Data Science

UNIT - 1

Data Science

→ data can be structured, unstructured or semi-structured.

- Domain of study that deals with vast volumes of data using modern tools and techniques.
- Used to find unseen patterns, derive meaningful information and make business decisions.
- Also known as data-driven science.
- Data used for analysis can be from multiple sources & present in various formats.

⇒ It let's you,

- Find the leading cause of a problem by asking the right question.
- Perform exploratory study of data
- Model the data using various algorithms.
- Communicate & visualize the results via graphs, dashboards etc.

DATA SCIENTIST

- A data scientest is a professional responsible for collecting, analyzing and interpreting extremely large amounts of data.
- A data scientist analyzes business data to extract meaningful insights.
- He solves business problems through series of step →

→ Ask right ques. to understand the problem.

→ Gather data from multiple sources - enterprise data, public data etc.

→ Process raw data & convert it into a format suitable for analysis.

→ Feed the data into the analytic system - ML algrorith or statistical model.

→ Prepare the results & insights to share with the appropriate stakeholders.

Student Choice
Date / /
Page No.

Turning Data into Product

- for most organizations, data science is employed to transform data into value or product.

value as in the form of

↓

improved revenue, reduced cost, improved customer experience etc.

Data Science gives the data collected by an organization a purpose.

Data Science & Big Data Relation

- Big Data → extremely large data sets that require specialized and often innovative techologies & techiques in order to use data efficiently.

data science ~~is used to~~ can be used to extract value from data of all sizes.

Big data ~~uses~~ is useful to data scientists bcz the more data you have, the more parameters you can ~~handle~~ include in a given model.

↑

more is not always better.

Date / /
Page No.

Life cycle of Data Science

1.) Setting Research Goal →

- Understand the problem by performing a study of the business model.
- Prepare a project charter which includes
  - ↳ what you are going to research on
  - ↳ data & resources needed
  - ↳ benefits of the organization from this
  - ↳ timetable.

2.) Retrieving data →

- Collecting data - finding & getting access to data needed in your project.
- This data is- either found within the company or retrieved from third party.

3.) Data preparation →

Checking & remediating data errors, enriching the data with data from other data sources, & transforming it into a suitable format for your models.

Student Choice
Date / /
Page No.

- It consists of a process which have following steps →

→ Data Integration – Resolve any conflicts in the dataset & eliminate redundancies.

→ Data Transformation – Normalize, transform & aggregate data using ETL (extract, transform, load) methods.

→ Data Reduction – Using various strategies, reduce size of the data without impacting the quality or outcome.

→ Data cleaning – Correct inconsistent data by filling out missing values and smoothing out noisy data.

4.) Data exploration →

- The data has to be examined before it is ready to use. The data and its characteristics require inspection.

↳ this is due to the different data types,

↓

nominal & ordinal data,
numerical data,
categoral data

needs different handling ←

- Graph, pie etc plots are done → Visualization process

- Then the descriptive statistics have to be computed in order to extract ~~important~~ features and test important variables.

  ↳ feature is used as it helps the data scientist to pick out the properties that represent the concerned data.

5:) Data Modeling →

- here, dimensions of the data set are minimized

  ↳ as every feature & variable is not necessary for prediction of the results.

- Data Scientists needs to choose the essential properties that will directly aid the prediction of the model.

- Machine learning and statistical ~~techin teni~~ techniques are used to achieve the project goal.

6:) Presentation and automation →

Presenting your results to the stakeholders & industrializing your analysis process for repetitive reuse & integration with other tools.

Student Choice
Date / /
Page No.

UNIT - 4

Types of Data → Structured, unstructured, Semi-Structured

1) Structured data →

- is the data which conforms to a data model
- has a defined structure
- follows a consistent order
- can be easily accessed & used by a person or a computer program.
- stored in a well-defined schemas eg → Database.
- It is generally Tabular with columns & rows that clearly defines its attributes.
- Data is well organised so, Definition, Format & meaning of data is explicitly known.
- Similar entities are grouped together to form relations.
- Data elements are addressable, so efficient to analyse & process.

2.) Unstructured Data →

- Does not conforms to a data model
- not easily identifiable structure such that it can not be used by a computer program easily
- data is not organised in a pre-defined manner or not have any pre-defined data model.

eg → Images, videos, Web pages etc.

- Supports data which lack a proper format or sequence.
- Very flexible due to absence of schema.
- Data is portable.

3.) Semi-Structured Data →

- does not conforms to a data model but has some structure.
- lacks fixed or rigid schema.
- It is the data that does not reside in a rational database but have some organisational properties that makes it easier to analyse.
- With some process, we can store them in a rational database.

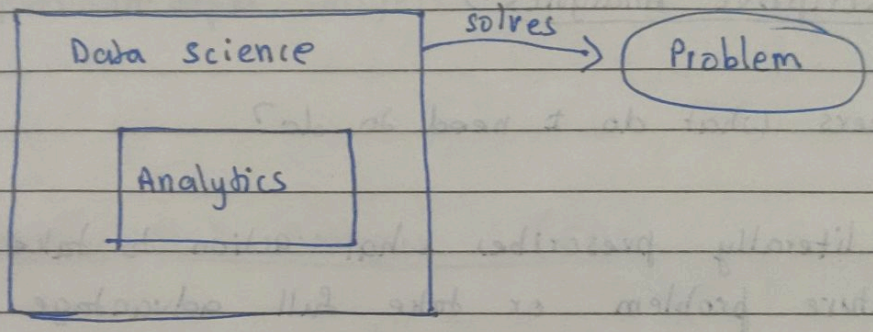
eg → E-mails, XML & other mark up languages.

Student Choice
Date / /
Page No.

Data Science Vs Analytics

data science → It takes output of analytics to solve problems.

Analytics → component of data science, used to understand what an organization's data looks like.



Types of Analytics

1.) Descriptive Analytics →

- Answers the question of what happened?
- It juggles raw data from multiple data sources to give valuable insights into the past.
- These findings signal that something is wrong or right without explaining why.

2.) Predictive Analytics →

- Tells what is likely to happen?
- It uses the findings of descriptive & diagnostic analytics to detect tendencies, clusters and exceptions and to predict future trends, which makes it valuable tool for forecasting.

3.) Prescriptive Analytics →

- Answers what do I need to do?
- It literally prescribes what action to take to eliminate a future problem or take full advantage of a promising trend.
- Requires historical data and external information.
- Uses sophisticated tools & techologies such as ML, business rules & algorithms makes it sophisticated to implement & manage.

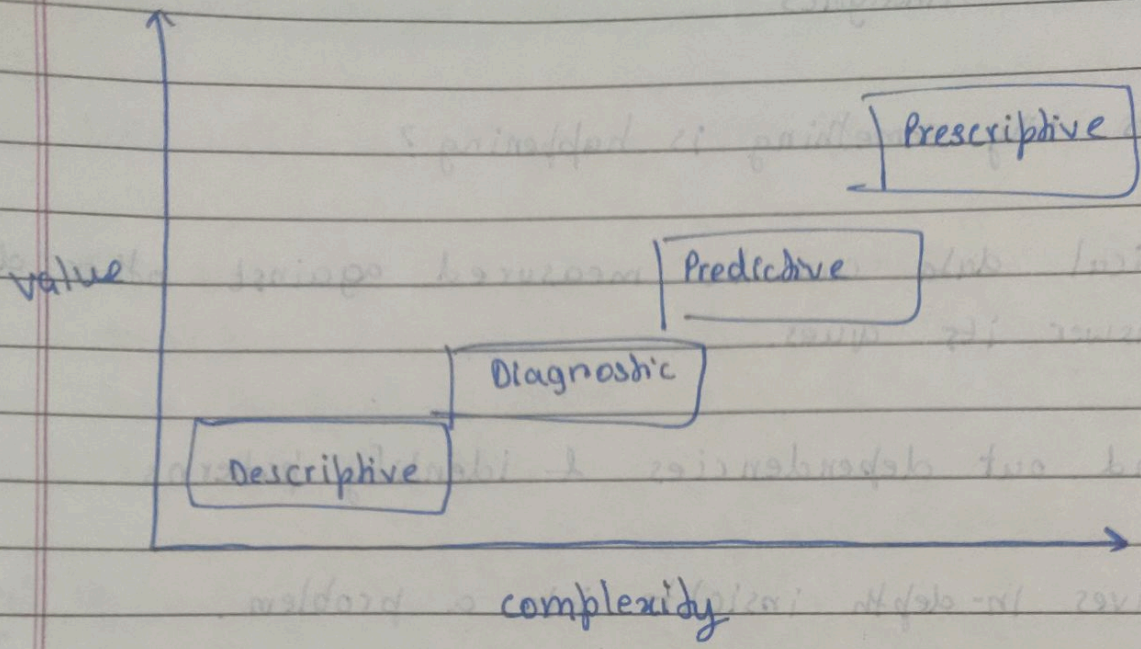
Student Choice
Date / /
Page No.

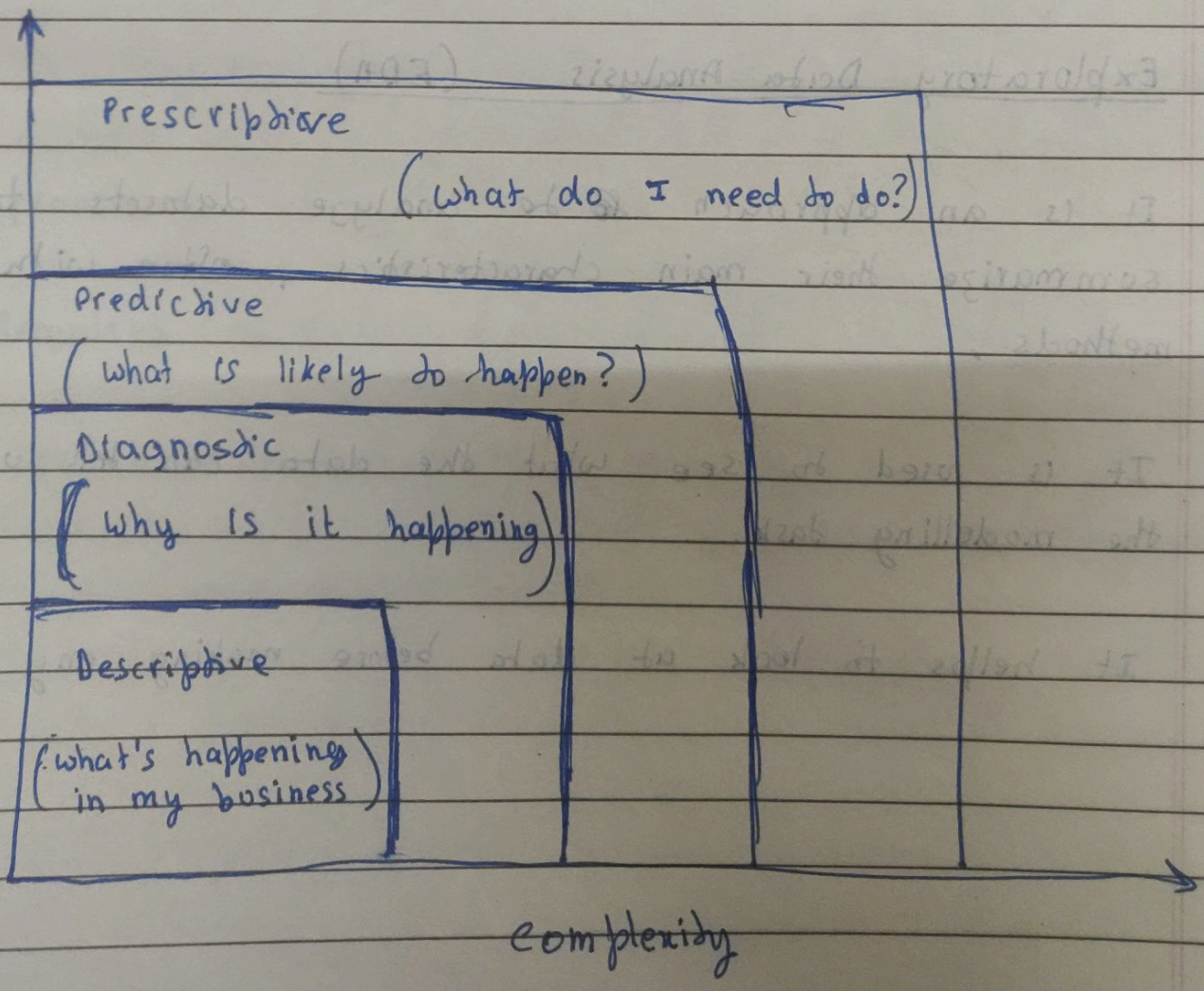
4:) Diagnostic Analytics

- Answers why something is happening?
- Historical data can be measured against other data to answer its ques.
- To find out dependencies & identify patterns
- It gives in-depth insights into a problem.

Exploratory Data Analysis (EDA)

- It is an approach to analyze datasets to summarize their main characteristics, often with usual methods.
- It is used to see what the data can tell us before the modelling task.
- It helps to look at data before making any assumptions.


Prescriptive
Predictive
Diagnostic
Descriptive
value
complexity


Prescriptive
(what do I need to do?)
Predictive
(what is likely to happen?)
Diagnostic
(why is it happening)
Descriptive
(what's happening in my business)
value
complexity