Data, Data Storage, Data Collection

Lecture 12: Recap

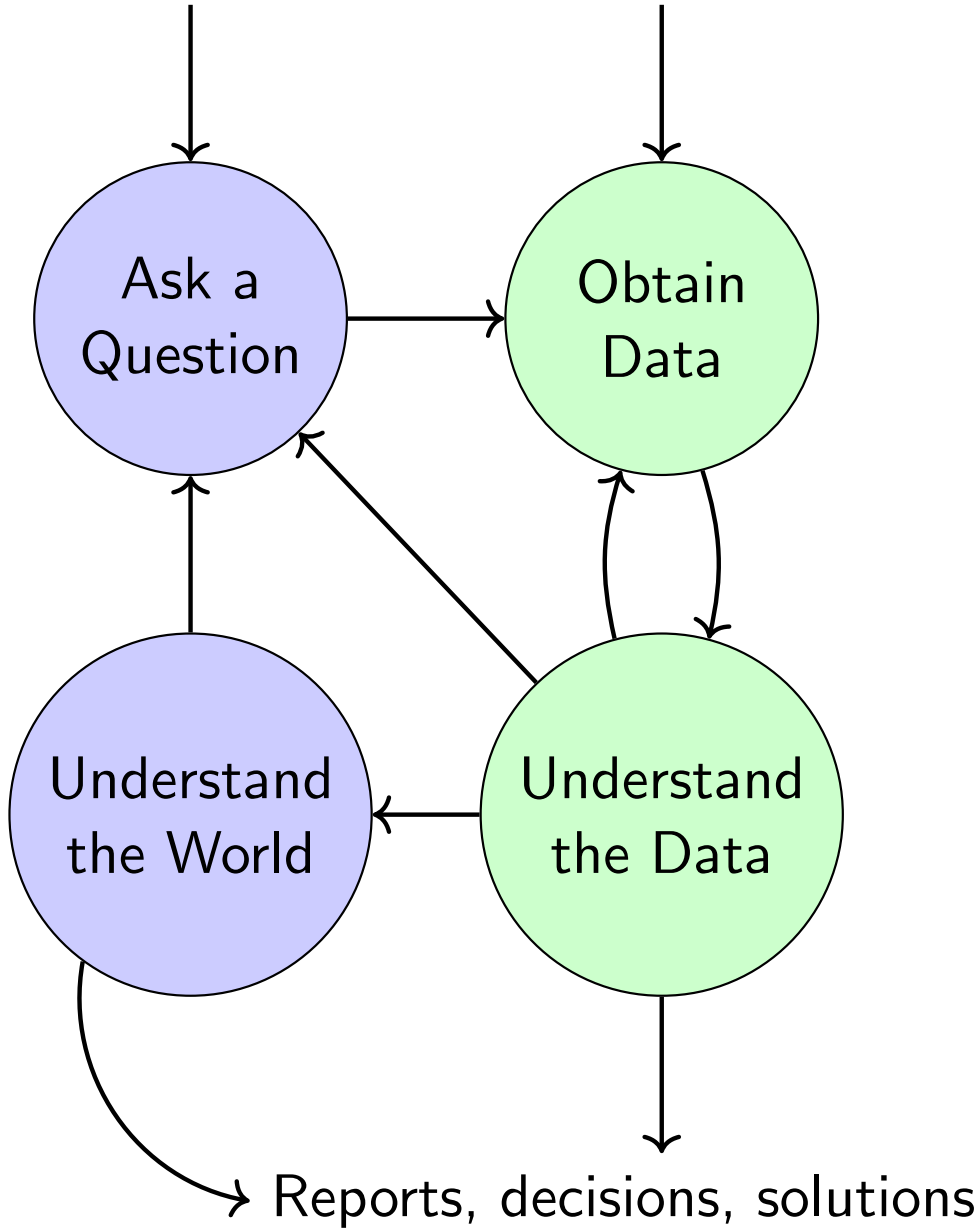
Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

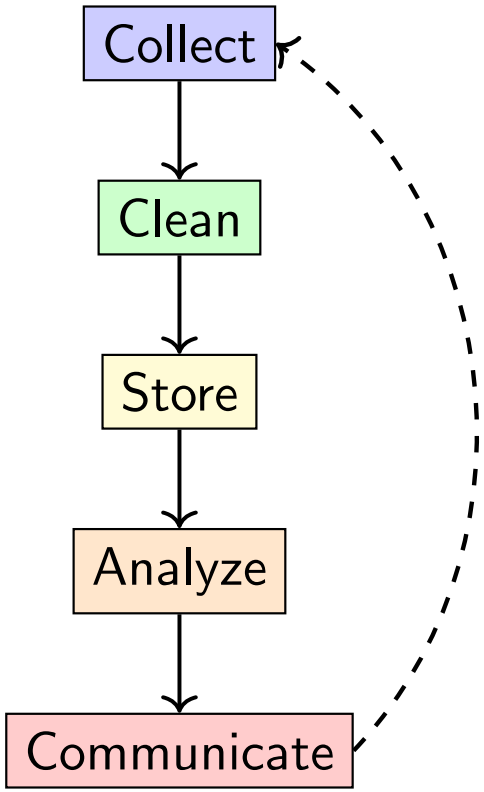
Session Objectives

The goal of today is to:

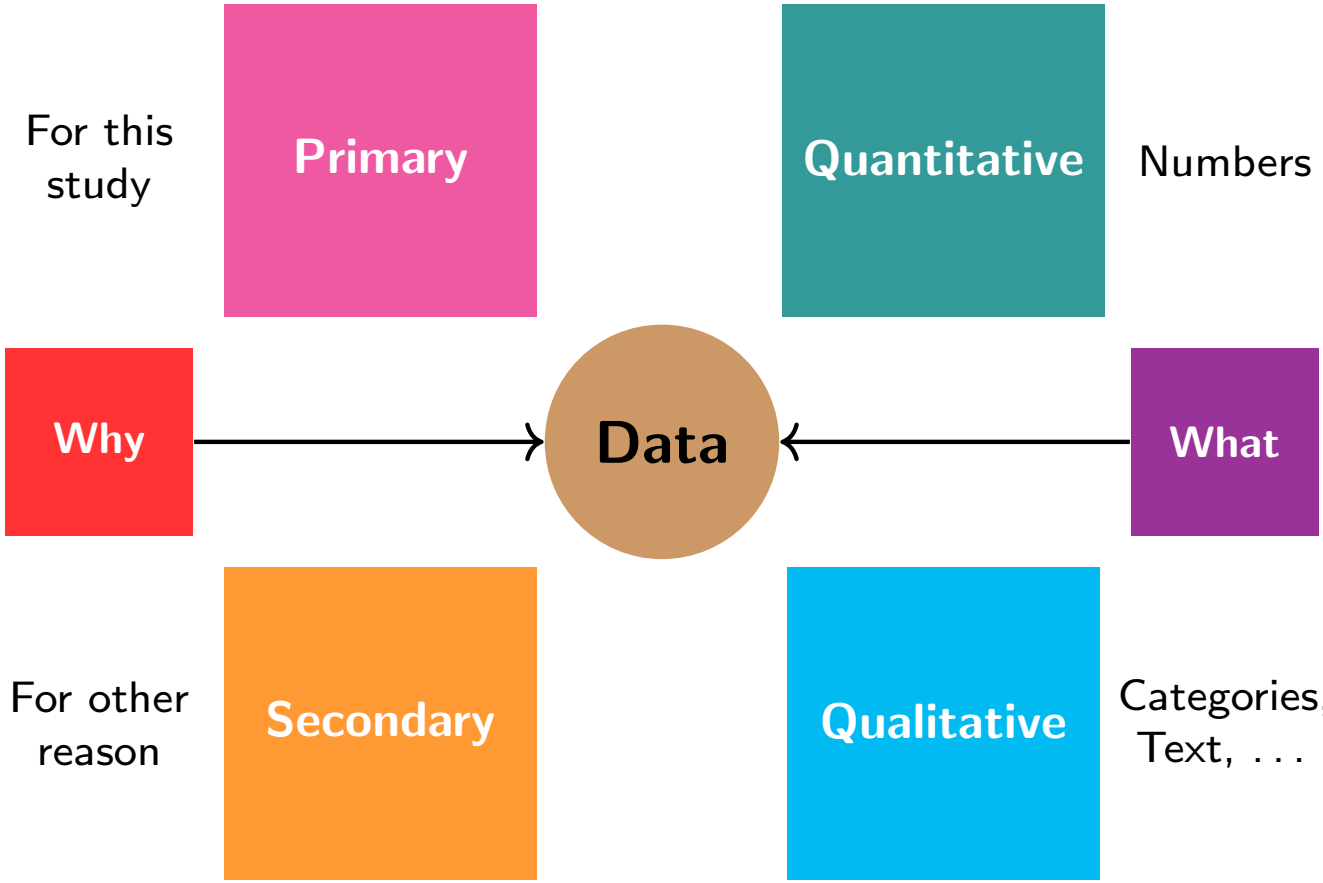
- Connect the dots across the data lifecycle
- Highlight essential concepts to retain long-term
- Prepare for the exam


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

The Lifecycle is a Loop



The process is **iterative**.


For this study
Primary
Quantitative
Numbers
Why
Data
What
For other reason
Secondary
Qualitative
Categories, Text, . . .

Overview of Data Collection Methods

Methods depend on the **study aim** and **type of data required**.

Experiments & Simulations

When to use: Test causal relationships

Surveys

When to use: Opinions, characteristics

Sensors & Logs

When to use: Quantitative, continuous data

APIs & Secondary Data

When to use: Use existing external data

Common Biases in Data Collection

Non-response Bias Participants who do not respond may differ from those who do (e.g., unhappy customers skip surveys).

Response Bias Participants answer in a way they think is "desirable" (e.g., overreporting healthy behaviors).

Selection Bias Certain groups are overrepresented (e.g., only highly motivated individuals respond).

Undercoverage Bias Some groups are excluded from the sample (e.g., only surveying students who attend lectures).

Data Quality

Definition (Data Quality)

How **fit for purpose** the data is for its intended use.

1. **Validity**: Does data conform to rules?
   - E.g., correct data types, or ranges.
2. **Accuracy**: Does data reflect the true value?
   - Hard to verify! E.g., self-reported survey data.
3. **Completeness**: Are all required values present?
   - Missingness often stems from **poor collection**.
4. **Consistency**: Is data equivalent across systems?
   - E.g., customer IDs matching in CRM and sales databases.
5. **Uniformity**: Are units standardized?
   - E.g., kg vs. lbs

Data Quality is further specified by the **ISO 25012** standard.

Strategies for Handling Missing Data

Deletion

- Simple but may introduce bias
- Only if features are mostly empty or irrelevant

Imputation

- Statistical methods
- Forward or backward fill
- Prediction

Flagging

- Add a binary column to indicate missingness.

You need to understand why data are missing, not just how many values are missing.

Data Transformation

Min-Max Normalization

Scales feature to $[\mathbf{0}, \mathbf{1}]$:

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)}$$

Use Case: Known data bounds.

Z-Score Standardization

Scales feature to $\mu = \mathbf{0}$; $\sigma = \mathbf{1}$:

$$x' = \frac{x - \mu}{\sigma}$$

Use Case: Gaussian-distributed data.

Label Encoding

Assigns integers to categories

Use Case: Ordinal categories.

⚠ False ordinality

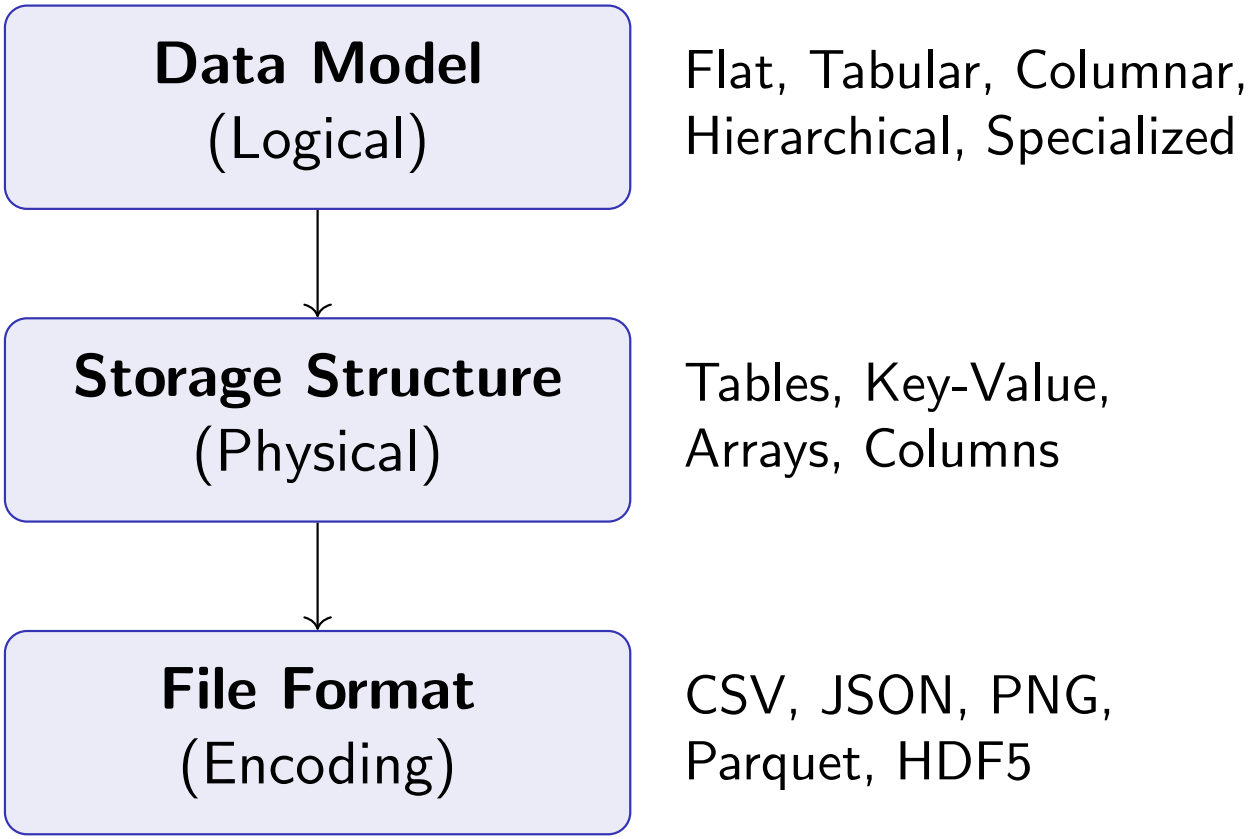
One-Hot Encoding

Creates binary columns

Use Case: Nominal categories.

⚠ High dimensionality

From Models to Formats



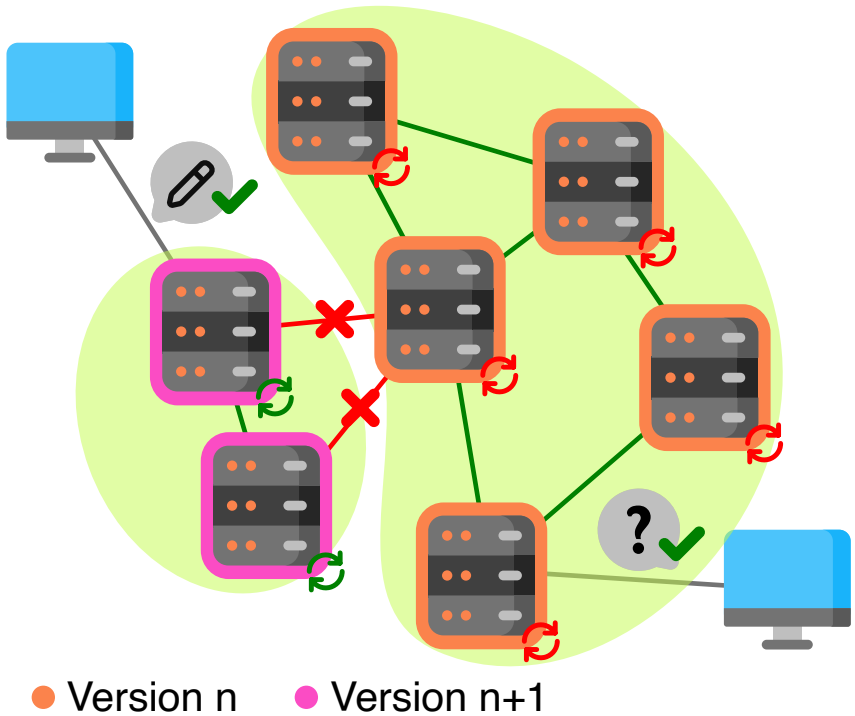
- **Data model** defines the logical structure.
- **Storage Structure** defines the physical organization.
- **File format** defines the encoding.

CAP Theorem

Theorem (CAP)

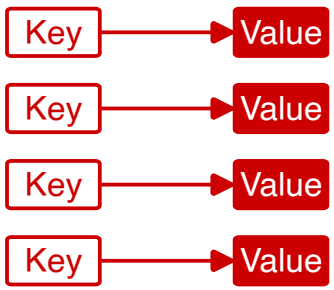
Any distributed system can guarantee **at most two** *of:*

- **C***onsistency: All nodes see the same data.*
- **A***vailability: Every request gets a response.*
- **P***artition Tolerance: System works despite network failures.*

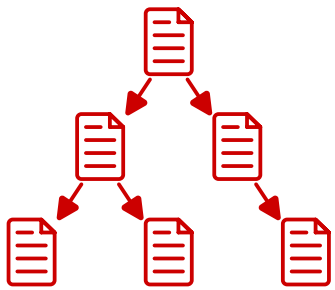


The purpose is to choose a trade-off under network trouble.

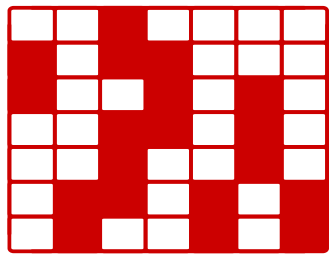
Four NoSQL Families



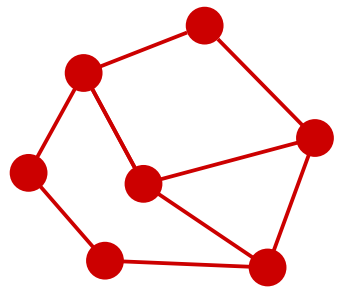
Key-Value



Document



Column



Graph

⚠ NoSQL databases **complement** SQL (polyglot persistence).

Aggregate: The Basic Unit of NoSQL

Definition (Aggregate)

A **collection of data** that interacts as a single unit.

```
{
  "user_id": 123,
  "name": "Alice",
  "balance": 1000,
  "orders": [
    {"order_id": 1, "amount": 99.99},
    {"order_id": 2, "amount": 49.99}
  ]
}
```

Data Normalization and Normal Forms

Definition (Normalization)

Process of organizing data into tables to **eliminate redundancy** and **prevent inconsistencies**.

ExpenseID	Expense	Paid by	Amount
9571	Internet	Alex	40
9571	Internet	Bill	40

⚠ The same fact appears twice with **conflicting values**.

Normal Forms (Edgar F. Codd)

- **1NF**: No repeating group.
- **2NF**: No partial dependencies.
- **3NF/BCNF**: No transitive dependencies.
- **4NF**: No multivalued dependencies.
- **5NF**: No join dependencies.

Decompose tables into **smaller, related tables**, so each table represents **one specific topic**.

Denormalization

Definition (Denormalization)

Intentional violation of normal forms to **accept redundancy** or **inconsistency risk**.

Practical Considerations

1. Data arrives from a **non-normalized system**.
2. Business rules might **evolve**.

Performance Considerations

3. To **optimize query speed** (e.g., pre-joined or duplicated data).

- Best suited for **read-heavy contexts**.
- Can violate **any normal form** (1NF to 5NF).

Exploratory Data Analysis (EDA)

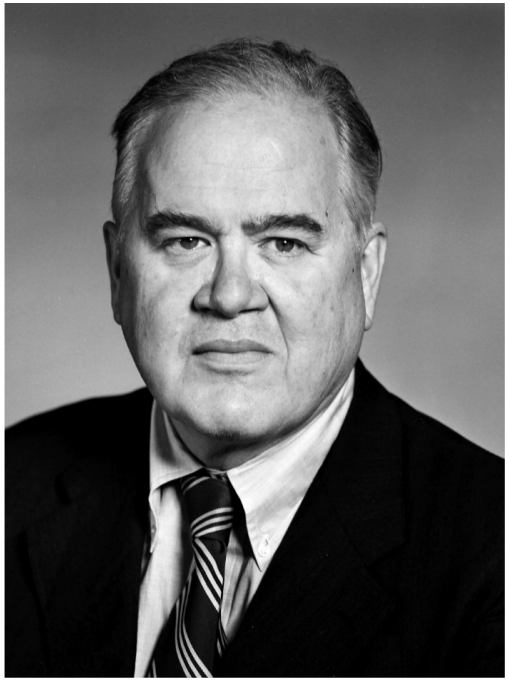


Photo Credit: Princeton University, Robert Matthews

John Tukey (1915-2000)

"EDA is **actively incisive**, rather than passively descriptive, with real emphasis on the **discovery of the unexpected**."

EDA relies on:

Summary statistics and **visualizations**

To discover:

Patterns, **structure**, and **anomalies**.

Questions Driving EDA

- How is Feature X distributed?
- Any unusual values in X?
- Are transformations helpful?
- How do X and Y relate?
- Does X behave differently across Z?

Summary Statistics

Measures of Center

Summarize data with a **single value**:

- Mean
- Median
- Mode

Measures of Spread

Summarize how **variable** the data is:

- Range
- IQR
- Standard Deviation

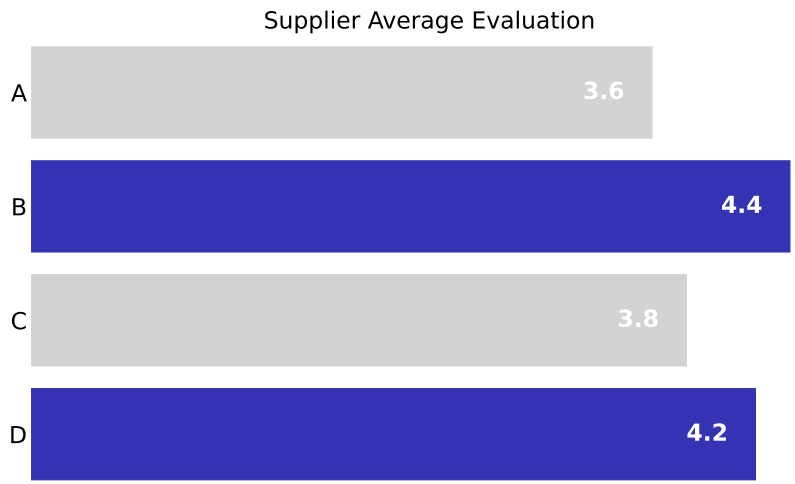
Key Insight

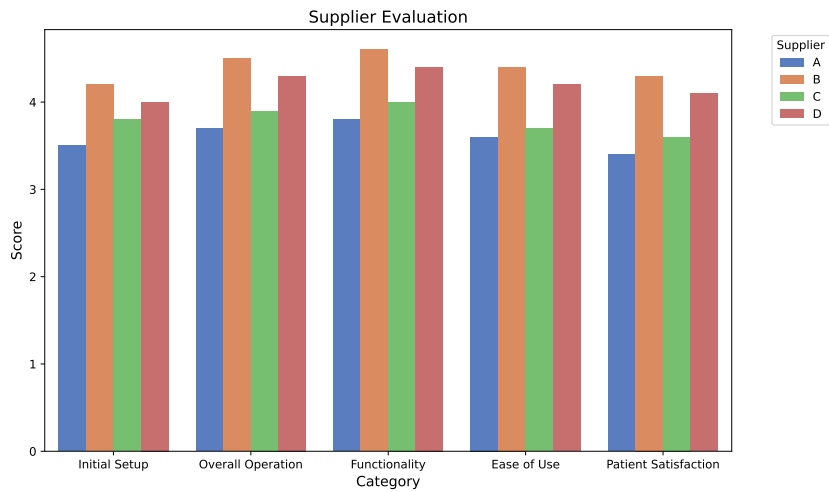
Measures of center and spread together provide a **minimalistic picture** of data distribution.

The Ruthless Minimalist Approach

Principles

- Remove all non-essentials
- Use clean, simple designs
- Maximize data-ink ratio
- Avoid chart junk





Data Storytelling Framework

1. Set the context
   "Our customer churn has increased by 15% this quarter."
2. Ask the question
   "What factors are contributing to this increase?"
3. Reveal evidence incrementally and logically
   "60% of churned customers had contacted support 3+ times."
4. Reveal the insight
   "This suggests customer service issues are a major driver of churn."
5. Recommend action
   "We recommend a customer success program for at-risk accounts."

What You Should Be Able To Do Now

- Identify and describe a data lifecycle
- Plan a small data collection strategy
- Evaluate data quality and clean a dataset
- Choose appropriate storage models and formats
- Normalize and denormalize a schema
- Analyse a dataset and communicate insights

Takeaways

1. Data has a lifecycle: every step requires decisions and trade-offs.
2. Quality, structure, and ethics matter as much as quantity.
3. Good communication is the final step: insights only exist if shared clearly.