Data Science and Analytics

Data science and its process

The notes for the lesson “Data science and its process”: chapters 1.1 to 1.3, with their figures, worked examples, exam traps, quizzes and exam question bank, in one file.

3 chapters | Chapters 1.1 to 1.3

About these notes

These are the notes for the University free n easy video lesson **Data science and its process** on Data Science and Analytics: chapters 1.1 to 1.3, with every figure, worked example, exam trap and quiz, and the long-answer exam questions on these chapters.

1.1 What data science is: need, benefits and uses

Data science is the work of turning data into knowledge and decisions. It joins three skills: programming, to collect and handle data; mathematics and statistics, to find patterns and say how sure we are; and knowledge of the field the data comes from, so that the questions and answers make sense. It exists because organisations now record almost everything they do, and no person can read millions of rows. It is used wherever evidence can improve a decision, from UPI fraud checks to rain forecasts.

Defining data science

A working definition: **data science** is the use of scientific methods, statistics and computing to extract useful knowledge from data and act on it. "Scientific" matters. A data scientist states a question, collects evidence, tests an answer and reports how certain it is, as in a laboratory. The product is not the data or the chart; it is a decision, a forecast or an app feature that is better because of the data.

The three skills are often drawn as overlapping circles, after a well-known diagram by Drew Conway. Each overlap of two skills without the third is something else. Programming plus statistics, without domain knowledge, is machine learning: powerful methods with no feel for which question matters. Statistics plus domain knowledge, without programming, is traditional research: sound, but slow on large data. Programming plus domain knowledge, without

statistics, is the **danger zone**: code that produces confident numbers nobody has tested. Data science is the centre, where all three meet.



Figure 1.1.1 Messaging everyone costs ₹2,00,000 and nets ₹1,00,000; messaging only the model's top 20,000 costs ₹40,000 and nets ₹1,70,000. Those customers order 3.5 times as often.

Data science and its neighbours

Examiners ask how data science differs from the fields around it.

FIELD	MAIN FOCUS	TYPICAL QUESTION
Statistics	Drawing conclusions from samples and measuring uncertainty	Did the new teaching method raise pass rates, or is the rise chance?
Data analytics, business intelligence	Reports and dashboards on what has happened	What were last month's sales by state?
Machine learning	Algorithms that learn patterns from data to predict	Is this email spam?
Artificial intelligence	Machines that act intelligently; machine learning is one route	Can a voice assistant answer in Tamil?
Big data	Data too large or too fast for one machine, and its tools (a later lesson)	How do we store a year of app logs?
Data science	The whole path from question to data to model to decision	Which customers should get this offer, and how much will it earn?

The need for data science

1. **Data is produced everywhere.** Every UPI payment, IRCTC booking, phone call and app tap leaves a record.
2. **Storage and computing became cheap.** Cloud services rent thousands of processors by the hour, so analysing all the data, not a small sample, is affordable.
3. **Free tools.** Python (pandas, scikit-learn) and R put strong methods on every laptop.
4. **Decisions under uncertainty.** Managers must decide before the future is known. Data science puts a number on the risk ("80% chance this loan is repaid").
5. **Competition.** A firm that sets prices and stock from evidence beats one that relies on habit.

Data science tasks fall into a few classes. **Classification** predicts a category (is this payment fraud?). **Regression** predicts a number (the price of a used car). **Clustering** groups similar items without labels (customer segments). **Association analysis** finds items that go together (bread with butter). **Anomaly detection** flags the unusual (a sudden spike in a meter reading). **Time series forecasting** predicts the next values of a sequence (next week's demand). Tasks that learn from past examples with known answers are **supervised**; clustering and association, which have no answers to learn from, are **unsupervised**. A later lesson teaches the algorithms for these tasks.

Benefits and uses

The benefits fall under six heads: better decisions, lower cost, personal service, early warning of risk, new products, and discovery of patterns nobody suspected. The uses, sector by sector:

- **Banking:** fraud scores for each card or UPI payment; credit scores for loans.
- **Retail:** recommendations ("customers who bought this also bought"), demand forecasts for warehouses.
- **Healthcare:** reading X-rays and eye scans, tracking dengue outbreaks.
- **Transport:** travel times in map apps, the chance a waitlisted ticket is confirmed.
- **Energy:** forecasting electricity demand so that a city like Chennai can buy enough power for hot afternoons and avoid power cuts.
- **Agriculture:** rainfall forecasts, crop yield estimates, pest alerts by SMS.
- **Sports:** cricket field placements, bowling plans, auction values.

WORKED EXAMPLE

Question (8 marks). An online grocery app has 1,00,000 customers. Each promotional SMS costs ₹2, and each customer who orders after it brings a margin of ₹150. If the app messages everyone, 2% order. A data science model ranks customers by their chance of ordering; if the app messages only the top 20,000, 7% of them order. (a) Find the net gain (margin minus SMS cost) of each plan. (b) Find the margin earned per rupee spent and the lift of the model. (c) State one more benefit and one risk of the targeted plan.

Solution.

1. Mass plan. Cost = 1,00,000 × ₹2 = ₹2,00,000. Orders = 2% of 1,00,000 = 2,000. Margin = 2,000 × ₹150 = ₹3,00,000. Net gain = ₹3,00,000 − ₹2,00,000 = **₹1,00,000**.
2. Targeted plan. Cost = 20,000 × ₹2 = ₹40,000. Orders = 7% of 20,000 = 1,400. Margin = 1,400 × ₹150 = ₹2,10,000. Net gain = ₹2,10,000 − ₹40,000 = **₹1,70,000**.
3. Margin per rupee: mass = 3,00,000 ÷ 2,00,000 = **₹1.50**; targeted = 2,10,000 ÷ 40,000 = **₹5.25**. Lift = response rate of the targeted group ÷ overall response rate = 7% ÷ 2% = **3.5**: the chosen customers are 3.5 times as likely to order.
4. The targeted plan gets 600 fewer orders but earns ₹70,000 more. Another benefit: 80,000 customers are not disturbed, so fewer of them block the app's messages. Risk: the model learns from past behaviour; if it is wrong or biased, good customers are never contacted. A small random control group, messaged anyway, shows whether the model still works.

EXAM TRAPS

- Data science is not a synonym for big data or machine learning. Machine learning is one tool inside it; big data is a problem of size and speed.
- "Explain the need for data science" wants reasons (volume, cheap computing, free tools, uncertainty, competition), not only a definition.
- For "benefits and uses", give sector-wise examples that name the decision being improved; a list of buzzwords scores low.
- Students swap the two side overlaps in the Venn diagram. Danger zone = programming + domain, no statistics. Traditional research = statistics + domain, no programming.
- Algorithm and model are different words: k-means is an algorithm; the cluster centres it finds for your data are the model.

CHECK YOURSELF

– REMEMBER **Define data science.** 2 marks

The use of scientific methods, statistics and computing to extract useful knowledge from data and act on it: question, evidence, tested answer, decision.

– REMEMBER **Name the three skills in the data science Venn diagram and the three pairwise overlaps.** 3 marks

Programming, maths and statistics, domain knowledge. Programming + statistics = machine learning; statistics + domain = traditional research; programming + domain = danger zone.

– UNDERSTAND **Explain the need for data science with four reasons.** 4 marks

Data is produced in huge volumes by payments, apps and sensors; storage and cloud computing are cheap; free tools exist; decisions must be made under uncertainty and data puts a number on the risk; competitors who use evidence win.

– UNDERSTAND **How does data science differ from statistics and from machine learning?** 4 marks

Statistics focuses on inference and uncertainty from samples; machine learning on algorithms that learn to predict. Data science uses both, adds programming and domain knowledge, and owns the full path from question to decision.

– APPLY **A bank checks 50,000 loan files a month by hand at ₹60 a file. A model clears 70% of files automatically and staff check the rest. Running the model costs ₹2,00,000 a month. Find the monthly saving.** 3 marks

Old cost = 50,000 × 60 = ₹30,00,000. Files left for staff = 30% = 15,000; cost 15,000 × 60 = ₹9,00,000; plus ₹2,00,000 = ₹11,00,000. Saving = **₹19,00,000 a month**.

– APPLY **Name the task class: (i) tomorrow's power demand, (ii) grouping shoppers by habit, (iii) a ₹90,000 UPI payment at 3 a.m. from a new phone, (iv) whether a student will pass.** 4 marks

(i) Time series forecasting (regression). (ii) Clustering. (iii) Anomaly detection, or classification if past fraud labels exist. (iv) Classification.

– ANALYZE **In the worked example, below what SMS cost does the mass plan earn more than the targeted plan?** 4 marks

With cost c rupees per SMS: mass net = 3,00,000 − 1,00,000c; targeted net = 2,10,000 − 20,000c. Equal when 90,000 = 80,000c, so c = **₹1.125**. Below about ₹1.13 per SMS, messaging everyone earns more; the value of a model depends on what each action costs.

BEYOND THE SYLLABUS

In industry the work is split into roles: data analysts build reports, data engineers build the pipelines that move data, data scientists build models, and machine learning engineers put models into apps and keep them healthy (MLOps). The ACM Data Science Task Force has published curriculum guidance for undergraduate data science, and ACM/IEEE CS2023 places much of this content under its Data Management and Artificial Intelligence knowledge areas. Search for "data engineer vs data scientist" and "MLOps". This is not for the exam.

1.2 Facets of data: structured, semi-structured and unstructured

Data comes in different forms, called its **facets**. The most important split is by how much structure it carries. **Structured** data fits a fixed table of rows and columns; **semi-structured** data carries its own labels, as in JSON and XML; **unstructured** data, such as free text, photos and sound, has no predefined model at all. The form decides how you store the data, how you query it and which methods can analyse it, so it is the first thing a data scientist checks.

Data, information and knowledge

Data is raw recorded values with no context: `38.2`. **Information** is data given context and meaning: "the temperature in Chennai at 3 p.m. on 12 May was 38.2 °C". **Knowledge** is a pattern learned from much information that guides action: "when the afternoon temperature passes 38 °C, Chennai's power demand peaks, so buy extra power for those days". Some books add a fourth level, **wisdom** (knowing when to apply the knowledge), and draw the four as the DIKW pyramid. Data science is the climb from the bottom of this pyramid to the top.

The three levels of structure

Structured data follows a **schema**, a fixed description of the columns and their types, decided before any data is stored. Every row of a railway booking table has the same columns: PNR, name, train number, fare. Relational databases (MySQL, Oracle, PostgreSQL), CSV files and spreadsheets hold structured data, and SQL queries it directly.

Semi-structured data has no fixed table, but each value carries a label (a key or a tag) that says what it is. Records may differ: one booking has a meal choice, another does not. The data is **self-describing**. JSON (from web APIs), XML, email headers and many log files are semi-structured. Document databases such as MongoDB store it (a later lesson).

Unstructured data has no predefined data model. A review, a scanned form, a voice note or a CCTV clip has internal pattern (grammar, pixels, sound waves) but no fields a database can query. Analysing it needs extra methods: natural language processing for text, computer vision for images, speech recognition for audio.

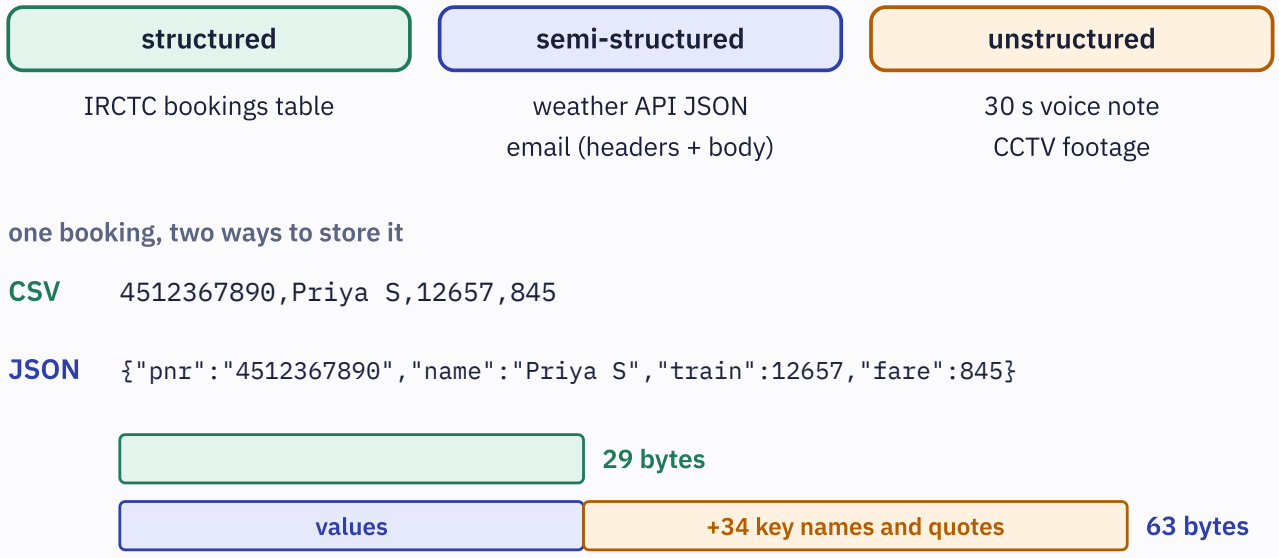


Figure 1.2.1 The five sources sorted by structure, and one booking stored as CSV and as JSON. JSON repeats its 34 bytes of keys and quotes in every record, so 10 lakh bookings take 63 MB against 29 MB.

FEATURE	STRUCTURED	SEMI-STRUCTURED	UNSTRUCTURED
Schema	Fixed, decided before storing	Flexible; carried in tags or keys	None predefined
Stored in	Relational database, CSV	Document store, JSON or XML files	File systems, object stores, data lakes

FEATURE	STRUCTURED	SEMI-STRUCTURED	UNSTRUCTURED
Queried with	SQL	JSON paths, XPath, document queries	Must be processed first (NLP, vision)
Example	Marks register	Weather API reply	Lecture video

The seven facets in the textbook

Cielen, Meysman and Ali's *Introducing Data Science* names seven facets. Each one maps onto the three levels above:

- **Structured:** tables with a schema.
- **Unstructured:** content with no model, such as free-form documents.
- **Natural language:** a special, hard kind of unstructured data (reviews, chats, tweets). Meaning depends on context, sarcasm and language; Indian text mixes Hindi, Tamil and English in one sentence.
- **Machine-generated:** data created by computers and devices without a person typing: server logs, call records, smart-meter readings. Huge in volume, usually structured or semi-structured.
- **Graph-based (network):** things joined by relationships, stored as nodes and edges: friends on a social network, stations on a metro map. Graph databases query it.
- **Audio, image and video:** unstructured media that need speech and vision models, such as reading number plates from traffic cameras.
- **Streaming:** data that arrives continuously and is processed as it flows, such as live cricket scores or a stream of UPI payments. Streaming describes how data arrives, not its structure.

Reading semi-structured data in Python

The Python program below reads one booking in JSON and the same booking in XML, using only the standard library. In XML, values sit inside named tags (`<name>`) or in attributes (`train="12657"`).

```
import json
import xml.etree.ElementTree as ET

js = '{"train": 12657, "name": "Priya", "fare": 845, "meals": ["veg"]}'
rec = json.loads(js)                  # JSON text -> Python dict
print(rec["name"], rec["fare"] + 25, rec["meals"][0])

xml = """<booking train="12657">
  <name>Priya</name>
  <fare currency="INR">845</fare>
</booking>"""
root = ET.fromstring(xml)             # XML text -> element tree
```

```
fare = root.find("fare")
print(root.tag, root.get("train"), root.find("name").text,
      fare.get("currency"), int(fare.text) + 25)
```

Output:

```
Priya 870 veg
booking 12657 Priya INR 870
```

JSON numbers arrive as numbers, so `845 + 25` works at once. XML text always arrives as a string, so the program converts it with `int()` first.

WORKED EXAMPLE

Question (6 marks). (a) Classify each as structured, semi-structured or unstructured: (i) the IRCTC bookings table in a relational database, (ii) a weather API reply in JSON, (iii) a 30-second voice note, (iv) an email, (v) CCTV footage. (b) A booking is stored as the CSV line `4512367890,Priya S,12657,845` or as the JSON line `{"pnr":"4512367890","name":"Priya S","train":12657,"fare":845}`. Each line ends with one newline byte, and each character takes one byte. Find the storage for 10 lakh (1,000,000) bookings in each form and explain the difference.

Solution.

1. (a) (i) Structured: fixed columns and a schema. (ii) Semi-structured: keys label each value. (iii) Unstructured: audio. (iv) Semi-structured: the headers (From, To, Date, Subject) are labelled fields, the body is free text. (v) Unstructured video; streaming if watched live.
2. (b) CSV line: count the characters. 10 (PNR) + 1 + 7 ("Priya S") + 1 + 5 (train) + 1 + 3 (fare) = 28, plus the newline = **29 bytes**.
3. JSON line: the same values plus the keys, quotes, colons and braces = 62 characters, plus the newline = **63 bytes**.
4. For 1,000,000 bookings: CSV = 29,000,000 bytes = **29 MB**; JSON = 63,000,000 bytes = **63 MB**. Ratio = 63 ÷ 29 ≈ **2.17**.
5. The extra 34 bytes per booking are the key names and quotes, repeated in every record. This is the price of being self-describing: the CSV states its columns once (in a header line or a schema), JSON states them every time. Compression and column formats such as Parquet remove most of this overhead.

EXAM TRAPS

- Semi-structured is not "partly unstructured". It is data whose structure travels with it in tags or keys instead of a fixed schema.
- Unstructured does not mean "no pattern at all". A photo is a grid of pixels; it means no predefined model a database can query.
- Machine-generated and streaming are facets of origin and arrival, not levels of structure. A sensor stream can be perfectly structured.
- Email is the examiner's favourite borderline case. Answer "semi-structured" and say why: labelled headers, free-text body.
- Data, information and knowledge: a bare number is data; once it has context it is information.

CHECK YOURSELF

– REMEMBER **List the seven facets of data in Cielen et al.** 3 marks

Structured; unstructured; natural language; machine-generated; graph-based (network); audio, image and video; streaming.

– REMEMBER **What is a schema?** 1 mark

A fixed description of the columns (fields) of a dataset and their types, decided before the data is stored.

– UNDERSTAND **Distinguish data, information and knowledge with one example.** 4 marks

Data: raw values (38.2). Information: data with context (Chennai, 3 p.m., 12 May, 38.2 °C). Knowledge: a pattern that guides action (above 38 °C, demand peaks; buy extra power).

– UNDERSTAND **Why is natural language harder to analyse than other unstructured data?** 3 marks

Meaning depends on context, word order, sarcasm and domain; the same word can mean different things; Indian text often mixes languages and scripts. Models must learn all of this from many examples.

– APPLY **Classify: a smart-meter reading table, a tweet, a hospital discharge summary in XML, an X-ray image.** 4 marks

Structured (machine-generated); unstructured (natural language); semi-structured (tagged); unstructured (image).

– APPLY **In the worked example, how many bytes would 5 lakh bookings take in JSON, and how many bytes do the keys add in total?** 2 marks

JSON: 5,00,000 × 63 = **31,500,000 bytes** (31.5 MB). Extra over CSV: 5,00,000 × 34 = **17,000,000 bytes** (17 MB).

– ANALYZE **A food-delivery app keeps orders, menus that differ by restaurant, customer reviews and rider GPS pings. Suggest a form and a store for each.** 6 marks

Orders: structured, relational database (fixed fields, money needs transactions). Menus: semi-structured JSON in a document store (fields vary). Reviews: unstructured text in a data lake or search index, analysed with NLP. GPS pings: machine-generated stream, structured rows in a time-series store.

BEYOND THE SYLLABUS

Modern platforms keep all three forms side by side in a data lake or "lakehouse", reading structure only when a query needs it (schema-on-read). Text, images and audio are now often turned into embeddings, lists of numbers that capture meaning, and stored in vector databases so that similar items can be found fast. Search for "Apache Parquet", "schema-on-read" and "vector database". This is not for the exam.

1.3 The data science process: goal to presentation

The data science process is the sequence of steps that takes a project from a question to a working result: set the research goal, retrieve the data, prepare it, explore it, build models, and present and automate the result. Following it stops a team from jumping to models before the question is clear or the data is clean.

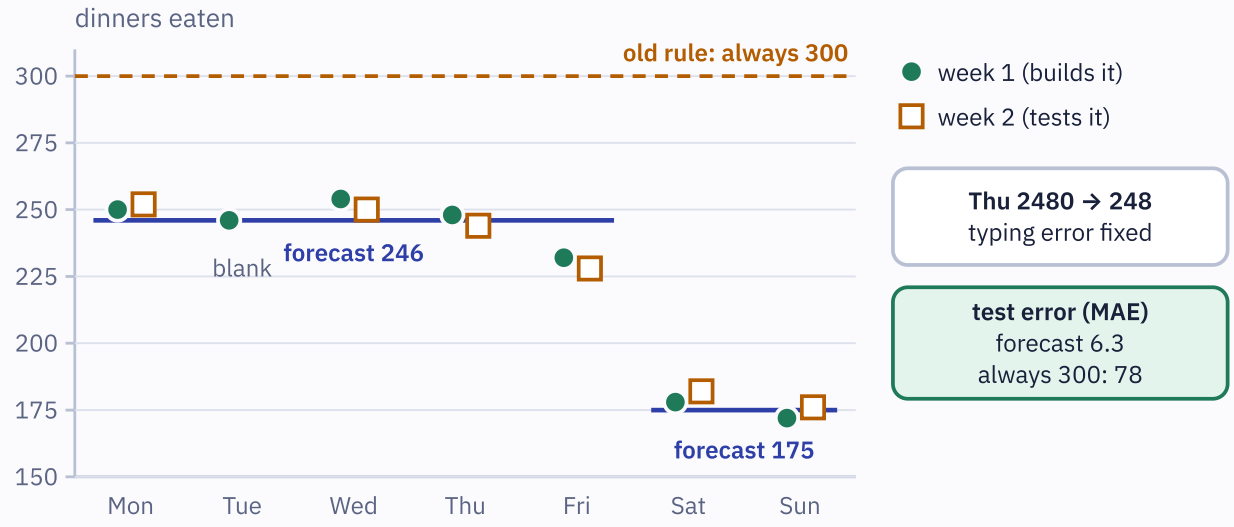
You need: the forms data takes (Chapter 1.2).

The six steps

1. **Set the research goal.** Agree with the people who need the result (the **stakeholders**) on what is to be found, why, and how success will be measured. Write it in a **project charter**: the research goal, the mission and context, how the analysis will be done, the resources

needed, proof that it is achievable, the deliverables with a measure of success, and a timeline. "Cut food waste in the hostel mess by 15% this term" is a goal; "use AI in the mess" is not.

2. **Retrieve data.** Find the data and get access to it. Look inside the organisation first: databases, data warehouses (cleaned data kept for reporting), data marts (a warehouse slice for one department) and data lakes (raw data in any form). Then look outside: open data, web APIs, purchased data (Chapter 1.5). Check quality as soon as it arrives.
3. **Prepare data.** Three sub-steps. **Cleansing** removes errors: typing mistakes, impossible values (an age of 250), extra spaces, mixed codes ("M", "Male"), mixed units (grams and kilograms), duplicates, missing values and outliers. **Integrating** combines sources: joining two tables on a common key, or appending tables with the same columns. **Transforming** reshapes variables for modelling: new columns, log scales, dummy (0/1) variables for categories, fewer variables. A later lesson teaches the tools.
4. **Explore data.** Exploratory data analysis (EDA) looks at the data through summaries and plots (histograms, box plots, scatter plots) to see distributions, relationships and oddities. It often finds errors that step 3 missed. A later lesson covers it.
5. **Build the model.** Choose a model type and its input variables (features), fit it on part of the data, then diagnose it and compare models on data the model has not seen (a **hold-out** or test set). Later lessons teach the methods.
6. **Present and automate.** Report the result to the stakeholders in plain terms, with charts and the measure of success agreed in the charter. If the result is needed again, automate it: a script or dashboard reruns the analysis on new data without an analyst. In the mess example, a script reads each day's token count and prints the next day's number for the cook.



Cook 259 on weekdays and 184 at weekends (5% buffer): 1,663 plates a week instead of 2,100, **saving 437 plates, ₹17,480 a week, and no day in week 2 runs short.**

Figure 1.3.1 Week 1 sets the forecast, 246 on weekdays and 175 at weekends; week 2 tests it with an average error of 6.3 plates, against 78 for always cooking 300.

Other versions of the same process

Textbooks draw the process with different names. The steps match closely; name the version you use in an exam answer.

CIELEN ET AL.	CRISP-DM	KOTU AND DESHPANDE
1 Set the research goal	Business understanding	Prior knowledge (objective, subject area)
2 Retrieve data	Data understanding	Prior knowledge (data)
3 Prepare data	Data preparation	Data preparation
4 Explore data	Data understanding	Data preparation (exploration)
5 Build models	Modeling; Evaluation	Modeling
6 Present and automate	Deployment	Application; Knowledge

CRISP-DM (Cross-Industry Standard Process for Data Mining) is the industry standard with six phases, drawn as a circle around the data. Its separate Evaluation phase asks whether the model meets the business goal, not only whether it is accurate, before anything is deployed. In Kotu and Deshpande's version, the last step, **knowledge**, is the insight the organisation keeps after the model runs.

WORKED EXAMPLE

Question (10 marks). A hostel mess for 320 students cooks 300 dinners every night and throws away the leftovers. The token system gives the dinners eaten over two weeks:

WEEK	MON	TUE	WED	THU	FRI	SAT	SUN
1	250	246	254	2480	232	178	172
2	252	blank	250 (two rows)	244	228	182	176

Apply the six steps to plan how many dinners to cook. Use week 1 to build the forecast and week 2 to test it. Food costs ₹40 a plate.

Solution.

1. **Goal.** Forecast dinners for each day and cook the forecast plus a 5% buffer. Success: no day runs short, and at least 15% fewer plates are cooked than the 2,100 a week cooked now (7 × 300).
2. **Retrieve.** Export the 15 rows from the token system as a CSV file (structured data).
3. **Prepare.** Thursday week 1 = 2,480 is impossible with 320 students; the paper register shows 248, a typing error, so correct it. Delete the duplicate Wednesday row. Tuesday week 2 is blank (machine down): leave it out. Transform: add a column "day type" = weekday (Mon–Fri) or weekend. 13 clean rows remain.
4. **Explore.** Week 1 weekday mean = (250 + 246 + 254 + 248 + 232) ÷ 5 = 1,230 ÷ 5 = **246**. Weekend mean = (178 + 172) ÷ 2 = **175**. Pattern: about 70 fewer students eat on weekends.
5. **Model.** Forecast = week 1 mean for that day type. Test on week 2's six days. Absolute errors: weekdays |252 − 246| = 6, |250 − 246| = 4, |244 − 246| = 2, |228 − 246| = 18; weekend |182 − 175| = 7, |176 − 175| = 1. Mean absolute error (MAE) = 38 ÷ 6 = **6.3 plates**. The old rule "always 300" has errors 48, 50, 56, 72, 118, 124: MAE = 468 ÷ 6 = **78 plates**.
6. **Present.** Cook 246 × 1.05 = 258.3 → **259** on weekdays and 175 × 1.05 = 183.75 → **184** on weekends (round up). Weekly plates = 5 × 259 + 2 × 184 = **1,663**, a saving of 2,100 − 1,663 = 437 plates (20.8%), worth 437 × ₹40 = **₹17,480 a week**. Check: week 2's highest weekday (252) and weekend (182) counts are below 259 and 184, so no day runs short. Goal met. Automate: a script reads each day's token count and prints the next day's number for the cook.

EXAM TRAPS

- The process is a loop, not a straight line. A diagram without feedback arrows loses marks.
- Students skip step 1. Without a measurable goal you cannot say whether the project worked.
- Cleansing, integrating and transforming are three sub-steps of data preparation, not three separate steps of the process.
- Testing a model on the same rows it was built from makes it look better than it is; always test on unseen data.
- Do not mix Cielen's six steps with CRISP-DM's six phases in one answer. Name your version.

CHECK YOURSELF

– REMEMBER **List the six steps of the data science process.** 3 marks

Setting the research goal; retrieving data; data preparation (cleansing, integrating, transforming); data exploration; data modeling; presentation and automation.

– REMEMBER **What does a project charter contain?** 3 marks

Research goal; mission and context; how the analysis will be done; resources; proof the project is achievable; deliverables and a measure of success; timeline.

– UNDERSTAND **Why is the data science process iterative? Give two examples of going back.** 4 marks

Later steps reveal problems the earlier ones missed. EDA finds a unit mismatch, so you return to cleansing; the model is weak, so you retrieve more data; the presentation shows stakeholders wanted a different measure, so the goal is refined.

– UNDERSTAND **Explain cleansing, integrating and transforming with one example each.** 6 marks

Cleansing: correct 2,480 to 248, remove duplicate rows. Integrating: join the marks table and the attendance table on roll number. Transforming: turn "weekday/weekend" into a 0/1 column, or take the log of incomes.

– APPLY **In the worked example, suppose week 2's Tuesday is found to be 238. Find the new MAE of the forecast.** 3 marks

New error $|238 - 246| = 8$. Total $= 38 + 8 = 46$ over 7 days. MAE $= 46 \div 7 \approx$ **6.6 plates**.

– APPLY **Name the step: (i) deleting duplicate rows, (ii) a histogram of fares, (iii) writing the success measure, (iv) a nightly script that emails the forecast, (v) downloading a CSV from data.gov.in, (vi) comparing two models on a test set.** 3 marks

(i) Prepare. (ii) Explore. (iii) Goal. (iv) Present and automate. (v) Retrieve. (vi) Build models.

– ANALYZE **CRISP-DM has a separate Evaluation phase; Cielen's process does not. Where does evaluation happen in Cielen's process, and what does a separate phase add?** 4 marks

In Cielen's step 5 (model diagnostics and comparison on hold-out data). A separate phase checks the model against the business goal (cost, fairness, whether stakeholders can use it) and decides whether to deploy or go back, not just which model is most accurate.

BEYOND THE SYLLABUS

Teams now run the process continuously. Data and models are versioned like code, experiments are logged so they can be repeated, and deployed models are watched for drift, the slow change in data that makes a model worse over time. Microsoft's Team Data Science Process is one published variant of the steps above. Search for "MLOps", "MLflow experiment tracking" and "model drift". This is not for the exam.

Key terms of this lesson

TERM	MEANING OR FORMULA	CHAPTER
Data science	Scientific methods, statistics and computing used to turn data into knowledge and decisions	1.1
Danger zone	Programming + domain knowledge without statistics	1.1
Lift	Response rate of the targeted group ÷ overall response rate	1.1
Data, information, knowledge	Raw values → values with context → patterns that guide action	1.2
Three levels of structure	Structured (fixed schema), semi-structured (self-describing keys or tags), unstructured (no predefined model)	1.2
Seven facets	Structured, unstructured, natural language, machine-generated, graph-based, audio-image-video, streaming	1.2
Data science process	Goal → retrieve → prepare (cleanse, integrate, transform) → explore → model → present and automate; iterative	1.3
CRISP-DM	Business understanding, data understanding, data preparation, modeling, evaluation, deployment	1.3
Mean absolute error	MAE = Σ|actual − forecast| ÷ n	1.3

Exam question bank

Long-answer exam questions on these chapters, each with an outline of the points an examiner looks for. Answer one in full on paper before you read its outline.

Remember and understand

– UNDERSTAND **Explain the need for data science. Describe its benefits and uses with examples from five sectors.** 13 marks

Outline: definition with the three skills (Venn diagram); need: data volume, cheap storage and computing, free tools, decisions under uncertainty, competition; benefits: better decisions, lower cost, personalisation, early risk warning, new products, discovery; five sectors, each with the decision improved (banking fraud, retail recommendations, healthcare imaging, transport travel times, energy demand forecasts); a one-line caution on bias.

– UNDERSTAND **Explain the facets of data with an example of each.** 13 marks

Outline: the seven facets (structured, unstructured, natural language, machine-generated, graph-based, audio-image-video, streaming); an example and the usual storage or method for each; map them to the three levels of structure; note that streaming describes arrival, not structure.

– UNDERSTAND **Compare structured, semi-structured and unstructured data. Show one record in XML and explain how Python reads it.** 8 marks

Outline: comparison table (schema, storage, query tool, example, ease of analysis); an XML record with elements and an attribute; `xml.etree.ElementTree.fromstring`, `find`, `.text`, `.get`; XML text arrives as strings and must be converted.

– UNDERSTAND **Explain the data science process with a neat diagram.** 16 marks

Outline: diagram of six steps with feedback arrows; for each step, what is done and what it produces: project charter contents; internal and external retrieval with quality checks; cleansing, integrating, transforming with examples; EDA with plots; model selection, fitting, hold-out testing; presentation and automation; why the process is iterative.

Apply

– APPLY **A city bus operator wants to cut waiting time at stops. Plan the project through the six steps, naming the data sources and the type of analytics used at each stage.** 16 marks

Outline: measurable goal (average wait under 10 minutes at peak); sources: GPS on buses (primary, machine-generated stream), ticket machine logs (internal), passenger survey (primary), road traffic API (secondary); preparation: remove GPS jumps, join trips to stops; EDA: wait time by hour and route (descriptive), why some routes bunch (diagnostic); model: forecast passengers per stop per hour (predictive); present: a timetable that minimises waiting within the fleet size (prescriptive) on a dashboard. Spans Chapters 1.3, 1.4 and 1.5.

Analyze and evaluate

– EVALUATE **"Data science is just machine learning with a new name." Do you agree? Justify.** 10 marks

Outline: disagree in part; machine learning is the modelling step; data science also covers the goal, collection, cleaning, exploration, communication and deployment, plus statistics for uncertainty and domain knowledge (Venn diagram, danger zone); many projects need only descriptive analytics and no machine learning; conclude with a balanced view.

– ANALYZE **Compare Cielen's six-step process, CRISP-DM and Kotu and Deshpande's process. Which would you use for a final-year project, and why?** 12 marks

Outline: mapping table of the steps; CRISP-DM's separate evaluation and deployment phases; Kotu's prior knowledge and knowledge steps; all are iterative; choice with a reason (for example, Cielen's steps for a short project with a clear goal, CRISP-DM where a client must approve deployment).

– EVALUATE **A hospital wants to predict which patients will miss follow-up appointments. Identify the facets of data available, the collection and ethical issues, and the main risk at each step of the process.** 16 marks

Outline: facets: appointment tables (structured), discharge summaries (semi-structured or text), doctors' notes (natural language), SMS delivery logs (machine-generated); collection: internal records, consent, data protection law, removing names; risks: vague goal, missing visits recorded as no-shows, leakage of future information into the model, unfair results for poorer patients, a model nobody acts on. Spans Chapters 1.2, 1.3 and 1.5.

Glossary (terms used in this lesson)

Algorithm	A step-by-step recipe for a task, such as k-means for clustering.
Analytics	The use of data to answer questions and support decisions.
CRISP-DM	Cross-Industry Standard Process for Data Mining: a six-phase project model.
Data lake	A store that keeps raw data of any form until it is needed.
Data mart	A slice of a data warehouse for one department.
Data warehouse	A store of cleaned, integrated data kept for reporting and analysis.
EDA	Exploratory data analysis: summaries and plots to understand data before modelling.
Facets of data	The different forms data takes, by structure, origin and arrival.
JSON	JavaScript Object Notation: text of keys and values, the common format of API replies.
Model	What an algorithm produces from data; it makes predictions or summaries.
Project charter	The document that fixes a project's goal, method, resources, deliverables, success measure and timeline.
Schema	A fixed description of a dataset's fields and their types.
Semi-structured data	Data that carries its own labels (keys or tags) instead of a fixed schema.
Unstructured data	Data with no predefined model, such as text, images, audio and video.

Sources and further reading

Textbooks

- Davy Cielen, Arno D. B. Meysman and Mohamed Ali, *Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools*, Manning, 2016.
- Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed., Morgan Kaufmann, 2019.
- Foster Provost and Tom Fawcett, *Data Science for Business*, O'Reilly, 2013. Data-analytic thinking and the CRISP-DM process.
- Wes McKinney, *Python for Data Analysis*, O'Reilly (3rd ed., 2022). The chapter on data loading, storage and file formats: JSON, XML, HTML and web APIs.
- Joel Grus, *Data Science from Scratch*, 2nd ed., O'Reilly, 2019. The chapter on getting data: files, scraping and APIs.

Web resources

The writing session's network could not open web pages, so these well-known open resources are listed from memory and were not opened.

- Wikipedia, "Cross-industry standard process for data mining": the six CRISP-DM phases.
- Open Government Data Platform India, data.gov.in: open datasets and APIs from Indian government bodies.
- Python documentation, json and xml.etree.ElementTree: the modules used in Chapters 1.2 and 1.5.
- Jake VanderPlas, *Python Data Science Handbook* (free online edition): NumPy, pandas and plotting for the later lessons.

Syllabus: the University free n easy universal syllabus, built from top universities' syllabi.

Built for the University free n easy YouTube channel from the universal syllabus of Data Science and Analytics (Oct 2026). Written with top universities' syllabi and standard textbooks; check a formula against your prescribed book before an exam, and report mistakes on the channel.