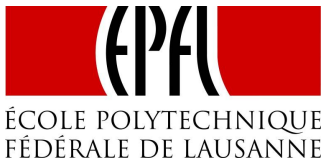
Applied Data Analysis (CS401)

Lecture 4
Data Visualization

Kiran Garimella



Visualization definitions

- "*Transformation of the symbolic into the geometric*" [McCormick et al. 1987]

- "*... finding the artificial memory that best supports our natural means of perception.*" [Bertin 1967]

- "*The use of computer-generated, interactive, visual representations of data to amplify cognition.*" [Card, Mackinlay, & Shneiderman 1999]

Uses for data visualization

Support reasoning about information **(analysis)**

- Finding relationships
- Discover structure
- Quantifying values and influences
- Should be part of a query/analyze cycle

Inform and persuade others **(communication)**

- Capture attention, engage
- Tell a story visually
- Can focus on certain aspects and omit others

Data visualization history

- Early data visualization (clay tablet maps) dates back to 5,000 years!

Cholera outbreak of 1854

- Cholera outbreak in London
- Over 500 people died
- At that time such diseases were believed to be caused by bad air


GREAT MARLBOROUGH STREET
MARLBOROUGH MEWS
PUMP
WORK HOUSE
PORTLAND STREET
PORTLAND MEWS
BREWERY
RICHMOND BUILDINGS
PETER STREET
GOLDEN SQUARE
HOPKINS STREET
ARCHER STREET
UP RUPERT ST
BRIDLE STREET
CAMBRIDGE STREET
NEW STREET
KING STREET
MARSHALL STREET
SILVER STREET
TYLER STREET
LITTLE PULTENEY STREET
GREAT PULTENEY STREET
BURLINGTON STREET

Data viz today

https://paulbutler.org/archives/visualizing-facebook-friends/

Old-fashioned viz

Great for data exploration, developed throughout the last few centuries…

Interactive viz

More and more common when delivering the results. New frameworks are the key enabler.

Want to learn more?

Dedicated course:

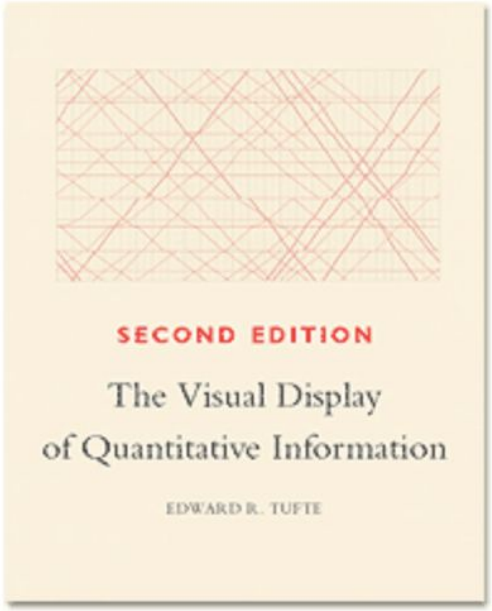
COM-480: Data Visualization

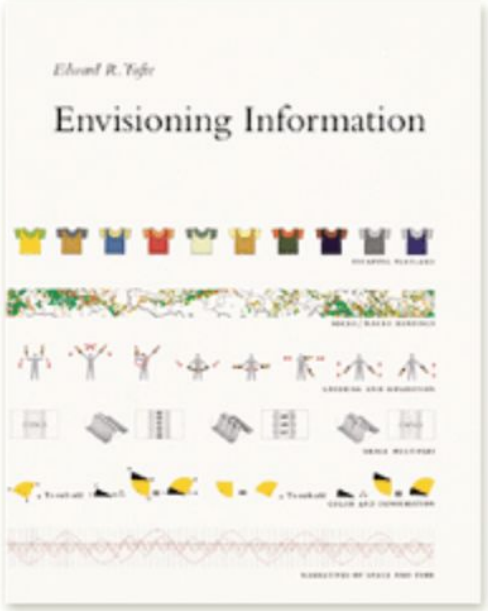
Today’s lecture

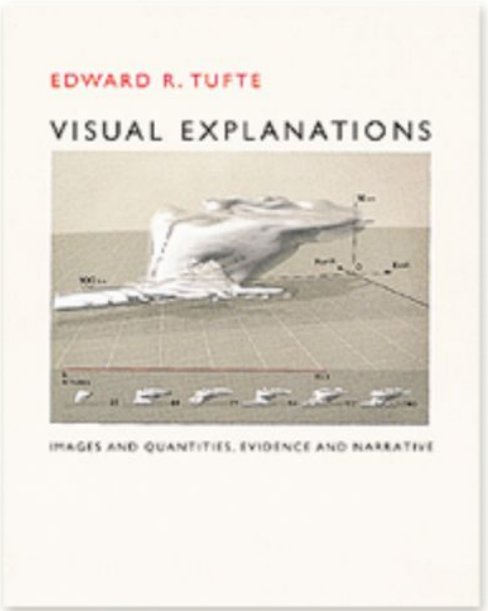
- Principles of data visualization
- Visualization for data exploration
- Examples
- Tools (more in lab session)

Principles of data visualization

Edward Tufte







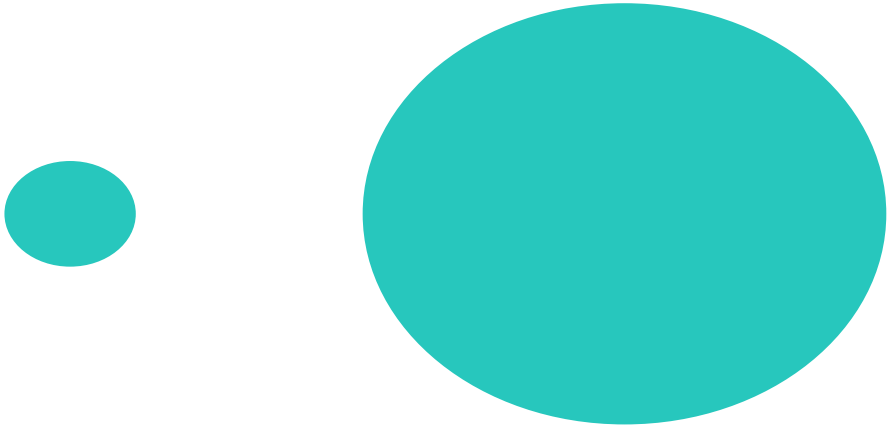
The Rules:

1. Show Your Data
2. Use Graphics
3. Avoid Chartjunk
4. Utilize Data-ink
5. Use Labels
6. Utilize Micro/Macro
7. Separate Layers
8. Use Multiples
9. Utilize Color
10. Understand Narrative

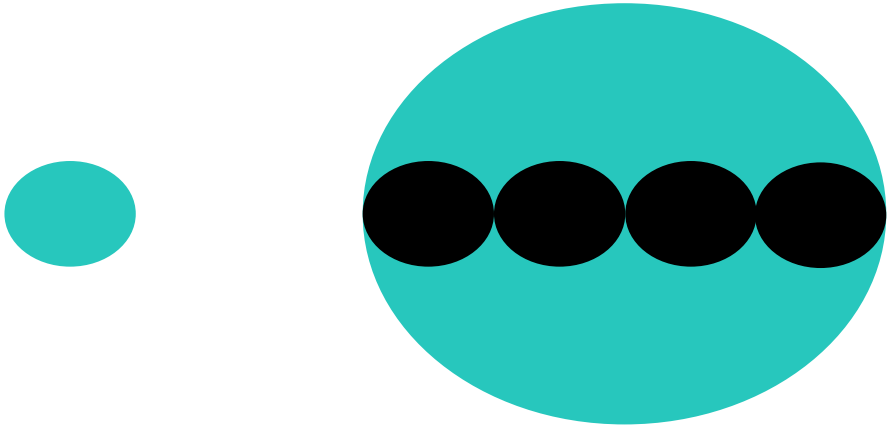
Tufte's Rules

http://www.sealthreinhold.com/school/tuftes-rules/

Perception of magnitudes



Compare area of circles



Compare area of circles

Perception of magnitudes

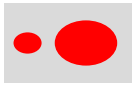
Most accurate

Position

Length

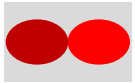
Slope

Angle



Area

Volume



Color hue-saturation-density

Least accurate

Cleveland, McGill (1984)
Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

Perception of color

(128, 128, 128)

(144, 144, 144)

Which is brighter?

Just Noticeable Difference

- JND (Weber's Law)

$$\frac{\Delta I}{I} = k,$$

- I : intensity; ΔI : increase from I; k : constant factor
- Required increase ΔI depends on original intensity I
- Most continuous variations in stimuli are perceived in discrete steps

Use colors wisely

Choose colors based on the information you want to convey

- Sequential
- Diverging
- Categorical

Use online resources to discover and record your color schemes

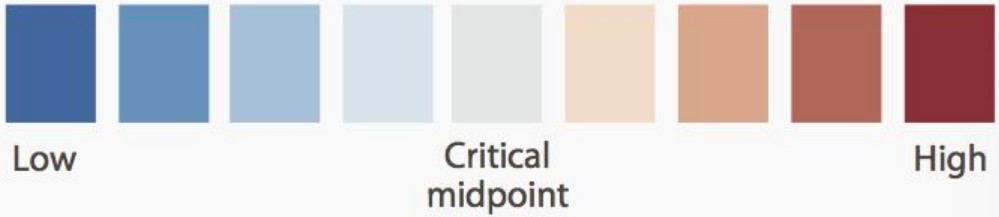
- Color Brewer
- Kuler
- Colour Lovers

Use colors wisely

Use colors wisely

Diverging

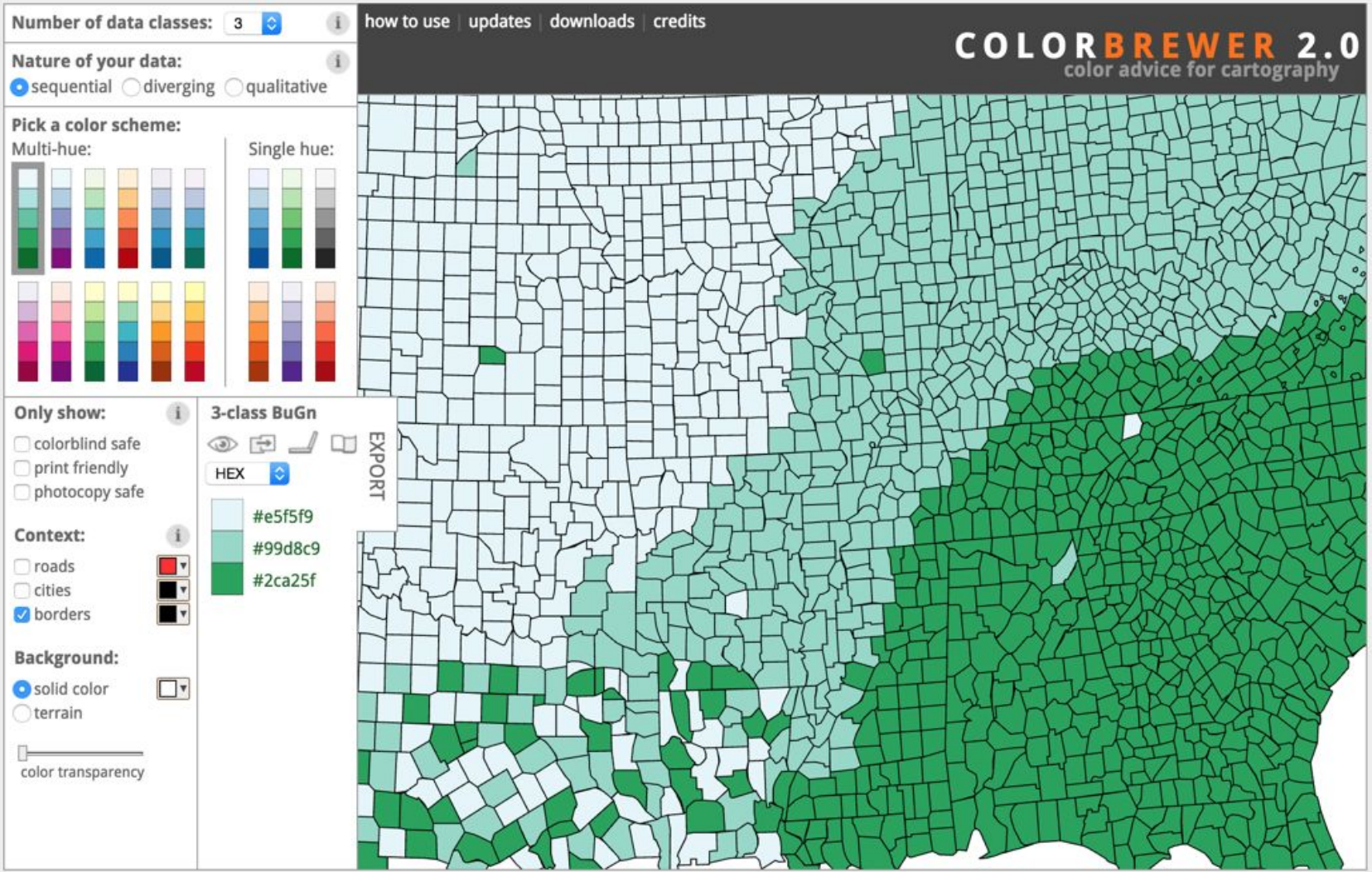
Two sequential schemes extended out from a critical midpoint value



Use colors wisely

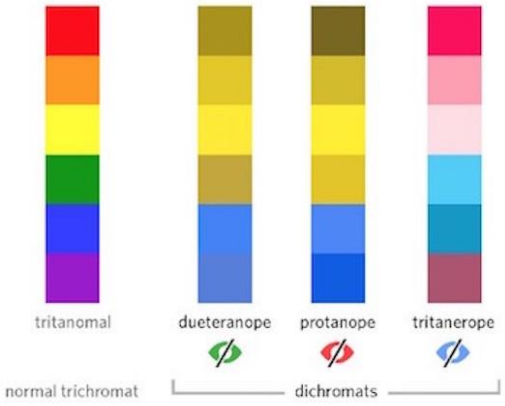
Categorical

Lots of contrast between each adjacent color


Number of data classes: 3
Nature of your data:
sequential diverging qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
3-class BuGn
HEX
#e5f5f9
#99d8c9
#2ca25f
EXPORT
how to use | updates | downloads | credits
COLORBREWER 2.0
color advice for cartography

Colorblind-friendly pallettes

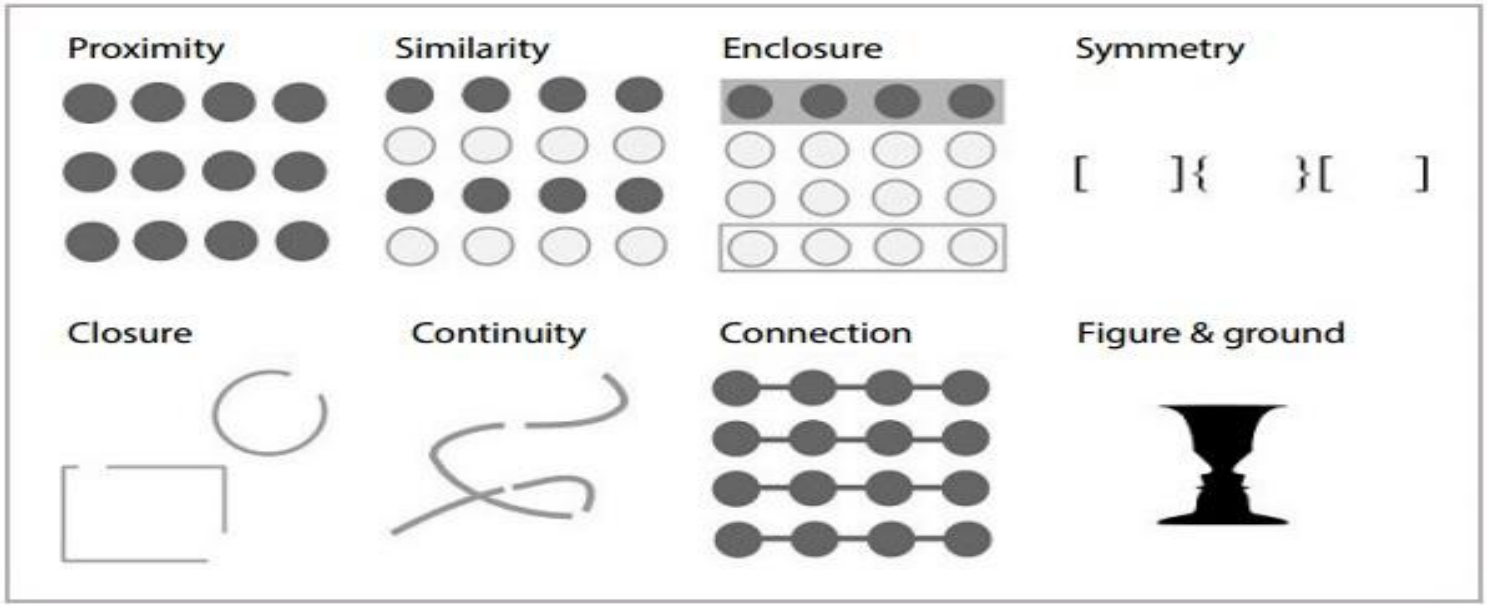
- 10% of males (i.e., ~10% of your reviewers...) have some form of colorblindness



	Original	Simulation: Protan	Simulation: Deutan	Simulation: Tritan		Hue	for Photoshop, Illustrator, Freehand, etc.: C,M,Y,K (%)	for Photoshop, Illustrator, Freehand, etc.: R,G,B (0-255)	for Word, Power Point, Canvas, etc.: R,G,B (%)
1					Black	– °	(0,0,0,100)	(0,0,0)	(0,0,0)
2					Orange	41°	(0,50,100,0)	(230,159,0)	(90,60,0)
3					Sky Blue	202°	(80,0,0,0)	(86,180,233)	(35,70,90)
4					bluish Green	164°	(97,0,75,0)	(0,158,115)	(0,60,50)
5					Yellow	56°	(10,5,90,0)	(240,228,66)	(95,90,25)
6					Blue	202°	(100,50,0,0)	(0,114,178)	(0,45,70)
7					Vermilion	27°	(0,80,100,0)	(213,94,0)	(80,40,0)
8					reddish Purple	326°	(10,70,0,0)	(204,121,167)	(80,60,70)

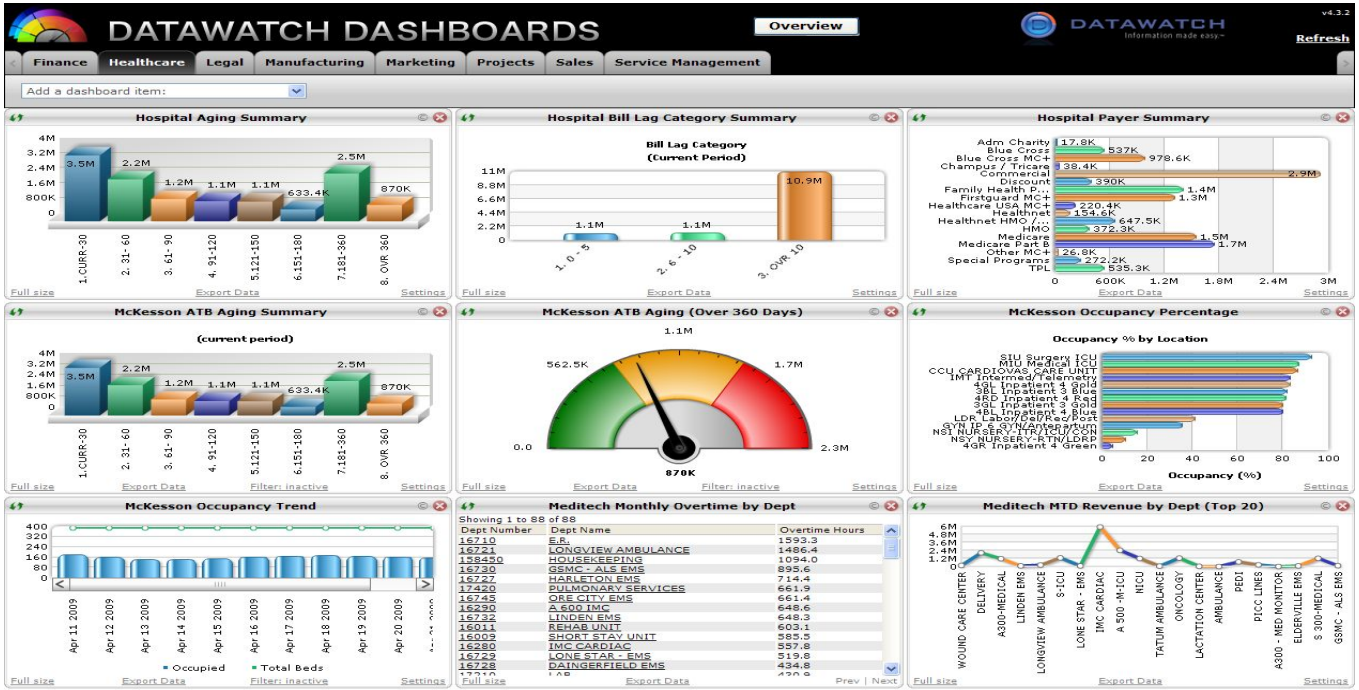
Use structure

Gestalt psychology principles (1912)



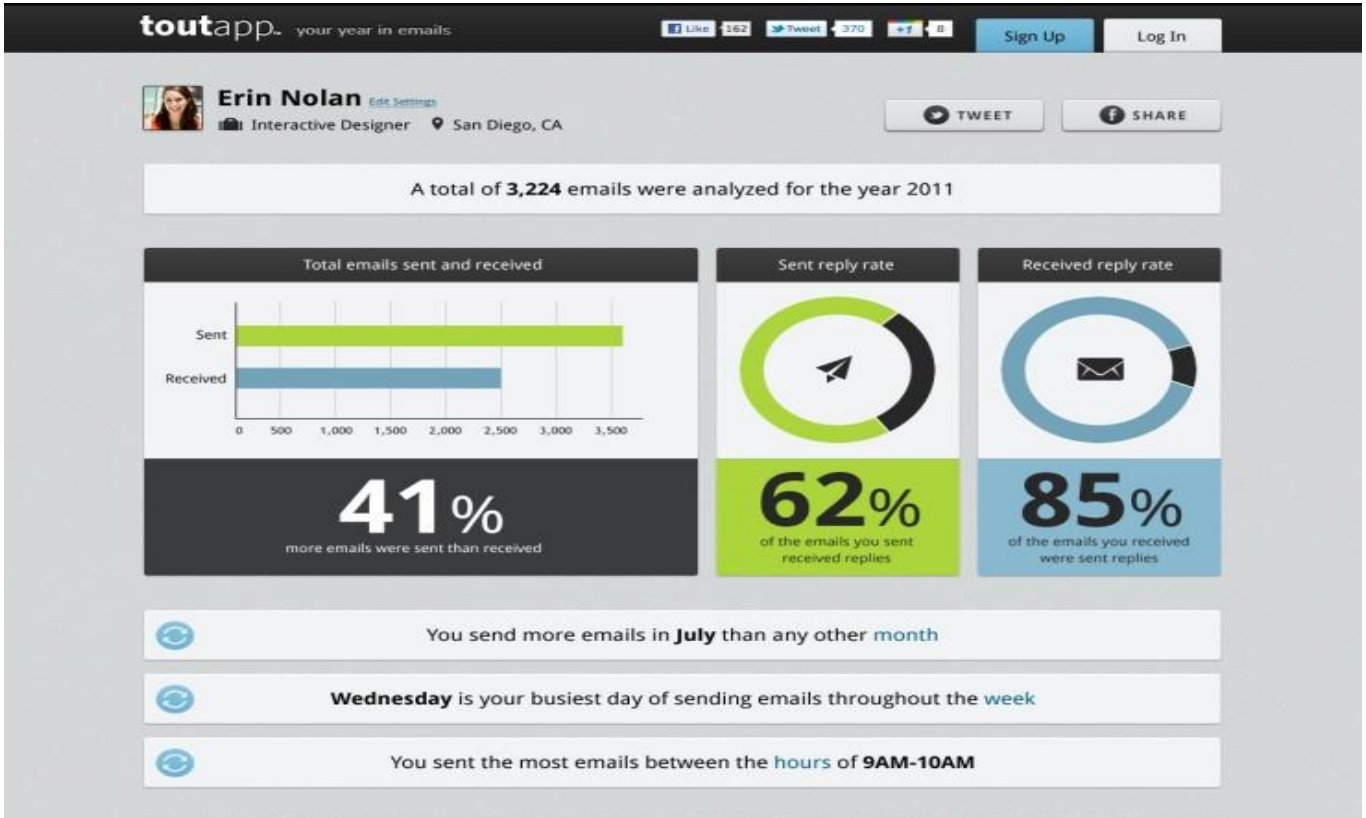
Source http://blog.fusioncharts.com/2014/03/how-to-use-the-gestalt-principles-for-visual-storytelling-podv/

Use structure
(but not like this!)



Source
https://www.vocalabs.com/blog/my-dashboard-pet-peeve

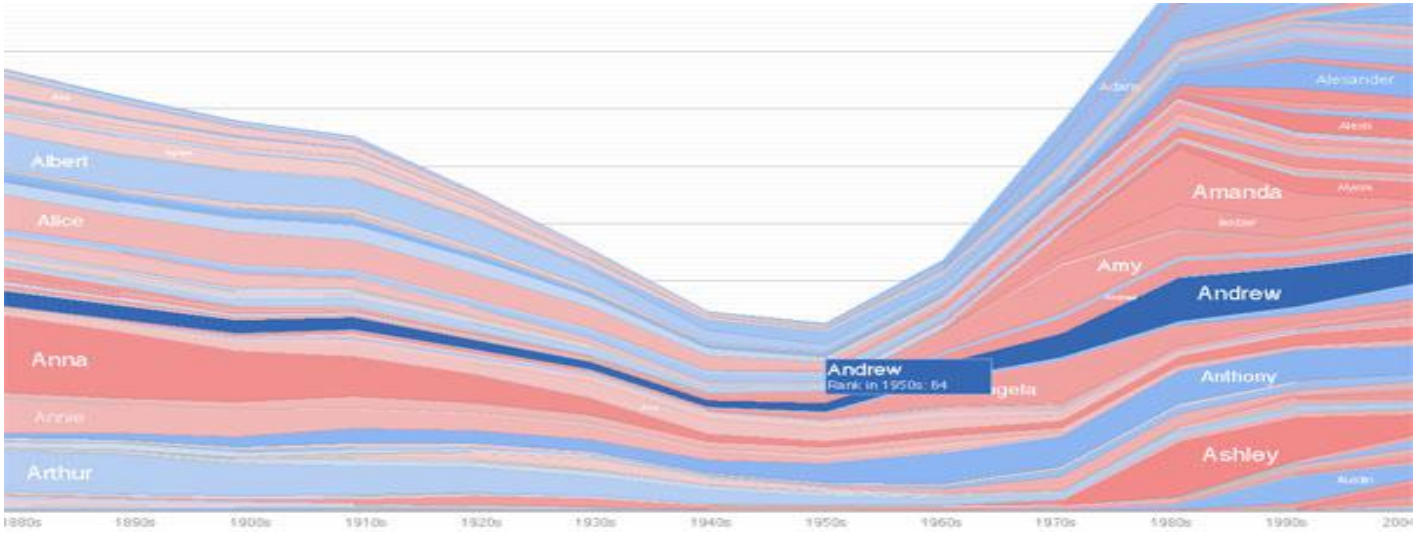
Less is more



Source
https://www.vocalabs.com/blog/my-dashboard-pet-peeve

Interactive chart design: simplifying

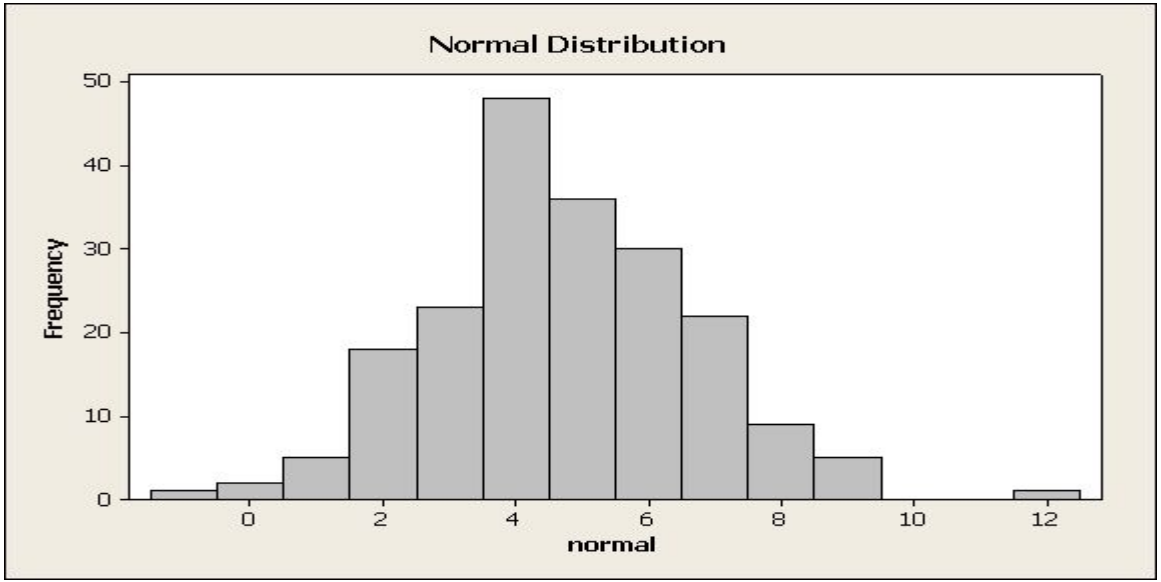
- With interactive charts you can keep things very simple by **hiding** and **dynamically revealing** important structure.
- On an interactive chart, you reveal the information most useful for **navigating** the chart.



Visualization for data exploration

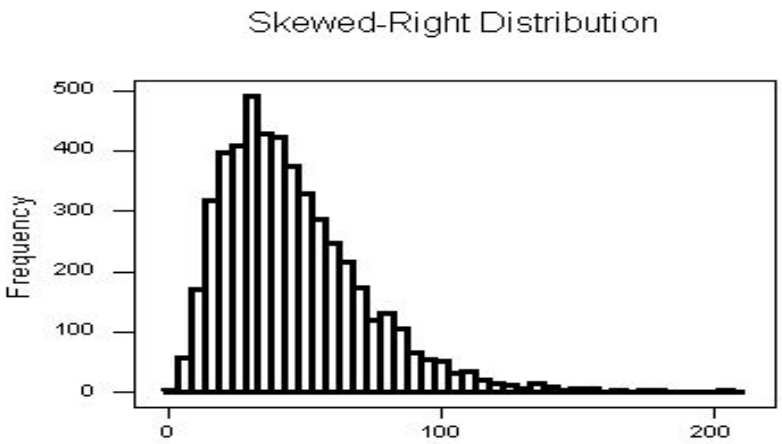
Histograms

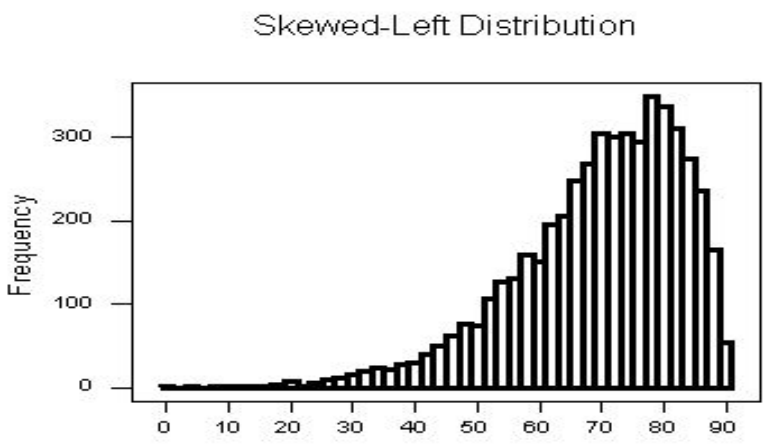
Histograms can tell you a lot about a single variable, discrete or continuous



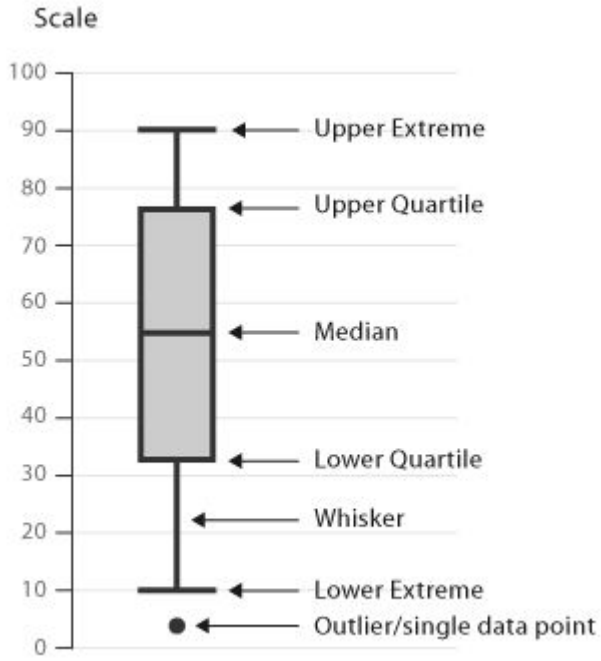
Histograms

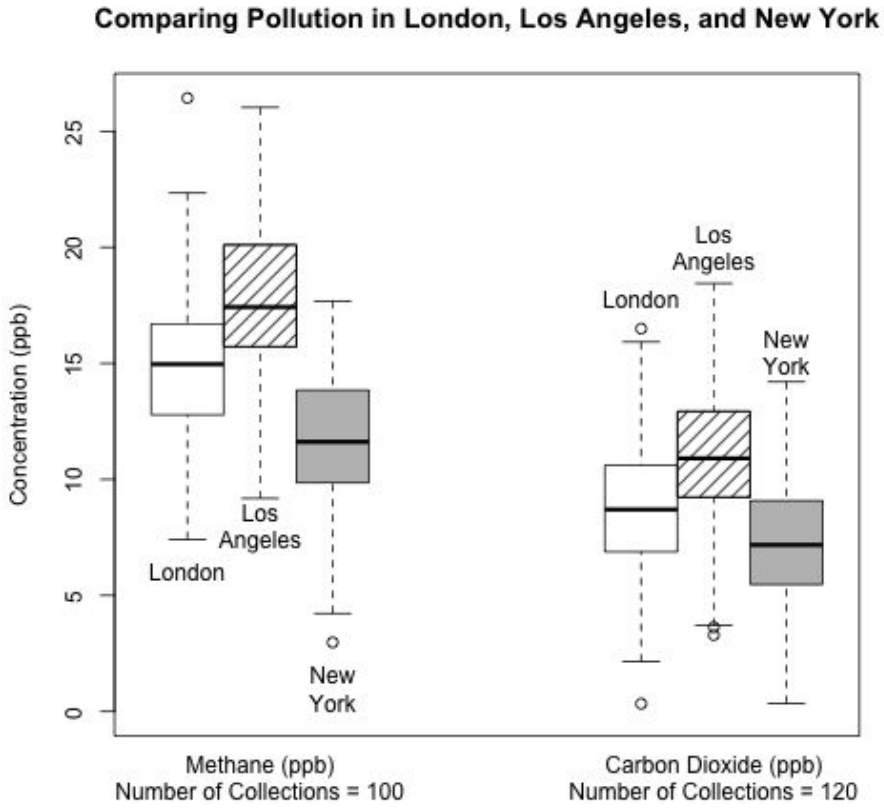
Skewed distributions



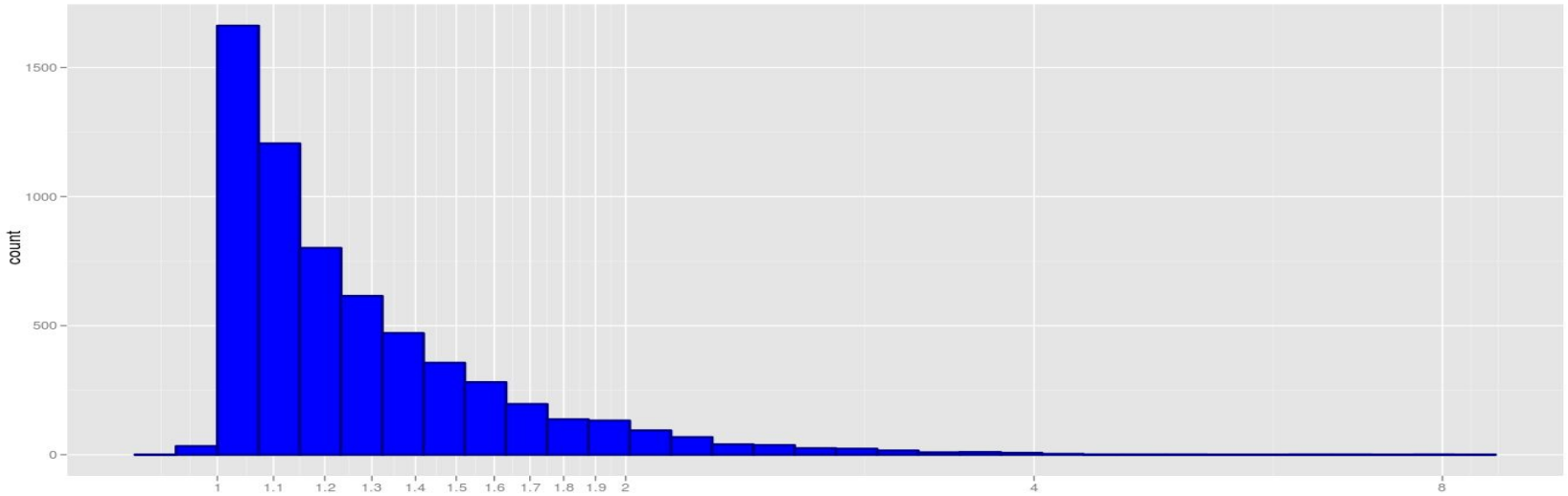


Box plots


Scale
100
90
80
70
60
50
40
30
20
10
0
Upper Extreme
Upper Quartile
Median
Lower Quartile
Whisker
Lower Extreme
Outlier/single data point

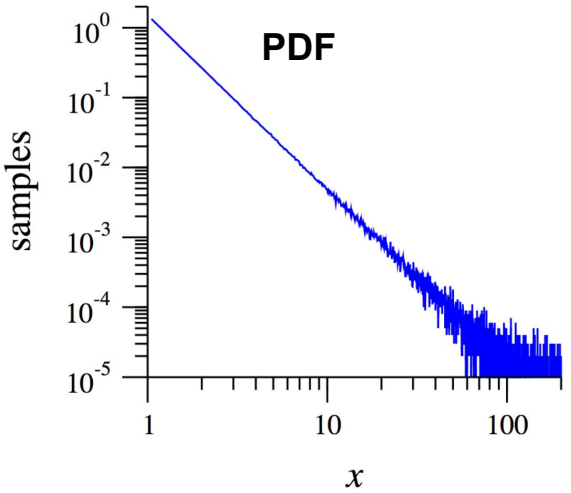

Comparing Pollution in London, Los Angeles, and New York
Concentration (ppb)
0
5
10
15
20
25
London
Los Angeles
New York
London
Los Angeles
New York
Methane (ppb)
Number of Collections = 100
Carbon Dioxide (ppb)
Number of Collections = 120

Heavy-tailed data



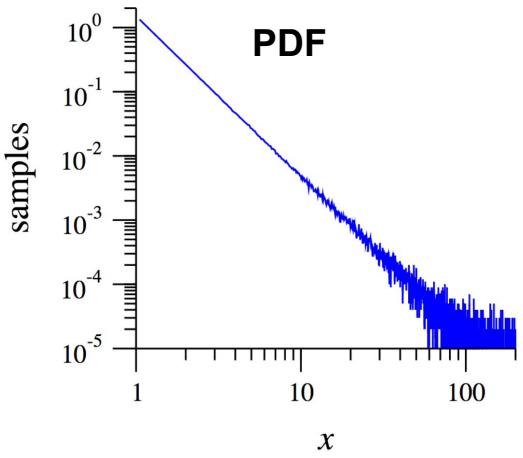
Heavy-tailed data: power laws

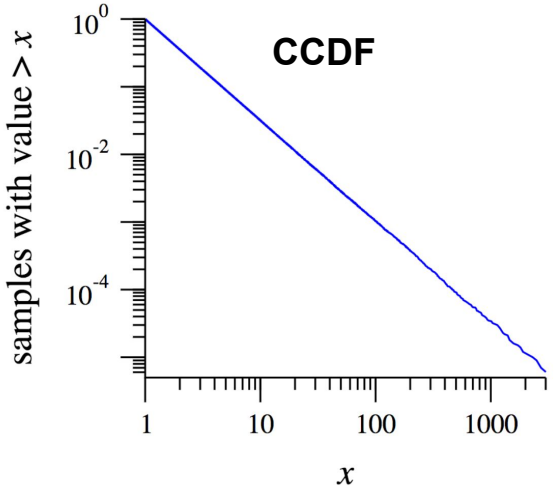
- $p(x) = Cx^{-\alpha},$
  - very very large values are rare, “but not very rare”
  - E.g. Number of friends
  - Many natural phenomena are power laws
  - For dealing with them, need to know some tricks
  - E.g., straight line on log-log axes:
    - $y = C\, x^{-\alpha} \leftrightarrow \log(y) = \log(C) - \alpha \log(x)$



Heavy-tailed data: power laws

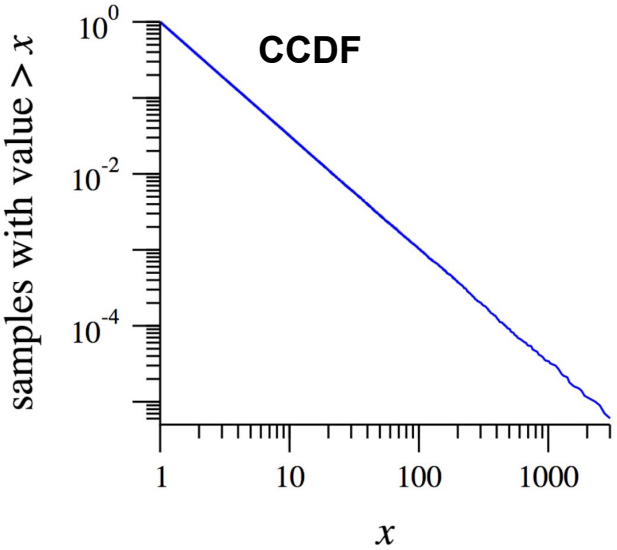
- Complementary cumulative distribution function (CCDF):
  - P(x) := Pr{X >= x}
- CCDF of power law is also a power law (with exponent α – 1)
  - $$P(x) = C \int_x^\infty x'^{-\alpha}\, \mathrm{d}x' = \frac{C}{\alpha - 1} x^{-(\alpha-1)}.$$



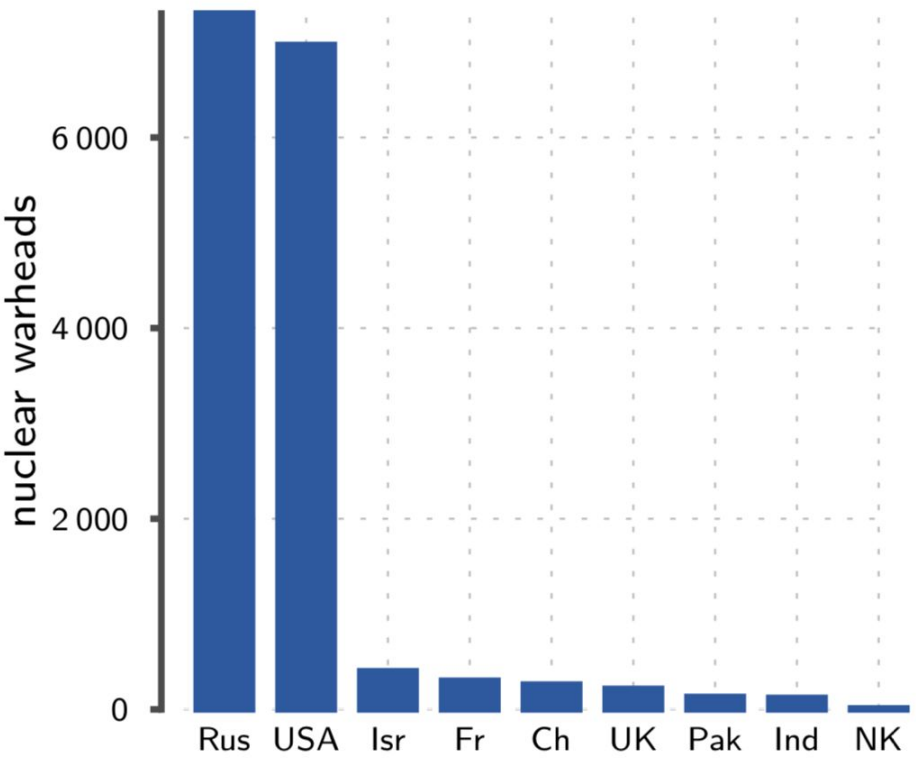


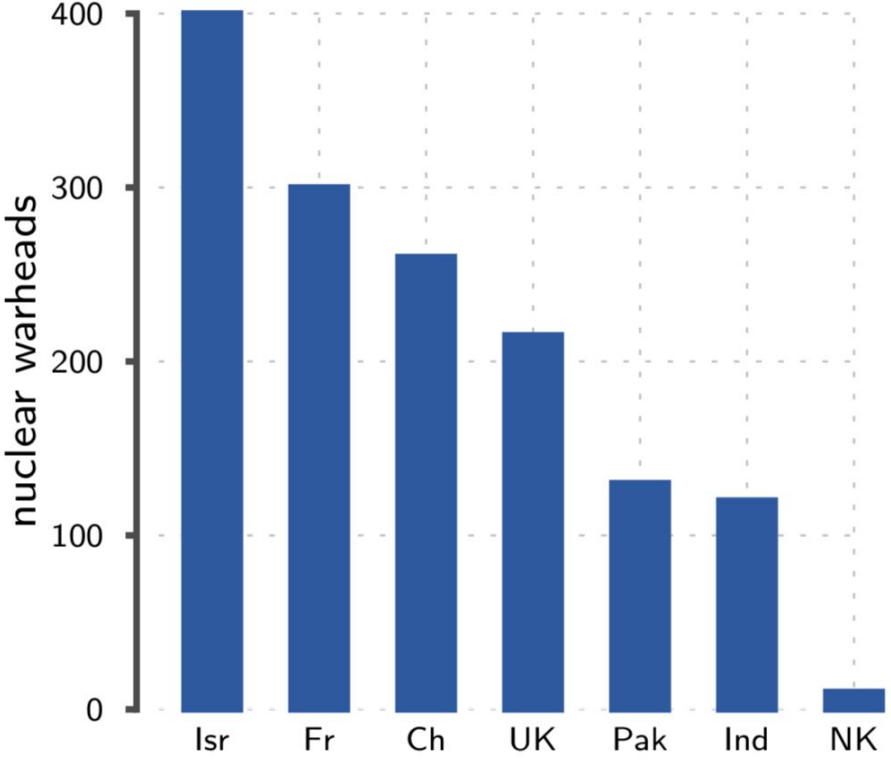
Heavy-tailed data: power laws

- Smart trick for plotting CCDF of any distribution:
  - x-axis: data sorted in ascending order
  - y-axis: (n:1)/n (where n is number of data points)

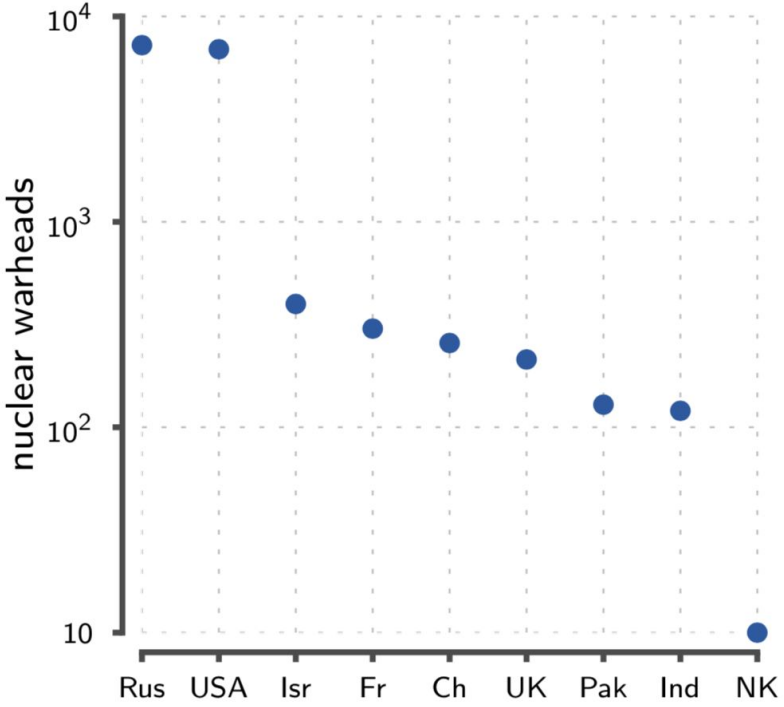


Log scale


nuclear warheads
6 000
4 000
2 000
0
Rus
USA
Isr
Fr
Ch
UK
Pak
Ind
NK

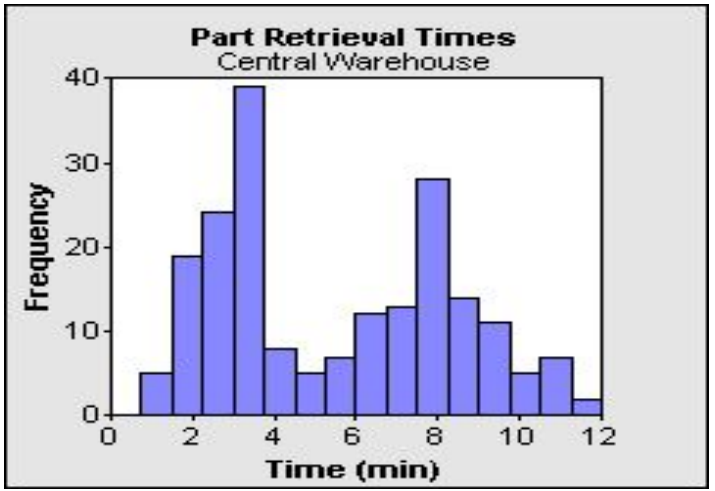

nuclear warheads
400
300
200
100
0
Isr
Fr
Ch
UK
Pak
Ind
NK

Log scale



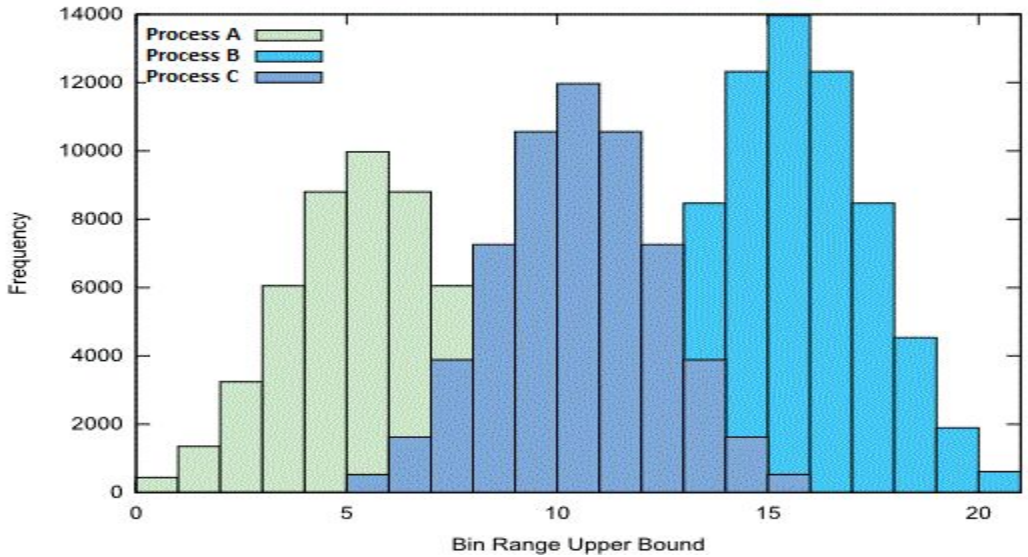
Multimodal data

- Two or more distinct peaks in a histogram.
- **Suggests two or more distinct populations** of samples.
- Often arise from gender/political views, or other binary factors.
- But don’t guess! Explore further by using, e.g., color and a histogram of multiple populations.



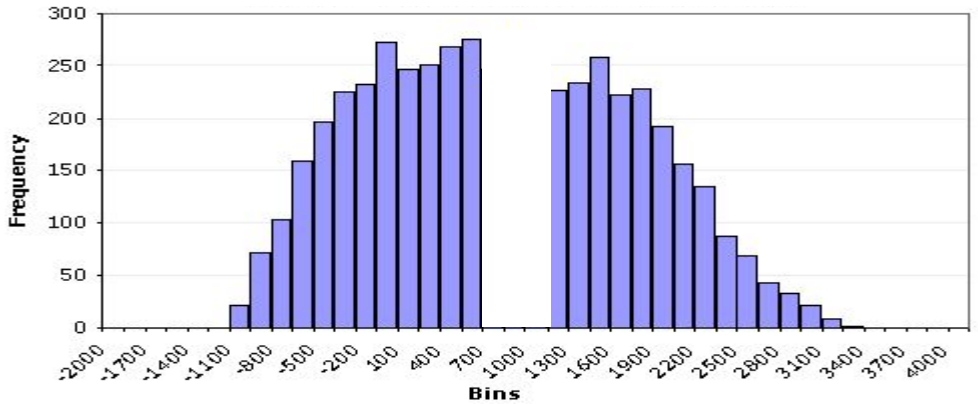
Multimodal data

Explore further by using, e.g., color and a histogram of multiple populations



Weird data

- Some data is **very hard to explain**.
- Don’t guess! Trace through the data pipeline to find where the strangeness comes from. Usually it’s a processing bug.

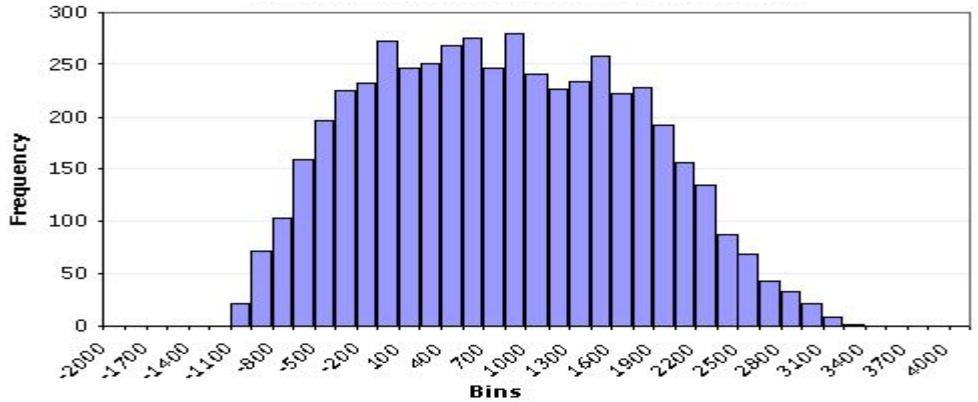


Proactive “weird-data detection”

If data looks ok, take a picture and save it for later…

Then **periodically compare new data with old** whenever there is a pipeline update.

Always try to have a **theory of what the data should look like**.

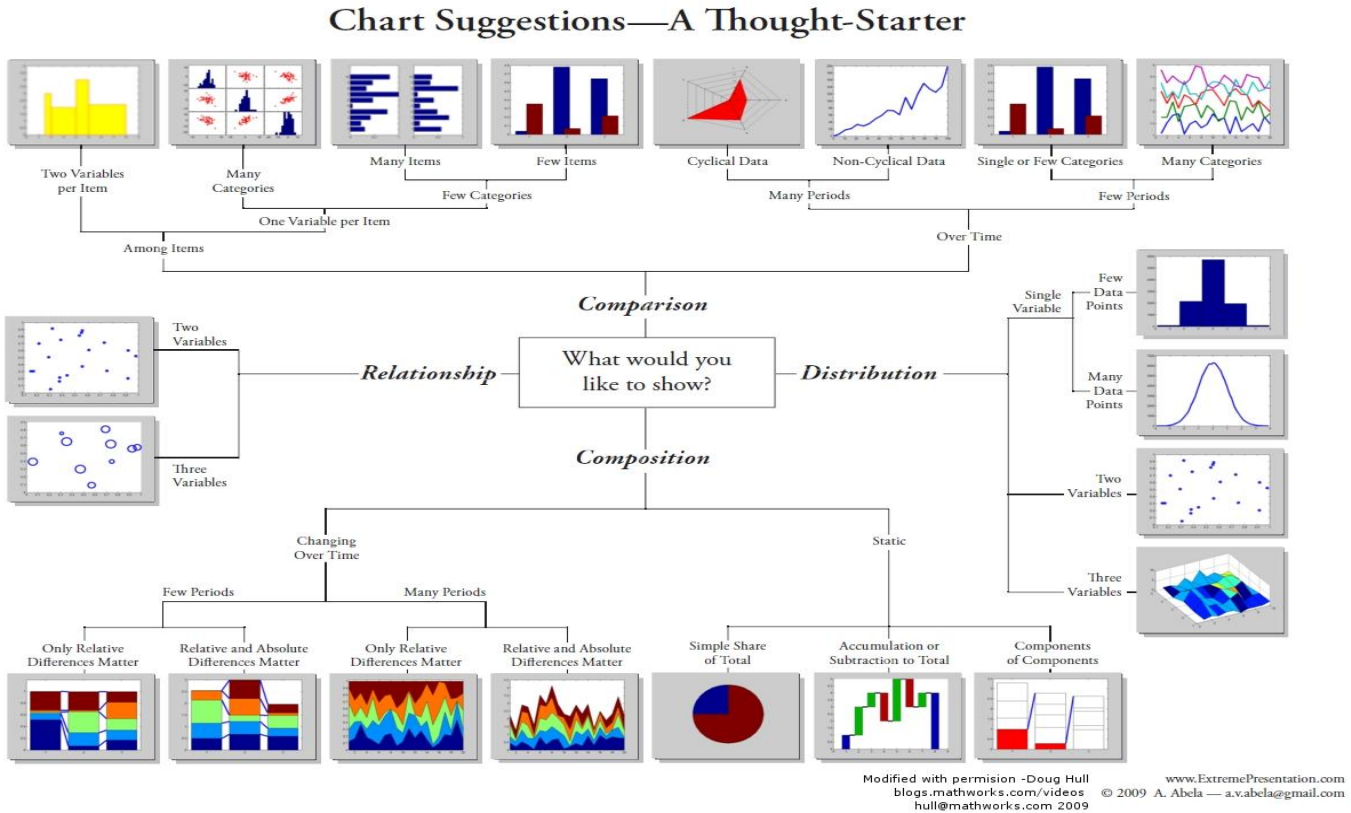


Remarks on exploration

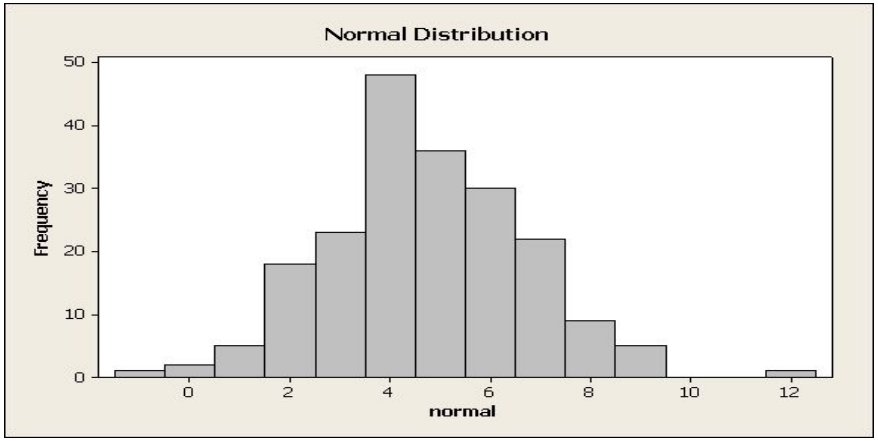
- **Form expectations** of what the data should look like. This helps you guard against pipeline errors and to identify interesting patterns
- But **expect the unexpected**

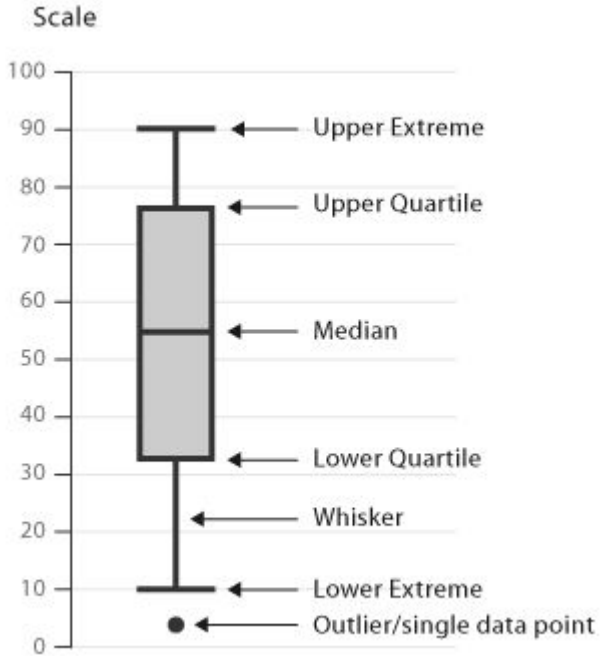
Navigating the chart landscape

Chart selection (by Andrew Abela)



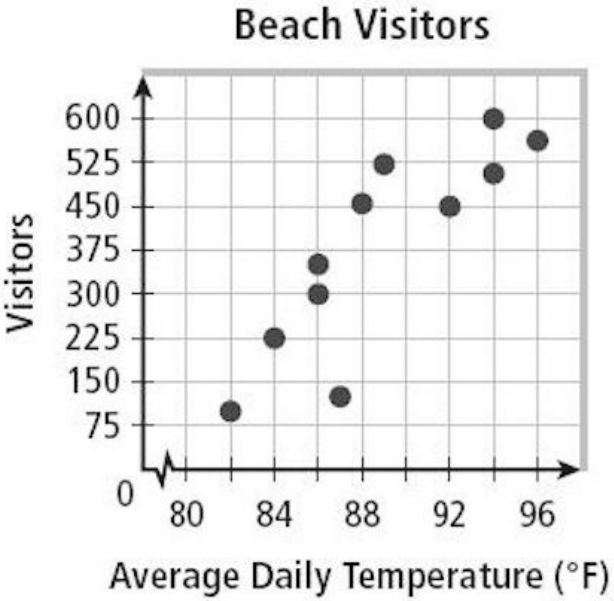
One variable: histograms, box plots

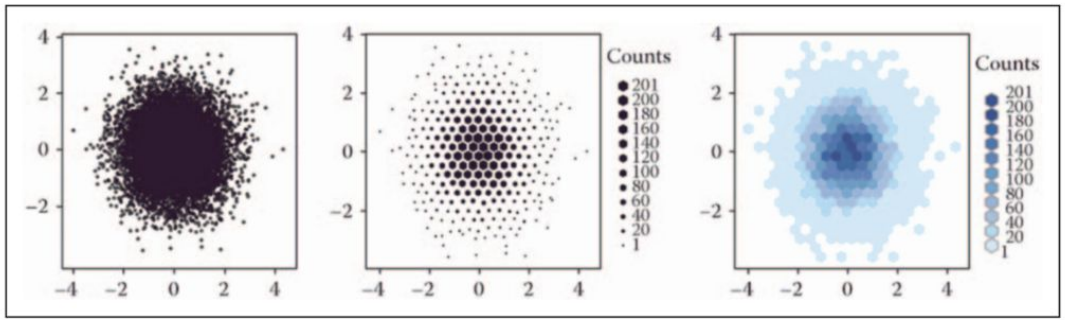




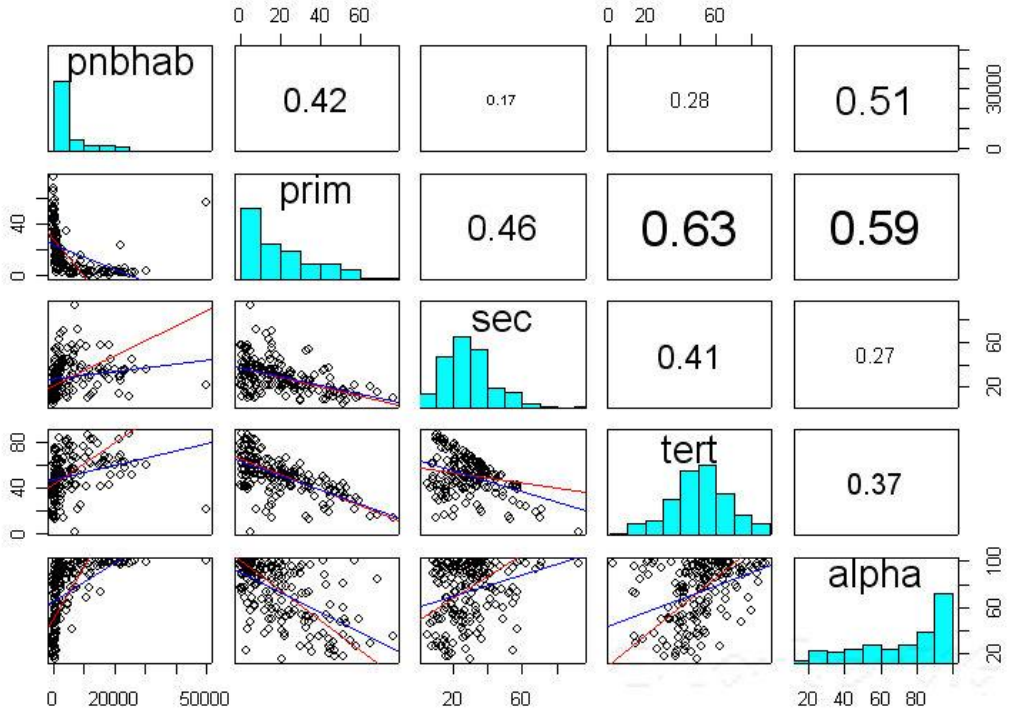
Two variables: scatter plots

Scatter plots quickly expose the relationships between two variables





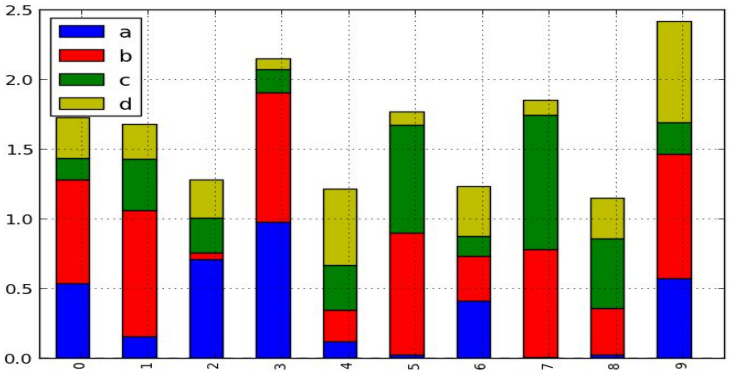
> 2 variables: scatter plot matrix

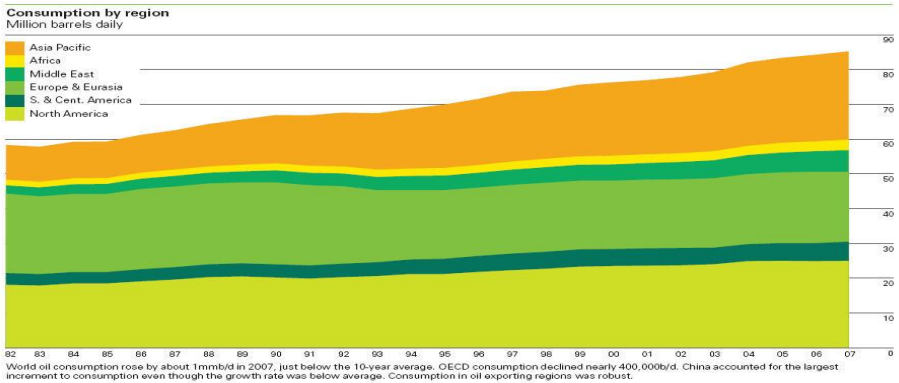


> 2 variables: stacked plots

Stack index, color, height

Stack variable and color variable: categorical

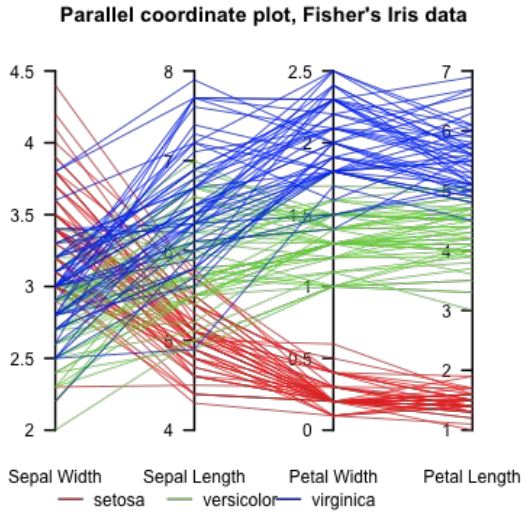




> 2 variables: parallel-coord. plots

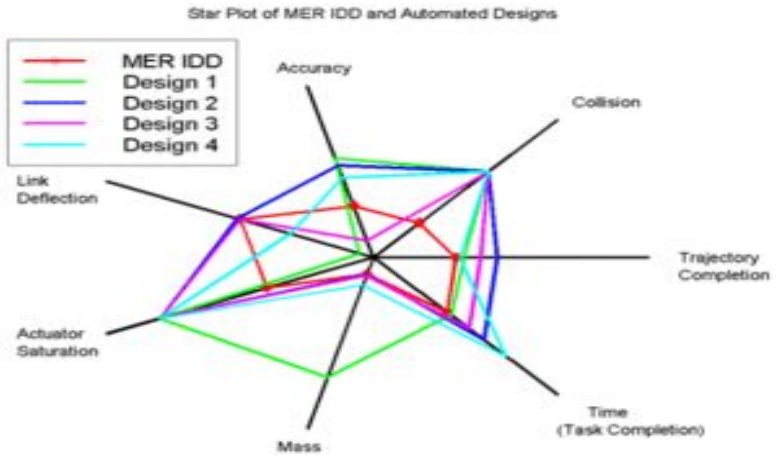
Color, x, y

Color variable is categorical, others arbitrary



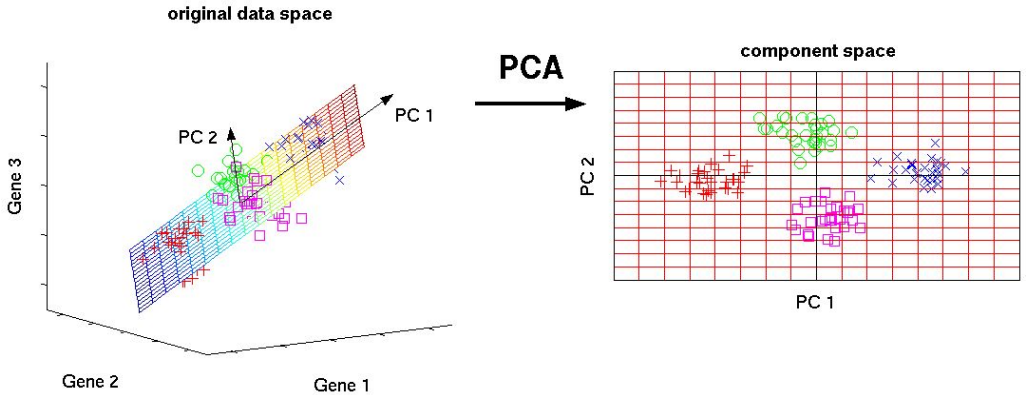
> 2 variables: radar charts

- Similar to parallel-coord. plots
- Doesn’t pretend that x axis has meaningful order
- Good for periodic data

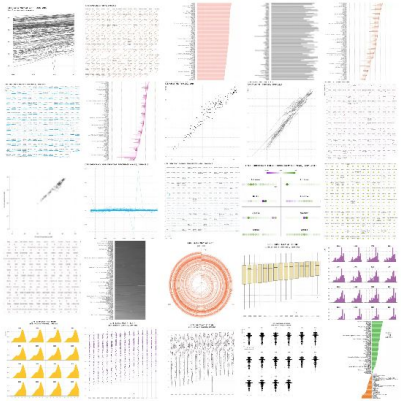


Dimensionality reduction

- **One example, PCA:** allows visualization of high-dimensional continuous data in 2D using principal components
- The principal components are the strongest (highest variation) dimensions in the dataset, and are orthogonal



One Dataset, visualized 25 ways

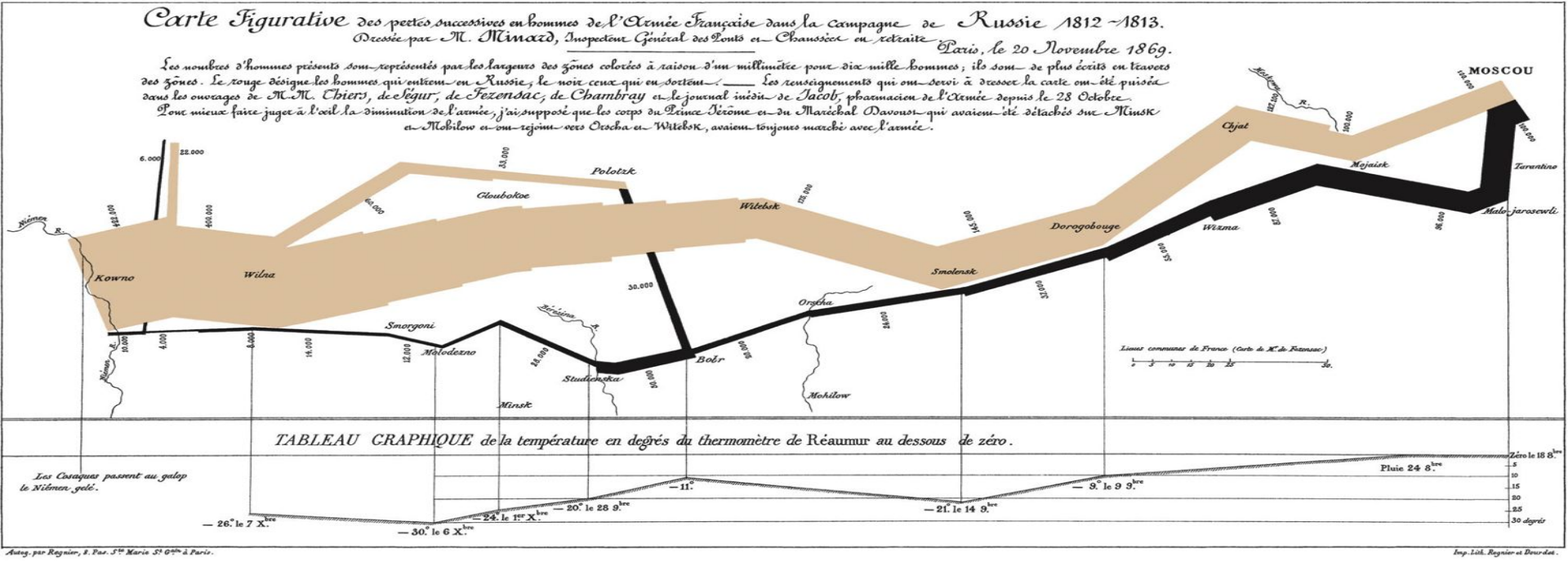


http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/

"You must help the data focus and get to the point. Otherwise, it just ends up rambling about what it had for breakfast this morning and how the coffee wasn't hot enough."

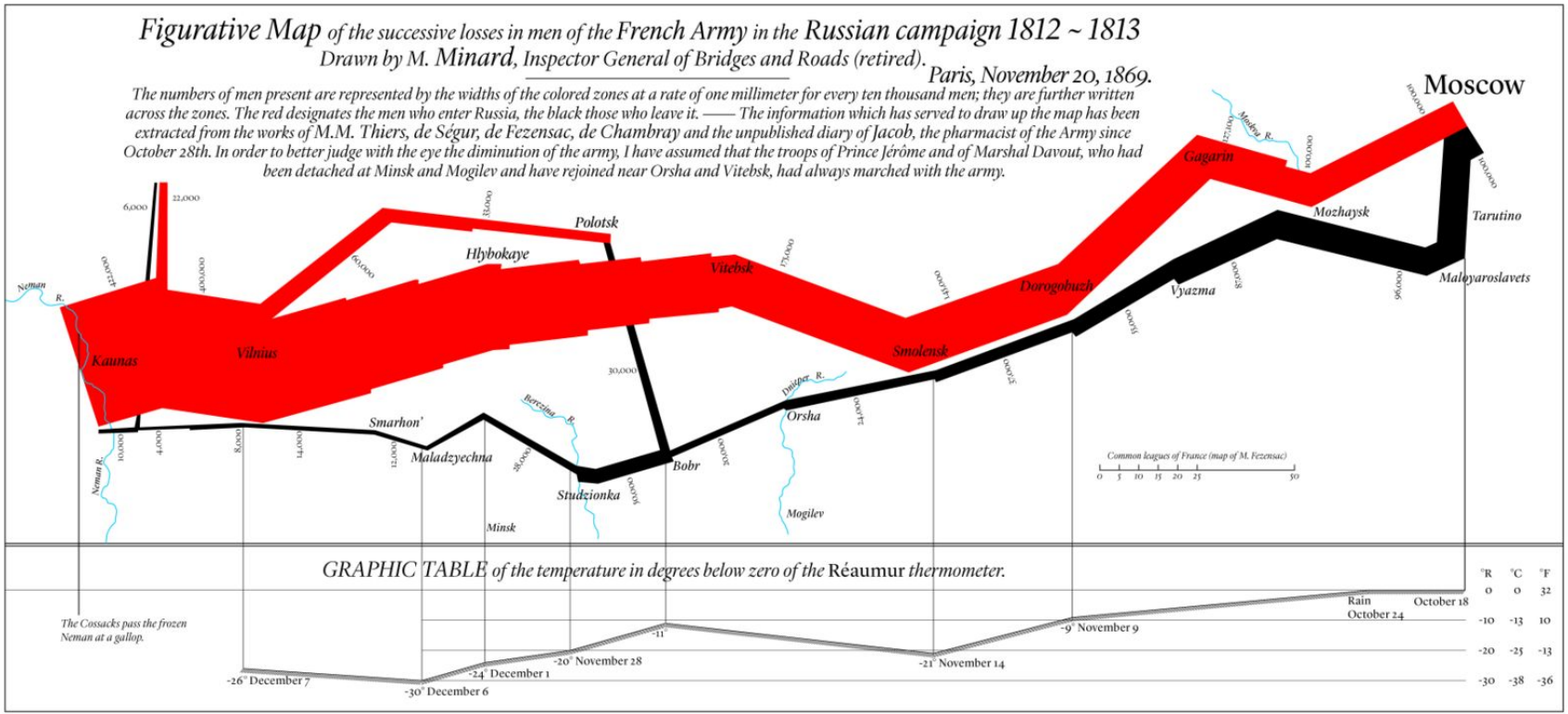
Good examples

Charles Joseph Minard 1869
Napoleon's march



According to Tufte: "*It may well be the best statistical graphic ever drawn.*"
5 variables: army size, location, dates, direction, temperature during retreat

In English


Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813
Drawn by M. Minard, Inspector General of Bridges and Roads (retired).
Paris, November 20, 1869.
The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. —— The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.
Moscow
Gagarin
Mozhaysk
Tarutino
Maloyaroslavets
Vyazma
Dorogobuzh
Smolensk
Vitebsk
Polotsk
Hlybokaye
Kaunas
Vilnius
Neman R.
Smarhon'
Maladzyechna
Neman R.
Berezina R.
Studzionka
Bobr
Minsk
Dnieper R.
Orsha
Mogilev
Moskva R.
Common leagues of France (map of M. Fezensac)
0 5 10 15 20 25 50
GRAPHIC TABLE of the temperature in degrees below zero of the Réaumur thermometer.
°R °C °F
0 0 32
-10 -13 10
-20 -25 -13
-30 -38 -36
The Cossacks pass the frozen Neman at a gallop.
-26° December 7
-30° December 6
-24° December 1
-20° November 28
-11°
-21° November 14
-9° November 9
Rain October 24
October 18

Interactivity to educate

Hans Rosling:

200 Countries, 200 Years, 4 Minutes

https://www.youtube.com/watch?feature=player_embedded&v=jbkSRLYSojo

Examples: Map based visualizations

Map-based visualizations, such as NYTimes visualization of your friends location on Facebook (US)

https://www.nytimes.com/interactive/2018/09/19/upshot/facebook-county-friendships.html

The future of journalism?

NY Times interactive visualizations (recession/recovery 2014)

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

https://www.nytimes.com/interactive/2018/03/19/upshot/race-class-white-and-black-men.html

And 2014 “the year in interactive storytelling”

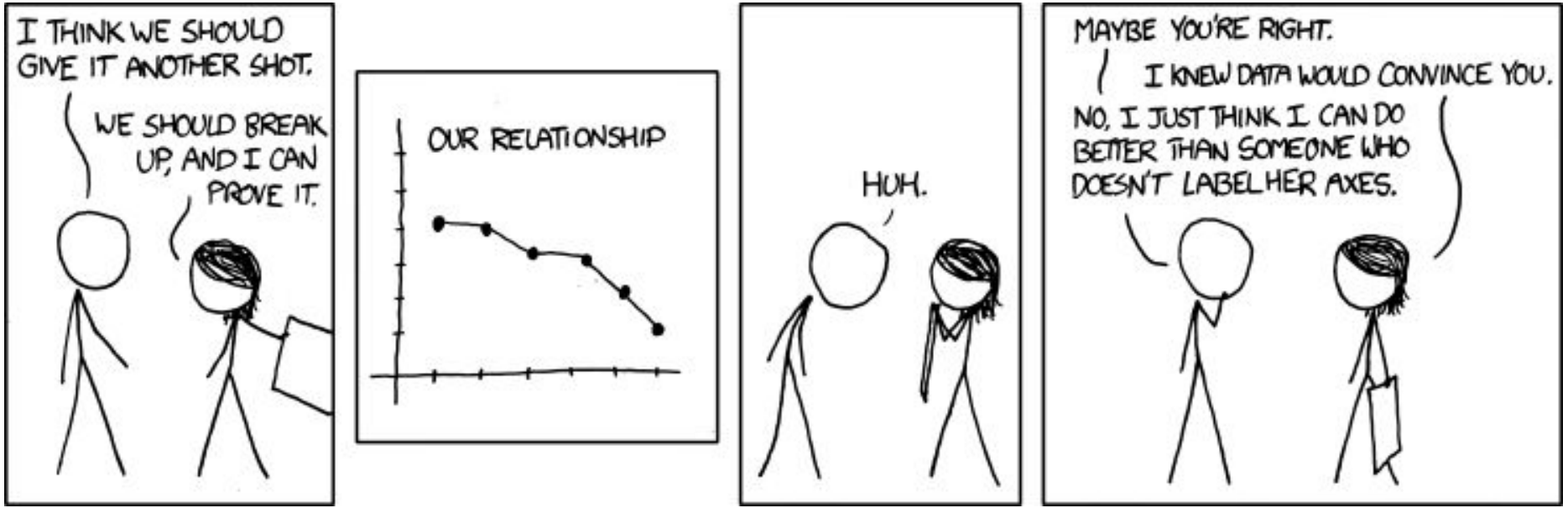
http://www.nytimes.com/interactive/2014/12/29/us/year-in-interactive-storytelling.html?_r=0

NY Times graphics are a great source of **best practices** in viz (except for when they’re not…)

Bad examples

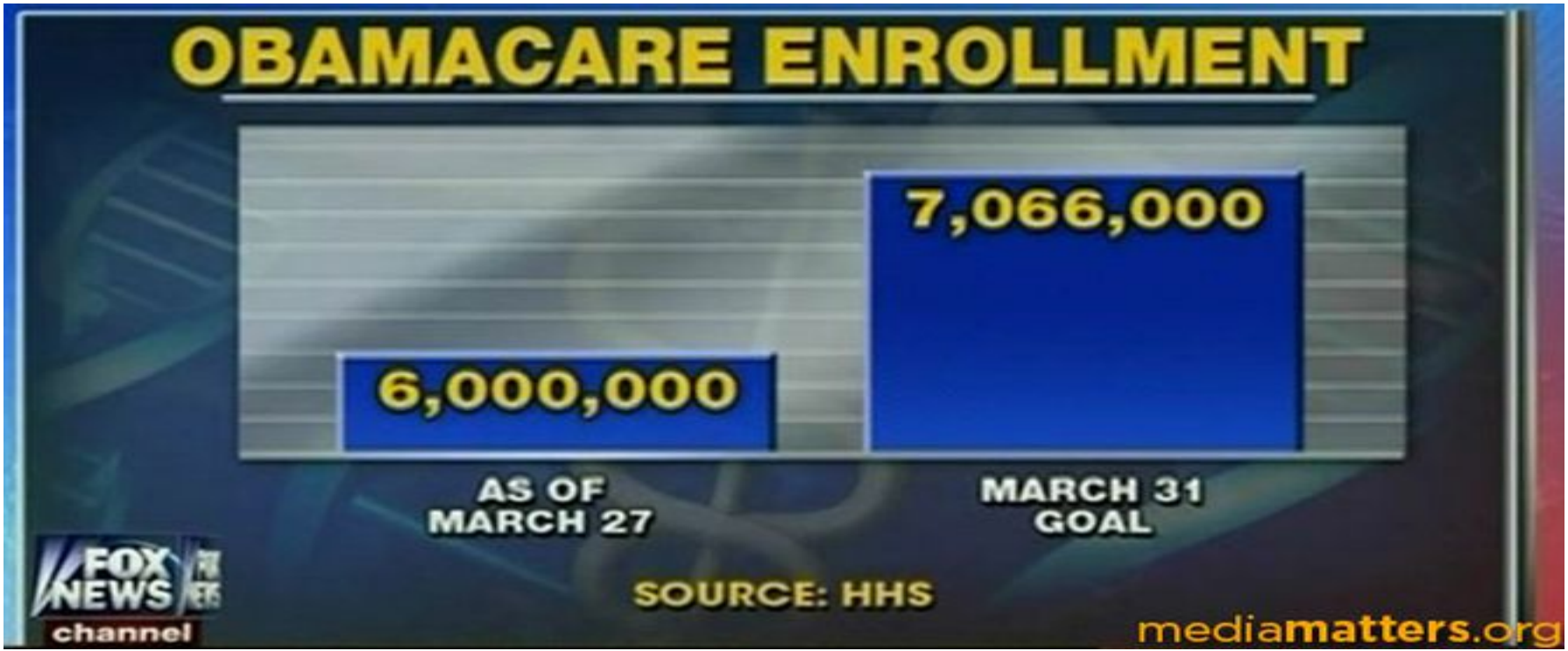
Courtesy of viz.wtf

Label your axes

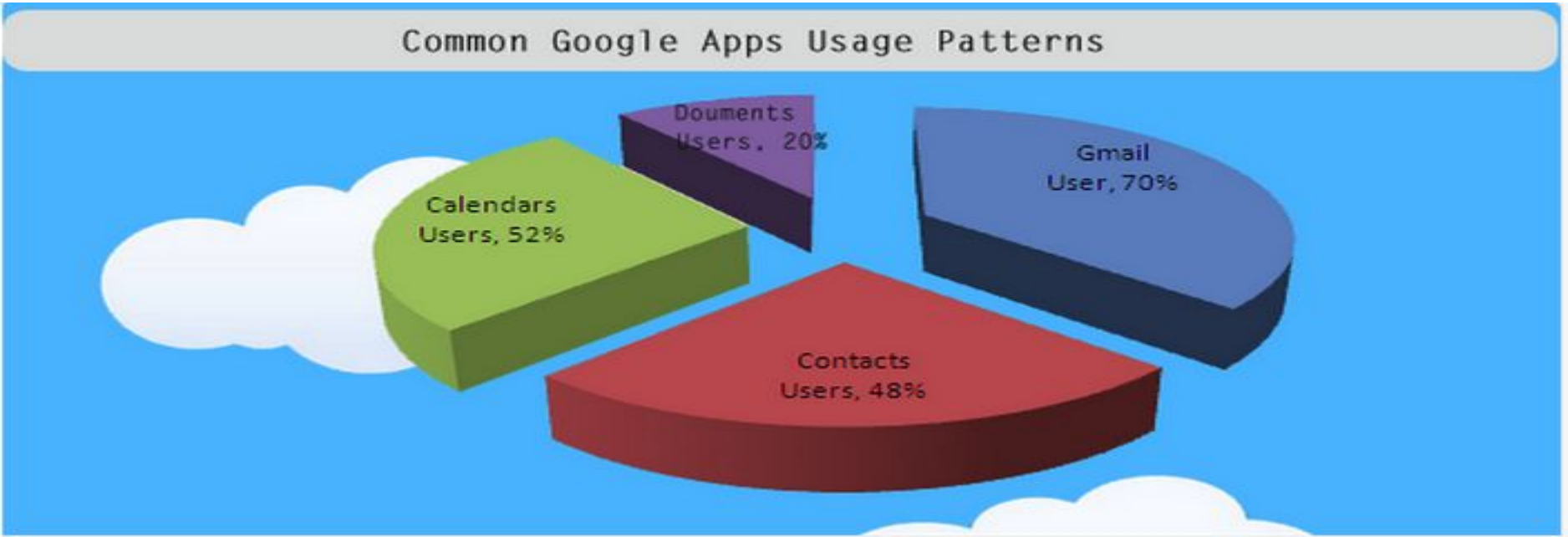


https://www.xkcd.com/833/

Visualization to educate?



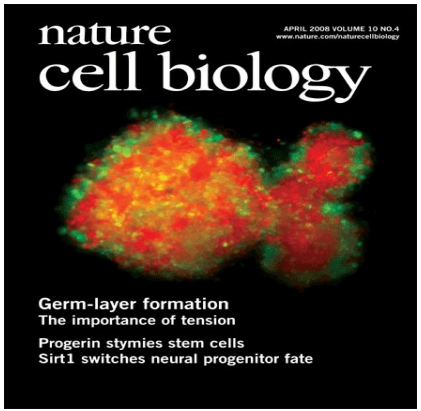
Pie in the sky?

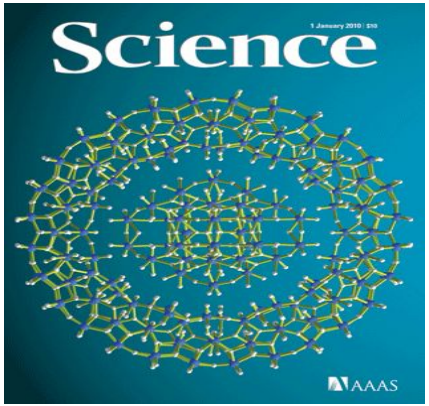


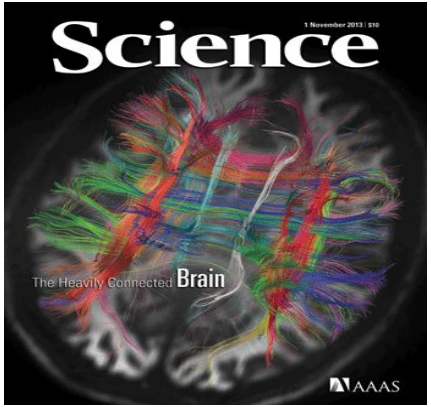
90% of US Households Consume Peanut Butter

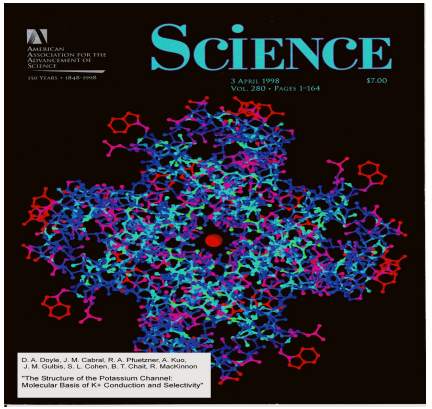
Needs fixing

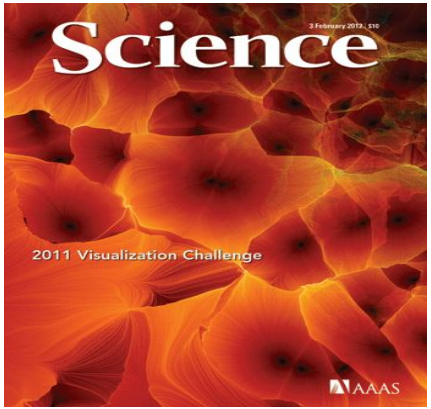
Data viz in the sciences











National Science Foundation
WHERE DISCOVERIES BEGIN

QUICK LINKS

SEARCH

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT NSF | FASTLANE

THE VIZZIES
VISUALIZATION CHALLENGE

The most beautiful visualizations from the worlds of science and engineering

Home | About | Timeline | Categories | Guidelines | Prizes | Winners

Important Dates

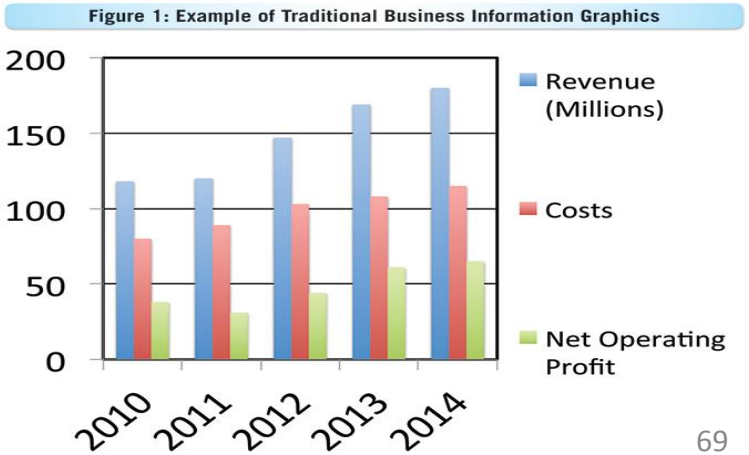
- The Visualization Challenge competition closes September 30, 2014.
- The deadline for all entries is 11:59 p.m. PST on Sept. 30, 2014.
- Competition judging rounds take place in October 2014.
- The 2014 winning entries will be announced in March 2015.
- Contest results will be publicly announced in *Popular Science* and on popsci.com in March 2015, and *Popular Photography* will recognize the winning photo. NSF will also publish the names of the winners on its website.

A case for ugly visualizations

People instinctively gravitate to attractive visualizations, and they have a better chance of getting on the cover of a journal.

But does this **conflict with the goals of visualization?**

- Rapid exploration
- Focus on most important details
- Easy and fast to develop and customize



Tools

Interactive toolkits: D3

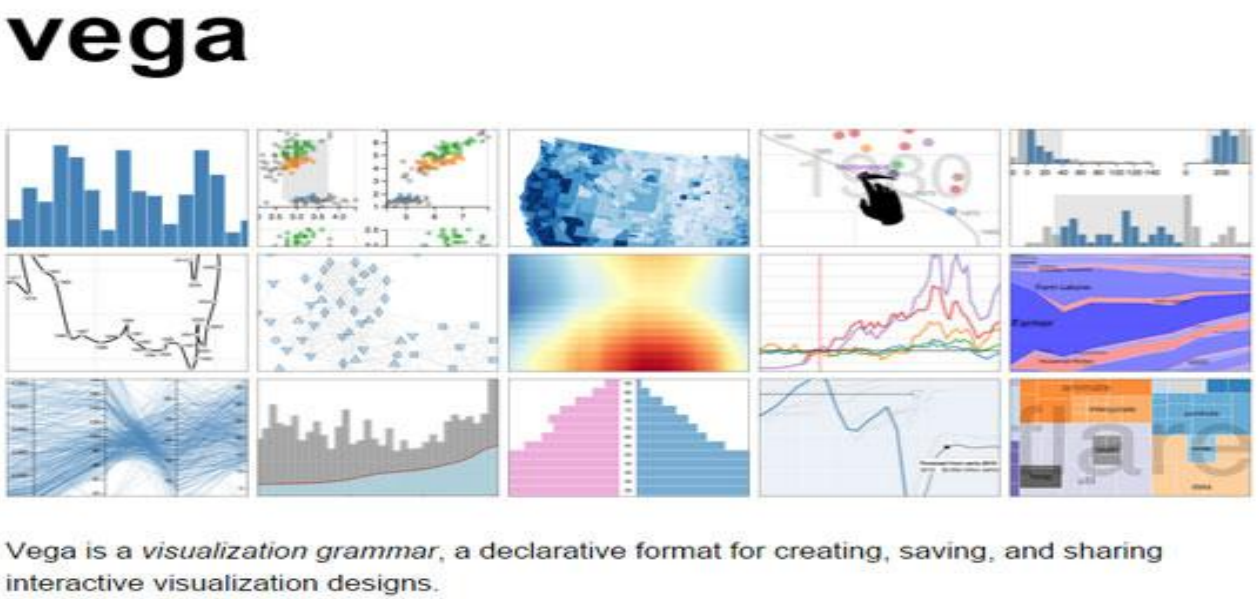
Without doubt, the most widely used interactive visualization framework is **D3**, developed around 2011 by Jeff Heer, Mike Bostock, and Vadim Ogievetsky.

Note from the authors: *D3 is intentionally a low-level system. During the early design of D3, we even referred to it as a "visualization kernel" rather than a "toolkit" or "framework"*

Interactive toolkits: Vega

Vega is a “visualization grammar” developed on top of D3.js

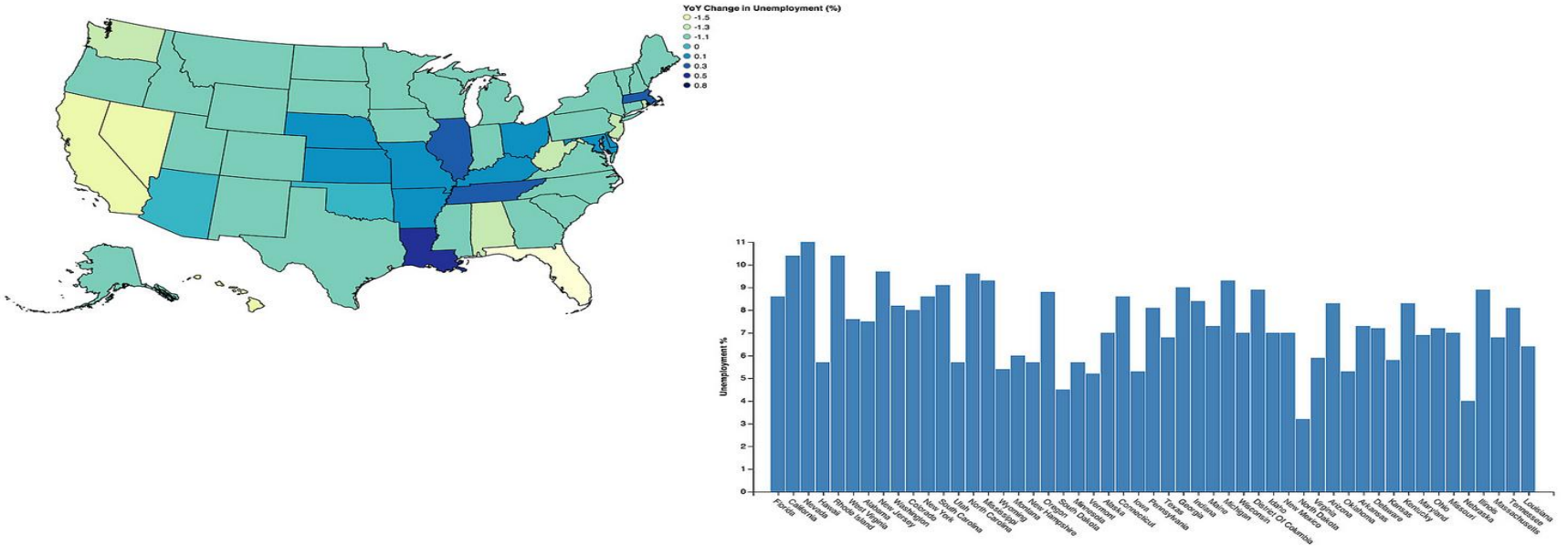
It specifies graphics in JSON format.



Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

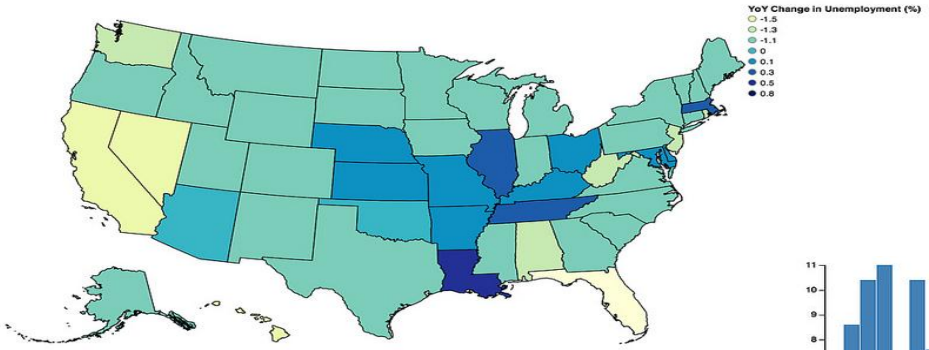
Trivia question: why is it called Vincent? Hint: Vincent+Vega= ?

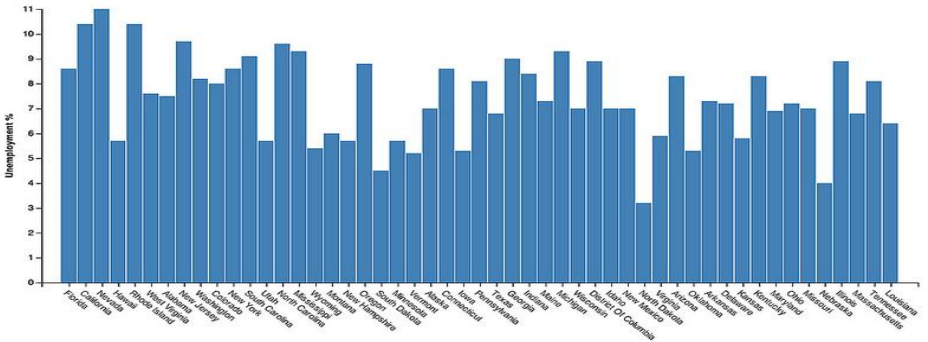


Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

Trivia question: why is it called Vincent? Hint: Vincent+Vega= ?



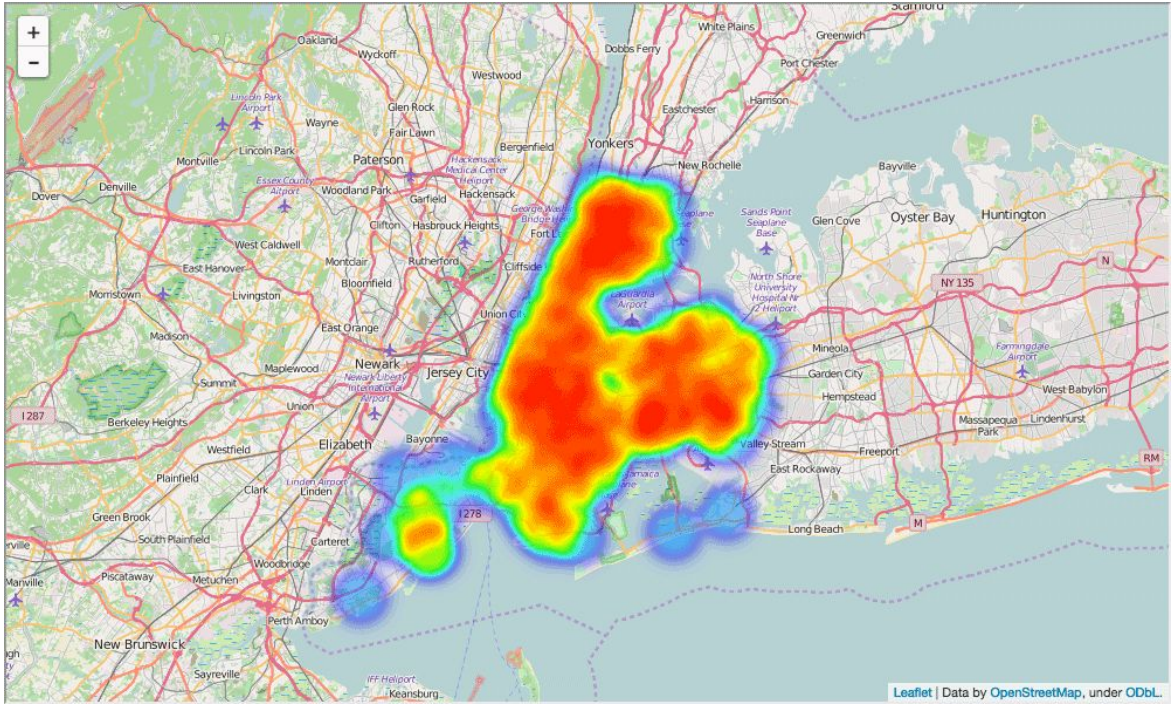


Bokeh: another interactive viz library

Bokeh is an independent Viz library focused more heavily on big data visualization. Has both Python and Scala bindings.

Visualizing maps: Folium

More in tomorrow's lab session!



Visualizing graphs

https://graphviz.org/

https://gephi.org/

Credits

- Last year's version of these slides