


# Complete Data Science and Machine Learning Using Python

By  
Jitesh Khurkhuriya



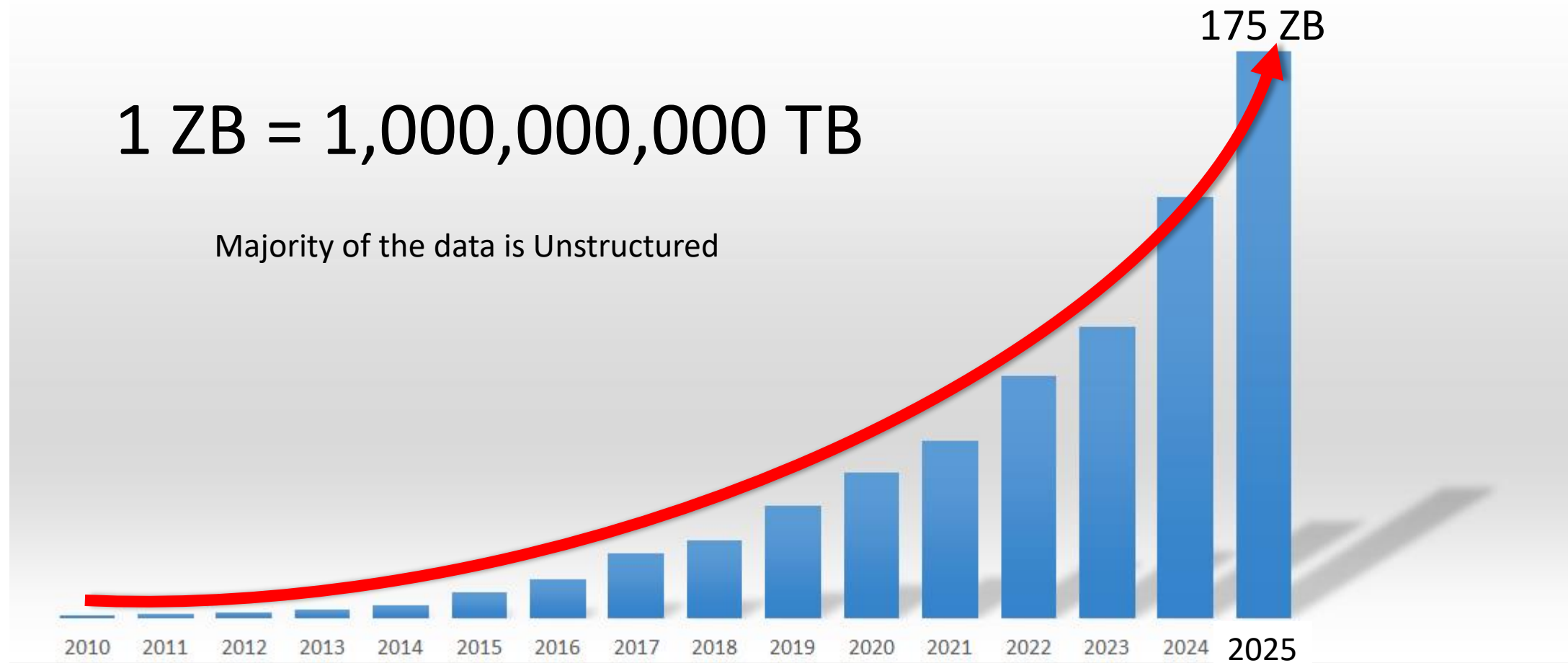
## The Digitization of the World From Edge to Core

David Reinsel - John Gantz - John Rydning  
November 2018

An IDC White Paper - #US44413318, Sponsored by  SEAGATE



# Data Growth – IDC-Seagate November, 2018



It's not just about.....

# NETFLIX

Customers who viewed this item also viewed these products



Dualit Food XL1500  
Processor

\$560

 Add to cart



Kenwood kMix Manual  
Espresso Machine

★★★★☆

\$250

 Select options



Weber One Touch Gold  
Premium Charcoal  
Grill-57cm

\$225

 Add to cart

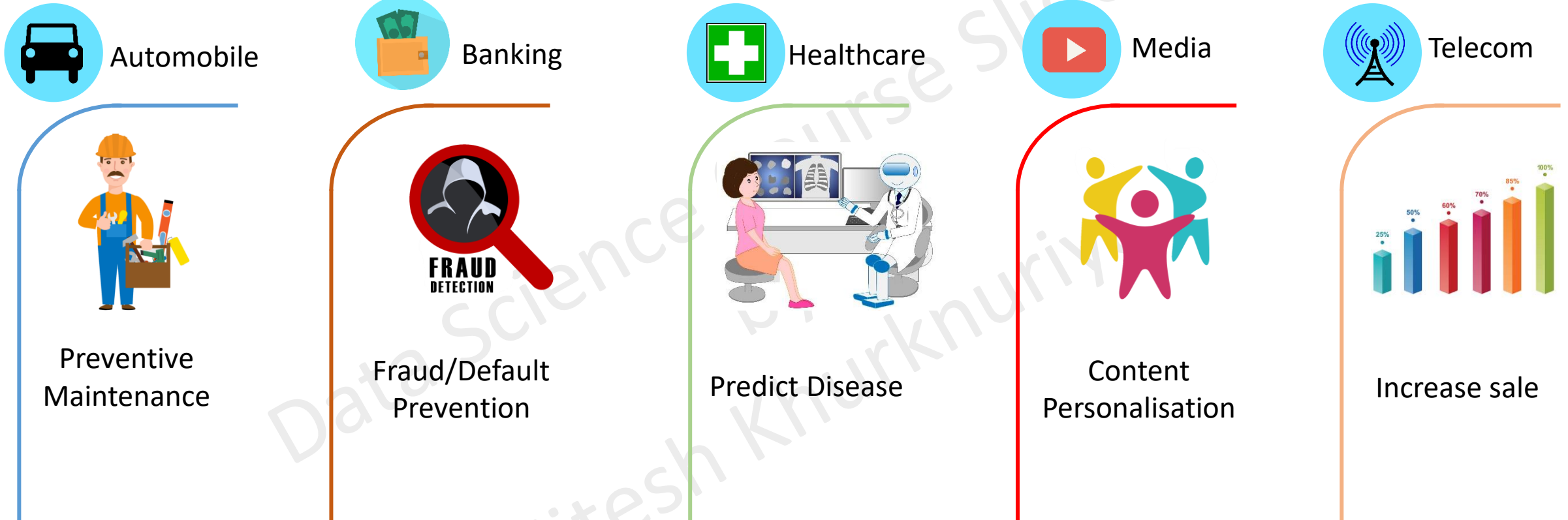


NoMU Salt Pepper and  
Spice Grinders

\$3

 View options

# Application of Data Science and Machine Learning



# Heard On The Streets

- **IDC Futurescape** - Two-thirds of Global 2000 Enterprises CEOs will centre their corporate strategy on digital transformation including machine learning (ML) solutions.
- **Harvard Business Review** – Data Scientist: The Sexiest Job of the 21st Century
- **McKinsey Report** – 45 percent of work activities could potentially be automated by currently demonstrated technologies; machine learning can be an enabling technology for the automation of 80 percent of those activities.
- **Microsoft CEO Satya Nadella** – called out machine learning -- and the big data that powers it -- as a key development in his memo to Microsoft last July.

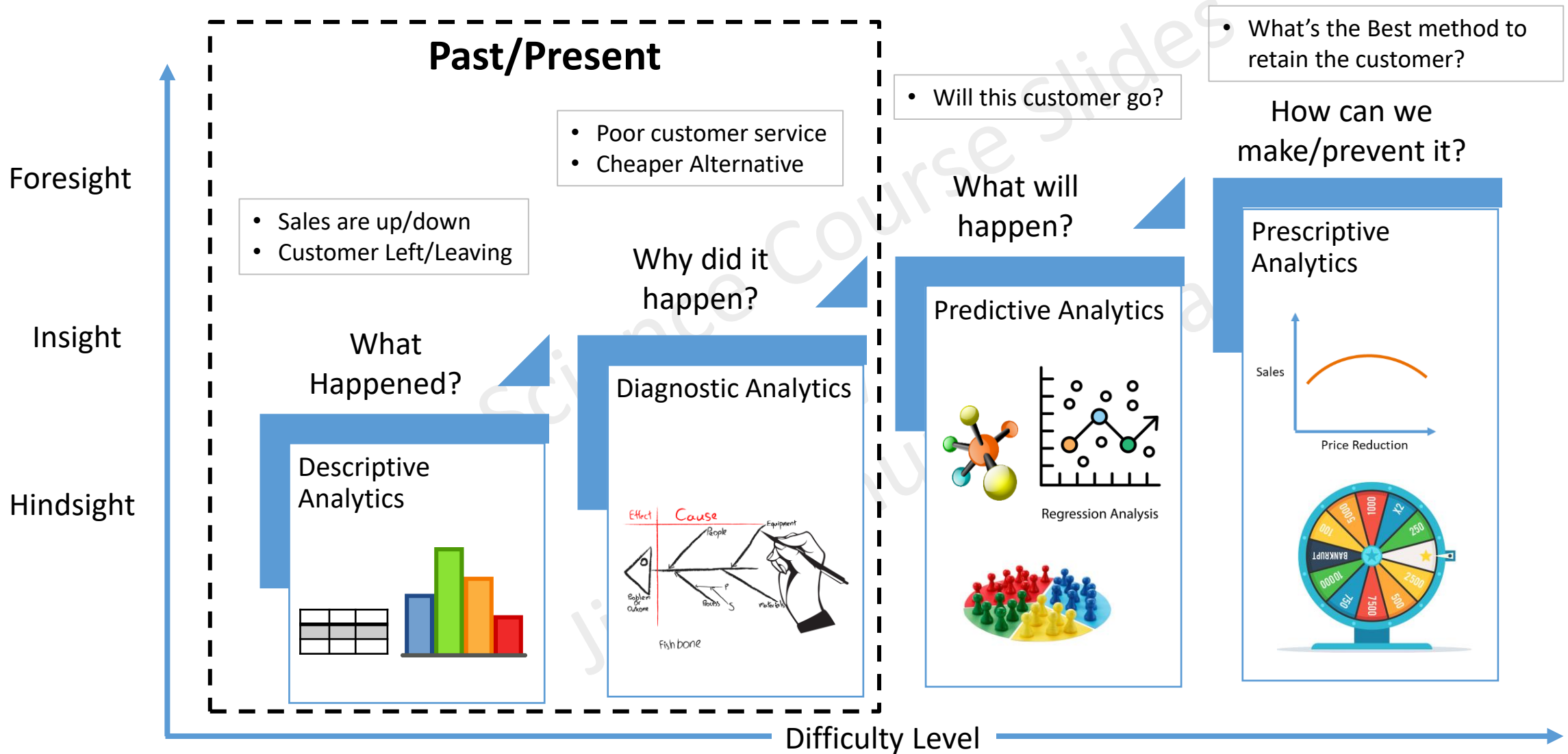


# Benefits of Data Science and Machine Learning

---

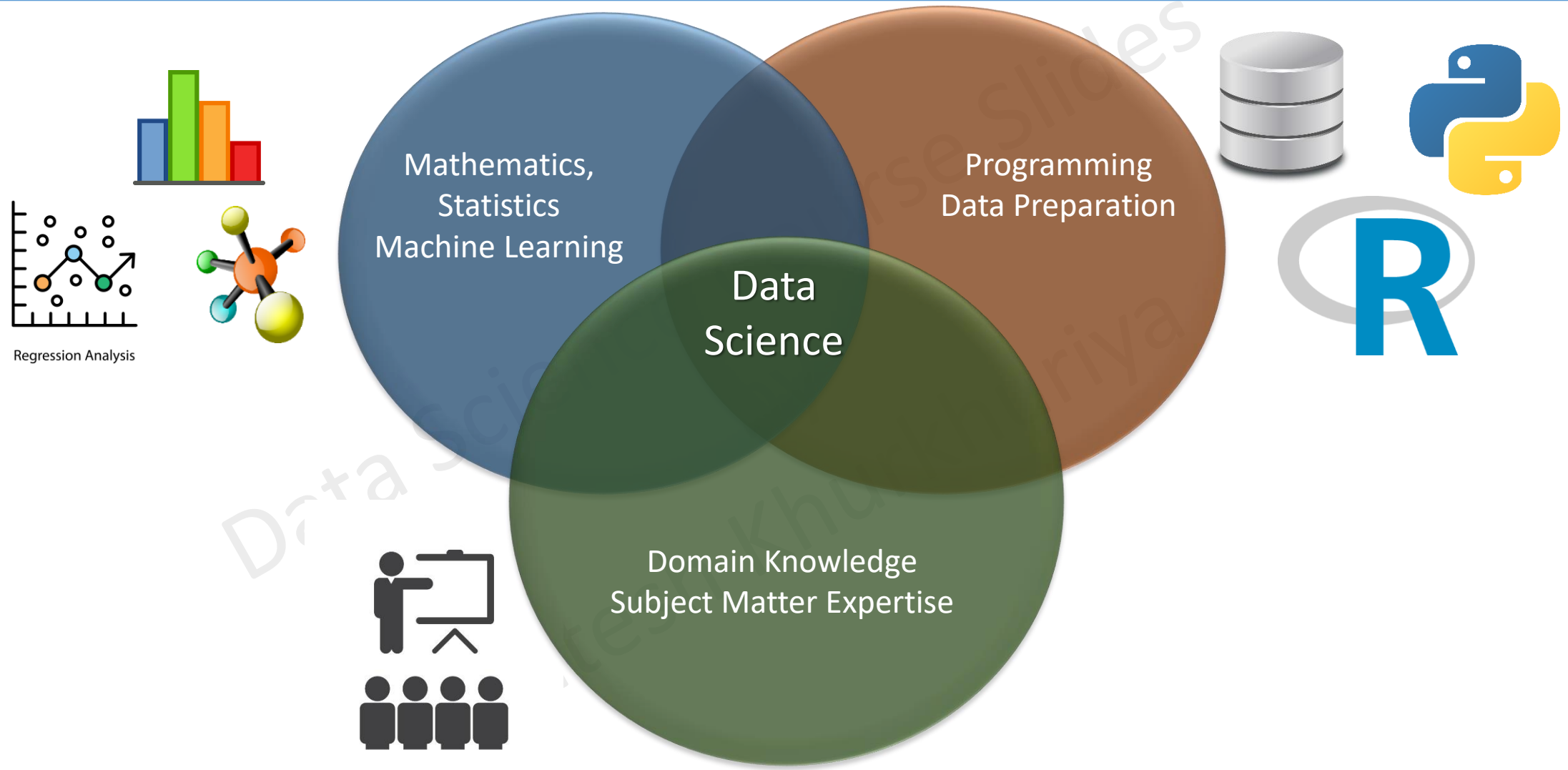
- ✓ Faster decisions
- ✓ Develop insights that are beyond human capabilities
- ✓ Act at the right time and take advantage of opportunities, converting them into closed deals.

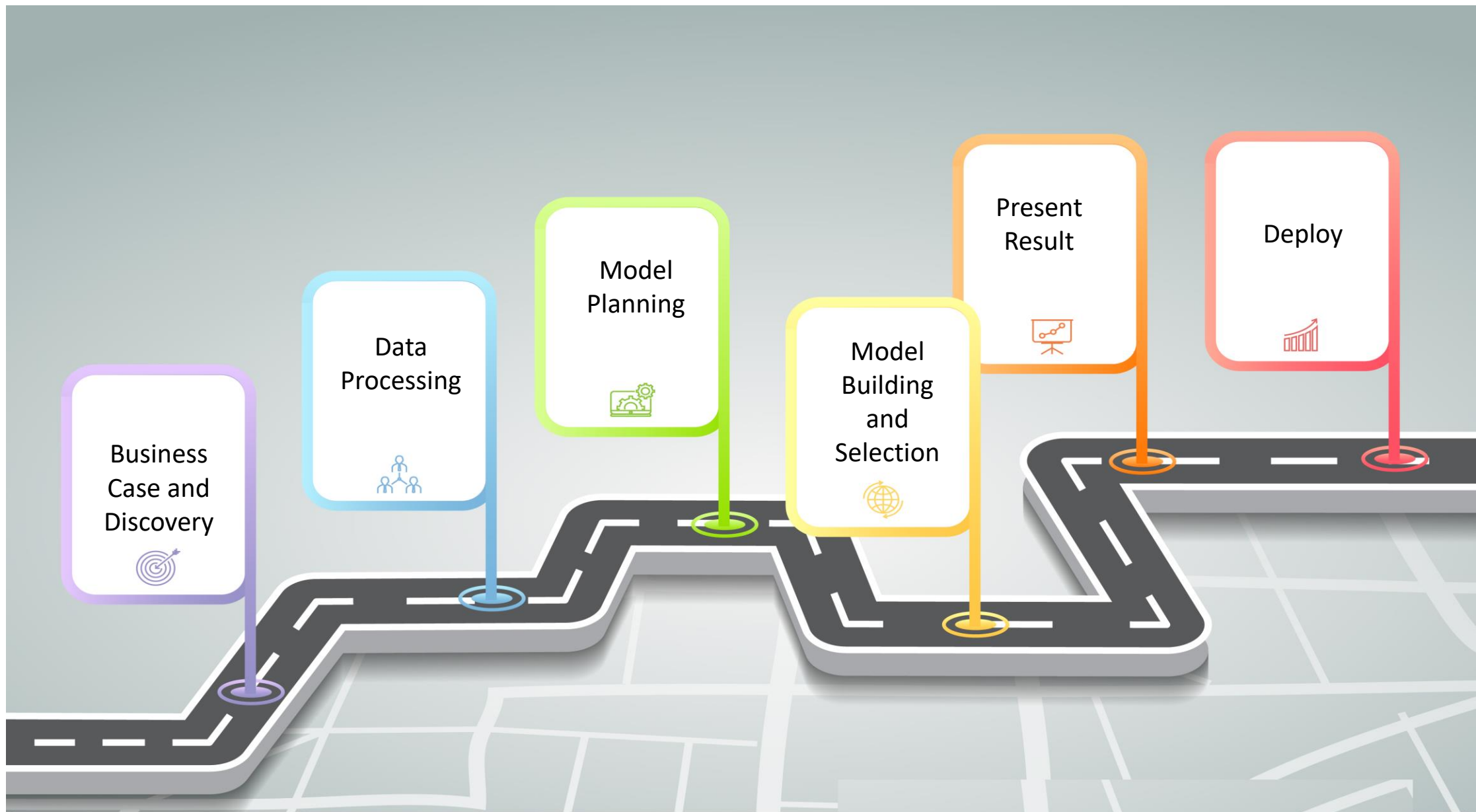
# Types of Analytics





# What is Data Science?





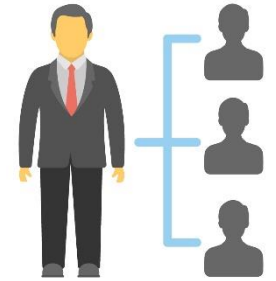
# Business Case and Discovery



Stakeholders Discussions



What's the End Goal?



How much time and budget we have



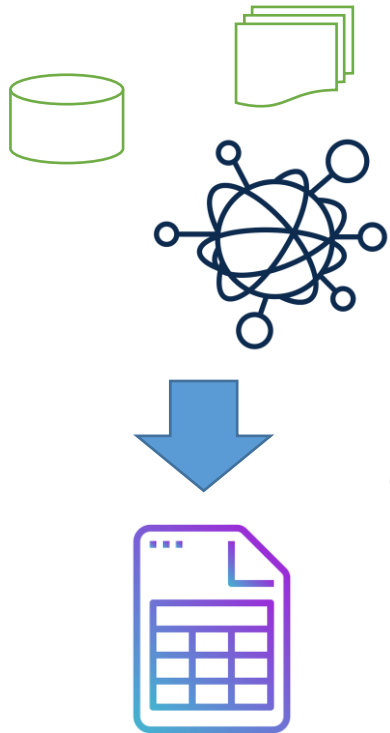
Past attempts



What kind of data is available

# Data Processing

## Data Mapping

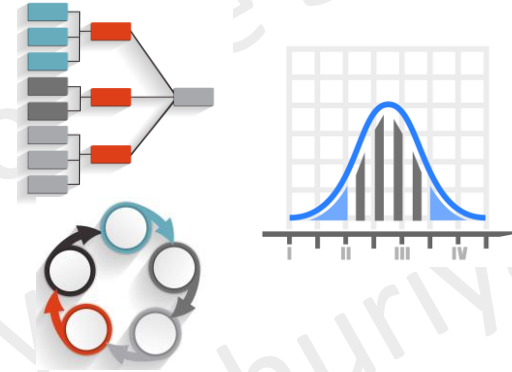


## Data Cleaning



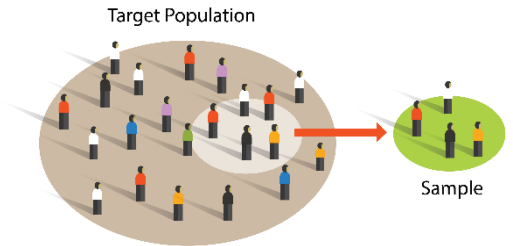
- Data Quality
- Missing Data
- Noisy Data
- Outlier Treatment

## Data Transformation



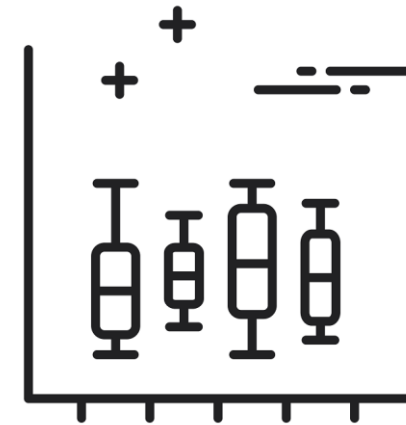
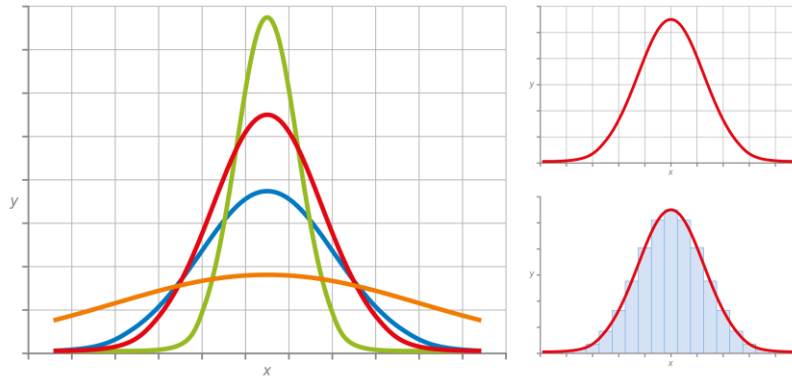
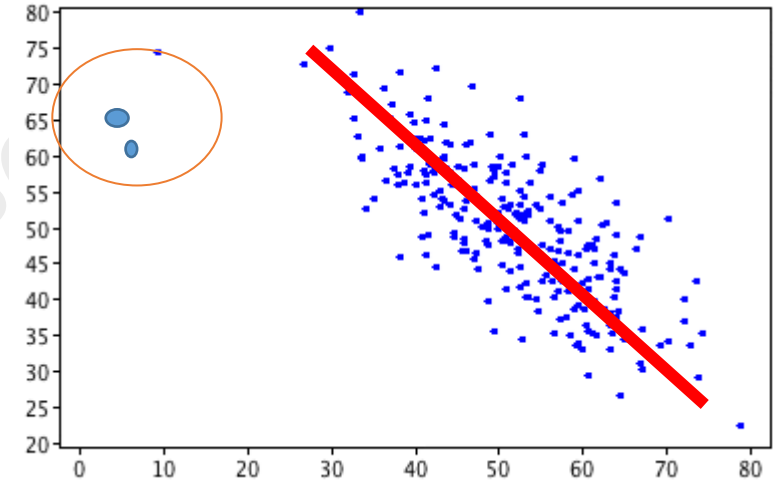
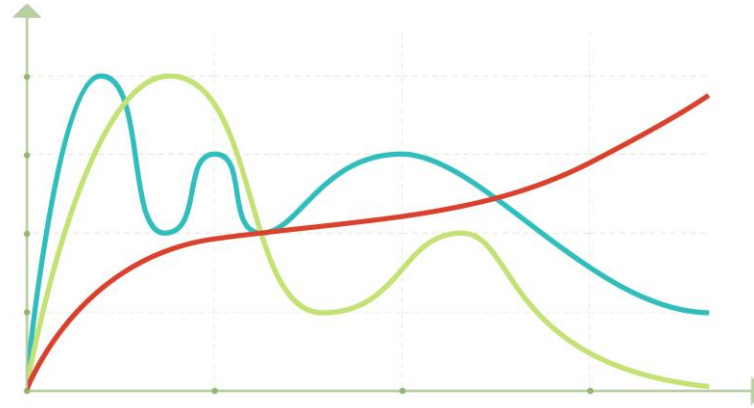
- Format conversion
- Data Normalization
- Statistical imputation
- Feature Engineering

## Sample the Data



- Data Sampling
- Data Split
- Data Binning

# Exploratory Data Analysis



## ANOMALY DETECTION

One Class SVM

> 100 Features

PCA Based Anomaly Detection

Fast Training

## CLUSTERING

K-MEANS

## MULTI-CLASS CLASSIFICATION

Fast Training, Linear Model

Multi-Class Logistic Regression

Accuracy, Long Training Times

Multi-Class Neural Network

Accuracy, Fast Training

Multi-Class Decision Forest

Accuracy, Small Memory Footprint

Multi-Class Decision Jungle

Depends on Two-Class

One-V-All Multiclass

## REGRESSION

Ordinal Regression

Data in Rank Order  
categories

Poisson Regression

Predicting Event Counts

Fast Forest Quantile Regression

Predicting a  
Distribution

Linear Regression

Fast Training, Linear  
Model

Bayesian Linear Regression

Linear Model, Small  
datasets

Neural Network Regression

Accuracy, Long Training  
Time

Decision Forest Regression

Accuracy, Fast Training

Boosted Decision Tree Regression

Accuracy, Fast Training,  
large Memory

Start

## TWO-CLASS CLASSIFICATION

Two Class SVM

>100 Features,  
Linear Model

Two-Class Averaged  
Perceptron

Fast Training,  
Linear Model

Two Class Logistic  
Regression

Fast Training,  
Linear model

Two Class Bayes  
Point Machine

Fast Training,  
Linear Model

Accuracy, Fast  
Training

Two-Class Decision  
Forest

Accuracy, Fast  
Training, LargeM

Two-Class Boosted  
Decision Tree

Accuracy, SmallM

Two Class Decision  
Jungle

>100 Features

Two Class Locally Deep  
SVM

Accuracy, Long  
Training Times

Two Class Neural  
Network

# What to consider while choosing an algorithm?

---

Predicting Categories

Predicting Continuous Value

Finding Unusual Data Points

Discovering Structure

© Jitesh Khurkhuriya



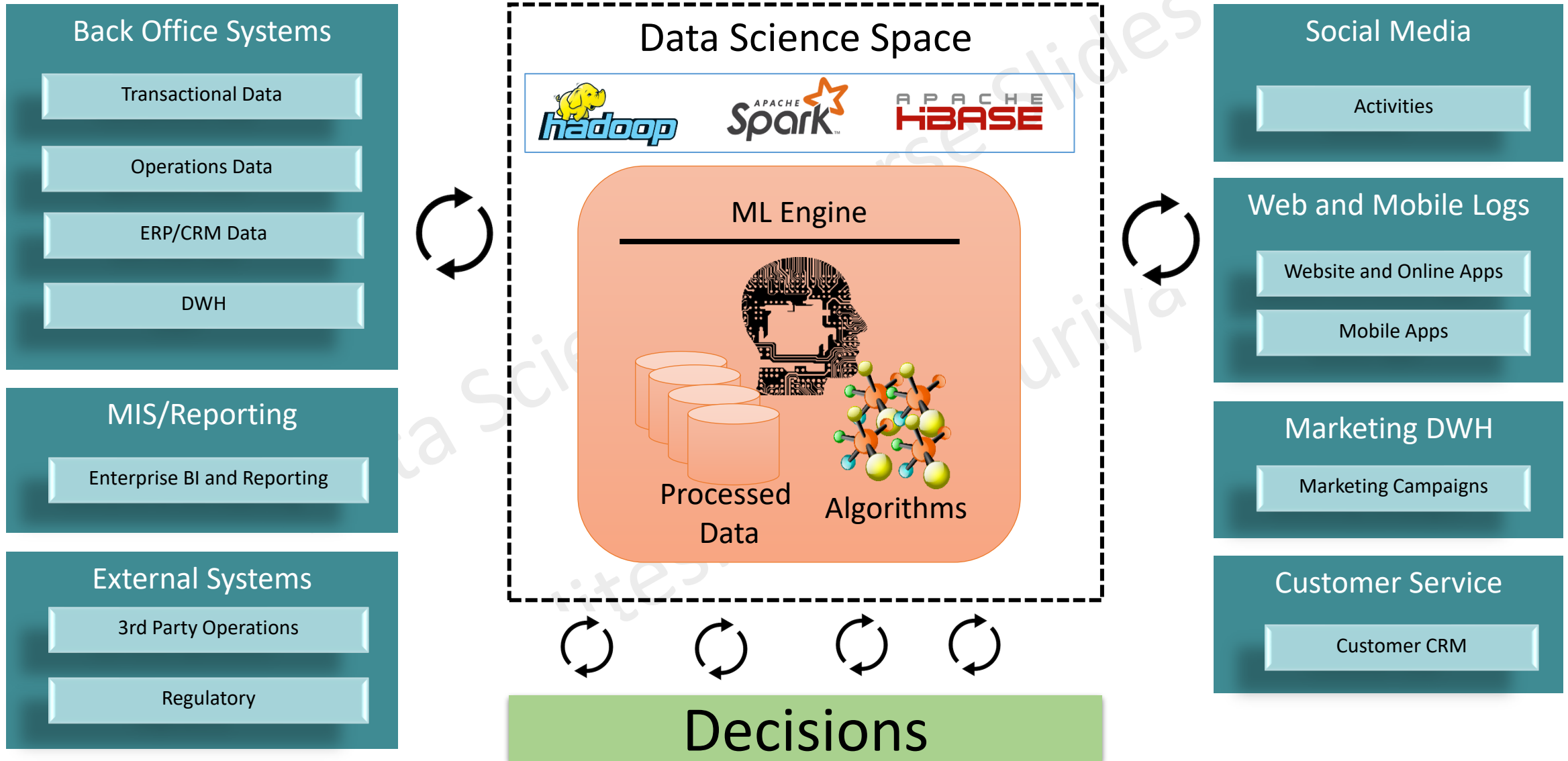


# Present the results

---

- Explain the process of model planning and selection
- Explain the findings; correlations, causes, variable selections
- Communicate the results
- Explain the process of operationalization

# Deployment

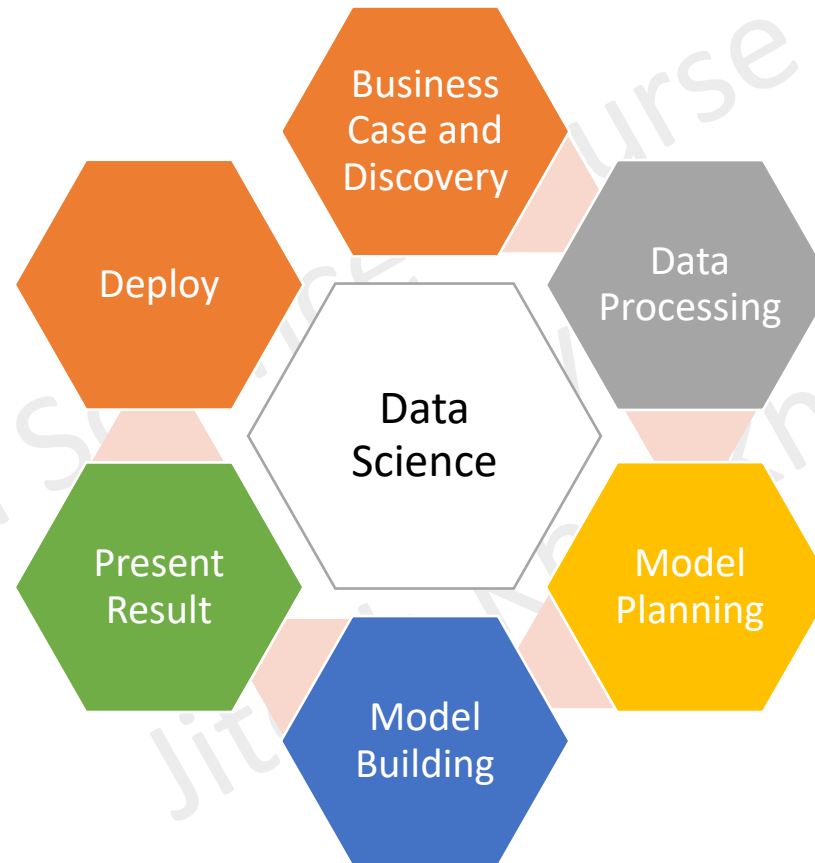


# Skills Required to be a Data Scientist

---

- Soft Skills

- Domain knowledge
- Communication
- Analytical skills
- Curiosity
- Common Sense



- Technical Skills

- Mathematics
- Statistics
- File handling or database
- Machine Learning
- Python or similar
- Tableau or similar visualization

# Soft Skills

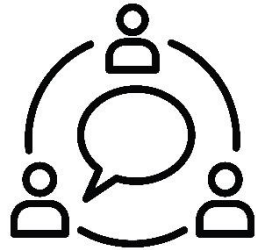
---

Understanding of the data elements based on domain expertise



Domain knowledge

Discovery phase as well as presenting findings to the stakeholders



Communication

Analyse various relationships among data features.



Analytical Skills

Asking the right questions to gain deeper understanding.



Curiosity

Is it making sense on normal beliefs?



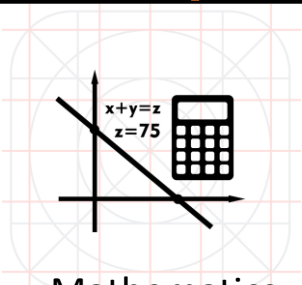
Common Sense

# Technical Skills

Math as the basis for algorithms. Helps for own implementations.

Helps in dealing with the imperfections of data as well as data transformation

Build models using either Python, R, SAS, Azure ML



Mathematics



Statistics



Data Wrangling



Machine Learning



Programming Languages



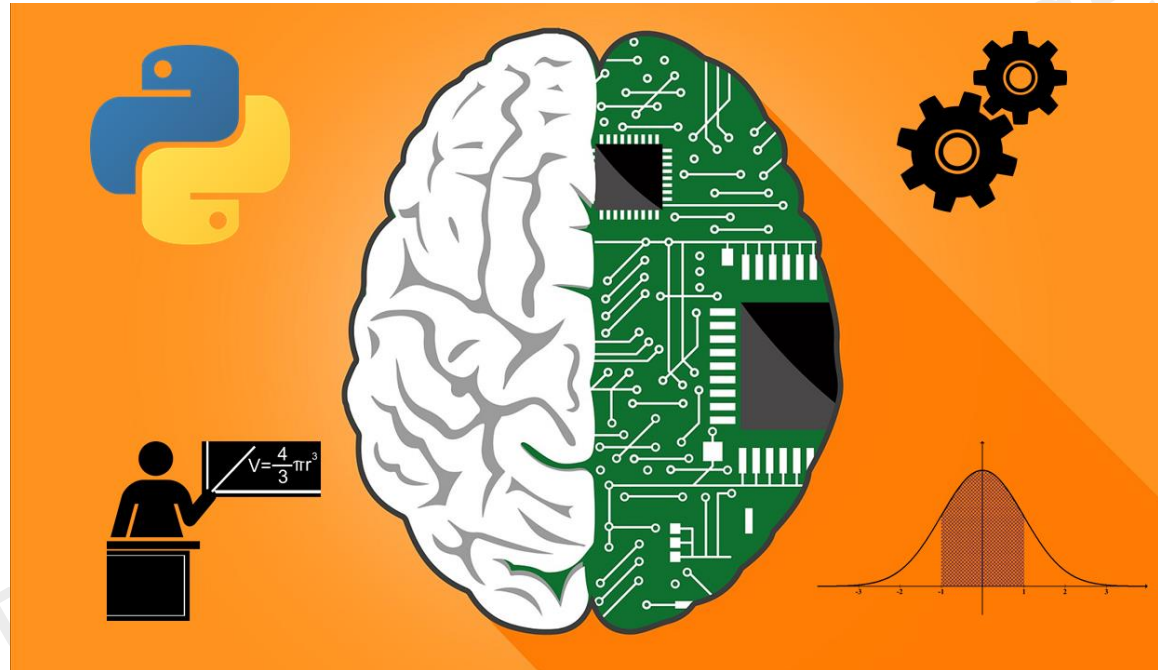
Data Visualisation

Helps in data imputation as well as validate the results of an experiment

Heart of Data Science. Various algorithms for predictions of the outcome.

Visual understanding of data as well as communication of findings.

# Complete Data Science and Machine Learning Using Python



# Thank You!