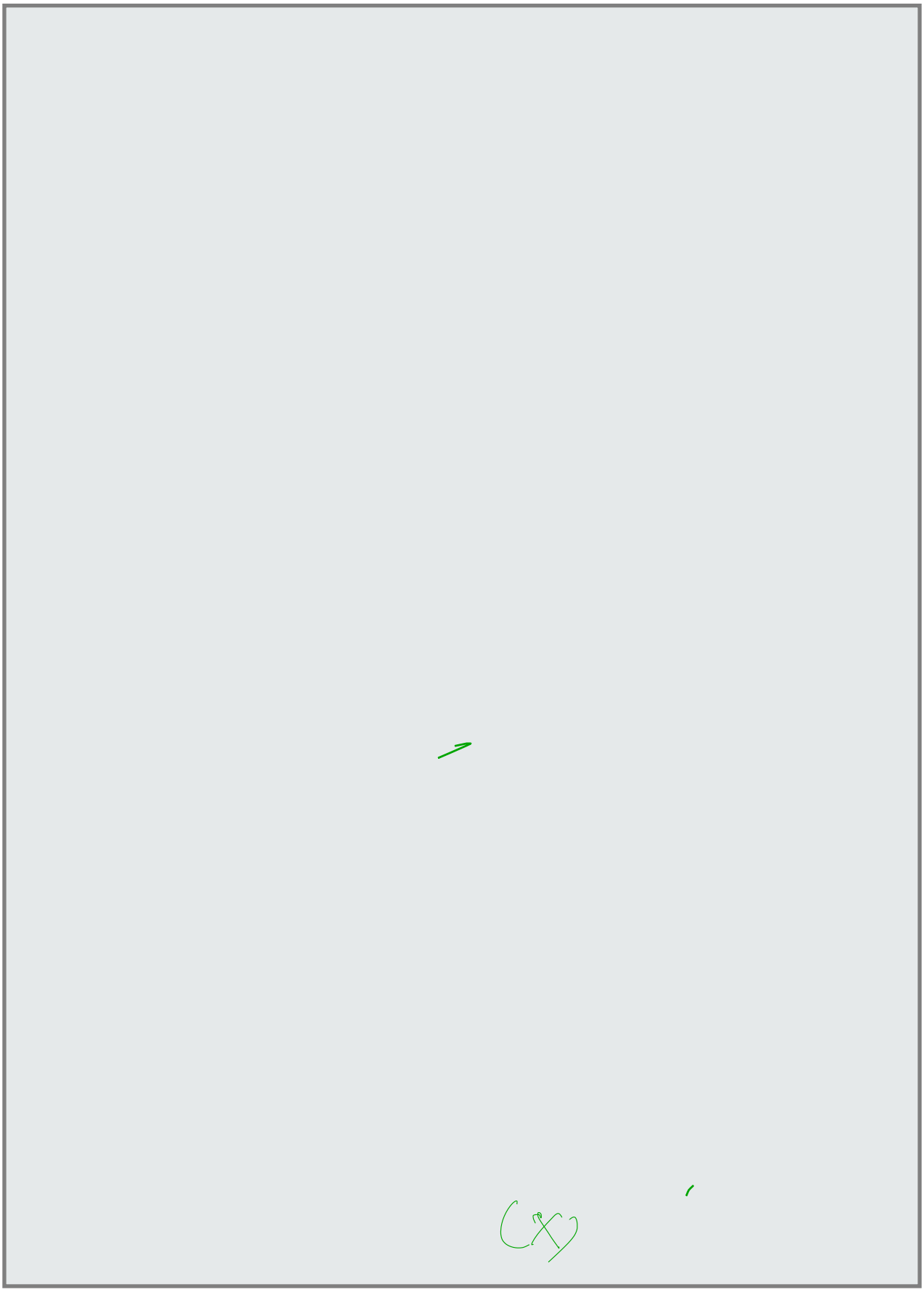
Workspace 'Workspace' in 'Intro'
Page 1 (row 1, column 1)



CMSE 381: Fundamentals of Data Science Methods

Jan 9th , 2023

About me

谢於颖 → YuYing Xie → Xie YuYing → XYY → Chr X Chr Y Y → Clark

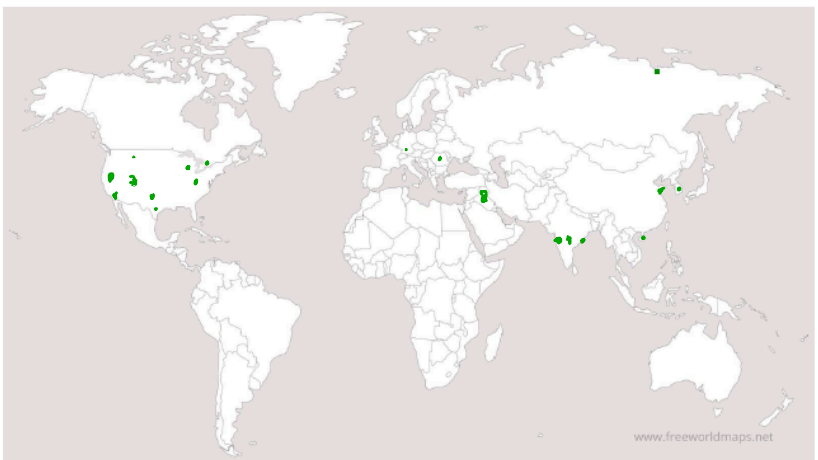
xyy@msu.edu

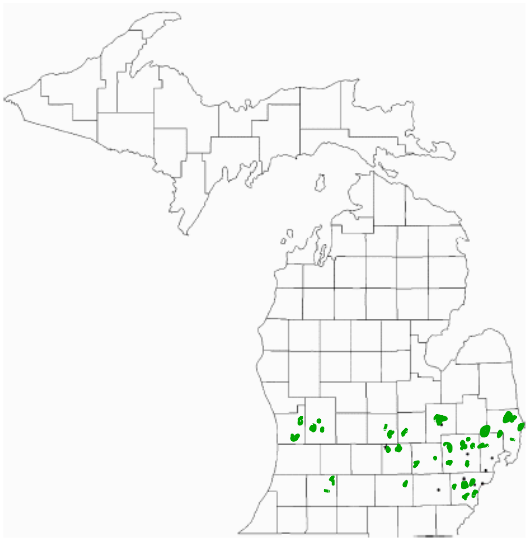
XX@msu.edu

About TA: Robert Termuhlen (termuhle@msu.edu)

Office Hours: Tue & Thur 2 – 3 pm

Introduce yourself. Annotate you





Get Printable Maps From:
Waterproof Paper.com

What is the course about?

- Topics:
  - Fundamental concepts of data science
  - Regression
  - Classification
  - Dimension reduction
  - Resampling methods
  - Tree-based methods, etc.

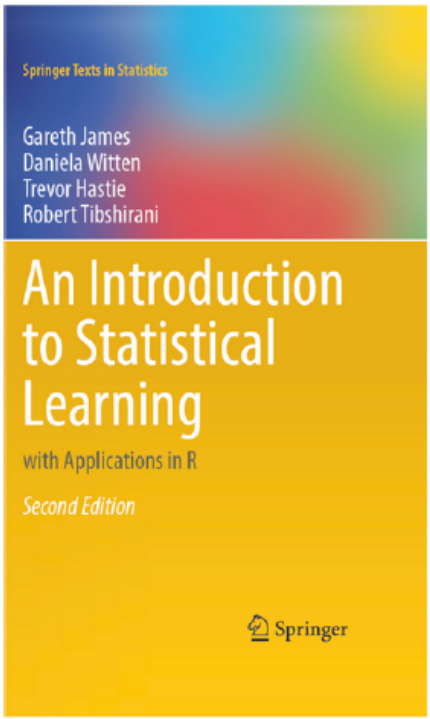
Deep learning

- Textbook:
  - An Introduction to Statistical Learning, with applications in R.
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction

paper

  - Also have a youtube list

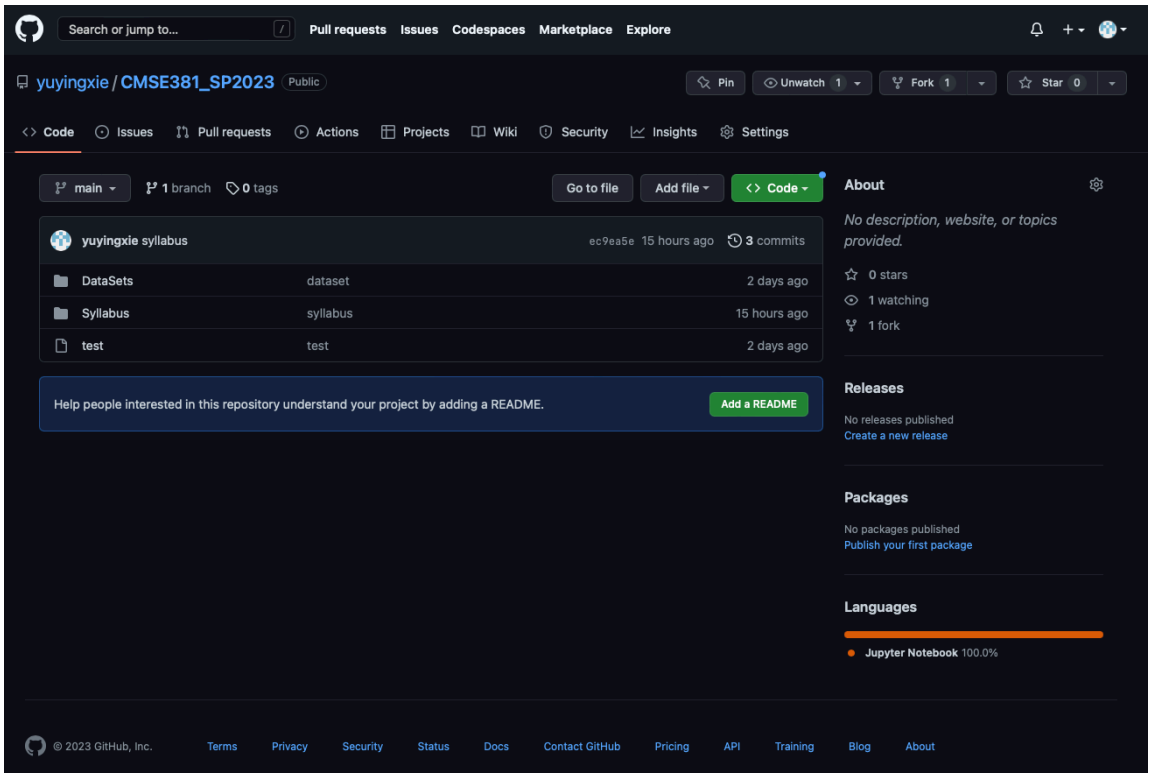
- Textbook:
  - An Introduction to Statistical Learning, with applications in
  - Online version available!
  - https://www.youtube.com/user/joshstarmer



Github and where to find slides and jupyter notebooks

MICHIGAN STATE UNIVERSITY

https://github.com/yuyingxie/CMSE381_SP2023



Office Hours

MICHIGAN STATE UNIVERSITY

Zoom Link: shorturl.at/ghpNR

Dr. Xie
Time: Tues 12:00-1:30 pm
Zoom or Egr 1513

Bob Termuhlen
Time: ~~TBA~~ Tue & Thur 2-3 pm
Zoom or Egr 1513
1508A

Instruction Format

- **In Class:**
- Stop me if you have any question!
- We will have coding section in most of the classes (bring your laptop)
- What happens in CMSE 381 stays in CMSE 381

- **Grade Breakdown:**
- 20% for Homework
- 15% for Quizzes
- 20% Midterm I
- 20% Midterm II
- 25% for the Final Project

Homework Policies

- Homework assigned bi-weekly, it is due before midnight on Friday.
- 20% of final grade.
- Late homework will have penalties (TA is full in charge.).
  - 24 hours late: 5% penalty
  - 48 hours late: 15% penalty
  - > 48 hours: No late work accepted
- Allowed to work with other students but homework should be in your own words.
- Each homework carries equal weight.
- Code needs to follow Google Style

Quiz Policies

- Once a week, there will be a 10 minutes Quiz at the beginning of the class.
- This will be basic content related to lectures since the last class. Possible questions include checking on definitions, or basic understanding of major ideas.
- 10 points per quiz
  - Drop two lowest grades
  - There will be some 3-minutes bonus-quizzes (1 credit) at the end of some classes. These credits can be added to your formal quizzes but your total Quiz-score is topped.

Graduate School

1. Research experience
2. Letters
3. Fellowship (NSF)
4. Course work
5. Broader Impact

Industrial

1. Research experience
2. Internship
3. Programming skill (Leetcode)
4. Letters

Make yourself different!

What is Machine Learning?

- Arthur Samuel (1959, IBM): *Field of study that gives computers the ability to learn without being explicitly programmed*

What is Statistical Learning?

- Subfield of Statistics
- Statistical learning emphasizes models and their interpretability, and precision and uncertainty
- Machine learning has a greater emphasis on large scale applications and prediction accuracy.
- The distinction between machine learning and statistical learning has become more and more blurred.

Why should you care?

- Nowadays, data is cheap (or even free), learning how to analyze data is critical.
  - Web data, e-commerce (Amazon, JD, Alibaba)
  - Car sales (Tesla, Ford, and GM)
  - Sport team (MSU, Lions, etc)
- Interpreting results and prevent any **bad decisions**.

Example: Presidential election

MICHIGAN STATE UNIVERSITY

http://fivethirtyeight.com/

On 2008, he correctly predicted the winner of 49 of the 50 states and all 35 U.S. Senate races .

On 2012, he correctly predicted the winner of all 50 states and the District of Columbia. That same year, Silver's predictions of U.S. Senate races were correct in 31 of 33 states.

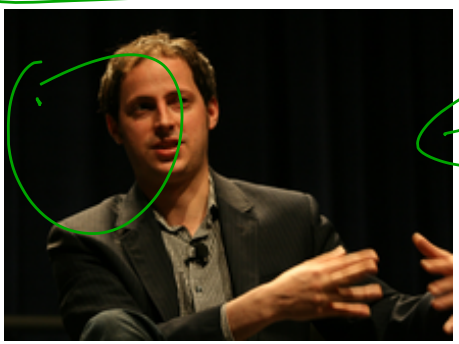
All the data he used came from internet!

Dow down 313 points after Obama won

On 2016, oops…….

Example: Presidential election 2016

MICHIGAN STATE UNIVERSITY



FiveThirtyEight

Politics Sports Science & Health Economics Culture

NOV. 11, 2016, AT 4:09 PM

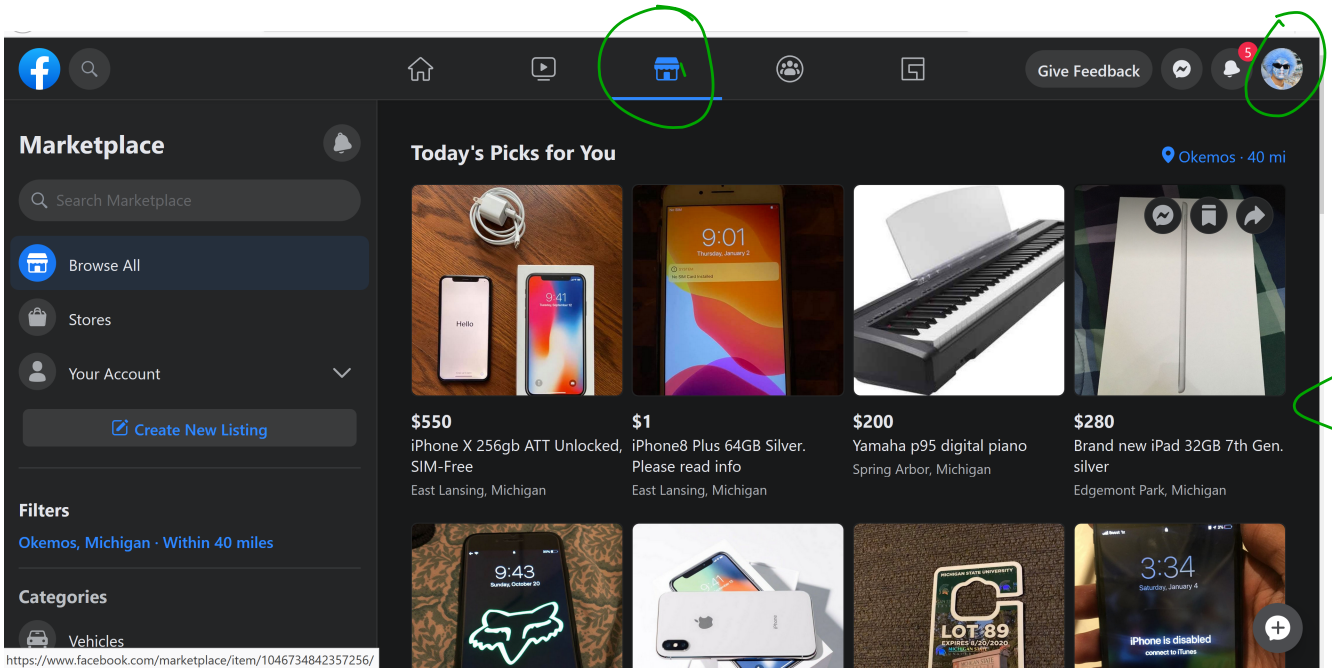
Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else

Example: Presidential election

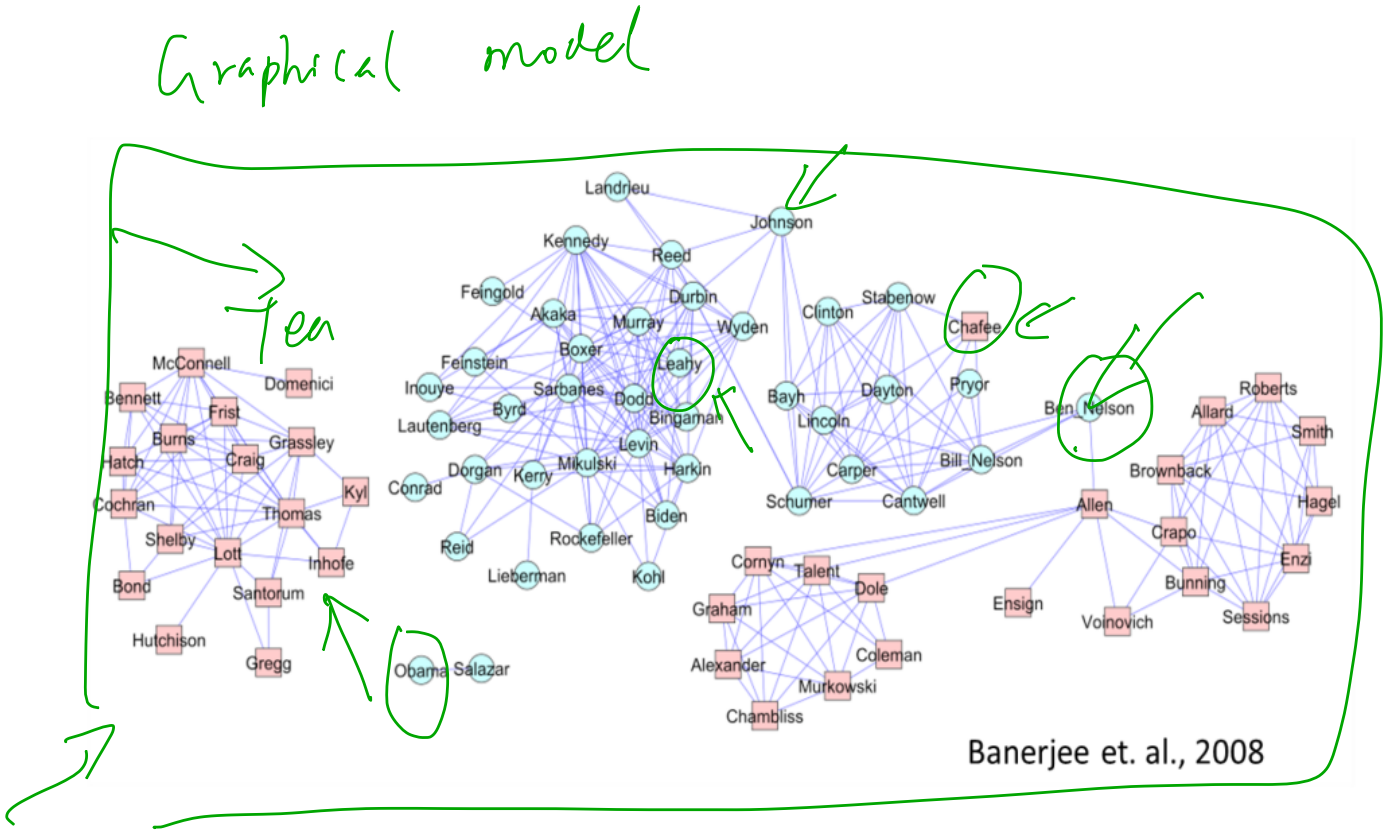
MICHIGAN STATE UNIVERSITY

Example: Big Data

MICHIGAN STATE UNIVERSITY



Politics


Graphical model
Landrieu
Johnson
Kennedy
Reed
Feingold
Durbin
Akaka
Murray
Wyden
Clinton
Stabenow
Chafee
Boxer
Feinstein
Leahy
Inouye
Sarbanes
Dodd
Byrd
Bingaman
Lautenberg
Levin
Dorgan
Kerry
Mikulski
Harkin
Conrad
Biden
Reid
Rockefeller
Lieberman
Kohl
Bayh
Dayton
Pryor
Lincoln
Carper
Bill Nelson
Schumer
Cantwell
Ben Nelson
Allard
Roberts
Smith
Brownback
Hagel
Allen
Crapo
Enzi
Bunning
Sessions
Ensign
Voinovich
Cornyn
Talent
Dole
Graham
Alexander
Coleman
Murkowski
Chambliss
McConnell
Domenici
Bennett
Frist
Burns
Grassley
Hatch
Craig
Kyl
Cochran
Thomas
Shelby
Lott
Inhofe
Bond
Santorum
Hutchison
Gregg
Obama
Salazar
Banerjee et. al., 2008

The Supervised Learning

MICHIGAN STATE UNIVERSITY

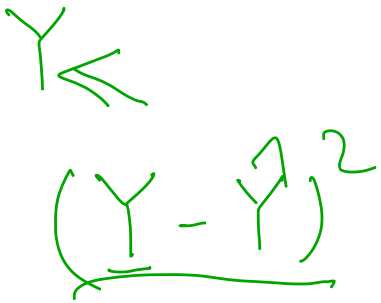
- Outcome measurement Y (also called dependent variable, response, target, label).
- Vector of *p* predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

MICHIGAN STATE UNIVERSITY

On the basis of the training data we want to?

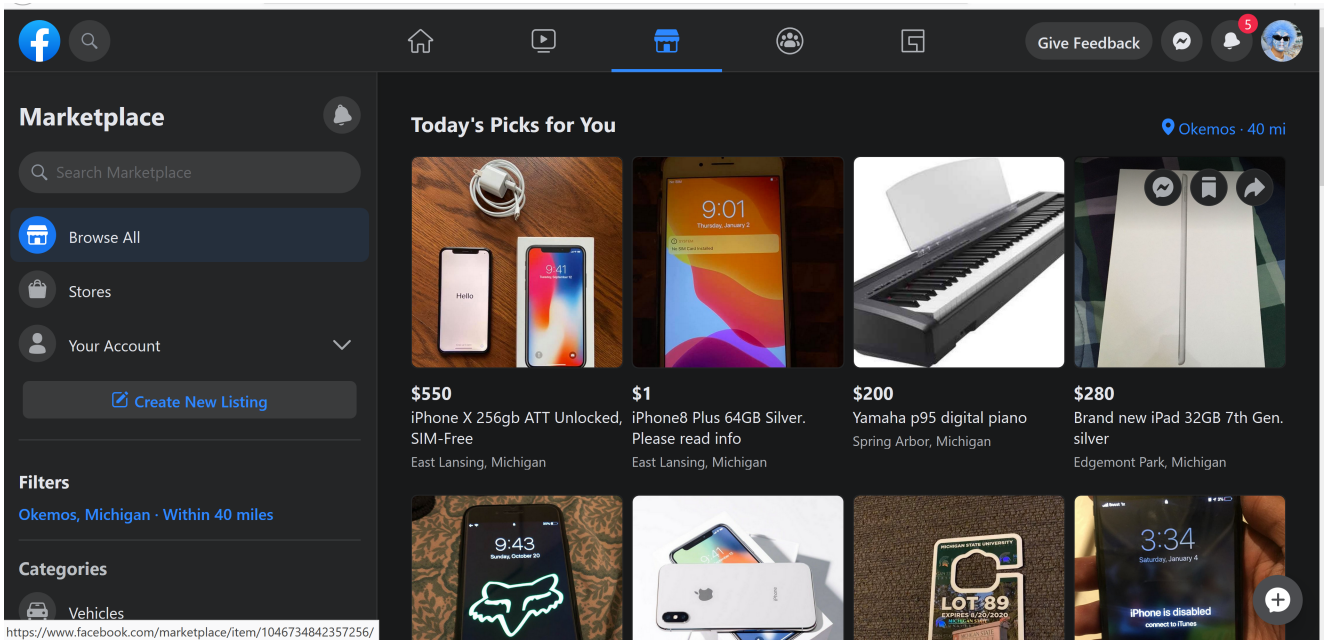
- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences



Unsupervised Learning

MICHIGAN STATE UNIVERSITY

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is fuzzier: find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Difficult to know how well you are are doing.
- Different from supervised learning but can be useful as a pre-processing step for supervised learning.


Give Feedback
Marketplace
Search Marketplace
Browse All
Stores
Your Account
Create New Listing
Filters
Okemos, Michigan · Within 40 miles
Categories
Vehicles
Today's Picks for You
Okemos · 40 mi
$550
iPhone X 256gb ATT Unlocked, SIM-Free
East Lansing, Michigan
$1
iPhone8 Plus 64GB Silver. Please read info
East Lansing, Michigan
$200
Yamaha p95 digital piano
Spring Arbor, Michigan
$280
Brand new iPad 32GB 7th Gen. silver
Edgemont Park, Michigan
https://www.facebook.com/marketplace/item/1046734842357256/

It obtains information related to customers via their mobile devices using 17 trillion location markers and various mobile applications that partner with SafeGraph.

REPORT

The effect of partisanship and political advertising on close family ties

M. KEITH CHEN AND RYNE ROHLA

Curtailed conversations

Most articles written about U.S. politics in the past few years have mentioned the increasing polarization of the electorate. But is this real, or does it merely reflect the increasing polarization of the media? Chen and Rohla estimate that in 2016, Thanksgiving dinners in which the hosts and guests lived in oppositely voting precincts were up to 50 minutes shorter than same-party-precinct dinners. That is, family members, adjured to avoid talking about contentious subjects, may have simply talked less.

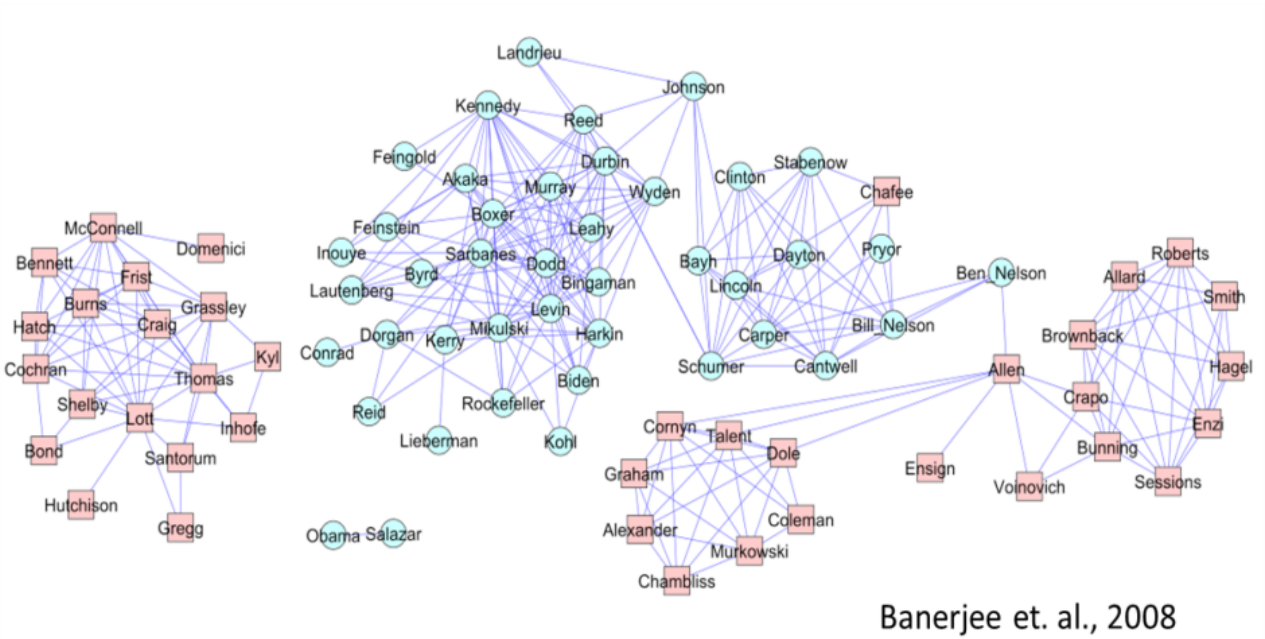
It obtains information related to customers via their mobile devices using 17 trillion location markers and various mobile applications that partner with SafeGraph.

How can you use this data?

Plan for the lab

- Find a group of 4 or so.
- Download the jupyter notebook and the csv file from github.
- Get started!

Question



Banerjee et. al., 2008