CSE/STAT 416

Course Wrap Up

Pemi Nguyen
Paul G. Allen School of Computer Science & Engineering
University of Washington

June 1, 2022

Administrivia

Final Exam

- Final Exam released next Monday 12pm and due Thursday 12 pm
- Final section review tomorrow

Please fill out the course evals! (for both your section TAs and me)

One Slide

Regression

Overfitting

Training, test, and true error

Bias-Variance tradeoff

Ridge, LASSO

Cross validation

Gradient descent

Classification

Logistic regression

Decision trees

Random Forest and Boosting

Precision and recall

k-Nearest Neighbor

Locality Sensitive Hashing

Document embeddings: TF-IDF, Bag of Words

Distance / Similarity metrics:

Unsupervised v. supervised

Dimensionality reduction, PCA

k-means clustering

Other forms of clustering

Recommender systems

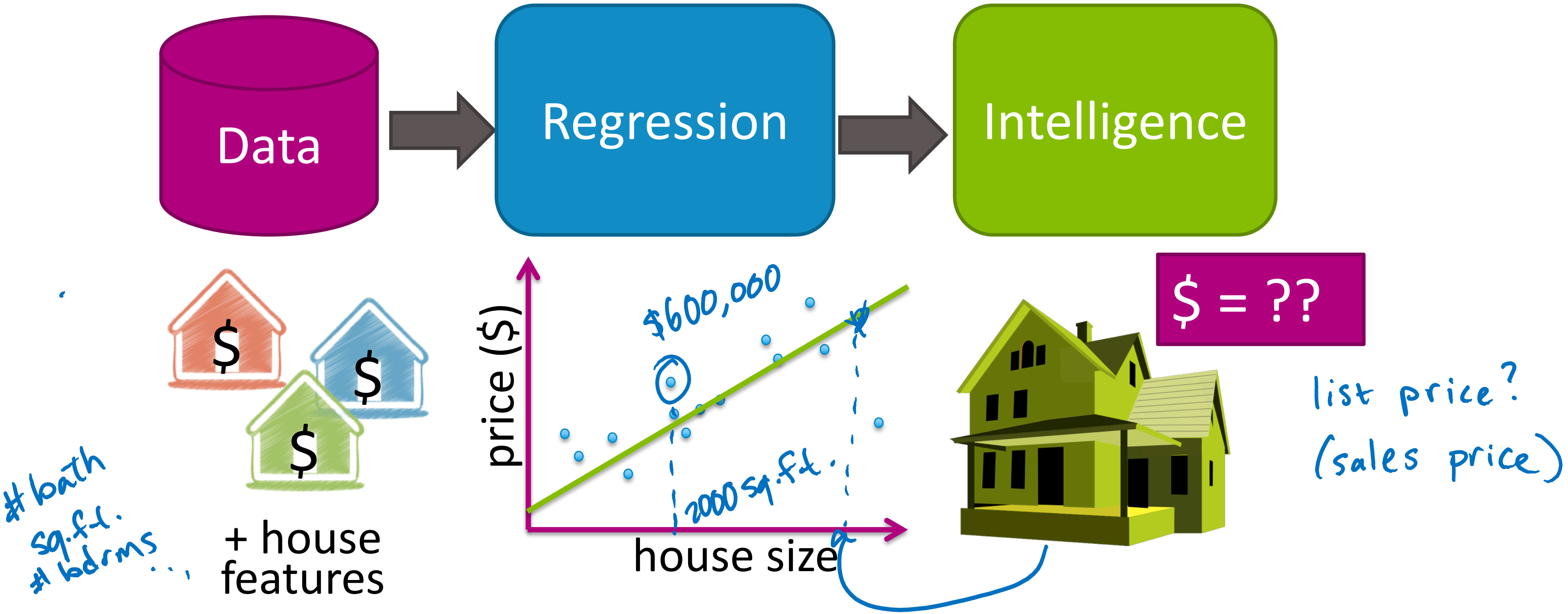
Matrix factorization

Neural networks

Convolutional neural networks

Transfer learning for deep learning

Case Study 1:
Predicting house prices



Regression

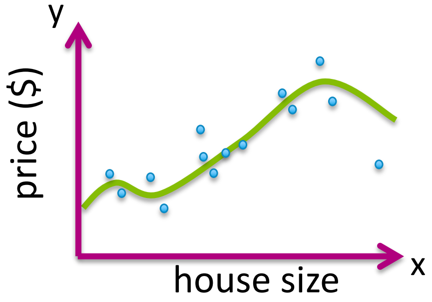
Case study: Predicting house prices

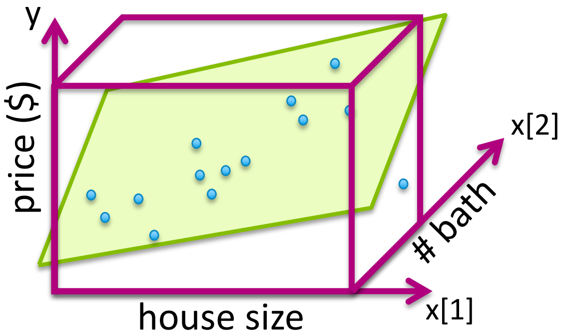
Models

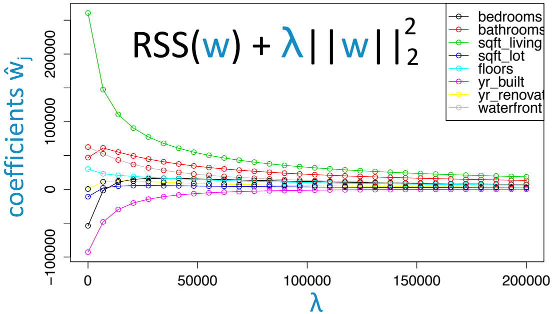
- Linear regression
- Regularization:
  Ridge (L2), Lasso (L1)

Including many features:

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- …







Regression

Case study: Predicting house prices

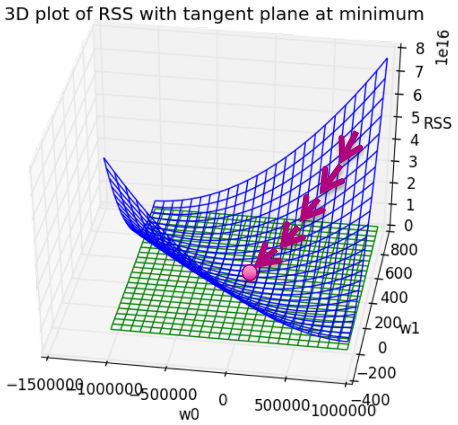
Algorithms

- Gradient descent

$$\text{RSS}(w_0,w_1) = (\$_{\text{house 1}}-[w_0+w_1 \text{sq.ft.}_{\text{house 1}}])^2 + (\$_{\text{house 2}}-[w_0+w_1 \text{sq.ft.}_{\text{house 2}}])^2 + (\$_{\text{house 3}}-[w_0+w_1 \text{sq.ft.}_{\text{house 3}}])^2 + \ldots \text{[include all houses]}$$

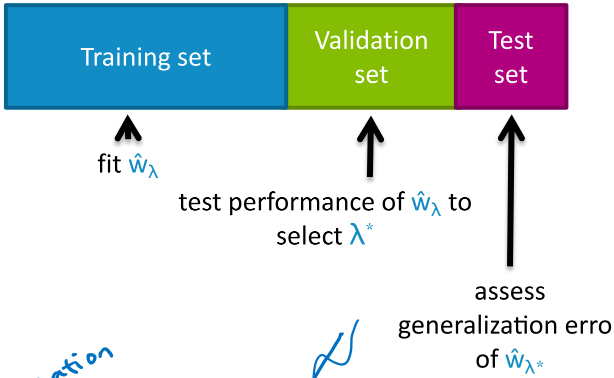
$$\downarrow$$

$$\hat{w}$$



Regression

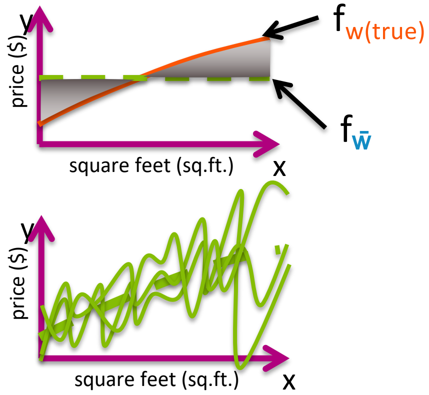
Case study: Predicting house prices

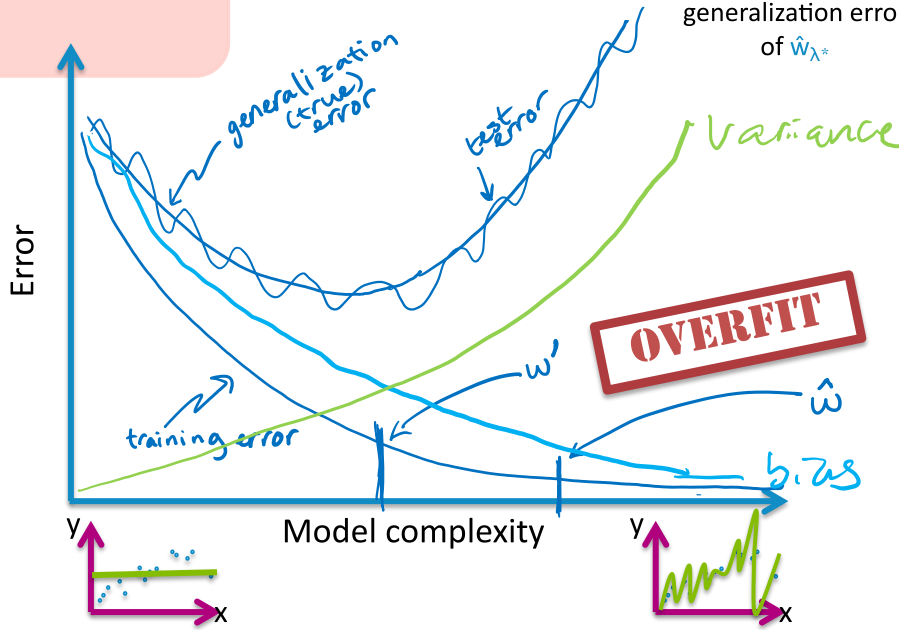


Concepts

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

1. Noise
2. Bias
3. Variance





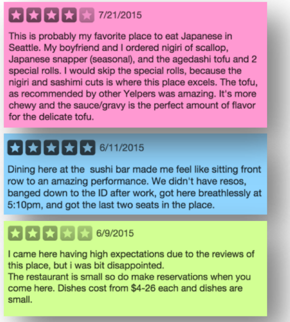
Case Study 2:
Sentiment analysis

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

stars / +/-

text

All reviews:



7/21/2015

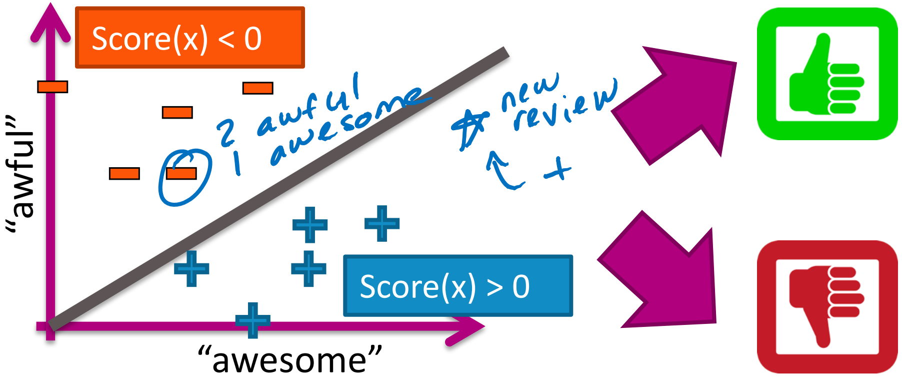
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended 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.

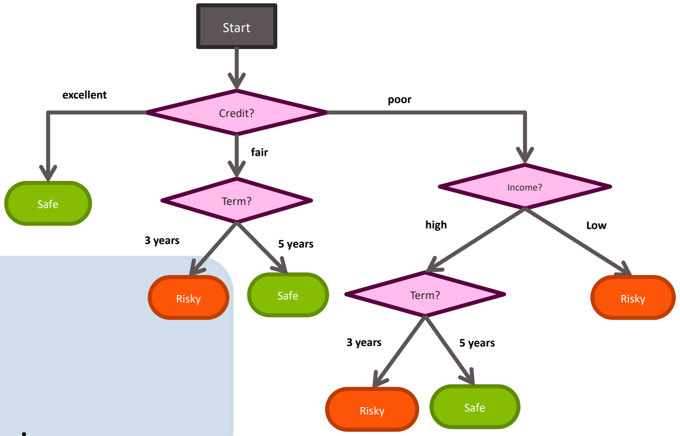
6/9/2015

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.



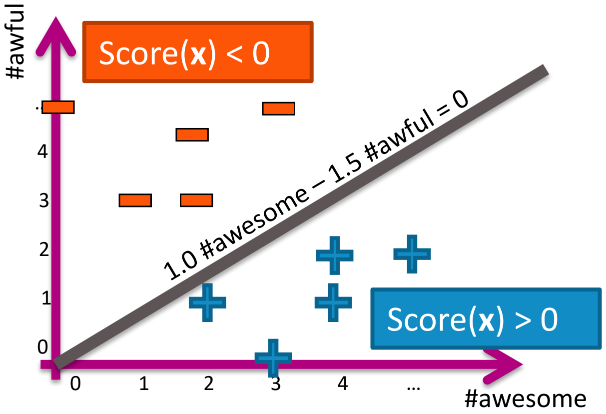
Classification

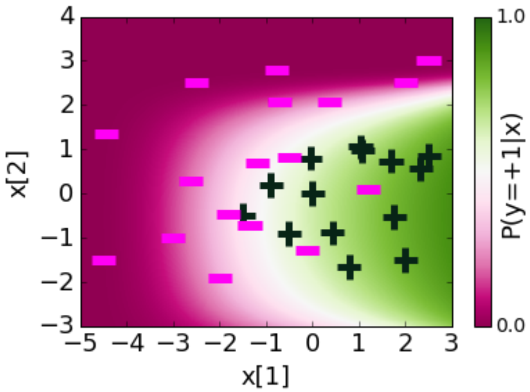
Case study: Analyzing sentiment



Models

- Linear classifiers (logistic regression)
- Multiclass classifiers
- Decision trees
- Boosted decision trees and random forests



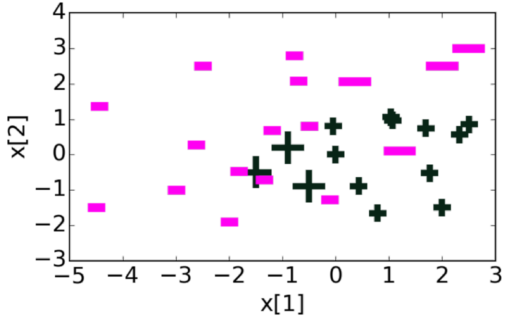


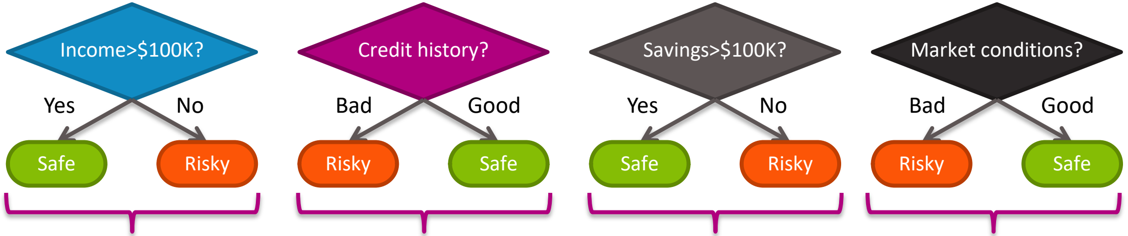
Classification

Case study: Analyzing sentiment

Algorithms

- Boosting
- Learning from weighted data α_i





weighted_error = 0.2

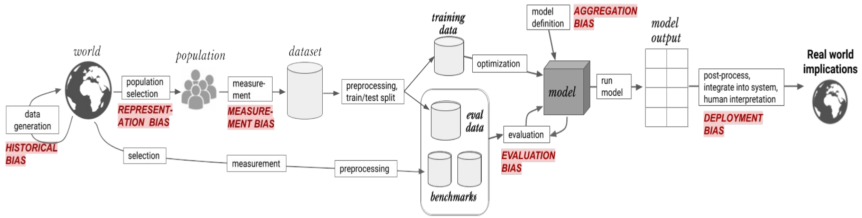
weighted_error = 0.35

weighted_error = 0.3

weighted_error = 0.4

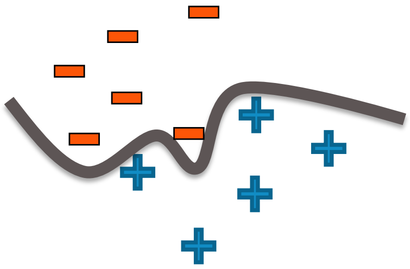
Classification

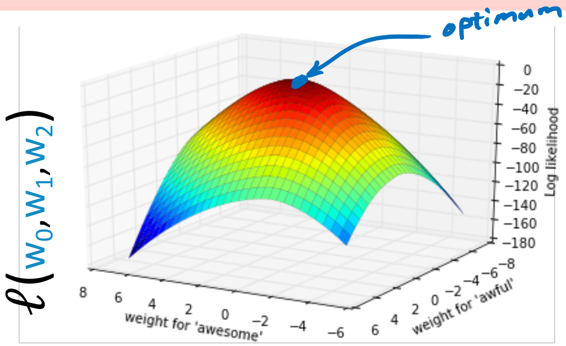
Case study: Analyzing sentiment

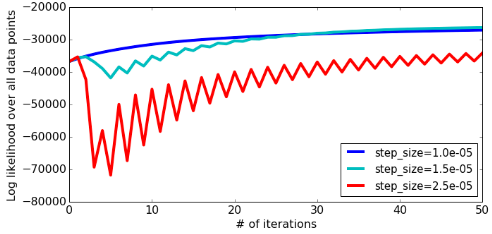


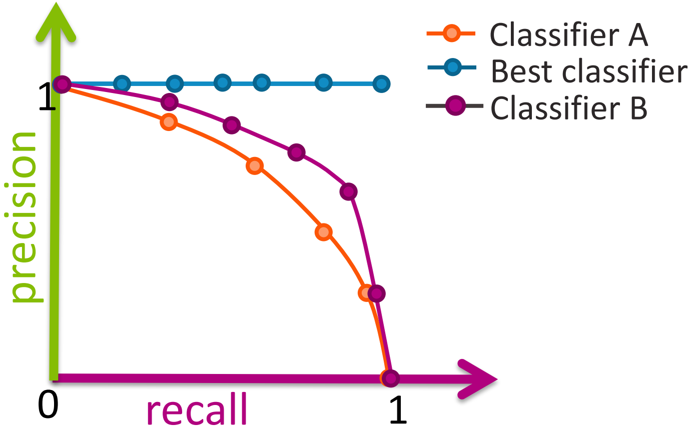
Concepts

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









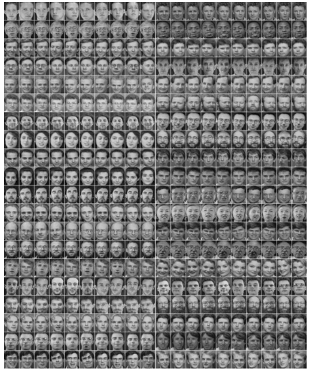
Case Study 3:
Document retrieval

Case Study 3+:
Document structuring for retrieval

Case Study 3++:
Dimensionality reduction

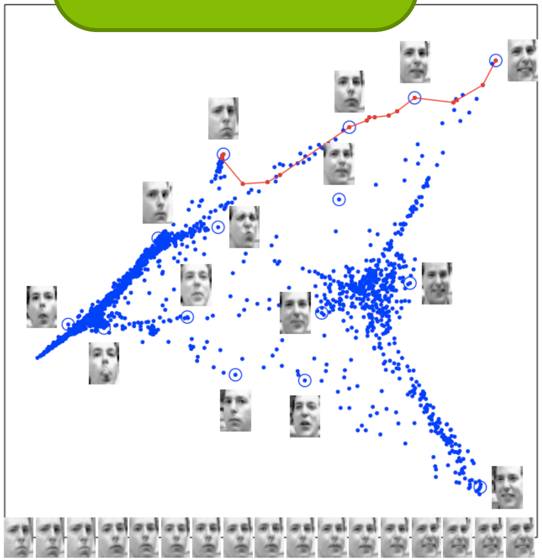
Data → PCA → Intelligence

simple



Images with thousands or millions of pixels

Can we give each image a coordinate, such that similar images are near each other?



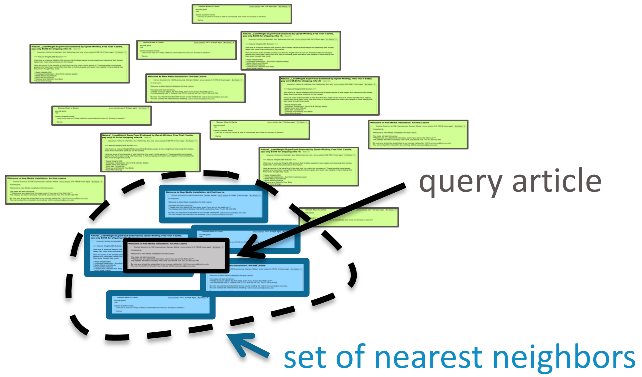
[Saul & Roweis ‘03]

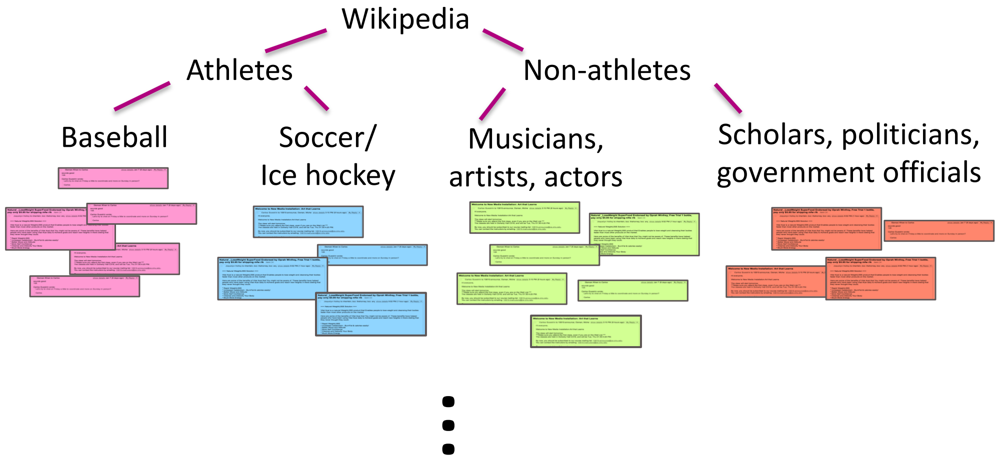
Clustering & Retrieval

Case study: Finding documents

Models

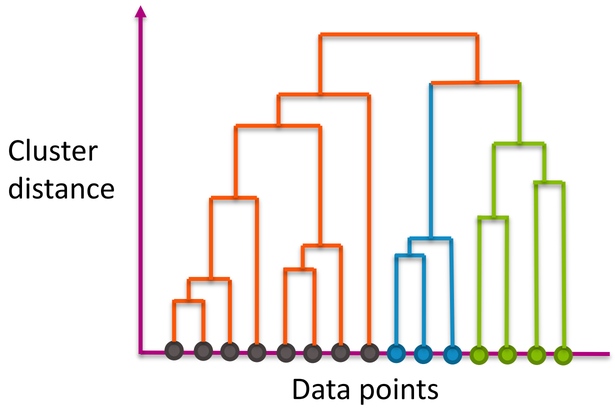
- Nearest neighbors
- Clustering
- Hierarchical clustering





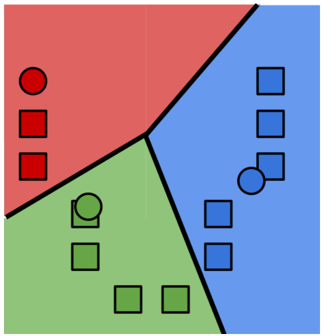
Clustering & Retrieval

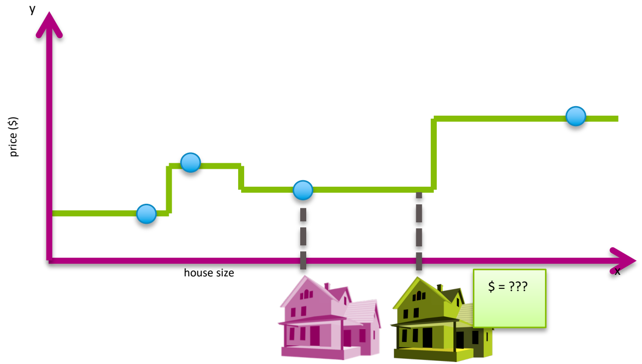
Case study: Finding documents

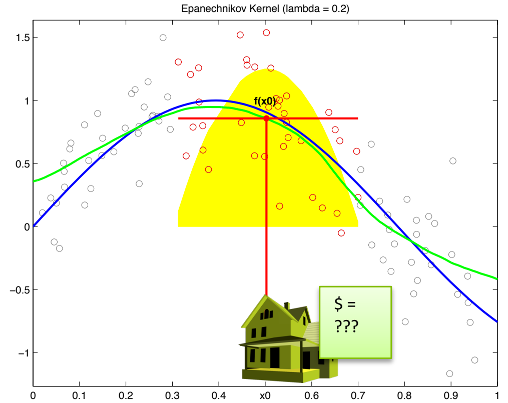


Algorithms

- k-means
- Locality-sensitive hashing (LSH)
- NN regression and classification
- Kernel regression
- Agglomerative and divisive clustering
- PCA





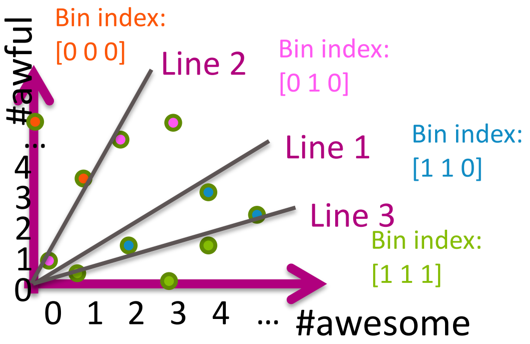


Clustering & Retrieval

Case study: Finding documents

Concepts

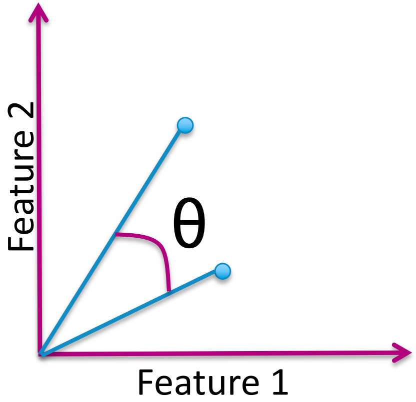
- Distance metrics, kernels, approximation algorithms, dimensionality reduction



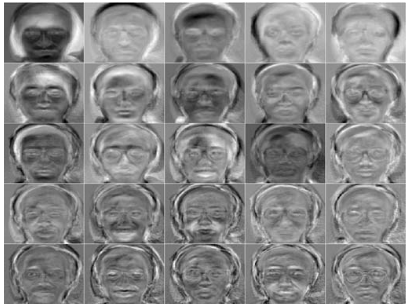
1 0 0 0 5 3 0 0 1 0 0 0 0

1*3
+
5*2
= 13

3 0 0 0 2 0 0 1 0 1 0 0 0



Principal components:

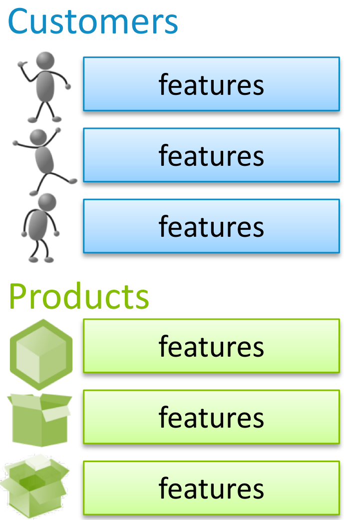


Reconstructing:

Case Study 4:
Product recommendation

Your past purchases:

+ purchase histories of all customers



Recommended items:

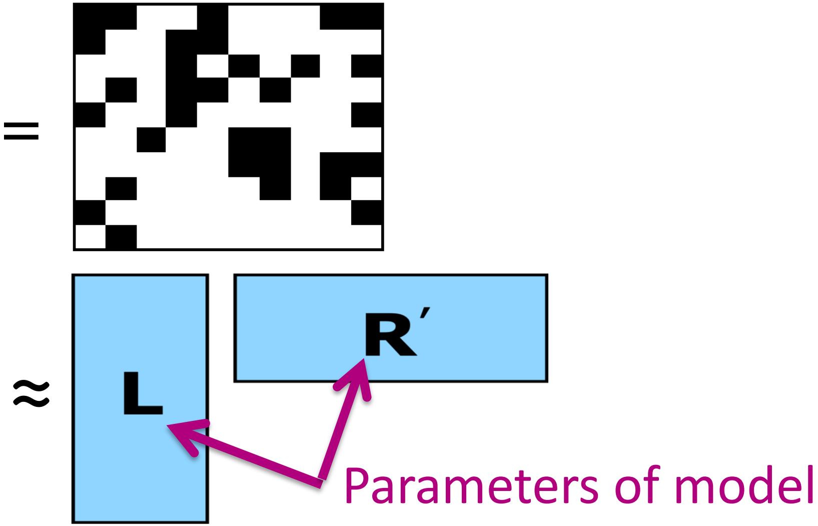
Recommender Systems & Matrix Factorization

Case study: Recommending Products

Models

- Collaborative filtering
- Matrix factorization

Rating =



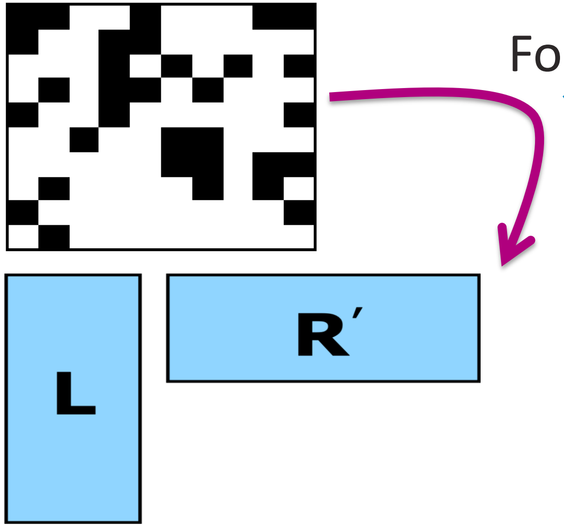
Recommender Systems & Matrix Factorization

Case study: Recommending Products

Algorithms

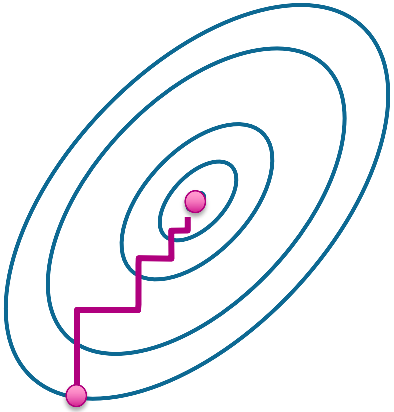
- Coordinate descent

Rating =



$\approx$

Form estimates $\hat{L}_u$ and $\hat{R}_v$

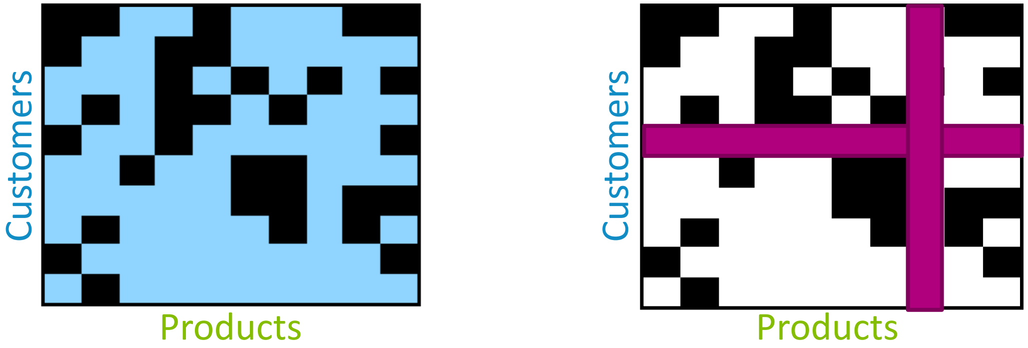


Recommender Systems & Matrix Factorization

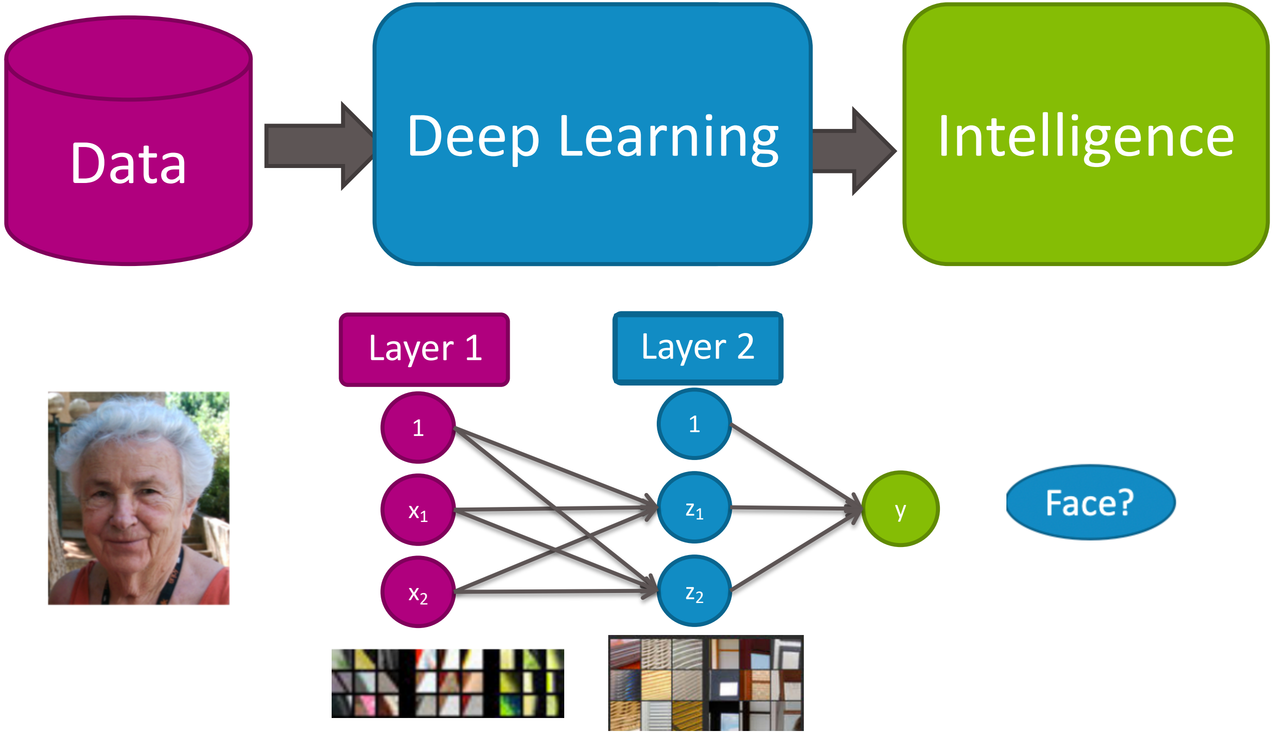
Case study: Recommending Products

Concepts

- Matrix completion, cold-start problem



Case Study 5:
Image classification

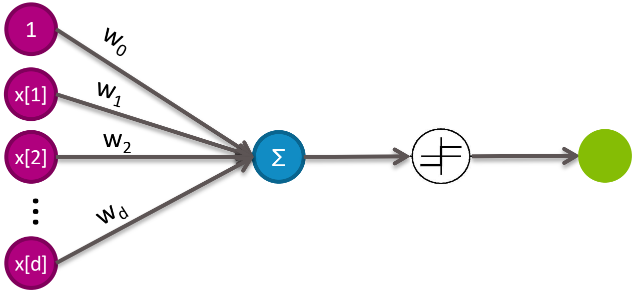


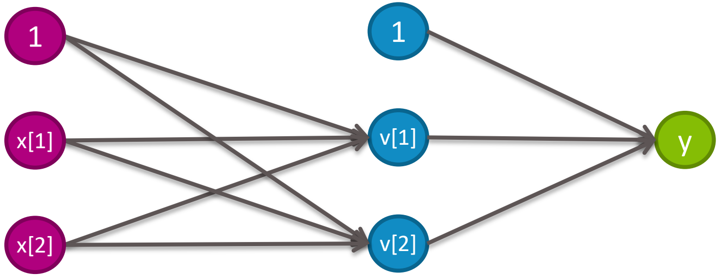
Deep Learning

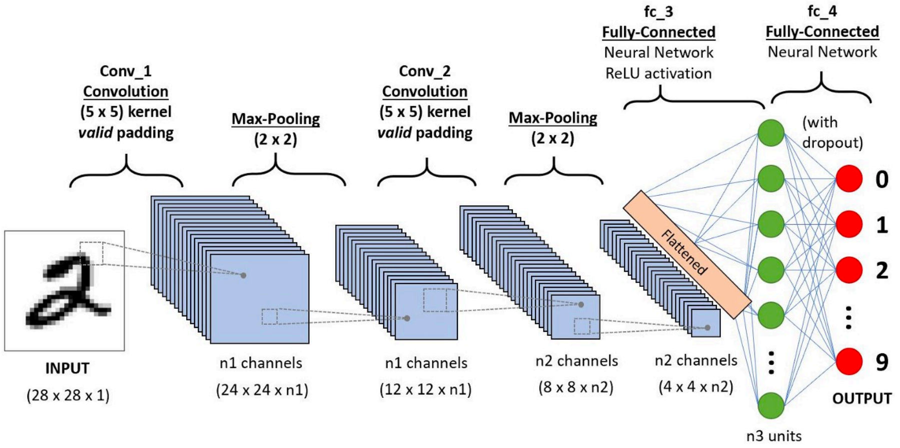
Case study: Image classification

Models

- Perceptron
- General neural network
- Convolutional neural network





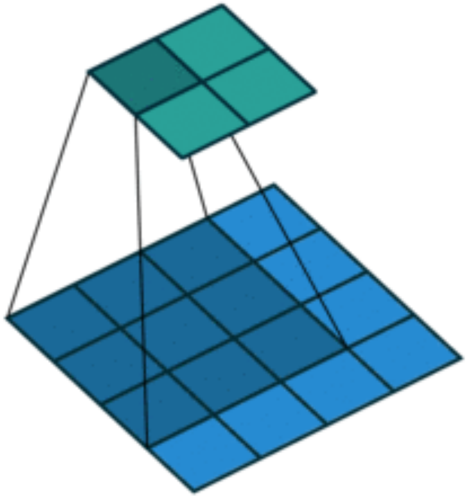


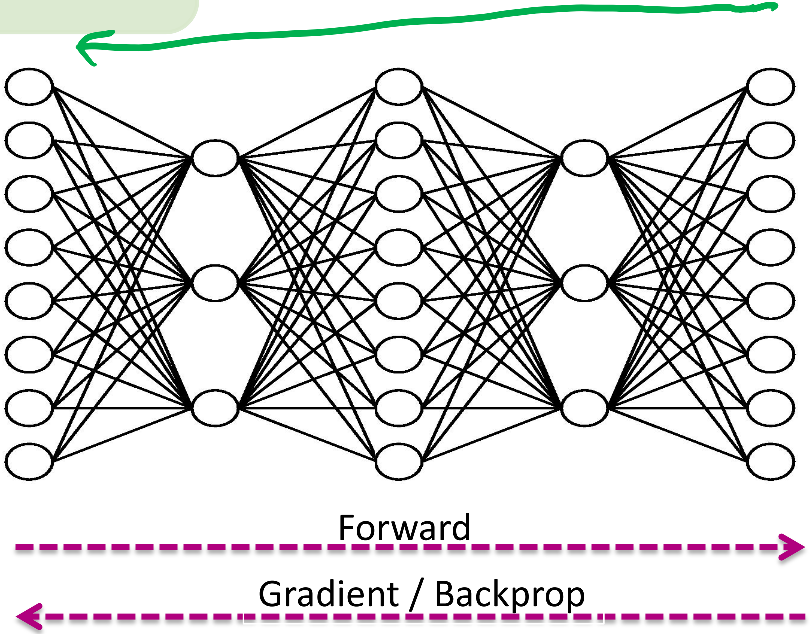
Deep Learning

Case study: Image classification

Algorithms

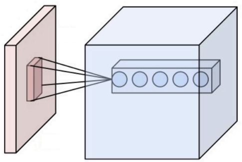
- Convolutions
- Backpropagation (high level only)

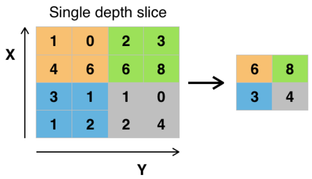


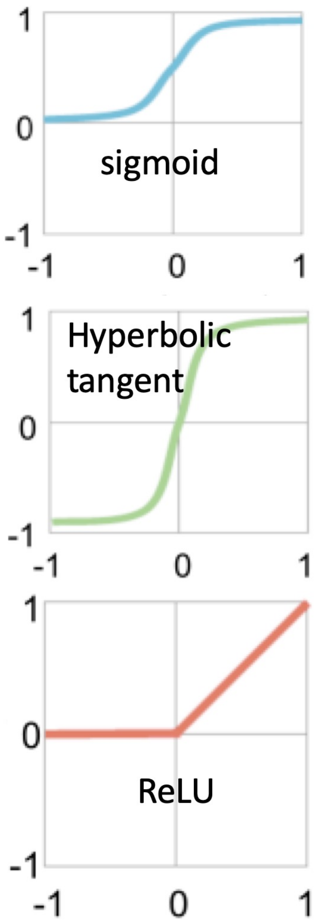


Deep Learning

Case study: Image classification

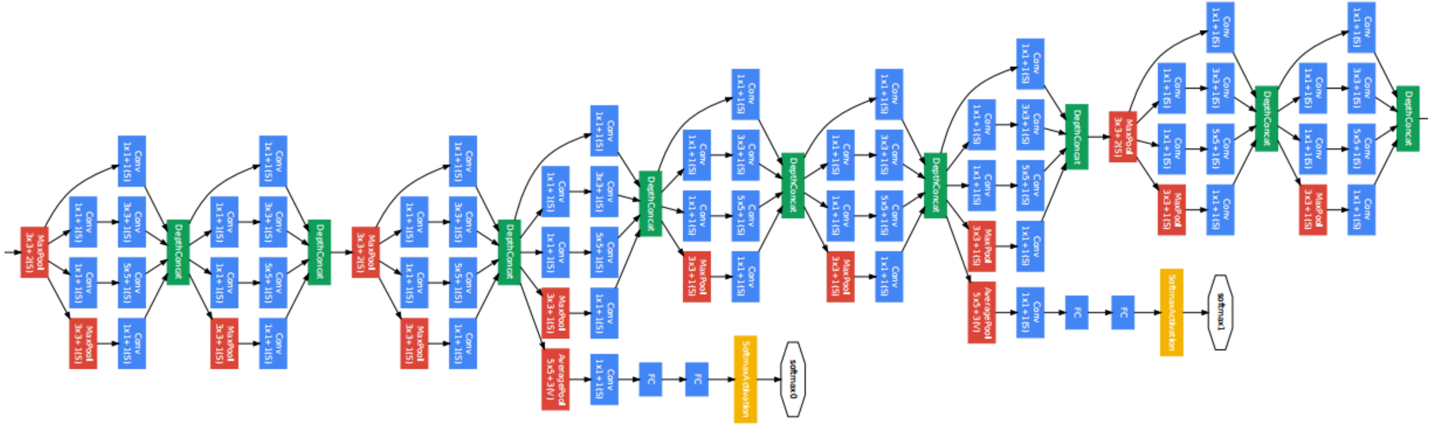


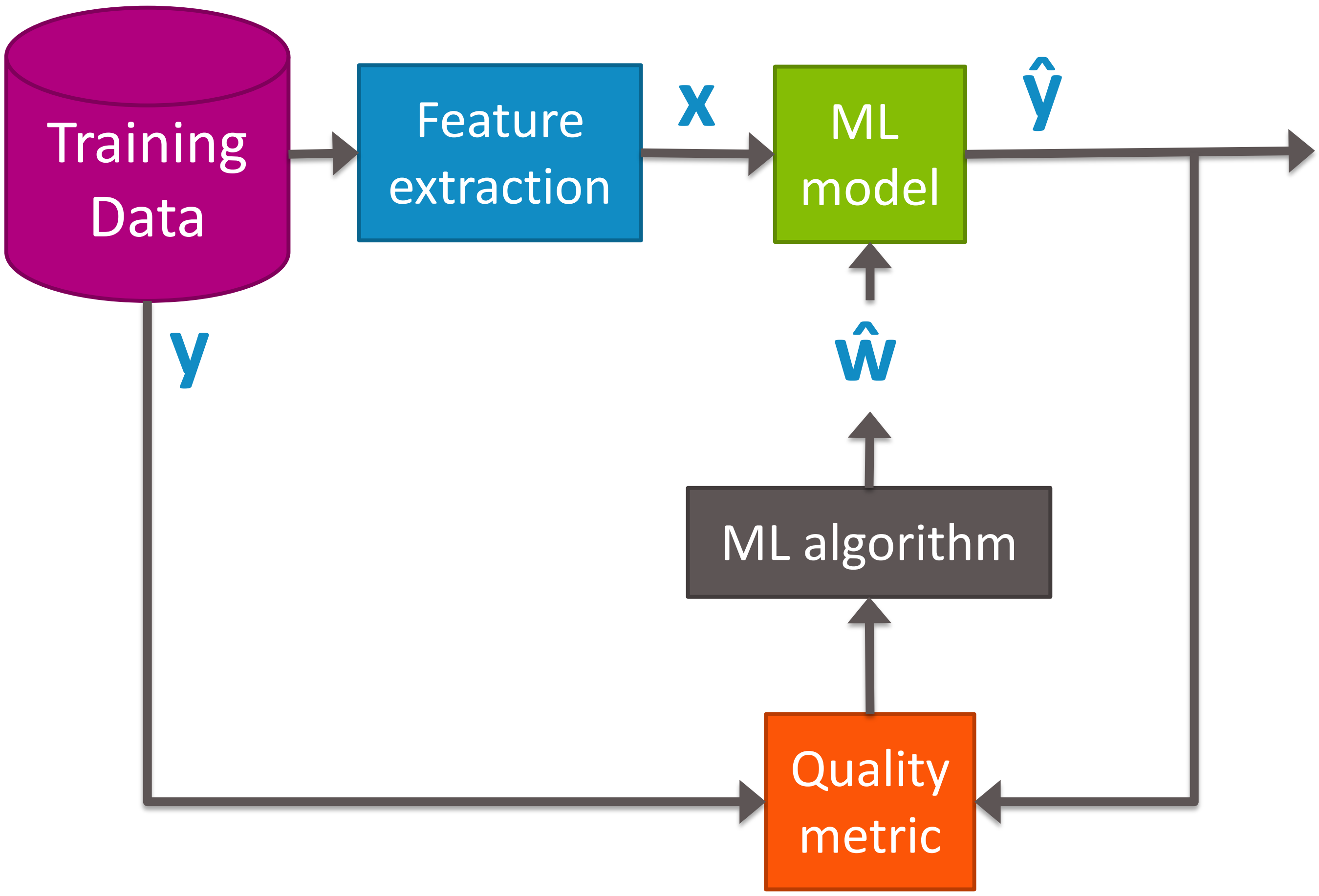




Concepts

- Activation functions, hidden layers, architecture choices





Training Data
Feature extraction
x
ML model
ŷ
y
ŵ
ML algorithm
Quality metric

Future Directions

Classes

There isn't a clear, "one right class" to take next! If you want to take course work, you can take anything that you are interested in to apply your ML knowledge there!

Fairly comprehensive list of data science class at UW: https://escience.washington.edu/data-science-courses-at-the-university-of-washington/

Some recommended classes in CSE:

- CSE 447 / 517: Natural Language Processing
- CSE 455: Computer Vision
- CSE 490 G1 / 599 G1: Deep Learning
- CSE 446 / 546: Machine Learning (super recommended, the more mathy version of this course)
- CSE 547: Machine Learning for Big Data
- Some other interesting but loosely related courses:
    - CSE 414 / 444: Database Management
    - CSE 412 / 442: Data Visualization

Future Directions

This is a (very insufficient) attempt to outline some interesting directions ML research is going. This list fails to provide breadth of coverage and depth of all the ways ML can be applied.

Something not showing up in this list doesn't mean ML can't be used for that task! I'm just one opinion about what I'm excited about in ML!

Applied ML

ML applied to basically any problem we might care about (and the tough challenges that come with that)

- Natural Language Processing (NLP)
- Computer Vision
- Computational Biology
- Medical Imaging / Health

ML Systems

Construed broadly, trying to build systems to efficiently implement and scale ML models.

Hardware: GPU

Energy Efficiency:

- GreenAI
- TinyML

Distributed Systems to store and deploy ML models, to solve big data challenges

ML Theory

Building foundational understanding for why/how ML works.

Machine Learning from a statistical perspective

- E.g. Analyzing a lot of concepts like overfitting, bias / variance mathematically

Theory of Deep Learning

- Video on Universal Approximation Theorem youtube.com/watch?v=Ijqkc7OLenI&t=120s

Optimization (convex and non-convex)

And more!

Interactive Learning / Reinforcement Learning

How do we design models that interact with the environment?

Examples: Self driving cars and robotics

Video on AlphaGo: https://www.youtube.com/watch?v=8dMFJpEGNLQ&t=1s

Areas of study:

- Interactive Learning: Multi-armed bandits
- Reinforcement Learning: Q-learning, deep reinforcement learning

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?

From what experience?

What model?

What loss function are you optimizing?

With what optimization algorithm?

Are there any guarantees?

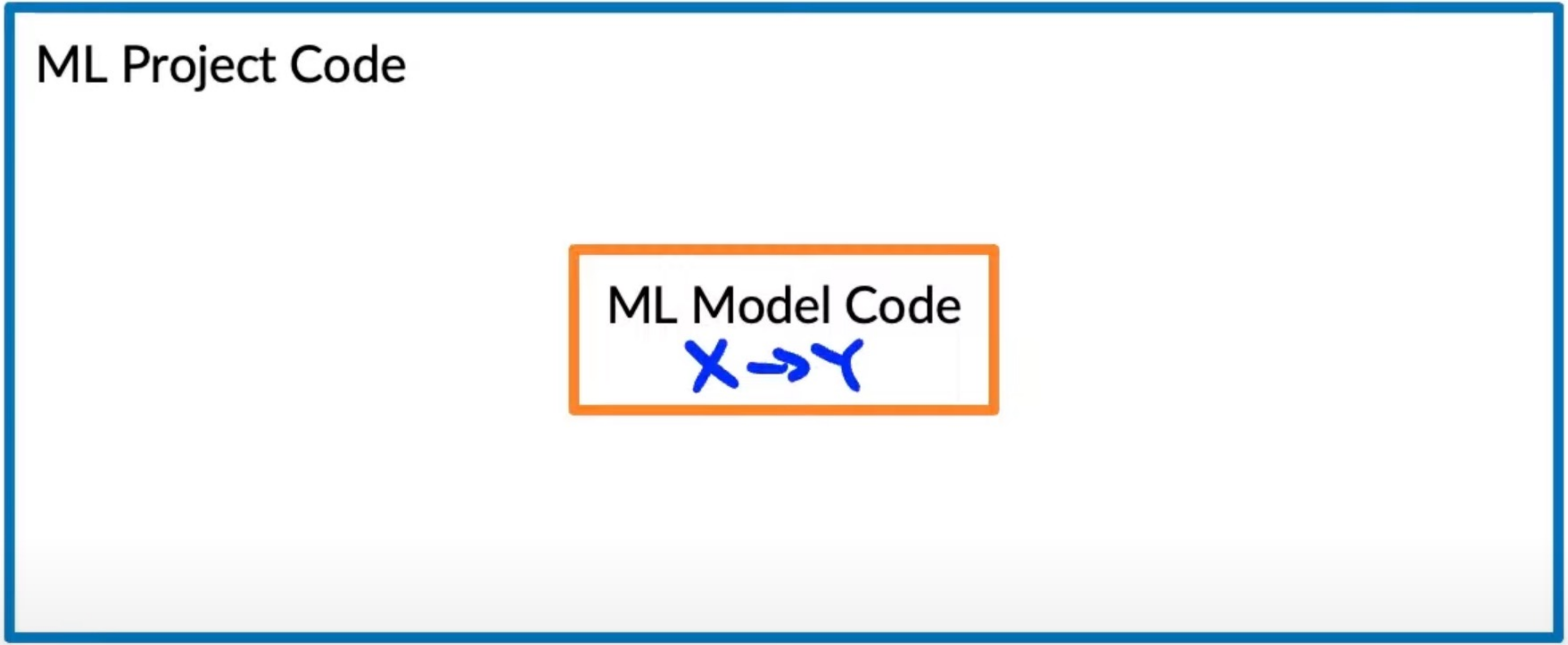
How will you evaluate the model?

Who will it impact and how?

MLOps

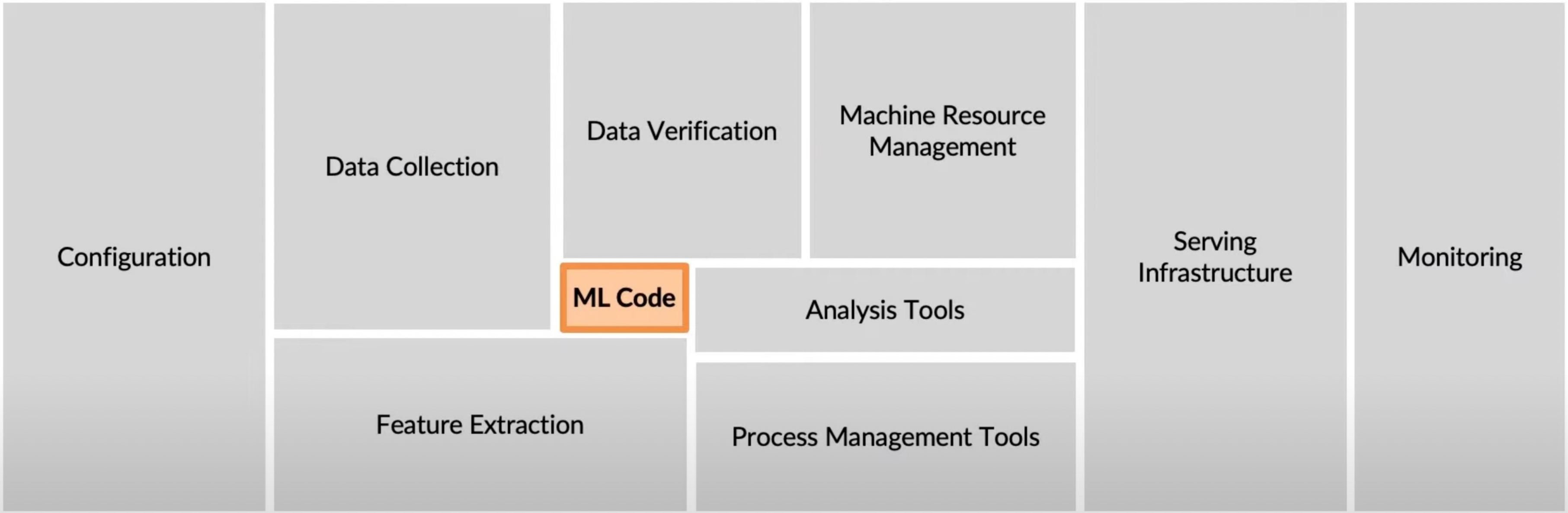
Source: Machine Learning Engineering for Production (MLOps) Specialization, Andrew Ng

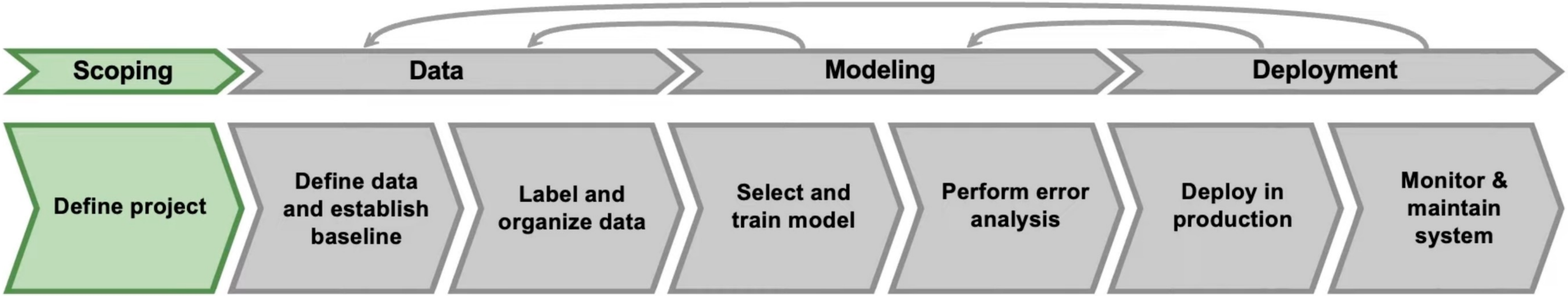
ML in production



Source: Machine Learning Engineering for Production (MLOps) Specialization, Andrew Ng

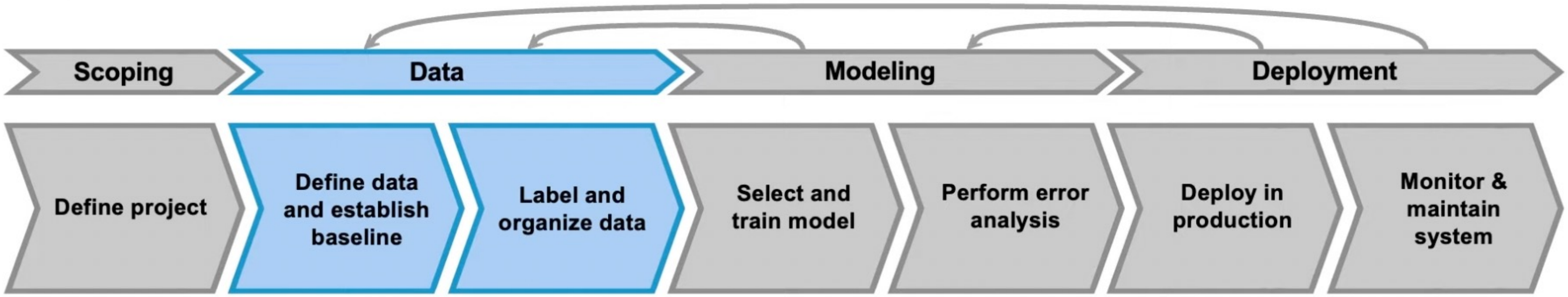
ML infrastructure





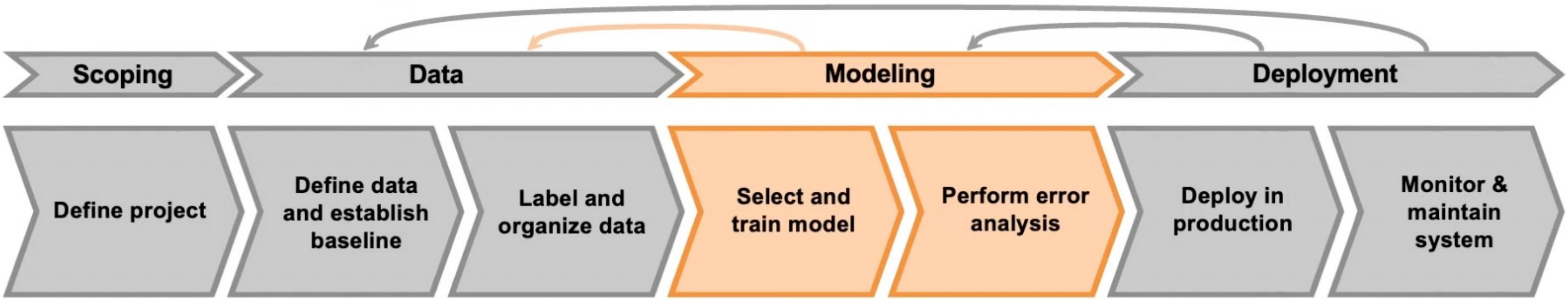
- What tasks do you plan to work on
- Key metrics: Accuracy, throughput / latency
- Estimated computing resources
- Assess feasibility of the solutions (by looking at benchmark papers/ open-source implementations)
- Timeline

Source: Machine Learning Engineering for Production (MLOps) Specialization, Andrew Ng



- Type of data (structured vs unstructured data such as texts, images, audio files)
- Is the data labelled consistently? Labelers must agree on a standard procedure for labeling to ensure uniformity
- Where to get data (open-source datasets are an option if you don't have financial resources)
- Establish a baseline model
- Data-centric approach vs model-centric approach (the model-centric approach focuses on improving the algorithm, code, and model architecture used to train the model, whereas the data-centric approach where the focus is to develop systematic engineering practices for improving data in ways that are reliable, efficient, and systematic)

Source: Machine Learning Engineering for Production (MLOps) Specialization, Andrew Ng



- Basically what we do in this class ☺
- Find suitable models for your tasks
- Fine tuning hyperparameters and analyze train / validation / test metrics

Shadow deployment: In this type of deployment, the model is deployed but the final decision is taken by a human, irrespective of what the model predicts. This is done usually to gauge how well the model is doing and where it is failing.

Source: Machine Learning Engineering for Production (MLOps) Specialization, Andrew Ng

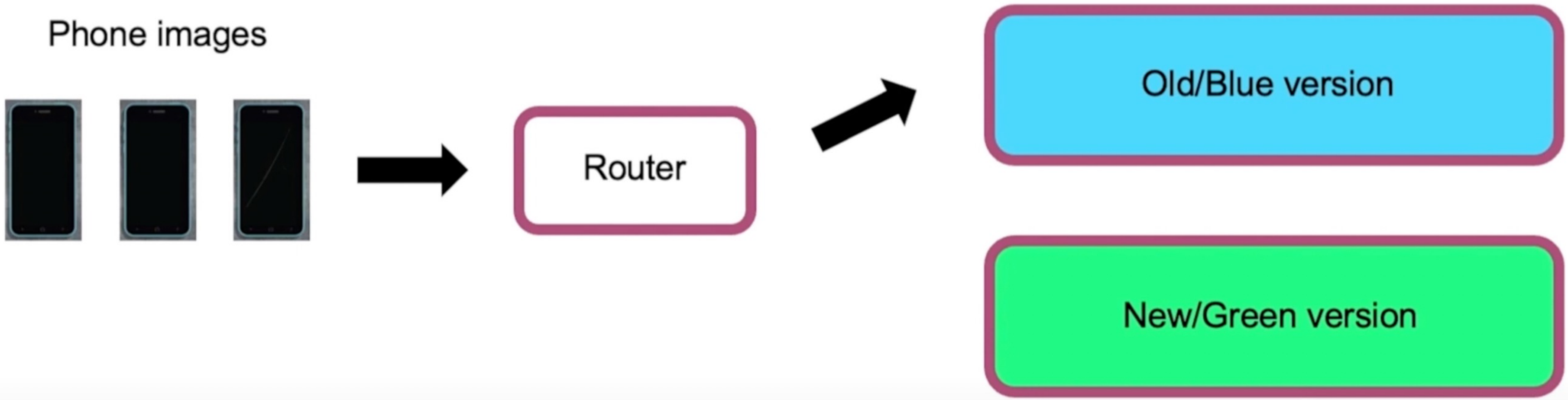
Canary deployment:

- Roll out to small fraction of the traffic initially
- Monitor the system and adjust the traffic depending on the model performance

Source: Machine Learning Engineering for Production (MLOps) Specialization, Andrew Ng

Source: Machine Learning Engineering for Production (MLOps) Specialization, Andrew Ng

Blue Green deployment: The traffic is gradually migrated from the old or *blue* version to the new or *green* version. This helps prevent any sort of downtime and in case of any bugs/errors, the application can easily be rolled back to the previous stable version or the blue version.



Concept Drift

Concept drift in machine learning and data mining refers to the change in the relationships between input and output data in the underlying problem over time.

This happens especially during COVID when a lot of economic and social patterns have undergone significant changes.

Model performance might deteriorate over time. Thus, ML model deployment is an iterative process.

We need to continuously update models once new data instances are received Or update models once a concept is detected.

Final words

This course is just covers high-level ideas about different machine learning models. There is way a lot more to this field.

Be aware of (sensationalized) online articles on Medium. Due to the popularity of Machine Learning these days, there might be an influx of incorrect information from those without expertise in ML. I always recommend you to read peer-reviewed articles from top conferences to confirm the information you are receiving.

This is the first time I'm teaching, and I'm aware that I might or might not live up to your expectations, and there is a lot I need to improve. However, I hope you're inspired by the marvelous wonders that this field can bring. I hope you enjoy this field as much I do.

I'll work in the industry after the course ends. Feel free to reach out to me at peminguyen0805@gmail.com if you wish to stay in touch or ask for advice on jobs/ research / career / pretty much anything ☺! Hopefully this will not be the last time we see each other.

Thanks for making my teaching experience memorable!

Congrats on finishing CSE/STAT 416!
Thanks for the hard work!
Good luck on the final exam next week ☺!