CSCI4622 Machine Learning 2020 Spring

Introduction

Geena Kim

University of Colorado **Boulder**

Class Info & Logistics

Class Meeting Time & Location

(In-Person) MWF 2-2:50 pm MST in Eaton Humanities Bldg 1B80

(Distance) https://cuboulder.zoom.us/j/972787367

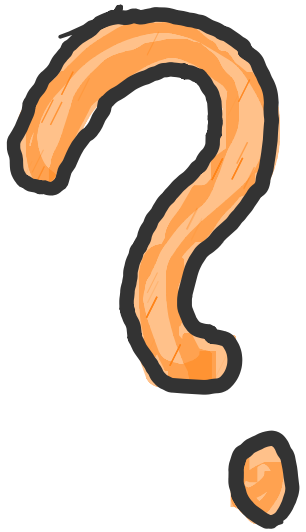
Course Staffs

Instructor: Geena Kim

Teaching Assistant: Shruthi Sukumar

Grader: Abhinivesh Palusa

The BUZZ words



Data Science

Artificial Intelligence

Machine Learning

Deep Learning

Data Science

- Interdisciplinary field concerning Data
- Almost anything to do with Data:

  (e.g.) Data Pipelining, Data Munging, Data Analysis, EDA…
- Includes Soft or Hard science, Small or Big data
- On the Job, Data Scientists do…

What does a Data Scientist do?

Data Scientists are responsible for the collection, cleaning and munging of data to meet the company's purpose. Duties vary according to the industry and may include experimental frameworks for product development and machine learning with the aim to lay a strong data foundation for robust analytics to be performed.

source: indeed.com

Artificial Intelligence

- One of the oldest and core subjects in CS
- About Problem-solving with Intelligence
- Theoretical and Practical
- On the Job, AI engineers/experts need skills..

Math and programming skills, Problem-solving skills, Machine learning....

and work on...

Build and deploy an AI system, build ML models, NLP, Robotics, Computer Vision...

Machine Learning

- Subfield of Artificial Intelligence
- About various statistical models and learning algorithms and training from Data
- Supervised/Unsupervised Learning
- Modern version of Statistical Learning
- On the Job, ML engineers do..

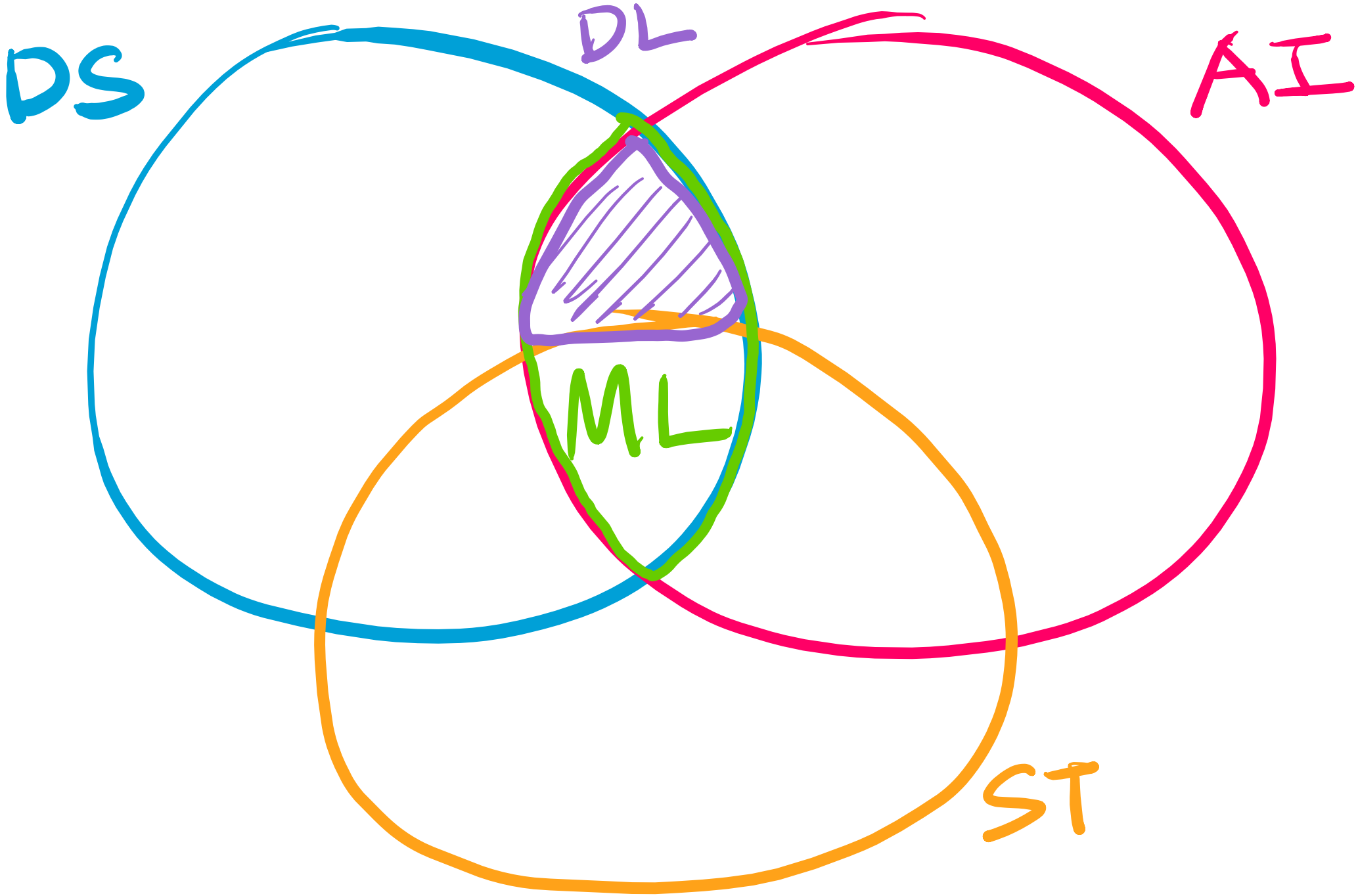
  Develop and test ML models, Design ML experiments, build ML system….

Deep Learning

- About Neural Network models, model training from Data, and all kinds of training techniques
- Subfield of Artificial Intelligence
- Subfield of Machine Learning
- On the Job, DL engineers do..

Solving complex technical challenges in various areas of deep learning such as object detection, segmentation, video understanding, sequence prediction, adaptive computing, memory networks, reduced precision training and inference, graph compilers, reinforcement learning, search distributed and federated training, and more

source: NVIDIA


DS
DL
AI
ML
ST

Good Time to Learn Machine Learning!

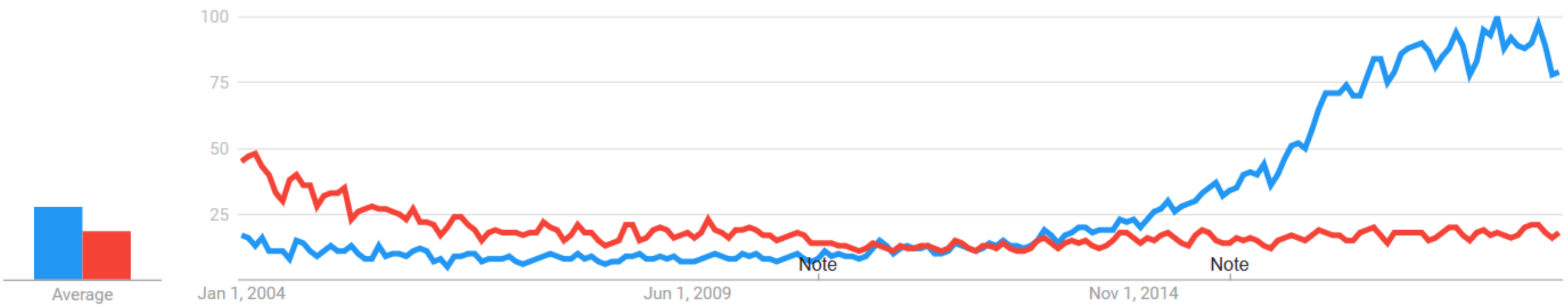
Machine learning
Field of study

Software engineering
Field of study

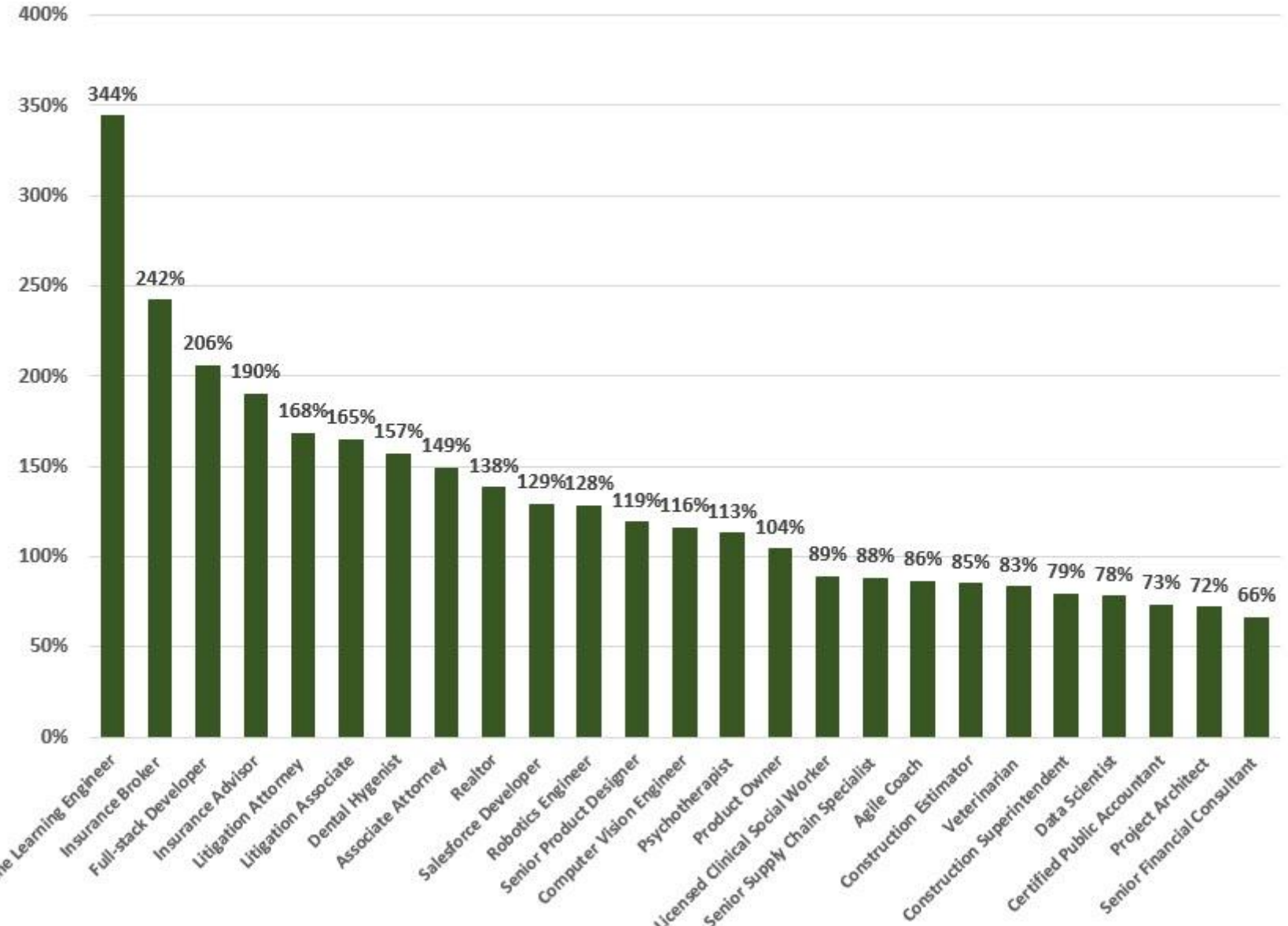
+ Add comparison

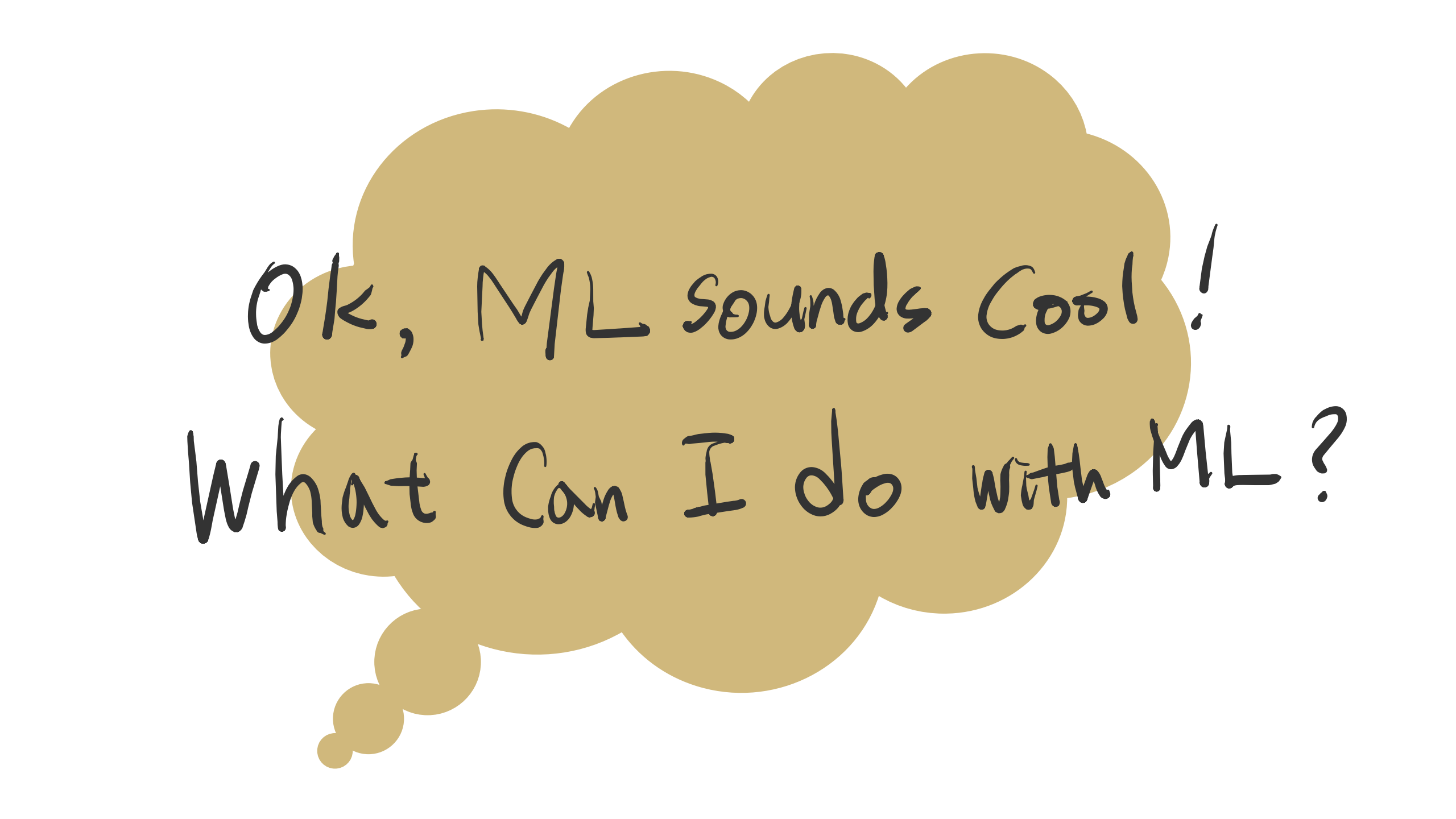
United States | 2004 - present | All categories | Web Search

Interest over time



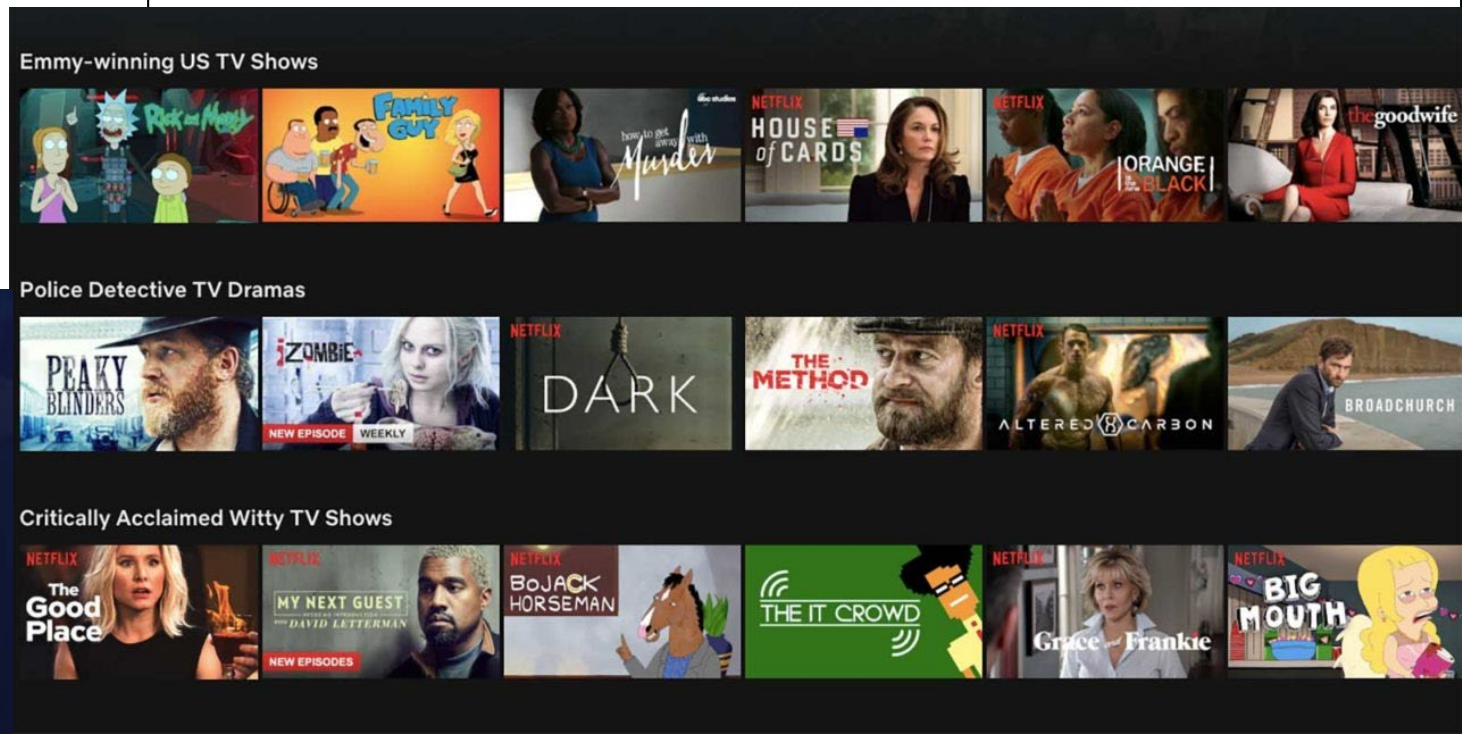
Indeed's Best Jobs In The U.S.
% Growth in # of postings, 2015 - 2018
March 14, 2019



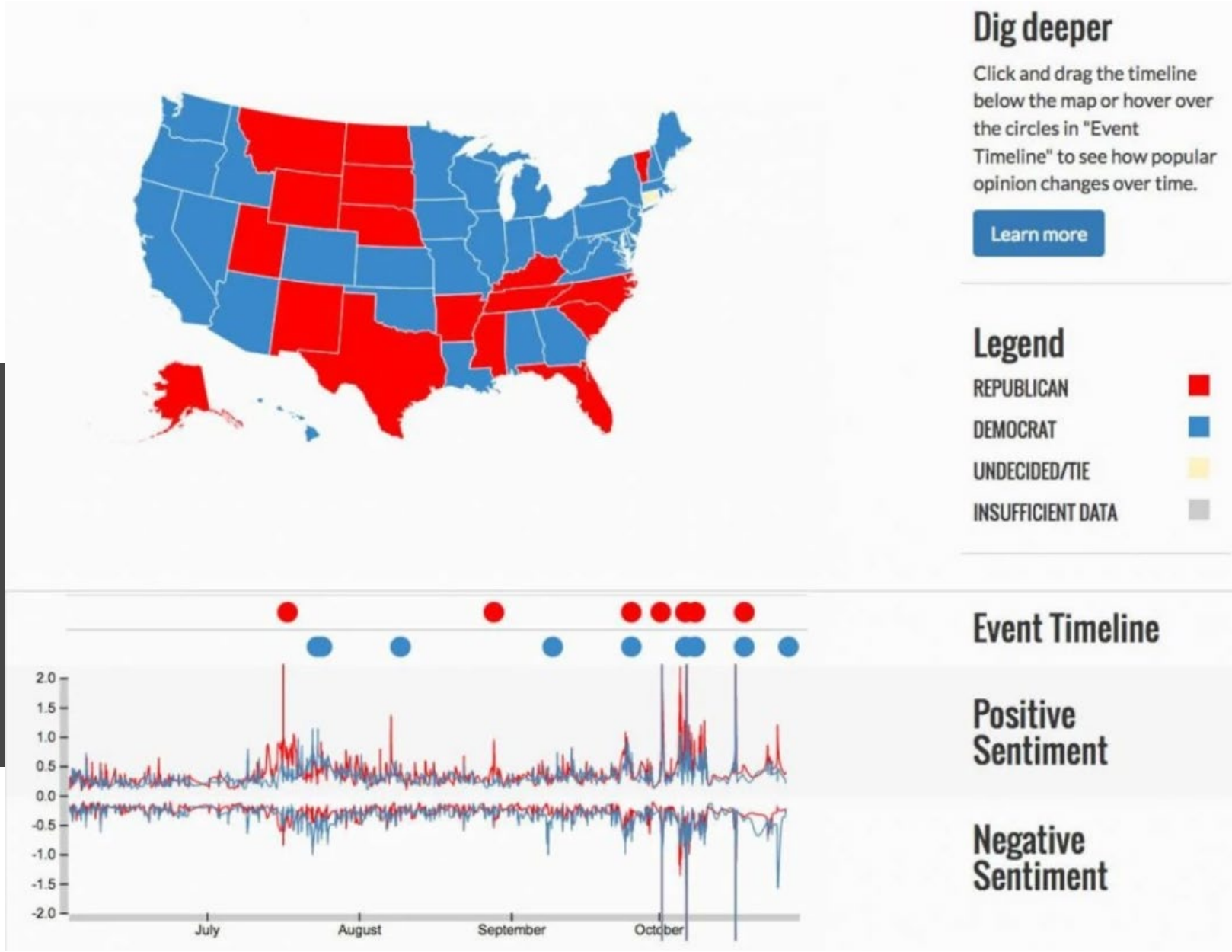

Ok, ML sounds Cool!
What Can I do with ML?

Machine Learning is Everywhere

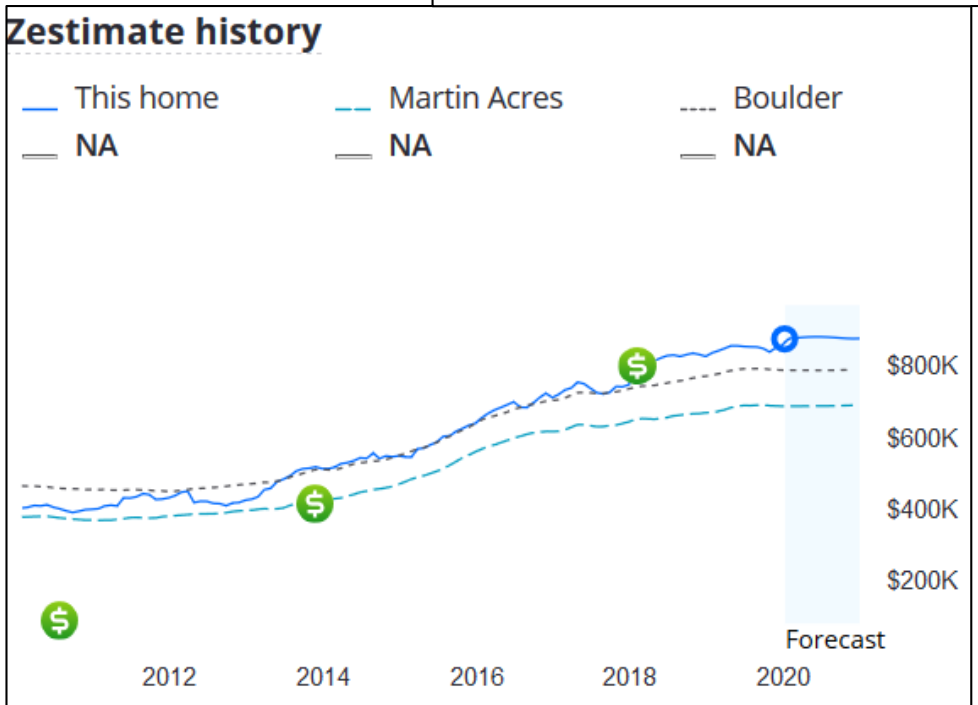
Product Recommendations



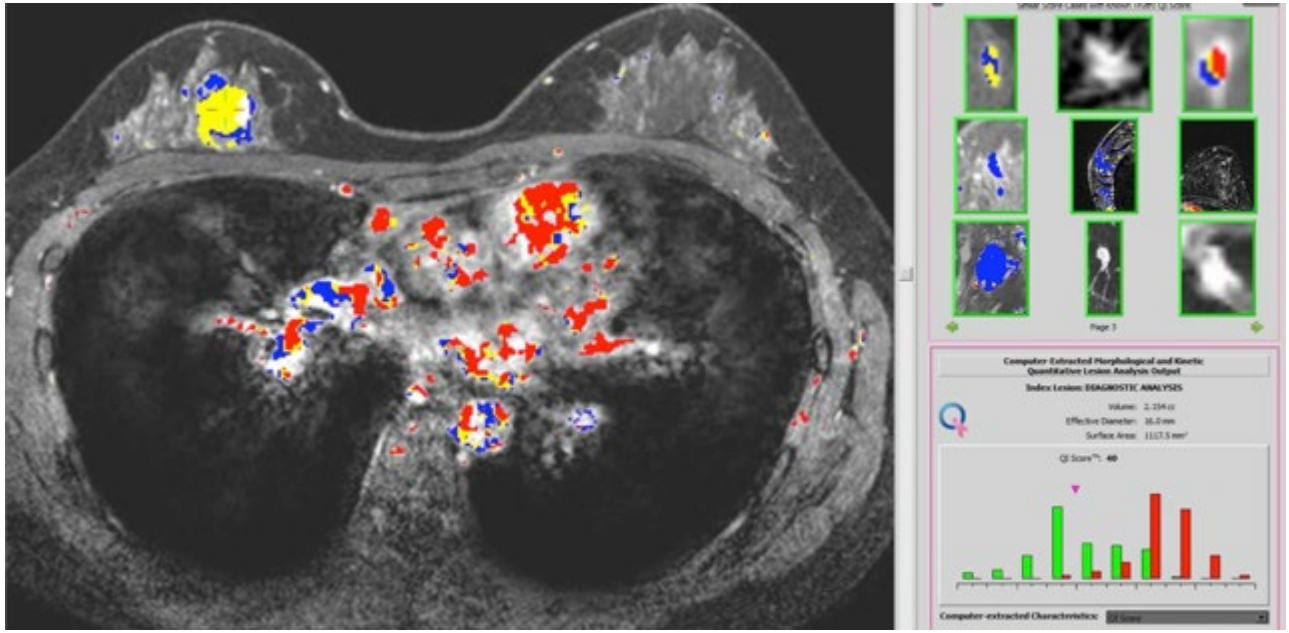
Sentiment Analysis

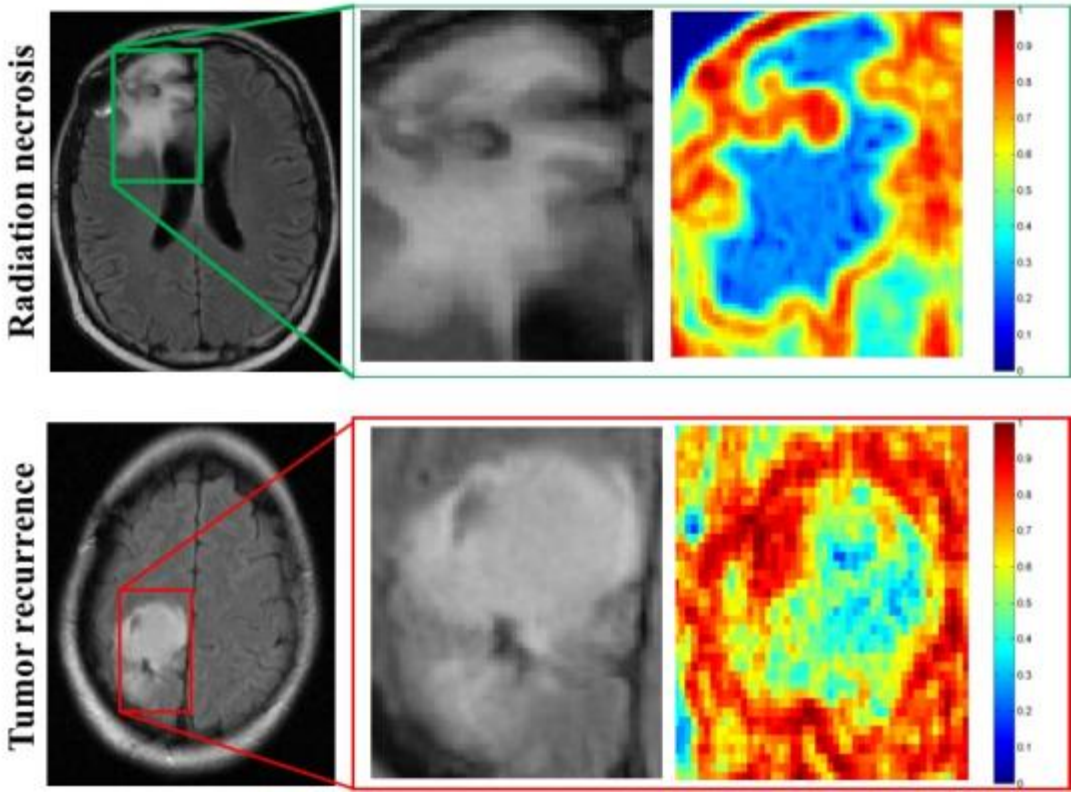


Price Forecasting

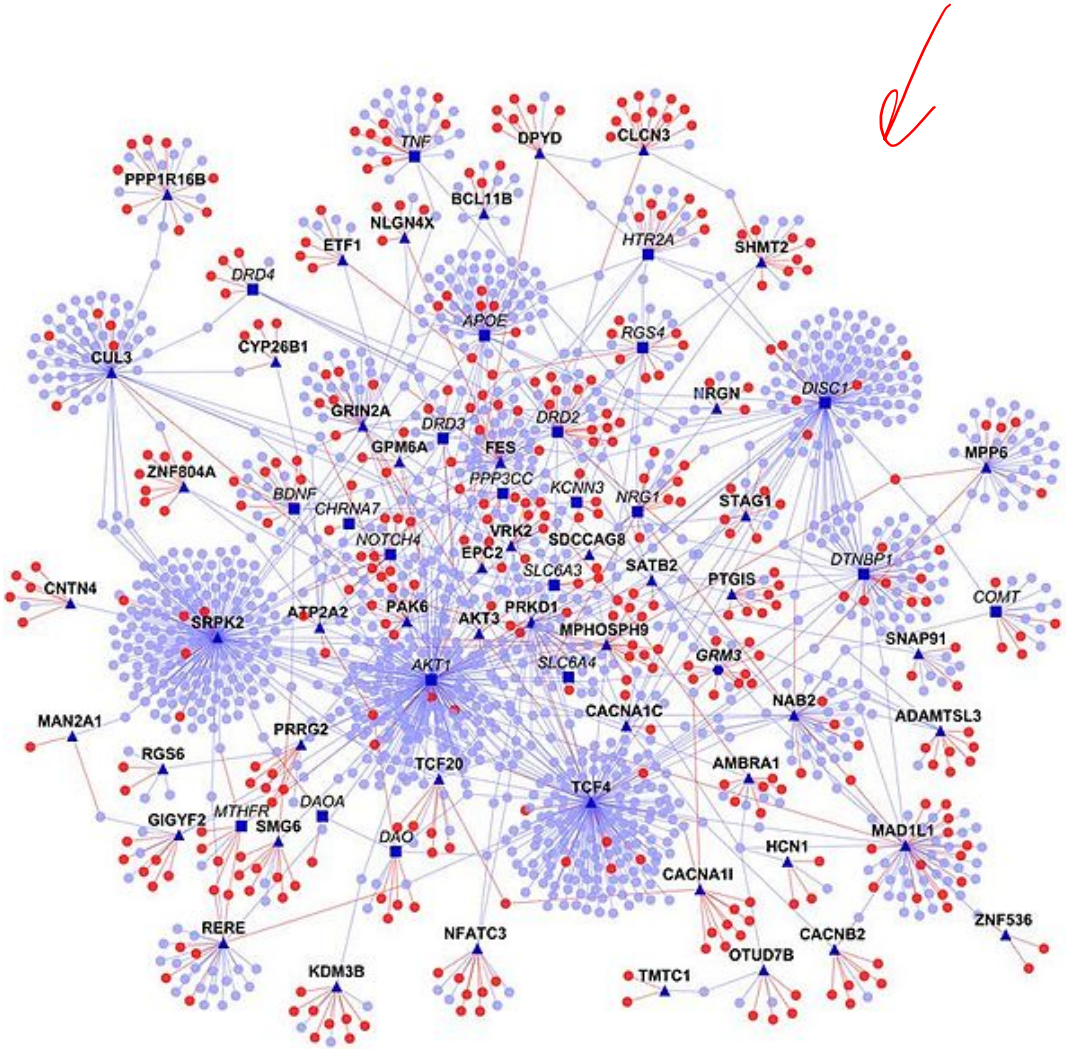


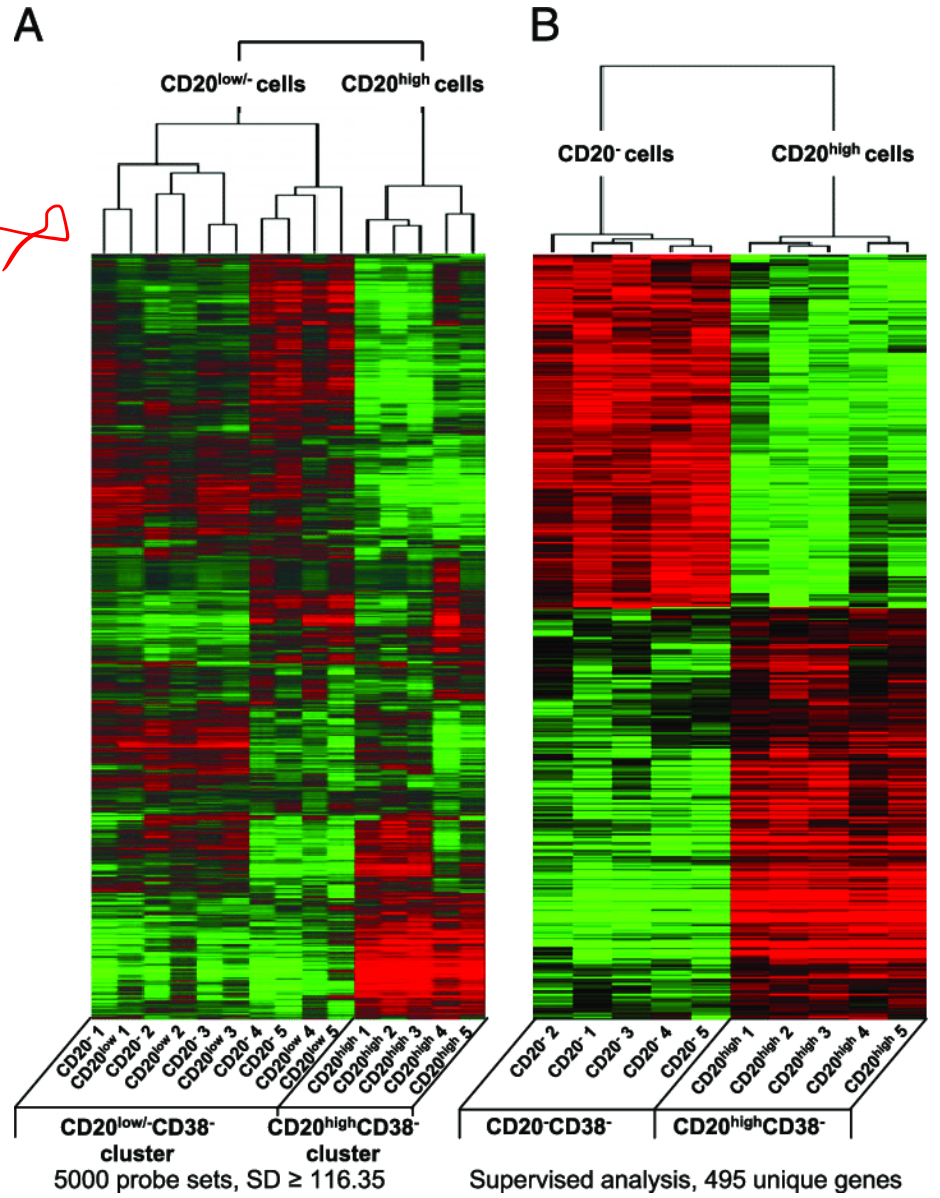
Medical Diagnosis





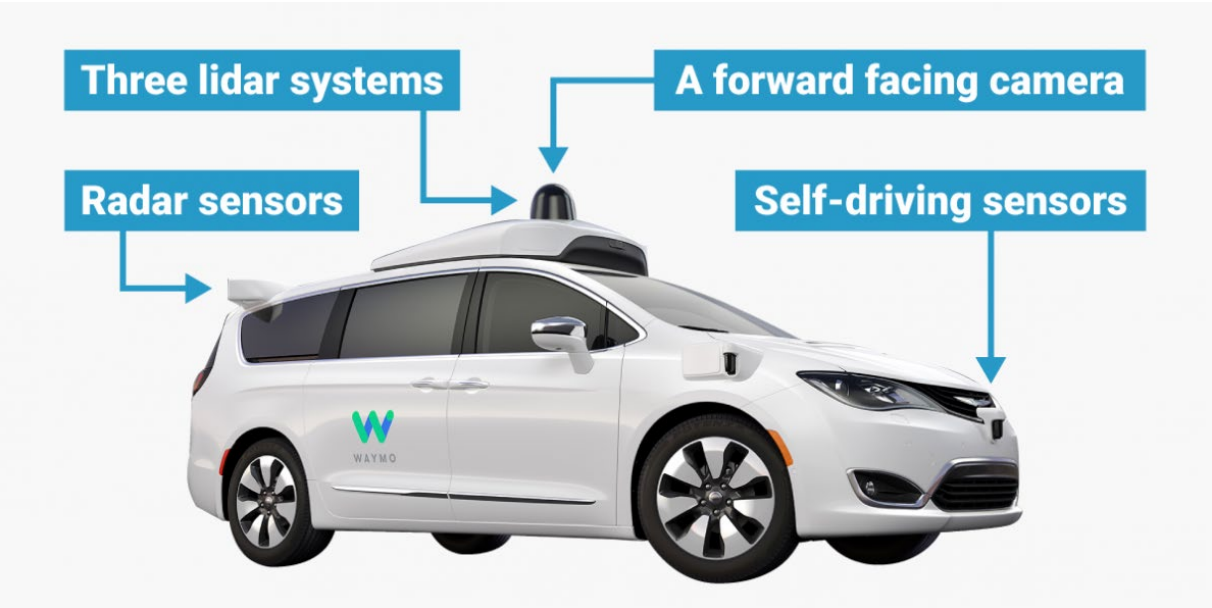
Bioscience Research

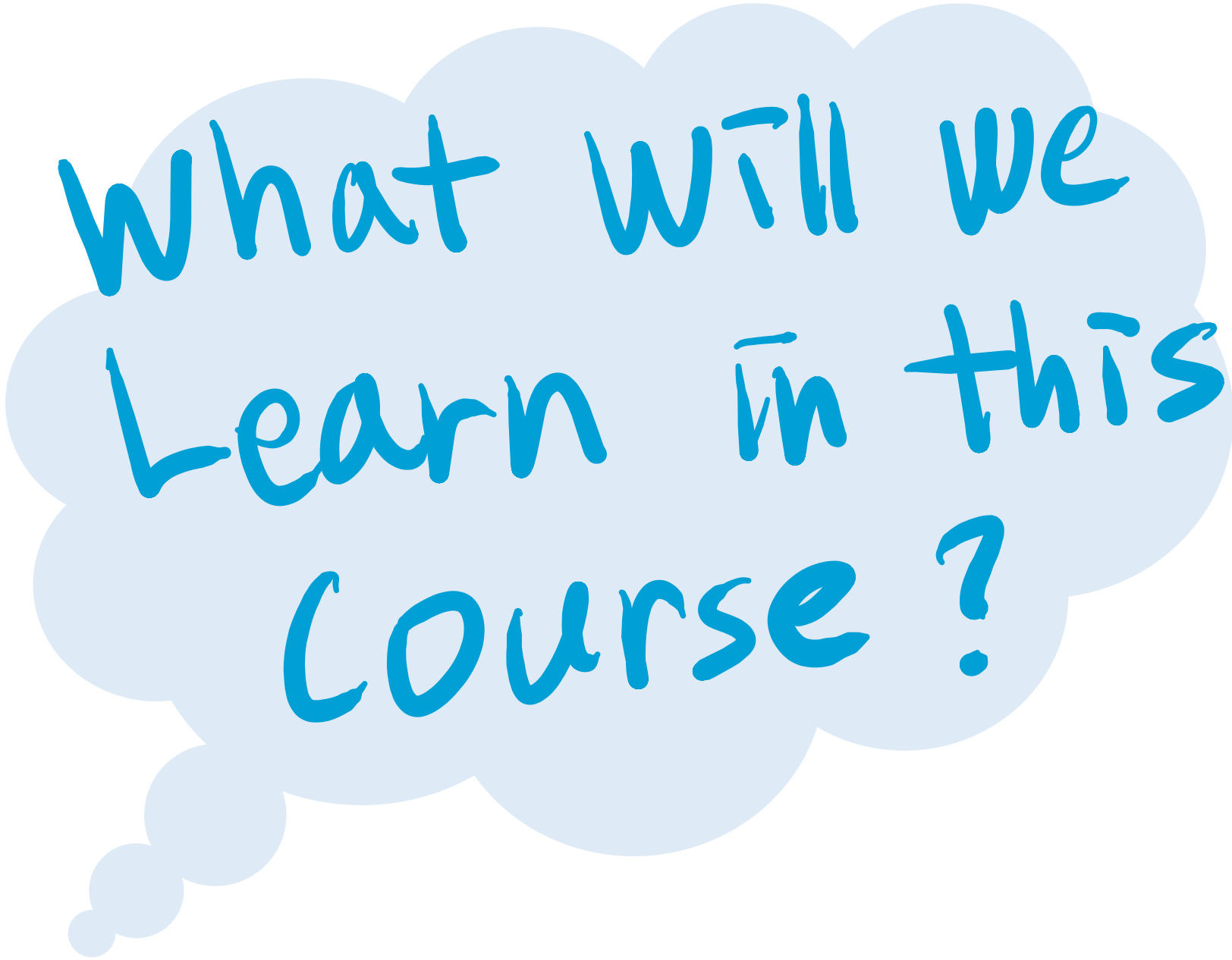




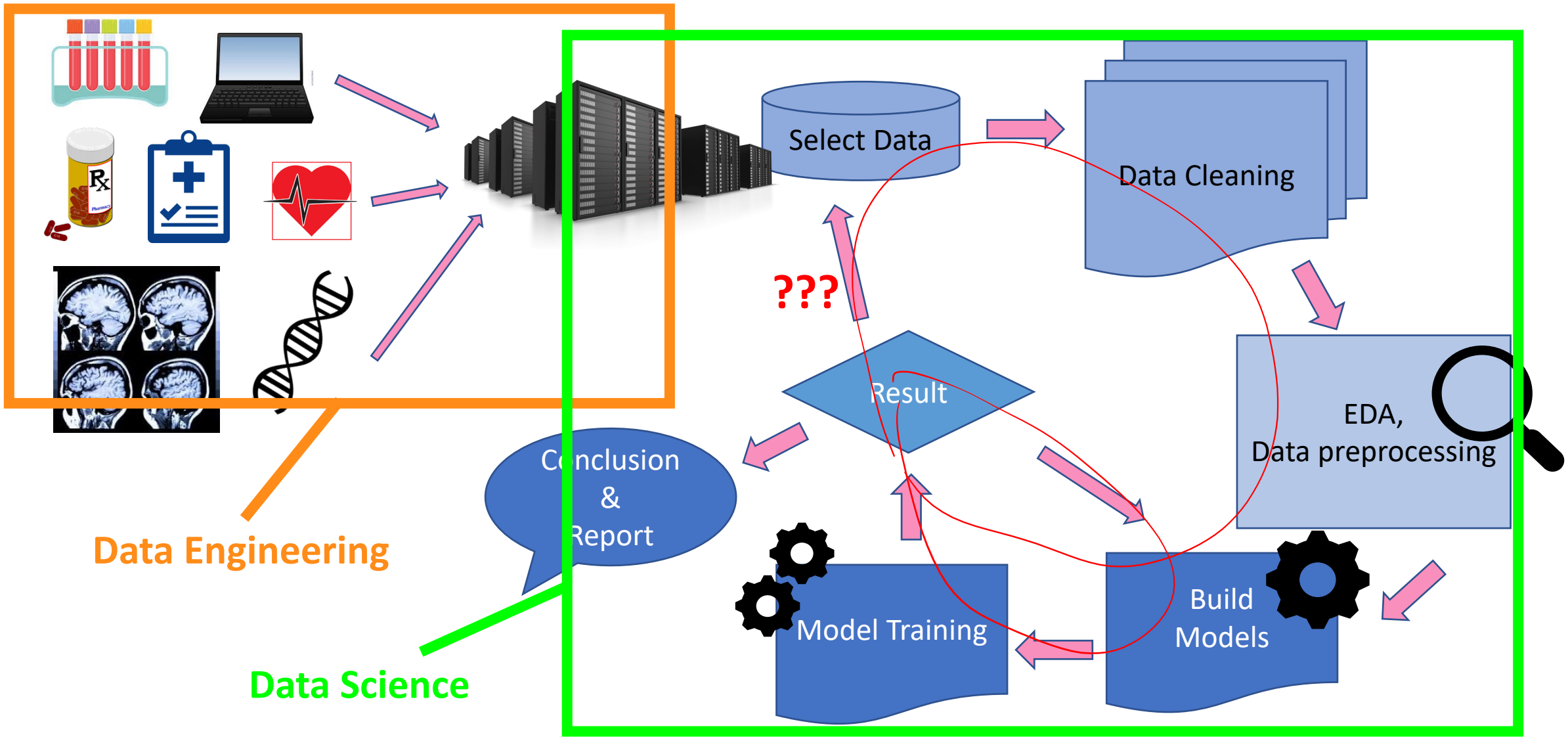
Internet of Things

Self-driving Car

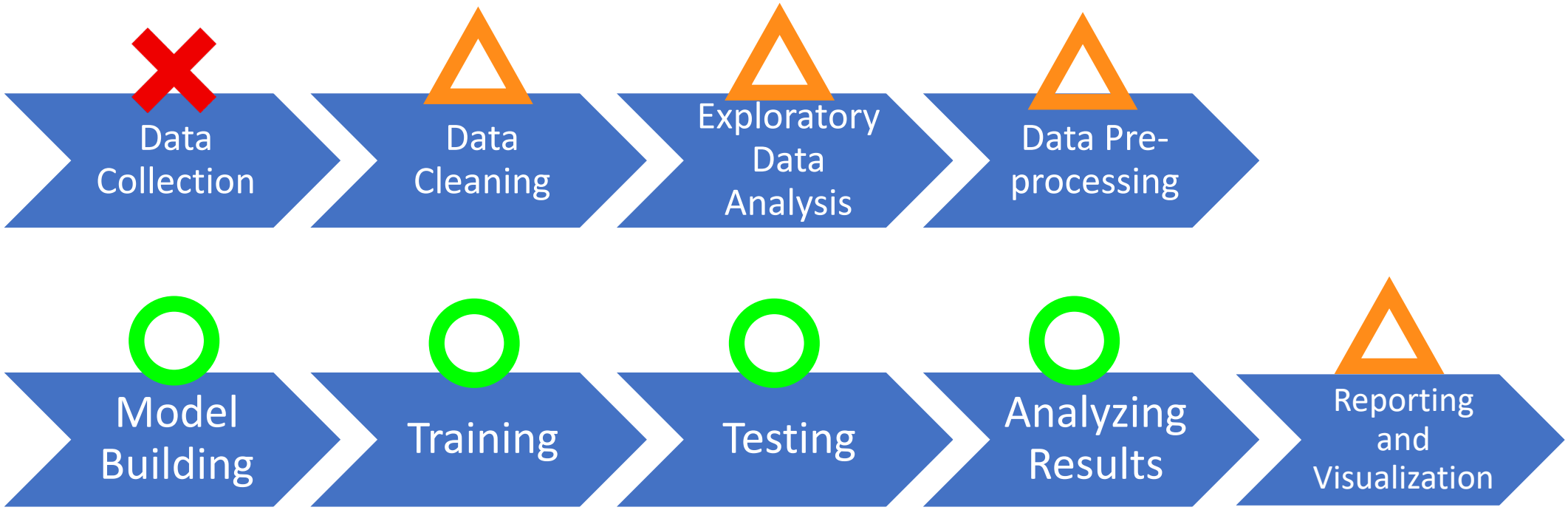



What will we
Learn in this
Course?

Data Science Project Life Cycle

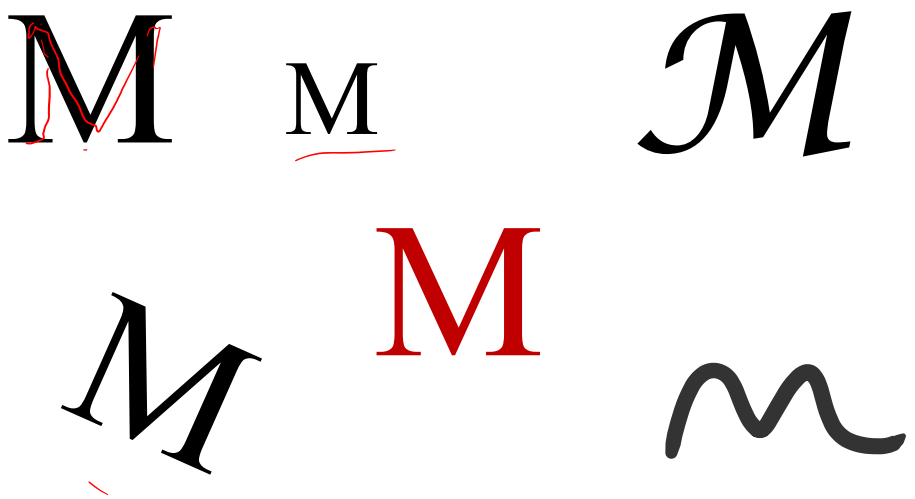


What is this course about?

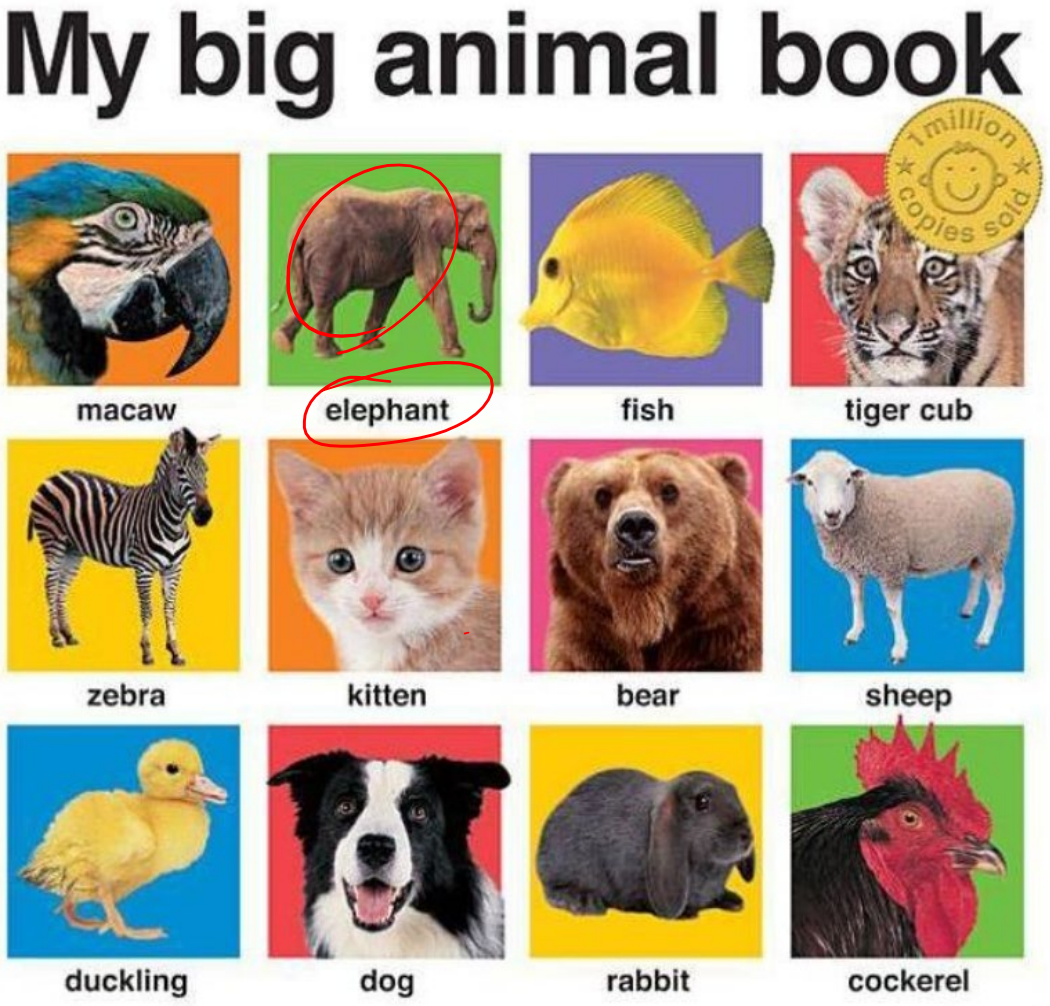


What is Learning?

Learn to generalize



Supervised Learning



?

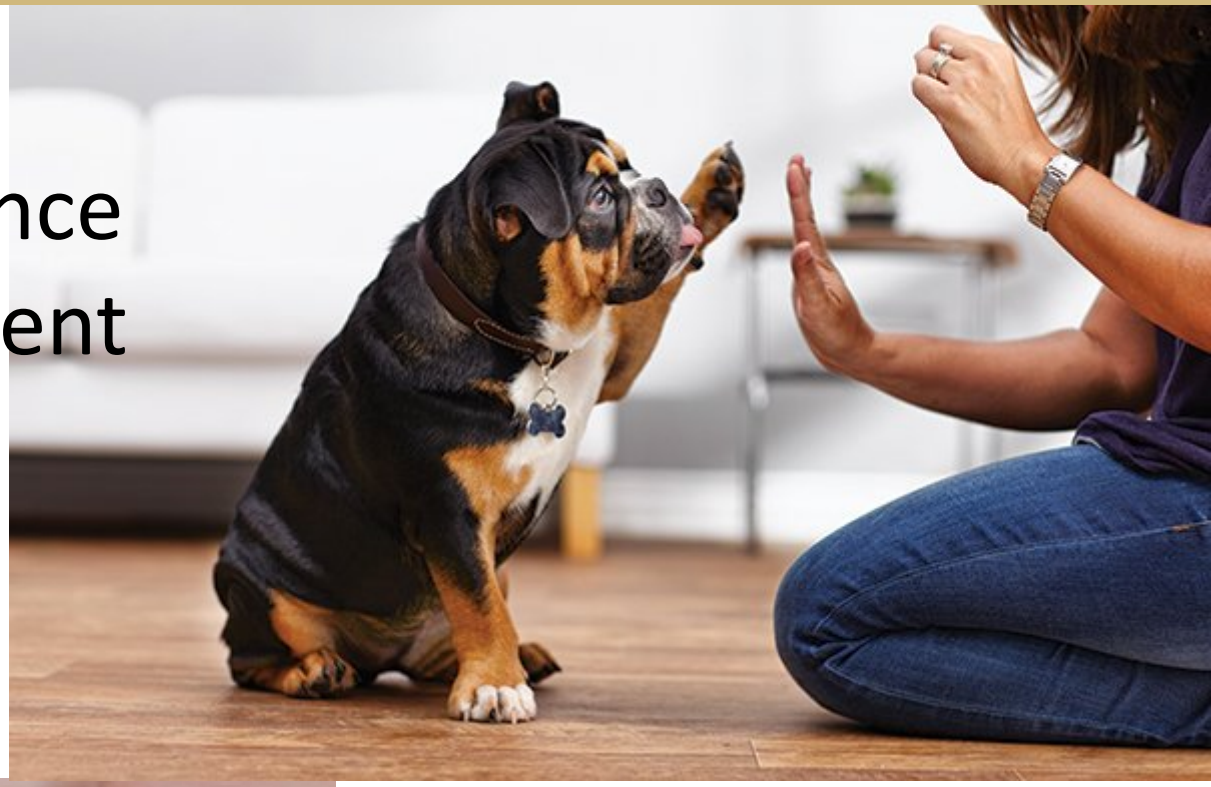
- Learn the data with labels
- Learn to predict

Unsupervised Learning

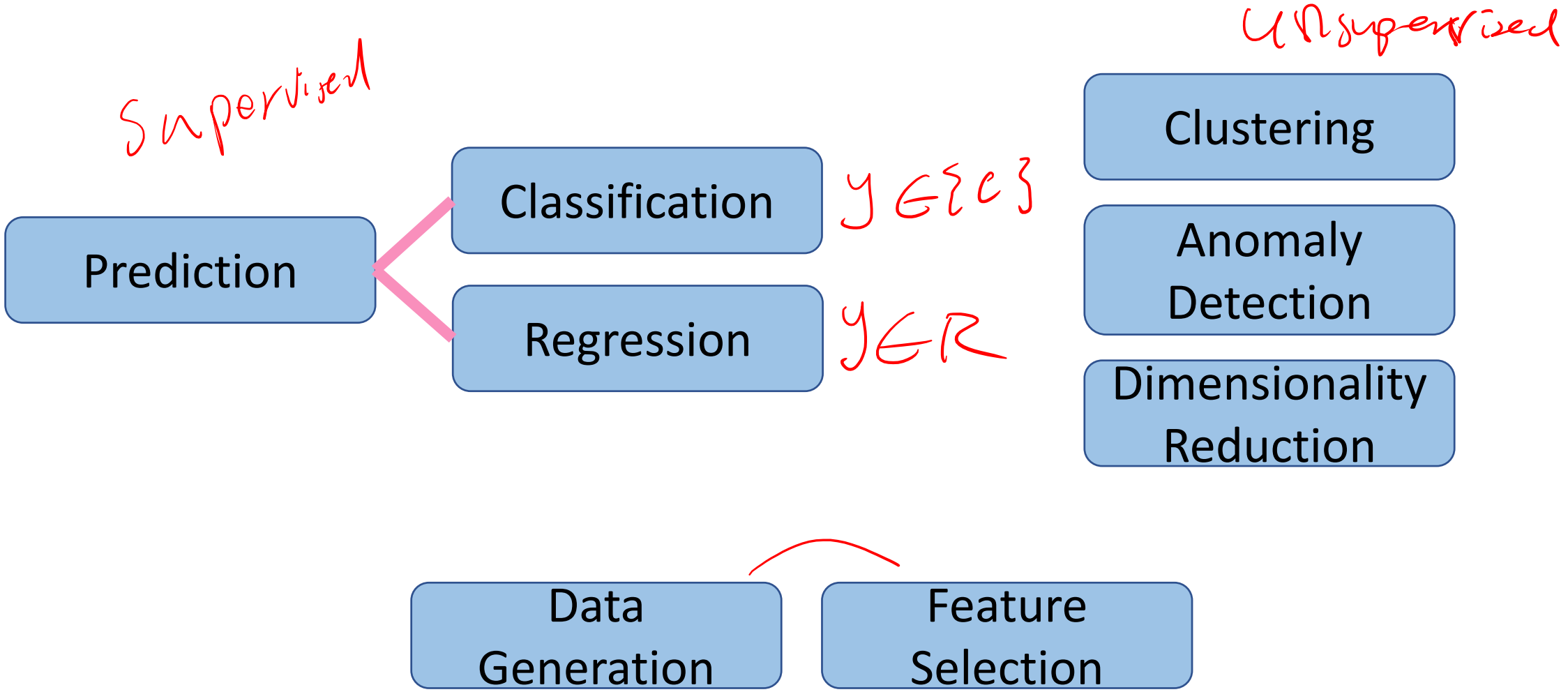
- Learn the data without labels
- Learn the underlying features/information

Reinforcement Learning

- Learn how to act from experience
- Experience = Reward/Punishment



Machine Learning Tasks



Machine Learning Models

Supervised

Parameteric

Linear Regression (R)
Logistic Regression (C)
Neural Networks (R, C)
Naïve Bayes Classifier (C)

Non-parameteric

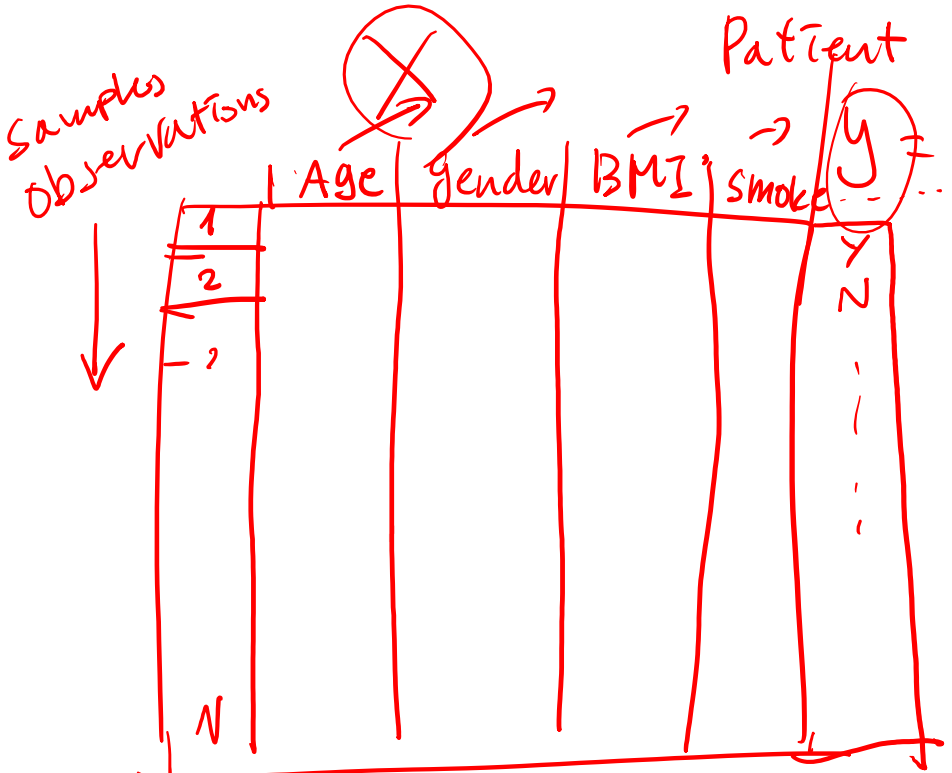
Support Vector Machine (R,C)
Decision Trees (R,C)
Nearest Neighbor (R, C)

Unsupervised

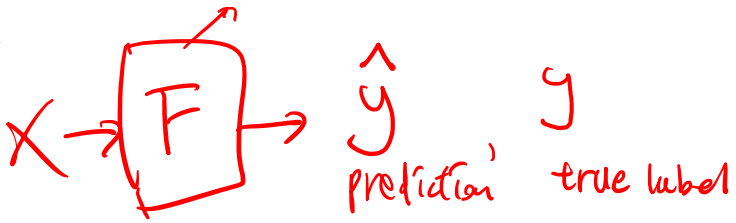
PCA (dimensionality reduction)
K-means clustering
Hierarchical clustering
Autoencoders (learning features)

Supervised Learning- variables

- Data: $\boldsymbol{X}$
- Target y
- Prediction $\hat{y}$
- Model $\mathcal{F}$
- Parameters $\boldsymbol{\theta}$
- Hyperparameters $\boldsymbol{\lambda}$
- Loss $\mathcal{L}$



Training (X, y)



$$\mathcal{L} = f(\hat{y}, y)$$

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2$$

Supervised Learning- model types

Data: $\boldsymbol{X}$
Target y
Prediction $\hat{y}$
Model $\mathcal{F}$
Parameters $\boldsymbol{\theta}$
Hyperparameters $\boldsymbol{\lambda}$
Loss $\mathcal{L}$

$$\hat{y} = \mathcal{F}(\boldsymbol{X})$$ → Newest Neighbor

$$\hat{y} = \mathcal{F}(\boldsymbol{X}, \theta_1, \theta_2, \cdots, \theta_n)$$ → LR Logr

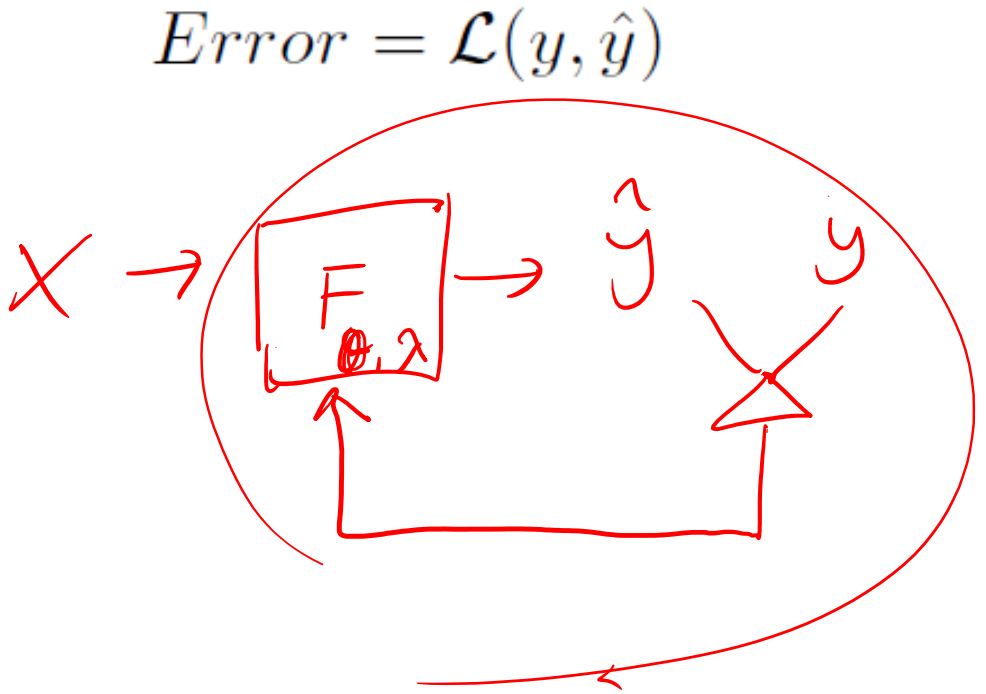
→ NN

$$\hat{y} = \mathcal{F}(\boldsymbol{X}, \theta_1, \theta_2, \cdots, \theta_n, \lambda_1, \lambda_2, \cdots, \lambda_m)$$

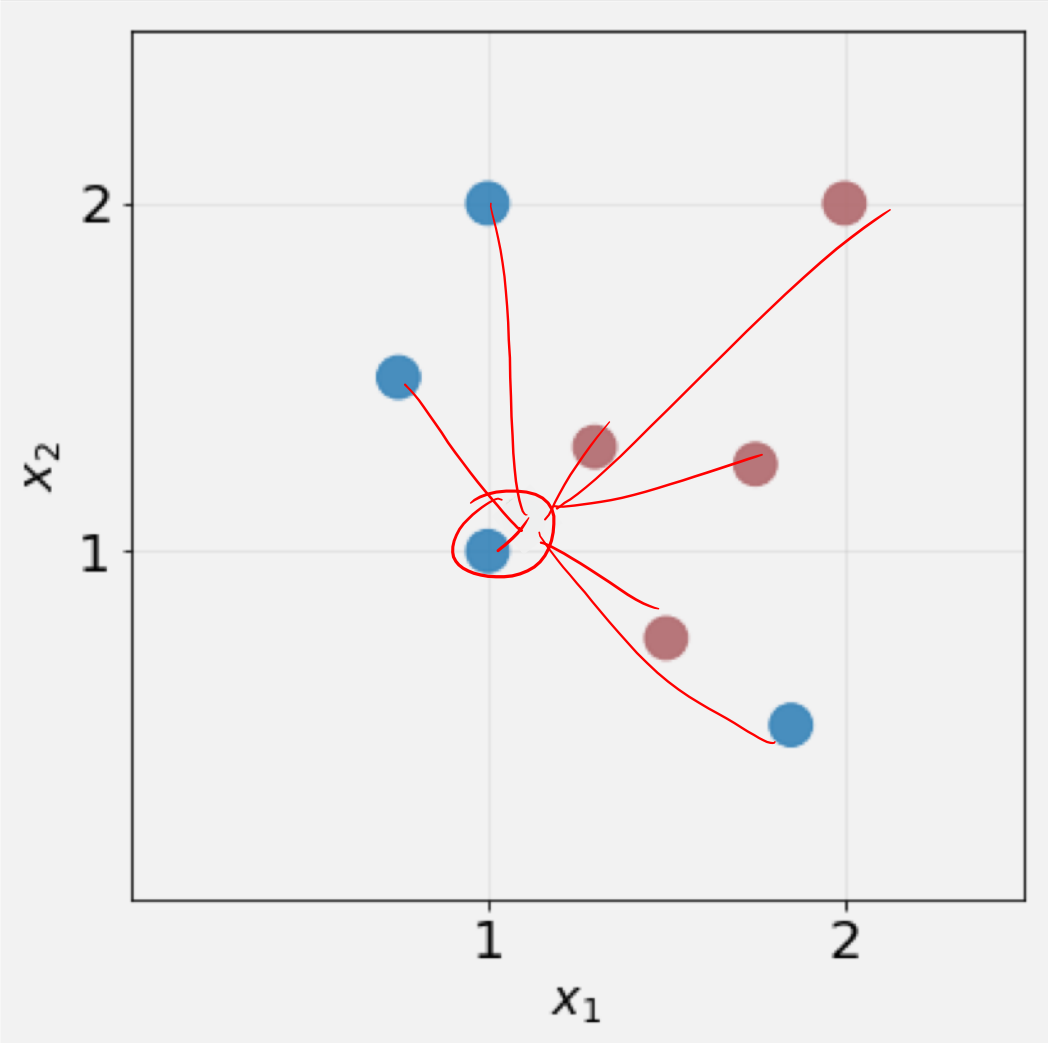
$$\hat{y} = \mathcal{F}(\boldsymbol{X}, \lambda_1, \lambda_2, \cdots, \lambda_m)$$ DT SVM

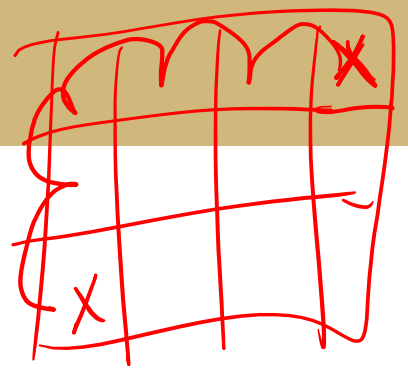
Supervised Learning- training

Data: $\boldsymbol{X}$
Target y
Prediction $\hat{y}$
Model $\mathcal{F}$
Parameters $\boldsymbol{\theta}$
Hyperparameters $\boldsymbol{\lambda}$
Loss $\mathcal{L}$



Nearest Neighbor





Manhattan distance

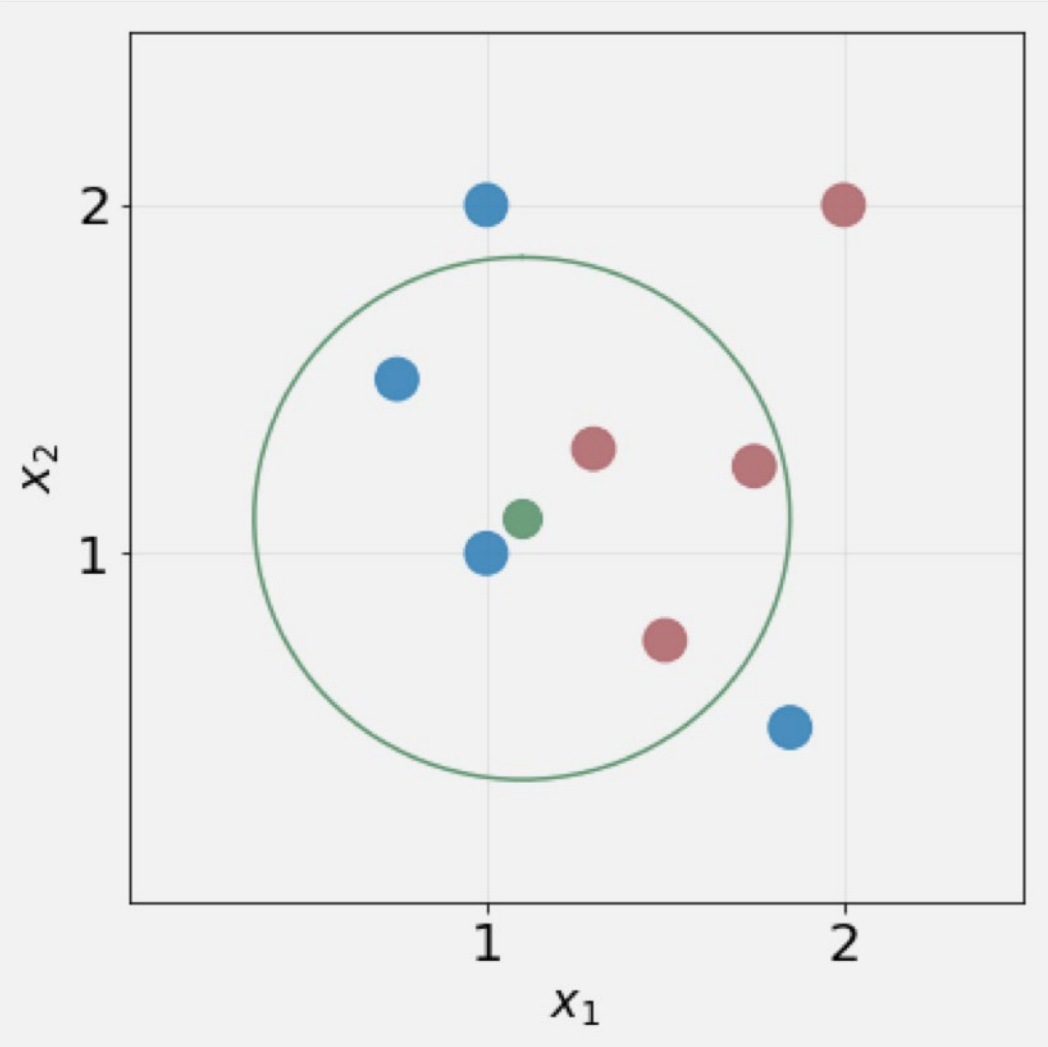
$$||\vec{X}_1 - \vec{X}_2||_1$$

Euclidean distance

$$||\vec{X}_1 - \vec{X}_2||_2$$

$\sqrt{x^2 + y^2 + z^2}$

K-Nearest Neighbor

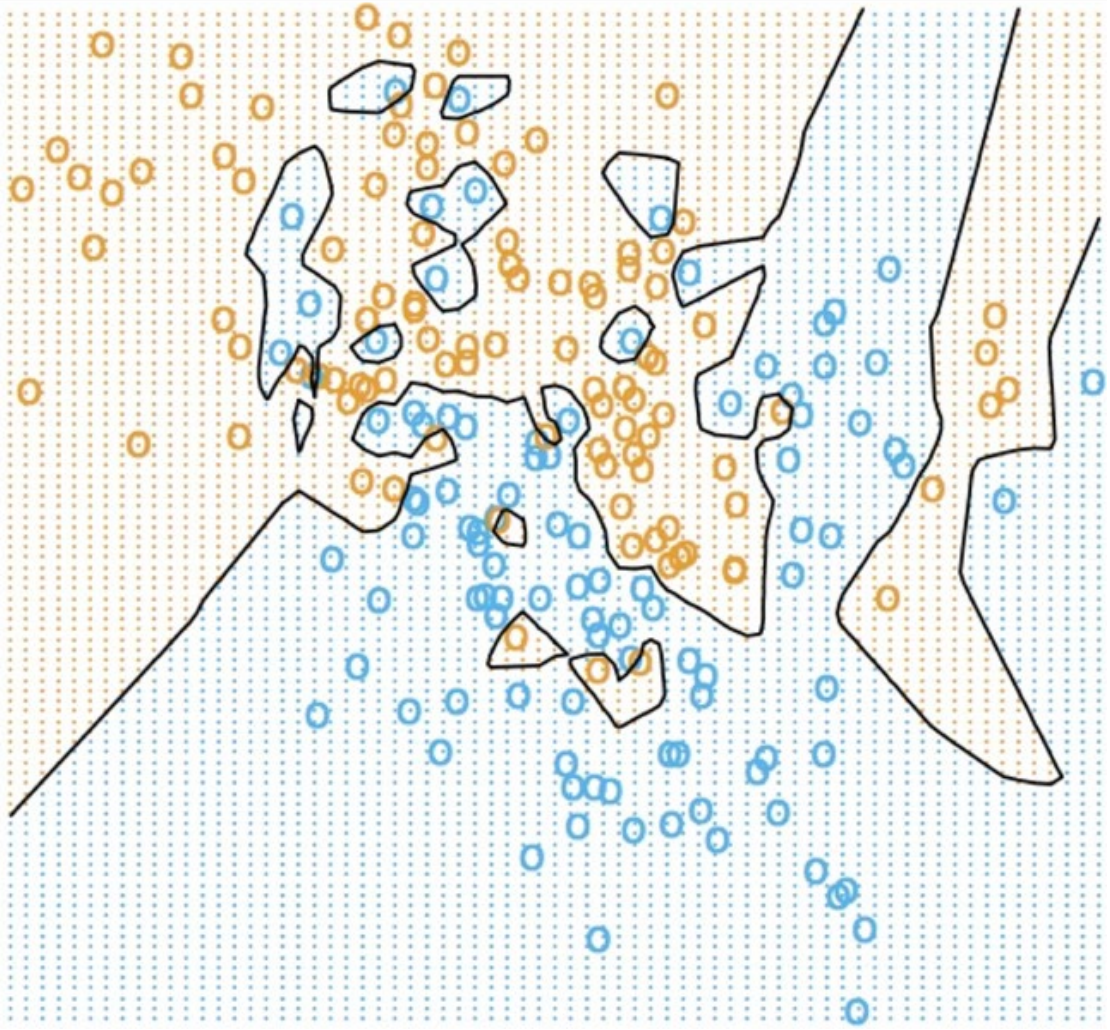


- Average /Majority vote
- k becomes a hyperparameter

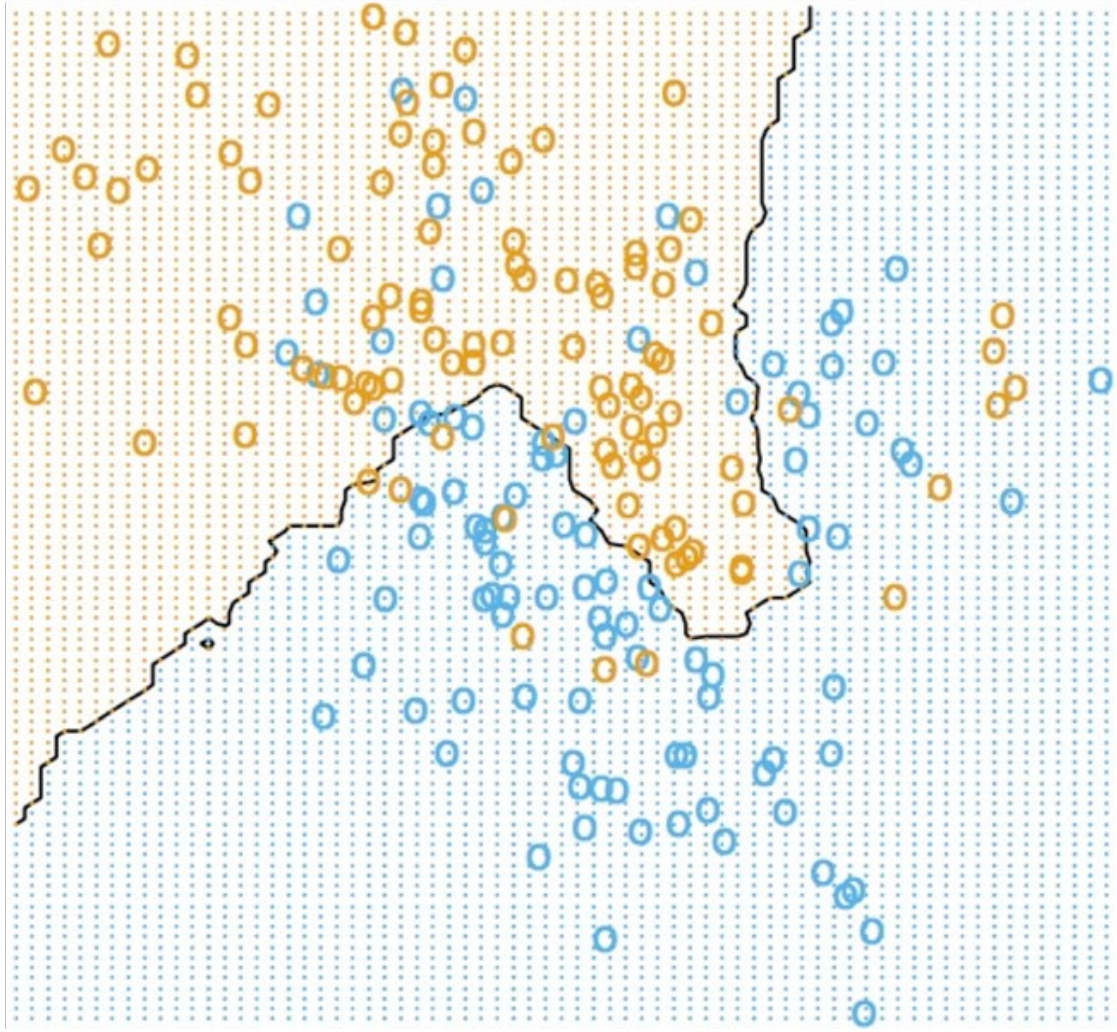
  - 1-NN predicts blue
  - 2-NN predicts tie
  - 5-NN predicts red

KNN { C, R

How to choose k?

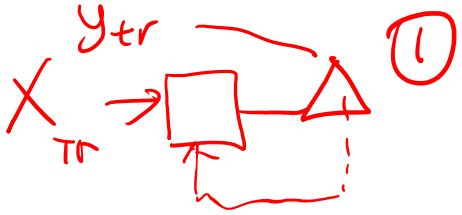


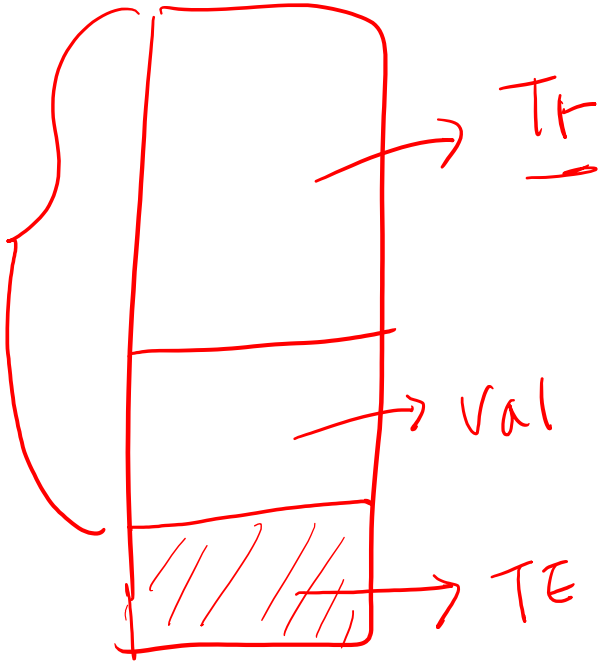
k=1

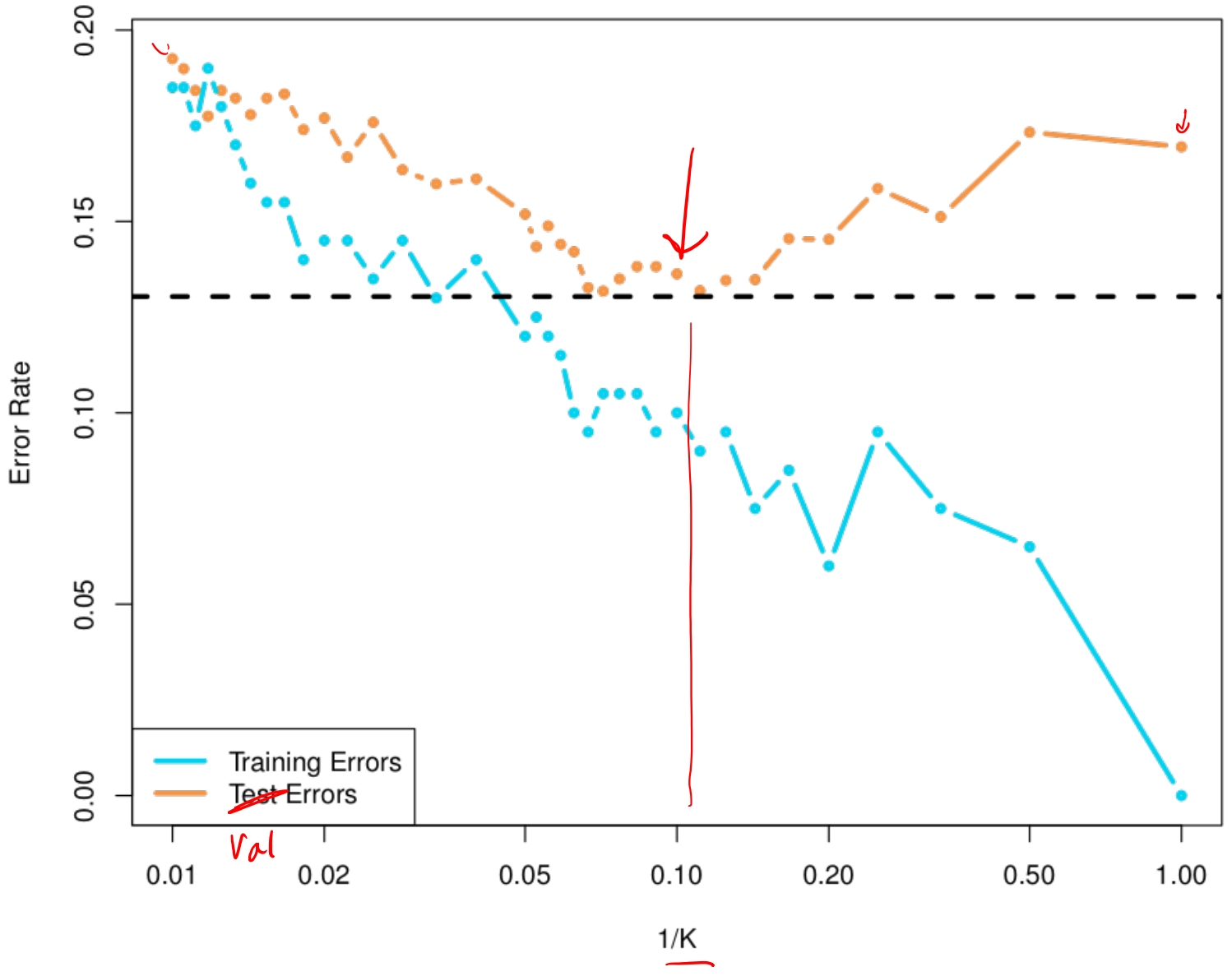


k=15

How to choose k?







KNN Properties

- Very Simple Algorithm
- Non-parameteric
- k is a Hyperparameter
- Very slow $O(N \times d)$
- Curse of dimensionality