Reminders

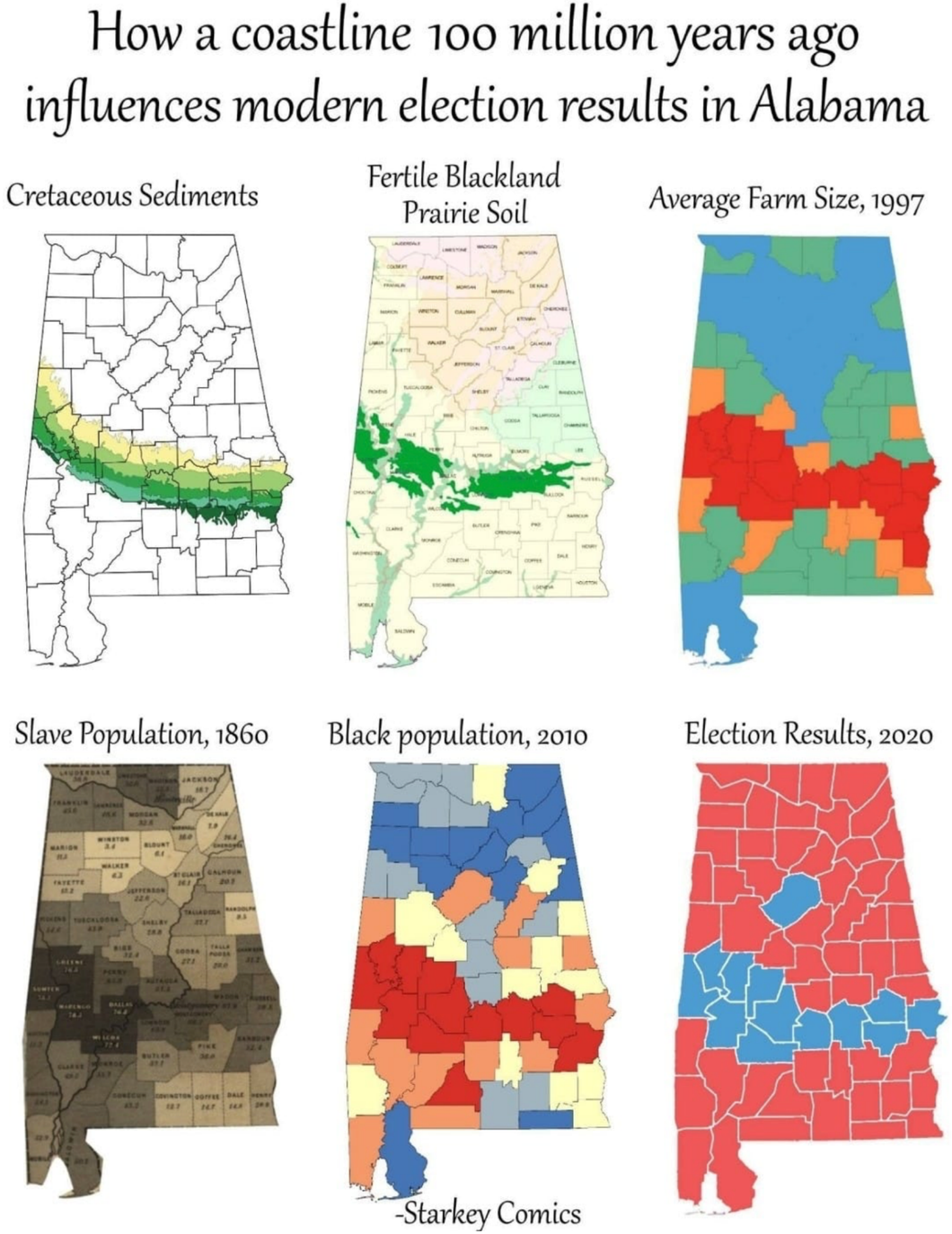
Upcoming due dates
Wed Oct 22nd Assignment 1
Thur Oct 23rd* Project review (1 per group)
Fri Oct 27th Discussion Lab 3

Repo invites: Click accept before it expires next week!
* - delayed due to Canvas outage on Monday

Geospatial: Maps as EDA

Data Science in Practice

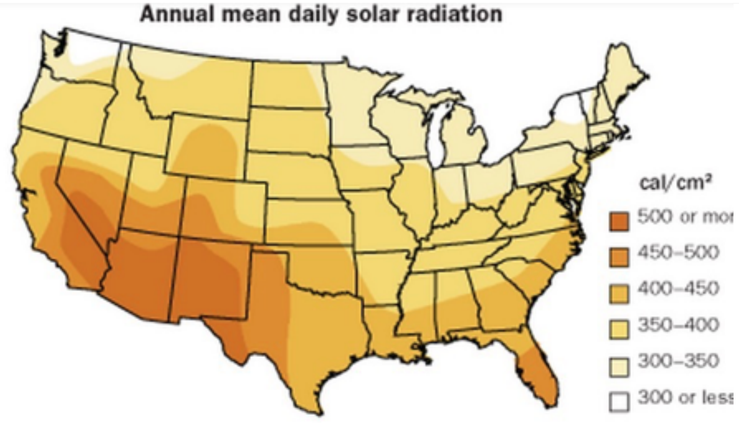
Jason G. Fleischer, PhD **Dept. of Cognitive Science** **UC San Diego** **https://jgfleischer.com**

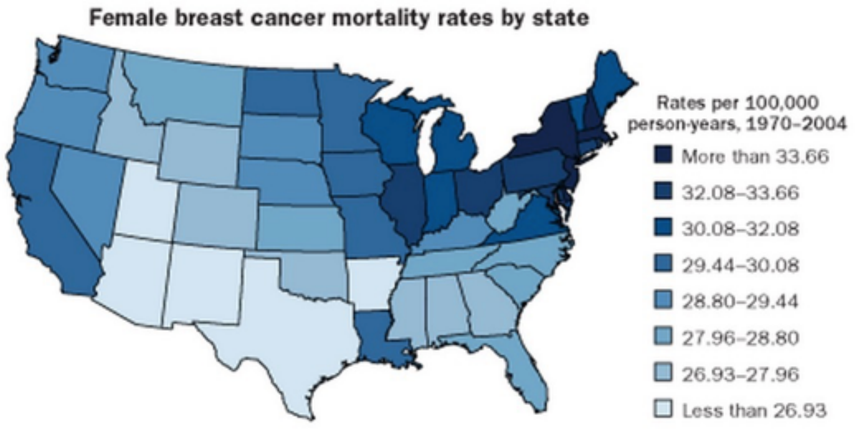


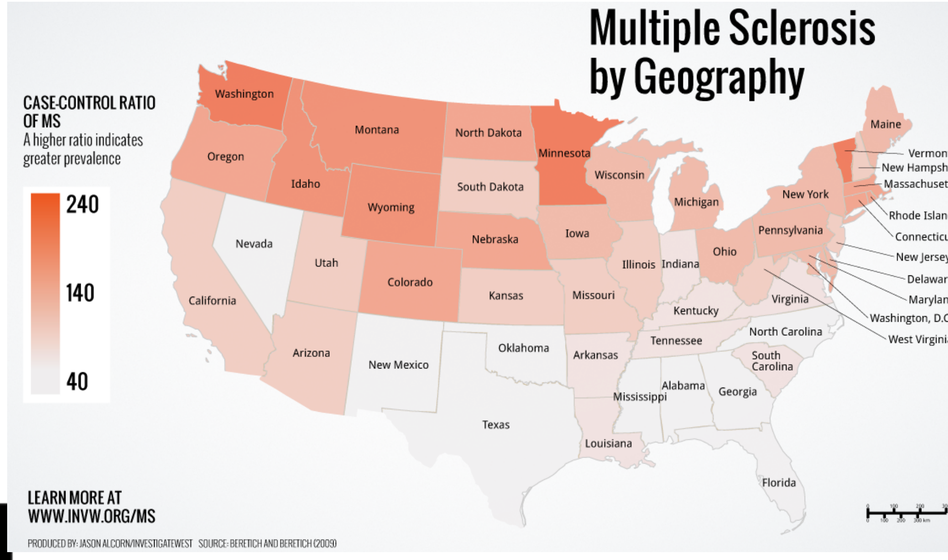
Why Geospatial Analysis?

‘Everything is related to everything else, but near things are more related than distant things.” -Tobler 1979

“...the purpose of geographic inquiry is to examine relationships between geographic features collectively and to use the relationships to describe the real-world phenomena that map features represent” -Clarke 2001



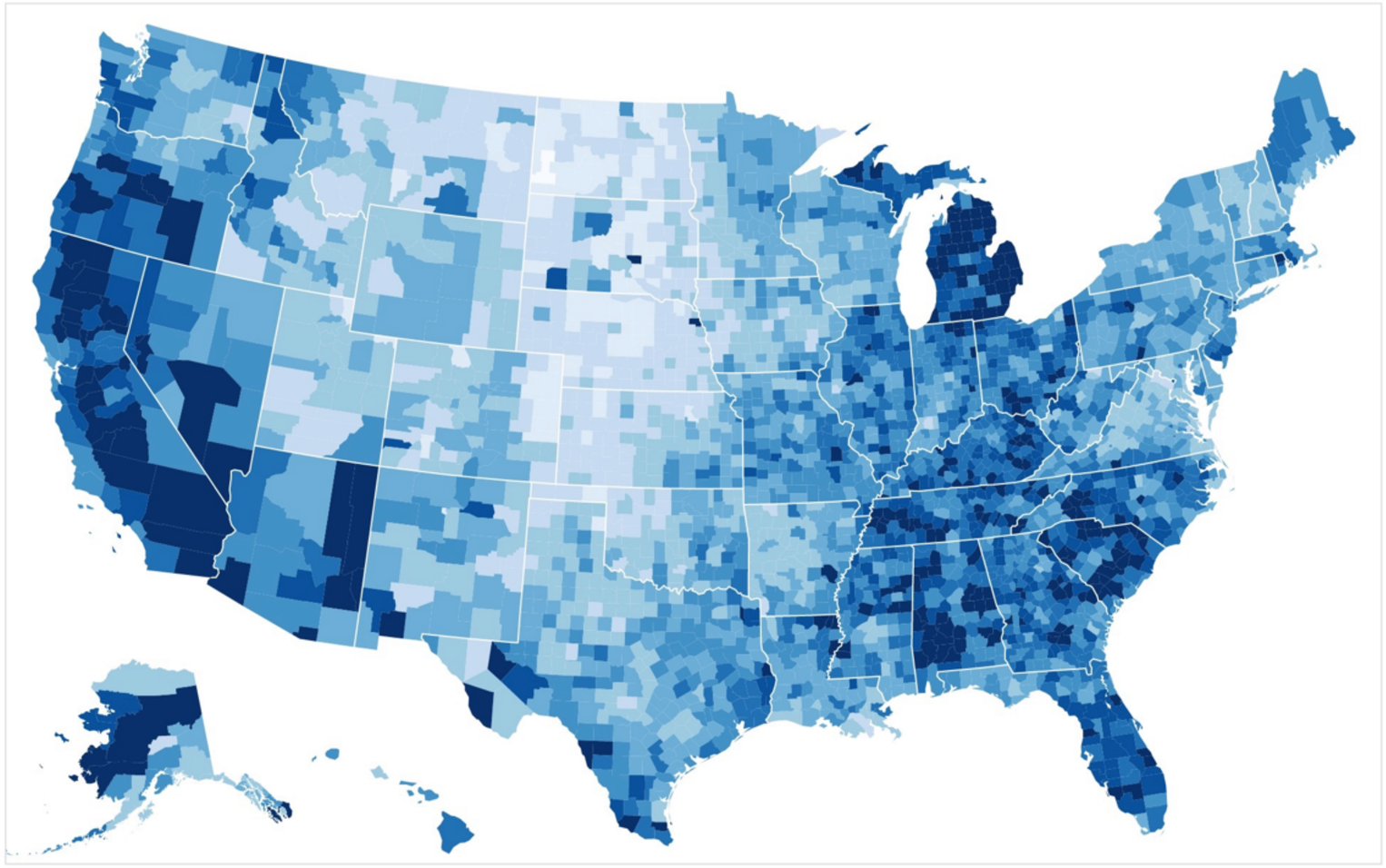




ON THE MAP Scientists who study vitamin D can't help but notice that a host of diseases seem to vary with latitude. Type 1 diabetes, multiple sclerosis and even some cancers appear to be more common in areas that get less sun -- meaning less opportunity for the body to produce vitamin D. The maps above illustrate the apparent link between solar radiation and breast cancer mortality rates.

SOURCE, FROM TOP: D. M. HARRIS AND V.L.W. GO / /J. OF NUTRITION 2004; NATIONAL CANCER INSTITUTE

Unemployment rate by county (August 2016)



This choropleth encodes unemployment rates from 2008 with a quantize scale ranging from 0 to 15%. A threshold scale is a useful alternative for coloring arbitrary ranges.

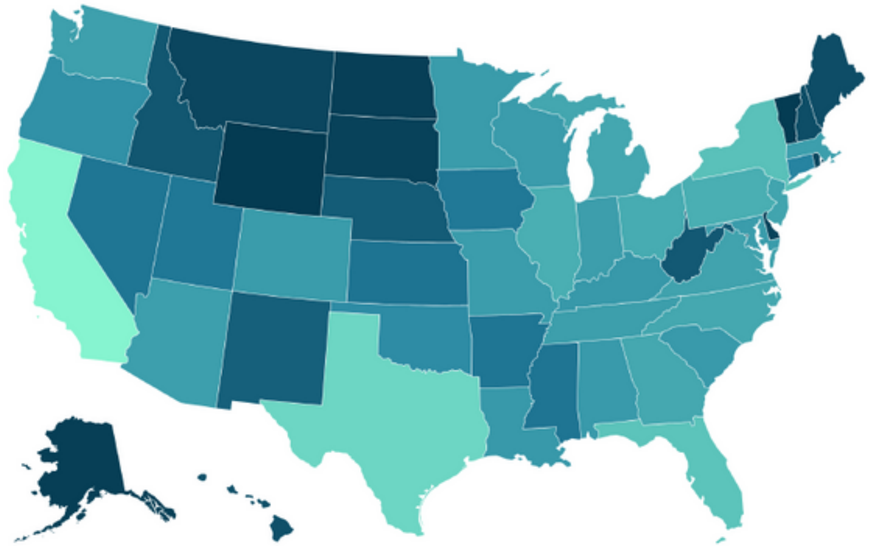
Open in a new window.

Choropleth maps are useful for visualizing *clear regional patterns* in the data

http://bl.ocks.org/mbostock/4060606

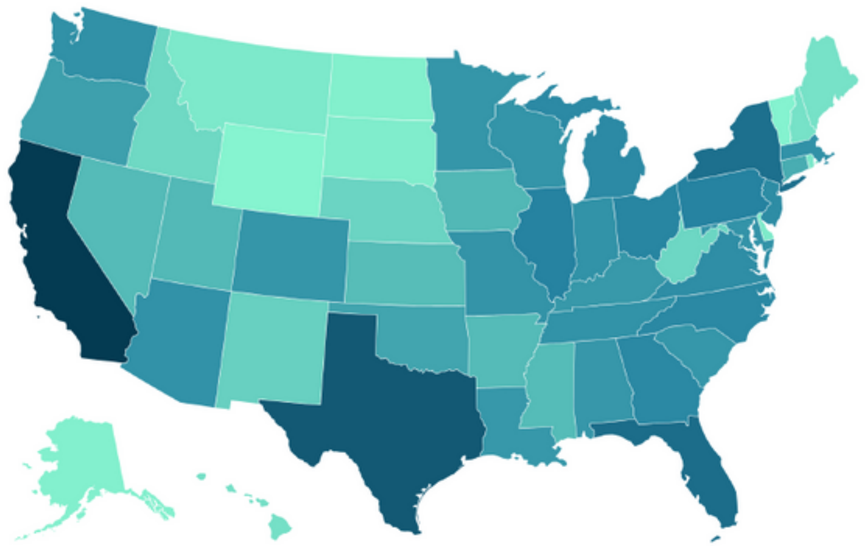
Use light colors for low values. Dark colors for high values.

NOT IDEAL



LOW POPULATION HIGH

BETTER

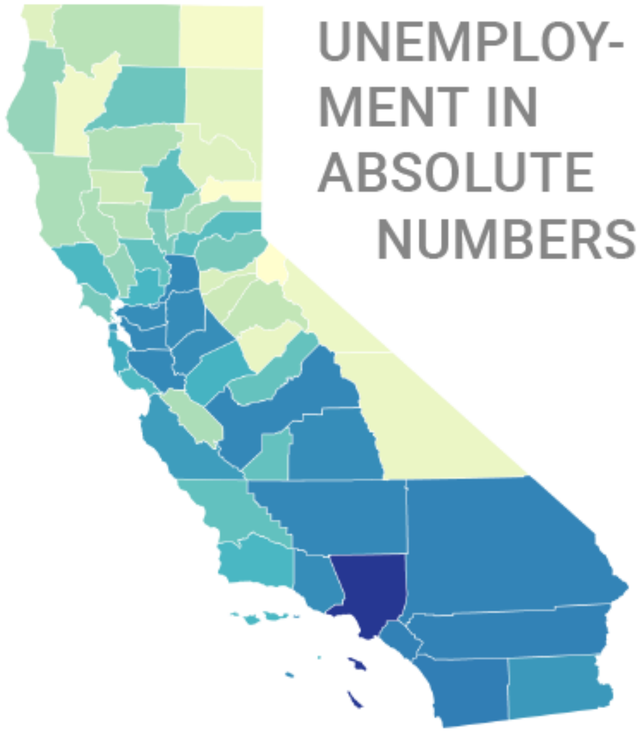


LOW POPULATION HIGH

https://blog.datawrapper.de/colors/

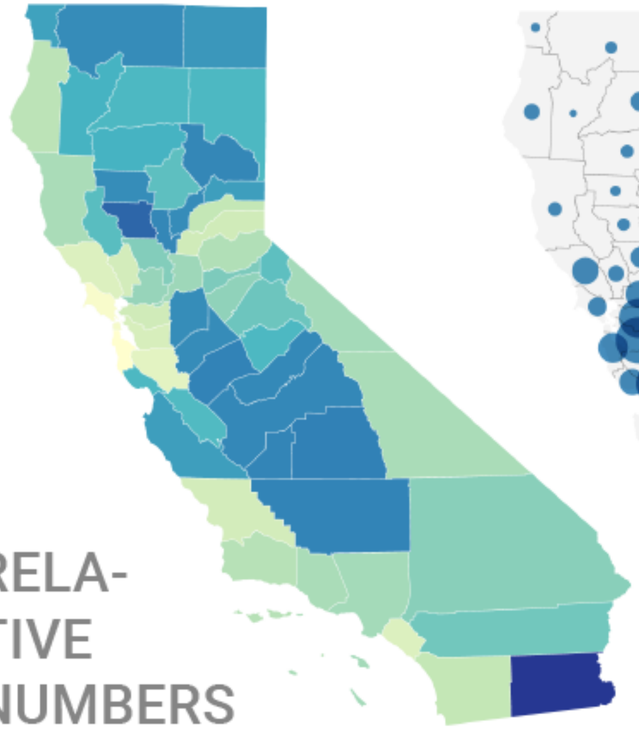
Choropleth should display relative differences, *not* absolute numbers

NOT IDEAL

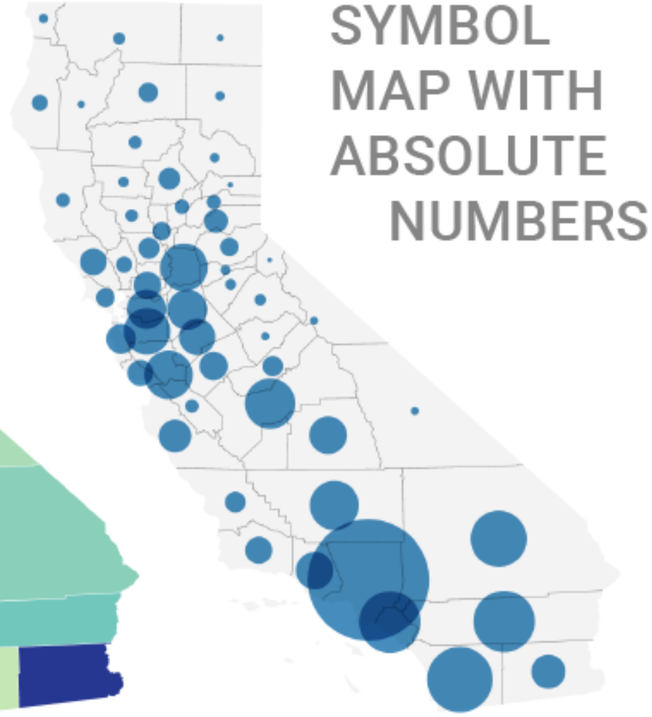


UNEMPLOY-
MENT IN
ABSOLUTE
NUMBERS

BETTER



RELA-
TIVE
NUMBERS

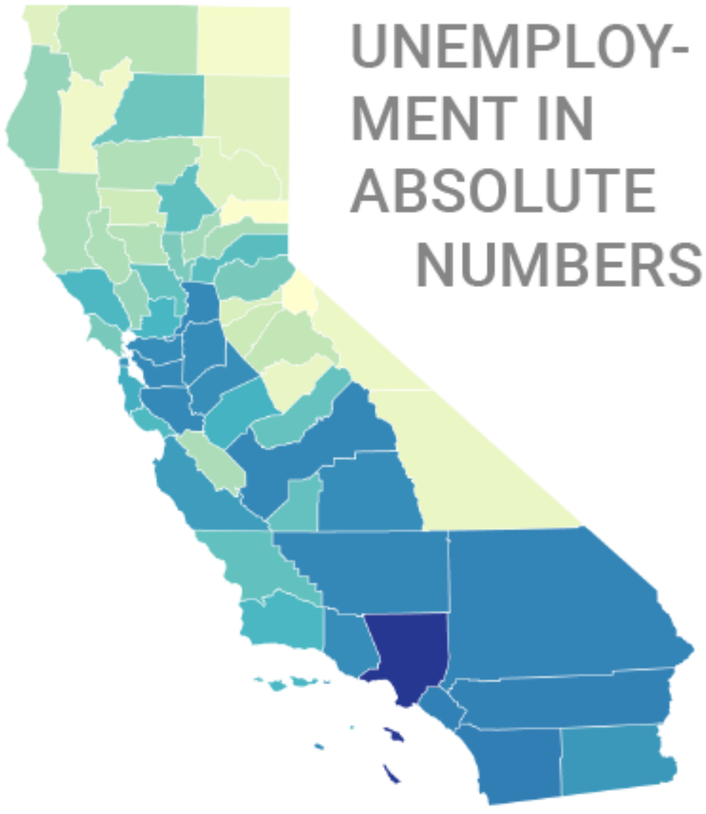


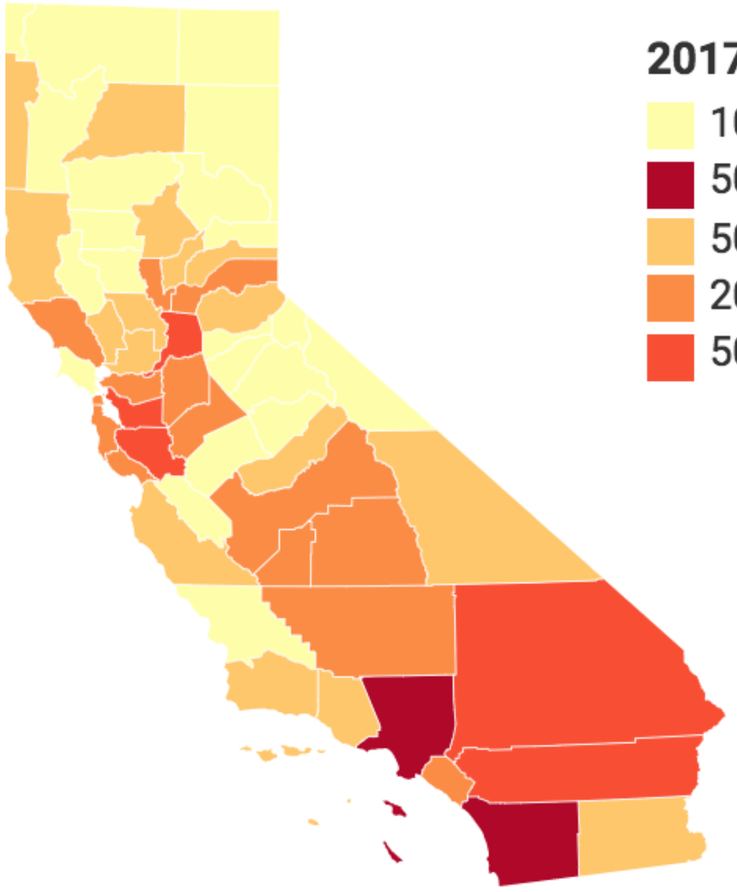
SYMBOL
MAP WITH
ABSOLUTE
NUMBERS

https://blog.datawrapper.de/choroplethmaps/

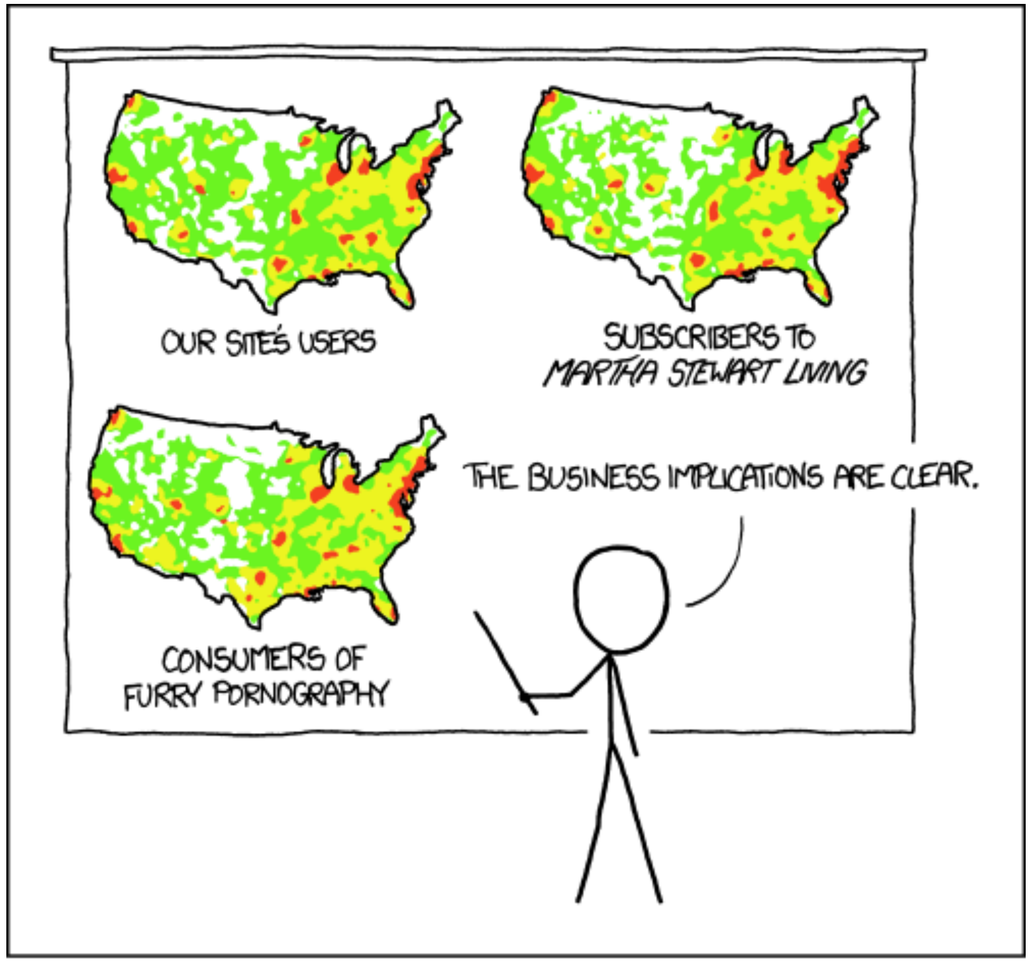
Map: Where Are Students Attending Charter Schools?

The majority of California's charter school student population is concentrated in Los Angeles, San Diego and Bay Area counties. Hover through the counties on each map for more information on their





https://calmatters.org/articles/california-charter-school-data-enrollment-cost-teachers-unions/



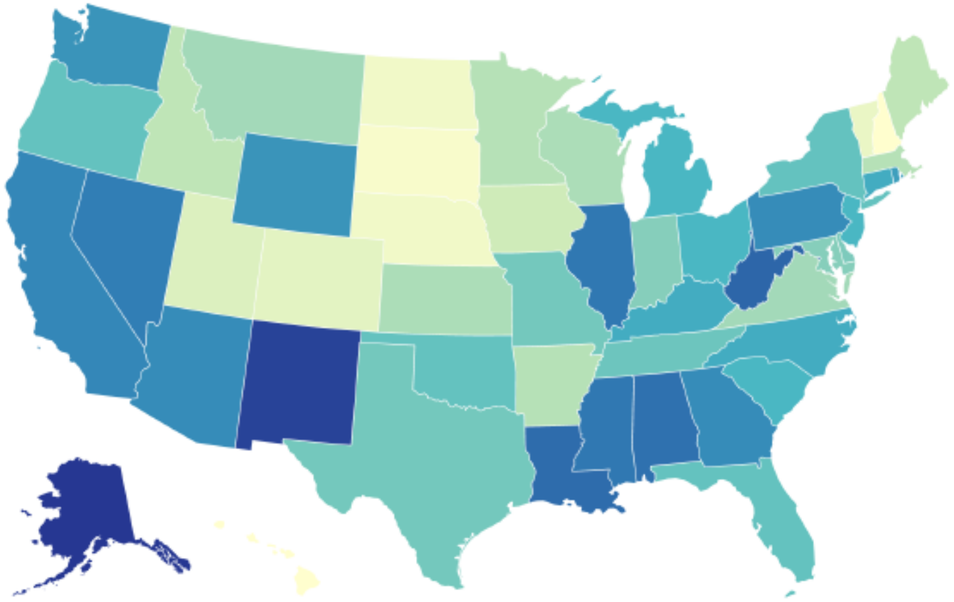
Choropleth maps can be misleading

https://xkcd.com/1138/

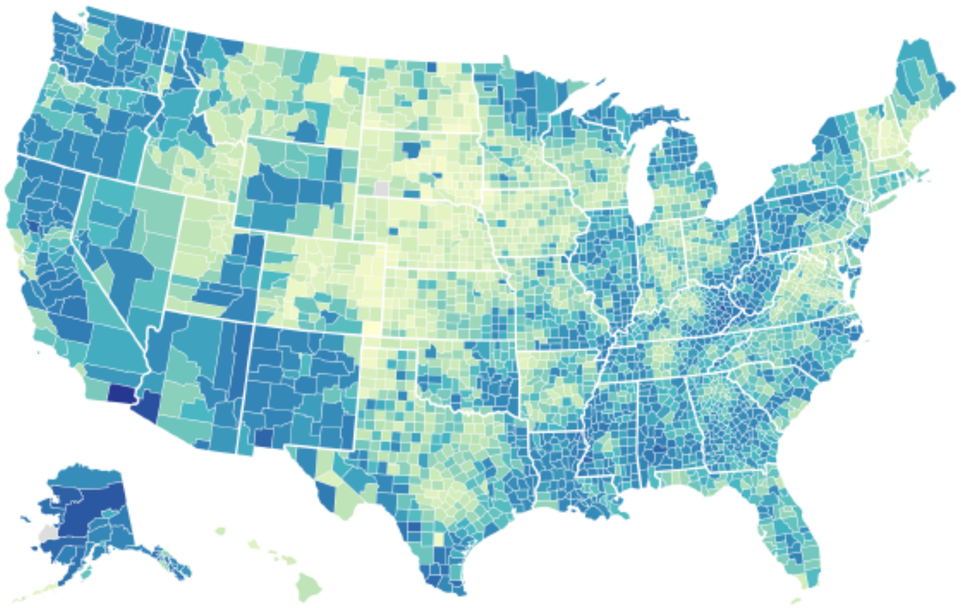
Consider using the smallest unit possible

(but there are exceptions!)

NOT IDEAL

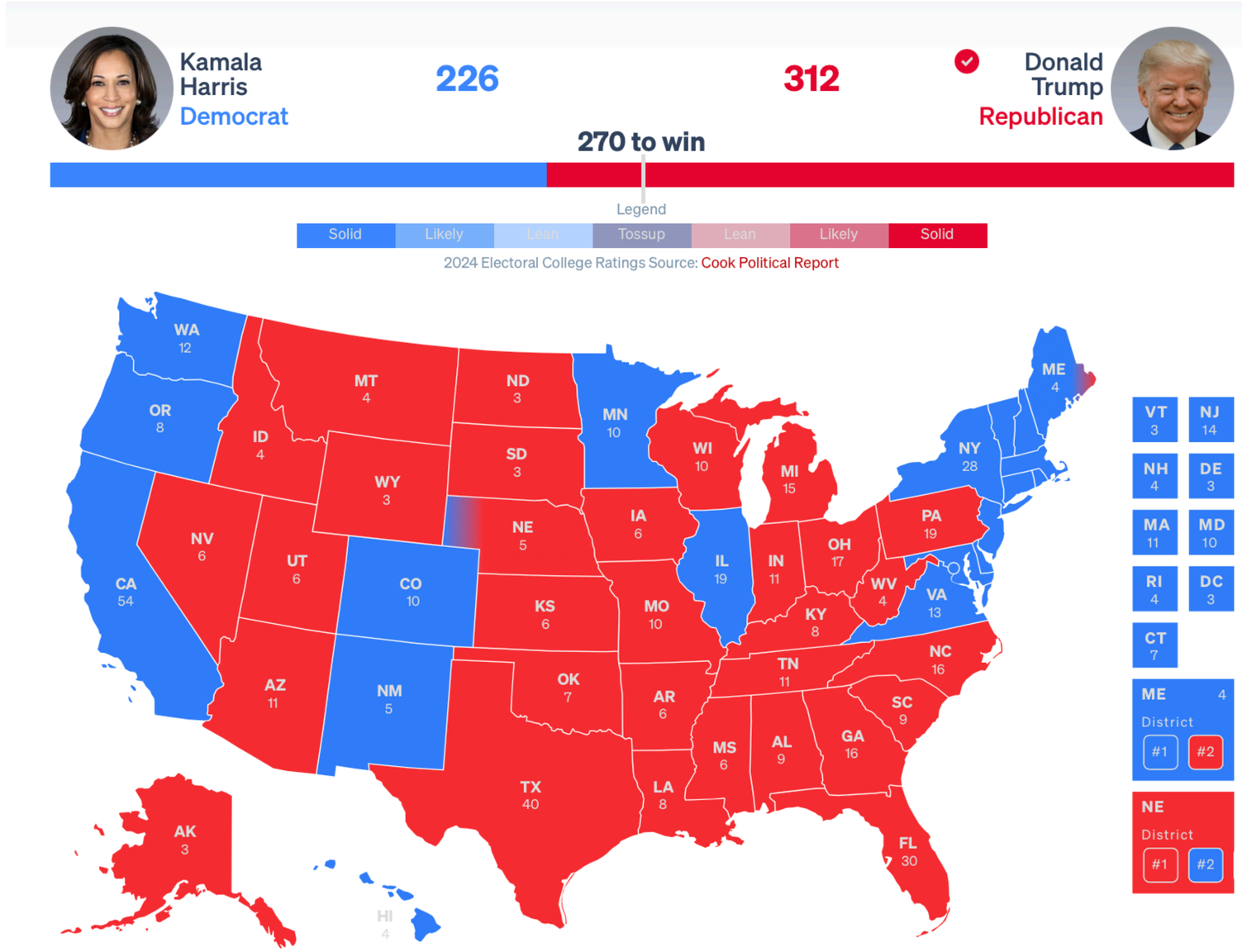


BETTER



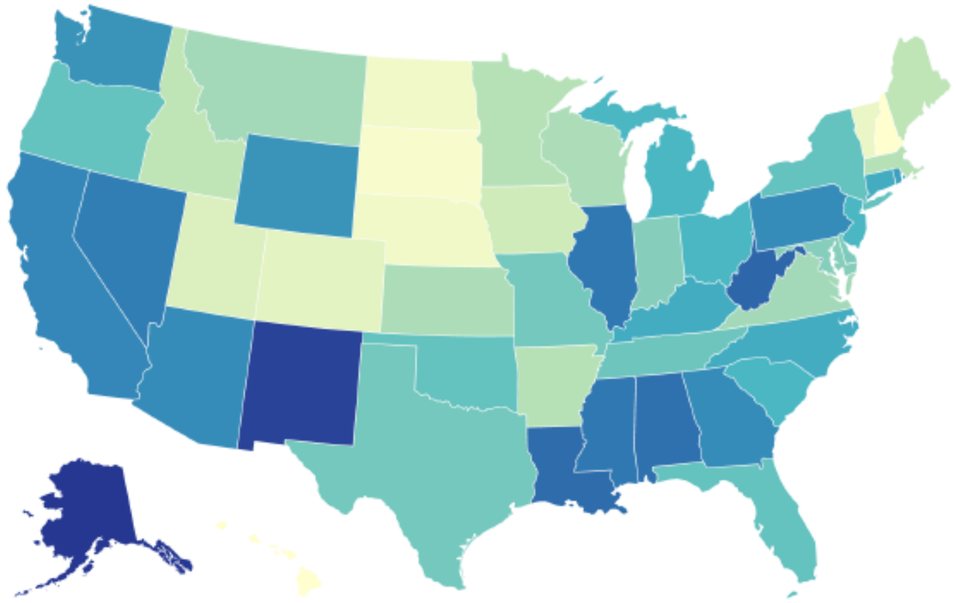
https://blog.datawrapper.de/choroplethmaps/

Sometimes summarizing at the state level is OK



Cartograms should be considered when displaying how many people were affected

NOT IDEAL

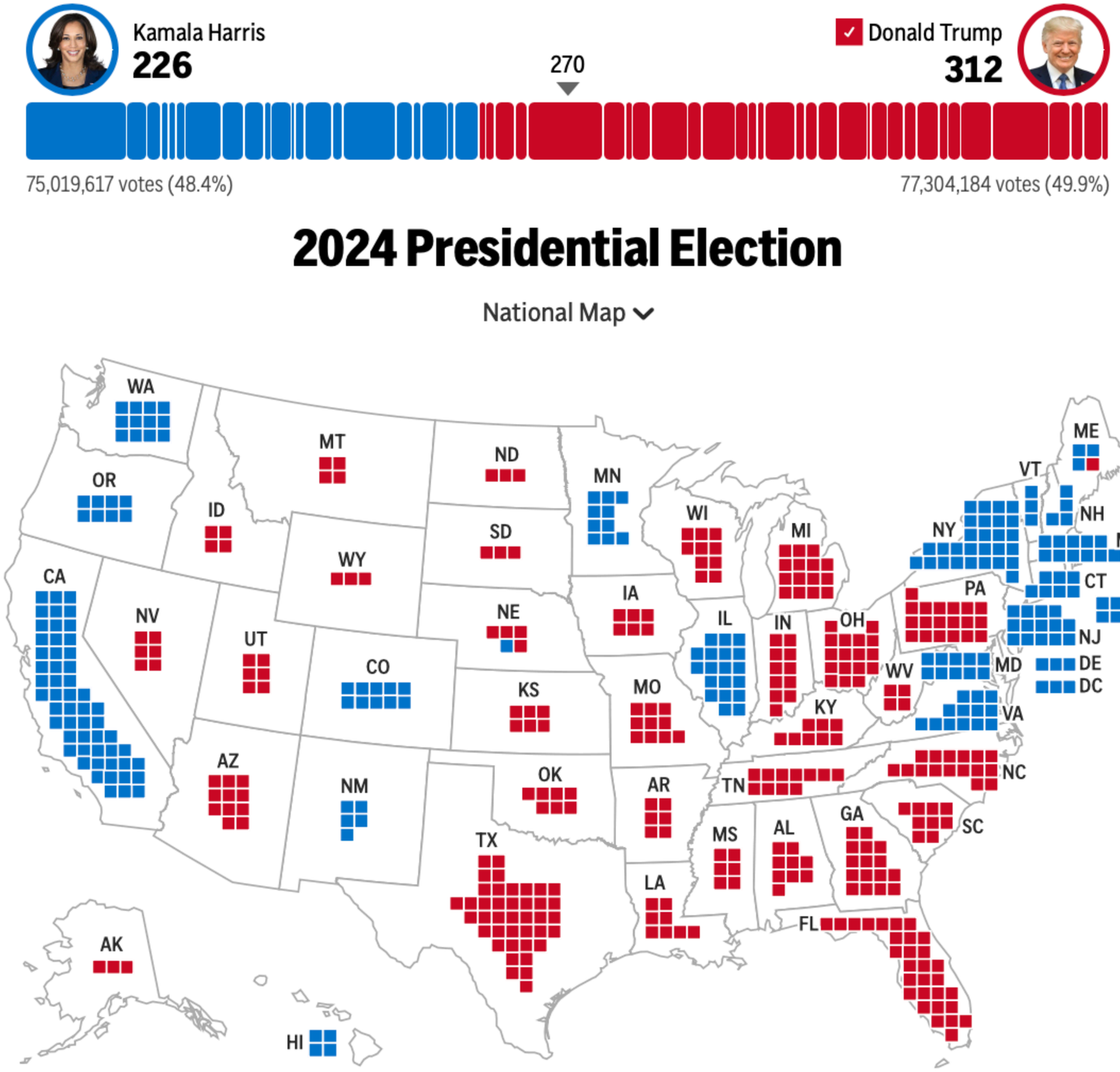


Choropleths answer “How much area was affected?”

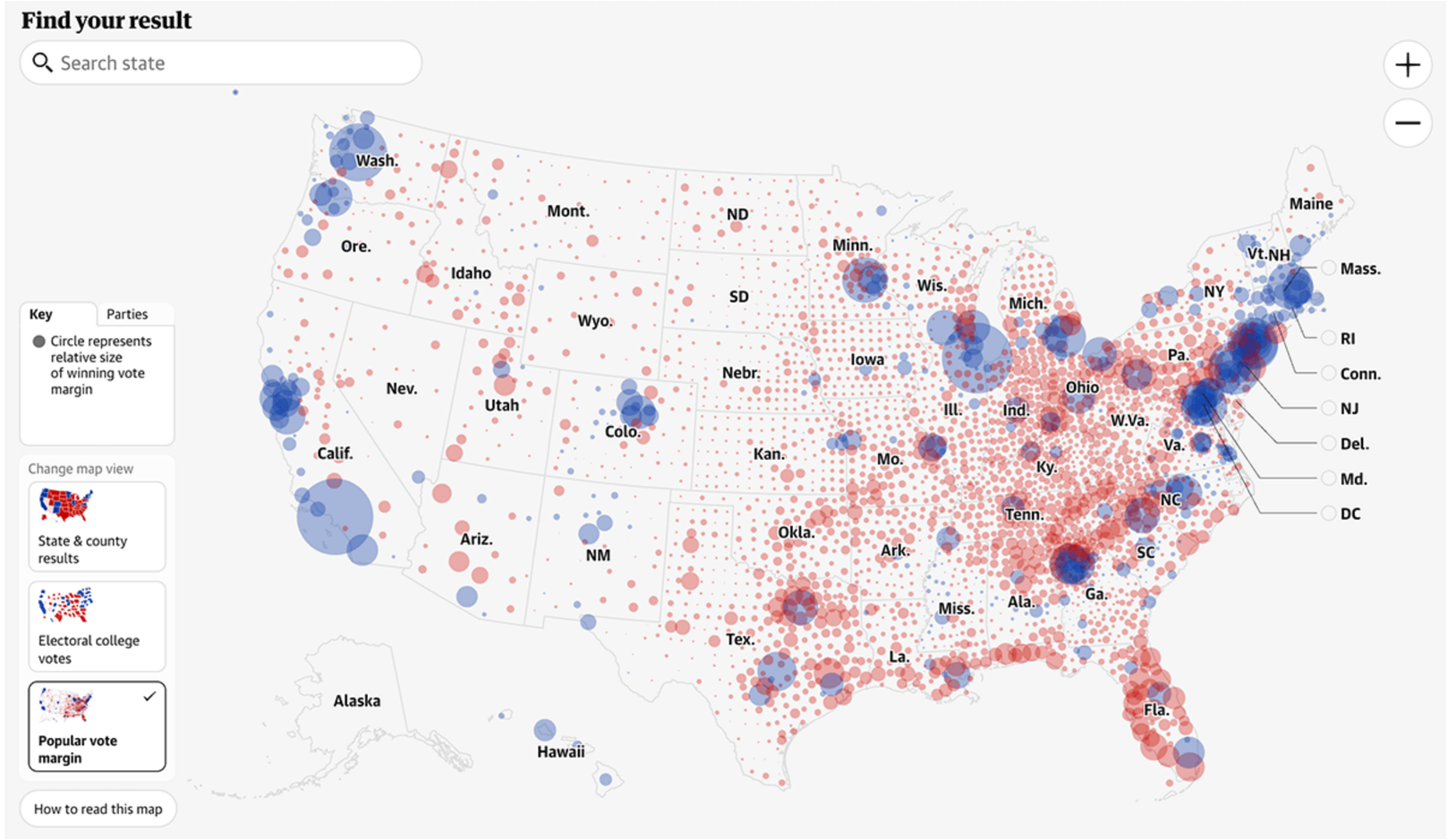
BETTER

Cartograms answer “How many people were affected?”

https://blog.datawrapper.de/choroplethmaps/



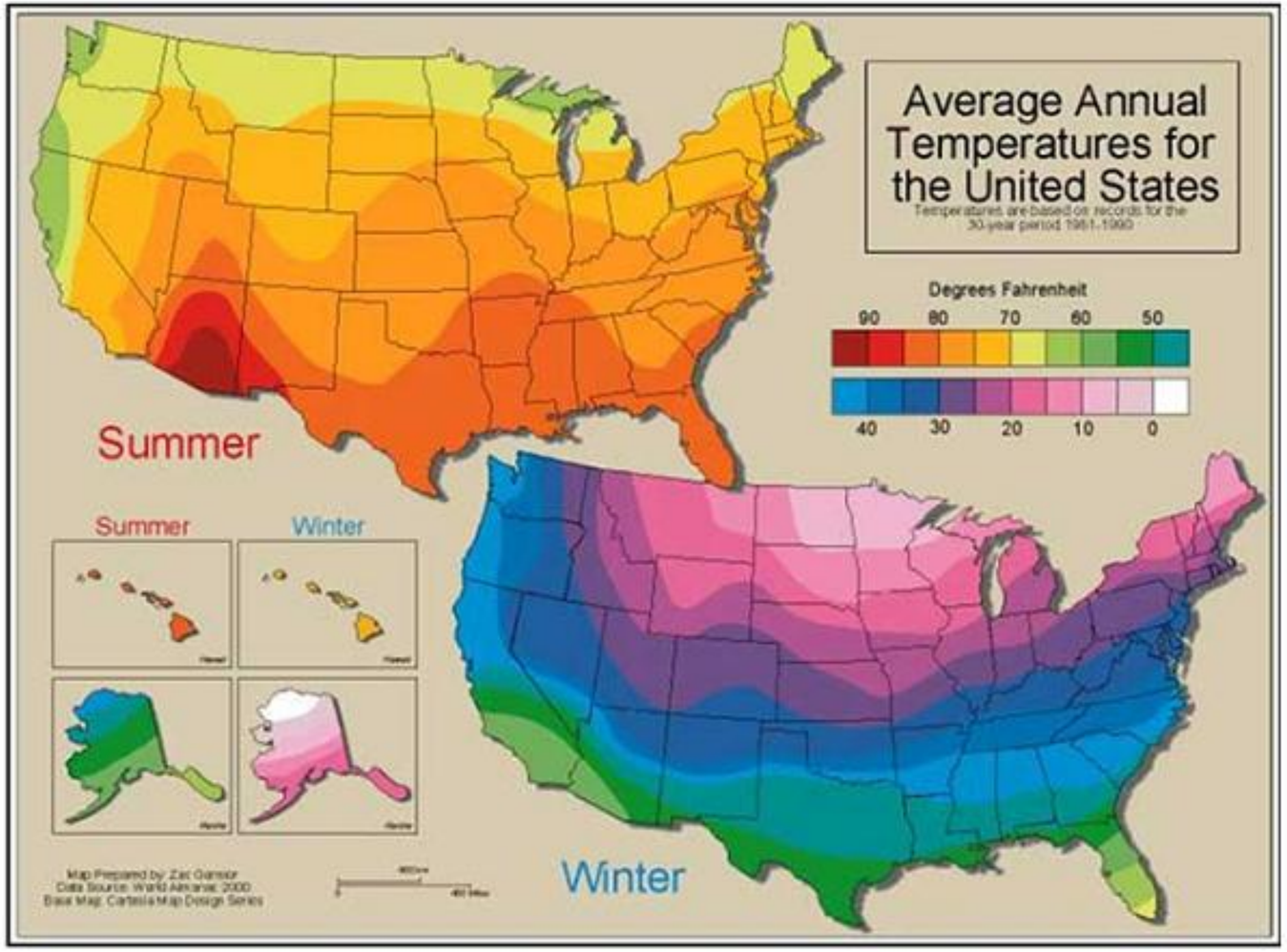
This **cartogram** more accurately shows how the populace is split since the number of tiles reflects state populations

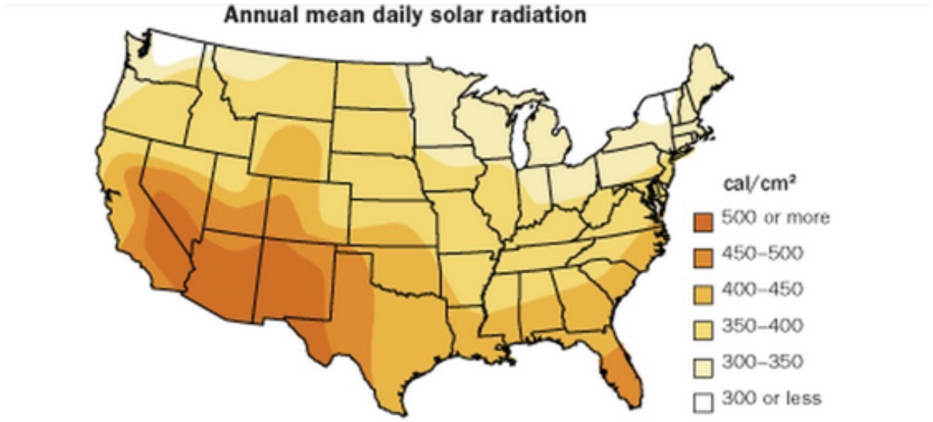


This county level **bubble graph** more accurately tells the full story, since the size of the bubbles reflects population while color shows the political split in that county

Isarithmic maps demonstrate smooth, continuous phenomena

(temperature, elevation, rainfall, etc.)





Isarithmic maps demonstrate smooth, continuous phenomena

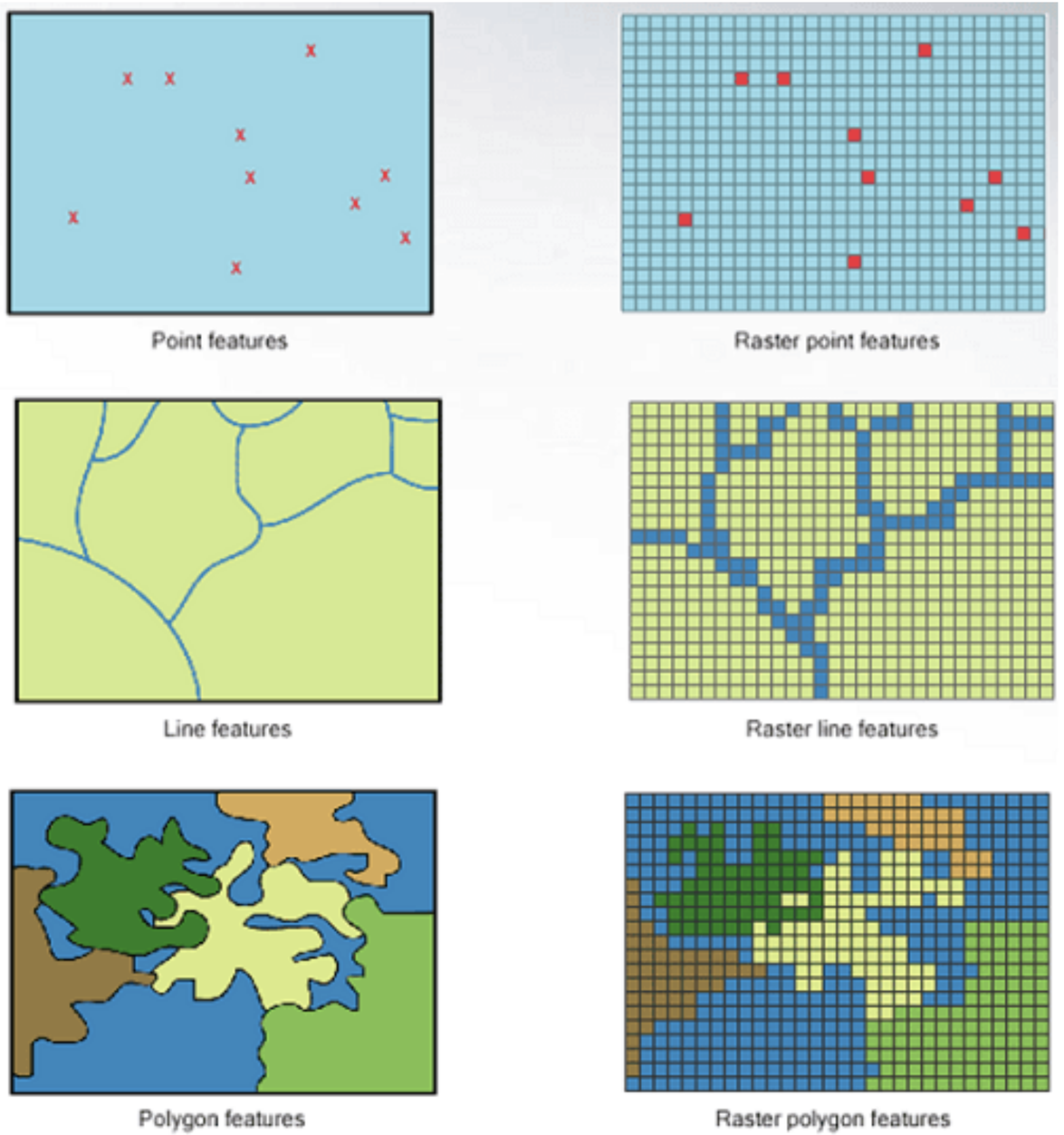
(temperature, elevation, rainfall, etc.)

Representing spatial data and making maps like these

Representing shapes on maps

- Vector data
  - Points, lines, polygons
- Raster data
  - Encodes the world as a continuous surface represented by a grid, such as the pixels of an image.

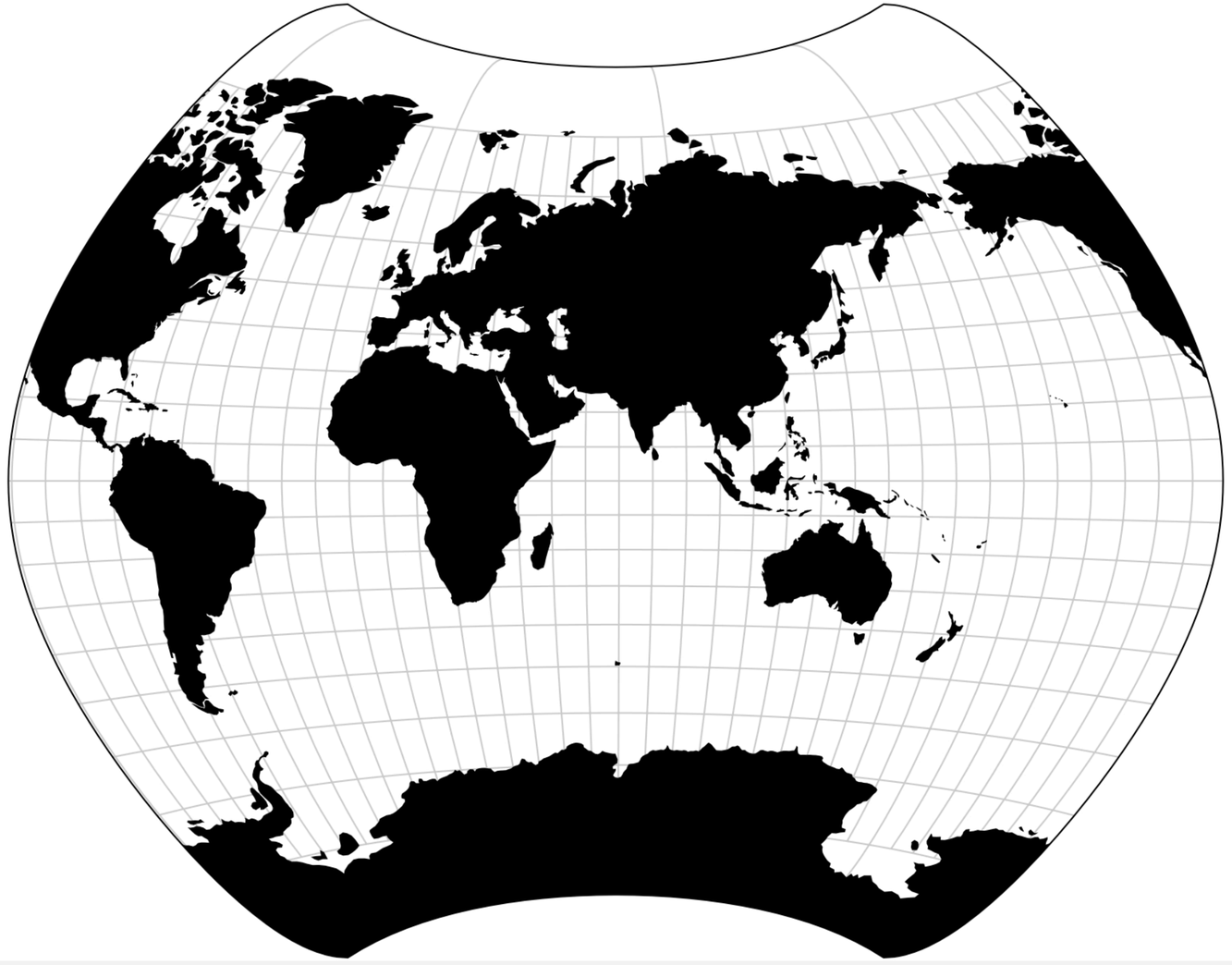
Source: Humboldt State University



Projection Transitions

This notebook interpolates smoothly between projections; this is easiest when both projections are well-defined over the given viewport (here, the world).

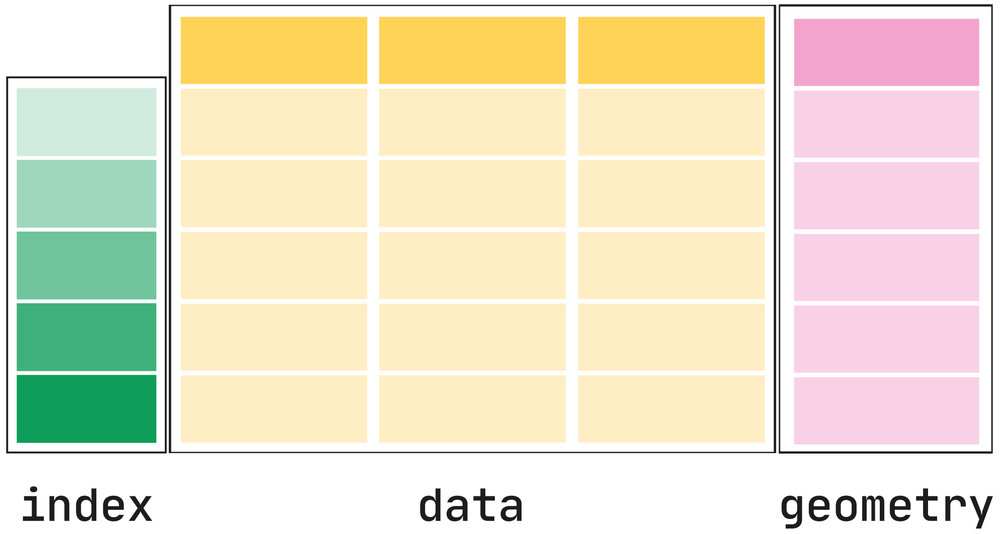
Laskowski tri-optimal *projection*



Source: https://observablehq.com/@d3/projection-transitions

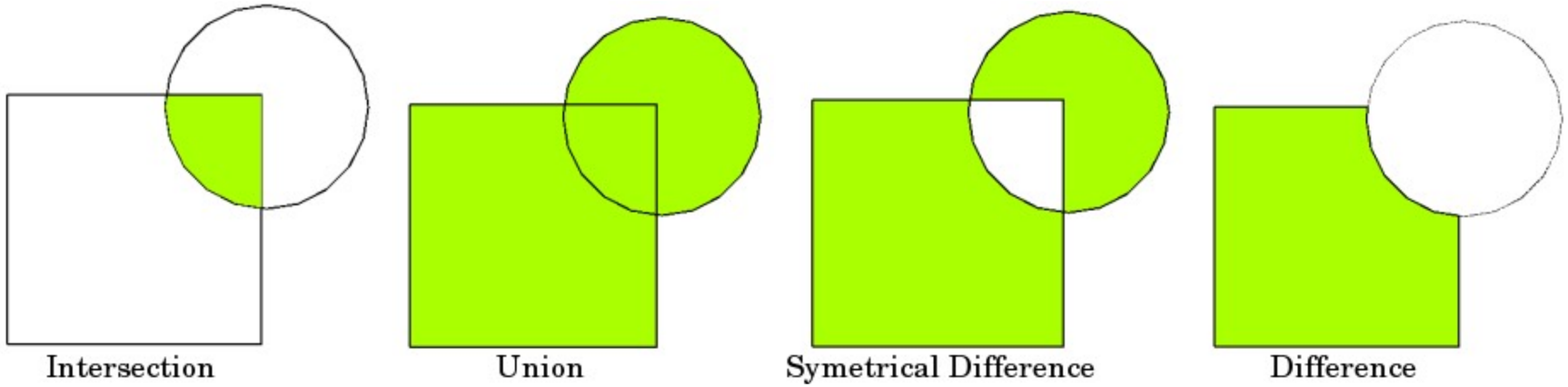
Geopandas

- Extends pandas, adds support for geospatial data
- Uses shapely library to represent vector geometry (e.g., POINT, LINE, POLYGON)
- Writes/reads shape files, GeoJSON and other formats
- Uses Coordinate Reference Systems (CRS) to represent how data aligns with the real world
- Different types of CRS: Geographic Coordinate Systems and Projected Coordinate Systems.



Geopandas skills

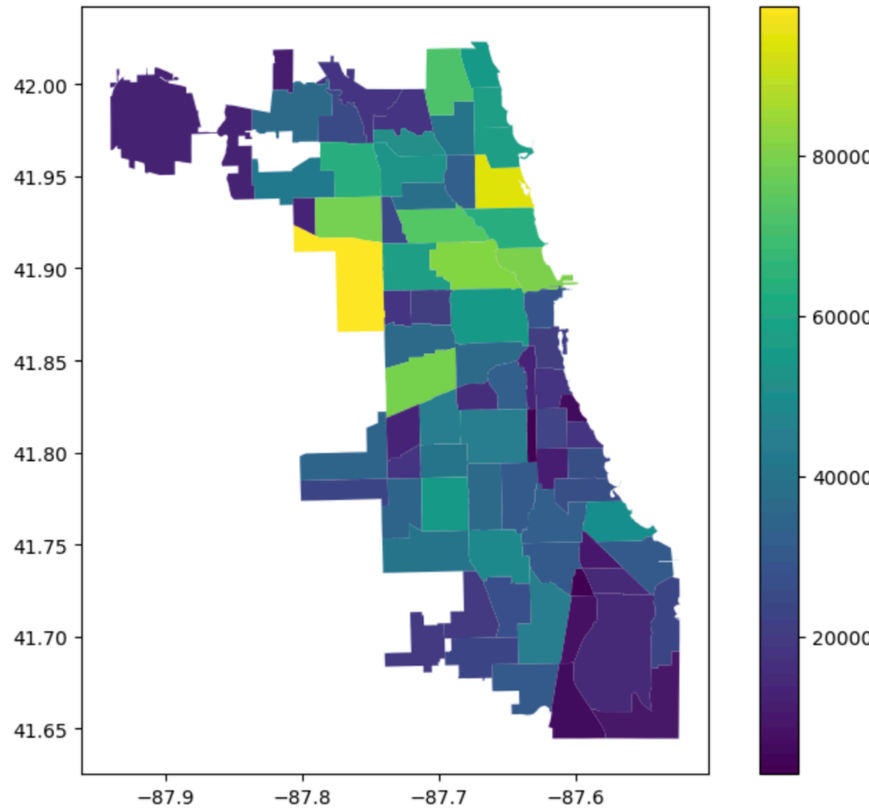
- Overlay spatial datasets: e.g. calculate what regions of Chicago are within 1km of a grocery store
- Join dataframes on shape information
- Aggregate across different subregions to get measurements at a more granular regional level



Geopandas tutorials

```
# Plot population estimates with an accurate legend
In [7]: chicago.plot(column='POP2010', legend=True);
```

- https://geopandas.org/en/stable/getting_started/introduction.html
- https://geopandas.org/en/stable/docs/user_guide/mapping.html
- https://geopandas.org/en/stable/docs/user_guide/set_operations.html
- https://geopandas.org/en/stable/gallery/index.html



Visualizing Geospatial Da

You want to visuali
been affected b

Best approach t

https://forms.gle/8iSWt8MdJdfM7PC69

Spatial Statistics : The Why

Spatial Statistics

The statistical techniques we've discussed so far don't work well when considering spatial distributions…

Spatial Statistics

The statistical techniques we've discussed so far don't work well when considering spatial distributions…

...which means we have a chance to take a look at data and the relationship between the data in new and interesting ways (distance, adjacency, interaction, and neighbor)

Spatial data violate conventional statistics:

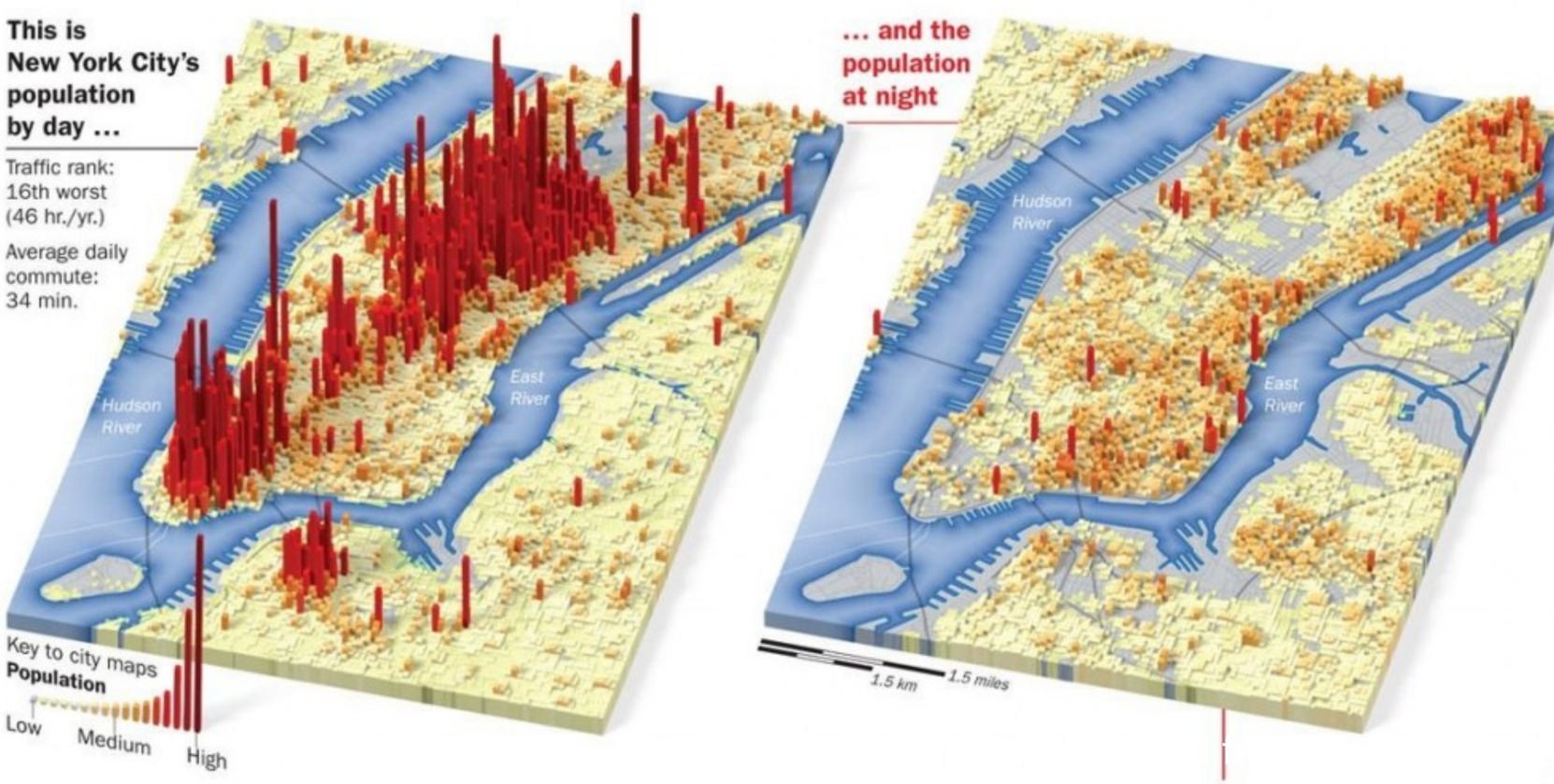
Violations of conventional statistics:

- Spatial autocorrelation
- Modifiable areal unit problem (MAUP)
- Edge effects (Boundary problem)
- Ecology fallacy
- Nonuniformity of space

Spatial Autocorrelation

Data from locations near one another in space are more likely to be similar than data from locations remote from one another:

- Housing market
- Elevation change
- Temperature



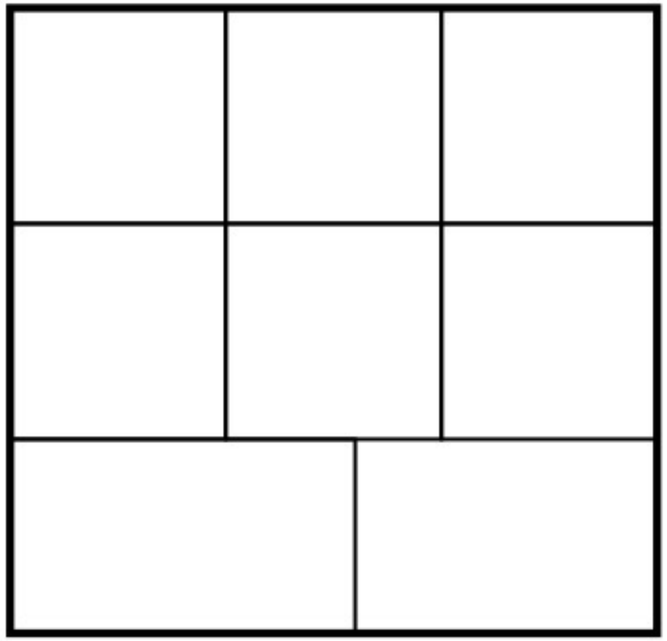
Modifiable Areal Unit Problem (MAUP)

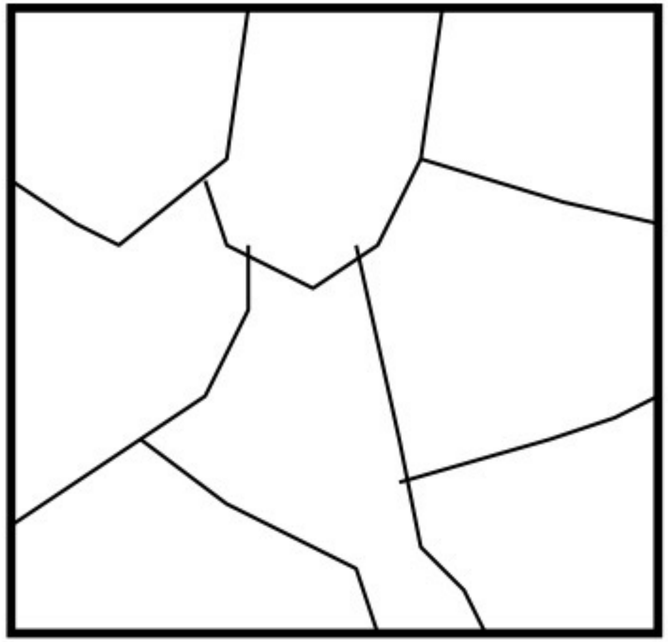
The aggregation units used are arbitrary with respect to the phenomena under investigation, yet the aggregation units used will affect statistics determined on the basis of data reported in this way.

If the spatial units in a particular study were specified differently, we might observe very different patterns and relationships.

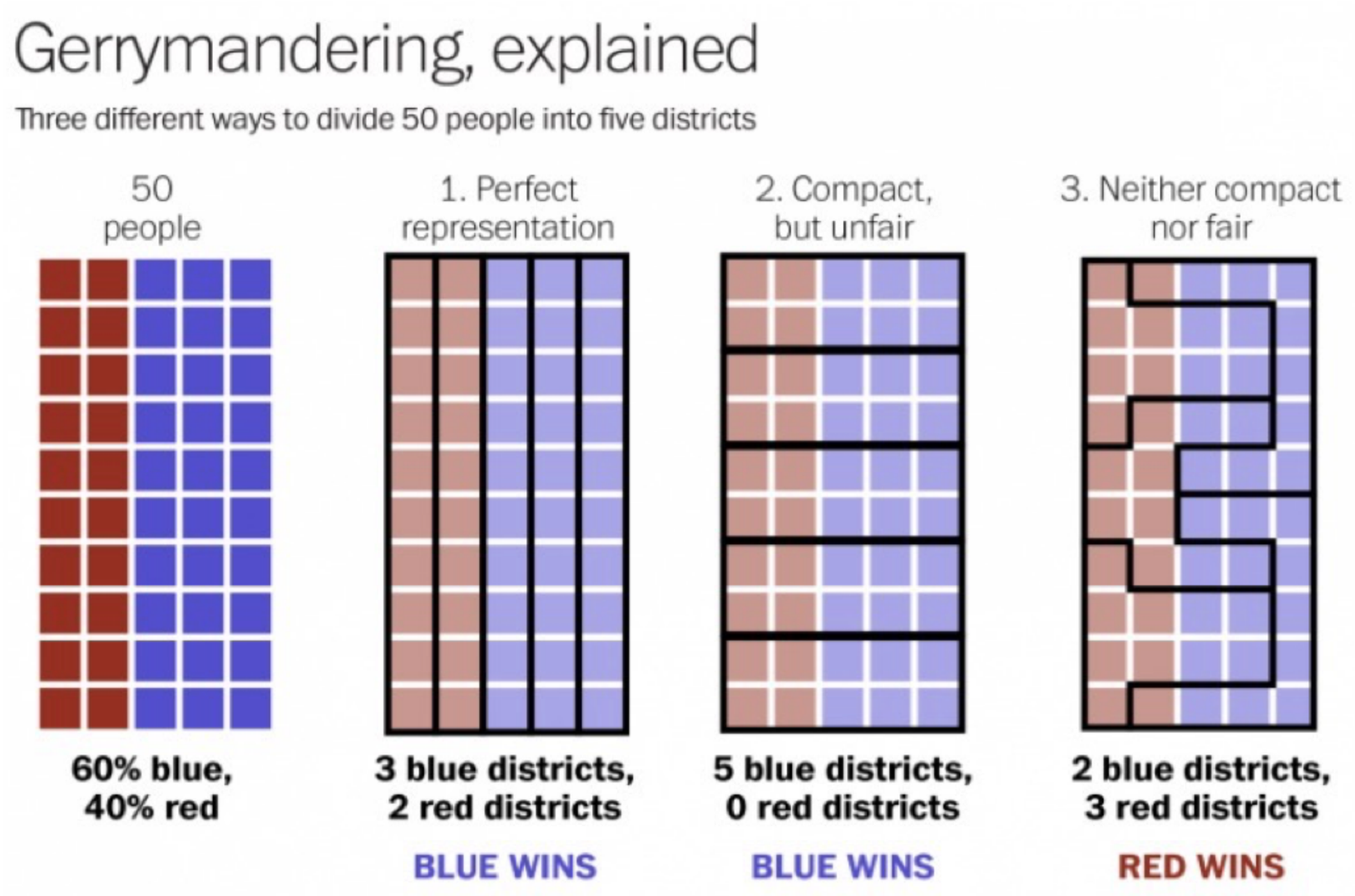
Modifiable Areal Unit Problem (MAUP)

modifiable area: Units are arbitrary defined and different organization of the units may create different analytical results.





For example...gerrymandering



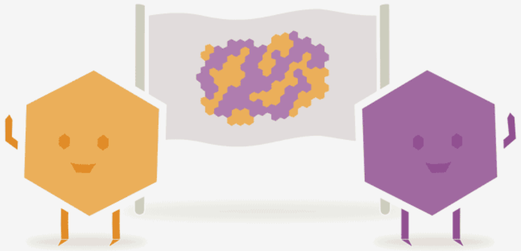
https://www.washingtonpost.com/news/wonk/wp/2015/03/01/this-is-the-best-explanation-of-gerrymandering-you-will-ever-see/

For example...gerrymandering

North Carolina

DISTRICTS REDRAWN TO OPTIMIZE COMPACTNESS

SOURCE: U.S. Census Bureau (top), Brian Olson (bottom)
GRAPHIC: The Washington Post. Published June 3, 2014



Welcome to Hexapolis

Every 10 years, Hexapolis redraws its congressional district lines — just like the United States does. But Hexapolis is a simpler place.

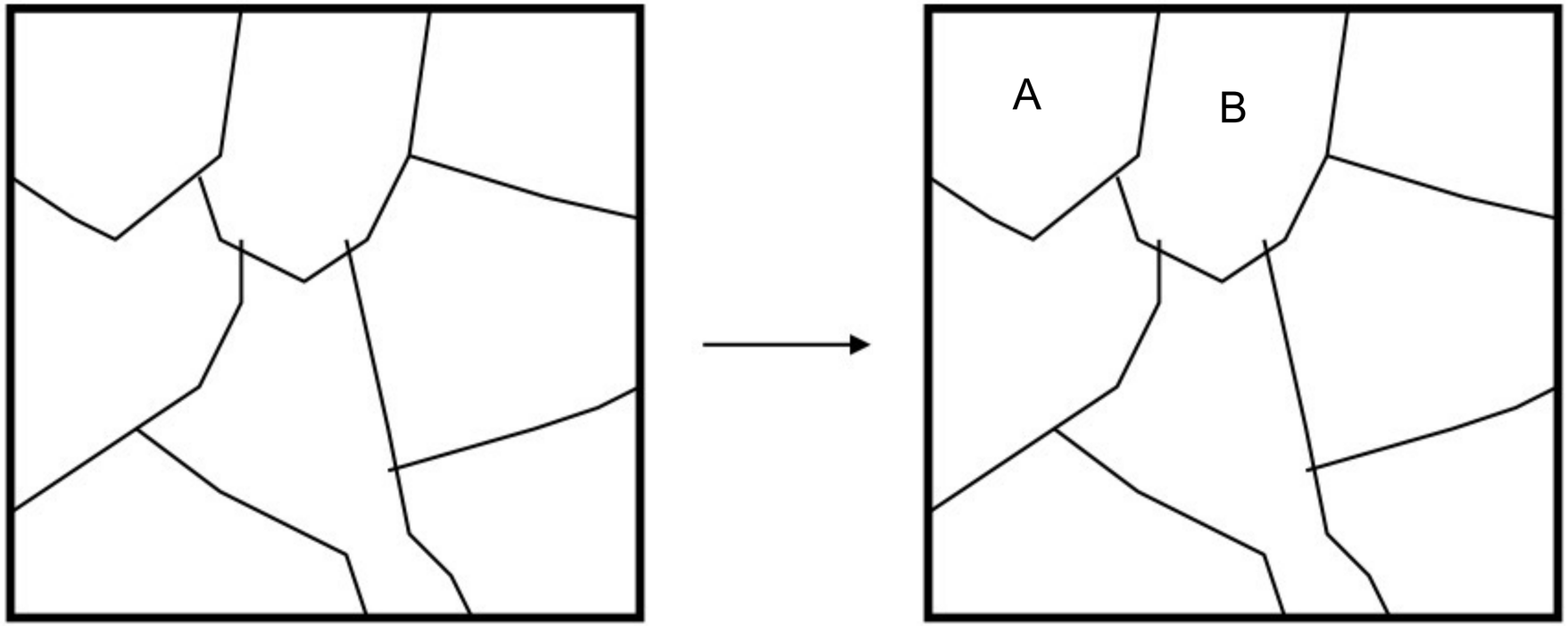
Lawmakers in either the **Purple Party** or **Yellow Party** control redistricting. To increase their advantage in upcoming elections, they have been known to gerrymander egregiously — even if it means leaving some voters disenfranchised.

Hexapolis has nine districts. Even though a majority of voters favor the Purple Party, that does not mean that the Yellow Party can't shift the state's partisan tilt.

https://www.nytimes.com/interactive/2022/01/27/us/politics/congressional-gerrymandering-redistricting-game-2022.html

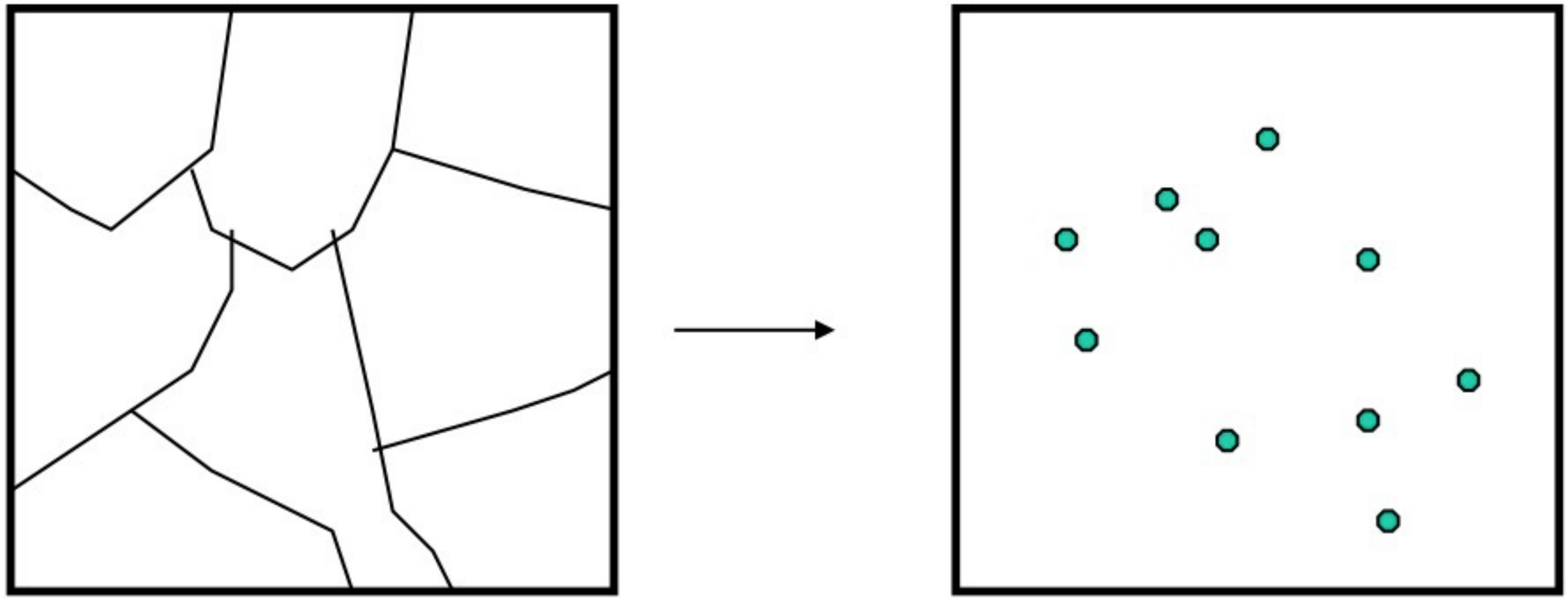
Edge Effects (The Boundary Problem)

Analyzing A vs B ignores similarities between the two based on their shared boundary



Ecological Fallacy

The Ecological Fallacy is a situation that can occur when a researcher or analyst makes an inference about an individual based on aggregate data for a group.



Ecological Fallacy

Example: we might observe a *strong relationship between income and crime at the county level*, with lower-income areas being associated with higher crime rate.

<u>Conclusions we might draw</u>:

- Lower-income persons are more likely to commit crime
- Lower-income areas are associated with higher crime rates
- Lower-income counties tend to experience higher crime rates The only valid conclusion!

Ecological Fallacy

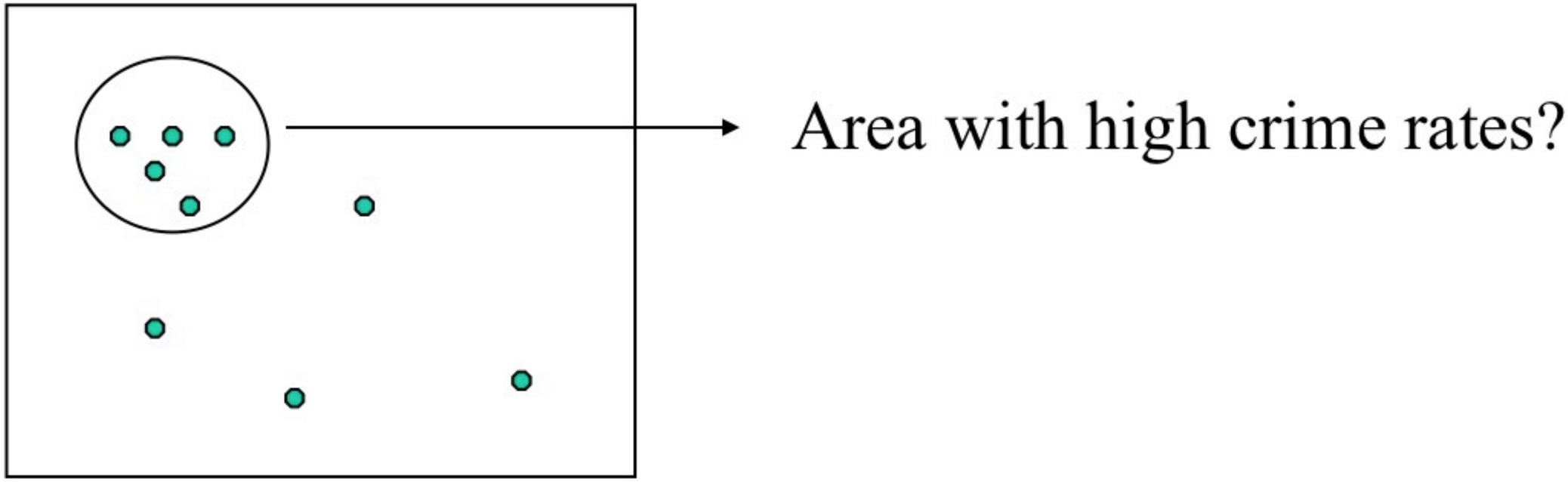
Issues:

Inferences drawn about associations between the characteristics of an aggregate population and the characteristics of sub-units within the population are wrong. That is: *results from aggregated data (e.g. counties) cannot be applied to individual people*

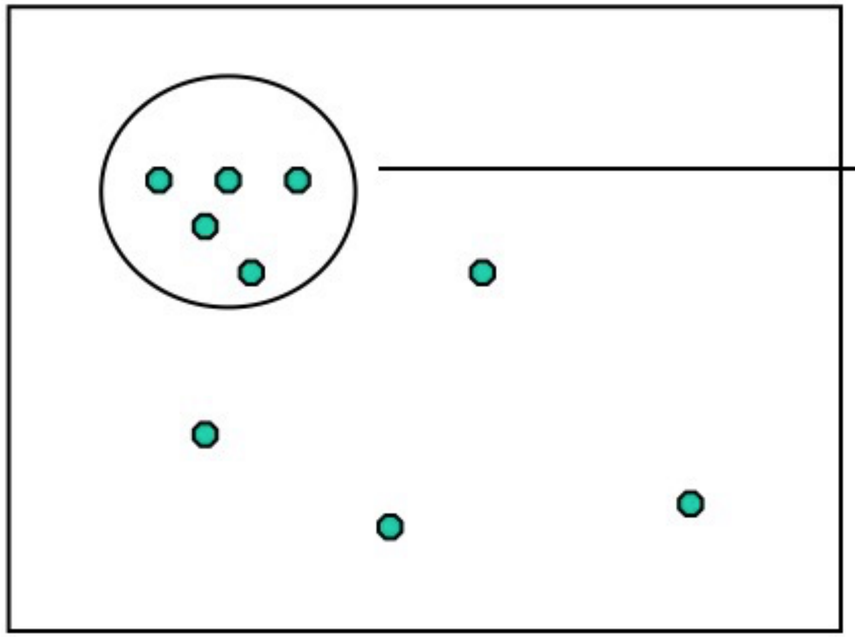
What should we do?

Be aware of the process of aggregating or disaggregating data may conceal the variations that are not visible at the larger aggregate level

Nonuniformity



Nonuniformity

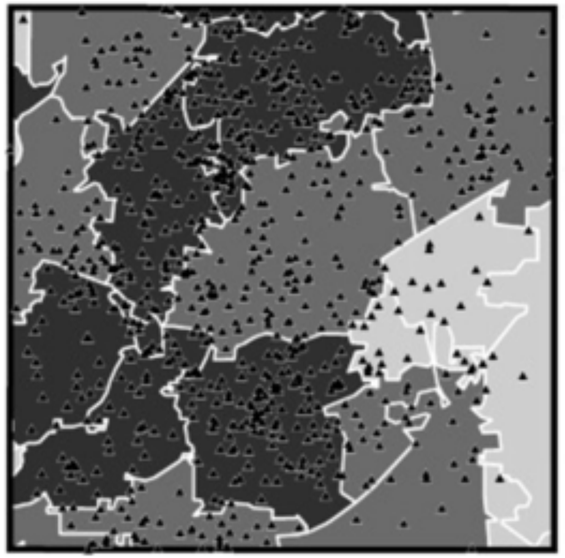


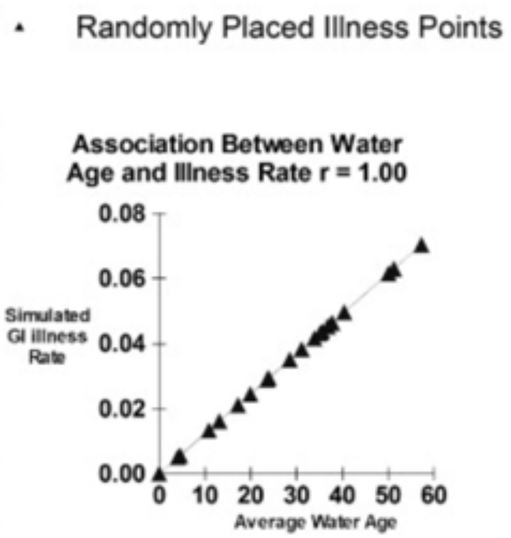
Crime locations

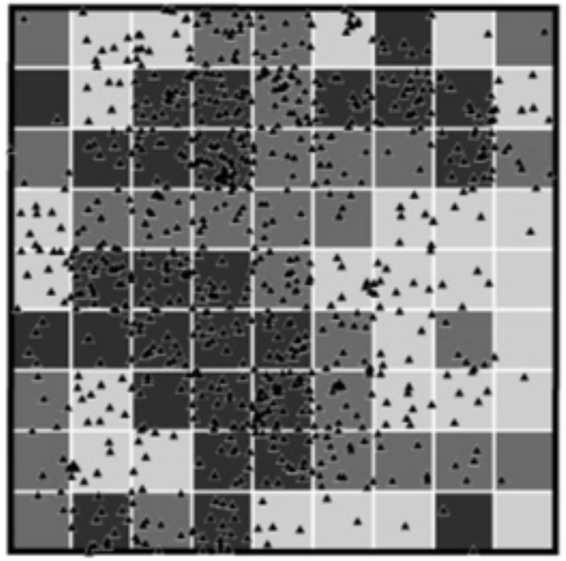
Area with high crime rates?

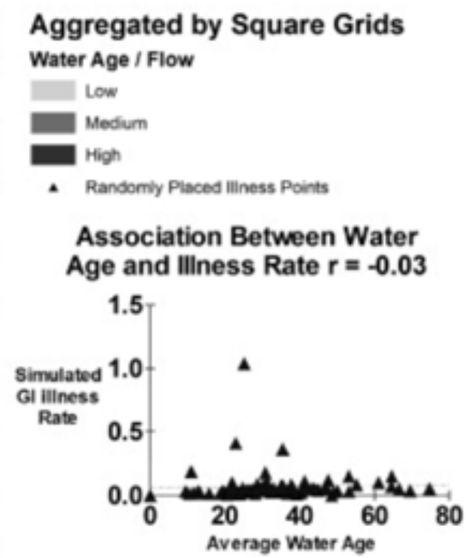
<u>Conclusion</u>: Bank robberies are clustered
….but only because banks are clustered!

Spatial Statistics









What explains what's going on here?

A
Spatial Autocorrelation

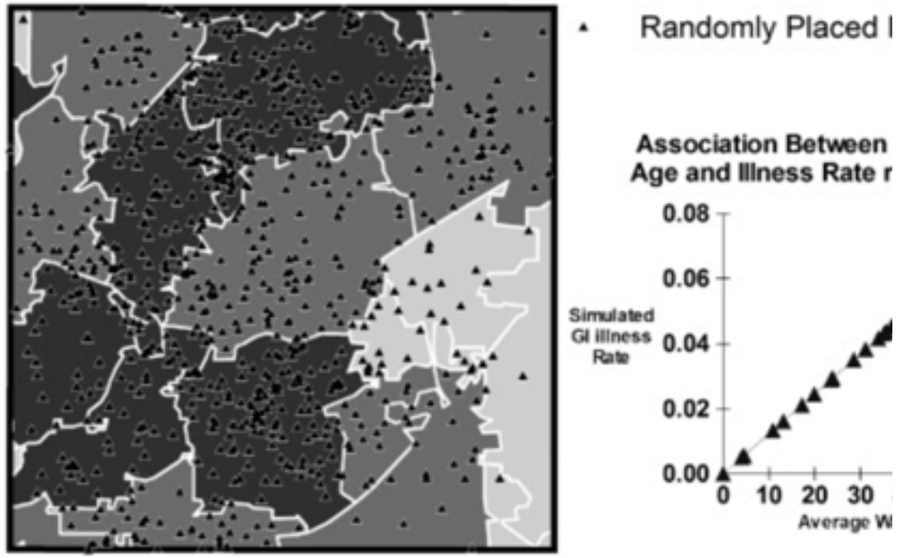
B
MAUP

C
Edge Effects

D
Ecological Fallacy

E
Nonuniformity

Spatial Statistics



What explains

A
Spatial
Autocorrelation

B
MAUP

https://forms.gle/nCTx4ud81GiRtLY48