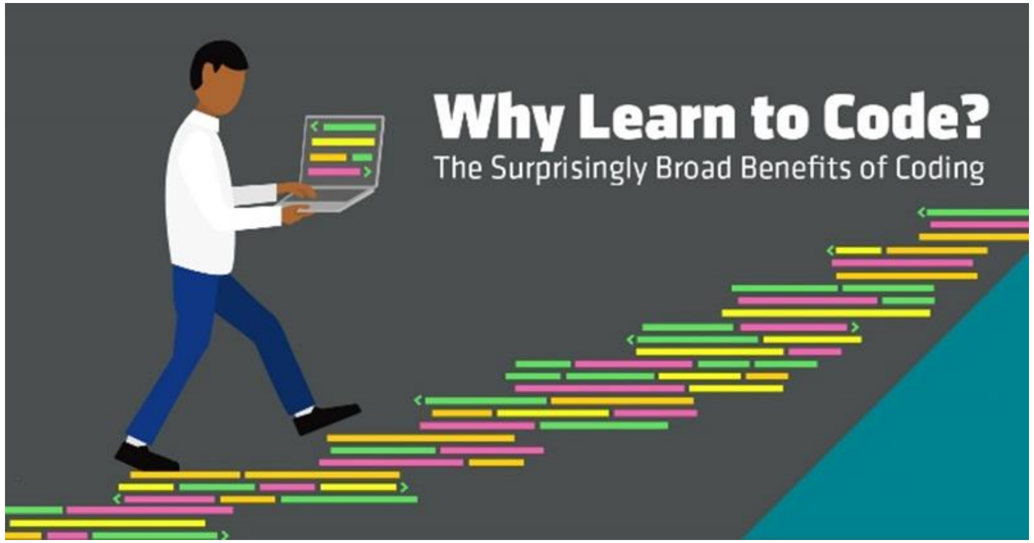
Introduction to Data Science for Geoscientists

Al-Amal Program

Amr. Moslim



Today's Agenda

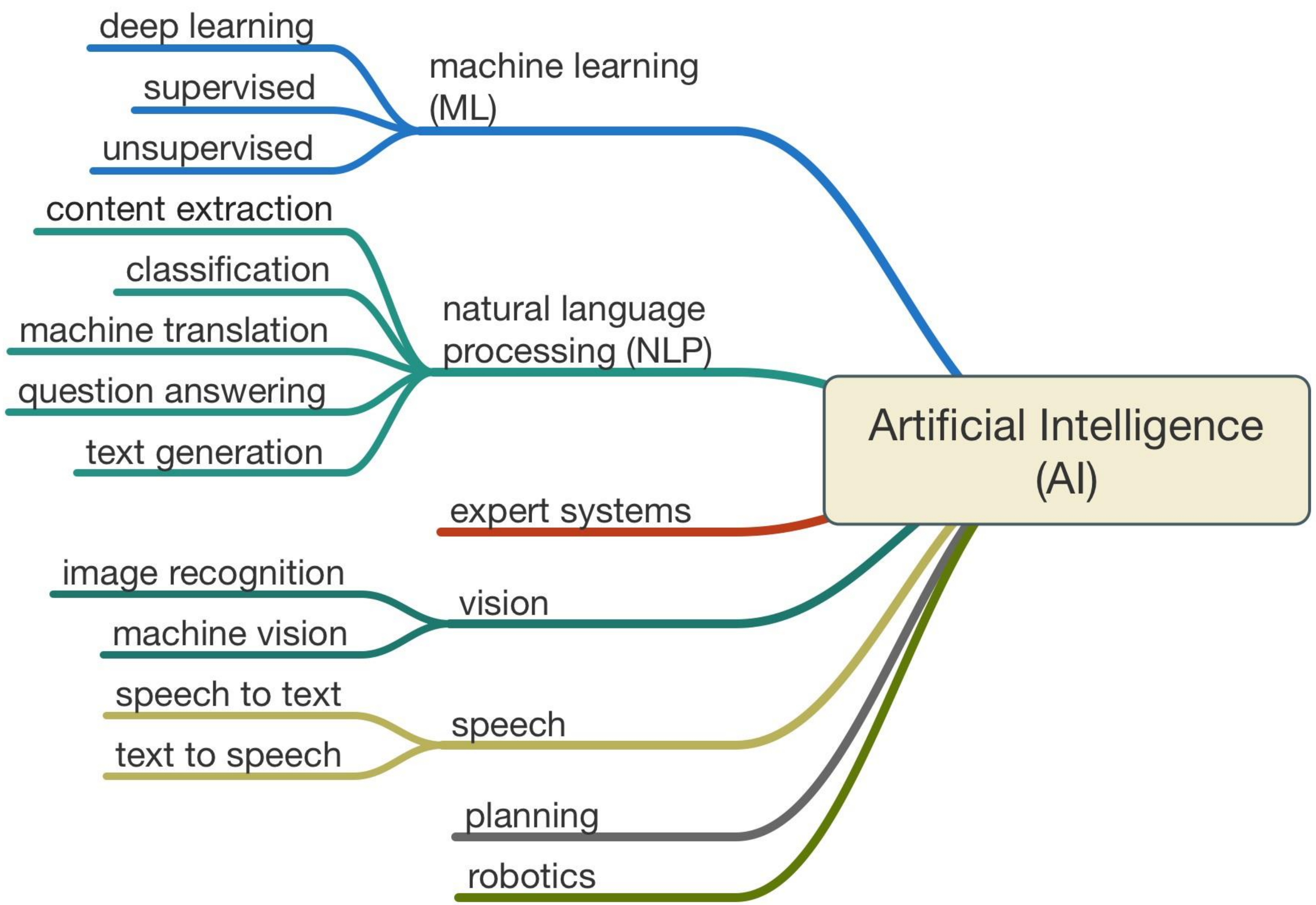
- What is Data science?
- What is Machine learning?
- DS VS ML
- DS Skills
- Why should I learn Python?
- How to Code?
- Machine Learning technique
- How does ML work?
- ML Models Evaluation
- ML Algorithms (Kmeans – KNN - Random Forest)

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.
In AI, we make intelligent systems to perform any task like a human.	In ML, we teach machines with data to perform a particular task and give an accurate result.
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.
On the basis of capabilities, AI can be divided into three types, which are, Weak AI, General AI, and Strong AI.	Machine learning can also be divided into mainly three types that are Supervised learning, Unsupervised learning, and Reinforcement learning.
It includes learning, reasoning, and self-correction.	It includes learning and self-correction when introduced with new data.
AI completely deals with Structured, semi-structured, and unstructured data.	Machine learning deals with Structured and semi-structured data.

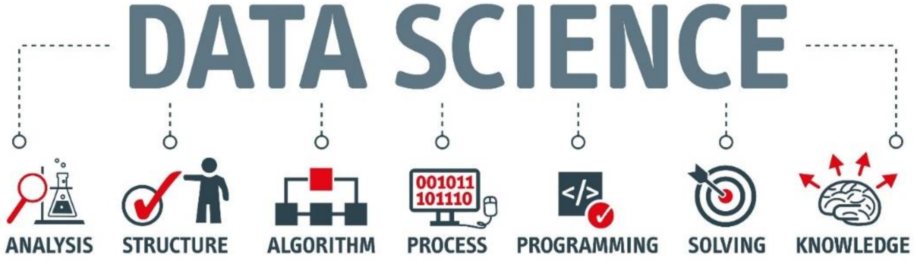
Artificial Intelligence

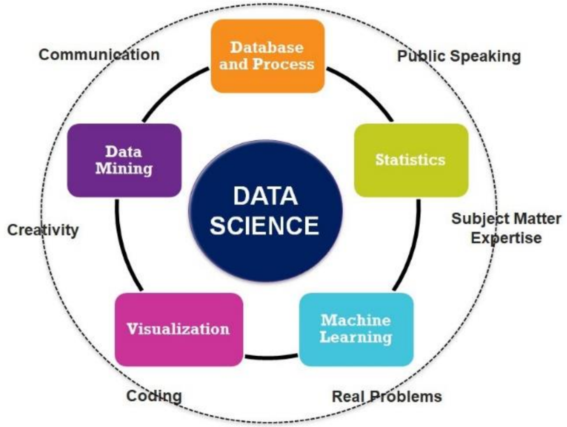
The ability of a digital **computer** or computer-controlled **robot** to perform tasks commonly associated with intelligent beings.



What's Data Science ?

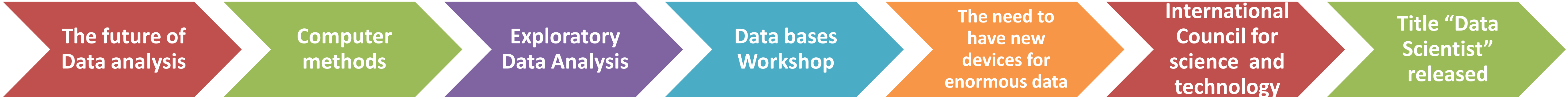
Data science is the field of study that uses modern tools and techniques to process, clean, analyze, model and visualize large data sets to get insights that are reliable to help organizations to understand certain criteria or condition and make business decisions





Amr.Moslim

Data Science History



1962

The future of Data analysis

John W. Tueky released a paper about the future and importance of Data Analysis

1974

Computer methods

Peter Naur used term Data Science for the first time

1977

Exploratory Data Analysis

John W. Tueky Introduced the term of EDA

1989

Data bases Workshop

George Shapiro Introduced a workshop about data bases

1999

The need to have new devices for enormous data

Jacob Zahavi Introduced the need to power devices for the enormous data "Big Data"

2001

International Council for science and technology

Committee on Data for science and Technology

2008

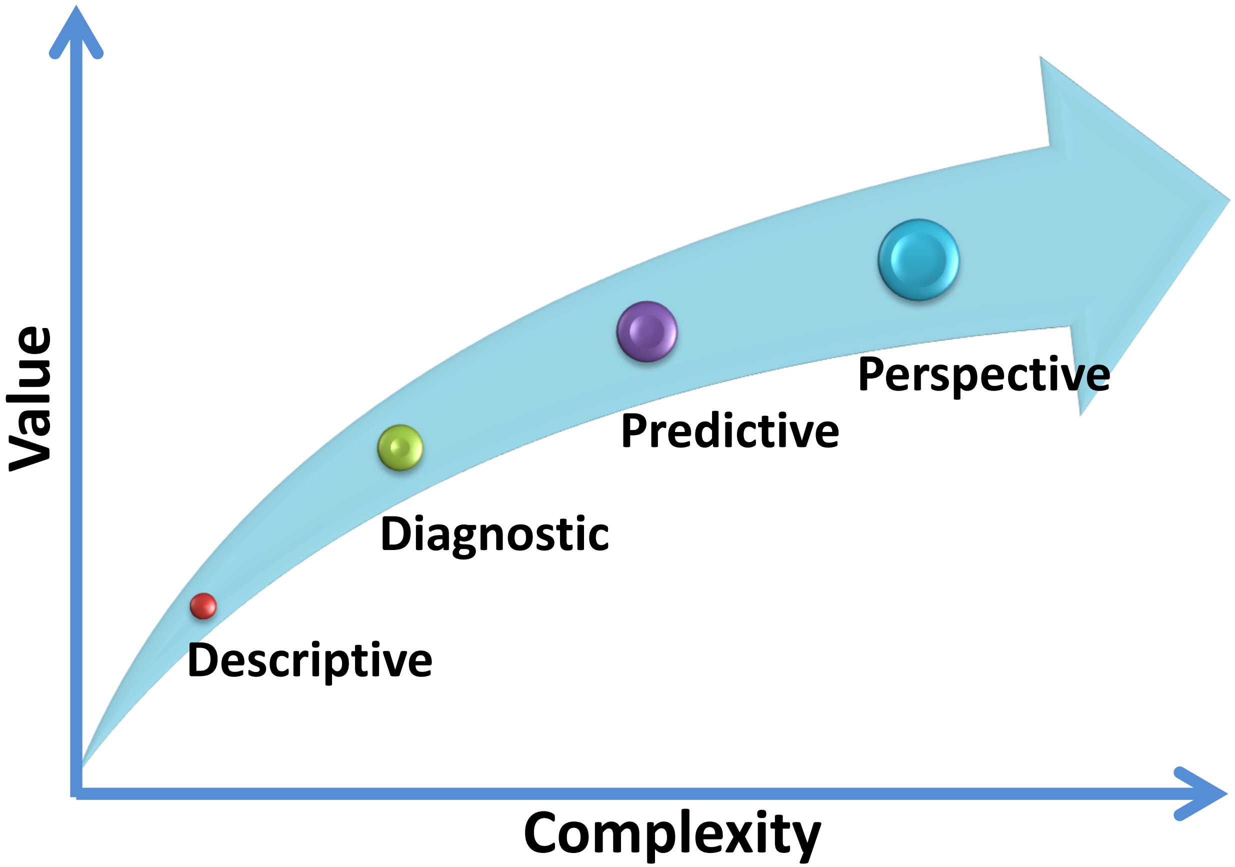
Title "Data Scientist" released

A trendy expression for who works in data analysis and analytics

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

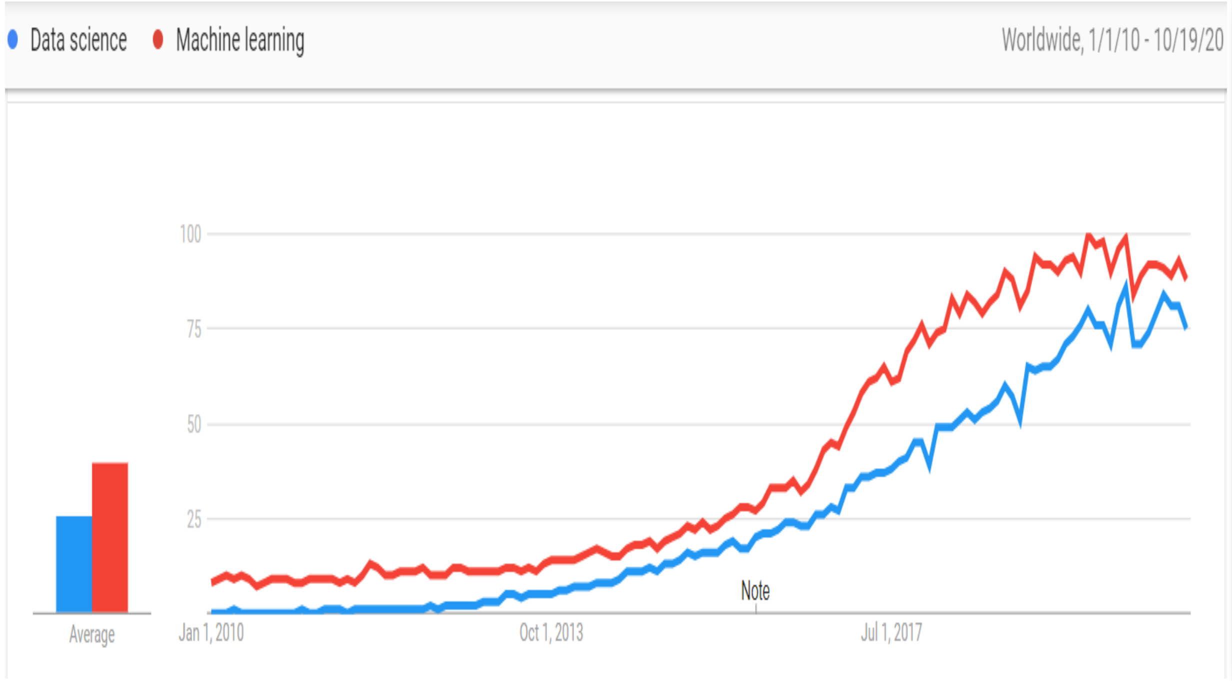
Perspective " What should I do":

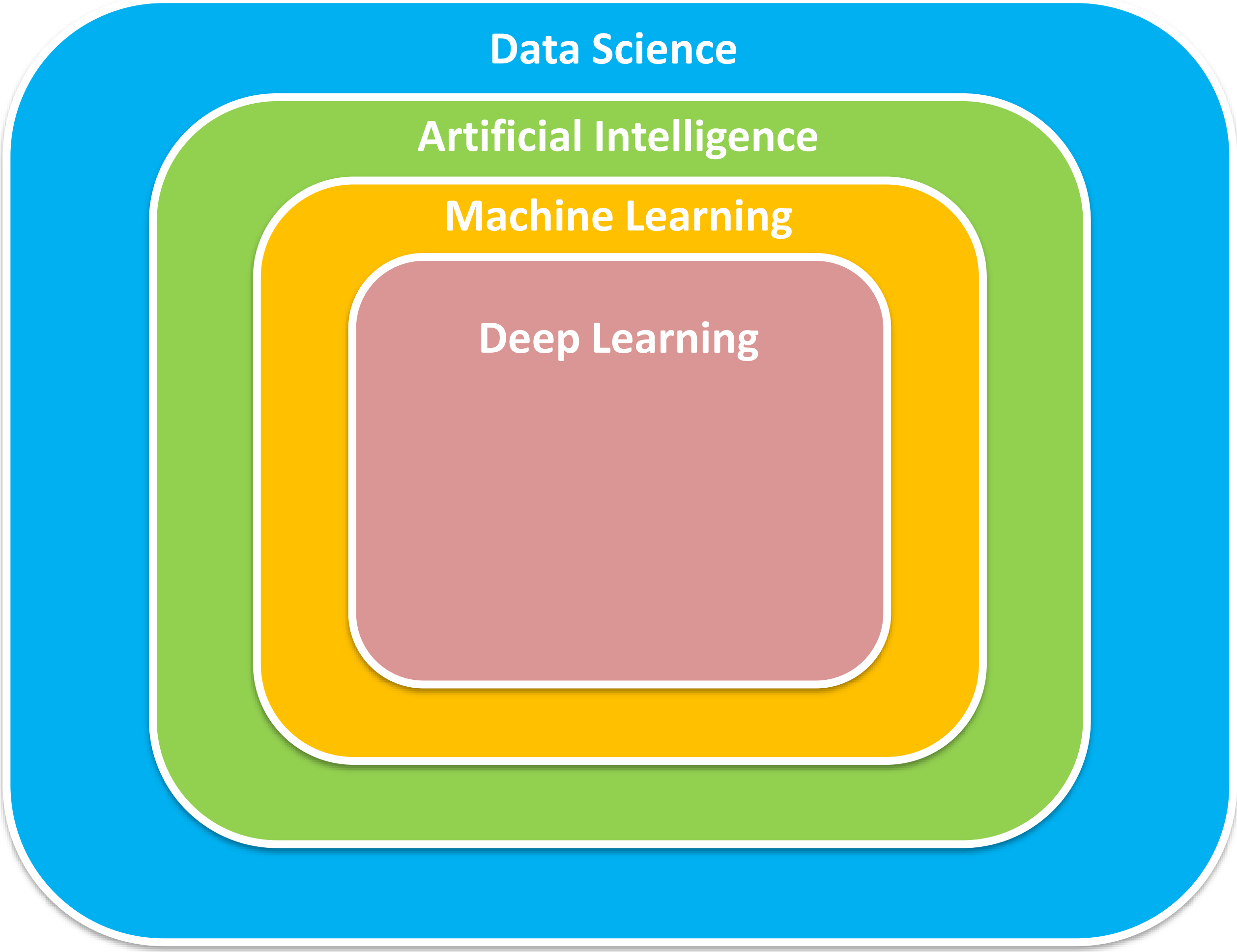
- Recommended actions and strategies based on testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

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?

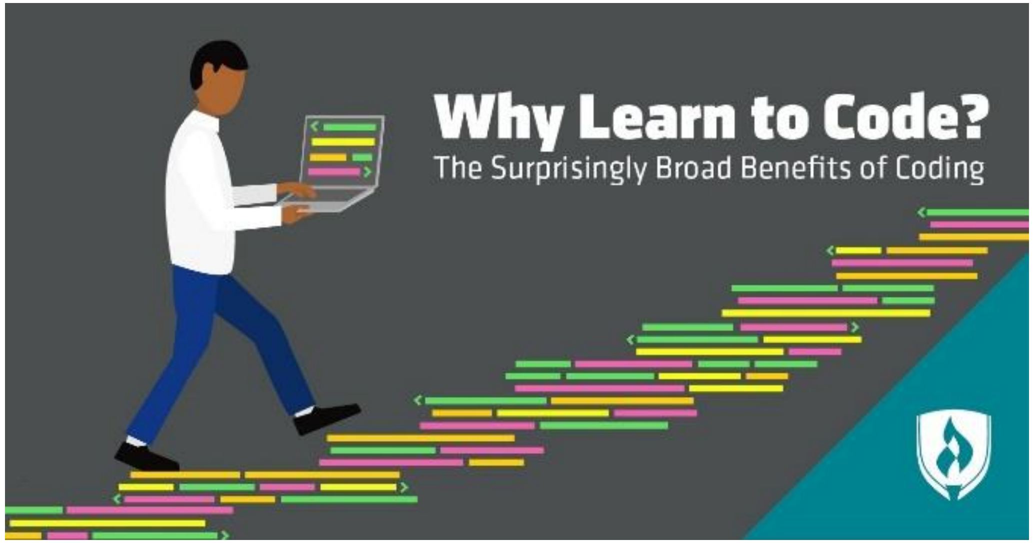




Amr.Moslim

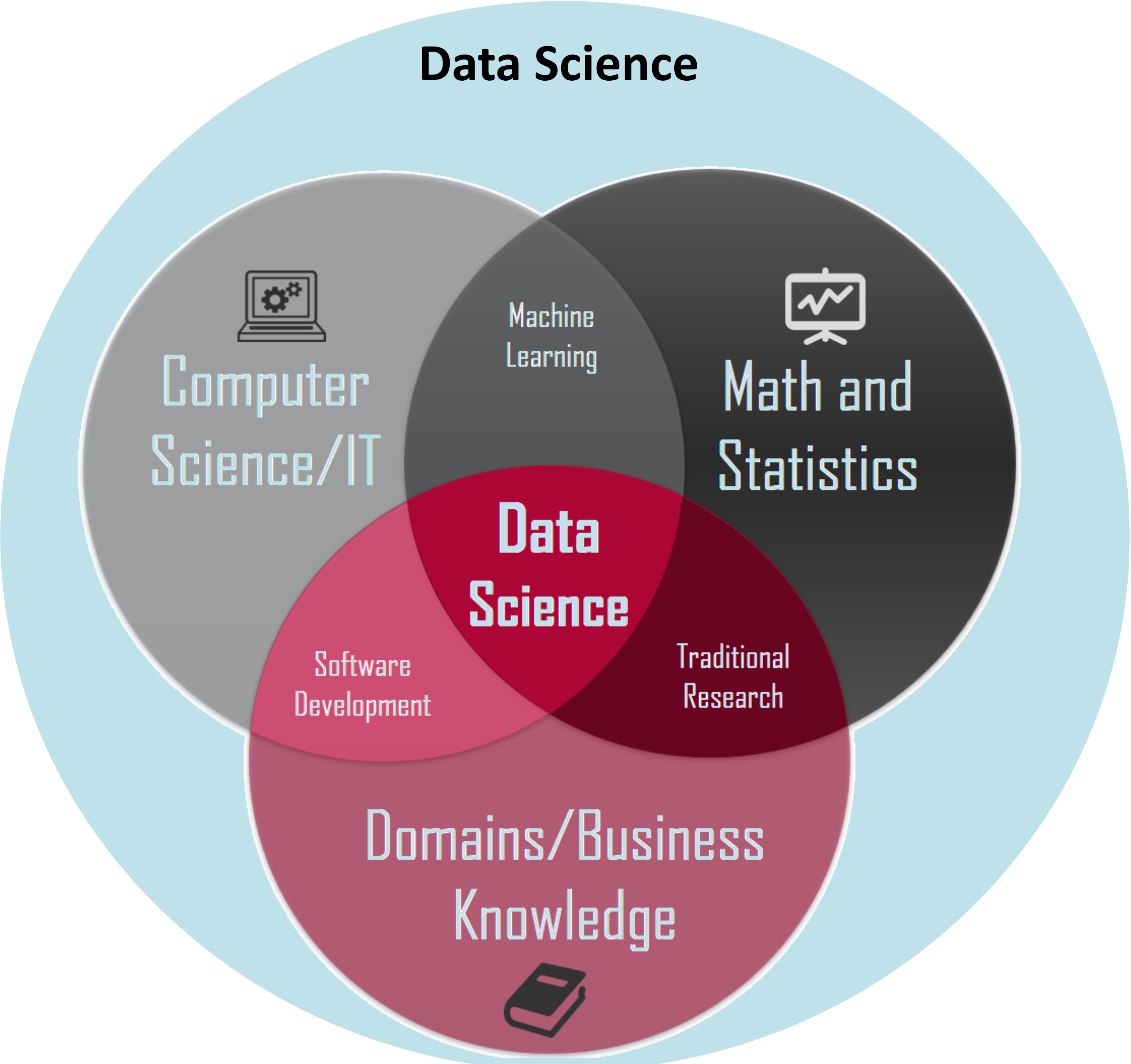
Why Learn to Code?

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

Data Science knowledge domains



Amr.Moslim

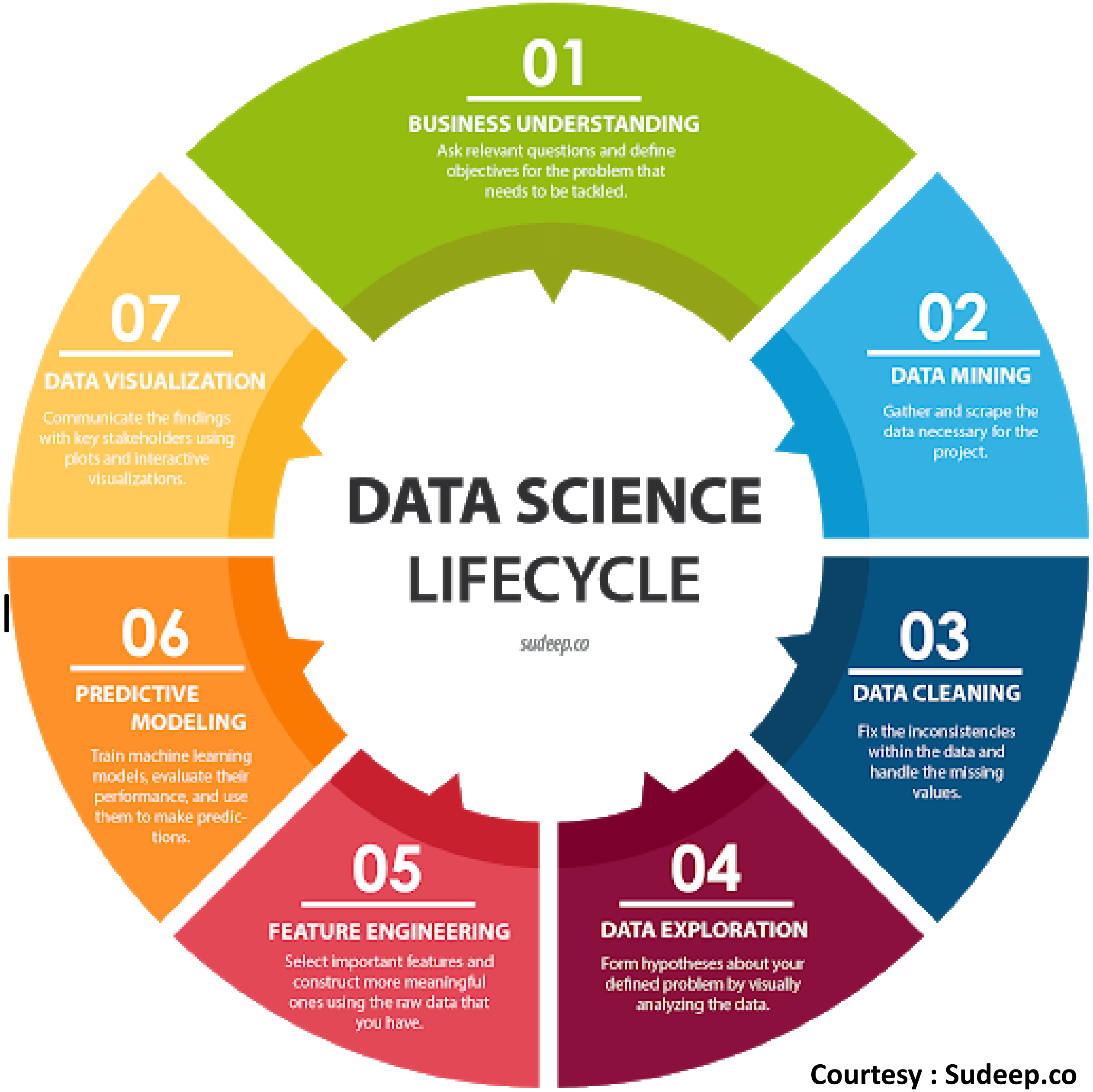
Data Science Life Cycle

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

- ✓ Model Building
- ✓ Model Evaluation



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.

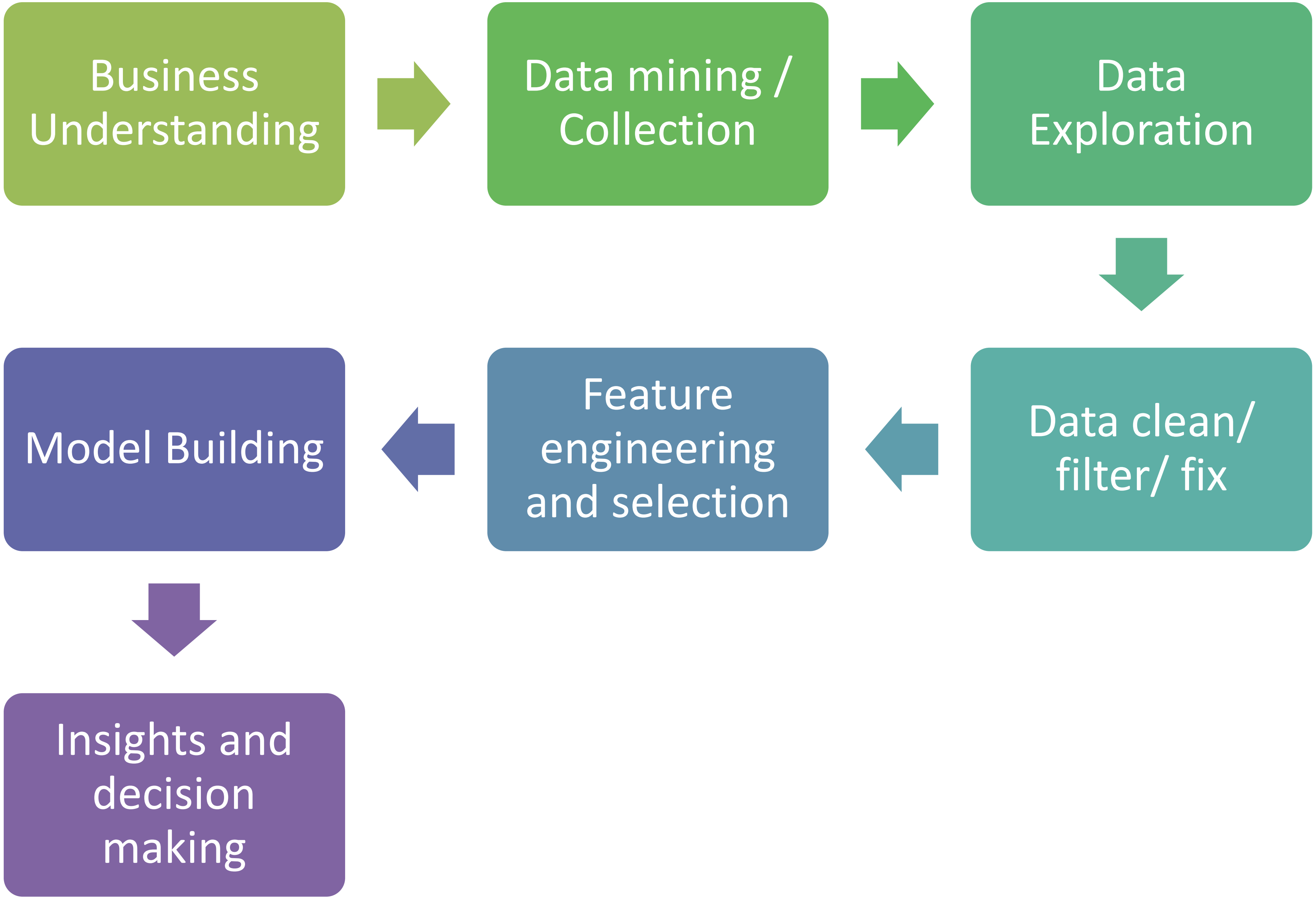
07 DATA VISUALIZATION

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

Amr.Moslim

Courtesy : Sudeep.co

Data Science Workflow



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

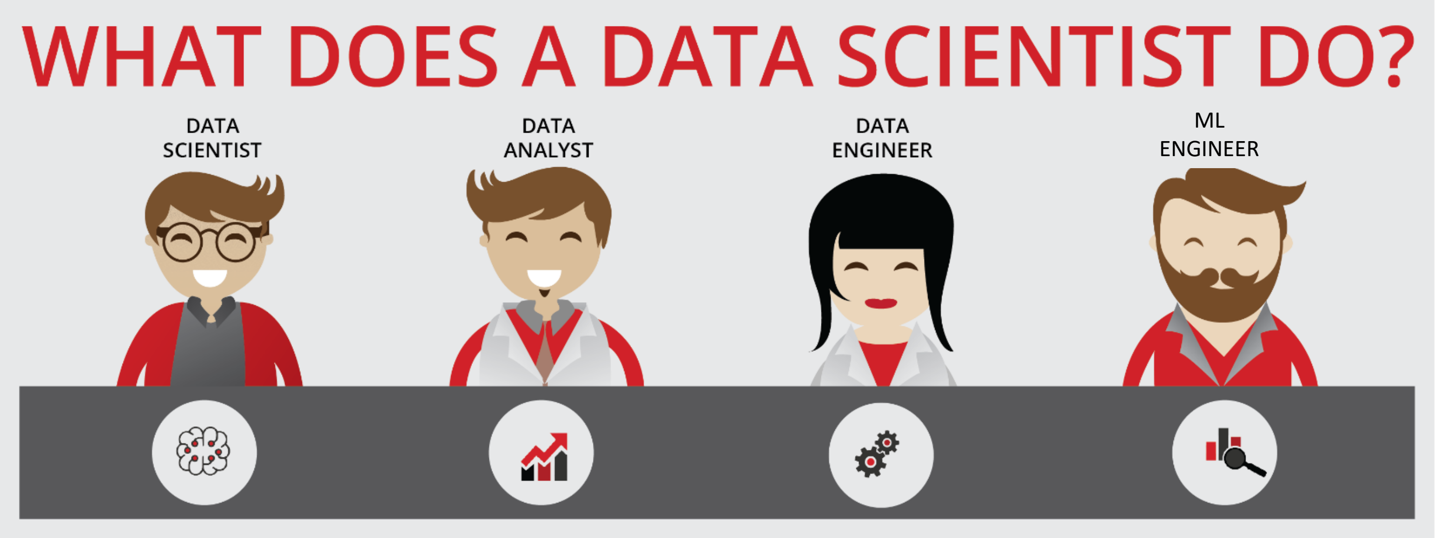
Amr.Moslim

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

Data science Skills

Data science Jobs



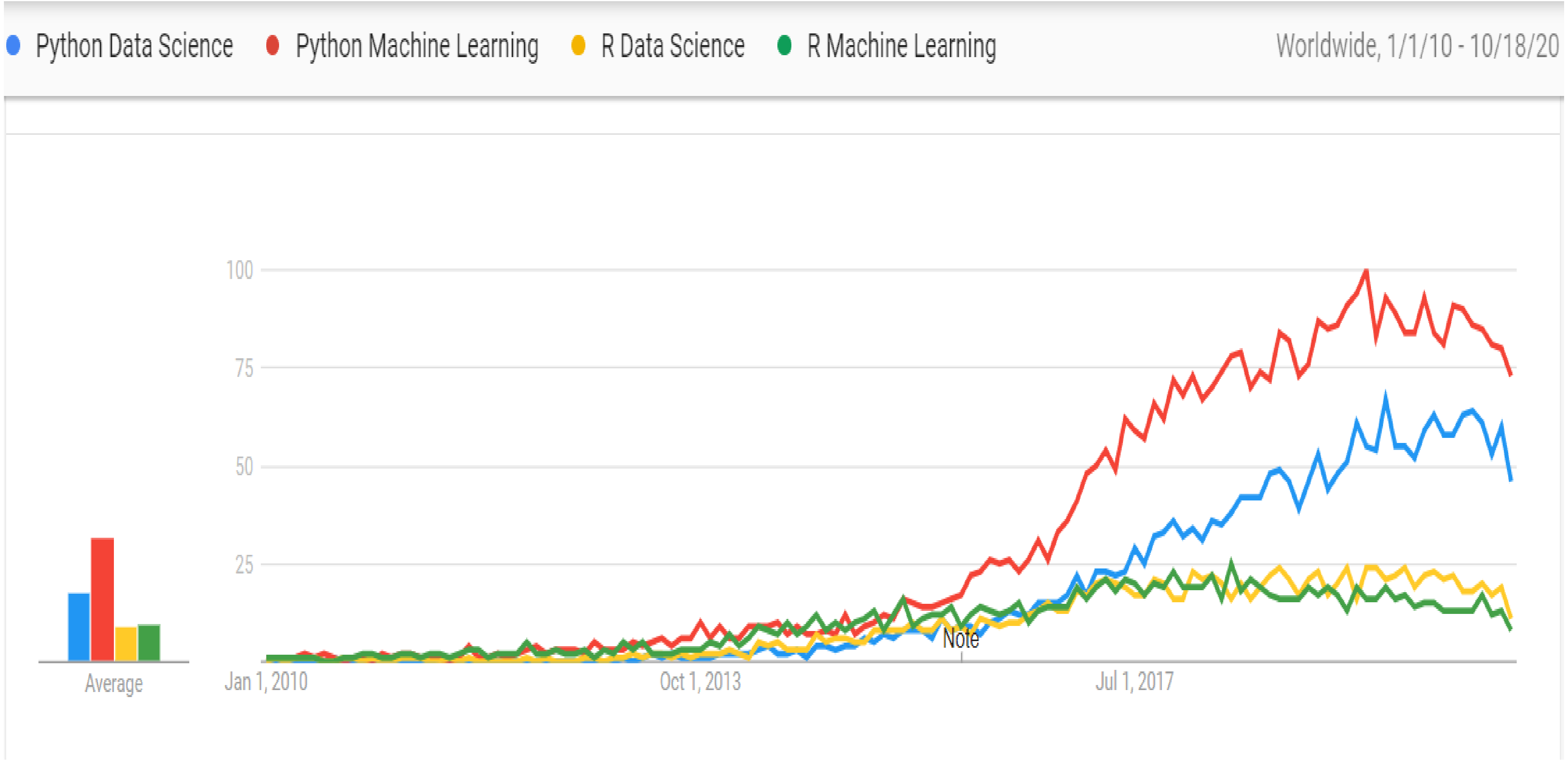
- DATA ANALYST
- DATA ENGINEER
- MACHINE LEARNING ENGINEER
- DATA SCIENCE GENERALIST

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

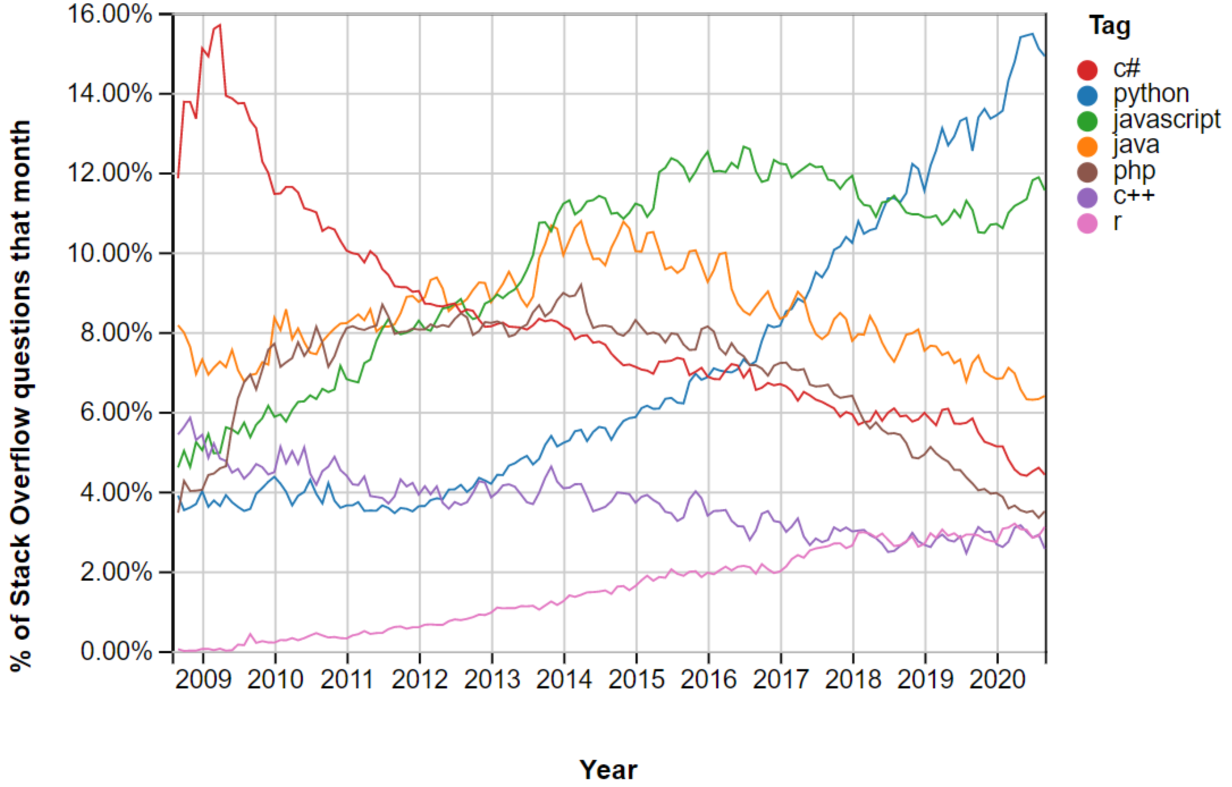
Why should I learn Python?

Why I should Learn Python ?



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



Norsk olje&gass Schlumberger AkerBP ProWellPlan AAPG Advancing the World of Petroleum Geosciences Agile* Woodside wintershall dea equinor ConocoPhillips EXPLO CROWD cegal

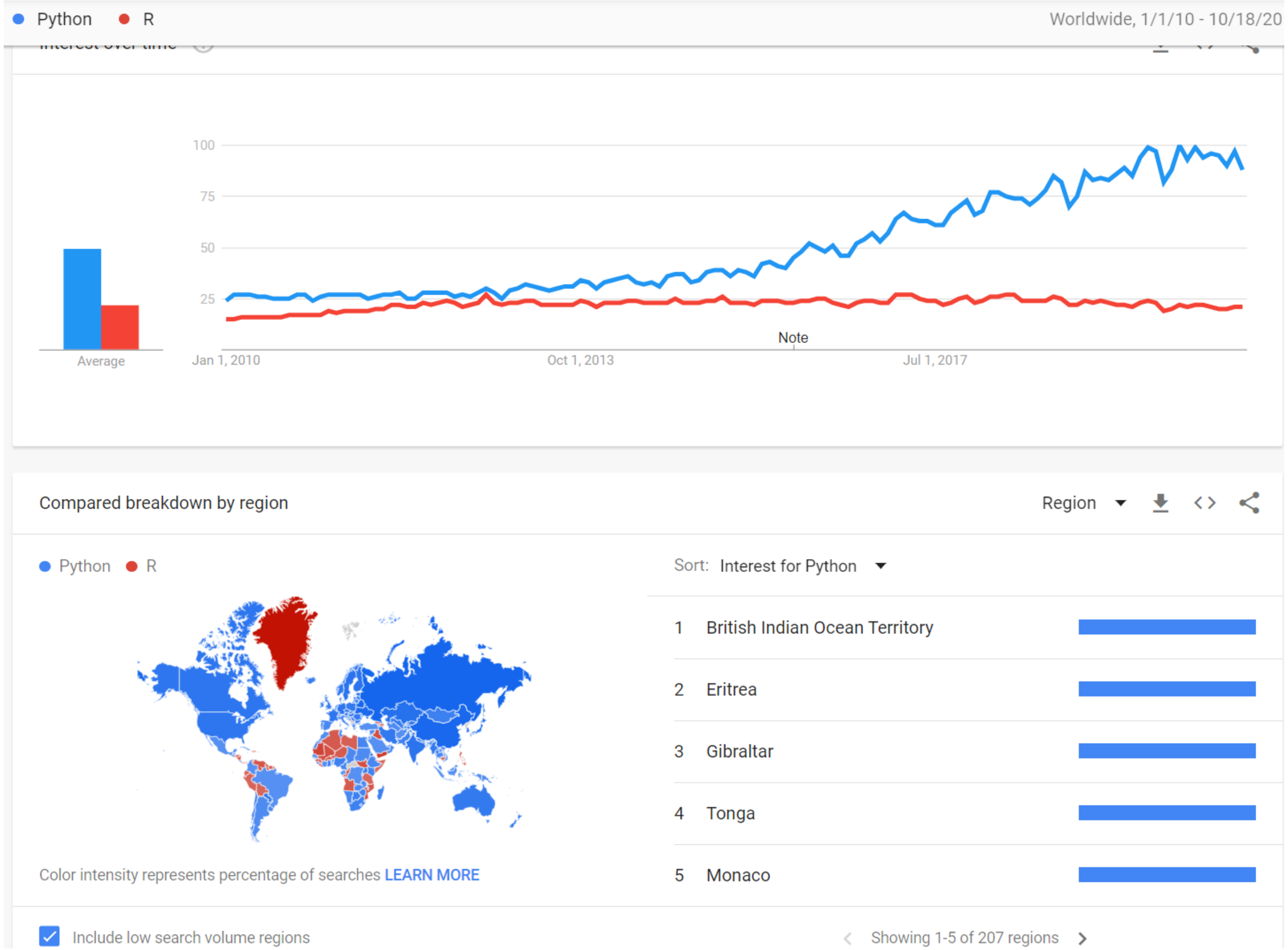
MAJOR COMPANIES THAT USE

python

Google NETFLIX facebook Instagram
amazon Quora slack intel NASA
Dropbox ebay Spotify Capital One

Amr.Moslim

Why I should Learn Python ?



Compared breakdown by region

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

Thank You for Your Attention