Python for Data Science

Fabiyi Opeyemi

scenarioacademy

What's the fuss anyway?

- Data — the new oil? *A 2017 post by The Economist said "the world's most valuable resource is no longer oil, but data."*
- 90% of the world's data was generated in the last two years.
- The 5 most valuable companies in the world have data at their core: Alphabet (Google's parent company), Apple, Amazon, Facebook and Microsoft.



Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms. It unifies Statistics, Mathematics, Computer Science and other fields to understand and analyze actual phenomena with data.

- ***Statistics:*** Collection, analysis, interpretation, presentation and organization of data using Mathematical methods.
- ***Artificial Intelligence (AI):*** Allowing machines the ability to mimic cognitive human functions like learning and problem solving.
- ***Machine Learning (ML):*** a branch of AI that gives machines the ability to 'learn" with data, without being explicitly programmed.
- ***Big Data:*** Datasets so voluminous and complex that they cannot be analyzed with traditional data analysis software.

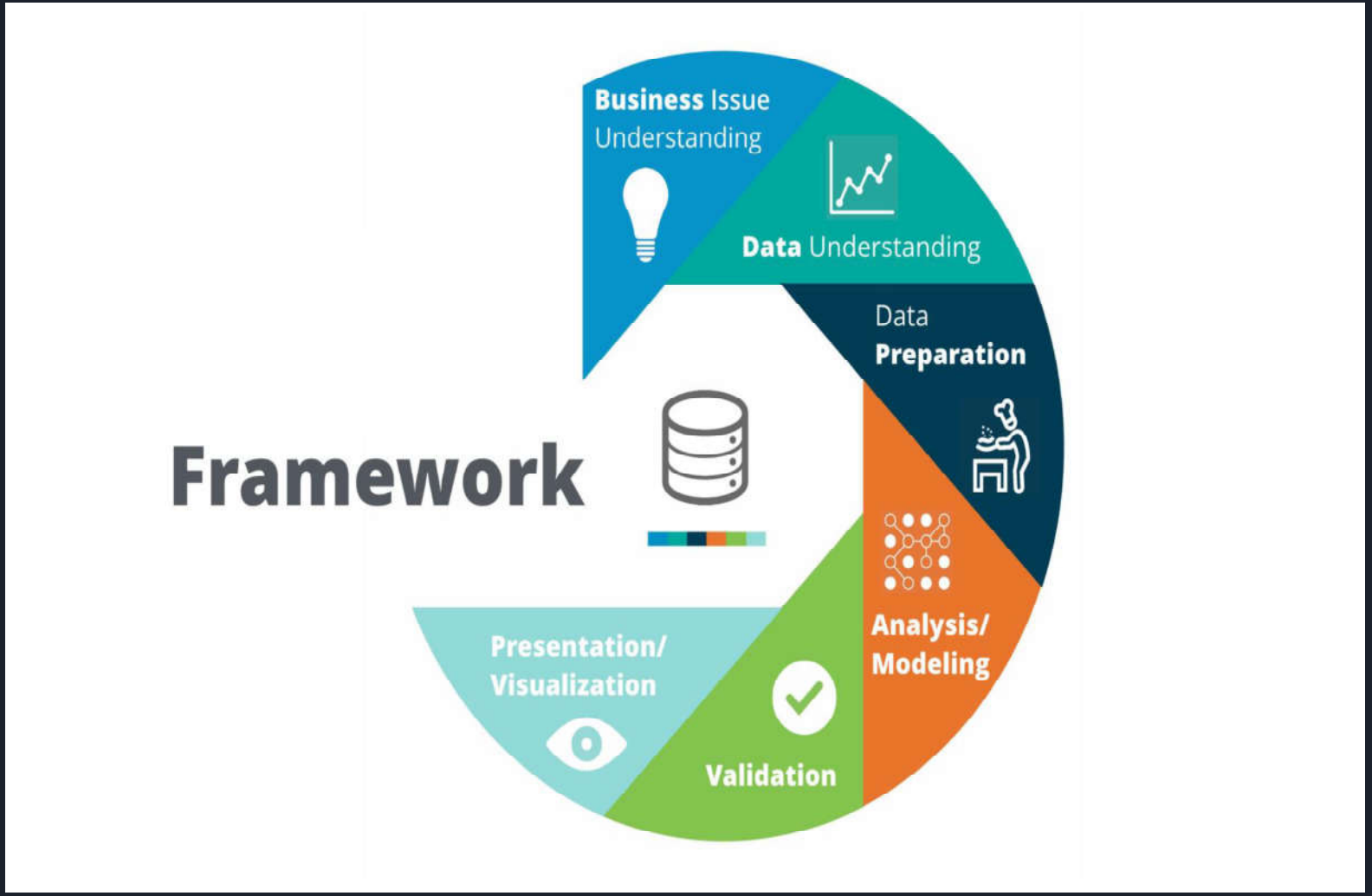
Data Analytics

Inspecting, cleansing, transforming, and modeling data to discover useful information, inform conclusions and support decision-making.

- *Goal:* To discover useful information, inform conclusions and support decision-making.

Analytics - Process

Cross Industry Process For Data Mining (CRISP-DM)



Business Understanding

Understanding the project objectives and requirements from a business perspective.

Data Understanding

Covers steps from initial data collection, understanding, and exploration for quick insights. Hypotheses can be formed here.

Data Understanding – Data Structures

- Structured Data
- Semi-Structured Data
- Unstructured Data

Data Understanding – Data Types

- Numerical
- Strings
- Boolean
- Date-Time
- Special Data: Images, Sounds etc

Data Preparation

Involves all activities to construct the final dataset from the initial data. Tasks include cleaning and transformation of the data.

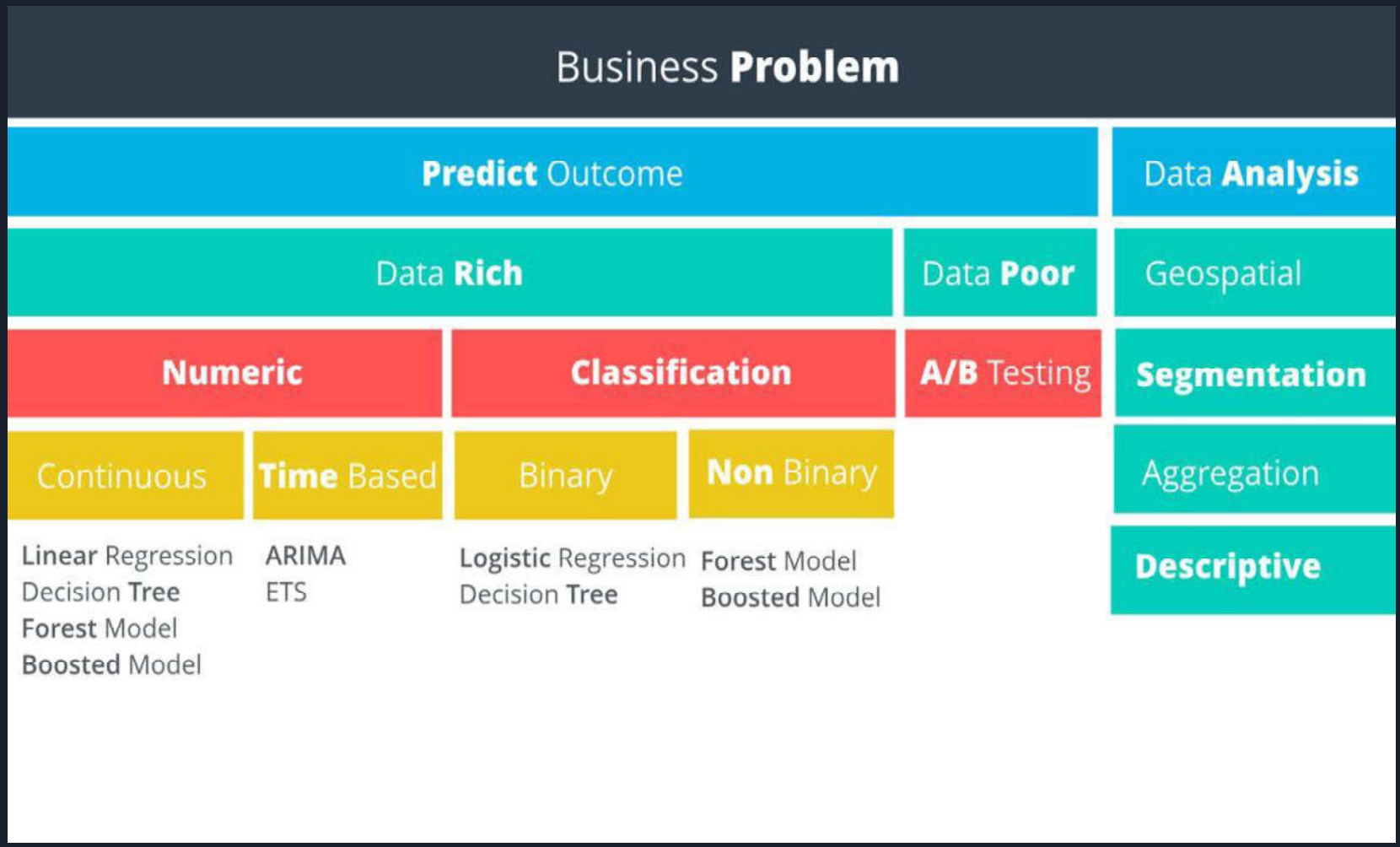
Data Preparation

- Data Cleaning
- Combining Data
- Dealing with Outliers

Analysis/Modeling

This is where the appropriate modeling and analysis techniques are applied to the prepared dataset. It is usually an iterative process with validation.

The Analyst's Methodology Map



Validation

Testing the models' results and ascertaining they meet fulfill the business objectives before final model deployment.

Presentation/Visualization

Presenting the results of the analysis, in relation to the original business problem, and making recommendations.

Analytics - Tools

R

Tableau Public

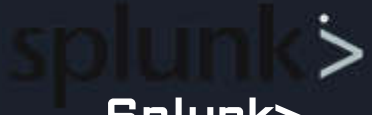
Python

SAS

Rapidminer

Knime

Apache Spark



Splunk>

Qlikview

Analytics - Types

Descriptive

What has happened?

Predictive

What could happen?

Prescriptive

What should we do?

Descriptive Analytics

What has happened?

Describes past events using statistical concepts like measures of central tendency, measures of variability, modality etc.

Example: a dashboard or a report that shows the number or percentage of sales people that have left the organization over the past year.

Predictive Analytics

What could happen?

Predicts future events using data from the past. Helps in planning for the future.

Examples: Using past data to predict the number of guests that a hotel would receive next Christmas.

Prescriptive Analytics

What should we do?

Predicts multiple future scenarios and proffers advice and recommendations on what the next steps should be.

Example: Recommending a system for handling the surge in number of guests.

Data Visualization & Storytelling

Data Visualization: The use of statistical graphics, plots, information graphics and other visual media to represent data.

Data Storytelling: A structured approach for communicating data insights, and it involves a combination of three key elements: data, visuals, and narratives.

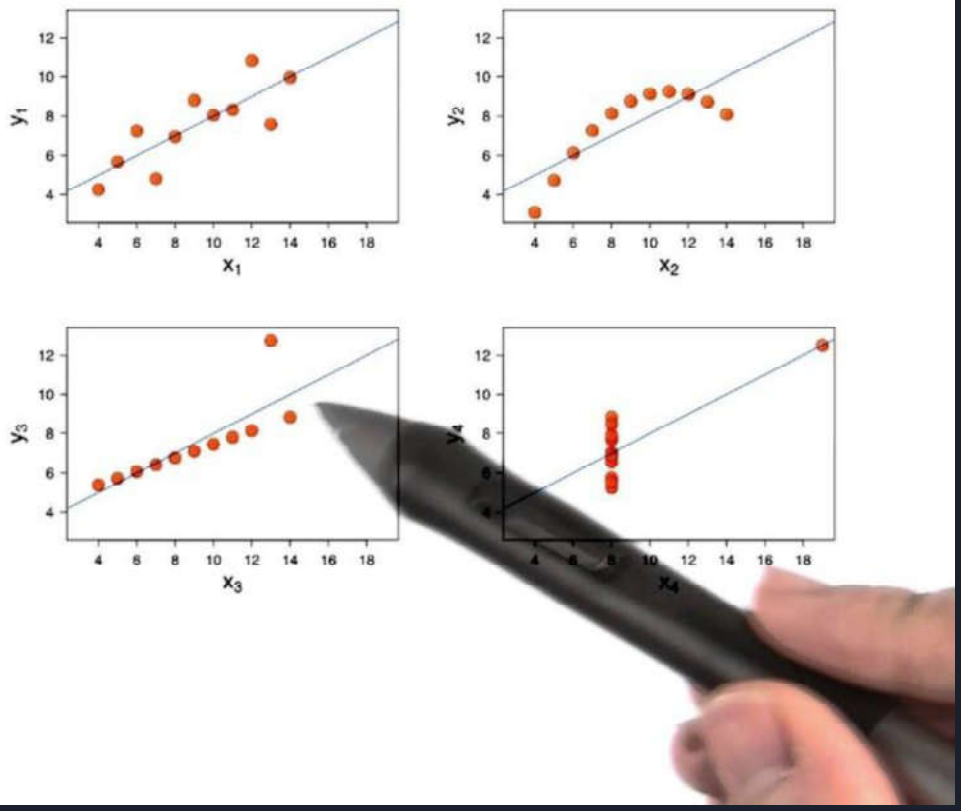
Exploratory Analysis

Summarizing the main properties of datasets with visual methods. Trying to get a sense of the data you're working with — and notice a few trends.

Sometimes, numbers alone don't tell the full story. Visuals show insights that are not always obvious from just descriptive statistical values.

Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Explanatory Analysis

Presenting your findings to the audience.
Audience could be your boss, customer, the public etc.
The visualization and story should be crafted to suit the specific audience being presented to.

Analytics - Applications

Industry	Applications
Banking/Finance	Credit-risk analysis, stock-price forecasting
Marketing	Consumer targeting, optimization of marketing campaigns, brand sentiment analysis
Human Resources Hiring	Hiring, Predicting employees' churn
Entertainment	Fan targeting, recommendations
Policing & Security	Predicting crime rates
Hospitality and Tourism	Forecasting hotel guests

Let’s talk