Big data

El full de càlcul amb les notes de la teva classe és big data? És clar que no. I cada vegada que una targeta de transport passa per cada bus de Mallorca durant un any sencer? O cada cerca d'un vol i un hotel que rep un web de viatges de Palma a l'agost? En algun punt entre aquests exemples hi ha una línia. On és?

El big data és una de les tecnologies habilitadores digitals que vas conèixer a la Unitat 2, i a la Unitat 4 vas veure que la IA aprèn de dades. Aquesta lliçó explica què fa que les dades siguin "grans", per què "moltes dades" no és la resposta correcta i com encaixen el big data, l'anàlisi de dades, l'aprenentatge automàtic, l'aprenentatge profund i la IA, perquè deixis de fer servir aquests cinc termes com si fossin sinònims.

Què aprendràs

- Descriure les cinc V que defineixen el **big data**: volum, velocitat, varietat, veracitat i valor
- Distingir les dades **estructurades**, **semiestructurades** i **no estructurades**
- Decidir si un cas és un problema de big data, i quina V hi pesa més
- Explicar com es relacionen el big data, l'**anàlisi de dades**, l'**aprenentatge automàtic** (*machine learning*), l'**aprenentatge profund** (*deep learning*) i la IA
- Anomenar els quatre tipus d'anàlisi per la pregunta que responen
- Evitar la trampa: big data no vol dir només "moltes dades"

Què fa que les dades siguin "grans"

> El **big data** són conjunts de dades que, pel seu volum, per la velocitat amb què arriben o per la seva varietat, no es poden emmagatzemar, processar ni analitzar de manera eficient amb les eines tradicionals (un full de càlcul, una sola base de dades en un servidor) i necessiten una arquitectura **escalable**.

Al joc de construir ciutats, un poble petit té unes quantes desenes de sensors i els podries llegir un a un. Una ciutat d'un milió d'habitants té comptadors de trànsit, comptadors d'aigua, GPS als busos i ciutadans queixant-se a les xarxes socials, tot alhora, cada segon. Les capes continuen funcionant, però al darrere ja no hi ha un ordinador: hi ha un centre de dades sencer.

Les tres primeres V les va proposar l'analista Doug Laney el 2001 en una nota de recerca per al META Group (avui part de Gartner). La majoria de llibres de text d'avui n'hi afegeixen dues més.

V	Què vol dir	Exemple
Volum	La quantitat de dades és molt gran i no para de créixer	Cada cerca de vols i hotels que rep cada dia un web de viatges
Velocitat	Les dades arriben de pressa i s'han de tractar gairebé en temps real	La posició GPS de cada bus d'una xarxa, enviada cada pocs segons

V	Què vol dir	Exemple
Varietat	Molts formats i fonts barrejats	Reserves (taules), ressenyes (text), fotos, registres del servidor (JSON), sensors
Veracitat	Qualitat i fiabilitat: dades incompletes, duplicades, errònies o esbiaixades	Ressenyes falses, sensors mal calibrats, el mateix client escrit dues vegades
Valor	Les dades només val la pena guardar-les si aporten un benefici	Preveure l'ocupació per ajustar preus i personal

Algunes fonts hi afegeixen més V (variabilitat, validesa, visualització…). L'Institut Nacional d'Estàndards i Tecnologia dels Estats Units (NIST), per exemple, hi afegeix la **variabilitat**. No cal que les memoritzis: el volum, la velocitat i la varietat surten gairebé a totes les definicions, i la veracitat i el valor completen les cinc habituals.

Estructurades, semiestructurades, no estructurades

La varietat s'entén millor quan classifiques les dades segons quanta estructura tenen.

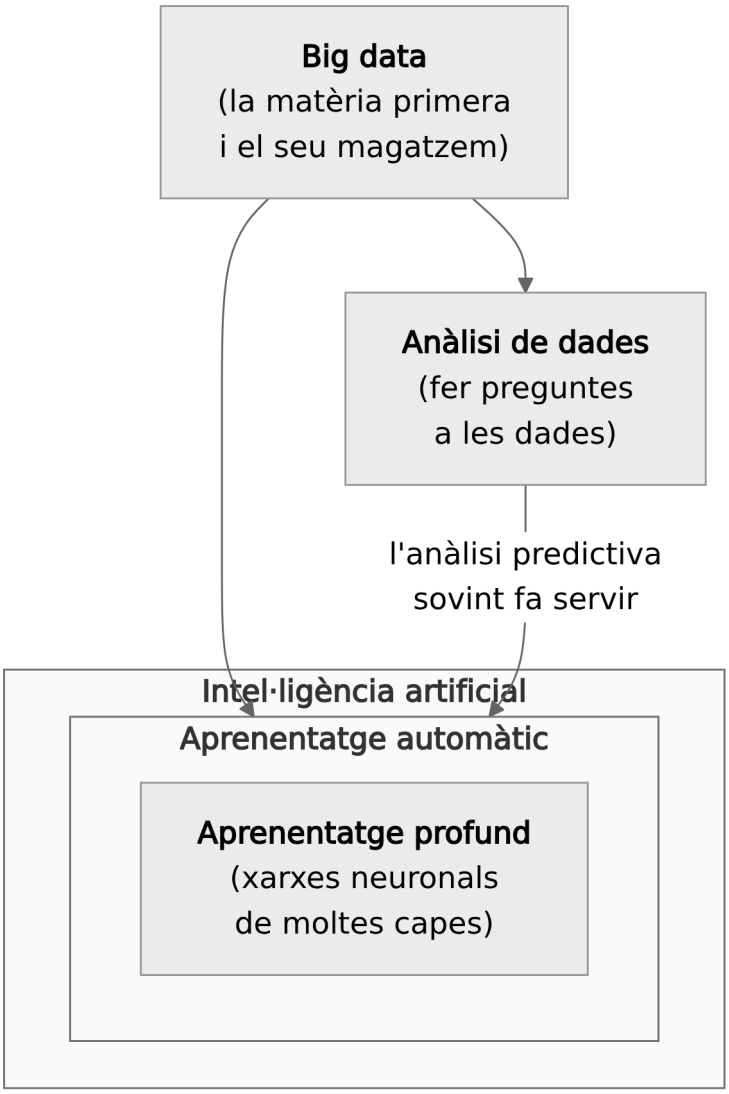
Tipus	Com són	Exemple
Estructurades	Files i columnes amb un esquema fix	Una taula SQL de reserves
Semiestructurades	Etiquetes o claus, però sense un esquema rígid	El JSON d'una API, XML, un registre del servidor
No estructurades	Contingut lliure sense camps predefinits	El text de les ressenyes, fotos, àudio, vídeo

> **Compte: big data no és "moltes dades"**
> Un fitxer enorme que és uniforme i no canvia mai no és el mateix problema que un flux continu de dades barrejades. I al revés: **no totes les empreses necessiten big data**. Moltes petites i mitjanes empreses en tenen prou amb una bona base de dades i algú que sàpiga fer-li preguntes. La prova no és "quants gigabytes?" sinó "l'eina tradicional deixa de funcionar aquí, i per què?"

Qui és qui: big data, anàlisi, ML, DL i IA

Terme	Què és
Intel·ligència artificial (IA)	Sistemes que fan tasques que associem amb la intel·ligència humana (la vas veure a la Unitat 4)
Aprenentatge automàtic (ML)	La part de la IA en què el sistema **aprèn de les dades** en lloc de seguir regles escrites a mà
Aprenentatge profund (DL)	La part de l'ML basada en xarxes neuronals de moltes capes; sol necessitar moltes dades i molta potència de càlcul
Big data	No és IA. És la **matèria primera** i la infraestructura per guardar-la i moure-la
Anàlisi de dades	El procés d'examinar dades per treure'n informació; pot fer servir estadística senzilla o ML

Big data sense anàlisi és un magatzem ple que ningú no mira. L'anàlisi sense bones dades dona males respostes. I bona part de la IA que fan servir de veres les empreses és anàlisi predictiva feta a gran escala. Per això les tres coses solen anar juntes.



Quatre preguntes, quatre tipus d'anàlisi

Tipus	Pregunta	Exemple a l'hotel	Eina típica
Descriptiva	Què ha passat?	Ocupació mitjana al juliol	Informes, gràfics
Diagnòstica	Per què ha passat?	L'ocupació va baixar la setmana que es va cancel·lar un esdeveniment	Creuar fonts de dades
Predictiva	Què passarà?	Reserves previstes per a la setmana que ve	ML
Prescriptiva	Què hem de fer?	Preu recomanat per habitació i dia	ML més optimització

A classe: big data al sector de la teva empresa

Treballa **en parella**, sobre el sector de l'empresa del projecte del teu grup. És una recerca web guiada: cada afirmació necessita una font que funcioni. **Cap eina d'IA.**

Pas	Fes això	Hauries de veure
1	Cerca un cas **real i documentat** de big data o d'analítica de dades al sector de la teva empresa, idealment a les Illes Balears o a Espanya (una nota de	Dos o tres articles candidats

Pas	Fes això	Hauries de veure
	premsa d'una empresa, una notícia, un projecte públic, un cas d'estudi). Prova consultes com `big data turismo Baleares`, `analítica de datos agua contadores inteligentes`, `big data puerto`	
2	Tria el més fiable: qui el publica, quan, i si anomena l'empresa i les dades. Copia'n l'URL i escriu la cita completa.	Una font de confiança
3	Omple una taula: quines **V** hi són presents i per què (almenys dues), quins tipus de dades (estructurades, semiestructurades, no estructurades) i quin **tipus d'anàlisi** és (descriptiva, diagnòstica, predictiva, prescriptiva).	Una taula de cinc files
4	Digues si el cas fa servir **ML** o **DL**, o només estadística, i com ho saps per la font. Si la font no ho diu, escriu "no s'indica".	Una frase
5	Ara la teva empresa: té, o podria tenir, un problema de big data? Justifica-ho amb almenys dues V. Escriu una **pregunta predictiva** que voldria respondre amb les seves dades.	Dos paràgrafs curts
6	Lliura el document a Classroom (un per parella, amb els dos noms).	

Per al teu projecte

Els passos 3 a 5 són matèria primera per a l'**apartat B (minicircuit de ciència de dades)** del Lliurament 5: l'apartat B parteix d'una pregunta, i la pregunta que escrius avui és una bona candidata. El cas real que has trobat també pot anar a la part de grup com a context del sector. Cita'l.

Idees clau

- El **big data** són dades que les eines tradicionals no poden tractar de manera eficient pel seu **volum**, la seva **velocitat** o la seva **varietat**; la **veracitat** i el **valor** completen les cinc V habituals.
- Les dades poden ser **estructurades**, **semiestructurades** o **no estructurades**; la varietat vol dir barrejar-les.
- Big data no és "moltes dades", i no totes les empreses el necessiten.
- L'**aprenentatge profund** és dins l'**aprenentatge automàtic**, que és dins la **IA**; el big data és la matèria primera, i l'**anàlisi de dades** és fer preguntes a les dades.
- Els quatre tipus d'anàlisi responen quatre preguntes: què ha passat, per què, què passarà i què hem de fer.

Comprovació ràpida

1 Quines tres V surten gairebé a totes les definicions de big data?

- **A** Valor, veracitat i visualització
- **B** Volum, velocitat i varietat
- **C** Volum, valor i validesa
- **D** Velocitat, veracitat i variabilitat

2 Text de ressenyes, fotos, lectures de sensors i taules de reserves, tot barrejat en un sol sistema. Quina V és?

- **A** Volum
- **B** Velocitat
- **C** Varietat
- **D** Valor

3 Una gestoria té els seus 800 clients en un full de càlcul. És big data?

- **A** Sí, perquè conté dades personals
- **B** No: poques dades, estructurades i estables, que una base de dades normal gestiona bé
- **C** Sí, perquè qualsevol llista de clients és big data
- **D** Només depèn de la mida del fitxer en megabytes

4 Quina afirmació sobre la IA, l'ML i el DL és correcta?

- **A** L'aprenentatge automàtic és dins l'aprenentatge profund, que és dins la IA
- **B** L'aprenentatge profund és dins l'aprenentatge automàtic, que és dins la IA
- **C** El big data és un tipus d'IA
- **D** IA, ML i DL són tres noms per a la mateixa cosa

5 "Quantes reserves tindrem la setmana que ve?" Quin tipus d'anàlisi ho respon?

- **A** Descriptiva
- **B** Diagnòstica
- **C** Predictiva
- **D** Prescriptiva

6 Quina V tracta de si ens podem refiar de les dades?

- **A** Velocitat
- **B** Veracitat
- **C** Volum
- **D** Varietat

7 Una empresa d'aigües de Mallorca llegeix milers de comptadors intel·ligents cada hora i també rep queixes telefòniques per fuites. Tria les dues afirmacions certes.

- **A** La velocitat i la varietat hi pesen molt: lectures freqüents i, a més, queixes de veu o de text
- **B** No pot ser big data perquè l'aigua és un sistema OT
- **C** Un model que avisi d'una fuita probable abans que algú la comuniqui seria anàlisi predictiva
- **D** Només és big data si fa servir aprenentatge profund

Ja saps què és el big data i com alimenta l'anàlisi i la IA. Tot seguit faràs tu mateix l'anàlisi, pas a pas, sobre un conjunt de dades real, i decidiràs on han de viure les dades: Ciència de dades i núvol.

Solucions

1. B · Les tres clàssiques, proposades per Doug Laney el 2001, són el **volum**, la **velocitat** i la **varietat**. La veracitat i el valor s'hi van afegir més tard per fer les cinc habituals.

2. C · Molts **formats i fonts** alhora, estructurats i no estructurats: això és la **varietat**.

3. B · Aquí cap eina tradicional no deixa de funcionar. Les dades són poques, **estructurades** i estables. El big data el defineix el problema, no la paraula "dades".

4. B · El **DL** (xarxes neuronals de moltes capes) és una part de l'**ML** (aprendre de les dades), que és una part de la **IA**. El big data és la matèria primera, no IA.

5. C · "Què passarà?" és la pregunta **predictiva**. "Què hi hem de fer?" seria prescriptiva.

6. B · La **veracitat** tracta de la qualitat i la fiabilitat: les ressenyes falses, els clients duplicats i els sensors mal calibrats la fan baixar.

7. A, C · Sector nou, mateixes eines. Les lectures freqüents dels comptadors són **velocitat**, lectures més queixes són **varietat**, i avisar abans que passi és **anàlisi predictiva**. El big data no requereix aprenentatge profund.