CSE 158

Web Mining and Recommender Systems

Introduction

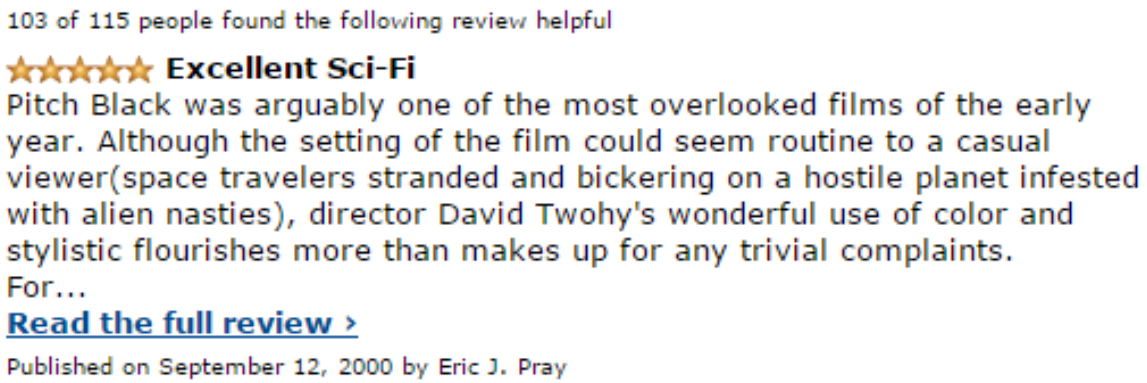
What is CSE 158?

In this course we will build models that help us to **understand data** in order to gain **insights** and make **predictions**

Examples – Recommender Systems

Prediction: what (star-) rating will a person give to a product?
e.g. rating(julian, Pitch Black) = ?

Application: build a system to recommend products that people are interested in


103 of 115 people found the following review helpful

Excellent Sci-Fi

Pitch Black was arguably one of the most overlooked films of the early year. Although the setting of the film could seem routine to a casual viewer(space travelers stranded and bickering on a hostile planet infested with alien nasties), director David Twohy's wonderful use of color and stylistic flourishes more than makes up for any trivial complaints. For...

Read the full review ›

Published on September 12, 2000 by Eric J. Pray

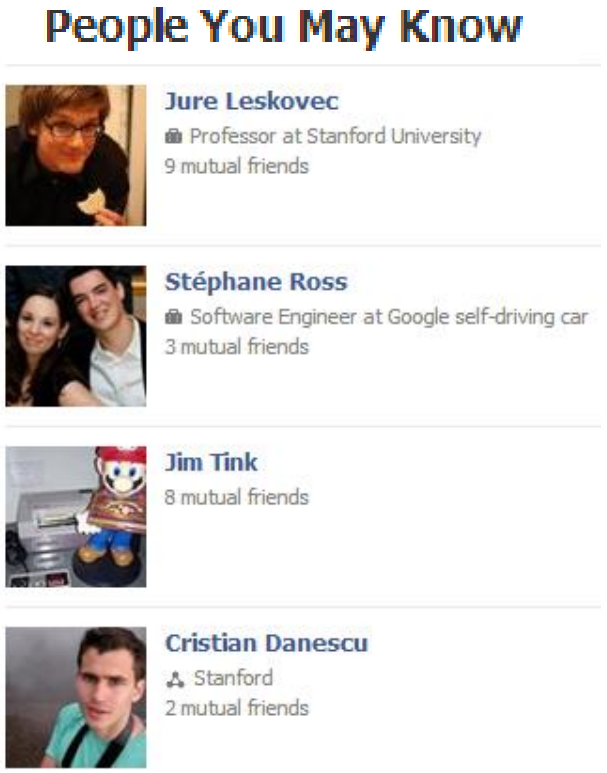
Insights: how are opinions influenced by factors like time, gender, age, and location?

Examples – Social Networks

Prediction: whether two users of a social network are likely to be friends

Application: "people you may know" and friend recommendation systems

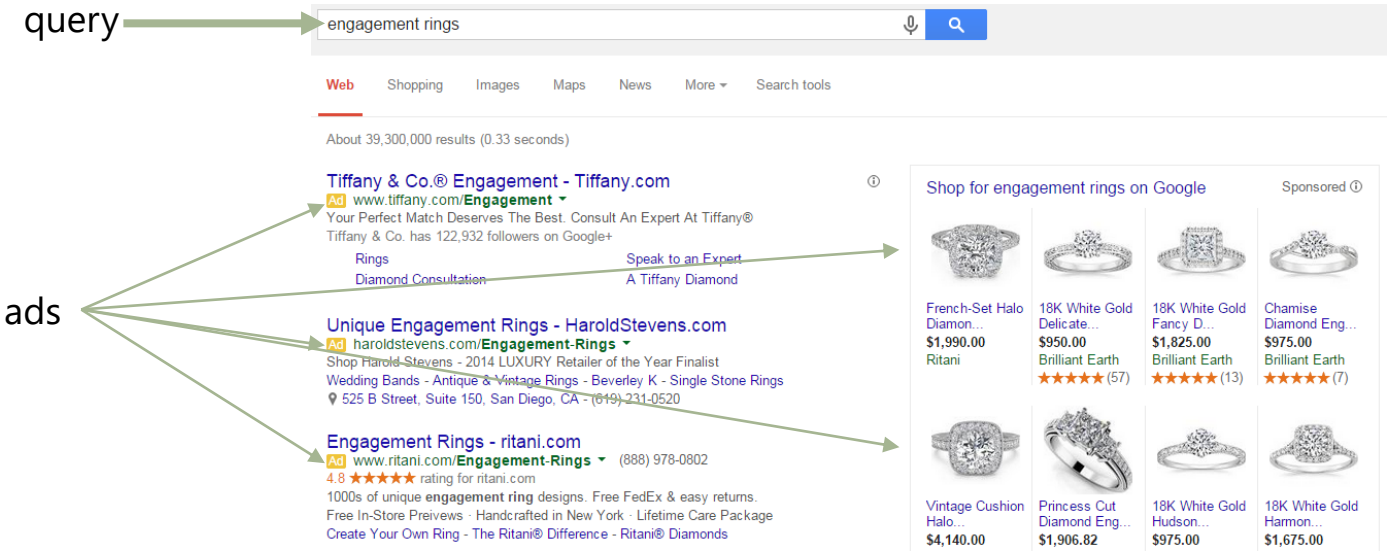
Insights: what are the features around which friendships form?



Examples – Advertising

Prediction: will I click on an advertisement?

Application: recommend relevant (or likely to be clicked on) ads to maximize revenue

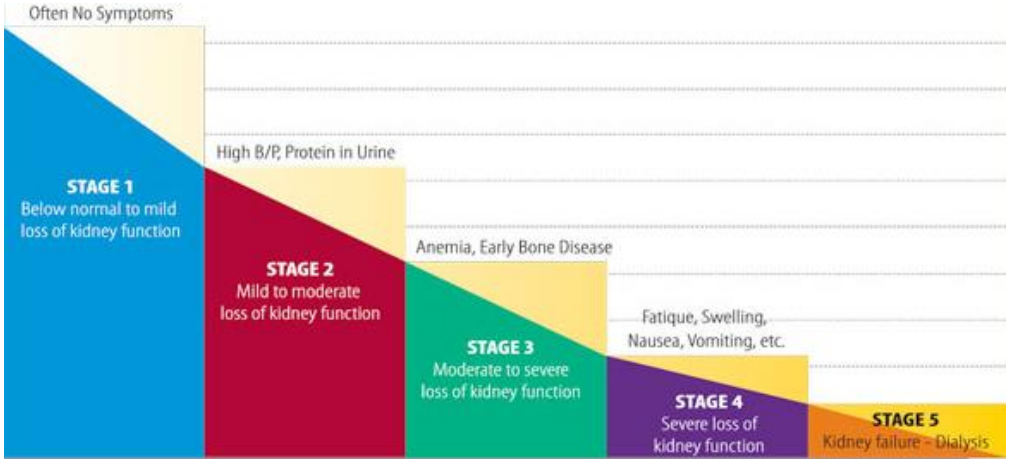


Insights: what products tend to be purchased together, and what do people purchase at different times of year?

Examples – Medical Informatics

Prediction: what symptom will a person exhibit on their next visit to the doctor?

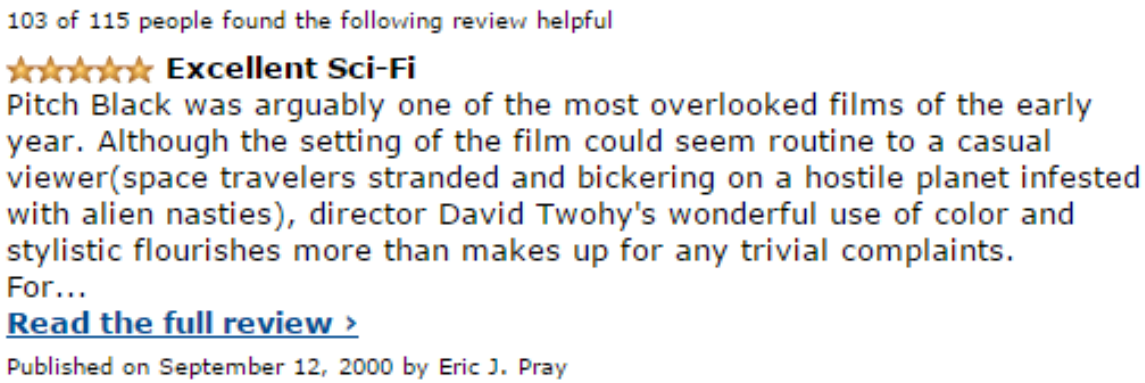
Application: recommend preventative treatment



Insights: how do diseases progress, and how do different people progress through those stages?

What we need to do data mining

1. Are the data associated with meaningful outcomes?
   - Are the data **labeled**?
   - Are the instances (relatively) independent?



103 of 115 people found the following review helpful

★★★★★ **Excellent Sci-Fi**

Pitch Black was arguably one of the most overlooked films of the early year. Although the setting of the film could seem routine to a casual viewer(space travelers stranded and bickering on a hostile planet infested with alien nasties), director David Twohy's wonderful use of color and stylistic flourishes more than makes up for any trivial complaints. For...

Read the full review ›

Published on September 12, 2000 by Eric J. Pray

e.g. who likes this movie?

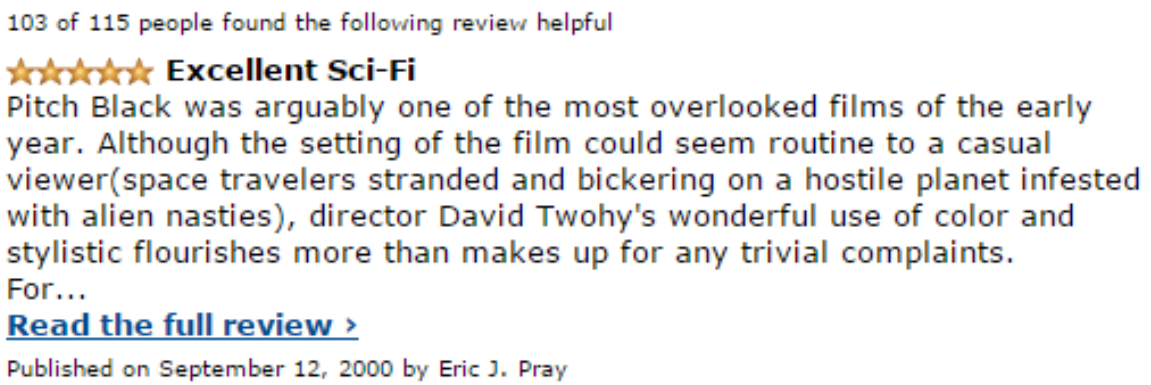
Yes! "Labeled" with a rating

e.g. which reviews are sarcastic?

No! Not possible to objectively identify sarcastic reviews

What we need to do data mining

2. Is there a clear objective to be optimized?
   - How will we **know** if we've modeled the data well?
   - Can actions be taken based on our findings?



e.g. who likes this movie?

How wrong were our predictions on average?

$$\frac{1}{N}\sum_{\text{ratings}}^{N} r_{u,i}(r_{u,i} - prediction(u,i))^2$$

What we need to do data mining

3. Is there enough data?
    - Are our results statistically significant?
    - Can features be collected?
    - Are the features useful/relevant/predictive?

What is CSE 158?

This course aims to teach

- How to **model** data in order to make **predictions** like those above
- How to **test and validate** those predictions to ensure that they are meaningful
- How to **reason about** the findings of our models

(i.e., "data mining")

What is CSE 158?

But, with a focus on applications from **recommender systems and the web**

- **Web** datasets

- Predictive tasks concerned with human **activities, behavior, and opinions**

(i.e., recommender systems)

Expected knowledge

Basic data processing

- Text manipulation: count instances of a word in a string, remove punctuation, etc.
- Graph analysis: represent a graph as an adjacency matrix, edge list, node-adjacency list etc.
- Process formatted data, e.g. JSON, html, CSV files etc.

Expected knowledge

Basic mathematics

- Some linear algebra $Ax = y \rightarrow x = (A^T A)^{-1} A^T y$
- Some optimization $\frac{d}{dx}(Ax - y)^2$
- Some statistics (standard errors, p-values, normal/binomial distributions)

Expected knowledge

All coding exercises will be done in **Python** with the help of some libraries (numpy, scipy, NLTK etc.)

Expected knowledge

Idea with "expected knowledge" is not that you know all of these things, but rather than you learn those that you don't **on your own**

See e.g. some student comments on the course:

Comment 1: "*I felt that the first four weeks of the course was slow... similar to all ML courses taught here, they review the same material on fundamentals of data science/machine learning*"

Comment 2: "*Difficult if you have not had any machine learning/data mining experience*"

CSE 158 vs. CSE 150/151

The two most related classes are

- CSE 150 ("Introduction to Artificial Intelligence: Search and Reasoning")
- CSE 151 ("Introduction to Artificial Intelligence: Statistical Approaches")

None of these courses are prerequisites for each other!

- CSE 158 is more "hands-on" – the focus here is on applying techniques from ML to real data and predictive tasks, whereas 150/151 are focused on developing a more rigorous understanding of the underlying mathematical concepts

CSE 158 vs. CSE 258

CSE 258 is the **graduate** version of this class. It is roughly the same, though there are some differences:

- CSE 258 will have more on graphical models (we'll cover it a little bit in 158, but not much)
- CSE 258 will have a little bit more on optimization (e.g. gradient based methods). We'll cover these too, but not really with complex derivations – in this class some of the more complex linear algebra / calculus will be treated in more of a "black box" way
- CSE 258 will cover more academic papers

- As long as you do the CSE 158 assessments, you're welcome to attend either class (but not this week!)

CSE 158 vs. CSE 258

Both classes will be podcast in case you want to check out the more advanced material:

(last year's links)

CSE158:
http://podcasts.ucsd.edu/podcasts/default.aspx?PodcastId=3746&v=1

CSE258:
http://podcasts.ucsd.edu/podcasts/default.aspx?PodcastId=3747&v=1

Lectures

In Lectures I try to cover:

- The basic material (obviously)
- **Motivation** for the models
- **Derivations** of the models

- Code examples
- Difficult homework problems / exam prep etc.
- **Anything else you want to discuss**

CSE 158

Web Mining and Recommender Systems

Course outline

Course webpage

The course webpage is available here:
http://cseweb.ucsd.edu/classes/fa18/cse158-a/

This page will include data, code, slides,
homework and assignments

Course webpage

(the previous course webpage is here):
http://cseweb.ucsd.edu/classes/fa17/cse158-a/

This quarter's content will be (roughly) similar

Course outline

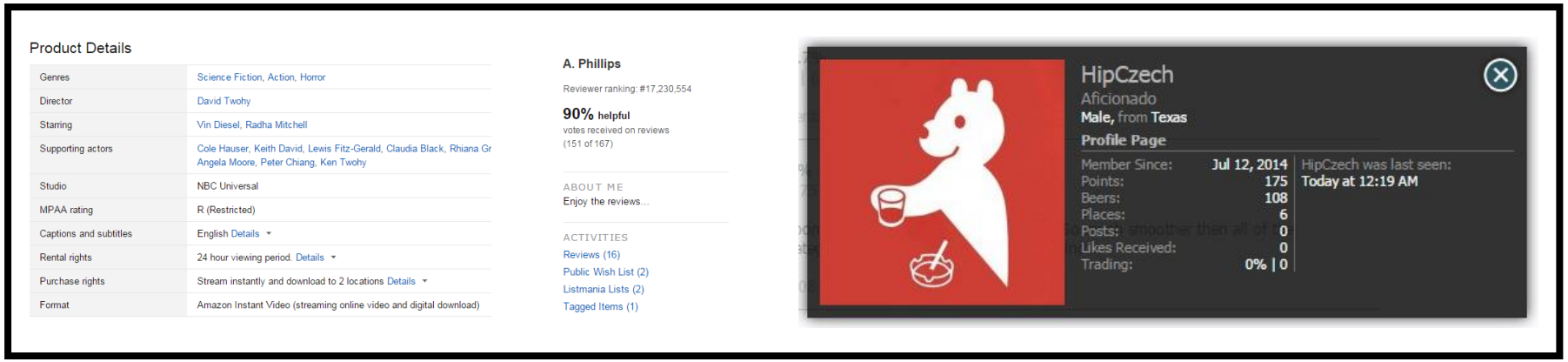
This course in in two parts:

1. **Methods** (weeks 1-3):
   - Regression
   - Classification
   - Unsupervised learning and dimensionality reduction
2. **Applications** (weeks 4-):
   - Recommender systems
   - Text mining
   - Social network analysis
   - Mining temporal and sequence data
   - Something else?

Week 1: Regression

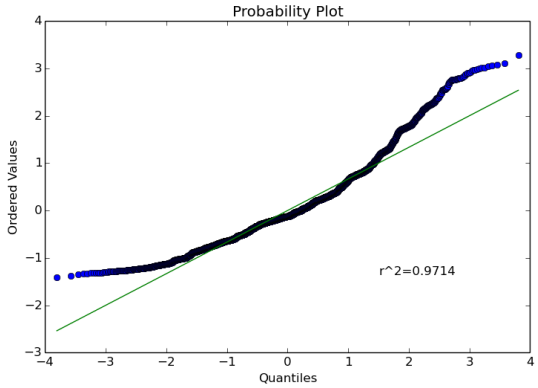
- Linear regression and least-squares
- (a little bit of) feature design
- Overfitting and regularization
- Gradient descent
- Training, validation, and testing
- Model selection

Week 1: Regression



How can we use **features** such as product properties and user demographics to make predictions about **real-valued** outcomes (e.g. star ratings)?

How can we prevent our models from **overfitting** by favouring simpler models over more complex ones?



How can we assess our decision to optimize a particular error measure, like the MSE?

Week 2: Classification

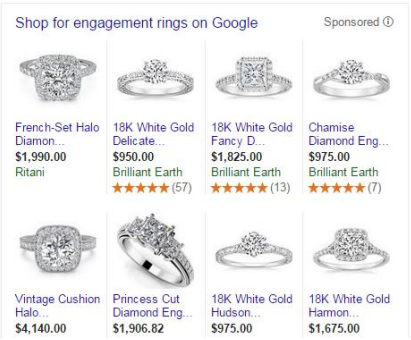
- Logistic regression
- Support Vector Machines
- Multiclass and multilabel classification
- How to evaluate classifiers, especially in “non-standard” settings

Week 2: Classification

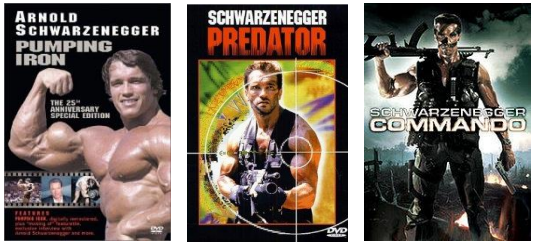
Next we adapted these ideas to **binary** or **multiclass** outputs

What animal is in this image?

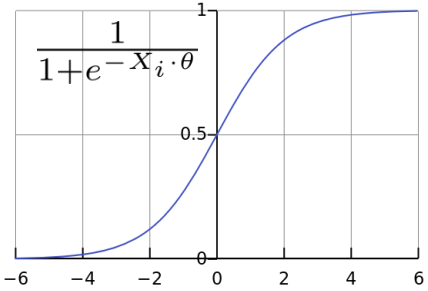
Will I **purchase** this product?



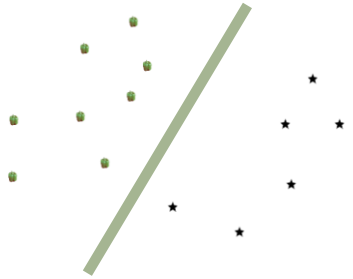
Will I **click on** this ad?



Combining features using naïve Bayes models



Logistic regression

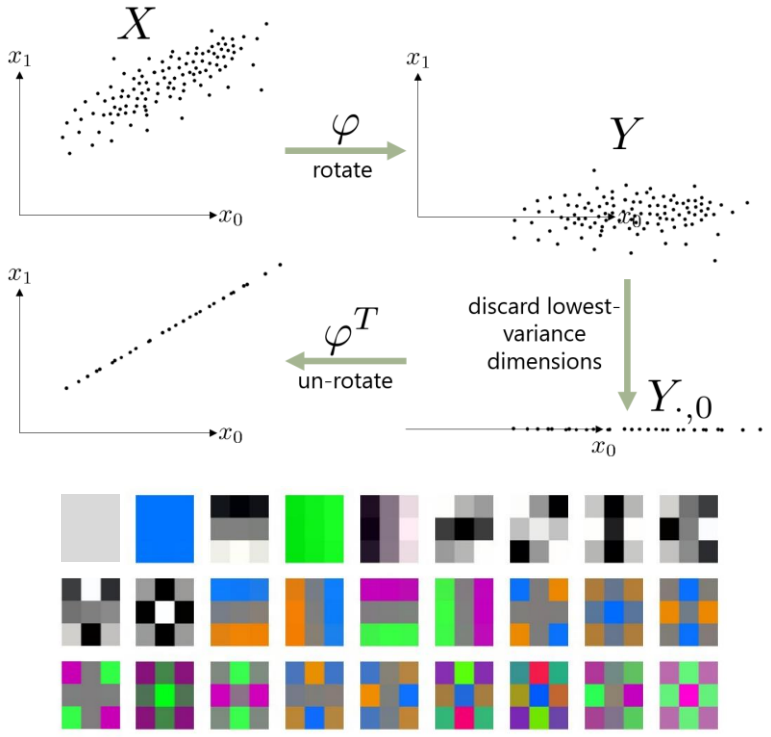


Support vector machines

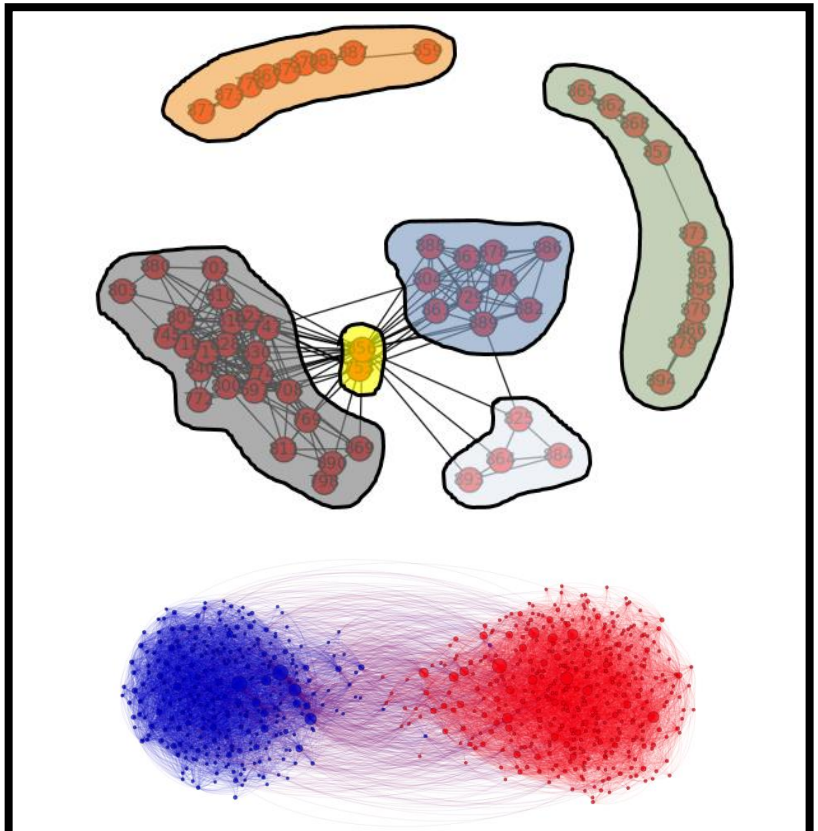
Week 3: Dimensionality Reduction

- Dimensionality reduction
- Principal component analysis
- Matrix factorization
- K-means
- Graph clustering and community detection

Week 3: Dimensionality Reduction



Principal component analysis



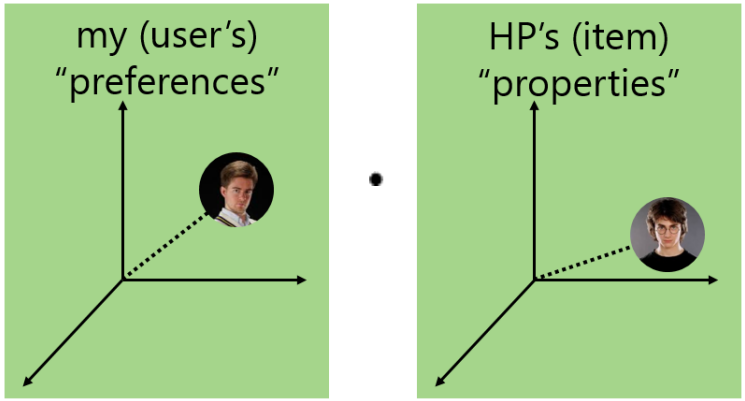
Community detection

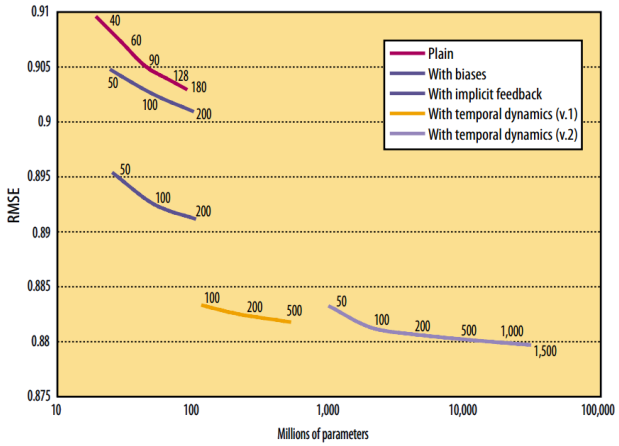
Week 4: Recommender Systems

- Latent factor models and matrix factorization (e.g. to predict star-ratings)
- Collaborative filtering (e.g. predicting and ranking likely purchases)

Week 4: Recommender Systems

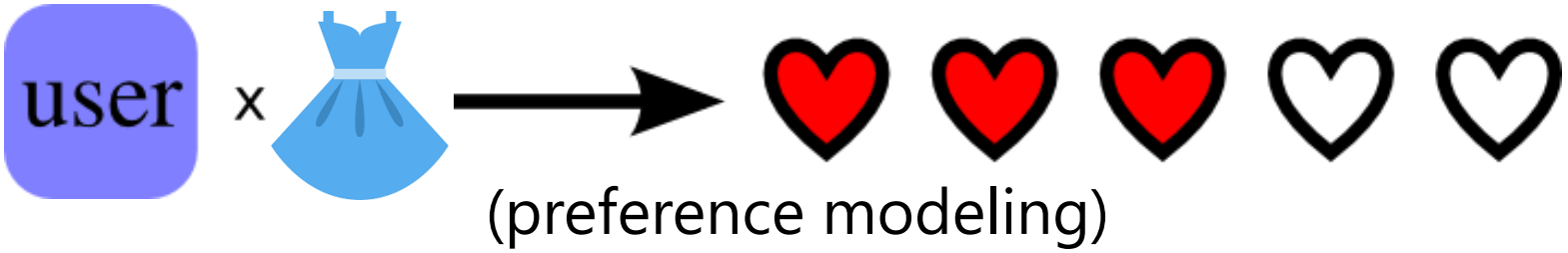
Rating distributions and the missing-not-at-random assumption



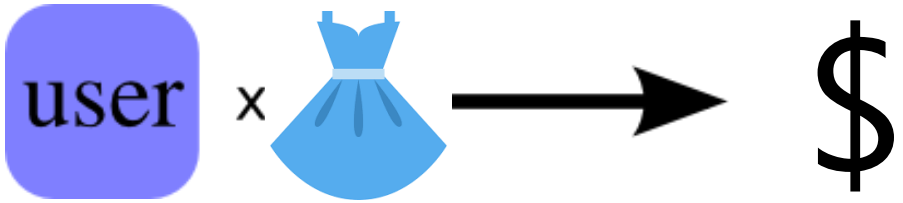


Latent-factor models

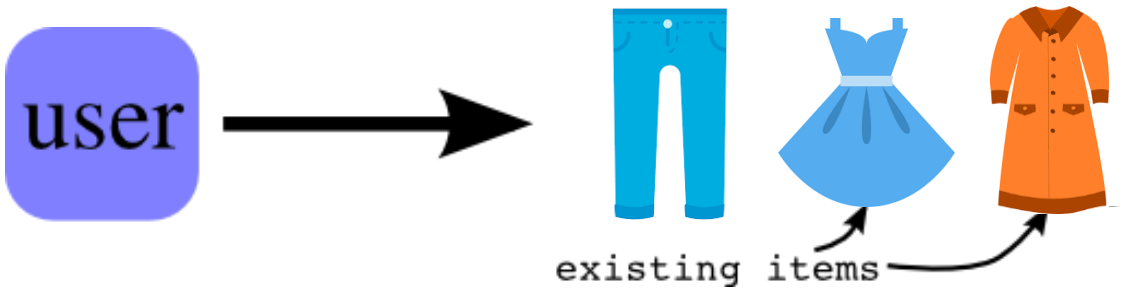
Week 4: Recommender Systems



(preference modeling)



(pricing)



(retrieval)

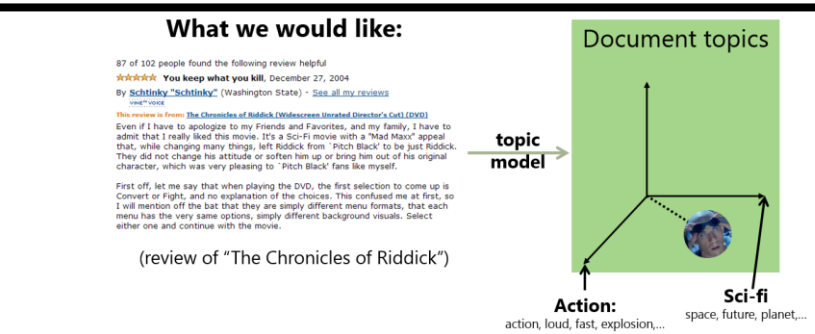
Week 5: Text Mining

- Sentiment analysis
- Bag-of-words representations
- TF-IDF
- Stopwords, stemming, and (maybe) topic models

Week 5: Text Mining

yeast and minimal red body thick light a Flavor sugar strong quad. grape over is molasses lace the low and caramel fruit Minimal start and toffee. dark plum, dark brown Actually, alcohol Dark oak, nice vanilla, has brown of a with presence. light carbonation. bready from retention. with finish. with and this and plum and head, fruit, low a Excellent raisin aroma Medium tan

Bags-of-Words



Topic models

Sentiment analysis

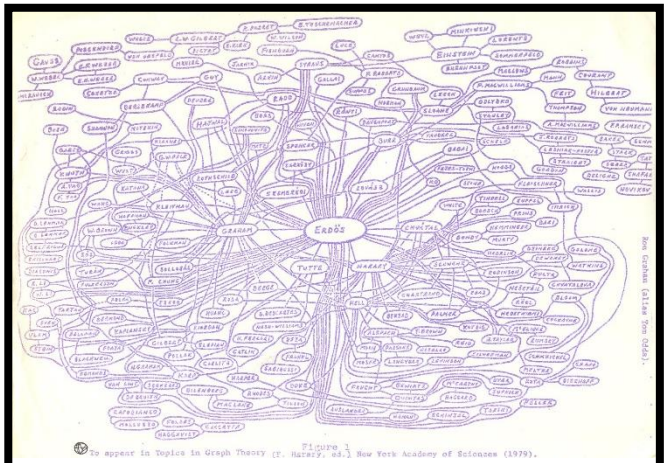
Week 6: Midterm (Nov 7)!

(More about grading etc. later)

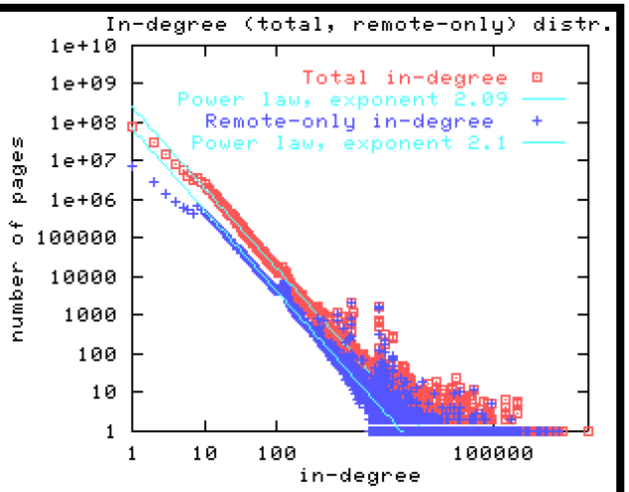
Week 7-8: Social & Information Networks

- Power-laws & small-worlds
  - Random graph models
  - Triads and “weak ties”
- Measuring importance and influence of nodes (e.g. pagerank)

Week 7-8: Social & Information Networks

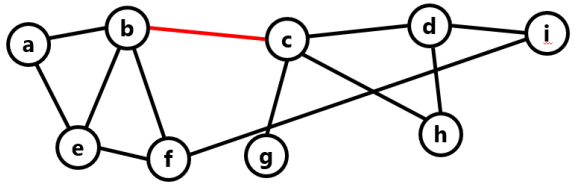


Hubs & authorities



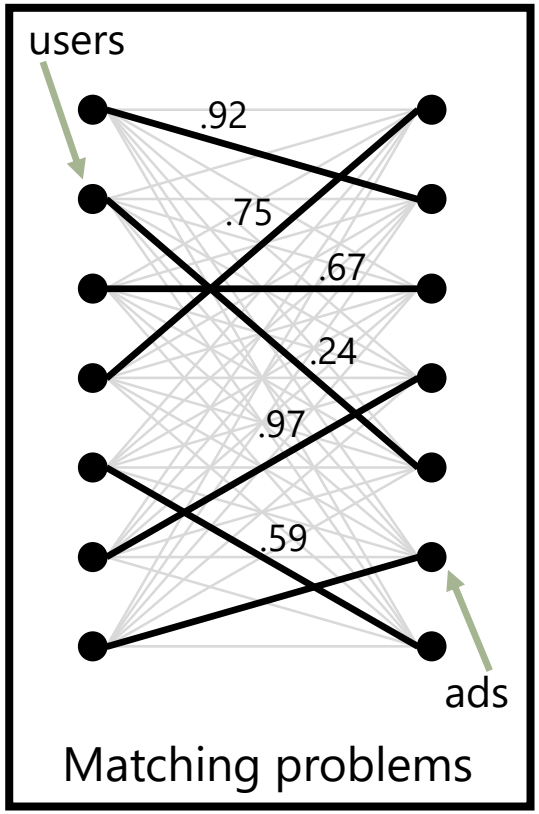
Power laws

Small-world phenomena



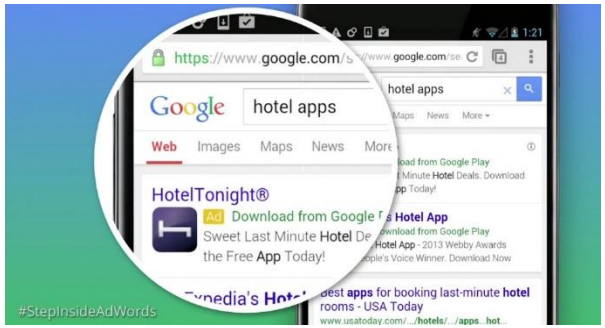
Strong & weak ties

Week 9: Advertising



Matching problems

AdWords

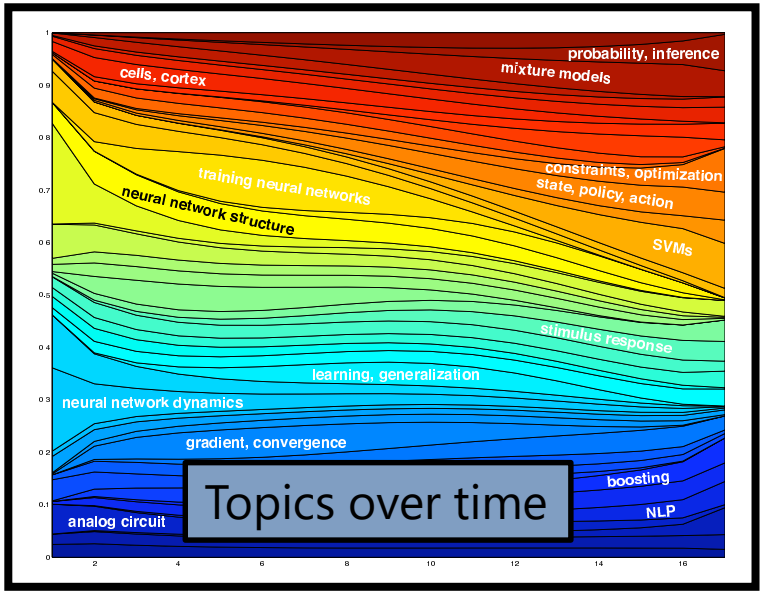


Bandit algorithms

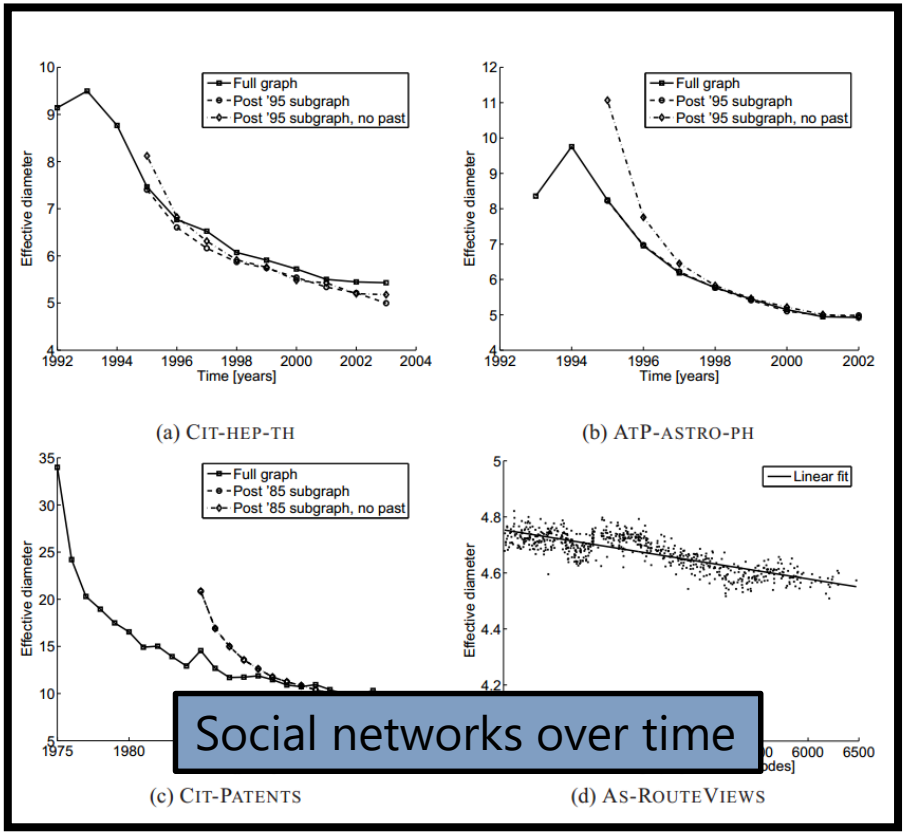
Week 10: Temporal & Sequence Data

- Sliding windows & autoregression
- Hidden Markov Models
- Temporal dynamics in recommender systems
- Temporal dynamics in text & social networks

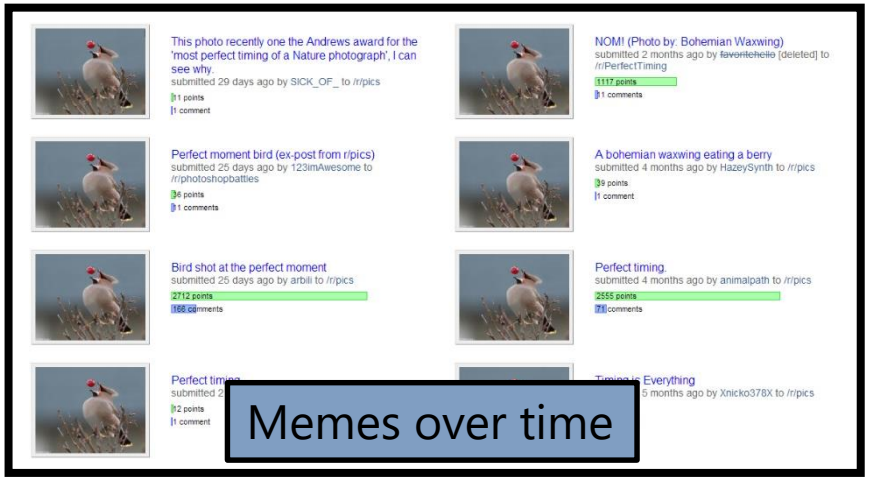
Week 10: Temporal & Sequence Data



Topics over time



Social networks over time

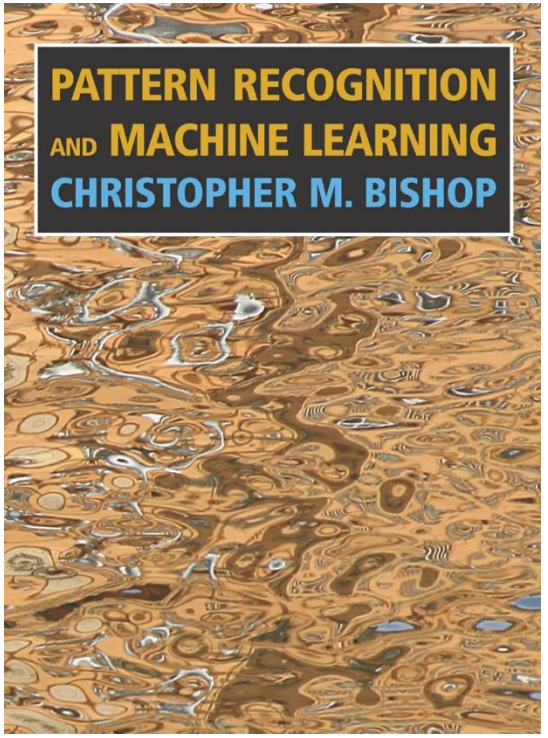


Memes over time

Reading

There is **no textbook** for this class

- I will give chapter references from *Bishop: Pattern Recognition and Machine Learning*
- I will also give references from Charles Elkan's notes (http://cseweb.ucsd.edu/classes/fa18/cse158-a/files/elkan_dm.pdf)



Evaluation

- There will be **four** homework assignments worth 8% each. Your **lowest grade** will be dropped, so that 4 homework assignments = 24%
- There will be a midterm in week 6, worth 26%
- One assignment on recommender systems (after week 5), worth 25%
- A short open-ended assignment, worth 25%

Evaluation

HW = 24%

Midterm = 26%

Assignment 1 = 25%

Assignment 2 = 25%

Actual goals:

- Understand the basics and get comfortable working with data and tools (HW)
- Comprehend the **foundational** material and the motivation behind different techniques (Midterm)
- Build something that **actually works** (Assignment 1)
- Apply your knowledge creatively (Assignment 2)

Evaluation

- Homework should be delivered by **the beginning of the <u>Monday</u> lecture in the week that it's due**
- All submissions will be made **electronically** (instructions will be in the homework spec, on the class webpage)

Evaluation

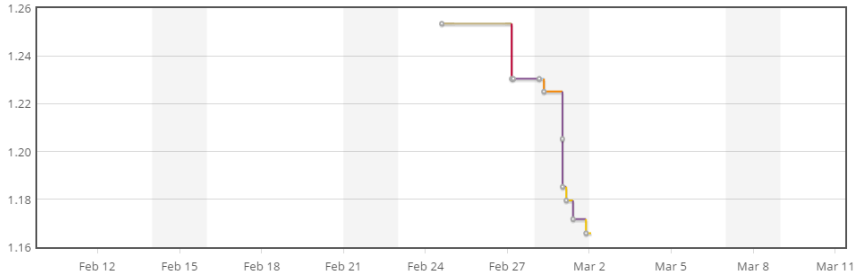
Schedule (subject to change but hopefully not):

Week 1: Hw 1 out
Week 3: Hw 1 **due**, Hw2 out
Week 5: Hw 2 **due**, Hw3 out, Assign. 1 out
Week 6: **midterm**
Week 7: Hw 3 **due**, Hw4 out, Assign. 2 out
Week 8: Assignment 1 **due**
Week 9: Hw4 **due**
Week 10: Assignment 2 **due**

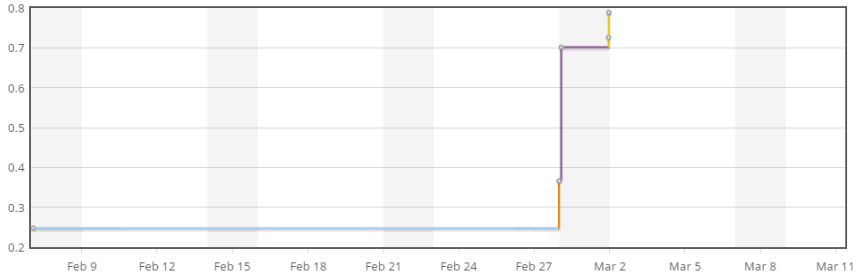
Previous assignments…

Assignment 1

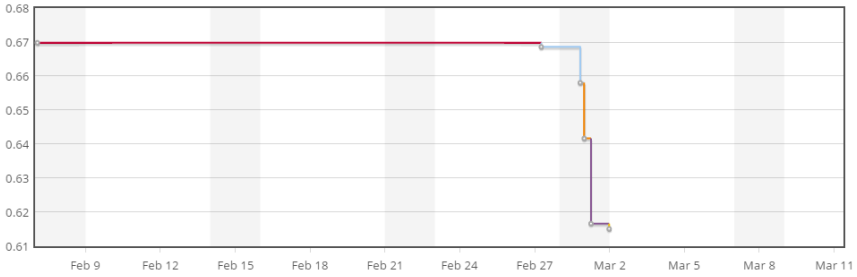
- Prediction tasks on Amazon clothing data, run as a competition on Kaggle



Rating prediction



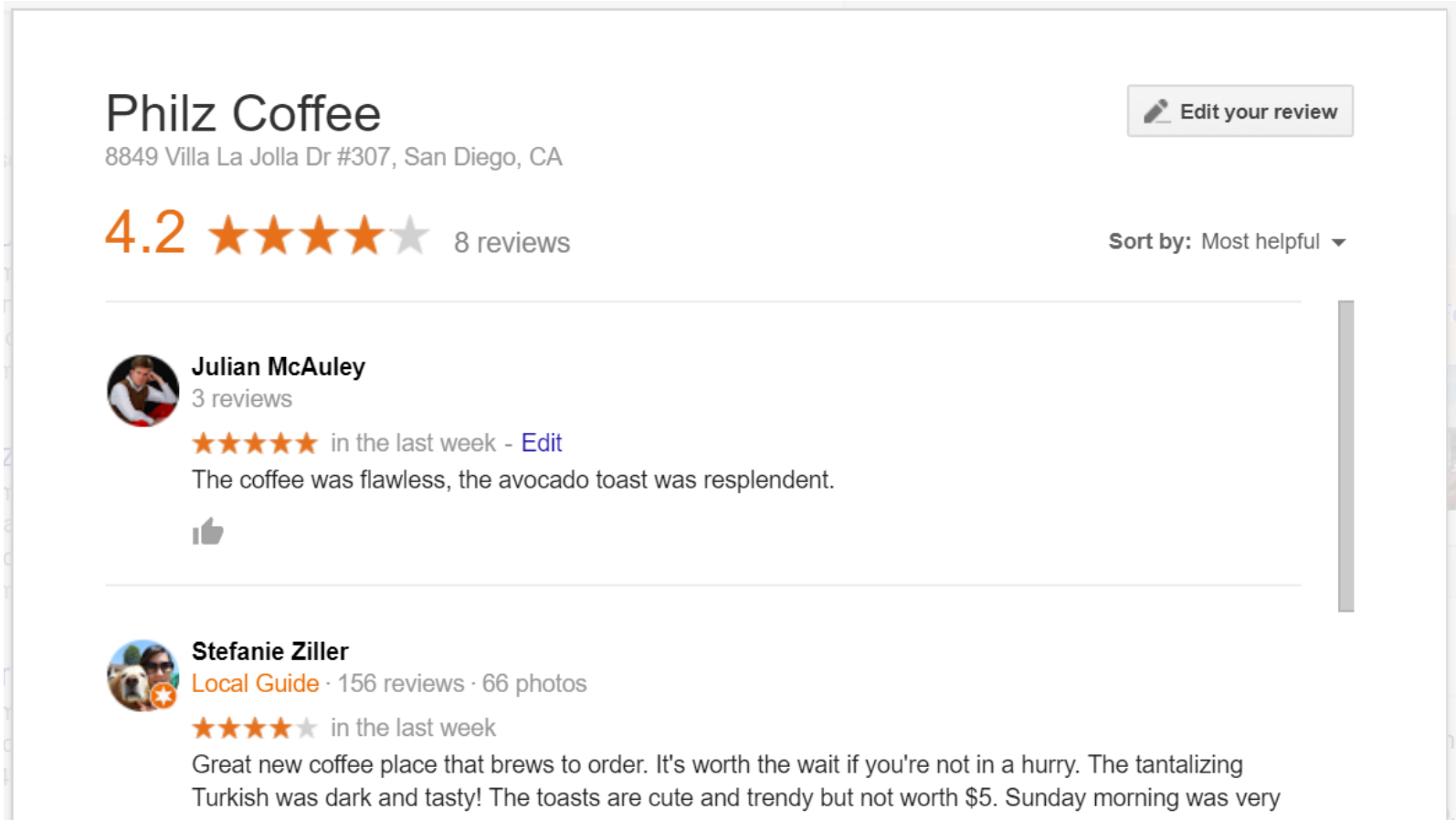
Purchase prediction



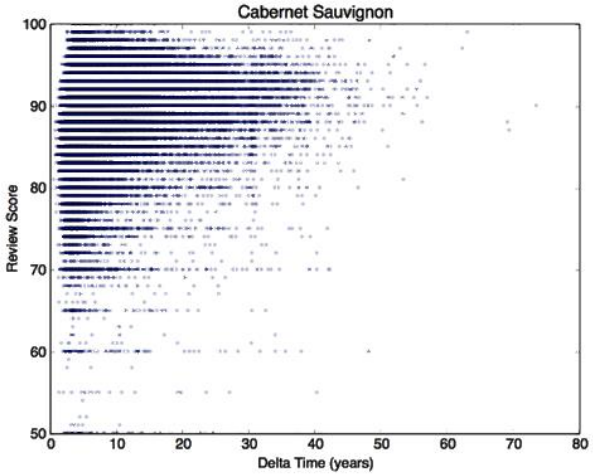
Helpfulness prediction

Assignment 1

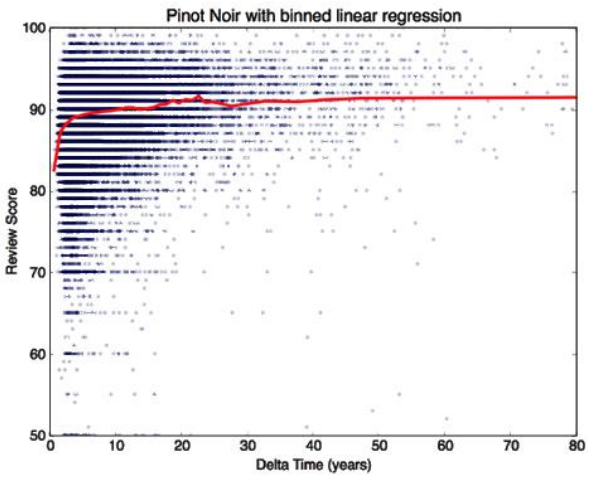
- We'll do something similar this year, but on Google Local data



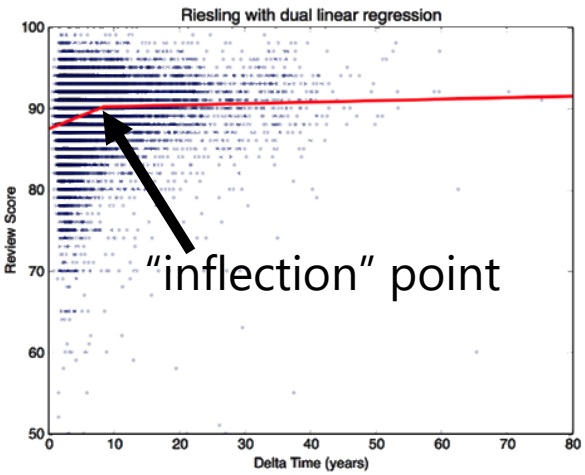
Assignment 2



Raw rating data



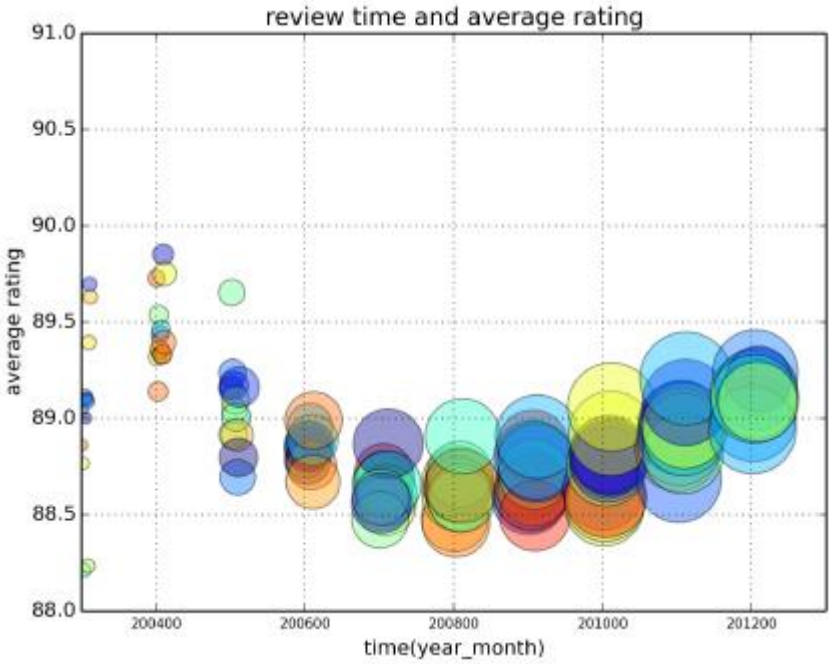
binned regression



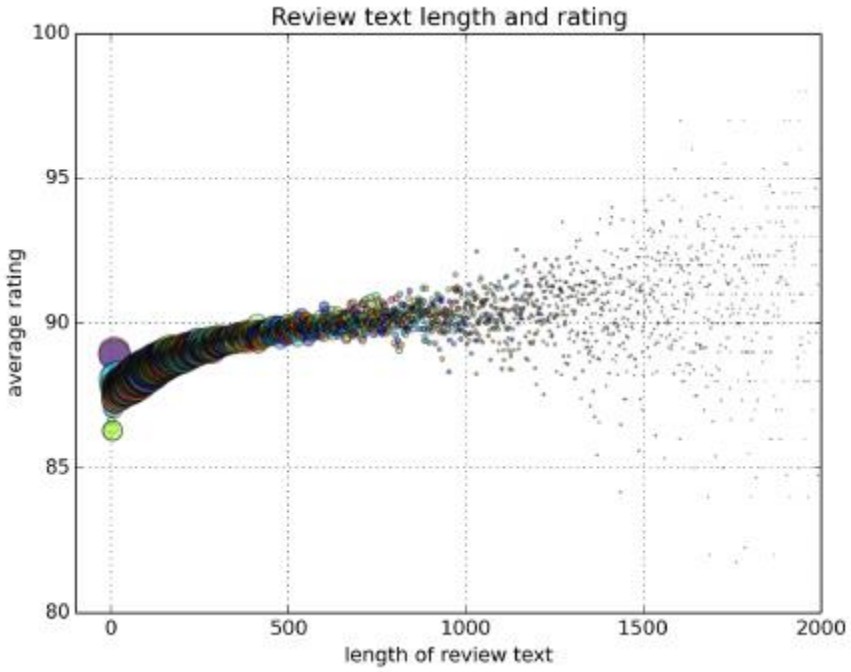
dual regression

Andrew Prudhomme – "Finding the Optimal Age of Wine"

Assignment 2



ratings vs. time

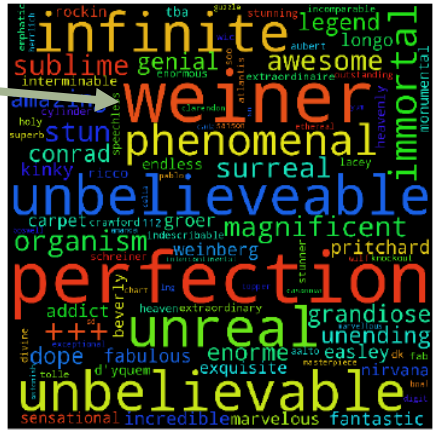


ratings vs. review length

Assignment 2

?

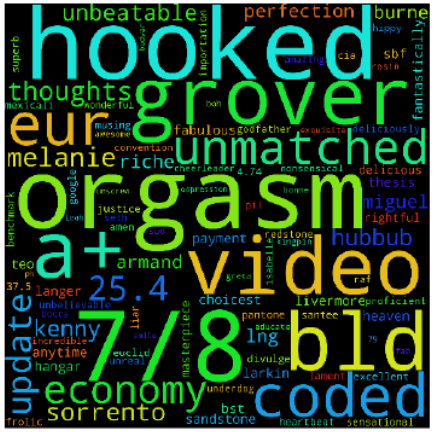
cellartracker:



positive words in wine reviews

negative words in wine reviews

RateBeer:

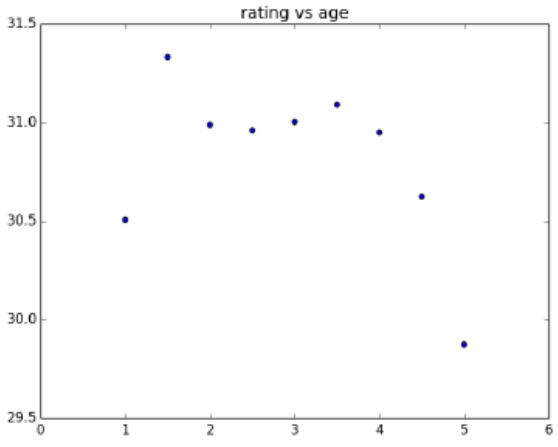


positive words in beer reviews

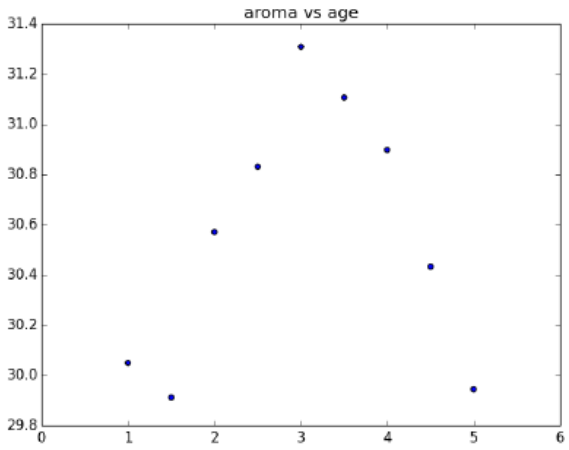
negative words in wine reviews

Ben Braun & Robert Timpe – "Text-based rating predictions from been and wine reviews"

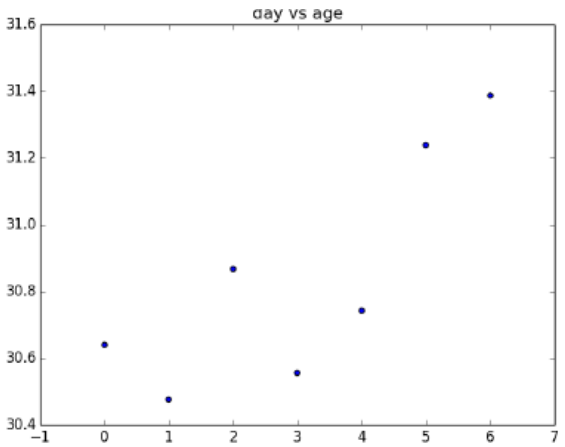
User age



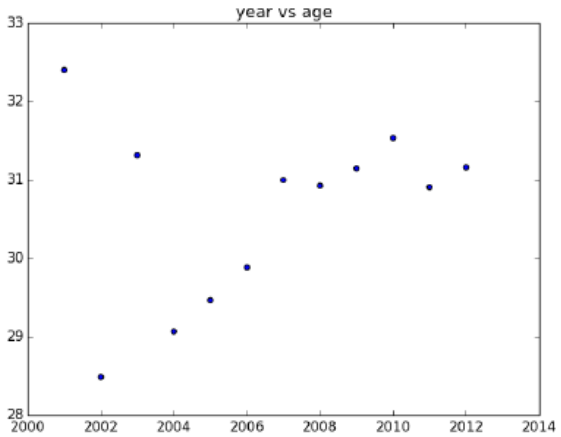
Rating vs. age



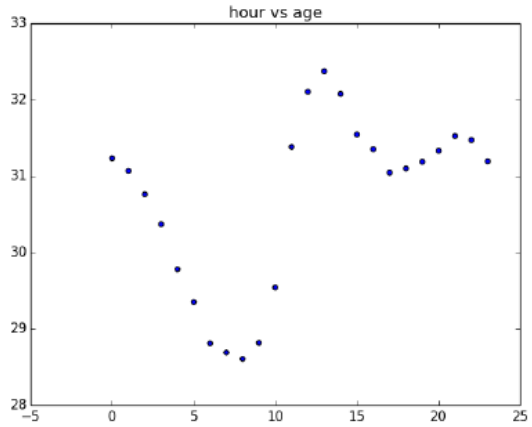
Aroma vs. age



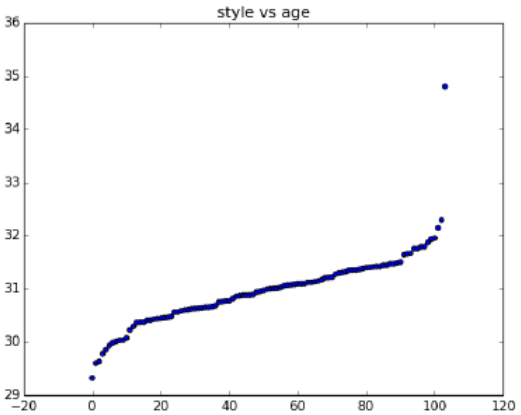
Day of week vs. age



Year vs. age



Hour of day vs. age



Category vs. age

Joseph Luttrell, Spenser Cornett

Assignment 2

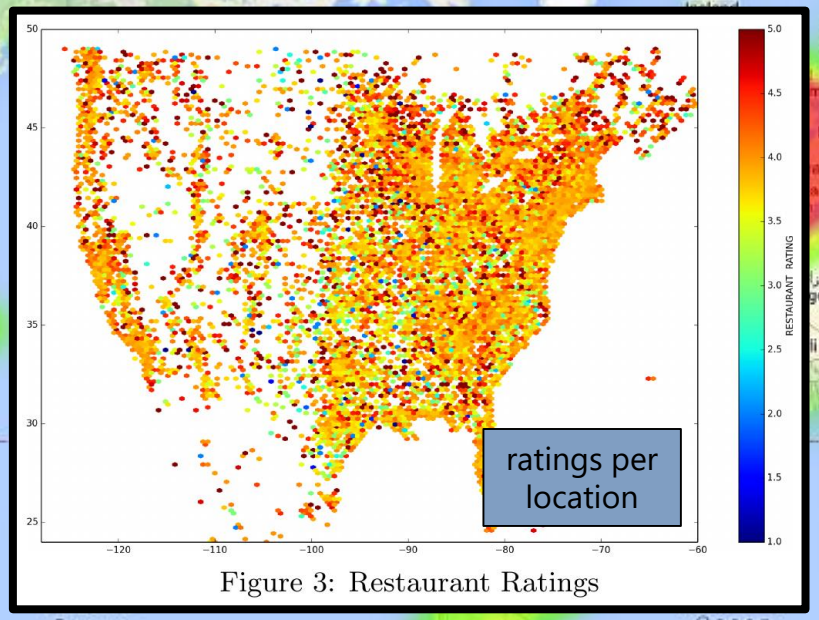


Figure 3: Restaurant Ratings

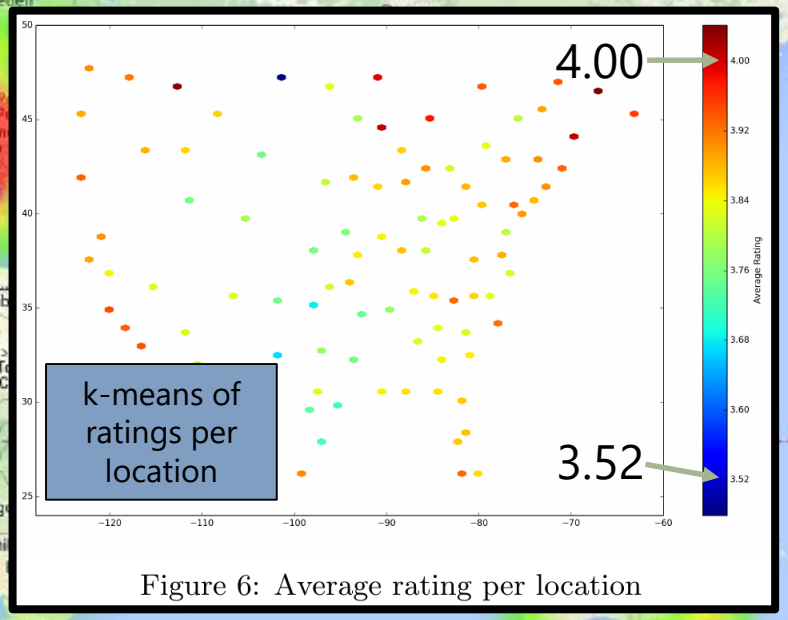
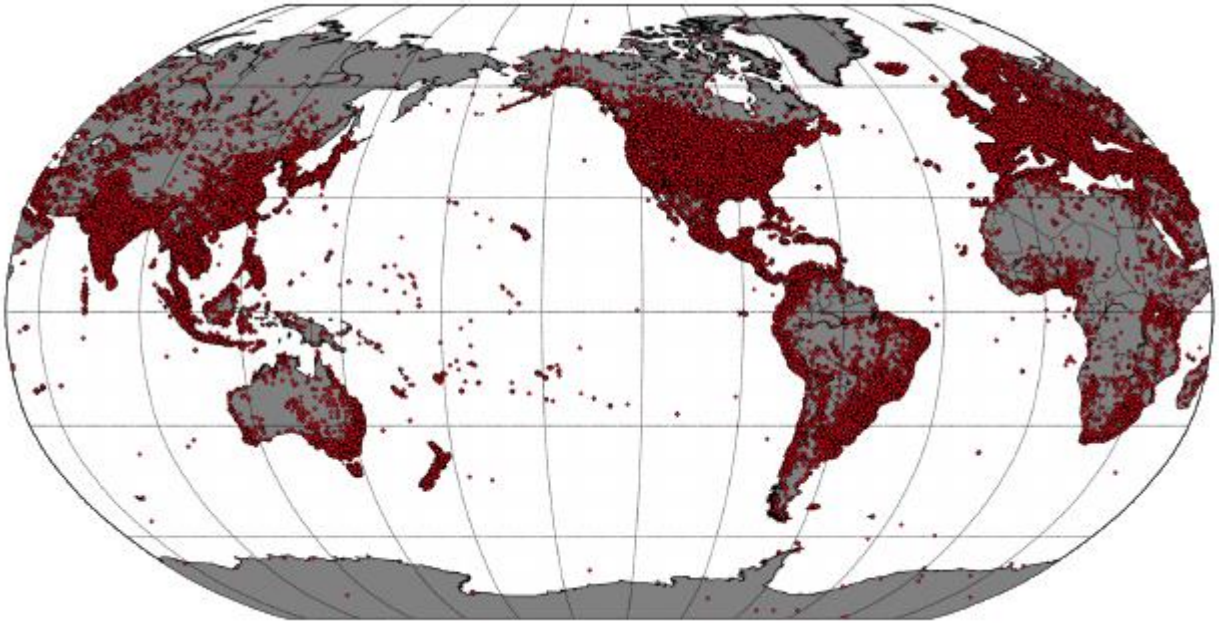


Figure 6: Average rating per location

Diego Cedillo & Idan Izhaki – "User Score for Restaurants Recommendation System"

Assignment 2



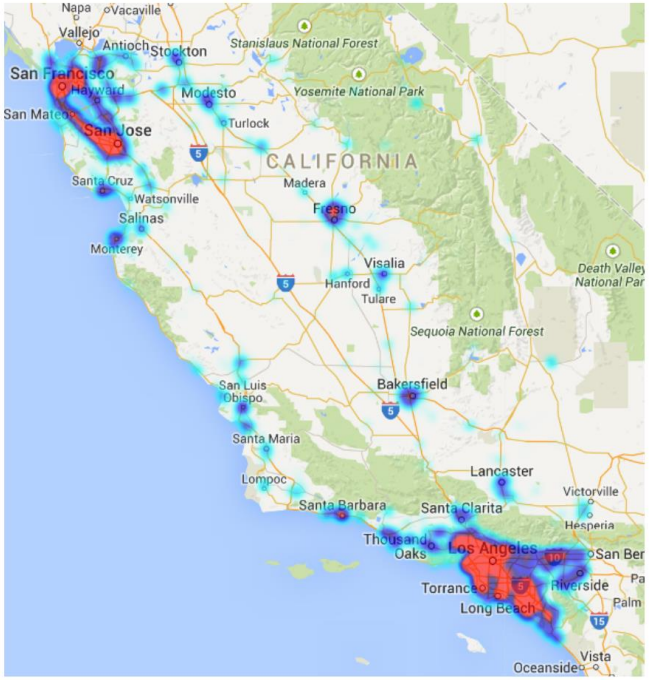
$$\widehat{r_{ui}} = \mu + b_u + b_i + (q_i + \frac{1}{|M(i)|} \sum_{n \in M(i)} |s_n|)^T p_u$$

set of geographic neighbours

impact of neighbours

Long Jin & Xinchi Gu – "Rating Prediction for Google Local Data"

Assignment 2



"Fitness"	*"Italian Restaurants"*	*"Airport & Rentals"*	*"Computer Repairs"*	*"Mexican"*
gym	food	san	computer	food
training	restaurant	francisco	store	mexican
fitness	wine	car	phone	tacos
classes	menu	airport	system	burrito
equipment	great	jose	buy	good
class	delicious	time	laptop	salsa
life	service	rental	apple	taco
great	dinner	driver	repair	chips
workout	dishes	service	problem	burritos
weight	excellent	bus	back	fish
ve	dining	shuttle	fixed	chicken
work	meal	taxi	pc	place
body	italian	trip	drive	delicious
yoga	experience	city	price	love
trainers	amazing	cab	data	fresh
people	wonderful	lax	fix	great
years	atmosphere	area	iphone	beans
feel	small	experience	screen	restaurant
instructors	decor	company	bought	asada

Topic model from Google Local business reviews

Mohit Kothari & Sandy Wiraatmadja – "Reviews and Neighbors Influence on Performance of Business"

Assignment 2

Wikispeedia navigation traces:

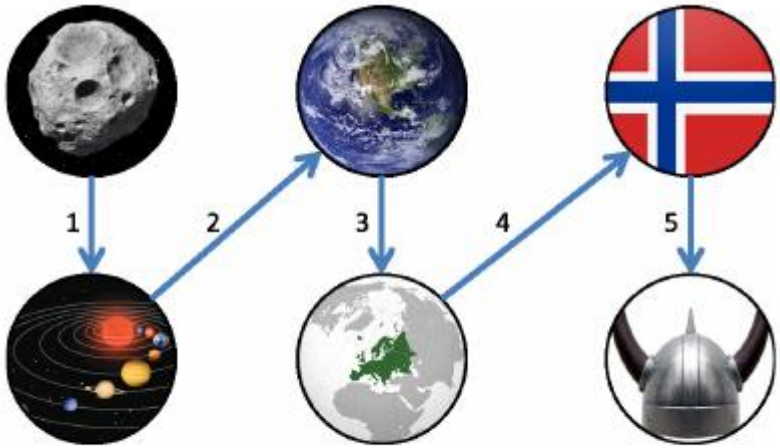


Figure 5: Graph of a complete path

	Average Click	Average Time
Finish Path	4.72	158.27
Finished Path Back	6.75	158.31
Unfinished Path	2.97	835.29
Unfinished Path Back	5.2	836.00

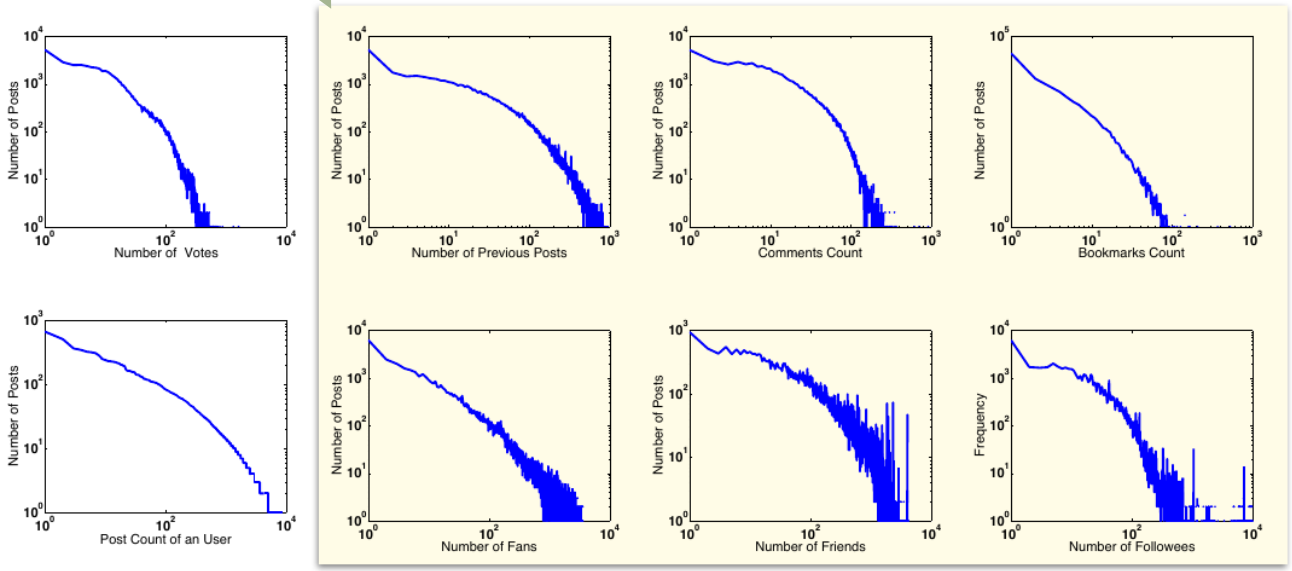
Shelby Thomas & Moein Khazraee – "Determining Topics in Link Traversals through Graph-Based Association Modeling"

Assignment 2

Images from Chictopia

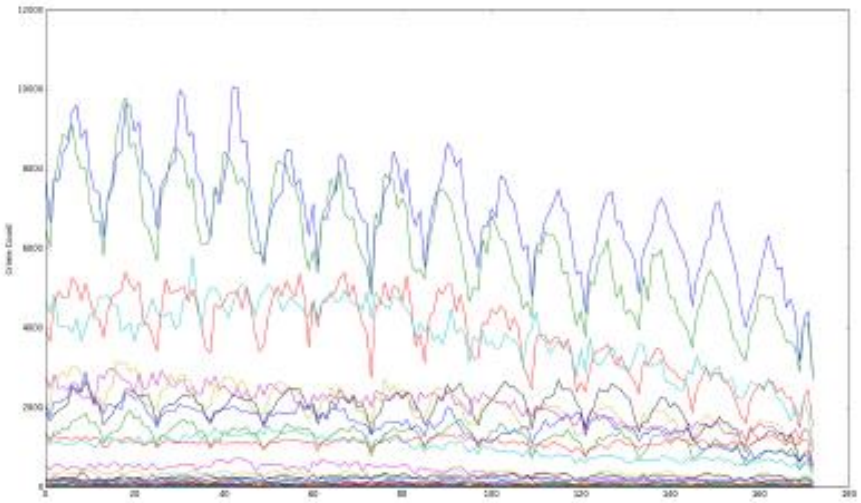
Tags	electric, every day, summer, cute, T-shirt, chic
Clothes	Chartreuse Uniqlo Socks Light Blue Uniqlo T-Shirt Bubble Gum Tie-Ups Belt White Christian Louboutin Heels
User Information	1369 friends 15 followees 2245 fans
Popularity	129 votes 62 comments 15 bookmarks

Power laws!

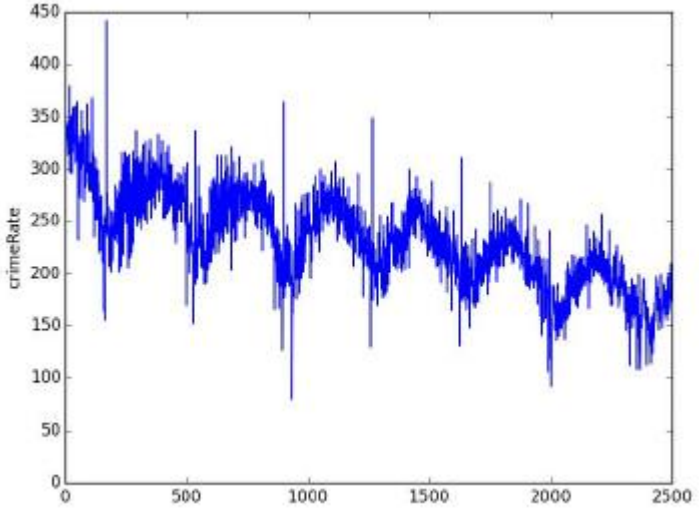


Wei-Tang Liao & Jong-Chyi Su – "Image Popularity Prediction on Social Networks"

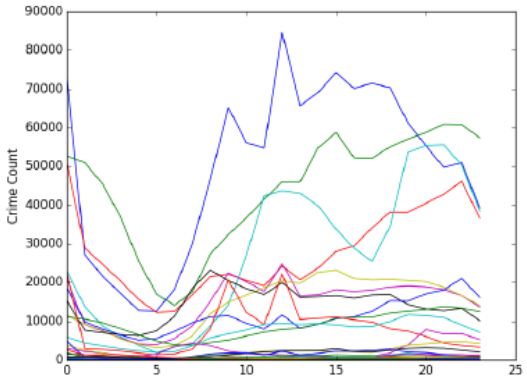
Crime (Chicago)



Over 15 years



Over 7 years

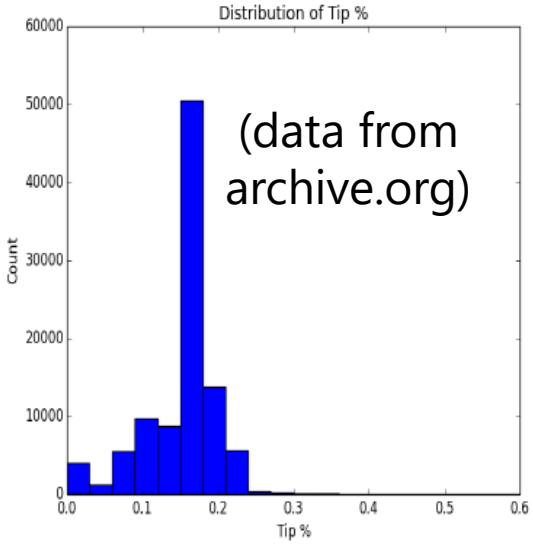


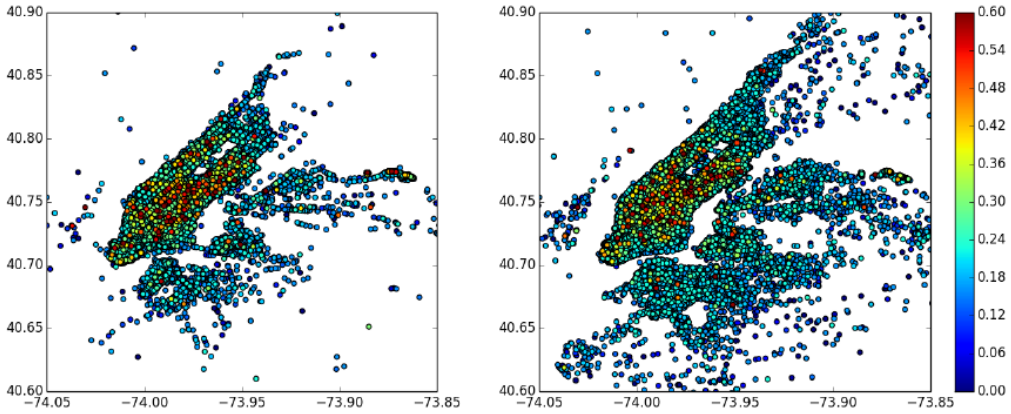
Hour of the day

Goal: to predict the number of incidents of crime on a given day

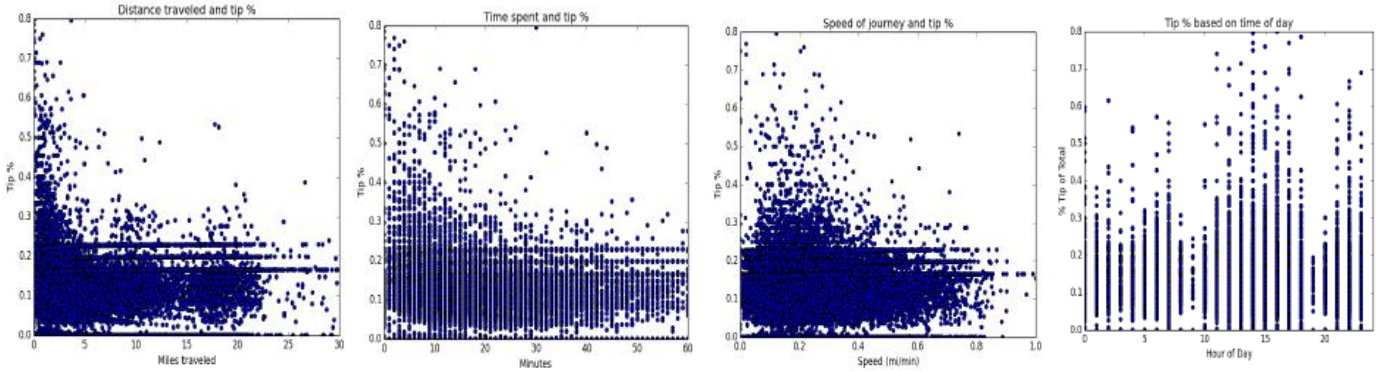
Joshua Wheeler, Nathan Moreno, Anjali Kanak

Predicting Taxi Tip-Rates in NYC





(pickup and dropoff)



Distance, time taken, speed, and time of day (also on geo)

Sahil Jain, Alvin See, Anish Shandilya

TAs

- Siddharth Sankaran Dinesh, sdinesh@ucsd.edu
- Ashin George, asg043@ucsd.edu
- Ashutosh Kumar Giri, a2giri@ucsd.edu
- Prashant Pandey, p3pandey@ucsd.edu
- Varun Syal, vsyal@ucsd.edu
- Yue Yu, yuy079@ucsd.edu
- Xiqiang Lin (Tutor) xil307@ucsd.edu
- Karl Rummler (Tutor) krummler@ucsd.edu

TAs will do most of the grading, and run office hours (in addition to my own)

Office hours

- I will hold office hours on Tuesday mornings (9:30am-1:00pm, CSE 4102)
- TA office hours will be held on Mondays and Fridays (exact time and location will be posted to the class webpage & Piazza)

Questions?

Most announcements will be posted to Piazza

https://piazza.com/ucsd/fall2018/cse158/home

please participate!