Data, Data Storage, Data Collection

Project Guidelines

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

Project: Reproducible Data Workflow

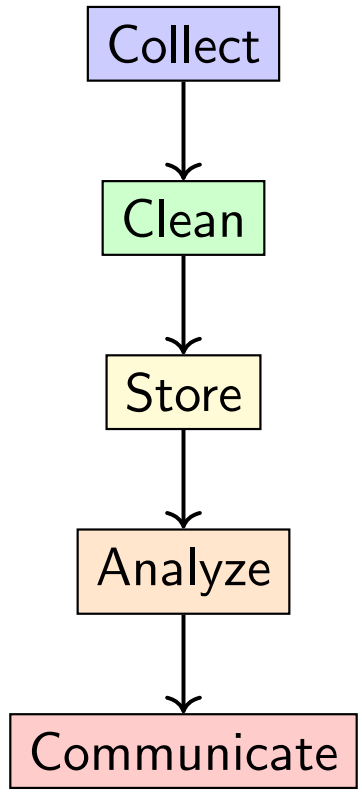
Start from a **question**, and build a full workflow to answer it: collection, cleaning, storage, analysis, and communication over **real data**.

This is your chance to bring together concepts from class in a practical, open-ended task:

- Groups of **five**.
- You choose the topic, the question, the data, and the technical approach

Goal

1. Start from a question, then acquire the data you need. **At least one source must be data you acquire or generate yourselves**. Proprietary data is not allowed.
2. Apply the **lifecycle steps**: collect, clean, store, analyze, communicate. Even choices should be documented and explained..
3. The outcome should not just be numbers: your project should **tell a story** and generate new insights. An inconclusive answer is fine if that's what the evidence supports.



What You Should Do

Collection : acquire or generate data yourselves (at least one source).

Cleaning : diagnose quality issues and solve them.

Storage : choose a format and justify your decision by compare at least two storage options against your data's needs.

Analysis : answer your question with evidence but observation from interpretation.

Communication : produce visualizations and a short written report.

What You Should Not Do

You should not **start from an existing** analysis-ready **dataset**.

You are not expected to **build complex machine learning models**.

You are not expected to use **advanced statistics** beyond what we cover in class.

The emphasis is on careful data handling, transparent decision-making, and clear communication.

Focus on **reproducibility and clarity** rather than **complexity**.

Process & Milestones

1. **16.09**: Project proposal
2. **29.09**: Raw data acquisition (1^{st} peer-review checkpoint)
3. **06.10**: Peer-review round 1 due
4. **03.11**: Intermediate version with a least cleaning done (2^{nd} peer-review checkpoint)
5. **10.11**: Peer-review round 2 due
6. **27.11**: Final deliverables
7. **02.12**: Group presentations

Peer Review

At two points in the term (29.09 and 03.11), you will review two other groups' work-in-progress using a structured review form.

The goal is **not** to grade the other group but to help them identify:

- something particularly well done,
- something that should be improved,
- a question or issue that remains unclear.

You will also have reflect on the feedback **you** receive and explain what you changed as a result.

Review quality, and the quality of questions you ask at final presentations, count toward your grade.

Project Proposal: 16.09

Each group will submit a short one-page proposal.

The proposal should include:

- **Group composition**,
- The central **question** you will explore,
- A description of the **data** you expect to need,
- **Challenges** you anticipate.

The proposal fixes the scope of your work and must to be approved (by me).

Final Deliverables: 27.11

Each group will submit

- a **report** of up to 5 pages
- the **code** (Jupyter notebook or runnable software project) that can reproduce your analysis.

Besides presenting the work, the report should offer a **reflection** on the various steps of the data lifecycle, putting forward the **difficulties** encountered and the **lessons** learned.

Group Presentation 02.12

Each group presents for 15 minutes, followed by questions.

The presentation, report, and code must be consistent.

The presentation also has to be submitted.

This is your chance to share results and highlight the most interesting aspects of your project.

Evaluation Criteria

Evaluated on the **appropriateness and justification** of your choices and not technical complexity for its own sake.

- Understanding and application of the data lifecycle.
- Clarity and reproducibility of the codebase.
- Quality and readability of the report and presentation.
- Teamwork and fair contribution.

Evaluation Criteria

Evaluated on the **appropriateness and justification** of your choices and not technical complexity for its own sake.

- Understanding and application of the data lifecycle.
- Clarity and reproducibility of the codebase.
- Quality and readability of the report and presentation.
- Teamwork and fair contribution.
- Quality of the peer reviews.
- Quality of the questions asked to peers during their presentations.

Support

If you have questions:

- Ask at the end of lectures,
- Or contact me by email.

It is normal to refine your project idea as you go.

Use feedback from peers and from me to improve your work.

The aim is not perfection, but a clear, well-structured process.