

18<sup>th</sup> 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](https://in.linkedin.com/in/sridharvaithianathan)

**Email:** [prof.sridhar.we@gmail.com](mailto: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

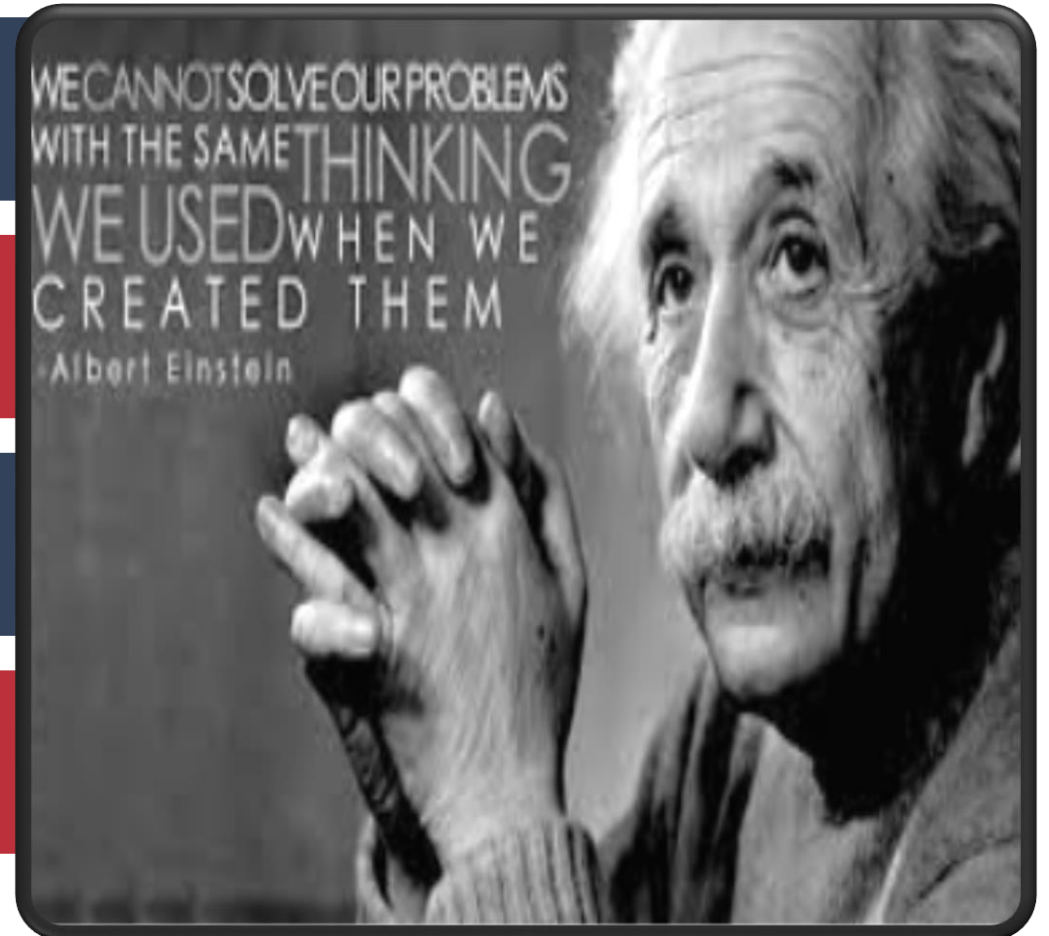
# SESSION OUTLINE

**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**



© Do not use or quote without explicit authorization of the author (Prof. Sridhar Vaithianathan)

# 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

**WITH GREAT DATA**



**COMES GREAT RESPONSIBILITY**

r quote without explicit  
he author (Prof. Sridhar  
Vaithianathan)

**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

The screenshot shows a web browser displaying a Forbes article. The browser's address bar shows the URL: [www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/](http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/). The Forbes logo is in the top left, and navigation links for 'New Posts', 'Most Popular', 'Lists', 'Video', and '10 Stocks to Buy Now' are in the top right. Below the navigation bar, there are links for 'Log in', 'Sign up', and 'Connect' with social media icons. The article itself is by Kashmir Hill, a Forbes Staff member, and is categorized under 'TECH'. It was published on 2/16/2012 at 11:02AM and has 2,798,481 views. The article title is 'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did'. The main text begins with 'Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what'. A large red Target logo is positioned to the right of the text. Below the text, there is an advertisement for LifeCell, featuring a woman and the text 'Expecting a Baby? Stem Cell Banking Now At just ₹ 9,990\*'. A 'Share' button is visible on the left side of the article.

© Do not use or quote without explicit authorization of the author (Prof. Sridhar Vaithianathan)

# lenddo

The screenshot shows a web browser window with the URL <https://lenddo.com>. The browser tabs include "Home - Socrative", "Socrative", "My watchlist | Microsoft Stream", and "Leveraging Technology Solution". The website header features the Lenddo logo and navigation links: PRODUCTS, SERVICES, ABOUT US, RESOURCES, and CONTACT US. A video player is overlaid on the page, displaying the title "How does Lenddo work?". The video player interface includes a play button, a progress bar at 0:00 / 2:38, and a YouTube logo. The video content shows a green background with the text "How does Lenddo work?" and the Lenddo logo. The Omidyar Network Learning Series logo is also visible. The video player has a close button (X) in the top right corner. The background of the website shows a world map and a section titled "At a glance" with the text "lending experience" and "financial inclusion". On the right side, it says "15+ countries covered". The Windows taskbar at the bottom shows various application icons and the system tray with the time 12:07 PM and language ENG.

# Amazon

The screenshot shows the Amazon India product page for 'Into Thin Air: A Personal Account of the Everest Disaster' by Jon Krakauer. The page includes the Amazon logo, search bar, navigation menu, and product details. The main product is 'Into Thin Air: A Personal Account of the Everest Disaster Paperback – 1 July 2011' by Jon Krakauer, priced at ₹319.00. The page also features a 'Customers who bought this item also bought these digital items' section with a carousel of related books, including 'Touching the Void' by Joe Simpson. A blue arrow points from a text box 'Touching the Void – Joe Simpson (1988)' to the 'Touching the Void' book in the carousel.

**Into Thin Air: A Personal Account of the Everest Disaster Paperback – 1 July 2011**  
by Jon Krakauer (Author)  
★★★★☆ 2,530 ratings

See all 20 formats and editions

Kindle Edition ₹ 267.75 Read with Our Free App	Audiobook ₹ 398.30 or 1 credit	Hardcover from ₹ 1,795.36 1 Used from ₹ 1,795.36 4 New from ₹ 2,215.00	<b>Paperback</b> ₹ 319.00 21 New from ₹ 297.00	Mass Market Paperback ₹ 3,065.28 5 Used from ₹ 490.00 2 New from ₹ 5,985.28	Audio CD ₹ 1,836.54 1 Used from ₹ 1,547.35 1 New from ₹ 1,836.54	Audio Cassette from ₹ 1,570.09 1 New from ₹ 1,570.09
--	-----------------------------------	---	--	--	---	--

**Jon Krakauer (1997)**

**Touching the Void – Joe Simpson (1988)**

Customers who bought this item also bought these digital items

- The Climb: Tragic Ambitions on Everest
- Into Thin Air: A Personal Account of the Everest Disaster
- Edger: Osprey Ventures Among Men and Mountains
- No Shortcuts to the Top: Climbing the World's 14 Highest Peaks
- Into the Wild (Picador Classic)
- Touching the Void
- Touching My Father's Soul: A Sheep's Sacred Journey to the Top of Everest
- No Way Down: Life and Death on K2
- Himalaya: Adventures, Meditations, Life
- The Snow Leopard (Penguin Classics)
- Under the Banner of Heaven: A Story of Violent Faith
- Becoming a Mountain
- High Adventure
- Nanda Devi: A Journey to the Last Sanctuary (The Hungry Student)

# Personalized Ads and Pop-Ups

The screenshot displays a Yahoo! Mail inbox with several personalized advertisements. The ads include:

- Term Life Insurance Ad:** Born b/w 1965-85? Get 1 Cr Life Cover @490/month\*. Cover upto ...
- LockdownDeal:** Monthly Grocery Stote Flat 40-50% | Fashion always ...
- Bajaj Finserv:** Avail Personal Loan of upto ₹25 Lakhs Today in just 3 Minut...
- Great Learning:** The easiest way to become an in-demand professio...
- AMA:** ELMAR Monday Musings - 6 Jul 20... Got lots of news to get the w...
- Giridhari Executive Par...:** Daily Digest Giridhari Executive Park Daily Digest Conversations Vie...
- LinkedIn Job Alerts:** Sridhar: 1 new job for 'director' in Hyderabad, Telangana, In...
- Courseera:** Recommended: Introduction to Data Science in Pyth...
- Reddit:** "1 of 2 protesters hit by car on closed Seatt... r/news - Posted by ...
- Analytics Vidhya:** [Analytics Vidhya] Weekly Progress Remin... Analytics Vidhya Hi ...

Below the ads, there are messages from "Yesterday" including:

- Country Vacation:** Dear MrSRIDHAR VAITHIANATHAN Update on your AMC Due (An...
- ResearchGate:** Sridhar, you were recently cited by an author from FEDERAL POLY...
- Magzter:** Read 5000+ Best-Selling Magazines Online No... Dear sridhar\_we...
- Netflix:** What's playing next, SRIDHAR? Watch one of our top picks for you.

On the right side of the inbox, a large orange pop-up advertisement is visible for "Coding for Kids". The ad features a young boy and the text:

Kids who learn to code are better problem solvers!

**Coding for Kids. Grade 1-8**

100% Improvement In Maths Performance, IQ & Logical Thinking. Book a Free Trial.

## 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  
2. Business Analytics

3. Data Mining  
4. Data Science

5. Business Intelligence  
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 Algorithms – Classification

### ML Pipeline

- Supervised Learning
- Unsupervised Learning

### Analytics Career Planning

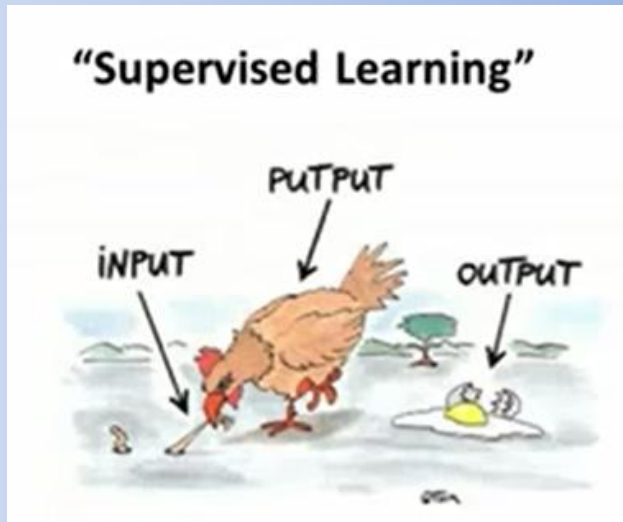
- TO DO LIST!



# STANDARD CLASSIFICATION OF MACHINE LEARNING TECHNIQUES



# MACHINE LEARNING TECHNIQUES



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



- DIMENSION REDUCTION
- SEGMENTATION
- “WHAT GOES WITH WHAT”

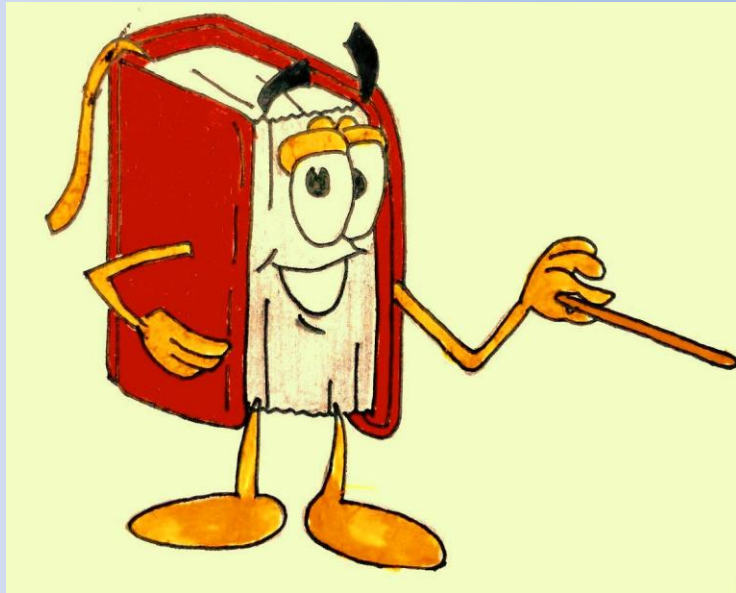
## **REINFORCEMENT LEARNING :**

Example : Spell Check – “Buutiful” – “Beautiful / Bountiful/Dutiful

Technique : Markov Chain.

(Sequential Action to maximize a cumulative Reward)

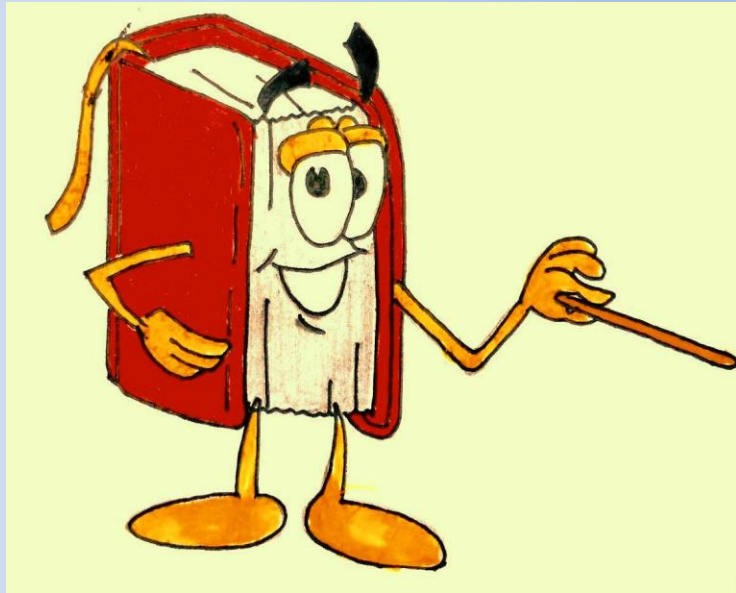
# MLA Tasks



- **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).

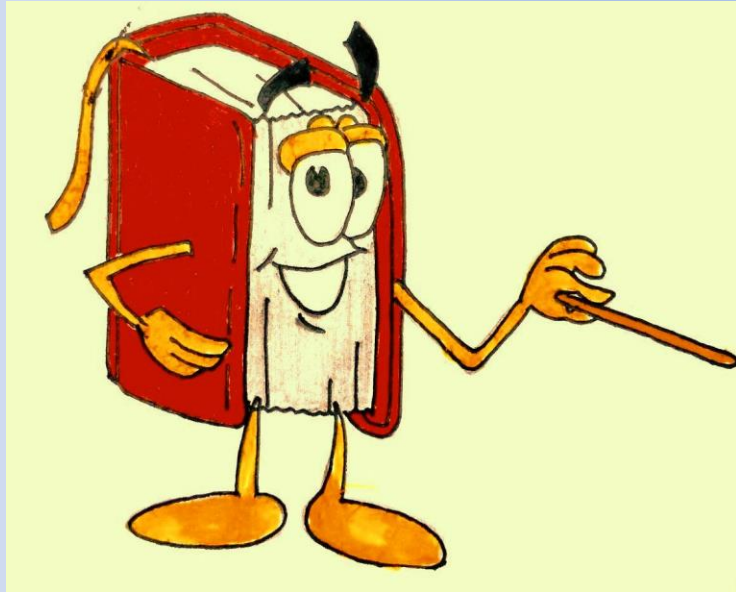
# MLA Tasks



- **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.

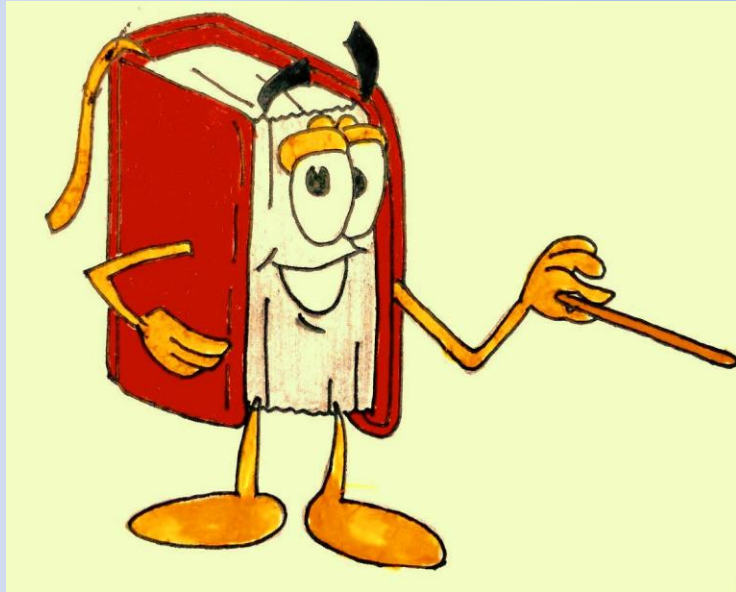
# MLA Tasks



- **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

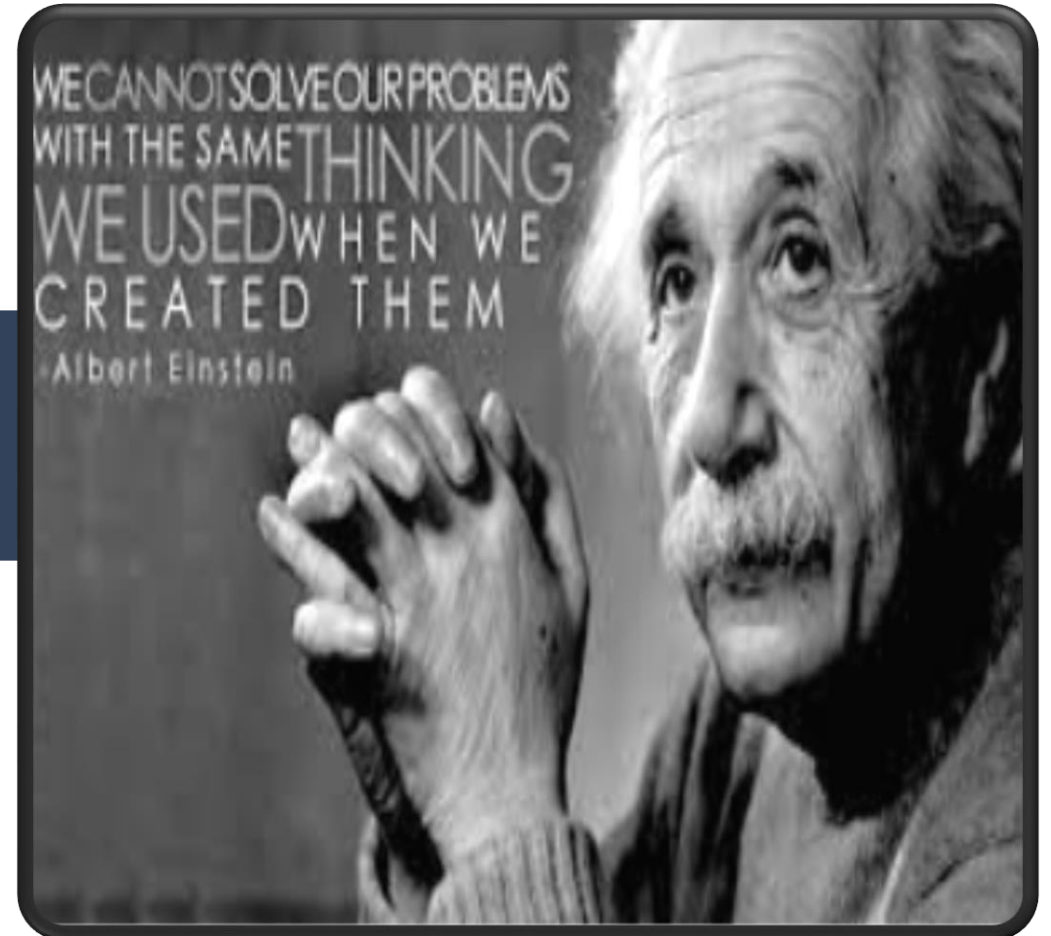


- **Answer:** This is supervised learning, because the status of the similar firms is known.

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

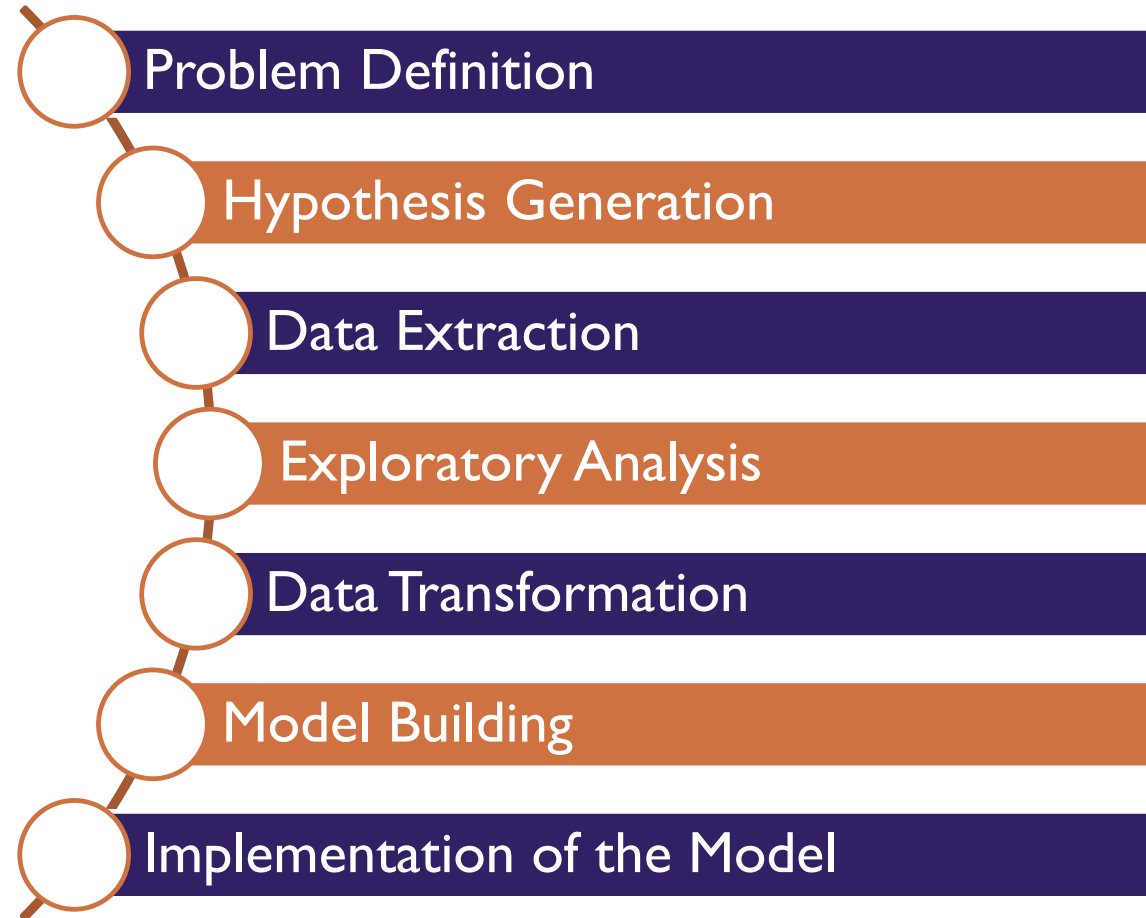


- Problem Definition
- Hypothesis Generation
- Data Extraction
- Exploratory Analysis
- Data Transformation
- Model Building
- Implementation of the Model

THE CRISP-DM PROCESS

# 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

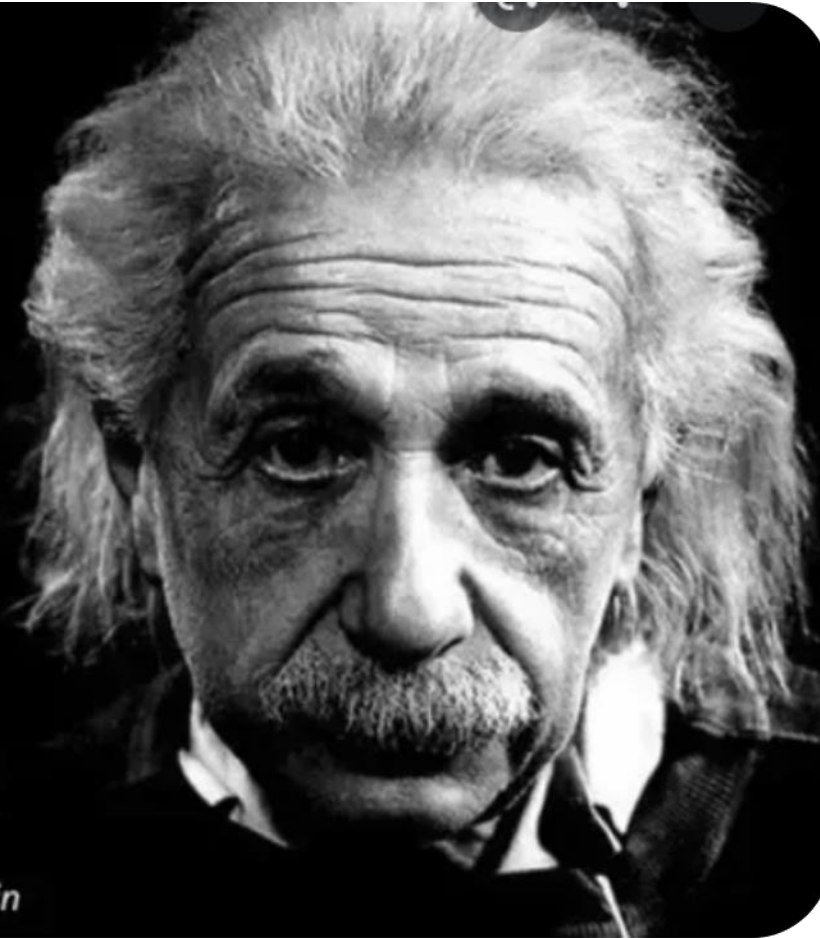
## EXAMPLE - ANALYTICS PIPELINE MANAGEMENT

# ANALYTICS PIPELINE MANAGEMENT

## I. PROBLEM DEFINITION

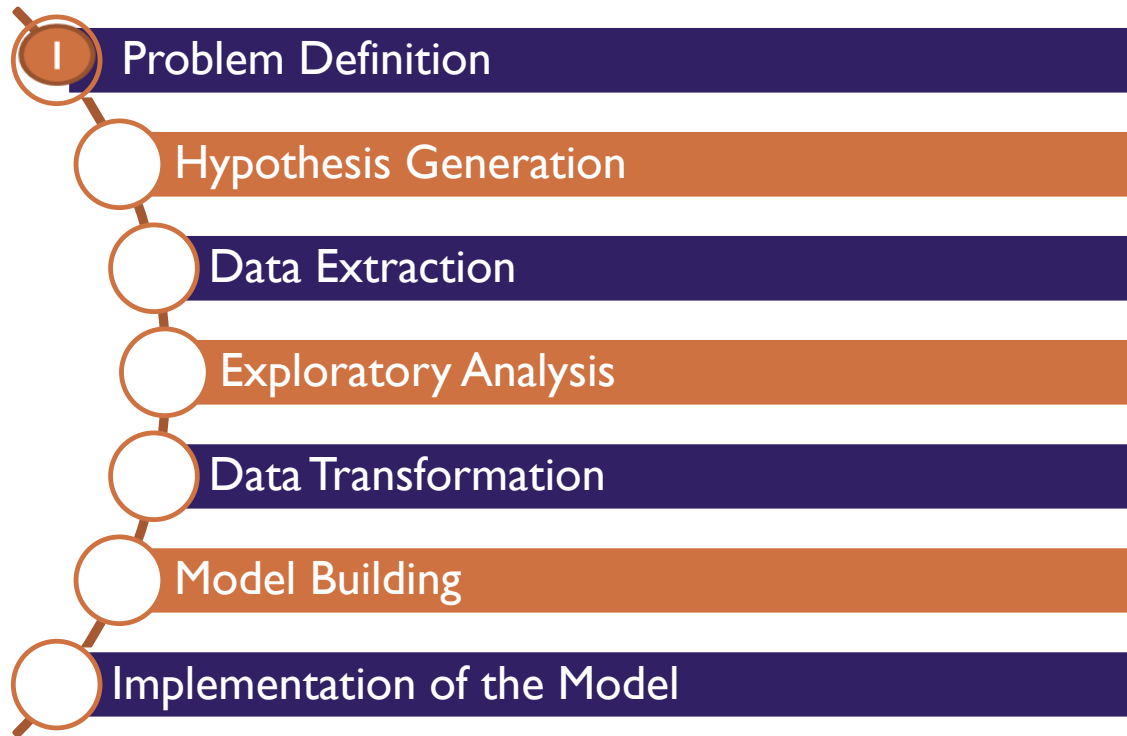
"If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than 5 minutes."

- Albert Einstein



# ANALYTICS PIPELINE MANAGEMENT

## I. PROBLEM DEFINITION



## **BUSINESS PROBLEM TO DATA PROBLEM**

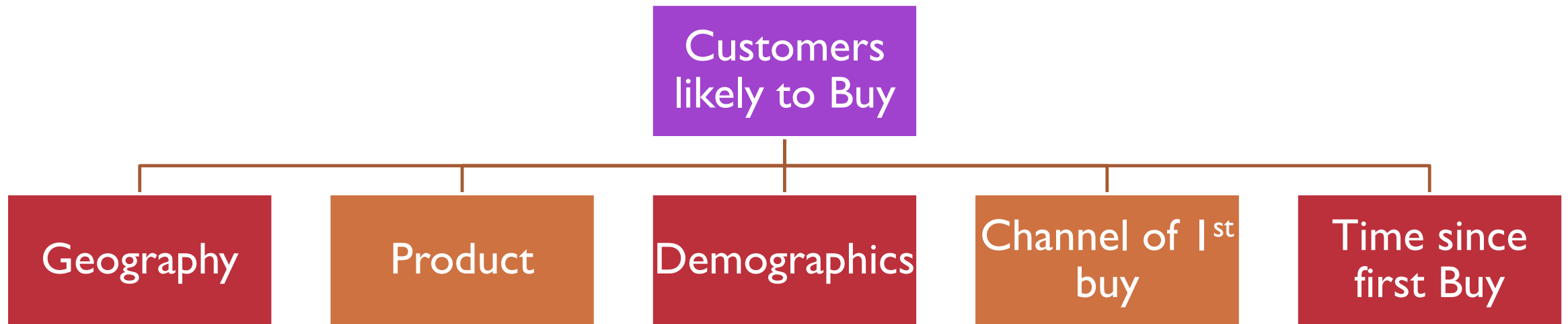
### **DATA PROBLEM :**

- 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

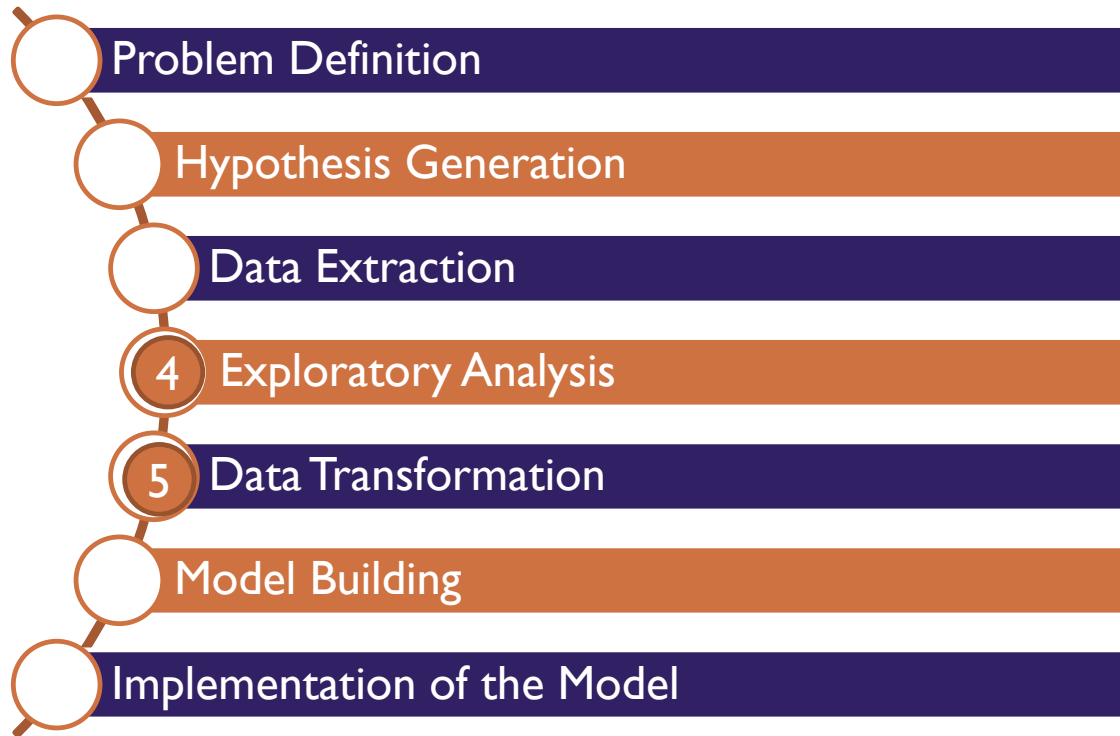
What Data do you need (based on the hypothesis)

The diagram consists of two large, stylized arrows pointing towards each other. The left arrow is purple and points right, containing the text 'What Data do you need (based on the hypothesis)'. The right arrow is red and points left, containing the text 'What data is actually available'. The space between the two arrows represents the data extraction process.

What data is actually available

# 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?



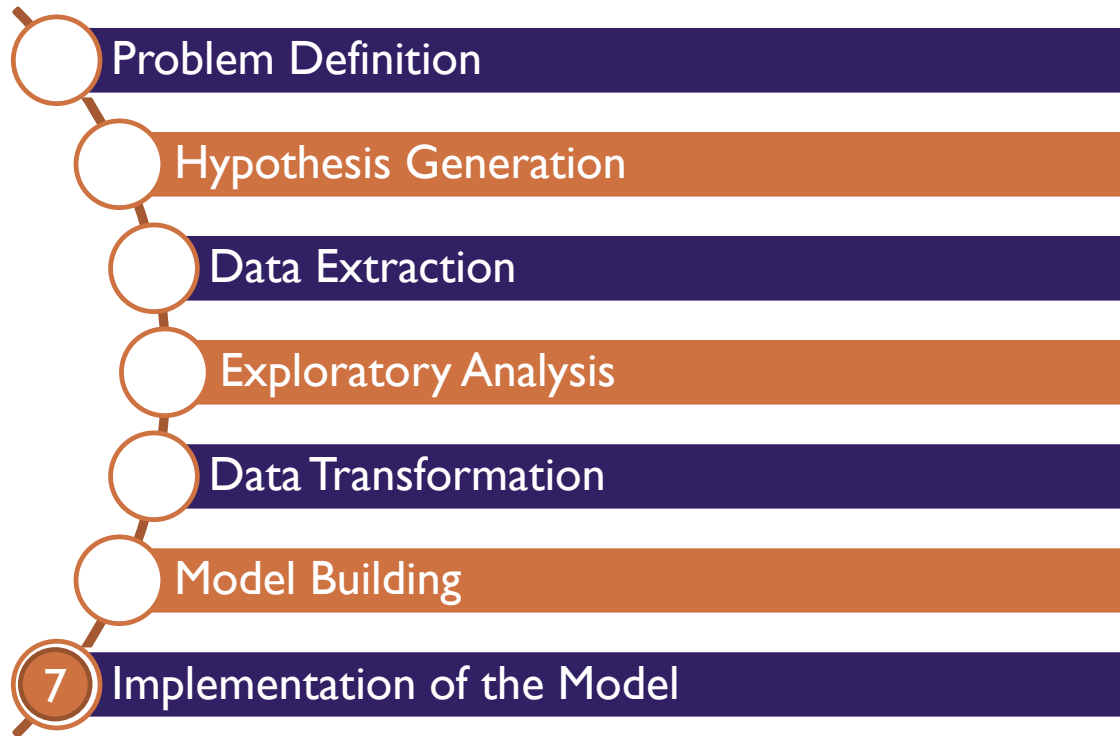
What kind of models should we build?



What's our model validation strategy?

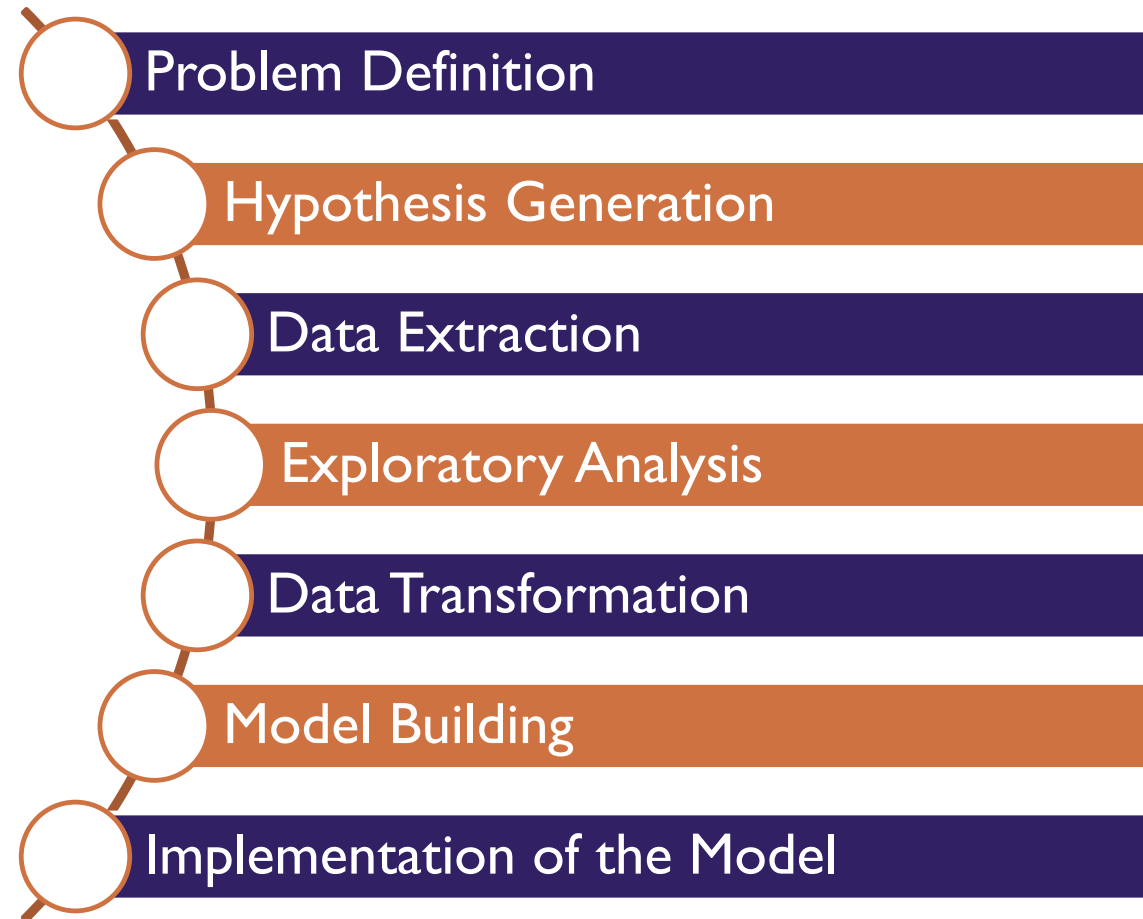
# ANALYTICS PIPELINE MANAGEMENT

## 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



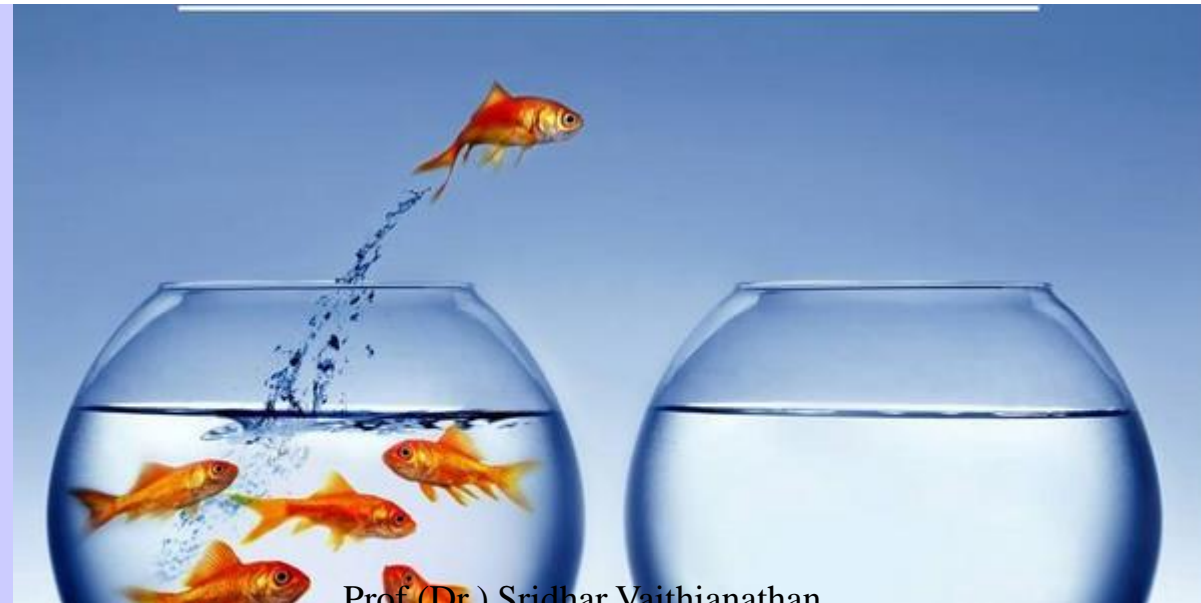
# ANALYTICS PIPELINE MANAGEMENT APPLICATION - EXAMPLE

CASE  
OF  
NETFLIX



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

**THANKS FOR LISTENING &  
ALL THE VERY BEST !**



Prof.(Dr.) Sridhar Vaithianathan,