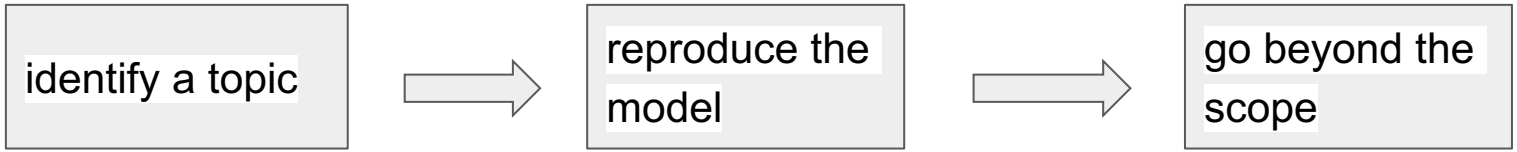
Data 2050 Pitch

Zhirui Li

Introduction

- Replicate and extend a paper that involves **public health** using **publicly available data**



How to Start?

Where to search?

- Plos ONE
- JAMA Network
- Journal of Applied Statistics
- American Journal of Epi

Where to find datasets?

- Kaggle
- Google dataset search
- UCI Machine Learning Repository
- National Center for Biotechnology Information

Topics of interests

- COVID-19
- Diabetes
- Abortion
- Gun Legislation
- Nursing Homes
- Foodborne Diseases
- Vaping
- Opioids

First Possible Paper

Machine Learning Prediction of Foodborne Disease Pathogens: Algorithm Development and Validation Study

- Predicted foodborne disease pathogens (total four types) and investigated what predictors play significant roles
- Models trained: a single decision tree, random forest, gradient boosted decision tree, and adaptive boosted decision tree
- Features selected: patients age, patients gender, time of illness, home address recorded as three variables (province, city, and district), symptom, diagnosis, food name, and food type name
- The gradient boosted decision tree method has the highest accuracy, with a macro-averaged F1 score nearly 69%
- Predictors such as latitude, longitude, sick time, and patient age have the most effect in predictions

Second Possible Paper

Mortality predictors of hospitalized patients with COVID-19: Retrospective cohort study from Nur-Sultan, Kazakhstan

- Identified predictors of mortality associated with COVID-19
- All cohorts were divided into two groups survivors and deceased patients
- Categorical variables were compared using the chi-square test
- Continuous variables were compared using t-test
- Age, respiratory rate, and CRP were found to be useful predictors of mortality

Thank You for Watching