

**zenius**



**Kampus  
Merdeka**  
INDONESIA JAYA

# Introduction to Data Science

Selasa, 14 Maret 2023

Data Analytics

Program Zenius Studi Independen Bersertifikat Bersama Kampus  
Merdeka



# Quick Intro

## Theo Jeremiah

### Roles:

- CURRENTLY | Data Scientist at AirAsia
- 20 - 23 | Data Scientist at Allianz Indonesia
- 19 - 20 | Business Development at Mineski Indonesia
- 18 - 19 | Data Analyst at Excite Indonesia



[www.linkedin.com/in/theojeremiah/](https://www.linkedin.com/in/theojeremiah/)

What comes to your mind when you hear the word

# **“Data Science.”**

- 1. What is Data Science ?**
- 2. Why Data Science ?**
- 3. Data Product**
- 4. Data Science Methodologies**



# What is Data Science ?

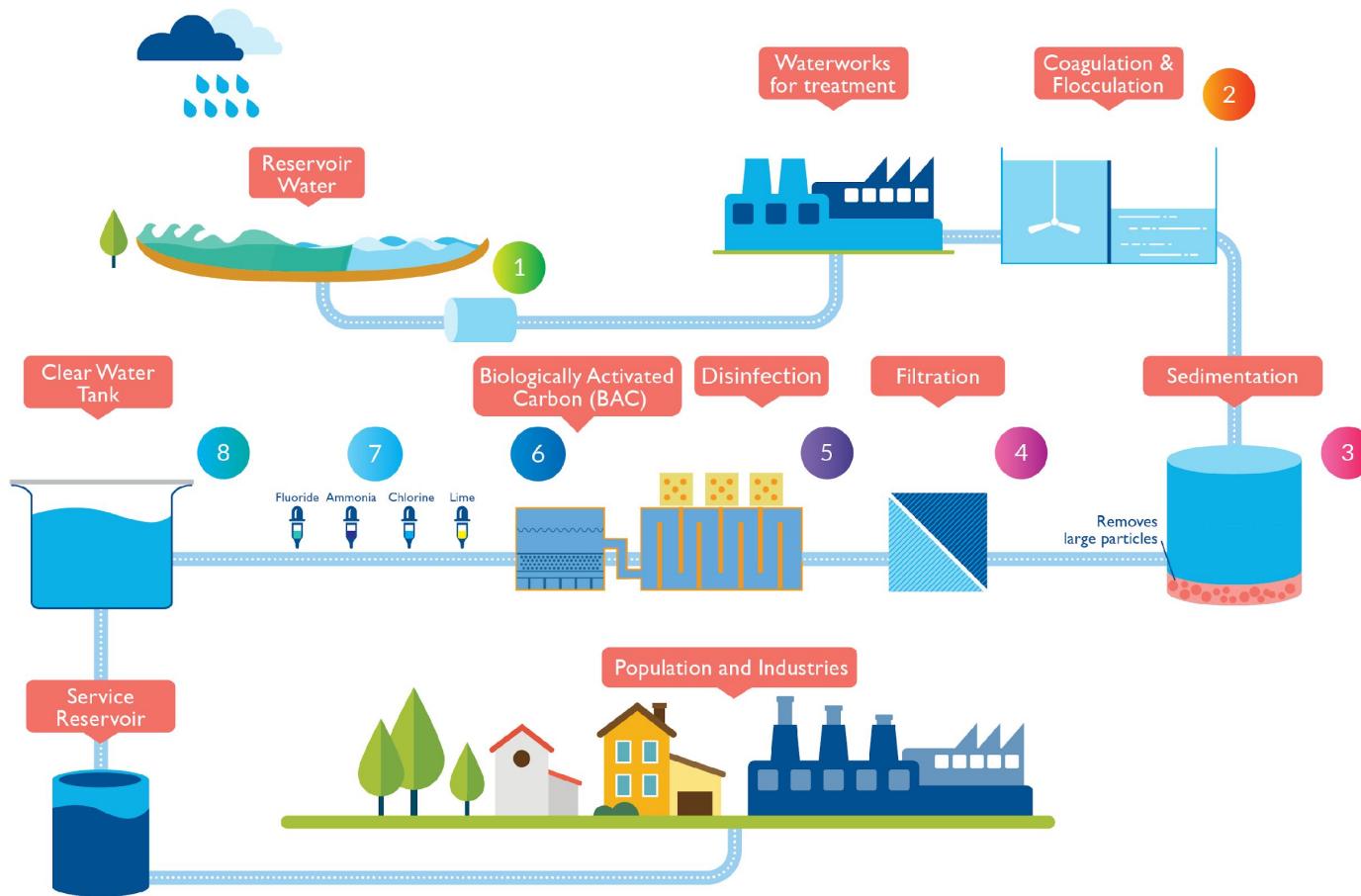
Data Science ?

Data Scientist ?

Machine Learning ?

**“Data is the new Oil.”**





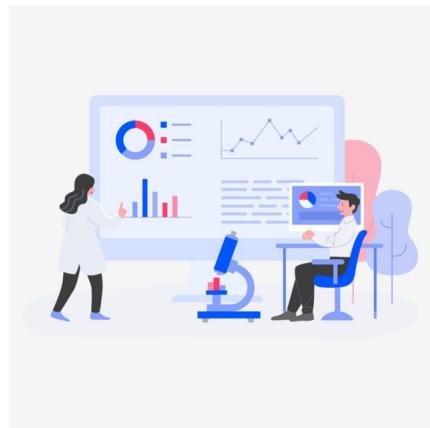
Data science **combines multiple “Science” fields**, including **statistics, scientific methods, artificial intelligence (AI)**, and **data analysis**, with an aim **to extract value from “data”**.

Those **who practice data science are called data scientists**, and they combine a range of skills to analyze data collected from the web, smartphones, customers, sensors, and other sources to derive actionable insights.

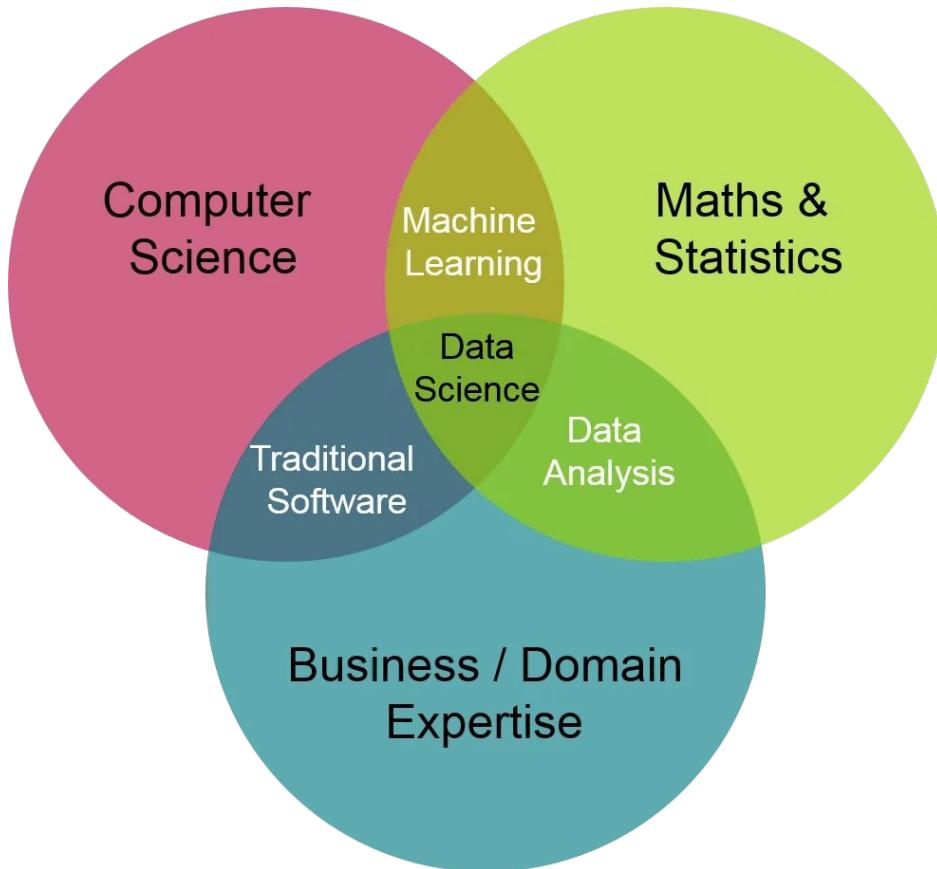
So, it could be concluded that :

**Data Science : The Field**

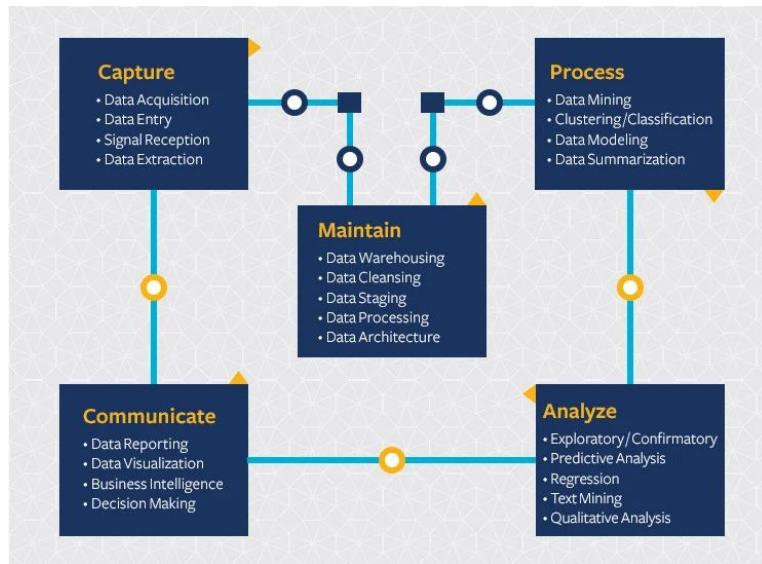
**Data Scientist : The Person**



Source : <https://www.oracle.com/data-science/what-is-data-science/>



Data science encompasses **preparing data for analysis**, including **cleansing, aggregating, and manipulating** the data to perform **advanced data analysis**. **Analytic applications** and **data scientists** can then **review the results** to **uncover patterns** and **enable business leaders to draw informed insights**.



Source : <https://ischoolonline.berkeley.edu/data-science/what-is-data-science/>

The central figure is a cartoon illustration of a man with dark hair, wearing a blue suit, red tie, and white shirt, sitting in a meditative lotus pose on a white cloud. He has his eyes closed and is holding a magnifying glass over a bar chart icon. To his left is a donut chart icon, and to his right is a bar chart icon. The background is a dark blue gradient with light blue wavy patterns in the corners.

## DATA SCIENTIST MUST-HAVE SKILLS

**MATH & STATISTICS**

- Machine Learning
- Statistical Modeling
- Exploratory Analysis
- Clustering
- Regression Analysis

**DOMAIN KNOWLEDGE & SOFT SKILLS**

- Inclination towards business operations
- Keen on working with data
- Problem solver
- Strategic, proactive, and cooperative
- Interested in hacking

**PROGRAMMING & DATABASE**

- Computer Science Fundamentals
- Database Management System
- Data Visualization
- Python
- Big Data

**COMMUNICATION & VISUALIZATION**

- Storytelling skills
- Convert data-based insights into decisions
- Collaborative with Sr. Management
- Knowledge of tools like Tableau
- Visual art design

SPOTLIGHT ON BIG DATA

## Spotlight

ARTWORK Tamar Cohen, Andrew J Buboltz  
2011, silk screen on a page from a high school  
yearbook, 8.5" x 12"

# Data Scientist: *The Sexiest Job of the 21st Century*

**Meet the people who  
can coax treasure out of  
messy, unstructured data.**

by Thomas H. Davenport  
and D.J. Patil

70 Harvard Business Review October 2012



Source : <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

# Why Data Science ?





Observation



- Customer Loyalty
- Customer Satisfaction

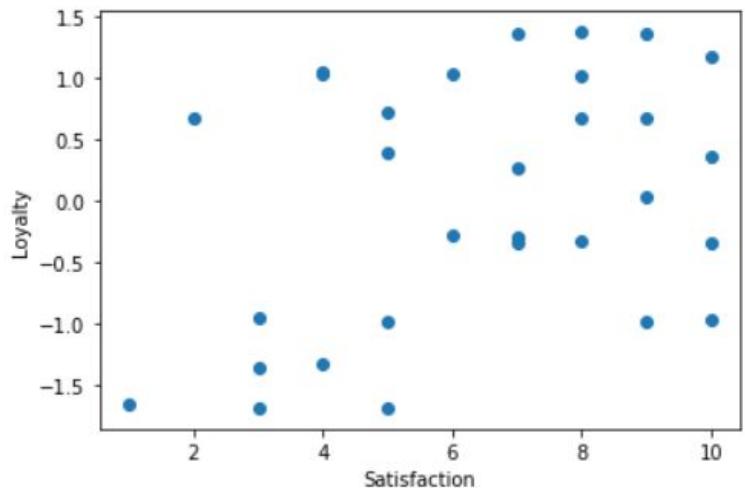
## 30 Customer Observation

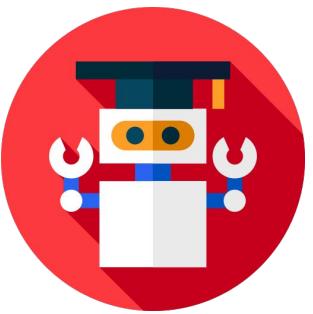
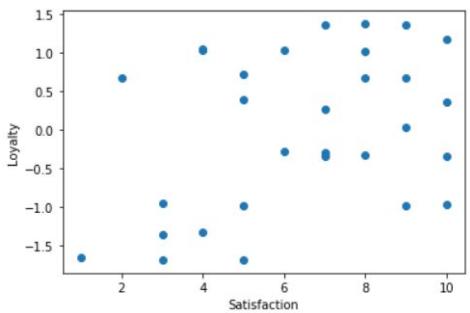


- Customer Loyalty
- Customer Satisfaction

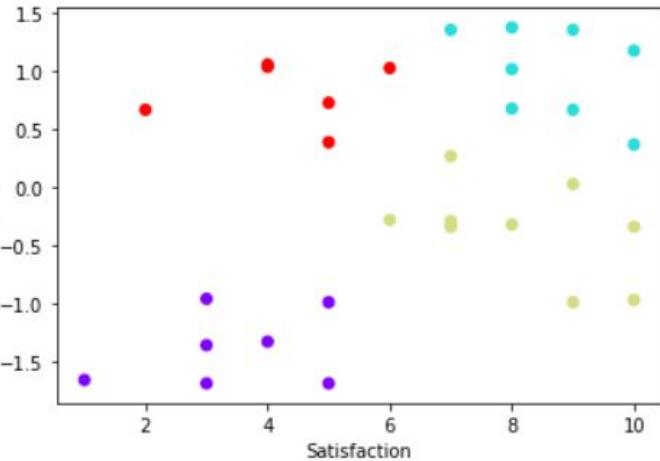
DATA

## Scatter Plot





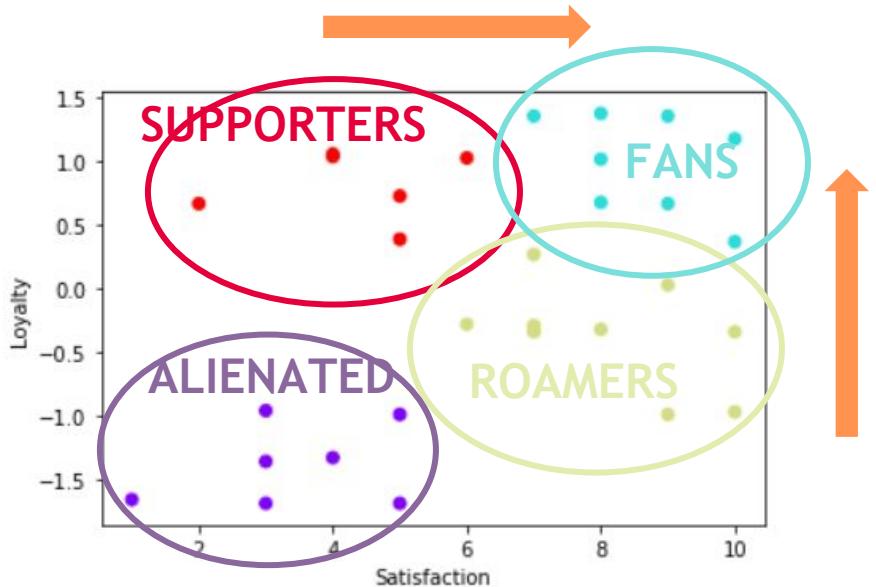
### Machine Learning



# SCIENCE

## Increase Customer Satisfaction:

- High Prices
- Long Queues
- Unfriendly Staff

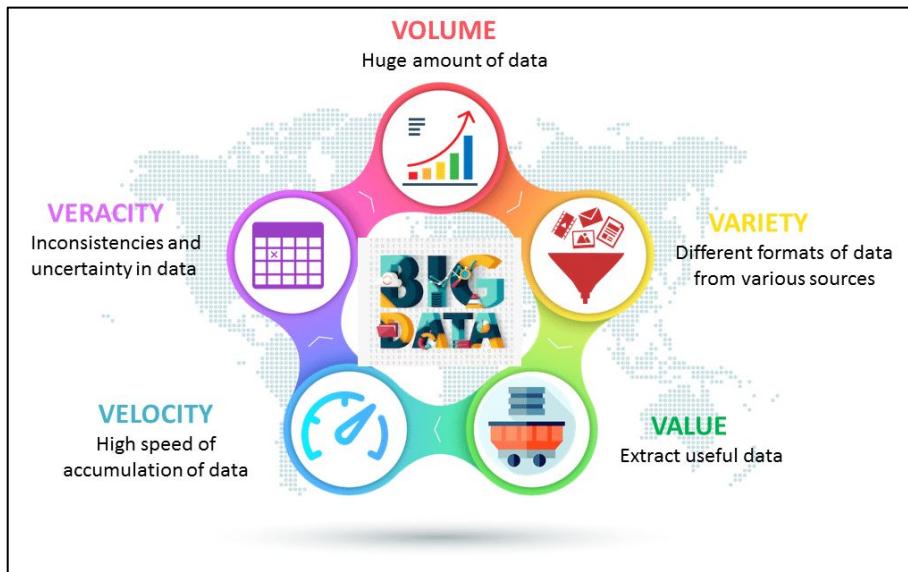


## INSIGHTS

## Increase Customer Loyalty:

- Discounts
- Loyalty Cards
- Raffles

# Why Data Science ?



# Why Data Science ?





357,762

TERKONFIRMASI  
+4,301 Kasus

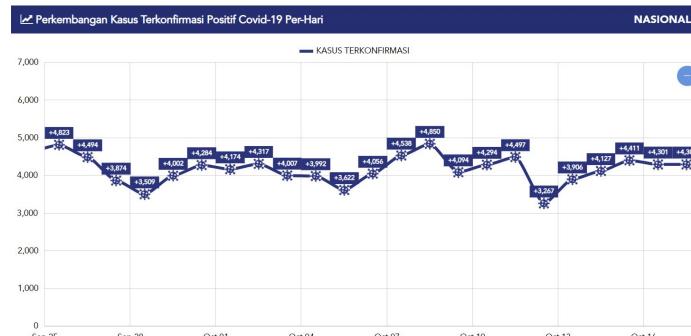
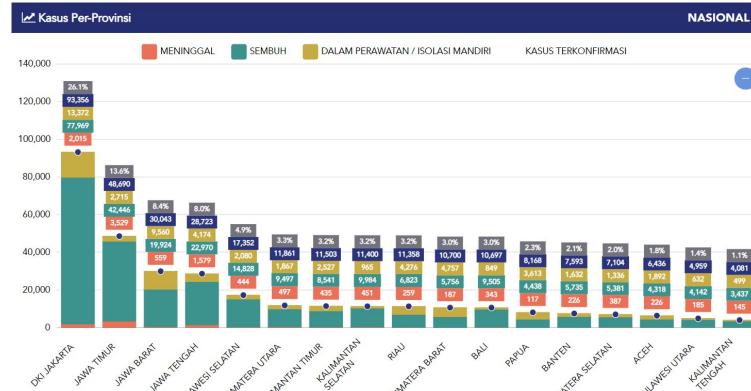
63,739

KASUS AKTIF  
17.8% dari Terkonfirmasi

281,592

SEMBUH  
78.7% dari Terkonfirmasi

12,431

MENINGGAL  
3.5% dari Terkonfirmasi

# Why Data Science ?

Data Science enables enterprises to **measure, track, and record performance metrics** for **facilitating enterprise-wide enhanced decision making**.

Companies can analyze trends to make critical decisions to engage customers better, enhance company performance, and increase profitability.

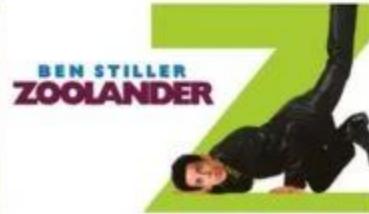
Data Science helps organizations identify and refine target audiences by combining existing data with other data points for developing useful insights. Data Science also helps recruiters by combining data points to identify candidates that best fit their company needs.

# Data Science Use Cases

NETFLIX

# Data Science Use Cases

NETFLIX



# Data Science Use Cases



# Data Product



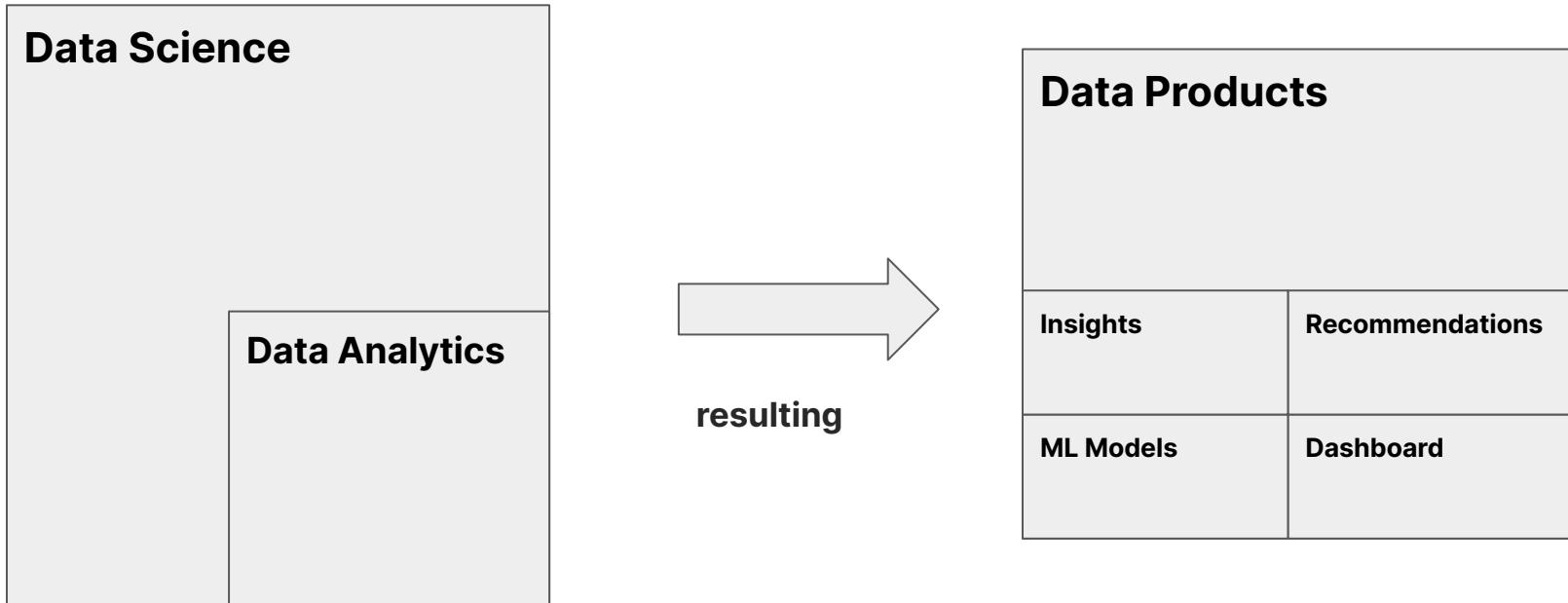
# Data Product

DJ Patil, former United States Chief Data Scientist, defined a **data product** as “**a product that facilitates an end goal through the use of data**”

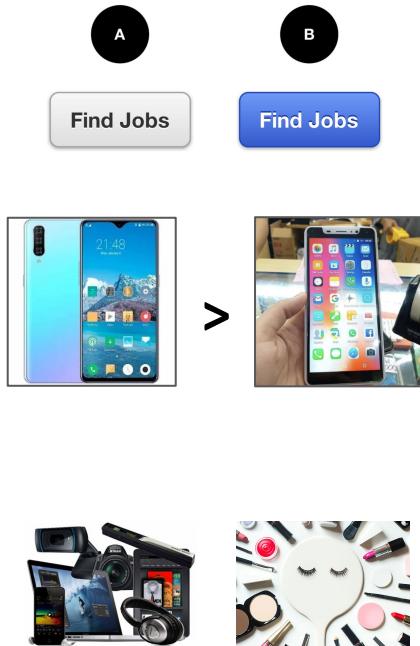
(from his book Data Jujitsu: The Art of Turning Data into Product, 2012).



We can say that :



# Insights



"In this page **Button B 33% is clicked more often** compared to **Button A**"

"In our website, **a product that has no background 40% sold more than ones with background**"

"In our website, **automotive and electronics category** is visited more by **men**. On the other hand, **baby products and beauty** is visited more by **women**"

A/B Testing

Analysis

# Machine Learning

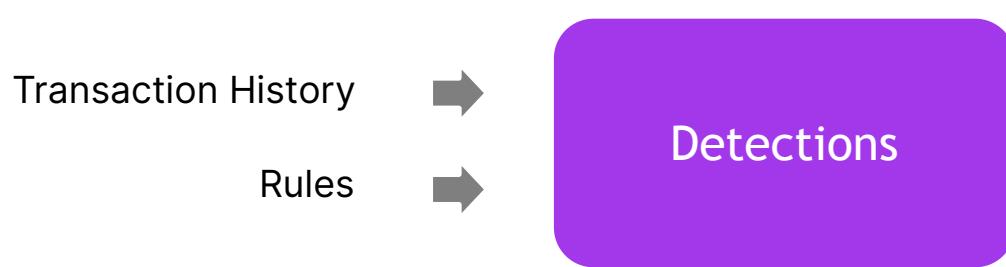
## Business need

You need to **identify** fraudulent credit card transactions before they have a larger impact on your company



# Machine Learning

## Classical programming approach

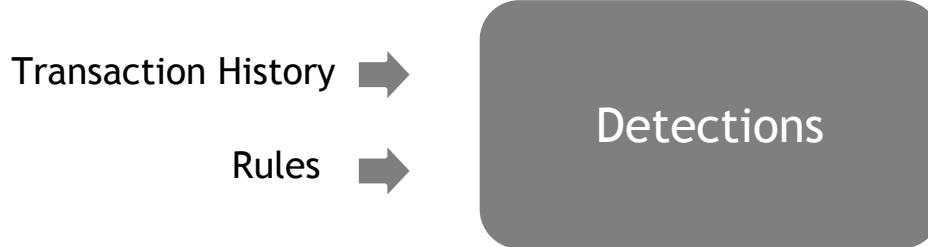


### Problems:

- Requires programmers to explicitly set rules
- Fraudulent cases are extremely unpredictable
- Each rule adds more time to the process

# Machine Learning

Classical programming approach



Machine learning approach

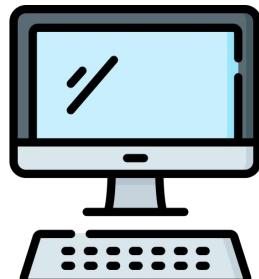


# Machine Learning

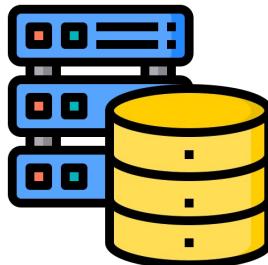
## Classical programming approach



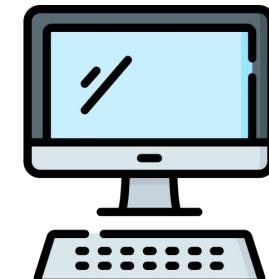
Very specific  
instructions



## Machine learning approach



Data



# Machine Learning

What about Machine learning:

- Derives rules from the data itself
- Adapts to complexities in Fraudulent transactions
- Trains model ahead of time, saving time towards making predictions

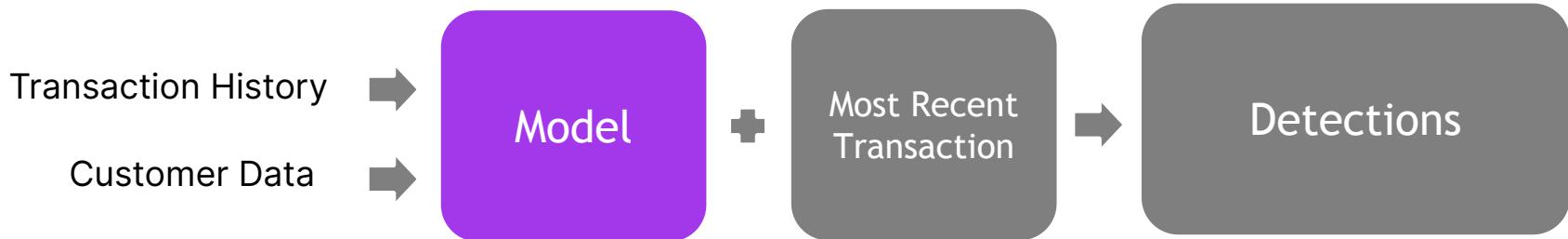
## Machine learning approach



# Machine Learning

A **trained algorithm** which:

- Is used to identify patterns in your data, and
- Does not require explicit, manually set rules.



# What is Machine Learning



$$f(\textcolor{violet}{x}) = \textcolor{teal}{a}_0 x_0 + \textcolor{teal}{a}_1 x_1 + \cdots + \textcolor{teal}{a}_n x_n$$

Features and **weights**

**Feature:** An important category of your data

**$x_0$ :** Is this product a hat?

Yes = 1

# What is Machine Learning



If  $f(\textcolor{violet}{x}) > 1$ , recommend the product.

---

$$f(\textcolor{violet}{x}) = \textcolor{cyan}{0.8} * \textcolor{violet}{1} + \textcolor{cyan}{0.25} * \textcolor{violet}{1}$$

$$f(\textcolor{violet}{x}) = 1.05$$

*What kinds of algorithms are used?*

Quoting a Data Scientist from a unicorn startup on 2019 :

"We split 2 type of data products into **machine analytics** and **human analytics**."

### Machine Analytics

Type of data products that relies dominantly on the usage of unstructured data, and the form of them frequently delivered as an application and machine is the pure decision maker

#### Example :

- Image Recognition
- Product Recommendation
- Fraud Detection

### Human Analytics

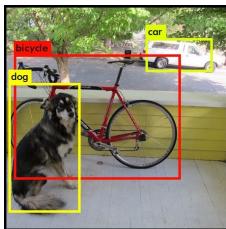
Type of data products that combines both the balance of unstructured and structured data to find insights and human will act as the decision maker

#### Example :

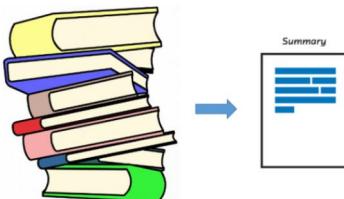
- Demand Forecasting
- Customer Lifetime Behavior Analysis

# Data Product Example

Image  
Recognition



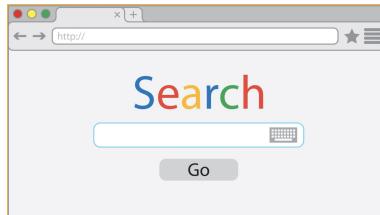
Text Mining



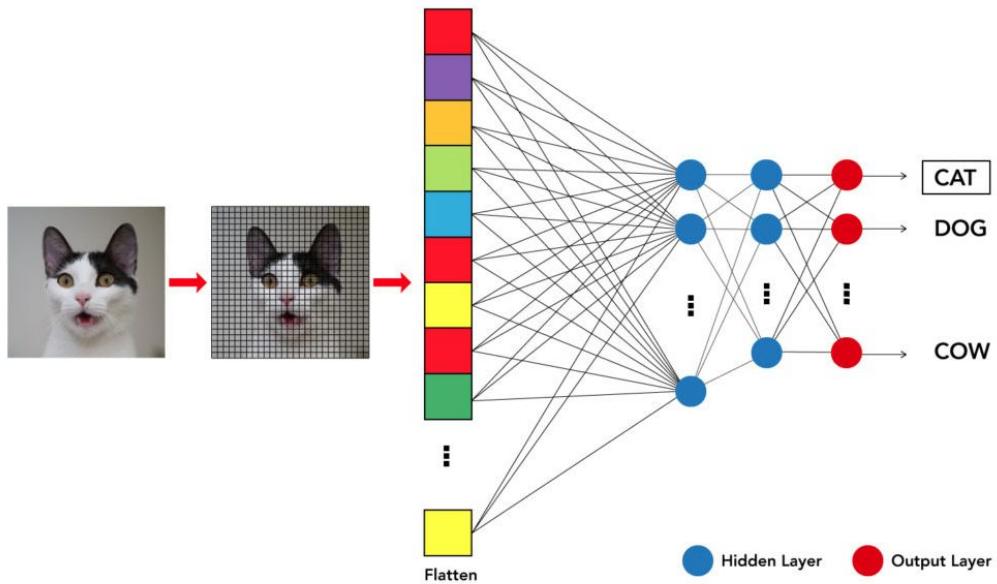
Route  
Optimizer



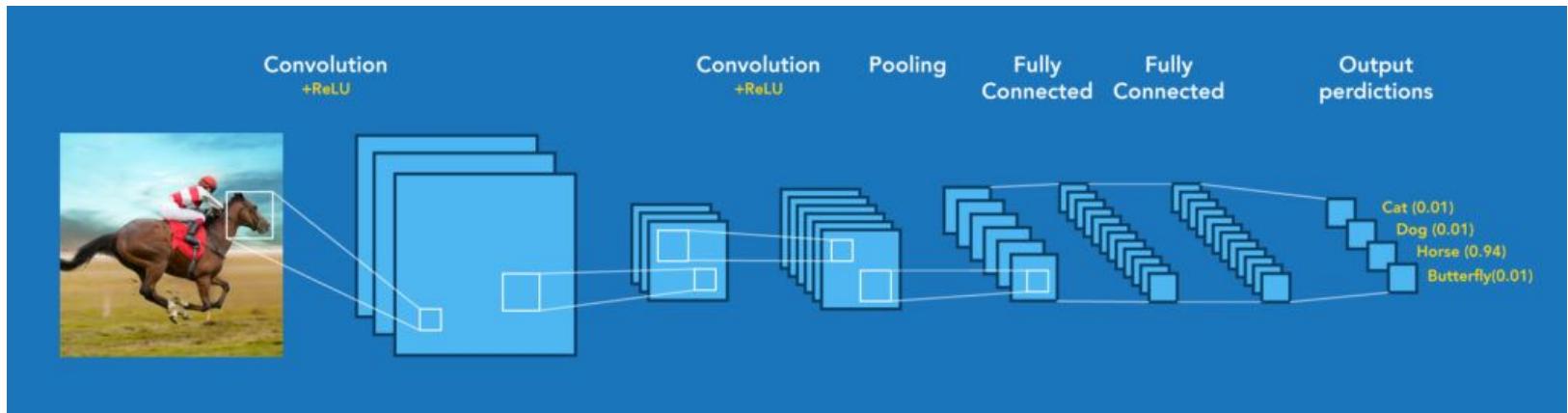
Search Engine



# Image Recognition

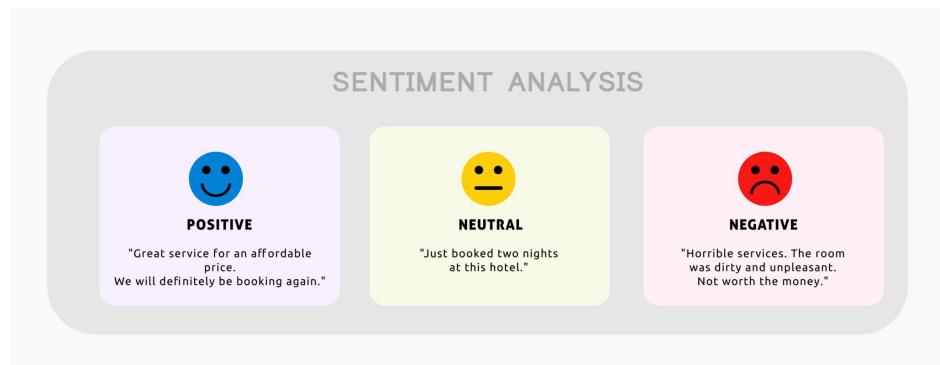
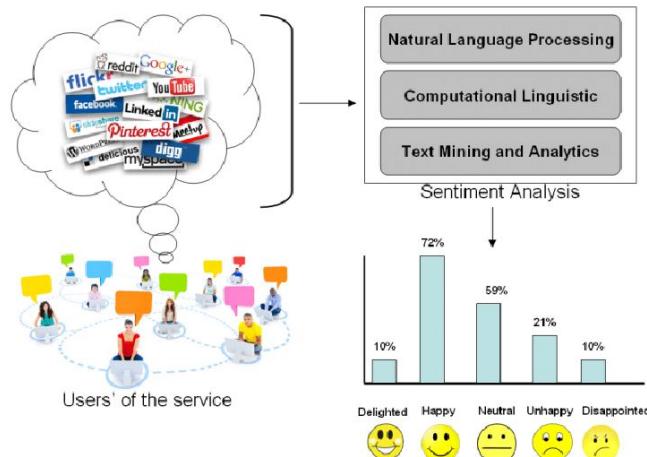


# Image Recognition



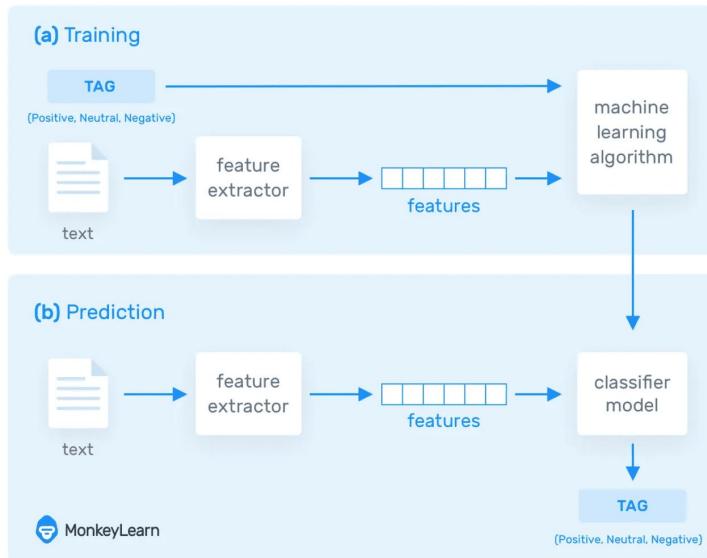
# Text Mining

## Sentiment Analysis



# Text Mining

## Sentiment Analysis



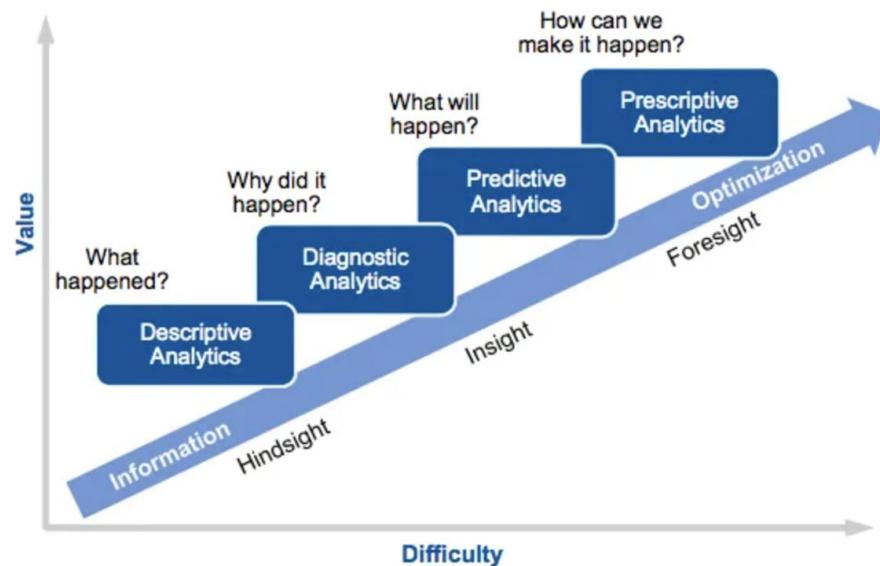
Source : <https://monkeylearn.com/blog/sentiment-analysis-machine-learning/>

**Pop Quiz! Can you mention one of  
data science use cases around us?**

# Data Science Methodologies



# Type Of Analytics

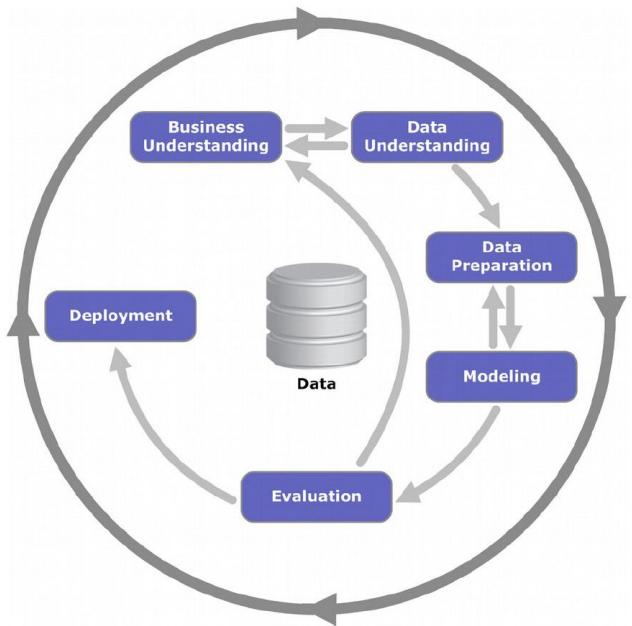


Source : Gartner Analytics Ascendancy Model

<https://www.clickz.com/how-can-ai-allow-marketers-to-predict-the-future/112268/gartner-analytic-ascendancy-model/>

<https://www.gartner.com/en/topics/data-and-analytics>

# CRISP-DM



Source : CRISP-DM on IBM Watson

<https://developer.ibm.com/articles/introduction-watson-studio/>

# CRISP-DM (Detail)

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p><b>Determine Business Objectives</b>  <i>Background Business Objectives</i>  <i>Business Success Criteria</i></p> <p><b>Assess Situation</b>  <i>Inventory of Resources Requirements, Assumptions, and Constraints</i>  <i>Risks and Contingencies</i>  <i>Terminology</i>  <i>Costs and Benefits</i></p> <p><b>Determine Data Mining Goals</b>  <i>Data Mining Goals</i>  <i>Data Mining Success Criteria</i></p> <p><b>Produce Project Plan</b>  <i>Project Plan</i>  <i>Initial Assessment of Tools and Techniques</i></p>	<p><b>Collect Initial Data</b>  <i>Initial Data Collection Report</i></p> <p><b>Describe Data</b>  <i>Data Description Report</i></p> <p><b>Explore Data</b>  <i>Data Exploration Report</i></p> <p><b>Verify Data Quality</b>  <i>Data Quality Report</i></p>	<p><b>Select Data</b>  <i>Rationale for Inclusion/Exclusion</i></p> <p><b>Clean Data</b>  <i>Data Cleaning Report</i></p> <p><b>Construct Data</b>  <i>Derived Attributes</i>  <i>Generated Records</i></p> <p><b>Integrate Data</b>  <i>Merged Data</i></p> <p><b>Format Data</b>  <i>Reformatted Data</i></p> <p><b>Dataset</b>  <i>Dataset Description</i></p>	<p><b>Select Modeling Techniques</b>  <i>Modeling Technique</i>  <i>Modeling Assumptions</i></p> <p><b>Generate Test Design</b>  <i>Test Design</i></p> <p><b>Build Model</b>  <i>Parameter Settings</i>  <i>Models</i>  <i>Model Descriptions</i></p> <p><b>Assess Model</b>  <i>Model Assessment</i>  <i>Revised Parameter Settings</i></p>	<p><b>Evaluate Results</b>  <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i>  <i>Approved Models</i></p> <p><b>Review Process</b>  <i>Review of Process</i></p> <p><b>Determine Next Steps</b>  <i>List of Possible Actions</i>  <i>Decision</i></p>	<p><b>Plan Deployment</b>  <i>Deployment Plan</i></p> <p><b>Plan Monitoring and Maintenance</b>  <i>Monitoring and Maintenance Plan</i></p> <p><b>Produce Final Report</b>  <i>Final Report</i>  <i>Final Presentation</i></p> <p><b>Review Project</b>  <i>Experience Documentation</i></p>

# Data Analytics Workflow



Requirements Scope	What to measure How to measure	Source Table Relationship Diagram Querying Data	Cleansing - Missing Values - Anomalies - Outliers	Techniques Tools	Presentation Visualization Insights
--------------------	-----------------------------------	---	--	------------------	--

**Tools :**  
SQL, Excel

**Scope :**  
Data Exploration  
Interviews to User

**Tools :**  
SQL, Excel, Python, R

**Scope :**  
Cleansing, Analysis, Finding Insights

**Tools :**  
Dashboarding (Tableau,  
PowerBI, Google Data Studio),  
Excel, Python

**Scope :**  
Delivering Data Product

Thanks!  
Any Questions?

