18th Sep 2022

# Analytics Pipeline Management


**Prof. (Dr. ) Sridhar Vaithianathan,**
**Director & Professor (Analytics),**
**Centre of Excellence in Analytics & Data Science,**
**SVKM's NMIMS (Deemed to be University), Mumbai.**



**LinkedIn:** in.linkedin.com/in/sridharvaithianathan

**Email:** prof.sridhar.we@gmail.com
**Mobile:** 99899 04245

# Disclaimer

(c) Do Not use or quote without the permission of the author

Personal use is permitted, but republication/redistribution requires Author’s permission

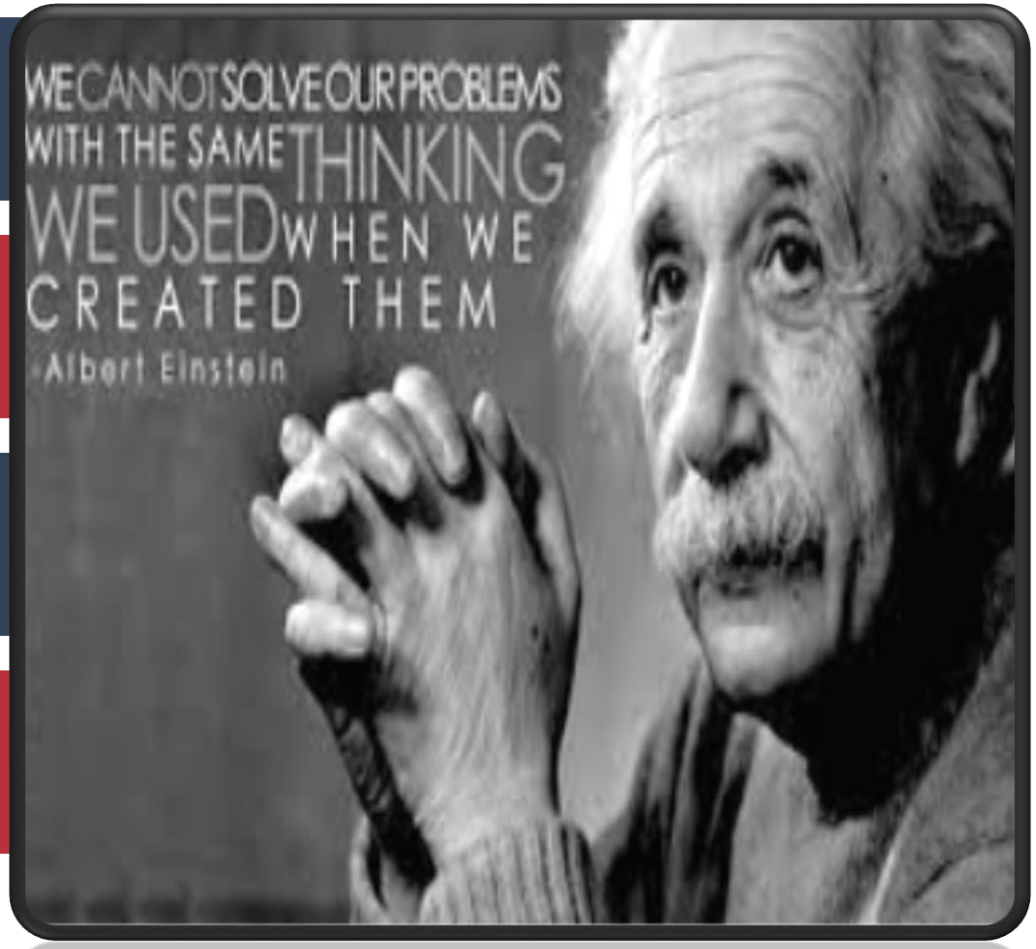
# SESSION OUTLINE

**Data Science – Redefining Living**

**Why Machine Learning ?**

**Machine Learning – Classification**

**ANALYTICS PIPELINE MANAGEMENT**



# PRESENTATION PROGRESS

**Data Science – Redefining Living**

**Why Machine Learning ?**

# HOW DATA SCIENCE IS REDEFINING LIVING ?

# THIS IS THE WORLD YOU KNOW…

# THIS IS THE WORLD YOU ARE ACTUALLY LIVING

## Since 2012 IN…..



DATA REVOLUTION…


TREASURE

# IT's all about DATA

## DATA EXPLOSION

- Each day, our society creates 2.5 quintillion bytes of data (that's 2.5 followed by 18 zeros). With this glut of data, the need to make sense of it becomes more acute.

## DATA QUOTES

- Data is the new Oil – Dunn Humby , UK-based customer science company.
- Data is the new raw material of business – Microsoft
- The world is now awash in data and we can see consumers in a lot clearer ways – Paypal
- The goal is to turn data into information, and information into insight - HP

**Self driving cars**: Google, Baidu, Tesla have implemented this technology.

**Speech recognition**: Google now, Siri, Cortana


**Genetics**: Clustering algorithms are used in genetics to help find genes associated with a particular disease.

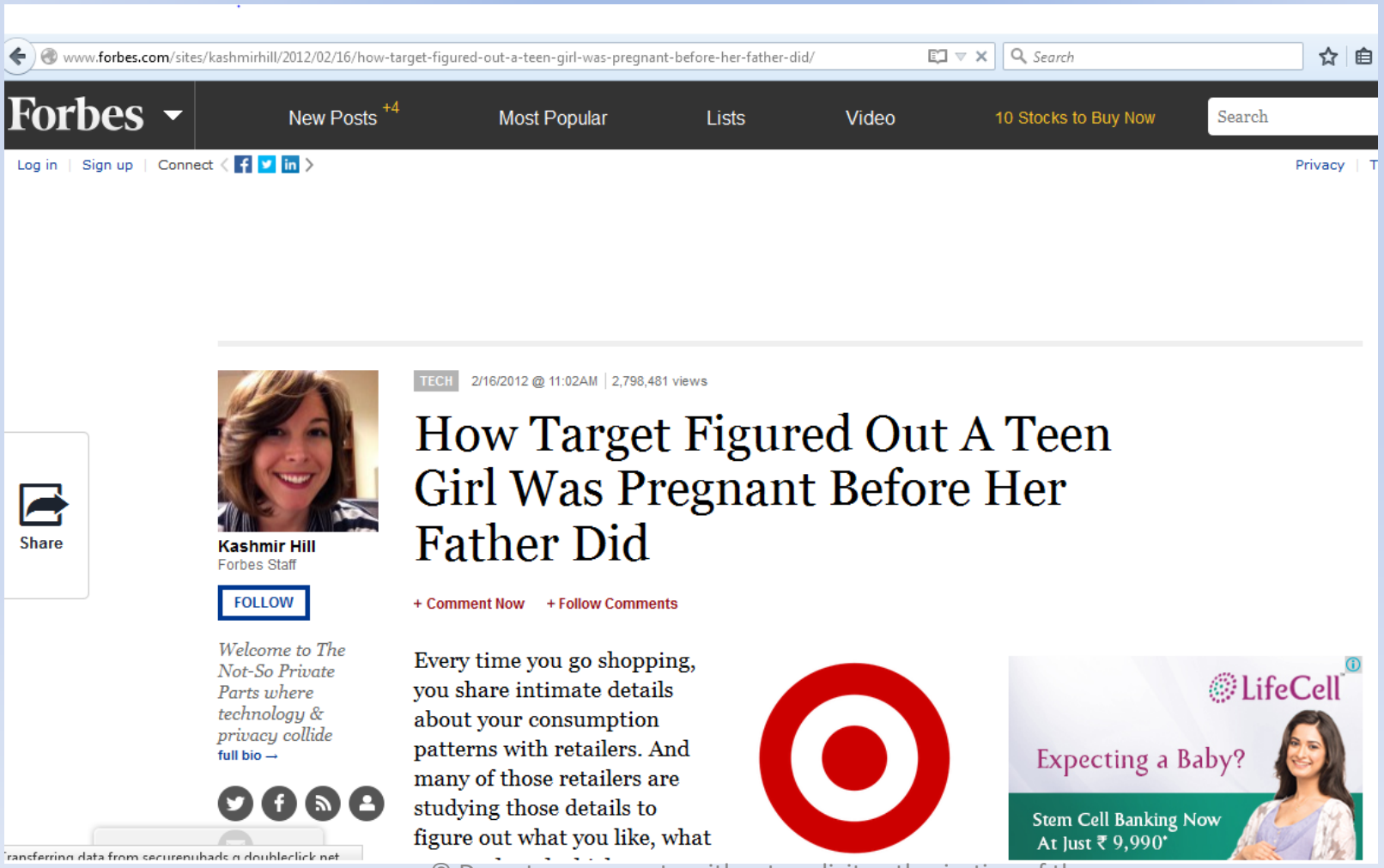
**Face recognition**: Facebook automatically tags people in photos where they appear.

# PLAYERS

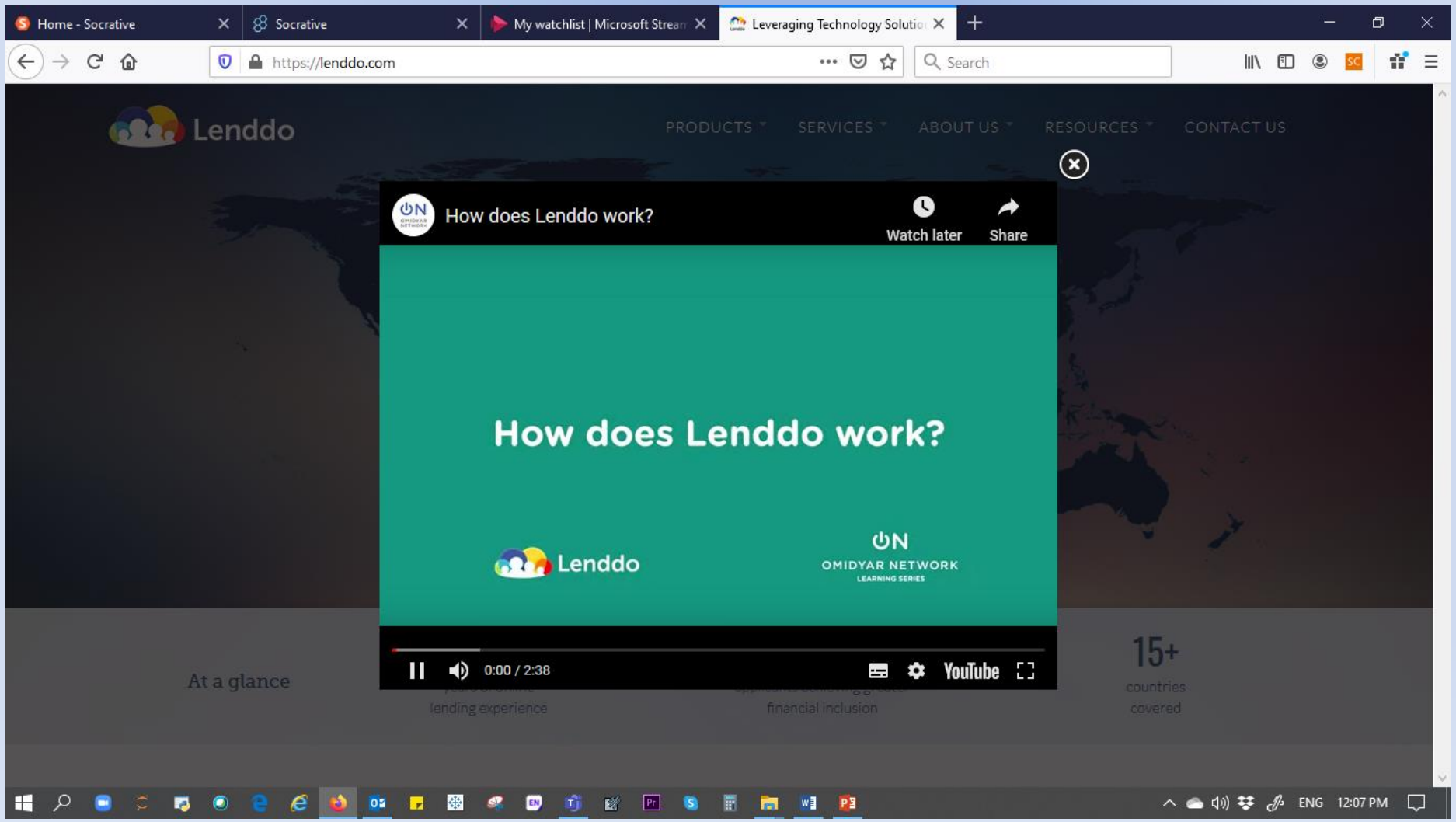
- Thought Leaders
  - Google, Facebook, Amazon
- Data driven firms
  - Uber, Twitter, NBC, Flipkart
- IT giants
  - Catching up to the buzz
  - Infosys, Cognizant, IBM, Accenture.....
- Data analytics focused startups/companies
  - Arcadia, DataHero, Walmart Labs, Mu sigma, Fractal Analytics, Flutura
- Traditional Businesses
  - DNV, Wal-Mart, Sears, DHL

# DATA SCIENCE - APPLICATIONS

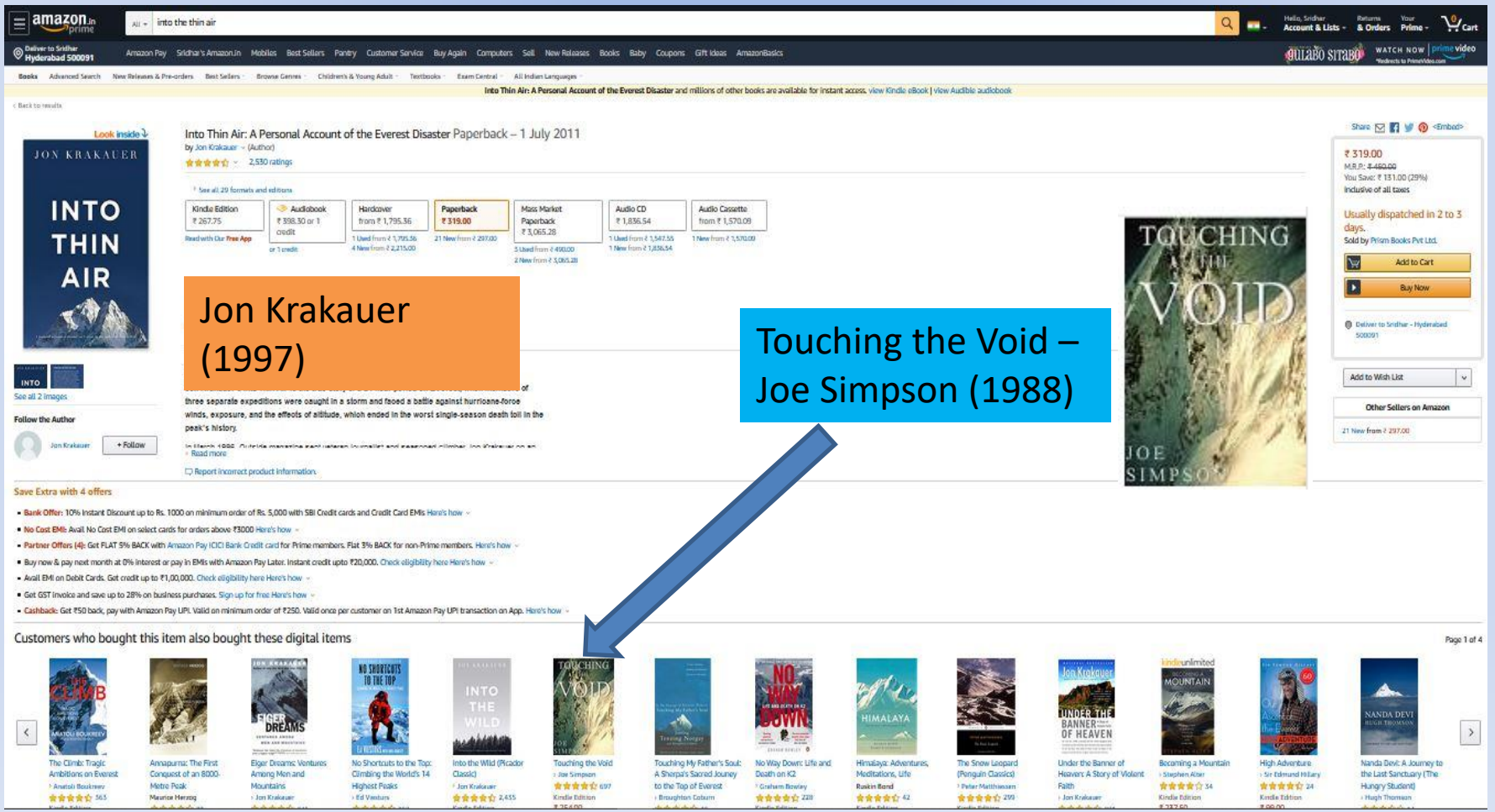
# Target – the mega Store – Pregnancy Prediction Score



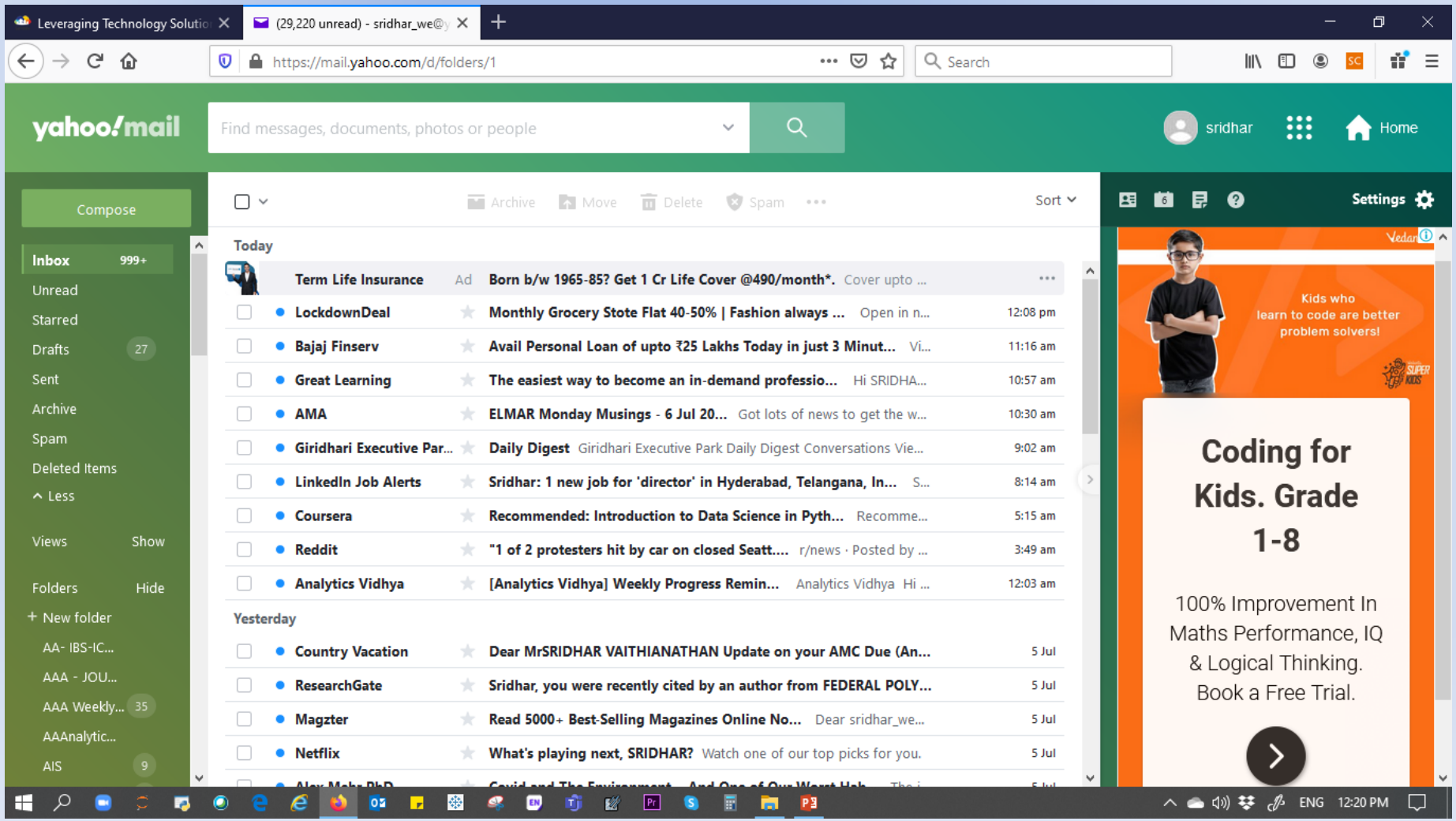
# lenddo



# Amazon



# Personalized Ads and Pop-Ups



# WHY MACHINE LEARNING?

# DATA DRIVEN DECISIONS

- The art and science of leveraging your data to get actionable insights and make better decisions is known as making data driven decisions.

Example (Video – Sophia)

- Context of Discussion
- Recognizing people faces
- Deciding to approve or reject a business deal

# PRESENTATION PROGRESS

**WHAT IS MACHINE LEARNING?**

**ML Definitions**

**TERMINOLOGIES**

**AI, ML & DL**

**DESCRIPTIVE, PREDICTIVE & PRESCRIPTIVE ANALYTICS**

# WHAT IS MACHINE LEARNING?

# What is Machine Learning?

- Arthur Samuel described it as: "the field of study that gives computers the ability to learn without being explicitly programmed."

- Tom Mitchell provides a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

# Basic Example

**Credit Classification - Good, Medium and Bad BY COMPUTER/MACHINE.**

- **E** = the EXPERIENCE of understanding from Datasets (N=10000 Records) with Outcome Variable.
- **T** = the TASK of Credit Classification.
- **P** = the accuracy of Classification (Predicted VS Actual).

# TERMINOLOGIES

- --1-- is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals.
- --3-- is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
- --5-- comprises the strategies and technologies used by enterprises for the data analysis of business information.
- --2-- is the scientific process of transforming data into insights for making better decisions
- --4-- is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining.
- --6-- is the field of study that gives computers the ability to learn without being explicitly programmed.

| | | |
|---|---|---|
| 1. Artificial Intelligence | 3. Data Mining | 5. Business Intelligence |
| 2. Business Analytics | 4. Data Science | 6. Machine Learning |

# AI, ML & DL

- **Artificial Intelligence :** Algorithms and Systems that exhibit human-like intelligence.

- **Machine Learning :** Subset of AI that can learn to perform a task with extracted data and/or models.

- **Deep Learning :** Subset of machine learning that imitate the functioning of human brain to solve problems.

# Descriptive, Predictive and Prescriptive Techniques

- **Descriptive Techniques**
  - **What the data is Telling ?**

- **Predictive Techniques**
  - **Why it is happening ?**

- **Prescriptive Techniques**
  - **What we should be doing?**

# PRESENTATION PROGRESS

**ML AIgorithms – Classification**

**ML Pipeline**

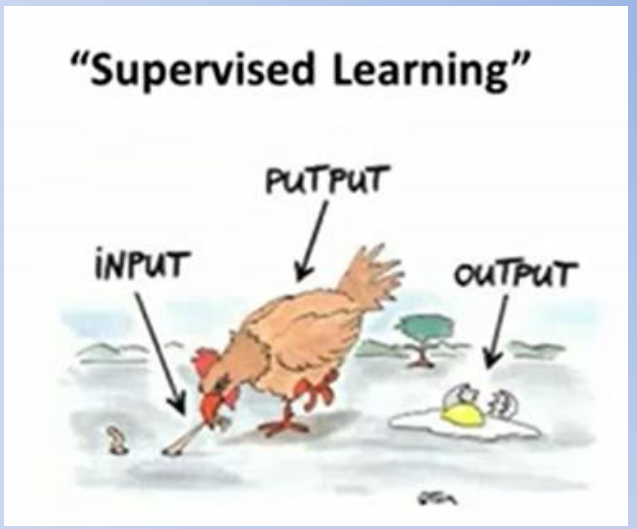
- Supervised Learning
- Unsupervised Learning

Analytics Career Planning

- TO DO LIST!

# STANDARD CLASSIFICATION OF MACHINE LEARNING TECHNIQUES

# MACHINE LEARNING TECHNIQUES





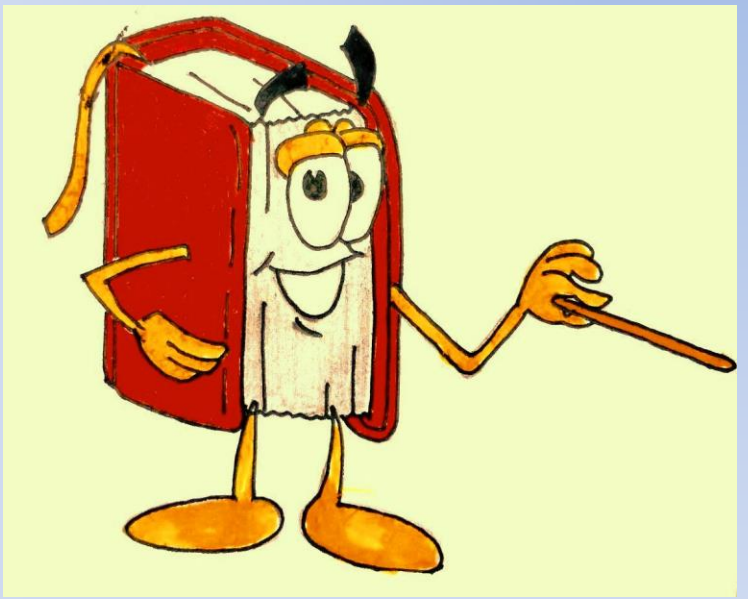
- PREDICTION (NUMERICAL Y)
- CLASSIFICATION (CATEGORICAL Y)

- DIMENSION REDUCTION
- SEGMENTATION
- "WHAT GOES WITH WHAT"

## REINFORCEMENT LEARNING :

**Example : Spell Check – "Buutiful" – "Beautiful / Bountiful/Dutiful**
**Technique : Markov Chain.**
**(Sequential Action to maximize a cumulative Reward)**

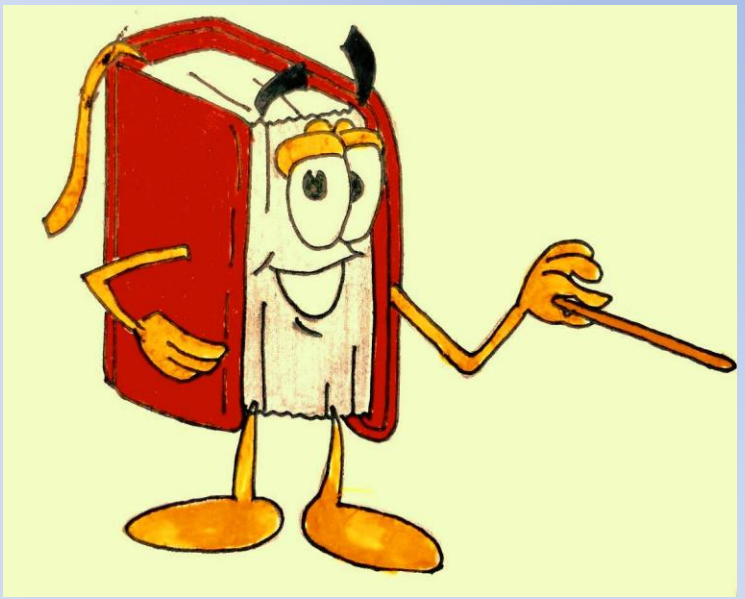
# MLA Tasks



- Answer: This is supervised learning, because the database includes whether the loan was approved or not.

Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).

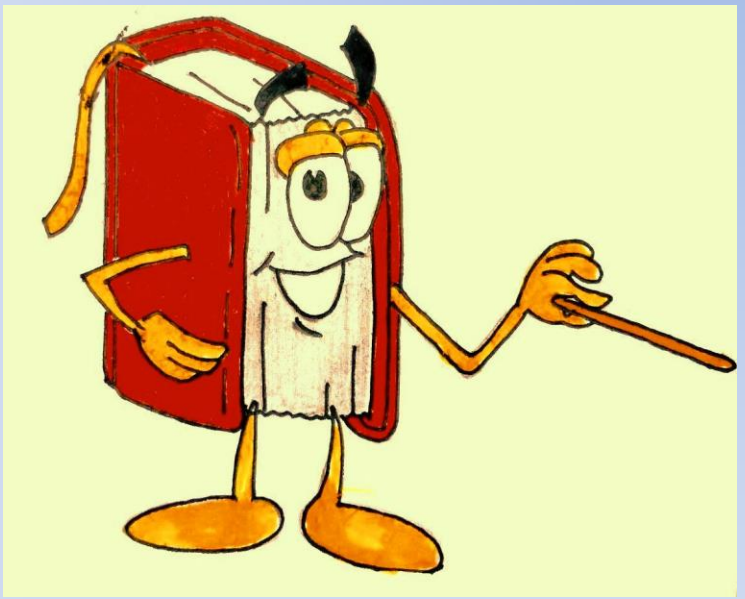
# MLA Tasks



- Answer: This is unsupervised learning, because there is no apparent outcome (e.g., whether the recommendation was adopted or not).

In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.

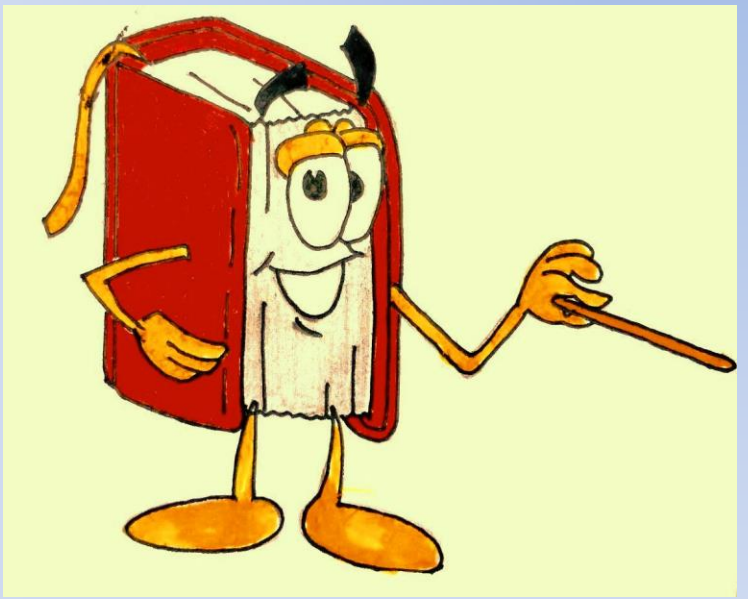
# MLA Tasks



- Answer: This is unsupervised learning because there is no known outcome (though once you use unsupervised learning to identify segments, you could use supervised learning to classify new customers into those segments).

Identifying segments of similar customers.

# MLA Tasks

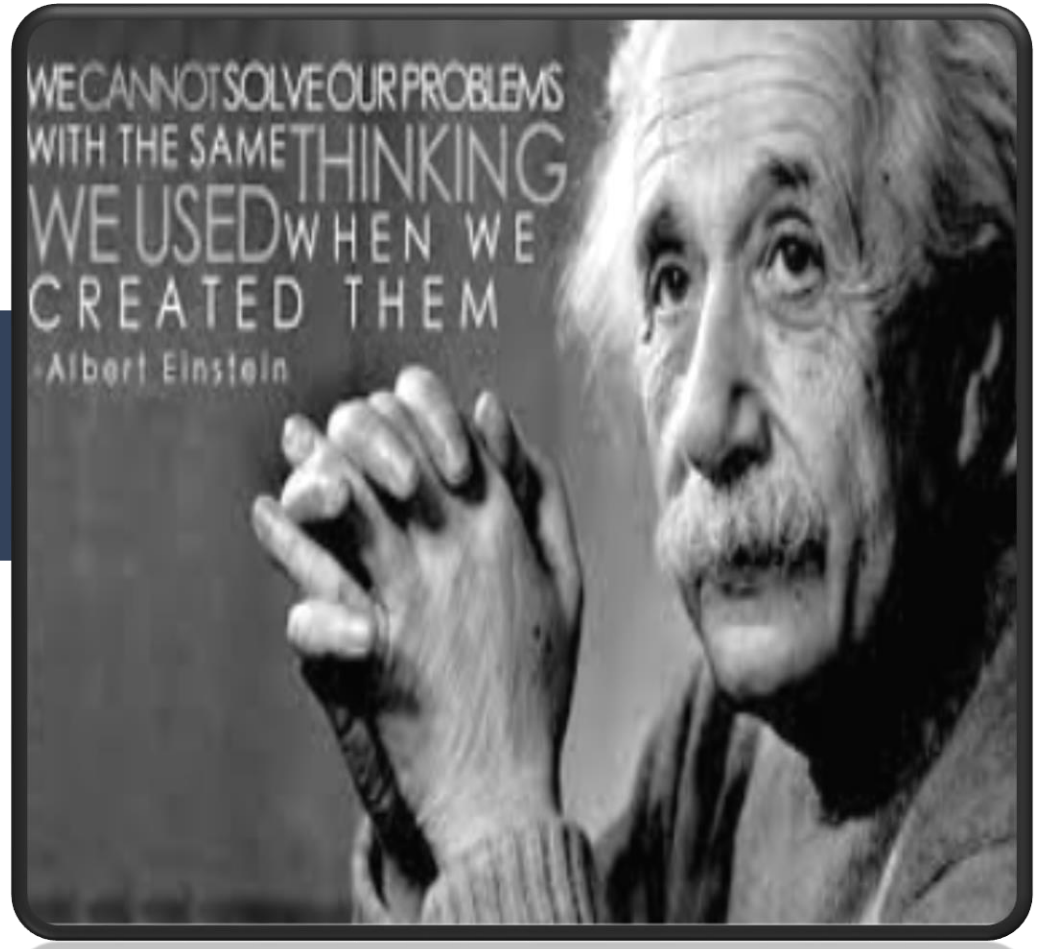


- Answer: This is supervised learning, because the status of the similar firms is known.

Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.

# SESSION OUTLINE

## ANALYTICS PIPELINE MANAGEMENT

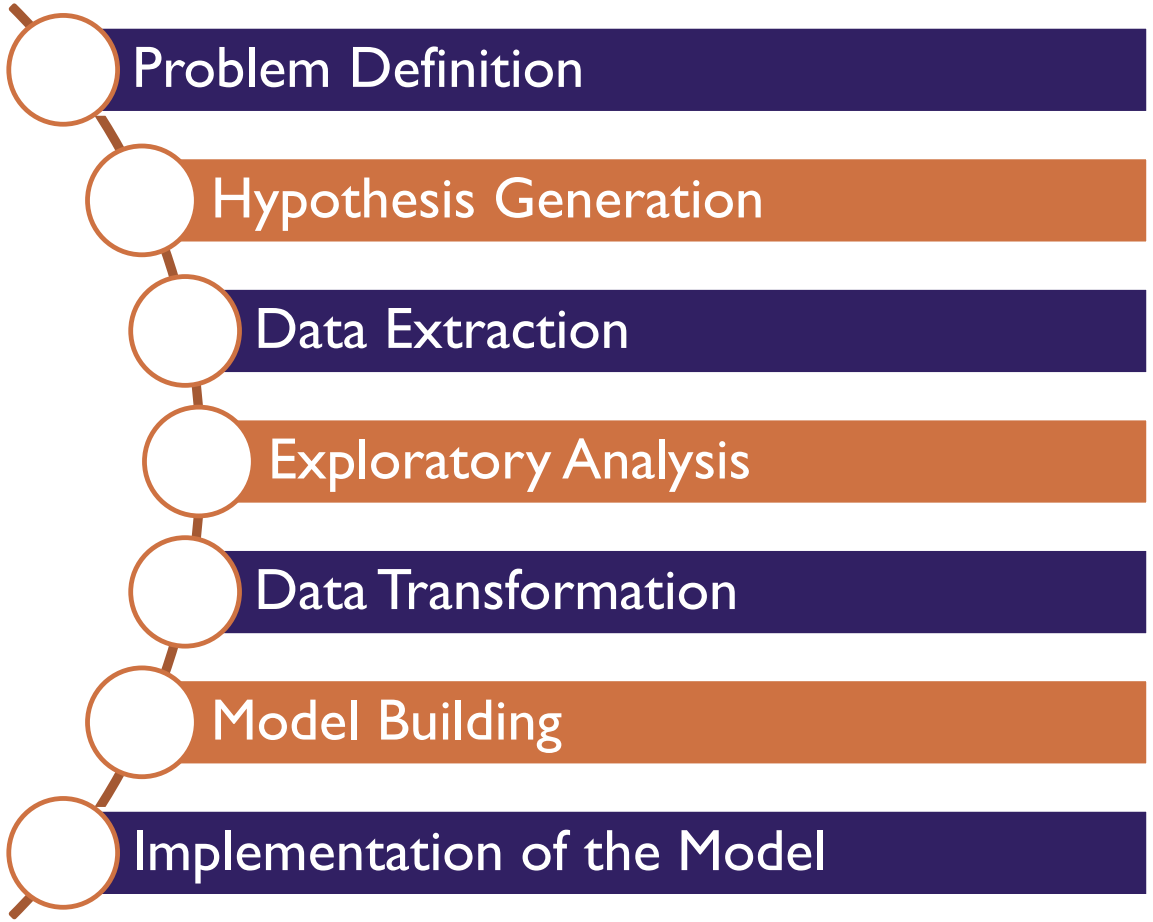


# ANALYTICS PIPELINE MANAGEMENT


THE CRISP-DM PROCESS

- Problem Definition
- Hypothesis Generation
- Data Extraction
- Exploratory Analysis
- Data Transformation
- Model Building
- Implementation of the Model

# ANALYTICS PIPELINE MANAGEMENT
## - 7 STEPS



# ANALYTICS PIPELINE MANAGEMENT

## BUSINESS LEADERS & DATA SCIENTISTS

- Business leaders /owners expect Data Scientists to solve Business Problems
- It is the responsibility of the data scientist to convert these ambiguous business problems to data problems

## EXAMPLE – ECOMMERCE COMPANY

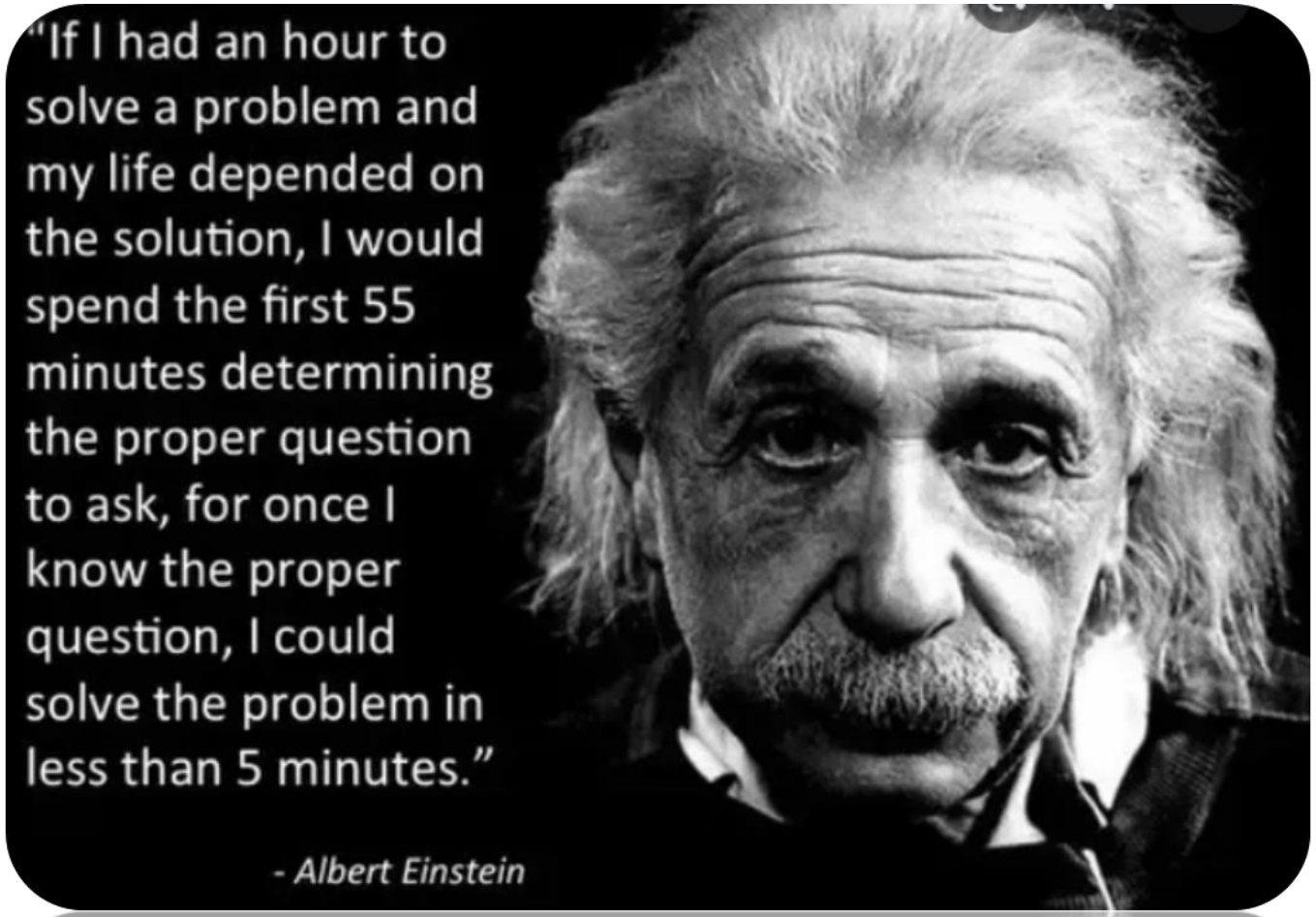
LEADERS: Business Problem:

I want to increase the Revenue of our Business by 25% without increasing the costs

Refer : The DS LC – Project Template

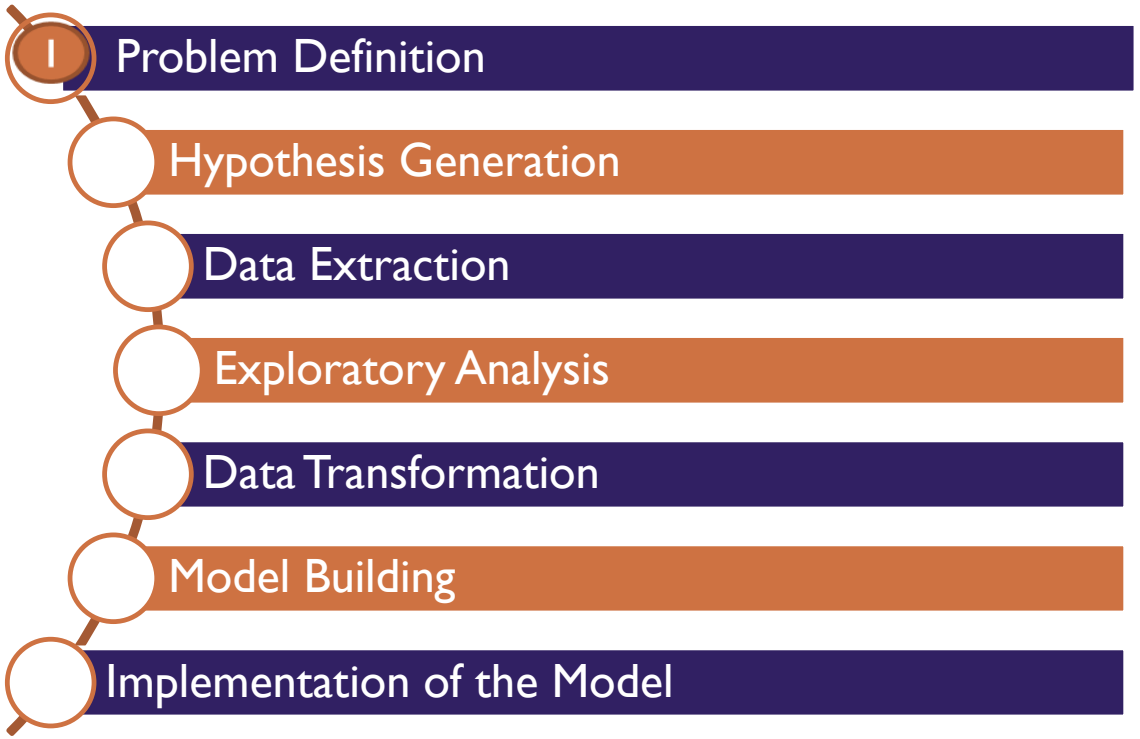
# EXAMPLE - ANALYTICS PIPELINE MANAGEMENT

# ANALYTICS PIPELINE MANAGEMENT
## I. PROBLEM DEFINITION



# ANALYTICS PIPELINE MANAGEMENT

## 1. PROBLEM DEFINITION



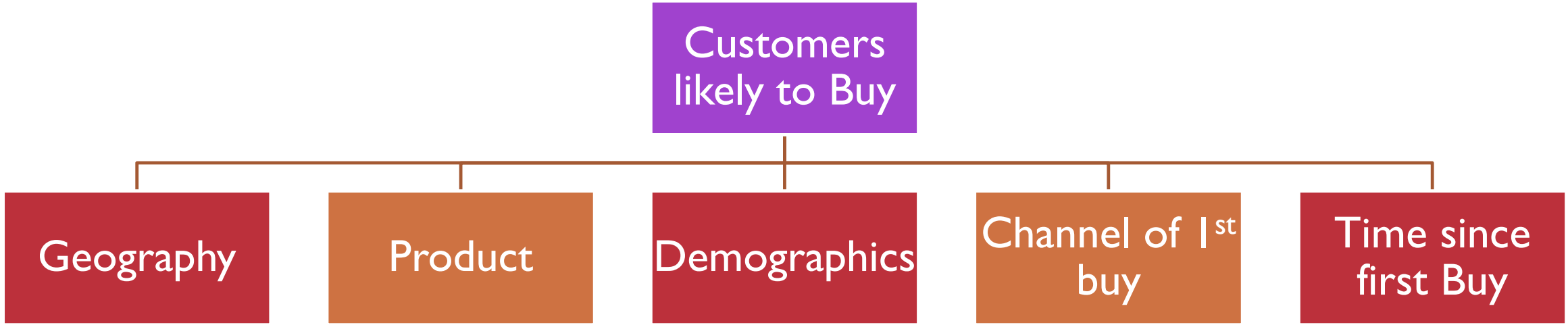

**BUSINESS PROBLEM TO DATA PROBLEM**

**DATA PROBLEM :**

- Can I identify FROM THE EXISTING CUSTOMERS who are likely to buy more products from us?

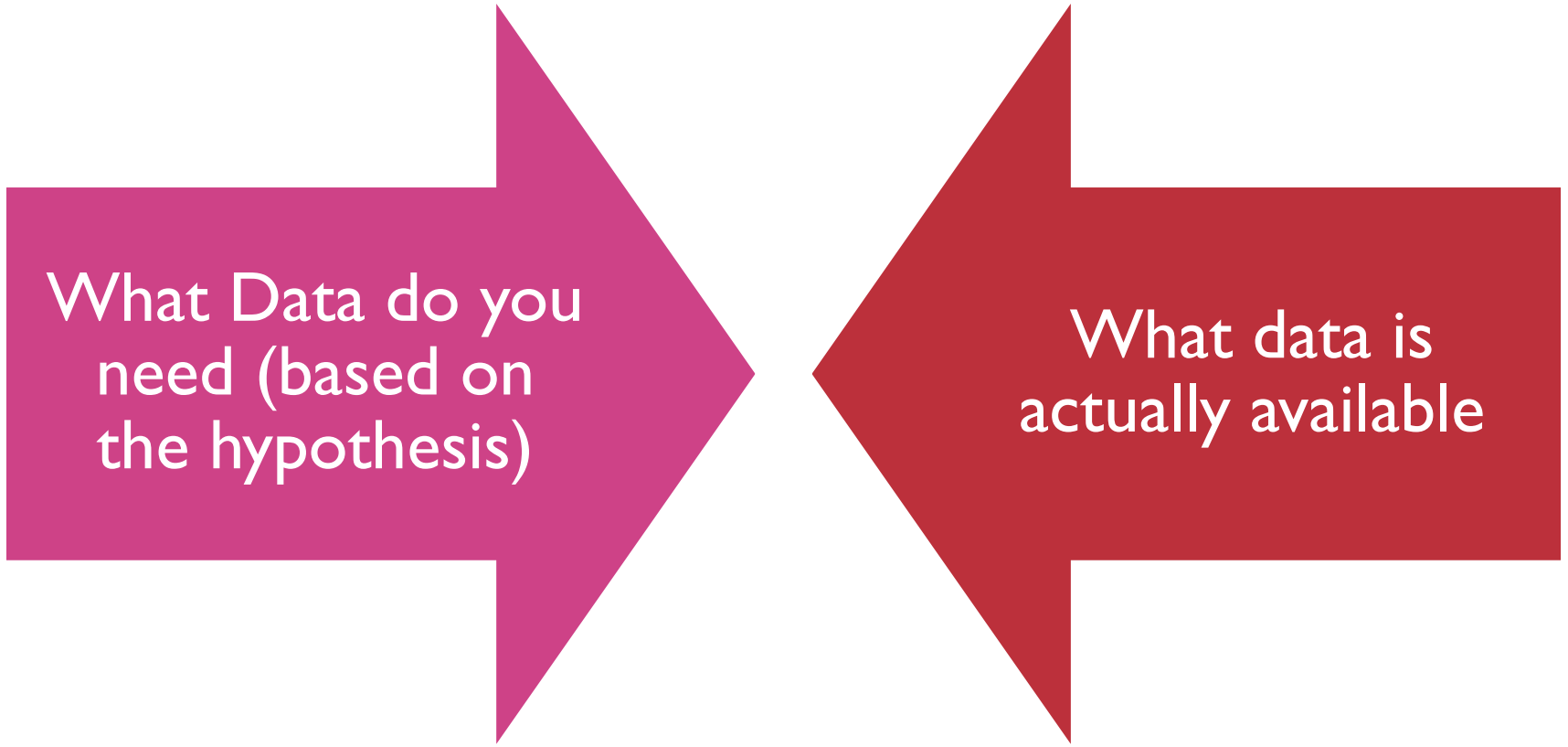
# ANALYTICS PIPELINE MANAGEMENT
## 2. HYPOTHESIS GENERATION

Which existing customers are more likely to buy again from us?



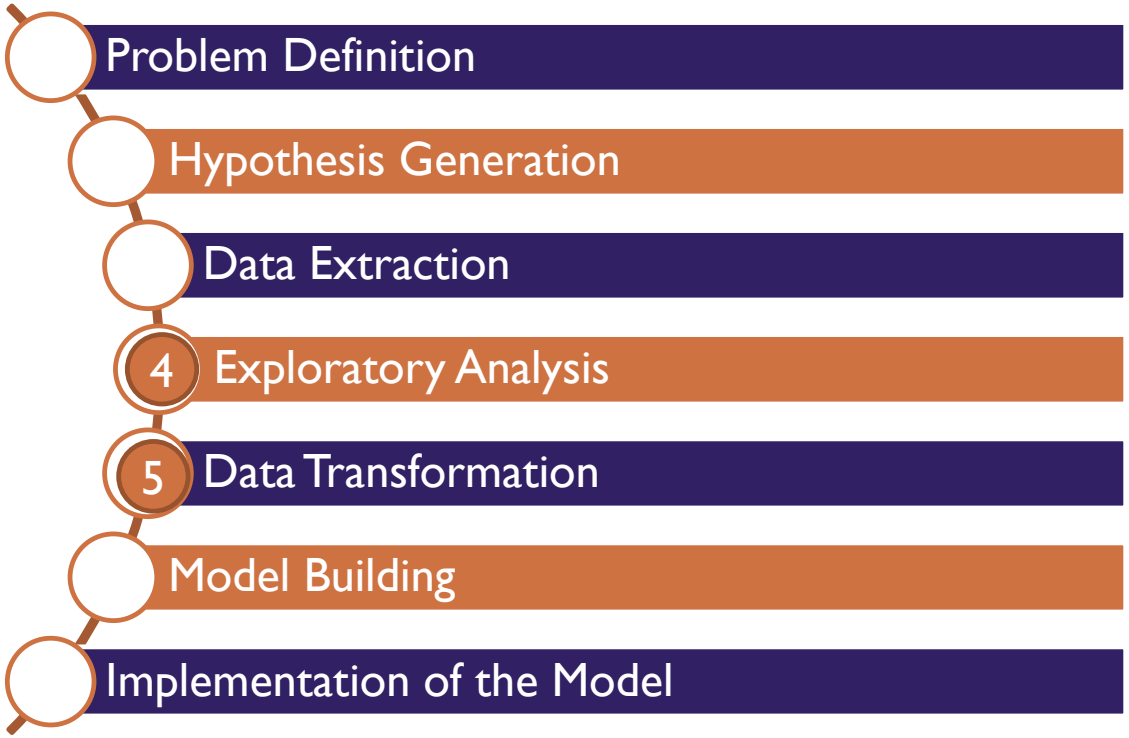
# ANALYTICS PIPELINE MANAGEMENT

## 3. DATA EXTRACTION



# ANALYTICS PIPELINE MANAGEMENT

## 4. EXPLORATORY ANALYSIS & 5. DATA TRANSFORMATION




- What is the relationship between different variables?
- What transformations are required?

# ANALYTICS PIPELINE MANAGEMENT

## 6. MODEL BUILDING

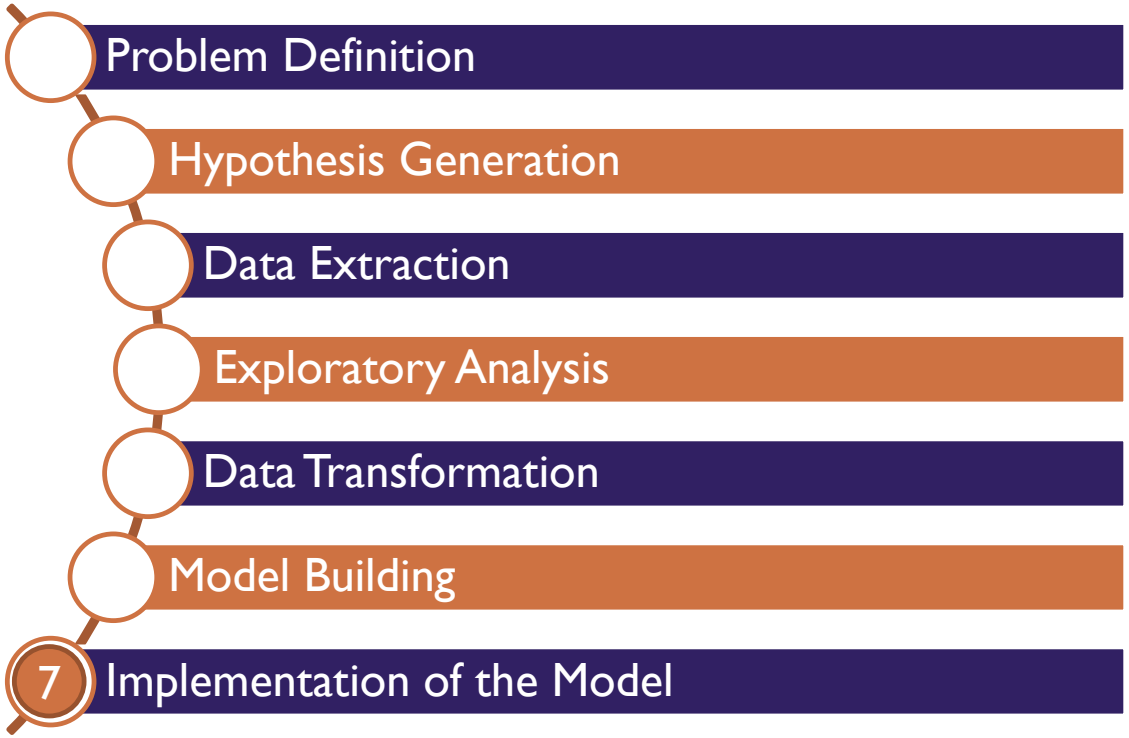
Which is the right evaluation metric for our problem?

What kind of models should we build?

What's our model validation strategy?

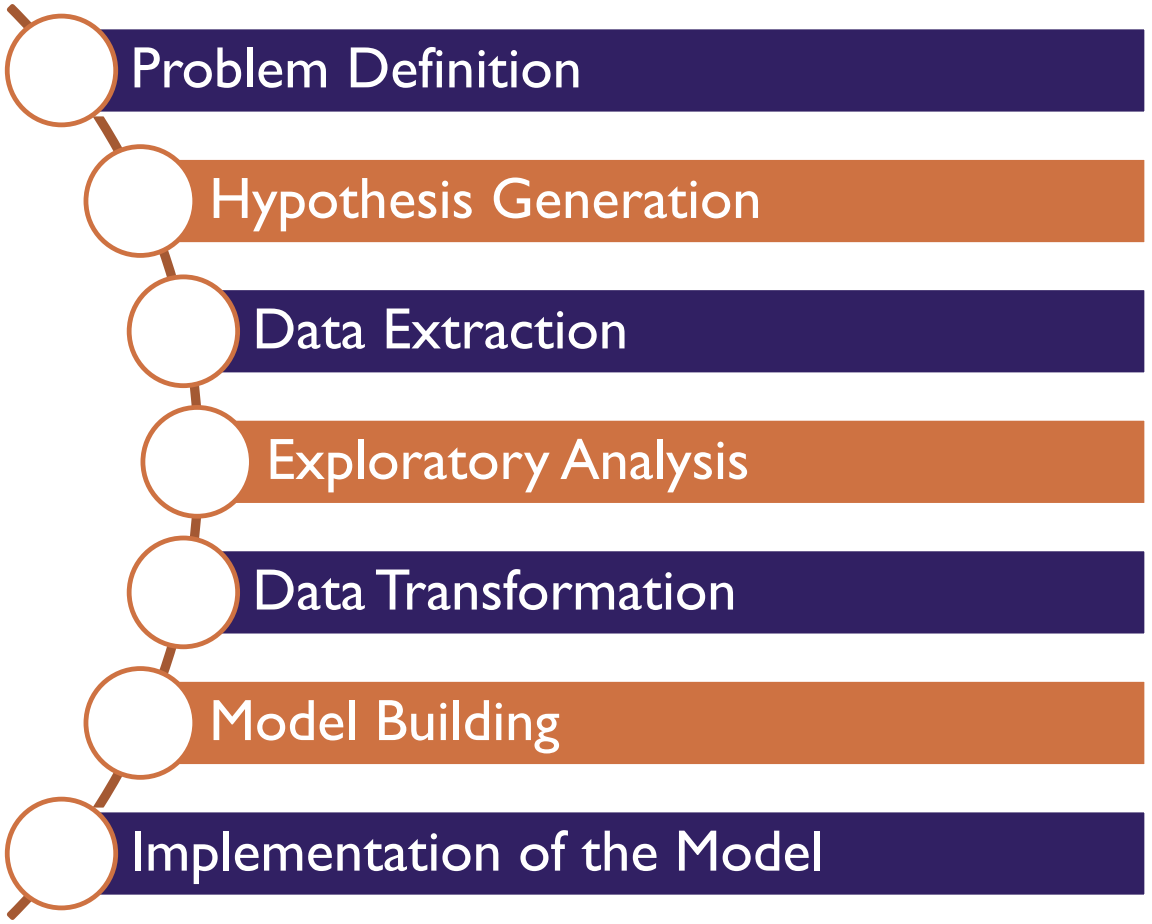
# ANALYTICS PIPELINE MANAGEMENT

## 7. IMPLEMENTATION OF THE MODEL




- Which model should we choose?
- Understand the business metric versus the statistical metric
- Finalize a framework for monitoring our model.

# SUMMARY - **ANALYTICS PIPELINE MANAGEMENT**



# ANALYTICS PIPELINE MANAGEMENT APPLICATION - EXAMPLE

CASE

OF

NETFLIX


If you don’t challenge yourself, you will never realize what you can become

# THANKS FOR LISTENING &
# ALL THE VERY BEST !