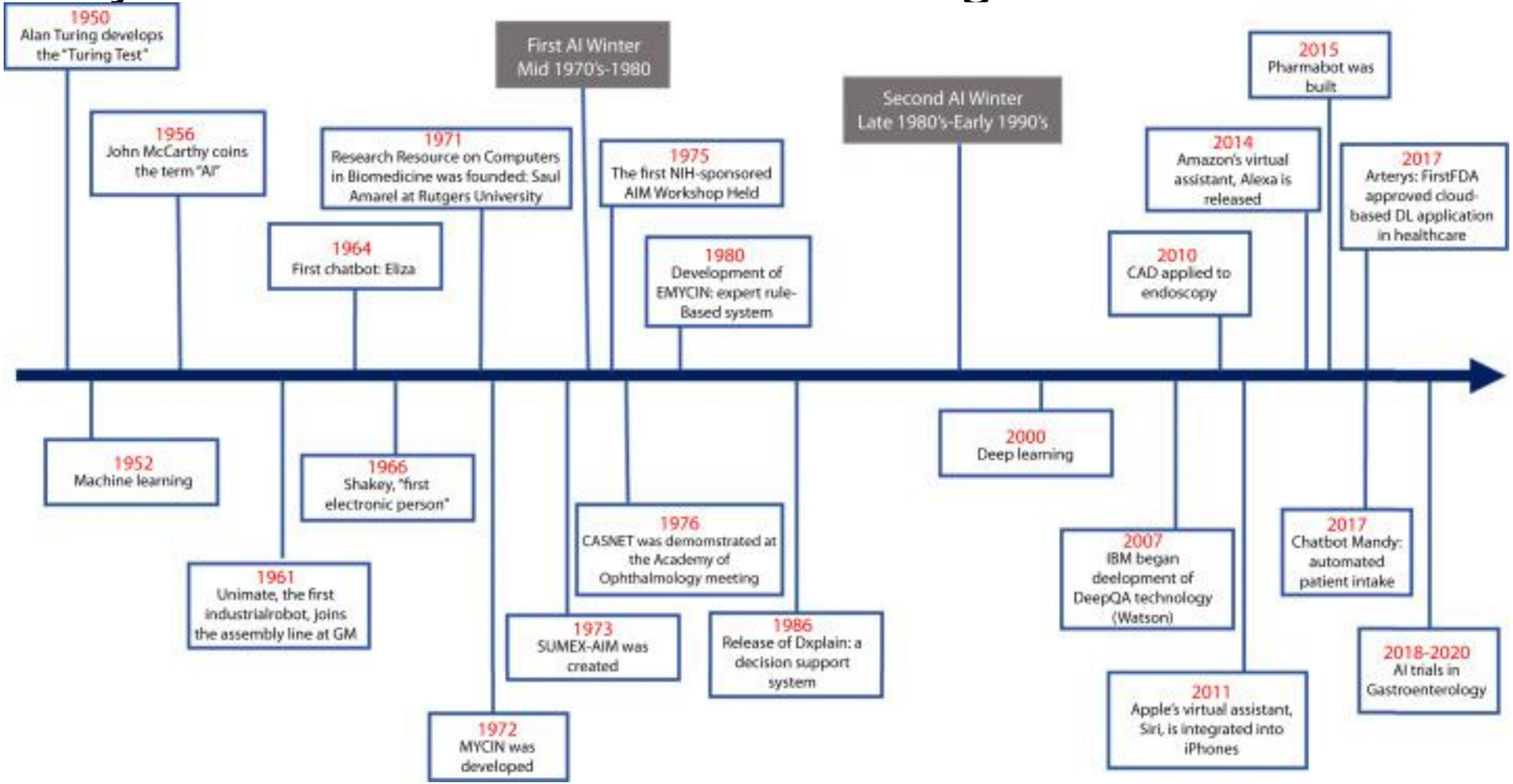
CYBER 207
Applied Machine Learning for Cybersecurity

Summer 2023

Week 1

History and Timeline of Artificial Intelligence



1950
Alan Turing develops the "Turing Test"

1952
Machine learning

1956
John McCarthy coins the term "AI"

1961
Unimate, the first industrialrobot, joins the assembly line at GM

1964
First chatbot: Eliza

1966
Shakey, "first electronic person"

1971
Research Resource on Computers in Biomedicine was founded: Saul Amarel at Rutgers University

1972
MYCIN was developed

1973
SUMEX-AIM was created

First AI Winter
Mid 1970's-1980

1975
The first NIH-sponsored AIM Workshop Held

1976
CASNET was demomstrated at the Academy of Ophthalmology meeting

1980
Development of EMYCIN: expert rule-Based system

1986
Release of Dxplain: a decision support system

Second AI Winter
Late 1980's-Early 1990's

2000
Deep learning

2007
IBM began deelopment of DeepQA technology (Watson)

2010
CAD applied to endoscopy

2011
Apple's virtual assistant, Siri, is integrated into iPhones

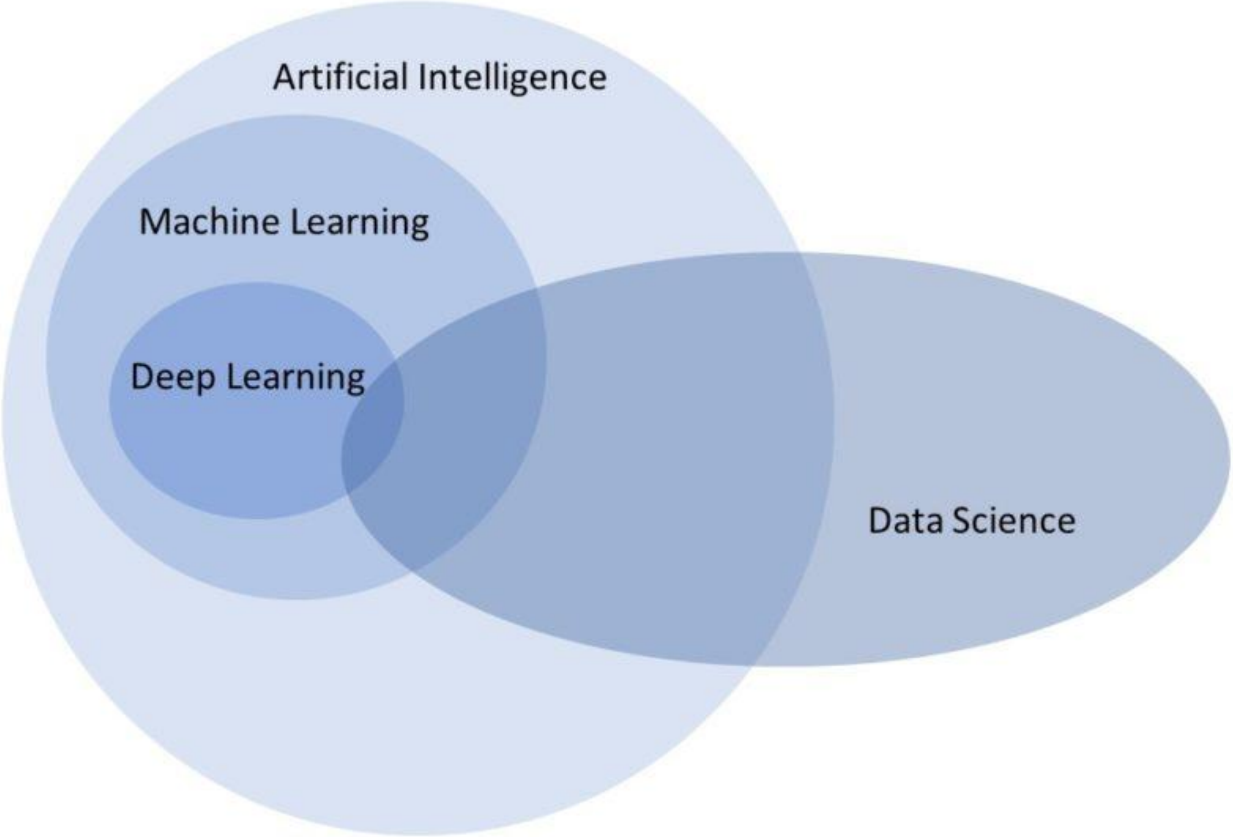
2014
Amazon's virtual assistant, Alexa is released

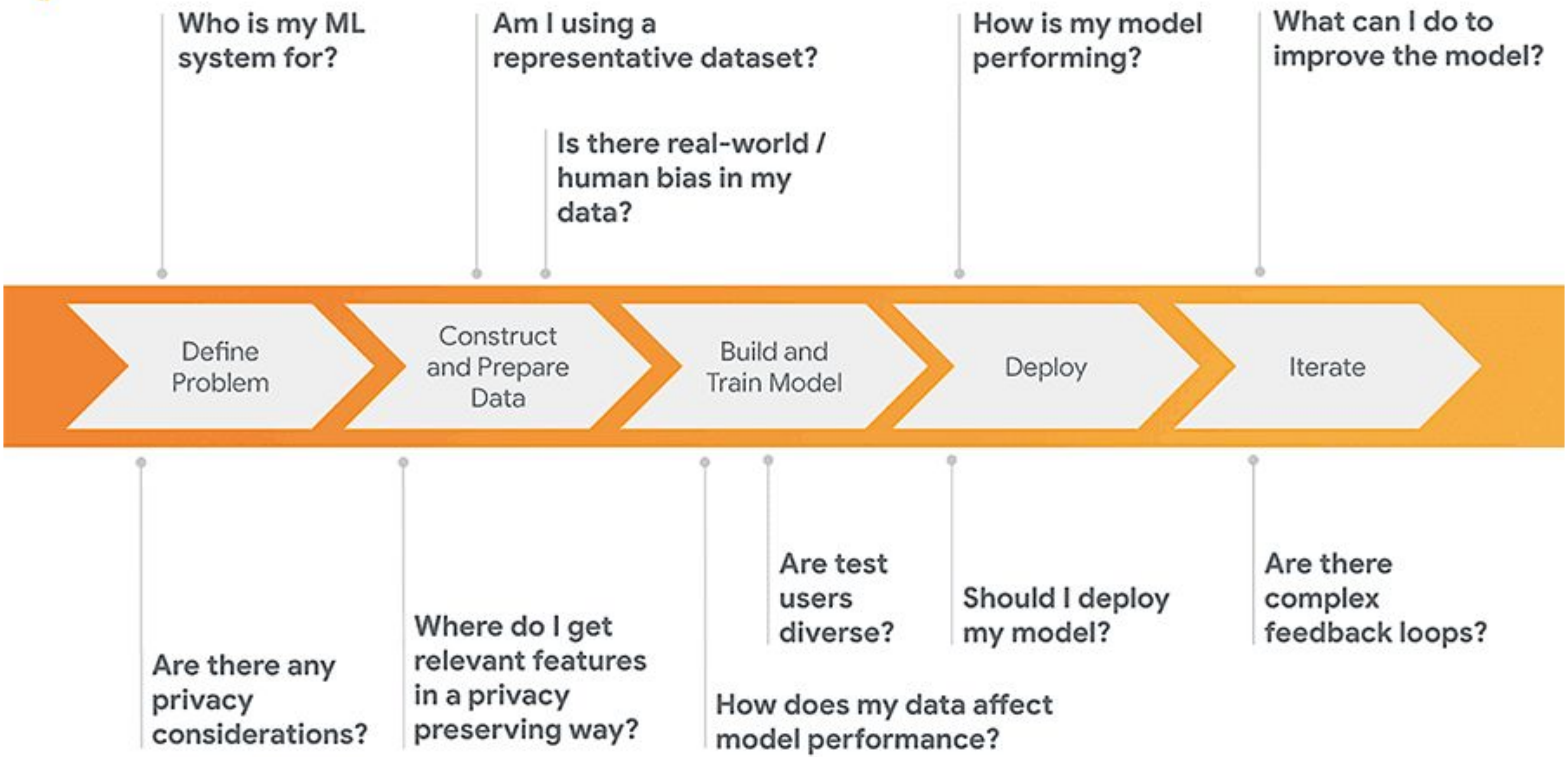
2015
Pharmabot was built

2017
Arterys: FirstFDA approved cloud-based DL application in healthcare

2017
Chatbot Mandy: automated patient intake

2018-2020
AI trials in Gastroenterology


Artificial Intelligence
Machine Learning
Deep Learning
Data Science



Who is my ML system for?

Am I using a representative dataset?

Is there real-world / human bias in my data?

How is my model performing?

What can I do to improve the model?

Define Problem

Construct and Prepare Data

Build and Train Model

Deploy

Iterate

Are there any privacy considerations?

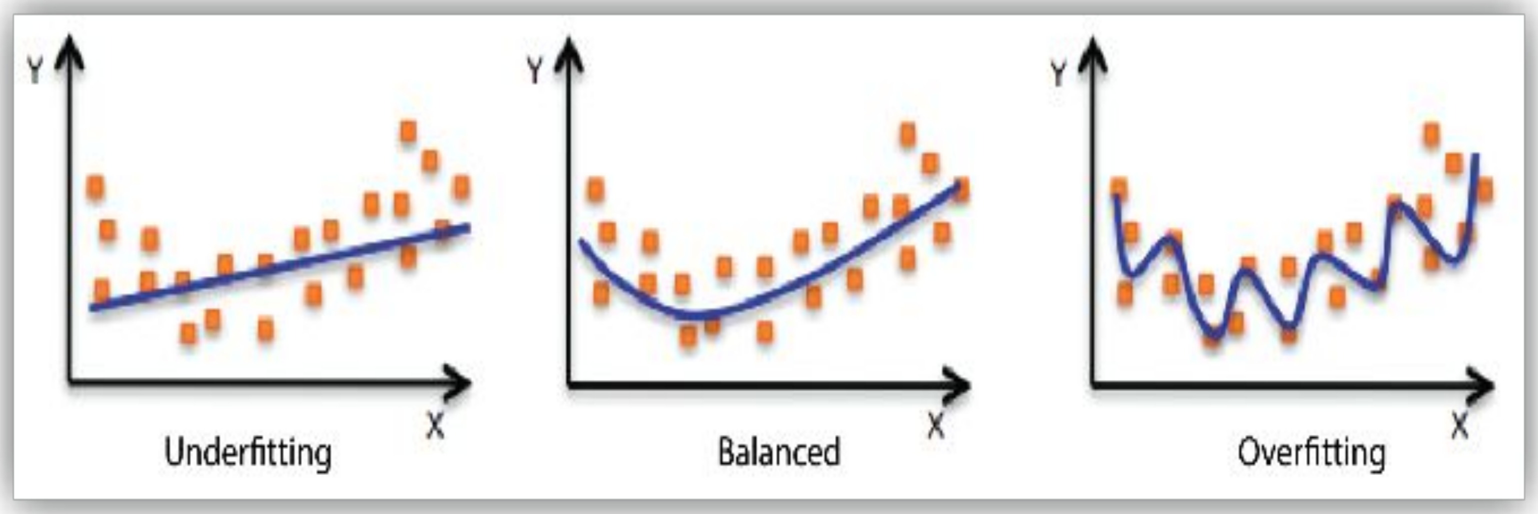
Where do I get relevant features in a privacy preserving way?

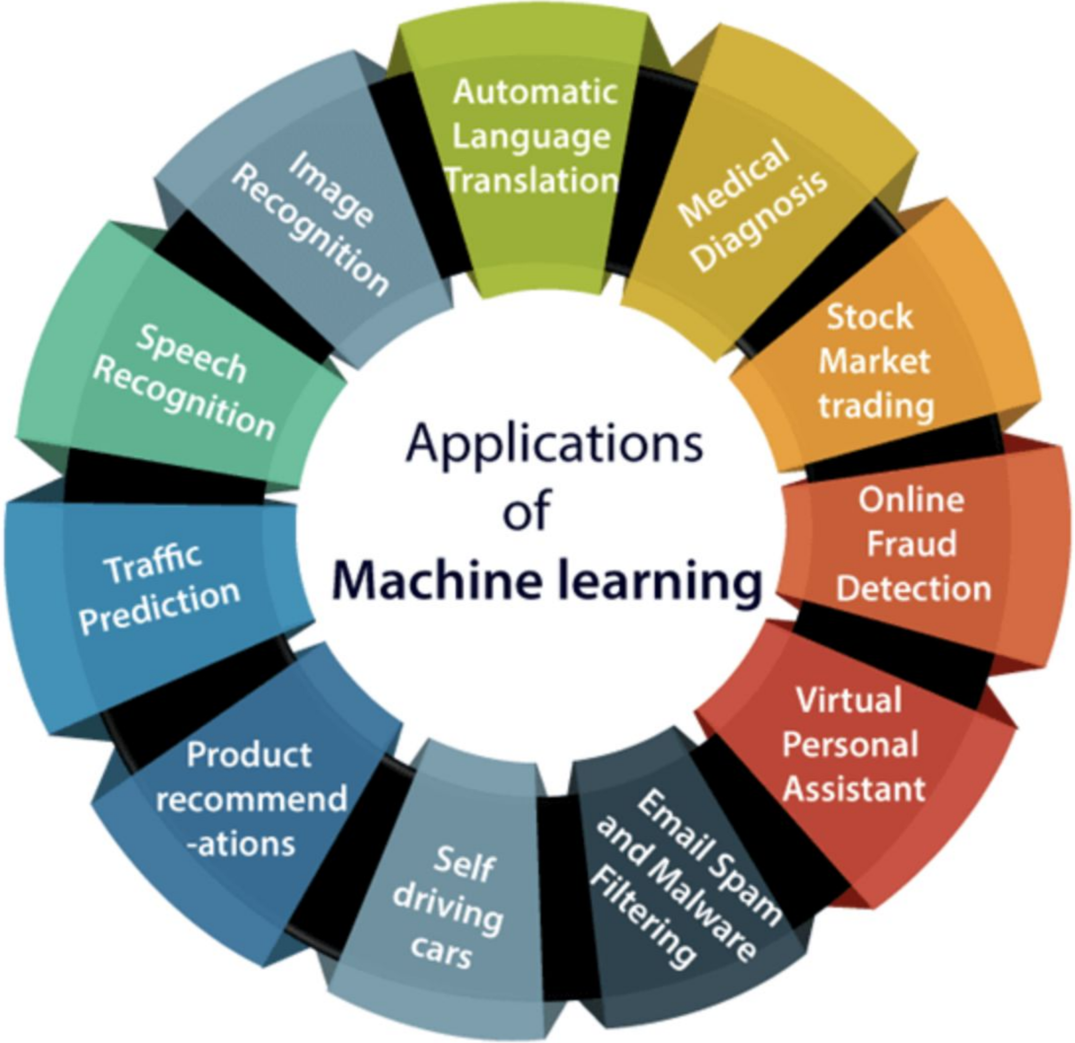
How does my data affect model performance?

Are test users diverse?

Should I deploy my model?

Are there complex feedback loops?


Y
X
Underfitting
Y
X
Balanced
Y
X
Overfitting



Applications of **Machine learning**

Automatic Language Translation

Medical Diagnosis

Stock Market trading

Online Fraud Detection

Virtual Personal Assistant

Email Spam and Malware Filtering

Self driving cars

Product recommend-ations

Traffic Prediction

Speech Recognition

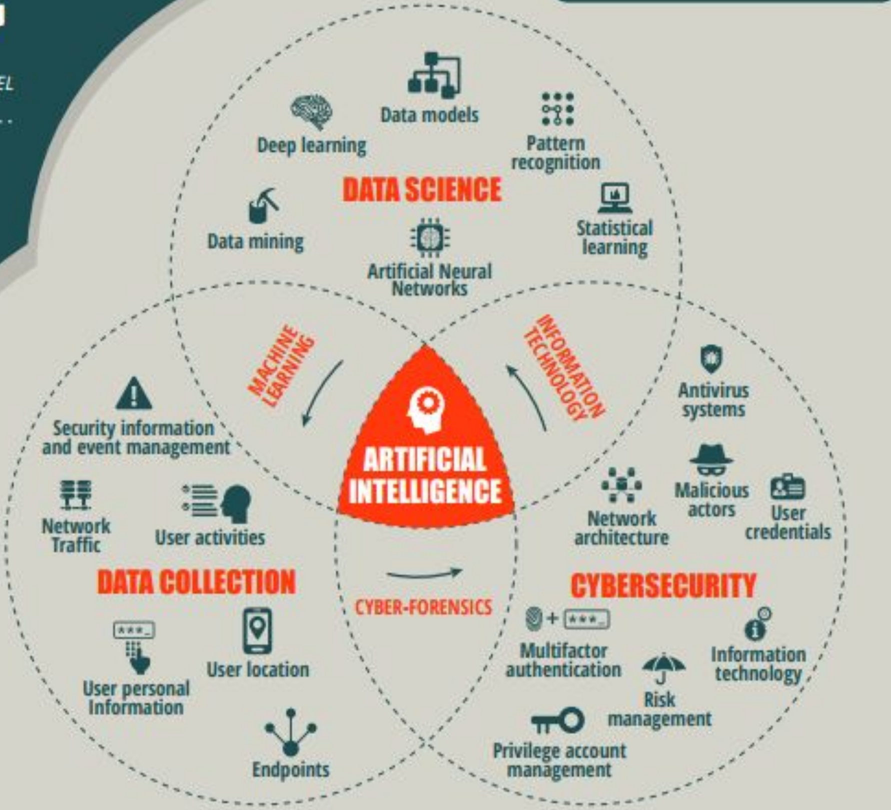
Image Recognition

Analytics and Forensics

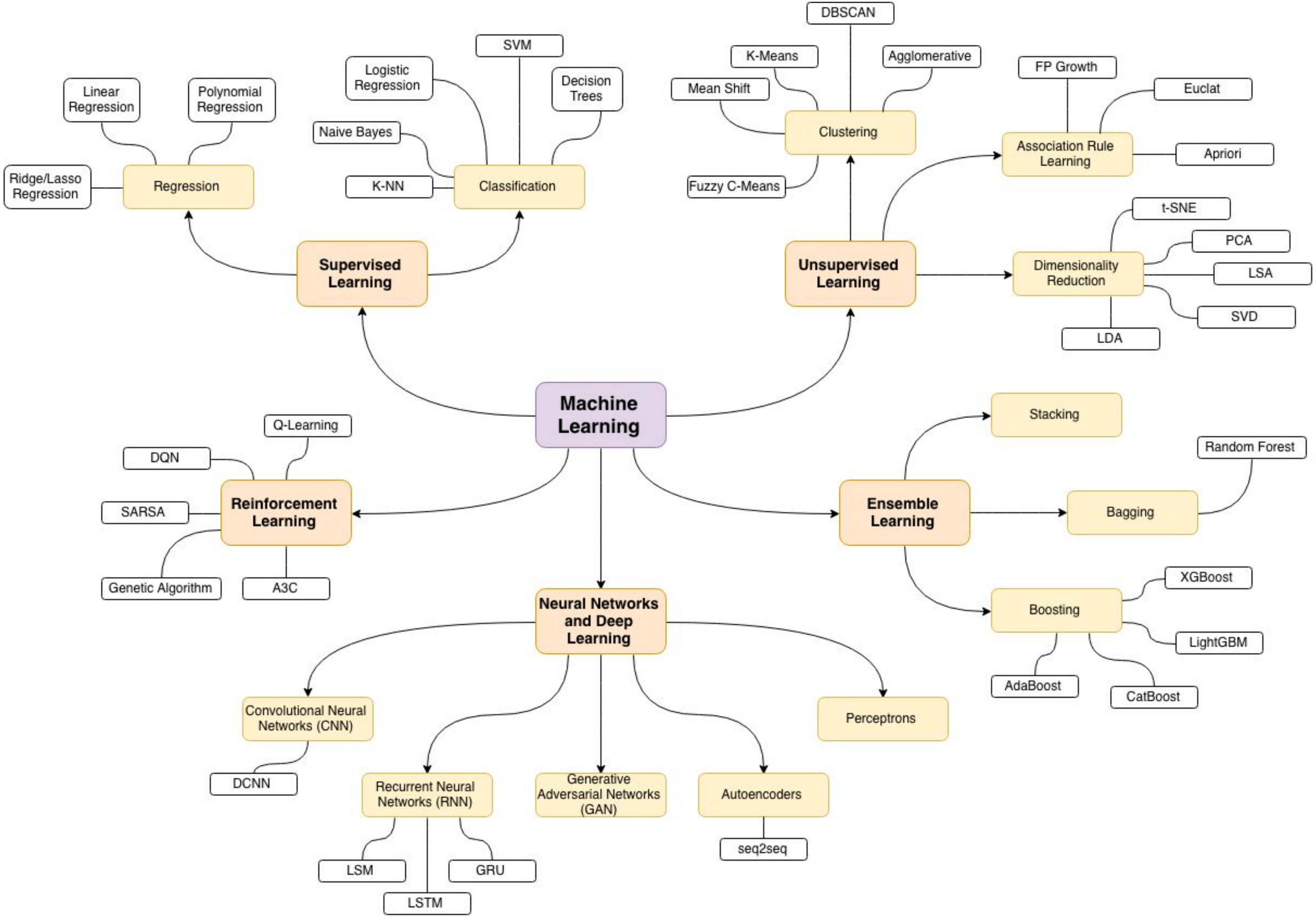
WHAT CAN MACHINE LEARNING DO FOR CYBERSECURITY?

A POTENT NEW ARSENAL FOR IT AND CYBERSECURITY PERSONNEL

- User entity behavioral analytics, deep learning, automation
- Assist IT professionals and defend against new cyberthreats
- Better predictive models, lower FPR, distill new metrics
- Fraud and anomaly detection
- Defend against new cyberthreats
- Better use of internal data and global repositories
- Tackle device influx and enhanced data loss prevention (DLP) solutions



- ▲ **Data Science**: Applying machine learning and creating new data models to combat new threats
- ▲ **Data Collection**: Harnessing the power of data from a wide spectrum of sources
- ▲ **Cybersecurity**: Domain-specific knowledge and versatility in an ever-changing environment



Machine Learning

Supervised Learning
- Regression
  - Linear Regression
  - Polynomial Regression
  - Ridge/Lasso Regression
- Classification
  - Logistic Regression
  - SVM
  - Decision Trees
  - Naive Bayes
  - K-NN

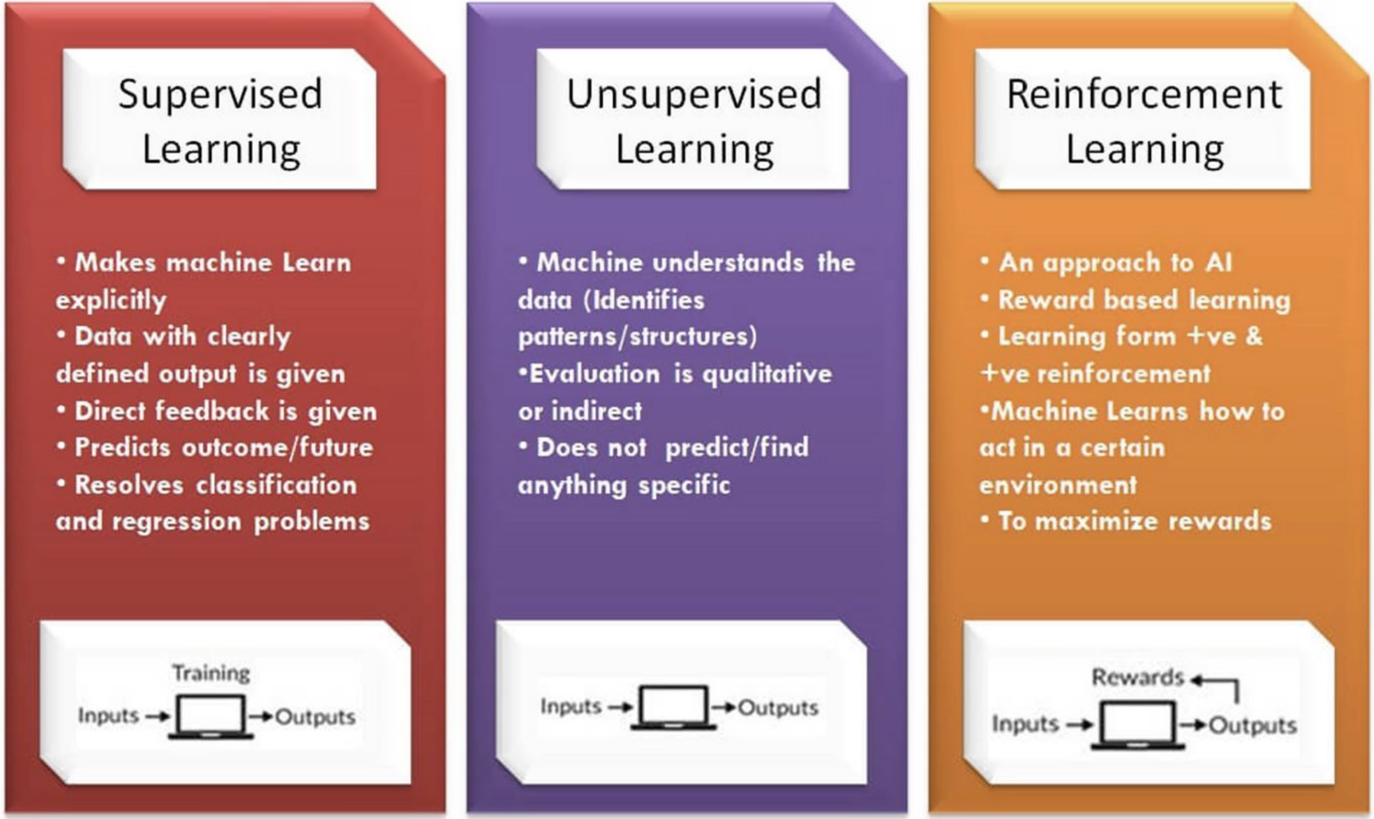
Unsupervised Learning
- Clustering
  - DBSCAN
  - K-Means
  - Agglomerative
  - Mean Shift
  - Fuzzy C-Means
- Association Rule Learning
  - FP Growth
  - Euclat
  - Apriori
- Dimensionality Reduction
  - t-SNE
  - PCA
  - LSA
  - SVD
  - LDA

Reinforcement Learning
- Q-Learning
- DQN
- SARSA
- Genetic Algorithm
- A3C

Ensemble Learning
- Stacking
- Bagging
  - Random Forest
- Boosting
  - XGBoost
  - LightGBM
  - AdaBoost
  - CatBoost

Neural Networks and Deep Learning
- Convolutional Neural Networks (CNN)
  - DCNN
- Recurrent Neural Networks (RNN)
  - LSM
  - LSTM
  - GRU
- Generative Adversarial Networks (GAN)
- Autoencoders
  - seq2seq
- Perceptrons

Fundamentals



Supervised Learning

- Makes machine Learn explicitly
- Data with clearly defined output is given
- Direct feedback is given
- Predicts outcome/future
- Resolves classification and regression problems

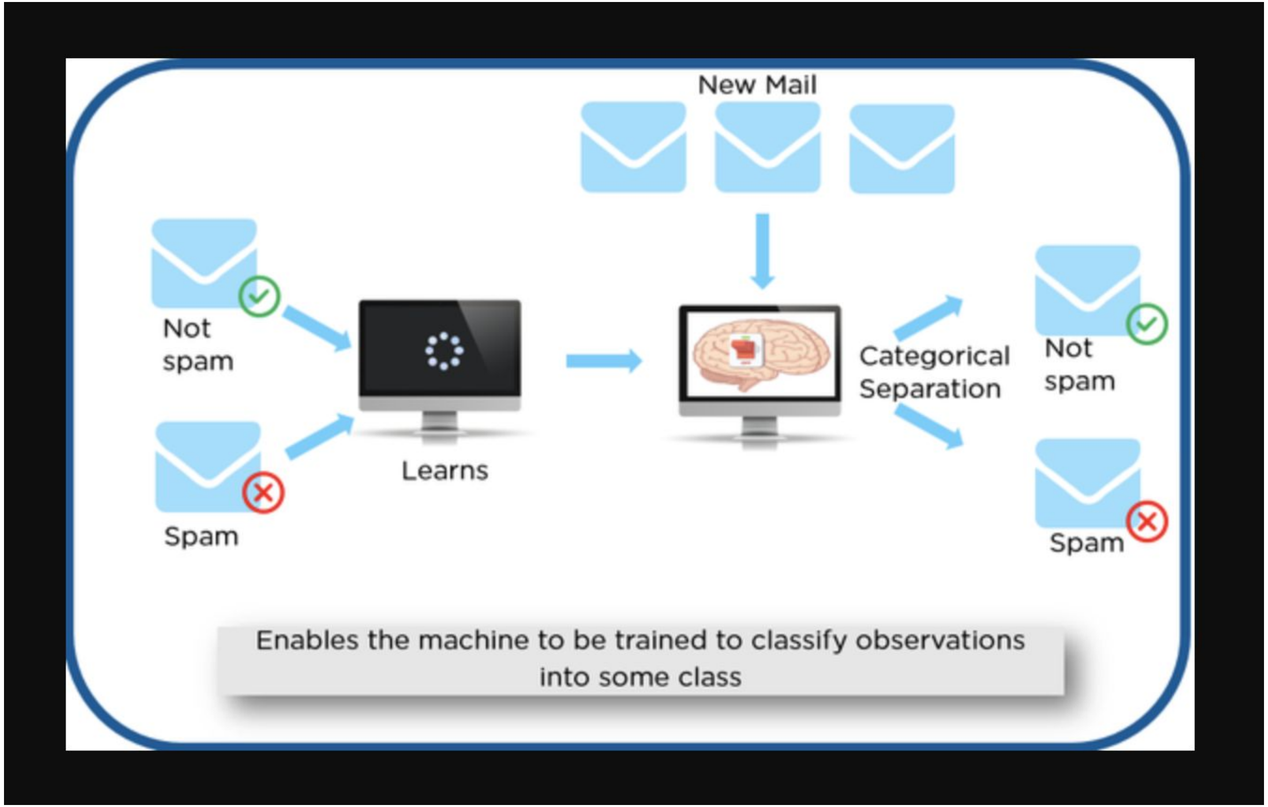
Unsupervised Learning

- Machine understands the data (Identifies patterns/structures)
- Evaluation is qualitative or indirect
- Does not predict/find anything specific

Reinforcement Learning

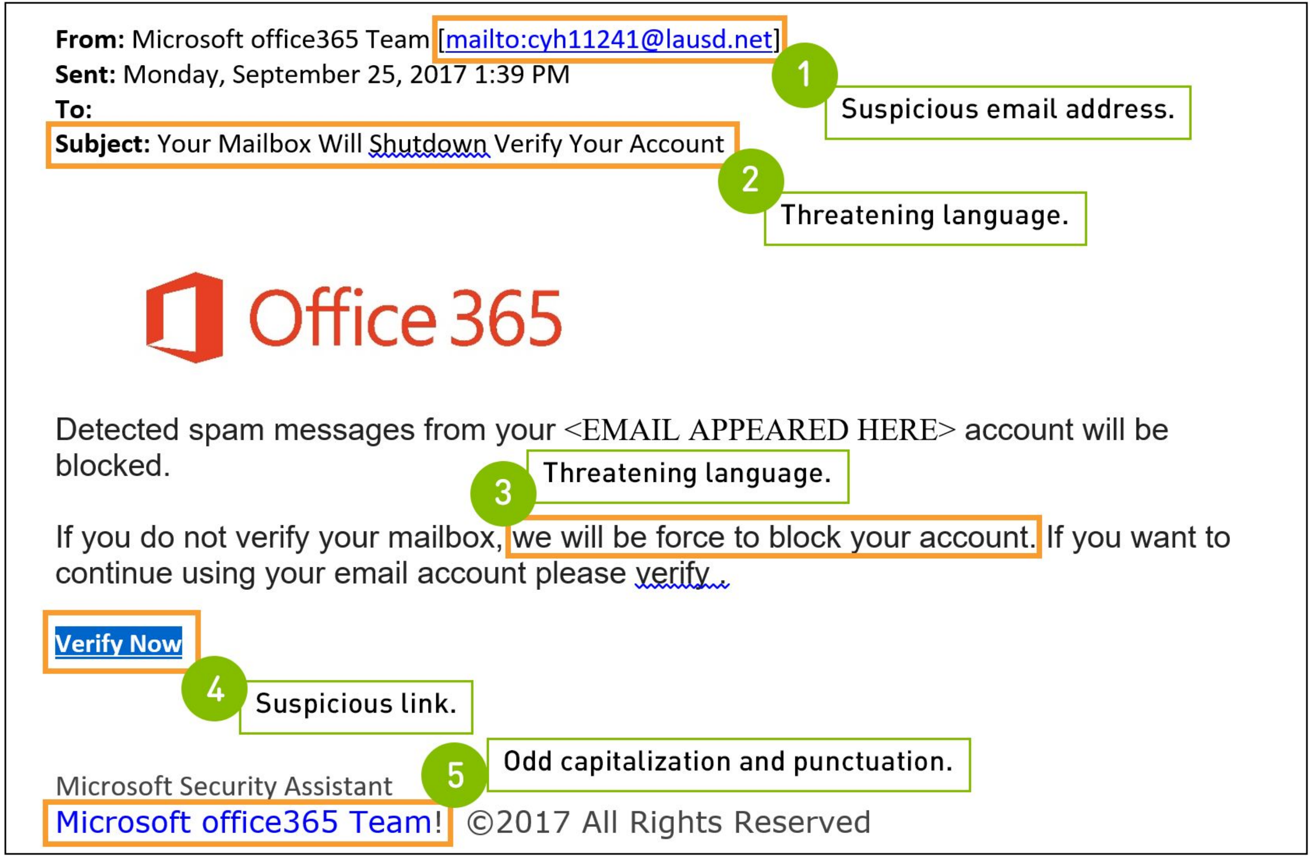
- An approach to AI
- Reward based learning
- Learning form +ve & +ve reinforcement
- Machine Learns how to act in a certain environment
- To maximize rewards

Applications: Spam and Non Spam



Enables the machine to be trained to classify observations into some class

Applications: Spam and Non Spam



From: Microsoft office365 Team [mailto:cyh11241@lausd.net]
Sent: Monday, September 25, 2017 1:39 PM
To:
Subject: Your Mailbox Will Shutdown Verify Your Account

Office 365

Detected spam messages from your <EMAIL APPEARED HERE> account will be blocked.

If you do not verify your mailbox, we will be force to block your account. If you want to continue using your email account please verify .

Verify Now

Microsoft Security Assistant
Microsoft office365 Team! ©2017 All Rights Reserved

1 Suspicious email address.

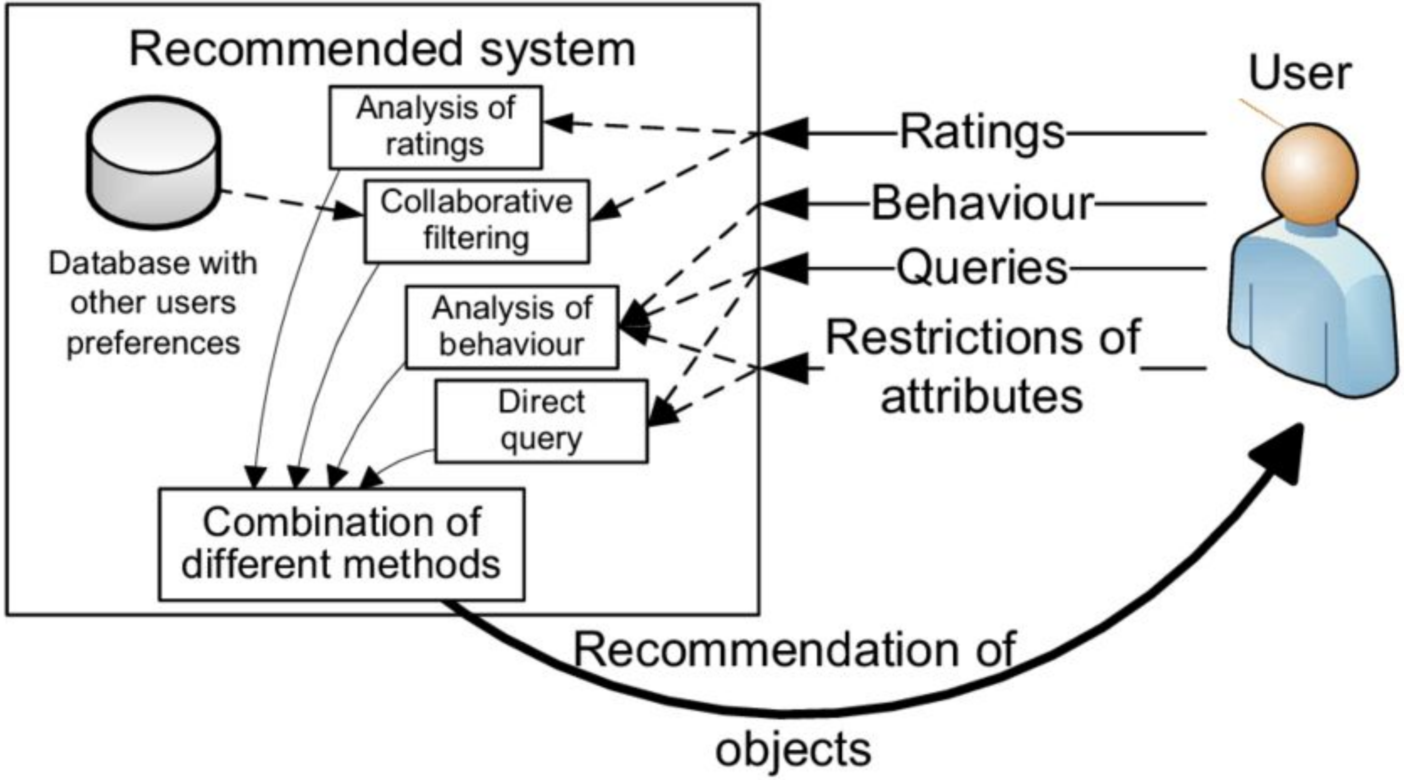
2 Threatening language.

3 Threatening language.

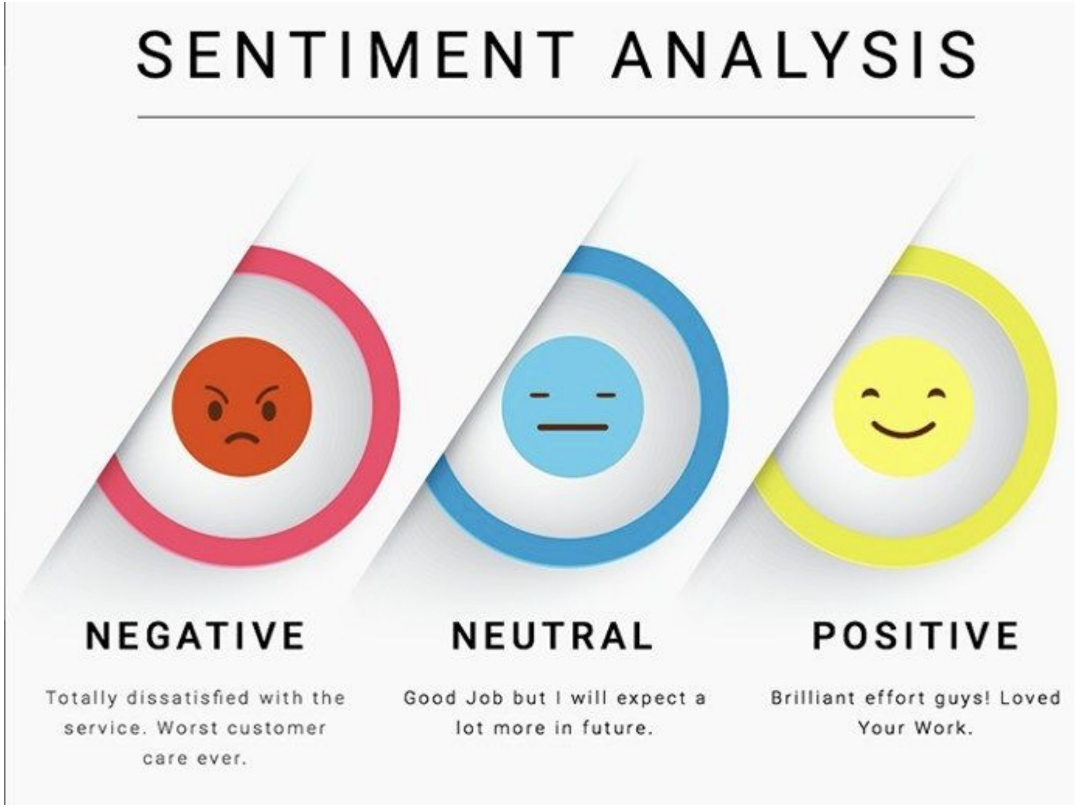
4 Suspicious link.

5 Odd capitalization and punctuation.

Applications: Recommendation Systems


Recommended system
User
Database with other users preferences
Analysis of ratings
Collaborative filtering
Analysis of behaviour
Direct query
Combination of different methods
Ratings
Behaviour
Queries
Restrictions of attributes
Recommendation of objects

Applications: Sentiment Analysis

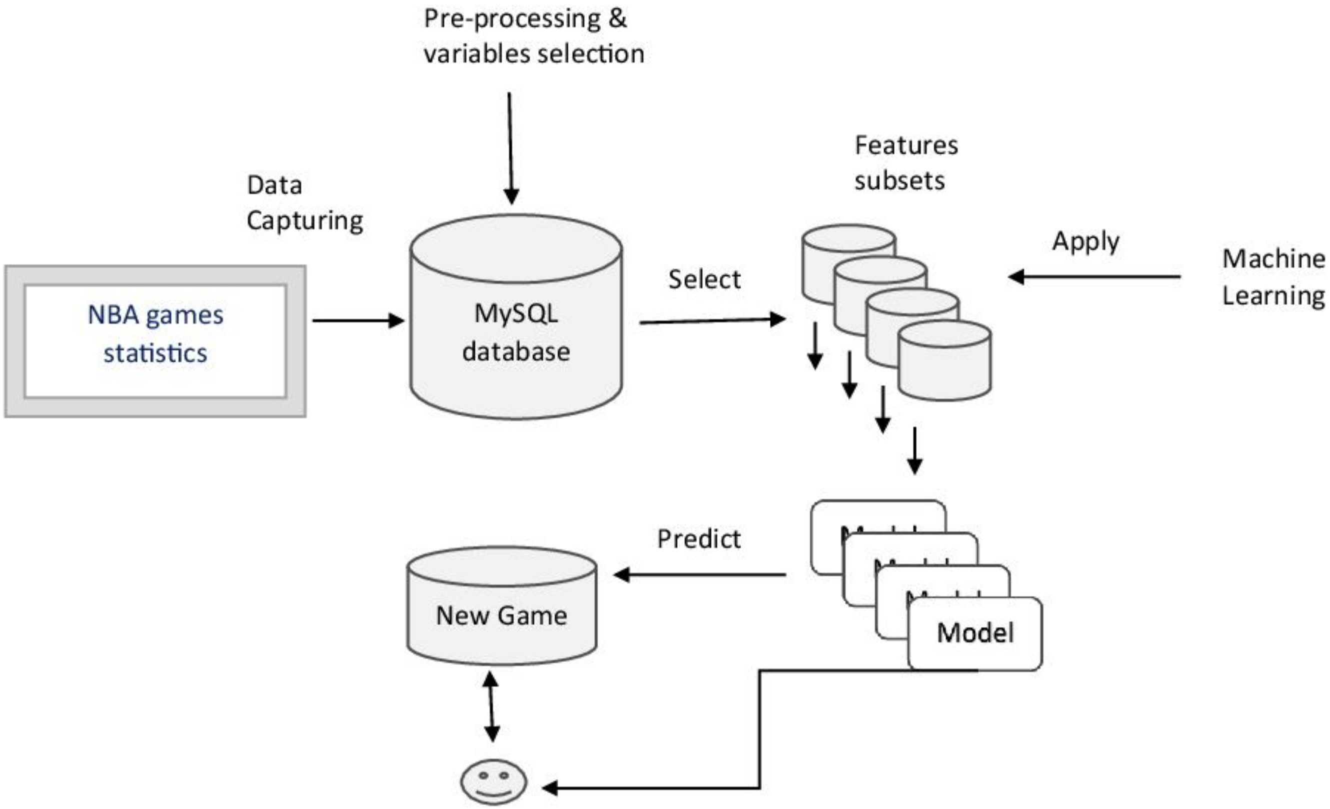


Totally dissatisfied with the service. Worst customer care ever.

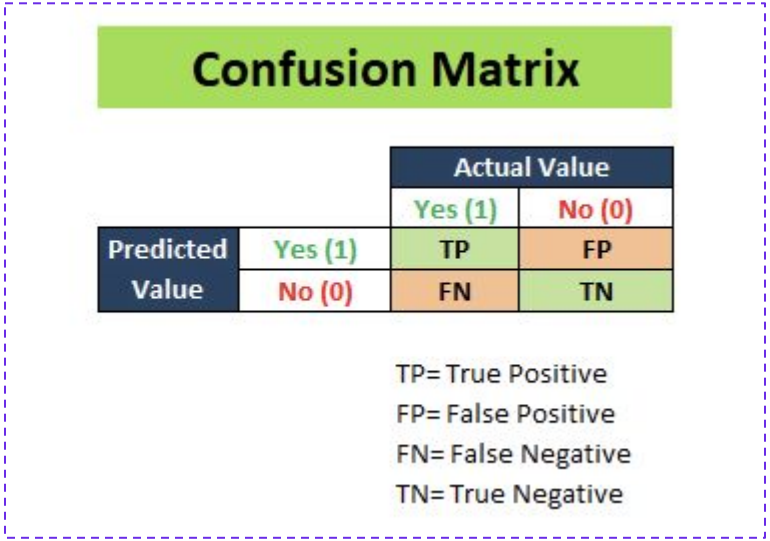
Good Job but I will expect a lot more in future.

Brilliant effort guys! Loved Your Work.

Applications: Sports Prediction



Evaluating the Models



Confusion Matrix

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	TP	FP
	No (0)	FN	TN

TP= True Positive

FP= False Positive

FN= False Negative

TN= True Negative

- If you have supervised data, you will want to maximize an objective function.
    - **Precision**: $TP \div (TP + FP)$ % positives correctly identifed
    - **Recall**: $TP \div (TP + FN)$ % existing positives identified
    - **Optimal point** on ROC (precision/recall) curve
    - **Accuracy**: $(TP + TN) \div (TP + TN + FP + FN)$
    - **F-test**: $2 \cdot (P \cdot R) \div (P + R)$