STA303: Artificial Intelligence

Machine Learning Basics

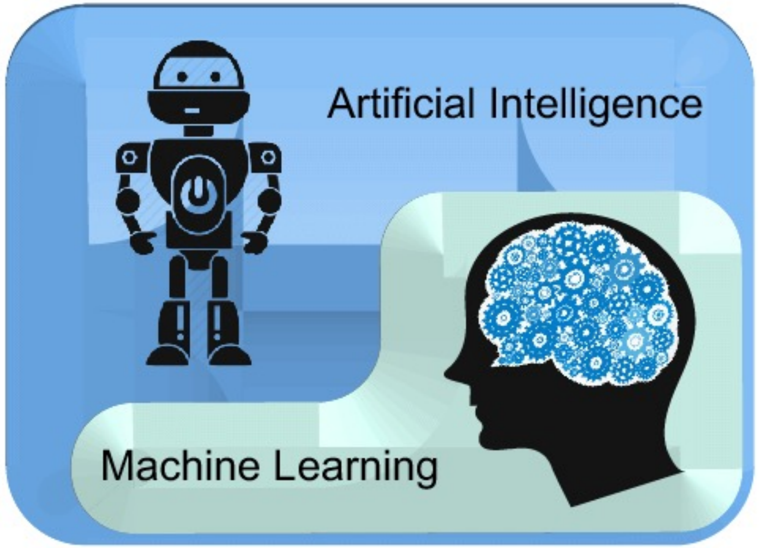
Fang Kong

https://fangkongx.github.io/

Part of slide credits: ai.berkeley.edu; Stanford CS229; https://www.geeksforgeeks.org/

Recap: What is Machine Learning?

- Subfield of Artificial Intelligence (AI)
  - Machine learning is an application or **subset** of AI that allows machines to learn from data without being programmed explicitly



https://www.javatpoint.com/difference-between-artificial-intelligence-and-machine-learning

Definition of Machine Learning

Tom Mitchell (1998): a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Experience (data): games played by the program (with itself)

Performance measure: winning rate

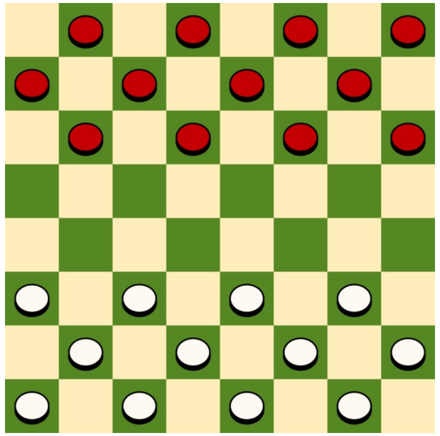
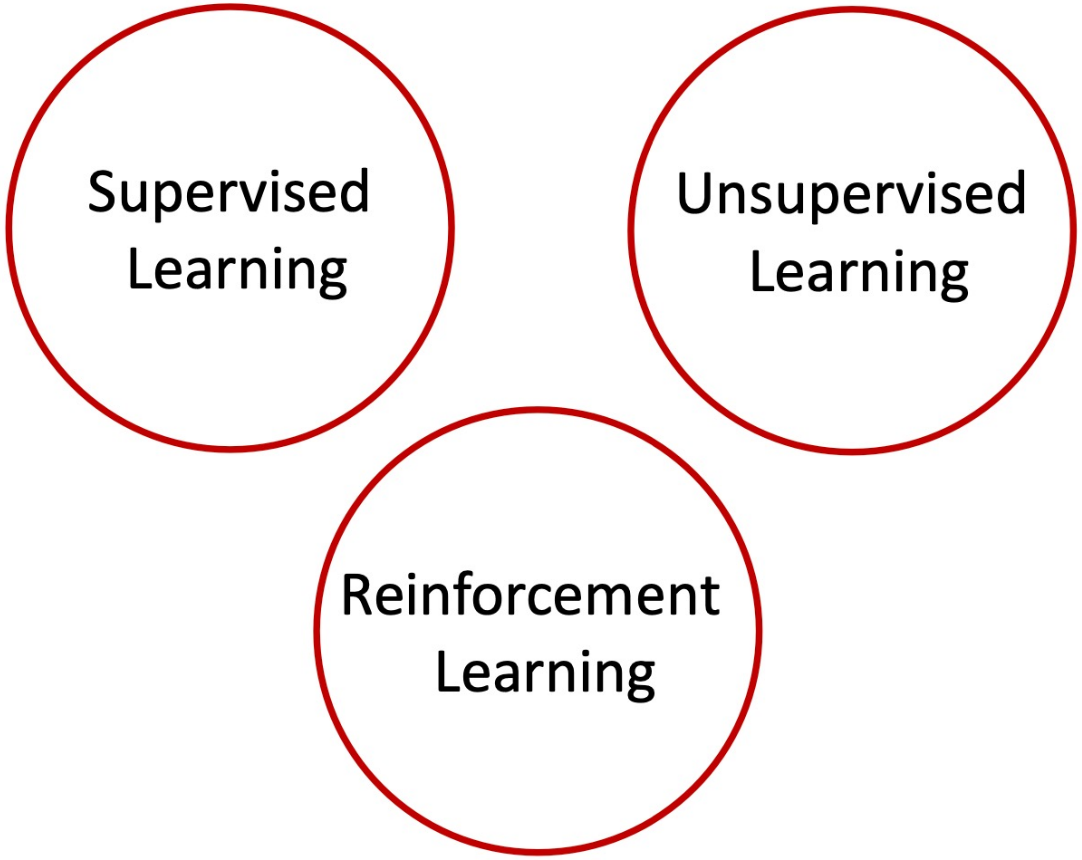


Image from Tom Mitchell's homepage

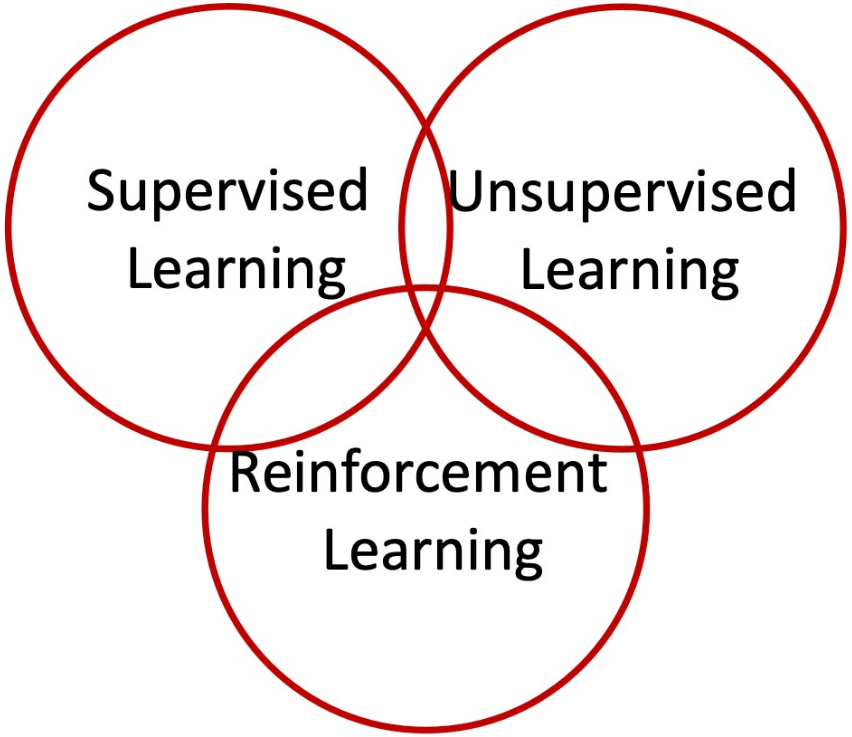
Taxonomy of ML

- A simplistic view based on tasks



Taxonomy of ML

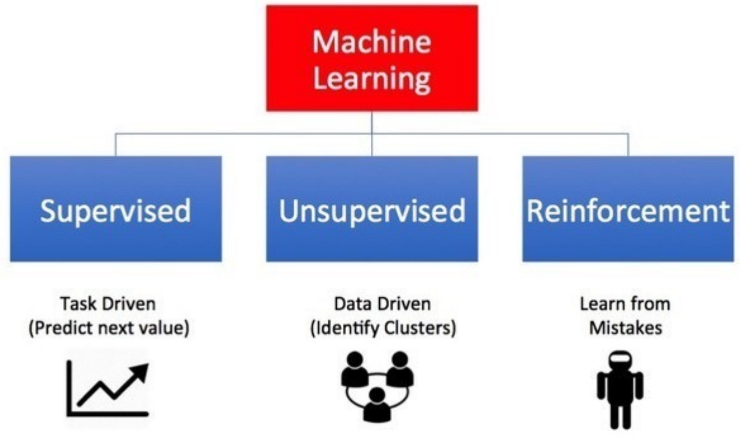
- A simplistic view based on tasks



can also be viewed as tools/methods

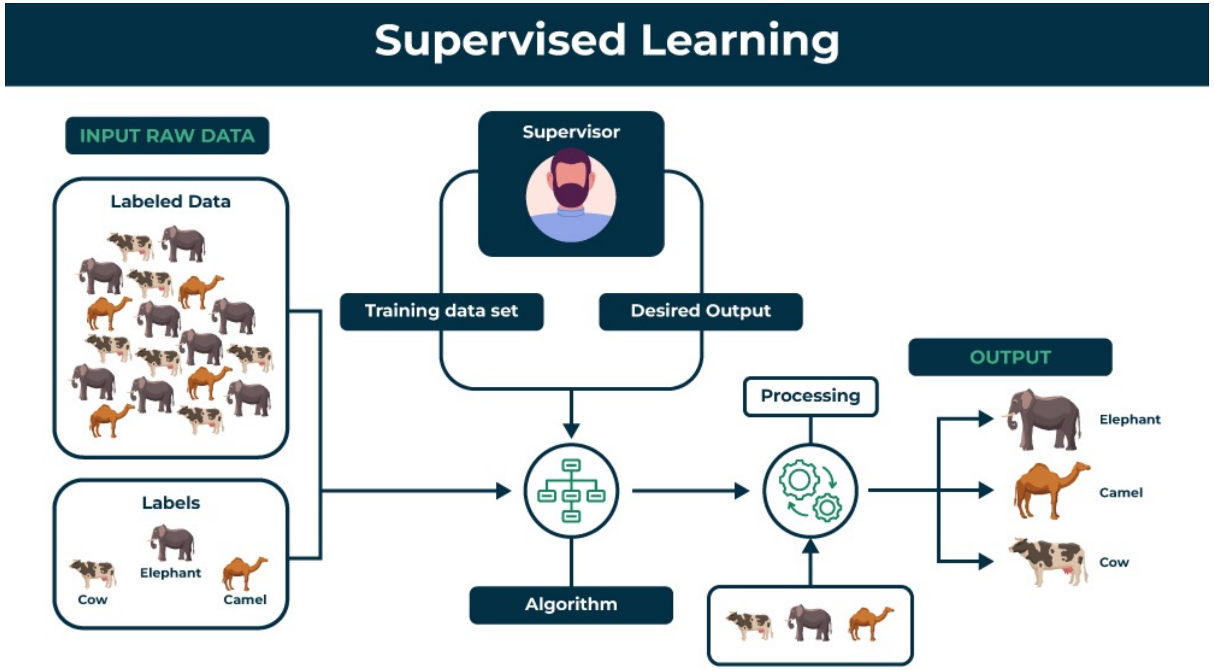
Types of Machine Learning

- Supervised learning
  - Use labeled data to predict on unseen points
- Unsupervised learning
  - No labeled data
- Reinforcement learning
  - Sequentially collect data and learn from feedback

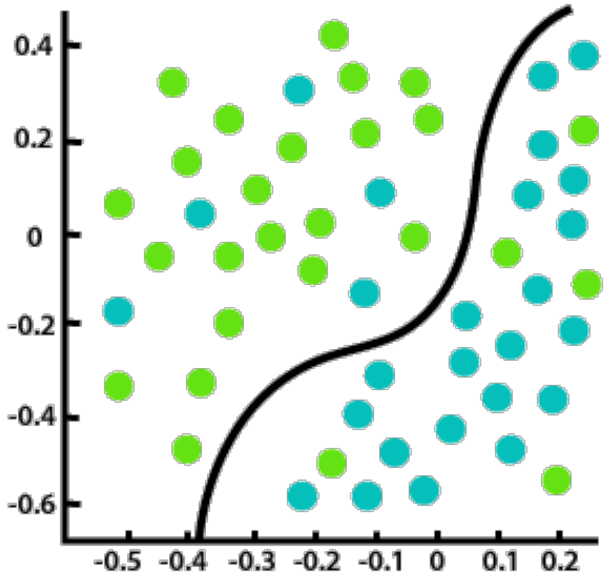


Supervised Learning

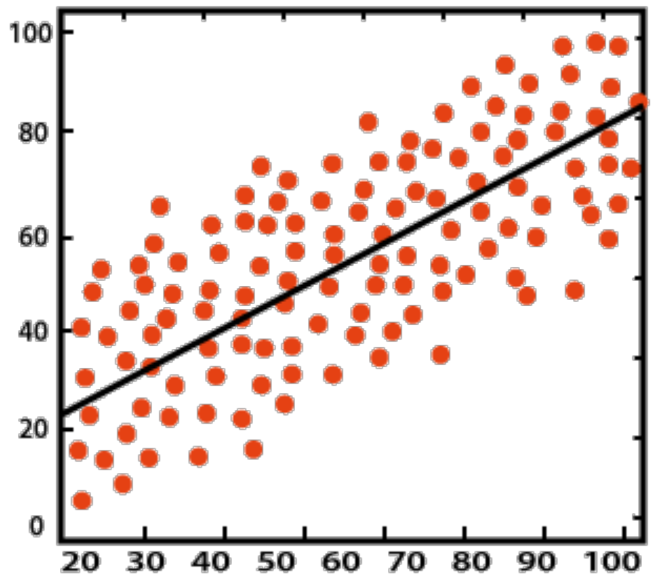
- Trained on a “Labelled Dataset”
- Labelled datasets have both input and output parameters



Tasks in Supervised Learning



Classification



Regression

Classification example: Spam Filter

- Input: an email
- Output: spam or not

- Setup:
  - Get a large collection of example emails, each labeled "spam" or "not"
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails

- Features: The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts, WidelyBroadcast
  - …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

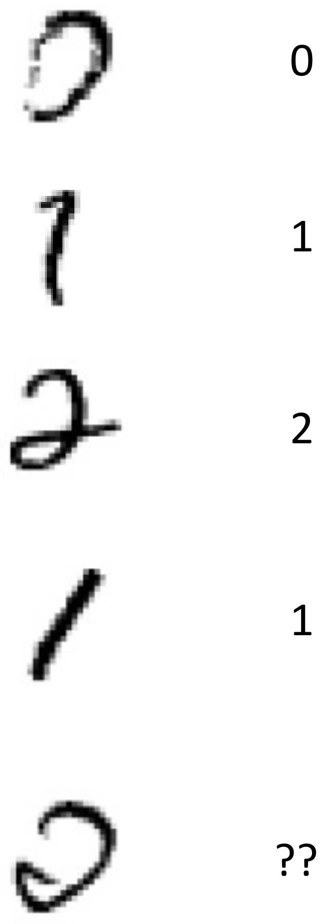
Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Classification example: Digit Recognition

- Input: images / pixel grids
- Output: a digit 0-9

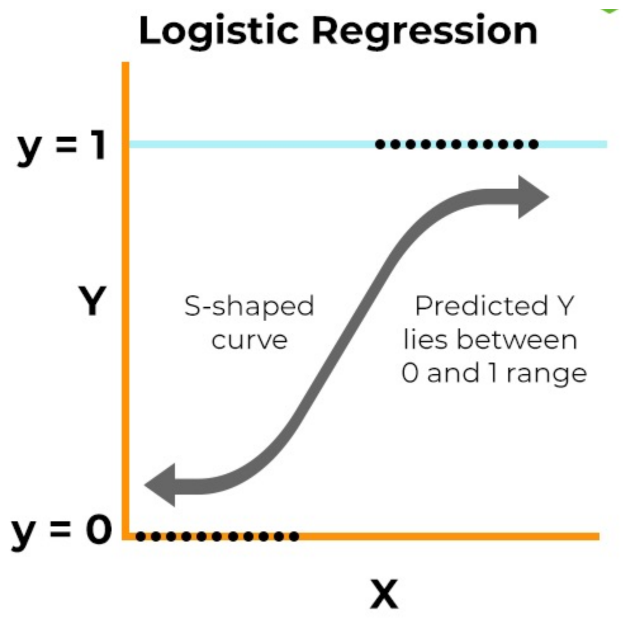
- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future digit images

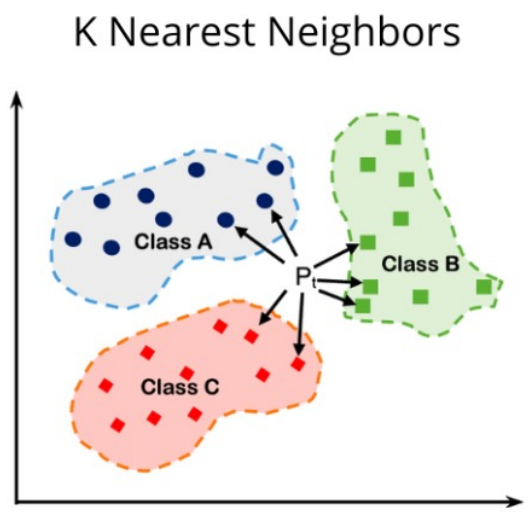
- Features: The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: NumComponents, AspectRatio, NumLoops
  - …

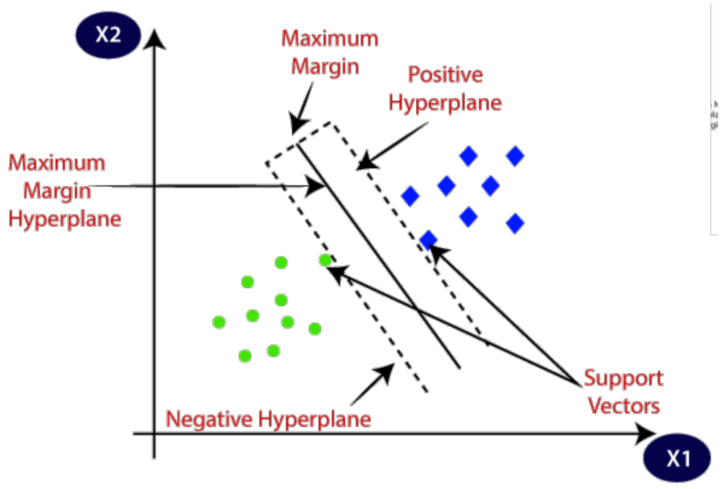


Common classification algorithms

- K-Nearest Neighbors (KNN)
- Logistic Regression
- Support Vector Machine
- Random Forest
- Decision Tree
- Naive Bayes

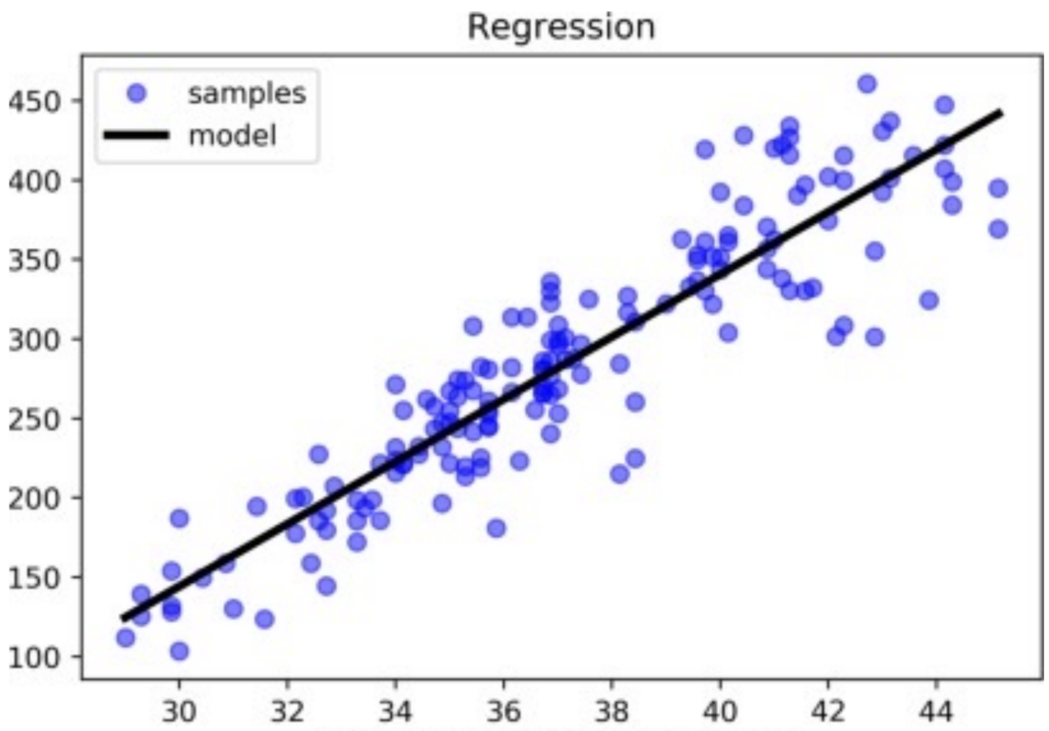






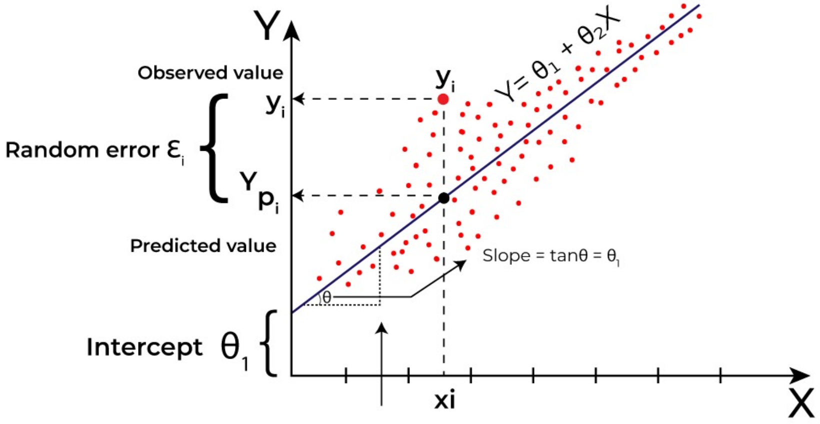
Regression example

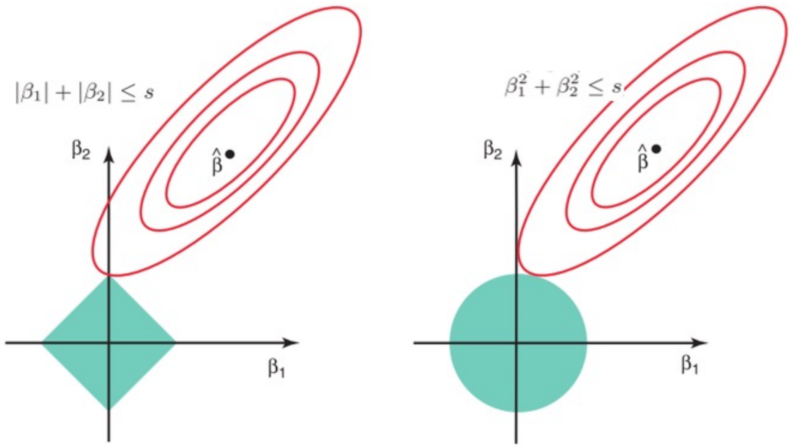
- Predicting the price of a house based on its size, location, and amenities
- Forecasting the sales of a product



Common regression algorithms

- Linear Regression
- Ridge Regression
- Lasso Regression
- Polynomial Regression
- Decision tree
- Random Forest



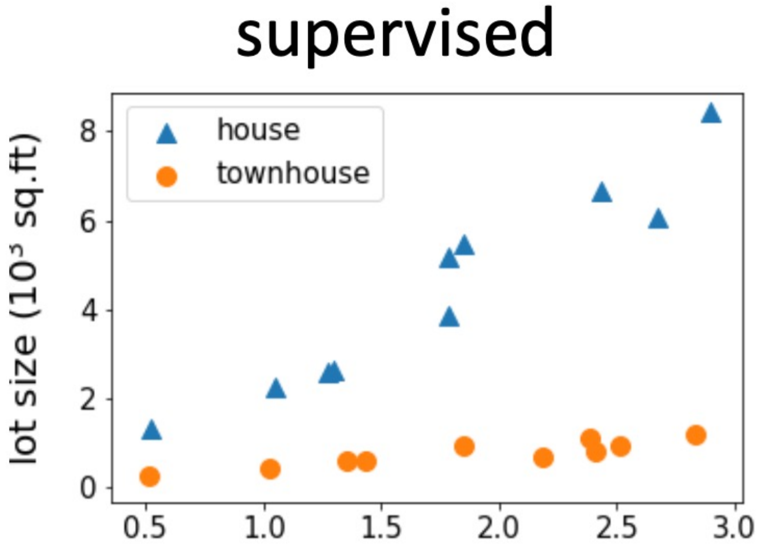


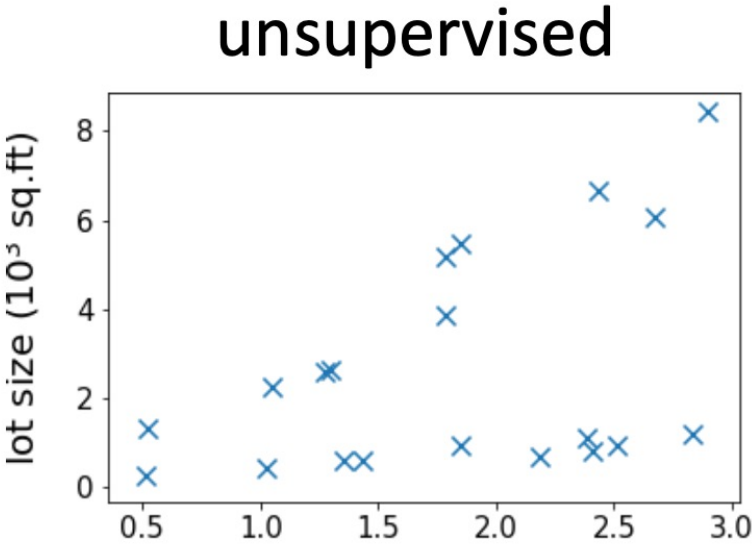
Lasso Regression

Ridge Regression

Unsupervised Learning

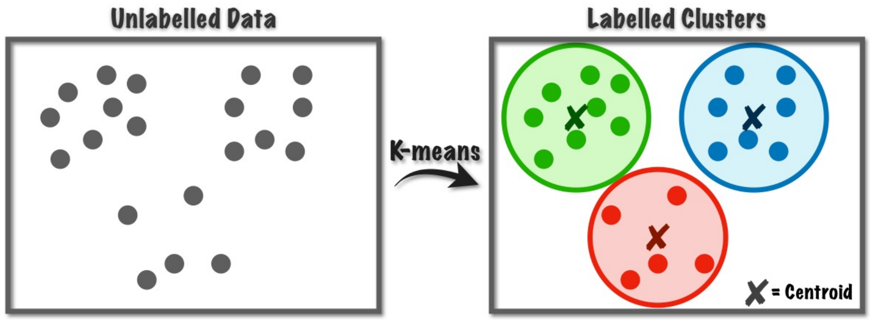
- Discover patterns and relationships using unlabeled data
- Without labeled target outputs

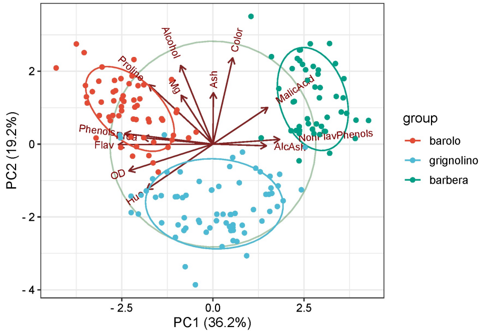




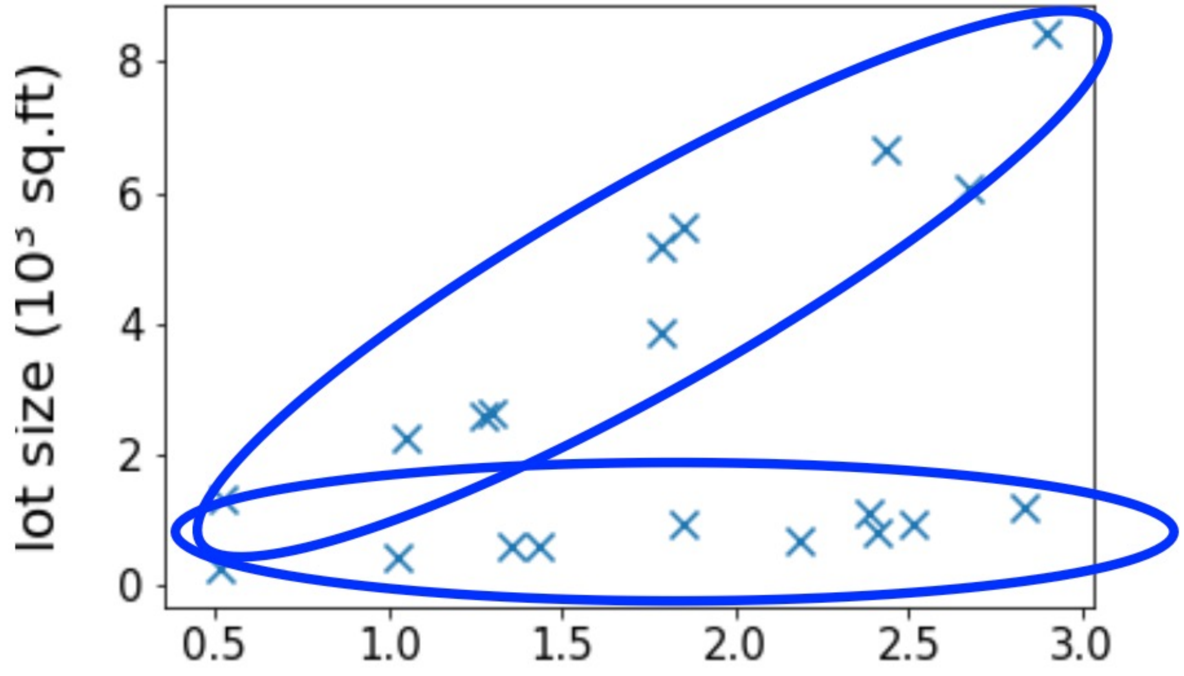
Tasks in Unsupervised Learning

Clustering	Dimensionality Reduction	Association (Data Mining)
- K-Means - Polynomial - Hierarchical - Fuzzy C-Means	- Principal Component Analysis - Kernel Principal Analysis	- Apriori Algorithm - Eclat Algorithm - FP-Growth Algorithm
Grouping data points into clusters based on their similarity	Reduce the dimensionality of data while preserving its essential information	Find the relationships between variables in the large database

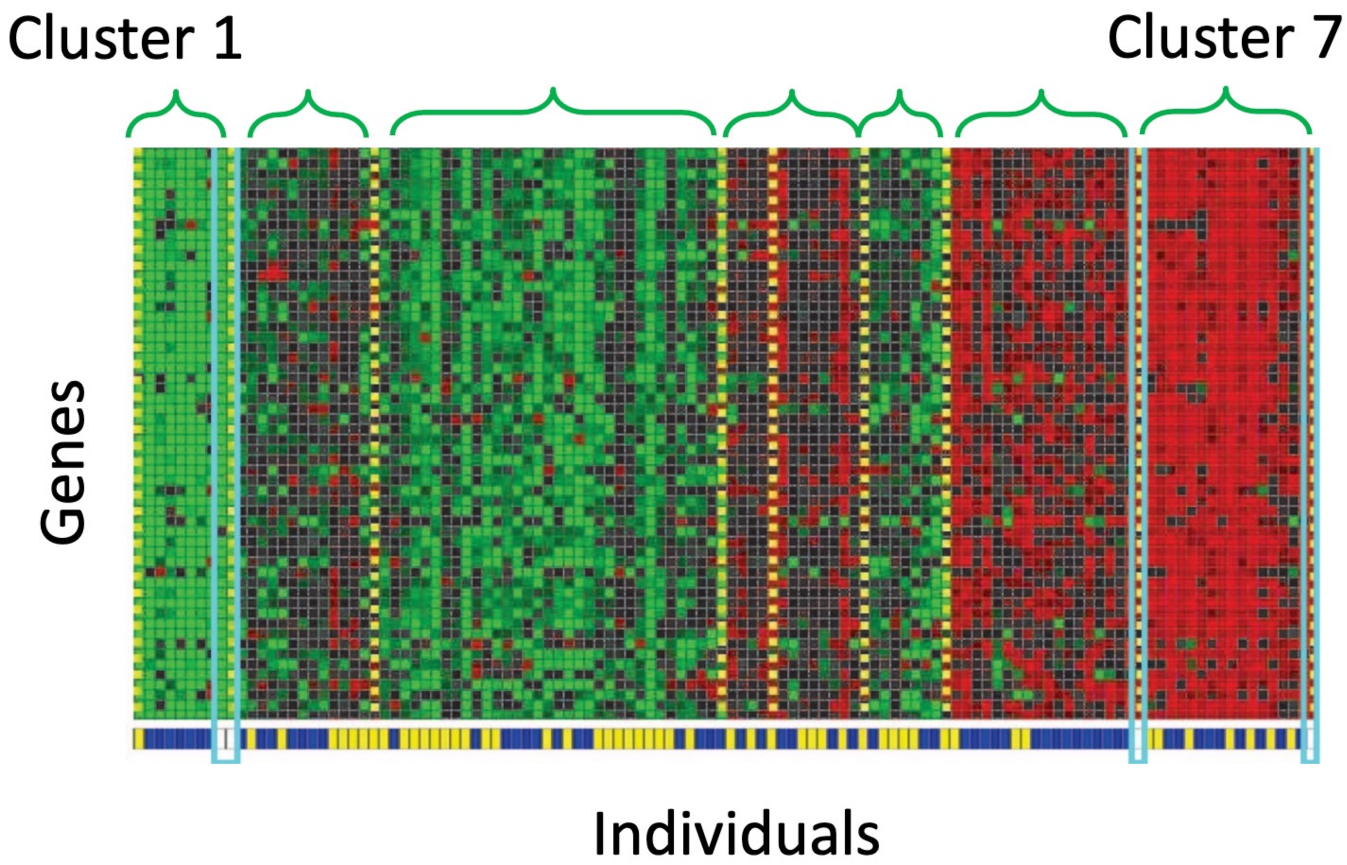




Clustering

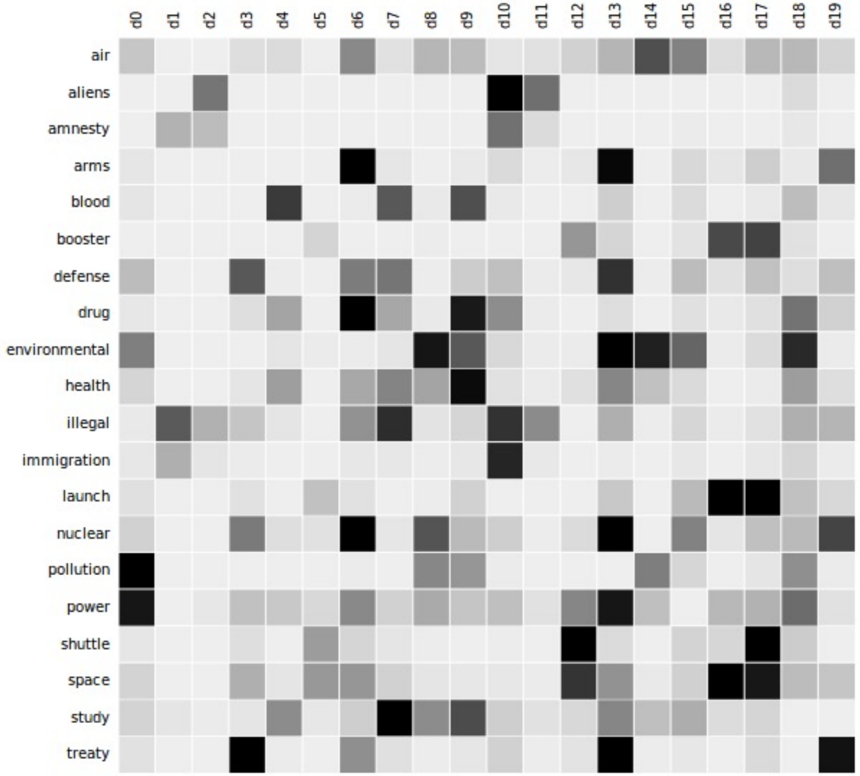


Clustering Genes



Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. [Su-In Lee, Dana Pe'er, Aimee M. Dudley, George M. Church and Daphne Koller. ’06]

Latent Semantic Analysis (LSA)

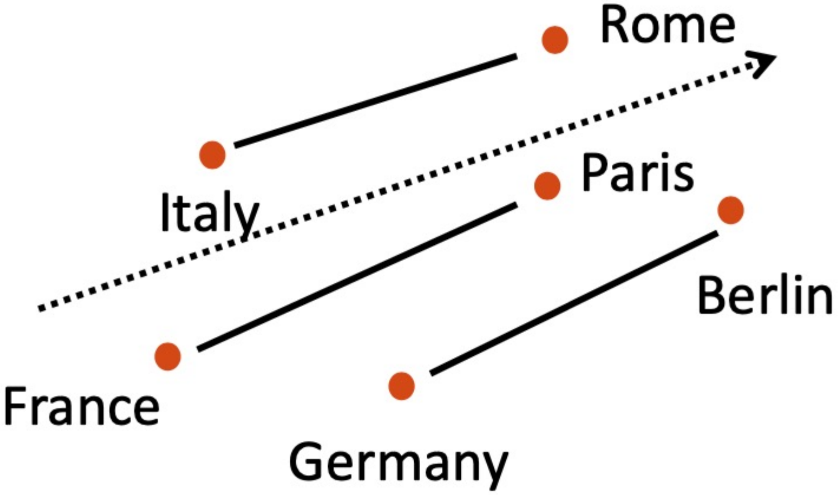


https://upload.wikimedia.org/wikipedia/commons/6/6e/Topic_detection_in_a_document-word_matrix.gif

Word Embeddings

Represent words by vectors

- word $\xrightarrow{\text{encode}}$ vector
- relation $\xrightarrow{\text{encode}}$ direction

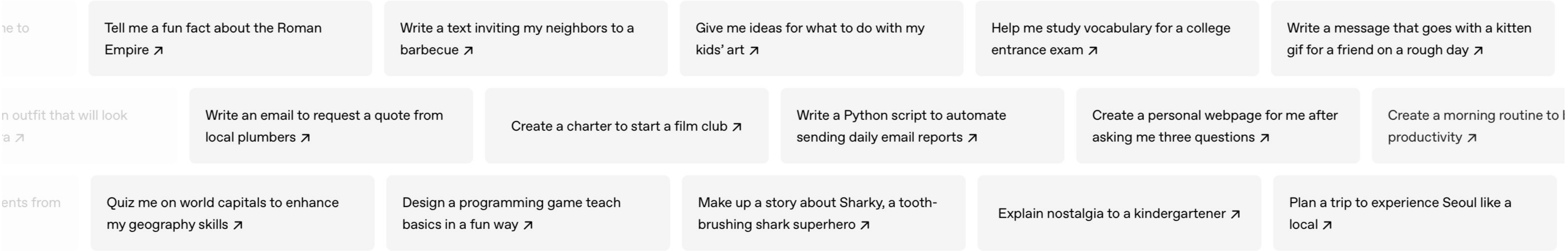


Unlabeled dataset

Word2vec [Mikolov et al'13]
GloVe [Pennington et al'14]

Large Language Models

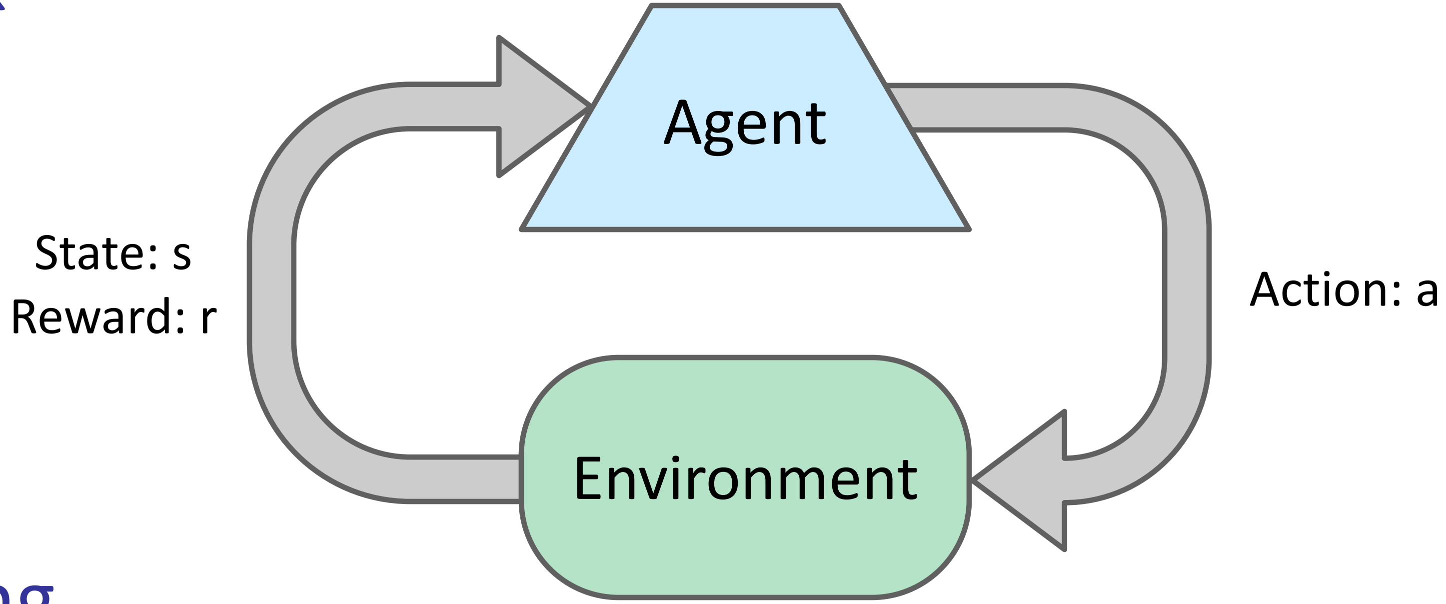
- Machine learning models for language learnt on large-scale language datasets
- Can be used for many purposes



https://openai.com/chatgpt/overview/

Reinforcement Learning

- Interact with the environment by producing actions and receiving feedback



- Q-learning
- Deep Q-learning
- PPO

Example: Learning to Walk

Initial

[Kohl and Stone, ICRA 2004]

[Video: AIBO WALK – initial]

Example: Learning to Walk

Finished

[Kohl and Stone, ICRA 2004]

[Video: AIBO WALK – finished]

Machine Learning Workflow

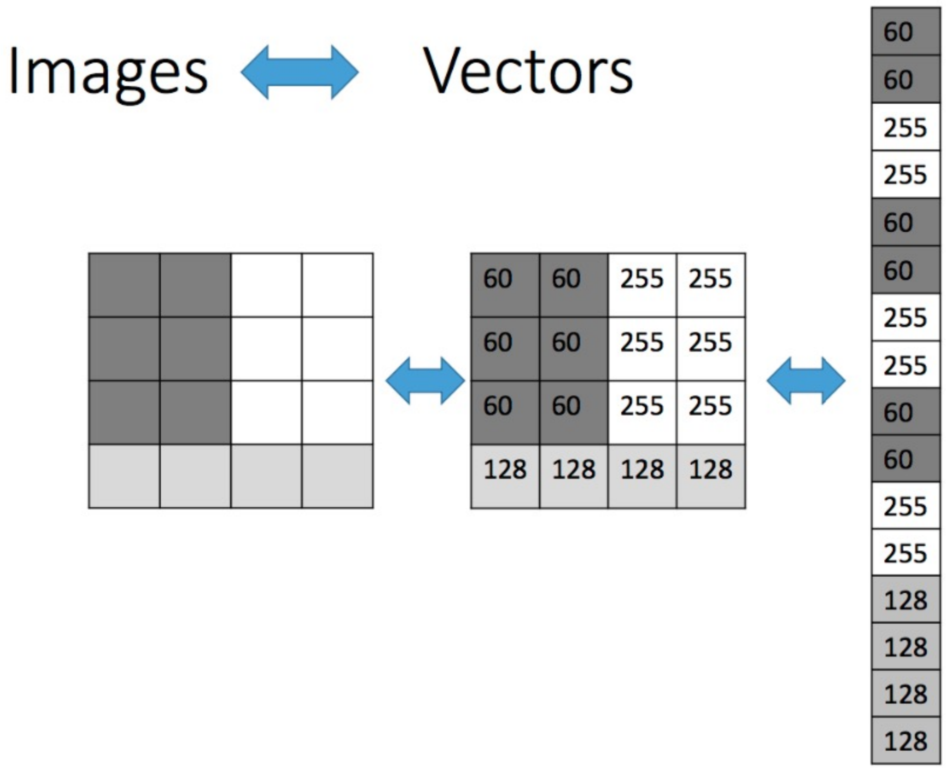
- 1. Gather and organize data
  - Preprocessing, cleaning, visualizing
- 2. Choose a model
- 3. Train and test your model, or iterate back to step 2 or 1
- 4. Deploy your model

Step 1: Gather and organize data

- Lots of types of data: images, text, audio waveforms, credit card transactions, etc.
- Common strategy: represent the input as an input vector in $\mathbb{R}^d$
  - Representation = mapping to another space that's easy to manipulate
  - Vectors are a great representation since we can do linear algebra

Step 1: Gather and organize data – Input vectors

- Such as raw pixels



- Better representations if you compute a vector of meaningful features.

Step 1: Gather and organize data – Input formulation

- Consider the classification problem in supervised learning

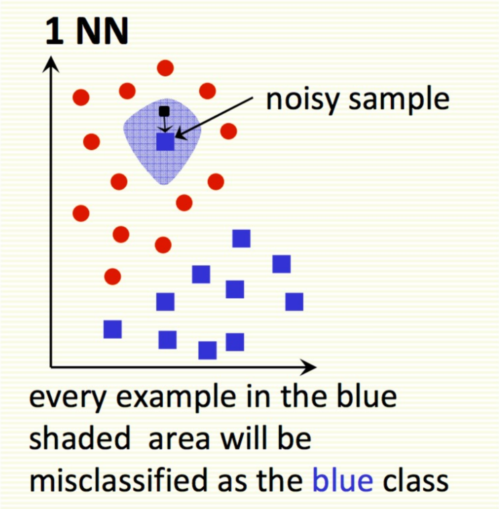
- Training set
  - A collection of $\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \dots, (\boldsymbol{x_N}, y_N)\}$
  - $\boldsymbol{x_i} \in \mathbb{R}^{\mathrm{d}}$ is the input feature of a data point
  - y_i is the corresponding label

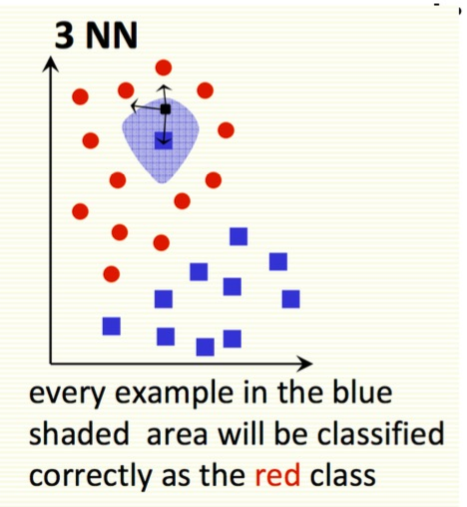
Step 2: Choose a model - Nearest Neighbors

- Suppose we are given a new input vector $\boldsymbol{x}$
- The idea of Nearest Neighbors (NN):
  - Select the nearest input vector of $\boldsymbol{x}$ in the training set
  - Use the label of the neighbor to predict the label of $\boldsymbol{x}$
  - How to formalize "nearest"?
    - Euclidean Distance, Manhattan Distance, Cosine Similarity

K-Nearest Neighbors (KNN)

- Nearest neighbors sensitive to noise or mis-labeled data
- Smooth by having k nearest neighbors vote





- Voting over k nearest neighbors: classification
- (Weighted) average over k nearest neighbors: regression

[Pic by Olga Veksler]

Step 3: Training and Testing

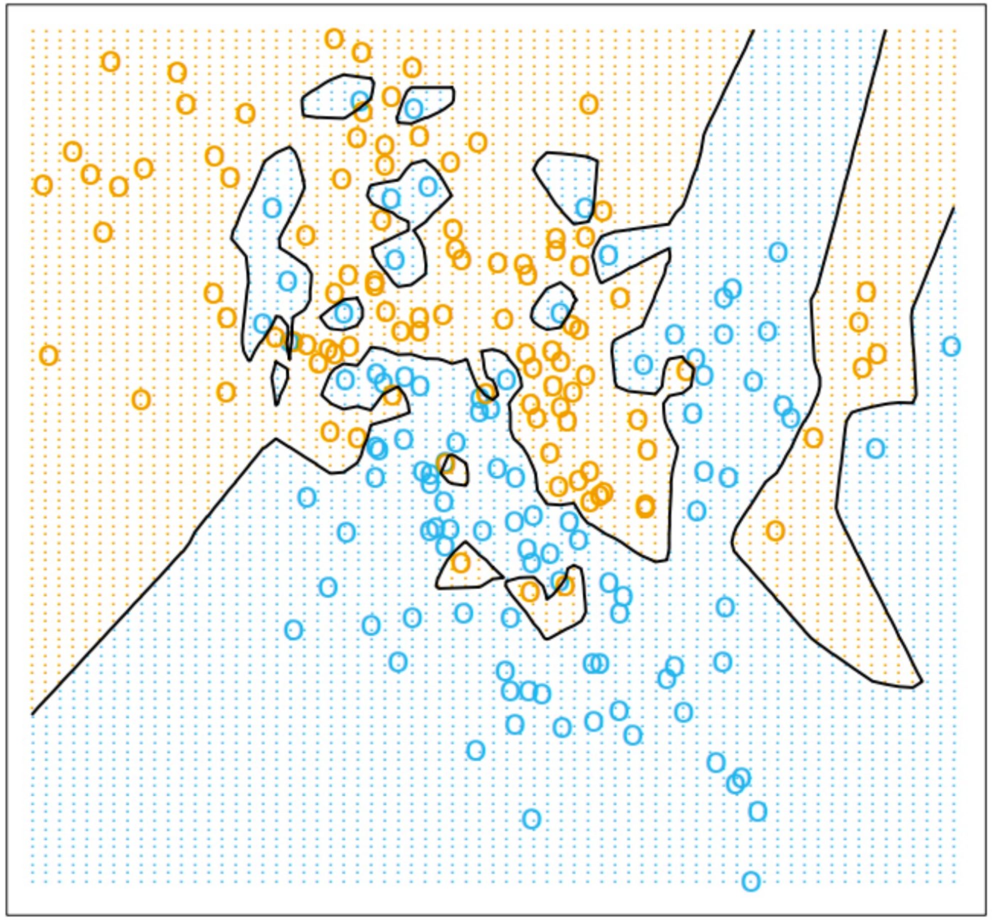
Empirical Risk Minimization

- Empirical risk minimization
  - Basic principle of machine learning
  - We want the model (classifier, etc) that does best on the true test distribution
  - Don’t know the true distribution so pick the best model on our actual training set
  - Finding “the best” model on the training set is phrased as an optimization problem

- Main worry: overfitting to the training set
  - Better with more training data (less sampling variance, training more like test)
  - Better if we limit the complexity of our hypotheses (regularization and/or small hypothesis spaces)

Overfitting

- K=1



[Image credit: ”The Elements of Statistical Learning”]

How to select a model?

- To solve a problem, which model should we choose?
  - KNN or logistic regression?
  - For KNN, which parameter k?

- Denote $\mathcal{M} = \{M_1, \ldots, M_d\}$ as all the models to choose

Select the one with the minimum training loss?

- Given the training set S

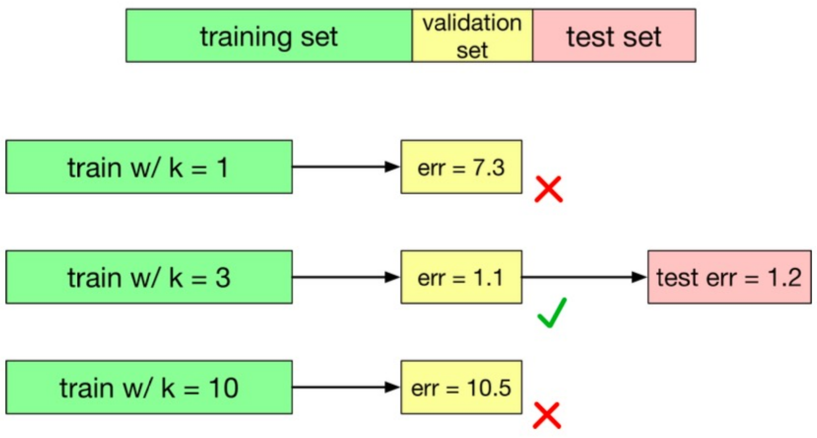
1. Train each model M_i on S, to get some hypothesis h_i.

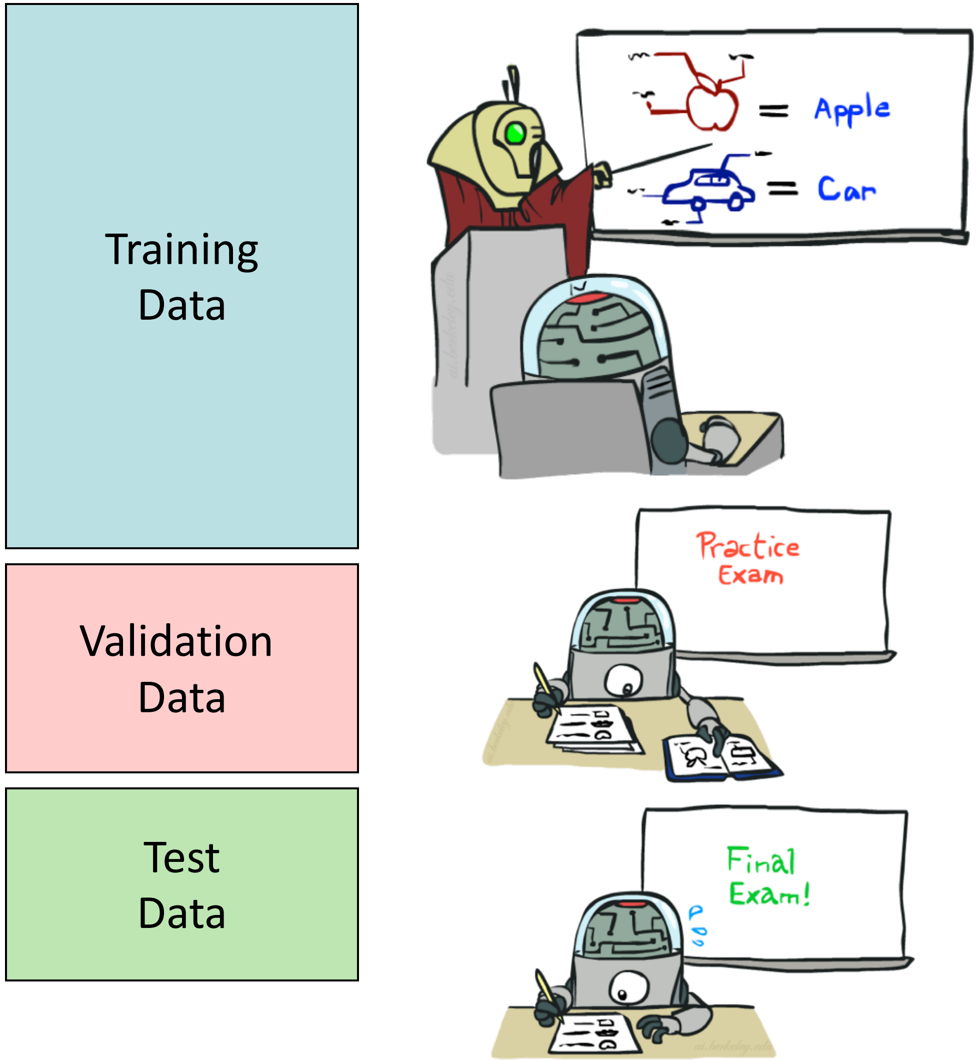
2. Pick the hypotheses with the smallest training error.

- What's the problem?
  - Lower training error prefers complex models
  - These models usually overfits

Solution: Hold-out cross validation

- How do we check that we’re not overfitting during training?
- Split training data into 3 different sets:
  - Training set
  - Validation set
  - Test set
- Experimentation cycle
  - Learn parameters on training set
  - Evaluate models on validation set
  - Very important: never “peek” at the test set!

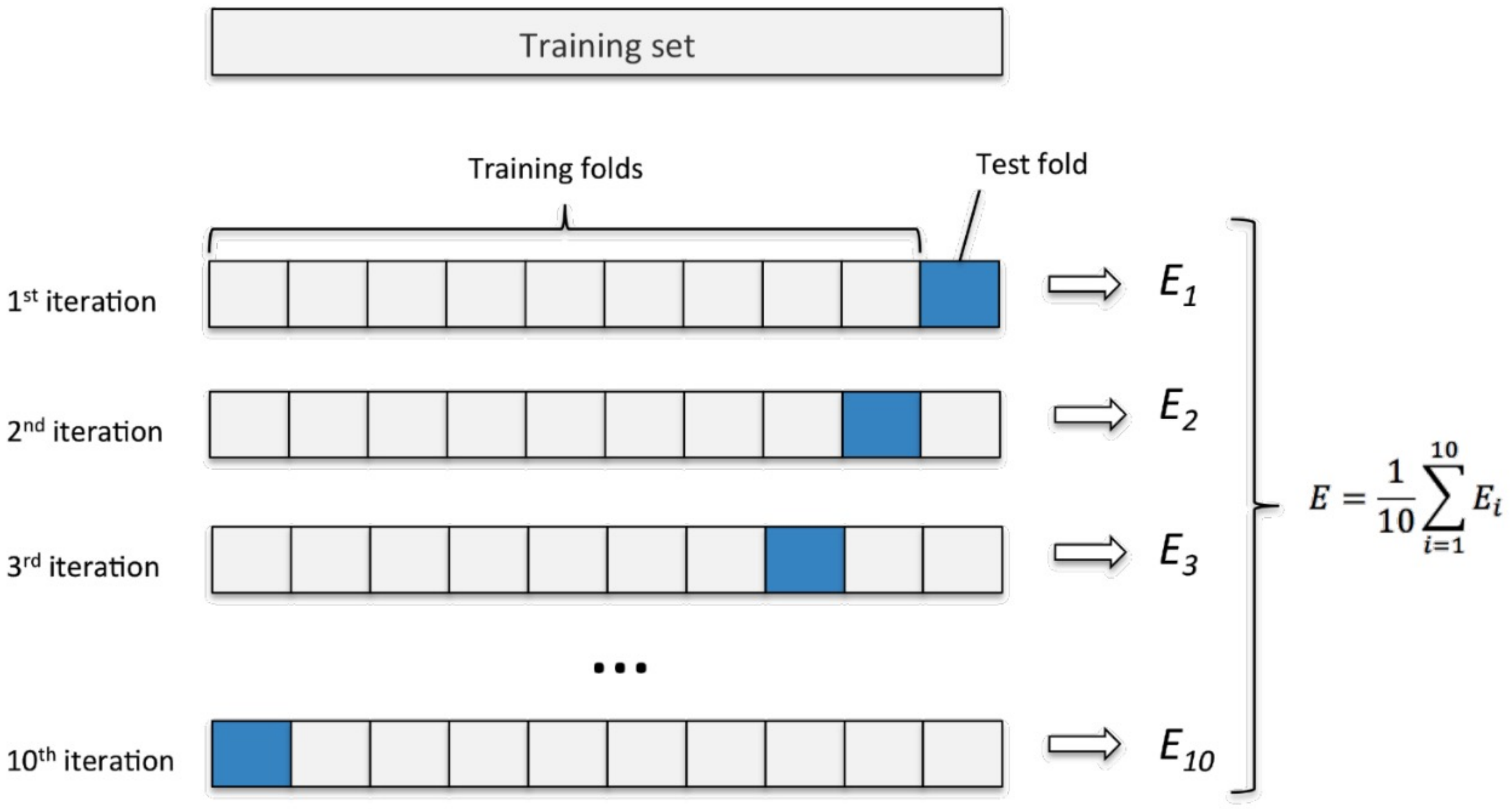




Hold-out cross validation (cont’d)

- The final model is only trained on 70% of the training set
- Especially in the case with small training set
  - Waste about 30% of the data

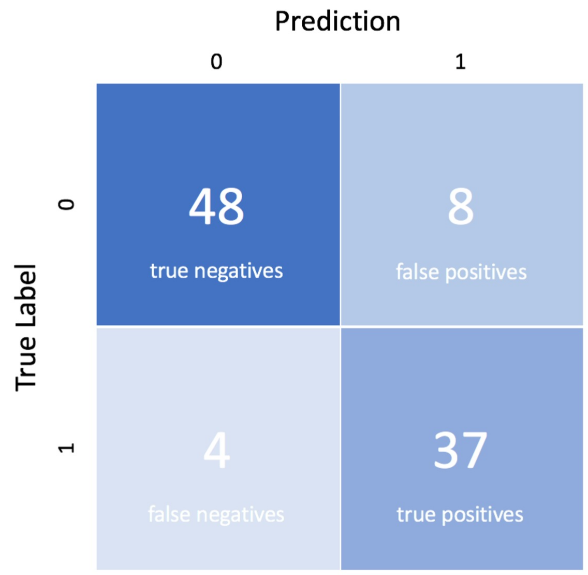
Improvement: k-fold cross validation

- 

Evaluation: Confusion matrix

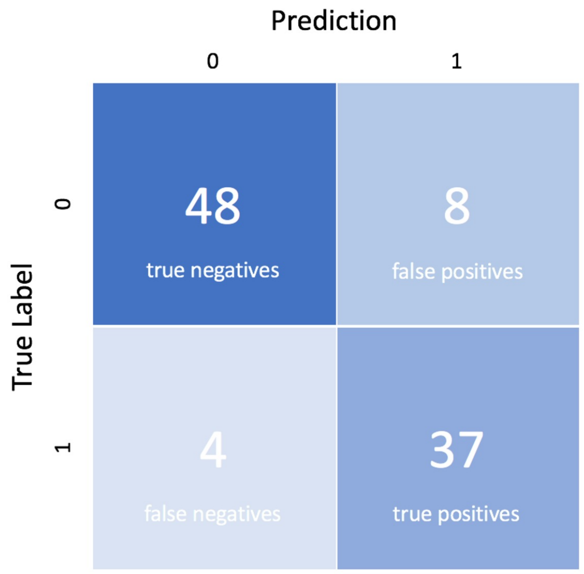
- Given a set of records containing positive and negative results, the computer is going to classify the records to be positive or negative

- Positive: The computer classifies the result to be positive
- Negative: The computer classifies the result to be negative
- True: What the computer classifies is true
- False: What the computer classifies is false



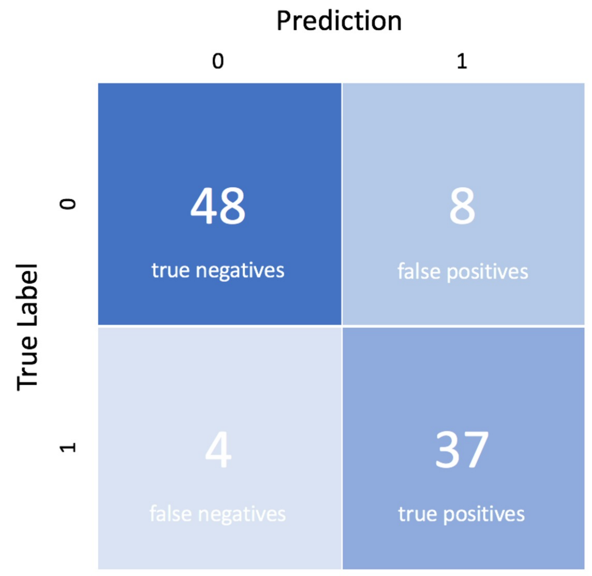
Accuracy

- Accuracy = $\frac{TN+TP}{TN+TP+FN+FP} = \frac{48+37}{48+37+4+8}$



Accuracy

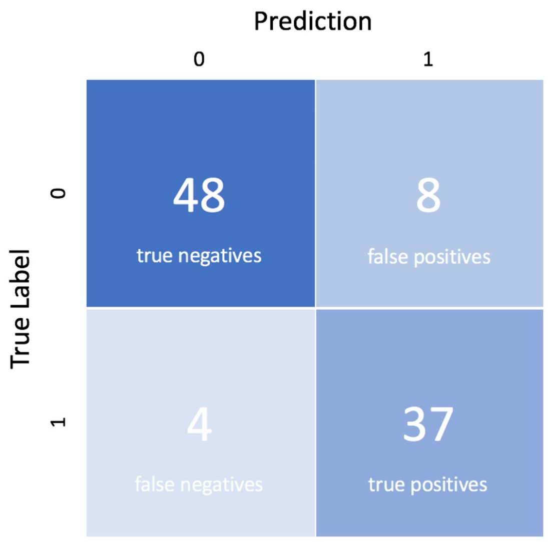
- Accuracy = $\frac{TN+TP}{TN+TP+FN+FP} = \frac{48+37}{48+37+4+8}$

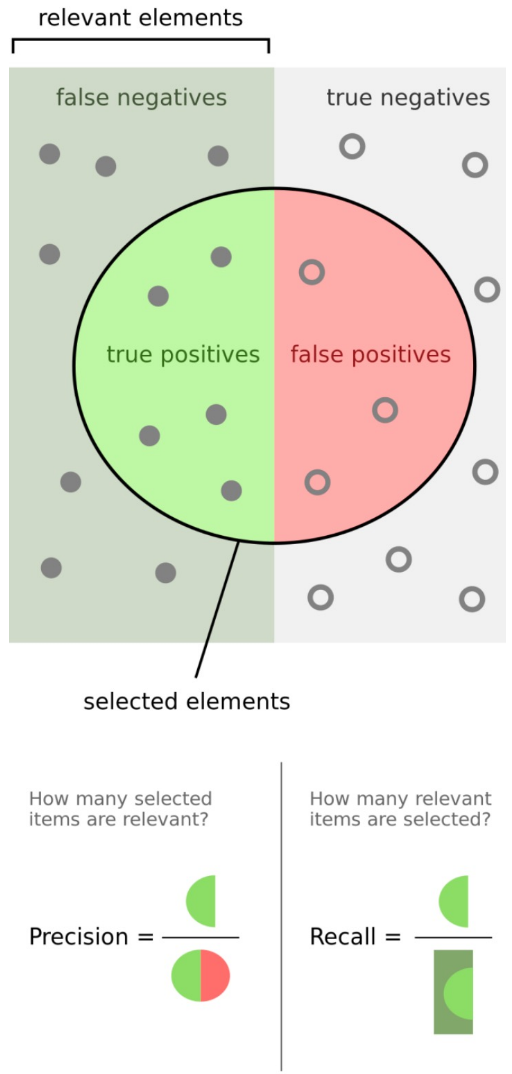


- Limitation
  - Suppose number of class 0 examples = 9990
  - Number of class 1 examples = 10
  - The model predicts every example as 0
  - Then the accuracy is 9990/10000=99.9%
  - The accuracy is misleading because the model does not detect any example in class 1

Other metrics

- Precision = $\frac{TP}{TP+FP} = \frac{37}{37+8}$
- Recall = $\frac{TP}{TP+FN} = \frac{37}{37+4}$
- F-measure = $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$





How to understand?

- A school is running a machine learning primary diabetes scanon all of its students
  - Diabetic (+) / Healthy (-)
  - False positive is just a false alarm
  - False negative
    - Prediction is healthy but is diabetic
    - Worst case among all 4 cases
- Accuracy
  - Accuracy = (TP+TN)/(TP+FP+FN+TN)
  - How many students did we correctly label out of all the students?

How to understand?

- A school is running a machine learning primary diabetes scanon all of its students
  - Diabetic (+) / Healthy (-)
  - False positive is just a false alarm
  - False negative
    - Prediction is healthy but is diabetic
    - Worst case among all 4 cases
- Precision
  - Precision = TP/(TP+FP)
  - How many of those who we labeled as diabetic are actually diabetic?

How to understand?

- A school is running a machine learning primary diabetes scanon all of its students
  - Diabetic (+) / Healthy (-)
  - False positive is just a false alarm
  - False negative
    - Prediction is healthy but is diabetic
    - Worst case among all 4 cases
- Recall (sensitivity)
  - Recall = TP/(TP+FN)
  - Of all the people who are diabetic, how many of those we correctly predict?

F1 score (F-Score / F-Measure)

- F1 Score = 2*(Recall * Precision) / (Recall + Precision)
- F1 Score=$\frac{1}{2}((1/\text{Recall} + 1/\text{Precision}))^{-1}$

- Harmonic mean (average) of the precision and recall
- F1 Score is best if there is some sort of balance between precision (p) & recall (r) in the system.
- Oppositely F1 Score isn't so high if one measure is improved at the expense of the other.
- For example, if P is 1 & R is 0, F1 score is 0.

Which to choose?

- Accuracy
  - A great measure
  - But only when you have symmetric datasets

- Precision
  - Want to be more confident of your TP
  - E.g. spam emails. We’d rather have some spam emails in inbox rather than some regular emails in your spam box.

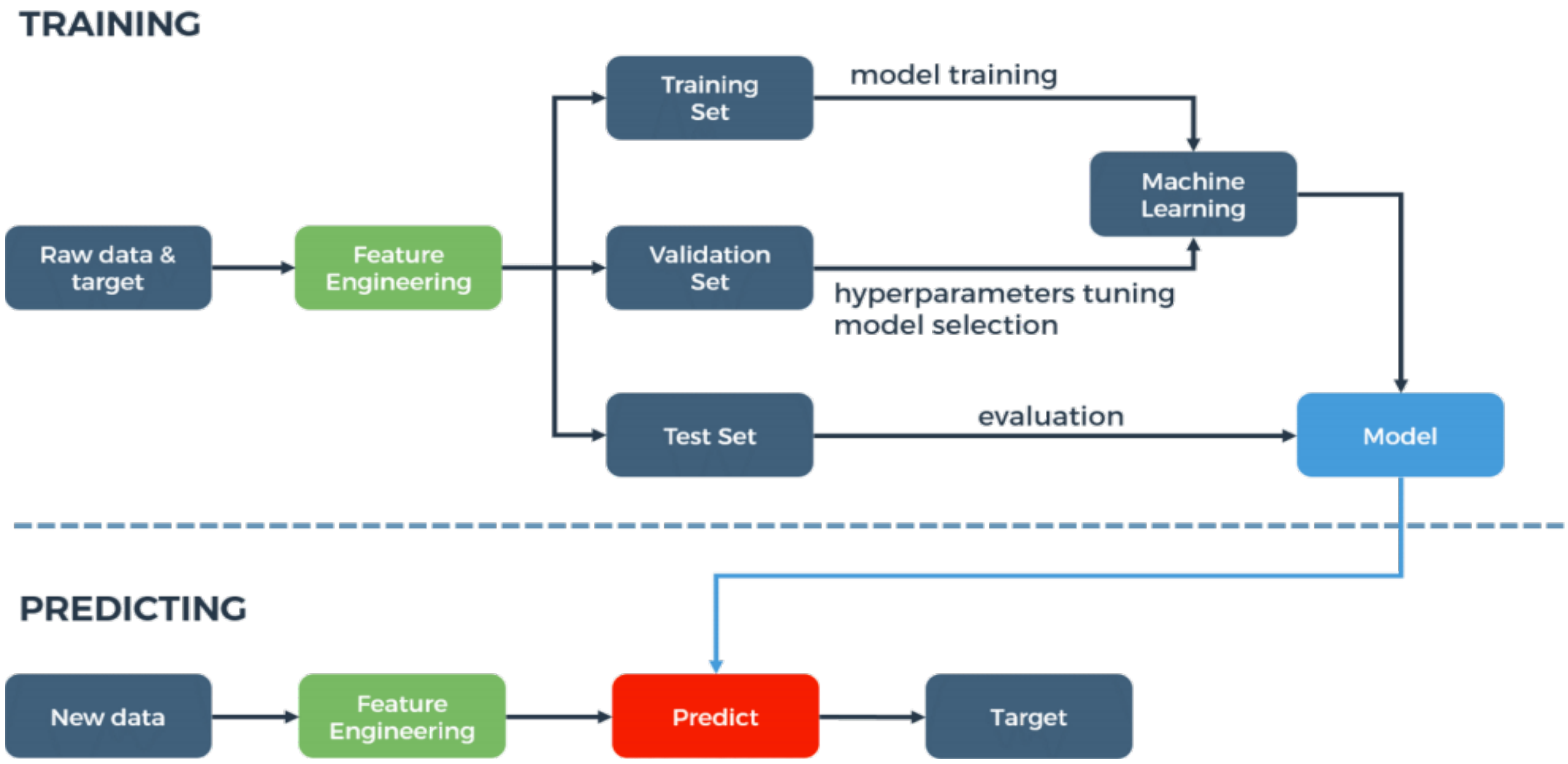
Which to choose?

- Recall
  - If FP is far better than FN or if the occurrence of FN is unaccepted/intolerable
  - Would like more extra FP (false alarms) over saving some FN
  - E.g. diabetes. We’d rather get some healthy people labeled diabetic over leaving a diabetic person labeled healthy
- F1 score
  - If the costs of FP and FN are both important

Quiz

- Suppose we have a test for a disease. There are 800 test persons. 50 of them are sick. The test yields 100 positive results. 40 of the positively tested persons are positive in reality.
- Create the confusion matrix and calculate the Accuracy, Precision, Recall, and the F1-score.

Machine Learning Process



Summary

- Types of machine learning
  - Supervised/Unsupervised/Reinforcement

- Machine learning process
  - Feature representation
  - KNN algorithm
  - Model selection
  - Evaluation metrics: Accuracy/Precision/Recall/F1-score