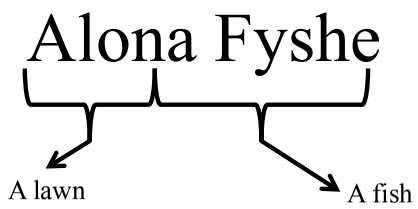
Introduction to Data Mining

Instructor: Dr. Alona Fyshe



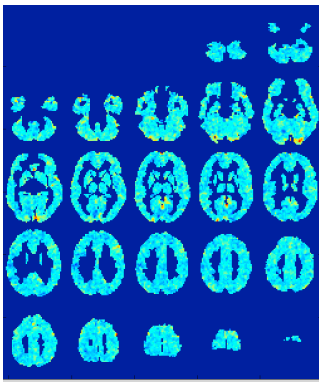
Me

- From Edmonton, AB
- Most recently lived in Pittsburgh, PA

About Me

Alona Fyshe
Email: afyshe at uvic.ca
Office: ECS 618
Office Hours: Friday 9:30-11 a.m. (right after Friday's lecture)

I study language and the brain



Why are you here?

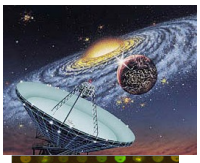
Social Media Data

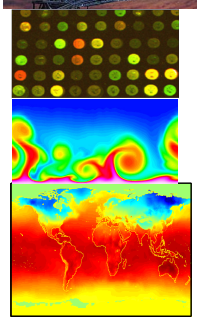
- 2 million emails per SECOND.
- 500 million tweets on Twitter a day.
- 500 million new videos viewed to YouTube a minute.
- …

- http://www.internetlivestats.com/one-second/

Scientific Data

- Data collected and stored at enormous speeds (GB/hour). E.g.
  - remote sensors on a satellite
  - telescopes scanning the skies
  - scientific simulations generating terabytes of data



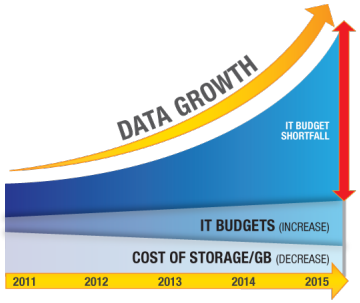


Other Data sources?

Need for Data Mining

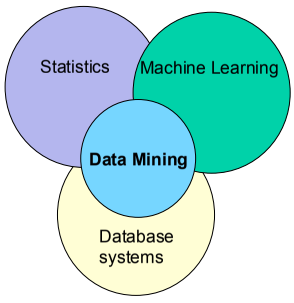
- Human analysts may take weeks to discover useful information.
  - Much of the data is never analyzed at all.

- Data mining is the process of automatically discovering useful information in large data repositories.



Origins of Data Mining

- Draws ideas from: machine learning/AI, statistics, and database systems



Data Mining Tasks

Data mining tasks are generally divided into two major categories:

- Predictive tasks [Use some attributes to predict unknown or future values of other attributes.]
  - Classification
  - Regression
  - e.g. differentiate between spam and non-spam email

- Descriptive tasks [Find human-interpretable patterns that describe the data.]
  - Association Discovery
  - Clustering
  - e.g. cluster emails by topic (school, friends, etc)

Predictive Data Mining or Supervised *learning*

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
  - e.g. for emails:
    - the attributes could be the words of the email
    - class would be spam/not spam
- Find ("learn") a *model* for the class attribute as a *function* of the values of the other attributes.
  - e.g. predict spam/not spam based on the words of the email
- Goal: Assign a class to **previously unseen** records as accurately as possible.

Learning

We can think of at least three different problems being involved in learning:

- memory
- averaging
- generalization

Example problem

(Adapted from Leslie Kaelbling's example in the MIT courseware)

- Imagine I'm trying predict whether my neighbor is going to drive into work, so I can ask for a ride.
- Whether she drives into work seems to depend on the following attributes of the day:
  - temperature
  - expected precipitation
  - day of the week
  - what she's wearing

Memory

- Okay. Let's say we observe our neighbor on three days:

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**
15	Snow	Mon	Casual	**Walk**

Memory

- Now, we find ourselves on a snowy "-5" degree Monday, and the neighbor is wearing casual clothes.
- Do you think she's going to drive?

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**
15	Snow	Mon	Casual	**Walk**
-5	Snow	Mon	Casual	

Memory

- Standard answer in this case is "yes".
  - This day is just like one of the ones we've seen before, and so it seems like a good bet to predict "yes."
- This is the most rudimentary form of learning, which is just to memorize the things you've seen before.

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**
15	Snow	Mon	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**

Noisy Data

- Things aren't always as easy as they were in the previous case. What if you get this set of noisy data?

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**?**

- We have certainly seen this case before, but the problem is that it has had different answers. Our neighbor is not entirely reliable.

Averaging

- One strategy would be to predict the majority outcome.

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**

Generalization

- Dealing with previously unseen cases
- Will she walk or drive?

Temp	Precip	Day	Clothes	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	Rain	Mon	Casual	**?**

We might plausibly make any of the following arguments:

- She's going to walk because it's raining today and the only other time it rained, she walked.
- She's going to drive because she has always driven on Mondays…

Classification: Fraud Detection

Goal: Predict fraudulent cases in credit card transactions.

Approach:

- Collect data about past transactions
  - **when** does a customer buy,
  - **what** does he buy,
  - **where** does he buy, etc.
- Label some past transactions as **fraud** or **fair** transactions.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

Trump's twitter account

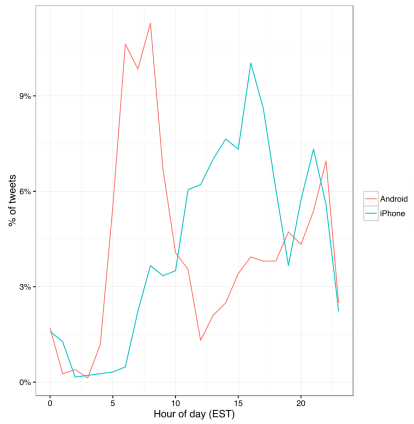
http://varianceexplained.org/r/trump-tweets/

There is some debate about who sends which of Donald Trump's tweets

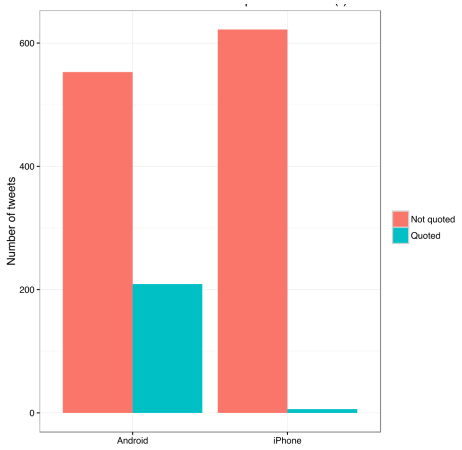
Android

Trump's twitter account

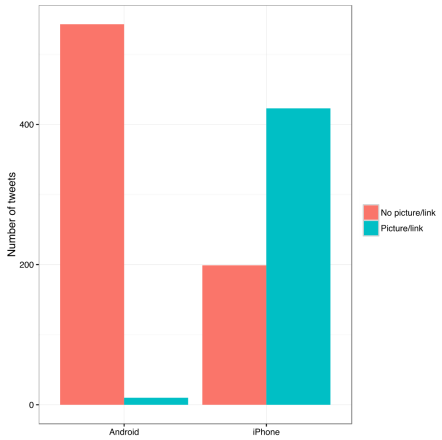
http://varianceexplained.org/r/trump-tweets/



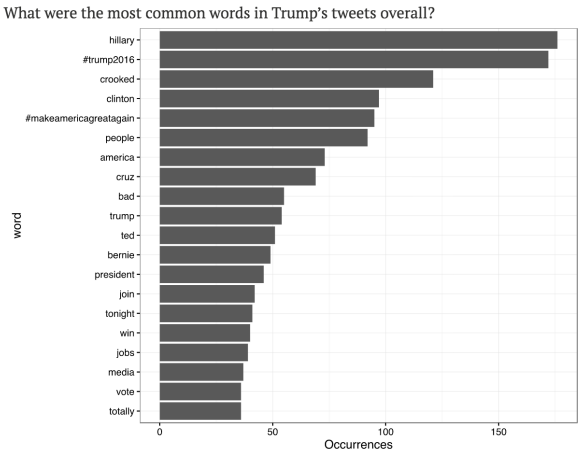
Trump's twitter account



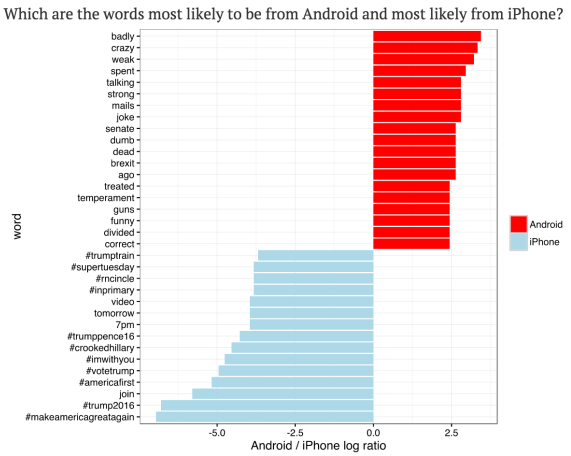
Trump's twitter account



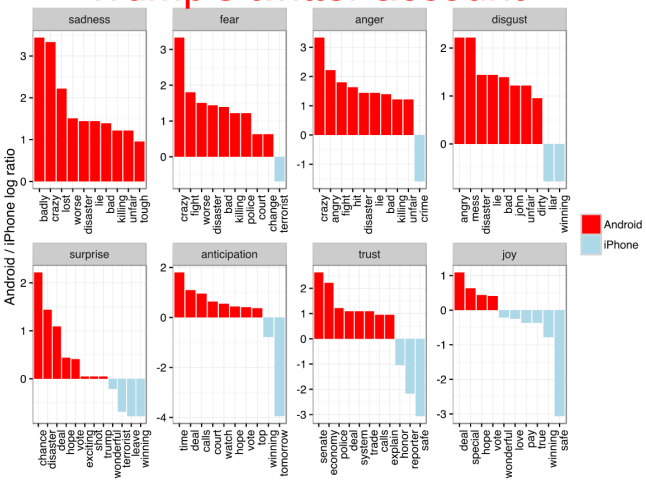
Trump's twitter account



Trump's twitter account



Trump's twitter account



Sentiment Visualization

https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Gender on Rate My Professor

http://benschmidt.org/profGender

Data Mining

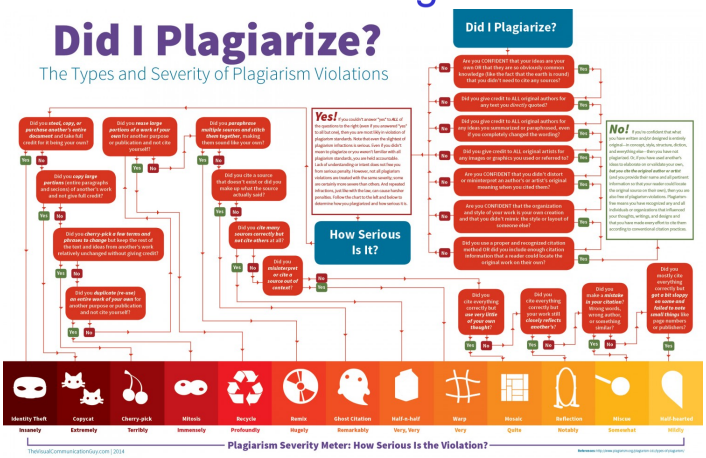
- Connex site:
  https://connex.csc.uvic.ca/portal/site/87903f62-7b6f-4f24-82ba-4c73f9e29d50
- Labs are Thursday
  - 11:30, 12:30, 1:30
- Office hours
  - Friday after class (9:30-11:00)
  - Or by appointment

Where can I find information?

- https://www.facebook.com/jeremy.hunsinger/videos/10104650538191043/
- https://heat.csc.uvic.ca/coview/outline/2016/Fall/CSC/578D

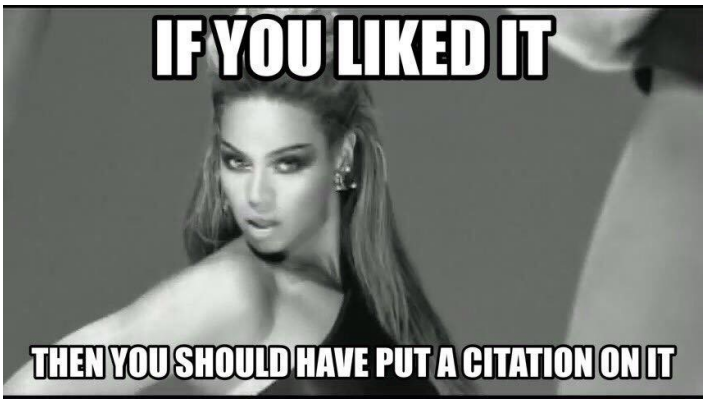
Cheating

http://visual.ly/did-i-plagiarize-types-and-severity-plagiarism-violation



Arguments I have had

- I googled the question and copied the answer I found
  - This is not “research”
- I let someone copy my assignment
- I copied their code and changed the variable names
- I copied several paragraphs from a published research paper, but I cited the paper
- P.S. I don’t decide cheating cases, they go to the academic integrity committee.
- See the full policy:
  - http://web.uvic.ca/calendar2016-09/undergrad/info/regulations/academic-integrity.html



Data Mining

- Data mining and Machine Learning
  - family of algorithms/techniques to formalize the process of finding patterns in data

- I have sad news for you.
  - Data mining is not magic, it's math!
  - A strong math/stats background will be very helpful

Data Mining

- This class involves:
  - **Python programming**. If you are rusty, or have not used Python, there are many online tutorials to help you get started. https://wiki.python.org/moin/BeginnersGuide/Programmers
  - **Linear Algebra**. If you have not taken linear algebra, the tools that you will need can be learned through this set of short videos: http://www.cs.cmu.edu/~zkolter/course/linalg/
  - **Statistics and Probability**. If you need some statistics review, here are some videos: https://www.youtube.com/playlist?list=PLRCdqbn4-qwoRTW3OpaB8-GnQwr6ta756

What Will We Cover?

- Supervised and unsupervised learning
  - supervised = ?
  - unsupervised = ?
- Supervised Learning
  - Decision Trees
  - SVMs
  - Linear/logistic regression
  - Neural networks / deep learning
- Unsupervised learning
  - Clustering algorithms
  - Recommender systems (Like Amazon/Netflix use)
  - Expectation Maximization (EM)
  - Neural networks / deep learning

Textbook

- **The textbook is *Optional:***

Introduction to Data Mining (First Edition)

Pang-Ning Tan, Michael Steinbach, Vipin Kumar

Addison Wesley (2005), ISBN: 0321321367

You do not need this textbook to do well the course. There will be no assigned readings, for example. But, if you find you learn well from textbooks, this could be a good investment.

Evaluation

- 3 assignments (with programming components), 15% total
- 1 project (30%)
- 1 midterm (15%)
- 1 final exam (40%)

Project

- Worth 30% of your mark
  - 5% Pitch & Proposal
  - 5% Mid-semester report
  - 5% Final Presentation (in-class)
  - 15% Final report

- Groups of ~5
- All Ugrad or all grad groups only please
  - it is possible students in the same group will not share the same mark

Previous Projects

- Some good projects from last year
  - http://web.uvic.ca/~afyshe/dm_projs_012016.html

Project Pitches

- Week of Sept 20 (drop deadline)
- During your lab
  - **required**
  - ~1 minute
- Convince other students (and the TA) that your project idea is interesting and worth working on.

Homework for next class

- Self study if you need a refresher in these areas:
  - **Python programming**. https://wiki.python.org/moin/BeginnersGuide/Programmers
  - **Linear Algebra**. http://www.cs.cmu.edu/~zkolter/course/linalg/
  - **Statistics and Probability**. https://www.youtube.com/playlist?list=PLRCdqbn4-qwoRTW3OpaB8-GnQwr6ta756