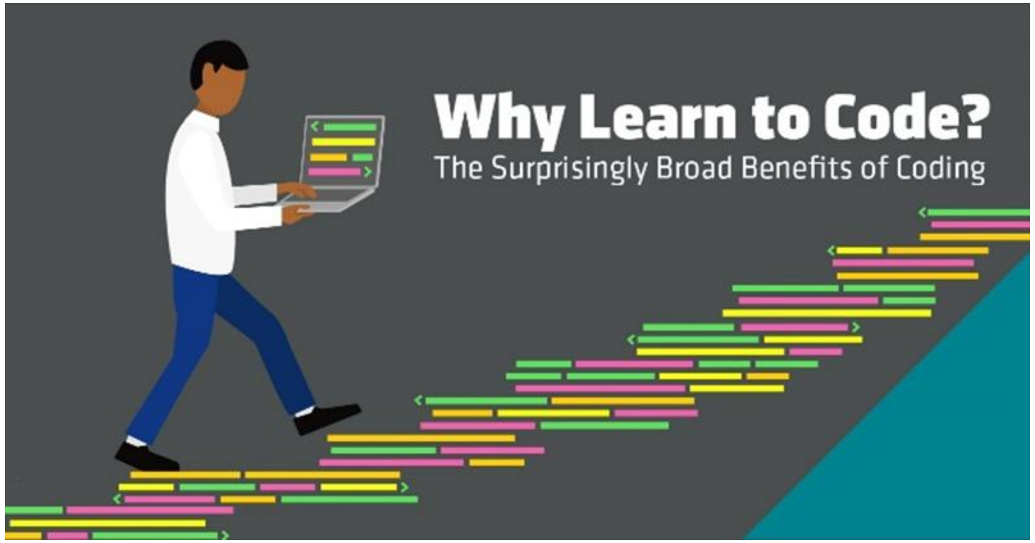
Machine Learning for Geoscientists

Amr. Moslim

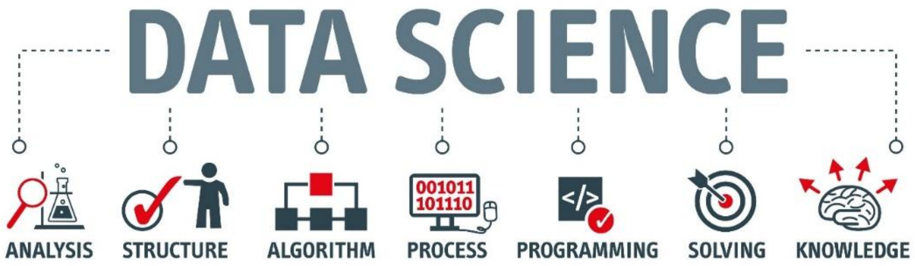


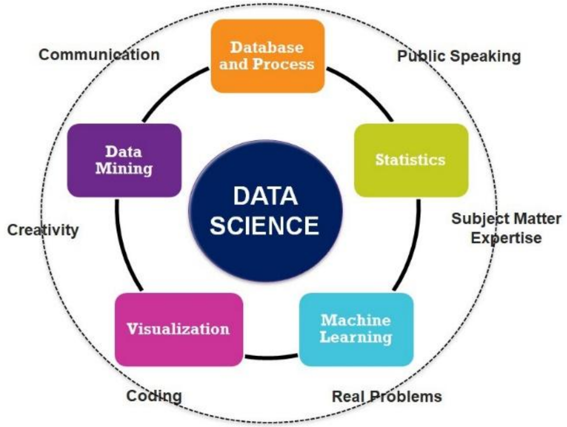
Today’s Agenda

- What is Data science?
- What is Machine learning?
- DS VS ML
- DS Skills
- Why should I learn Python?
- How to Code?
- Machine Learning technique
- How does ML work?
- ML Models Evaluation
- ML Algorithms (Kmeans – Random Forest)

What’s Data Science ?

Data science is the field of study that uses modern tools and techniques to process, clean, analyze, model and visualize large data sets to get insights that are reliable to help organizations to understand certain criteria or condition and make business decisions

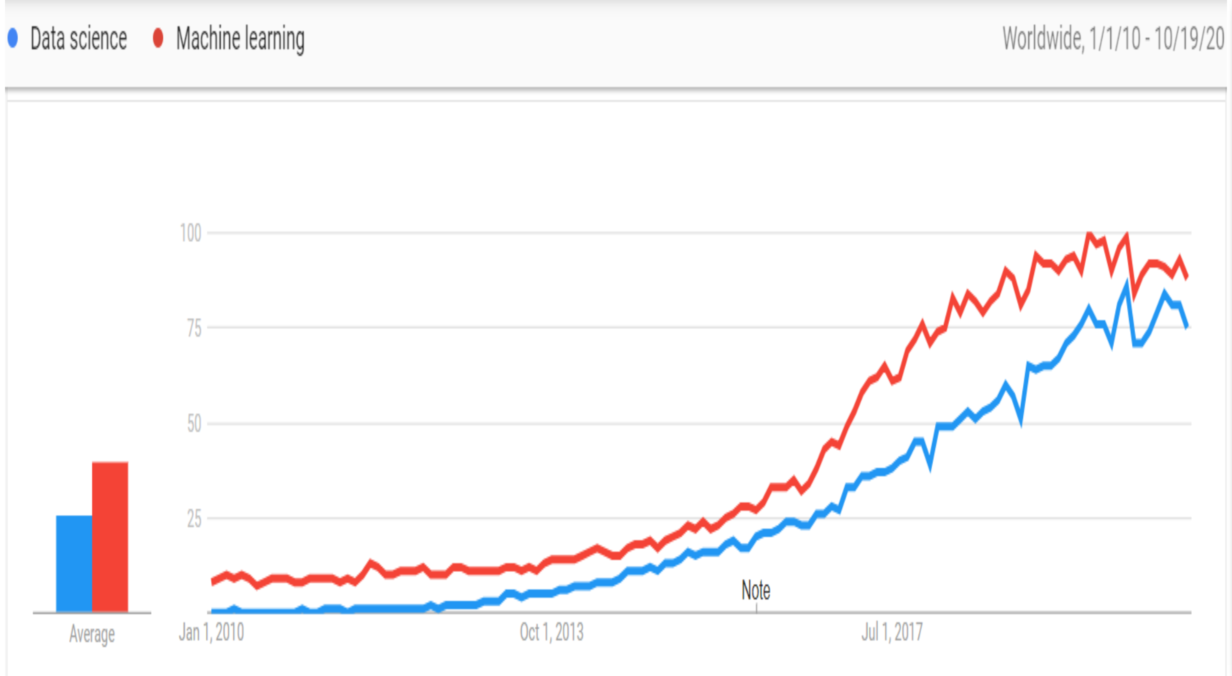


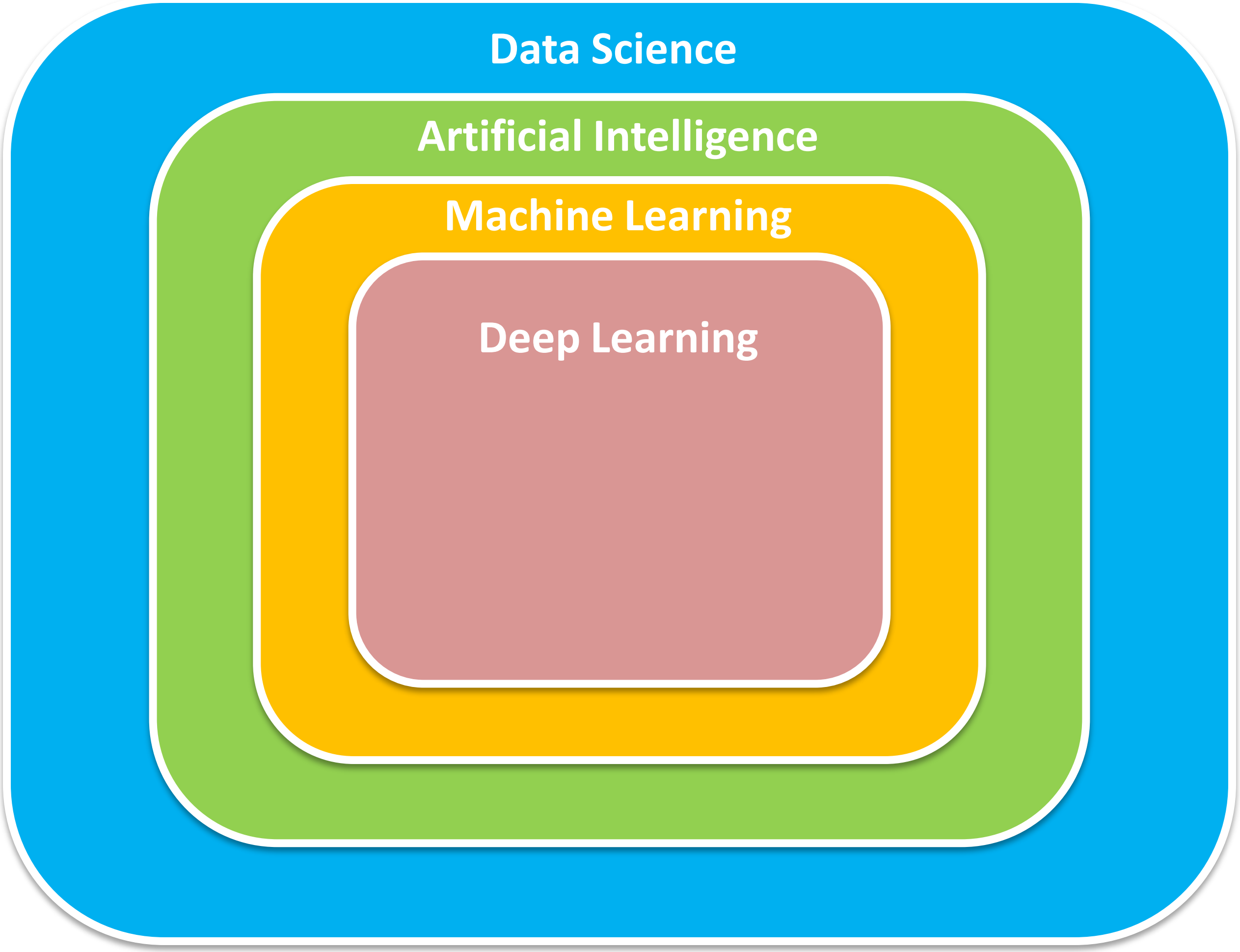


What’s Machine Learning?

Machine learning is the field of AI that allows systems to learn from past data and make intelligent decisions on their own using algorithms without explicitly programed and improve its experience

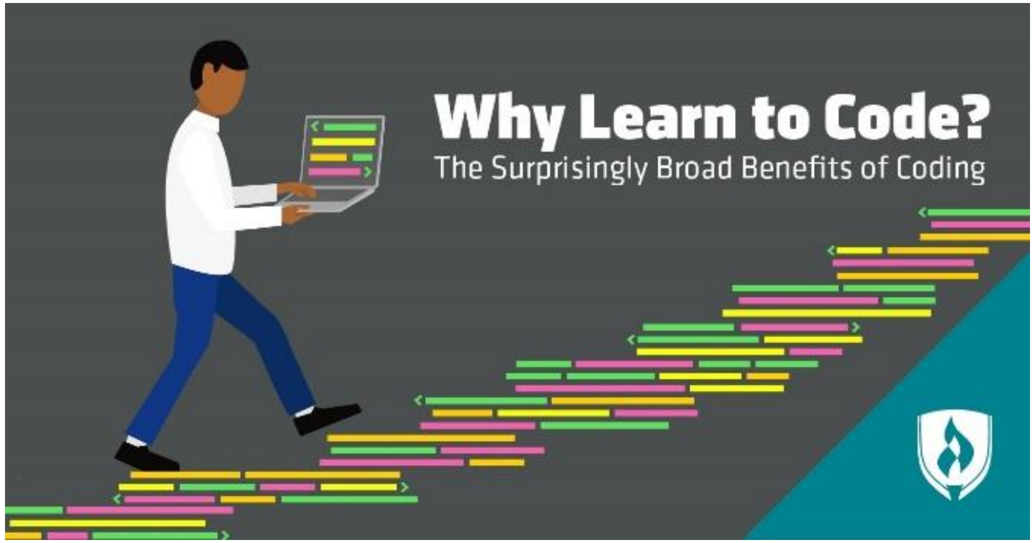
Data Science vs Machine Learning?





Why Learn to Code?

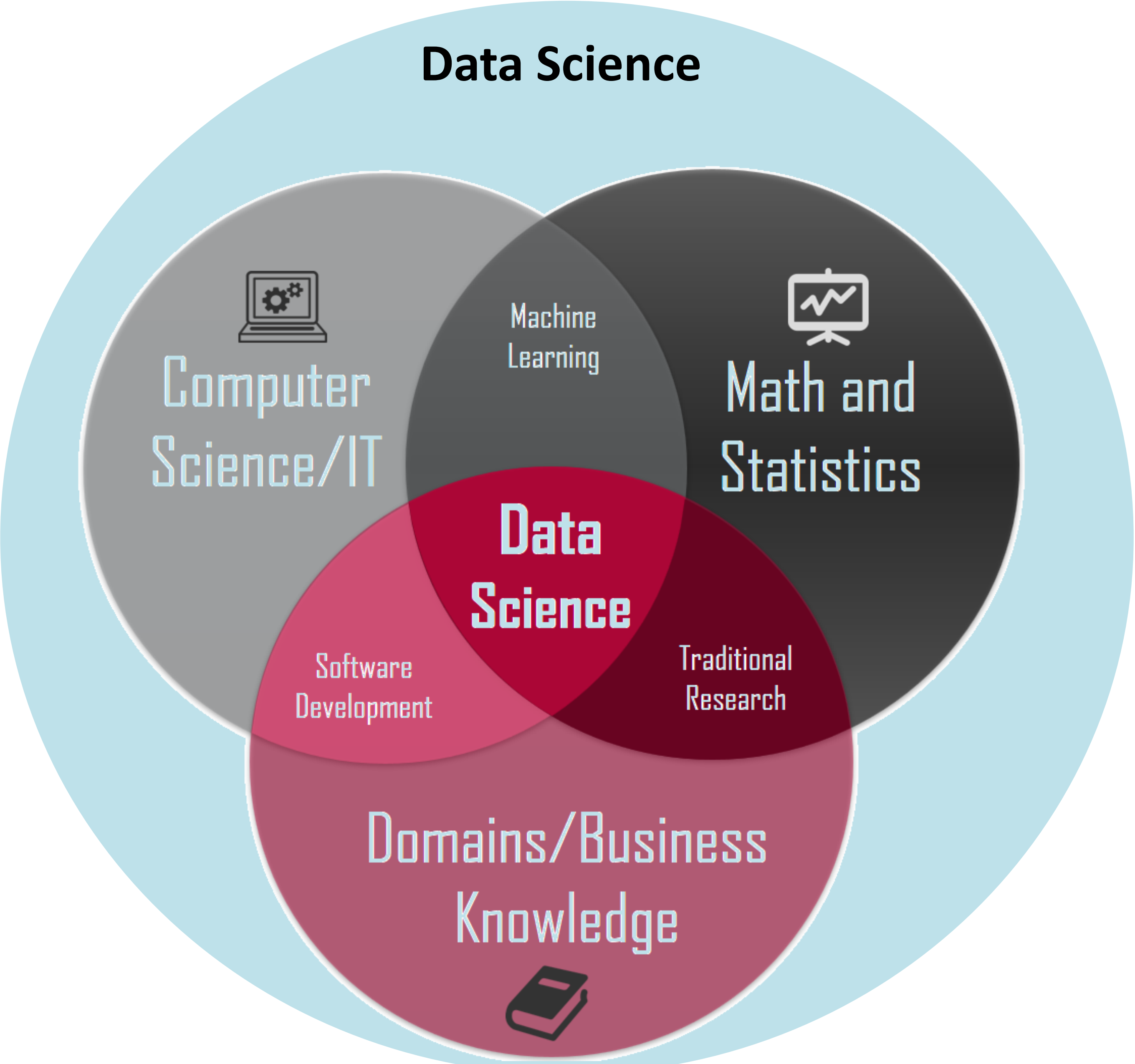
1. Coding is another language.
2. Coding fosters creativity.
3. Coding helps learn Math skills and makes sense of it.
4. Coding improves writing academic performance.
5. Coding can lead to software development jobs
6. It open up other job opportunities
7. Coding can make your job application stand out
8. Coding literacy can help you understand other aspects of tech
9. It could lead to freelance work
10. Coding can allow you to pursue passion projects
11. Coding can boost problem solving and logic skills
12. Coding improves interpersonal skills
13. Being a skilled coder can build confidence
14. Freedom to Make My Own projects
15. People Come to ME Asking if I Can Work for THEM
16. You can do work remotely any where.
17. I Am Part of a Top Secret Club (a.k.a., the Tech Community)
18. I Have a Sense of Self-Reliance and Empowerment



It’s a great Empowering tools and Skills

Amr.Moslim

Data Science knowledge domains



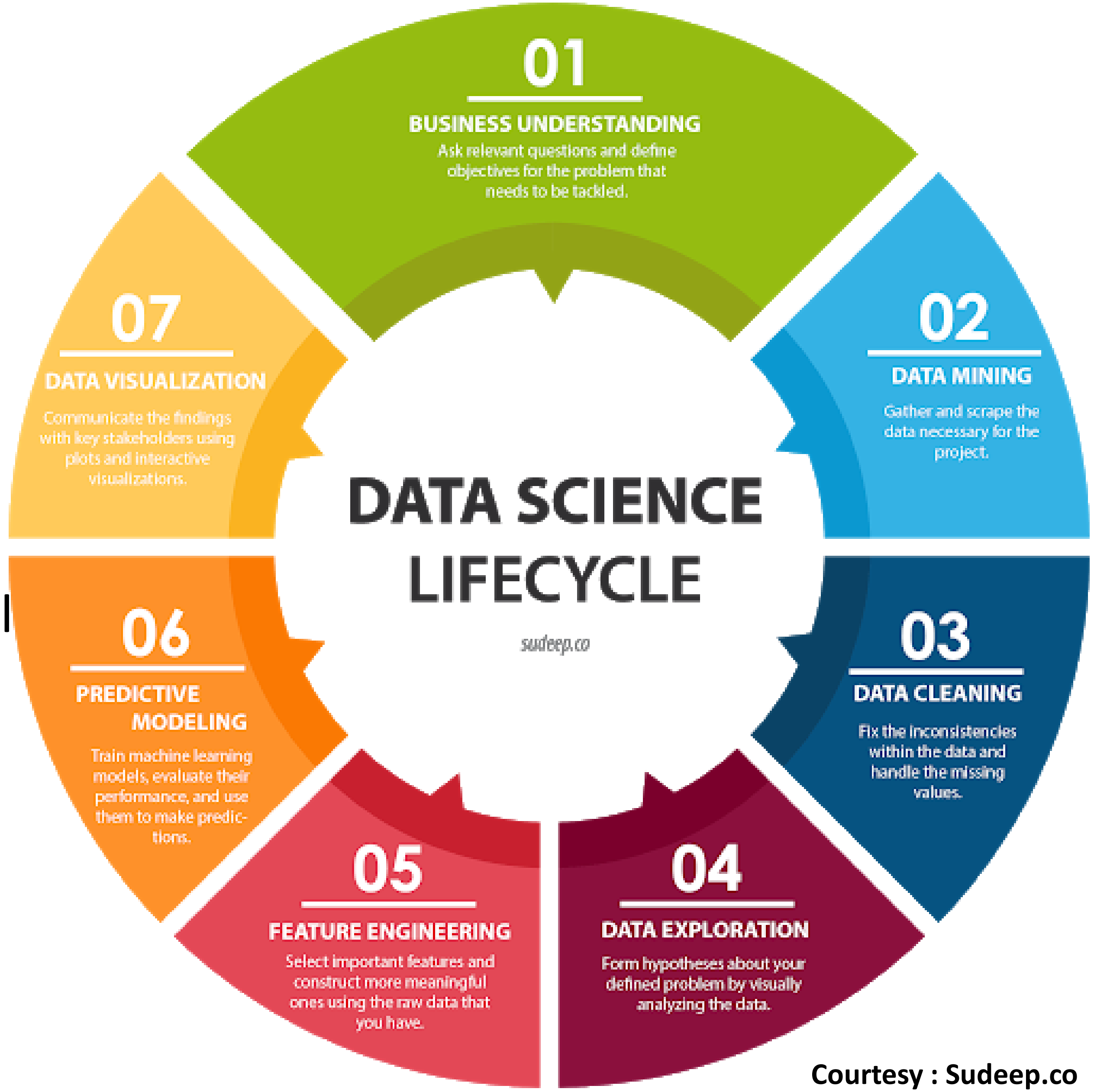
Amr.Moslim

Data Science Life Cycle

80% of the Data scientist time is dedicated to

- Data collection
- Data cleaning
- Data exploration
- Feature engineering

20 % of the data scientist time is for model selection and building



DATA SCIENCE LIFECYCLE

sudeep.co

01 BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING
Gather and scrape the data necessary for the project.

03 DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.

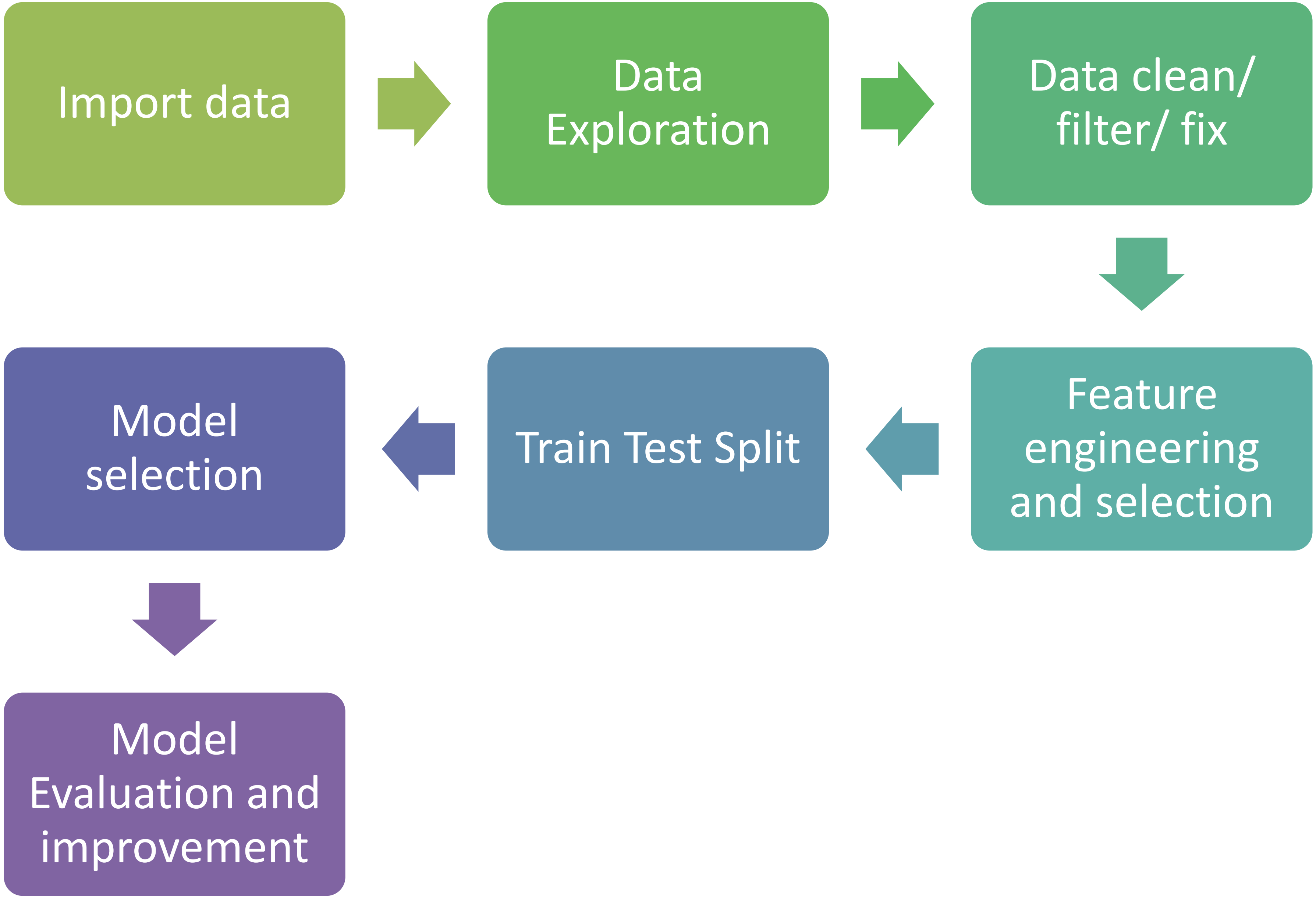
05 FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Courtesy : Sudeep.co

Machine Learning Work flow



Import data

Data Exploration

Data clean/ filter/ fix

Feature engineering and selection

Train Test Split

Model selection

Model Evaluation and improvement

Data Science Vs Machine Learning

Data Science Vs Machine Learning

Characteristics	Data Science	Machine Learning
Objective	Focus on find unforeseen and hidden trends to understand the data pattern	Focuses on making predictions and classifications to get new data points
Tools	**Python**, **R**, SAS, Spark, Excel, MATLAB, MySQL, Tableau	**Python**, **R**, Scikit Learn, ML Studio, MS Azure
Applications in O& G	• Time series analysis • Production forecast • Oil price prediction	• S-wave log predication • Facies classification • Porosity logs prediction using seismic attributes
Skills	• Database and SQL • Mathematics and statistics • Knowledge of programming • Data mining, data wrangling • Data visualization • Machine Learning	• Programming (Python , R) • Mathematics and statistics • Machine Learning algorithms • Data Modeling • NLP

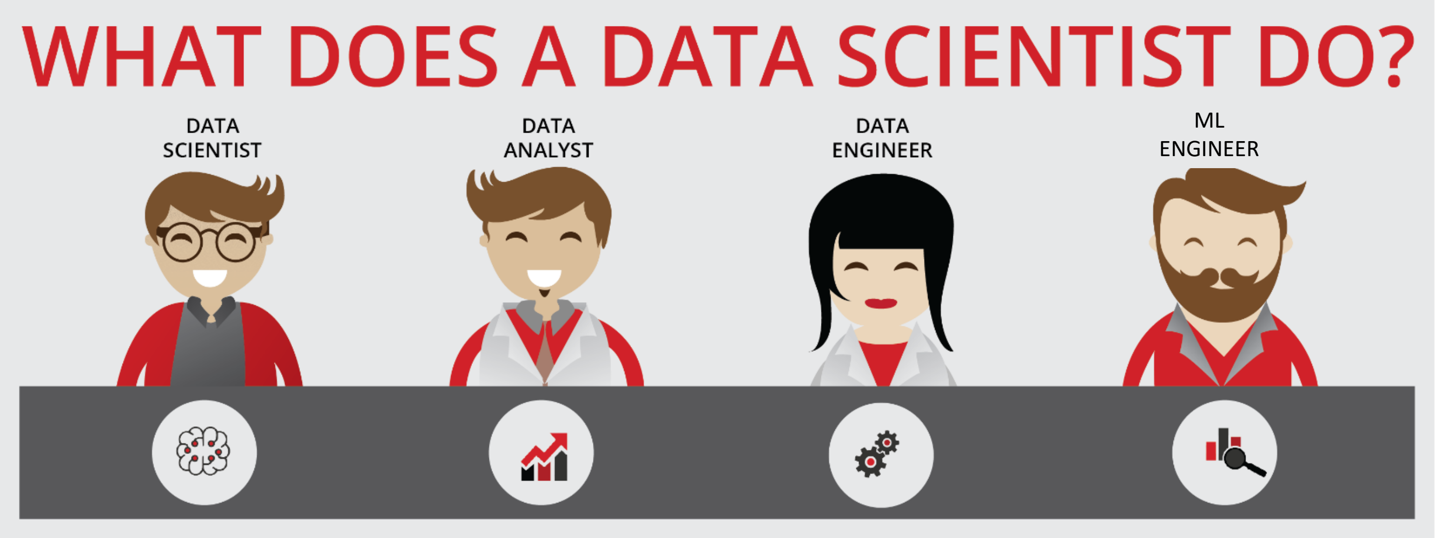
Amr.Moslim

ML vs DS

Machine Learning	Data Science
Data structured - unstructured	Any type of data
No specified rules for each problem	Has specified approach and workflow for each problem
Generate generalized models for each problem type	Generate specific insights for each problem
Understanding algorithms and maths is crucial.	Domain expertise is the king
Classifies / predicts for new data points / patterns from historical data	Create insights from world complexities
Input data should be transformed specifically for the algorithm	Input data can be used directly which is to be read and analyzed

Data science Skills

Data science Jobs



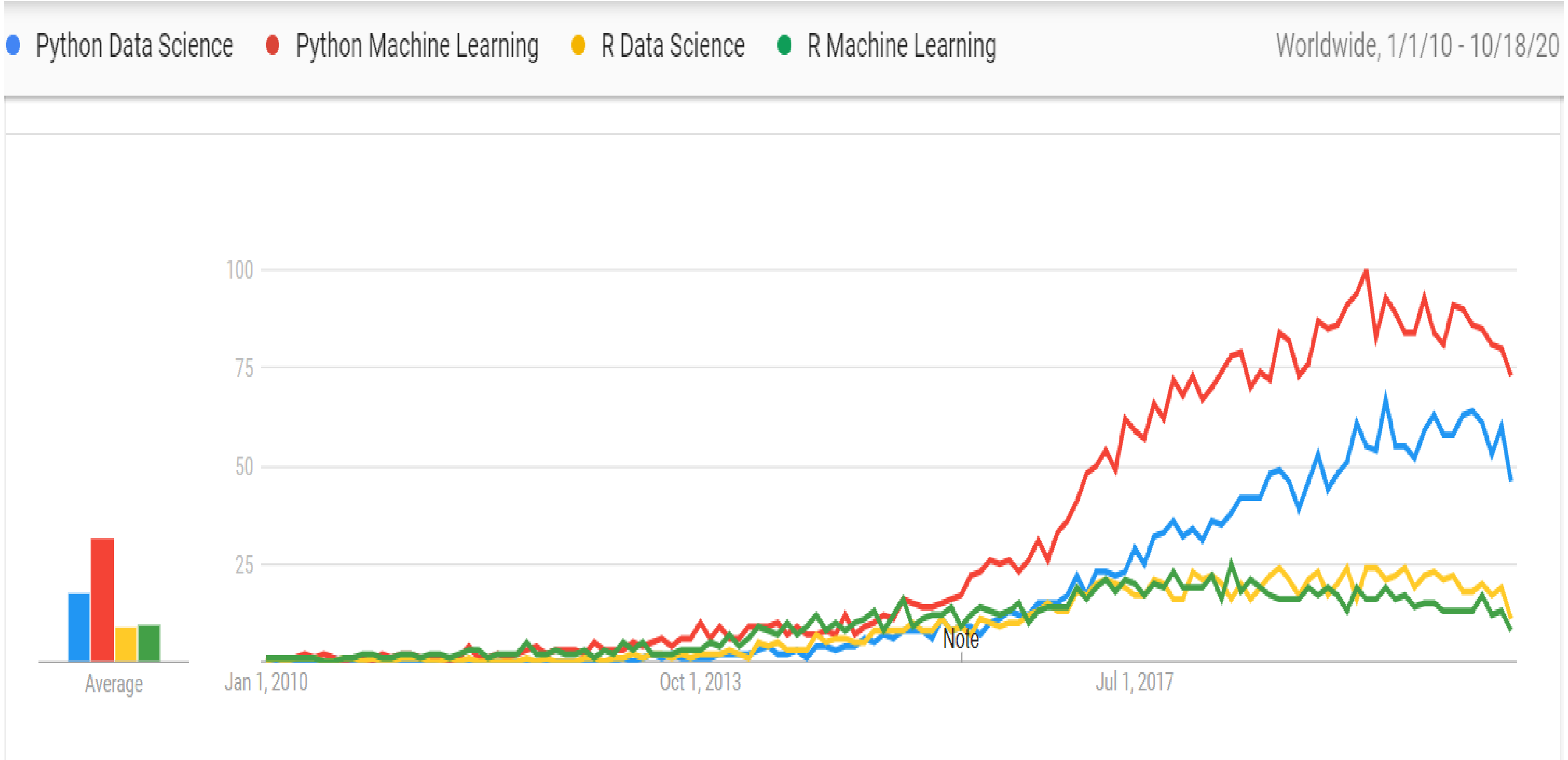
- DATA ANALYST
- DATA ENGINEER
- MACHINE LEARNING ENGINEER
- DATA SCIENCE GENERALIST

Data Scientist Skills/Experience

- **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/ Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods or Business Process Management
  - Application/subject domain related (research or business)
  - Mathematics and Statistics
- **Group 2: Big Data (Data Science) tools and platforms**
  - Big Data Analytics platforms
  - Mathematics & Statistics applications & tools
  - Databases (SQL and NoSQL)
  - Data Management and Curation platform
  - Data and applications visualization
  - *Cloud based platforms and tools*
- **Group 3: Programming and programming languages and IDE**
  - General and specialized development platforms for data analysis and statistics
- **Group 4: Soft skills or Social Intelligence**
  - Personal, inter-personal communication, team work, professional network

Why should I learn Python?

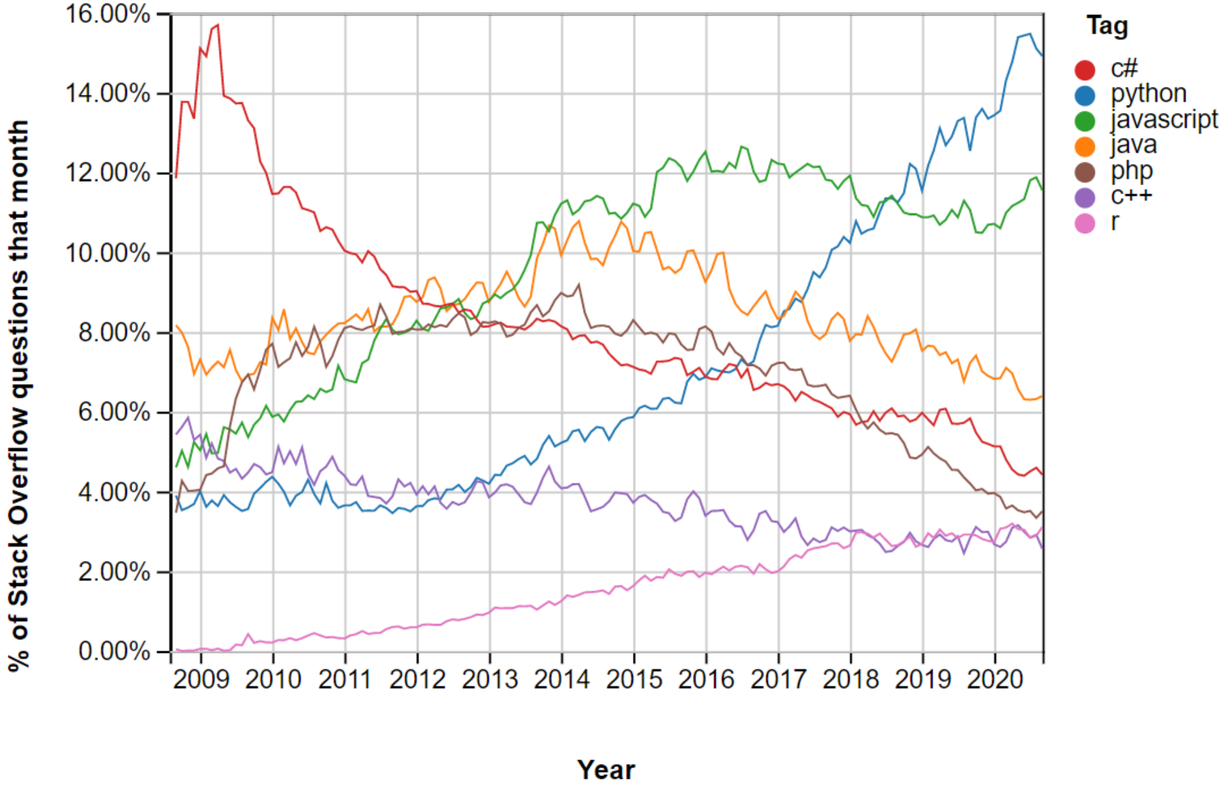
Why I should Learn Python ?



Amr.Moslim

Why I should Learn Python ?

1. Python is the fastest growing programming language
2. Python is easy to read, write, and learn
3. Python has an incredibly supportive community
4. Open source package (free)
5. Multi purpose programming language
6. Big companies uses python in their main frame work
7. High in demand in the market of data science
8. Hundreds of applications & libraries
9. Python developers make great money
10. Great tool for reproducibility
11. Collaborative language to build complex tasks

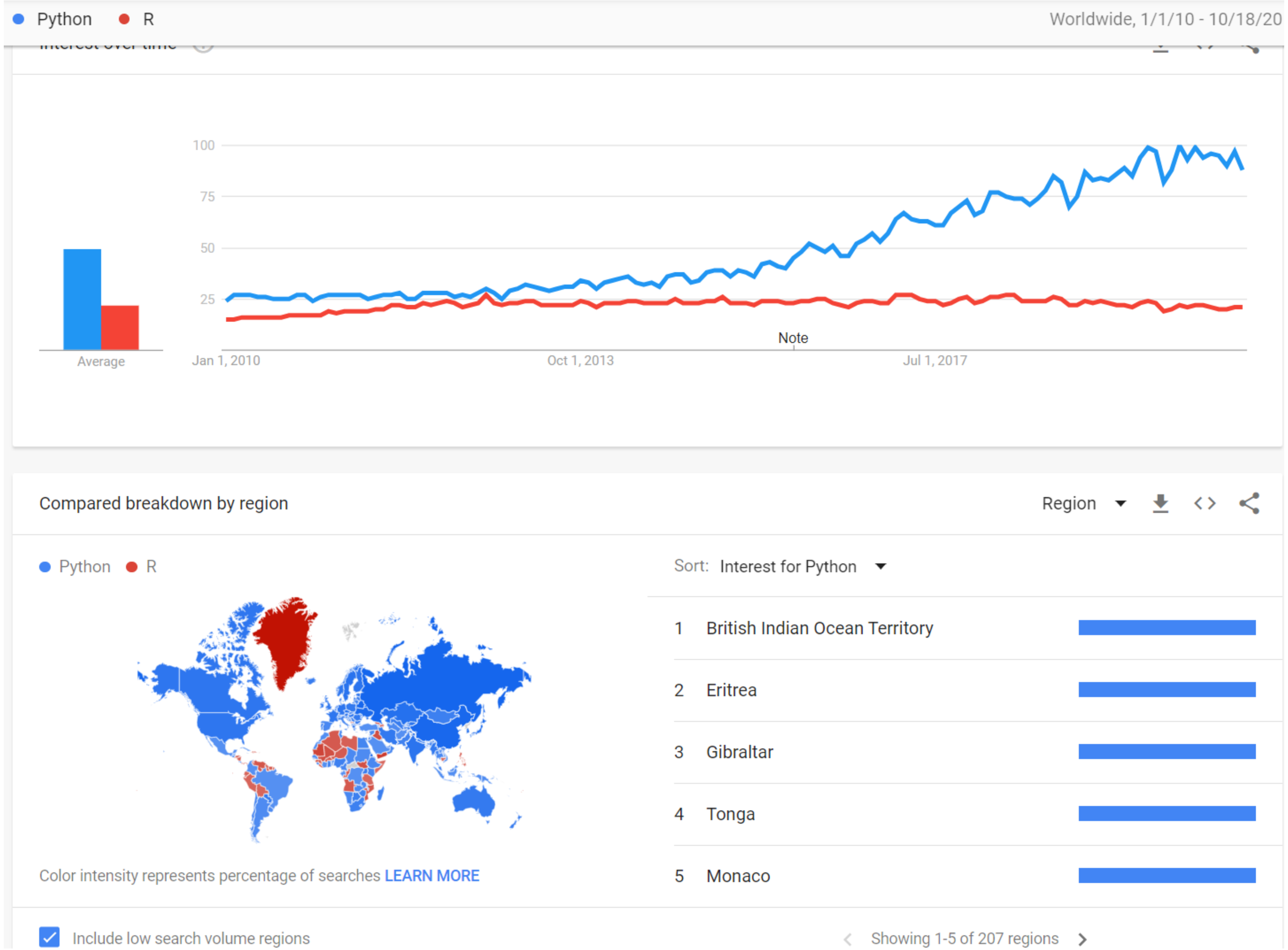


MAJOR COMPANIES THAT USE

python

Amr.Moslim

Why I should Learn Python ?



● Python ● R   Worldwide, 1/1/10 - 10/18/20

Compared breakdown by region   Region

● Python ● R

Sort: Interest for Python

1	British Indian Ocean Territory
2	Eritrea
3	Gibraltar
4	Tonga
5	Monaco

Color intensity represents percentage of searches LEARN MORE

☑ Include low search volume regions

Showing 1-5 of 207 regions

How to code?

Coding Workflow Basic Aspects

- **Assignment:**
  - Types of data structure (integer, float, String, Boolean)
- **Control flow:**
  - If statement
  - While loops
  - For loops
- **Mathematical Operators:**
  - (+, -, *, /)
  - (>, <, =, >=, <=, !=)
  - Logical operators:
    - (+=, -=, //, %, %%)
- **Functions:**

  A set of commands that works in sequence to perform a certain task that can include assignment, flow control tools and or mathematical expressions.
  - def: in Python
  - Function (x) in R
- **Error handling:**
  - Avoid having user errors
  - Handling errors
- **Reviewing:**
  - Debugging : to check that all the results as it should be even if you didn't get any errors explicitly

Python most popular packages

- **Analysis packages**
  - Numpy : Numerical Manipulation and linear alegabra
  - Pandas : building & Manipulating DataFrames

- **Visualization packages**
  - Matplotlib : plots and contours
  - Seaborn : beautiful plots
  - Plotly : interactive plotting

- **Machine Learning packages**
  - Tensorflow : Neural NetWork and Deep learning
  - Keras: ML algorithms
  - Scikit Learn: ML algorithms and model evaluations

- **Scientific packages**
  - Scipy : scientific equations in python
  - Obspy : seismic manipulation and reading segy

- **Geoscience Package**
  - Welly : reading / write well logs las files
  - Lasio : reading / write well logs las files
  - Segyio : seismic Segy files reading / writing and manupliation.
  - Petopy : Petrophysical evaluation

Machine Learning technique

Machine Learning Algorithm Classification

Supervised Learning

Labeled data prediction

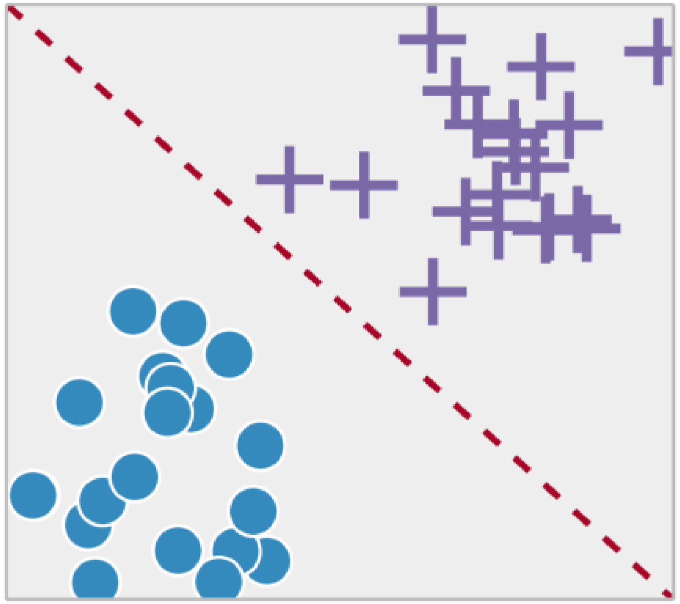
- Regression
- Classification

Unsupervised Learning

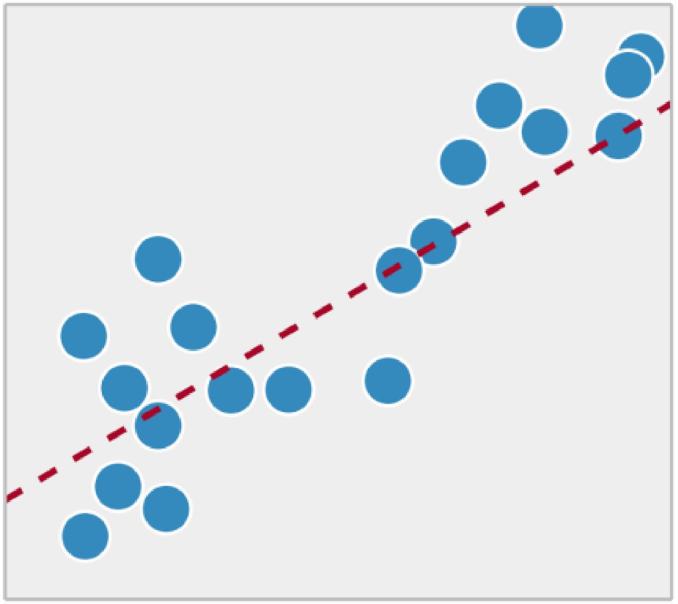
unlabeled data

- Dimensionality reduction
- Clustering

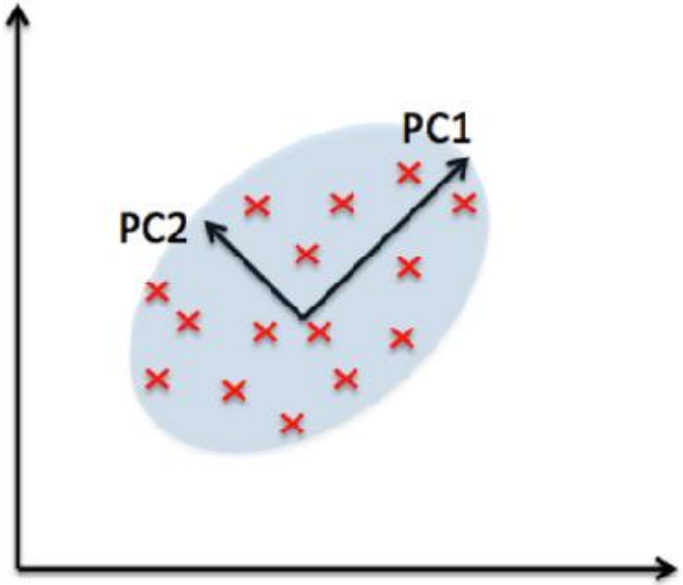
Classification



Regression



Dimensionality reduction



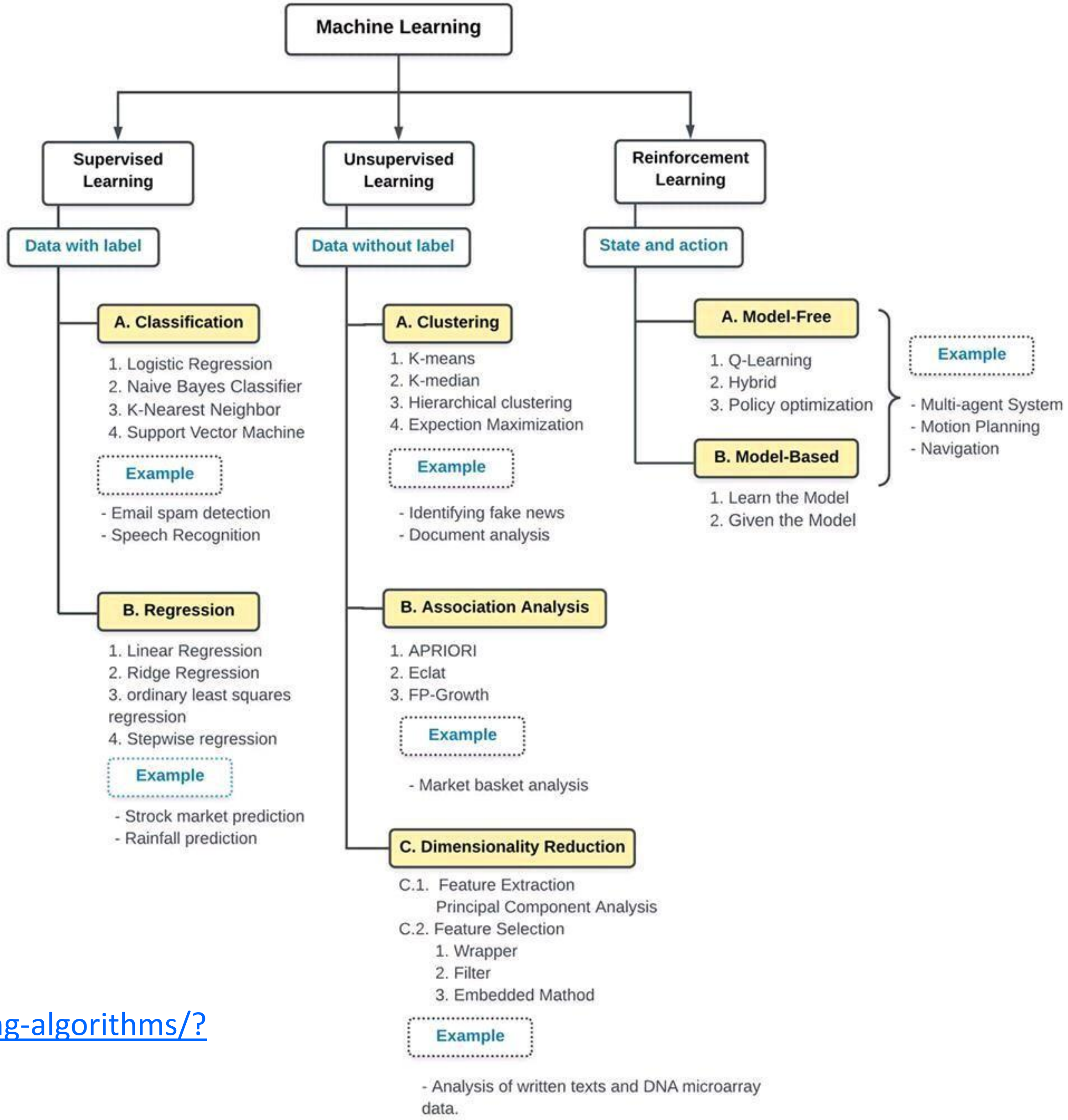
Machine Learning Algorithms

<u>Most commonly used Machine learning algorithms:</u>

1.Linear Regression
2.Logistic Regression
3.Decision Tree
4.SVM
5.Naive Bayes
6.kNN
7.K-Means
8.Random Forest
9.Dimensionality Reduction Algorithms PCA
10.Gradient Boosting algorithms

    1. GBM
    2. XGBoost
    3. LightGBM
    4. CatBoost

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?



Machine Learning

Supervised Learning

Data with label

A. Classification

1. Logistic Regression
2. Naive Bayes Classifier
3. K-Nearest Neighbor
4. Support Vector Machine

Example

- Email spam detection
- Speech Recognition

B. Regression

1. Linear Regression
2. Ridge Regression
3. ordinary least squares regression
4. Stepwise regression

Example

- Strock market prediction
- Rainfall prediction

Unsupervised Learning

Data without label

A. Clustering

1. K-means
2. K-median
3. Hierarchical clustering
4. Expection Maximization

Example

- Identifying fake news
- Document analysis

B. Association Analysis

1. APRIORI
2. Eclat
3. FP-Growth

Example

- Market basket analysis

C. Dimensionality Reduction

C.1. Feature Extraction
Principal Component Analysis
C.2. Feature Selection
1. Wrapper
2. Filter
3. Embedded Mathod

Example

- Analysis of written texts and DNA microarray data.

Reinforcement Learning

State and action

A. Model-Free

1. Q-Learning
2. Hybrid
3. Policy optimization

B. Model-Based

1. Learn the Model
2. Given the Model

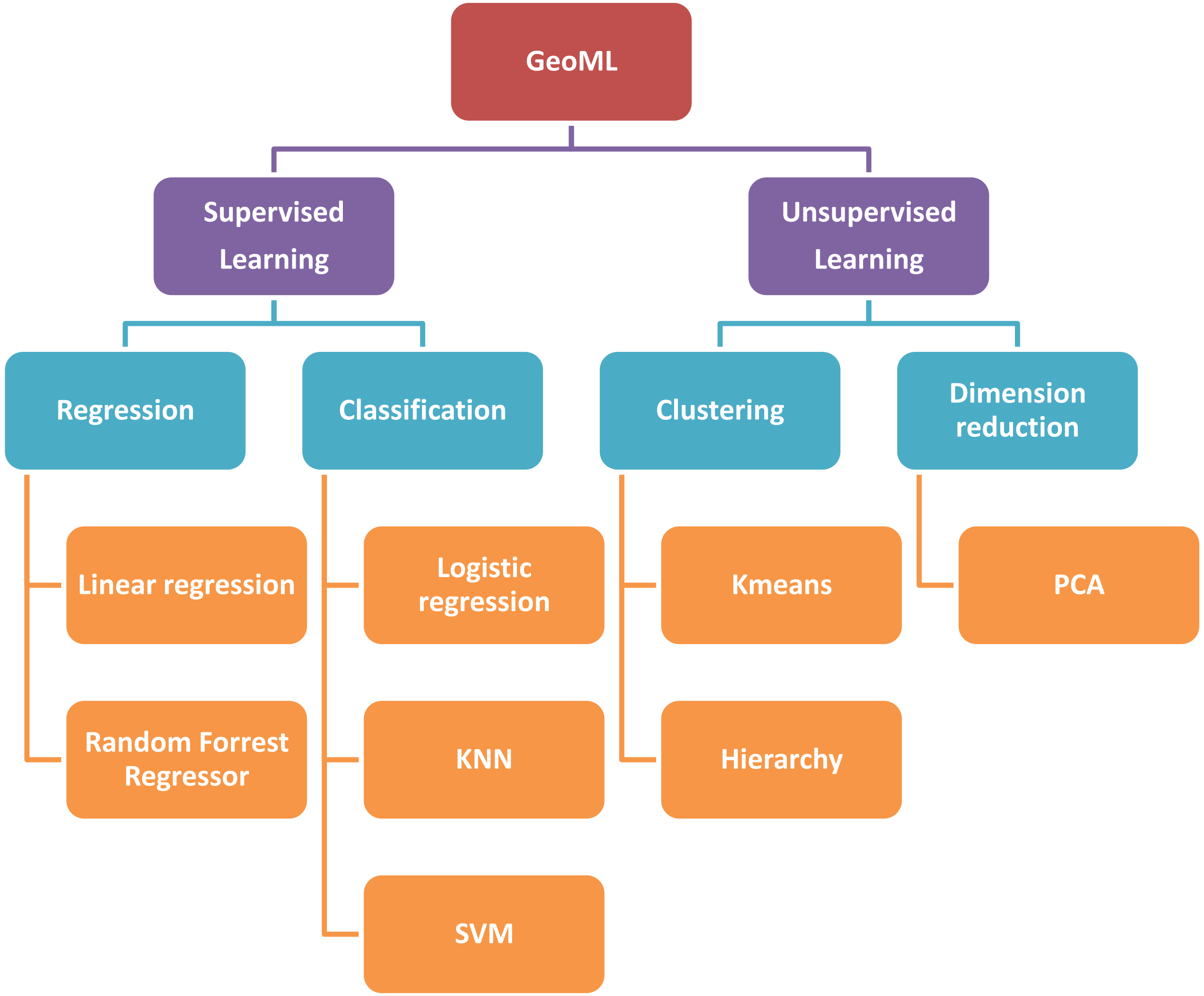
Example

- Multi-agent System
- Motion Planning
- Navigation

Machine Learning Algorithms

Most commonly used Machine learning algorithms:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms (PCA)
10. Gradient Boosting algorithms
    1. GBM
    2. XGBoost
    3. LightGBM
    4. CatBoost



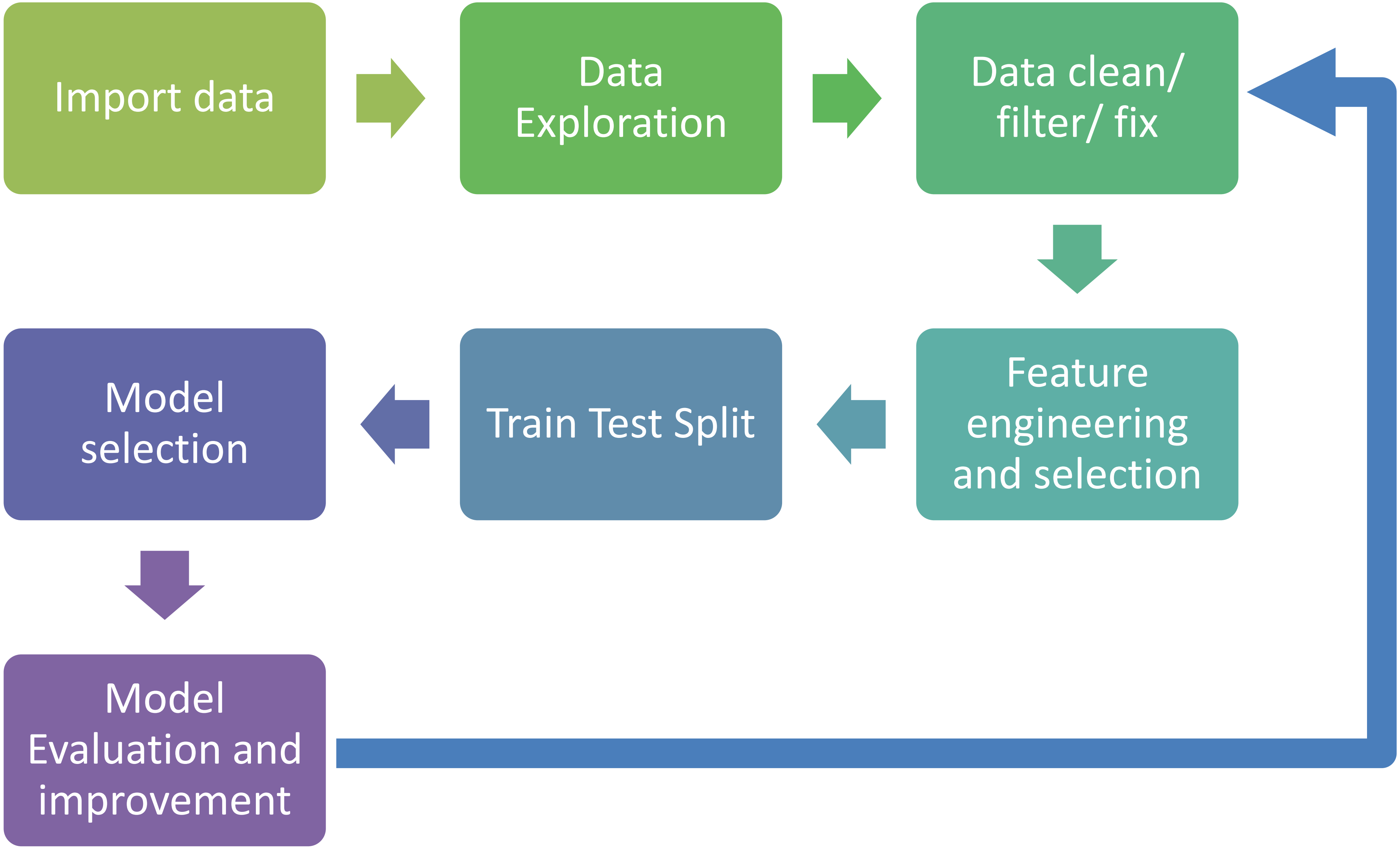
Machine Learning vs Coding

Characteristics	Machine Learning Algorithms	Common coding
Objective	To teach the machine to create models to solve the problem without hard coding using data patterns	To use programming language to explicitly code the solution to the problem

Example: v = d/t	Data = (mass, height, width, velocity)
	Lm = linearregression()
	Lm.fit()
	Lm.predict()
		Data = (d, t)
		def velocity(d,t):
		v = d/t
		return (v)

Tools	Python, R, Scikit learn, Tensorflow, etc…	Python, R, Visual Basic, Java, Go, Excel
Running time	Most of time in data wrangling and model evaluation	Most of the time in coding the problem and solution
Output	ML model and forecast	Data table, graphs, dashboards
Reproducibility	Yes with the same data formats	Yes with the same data formats
Domain knowledge	It is very important and highly recommended	a must

Machine Learning Work flow



Amr.Moslim

How does ML work?

Linear Regression

- ***Objective:***

  model the expected value of a continuous variable, ***Y***, as a linear function of the continuous predictor, ***X***

- ***Model structure:***

  $Y = Ax + B$

- ***Model assumptions:***

  Y is normally distributed, errors are normally distributed, and independent

- ***Parameter estimates and interpretation:***

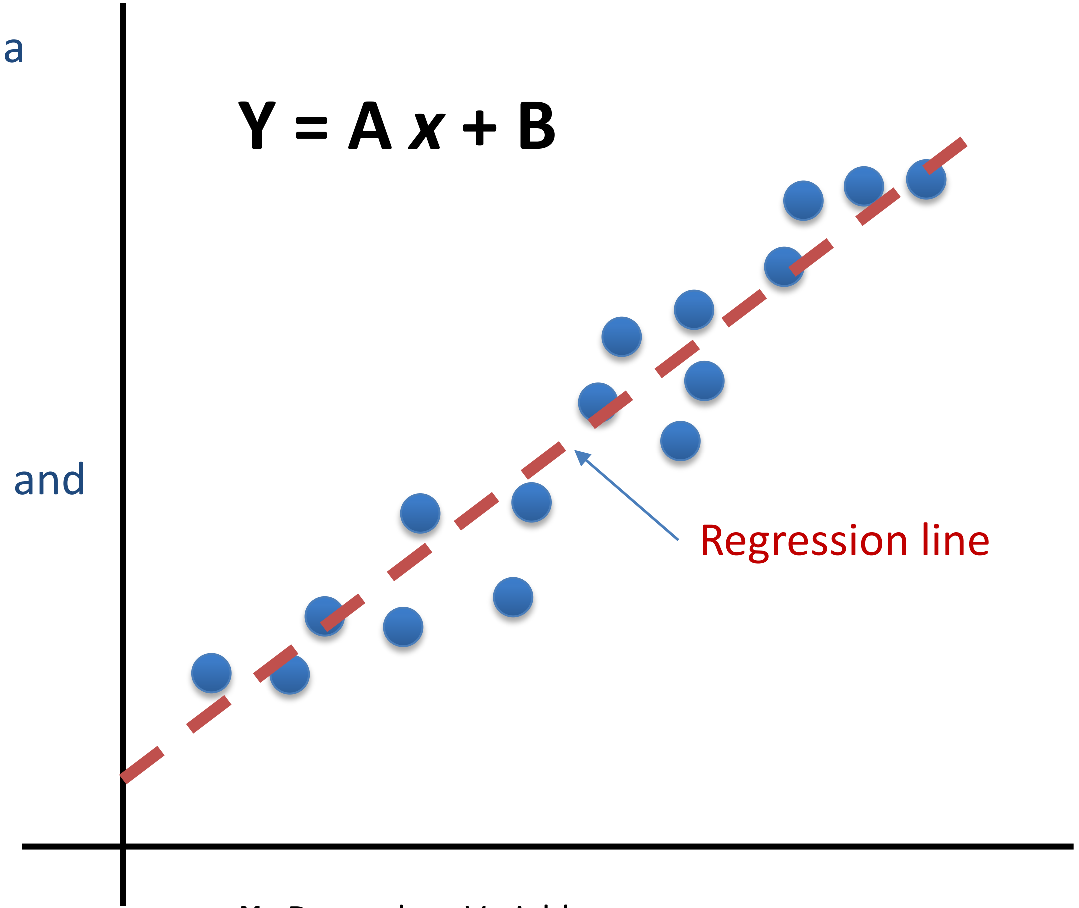
  B the intercept, and **A** is estimate of the slope

- ***Model fit:***

  R^2, residual analysis

- ***Model selection:***

  possible predictors, which variables to include?



Y : Dependent Variable
A : Slope
x **:** Independent variable
B : Intercept

Amr.Moslim

Linear Regression - Gradient Descent

- ***Objective:***

  To minimize the error function to close to zero (Cost Function) If possible.

- ***Function structure:***

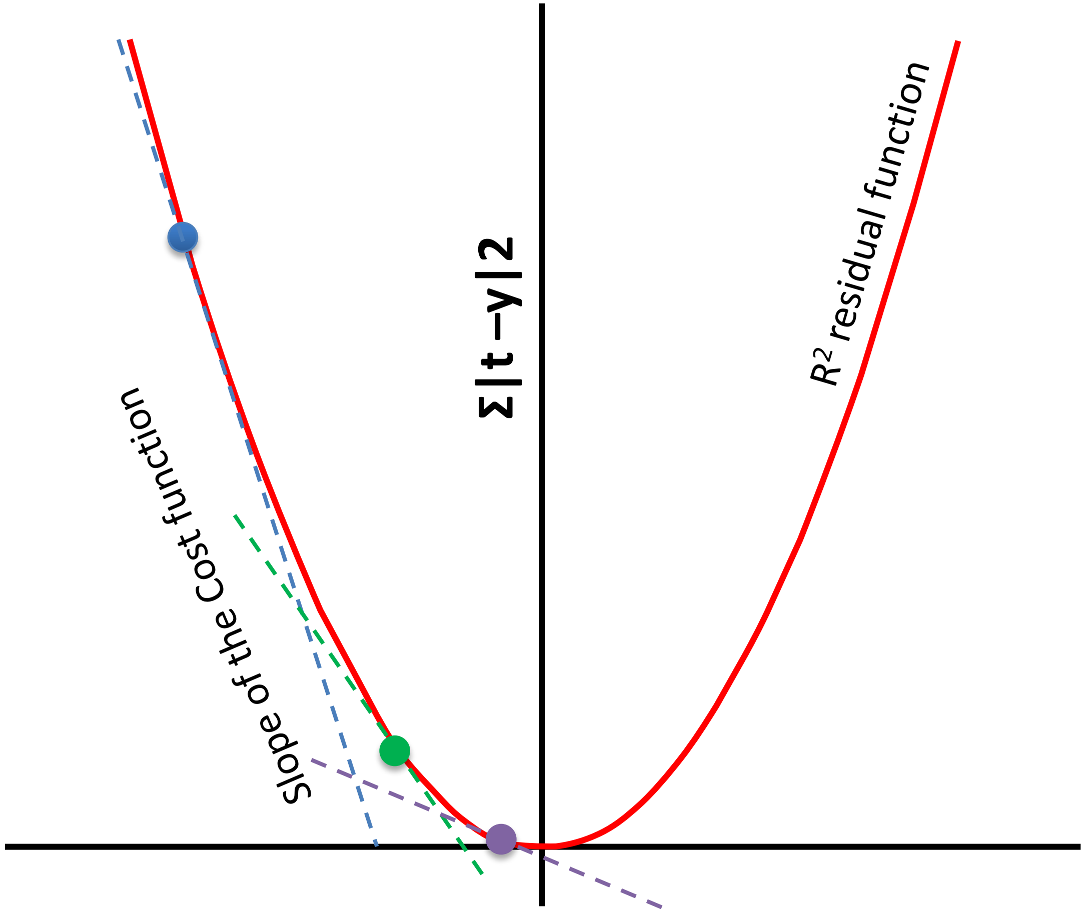
  Cost function : $\Sigma|t - y|^2$

- ***Model assumptions:***

  Slope of the *cost function* ~= Zero, then it is the best prediction

- ***Parameter estimates and interpretation:***

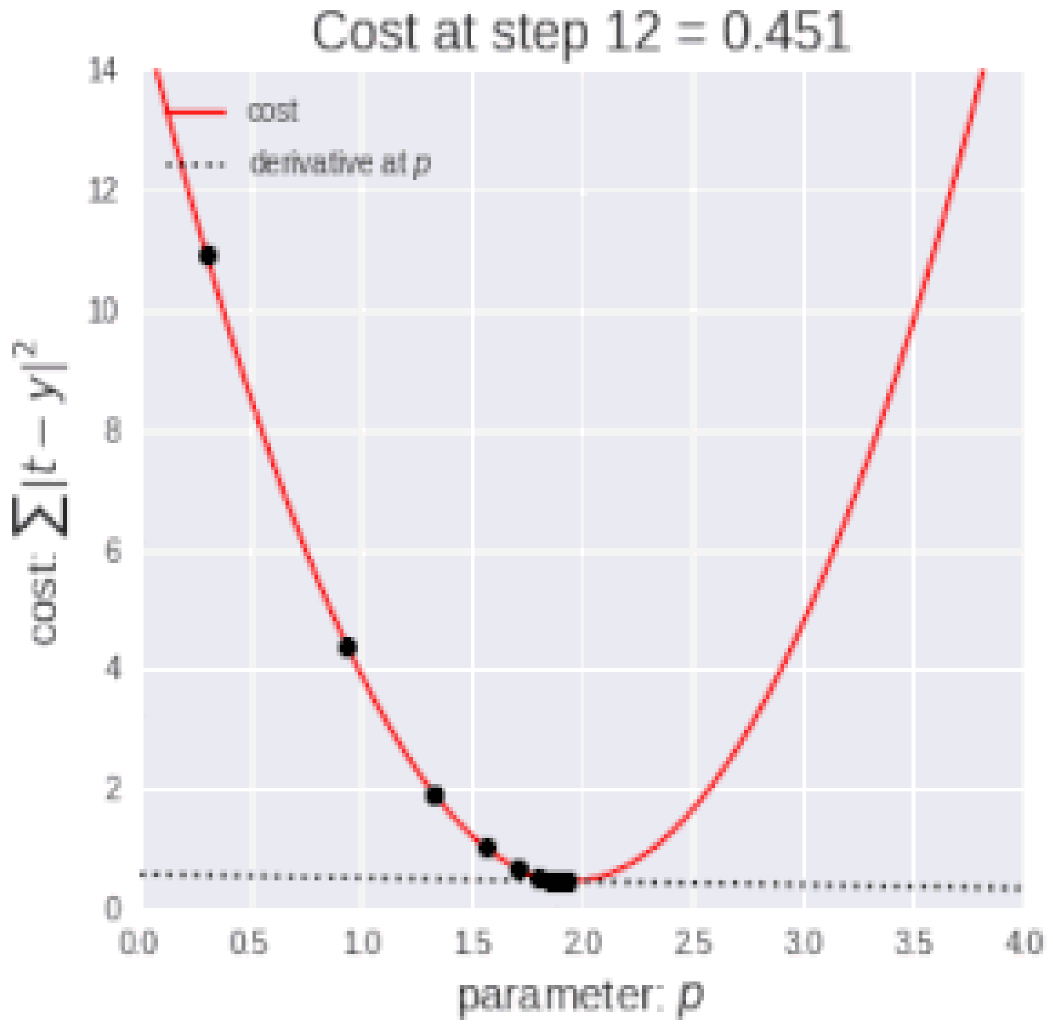
  - Slope first derivative over certain iterations,
  - Learning rate

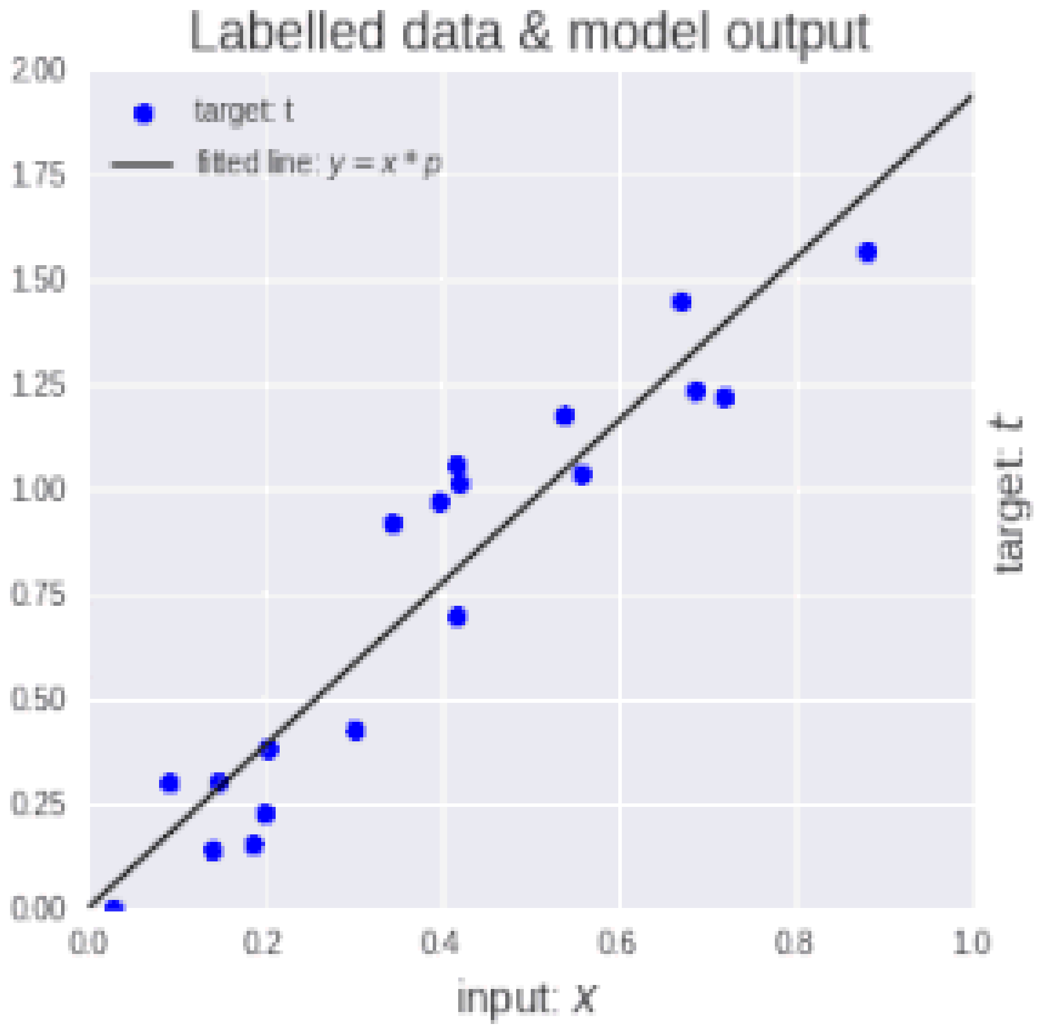


Y : Cost Function (Loss function, Error)
A : Slope
x **:** N# of iterations

Amr.Moslim

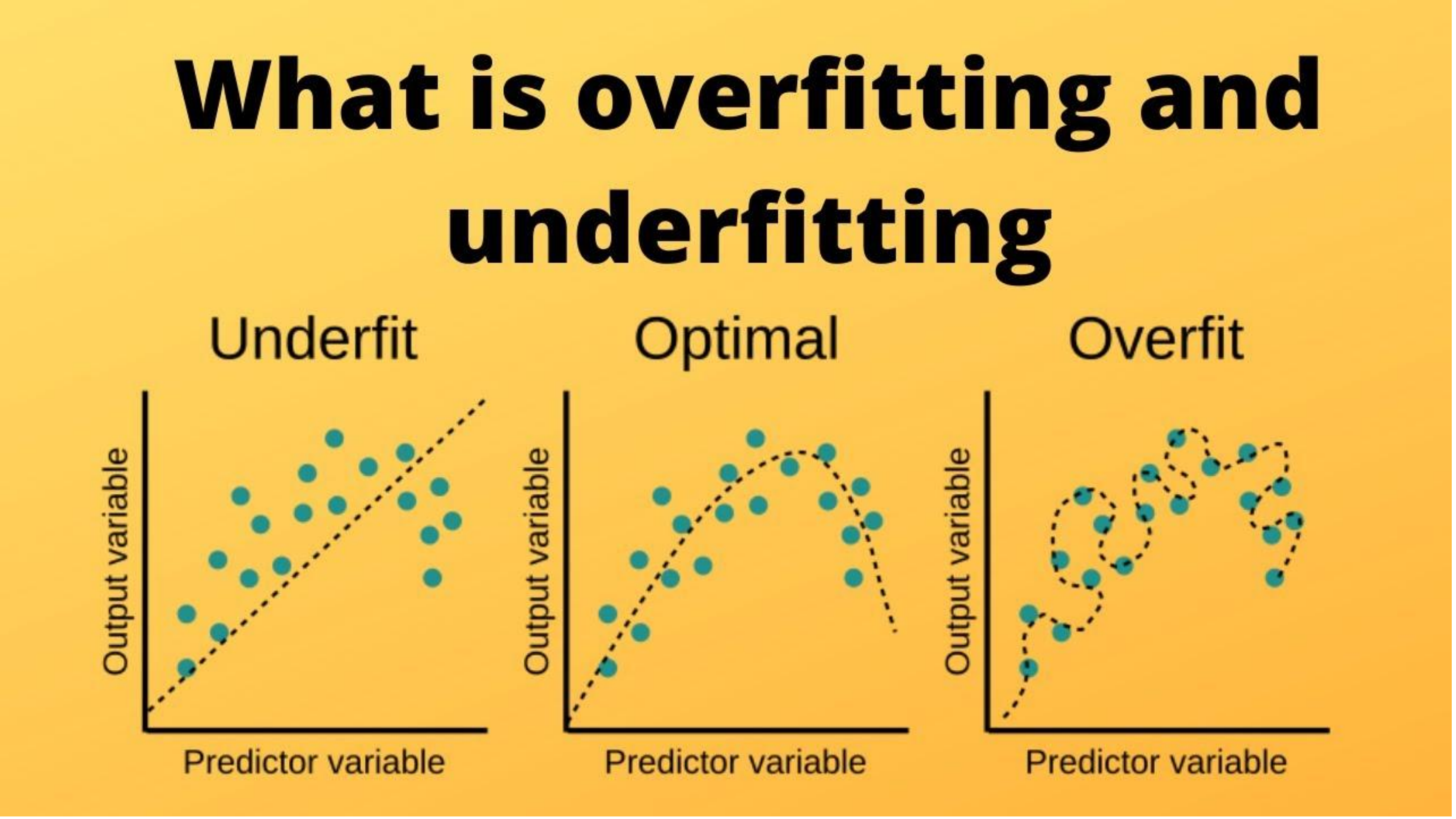
Linear Regression - Gradient Descent





ML Models Evaluation

Model Evaluation

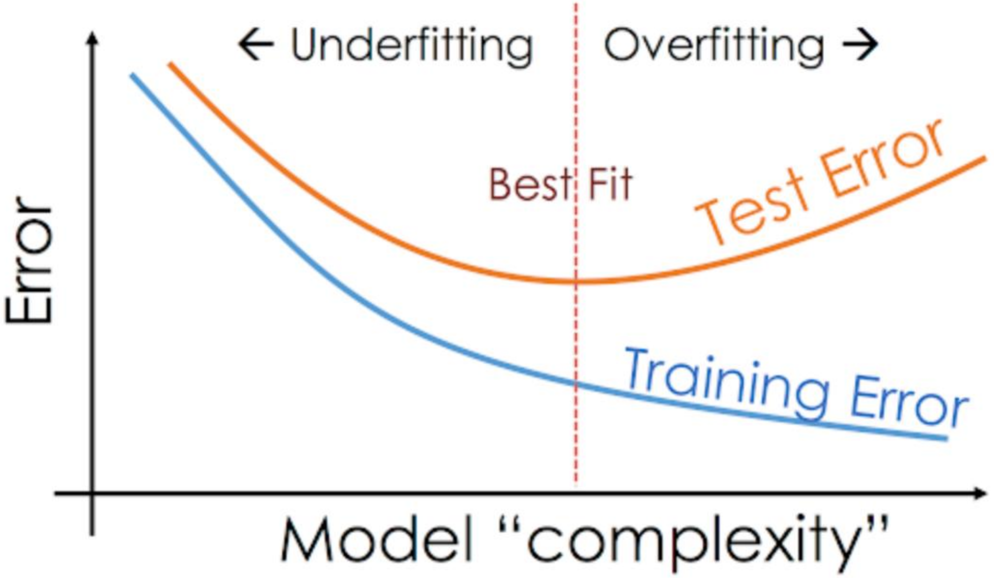


What is overfitting and underfitting

Model Evaluation

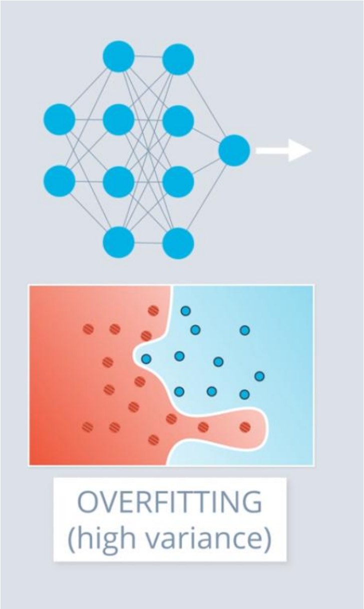
Overfitting means the model has been trained too well

- The model knows too much details for every data point
- The model includes noise as well as the data
- Negatively impact the models ability to generalize
- More likely to happen in nonlinear / nonparametric data



Characteristics of Overfitting:

- High Variance
- Low Bias
- Low standard deviation
- No generalization



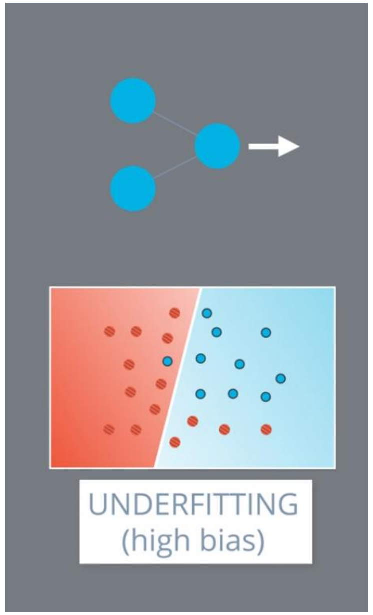
Model Evaluation

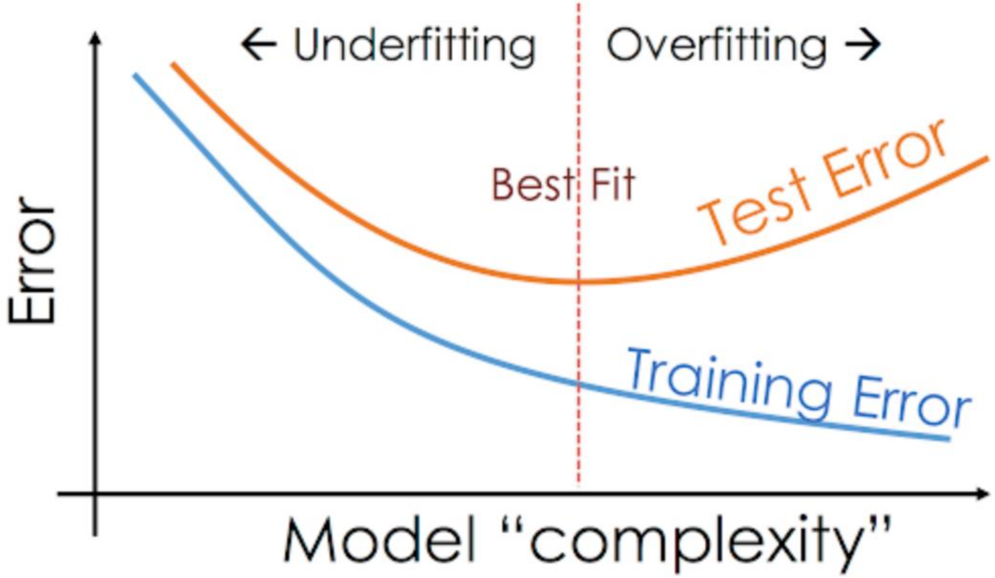
under-fitting means the model is too simple for the training data and test data

- The model knows so little for the whole data set
- It will have poor performance on the training data
- Negatively impact the models ability to generalize
- It is easy to detect given a good performance metric

Characteristics of Overfitting:

- Low Variance
- High Bias
- High standard deviation
- Low generalization





K-Means Classification

- ***Objective:***

  To be able to cluster the data based on the input variable or variables and find cluster centroids

- ***Model structure:***

  Euclidean distance = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

- ***Model assumptions:***
  - ➢ Cluster centroids are in the middle of each cluster
  - ➢ Each cluster has a centroid and data is scattered around it

- ***Parameter estimates and interpretation:***

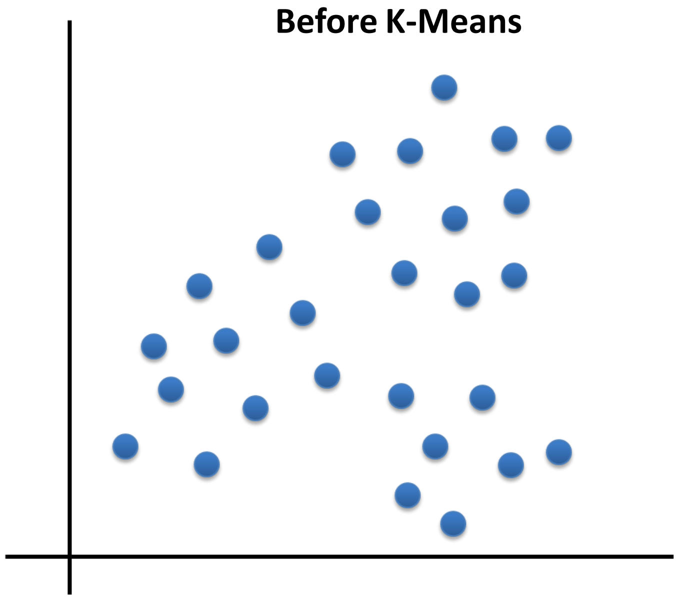
  Find Euclidian distance that correspond to each centroid

- ***Model fit:***

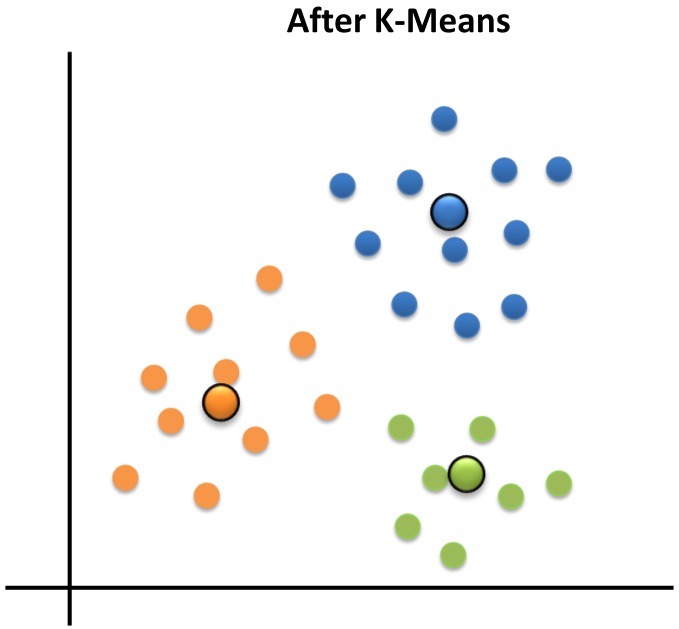
  R^2, residual analysis

- ***Model selection:***

  possible predictors, which variables to include?

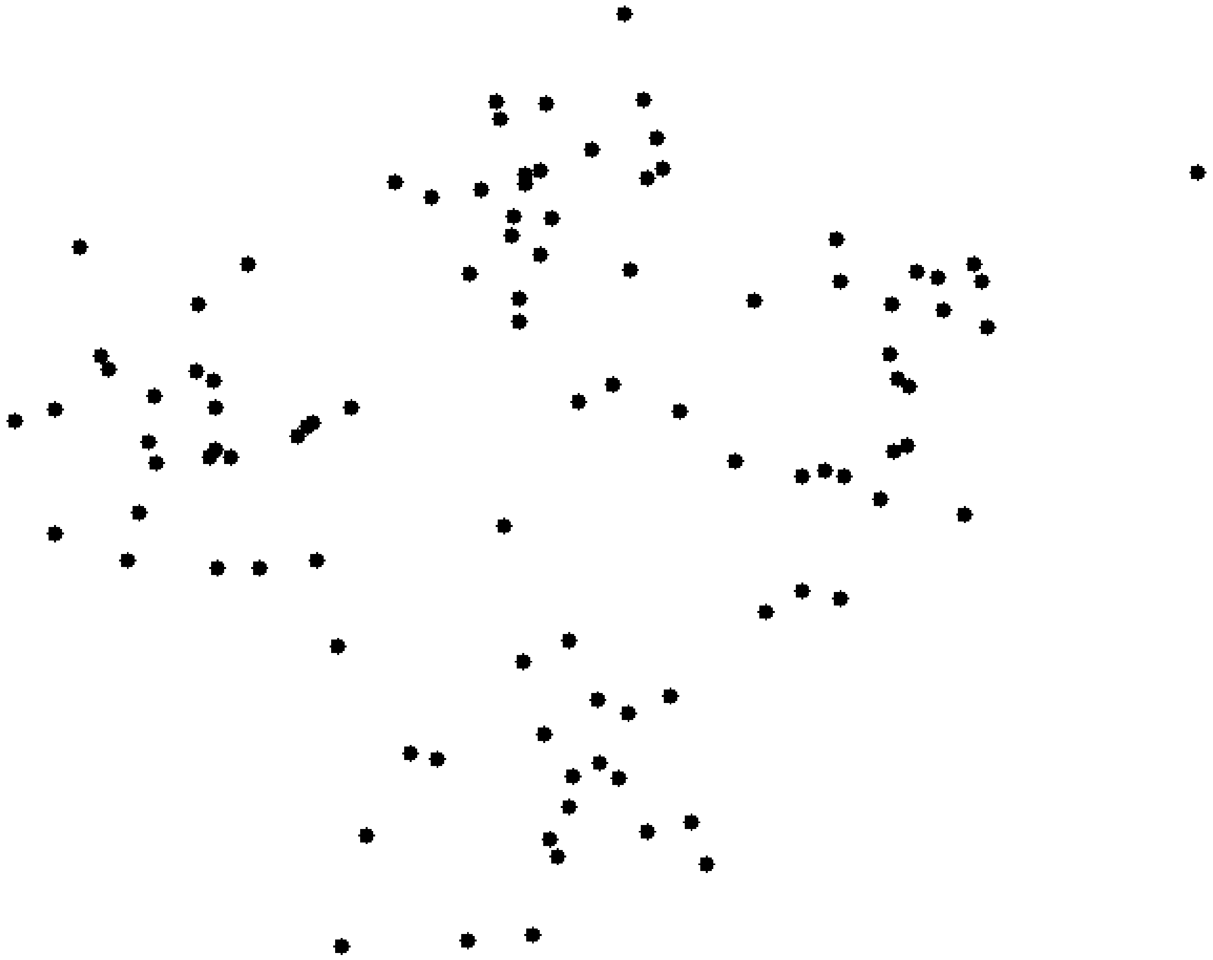


Before K-Means



After K-Means

K-Means Classification



Amr.Moslim

Random Forest

- ***Objective:***

  To be able to build decision boundaries based on the maximum variance between variables

- ***Model structure:***

  <u>Step 1 :</u> *Select random samples from a given dataset.*

  <u>Step 2 :</u> *Construct a decision tree for every sample.*

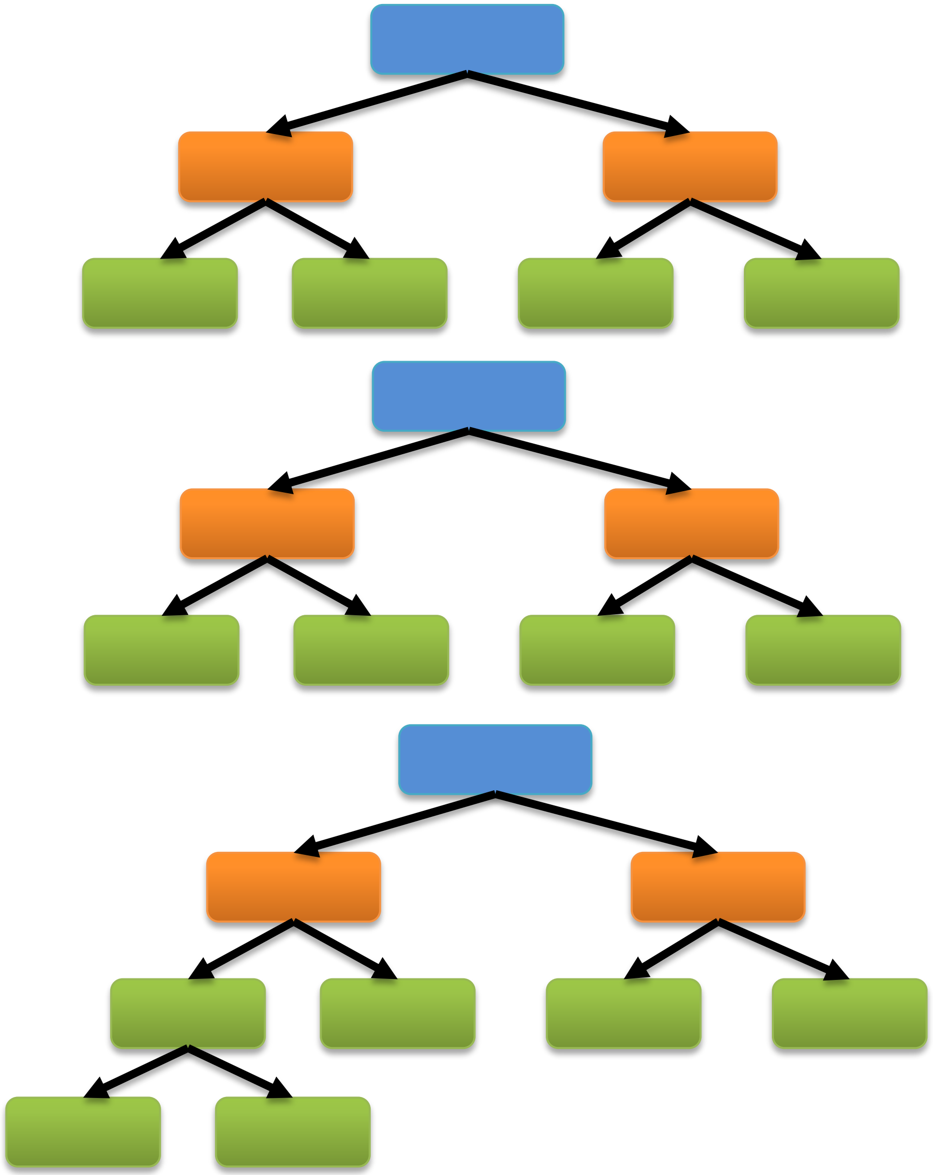
  <u>Step 3 :</u> *Get the prediction result from every decision tree.*

  <u>Step 4 :</u> *Voting will be performed for every predicted result.*

  <u>Step 5 :</u> *Select the most voted prediction result as the final prediction result.*

- ***Model fit:***

  R^2, Confusion Matrix



Amr.Moslim

Confusion Matrix

- ***Confusion Matrix KPI:***
  - ***Precision:*** *true positive rate*

$$\frac{TP}{TP+FP}$$

  - ***Recall:*** *true positive over the 1 class predict*

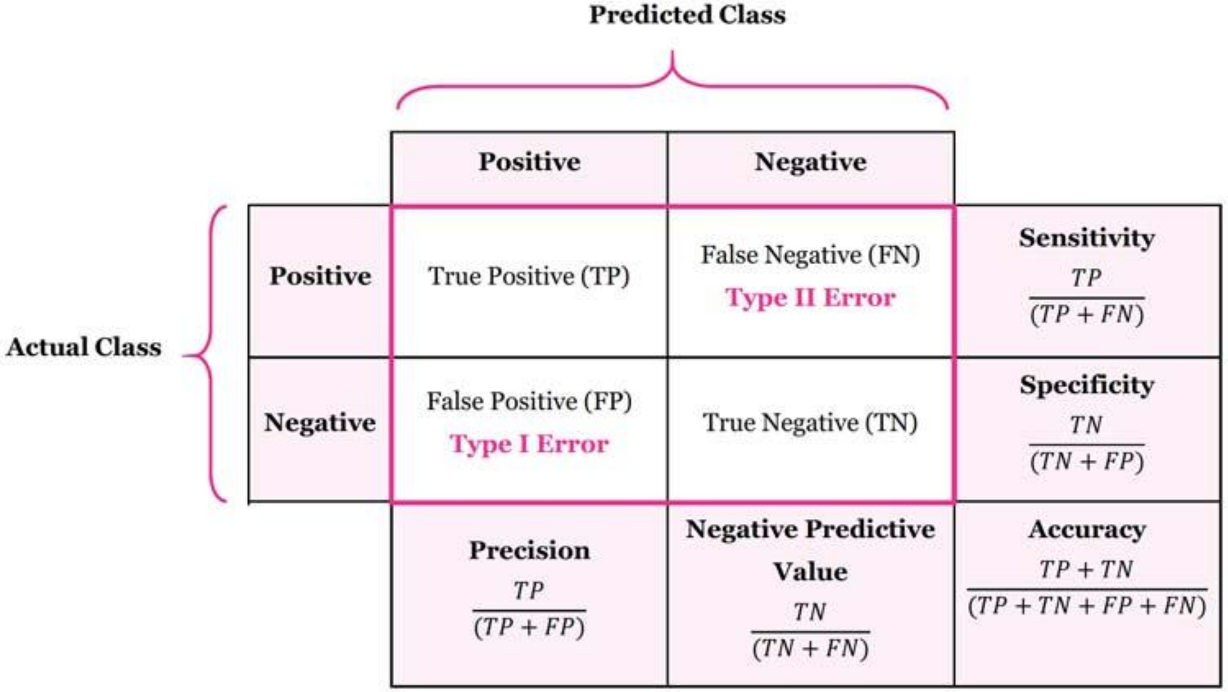
$$\frac{TP}{TP+FN}$$

  - ***F1 Score:***

$$\frac{2*precision*Recall}{Precision+Recall}$$

  - ***Accuracy:***

$$\frac{TP+TN}{TP+TN+FP+FN}$$



		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Thank You for Your Attention