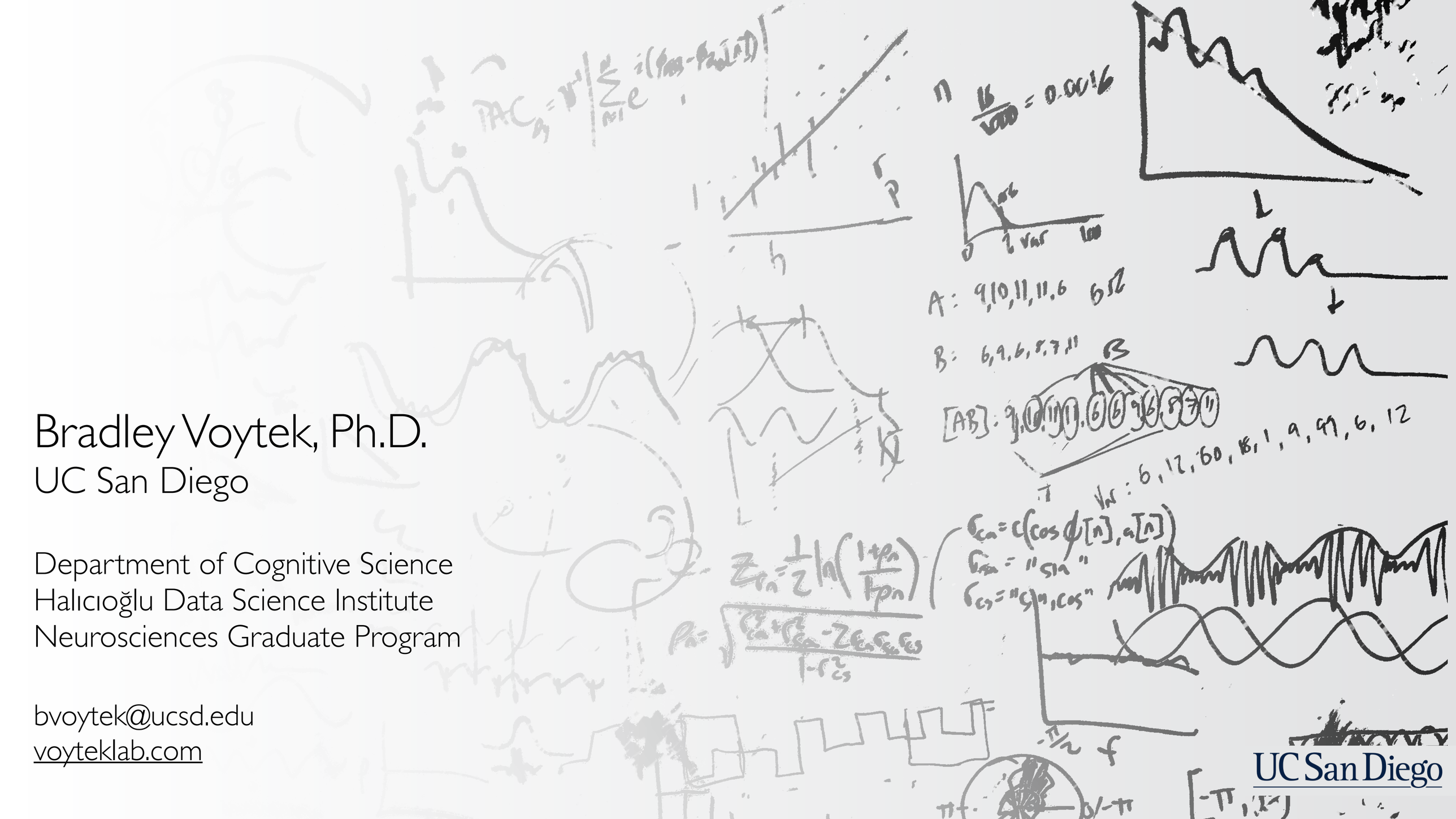
Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
voyteklab.com



UC San Diego

Decision trees

- A sequence of tests.
- Representation very natural for humans.
- Style of many "How to" manuals and trouble-shooting procedures.

Decision trees

http://r2d3.us/visual-intro-to-machine-learning-part-1/

COGS 9
Introduction to Data Science

Geography, MAUP, and the ecological fallacy

UC San Diego

Today's learning objective

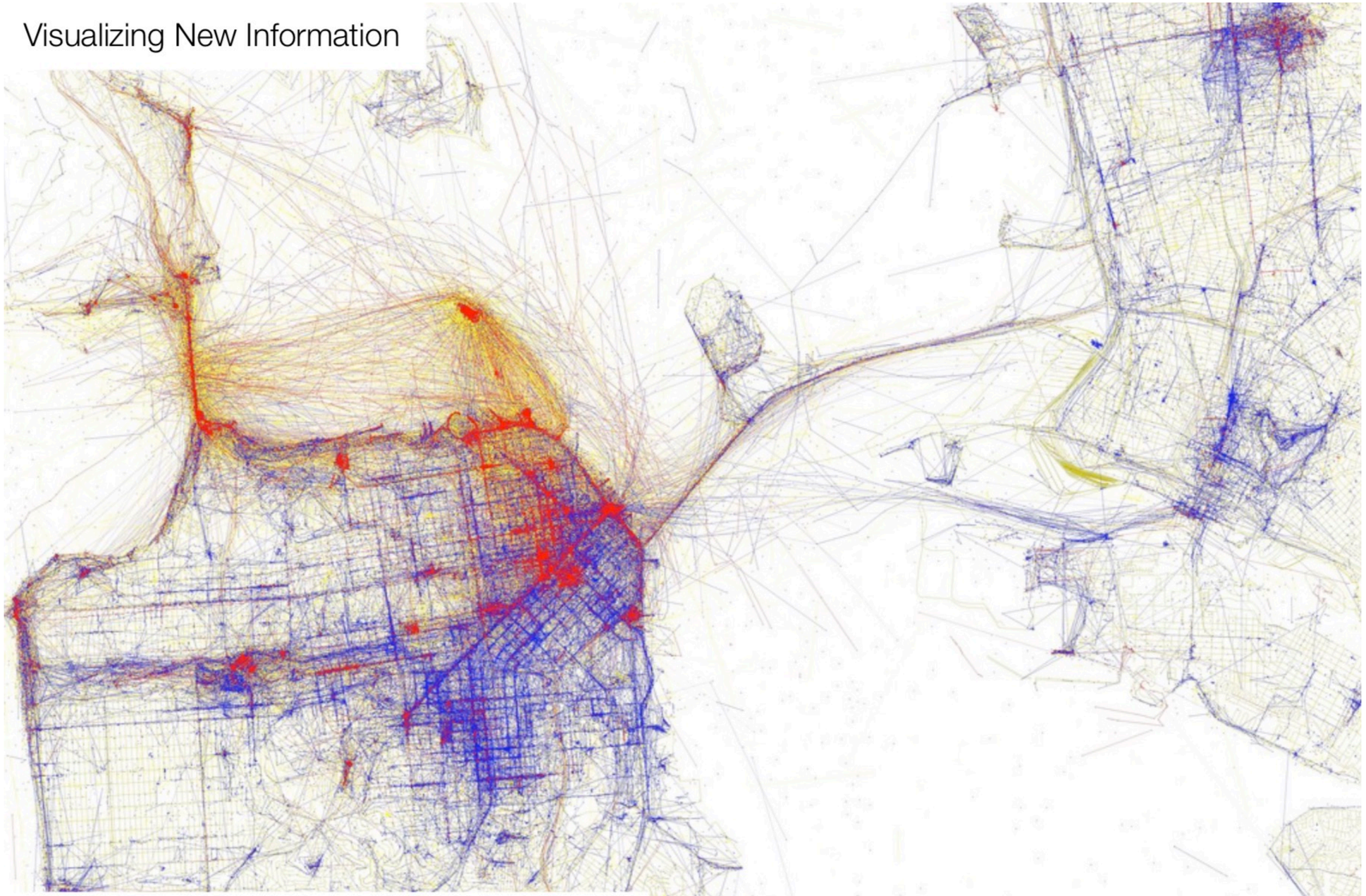
Describe the benefits and limitations of various geospatial visualizations.
Define and explain the various statistical challenges to geospatial analysis.

Why geospatial analyses?

"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

Visualizing New Information



"Tourists vs. Locals," Eric Fischer, (2010)

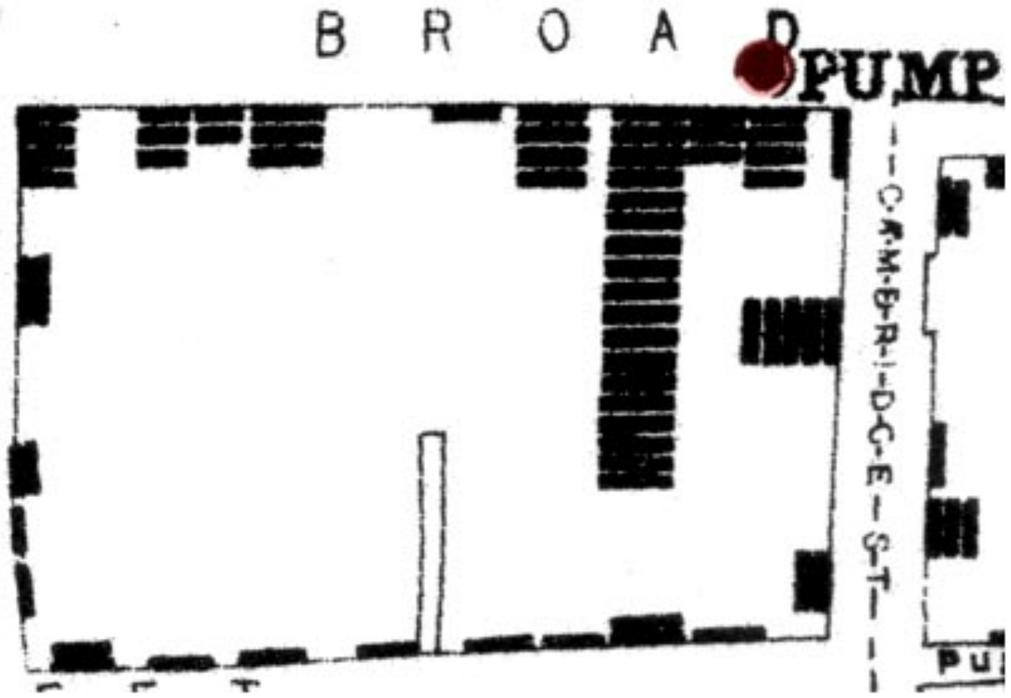
http://www.flickr.com/photos/walkingsf/sets/72157624209158632/

Source: *New York Times*


N Torrey Pines Rd
University of California, San Diego Park
University City
Scripps Coastal Reserve
San Diego Fwy
Expedition Way
La Jolla Athletic Park

Geospatial distributions to understand disease

Cholera



John Snow 1850s map of cholera in London

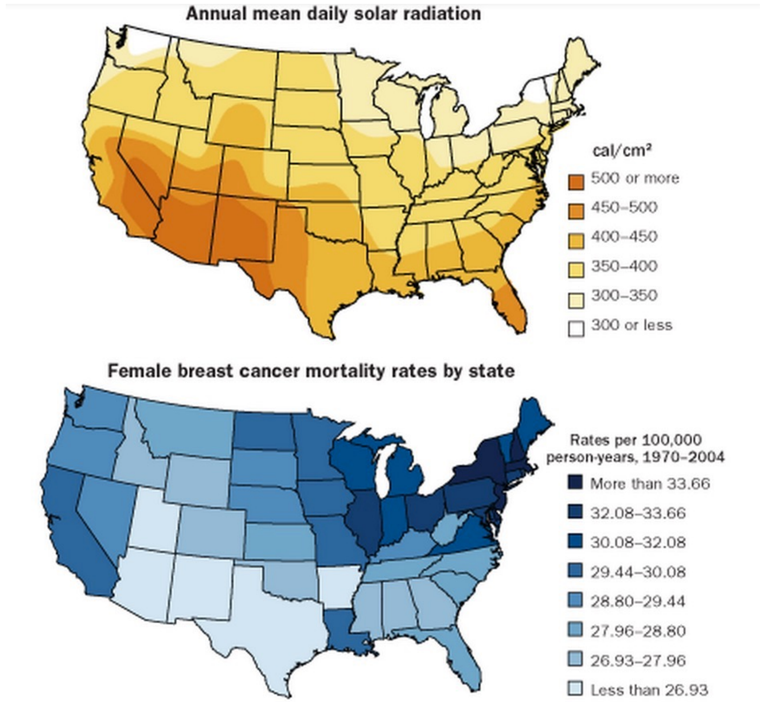
EDA or hypothesis testing?

Did he discover the association between water and cholera after drawing the map?

Or…

Did he draw the map in order to prove the association?

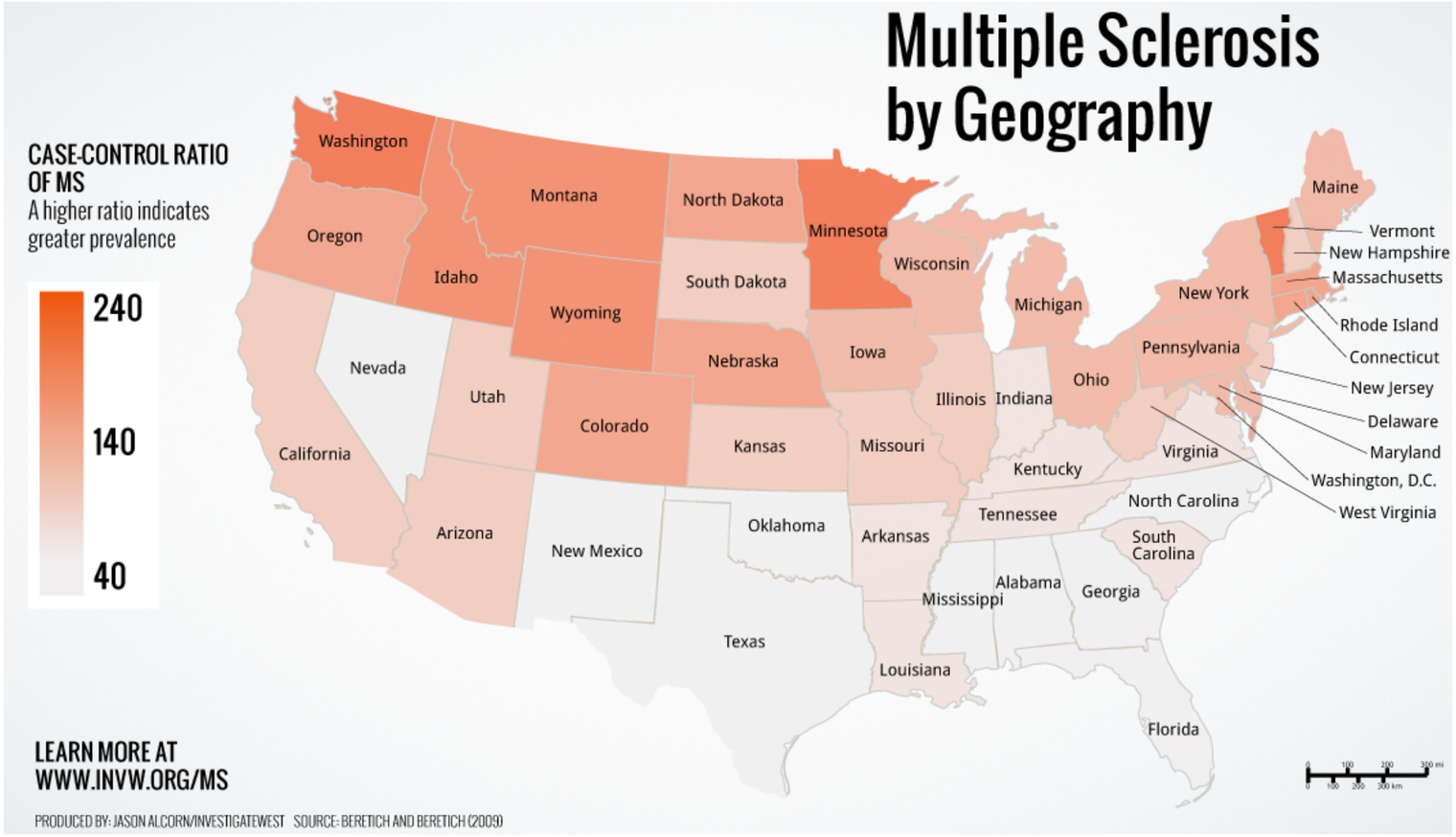
Geospatial disease distribution



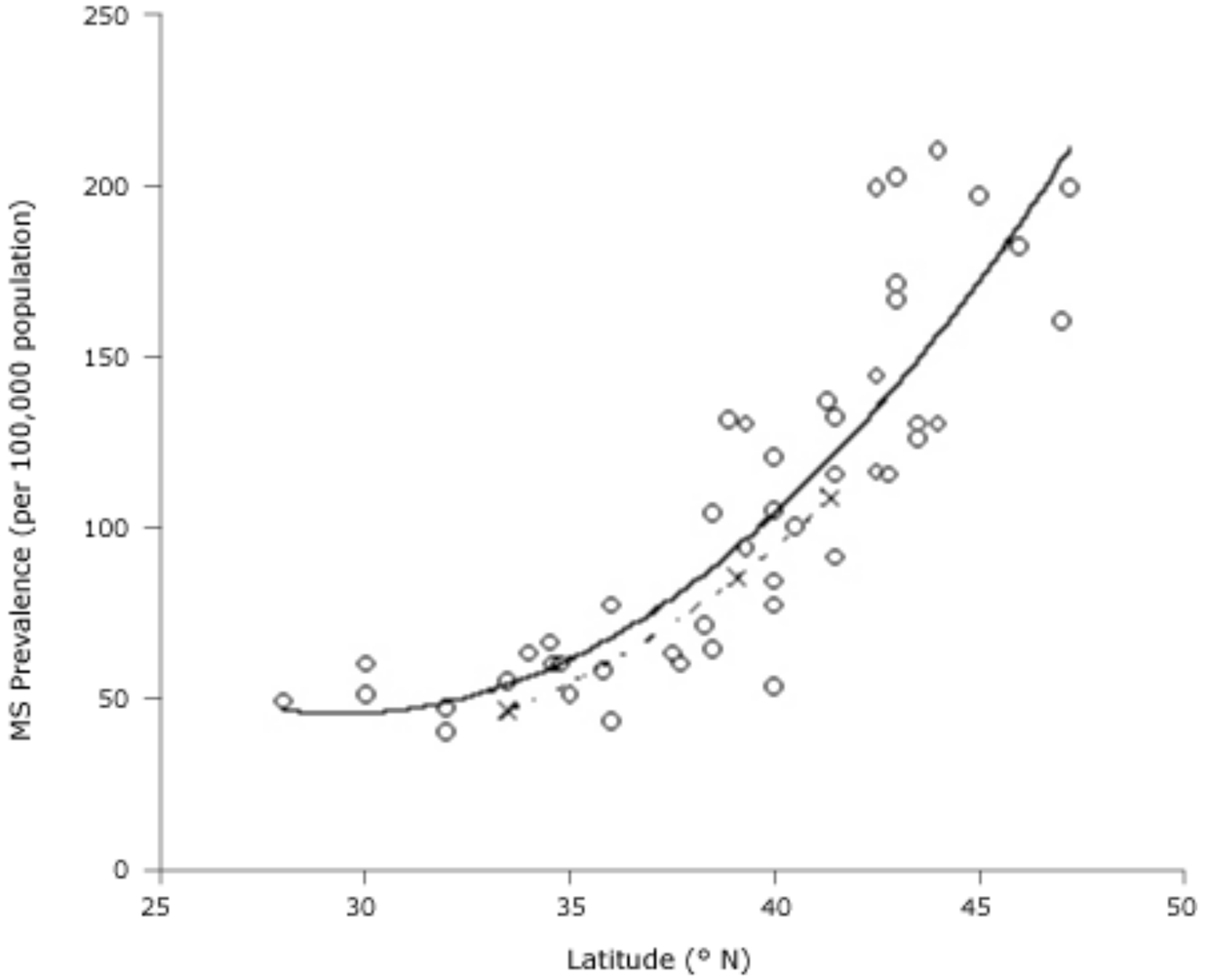
ON THE MAP Scientists who study vitamin D can't help but notice that a host of diseases seem to vary with latitude. Type 1 diabetes, multiple sclerosis and even some cancers appear to be more common in areas that get less sun -- meaning less opportunity for the body to produce vitamin D. The maps above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

Why geospatial analyses?



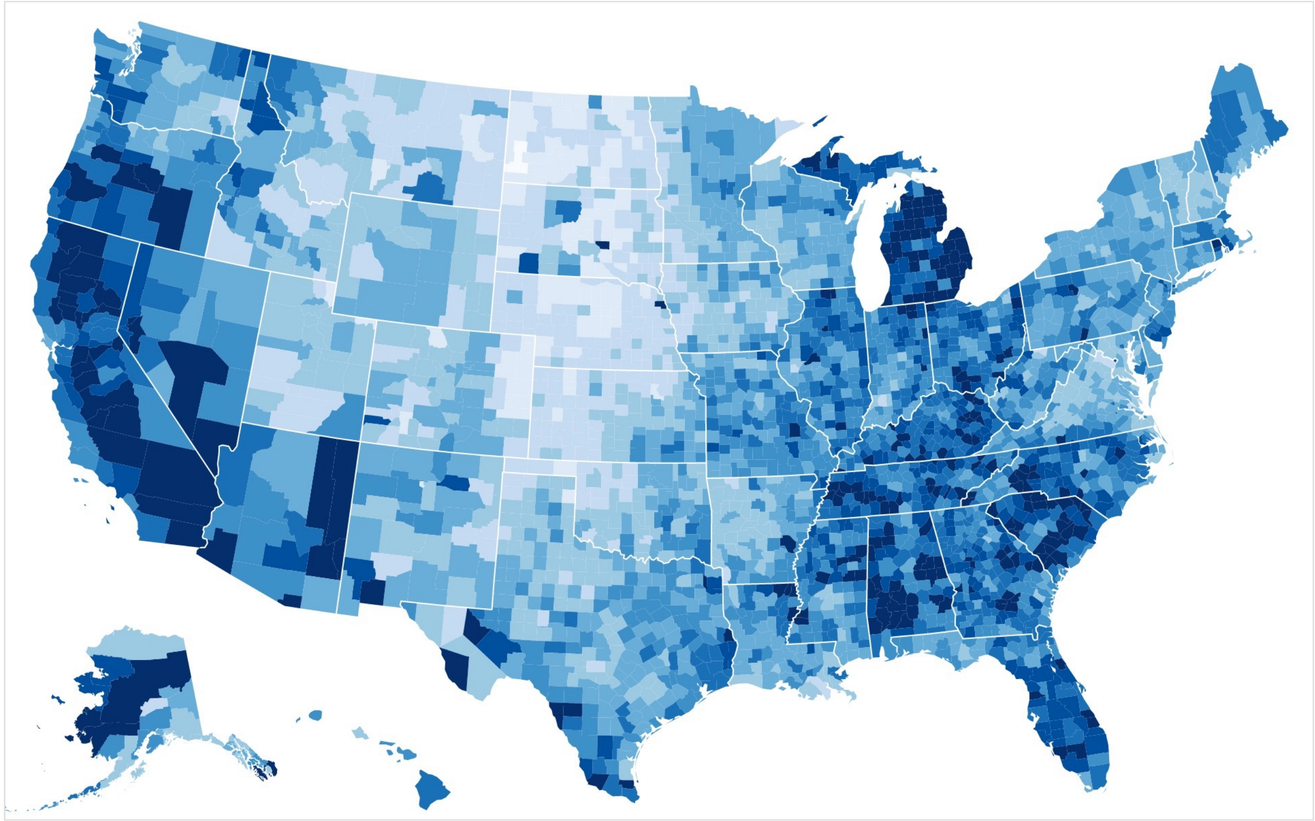
Why geospatial analyses?



Source: CDC

Choropleth maps

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.

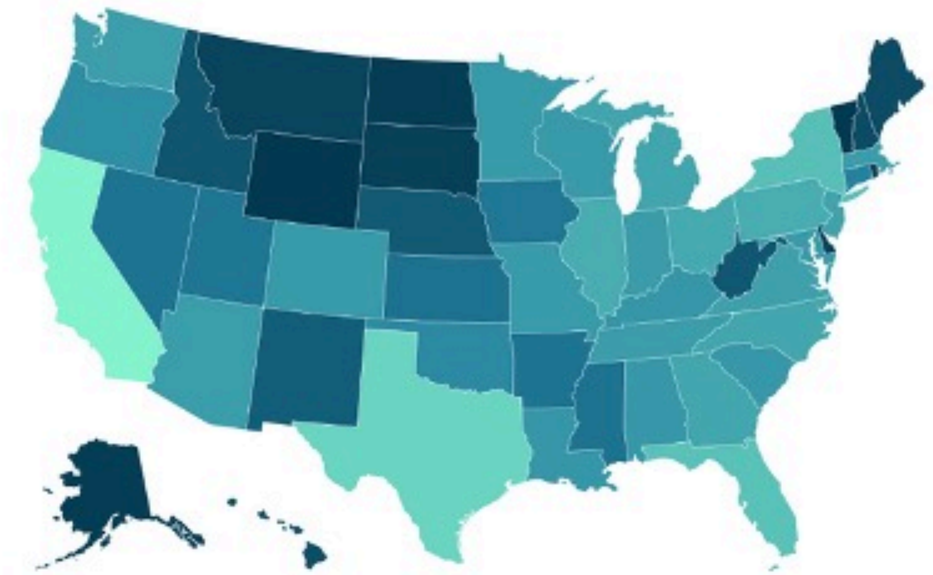
Open in a new window.

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

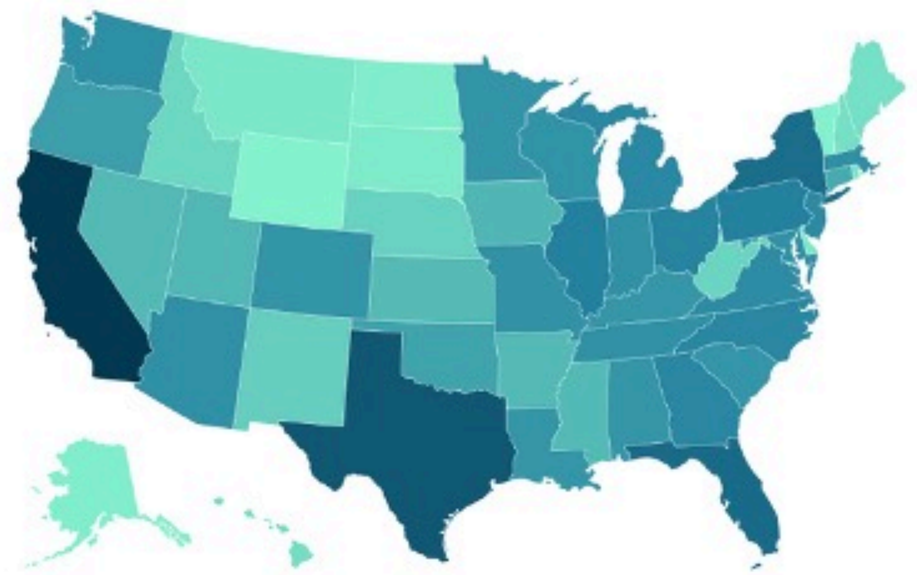
Source: http://bl.ocks.org/mbostock/4060606

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

NOT IDEAL



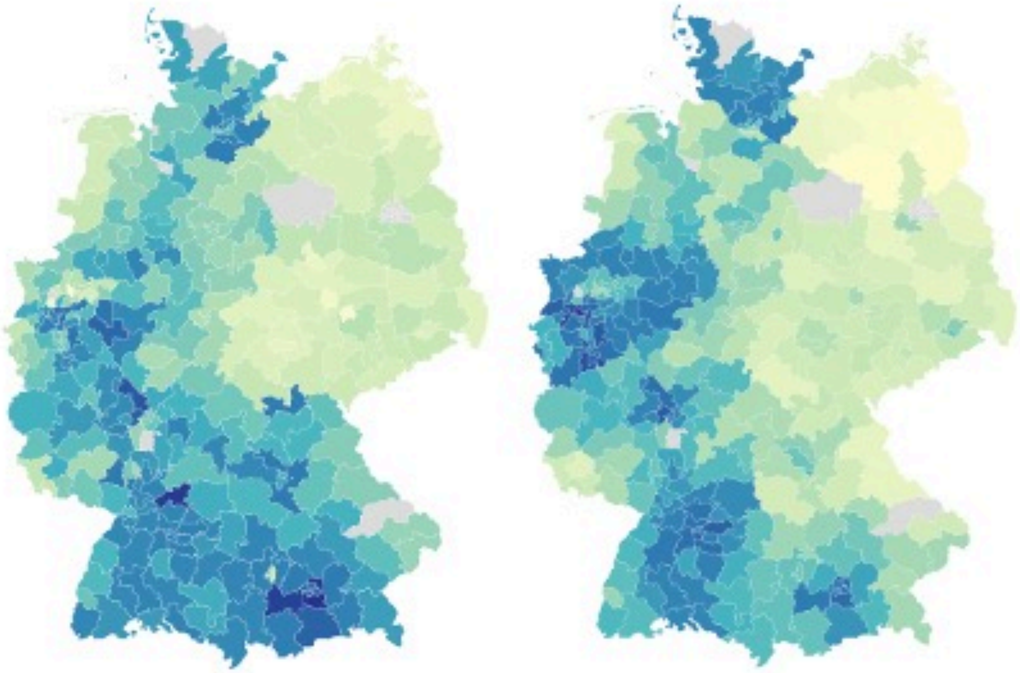
BETTER



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

Choropleth Maps shine when displaying a *single* variable

NOT IDEAL



INCOME

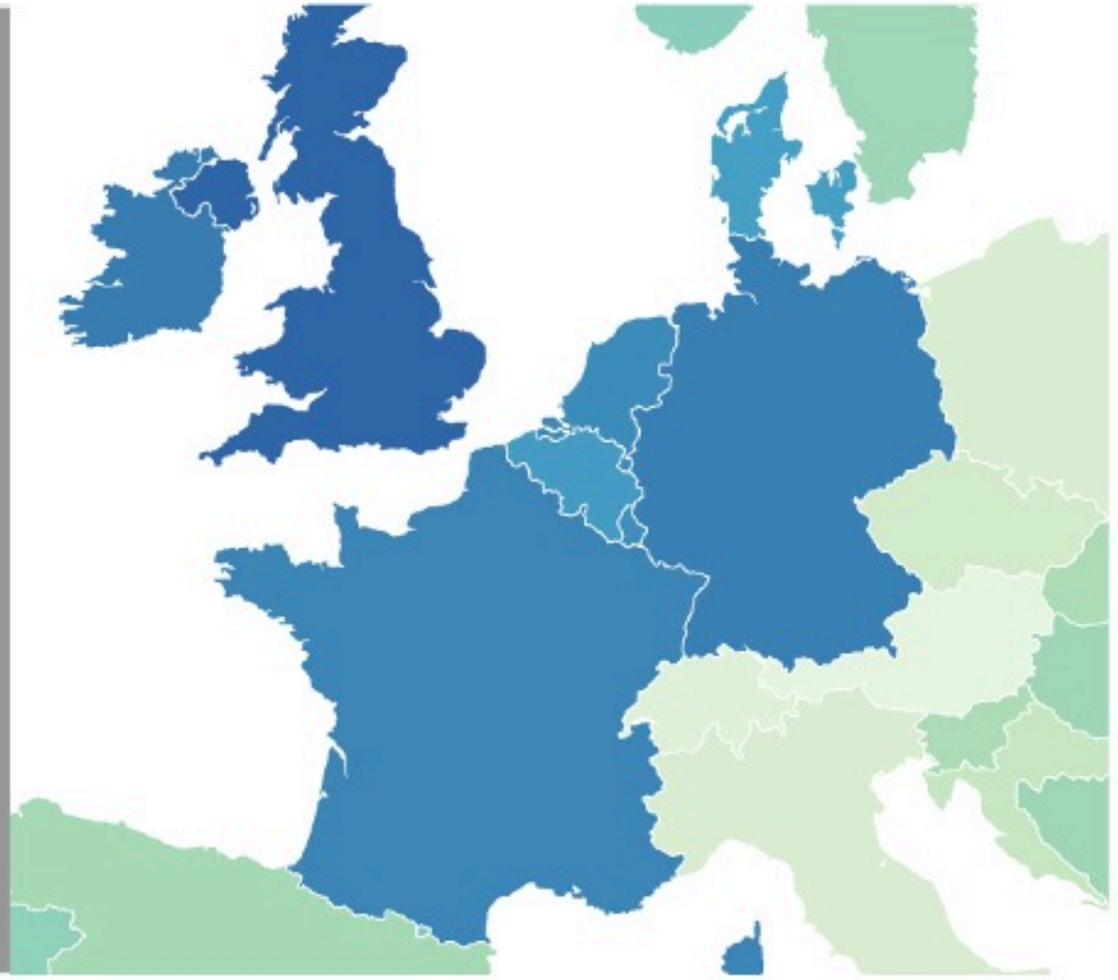
VOTES FOR THE LIBERAL PARTY

BETTER

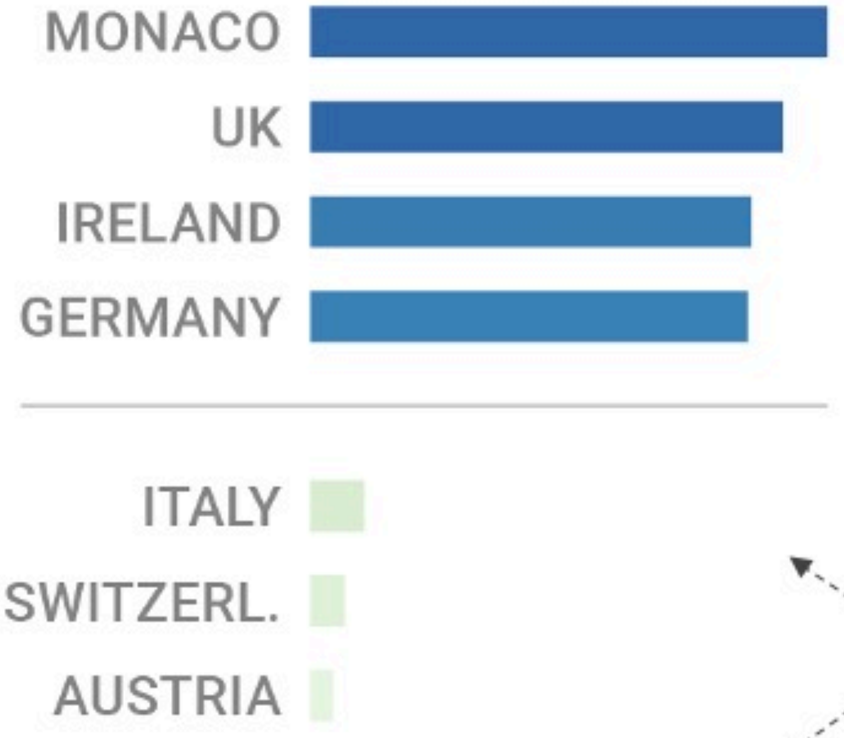
Source: https://blog.datawrapper.de/colors/

Choropleth excel at displaying the big picture, *not* subtle differences

NOT IDEAL



BETTER



If your goal is to point out the numeric differences between regions, choose something other than a choropleth

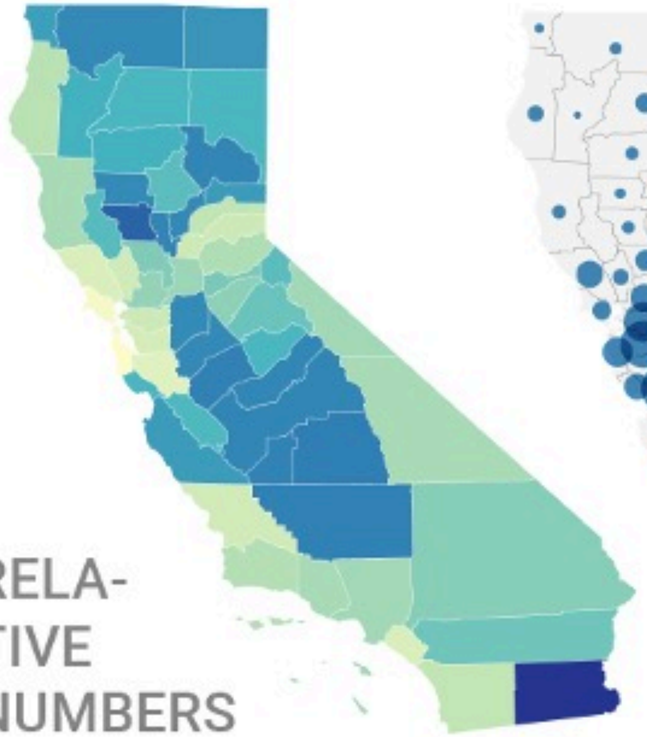
Source: https://blog.datawrapper.de/colors/

Choropleth should display relative differences, *not* absolute numbers

NOT IDEAL

UNEMPLOY-
MENT IN
ABSOLUTE
NUMBERS

BETTER

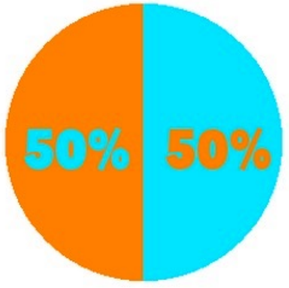


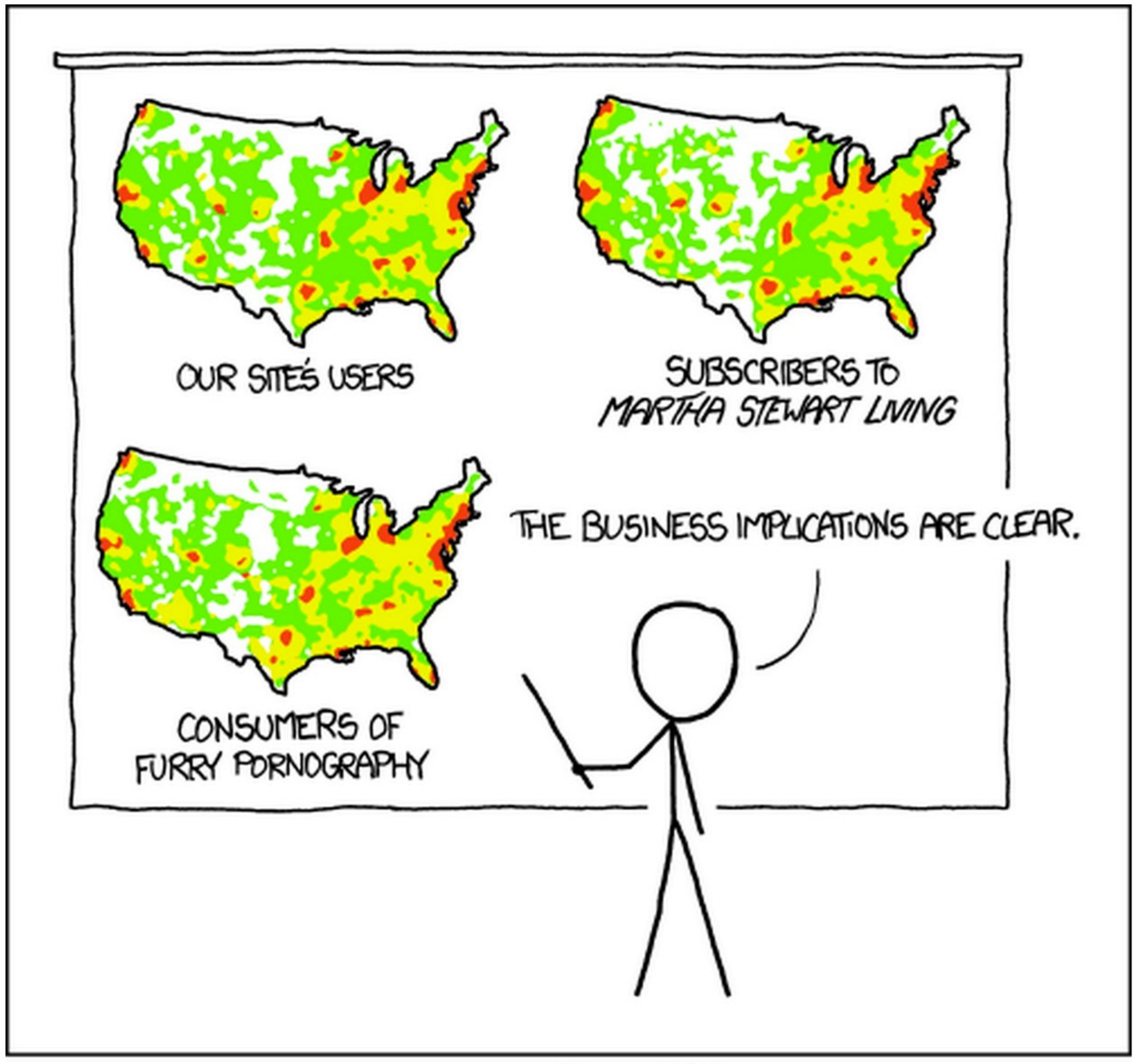
RELA-
TIVE
NUMBERS

SYMBOL
MAP WITH
ABSOLUTE
NUMBERS

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

UNITED STATES ECONOMIC ACTIVITY, SPLIT IN HALF





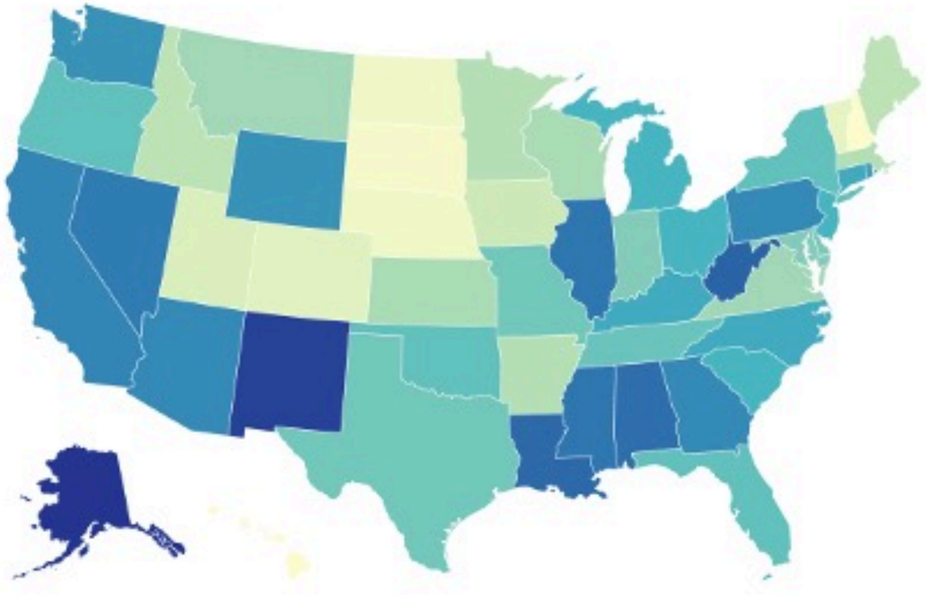
PET PEEVE #208:

GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

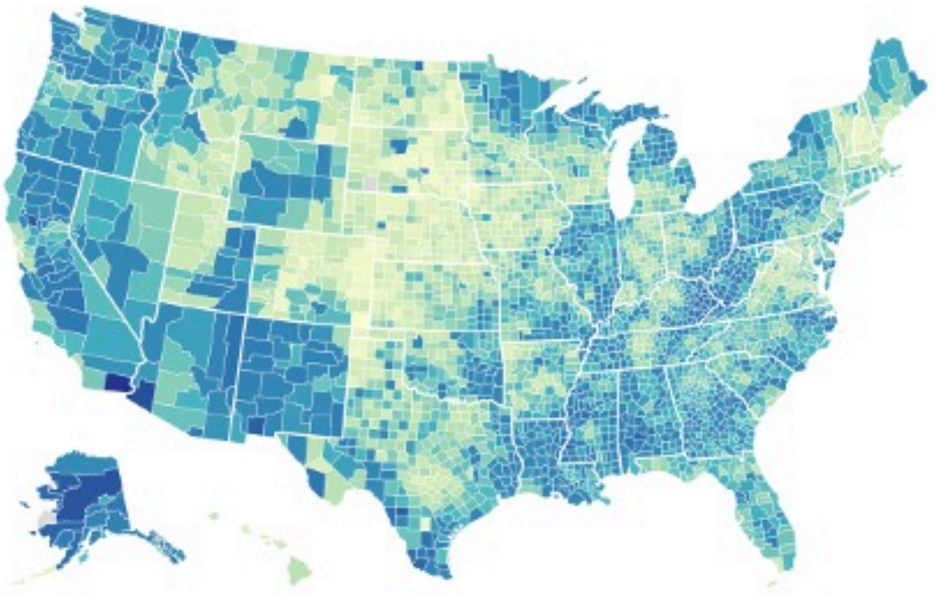
Consider using the smallest unit possible

(but there are exceptions!)

NOT IDEAL

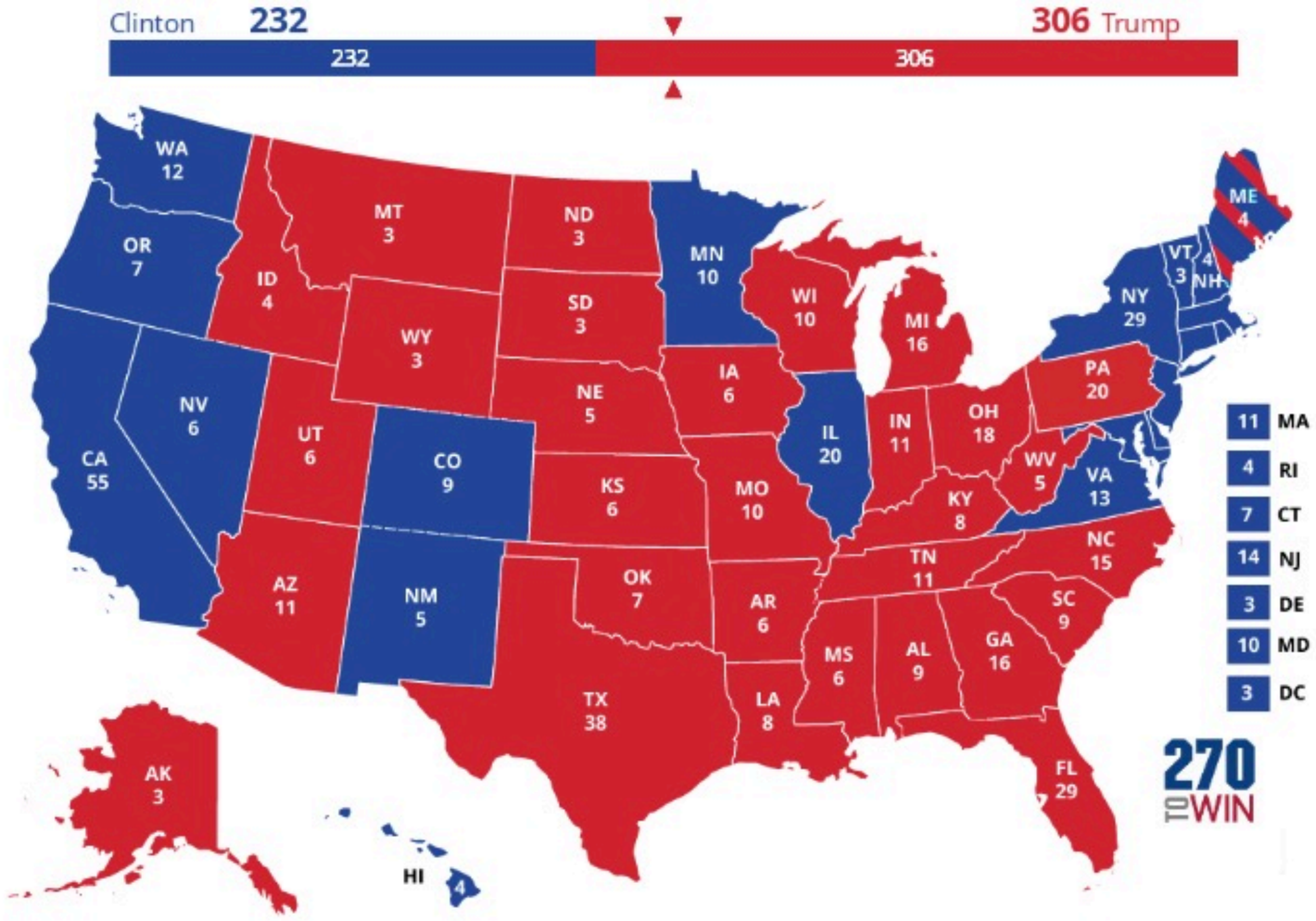


BETTER

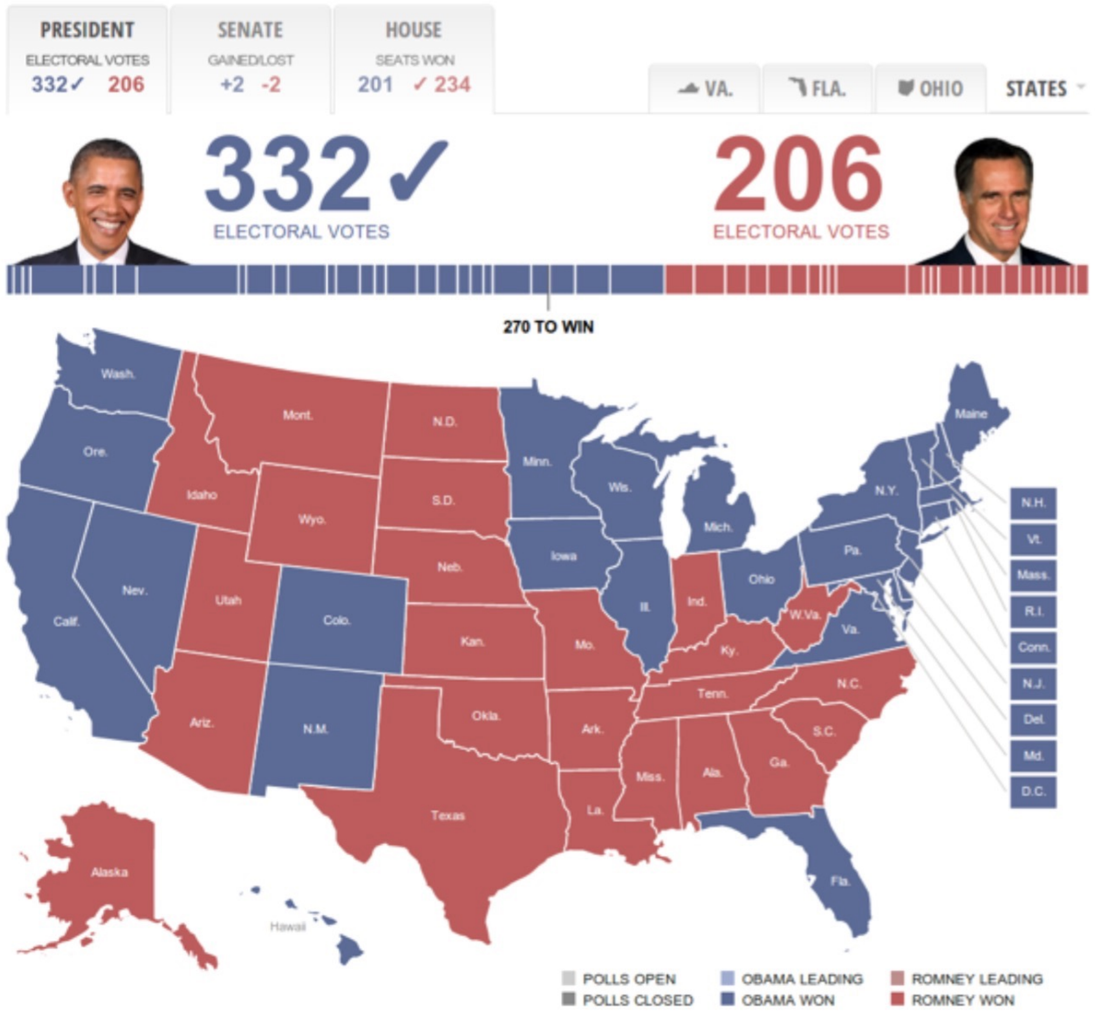


Source: https://blog.datawrapper.de/colors/

Misleading Choropleth maps



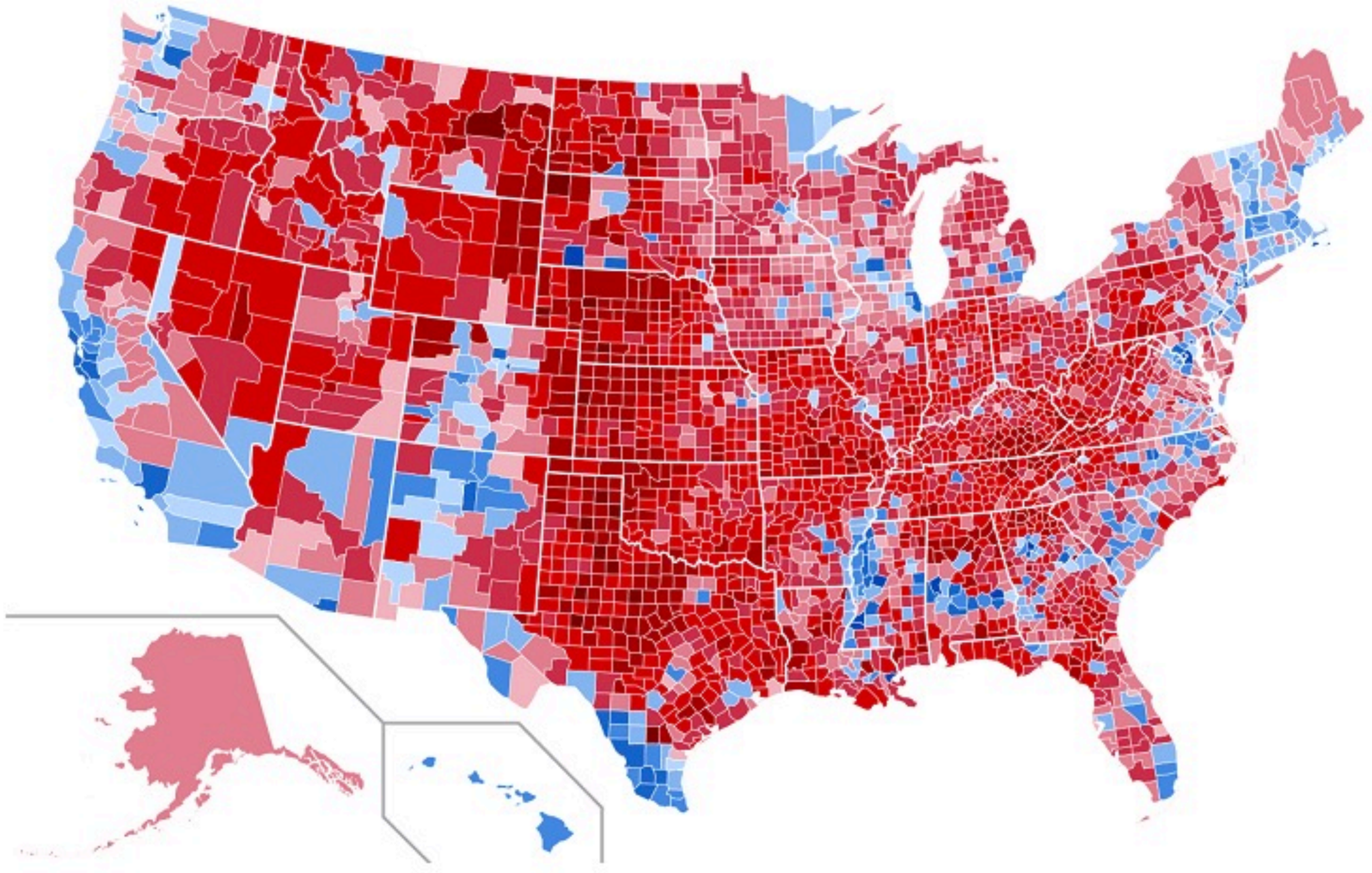
Misleading Choropleth maps



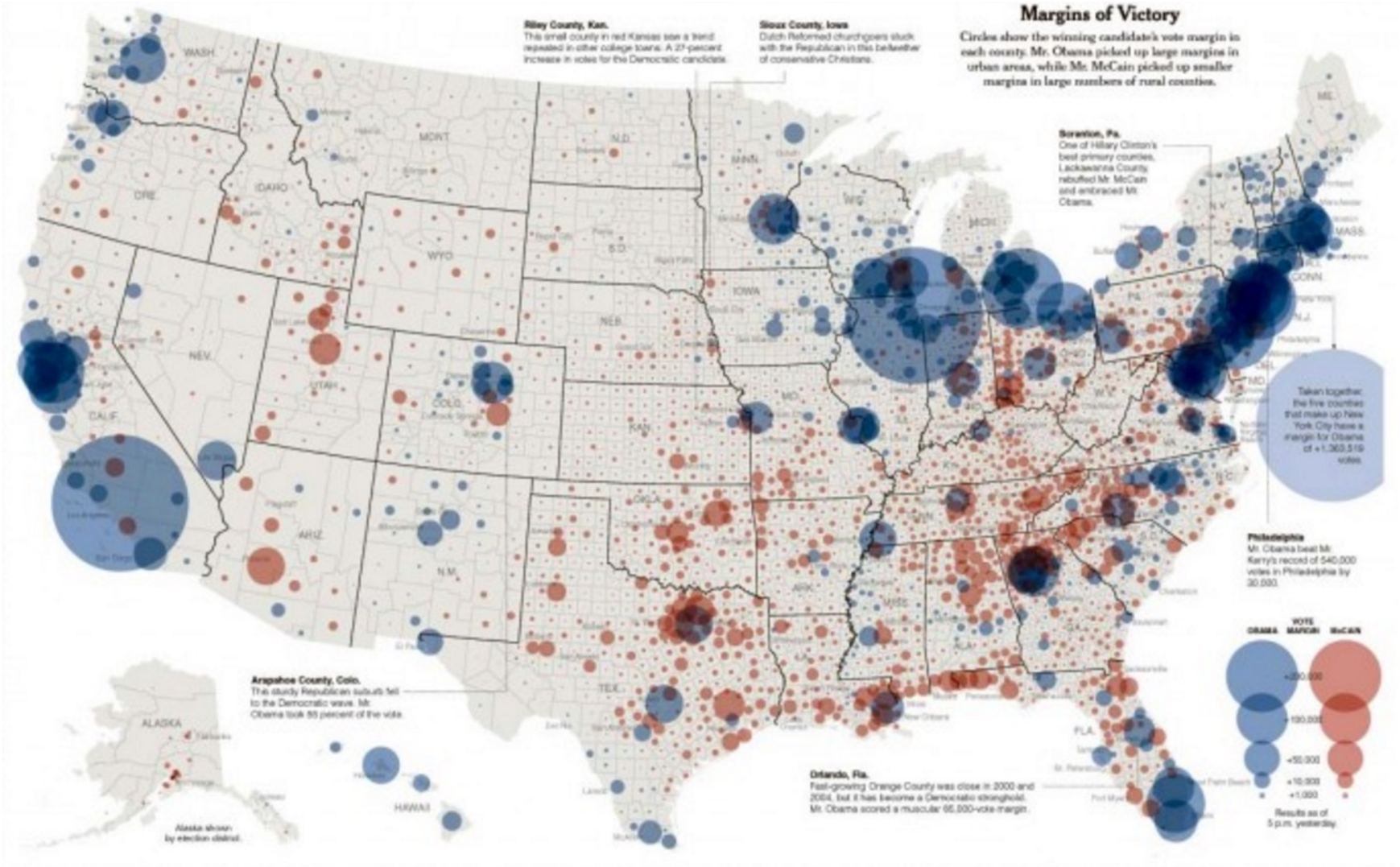
Source: *Huffington Post*, http://elections.huffingtonpost.com/2012/results

UC San Diego

Misleading Choropleth maps



Better maps

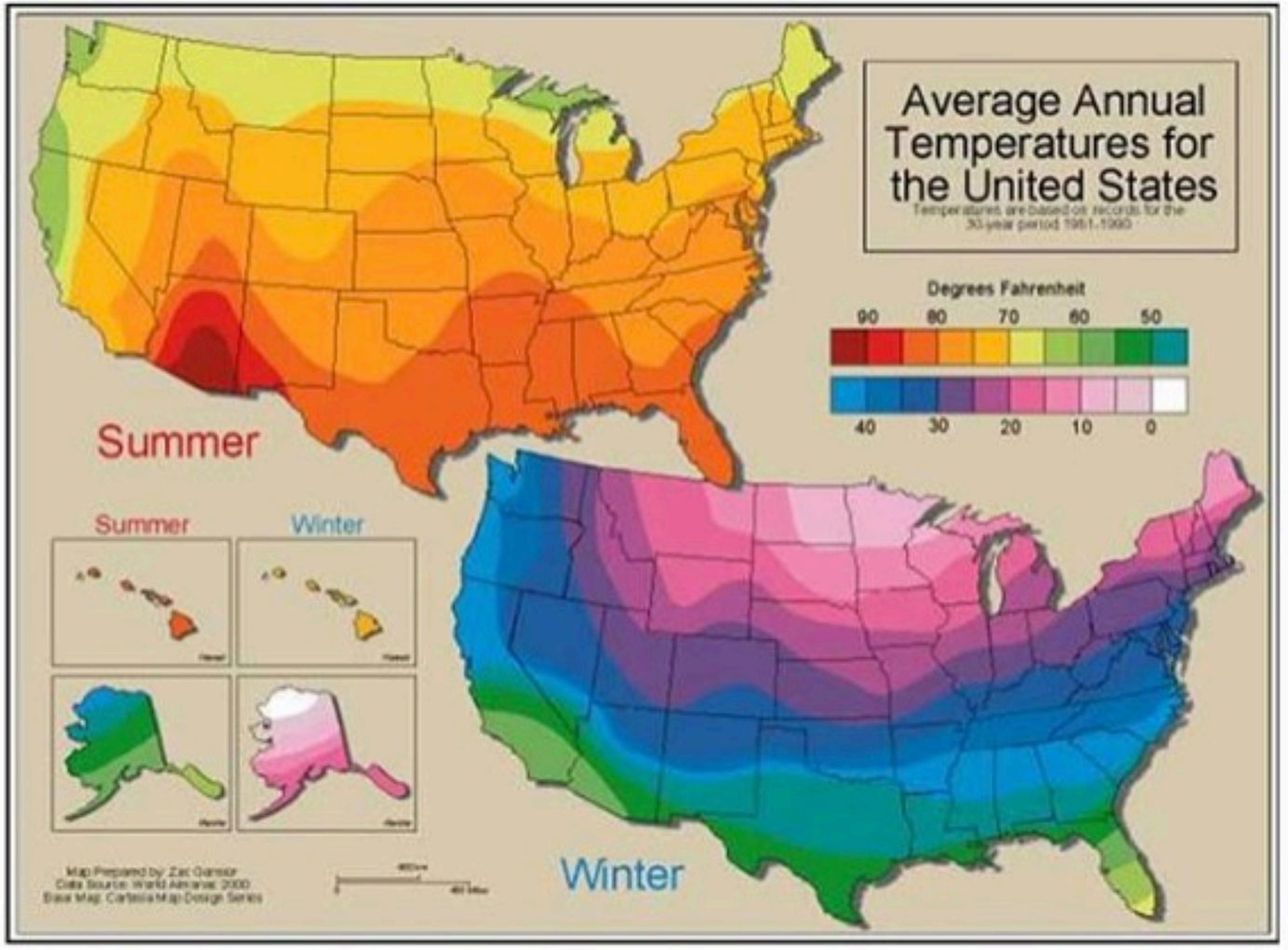


Source: http://www.thefunctionalart.com/2012/11/when-choropleth-maps-deceive.html

UC San Diego

Isarithmic maps demonstrate smooth, continuous phenomena

(temperature, elevation, rainfall, etc.)



Spatial statistics

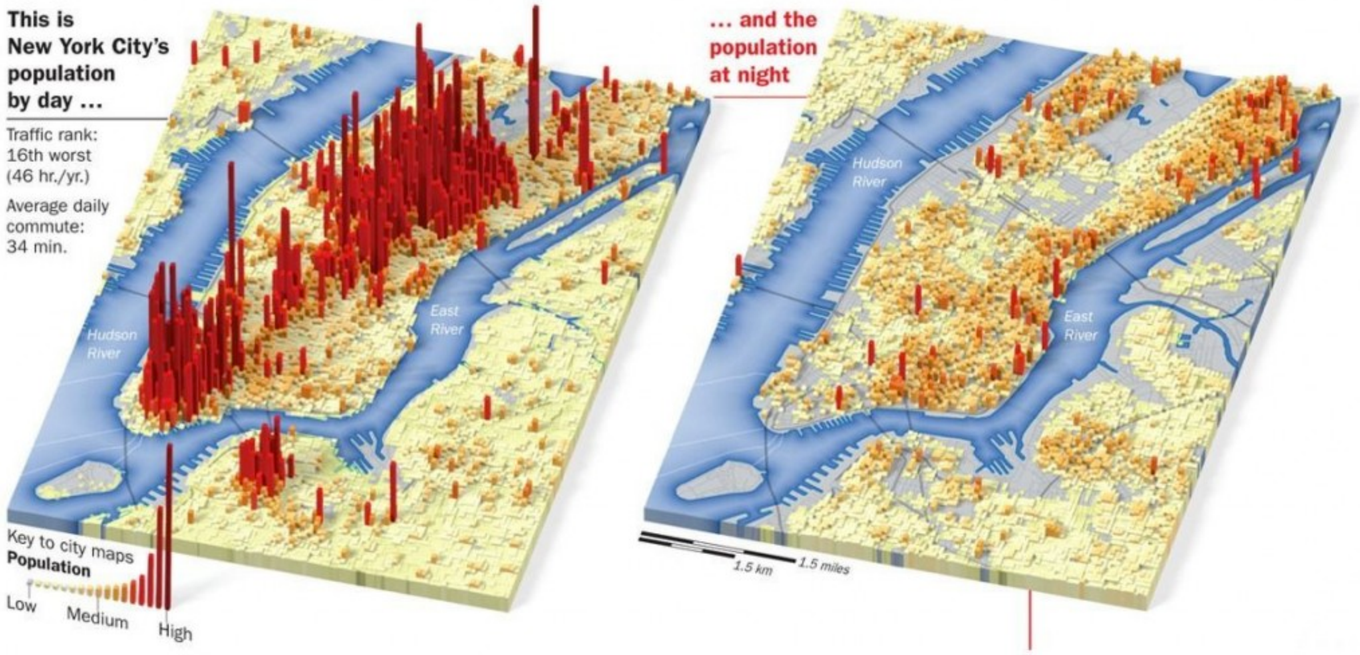
Spatial data violate the requirements of conventional statistics:

- Spatial autocorrelation
- Modifiable areal unit problem
- Ecology fallacy
- Scale
- Nonuniformity of space
- Edge effects

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

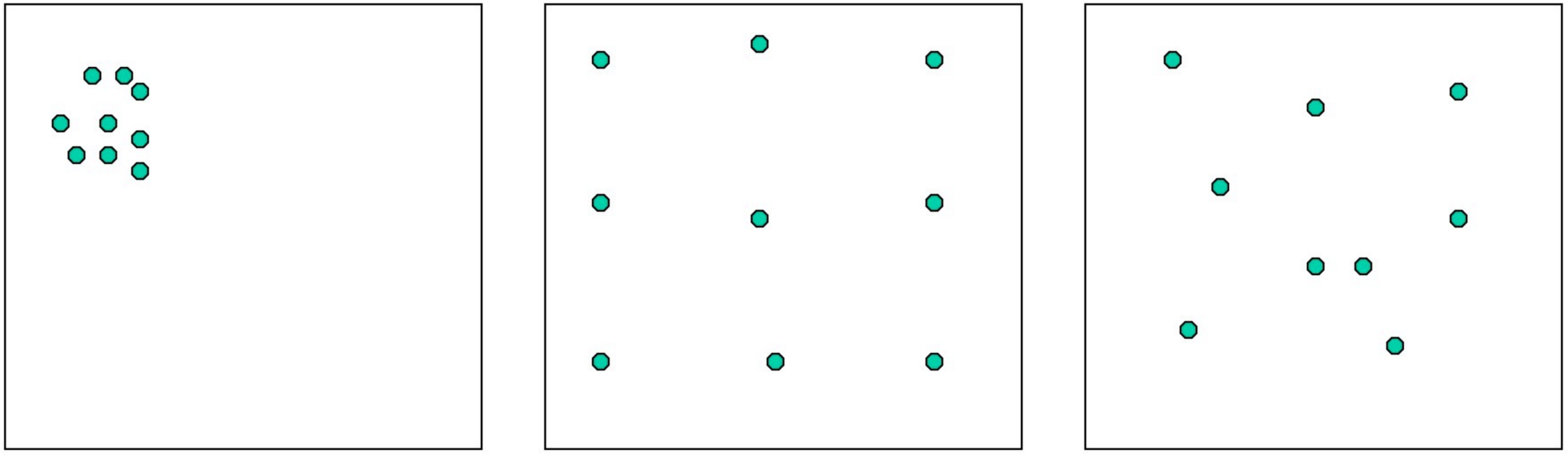


Spatial autocorrelation

Three general possibilities:

- Positive autocorrelation: nearby locations are *more* likely to be similar to one another.
- Negative autocorrelation: nearby locations are *less* likely to be similar to one another.
- Zero autocorrelation: no spatial effect is discernible, and observations seem to vary randomly through space

Spatial autocorrelation



Positive

Negative

Zero (Random)

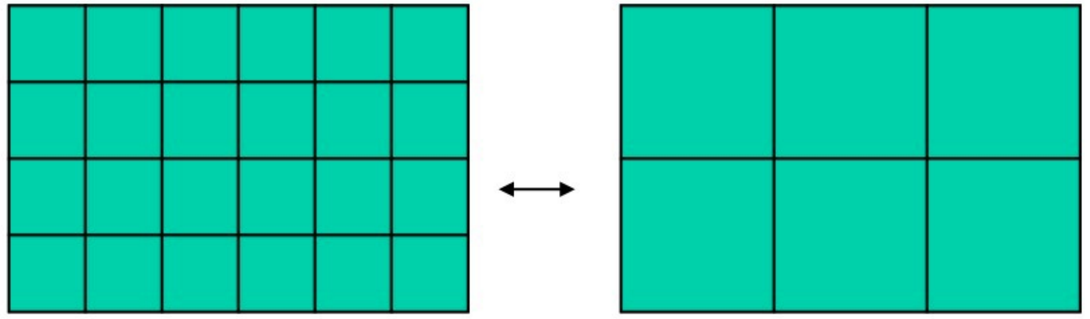
Modifiable Areal Unit Problem (MAUP)

Modifiable Areal Unit Problem: 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)

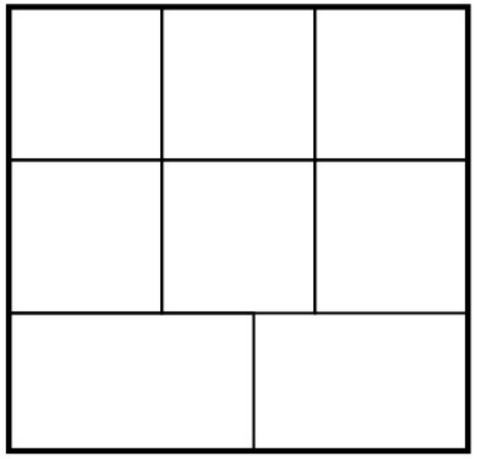
Scale issue: involves the aggregation of smaller units into larger ones. Generally speaking, the larger the spatial units, the stronger the relationship among variables.

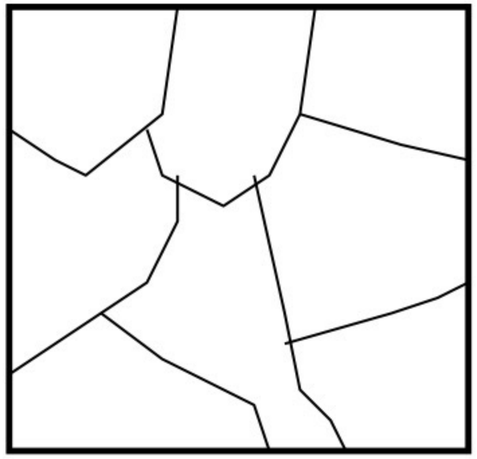


Aggregation (smoothed)

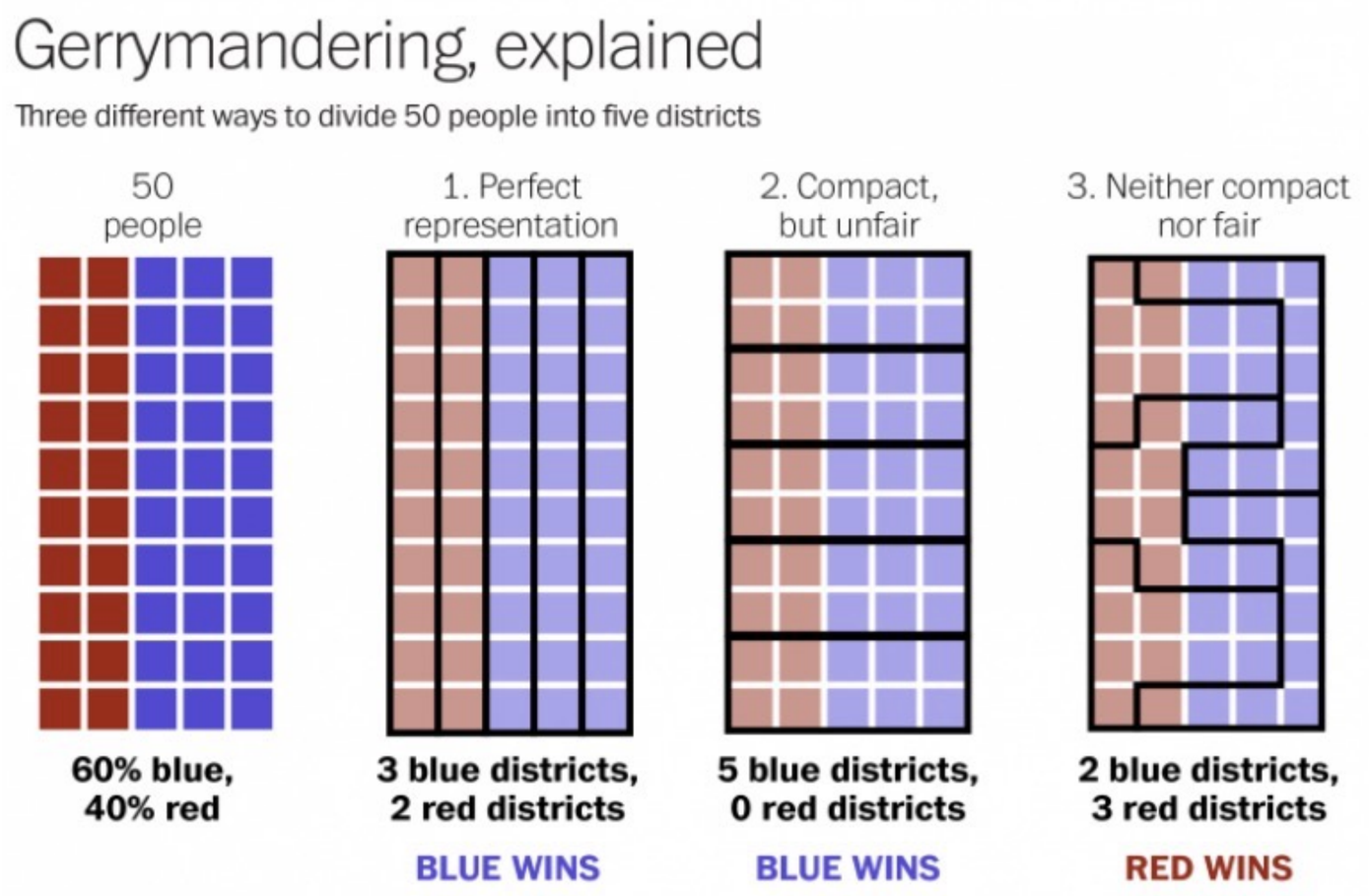
Modifiable Areal Unit Problem (MAUP)

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





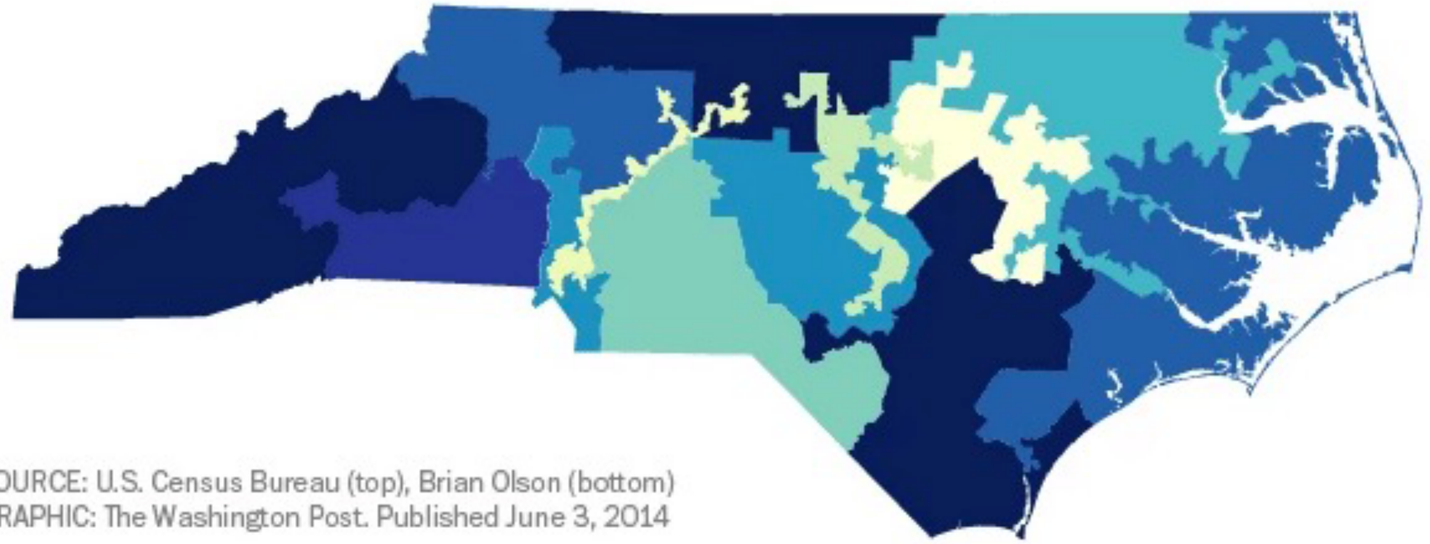
Gerrymandering



Gerrymandering

North Carolina

CURRENT CONGRESSIONAL DISTRICTS



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

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

Modifiable Areal Unit Problem (MUAP)

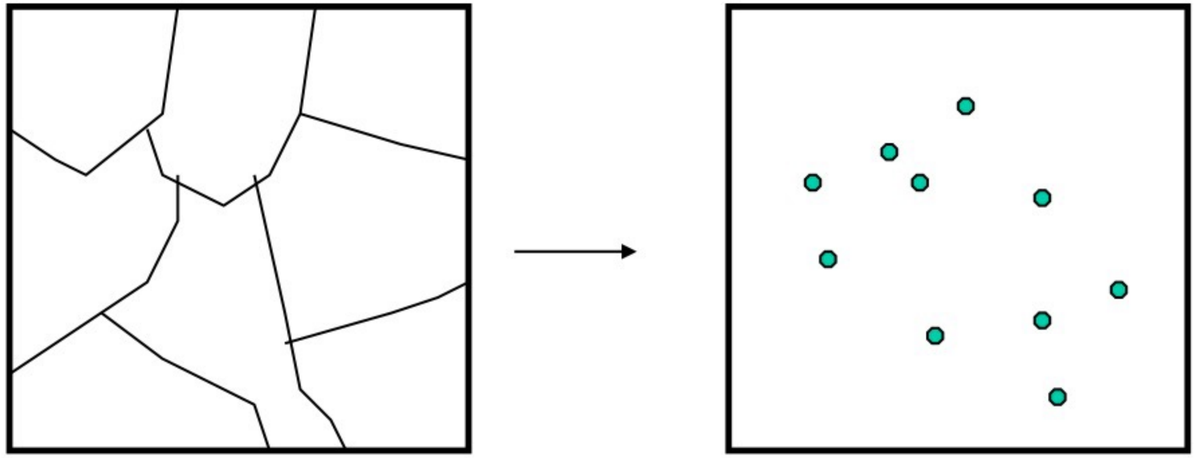
Potential problems in almost every field that utilizes spatial data.

In the 2000 U.S. presidential election, Al Gore, with more of the population vote than George Bush, but failed to become president.

A different aggregation of U.S. counties into states could have produced a different outcome (switch just one northern Florida county to Georgia or Alabama would have produced a different outcome).

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.

Conclusion:

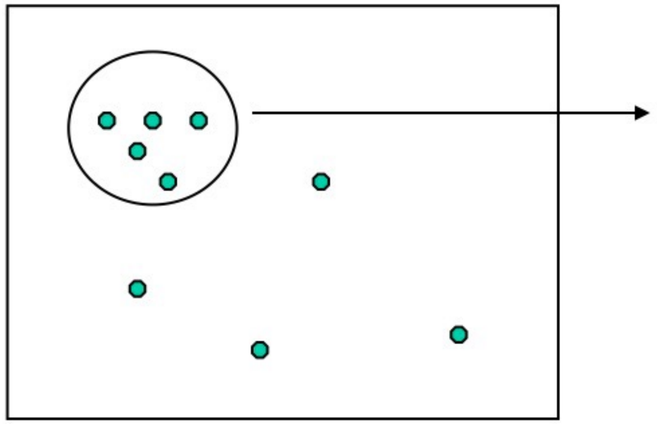
- Lower-income persons are more likely to commit crime
- Lower-income areas are associated with higher crime rates
- Lower-income counties tend to experience higher crime rates

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



Area with high crime rates?

Crime locations

Bank robberies are clustered
But only because banks are clustered!

COGS 9
Introduction to Data Science

Text mining

UC San Diego

Today's learning objective

Explain analytical approaches to analyzing textual data.
Discuss limitations to analyzing text data.

Plurals

Source: McWhorter J, *TED-Ed* 2013

Verb regularization

Table 1 | The 177 irregular verbs studied

Frequency	Verbs	Regularization (%)	Half-life (yr)
10^{-1}–1	be, have	0	38,800
10^{-2}–10^{-1}	come, do, find, get, give, go, know, say, see, take, think	0	14,400
10^{-3}–10^{-2}	begin, break, bring, buy, choose, draw, drink, drive, eat, fall, fight, forget, grow, hang, help, hold, leave, let, lie, lose, reach, rise, run, seek, set, shake, sit, sleep, speak, stand, teach, throw, understand, walk, win, work, write	10	5,400
10^{-4}–10^{-3}	arise, bake, bear, beat, bind, bite, blow, bow, burn, burst, carve, chew, climb, cling, creep, dare, dig, drag, flee, float, flow, fly, fold, freeze, grind, leap, lend, lock, melt, reckon, ride, rush, shape, shine, shoot, shrink, sigh, sing, sink, slide, slip, smoke, spin, spring, starve, steal, step, stretch, strike, stroke, suck, swallow, swear, sweep, swim, swing, tear, wake, wash, weave, weep, weigh, wind, yell, yield	43	2,000
10^{-5}–10^{-4}	bark, bellow, bid, blend, braid, brew, cleave, cringe, crow, dive, drip, fare, fret, glide, gnaw, grip, heave, knead, low, milk, mourn, mow, prescribe, redden, reek, row, scrape, seethe, shear, shed, shove, slay, slit, smite, sow, span, spurn, sting, stink, strew, stride, swell, tread, uproot, wade, warp, wax, wield, wring, writhe	72	700
10^{-6}–10^{-5}	bide, chide, delve, flay, hew, rue, shrive, slink, snip, spew, sup, wreak	91	300

177 Old English irregular verbs were compiled for this study. These are arranged according to frequency bin, and in alphabetical order within each bin. Also shown is the percentage of verbs in each bin that have regularized. The half-life is shown in years. Verbs that have regularized are indicated in red. As we move down the list, an increasingly large fraction of the verbs are red; the frequency-dependent regularization of irregular verbs becomes immediately apparent.

Source: Lieberman *et al.*, *Nature* 2007

Culturomics

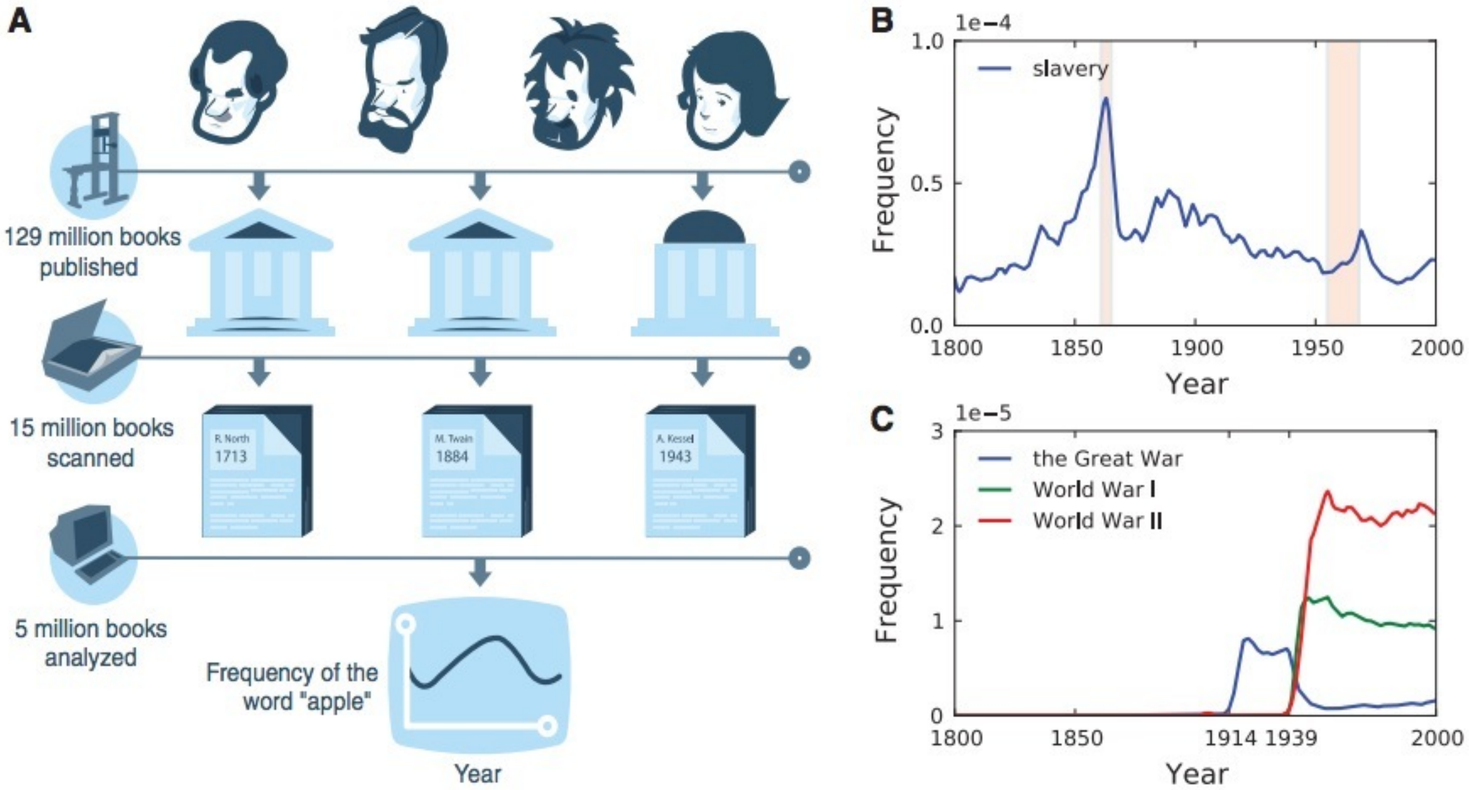
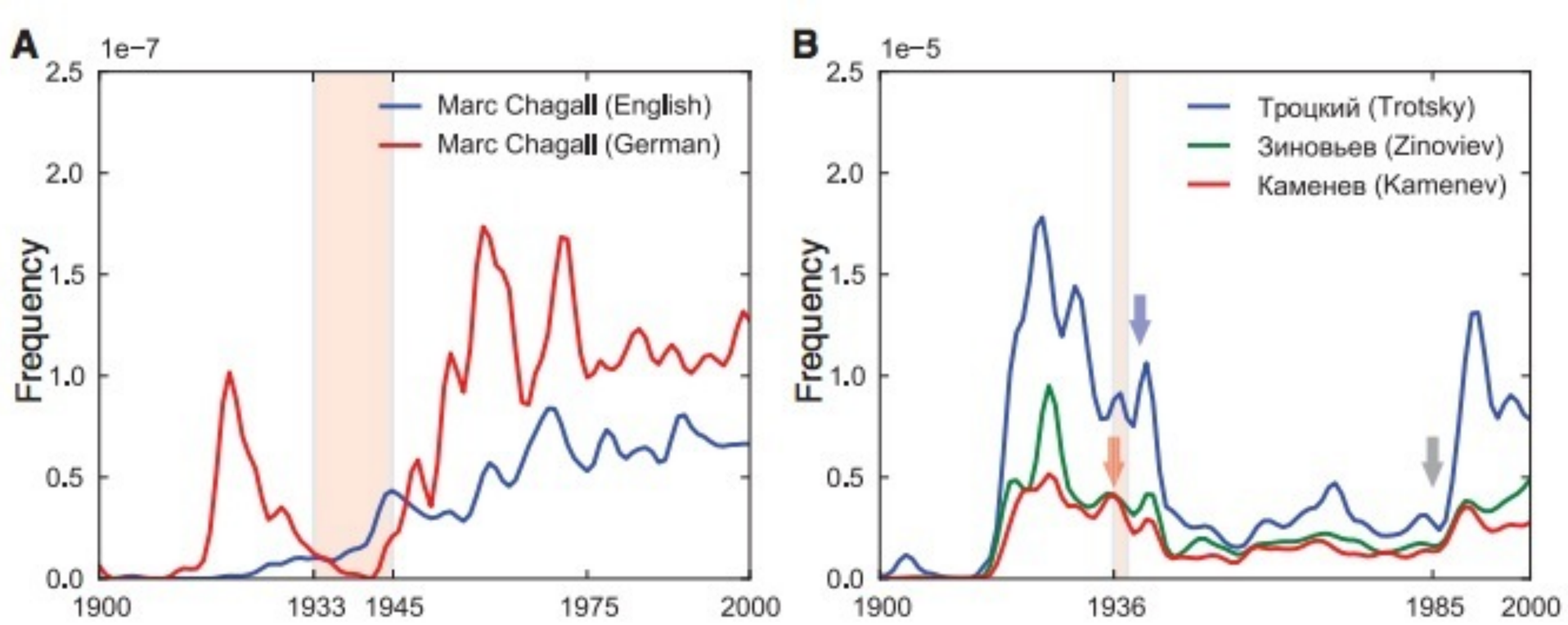


Fig. 1. Culturomic analyses study millions of books at once. **(A)** Top row: Authors have been writing for millennia; ~129 million book editions have been published since the advent of the printing press (upper left). Second row: Libraries and publishing houses provide books to Google for scanning (middle left). Over 15 million books have been digitized. Third row: Each book is associated with metadata. Five million books are chosen for computational analysis (bottom left). Bottom row: A culturomic time line shows the frequency of "apple" in English books over time (1800–2000). **(B)** Usage frequency of "slavery". The Civil War (1861–1865) and the civil rights movement (1955–1968) are highlighted in red. The number in the upper left (1e-4 = 10^{-4}) is the unit of frequency. **(C)** Usage frequency over time for "the Great War" (blue), "World War I" (green), and "World War II" (red).

Source: Michel *et al.*, *Science* 2011

UC San Diego

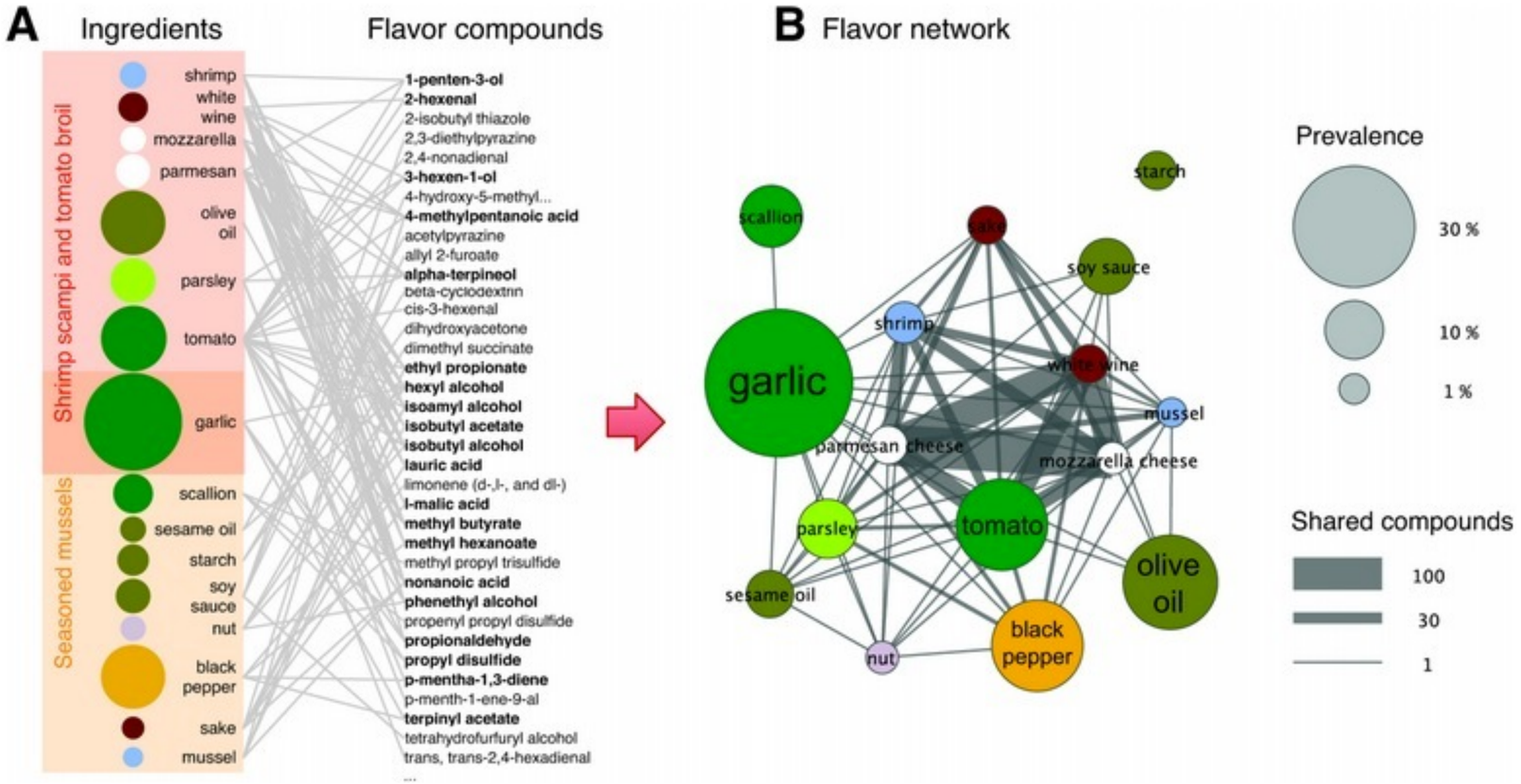
Culturomics



Source: Michel *et al.*, *Science* 2011

UC San Diego

Flavor networks



Source: Ahn *et al.*, *SciRep* 2011

UC San Diego

What does data mining get us?

Google