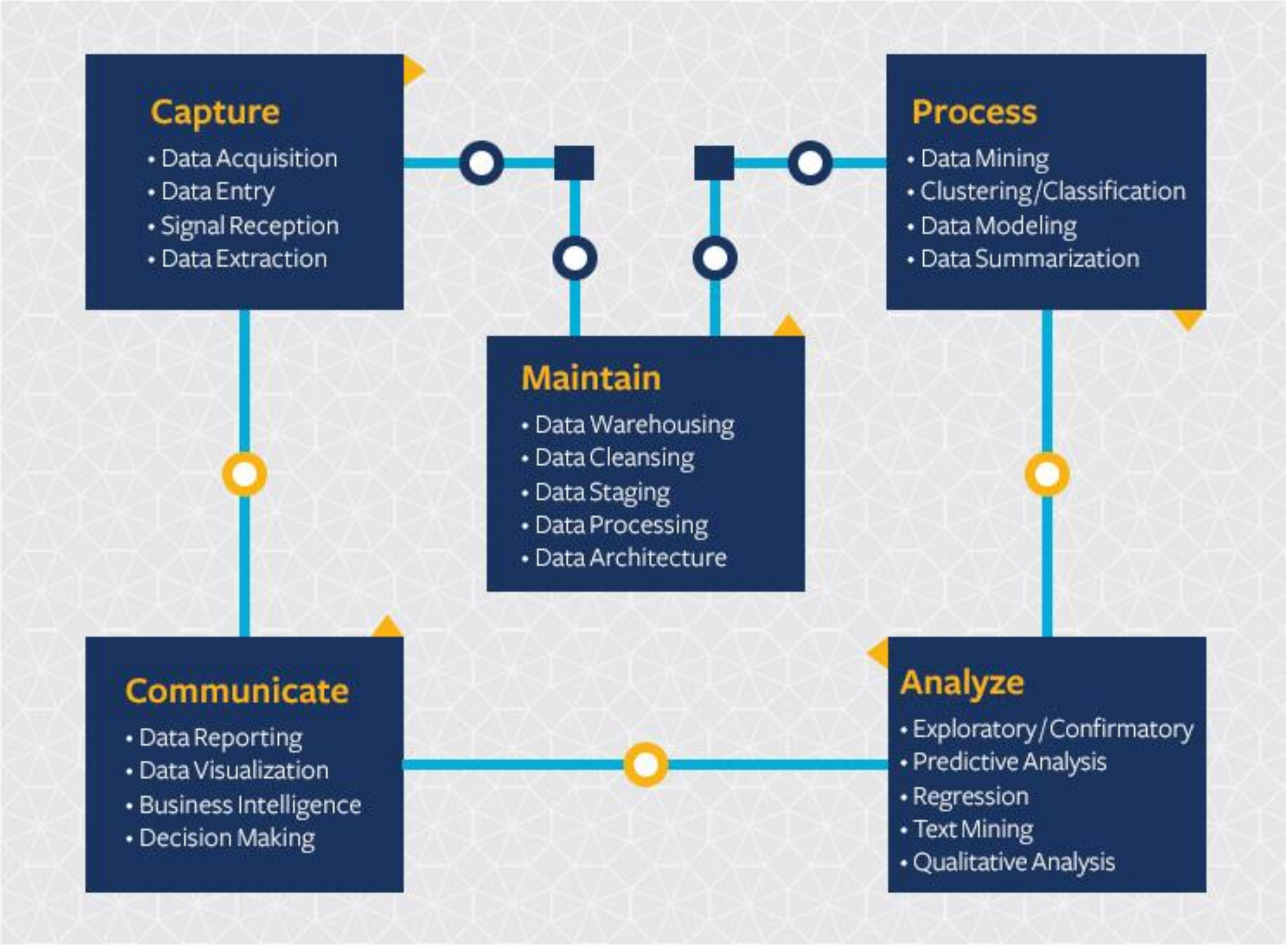
Machine Learning

Introduction to Data Science and Machine Learning

Data Science

- What is Data science?
  - Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate **knowledge and insights** from noisy, **structured, and unstructured data**.
  - Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.
  - https://en.wikipedia.org/wiki/Data_science

Data Science



Capture

- Data Acquisition
- Data Entry
- Signal Reception
- Data Extraction

Process

- Data Mining
- Clustering/Classification
- Data Modeling
- Data Summarization

Maintain

- Data Warehousing
- Data Cleansing
- Data Staging
- Data Processing
- Data Architecture

Communicate

- Data Reporting
- Data Visualization
- Business Intelligence
- Decision Making

Analyze

- Exploratory/Confirmatory
- Predictive Analysis
- Regression
- Text Mining
- Qualitative Analysis

- Ref: https://datascience.berkeley.edu/about/what-is-data-science/

Data Science

- Why is there a sudden increased interest Data Science?
  - Burst in Data – Internet, electronic devices
  - Technological advancements – data storage, processing power, cloud based storage and computing
  - Businesses looking to use data to gain competitive advantage
  - “The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”

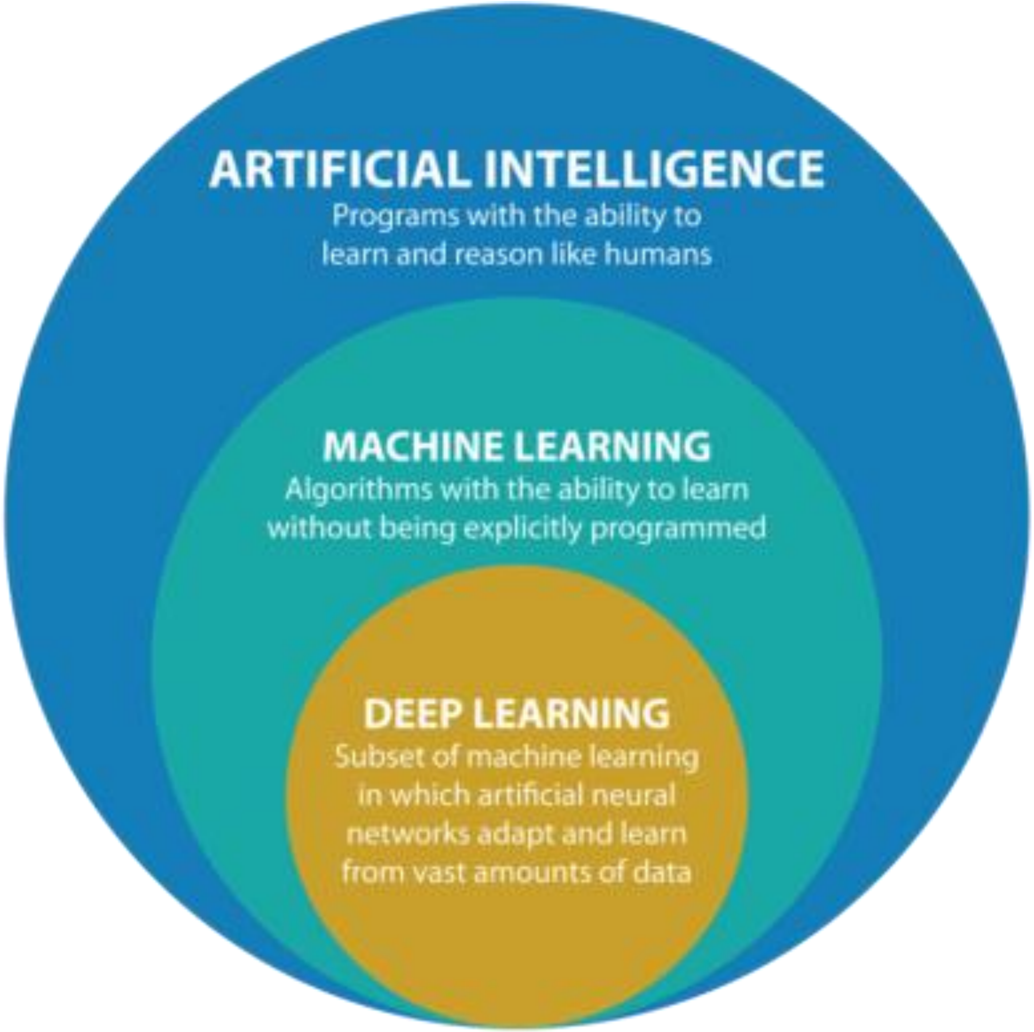
    - Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

Machine Learning

- What is Machine Learning?

- Machine learning is a study of algorithms and statistical models that computer systems use in order to perform a specific task effectively **without being explicitly programmed, relying on patterns** instead

AI, ML, Deep Learning



ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

https://www.quora.com/How-do-artificial-intelligence-and-machine-learning-impact-automation

Machine Learning

- Process of enabling computer to learn to do tasks (for example, prediction) based on well defined statistical and mathematical methods
- The ability to do the prediction is built in form of a “model”.
- A model is the result of the learning process
- The model represents the process which generated the data used to build the model
- The more representative data is of the real world in which the process is executed, the better the model would be

Machine Learning

- How does machine learning work?
    - It searches through data to look for patterns
    - The patterns are expressed as statistical / mathematical structures, for example polynomial equations
    - These statistical / mathematical structures, which can be used to perform predictions, are called models

Use of Machine Learning

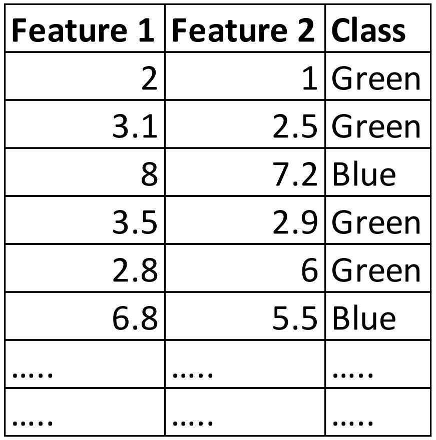
- Machine learning is useful when (few examples)
  - Data patterns are too complex and constantly changing. E.g. weather forecasting
  - We find it hard to express our knowledge about patterns as a program. e.g. Character recognition
  - We do not readily have an algorithm to identify a particular pattern e.g. spam mail detection

Feature Space

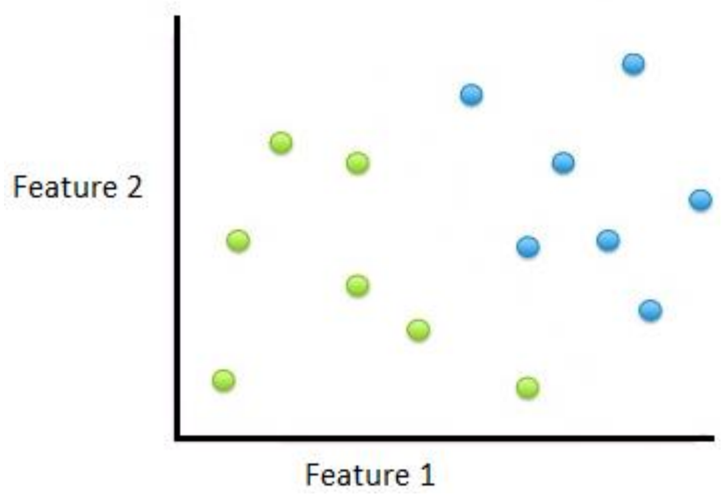
- Each record represents data collected on various attributes

- These values, when plotted, are called feature space or mathematical space

- Following is an example of 2-dimensitonal feature spce

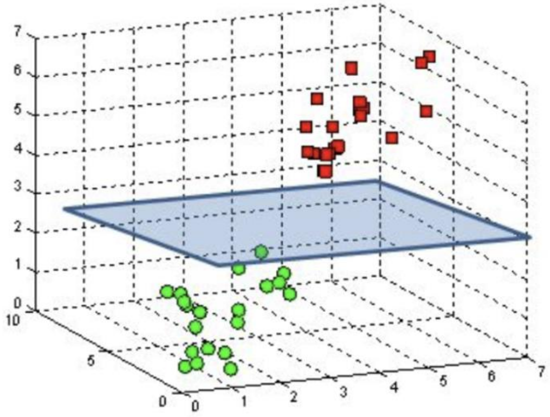


Feature 1	Feature 2	Class
2	1	Green
3.1	2.5	Green
8	7.2	Blue
3.5	2.9	Green
2.8	6	Green
6.8	5.5	Blue
…..	…..	…..
…..	…..	…..



Feature Space

- In a feature space, each attribute becomes a dimension and each record becomes a point in the space
- Feature space can be 3-dimensional or multi-dimensional. In real world, typically there will be multi-dimensional feature space.
- Beyond 3-dimension, we cannot visualize the feature space and depend on statistical and mathematical concepts to derive meaning from it



Terminology

The value which we want to predict:

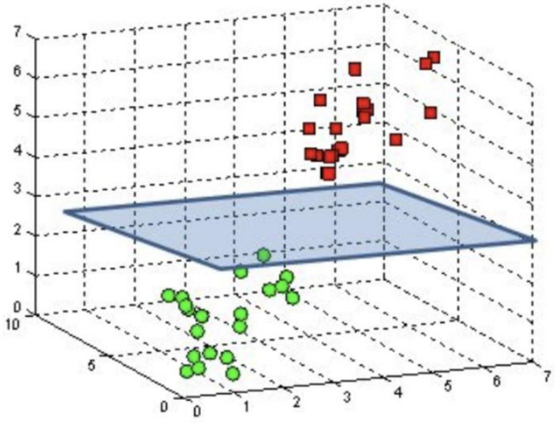
- Target variable, Dependant variable, Y, Predicted variable, Label

The values using which we will attempt to predict:

- Features, Dimension, Independent Variables, Xs, Predictor variable

A Model

- A plane shown in the diagram below is in example of a classification model
- This model attempts to classify data points as Blue or Red (e.g. diabetic or non diabetic)
- The equation of the place can be used to predict the classification of new records
- Thus, if we provide the three dimensions of a point, i.e. values of three attributes, then the model can predict classification of the point
- Proportion of the records that are correctly classified by a model decides accuracy of the model



Machine Learning Categories

- Popular Machine Learning models:

Supervised Learning	
Regression	**Classification**
Linear Regression	K-nearest Neighbors*
Artificial Neural Network	Logistic Regression
	Decision Tree*, Random Forest*
	Naïve Bayes classifier
	Support Vector Machine*
	Artificial Neural Network*

** : Also for Regression*

Unsupervised Learning	
Cluster Analysis	**Dimension Reduction**
K-Means Clustering	Principle Component Analysis
Hierarchical Clustering	

Machine Learning

Dimension Reduction

Dimensions

- Each predictor variable is called a Dimension or a Feature

- Variation of values in each dimension can affect value of the target variable and hence can be useful in predicting target variable

- More the dimensions, possibly, more is the information the dataset

- However, too many dimensions **can** become challenge to machine learning algorithms

- Adding more dimensions, not always result in improvement of performance of models.

Dimension Reduction

- Objective of dimension reduction is to convert data with high number of dimensional into data with fewer dimensions, with **minimal loss** of information.

- Following types of dimension are candidates for dimension reduction:

  - Dimensions with low variance carry little information. Such dimensions (columns) can be considered for dropping.

  - Dimensions with strong correlation are likely to contain similar information. In such case, it may not be necessary to include both the columns in a model.

Principle Component Analysis

- Principle Component Analysis (PCA) is useful when there is a strong correlation between predictor variables (dimensions)

- PCA transforms existing dimensions into new dimensions

- Helps remove information redundancy between dimension

Machine Learning

Examples of Roles

What is needed to Build ML Models

- Good quality data that is representative of the real-world process is the key starting point. Without data, we cannot build machine learning models

- Domain knowledge – without domain knowledge, it is not possible to understand data, check data quality etc. which is essential while building a model

- Understanding of basic mathematics, statistics and machine learning algorithms

- Technical programming skills

Roles

- Data Scientist:
  - Understanding business challenges
  - Thorough understanding of machine learning algorithms
  - Predictive analytics. Define models to be used.
  - Create valuable actionable insights using data.
  - Effectively communicate findings to the business.
  - Ability to understand Big picture, in-depth knowledge of Statistics techniques and technical competency to work with data.

- Machine Learning Engineer:
  - Design and develop machine learning algorithms
  - Run machine learning tests and experiments
  - Optimize models
  - Implement appropriate ML algorithms

Other Related Roles

- Data engineer / Big Data engineer, Data Architect
- Business Analyst
- Visualization expert

Machine Learning

Python for Machine Learning

Machine Learning Languages

- Python and R are suited for data science functions.

- Go is emerging as an alternative but is not yet as well supported as Python.

- In practice, data science teams use a combination of languages to play to the strengths of each one, with Python and R used in varying degrees

- As of now, Python stands out as the preferred language for machine learning framework

Python

We will mainly use following libraries:

- NumPy - for Array operations, and basic mathematical and statistical functions etc
- SciPy - It builds on NumPy. Add a a collection of algorithms and functions for probability distributions, computing integrals numerically, solving differential equations, optimization etc
- Pandas - for Data-frame operations, reading excel etc
- Matplotlib and Seaborn – for various plots such as histogram, boxplot, scatterplot etc
- Scikit-learn - For machine learning algorithms, including unsupervised learning, regression and classification. For measuring performance of models, performing data split etc

Jupyter Notebook

- The Jupyter Notebook is an open-source application that makes it convenient to learn concepts using Python in interactive interpreter mode
- It is a preferred environment for learning new concepts using Python

Machine Learning

Supervised Machine Learning

Supervised Machine Learning

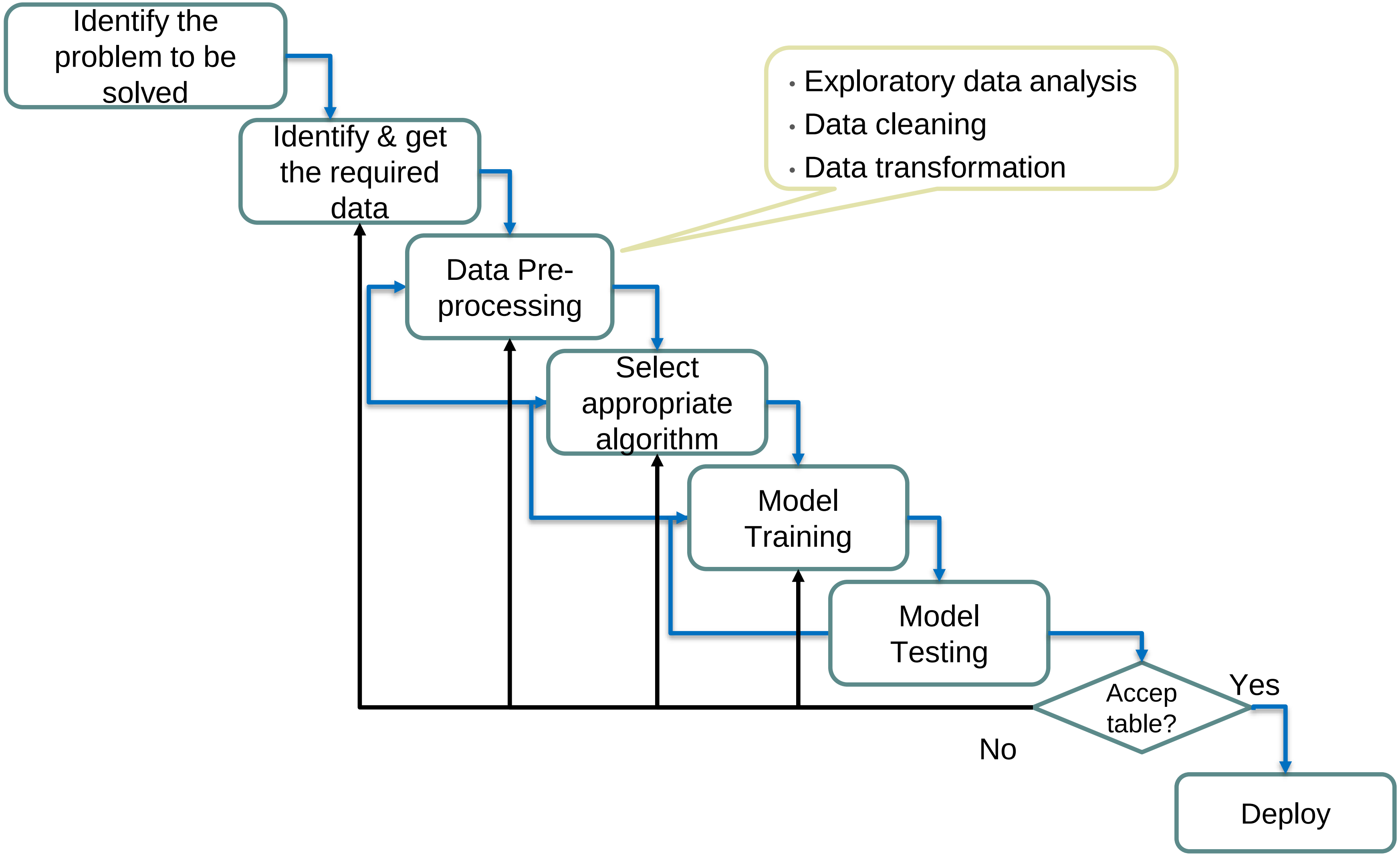
- Supervised Machine Learning is a class of machine learning algorithms where a target variable is to be predicted based on values of predictor variables

- For a given business problem, data needed to perform prediction is identified

- Model is trained using data that contains predictor and target values (training data)

- The model is tested for using test data where only predictor variables are supplied to the model. Predicted values are compared with actual values to evaluate performance of the model

Supervised Machine Learning


Identify the problem to be solved
Identify & get the required data
Data Pre-processing
Select appropriate algorithm
Model Training
Model Testing
Accep table?
Yes
No
Deploy

- Exploratory data analysis
- Data cleaning
- Data transformation