Module 1

Introduction to Big Data Platform

Prepared by, Nimmy Francis , AP @ MCA, AJCE

Module- I

Introduction to Big Data Platform–History of Data Management-Structuring Bigdata-Elements of Big Data, Big data Analytics

Introduction to Data Mining: Data Mining Process, Data Mining Techniques - Clustering - Classification - Regression - Outlier detection - Association rules - Prediction, Big Data Vs Data Mining

Introduction

- *A Collection of large and complex datasets which are difficult to store and process using the traditional database and data processing tools* is considered as **big data**.
- *Big data is collected from traditional and digital sources which, when refined properly can be used for research and analysis.*
- With time, organizations are growing and with this data generated from these organizations are also increasing exponentially.

https://www.geeksforgeeks.org/the-big-data-world-big-bigger-and-biggest/?ref=lbp

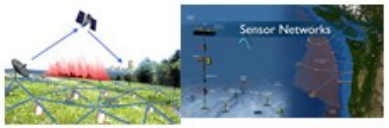
- *The challenge is* to have a platform which can provide a single, consistent view of the complete data.
- *Another challenge* is to organize this data so that it makes sense and can be utilized as useful information.
- *Everything around us generates big data continuously.*
- *Social media websites and digital sources are responsible for producing such huge amount of data.*
- How this huge amount of data is transmitted – sensors, mobile and systems are the answer.

Who's Generating Big Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- *Data is everywhere, in every industry, in the form of numbers, images, videos and text.*
- As data continues to grow, so does the need to organize it.
- The process of capturing or collecting Big data is known as '**datafication**'.
- Big data is 'datafied' so that it can be used productively.
- *To extract meaningful value from Big data*, we need optimal processing power, analytical capabilities and skills.

- ***It becomes difficult for computing systems to manage 'Big data'*** because of the immense speed and volume at which it is generated.
- Traditional data management, warehousing systems are not capable to analyze this type of data.
- Due to its complexity, big data is stored in distributed architecture file system(files will be broken into blocks and stored in nodes over the distributed architecture).
- **Hadoop by Apache** is widely used for storing and managing big data; uses HDFS as the storage system.

Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing. It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

https://www.javatpoint.com/what-is-hadoop

Features of Big data

Big Data

- Is a new data challenge that requires leveraging existing systems differently
- Is classified in terms of 4 Vs
  - Volume
  - Variety
  - Velocity
  - veracity
- Is usually unstructured and qualitative in nature

Leveraging - use (something) to maximum advantage.

Types and sources of data

- Big data is a pool of huge amounts of data of all types, shapes and formats collected from varied sources.
- Some common types of data and their sources:

Type	Description	Source
Social Data	Information collected from various social networking sites and online portals	Facebook, Twitter and LinkedIn
Machine Data	Information generated from RFID chips, bar code scanners, and sensors	RFID chip readings, Global Positioning System (GPS) results
Transactional Data	Information generated from online shopping sites, retailers, and B2B transactions	Retail websites like eBay and Amazon

History of Data Management – Evolution of Big Data

- Big data is the new term of data evolution directed by the enormous velocity, variety, and volume of data.
- **Velocity** implies the speed with which the data flows in an organization; **variety** refers to the varied forms of data, such as structured, semi structured, or unstructured; and **volume** defines the amount or quantity of data an organization has to deal with.
- **Figure 1.2** shows the challenges faced while handling data over the past few decades:

Fig 1.2 : Evolution of Big data

In the early 60s, technology witnessed problems with velocity or real-time data assimilation.

This need inspired the evolution of databases.

In the 90s, technology witnessed issues with **variety** (e-mails, documents, videos) leading to the emergence of non-SQL stores.

Today, technology is facing issues related to huge **volume**, leading to new storage and processing solutions.

Some of the *major milestones in the evolution of big data*:

- **1940s :** An American librarian speculated the potential shortfall of shelves and cataloging staff, realizing the **rapid increase in information and limited storage.**
- **1960s :** Automatic Data Compression was published in the communications of the ACM (Association of Computing Machinery).
  - The paper described '**Automatic Data Compression**' as a complete automatic and fast three-part compressor that can be used for any kind of information in order to reduce the slow external storage requirements and increase the rate of transmission from a computer system.

- **1970s :** In Japan, the Ministry of Posts and Telecommunications initiated a project to study information flow in order to track the volume of information circulating in the country.
- **1980s :** A research project was started by the Hungarian Central Statistics Office to account for the country's information industry. It measured the volume of information in bits.
- **1990s :** Digital storage systems became more economical than paper storage.
- **2000 onwards :** Various methods were introduced to streamline information.

Techniques for controlling the volume, velocity and variety of data emerged.

Structuring Big Data

- ***Structuring big data*** means, arranging the available data in a manner such that it becomes easy to study, analyze and derive conclusion from it.
- **Structuring data helps in understanding** user behaviors, requirements and preferences *to* ***make personalized recommendations*** *for every individual.*
- *When a user regularly visits or purchases from online shopping sites, say eBay, each time he/she logs in, the system can present a recommended list of products that may interest the user on the basis of his/her purchases or searches.*

- Thus presenting a specially customized recommendation set for every user.
- This is the power of *big data analytics*.

Types of data

- Data that comes from multiple sources, such as databases, Enterprise Resource Planning (ERP) systems, weblogs, chat history, and GPS maps, *varies in format*.
- Different formats of data need to be made consistent and clear to be used for analysis.
- **Types of Data Sources**:
  - **Internal sources**, such as *organizational or enterprise data*.
  - **External sources**, such as *social data*.

Table 1.3 :Comparison between internal & external resources of data

Data source	Definition	Examples of sources	Application
Internal	Provides structured or organized data that originates from within the enterprise and helps run business.	Customer Relationship Management (CRM), Enterprise Resource Planning (ERP) systems, customer details, products and sales details	This data is used to support daily business operations of an organization.
External	Provides unstructured or unorganized data that originates from the external environment of an organization.	Business partners, Internet, Government, Market research organizations	This data is often analyzed to understand the entities such as customers, competitors, market and environment.

- Figure 1.3 shows the various parts of Big data.

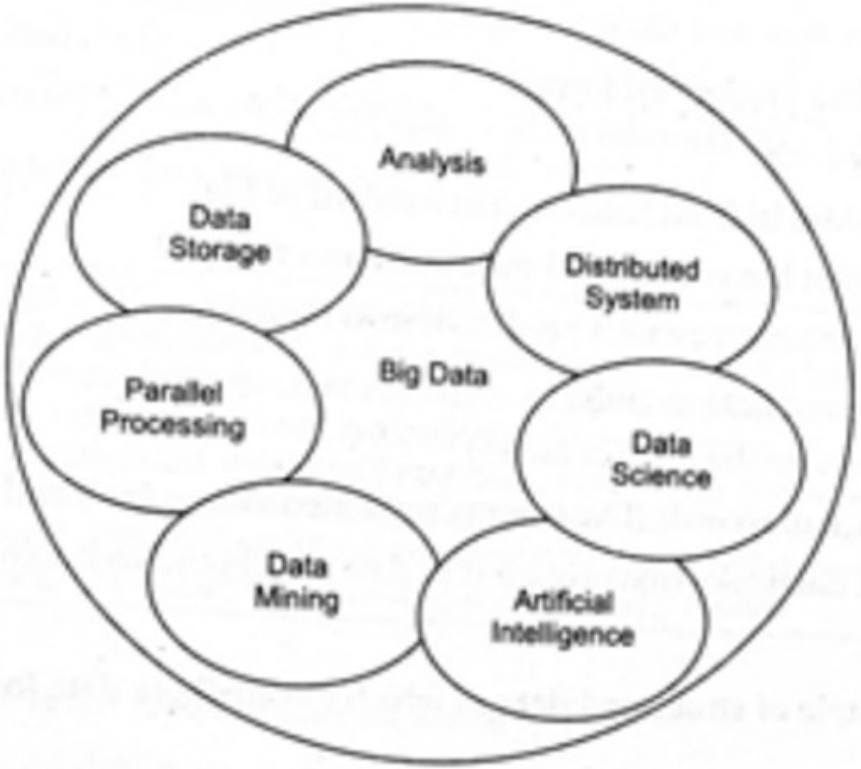


Figure 1.3: Concepts of Big Data.

- **Based on the data received from the sources** mentioned in Table 1.3 ,Big data comprises- the *three types of data*:
  1. *Structured data*
  2. *Unstructured data*
  3. *Semi-structured data*
  - In a real-world scenario, the unstructured data is larger in volume than the structured and semi-structured data, approximately 70% to 80% of data is in unstructured form.

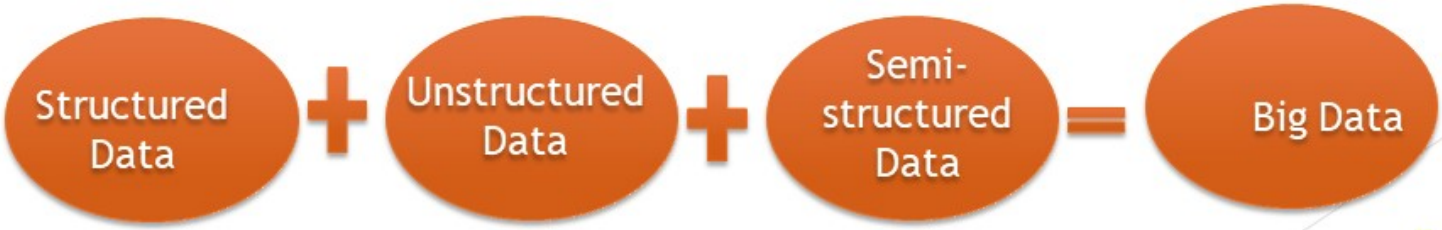


Figure 1.4 : Types of Big data

1. Structured Data

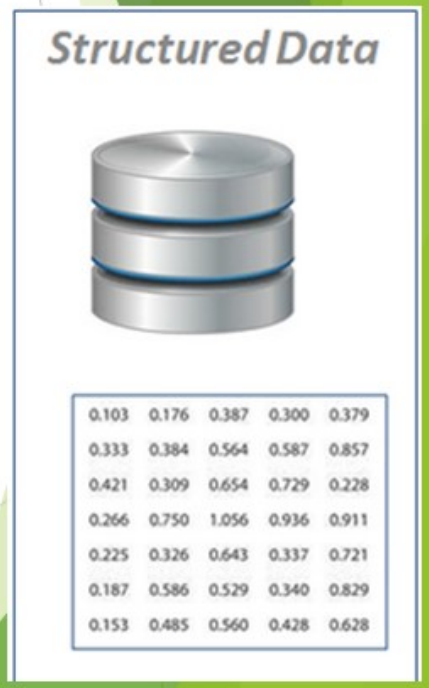
- **Structured Data has a defined pattern**.
- This pattern makes it easier for any program to sort, read and process the data.
- **Structured data**:
  - Is organized data in a predefined format
  - Is stored in tabular form
  - Is the data that resides in fixed fields within a record or file
  - Is formatted data that has entities and their attributes mapped
  - Is used to query and report.
- Processing structured data is much easier and faster than processing data without any specific repeating patterns.

Sources of structured data:

1. Relational databases
2. Flat files in the form of records (A flat file database is a database that stores data in a plain text file. Each line of the text file holds one record, with fields separated by delimiters, such as commas or tabs)
3. Multidimensional databases used in data warehouse technology.
4. Legacy Databases (It can refer to any of the older database technologies. Flat file, hierarchy, and network databases are usually referred as *legacy* databases.

- Table 1.4 shows a *sample of structured data* in which the attribute data for every customer is stored:

Customer ID	Name	Product ID	City	State
22465	Smitha	141	Bangalore	Karnataka
45378	George	177	Kochi	Kerala
87231	John	445	Chennai	Hyderabad



https://www.youtube.com/watch?v=44BuY-xtzVE&t=1194s

2. Unstructured Data

- Unstructured data is a *set of data that might or might not have any logical or repeating patterns.*
- **Unstructured Data**:
  - Consists of metadata
  - Comprises inconsistent data, such as data obtained from files, social media websites, satellites etc.
  - Consists of data in different formats such as e-mail, text, audio, video or images.

Sources of Unstructured Data:

- Internal and external to an organization: documents, logs, survey results, feedback and e-mails from both within and across the organization.
- Social media : data obtained from social networking platforms, including YouTube, Facebook, Twitter, LinkedIn and so on..
- Mobile data : data such as text messages and location information.

Challenges associated with unstructured data:

- Identifying the unstructured data that can be processed.
- Sorting, organizing and arranging unstructured data in different sets and formats.
- Combining and linking unstructured data in a more structured format to derive any logical conclusions.
- Costing in terms of storage and human resource needed to deal with exponential growth of unstructured data.

3. Semi-Structured Data

- Semi-structured data, also known as having a schema-less or self describing structure, refers to a *form of structured data that contains tags or markup elements in order to separate elements and generate hierarchies of records and fields in the given data.*
- Such type of data does not follow the proper structure of data models as in relational databases.
- In other words, data is stored inconsistently in rows and columns of a database.

Sources:

- Files systems such as web data in the form of cookies
- Data exchange formats such as JavaScript Object Notation (JSON) data.

Table 1.5 : Semi structured data

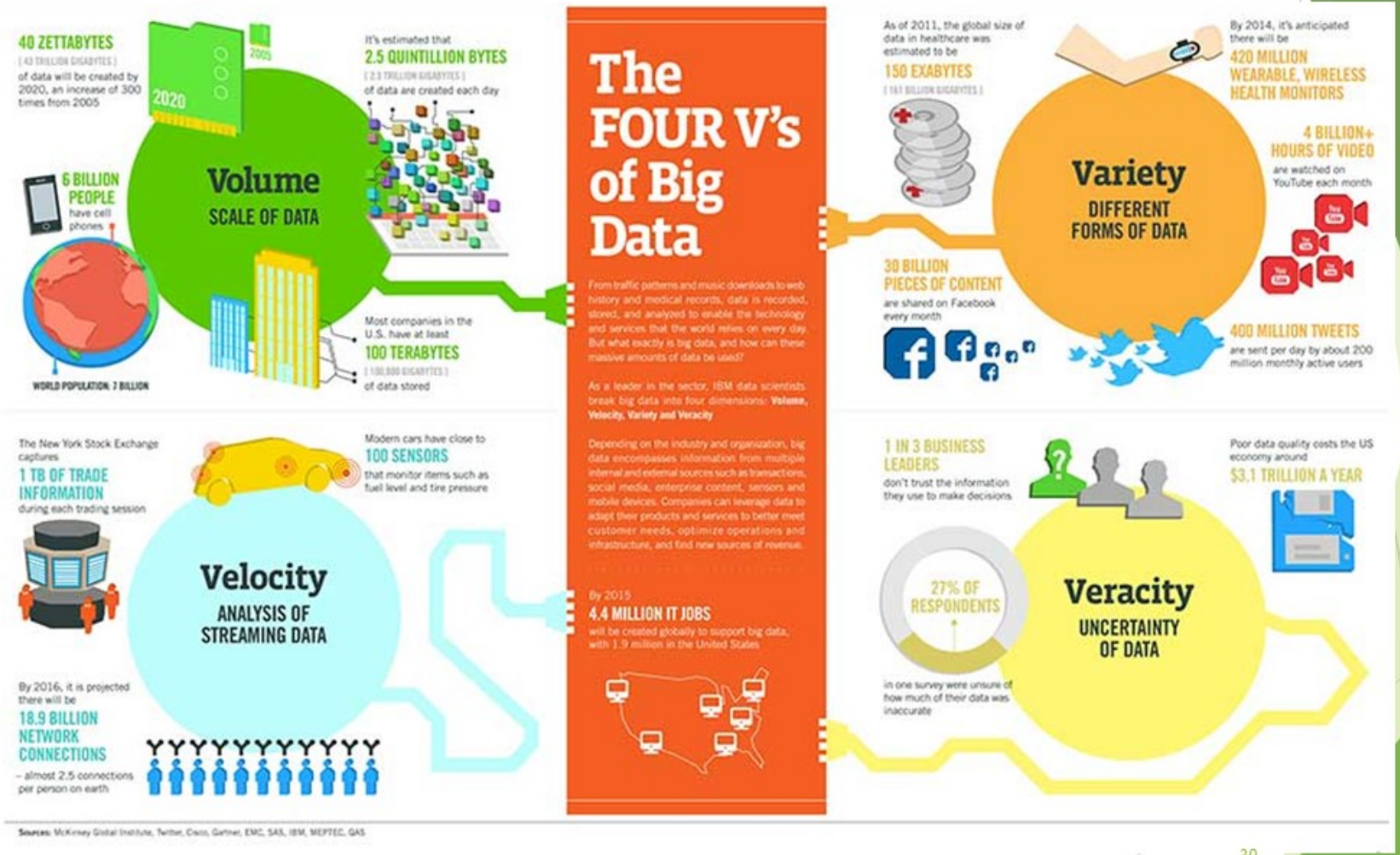
SL.NO	NAME	E-MAIL
1	Sam Jacobs	smj@xyz.com
2	First name: David Last name : Lee	davidlee@xyz.com

Elements of Big Data

- The growth of data can be represented in terms of **four Vs**:

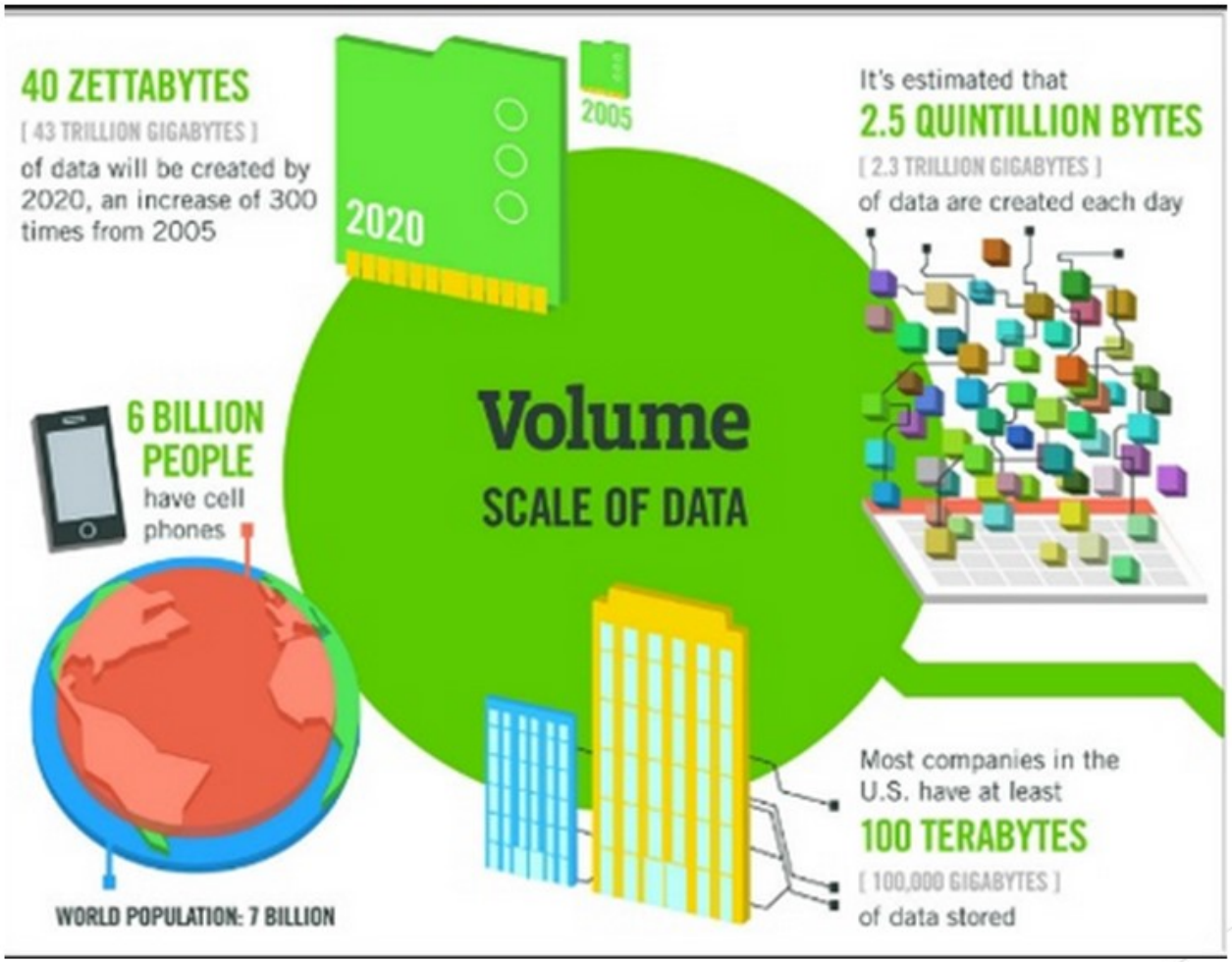
- *Volume*
- *Velocity*
- *Variety*
- *Veracity*

1. **Volume**: Amount of quantity of data an organization has to deal with.
2. **Velocity:** The speed which the data flows in an organization.
3. **Variety:** Varied forms of data, such as structured, semi-structured, or unstructured.
4. **Veracity**: Certainty of data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, GAS

1. Volume

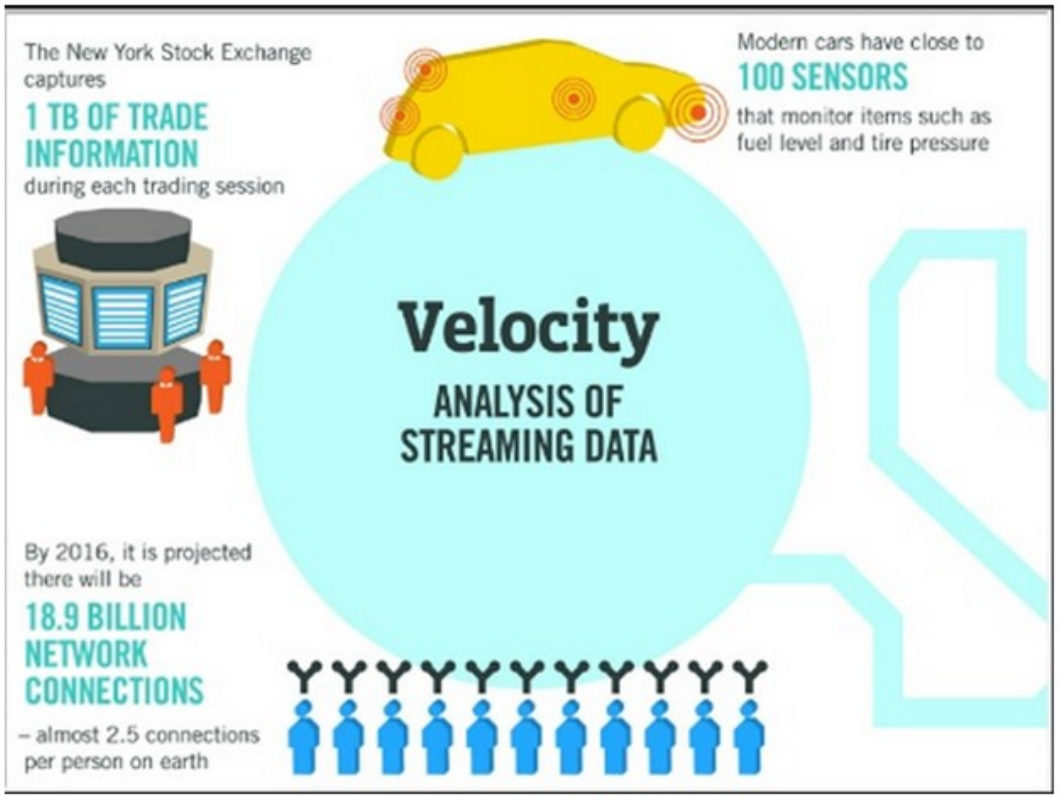
- *Volume is the amount of data generated by organizations or individuals.*
- Today, the volume of data in most organizations is approaching Exabytes.
- Every minute, over 571 new websites are being created.
- It is estimated that by 2020, online business transactions will reach up to 450 billion per day.
- *Internet alone generates a huge amount of data.*
- *Internet has around 14.3 trillion live web pages.*
- *Internet has around 672 exabytes of accessible data.*
- *Total data stored on the internet is over 1 yottabyte.*


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

2. Velocity

- *Velocity refers to the speed at which new data is generated and the speed at which data moves around.*
- In other words, the rate at which data is generated, captured and shared in real time.
- The *sources of high velocity data include* :
- IT devices, including routers,switches,firewalls etc. constantly generate valuable data.
- Social media including facebook posts, tweets and other social media activities , create huge amount of data, which is to be analysed instantly at a fast speed because the value degrades quickly with time.

- Portable device, including mobile,PDA etc. also generate data at a high speed.



3.Variety

- *Variety refers to the different types of data.*
  - Structured data such as financial data Vs unstructured (80% of the world's data - text, images, video, voice, etc.)
- With big data technology ,we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

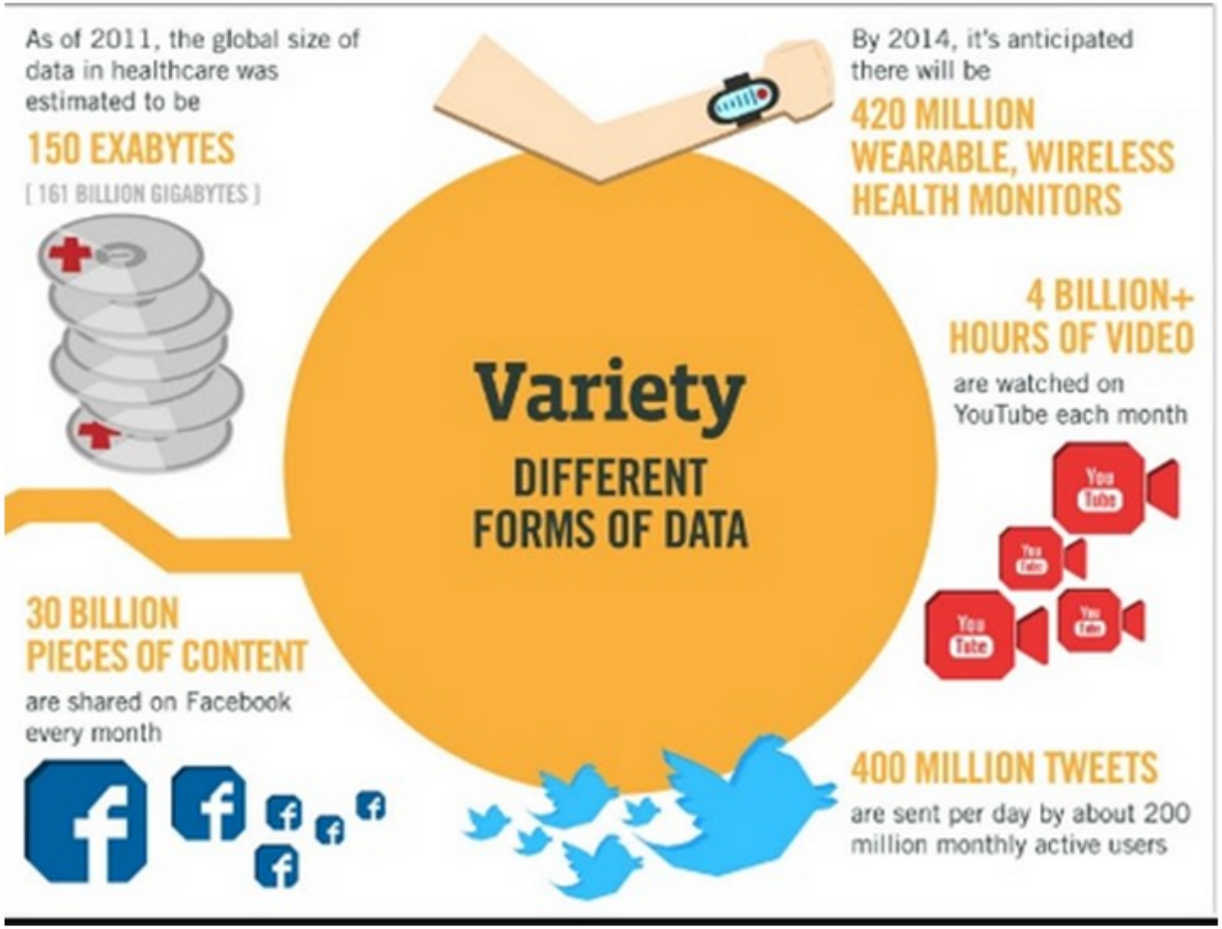
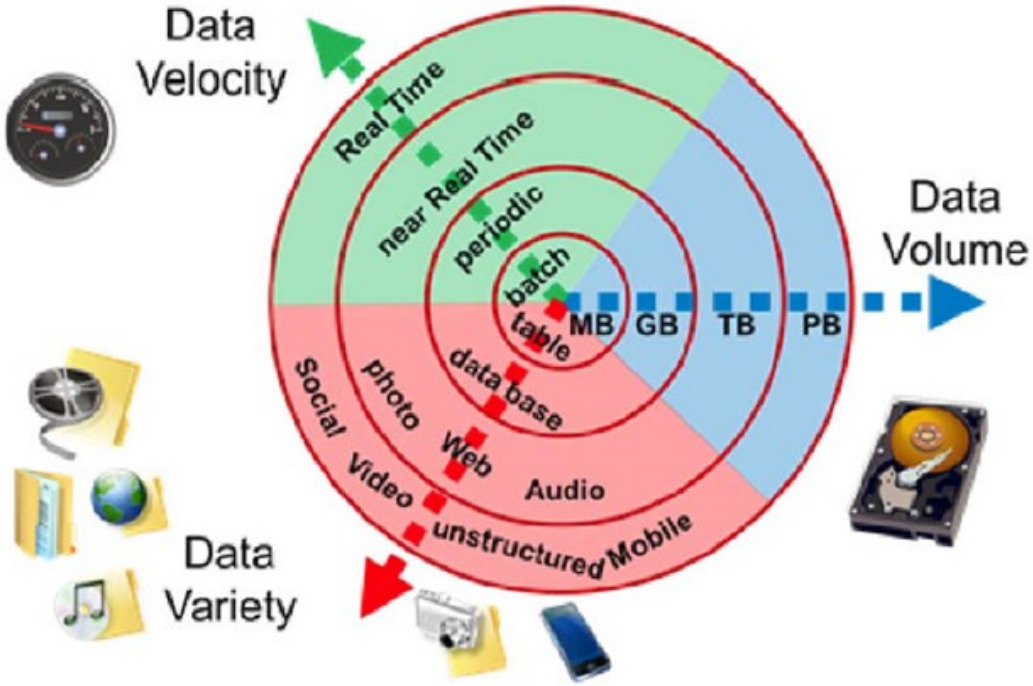
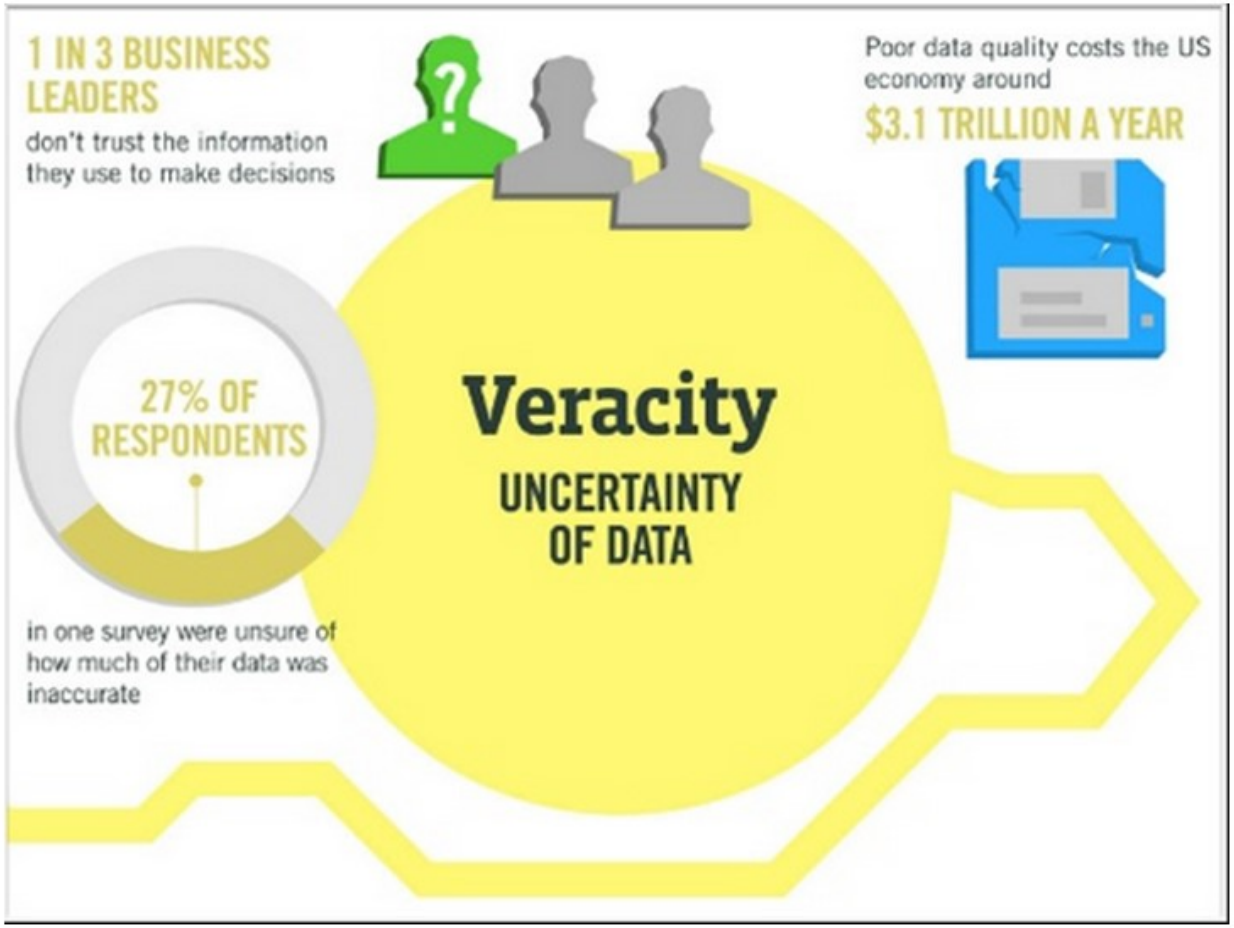

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Figure 1.7 : Various types of data



4. Veracity

- *Veracity refers to the uncertainty or trustworthiness of the data; i.e., whether the obtained data is correct or consistent.*
- Out of the huge amount of data that is generated in almost every process, only the data that is correct and consistent can be used for further analysis.
- Data when processed becomes information; however, a lot of effort goes in processing the data.
- Big data especially in the unstructured and semi structured forms, is ,messy in nature, and it takes a good amount of time and expertise to clean that data and make it suitable for analysis.


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Big Data Analytics

- Big data analytics *helps to conduct business in many ways*, such as it improves decision making, business process management etc.
- Business analytics **uses the data** and **different techniques** like information technology, features of statistics, quantitative methods and **different models** to provide results.
- There are ***three main types of business analytics*** : descriptive analytics, predictive analytics, and prescriptive analytics.

1. Descriptive Analytics

- It is the most widespread form of analytics and it serves as a base for advanced analytics.
- It answers the question **"What happened in the business".**
- It analyzes a database to provide information on the trends of past or current business events that can help managers, planners, leaders etc. to develop a road map for future actions.
- ***It performs an in-depth analysis of data to reveal details such as frequency of events, operation costs, and the underlying reason for failures.***
- It helps in identifying the root cause of the problem.

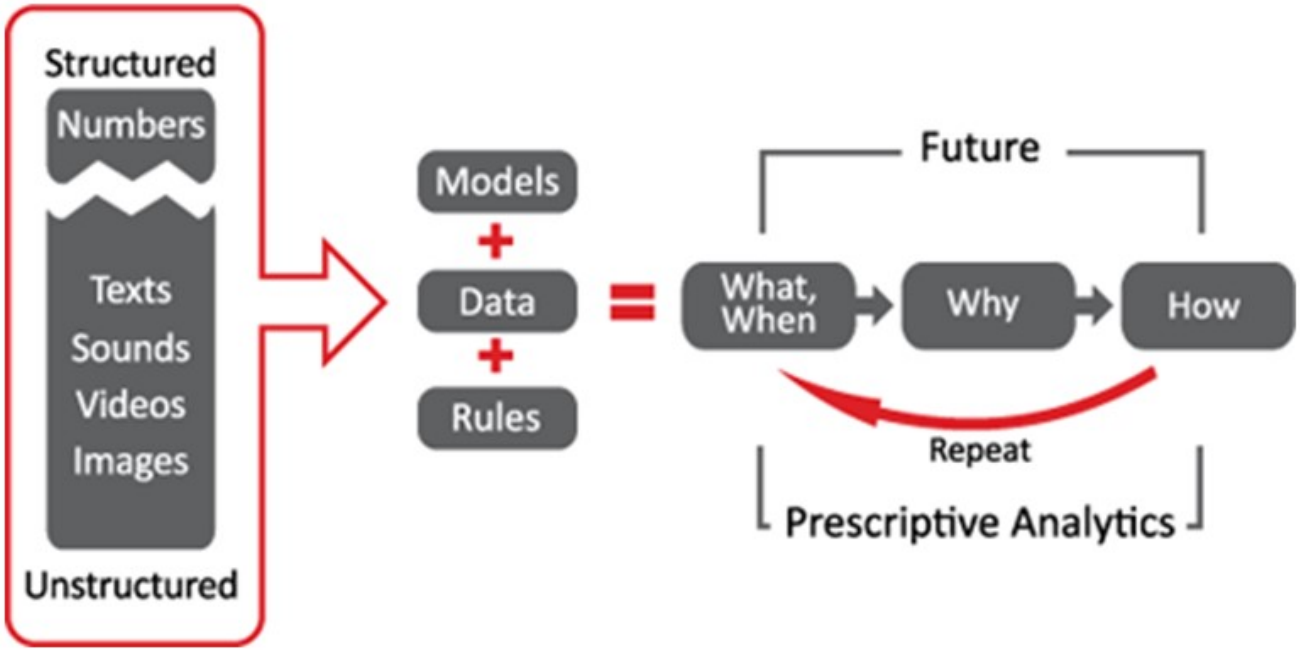
2. Predictive analytics

- Predictive analytics is about understanding and predicting the future and answers the question **"what could happen"** by using statistical models and different forecast techniques.
- It predicts the near future probabilities and trends and helps in what-if analysis.
- (**What-If Analysis** is the process of changing the values in cells to see how those changes will affect the outcome).
- In predictive analytics, we use different statistics, data mining techniques and machine learning to analyze the future.

3.Prescriptive analysis

- Prescriptive analysis answers "**what should we do**" **on the basis of complex data** ***obtained from descriptive and predictive analysis.***
- By using the optimization technique, prescriptive analytics determines the finest substitute to minimize or maximize some equitable finance, marketing and many other areas.
- **For example**, if we have to find the best way of shipping goods from a factory to a destination, to minimize costs, we will use the prescriptive analytics.

- Figure 1.9 shows a diagrammatic representation of the **stages involved in the prescriptive analytics** :



Advantages of Big data Analytics

- The right analysis of the available data can improve major business processes in various ways.
- **For example**, in a manufacturing unit, data analytics can improve *the functioning of the following processes*.
- Procurement**:** To find out which suppliers are more efficient and cost-effective in delivering products on time.
- Product Development**:** To draw insights or innovative product and service formats and designs for enhancing the development process and coming up with demanded products.

- Manufacturing: To identify machinery and process variations that may be indicators of quality problems.
- Distribution: To enhance supply chain activities.
- Marketing: To identify which marketing campaigns will be the most effective in driving and engaging customers and understanding customer behaviors and channel behaviors.
- Human resources: To find out the characteristics and behaviors of successful and effective employees, as well as other employee insights for managing talent better.

- **Price management:** To optimize prices based on the analysis of external factors.
- **Merchandising:** To improve merchandise breakdown on the basis of current buying patterns and increase inventory levels and product interest insights on the basis of the analysis of various customer behaviors.
- **Sales:** To optimize assignment of sales resources and accounts, product mix, and other operations.
- **Storage operations:** To adjust inventory levels on the basis of predicted buying patterns, study of demographics, weather, key events and other factors.

Merchandising-the activity of promoting the sale of goods

Transportation : Big data has greatly improved transportation services.

- *The data containing traffic information is analyzed to identify traffic jam areas*.
- Distributed sensors are installed in handheld devices, on the roads and on vehicles to provide real-time information.
- *This data is analyzed and passed to the traffic control authority*.

Education : e-learning *helps to analyze the students' ability and thus impart education effectively in accordance with each student's needs.*

- The analysis can be done by studying the responses to questions, recording the time consumed in attempting those questions, and analyzing other behavioral signals of the students.

Travel : Maintains complete details of all customer records that are then analyzed to determine certain behavior patterns in customers.

- **For example**, *in the airline industry, big data is analyzed for identifying personal preferences or spotting which passengers like to have window seats*.
- Big data track customers who regularly fly between specific routes to understand their behavior, what kind of items are preferred.

Healthcare: *The pharmacy and medical device companies use big data to improve their research and development practices.*

- Big data helps doctors to analyze the requirement and medical history of every patient and provide individualistic services to them, depending on their medical condition.

Telecom : The mobile revolution and the internet usage on mobile phones have led to a tremendous increase in the amount of data generated in the telecom sector.

- Big data analytics *allows telecom industries to utilize this data for extracting meaningful information that could be used to understand their performance, improving customer service, and generating more business opportunities*.

Consumer Goods Industry: Consumer goods companies generate huge volume of data in varied formats from different sources, such as transactions, billing details, feedback forms etc.

- **For example** , *the data generated from the Point-of-Sale (POS) systems provides significant real-time information about customer's preferences, current market trends, the increase and decrease in demand of different products at different regions* etc.

- **Aviation Industry** : Big data also plays a significant role in the commercial aviation industry.
- This industry also maintains a detailed record of all their customers that includes their personal information, flying preferences and other trends and patterns.
- *The organization analyses this data to improve their customer services and thus brand image.*
- In addition, every aircraft generates a significant amount of data during operation.
- *This data is then analysed for enhancing operational efficiencies., identifying the parts that require repairs and taking any necessary constructive or preventive measures on time.*

Introduction to Data Mining

- **Data mining** is one of the most useful techniques that help entrepreneurs, researchers, and individuals ***to extract valuable information from huge sets of data***.
- Data mining is also called ***Knowledge Discovery in Database (KDD)***.
- The **knowledge discovery process** includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

https://www.javatpoint.com/data-mining

What is Data Mining?

- The *process of extracting information* to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called **Data Mining**.

https://www.youtube.com/watch?v=q7AfJDI2Onk

- Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures.
- Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events.
- Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems.
- It primarily turns **raw data into useful information.**

Benefits of data mining

- The business benefits of data mining come from the increased ability to uncover hidden patterns, trends, correlations and anomalies in data sets.
- *That information can be used to improve business decision-making and strategic planning through a combination of conventional data analysis and predictive analytics.*

Specific data mining benefits include the following:

- **More effective marketing and sales :** Data mining helps marketers better understand customer behavior and preferences, which enables them to create targeted marketing and advertising campaigns.

https://www.techtarget.com/searchbusinessanalytics/definition/data-mining#:~:text=Data%20mining%20is%20the%20process,make%20more%2Dinformed%20business%20decisions.

- Similarly, sales teams can use data mining results to improve lead conversion rates and sell additional products and services to existing customers.
- **Better customer service** :Companies can identify potential customer service issues more promptly and give contact center agents up-to-date information to use in calls and online chats with customers.
- **Improved supply chain management :** Organizations can spot market trends and forecast product demand more accurately, enabling them to better manage inventories of goods and supplies.
- Supply chain managers can also use information from data mining to optimize warehousing, distribution and other logistics operations.

- **Increased production uptime :** Mining operational data from sensors on manufacturing machines and other industrial equipment supports predictive maintenance applications to identify potential problems before they occur, helping to avoid unscheduled downtime.
- **Stronger risk management :** Risk managers and business executives can better assess financial, legal, cybersecurity and other risks to a company and develop plans for managing them.
- **Lower costs :** Data mining helps drive cost savings through operational efficiencies in business processes and reduced redundancy and waste in corporate spending.

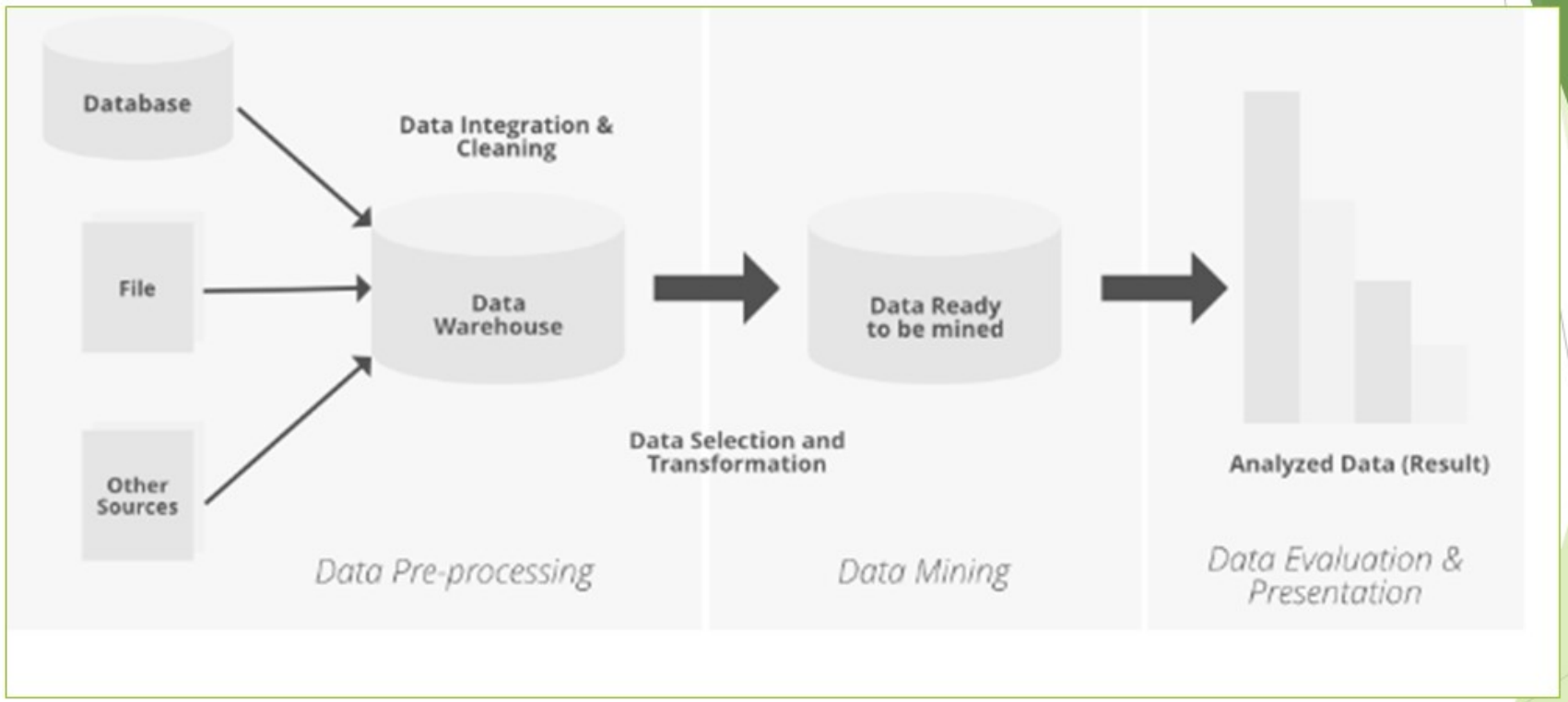
Data Mining Process

KNOWLEDGE DISCOVERY (KDD) PROCESS

Here is the list of steps involved in the knowledge discovery process:

- **Data Cleaning** – In this step, the noise and inconsistent data is removed.
- **Data Integration** – In this step, multiple data sources are combined.
- **Data Selection** – In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** – In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** – In this step, intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** – In this step, data patterns are evaluated.
- **Knowledge Presentation** – In this step, knowledge is represented.

https://www.youtube.com/watch?v=1NjPTh0Eoeg


Database
File
Other Sources
Data Integration & Cleaning
Data Warehouse
Data Selection and Transformation
Data Ready to be mined
Analyzed Data (Result)
Data Pre-processing
Data Mining
Data Evaluation & Presentation

1.Data Cleaning
2. Data Integration

1. Data Cleaning:

- Removing noisy data
- Fill the missing values
- Remove inconsistent data

• Remove Noise and Inconsistent Data

2. Data Integration:

- The process of combining data from multiple heterogeneous data sources is known as Data integration

• Where multiple data sources may be combined

https://www.youtube.com/watch?v=9bPvfd63k2M

3. Data Selection
4. Data Transformation

- Where data relevant to the analysis task are retrieved from the data base

3. Data Selection:

- **The process of selecting task relevant data.**

4. Data Transformation:

- **Data need to be transferred int appropriate form for variou applications**

- Where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operation

5. Data Mining
6. Pattern Evaluation

5. Data Mining :

- The functionalities such as characterization, classification, regression, association, clustering, outlier analysis applied to mine data.

An essential Process where intelligent methods are applied to extract data patterns

6. Pattern evaluation:

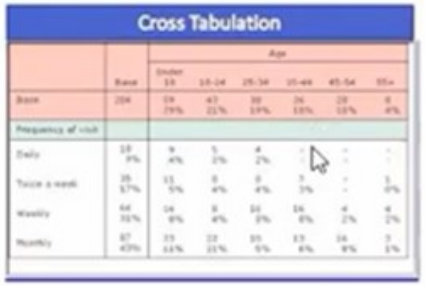
- All the discovered data may not be interested or we may discovered huge amount data as interested. Thus interesting measures such as support, confidence, lift etc used for finding only interested patterns.

- To identify the truly interesting patterns representing knowledge based on interestingness measures

7.Knowledge representation

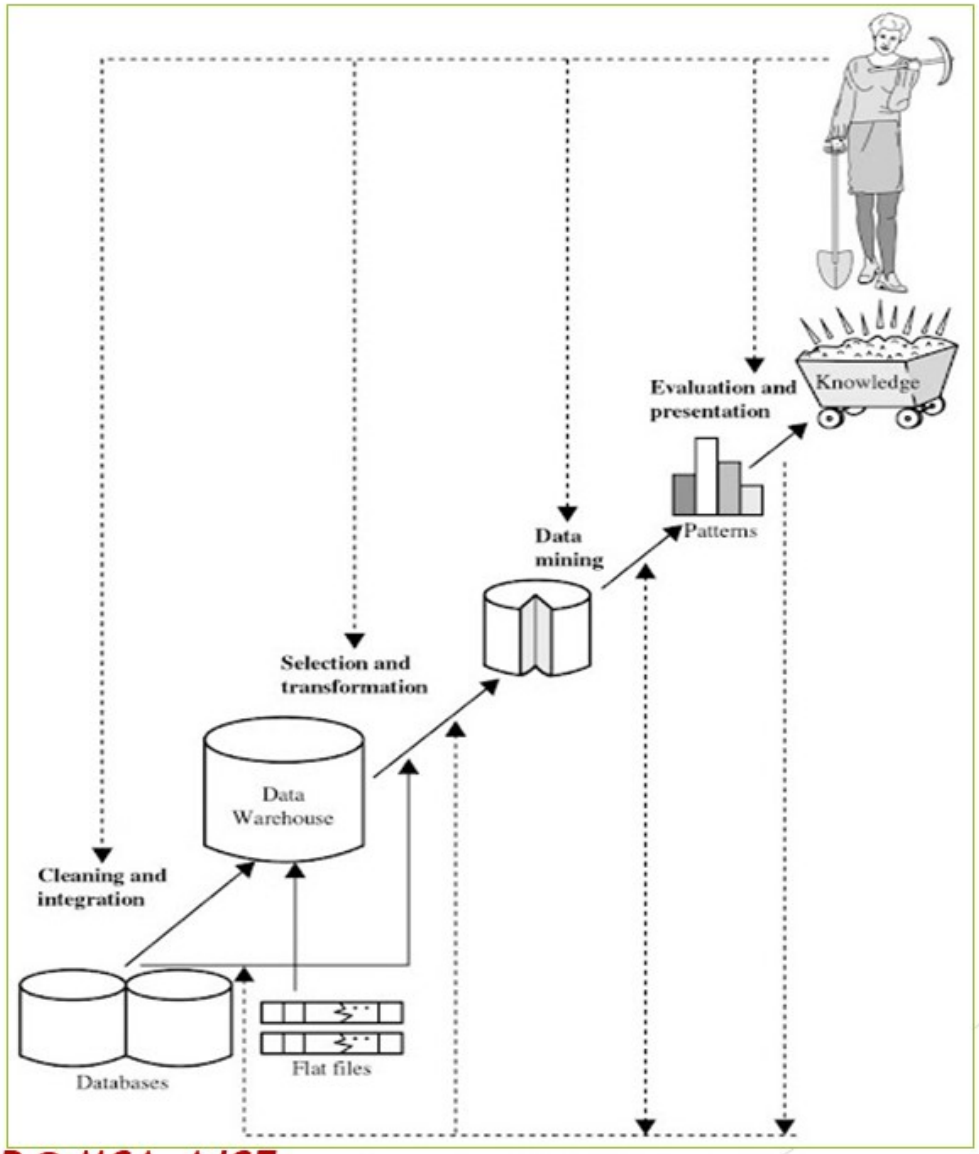
7. knowledge presentation:

- Discovered knowledge presented visualization of discovered knowledge in the form of pie charts, bar chats, curves, plots, tables, cross tabs, trees etc.



- IF-THEN
- Independent part, combined with other pieces, to produce better result
- Model of human behavior
- Examples
  - IF condition, THEN conclusion
  - Conclusion, IF condition
  - If condition, THEN conclusion1 (OR) ELSE conclusion2

KDD process

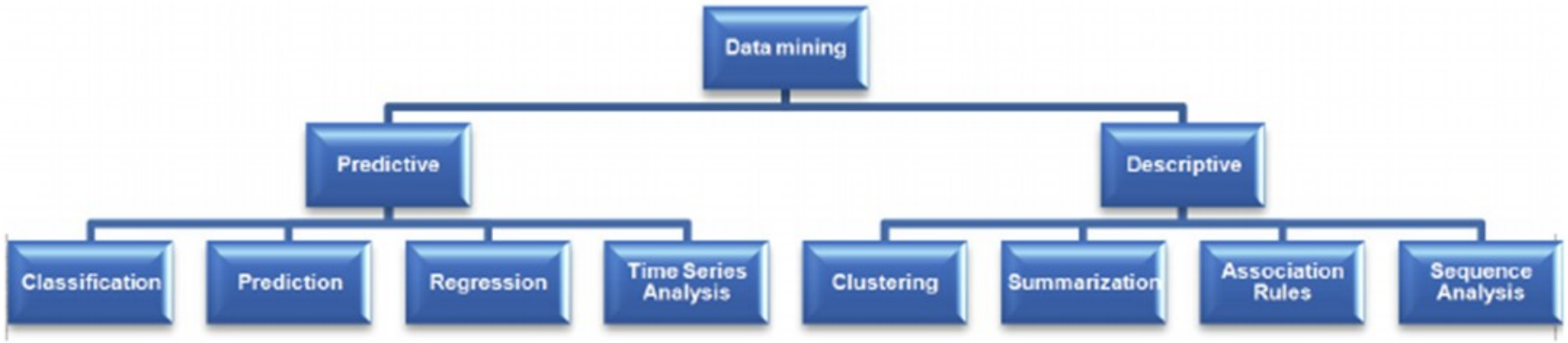

Evaluation and presentation
Knowledge
Patterns
Data mining
Selection and transformation
Data Warehouse
Cleaning and integration
Databases
Flat files

Data Mining Techniques

- **To extract information from data**, a wide variety of **data mining techniques** are employed.

These include:

1. Classification
2. Clustering
3. Association rule
4. Regression
5. Anomaly(outlier) detection
6. Prediction



https://www.analyticsvidhya.com/blog/2021/05/introduction-to-data-mining-and-its-applications/

1.Clustering

- **Clustering** is a *technique for grouping related entries in a database into clusters based on their similarities.*
- Whereas classification assigns variables into known categories, **the clustering technique first singles out these clusters in the dataset and then groups variables based on their characteristics.**
- **For example**, you can cluster customers into groups according to the sales data – *those who regularly buy pet food* **or** *specific drinks* **and** *who are stable in their preferences* and *customer behavior.*
- Once you establish these clusters, you can easily target them with customized advertisements.

https://serokell.io/blog/guide-to-data-mining-techniques

2. Classification

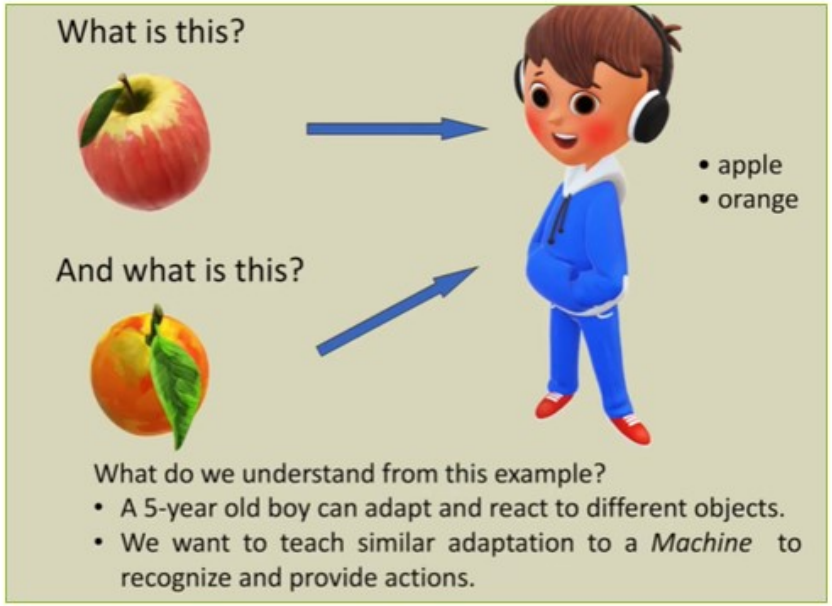
- Classification is used to divide data **into predetermined groups or classes.**
- *This data mining technique determines* the class to which a record belongs based on the values of several attributes.
- The goal is to sort data into predefined classes.

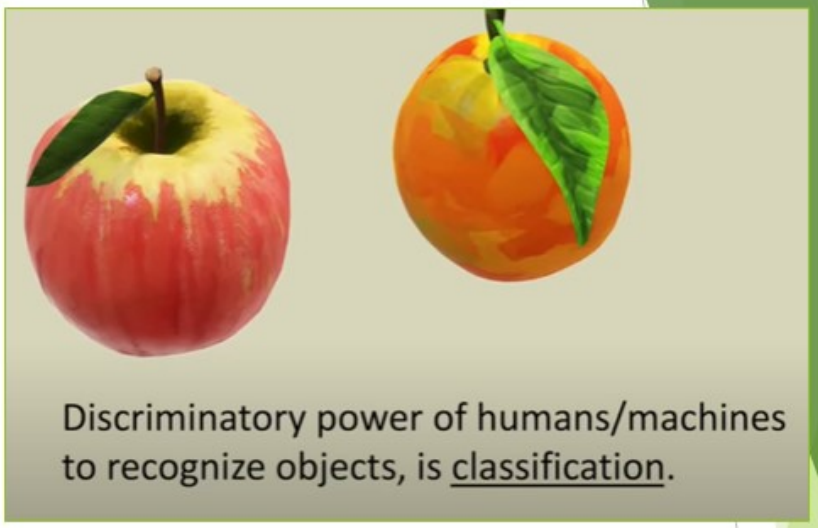
Clustering has a wide range of applications:

- medical diagnostics
- computational biology
- text mining
- web analytics
- **Another example of clustering**, there are two clusters named as mammal and reptile. A mammal cluster includes human, leopards, elephant, etc.
- On the other hand, reptile cluster includes snakes, lizard, komodo dragon etc.
- **The tools mainly used in cluster analysis** are k-mean, k-medoids , density based, hierarchical and several other methods.

- **For example**, in a banking application, the customer who applies for a loan may be classified as a safe and risky according to his/her age and salary.
- This type of activity is also called **supervised learning**.
- The constructed model can be used to classify new data.
- The **learning step** can be accomplished by using already defined training set of data.
- Each record in the training data is associated with an attribute referred to as a class label, that signifies which class the record belongs to.
- Some **common classification algorithms** are *decision tree, neural networks, logistic regression*, etc.

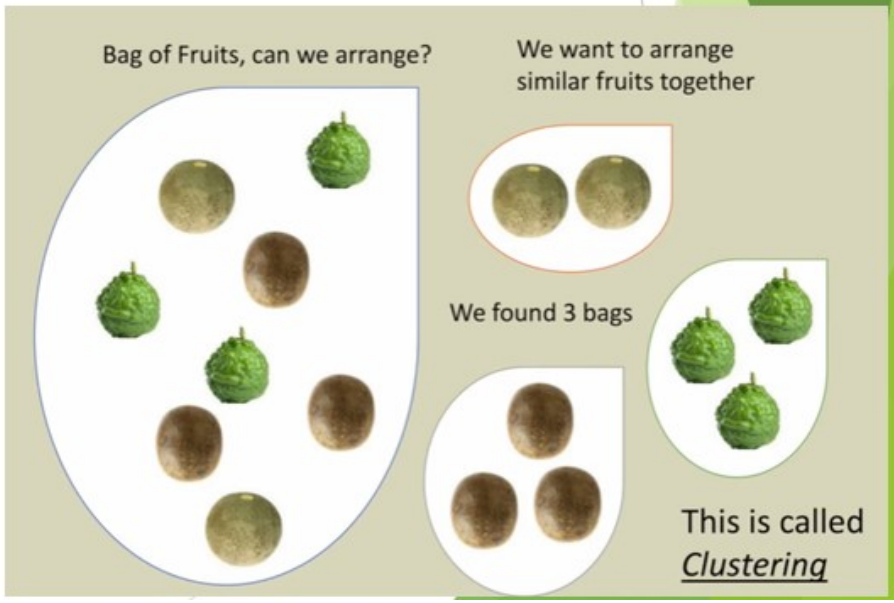
https://techdifferences.com/difference-between-classification-and-clustering.html





// extra.. Slide 72 - Slide 75

https://www.youtube.com/watch?v=LaqTK52rgks



Key Differences Between Classification and Clustering

1. Classification is the process of classifying the data with the help of class labels. On the other hand, Clustering is similar to classification but there are no predefined class labels.
2. Classification is geared with supervised learning. As against, clustering is also known as unsupervised learning.
3. Training sample is provided in classification method while in case of clustering training data is not provided.

https://techdifferences.com/difference-between-classification-and-clustering.html

- **Classification and clustering are the methods used in data mining** ***for analyzing the data sets and divide them on the basis of some particular classification rules or the association between objects.***
- **Classification categorizes the data** *with the help of provided training data.*
- On the other hand, **clustering uses** *different similarity measures to categorize the data.*

https://techdifferences.com/difference-between-classification-and-clustering.html

Examples

Netflix

A well-known application of clustering algorithms are Netflix recommendation systems. Although the company is quite discreet with its algorithms, it is confirmed that there are about 2,000 clusters or communities that have common audiovisual tastes. Cluster 290 is the one that includes people who like the series "Lost", "Black Mirror" and "Groundhog Day". Netflix uses these clusters to refine its knowledge of the tastes of viewers and thus **make better decisions in the creation of new original series**.

Fraud Detection

Classification is commonly used in the financial sector to guarantee data security. In the era of online transactions where the use of cash has decreased markedly, it is necessary to determine whether movements made through cards are safe. Entities can classify transactions as correct or fraudulent using historical data on customer behavior to **detect fraud very accurately**.

https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation

3. Regression

- Regression analysis is the data mining process, that is used to identify and analyze the relationship between variables because of the presence of the other factor.
- It is used to define the probability of the specific variable.
- Regression, primarily a form of planning and modeling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition.
- Primarily it gives the exact relationship between two or more variables in the given data set.

https://www.youtube.com/watch?v=Y8-85kblCXg&t=133s

Some more examples are:

Example 3:

Based on the expected number of customers and the previous days' data given below, predict the number of burgers that will be sold by a KFC outlet .

No. of Customers	Burgers Sold
798	444
450	324
5067	1054

Example 4:

Calculate the life expectancy for a group of people with the average length of schooling based on the data given below:

Average Length of Schooling (in years)	Life Expectancy
14	63
20	80.4
10	57.1

https://www.youtube.com/watch?v=DtOYBxi4AIE

- One example is forecasting customers' age based on their purchase history.
- We can also predict costs based on such variables as consumer demand – for example, a surge of prices on the secondary market due to the increased demand for cars in the US.

*4. Anomaly detection(*Outlier Analysis)

- **Anomaly detection** is a *data mining technique* **used to identify outliers** *(values that deviate from the norm).*
- **For example**, in e-commerce datasets, *it can detect unusual sales during a given week at a store location*.
- Among other things, it can be used to discover credit or debit fraud and identify intrusion or interruption in the network.

https://www.youtube.com/watch?v=aSwQjBS-Osg&list=PLmAmHQ-_5ySxFoIGmY1MJao-XYvYGxxgj&index=28

5. Association rule

- **Association rule** learning **discovers** **if-then patterns** between two or more variables.
- The simplest example is the association between buying bread and butter.
- *People that buy bread usually get butter with it, and vice versa.*
- *That is why you will find these two products close to one another in a grocery store.*
- However, the link may be not that direct.

- While the concepts behind association rules can be traced back earlier, association rule mining was defined in the 1990s, when computer scientists Rakesh Agrawal, Tomasz Imieliński and Arun Swami developed an algorithm-based way to find relationships between items using point-of-sale (POS) systems.
- Applying the algorithms to supermarkets, the scientists were able to discover links between different items purchased, called *association rules*, and ultimately use that information to predict the likelihood of different products being purchased together.
- *For retailers, association rule mining offered a way to better understand customer purchase behaviors.*
- Because of its retail origins, association rule mining is often referred to as *market basket analysis*.

https://www.techtarget.com/searchbusinessanalytics/definition/association-rules-in-data-mining#:~:text=Association%20rule%20mining%2C%20at%20a,and%20a%20consequent%20(then).

6. Prediction

- Another process of data analysis is prediction.
- It is used to find a numerical output.

https://www.javatpoint.com/classification-and-predication-in-data-mining

- Prediction used a combination of other data mining techniques such as clustering, classification, etc.
- It analyzes past events or instances in the right sequence to predict a future event.

https://www.youtube.com/watch?v=Y8-85kblCXg&t=133s

- Predicting the value of a house depending on the facts such as the number of rooms, the total area, etc., is an example for prediction.
- *For example*, suppose the marketing manager needs to predict how much a particular customer will spend at his company during a sale.

Big Data Vs Data Mining

Data Mining	Big Data
It is one of the method in the pipeline of Big Data.	Big Data is a technique to collect, maintain and process the huge information. It explains the data relationship.
Data mining is a part of Knowledge Discovery of the Data. It is close view of the data.	It is about extracting the vital and valuable information from huge amount of the data.It is a technique of tracking and discovering of trends of complex data sets. It is a large or overall view of the data.
The goal is same as Big Data as it is one of the tool of Big Data.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.

https://www.geeksforgeeks.org/difference-between-big-data-and-data-mining/

It is manual as well as automated in nature.	It is only automated as computing huge data is difficult.
It only focuses on only one form of data. i.e. structured.	It focuses and works with all form of data i.e. structured, unstructured or semi-structured.
It is used to create certain business insights. Data mining is a manager of the mine.	It is mainly used for business purposes and customer satisfaction. Big Data is a mine.
It is a sub set of Big Data. i.e. one of the tools.	It is a super set of Data Mining.
It is a tool to dig up the vital information from the large data. Data can be large as well as small.	It is more involved with the processes of handling voluminous data. Data can only be large.