Big Data: Challenges and Opportunities

Lecture 1 — Introduction

Stéphane Derrode · Département MI · Centrale Lyon · 2025–2026

Module Overview

📚 Structure

- 8 × 2h lectures — Monday 8–10am
- 3 × 4h lab sessions — Wednesday afternoons
- 1 bibliography report (groups of 4–6)
- 1 written exam — Dec. 15, 1h30

📋 Key Info

- Course page: perso.ec-lyon.fr/derrode.stephane
- Submissions: Pedagogie3 platform
- Topic registration: by Oct. 6
- Attendance is mandatory

Part 1

Data is Everywhere

Data is Everywhere

«Every two days we create as much information as we did from the dawn of civilization up until 2003.»

— Eric Schmidt, Google CEO, 2010

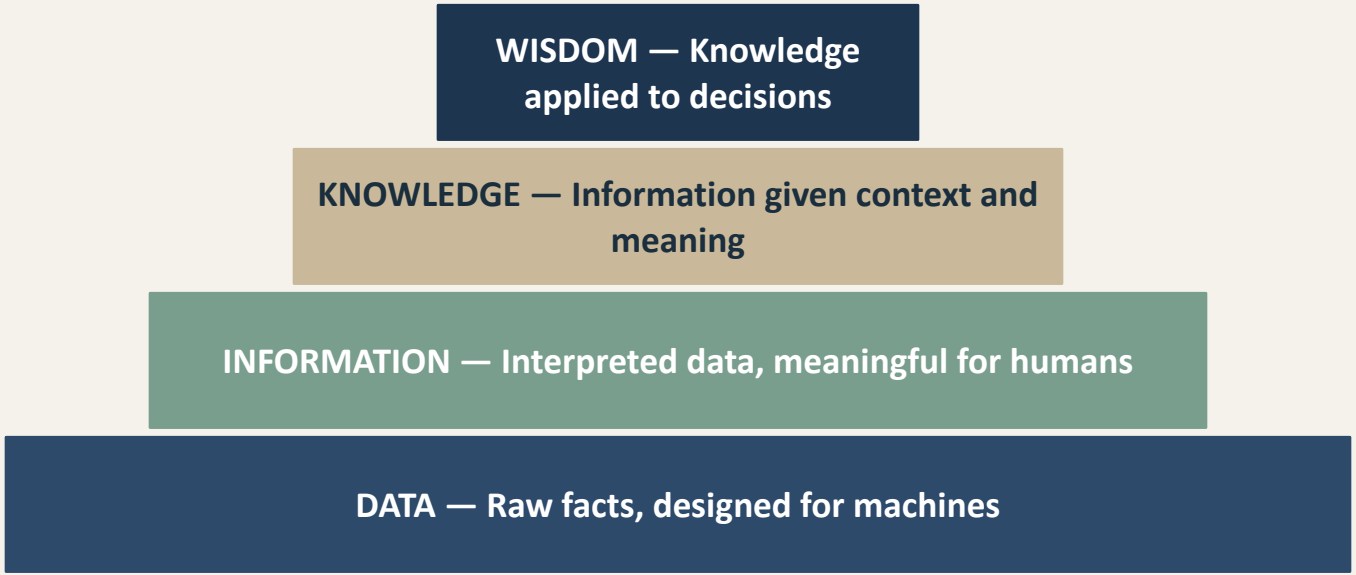
Every device around us collects, stores and transmits data continuously.

💡 **Think about it:**

Can you identify 3 devices near you right now that are actively collecting data?

→ What data?
→ Stored where?
→ Used how?

From Data to Wisdom



Attributed to Russell Ackoff, 1989

The Data Economy

328 million TB
of data created every day in 2023

$103 billion
global Big Data market in 2023

11 million
data science jobs projected by 2026

Sources: Statista, IDC, World Economic Forum 2023–2024

Part 2

What is Big Data?

The Data Deluge

«With enough data, the numbers speak for themselves.»
— Chris Anderson, Wired Magazine, 2008

Traditional approach
Hypothesis → Experiment → Analysis → Theory

The data-driven claim
Collect everything → Find patterns → Correlation is enough?

⚠️ *Critical thinking: correlation ≠ causation — always question your data.*

Defining Big Data

«Big Data is a new generation of technologies and architectures designed to extract economic or scientific value from a large variety of data, by enabling high-velocity capture, discovery and/or analysis.»

— Gartner, 2001

This definition introduced the now-classic 3 Vs model → expanded to 6 Vs.

The 3 Vs of Big Data

📦 Volume

From terabytes to petabytes

Twitter: 7 TB/day (2013)
→ 500 TB/day (2023)

⚡ Velocity

Real-time generation & processing

Stock trading, IoT sensors
Social media feeds

🗂️ Variety

Structured, semi-structured, and unstructured data

Text, images, video, logs

Beyond the 3 Vs — The 6 Vs

📦 **Volume**

⚡ **Velocity**

🗂️ **Variety**

+ extended with:

💎 **Value**

The ultimate goal
Economic, societal, scientific impact of data

📉 **Variability**

Data inconsistency over time
Complicates management and pipelines

✅ **Veracity**

Quality & trustworthiness
Garbage in → Garbage out

🎯 Quick Check — Parts 1 & 2

Which of the following is NOT one of the original 3 Vs of Big Data defined by Gartner?

A — Volume

B — Velocity

C — Veracity ✓

D — Variety

Answer: C — Veracity was added later as a 4th V (beyond the original Gartner model)

Part 3

Data Lifecycle

The Data Lifecycle

Data = 'black gold' — value must be extracted and refined at every stage

Storage: Data Centers

- Tens to hundreds of petabytes stored per facility
- Industrial constraints: energy supply, cooling, connectivity
- Consume 1–2% of global electricity — rising fast with AI
- New trends: immersion cooling, Arctic locations, underwater DCs
- Major players: Google, Microsoft, Amazon, Meta, OVH

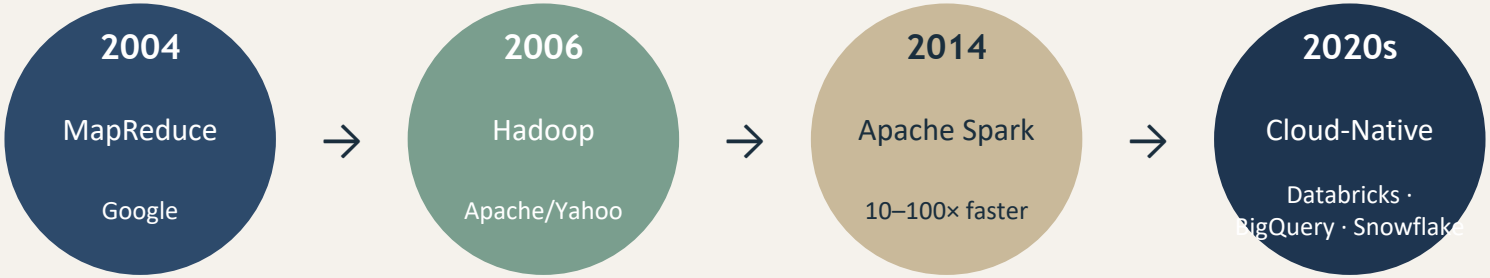
🌍 10,978 data centers worldwide (2024)

⚡ 200+ TWh/year consumed

🤖 AI is driving +10%/year growth

🌊 Microsoft Project Natick: underwater DC

Processing: From MapReduce to Cloud-Native



Core idea: separate computation from storage infrastructure — distribute everything

Today: serverless pipelines, auto-scaling, no infrastructure to manage

🎯 Quick Check — Part 3

What was the key conceptual innovation of MapReduce (Google, 2004)?

A — Inventing the relational database

B — Separating computation from infrastructure management ✓

C — Creating the first data center

D — Introducing real-time data streaming

Answer: B — MapReduce separates the 'what to compute' from 'how to distribute it'

Part 4

Risks & Ethics

The Privacy Paradox

What apps ask for:

📍 Location (always on)
🎙️ Microphone access
📷 Camera access
👥 Full contacts list
📅 Calendar access
→ *…for a flashlight app.*

💬 **Discussion:**

Would you trade your precise location data for a free coffee?

What about your health data for a free gym membership?

→ *Why do we accept these trade-offs?*

🎬 Spécial Investigation (2015)

🎬 RTS Geopolitis GAFA

🎬 RTS Big Data données

🎬 Arte: exploitation données clients

🎬 Arte: Big Data & océans

🎬 Arte: Big Data au CERN

79% of Americans are concerned about how companies use their data — Pew Research Center, 2023

Big Tech & Data Power

- GAFAM control 80%+ of global personal data (est.)
- Combined revenue > $1.5 trillion in 2023
- New entrants: NVIDIA (AI compute), OpenAI, ByteDance
- Acronym evolution: GAFAM → MAAMA → 'Magnificent 7'
- Network effect: more users → more data → better AI → more users

Data moats: accumulated data becomes an insurmountable competitive advantage

Regulatory response: DMA (EU Digital Markets Act, 2023) aims to break these moats

Videos to Watch — Freedoms & Big Tech

The following videos are in French — recommended for viewing outside of class

🎬 Big Data : les nouveaux devins (Spécial Investigation, 2015)	50 min
🎬 GAFA : une domination abusive (RTS Geopolitis)	15 min
🎬 Big Data : que fait-on de nos données ? (RTS/TV5monde)	15 min
🎬 La course à l'exploitation des données clients (Arte Future)	6 min
🎬 Big Data et la protection des océans (Arte Future)	5 min
🎬 L'utilisation du Big Data au CERN (Arte.tv, 2017)	2 min 🔗 CNIL website

Part 5

Regulations

Protecting Personal Data — A Legal Timeline

1978 — French 'Informatique et Libertés' — Creation of CNIL

2003 — EU Directive — Right of access becomes obligation to publish

2011 — French Decree — Free reuse of public data

2016 — Loi République Numérique — Open data commitment

2018 — GDPR enforced — EU-wide personal data protection

2024 — AI Act — World's first AI regulation framework

GDPR — General Data Protection Regulation (2018)

Any data concerning a European citizen, even processed outside the EU, falls under this regulation.

→ Extra-territorial scope: applies to any company worldwide handling EU citizens' data

Your 6 rights:

- Right to information
- Right of access
- Right to erasure
- Right to rectification
- Right to portability
- Right to restrict processing

⚠️ Fines: up to €20M or 4% of global annual turnover — whichever is greater

🔗 GDPR full text (EUR-Lex)

The EU AI Act (2024) — World's First AI Regulation

A risk-based approach: four levels of regulation

🚫 **PROHIBITED — Unacceptable Risk** Social scoring · Cognitive manipulation · Real-time facial recognition in public spaces

⚠️ **HIGH RISK — Strictly Regulated** Healthcare · Education · Justice · Employment · Critical infrastructure

ℹ️ **LIMITED RISK — Transparency Required** Chatbots, deepfakes and AI-generated content must be clearly labeled

✅ **MINIMAL RISK — Free to Use** Spam filters · Music recommendations · Video games

🔗 AI Act — European Parliament

Fines up to €35M or 7% of global turnover

Why the AI Act Matters

🛡️ Protect citizens

Guard against harmful AI while preserving fundamental rights and human dignity

🚀 Foster innovation

Clear, predictable rules create a stable environment for European AI development

🌍 Set global standards

Like GDPR for data, the AI Act aims to become the international benchmark

The EU's dual goal: protect citizens AND compete globally in AI — the so-called 'Brussels Effect'

🎯 Quick Check — Parts 4 & 5

Under the EU AI Act, which of the following uses of AI is PROHIBITED?

A — Spam email filters

B — AI-assisted medical diagnosis

C — Social scoring systems ✓

D — Music recommendation algorithms

Answer: C — Social scoring (ranking citizens based on behavior) is explicitly prohibited under Article 5

What's Next

📝 **Bibliography Report**

- Groups of 4–6 students
- ~10 pages on a Big Data topic
- Deadline: Friday Dec. 19, 2025
- Register your group & topic by Oct. 6
  → shared file on course page

🌐 **Course page:**
perso.ec-lyon.fr/derrode.stephane
→ Teaching → MOD 2.1

📁 Submissions: Pedagogie3 platform

🔗 perso.ec-lyon.fr/derrode.stephane

🔗 Pedagogie3 platform

Next lecture: Open Data & Linked Open Data — Oct. 6