UNIT-I
Introduction to Data Science and Big Data

Dr. Vijay A. Kotkar
Assistant Professor

Department of Computer Engineering
Pimpri Chinchwad College of Engineering & Research,
Ravet, Pune

Basics and Need of Data Science

- **Data science:** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.

- Data science encompasses preparing data for analysis, including cleaning, aggregating, and manipulating the data to perform advanced data analysis.

- Examples of Data Science - Identification and prediction of disease, Optimizing shipping and logistics routes in real-time, detection of frauds, healthcare recommendations, automating digital ads etc.

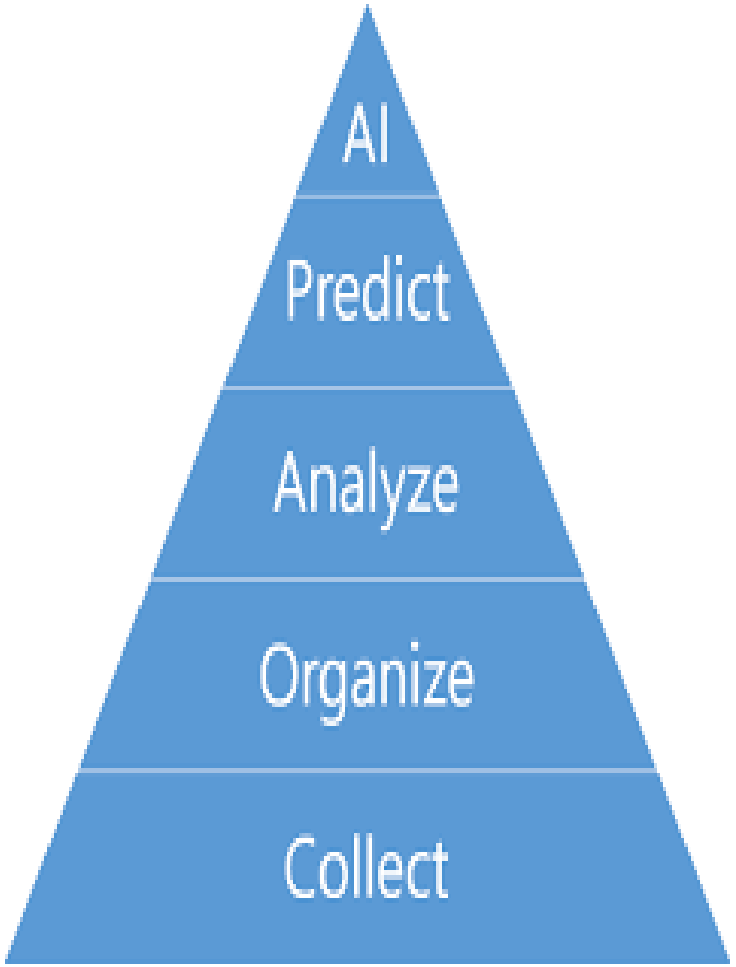
Basics and Need of Data Science

Need:

1) Data Science is the ability to process and interpret data.

2) This enables companies to make informed decisions around growth, optimization, and performance.

3) Demand for skilled data scientists is on the rise now and in the next decade.



For example -

- Machine Learning is now being used to make sense of every kind of data – big or small.

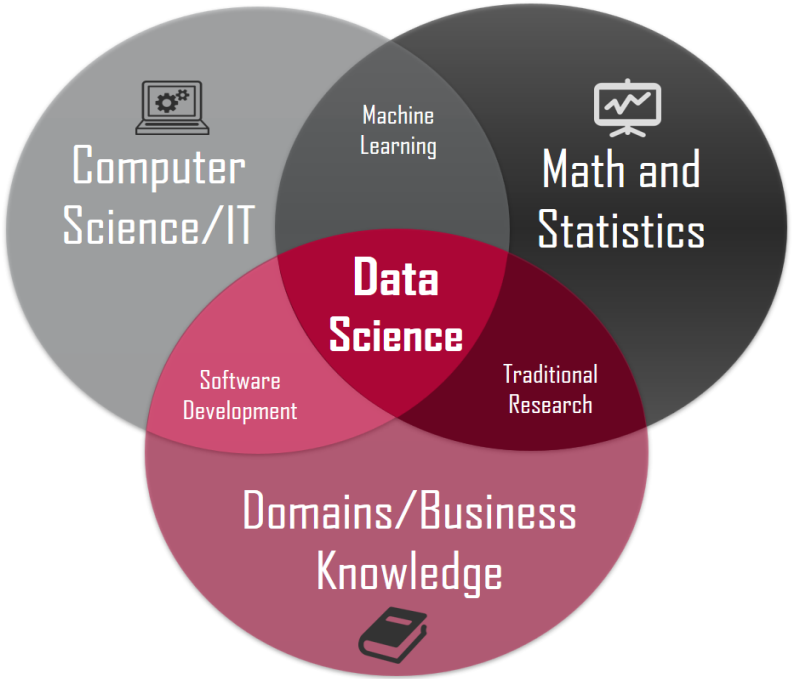
- Data metrics are driving every business decision.

- Companies are busy ramping up their data science workforce to enable higher efficiency and planning.

Basics and Need of Data Science

Purpose:

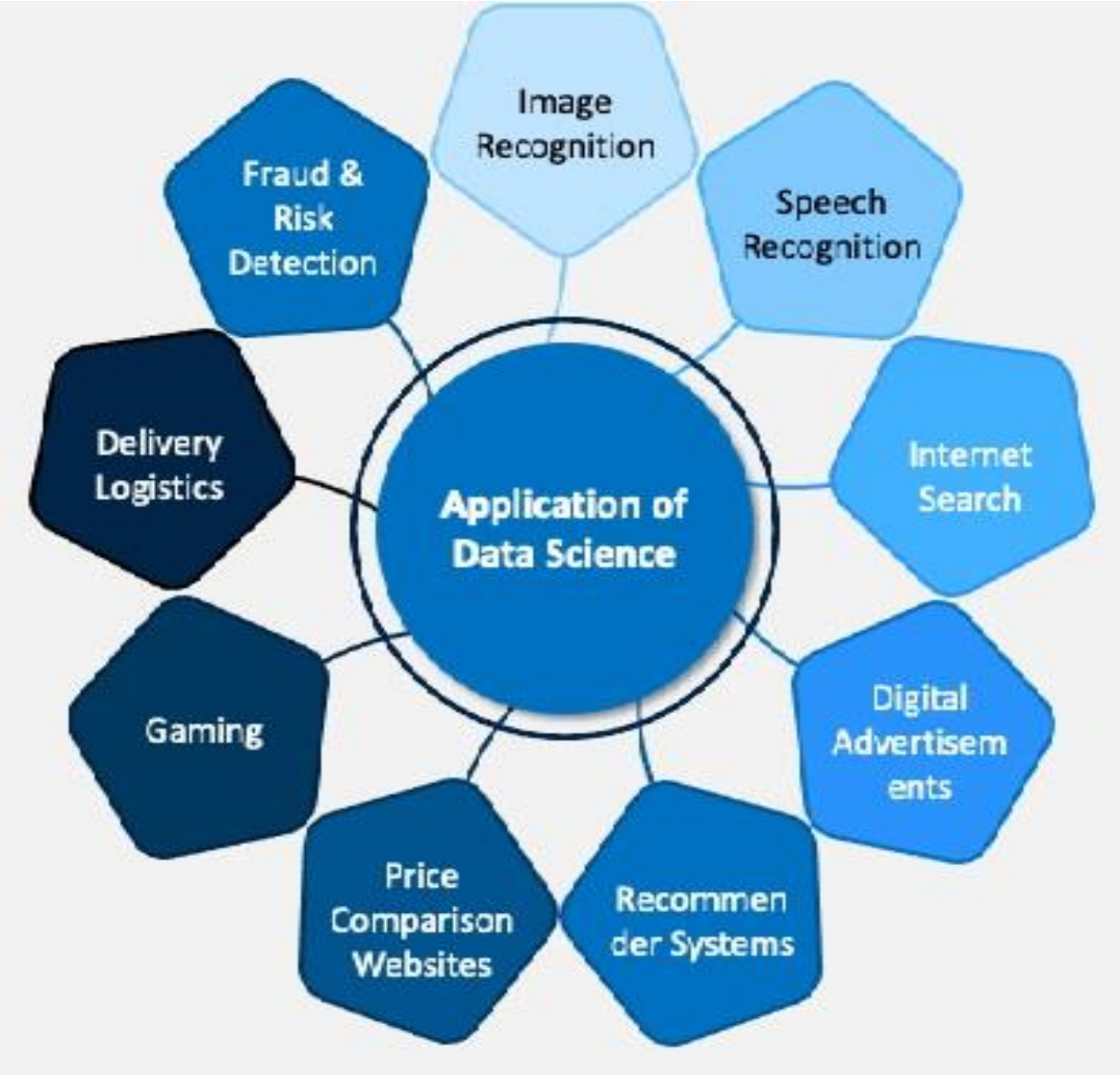
- To find patterns within data.
- It uses various statistical techniques to analyze and draw insights from the data.
- The goal of a Data Scientist is to derive conclusions from the data. Through these conclusions, he is able to assist companies in making smarter business decision.



Basics and Need of Data Science

Applications of Data Science:

1) Fraud and Risk Detection
2) Healthcare
3) Internet Search
4) Targeted Advertising
5) Website Recommendations
6) Advanced Image Recognition
7) Speech Recognition
8) Airline Route Planning
9) Gaming
10) Augmented Reality



Basics and Need of Big Data

Big Data: is a combination of structured, semi-structured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

The world's technological per-capita capacity to store information doubled every 40 months.

- As of 2012, 2.5 exabytes (2.5×1018) of data/day
- Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data.

- Big Data: new driver for digital economy & society
- Gartner: hundreds of billions of GDP by 2020.
- Intangible factor after labor and capital
- Data Science: The fourth paradigm

Basics and Need of Big Data

Applications of Big Data:

1) Tracking Customer Spending Habit, Shopping Behavior
2) Recommendation
3) Smart Traffic System Secure
4) Air Traffic System
5) Auto Driving Car
6) Virtual Personal Assistant Tool
7) IoT
8) Education Sector
9) Energy Sector
10) Media and Entertainment Sector

Data Explosion

• The world is currently used to sparing everything without exception in the electronic space. Processing power, RAM speeds and hard-disk sizes have expanded to level that has changed our viewpoint towards data and its storage

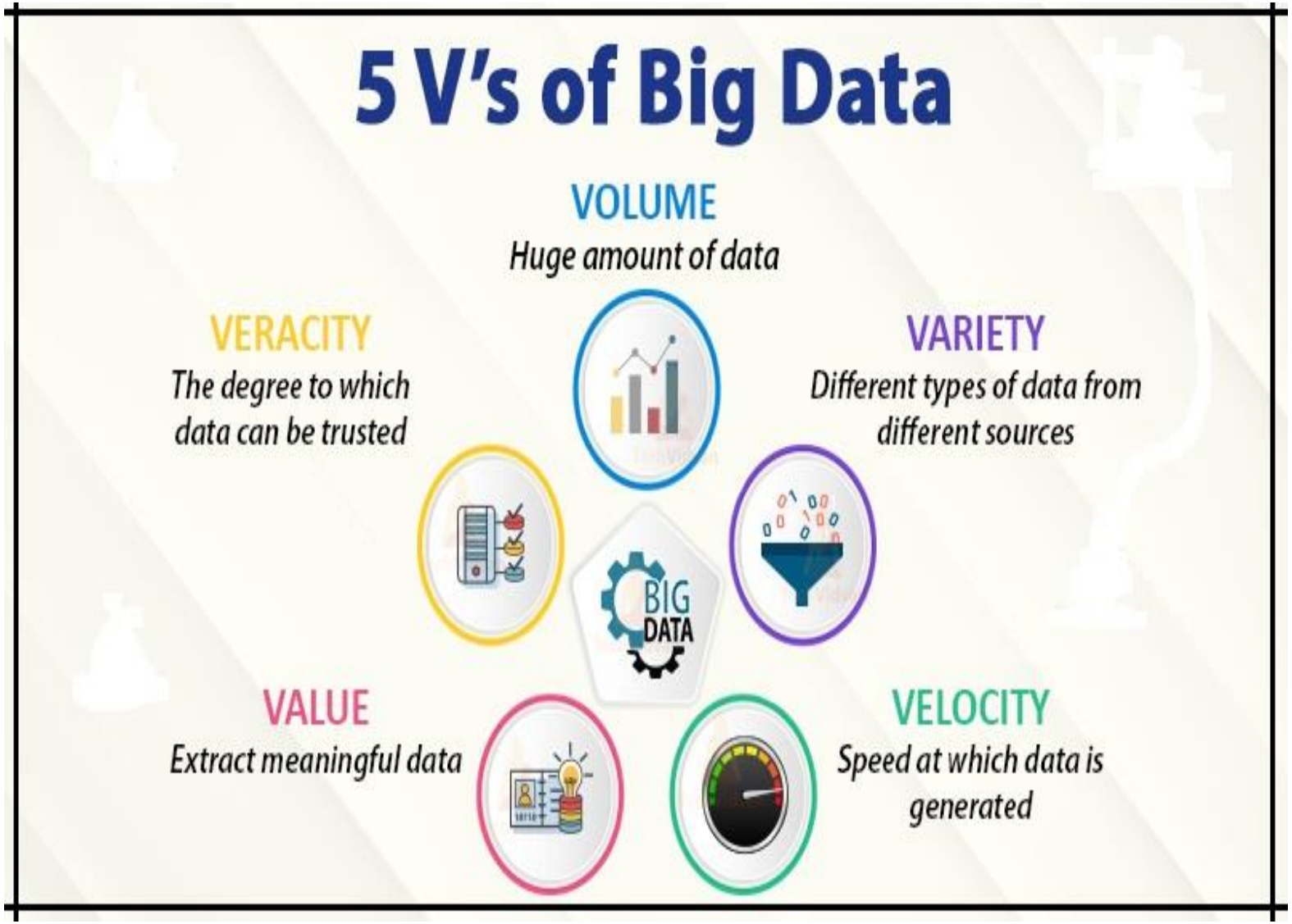
• The rapid or exponential increase in the amount of data that is generated and stored in the computing systems, that reaches level where data management becomes difficult, is called “Data Explosion”.

• Data explosion forms data nature in computer systems. To explore data nature, new theories and methods are required

• The key drivers of data growth are following :

1) Increase in storage capacities
2) Cheaper storage
3) Increase in data processing capabilities by modern computing devices
4) Data generated and made available by different sectors

Data Explosion


5 V's of Big Data
VOLUME
Huge amount of data
VERACITY
The degree to which data can be trusted
VARIETY
Different types of data from different sources
BIG DATA
VALUE
Extract meaningful data
VELOCITY
Speed at which data is generated

5 V's of Big Data

1. Volume:

• The name 'Big Data' itself is related to a size which is enormous.
• Volume is a huge amount of data.
• Example: In the year 2016, the estimated global mobile traffic was 6.2 Exabytes (6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 ExaBytes of data.

2. Velocity:

• Velocity refers to the high speed of accumulation of data.
• In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
• There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
• Example: There are more than 3.5 billion searches per day are made on Google. Also, Facebook users are increasing by 22%(Approx.) year by year.

5 V's of Big Data

3. Variety:

- It refers to nature of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.

1) Structured data: This data is basically an organized data. It generally refers to data that has defined the length and format of data.

2) Semi- Structured data: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.

3) Unstructured data: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database.

4. Veracity:

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy, quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- Example: Data in bulk could create confusion whereas less amount of data could convey half or incomplete information.

5 V’s of Big Data

5. Value:

- After having the 4 V’s into account there comes one more V which stands for Value!. The bulk of data having no Value is of no good to the company, unless you turn it into something useful.

- Data in itself is of no use or importance but it needs to be converted into something valuable to extract information. Hence, you can state that Value! is the most important V of all the 5V’s.

Data Science and Information Science

Data Science:

- Data Science is the discovery of knowledge or actionable information in data.
- Data Science is heavy on Computer Science and Mathematics.
- Data Science is used in Business functions such as strategy formation, decision making and operational processes.

Information Science:

- Information Science is the design of practices for storing and retrieving information.
- Information Science is more concern with areas such as library science, cognitive science and communications.
- Information Science is used in knowledge management, data management and interaction design.

Business Intelligence

Business Intelligence:

•Business Intelligence is a process of **collecting, integrating, analyzing and presenting the data**. With Business Intelligence, executives and managers can have a better understanding of decision-making. This process is carried out through software services and tools.

•Using Business Intelligence, organizations are able to several **strategic and operational business decisions**. Furthermore, BI tools are used for analysis and creation of reports.

•They are also used for **producing graphs, dashboards, summaries, and charts** to help the business executives to make better decisions.

Business Intelligence

Uses of Business Intelligence:

1) Measuring Performance and quantifying the progress towards reaching the business goal.
2) Performing quantitative analysis through predictive analytics and modeling.
3) Visualizing data and storing data in data warehouses and its further processing in OLAP.
4) Using knowledge management programs to develop effective strategies in order to gain insights about learning management and raise compliance issues.

Business Intelligence *vs* Data Science

Factors	Business Intelligence	Data Science
Concept	Deals with data analysis on the business platform.	Consists of several data operations in various domains.
Scope	BI analyzes past data	Past data is analyzed for future predictions.
Data	Handling static and structured data	Both structured & unstructured data that is also dynamic.
Data Storage	Data stored mostly in data-warehouses	Data utilized is distributed in real time clusters.
Procedure	BI helps companies to solve questions.	Questions are both curated and solved by data scientists.
Tools	MS Excel, SAS BI, Sisense, Microstrategy	Python, R, Hadoop/Spark, SAS, TensorFlow.

Data Science Life Cycle

Data Science Life Cycle Steps:

1) Business Understanding:

- To ensure every decision that is being taken should have accurate data that high probability in achieving results.

- Following five types of questions every data scientist should question himself to find the better solution for the problem:

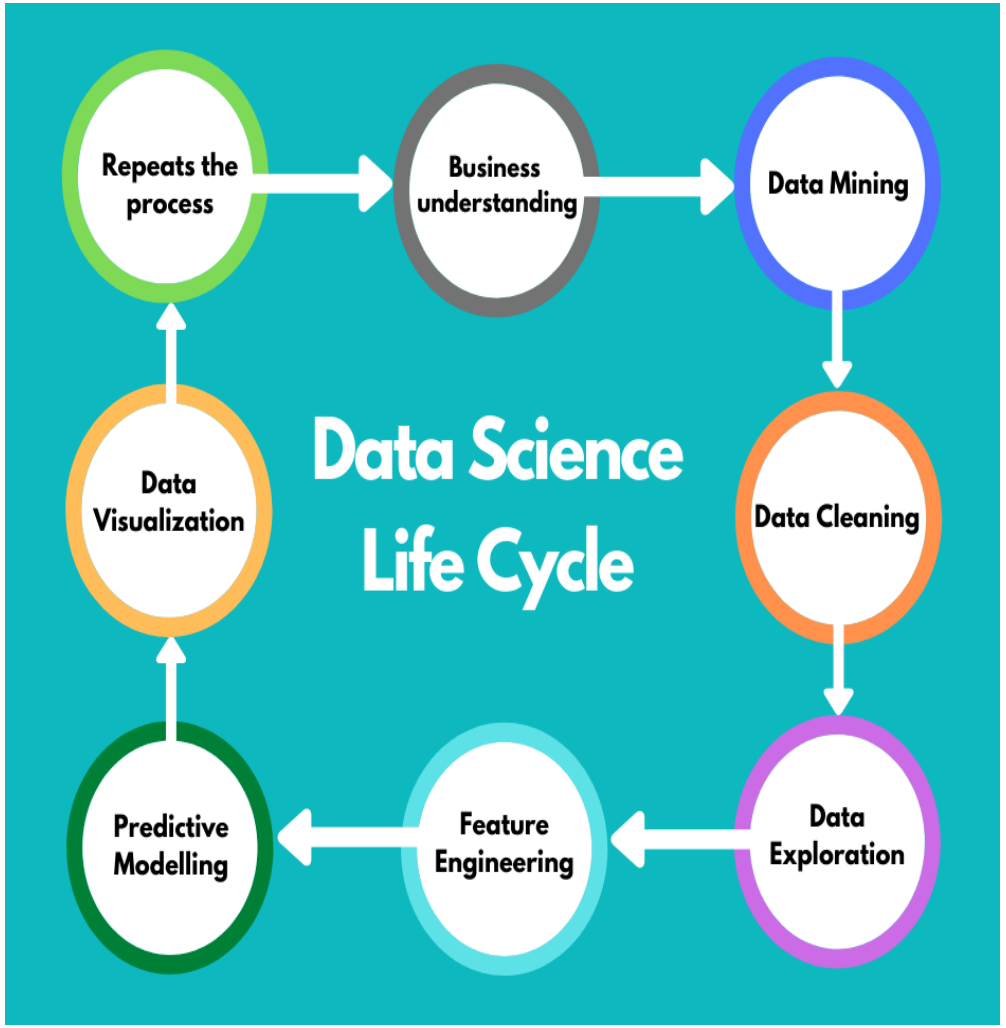
Q.1) How much or How many? (Regression)

Q.2) Which category? (Classification)

Q.3) Which group? (Clustering)

Q.4) Is this weird? (Anomaly Detection)

Q.5) Which option should be taken? (Recommendation)



Data Science Life Cycle

2) Data Mining:

- To start gathering relevant data, the process includes the collection of data from different sources is called Data Mining.
- Mining of data can be done either by data retrieval or cleaning and sometimes both.

3) Data Cleaning:

- Now, we got all required data.
- Preparing and cleaning the unwanted or irrelevant data.

4) Data Exploration:

- Now, our data is clean and ready for analysis.
- In this stage, we will understand the bias and patterns in our data.
- Considering all the information, we can now start to assume our data and methods to tackle our problems.

Data Science Life Cycle

5) Feature Engineering:

- In Machine Learning, a feature is a measurable property of a phenomenon that is being observed.
- Ex. In character recognition, the feature could be histogram counting the number of white pixels.
- The feature engineering is the method of using domain knowledge to change our unstructured data into informative features that represent our problem that we are trying to solve.

- There are two types of tasks that we perform in Feature Engineering:

1) Feature Selection – Filter methods, Wrapper methods, Embedded methods

2) Feature Construction - Creating new features, Continuous change in input variable

Data Science Life Cycle

6) Predictive Modeling:

- In this phase machine learning comes in picture where the concept of predictive modeling is used in data science projects.
- Decide and pick the right prototype that suits our project.

7) Data Visualization:

- Data obtained is a combination of different fields like statistics, communication. art, psychology etc.
- The goal is to communicate the data in an effective and in simple way.
- Once we obtain required insights from our prototype, we need to represent them properly so that various customers in the project should understand.

Data Science Life Cycle

8) Business Understanding:

- Now, we will get back to the first stage of the lifecycle ,it is an iterative process.
- Final phase to know how successful we have completed our project.

- Our model should be able to answer the following questions:

Q.1) Does the model tackle the problems identified?

Q.2) Does the analysis yield any relevant solution?

- If we encounter any new insights while the first iteration of the lifecycle, we can now include them and continue the next iteration, this process continues until the exact answer to the problem is identified.

Data

Data:

- Just as trees are the raw material from which paper is produced, So data be viewed as the raw material from which information is obtained.
- Ex. The height-weight data – numerical and structured data,

  Post a picture using Smartphone – multimedia data

Data Types: There are two types of data types

1) Structured data - Does not need to be strictly numbers

   (Ex. Numbers, Text, Boolean type, Categorical)

2) Unstructured data - Data without labels

   (Ex. Height between 65 inches and 67 inches, IQ 125-130)

Data Collections

There are many places online to look for sets or collections of data.

1) Open Data:

- Data should be freely available in a public domain that can be used by anyone as they wish, without restrictions from copyrights, patents or other mechanisms of control.
- List of principles associated with open data:

1) Public
2) Accessible
3) Described
4) Reusable
5) Complete
6) Timely
7) Managed post release

Data Collections

2) Social Media Data:

- Collecting data to analyze for research or marketing purpose.
- Facilitated by Application Programming Interface (API)
- Ex. Facebook Graph API used by any individual or organization to collect and use this data to accomplish a variety of tasks, such as developing new socially impactful application, research on human information behavior etc.

3) Multimodal Data:

- More devices are exist – need to collect and explore multimodal (different forms) and multimedia (different media) data such as images, music, other sounds, gestures, body posture etc.

Data Collections

4) Data Storage and Presentation:

- Depending on its nature, data is stored in various formats.
- Commonly used formats that stores data as simple text:

1) CSV (Comma-Separated Values) – import and export format for spreadsheet and databases

2) TSV (Tab-Separated Values) – raw data and can be imported into and exported from spreadsheet software

3) XML (eXtensible Markup Language) – store and transport data

4) RSS (Really Simple Syndication) – share data between services

5) JSON (JavaScript Object Notation) – lightweight data interchange format, not only easy for humans to read and write, but also for machines for parse and generate

Need of Data Wrangling

Data Wrangling:

- It is the process of cleaning, structuring and enriching raw data into a desired format for better decision making in less time.
- **Need:** Data has become more diverse, unstructured, demanding increased time spent, culling, cleaning and organizing data ahead broader analysis.

Data Pre-processing

Data Pre-processing:

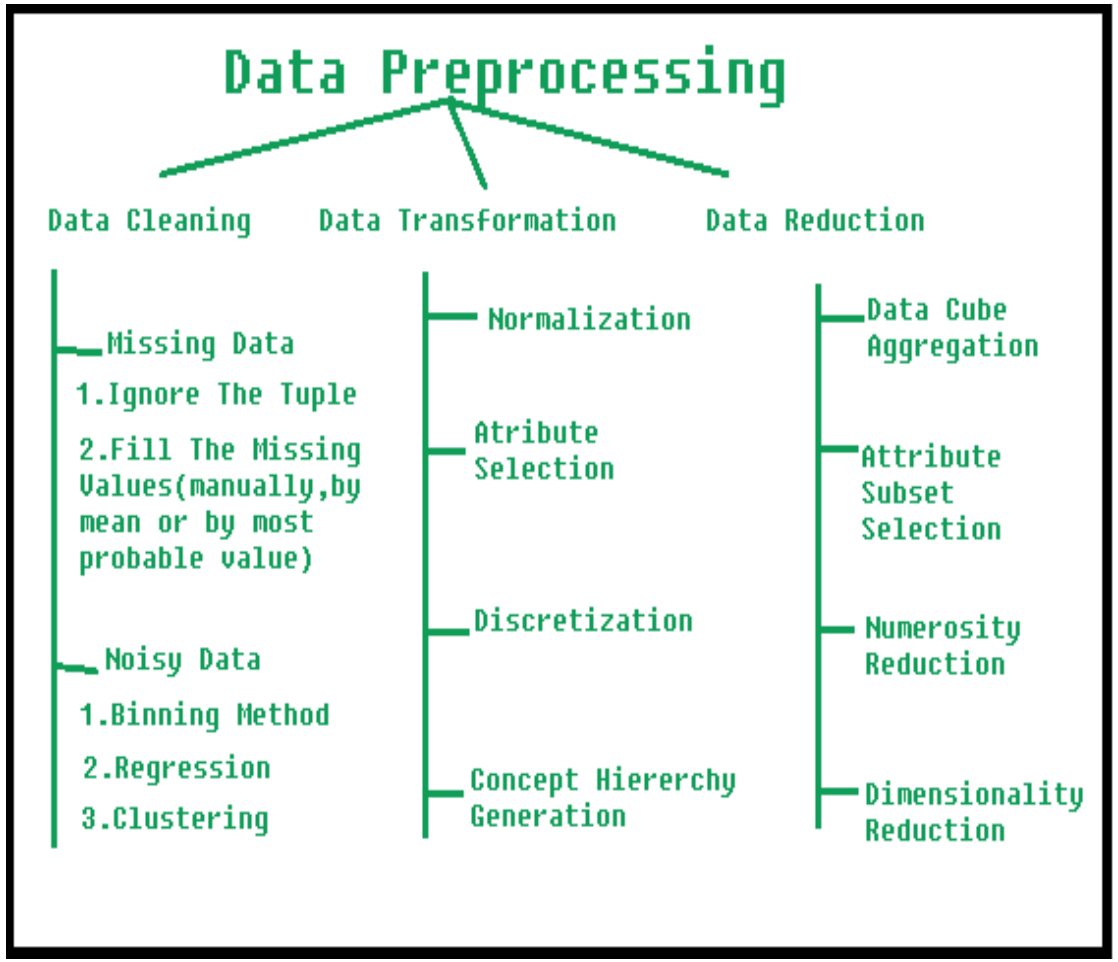
- Data need to be cleaned up before it can be used for a desired purpose, is called data pre-processing.

- Factors that indicate data is not clean or ready to process:

1) Incomplete – some of the attribute values are lacking

2) Noisy – errors or outliers

3) Inconsistent – discrepancies in codes or names



Data Cleaning

- Three key methods that describe ways in which data may be cleaned or better organized:

1) Data Munging (Manipulating/Wrangling) –

- The data is not in a format that is easy to work with.
- Convert data into something more suitable for a computer to understand.

2) Handling Missing Data –

- Sometimes data may be in the right format, but some of the values are missing.
- Data may be missing due to problems with the process of collecting data or equivalent malfunction.

3) Smooth Noisy Data –

- Data is not missing, but it is corrupted for some reason.
- Data corruption may be result of faulty data collection instruments, data entry problems or technology limitations.
- Methods: Binning, Regression, Clustering

Data Integration

Data Integration:

- Data from various sources commonly needs to be integrated.

- The following steps describe how to integrate multiple databases or files:

1) Combine data from multiple sources into a coherent storage.

2) Engage in schema integration or combining of metadata.

3) Detect and resolve data value conflicts.

4) Address redundant data in data integration.

Data Reduction

Data Reduction:

- It aims to increase the storage efficiency and reduce data storage and analysis costs.
- The various steps to data reduction are:

1) Data Cube Aggregation:

Aggregation operation is applied to data for the construction of the data cube.

2) Attribute Subset Selection:

The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

3) Numerosity Reduction:

This enable to store the model of data instead of whole data, for example: Regression Models.

4) Dimensionality Reduction:

This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

Data Transformation

Data Transformation:

- This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1) Normalization:

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2) Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3) Discretization:

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

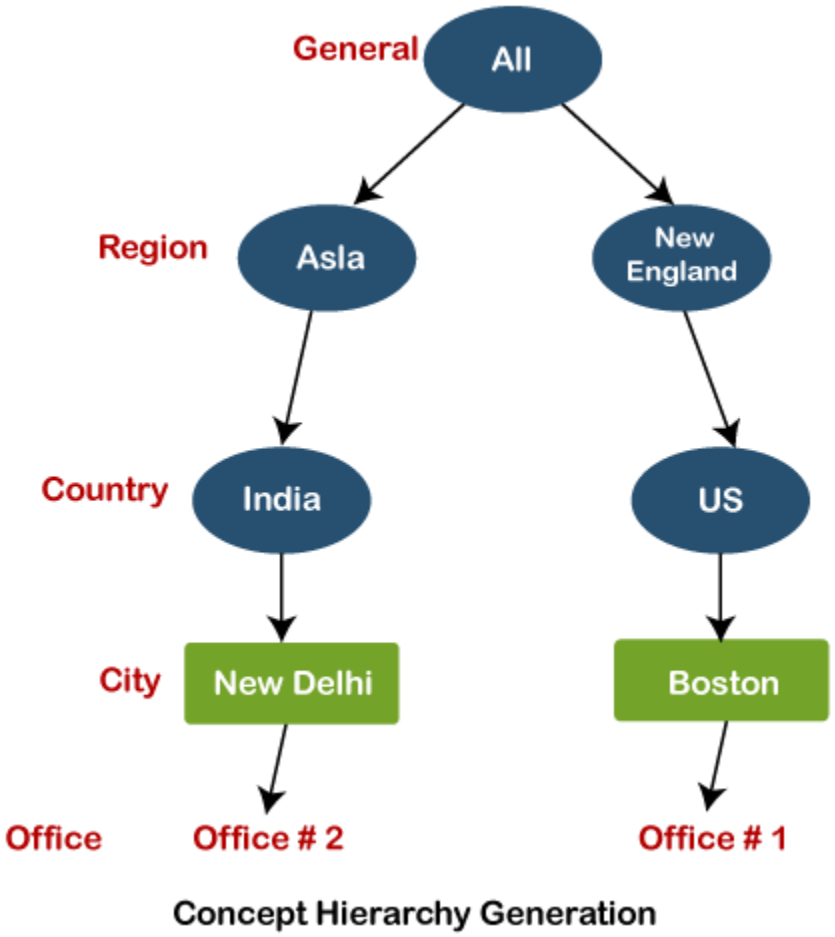
4) Concept Hierarchy Generation:

Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Data Discretization

Data Discretization:

- Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals with minimal loss of information.
- Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. Another example is analytics, where **we gather the static data of website visitors**.



Concept Hierarchy Generation

Case Study

Create academic performance dataset of students and perform data pre-processing using techniques of data cleaning and data transformation.