OPTIMIZING FOOD SAFETY AT THE CITY OF CHICAGO

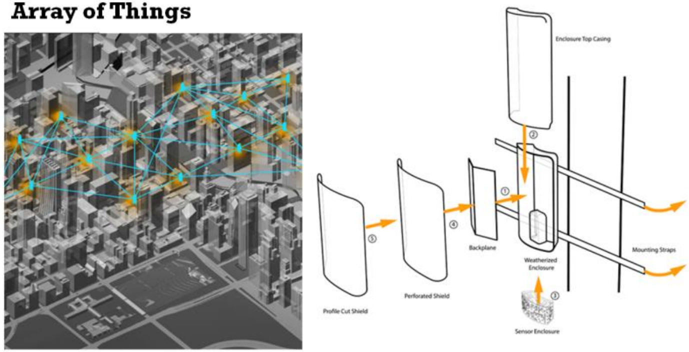
GENE LEYNES

Chicago R User Group Oct 2016

https://github.com/Chicago/food-inspections-evaluation

CITY OF CHICAGO
DATA SCIENCE INIATIVES

Array of Things

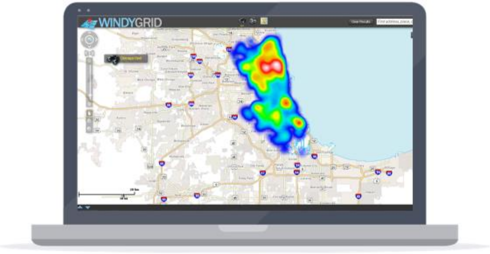


Open Source Sensor Platform

Chicago Open Data Portal

Research Partnerships

Kaggle Competition for
West Nile Virus



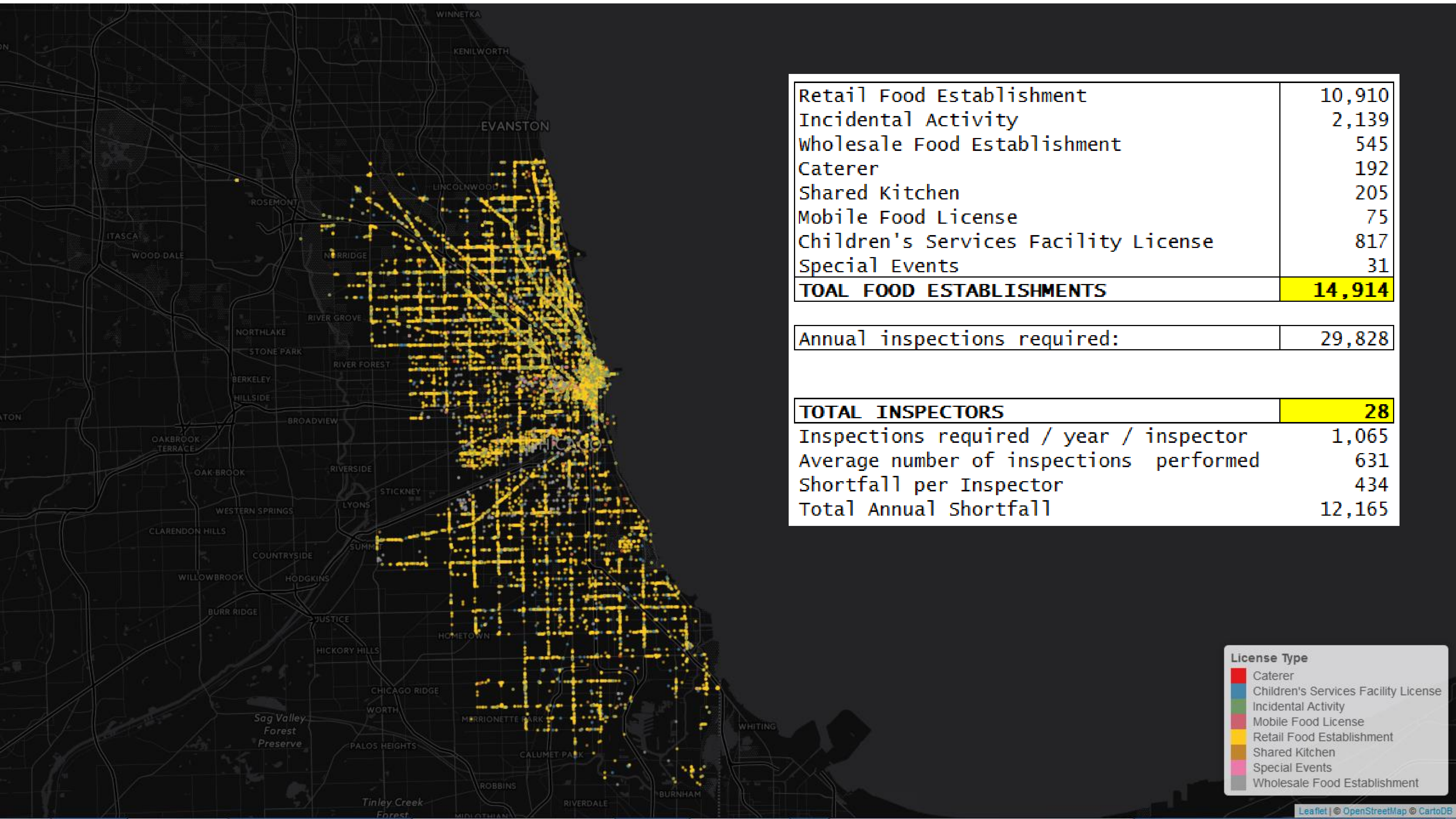
Open Grid BI Tool

FOOD INSPECTIONS PROBLEM STATEMENT

- By law, the City of Chicago is required to inspect food establishments 2x / year
  + Additional inspections for new businesses
  + Additional inspections for consumer complaints

- There are approximately 15,000 businesses
- There were less than 30 food inspectors

- Not every restaurant has the same risk of causing food borne illness



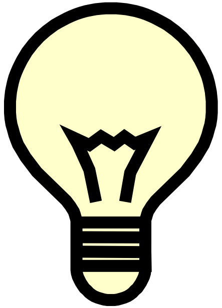
Retail Food Establishment	10,910
Incidental Activity	2,139
Wholesale Food Establishment	545
Caterer	192
Shared Kitchen	205
Mobile Food License	75
Children's Services Facility License	817
Special Events	31
TOAL FOOD ESTABLISHMENTS	**14,914**

Annual inspections required:	29,828

TOTAL INSPECTORS	**28**
Inspections required / year / inspector	1,065
Average number of inspections performed	631
Shortfall per Inspector	434
Total Annual Shortfall	12,165

PROPOSAL

Specifically…



Can we use historical data to predict which inspections are most likely to have a critical violation?

- Develop a binary response model where
- A positive outcome is the presence of any violation numbered 1 to 14 "critical violations"
- Where the observations used **to build the model** are historical food inspections, and
- The observations **to build the prediction** are current food establishment business licenses

DATA SOURCES

Business Licenses

Sanitation Complaints

Food Inspection History

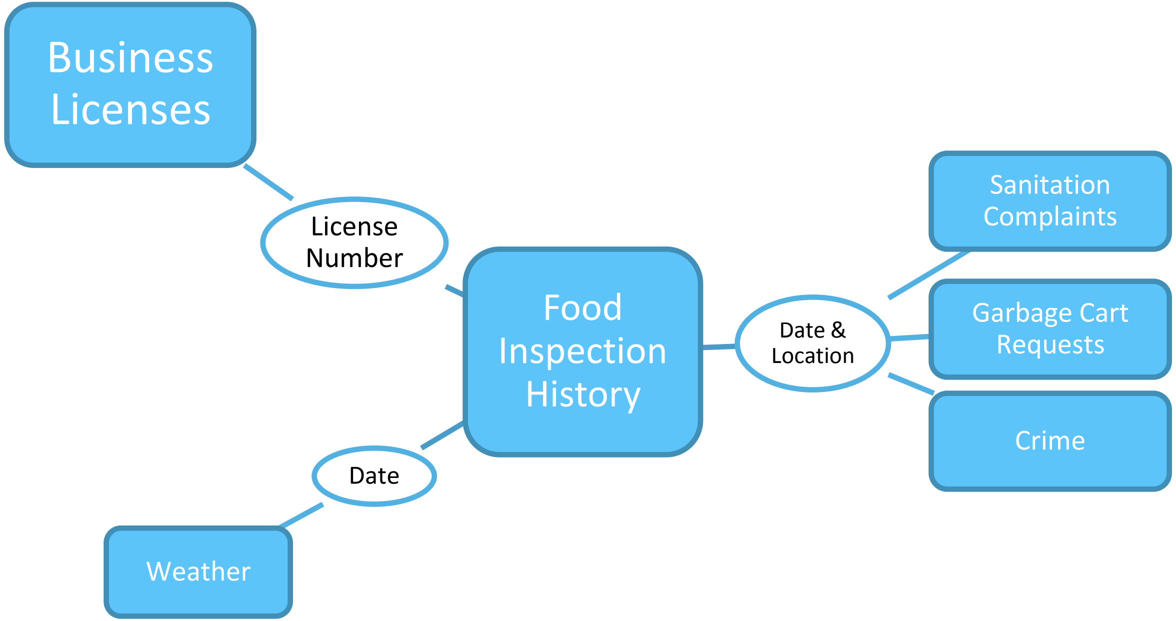
Garbage Cart Requests

Crime

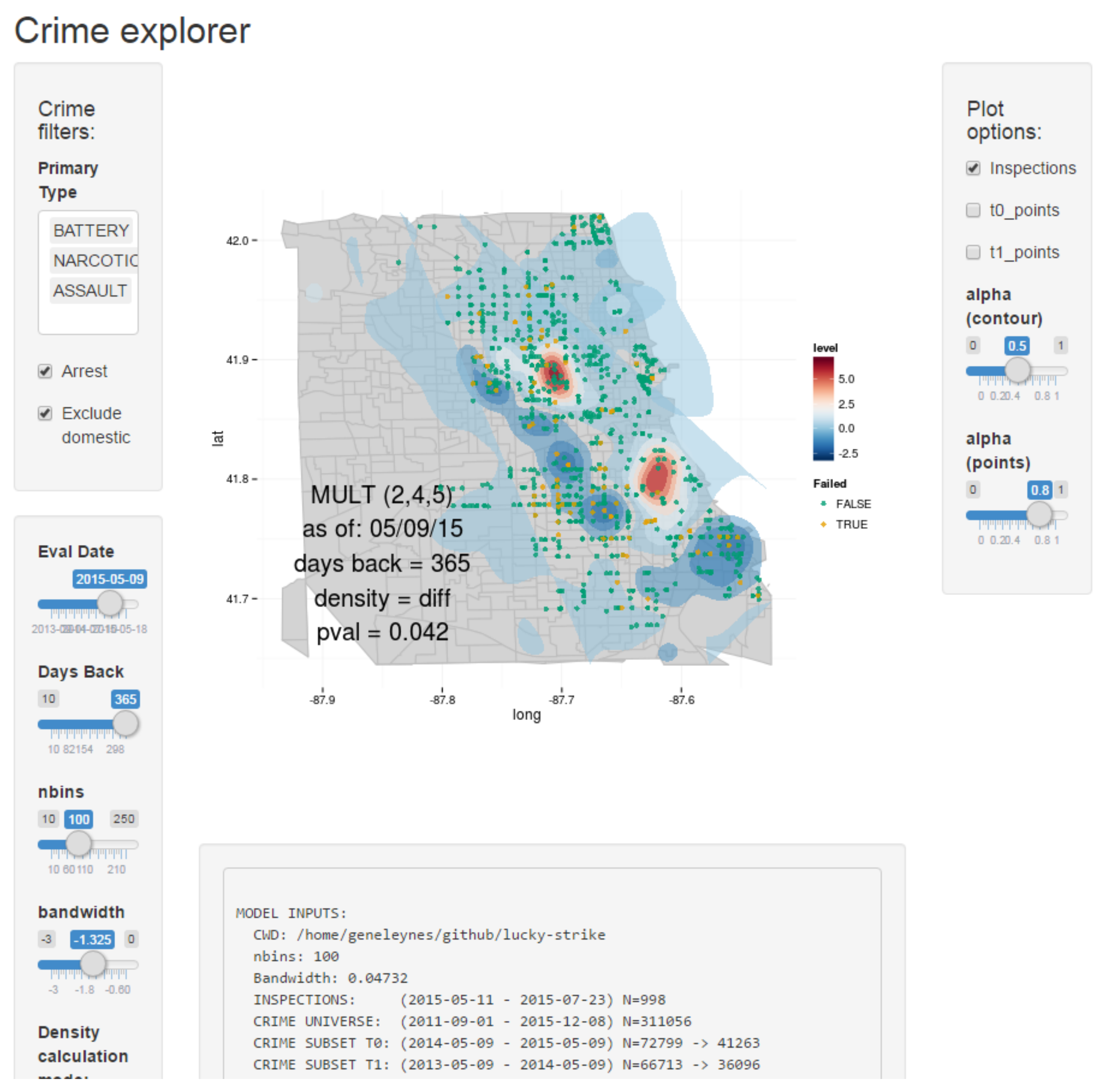
Weather

https://data.cityofchicago.org/

DATA SOURCES



https://data.cityofchicago.org/



311 / 911 Calls are a rich source of high quality data

Linking to other events requires several assumptions

Used Shiny to explore KDE assumptions

The model predicts the likelihood of finding a critical violation, which is the type most likely to cause illnesses.

Ultimately, eleven different variables were used in the final model.

GLM Elastic Net model.

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1+e^{(\beta_0+x_i^T\beta)})\right] + \lambda\big[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\big]$$

Significant Predictors:

- Inspectors
- Restaurants with previous serious and critical violations
- Three-day average high temperature
- Location of restaurant
- Nearby garbage and sanitation complaints
- Nearby burglaries
- Whether the establishment has a tobacco or has an incidental alcohol consumption license.
- Length of time since last inspection.
- Length of time the restaurant has been open.

Technical Keys to Success:

- R / R Studio
- Git / GitHub
- data.table
- knitr
- glmnet

WORKFLOW

GitHub was essential for issue tracking, branch management, and communication.

TOOLS

The data.table package was instrumental for fast processing and feature generation. The foverlaps function was particularly useful for linking records.

COMMUNICATION

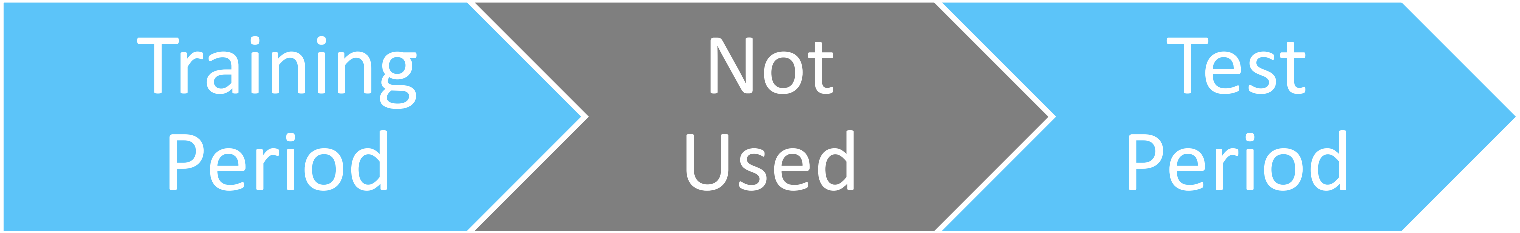
We used knitr to produce intermediate reports and final documentation, also used github.io.

TEST / TRAIN FRAMEWORK

- Initial model was built on 2011 – 2013 data, tested in early 2014
- First experiment failed, mostly because of inspector effects
- Second model was completed later in 2014, tested in 2014, released in 2015

Sept 2011 – Feb 2014

Sept 2014 – Oct 2014

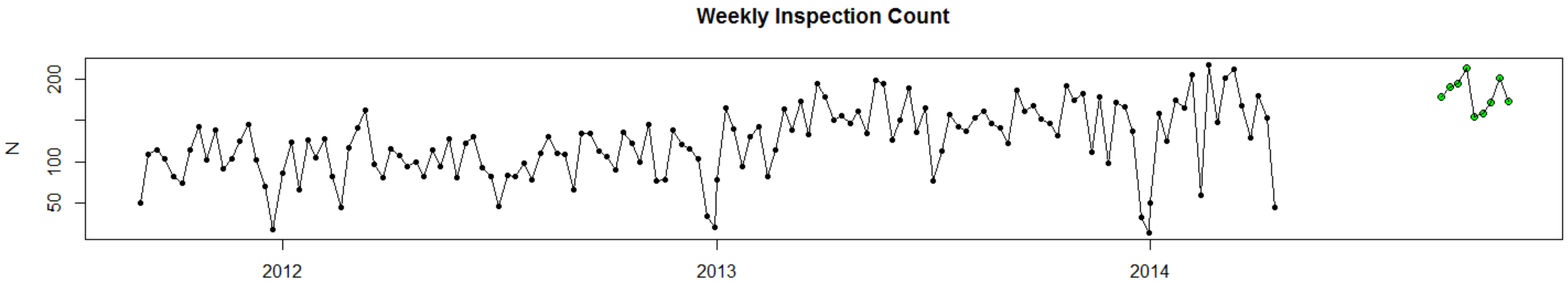


A model is built on *historically* available data

The model is tested on *future* data

TEST / TRAIN FRAMEWORK

- Initial model was built on 2011 – 2013 data, tested in early 2014
- First experiment failed, mostly because of inspector effects
- Second model was completed later in 2014, tested in 2014, released in 2015

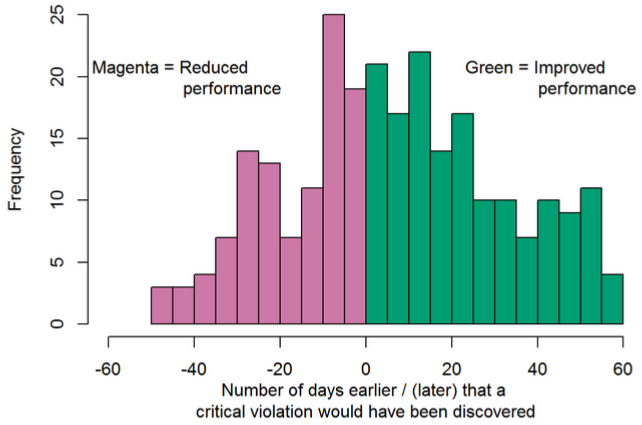


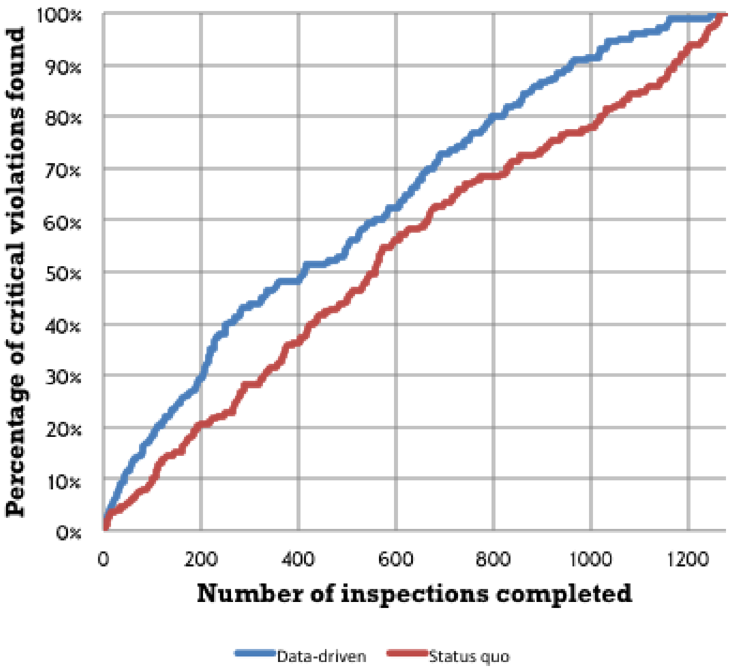
A model is built on *historically* available data

The model is tested on *future* data

Model Evaluation

During the test the data driven approach would have generally found critical violations sooner





~~Our model has an AUC of 0.67226~~

"By using a data driven approach we would have found critical violations 7 days sooner during the test period."

Feature Generation Example

Example from:
23_generate_model_dat.R

- Create a basis for the model data, dat_model

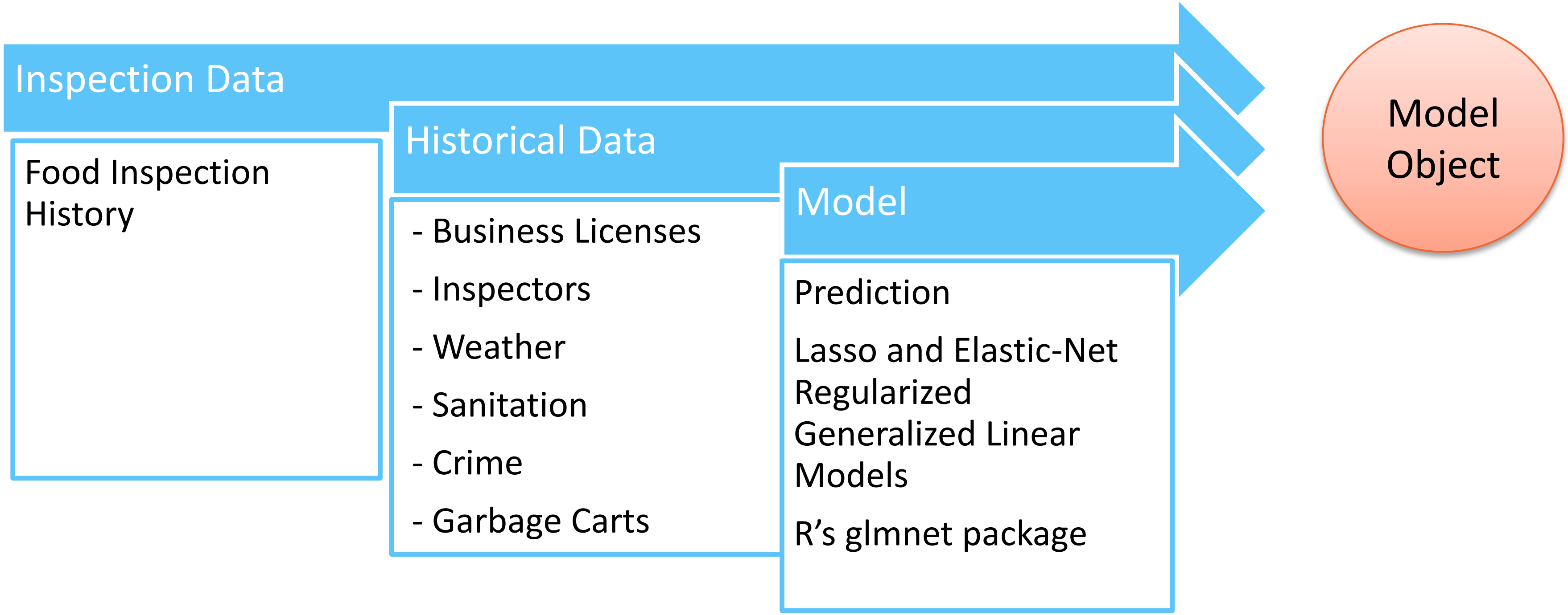
```
##==============================================================================
## Create basis for dat_model, which is the data that will be used in the model
##==============================================================================
dat_model <- foodInspect[i = TRUE ,
                         j = list(Inspection_Date,
                                  License,
                                  Inspection_Type,
                                  Results),
                         keyby = Inspection_ID]
```

- Calculate “minDate”, which is the earliest date seen for a particular License Number

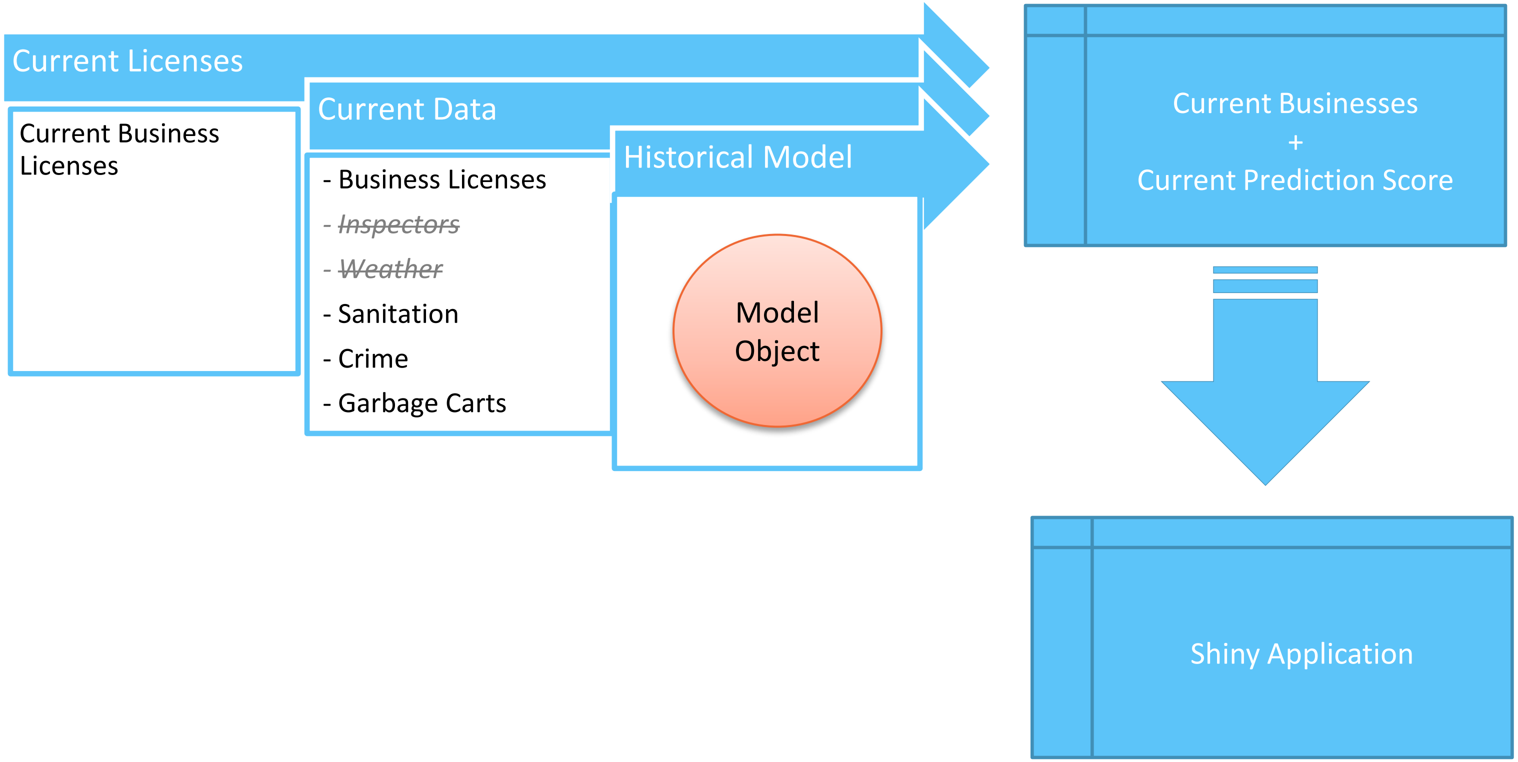
- Use minDate to calculate the age at inspection, which is used in the model

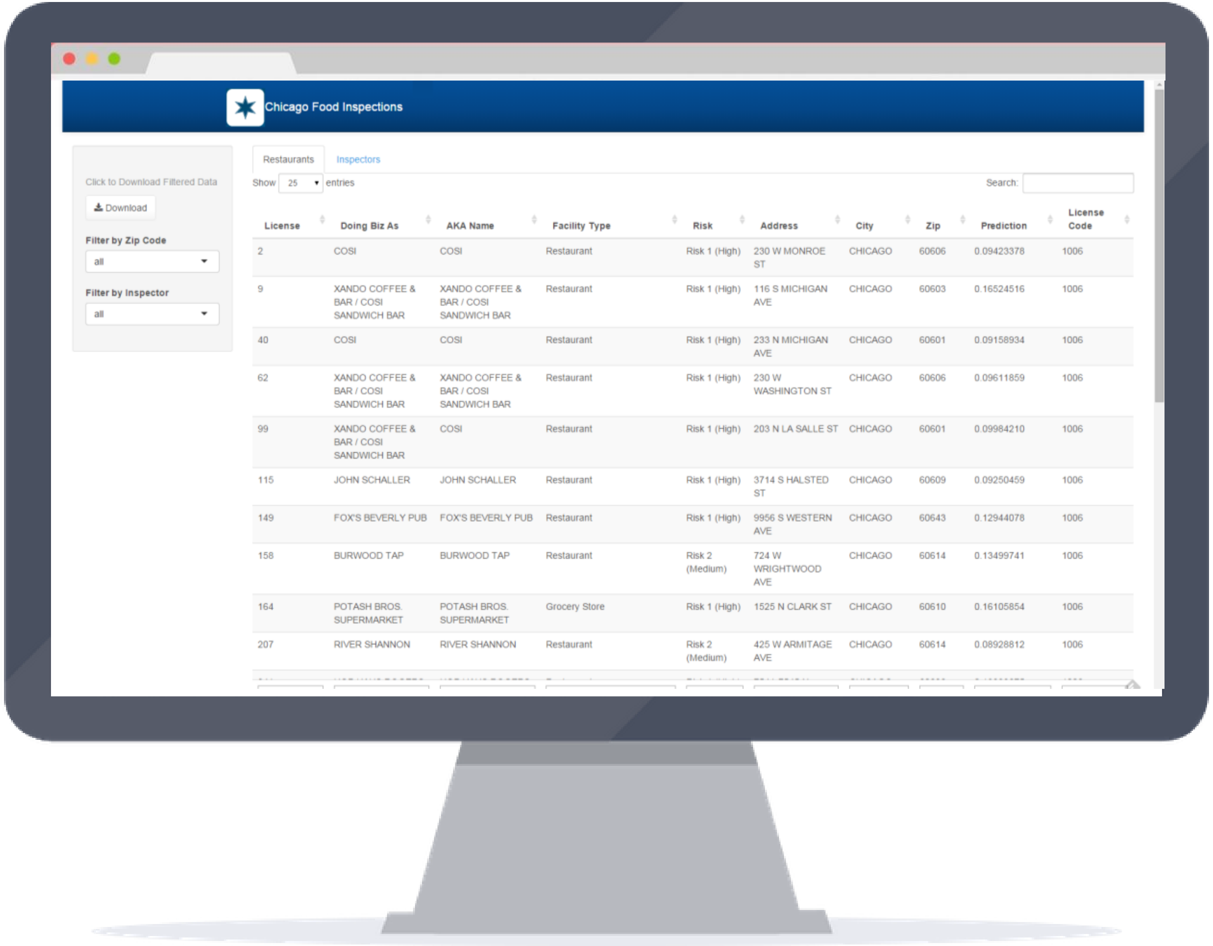
```
## Calculate min date (by license)
business[ , minDate := min(LICENSE_TERM_START_DATE), LICENSE_NUMBER]
business[ , maxDate := max(LICENSE_TERM_EXPIRATION_DATE), LICENSE_NUMBER]

## Calculate age at inspection:
## Add minDate to dat_model
dat_model <- merge(x = dat_model,
                   y = business[ , list(Business_ID = ID,
                                        minDate,
                                        maxDate)], # maxDate's just nice to have
                   by = "Business_ID",
                   all.x = TRUE)
## Use minDate to calculate age
dat_model[ , ageAtInspection := as.numeric(Inspection_Date - minDate) / 365]
```

Model


Inspection Data
Food Inspection History
Historical Data
- Business Licenses
- Inspectors
- Weather
- Sanitation
- Crime
- Garbage Carts
Model
Prediction
Lasso and Elastic-Net Regularized Generalized Linear Models
R’s glmnet package
Model Object

PREDICTION AND APPLICATION





The Final Result:

A simple Shiny application that lists

- Business details
- Zip codes
- Predictions

That's it, no fancy maps!

(Also has performance summaries, not shown)

THANK YOU

Contact & Info:

Gene Leynes
gene.leynes@cityofchicago.org
@geneorama

https://chicago.github.io/food-inspections-evaluation/
https://github.com/Chicago/food-inspections-evaluation
https://data.cityofchicago.org/

PBS Newshour
The Economist

Thank you:

Bloomberg Philanthropies

Allstate Insurance

Civic Consulting Alliance

The Chicago Department of Public Health