Big data

Is the spreadsheet with your class marks big data? Of course not. And every tap of every travel card on every bus in Mallorca for a whole year? Or every search for a flight and a hotel that a Palma travel website receives in August? Somewhere between those examples there is a line. Where is it?

Big data is one of the enabling technologies you met in Unit 2, and in Unit 4 you saw that AI learns from data. This lesson explains what makes data "big", why "a lot of data" is not the right answer, and how big data, data analysis, machine learning, deep learning and AI fit together, so that you can stop using the five terms as if they were synonyms.

What you will learn

- Describe the five V's that define **big data**: volume, velocity, variety, veracity and value
- Tell **structured**, **semi-structured** and **unstructured** data apart
- Decide whether a case is a big data problem, and which V weighs most
- Explain how big data, **data analysis**, **machine learning**, **deep learning** and AI are related
- Name the four types of analysis by the question they answer
- Avoid the trap: big data does not just mean "lots of data"

What makes data "big"

> **Big data** are datasets that, because of their volume, the speed at which they arrive or their variety, cannot be stored, processed or analysed efficiently with traditional tools (a spreadsheet, a single database on one server) and need a **scalable** architecture.

In the city-builder game, a small town has a few dozen sensors and you could read them one by one. A city of a million inhabitants has traffic counters, water meters, bus GPS units and citizens complaining on social media, all at once, every second. The overlays still work, but behind them there is no longer one computer: there is a whole data centre.

The first three V's were proposed by the analyst Doug Laney in a 2001 research note for the META Group (now part of Gartner). Most textbooks today add two more.

V	What it means	Example
Volume	The amount of data is very large and keeps growing	Every flight and hotel search a travel website receives each day
Velocity	Data arrive fast and must be handled almost in real time	The GPS position of every bus in a network, sent every few seconds
Variety	Many formats and sources mixed together	Bookings (tables), reviews (text), photos, server logs (JSON), sensors
Veracity	Quality and reliability: incomplete, duplicated, wrong or biased data	Fake reviews, badly calibrated sensors, the same customer typed twice

V	What it means	Example
Value	Data are only worth keeping if they bring a benefit	Forecasting occupancy to adjust prices and staff

Some sources add more V's (variability, validity, visualisation...). The US National Institute of Standards and Technology (NIST), for example, adds **variability**. You do not need to memorise them: volume, velocity and variety are in almost every definition, and veracity and value complete the usual five.

Structured, semi-structured, unstructured

Variety becomes clearer when you sort data by how much structure they have.

Kind	What it looks like	Example
Structured	Rows and columns with a fixed schema	A SQL table of bookings
Semi-structured	Labels or keys, but no rigid schema	JSON from an API, XML, a server log
Unstructured	Free content with no predefined fields	Review text, photos, audio, video

> **Careful: big data is not "a lot of data"**
> A huge file that is uniform and never changes is not the same problem as a continuous stream of mixed data. And the other way round: **not every company needs big data**. Many small and medium companies are perfectly served by a good database and somebody who knows how to ask it questions. The test is not "how many gigabytes?" but "does the traditional tool stop working here, and why?"

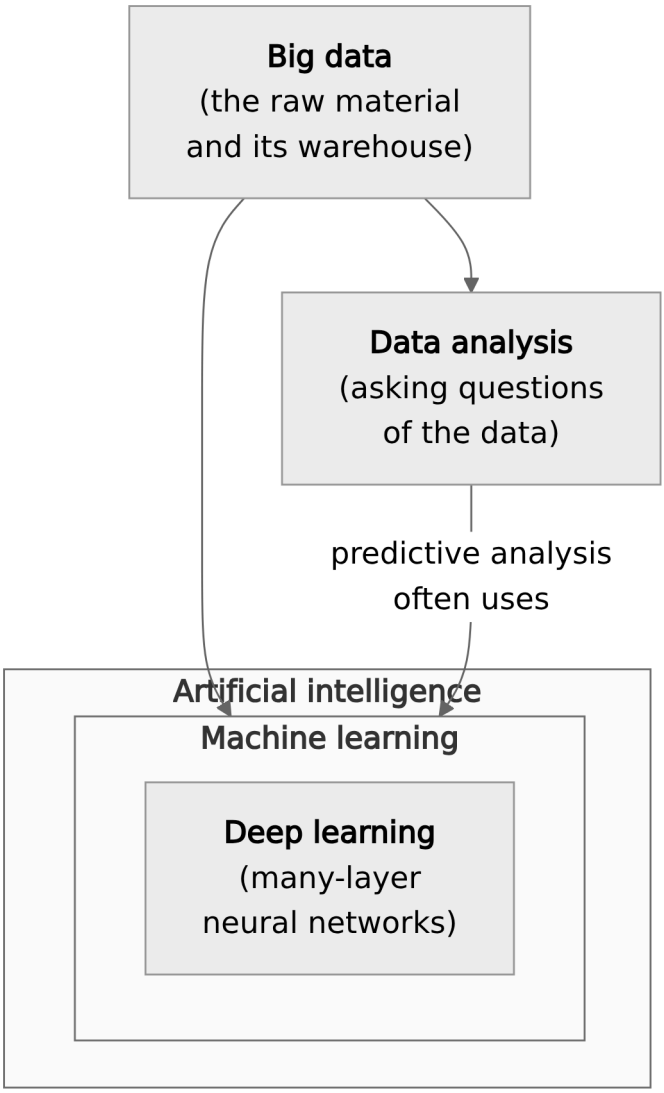
Who is who: big data, analysis, ML, DL and AI

Term	What it is
Artificial intelligence (AI)	Systems that do tasks we associate with human intelligence (you saw it in Unit 4)
Machine learning (ML)	The part of AI where the system **learns from data** instead of following rules written by hand
Deep learning (DL)	The part of ML based on neural networks with many layers; it usually needs a lot of data and computing power
Big data	Not AI. It is the **raw material** and the infrastructure to store and move it
Data analysis	The process of examining data to get information out of them; it can use simple statistics or ML

Big data without analysis is a full warehouse that nobody looks at. Analysis without good data gives bad answers. And a lot of the AI that companies actually use is predictive analysis done at scale. That is why the three usually go together.

Four questions, four types of analysis

Type	Question	Hotel example	Typical tool
Descriptive	What happened?	Average occupancy in July	Reports, charts



Type	Question	Hotel example	Typical tool
Diagnostic	Why did it happen?	Occupancy fell the week an event was cancelled	Crossing data sources
Predictive	What will happen?	Expected bookings for next week	ML
Prescriptive	What should we do?	Recommended price per room and day	ML plus optimisation

In class: big data in your company's sector

Work **in pairs**, on the sector of your group's project company. This is guided web research: every claim needs a working source. **No AI tools.**

Step	Do this	You should see
1	Search for one **real, documented** case of big data or data analytics in your company's sector, ideally in the Balearic Islands or in Spain (a company's press release, a news article, a public project, a case study). Try queries like `big data turismo Baleares`, `analítica de datos agua contadores inteligentes`, `big data puerto`	Two or three candidate articles

Step	Do this	You should see
2	Pick the most reliable one: who publishes it, when, and does it name the company and the data? Copy the URL and write the full citation.	One source you trust
3	Fill in a table: which **V's** are present and why (at least two), which kinds of data (structured, semi-structured, unstructured), and which **type of analysis** it is (descriptive, diagnostic, predictive, prescriptive).	A five-row table
4	Say whether the case uses **ML** or **DL**, or only statistics, and how you know from the source. If the source does not say, write "not stated".	One sentence
5	Now your own company: does it have, or could it have, a big data problem? Justify it with at least two V's. Write one **predictive question** it would like to answer with its data.	Two short paragraphs
6	Hand in the document on Classroom (one per pair, both names).	

For your project

Steps 3 to 5 are raw material for **section B (mini data-science pipeline)** of Deliverable 5: section B starts from a question, and the question you write today is a good candidate. The real case you found can also go into the group part as context for the sector. Cite it.

Key takeaways

- **Big data** are data that traditional tools cannot handle efficiently because of their **volume**, **velocity** or **variety**; **veracity** and **value** complete the usual five V's.
- Data can be **structured**, **semi-structured** or **unstructured**; variety means mixing them.
- Big data is not "lots of data", and not every company needs it.
- **Deep learning** is inside **machine learning**, which is inside **AI**; big data is the raw material, and **data analysis** is asking the data questions.
- The four types of analysis answer four questions: what happened, why, what will happen, what should we do.

Quick check

1 Which three V's appear in almost every definition of big data?

A Value, veracity and visualisation
B Volume, velocity and variety
C Volume, value and validity
D Velocity, veracity and variability

2 Review text, photos, sensor readings and booking tables, all mixed in one system. Which V is that?

A Volume
B Velocity
C Variety
D Value

3 A gestoria keeps its 800 clients in one spreadsheet. Is it big data?

A Yes, because it contains personal data
B No: few, structured, stable data that a normal database handles well
C Yes, because any customer list is big data
D It depends only on the size of the file in megabytes

4 Which statement about AI, ML and DL is correct?

- **A** Machine learning is inside deep learning, which is inside AI
- **B** Deep learning is inside machine learning, which is inside AI
- **C** Big data is a kind of AI
- **D** AI, ML and DL are three names for the same thing

5 "How many bookings will we get next week?" Which type of analysis answers it?

- **A** Descriptive
- **B** Diagnostic
- **C** Predictive
- **D** Prescriptive

6 Which V is about whether the data can be trusted?

- **A** Velocity
- **B** Veracity
- **C** Volume
- **D** Variety

7 A water company in Mallorca reads thousands of smart meters every hour and also receives phone complaints about leaks. Pick the two statements that are true.

- **A** Velocity and variety weigh heavily: frequent readings plus voice or text complaints
- **B** It cannot be big data because water is an OT system
- **C** A model that warns of a likely leak before it is reported would be predictive analysis
- **D** It is big data only if it uses deep learning

You know what big data is and how it feeds analysis and AI. Next you will do the analysis yourself, step by step, on a real dataset, and decide where the data should live: Data science and the cloud.

Solutions

1. B · The classic three, proposed by Doug Laney in 2001, are **volume**, **velocity** and **variety**. Veracity and value were added later to make the usual five.

2. C · Many **formats and sources** at once, structured and unstructured: that is **variety**.

3. B · No traditional tool stops working here. The data are few, **structured** and stable. Big data is defined by the problem, not by the word "data".

4. B · **DL** (many-layer neural networks) is a part of **ML** (learning from data), which is a part of **AI**. Big data is the raw material, not AI.

5. C · "What will happen?" is the **predictive** question. "What should we do about it?" would be prescriptive.

6. B · **Veracity** is about quality and reliability: fake reviews, duplicated customers and badly calibrated sensors all lower it.

7. A, C · New sector, same tools. Frequent meter readings mean **velocity**, readings plus complaints mean **variety**, and warning before it happens is **predictive analysis**. Big data does not require deep learning.