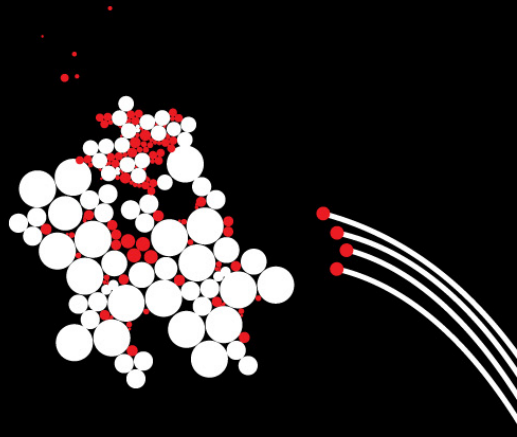
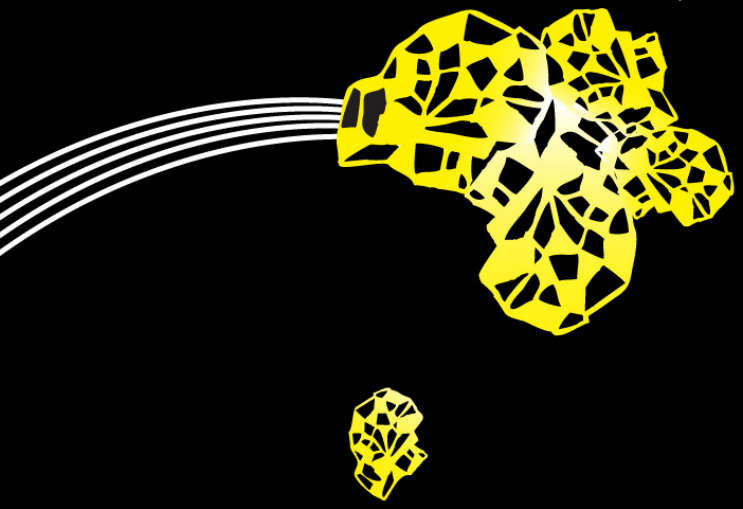
UNIVERSITY OF TWENTE.



M-SE Q2 DATA MASTERY CHALLENGE

THE WHY AND THE HOW – RATIONALE, LOGISTICS, AND ORGANIZATION

FRANK OSTERMANN, COURSE COORDINATOR

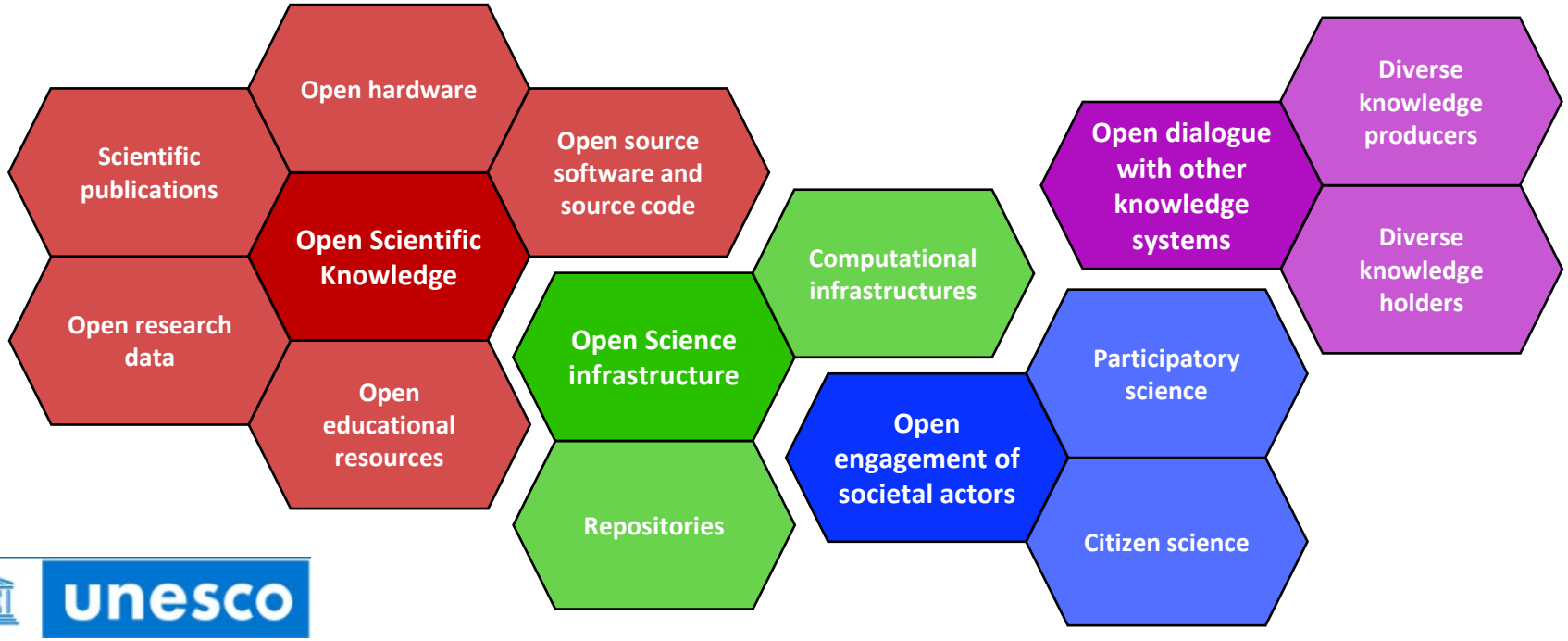


OVERVIEW

- Topic
- Learning objectives
- Schedule
- Milestones
- Assessment and test plan
- Canvas structure
- Course policies

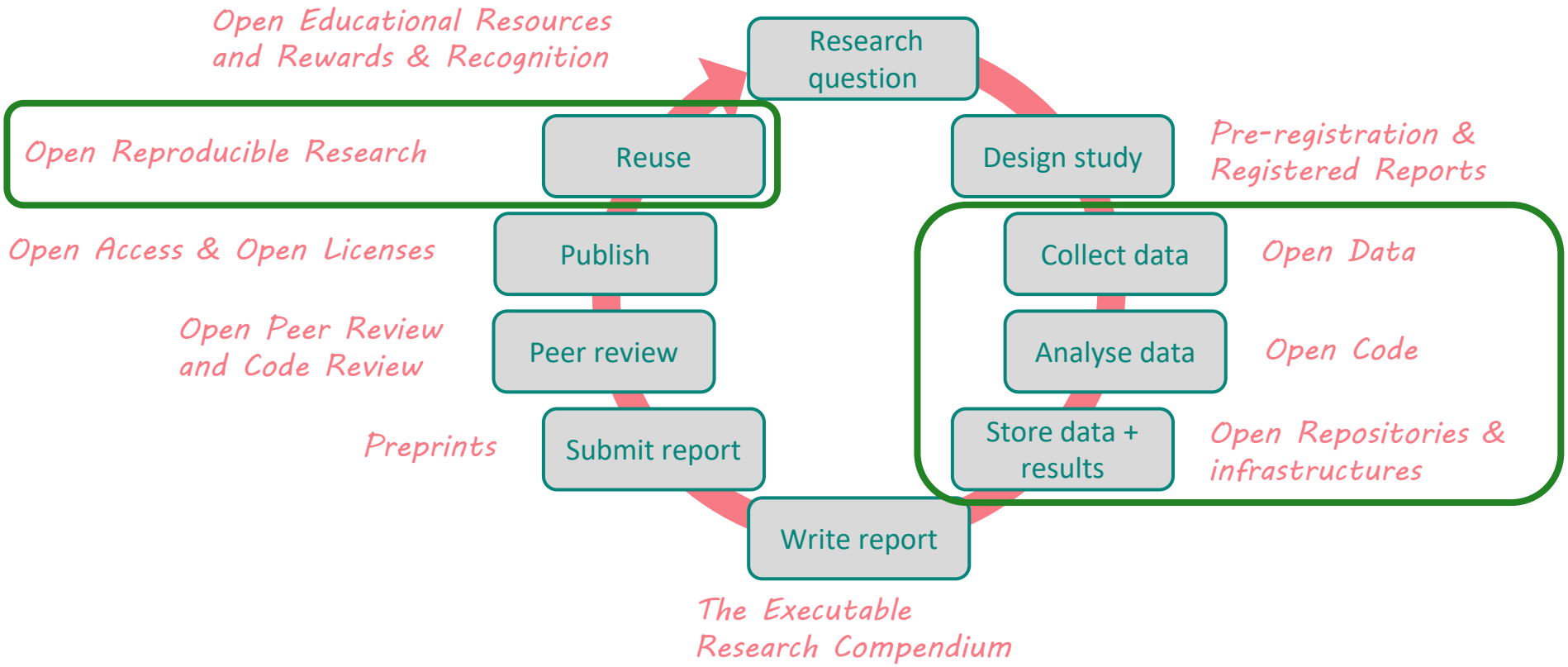
THE BIG PICTURE: OPEN SCIENCE

What is Open Science, and why does it matter?



unesco

OPEN AND REPRODUCIBLE RESEARCH – CHALLENGE!



WHY “DATA MASTERY”

“4th Paradigm”: Data (driven) science

+

Heterogeneous (geospatial) data

+

Open (and reproducible) research

=>

Need to master data {handling, engineering}

DATA MASTERY – CHALLENGE!

What types of data do you know?

- Text, numbers, audio, video
- Static, dynamic/streaming
- XLSX, CSV, TSV
- (Geo)JSON, Shapefiles, GML
- DOCX, PDF
- DBMS (relational, document, key-value, …)
- Parquet

WHY “CHALLENGE”

You get to work on

- Real research problems
- Develop hands-on skills based on solid knowledge

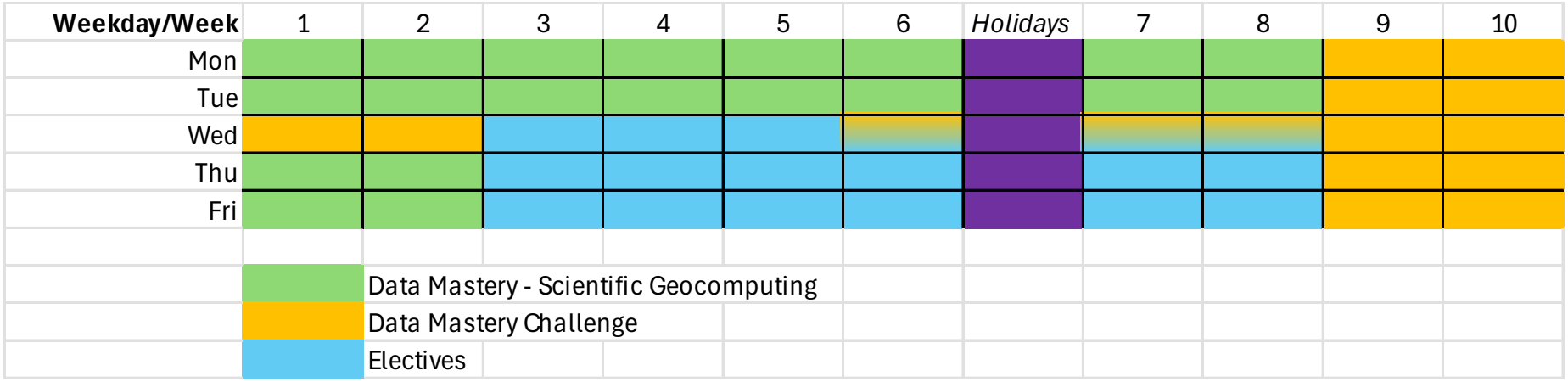
LEARNING OBJECTIVES

LO1 Develop a conceptual analysis workflow and data management and sharing plan

LO2 Share a reproducible package of the implemented workflow, containing all required data and code and sufficient documentation

LO3 Reflect on your approach and implementation, and evaluate the degree of success of your reproduction or replication

SCHEDULE – THE BIG PICTURE



SCHEDULE

Week	Date (week)	Content
01	13 Nov	Course Intro: Data Mastery as prerequisite for open and reproducible research; skill survey Open science, reproducibility and replicability in the geosciences; best practices
02	22 Nov	Assessing a study's reproducibility; finding/choosing a challenge ***Introduction of challenges***
06	18 Dec	Recap best practices & discussion, preregistration as a planning tool ***Choice of challenge, group formation***
07	08 Jan	Work session on preregistration ***Hand in preregistration***
08	15 Jan	Interoperability, reproducibility tools (Github, data citations, markdown, project sharing) ***Peer feedback on preregistration***
09	20 Jan	Question hour
09	22 Jan	***Mid-term group presentations***, peer feedback
10	27 Jan	Brief group reporting, peer feedback, tool support
10	29 Jan	***Final group presentations***, CodeCheck, and wrap up
10	31 Jan	*H**and in project***

ASSESSMENT AND TEST PLAN

- There is only one deliverable: the project package, a.k.a. final assignment
- However, presentation of your project is mandatory (not graded, but without it, fail mark)
- All group work, except for a short individual reflection part.

CANVAS STRUCTURE

- Home page: General information, static
- Announcements: Important notifications, **make sure to get/read them**!
- Syllabus: Schedule, materials, rubric, grading
- Discussions: main mode of asking questions outside of face-to-face sessions

COURSE POLICY =! COURSE POLICE

- You are here to learn – I will do my best to give you lots of opportunity, but I will not try to force you…
- Academic integrity: No plagiarism (and that includes undocumented use of generative AI, more on this next slides)
- Class attendance is expected, please communicate if you can't attend
- No Canvas mail! Please use regular e-mail for communication

A WORD OR TWO ON CHATGPT AND FRIENDS

ChatGPT is an awesome tool for professionals, i.e., those who have already the necessary skills and knowledge, to become more productive.

It is a problematic tool for those who want to become professionals.

Remember:

You haven't understood something until you can explain it in your words!

OR, AS THE SYLLABUS SAYS

This course follows the general UT rules on generative AI in education, outlined at AI In Education | Learning & Teaching Portal (utwente.nl). However, those are by necessity quite generic. Below the concrete rules that govern this course in addition to the general UT rules:

- You are allowed to use LLMs/genAI such as ChatGPT
- Just like any other source that is not you, you need to reference anything and everything that you copy and paste or even paraphrase from LLM queries.
- You can do that in the same style as if you were referencing a conversation with an expert, e.g., with a footnote and "ChatGPT communication, 11.11.2222; for details, see Appendix X".
- Note the appendix: When you use (parts of) a genAI answer, you need to document the entire process completely. That means, every (and I mean every) query and response in that conversation (from which you use the output). Or, if you use the API, the exact query with all parameters. If possible, use options like a seed parameter or fingerprint values (compare How to make your completions outputs consistent with the new seed parameter | OpenAI CookbookLinks to an external site.)
- Do NOT rely on LLMs/genAI for summaries, evaluations, or other text for your report. You need all the training in analytical thinking and argumentative skills that you can get. For all the crucial moments in your life, there will ALWAYS be a face-to-face component happening in real-time. If you don't train your thinking and argumentations skills, you will do worse in those situations.

AND NOW FOR SOMETHING COMPLETELY DIFFERENT

… WELL, NOT REALLY

- Let’s do the test!