

Applied Data Analysis (CS401)



Lecture 3
Visualizing data
4 Oct 2023

EPFL

Robert West



Announcements

- You must find 4 team mates and [register](#) by this Fri 6 Oct 23:59
 - Every student must register individually
 - Still looking for a team or for team members? Use Ed! If need be, talk to Sepideh [TA]
- Project milestone 1 was released last Fri; due Fri 13 Oct 23:59
- Homework 1 released Fri 13 Oct (due two weeks later)
- Friday's lab session:
 - 13:15–13:30 (on Moodle): Quiz 2 (the first one that counts)
 - 13:30–14:30: Project milestone P1 office hours (Zoom)
 - In parallel: Exercise on data visualization and working with data from the Web

Announcements

Want to contribute to a (fun!) research study?

https://go.epfl.ch/wikispeedia_llm_ada_doc

Your mission: Madagascar >> British Empire
So far you've clicked 0 links: Madagascar

You think this article should link to [British Empire](#), but it doesn't? — [Report!](#)

Wikispeedia [Start over](#)

Madagascar

2007 Schools Wikipedia Selection. Related subjects: African Countries; Countries

SOS Children works in Madagascar. For more information see [SOS Children in Madagascar, Africa](#)

Madagascar (officially the **Republic of Madagascar**) or **Malagasy Republic**, is an island nation in the [Indian Ocean](#), off the southeastern coast of [Africa](#). The main island, also called Madagascar, is the fourth largest island in the world, and is home to five percent of the world's [plant](#) and [animal](#) species, (more than 80 percent of which are indigenous to Madagascar.) Most notable are the lemur infraorder of primates, the carnivorous fossa, three endemic bird families and six endemic baobab species. The adjective for Madagascar is *Malagasy* (pronounced "ma-la-gas-ee" or "mal-a-gash"), and the official national language is the Malagasy language.

History

The first settlers came from Asia, rather than Africa, circa 700 AD. The culture shows the influence of both Africa and Asia. The settlement represented the western-most branch of the great Austronesian expansion. Some of the strongest evidence indicating that the settlers of Madagascar came from this region is the linguistic similarity between the Malayo-Polynesian and Malagasy languages.

Republikan'i Madagasikara
République de Madagascar
Republic of Madagascar

 Flag
 Coat of arms

Motto: "Tanindrazana, Fahafahana, Fandrosoana" (Malagasy)
"Ancestral-land, Liberty, Progress"

Anthem: *Ry Tanindrazana nay malala ô*
Oh, Our Beloved Ancestral-land

Feedback

Give us feedback on this lecture here:

<https://go.epfl.ch/ada2023-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?](#)
- ...

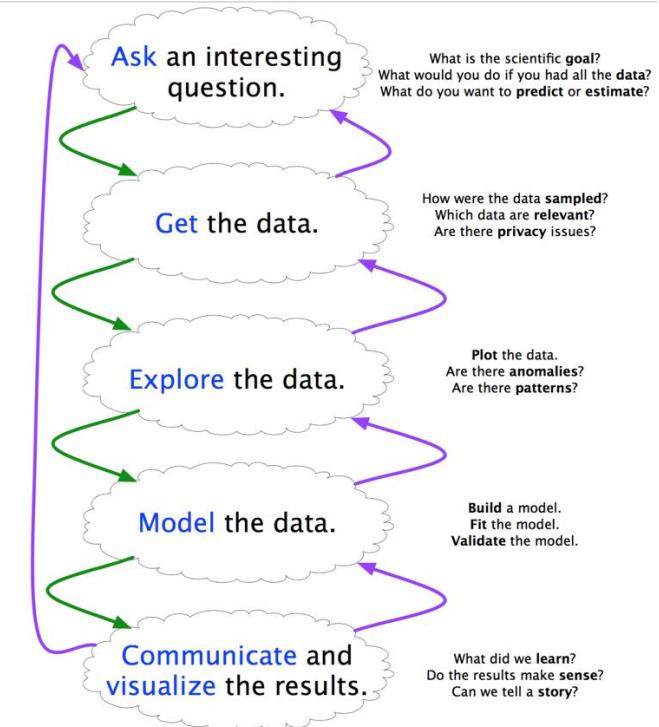
Uses for data visualization

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



Make it easier to evaluate potential courses of action (**decision-making**)

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

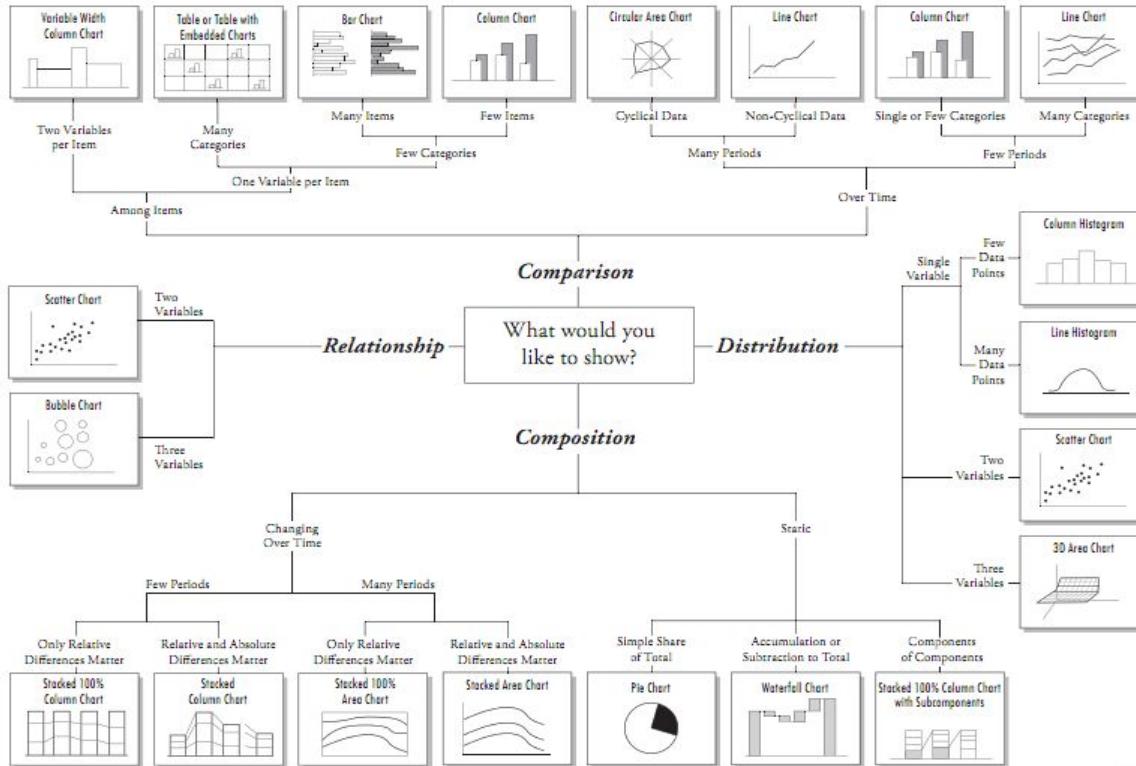
- Part 1: Navigating the chart landscape
- Part 2: Principles and best practices
- Part 3: A (small) selection of use cases for data visualization

Part 1

Navigating the chart landscape

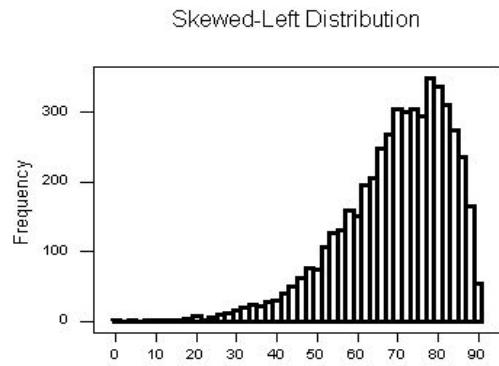
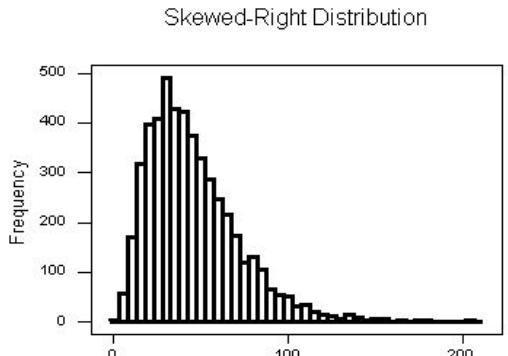
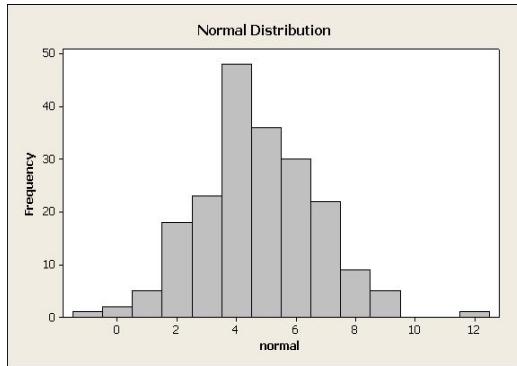
Chart selection

Chart Suggestions—A Thought-Starter

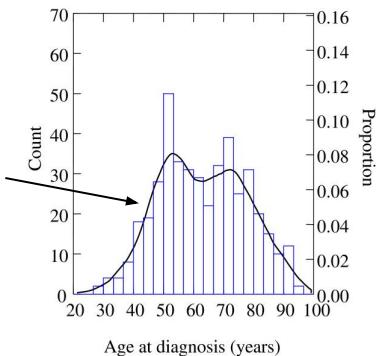


One variable: histograms

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

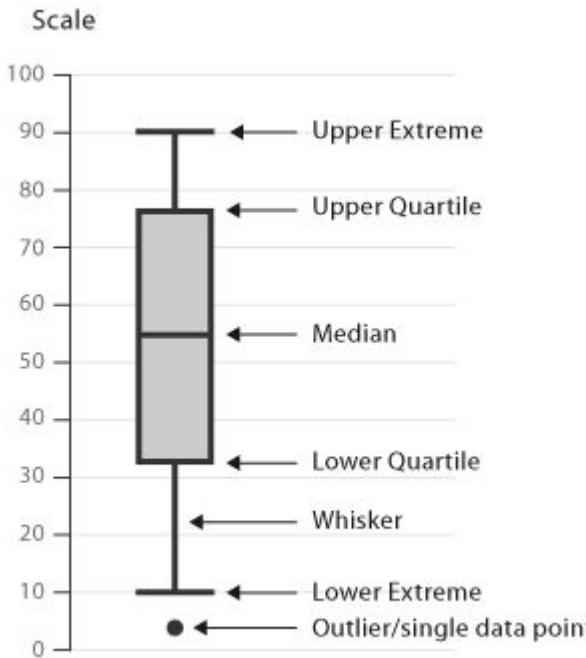


Smoothed histogram (a.k.a. kernel density estimate)

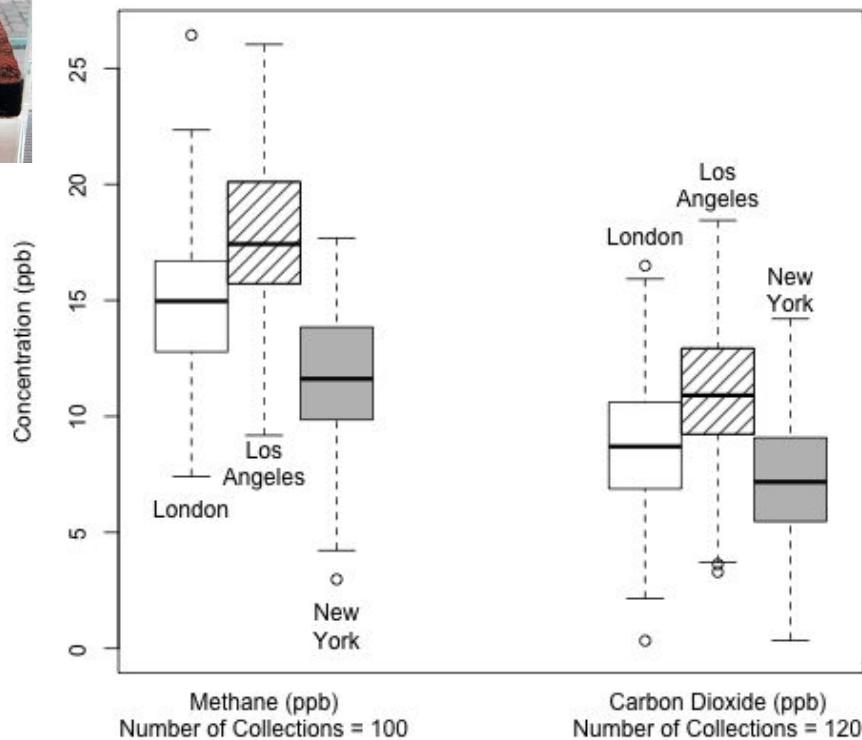


Easy to
recognize
skewed
distributions!

One variable: box plots

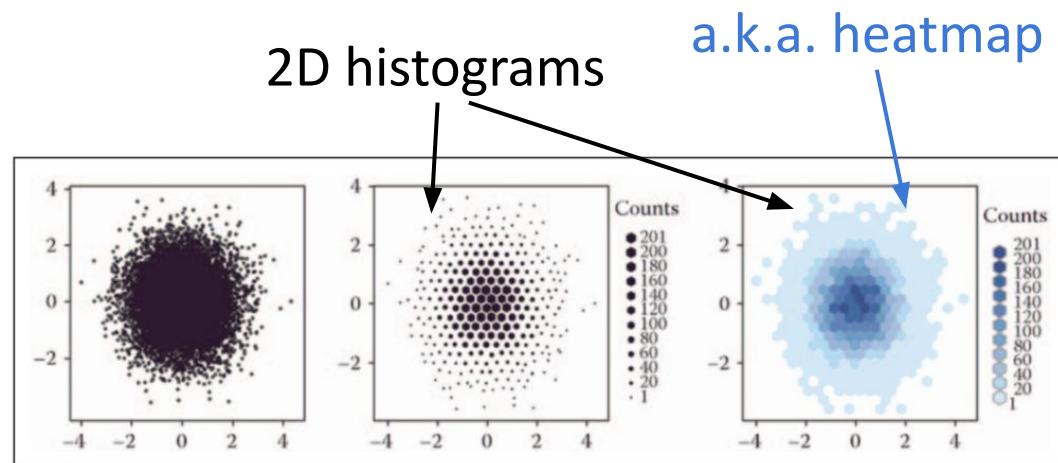
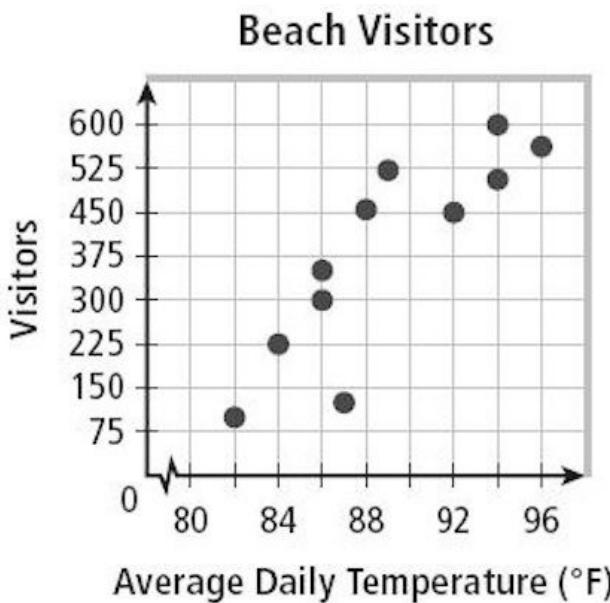


Comparing Pollution in London, Los Angeles, and New York



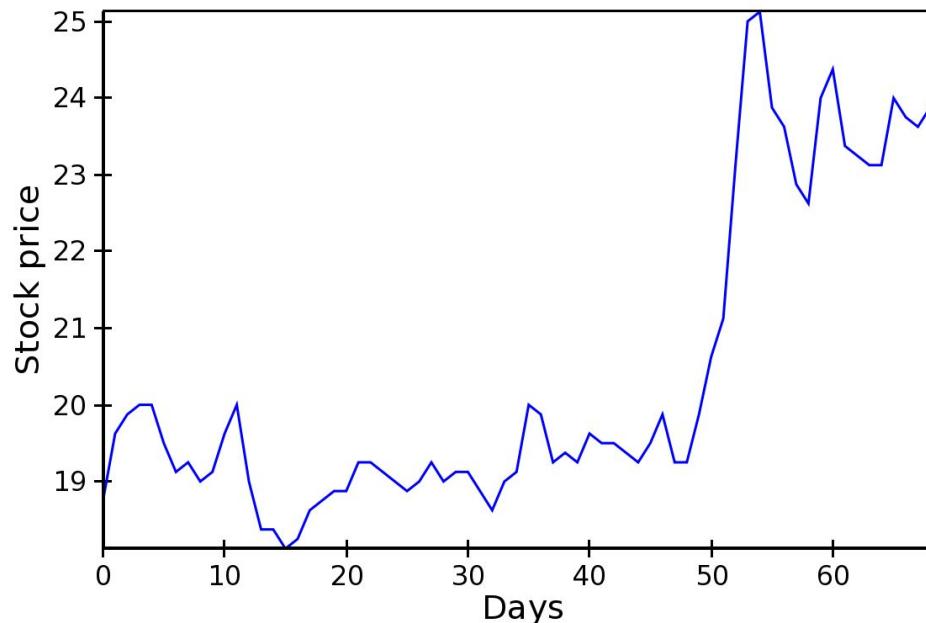
Two variables: scatter plots

Scatter plots quickly expose the relationships between two variables

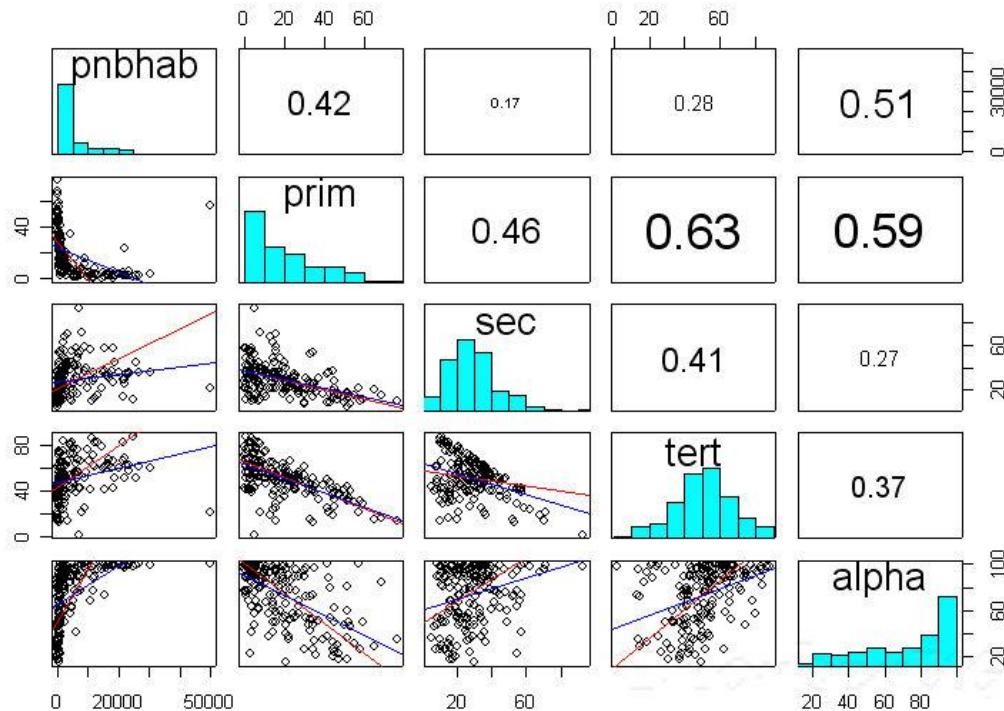


Two variables: line plots

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



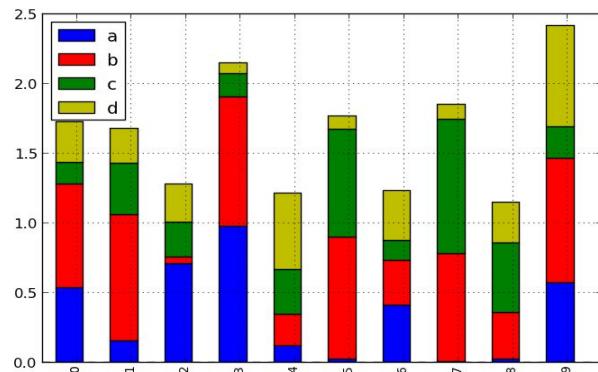
> 2 variables: scatter plot matrix



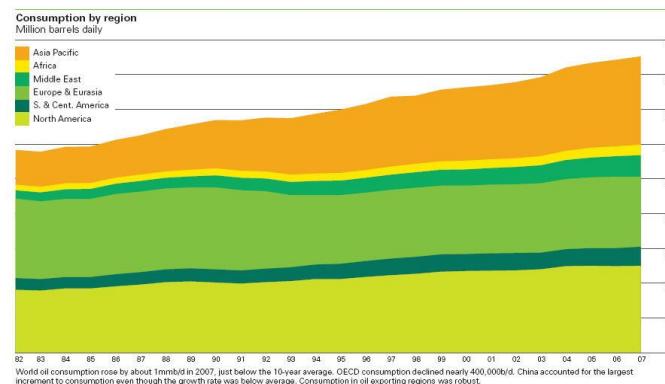
> 2 variables: stacked plots

Here: 3 variables: stack index, height, color

Stack variable and color variables categorical,
height variable continuous:

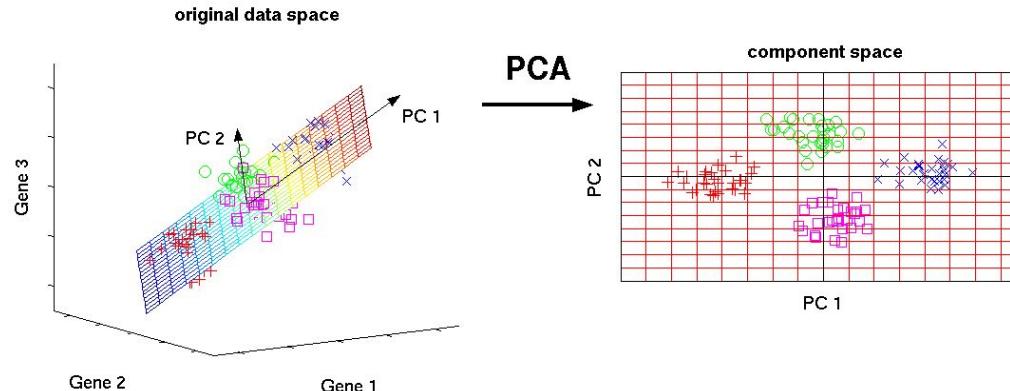


Color variable categorical,
stack and height variables continuous:

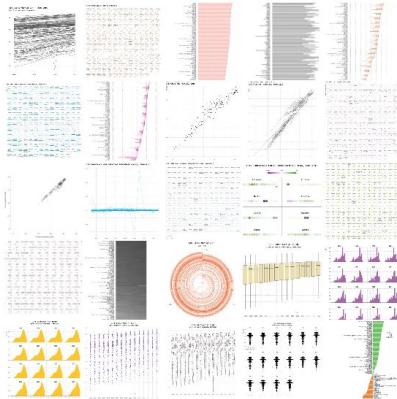


Dimensionality reduction

- 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



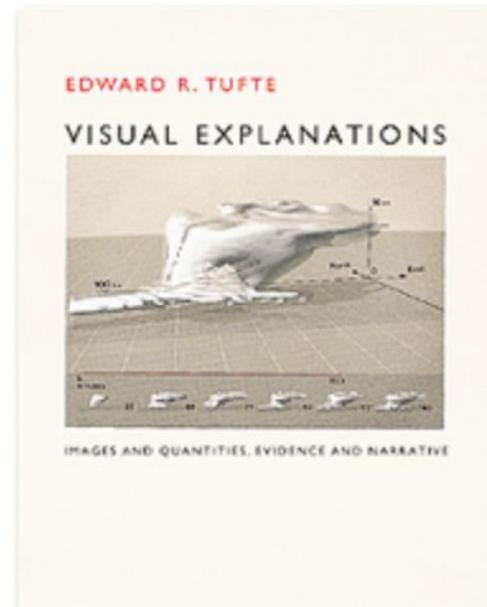
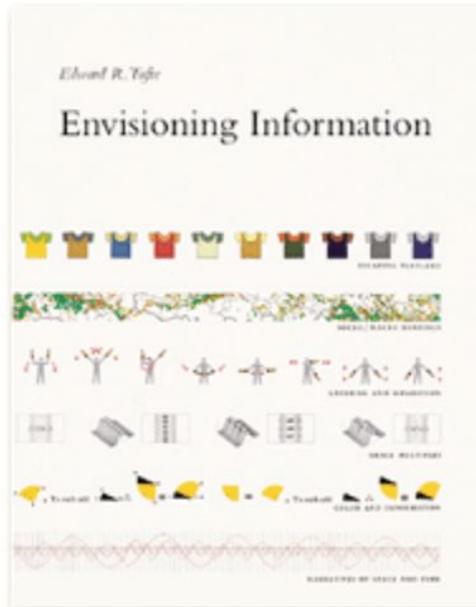
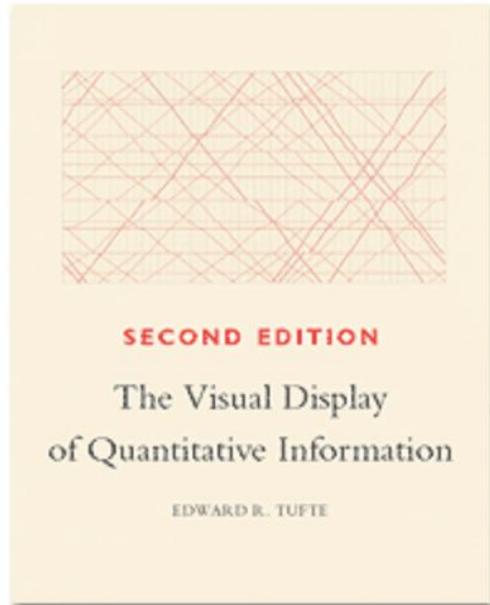
<http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways>

“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 Tufte



[[another great coffee table book](#)]

Perception of magnitudes

(134, 134, 134)



(144, 144, 144)



Which is brighter?

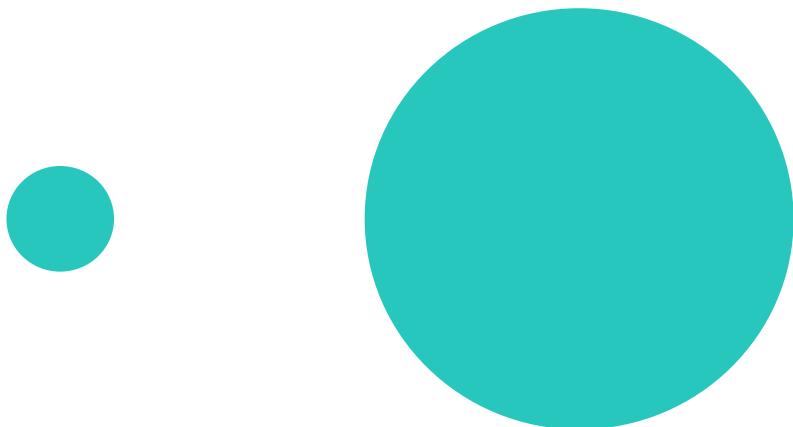
Just noticeable difference (JND)

- Weber's law:

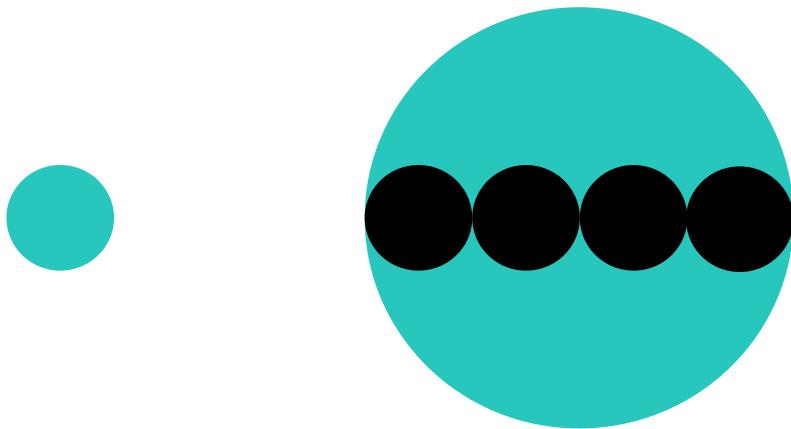
$$\frac{\Delta I}{I} = k,$$

- 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



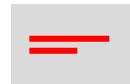
Compare area of circles

Perception of magnitudes

Most accurate



Position



Length



Slope



Angle



Area



Volume

Least accurate

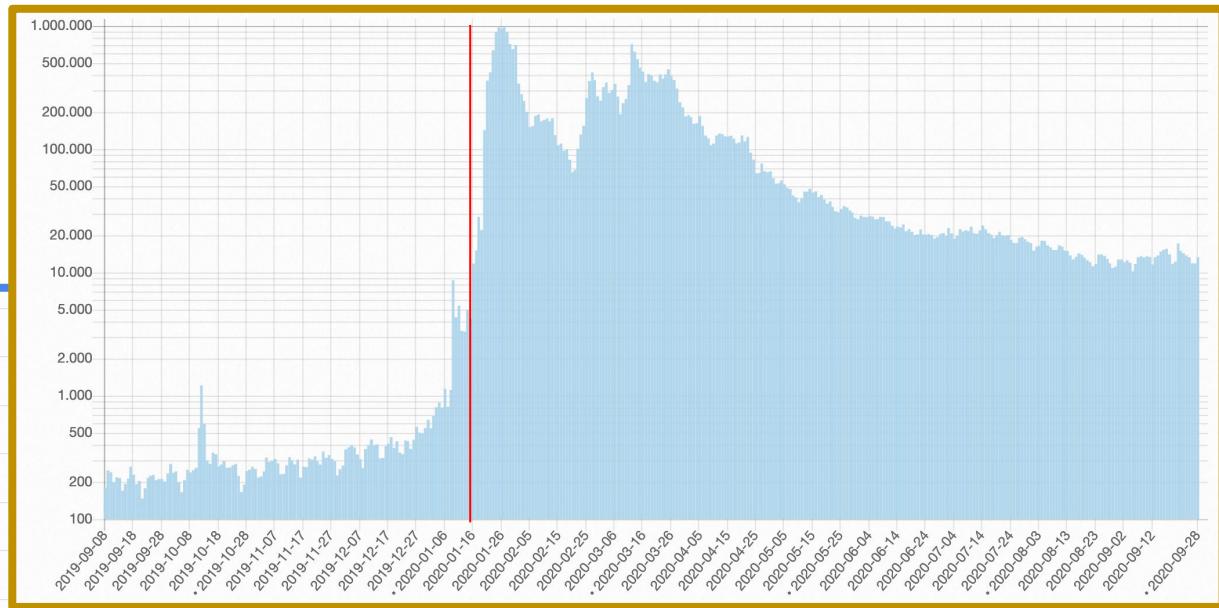
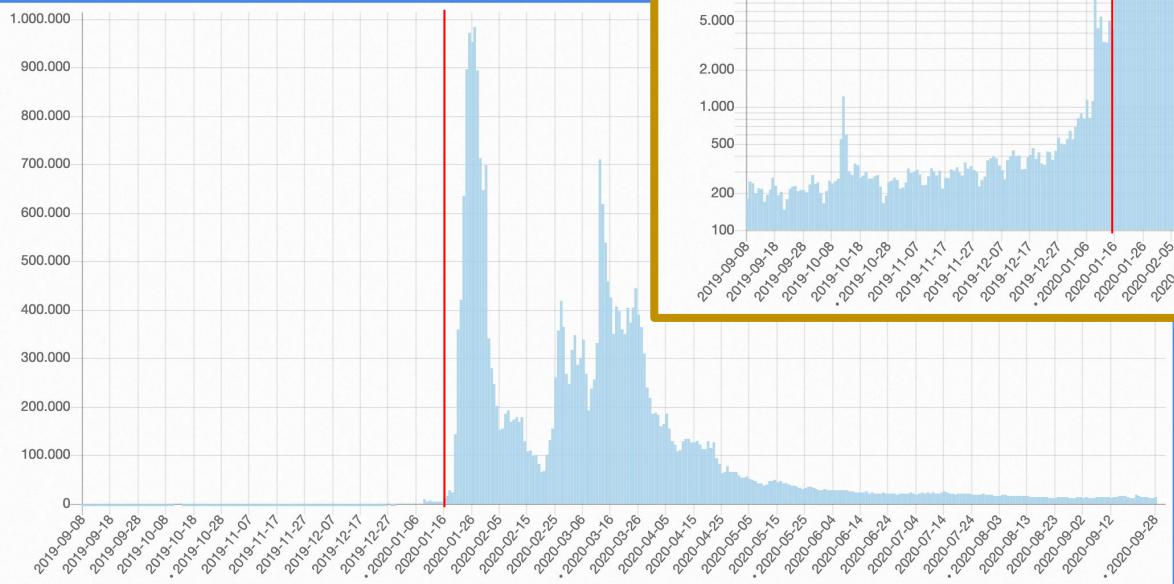


Color hue-saturation-density

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

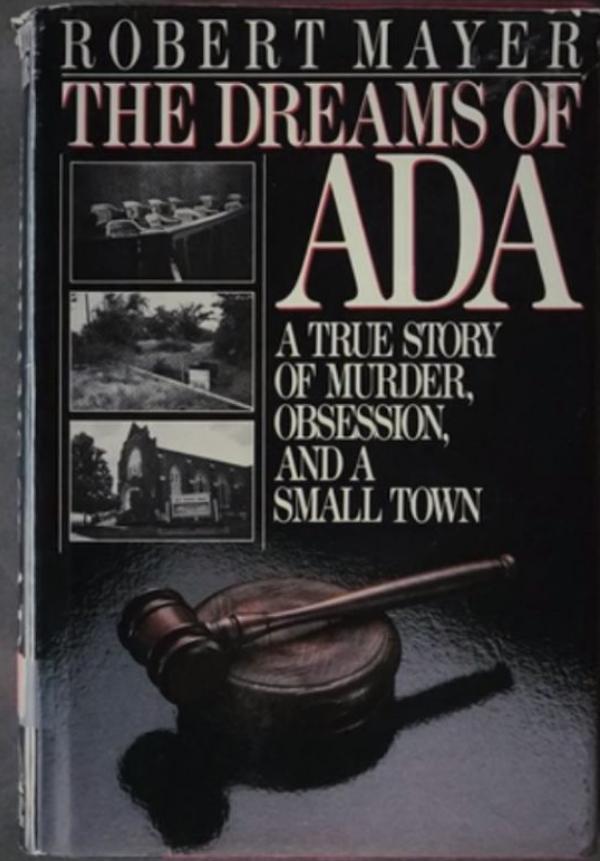
Choose your axes wisely!

Time series of pageviews
of Wikipedia article about
“Coronavirus”
(linear y-axis)

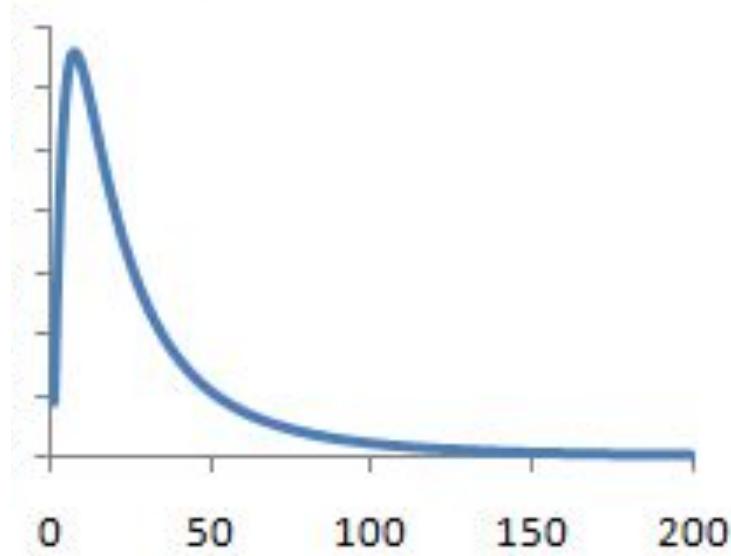


(logarithmic y-axis)

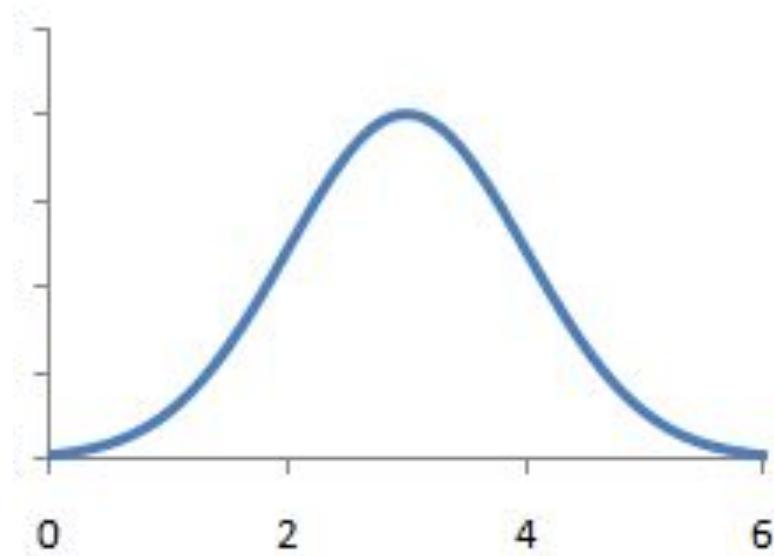
Commercial
break



Choose your axes wisely: Visualizing heavy-tailed distributions



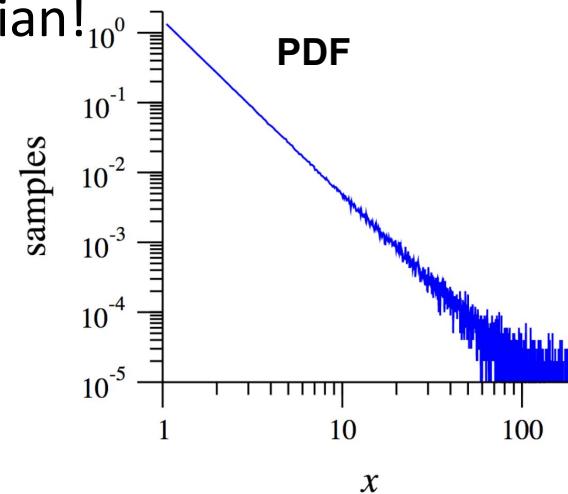
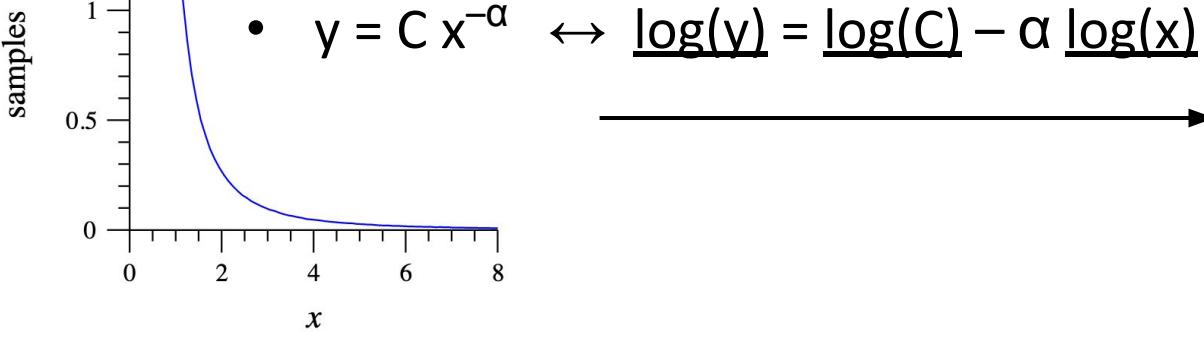
Linear x-axis



Logarithmic x-axis

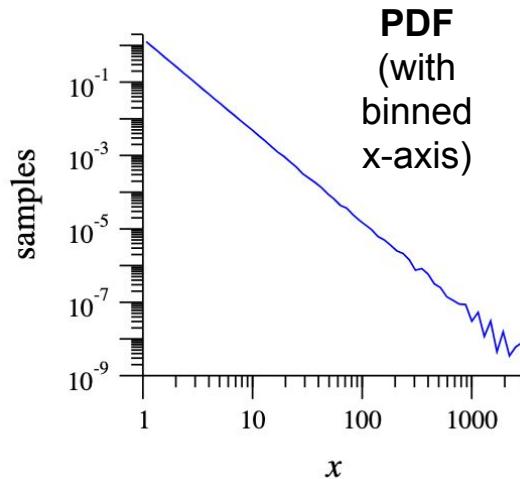
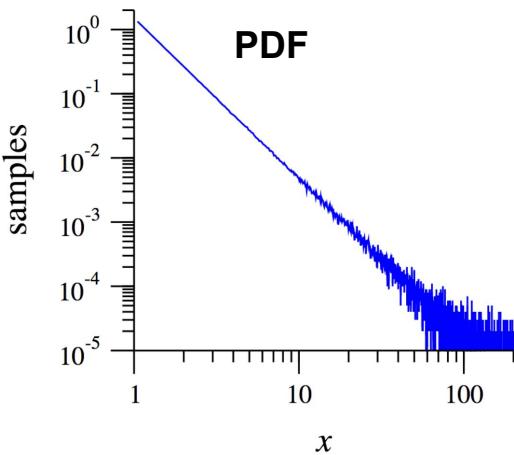
Heavy-tailed data: power laws

- Probability of x : $p(x) = Cx^{-\alpha}$,
 - 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 = $\infty \rightarrow$ use median!
 - 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):
 $P(x) := \Pr\{X \geq x\}$



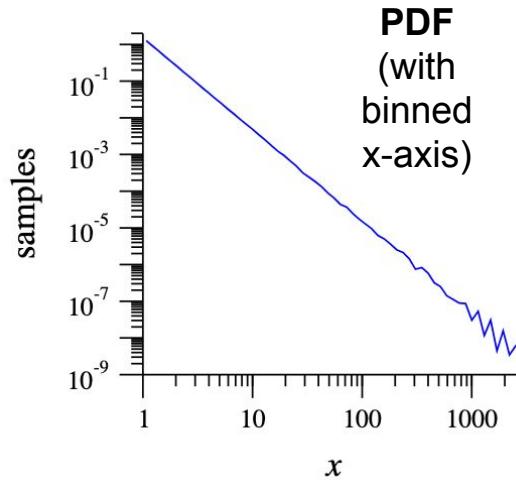
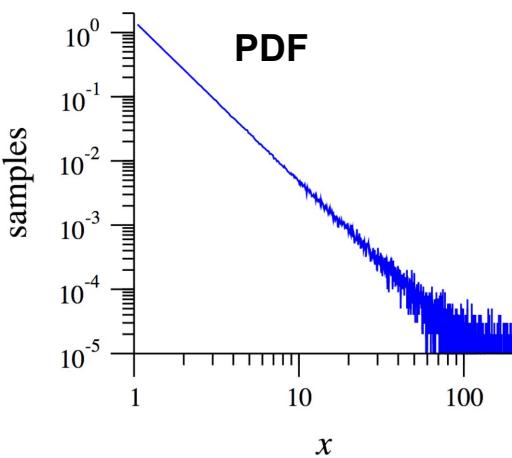
POLLING TIME

- “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>

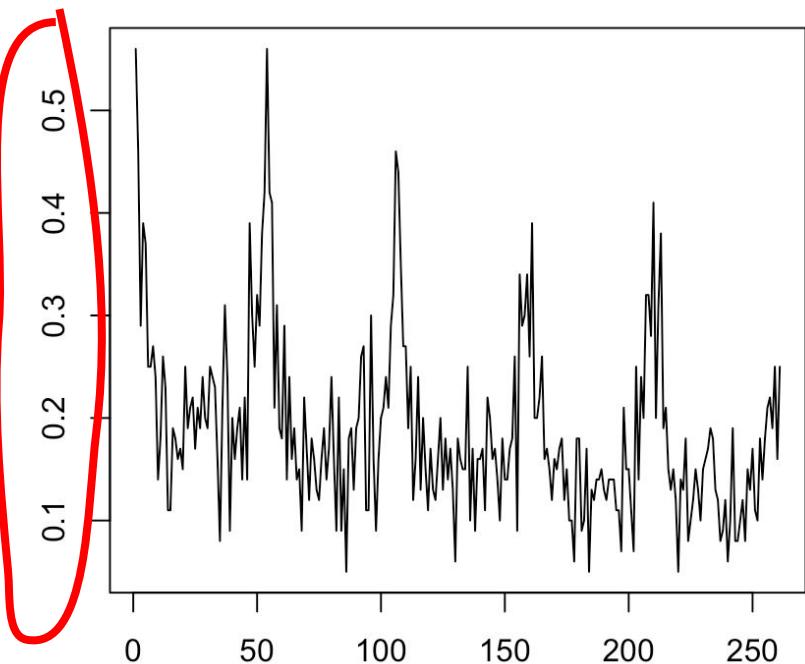
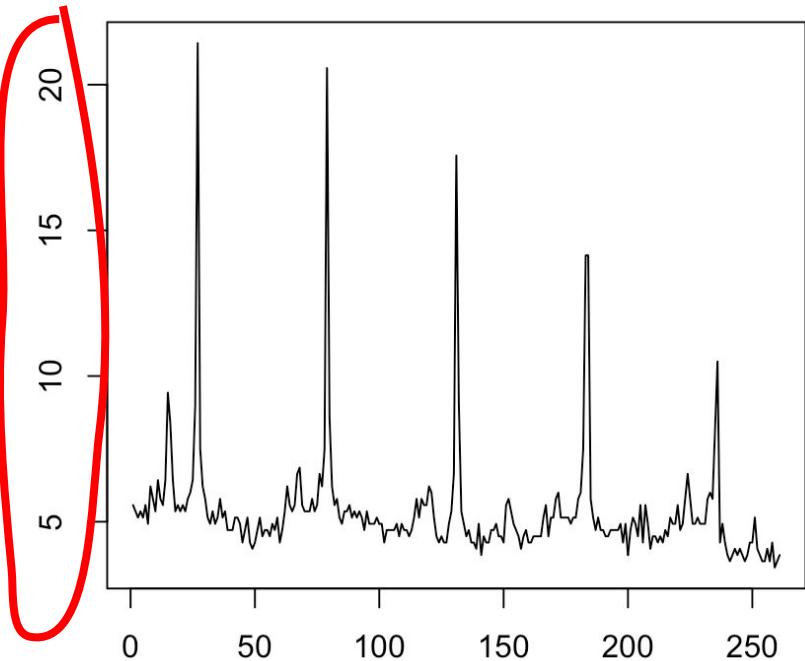


Heavy-tailed data: power laws

- Complementary cumulative distribution function (CCDF):
 $P(x) := \Pr\{X \geq x\}$
- CCDF of power law is also a power law (with exponent $\alpha - 1$)
$$P(x) = C \int_x^{\infty} x'^{-\alpha} dx' = \frac{C}{\alpha - 1} x^{-(\alpha - 1)}.$$
- CCDF plot is monotonically decreasing (even without binning)

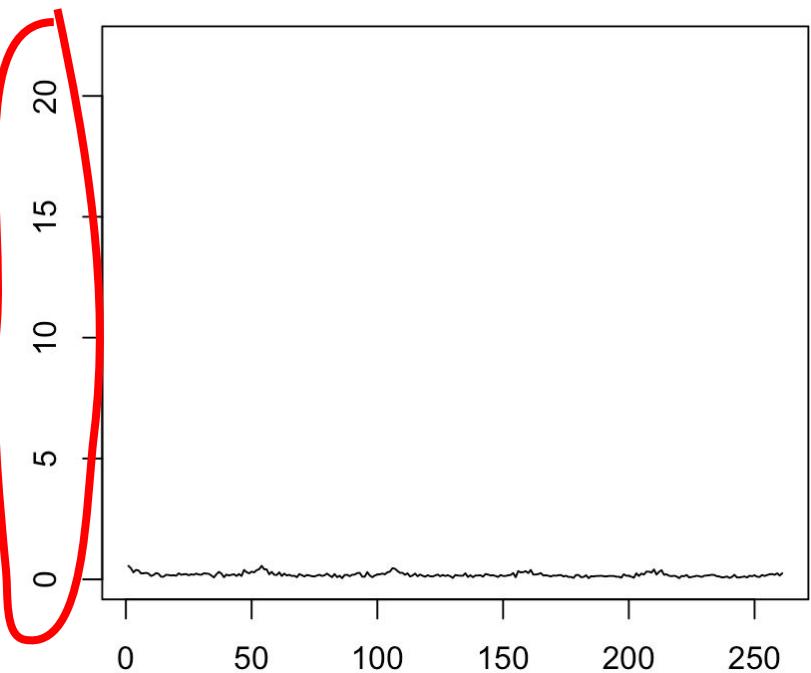
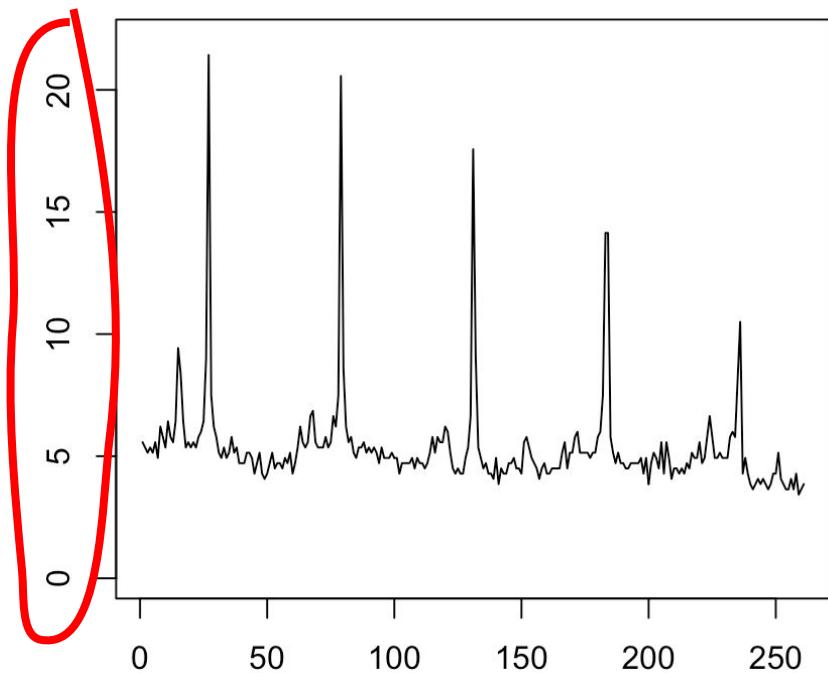


Answer fast: which time series has a higher mean value?

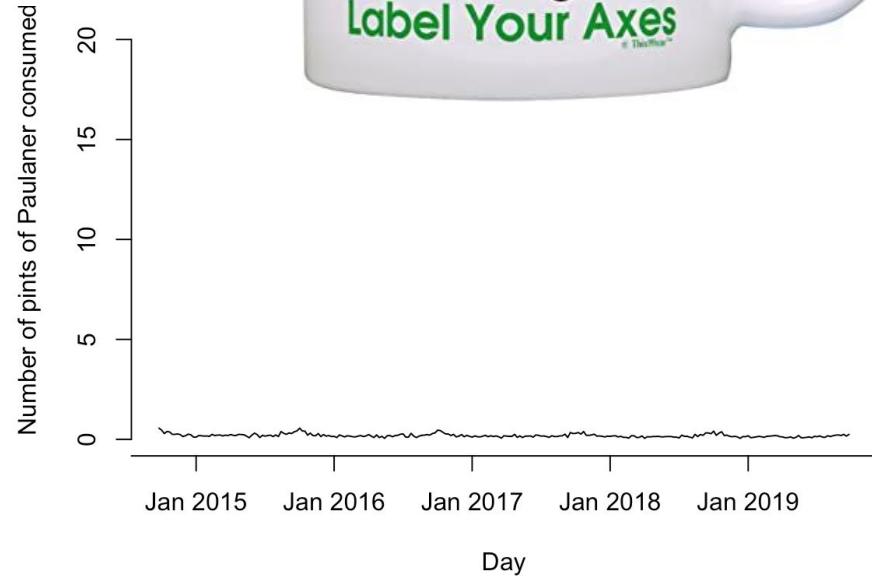
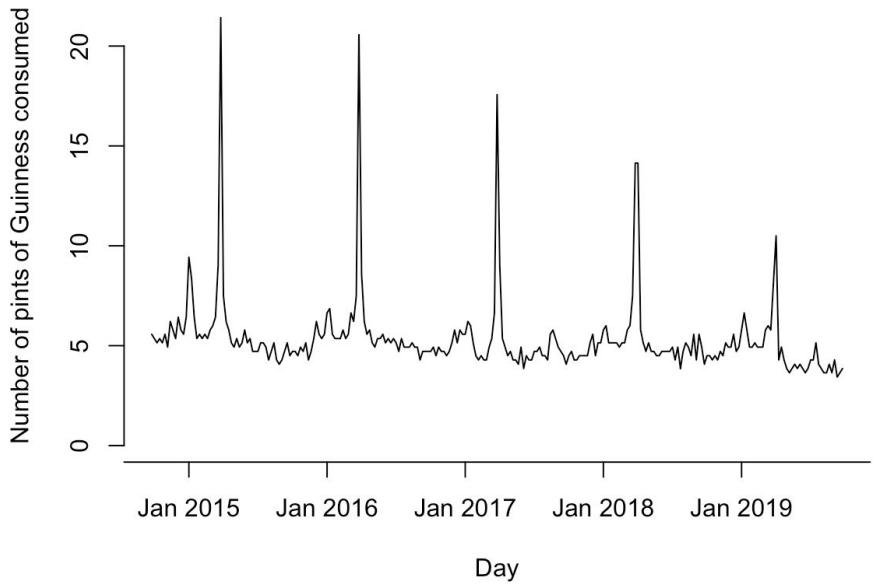


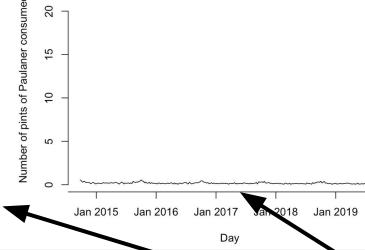
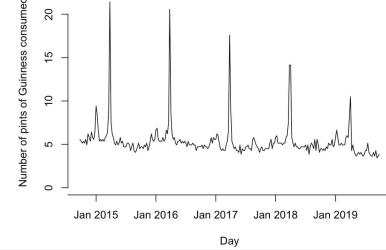
Use consistent axes!

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



Label your axes!

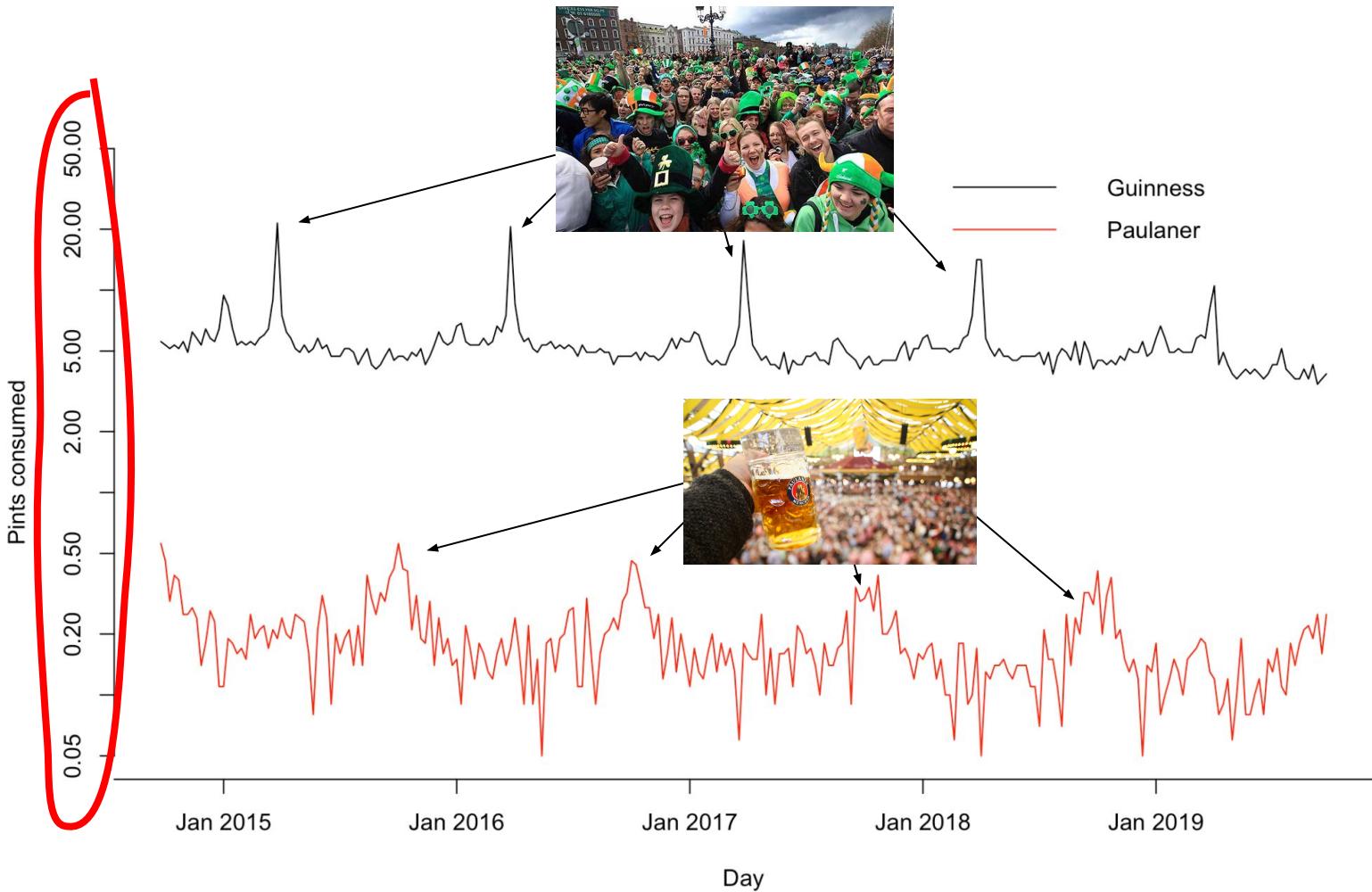




THINK FOR A MINUTE:

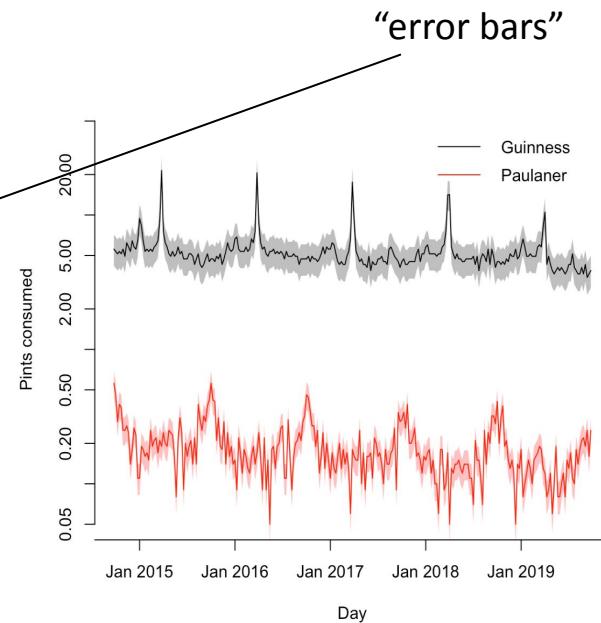
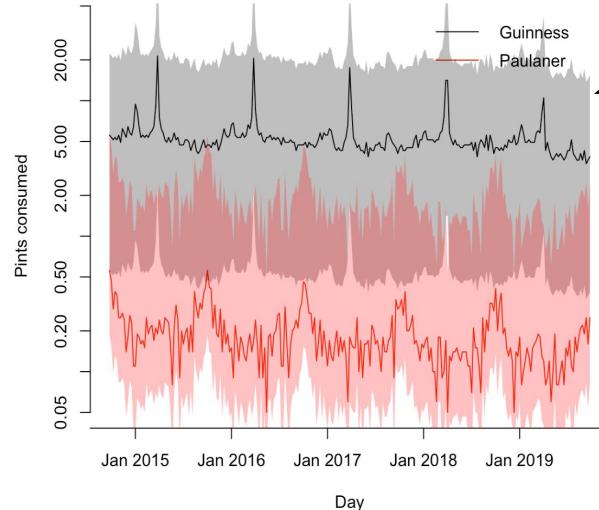
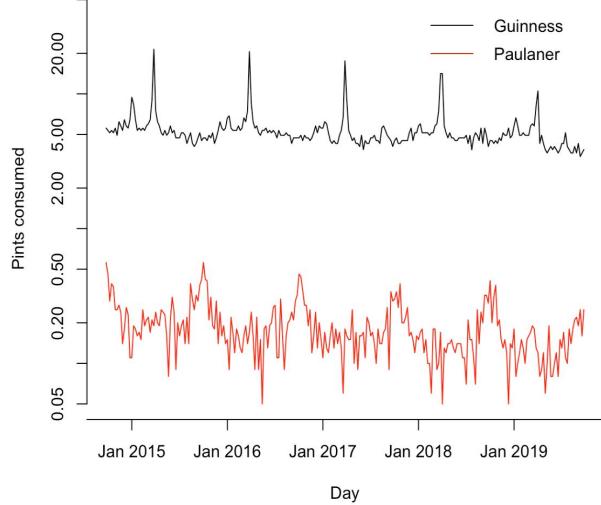
How could we show details of both time series without using different y-axes?

(Feel free to discuss with your neighbor.)

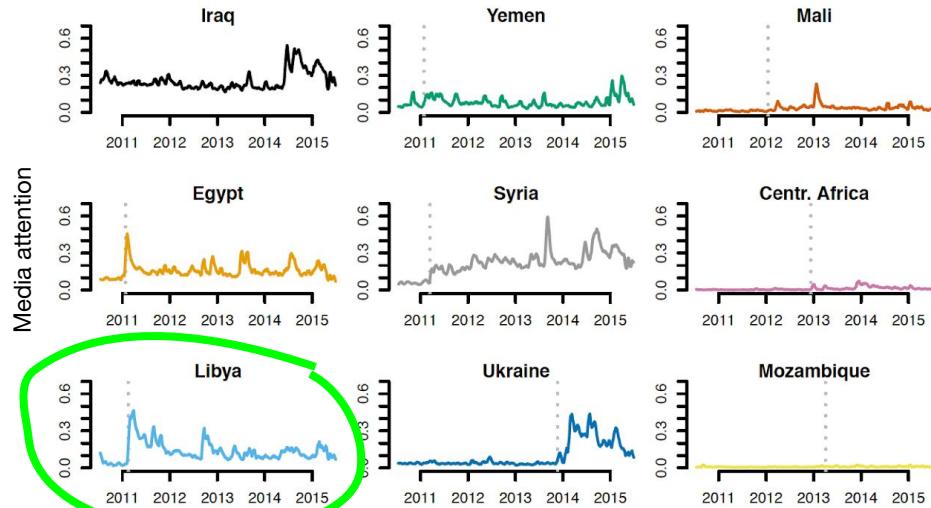


Show data uncertainty!

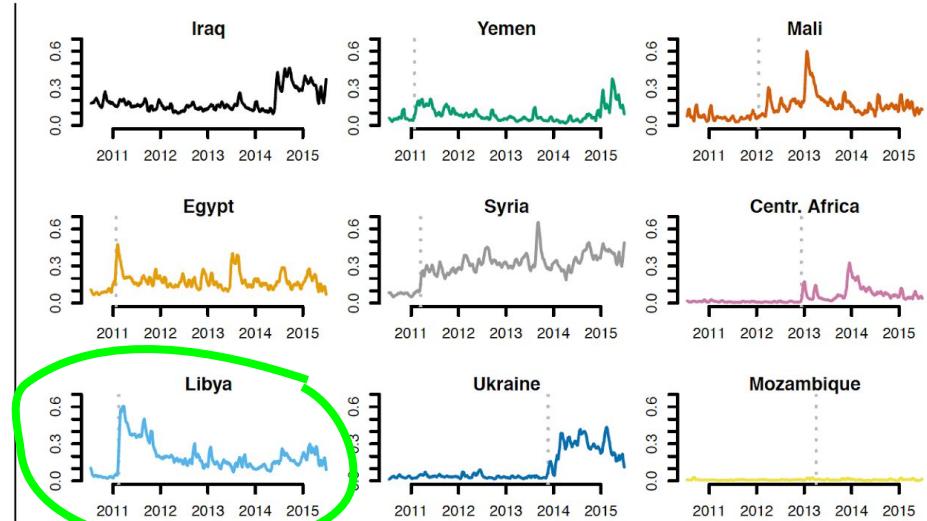
Which beer is more popular, **Guinness** or **Paulaner**?



Consider using small multiples!



(a) English



(b) French

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



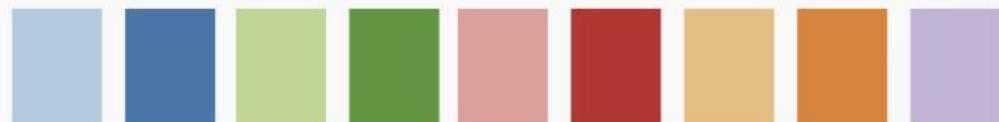
Diverging

Two sequential schemes extended out from a critical midpoint value



Categorical

Lots of contrast between each adjacent color



Number of data classes: 3

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

Nature of your data:

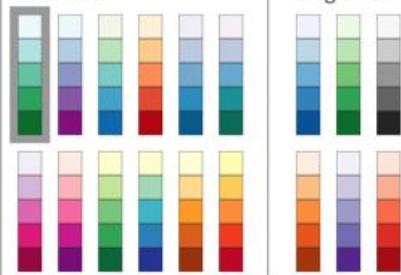
sequential diverging qualitative

COLORBREWER 2.0

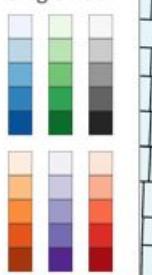
color advice for cartography

Pick a color scheme:

Multi-hue:



Single hue:



Only show:

- colorblind safe
- print friendly
- photocopy safe

Context:

- roads
- cities
- borders

Background:

- solid color
- terrain

3-class BuGn



#e5f5f9

#99d8c9

#2ca25f

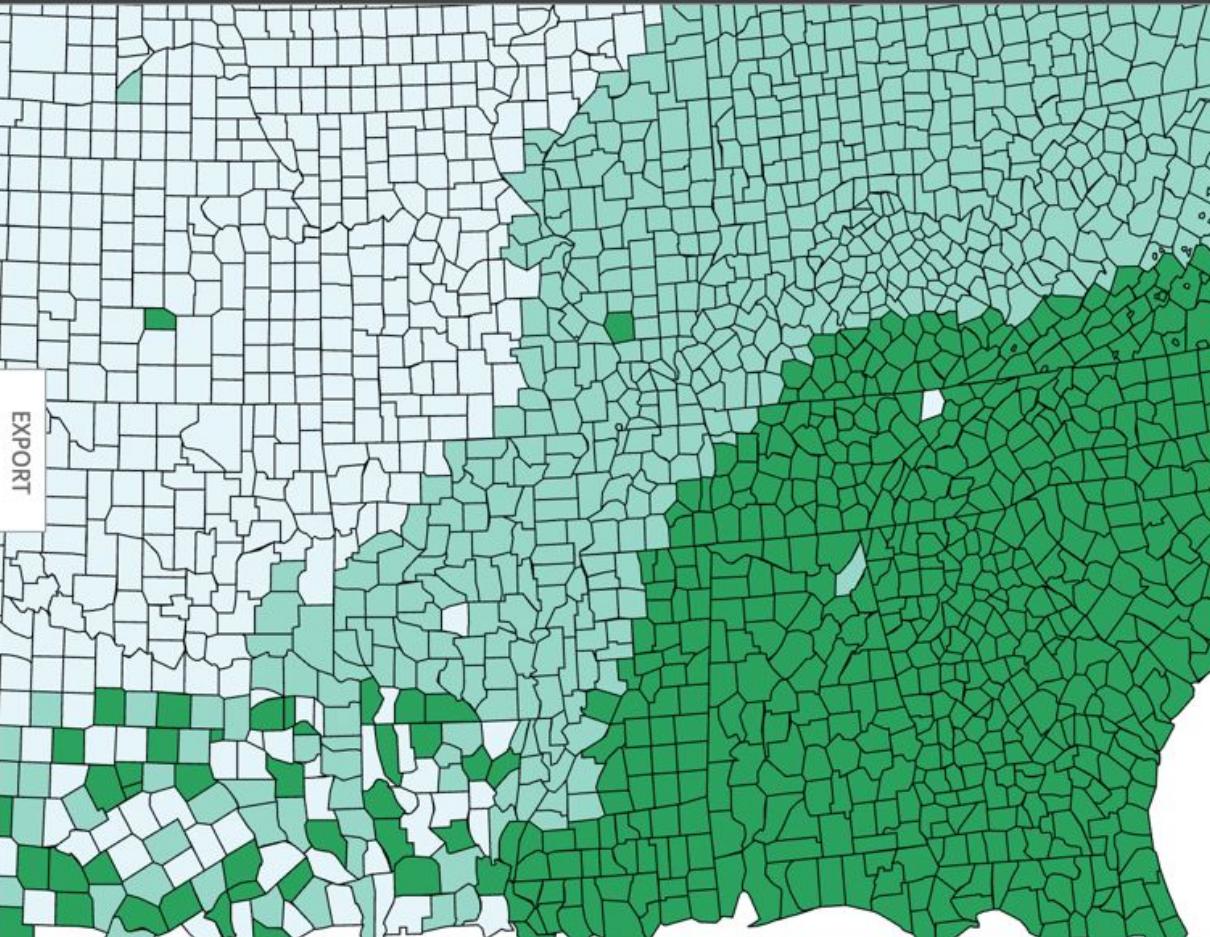
EXPORT

#e5f5f9

#99d8c9

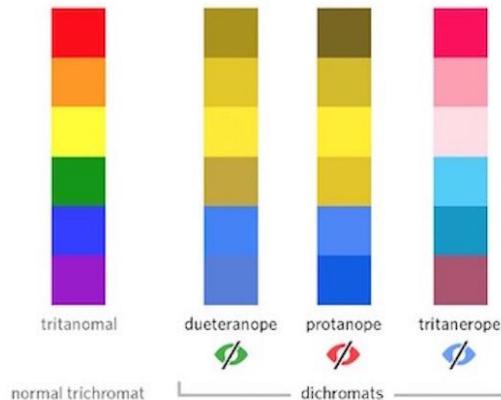
#2ca25f

color transparency



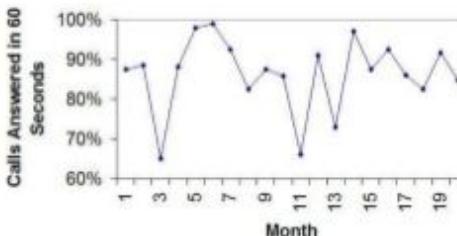
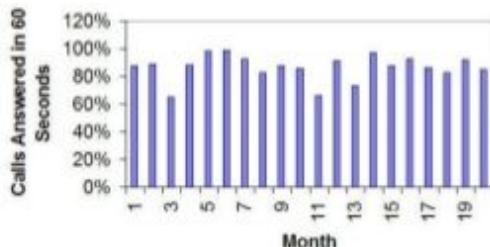
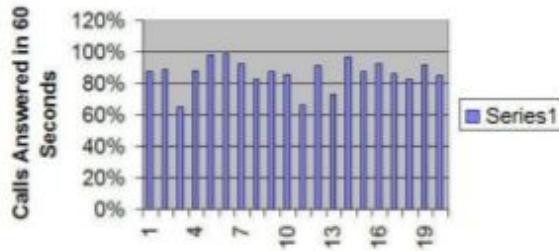
Use colorblind-safe palettes!

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



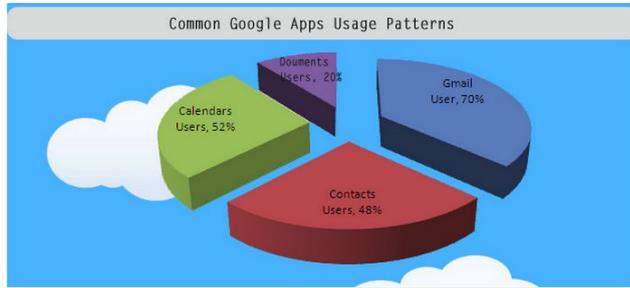
Original	Simulation			Hue	for Photoshop, Illustrator, Freehand, etc.		for Word, Power Point, Canvas, etc.	
	Protan	Deutan	Tritan		C,M,Y,K (%)	R,G,B (0-255)	R,G,B (%)	
1 Black	—°	(0,0,0,100)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	
2 Orange	41°	(0,50,100,0)	(230,159,0)	(90,60,0)	(0,50,100,0)	(230,159,0)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

Use data ink wisely! Avoid chart junk!

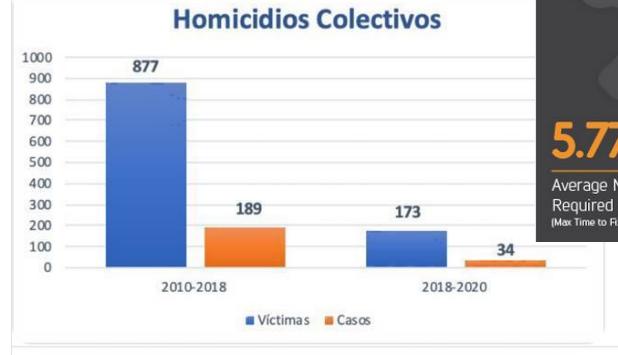


Graphical excellence gives
the viewer the greatest
number of ideas in the
shortest time with the least
ink in the smallest space.
-Edward Tufte

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



90% of US Households Consume Peanut Butter

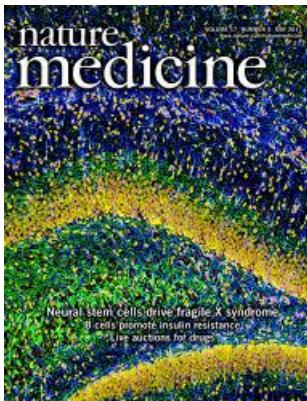
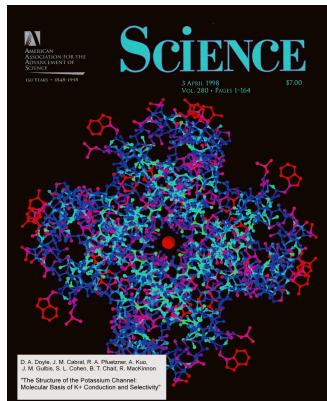
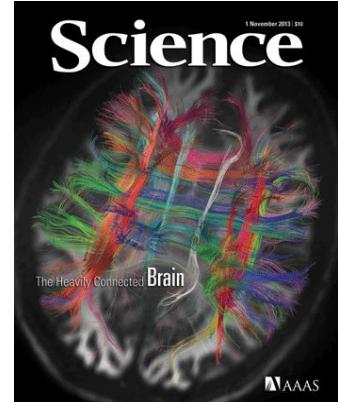
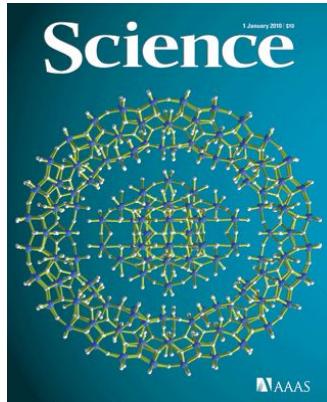
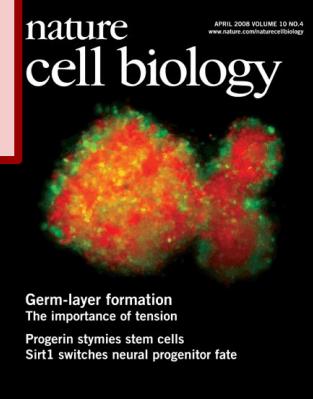


Courtesy of viz.wtf

Part 3

A (small) selection of use cases for data visualization

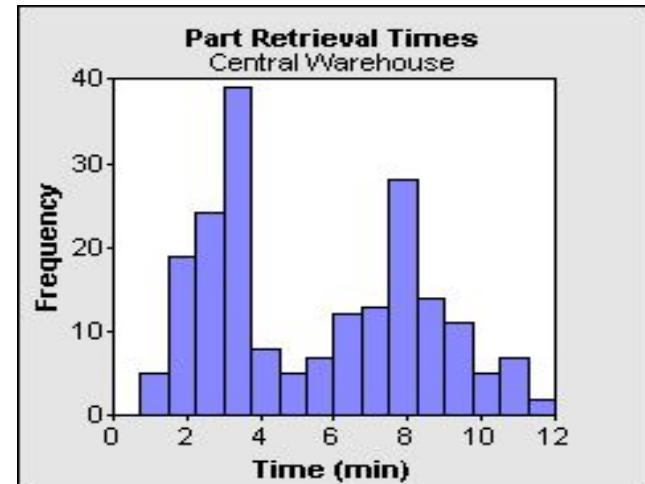
Use case: Presenting scientific results



Use case: Data wrangling

Multimodal data

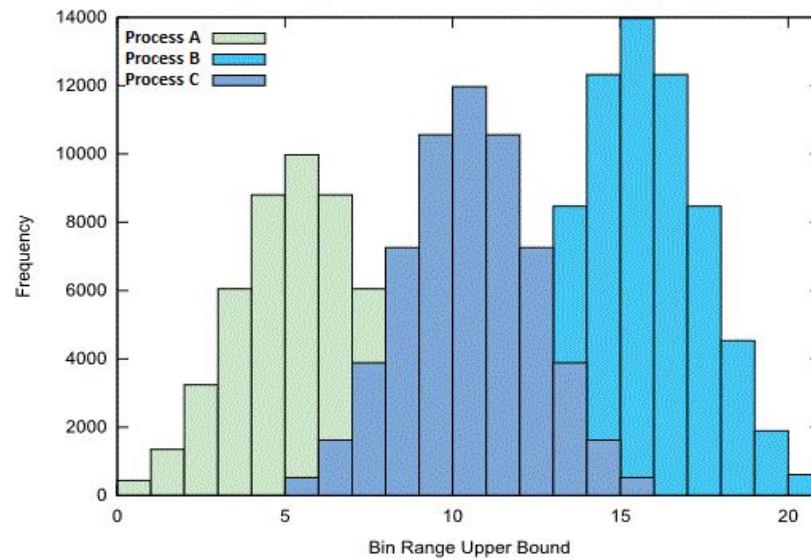
- Two or more distinct peaks in a histogram often **suggest 2 or more distinct populations** of samples.
- But don't guess! Explore further by using, e.g., color and a histogram of multiple populations (p.t.o.).



Use case: Data wrangling

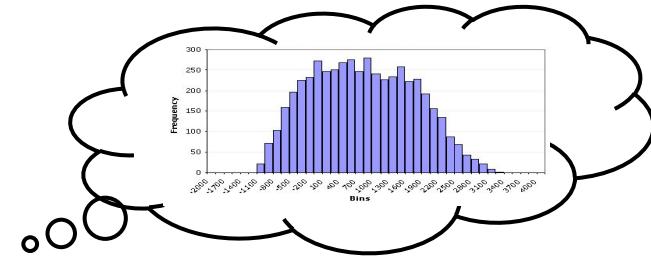
Multimodal data

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

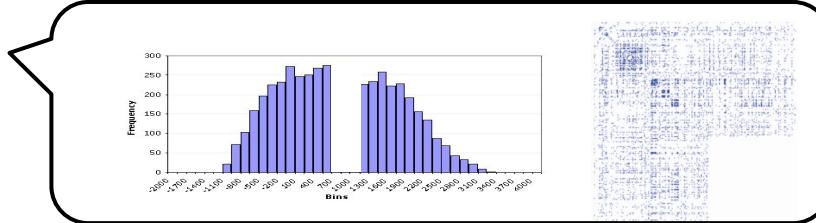


Use case: Data wrangling

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

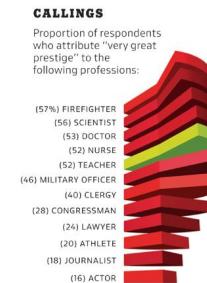
NY Times interactive visualizations (recession/recovery 2014)

<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:

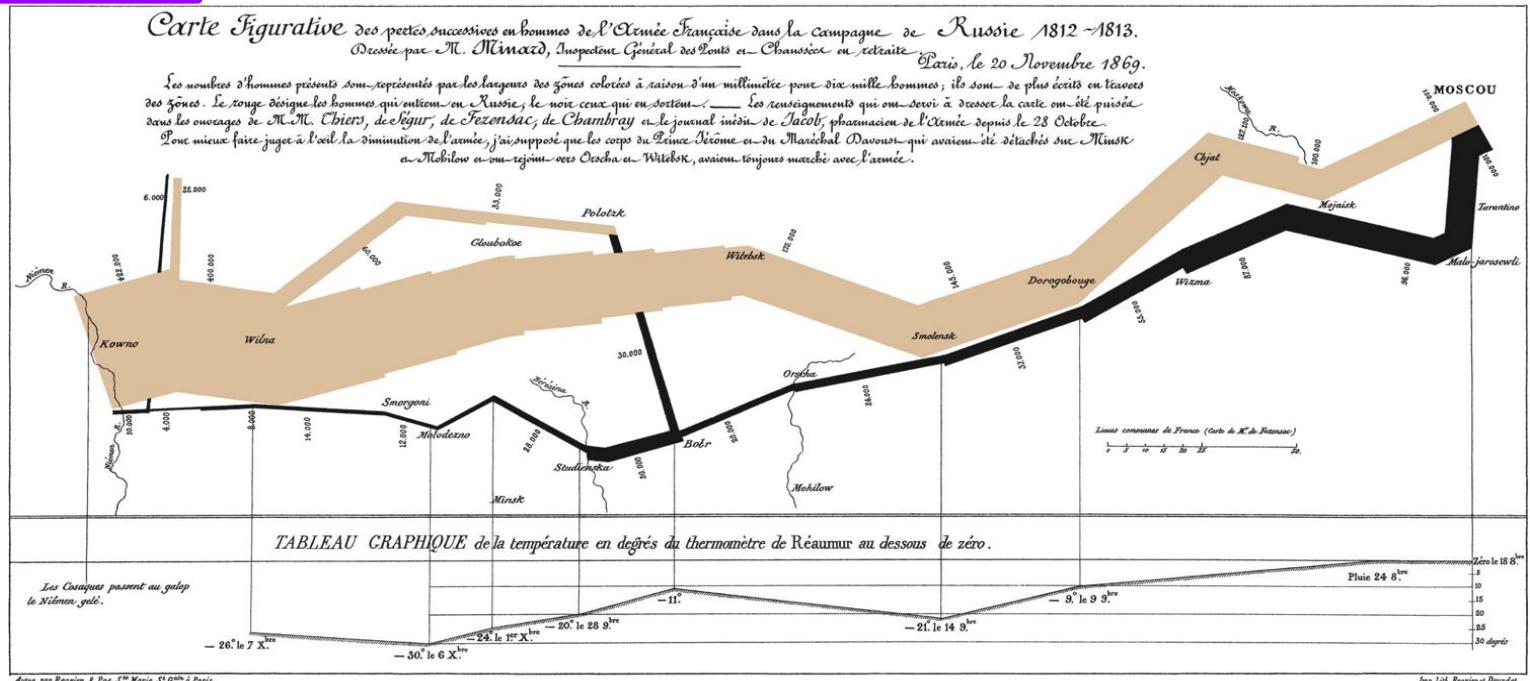
200 countries, 200 years, 4 minutes

<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**.

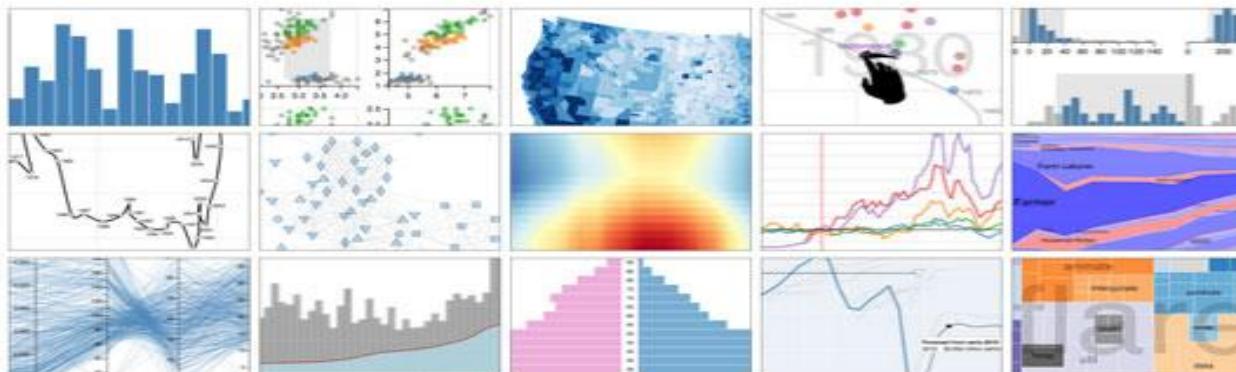
Note from the authors: *D3 is intentionally a low-level system.*
During the early design of D3, we even referred to it as a "visualization kernel" rather than a "toolkit" or "framework"

Interactive toolkits: Vega

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

It specifies graphics in JSON format.

vega

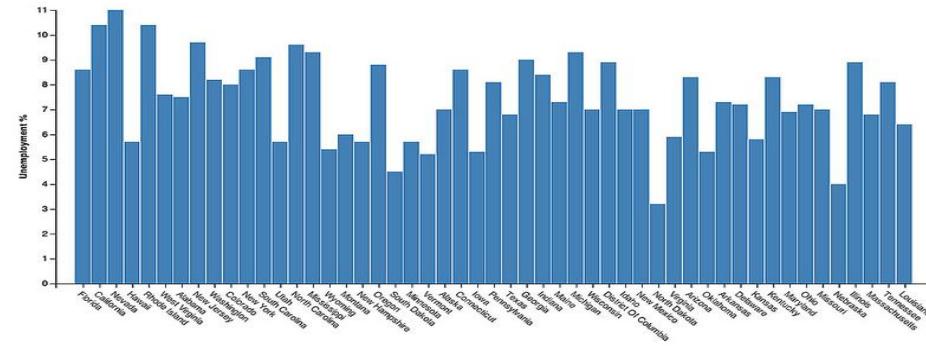
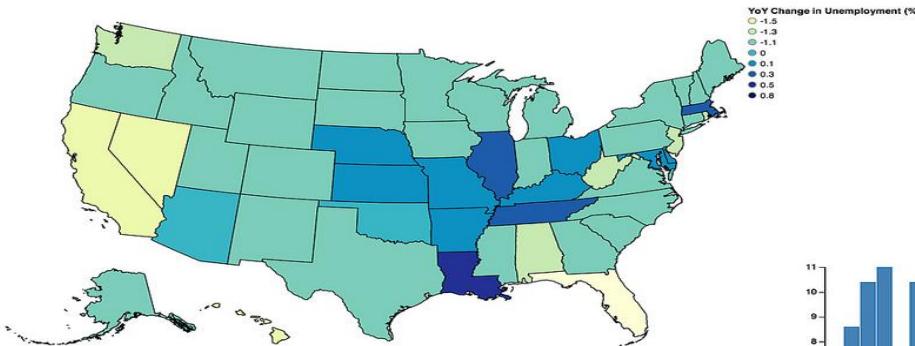


Vega is a *visualization grammar*, a declarative format for creating, saving, and sharing interactive visualization designs.

Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

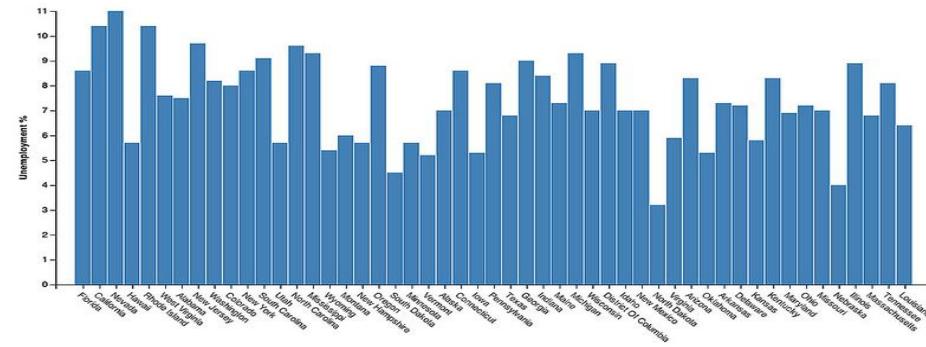
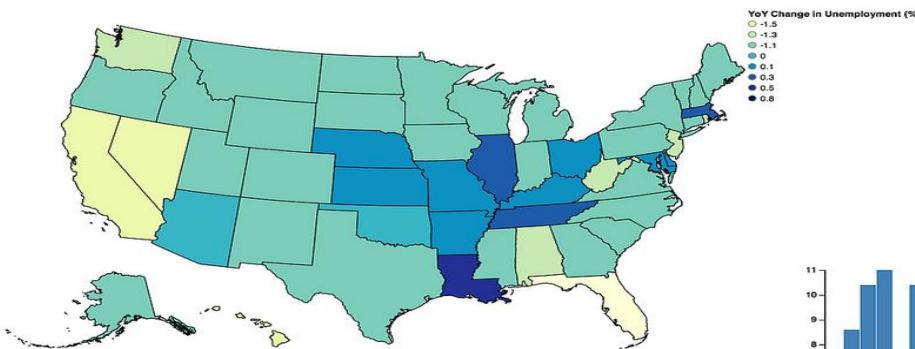
Trivia question: why is it called Vincent? Hint: Vincent+Vega= ?



Interactive toolkits: Vincent

Vincent is a Python-to-Vega translator.

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



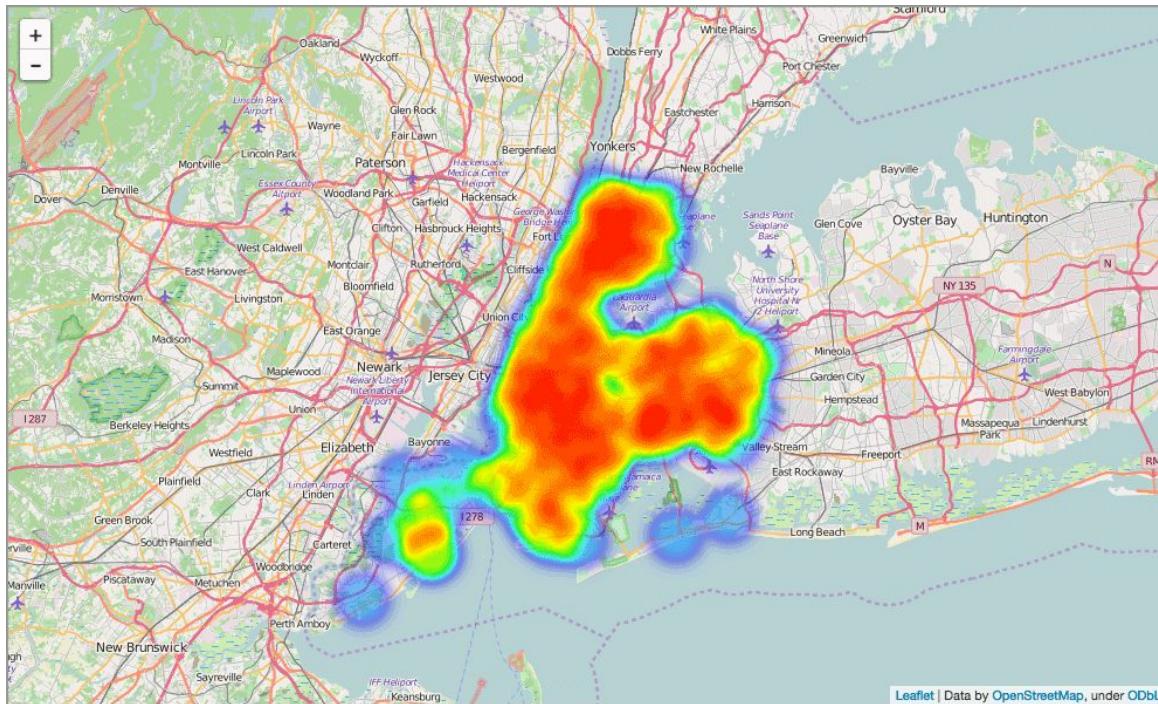
Bokeh: another interactive viz library

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:

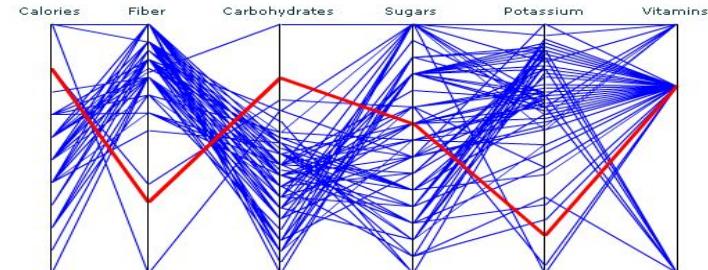
<https://go.epfl.ch/ada2023-lec3-feedback>

- 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

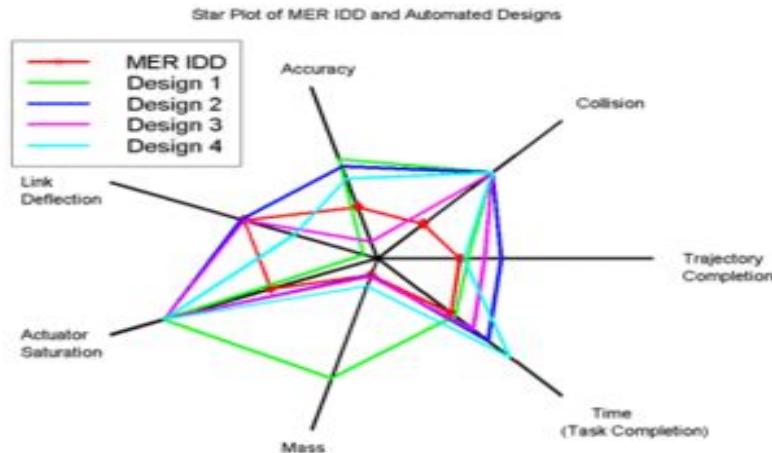
Color, x, y

Color variable is categorical, others arbitrary



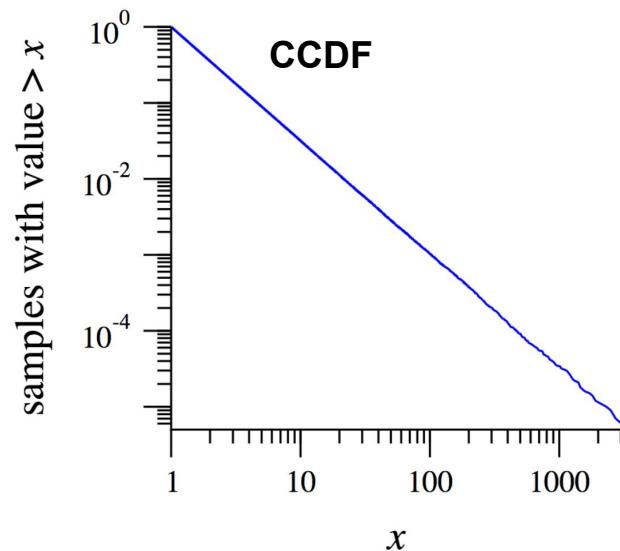
> 2 variables: radar charts

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



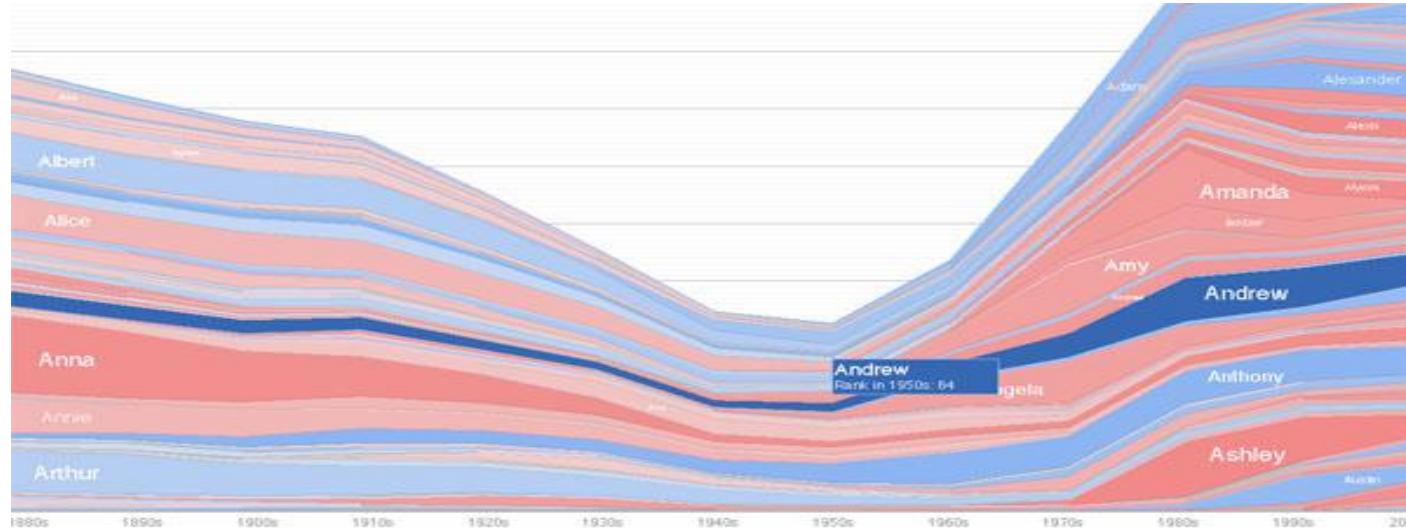
Heavy-tailed data: power laws

- 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)



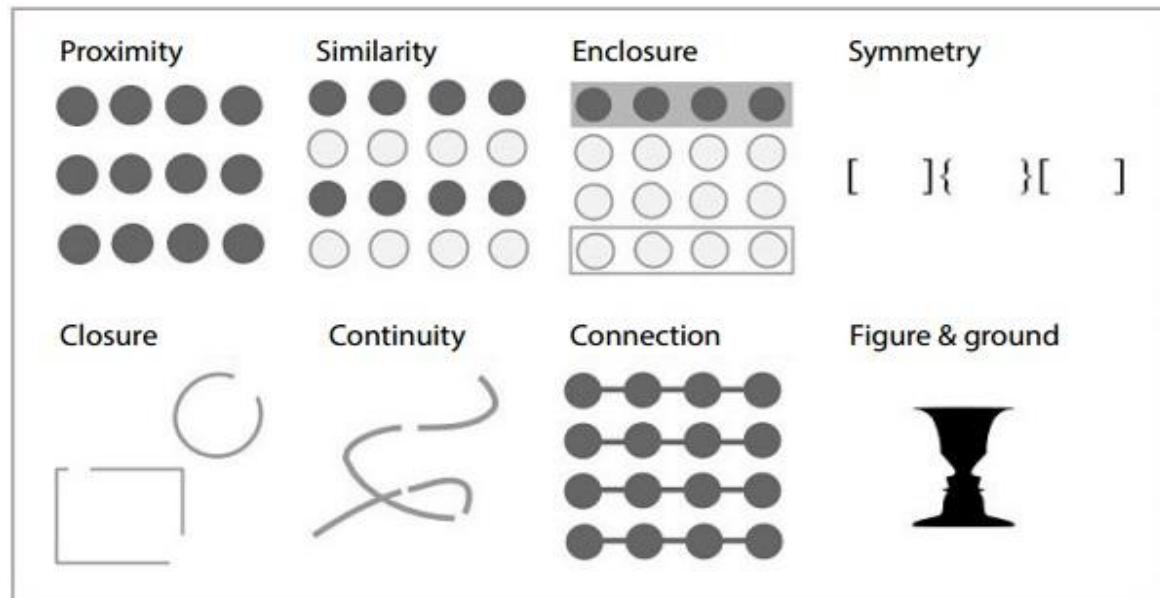
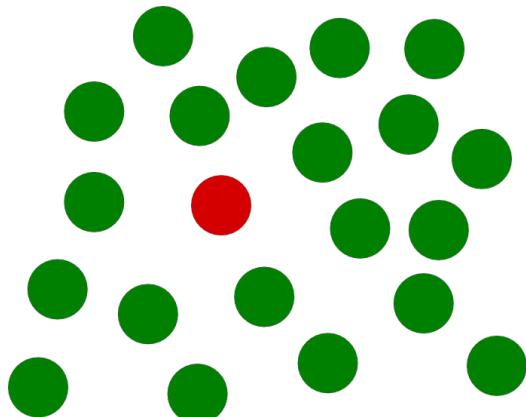
Interactive chart design: simplifying

- 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)

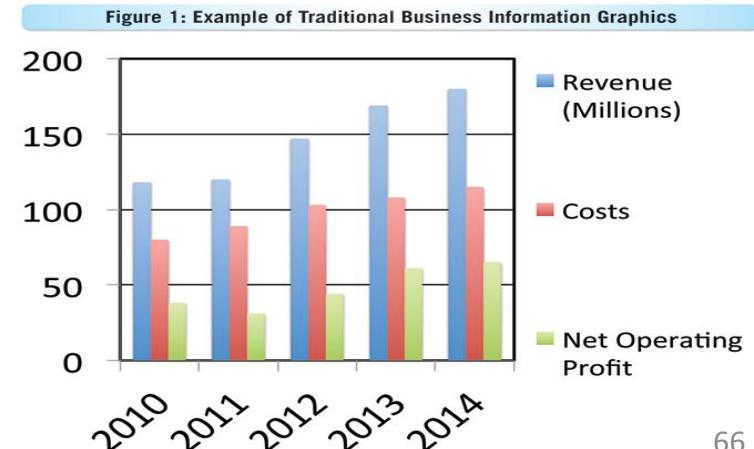


A case for ugly visualizations

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!

