Predicting Food Desert via Social Media

Chuan Hong

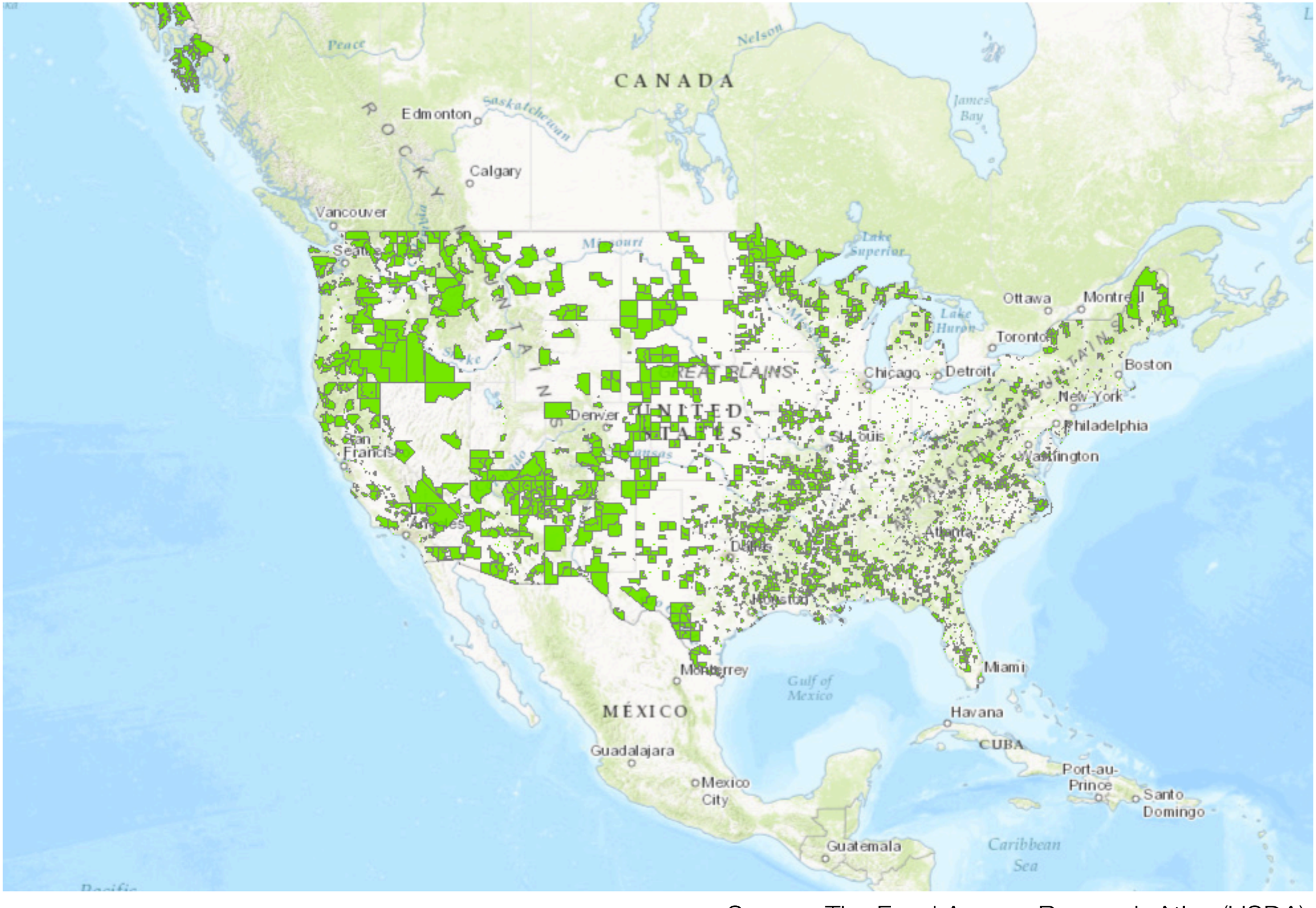
Food Desert

- Food deserts are areas that lack access to affordable fruits, vegetables, whole grains, low fat milk, and other foods that make up the full range of a healthy diet (CDC, 2016)

- Health problem: food deserts are heavy on local quickie marts that provide a wealth of processed, sugar, and fat laden foods that are known contributors to our nation's obesity epidemic

- At least **500 people** and/or at least **33 percent** of the census tract's population reside more than **one mile** from a supermarket or large grocery store (for rural census tracts, the distance is more than **10 miles**) (ANA, 2011)

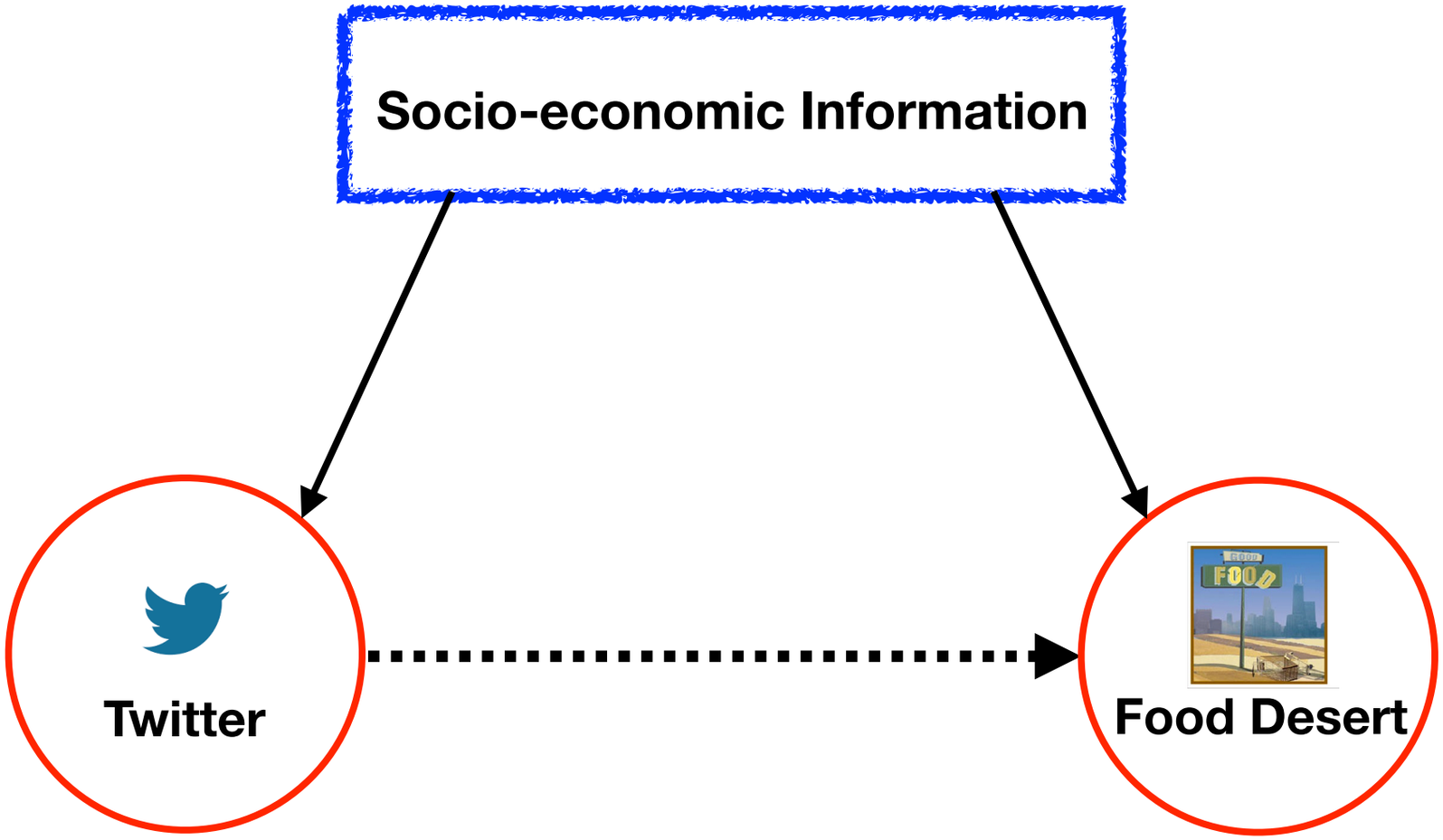
Food deserts mapped from coast to coast (census tract)



Source: The Food Access Research Atlas (USDA)

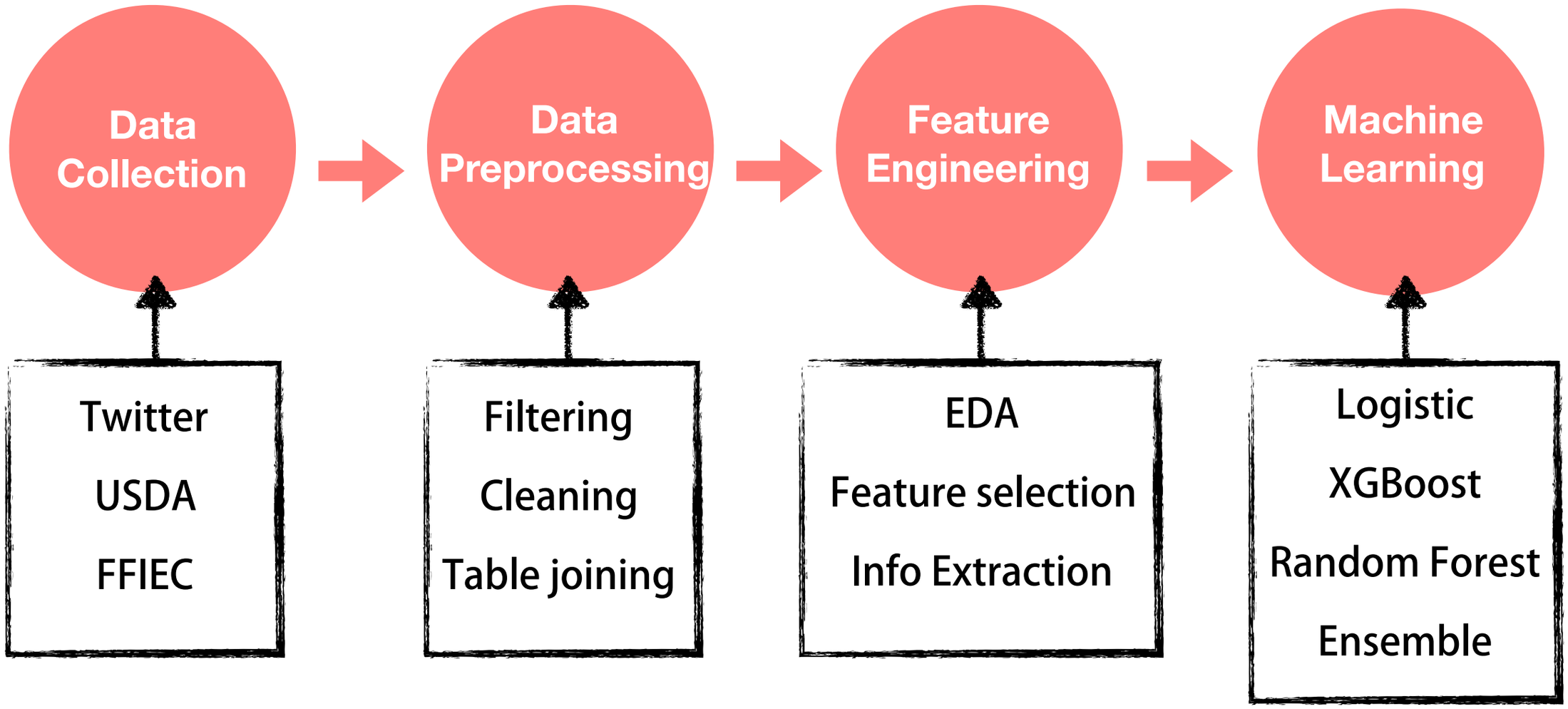
You Post What You Eat

Motivation



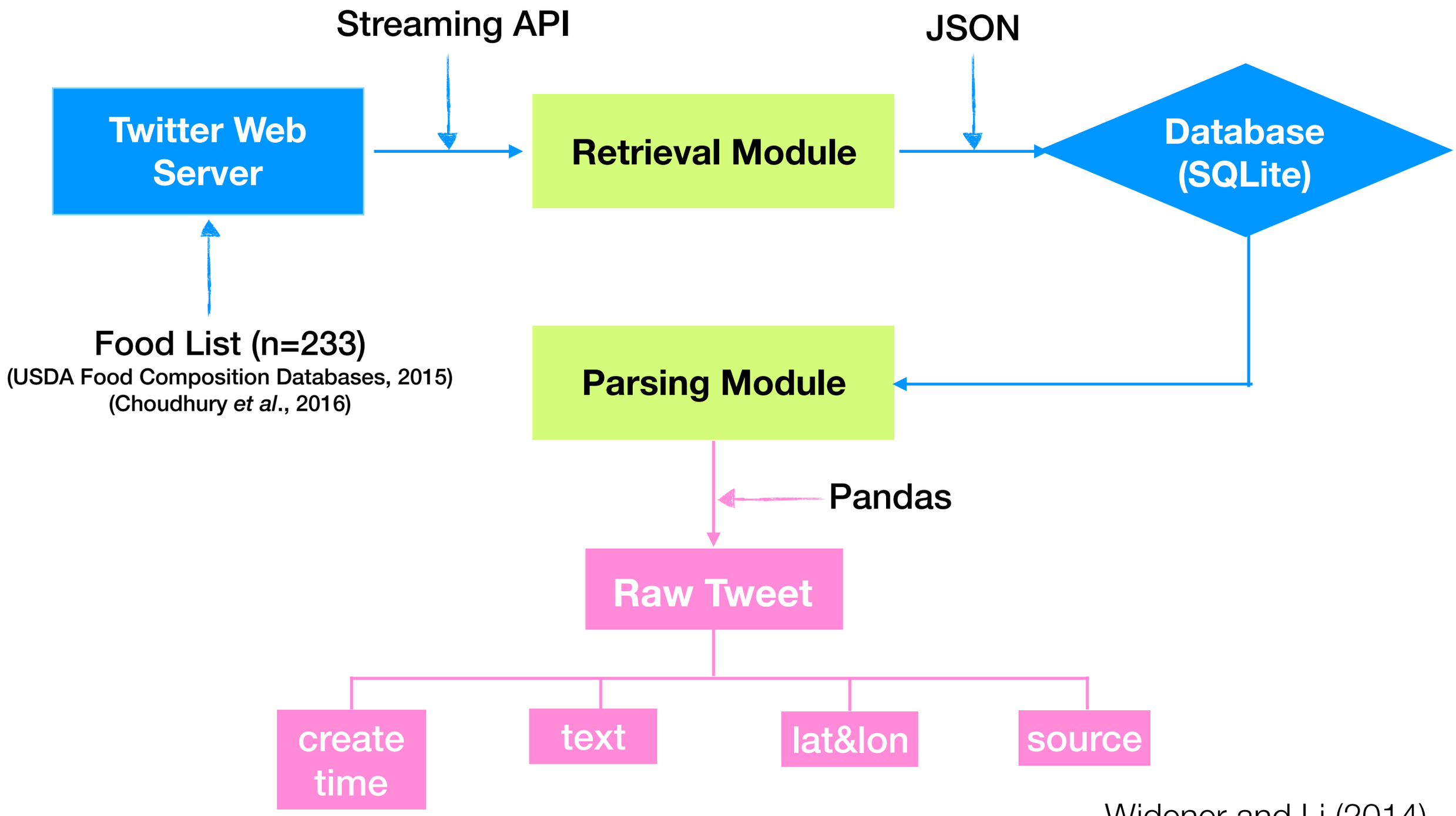
Directed acyclic graph of data needed

Flowchart


Data Collection
Twitter
USDA
FFIEC
Data Preprocessing
Filtering
Cleaning
Table joining
Feature Engineering
EDA
Feature selection
Info Extraction
Machine Learning
Logistic
XGBoost
Random Forest
Ensemble

Data Collection

Twitter Data



Widener and Li (2014)

Food Dessert Data

- Source: Food Access Research Atlas (USDA, 2013) (download)
- Data Description:
  - Census tract: food deserts are defined at tracts (n=72,217)
  - Food Desert (dependent var): urban: > 1 mile, rural: >10 miles

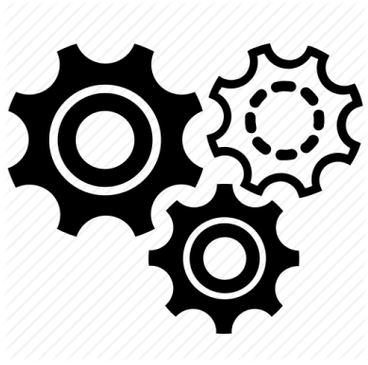
    #Low access and low income 1 and 10 = 7,436

Social-economic Data

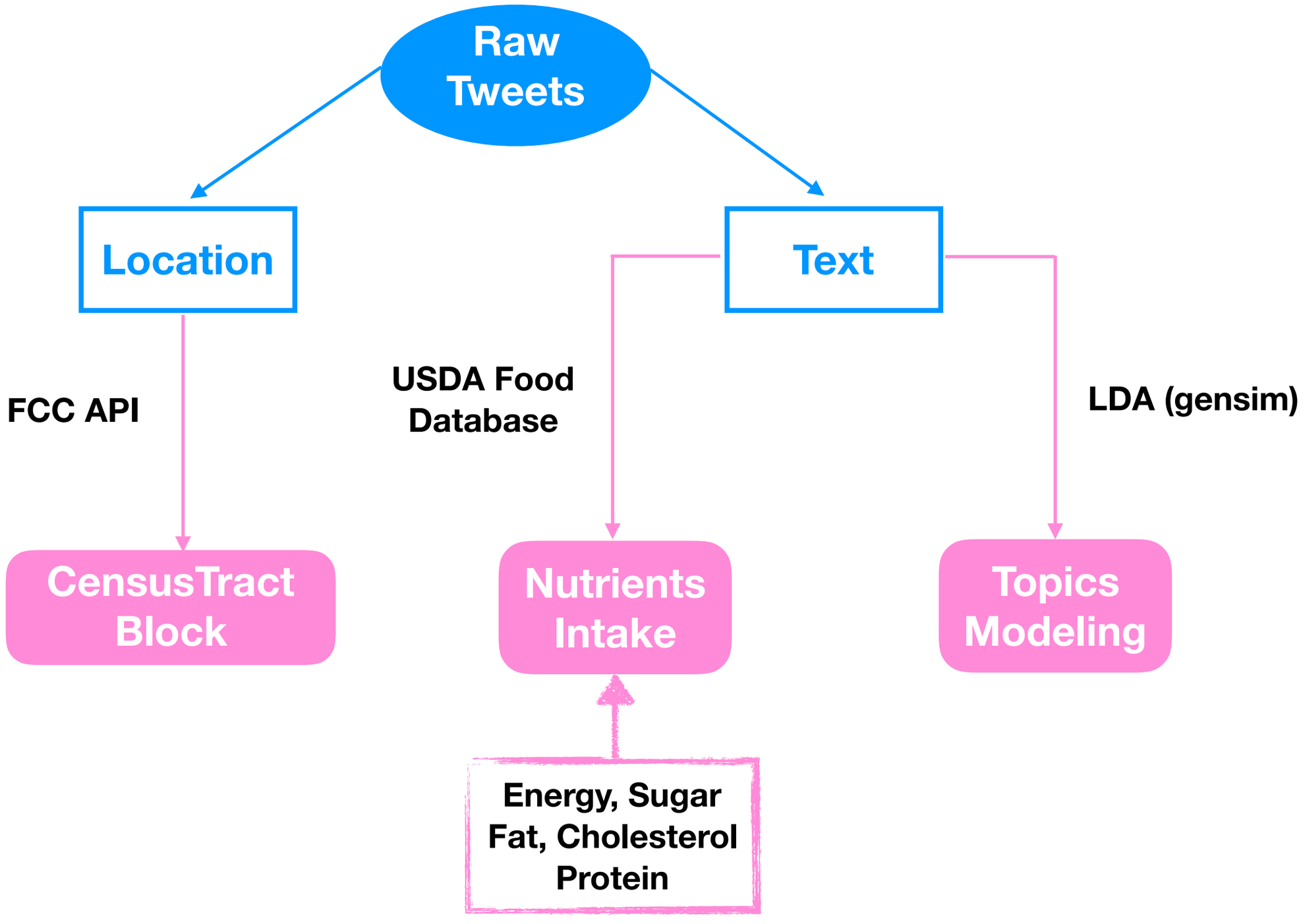
- Food Access Research Atlas (USDA, 2013)
    - Method: download
    - Data: food desert status features

- Federal Financial Institutions Examination Council's Census Reports (2016)
    - Method: python and selenium
    - Tables: demographic, population, housing, and income

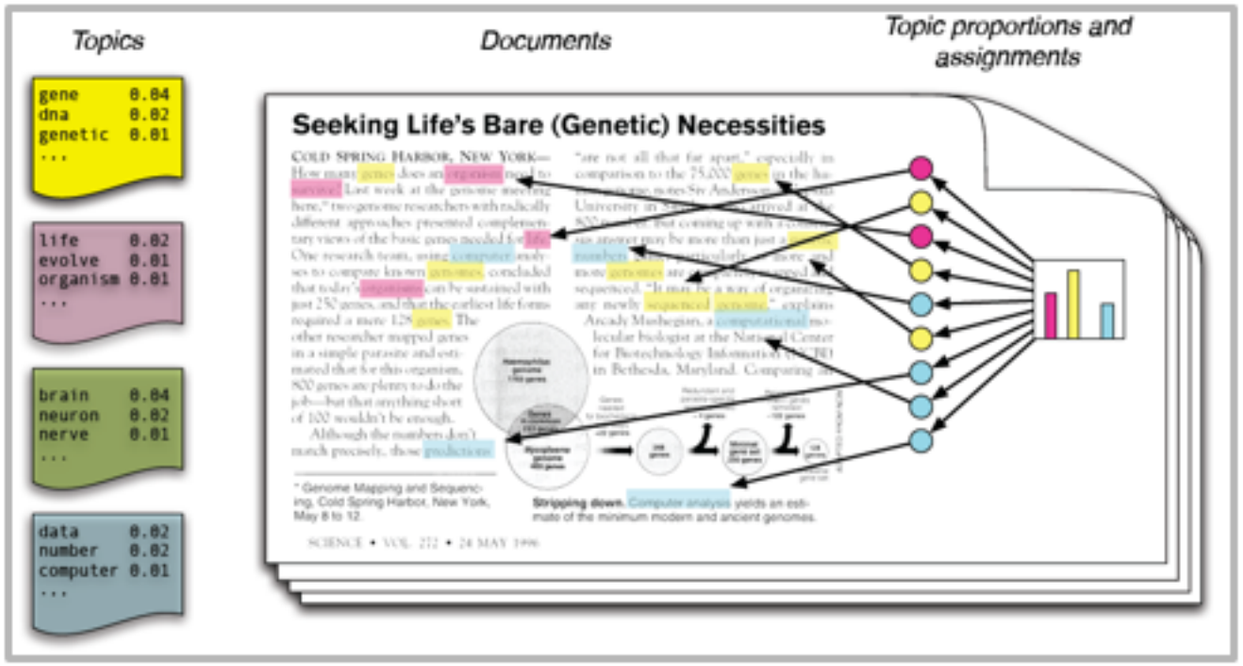
Data Preprocessing and Feature Engineering



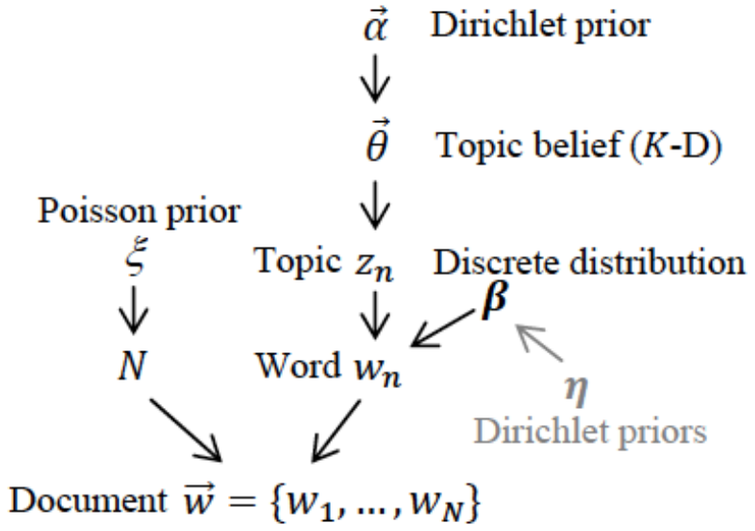
Extracting Information from Tweets



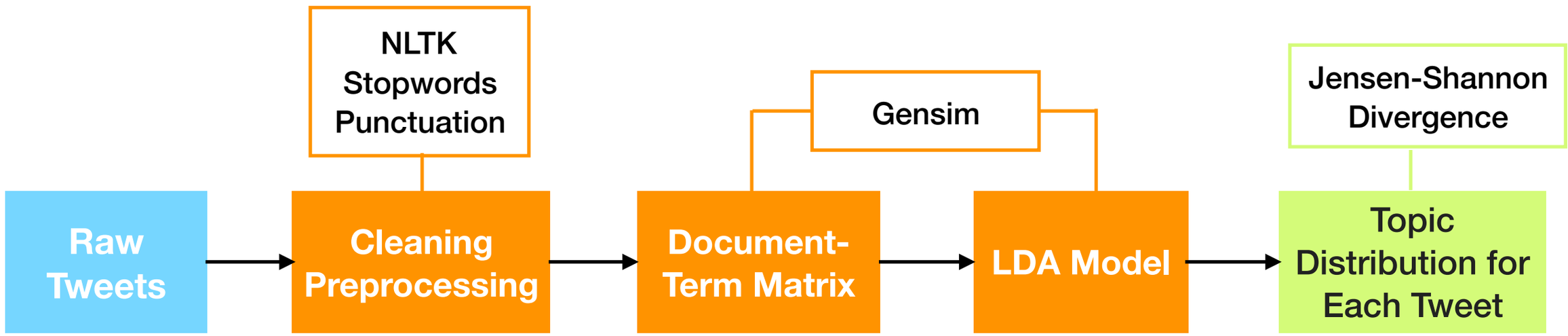
Topic Modeling - Latent Dirichlet Allocation



AnalyticsVidhya.com



Graphical representation of LDA



Summary of Dataset

Nutrients Intake

census track
energy intake
sugar intake
protein intake
cholesterol intake
lipid intake

census tract
topic distribution
(n =25)

Tweets Topics

census track (#tweets = 4,6617) → **Food Desert**

SES

census track

% minority	#households
% non-Hispanic white	median family income
median house age	#families
owner housing units	% below poverty line
distressed/underserved tract	% low access, low income
population	urban/rural
%low access 0 -17 yrs	vehicle access
%low access 65+ yrs	% group quarters population

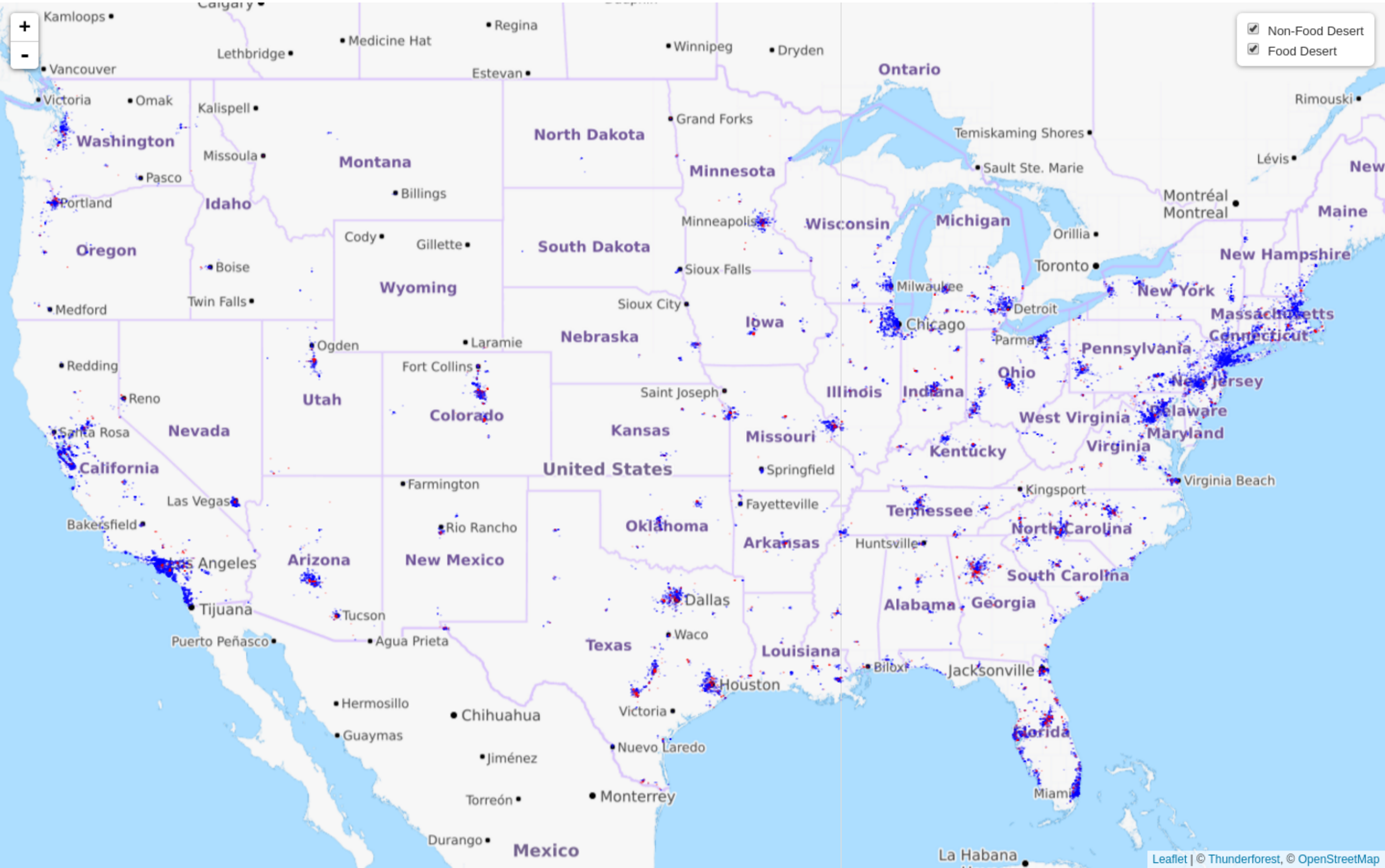
Data Exploring

Tweets: food desert vs. other tracts

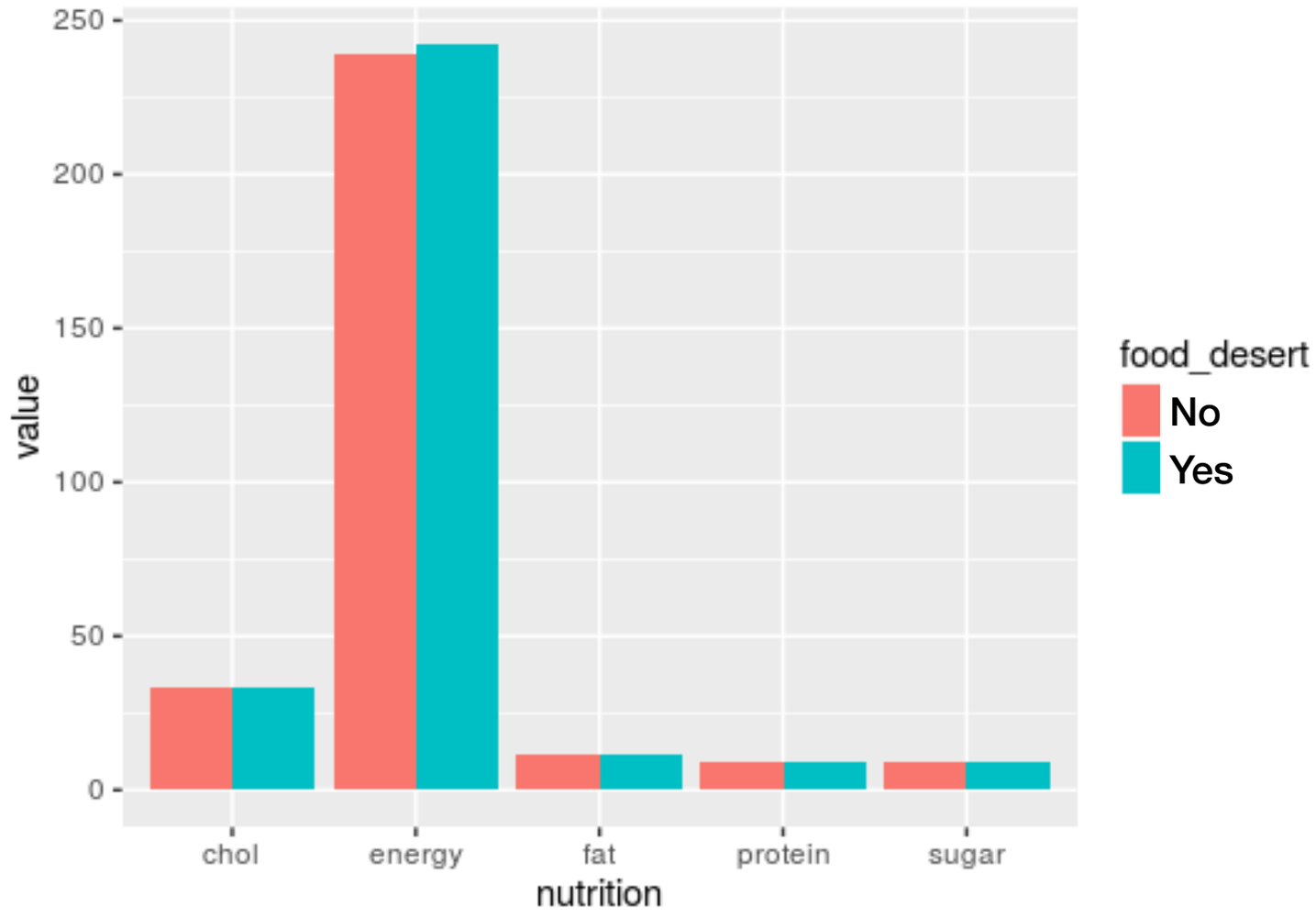
	Tweets	Tweets%
Food Desert	3972	8%
Non-food Desert	42645	92%
Total	46617	100%

→ Imbalanced dataset

Mapping Tweets in Food Deserts and Other Tracts

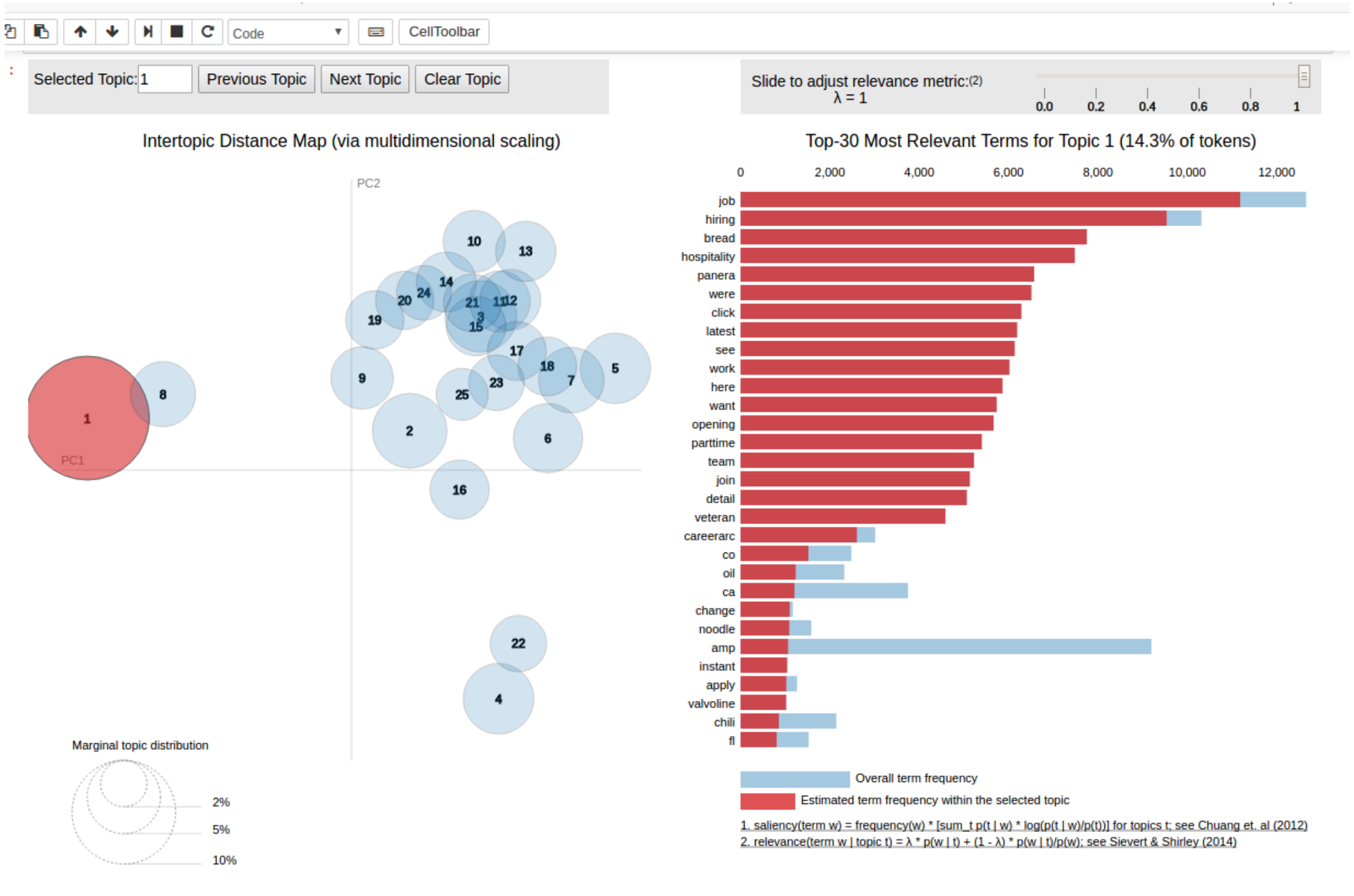


Nutrients ingestion: food desert vs. other tracts



Topics of tweets

- LDAvis

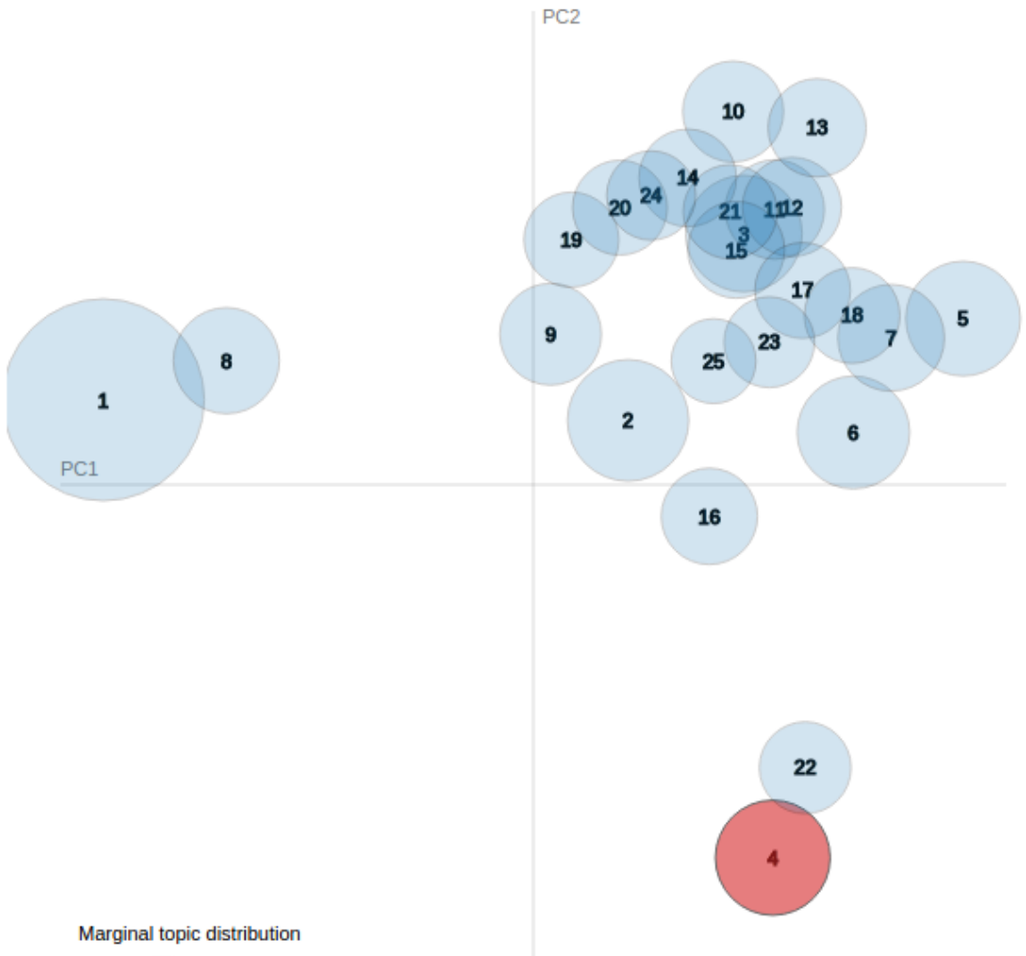


Selected Topic: 4 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)
λ = 1

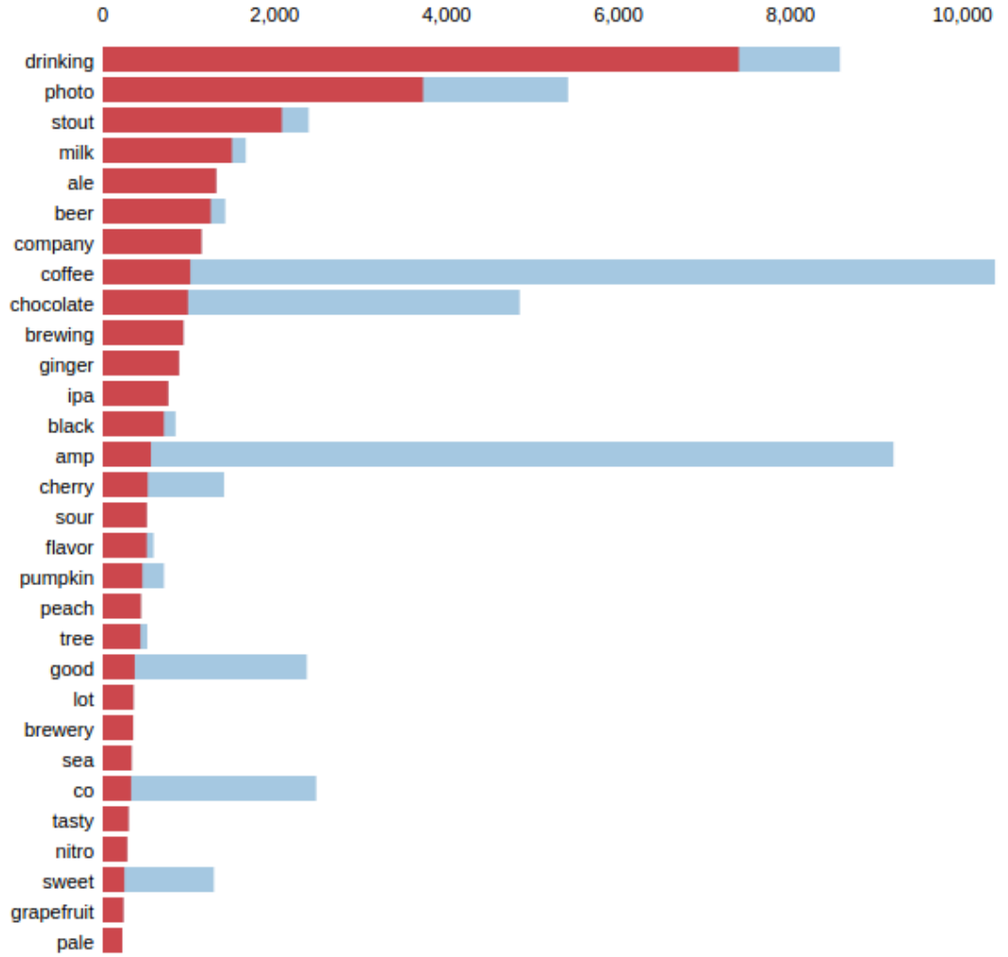
0.0 0.2 0.4 0.6 0.8 1

Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution

Top-30 Most Relevant Terms for Topic 4 (4.6% of tokens)



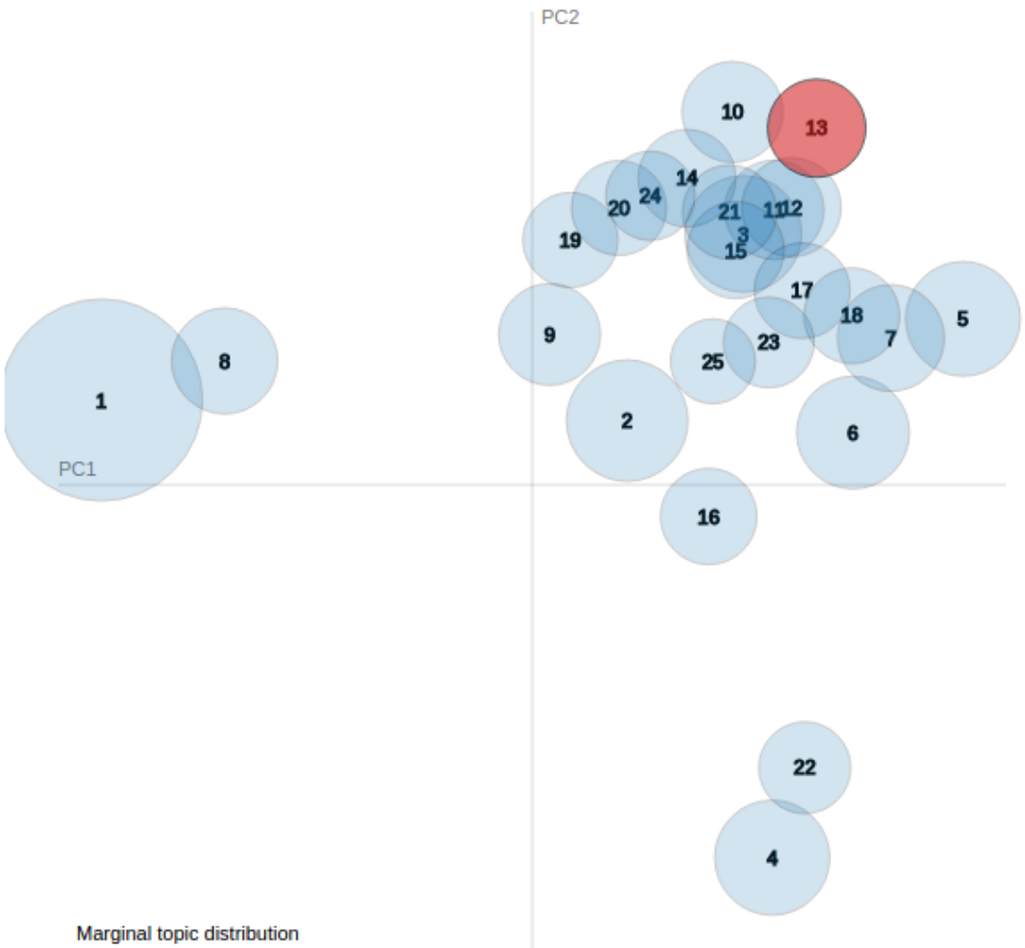
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Selected Topic: 13 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)
λ = 1

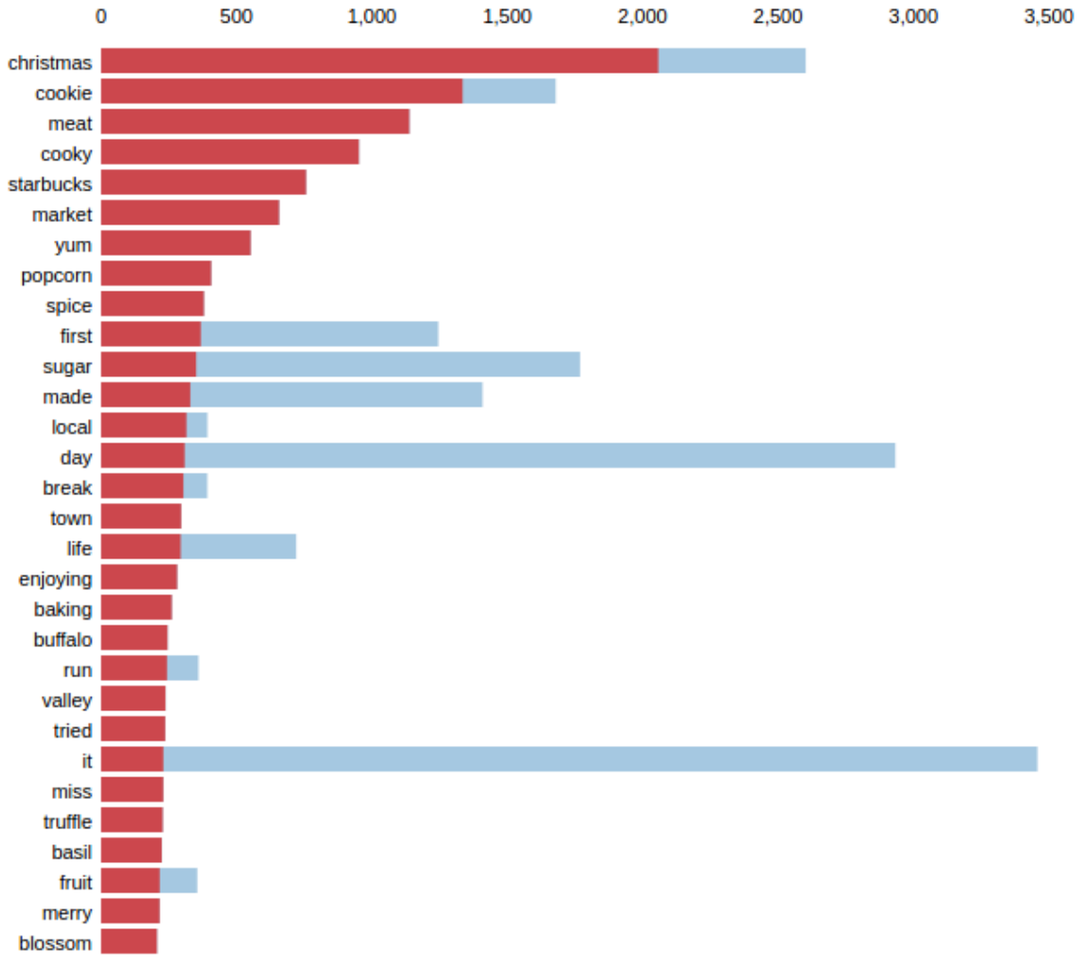
0.0 0.2 0.4 0.6 0.8 1

Intertopic Distance Map (via multidimensional scaling)



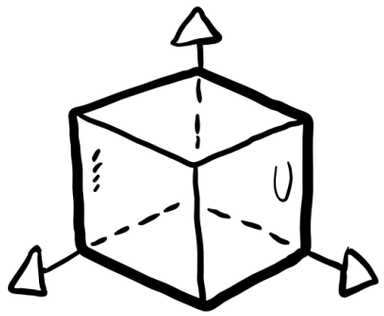
Marginal topic distribution

Top-30 Most Relevant Terms for Topic 13 (3.4% of tokens)



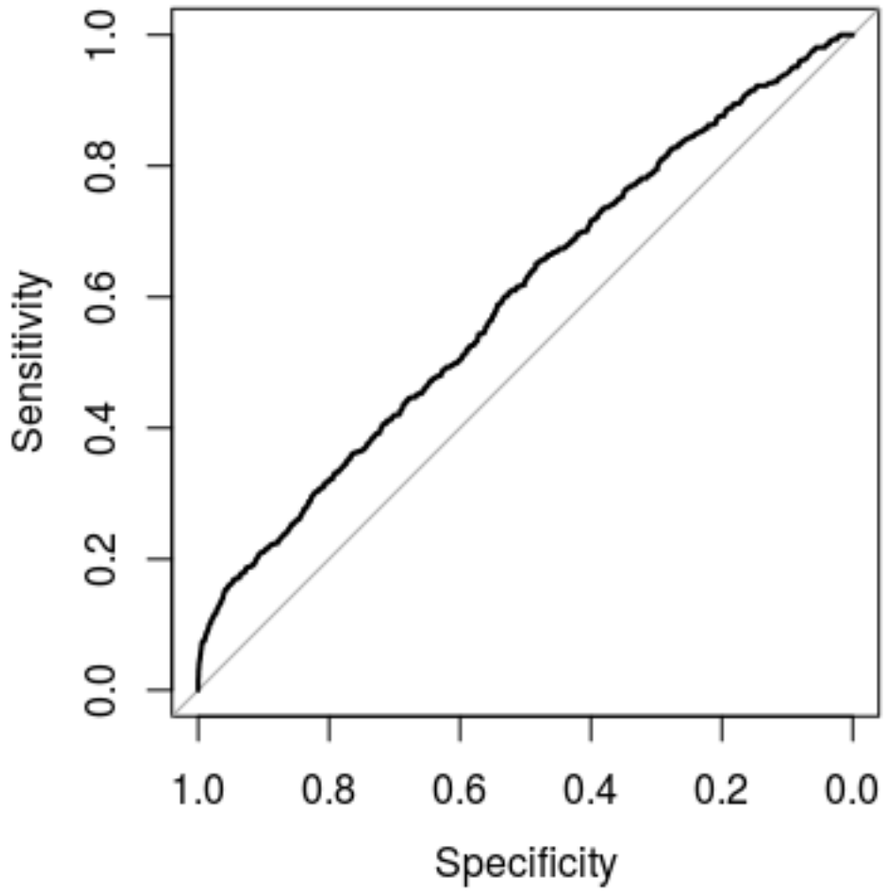
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Modeling



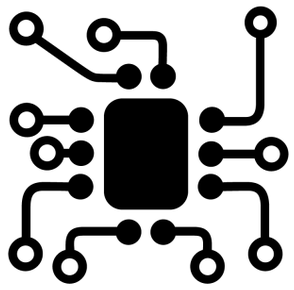
Modeling Results (test dataset)

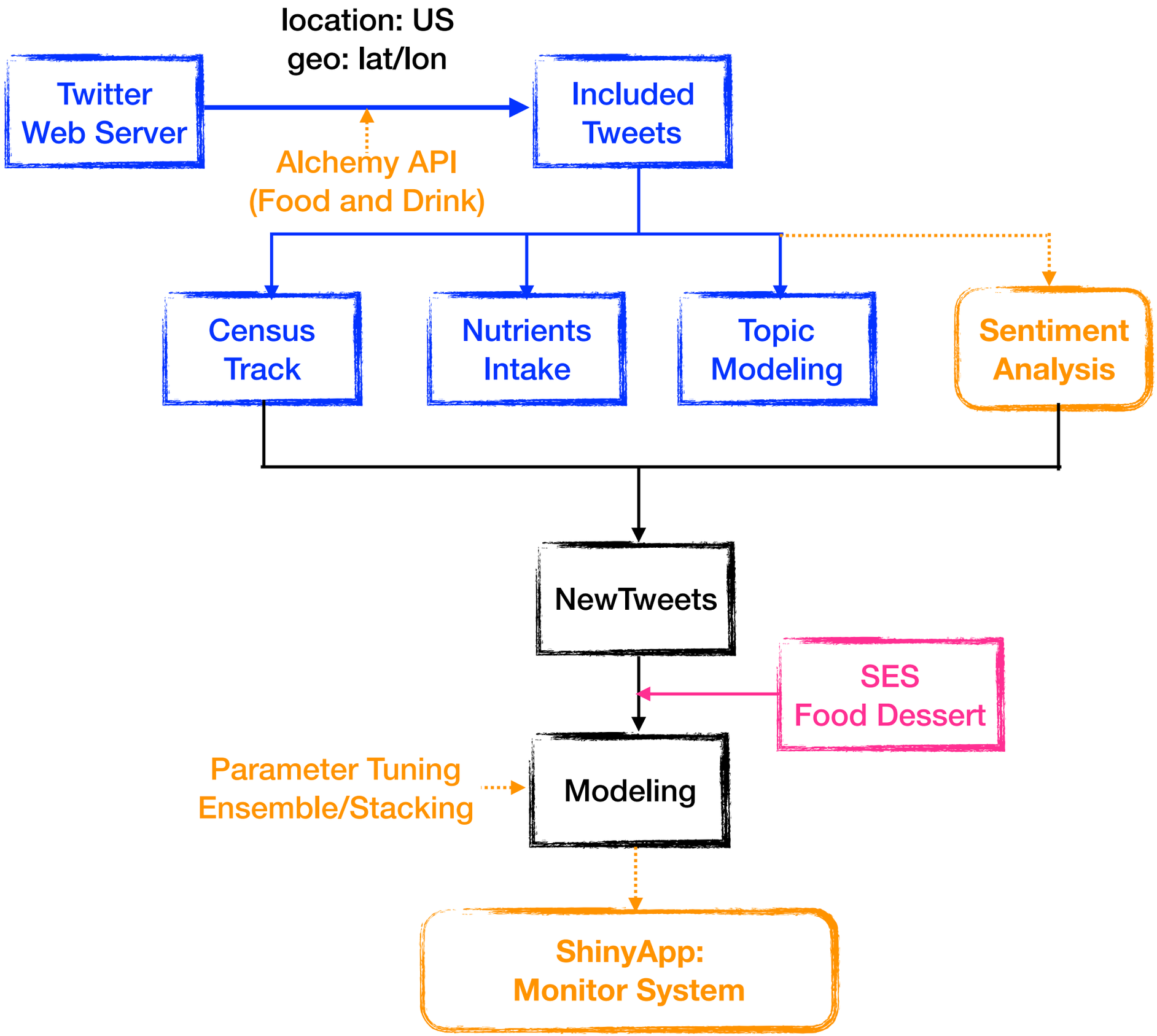
	ROC (tweets only)	ROC (tweets + SES)
logistic regression	0.524	0.95
Random Forest	0.587	0.99
XGBoost	0.567	0.99
Ensemble	0.598	



ROC Curves for “Tweets Only”

Pipeline and Future Work




location: US
geo: lat/lon
Twitter
Web Server
Alchemy API
(Food and Drink)
Included
Tweets
Census
Track
Nutrients
Intake
Topic
Modeling
Sentiment
Analysis
NewTweets
SES
Food Dessert
Parameter Tuning
Ensemble/Stacking
Modeling
ShinyApp:
Monitor System

Limitation and future work

- Tweets + Image ⟶ improve accuracy
- Noisy tweets: Alchemy API ⟶ “Food and Drink” only
- Geo-cultural difference: North East, West, Middle West ….

- Modeling: parameter tuning for “tweets only” models
- Application: streaming pipeline to monitor ingestion/food access

“Tweet! Tweet!” Coming soon !!!