Yousef Elbaroudy

Introduction to Data Science

GUIDELINES

- Just focus, you don't need to memorize all things will be mentioned
- This introduction to make the next step into data science
- Don't mind to ask, just say the question whenever you want
- You are not about to get into an exam, so don't get confused
- Try to make a notes about the important things that will be mentioned

Enjoy your trip ☺

WHAT IS MENTORSHIP ?

Mentorship is about mentoring through some specific field, so you are not here for training as long as you are here to gain more experience, practice and have some tasks in pocket to implement, you should have previous experience in at least one field, no certificates, just practicing.

WHAT IS INTERNSHIP ?

Internship is provided by companies, organizations or educational governments to be trained for a predefined duration, some require no prerequisites or no experience to be accepted, and at the end of the training you gain a certificate.

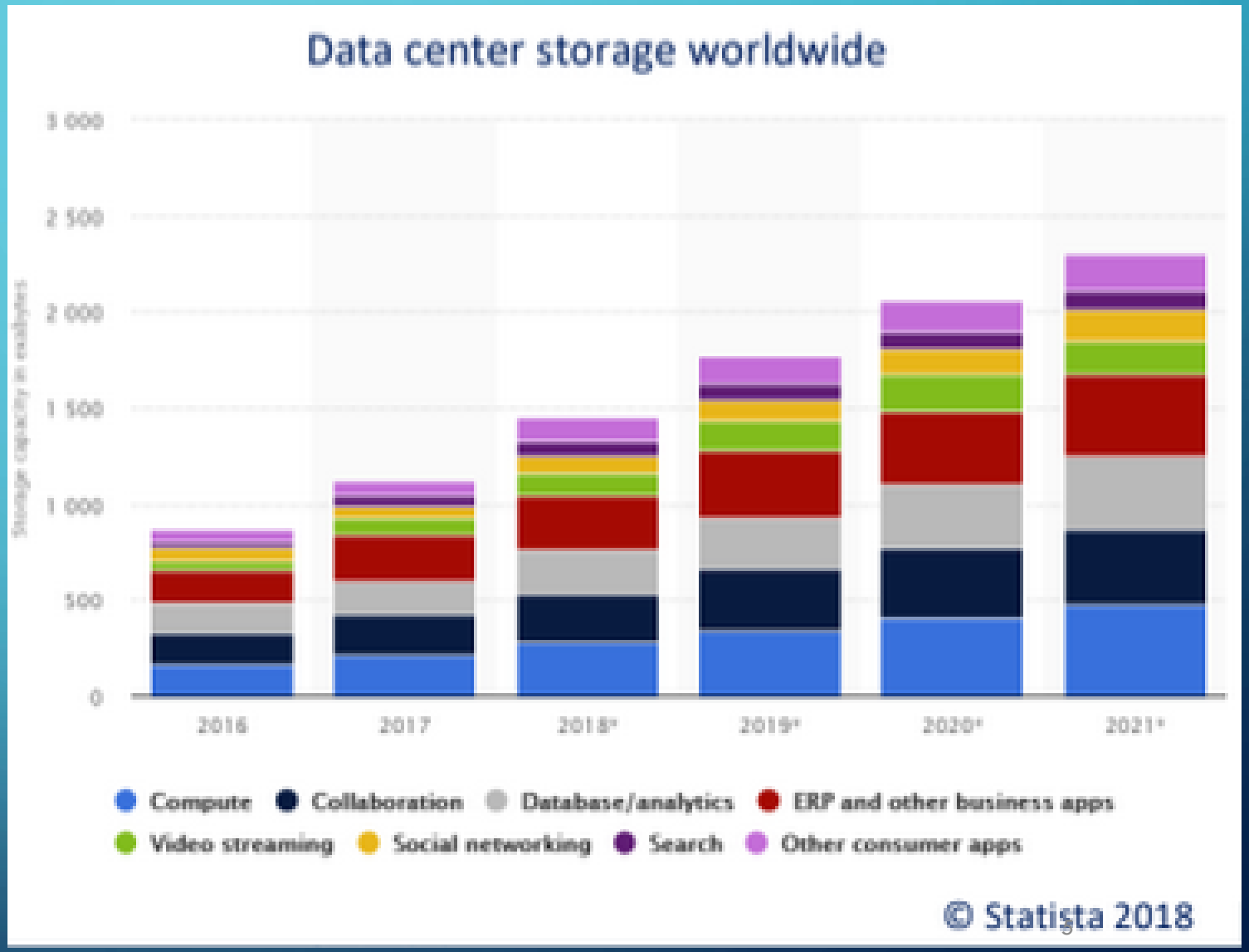
BIG DATA

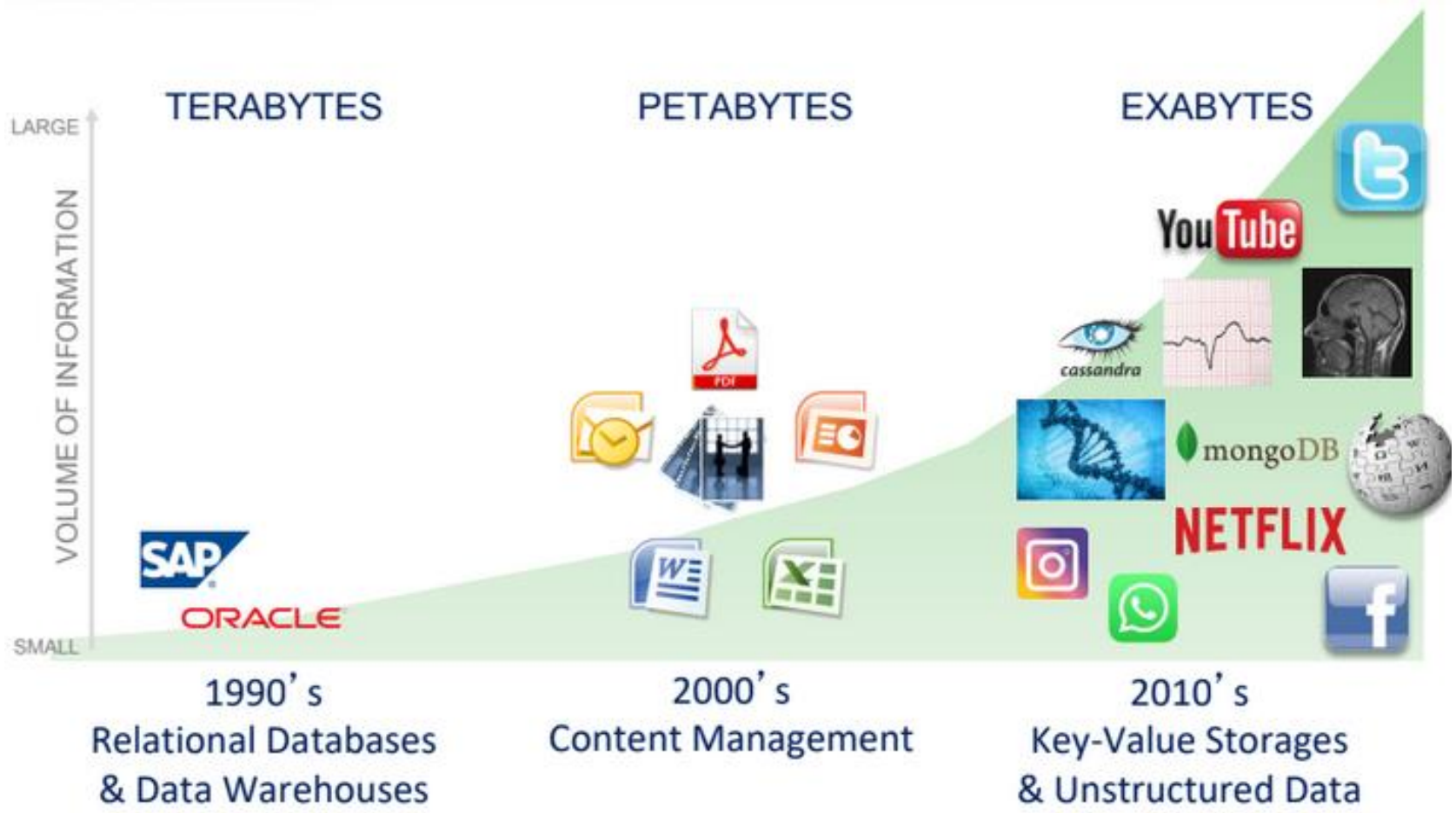
Big Data is **high-volume**, **high-velocity** and/or **high-variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

This is the Concept of 3 Vs

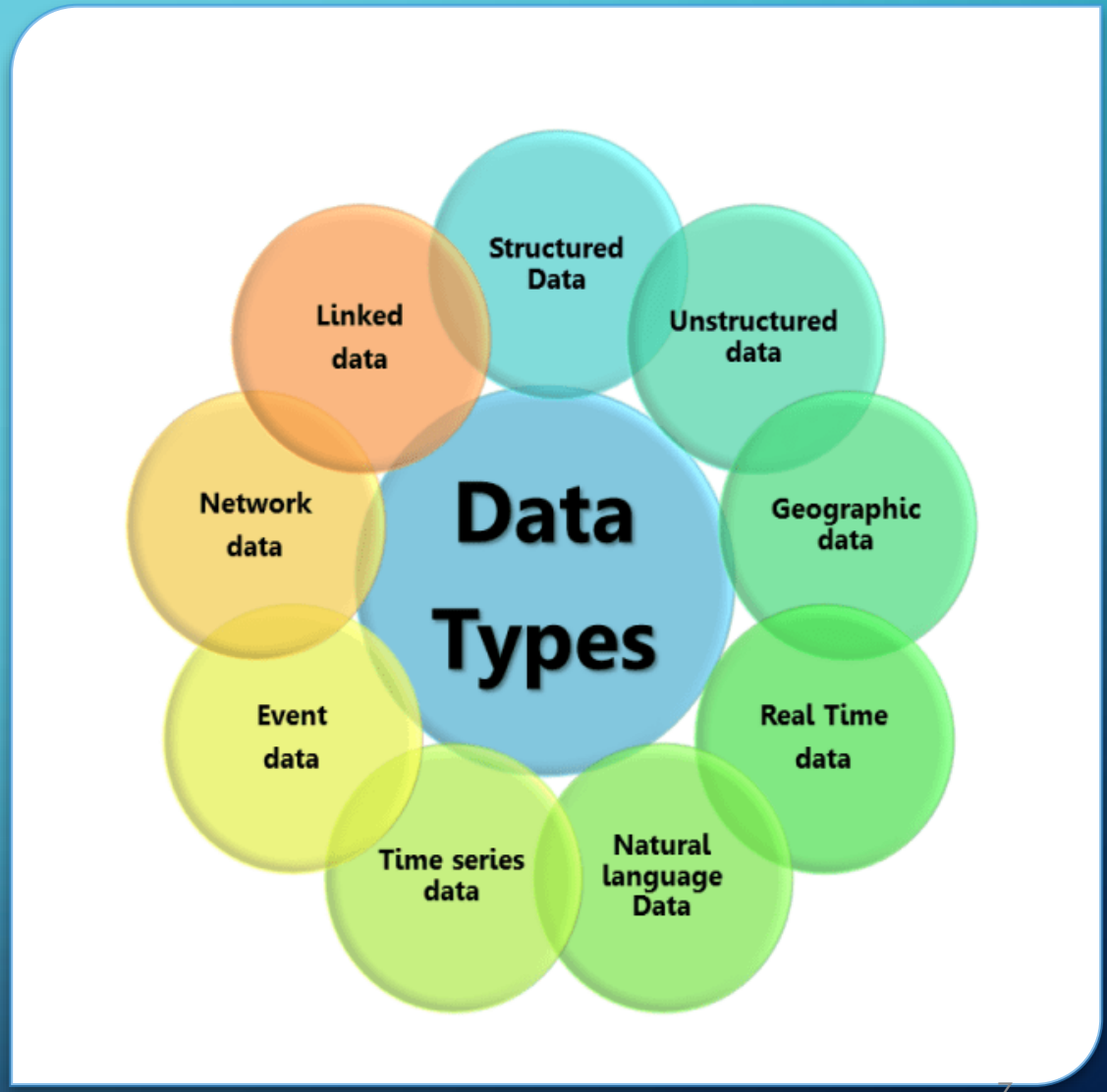
Volume	Velocity	Variety
Scale of the data must be big	Speed at which new data is created and at which data must be processed and analyzed	Diversity in data types and data sources

VOLUME

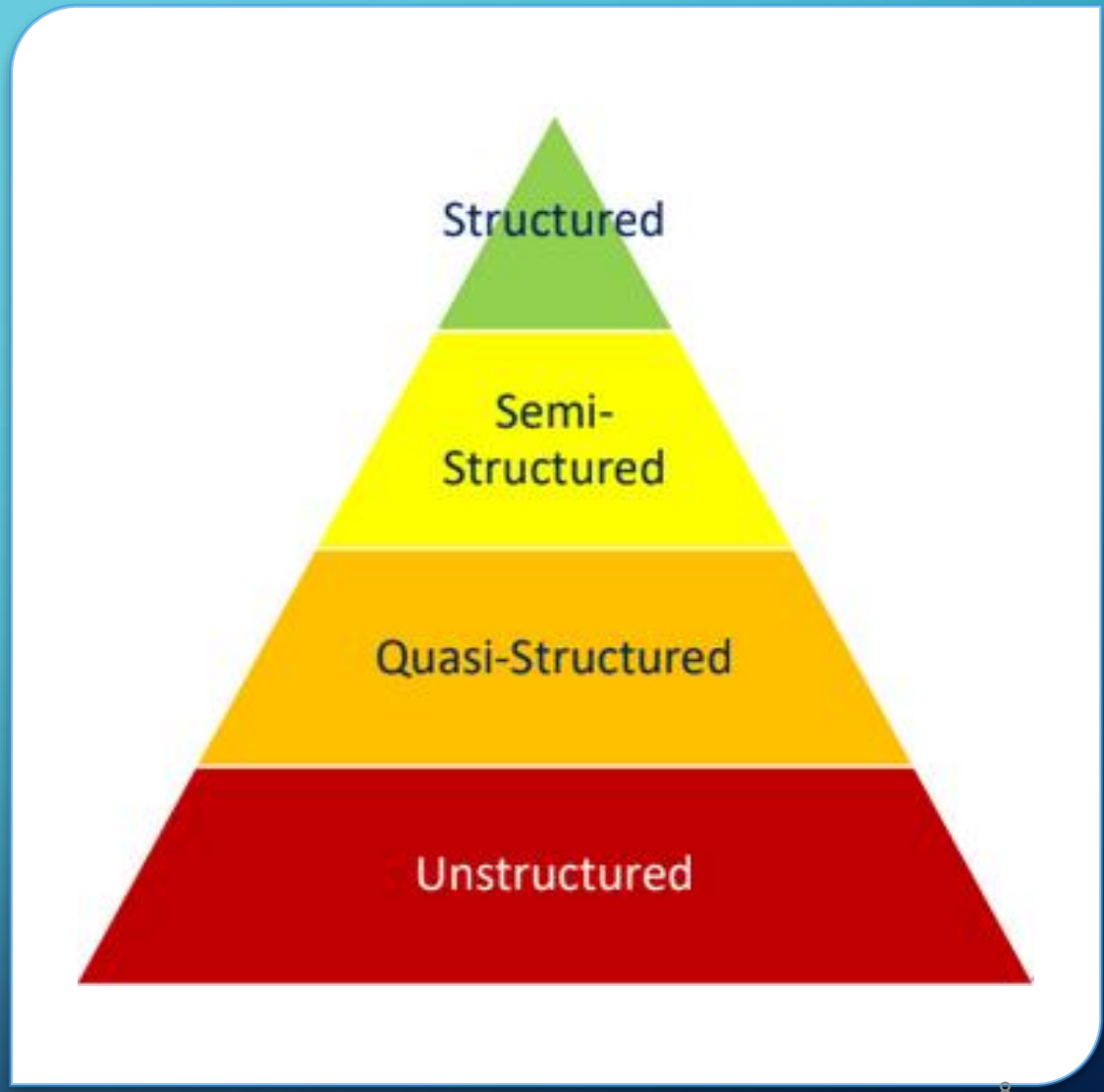

Data center storage worldwide
3 000
2 500
2 000
1 500
1 000
500
0
2016
2017
2018*
2019*
2020*
2021*
Compute
Collaboration
Database/analytics
ERP and other business apps
Video streaming
Social networking
Search
Other consumer apps
© Statista 2018


TERABYTES
PETABYTES
EXABYTES
LARGE
VOLUME OF INFORMATION
SMALL
SAP
ORACLE
YouTube
cassandra
mongoDB
NETFLIX
1990's
Relational Databases
& Data Warehouses
2000's
Content Management
2010's
Key-Value Storages
& Unstructured Data

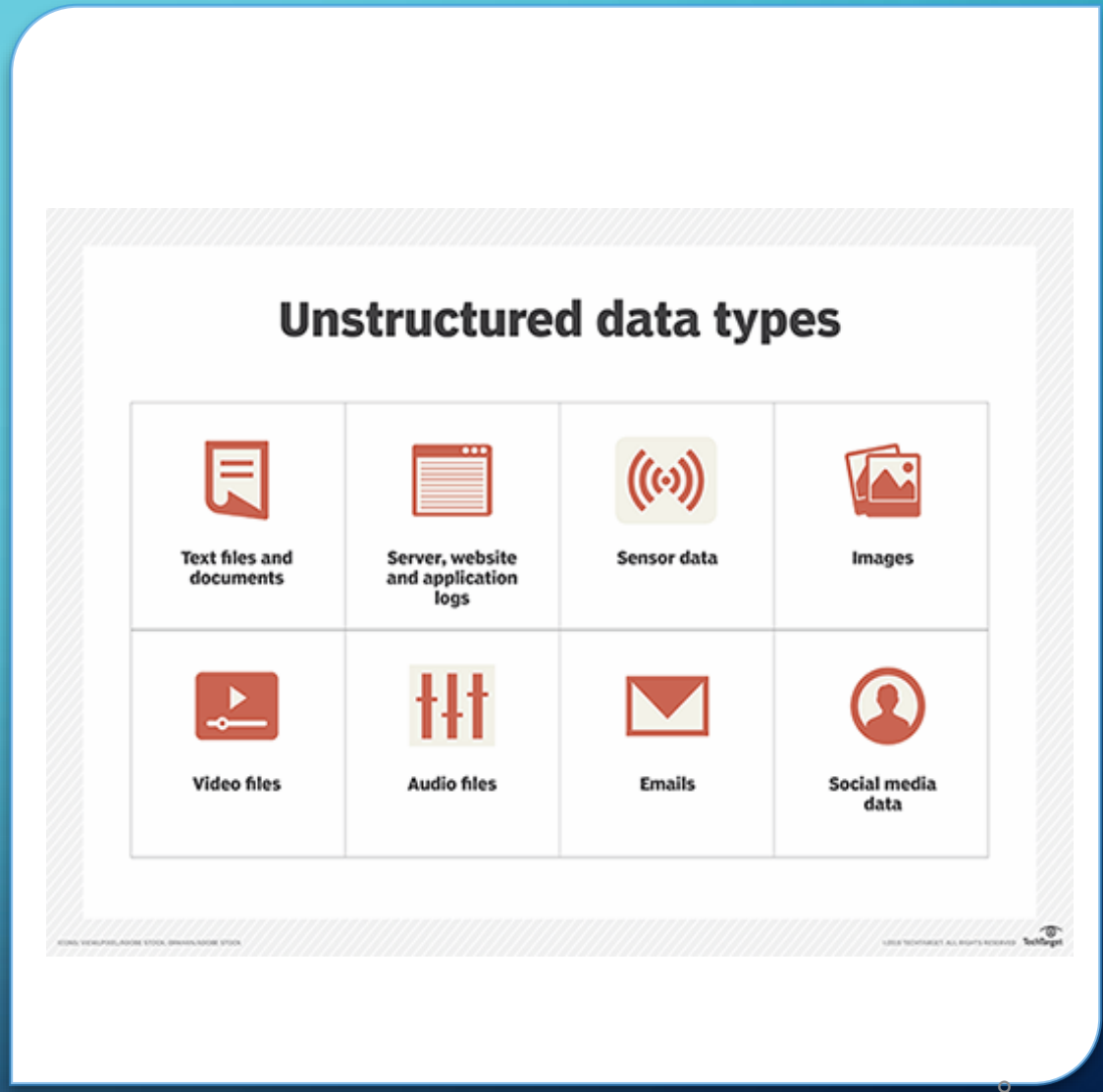
TYPES OF DATA



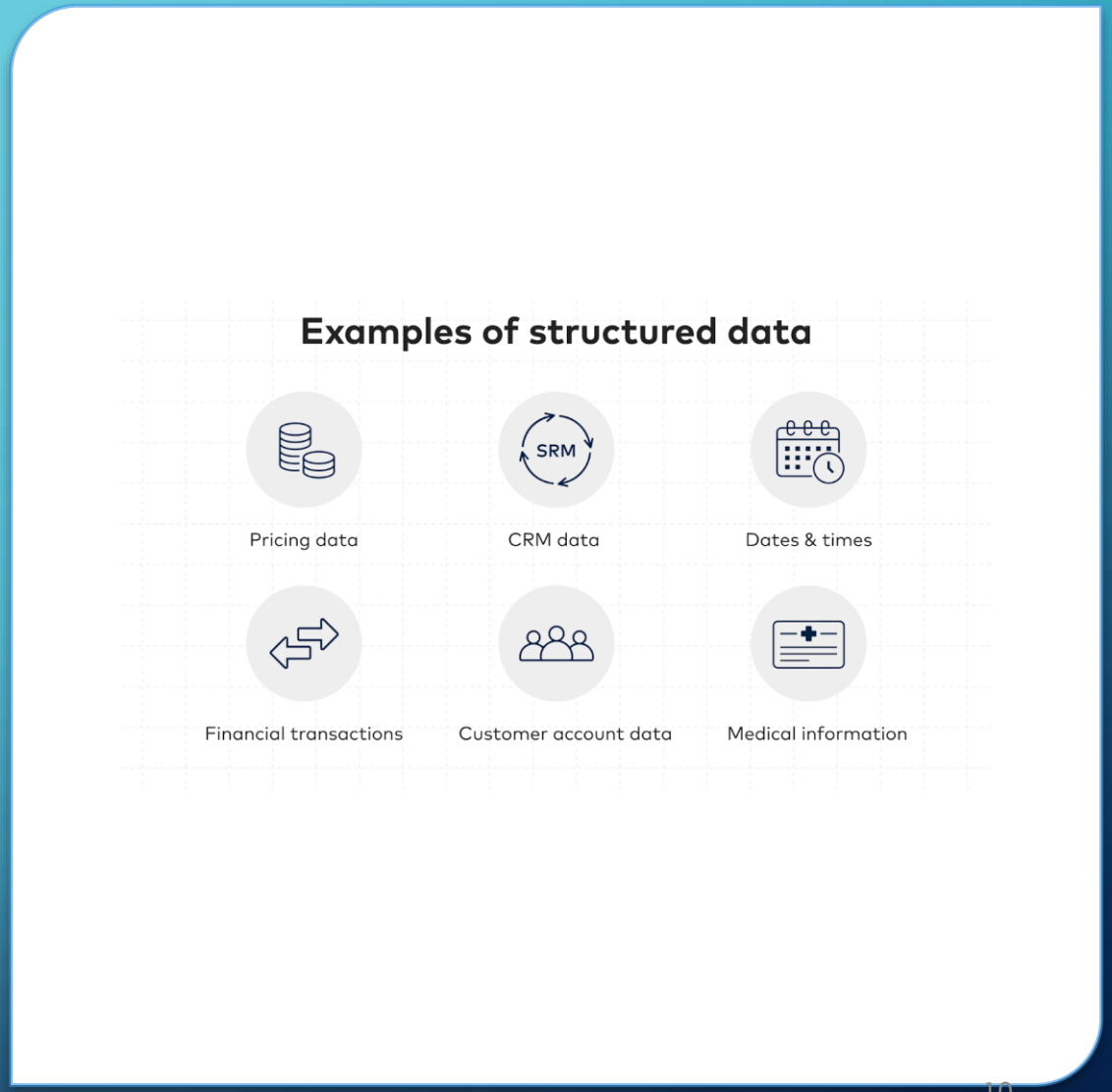
TYPES OF DATA



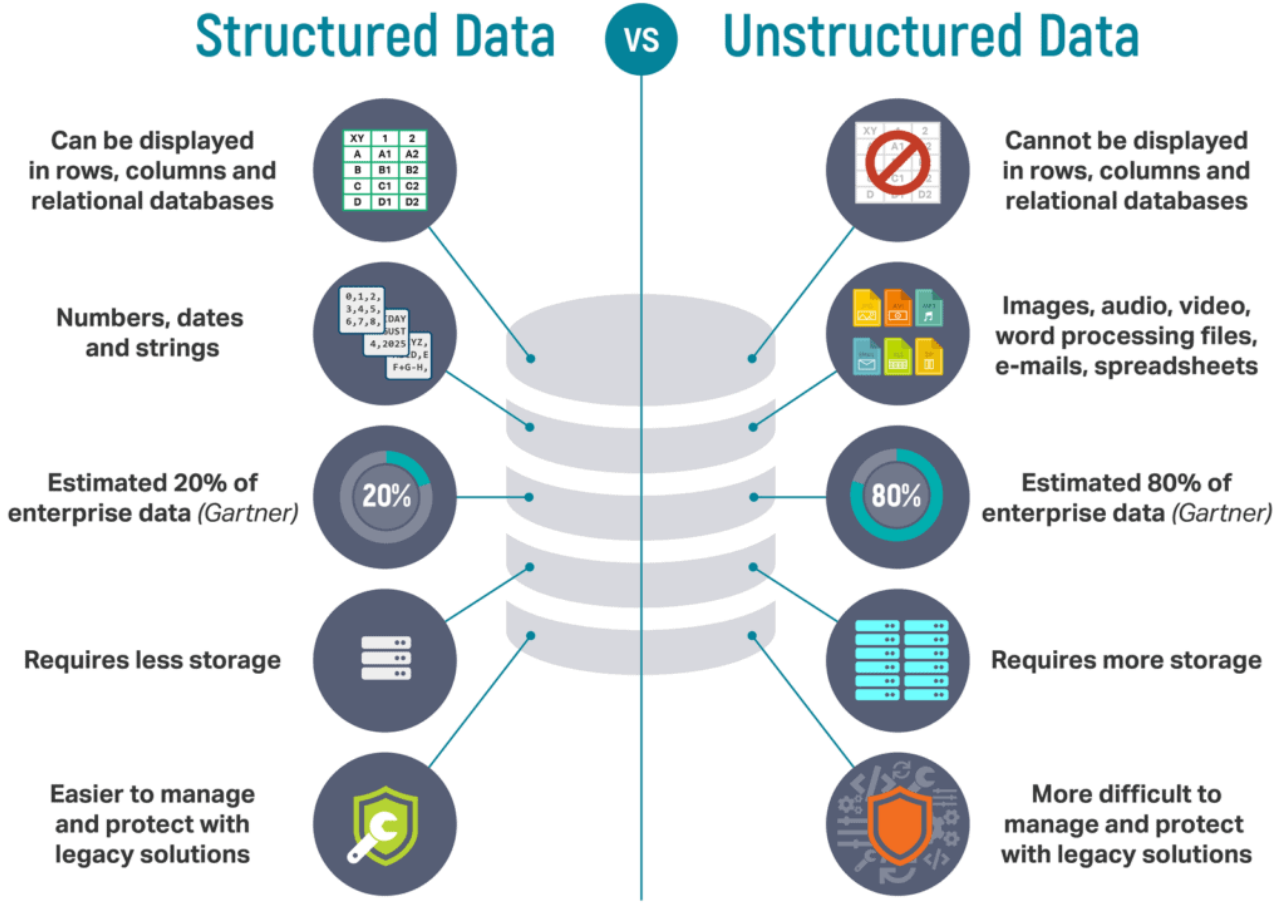
UNSTRUCTURED DATA



STRUCTURED DATA



STRUCTURED VS. UNSTRUCTURED



EACH TYPE OF DATA IS DIFFERENT TYPE OF FIELD !

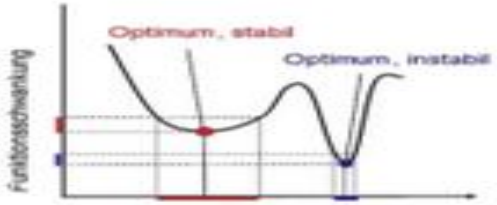
DATA SCIENCE

There is **NO** definition for the data science itself, since it has multiple aspects of disciplines **(Data Science does not mean ML or AI)**

What is the goal ?

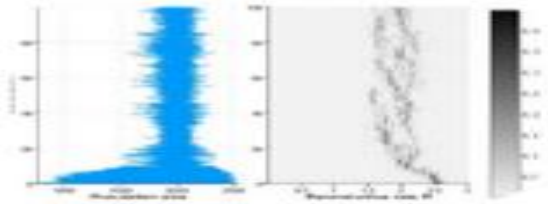
The Extraction of Knowledge from data, and the automation process in the next step with the integration of Software Engineering or Business Intelligence.

Computational Geometry



Optimization

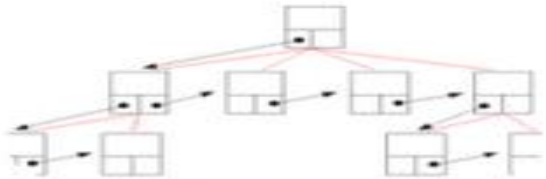
Stochastics



Scientific Computing

Machine Learning

MATHEMATICAL ASPECT (SCIENTIFICALLY)



Data Structures and Algorithms

Databases

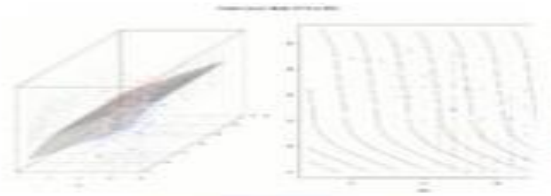
Distributed Computing

Software Engineering

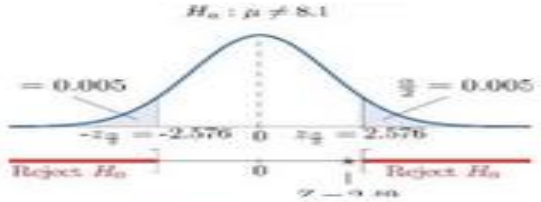
Artificial Intelligence

Machine Learning

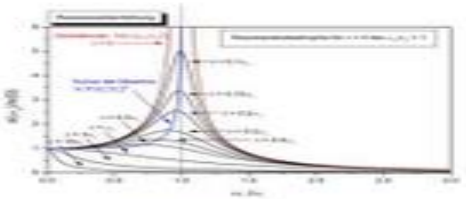
COMPUTER SCIENCE ASPECT (TECHNICALLY)



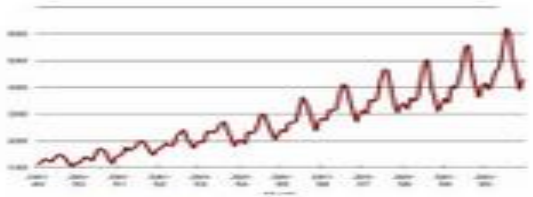
Linear Models



Statistical Tests



Inference

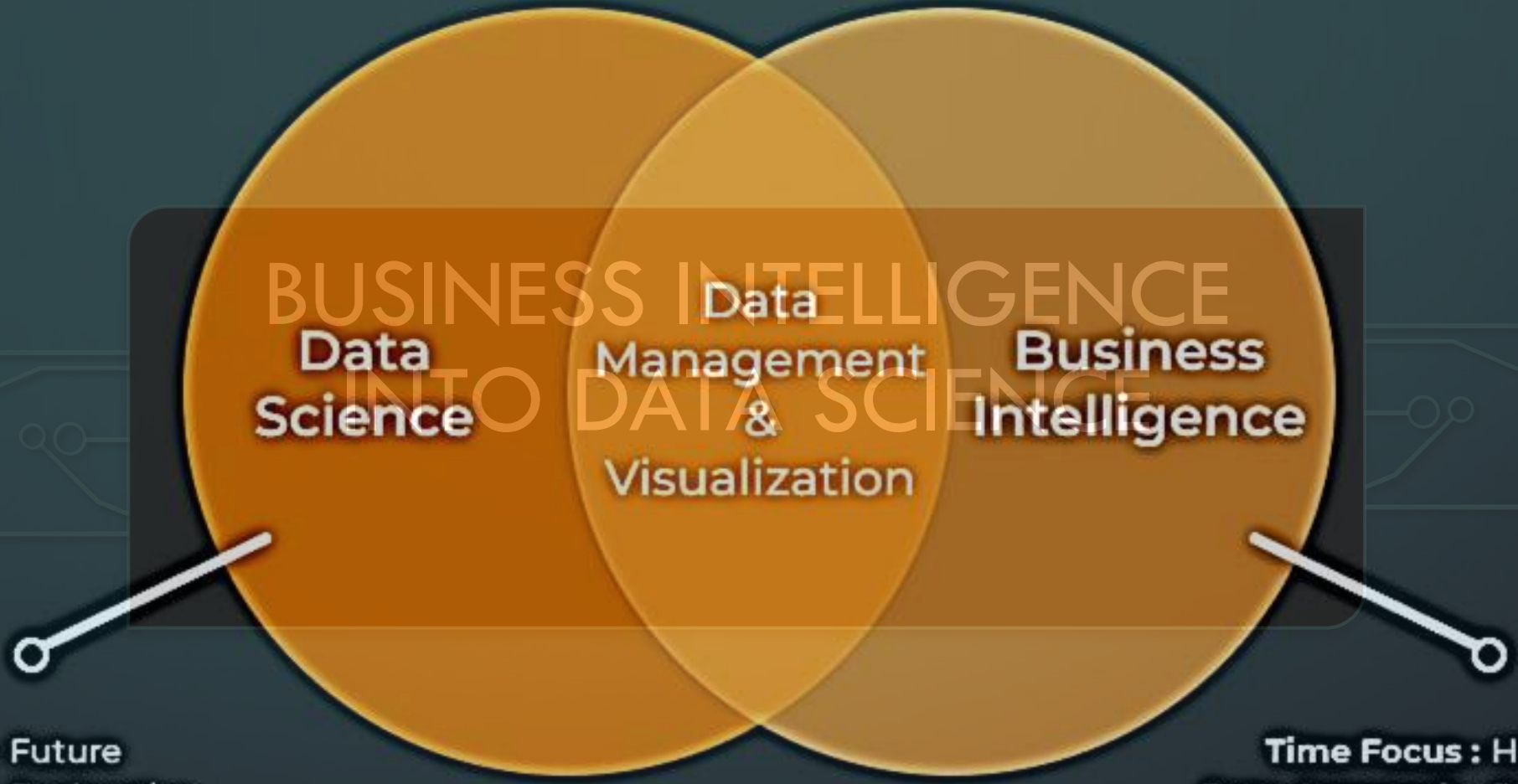


Time Series Analysis

Machine Learning

STATISTICAL ASPECT (BUSINESS)

Data Science vs. Business Intelligence



Time Focus : Future
Data Approach : Explorative
Automation : Low
Business Driver : Planning
Business Value : Hypothesis Testing

Time Focus : Historic
Data Approach : Descriptive
Automation : High
Business Driver : Decision Support
Business Value : Trend Identification

Data Science vs. Business Intelligence

FACTORS	DATA SCIENCE	BUSINESS INTELLIGENCE
Concept	Consists of several data operations in various domains	Deals with data analysis on the business platform
Scope	Past data is analyzed for future predictions	BI analyzes past data
Data	Both structured & unstructured data that is also dynamic	Handling static and structured data
Data Storage	Data utilized is distributed in real time clusters	Data stored mostly in data-warehouses
Procedure	Questions are both curated and solved by data scientists	BI helps companies to solve questions

Data source: https://data-flair.training/blogs/business-intelligence-vs-data-science/

axisbits

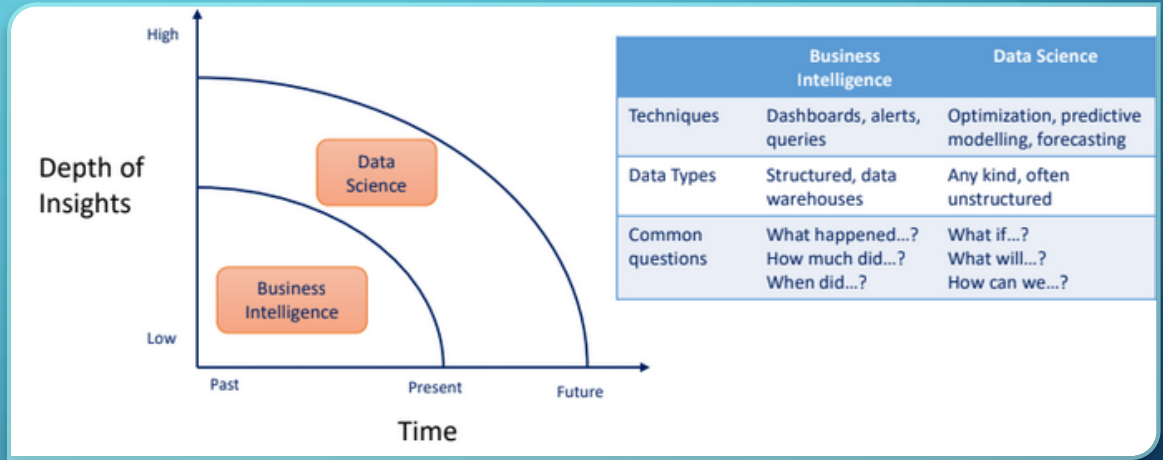
Business Intelligence *vs* Data Science

Factors	Business Intelligence	Data Science
Concept	Deals with data analysis on the business platform.	Consists of several data operations in various domains.
Scope	BI analyzes past data	Past data is analyzed for future predictions.
Data	Handling static and structured data	Both structured & unstructured data that is also dynamic.
Data Storage	Data stored mostly in data-warehouses	Data utilized is distributed in real time clusters.
Procedure	BI helps companies to solve questions.	Questions are both curated and solved by data scientists.
Tools	MS Excel, SAS BI, Sisense, Microstrategy	Python, R, Hadoop/Spark, SAS, TensorFlow.

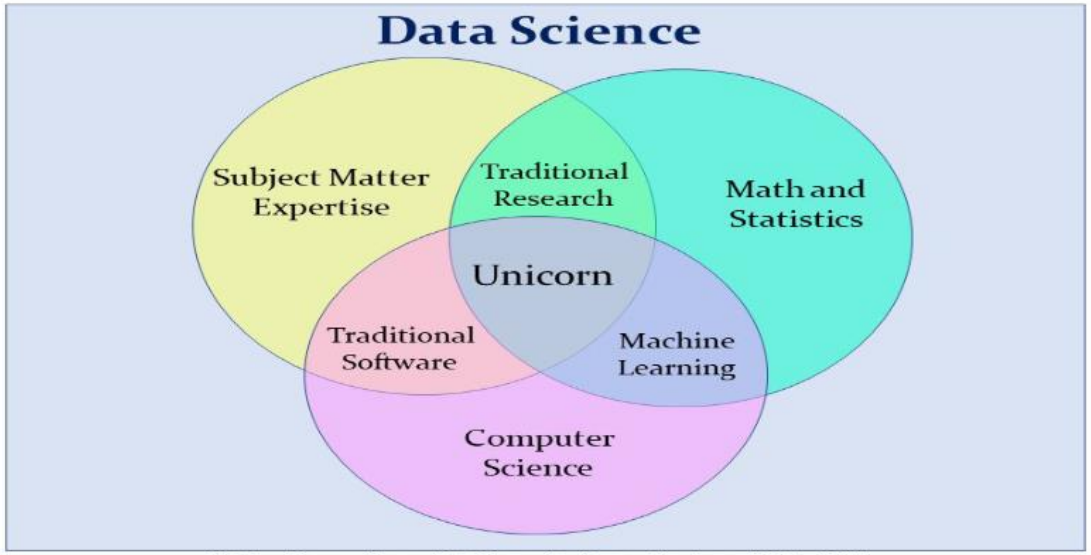
DATA SCIENCE VS BUSINESS INTELLIGENCE

- **Business Intelligence** best practices that enable access to and analysis of information to improve and optimize decisions and performance.

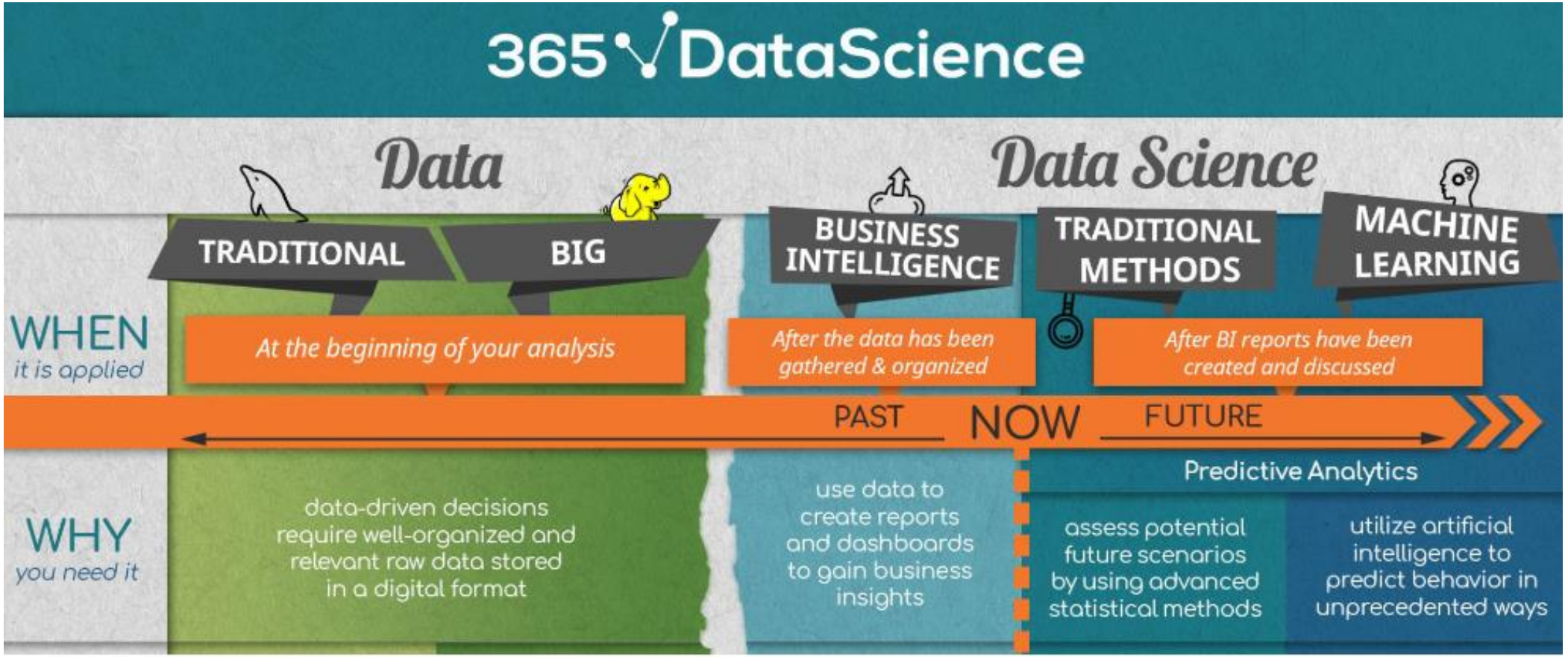
Secret Note
That's make sense, since there is a difference between Data Analysis and Data Analytics



	Business Intelligence	Data Science
Techniques	Dashboards, alerts, queries	Optimization, predictive modelling, forecasting
Data Types	Structured, data warehouses	Any kind, often unstructured
Common questions	What happened...? How much did...? When did...?	What if...? What will...? How can we...?



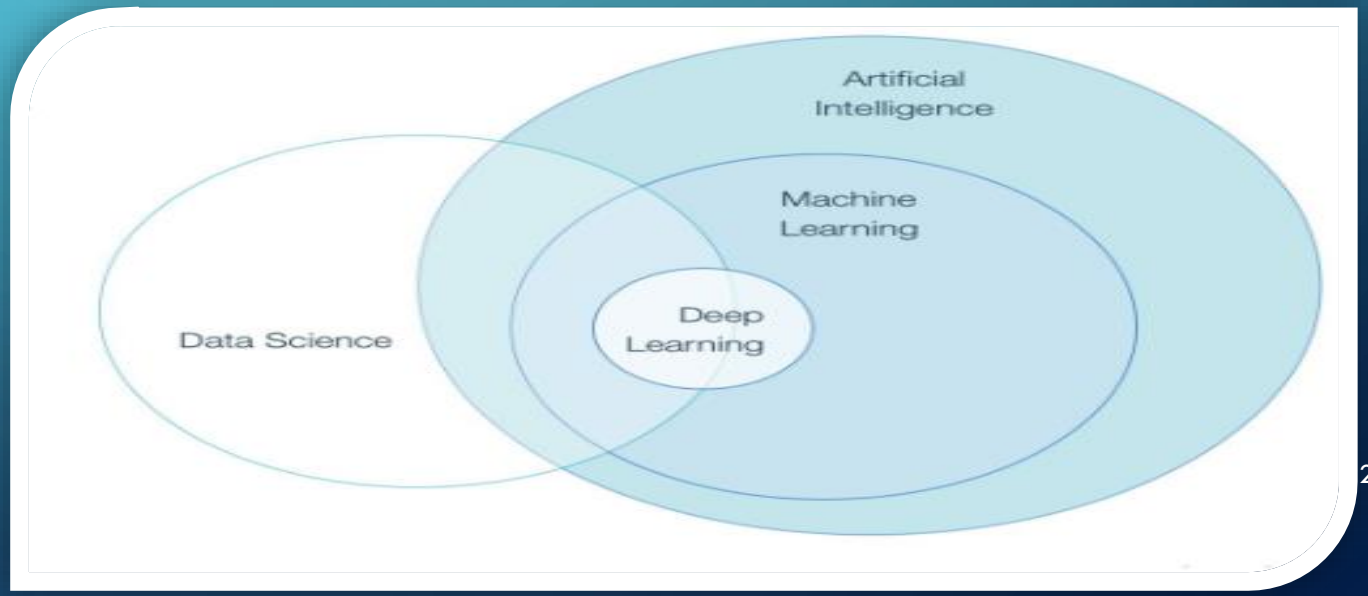
FOCUSING ON DATA SCIENCE


365 DataScience
Data
Data Science
TRADITIONAL
BIG
BUSINESS INTELLIGENCE
TRADITIONAL METHODS
MACHINE LEARNING
WHEN
it is applied
At the beginning of your analysis
After the data has been gathered & organized
After BI reports have been created and discussed
PAST
NOW
FUTURE
Predictive Analytics
WHY
you need it
data-driven decisions require well-organized and relevant raw data stored in a digital format
use data to create reports and dashboards to gain business insights
assess potential future scenarios by using advanced statistical methods
utilize artificial intelligence to predict behavior in unprecedented ways

DATA SCIENCE & AI

According to the Turing Test

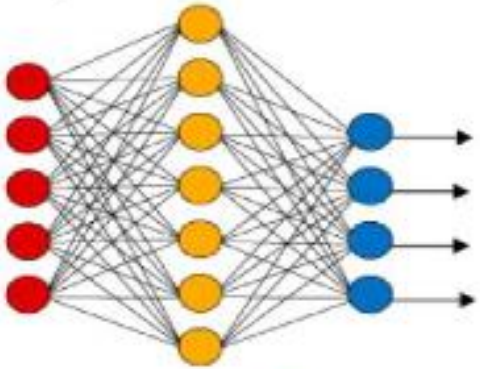
Artificial Intelligence is the machine's ability to exhibit intelligence behavior indistinguishable from that of a human.



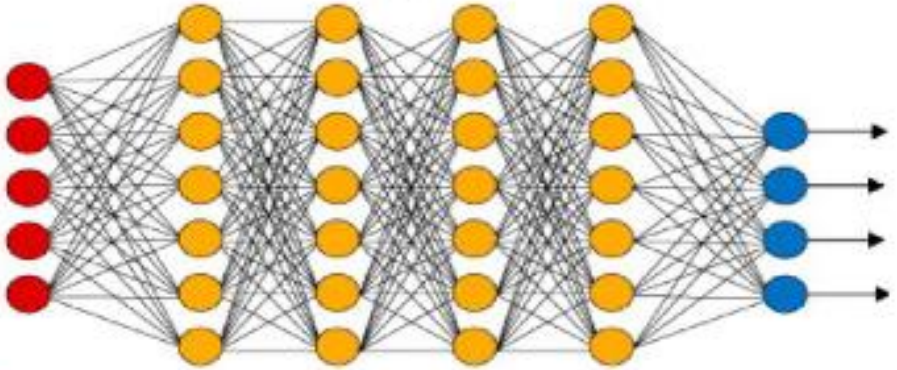
TRADITIONAL PROGRAMMING

MACHINE LEARNING

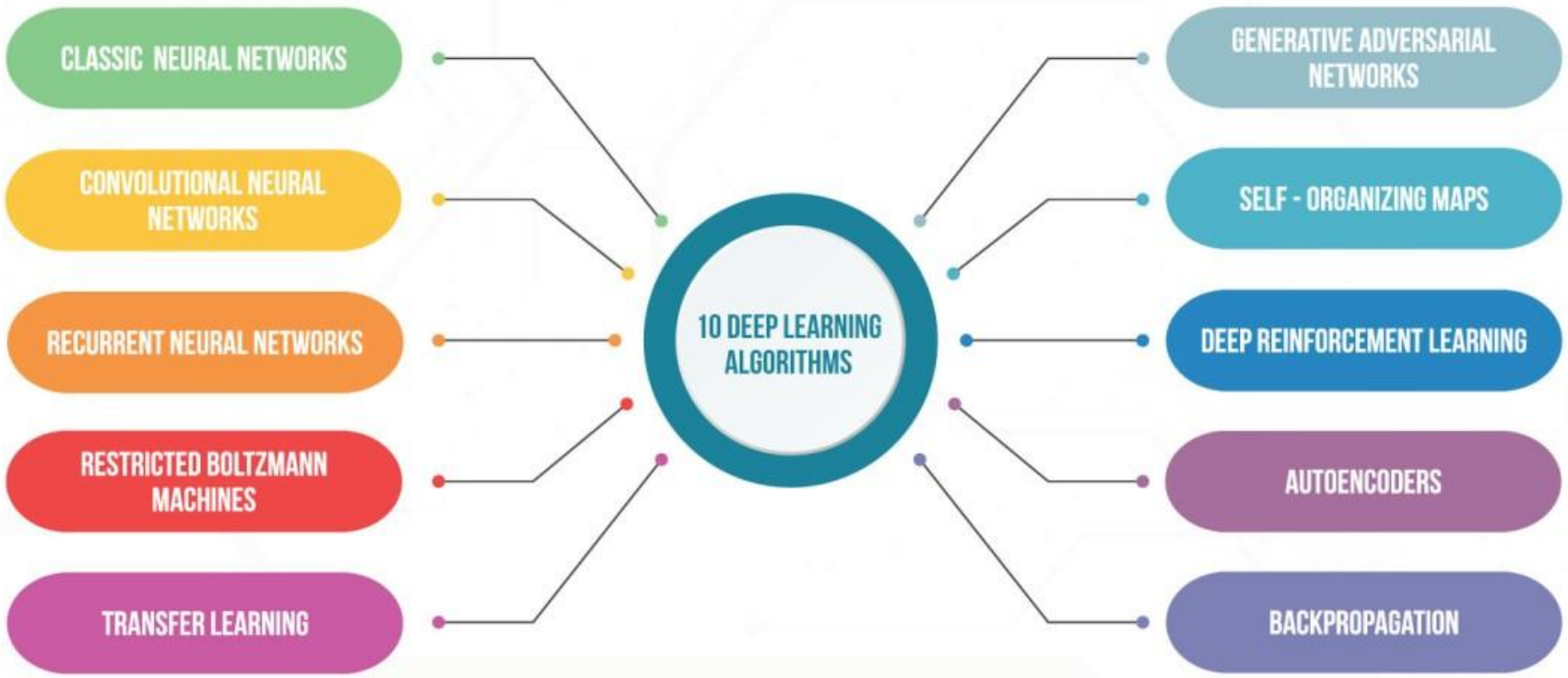
Simple Neural Network



Deep Learning Neural Network

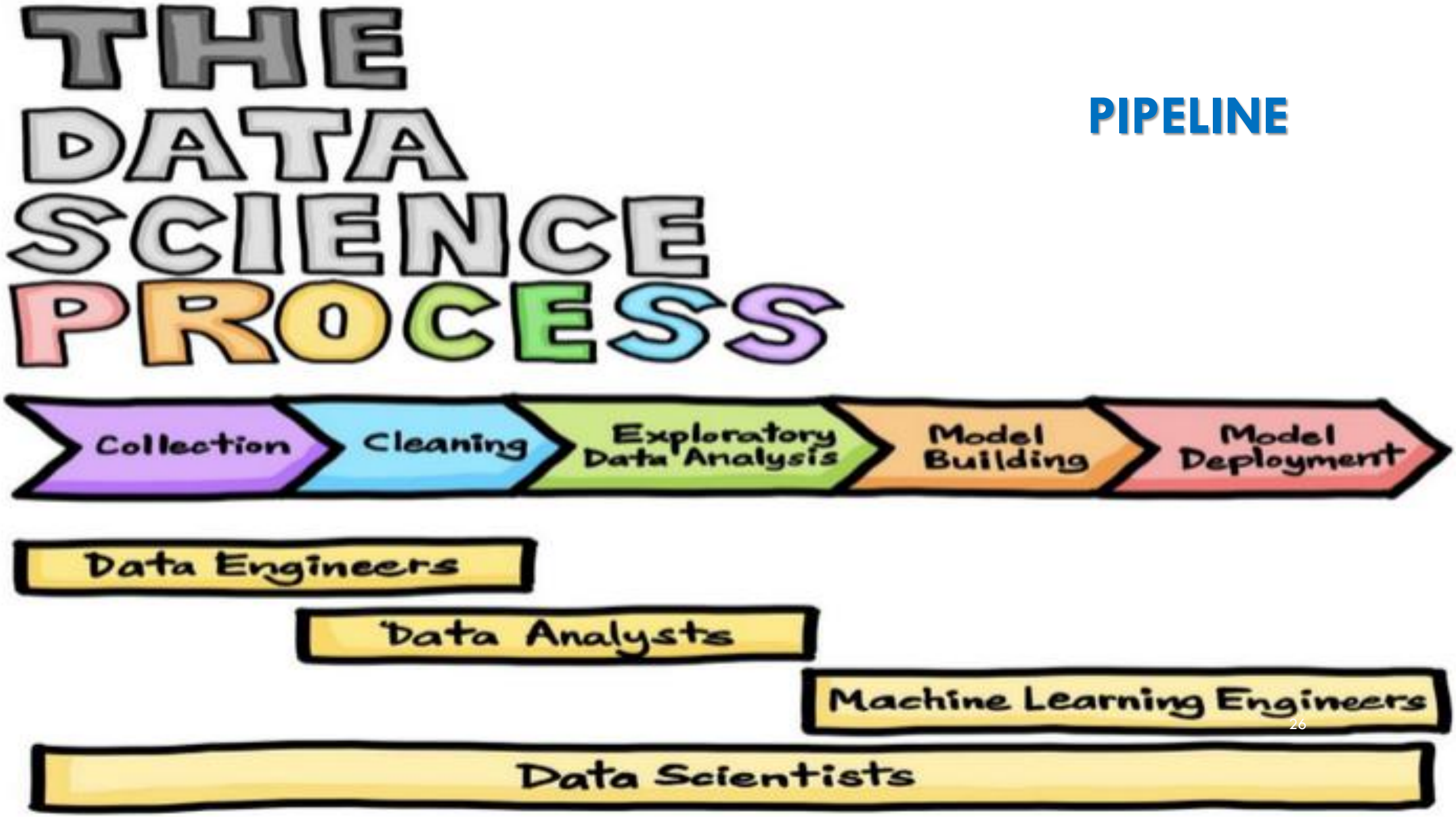


Input Layer Hidden Layer Output Layer


CLASSIC NEURAL NETWORKS
CONVOLUTIONAL NEURAL NETWORKS
RECURRENT NEURAL NETWORKS
RESTRICTED BOLTZMANN MACHINES
TRANSFER LEARNING
10 DEEP LEARNING ALGORITHMS
GENERATIVE ADVERSARIAL NETWORKS
SELF - ORGANIZING MAPS
DEEP REINFORCEMENT LEARNING
AUTOENCODERS
BACKPROPAGATION
SOURCE: SHUTTERSTOCK
© 2022 ALL RIGHTS RESERVED
NEIL SAHOTA

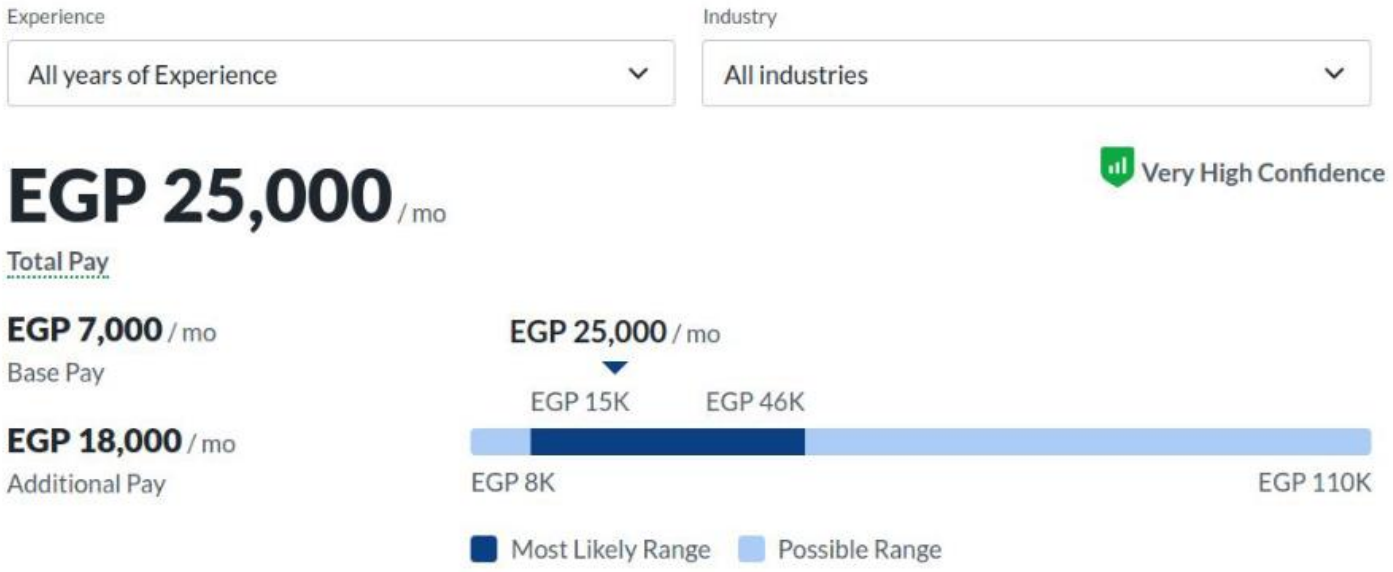
WHAT IS THE PIPELINE ?

PIPELINE


THE DATA SCIENCE PROCESS
Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists

DATA ANALYST

- Job might consist of tasks like:
  - pulling data out of SQL databases
  - becoming an Excel or Tableau master
  - producing basic Data Visualizations and Reporting Dashboards
  - On occasion:
    - analyze the results of an A/B Test
    - take the lead on company's Google Analytics account
- Some companies: Data Scientist is synonymous to Data Analyst



https://www.glassdoor.com/Salaries/cairo-data-analyst-salary-SRCH_IL.0,5_IM1175_KO6,18.htm

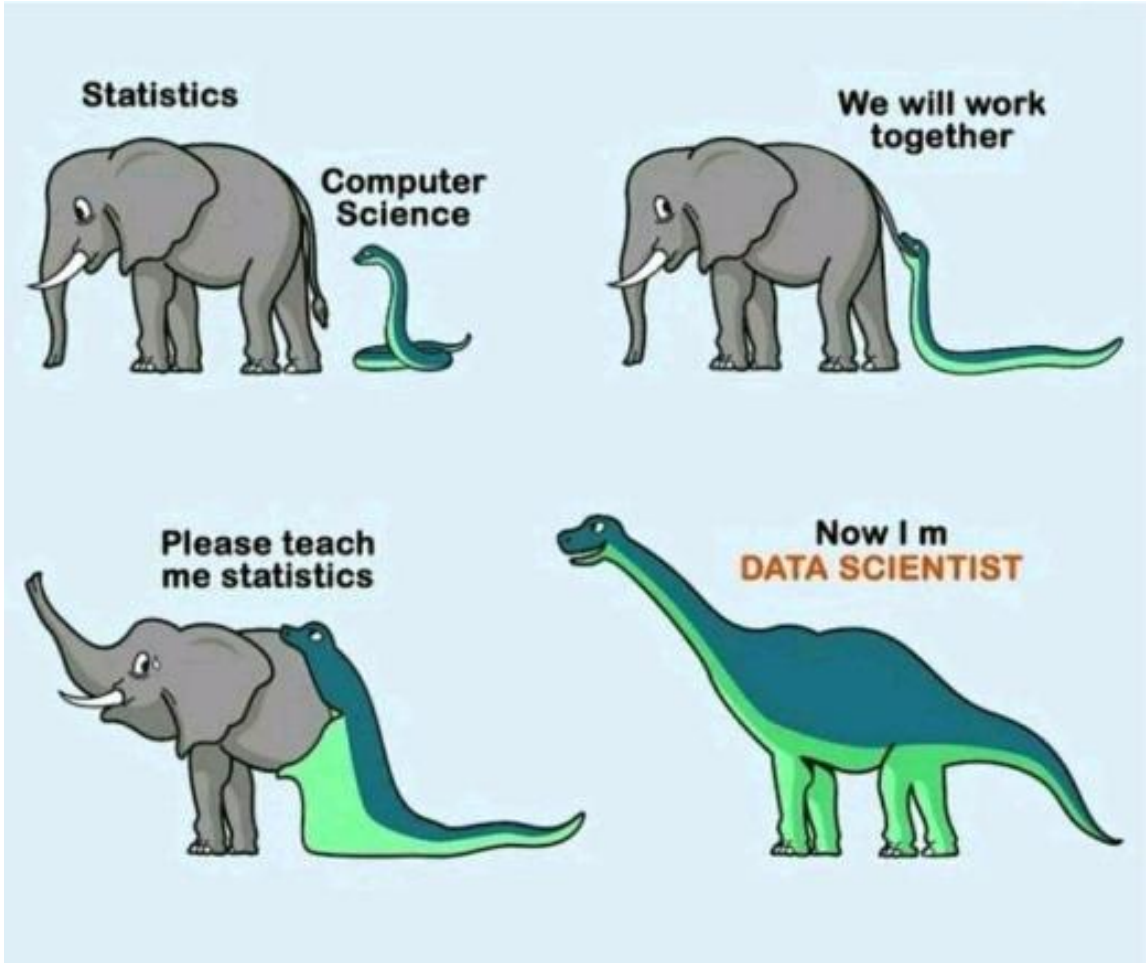
DATA ANALYST AVERAGE SALARIES: EGYPT 16/2/23

ML ENGINEERS

- Some companies: data or data analysis platform is the product
  - so, Data Analysis or ML can be pretty intense
  - so, there is a need for someone who:
    - has a formal mathematics or statistics background
    - is hoping to continue down a more academic path
- **ML Engineers**
  - often focus more on producing great data-driven products
  - less focus on answering operational questions for a company

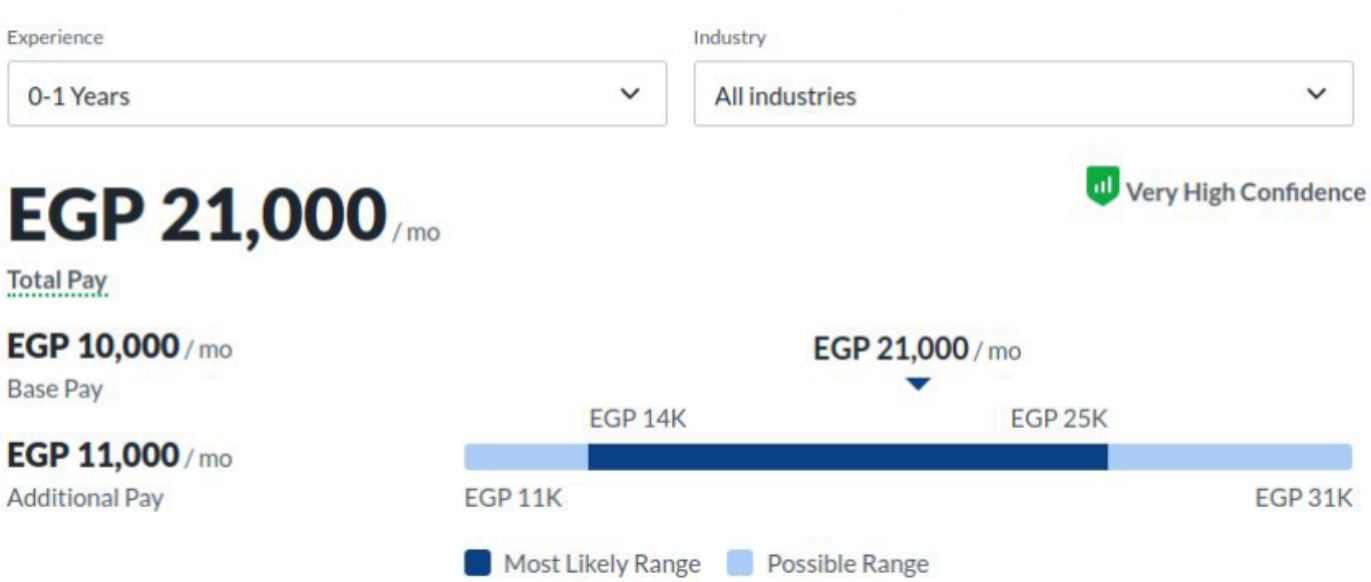
DATA ENGINEERS

- Some companies: have lot of traffic and large amount of data
  - so, there is a need for someone who:
    - set up lot of data infrastructure that the company will need moving forward
    - can provide analysis
- Job postings listed under Data Scientist and Data Engineer
  - strong software engineering skills are more important
  - heavy statistics and ML expertise are less important



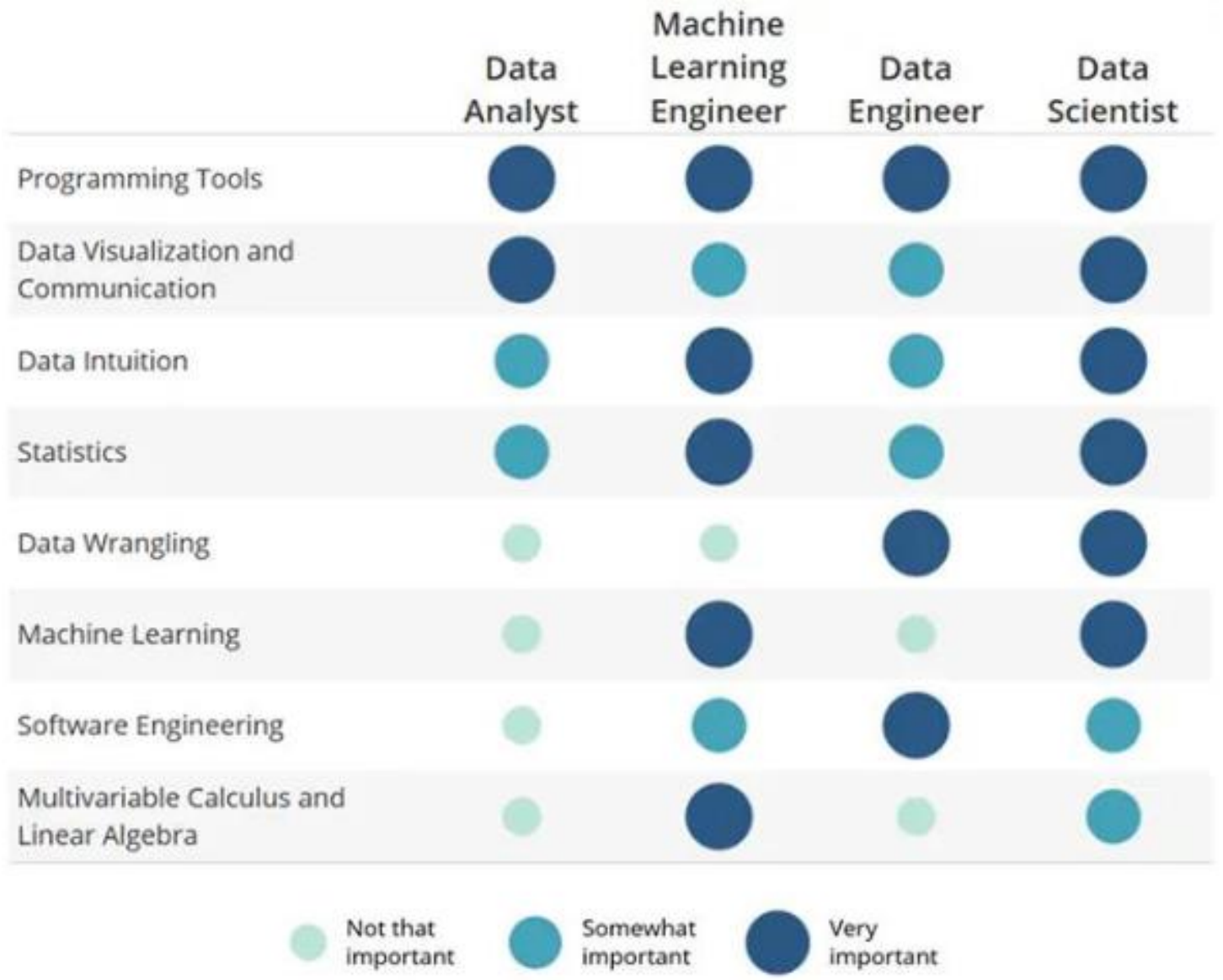
DATA SCIENTISTS

- Often used as a blanket title to describe jobs that are drastically different



https://www.glassdoor.com/Salaries/cairo-data-scientist-salary-SRCH_IL.0,5_IM1175_KO6,20.htm

DATA SCIENTIST AVERAGE SALARIES: EGYPT 2022


Data Analyst
Machine Learning Engineer
Data Engineer
Data Scientist
Programming Tools
Data Visualization and Communication
Data Intuition
Statistics
Data Wrangling
Machine Learning
Software Engineering
Multivariable Calculus and Linear Algebra
Not that important
Somewhat important
Very important

Types of Data Professionals

Where are you?

Data Engineer

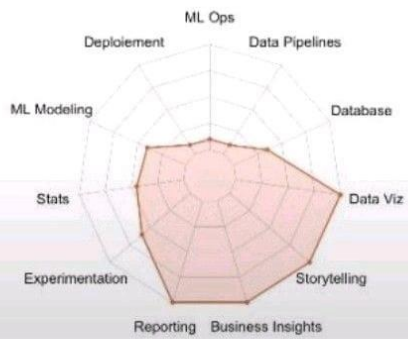
Data Scientists
ML Ops

ML Engineer

Data Analysts
ML Ops

Data Scientist

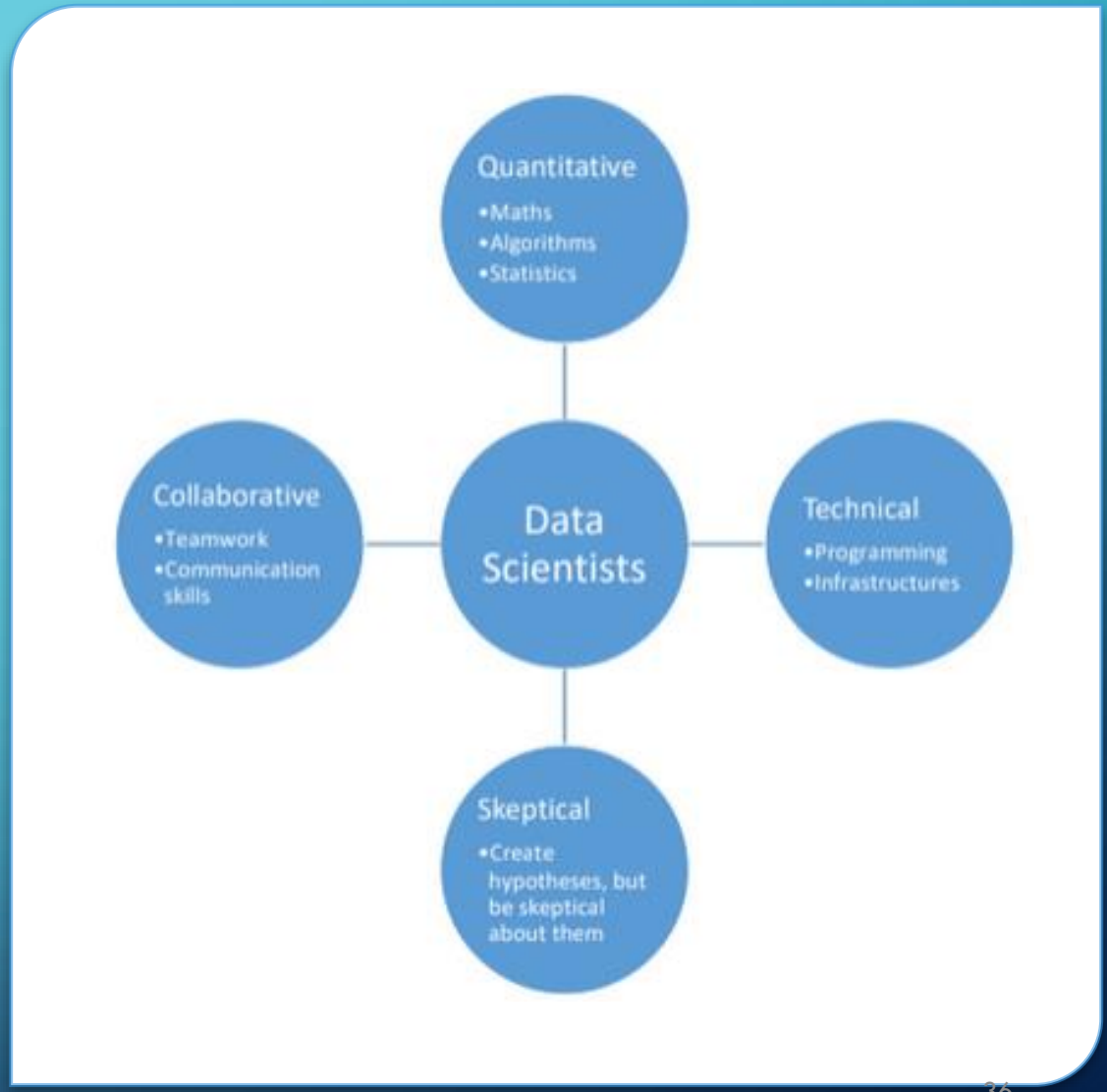
Data Analyst



WHAT ARE DATA SCIENTISTS ?

- Not computer scientists
  - ❑ But should know about databases, data structures, algorithms, etc.
- Not mathematicians
  - ❑ But should know about optimization, stochastics, etc.
- Not statisticians
  - ❑ But should know about regression, statistical tests, etc.
- Not domain experts
  - ❑ But must work together with them

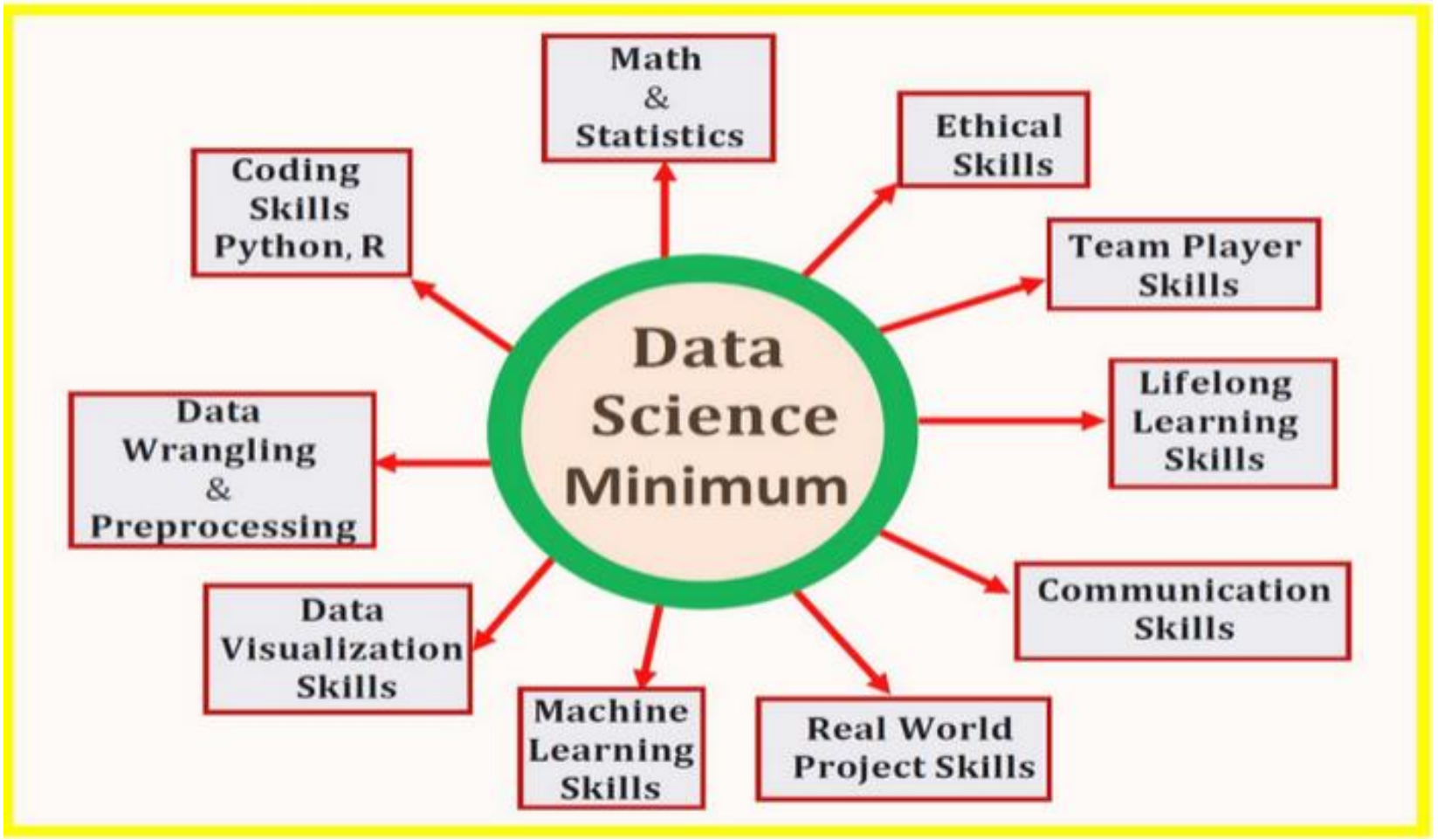
SKILLS


Quantitative
•Maths
•Algorithms
•Statistics
Collaborative
•Teamwork
•Communication skills
Data Scientists
Technical
•Programming
•Infrastructures
Skeptical
•Create hypotheses, but be skeptical about them

DIFFERENT TYPES OF DATA SCIENTISTS

According to Microsoft Research:

- Polymath
  - „Do it all"
- Data Evangelist
  - Data analysis, disseminating and acting on insights
- Data Preparer
  - Querying existing data, preparing data for analysis
- Data Shapers
  - Analyzing and preparing data
- Data Analyzer
  - Analyzing data
- Platform Builder
  - Collect data and create infrastructures
- Moonlighters (50%/20%)
  - „Spare time" data scientists
- Insight Actors
  - Use the outcome and act on insights.


Data Science Minimum
Math & Statistics
Ethical Skills
Coding Skills Python, R
Team Player Skills
Data Wrangling & Preprocessing
Lifelong Learning Skills
Data Visualization Skills
Communication Skills
Machine Learning Skills
Real World Project Skills

APPLICATIONS OF DATA SCIENCE

Intelligent Systems

Robotics

Marketing

Medicine

Autonomous Driving

Social Networks

DATA SCIENCE PROJECT ROLES

Role	Description
Business User	• Someone who uses the end results • Can consult and advise project team on value of end results and how these will be operationalized
Project Sponsor	• Responsible for the genesis of the project • Generally provides the funding • Gauge the value from the final outputs
Project Manager	• Ensure key milestones and objectives are met on time and at expected quality • Plans and manages resources
Business Intelligence Analyst	• Business domain expertise with deep understanding of the data • Understands reporting in the domain, e.g., Key Performance Indicators (KPIs)
Data Engineer	• Deep technical skills to assist with data management and ETL/ELT
Database Administrator	• Provisions and configures database environment to support the analytical needs of the project
Data Scientist	• Expert on analytical techniques and data modeling • Applies valid analytical techniques to given business problems • Ensures analytical objectives are met

DELIVERABLES

Role	Deliverable
Business User	Expects a sponsor presentation: ➢ Are the results good for me? ➢ What are the benefits for me? ➢ What are the implications for me?
Project Sponsor	Expects a sponsor presentation: ➢ What is the impact of operationalizing the results? ➢ What are the risk and what is the potential ROI? ➢ How can this be evangelized within the organization (and beyond)?
Project Manager	• Responsible for the timely availability of all deliverables • Responsible for the sponsor presentations
Business Intelligence Analyst	Expects an analyst presentation: ➢ Which data was used? ➢ How will reporting change? ➢ How will KPIs change?
Data Engineer	• Responsible for data engineering code and technical documentation
Database Administrator	• Responsible for infrastructure code and technical documentation
Data Scientist	• May be the target audience for analyst presentations. • Responsible for data analysis code and technical documentation • Responsible for the analyst presentation • Support of the project management with the sponsor presentation


BREAK

PREREQUISITES

What do you think ?

PREREQUISITES

Skill	Tool/Topics
Programming Skills	Python, R or MATLAB (Python is Preferred)
Mathematics	Differential Calculus, Linear Algebra
Probability & Statistics	Descriptive Measurements (Mean, Mode, Median), Distributions, Event probability, Conditional Probability, Joint Distributions, Bayes Theorem, A/B Testing … etc
Database	SQL
Data Visualization	Python –Preferred- (Matplotlib, Seaborn), Tableau, PowerBI
Software Engineering	APIs, Data Structure, Algorithms, Web Development and Mobile Development

tools:

- statistical
- mathematical
- programming
- problem-solving
- data-management

OLD PLAN

Requirements	Topics	Duration (In Weeks)	Recommended Sources	Tasks
Python Programming Language (Jupyter Notebook)	Variables, I/O Commands Data Structures (Lists,Tuples,Dictionaries,Sets) Selective Structure (if - if else - if elif) Repetitive Structure (For & while loops) Functions Strings Standard Libraries (Math - csv - collection - json - random) Exception Handling Debugging Files (JSON , CSV) Object-Oriented Programming Algorithms (Big-O - Recursions - Searching - Sorting - DFS & BFS - Brute Force)	3	Intro to Python for Computer Science & Data Science (Chapter 1 to 6 & Chapter 8 to 11) **Book** Python for Data Analysis (Chapter 2 & 3) **Book** Cisco PCAP Course (Basics & Advanced) **Course** Coursera Python Course (Basics & Advanced) **Course** Codezilla & Al-Desouky & Al-Zero playlists **(Youtube)**	Tic-Tac-Toe Game (Without Machine Learning) First Week Sheets Second Week Sheets Last Week Sheets Challenging Exercises (in last week) **Discussion Each Week**
Basic Math (Pre-Caluclus & Calculus)	Number Theory Functions Summations Exponents Logarithms Euler's Number & Natural Logarithms Limits Derivatives (Partial Derivative & Chain Rule) Integrals	1	Essential Math for Data Science (Chapter 1) **Book** Thomas Calculus (Chatper 1 to 5 & Chapter 7) **Book**	Book Exercises (at least one exercise for each topic) **Discussion at the end of the week**
Linear Algebra	NumPy Library (using Python) * SciPy Library (using Python) * Vectors basics & Vector Space Linear transformation Matrices Matrix multiplications Determinants Dimensions & Rank Special Matrices Elementary row operations & Row Reduction System of equations and Inverse matrix Linear dependance Basis & Span Eigen Values & Eigen Vectors Orthogonality & Least-squares Gram-Schamidt process	3	Python for Data Science Handbook (Chapter 2) **Book** Python for Data Analysis (Chapter 4) **Book** Practical Linear Algebra for Data Science (Chapter 1 to 6 & Chapter 8 to 11) **Book** * Mathematics for Machine Learning (Chapter 2) **Book** Essential Math for Data Science (Chapter 4) **Book** Linear Algebra and its Application (Chapter 1 to 6) **Book** * Intro to Python for Computer Science & Data Science (Chapter 7) **Book**	System of Linear equations Calculator (Using python) Multiple Sheets **Discussion Each week**

OLD PLAN

Probability & Statistics	Pandas Library (using Python) * statistics Library (using Python) * StatsModels Liberary (using Python) * Populations, Samples and bias Normal Distribution Joint Distribution Union Distribution Conditional Probability Baye's Theorem Binomial Distribution Beta Distribution Poisson Distribution Descriptive Statistics (Mean - Median - Mode - Variance - Standard Deviation - Normal Distribution - The inverse CDF - Z-Score - Expected Value) The central Limit Theorem Confidence Interval Degrees of freedom Hypothesis Testing A/B Testing T-Distribution Chi-squared Distribution F-Distribution Simple Linear Regression Naive Bayes	4	Dr. Ahmed Hagag Playlist (Youtube) Python Data Science Handbook (Chapter 3) Book * Practical Statistics for Data Science (Chapter 1 to 3) Book * Python for Data Analysis (Chapter 5) Book Essential Math for Data Science (Chapter 2 & 3) Book * Mathematics for Machine Learning (Chpater 7) Book	Multiple Sheets Discussion Each Week
SQL (Database) -Optional-	SQL Datatypes Table creation Populating and Modifying Data Query Mechanics The (Select - where - from) clauses Condition Evaluation & Types Set operators & Rules Data generation, manipulation and generation Grouping & Aggregation Subqueries Querying multiple Tables	2	Learning SQL (Chapter 1 to 9) Book	
Data Visualization using Python	MatPlotLib (using Python) * Seaborn Library (using Python) * Simple Line Plots Simple Scatter Plots Visualizing Errors Density and Contour Plots Histograms, Binnings and Density Customizing Plot Legends Multiple subplots Text & Annotation Customizing Ticks Three-Dimensional Plotting Visualization with Seaborn	2	Python Data Science Handbook (Chapter 3) Book * Python for Data Analysis (Chapter 9) Book *	
Total Weeks	15			

PYTHON (JUPYTER NOTEBOOK)

Most Preferred Programming Language in 2023

Recently used in most applications of the Computer Science

Jupyter Notebook is an open-source web app and the preferred IDE of Python since it organize your code very well. Additionally, some similar platforms support it (Kaggle – Google Collab)

Interesting Fact
Jupyter is a reference to 3 Programming Languages
Julia, **Pyt**hon, **R**

jupyter covid_19_dashboard Last Checkpoint: Last Friday at 11:45 PM (unsaved changes) ✔

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code Voila

In [13]:
```
# importing libraries

from __future__ import print_function
from ipywidgets import interact, interactive, fixed, interact_manual
from IPython.core.display import display, HTML

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import folium
import plotly.graph_objects as go
import seaborn as sns
import ipywidgets as widgets
```

In [14]:
```
# loading data right from the source:
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co
recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co
country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')
```

In [15]:
```
confirmed_df.head()
```

In [16]:
```
recovered_df.head()
```

In [17]:
```
death_df.head()
```

In [18]:
```
country_df.head()
```

1. **Scientifics Computing** Libraries

Pandas
(Data structures & tools)

NumPy
(Arrays & matrices)

SciPy
(Integrals, solving differential equations, optimization)

2. Visualization Libraries

Matplotlib
(plots & graphs, most popular)

Seaborn
(plots : heat maps, time series, violin plots)

3. **Algorithmic** libraries

Scikit-learn
(Machine Learning : regression, classification,…)

Statsmodels
(Explore data, estimate statistical models, and perform statistical tests.)

TensorFlow

PyTorch

TENSORFLOW & PYTORCH

ANACONDA®

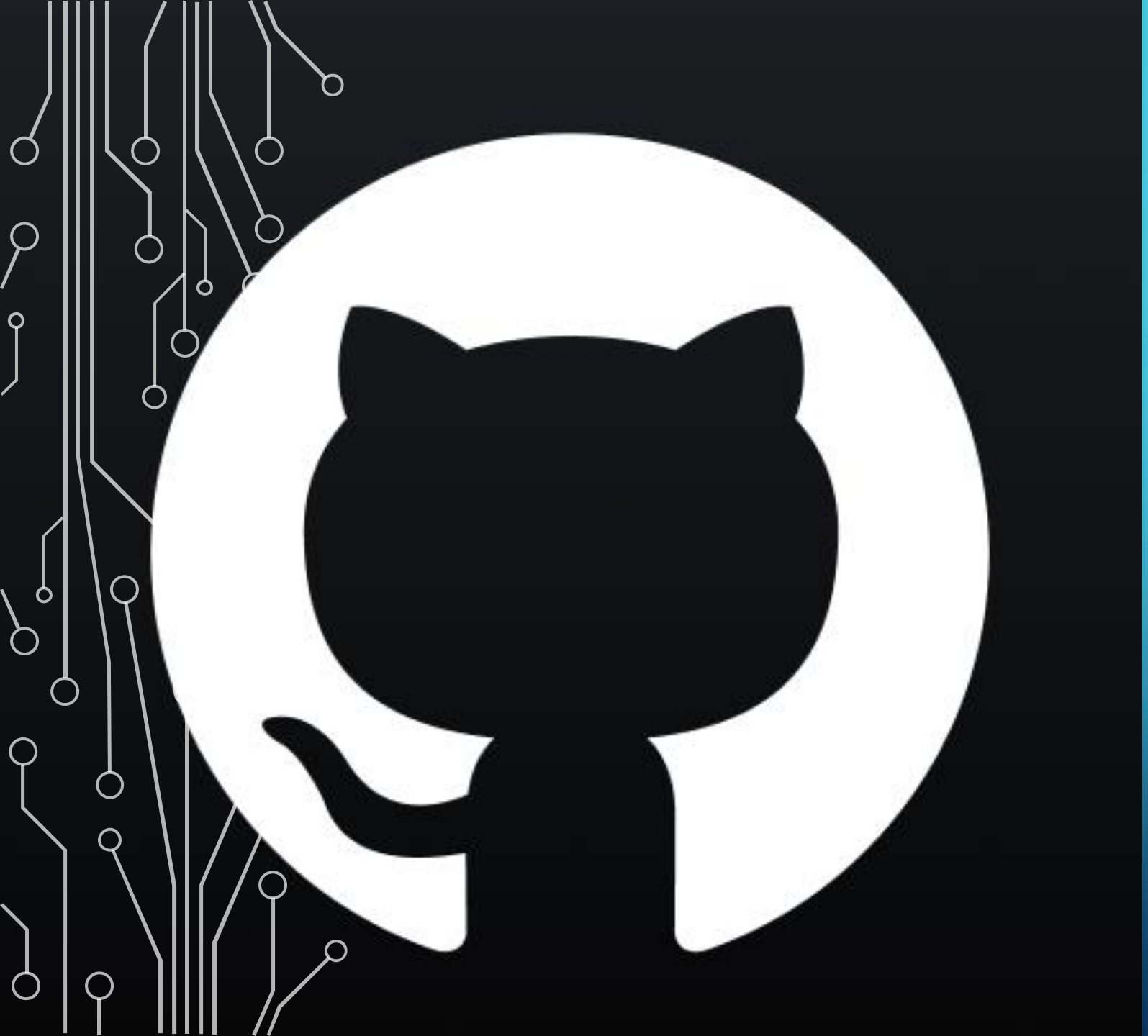
USING ANACONDA PLATFORM

SETUP YOUR KAGGLE ACCOUNT

The Biggest Data science and Machine Learning Platform in the world

It provide a **HUGE** number of Datasets freely to be used in the projects

There are some free courses for short hours of some topics such as (Data Visualization, Data Analysis, Machine Learning, Advanced Machine Learning … etc)



PREPARE YOUR **GITHUB**

HOW IT WILL BE GOING ? (EMC)

1. Discovery
2. Pre-processing data
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

DELL EMC

IN OTHER WORDS

1. Question
2. Wrangle
3. Explore
4. Draw Conclusions
5. Communicate

Step 1: Ask Questions

- Given data then ask questions, or
- Ask questions then **gather** data

Step 2: Wrangle Data

a. **Gather** data to answer question

b. **Assess** data to identify any problems in your data's quality or structure

c. **Clean** data by modifying, replacing, or removing data

UDACITY

Step 3: Perform Exploratory Data Analysis (EDA)

- **Explore then augment** data to maximize the potential of
  - analyses & visualizations & models
- **Exploring** involves:
  - finding **patterns** in data
  - **visualizing** relationships in data
  - building **intuition** about what you're working with
- **After Exploring** (**optional**)
  - **Remove Outliers:**
  - **Feature Engineering:** create better features from data

UDACITY

Step 4: Draw Conclusions (or even make predictions)

- typically approached with **ML** or **inferential statistics**

Step 5: Communicate Results

- often need to **justify** and **convey** meaning in the insights
- if your end goal is to build a system, you usually need to:
  - **share** what you've built
  - **explain** how you reached design decisions
  - **report** how well it performs
- communicate results by: report | slides | presentation | post | email | conversation
- **Data Visualization** will always be very valuable

UDACITY

IN OTHER WORDS (CONT.)

1. Ask
2. Prepare
3. Process
4. Analyse
5. Share
6. Act

1. **Ask**: Business Challenge/Objective/Question
2. **Prepare**: Data generation, collection, storage, and data management
3. **Process**: Data cleaning/data integrity
4. **Analyze**: Data exploration, visualization, and analysis
5. **Share**: Communicating and interpreting results
6. **Act**: Putting your insights to work to solve the problem

GOOGLE DATA ANALYTICS

DISCOVERY

- Learn the domain
  - Knowledge for understanding the data and the use cases of the project
  - Knowledge for the interpretation of the results
- Learn from the past
  - Identify past projects on similar issues
    - Differences, reasons for failures, weaknesses of past projects
  - Can also be projects of competitors, if reports are available

DISCOVERY (CONT.)

- Frame the problem
  - **Framing** is the process of stating the data analysis problem to be solved
  - Why is the problem important?
  - Who are the key stakeholders and what are their interests in the project?
  - What is the current situation and what are pain points that motivate the project?
  - **What are the objectives of the project?**
  - Business needs
  - Research goals
  - What needs to be done to achieve the objectives?
  - What are success criteria for the project?
  - What are risks for the project?

DOMAIN OF THE PROBLEM

Natural Language Processing

Computer Vision

Internet of Things

Optimizations

Speech Signals

DISCOVERY (CONT.)

- Begin learning the data
  - Get a high-level understanding of the data
    - May be even some initial statistics or visualizations of the data
  - Determine requirements for data structures and tools for processing the data

- Formulate hypothesis
  - Part of the "Science“ in "Data Science“
  - Should define expectations
    - "Feature X is well suited for the prediction of ...“
    - "The following patterns will be found in the data: ...“
    - "Deep learning will outperform ...“
    - "Decision trees will perform well and allow insights into ...“
  - Should be discussed with stakeholders

DISCOVERY (CONT.)

- Analyze available resources
  - Technologies
    - Resources for computation and storage
    - Licenses for analysis frameworks
  - Data
    - Is the available data sufficient for the use case?
    - Would other data be required, and could the additional data be collected within the scope of the project?
  - Timeframe
    - Scope in calendar time and person months
  - Human resources
    - Who is available for the project?
    - Is the skillset a good match for the tasks of the project?

Only start project if the resources are sufficient !

DATA PREPARATION/PREPROCESSING

- Create the infrastructure for the project
  - Usually different from infrastructure in which data is made available to you
  - Warehouse/csv-file/... ← → distributed storage that enables analysis
  - Could also be simpler, for small data sizes

- Extract – Transform – Load (ETL) the data
  - Define how to query existing database to extract required data
  - Determine required transformations of the raw data
    - Quality checking (e.g., filtering of missing data, implausible data)
    - Structuring (e.g., for unstructured data, differences in data structures)
    - Conversions (e.g., timestamps, character encodings)
  - Load the data into your analysis environment

DATA PREPARATION/PREPROCESSING (CONT.)

- ELT vs. ETL
  - Transformations can be very time-consuming for big data
  - Might not be possible without using the analysis infrastructure

→ Load raw data, transform afterwards → ELT!

- Also allows more flexibility with transformations
  - E.g., testing the effect of different transformations

- Allows access to raw data

DATA PREPARATION/PREPROCESSING (CONT.)

- Get a deep understanding of the data
  - Understand all data sources
  - E.g., what does each column in a relational database contain?
  - How can a structure be imposed on semi-/quasi-/unstructured data?
- Survey and visualize data
  - Descriptive statistics
  - Correlation analysis
  - Visualizations like histograms, density plots, pair-wise plots, etc.
- Clean and normalize data
  - Discard data that is not required
  - Can make the difference between a complex infrastructure and a single machine for analysis
  - Normalize to remove scale effects

DATA PREPARATION/PREPROCESSING (CONT.)

- Example:
  - 100 million measurements
  - 10 floating point features per measurement → 80 Bytes per measurement
  - 3 useful features ≈ 24 Bytes per measurement
  - 7.45 Gigabytes with all features, 2.23 Gigabytes with only useful features

  → Can use my laptop for cleaned data without problems

MODEL PLANNING

- Determine methods for data analysis
- Should be well-suited to meet objectives
  - Often determines the type of method
    - Classification, regression, clustering, association mining, ...
  - Other factors can also restrict the available methods
    - For example, if insight is important, **"Blackbox"** methods cannot be used
- Should be well-suited for the available data
  - Volume, structure, ...

A blackbox method is a method where you only get results, but do not really understand why the output is computed that way.
A whitebox method also explains why the output is as it is.

MODEL PLANNING (CONT.)

- Methods for data analysis may cover
  - Feature modeling, e.g., for text mining
  - Feature selection, e.g., based on information gain, correlations, etc.
  - Model creation, e.g., different models that may address the use case
  - Statistical methods, e.g., for the comparison of results
  - Visualizations, e.g., for the presentation of results
- Split data into different data sets
  - Training data, validation data, test data
  - "Toy“ data for local use in case of big data
    - Same structure, but very small

MODEL BUILDING

- Perform the analysis using the planned methods
  - Often iterative process!
- Separate phase, because this can be VERY time consuming
  - Use toy examples for model planning
  - Use real big data set with potentially lots of hyper parameters for tuning during model building
- Includes the calculation of performance indicators

COMMUNICATE RESULTS

- Main question: Was the project successful?
- Compare results to hypothesis from the discovery phase
- Identify the key findings
- Try to quantify the value of your results
  - Business value, e.g., the expected Return On Investment (ROI)
  - Advancement of the state of the art
- Summarize findings for different audiences

OPERATIONALIZE

- Implement results in operation
  - Only in case of successful projects
- Should run a pilot first
  - Determine if expectations hold during the practical application
  - All kinds of reasons for failures
  - Rejection by users, shift in data reduces model performance, ...
- Define a process to update and retrain model
  - Data gets older, models get outdated
  - Data driven models should be updated regularly
  - Process is required

REMEMBER !

Data Science is IMPOSSIBLE TO MASTER

QUESTIONS


THE END