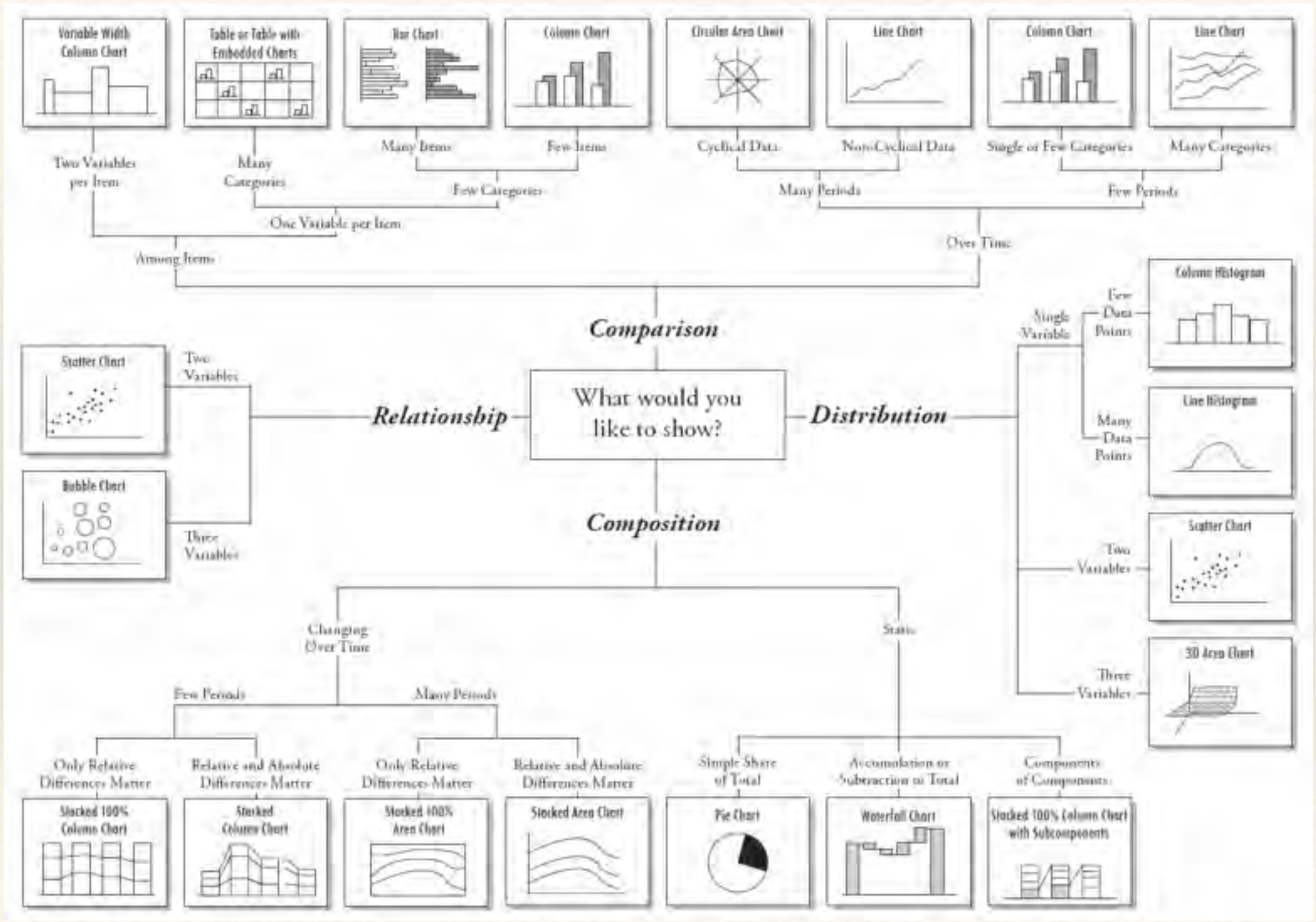


Practice Good Visual Design



Choosing The Right Tool for the Job

Choosing The Right Tool for the Job



http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Look at Good/Bad Visualizations

- Good Examples:
  - http://flowingdata.com/
  - http://flowingdata.com/2012/04/27/data-and-visualization-blogs-worth-following/
- Bad Examples:
  - http://wtfviz.net/
  - http://junkcharts.typepad.com/junk_charts

Practice

Data Sources

* Fivethirtyeight Data
* Quandl
* Datamob
* Reddit Datasets Lists
* Datahub
* Factual
* Census.gov
* Data.gov
* Dataverse Network
* Infochimps
* Linked Data
* Data Market
* Reddit Open Data
* Climate Data Sources
* Climate Station Records
* CDC Data
* World Bank Catalog
* StateMaster
* Socrata
* The UN
* Weatherbase
* ESPN
* Datamarket
* Google Public Data
* Million Song Database
* Hillary Mason's aggregation of Dataset links
* NASDAQ Data Store
* KDNuggets links
* Amazon Public Datasets
* Data NYC
* Freebase
* DBpedia
* Enigma
* Reuters Corpora
* World bank Data
* International Monitory Fund Data

Libre Office

- Load a spreadsheet with data
- Make a time series line plot

Python/Pandas/Matplotlib/Ipython

- Load a time series of data
- Make a line plot

Data Transformations

Can the data be visualized as-is?

- 1885 Hight data from Francis Galton on 928 (adult) children

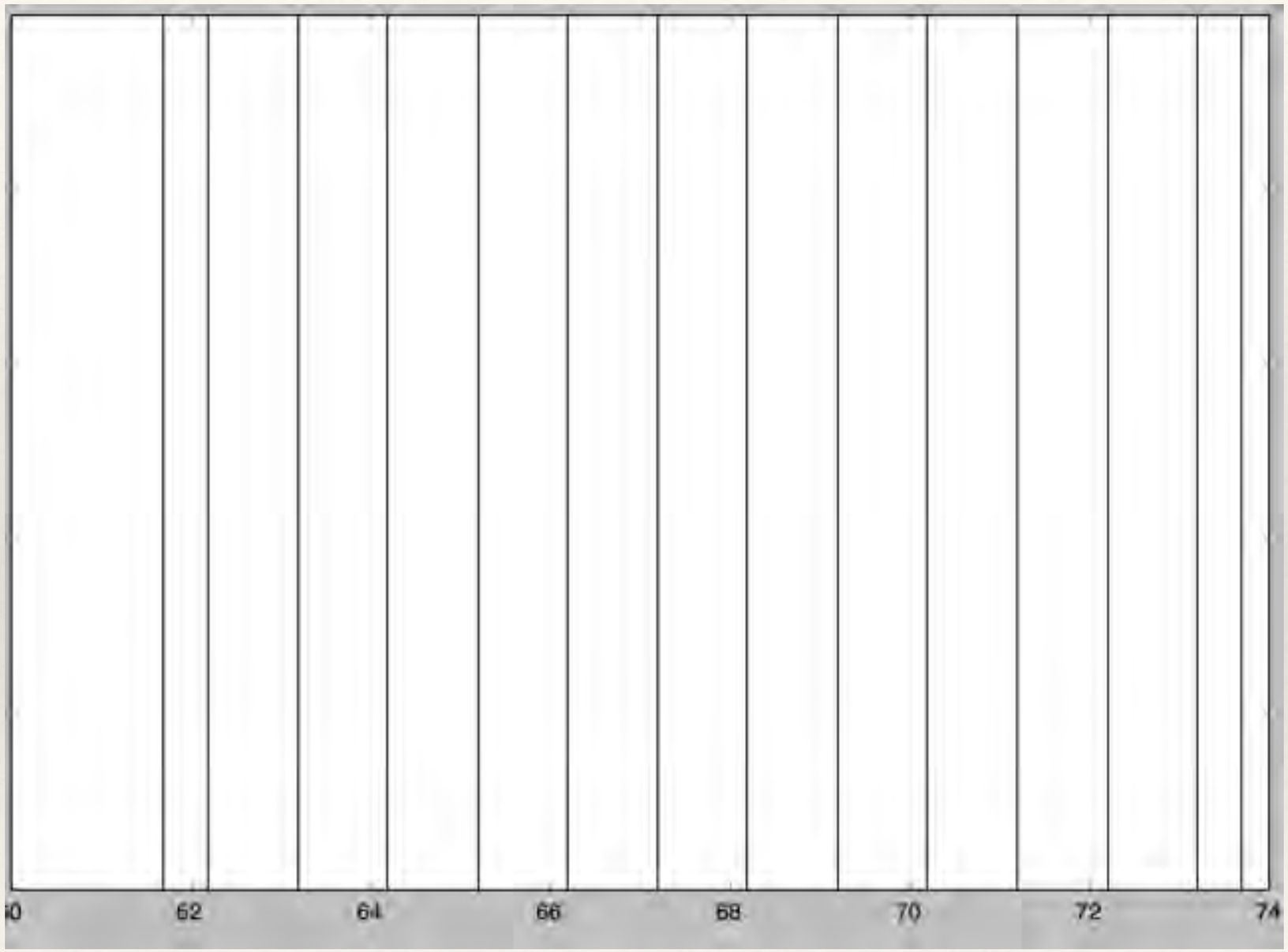
Yum Data

61.7, 61.7, 61.7, 61.7, 61.7, 62.2, 62.2, 62.2, 62.2, 62.2, 62.2, 62.2, 63.2, 64.2, 65.2, 66.2, 67.2, 68.2, 69.2, 70.2, 71.2, 72.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.2, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7, 73.7

Now what??

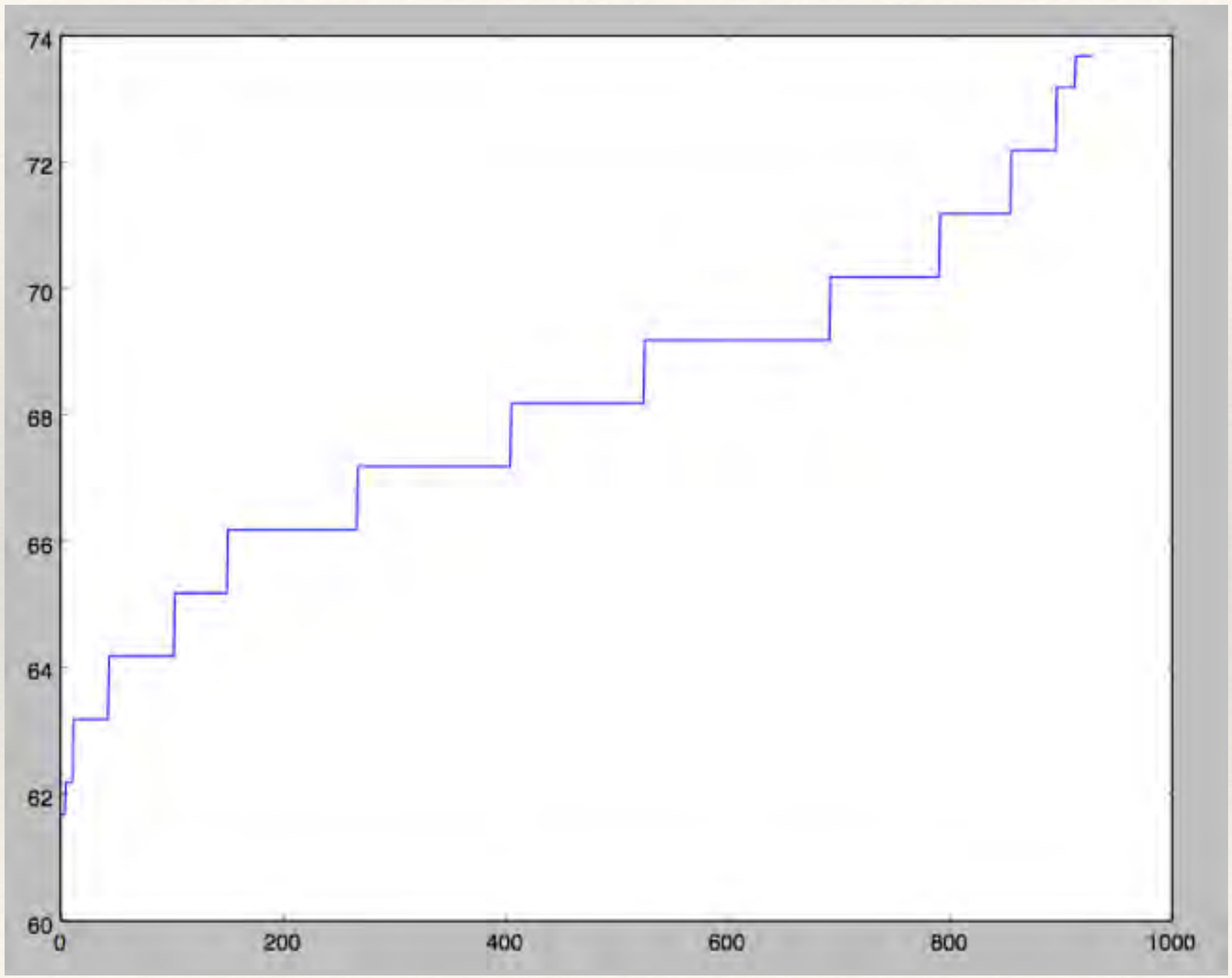
Vertical lines (1D data)

Not too illuminating

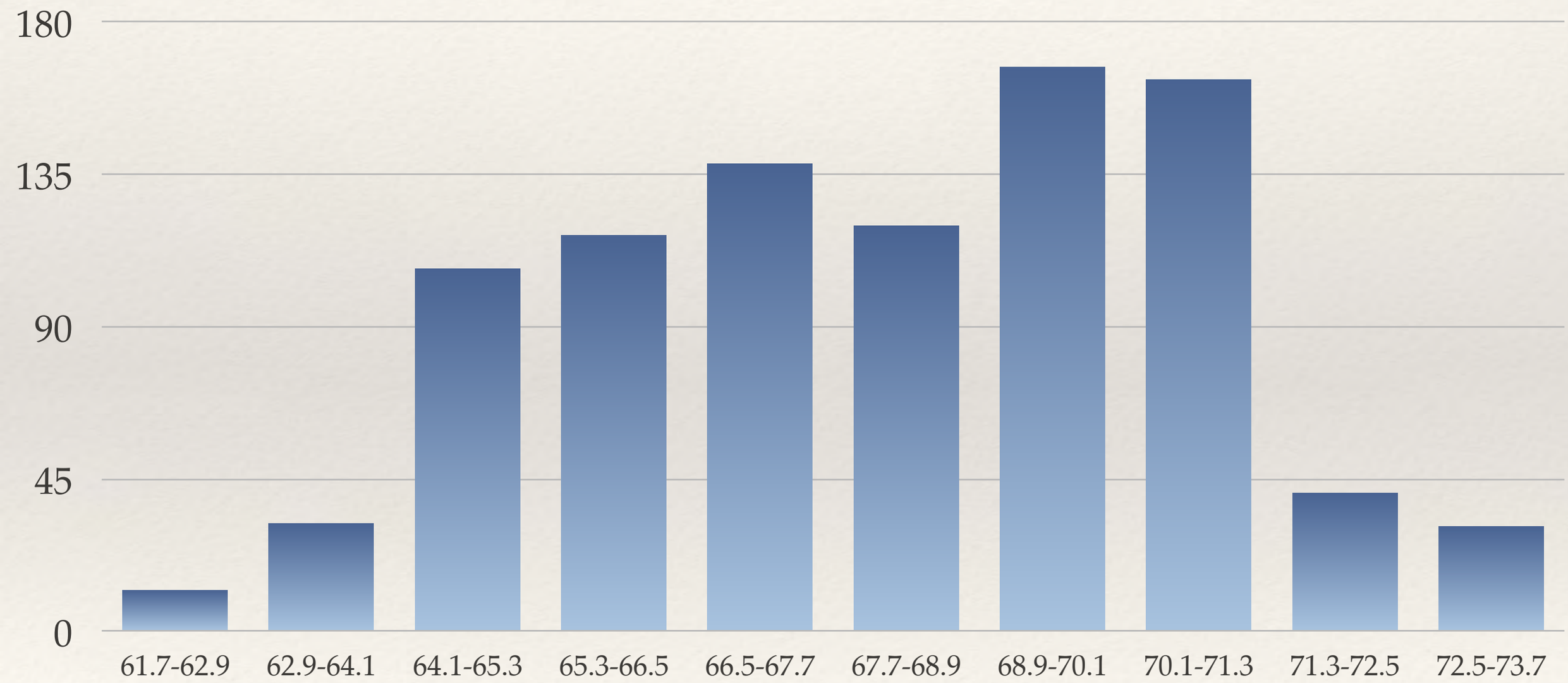


Height

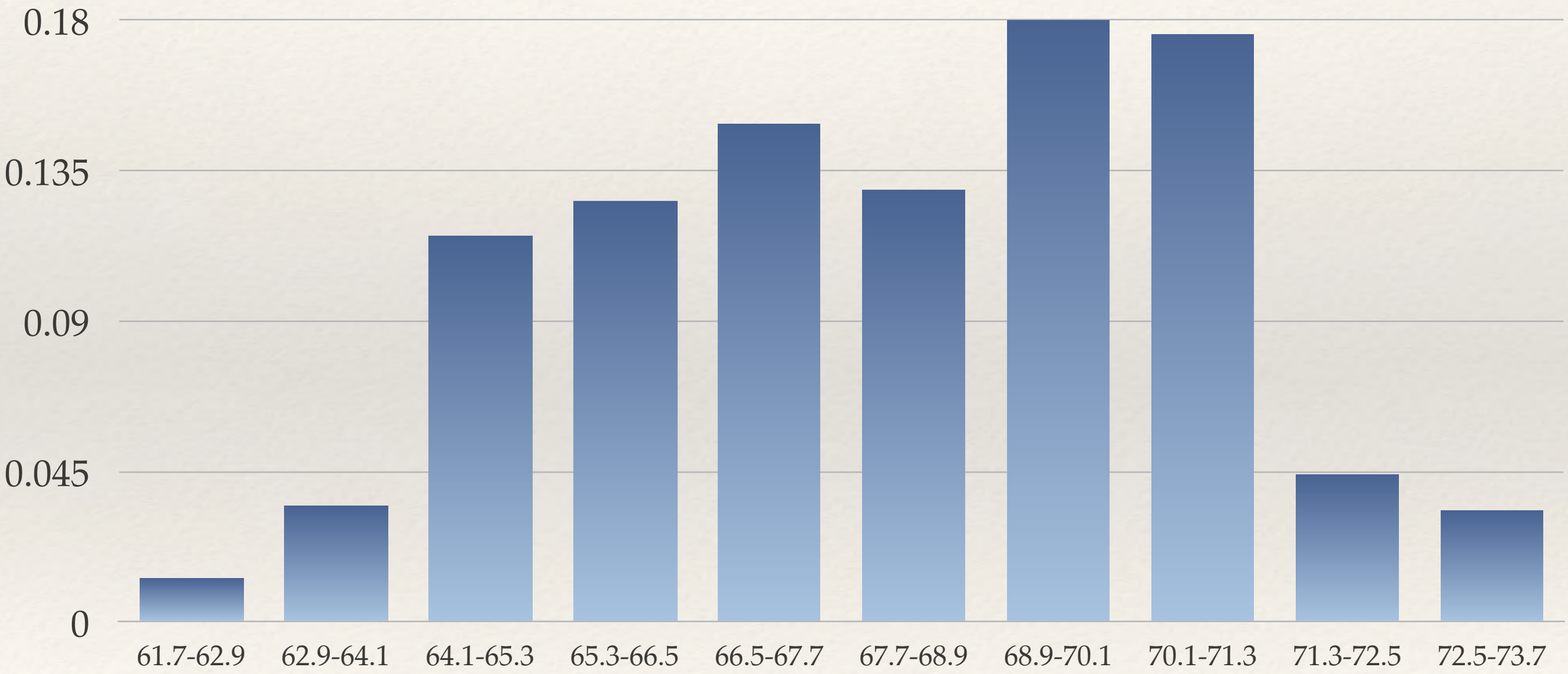
Sort and Plot



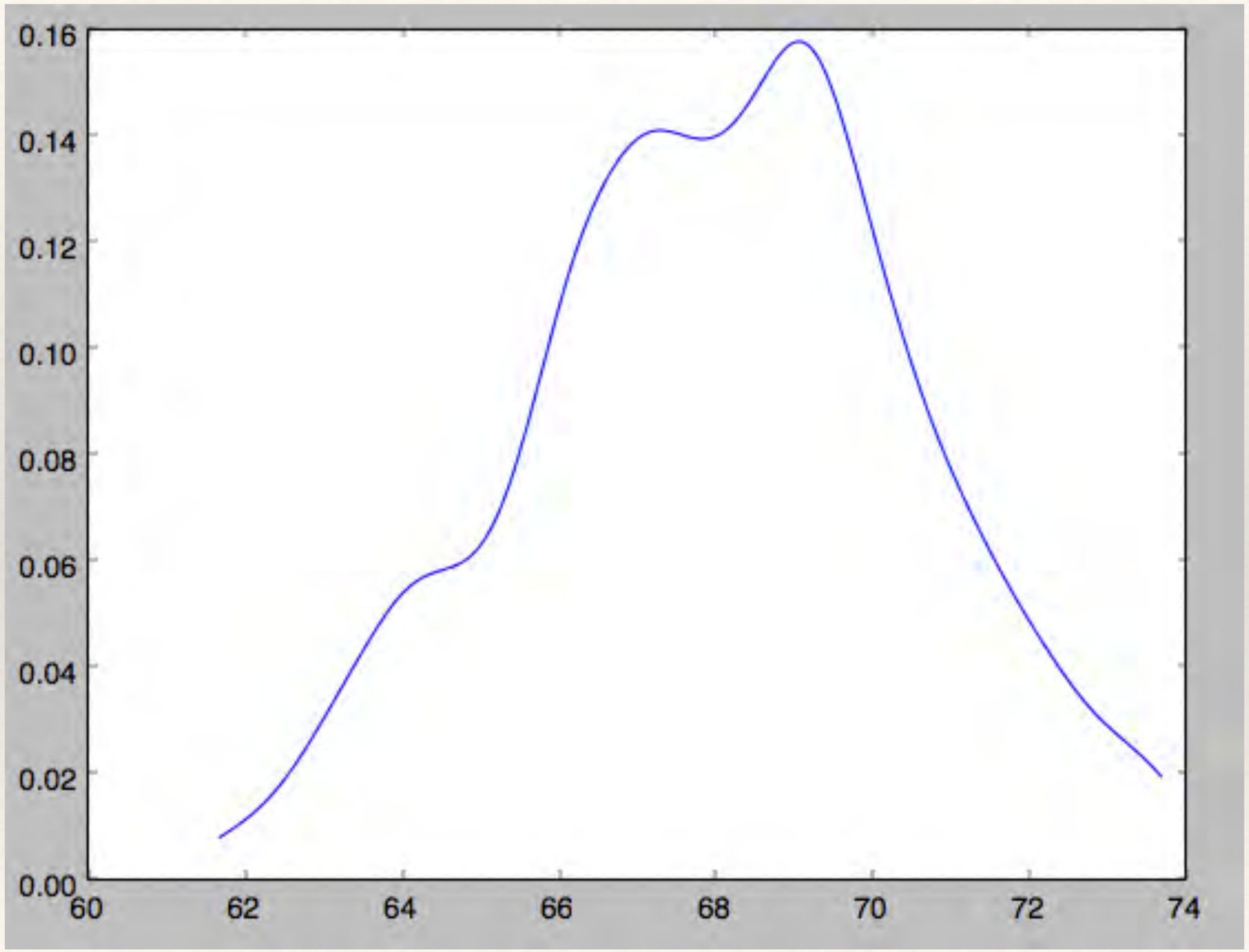
Distribution (histogram)



Probability



Probability using Using KDE

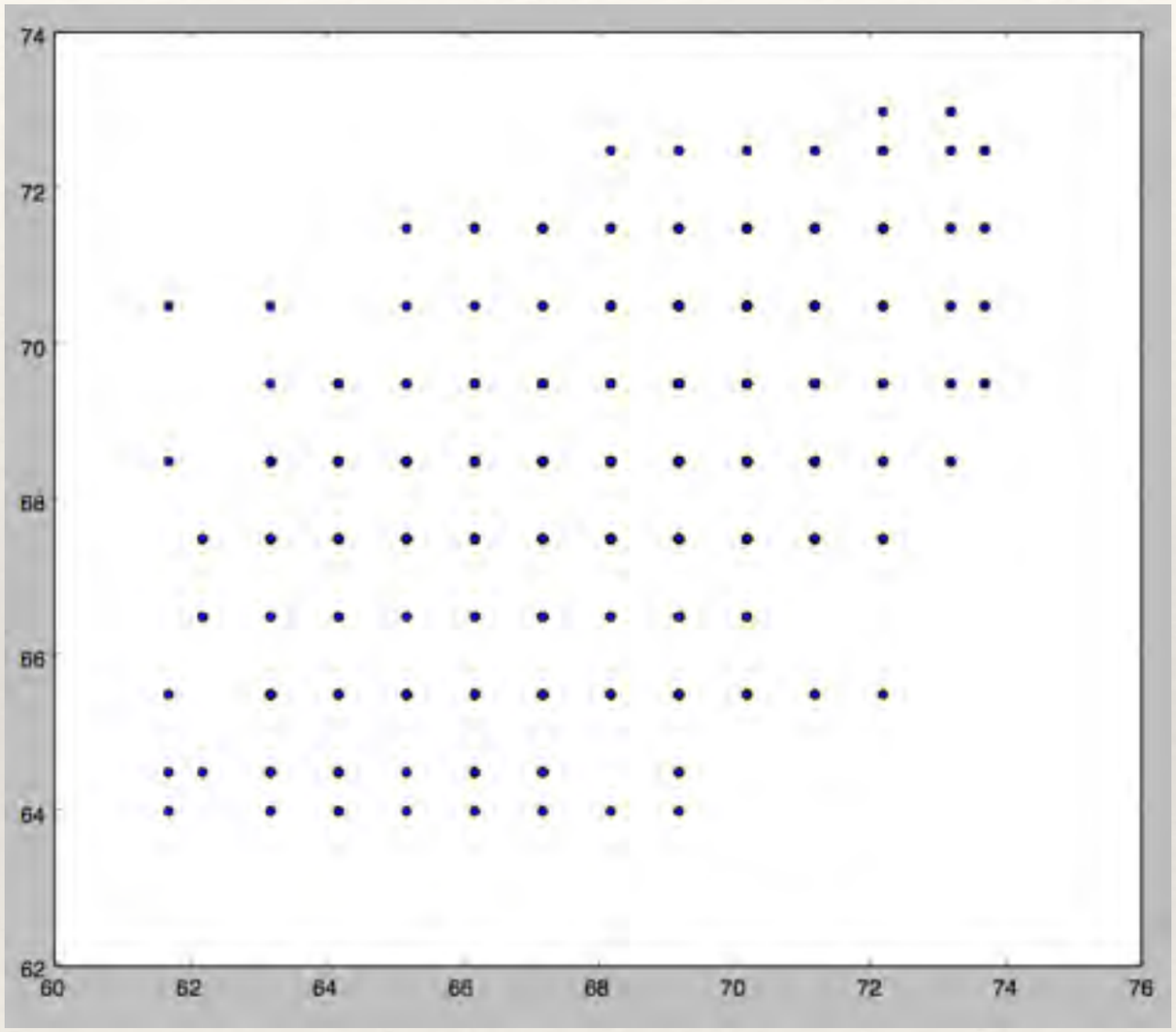


Galton Data also has "midparent" height.

- Mid-parent height = mean(father hight, 1.08* mother height)

How do we show relationship?

Scatter Plot

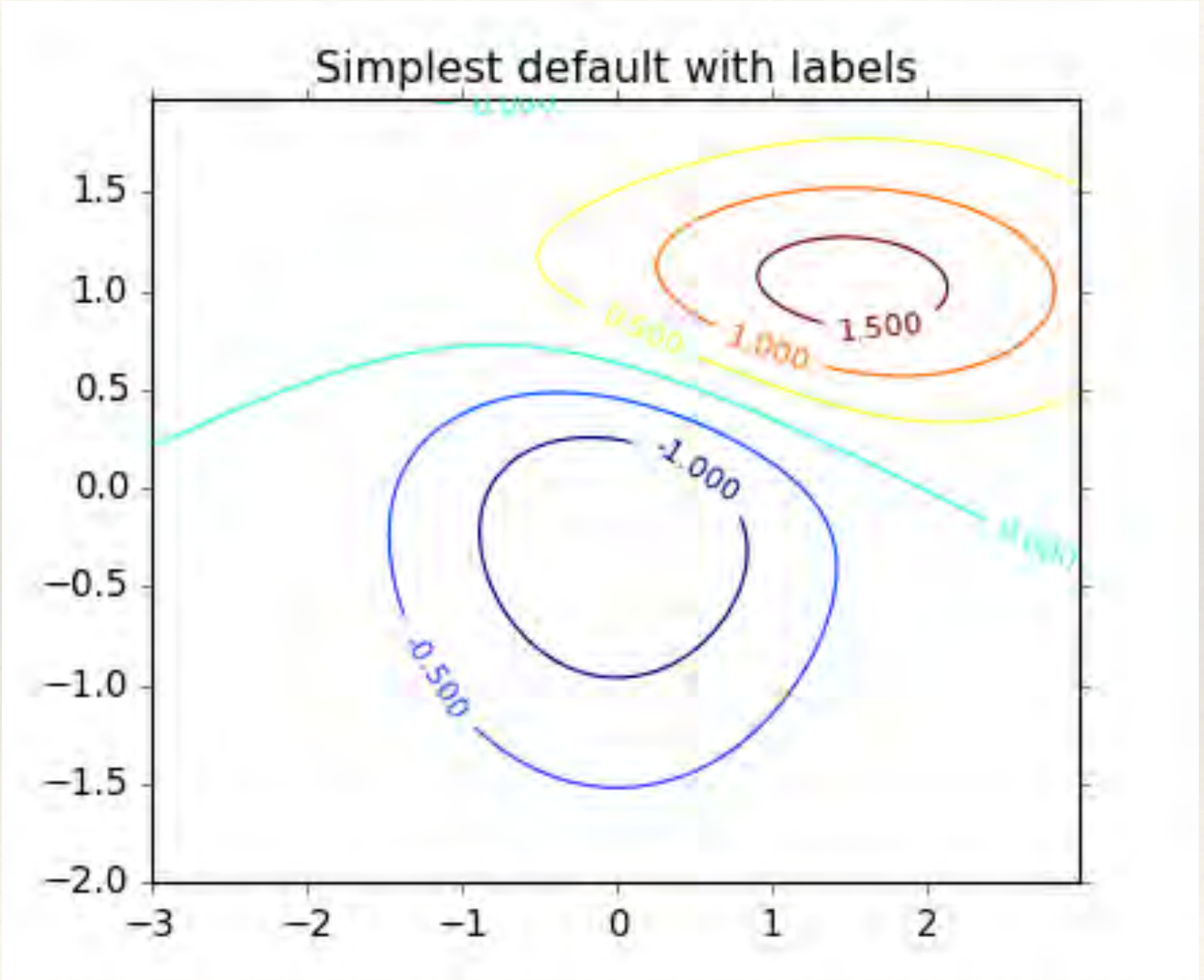


Uggh! Data heavily quantized. Blah.

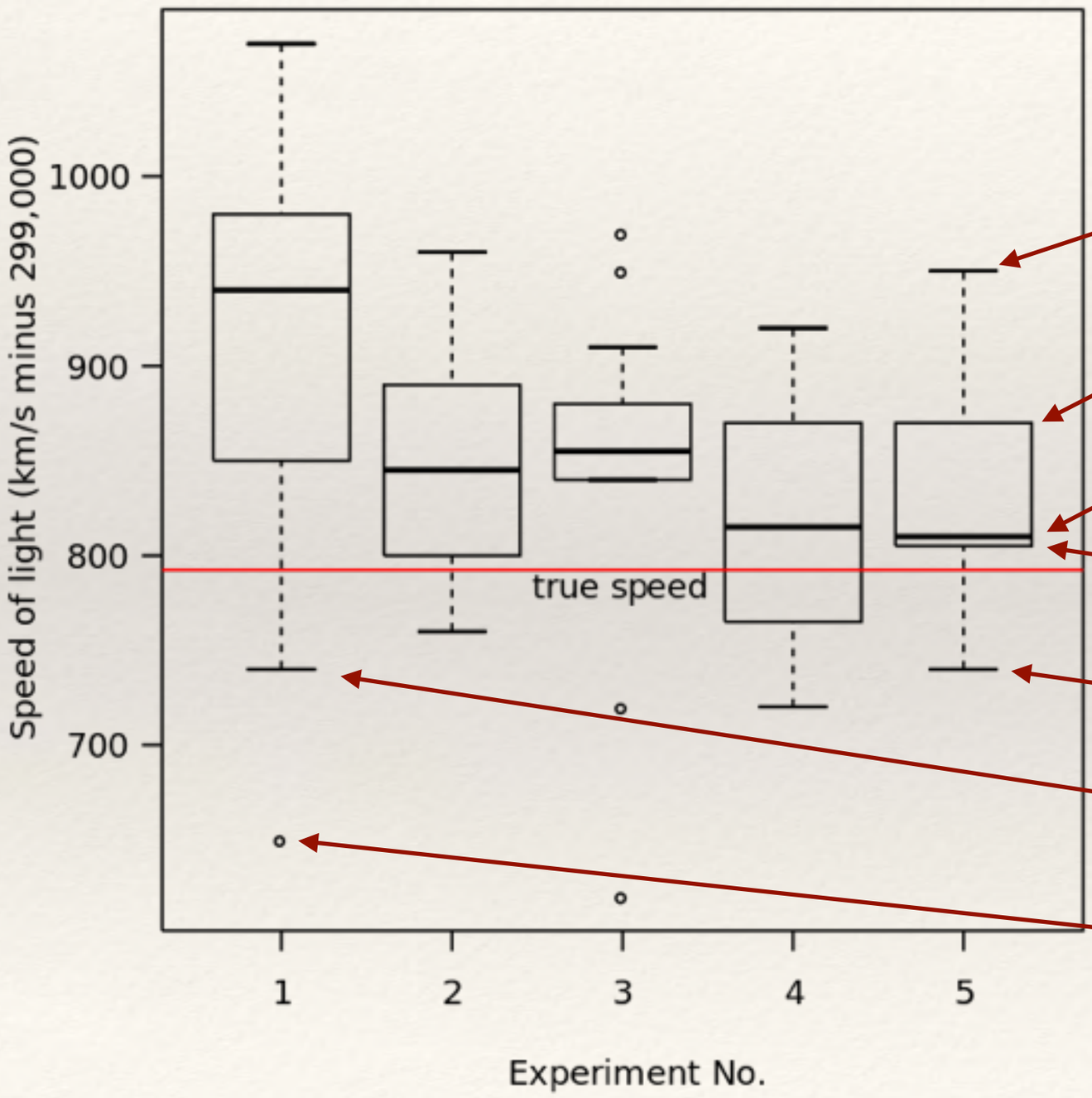
KDE also possible

- We can do a Kernel Density estimator to find surface

Contour Plot (2D)



Box and Whiskers



Max

First Quartile

Median (Second Quartile)

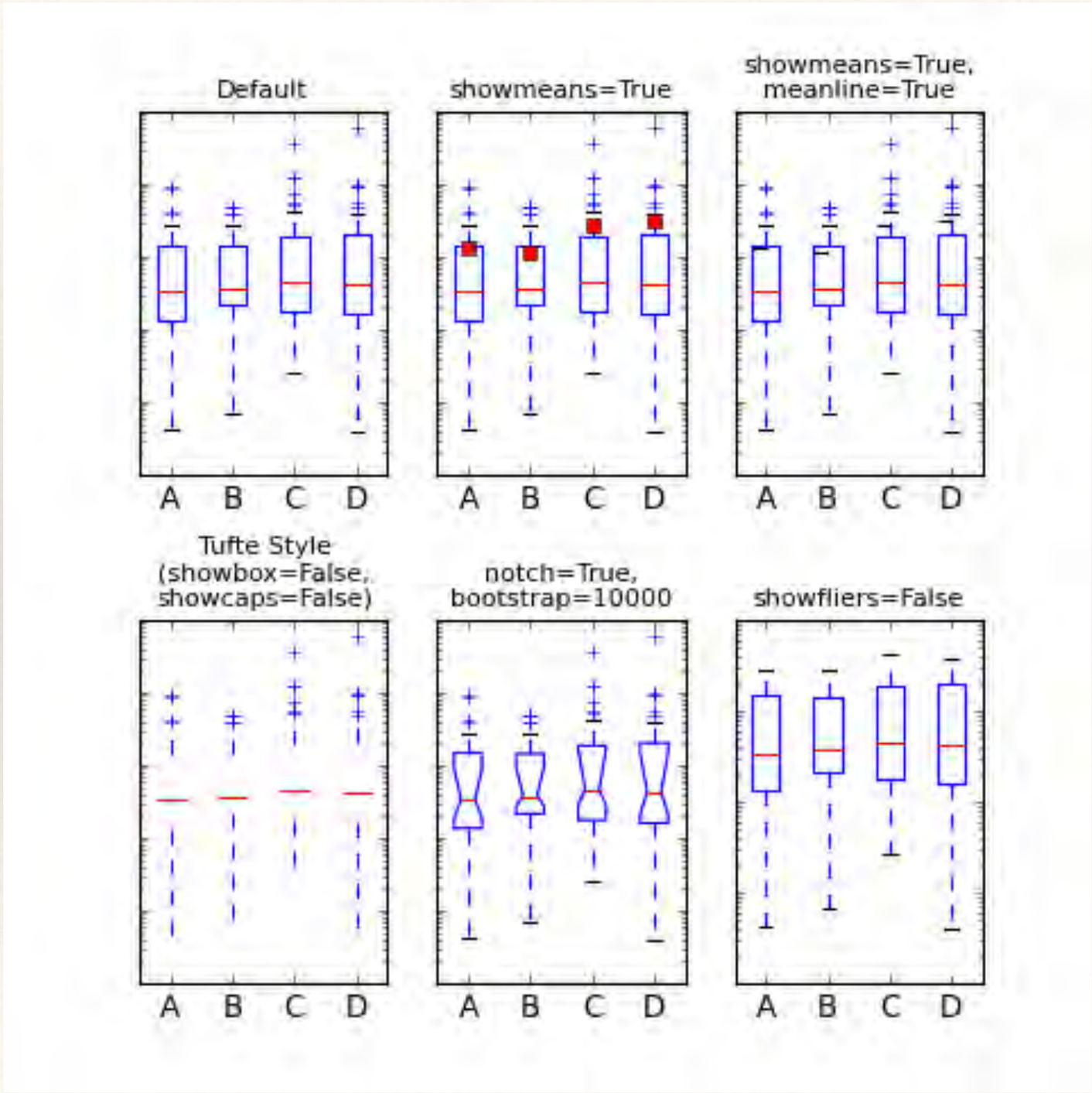
Third Quartile

Min

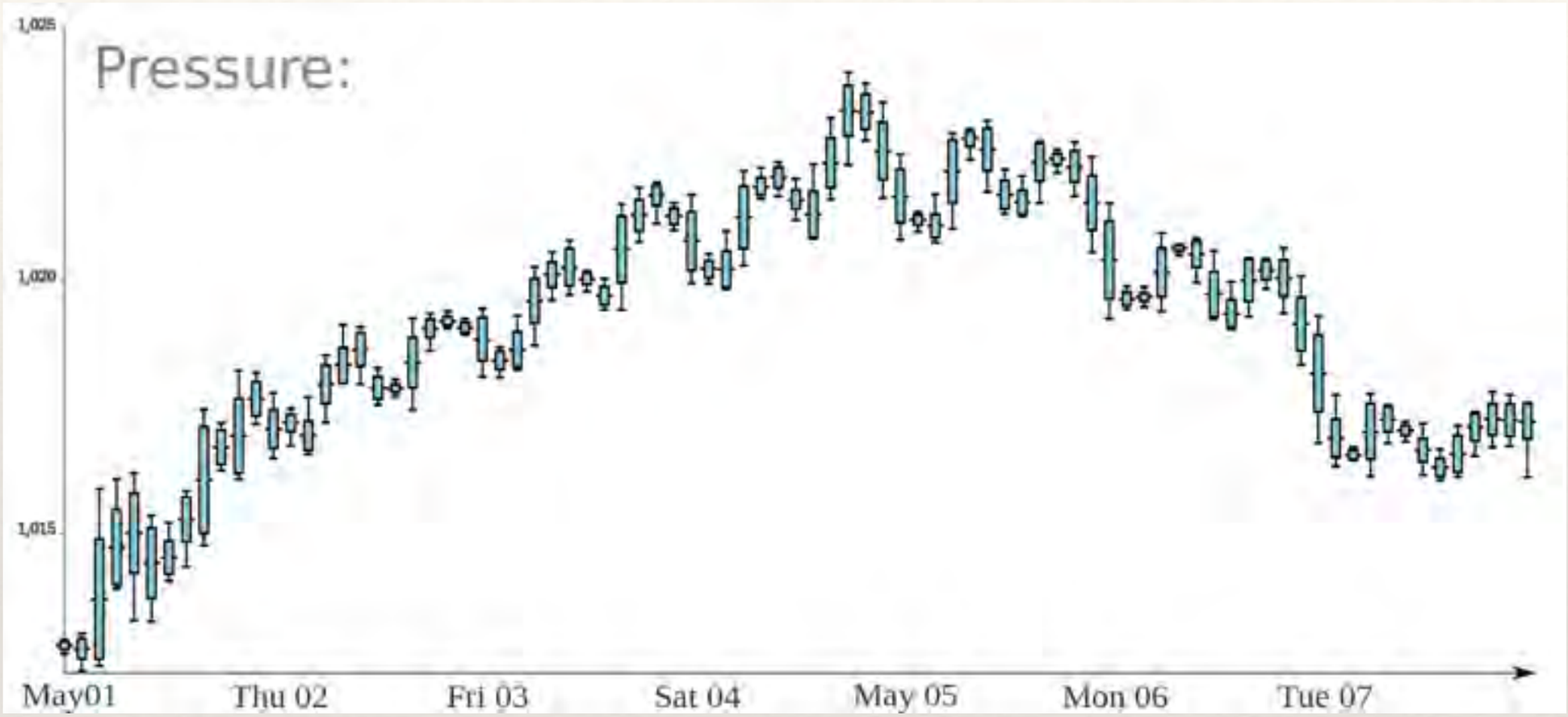
lowest datum still within 1.5 IQR

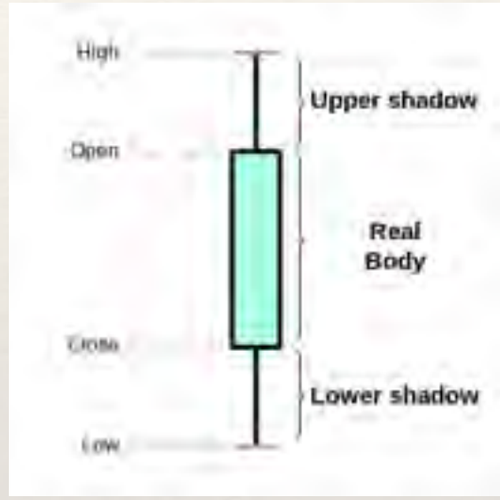
outlier

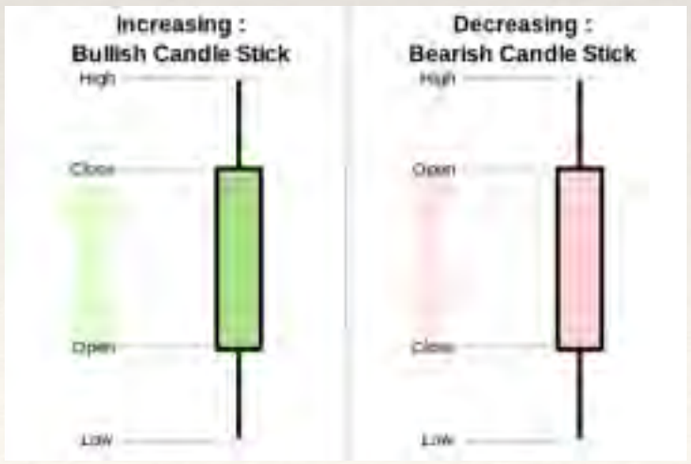
Matplotlib Boxplot



Candle Stick Chart







Finance

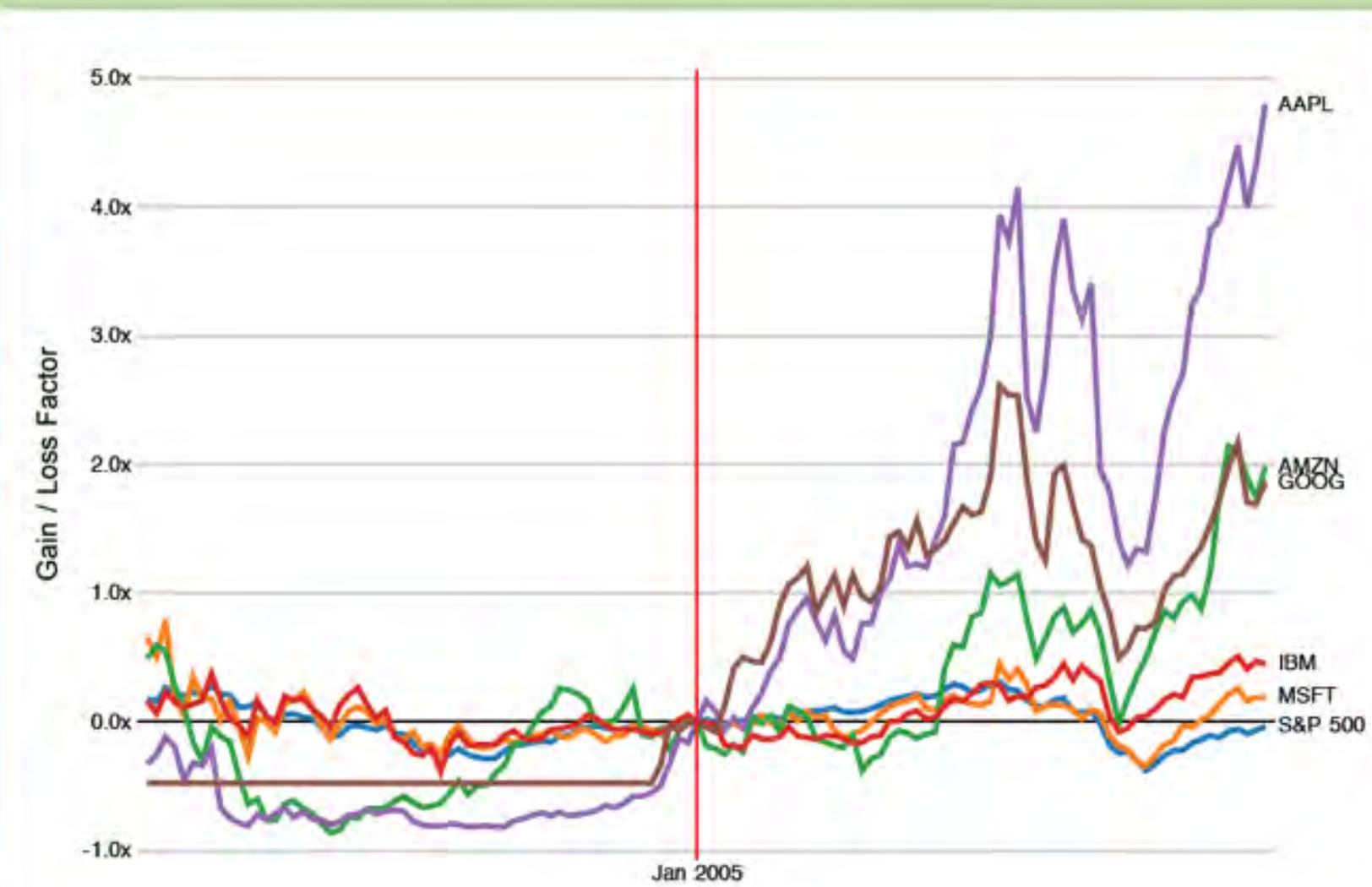
Moving Averages Smoothing

Visualization Zoo (Heer,Bostock, et al)

Time Series: Index Charts

FIGURE 1A

Index Chart of Selected Technology Stocks, 2000-2010

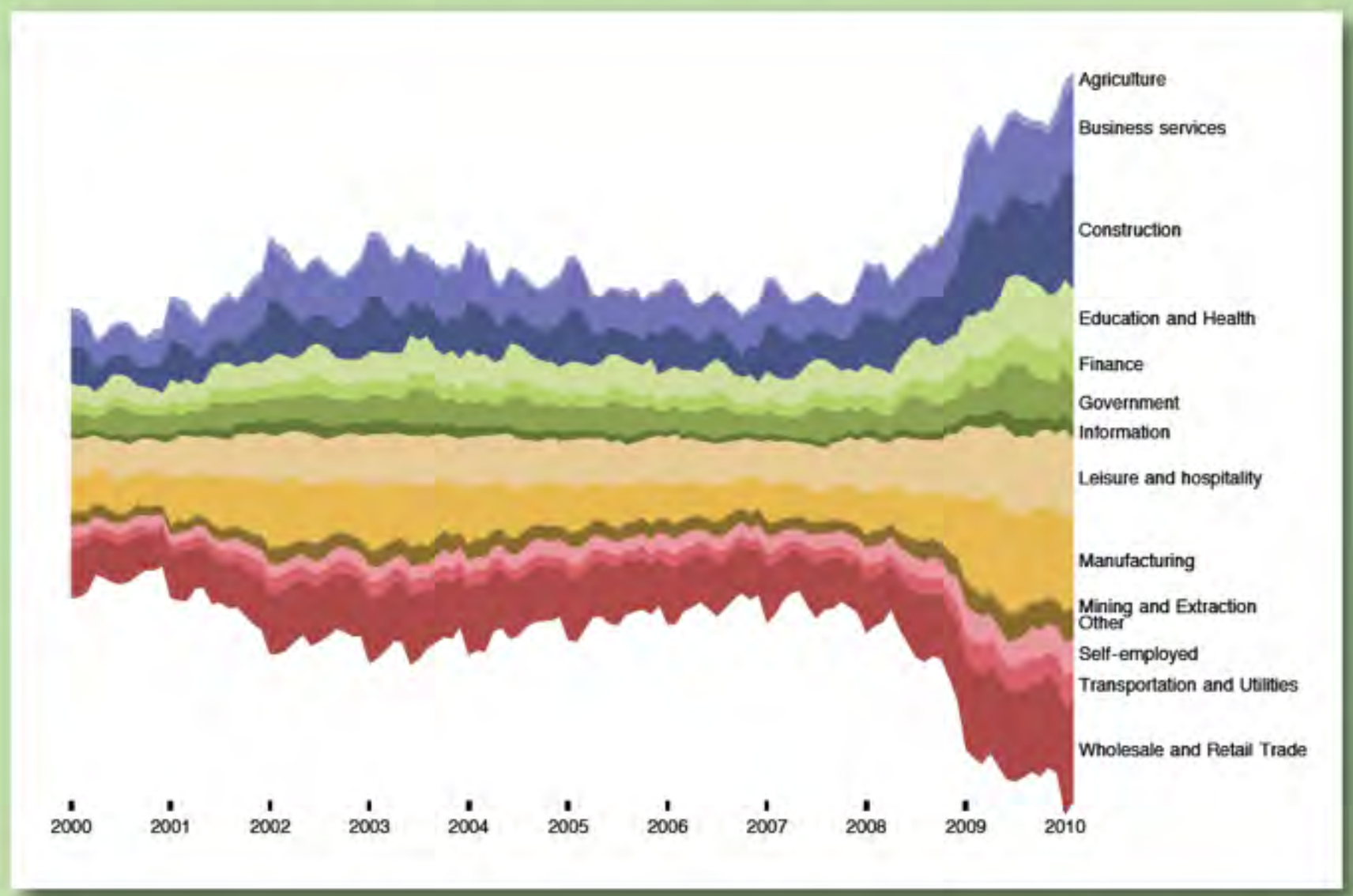


Source: Yahoo! Finance; http://hci.stanford.edu/jheer/files/zoo/ex/time/index-chart.html

Time Series: Stacked Graph

FIGURE 1B

Stacked Graph of Unemployed U.S. Workers by Industry, 2000-2010

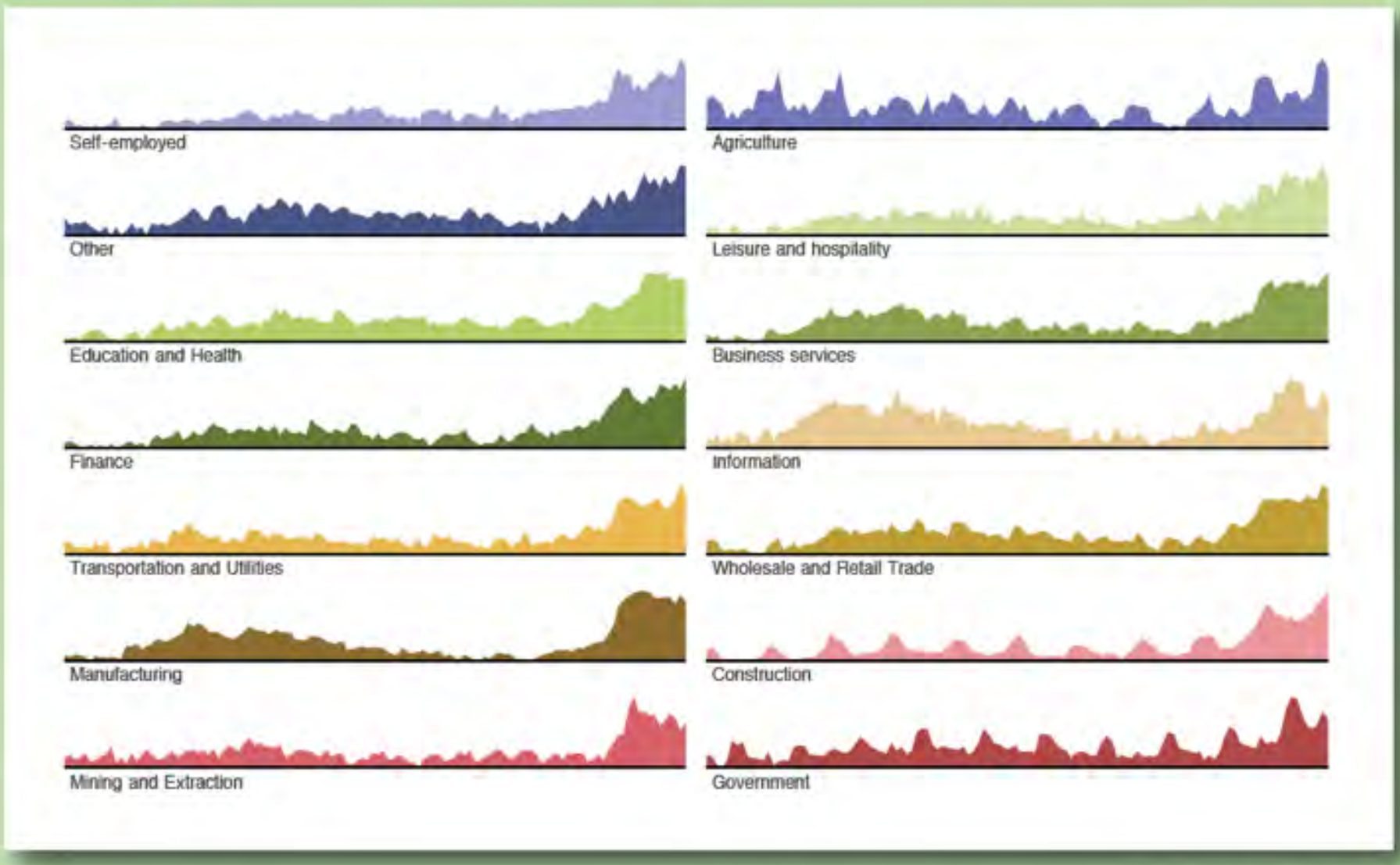


Source: U.S. Bureau of Labor Statistics
http://hci.stanford.edu/jheer/files/zoo/ex/time/stack.html

Small Multiples

FIGURE 1C

Small Multiples of Unemployed U.S. Workers Normalized by Industry, 2000-2010

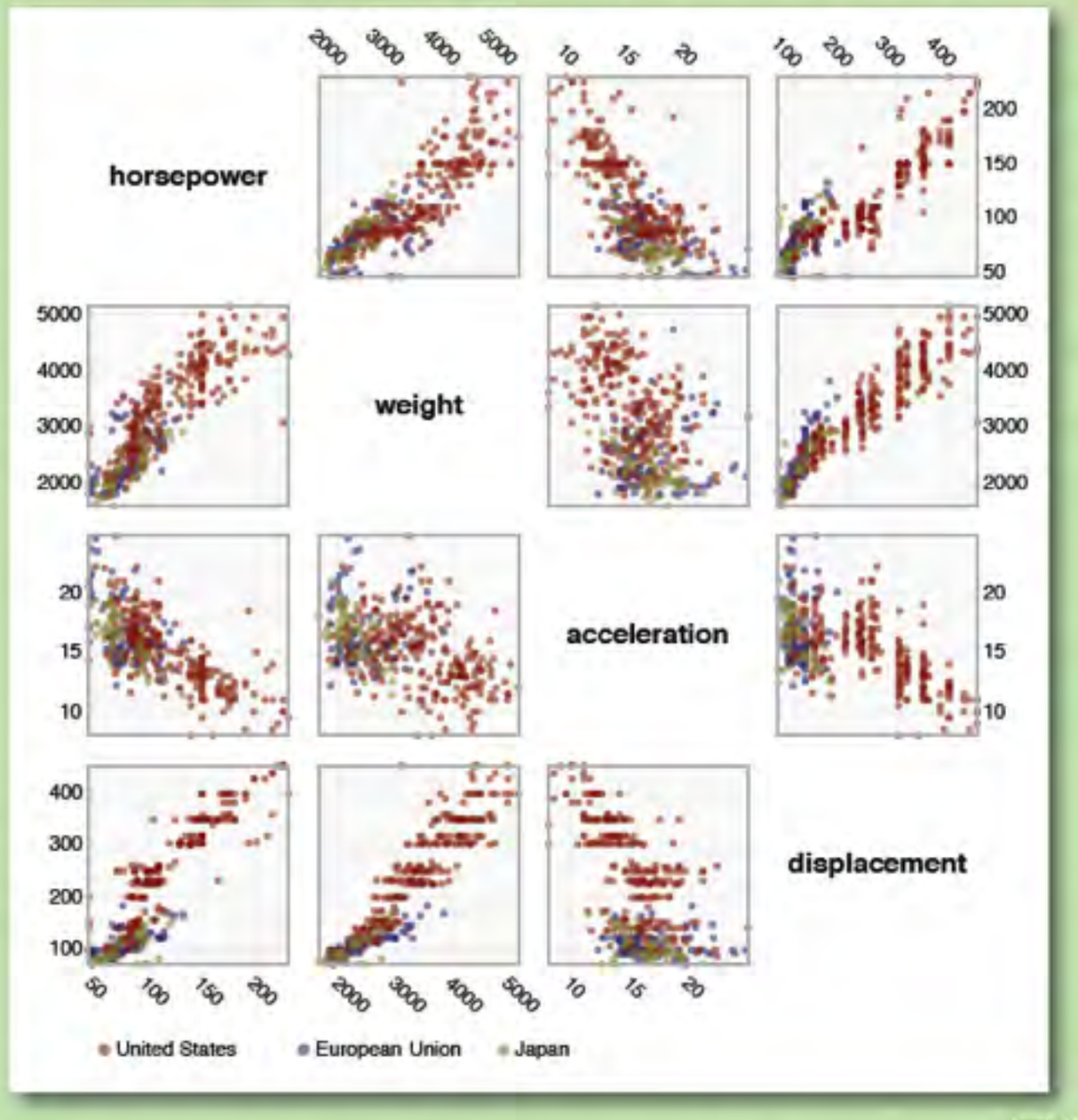


Source: U.S. Bureau of Labor Statistics
http://hci.stanford.edu/jheer/files/zoo/ex/time/multiples.html

Scatter Plot

FIGURE 2C

Scatter Plot Matrix of Automobile Data

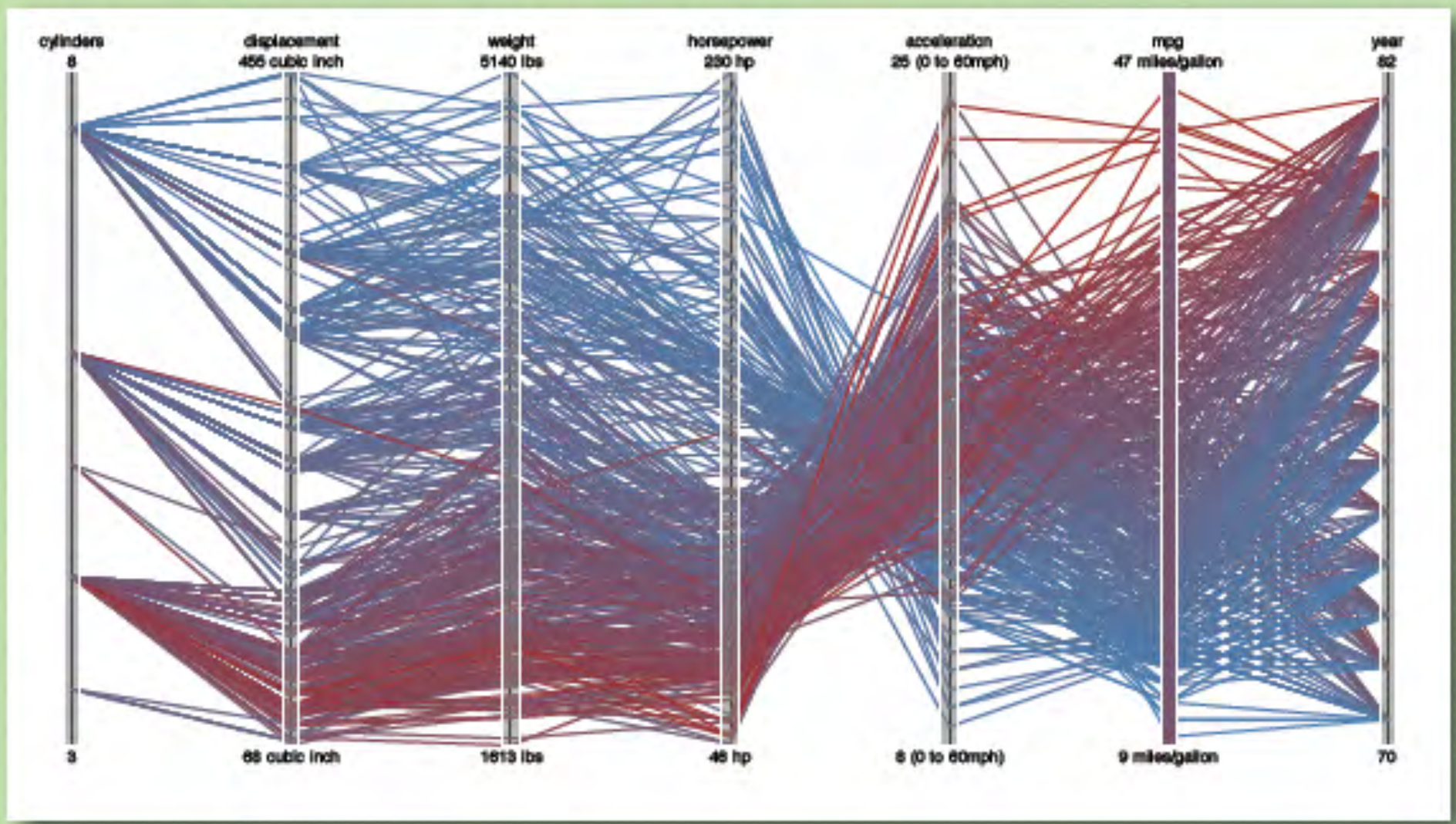


Source: GGobi
http://rci.stanford.edu/jwee/files/zoo/ex/stats/spiom.html

Parallel Coordinates

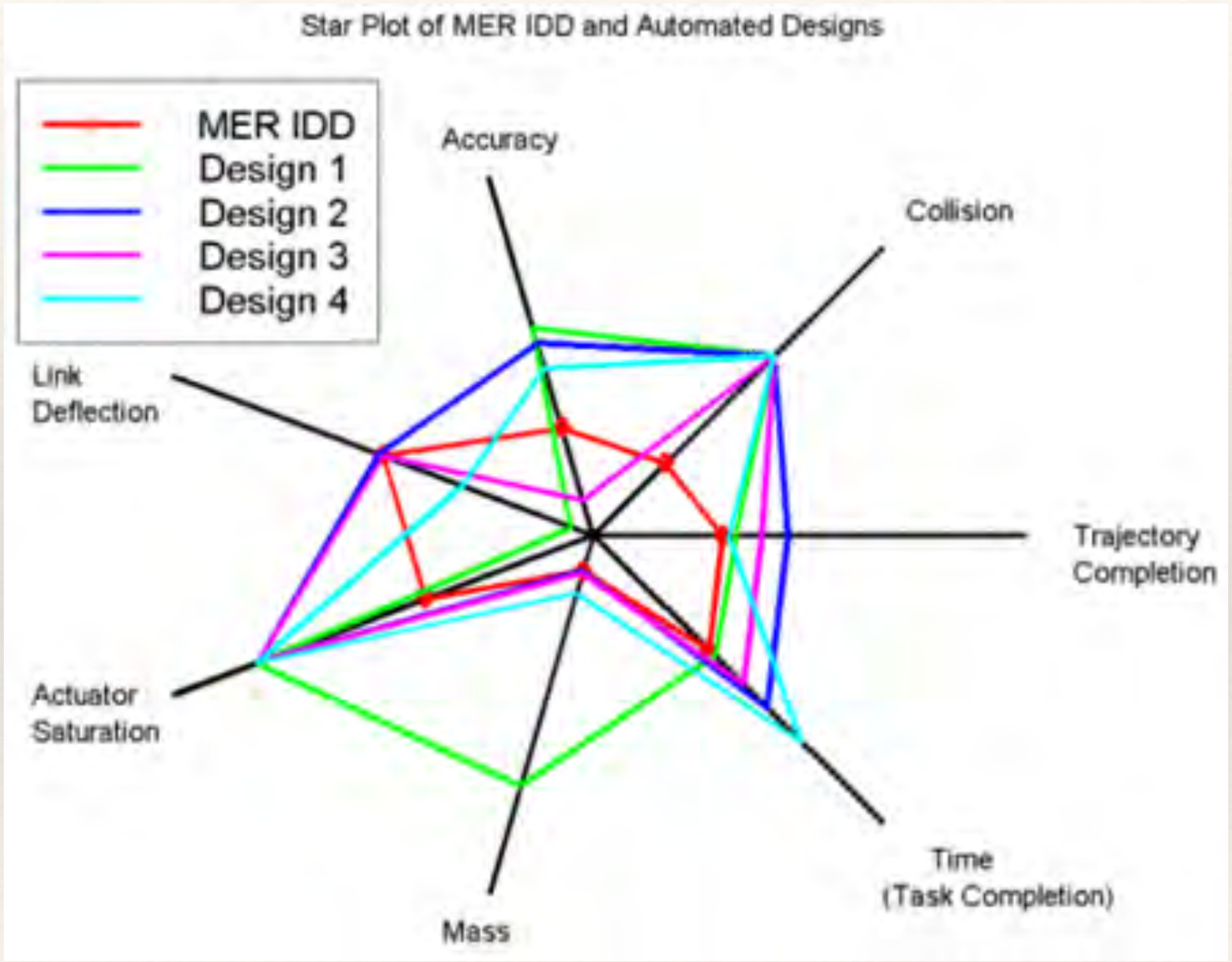
FIGURE 2D

Parallel Coordinates of Automobile Data



Source: GGobi
http://hci.stanford.edu/jheer/files/zoo/ex/stats/parallel.html

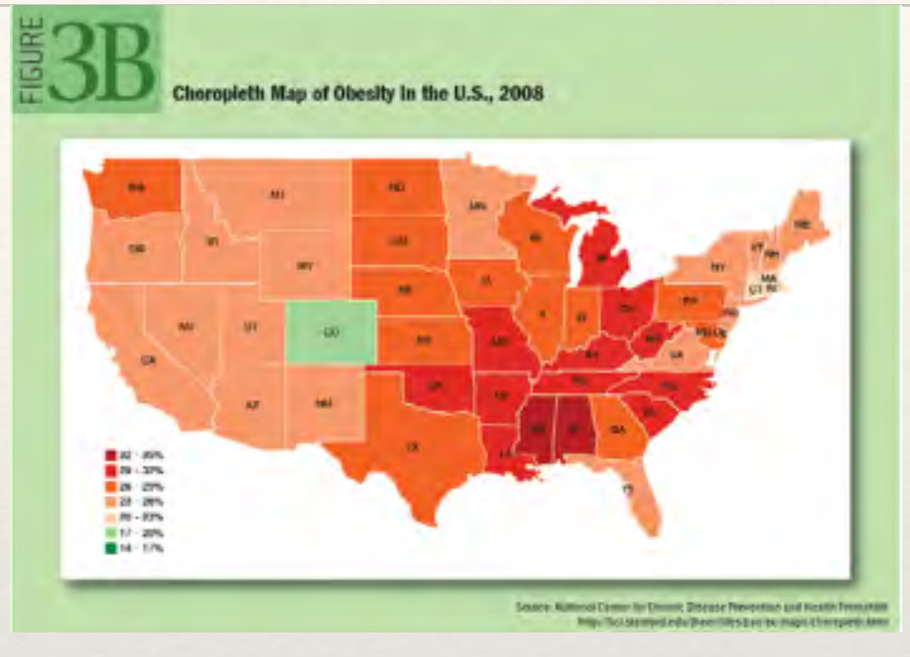
Radar Chart



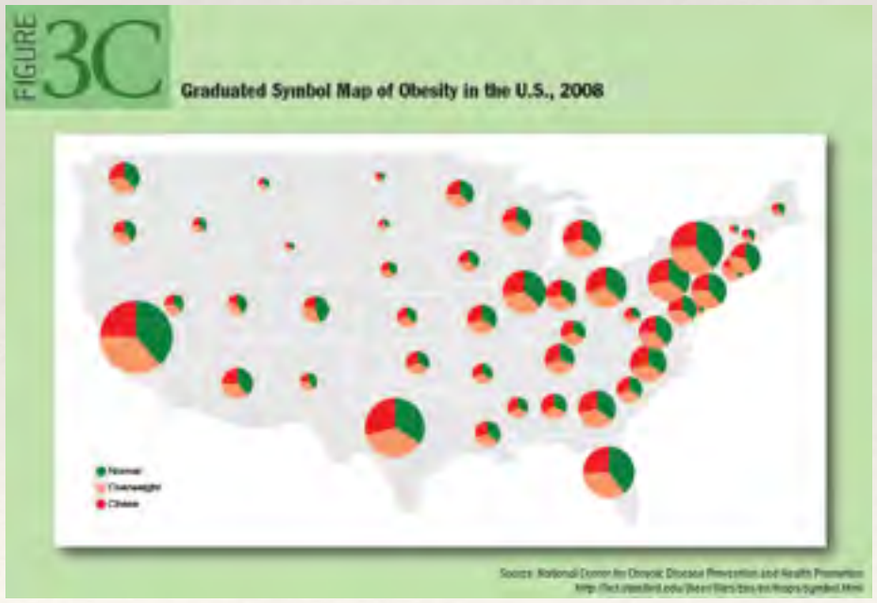
Typically Positive data

Map(s)

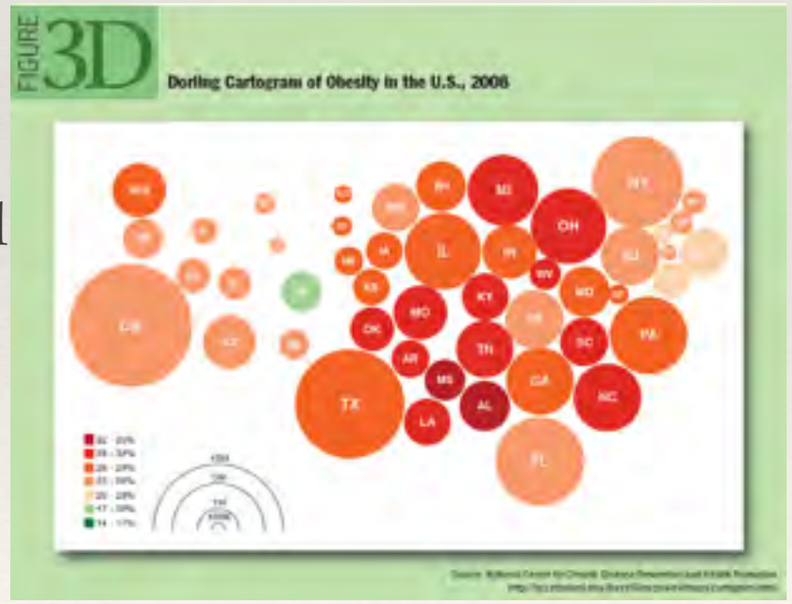
Flow



Choropleth



Graduated Symbol Map



Cartogram

More in Later Lecture

Hierarchies

FIGURE 4A

Radial Node-link Diagram of the Flare Package Hierarchy

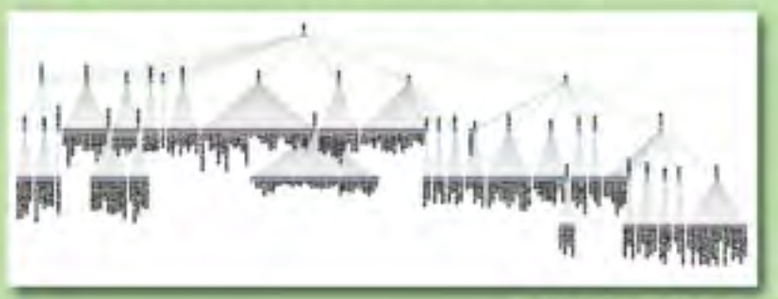


FIGURE 4B

Cartesian Node-link Diagram of the Flare Package Hierarchy

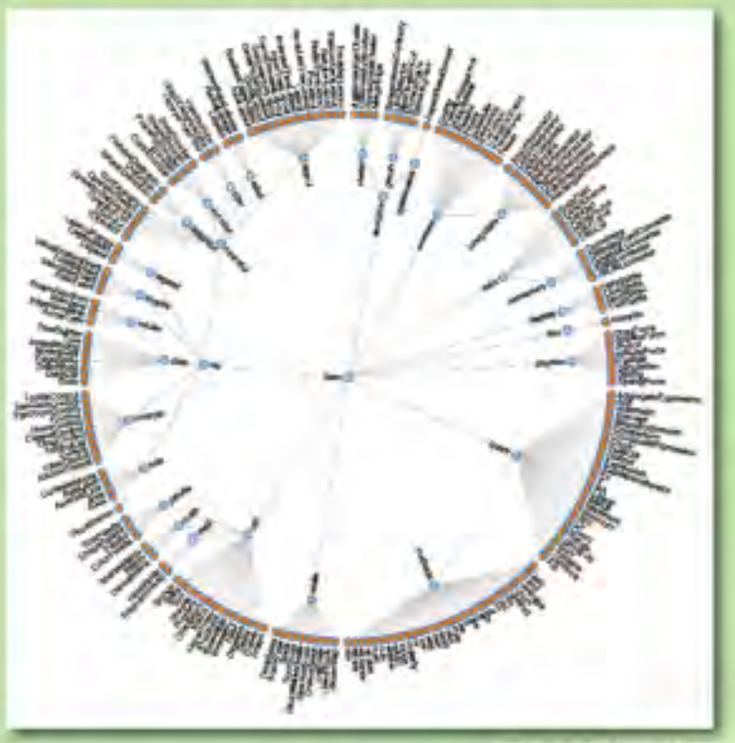


FIGURE 4C

Indented Tree Layout of the Flare Package Hierarchy

FIGURE 4D

Icicle Tree Layout of the Flare Package Hierarchy

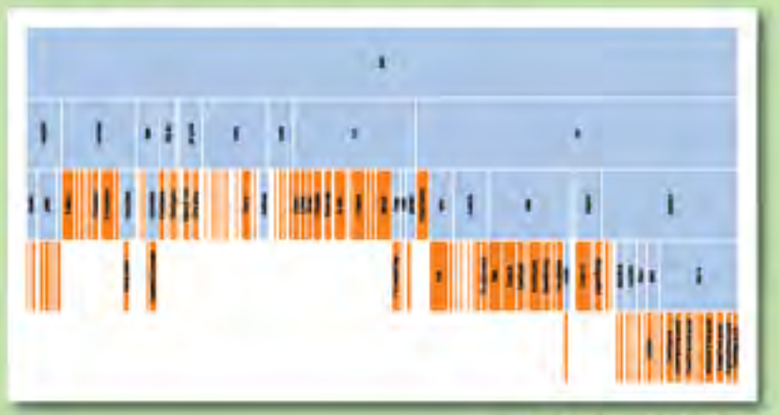


FIGURE 4E

Sunburst (Radial Space-filling) Layout of the Flare Package Hierarchy

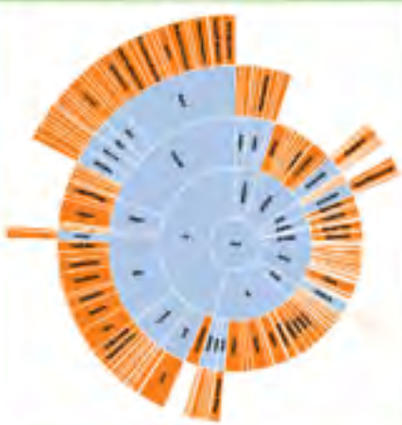


FIGURE 4F

Treemap Layout of the Flare Package Hierarchy

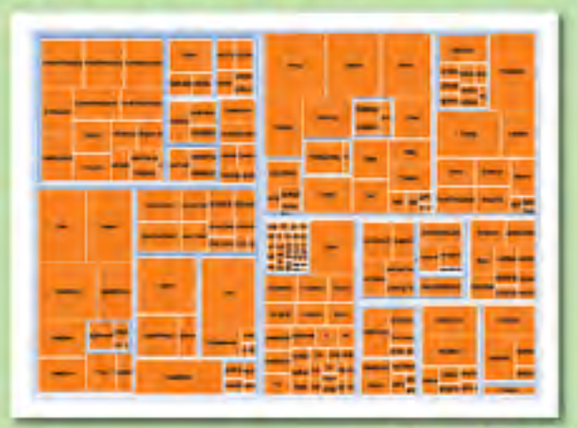
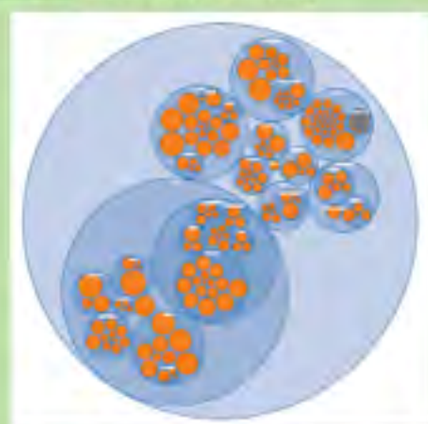


FIGURE 4G

Nested Circles Layout of the Flare Package Hierarchy



Network

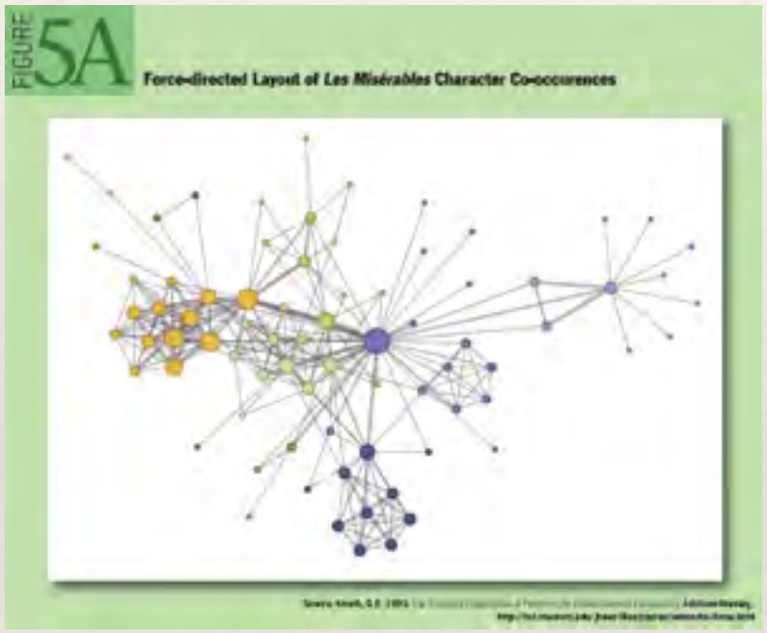


FIGURE 5A
Force-directed Layout of Les Misérables Character Co-occurences

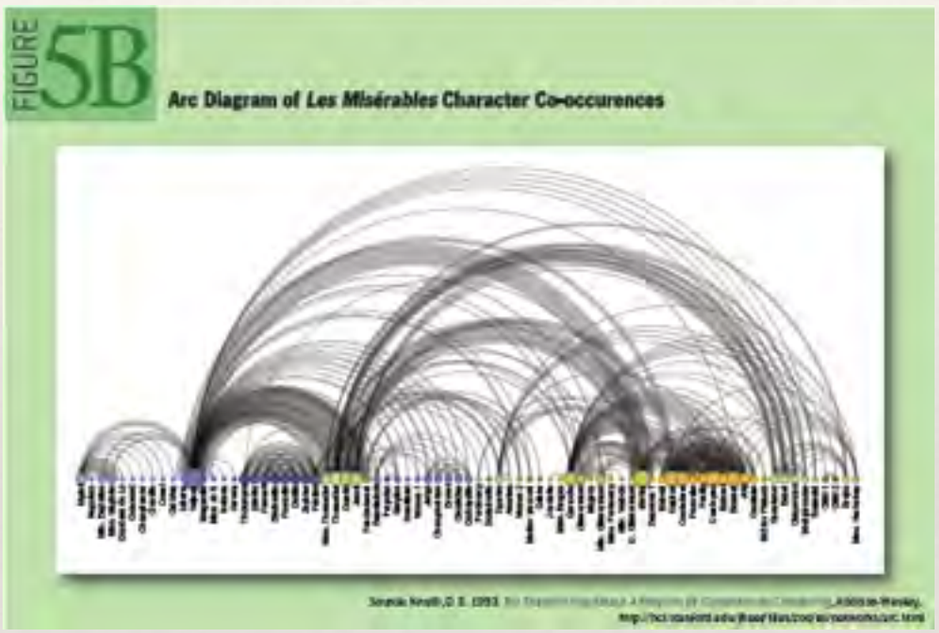


FIGURE 5B
Arc Diagram of Les Misérables Character Co-occurences

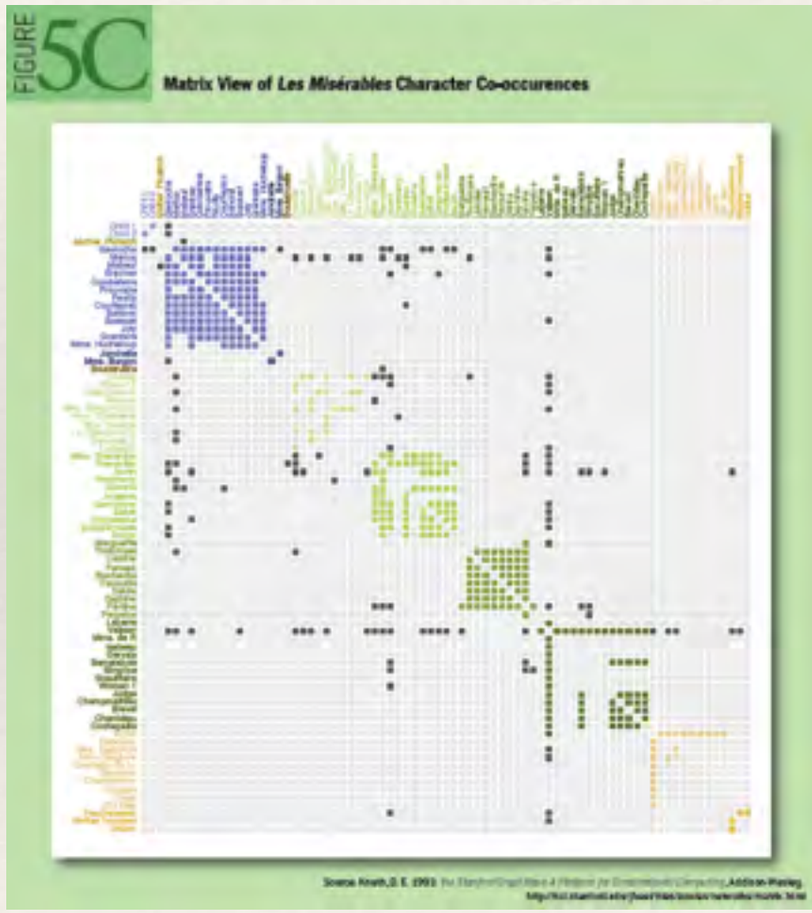


FIGURE 5C
Matrix View of Les Misérables Character Co-occurences

Recommended Tool for Static Plots

matplotlib

http://matplotlib.org/

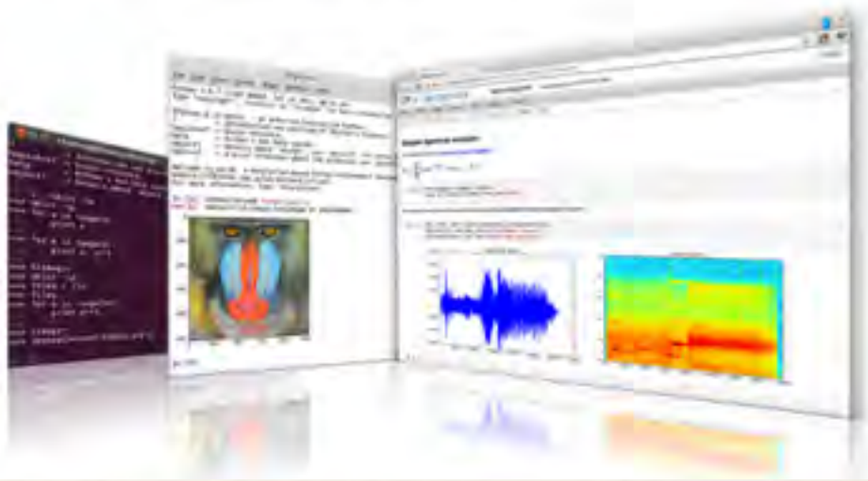
Rapid Prototyping: Use Ipython

IP[y]: IPython
Interactive Computing

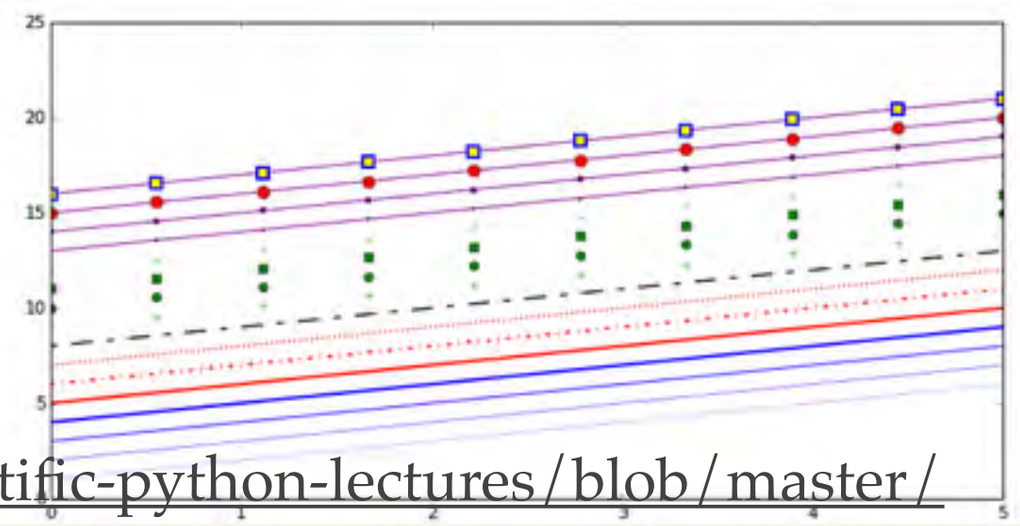
Install · Docs · Videos · News · Cite · Sponsors · Donate

IPython provides a rich architecture for interactive computing with:

- Powerful interactive shells (terminal and Qt-based).
- A browser-based notebook with support for code, text, mathematical expressions, inline plots and other rich media.
- Support for interactive data visualization and use of GUI toolkits.
- Flexible, embeddable interpreters to load into your own projects.
- Easy to use, high performance tools for parallel computing.



```
In [33]: fig, ax = plt.subplots(figsize=(12,6))

ax.plot(x, x+1, color="blue", linewidth=0.25)
ax.plot(x, x+2, color="blue", linewidth=0.50)
ax.plot(x, x+3, color="blue", linewidth=1.00)
ax.plot(x, x+4, color="blue", linewidth=2.00)

# possible linestype options '-', '--', '-.', ':', 'steps'
ax.plot(x, x+5, color="red", lw=2, linestyle='-')
ax.plot(x, x+6, color="red", lw=2, ls='-.')
ax.plot(x, x+7, color="red", lw=2, ls=':')

# custom dash
line, = ax.plot(x, x+8, color="black", lw=1.50)
line.set_dashes([5, 10, 15, 10]) # format: line length, space length, ...

# possible marker symbols: marker = '+', 'o', '*', 's', ',', '.', '1', '2'
, '3', '4', ...
ax.plot(x, x+ 9, color="green", lw=2, ls='*', marker='+')
ax.plot(x, x+10, color="green", lw=2, ls='*', marker='o')
ax.plot(x, x+11, color="green", lw=2, ls='*', marker='s')
ax.plot(x, x+12, color="green", lw=2, ls='*', marker='1')

# marker size and color
ax.plot(x, x+13, color="purple", lw=1, ls='-', marker='o', markersize=2)
ax.plot(x, x+14, color="purple", lw=1, ls='-', marker='o', markersize=4)
ax.plot(x, x+15, color="purple", lw=1, ls='-', marker='o', markersize=8, m
arkerfacecolor="red")
ax.plot(x, x+16, color="purple", lw=1, ls='-', marker='s', markersize=8,
        markerfacecolor="yellow", markeredgewidth=2, markeredgecolor="blue
");
```



http://ipython.org/

http://nbviewer.ipython.org/github/jrjohansson/scientific-python-lectures/blob/master/Lecture-4-Matplotlib.ipynb

http://www.erh.noaa.gov/okx/climate/almanacs/nycaug.htm

This data is unofficial and for informational purposes only.
For official certified climate data please contact the National Climatic Data Center (NCDC)
Page last updated: Sep 19th 2011

NATIONAL WEATHER SERVICE FORECAST OFFICE UPTON, NEW YORK

CENTRAL PARK NEW YORK CITY
RECORDS 1869-2011

TEMPERATURE/PRECIPITATION/SNOWFALL NORMALS 1971-2000
HEATING DEGREE DAY NORMALS 1961-1990

AUGUST

DAY	SUN EST		TEMPERATURE							ACCUMULATED NORMAL DEGREE DAYS (HEATING)		GREATEST DAILY	
			NORMAL			RECORD						precipi-	
	rise	set	max	min	mean	highest		lowest		month	season	tation	snowfall
1	4:52	7:12	84	69	76	100	1933	59	1964*	0	0	2.85 1878	
2	4:53	7:11	84	69	76	100	1955	57	1875	0	0	2.49 1973	
3	4:54	7:10	84	69	76	97	2005	55	1927*	0	0	2.44 1885	
4	4:55	7:08	84	69	76	100	1944	56	1886*	0	0	3.25 1915	
5	4:56	7:07	83	69	76	101	1944	56	1951*	0	0	1.44 1884	
6	4:57	7:06	83	69	76	97	1955	53	1869	0	0	3.31 1878	
7	4:58	7:05	83	69	76	104	1918	57	1994	0	0	2.18 1921	
8	4:59	7:04	83	69	76	99	2001	54	1903	0	0	2.60 1927	
9	5:00	7:02	83	69	76	103	2001	57	1989*	0	0	4.10 1942	
10	5:01	7:01	83	69	76	98	1949*	55	1879	0	0	4.64 1990	
11	5:02	7:00	83	69	76	102	1944	56	1962	0	0	2.39 1983	
12	5:03	6:59	83	69	76	97	1944	55	1889	0	0	3.62 1955	
13	5:04	6:57	83	69	76	99	2005	55	1930	0	0	2.70 1955	
14	5:05	6:56	83	68	76	99	1988	54	1964	0	0	5.81 2011	

Libre Office Method

- "Manually chop out data"
- Put in spreadsheet
- Use "chart" function
- Fix up

Manually Chop Out Data

- Here I use Vim block select to pull out data

	DAY	SUN EST rise	SUN EST set	TEMPERATURE NORMAL max	TEMPERATURE NORMAL min	TEMPERATURE NORMAL mean	TEMPERATURE RECORD highest		TEMPERATURE RECORD lowest		ACCUMULATED NORMAL DEGREE DAYS (HEATING) month	ACCUMULATED NORMAL DEGREE DAYS (HEATING) season
1	1	4:52	7:12	84	69	76	100	1933	59	1964*	0	0	2.85	1878
2	2	4:53	7:11	84	69	76	100	1955	57	1875	0	0	2.49	1973
3	3	4:54	7:10	84	69	76	97	2005	55	1927*	0	0	2.44	1885
4	4	4:55	7:08	84	69	76	100	1944	56	1886*	0	0	3.25	1915
5	5	4:56	7:07	83	69	76	101	1944	56	1951*	0	0	1.44	1884
6	6	4:57	7:06	83	69	76	97	1955	53	1869	0	0	3.31	1878
7	7	4:58	7:05	83	69	76	104	1918	57	1994	0	0	2.18	1921
8	8	4:59	7:04	83	69	76	99	2001	54	1903	0	0	2.60	1927
9	9	5:00	7:02	83	69	76	103	2001	57	1989*	0	0	4.10	1942
10	10	5:01	7:01	83	69	76	98	1949*	55	1879	0	0	4.64	1990
11	11	5:02	7:00	83	69	76	102	1944	56	1962	0	0	2.39	1983
12	12	5:03	6:59	83	69	76	97	1944	55	1889	0	0	3.62	1955
13	13	5:04	6:57	83	69	76	99	2005	55	1930	0	0	2.70	1955
14	14	5:05	6:56	83	68	76	99	1988	54	1964	0	0	5.81	2011
15	15	5:06	6:54	83	68	76	97	1988	54	1964	0	0	1.52	1911
16	16	5:07	6:53	83	68	76	96	1944	55	1880	0	0	4.80	1909
17	17	5:08	6:52	83	68	75	95	1944	56	1979*	0	0	2.86	1974
18	18	5:09	6:50	83	68	75	94	2002*	55	1915	0	0	3.95	1879

Libre Office (or spreadsheet prog)

- Paste data in
- Delete blank lines

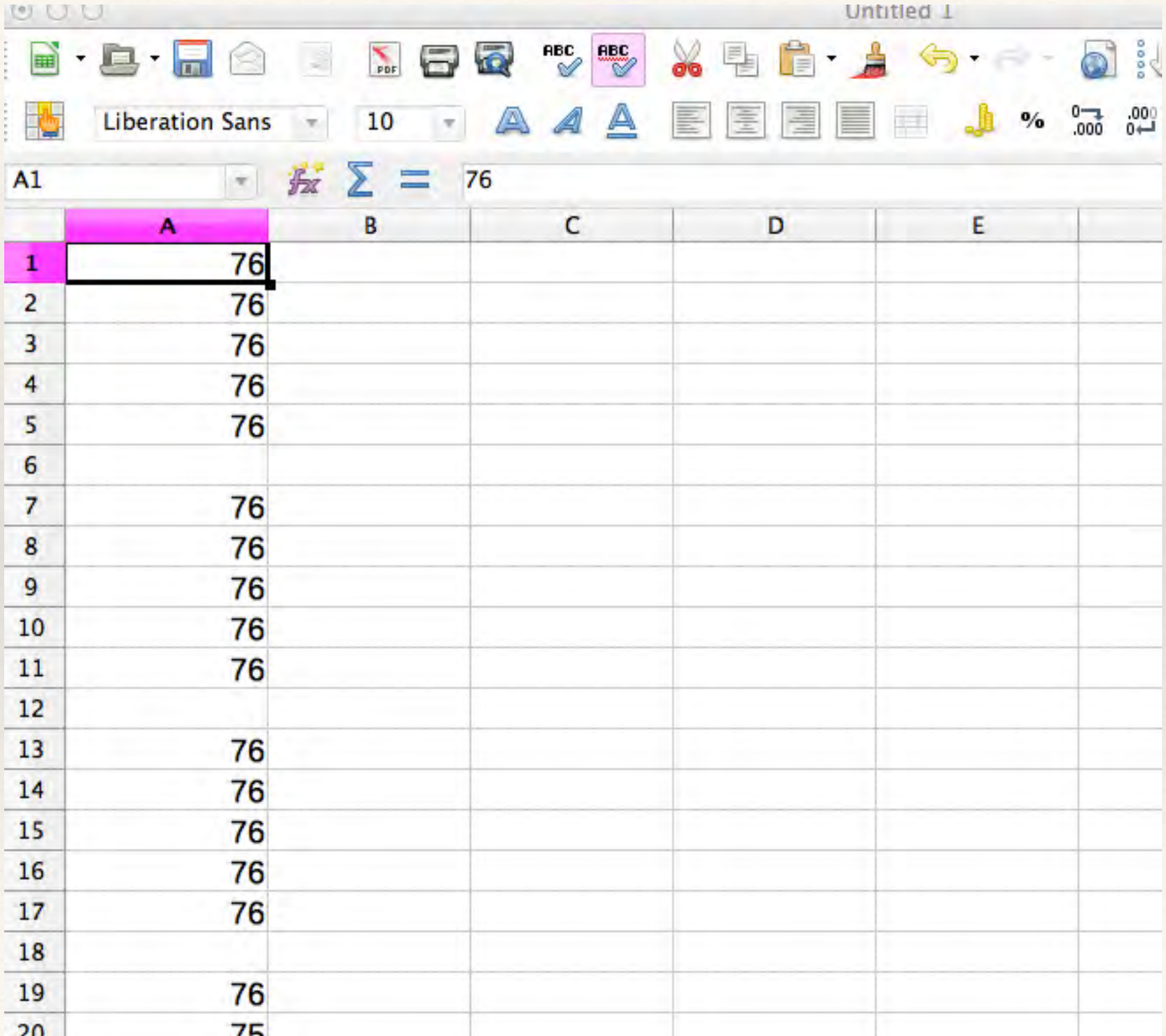
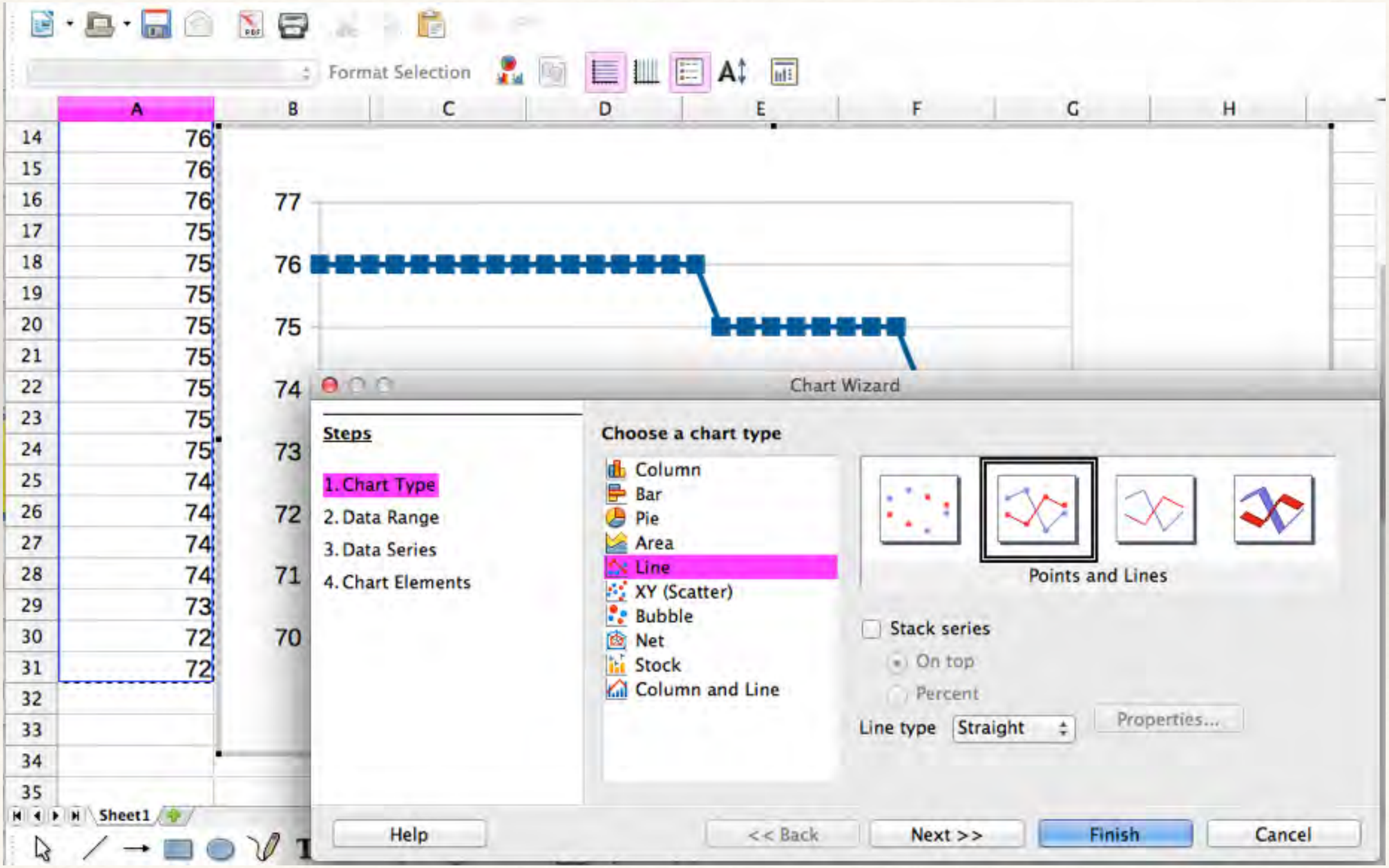
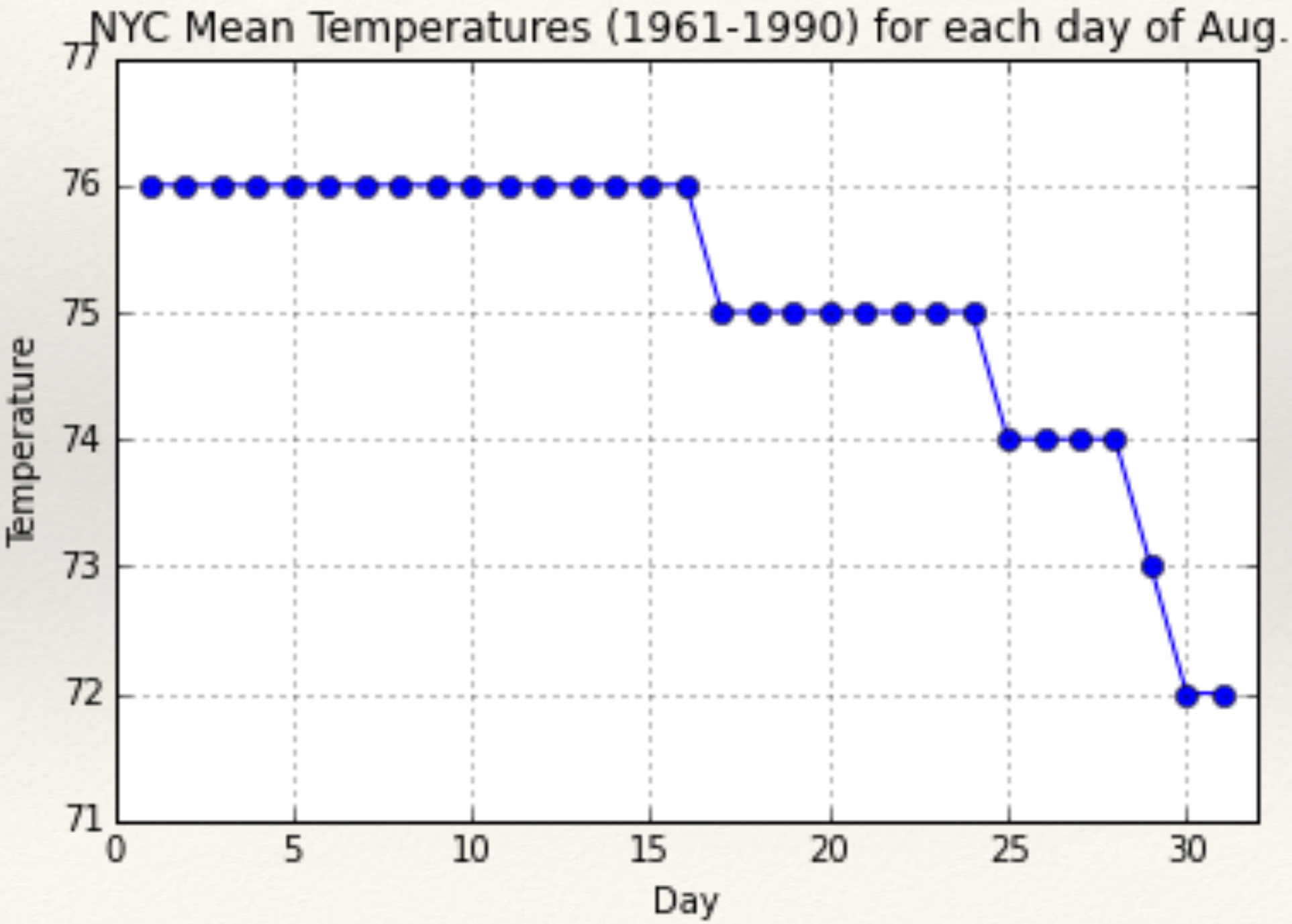


Chart Wizard



Label and adjust



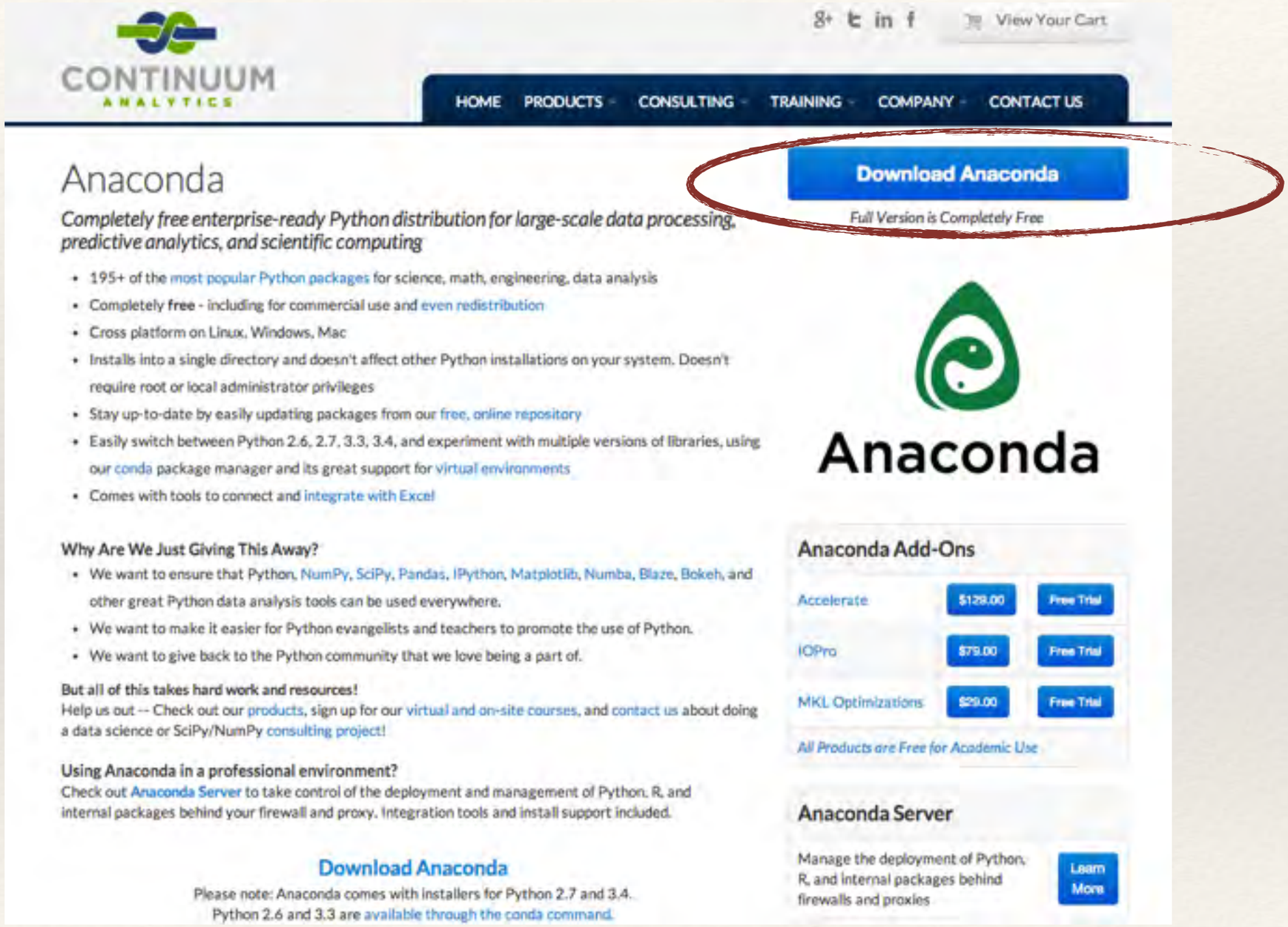
Pro/Cons

- Pros:
  - WYSIWYG
  - Can directly manipulate data
  - Easily try options
- Cons:
  - Difficult to automate
  - Limited flexibility
  - Limited processing options

Python

- Can interactively work with
  - ipython shell
  - ipython notebook
- Can save notebook or turn into script

Python Distribution (one choice)



https://store.continuum.io/cshop/anaconda/

https://store.continuum.io/cshop/academicanaconda

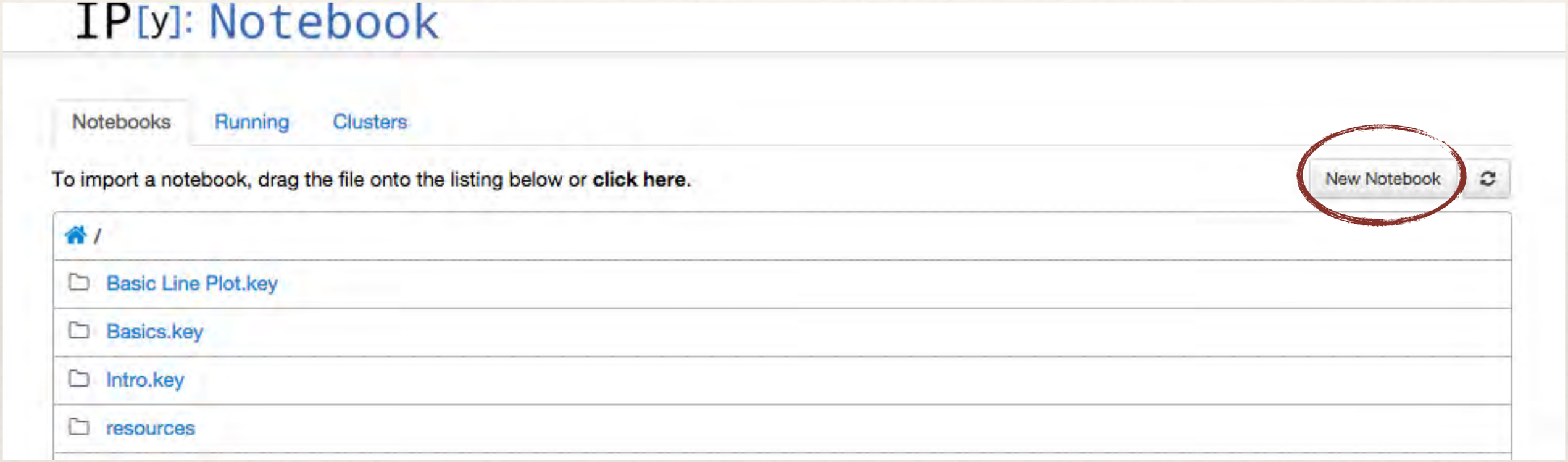
Other choices

- Mac OS: homebrew (http://brew.sh/) install python, then numpy, matplotlib, scipy using home-brew … everything else with pip
- Linux Ubuntu apt-get (yum for redhat) for numpy, matplotlib, scipy
- Windows: use anaconda (previous slide) or Ubuntu inside Virtual Box VM then see above

Start Notebook

- $ ipython notebook
- (assumes installation and set up ok)

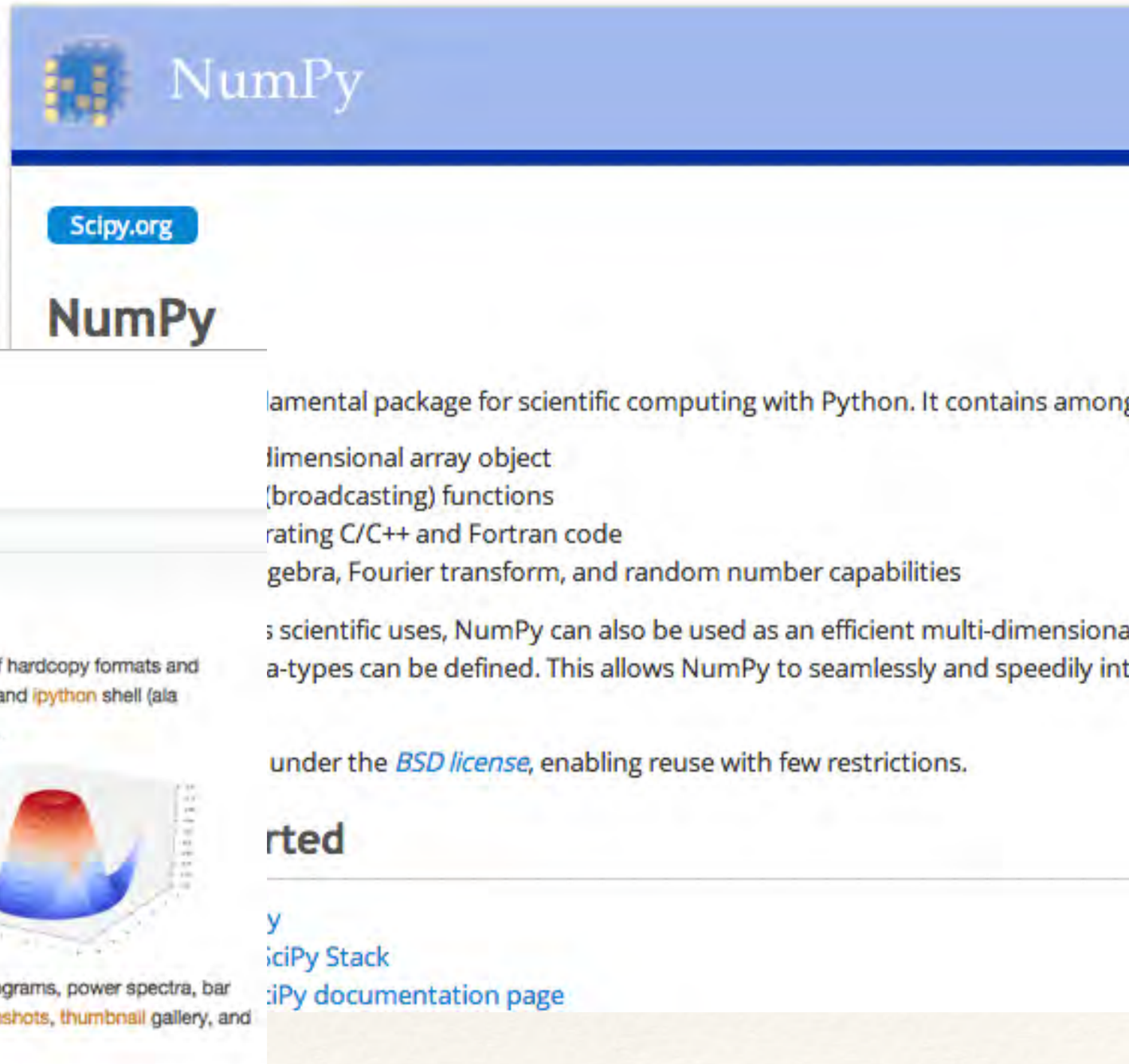
Open New Notebook

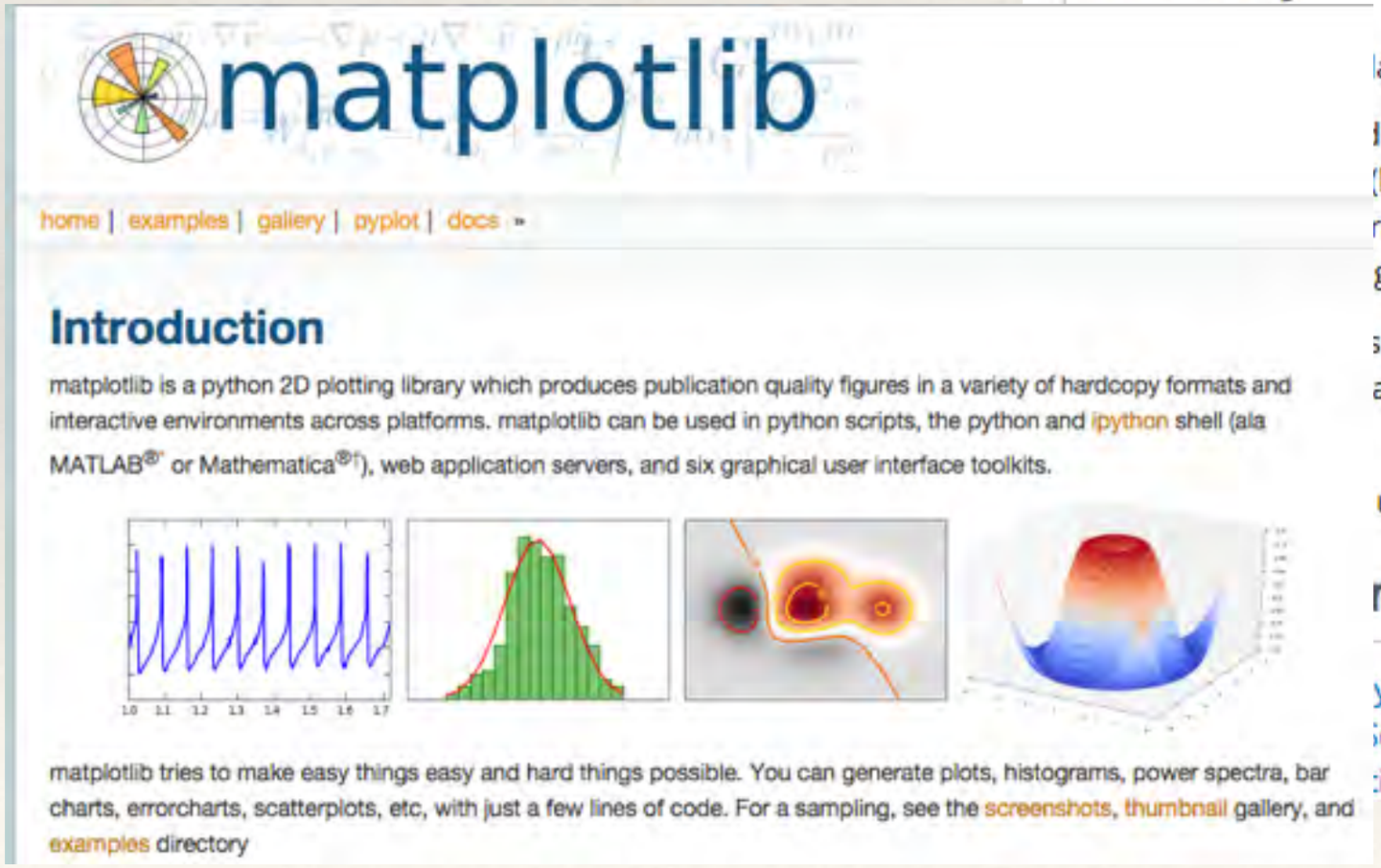


Initial Load Needed libraries

```
%matplotlib inline
```

```
import matplotlib.pyplot as plt
import numpy as np
```





Request Library for Loading from Web

Previous topic

9. Full Grammar spec

previous chapter

Next topic

1. Introduction

This Page

- 20. Internet Protocols and Support
  - 20.1. `webbrowser` — Convenient Web-browser controlle
  - 20.2. `cgi` — Common Gateway Interface support
  - 20.3. `cgitb` — Traceback manager for CGI scripts
  - 20.4. `wsgiref` — WSGI Utilities and Reference Implem
  - 20.5. `urllib` — Open arbitrary resources by URL
  - 20.6. `urllib2` — extensible library for opening URLs

Hard Way

Requests 2.4.0 documentation »

next |

Requests: HTTP for Humans

Easier API

Release v2.4.0. (*Installation*)

Requests is an *Apache2 Licensed* HTTP library, written in Python, for human beings.

Python's standard **urllib2** module provides most of the HTTP capabilities you need, but the AP
broken. It was built for a different time — and a different web. It requires an *enormous* amount o
method overrides) to perform the simplest of tasks.

Things shouldn't be this way. Not in Python.

Load Data from Web

IP[y]: Notebook Basic Line Plot Last Checkpoint: Sep 01 12:39 (unsaved changes)

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

In [23]:
```
%matplotlib inline
```

In [24]:
```
import matplotlib.pyplot as plt
import numpy as np
```

In [25]:
```
import requests
```

In [26]:
```
r = requests.get('http://www.erh.noaa.gov/okx/climate/almanacs/nycaug.htm')
r.text
```

Out[26]:
```
u'<html>\r\n\r\n<head>\r\n<meta name="GENERATOR" content="Microsoft FrontPage 3.0">\r\n<meta name="P
rogId" content="FrontPage.Editor.Document">\r\n<title>AUGUST CLIMATE SUMMARY - CENTRAL PARK NY</titl
e>\r\n</head>\r\n\r\n<body bgcolor="#FFFF80">\r\n\r\n<div align="center"><strong>This data is unoffi
cial and for informational purposes only.<br> \r\n  For official certified climate data please conta
ct the <a href="http://www.ncdc.noaa.gov/oa/ncdc.html">National Climatic Data Center (NCDC)</a></str
ong></div> \r\n<div align="center"><strong>Page last updated: Sep 19th 2011</strong></div>\r\n<div a
lign="center"><center>\r\n\r\n    <pre><b>NATIONAL WEATHER SERVICE FORECAST OFFICE UPTON, NEW YORK\r
\n\r\nCENTRAL PARK NEW YORK CITY\r\nRECORDS 1869-2011\r\n\r\nTEMPERATURE/PRECIPITATION/SNOWFALL NORM
ALS 1971-2000 \r\nHEATING DEGREE DAY NORMALS 1961-1990\r\n\r\nAUGUST\r\n\r\n
                    ACCUMULATED NORMAL                \r\n       DAY    SUN              TEMPERATURE
      DEGREE DAYS   GREATEST DAILY\r\n                      EST       NORMAL          RECORD         (HEATIN
G)  precipi-          \r\n                rise   set max min mean highest  lowest  month season   tati
on  snowfall\r\n\r\n 1  4:52  7:12  84  69  76 100 1933  59 1964*   0     0  2.85 1878\r\n 2  4:53
7:11  84  69  76 100 1955  57 1875    0     0  2.49 1973\r\n 3  4:54  7:10  84  69  76  97 2005  55
1927*   0     0  2.44 1885\r\n 4  4:55  7:08  84  69  76 100 1944  56 1886*   0     0  3.25 1915\r\n
 5  4:56  7:07  83  69  76 101 1944  56 1951*   0     0  1.44 1884\r\n\r\n 6  4:57  7:06  83  69  76
  97 1955  53 1869    0     0  3.31 1878\r\n 7  4:58  7:05  83  69  76 104 1918  57 1994    0     0
 2.18 1921\r\n 8  4:59  7:04  83  69  76  99 2001  54 1903    0     0  2.60 1927\r\n 9  5:00  7:02
```

Text Munging (scraping)

- Some string methods:
  - 'endswith',
  - 'expandtabs',
  - 'find',
  - 'index',
  - 'isalnum',
  - 'isalpha',
  - 'isdigit',
  - 'islower',
  - 'isspace',
  - 'istitle',
  - 'isupper',
  - 'lstrip',
  - 'partition',
  - 'replace',
  - 'rfind',
  - 'rindex',
  - 'rjust',
  - 'rpartition',
  - 'rsplit',
  - 'rstrip',
  - 'split',
  - 'splitlines',
  - 'startswith',
  - 'strip'

Regular Expressions (Regexp)

but …

Some people, when confronted with a problem, think "I know, I'll use regular expressions." Now they have two problems.

- Jamie Zawinski (?)

Regexp (very useful)

Table Of Contents

7.2. re — Regular expression operations ¶

This module provides regular expression matching operations similar to those found in Perl. Both patterns and strings to be searched can be Unicode strings as well as 8-bit strings.

Regular expressions use the backslash character (`'\'`) to indicate special forms or to allow special characters to be used without invoking their special meaning. This collides with Python's usage of the same character for the same purpose in string literals; for example, to match a literal backslash, one might have to write `'\\\\'` as the pattern string, because the regular expression must be `\\`, and each backslash must be expressed as `\\` inside a regular Python string literal.

The solution is to use Python's raw string notation for regular expression patterns; backslashes are not handled in any special way in a string literal prefixed with `'r'`. So `r"\n"` is a two-character string containing `'\'` and `'n'`, while `"\n"` is a one-character string containing a newline. Usually patterns will be expressed in Python code using this raw string notation.

It is important to note that most regular expression operations are available as module-level functions and `RegexObject` methods. The functions are shortcuts that don't require you to compile a regex object first, but miss some fine-tuning parameters.

See also:

Look for data lines with regexp

This data is unofficial and for informational purposes only.
For official certified climate data please contact the National Climatic Data Center (NCDC)
Page last updated: Sep 19th 2011

NATIONAL WEATHER SERVICE FORECAST OFFICE UPTON, NEW YORK

CENTRAL PARK NEW YORK CITY
RECORDS 1869-2011

TEMPERATURE/PRECIPITATION/SNOWFALL NORMALS 1971-2000
HEATING DEGREE DAY NORMALS 1961-1990

AUGUST

DAY	SUN EST		TEMPERATURE							ACCUMULATED NORMAL DEGREE DAYS (HEATING)		GREATEST DAILY	
			NORMAL			RECORD							
	rise	set	max	min	mean	highest		lowest		month	season	precipi-tation	snowfall
1	4:52	7:12	84	69	76	100	1933	59	1964*	0	0	2.85	1878
2	4:53	7:11	84	69	76	100	1955	57	1875	0	0	2.49	1973
3	4:54	7:10	84	69	76	97	2005	55	1927*	0	0	2.44	1885
4	4:55	7:08	84	69	76	100	1944	56	1886*	0	0	3.25	1915
5	4:56	7:07	83	69	76	101	1944	56	1951*	0	0	1.44	1884
6	4:57	7:06	83	69	76	97	1955	53	1869	0	0	3.31	1878
7	4:58	7:05	83	69	76	104	1918	57	1994	0	0	2.18	1921
8	4:59	7:04	83	69	76	99	2001	54	1903	0	0	2.60	1927
9	5:00	7:02	83	69	76	103	2001	57	1989*	0	0	4.10	1942

Split then filter

```
import re
```

```
lines = [line for line in r.text.split('\n') if re.match('(\d|\s)\d\s\s\d',line)]
```

```
lines
```

```
[u' 1  4:52  7:12  84  69  76 100 1933  59 1964*   0     0  2.85 1878\r',
 u' 2  4:53  7:11  84  69  76 100 1955  57 1875    0     0  2.49 1973\r',
 u' 3  4:54  7:10  84  69  76  97 2005  55 1927*   0     0  2.44 1885\r',
 u' 4  4:55  7:08  84  69  76 100 1944  56 1886*   0     0  3.25 1915\r',
 u' 5  4:56  7:07  83  69  76 101 1944  56 1951*   0     0  1.44 1884\r',
 u' 6  4:57  7:06  83  69  76  97 1955  53 1869    0     0  3.31 1878\r',
 u' 7  4:58  7:05  83  69  76 104 1918  57 1994    0     0  2.18 1921\r',
 u' 8  4:59  7:04  83  69  76  99 2001  54 1903    0     0  2.60 1927\r',
 u' 9  5:00  7:02  83  69  76 103 2001  57 1989*   0     0  4.10 1942\r',
 u'10  5:01  7:01  83  69  76  98 1949* 55 1879    0     0  4.64 1990\r',
 u'11  5:02  7:00  83  69  76 102 1944  56 1962    0     0  2.39 1983\r',
 u'12  5:03  6:59  83  69  76  97 1944  55 1889    0     0  3.62 1955\r',
 u'13  5:04  6:57  83  69  76  99 2005  55 1930    0     0  2.70 1955\r',
 u'14  5:05  6:56  83  68  76  99 1988  54 1964    0     0  5.81 2011\r',
 u'15  5:06  6:54  83  68  76  97 1988  54 1964    0     0  1.52 1911\r',
 u'16  5:07  6:53  83  68  76  96 1944  55 1880    0     0  4.80 1909\r',
```

Extract Numbers and Convert

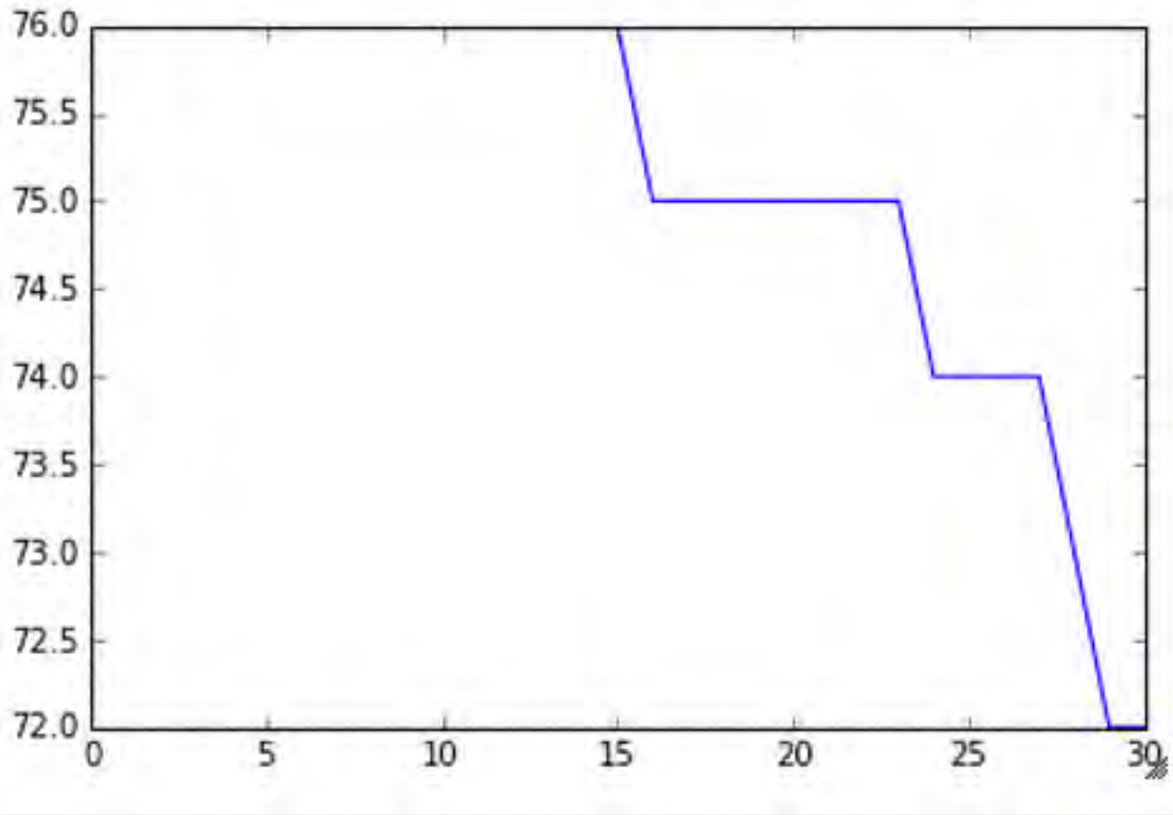
```
data = [ int(line.split()[5]) for line in lines]
```

```
data
```

```
[76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 76,
 75,
 75,
 75,
 75,
 75,
```

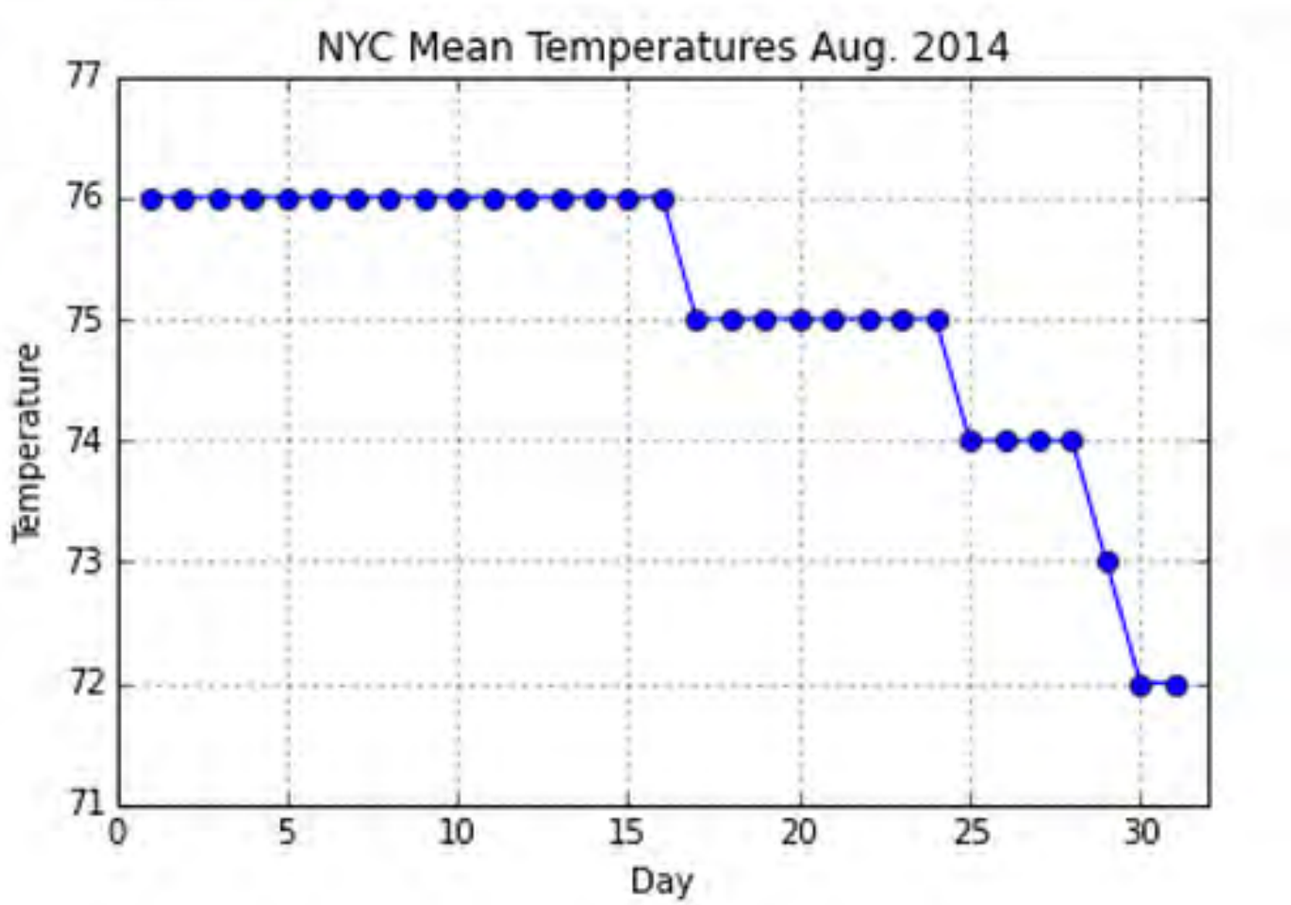
Quick Plot

```
: plt.plot(data)
: [<matplotlib.lines.Line2D at 0x10eaece90>]
```



Explicitly Set Properties

```
fig = plt.figure()
ax = fig.add_subplot(1,1,1)
ax.set_title('NYC Mean Temperatures Aug. 2014')
ax.set_ylabel('Temperature')
ax.set_xlabel('Day')
ax.plot(range(1,32),data,'-',marker='o')
ax.axis([0, 32, min(data)-1, max(data)+1])
ax.grid(True)
```



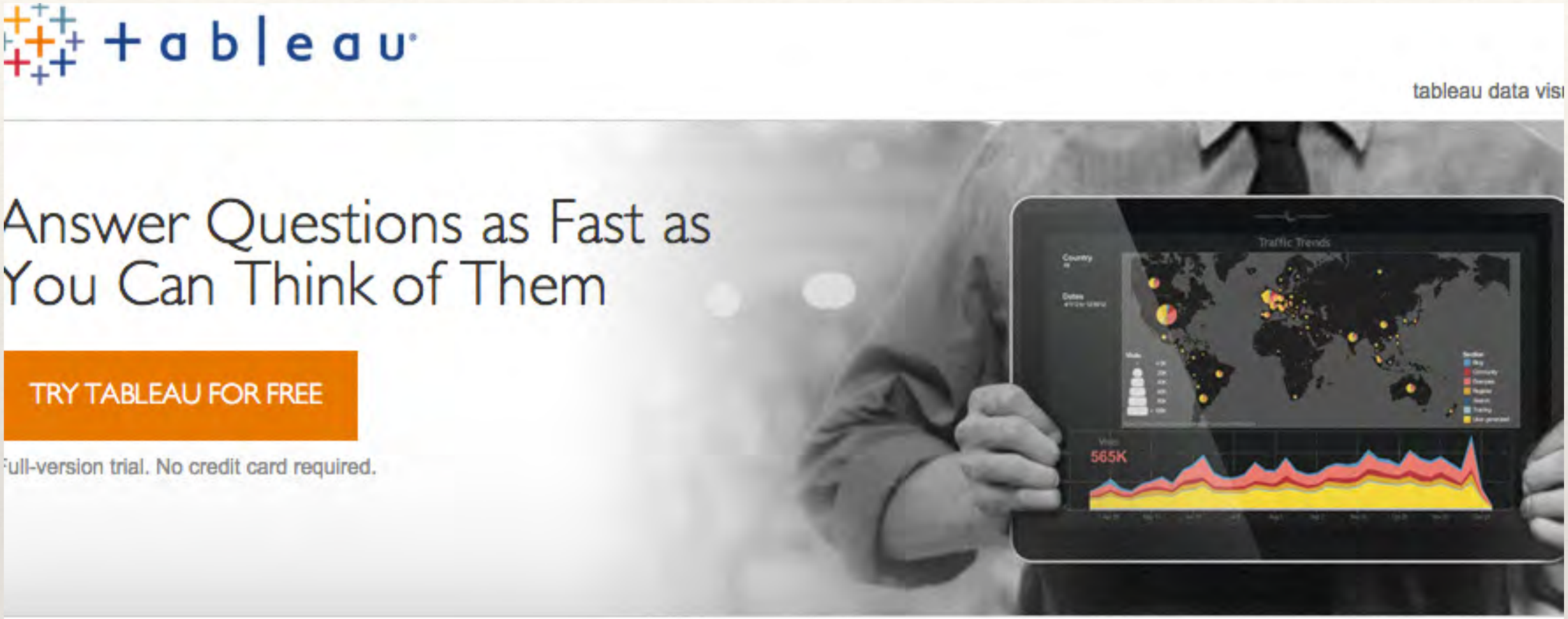
Imperative vs Object Approach

```
plt.plot(data)
```

vs

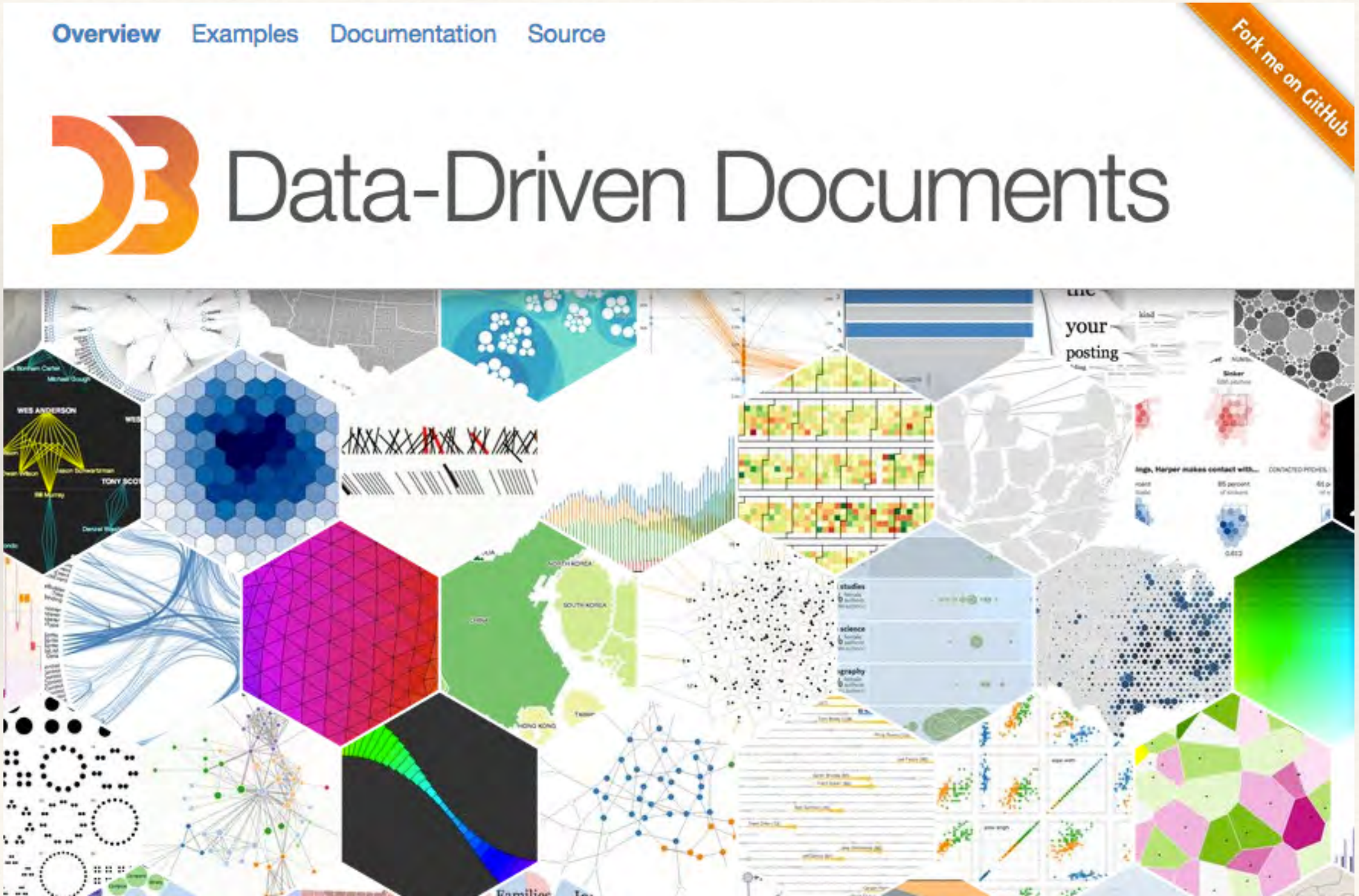
```
fig = plt.figure()
ax = fig.add_subplot(1,1,1)
ax.set_title('NYC Mean Temperatures Aug. 2014')
ax.set_ylabel('Temperature')
ax.set_xlabel('Day')
ax.plot(range(1,32),data,'-',marker='o')
ax.axis([0, 32, min(data)-1, max(data)+1])
ax.grid(True)
```

Tableau may also be helpful

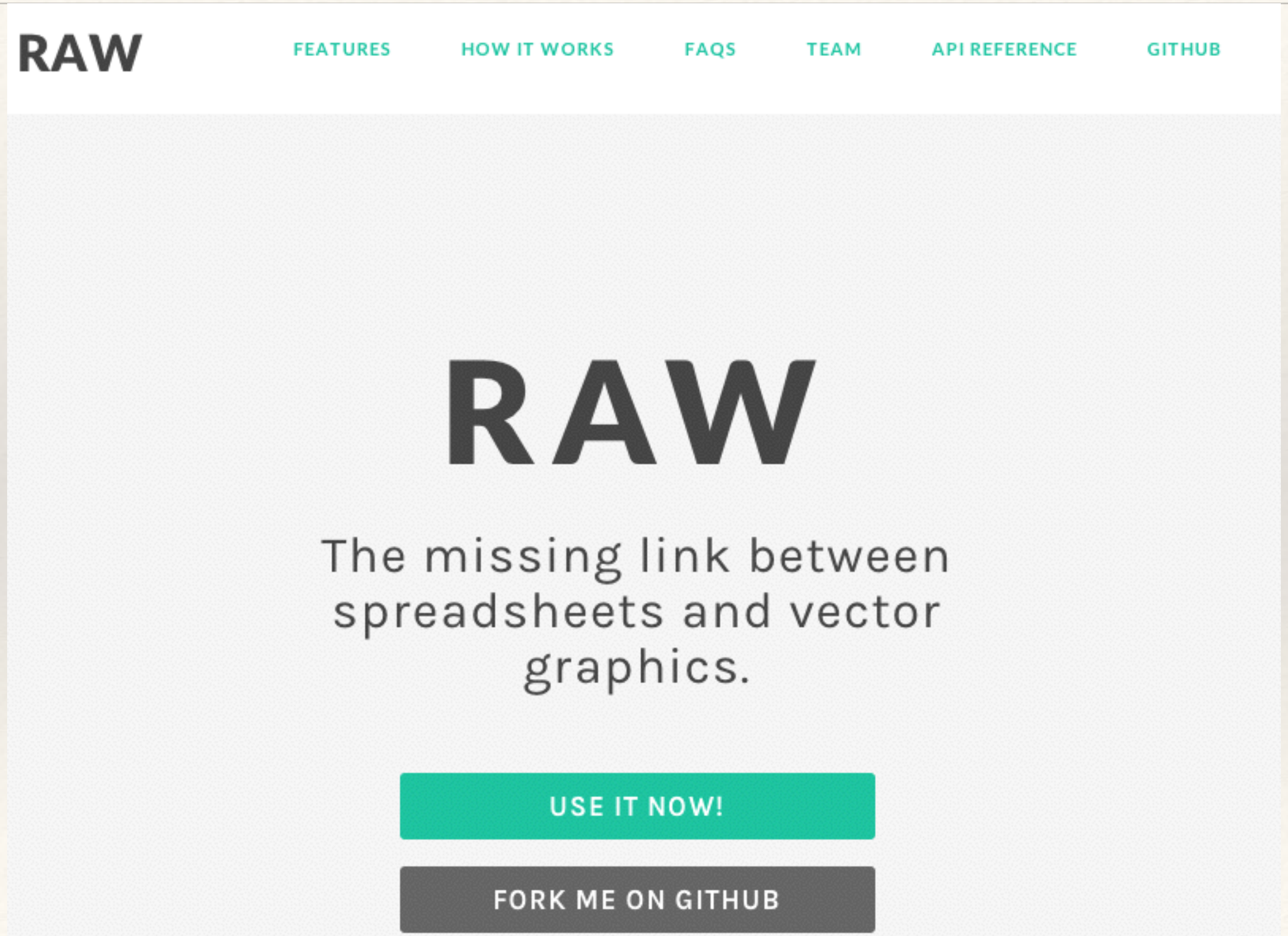


http://www.tableau.com/

Later use D3


Overview
Examples
Documentation
Source
Fork me on GitHub
Data-Driven Documents

Raw lets you do some D3 pro typing



Some Recommended Software Tools

- mercurial (bit bucket)/git github [version control]
- scientific python tools
  - python, numpy, scipy, matplotilib, pandas, ipython, basemap
  - linux (apt-get)/pip, mac (homebrew), mac/windows anaconda from continuum
- D3
- Editor, web browser (vim/sublime text2)

Supplemental Tools

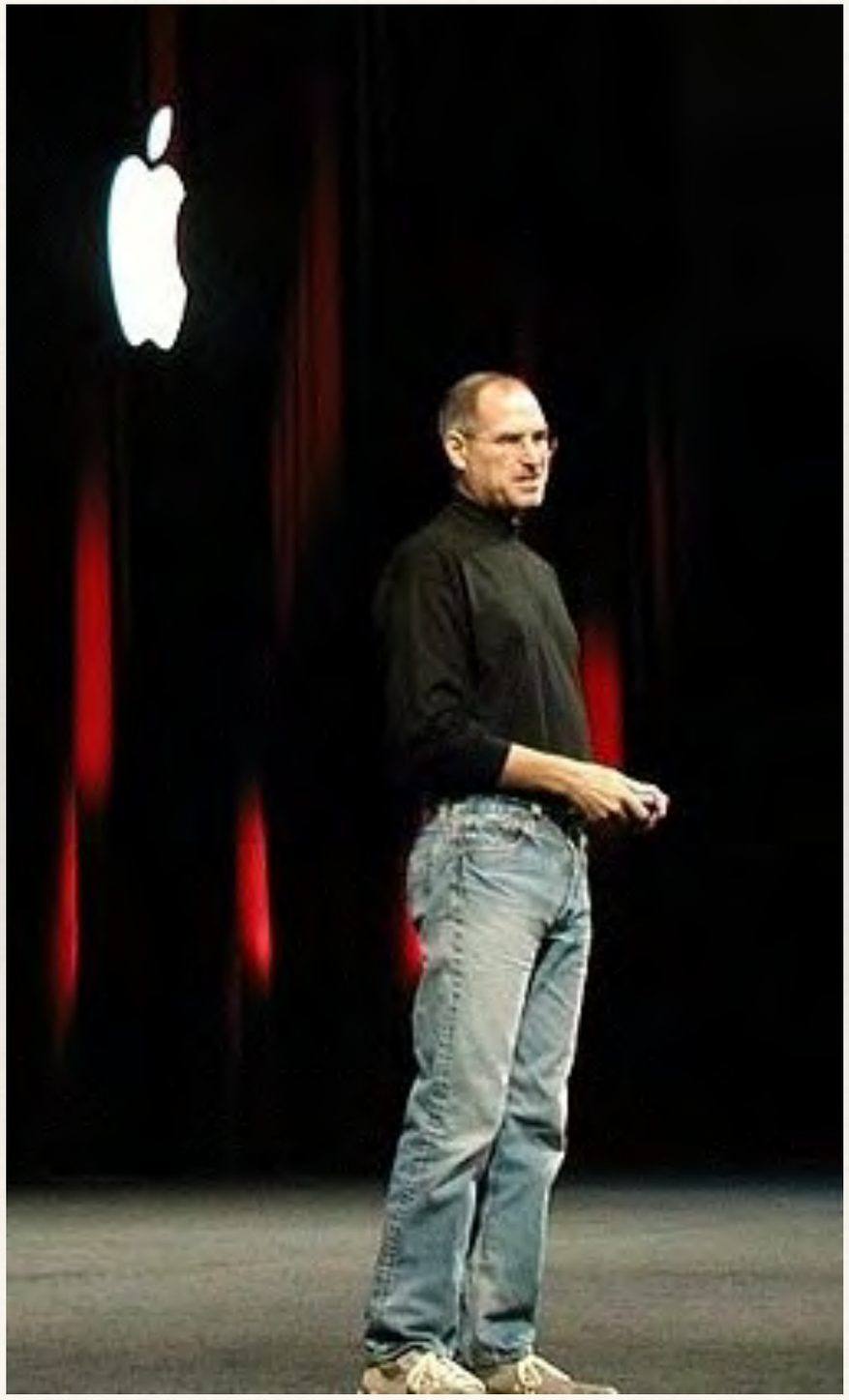
- Libre Office/Google Docs Spreadsheet
- Inkscape (for vector/svg editing)
- Gimp (for pixel editing)
- tableau (http://www.tableausoftware.com) free version
- Other python vis libraries: networkX, mayavi2 (3D), bokeh, seaborn, chaco, vincent, ggplot (python)
- Other Javascript libraries: three.js (3D), philogl (3d), processing.js, digraphs.js, polymaps.js, dimple.js
- R has ggplot2
- Gephi

Some Guiding Principles

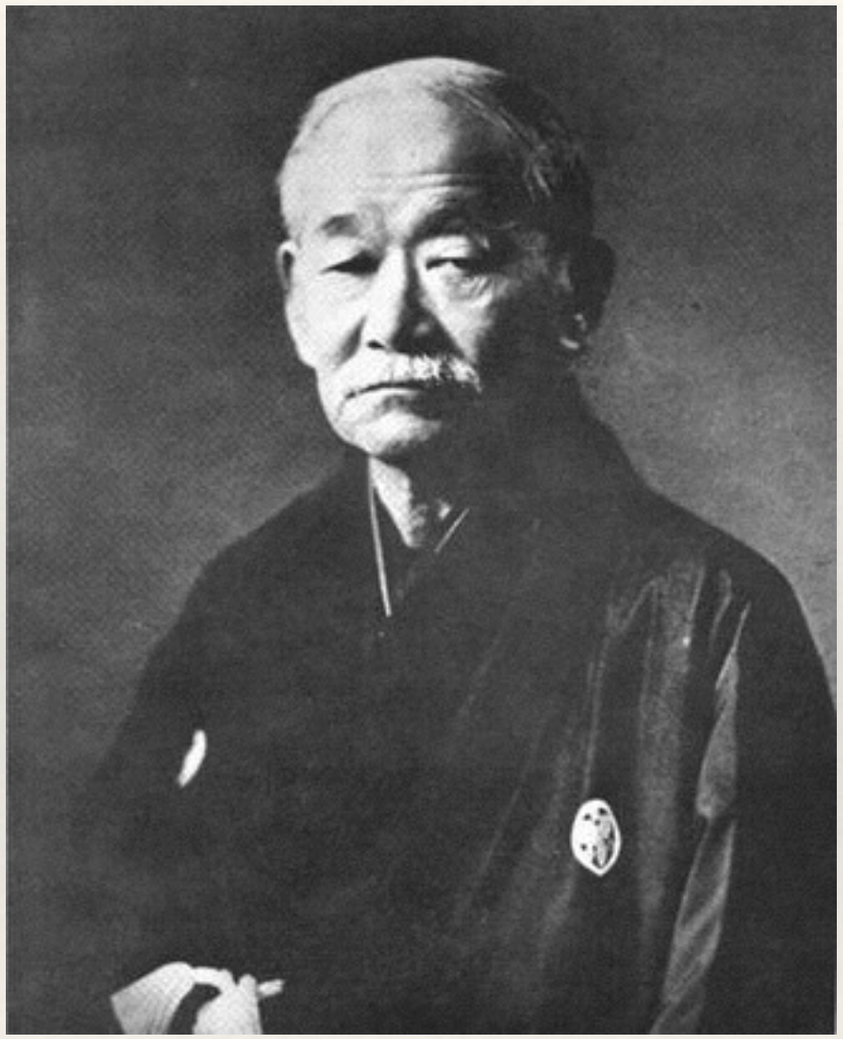
What to we mean by good design?

Design is a funny word.
Some people think design means how it looks. But of course, if you dig deeper, it's really how it works.

Steve Jobs



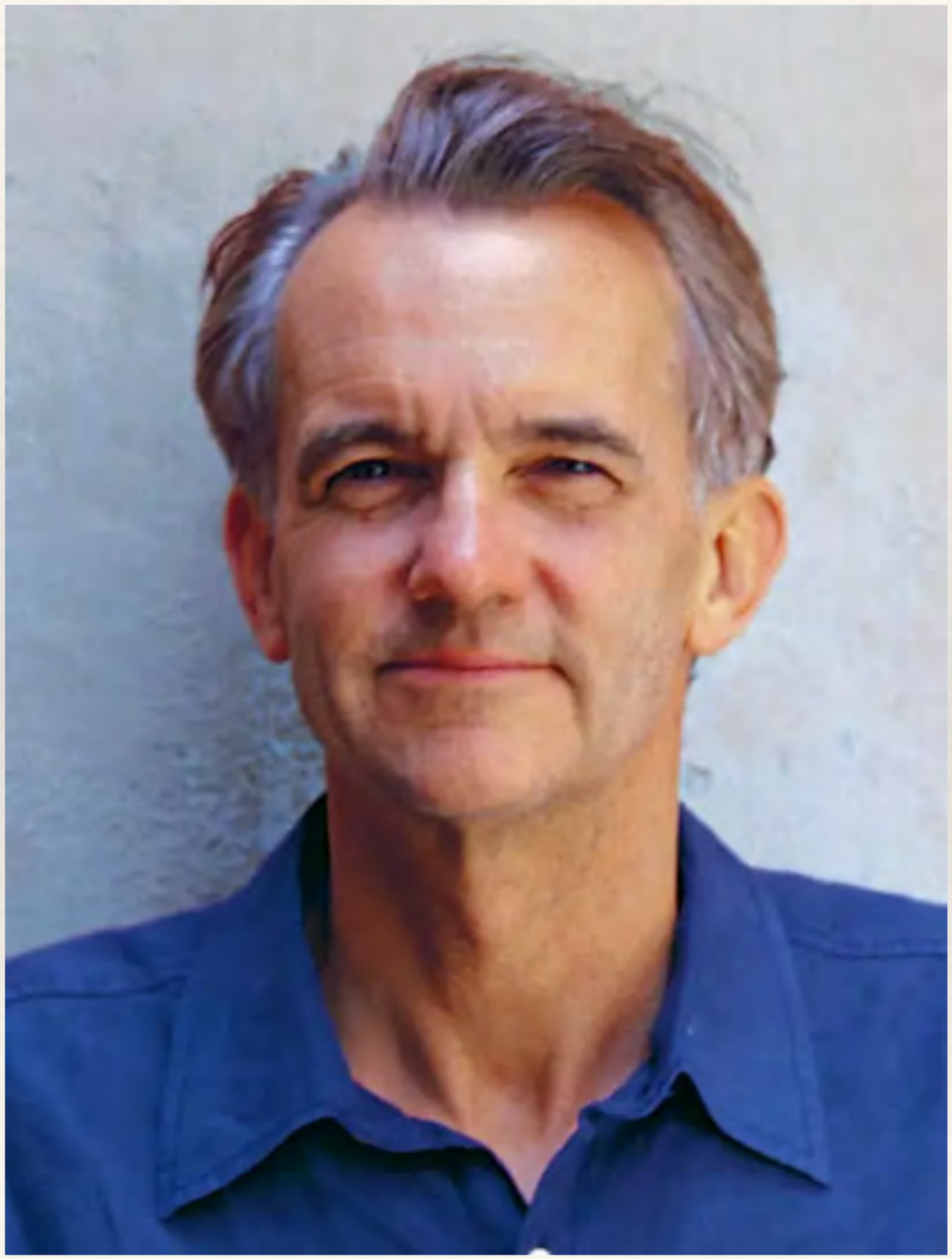
Attributes of good design



Judo Master: Kano Jigoro

Maximum Efficiency
with Minimum Effort

Edward Tufte



American Statistician

Pioneer

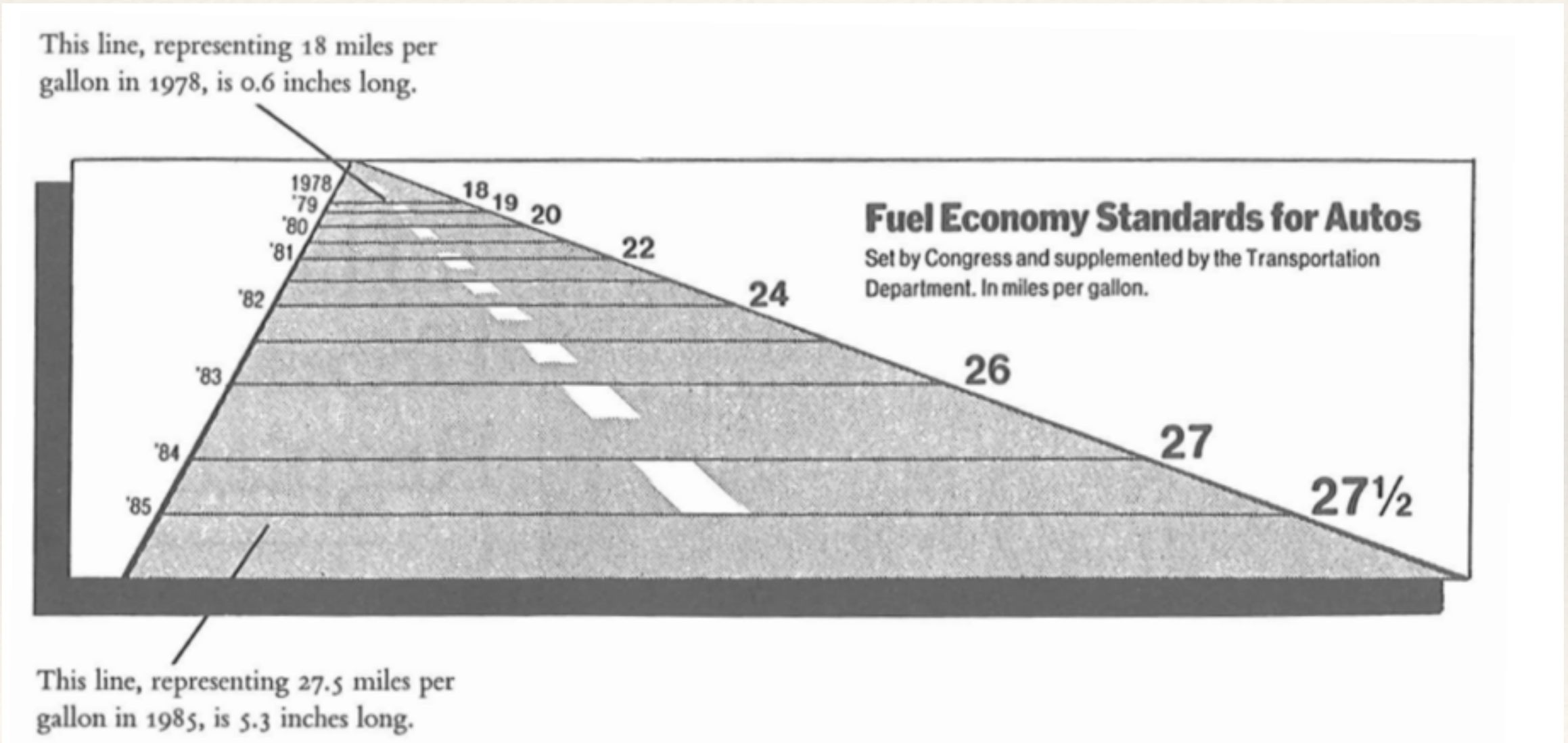
Can be controversial

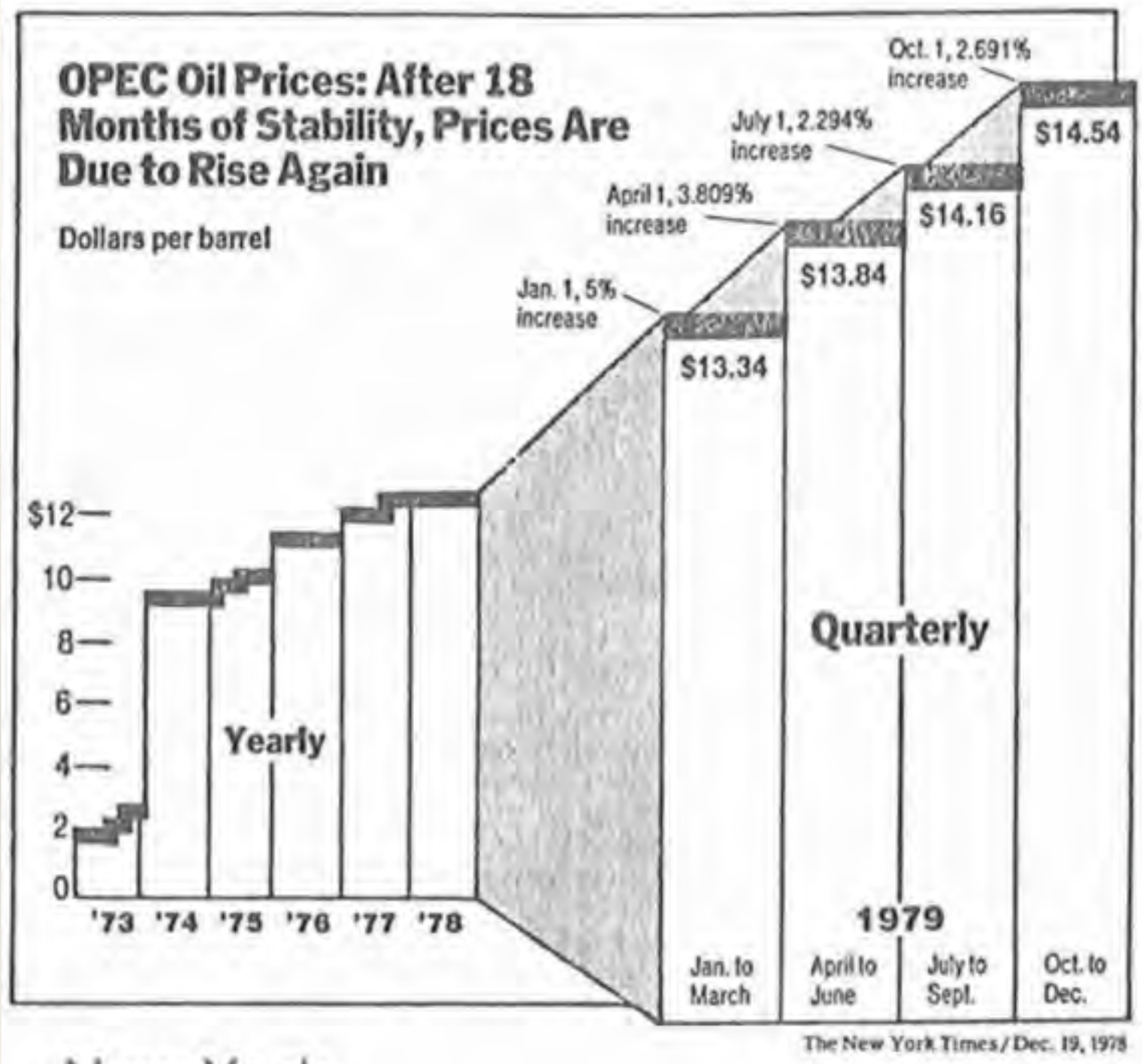
Hard to overstate importance

Principle Tufte: Graphical Integrity

Lie Factor

$$LieFactor = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$





During this time	one vertical inch equals
1973–1978	$8.00
January–March 1979	$4.73
April–June 1979	$4.37
July–September 1979	$4.16
October–December 1979	$3.92

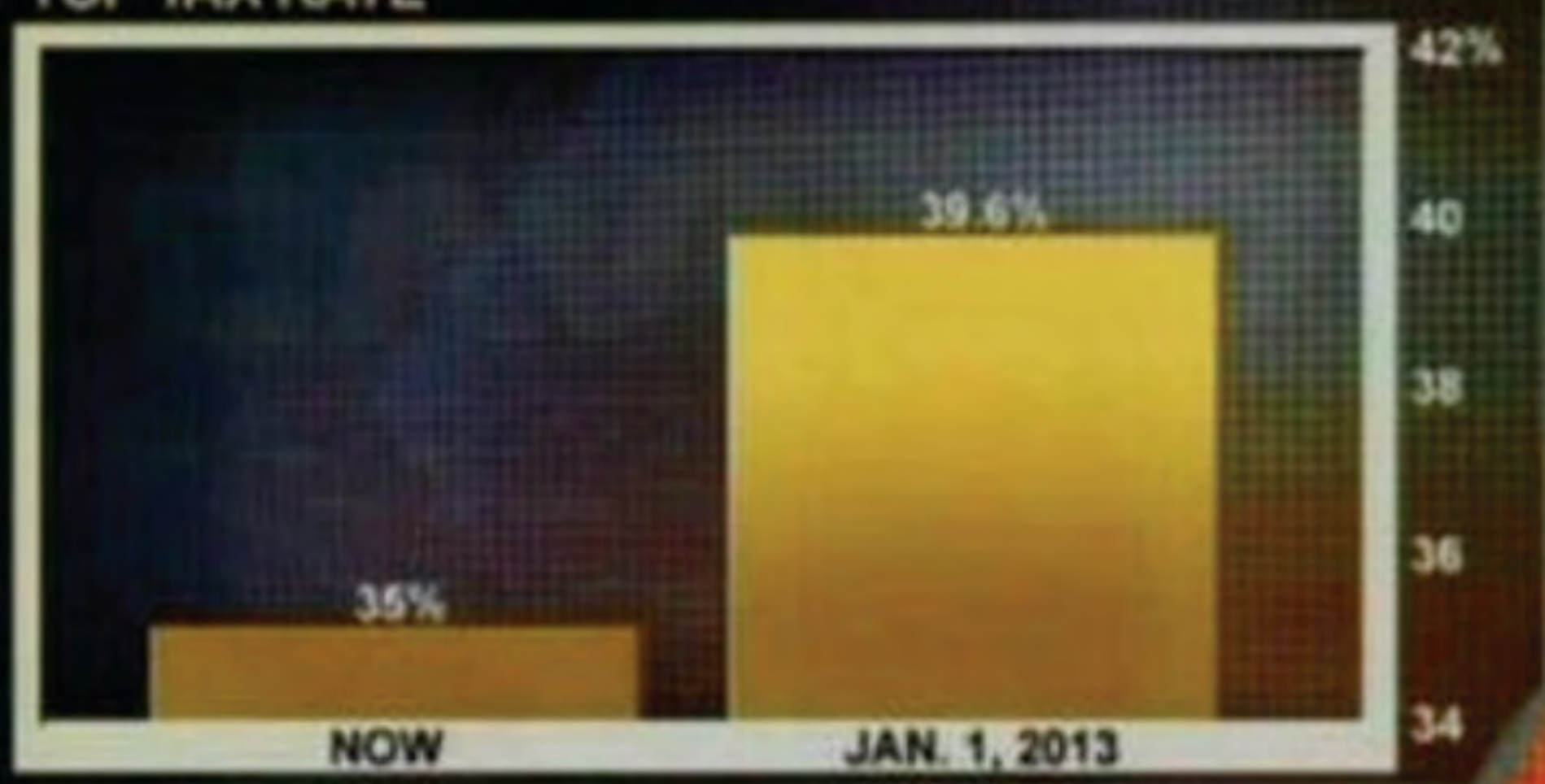
During this time	one horizontal inch equals
1973–1978	3.8 years
1979	0.57 years

- design variation

$$LieFactor = 15.1$$

New York

IF BUSH TAX CUTS EXPIRE



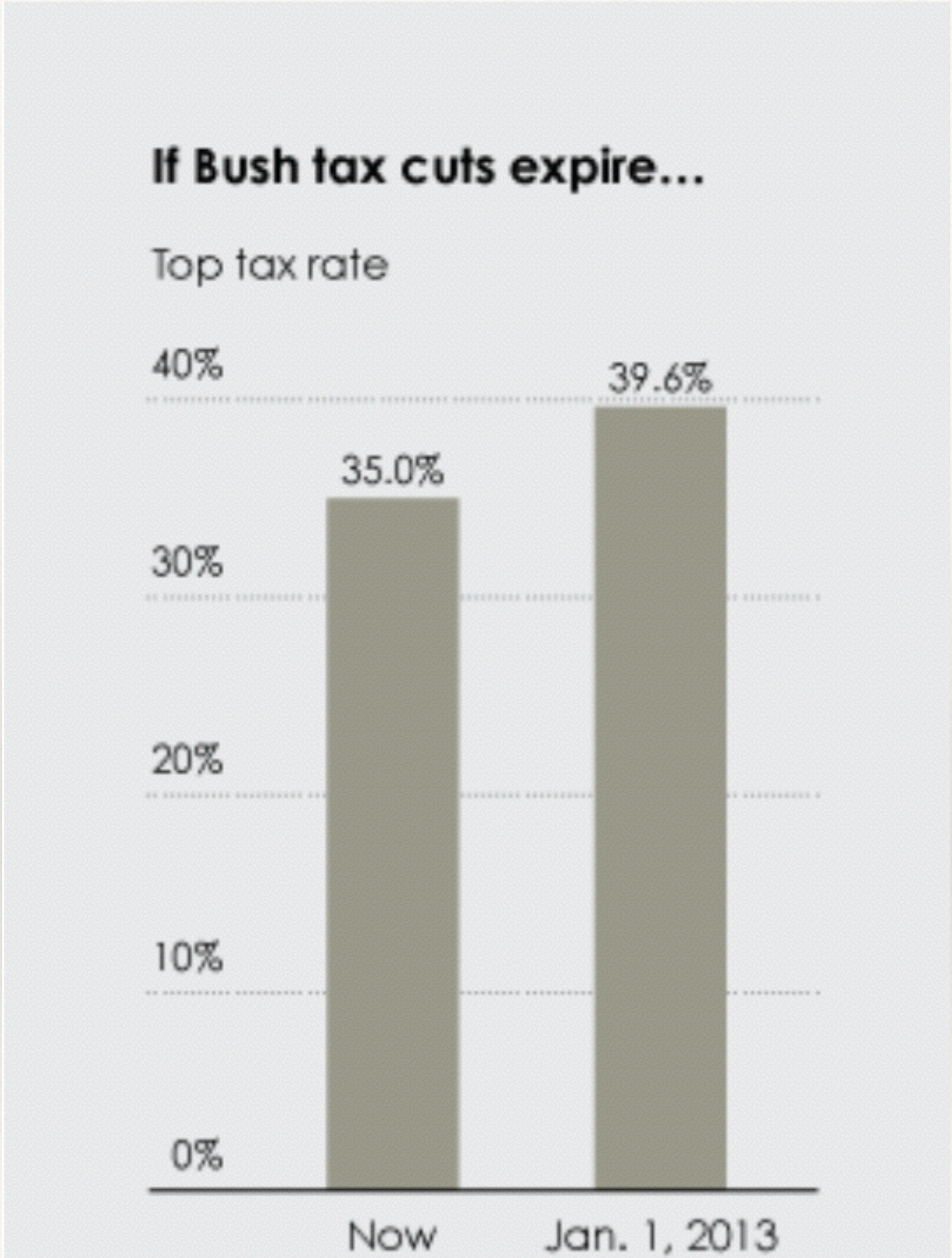
8:01p ET

FOX BUSINESS

TOP STORIES | TECHNOLOGY | CONSUMER

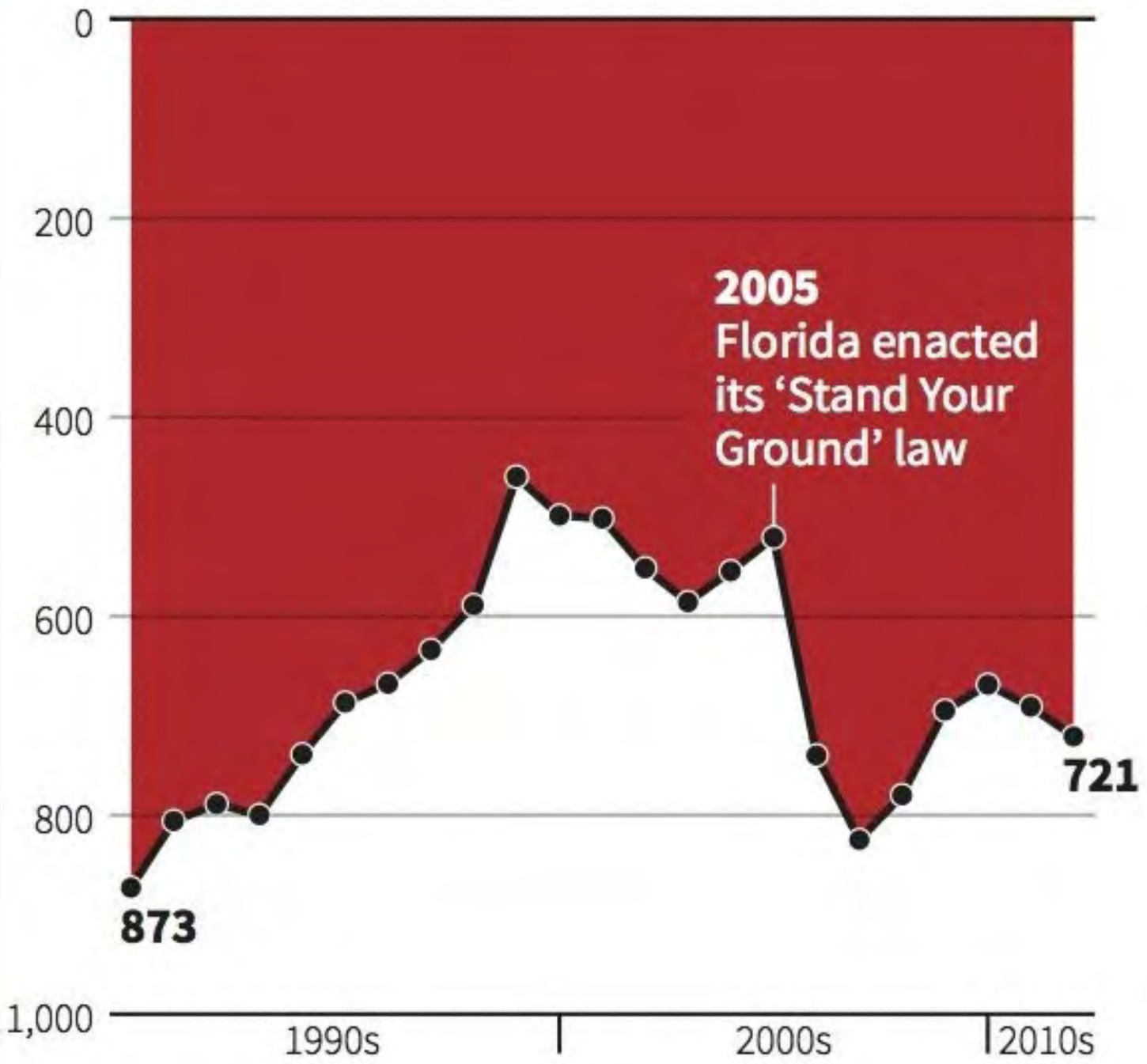
VITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T

DOW 13008.68 ▼ 64.33 | S&P 1379.32 ▼ 5.98 | NASDAQ 2939.52 ▼ 6.32


If Bush tax cuts expire...
Top tax rate
40%
30%
20%
10%
0%
35.0%
39.6%
Now
Jan. 1, 2013

Gun deaths in Florida

Number of murders committed using firearms

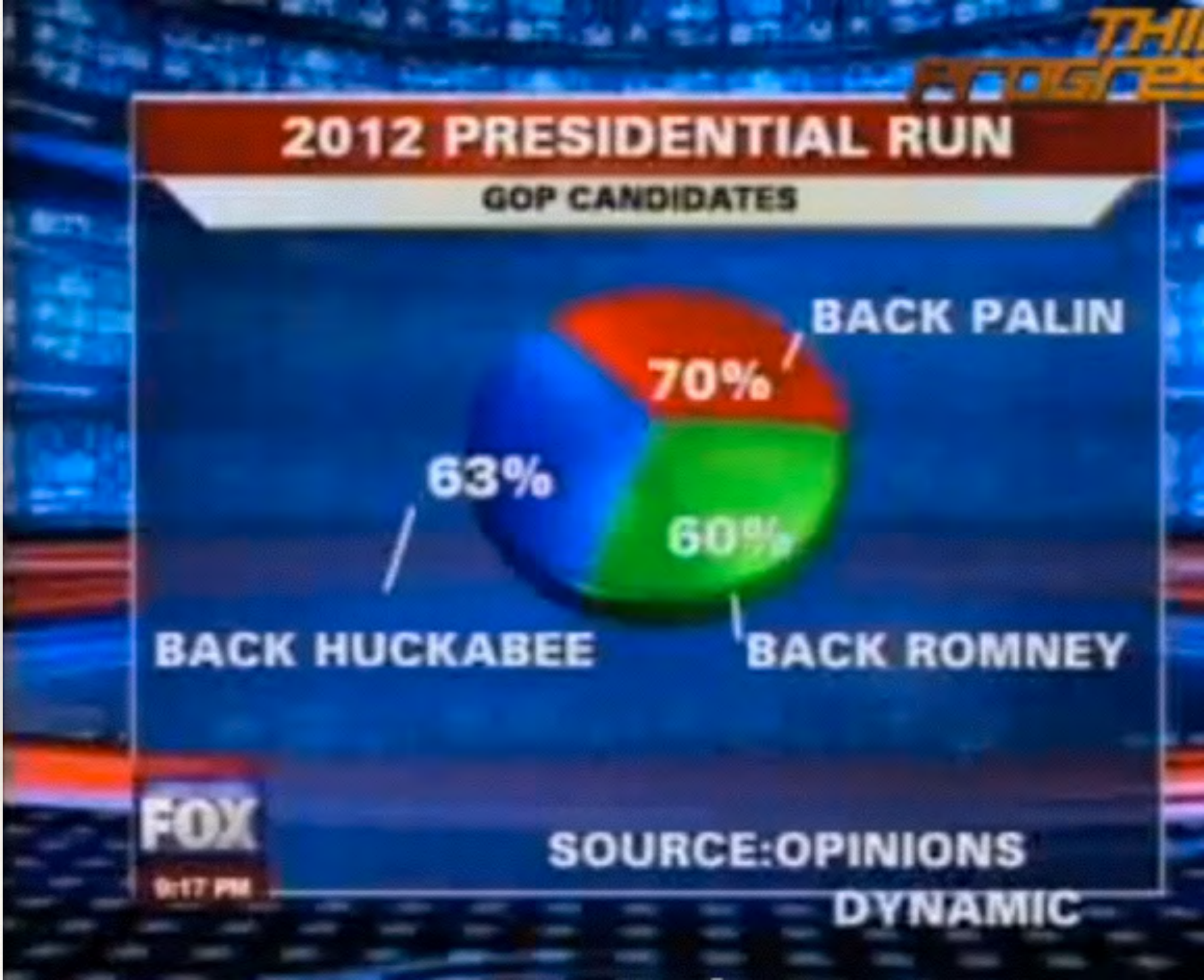


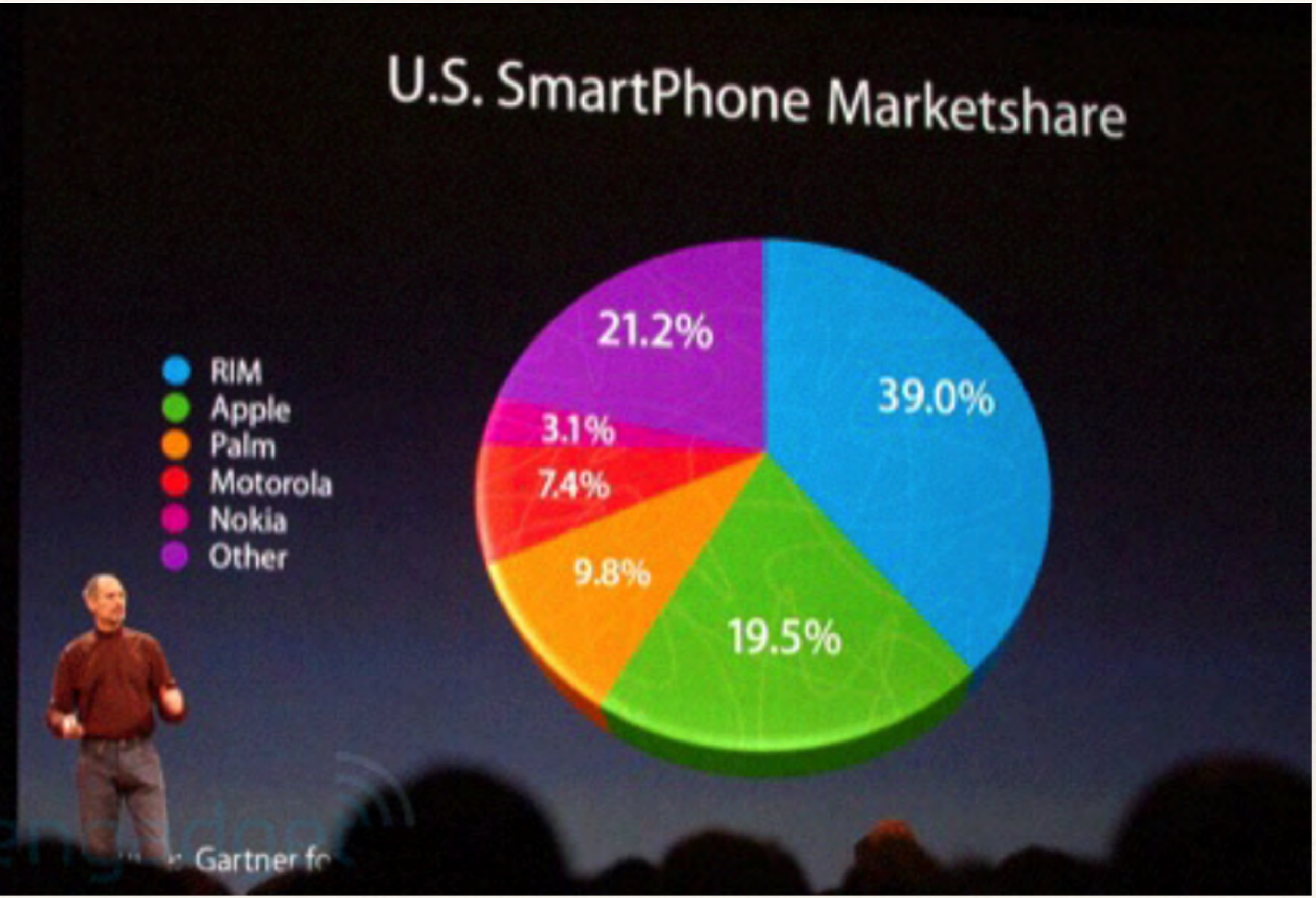
Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

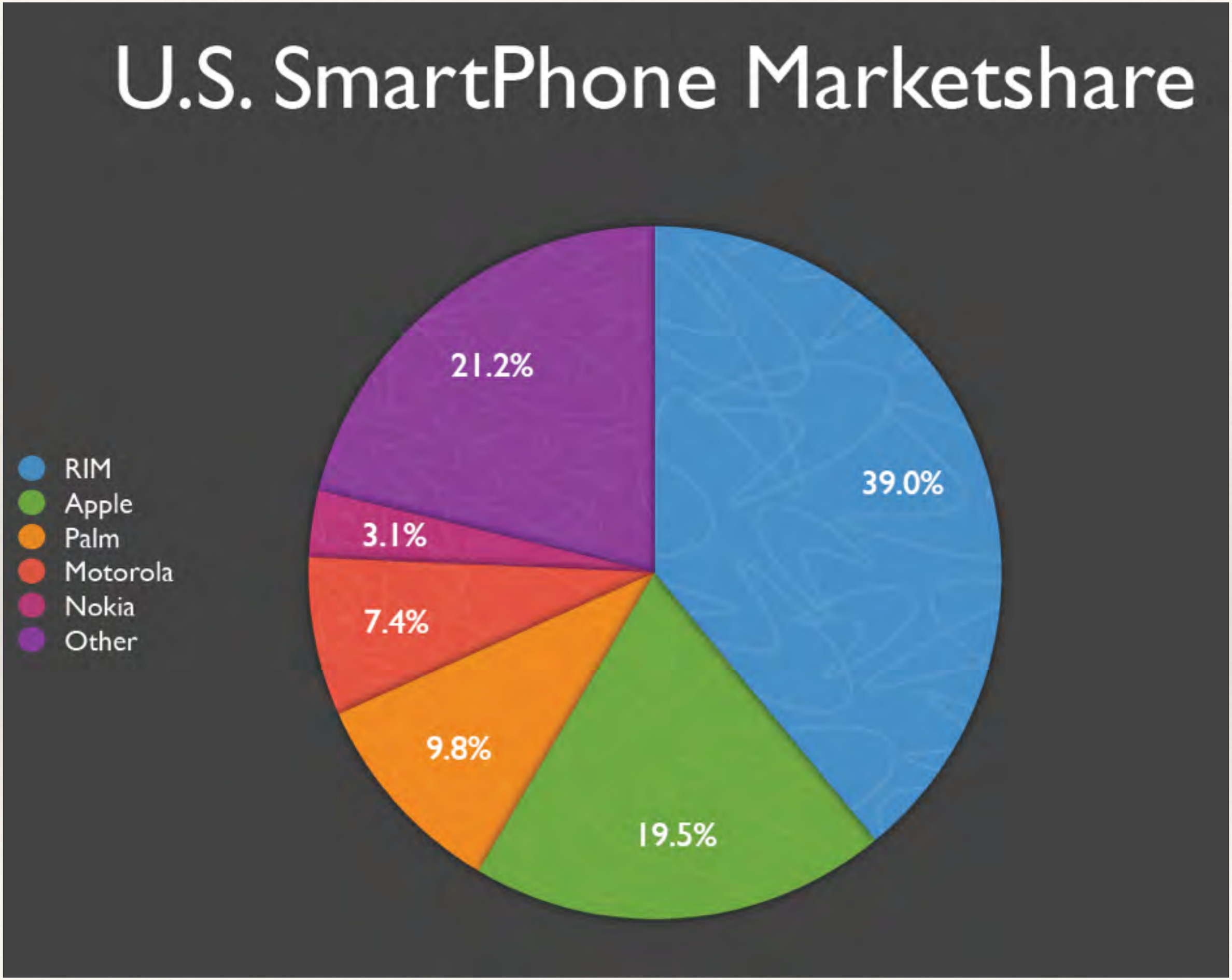
REUTERS

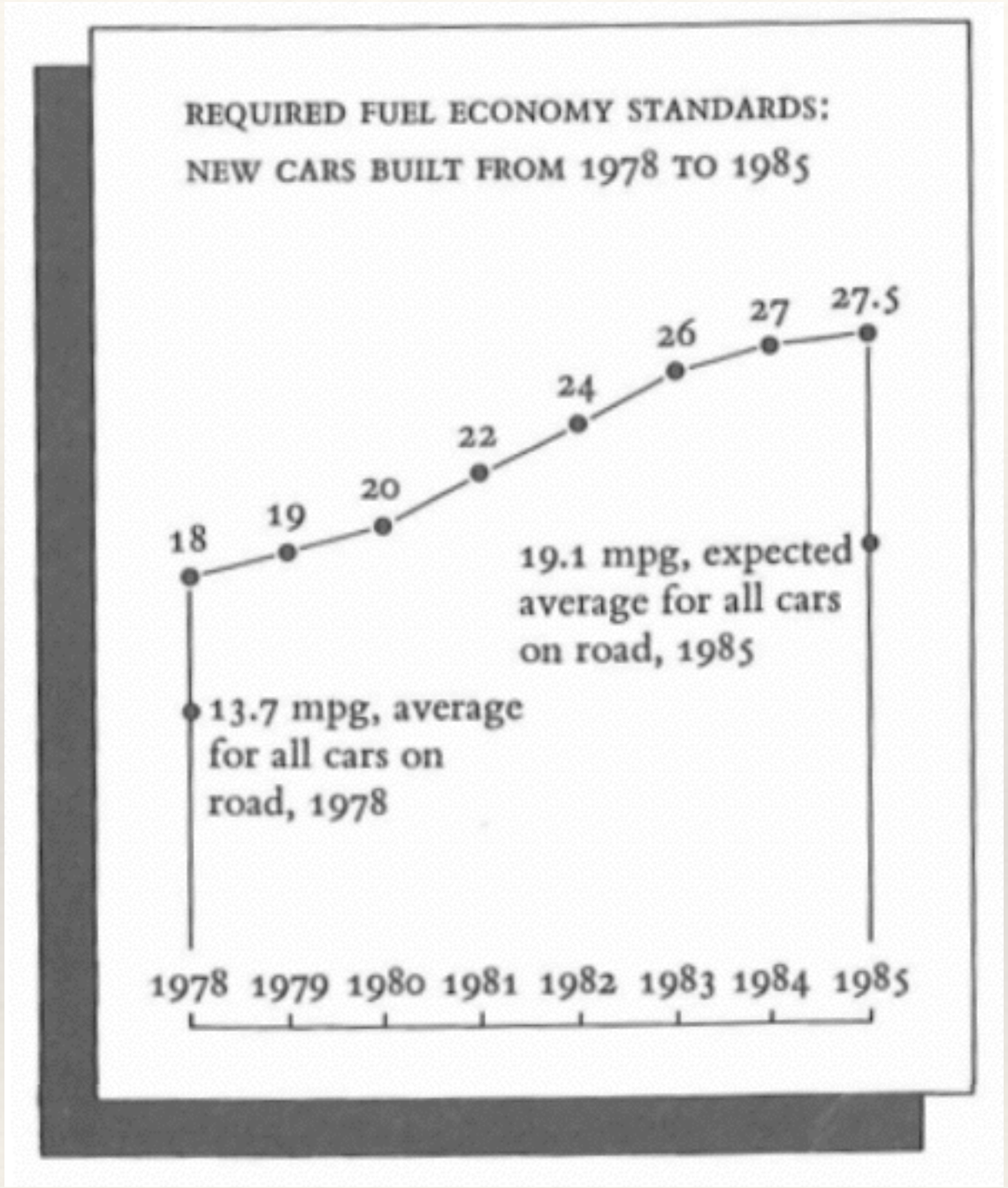
And then there are pie charts


2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
9:17 PM
SOURCE:OPINIONS DYNAMIC

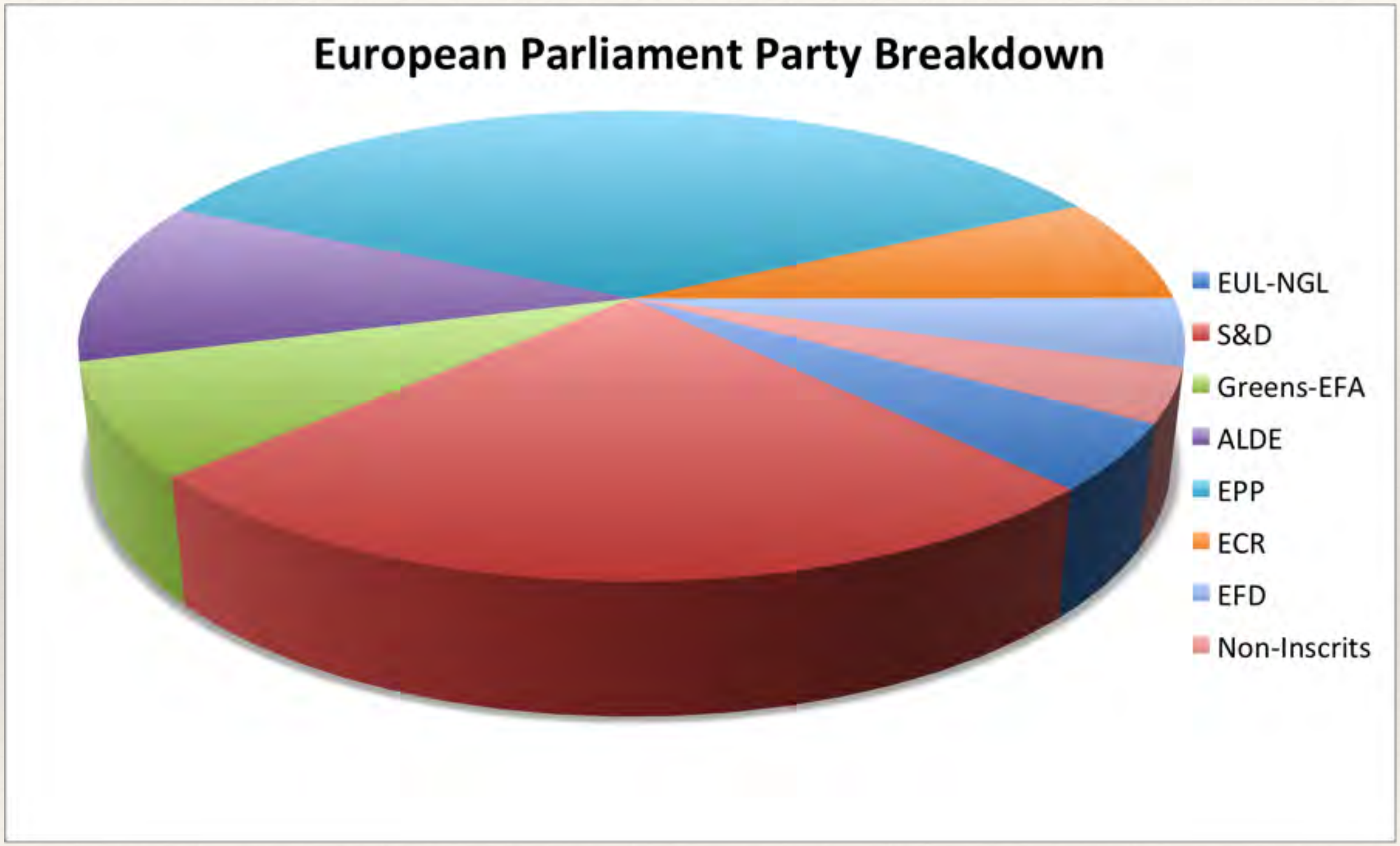

U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
21.2%
39.0%
3.1%
7.4%
9.8%
19.5%
Gartner fo

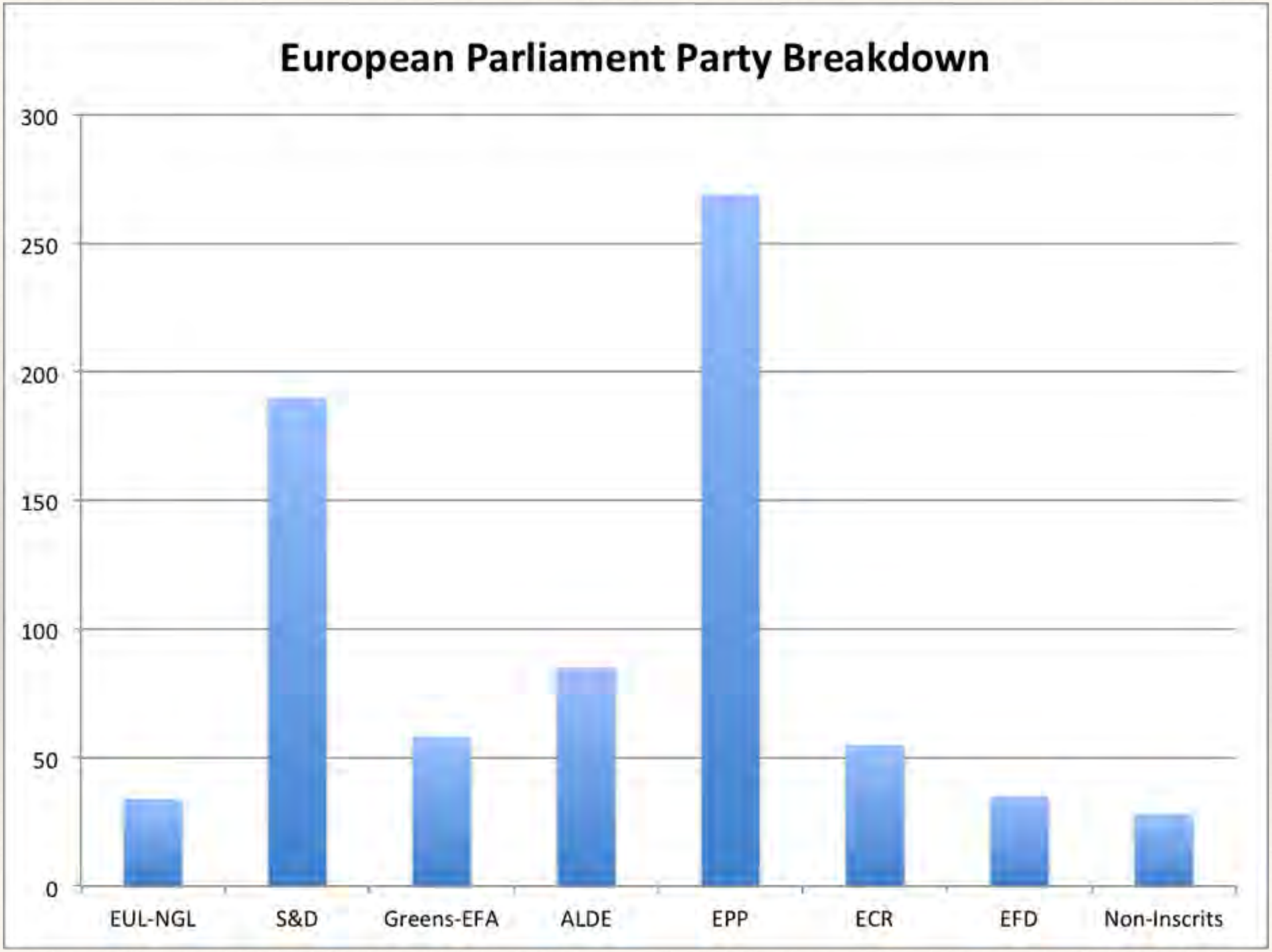
U.S. SmartPhone Marketshare




REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985
18
19
20
22
24
26
27
27.5
19.1 mpg, expected average for all cars on road, 1985
13.7 mpg, average for all cars on road, 1978
1978 1979 1980 1981 1982 1983 1984 1985

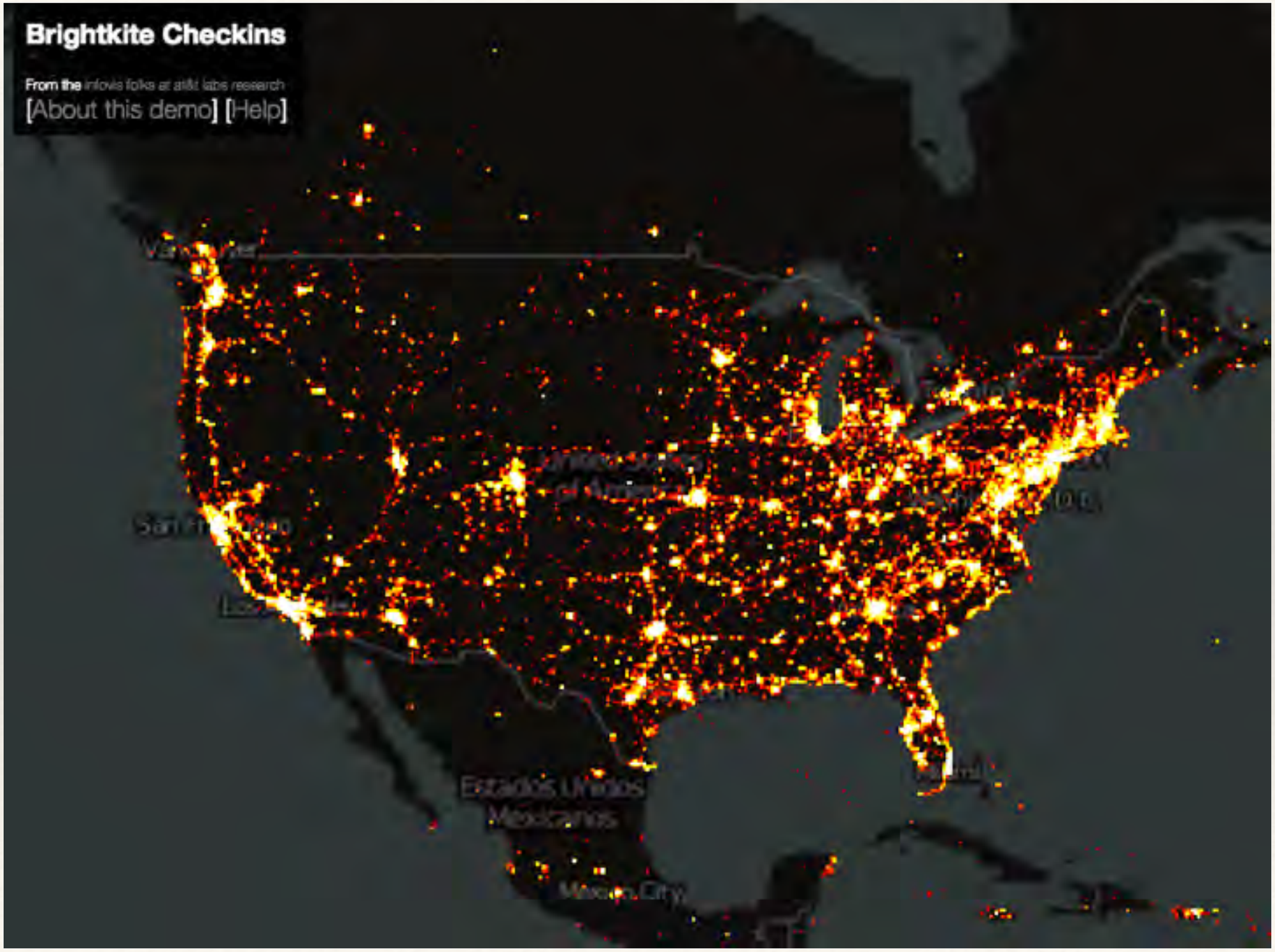
3D adds to Extra Distortion



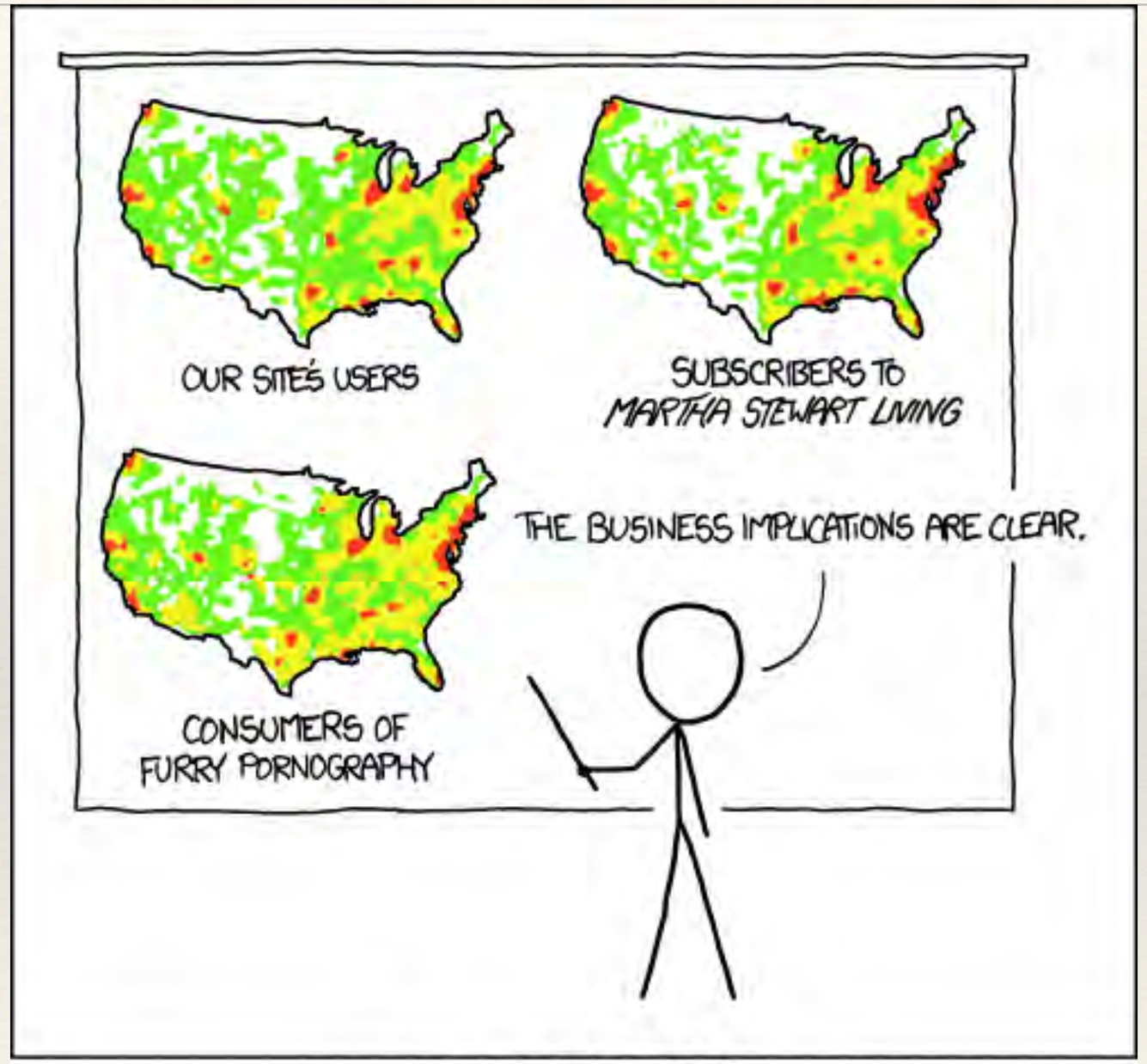

European Parliament Party Breakdown
300
250
200
150
100
50
0
EUL-NGL
S&D
Greens-EFA
ALDE
EPP
ECR
EFD
Non-Inscrits

More Baloney than Lies

nanocubes.net

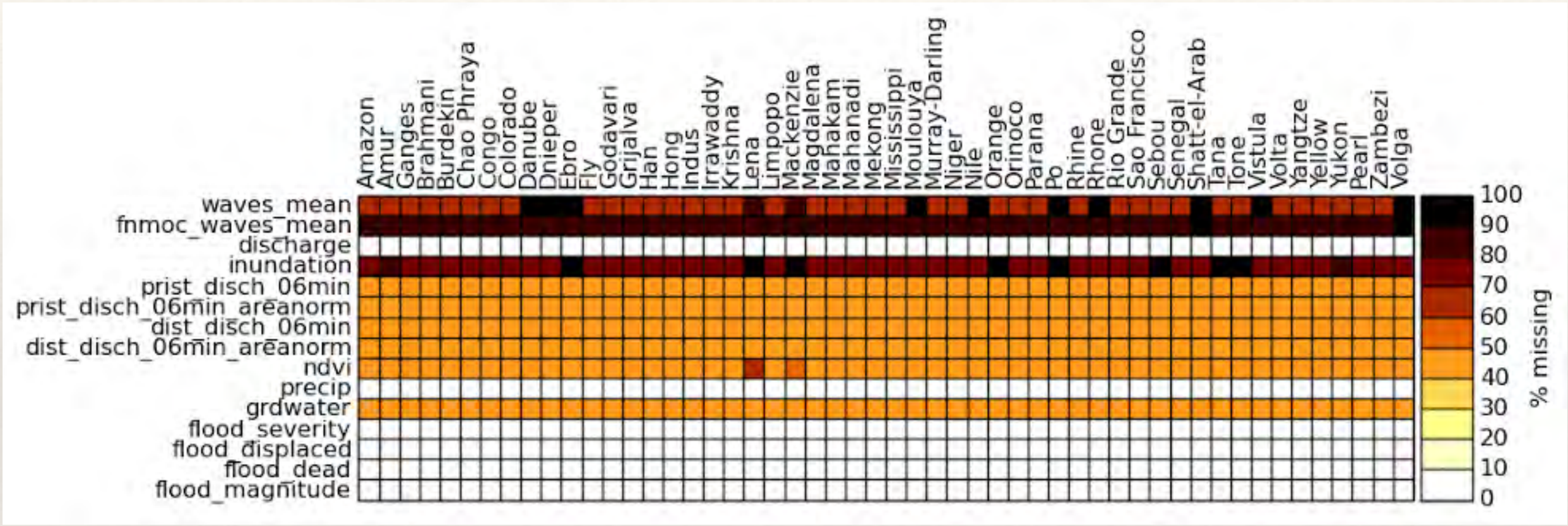


Actually Content Free



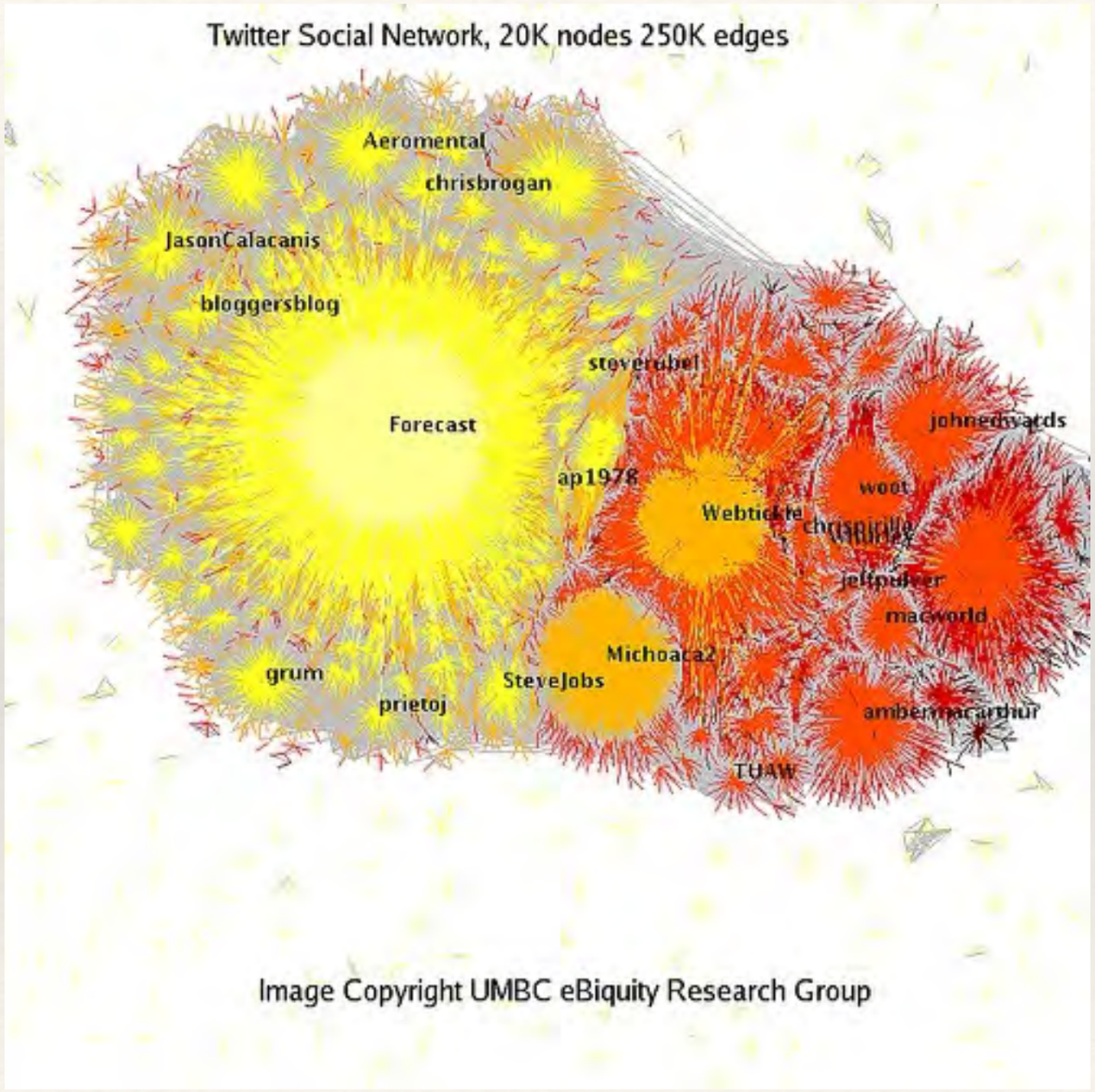
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Missing Data is also a Problem



Some hydrology data we were working with.

Learning from Social Networks



http://ebiquity.umbc.edu/blogger/2007/04/19/twitter-social-network-analysis/

Problems with Social Network Data

Daily Mail .com

Home | U.K. | News | Sports | U.S. Showbiz | Australia | Femail | Health | Science | Mor

Science Home | Pictures


America encased in ice: Stunning shot


I'm gay. And I want my kid to be gay

Man's fury at United Airlines after flight


Chris Christie failing behind

Rise of the Twitter bots: Social network admits 23 MILLION of its users tweet automatically without human input

- Twitter now has more than 270 million users who actively log in and tweet

THE NEW YORKER

NEWS | CULTURE | BOOKS & FICTION | SCIENCE & TECH | BUSINESS | HUMOR | MAGAZINE | VIDEO

NOVEMBER 14, 2013

THE RISE OF TWITTER BOTS

BY ROB DUBBIN

Share | Tweet | +1

Last Tuesday, Google decided that I was a spammer, and I lost access to my e-mail for twelve hours. It was my fault. One of my Twitter accounts, RealHumanPraise, was mentioned on "The Colbert Report," where I work as a writer, at 11:46 P.M. In the course of the next hundred and twenty seconds, it

Bits

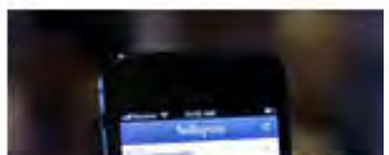
SOCIAL

Millions of Fake Instagram Users Disappe

By VINDU GOEL DECEMBER 18, 2014 8:59 PM

Email

Social media services like Facebook and Twitter are always



CNN

U.S. Edition | New York City, NY 41° | Sig

News | Video | TV | Opinions | More... | Search CNN

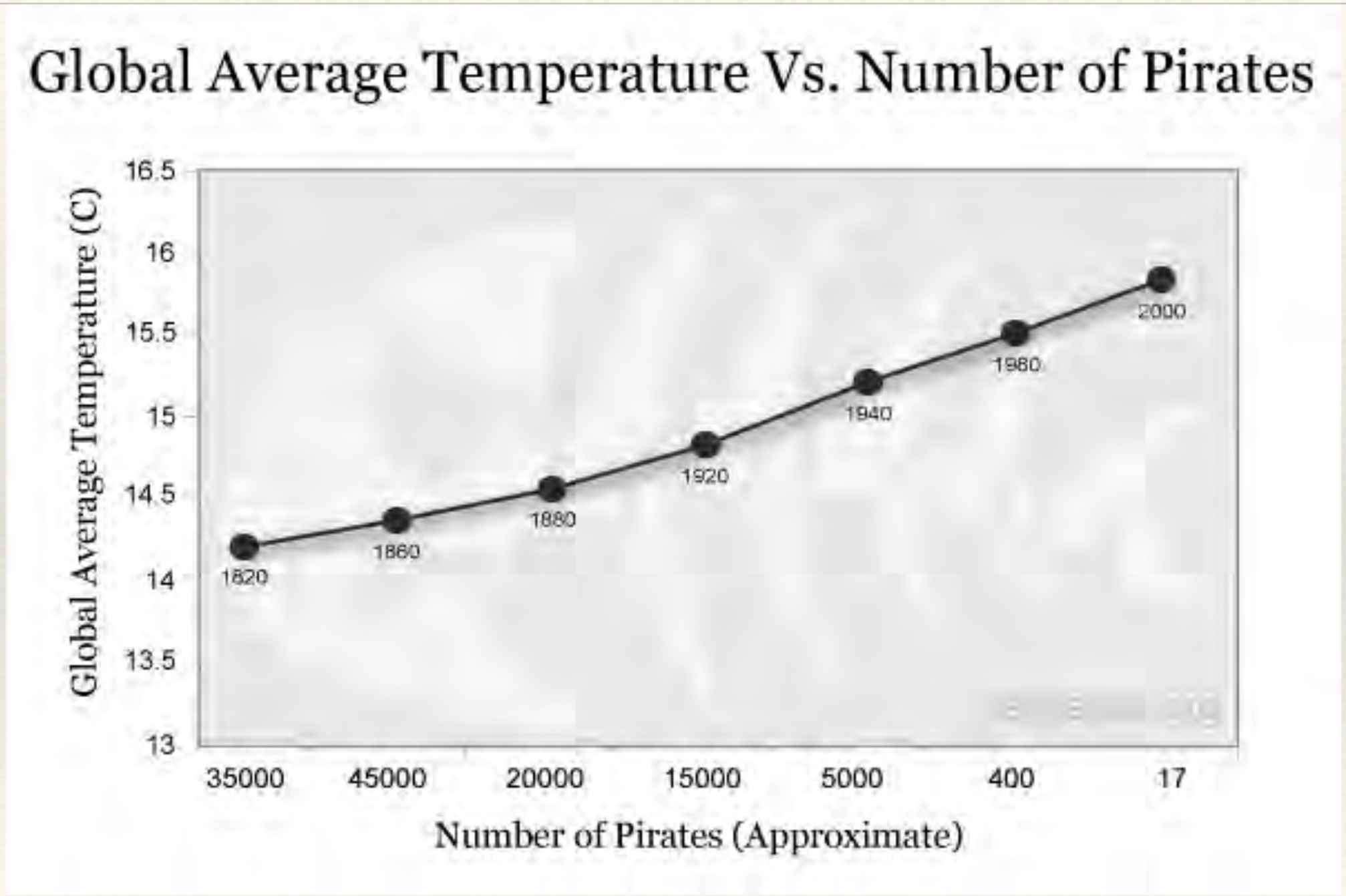
U.S. | World | Politics | Tech | Health | Entertainment | Living | Travel | Money | Sports

83 million Facebook accounts are fakes and dupes

By Heather Kelly, CNN

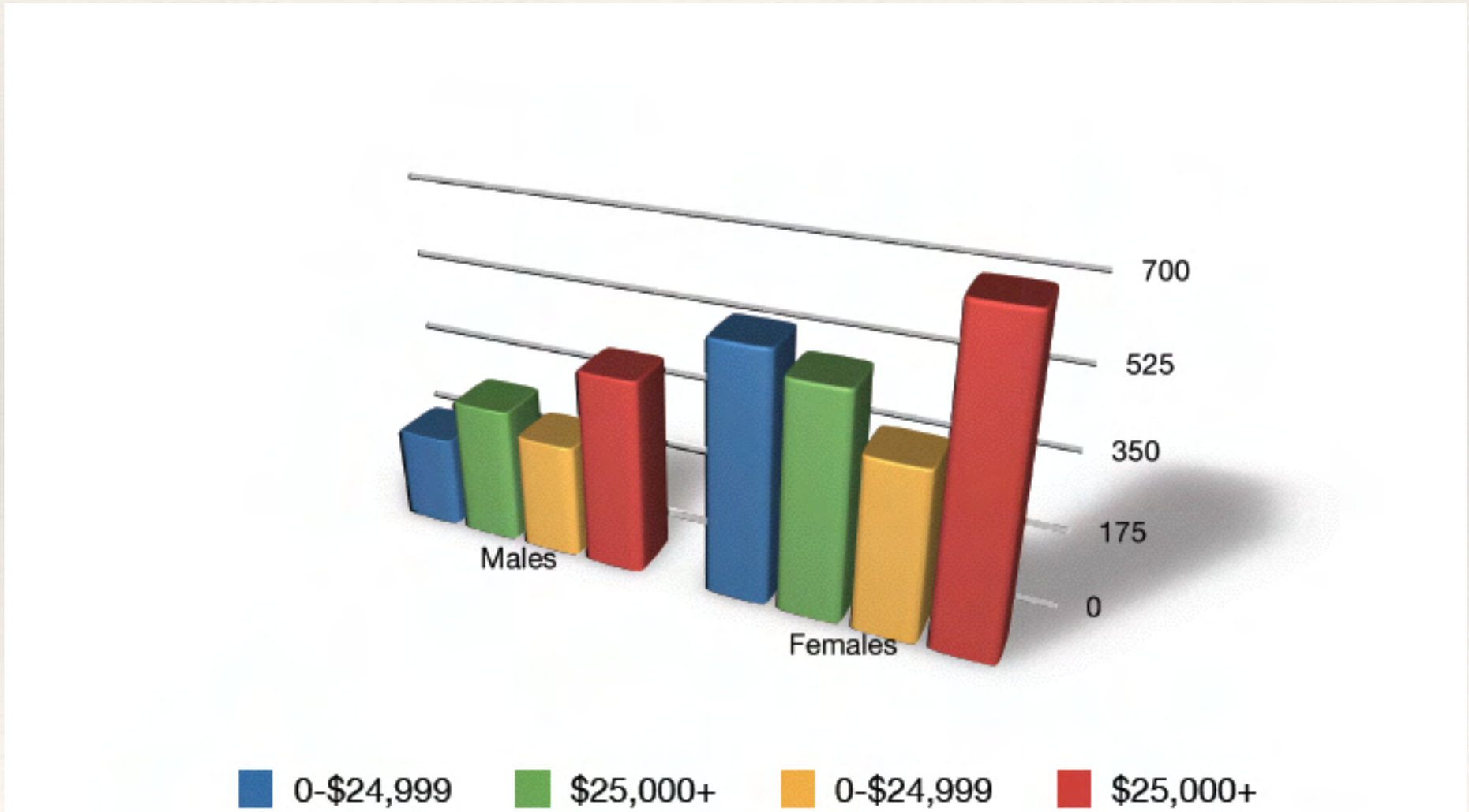
Updated 5:27 AM ET, Fri August 3, 2012

Numbers don't Lie?



Maximize: Data to Ink Ratio

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



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

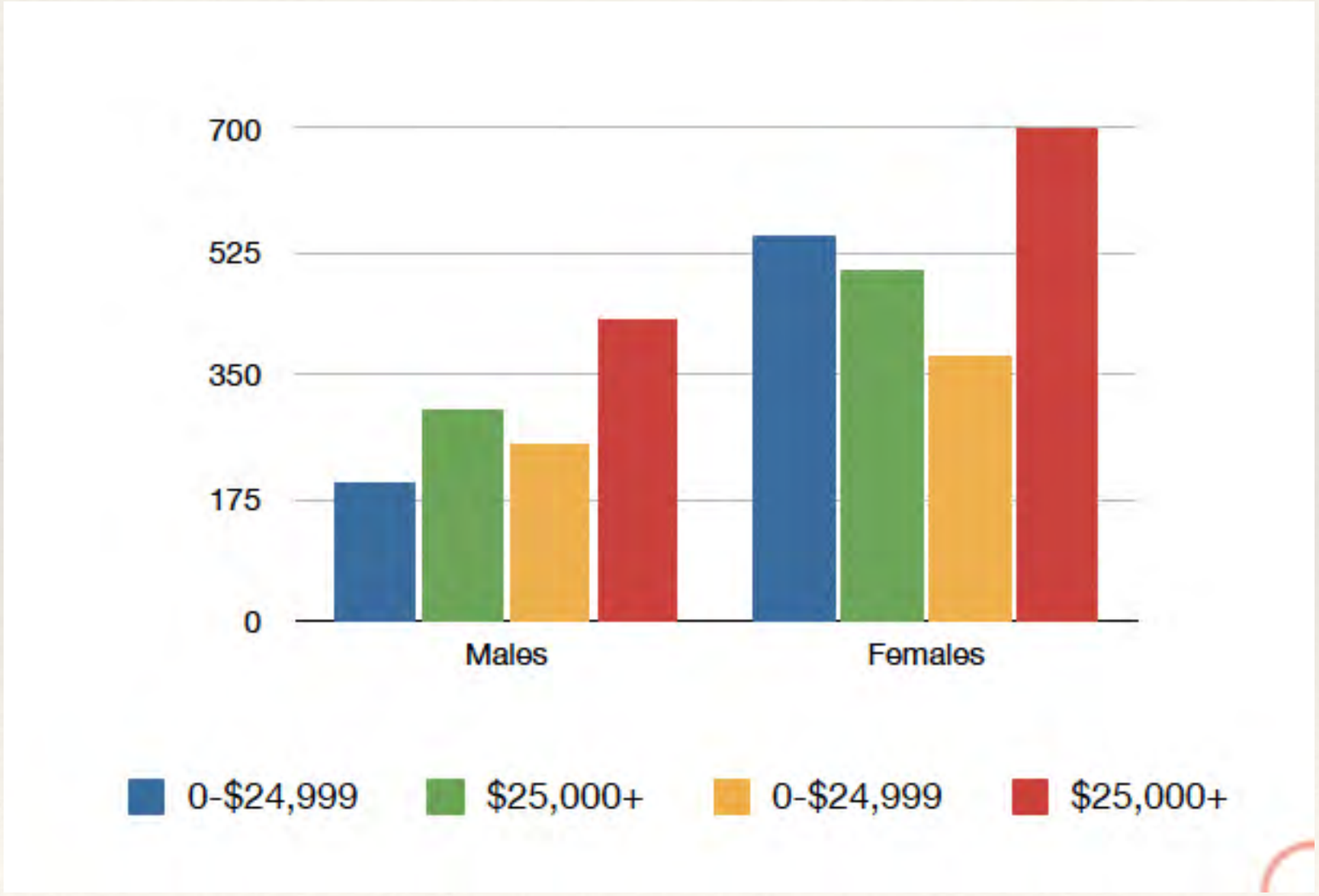
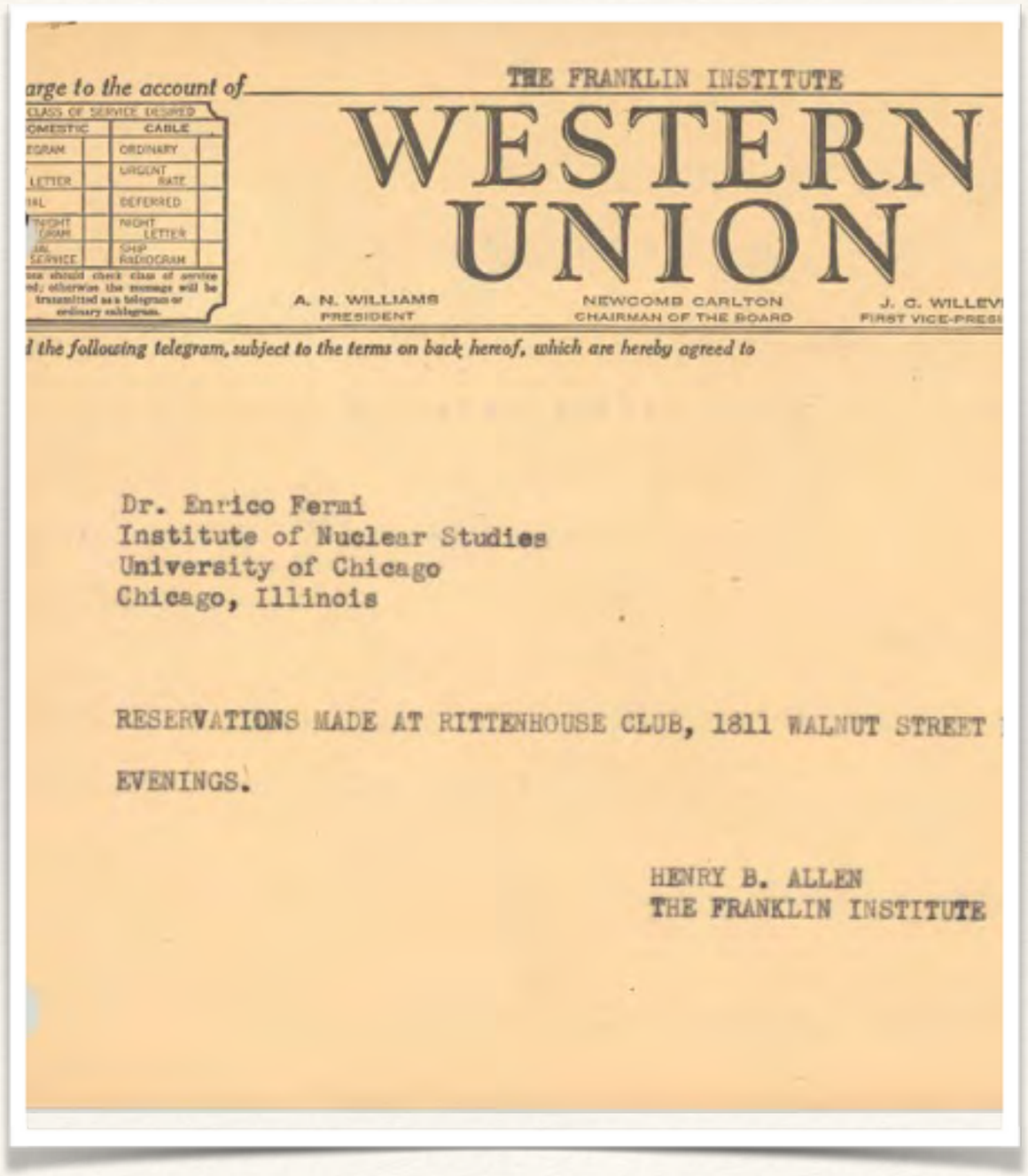


Chart JUNK

If you paid for decoration?

A designer knows he has achieved perfection not when there is nothing left to add, but when there is nothing left to take away.

Antoine de Saint-Exupery



arge to the account of

THE FRANKLIN INSTITUTE

CLASS OF SERVICE DESIRED

DOMESTIC		CABLE	
TELEGRAM		ORDINARY	
LETTER		URGENT RATE	
RL		DEFERRED	
NIGHT TELEGRAM		NIGHT LETTER	
SERVICE		SHIP RADIOGRAM	

WESTERN UNION

A. N. WILLIAMS, PRESIDENT — NEWCOMB CARLTON, CHAIRMAN OF THE BOARD — J. C. WILLEY, FIRST VICE-PRES

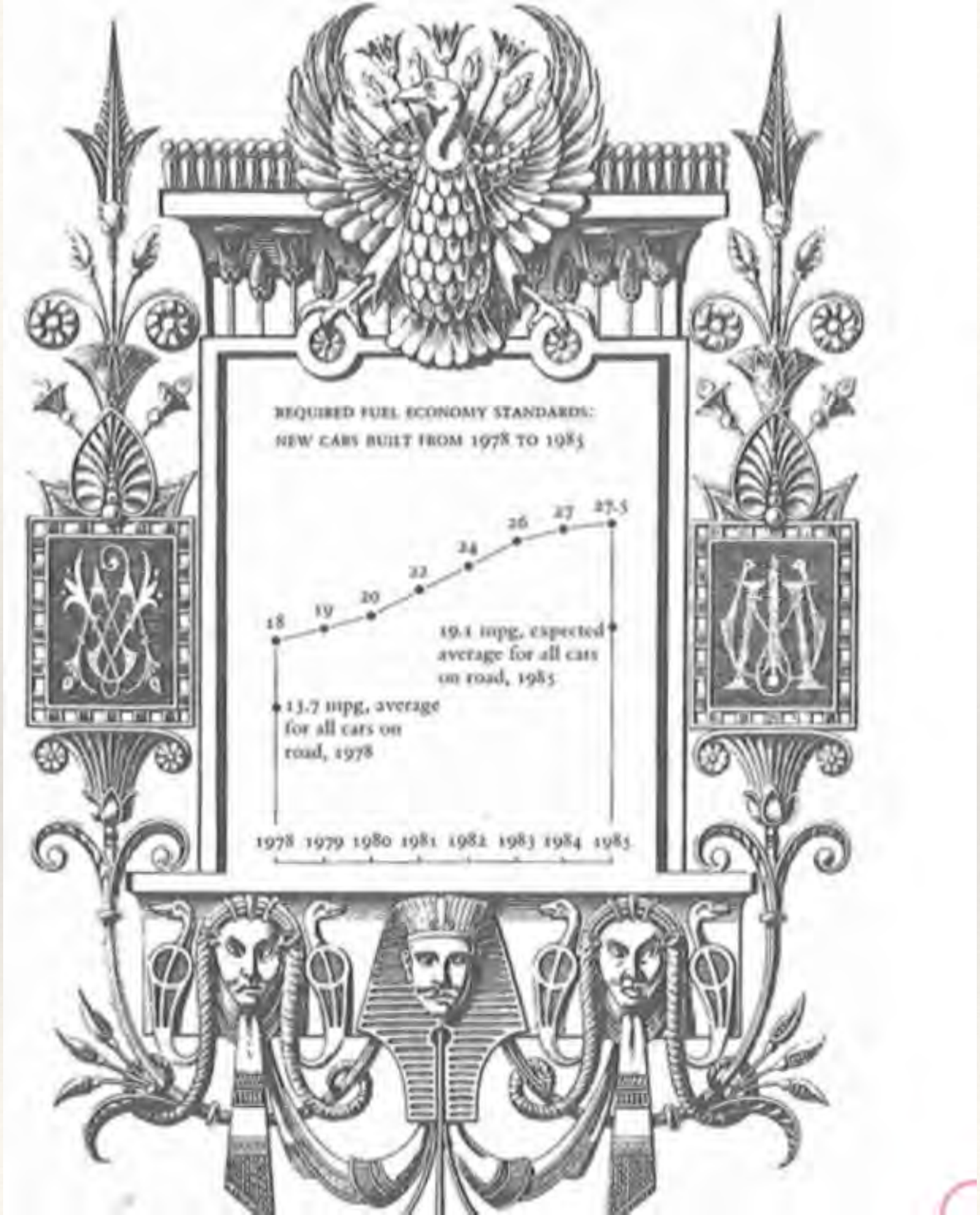
l the following telegram, subject to the terms on back hereof, which are hereby agreed to

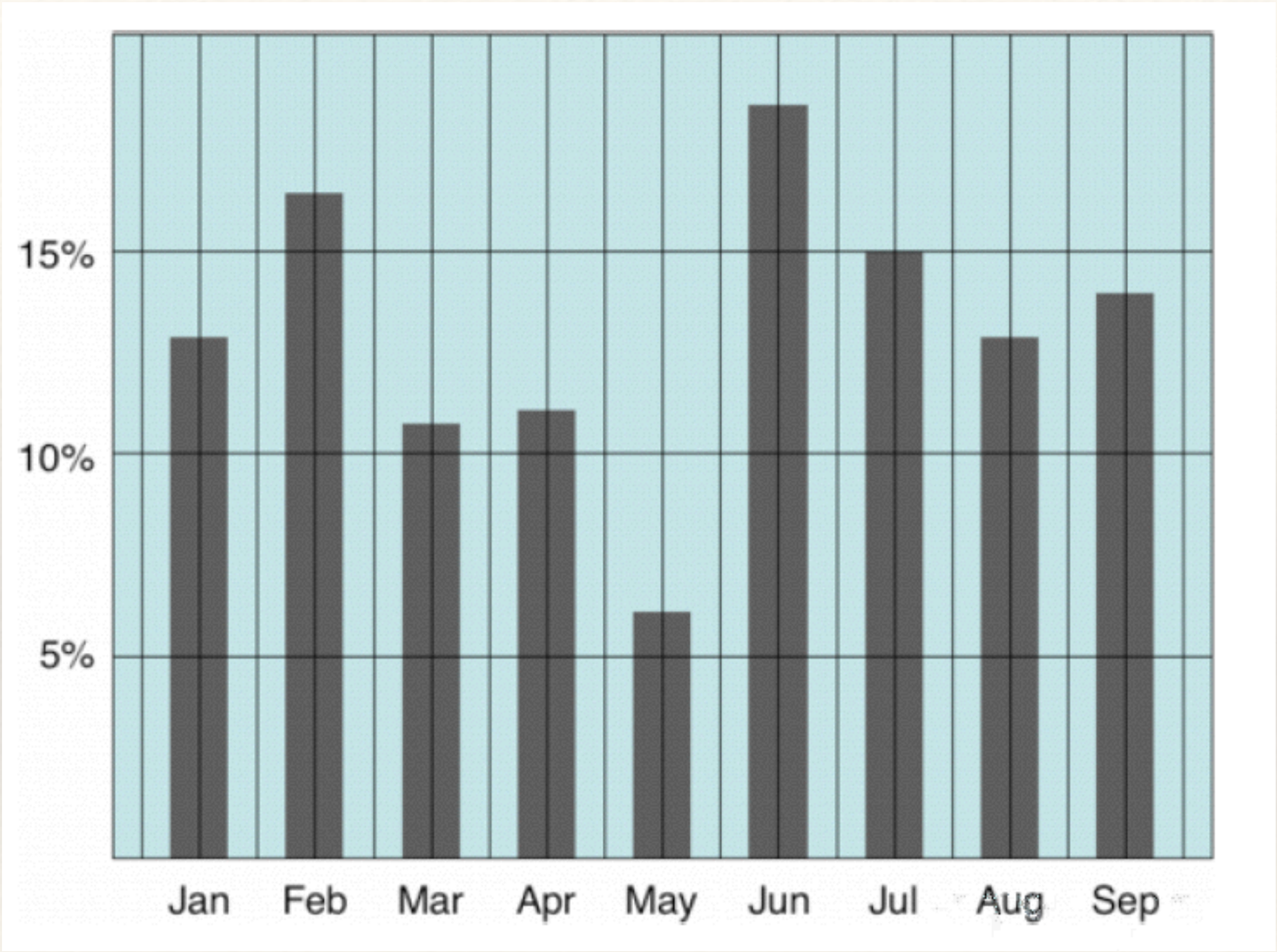
Dr. Enrico Fermi
Institute of Nuclear Studies
University of Chicago
Chicago, Illinois

RESERVATIONS MADE AT RITTENHOUSE CLUB, 1811 WALNUT STREET

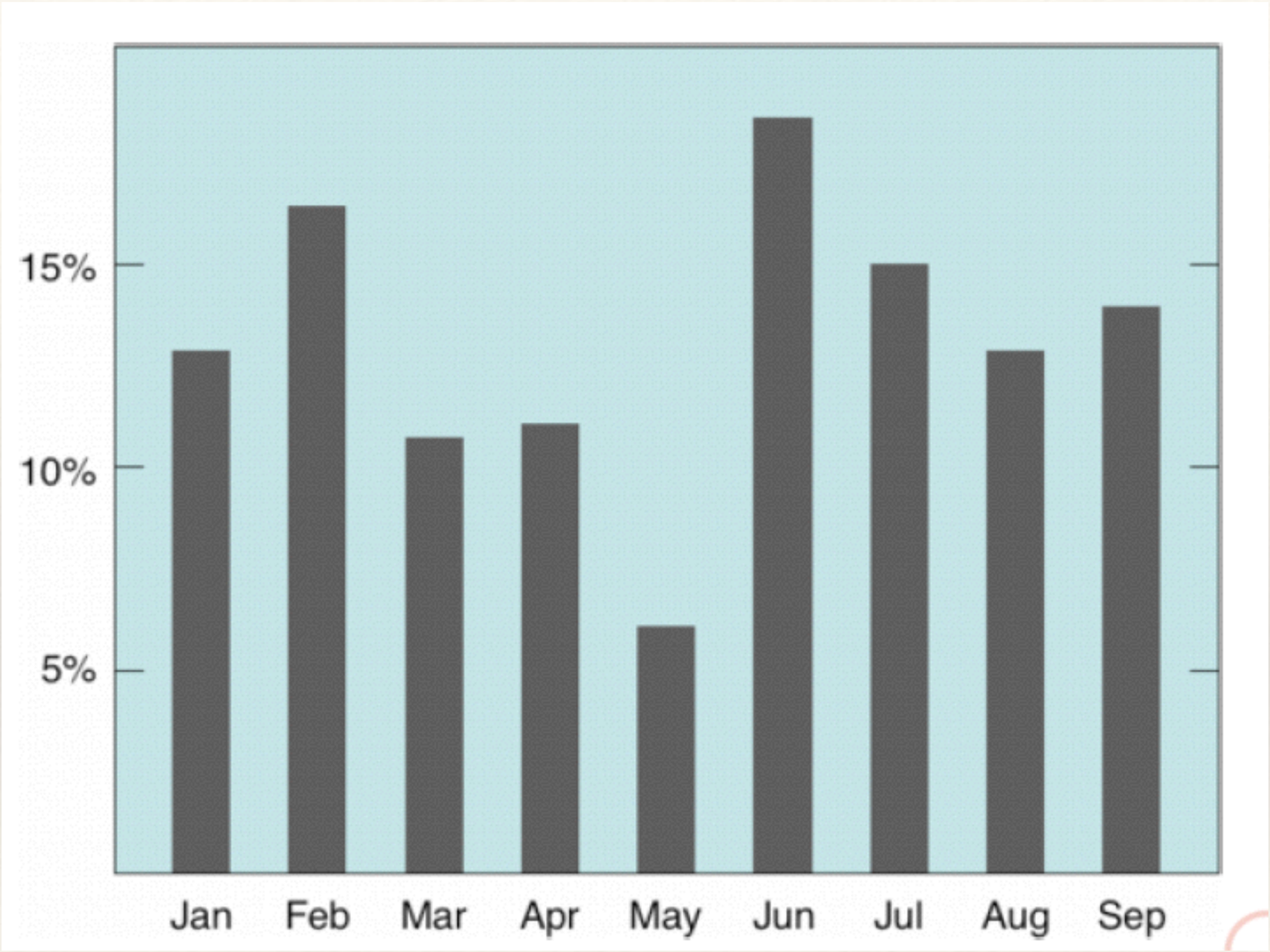
EVENINGS.

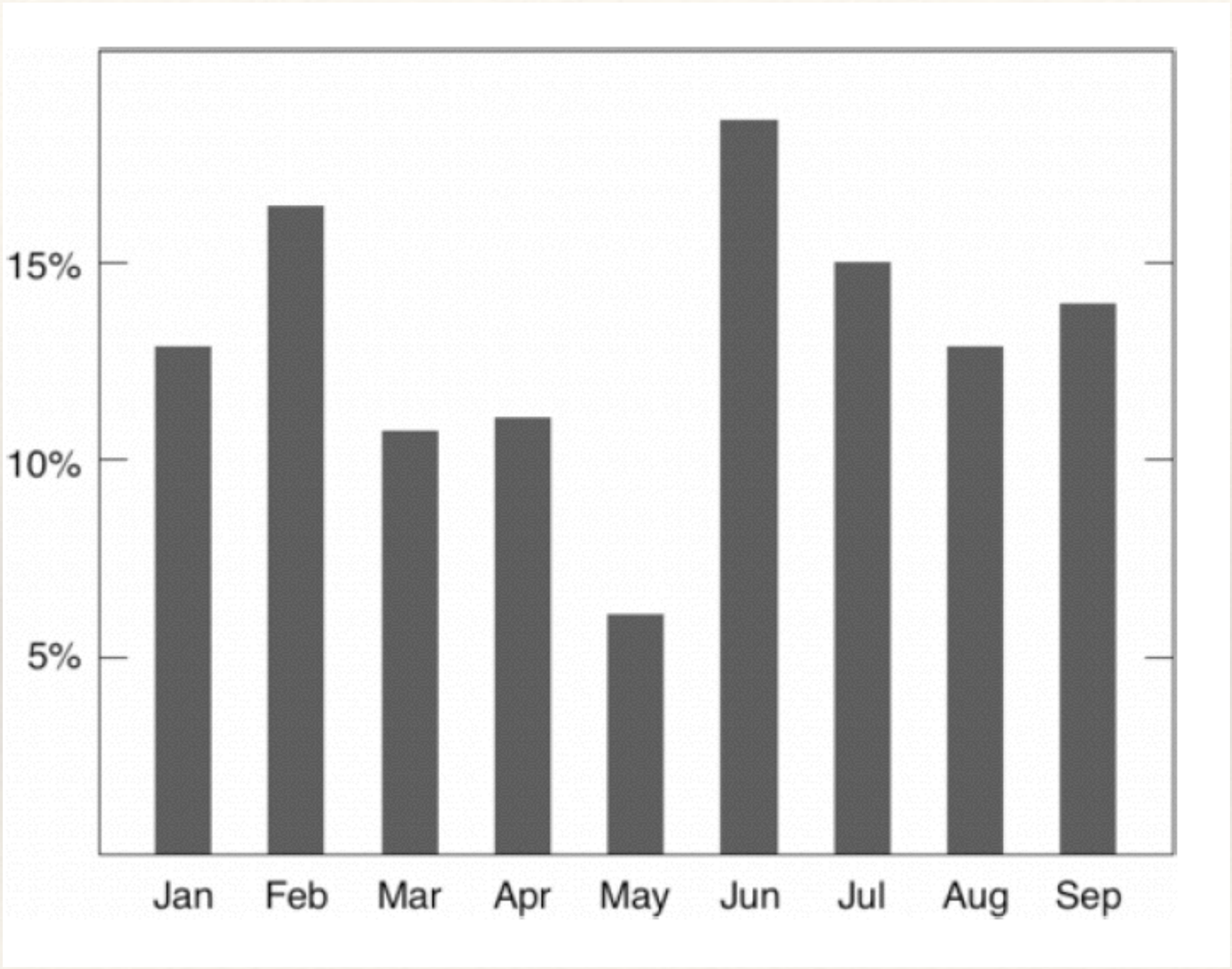
HENRY B. ALLEN
THE FRANKLIN INSTITUTE

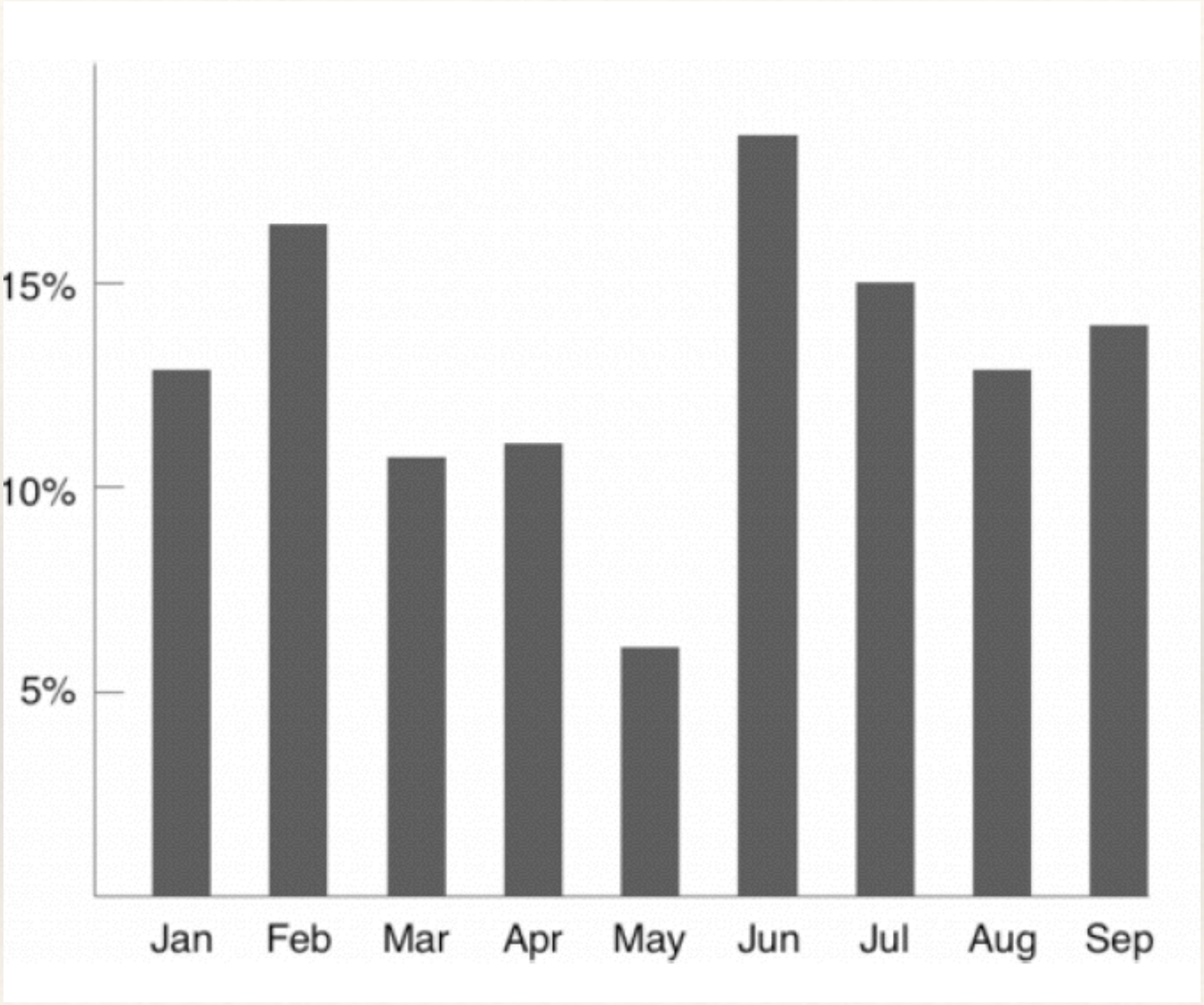

REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985
18
19
20
22
24
26
27
27.5
19.1 mpg, expected average for all cars on road, 1985
13.7 mpg, average for all cars on road, 1978
1978 1979 1980 1981 1982 1983 1984 1985

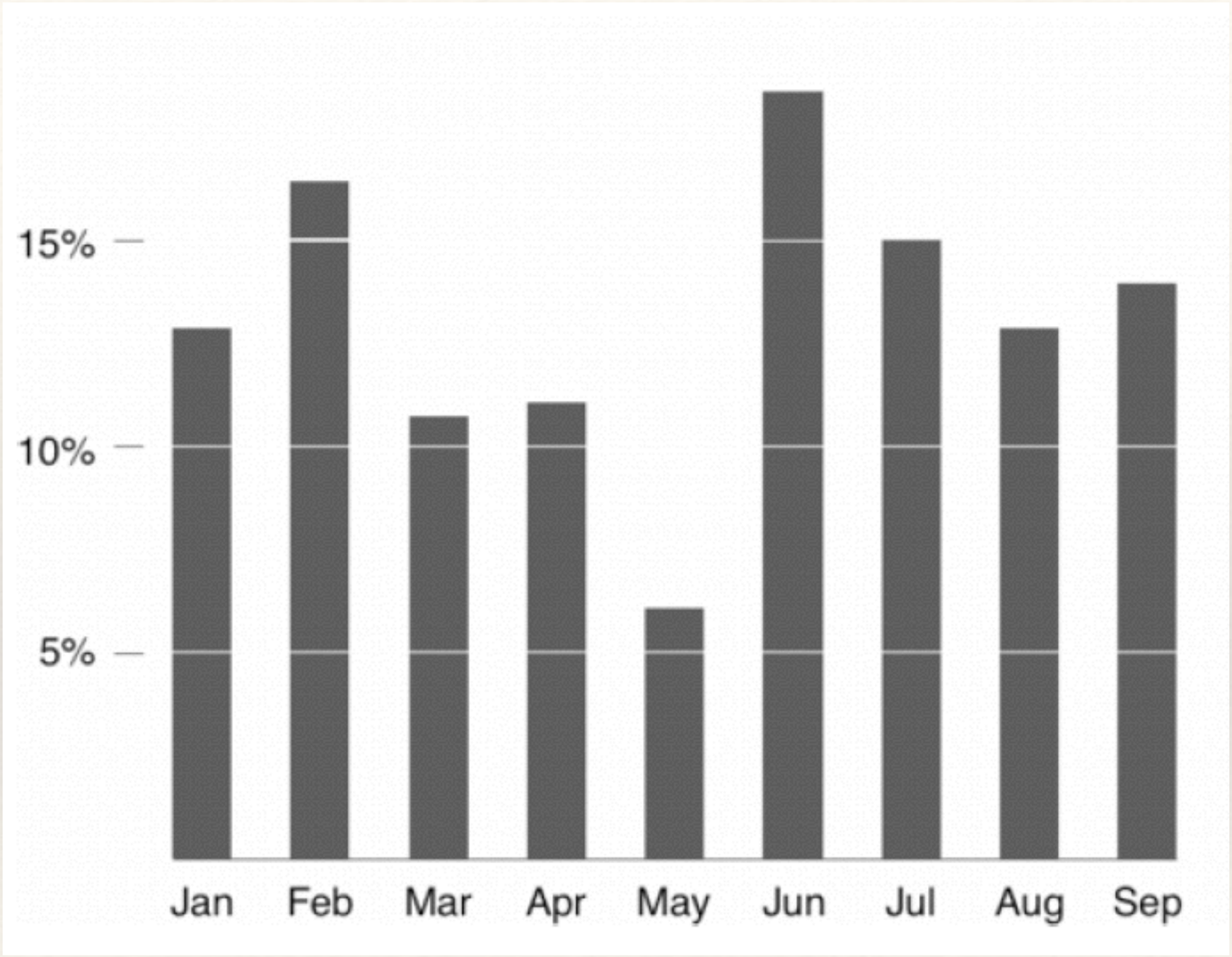


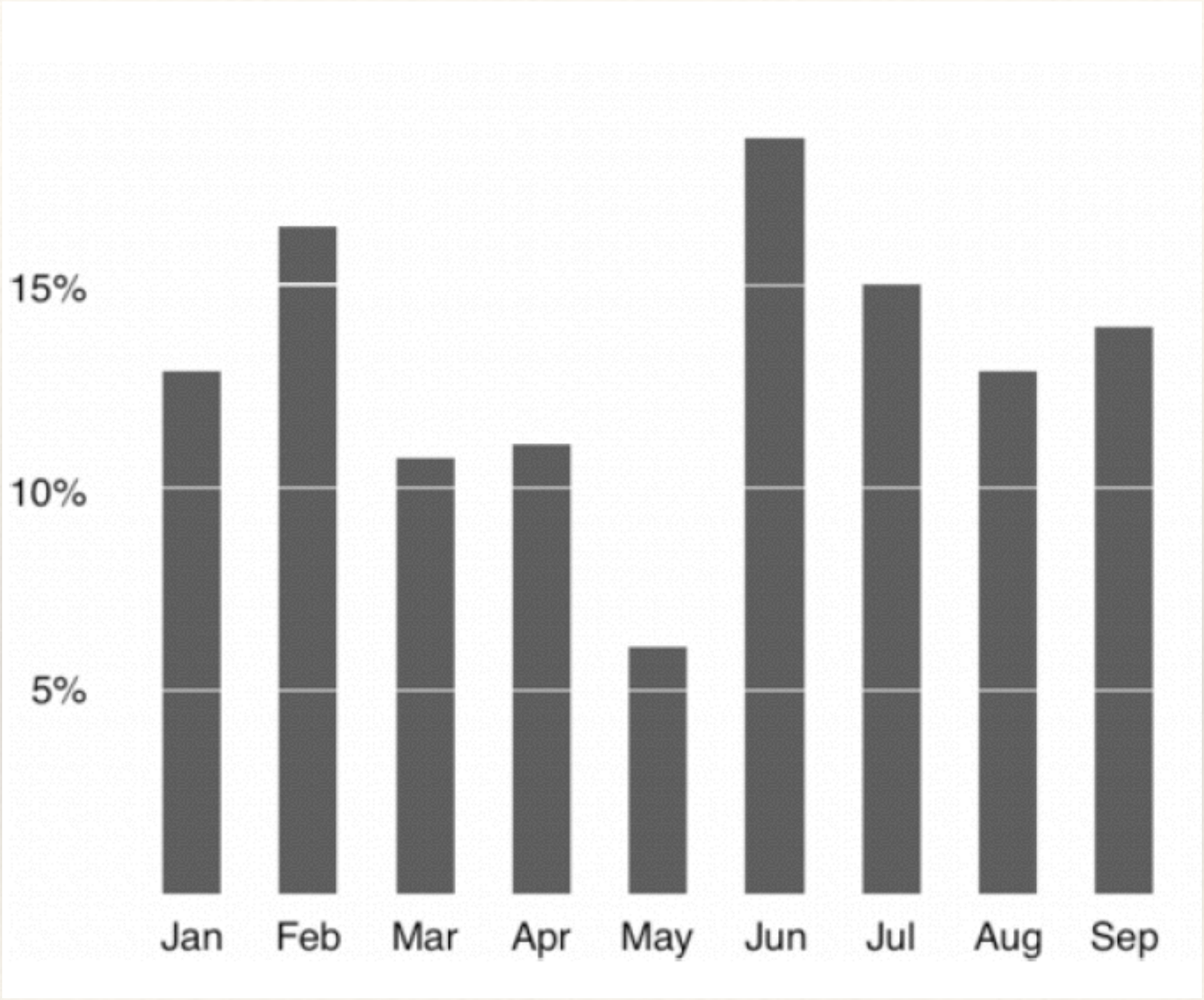
Tim Brey

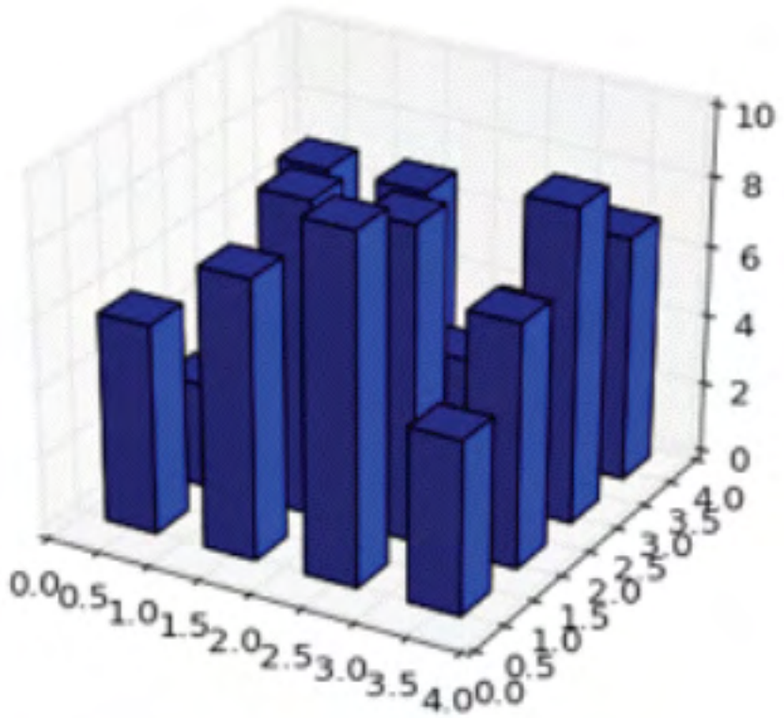

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

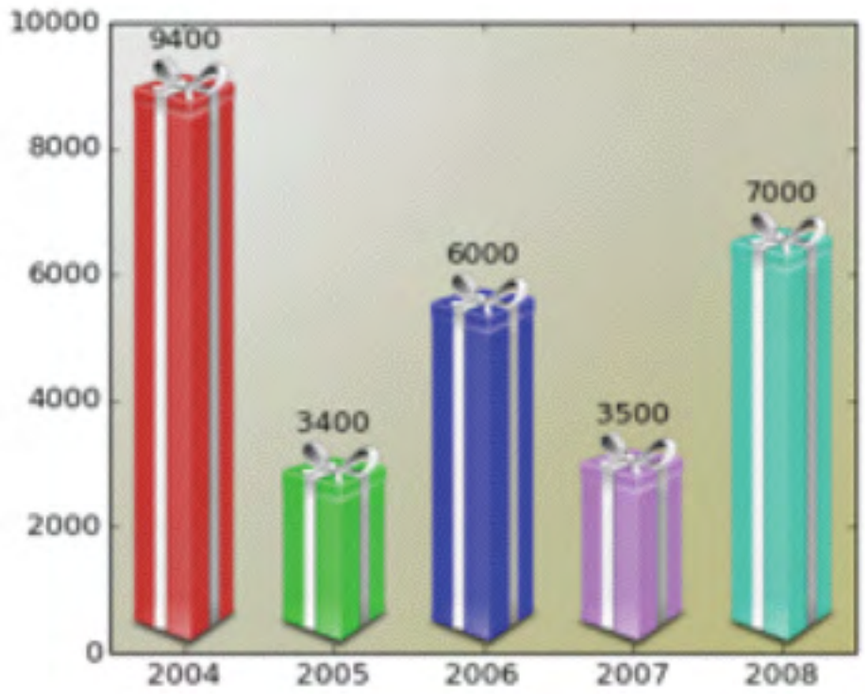

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep


15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

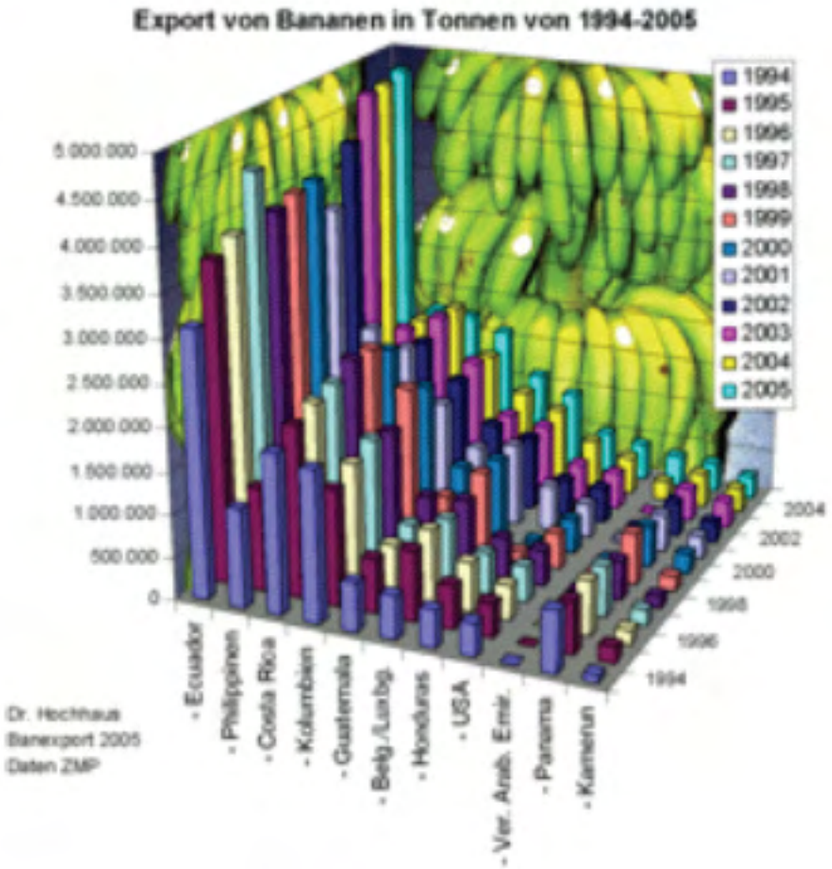

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

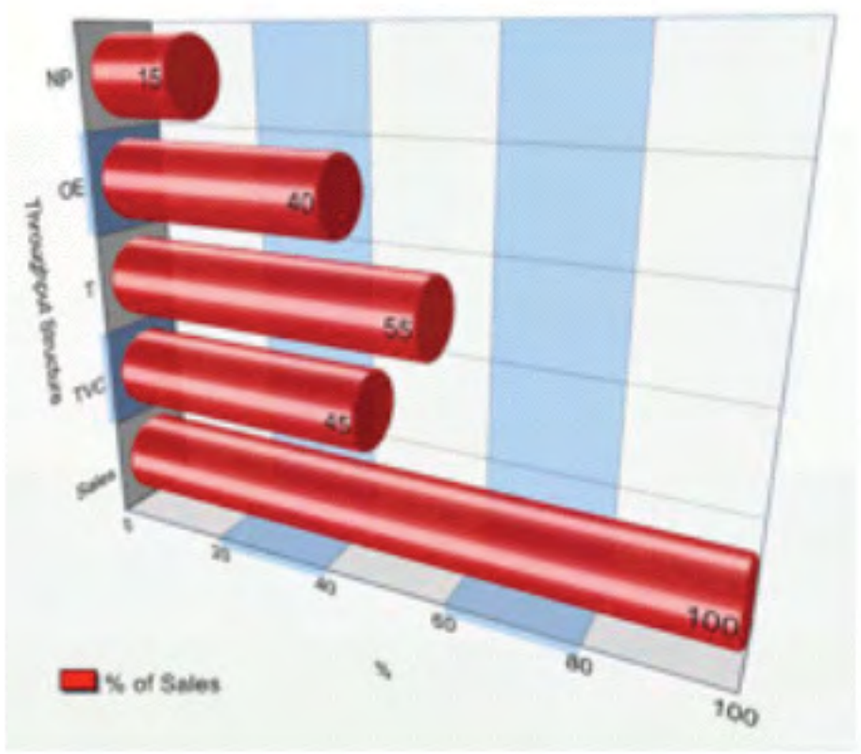

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep





matplotlib gallery

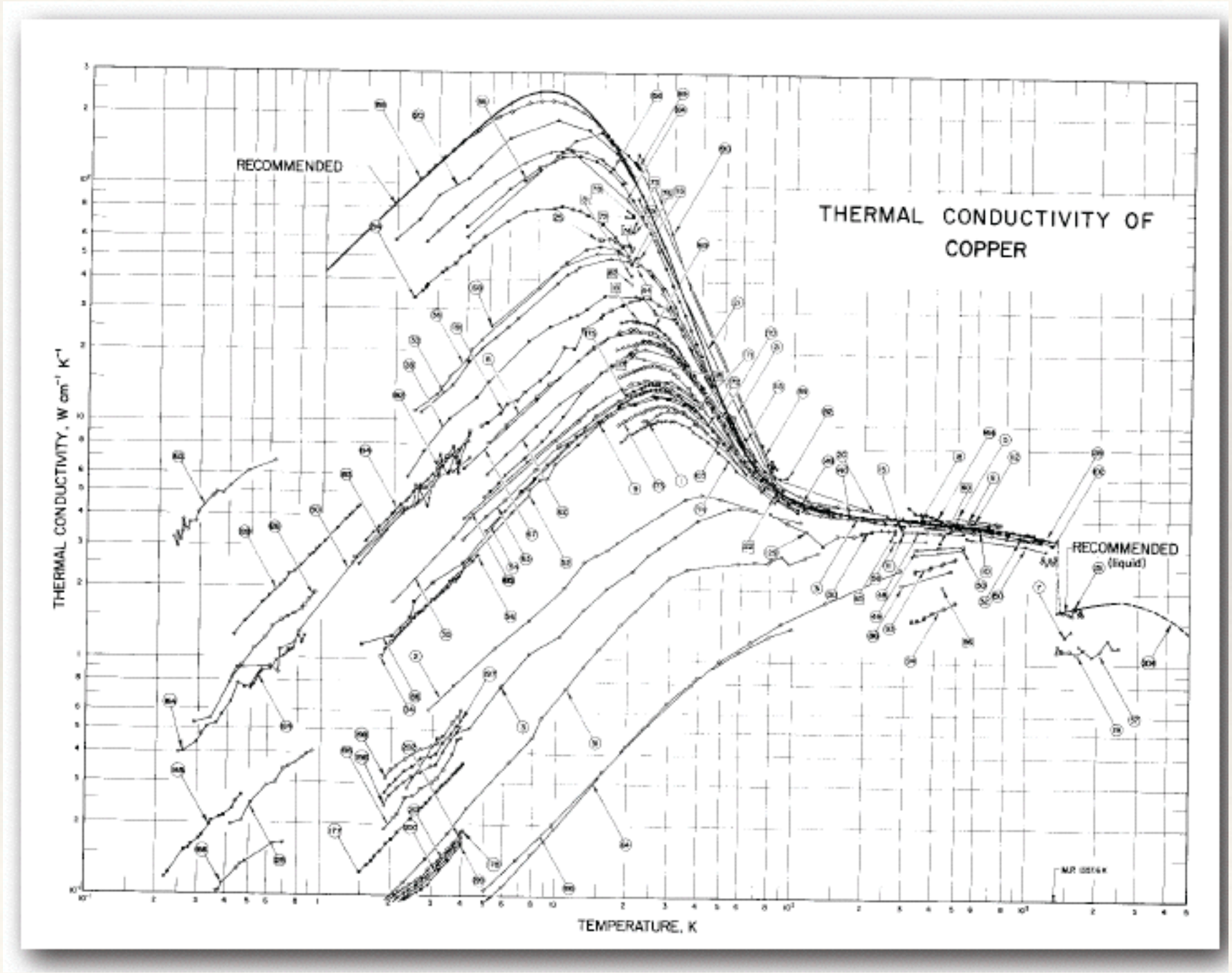




Excel Charts Blog

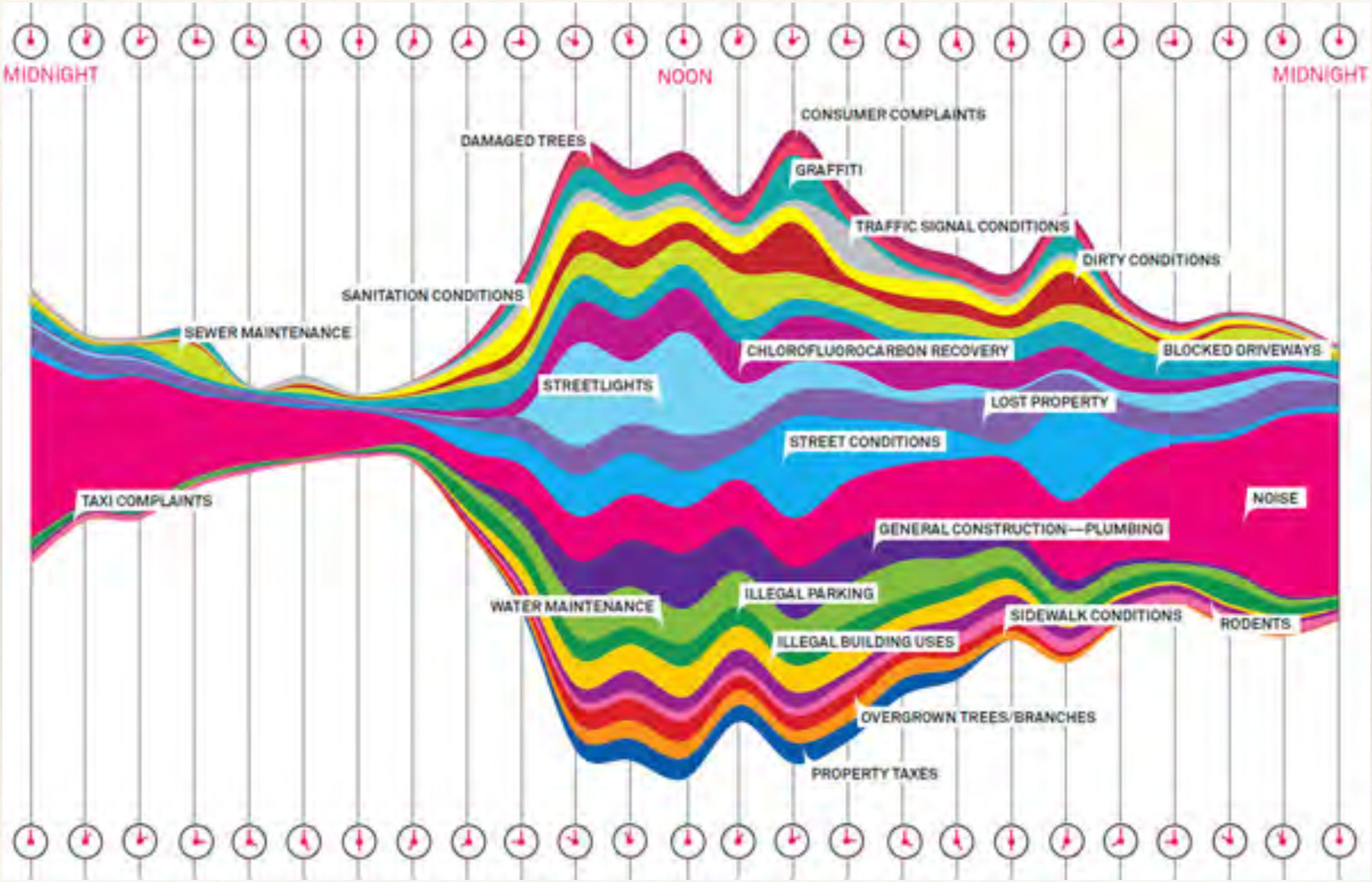
Principle: Increase Data Density

$$\text{Data density} = \frac{\text{Number data items}}{\text{Area of data in graphic}}$$



Ho et al., “Thermal Conductivity of the Elements: A Comprehensive Review” J. Phys. Chem. 1974

100 Million Calls to 311 by Steven Johnson 2011



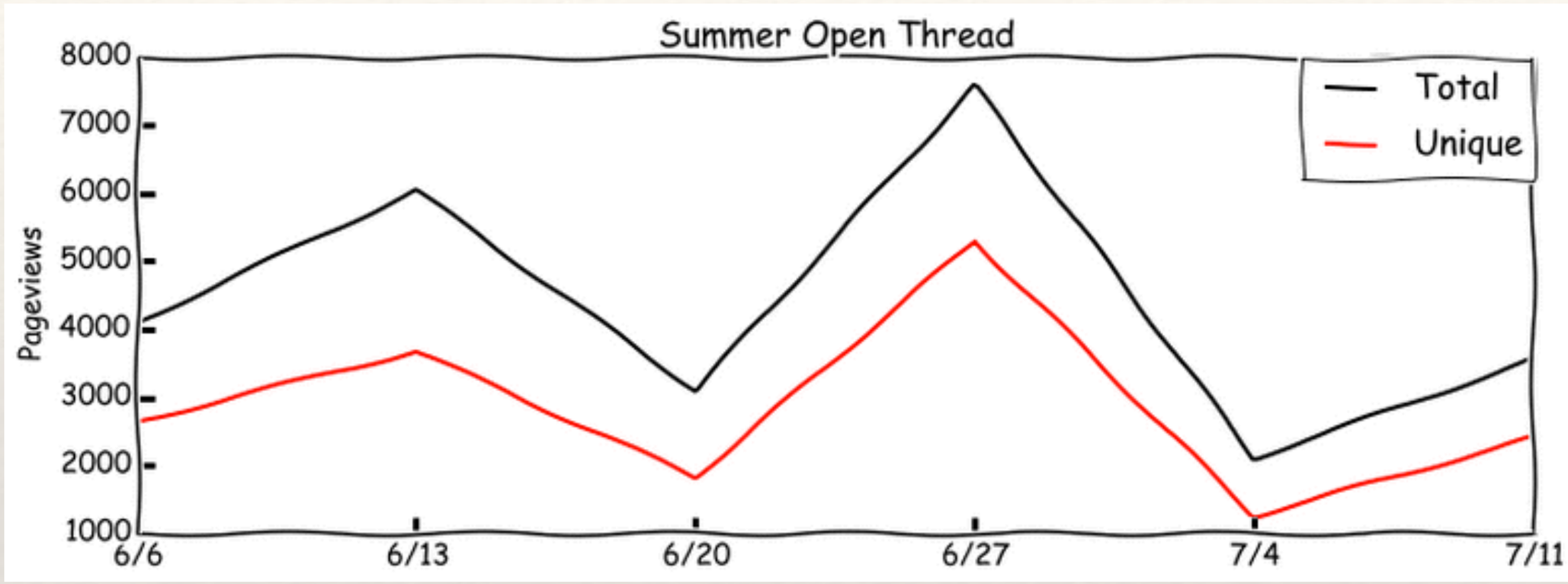
Tufte Principles

- Don’t Lie
- Maximize Data to Ink Ratio
- Avoid Chart Junk
- Increase Data Intensity

Hannah's Rules

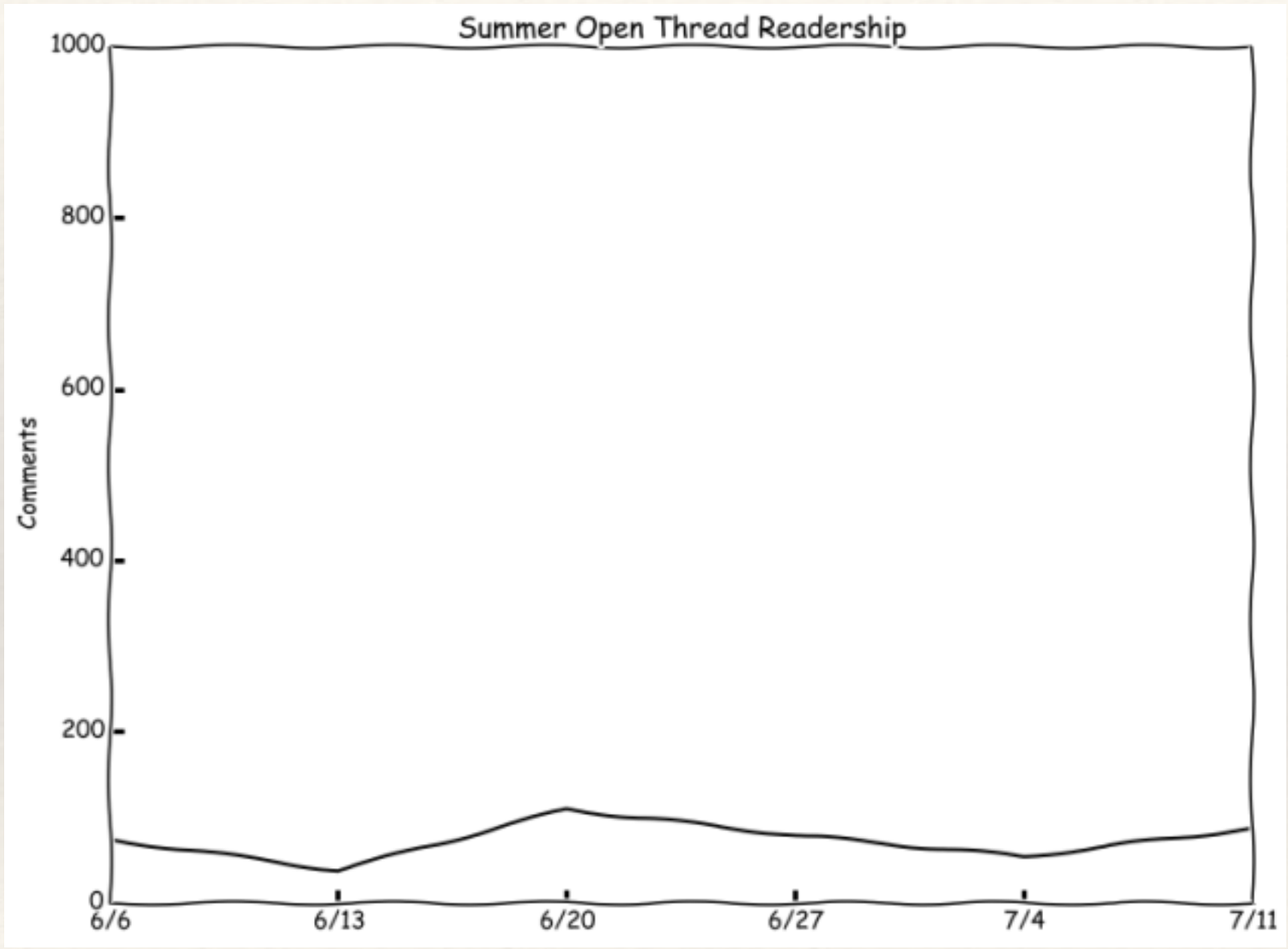
- http://hackerspace.lifehacker.com/5-rules-for-making-graphs-1605706367

1. Label Everything



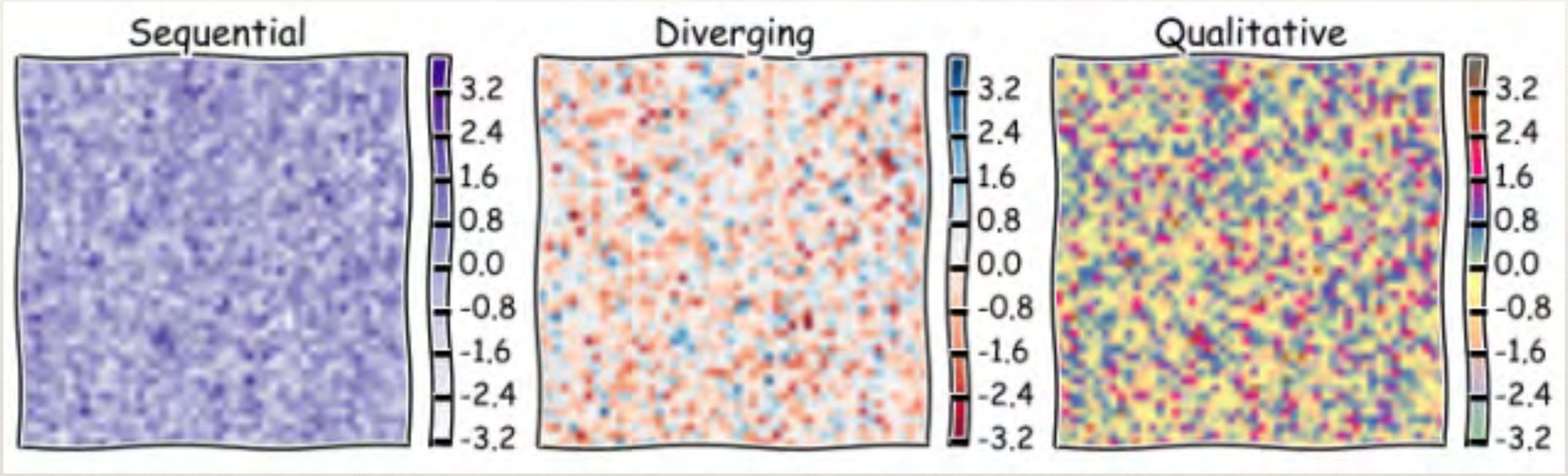
Important: Meaningful Titles
Label Axis
List data source

2. Work with the Numbers

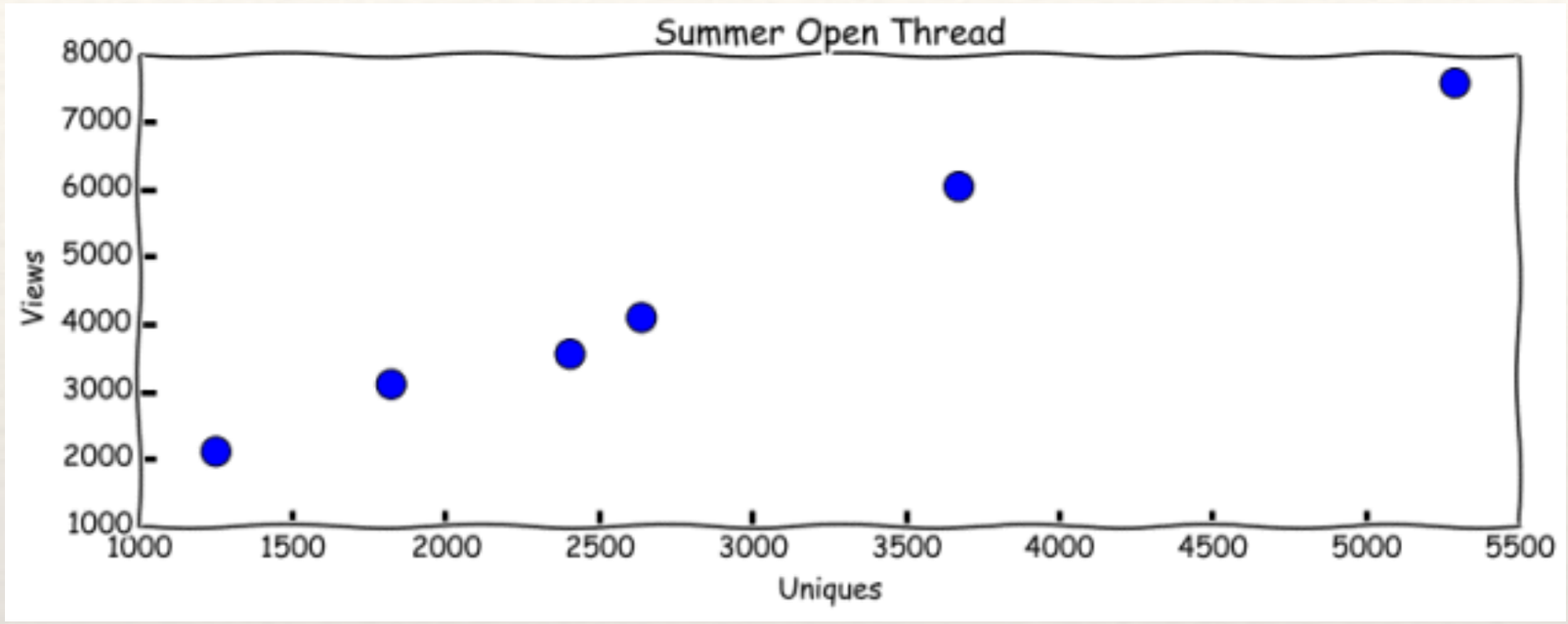


Should be zoomed on range of data

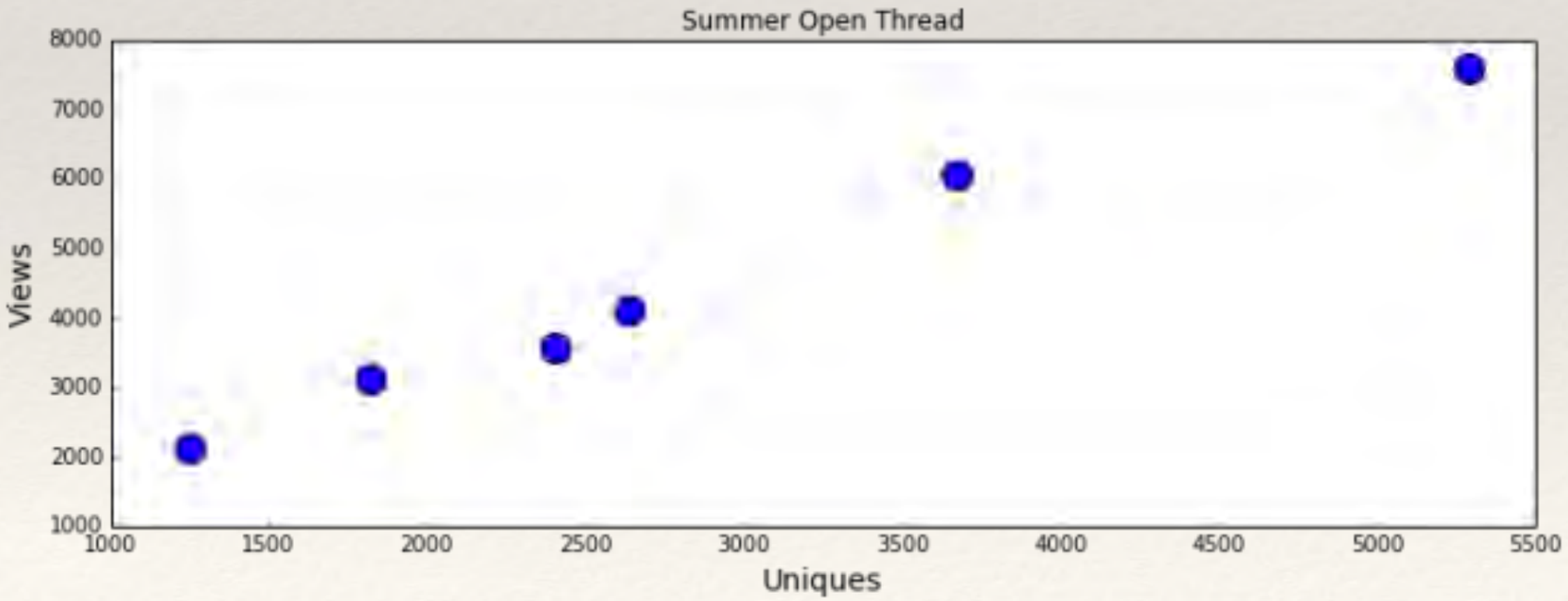
3. Choose Colors Carefully



4. Know your Audience

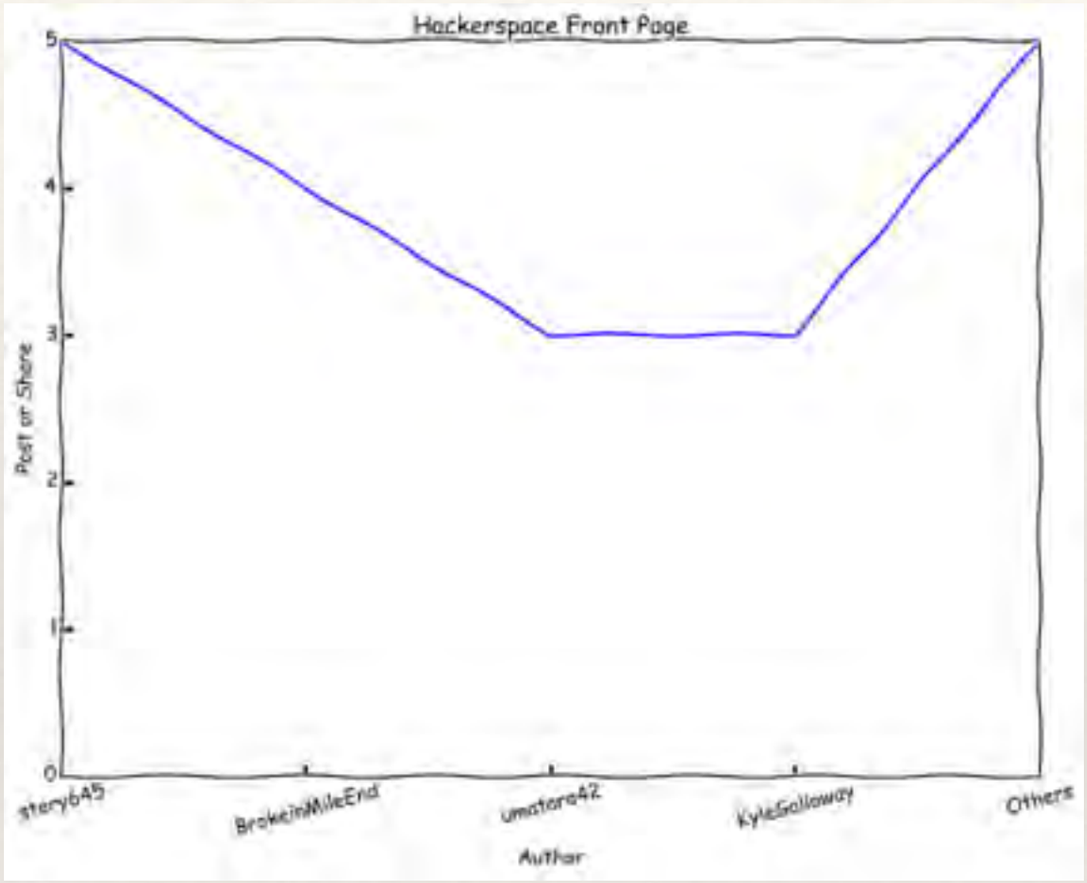


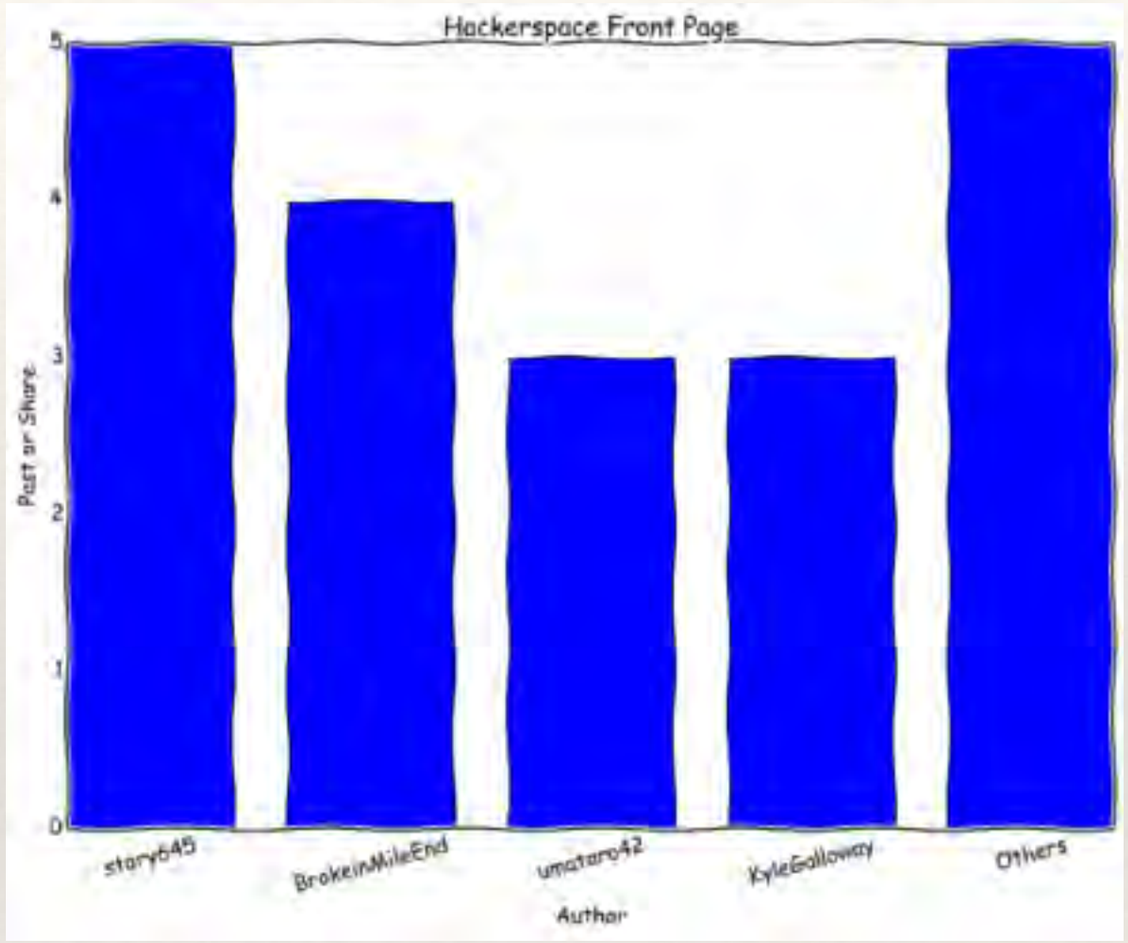
14 year olds



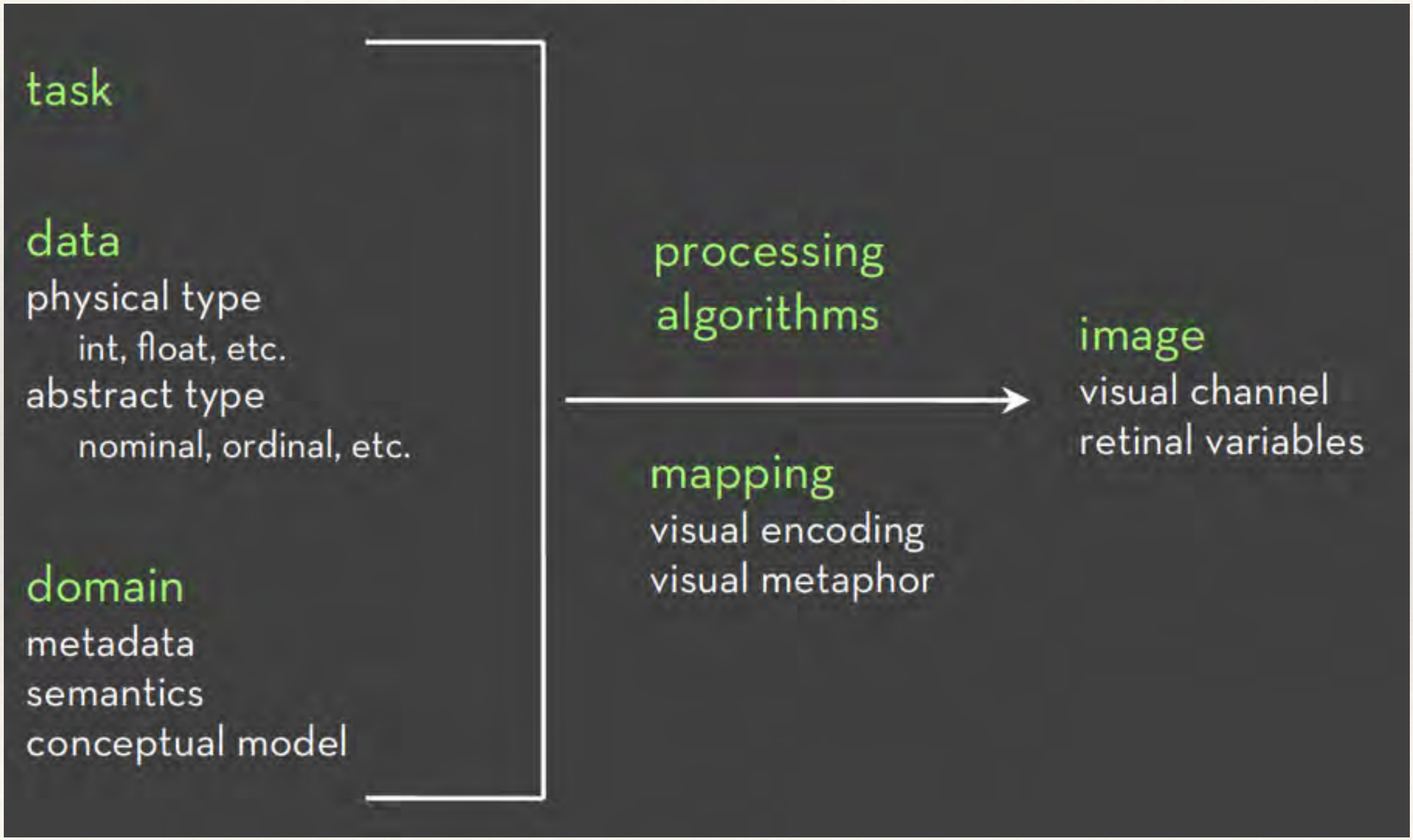
Professors

5. Use the Correct Graph





The Big Picture



SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens
Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues

by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL Categorical Qualitative	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Nominal, Ordinal and Quantitative

- N: Nominal (labels)
  - Eg. Animals, pigs, goats, cattle
- O: Ordered
  - Eg. XS, S, M, L, XL, XXL
- Q: Interval (zero irrelevant)
  - Eg. Dates, Location (lon, lat)
- Q: Ratio (linear scale)
  - Eg. Mass, charge, speed

Data Types (Operations)

- Nominal: =, ≠
- Ordinal: =, ≠ and <, >
- Interval: =, ≠, <, >, and - (distance between points), + (diff)
- Ratio: =, ≠, <, >,+,-, and x,÷

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Item

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack		2/22/08
32	7/16/07	2-High	Small Pack		7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box		7/18/07
35	10/23/07	4-Not Specified	Wrap Bag		10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Attribute aka Feature

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack		2/22/08
32	7/16/07	2-High	Small Pack		7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Semantics

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci	1 = Quantitative	0.6	6/6/05
70	12/18/06	5-Low	2 = Nominal	0.59	12/23/06
70	12/18/06	5-Low	3 = Ordinal	0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci	1 = Quantitative	0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low	2 = Nominal	0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium	3 = Ordinal	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Example: U.S. Census Data

- People: # of people in group
- Year: 1850 – 2000 (every decade)
- Age: 0 – 90+
- Sex: Male, Female
- Marital Status: Single, Married, Divorced, …

Census Data

- People
- Year
- Age
- Sex
- Marital Status
- 2348 data points

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534
30	1850	70	0	1	73677
31	1850	70	0	2	71762
32	1850	75	0	1	40834
33	1850	75	0	2	40229
34	1850	80	0	1	23449
35	1850	80	0	2	22949
36	1850	85	0	1	8186
37	1850	85	0	2	10511
38	1850	90	0	1	5259
39	1850	90	0	2	6569
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

Census: N, O, Q?

- People Count……
- Year……
- Age……
- Sex……
- Marital Status…..

Census: N, O, Q?

- People Count…… **Q-Ratio**
- Year…… **Q-Interval (O)**
- Age…… **Q-Ratio (O)**
- Sex…… **N**
- Marital Status….. **N**

Visual Variables

Jacques Bertin

- French cartographer [1918-2010]
- Semiology of Graphics [1967]
- Theoretical principles for visual encodings

Bertin's Visual Variables

Marks Points Lines Areas

Channels

Position

Size

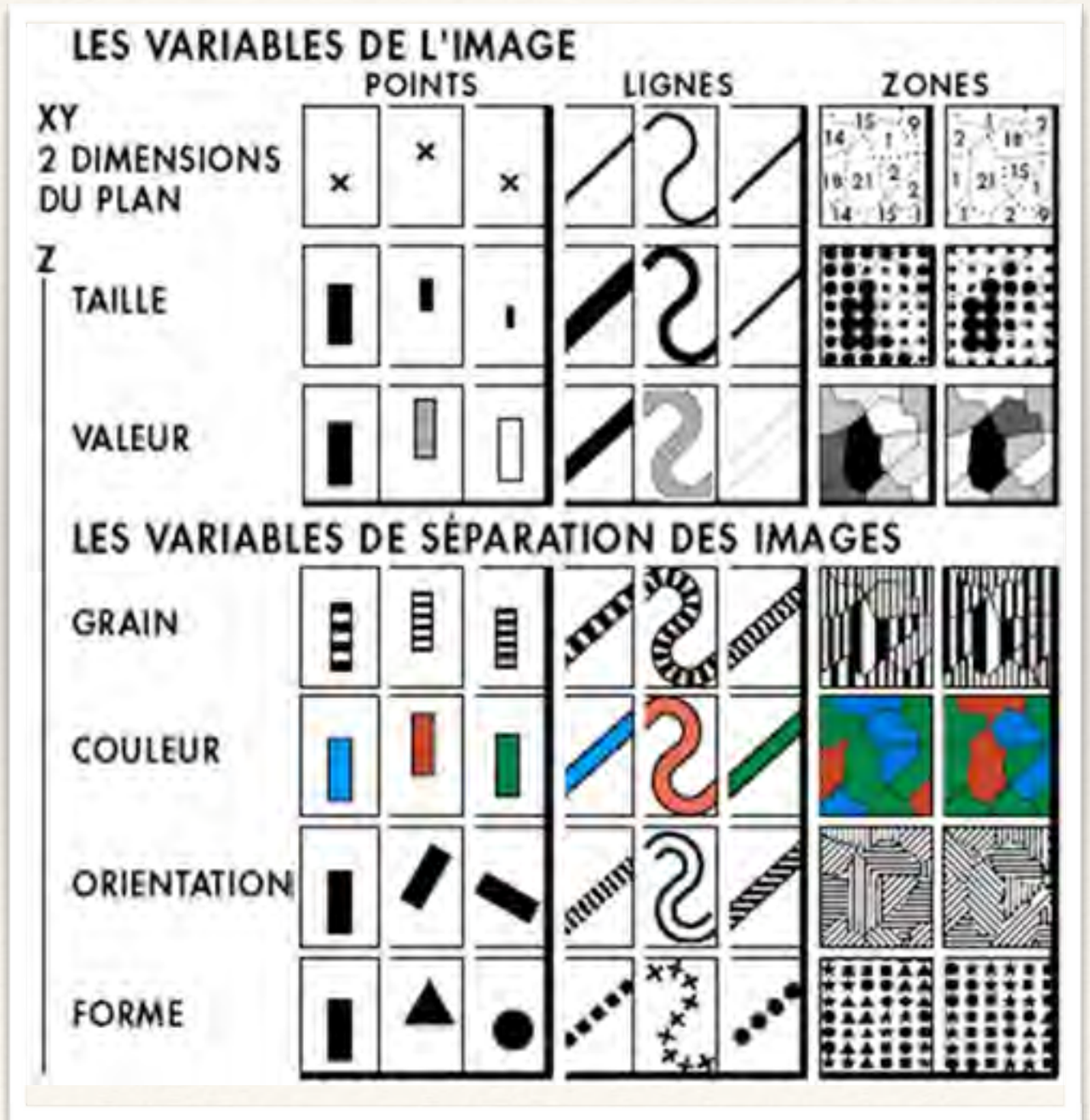
(Grey)Value

Texture

Color

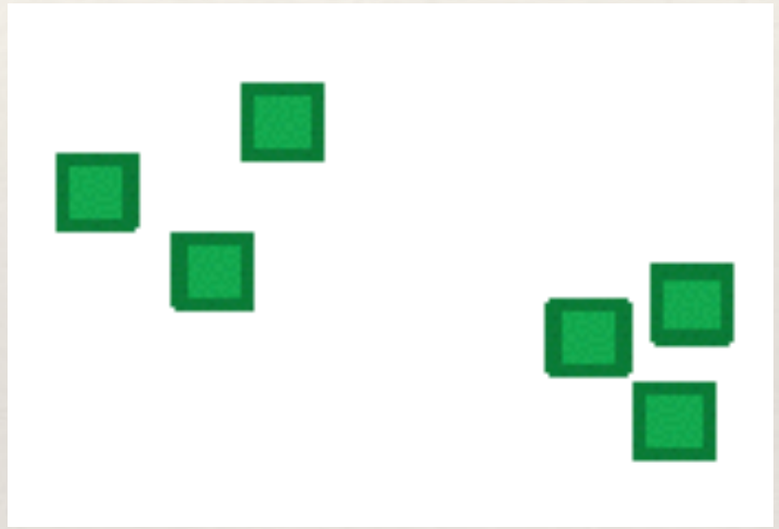
Orientation

Shape

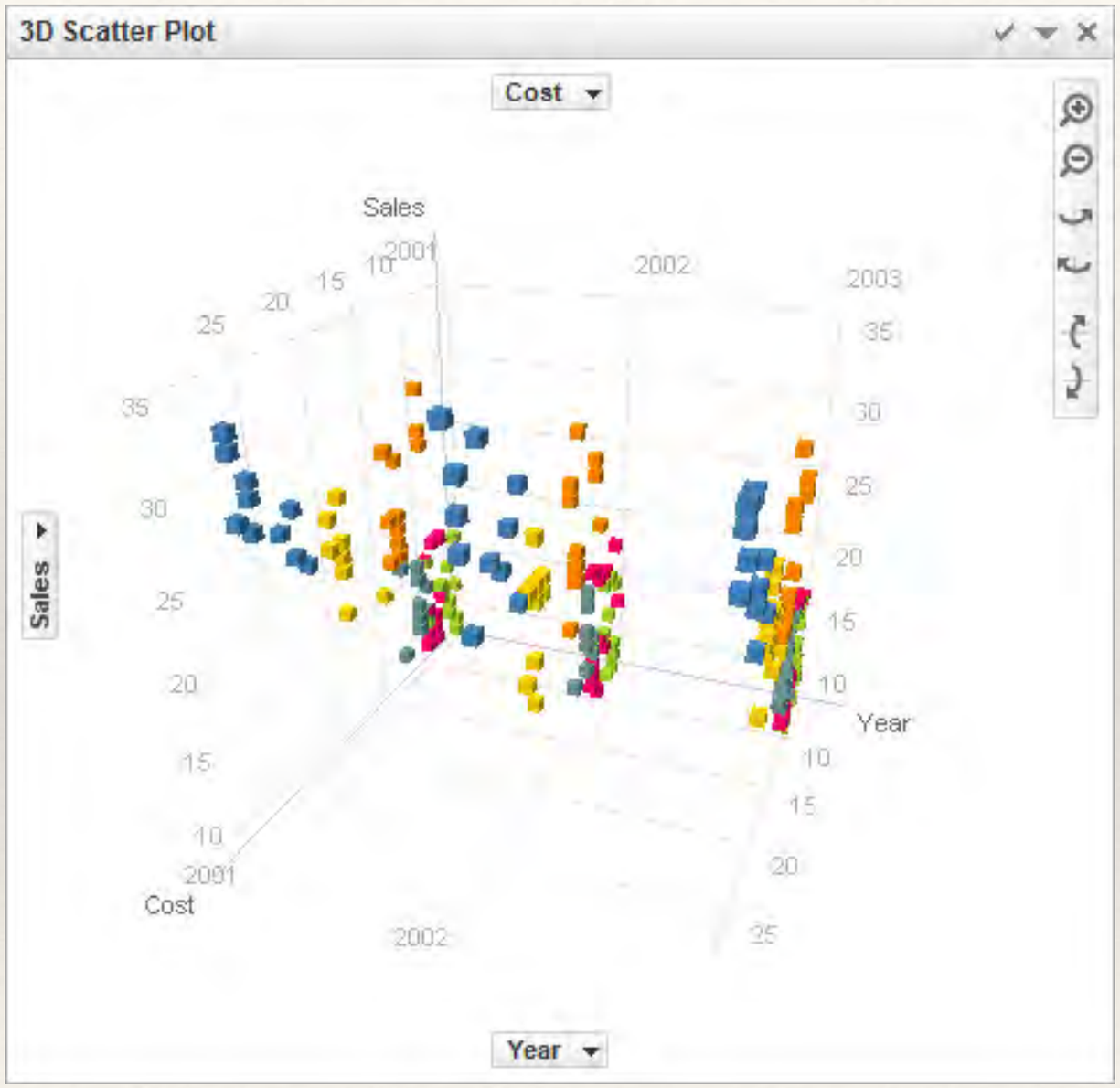


Position

- Strongest visual variable
- Suitable for all data types
- Problems:
  - Sometimes not available
  - Cluttering



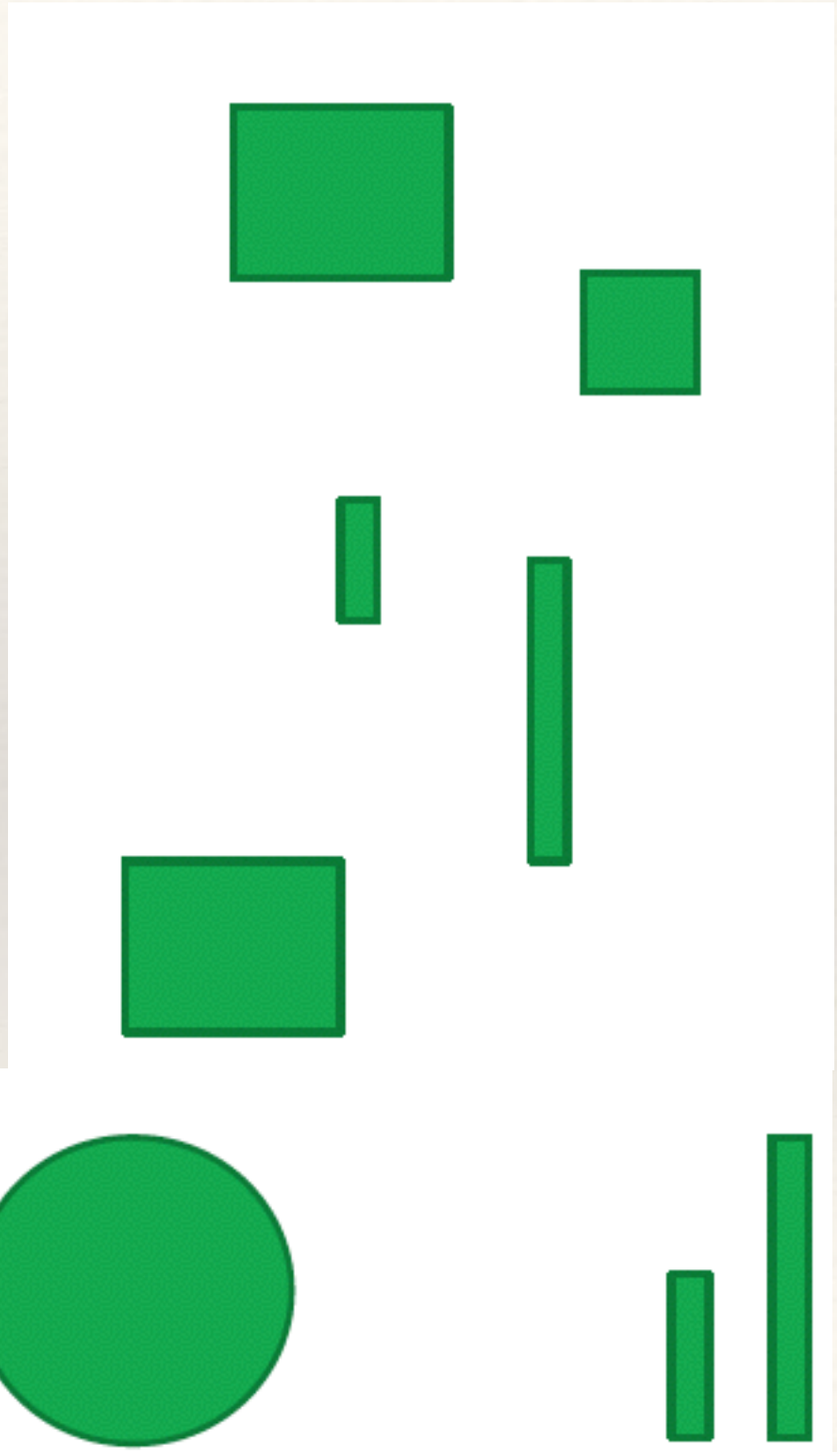
Position in 3D?



[Spotfire]

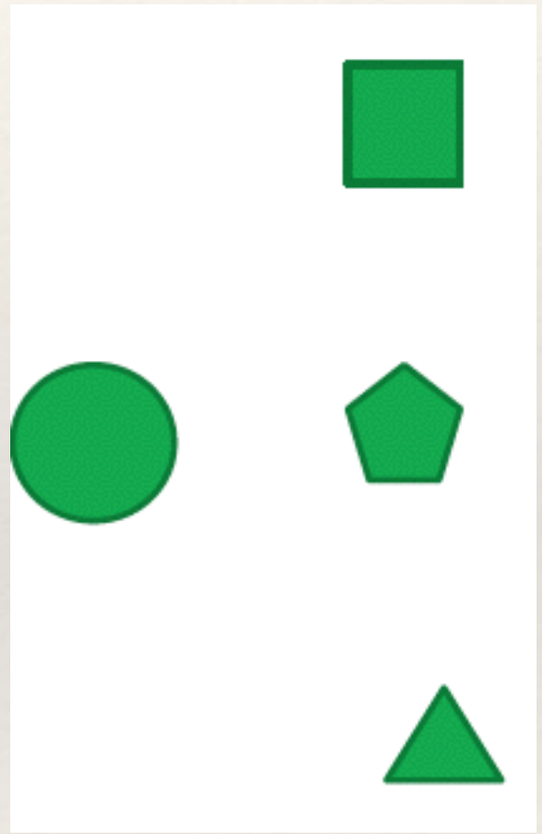
Size & Length

- Good visual variable
- Easy to see whether one is bigger
- Grouping works
- Judging differences
- Good for aligned bars (position)
- OK for changes in length
- Bad for changes in area



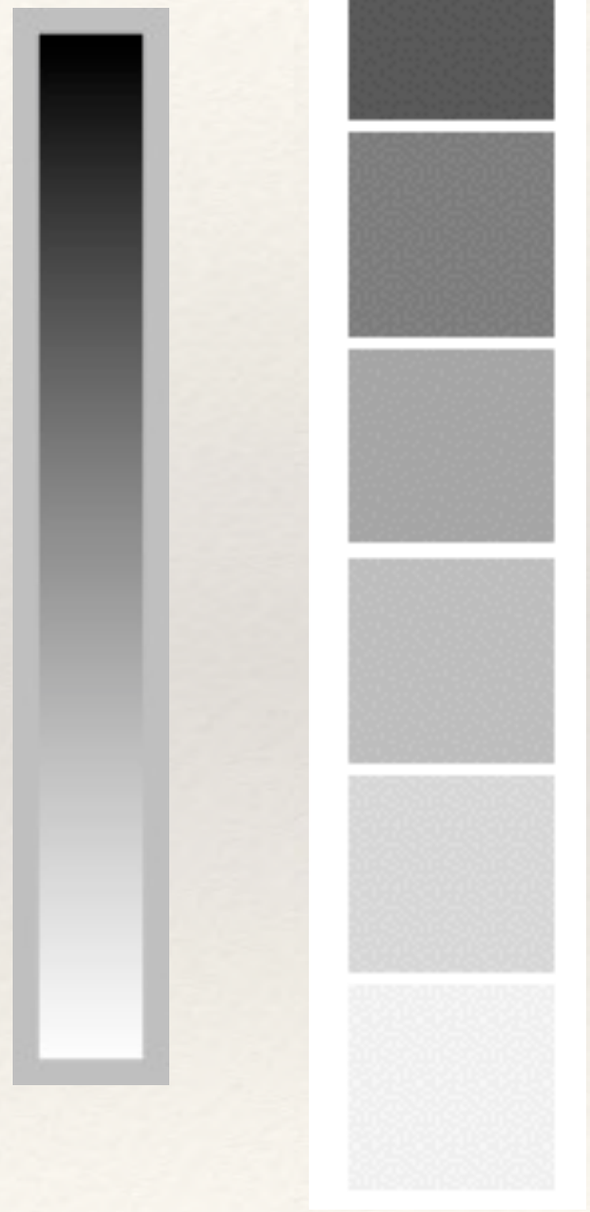
Shape

- Great to recognize many classes.
- No grouping, ordering.



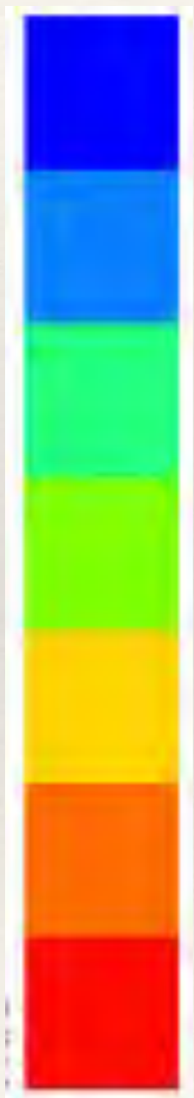
Value

- Good for quantitative data when length & size are used.
- Not very many shades recognizable
- Supports grouping
- Is pre-attentive (stands out) if sufficiently different



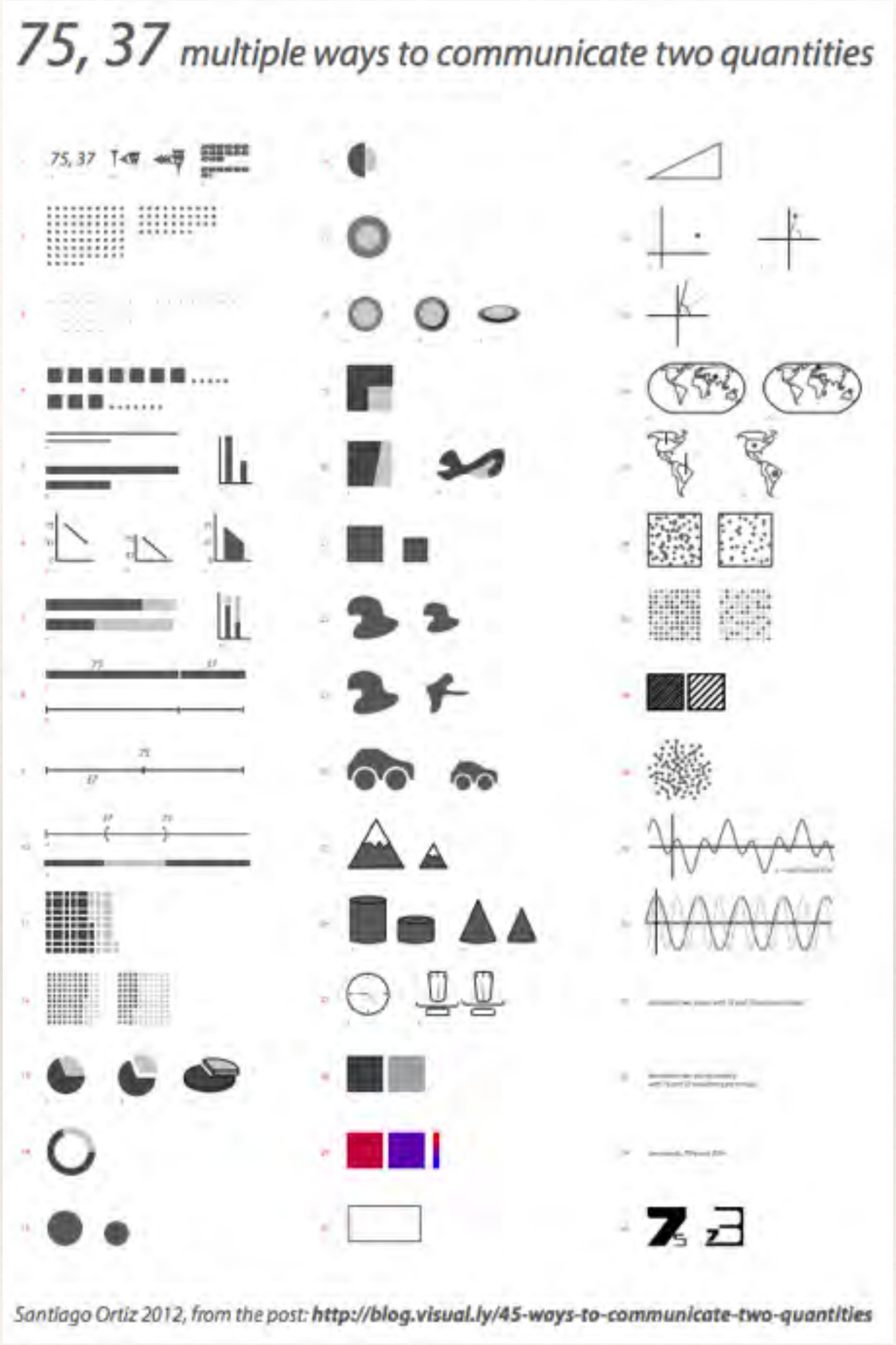
Color (Hue)

- Good for qualitative data
- Limited number of classes!
- Not good for quantitative data!
- Is pre-attentive if sufficiently different.
- Lots of pitfalls! Be careful!



Saturation (color)

- Good for Qualitative Data
- Good for Ordered Data
- Ok for Quantitative Data



http://blog.visual.ly/45-ways-to-communicate-two-quantities/

Bertin, 1967

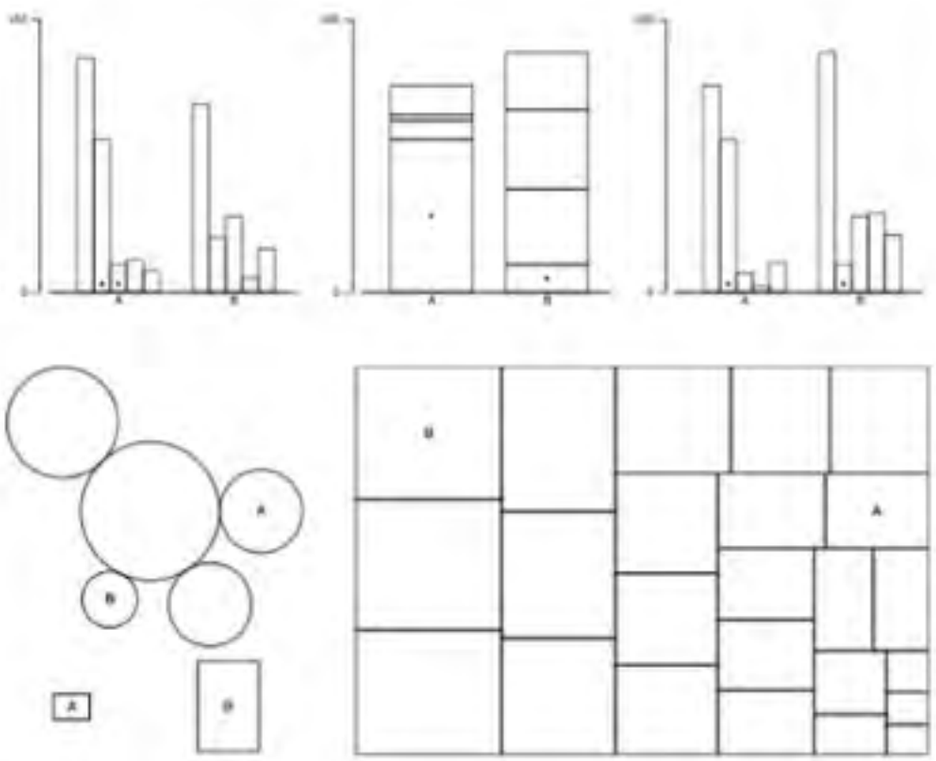
	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good
~ = OK
✗ = Bad

Heer & Bostock, 2010

Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design

Jeffrey Heer, Michael Bostock

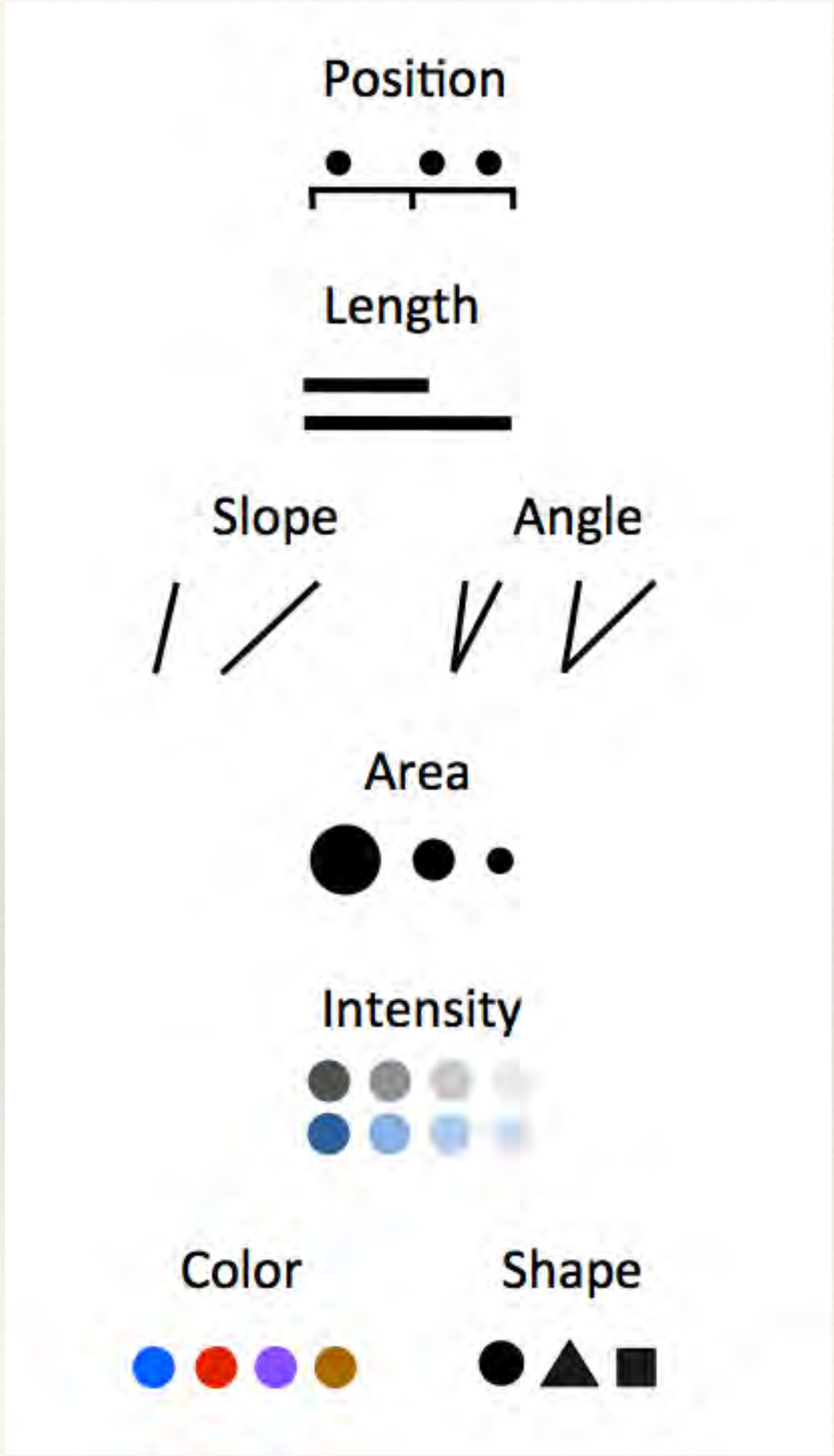


Experimental stimuli in which participants were asked to estimate what percentage the smaller value was of the larger.

ABSTRACT

Understanding perception is critical to effective visualization design. With its low cost and scalability, crowdsourcing presents an attractive option for evaluating the large design space of visualizations; however, it first requires validation. In this paper, we assess the viability of Amazon's Mechanical Turk as a platform for graphical perception experiments.

Most Efficient



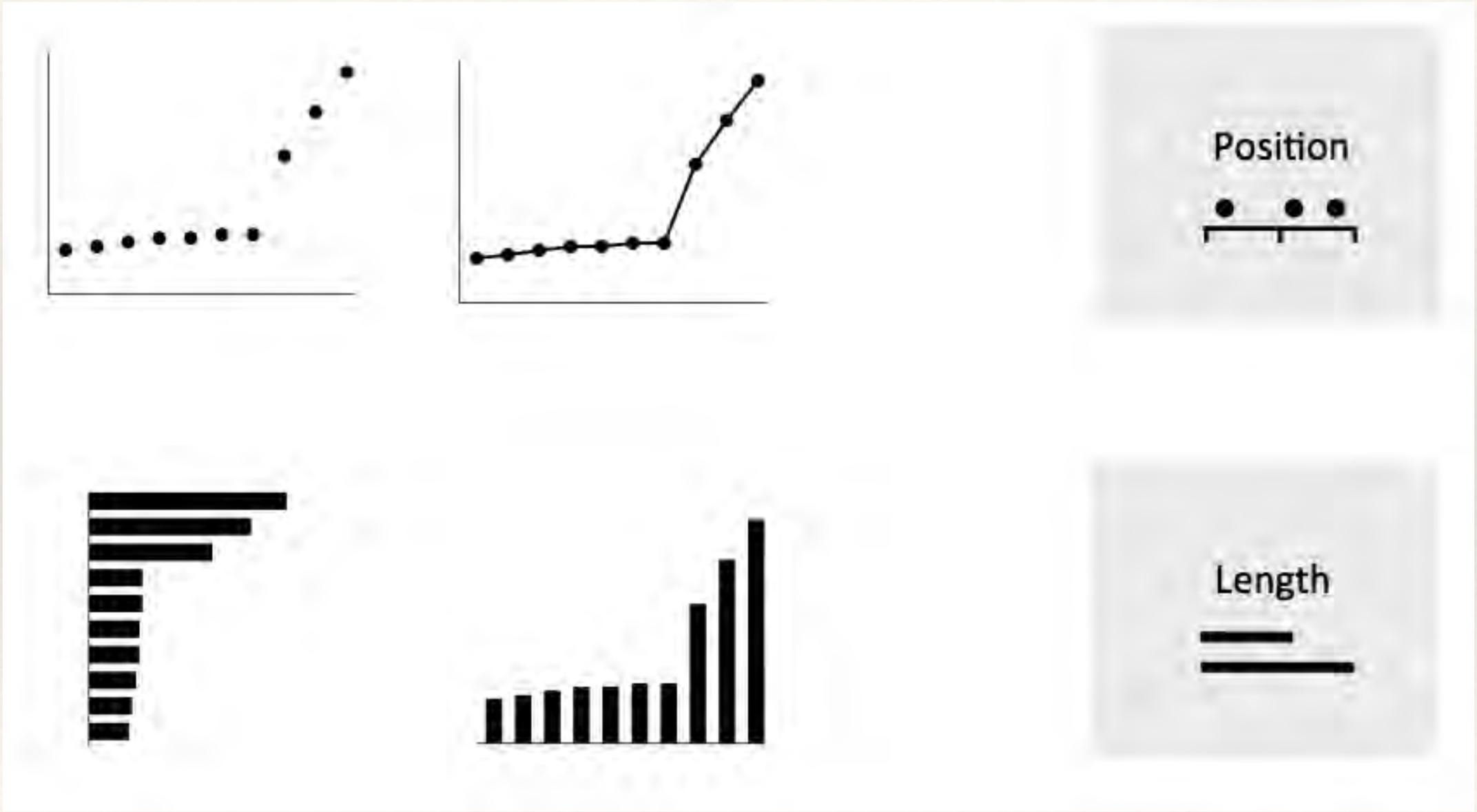
Quantitative

Ordinal

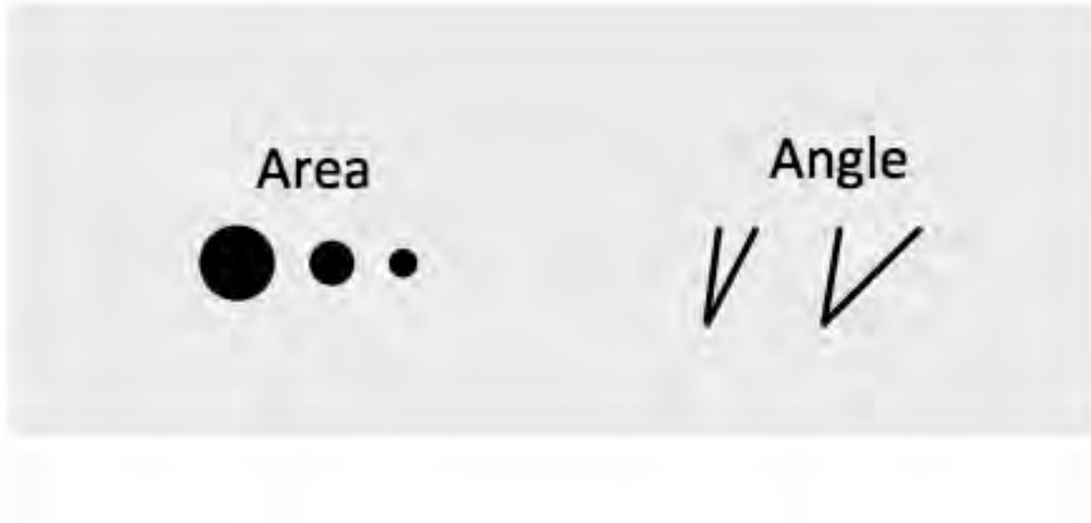
Nominal

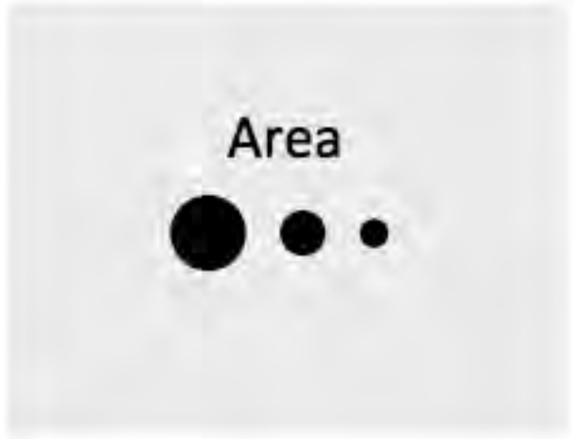
Least Efficient

Most Effective



Less Effective





Least Effective

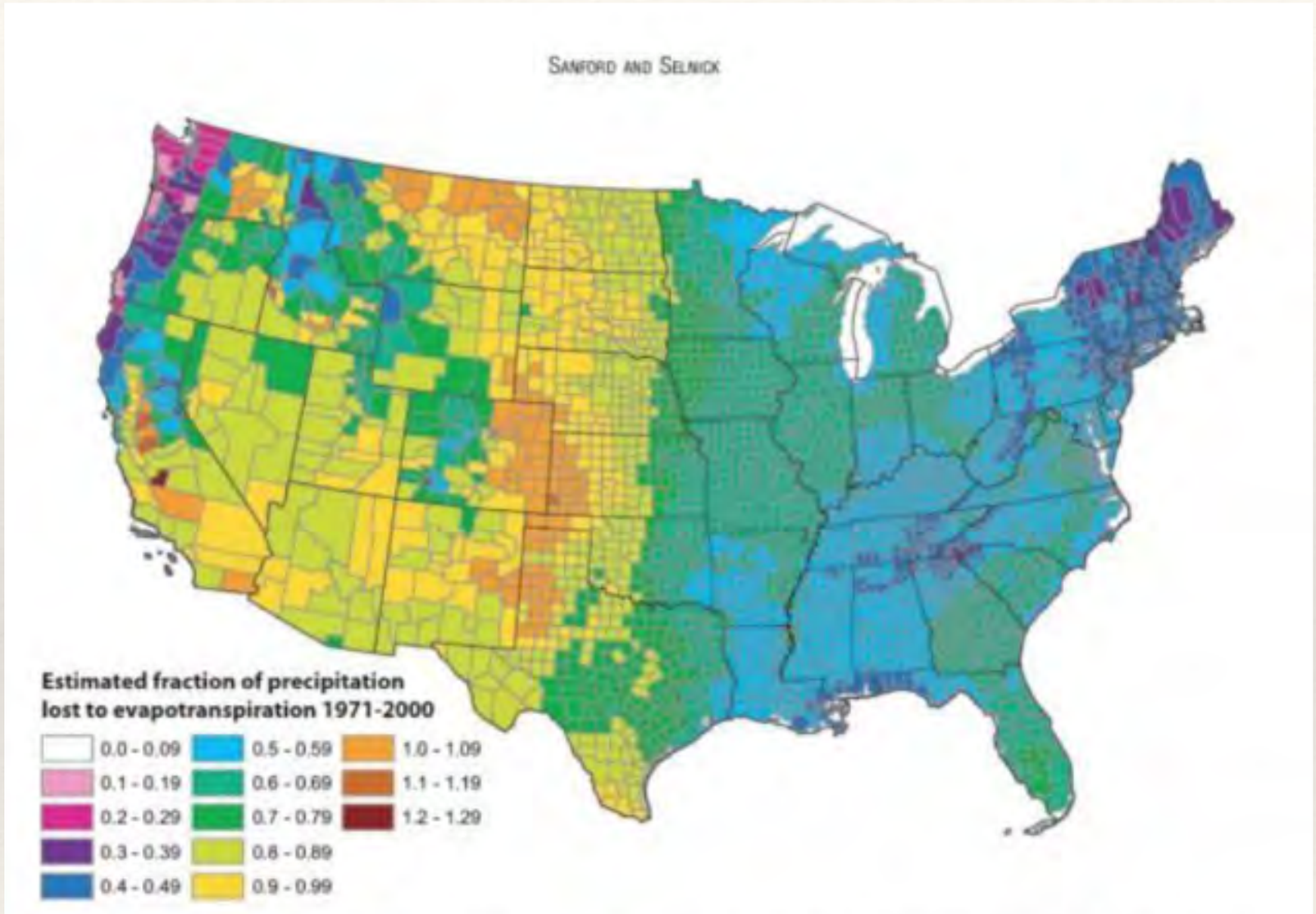


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.