Applied Data Analysis (CS401)

Lecture 3
Visualizing data
5 Oct 2022

EPFL

Robert West

dlab

Announcements

- You must find 3 team mates and register by this Fri 7 Oct 23:59
  - Every student must register individually
  - Still looking for a team or for team members? Use Ed (see post)!
- Project milestone 1 due Fri 14 Oct 23:59
- Homework 1 released Fri 14 Oct (due two weeks later)
- Friday's lab session:
  - From now on: Lab session materials to be released on Wed before lab
  - 13:15 - 13:30 (on Moodle): Quiz 2 (the first one that counts)
  - More on data visualization; working with data from the Web
  - Project milestone P1 office hours

Feedback

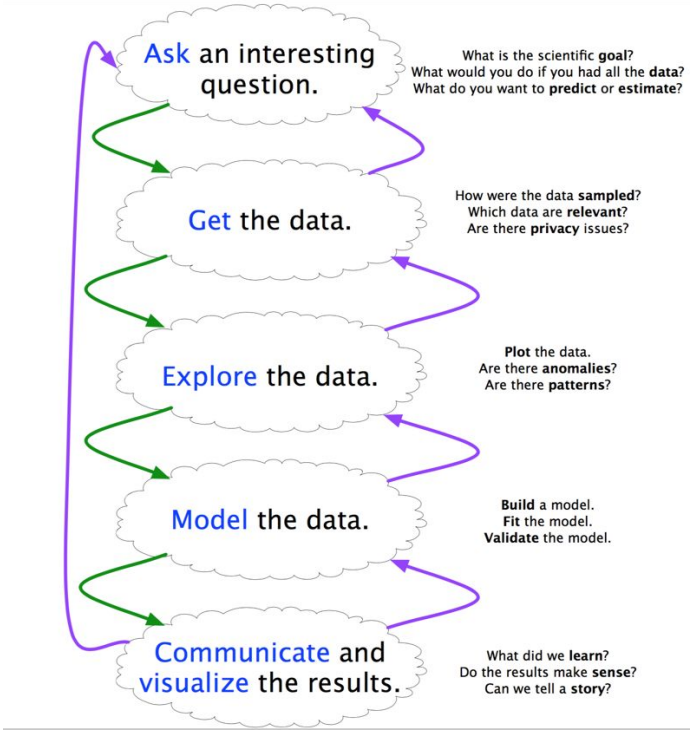
Give us feedback on this lecture here:
https://go.epfl.ch/ada2022-lec3-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- What's your favorite color, baby?
- …

Uses for data visualization

Support reasoning about information **(analysis)**

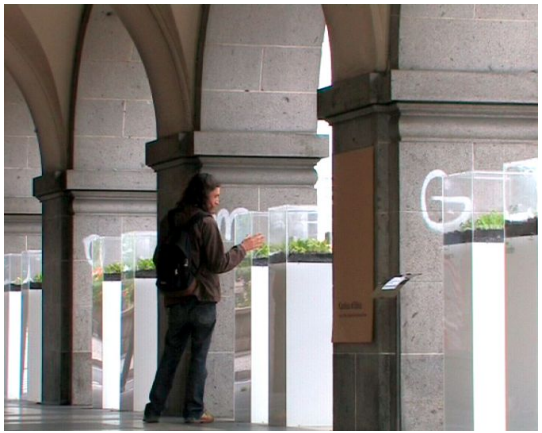
- Finding relationships
- Discover structure
- Quantifying values and influences
- Should be part of the data analysis cycle ➡



Inform and persuade others **(communication)**

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

An unconventional example



"Garden of Eden": 8 lettuces, each of which is enclosed in its own airtight plexiglas box and represents a major city. The concentration of ozone in each box is controlled in real-time to reflect the current pollution level in the city.

Static viz

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

Interactive viz

More and more common when delivering the results (and also during exploration). New frameworks are the key enabler.

Want to learn more?

Dedicated course:

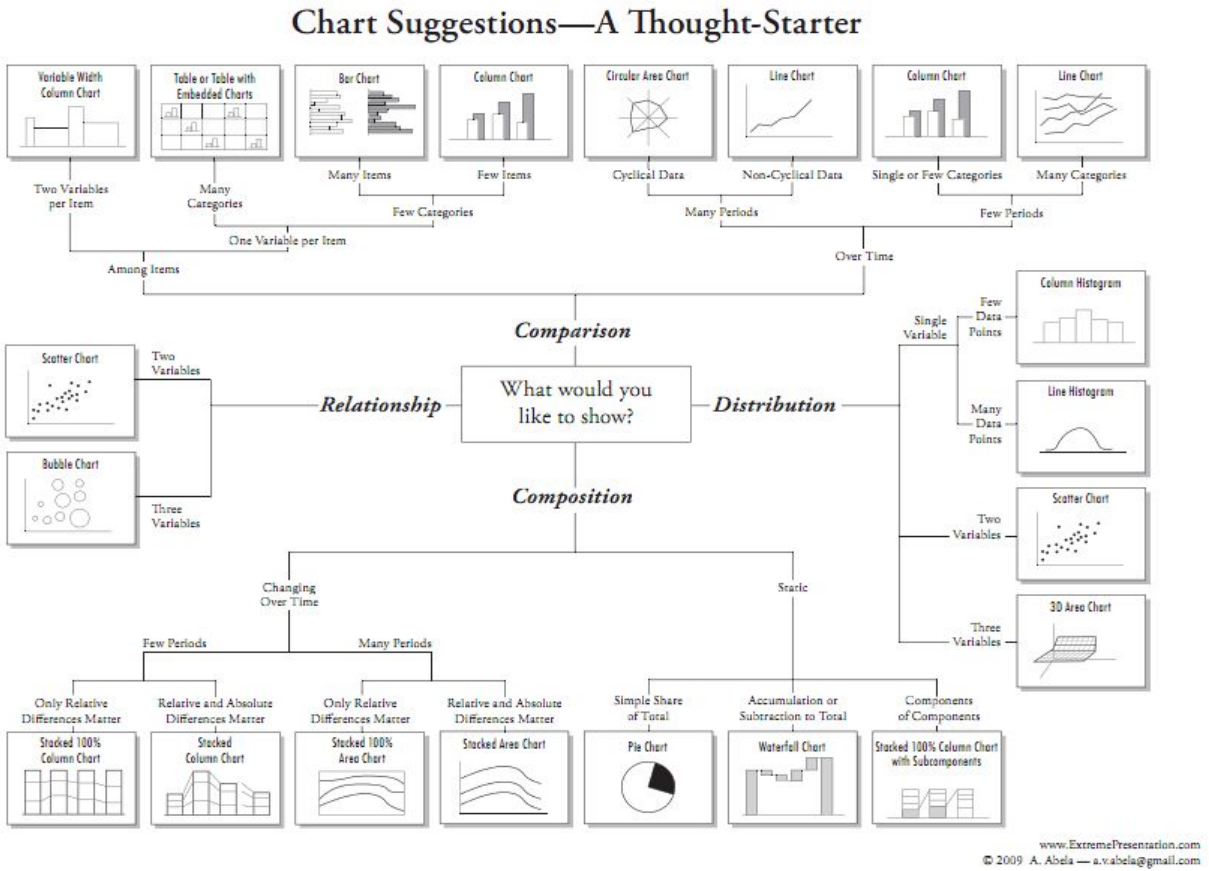
COM-480: Data Visualization

Today’s lecture

- Part 1: Navigating the chart landscape
- Part 2: Principles and best practices
- Part 3: A (small) selection of use cases for data visualization

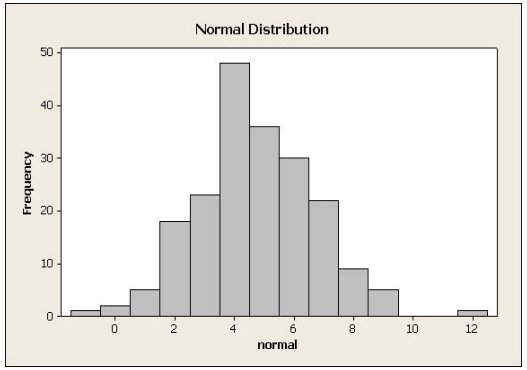
Part 1
Navigating the chart landscape

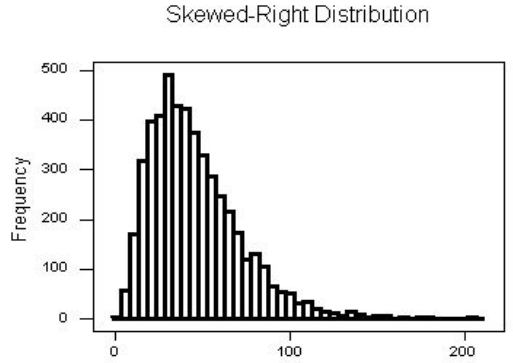
Chart selection

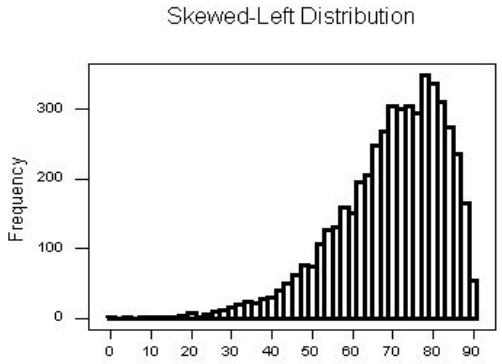


One variable: histograms

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

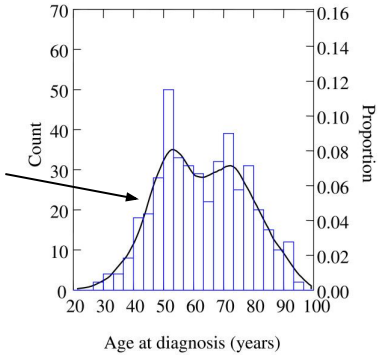




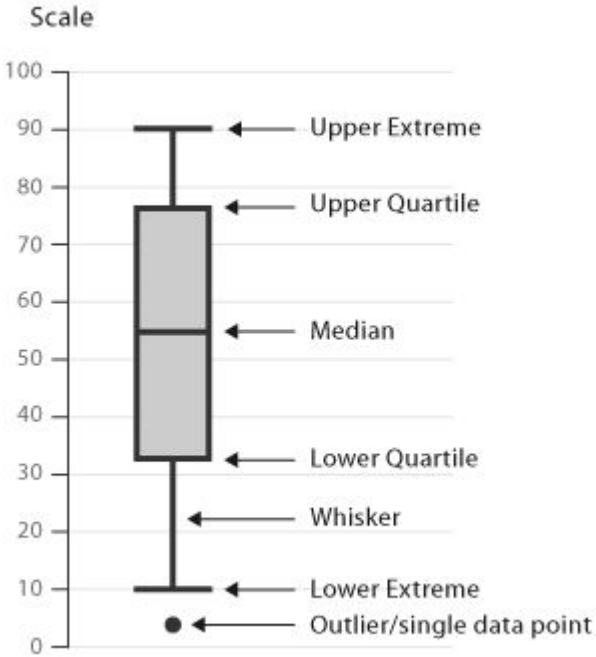


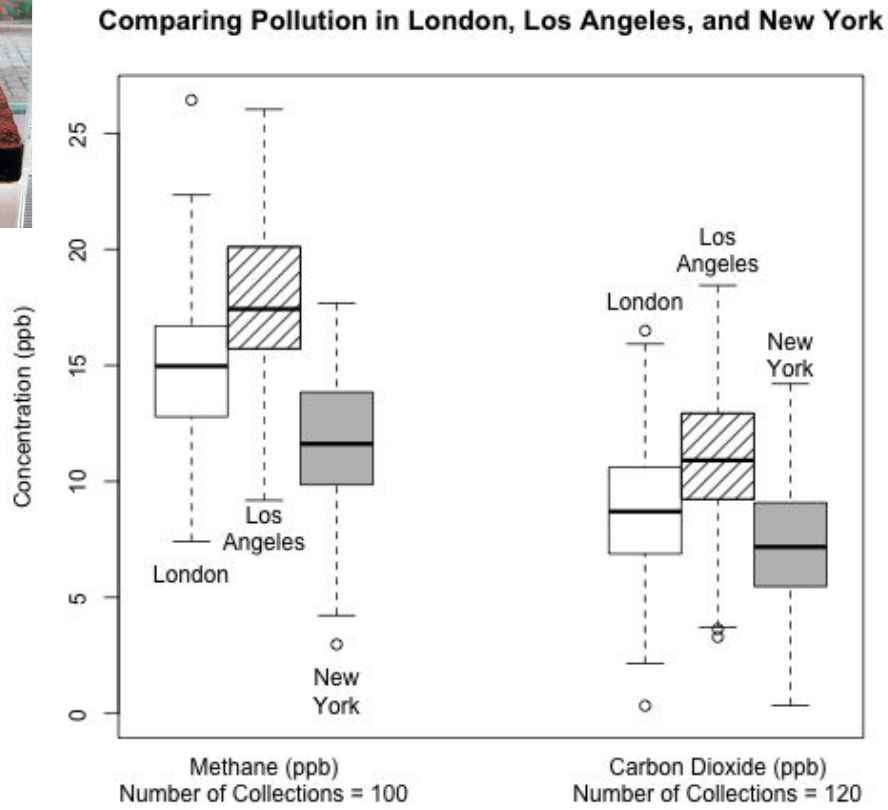
Easy to recognize skewed distributions!

Smoothed histogram (a.k.a. kernel density estimate)



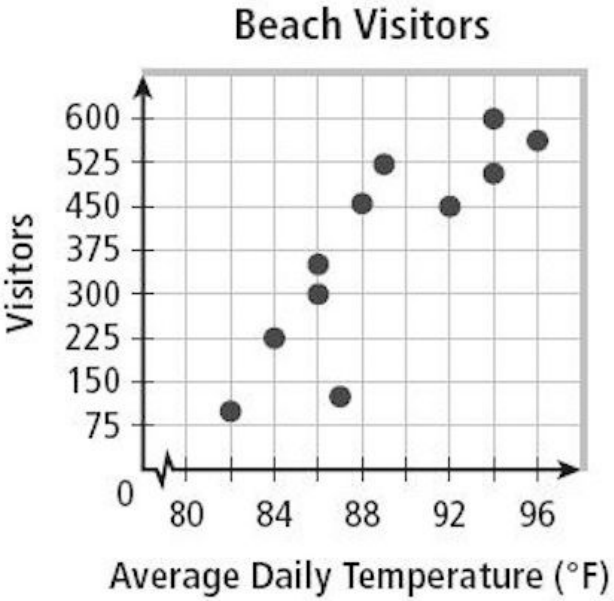
One variable: box plots





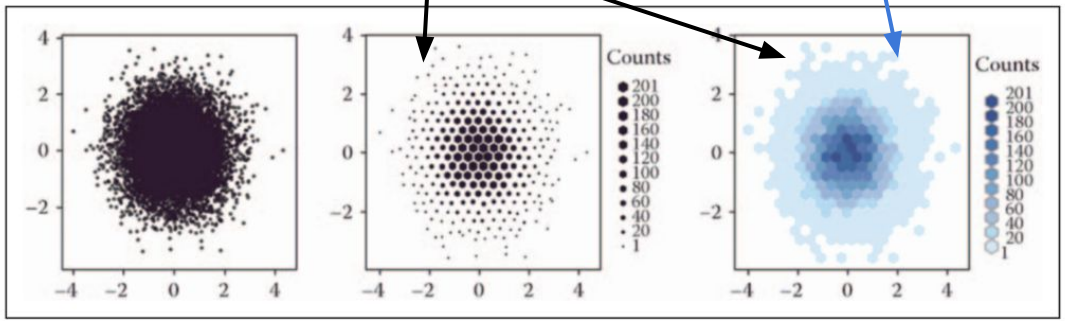
Two variables: scatter plots

Scatter plots quickly expose the relationships between two variables



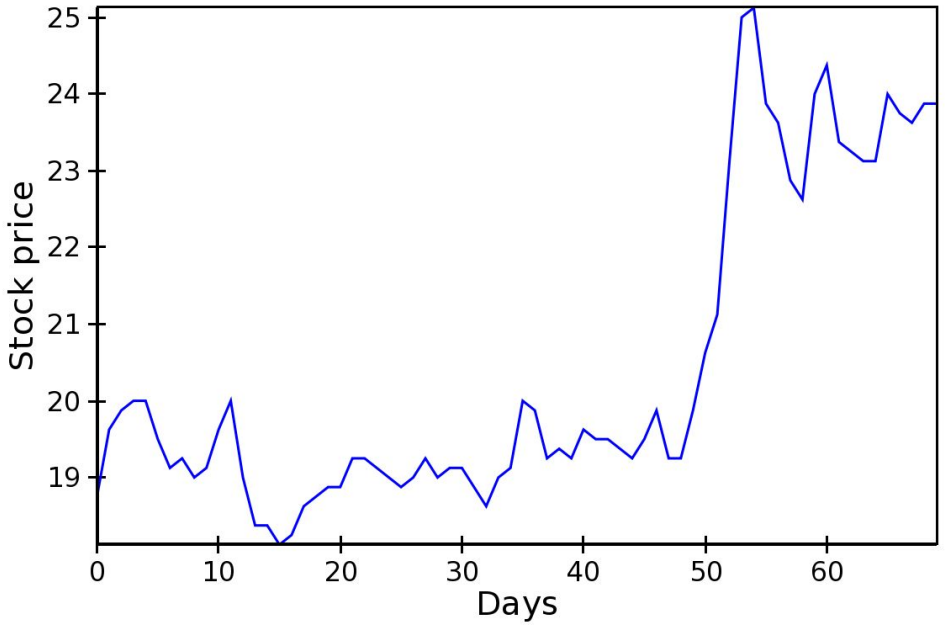
2D histograms

a.k.a. heatmap

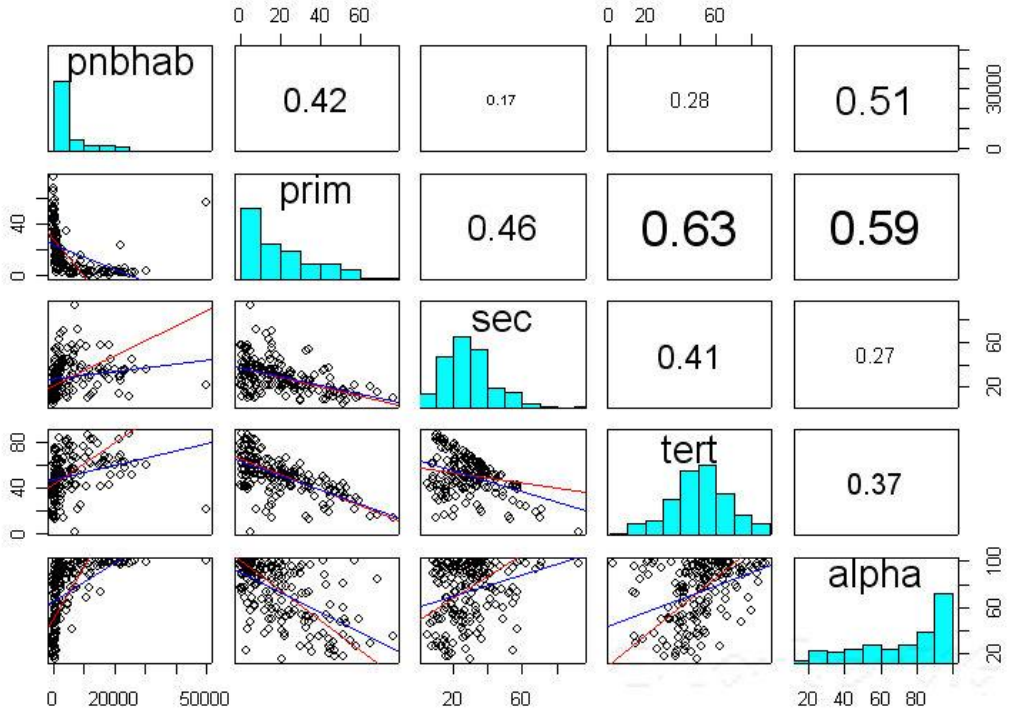


Two variables: line plots

If relationship is functional (for instance, after binning and aggregating)



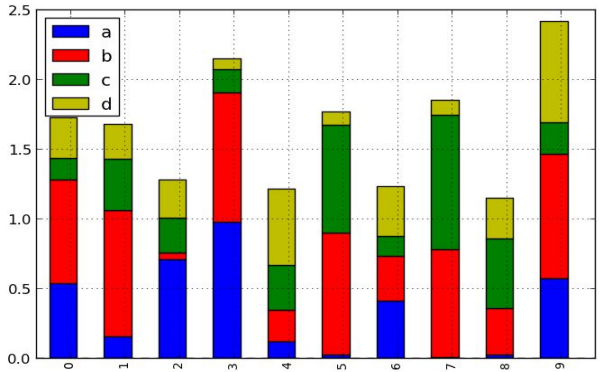
> 2 variables: scatter plot matrix



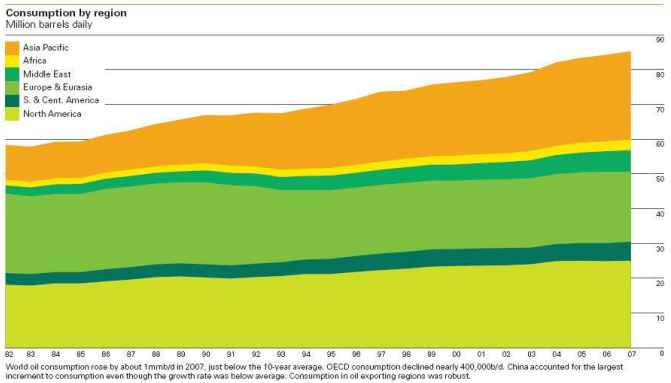
> 2 variables: stacked plots

Here: 3 variables: stack index, height, color

Stack variable and color variables categorical, height variable continuous:

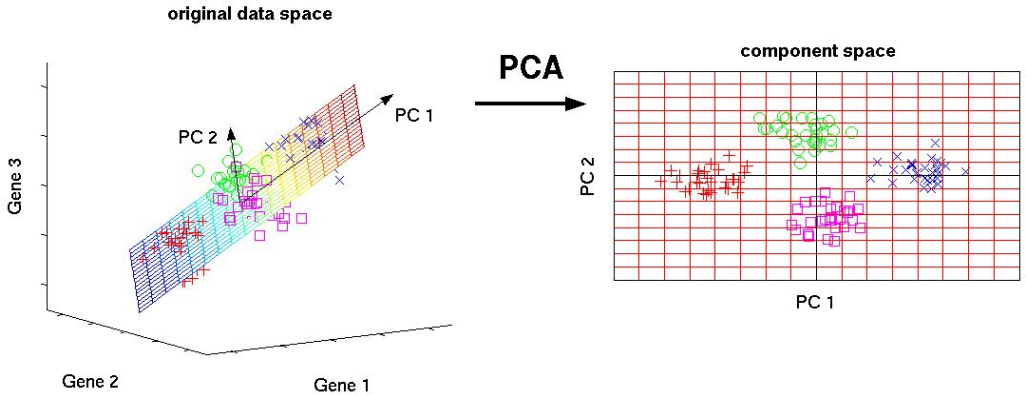


Color variable categorical, stack and height variables continuous:

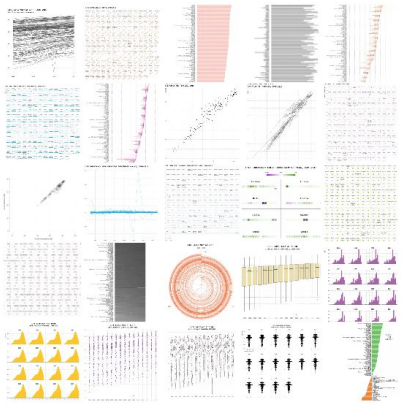


Dimensionality reduction

- For 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

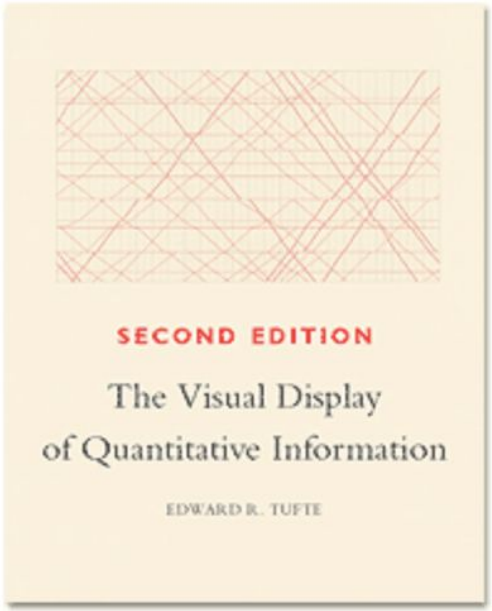


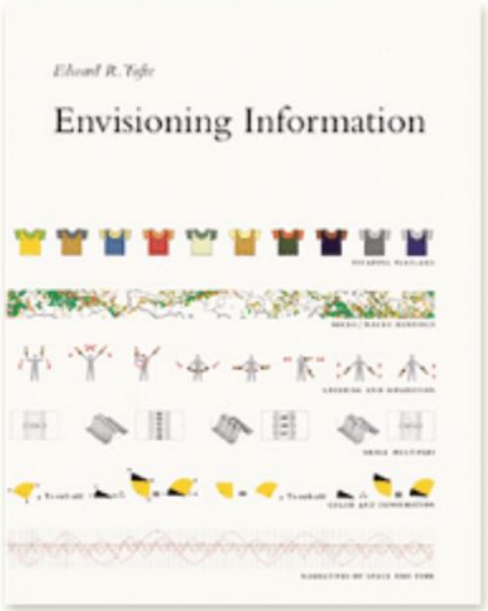
http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways

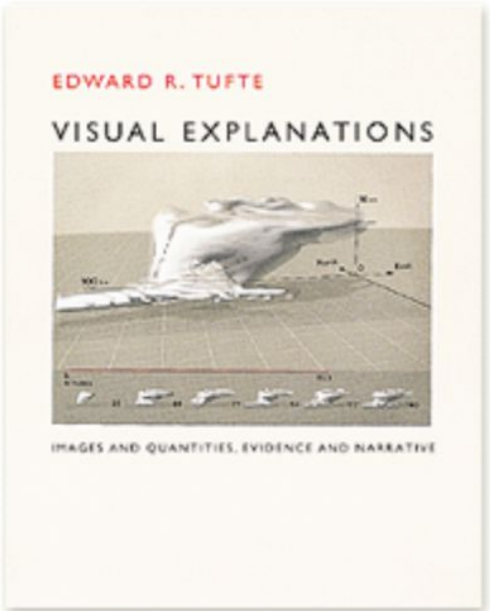
"You must help the data focus and get to the point. Otherwise, it just ends up rambling about what it had for breakfast this morning and how the coffee wasn't hot enough."

Part 2
Principles and best practices

Instructive coffee table books by Edward







[demo] [another great coffee table book]

Perception of magnitudes

(128, 128, 128)

(144, 144, 144)

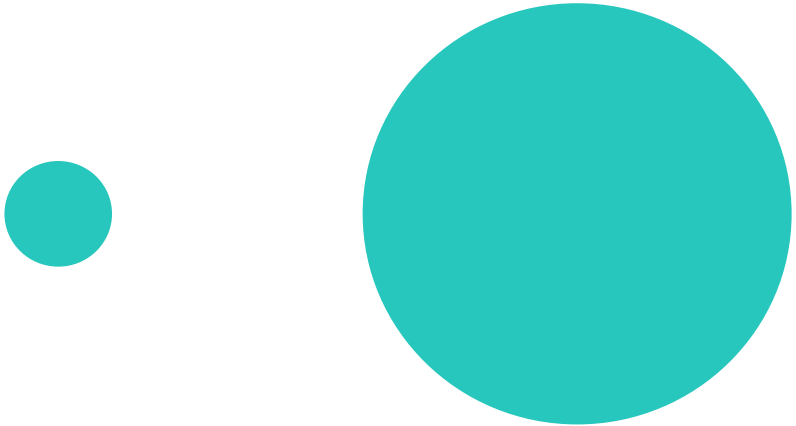
Which is brighter?

Just noticeable difference (JND)

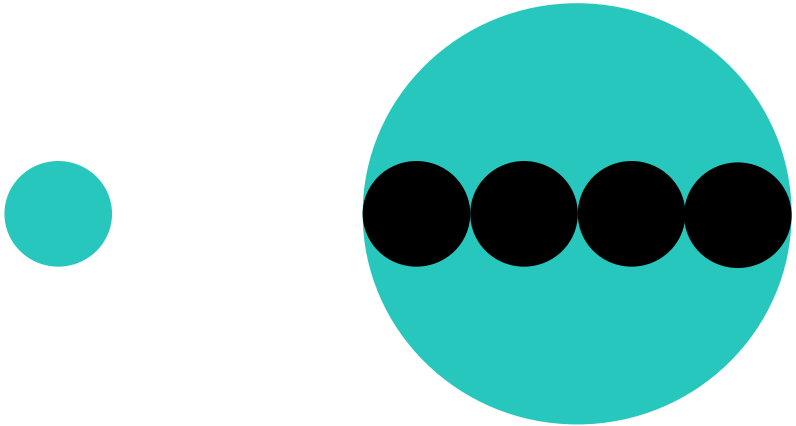
- Weber's law:

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

- I: intensity; ΔI: increase from I to notice a difference; k: constant
- Required increase ΔI depends on original intensity I
- Most continuous variations in stimuli are perceived in discrete (multiplicative) steps



Compare area of circles



Compare area of circles

Perception of magnitudes

Most accurate

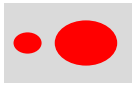
Least accurate

Position

Length

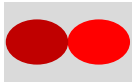
Slope

Angle



Area

Volume



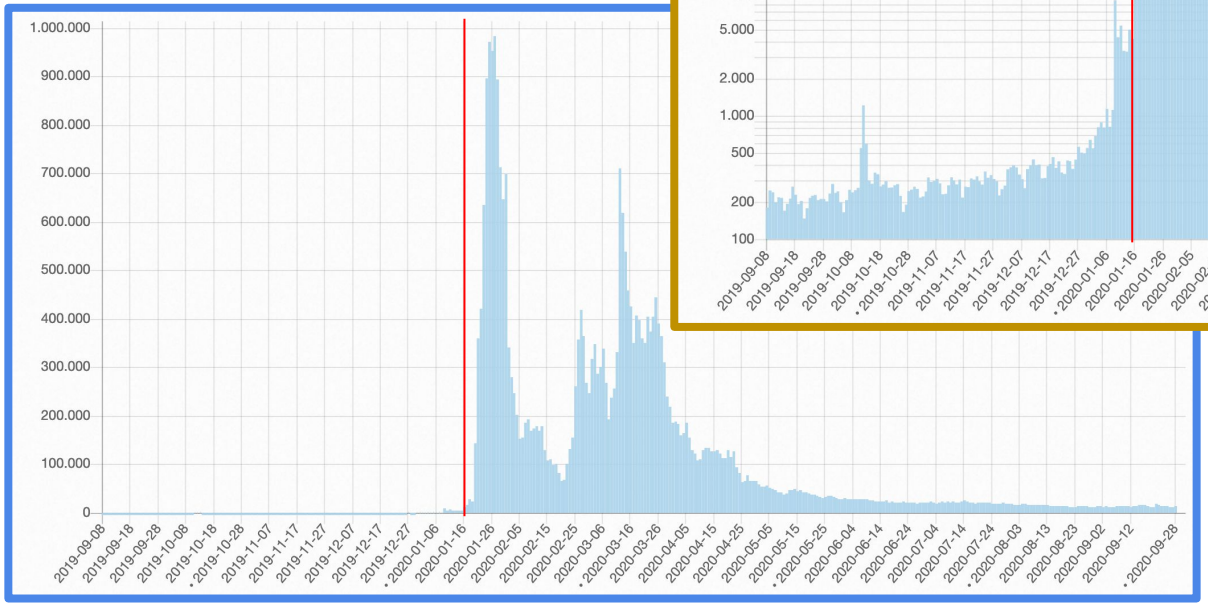
Color hue-saturation-density

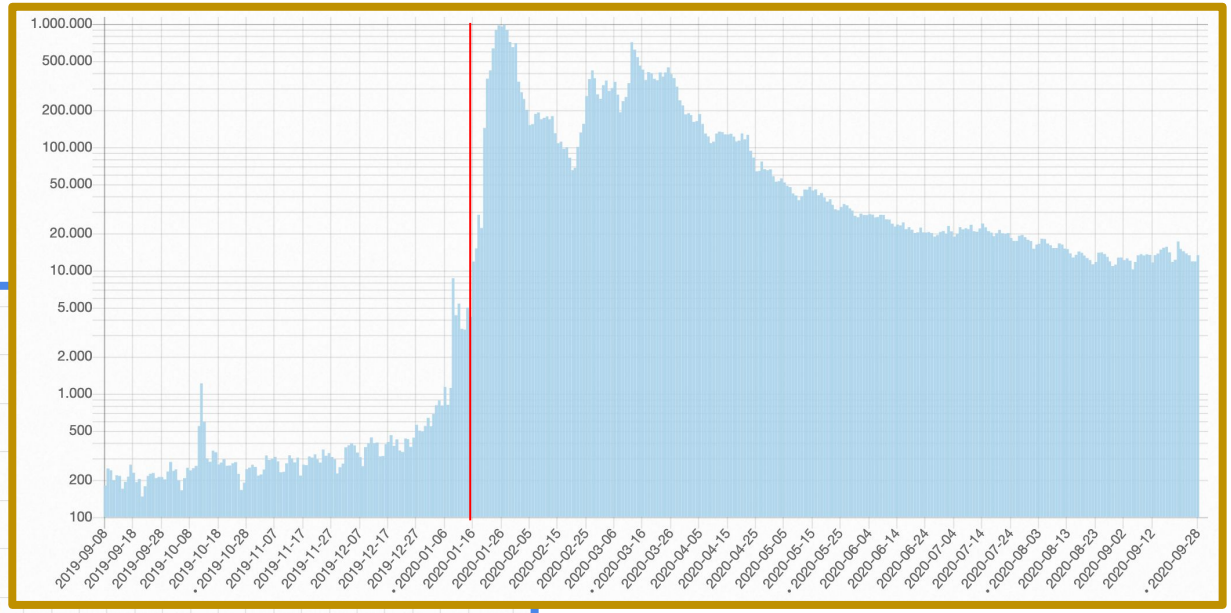
Cleveland & McGill (1984)
Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

Choose your axes wisely!

Time series of pageviews of Wikipedia article about “Coronavirus”

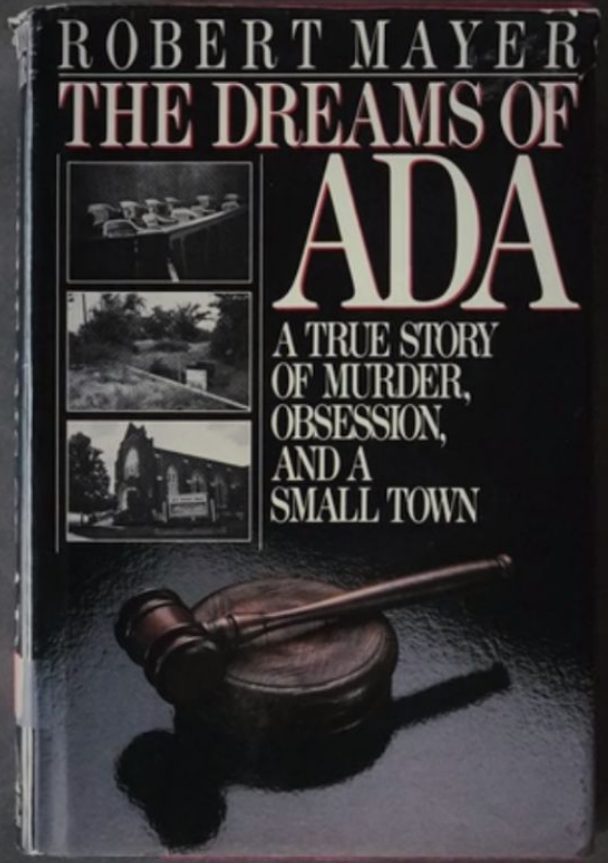
(linear y-axis)



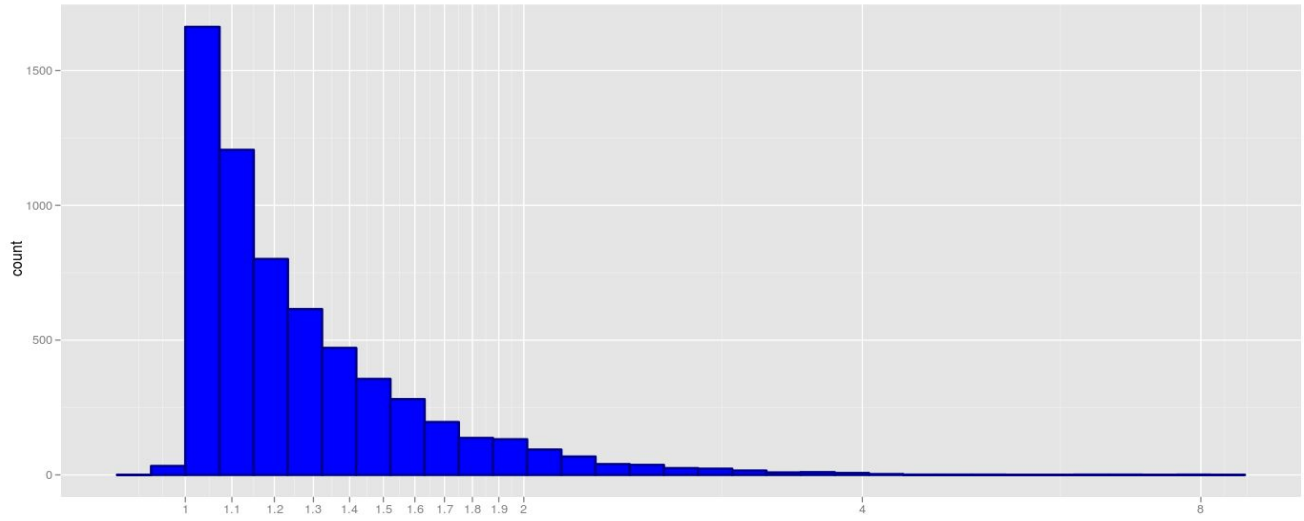


(logarithmic y-axis)

[link]

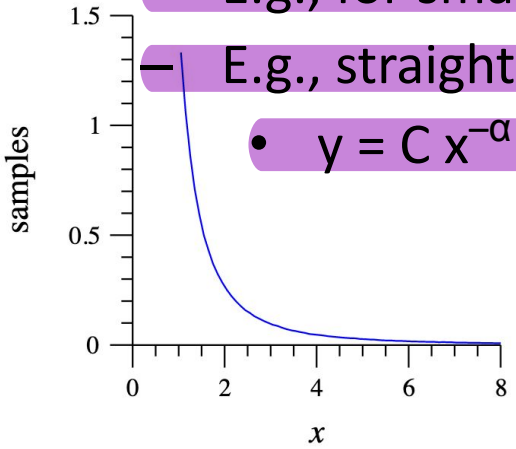

ROBERT MAYER
THE DREAMS OF ADA
A TRUE STORY OF MURDER, OBSESSION, AND A SMALL TOWN

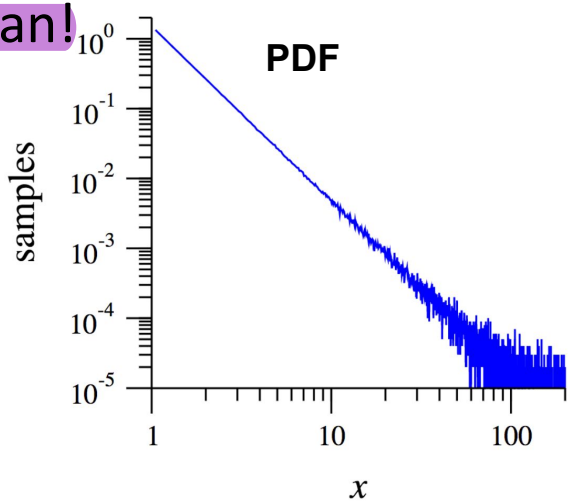
Choose your axes wisely: Visualizing heavy-tailed distributions



Heavy-tailed data: power laws

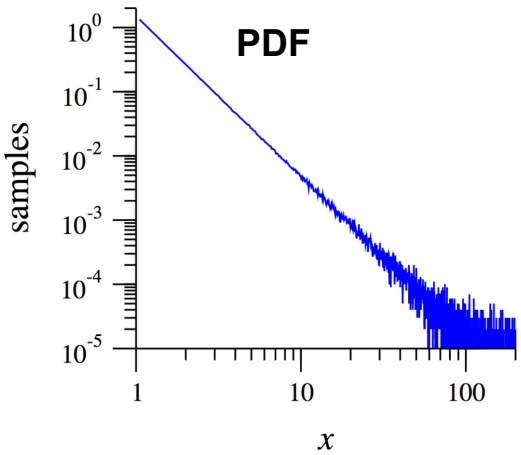
- $p(x) = Cx^{-\alpha}$,
  - Very large values are rare, “but not very rare”
  - Body size vs. city size
  - Many natural phenomena are power laws (e.g., # of friends)
  - For dealing with them, need to know some tricks
  - E.g., for small α, mean & var = ∞ ⟶ use median!
  - E.g., straight line on log-log axes:
    - $y = C\,x^{-\alpha} \leftrightarrow \underline{\log(y)} = \underline{\log(C)} - \alpha\,\underline{\log(x)}$

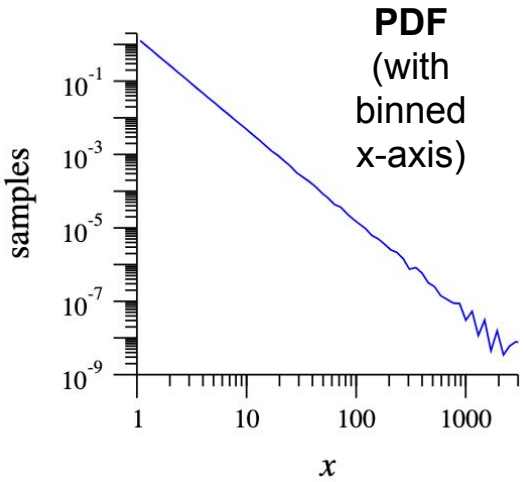




Heavy-tailed data: power laws

- Complementary cumulative distribution function (CCDF):
  P(x) := Pr{X >= x}





POLLING TIME

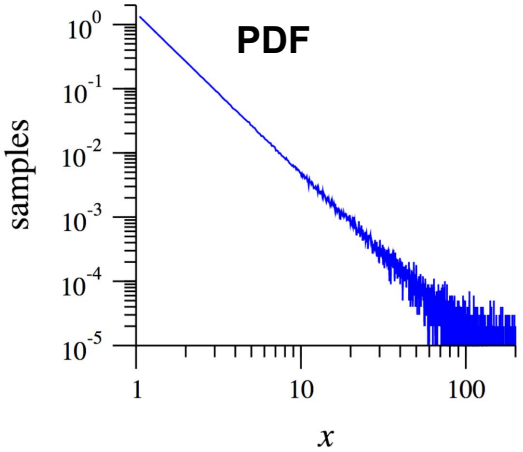
- "What shape does the CCDF of a power law have when plotted on log-log axes?"
- Scan QR code or go to https://web.speakup.info/room/join/66626

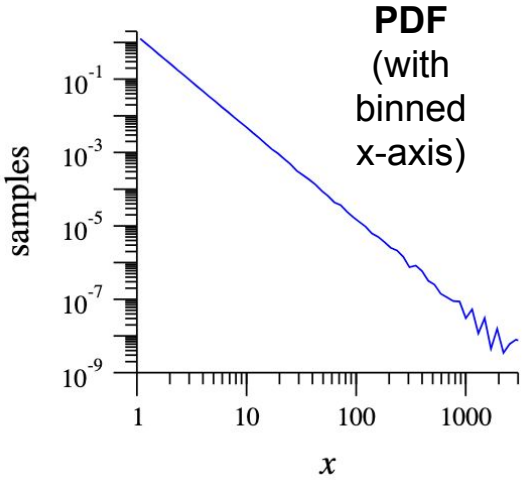
Heavy-tailed data: power laws

- Complementary cumulative distribution function (CCDF):
  P(x) := Pr{X >= x}
- CCDF of power law is also a power law (with exponent α – 1)

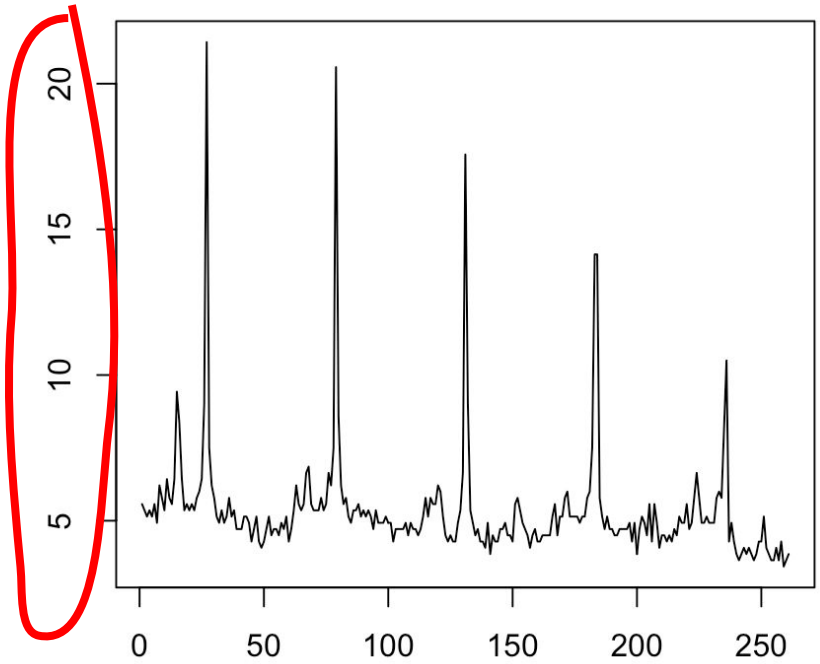
$$P(x) = C \int_x^\infty x'^{-\alpha}\, \mathrm{d}x' = \frac{C}{\alpha - 1} x^{-(\alpha-1)}.$$

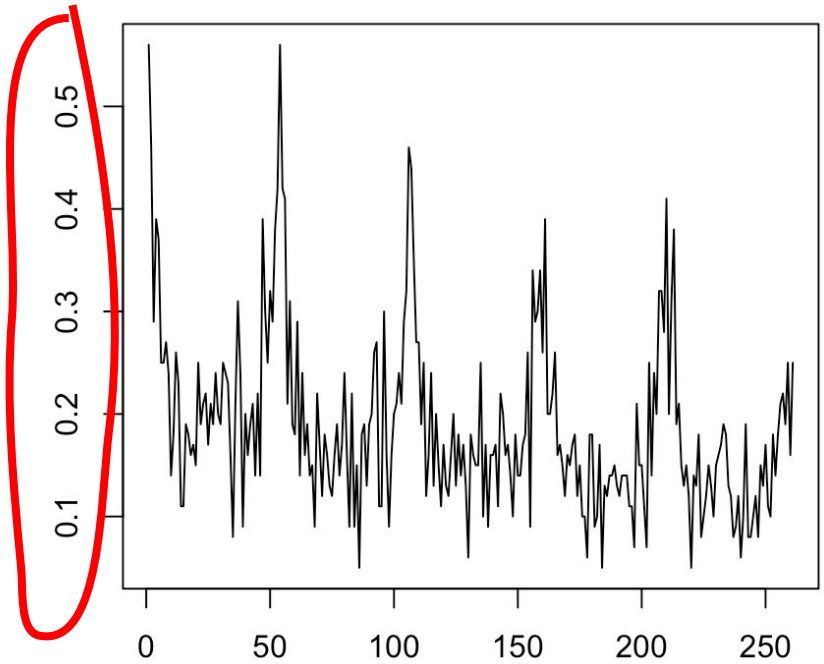
- CCDF plot is monotonically decreasing





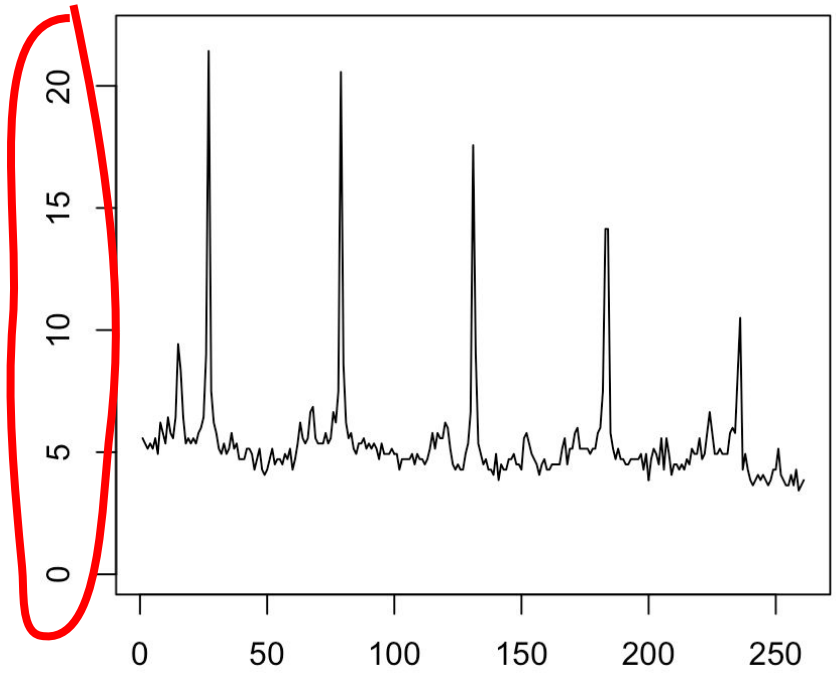
Answer fast: which time series has a higher mean value?

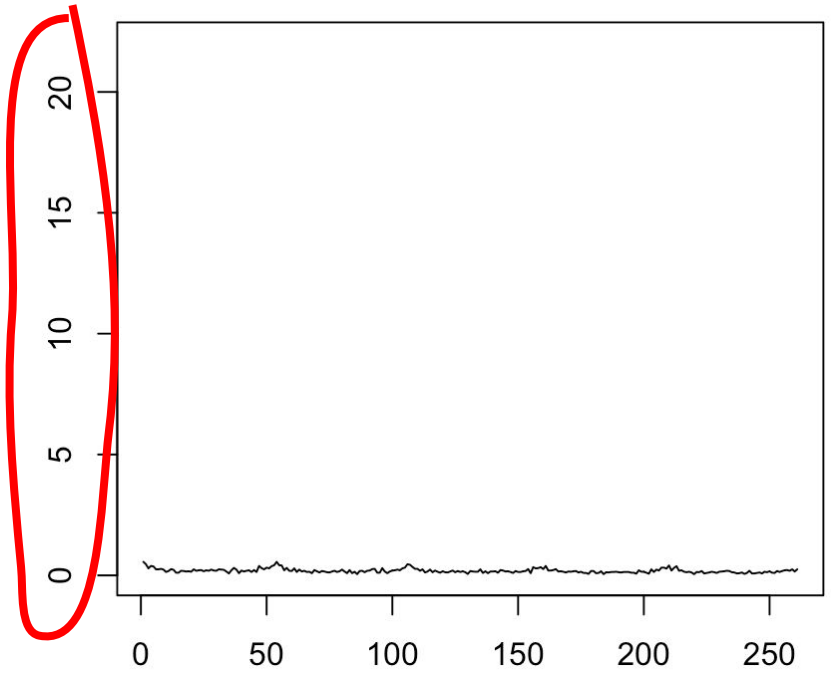




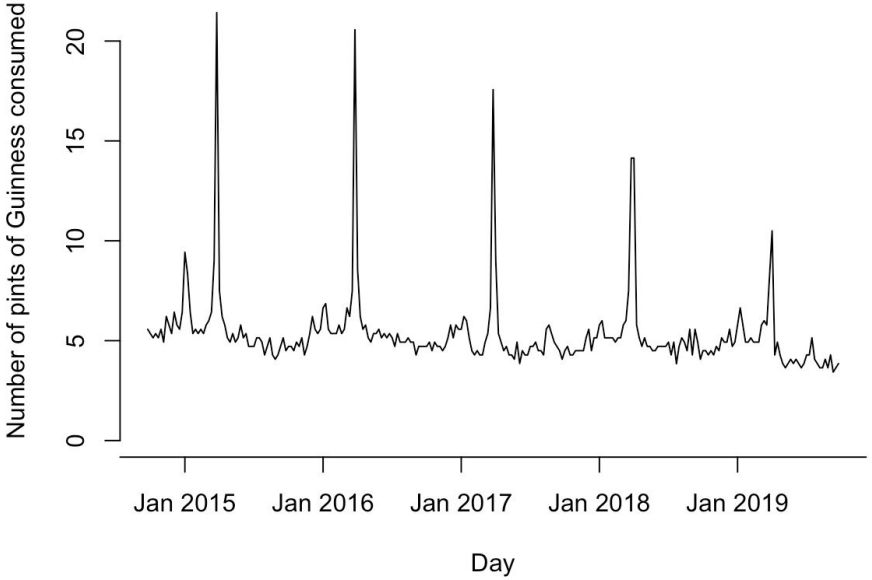
Use consistent axes!

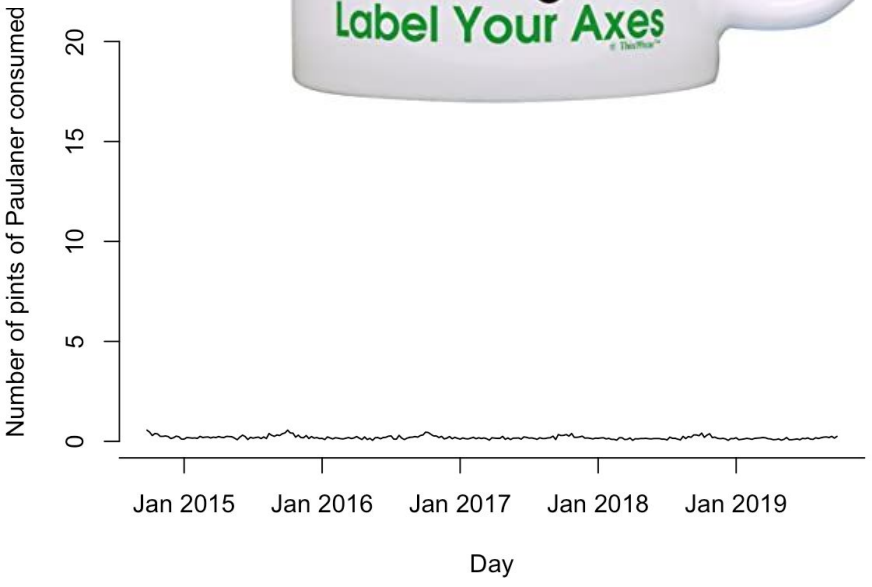
Answer fast: which time series has a higher mean value?

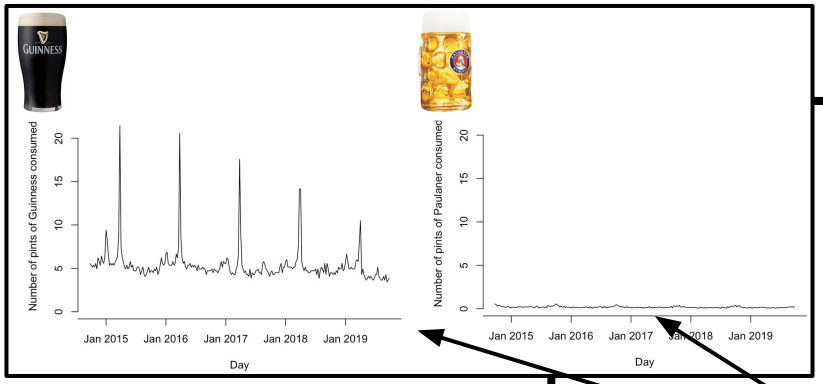




Label your axes!







THINK FOR A MINUTE:

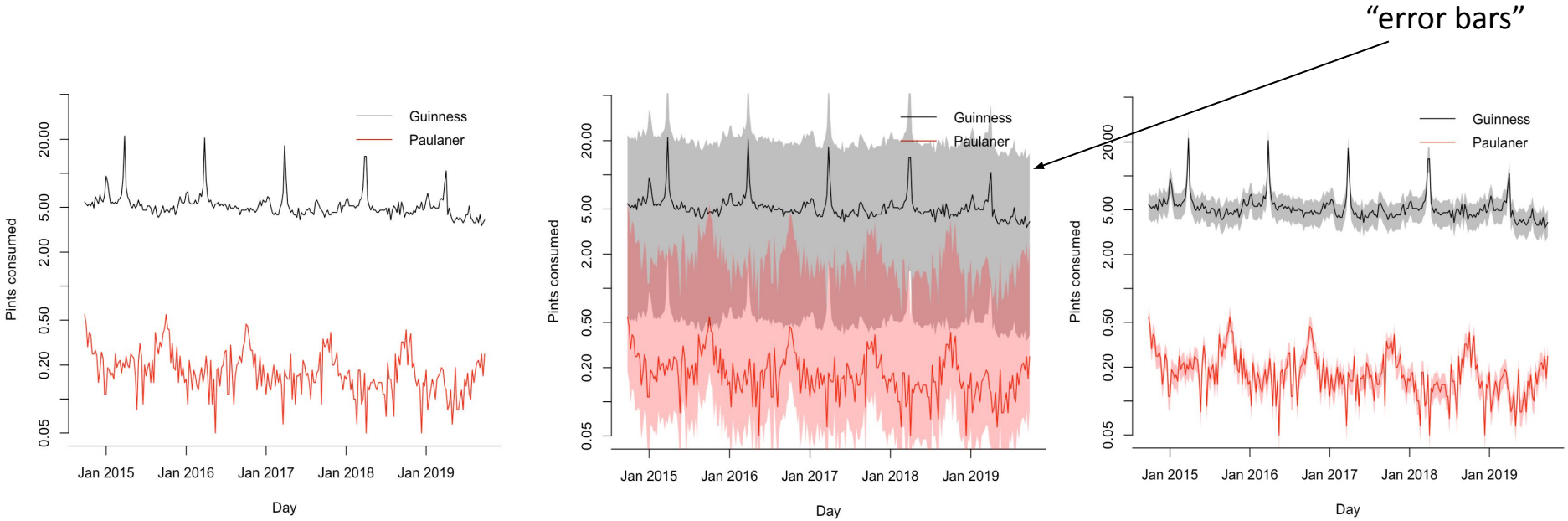
How could we show details of both time series without using different y-axes?

(Feel free to discuss with your neighbor.)

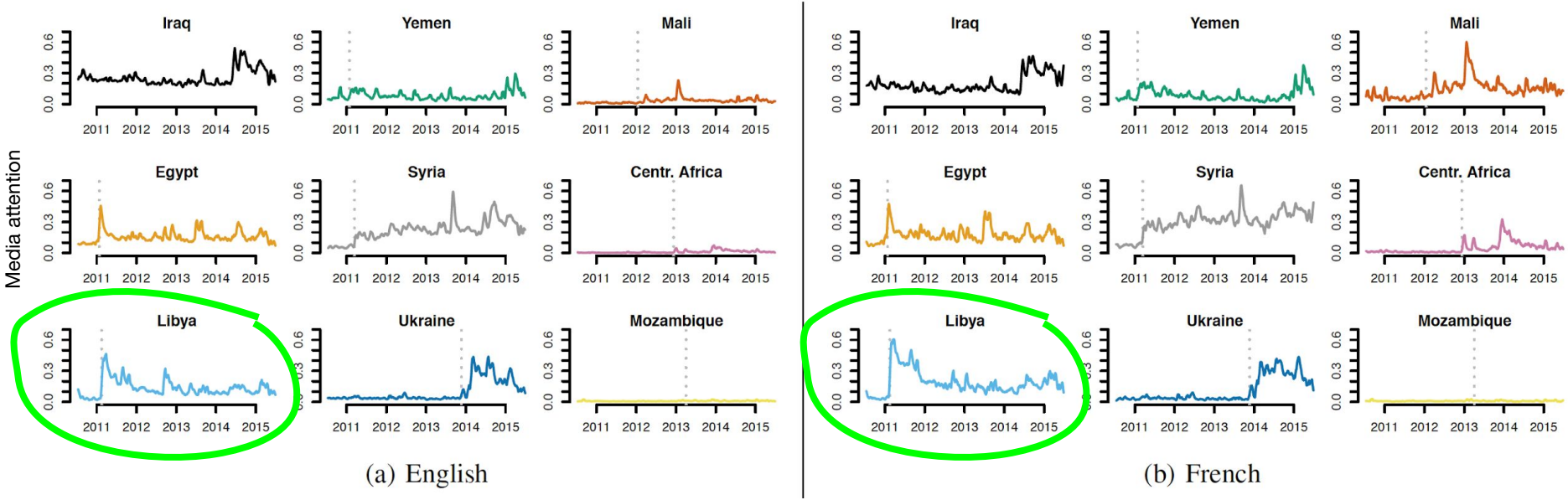
Guinness
Paulaner
Pints consumed
0.05
0.20
0.50
2.00
5.00
20.00
50.00
Jan 2015
Jan 2016
Jan 2017
Jan 2018
Jan 2019
Day

Show data uncertainty!

Which beer is more popular, Guinness or Paulaner?



Consider using small multiples!



Use colors consistently!

[link]

Use colors wisely!

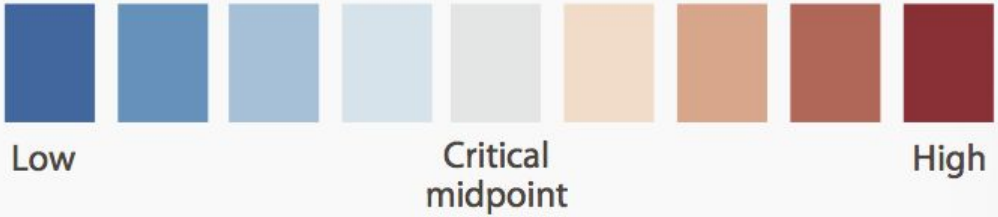
Choose colors based on the information you want to convey

Sequential

Colors can be ordered from low to high

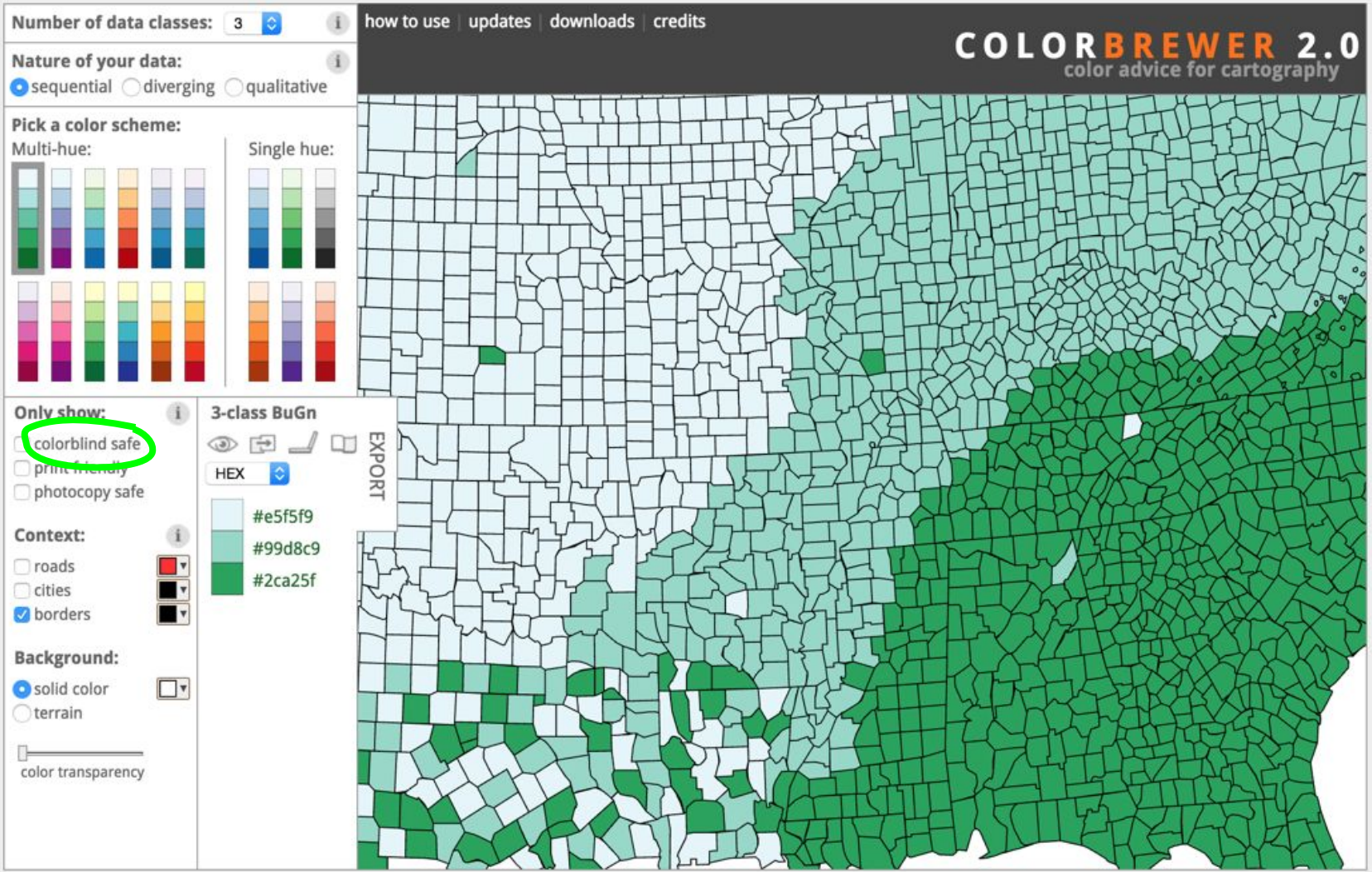
Diverging

Two sequential schemes extended out from a critical midpoint value



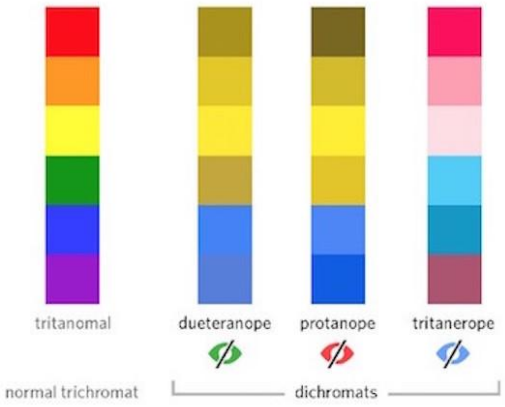
Categorical

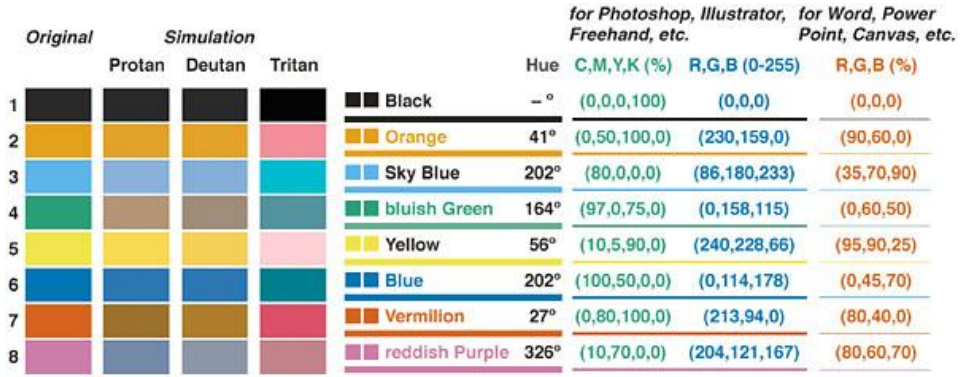
Lots of contrast between each adjacent color


Number of data classes: 3
Nature of your data:
sequential diverging qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
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

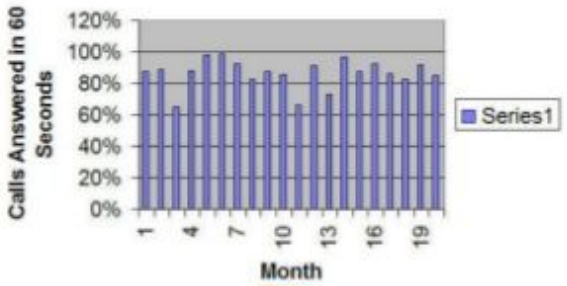
Use colorblind-safe palettes!

- Remember: 10% of males have some form of colorblindness

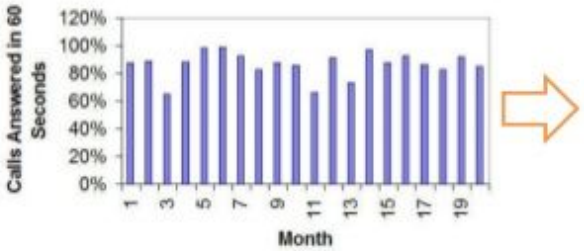


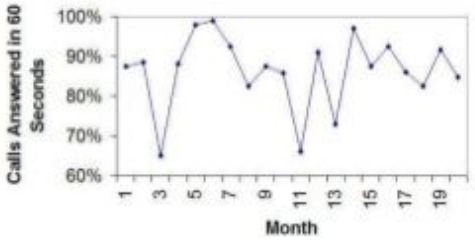


Use data ink wisely! Avoid chart junk!

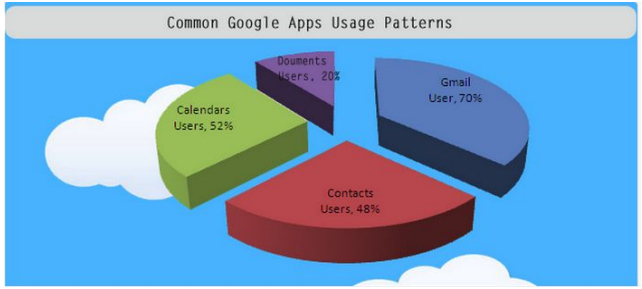


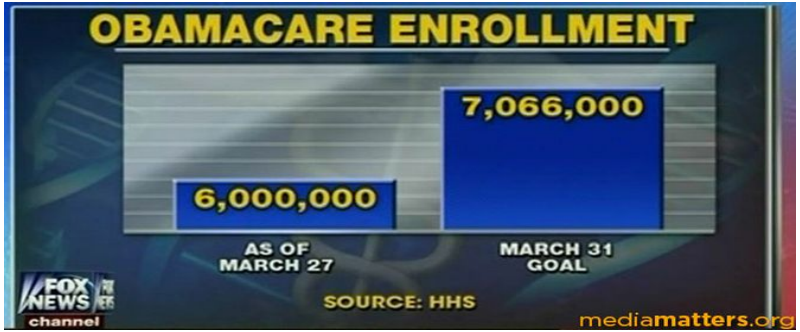
Graphical excellence gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
-Edward Tufte

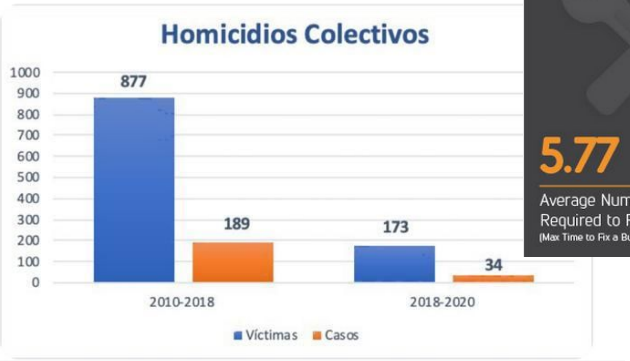


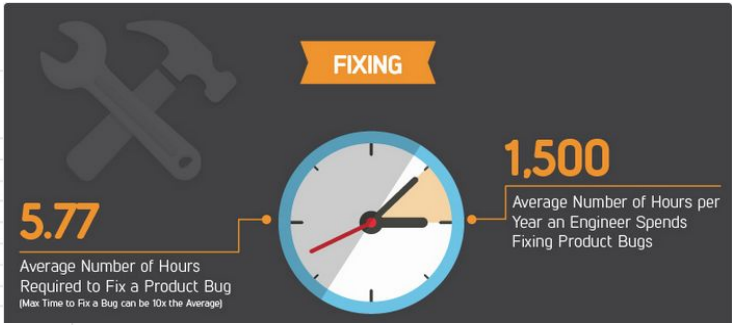


Toilet exercise: Which principles and best practices do these graphics violate?





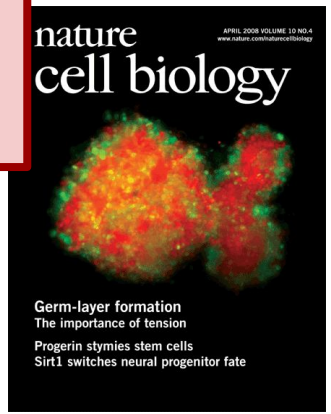


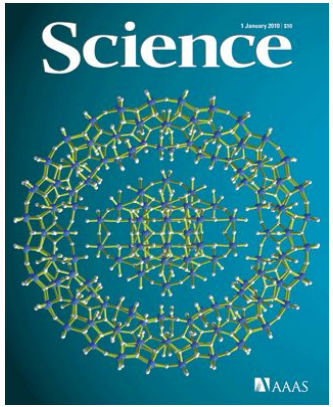


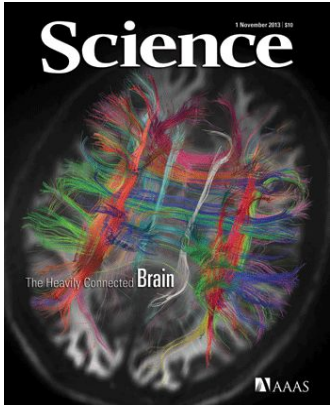
Courtesy of [viz.wtf](viz.wtf)

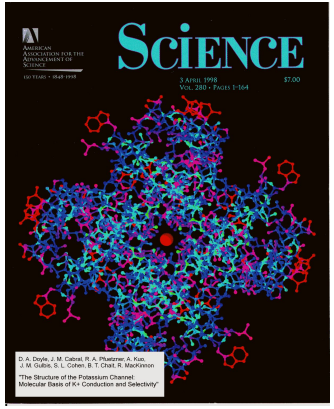
Part 3
A (small) selection of use cases for data visualization

Use case: Presenting scientific results





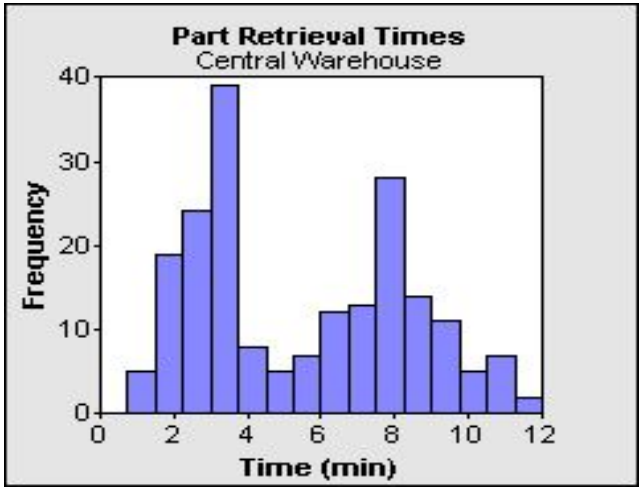




Use case:
Data wrangling

Multimodal data

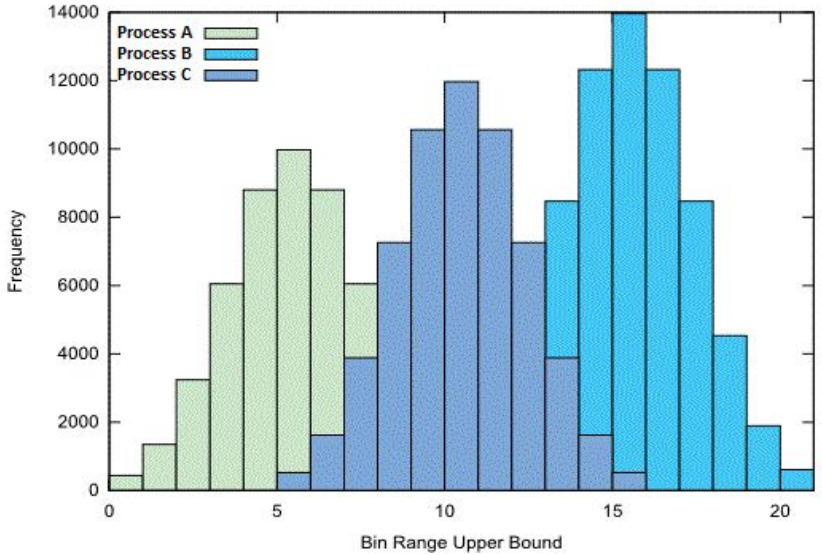
- Two or more distinct peaks in a histogram often **suggest 2 or more distinct populations** of samples.
- But don’t guess! Explore further by using, e.g., color and a histogram of multiple populations (p.t.o.).



Use case:
Data wrangling

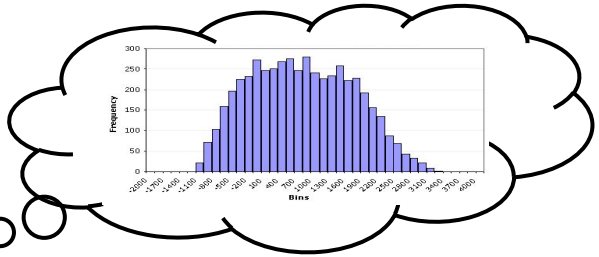
Multimodal data

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

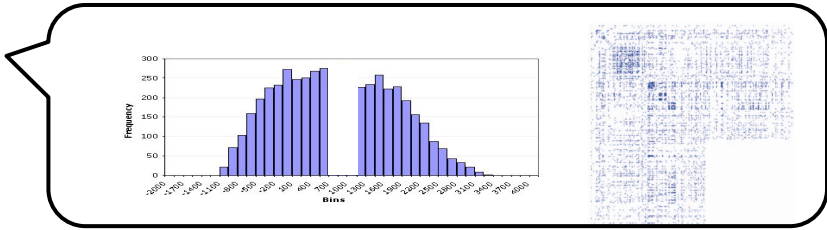


Use case:
Data wrangling

Weird data



- Maintain a **theory of what the data should look like**.
- Some data is **very hard to explain**.
- Never just blink it away!
- First, assume a bug. Try to fix it.
- If not a bug: you might have made an interesting discovery!
- Some of science's most important findings were made by not ignoring weird data, but dwelling on it!



[link]

Use case: 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

And 2014 “the year in interactive storytelling”

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

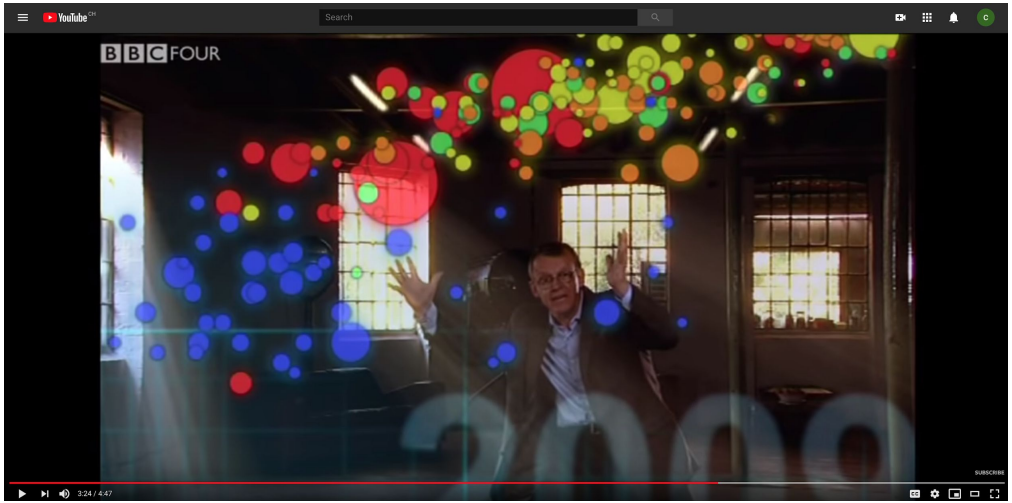
NY Times graphics are a great source of
best practices in viz (except for when they’re not…)

Use case: Educating the public

Hans Rosling:

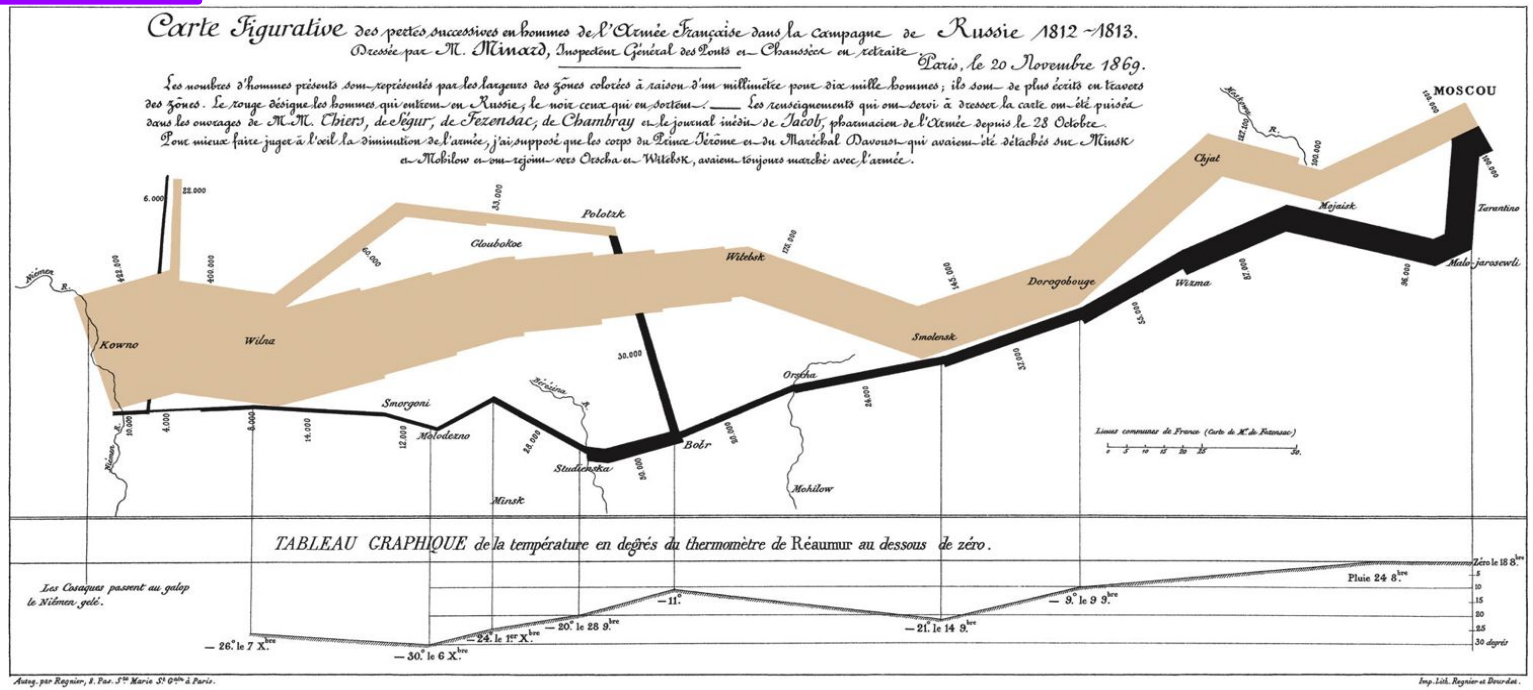
200 countries, 200 years, 4 minutes

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



Use case: Give new perspectives

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

Tools

(remaining slides for your personal perusal)

Interactive toolkits: D3

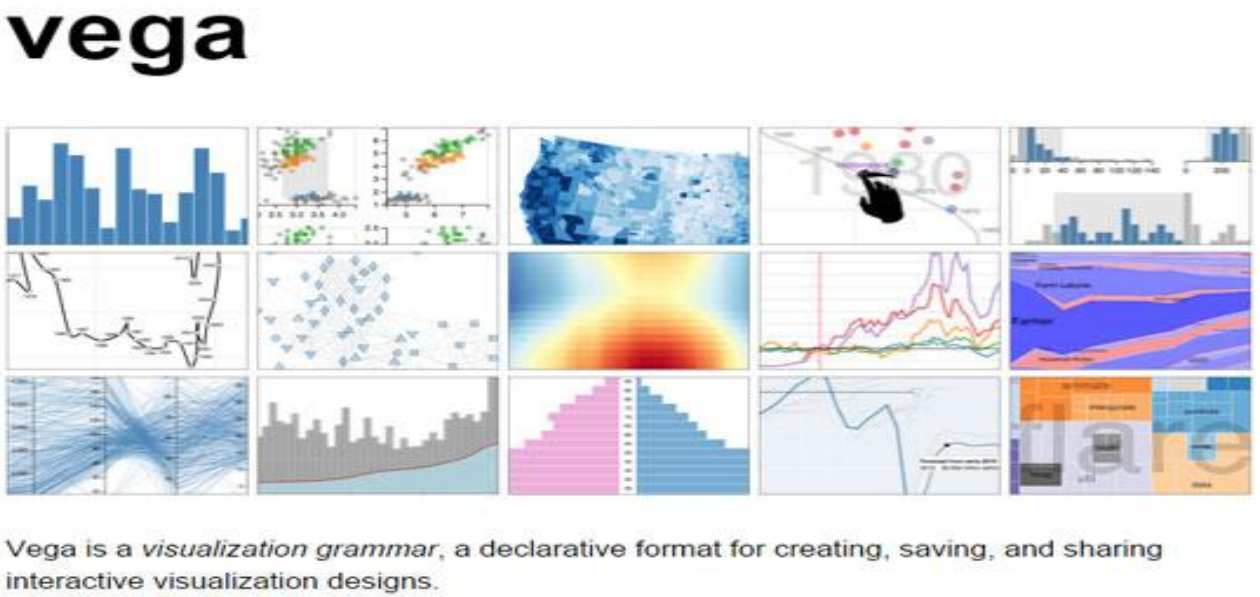
Without doubt, the most widely used interactive visualization framework is **D3**.

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

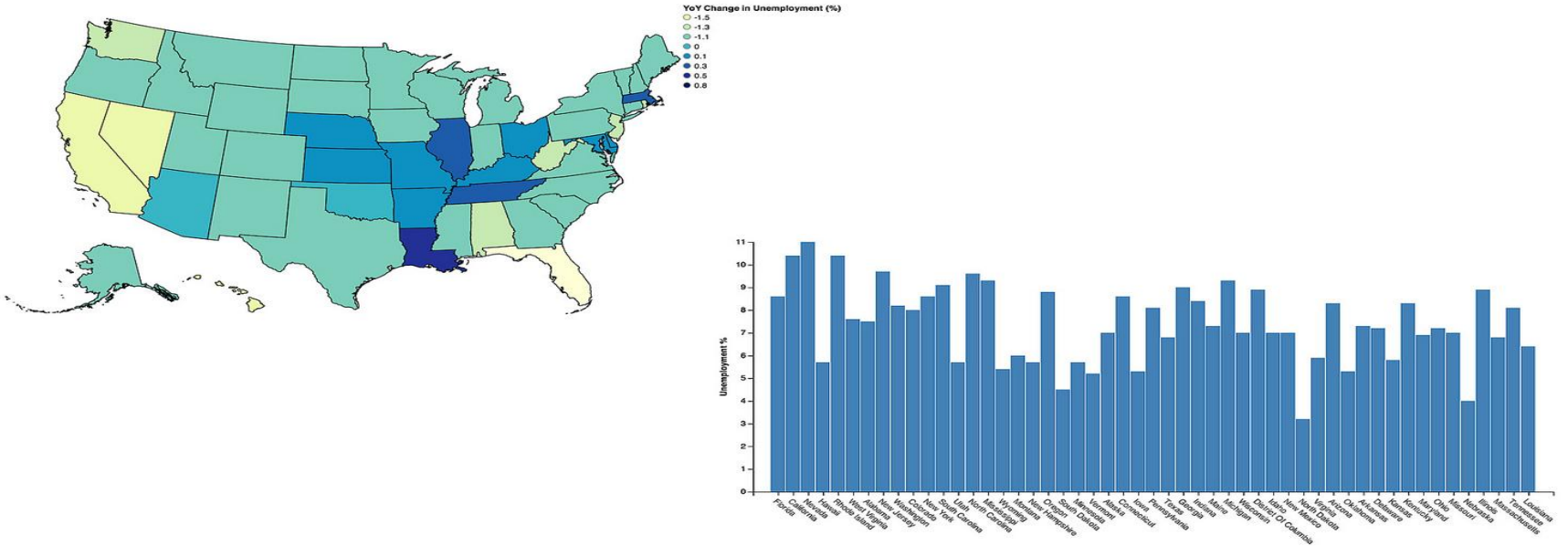
It specifies graphics in JSON format.



Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

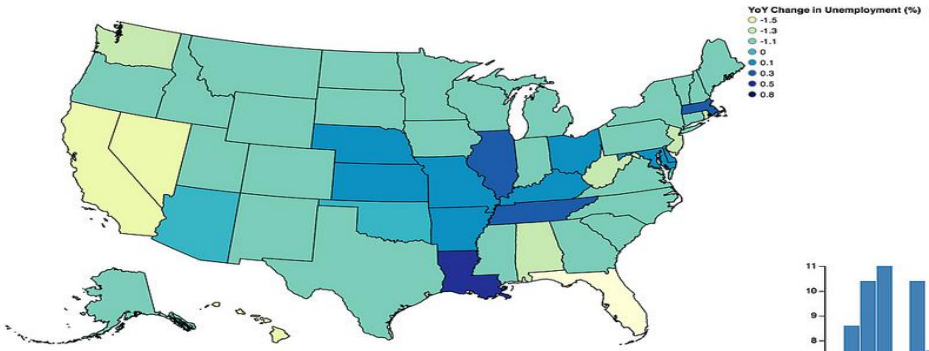
Trivia question: why is it called Vincent? Hint: Vincent+Vega= ?

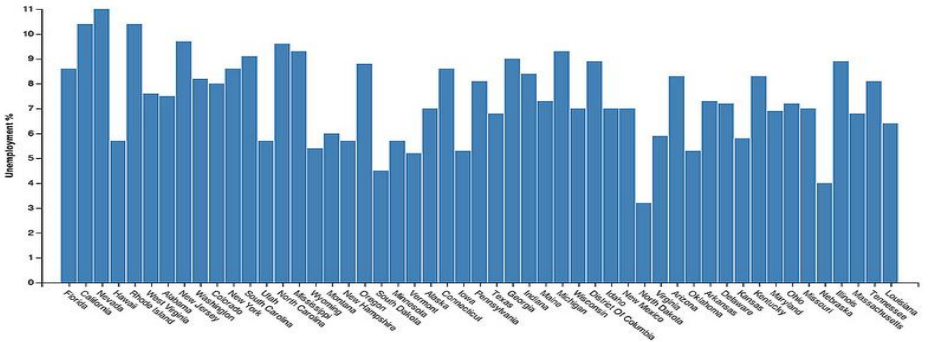


Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

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



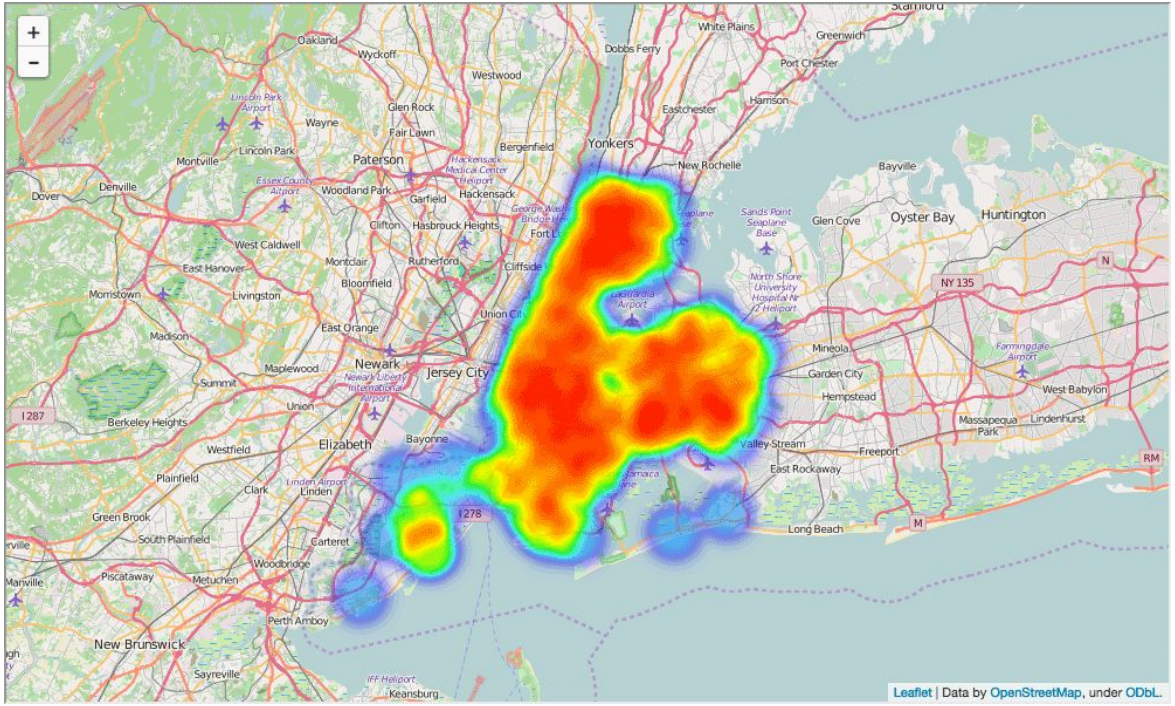


Bokeh: another interactive viz library

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!



Feedback

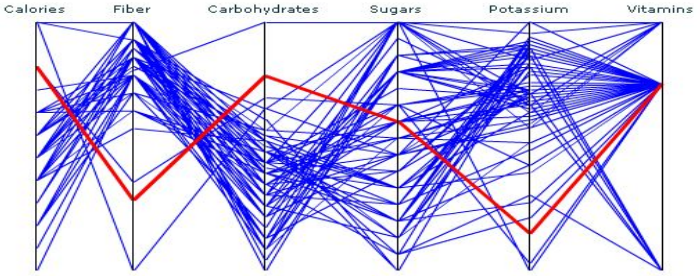
Give us feedback on this lecture here:
https://go.epfl.ch/ada2022-lec3-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- What's your favorite color, baby?
- …

> 2 variables: parallel-coord. plots

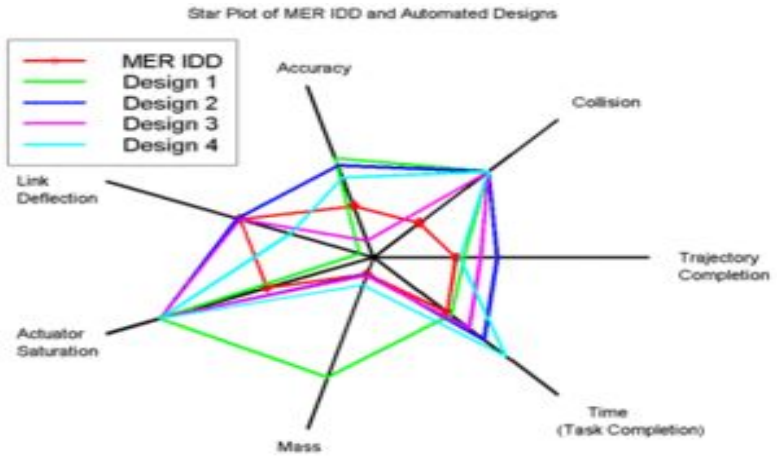
Color, x, y

Color variable is categorical, others arbitrary



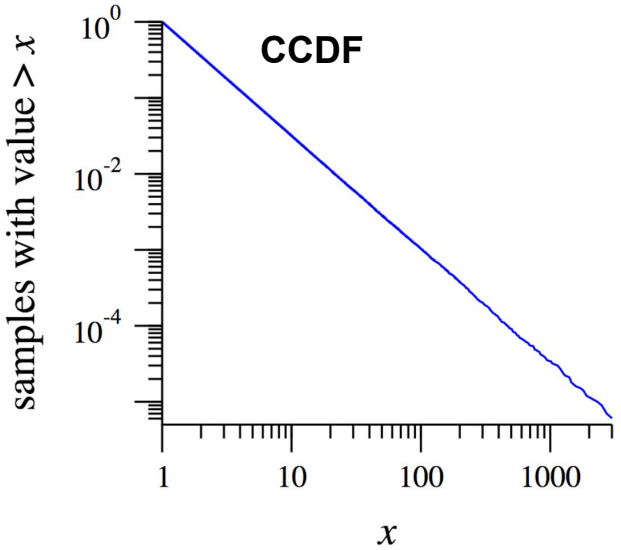
> 2 variables: radar charts

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



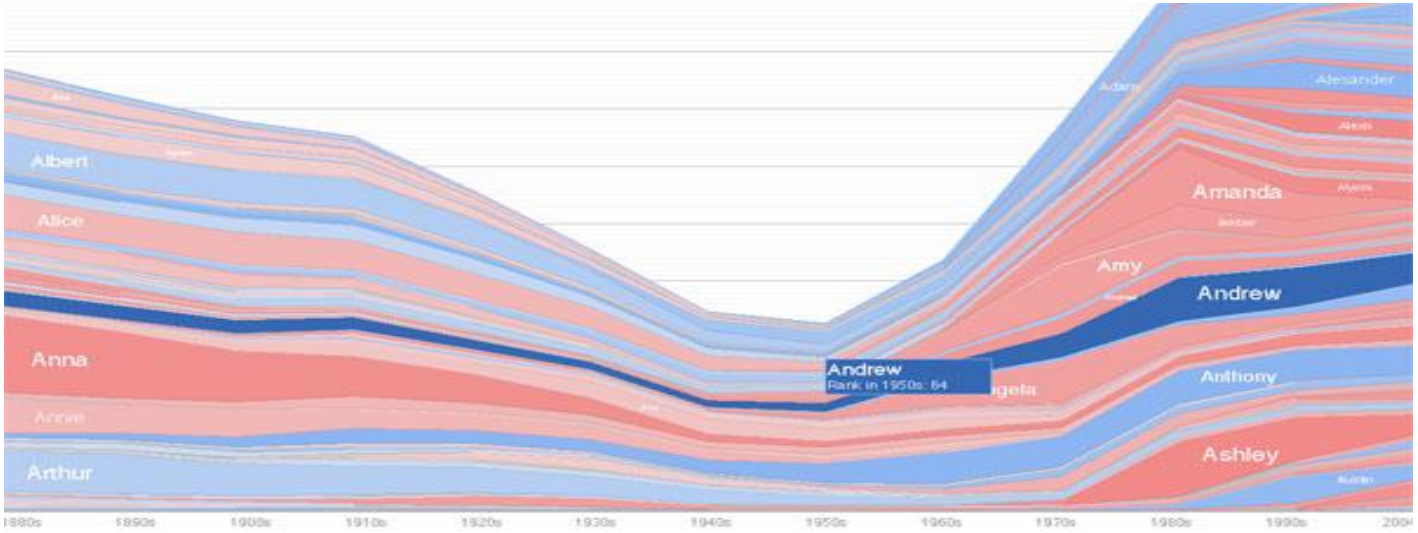
Heavy-tailed data: power laws

- 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)



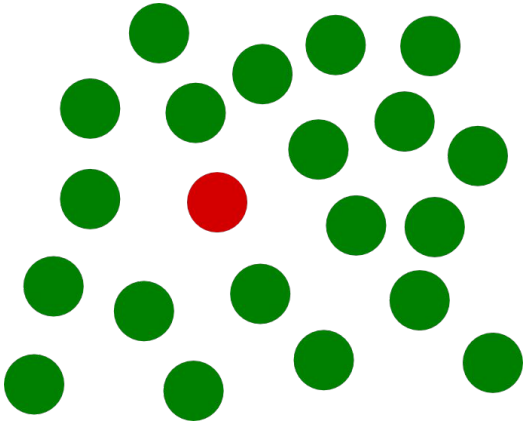
Interactive chart design: simplifying

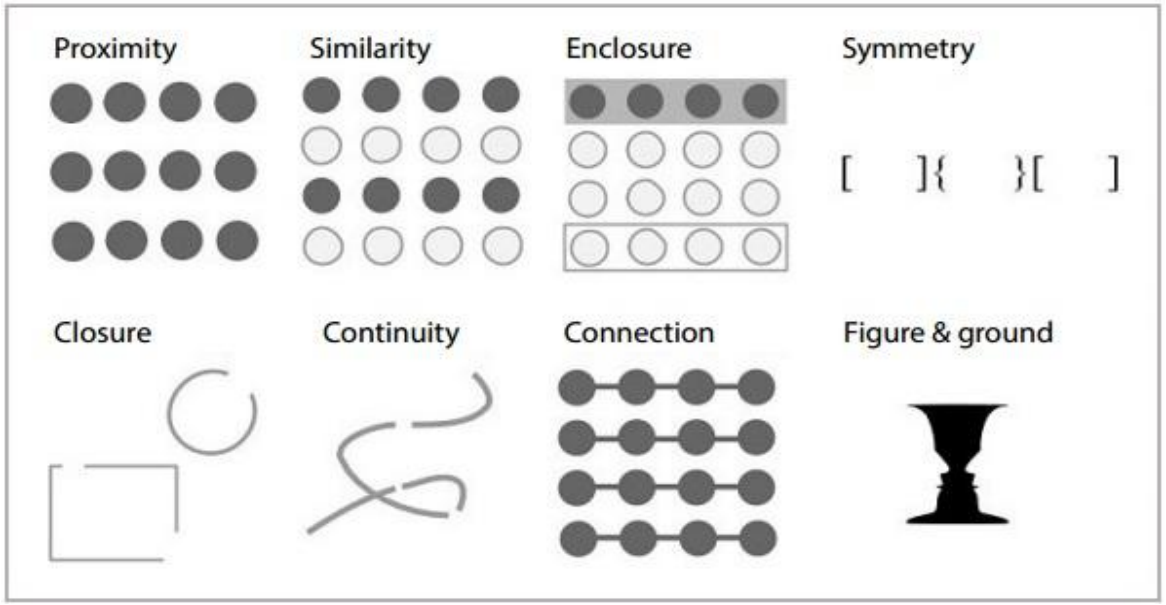
- With interactive charts you can keep things very simple by **hiding** and **dynamically revealing** important structure.
- On an interactive chart, you reveal the information most useful for **navigating** the chart.



Use structure!

Gestalt psychology principles (1912)



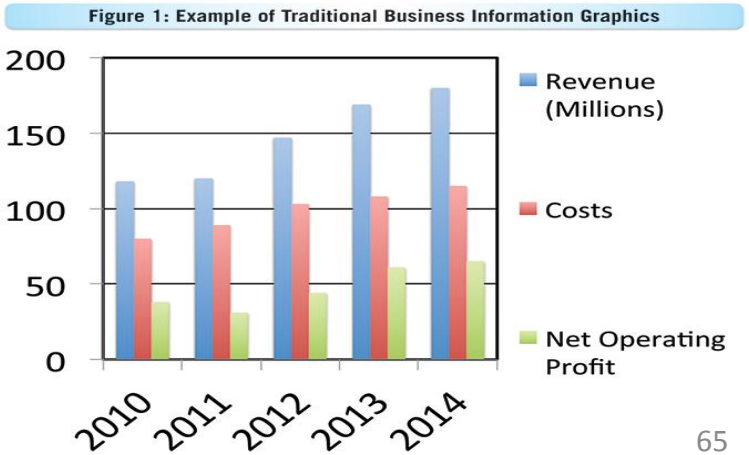


A case for ugly visualizations

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



Guide your audience!