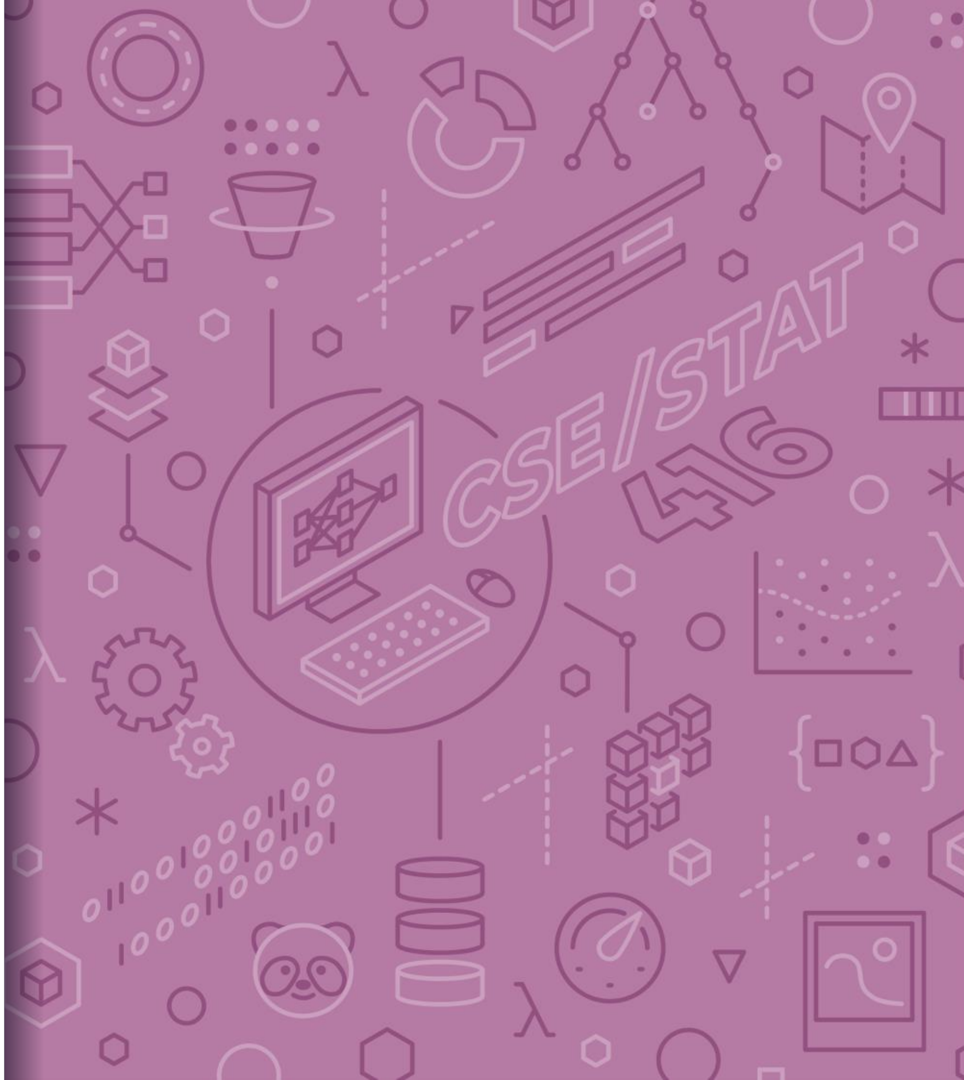
CSE/STAT 416

Course Wrap Up & Guest Lectures

Amal Nanavati
Paul G. Allen School of Computer Science & Engineering
University of Washington
August 15, 2022

Adapted from Hunter Schafer's slides



Upcoming Deadlines

- TOMORROW Tues 8/16 11:59PM: HW7 due
  - **NO LATE DAYS**!!!
- Wed 8/17 9AM: Final Exam released
- Thurs 8/18 11:59PM: Final Exam due
  - **NO EXTENSIONS**!!!
- Fri 8/19 11:59PM: Guest lecture extra credit
  - Worth 1 Conceptual Homework (3.57% of your grade)
  - Submit on Gradescope.
- No course work will be accepted after Fri Aug 19 11:59PM
  - e.g., late checkpoints

Final Exam Logistics

- Released Wed 8/17 **9AM**, Due Thurs 8/18 **11:59PM**
- On Gradescope, completed **individually**
- Expected Length: **2 hours** (with time pressure)
  - You can take it for any subset of the 38 hours 59 mins it is released, including in multiple sittings.
- **Allowable Resources**:
  - Your Learning Reflections
  - Lecture Slides & Personal Notes
  - Checkpoints
  - HW Assignments
- **Disallowed Resources**:
  - Google / the Internet
  - Your peers
- **Getting Help**:
  - Office hours are canceled Wed-Fri!
  - We will only respond to EdSTEM questions on logistics and clarifications
  - All EdSTEM responses will be public.

Final Exam Format

- 11 questions, each with several subquestions
- ~45 subquestions total
  - ~ 1/3 Free Response
  - ~ 1/3 Numeric Calculations
  - ~ 1/3 Multiple Choice Questions
  - (One question asks you to upload a file, other questions give you the option to upload a file showing your work)
- All Conceptual
  - Think of it like a cumulative conceptual assignment
- 15% of your course grade
- **BE SURE TO SAVE YOUR ANSWERS FREQUENTLY**!

Tips on Taking the Final Exam

- Take-home exams are like a gas; they expand to fill all the time you give it!
  - You can overthink every question, you can cross-reference course material for every question. This is not something you'd do for an in-class exam.
- To avoid this exam from taking up all your time:
  - Set a maximum amount of time you'll spend on the final. (e.g., 3 hours? 4 hours? Your choice.)
  - First pass:
    - Set a timer for 2 hours
    - Take it under time pressure. Submit your best answer given the time constraints.
    - Note down which questions you're less sure about.
  - Remaining pass(es):
    - Revisit the questions you were unsure about, try them with more time.
  - Submit and stop thinking about the exam when the max time has elapsed! At some point, spending more time won't help.

Poll Everywhere

Think

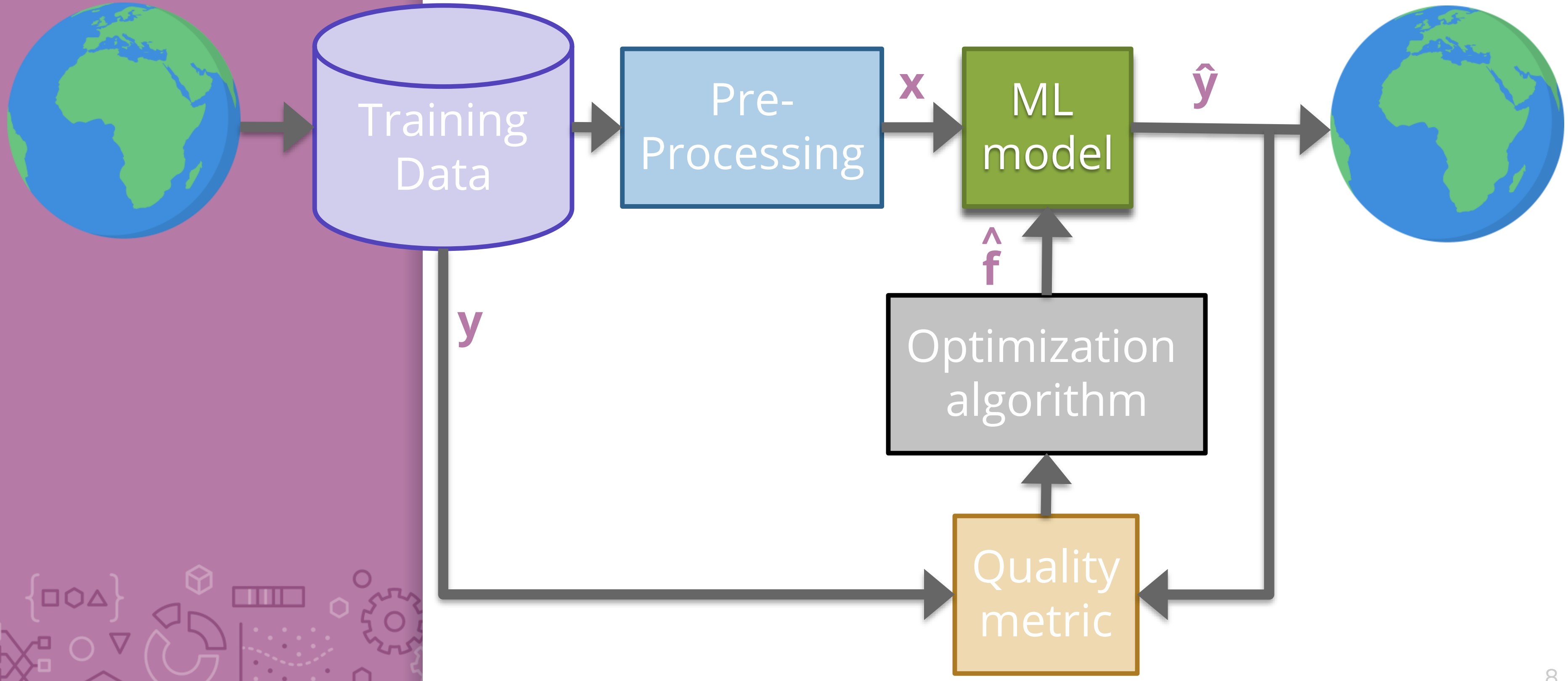
5 mins

Till 2:35 fill out course evals

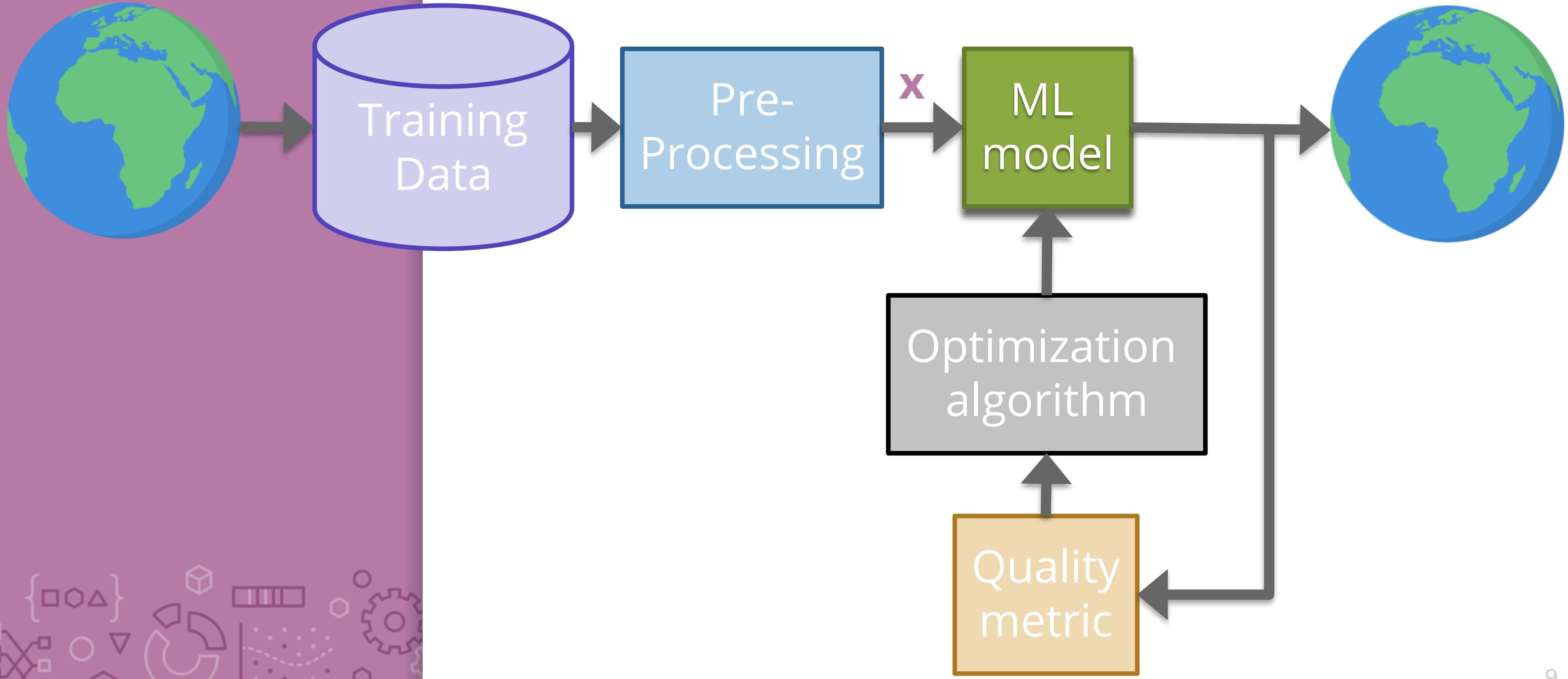
- There is extra credit available on the final if you complete the course evals.
- **Take 5 minutes right now to complete course and section evals:**
  - **Course**: https://uw.iasystem.org/survey/261325
  - **Section AA/BA** (Wuwei): https://uw.iasystem.org/survey/261326
  - **Section AB/BB** (Karman): https://uw.iasystem.org/survey/261327
  - **Section AC/BC** (Max): https://uw.iasystem.org/survey/261189

Course Recap

ML Pipeline (Supervised)



ML Pipeline (Unsupervised)



Poll Everywhere

Group

5 mins

pollev.com/cs416

- Let’s use the ML Pipeline to classify the concepts we’ve learnt in the course so far!
- For each component of the ML Pipeline below, contribute to the PollEv word cloud regarding what concepts fir into that component! (1 min each)
  - Pre-Processing
  - ML Models
  - Quality Metrics
  - Optimization Algorithms
  - Concepts that don’t fit neatly into one category of the pipeline

One Slide

House Prices

- Regression
- Overfitting
- Bias-Variance tradeoff
- Training, test, and validation error
- Cross validation
- Ridge, LASSO
- Standardization
- Gradient Descent

① Sentiment Analysis

② Loan Safety

③ Kaggle

- Classification
- Text Encodings (BoW, TF-IDF)
- Logistic Regression
- Social Bias & Fairness in ML
- k-NN Classification
- Decision Trees
- Random Forests
- AdaBoost
- Precision and Recall
- Handling Missing Data

Image Classification

- Neural Networks
- Convolutional Neural Networks
- Transfer Learning for deep neural networks

Document Clustering & Analysis

- Unsupervised v. supervised learning
- k-means clustering
- Hierarchical clustering
- Dimensionality reduction, PCA

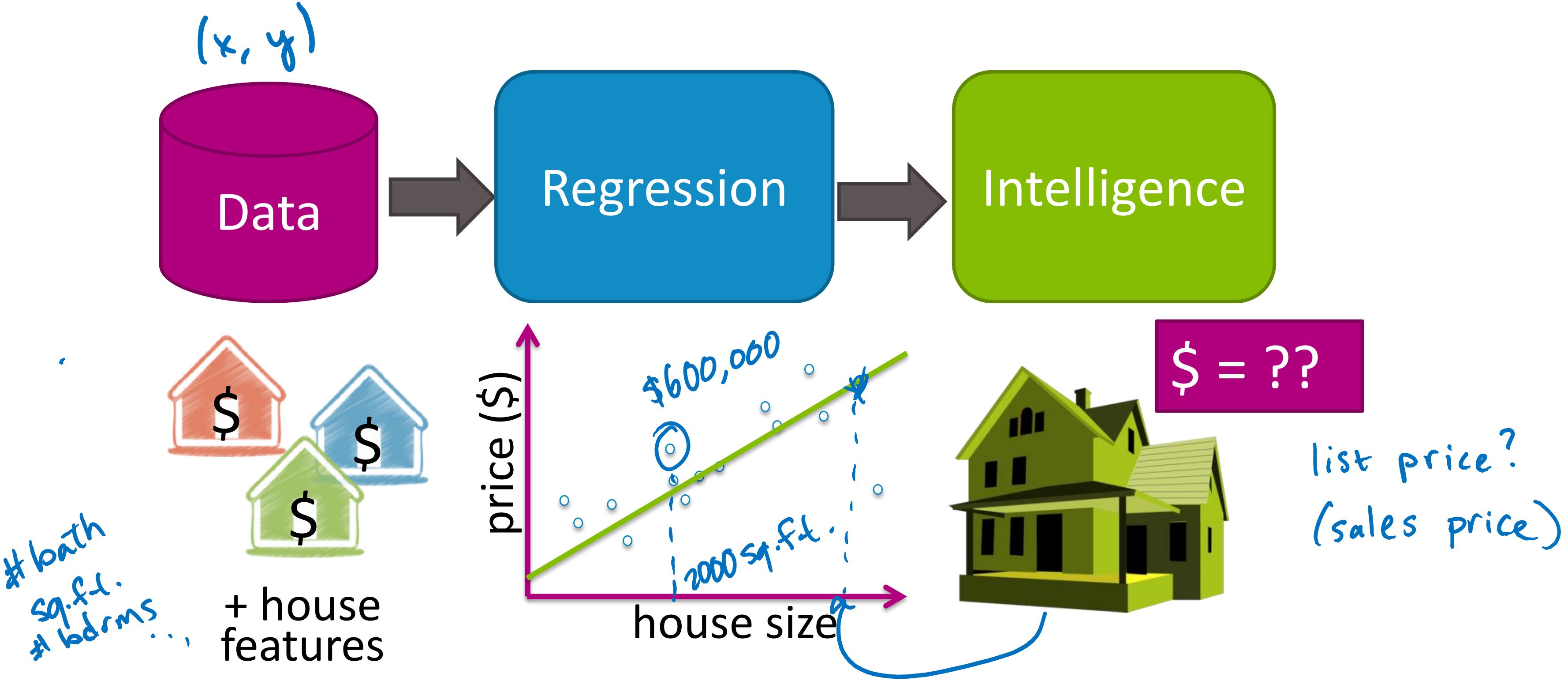
Product Recommendation

- Recommender systems
- Matrix factorization
- Coordinate descent

Case Study 1:
Predicting house prices

Model: $y_i = f(x_i) + \epsilon_i$

Predictor: $\hat{y}_i = \hat{f}(x_i)$



Data

Regression

Intelligence

+ house features

$ = ??

Regression

Ridge: $\arg\min_{w} L(w) + \lambda \|w\|_2^2$

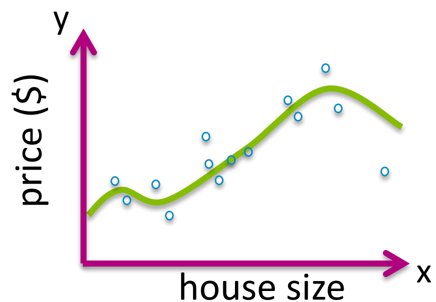
Case study: Predicting house prices

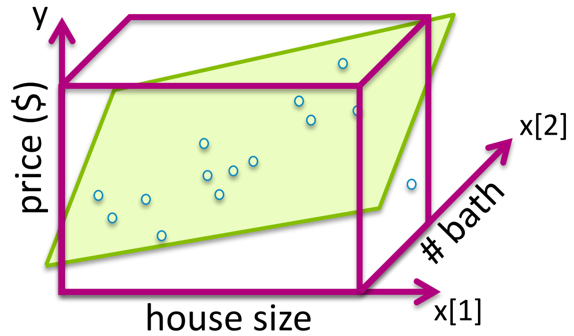
Models

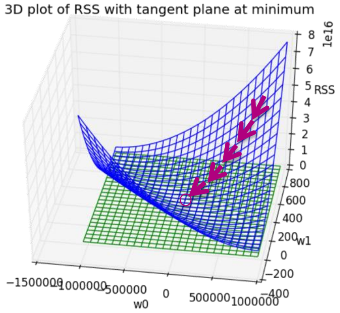
- Linear regression
- Regularization: Ridge (L2), Lasso (L1)

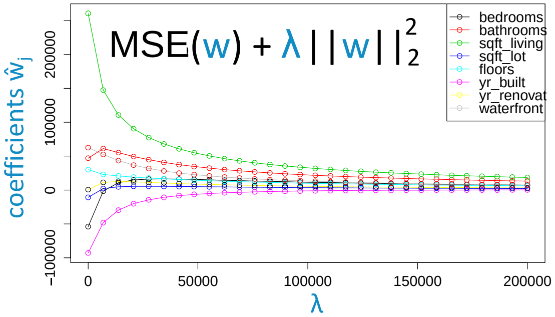
Algorithms

- Gradient descent







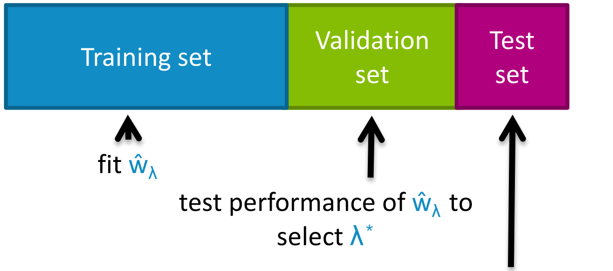


Regression

Case study: Predicting house prices

Concepts

- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection

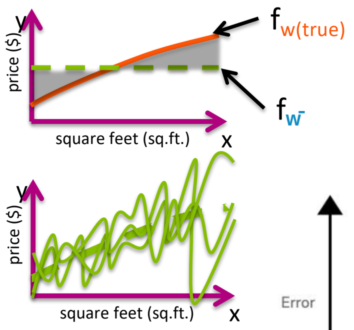


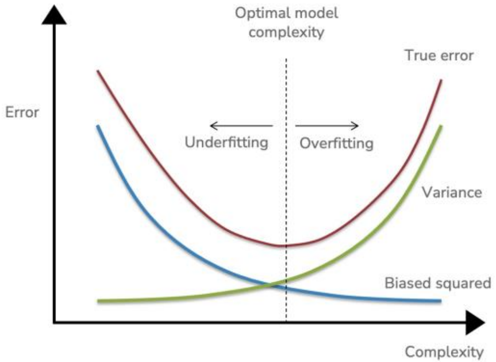
fit $\hat{w}_\lambda$

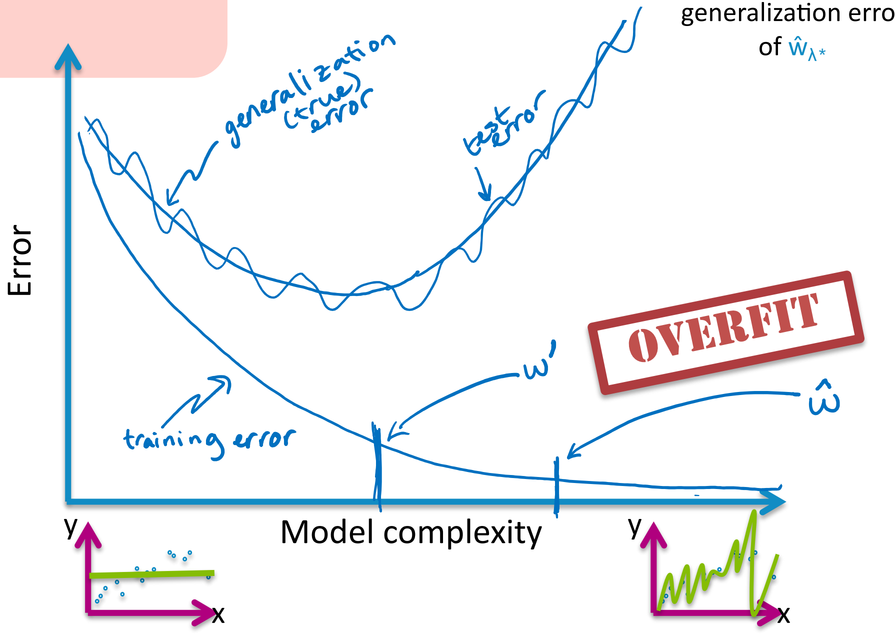
test performance of $\hat{w}_\lambda$ to select λ^*

assess generalization erro of $\hat{w}_{\lambda^*}$

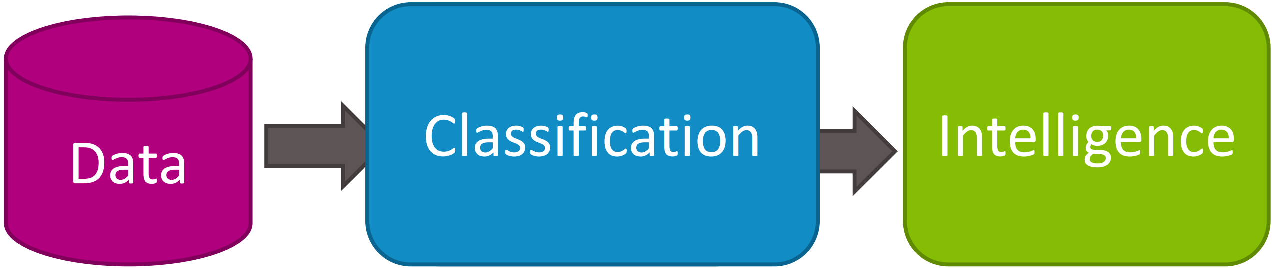
1. Noise
2. Bias
3. Variance







Case Study 2:
Sentiment analysis

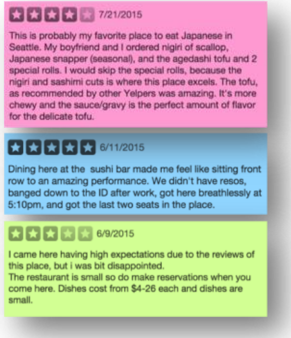


Sushi was awesome,
the food was awesome,
but the service was awful.

stars / +/-

text

All reviews:


7/21/2015

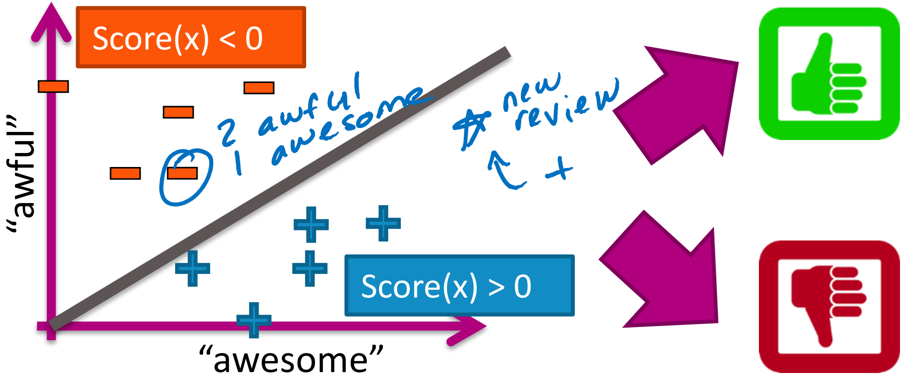
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended by other Yelpers was amazing. It's more chewy and the sauce/gravy is the perfect amount of flavor for the delicate tofu.

6/11/2015

Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have resos, banged down to the ID after work, got here breathlessly at 5:10pm, and got the last two seats in the place.

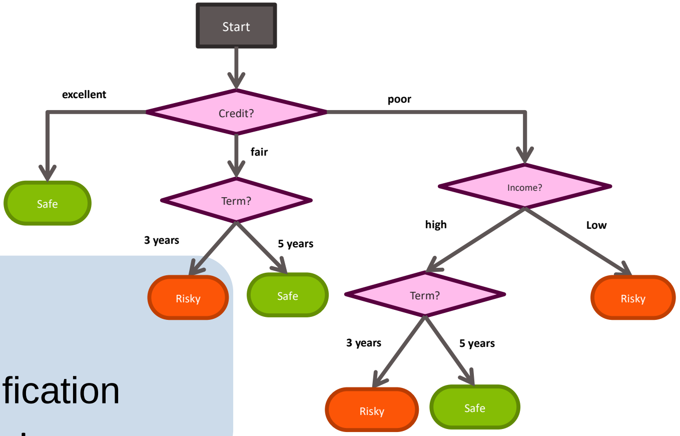
6/9/2015

I came here having high expectations due to the reviews of this place, but i was bit disappointed.
The restaurant is small so do make reservations when you come here. Dishes cost from $4-26 each and dishes are small.



Classification

Case study: Analyzing sentiment

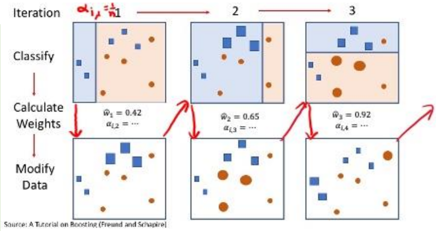


Models

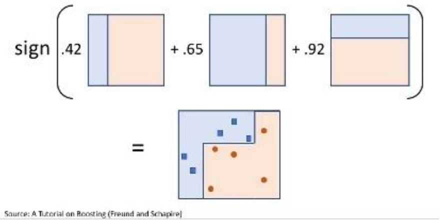
- Linear classifiers (logistic regression)
- Multiclass classifiers
- Decision trees, k-nearest neighbors classification
- Boosted decision trees and random forests

Algorithms

- Boosting
- Learning from weighted data



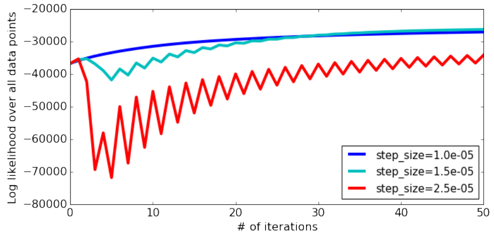
Source: A Tutorial on Boosting (Freund and Schapire)



Source: A Tutorial on Boosting (Freund and Schapire)

Concepts

- Decision boundaries, maximum likelihood estimation, ensemble methods, random forests
- Precision and recall

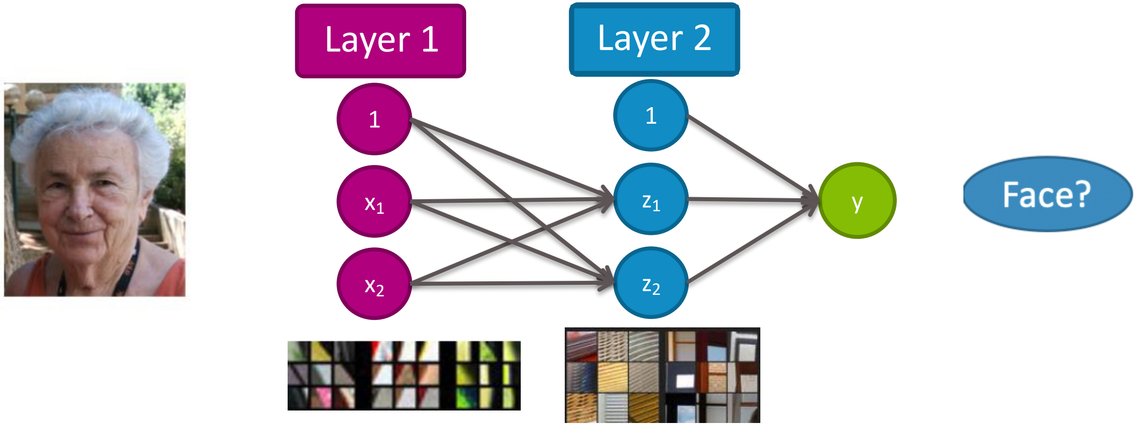


optimum

Case Study 3:

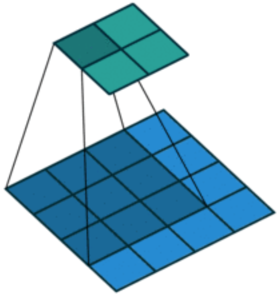
Image classification

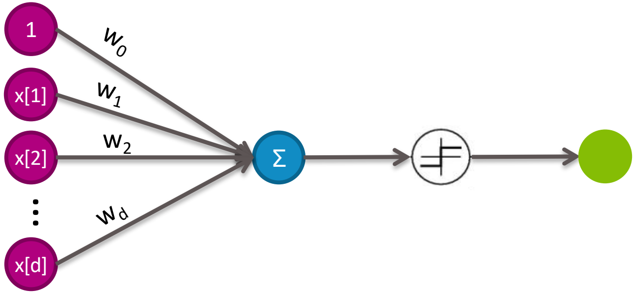
Data
Deep Learning
Intelligence



Deep Learning

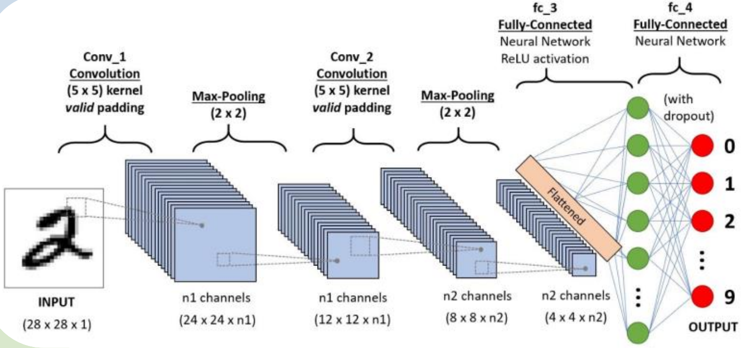
Case study: Image classification





Models

- Perceptron
- General neural network
- Convolutional neural network

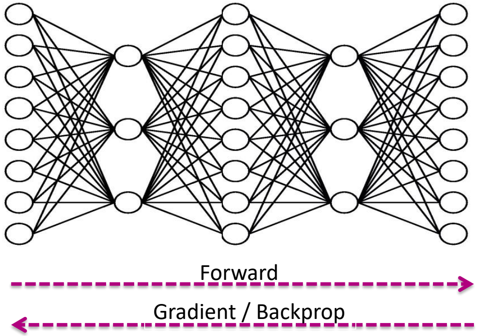


Algorithms

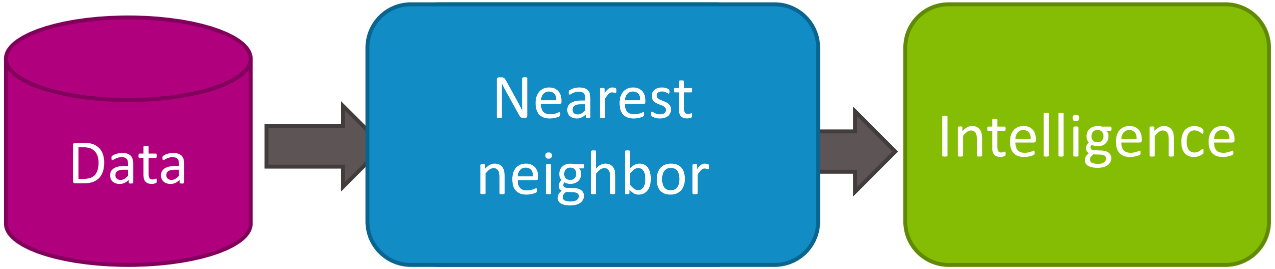
- Convolutions
- Backpropagation (high level only)

Concepts

- Activation functions, hidden layers, architecture choices

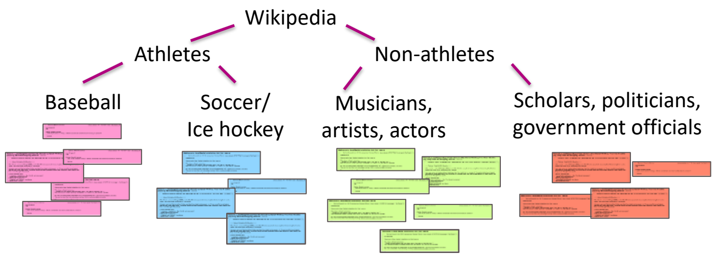


Case Study 4:
Document Clustering & Analysis



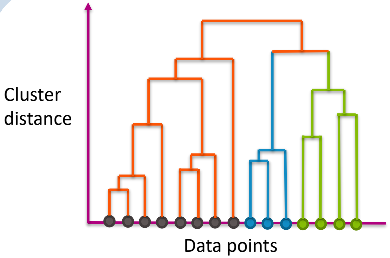
Clustering & Retrieval

Case study: Finding documents


Wikipedia
Athletes
Non-athletes
Baseball
Soccer/ Ice hockey
Musicians, artists, actors
Scholars, politicians, government officials

Models

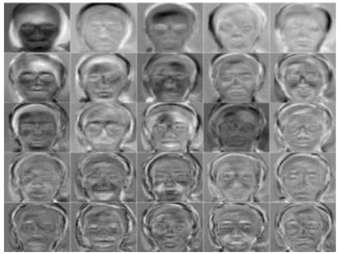
- Clustering
- Mixture Models
- Hierarchical Clustering

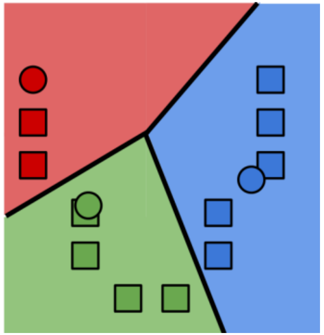


Algorithms

- k-means / k-means++
- Agglomerative & Divisive Clustering
- Principal Component Analysis

Principal components:



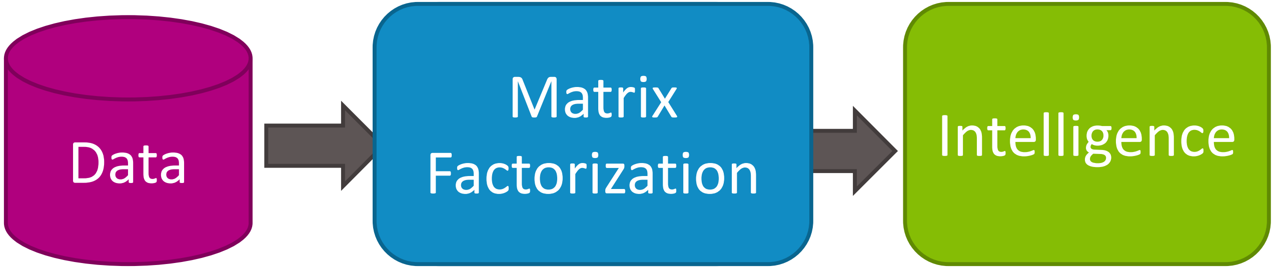


Concepts

- Unsupervised Learning
- Clustering
- Dimensionality Reduction

Reconstructing:

Case Study 5:
Product recommendation



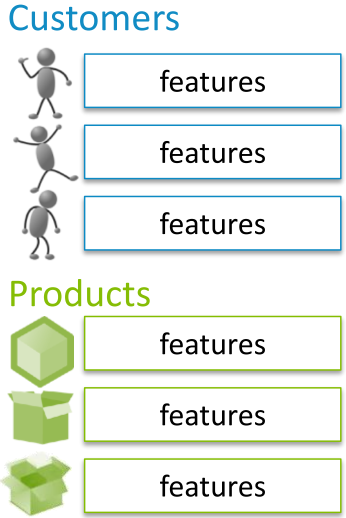
Data

Matrix Factorization

Intelligence

Your past purchases:

+ purchase histories of all customers



Customers

features

features

features

Products

features

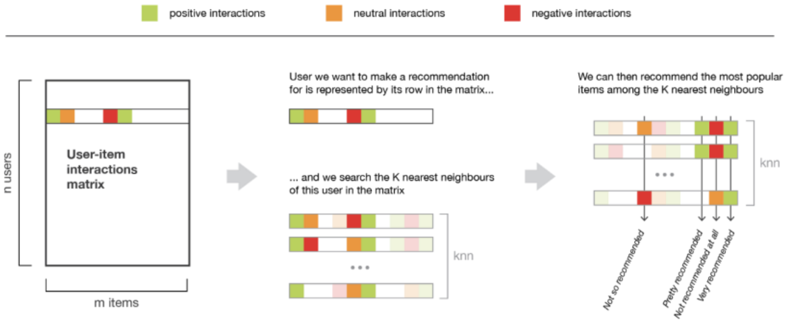
features

features

Recommended items:

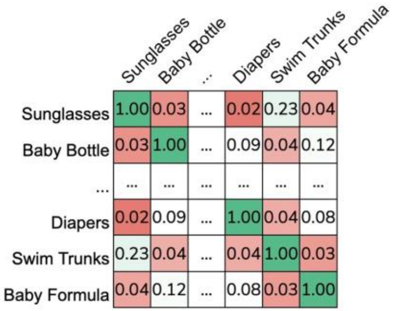
Recommender Systems & Matrix Factorization

Case study: Recommending Products


positive interactions
neutral interactions
negative interactions
n users
User-item interactions matrix
m items
User we want to make a recommendation for is represented by its row in the matrix...
... and we search the K nearest neighbours of this user in the matrix
knn
We can then recommend the most popular items among the K nearest neighbours
knn
Not so recommended
Pretty recommended
Not recommended at all
Very recommended

Models

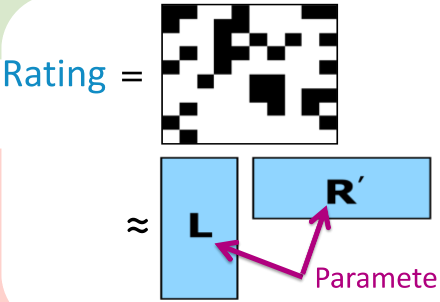
- Collaborative filtering
- Matrix factorization



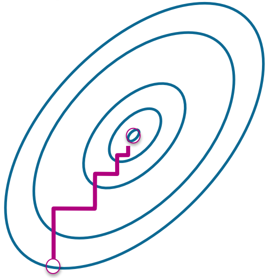
	Sunglasses	Baby Bottle	...	Diapers	Swim Trunks	Baby Formula
Sunglasses	1.00	0.03	...	0.02	0.23	0.04
Baby Bottle	0.03	1.00	...	0.09	0.04	0.12
...	...	...	...	...	...	...
Diapers	0.02	0.09	...	1.00	0.04	0.08
Swim Trunks	0.23	0.04	...	0.04	1.00	0.03
Baby Formula	0.04	0.12	...	0.08	0.03	1.00

Algorithms

- Coordinate descent



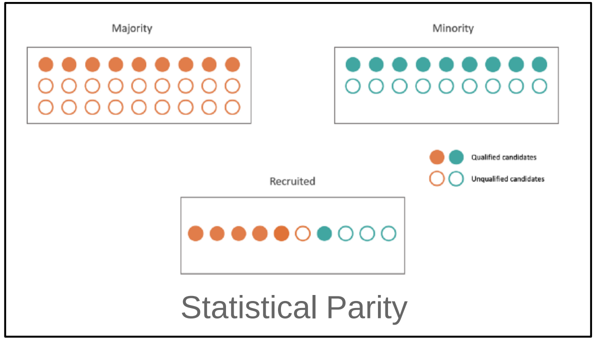
Parameters of model



Concepts

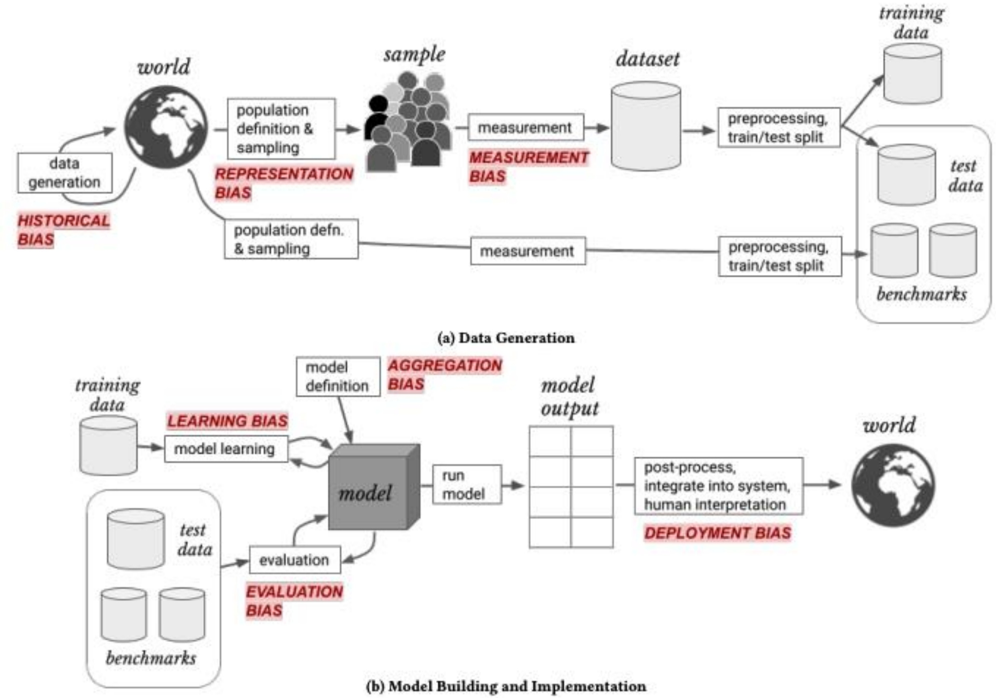
- Matrix completion, cold-start problem, co-occurence matrix, Jaccard Similarity

Bias & Fairness in ML



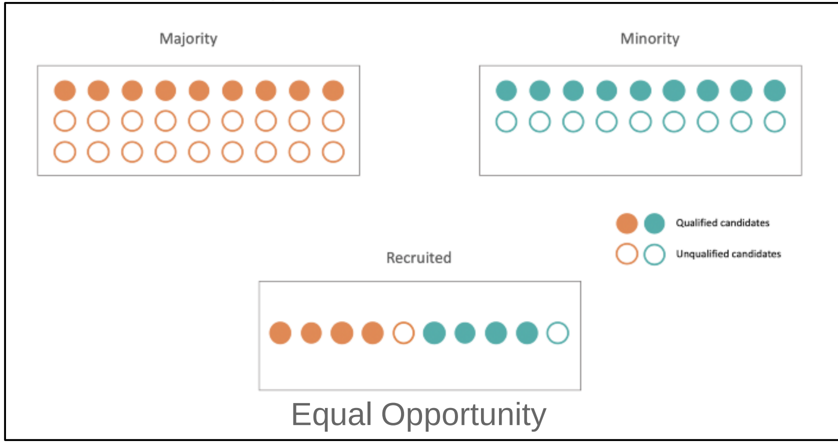
Statistical Parity

- Fairness Metrics:
  - Fairness through Unawareness
  - Statistical Parity
  - Equal Opportunity
- (Some) Potential Solutions:
  - Not developing the tech
  - Education ☺
  - More inclusive datasets
  - Incorporating Fairness Metrics into the Algorithm
  - Regulation



(a) Data Generation

(b) Model Building and Implementation



Equal Opportunity

Future Directions

Data Science courses offered at UW: https://escience.washington.edu/data-science-courses-at-the-university-of-washington/

A few directions of ML research that I'm excited by:

- FAccT (ACM Conference on Fairness, Accountability, and Transparency)
- Interpretability (how can we understand what deep networks are doing?)
- Interactive Learning, Online Learning
- Reinforcement Learning, Robot Learning
- Green AI, making learning more efficient
- ML for Healthcare, Computational Biology
- ML Education, training a generation of data scientists that are fluent in ethical & social considerations

Big Picture

Improving the performance at some task through experience!

- Before you start any learning task, remember fundamental questions that will impact how you go about solving it

What is the learning problem?

What model?

With what optimization algorithm?

How will you evaluate the model?

From what experience?

What loss function are you optimizing?

Are there any guarantees?

Who will it impact and how?

Congrats on finishing CSE/STAT 416!
Thanks for the hard work!

Brain Break

3:03