DATA SCIENCE Workshop

Are you one of the many who dreams of becoming a data scientist?

IF YES!

Then start today, the world of data needs you.

1st session

Created by: Moataz Elmesmary

@cisteam15

Agenda for all days

DAY 01

.Intro to Data Science
.Python Basics

DAY 02

.Pandas & Numpy
.Data Clean

DAY 03

.EDA & Visualizing
.Starting with your first project

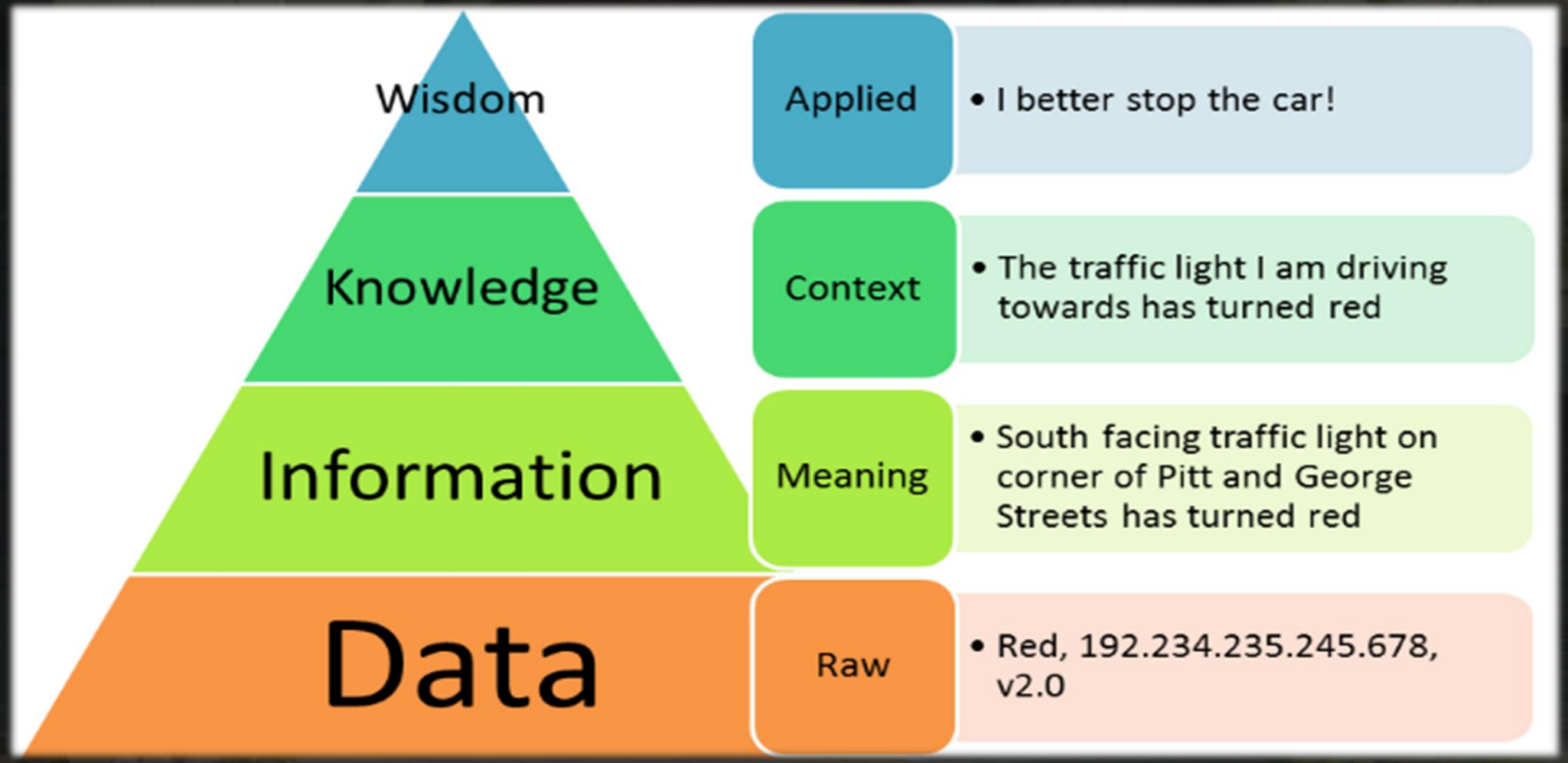
DAY 04

.Intro to Machine Learning
.Complete your first project

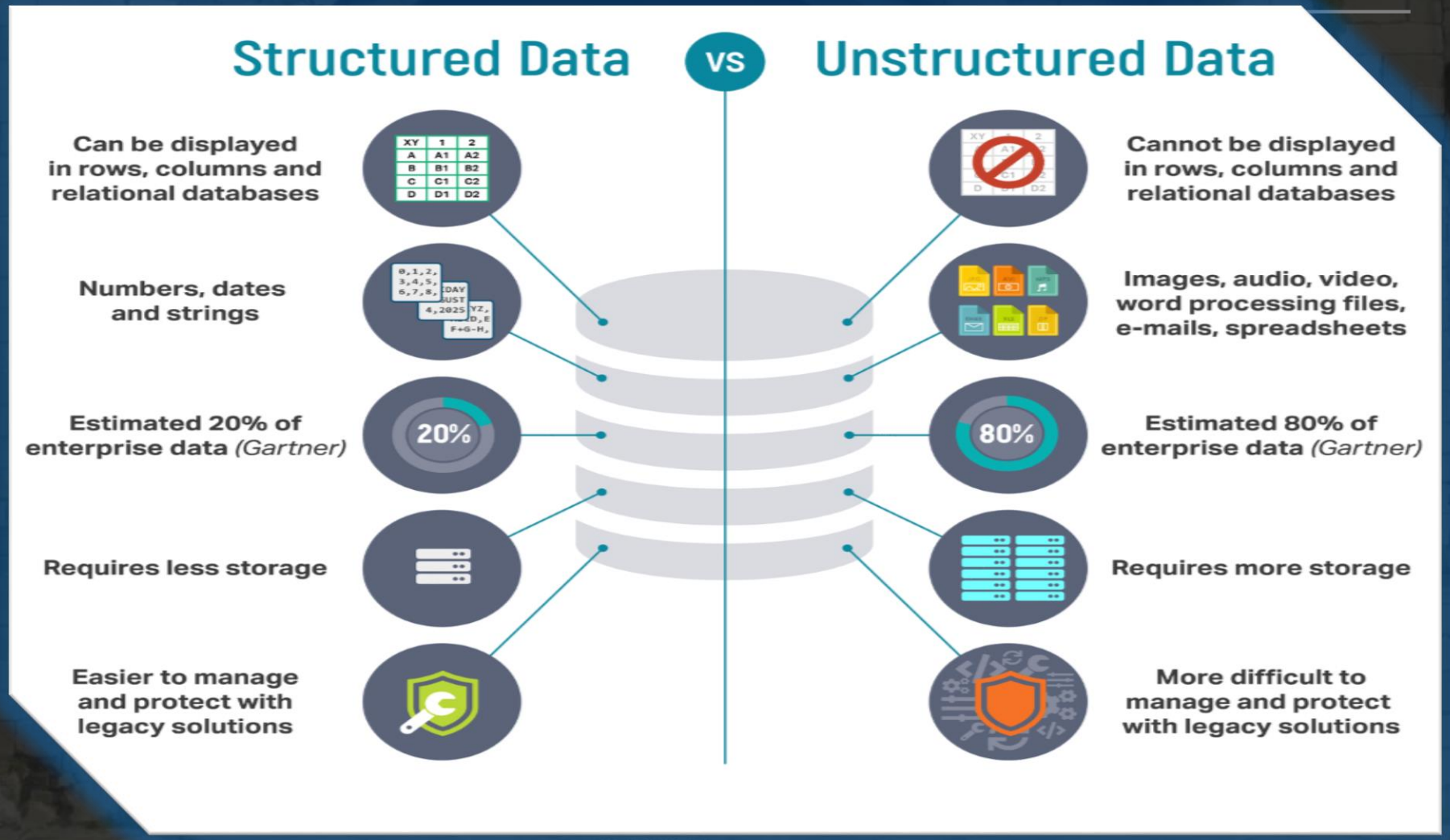
Agenda for first day

- About Data Science
- Projects Life Cycle
- Common Mistakes
- Libraries
- Roles – Job Description
- Important sites
- Hard & Soft skills
- RoadMap
- Python Basics

DIKW Pyramid



Data Types

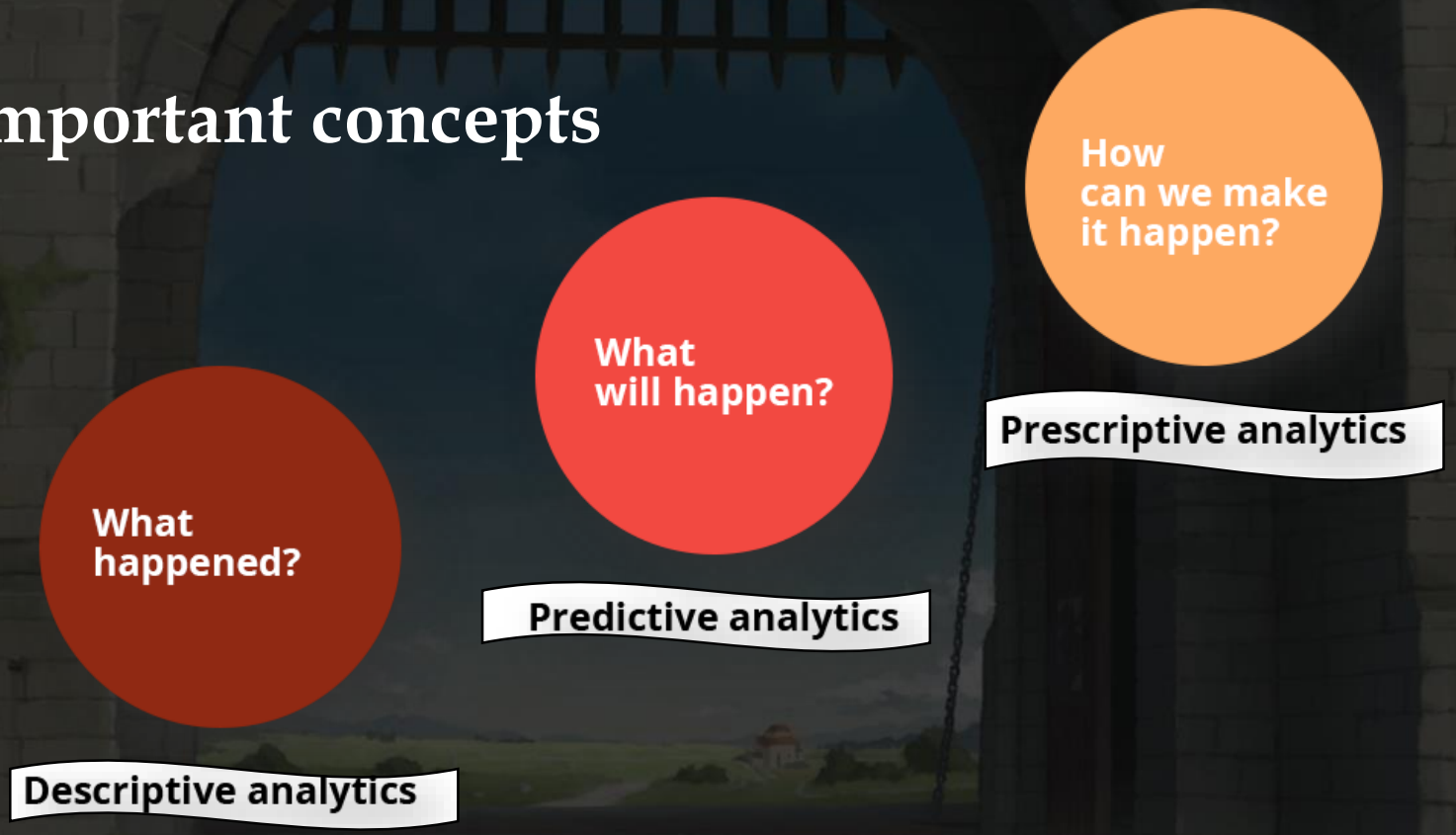

Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

Why Data Science?

1. **It’s in Demand**
2. **Very Challenging**
3. **Abundance of Positions**

What`s Data Science

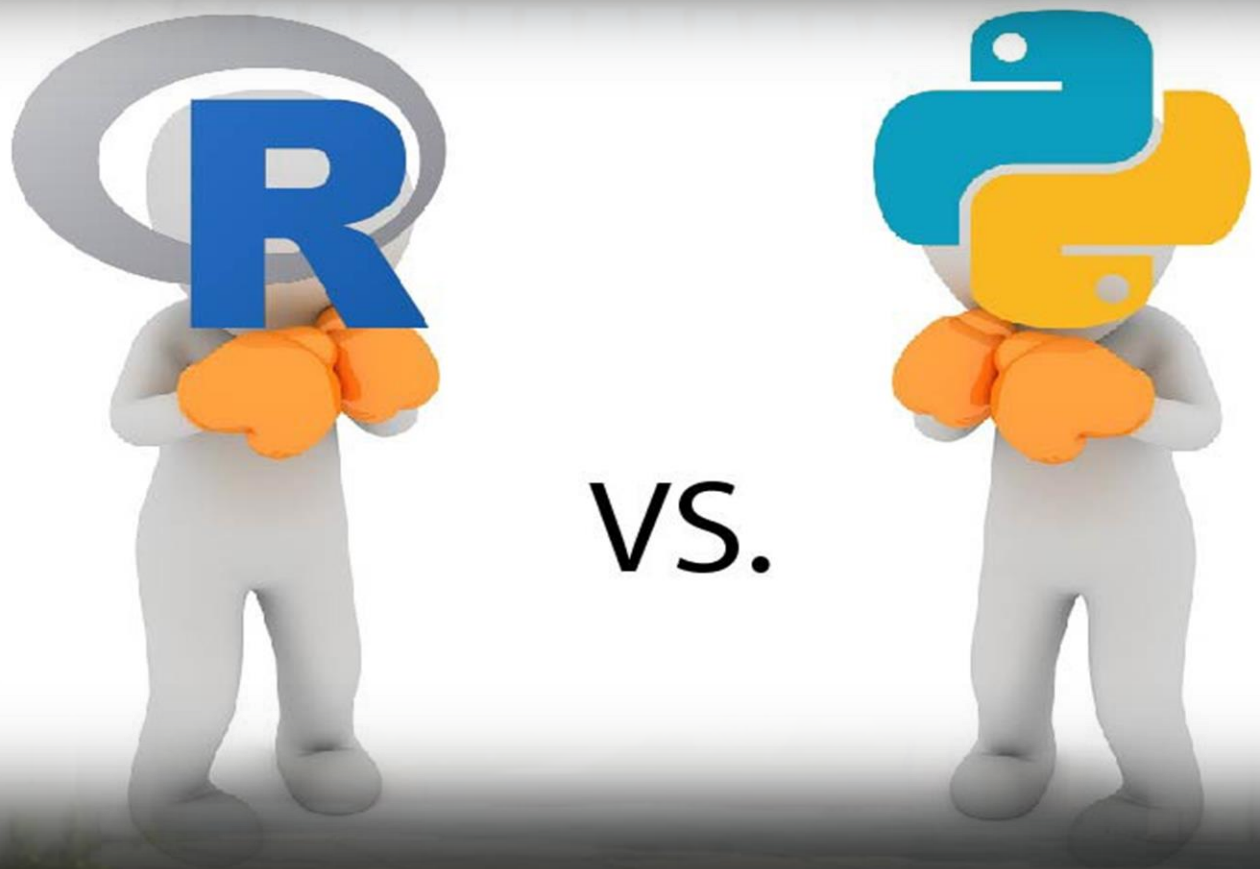
+ important concepts



Important sites

1. **Kaggle**
2. **GitHub**
3. **DataCamp**
4. **Khan Academy**
5. **Coursera**
6. **HackerRank**

Which is better to choose?



Common Mistakes

1. When When When???????
2. Implementation is a secret beast.
3. Learning everything at the same time. Like rushing for ML without knowing programming.
4. Never Give up (advice)
5. Neglecting communications skills.
6. Waiting till you get a new laptop/desktop.

Data Science Job Titles

THERE ARE VARIOUS ROLES OFFERED TO A DATA SCIENTIST LIKE:

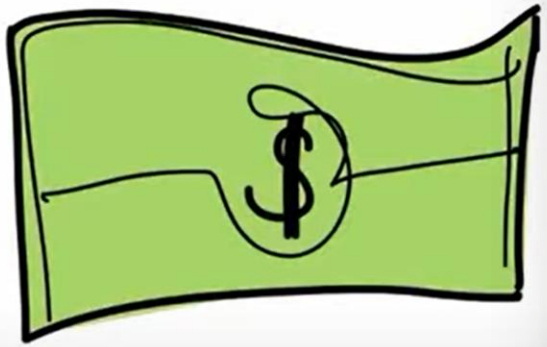
DATA ANALYST

MACHINE LEARNING ENGINEER

DEEP LEARNING ENGINEER

DATA ENGINEER

DATA SCIENTIST



$95,000 TO $165,000

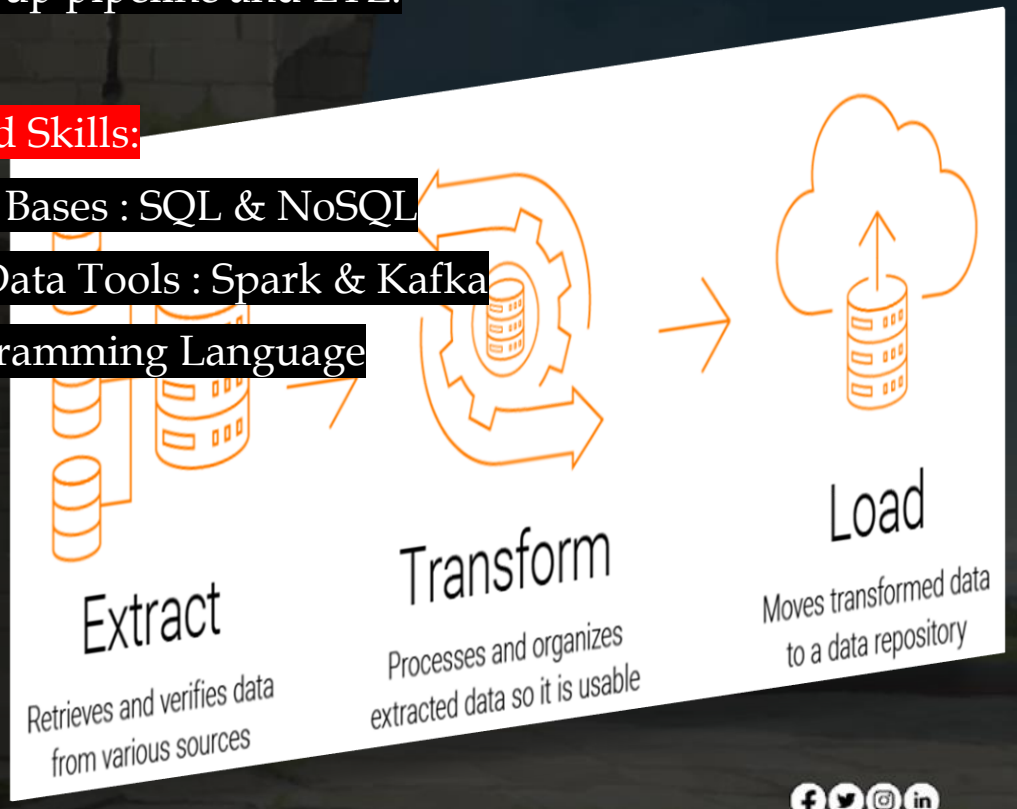
Data Science Job Titles

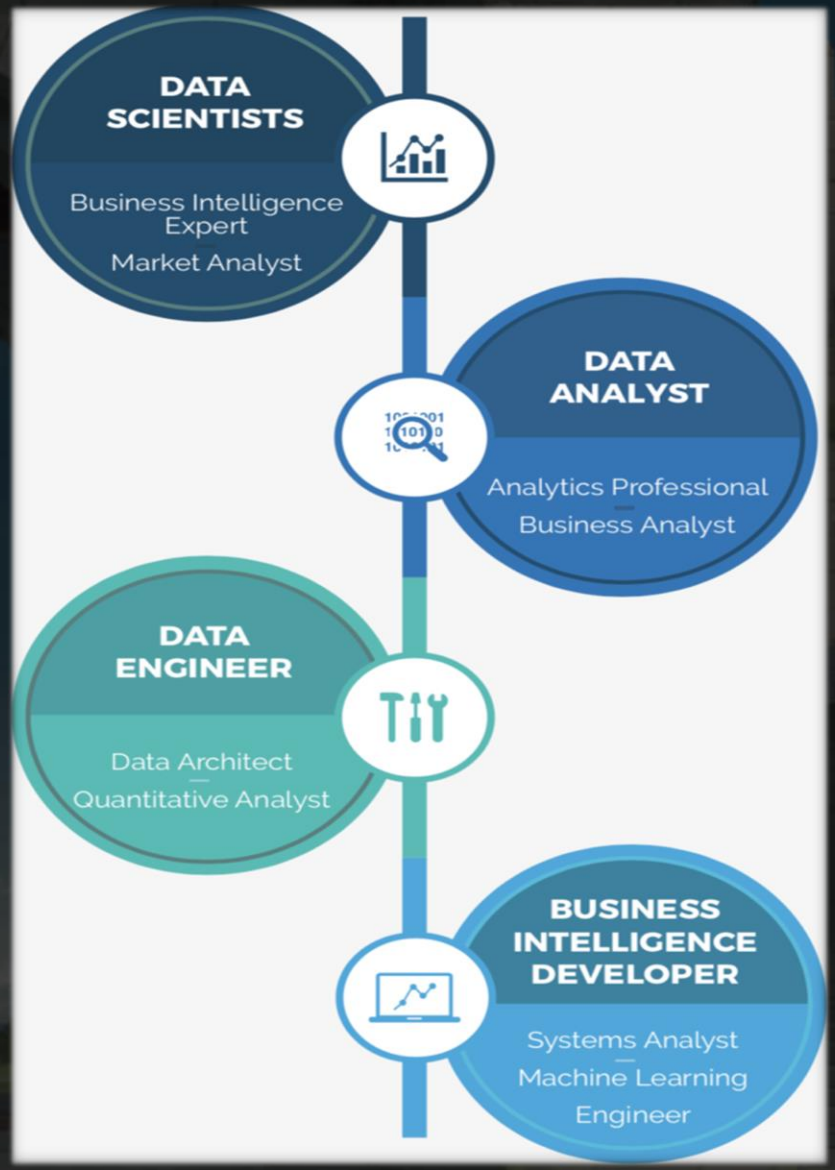
Data Engineer:

- data collection – part of (data pre-processing)
- brings data and put it in databases.
- build up pipeline and ETL.

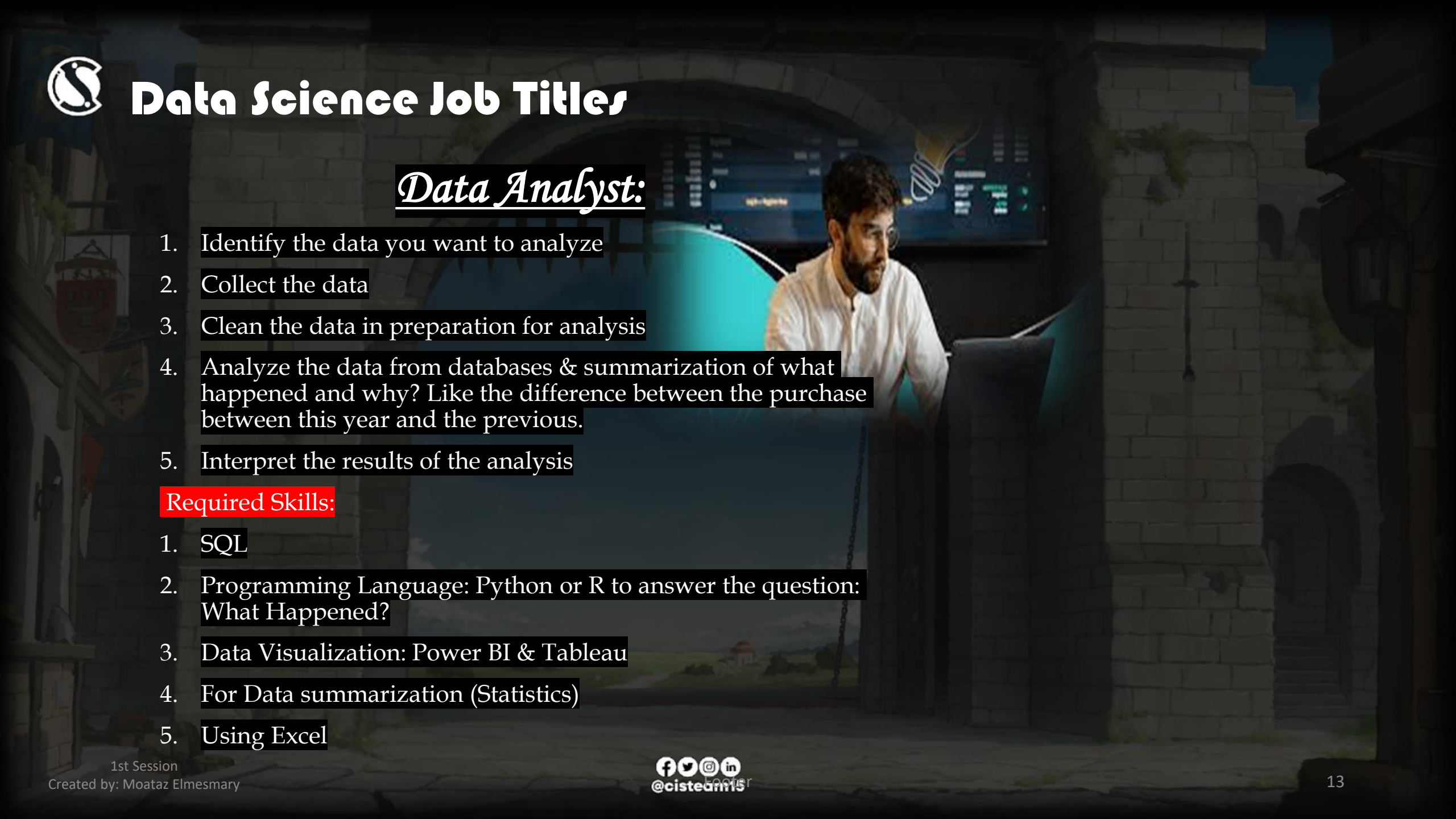
Required Skills:

1. Data Bases : SQL & NoSQL
2. Big Data Tools : Spark & Kafka
3. Programming Language





Data Science Job Titles



Data Analyst:

1. Identify the data you want to analyze
2. Collect the data
3. Clean the data in preparation for analysis
4. Analyze the data from databases & summarization of what happened and why? Like the difference between the purchase between this year and the previous.
5. Interpret the results of the analysis

Required Skills:

1. SQL
2. Programming Language: Python or R to answer the question: What Happened?
3. Data Visualization: Power BI & Tableau
4. For Data summarization (Statistics)
5. Using Excel

Data Science Job Titles

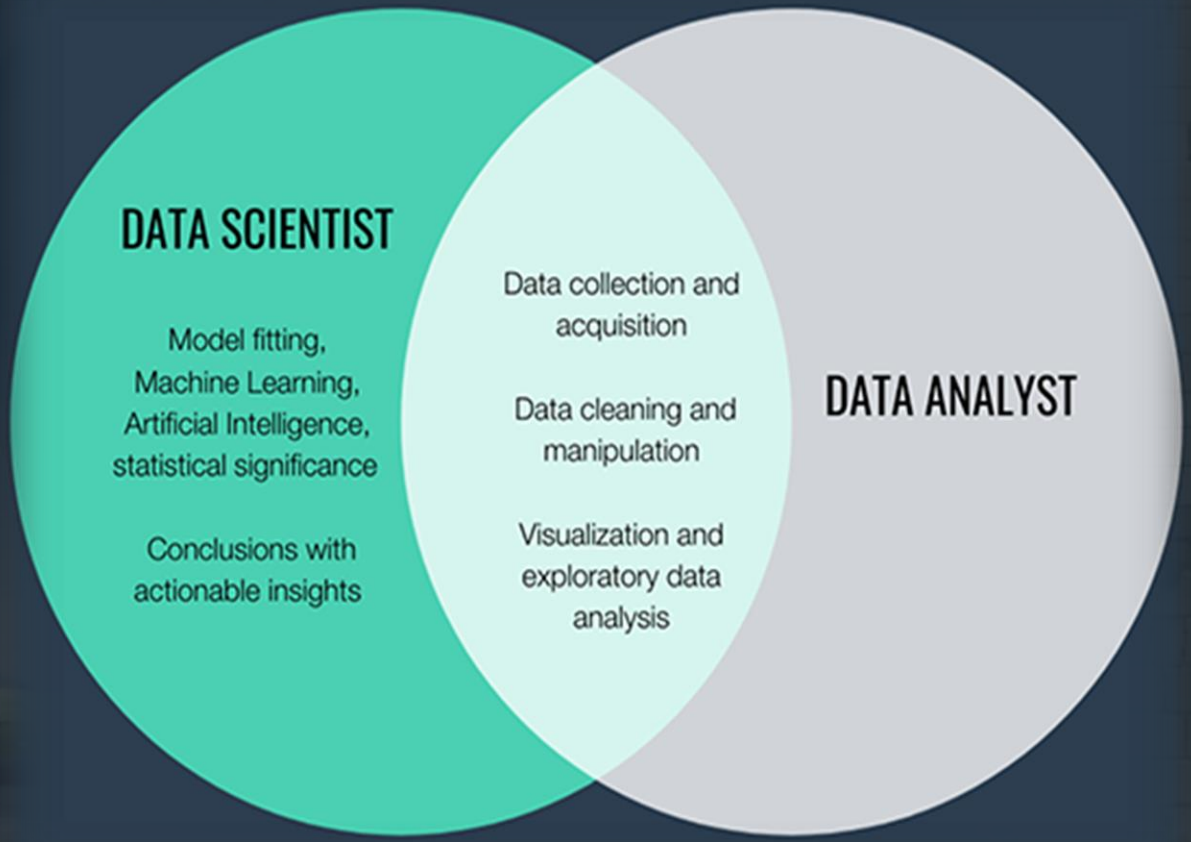
Data Scientist:

Data analyst looks at the past and answers " What happened and Why from the existent data.

- + Machine learning, data modeling

Required Skills:

1. Programming Language: Python or R
2. SQL
3. Data Visualization
4. Statistics



Data Science Job Titles

Machine Learning Engineer:

- Designing ML systems.
- Researching and implementing ML algorithms and tools.

Required Skills:

1. Solid programming background
2. Data Science skills
3. Advanced math and statistics skills, surrounding subjects such as linear algebra, calculus and Bayesian Statistics.
4. Experience in working with ML frameworks.
5. Experience working with ML libraries and packages.

IMPORTANT EDUCATIONAL OBJECTIVES

Can be divided into two categories: *Soft Skills* and *Hard Skills*.

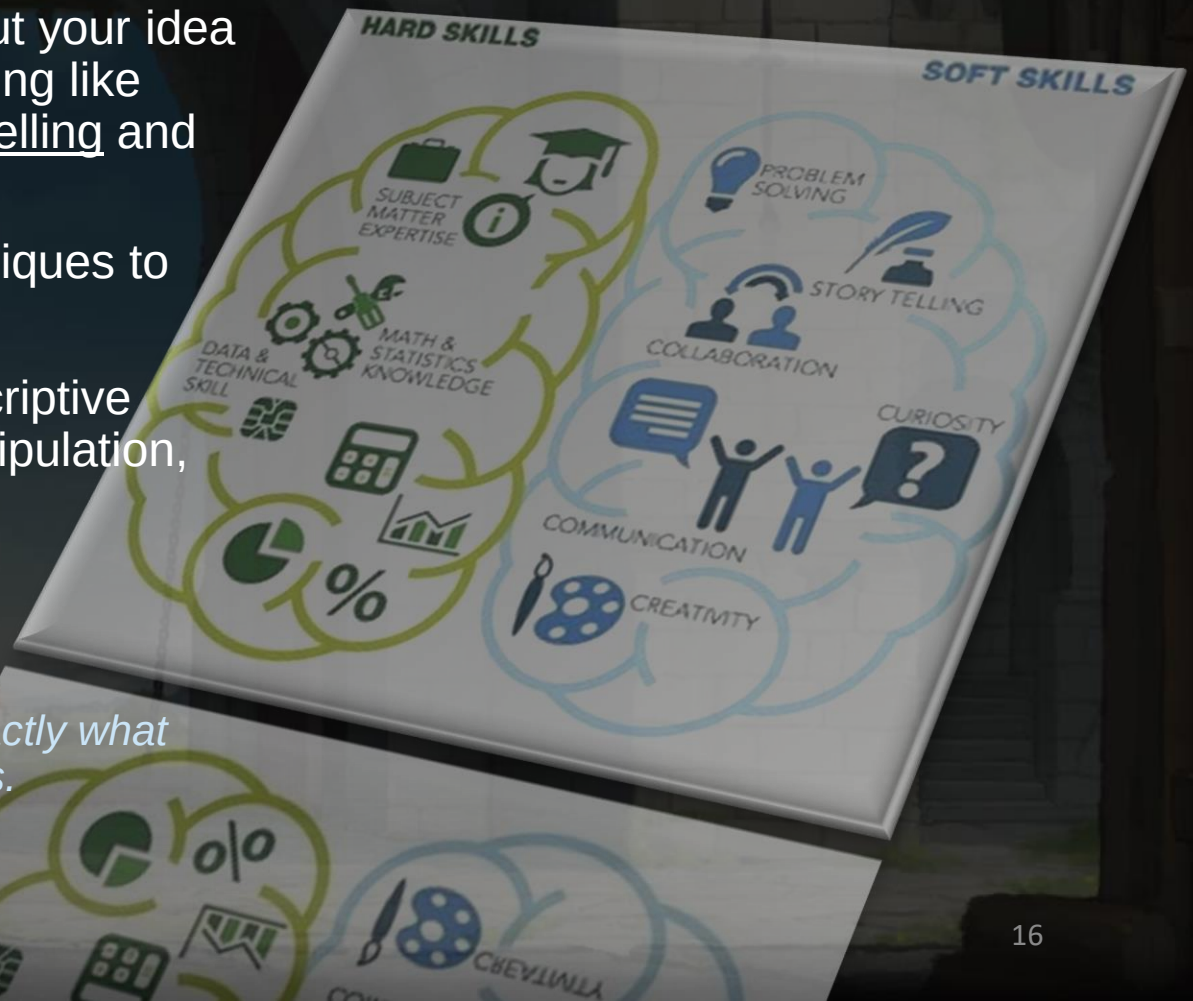
Soft skills include behavioral skills that help you put your idea on the table with sufficient explanation and convincing like *problem-solving*, *critical thinking*, Creativity, Story Telling and *Curiosity*.

Hard skills teach you to use all the tools and techniques to derive results from huge data sets like:

Foundation blocks (Programming language, Descriptive analytics and Visualization, Data handling and manipulation, Data wrangling and summarization),

statistical tools, algorithms, and machine learning.

A perfect amalgamation of soft skills and hard skills is exactly what enterprises are looking for in their in-house data scientists.

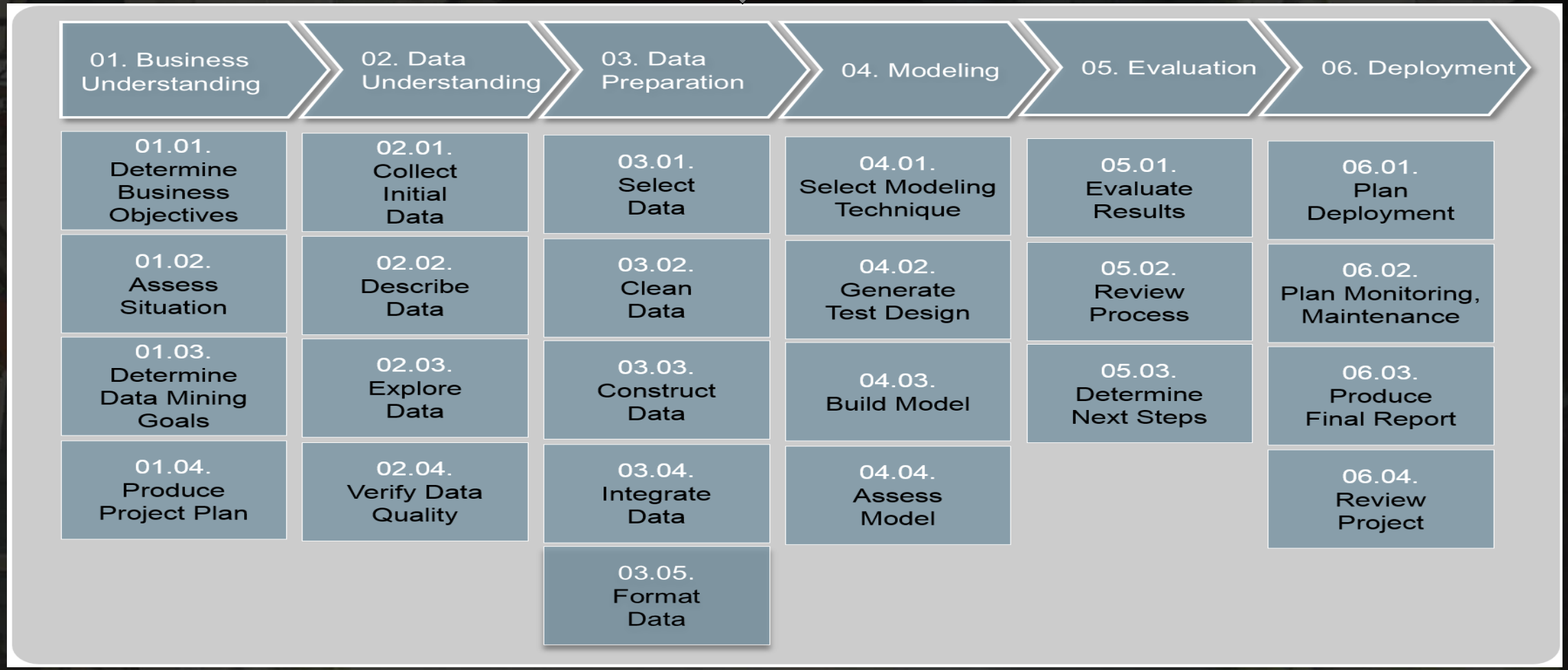


Syllabus and Roadmap

https://docs.google.com/document/d/1FgFLspZZ_wXbTVx_0Z5Akj6csHBCb2suhlofA3KG268/edit#heading=h.b5ucstccunfa

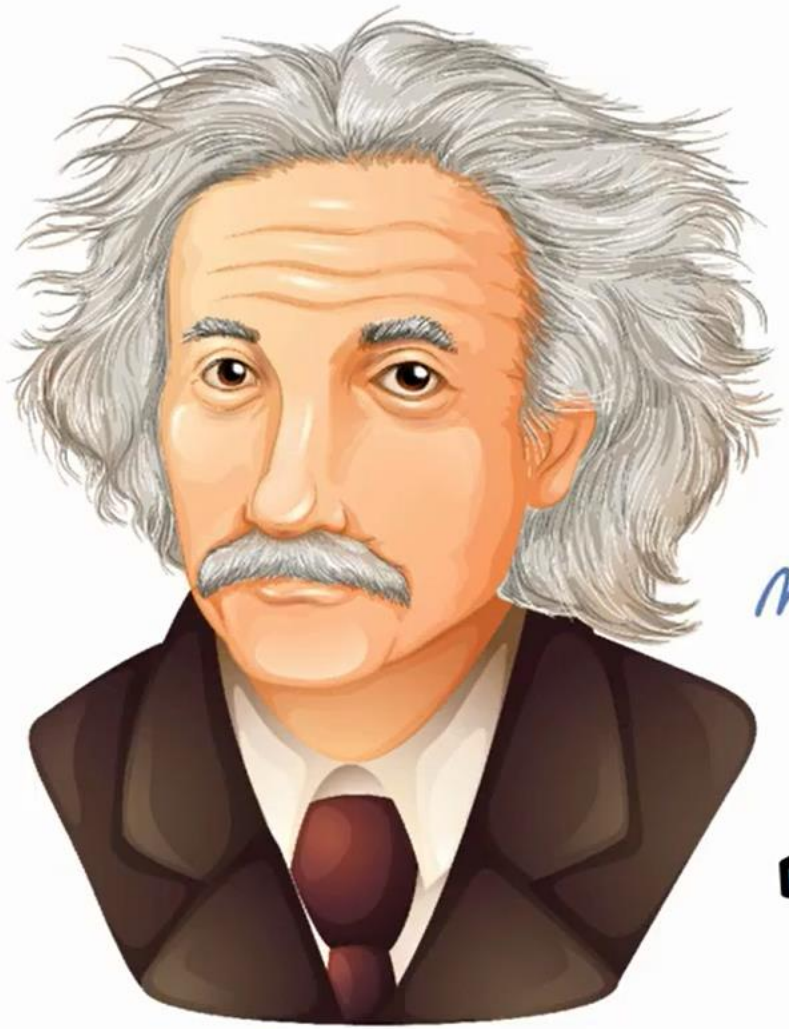
https://docs.google.com/document/d/1FgFLspZZ_wXbTVx_0Z5Akj6csHBCb2suhlofA3KG268/edit#

Data Science project life cycle



Business Understanding

Critical



59 Minutes

1 Hour

Reslove

1 Minute

Define

Well Defined Problem

Contains Solution

Obvious and Straightforward

Savings

Time

Money

Resources

1 BUSINESS PROBLEM

WHY?....WHY?....WHY?....

ONE OF THE MANY TRAITS OF

A GOOD DATA SCIENTIST!

2 DATA ACQUISITION

- WEB SERVERS
- LOGS
- DATABASES
- API'S
- ONLINE REPOSITORIES

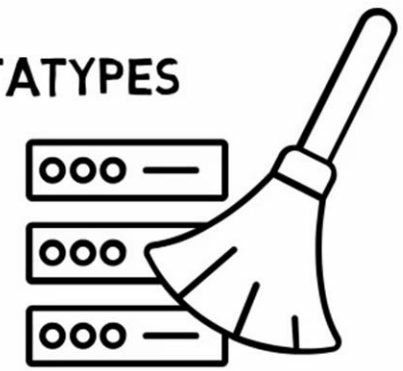
DATA PREPARATION

DATA CLEANING

INCONSISTENT DATATYPES

MISSPELLED ATTRIBUTES

MISSING AND DUPLICATE VALUES



TRANSFORMATION

talend

Informatica

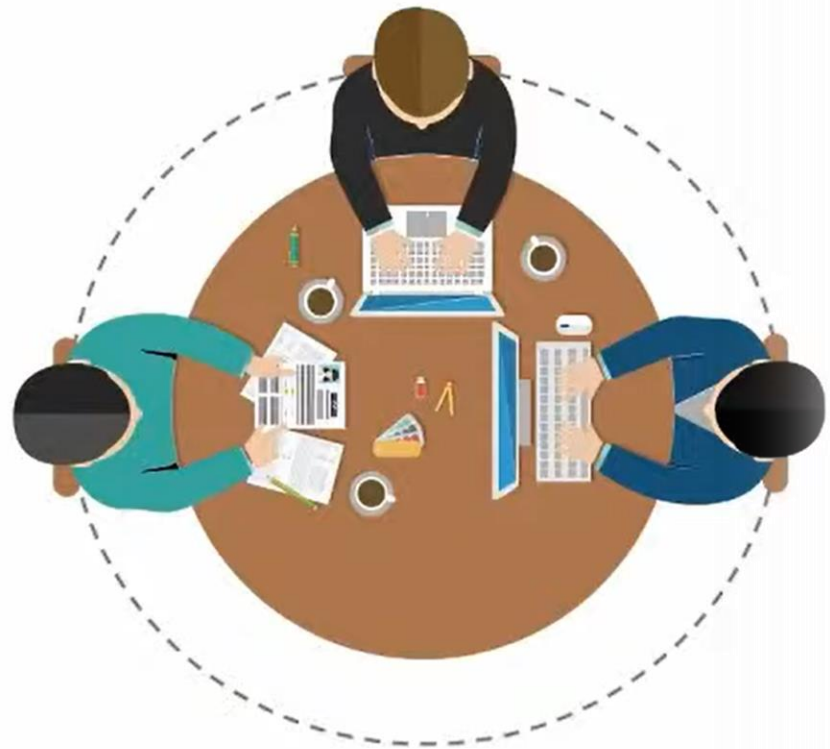
Data Preparation - Life cycle

Data Cleaning
Correcting inconsistent data by filling out missing values and smoothing out noisy data

Data Reduction
Using various strategies, reducing the size of data but yielding the same outcome

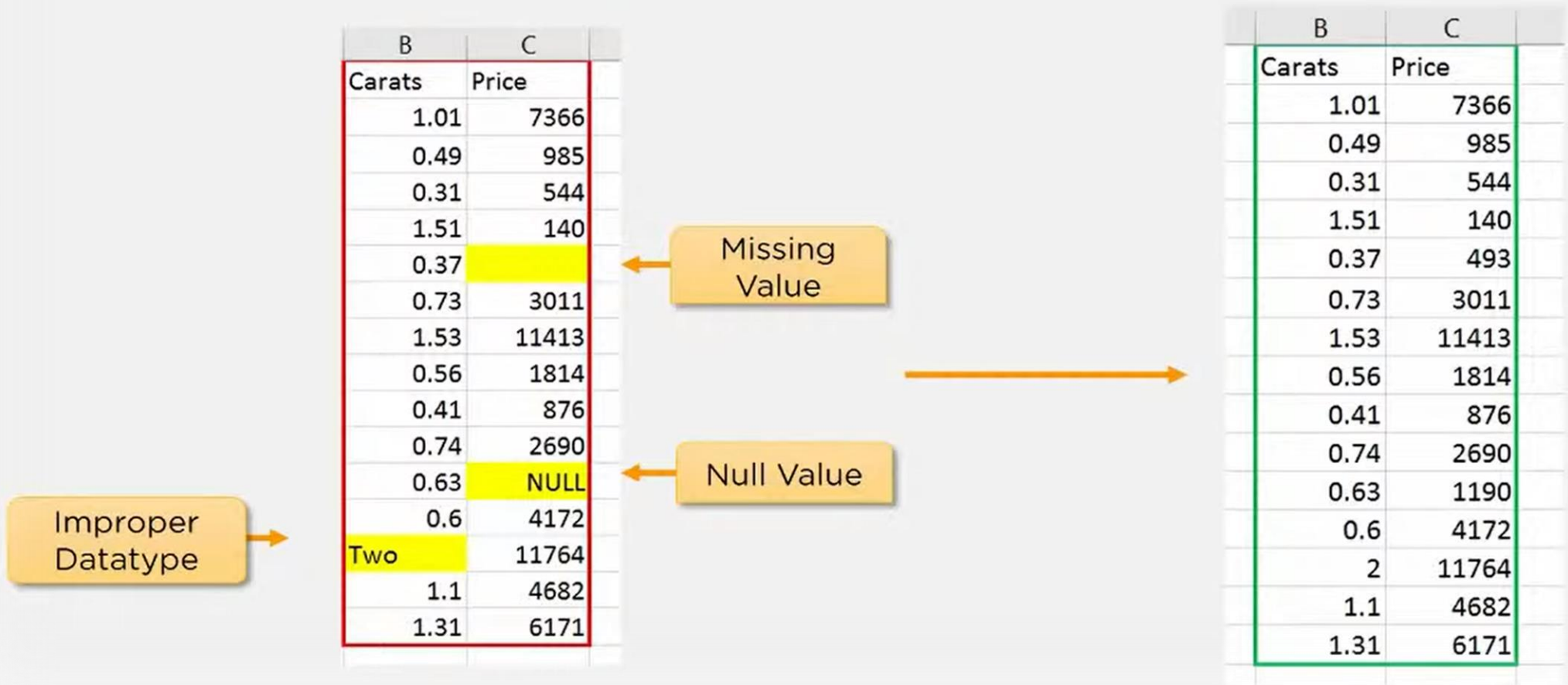
Data Transformation
It involves normalization, transformation and aggregation of data using ETL methods

Data Integration
Resolving any conflicts in the data and handling redundancies



Data Preparation - Use Case

Data preparation : Make the data clean and valuable.



Merging data

	group
employee	
Bob	Accounting
Jake	Engineering
Lisa	Engineering
Sue	HR

	hire_date
employee	
Lisa	2004
Bob	2008
Jake	2012
Sue	2014

	group	hire_date
employee		
Bob	Accounting	2008
Jake	Engineering	2012
Lisa	Engineering	2004
Sue	HR	2014

	Pet	Color	Eyes
0	Cat	Brown	Black
1	Dog	Golden	Black
2	Dog	Golden	Black
3	Dog	Golden	Brown
4	Cat	Black	Green

	Pet	Color	Eyes
0	Cat	Brown	Black
1	Dog	Golden	Black
3	Dog	Golden	Brown
4	Cat	Black	Green

Drop duplicates

Data Preparation - Use Case

Ways to fill missing data values:

If dataset is huge, we can simply remove the rows with missing data vales. It is the quickest way.
i.e. we use the rest of the data to predict the values.

We can substitute missing values with mean of rest of the data using pandas' dataframe in Python.

i.e. df.mean()
df.fillna(mean)

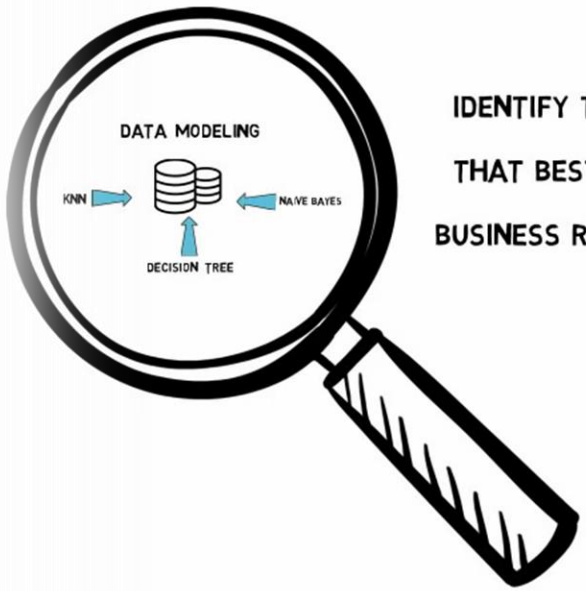
④ EXPLORATORY DATA ANALYSIS

DEFINES AND REFINES

THE SELECTION OF FEATURE

VARIABLES THAT WILL BE USED

IN THE MODEL DEVELOPMENT



IDENTIFY THE MODEL THAT BEST FITS THE BUSINESS REQUIREMENT

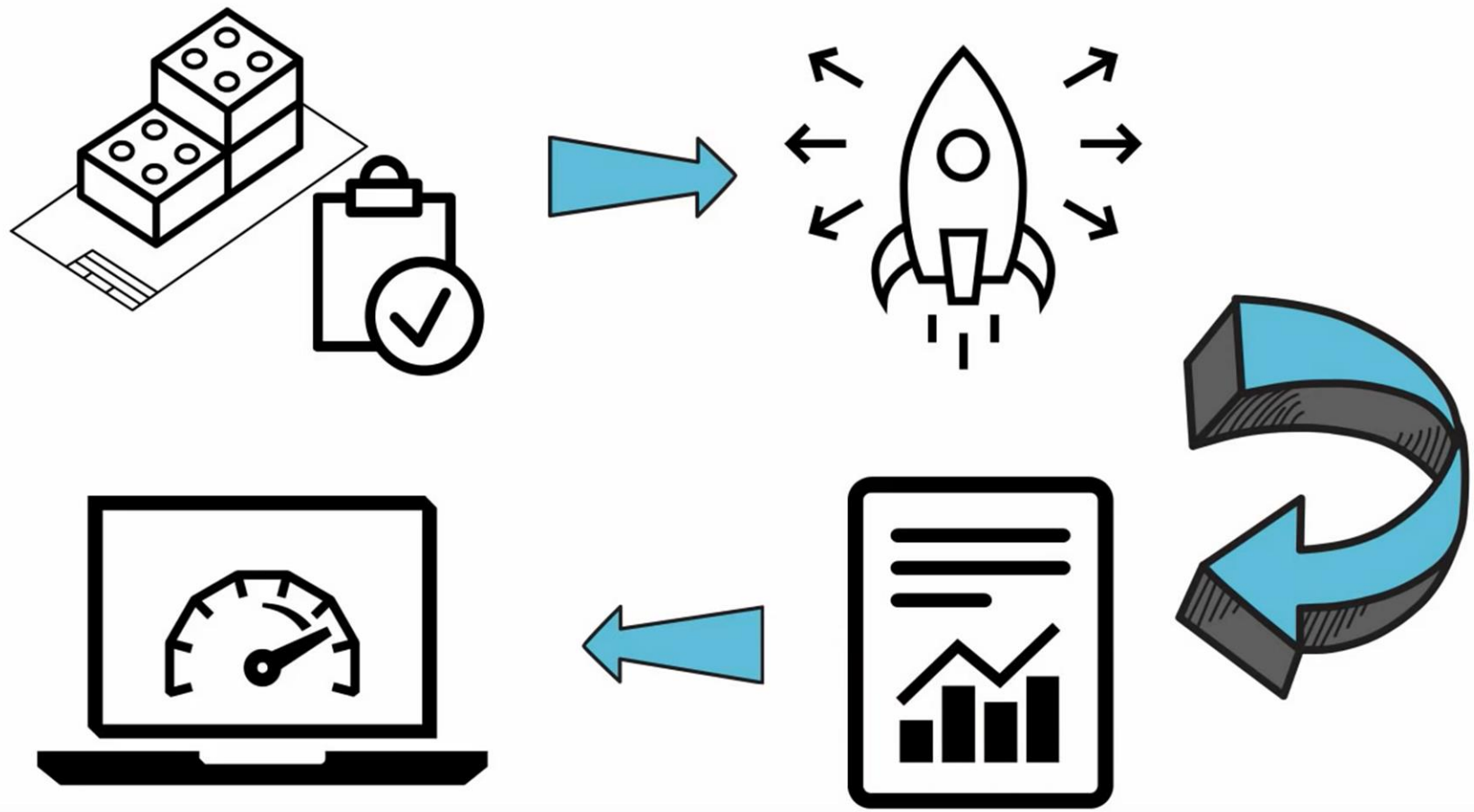
TRAINS THE MODELS ON THE TRAINING DATASET AND TEST

SELECT THE BEST PERFORMING MODEL

6 VISUALIZATION AND COMMUNICATION

tableau Power BI QlikView

7 DEPLOYS AND MAINTAINS



Data Science Applications

E-commerce

- Identifying Consumers
- Recommending Products
- Analyzing Reviews

Manufacturing

- Predicting Potential Problems
- Monitoring Systems
- Automating Manufacturing Units
- Maintenance Scheduling
- Anomaly Detection

Banking

- Fraud Detection
- Credit Risk Modeling
- Customer Lifetime Value

Healthcare

- Medical Image Analysis
- Drug Discovery
- Bioinformatics
- Virtual Assistants

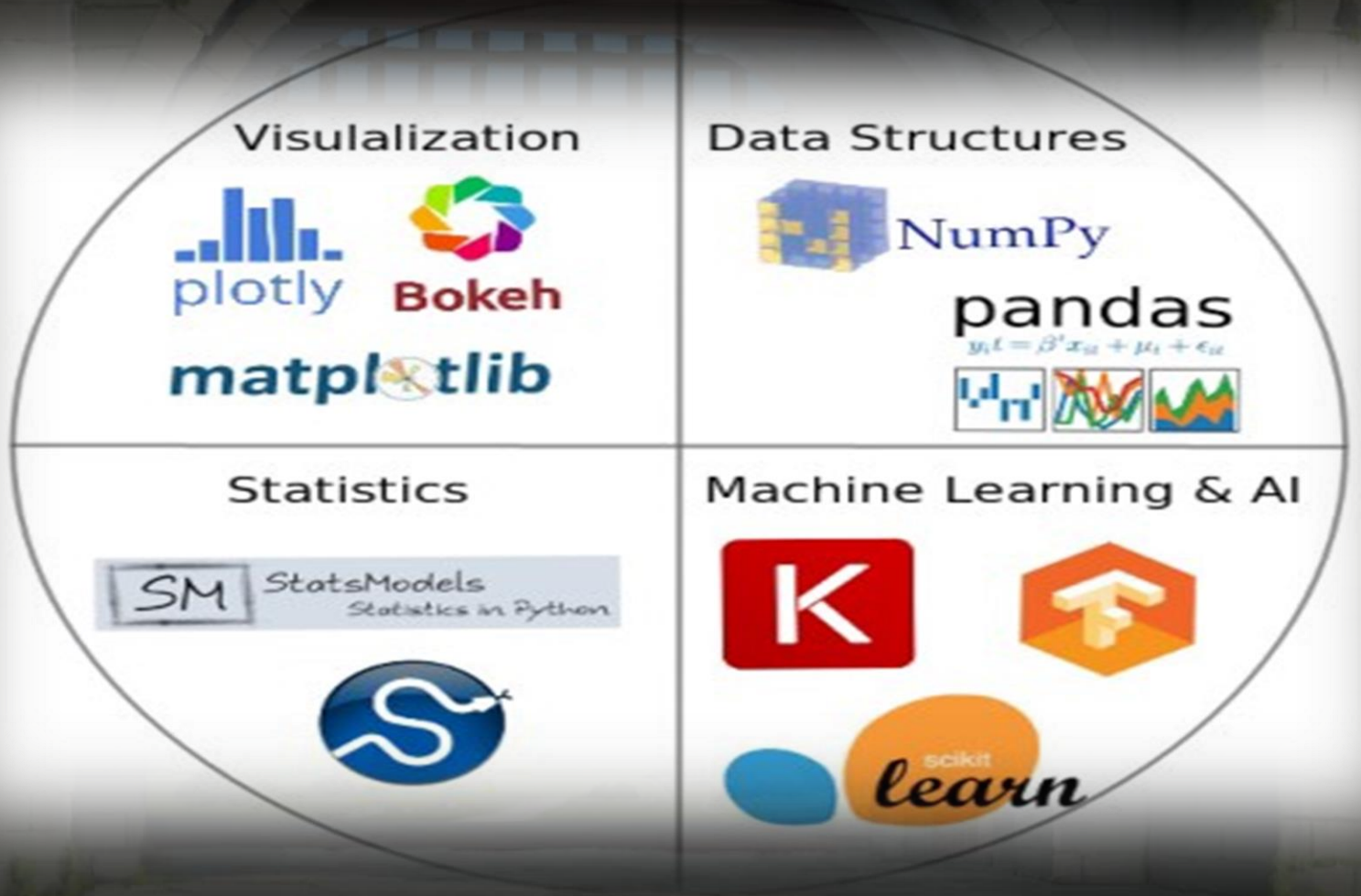
Transport

- Self Driving Cars
- Enhanced Driving Experience
- Car Monitoring System
- Enhancing the safety of passengers

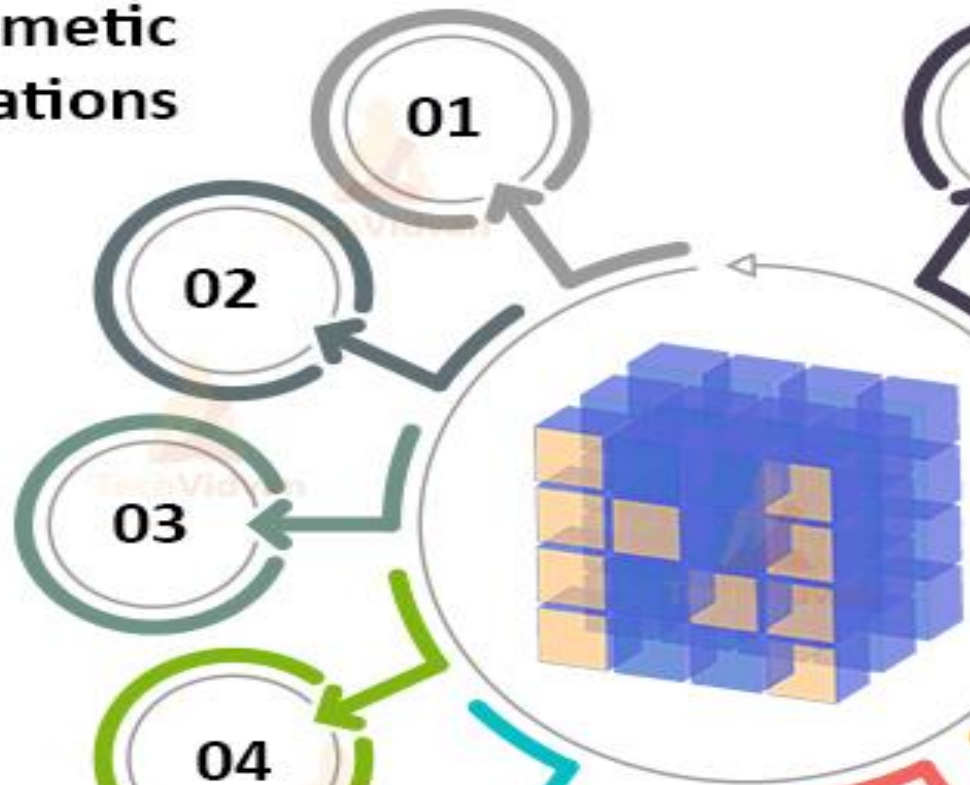
Finance

- Customer Segmentation
- Strategic Decision Making
- Algorithmic Trading
- Risk Analytics

Important Libraries



Uses of NumPy



01 Arithmetic operations

02 Statistical operations

03 Bitwise operators

04 Copying & viewing arrays

05 Stacking

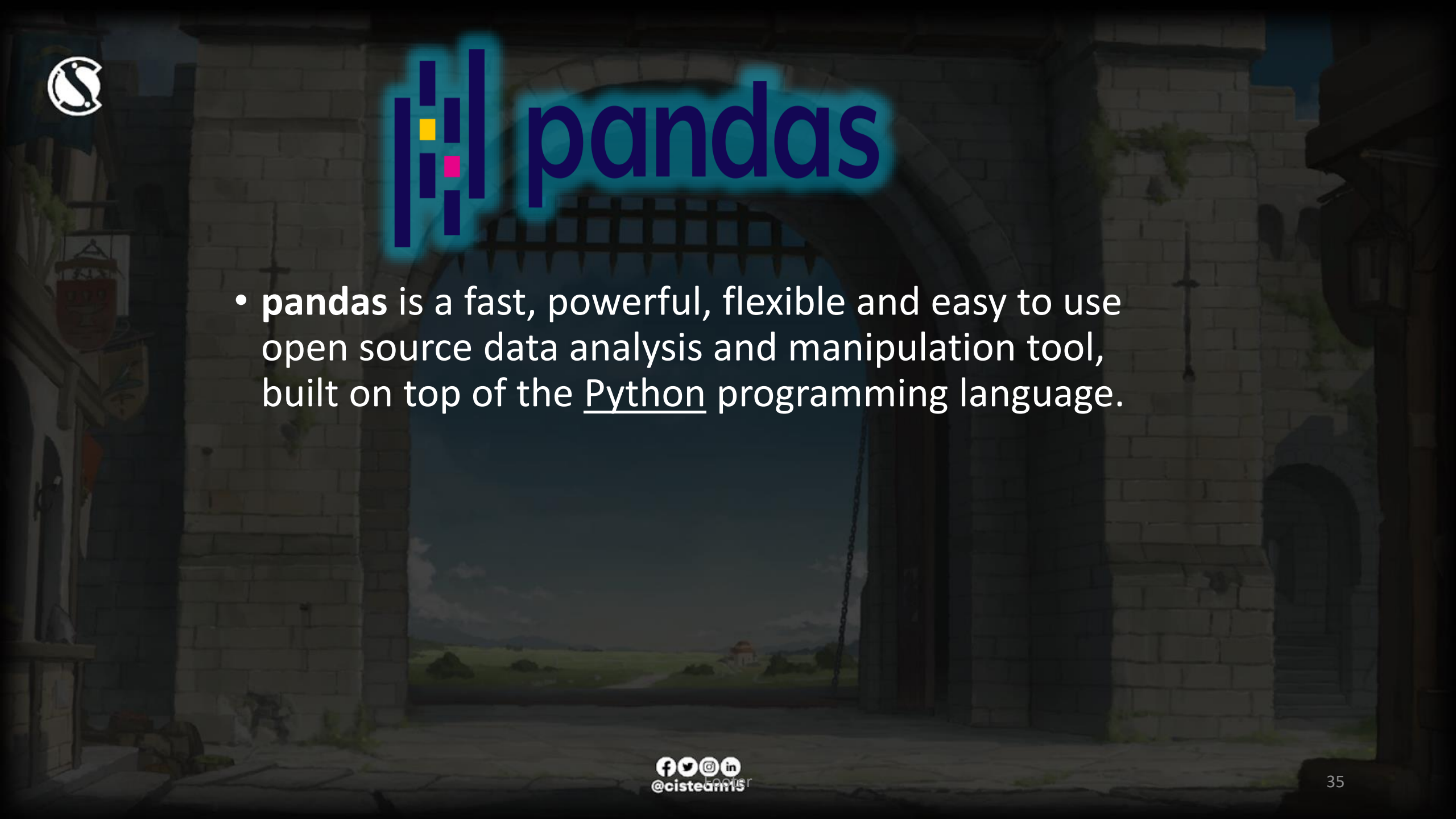
06 Matrix Operations

07 Linear algebra

08 Broadcasting

09 Mathematical operations

10 Searching, sorting & counting



- **pandas** is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

Machine Learning TALK

Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)

- ❑ Given email labeled as spam/not spam, learn a spam filter.
- ❑ Given a set of news articles found on the web, group them into sets of articles about the same stories.
- ❑ Given a database of customer data, automatically discover market segments and group customers into different market segments.
- ❑ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.