OPEN DATA NATION

FIVAR:

Food Inspection Violation, Anticipating Risk

Jonathan Boyle

Data Incubator Fellowship 3rd round presentation

29 July 2016

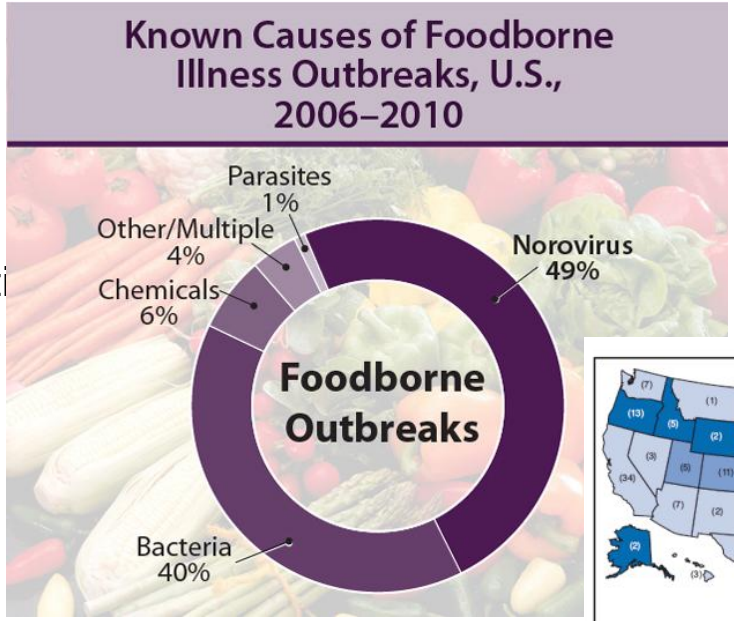
Outline

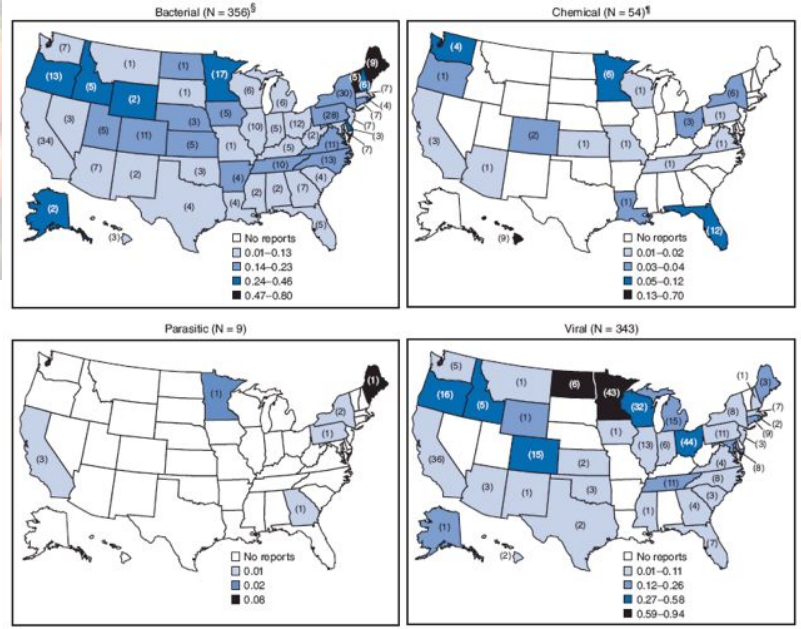
Issue

- Food borne illness
  - Better predicti

Approach

- More EDA
  - Clustering
  - Mapping
  - # restaurants, violations/area
- Model
  - Selection
  - Optimization



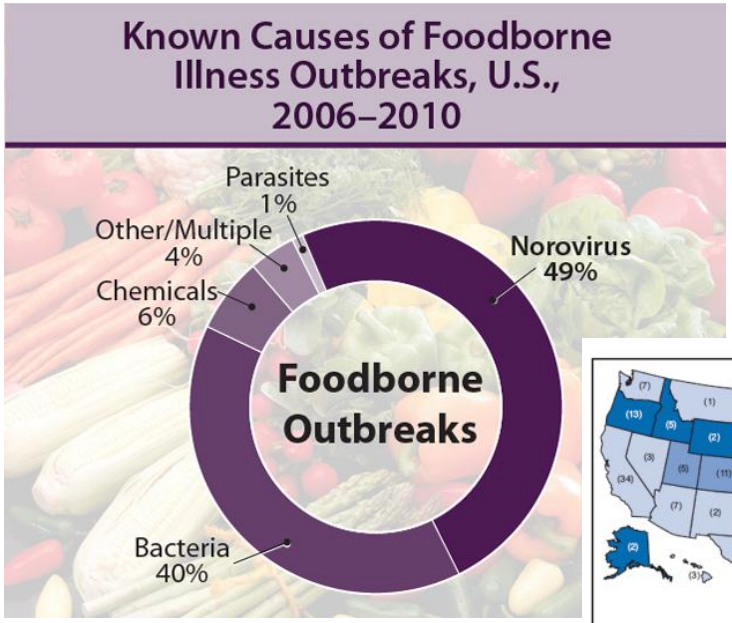


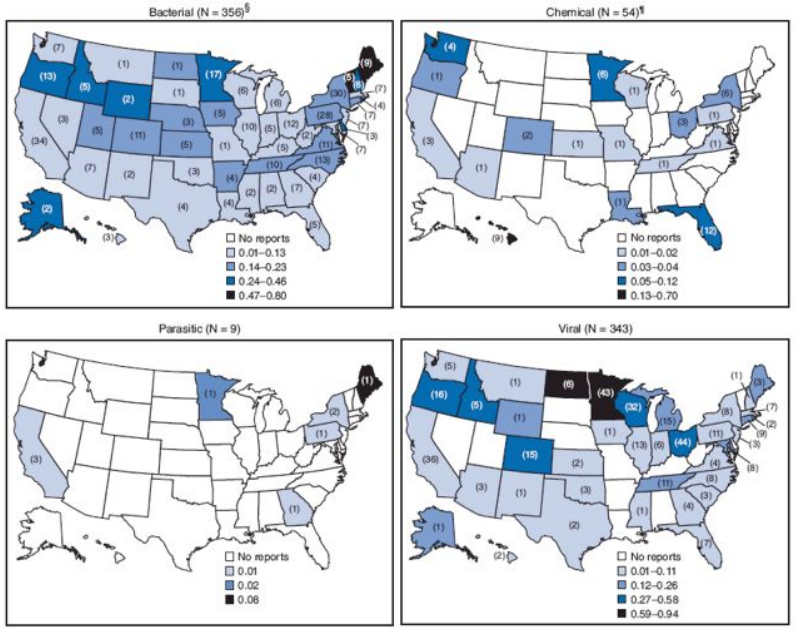
http://www.cdc.gov/features/dsnorovirus/figure3.html
http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5822a1.htm

Outline

Results

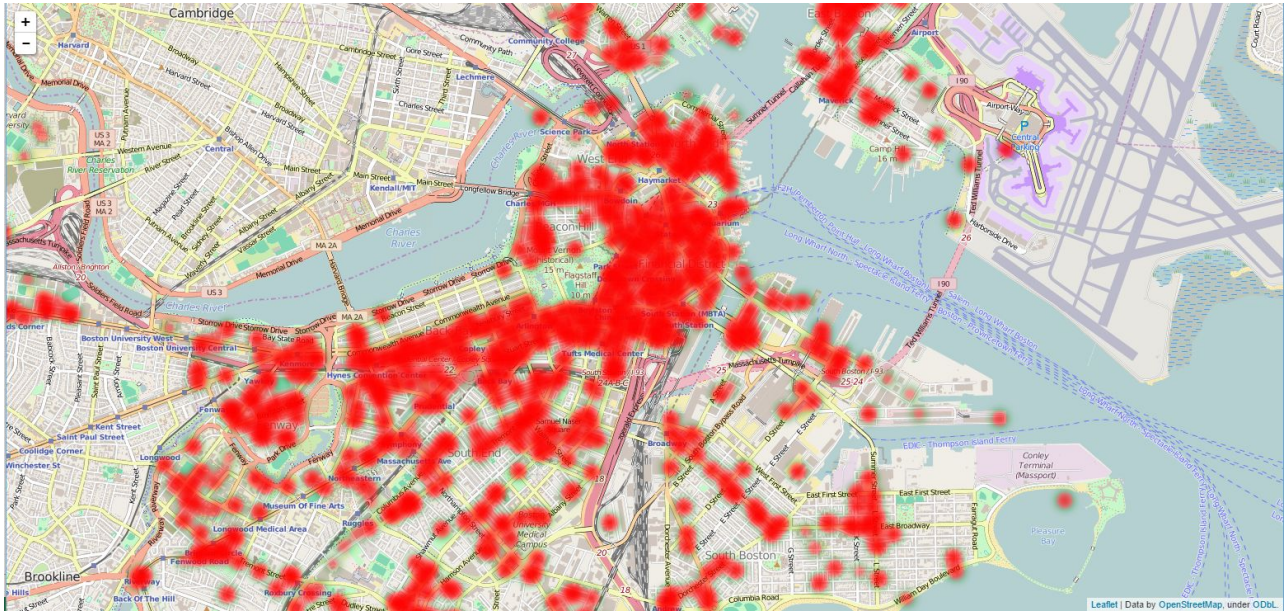
- *Predictions*
  - *Violations days sooner*
  - *Rank order for inspections*





http://www.cdc.gov/features/dsnorovirus/figure3.html
http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5822a1.htm

Issue



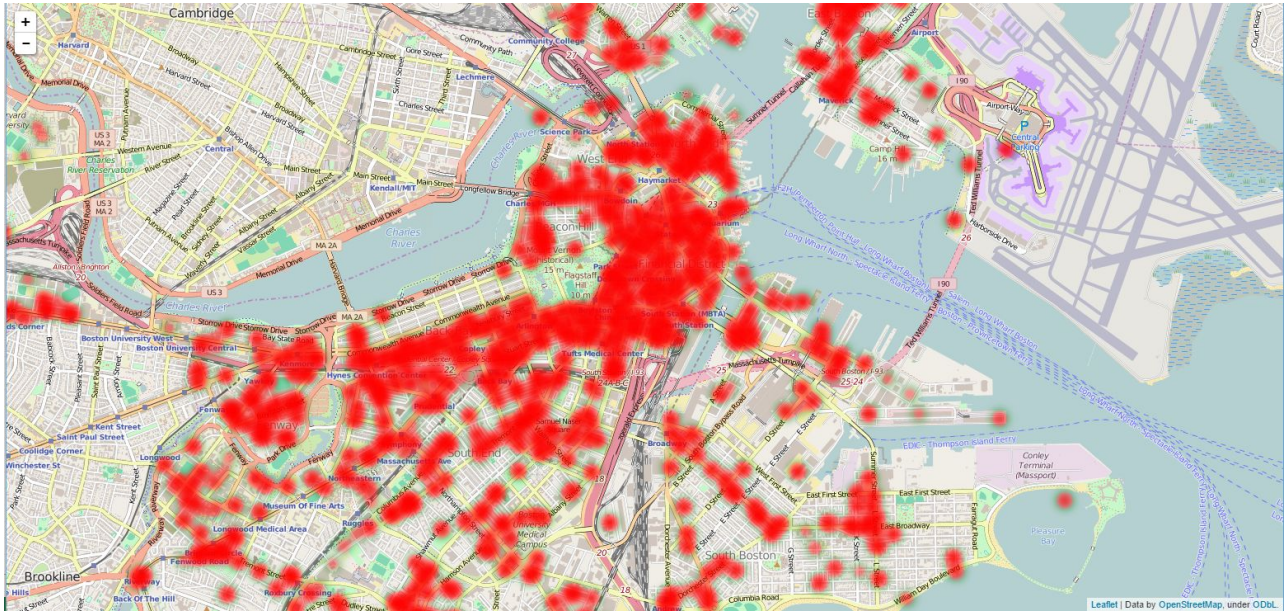
Food borne illness from restaurants

- Not enough health inspectors
- Annual inspections
- Present approach not optimal for public health protection

Different approach

- Machine Learning
  - Predict when violations will occur
    - Reduce illness
- Open Data Nation
  - FIVAR model

EDA



To Date

- Boston restaurant health inspection reports
  - Cleaned/filtered for initial analysis
    - 162 MB file → 21 MB file

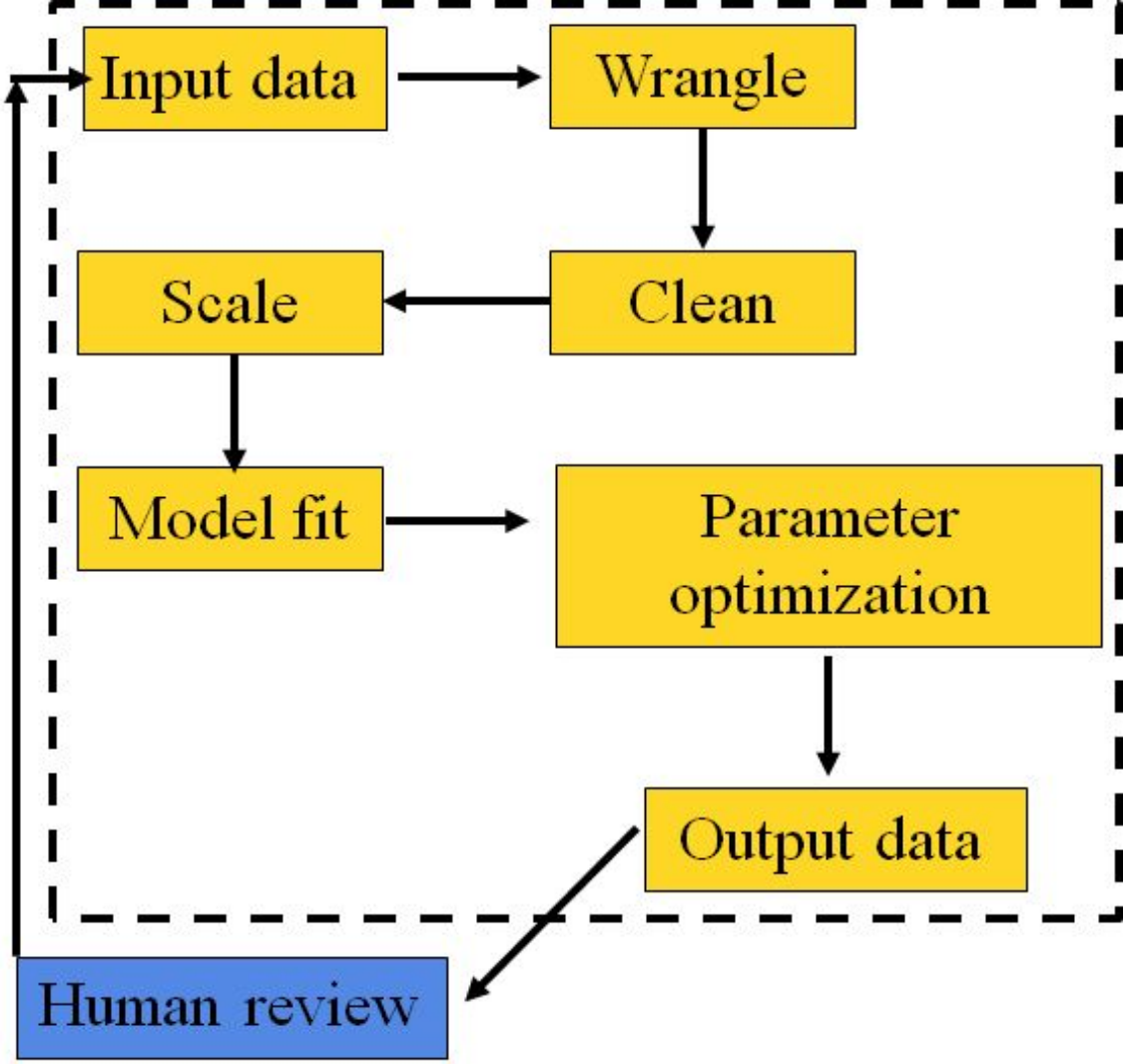
EDA

- Heatmaps
  - Failed inspections for 2007-2016
- Correlation heatmap, boxplots
  - ~56% fail rate, 44% pass rate
- Need to split by year, normalize for restaurant, region

Model

Data

- Multiple sources to compile
  - Boston restaurant health inspection reports
  - 311 complaints
  - Crime reports
  - Property assessments
  - Permits
    - Building
    - Liquor
    - Entertainment
  - Weather data
  - Yelp data



Model

Selection

- Binary classifier
  - RandomForest → Log Reg,KNN
- Regression
  - RandomForest → GLM
- Model(s) chosen based on prior work
  - Chicago
  - Montgomery County, MD
  - Washington, D.C.

Testing

- Test-train split
  - 2007-2015 data
- Out-of-sample
  - 2016 data