An Intro to Spatial Data Analysis in R

Nabil Abdurehman
github.com/nabilabd

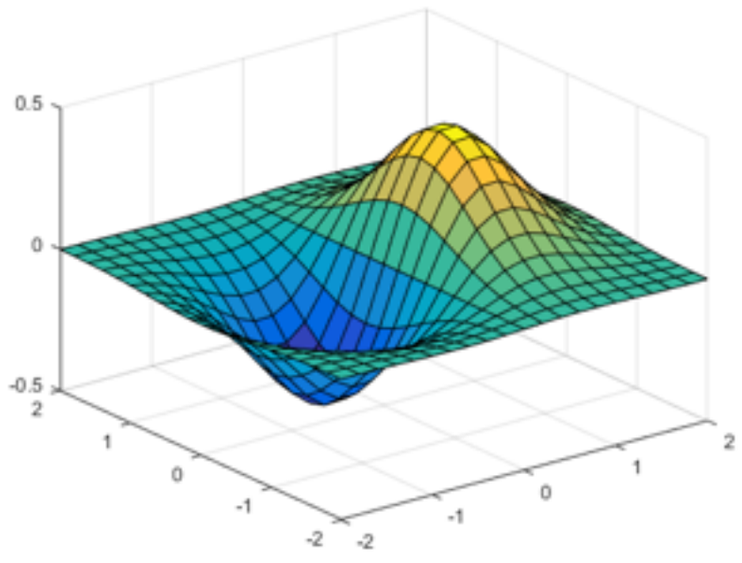
Background

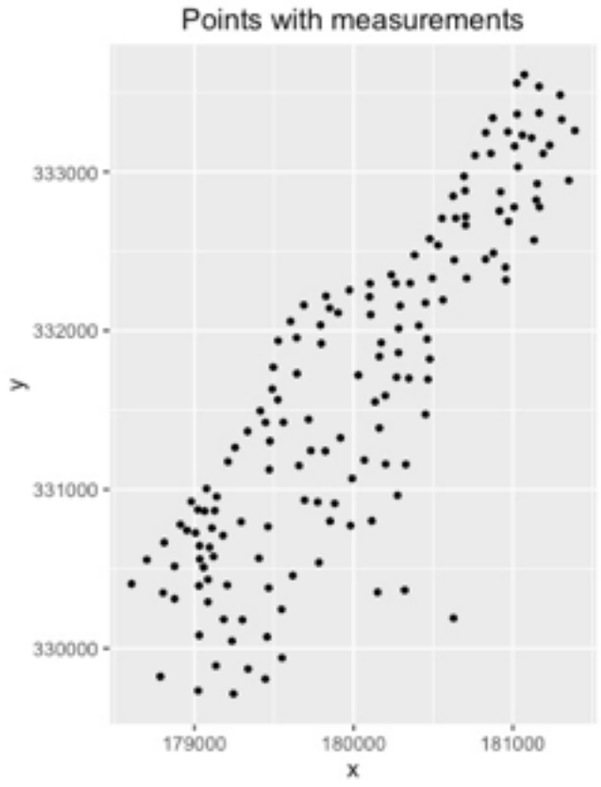
- B.A. in Math, Princeton ’11
- Taught English in China
- Data Analyst in Market Research
- M.S. in Statistics

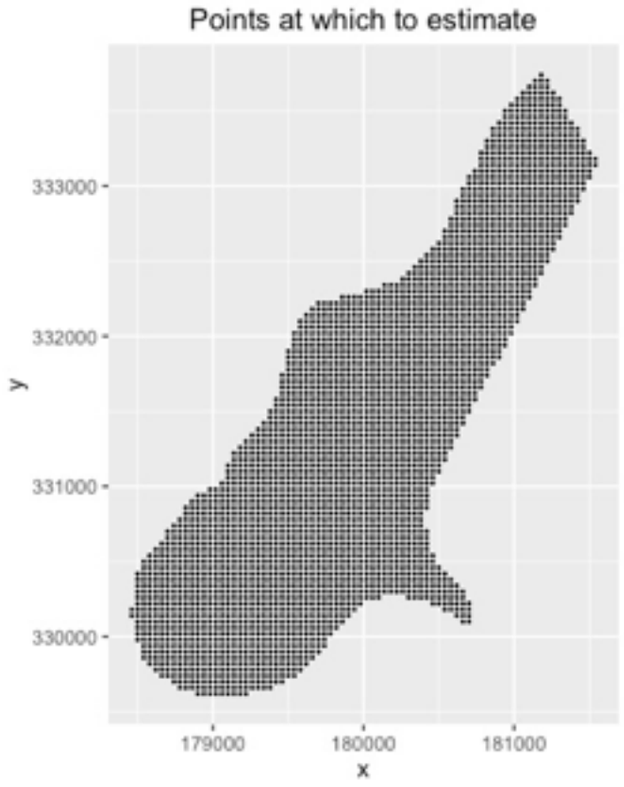
- Ask me about: ergonomics

Overview

- **Goal**: interpolation over space
- **Assume**: some Z = f(X, Y) which varies over region of interest
- **Question**: If you know the lead concentrations in soil at certain locations, how can you estimate concentrations elsewhere?





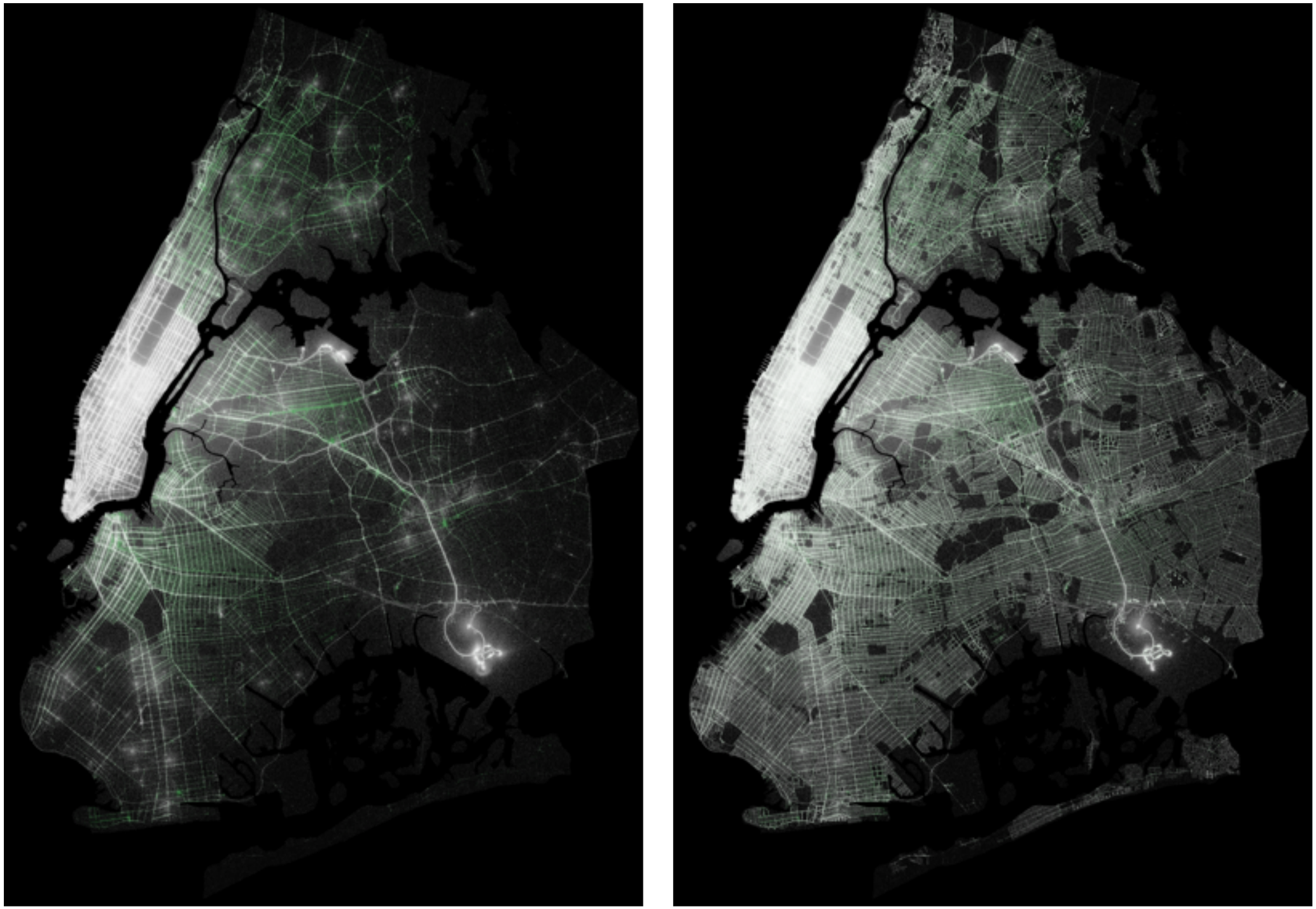


Contents

Motivation

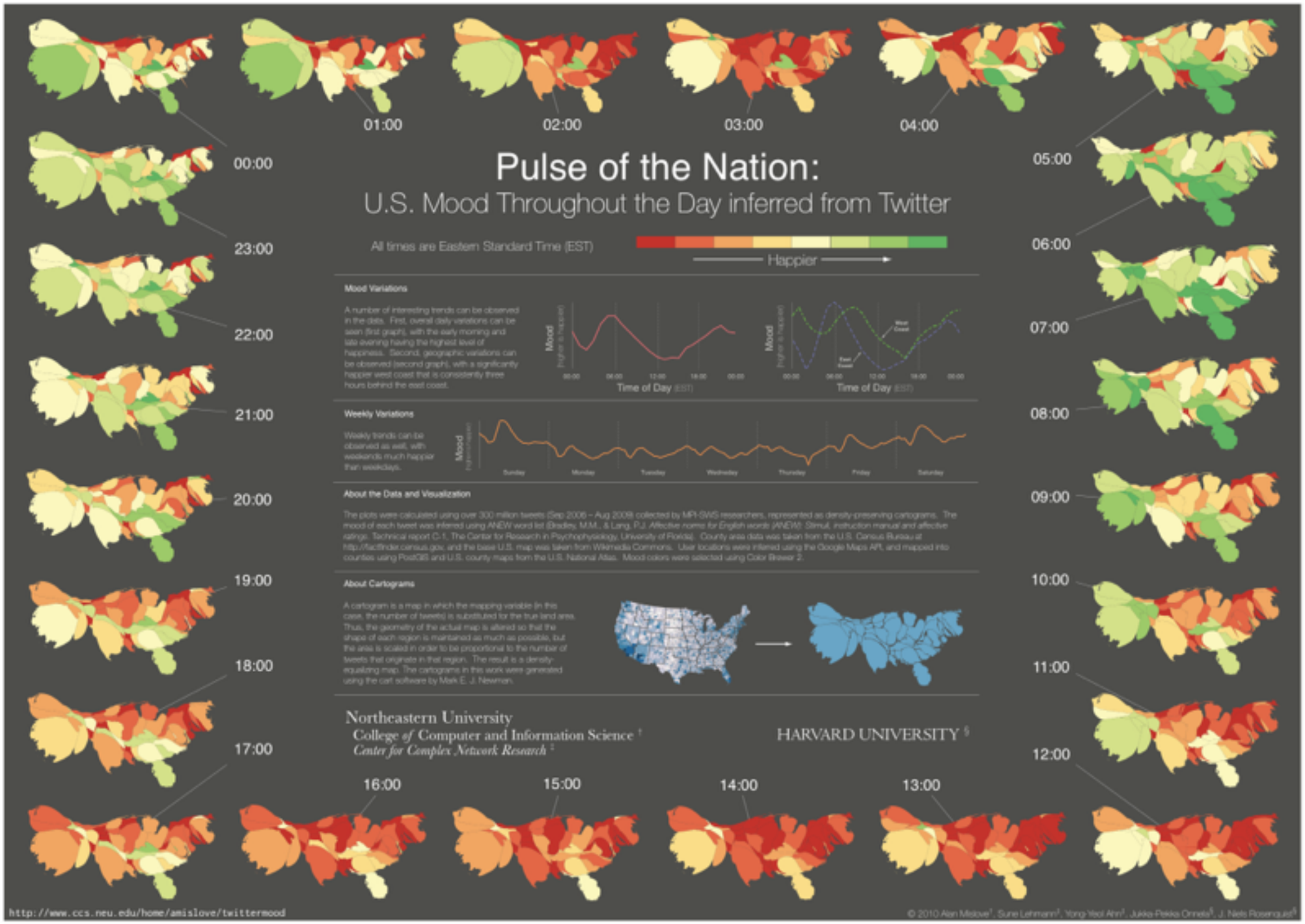
- Much data implicitly spatial (or, spatio-temporal)
- Contain values which vary by space and/or time
- Ex: tweet data, public transport data

NYC Pickups and Dropoffs



Source: http://toddwschneider.com/

Estimating Moods from Tweets



Source: http://www.ccs.neu.edu/home/amislove/twittermood/

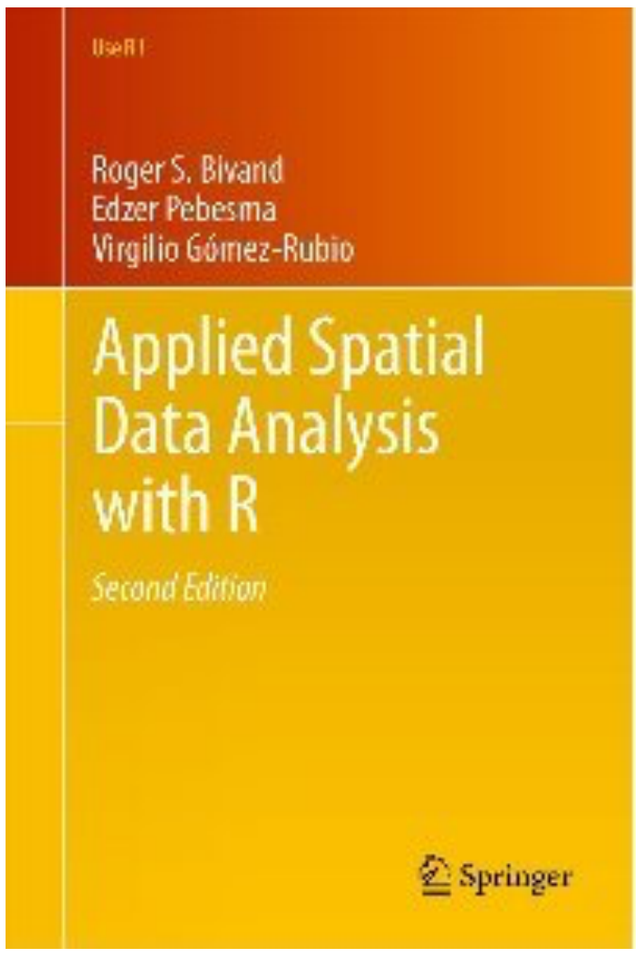
Motivation: Problems

- But, how to identify if data inherently spatial?
  - Ambiguous column names (e.g., i/j, Xm/Ym, row/col)
  - Unknown units (e.g., meters, degrees)
  - Different structures might be used (e.g., volcano dataset)
- Want structures with explicit representation
- **Ideal**: Structure to easily distinguish data vs. coords

R Spatial Toolbox

Reference ("ASDAR"):

- Luckily, R has packages capable of handling spatial (or ST) data
- Here, we focus on two:
    1. sp
    2. gstat



sp

sp: overview

- Early package (ca. 2003) standardizing S/ST data
- Widely adopted, built upon (~ 300 pkgs)
- Easier to first see examples of use, before theory
- See: 1-data.R, 2-intro_spdf.R

Spatial Class Hierarchy

data type	class	attributes	contains
points	`SpatialPoints`	No	`Spatial`
points	`SpatialPointsDataFrame`	`data.frame`	`SpatialPoints`
multipoints	`SpatialMultiPoints`	No	`Spatial`
multipoints	`SpatialMultiPointsDataFrame`	`data.frame`	`SpatialMultiPoints`
pixels	`SpatialPixels`	No	`SpatialPoints`
pixels	`SpatialPixelsDataFrame`	`data.frame`	`SpatialPixels` `SpatialPointsDataFrame`
full grid	`SpatialGrid`	No	`SpatialPixels`
full grid	`SpatialGridDataFrame`	`data.frame`	`SpatialGrid`
line	`Line`	No	
lines	`Lines`	No	`Line` list
lines	`SpatialLines`	No	`Spatial`, `Lines` list
lines	`SpatialLinesDataFrame`	`data.frame`	`SpatialLines`
polygons	`Polygon`	No	`Line`
polygons	`Polygons`	No	`Polygon` list
polygons	`SpatialPolygons`	No	`Spatial`, `Polygons` list
polygons	`SpatialPolygonsDataFrame`	`data.frame`	`SpatialPolygons`

Source: Vignette in `sp` package

Spatial Class Hierarchy

- Can see there are many spatial classes. Here, we only deal with one
- Organized in the abstract, with “top-down” approach
- “Spatial” is most general
  - Doesn’t hold data
  - Object class depends on kind of data used

Building Objects Manually

- Different ways to construct objects
- Can add data to spatial object, or directly make SPDF
- See: 3-spatial_classes.R

sp: Limitations

- Package designed to aid in organizing data
- Sometimes want to perform computations
- How to build off of that?

gstat

gstat: overview

- Functions for modelling spatial and ST data
- Includes different interpolation routines (e.g., IDW)
- In many applications, one approach is common…

Kriging

- Named after South African professor, Daniel Krige
- Sought to identify, estimate mineral deposits of gold
- First need to estimate spatial variability

Variogram

- "Describes degree of spatial dependence"
- Formally defined as,

$$2\gamma(s_1, s_2) = var(Z(s_1) - Z(s_2)) = \mathbb{E}\left[(Z(s_1) - Z(s_2))^2\right]$$

- Expect: points farther away less related than points nearby
- But, if we knew Z, then wouldn't need to estimate

Sample Variogram

- Can calculate a sample variogram:

    1. Consider all pairs of points in spatial domain

    2. Divide groups based on separation/distance

    3. Take average variance per group

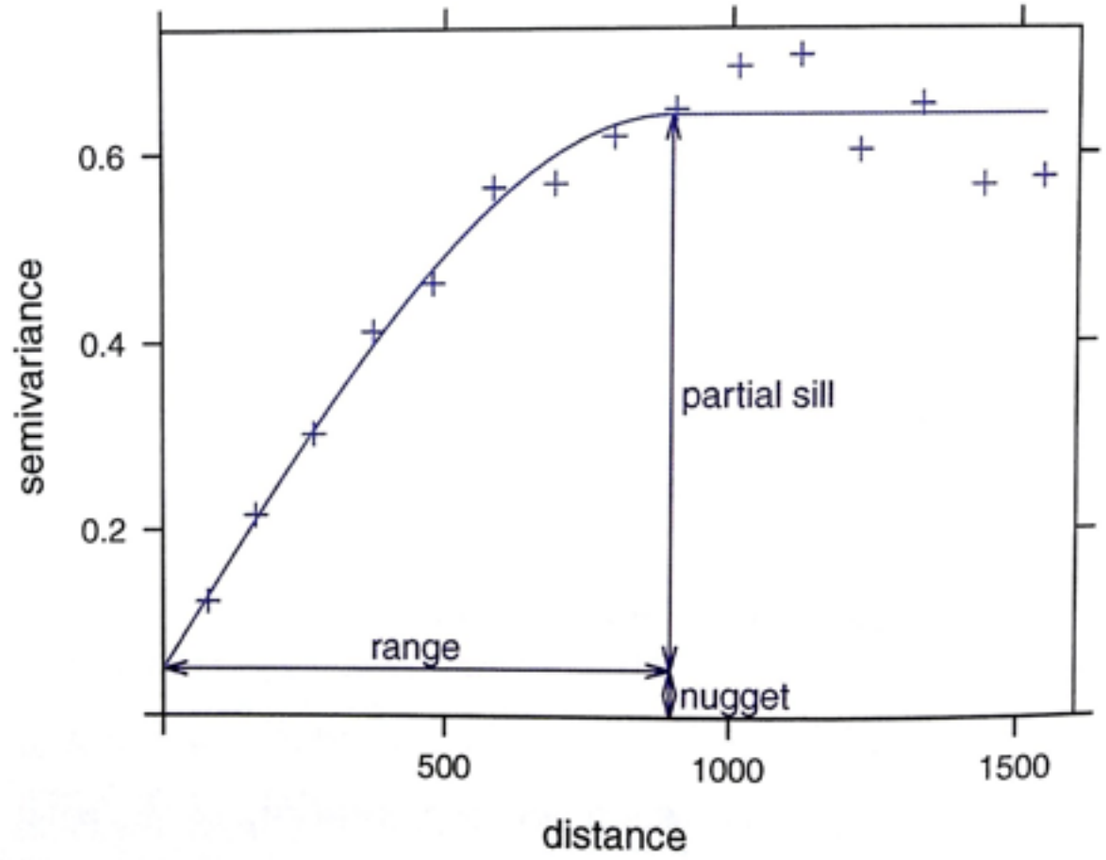
- Formally, defined as:

$$2 * \hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} |s_i - s_j|^2$$

- See: 4-variogram.R

Variogram Modeling

- Different variogram models available
- Typically characterized by three values: nugget, range, sill

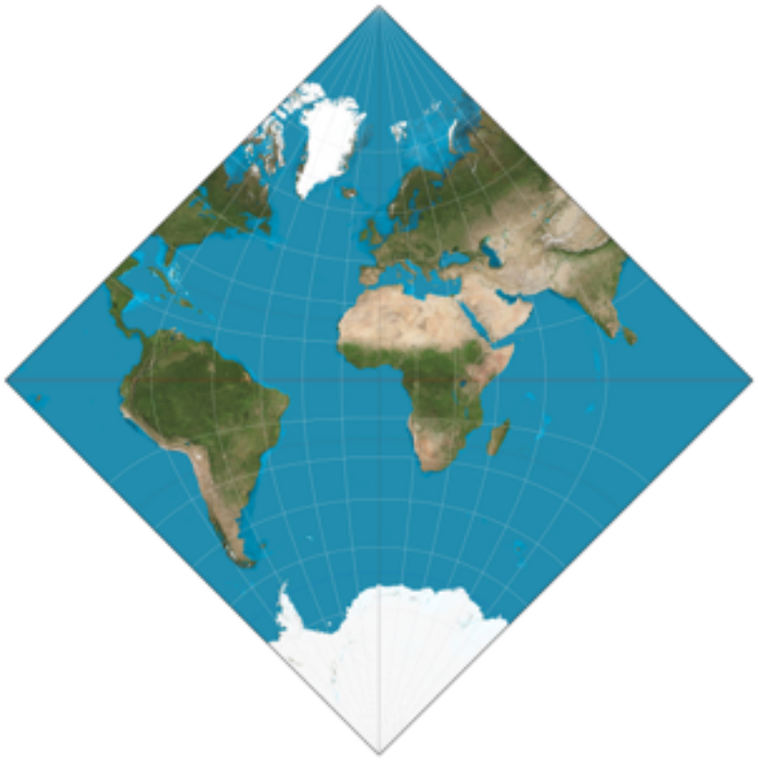


Source: ASDAR, 2ed

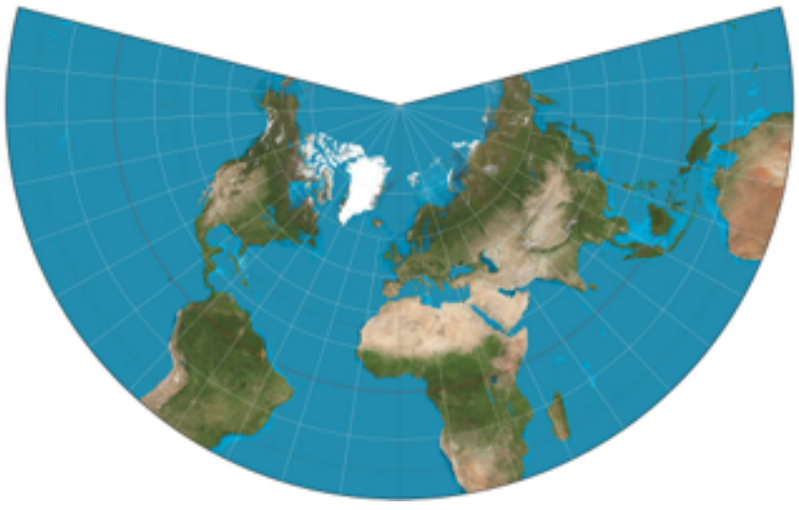
- For this, need accurate measure of distance

Projections

- Affect how distance is perceived and quantified



Hemisphere-in-a-Square



Lambert Conformal

Werner

Projections: Mercator

- Common, not always sufficient
  - Distorts size of regions
    - Distance hard to calculate
- Need to assign projections to data
- See: 5-projections.R

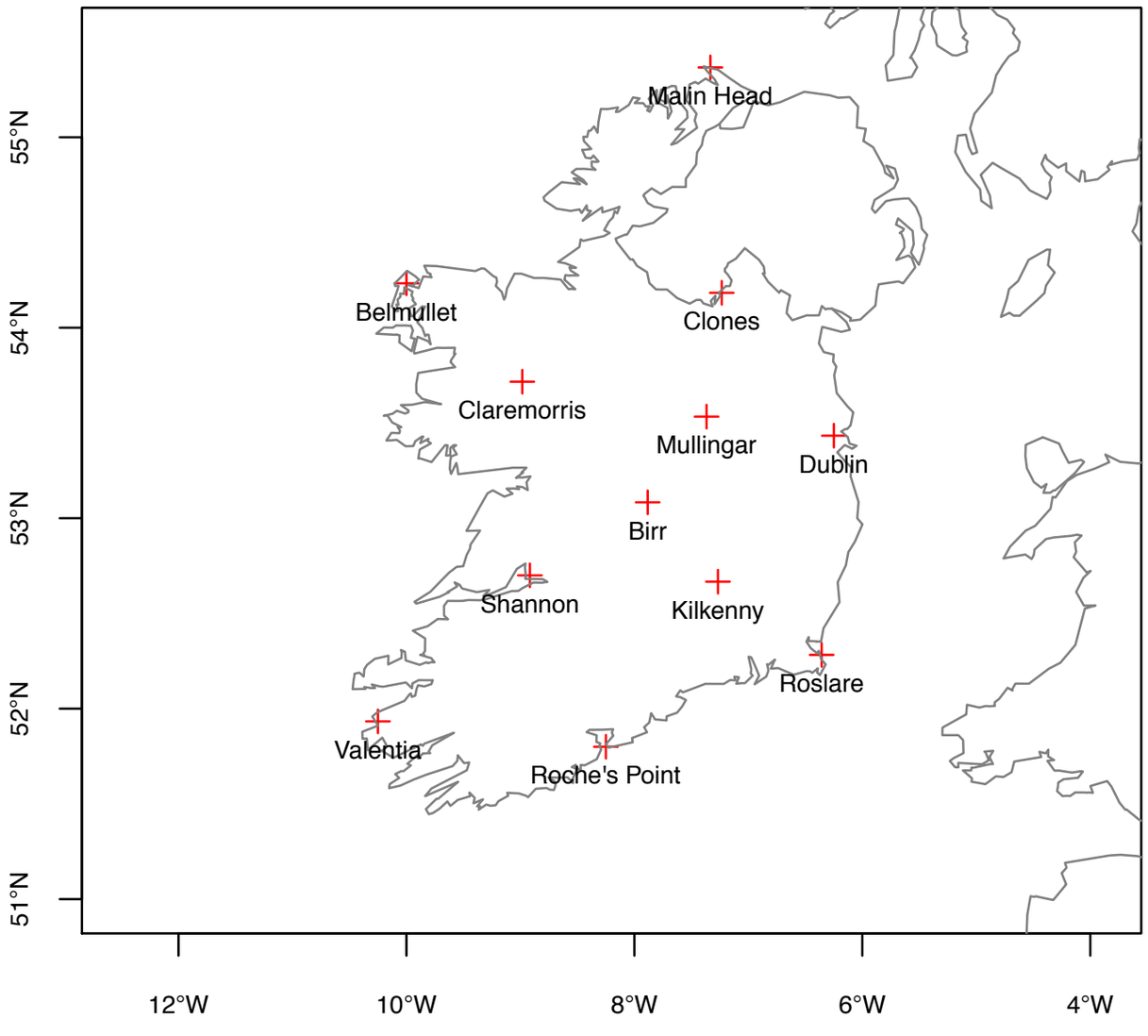
Kriging: Part 2

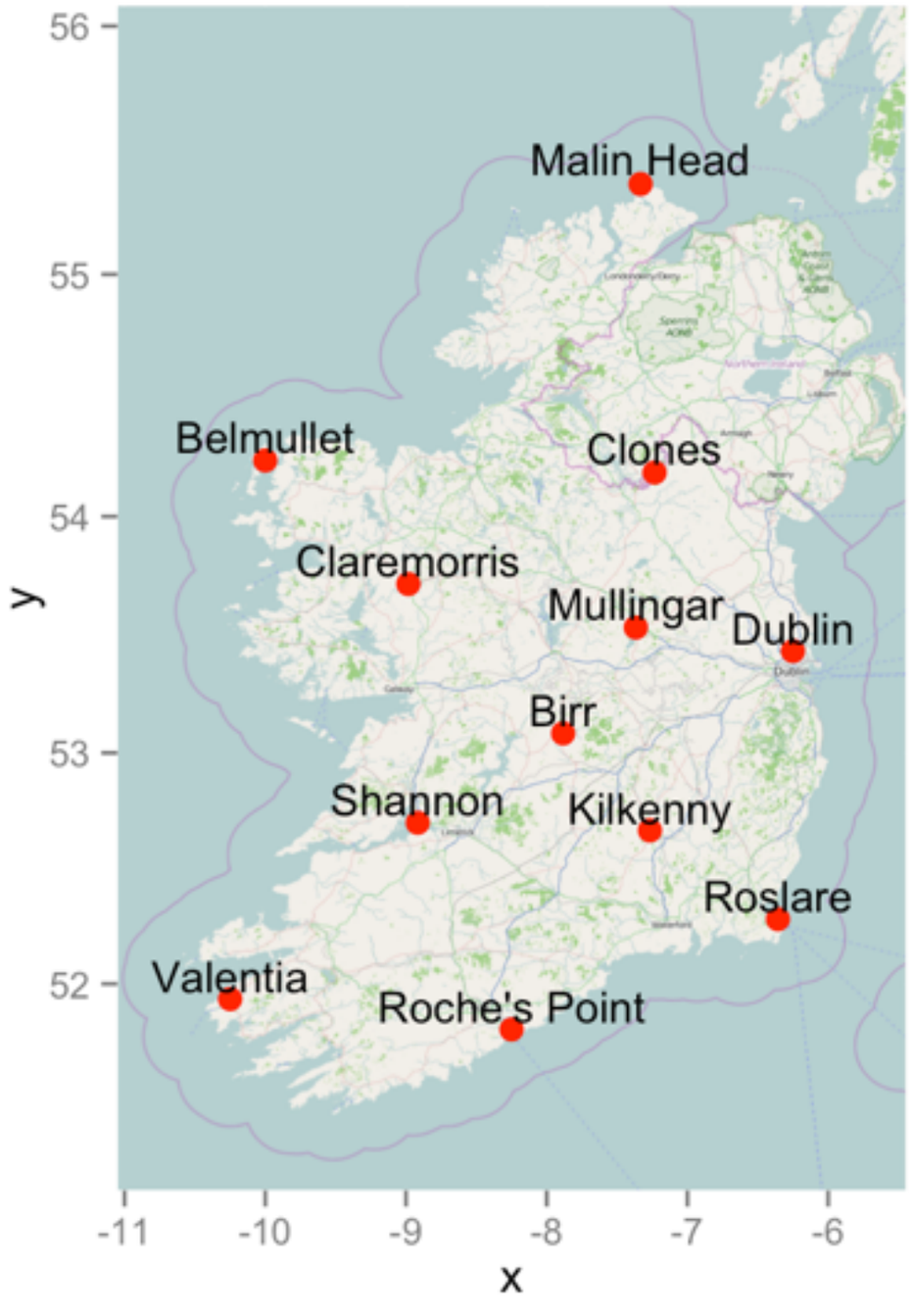
- Now, can measure distance, spatial variability
- Thus, can interpolate with kriging
- See: 6-kriging.R

Visualization

- With results, want to communicate them
- Different packages and tools available
- What are you trying to convey about your data?
- See: 7-visualizations.R

Examples: Base vs. ggmap





Source: `spacetime` vignette

Visualization: Questions

- Questions to ask:
  - Can the data be meaningfully aggregated? (e.g., counts by region) [see rjournal article on ggmap]
  - Interactivity important? (e.g., zoom)

Application: Source Apportionment

- Particulate matter formed from different sources
- Want to quantify various contributions to PM mass
- One method incorporates simulated and observed concentrations of ambient chemical elements
- Once revised estimates produced, can interpolate from point locations across the US.
- For more: https://github.com/nabilabd/hybridSA

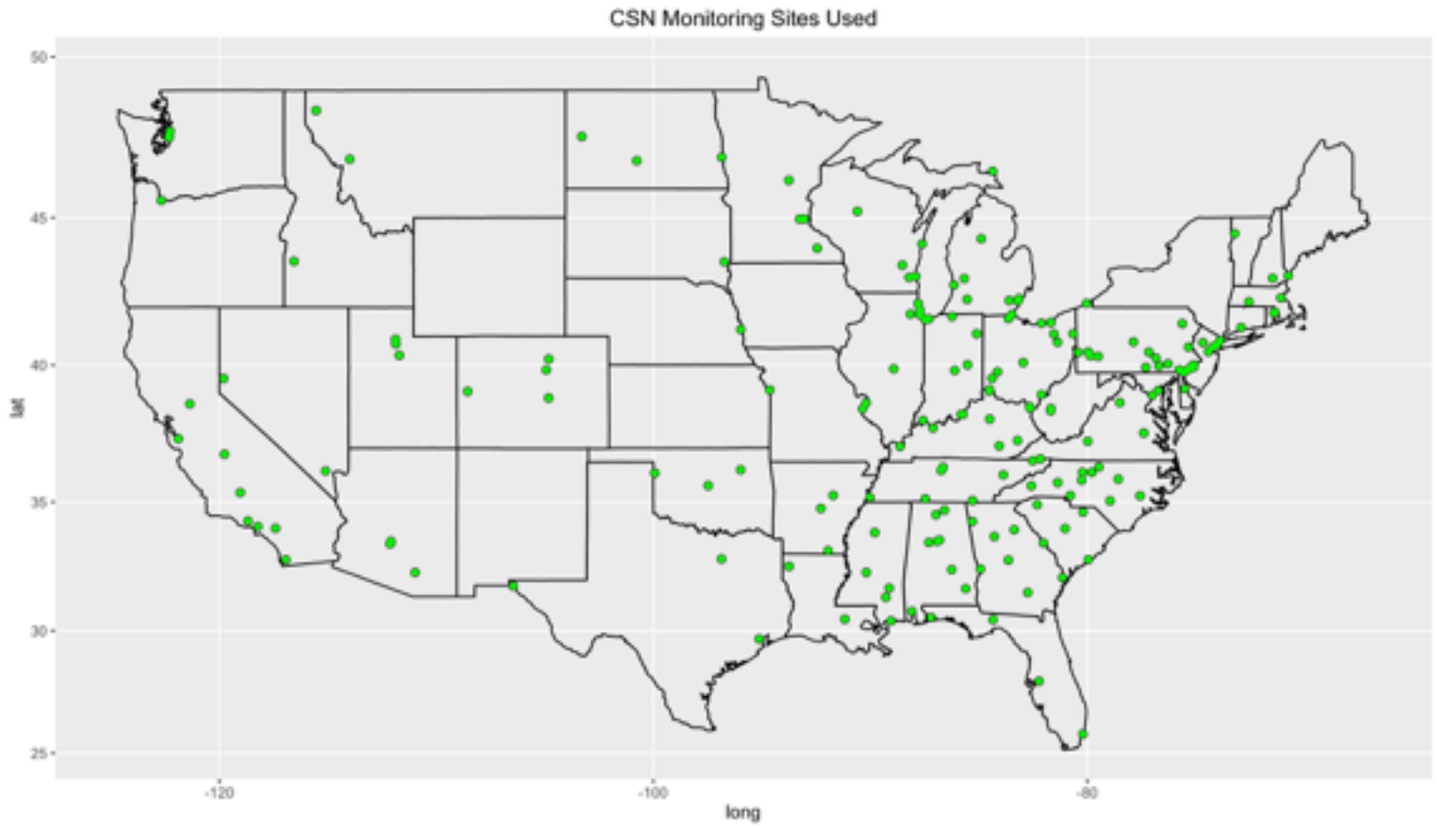
Data Description

- PM2.5 mass, as well as forty chemical species, including: iron, silicon, potassium, sodium.
- For 2007, apportionment was performed into sixteen sources, including: coal combustion, aircraft emissions, on-road gasoline, fire.
- Incorporates uncertainty from both estimates
- Revises simulated concentrations to account for observed phenomena
- Used this equation, see Hu et. al (2014) for more details

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{\left(c_i^{obs} - c_i^{sim} - \sum_{j=1}^{N} SA_{i,j}^{base}(R_j - 1)\right)^2}{\sigma_{obs}^2 + \sigma_{CTM}^2} \right] + \Gamma \sum_{j=1}^{J} \frac{ln(R_j)^2}{\sigma_{ln(R_j)}^2}$$

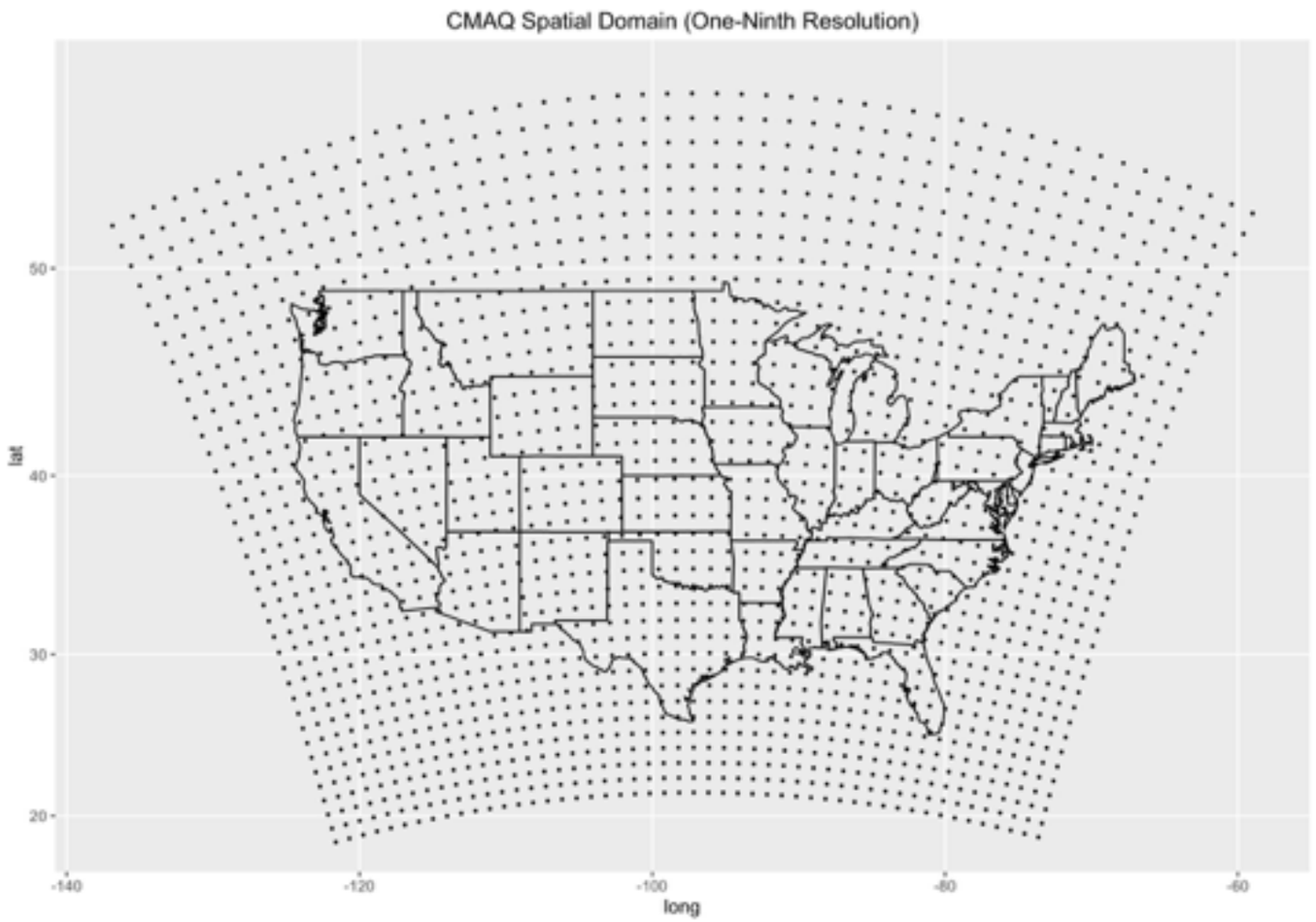
Observational Data

- This is complete set of sites
- Obtained via EPA values for CSN network
- Great irregularities on chemicals measured, and frequency
- This is the limitation for hybrid algorithm



Simulated Data

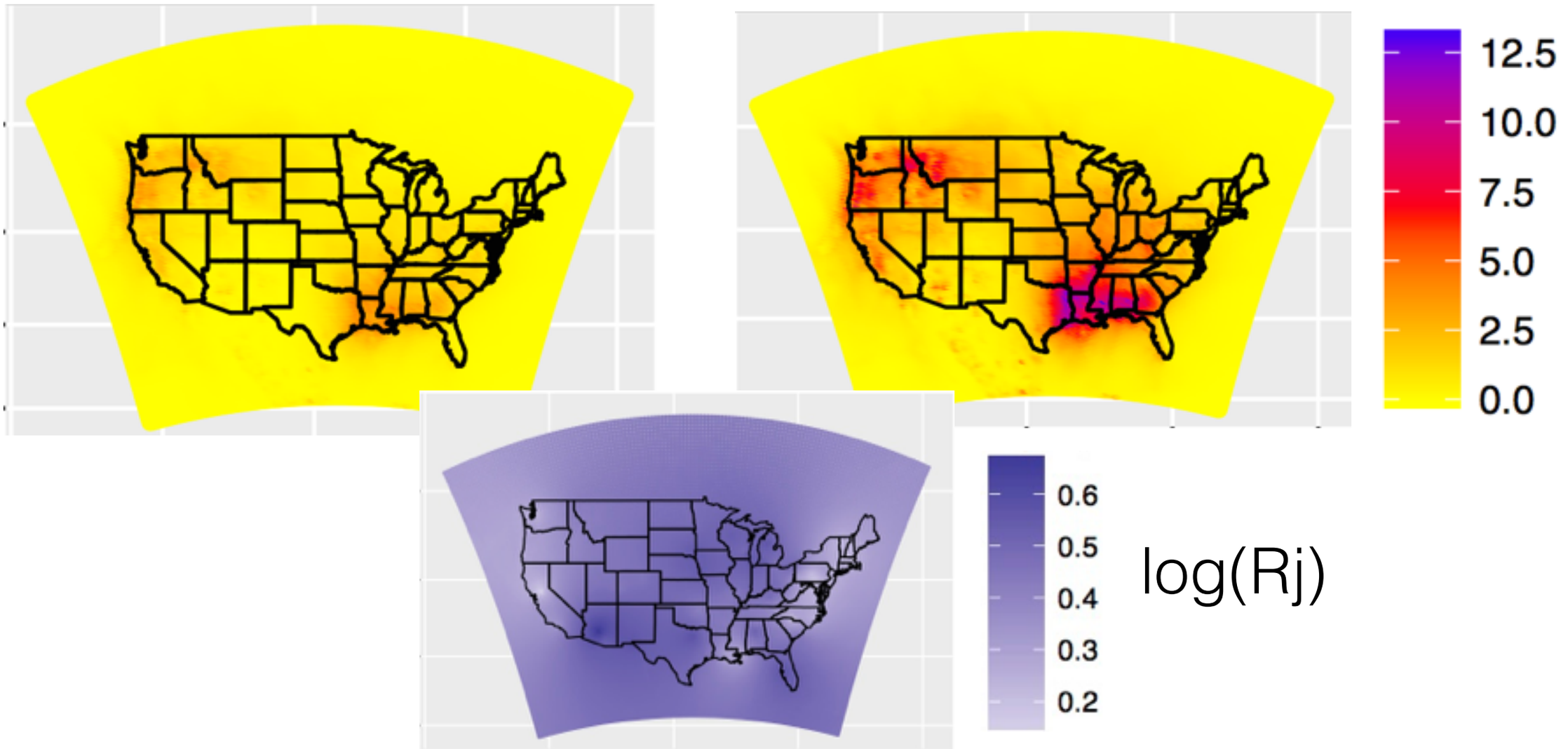
- Obtained from CMAQ model
- 36km x 36km grid cells
- Spatially and temporally, both regular and complete
- Mainly concerned with values in US



Revised Spatial Fields

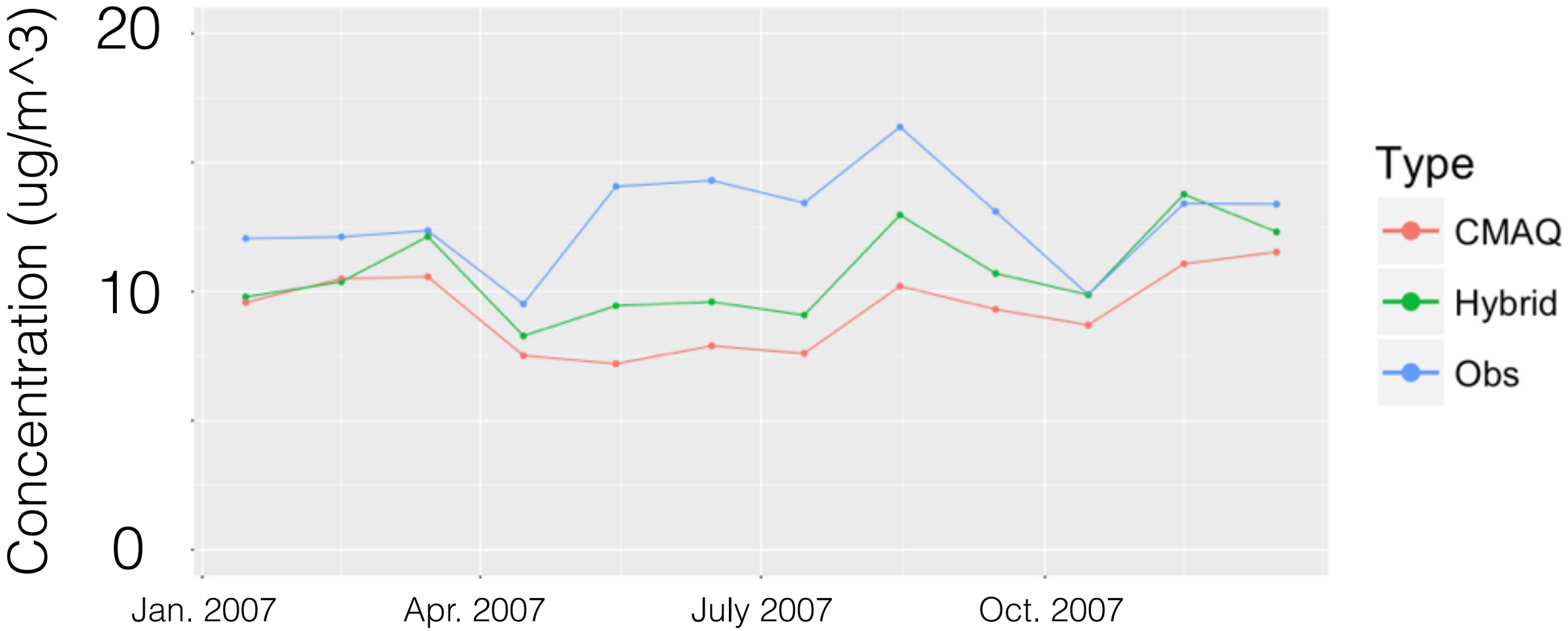
CMAQ Impacts

Hybrid Impacts



- Average source impacts for FIRE in Fall 2007
- Hybrid method estimates exceed CMAQ

Preliminary Results



- Monthly averages across all sites for 2007
- Here, hybrid estimates closer to observed PM than CMAQ
- Secondary species not accounted for

Alternatives

- When kriging can fail, inverse distance weighting can produce similar results

- For Bayesian methods, different packages (e.g., spatstat). Can look into spTimer:
    - Pros: Good documentation (paper in JSS). Efficient code, can use on larger datasets. ST interpolation in single step.

    - Cons: Not deterministic (or perhaps replicable).

- `spacetime` package for classes dealing with spatio-temporal data. Can use in conjunction with gstat (e.g., ST kriging, ST variograms)

    - NB: Recent (i.e., last couple of years), so documentation might be lacking, with steep learning curve for the uninitiated

References

- ASDAR, 2ed
- http://allisonlassiter.com/blog/
- vignettes for: sp, gstat, spacetime
- Hierarchical Modeling and Analysis for Spatial Data, Banerjee

Acknowledgments

- Profs. Ted Russell, Jim Mulholland
- Cesunica Ivey
- Derek Norton

Thanks!

- Questions?