

CS 584: Data Mining

Prof. Sanmay Das
Spring 2022

Plan for today

Lecture and Covid logistics


What is this class about?

Introductions


Class logistics



Lecture logistics

- Learning is best when there's an ongoing conversation.
 - Interruptions are always welcome
 - When I ask a question, I will not continue until someone tries to answer!
 - I will use a mix of slides and the board. Slides will be posted to Piazza a few days after lecture.
 - I encourage you to take handwritten notes in lecture.
- 


Covid Impacts

- Green checks and face masks are necessary.
 - No eating or drinking in lecture, please. We will take a break midway.
 - Contingency planning:
 - If we have to cancel lecture for any reason please check Piazza for instructions on what we will do then
 - If you cannot make it to class because you are sick or quarantined or the like, post a private note to me on Piazza and we'll figure out how to handle it best.
- 

What is data mining?

- If you ask social scientists, often a pejorative term!
 - These folks often have a serious focus on *causality*. Very important issue.
- For computer scientists, discovery of novel structure (patterns, relationships, trends) in datasets, and development of models that can be used for prediction or forecasting.
- The techniques are key to what many of the big tech companies do (Netflix, Amazon, FB) and are also increasingly prevalent throughout the economy (Target, banks, supermarkets...)

Example: Credit Cards

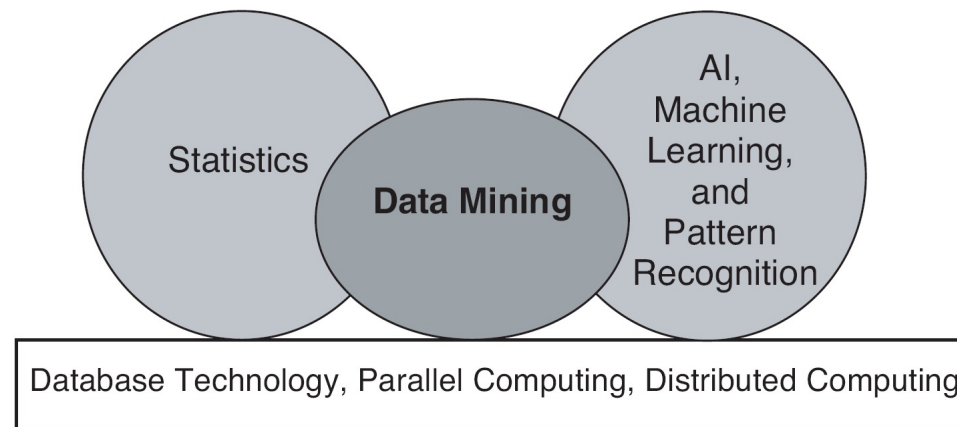
- A major bank combined credit card transaction data with credit bureau data to better predict delinquencies and defaults. [Khandani, Kim & Lo 2010]
 - What can you do with this information?
 - Prevent “run-up”
 - Estimated savings of 6-25% on credit card delinquency losses!
 - A major regulator is building on this to assess risk at broader (bank, economy) levels [Butaru et al 2016]
- 

Example: Medicare

- In the early 2010s, about 500,000 Medicare beneficiaries received hip or knee replacements every year
- Costs are both monetary and quality-of-life (first 6 months after surgery are very tough, but outcomes improve by 12 months)
- However, 1.4% of recipients die in the month after surgery, and 4.2% in months 1-12.
- These 4.2% are highly predictable using classification methods.
- Should they have gotten the surgery?
 - Discuss with each other: Key questions, whose perspective are you considering? What choices should you give individuals?
 - I'll ask some folks to discuss.

Kleinberg et al
(2015)

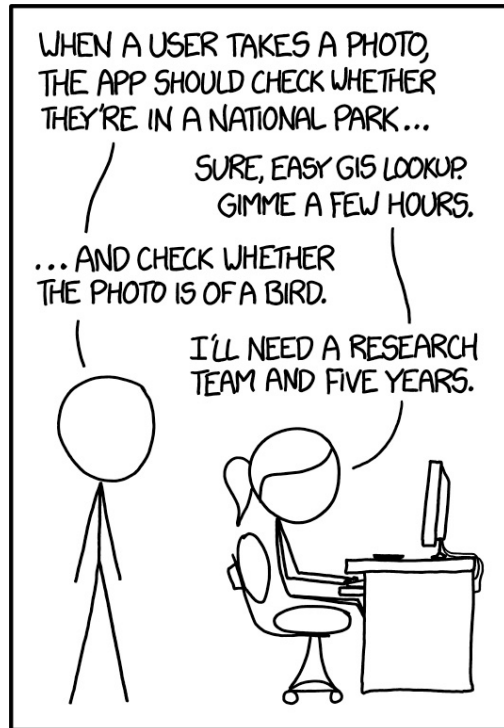
Origins of data mining



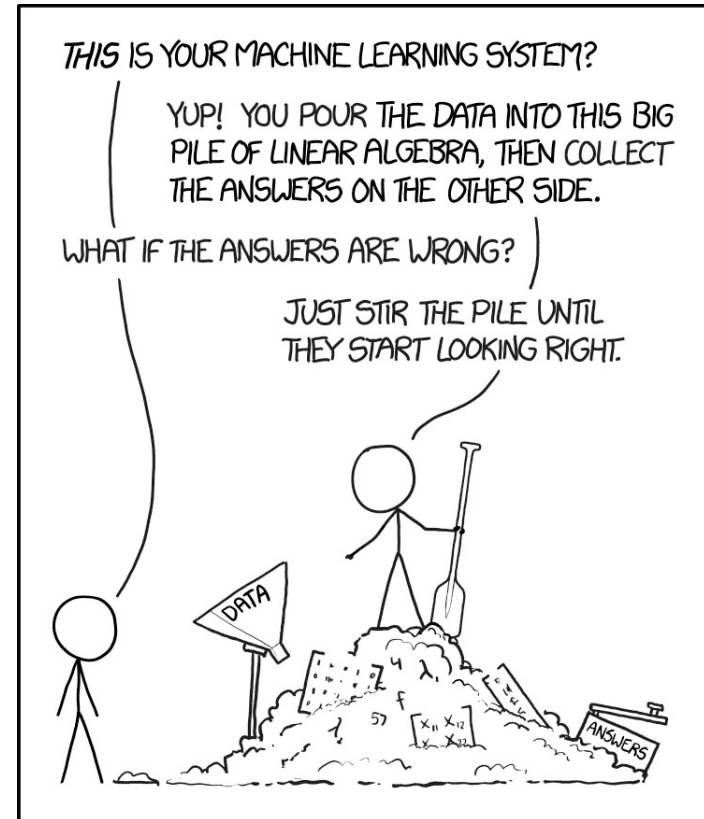
Deals with challenges of scale, dimensionality, heterogeneity, complexity, and distributed nature of data

Key component of the emerging area of data science

From the textbook




IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.



xkcd

Our main topics this semester

- Classification: Predict whether a customer will default on their mortgage or not
 - Regression: Predict how many inches of rain will fall in Fairfax tomorrow
 - Fairness, Accountability, Transparency, Ethics: How can we ensure that our algorithms aren't biased, e.g. against Black people?
 - Clustering: Divide media websites into groups that are similar to each other
 - Association Analysis: Given the items a customer has put in their shopping cart, what else might they buy?
 - Anomaly detection: Find a hacker posing as a student in my Zoom session
- 



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



Sections

The Washington Post
Democracy Dies in Darkness

sanmay

Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By **Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel**
October 17, 2016



Most Read Politics

1 House Republican leaders move to strip Rep. Steve King of his committee assignments over comments about white nationalism



2 Old land deal quietly haunts Mick Mulvaney as he serves as Trump's chief of staff



3 Trump and lawmakers paralyzed over shutdown as both sides remain dug in



4 The shutdown is giving some Trump advisers what they've long wanted: A



Clustering

Suppose you only have the features and no labels.
Still want to describe the data in some useful way

Example:

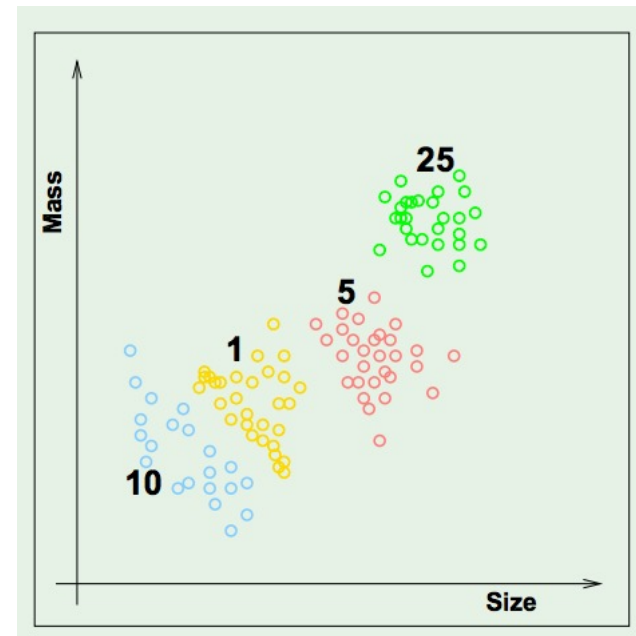
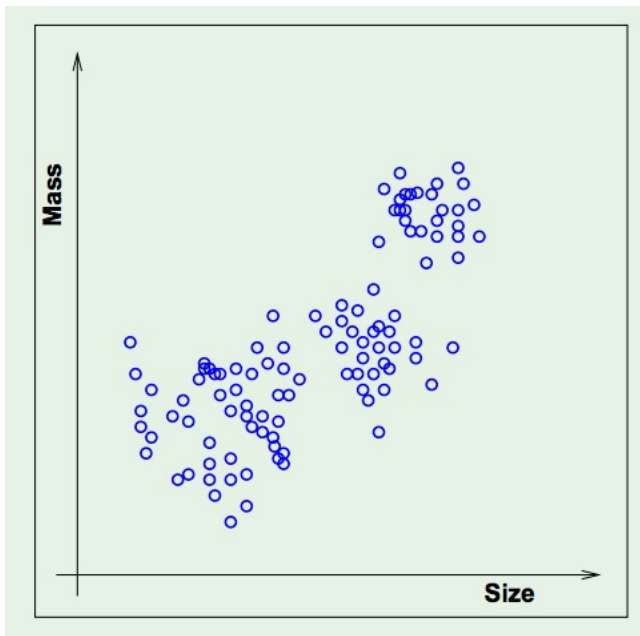
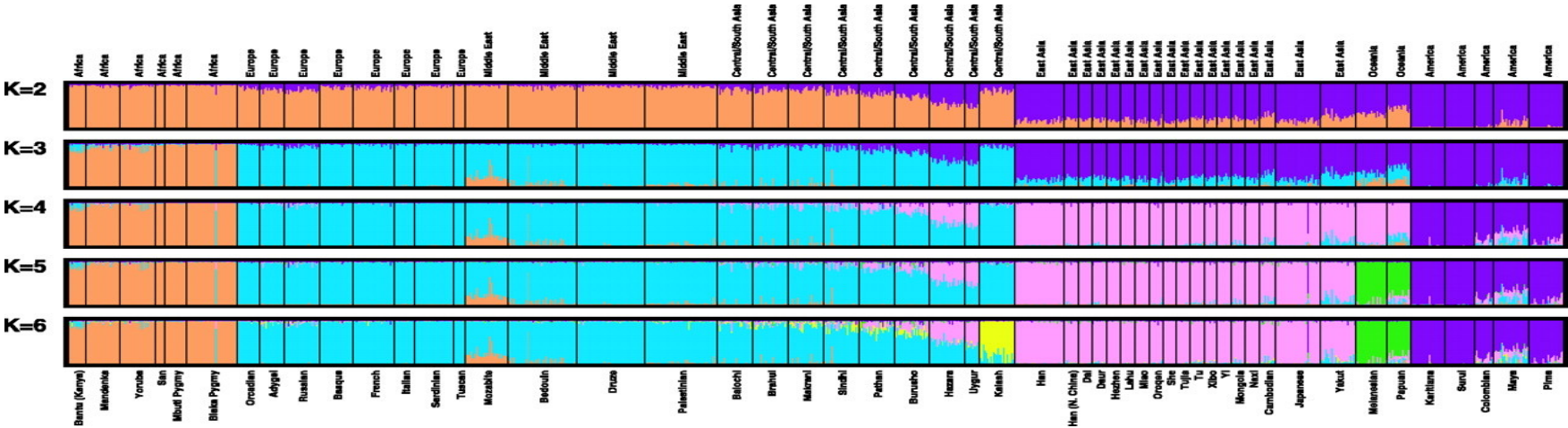


Figure 1 Estimated population structure.



N A Rosenberg et al. Science 2002;298:2381-2385



Association rule discovery

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$

Anomaly detection

- Detect deviations from “normal” behavior
- Useful for fraud detection, network intrusion detection, major changes in natural systems, etc.

Syllabus and course logistics

- Office hours will be announced soon and start next week
 - Please use Piazza as the main mode of asking questions electronically (rather than individual email)
 - Programming and math
 - Probably easiest to do the programming work in Python
 - We will heavily emphasize conceptual foundations, which require a solid grounding in math, particularly probability and statistics
 - Grading will partly be on the basis of writeups. Good communication is **critical** and at least as important as “doing the work correctly”
 - Let’s look over the syllabus together and discuss
- 