Data Science

Second Year

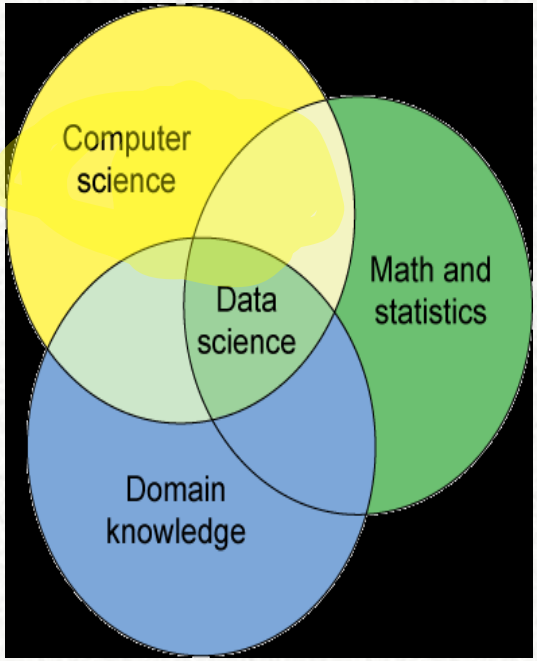
Why to study Data Science?

Lots of data everywhere around you.

- Social Media
- Banking
- Railways
- Economic and Finance
- Sports
- E-commerce
- Conversational Agents
- Transport
- Health Care
- Industries
- Google,Yahoo

What is Data Science?

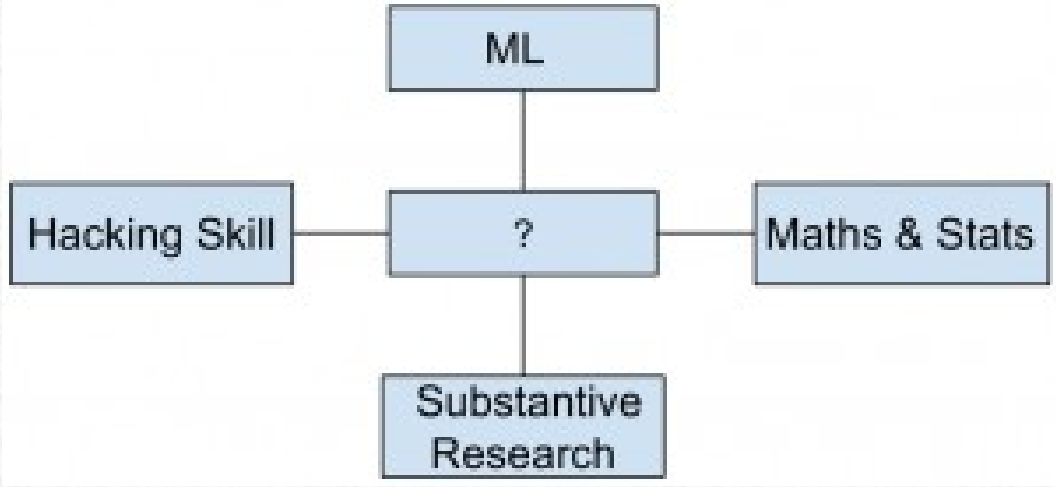
It is a combination of **Mathematics and Statistics**, **Domain Knowledge** and **Computer Science** to gain insight into a real world problem expressed using data.



Definition of Data Science

- Data Science is the application of **computational** and **statistical** techniques
- **Computational** because data science typically involves some sort of algorithm methods written in a code.
- **Statistical** because inferences based on statistics help us to build predictions that we make.

Which of the following would be more appropriate to be replaced with question mark in the following figure?



a) Data Analysis
b) Data Science
c) Descriptive Analytics
d) None of the mentioned

b) Data Science is a multidisciplinary field which involves extraction of knowledge from large volumes of data that are structured or unstructured.

What is Data?

- Data are the values of **qualitative** or **quantitative variables** belonging to a **set of items**.
- **Variables** are the measurement or characteristics of the item. They might be measured in qualitative terms or quantitative terms.
- **Qualitative terms** could be such as country of origin, medical treatment , gender etc.
- **Quantitative terms** could be height, weight , blood pressure etc.
- **Set of items** may be population or the set of items we are interested in. For example: To give the count of Indian census, the population of Indian people should be considered , population from other country will not make sense. Thus Indian citizen becomes our set of interests

Qualitative data

- Non numerical

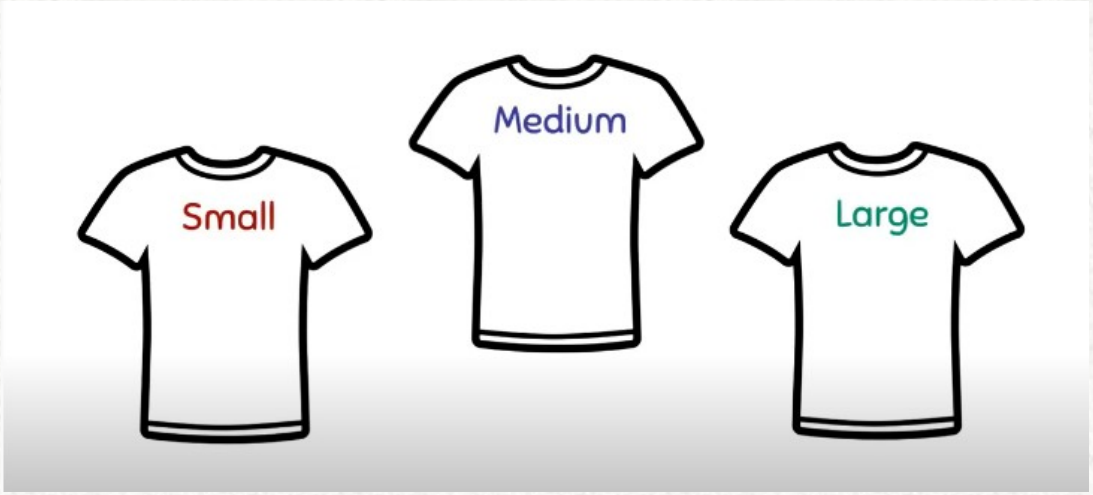
- Robust Aroma
- Frothy

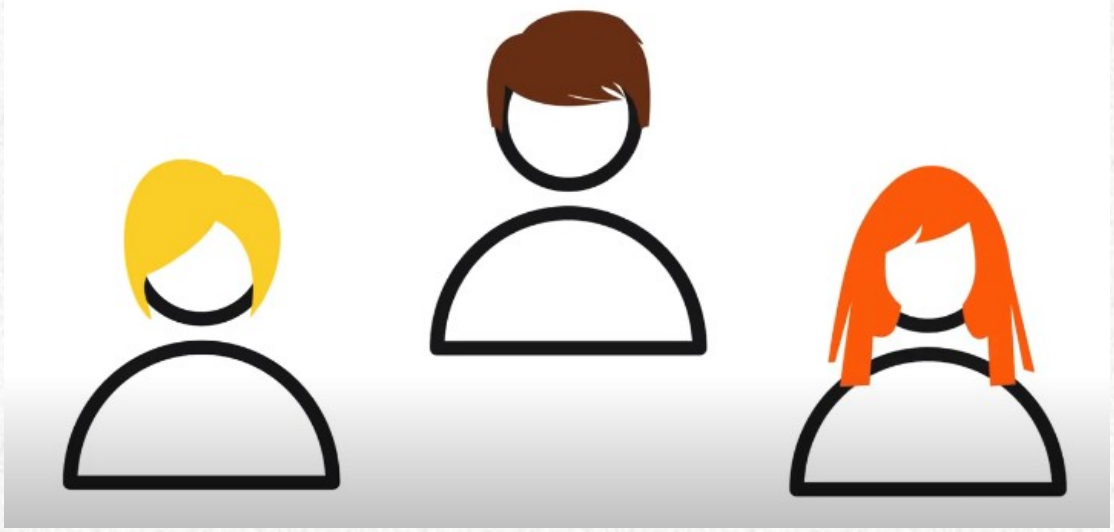
Appearance

- Strong Taste
- Blue Cup

Used for.....

Example.....





Quantitative data

- Referring to a number

- 200 ml
- 160 Degrees Fahrenheit
- 30 INR

Types of Quantitative Data

- Discrete data
  - a count
  - involves whole numbers
  - e.g. No of Children in your family

- Continuous data
  - Numerical data
  - any values in certain range
  - e. g. Temperature

Why do we use Data?

- To ask a question
- To tell a story
- To find patterns

How do we collect data?

- Primary sources
- Secondary sources

Data collected by researchers

Interviews
Observations
Case studies
Questionnaires

Data that already exists

Official statistics
Web information
Government Reports
Previous research

Example of Data with Different Variables

The above data set contains information about the drivers of different countries, their weight, gender, speed of vehicle(car) they drive. The data can be used to analyze the following :

i) What is the average speed of indian drivers?

ii) What is the general speed of female drivers from Italy?

iii) Whether Male drivers drive fast or female?

S.No	Name of driver	Country	Age	Avg Speed of vehicle	Gender
1	John	USA	28	80Km/Hr	M
2	Lisa	Italy	35	65 Km/Hr	F
...	.....	.....	.....	.....	.....
100	Rahul	India	48	70Km/Hr	M

Types of Variable

Numerical (Quantitative)- Takes numerical values which can be added, subtracted etc

- Continuous
- Discrete

Categorical (Qualitative)- Takes on a limited number of distinct categories.

- Ordinal
- Regular

Numerical Variable

Continuous Variable : Takes any of the value in given range e.g height.

Discrete Variable: Takes one of the specific set of numerical values e.g number of two wheelers ,no. of houses a person has.

Height is a continuous variable but we tend to report it as discrete variable by rounding its value. Thus rounding of continuous variable make them appear as discrete.

Categorical Variable

Ordinal categorical variable: The categorical variables having ordered level are ordinal variables e.g customer satisfaction survey for a service can be very satisfactory, satisfactory , neutral , unsatisfactory , very unsatisfactory or speed if low, medium , high etc.
Regular categorical variable: Does not have any order. E.g which chocolate do you like dark or white, gender is female or male etc.

Case Study : Google Transparency Report

Country	Cr_req	Cr_comply	Ud_req	Ud_comply	Hemisphere	hdi
Argentina	21	100	134	32	Southern	Very high
Australia	10	40	361	73	Southern	Very high
Belgium	6	100	90	67	Northern	Very high
Brazil	224	67	703	82	Southern	high
USA	92	63	5950	93	Northern	Very high

Description of variables

Country: Identifier variable, indicating name of the country for which data are gathered.
Cr_req: No.of content removal requests made by the respective country. It is Discrete numerical variable.
Cr_comply: Percentage of content removal requests that Google has complied with. It is Continuous numerical variable.
Ud_req: No.of user data requests by the country as a part of criminal investigation. It is discrete numerical variable.
Ud_comply: Percentage of user data requests that Google has complied with. It is continuous numerical data variable.
Hemisphere: Whether the country is in southern or northern hemisphere. It is regular categorical variable.

Univariate , Bivariate , Multivariate Data analysis

Univariate Data: Contains only one variable.
e.g: Height of applicants in army recruitment camp measured in cms.

Height	164	168	170	172	174	178	180	....

For the data measures of central tendency such as mean , mode etc or dispersion or spread of data such as range , minimum ,maximum ,quartiles , variance etc can be used.
Also frequency distribution tables ,pie charts ,bar charts , box plot , histogram etc can be used.

Bivariate Data

Bivariate contains two different variables.

Deals with cause and consequence relationships and the analysis is done to find out the relationship among two variables.

Temp in deg	20	25	35	43
Ice cream sales	2000	2500	5000	7800

It shows relationship between temperature and sales is directly proportional.

Bivariate Contd…

It involves comparisons ,relationships ,causes and explanations.

These are generally plotted on X and Y axis on the graph for better understanding.

One of the variable is independent (such as

Multivariate Data

It contains three or more variables.

For example research work of a doctor who has collected data on eating habits of the participants such as meat consumption , dairy products and chocolate consumed per week and their respective cholesterol ,blood pressure and weight data.

Techniques used are regression analysis ,

Case study 1- Display of categorical data

A Survey was conducted with a group of 20 persons. They were asked to inform about their hair and eye color.

A 2 way co[illegible]med as under-

Hair color	Eye Color				
	Blue	Green	Brown	Black	Total
Blonde	**2**	**1**	**2**	**1**	**6**
Red	**1**	**1**	**2**	**0**	**4**
Brown	**1**	**0**	**4**	**2**	**7**
Black	**1**	**0**	**2**	**0**	**3**

Questions

1. How many people have Brown eye color?
   10
2. How many people have Blonde hair?
   6
3. How many Brown haired people have Black eyes?
   2
4. What is the percentage of people with Green eyes?
   10
5. What percentage of people have red hair and Blue eyes?
   5

Case Study 2- Display of numerical data

Following data shows the experiment data to study the effect of different amounts of water on the germination of seeds.

For each amount of water,4 identical boxes were sown with 100 seeds each and the number of seeds having germinated after 2 weeks was recorded.

The experiment was repeated with the boxes covered to slow the evaporation.

There were six levels of watering, coded 1 to 6 with higher codes corresponding to more water.

Case Study 2- Display of numerical data Contd…

1. Uncovered boxes

	Amount of water					
	1	**2**	**3**	**4**	**5**	**6**
Number of seeds germinated per box	**2** **2**	**4** **1**	**6** **6**	**82**	**79**	**0**
	2 **5**	**4** **6**	**7** **2**	**73**	**68**	**0**

Case Study 2- Display of numerical data Contd…

2. Covered boxes

	Amount of water					
	1	**2**	**3**	**4**	**5**	**6**
Number of seeds germinated per box	**4** **5**	**6** **5**	**8** **1**	**55**	**31**	**0**
	4 **1**	**8** **0**	**7** **3**	**51**	**36**	**0**
	4	**7**	**7**	**49**	**45**	**0**

Questions

1. What is the average number of seeds germinated for the uncovered boxes with level of watering equal to 4?
   78
2. What is the median value for the data covered boxes?
   45
3. Establish conclusions on the basis of available data:
   a. Association of levels of watering with the number of germinating seeds in case of covered boxes as well as uncovered boxes.
   b. Association of number of germinating seeds with the fact that the boxes were covered or uncovered.

Case study 3- Numerical data display.

The data gives the production of wheat in years 2015 and 2016 for top 12 wheat producing states in India:

..\RData sets\wheat_box.xlsx

Question:

Plot the box plots for the wheat produce as applicable for 2015 and 2016.

To be carried out in Lab1

Case study 4

The data set is a part of a larger data set. The small dataset we consider , gives measurements of two flower parts sepal length and sepal width in cms on 50 specimens of each of two species of the iris flower , namely Iris setosa and Iris Virginica.

..\RData sets\iris data.xlsx

Question:

Plot a scatter plot with sepal length on X-axis and Sepal width on Y axis. The scatter plot should be a common single plot.

To be carried out in Lab1

Raw Data

- It is also called as source data or atomic data.
- It is unprocessed data.
- It is hard to parse or analyze.
- This information may be stored in file or it can be mere collection of numbers , characters etc.
- This data can be entered in computer or it can be generated by the computer.
- It is hard to parse or analyze.
- Data analysis includes the processing or cleaning of data.
- Data scientist’s job is to do processing of such data.
- The processing of raw data has to be carried out more than once, this record should be maintained.

Processed Data

- It is the data which is ready for analysis.
- Some kind of cleaning, transformation is performed on raw data to get processed data which can be analyzed and visualized.
- *For example*: Voltage signal of microphone is raw data and when the signal is filtered and noise is removed, it is processed data.
- Processing can include merging, subsetting, transforming etc.
- Depending on the field of work there can be standards of processing.
- It is important that all the steps should be recorded.

Raw data to Processed Data

Sources of Raw data

- A binary file generated by a machine
- Unformatted excel file
- JSON from Twitter API
- Hand entered No.s (Readings) you collected

Typical Attributes of Raw data

- No software is run on the data.
- No manipulation of any of the numbers in the data has taken place.
- No data is removed from the data set.
- The data set is not summarized in any possible way.

Expected Attributes of Processed data

- Each variable to be measured should be in one column.
- Each different observation of that variable should be in a different row.
- There should be one table for each kind of variable.
- There should be one table for each kind of variable. e.g. data collected from Twitter should be kept in a separate table & the data from Face book in another table etc.
- If you end up with multiple tables , they should include a column in the table that allows them to be linked. This is important for merging the dataset.

Minor Attributes:

- Include a row at the top of each file with variable names.
- Assign ’easily perceivable’ variables names. e.g. If a table is reporting User Experience as excellent, good, fair, poor etc. The variable name may be ‘user_exp’ instead of ‘ux’ etc.

Code Book or Meta data

It should contain following information about the variables:

- Units- whether the unit of column is in Rs in lacs or thousands.
- Summary choices-whether mean or median is used.
- Information about resource of data- which data base is used or structured survey etc is used.
- Valid link to the data base should be given.
- For structured survey- what was the population, whether it was observational or experimental design , selection of samples , confounding variables information biases if any , mathematical formulation should be mentioned in the code book.

1. Which of the following is an example of raw data?

a) original swath files generated from a sonar system
b) initial time-series file of temperature values
c) a real-time GPS-encoded navigation file
d) all of the mentioned

Answer: d : Raw data refers to data that have not been changed since acquisition.

2. Which of the following data is put into a formula to produce commonly accepted
results?
a) Raw
b) Processed
c) Synchronized
d) All of the Mentioned
Asnwer : b

3. Which of the following is another name for raw data?
a) destination data
b) eggy data
c) secondary
d) machine learning

Case Study

- “Growth in a Time of Debt”, Reinhart C. and Rogoff K , American Economic Review: Papers and Proceedings , Vol- 100.
- Another Economist “Thomas Herndon” got hold of raw excel file and metadata and proved that selective exclusion of available data and on conventional weighting of summary statistics lead to serious errors.
- Hence the representation of relationship between public debt and GDP is in accurate.
- “Does High public debt consistently stifle economic growth? A critique of Reinhart and Rogoff” , Herndon T, Ash m, Pollin r, Peri working paper series , no.32, April 2013.
- Case study highlights the importance of metadata in data processing and more importantly ethics in data science.

What is metadata & what is it used for?

What is metadata?

“Metadata creation = the art formerly known as cataloguing”?

delivery of resources by resource creators/owners
rather than (or as well as) by intermediary

remote access to resources for all
(potentially…)

emphasis on customer/user

information overload
quantity vs. quality?

the Google effect

What is metadata? (2)

“Data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics. A user might be a program or a person.”

Dempsey and Heery, 1998

“Machine understandable information about web resources or other things.”

Berners-Lee, 1997

Structured data about resources that can be used to help support a wide range of operations

Who/what uses metadata?

- Human agent
  - owner managing resources
  - researcher seeking resources
  - third party services

- Software agents
  - aggregators
  - “portals” presenting “landscape” to user
  - “brokers” performing query tasks on behalf of user

What resources, objects, things?

HTML documents
digital images
databases
books
museum objects
archival records

collections
services
physical places
People
concepts
events

What operations?

Different “flavours” of metadata serve different purposes
- simple, generic vs. rich, specific
- published widely vs. shared within community vs. used by resource owner/manager

Owner / manager / provider wants to
- establish control of resources
- administer/manage resources (through time)
- disclose/promote resources
- enable and control access/use of resources
- contextualise resources

What operations? (2)

End user wants to
- find
- identify
- select
- obtain/use
- Interpret

What information required in metadata?

No “one size fits all” solution

Depends on operation which metadata supports

Refer to **standards**:

- Benefit of others’ experience, expertise
- Provide basis for good practice
- Reflect consensus, so facilitate exchange, access, interoperability
- May have support in software tools

“Administrative” metadata

Extraction of data

- Real life situations offer data in a far from easier format.
- Data present in a file as line would be embedded with headers, strings , start , stop instructions etc. We should get the raw files figure out its structure and extract the relevant bits.
- Another challenge is data may be well organized but may be in some different format which is difficult to analyze. Hence it should be first converted to the format which can be easily analyzed- API of Twitter in JSON format.

Extraction of data Contd....

- Sometimes data is available in the form of free text. This data may be pliable with human intelligence but can be a challenge as a task.— Doctor's Prescription. It contains Dr's name registration no, etc.
- Sometimes data would be given by two different sources for combining and joint processing.
- MySQL and MongoDB are free popular free databases.
- Hence the knowledge of different formats is also necessary. types of

Data Formats

1. XML Format: Extensible Markup Language.
2. JSON Format: Java Script Object Notation.
3. MySQL Format:
4. HDF5 Format: Hierarchical Data Format Version 5.

XML - eXtensible Markup Language

- Designed to store and transport data.
- Used in internet applications.
- For Exchanging the data between incompatible systems or upgrading systems , data must be converted properly. Sometimes incompatible data is lost.
- XML stores data in plain text format which provides software and hardware independence of storing and transporting data.

XML Contd...

- XML also makes it easier to expand and upgrade new operating systems, new applications or new browsers without losing data

- It has two main parts:

1. Markup- labels that give the text structure
2. Content- the actual text of document.

   Also used for web scraping.

Example of XML script

```
<productListing title="My Products">
  <product>
    <name>Product One</name>
    <description>
      Product One is an exciting new widget that will simplify your life.
    </description>
    <cost>$19.95</cost>
    <shipping>$2.95</shipping>
  </product>
  <product>
    <name>Product Two</name>
  </product>
  <product>
    <name>Product Three</name>
    <description>
      This is such a terrific widget that you will most certainly want to buy one for your home and another one for your office!
    </description>
    <cost>$24.95</cost>
    <shipping>$0.00</shipping>
  </product>
</productListing>
```

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

JSON - JavaScript Object Notation

- Light weight (uses less words) format for storing and transporting.
- It is often used when data is sent from server to web page.
- It is self describing and easy to understand.
- It is language independent.
- It uses java script syntax but JSON format is text only.
- It can be read and used as a data format by any programming language.

If the data is mere plain text line like” Hello world” it can be easily send.
But assume your data contains the details of 100 employees with their attributes like employee, employee name, salary,department,age, etc. in java object. It becomes difficult transfer.
Hence you can convert this to JSON and send it to client.
At client side it can be converted back to java object.
GSON, JACKSON like libraries can be used to do this job.

XML VS JSON

Similarities:

- Both are self describing, human readable
- Both are Hierarchical i.e have values within values.
- Both can be parsed and used by lots of programming languages.

Differences:

- JSON is shorter and quicker to read and write.
- JSON can use arrays.

XML

```
{"employees":[
  { "firstName":"John", "lastName":"Doe" },
  { "firstName":"Anna", "lastName":"Smith" },
  { "firstName":"Peter", "lastName":"Jones" }
]}
```

JSON

```
<employees>
  <employee>
    <firstName>John</firstName> <lastName>Doe</lastName>
  </employee>
  <employee>
    <firstName>Anna</firstName> <lastName>Smith</lastName>
  </employee>
  <employee>
    <firstName>Peter</firstName> <lastName>Jones</lastName>
  </employee>
</employees>
```

MySQL

- Database system used on the web.
- It runs on the server.
- Ideal for both small and large applications.
- It is very fast, reliable and easy to use.
- It compiles on a number of Platforms.
- Data is stored in tables, which is a collection of related data.
- A query is a question or a request.
- We can query a database for specific information and have a record set returned.

HDF5 - Hierarchical Data Format

- Open source file format, supports large, complex , heterogenous data.
- Uses file directory like structure that allows to organize data within the file in many structured ways.
- Allows embedding of meta data which is self describing. It means that each file , group and dataset can have associated metadata that describes exactly what the data are.
- This feature facilitates automation without a separate metadata document.
- With language like R or Python we can get information from metadata that are associated with dataset.

HDF5 Contd...

- Two important terms used for HDF5 are

  1. Group : A folder like element within an HDF5 file that might contain other groups or datasets within it.

  2. Dataset: The actual data contained within the HDF5 file. Data sets are often stored within groups in the file.

- It is a compressed format as size of all data contained within HDF5 is optimized.
- However, it may contain big data and can be large in size.
- Powerful attribute is “data slicing”, extracting particular subset from data for processing.
- Hence entire data set need not be in RAM.

HDF5 Contd..

- It may contain different types of data sets within same file like text data set, numeric data set or a data set containing both the data types i.e heterogeneous data set.

Data base and DBMS

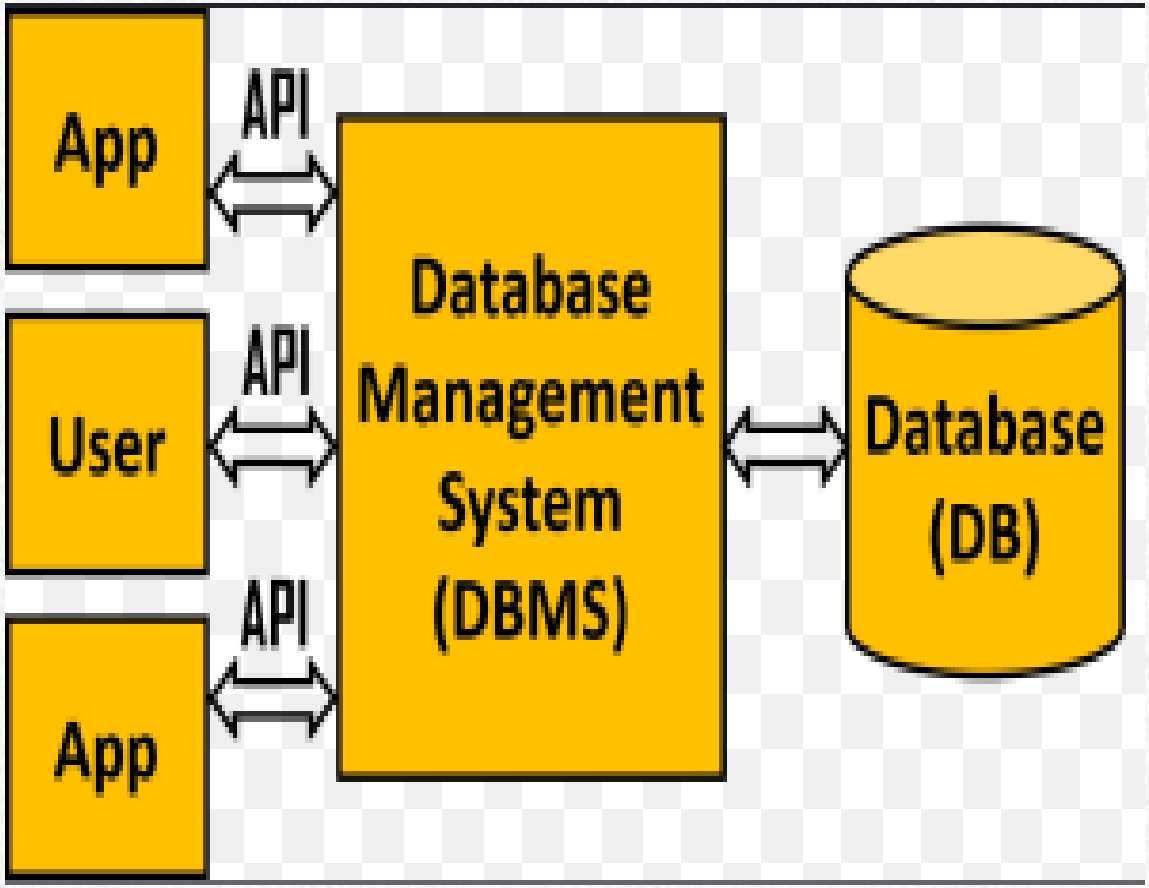
It is a collection of logically related data. Data base management system is a software (DBMS) which manages this data

It is a layer between programs and data

It makes it possible for user to create , read, update and delete data

It provides features for data like data retrieval, data integrity , data modification , data security

DBMS


App
API
User
API
App
API
Database
Management
System
(DBMS)
Database
(DB)

Advantages of DBMS

Reducing data redundancy: Being a single entity any change in it is reflected immediately. Because of this , there is no chance of encountering multiple data or duplicate data.

Sharing of Data: In a database, users of the database can share data amongst themselves. There are various levels of authorization to access the data, and consequently the data can only be shared based on correct authorization protocols followed. Remote users can also simultaneously access and share the data between themselves.

Data Integrity: It mean data is accurate and consistent.

Data Security: Only authorized users are allowed to access the data.

Backup and recovery: DBMS takes care of backup and recovery. It also restores the database after a crash or system failure to its previous condition.

Structured Query Language

SQL is a structured language for accessing and manipulating databases.

SQL can execute queries against a database. It can retrieve data from a database.

It can insert, update or delete records from a database

SQL can create new databases, or

Data Cleaning

- It is defined as the process to ensure the correctness , consistency and usability of the data.

- For any type of data, data quality is very important.

Meaning of Data Quality

Generally, you have a problem if the data doesn’t mean what you think it does, or should

- Data not up to spec : garbage in, glitches, etc.
- You don’t understand the spec : complexity, lack of metadata.

Many sources and manifestations

Data quality problems are expensive and pervasive

- DQ problems cost hundreds of billion $$$ each year.
- Resolving data quality problems is often the biggest effort in a data science study.

Example

T.Das|97336o8327|24.95|Y|-|0.0|1000
Ted J.|973-360-8779|2000|N|M|NY|1000

- Can we interpret the data?
  - What do the fields mean?
  - What is the key? The measures?
- Data glitches
  - Typos, multiple formats, missing / default values
- Metadata and domain expertise
  - Field three is Revenue. In dollars or cents?
  - Field seven is Usage. Is it *censored*?
    - Field 4 is a censored flag. How to handle censored data?

Data Glitches

Systemic changes to data which are external to the recorded process.

- Changes in data layout / data types
  - Integer becomes string, fields swap positions, etc.
- Changes in scale / format
  - Dollars vs. euros
- Temporary reversion to defaults
  - Failure of a processing step
- Gaps in time series
  - Especially when records represent incremental changes.

Conventional Definition of Data Quality

Accuracy
- The data was recorded correctly.
- The degree to which the data is close to the true values. e.g the address of a street is given in valid format but is not true i.e. the address does not exist.

Completeness
- All relevant data was recorded.
- The degree to which all required data is known.

Uniqueness
- Entities are recorded once.

Timeliness
- The data is kept up to date.

Conventional Definition of Data Quality

Consistency

The degree to which data is specified using same unit e.g currency of one country is different from other.

Inconsistency occurs when two values in the data set contradict with each other e.g Boy with age 10 years is defined as senior citizen.

Uniformity

The degree to which data is specified using same unit e.g currency of one country is different from other.

Validity

The degree to which the data conform to defined business rules or constraints. e.g. dates should fall in typical range, certain columns cannot be empty etc.

Finding a modern definition

We need a definition of data quality which

- Reflects the use of the data
- Leads to improvements in processes
- Is measurable (we can define metrics)

First, we need a better understanding of how and where data quality problems occur

- The data quality continuum

The Data Quality Continuum

Data and information is not static, it flows in a data collection and usage process

- Data gathering
- Data delivery
- Data storage
- Data integration
- Data retrieval
- Data mining/analysis

Steps of data cleaning

- Monitor errors
- Standardize processes
- Validate accuracy
- Scrub for duplicate data

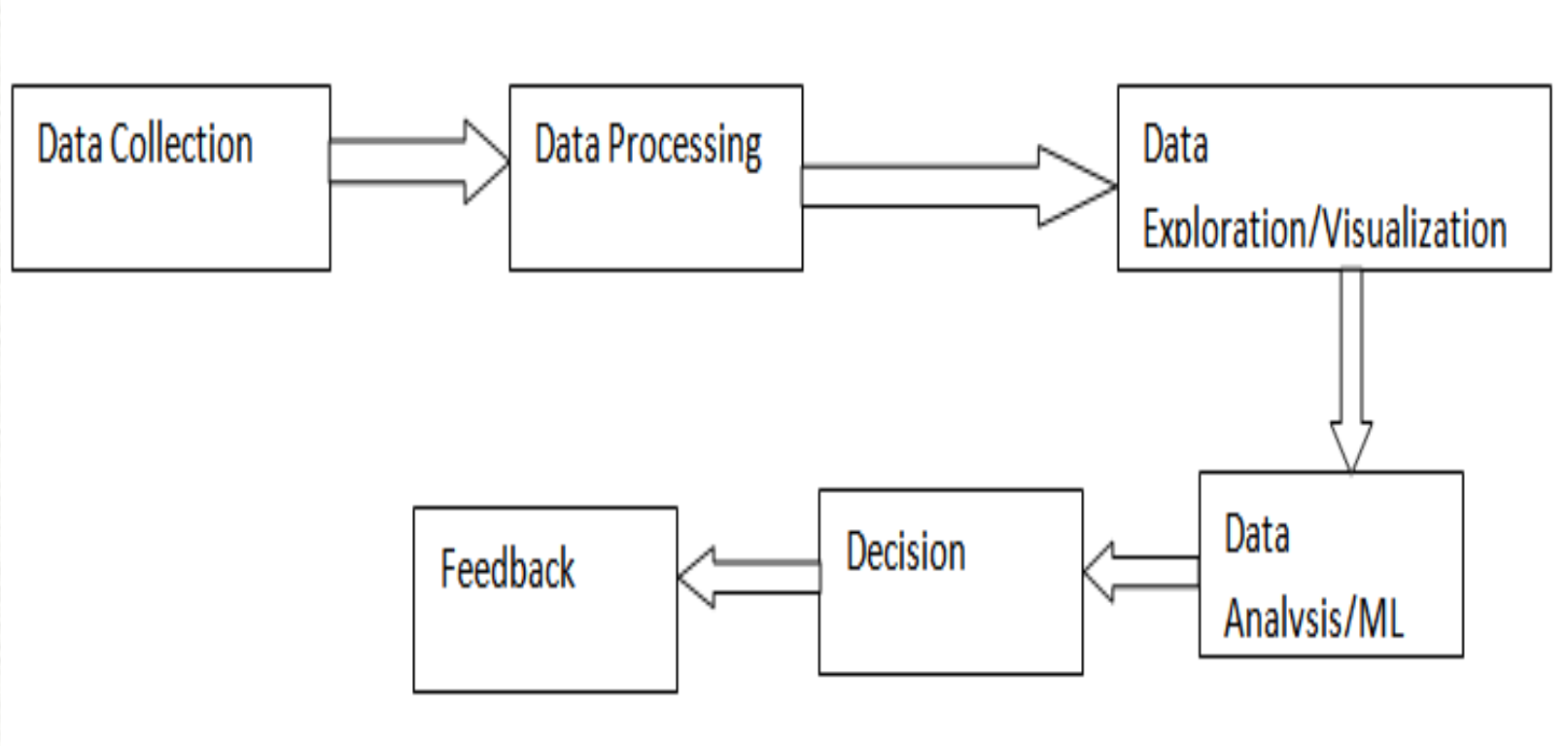
Different Techniques of Data Cleaning

1. Removing Irrelevant Data e.g analyzing health related data phone numbers not required.
2. Removing Duplicates: It happens when data are combined from different sources or say user has submitted submit button twice or request to online booking submitted twice etc.
3. Type Conversion: Ensure numbers are stored as numeric data type and not as string. If the value cannot be converted to any specific type then it should be converted to NA value and warning to be issued.
4. Syntax errors: Remove white spaces at the beginning or ending. “ Hello World “ instead it should be “Hello World”.

Different Techniques of Data Cleaning

5. Padding: Some times to adjust the width of a number zeros are required to be padded e.g 456 to convert it to 6 digits for data uniformity zeros can be padded at the beginning i.e 000456.
6. Fix Typos: let gender variable is actually having two classes as female and male. But if used as female, fem,fe_male,male,M . This should be removed.
7. Standardize: To recognize the typo and to put them in standard format . For example for all strings use only one format i.e either upper case or lower case.
8. Missing values: The data is not recorded. It can be ignored or impute i.e finding that missing data from the available data or copying values from another similar records.

Data Science Pipeline



Data Collection

- All available datasets are gathered from structured/unstructured data sources such as internet , internal/external data bases/third party sources etc.
- Then their data is extracted into a usable format such as CSV,JSON etc.
- Typical skills required are
  - Distributed storage: Hadoop , Apache spark
  - Database Management- MySQL , MongoDB
  - Handling relational data base queries.
  - Retrieving unstructured data- text , audio, video etc.

Data Processing

- Time consuming and laborious.
- The steps which we discussed for data cleaning, to be performed as per the requirement.

- Skills required are
  - Coding Language- R, Python
  - Data Modification Tools-NumPy , R, Pandas
  - Distributed Processing-Hadoop , Map Reduce

Data Exploration/Visualization

- Finds the patterns of the data.
- Different types of visualization and statistical techniques are used.
- Data reveals the trend through different graphs, charts and analysis.
- For understanding the visualization domain expertise is required.
- Skills required are
  - Python utilities-Numpy, Matplotlib, Pandas, Scipy
  - R utilities- GGplot2, Dplyr
  - Statistics- Inferential Statistics
  - Data Visualization- Tableau

Data Analysis/Machine Learning

- Its objective is to do the in-depth analytics , mainly the creation of relevant machine learning models such as predictive model or algorithm development.
- Second objective is it evaluates and refine the model. This involves multiple sessions of evaluation and optimization cycles.
- The accuracy of algorithm to be increased by training it with fresh ingestion of data, minimizing losses etc.

- Skill required are
  - Machine learning-supervised and unsupervised algorithms
  - Mathematics-Algebra , Calculus

Decision

- Interpreting the data is more like communicating the decision to interested parties.
- The objective of this step is to first identify the business insight and correlate it to the decision.
- Involve domain experts in correlating the decisions with business problems
- Domain experts help in visualizing the decisions according to the business dimensions which also assists in communicating facts to a non technical audiences.
- Skill required are
  - Business domain knowledge
  - Data visulisation tools- Tableau,D3.js,seaborn
  - Communication-Presentation , speaking , reporting, writing

Feedback

- It is important to revisit and update the model on a periodic basis, depending on the frequency of new data generation.

- The more that is received , the frequent the feedback is.

- Hence , regardless of use case ,persona ,context or data size , data processing pipeline must connect ,collect ,integrate ,cleanse, prepare, relate ,protect and deliver trusted data at scale and at the speed of business.