Dealing with Big Data and Machine Learning - Introduction

Al-Amal Program

Amr.Moslim

Dec.2021

Today's Agenda

- What's Big Data
- Evolution of Big Data
- Big Data Challenges
- Big Data Types
- Why Big Data is Important
- Big Data – Popular Use Cases
- Big Data Tools
- How Big Data Works
- Optimize The Value of Big Data
- Introduction to Machine Learning

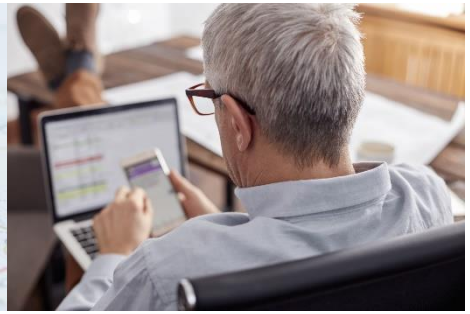
What's Big Data?

64,140
Instagram photos

336,480
Skype Calls

5,365,260
Youtube Videos

5,500,560
Google Searches



181,331,340
Emails Sent

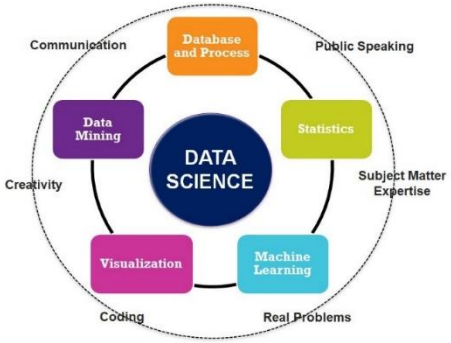
Every Single Second

2.5 Quintillion bytes of data generated Every Single Day

https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

What's Big Data?

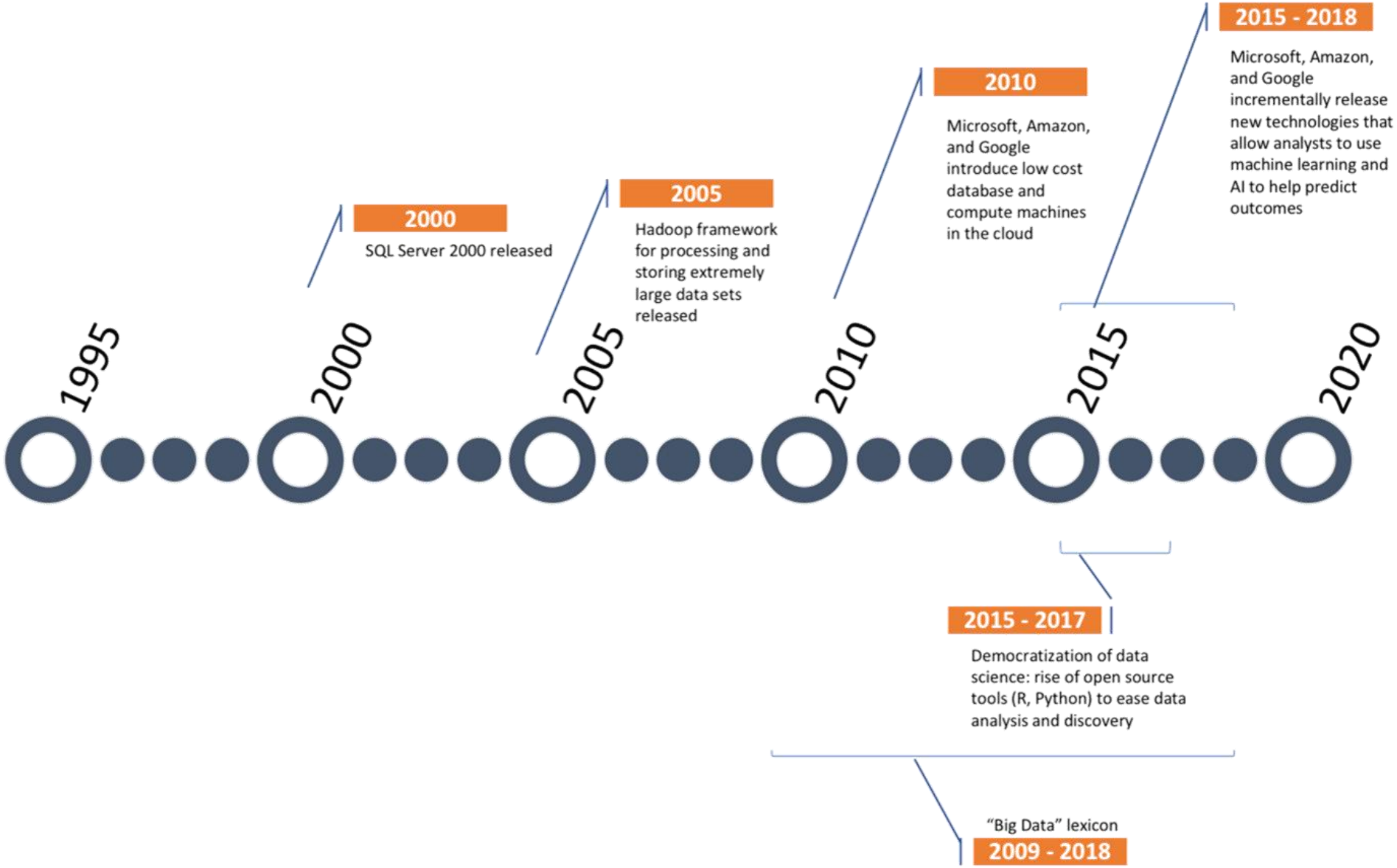
Big data is a term that describes large, hard-to-manage volumes of data – both structured and unstructured – that overwhelm businesses on a day-to-day basis. Big data can be analyzed for insights that improve decisions and give confidence for making strategic business decisions.



But it's not just the type or amount of data that's important, its what organizations do with the data that matters.

https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

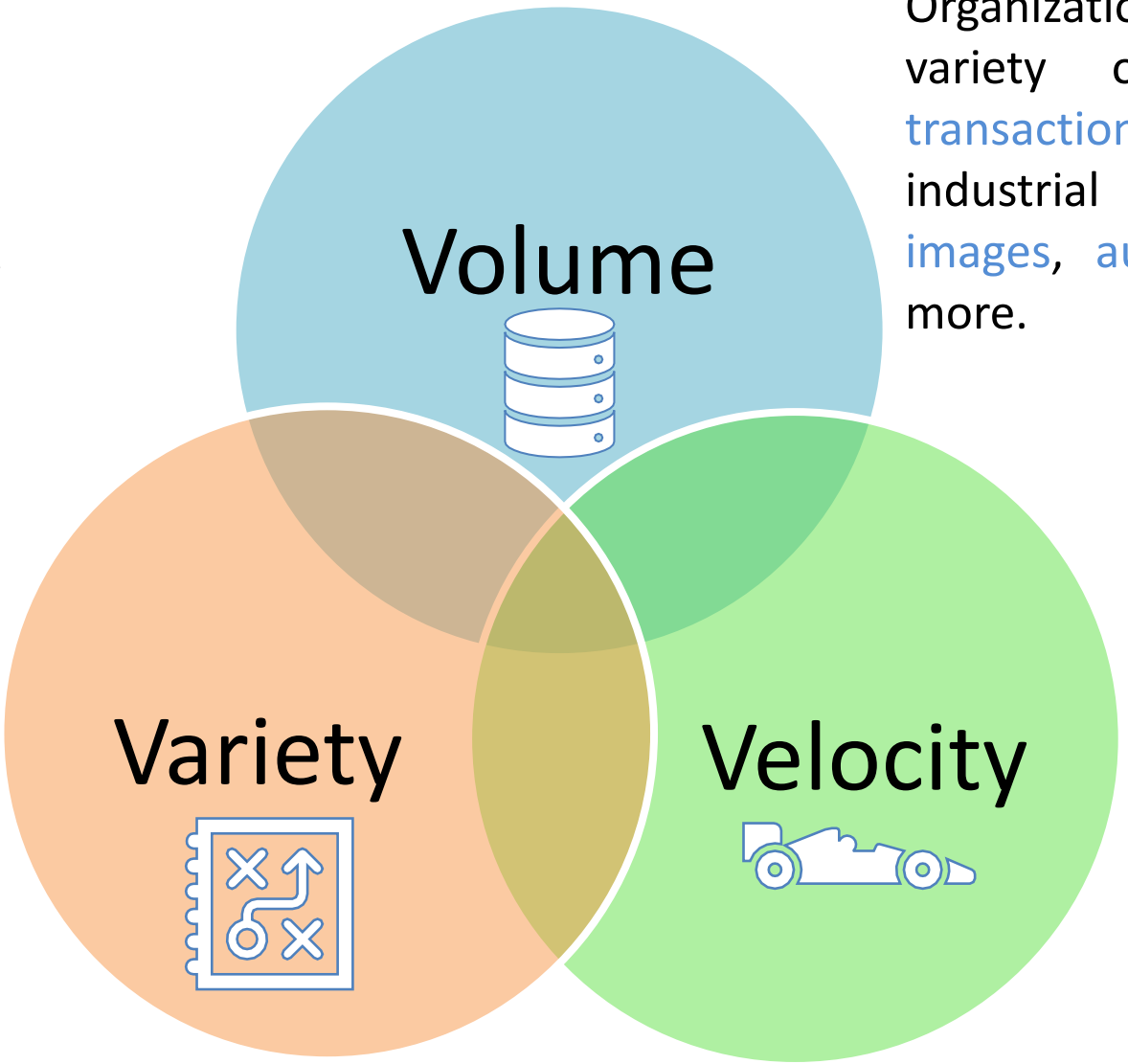
Evolution of Big Data



Big Data Challenges

Organizations collect data from a variety of sources, including transactions, smart (IoT) devices, industrial equipment, videos, images, audio, social media and more.

Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions.

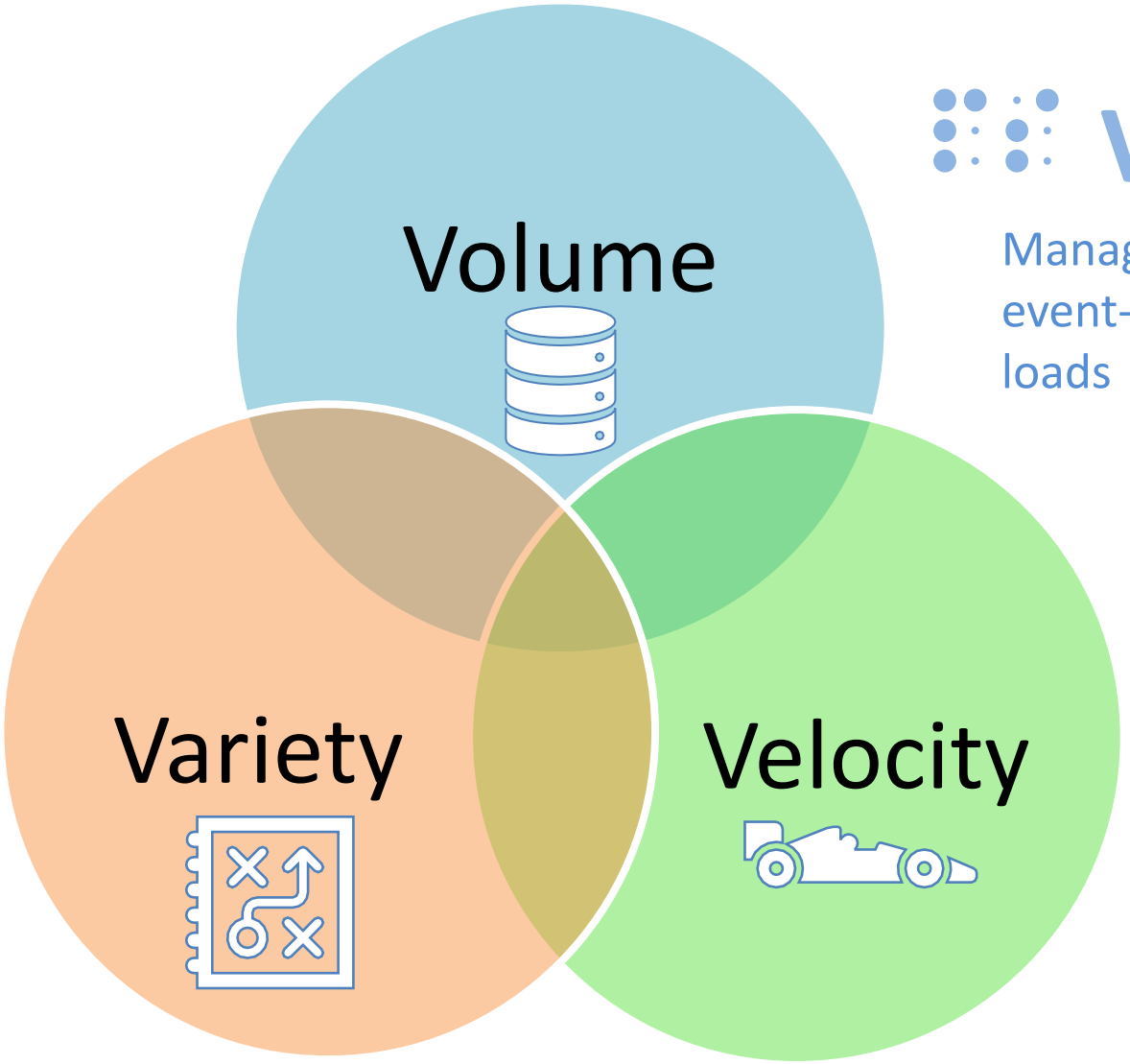


Growth on the Internet of Things, data streams into businesses at an unprecedented speed and must be handled in a timely manner. sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

Big Data Challenges

Veracity

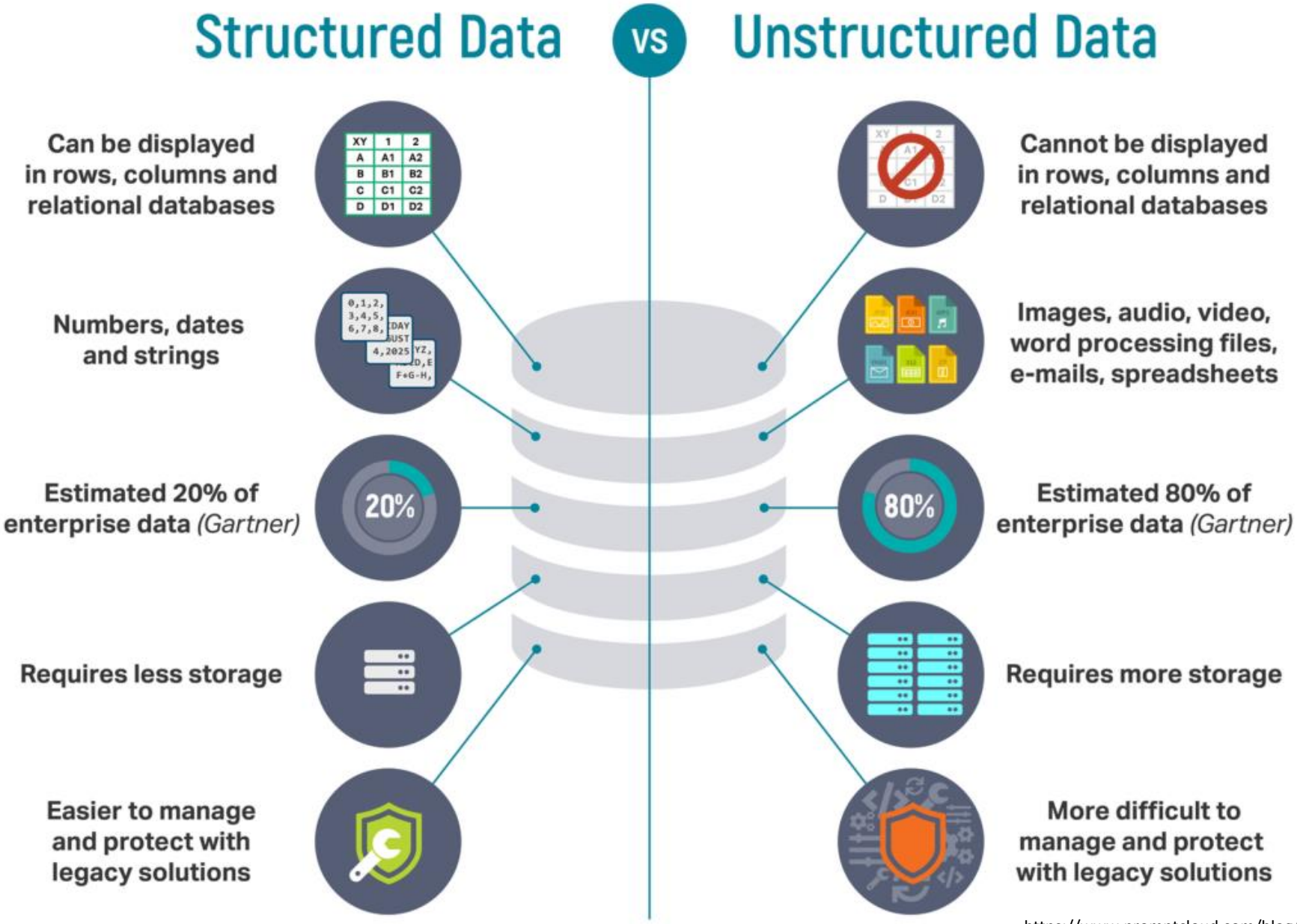
the quality of data. Businesses need to connect and correlate relationships, hierarchies and multiple data linkages. Otherwise, their data can quickly spiral out of control.



Variability

Manage daily, seasonal and event-triggered peak data loads

Big Data Types



https://www.promptcloud.com/blog/big-data-evolution-technology-modern/

Why Big Data is Important

- Enable smart decision making.
- Improve and fast track the operational efficiencies
- Optimize product development
- Drive new revenue and growth opportunities
- Streamline resource management
- Automate tasks that takes more time and resources so the individual can focus on the intelligence part

https://www.promptcloud.com/blog/big-data-evolution-technology-modern/

Big Data – Popular use cases

Internet of Things

Information Security

Gathering rich insights about business

Data warehouse optimization

Improving healing and public health

https://www.promptcloud.com/blog/big-data-evolution-technology-modern/

Big Data – Oil & Gas

Gathering real time information while drilling

Production monitoring sensors

Seismic data acquisition products

Geoscience data (well logs, cores, maps, remote sensing images)

Well test data and reservoir monitoring data

Data Storage and Management tools

SQL

mongoDB

cassandra

neo4j

APACHE HBASE

talend

APACHE hadoop

HDInsight Microsoft

Apache ZooKeeper™

A

MS Azure

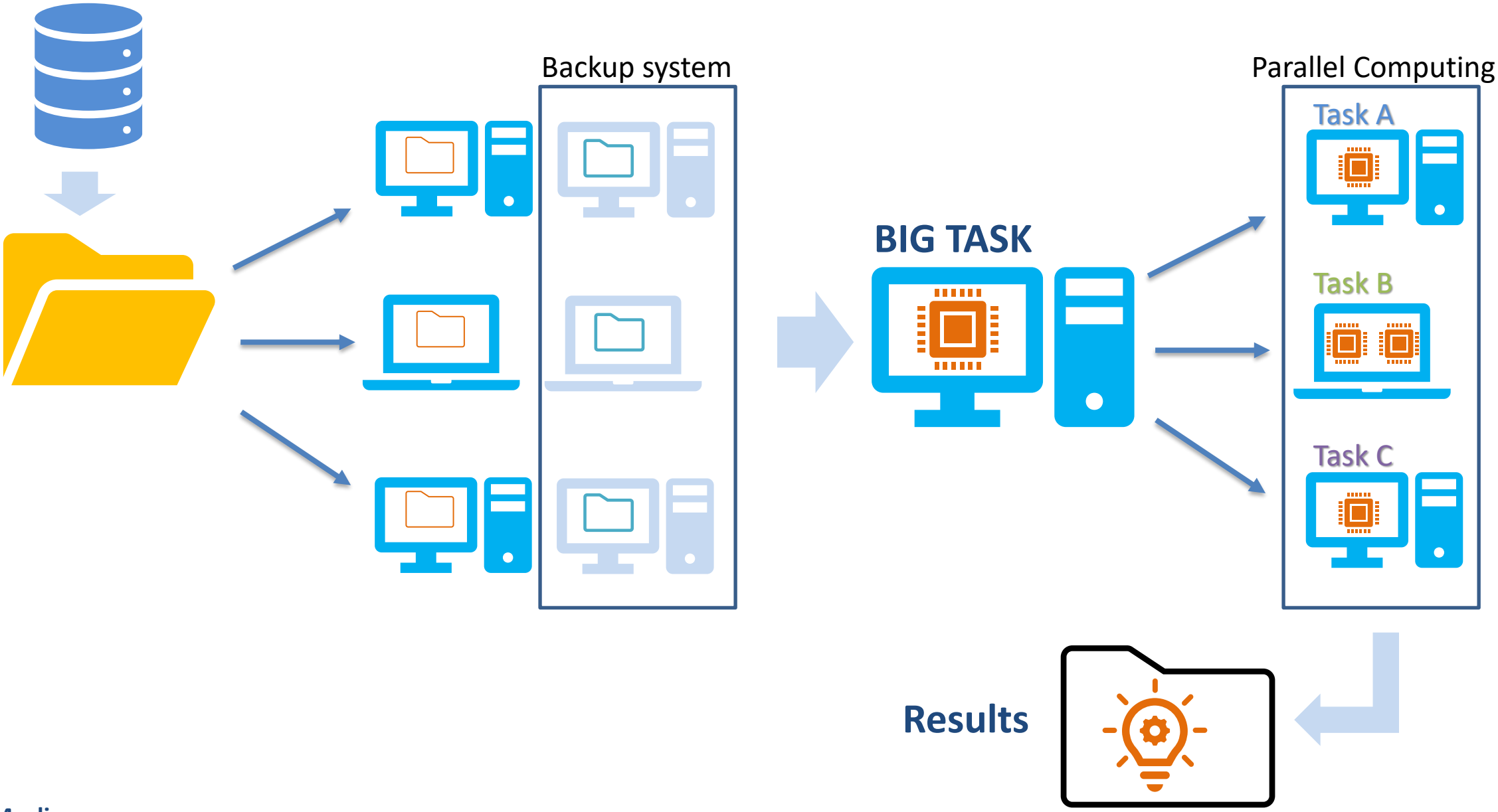
CloudDB

Data Cleaning and Processing tools

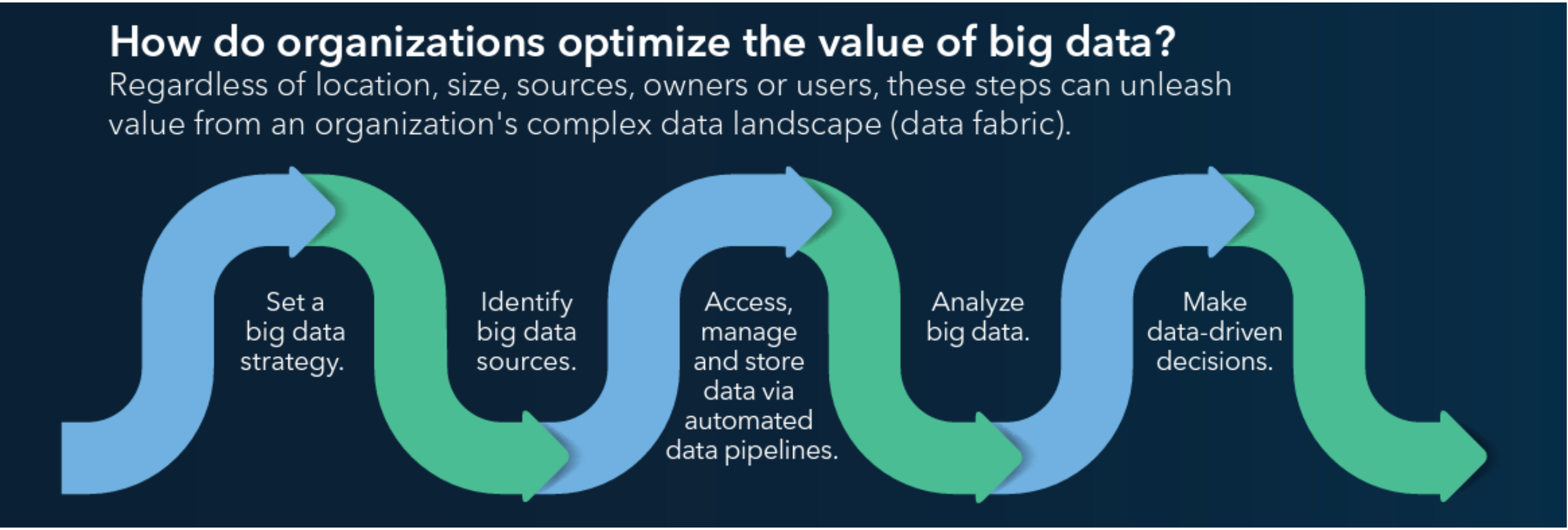
"Deep learning craves big data because big data is necessary to recognize hidden patterns and to find answers without overfitting the data. With deep learning, the more good quality data you have, the better the results"

ATLAS.ti python™ R OpenRefine

Big Data – how it works?



Optimize The Value of Big Data



Break

What's Machine Learning?

Machine learning is the field of Data Science and 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

Machine Learning Algorithm Classes

Supervised Learning

Labeled data

Unsupervised Learning

unlabeled data

- ***Data has labels (reference) model should learn.***
- ***Model should be continuously test based on the label prediction or classification.***

- ***Data has NO labels. Data learn from itself.***
- ***Model should be judged based on certain criteria.***

Machine Learning Algorithm Classification

Supervised Learning

Labeled data prediction

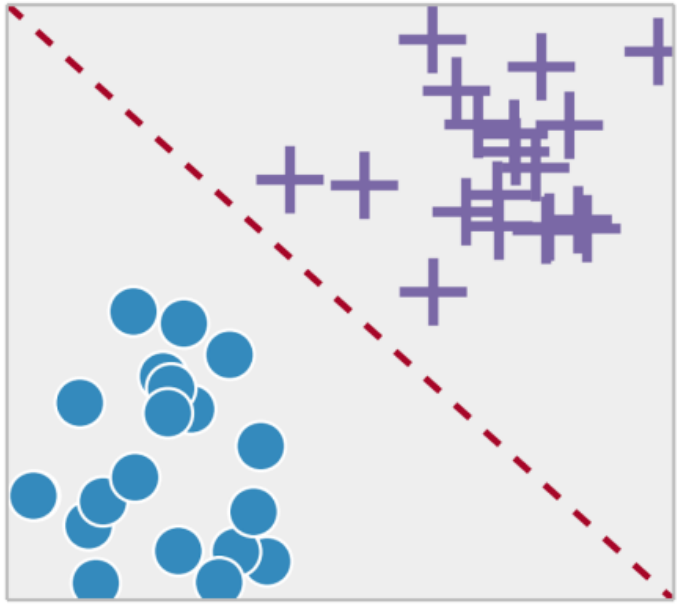
- Regression
- Classification

Unsupervised Learning

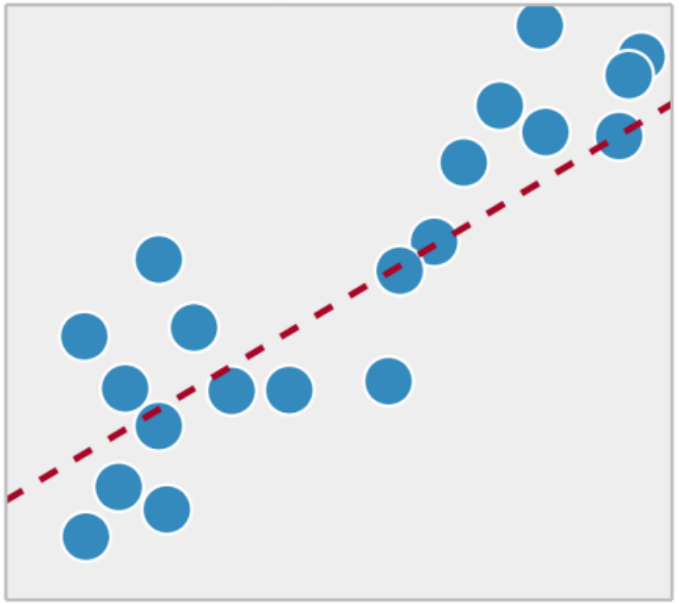
unlabeled data

- Dimensionality reduction
- Clustering

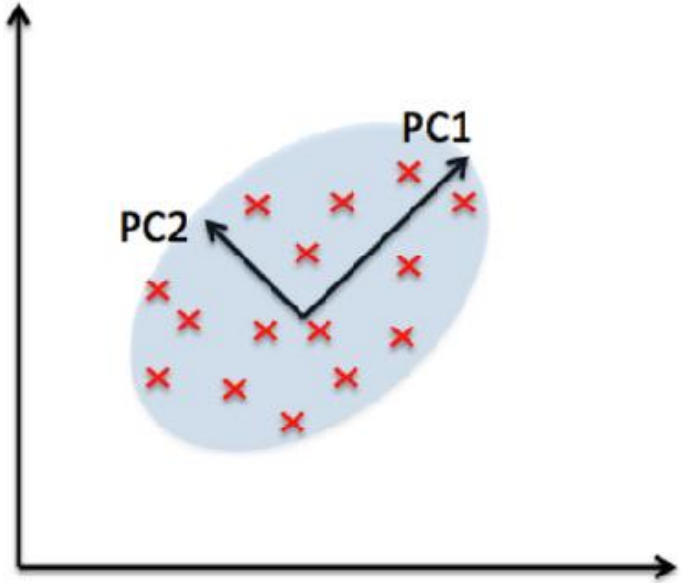
Classification



Regression



Dimensionality reduction



Machine Learning Algorithm Classification

Supervised Learning

Labeled data prediction

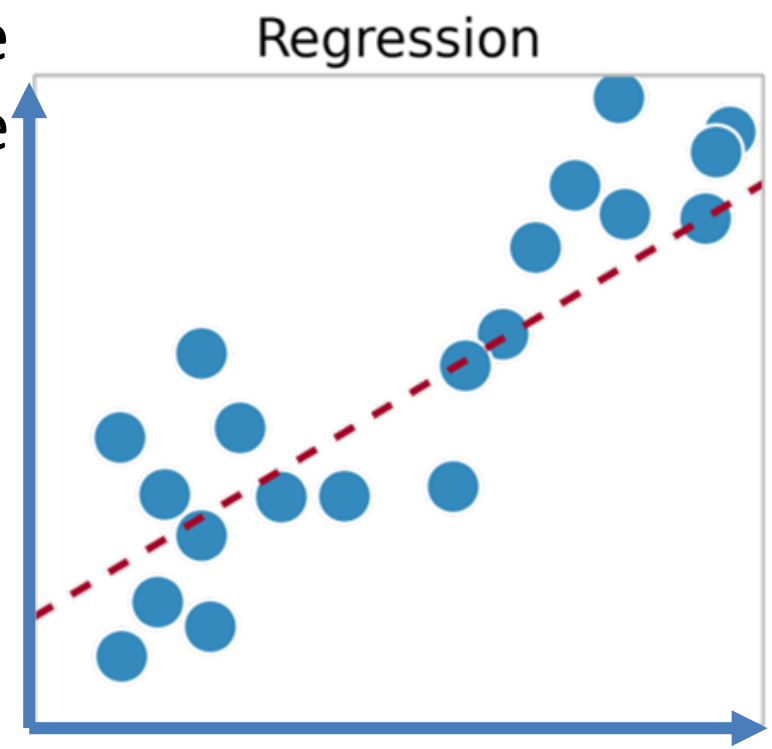
- Regression
- Classification

Regression:

is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

- Statistical Modeling Technique
- Types (Linear, Logistic, Polynomial, …)
- Data is numerical values (Not Categorical)

Example : missing logs predication



Machine Learning Algorithm Classification

Supervised Learning

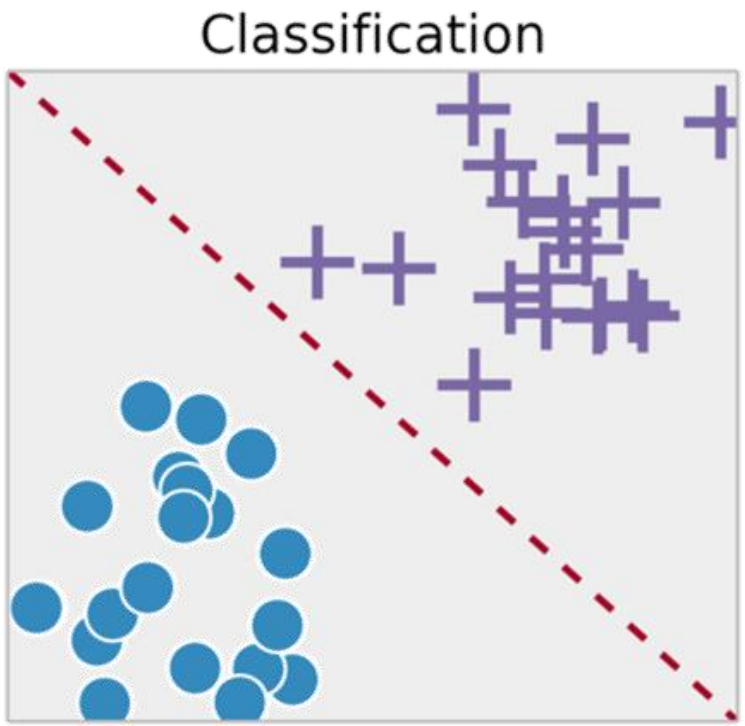
Labeled data prediction

- Regression
- Classification

Classification: (Categorization)

systematic arrangement in groups or categories according to established criteria

- Uses predefined classes
- Belongs to which class



Example : Fraud Detection (Spam / No Spam)
Facies Classification

Machine Learning Algorithm Classification

Unsupervised Learning

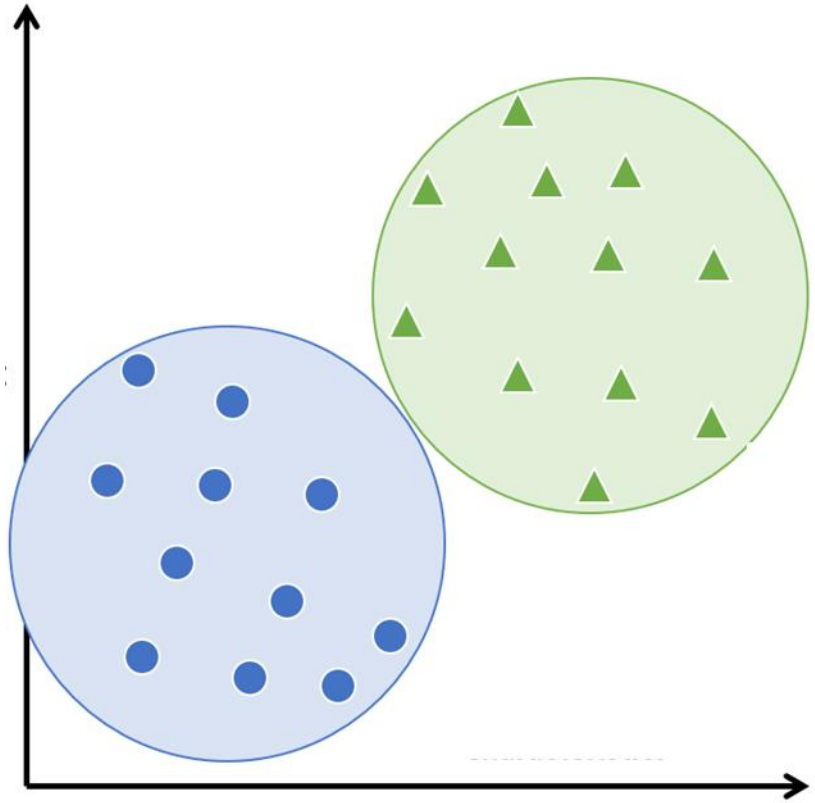
unlabeled data

- Dimensionality reduction
- Clustering

Clustering:

identifies similarities between objects, which it groups according to those characteristics in common and which differentiate them from other groups of objects. These groups are known as "clusters".

- NO predefined classes
- Similar data points properties clusters together



Example : Customer Segmentation

Facies Classification (first time ☺)

Machine Learning Algorithm Classification

Unsupervised Learning

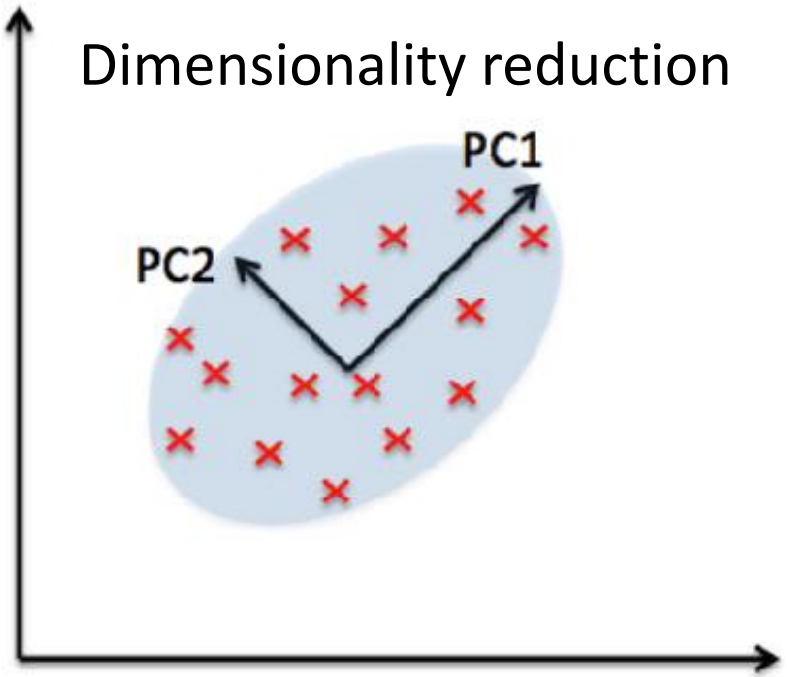
unlabeled data

- Dimensionality reduction
- Clustering

Dimensionality Reduction:

Analyzing the datasets with an extremely high number of features
is often performed to obtain better input features for machine learning algorithms.

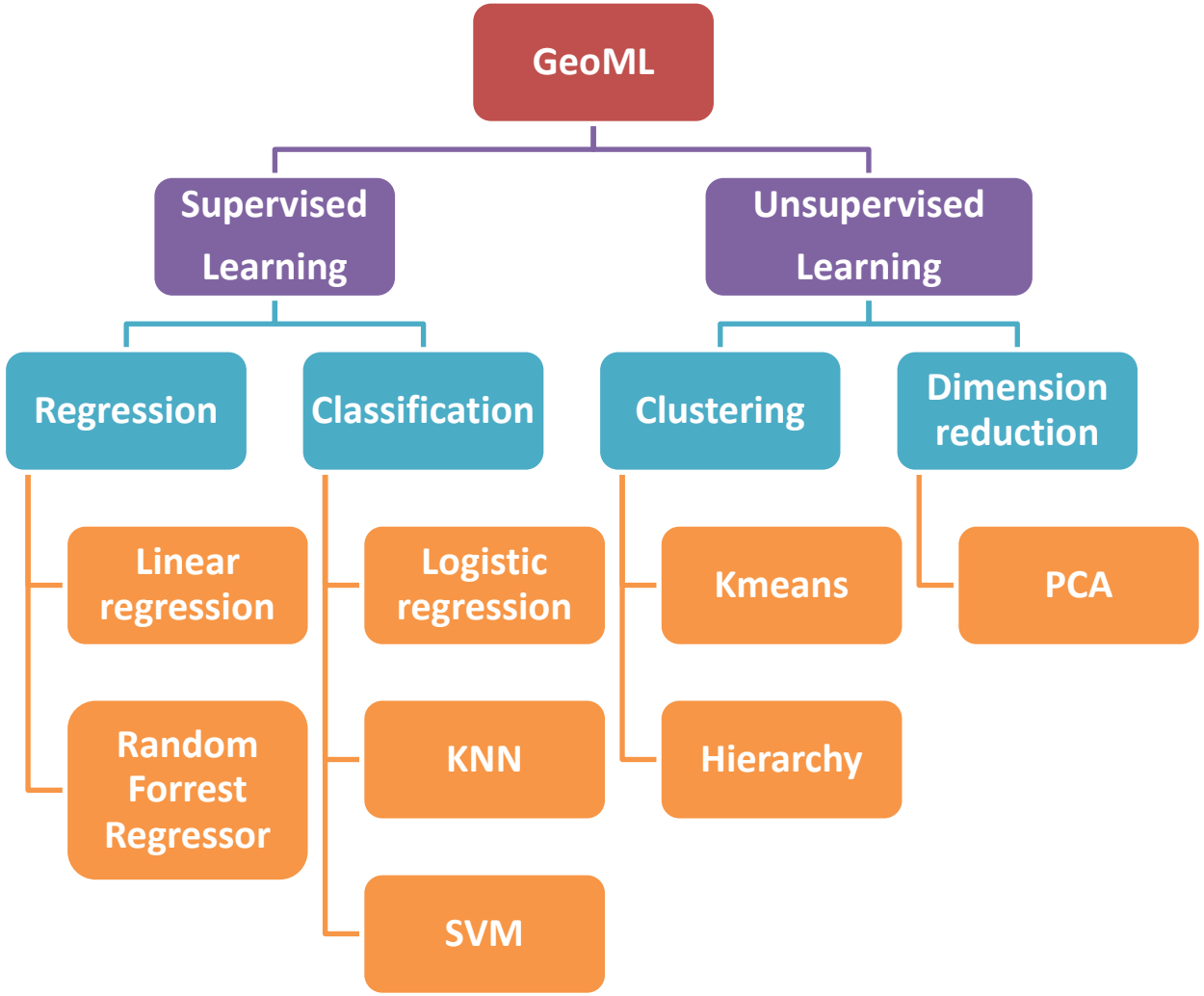
- It improves computational efficiency without sacrificing much on the prediction capability
- removes the collinearity



Machine Learning Algorithms

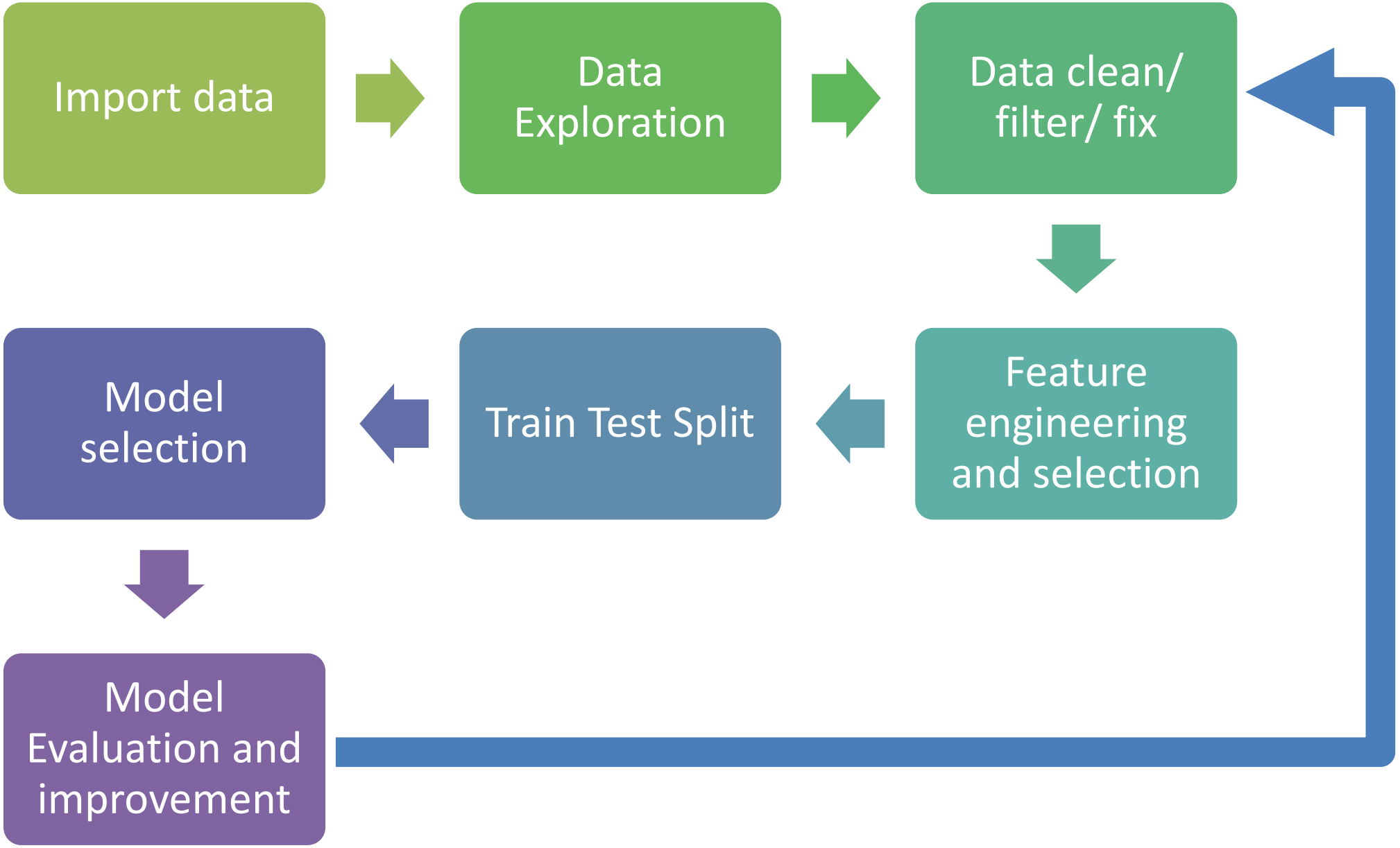
Most commonly used Machine learning algorithms:

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



MACHINE LEARNING WORKFLOW

Machine Learning Work flow



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:***

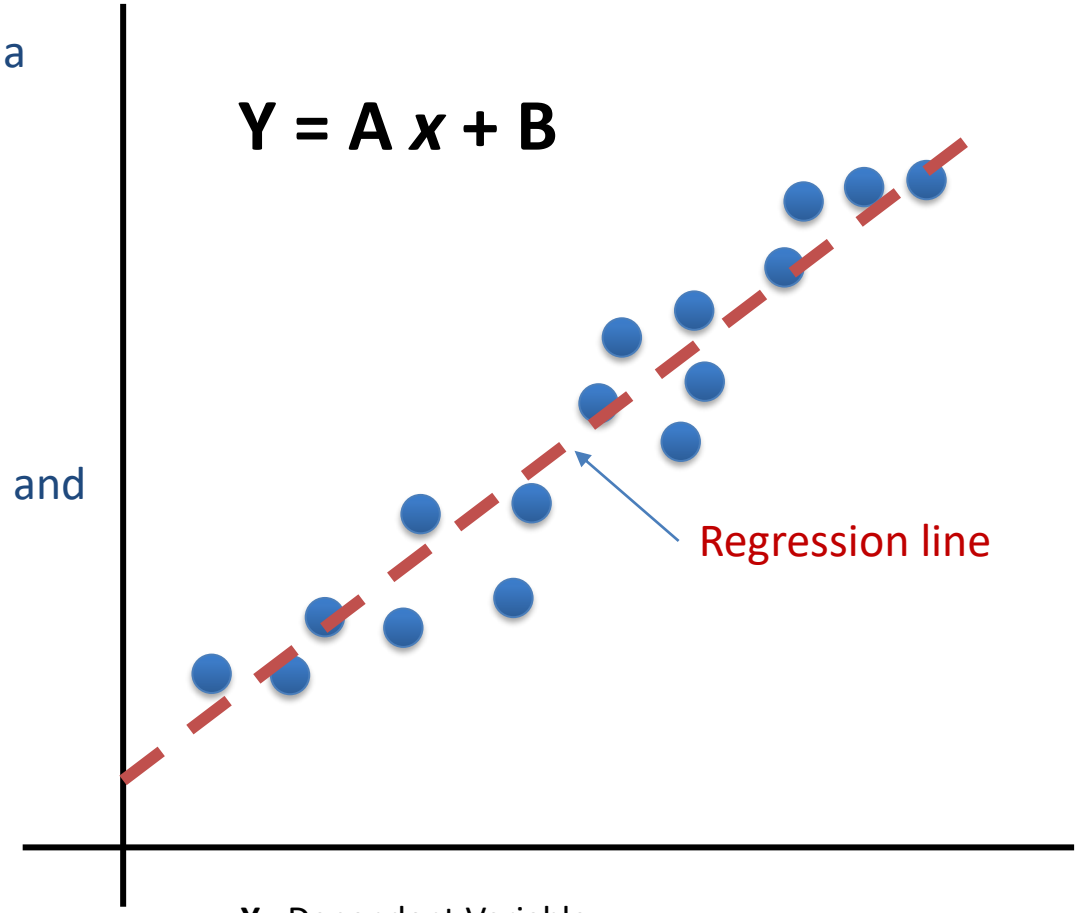
  B the intercept, and **A** is estimate of the slope

- ***Model fit:***

  R^2, residual analysis

- ***Model selection:***

  possible predictors, which variables to include?



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

Linear Regression - Gradient Descent

- ***Objective:***

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

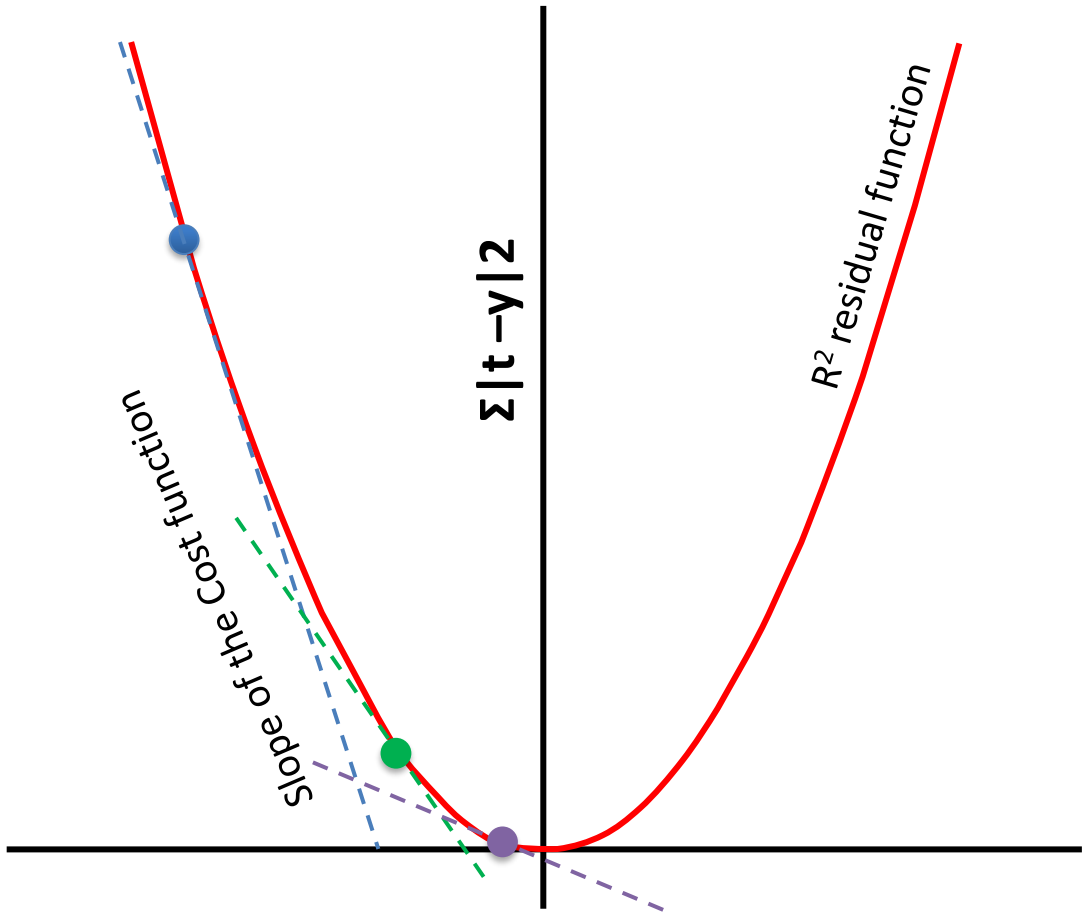
- ***Function structure:***

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

- ***Model assumptions:***

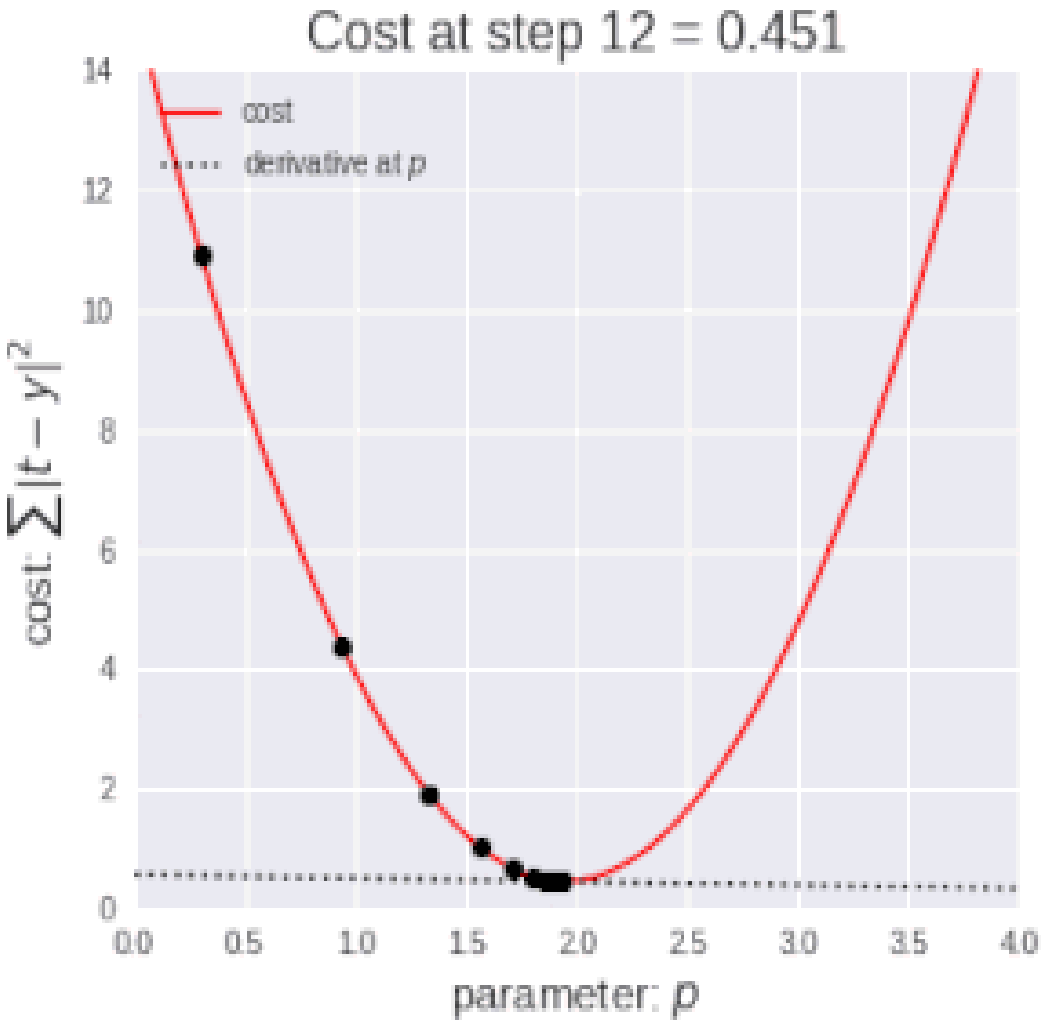
  Slope of the *cost function* ~= Zero, then it is the best prediction

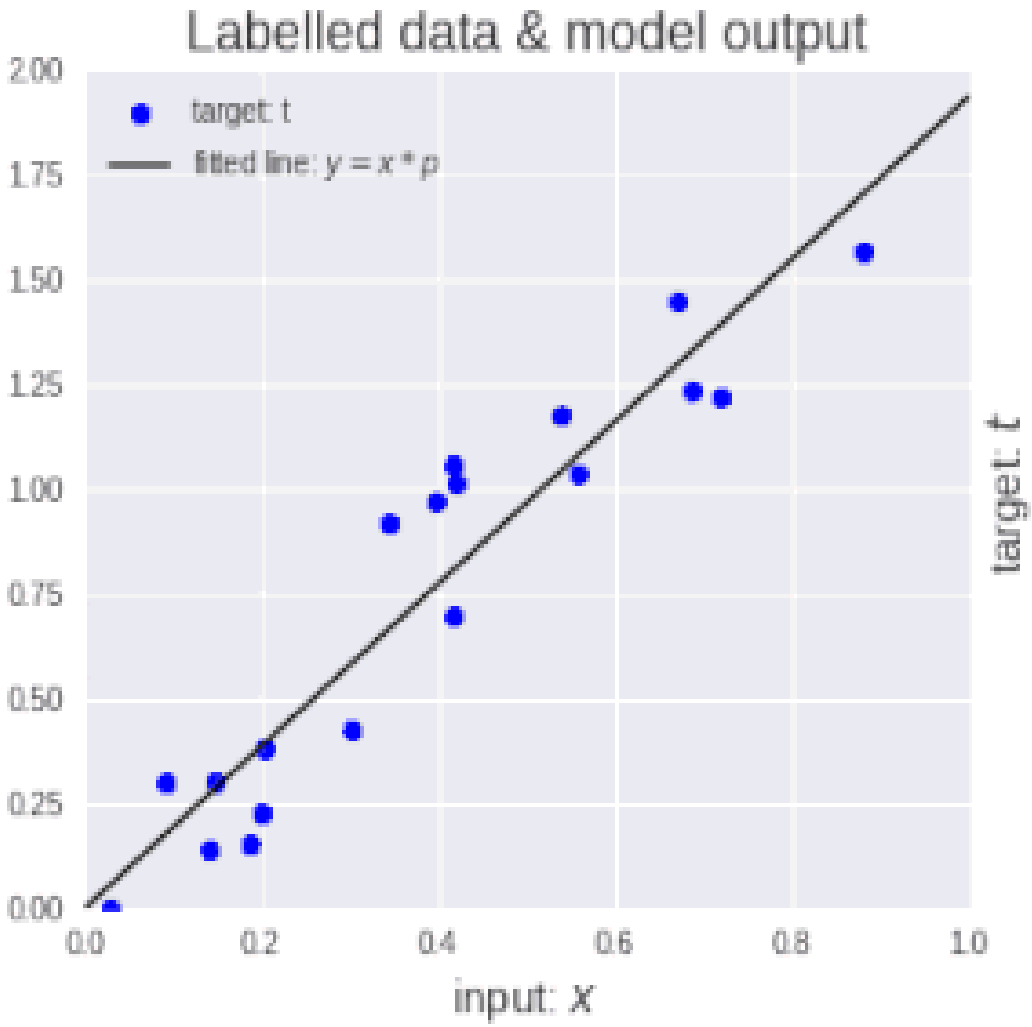
- ***Parameter estimates and interpretation:***
  - Slope first derivative over certain iterations,
  - Learning rate



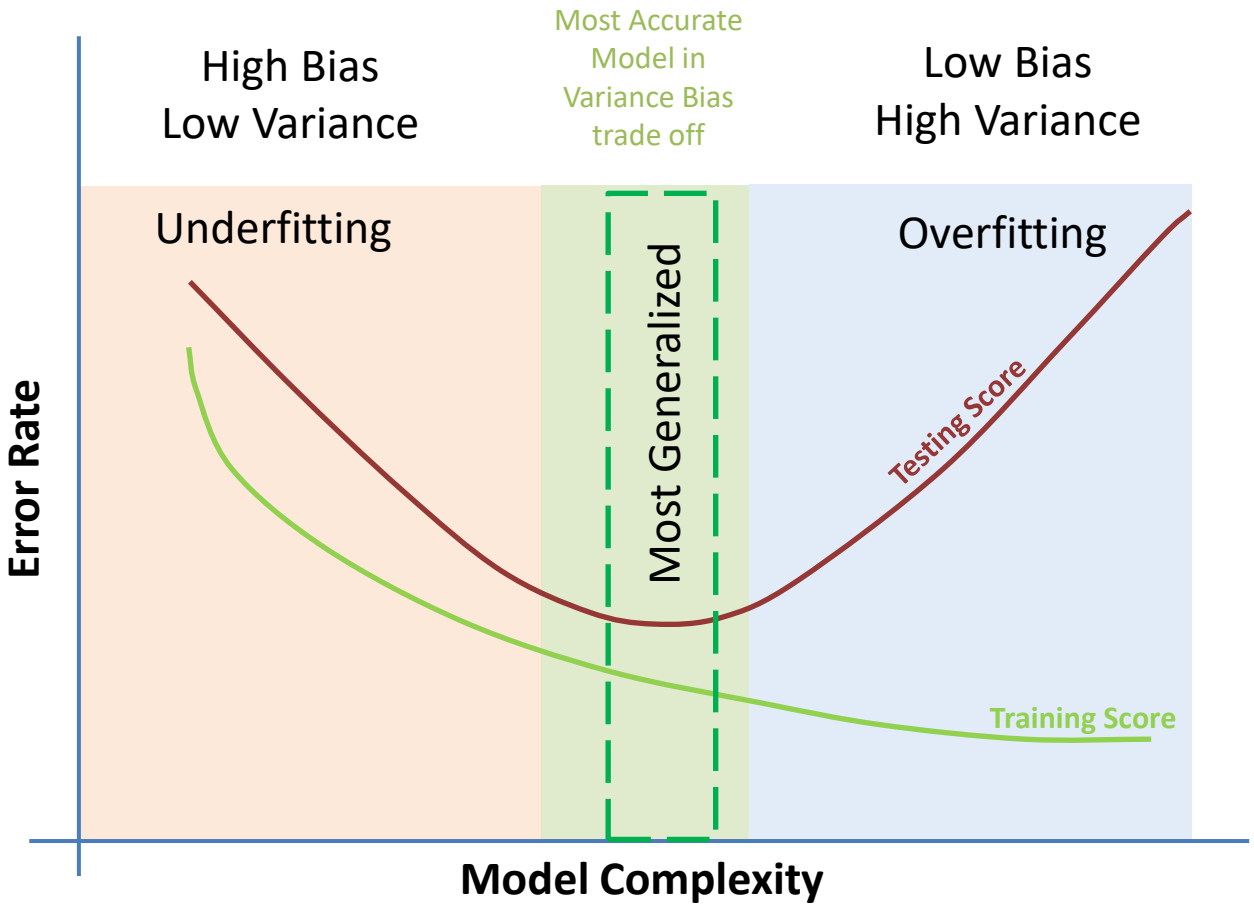
Y : Cost Function (Loss function, Error)
A : Slope
x **:** N# of iterations

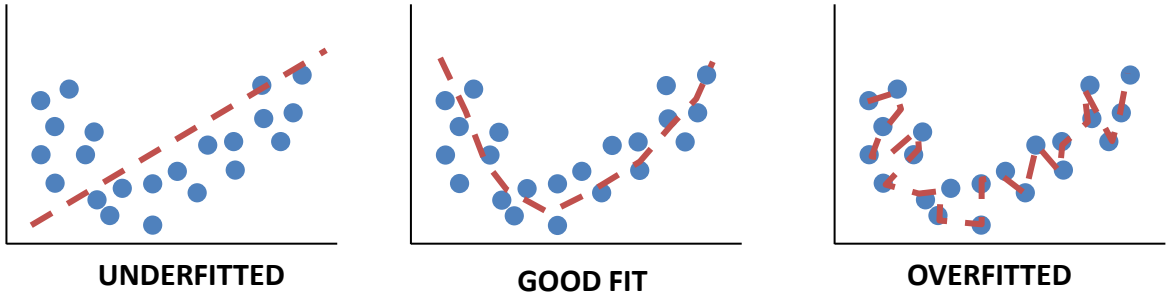
Linear Regression - Gradient Descent





Model Evaluation





Thank You for Your Attention

Amr Moslim

Chat

الله خيرا

Mohammed Essam to Everyone 07:34 PM

ME

ألف شكر يا باشمهندس

Alaa Abdel... to Hosts and panelists 07:34 PM

AA

شكرا جدا لحضرتك يا دكتور

Abdelrahman Maher to Everyone 07:36 PM

AM

شكرا جدا يا هندسة

ahmed wagdy to Everyone 07:36 PM

AW

شكرا يابشمهندس وربنا يكرمك ويوفقك دايما

Mona nady to Everyone 07:37 PM

MN

شكرا لحضرتك يا دكتور

Asmaa Salah to Everyone 07:37 PM

AS

شكرا جدا

Asmaa Abobaker to Everyone 07:37 PM

AA

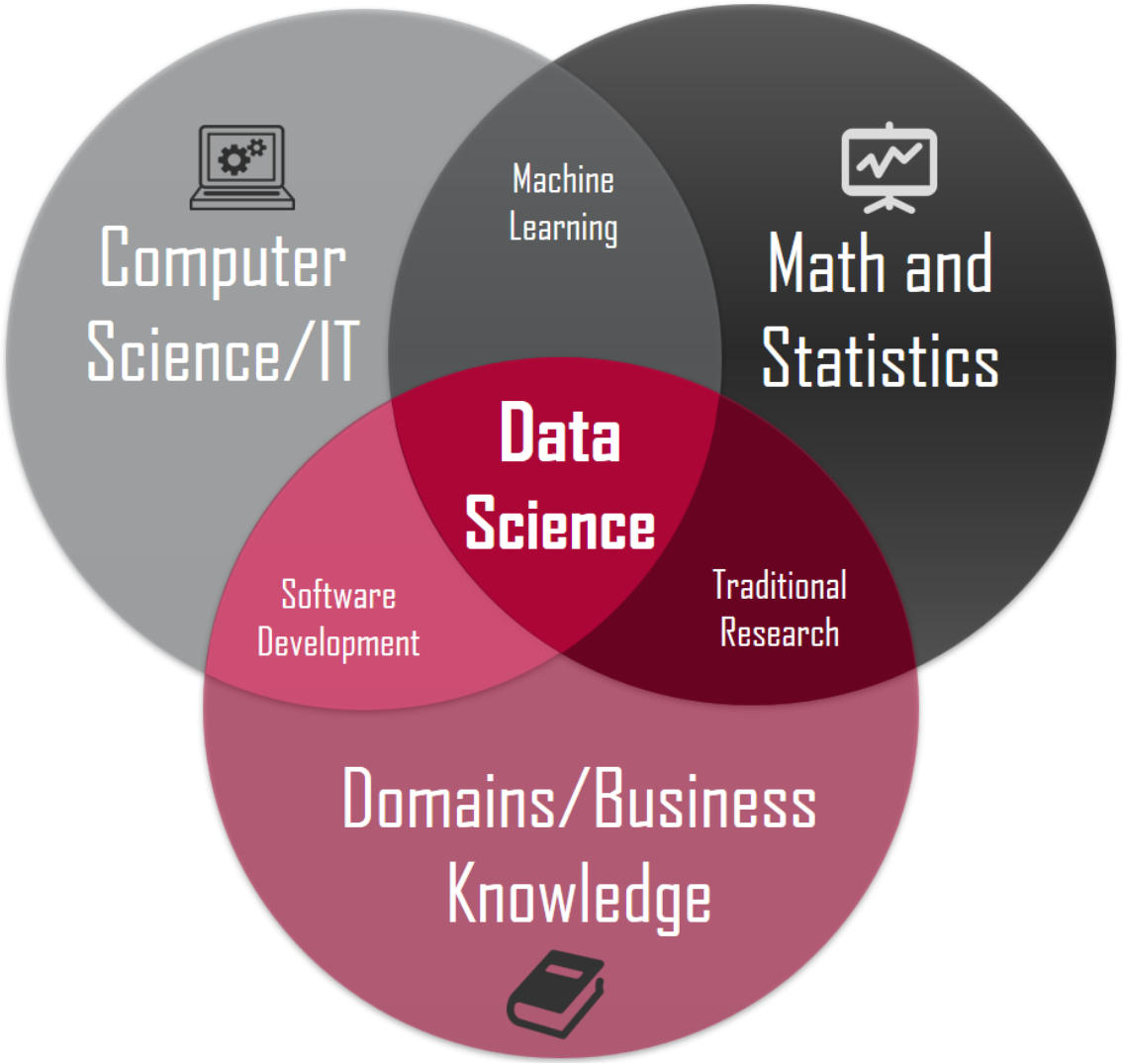
شكرا جدا لتعبك ي بشمهندس

Who can see your messages? Recording On

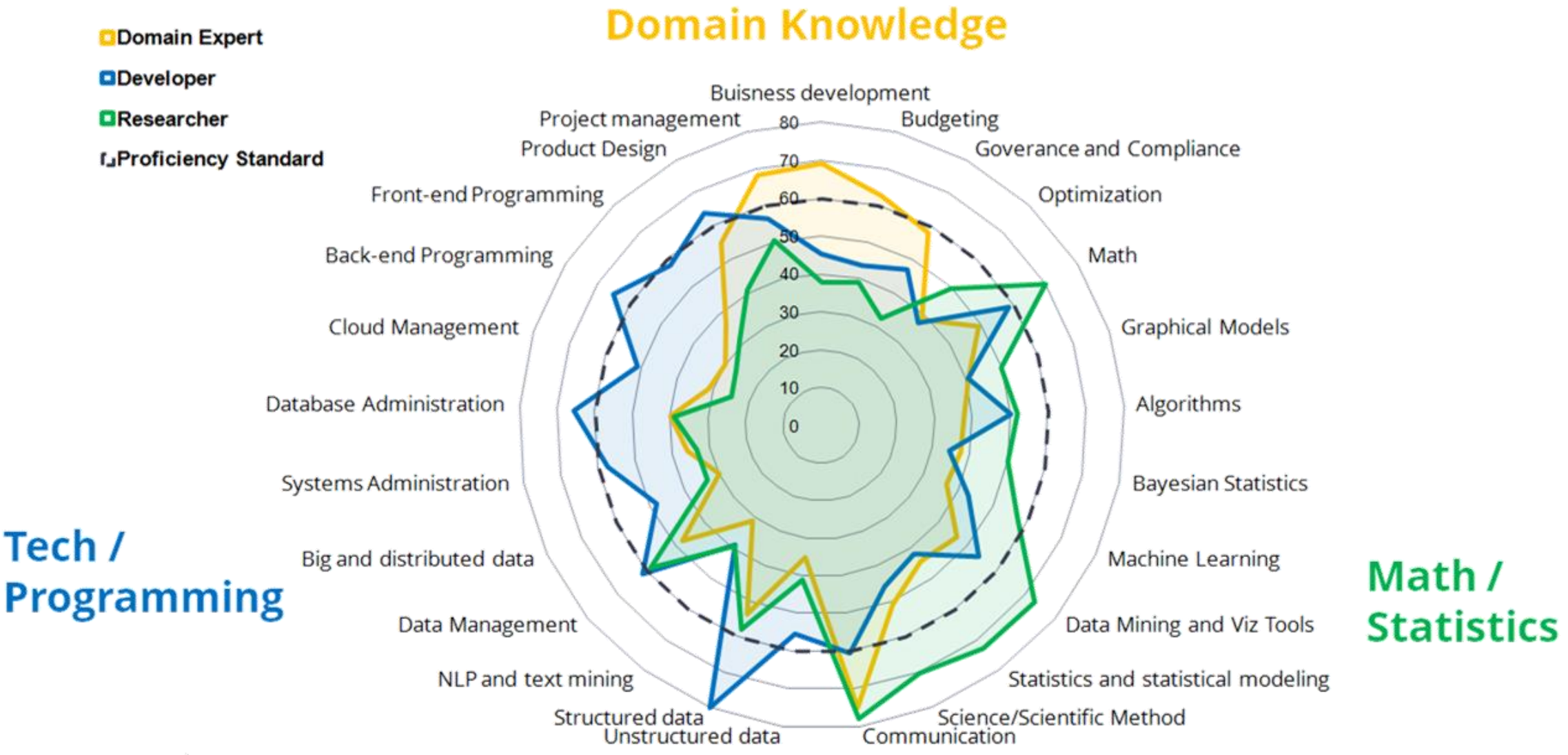
To: Ayman (Direct Message)

Type message here...

Data Science knowledge domains

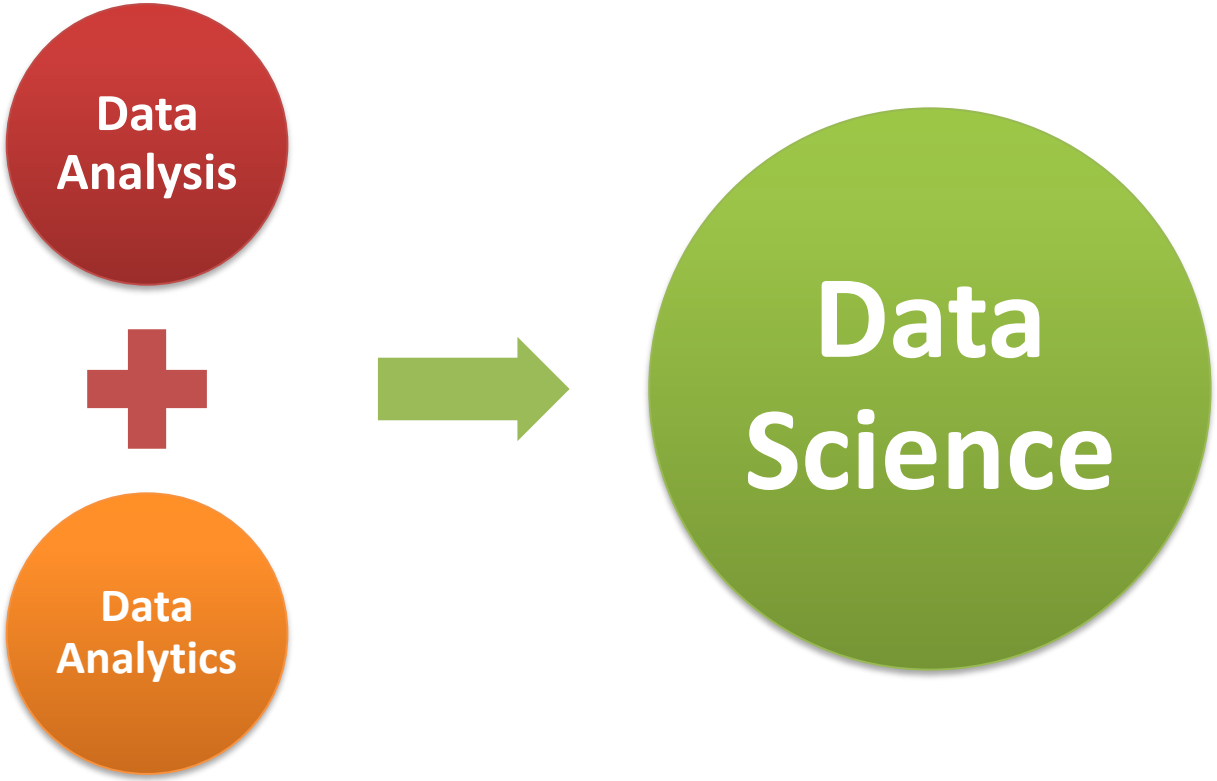


Data Science knowledge domains



Courtesy to Business Broadway

Data Science knowledge Types



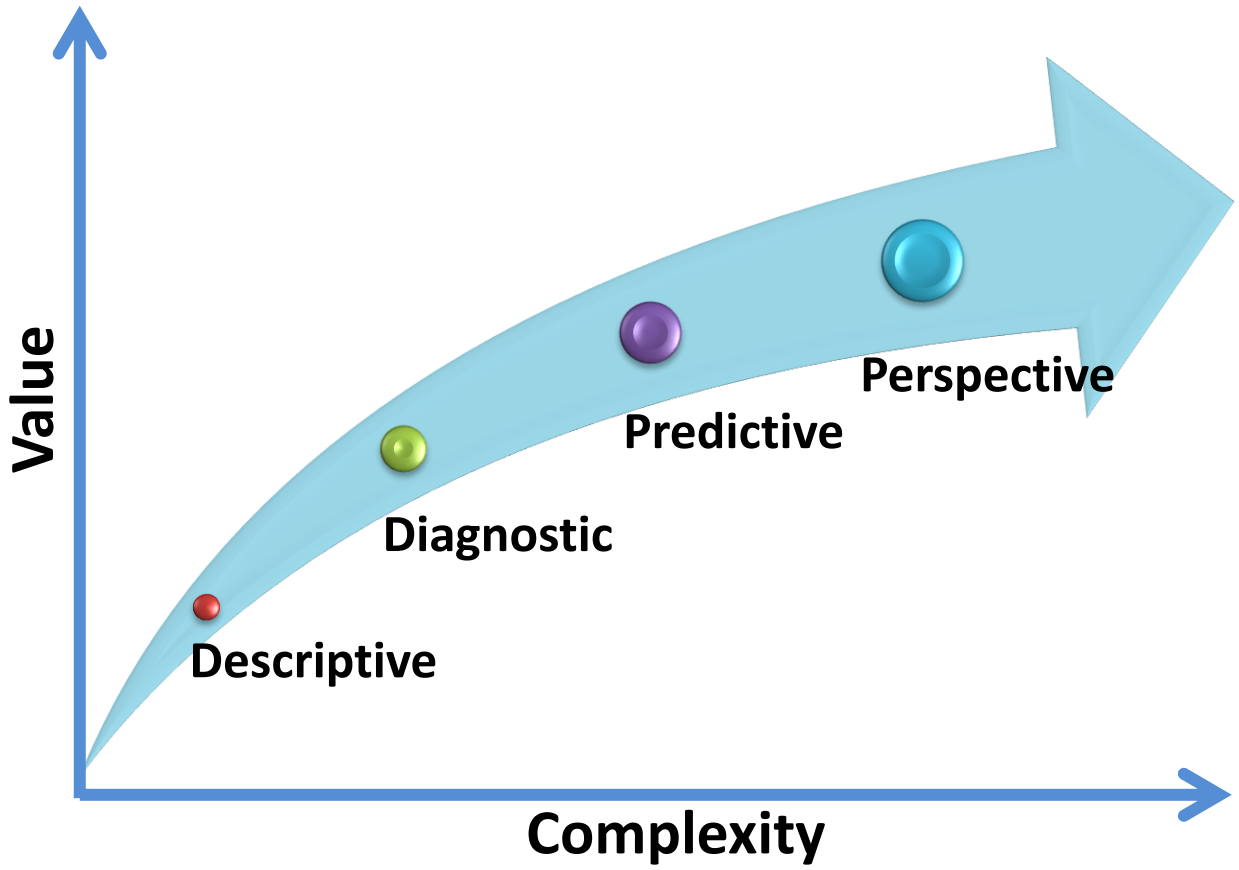
Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, insights, patterns, conclusions, to support decision-making.

Data analytics is the science of processing and analyzing raw data in order to make insights and to support decision making including the current (live / realtime) and past data.

Data Analysis

- Data preparations
- Data Cleaning / Editing
- Statistics: Frequencies, Means, standard deviation, correlation , probabilities , Variances, scaling, standardization, outlier removal
- Visualizations
- Interpretations of trends and patterns

Data Analytics



Descriptive "What's happening" :

- Data understanding & Exploration
- Data visualization

Diagnostic "Why it is happening":

- Dive into the root cause
- Isolate all factors and eliminate noise

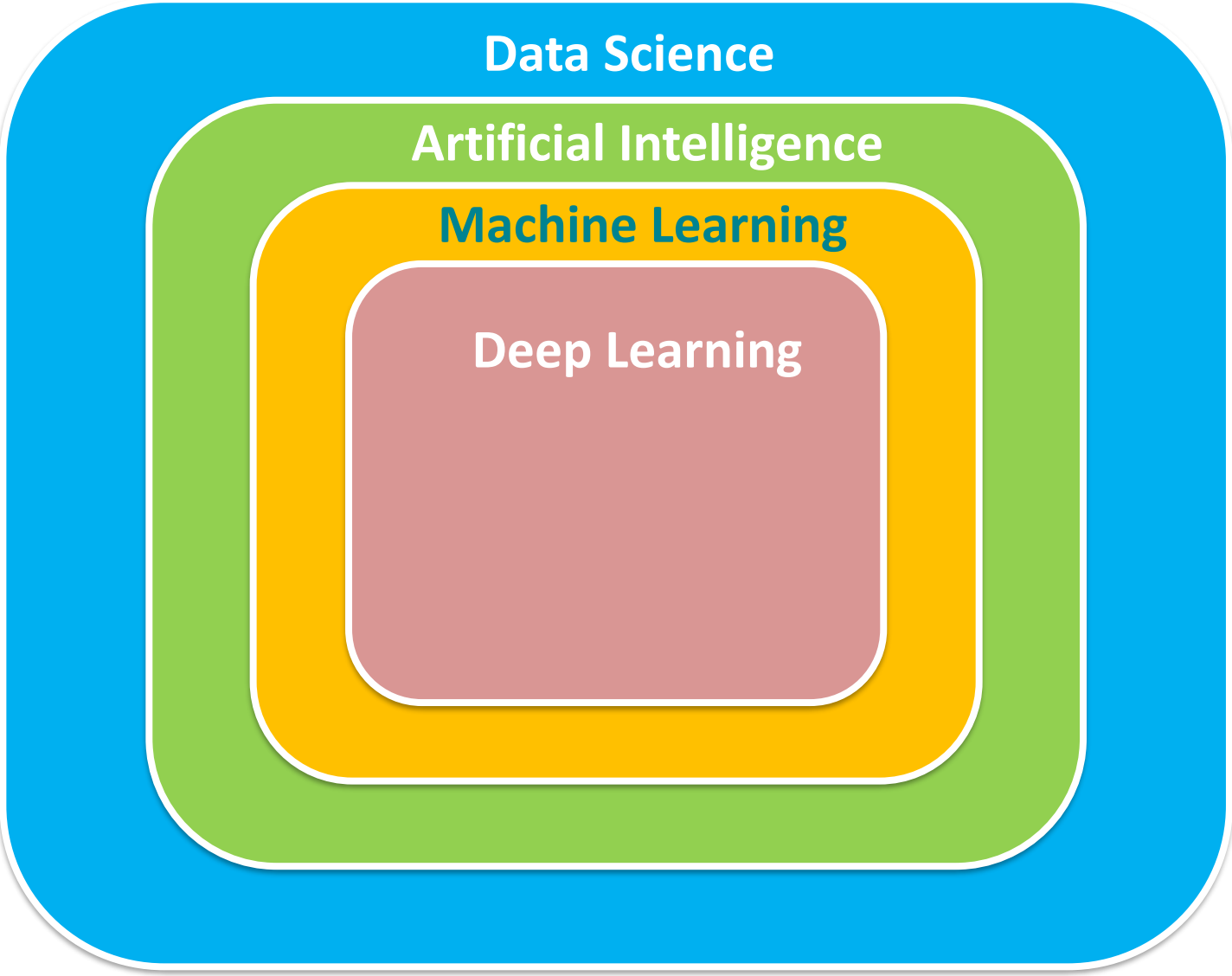
Predictive "What's going to happen":

- Historical patterns are used to predict specific future outcomes using algorithms
- Decisions are automated and updated using algorithms and technology

Perspective " What should I do":

- Recommended actions and strategies based on testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Data Science includes



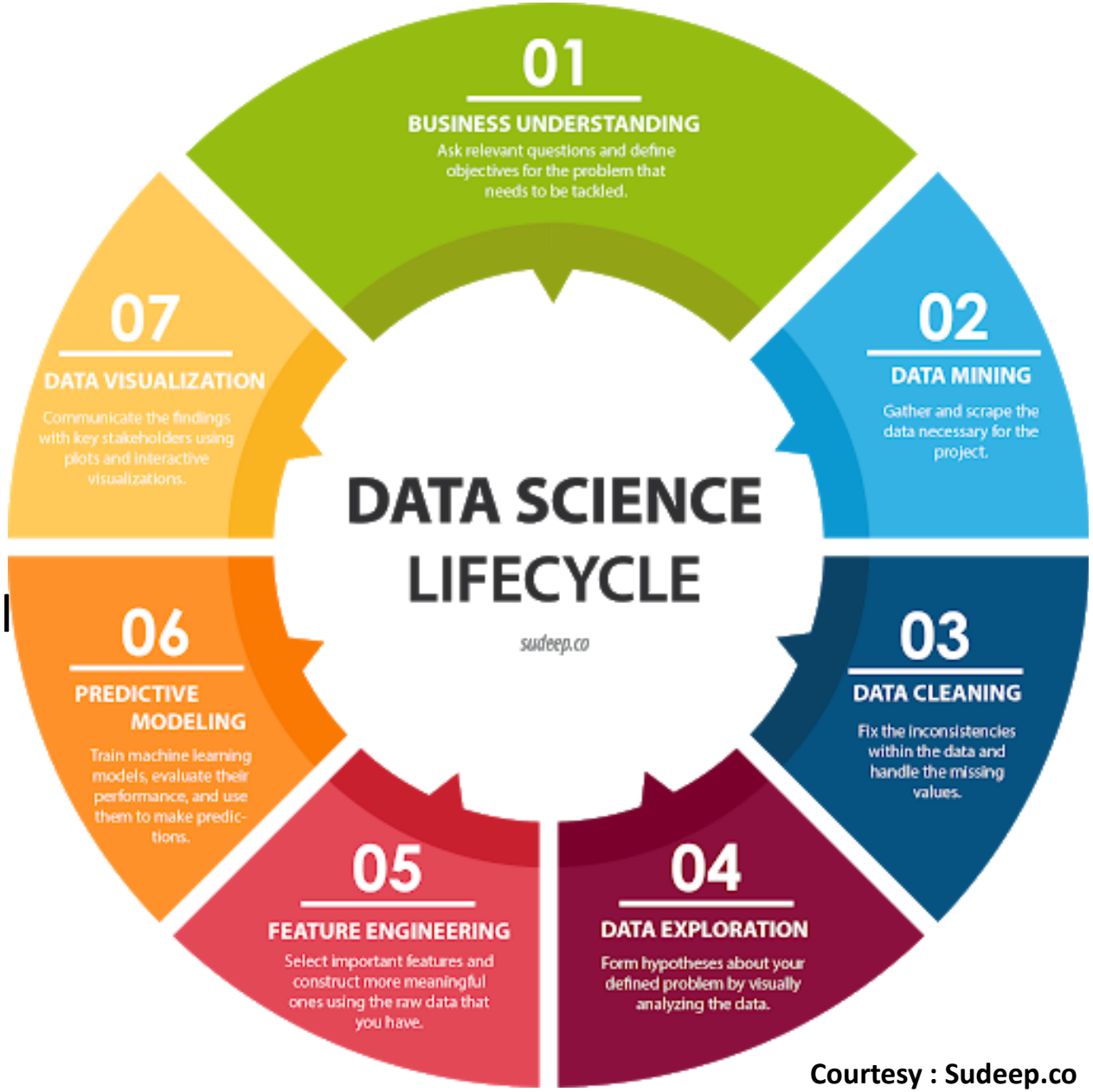
Data Science Life Cycle

80% of the Data scientist time is dedicated to

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

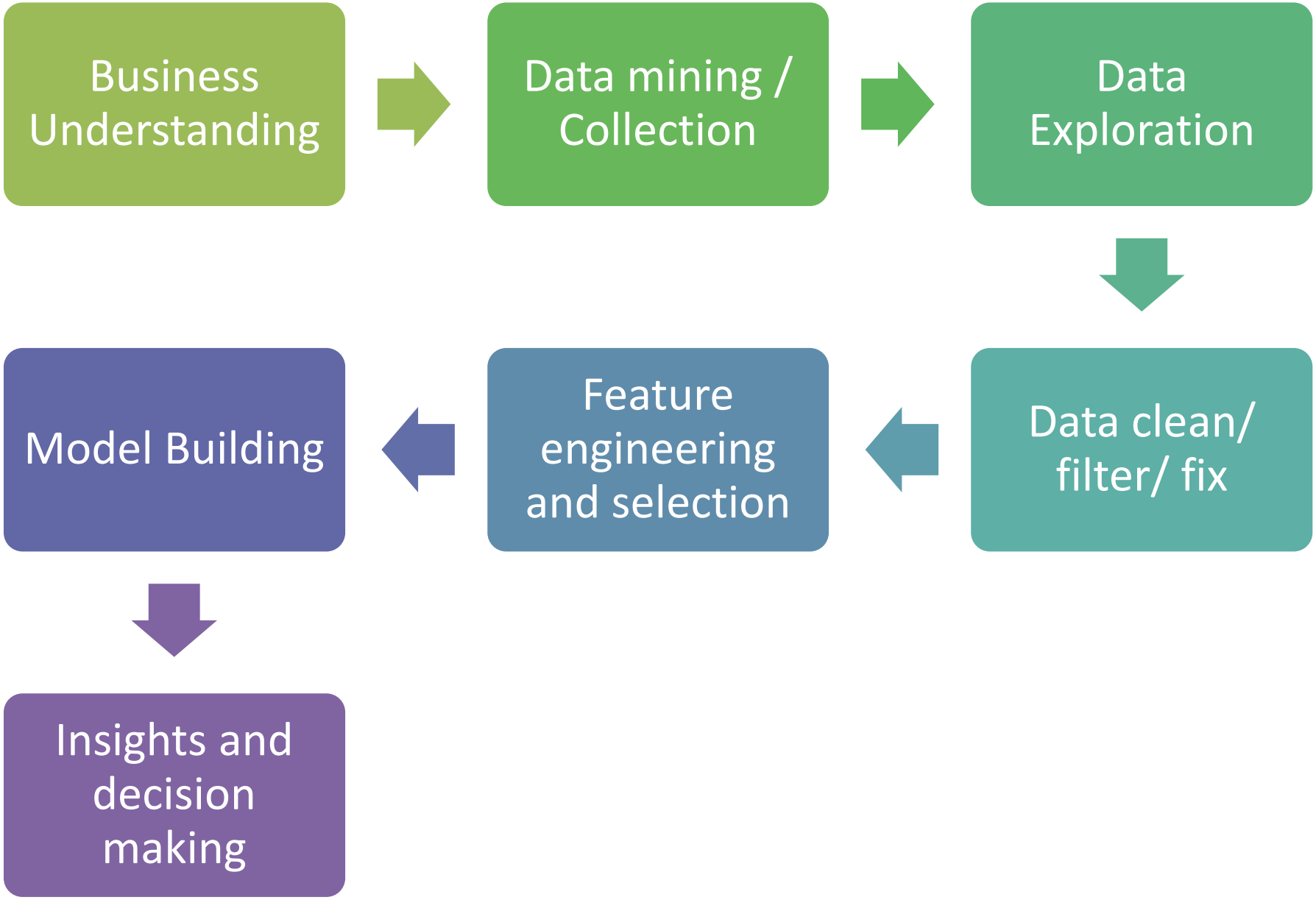
20 % of the data scientist time is for model selection and building

- ✓ Model Building
- ✓ Model Evaluation



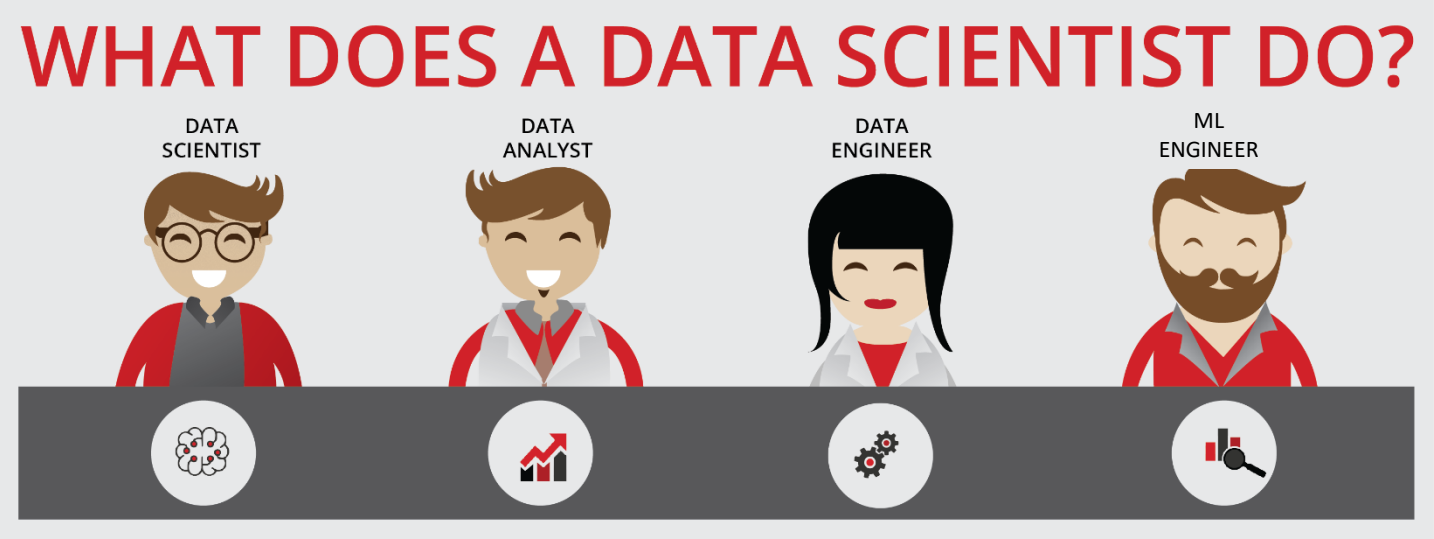
Courtesy : Sudeep.co

Data Science Workflow



Data science Skills

Data science Jobs



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

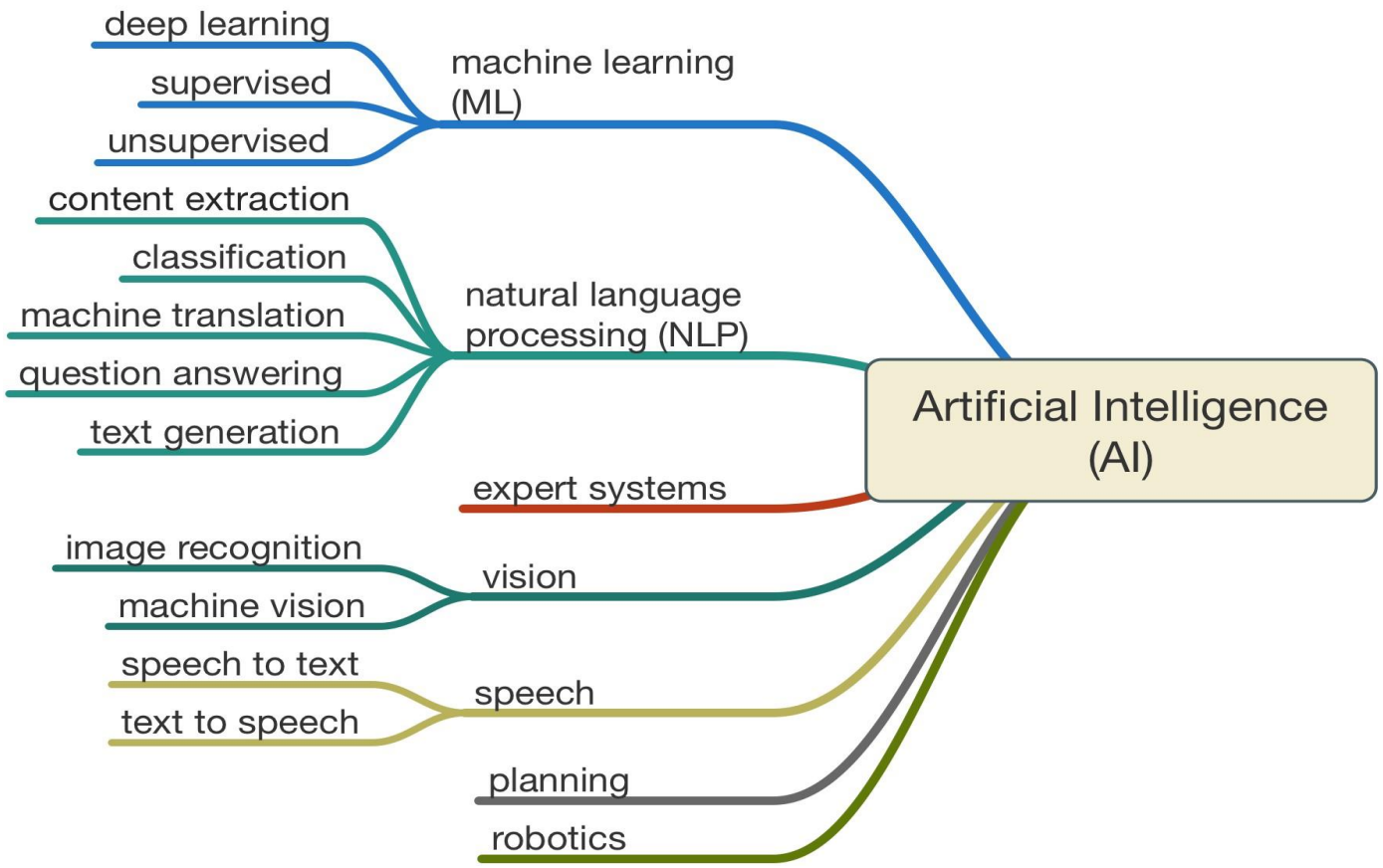
Data Scientist Skills/Experience

- **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

Artificial intelligence vs Machine Learning

Artificial Intelligence

The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The ability to imitate human behavior in thinking and decision making



What's Machine Learning?

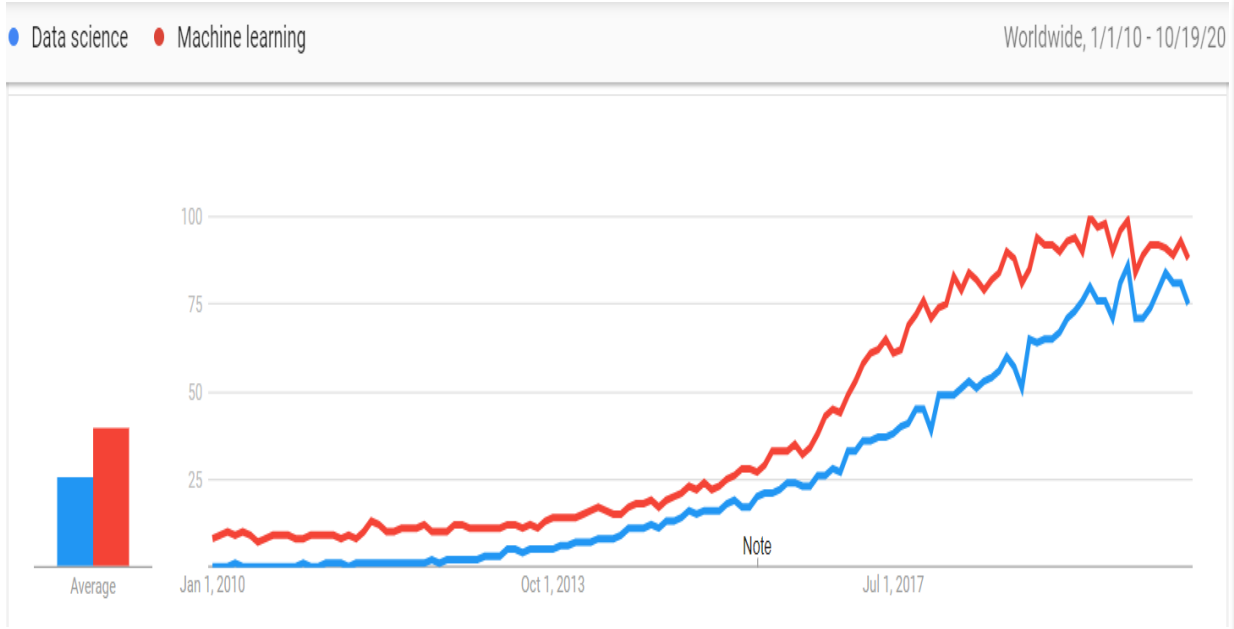
Machine learning is the field of AI that allows systems to learn from past data and make intelligent decisions on their own using algorithms without explicitly programed and improve its experience

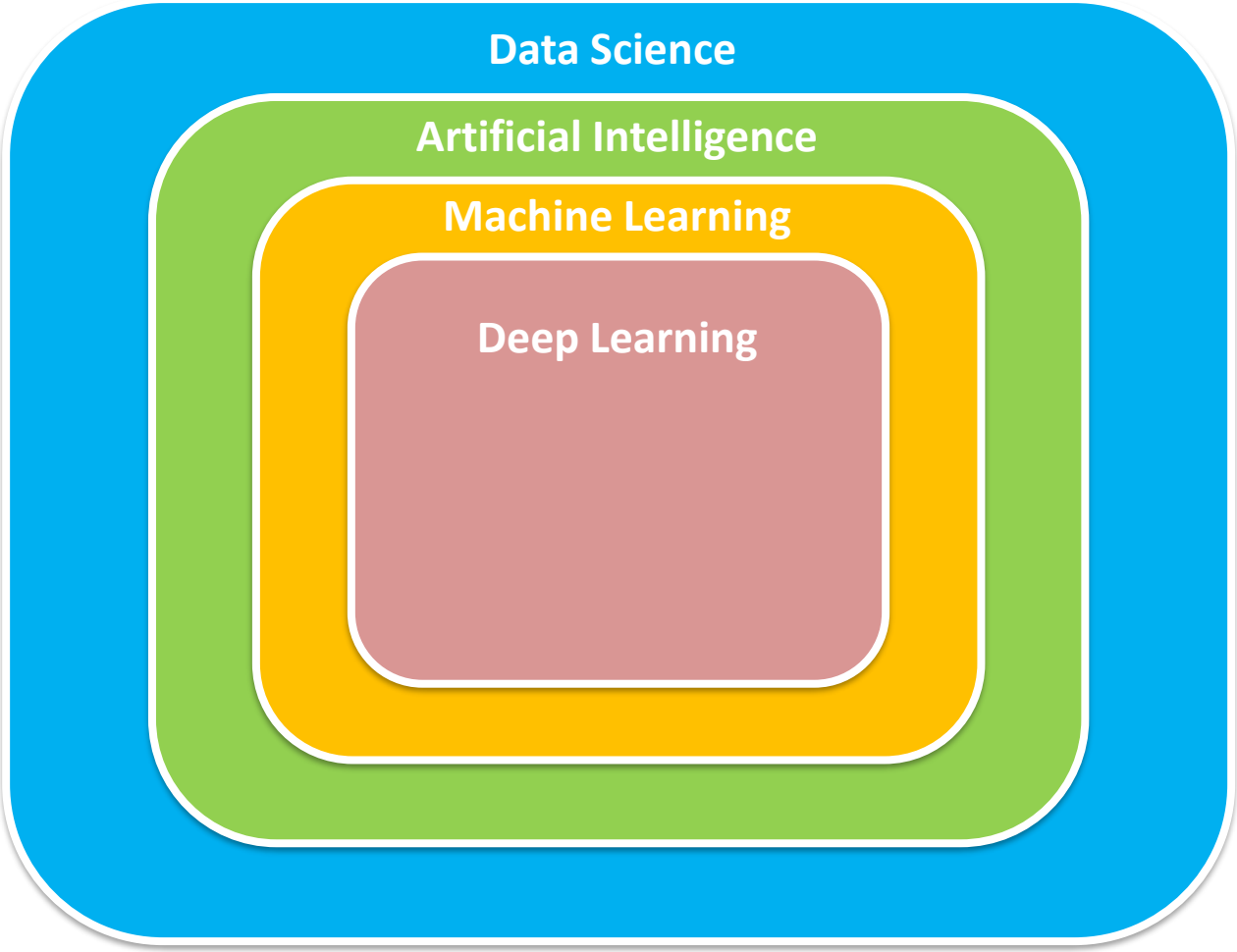
AI vs ML

Artificial Intelligence	Machine learning
Artificial intelligence is a technology which enables a machine to simulate human behavior.	Machine learning is a subset of AI which allows a machine to automatically learn from past data without programming explicitly.
The goal of AI is to make a smart computer system like humans to solve complex problems.	The goal of ML is to allow machines to learn from data so that they can give accurate output.
Machine learning and deep learning are the two main subsets of AI.	Deep learning is a main subset of machine learning.
AI has a very wide range of scope.	Machine learning has a limited scope.
AI is working to create an intelligent system which can perform various complex tasks.	Machine learning is working to create machines that can perform only those specific tasks for which they are trained.
AI system is concerned about maximizing the chances of success.	Machine learning is mainly concerned about accuracy and patterns.
The main applications of AI are Siri, customer support using catboats, Expert System, Online game playing, intelligent humanoid robot, etc.	The main applications of machine learning are Online recommender system, Google search algorithms, Facebook auto friend tagging suggestions, etc.
It includes learning, reasoning, and self-correction.	It includes learning and self-correction when introduced with new data.

Data Science Vs Machine Learning

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

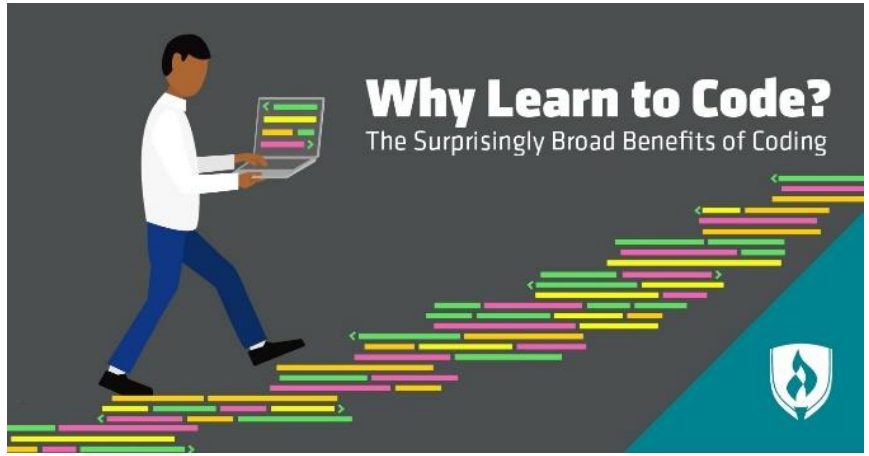
DS vs ML

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

Why learn to code?

Why Learn to Code?

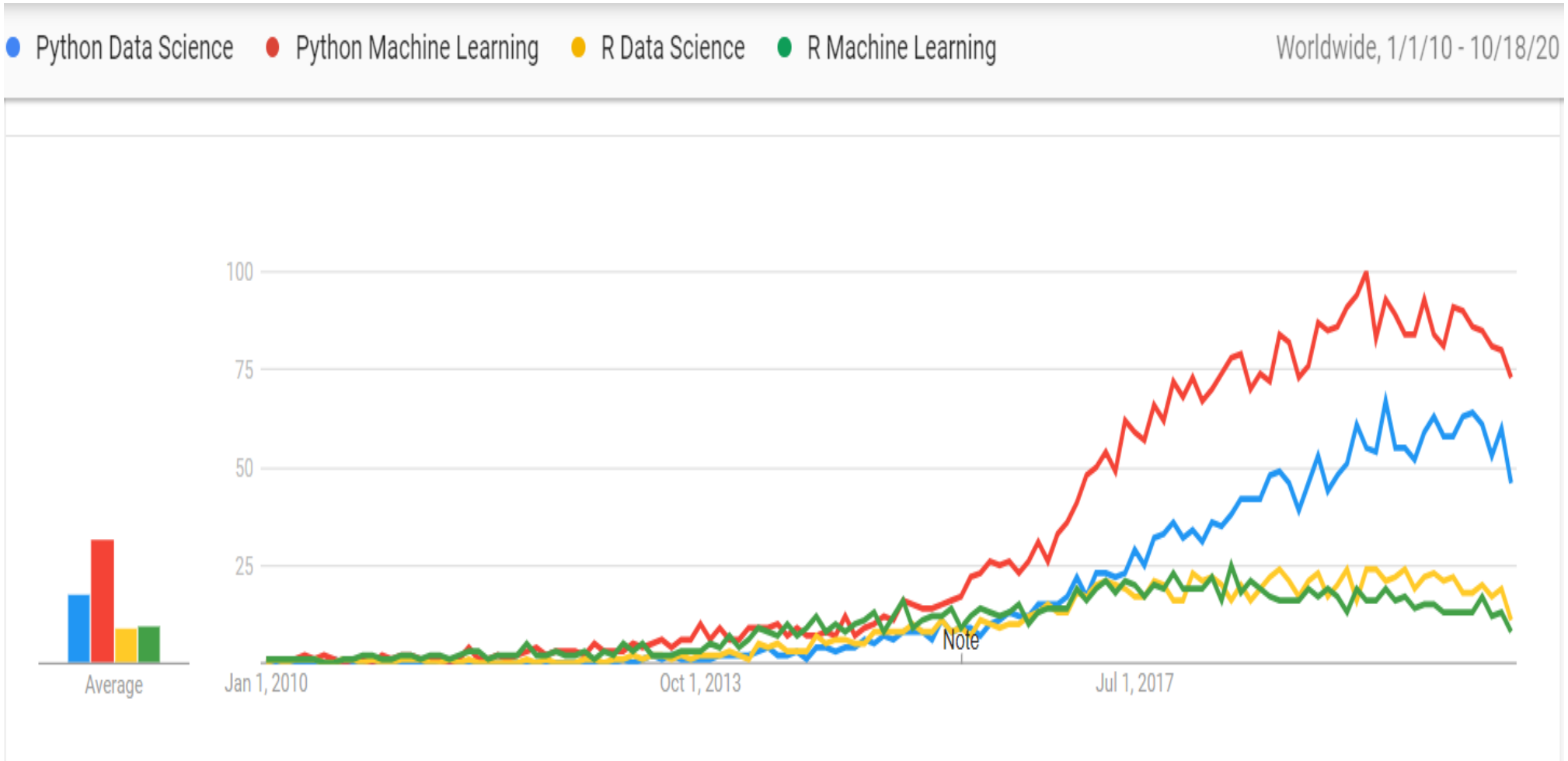
1. Coding is reproducible.
2. Coding fosters creativity.
3. Coding helps learn Math skills and makes sense of it.
4. Coding improves writing academic performance.
5. Coding can lead to software development jobs
6. It open up other job opportunities
7. Coding can make your job application stand out
8. 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

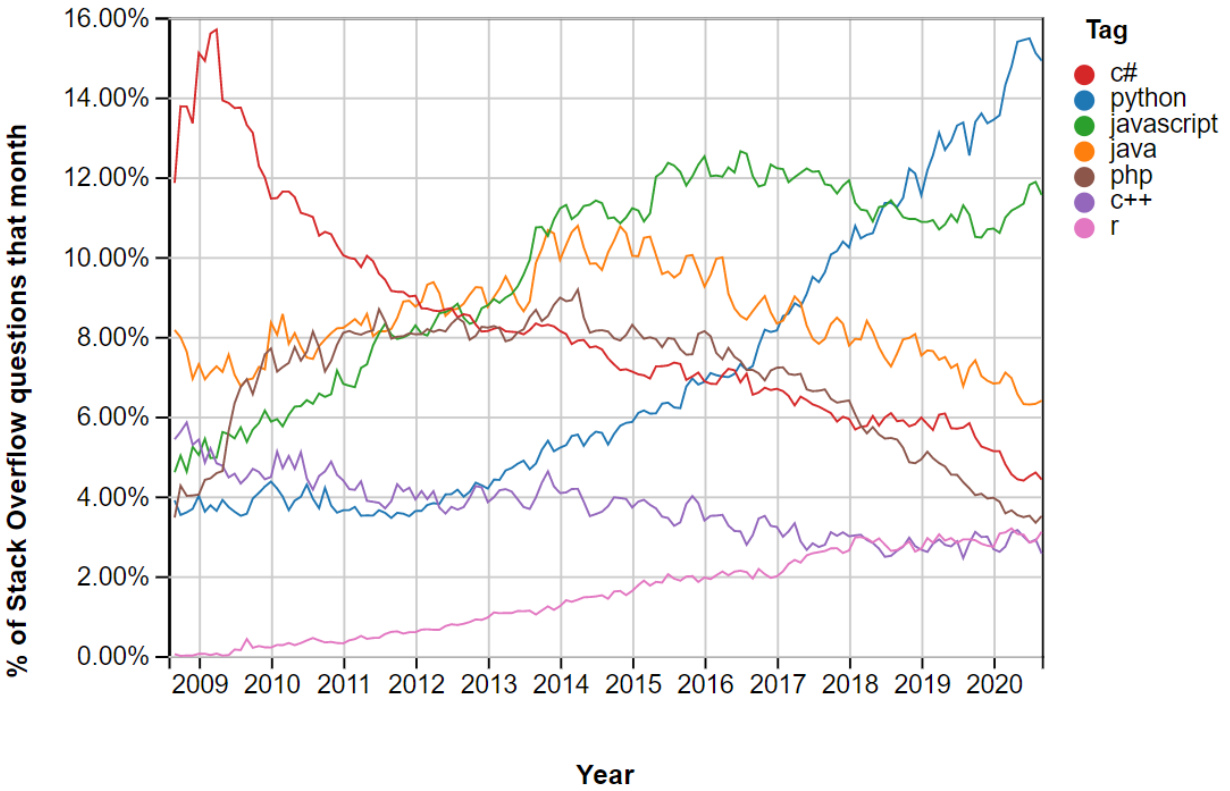
Why should I learn Python?

Why I should Learn Python ?



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

Google NETFLIX facebook Instagram

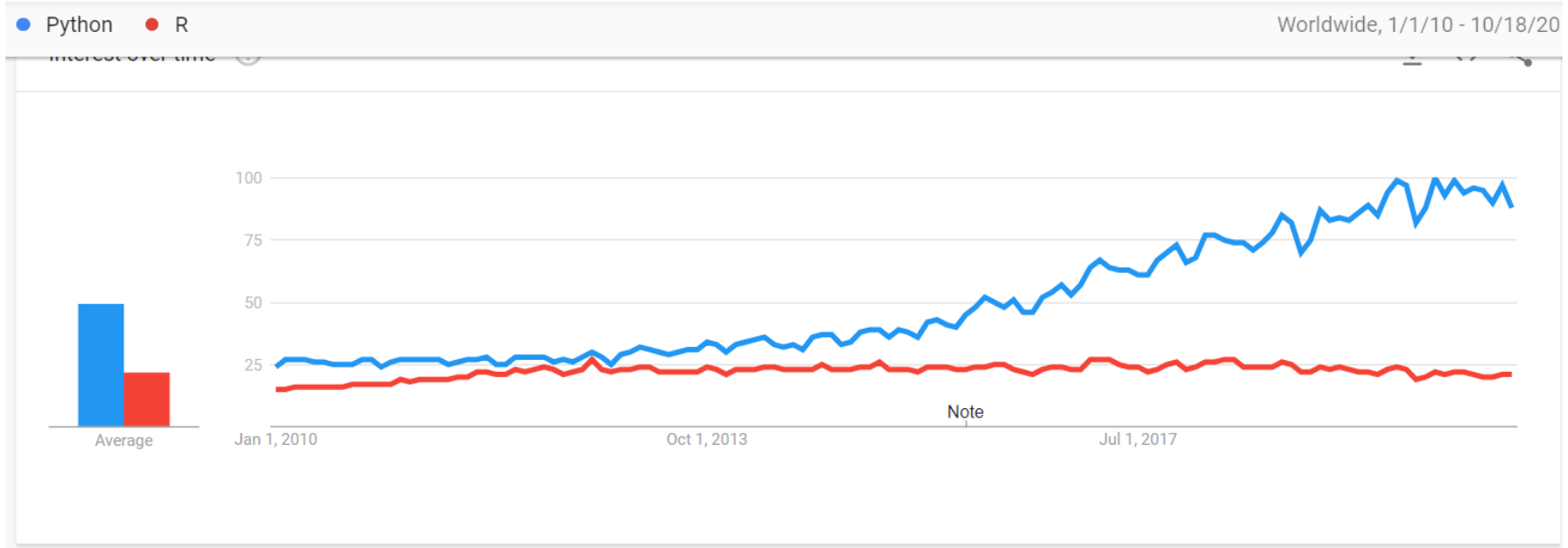
amazon Quora slack intel NASA

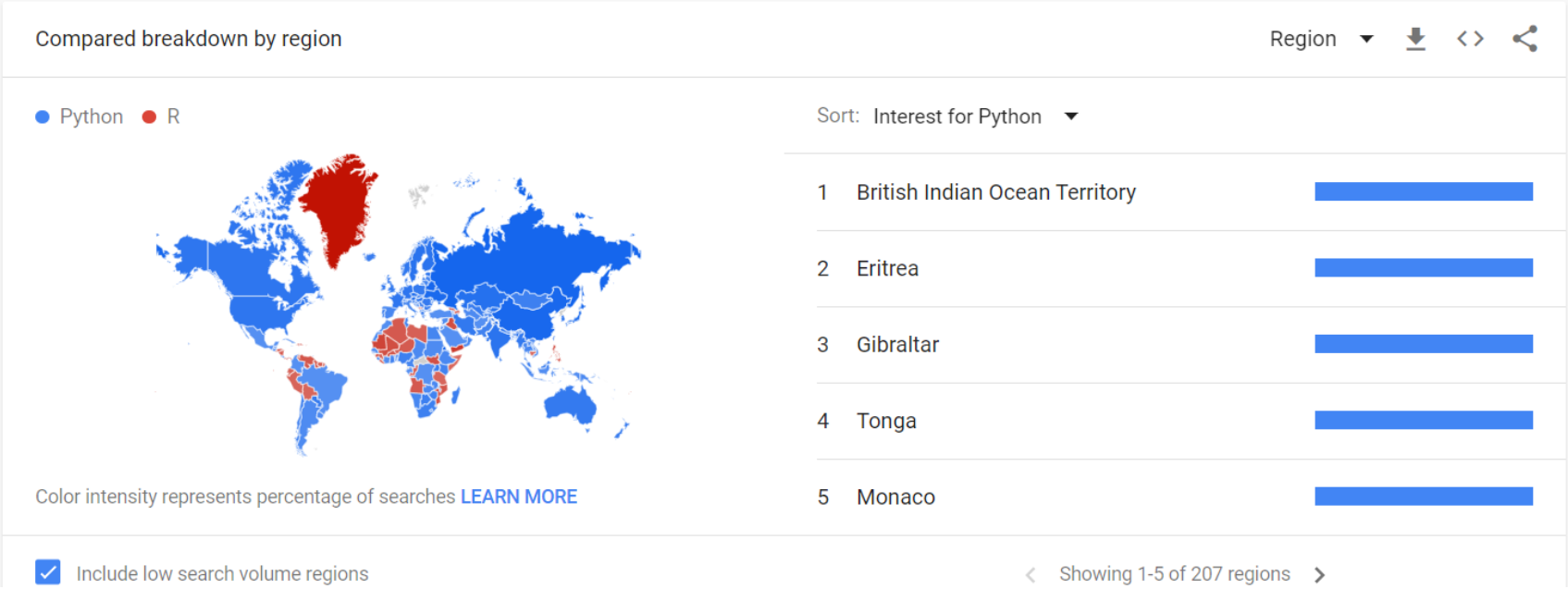
Dropbox ebay Spotify CapitalOne

Norskolje&gass Schlumberger AkerBP ProWellPlan AAPG Agile* Woodside wintershall dea equinor ConocoPhillips EXPLOCROWD cegal

Amr.Moslim

Why I should Learn Python ?





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

DS Applications

Thank You for Your Attention