

CSE 258

Web Mining and Recommender Systems

Introduction

What is CSE 258?

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?

People You May Know



Jure Leskovec

Professor at Stanford University
9 mutual friends



Stéphane Ross

Software Engineer at Google self-driving car
3 mutual friends



Jim Tink

8 mutual friends



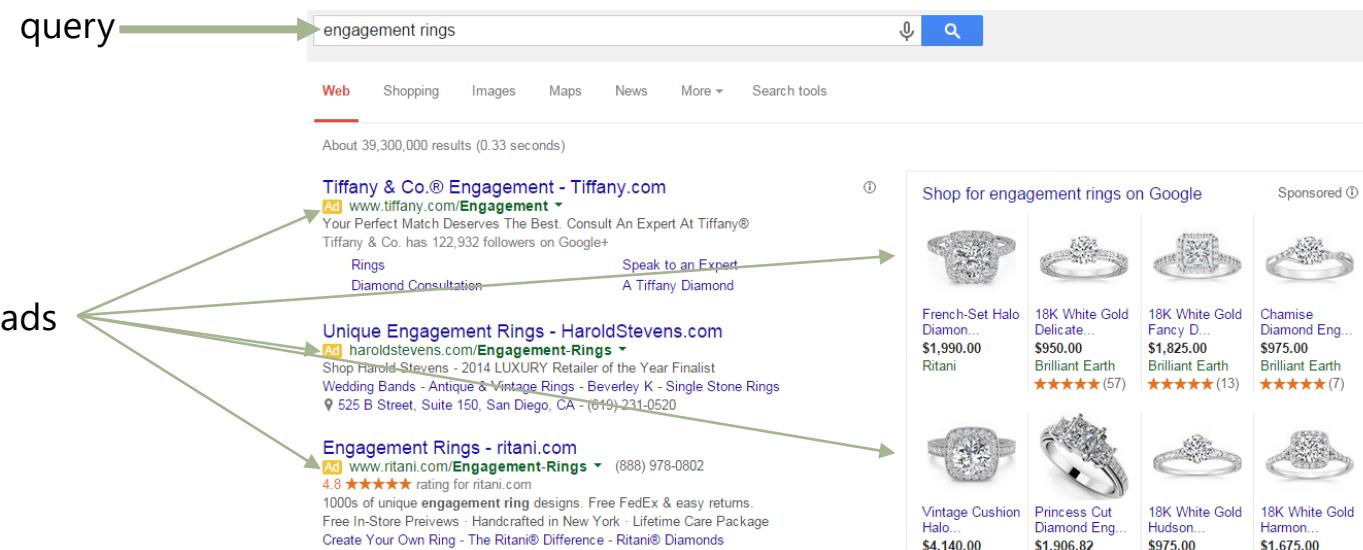
Cristian Danescu

Stanford
2 mutual friends

Examples – Advertising

Prediction: will I click on an advertisement?

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

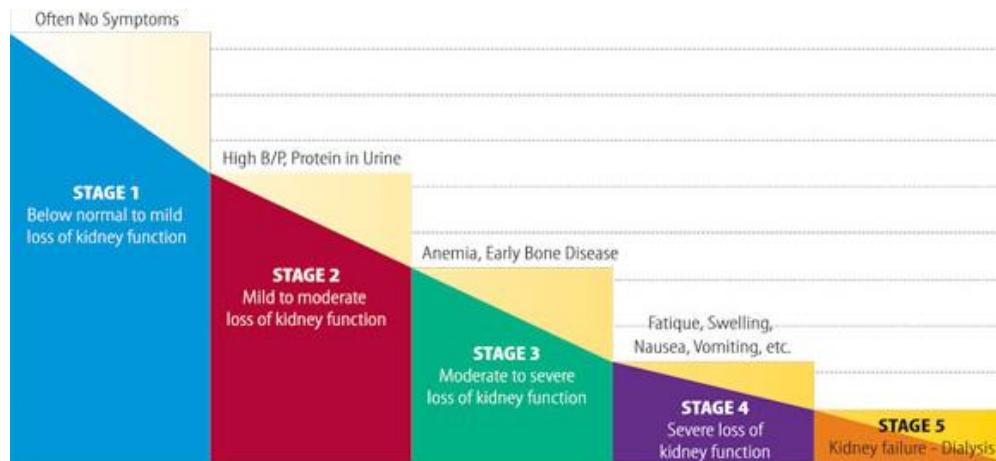


Insights: what products tend to be purchased together, and what do people purchase at different times of year?

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? **measures**
 - How will we **know** if we've modeled the data well?
 - Can actions be taken based on our findings?

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?

How wrong were our predictions on average?

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

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 258?

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 258?

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.)

CSE 258 vs. CSE 250A/B

The two most related classes are

- CSE 250A ("Principles of Artificial Intelligence: Probabilistic Reasoning and Decision-Making")
- CSE 250B ("Machine Learning")

None of these courses are prerequisites for each other!

- CSE 258 is more "hands-on" – the focus here is on applying techniques from ML to real data and predictive tasks, whereas 250A/B are focused on developing a more rigorous understanding of the underlying mathematical concepts

CSE 258 vs. CSE 158

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=3004&v=1>

CSE258:

<http://podcasts.ucsd.edu/podcasts/default.aspx?PodcastId=3003&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 258

Web Mining and Recommender Systems

Course outline

Course webpage

The course webpage is available here:

<http://cseweb.ucsd.edu/classes/wi17/cse258-a/>

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

Course webpage

(winter's course webpage is here):

<http://cseweb.ucsd.edu/classes/fa15/cse255-a/>

This quarter's content will be (roughly) similar (though the weighting of assignments/midterms etc. is different)

Course outline

This course is in two parts:

1. **Methods** (weeks 1-4):

- 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... visualization/crawling/online advertising etc.

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

Product Details

Genres	Science Fiction, Action, Horror
Director	David Twohy
Starring	Vin Diesel, Radha Mitchell
Supporting actors	Cole Hauser, Keith David, Lewis Fitz-Gerald, Claudia Black, Rhiana Grangaard, Angela Moore, Peter Chiang, Ken Watanabe
Studio	NBC Universal
MPAA rating	R (Restricted)
Captions and subtitles	English Details ▾
Rental rights	24 hour viewing period. Details ▾
Purchase rights	Stream instantly and download to 2 locations Details ▾
Format	Amazon Instant Video (streaming online video and digital download)

A. Phillips

Reviewer ranking: #17,230,554

90% helpful
votes received on reviews
(151 of 167)

ABOUT ME
Enjoy the reviews...

ACTIVITIES
Reviews (16)
Public Wish List (2)
Listmania Lists (2)
Tagged Items (1)



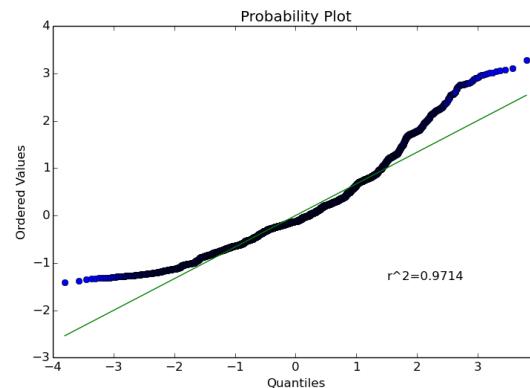
HipCzech
Aficionado
Male, from Texas

Profile Page

Member Since:	Jul 12, 2014	HipCzech was last seen:
Points:	175	Today at 12:19 AM
Beers:	108	
Places:	6	
Posts:	0	
Likes Received:	0	
Trading:	0% 0	

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 2/3: Classification

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

Week 2/3: Classification

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



Pitch Black - Unrated Director's Cut R

★★★★★ (777) IMDb 7.1/10

Watch Trailer

When their ship crash-lands on a remote planet, escaped convict Riddick (Vin Diesel) isn't the only one lurking in the shadows, waiting to attack in the dark.

See More

Starring: Vin Diesel, Radha Mitchell
Runtime: 1 hour, 53 minutes
Available to watch on [supported devices](#).

UNRATED

Shop for engagement rings on Google Sponsored

French-Set Halo Diamond Ring \$1,990.00 Ritani

18K White Gold Delicate Diamond Ring \$950.00 Brilliant Earth

18K White Gold Fancy Diamond Ring \$1,825.00 Brilliant Earth

Chamise Diamond Engagement Ring \$975.00 Brilliant Earth

Vintage Cushion Halo Diamond Ring \$4,140.00

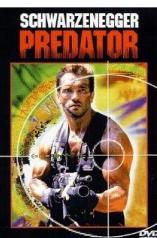
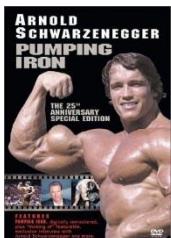
Princess Cut Diamond Engagement Ring \$1,906.82 Hudson

18K White Gold Harmon Diamond Ring \$975.00

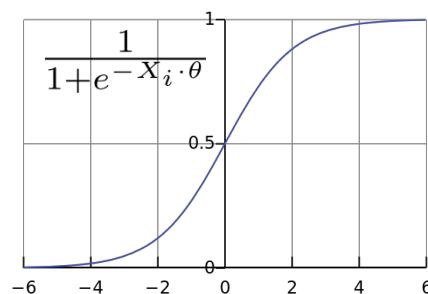
18K White Gold Harmon Diamond Ring \$1,675.00

What animal is in this image?

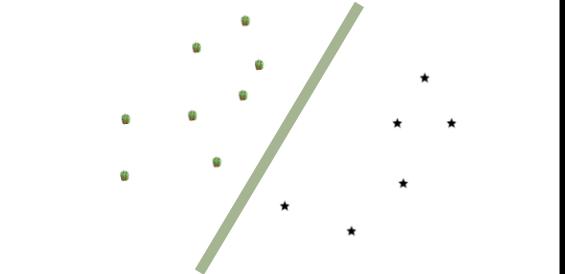
Will I **purchase** this product? Will I **click on** this ad?



Combining features using naïve Bayes models



Logistic regression

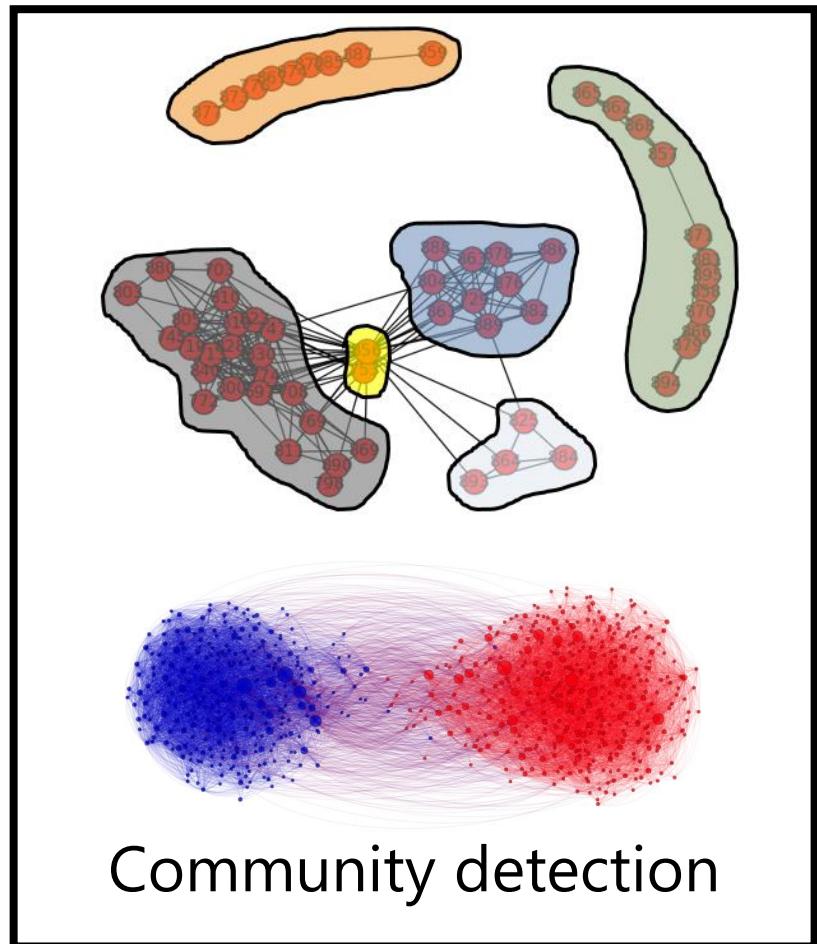
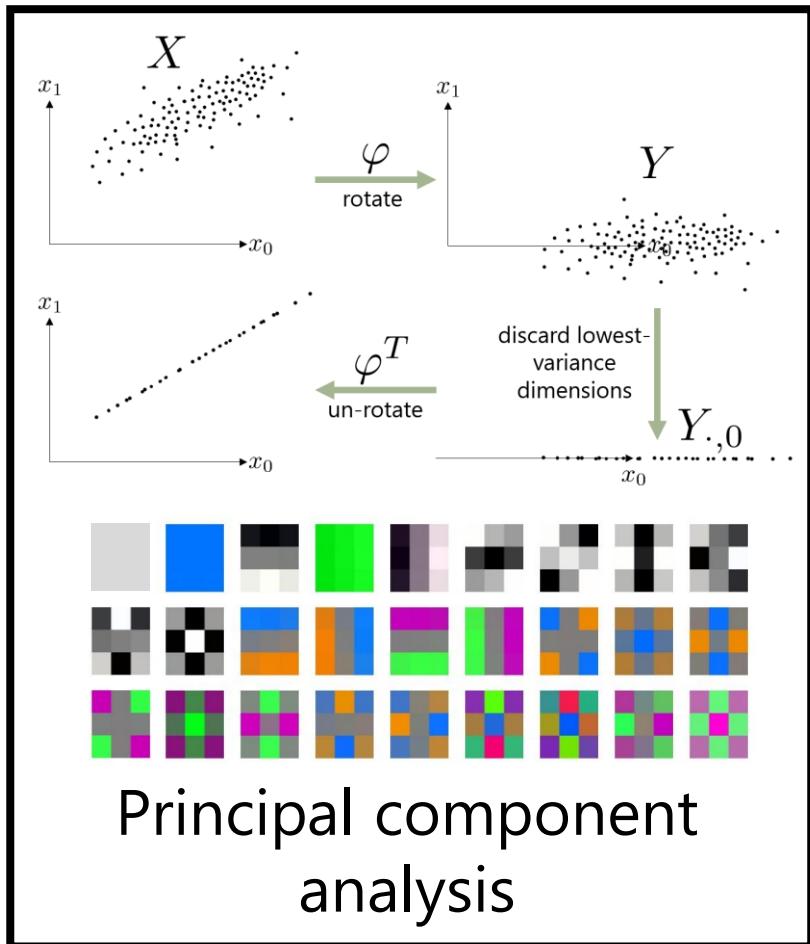


Support vector machines

Week 3/4: Dimensionality Reduction

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

Week 3/4: Dimensionality Reduction



Week 4/5: Recommender Systems

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

Week 4/5: Recommender Systems

Island WFM1000SCDLI Diamonds
old Case Black Leather Men's Watch

Men's 18K Gold Rolex Yachtmaster II Model # 116688
by Rolex

\$34,880.00

Show only Rolex items

★★★★★ 94

0 items

3.7 out of 5 stars

5 star	47
4 star	13
3 star	13
2 star	2
1 star	19

Now when I take him for a walk I know I am impressing people even more than I EVER did when I merely walked my monkey while wearing this wonderful watch.

Dr. Space | 11 reviewers made a similar statement

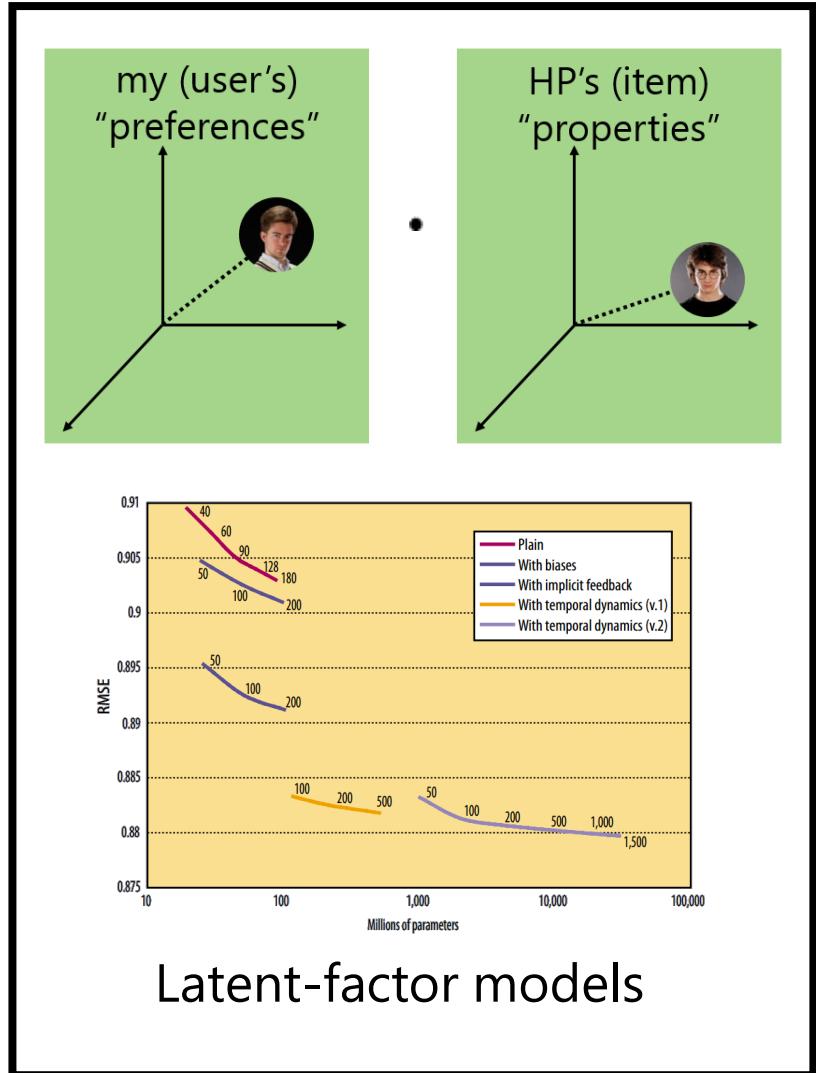
You also placed a review on a watch you don't own in order to spew.

A. Wright | 3 reviewers made a similar statement

See all 94 reviews

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

Rating Probability vs Rating Value for different rating levels.



Week 5/6: Text Mining

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

Week 5/6: 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

What we would like:

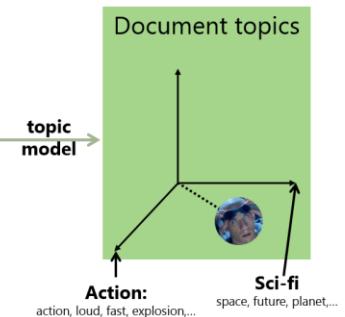
87 of 102 people found the following review helpful

★★★★★ You keep what you kill, December 27, 2004
By [Schtlinky "Schtlinky"](#) (Washington State) - [See all my reviews](#)

This review is from: **The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)**
Even if I have to apologize to my Friends and Favorites, and my family, I admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Max" that, while changing many things, left Riddick from "Pitch Black" to be just. They did not change his attitude or soften him up or bring him out of his character, which was very pleasing to "Pitch Black" fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")



Topic models



Sentiment analysis

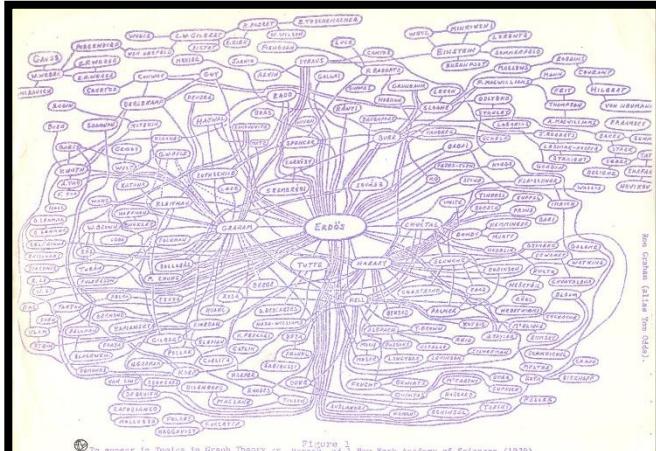
Week 6: Midterm (Feb 15)!

(More about grading etc. later)

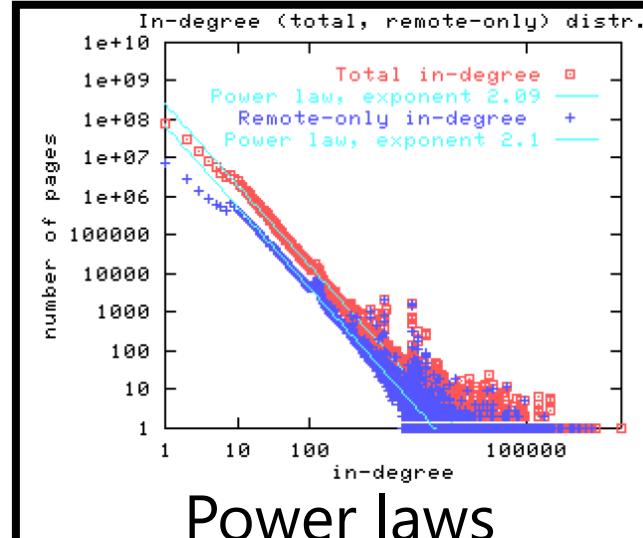
Week 8: Social & Information Networks

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

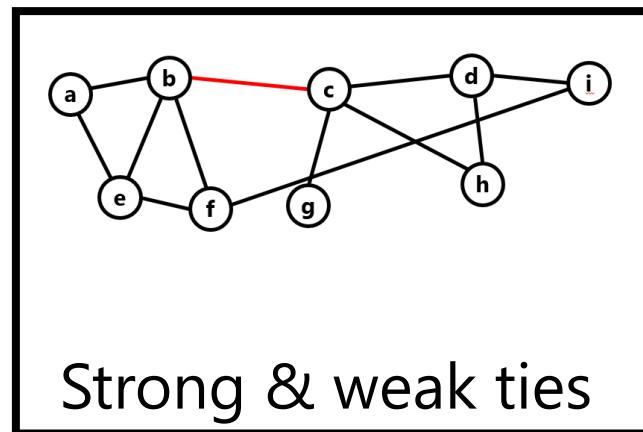
Week 8: Social & Information Networks



Hubs & authorities

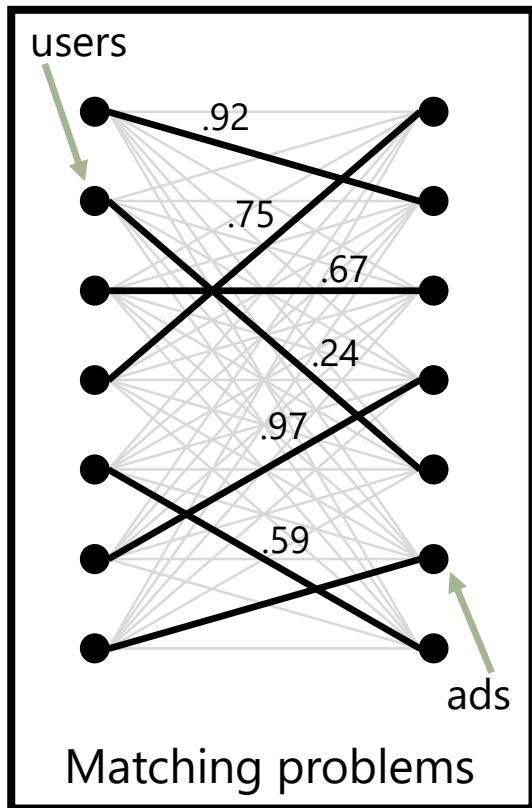


Small-world phenomena



Strong & weak ties

Week 9: Something else (advertising?)

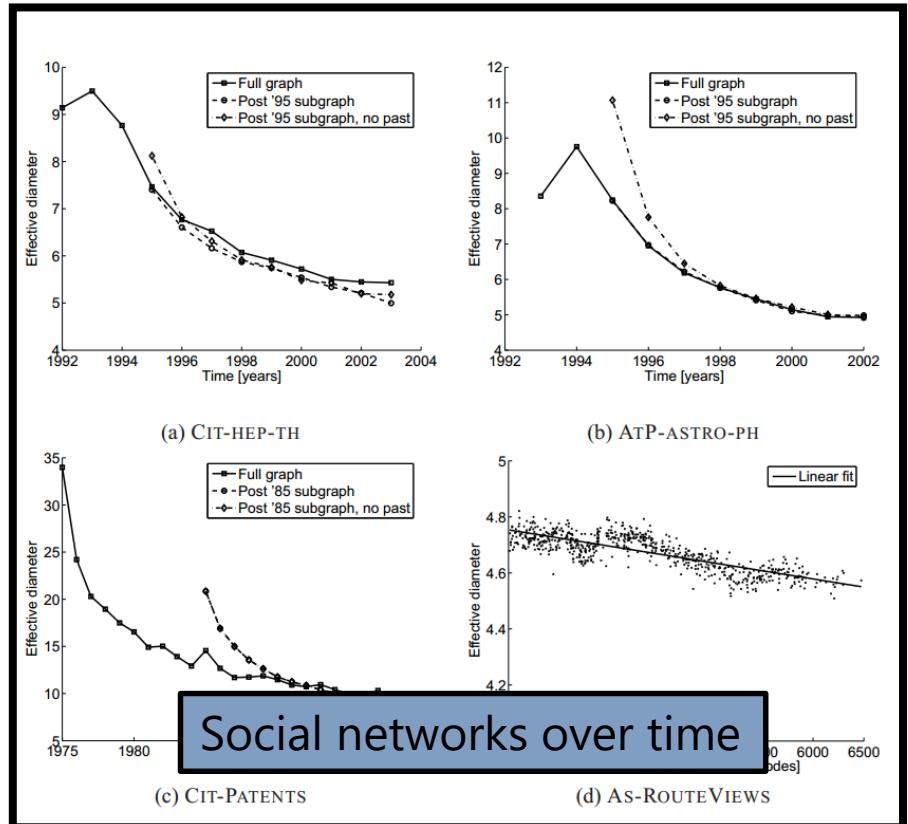
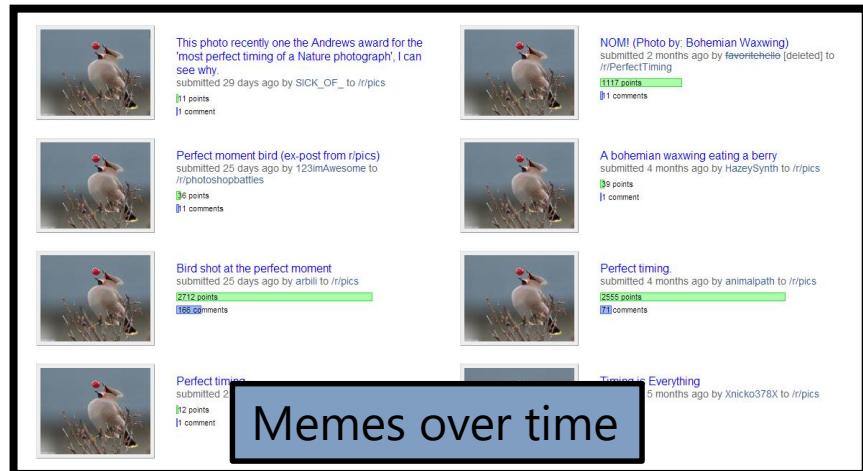
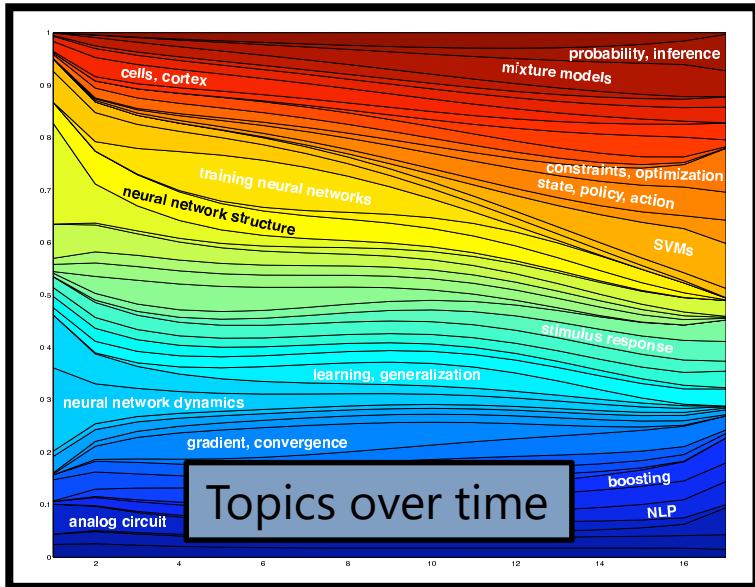


Bandit algorithms

Week 10: Temporal & Sequence Data

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

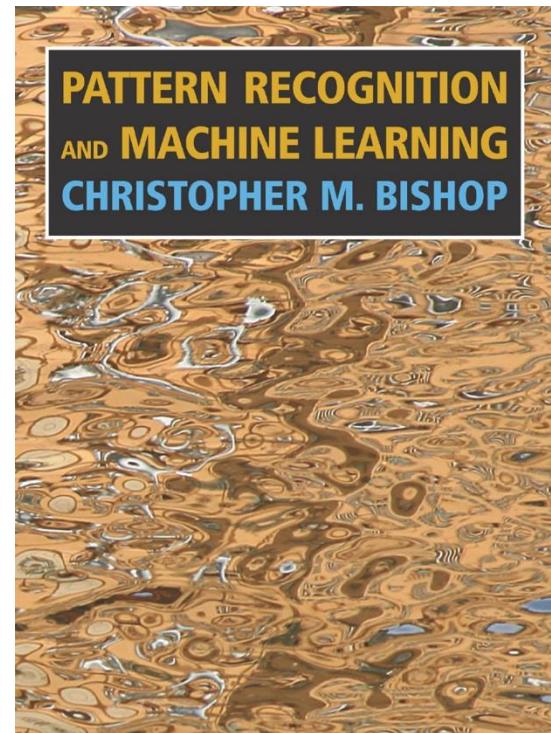
Week 10: Temporal & Sequence Data



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/wi17/cse258-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 Monday 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**

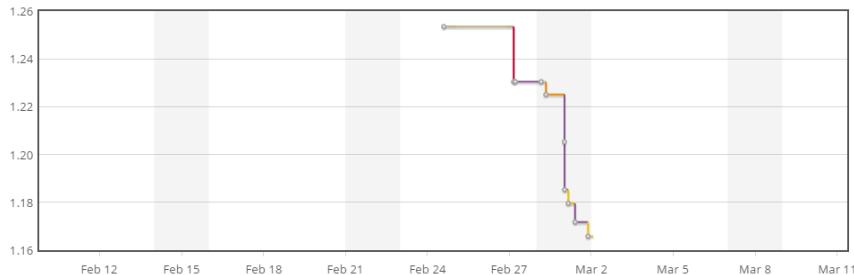
Week 9: Hw4 **due**

Week 10: Assignment 2 **due**

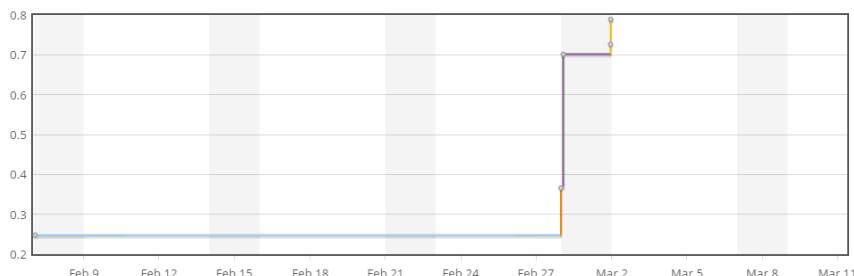
Previous assignments...

Assignment 1

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



Rating prediction



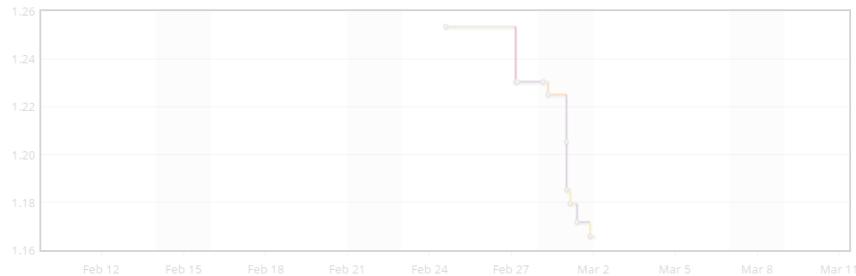
Purchase prediction



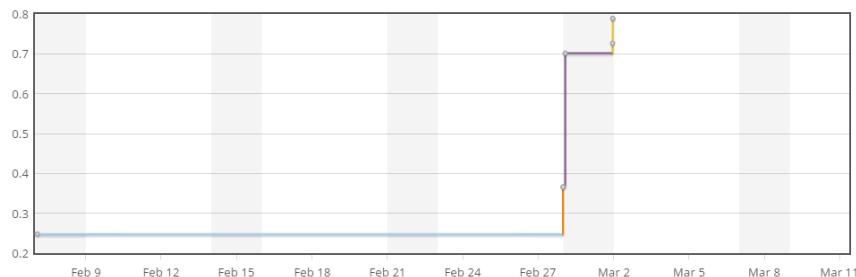
Helpfulness prediction

Assignment 1

- We'll definitely do this again, but with different data and possibly different tasks



Rating prediction

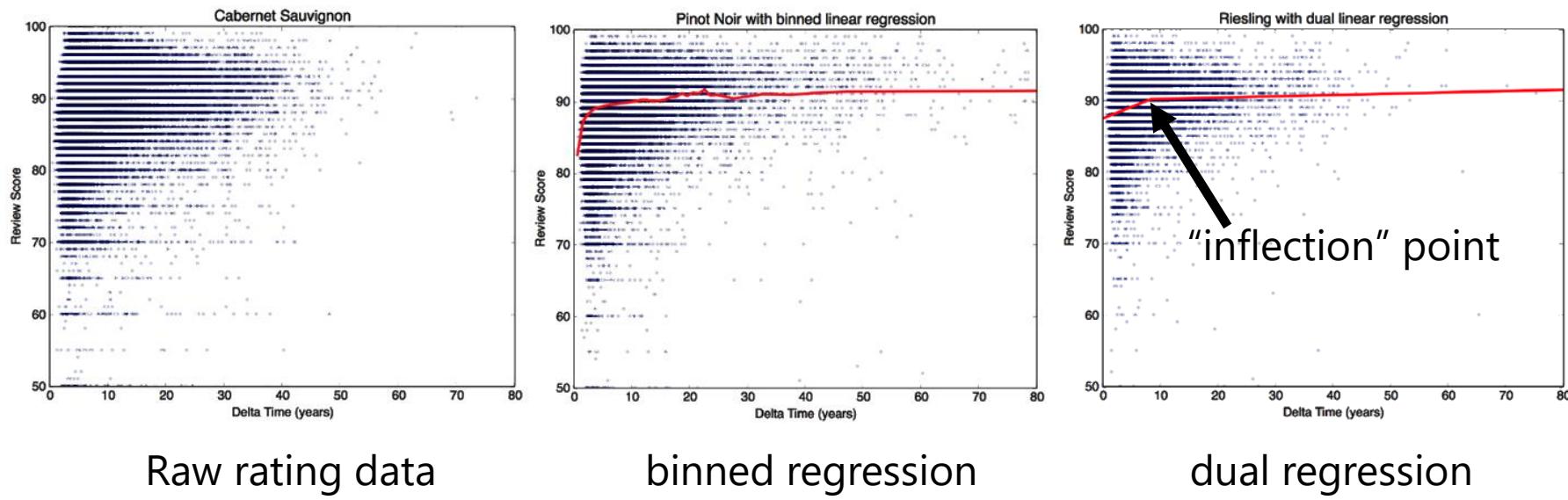


Purchase prediction

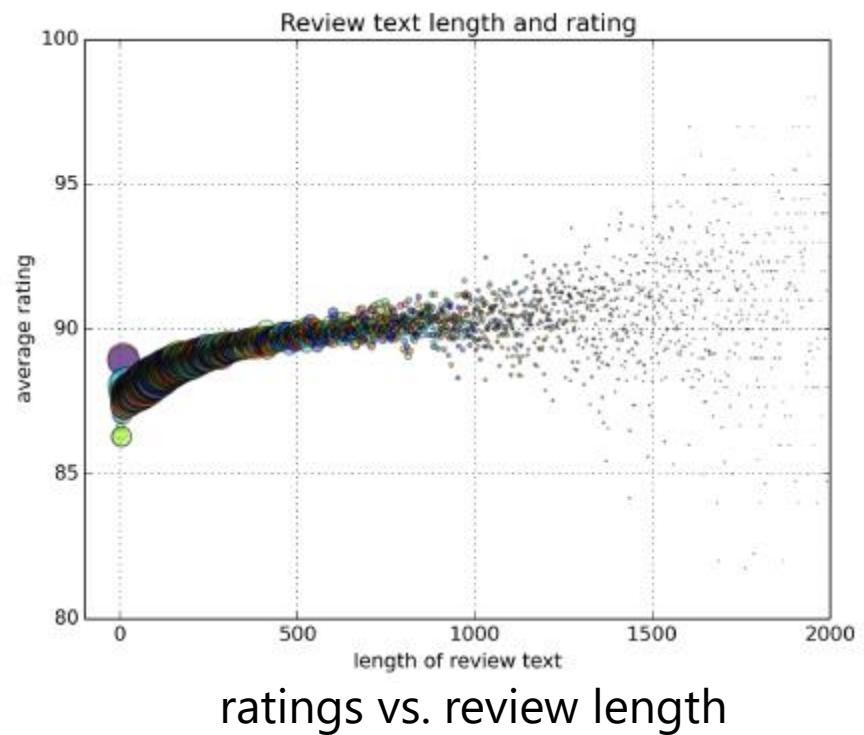
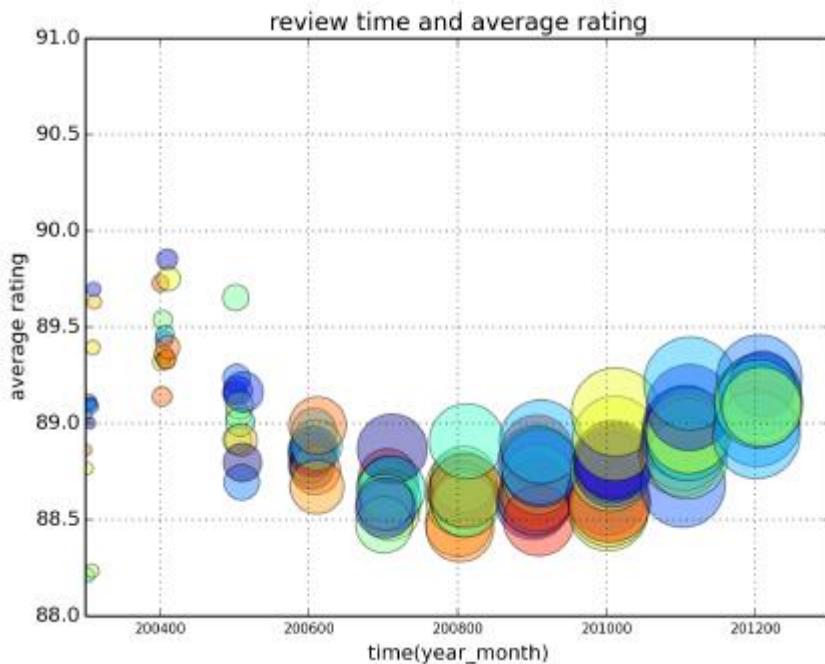


Helpfulness prediction

Assignment 2



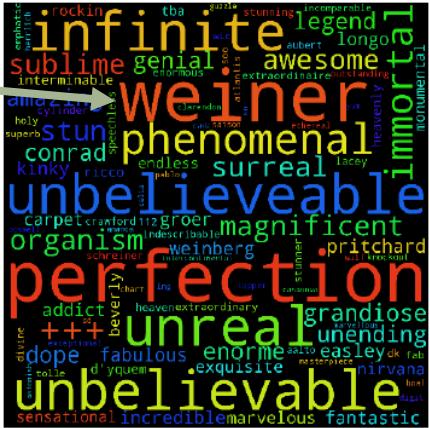
Assignment 2



Assignment 2

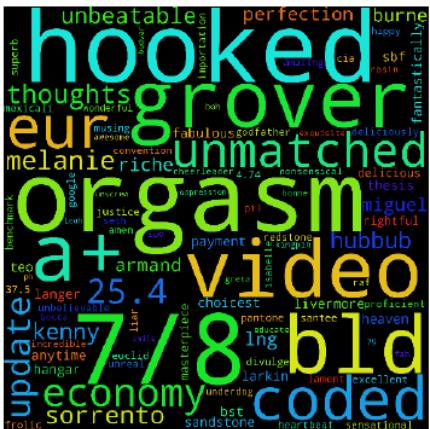
?

cellartracker:



positive words in wine reviews

RateBeer:



positive words in beer reviews

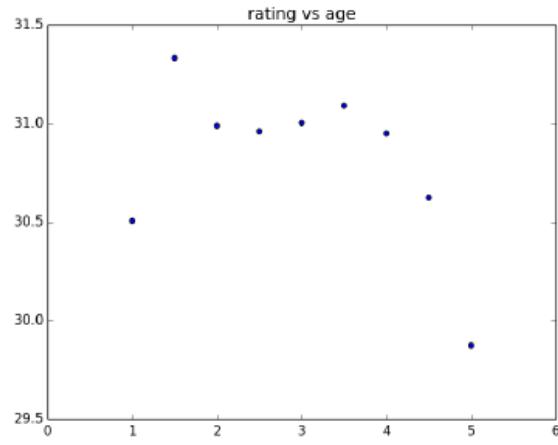


negative words in wine reviews

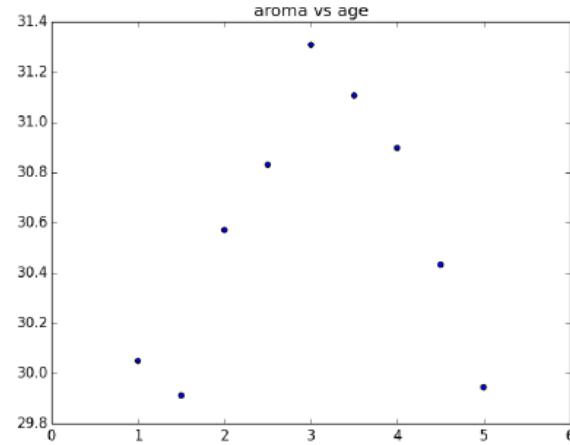


negative words in beer reviews

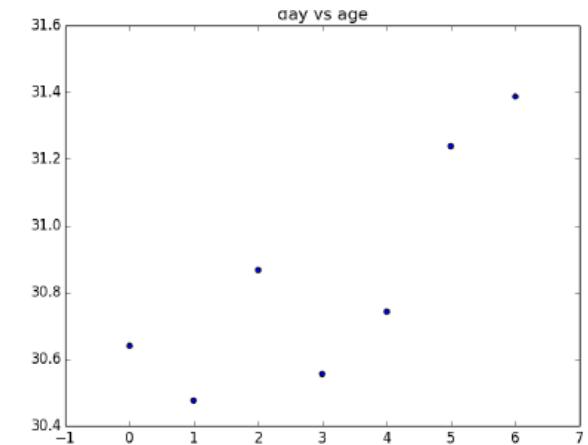
User age



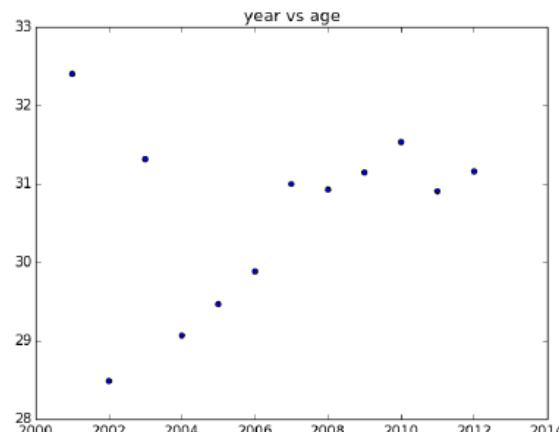
Rating vs. age



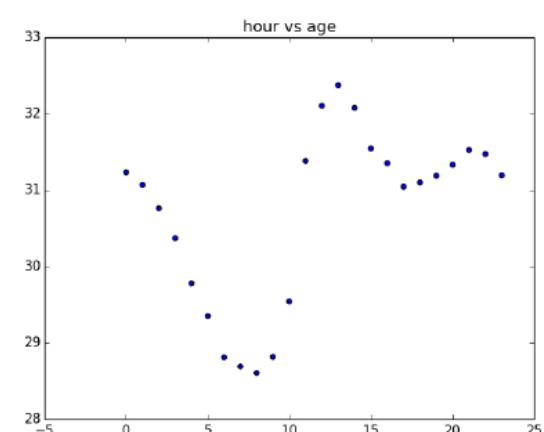
Aroma vs. age



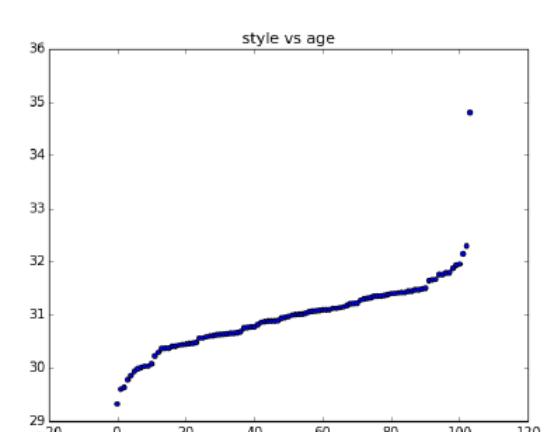
Day of week vs. age



Year vs. age

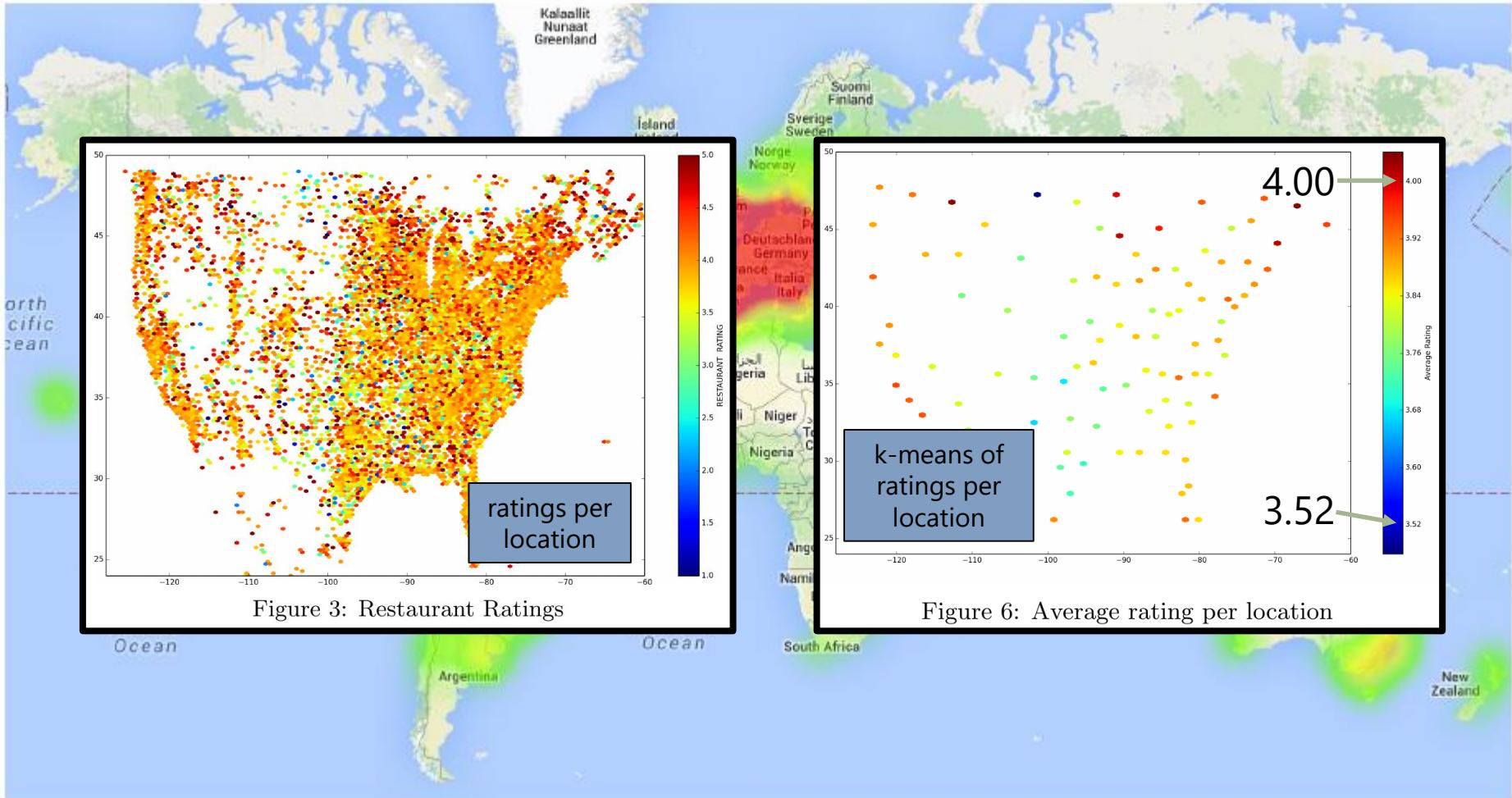


Hour of day vs. age

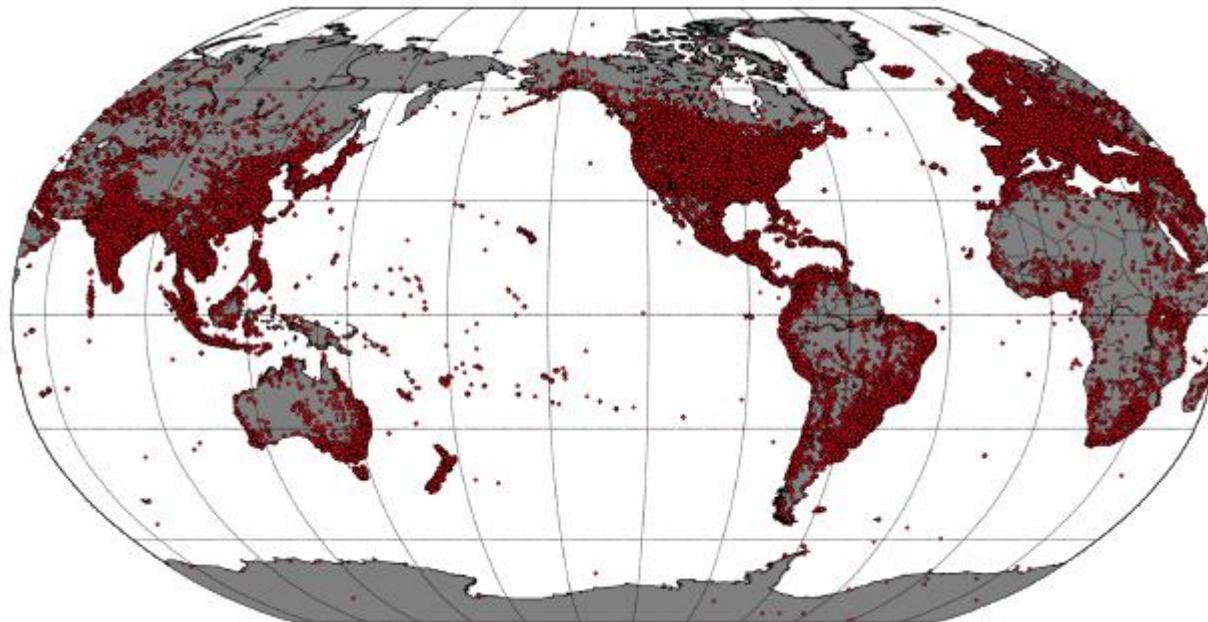


Category vs. age

Assignment 2



Assignment 2

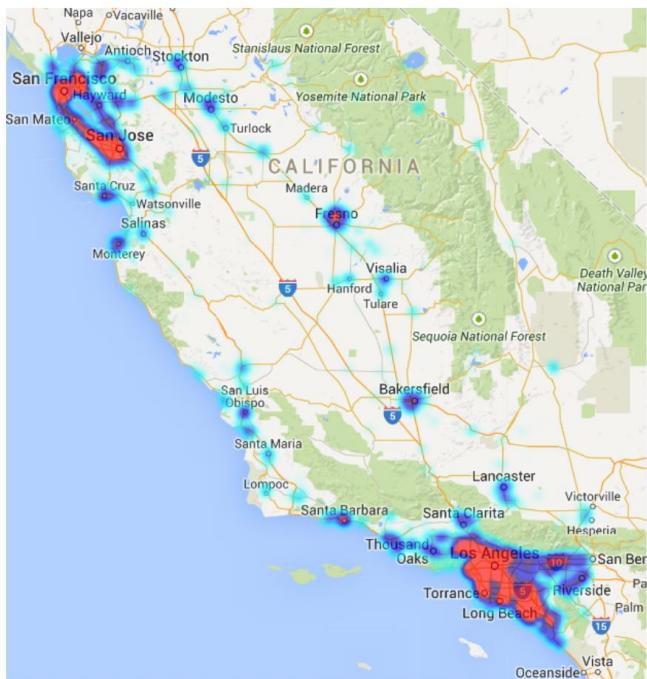


$$\widehat{r}_{ui} = \mu + b_u + b_i + \left(q_i + \frac{1}{|M(i)|} \sum_{n \in M(i)} |s_n| \right)^T p_u$$

set of geographic neighbours

impact of neighbours

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

Assignment 2

Wikispeedia
navigation
traces:

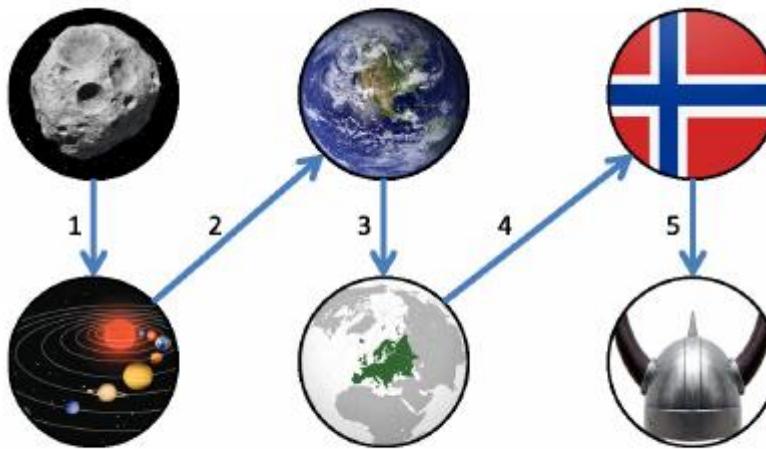


Figure 5: Graph of a complete path

	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

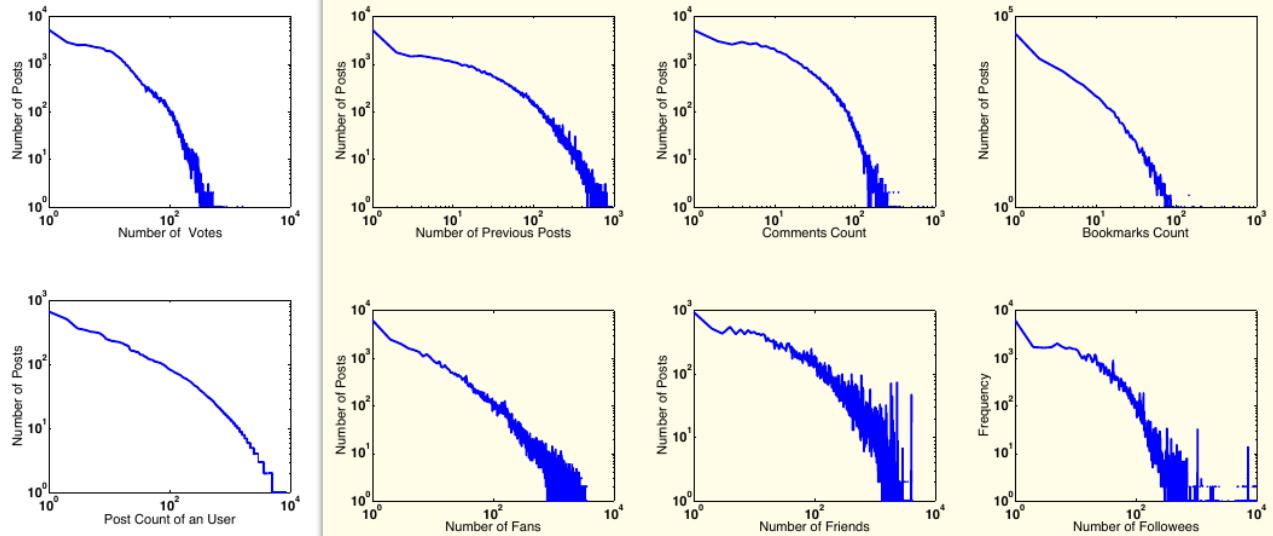
Assignment 2

Images from Chictopia

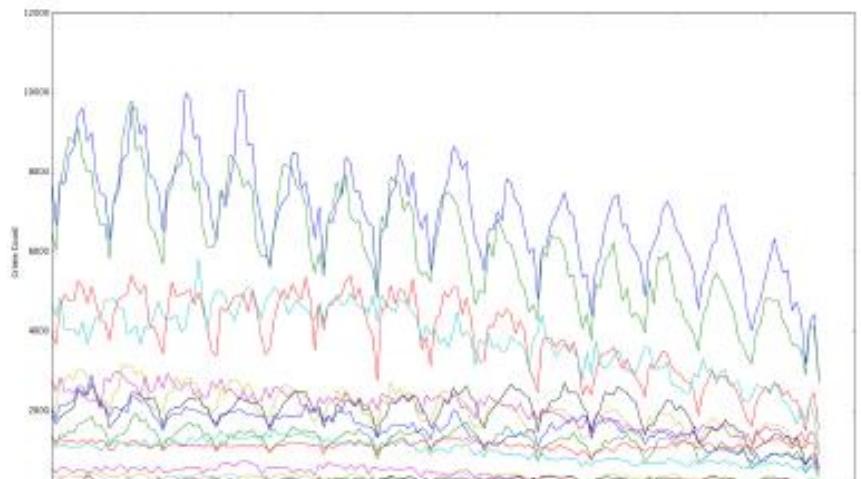


Tags	electric, every day, summer, cute, T-shirt, chic
Clothes	Chartreuse Uniqlo Socks Light Blue Uniqlo T-Shirt Bubble Gum Tie-Ups Belt White Christian Louboutin Heels
User Information	1369 friends 15 followees 2245 fans
Popularity	129 votes 62 comments 15 bookmarks

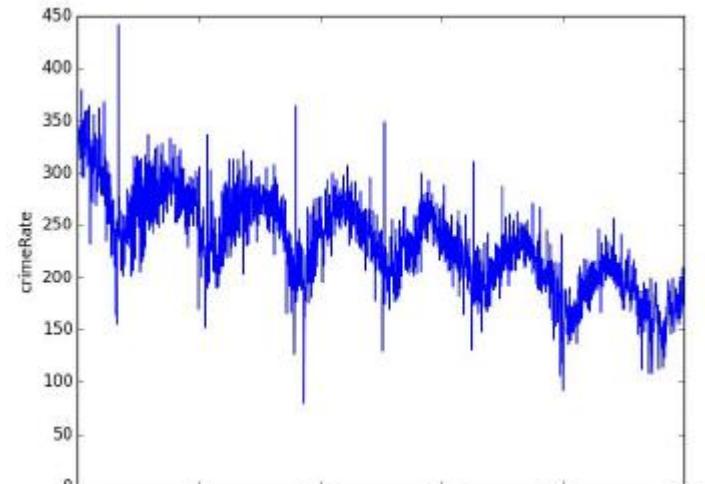
Power laws!



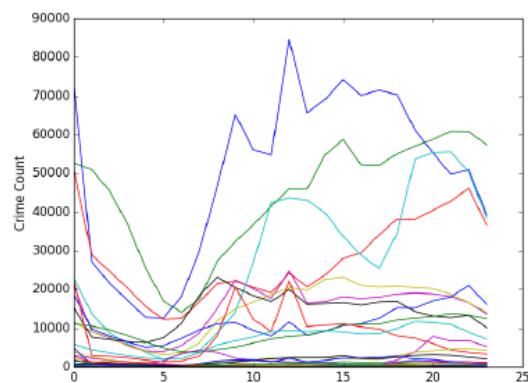
Crime (Chicago)



Over 15 years



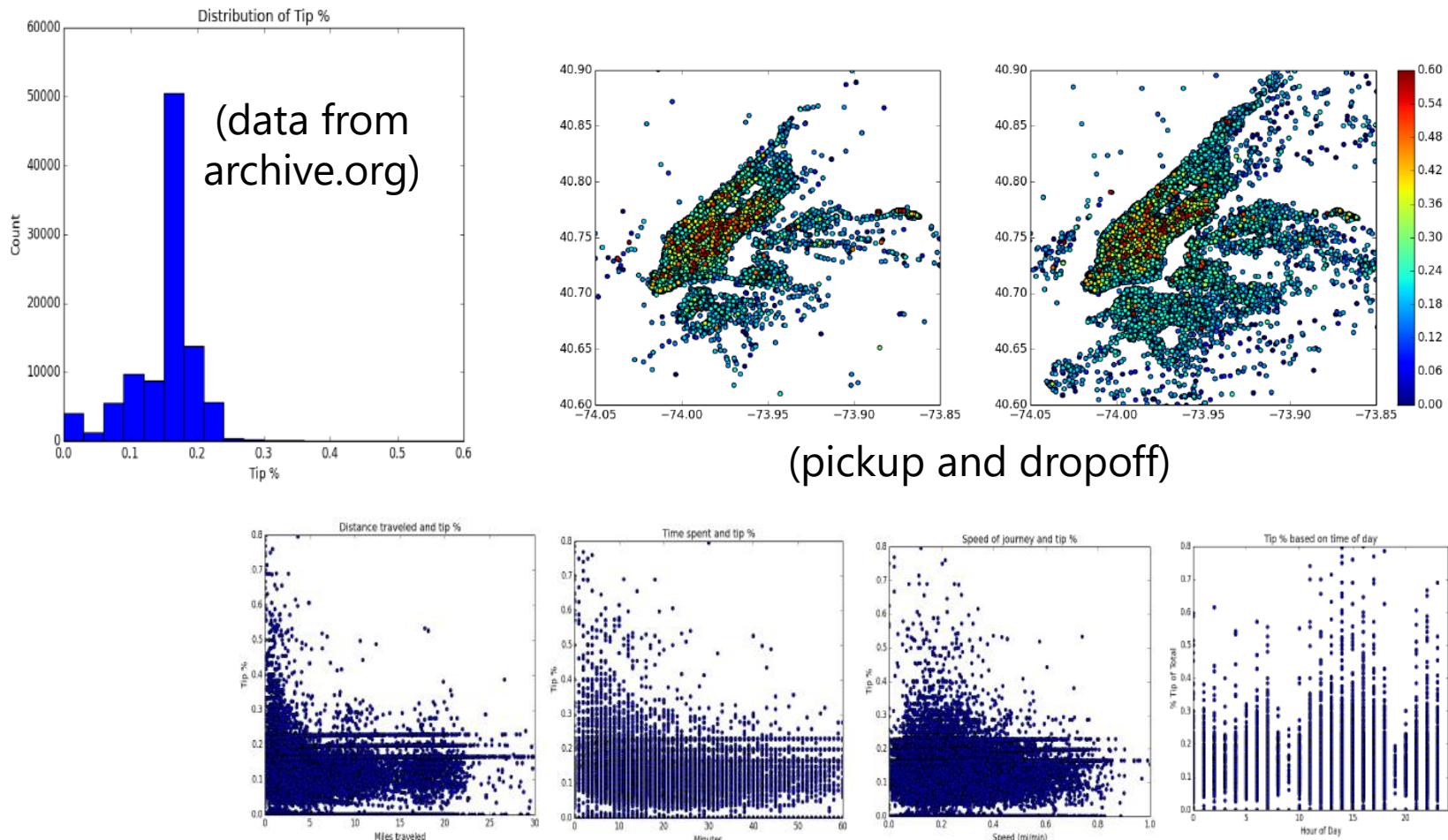
Over 7 years



Hour of the day

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

Predicting Taxi Tip-Rates in NYC



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

TAs

- Ruining He
- Zhanglong Ji
- Caroline Kim
- Zachary Lipton
- Ramkishore Swaminathan
- Mengting Wan

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 (9:00am-1:00pm, CSE 4102)
- TA office hours will be held on Mondays and Fridays from 10:00am-13:00pm in B250A

Questions?

Most announcements will be
posted to Piazza

<https://piazza.com/ucsd/winter2017/cse258/home>

please participate!