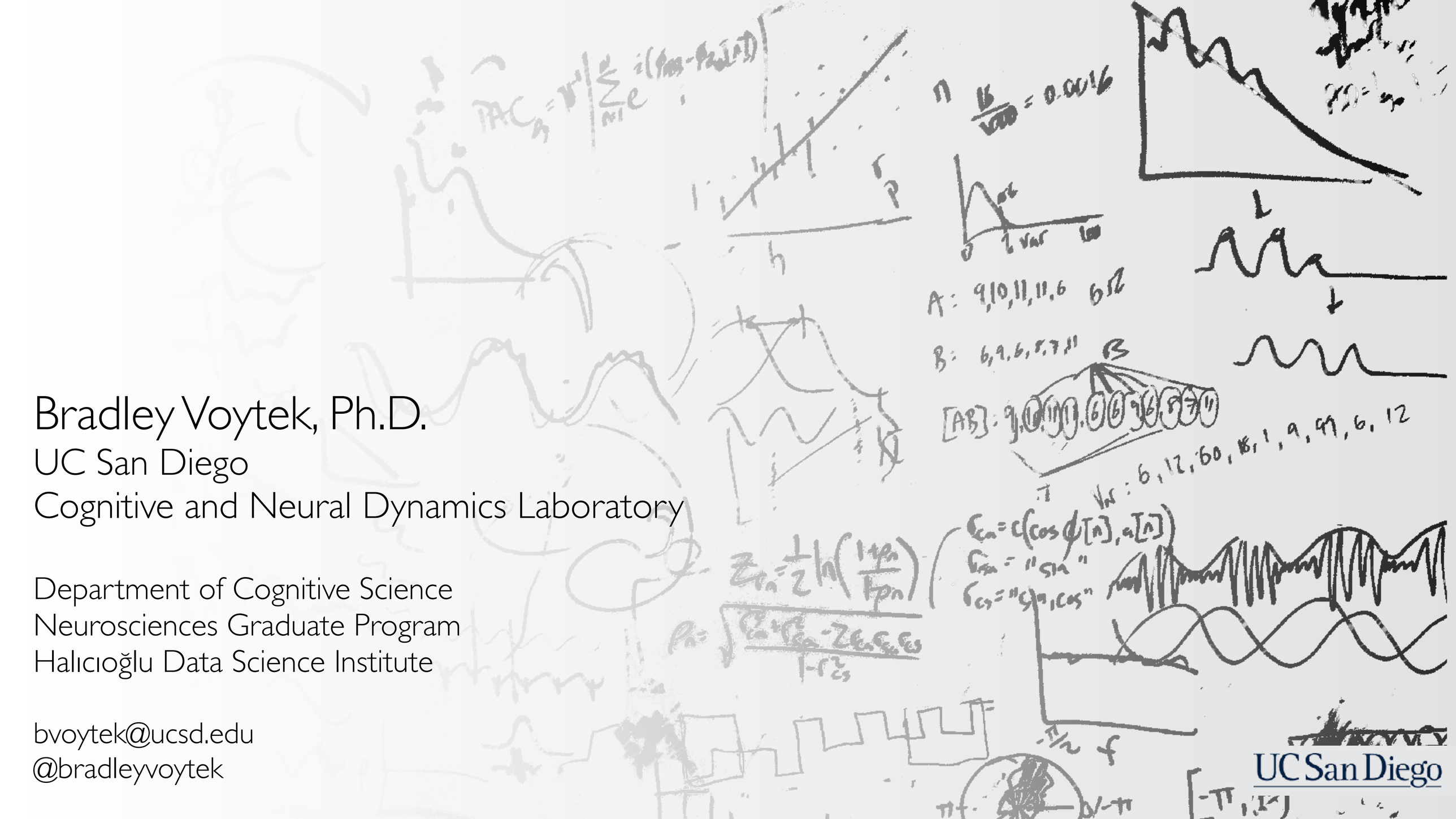


Bradley Voytek, Ph.D.

UC San Diego

Cognitive and Neural Dynamics Laboratory

Department of Cognitive Science

Neurosciences Graduate Program

Halıcıoğlu Data Science Institute

bvoytek@ucsd.edu

@bradleyvoytek

UC San Diego

COGS 108
Data Science in Practice

Geospatial Analyses

UC San Diego

Geocoding

Geocoding is the process of converting addresses (like a street address) into geographic coordinates (like latitude and longitude), which you can use to place markers on a map, or position the map.

Source: Google

Google Maps API

Client-side geocoding, which is executed in the browser, generally in response to user action. The Google Maps JavaScript API provides classes that make the requests for you.

UC San Diego

Source: Google

Google Maps API

When to use client-side geocoding

The short answer is "almost always.” The reasons are:

- Client-side request and response provide a faster, more interactive experience for users.
- A client-side request can include information that improves geocoding quality: user language, region, and viewport.

Source: Google

Google Maps API

HTTP server-side geocoding, which allows your server to directly query Google's servers for geocodes. The Google Maps Geocoding API is the web service that provides this functionality. Typically, you integrate this service with other code that is running server-side.

Source: Google

Google Maps API

When to use server-side geocoding

Server-side geocoding is best used for applications that require you to geocode addresses without input from a client. A common example is when you get a dataset that comes independently of user input, for instance if you have a fixed, finite, and known set of addresses that need geocoding. Server-side geocoding can also be useful as a backup for when client-side geocoding fails.

Source: Google

Google Maps API

The Google Maps Geocoding API has the following limits in place:

Standard Usage Limits

Users of the standard API:

- 2,500 free requests per day, calculated as the sum of client-side and server-side queries.
- 50 requests per second, calculated as the sum of client-side and server-side queries.

UC San Diego

Source: Google

Geocoding

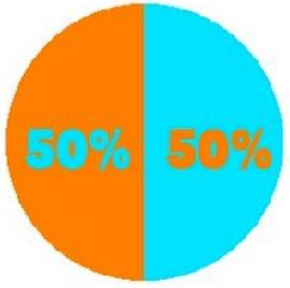
Reverse geocoding is the process of converting geographic coordinates into a human-readable address. The Google Maps Geocoding API's reverse geocoding service also lets you find the address for a given place ID.

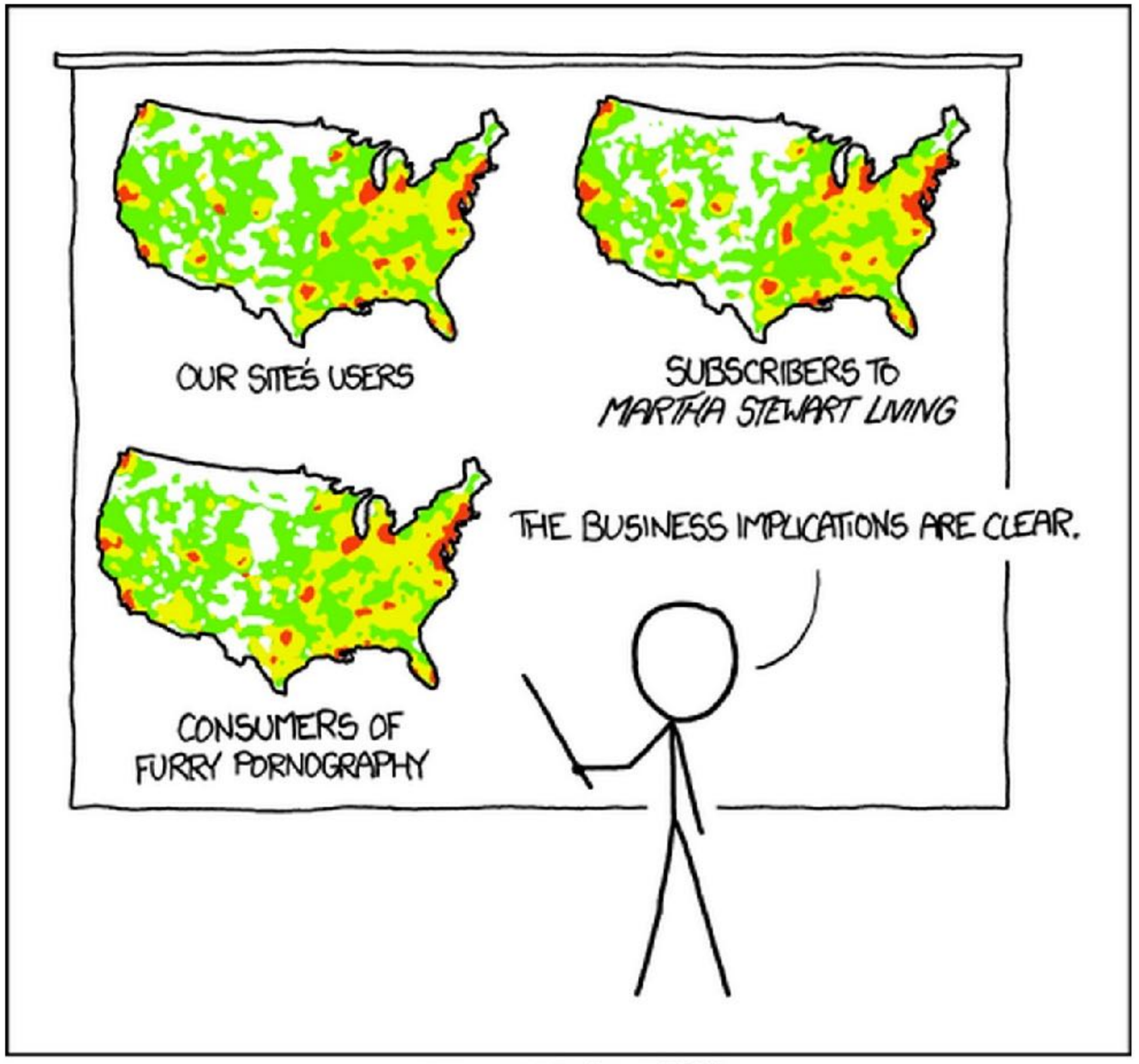
Source: Google

Geocoding

Jupyter Demo!

UNITED STATES ECONOMIC ACTIVITY, SPLIT IN HALF





PET PEEVE #208:

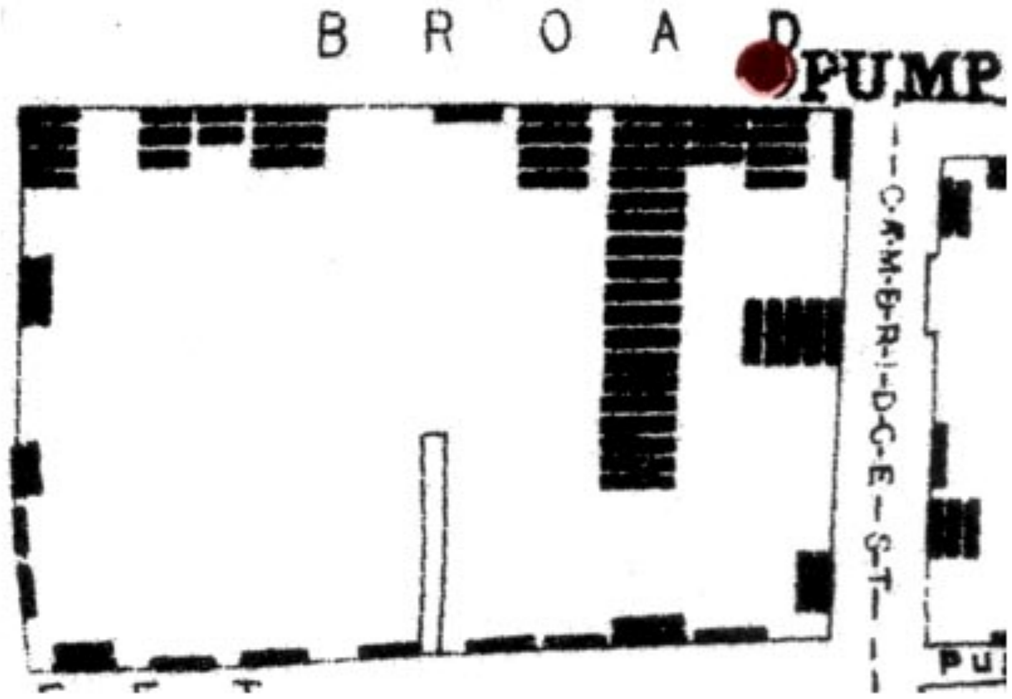
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

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

Cholera



John Snow 1850s map of cholera in London

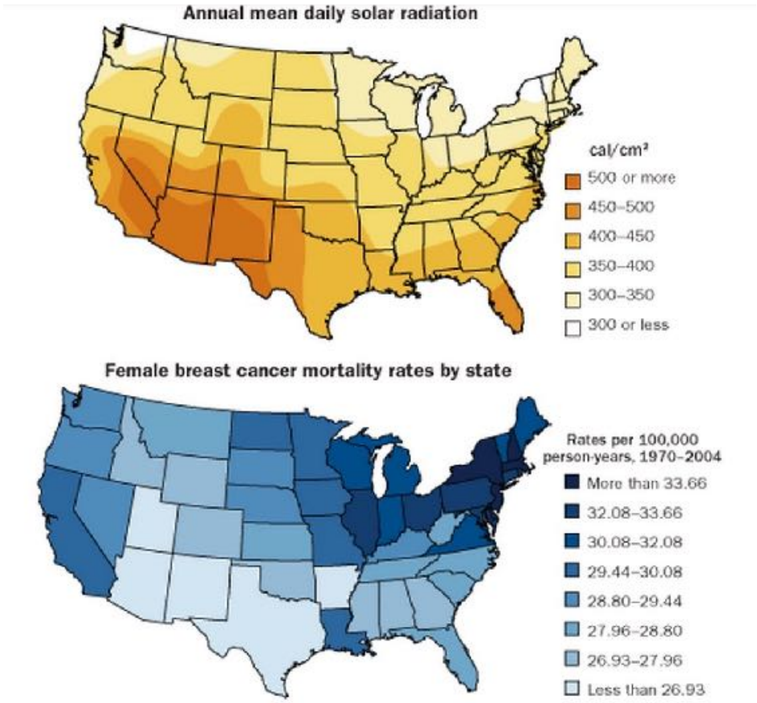
ESDA or hypothesis testing?

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?

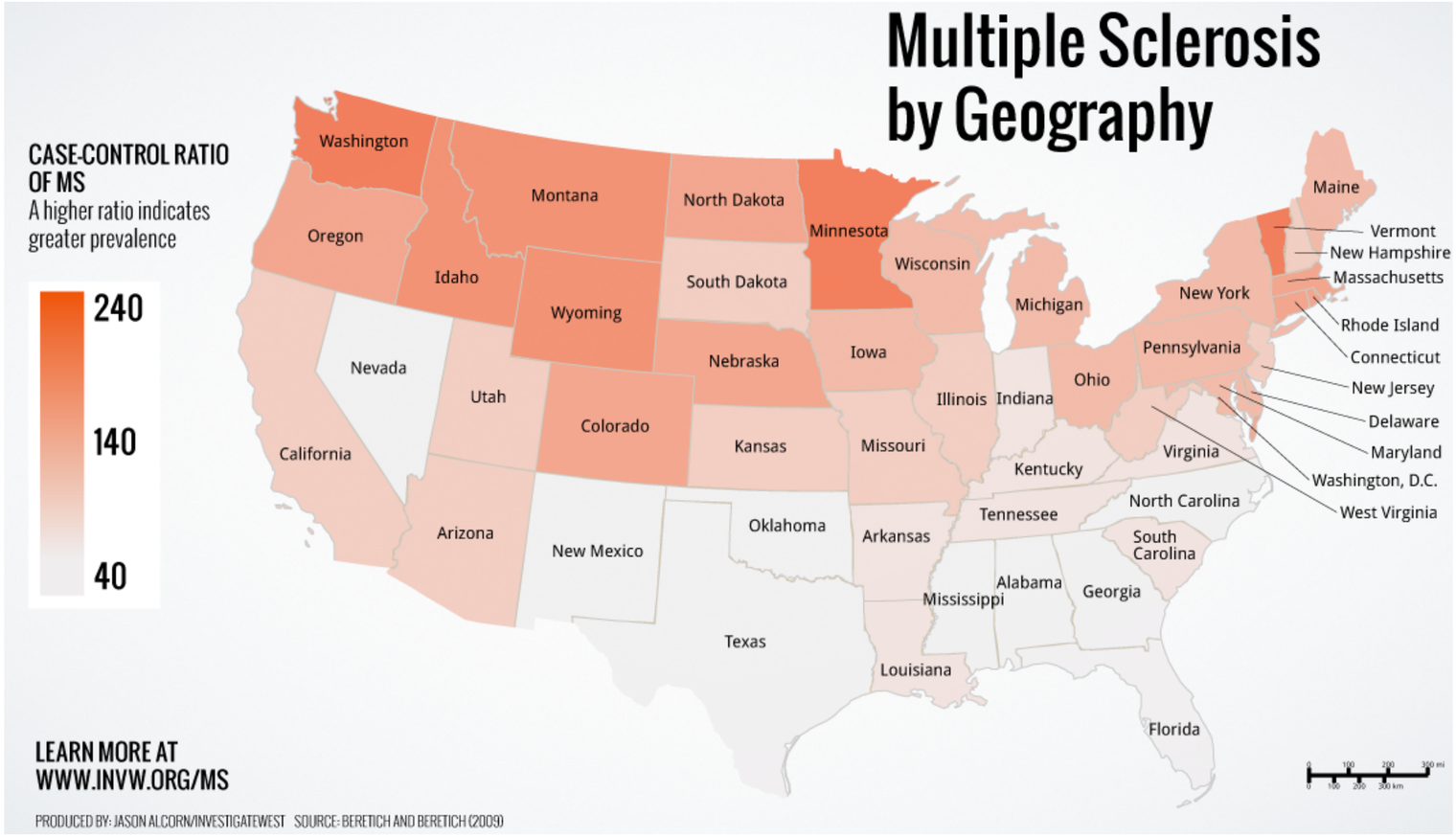
Geospatial disease distribution



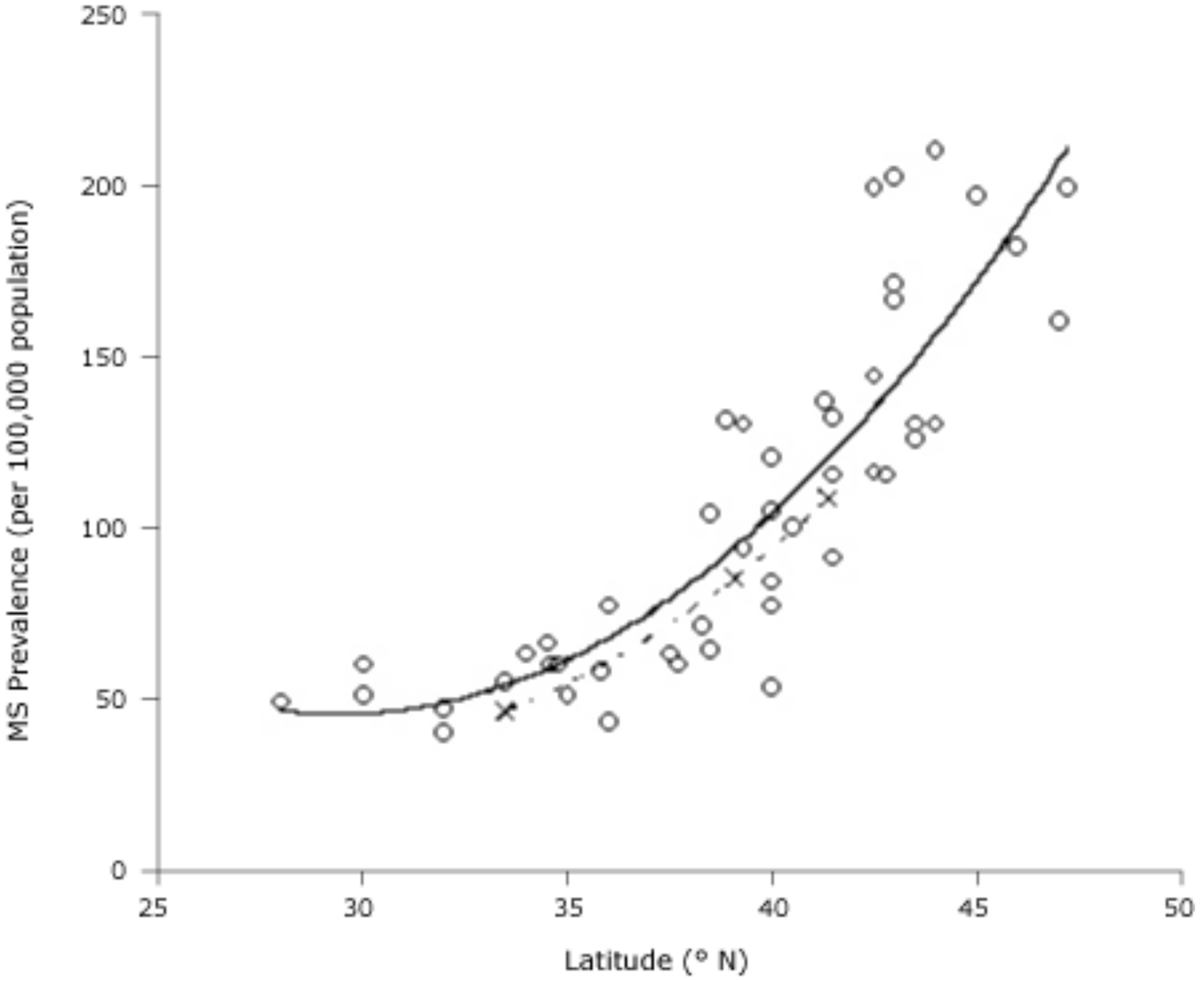
ON THE MAP Scientists who study vitamin D can't help but notice that a host of diseases seem to vary with latitude. Type 1 diabetes, multiple sclerosis and even some cancers appear to be more common in areas that get less sun -- meaning less opportunity for the body to produce vitamin D. The maps above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



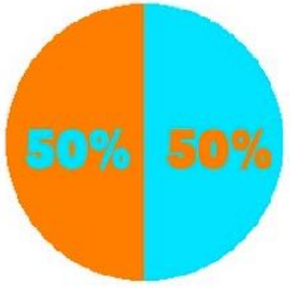
Source: CDC

Good news and bad news

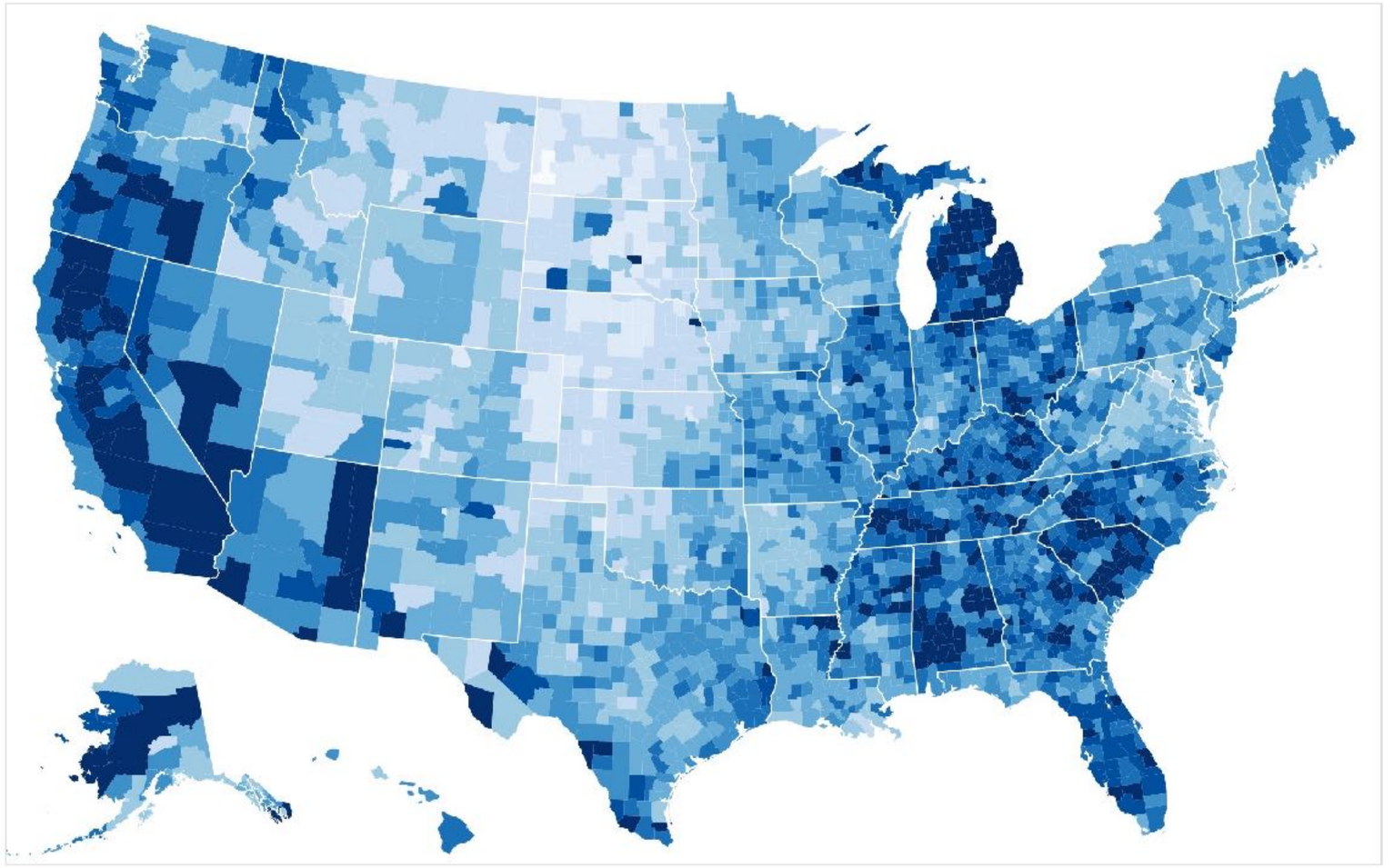
Bad news: Many of the standard techniques and methods documented in standard statistics textbooks have significant problems when we try to apply them to the analysis of the spatial distributions.

Good news: Geospatial referencing provides us with a number of new ways of looking at data and the relations among them. (e.g. distance, adjacency, interaction, and neighbor)

UNITED STATES ECONOMIC ACTIVITY, SPLIT IN HALF



Choropleth maps

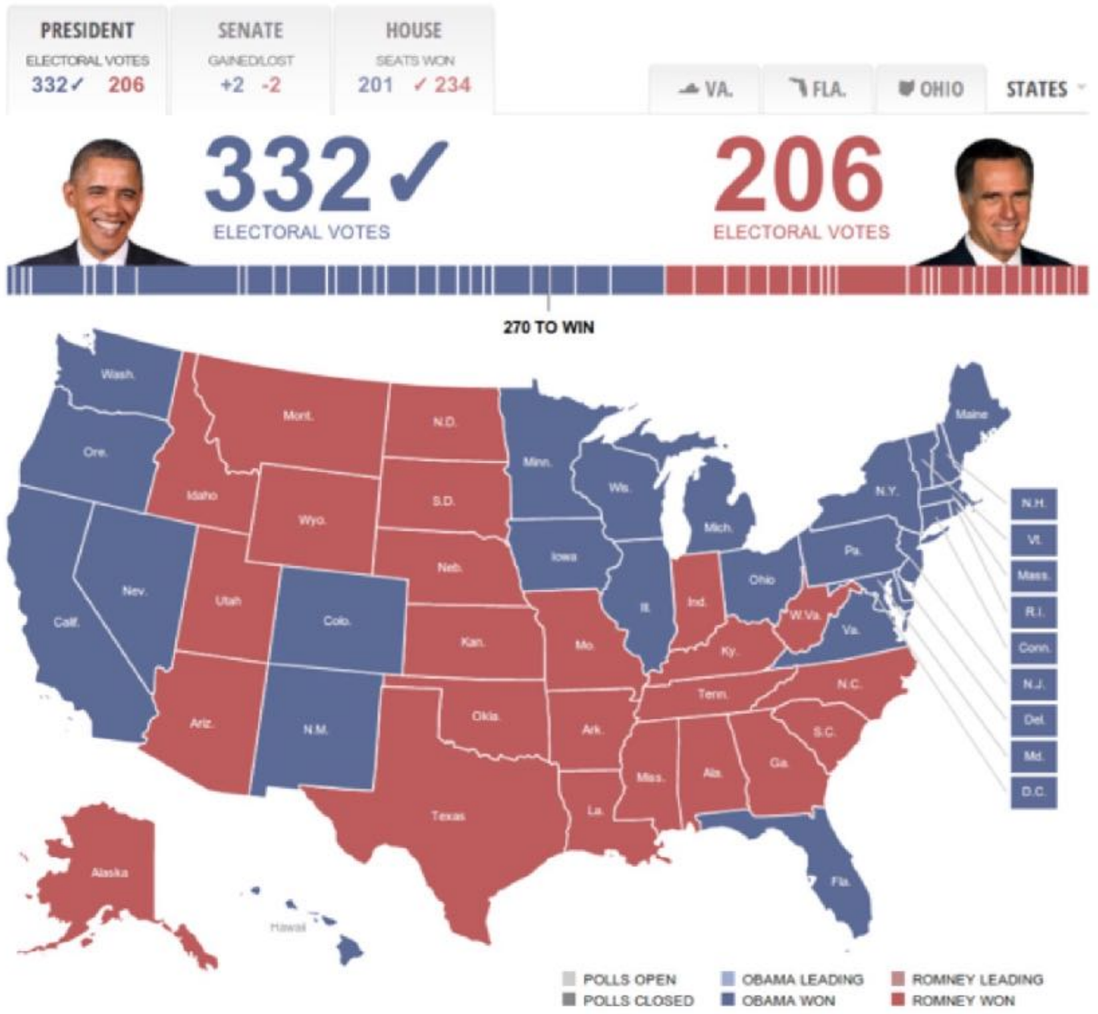


This choropleth encodes unemployment rates from 2008 with a quantize scale ranging from 0 to 15%. A threshold scale is a useful alternative for coloring arbitrary ranges.

UC San Diego

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

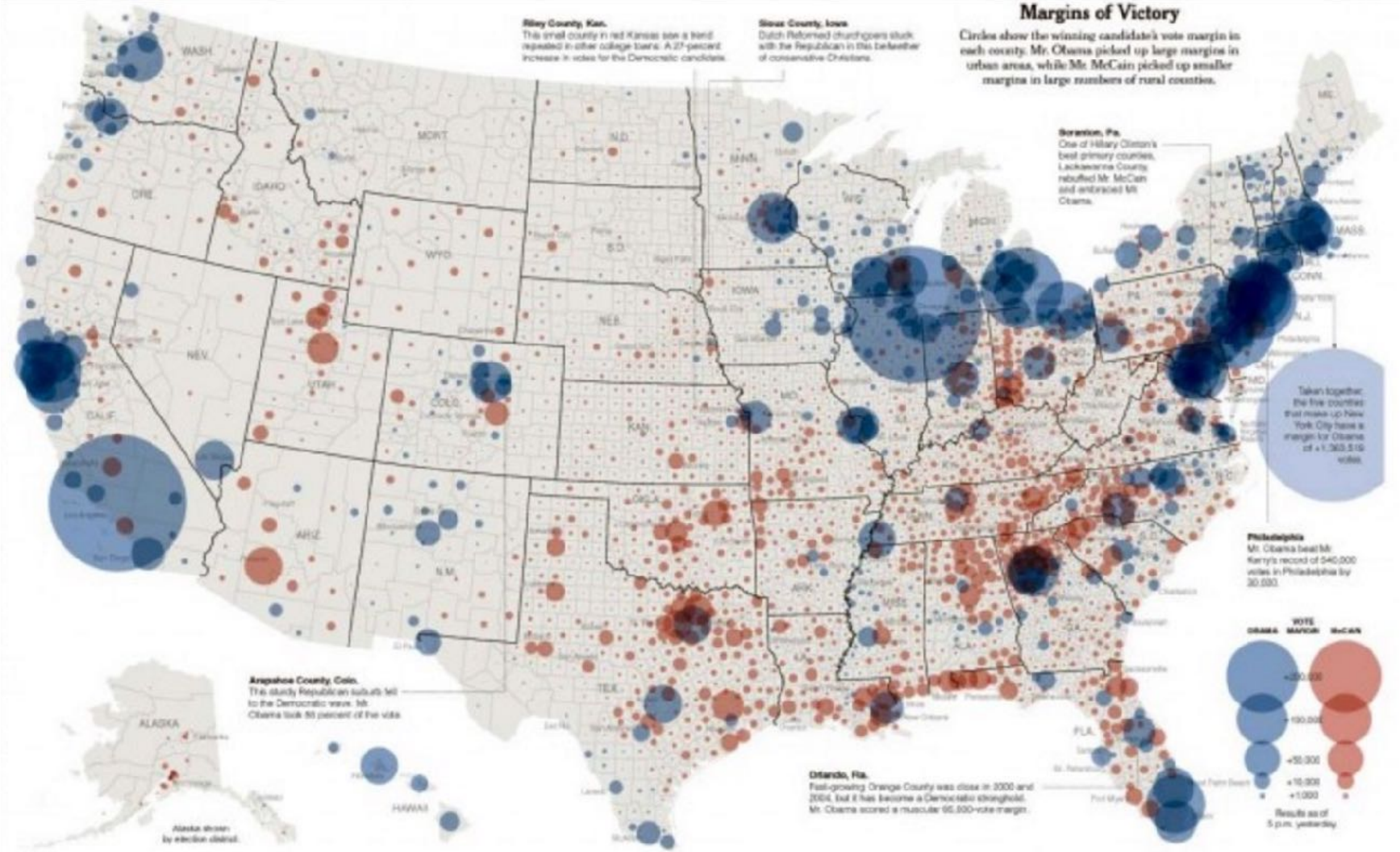
Misleading Choropleth maps



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

UC San Diego

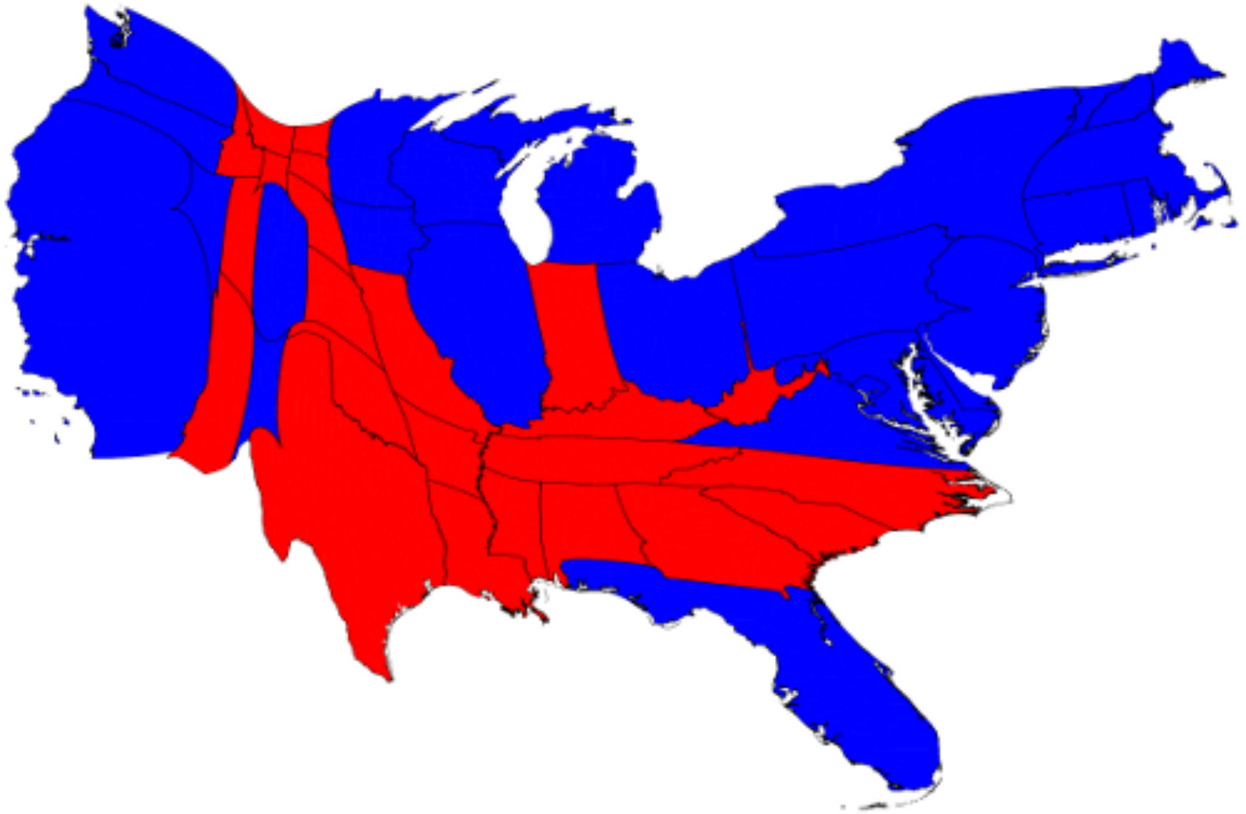
Better Choropleth maps



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

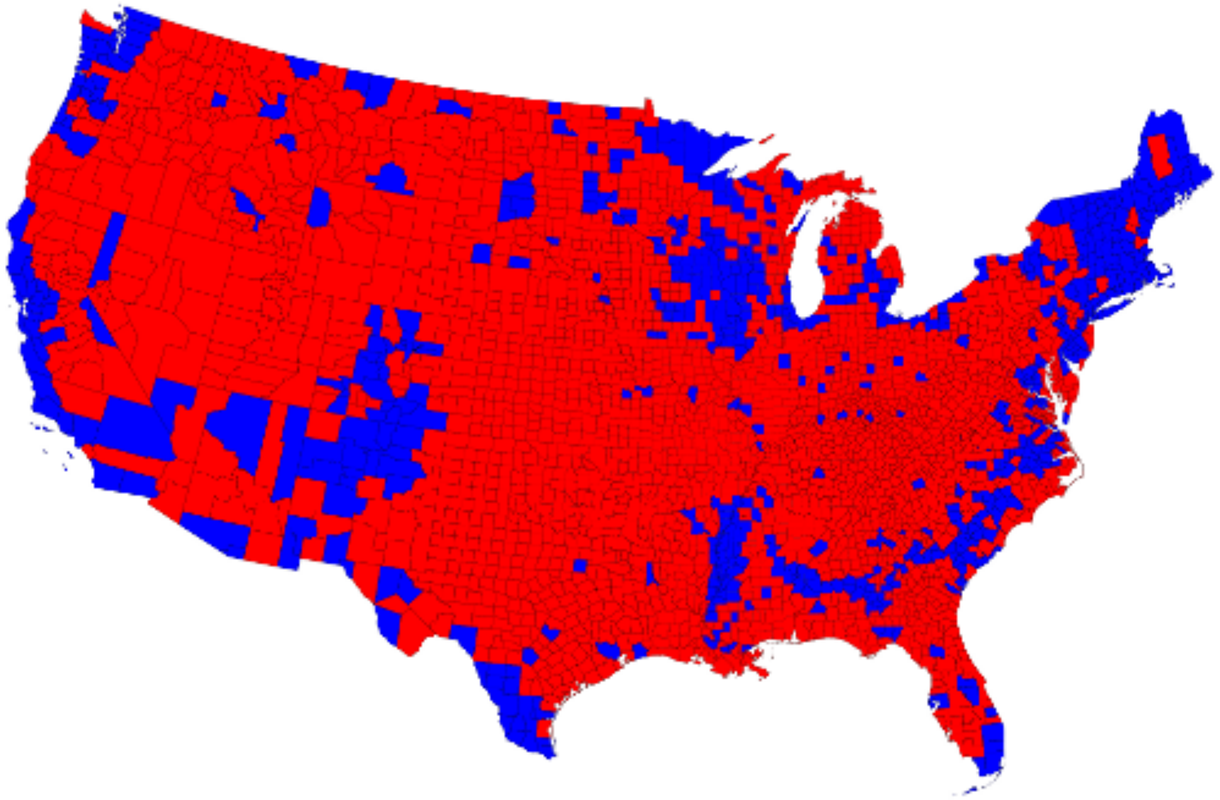
UC San Diego

Cartogram



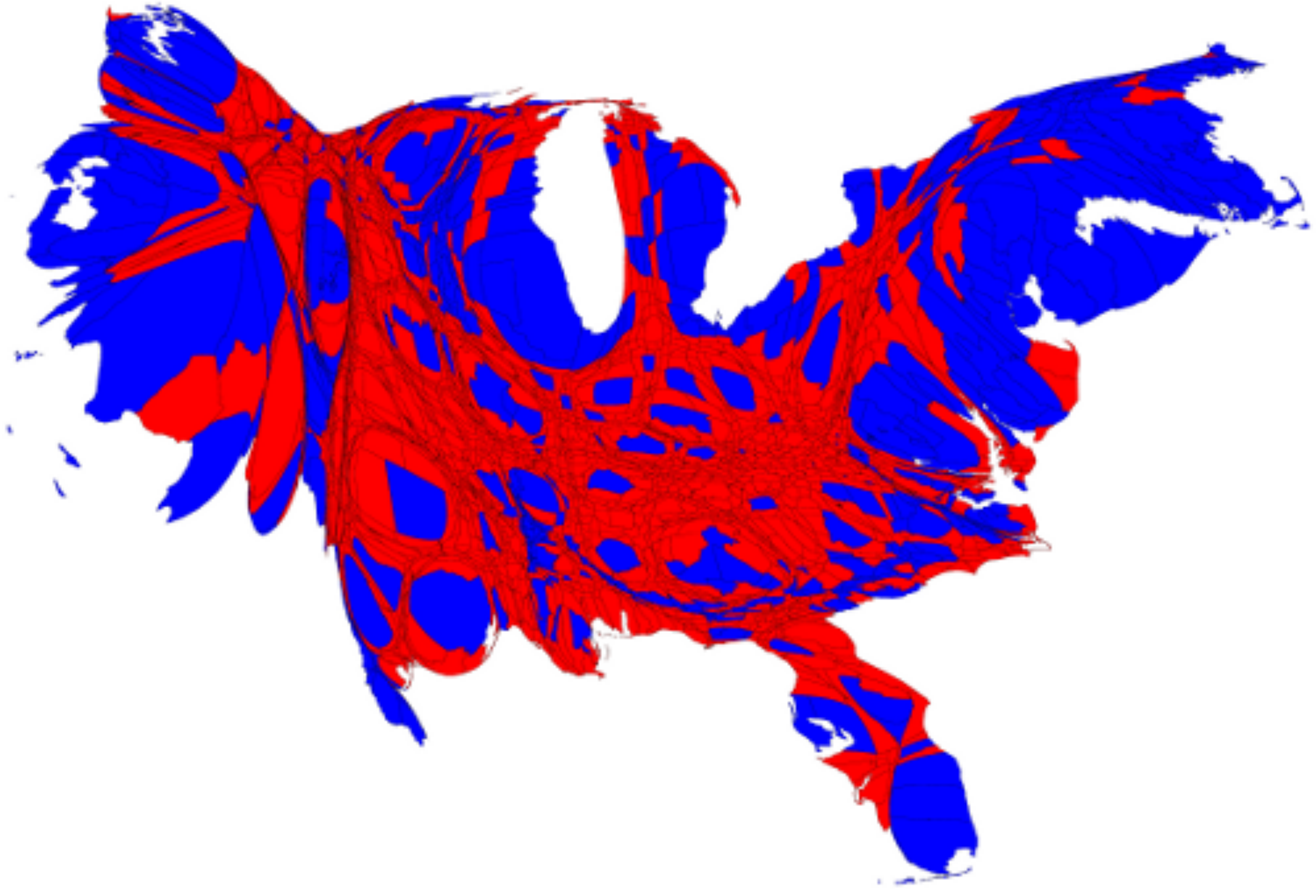
Source: http://www-personal.umich.edu/~mejn/election/2012/

Isarithmic map



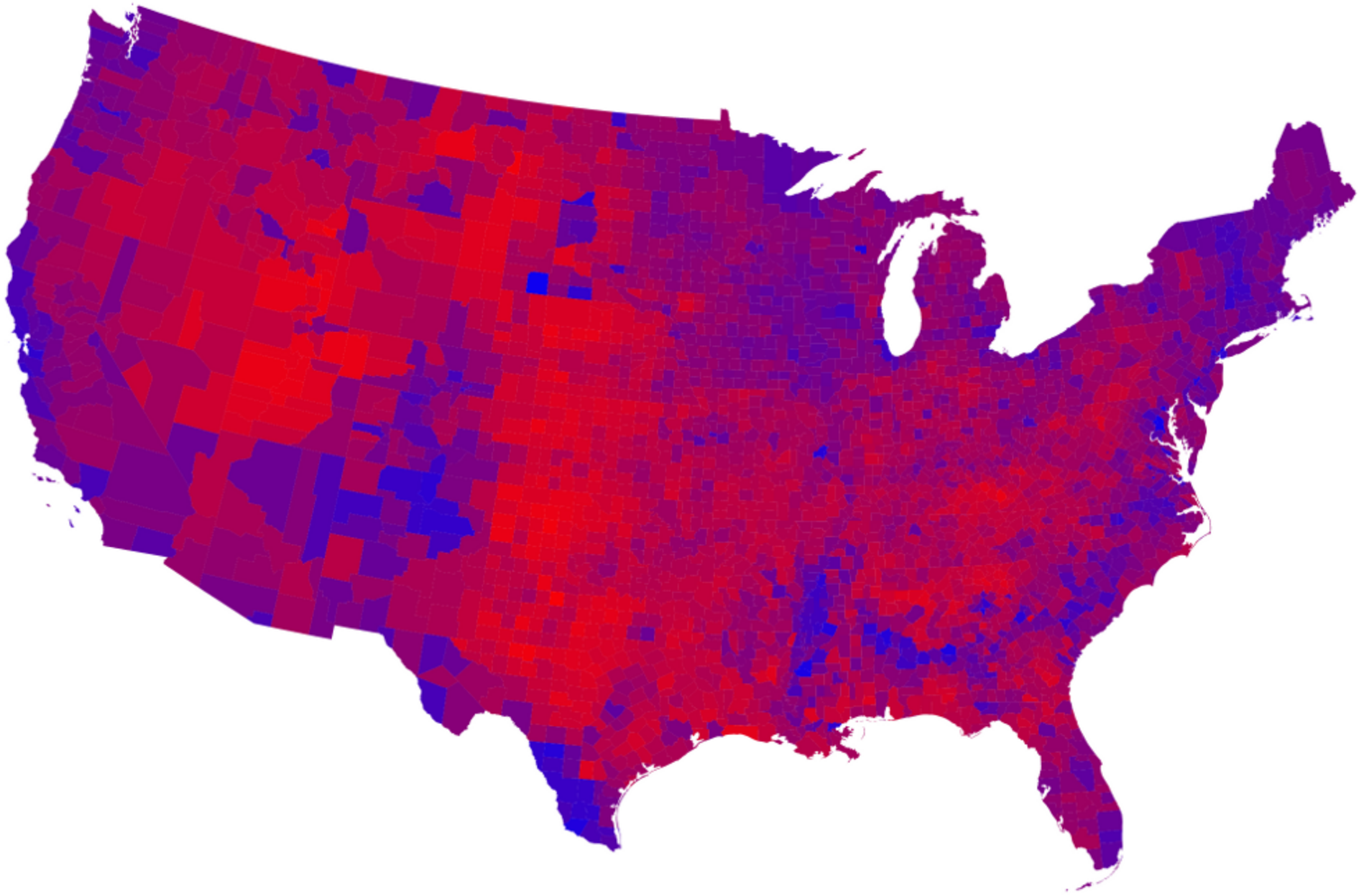
Source: http://www-personal.umich.edu/~mejn/election/2012/

Cartogram



Source: http://www-personal.umich.edu/~mejn/election/2012/

Cartogram



Source: http://www-personal.umich.edu/~mejn/election/2012/

Spatial statistics

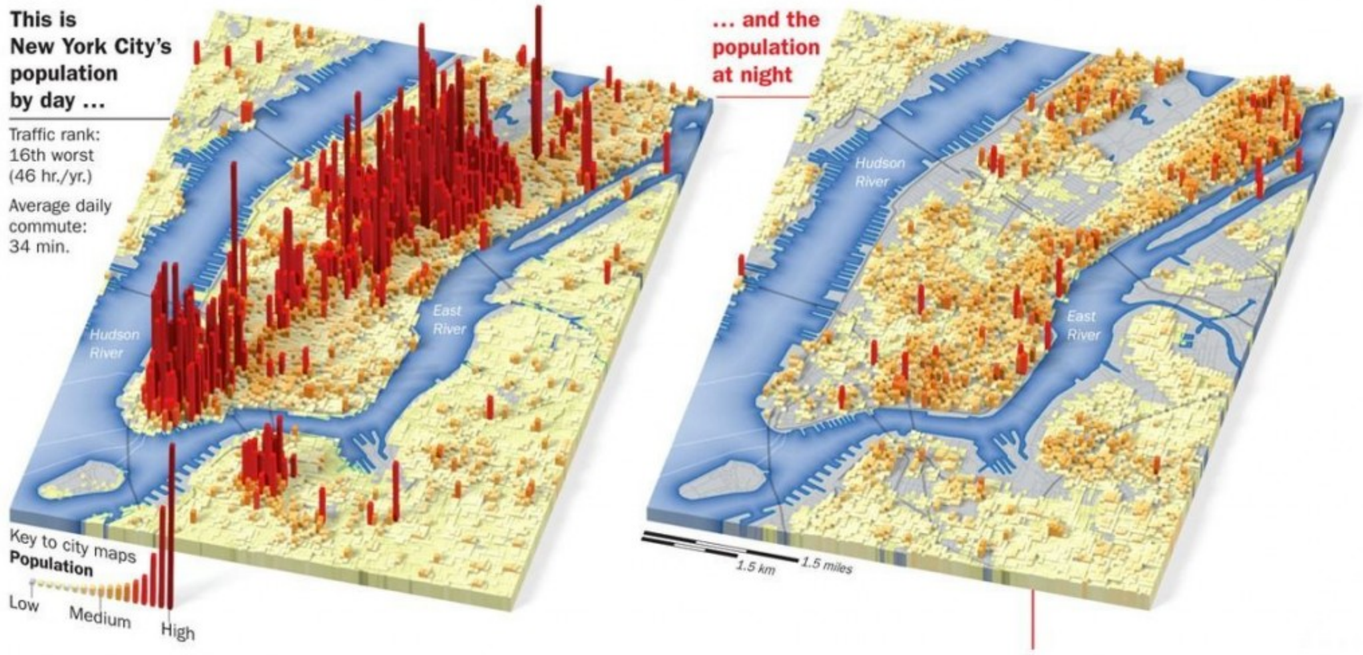
Spatial data violate the requirements of conventional statistics:

- 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

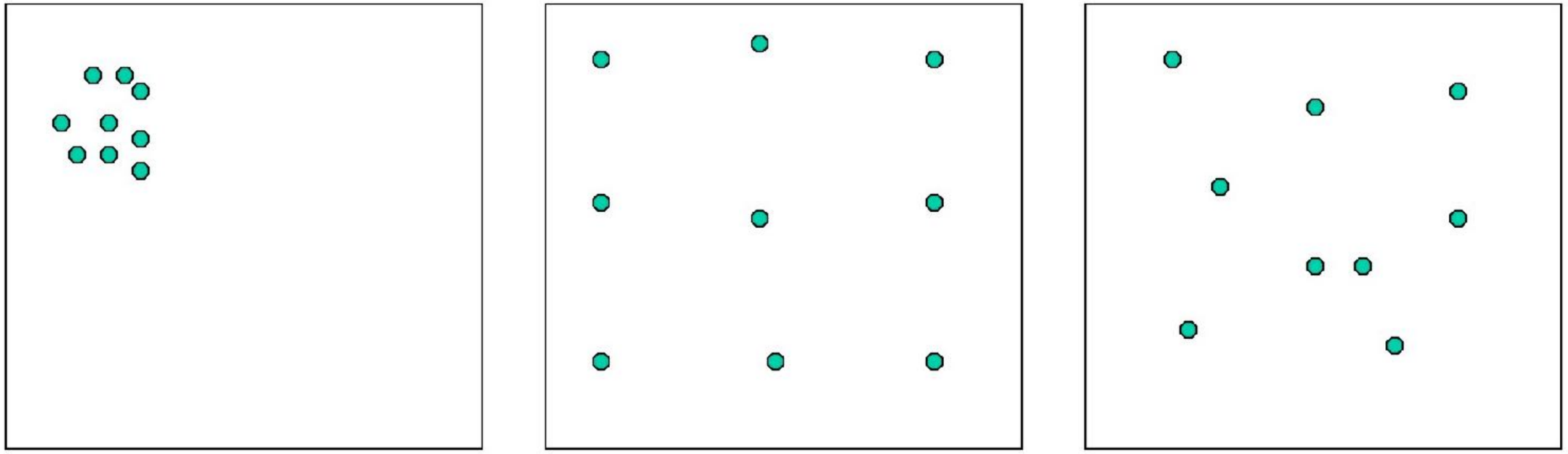


Spatial autocorrelation

Three general possibilities:

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

Spatial autocorrelation



Positive

Negative

Zero (Random)

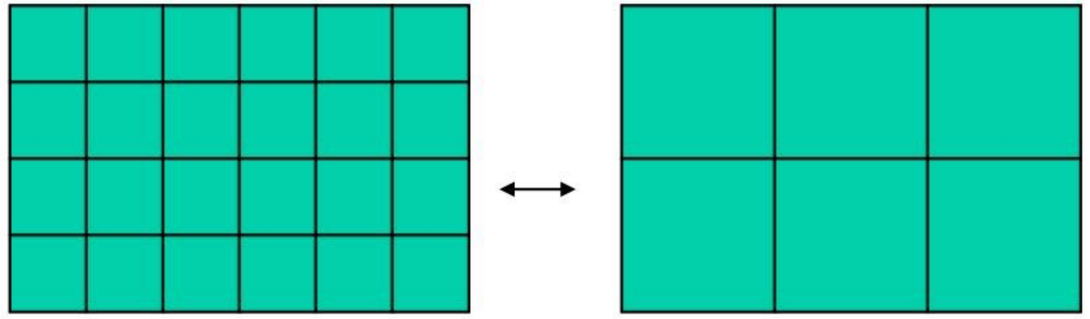
Modifiable Areal Unit Problem (MAUP)

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)

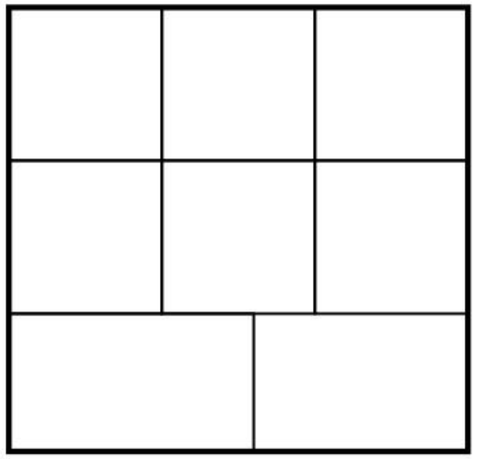
Scale issue: involves the aggregation of smaller units into larger ones. Generally speaking, the larger the spatial units, the stronger the relationship among variables.

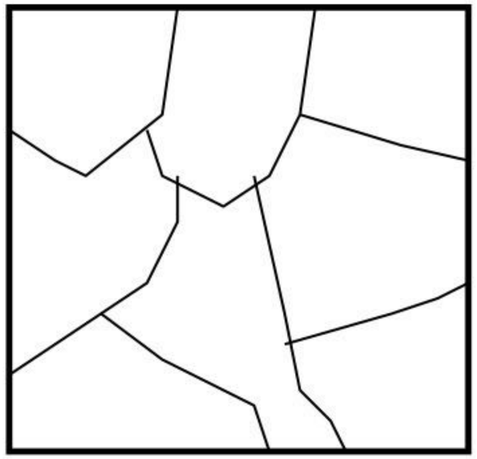


Aggregation (smoothed)

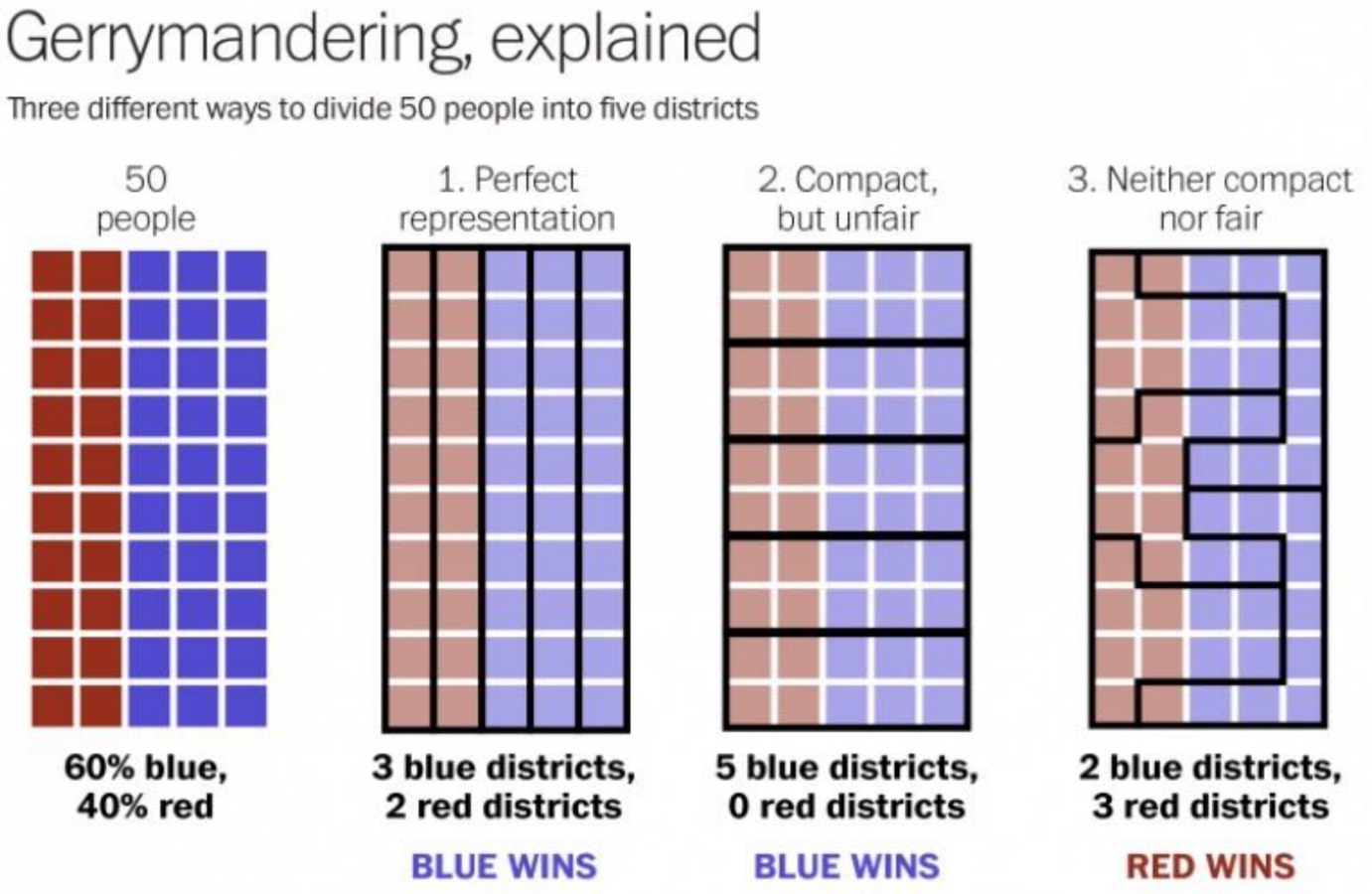
Modifiable Areal Unit Problem (MAUP)

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





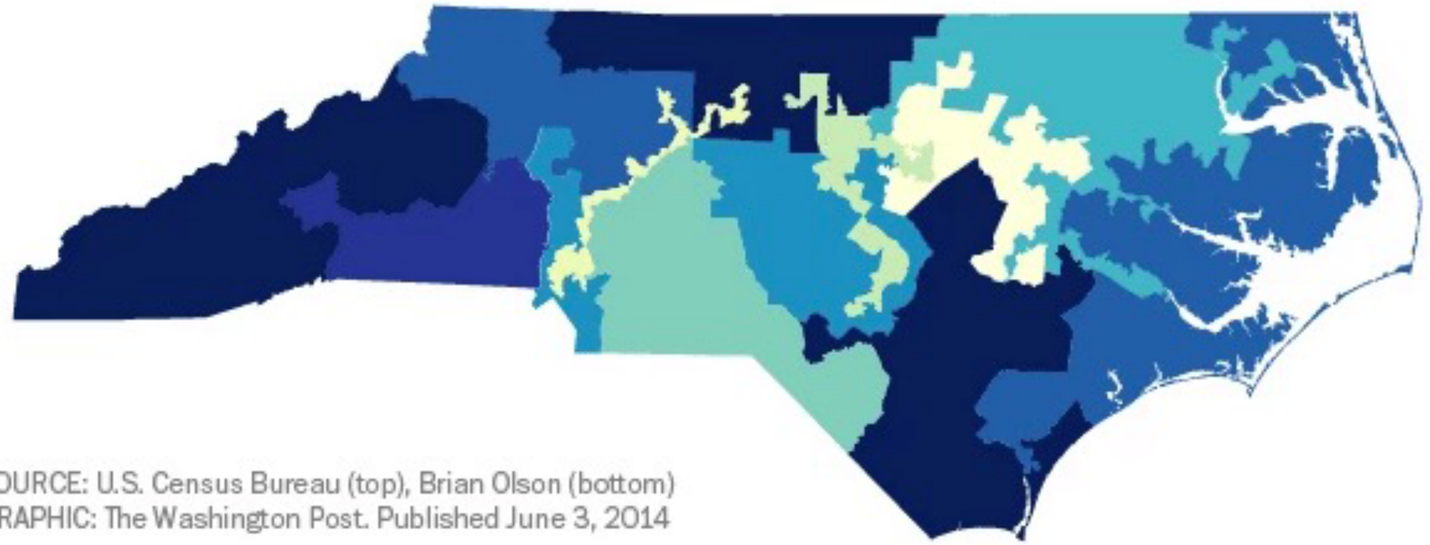
Gerrymandering



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 (MAUP)

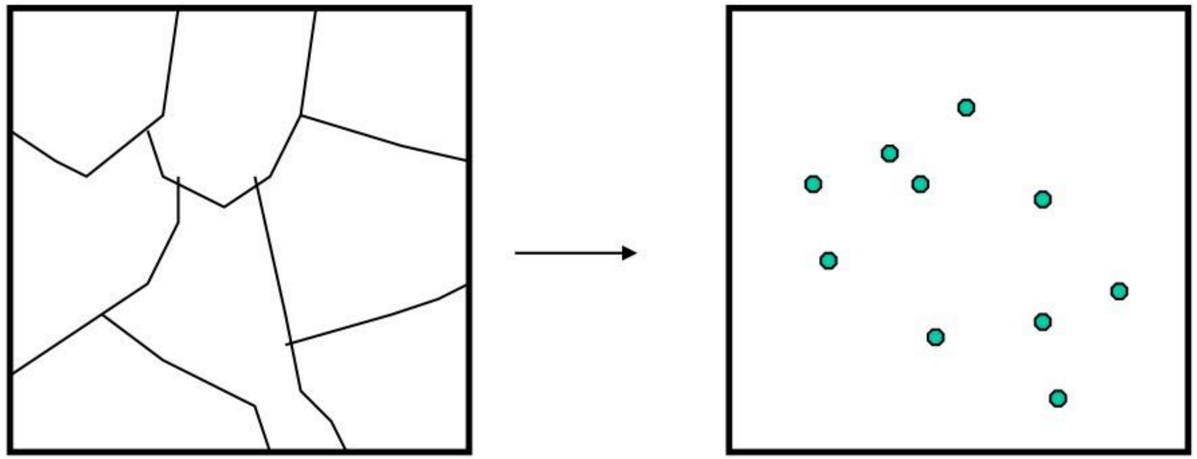
Potential problems in almost every field that utilizes spatial data.

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:

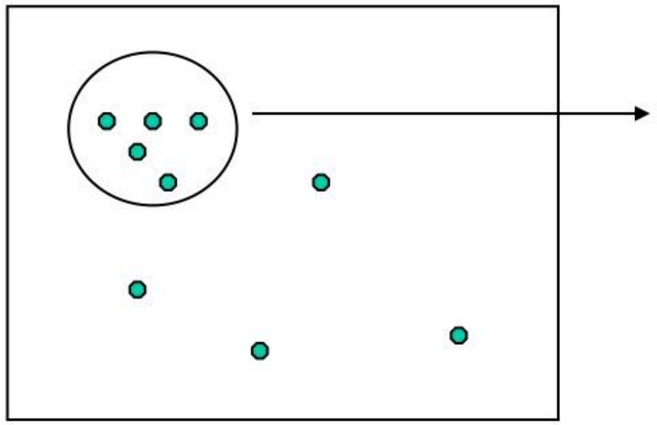
- Lower-income persons are more likely to commit crime
- Lower-income areas are associated with higher crime rates
- Lower-income counties tend to experience higher crime rates

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

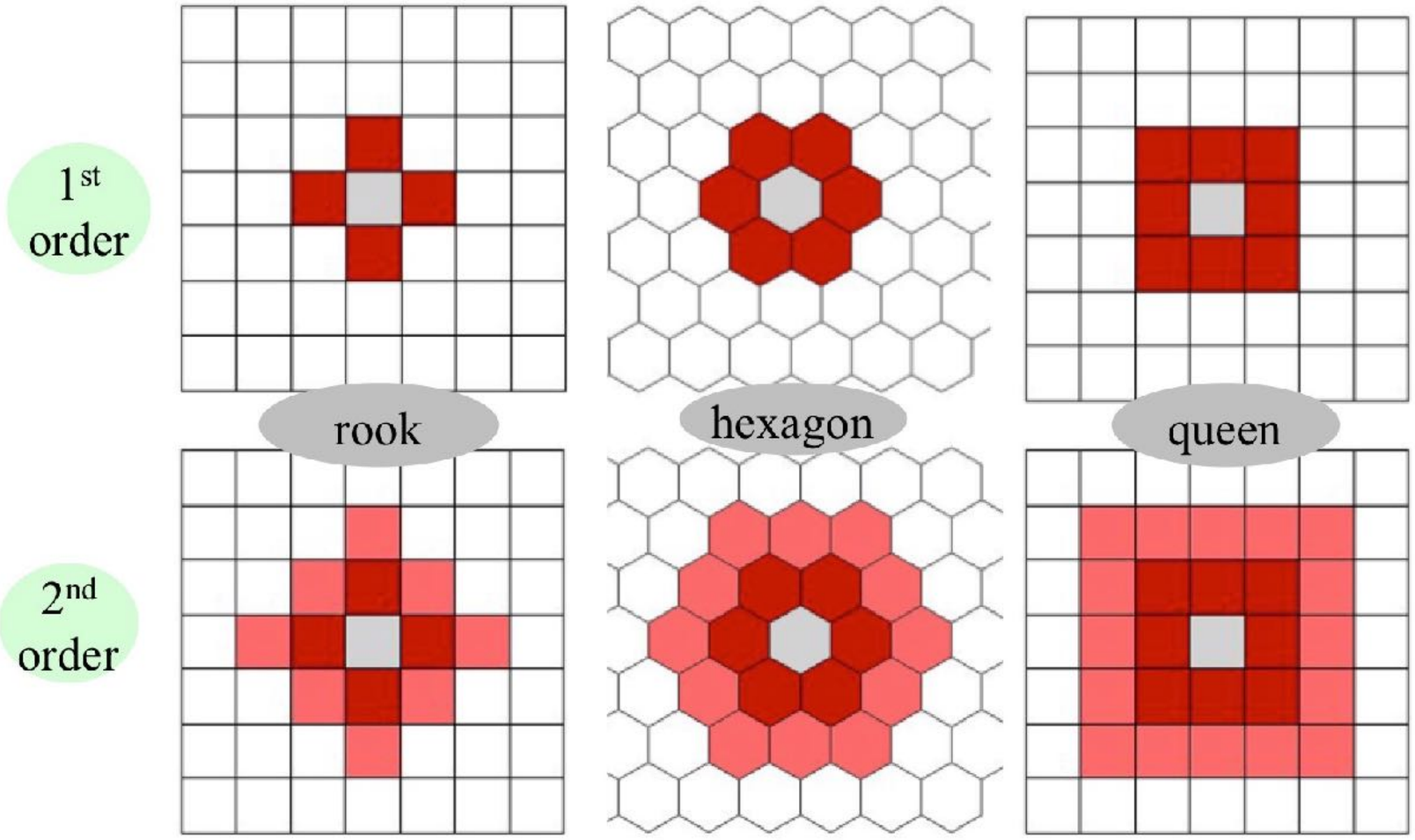


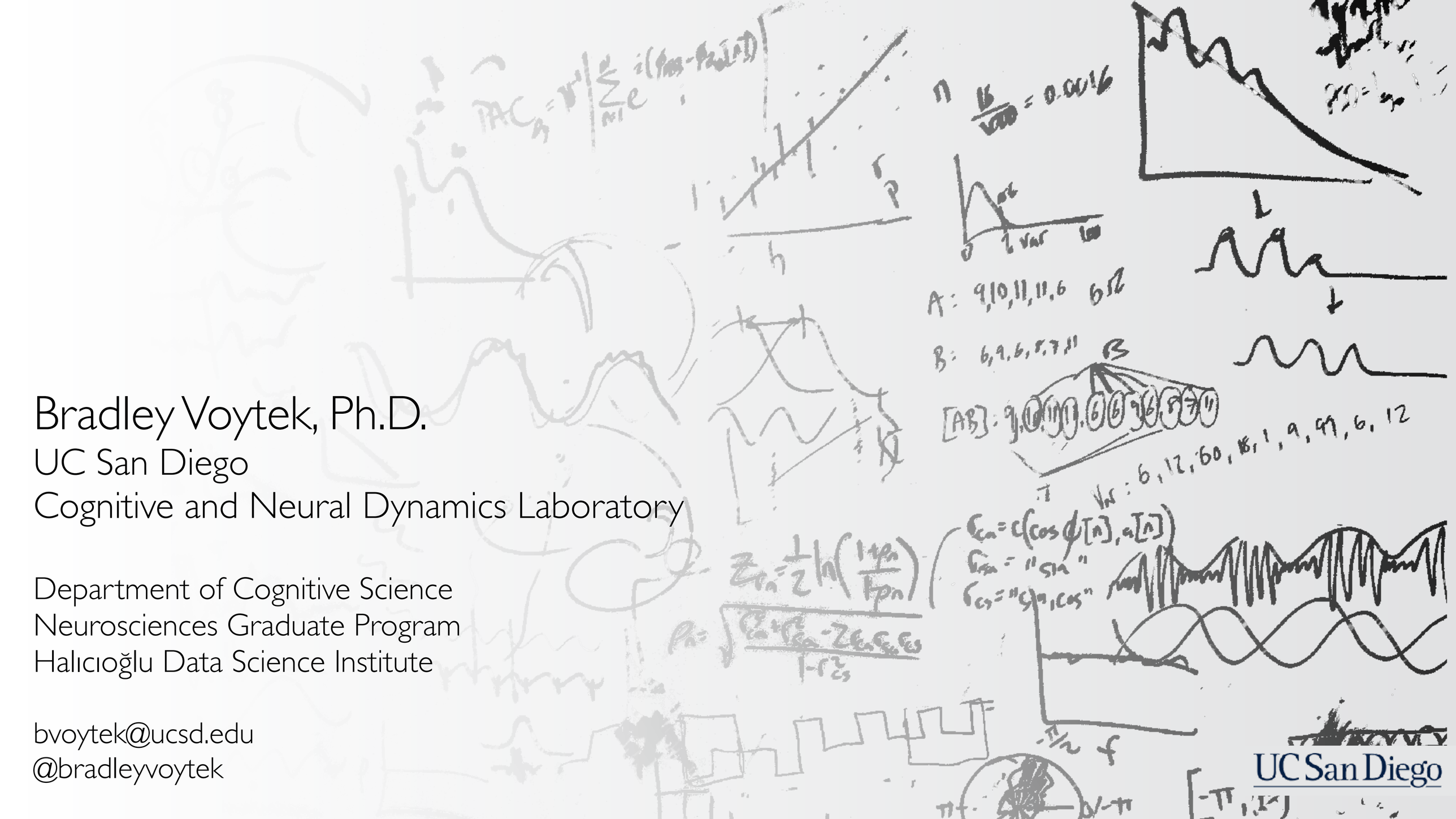
Area with high crime rates?

Crime locations

Bank robberies are clustered
But only because banks are clustered!

Clusters





Bradley Voytek, Ph.D.

UC San Diego

Cognitive and Neural Dynamics Laboratory

Department of Cognitive Science

Neurosciences Graduate Program

Halıcıoğlu Data Science Institute

bvoytek@ucsd.edu

@bradleyvoytek

UC San Diego