Welcome (Again!) to
MATH 4100/COMP 5360– Introduction to Data Science

(Some) Principles of Data Visualization, cont'd

February 28, 2023

Based on prior lectures from
Alex Lex and Bei Wang Phillips

+ others where noted

Announcements / Reminders / Schedule

HW 6 – Due Friday!

Thursday: Classification, Decision Trees!

Friday: Team formation check—in!

…and then Spring Break!

…and then… project proposal!

Team Formation

Please fill out Google Form by Friday:

https://docs.google.com/forms/d/e/1FAIpQLSc0foXMyjCu3haQHHEVjyN2dBnar9dswjkVLCinh0IuHBnZEA/viewform?usp=sf_link

Really quick! Just a topic + team members, emails, UID, and programming experience.

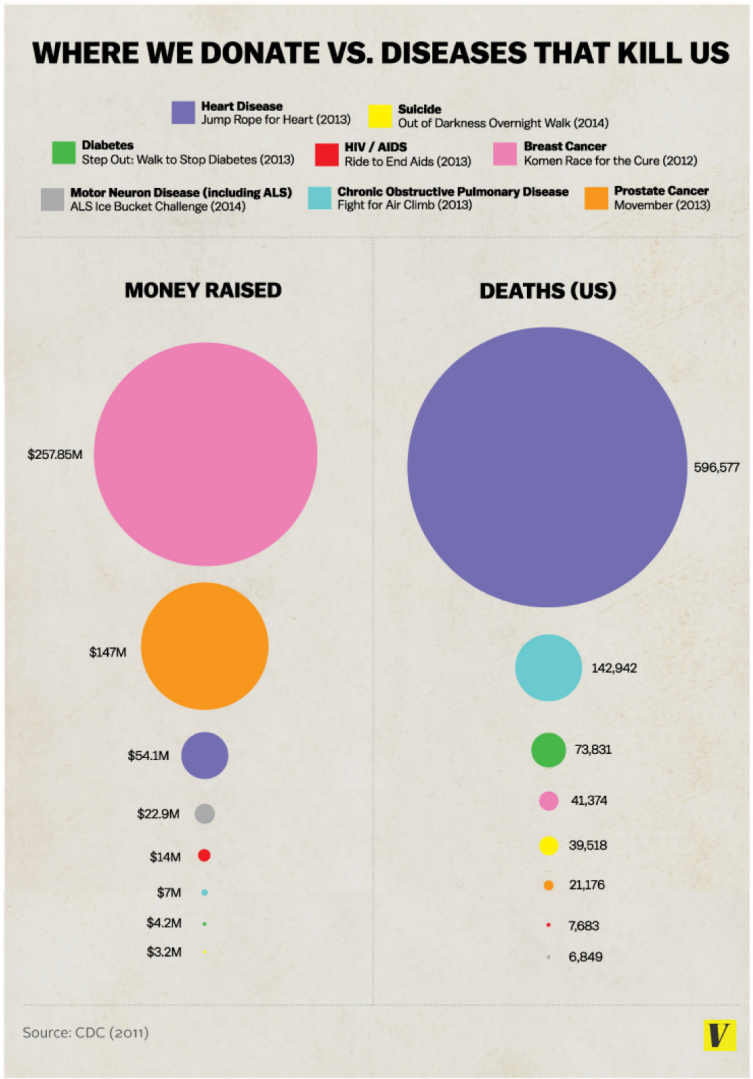
Tuesday Office Hours – Changing after Spring Break

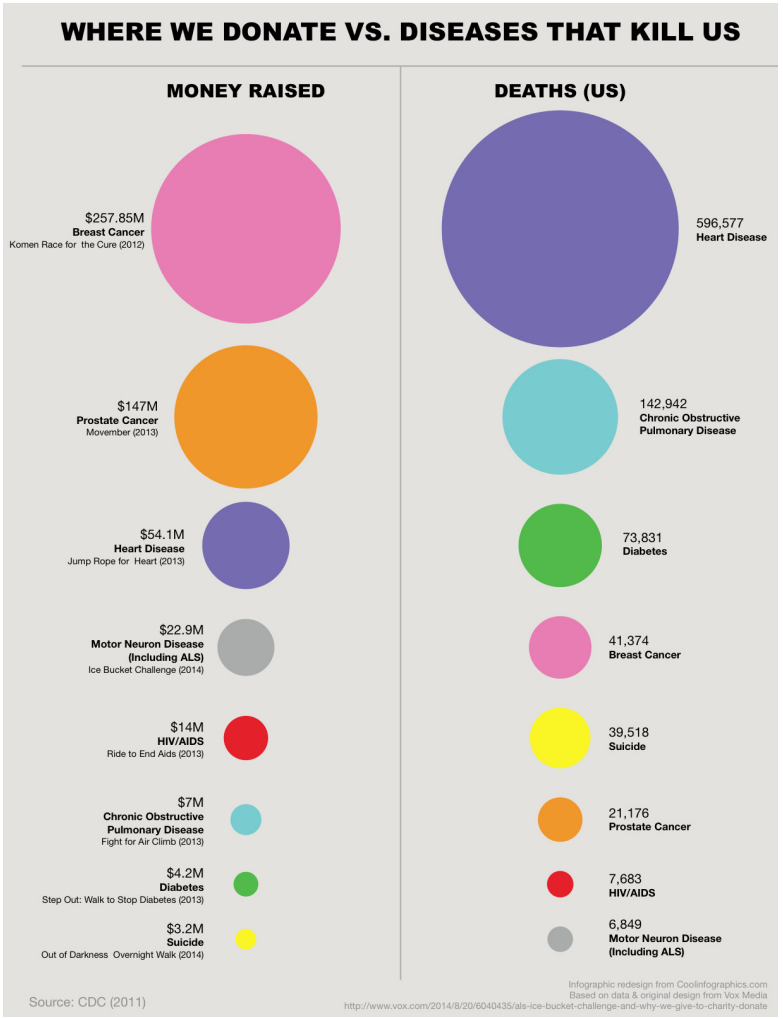
Today 6:30 Over Zoom

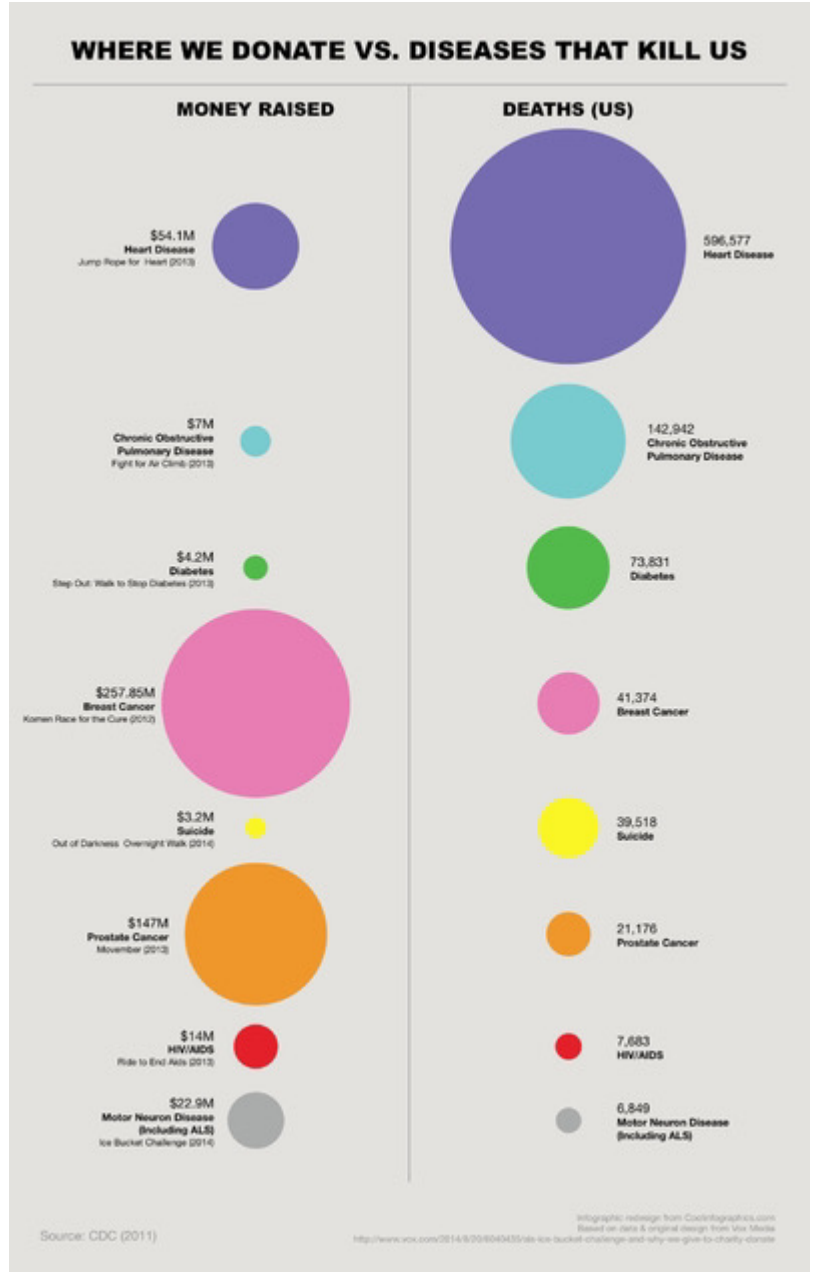
AFTER Spring Break, 5 - 6PM (Right after Class) in-person

- We'll try them here until we get in trouble ☺
- If the classroom doesn't work, WEB 4660

Recall -- Circles: Encode by Area not Radius

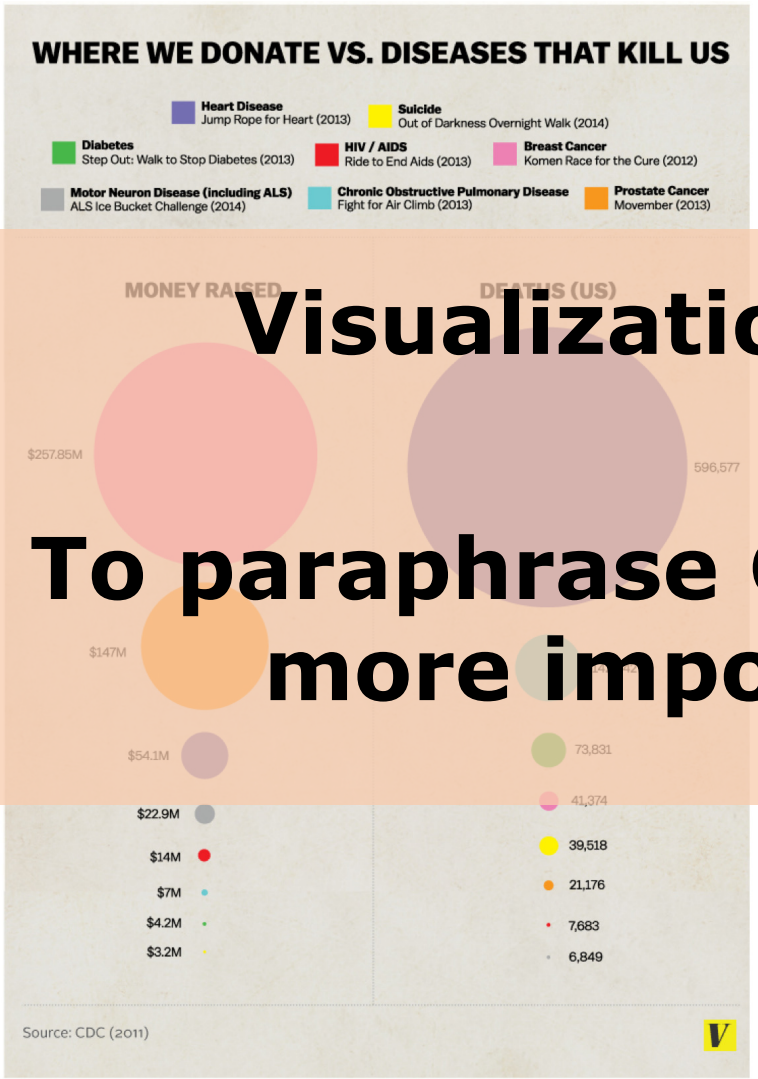


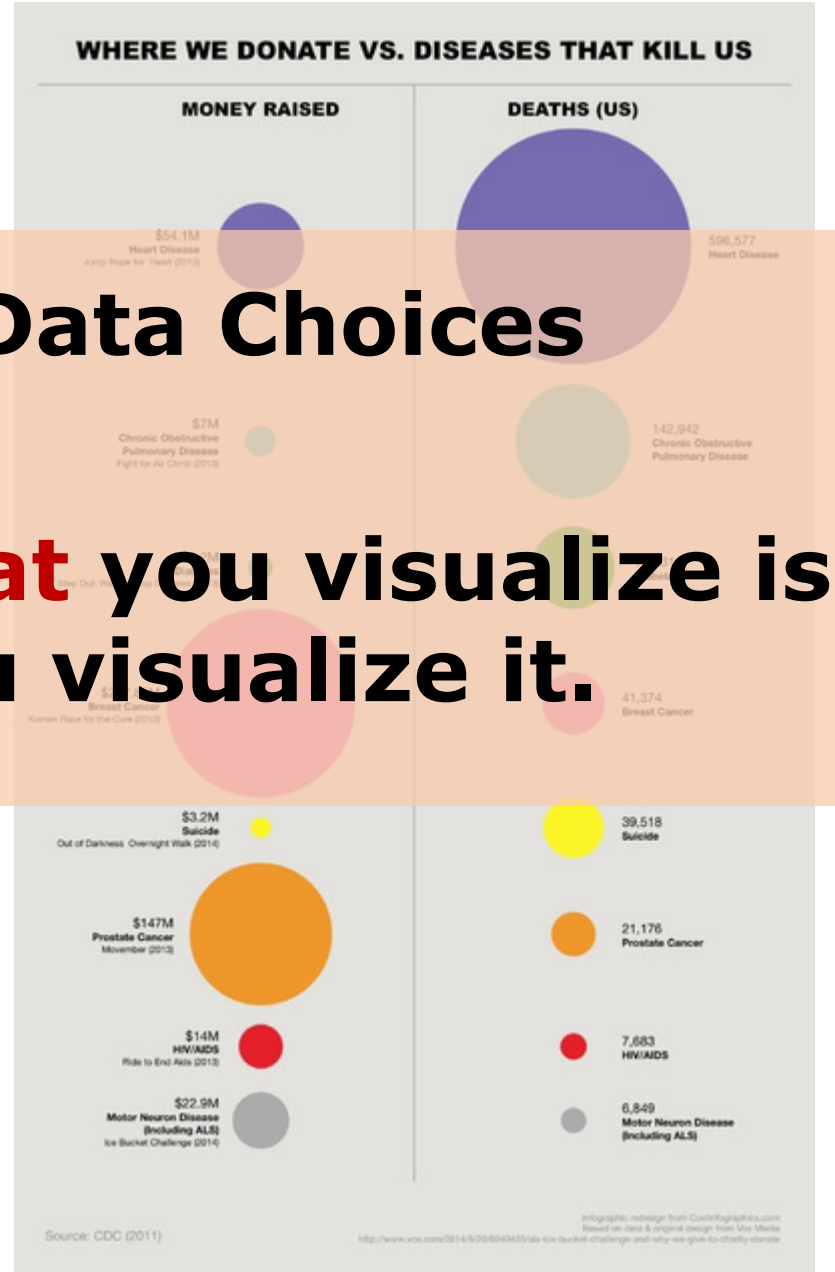




Images from Vox and http://coolinfographics.com/blog/2014/8/29/false-visualizations-sizing-circles-in-infographics.html

Circles: Encode by Area not Radius





Visualization can't overcome Data Choices

To paraphrase George Furnas, what you visualize is more important than how you visualize it.

Images from Vox and http://coolinfographics.com/blog/2014/8/29/false-visualizations-sizing-circles-in-infographics.html

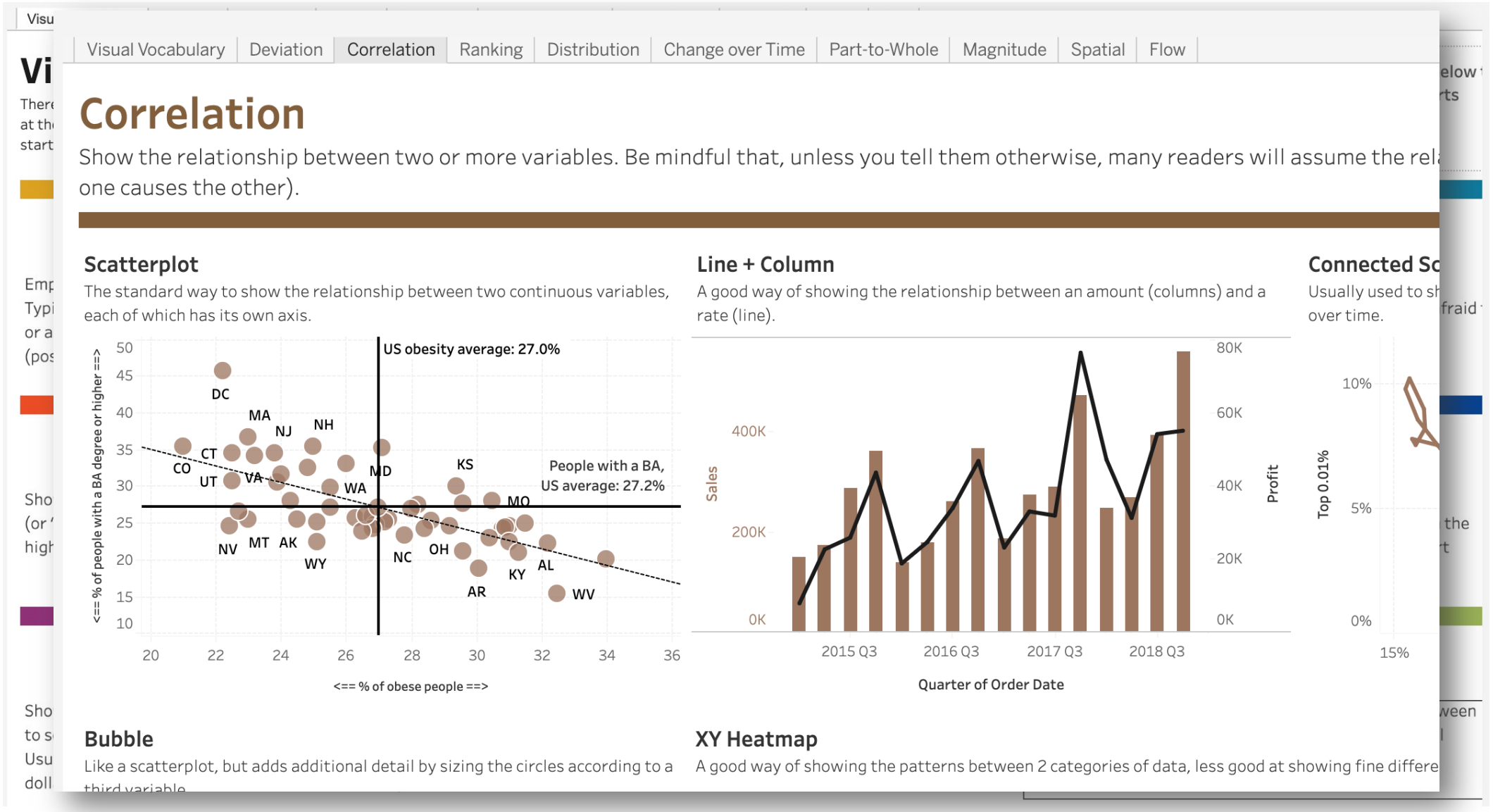
What chart should I use?

What are you trying to understand?

What is the nature of my data?

What is the nature of my task?

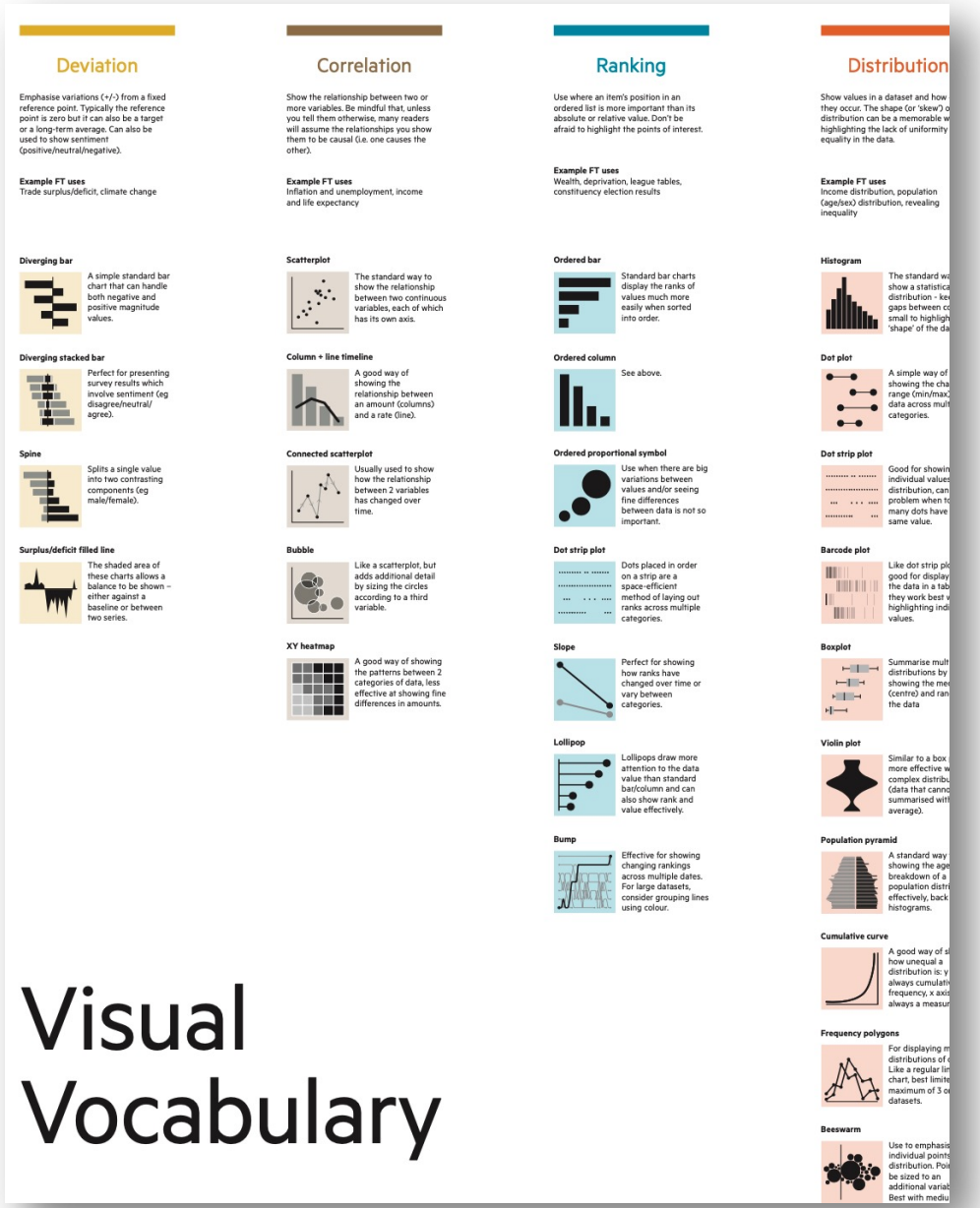
Some sources offer “visual vocabularies” for common abstract goals



https://public.tableau.com/en-us/gallery/visual-vocabulary

dataviscatalogue.com

Financial Times



What if I have tabular data?

…like a DataFrame or CSV

Tasks for tabular data…

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate c

Diverging bar

A simple sta
chart that ca
both negativ
positive mag
values.

Diverging stacked bar

Perfect for p
survey resul
involve sent
disagree/neu

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables,

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses

ts, economic time
es in a market

e standard way to
w a changing time
ies. If data are
gular, consider
rkers to represent
a points.

lumns work well for
owing change over
e - but usually best
h only one series of

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as

Spatial

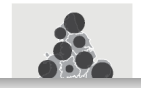
Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)



Use for totals rather than rates – be wary that small differences in data will be hard to

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information; relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically

https://github.com/Financial-Times/chart-doctor/tree/main/visual-vocabulary

Histograms

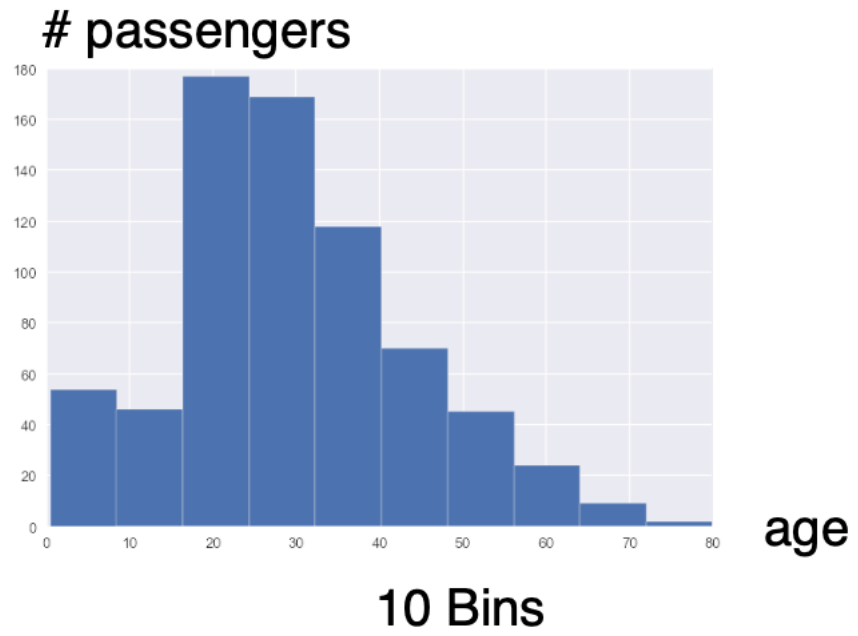
Hard to predict how many bins to use.

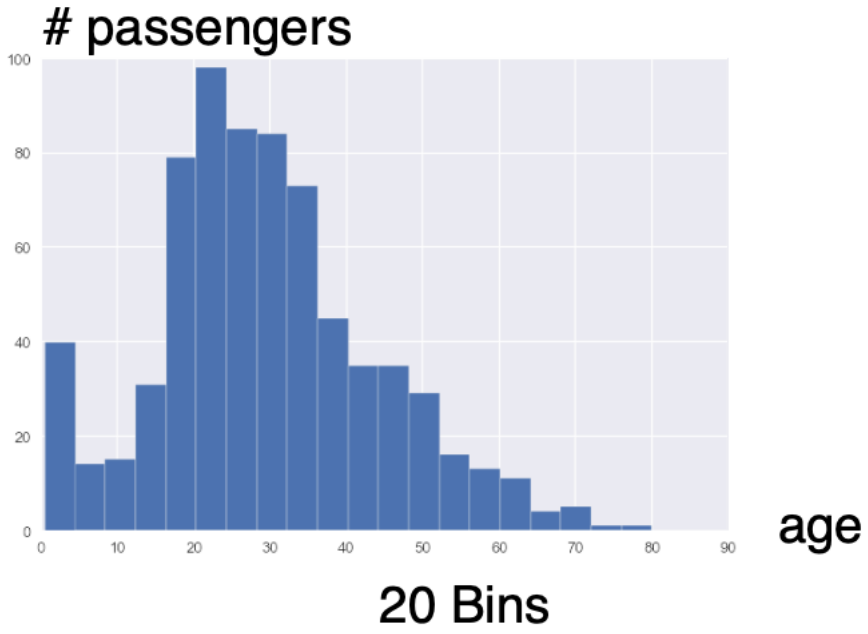
…possibly solve with interactivity!

Some rules of thumb:

#bins = sqrt(n)

#bins = log2(n) + 1





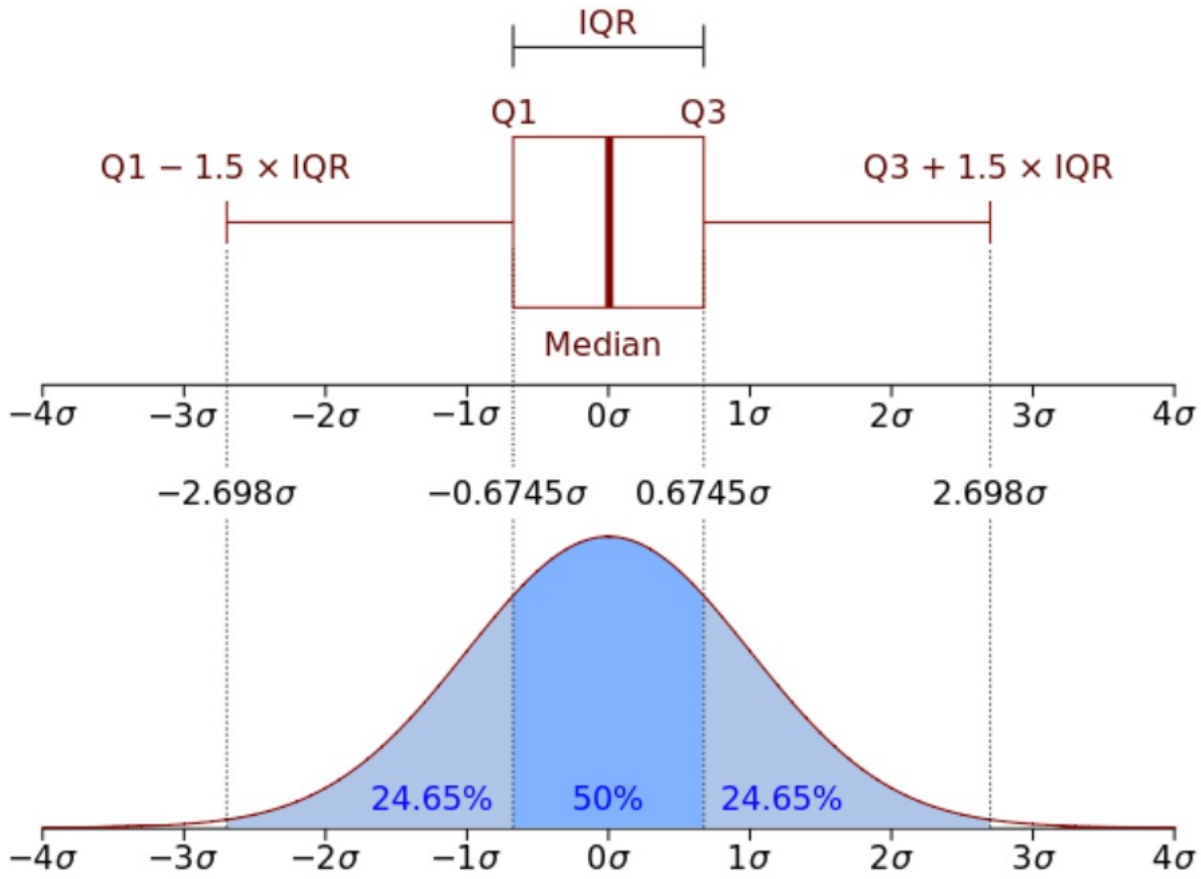
Box plots

a.k.a. “Box-and-whisker plots”

Helpful to show outliers as points

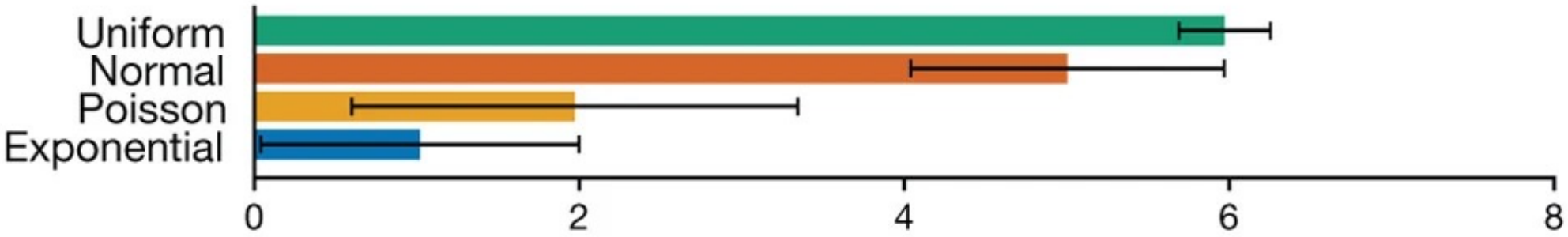
Less useful for data not following normal distribution

Especially bad for bimodal or multi-modal distributions.



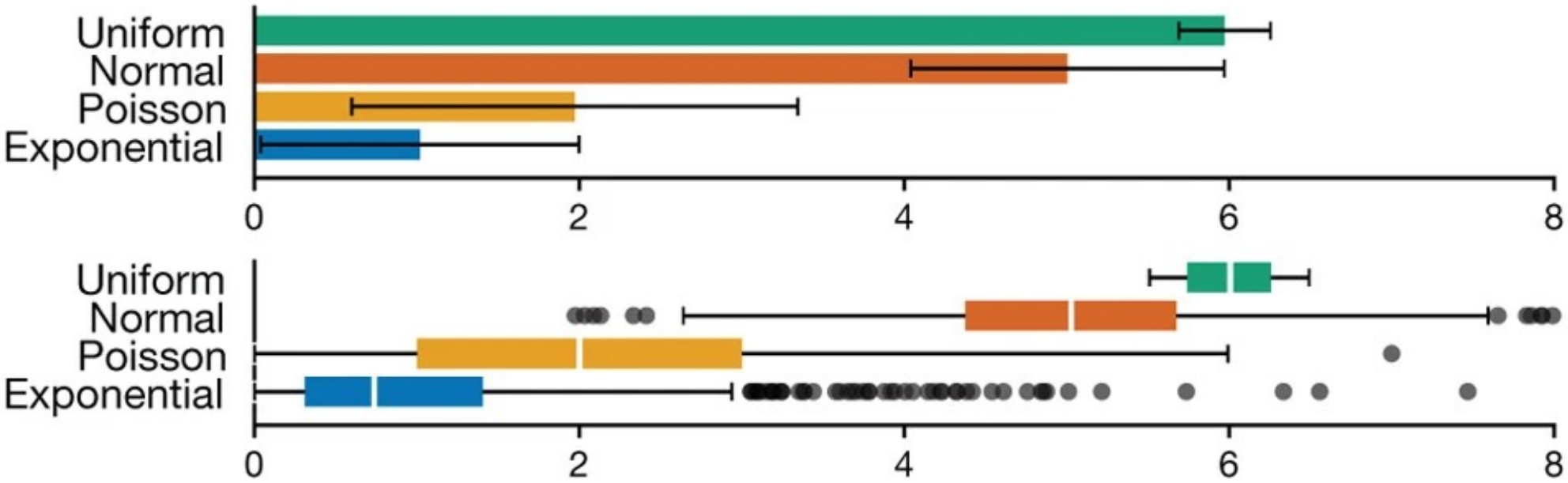
Wikipedia

Error bars vs. Boxplots on some test datasets



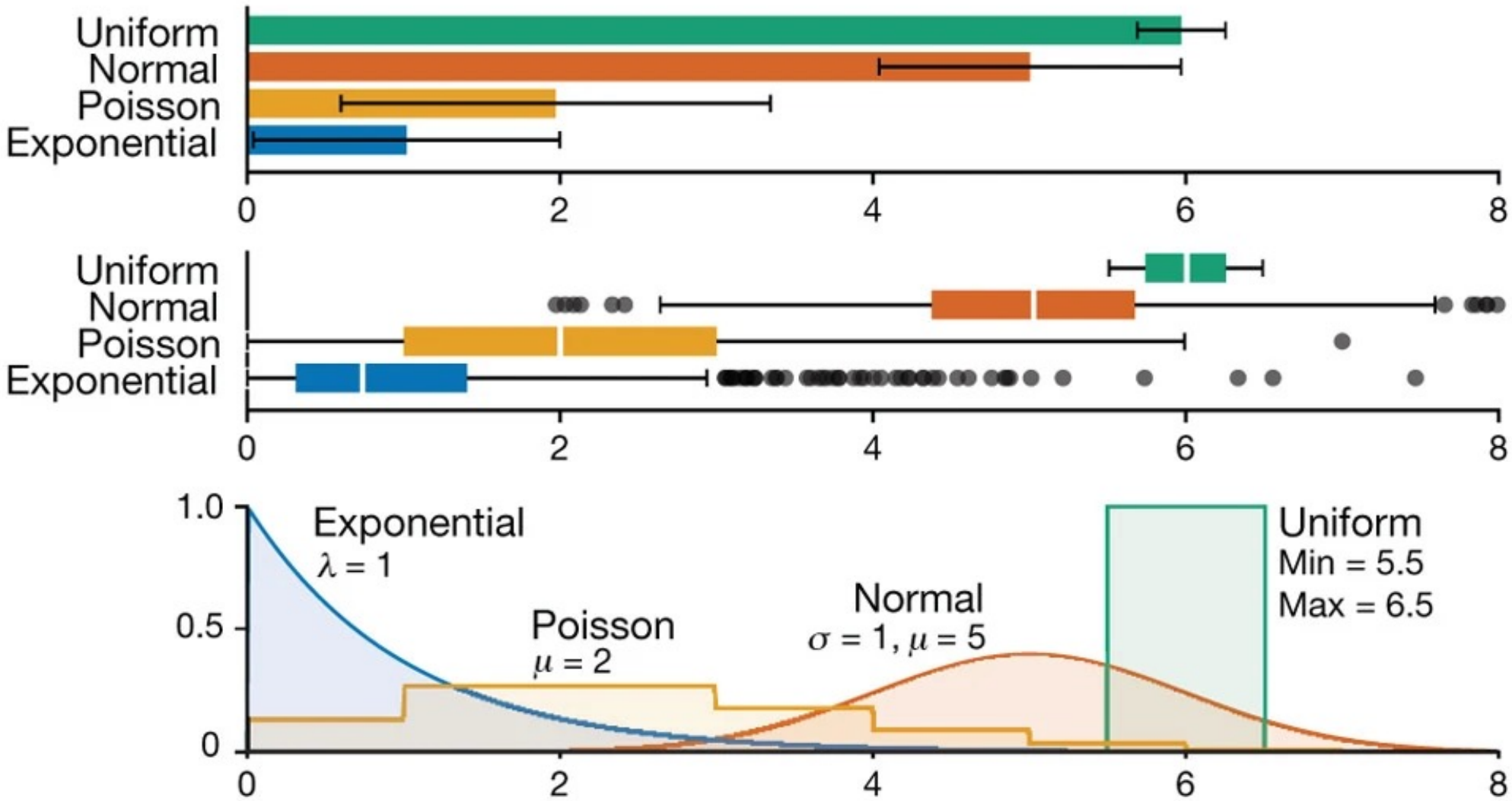
Streit & Gehlenborg, https://www.nature.com/articles/nmeth.2807

Error bars vs. Boxplots on some test datasets



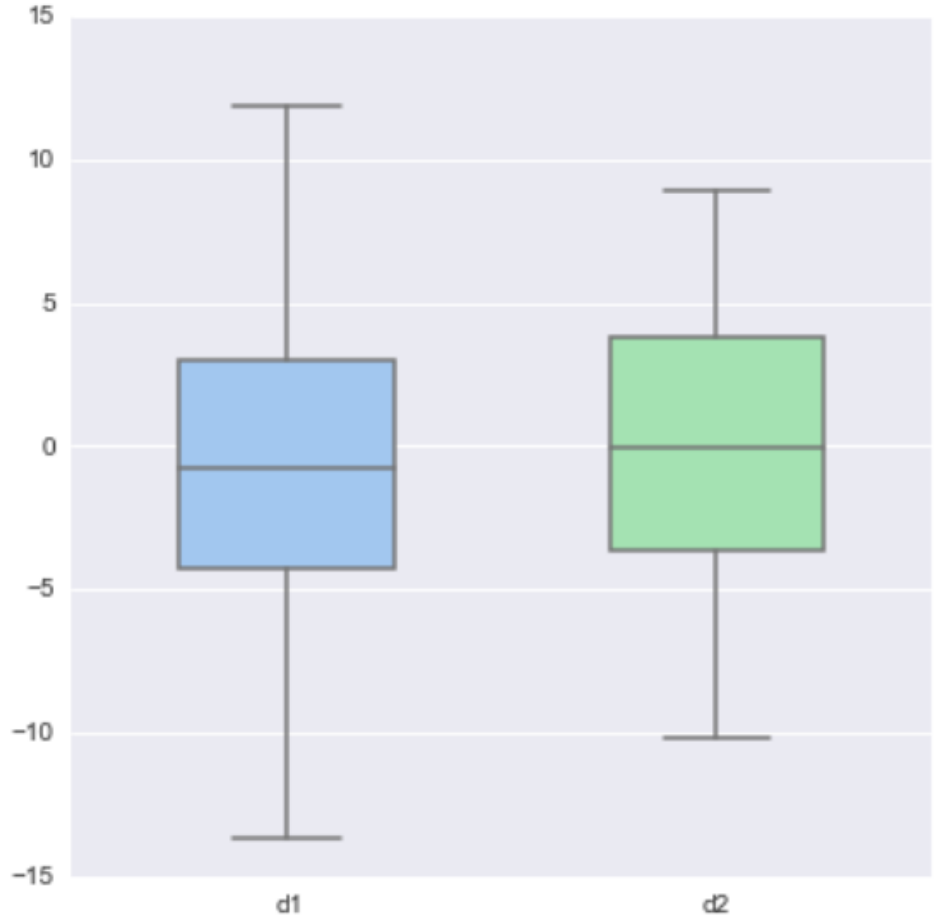
Streit & Gehlenborg, https://www.nature.com/articles/nmeth.2807

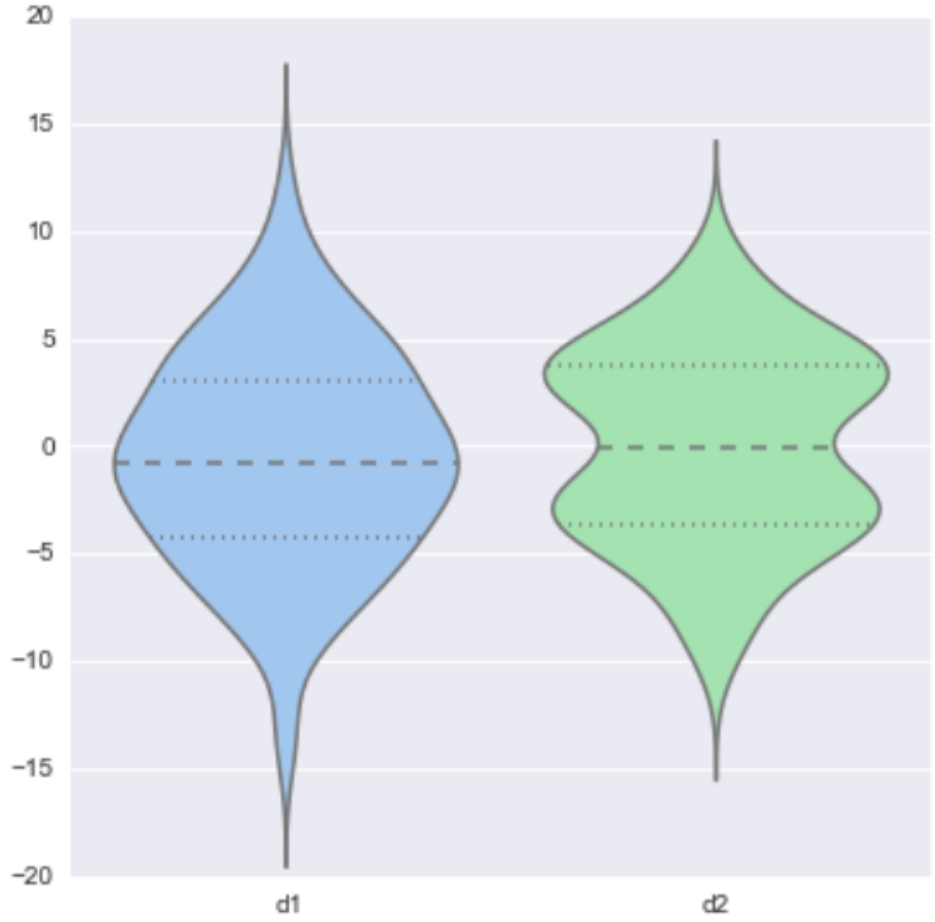
Error bars vs. Boxplots on some test datasets



Streit & Gehlenborg, https://www.nature.com/articles/nmeth.2807

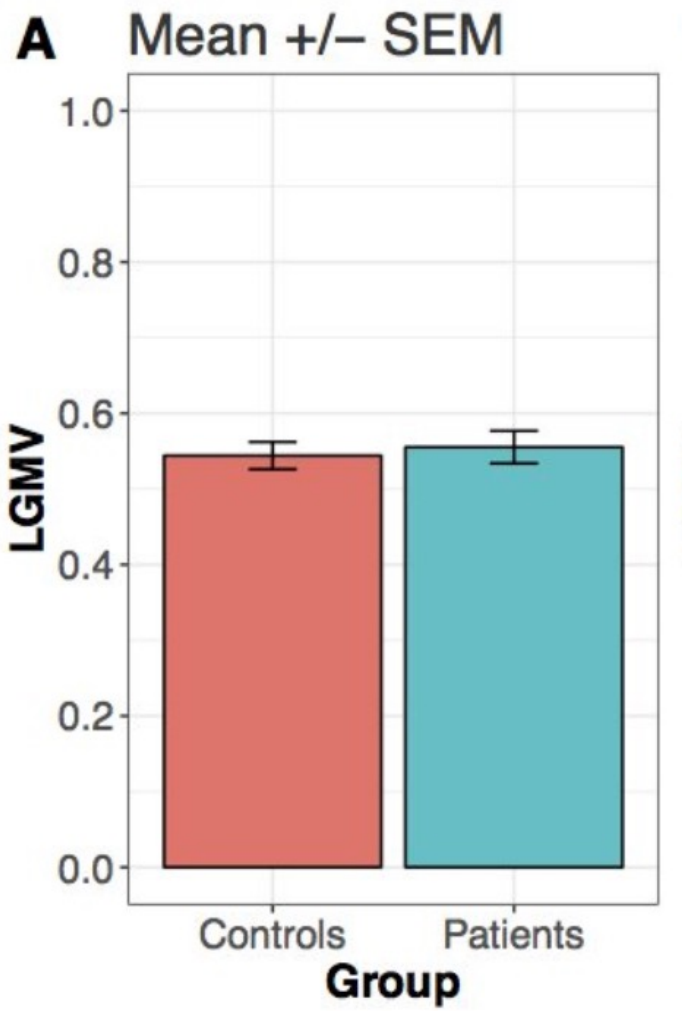
Violin plots are box plots with probability density functions





http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

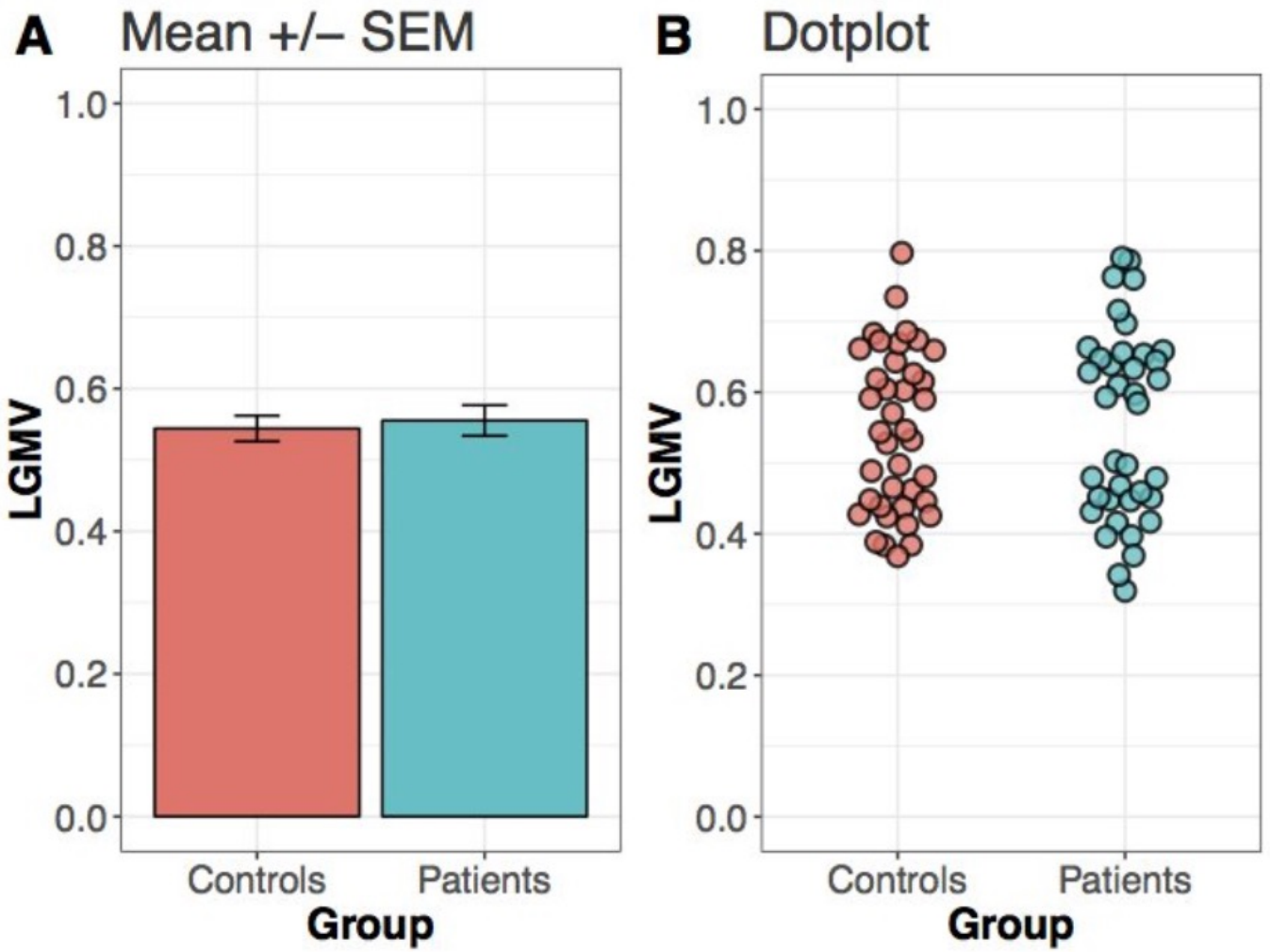
Box plots vs. Dot plots



Data Source https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-10-67

https://twitter.com/robustgar/status/859318971920769024

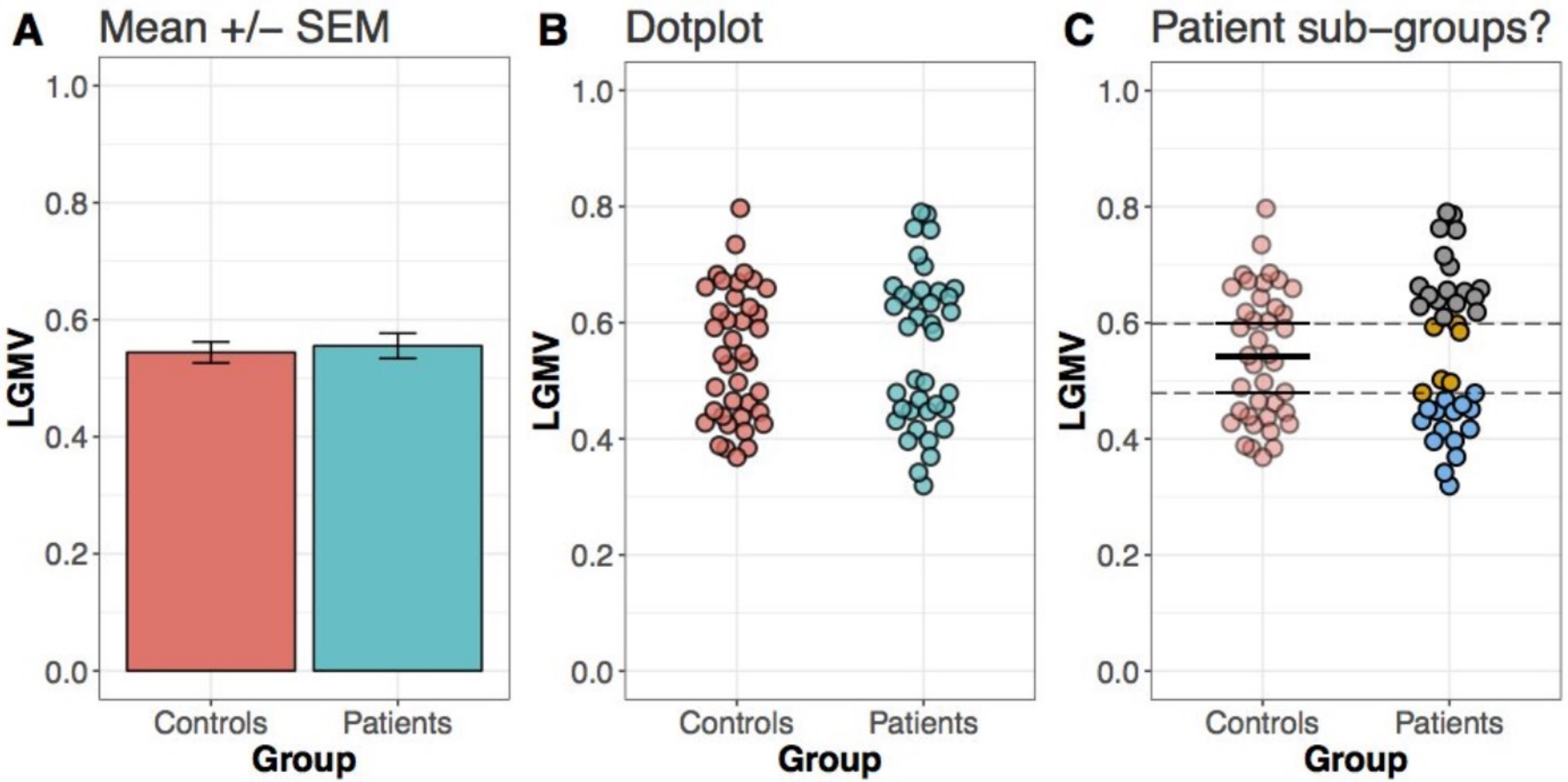
Box plots vs. Dot plots



Data Source https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-10-67

https://twitter.com/robustgar/status/859318971920769024

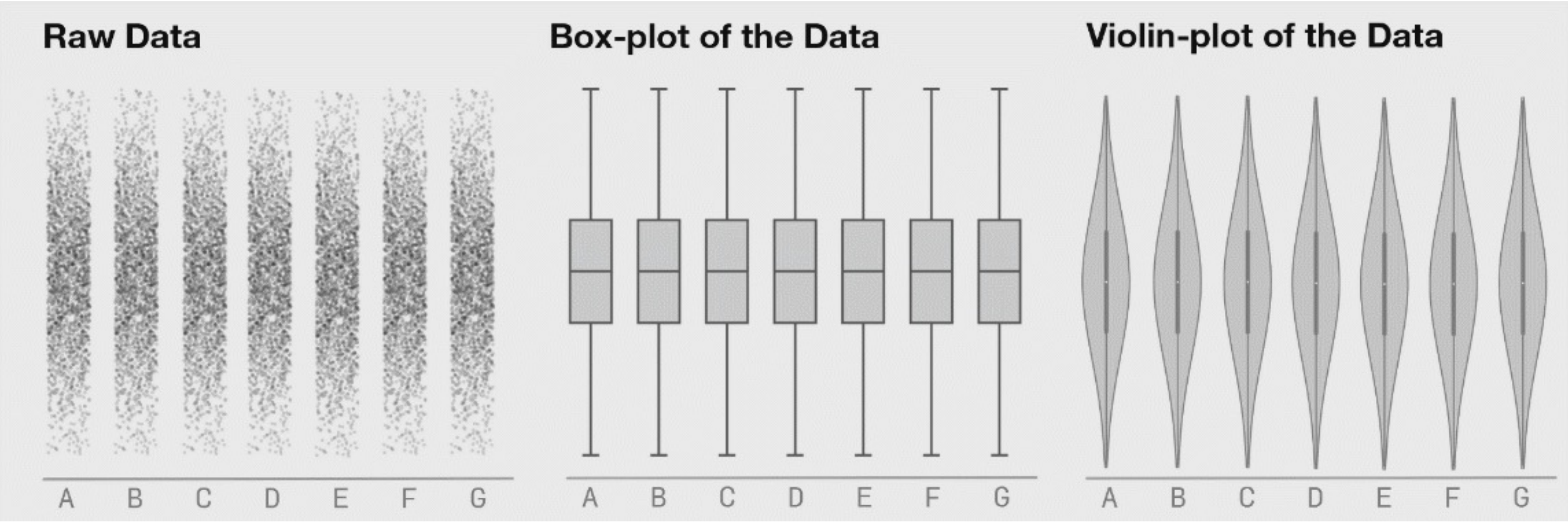
Box plots vs. Dot plots



Data Source https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-10-67

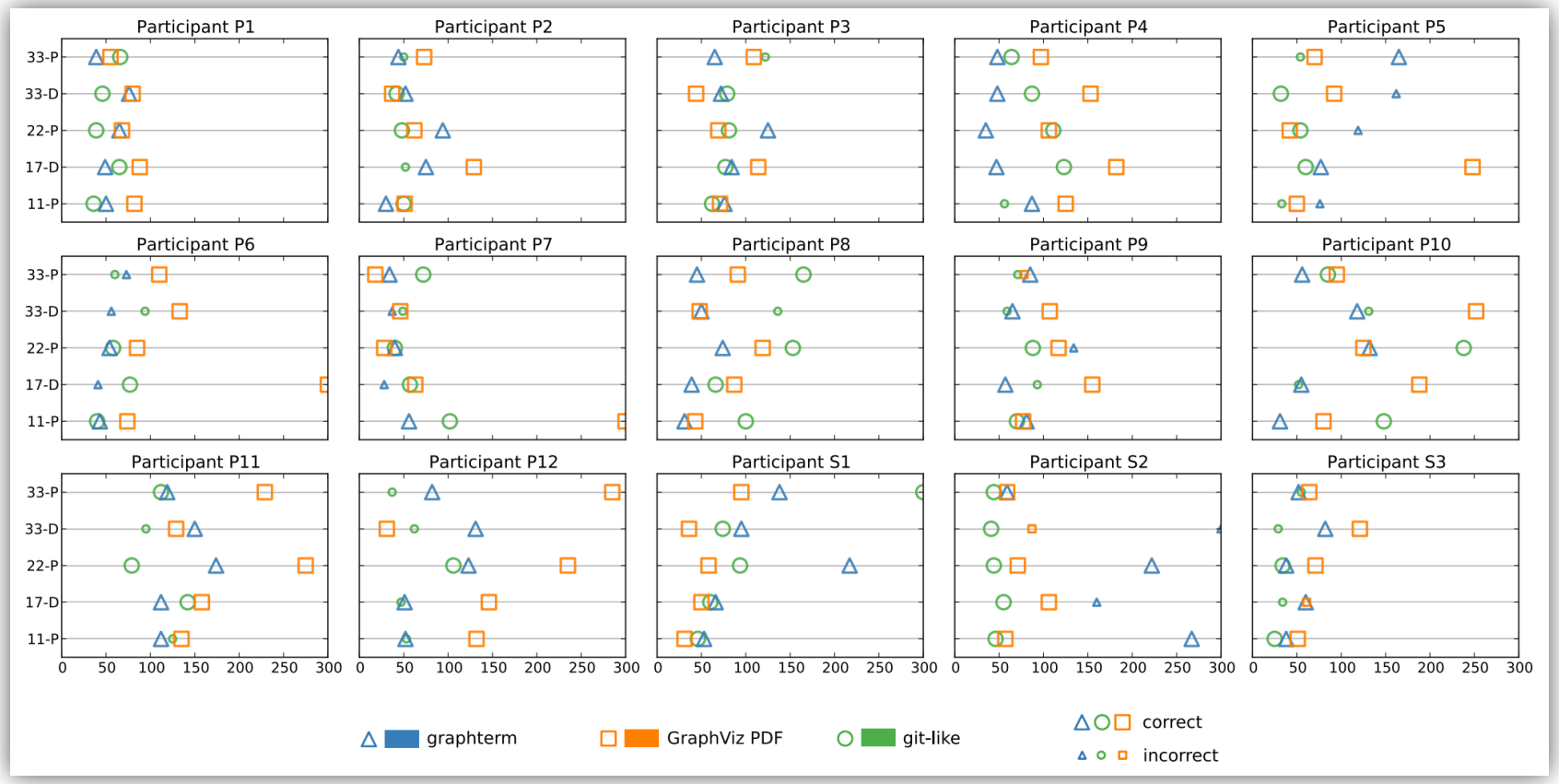
https://twitter.com/robustgar/status/859318971920769024

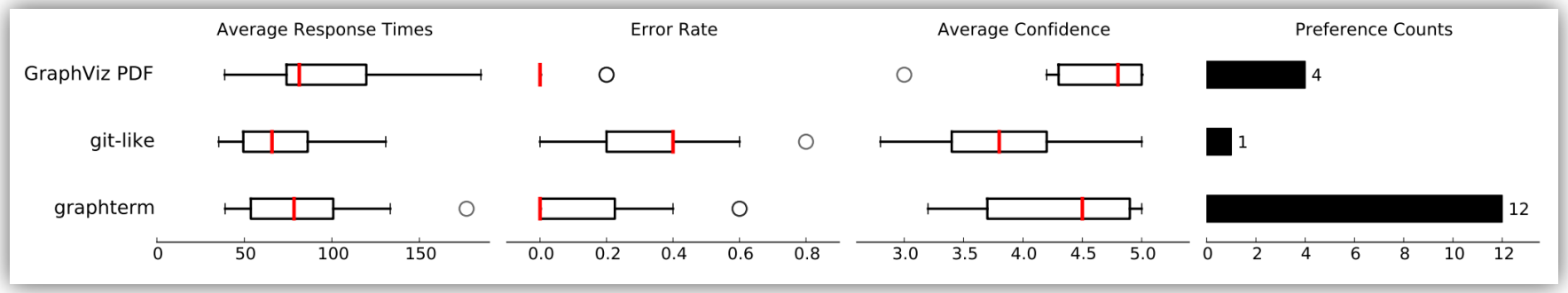
Different distributions, same box plot



https://blog.bioturing.com/2018/05/16/5-reasons-you-should-use-a-violin-graph/

When possible, show all the data

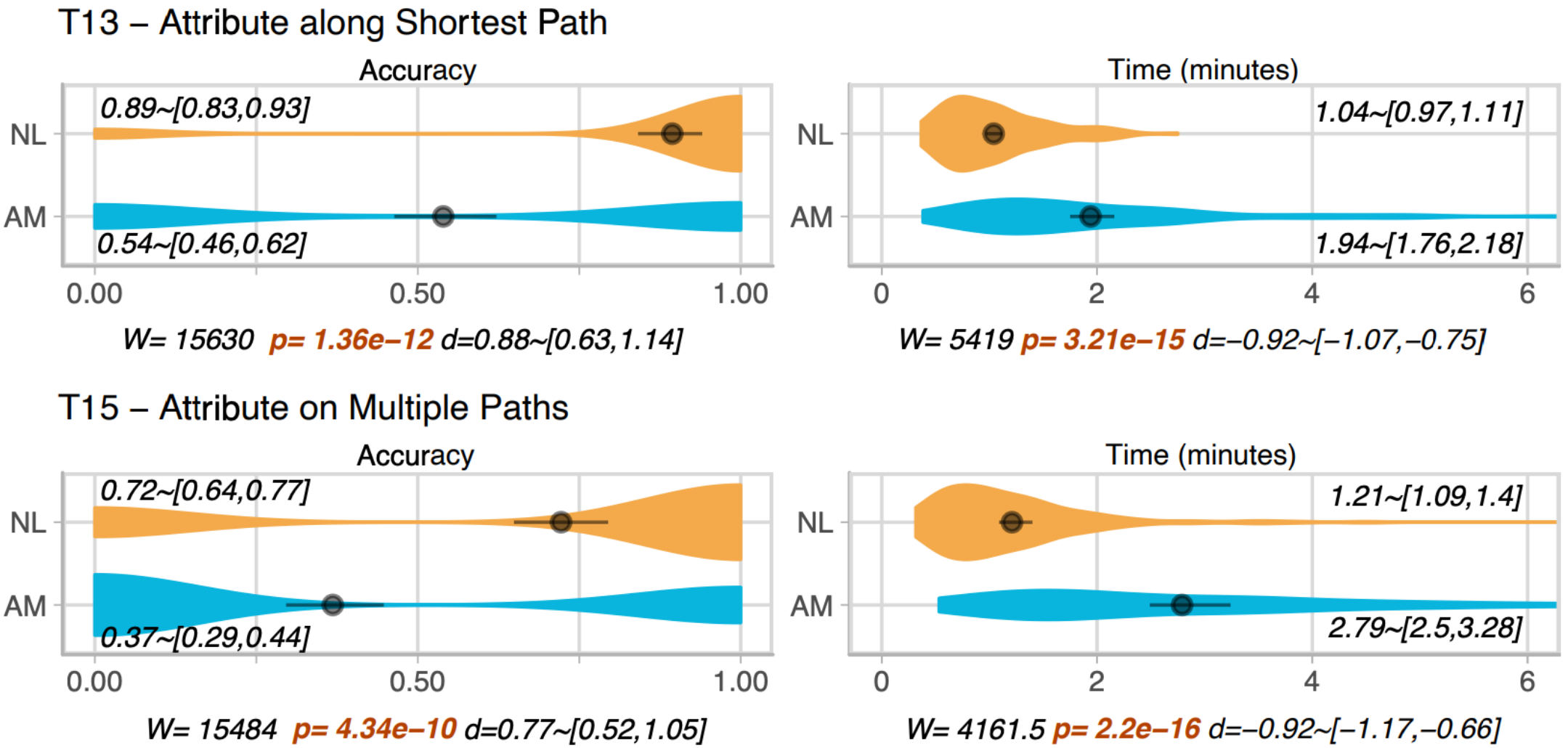




Can pair with summary

Isaacs & Gamblin, 2019

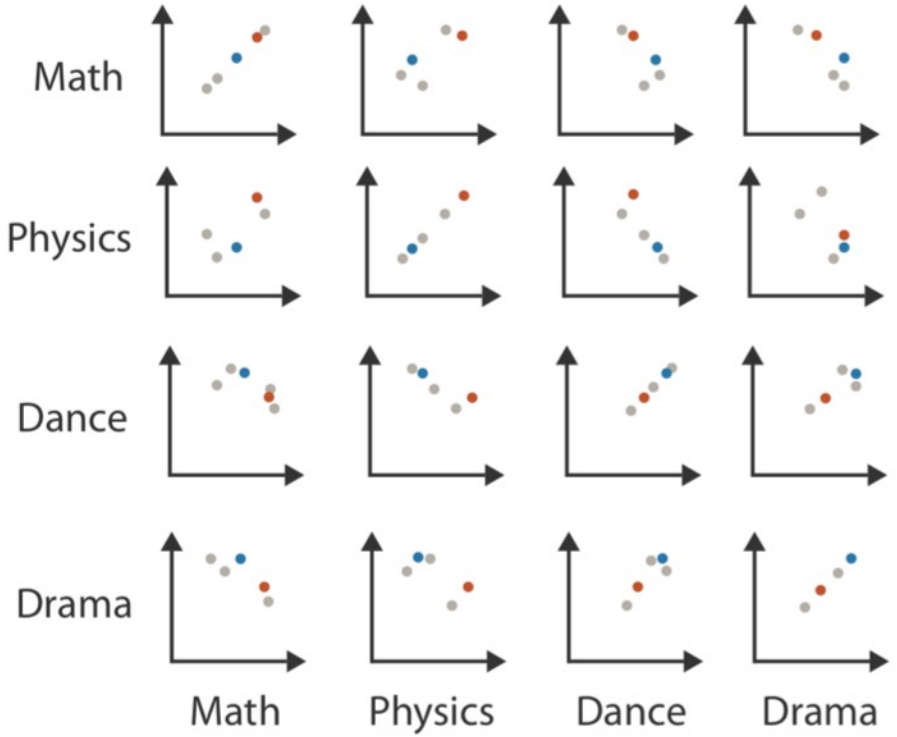
It’s not always possible to show all the data, so show confidence interval in context



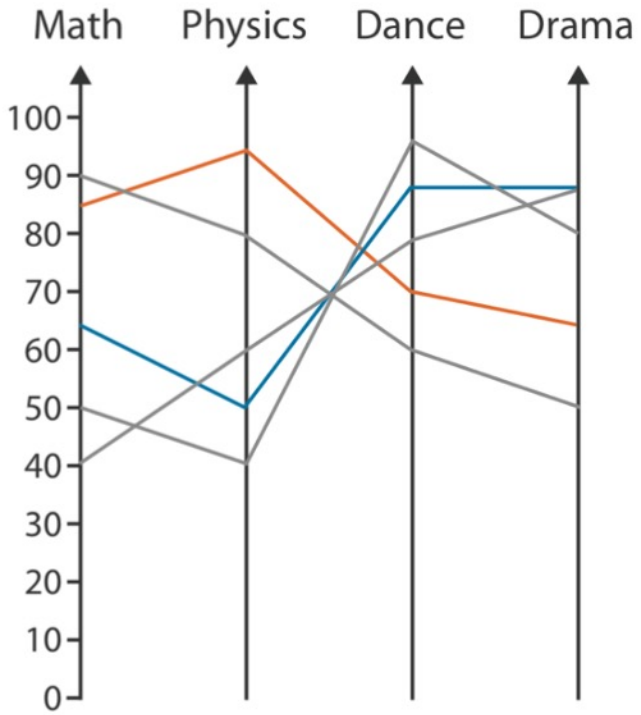
Nobre et al. Evaluating Multivariate Network Visualization Techniques Using a Validated Design and Crowdsourcing Approach

Compare many attributes…

Scatterplot Matrix



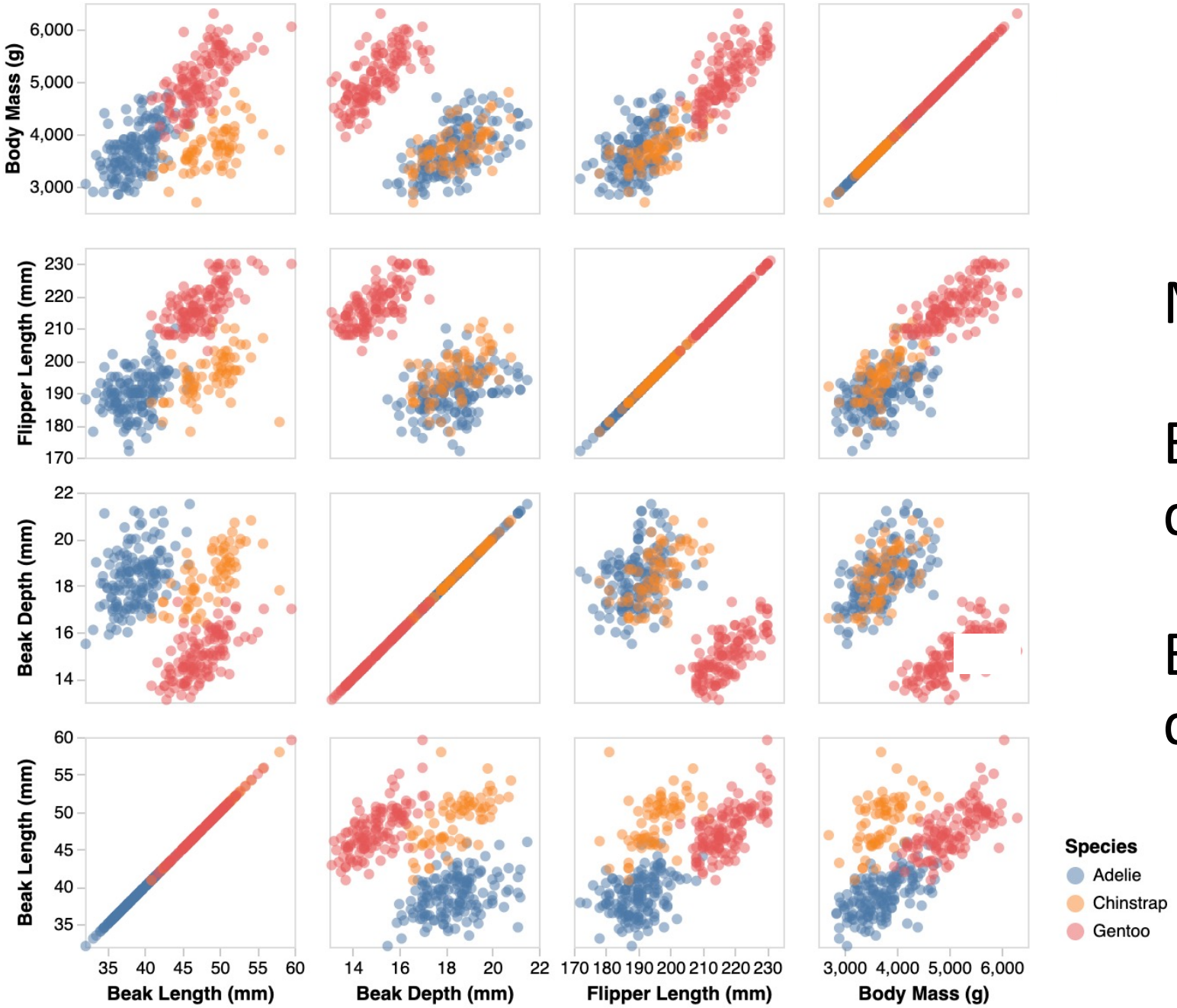
Parallel Coordinates



Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatterplot Matrix (SPLOM)



Matrix of size d x d

Each row/column is one dimension

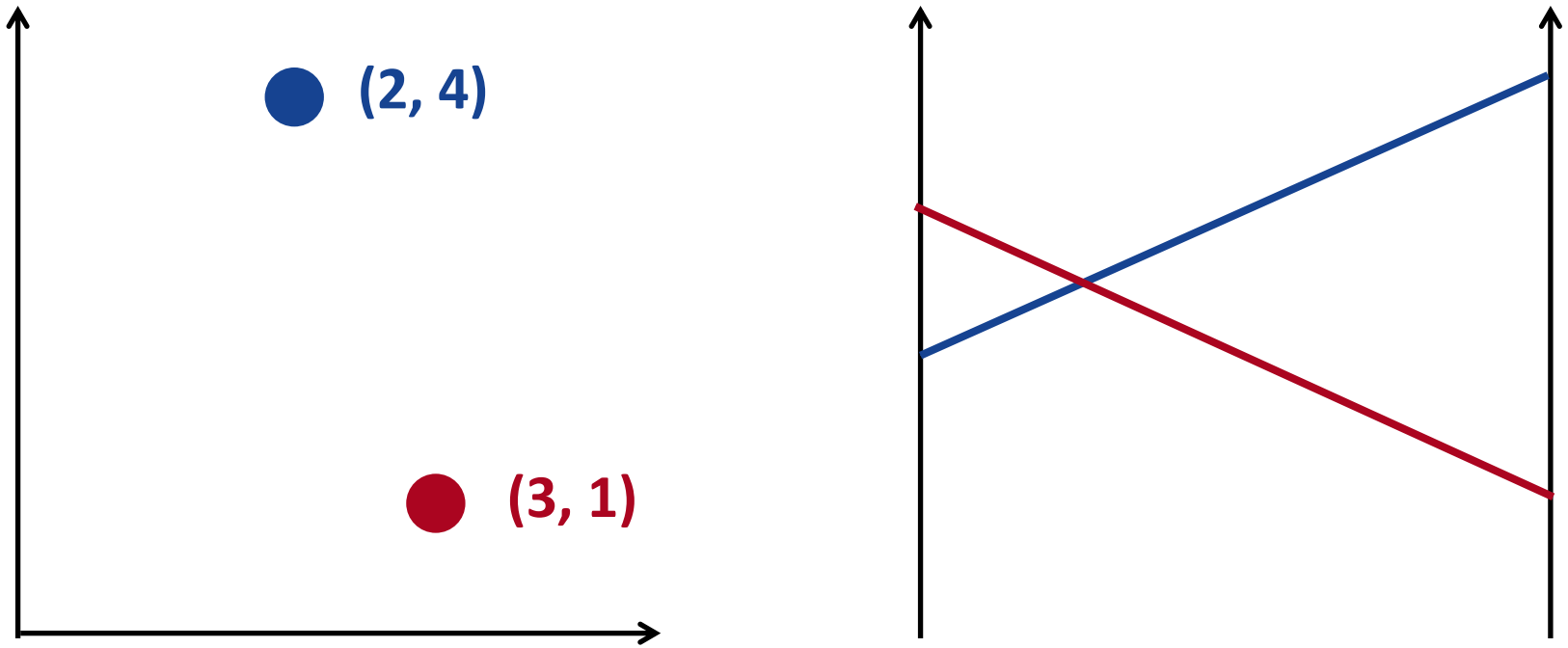
Each plot is two dimensions

https://vega.github.io/vega/examples/brushing-scatter-plots/

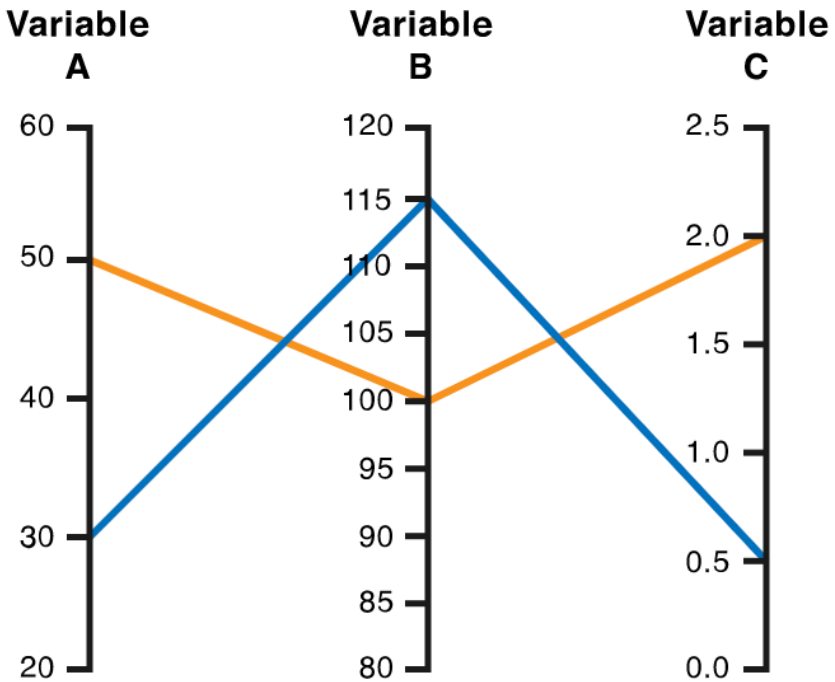
Parallel coordinates

Each dimension gets one parallel axis

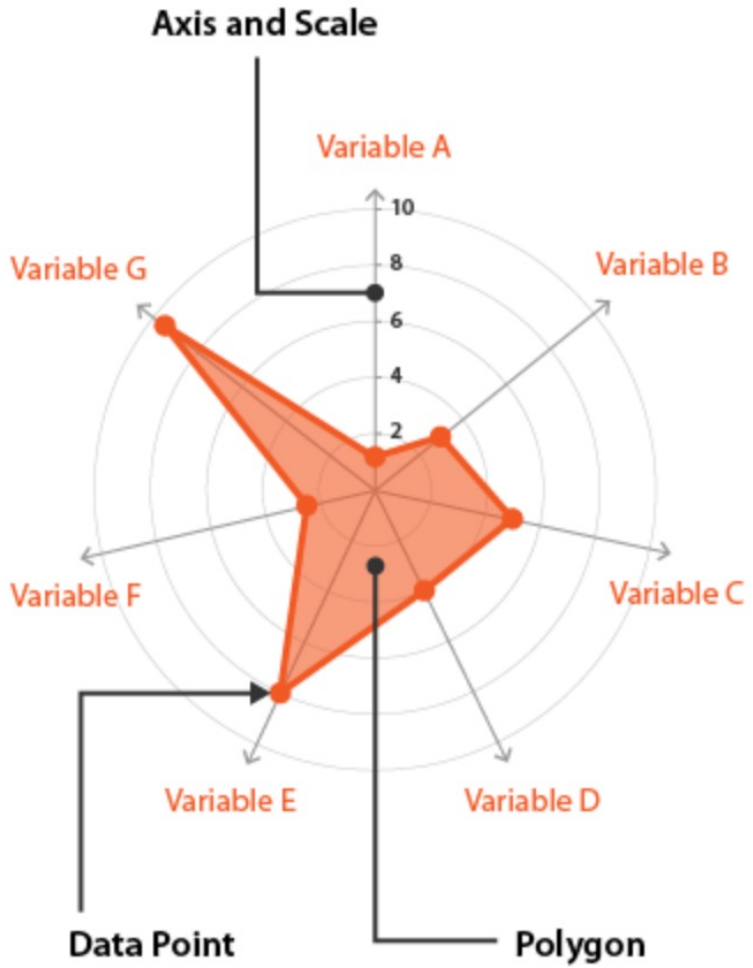
Items in dataset represented with a (poly-)line



Parallel coordinate plots are similar to Radar/Star charts but have common orientation.

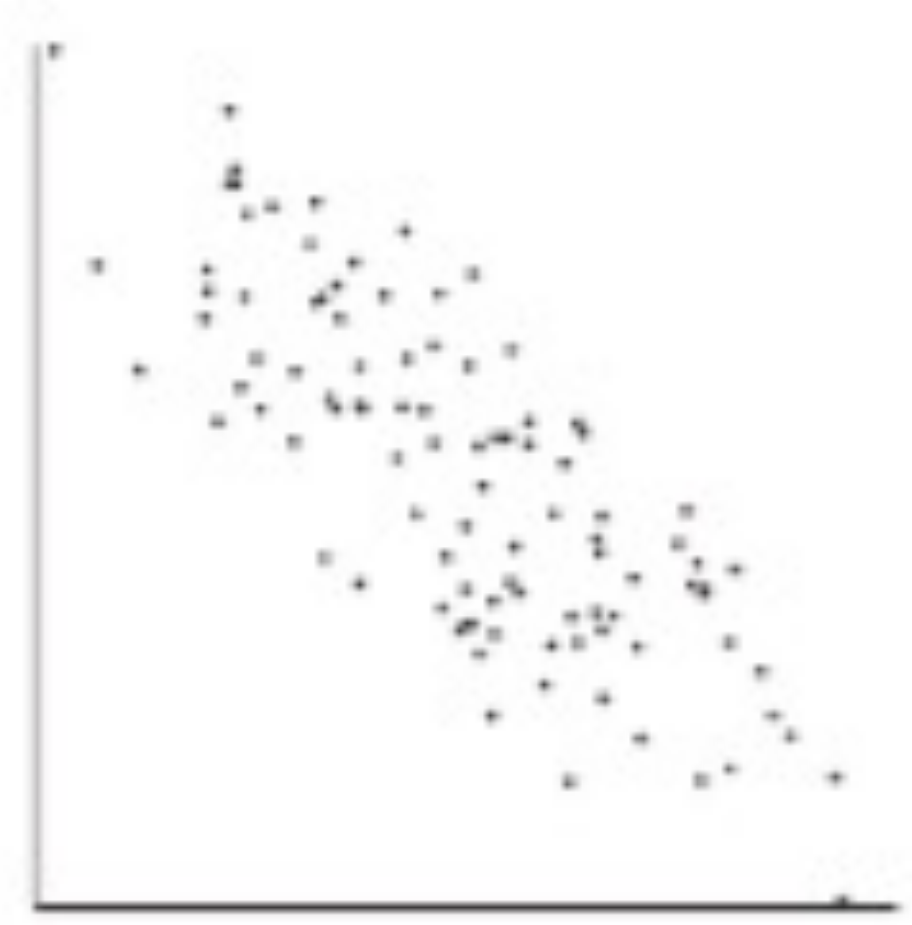


Data			
	Variable A	Variable B	Variable C
Item 1	50	100	2.0
Item 2	30	115	0.5



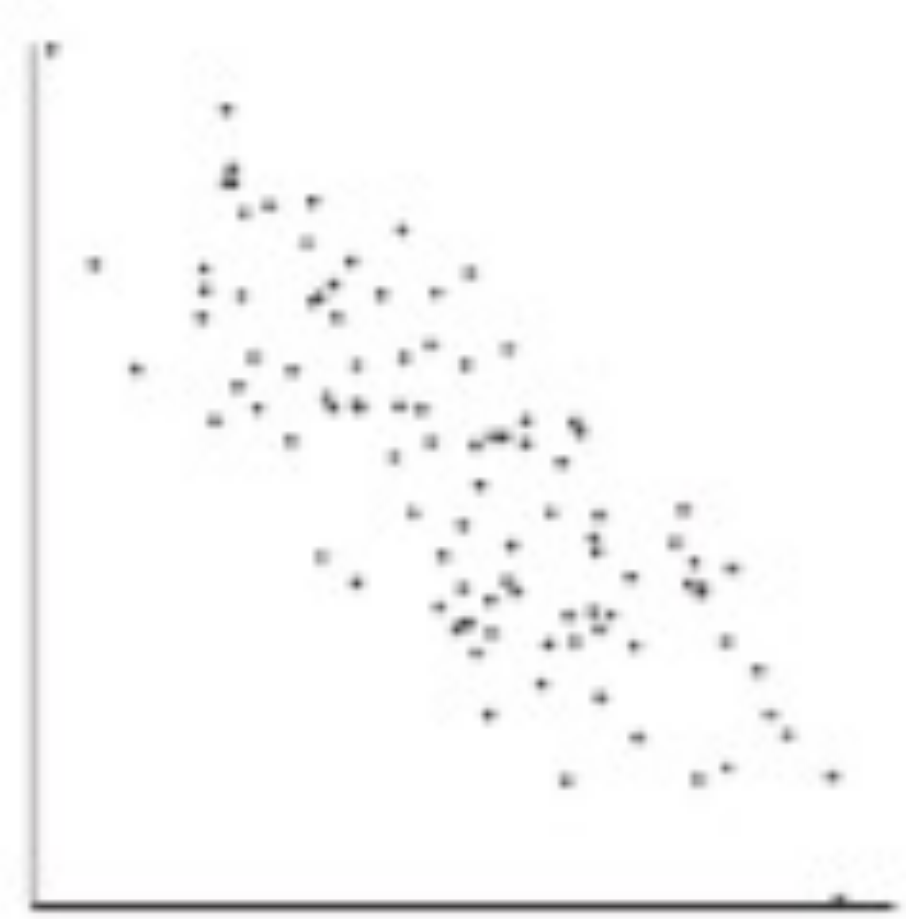
Images from datavizcatalogue.com

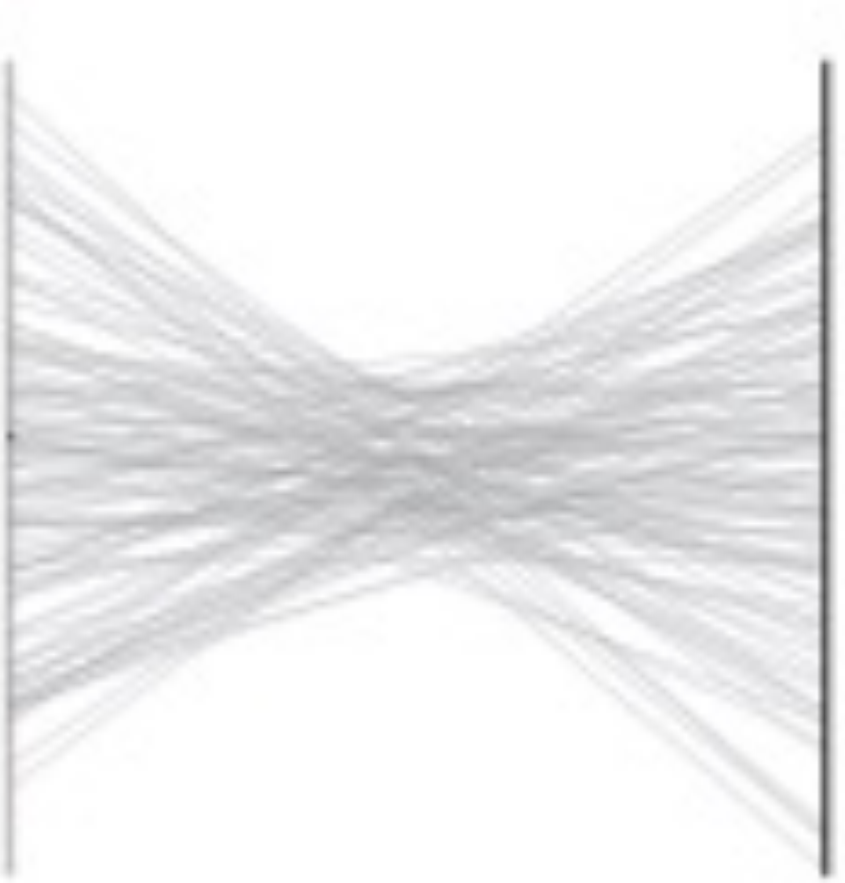
Scatterplots versus Parallel Coordinates



Harrison et al. Ranking Visualizations of Correlation Using Weber's Law 2014

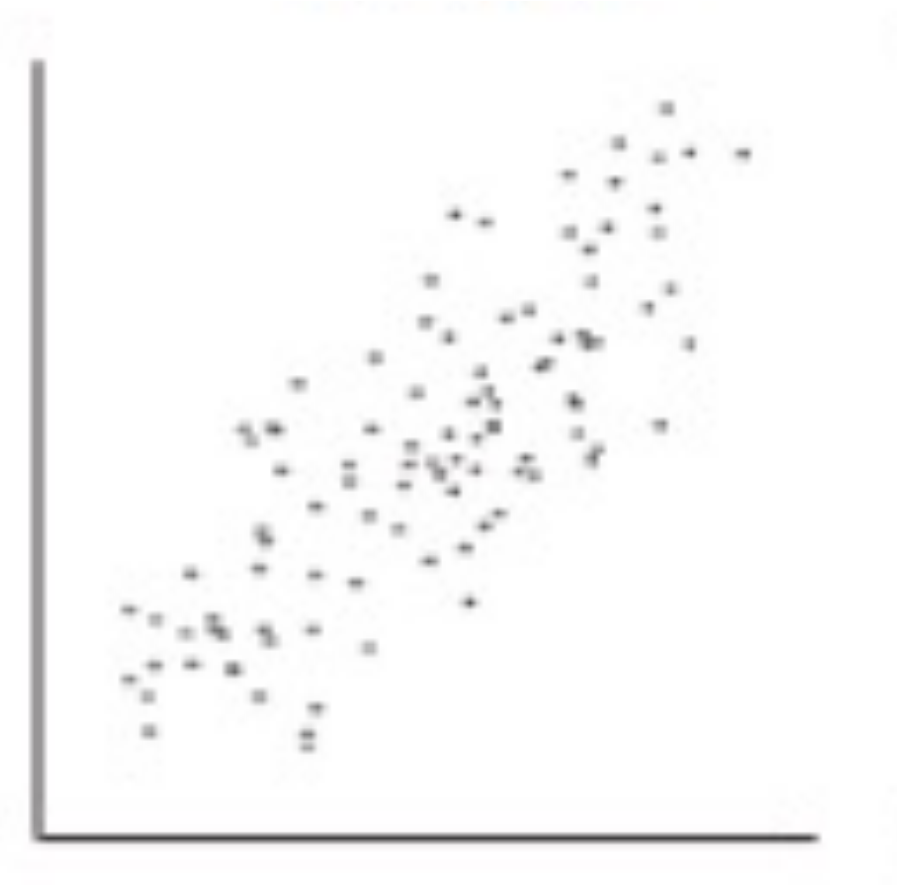
Scatterplots versus Parallel Coordinates





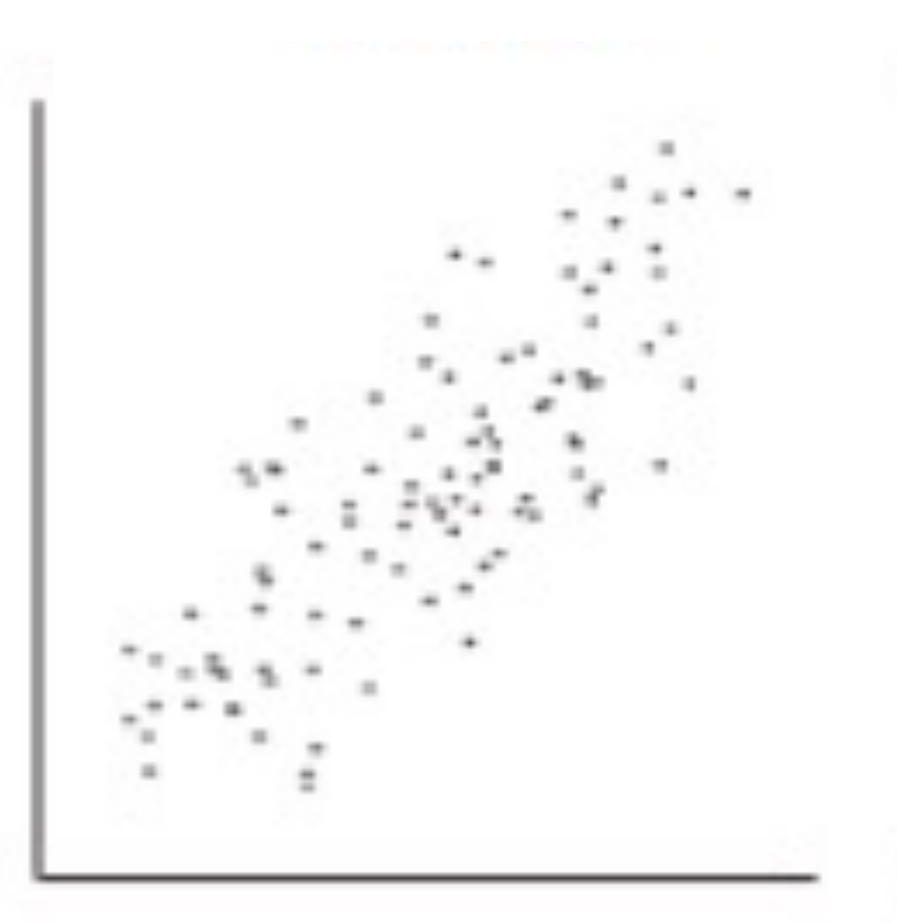
Harrison et al. Ranking Visualizations of Correlation Using Weber's Law 2014

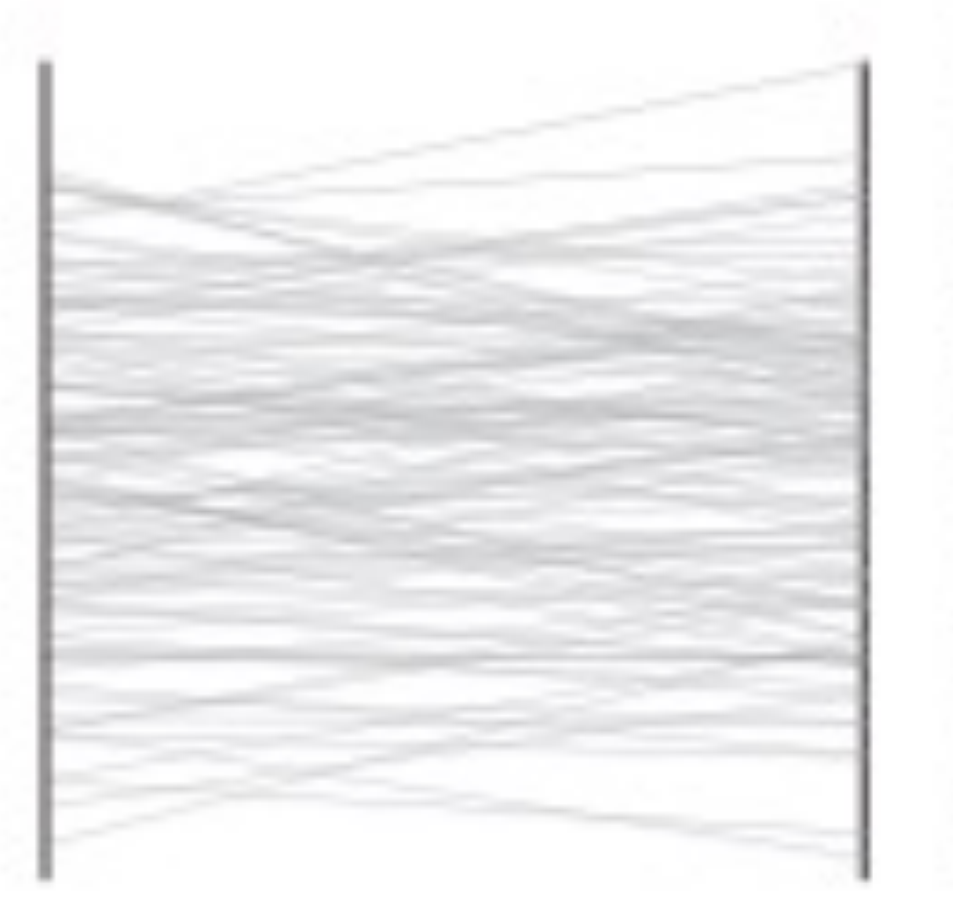
Scatterplots versus Parallel Coordinates



Harrison et al. Ranking Visualizations of Correlation Using Weber's Law 2014

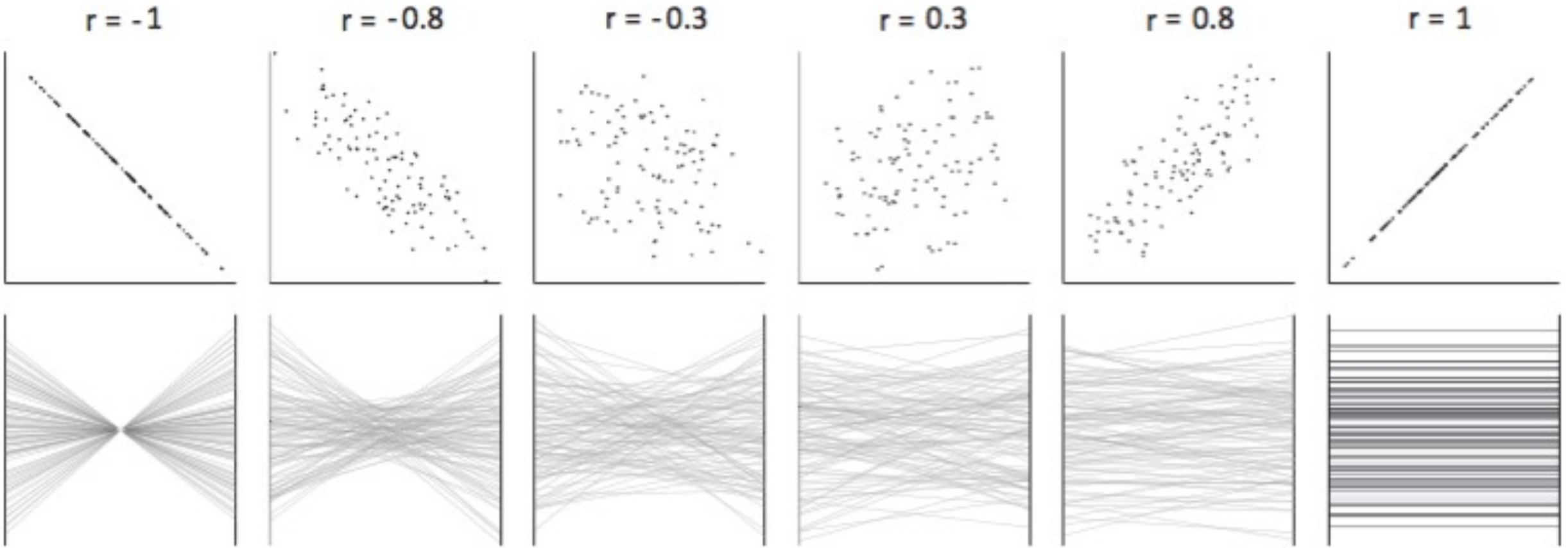
Scatterplots versus Parallel Coordinates





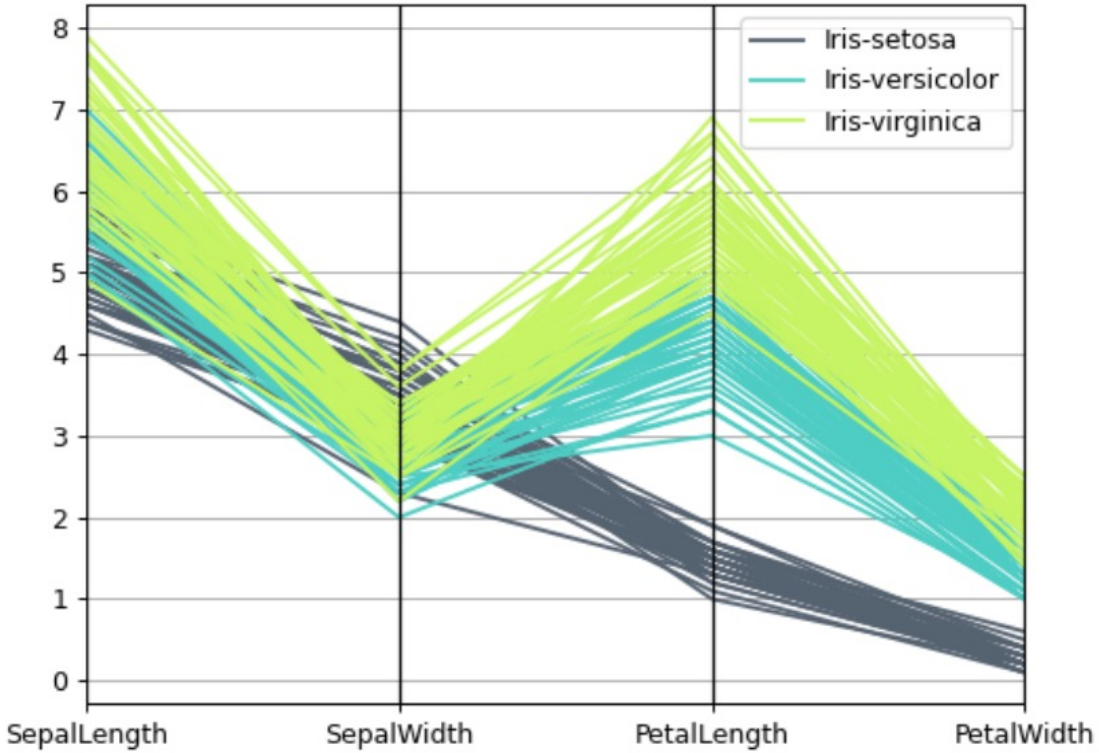
Harrison et al. Ranking Visualizations of Correlation Using Weber's Law 2014

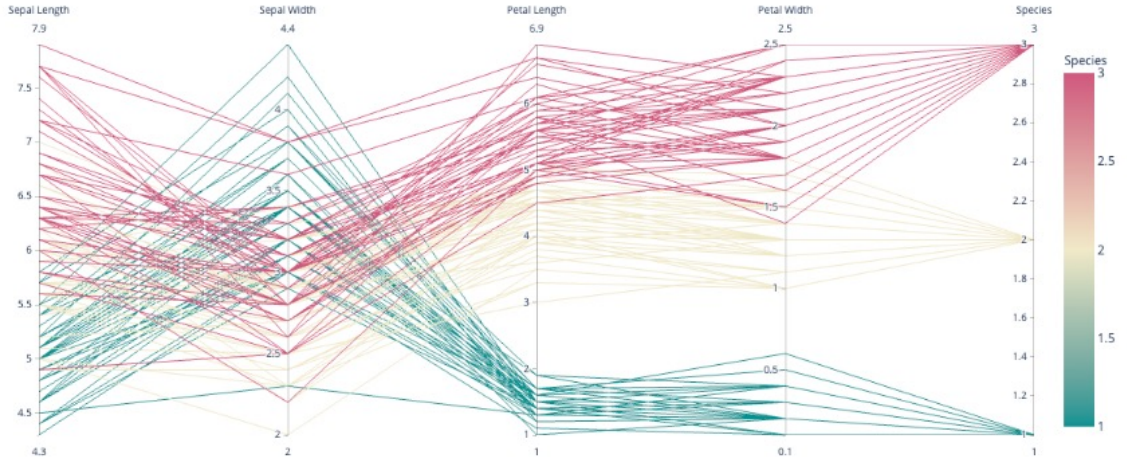
Scatterplots versus Parallel Coordinates



Harrison et al. Ranking Visualizations of Correlation Using Weber's Law 2014

Pandas and Plot.ly support





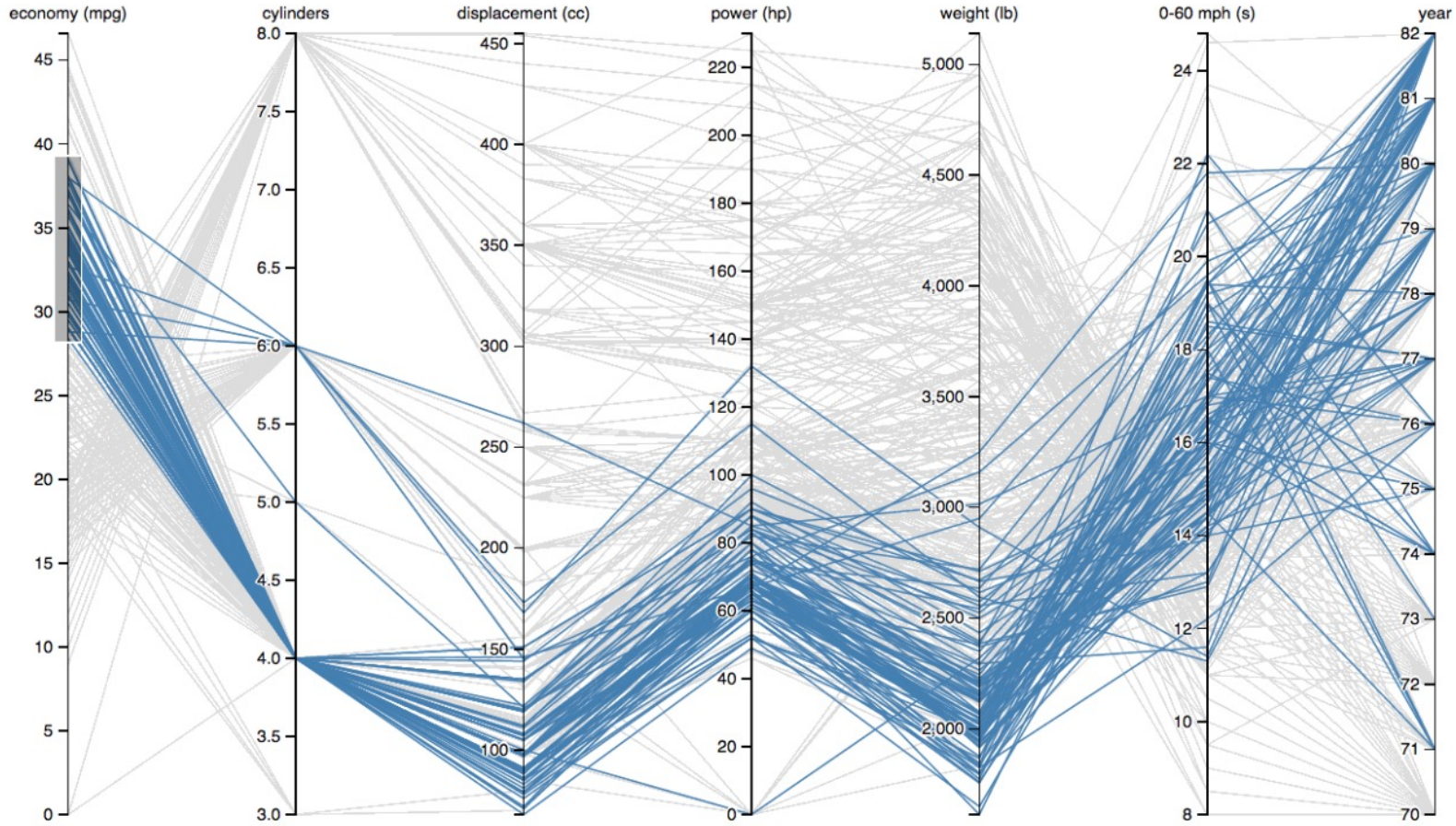
https://pandas.pydata.org/docs/reference/api/pandas.plotting.parallel_coordinates.html

https://plotly.com/python/parallel-coordinates-plot/

Parallel coordinates

Brushable across many axes at once

Axes may be re-ordered



https://bl.ocks.org/jasondavies/1341281

Pixel-based displays

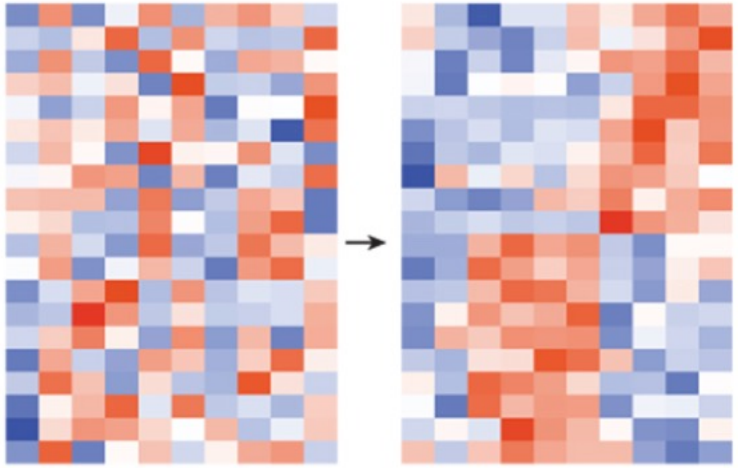
Each cell is a “pixel”
Value encoded with color

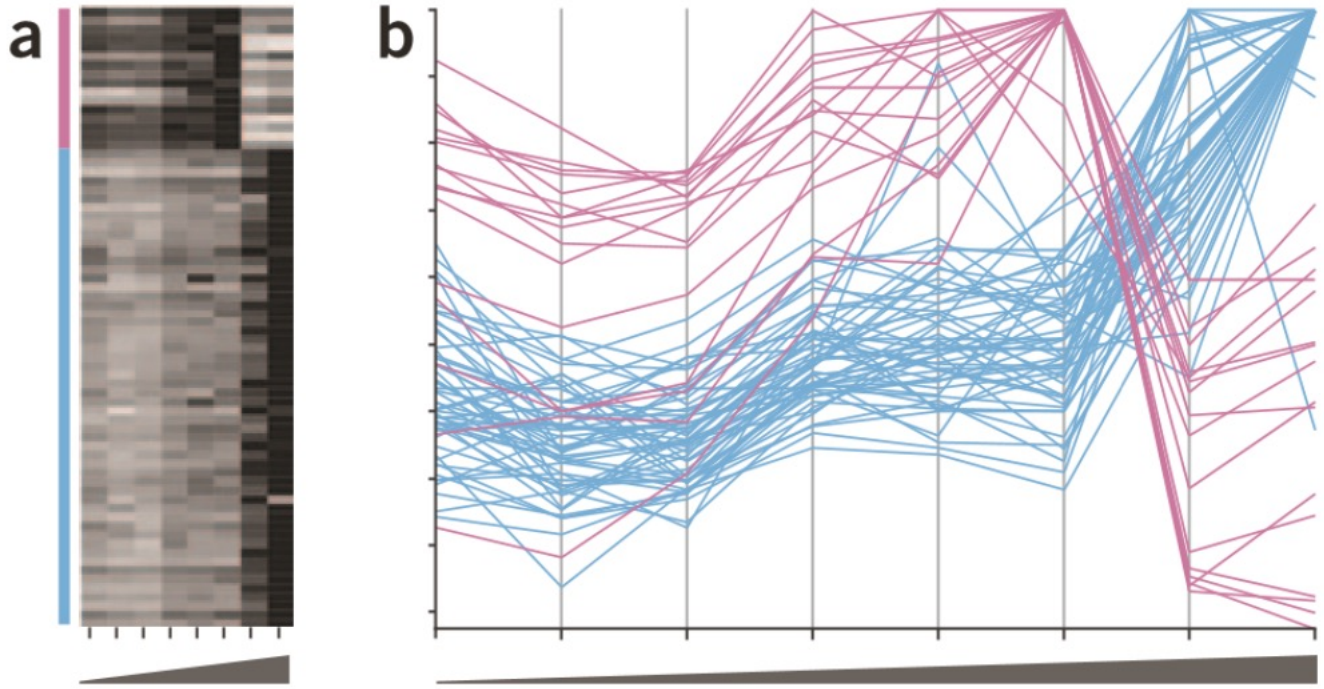
Relies on ordering for interpretability

If order isn’t obvious, use clustering

May scale to 1px / item

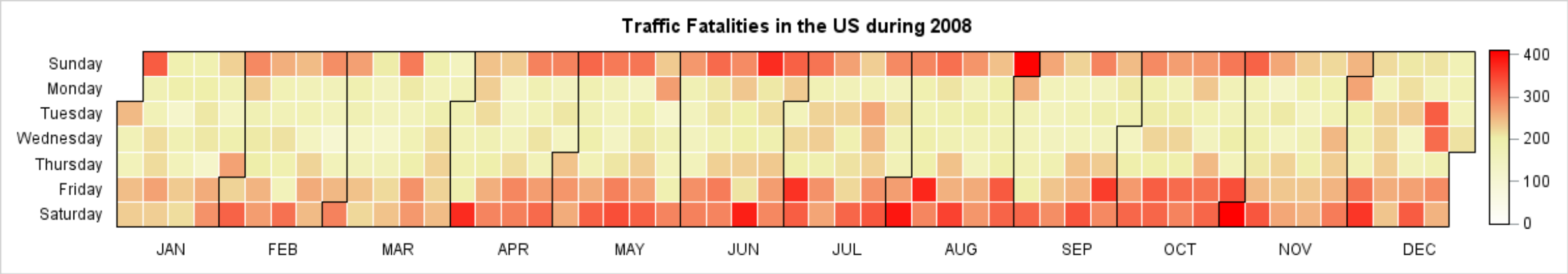
Good for data with same scale and type





[Gehlenborg & Wong 2012]

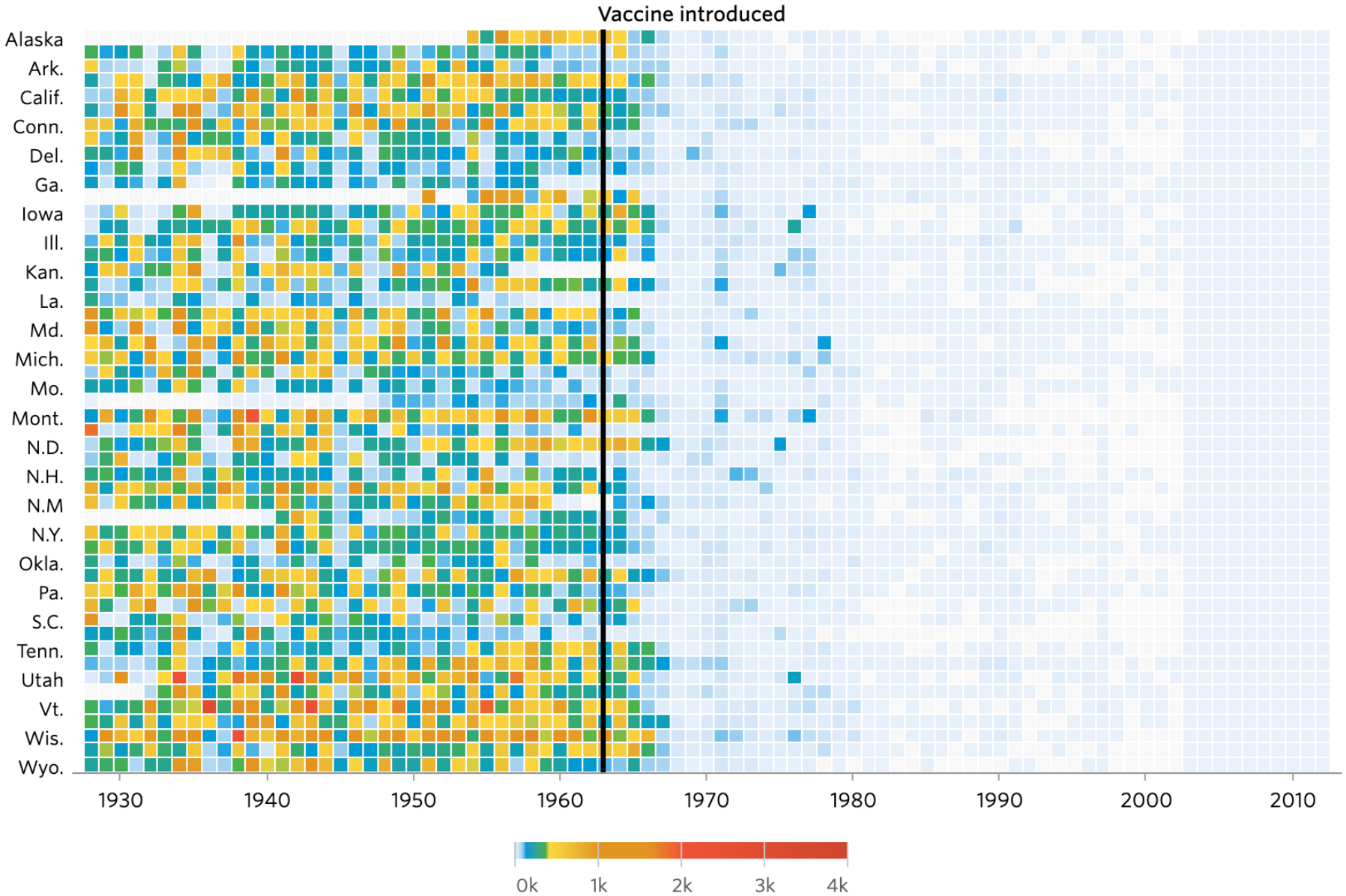
Heat map & calendar heat map



Seen something similar elsewhere?

https://blogs.sas.com/content/graphicallyspeaking/2011/12/08/calendar-heatmaps-in-gtl/

Measles

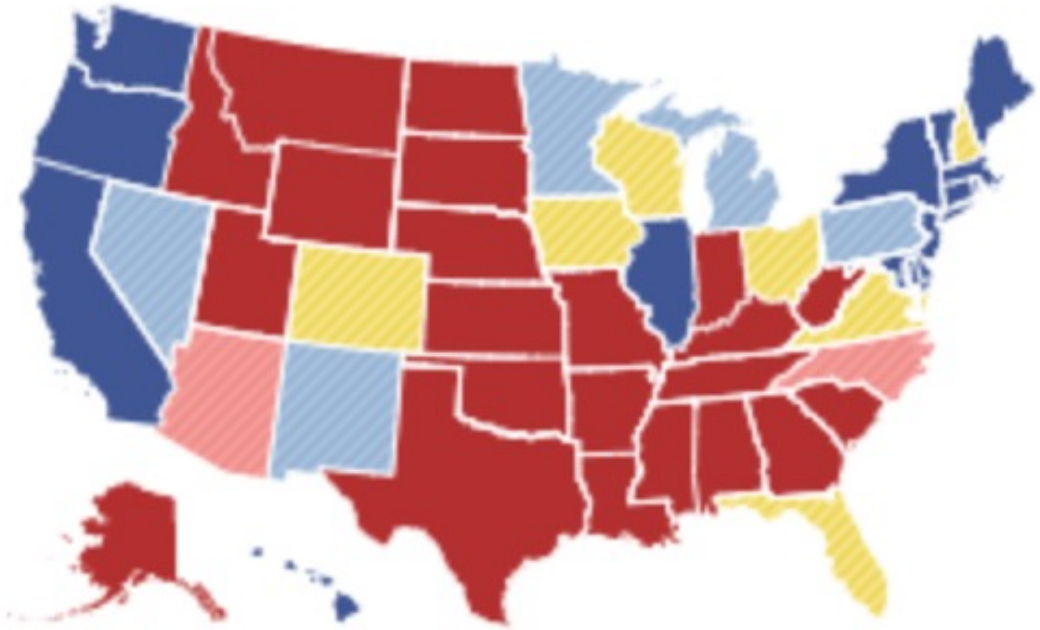


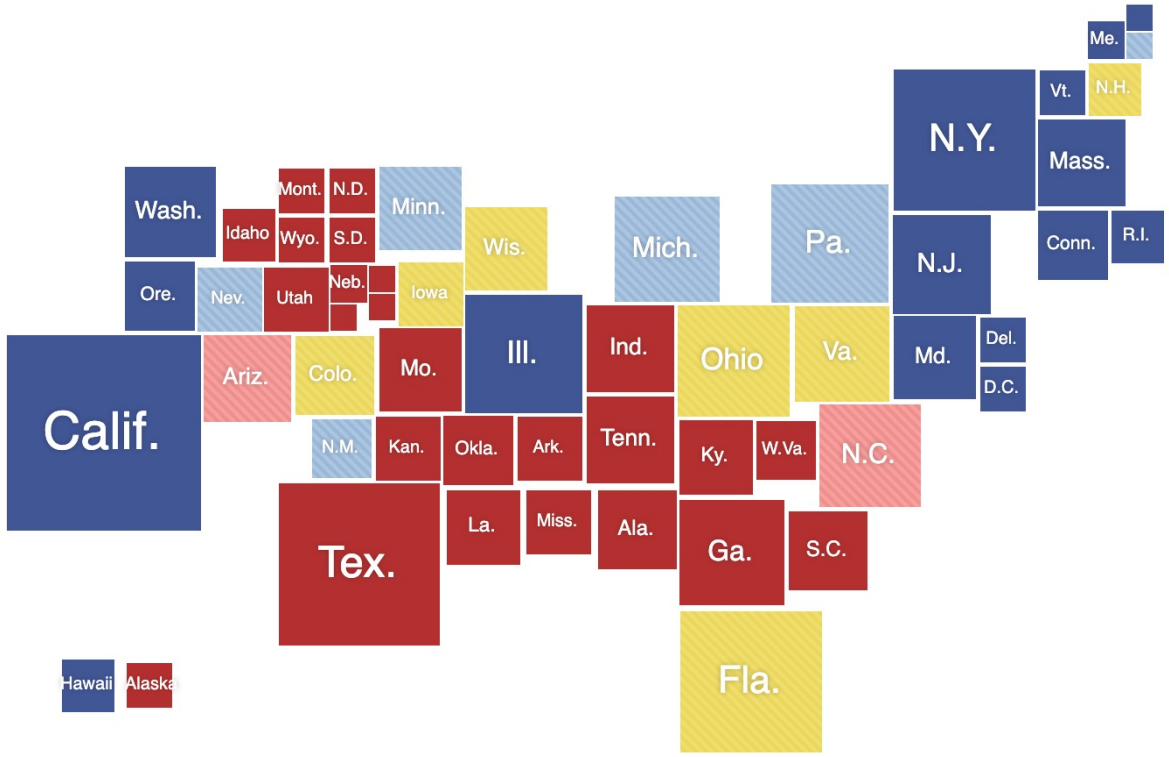
http://graphics.wsj.com/infectious-diseases-and-vaccines/

What if I have geospatial data?

…like lat/long

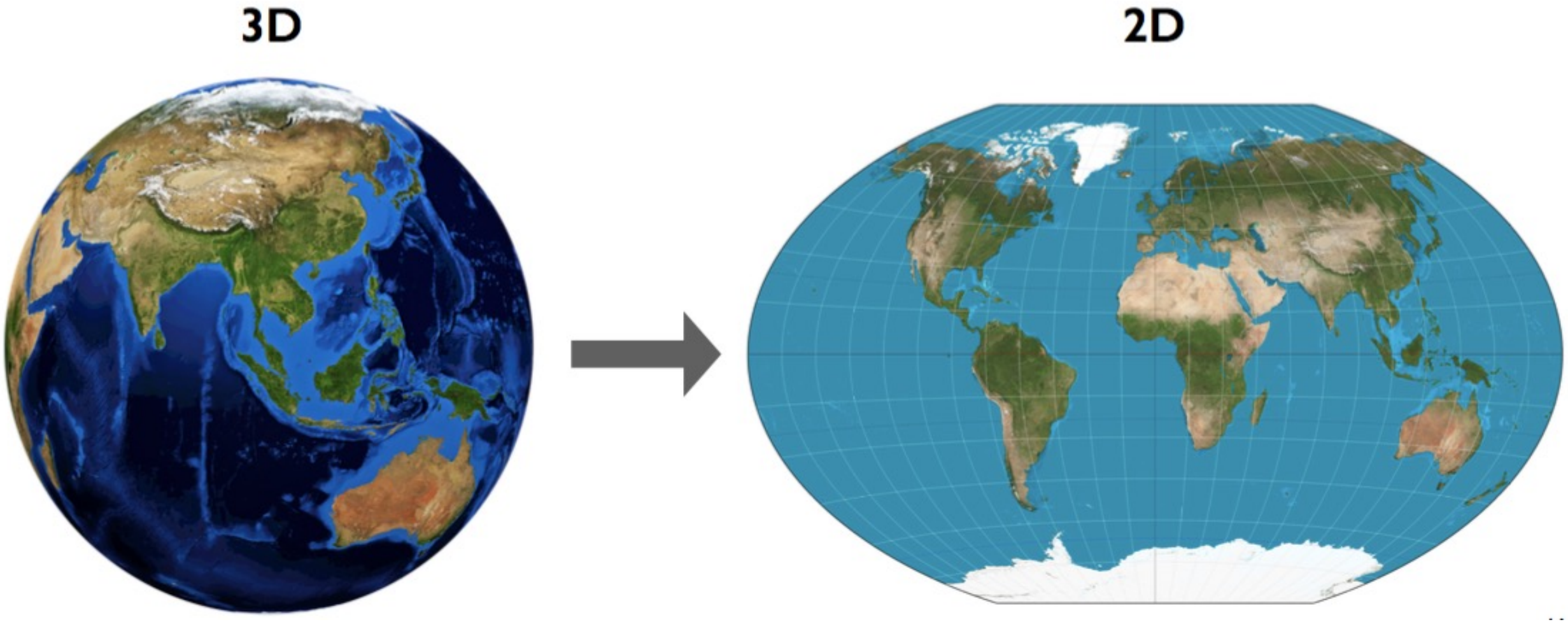
Maps & Cartograms





https://www.nytimes.com/elections/2012/ratings/electoral-map.html

Map projection is a form of dimensionality reduction



Adapted from slides of Michelle Borkin

Different projections preserve different things

- Areas
- Shapes
- Directions
- Bearings
- Distances
- Scales

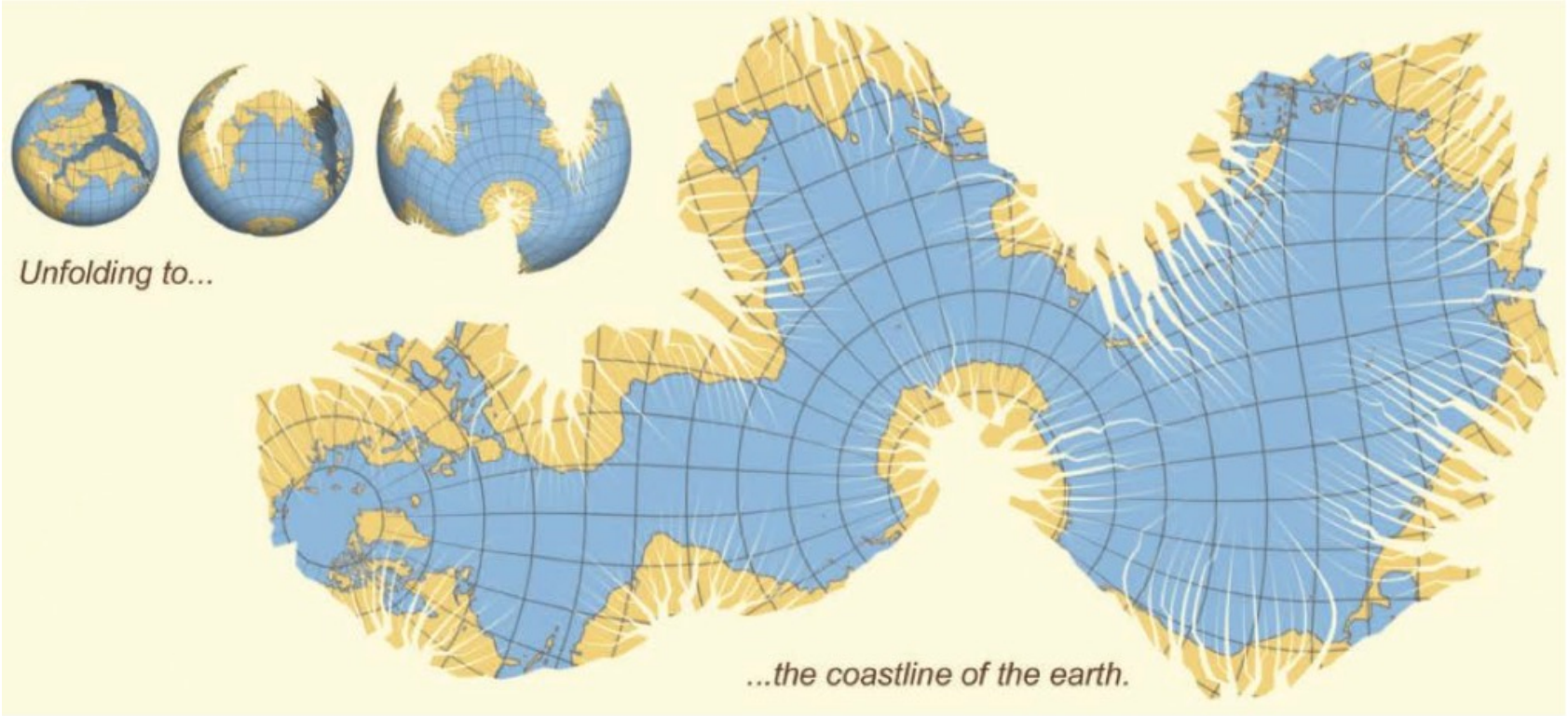
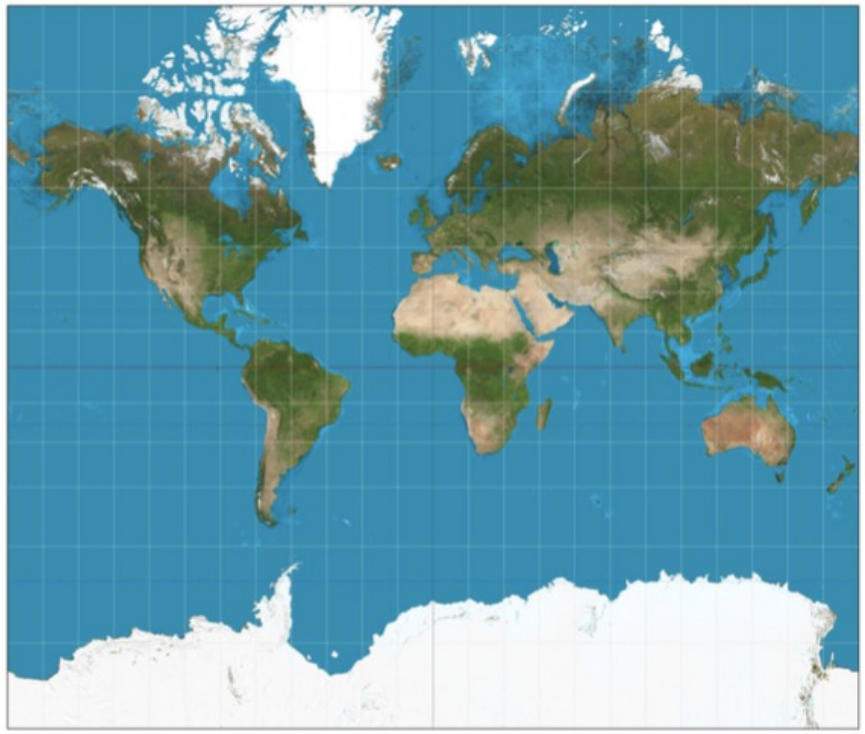


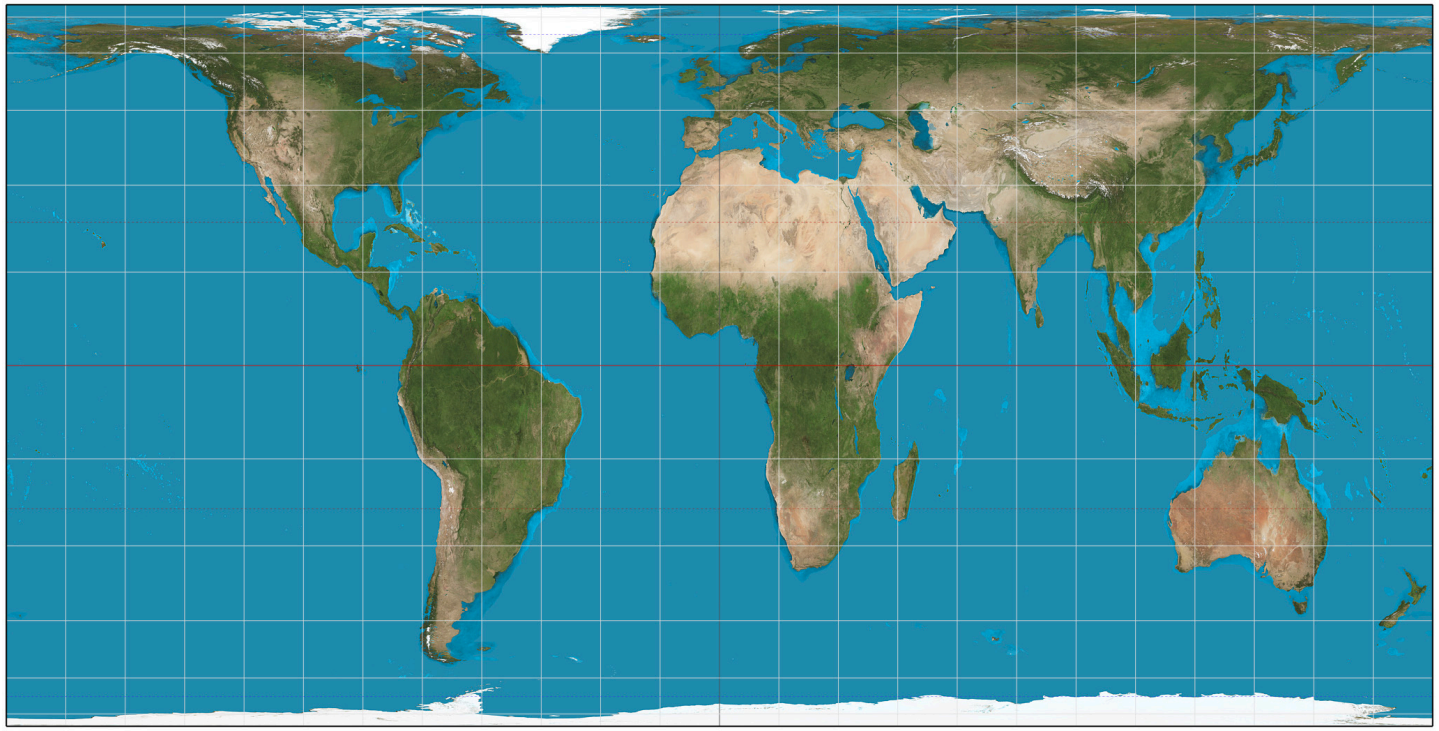
Image from van Wijk 2008 via slides of Josh Levine & Miriah Meyer.

Which projection should I use?

Mercator



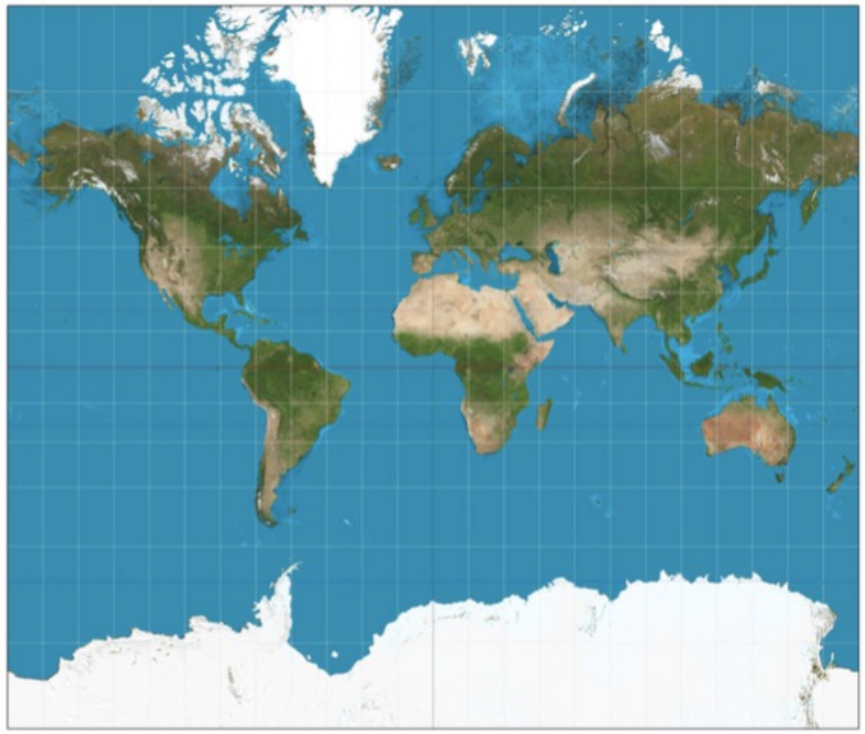
Hobo-Dyer



Adapted from slides of Michelle Borkin

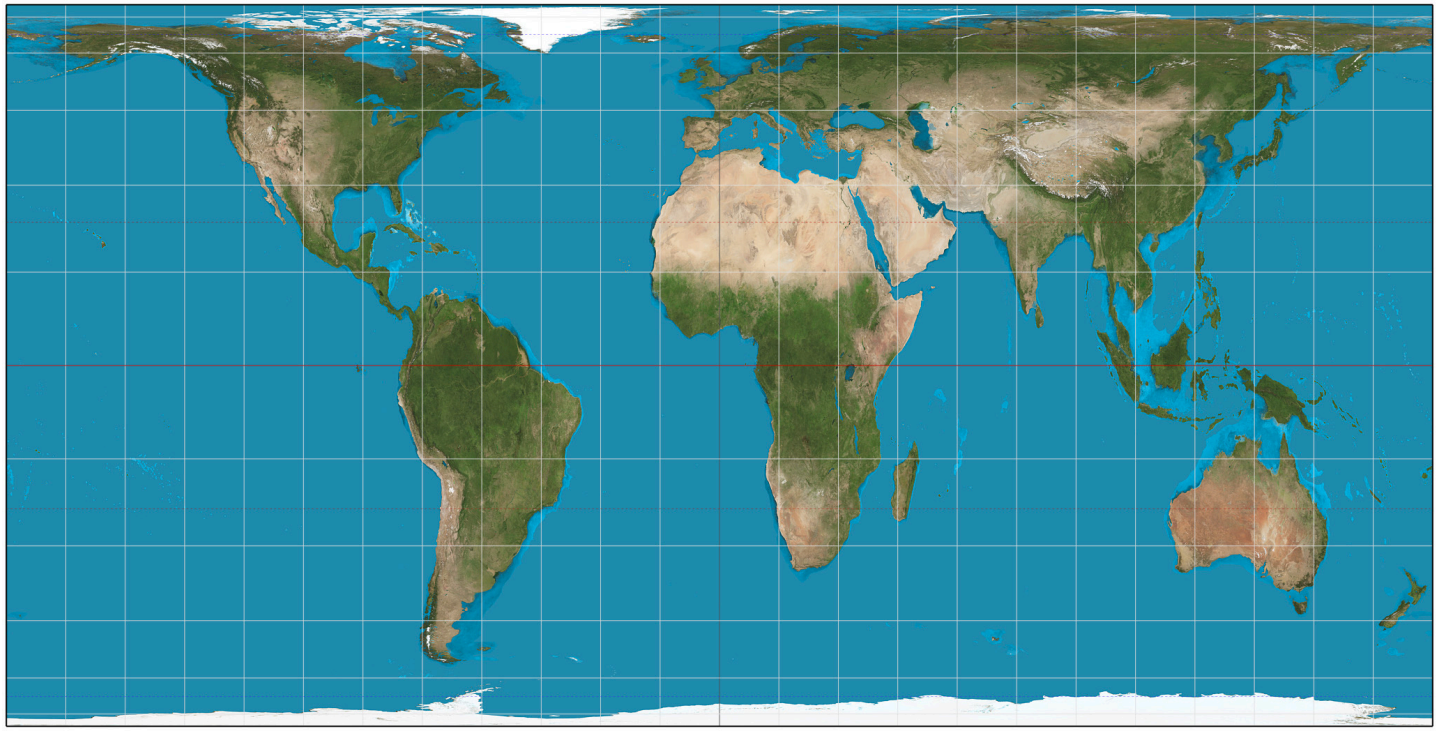
Which projection should I use?

Mercator



Navigation

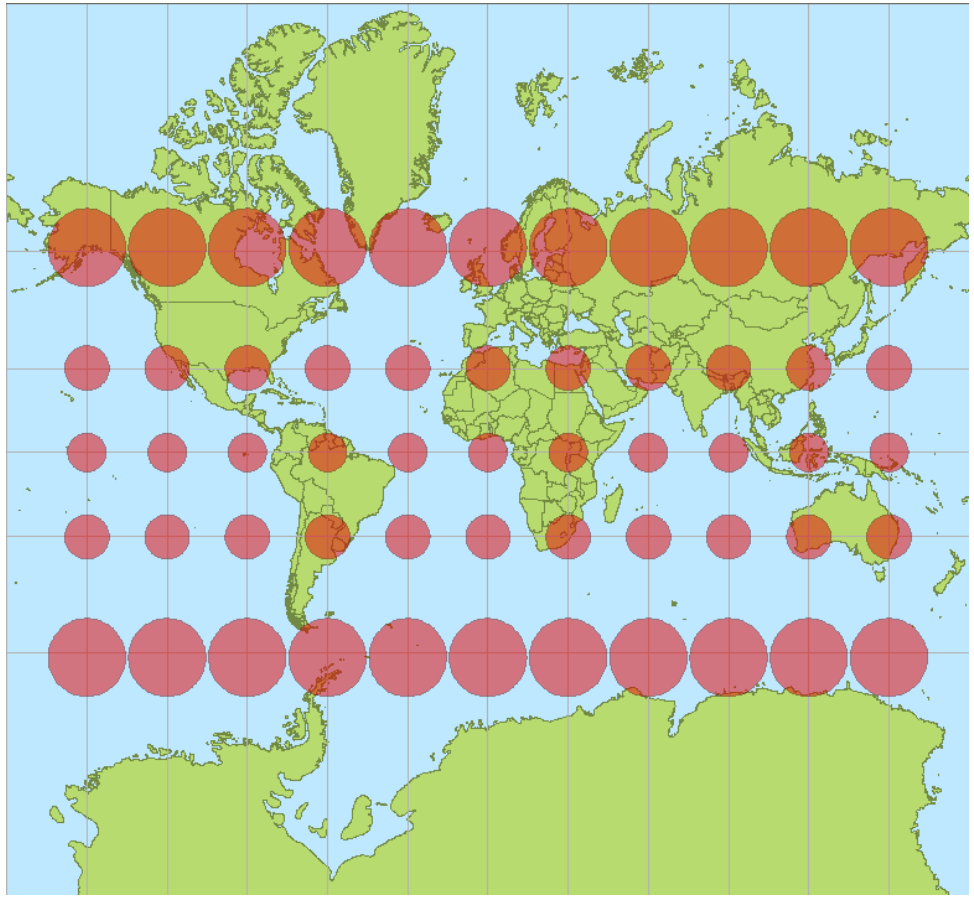
Hobo-Dyer

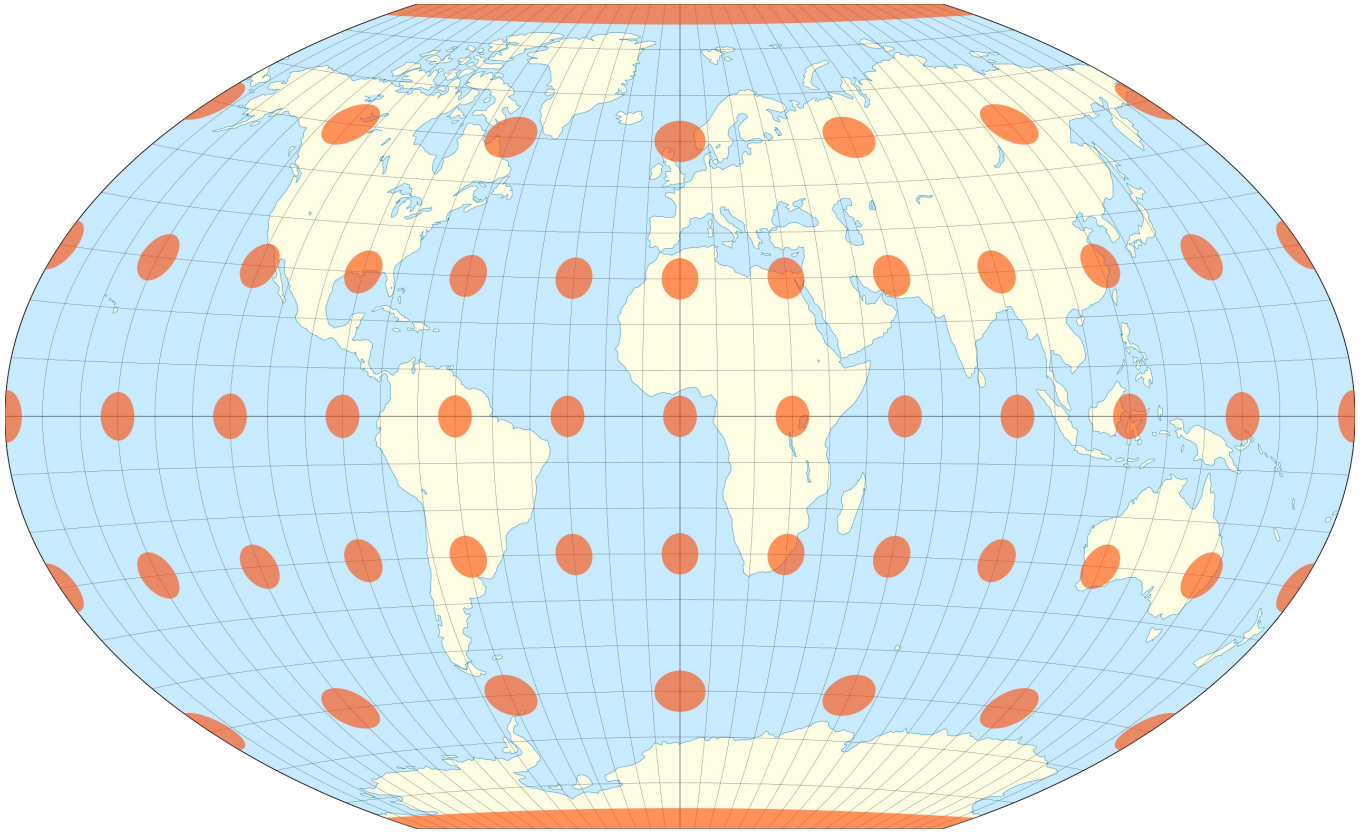


Area Comparison

Adapted from slides of Michelle Borkin

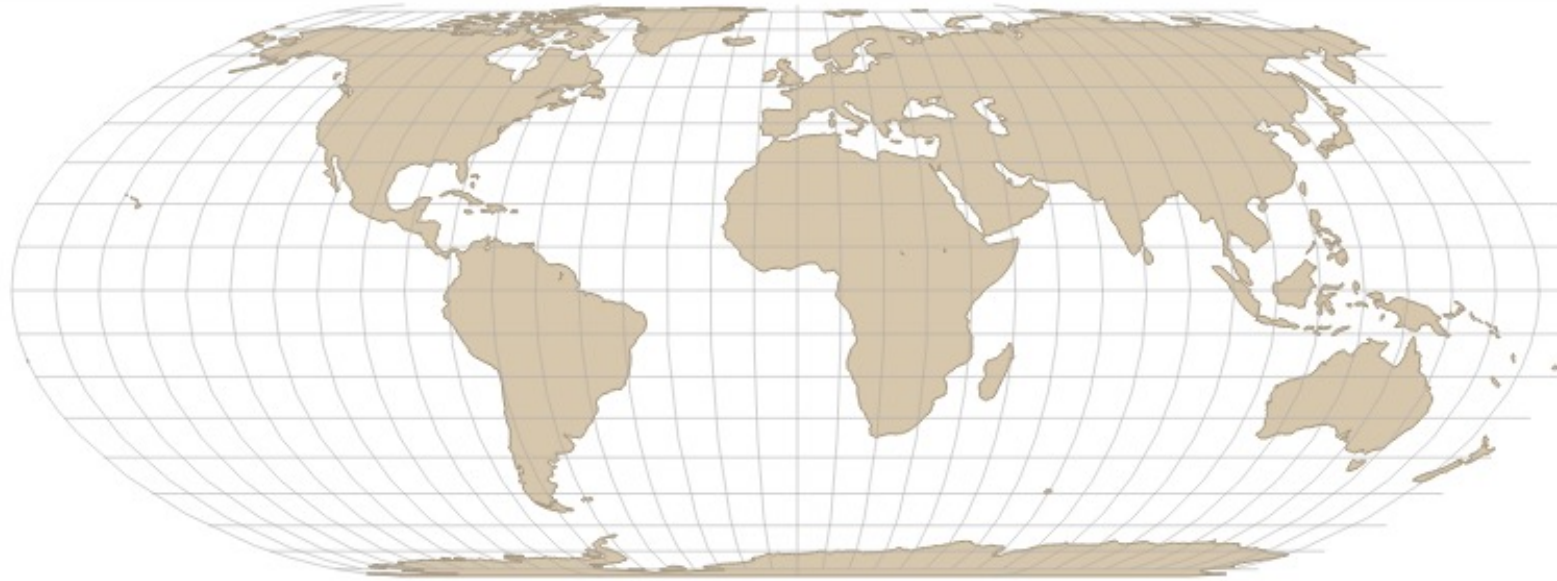
Comparing Map Projections – Tissot's Indicatrix

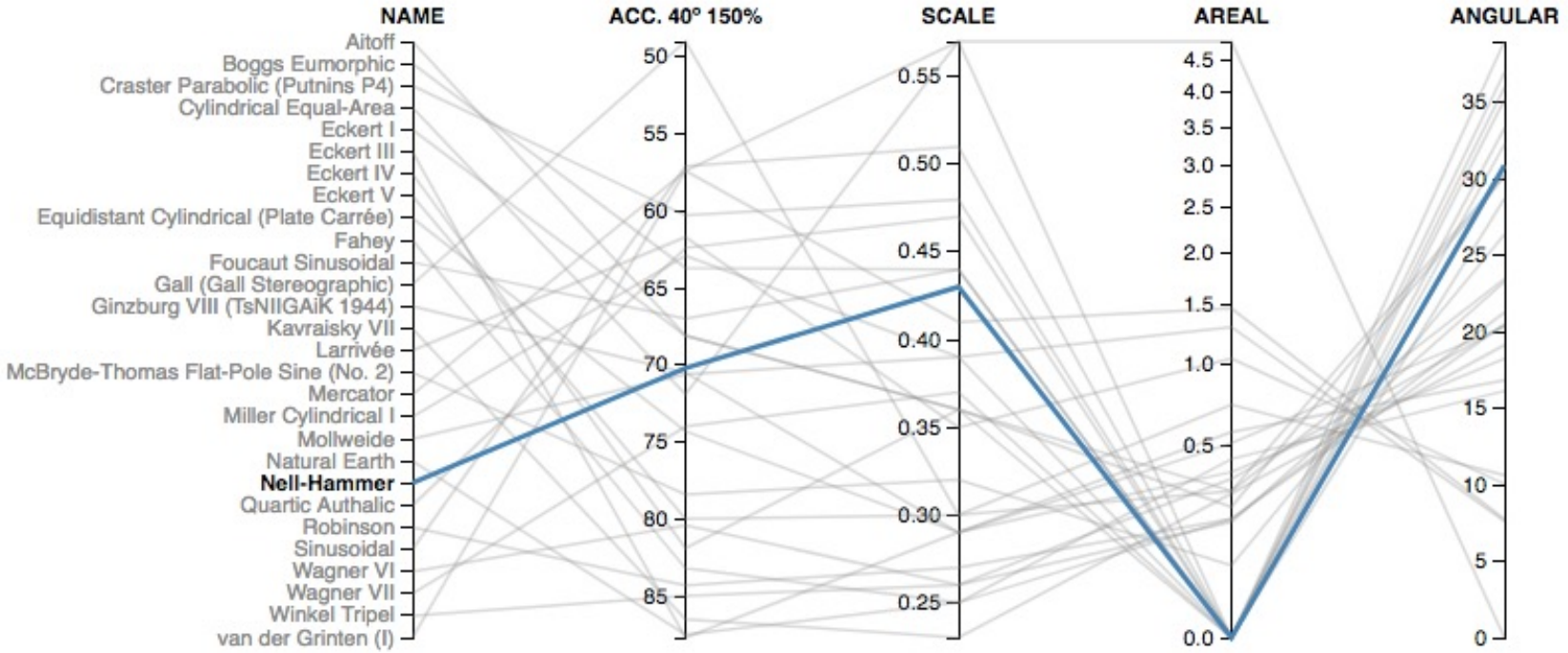




See also: https://observablehq.com/@d3/tissots-indicatrix

Comparing Map Projections – Distortions



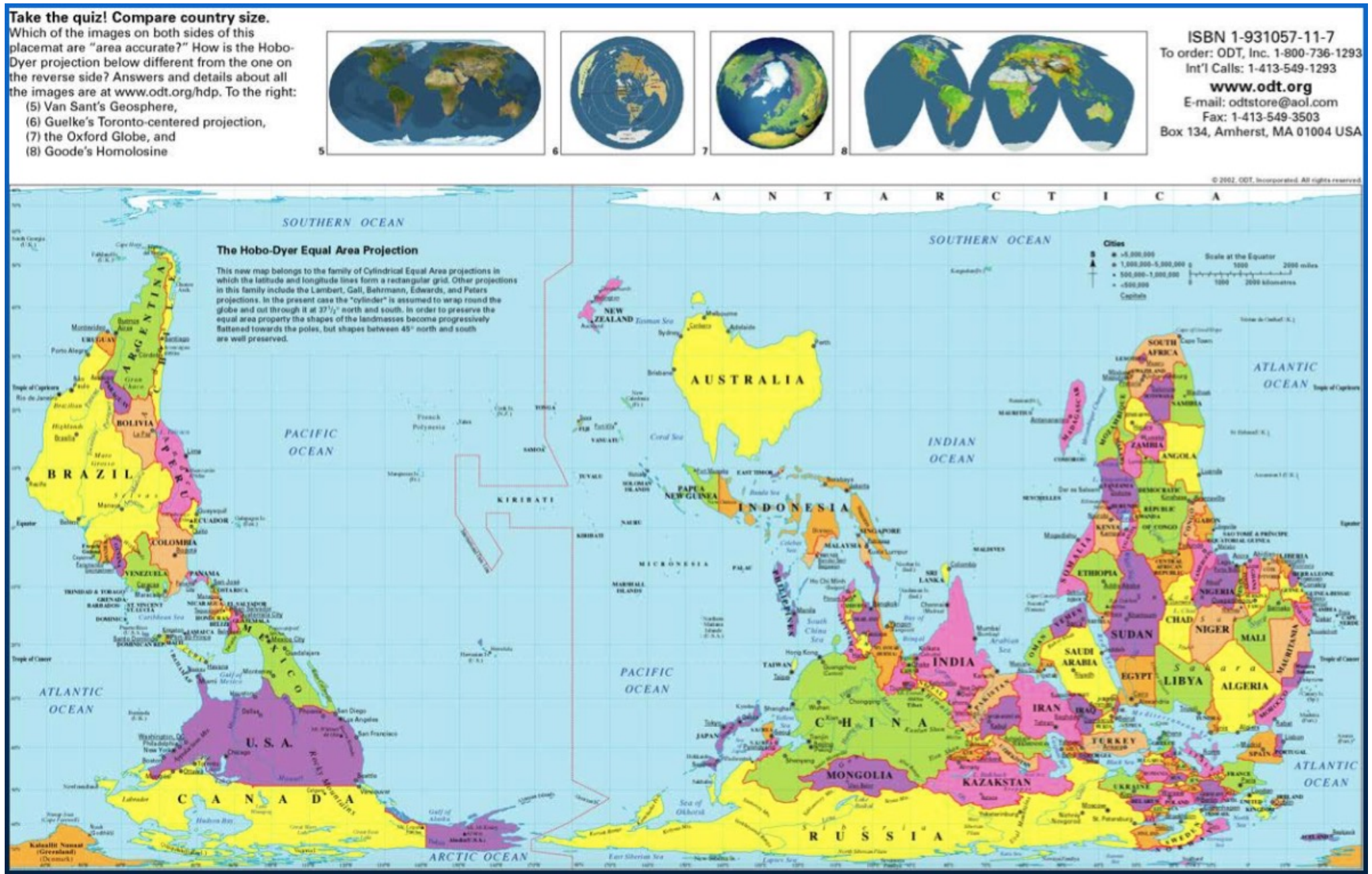


https://bl.ocks.org/syntagmatic/ba569633d51ebec6ec6e

Projection Choices & Orientation Change What We See

Adapted from slides of Josh Levine & Miriah Meyer.

Projection Choices & Orientation Change What We See



Adapted from slides of Josh Levine & Miriah Meyer.

How big is Africa?

How big is Africa?

The True Size of Africa

A small contribution in the fight against rampant Innumeracy, by Kai Krause

Graphic layout for visualization only (some countries are cut and rotated)
But the conclusions are very accurate: refer to table below for exact data

COUNTRY	AREA x 1000 km²
China	9.597
USA	9.629
India	3.287
Mexico	1.964
Peru	1.285
France	633
Spain	506
Papua New Guinea	462
Sweden	441
Japan	378
Germany	357
Norway	324
Italy	301
New Zealand	270
United Kingdom	243
Nepal	147
Bangladesh	144
Greece	132
TOTAL	30.102
AFRICA	30.221

In addition to the well known social issues of *illiteracy* and *innumeracy*, there also should be such a concept as "*innumpancy*", meaning *insufficient geographical knowledge*.

A survey with random American schoolkids let them guess the population and land area of their country. Not entirely unexpected, but still rather unsettling, the majority chose "*1-2 billion*" and "*largest in the world*", respectively.

Even with Asian and European college students, geographical estimates were often off by factors of *2-3*. This is partly due to the highly distorted nature of the predominantly used mapping projections (such as *Mercator*).

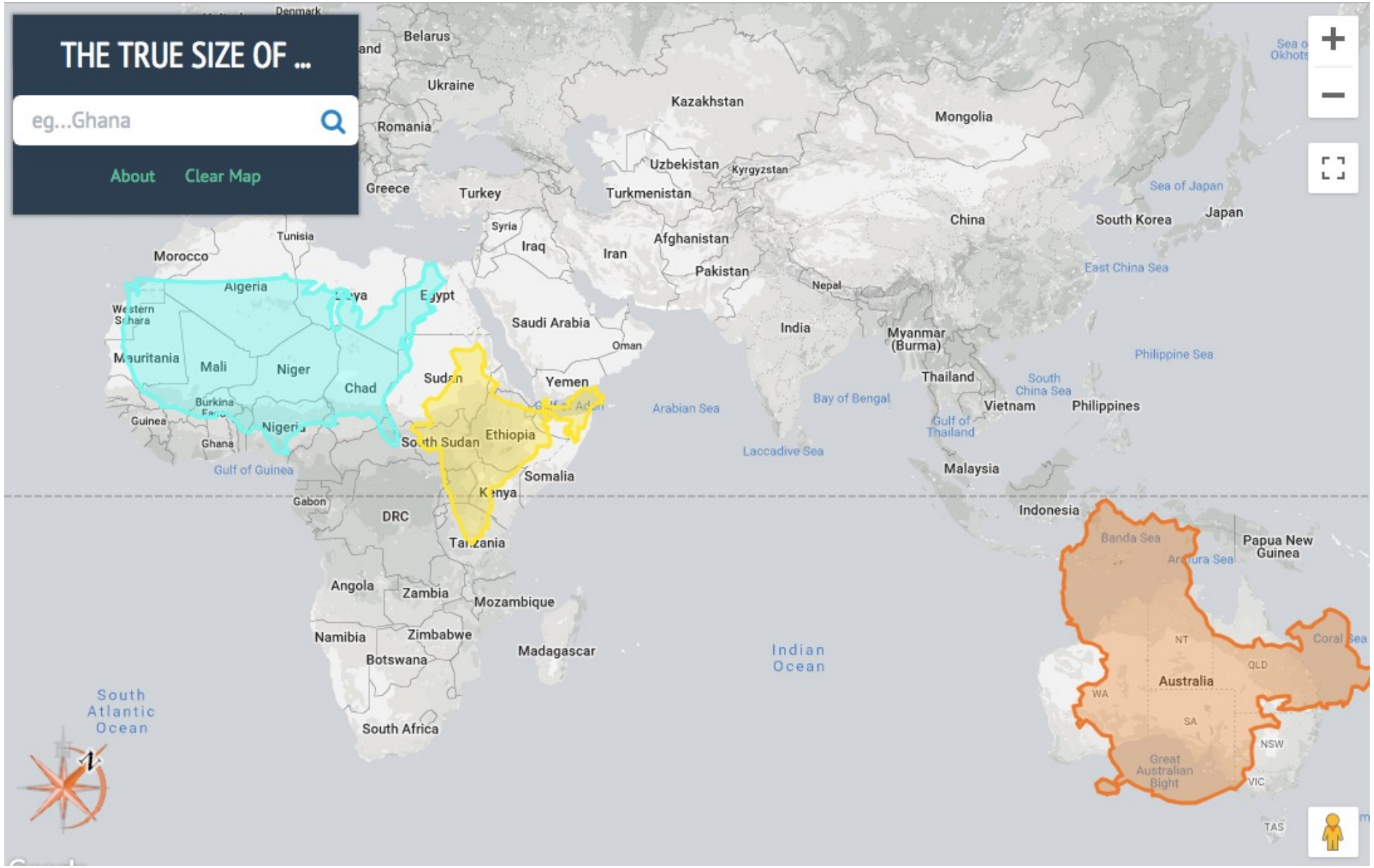
A particularly extreme example is the worldwide misjudgement of the true size of *Africa*. This single image tries to embody the massive scale, which is larger than the *USA, China, India, Japan* and *all of Europe..... combined!*

Top 100 Countries

Area in square kilometers, Percentage of World Total
Sources: Britannica, Wikipedia, Almanac 2010

Adapted from slides of Michelle Borkin, https://www.sciencealert.com/this-is-the-true-size-of-africa

thetruesize.com

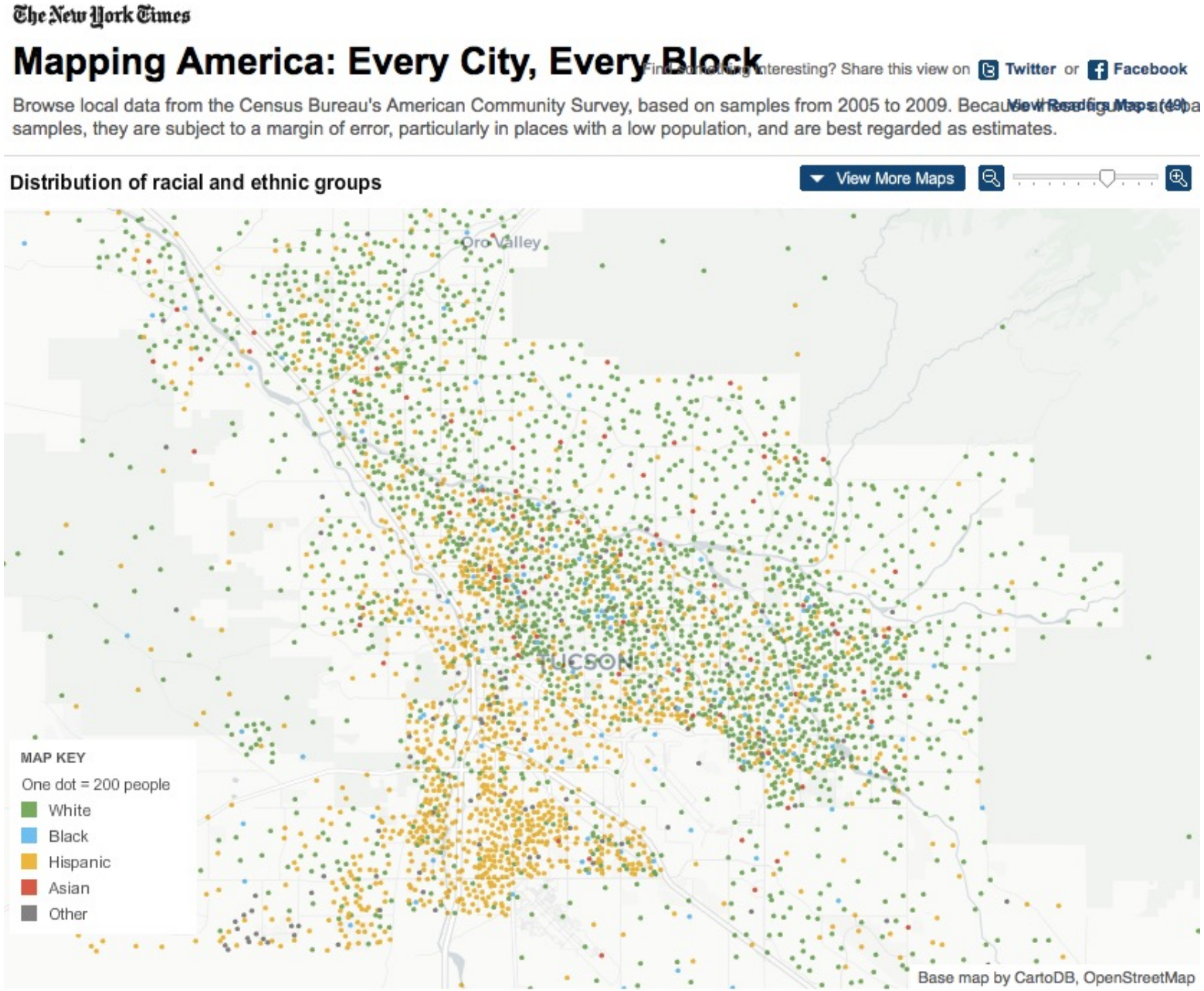


How big is everything else?

World Mercator projection with true country size and shape added

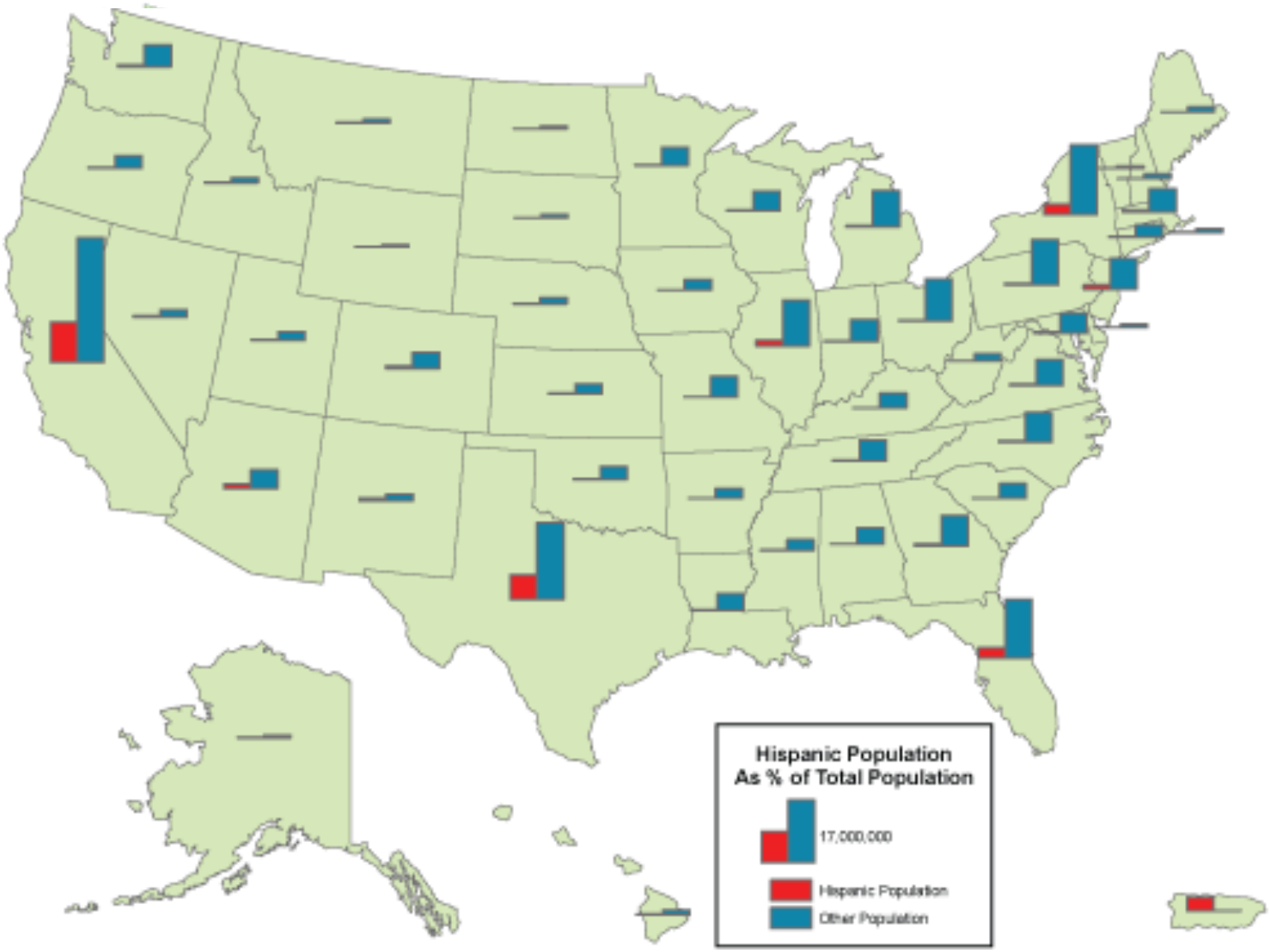
https://twitter.com/neilrkaye/status/1161610119165161473

Visualizing Data on Maps – Points



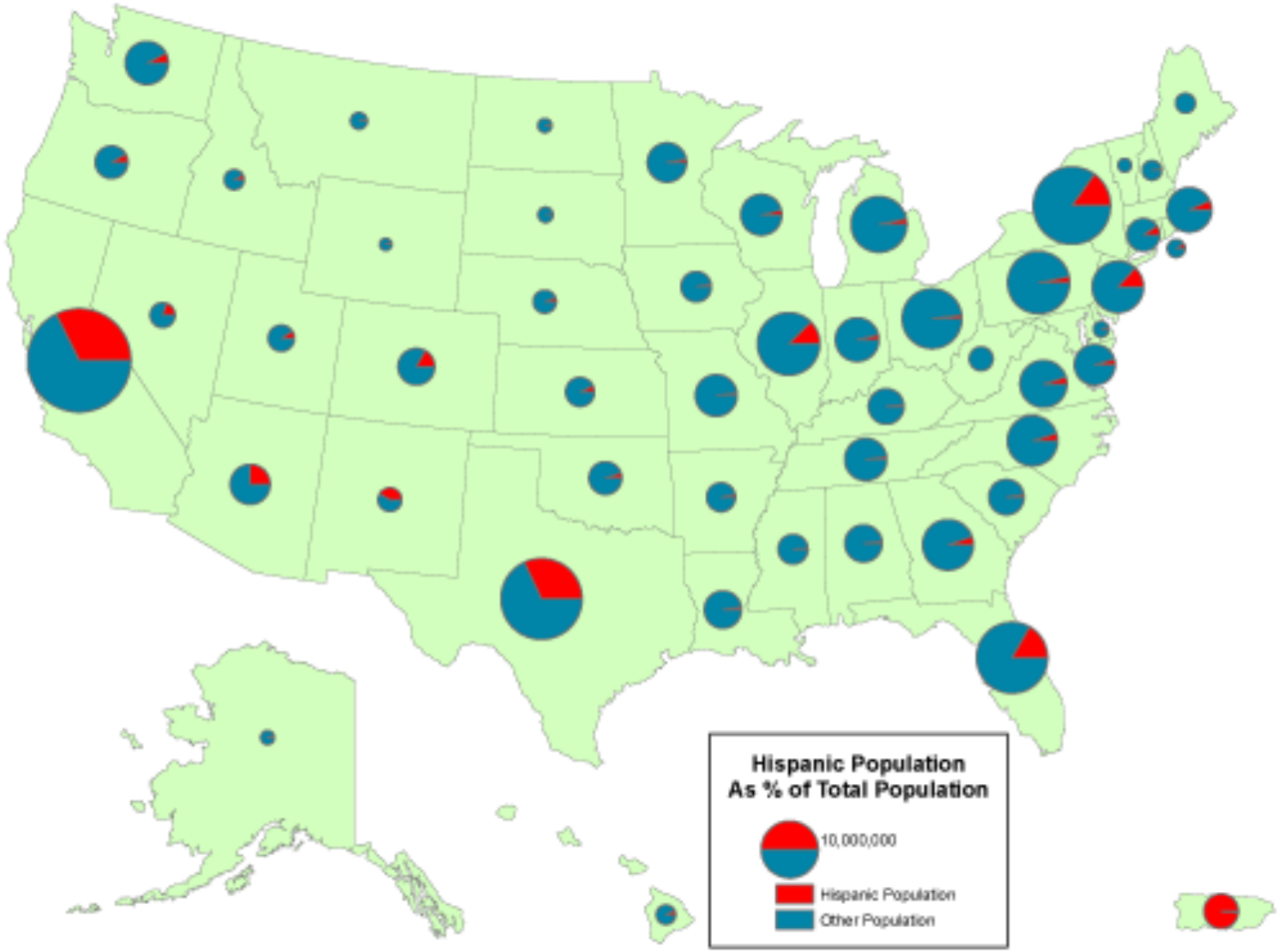
Adapted from slides of Meyer and Levine, See also http://racialdotmap.demographics.coopercenter.org/

Visualizing Data on Maps – Charts



https://www.e-education.psu.edu/natureofgeoinfo/c3_p17.html

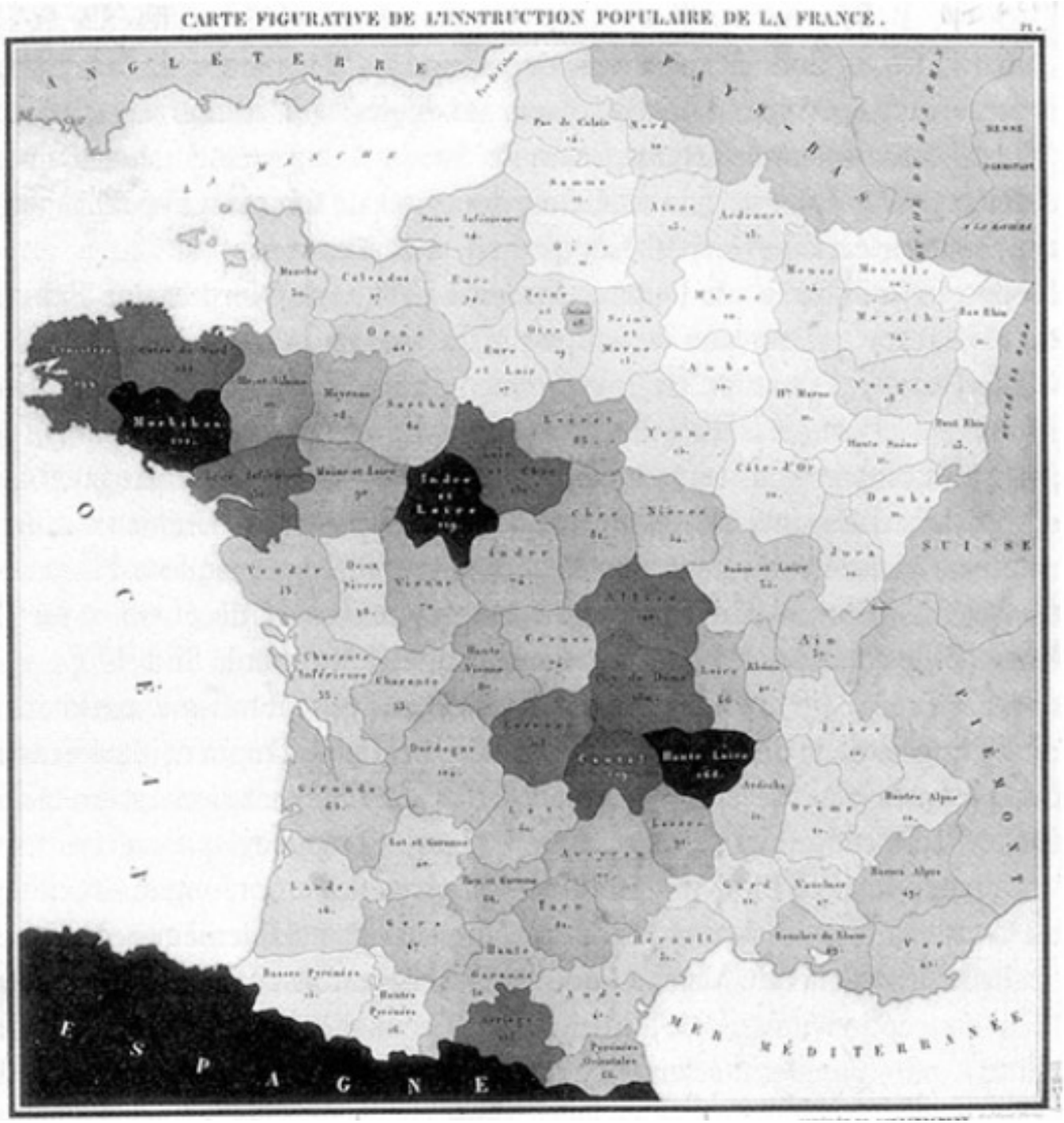
Visualizing Data on Maps – Charts



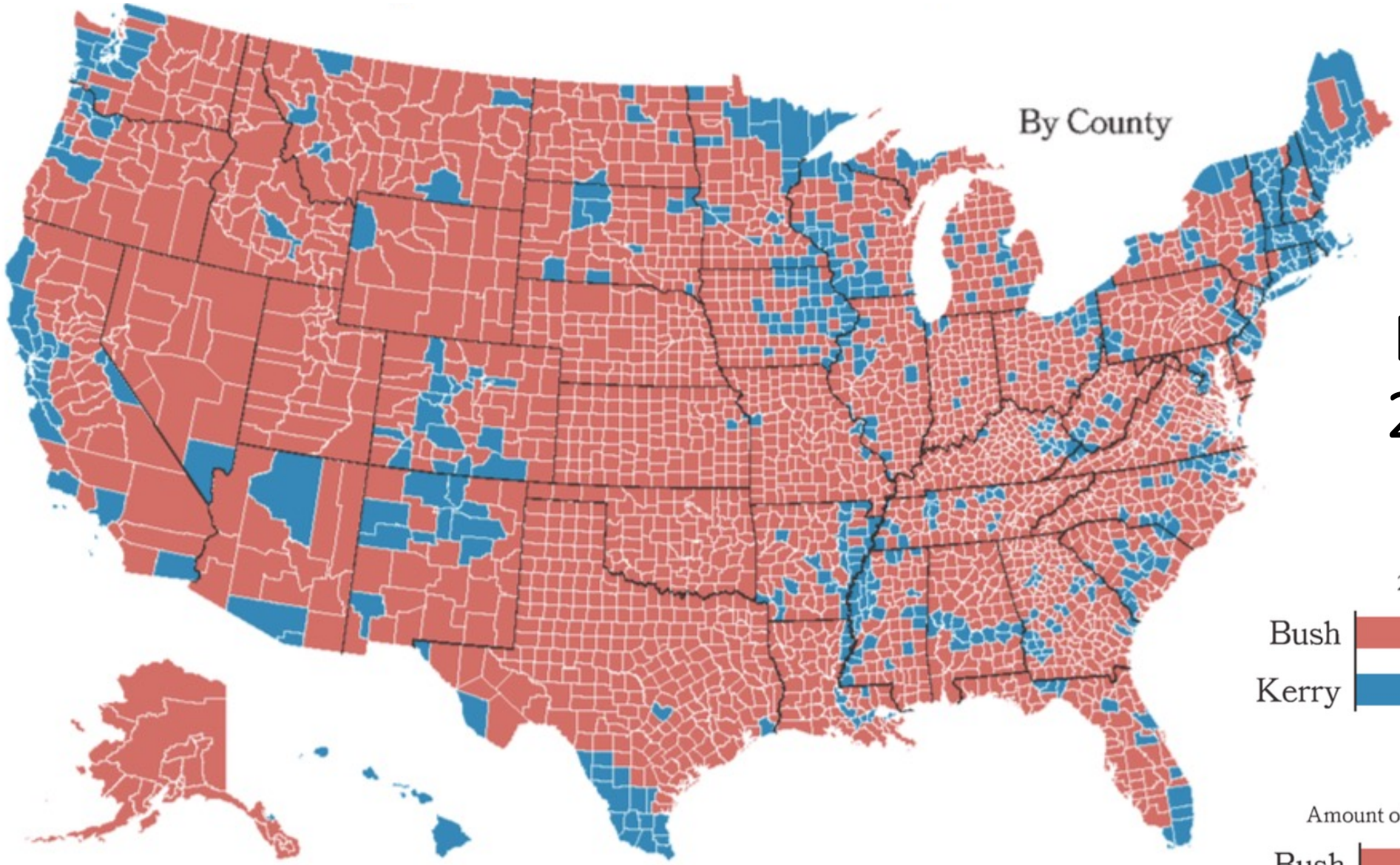
https://www.e-education.psu.edu/natureofgeoinfo/c3_p17.html

Choropleth maps color, shade, or fill areas to represent some (aggregated) data attribute

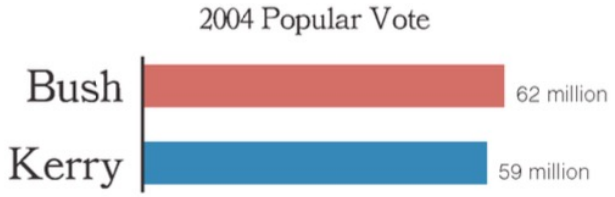
Illiteracy in France,
Dupin 1826

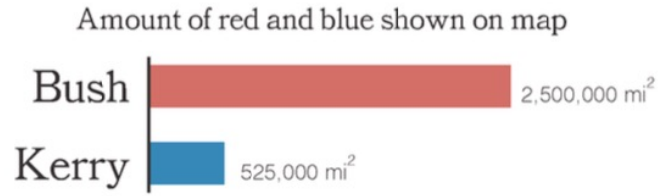


Adapted from slides of Meyer and Levine

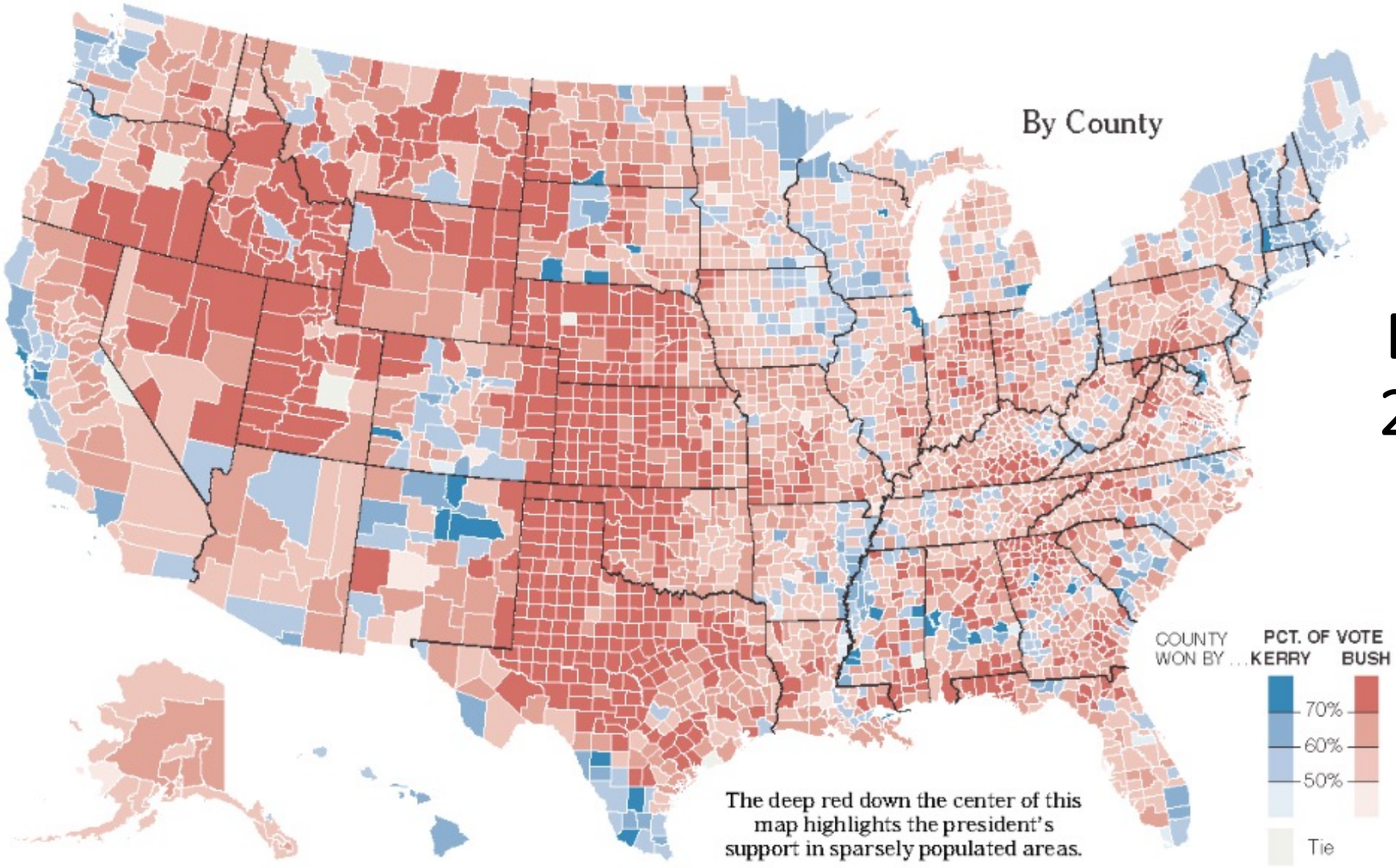


Bush v. Kerry, 2004 Election



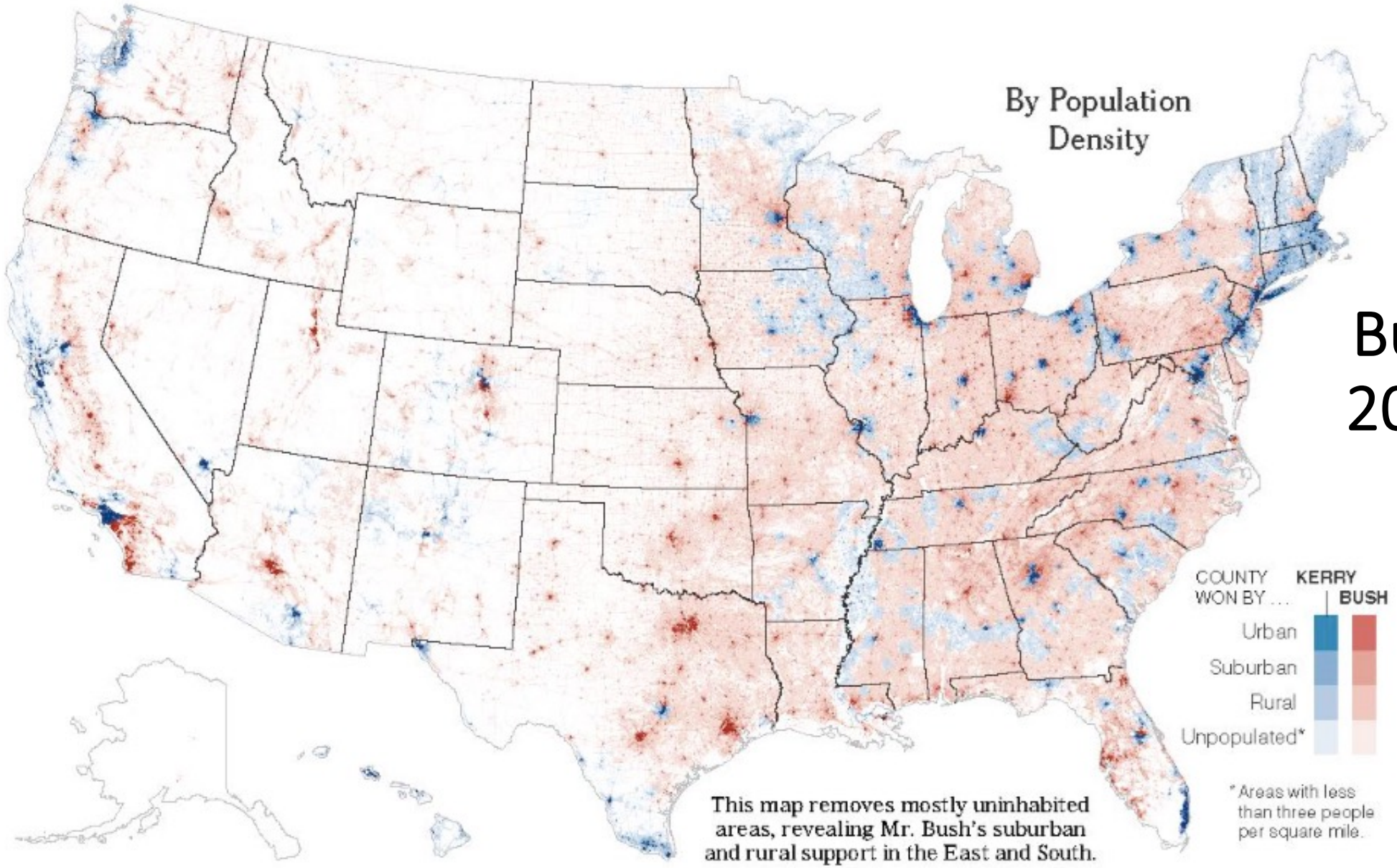


Matthew Ericson, NY Times, via slides of Meyer and Levine



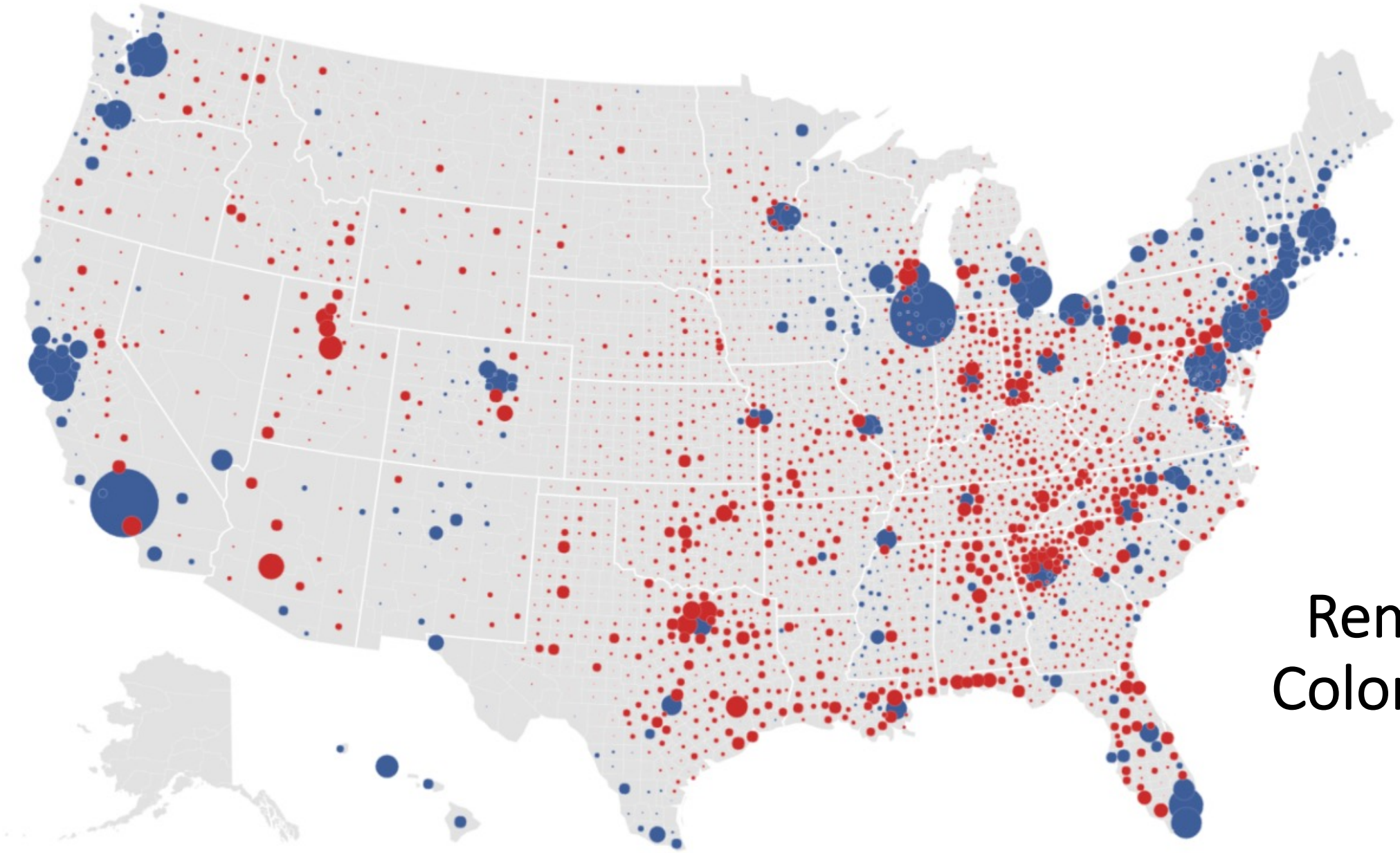
Bush v. Kerry, 2004 Election

Matthew Ericson, NY Times, via slides of Meyer and Levine



Bush v. Kerry, 2004 Election

Matthew Ericson, NY Times, via slides of Meyer and Levine



Remember:
Color is hard…

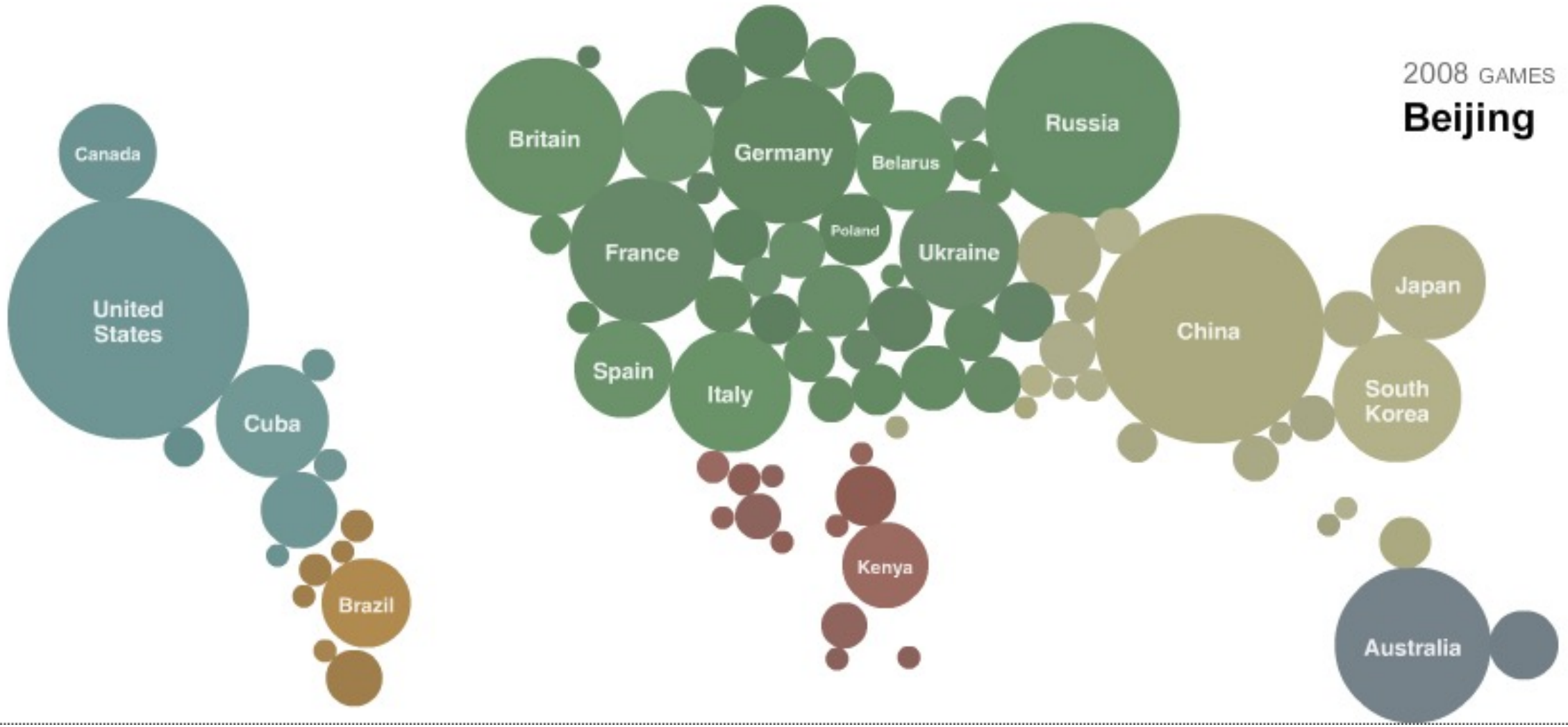
https://www.nytimes.com/interactive/2016/11/01/upshot/many-ways-to-map-election-results.html

Cartograms scale and distort areas to encode an attribute

Land areas of countries,
Emile Levasseur 1868,
Rectangular cartogram

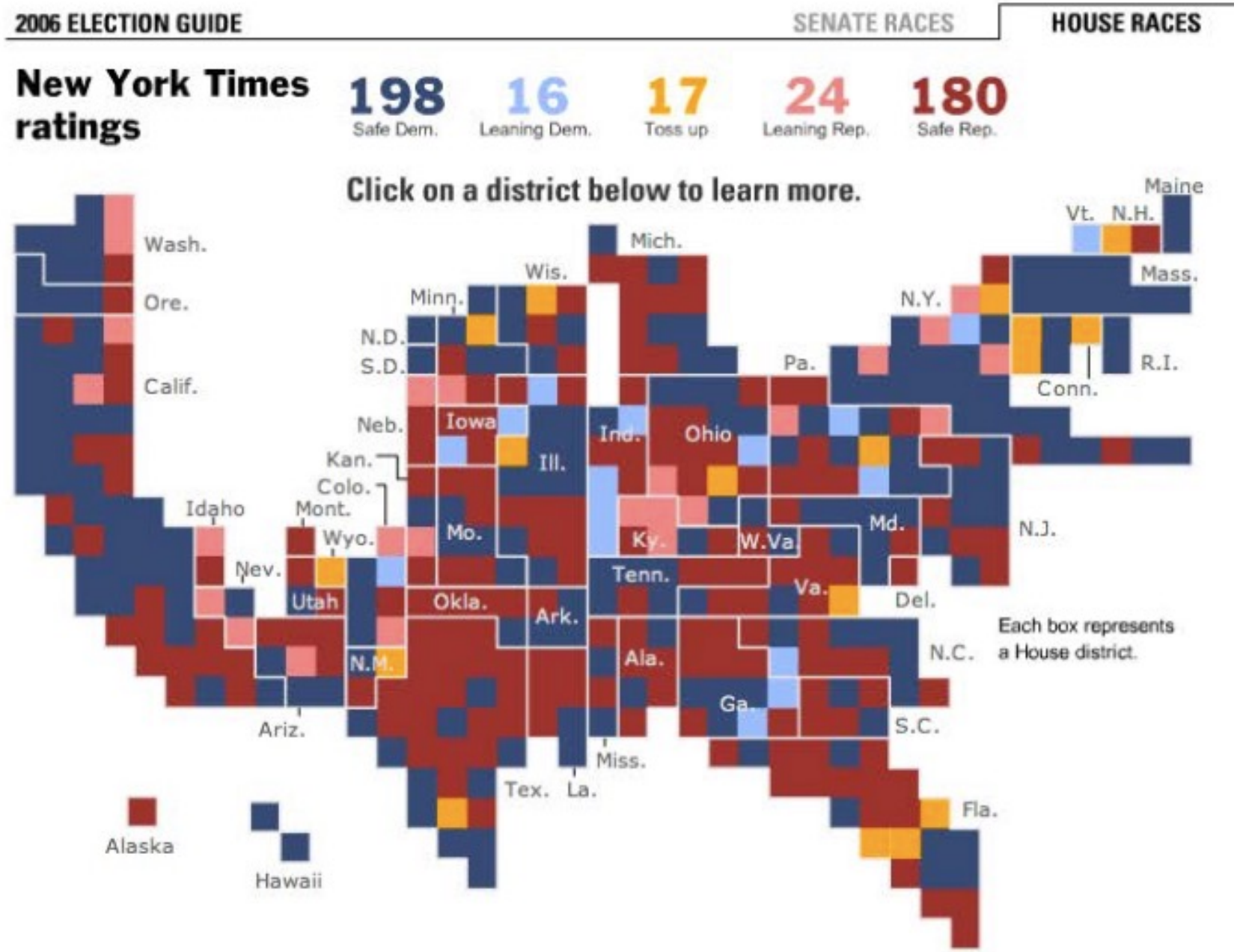
Adapted from slides of Meyer and Levine

Dorling cartogram of Olympic medal count



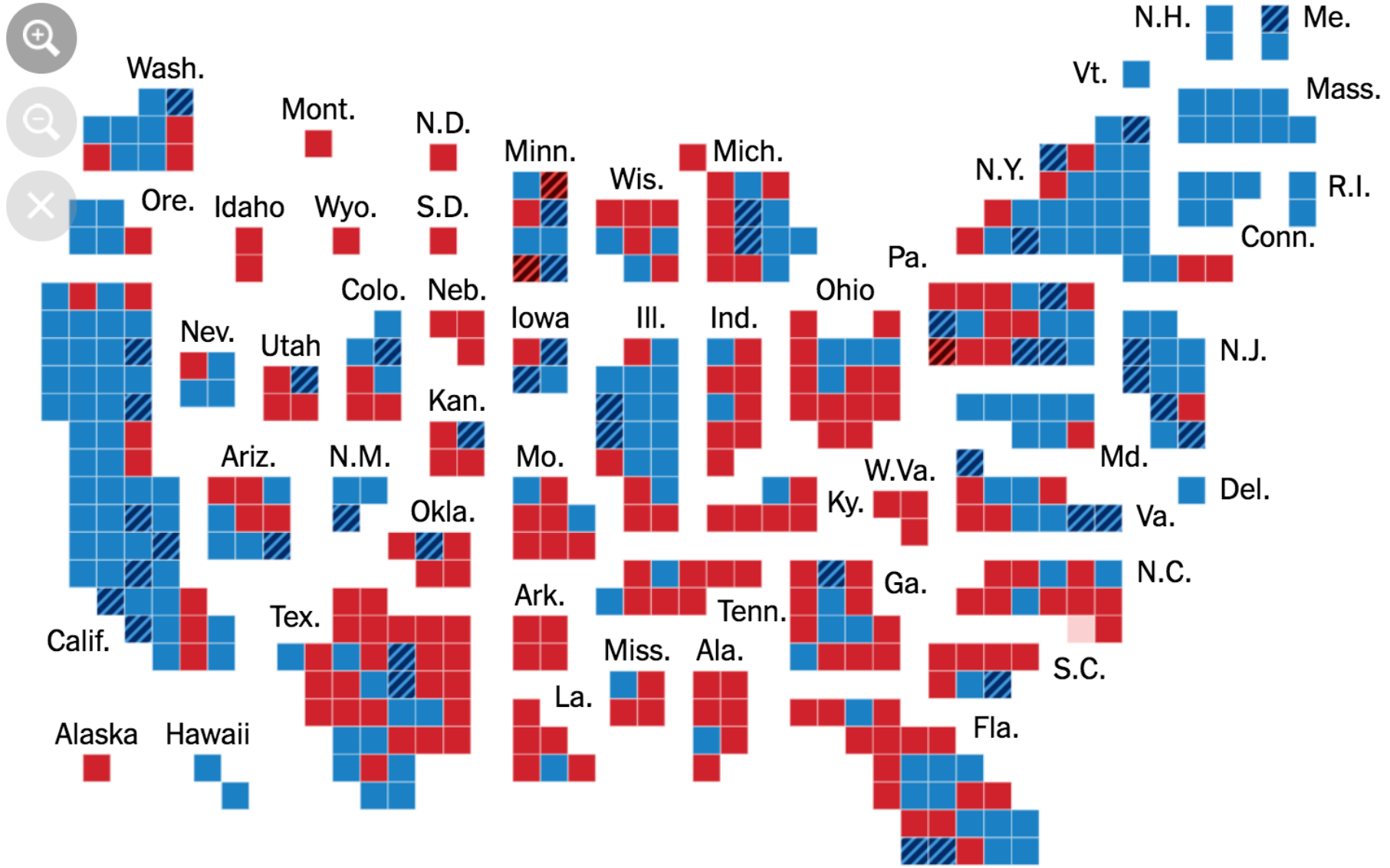
https://archive.nytimes.com/www.nytimes.com/interactive/2008/08/04/sports/olympics/20080804_MEDALCOUNT_MAP.html?_r=

Rectangular cartogram of by-county election predictions



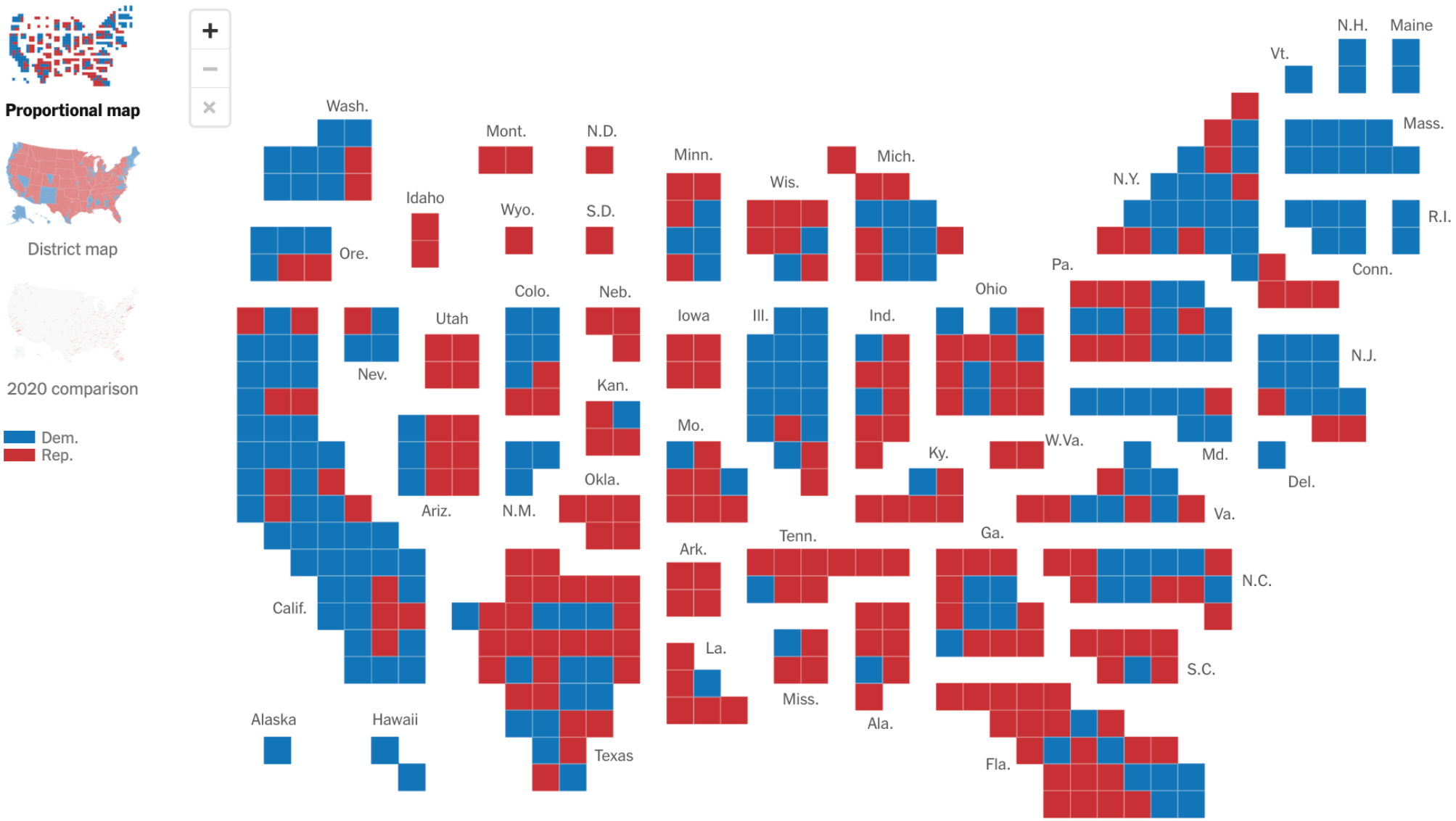
Adapted from slides of Meyer and Levine

12 years later…



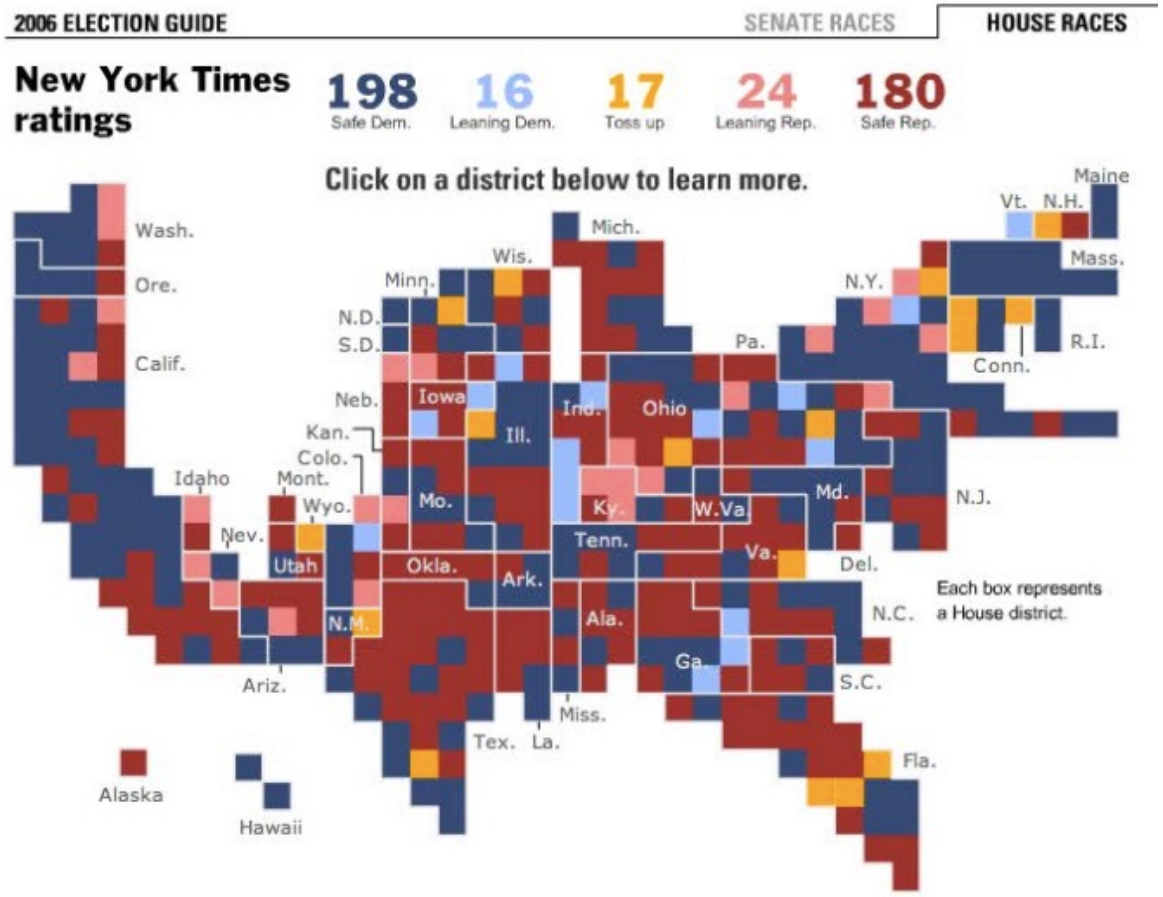
https://www.nytimes.com/interactive/2018/11/06/us/elections/results-house-elections.html

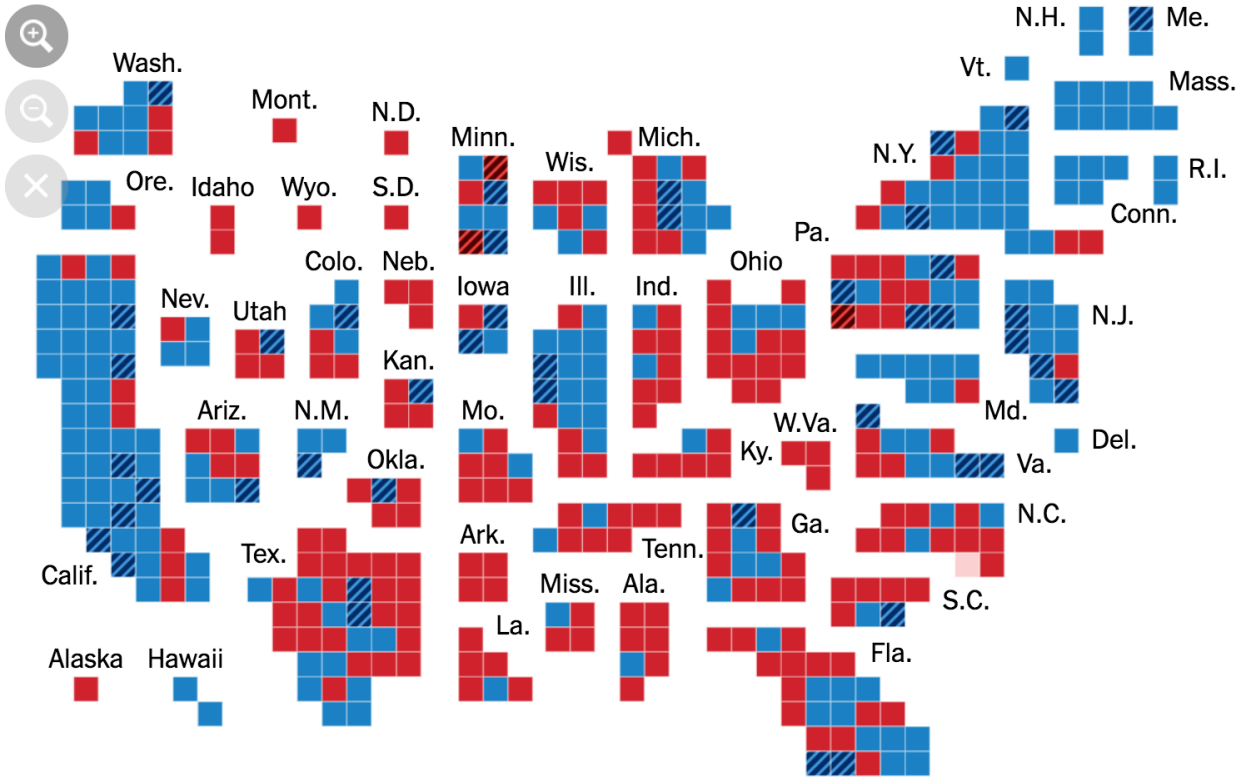
And in 2022…



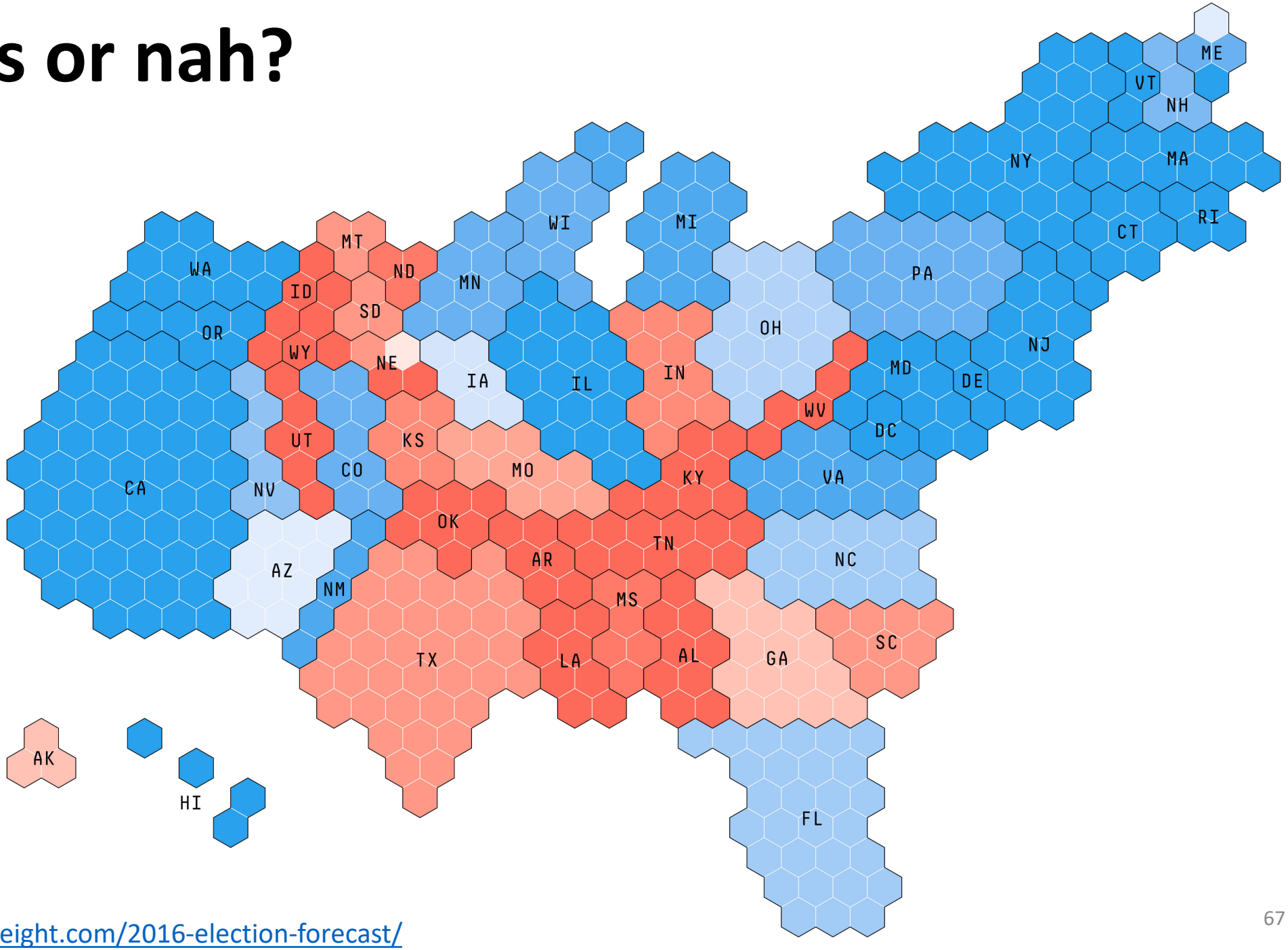
https://www.nytimes.com/interactive/2022/11/08/us/elections/results-house.html

Which do you prefer?





Hexagons or nah?

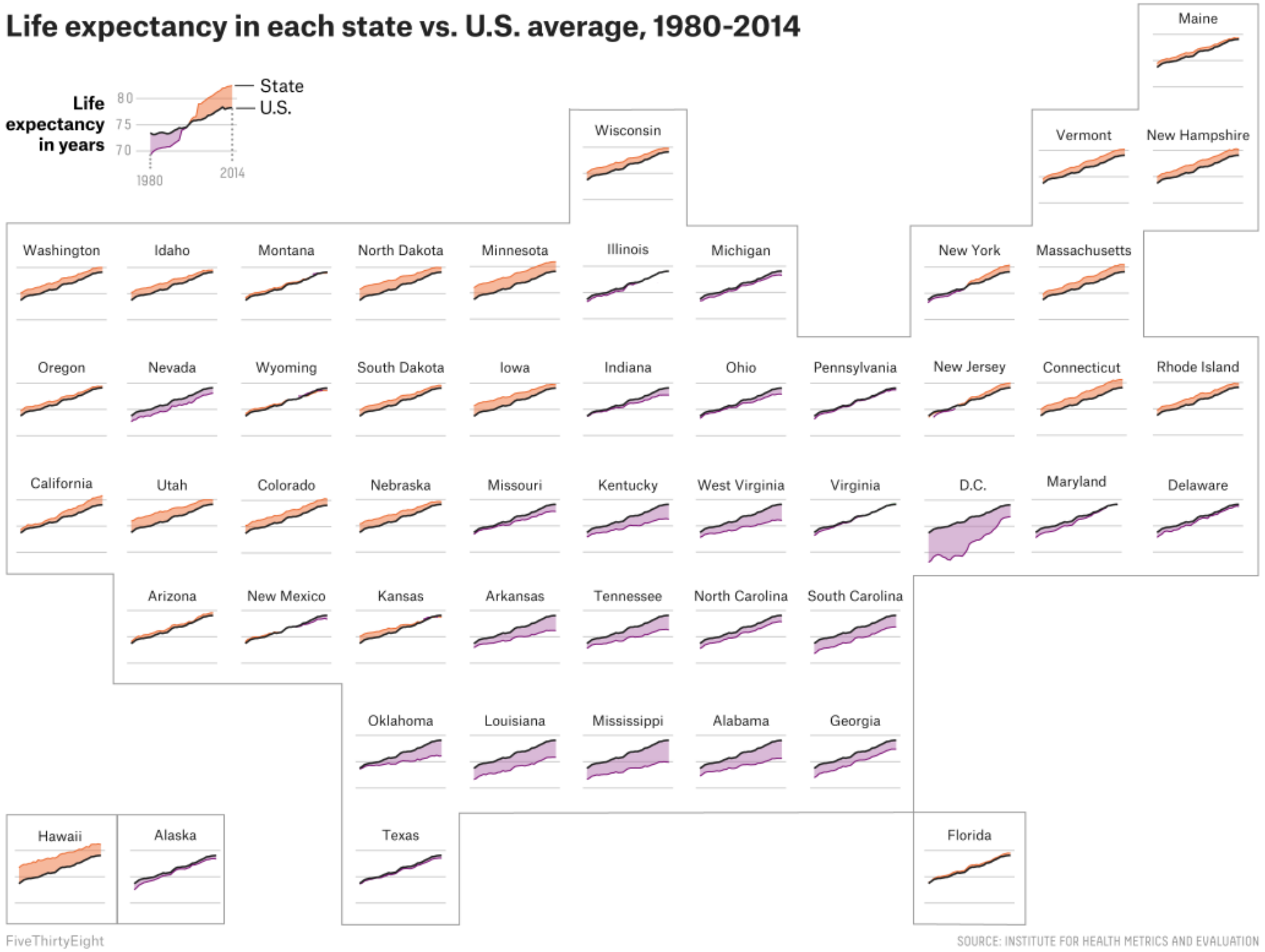


https://projects.fivethirtyeight.com/2016-election-forecast/

You could preserve shape

https://www.bouncymaps.com/#!/bouncymaps/world/-2102779806

Geo-grid + small multiple of charts



https://fivethirtyeight.com/features/as-u-s-life-expectancies-climb-people-in-a-few-places-are-dying-younger/