UNIT–1 : Introduction to Data Science (PYQ Solution)

PART–A Answers (≈70 Words Each)

Q1. What is data science? Name its main components. (Expected)

Data science is an interdisciplinary field that extracts useful knowledge and insight from structured and unstructured data using statistics, programming, machine learning and domain knowledge.

Main components:

- **Statistics and mathematics: probability, inference, linear algebra**
- **Computer science: programming in R or Python, databases, machine learning**
- **Domain knowledge: understanding the business or scientific problem**

Typical output is a report, a prediction model or a data-driven decision.

Q2. What is the digital universe? (Expected)

The digital universe is the total amount of digital data created, captured, replicated and consumed in the world. It grows exponentially, roughly doubling every two years, because of sensors, mobile apps, social media, online transactions, video and IoT devices.

It matters to a data scientist because more data means more opportunity for insight, but also more challenges in storage, quality, privacy and processing. Most of this data is unstructured (text, images, video).

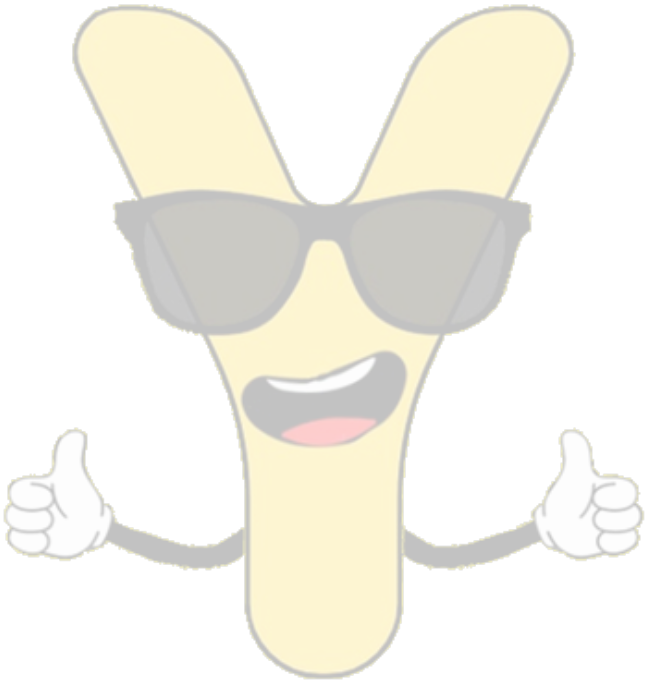
Q3. List four sources of data. What is an information commons? (Expected)

Four sources of data:

1. **Surveys and experiments (primary data)**

2. **Sensors and IoT devices**
3. **Company databases and transactions**
4. **Web, social media and APIs**

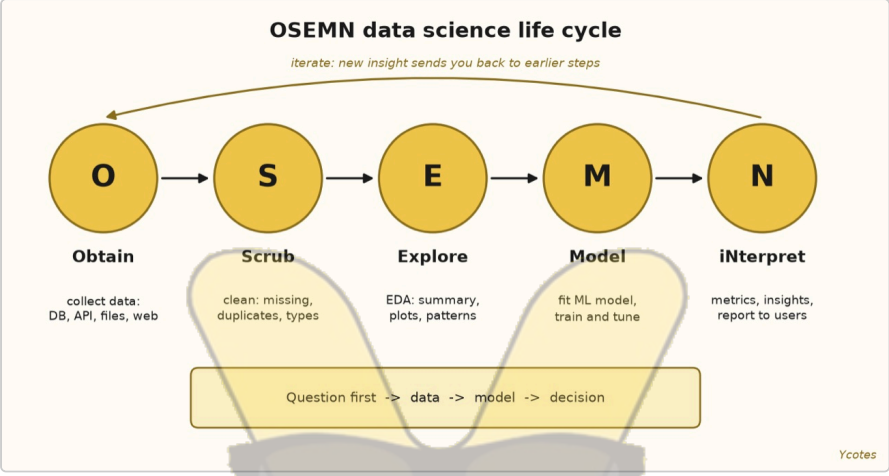
Information commons: a shared, openly accessible pool of data and knowledge that anyone may use and contribute to, such as data.gov.in, Wikipedia, OpenStreetMap and Kaggle datasets. It lowers the cost of getting data for research and learning.



PART–B Answers (≈250 Words Each)

Q1. Explain the OSEMN data science life cycle with a diagram. (Expected)

OSEMN (pronounced "awesome") is a five-step workflow followed in a data science project.



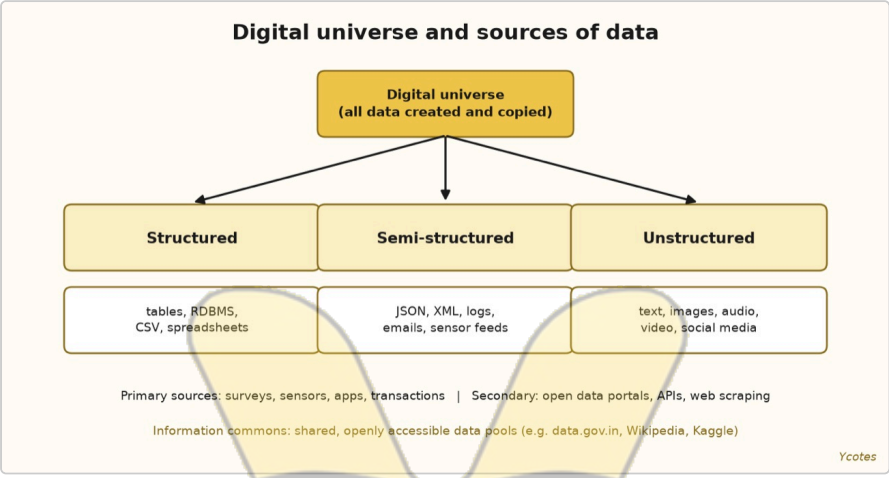
Step	What happens	R example
Obtain	Collect data from files, databases, web pages and APIs	read.csv("data.csv")
Scrub	Clean the data: fix missing values, duplicates, wrong data types and outliers	na.omit(df), as.numeric()
Explore	Summary statistics and plots to understand distributions and relationships	summary(df), hist(), cor()
Model	Choose, train and tune a machine learning model	lm(), rpart(), kmeans()
i**N**terpret	Evaluate with metrics, explain results and present to decision makers	RMSE, confusion matrix, charts

Key points:

- The cycle is **iterative**. Poor results during modelling send us back to scrubbing or to obtaining more data.
- Scrubbing normally takes the largest share of the time (often 60 to 80 percent).
- The final step is not only accuracy; the findings must be explained in simple language so that the business can act on them.

Example: predicting house prices: obtain a listings CSV, scrub missing areas, explore price against size, model with linear regression, interpret with RMSE and a report.

Q2. Explain the sources of data and the types of data (structured, semi-structured, unstructured). (Expected)



Sources of data:

- **Primary sources:** data collected first-hand for the purpose, such as surveys, interviews, experiments, sensors and IoT devices.
- **Secondary sources:** data already collected by others, such as government open-data portals, company databases, research papers, APIs and web scraping.

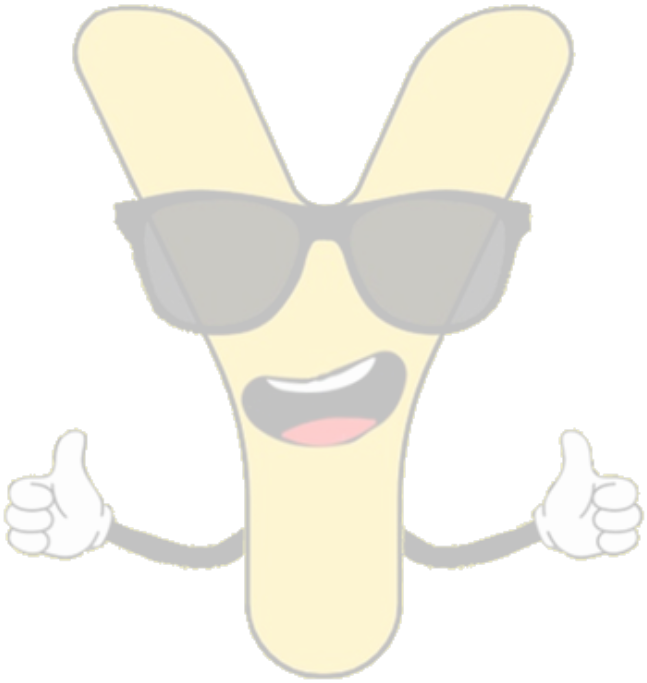
Types of data:

Type	Meaning	Examples
Structured	Fixed schema of rows and columns	RDBMS tables, CSV, Excel
Semi-structured	Has tags or keys but no rigid table	JSON, XML, logs, email
Unstructured	No predefined format	Text, images, audio, video, social media posts

Observations:

- About 80 percent of the digital universe is unstructured, which needs special tools (text mining, image processing).

- Structured data is easy to query with SQL; unstructured data needs preprocessing before modelling.
- Data quality (accuracy, completeness, consistency, timeliness) matters more than volume.



PART–C Answers (≈500 Words Each)

Q1. What is data science? Explain its components, applications and how it differs from data analytics and machine learning. (Expected)

Definition: data science is the practice of collecting, cleaning, analysing and modelling data to extract knowledge and support decisions. It combines **statistics**, **computer science** and **domain expertise**.

Components:

1. **Data collection and storage:** databases, data warehouses, data lakes, big-data platforms.
2. **Data preparation:** cleaning, integration, transformation, feature engineering.
3. **Analysis and statistics:** descriptive statistics, hypothesis testing, correlation.
4. **Machine learning:** regression, classification, clustering, deep learning.
5. **Visualization and communication:** charts, dashboards, storytelling.
6. **Tools:** R, Python, SQL, Tableau, Hadoop, Spark.

Applications:

Domain	Use
Banking	Fraud detection, credit scoring
Healthcare	Disease prediction, medical imaging
E-commerce	Recommendation systems, demand forecasting
Transport	Route optimisation, self-driving vehicles
Agriculture	Crop yield and disease prediction
Social media	Sentiment analysis, spam filtering

Comparison:

	Data science	Data analytics	Machine learning
Aim	Insight and prediction from data	Explain what happened	Learn patterns to predict
Scope	Widest, end-to-end	Narrow, mostly descriptive	A technique inside data science
Output	Models, products, decisions	Reports, dashboards	Trained model
Tools	R, Python, SQL, big data	Excel, SQL, BI tools	Algorithms and libraries

Relationship: analytics is often the Explore step and machine learning the Model step of the OSEMN life cycle, so data science contains both.

Conclusion: data science turns raw data into decisions, and the OSEMN cycle gives a repeatable way to do it.

Q2. Discuss the digital universe, sources of data and the information commons. Why are these important for a data scientist? (Expected)

Digital universe: all digital data created, captured and copied worldwide. It is growing exponentially because of smartphones, social media, sensors, online payments, streaming and IoT. Growth is dominated by unstructured data such as video, images and text.

Sources of data:

Category	Examples
Machine and sensor data	IoT devices, GPS, server logs, CCTV
Transactional data	Bank, retail, railway bookings
Social and web data	Tweets, reviews, clicks, web scraping
Open and government data	Census, data.gov.in, weather data
Surveys and experiments	Questionnaires, clinical trials

Information commons: shared knowledge and data resources that are open to everyone, with community rules for use and contribution. Examples are Wikipedia, OpenStreetMap, Kaggle, UCI Machine Learning Repository and government open-

data portals. They are valuable because they reduce cost, allow reproducible research and let students and start-ups build models without owning large datasets.

Why important for a data scientist:

1. Knowing where data comes from decides its **quality and bias**; a model is only as good as its data.
2. Different sources need different tools: SQL for structured, text mining for unstructured, APIs for web data.
3. Open commons give free benchmarks to compare algorithms.
4. The size of the digital universe requires scalable storage and processing (big-data tools).
5. Legal and ethical aspects such as **privacy, consent and licensing** must be respected when using data.

Challenges: volume, velocity, variety, veracity (the 4 Vs), plus missing values, duplication and security.

Conclusion: a data scientist must understand the digital universe, choose reliable sources and use the information commons responsibly to start the OSEMN cycle correctly.

