completed.

Page:
Date: / /

Data Science - 1.

Q.1. What is data science? What is its relationship to statistics?

→ - Data science is an interdisciplinary field that uses scientific methods, processes, algorithms & systems to extract knowledge and insights from data in various forms, both structured & unstructured similar to data mining.

- Data science is a concept to unify statistics, data analysis, machine learning & their related methods in order to understand & analyze actual phenomena with data.

- It employs techniques & theories drawn from many fields within the context of mathematics, statistics, info science & computer science.

- Data science is study of the extraction of knowledge from data.

↓

* Relationship to statistics.

- Many critical academics & journalists see no distinction bet data science & statistics.

- Data science is a buzzword without a clear definition and has simply

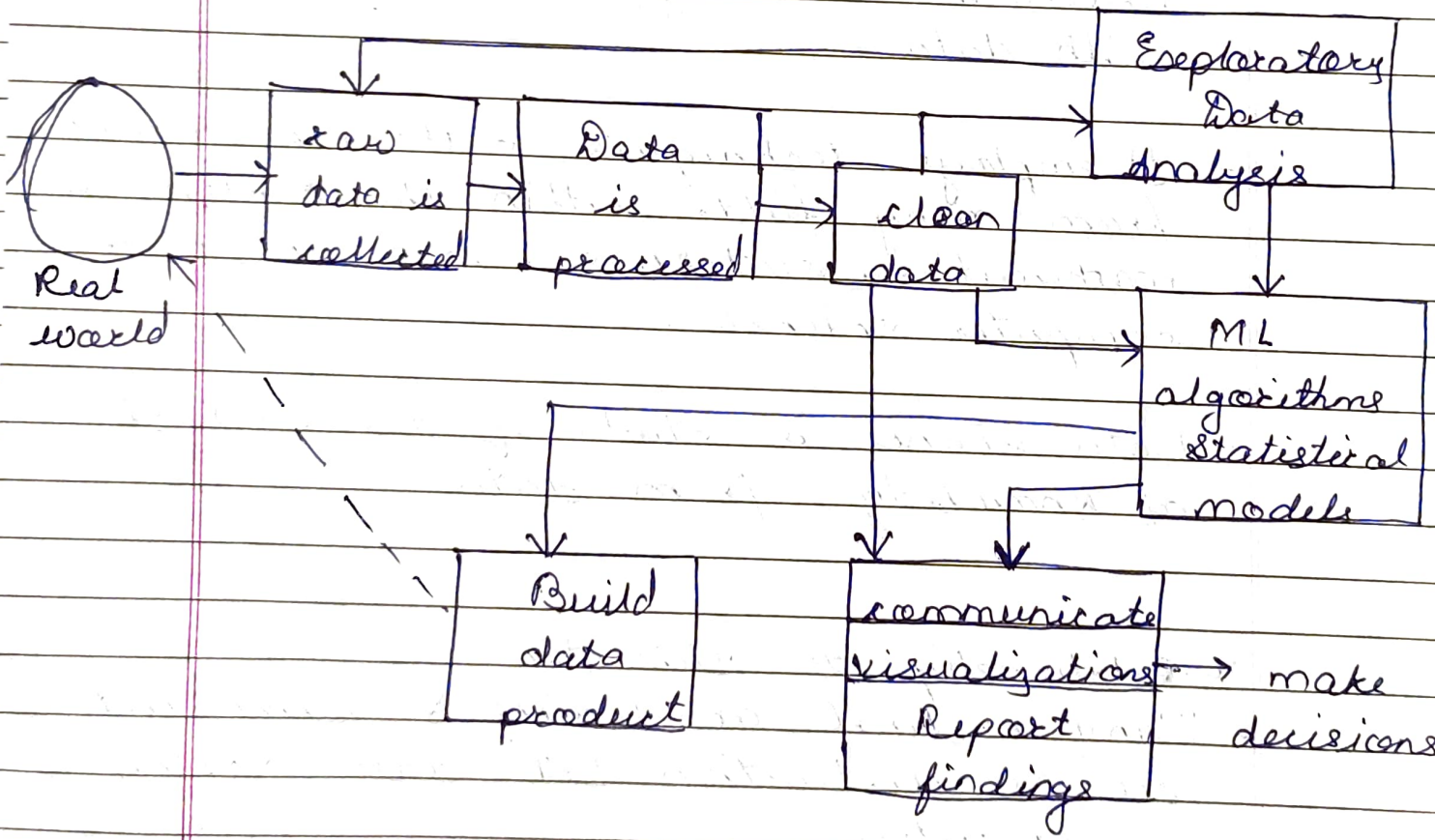
replaced business analytics.

- Data science, like any other interdisciplinary field, employs methodologies and practices from across the academia & industry.

- Data science is different from the existing practice of data analysis.

Q.2. With neat diagram, explain data science process.

→



Page :
Date : / /

3.3. Who is data scientists? What are typical job duties for data scientists?

→

- Data scientists are a new breed of analytical data expert who have the technical skills to solve complex problems - and the curosity to explore what problems need to be solved.

- A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

- Many data scientists began their careers as statisticians or data analysts

- A data scientist can also be divided into different roles based on their skill sets. - data researcher, data developers.

* Typical job duties -

① collecting large amounts of raw data and transforming it into a more usable format.

② solving business related problems using data driven techniques.

③ working with a variety of programming languages - ~~to~~ including R & Python.

(4) staying on top of analytical techniques such as ML, deep learning & text analytics.

(5) communicating & collaborating with IT & business.

(6) looking for order & patterns in data as well as spotting trends.

Q.4. What are applications of data science?

→ (1) internet search
(2) digital advertisements
(3) recommender systems
(4) image recognition
(5) speech recognition
(6) gaming
(7) price comparison websites
(8) airline route planning
(9) fraud & risk detection
(10) delivery logistics
(11) marketing, finance
(12) human resources, health care
(13) govt. policies.

Page :
Date : / /

Q.5. What are top 10 challenges to practicing data science at work?

→ (1) dirty data (36%)
(2) lack of data science talent (30%)
(3) company politics (27%)
(4) lack of clear question (22%)
(5) data inaccessible (22%)
(6) results not used by decision makers (18%)
(7) explaining data science to others (16%)
(8) privacy issues (14%)
(9) lack of domain expertise (14%)
(10) organization small & cannot afford data science team (13%)

Q.6. Compare study of data science with databases. What is the role of SQL in data science?

→

- A database simply is the place where you store the data.

- A data engineer is responsible for setting up & maintaining the infrastructure for database.

- In contrast, a data scientist is not concerned about storing data. His/her job is to actually derive meaningful insights.

- Databases are logical structures that are based on set theory that have

relationships based on unique primary & foreign keys. These links allow processing of data between tables while holding the integrity of data.

- Data science is the study & ~~and~~ analysis of data using methods, processes and insights that corresponds to structured or unstructured data.

- ~~The difference is that~~ databases are actual objects & data science are methods, processes that corresponds to structured or unstructured data.

* Role of SQL in DS.
- Data scientists works with data and all the structured data is stored in databases. So, if one needs to play with data, he must need SQL.

- For querying & manipulating the data we use a language similar to SQL known as Hive QL.

- For creating a table & test environment data scientists use SQL.

- For doing analytics tasks over the data that is stored in oracle DB or SQL server, we use SQL.

- When working with Big Data processing tools, we use SQL for data preparation.

Q.7. Define scientific computing. What are the capabilities of computational scientist? What are applications of scientific computing?

→ * Scientific computing -

- Scientific computing is the science of solving problems with computers.

- also called as computational science.
- It is rapidly growing multidisciplinary field that uses advanced computing capabilities to understand & solve complex problems.

- considered as third mode of science.

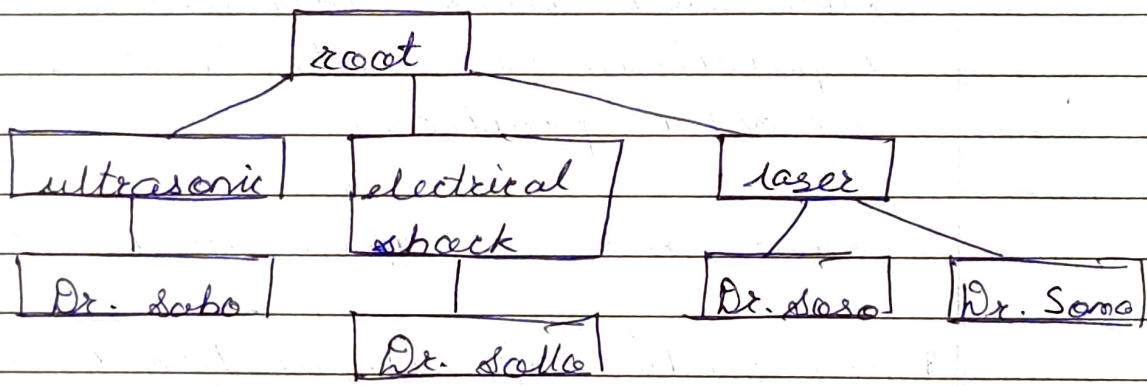
* capabilities of computational scientists -
1. recognizing complex problems.
2. adequately conceptualize the system containing these problems.
3. design a framework of algorithms suitable for studying system.
4. choose a suitable computing infrastructure.
5. adjust the conceptualization of system.

* Applications of computational science -
1. urban complex systems

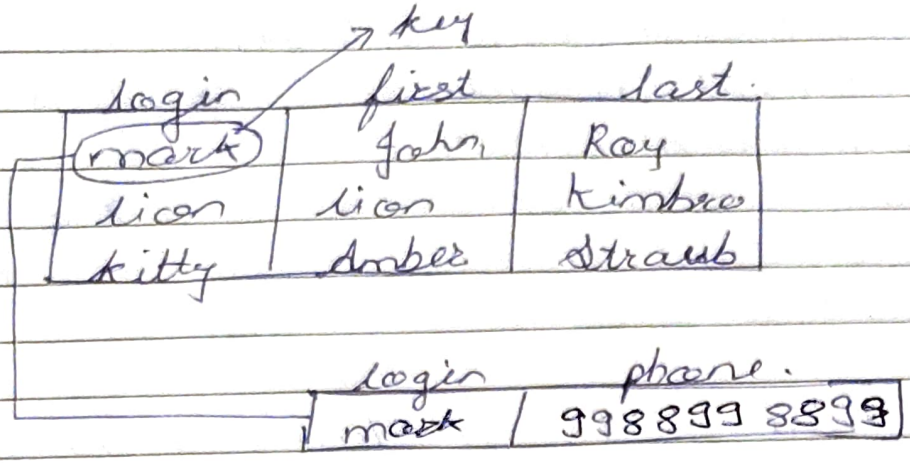
② computational finance
③ computational biology
④ computational science in engineering.

Q.8. Explain data modelling approaches.

→ 1] hierarchical data modelling.
- Array data in treelike, one-to-many arrangements
- Ex - IBM's information management system (IMS)
- Method is common in XML, geographic info systems.

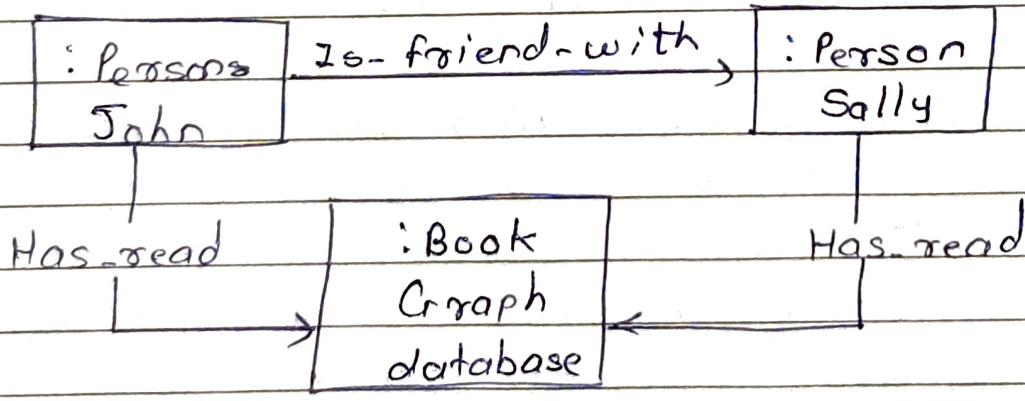


2] Relational data modeling.
- Relational date modeling was first described in 1970.
- Data segments are explicitly joined by use of tables.
- Relational data model was coupled with SQL.



3] Entity relationship model.
- closely integrated with relational data models.
- ER models use diagrams to graphically depict the elements in a database.
- Relationships are visually mapped.

4] graph data models.
- used with graph databases.
- used for describing complex relationships within data sets particularly in social media.



Page :
Date : / /

Q.9. Explain statistical data modelling techniques.

→ - A statistical model is a special class of mathematical model.
- statistical model is non - deterministic.
- The statistical model is the mathematical eq^n that is used.

- statistical data modeling techniques -
  1. linear regression
  2. non - linear regression
  3. logistic regression
  4. multivariate analysis.

Q.10. Explain Bonferroni's principle with suitable example.

→ - Bonferroni's principle is an informal presentation of a statistical theorem that states if your method of finding significant items returns significantly more items that you would expect in the actual population, you can assume most of the items you find with it are bogus.

- This means that an algorithm or method we think is useful for finding a particular set of data actually returns more false positives as it returns large portion of data than should be within that category.

- Applying Bonferroni's principle to an algorithm or system for ~~finding~~ identifying or classifying data gives an upper bound on the accuracy of your methods.

- Bonferroni's principle is a statistical method for accounting for random events.

- Ex - Suppose that evil-doers periodically gather at hotel to plot their evil. We want to detect them. To find evil-doers, we shall look for people who, on two different days, were both at same hotel.

10^9 people being ~~to~~ tracked.
total days = 1000.
Each person stays in a hotel 1% of the time (10 days out of 1000)
Hotels hold 100 people (so 10^5 hotels).

Probability that given persons p & q will be at same hotel on given day d:

$$\frac{1}{100} \times \frac{1}{100} \times 10^{-5} = 10^{-9}$$

Probability that p & q will be at the same hotel on given days d_1 & d_2:

$$10^{-9} \times 10^{-9} = 10^{-18}$$

Pairs of days: $5 \times 10^{+5}$

Page:
Date: / /

Probability that p & q will be at the same hotel on some two days:

$5 \times 10^{5} \times 10^{-18} = 5 \times 10^{-13}$.

Pairs of people: 5×10^{-17}

Expected no. of suspicious pairs of people:

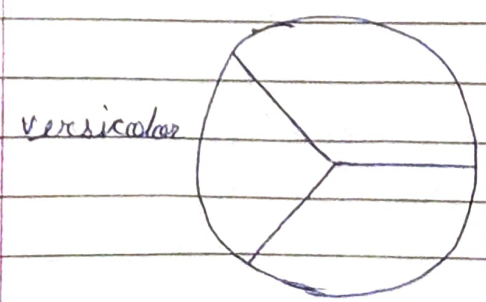
$5 \times 10^{17} \times 5 \times 10^{-13} = 250000$.

Q.11. Explain data visualization techniques available for data scientist.

→ - Data visualization is viewed by many disciplines as a modern equivalent of visual communication.
- It involves the creation & study of the visual representation of data.
- It makes complese data more accessible, understandable & usable.

- Data visualization techniques -

① pie charts -
- one of clearest way to present data.



② Bar charts -
- Same info as pie charts can be conveyed.

③ histograms -
- visualization tool, as it contains interesting information.
- histogram corresponds to classes of real - world entities.

④ Boxplots -
- convenient way to summarize the dataset by showing median quantities, min/max values for each of the variable.

⑤ Scatterplots -
- one of simplest but powerful ways to visualize relationship within dataset.