Data, Data Storage, Data Collection

Lecture 1: Introduction to Data

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

Administrative Information

Course Description and Organization

Data is the foundation of **intelligent systems** and **decision-making processes**.

Understanding **how to work with data**, from its collection to its use for decision making, is central to modern data-driven applications.

This course introduces students to the full **data lifecycle**. It explores how to **collect**, **clean**, **store**, and **analyse** data, before discussing how to **communicate** the findings.

Course Description and Organization

This course consists of **14 sessions**, consisting of a mix of lectures, group exercises, pen-and-paper exercises and labs.

Prerequisites:

- Understanding of relational databases and distributed systems
- Programming with Python
- Basic knowledge of algebra and statistics

Grading

The final grade will be based on three components:

- **Continuous Evaluation:** Quizzes throughout the course to monitor your progress.
- **Project:** Apply the data life-cycle to extract insights from a dataset, culminating in a presentation to your peers during the final sessions. (Details provided later today.)
- **Final Assessment:** Written exam at the end of the course to evaluate overall understanding.

The grading scheme is still pending validation (and will be shared as soon as possible).

Pedagogical content (slides, exercises, etc.)

Will be available on ESSEC's platform as soon as I have access.

In the meantime, on my website
`https://romainpascual.fr/teaching/dsaidams`

Introduction

Session Objectives

At the end of this session, you should be able to:

- Understand what is data and its relation to information and knowledge
- Understand what are the sources of data and their influence on data quality
- Know the three main representations of data and their properties

Why Data Matters?

" In short, ladies and gentlemen, my message today is that **data is gold**. We have a huge goldmine in public administration. Let's start mining it. " (Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda, 2011)

" **Data is the new oil**. " (Clive Humby, Mathematician, 2006)

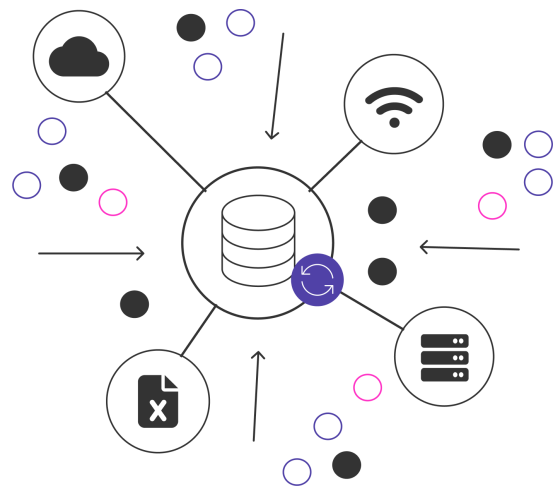
Data is Everywhere

Data comes from a wide range of **sources**: sensors, user activity, surveys, logs, transactions.

It is captured through many **means**: automatic systems (e.g., IoT), manual entry, and APIs (application programming interface).

Devices generating data include:

- Smartphones, laptops, wearable tech
- Industrial machines, satellites
- Medical scanners, CCTV systems



Data in Society

Data drives decision-making and policy across all domains.

- **Industry:** real-time dashboards for supply chain, customer behavior prediction
- **Sciences:** climate modeling, genomic research, pandemic tracking
- **Public Policy:** urban planning from mobility data, budget allocations from census

Data in Society

Data drives decision-making and policy across all domains.

- **Industry:** real-time dashboards for supply chain, customer behavior prediction
- **Sciences:** climate modeling, genomic research, pandemic tracking
- **Public Policy:** urban planning from mobility data, budget allocations from census

With increased usage comes increased responsibility, privacy and ethical concerns must be addressed. How can we balance innovation with individual rights to privacy?

What is Data?

Definition (Data)

Data is a collection of value (formally a sequence of symbols) that can be processed by a computer.

When the symbols are 0 and 1, we call it **digital data**. In modern computer systems, all data is digital.

Data comes in many forms, such as numbers, text, images, audio, and video.

Metadata is data about data, providing information about the data itself, such as its source, format, and structure.

Processing Data

Data can be seen as the smallest units that can be used as a basis for calculation, reasoning, or discussion

As such, data are both inputs, i.e., variables that can be manipulated, and outputs, i.e., results of computations. Thus, data can either be **raw** or **processed**.

Information

Data may represent abstract ideas or concrete measurements.

When collected and observed without interpretation, these elements remain mere data points-discrete and disorganized entities lacking inherent meaning or significance.

The meaning of data is not inherent; it depends on the context in which it is analysed.

Definition (Information)

Information is processed data with meaning.

Information: Examples

Example (Temperature)

For instance, if data points include daily temperature readings over a year, information is recognizing the trend of temperatures, understanding seasonal changes, and predicting future weather conditions.

Example (Sales)

A sequence of numbers "100, 150, 200" is just data. However, if you put it into context: "The sales of a product over the past three months were 100, 150, and 200 units," it becomes information because it provides meaning and context.

From Data to Information

Converting data into information involves the following steps:

Processing Data is transformed to prepare it for analysis. This may include cleaning, filtering, and aggregating data.

Organizing Data is structured and categorized to facilitate retrieval. This may involve creating databases, tables, or other data structures.

Interpreting Data is analyzed to extract meaning. This may involve statistical analysis, data mining, or machine learning techniques.

Visualizing Data is presented in a way that makes it easier to understand and use for decision-making. This may involve creating charts, graphs, or other visual representations of the data.

Knowledge

Knowledge is the understanding and awareness gained through experience, education, or analysis. It is the ability of applying information to specific contexts or problems. Knowledge allows individuals to make informed decisions, solve problems, and create new ideas.

Definition (Knowledge)

Knowledge is the application of information.

It can be **explicit**: documented and easily shared, such as in books or databases, or **tacit**: personal and acquired through practice and experience, such as skills or intuition.

Example

Imagine you see a car accelerate past you on the motorway. You recognize that this is due to the design of its engine, which is more effective than the one in your car, owing to factors such as the number of cylinders and gear ratios. You also understand that its sleek aerodynamic profile minimizes air resistance, allowing it to cut through the air more easily. This is the information you possess about the car's speed.

Knowledge, however, is the application of this information in a concrete context. You might also know that these characteristics don't matter much because the national speed limit is too low to fully utilize the car's potential. Thus, knowledge enables one to make decisions, solve problems, and predict outcomes.

From Information to Knowledge

The process of transforming information into knowledge involves:

Contextualization Placing information in a relevant context to enhance understanding.

Application Using information to solve problems or make decisions.

Reflection Evaluating the outcomes of applying information to refine understanding and improve future decision-making.

Sharing Communicating knowledge with others to foster collaboration and collective learning.

Knowledge is dynamic and evolves as new information is acquired and experiences are gained.

Data, Information, Knowledge

In summary, data is the raw material, information is the processed data in context, and knowledge is the ability to realized an action based on that information.

$\mathbf{K} = \mathbf{I} \cdot \mathbf{A}$

C = Knowledge
I = Information
A = Application

Knowledge
Information assimilated to perform an action

I am in Orléans today
so *I dress warmly*

Information
Data put into context

The temperature is 10°C in Orléans today

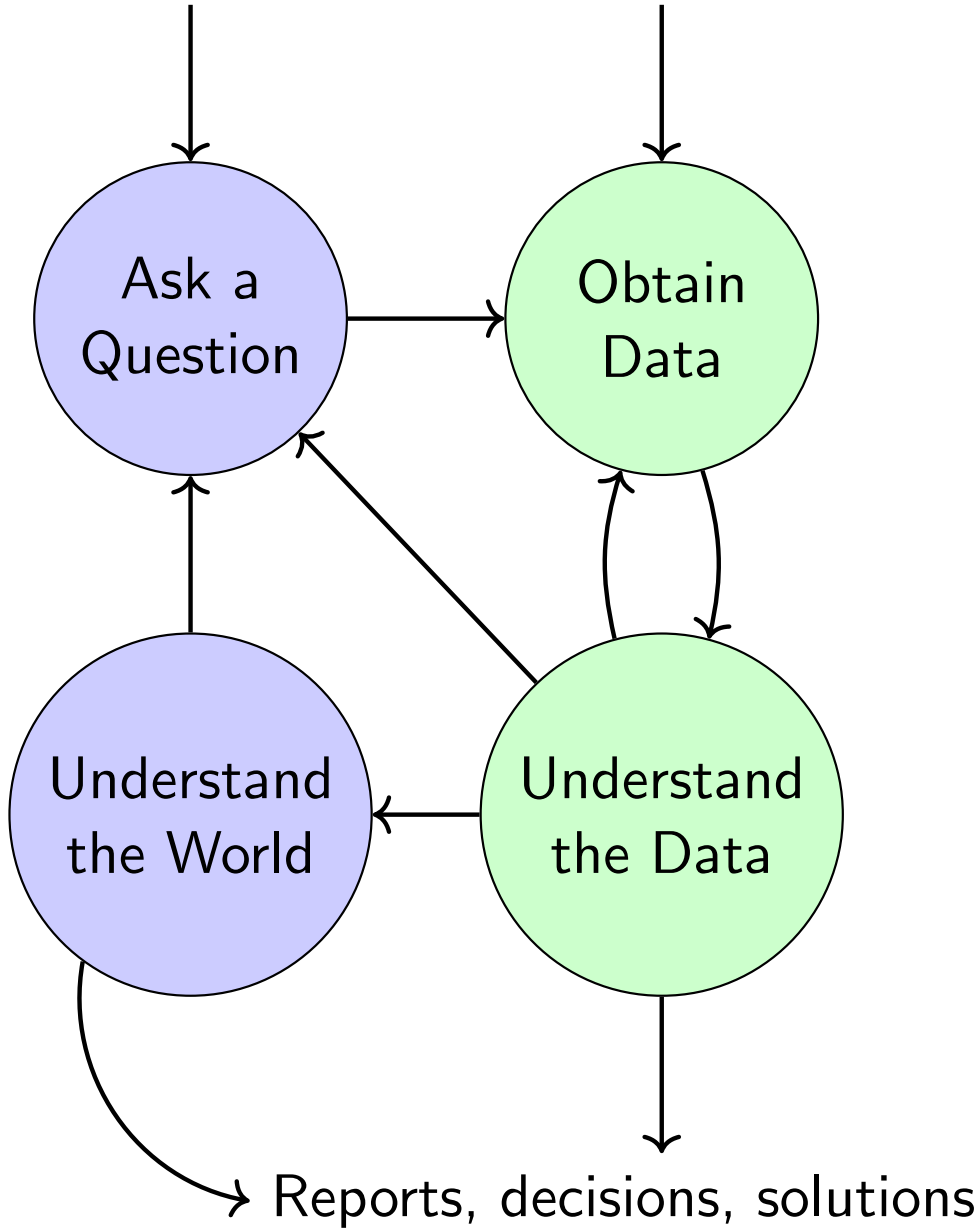
$\mathbf{I} = \mathbf{D} \cdot \mathbf{C}$

I = Information
D = Data
C = Context

Data
Raw elements, outside any context

10°C

Data Science in a Nutshell


Ask a Question
Obtain Data
Understand the World
Understand the Data
Reports, decisions, solutions

Why Questions Matter

Data without a question is just noise!

The purpose of data science is to **answer questions about the world**.

But any question might be addressable by data science.

Types of Questions in Data Science

Descriptive	What happened? How many patients have fever? What was the average sales revenue last quarter?
Exploratory	What patterns or relationships exist within the data? Is there correlation between physical activities and sleep? How do product ratings relate to review length?

Types of Questions in Data Science

Descriptive	What happened? How many patients have fever? What was the average sales revenue last quarter?
Exploratory	What patterns or relationships exist within the data? Is there correlation between physical activities and sleep? How do product ratings relate to review length?
Inferential	What generalizes beyond this dataset? Does a new marketing campaign increase average sales? Is there a statistically significant difference in test scores between two groups of students?

Types of Questions in Data Science

Descriptive	What happened? How many patients have fever? What was the average sales revenue last quarter?
Exploratory	What patterns or relationships exist within the data? Is there correlation between physical activities and sleep? How do product ratings relate to review length?
Inferential	What generalizes beyond this dataset? Does a new marketing campaign increase average sales? Is there a statistically significant difference in test scores between two groups of students?
Predictive	What will happen next? Will congestion occur tomorrow? What is the probability that a loan applicant will default.

Narrowing down a broad question into one that can be answered with data is a key element !

Understanding Data

Structured Data

Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. (See the lecture on relational databases and SQL.)

Examples of structured data format include:

- Relational databases (SQL)
- Spreadsheets (Excel, CSV)
- Tables in documents

Structured Data: Fields and Records

When data is structured, it can be seen as a collection of **records**, each record storing a **value** for a collection of **fields**. Each field has a specific data type (e.g., integer, string, date).

⚠ Do not confuse the field and its value: A field is a column in a table, while its value is the symbol stored in that field for a specific record.

Example: A table of students with fields such as name, age, and grade. Each student is a record in the table, and each field contains specific information about that student.

Name	**Age**	**Grade**	**Major**
Alice	20	18	Computer Science
Bob	22	12	Mathematics
Charlie	21	16	Physics

Semi-structured Data

Semi-structured data is a form of data that does not conform to a fixed schema or structure, but still contains some organizational properties that make it easier to analyze than unstructured data. It is often represented in formats that allow for flexibility in the organization of data elements.

With some processes, you can store them in a relational database (it could be very hard for some kind of semi-structured data), but semi-structured exist to ease space.

Examples of semi-structured data format include:

- JSON (JavaScript Object Notation)
- XML (eXtensible Markup Language)
- YAML (YAML Ain't Markup Language)
- Key-Value stores (e.g., Redis, MongoDB)

Unstructured Data: No Fixed Schema

Most of the time, semi-structured data is stored in NoSQL databases.

Example (Using YAML):

```
- Name: Alice
  Age: 20
  Grade: 18
  Major: Computer Science
- Name: Bob
  Age: 22
  Grade: 12
  Major: Mathematics
- Name: Charlie
  Age: 21
  Grade: 16
  Major: Physics
```

Unstructured Data

Unstructured data is a data which is not organized in a predefined manner, thus it is not a good fit for a mainstream relational database. It is typically stored in file systems and media storage systems. It requires specialized tools and techniques for processing and analysis, such as natural language processing (NLP) for text data or computer vision for image data (see the NLP and Computer Vision lectures).

Examples of unstructured data include:

- Text documents (e.g., Word, PDF)
- Multimedia files (e.g., images, audio, video)
- Social media posts (e.g., tweets, Facebook posts)
- Emails and messages
- Web pages and blogs

Unstructured Data: No Fixed Schema

Unstructured data does not have a fixed schema, meaning it does not have a predefined structure.

Example (Using a text document):

```
Alice is a student in Computer Science. She is 20
years old and has a grade of 18. Bob is a student
in Mathematics. He is 22 years old and has a grade
 of 12. Charlie is a student in Physics. He is 21
 years old and has a grade of 16.
```

⚠ The same information can be represented in different data formats.

Comparison of Data Representations

Repr.	**Structured**	**Semi-structured**	**Unstructured**
Format	Tables	Hierarchy	None
Example	SQL Tables, Excel, CSV	JSON, XML	Text, Audio, Images
Uses	Easy to store, analyze and query	Keeps some structure but with flexibility	Easy to aggregate but processing needed for analysis
Stored in	Databases or Spreadsheets	NoSQL, APIs	Files and Media Storage

Real-World Balance

In practice, most data is **semi-structured** (e.g., JSON) or **unstructured** (e.g., text, images).

These formats are harder to analyze directly.

A central challenge in data management is to **restructure** or **preprocess** data into usable formats.

Common Mistakes and Misconceptions

- Assuming all data can be directly loaded into tables
- Ignoring missing or inconsistent values
- Treating semi-structured data as structured
- Overlooking the need for preprocessing unstructured data
- Assuming file format implies structure (e.g., a ‘.csv‘ file may contain nested data in strings)

Trade-offs and Alternatives

- **Flexibility vs performance**: JSON is flexible but costly to query
- **Manual vs automated parsing**: Trade speed for robustness
- **Preprocessing time vs runtime complexity**: Flatten early vs query nested structures

Where does data come from?

Where Data Comes From?

Human-generated vs Machine-generated

- **Human-generated:** Created by people. Can be structured, semi-structured, or unstructured.
- **Machine-generated:** Produced automatically by devices or software. Mostly structured or semi-structured.

Where Data Comes From?

Human-generated vs Machine-generated

- **Human-generated:** Created by people. Can be structured, semi-structured, or unstructured.
- **Machine-generated:** Produced automatically by devices or software. Mostly structured or semi-structured.

Internal vs External

- **Internal:** Collected inside an organization. Usually controlled and accessible only to employees.
- **External:** Collected outside the organization.

Where Data Comes From?

Human-generated vs Machine-generated

- **Human-generated:** Created by people. Can be structured, semi-structured, or unstructured.
- **Machine-generated:** Produced automatically by devices or software. Mostly structured or semi-structured.

Internal vs External

- **Internal:** Collected inside an organization. Usually controlled and accessible only to employees.
- **External:** Collected outside the organization.

Open vs Proprietary

- **Open:** Freely accessible and reusable (e.g., under open licenses)
- **Proprietary:** Owned and restricted, (e.g., by copyrights)

Example: Customer Satisfaction Survey

A company collects a customer satisfaction survey, including numeric ratings and free-text comments.

Is the data human-generated or machine-generated? internal or external?

Example: Customer Satisfaction Survey

A company collects a customer satisfaction survey, including numeric ratings and free-text comments.

Is the data human-generated or machine-generated? internal or external?

- **Human-generated:** Collected from people.
- **Internal:** Collected within the company.

Example: Social Media Data

Public tweets about a new product are collected for sentiment analysis.

Is the data human-generated or machine-generated? internal or external?

Example: Social Media Data

Public tweets about a new product are collected for sentiment analysis.

Is the data human-generated or machine-generated? internal or external?

- **Human-generated:** People write the tweets.
- **External:** Collected outside the organization.

Example: Factory Sensor Data

Sensors in a factory record temperature and machine vibrations every minute.

Is the data human-generated or machine-generated? internal or external?

Example: Factory Sensor Data

Sensors in a factory record temperature and machine vibrations every minute.

Is the data human-generated or machine-generated? internal or external?

- **Machine-generated:** Automatically recorded by devices.
- **Internal:** Collected within the factory.

Example: Public Weather Data

Hourly temperature and humidity readings downloaded from a public weather station API.

Is the data human-generated or machine-generated? internal or external?

Example: Public Weather Data

Hourly temperature and humidity readings downloaded from a public weather station API.

Is the data human-generated or machine-generated? internal or external?

- **Machine-generated:** Automatically collected by sensors.
- **External:** Collected outside the organization.

Example: Open vs Proprietary Data

1: Government census dataset under an open license.
2: Company customer relationship management dataset.

Is the data human-generated or machine-generated? internal or external? Which is open and which is proprietary?

Example: Open vs Proprietary Data

1: Government census dataset under an open license.
2: Company customer relationship management dataset.

Is the data human-generated or machine-generated? internal or external? Which is open and which is proprietary?

- **Human- vs Machine-generated:** Census is made by people; CRM might include both.
- **Internal vs External:** Census is made public (external); CRM is mostly for the company (internal).
- **Open vs Proprietary:** Census is open; CRM is proprietary.

Some Questions

Why do you think enterprises do not always share their data?

Some Questions

Why do you think enterprises do not always share their data?

Can I use publicly available social media data without consent?

Some Questions

Why do you think enterprises do not always share their data?

Can I use publicly available social media data without consent? What if it was anonymized?

Some Questions

Why do you think enterprises do not always share their data?

Can I use publicly available social media data without consent? What if it was anonymized?

Can I use information leaked publicly?

Some Questions

Why do you think enterprises do not always share their data?

Can I use publicly available social media data without consent? What if it was anonymized?

Can I use information leaked publicly? after a cyberattack?

Some Questions

Why do you think enterprises do not always share their data?

Can I use publicly available social media data without consent? What if it was anonymized?

Can I use information leaked publicly? after a cyberattack? due to an internal error?

Some Questions

Why do you think enterprises do not always share their data?

Can I use publicly available social media data without consent? What if it was anonymized?

Can I use information leaked publicly? after a cyberattack? due to an internal error?

Where data comes from affects its reliability, the right to exploit it and its potential biases.

Ethics in Data Sciences

Ethics in Data (Preview)

Data is never neutral. Ethical considerations arise in multiple dimensions:

Bias	Pre-existing social, cultural, or technical biases embedded in data.
Privacy	Collection, storage, and use of personal data; anonymization; GDPR
Ownership	Who has rights to the data; commercial vs public

Dedicated ethics session later

Cambridge Analytica (2018)

Background: Cambridge Analytica, a political consulting firm, harvested data from millions of Facebook users without their knowledge.

How it happened:

A personality quiz app ("thisisyourdigitallife") collected data from users.

Facebook's API at the time also allowed access to the quiz takers' friends' profiles.

Estimated 50–87 million user profiles were harvested.

Key issue: Users never gave informed consent.

Use of the Data

Built detailed psychological profiles of voters (e.g., introverted vs extroverted, politically leaning).

Used microtargeted ads during political campaigns (e.g., Brexit referendum, 2016 U.S. presidential election).

Created personalized messages designed to influence behavior and opinions.

Key issue: Manipulation of political processes through opaque algorithms.

Impact and Consequences

Consequences:

Public outrage and decline in trust toward Facebook.
5 billion fine by the U.S. Federal Trade Commission (FTC).
Strengthened calls for regulation (GDPR, data protection laws).
Facebook restricted API access and tightened data policies.

Lessons:

- Privacy breaches can have societal-level consequences.
- Data collection practices must be transparent.
- Informed consent and accountability are central to ethical data science.