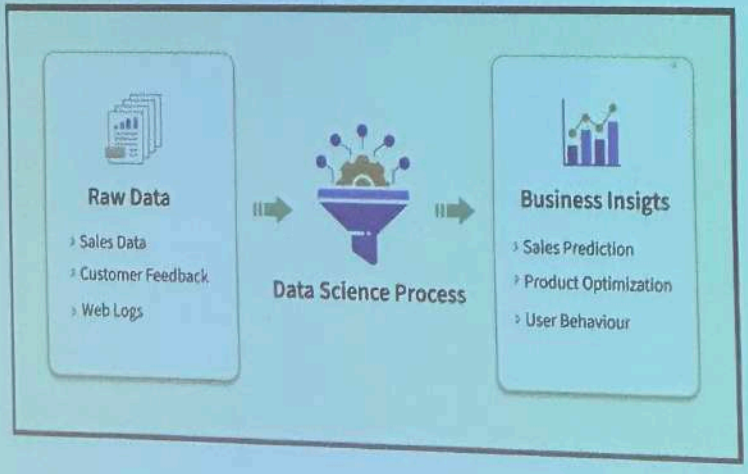
Data Science

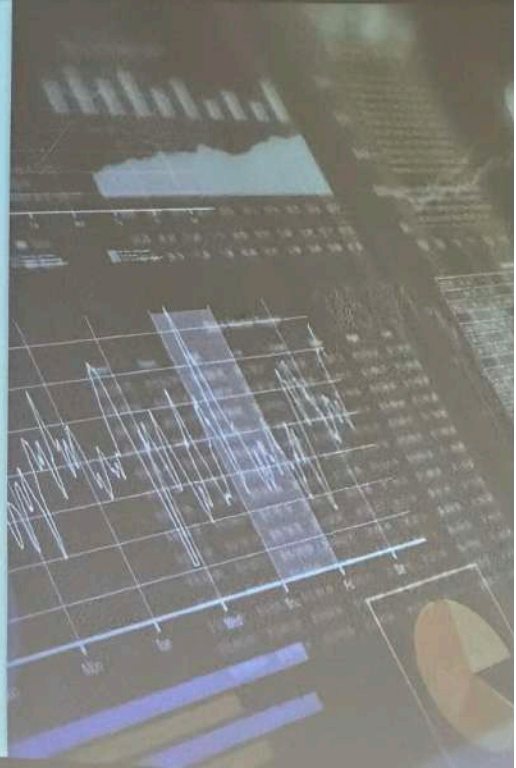
- Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms and systems to extract or extrapolate knowledge from potentially noisy, structured, or unstructured data (wiki)



Data Mining

- Data mining is the process of discovering valuable insights, patterns, and information from vast datasets by employing various techniques, algorithms, and tools

- Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data



- In particular, data mining draws upon ideas, such as
  - sampling, estimation, and hypothesis testing from statistics
  - search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning.

Types of Patterns

Associations

- Coffee buyers usually also purchase sugar

Clustering

- Segments of customers requiring different promotion strategies

Classification

- Bank customer is applying for loan, will the applicant be defaulter?

Association Rules

- antecedent and a consequent
- $\boldsymbol{AR}: X \rightarrow Y$ $\{x_1, x_2, \ldots, x_n\} \rightarrow \{y_1, y_2, \ldots, y_k\}$
- ***itemset*** is the list of all the items in the antecedent and the consequent, i.e. $X \cup Y$

- To evaluate 'association rule' three important measures
  - Support(AR) $= Prob\ (X \cap Y)$

$$\frac{\#\ transactions\ containing\ X \cap Y}{\#transactions\ in\ D}$$

  - Confidence(AR) $= Prob\ (Y|X)$

$$\frac{\#\ transactions\ containing\ X \cap Y}{\#transactions\ in\ X}$$

  - Lift(AR) $= \frac{Prob(Y|X)}{Prob\ (Y)}$

$$\frac{\frac{\#\ transactions\ containing\ X \cap Y}{\#\ transactions\ in\ X}}{\frac{\#transactions\ in\ Y}{\#transactions\ in\ D}}$$

Applications of Association Rules

- Market Basket Analysis
  - People who have bought Sundara Kandam have also bought Srimad Bhagavatham
- Recommendation Systems
- Fraud Detection
- Healthcare and Medical Research
  - Allergy to latex rubber usually co-occurs with allergies to banana and tomato
- Census analysis
  - Immigrants are usually male
- Sports
  - A chess end-game configuration with "white pawn on A7" and "white knight dominating black rook" typically results in a "win for white".

Transaction id	Items
1	Tomato, Potato, Onion
2	Tomato, Potato, Brinjal, Pumpkin
3	Tomato, Potato, Onion, Chilly,
4	Tamarind, Lemmon

AR: Tomato, Potato → Onion

What is itemset here?
Support, Confidence and Lift of the above AR

Apriori Algorithm

- Invented by Rakesh Agrawal and Ramakant Srikant (1994)
- Can we speed up than pure brute force?

- Apriori: acknowledges the prior knowledge
  - If any itemset is not frequent, its superset cannot be frequent
  - An itemset can be frequent only if all its subsets are frequent

How does it work?

- Step 0: create 1-size frequent *itemsets* list that meet threshold support, k=1
- Step 1: Expand the *itemsets* list
  - From the k sized *itemsets* list combine overlapping sets to k+1 size *itemsets* list
- Step 2: Prune the expanded *itemsets* list using apriori property,
  - k=k+1
- Step 3: remove infrequent *itemsets* from the list
- Repeat Step 1,2,3 till no more further expansion possible