Unit-I: Introduction to Data Science and Big Data

Syllabus -

Basics and need of Data Science and Big Data, Applications of Data Science, Data explosion, 5 V's of Big Data, Relationship between Data Science and Information Science, Data Science Life Cycle, Data: Data Types, Data Collection. Need of Data wrangling, Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization.

Data Science

- It is the area of study which involves **extracting insights** from vast amounts of data by the use of **various scientific methods, algorithms, and processes**.
- It helps you **to discover hidden patterns** from the raw data.
- The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.
- It allows to extract knowledge from structured or unstructured data.
- **It enables to translate a business problem into a research project and then translate it back into a practical solution.**

Why Data Science?

Advantages:

- With the right tools, technologies, algorithms, we can use data and **convert it into a distinctive business advantage**
- It can help to **detect fraud** using advanced machine learning algorithms
- It helps to **prevent** any significant **monetary losses**
- Allows to **build intelligence** ability in **machines**
- You can perform **sentiment analysis** to gauge customer brand loyalty
- It enables to take **better and faster decisions**
- Helps to **recommend** the right product to the right customer to enhance your business

Data Science Components

Domain Expertise

Statistics

$$z = \frac{x - \mu}{\sigma}$$

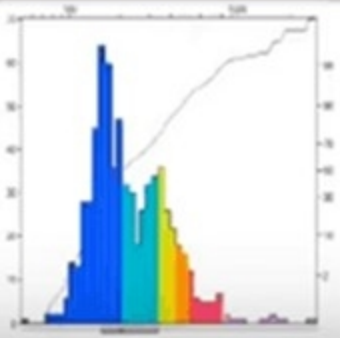
μ = Mean

σ = Standard Deviation

Data Engineering

Data Science

Visualization



Advanced Computing

Applications of Data Science

Internet Search:

- Google search use Data science technology to search a specific result within a fraction of a second

Recommendation Systems:

- To create a recommendation system. Example, "suggested friends" on Facebook or suggested videos" on YouTube

Image & Speech Recognition:

- Speech recognizes system like Siri, Google assistant, Alexa run on the technique of Data science. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

Applications of Data Science

- Gaming world:
  - EA Sports, Sony, Nintendo, are using Data science technology.
  - Games are now developed using Machine Learning technique. It can update itself when you move to higher levels.
- Online Price Comparison:
  - PriceRunner, Junglee, Shopzilla work on the Data science mechanism.

Big Data Overview

- Industries that gather and exploit data
  - Credit card companies monitor purchase
    - Good at identifying fraudulent purchases
  - Mobile phone companies analyze calling patterns – e.g., even on rival networks
    - Look for customers might switch providers
  - For social networks data is primary product
    - Intrinsic value increases as data grows

Attributes Defining Big Data Characteristics

- Huge volume of data
  - Not just thousands/millions, but billions of items
- Complexity of data types and structures
  - Varity of sources, formats, structures
- Speed of new data creation and grow
  - High velocity, rapid ingestion, fast analysis

Attributes Defining Big Data Characteristics

1. Volume
   - Big Data observes and tracks what happens from various sources which include business transactions, social media and information from machine-to-machine or sensor data. This creates large volumes of data.
2. Variety
   - Data comes in all formats that may be structured, numeric in the traditional database or the unstructured text documents, video, audio, email, stock ticker data.
3. Velocity
   - The data streams in high speed and must be dealt with timely. The processing of data that is, analysis of streamed data to produce near or real time results is also fast.

4. Veracity:

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and 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. 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.

Big Data Analytics Importance

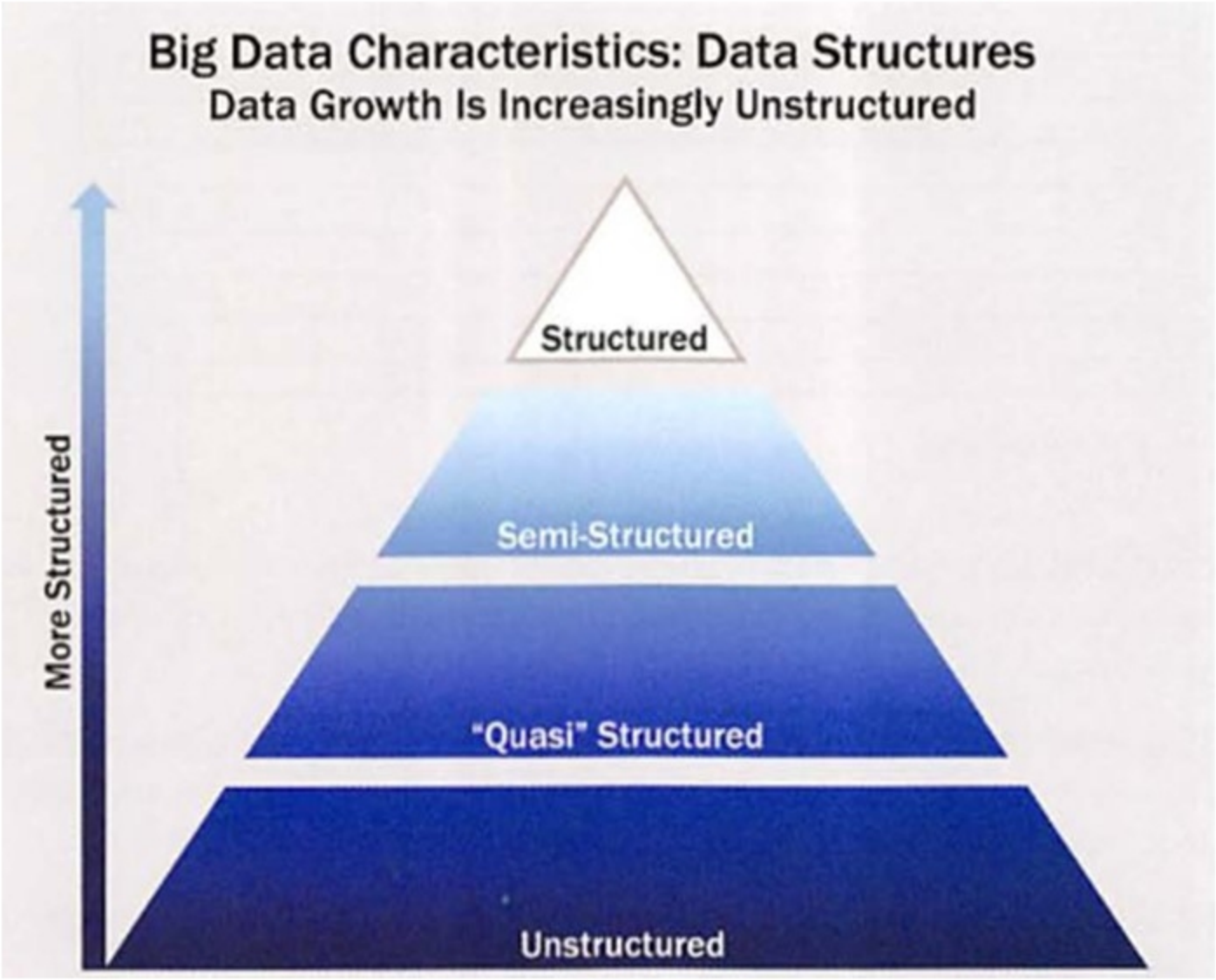
- **Cost Savings :** help in identifying more efficient ways of doing business.
- **Time Reductions :** helps businesses analyzing data immediately and make quick decisions based on the learnings.
- **New Product Development :** By knowing the trends of customer needs and satisfaction through analytics you can create products according to the wants of customers.
- **Understand the market conditions :** By analyzing big data you can get a better understanding of current market conditions.
- **Control online reputation:** Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company.

Sources of Big Data Deluge

- **Mobile sensors** – GPS, accelerometer, etc.
- **Social media** – 700 Facebook updates/sec in 2012
- **Video surveillance** – street cameras, stores, etc.
- **Video rendering** – processing video for display
- **Smart grids** – gather and act on information
- **Geophysical exploration** – oil, gas, etc.
- **Medical imaging** – reveals internal body structures
- **Gene sequencing** – more prevalent, less expensive, healthcare would like to predict personal illnesses

Sources of Big Data Deluge

Data Structures: Characteristics of Big Data



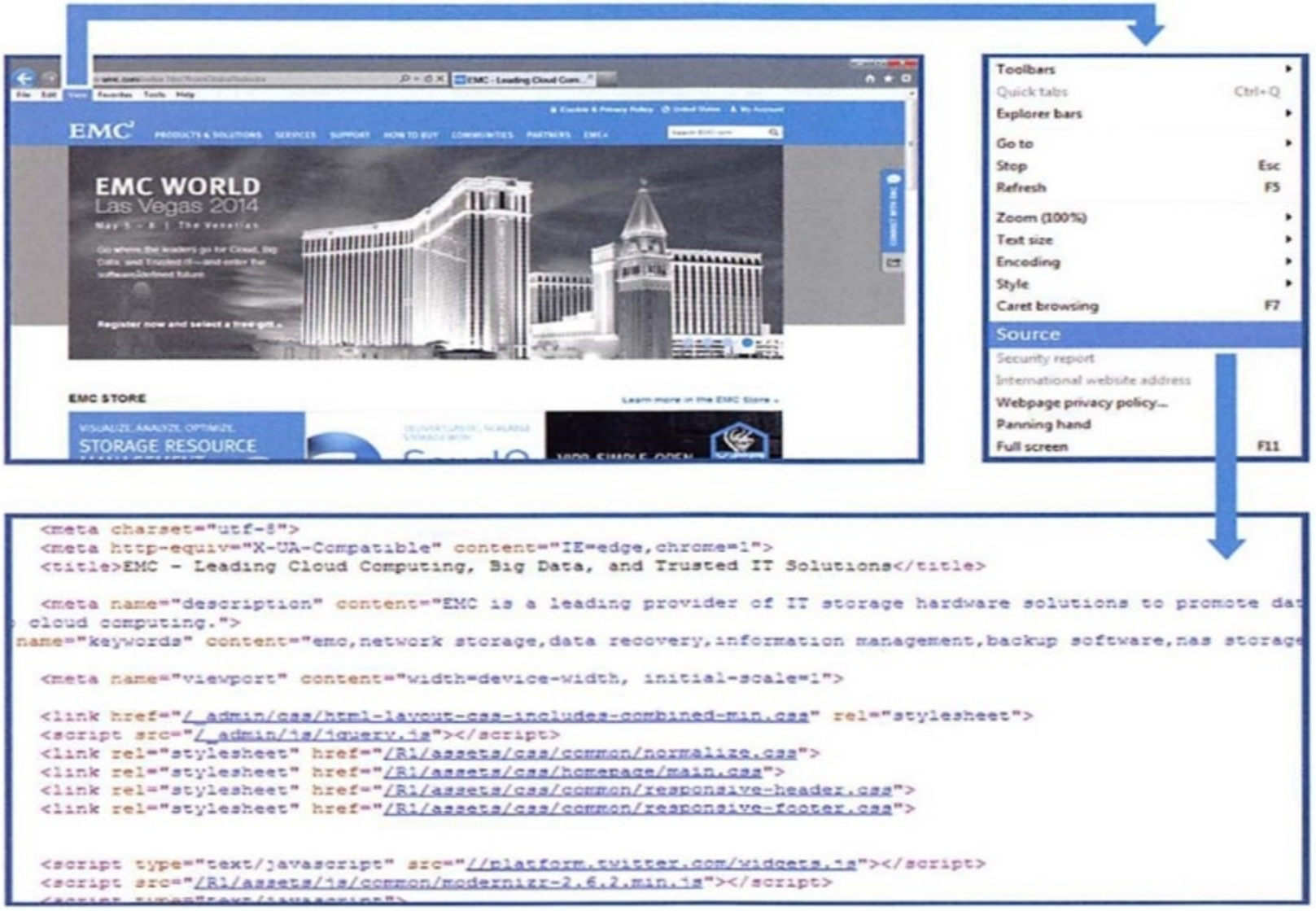
Data Structures: Characteristics of Big Data

- Structured – defined data type, format, structure
  - *Transactional data, OLAP cubes, RDBMS, CVS files, spreadsheets*
- Semi-structured
  - *Text data with discernable patterns – e.g., XML data*
- Quasi-structured
  - *Text data with erratic data formats – e.g., clickstream data*
- Unstructured
  - *Data with no inherent structure – text docs, PDF's, images, video*

Example of Structured Data

Rno	Name	Address	Phone no
1	Amit	Nashik	9766543267
2	Neha	Pune	-
3	Jiya	Mumbai	-
4	Riya	Aurangabad	8990765432

Example of Semi-Structured Data



Example of Quasi-Structured Data
visiting 3 websites adds 3 URLs to user's log files

1

https://www.google.com/#q=EMC+data+science

2

https://education.emc.com/guest/campaign/data_science.aspx

3

https://education.emc.com/guest/certification/framework/stf/data_science.aspx

Example of Unstructured Data Video about Antarctica Expedition

Types of Data Repositories from an Analyst Perspective

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.
Data Warehouses	Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

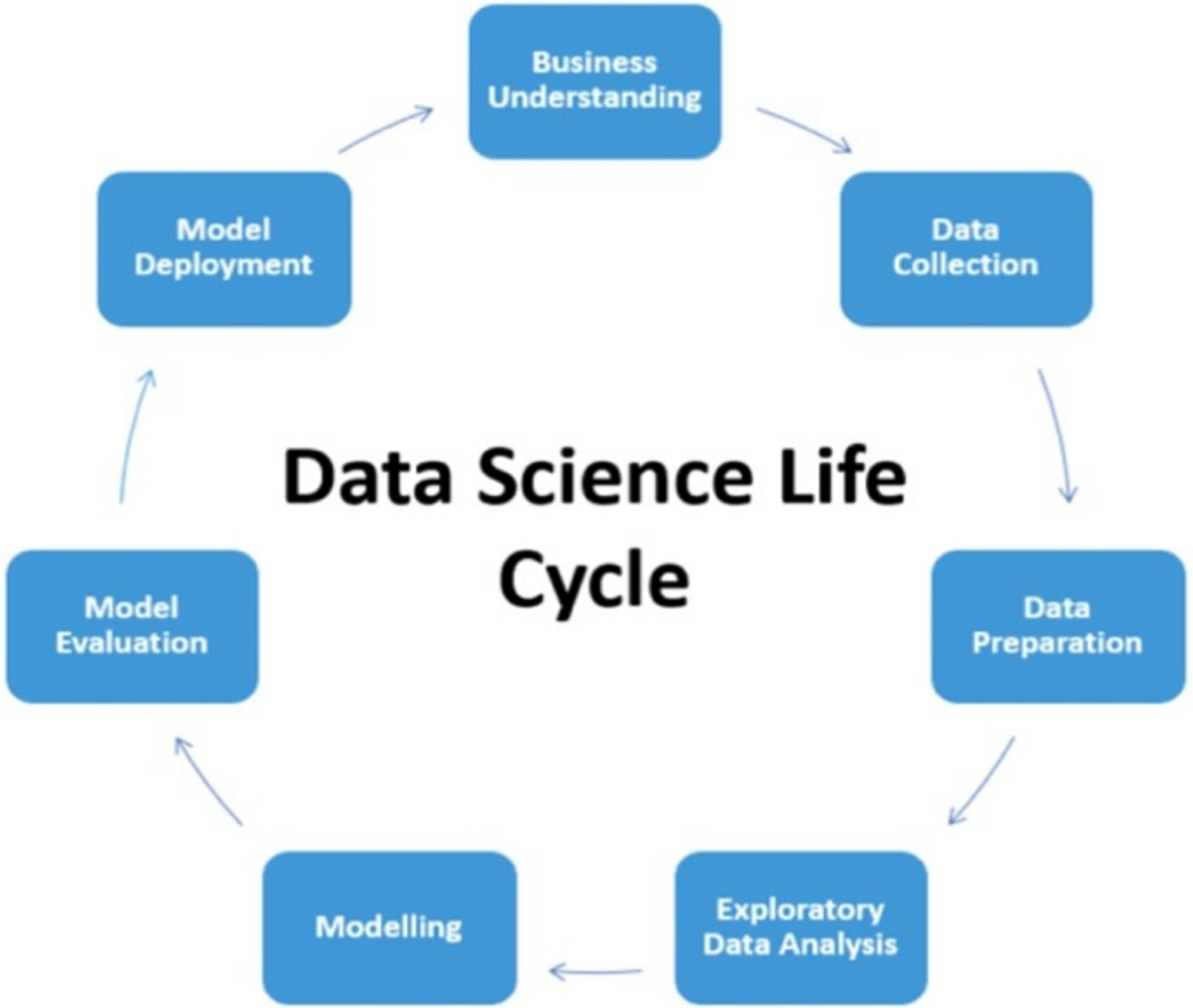
The Relationship between Data Science and Information Science

- The field of information science, which often stems from computing, computational science, informatics, information technology, or library science, often represents and serves such application areas.
- Covers people studying, accessing, using, and producing information in various contexts.

Information vs. Data

- A traditional view used to be that data is something raw, meaningless, an object that, when analyzed or converted to a useful form, becomes information.
- Information is also defined as "data that are endowed with meaning and purpose."
- The Data, Information, Knowledge, and Wisdom (DIKW) model differentiates the meaning of each concept and suggests a hierarchical system among them.

Data Science Life Cycle



Data Science Life Cycle

1. Business Understanding

- Plays a key role in success of any project.
- Success of any project depends on the quality of questions asked for the dataset.
- Every domain and business work with a set of rules and goals.
- In order to acquire the correct data, we should be able to understand the business.
- Asking questions about dataset will hep in narrowing down to correct data acquisition.

2. Data Collection

- Data serves important ingredient for making any Data Science project.
- Data could be from various sources which could be –
  - *logs from webservers,*
  - *data from online repositories,*
  - *data from databases,*
  - *social media data,*
  - *data in excel sheet, etc.*
- A major **challenge** faced by data professionals in data acquisition step is to understand where the data comes from and whether it is the **latest** data or not.
- It makes it a crucial step to keep a **track** all through the project life cycle as data might to be re-acquired to do analytics and reach to conclusions.

3. Data Preparation

- Data may be or may not be in required format.
- It could also be said that data needs to be cleaned before processing any further. Thus, this step is also known as **Data Cleaning or Data Wrangling**.
- Data acquired in previous step might not give clear analytical picture or patterns in the data. So, to understand this data needs to be **structured and cleaned.**
- Might be data is obtained from different sources but for analysis data need to be **clubbed** together from different sources. This is also referred as **structuring** the data.
- Apart from this data might have missing values which will cause obstruction in analysis and model building.
- **Exploratory Data Analysis (EDA)** plays an important role at this stage as summarization of clean data helps in identifying the structure, outliers, anomalies and patterns in the data. These insights could help in **building the model.**

4. Data Modelling

- This stage seems to be most interesting one to almost all of the data scientists.
- **Feature selection** is one of the first things that you would like to do in this stage.
- Not all features might be essential for making the predictions. What needs to be done here is to **reduce the dimensionality** of the dataset.
- It should be done such that features contributing to the prediction results should be selected.

- Based on the business problem models could be selected.
- It is essential to identify what is the **ask**, is it a classification problem, regression or prediction problem, time series forecasting or a clustering problem.
- After the modelling process, **model performance measurement** is required.
- For this precision, recall, F1-score for classification problem could be used. For regression problem R2, MAPE (Moving Average Percentage Error) or RMSE (Root Mean Square Error) could be used.

5. Interpreting Data

- This is the last step of any Data Science project and also the most important step.
- The predictive power of the model lies in its ability to **generalize**.
- **Visualization** of findings should be done. It should be in line with business questions.
- It should be meaningful to the organization and the stakeholders.
- All the above steps make a complete Data Science project but it is an iterative process and various steps are repeated until we are able to fine tune the methodology for a specific business case.

Data Wrangling

- It is the process of taking disorganized or incomplete raw data and **standardizing** it so that you can easily access, consolidate, and analyze it.
- It also involves **mapping** data fields from source to destination, for example, targeting a field, row, or column in a dataset and implementing an action like joining, parsing, cleaning, consolidating, or filtering to produce the required output.
- The process ensures that the data is ready for automation and further analysis.

Why Do You Need Data Wrangling?

- Data professionals spend almost **73% of their time** just wrangling the data.
- It helps business users make **concrete, timely decisions** by cleaning and structuring raw data into the required format.
- Accurately wrangled data ensures that quality data is entered into analytics or downstream processes for **consolidation and collaboration.**
- Data wrangling can be arranged into a consistent and repeatable **procedure** using data integration tools with automation capabilities that clean and convert source data into a reused format as per the end requirements.

Data Pre-processing

- Data in the real world is often dirty; that is, it is in need of being cleaned up before it can be used for a desired purpose.
- This is often called data pre-processing.
- The factors that indicate that data is not clean or ready to process:
  - ***Incomplete***. *When some of the attribute values are lacking, certain attributes of interest are lacking, or attributes contain only aggregate data.*
  - ***Noisy***. *When data contains errors or outliers. For example, some of the data points in a dataset may contain extreme values that can severely affect the dataset's range.*
  - ***Inconsistent***. *Data contains discrepancies in codes or names.*

Data Cleaning

Data Cleaning

Data Cleaning

- Since there are several reasons why data could be "dirty," there are just as many ways to "clean" it.
- Reasons for "dirty" or "unclean" data
  - Dummy values
  - Absence of data
  - Violation of business rules
  - Data integration problems
  - Contradicting data
  - Inappropriate use of address line
  - Reused primary keys
  - Non-unique identifiers

Data Cleaning

What to do to clean data?

- Handle Missing Values
- Handle Noise and Outliers
- Remove Unwanted data

Data Cleaning

Handle Missing Values

- **Ignore the data row:** This method is suggested for records where maximum amount of data is missing.
- This method is usually avoided where only less attribute values are missing. If all the rows with missing values are ignored i.e. removed, it will result in poor performance.
- **Fill the missing values manually:** This is a very time consuming method and hence infeasible for almost all scenarios.
- **Use a global constant to fill in for missing values:** A global constant like "NA" or 0 can be used to fill all the missing data. This method is used when missing values are difficult to be predicted.

Data Cleaning

Handle Missing Values

- **Use attribute mean or median:** Mean or median of the attribute is used to fill the missing value.
- **Use forward fill or backward fill method**: In this, either the previous value or the next value is used to fill the missing value.
- A mean of the previous and succession values may also be used.
- **Use a data-mining algorithm to predict the most probable value**

Data Cleaning

Handle Noise and Outliers

- Noise in data may be introduced due to **fault in data collection, error during data entering** or due to data transmission errors, etc..
- Noise can be handled using **binning**. In this technique, sorted data is placed into bins or buckets. Bins can be created by equal-width (distance) or equal-depth (frequency) partitioning. On these bins, smoothing can be applied. Smoothing can be by bin mean, bin median or bin boundaries.

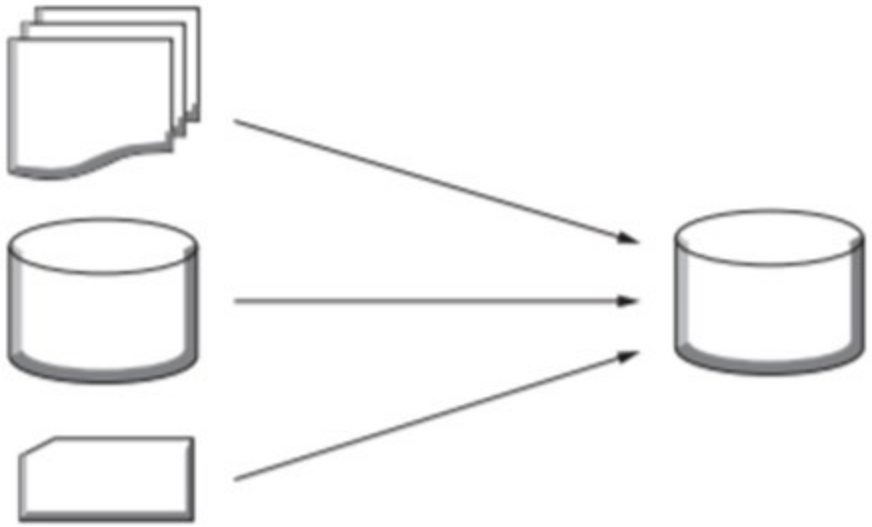
Data Cleaning

Remove Unwanted Data

- Unwanted data is **duplicate or irrelevant** data.
- Scraping data from **different sources** and then **integrating** may lead to some duplicate data if not done efficiently.
- This redundant data should be removed as it is of no use and will only **increase the amount of data and the time to train the model**.
- Due to redundant records, the model may not provide accurate results as the duplicate data interferes with analysis process, giving more importance to the repeated values.

Data Integration

Data Integration



Data Integration

- To be as efficient and effective for various data analyses as possible, 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 place (e.g., a single file or a database).
2. Engage in schema integration, or the combining of metadata from different sources.
3. Detect and resolve data value conflicts.
   a. For example: A conflict may arise; for instance, such as the presence of different attributes and values from various sources for the same real-world entity.
   b. Reasons for this conflict could be different representations or different scales; for example, metric vs. British units.

Data Integration

4. Address redundant data in data integration. Redundant data is commonly generated in the process of integrating multiple databases.

For example:

a. The same attribute may have different names in different databases.
b. One attribute may be a "derived" attribute in another table; for example, annual revenue.
c. Correlation analysis may detect instances of redundant data.

Data Transformation

Data Transformation –17, 25, 39, 128, –39 ⟶ 0.17, 0.25, 0.39, 1.28, –0.39

Data Transformation

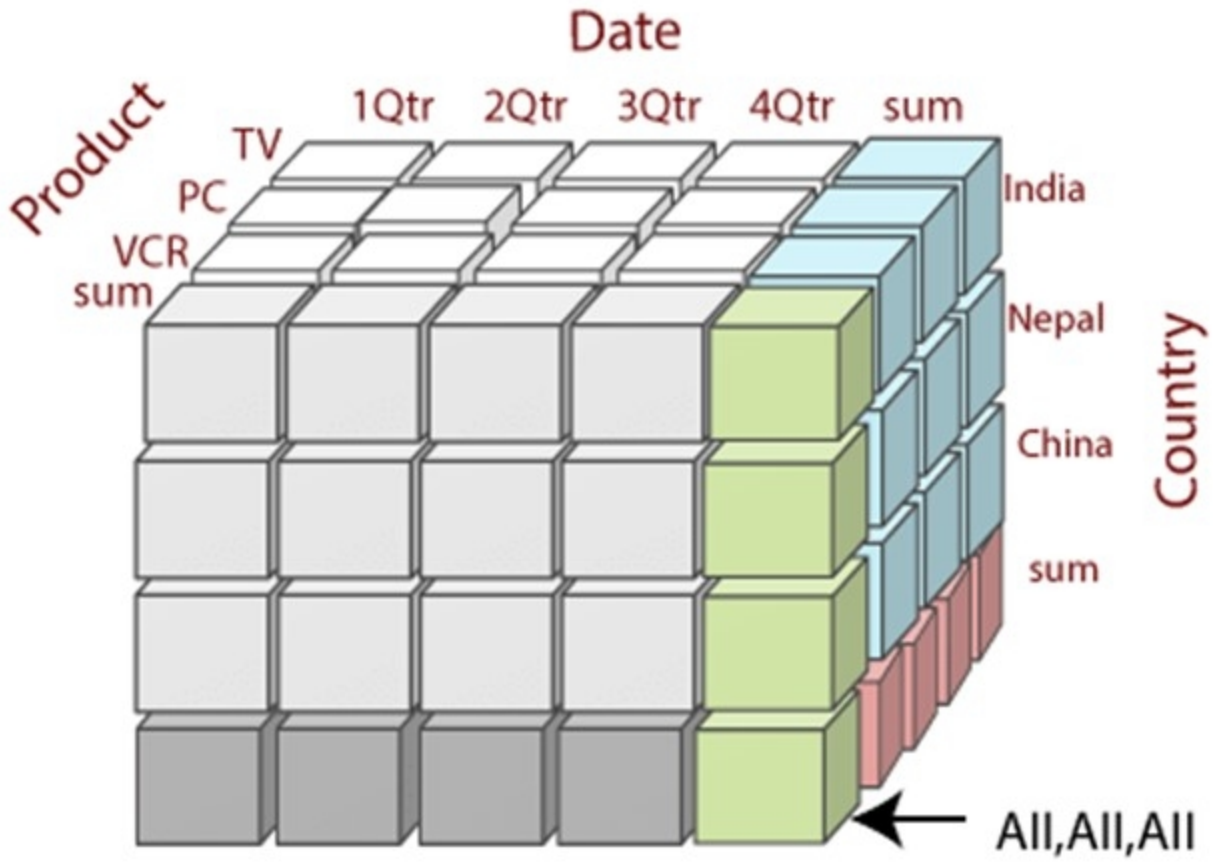
- Data must be transformed so it is consistent and readable (by a system).

The following five processes may be used for data transformation.

1. Smoothing: Remove noise from data.
2. Aggregation: Summarization, data cube construction.
3. Generalization: Concept hierarchy climbing.
4. Normalization: Scaled to fall within a small, specified range and aggregation.
5. Attribute or feature construction.

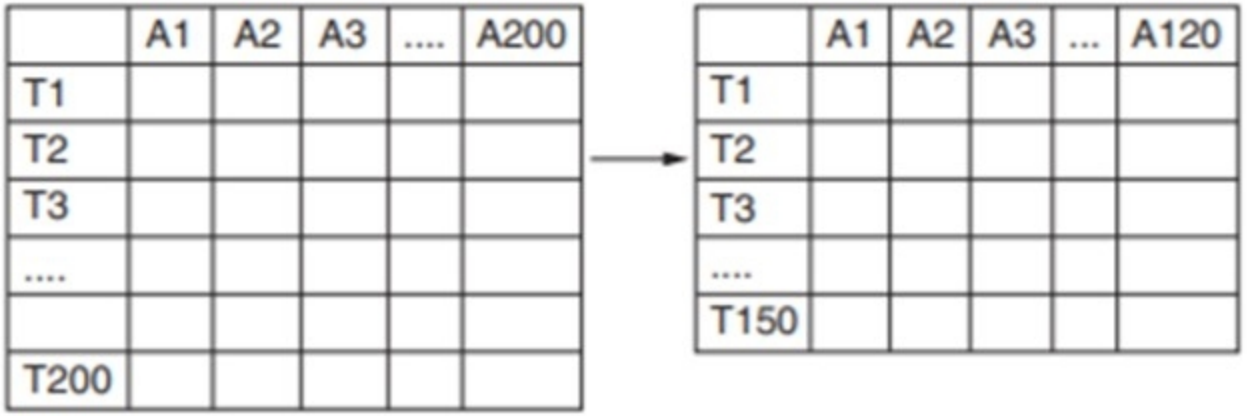
   New attributes constructed from the given ones

Data Cube


Date
1Qtr
2Qtr
3Qtr
4Qtr
sum
Product
TV
PC
VCR
sum
India
Nepal
China
sum
Country
All,All,All

Data Reduction

Data Reduction



	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Data Reduction

- Data reduction is a key process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained.
- One example of a large dataset that could warrant reduction is a data cube.
- Data cubes are multidimensional sets of data that can be stored in a spreadsheet.
- A data cube could be in two, three, or a higher dimension.
- Each dimension typically represents an attribute of interest.
- We could reduce information from all those dimensions to much smaller and manageable without losing much.
- The most common techniques used for data reduction

1. **Data Cube Aggregation.** The lowest level of a data cube is the aggregated data for an individual entity of interest.

- We reduce the data to its more meaningful size and structure for the task at hand.

2. **Dimensionality Reduction.** In contrast with the data cube aggregation method, where the data reduction was with the consideration of the task, dimensionality reduction method works with respect to the nature of the data.

- A dimension or a column in the data spreadsheet is referred to as a "**feature**," and the goal of the process is to identify which features to remove or collapse to a combined feature.
- This requires identifying redundancy in the given data and/or creating composite dimensions or features that could sufficiently represent a set of raw features.
- **Strategies** for reduction include **sampling, clustering, principal component analysis**, etc.

Data Discretization

- We are often dealing with data that are collected from processes that are continuous, such as temperature, ambient light, and a company's stock price.
- But sometimes we need to convert these continuous values into more manageable parts. This mapping is called **discretization**.
- And as you can see, in undertaking discretization, we are also essentially reducing data.
- This process of discretization could also be perceived as a means of data reduction, but it holds particular importance for numerical data.

Data Discretization

- There are three types of attributes involved in discretization:
  - *a. Nominal: Values from an unordered set*
  - *b. Ordinal: Values from an ordered set*
  - *c. Continuous: Real numbers To achieve discretization, divide the range of continuous attributes into intervals.*
- For instance, we could decide to split the range of temperature values into cold, moderate, and hot, or the price of company stock into above or below its market valuation.