

# Data Mining & Analytics

INFO 254 / INFO 154

School of Information / Spring 2019

Prof. Zach Pardos

What will be learned in class?

# What will be learned in class?

- Common sense in data mining and machine learning

Zachary A. Pardos  
Assistant Professor

Graduate School of Education  
School of Information

Research -> [zachpardos.com](https://zachpardos.com)

### Areas of study (Big Data in Education)

- Knowledge representation
- Engineering personalized, adaptive affordances

### Courses taught

- Digital Learning Environments (EDU Online Fall/Summer)
- Machine Learning in Education (EDU/INFO Fall)
- Data Mining & Analytics (INFO Spring)

### Training

PhD in Computer Science (WPI)  
Postdoc at MIT CS AI Lab

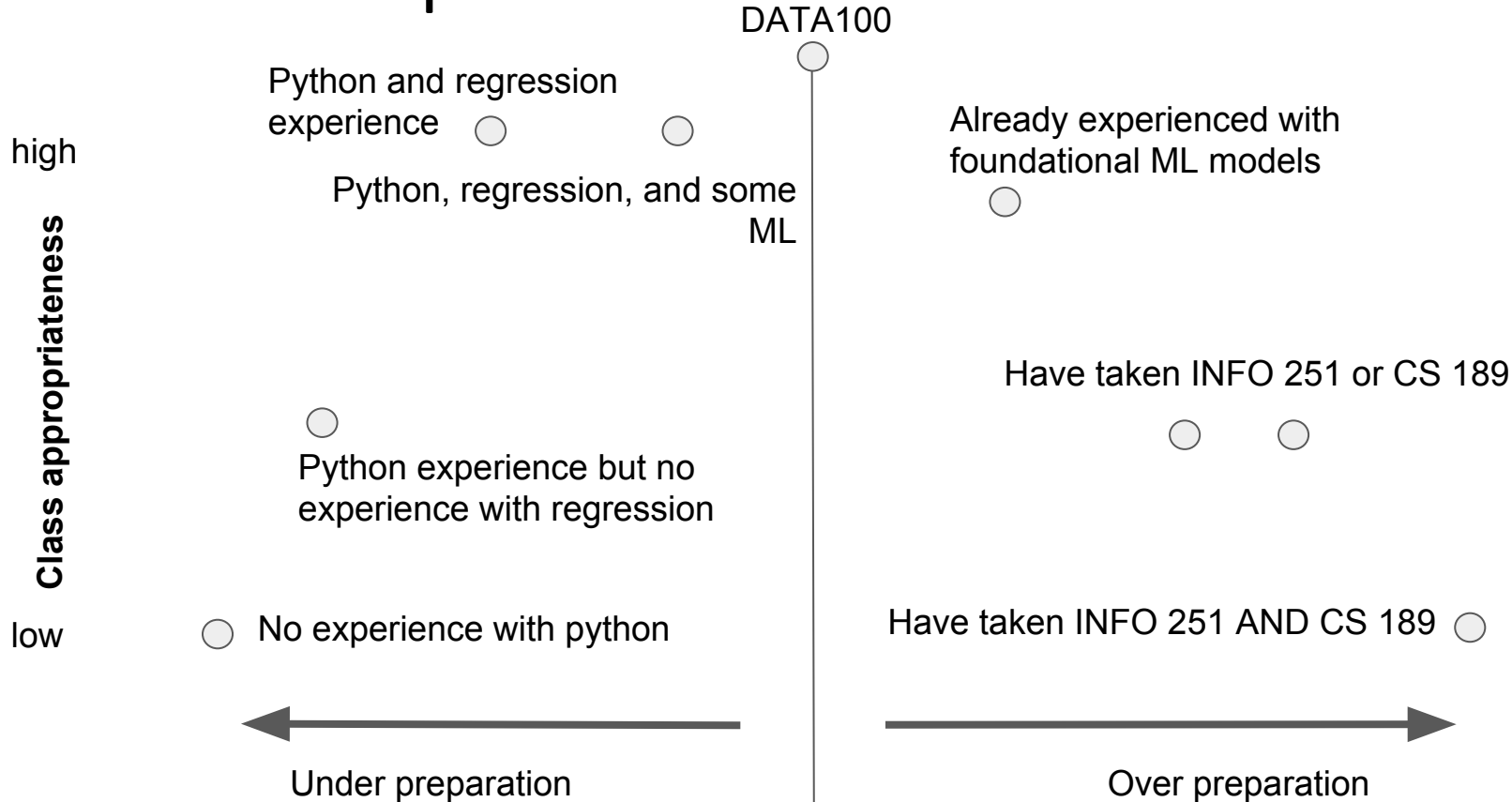
Research Lab: [github.com/CAHLR](https://github.com/CAHLR)

**CAHL** Computational Approaches  
to Human Learning (CAHL)

# Survey results

Bonus question: Piazza or bCourses?

# Class Preparation Guidance



# Overview of topics covered in class

- First half of the semester
  - Foundational models
  - Culminates in prediction competition & midterm
- Second half of the semester
  - Representation learning approaches (neural networks)
  - Culminates in final project presentation (info 154) and paper (info 254)
- Tools
  - Python/pandas
  - Python/numpy
  - Python/scikit-learn
  - Python/keras

Example Final Projects  
from Data Mining and Analytics '16 & '17



# Predicting Basketball Player Salaries

Zubair Marediya, Naya Olmer, Nadar Azari, Boris Lo  
Info 290T | Data Mining and Analytics | Spring 2016

## Goals

- Predicting **basketball players contract values** using historical contracts and player statistics.
- Comparing predicted contract values with actual contract values to **identify which players are undervalued and overvalued.**



## Data

- Player Salaries
- Player Statistics (Minutes, Field Goals, Three Pointers, Rebounds, etc.)

# Features

- Games Played
- Team
- Minutes
- Field Goals
- Field Goal Attempts
- Three Pointers
- Three Point Attempts
- Free Throws
- Free Throw Attempts
- Plus/Minus
- Year
- Salary Cap
- Offensive Rebounds
- Defensive Rebounds
- Total Rebounds
- Assists
- Steals
- Blocks
- Turnovers
- Fouls
- Points
- Age
- Salary
- Percentage of Salary Cap

## Outcome

Predicted player salaries with an RMSE of only **\$2.9 million dollars**.

**40%** of players found over-valued.

**60%** of players found under-valued.

## Example 1: Stephen Curry (2014-2015)

- During the 2014-2015 NBA season, Stephen Curry was named league MVP.
- During this season, Stephen Curry made \$10,629,214.
- The ML model predicts that Curry should have earned \$12,696,622 for his work in 2014-2015, meaning that Curry far exceeded his team's projected value.



**Example 2: James Harden (2014-2015)**

- During the 2014-2015 season, James Harden was the MVP runner up.
- During the season, Harden made \$14,728,884.
- The ML model predicts Harden should have made \$14,441,376, almost exactly what he actually made. This mean that he met the Rockets' expectations.





# [Eudae] Sensing Mood From Fitbit Data



Info 290T Final Project, Spring 2016  
April Dawn Kester, Audrey Leung, Heidi Huang, Laura Montini, Yiwen Tang



## Goal

Predicting a person's mood from their Fitbit data using machine learning.

# Example Final Project: Mood from Fitbit Data

## Data

18 Participants

2 Weeks

4 Mood Surveys per day

Fitbit device continuously tracking

### Mood Features

- Arousal Dimension
- Arousal Dimension Slider
- Pleasure Dimension
- Pleasure Dimension Slider

### Activity Features

- Steps
- Distance
- Floors
- Minutes Asleep
- Heart Rate

The screenshot shows a mobile application interface for mood tracking. At the top, there is a status bar with icons for signal, Wi-Fi, and battery, and the time 11:56. Below the status bar, the text "Please indicate how you feel right now." is displayed. Underneath this text, there are five purple buttons stacked vertically, each containing a mood label: "Very Displeased", "Displeased", "Neutral", "Pleased", and "Very Pleased".

# Outcome

Achieved **80% accuracy** in predicting mood based on Fitbit data.

Best results on aggregate, discrete data

**45%**

Baseline Accuracy  
(Majority Class = Neutral)

**62%**

Random Forest

**80%**

Neural Network

**80%**

Logistic Regression

- NYCe TAXI !

Anand

Anubhav

Sindhuja

## Motivation

Predicting tips for NYC taxi drivers based on the features and characteristics of the taxi rides.

### ● PROJECT GOALS

- - Maximize tip benefits for taxi drivers
  - A generic algorithm that can be extended to any city

## Example Final Project: Maximizing Taxi Tips

### ● RELEVANT FIELDS



#### Unique Fields

- medallion
- hack\_license
- vendor\_id

#### Trip Fields

- pickup\_datetime
- dropoff\_datetime
- trip\_time
- trip\_distance
- pickup\_latitude
- pickup\_longitude
- dropoff\_latitude
- dropoff\_longitude
- passenger\_count

#### Fare Fields

- fare\_amount
- **tip\_amount**
- tolls\_amount
- total\_amount
- mta\_tax
- payment\_type

## Outcome

- 98% accuracy on Tip vs. No Tip
- 81% accuracy on Tip Class
- 68% accuracy on Tip Percentage



### ● INFERENCES

- ◦ Card Payments have higher tip
- Tip varies directly with fare
- Routes with tolls yield less tip
- Tip directly varies with the cost of living index



# Song2Vec

By  
Kartik Gupta, Vishnu Murthy, Elias  
Orellana & Samridh Saluja,

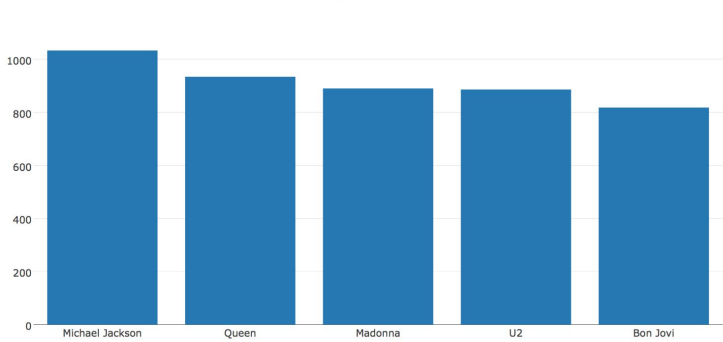


# Summary Statistics

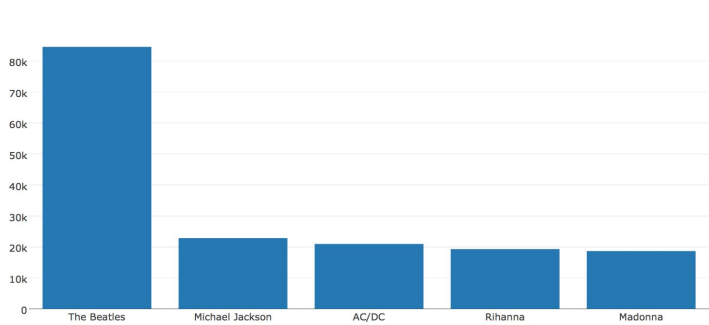
Total No. of Artists	Total No. of Stations	Total No. of Plays
54,592	3,535	8,477,970

Unique Artist	Mean	Median	Max
No. of Stations	22	9	1033
No. of Plays	156	40	84,587

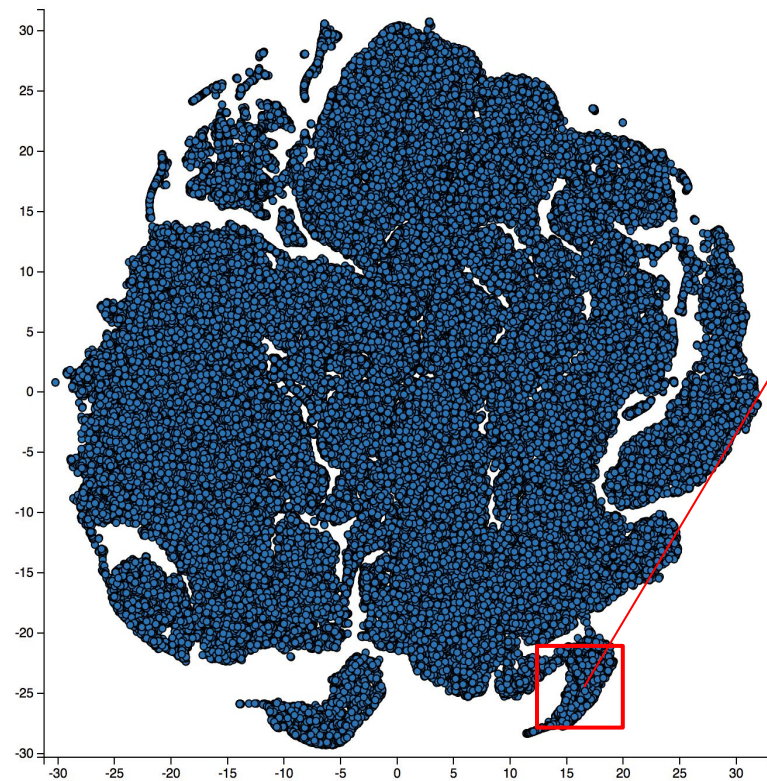
Top 5 Artists Played on Most # of Stations



Top 5 Artists Played based on Most # of Plays



# Visualization of the Data - Unique Songs

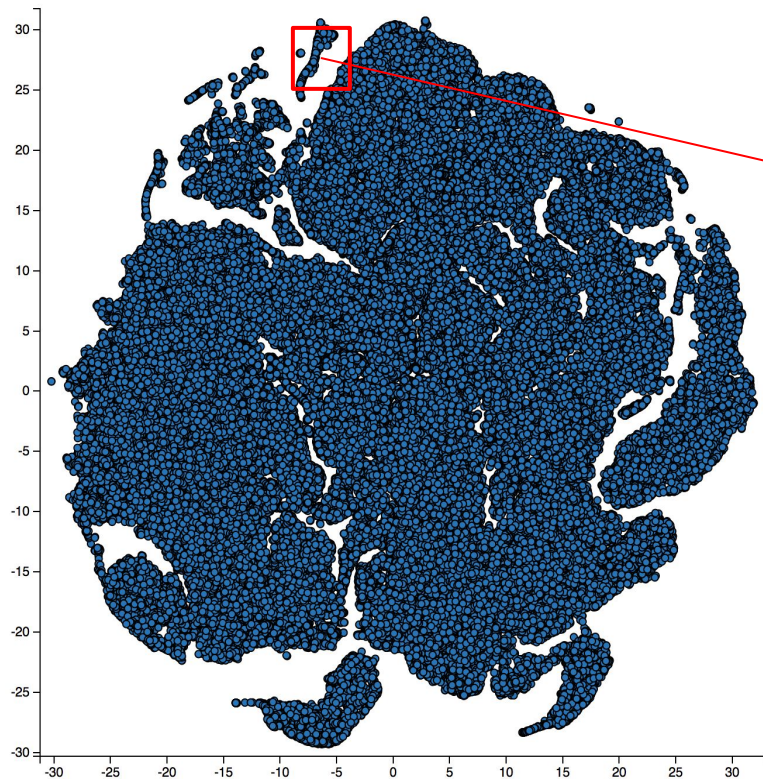


## Genre-based clusters

Rap/contemporary

- 1) Jay-Z - Run This Town feat. Rihanna & Kanye West
- 2) 2Pac - Hit 'Em Up
- 3) Dr. Dre - Xxplosive
- 4) Young Money - Bedrock (Ft Lloyd)
- 5) Chris Brown - No Bull

# Visualization of the Data - Unique Songs



## Genre-based clusters

### Classic Rock

- 1) The Beatles - Mr. Moonlight
- 2) Paul McCartney - Ever Present Past
- 3) John Lennon - Real Love
- 4) Johnny Cash - In My Life
- 5) George Harrison - Here Comes The Sun

# Sanity Check

Same genre

```
model.most_similar(positive=[ 'Drake' ])
```

Same sound

```
[('Lil Wayne', 0.9860175848007202),  
 ('T-Pain', 0.9685032367706299),  
 ('DJ Khaled', 0.9681740999221802),  
 ('Fabolous', 0.9598538279533386),  
 ('Big Sean', 0.9577240347862244),  
 ('Cee-Lo', 0.9560977220535278),  
 ('diddy dirty money', 0.9537268877029419),  
 ('Nicki Minaj', 0.9485852718353271),  
 ('J Cole', 0.9462763071060181),  
 ('Mindless Behavior', 0.9460246562957764)]
```

Record label

# Findings

Artist Specific



Daddy Yankee  
Reggaeton



Drake  
Hip Hop



Reily (Reyli)  
Latin Pop  
0.80

# DMA '17 Class Final Project Titles (1 of 2)

Exploring Online Political Climates

Using Machine Learning to Predict Asthma Risk

Anomaly Detection in time-series data (Water Usage)

Predicting Reasons for Medical Non-adherence

Representation Learning and ML on Materials Data Set

Boston Airbnb Neighborhood Price Analytics

The Effects of Climate Change on East Coast Hurricanes

Sequence to Sequence Approach to Finding Relevant answers in MOOC forums



# DMA '17 Class Final Project Titles (2 of 2)

Indoor Localization

Asmi, Artificial Social Media Intern

song2vec

Generating Semantic Descriptions for Images

Computer Vision Classification of Cervix Types

How Does Temperature Affect Carbon Dioxide Release

Sentiment Analysis : Emotion in Text

Ipredict

Fake News

# Structure of the class

Tuesdays: Concepts introduced, quiz answers reviewed

Thursdays: Quiz (10m), software tutorial, lab begins

Grading:

- Homeworks/Labs - 30%
- Midterm - 25%
- Quizzes - 10%
- Final Project - 35%

# Local optimization vs. Strategic goals

- Local optimization
  - Increasing accuracy (hyper parameter tuning, feature selection, ensembling)
- Strategic goals
  - Who does the predictive analysis benefit and why is it needed instead of descriptive stats?

Strategy comes first!

Example 1

What outcome might you want to achieve with your selection criteria?

- Maximize **engagement** (likes/comments/shares)
- Maximize **network growth** (friend invites)

Kevin L. Smith

Edit Profile

Update Status

Add Photos/Video

nind?

SORT: MOST RECENT

ng Events

[?]

JACQUES RENAULT & JUSTIN MILLER + CALE PARKS

March 9 at 10:00pm

Bossa Nova Civic Club in Brooklyn, New York

Join - 81 people are going

See 1 more

Photos

Browse

ADS

Iman Jordan and 1 other

Create Event

3 requests from Lo Marie

Games You May Like

Zombie Lane

1,000,000 people play Zombie Lane.

Play Now

Facebook © 2013

English (US) · Privacy · Terms · Cookies · More

Thinking Of You"

mugpie.

George Fitzgerald, "Thinking Of You"

www.lphnyc.com

George Fitzgerald, "Thinking Of You"

Like · Comment · Share · 2 · View post · 20 minutes ago

Jennifer Caitlin Welsh

Oh good morning cats!! Are you telling me to get off my phone and feed you? Do you promise not to jump on me if I do?

Like · Comment · Share · 48 minutes ago via mobile

Class of 2010

DC CAP Temple Universi...

Add Group...

MORE

Friends on Chat

Chat (44)

# General ML approach to classification:

1. Create features of the post ( $p_n$ )
2. Create features of the reader/user ( $u_n$ )
3. Choose outcome to optimize ( $Y$ )
4. Learn historical relationship between features and outcome

features of post		features of user		outcome
$p_1$	$p_n$	$u_1$	$u_n$	$Y$

$$f(p_1, \dots, p_n, u_1, \dots, u_n) = Y$$

# Example 2

ucberkeleyofficial  
Sproul Plaza

16h



© Ron Riesterer/Photoshelter

4,492 likes

ucberkeleyofficial Happy #MLKDay, everyone! 🌟 #FiatLux  
Dr. King spoke on Sproul Plaza on May 17, 1967, to a crowd of 7,000.  
Photo copyright Ron Riesterer/Photoshelter)  
[#ucberkeley](#) [#martinlutherking](#) [#sproulplaza](#) [#crowd](#)

[view all 14 comments](#)

das\_the\_shef incredible

catstanton i love this school (:

Comment  
highlighting →

# General ML approach to sentiment classification

1. Create word frequency featureset
2. Hand code sentiment of a sample of comments
3. Learn the influence (weights) of each word wrt the code

frequency of each words in the comment				outcome
$w_1$	$w_2$	...	$w_n$	$Y$

$$f(w_1, w_2, \dots, w_n) = Y$$

# Example 3



## Personalized recommendation

- ▶ [Getting started](#)
- ▶ [1. Balance equation: Our working horse](#)
- ▶ [2. Energy balances](#)
- ▼ [3. Drag force](#)
  - Introduction
  - 3.1 Flow around objects
    - Exercises
  - 3.2 Stokes' Law
    - Exercises
  - 3.3 Terminal velocity**
    - Exercises
  - Graded questions
    - Weekly Exam
- ▶ [4. First steps into heat and mass transfer](#)
- ▶ [5. Newton's law of cooling](#)

Progress

3. Drag force > 3.3 Terminal velocity > Discussion board

[< Previous](#)

[Next >](#)

### Discussion board

[Bookmark this page](#)

This is the place to ask your questions about the introduction, ending, or anything in between. If you have any other feedback, feel free to leave it here as well. Don't forget to help your peers if you understood all the questions!

#### Subsection 3.3 Forum

Topic: Week 3 / Subsection 3: Terminal Velocity

[Show Discussion](#)

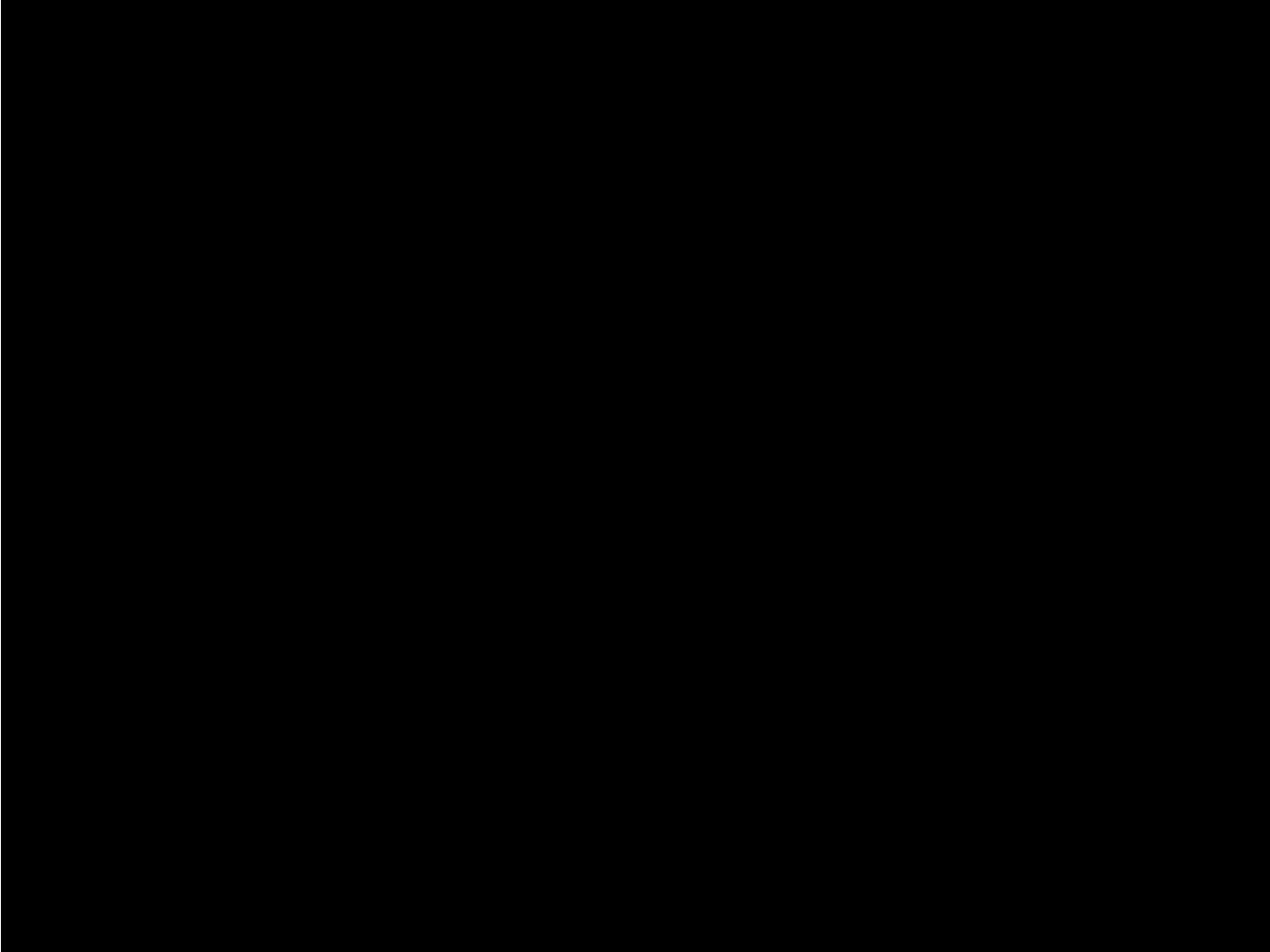
#### Suggestion for You

Consider visiting:  
**3.3 Terminal velocity:** [Answers](#)

[\[?\]](#)

Go







AskOski

[Explore](#)

[Requirements](#)

[Contact](#)

[CalNet Login](#)

# AskOski

Explore personalized course information based on historic enrollments.



Project driven by data science research

[CalNet Login](#)

[AskOski.berkeley.edu](https://AskOski.berkeley.edu)

# Overview of topics covered in class

- First half of the semester
  - Foundational models
  - Culminates in prediction competition & midterm
- Second half of the semester
  - **Representation learning approaches** (neural networks)
  - Culminates in final project presentation (info190) and paper (info254)
- Tools
  - Python/pandas
  - Python/numpy
  - Python/scikit-learn
  - Python/keras

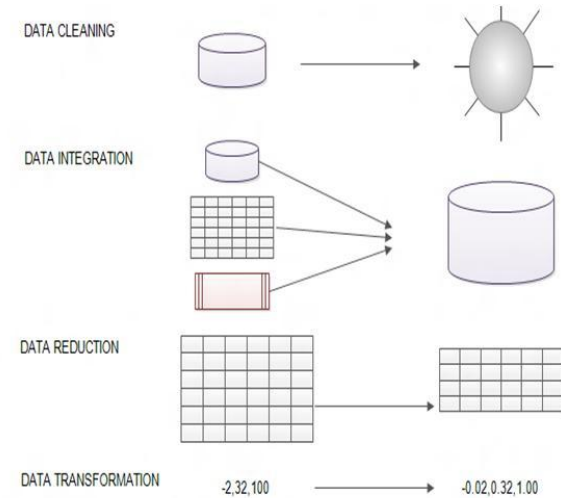
# Data Preprocessing

- “Garbage in, garbage out”
- Majority of data miners spend 60% or more of their time on data cleaning and preparation) \*

[http://www.kdnuggets.com/polls/2003/data\\_preparation.htm](http://www.kdnuggets.com/polls/2003/data_preparation.htm)

## Topics

- Data Cleaning ( missing values, noisy data )
- Data Reduction ( PCA, Regression and Binning )
- Data Transformation ( Aggregation, Smoothing )

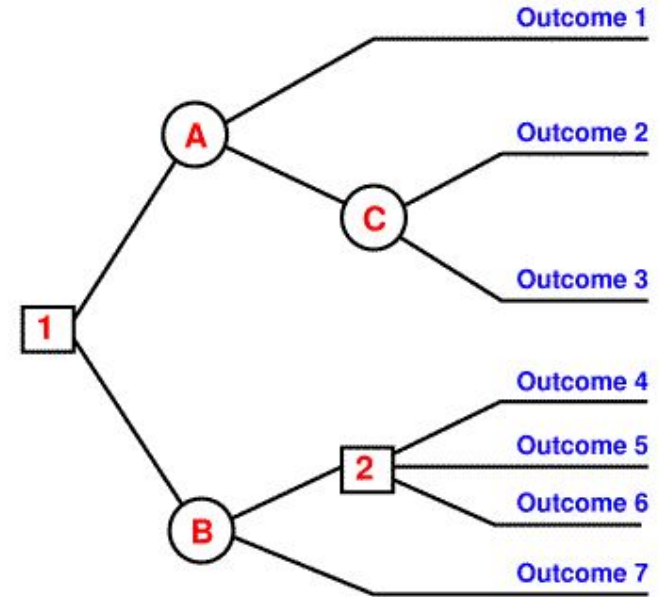


# Decision Trees

- Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from features.

## Topics

- Decision Tree Induction
- Attribute Selection Measures - Information Gain, Gain Ratio, Gini Index
- Tree Pruning

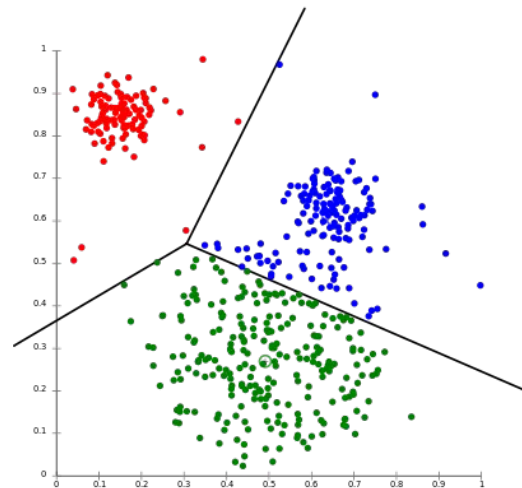


# Clustering

- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (by some measure) to each other than to those in other groups. This is a popular form of unsupervised learning.

## Topics

- K-means Clustering
- Spectral Clustering
- Cluster Quality ( Elbow method, Silhouette Coefficient )



# Error Metrics

- Metrics for evaluating classifier performance

## Topics

- Confusion matrix
- Accuracy
- AUC
- Specificity
- Precision and Recall

precision: TP/cancer diagnoses

		Diagnosis	
		No cancer	Cancer
True state	No cancer	TN	FP
	Cancer	FN	TP

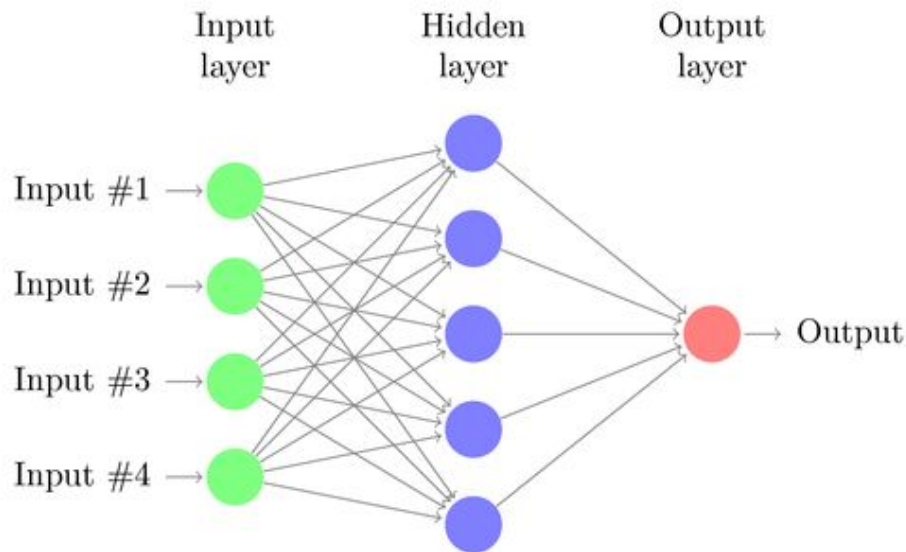
recall: TP/cancer true states

# Neural Networks

- Artificial neural networks are a series of nodes (w/activations) and edges (w/weights) that can learn arbitrary functions and were loosely inspired by the neural structure of the brain.

## Topics

- Multi-perceptron FFNN
- Back propagation
- Recurrent Neural Networks



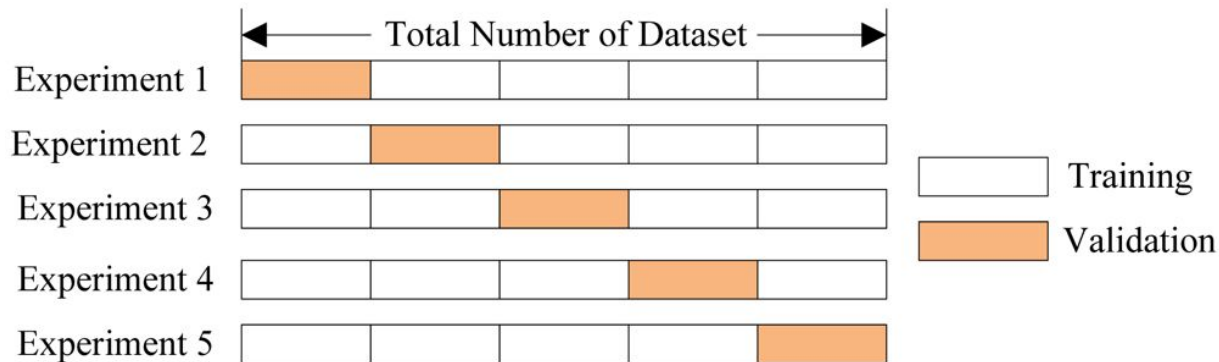


# Cross-validation and Classifier Enhancement

- An evaluation to improve reliability of performance

## Topics

- Hold out strategies
- Generalization
- Statistical significance

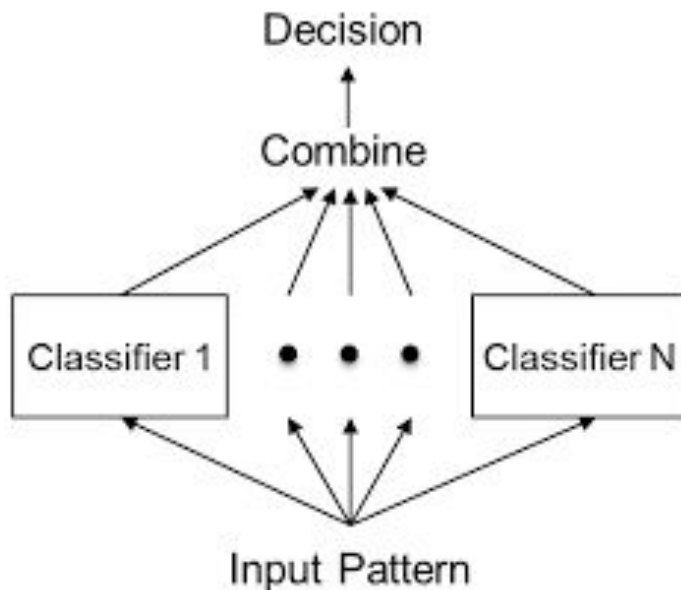


# Ensemble Learning

- Ensemble learning is the combination of predictors that often results in better performance than the individual predictors could achieve alone

## Topics

- Bagging
- Boosting (Adaboost)
- Random forests

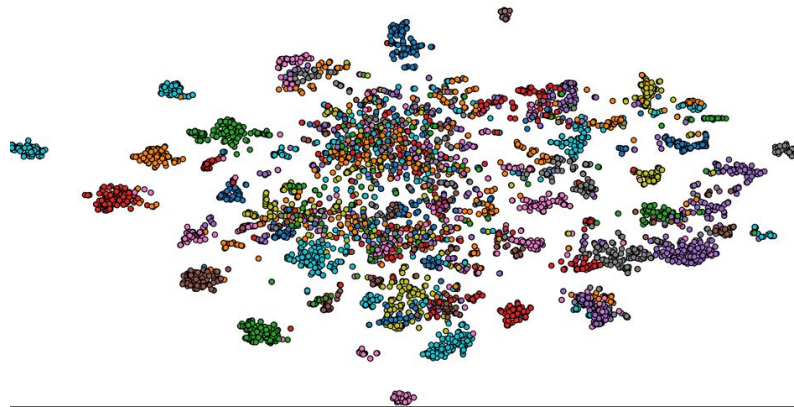


# Representation Learning

- New approach to ML: Instead of hand engineering features, learn them from the raw data

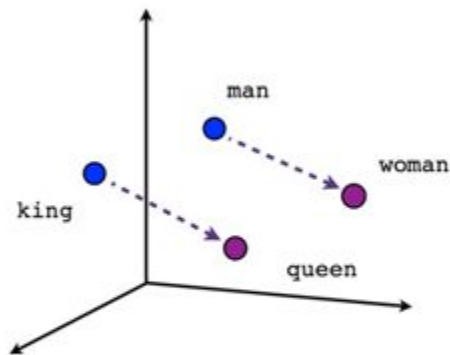
## Topics

- Skip-grams
- RNNs (“deep” learning)
- Dimensionality reduction
- Visualization (cluster analysis)

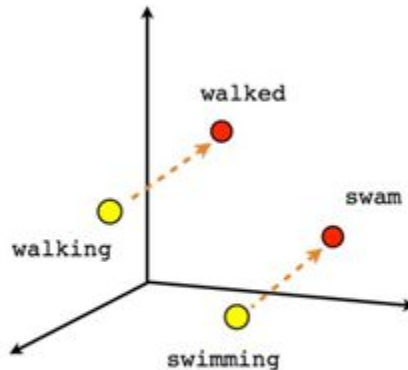


Berkeley Course Representation

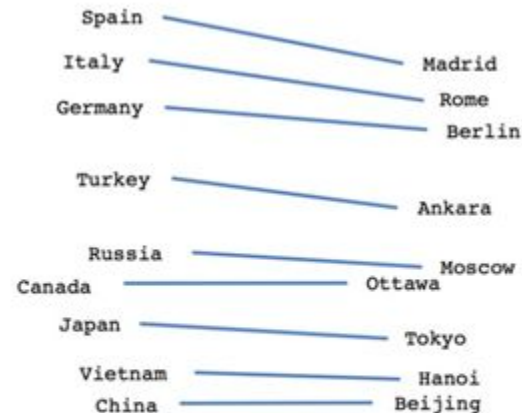
# Word2Vec Representation Learning



Male-Female



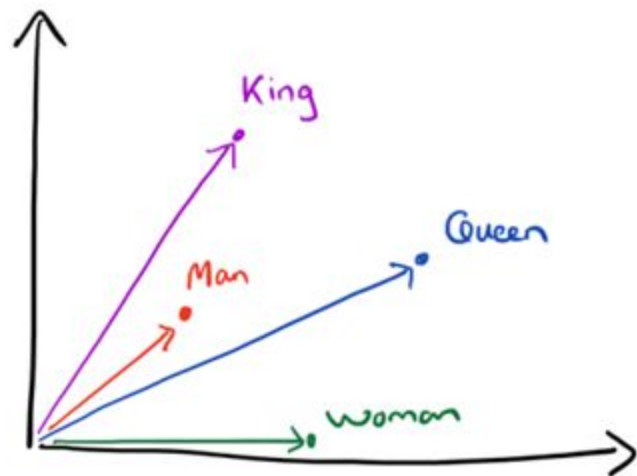
Verb tense



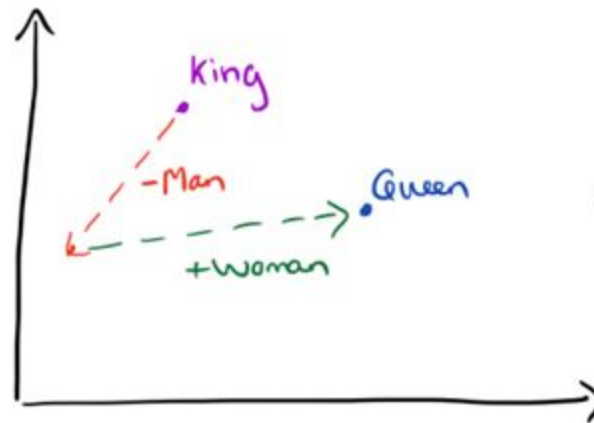
Country-Capital

Mikolov, T., & Dean, J. (2013)

# Word2Vec Representation Learning



Word  
Vectors



Vector  
Composition

(Adrian Colyer, 2016)

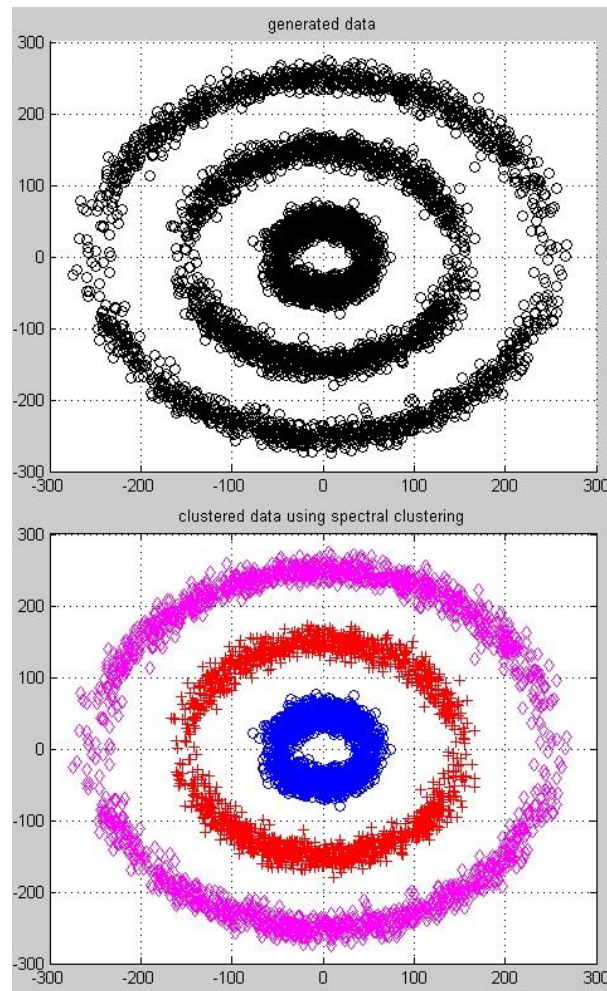
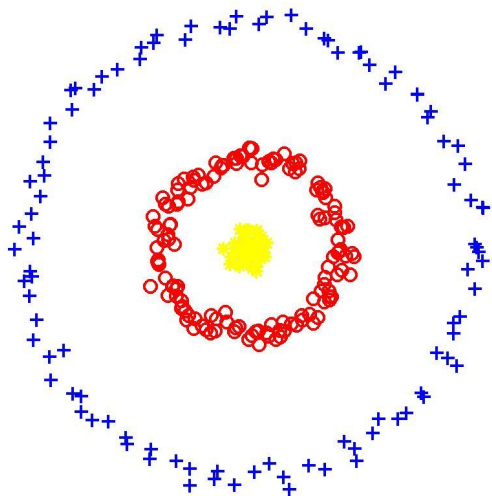
$$\text{KING}[\text{vec}] - \text{MAN}[\text{vec}] + \text{WOMAN}[\text{vec}] \approx \text{QUEEN}[\text{vec}]$$

Finding implicit bias in language

Bolukbasi (2016); Caliskan (2017)

# Advanced Clustering

- Graph theoretic form of clustering
- Can capture geometric patterns instead of only spherical gaussians (*k-means*)



[see you Tuesday]

# Questions?

INFO 254 / INFO 154

School of Information / Spring 2019

Prof. Zach Pardos