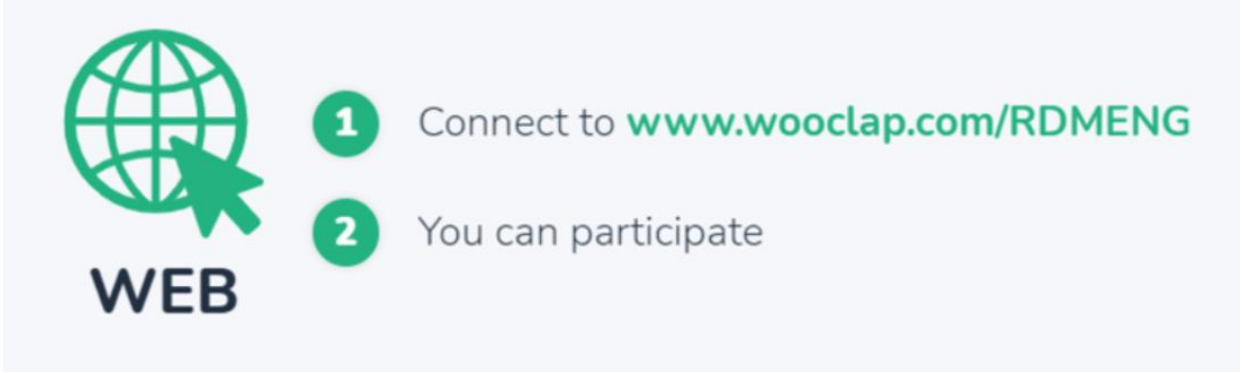
RESPONSIBLE RESEARCH DATA MANAGEMENT

Welcome!

This session is using Wooclap. Please open the Wooclap event page (see left panel), answer the test question, and keep the Wooclap page open during the whole training.



RESPONSIBLE RESEARCH DATA MANAGEMENT

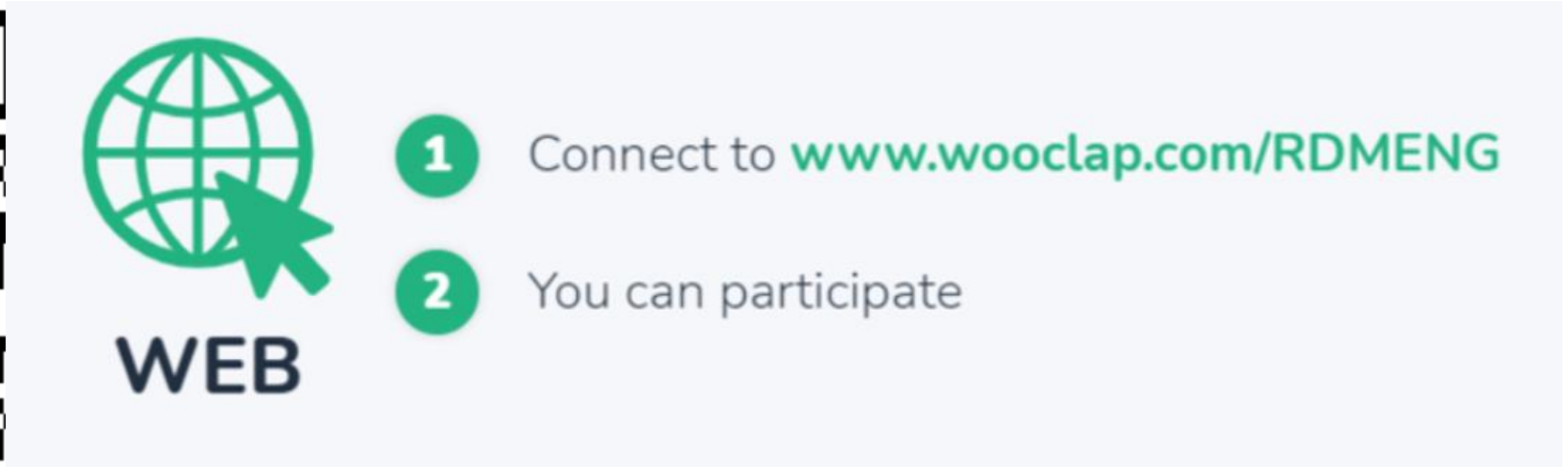
Euraxess Webinar Trainings for PhD Candidates
28/03/2022

Judith Biernaux, Dr.
Research Data Officer
Place du XX Août, 7 (Bât. A1)
B-4000 Liège
+32 (0)4 366 55 14
jbiernaux@uliege.be

euraxess
RESEARCHERS IN MOTION

LIÈGE université

What is research data?



What are we going to discuss today?

What is research data?

Factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.

Recommendation of the Council concerning Access to Research Data from Public Funding (OECD, 2006)

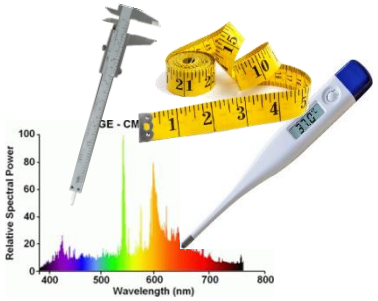
Problems with this definition:

- Only takes into account data with a digital format and excludes physical data (samples, paintings, manuscripts, papers, ...)
- Stems from a sharing point of view, not a user point of view

-> Needs to be broadened

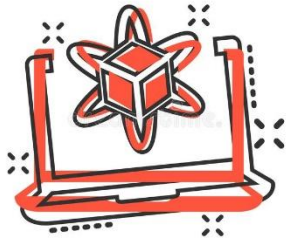
What is research data?

Documents, tables, maps

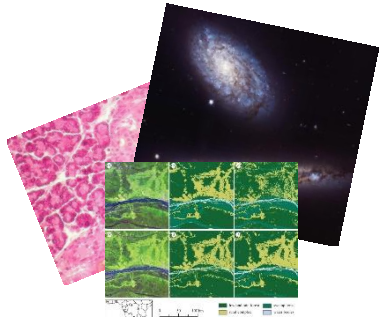


Experimental results

Polls, forms

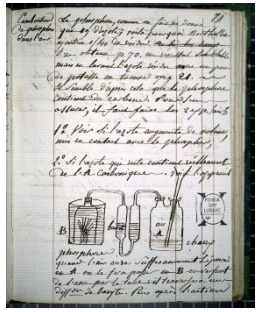


Simulation results



Images, videos

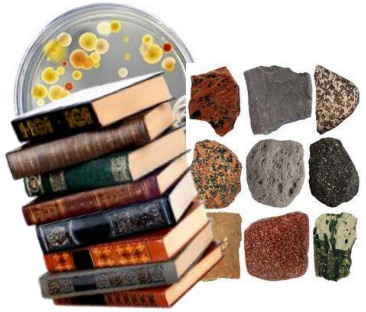
Audio files



Lab notes, field work notes

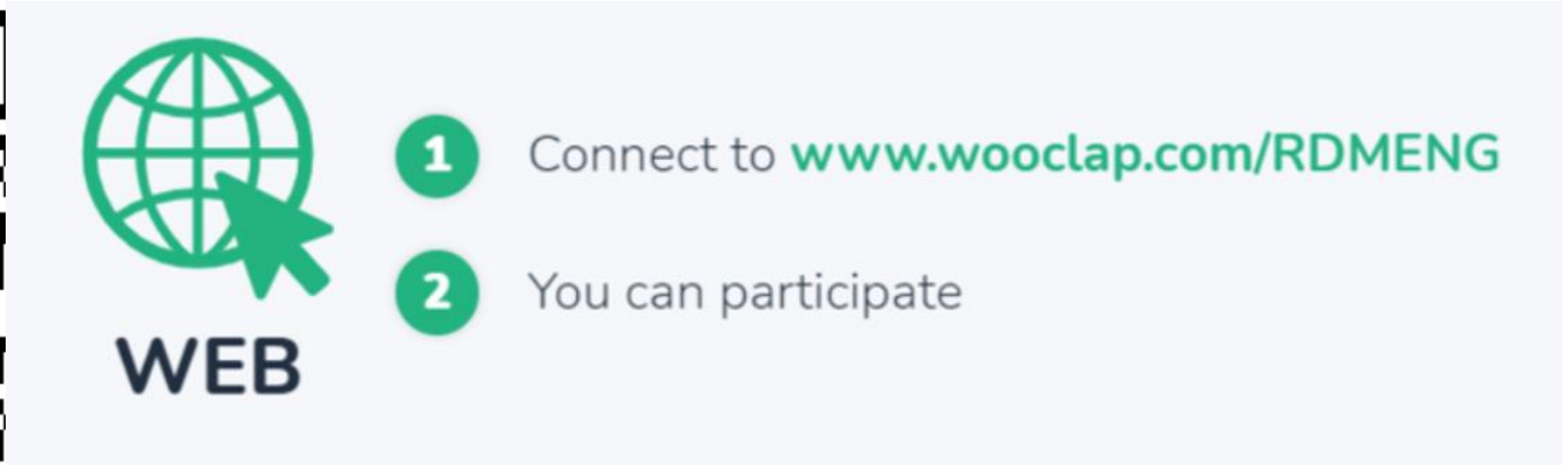
DNA sequences

Papers, publications



Physical samples

What is research data?



Is your data qualitative, quantitative, a bit of both, or are you unsure?
Write down a few words describing your data

What is research data?

Data are at the **core of your research:**

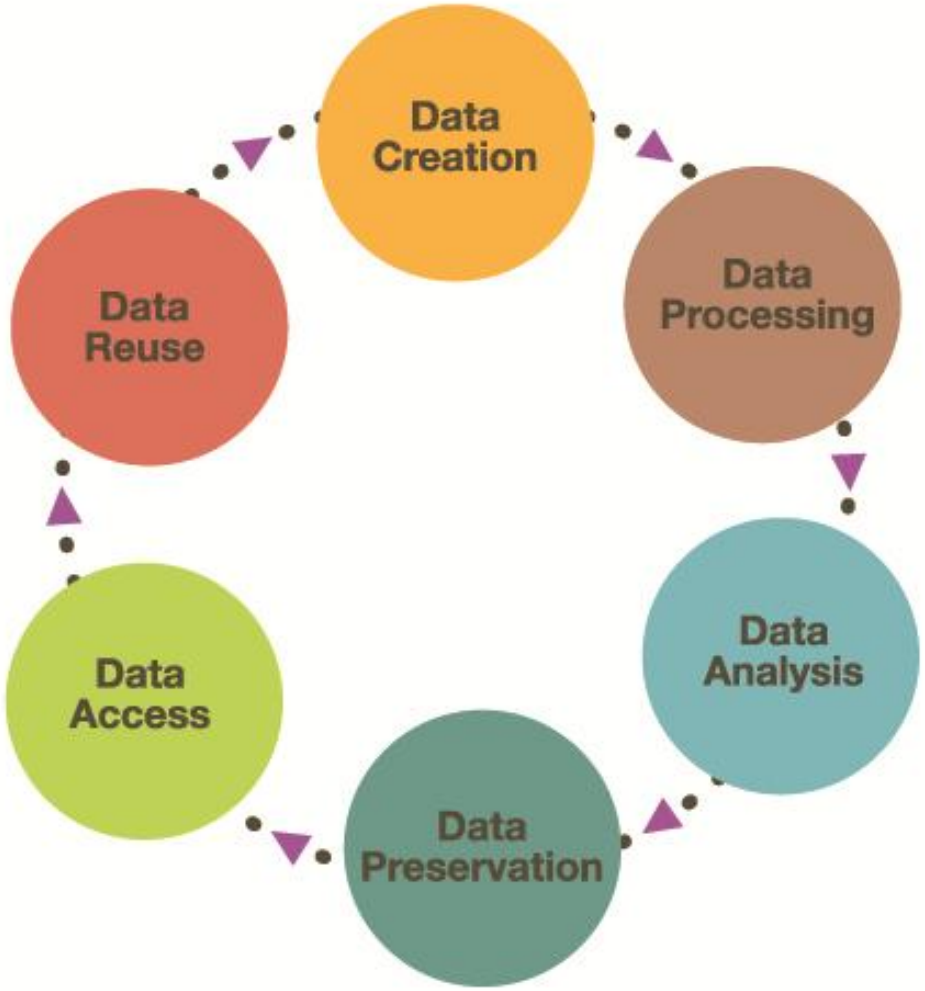
-> they enable the research process (answering your research question)
-> they support or disprove hypotheses
-> they usually contribute to the choice of methodology
-> they may have an impact on the quality of your results
-> they sometimes carry an economical value

They go through all steps of a research project and ought to go through processes of **high quality**

-> research data management

What does responsible research data management mean?

Data Life Cycle



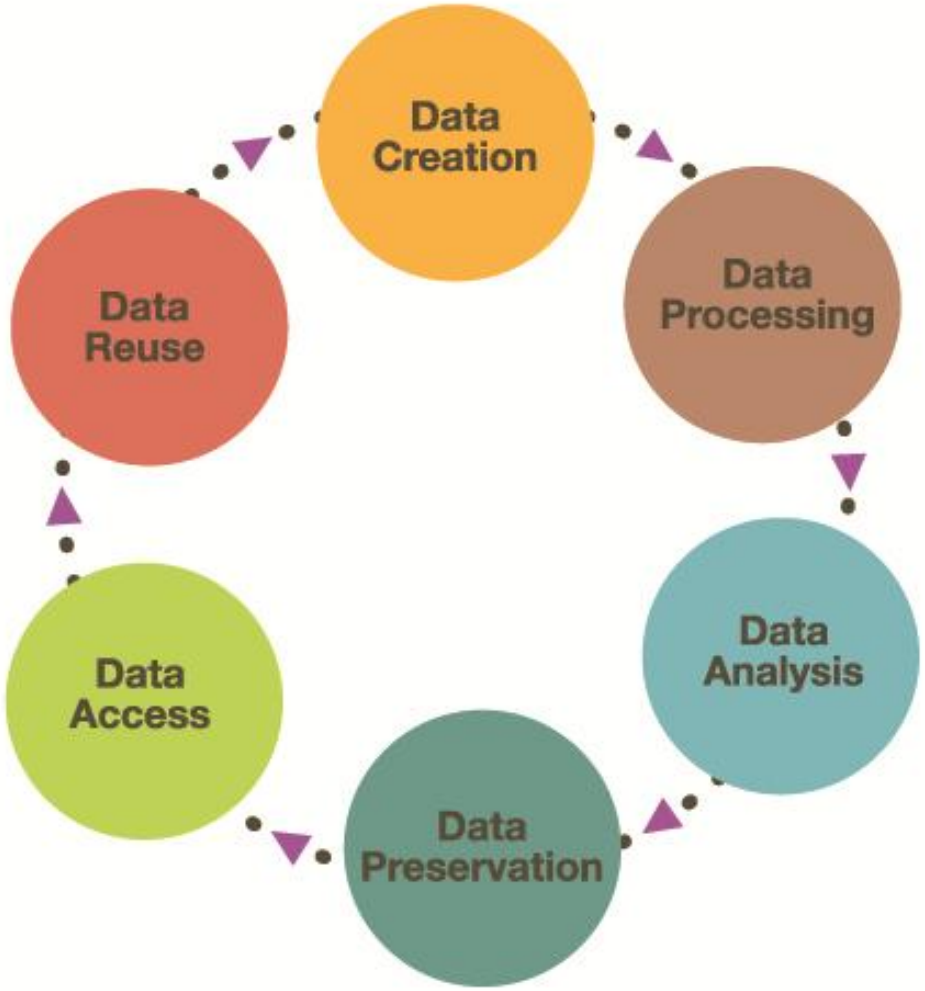
Set of practices around research data, including but not limited to:

-> collecting (first / second hand)
-> storage, curation
-> documentation
-> formatting
-> filtering, sampling
-> analysis
-> publishing, sharing

Responsible data management:
-> Adopting **good habits** for each of these tasks, so that research data get easier to use, to share, and to re-use

What does responsible research data management mean?

Data Life Cycle



EARLY ON : DATA PLANNING

What is my data like? What are the applicable regulations? Who do they belong to?

Nature, format, volume, source, collection, access...

DURING : DATA HANDLING

How should I store them?

Safety, size, security, backups, documentation...

TOWARDS THE END: DATA SHARING

How should I share my data? What happens to my data after my project is over?

Open and FAIR data, licences, data sustainability and re-use

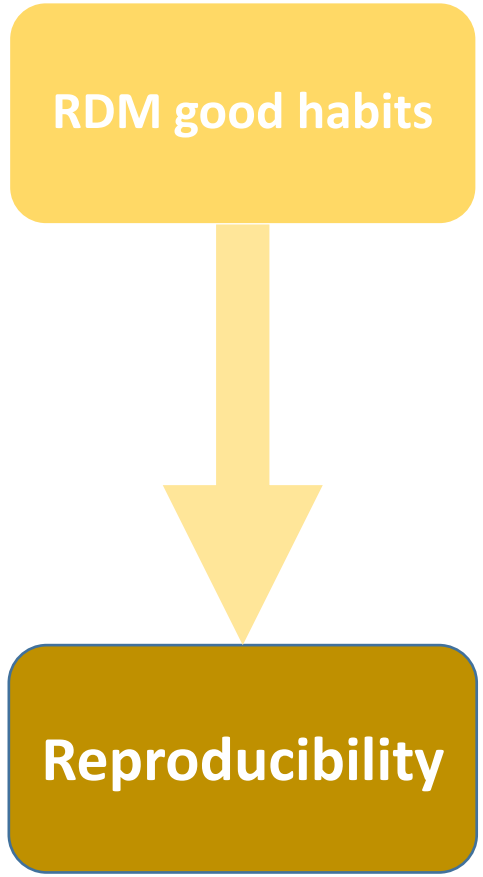
What does responsible research data management mean?

Good habits in RDM...

- Guarantee the **safety** and **security** of your data
- Make it compliant to **regulations** and to **ethical** codes
- Enable easy data **re-use** and therefore, optimises the scientific effort
- Enable, in summary, scientific **reproducibility**

Reproducibility is the possibility for a published study to be verified, reused, continued, or widened, by fellow researchers. It is applicable to **data**, **results** and **methods**.

It is what brings **life** to your research



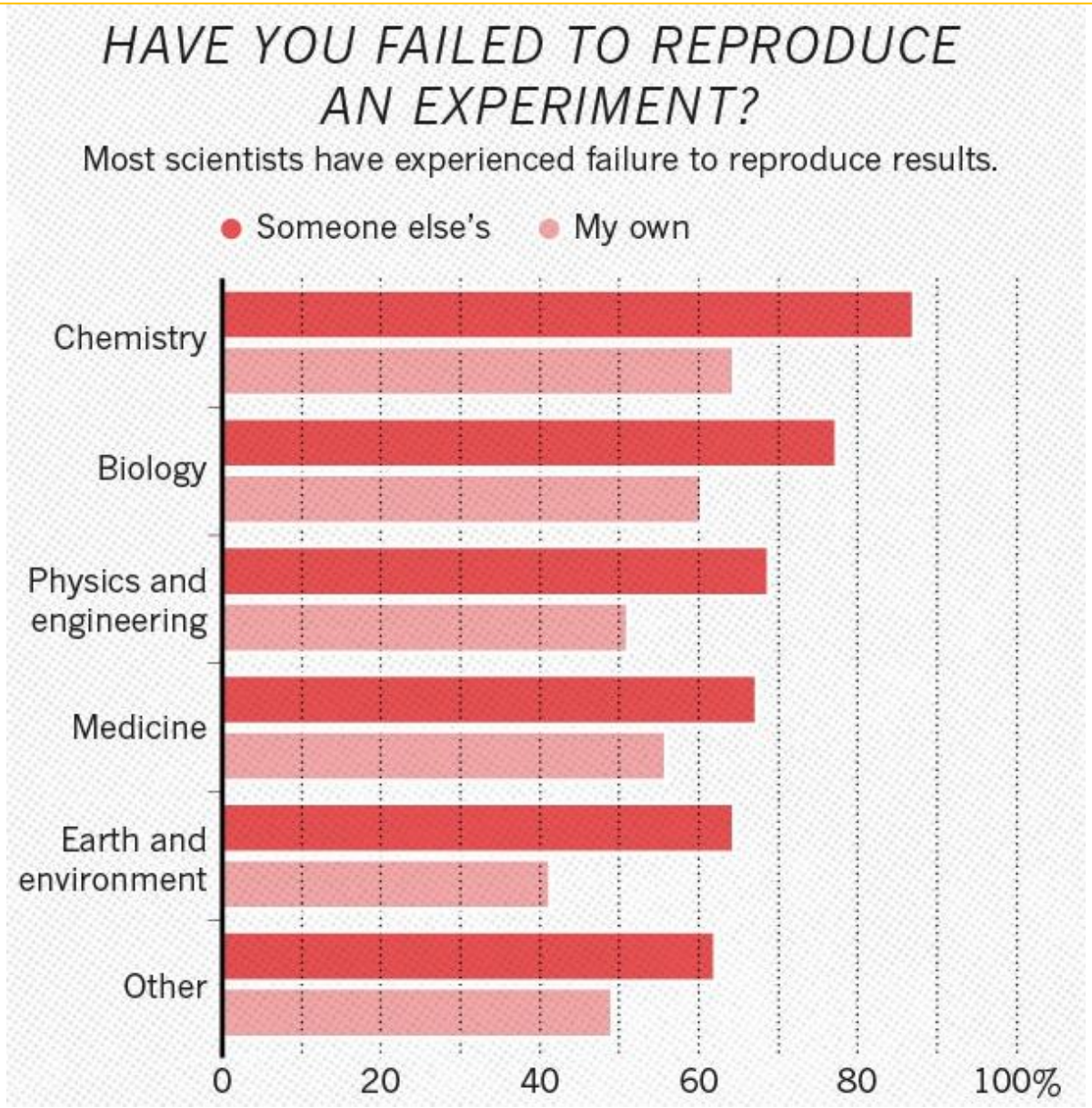
Why is it hard to make research reproducible ?

Scientific Reproducibility

https://www.youtube.com/watch?v=FpCrY7x5nEE

Scientific Reproducibility

Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a



Scientific Reproducibility

Reproducibility crisis

- Most scientific results are difficult, even **impossible**, to reproduce and/or replicate [*]

- This issue stems from a **paradoxical context** that does not favour scientific reproducibility but pushes towards a lack of **quality** (precision, transparency, integrity) in scientific methodology

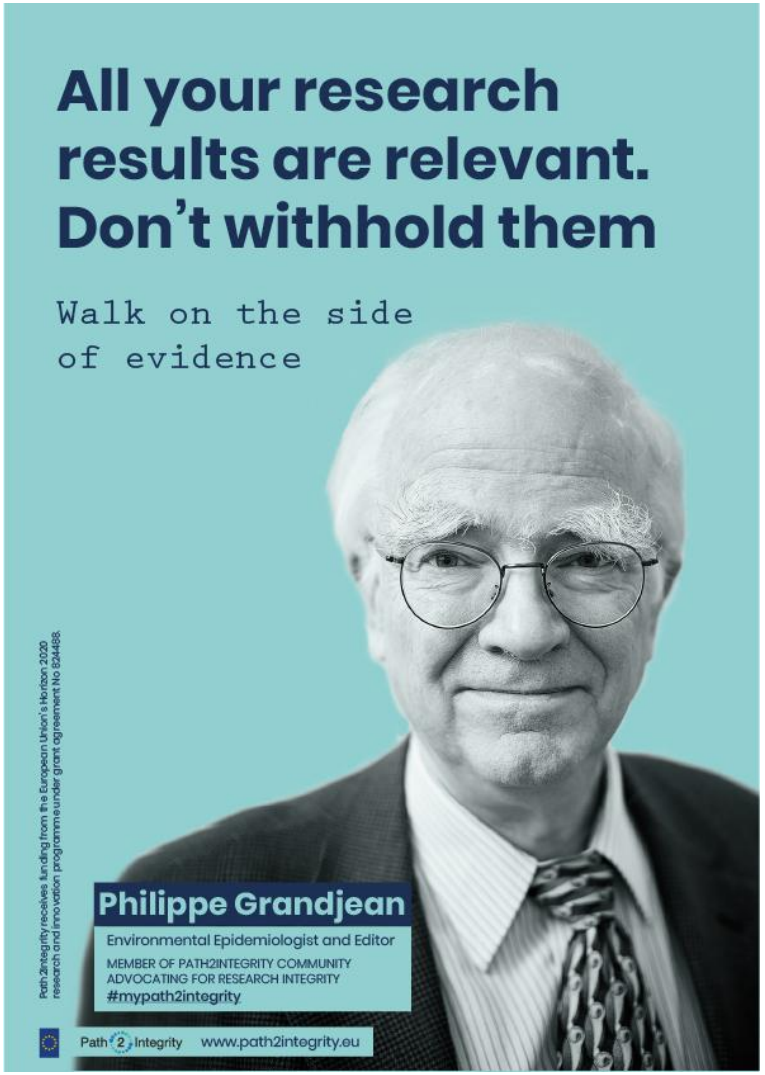
Scientific Reproducibility

Reproducibility crisis

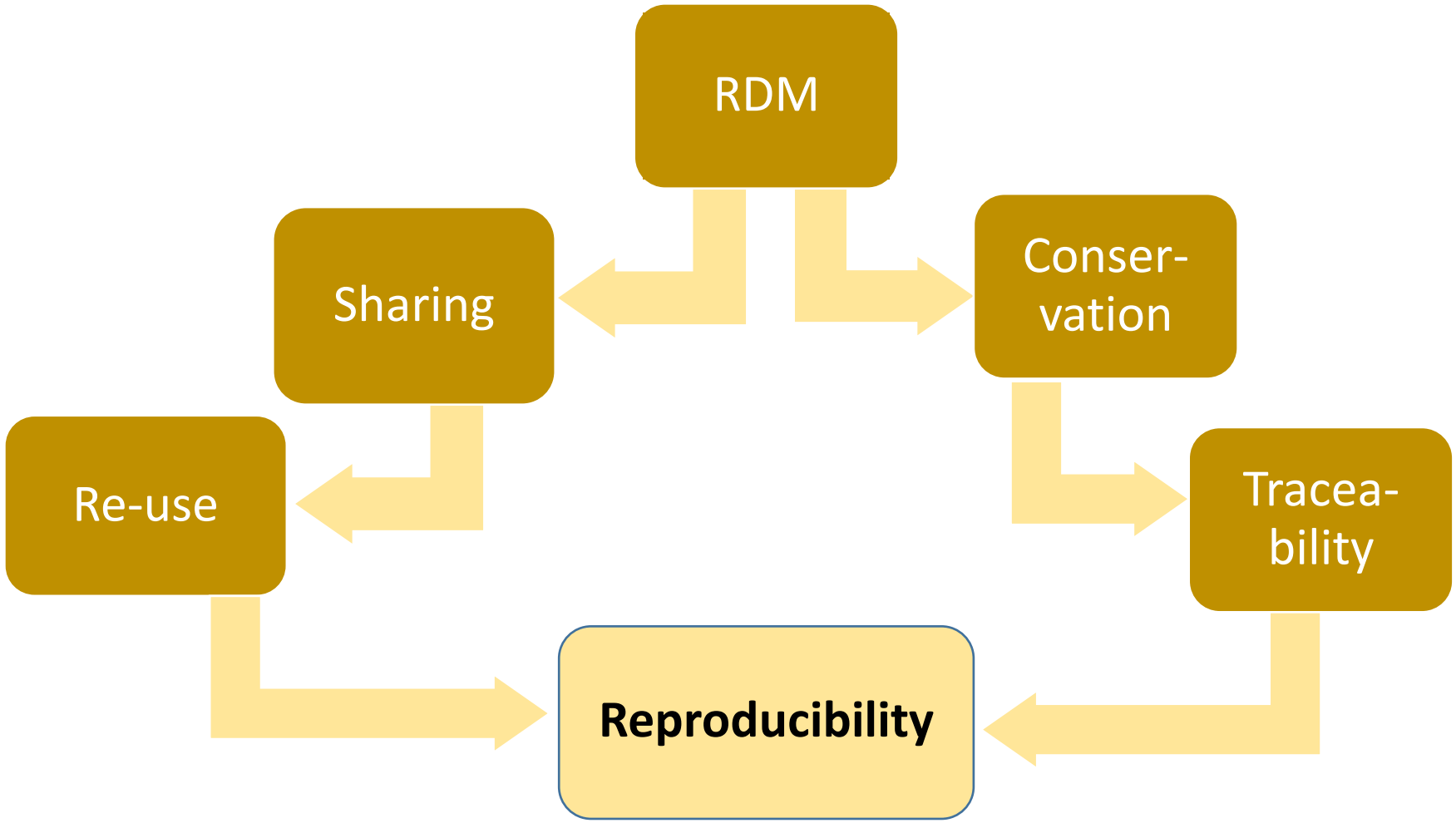
- Most scientific results are difficult, even **impossible**, to reproduce and/or replicate [*]
- This issue stems from a **paradoxical context** that does not favour scientific reproducibility but pushes towards a lack of **quality** (precision, transparency, integrity) in scientific methodology
- This is **not a decrease** in researchers skills but **a cultural phenomenon**, because of the paradoxical system that rules research culture (publish or perish)
- More and more **stakeholders** are nudging a **cultural change** towards more transparency and accountability

You can be the change

Scientific Reproducibility



Scientific Reproducibility

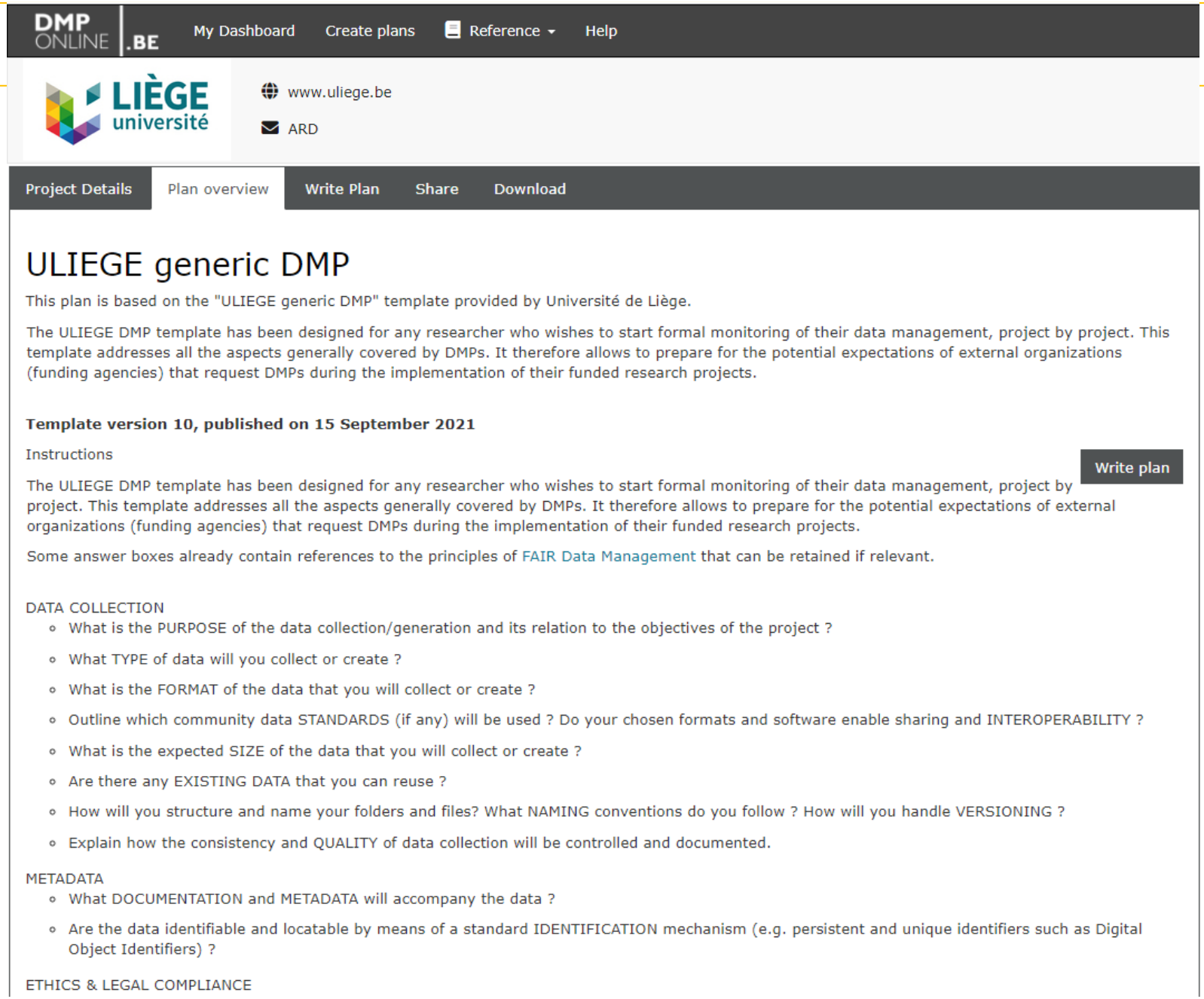


Data planning

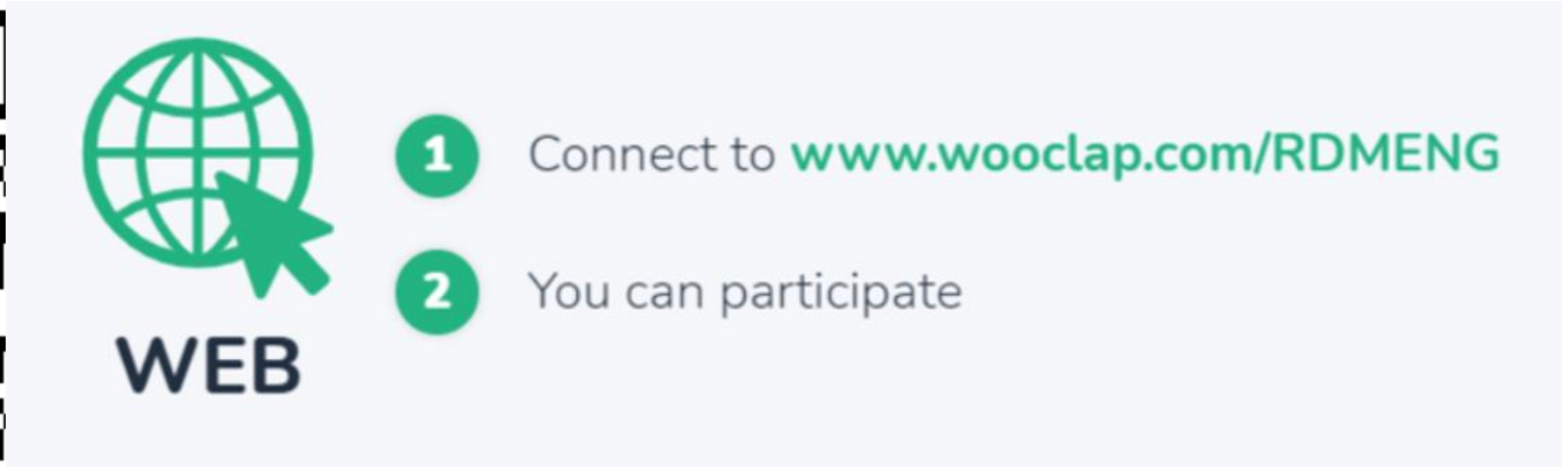
A **Data Management Plan** is a set of questions, usually web-based, that works as a checklist of attention points to guide the researcher through the data lifecycle.

- Some funders require to fill in template DMPs as deliverables, and those are therefore reviewed
- Some DMP templates are made available online, as examples, for researchers to use without any obligation or without any review
- These online tools usually provide guidance and examples of best practice
- A DMP is a living thing, it can evolve along the research project
- Most institutions (should) encourage the use of DMPs
- Recommended tool : the Data Curation Center (UK) provides DMP templates in text form, guidance and examples, as well as the DMPonline portal – other institutions may use other tools

Data planning



Data planning



Have you filled in a DMP for your PhD thesis?

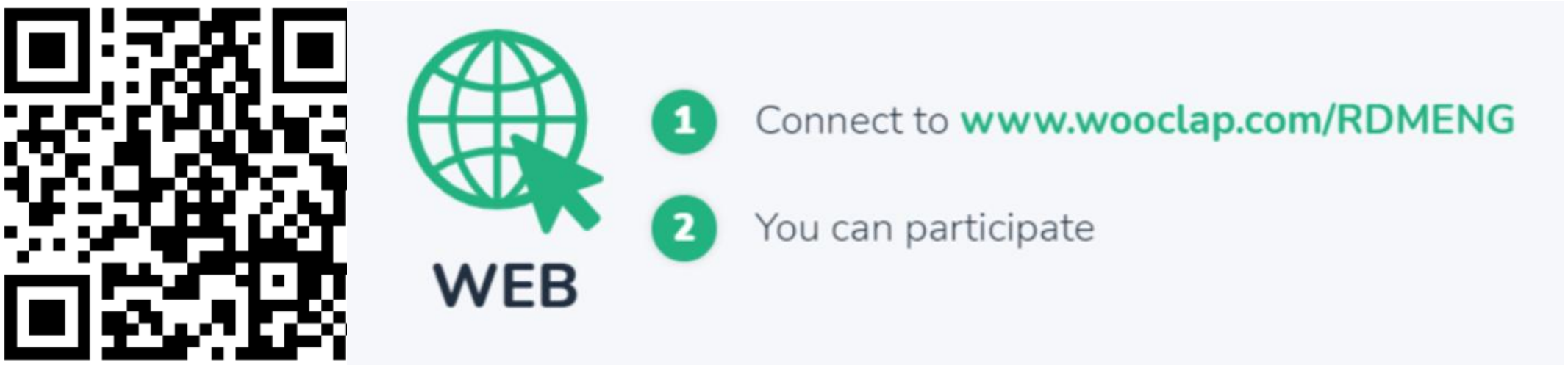
Data planning

Questions break

Data planning: rules and regulations

Many questions of data management, specifically access, storage, protection and sharing, have roots in applicable rules and regulations
-> awareness is a good start !

Data planning: rules and regulations



Which regulations do you think might apply to these examples?

Content : Catherine Thiry

Data planning: rules and regulations

Case	Applicable regulations
A history researcher collects representations of the pope through the ages	Copyright, intellectual property regulations
A PhD student in applied sciences takes videos of volunteers during fatigue tests while driving a simulated car	GDPR
A social sciences research records interviews with study subjects and writes them down using pseudonyms	GDPR
A biology lab requests lion blood samples from a Tanzanian nature reserve	Nagoya protocol
In an economics research project, someone collects confidential industrial information from different companies	Confidentiality agreement / non disclosure agreement

Content : Laurence Thys

Data planning: rules and regulations

General Data Protection Regulations (GDPR - 2016) is a set of regulations protecting the **privacy** of humans. It addresses the collection, processing, storage, transfer of personal data inside and outside EU.

When applied to research, it results in a few obligations:

- **Informing** subjects of the use of their data and leaving them the right to withdraw (with some limitations)
- **Protecting** their data from transferring, leaking, publishing...
- Building their study protocol and data collection on a **legal basis** (“why do you need personal data”)

! Be mindful of anonymization

Content : Catherine Thiry

Data planning: rules and regulations

Data relating to these sectors are subject to specific codes of conduct:

- Agricultural data
- Energy consumption data;
- Clinical trial regulation/ pharmaceutical testing data;
- Geographical data;
- Road safety-related minimum universal traffic data;
- Electronic communication data;
- Chemical safety data…

-> Talk to your local legal affairs dept

Data planning: rules and regulations

Questions break

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

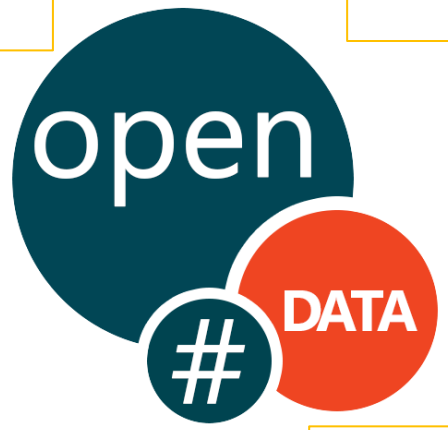
The aim is therefore to practice as much **open data** as possible.

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

Increases **visibility**, impact and improves reputation

Facilitates scientific **reproducibility**



Ensures data **sustainability,** even after staff rotation for example

Optimises return on investment from funders

Facilitates **collaborations**

Selection criterion for some funding programs (or awards)

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

The aim is therefore to practice as much **open data** as possible.

However, open data is **not always possible** or not always the best way to go, or even not the only recommendation that should be observed **(why?)**

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

The aim is therefore to practice as much **open data** as possible.

However, open data is **not always possible** or not always the best way to go, or even not the only recommendation that should be observed **(why?)**

Data that cannot be shared

For legal reasons (GDPR, NDA...)
For strategic reasons (patents -> embargo)

Note : good RDM habits are also for oneself ☺

Open data

Not always a token of quality

Not always re-usable straight away (it is not just about posting online)

Should be the direction if not the destination

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

The aim is therefore to practice as much **open data** as possible.

However, open data is **not always possible** or not always the best way to go, or even not the only recommendation that should be observed **(why?)**

The data FAIRness spectrum

Findable

Accessible

Interoperable

Reusable

FAIR data

The data FAIRness spectrum

Findable

Data are discoverable and easy to find, by both humans and computers.

- **Metadata**
- Digital Object Identifier
- Other standard identifier

Accessible

Interoperable

Reusable

FAIR data

The data FAIRn

Paper metadata

The Location of Young Pulsar PSR J0837−2454: Galactic Halo or Local Supernova Remnant?

Show affiliations

Pol, Nihan; Burke-Spolaor, Sarah; Hurley-Walker, Natasha; Blumer, Harsha; Johnston, Simon; Keith, Michael; Keane, Evan F.; Burgay, Marta; Possenti, Andrea; Petroff, Emily; Bhat, N. D. Ramesh

We present the discovery and timing of the young (age ~ 28.6 kyr) pulsar PSR J0837−2454. Based on its high latitude ($b = 9.8^{o}$) and dispersion measure (DM $= 143$~pc~cm^{-3}), the pulsar appears to be at a z-height of >1 kpc above the Galactic plane, but near the edge of our Galaxy. This is many times the observed scale height of the canonical pulsar population, which suggests this pulsar may have been born far out of the plane. If accurate, the young age and high z-height imply that this is the first pulsar known to be born from a runaway O/B star. In follow-up imaging with the Australia Telescope Compact Array (ATCA), we detect the pulsar with a flux density $S_{1400} = 0.18 \pm 0.05$ mJy. We do not detect an obvious supernova remnant around the pulsar in our ATCA data, but we detect a co-located, low-surface-brightness region of $\sim 1.5^{o}$ extent in archival Galactic and Extragalactic All-sky MWA Survey data. We also detect co-located Hα emission from the Southern Hα Sky Survey Atlas. Distance estimates based on these two detections come out to ~0.9 kpc and ~0.2 kpc respectively, both of which are much smaller than the distance predicted by the NE2001 model (6.3 kpc) and YMW model (> 25 kpc) and place the pulsar much closer to the plane of the Galaxy. If the pulsar/remnant association holds, this result also highlights the inherent difficulty in the classification of transients as "Galactic" (pulsar) or "extragalactic" (fast radio burst) toward the Galactic anti-center based solely on the modeled Galactic electron contribution to a detection.

Publication: eprint arXiv:2104.11680

Pub Date: April 2021

arXiv: arXiv:2104.11680

Bibcode: 2021arXiv210411680P

Keywords: Astrophysics - High Energy Astrophysical Phenomena

E-Print Comments: Published in ApJ. 12 pages, 9 figures, 2 tables; doi:10.3847/1538-4357/abe70d

The data FAIRnes

Dataset metadata

Files | Metadata | Terms | Versions

Export Metadata

Citation Metadata

Dataset Persistent ID	doi:10.14428/DVN/FT2TX8
Publication Date	2021-02-05
Title	Replication Data for: Expectancy-value-cost motivational theory to explore final year medical students' research intentions and past research experience: a multicentre cross-sectional questionnaire study
Author	Van Maele, Louis (IRSS, CAMG, Université catholique de Louvain, Belgium) - ORCID: 0000-0003-1683-1207 Devos, Christelle (IPSY, Université catholique de Louvain, Belgium) Guisset, Séverine (SMCS, LIDAM, Université catholique de Louvain, Belgium) Leconte, Sophie (IRSS, CAMG, Université catholique de Louvain, Belgium) Macq, Jean (IRSS, Université catholique de Louvain)
Contact	Use email button above to contact. Van Maele, Louis (IRSS, CAMG, Université catholique de Louvain)
Description	The purpose of this dataset was to study final-year medical students' research intentions and motivation based on the Expectancy-Value-Cost motivational theory. The data comes from an online questionnaire sent in February 2017 to final-year medical students in three French-speaking Belgian universities (ULB, UCLouvain and ULg).
Subject	Medicine, Health and Life Sciences; Other
Keyword	Motivation (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=motivation Medical Students (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=medical+student Research (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=activities%2C+research Questionnaire (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=design%2C+questionnaire Belgium (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=belgium
Production Date	2017-03-30
Production Place	Belgium
Depositor	Van Maele, Louis
Deposit Date	2020-12-02

Live example : publishing on Zenodo

The data FAIRness spectrum

Findable

Data are discoverable and easy to find, by both humans and computers.

- **Metadata**
- Digital Object Identifier
- Other standard identifier

In most cases, at least the metadata can be shared

Accessible

Interoperable

Reusable

FAIR data

The data FAIRness spectrum

Findable

Data are discoverable and easy to find, by both humans and computers.

- **Metadata**
- Digital Object Identifier
- Other standard identifier

In most cases, at least the metadata can be shared

Accessible

Data are made available in a **sustainable** way, even after the project is over:

- The (meta)data are retrievable with a flexible protocol in an **open directory** (harvesting)
- If the data cannot be shared, it has to be justified

Using a data repository usually checks most boxes

Interoperable

Reusable

FAIR data

The data FAIRness spectrum

Findable

Data are discoverable and easy to find, by both humans and computers.

- **Metadata**
- Digital Object Identifier
- Other standard identifier

In most cases, at least the metadata can be shared

Accessible

Data are made available in a **sustainable** way, even after the project is over:

- The (meta)data are retrievable with a flexible protocol in an **open directory** (harvesting)
- If the data cannot be shared, it has to be justified

Using a data repository usually checks most boxes

Interoperable

Data are able to be operated / exchanged / compared between a variety of institutions, workflows, software, applications, systems, ...

- The (meta)data use a broadly compatible format (not proprietary if possible)
- The documentation is in English

Reusable

FAIR data

The data FAIRness spectrum

Findable

Data are discoverable and easy to find, by both humans and computers.

- **Metadata**
- Digital Object Identifier
- Other standard identifier

In most cases, at least the metadata can be shared

Accessible

Data are made available in a **sustainable** way, even after the project is over:

- The (meta)data are retrievable with a flexible protocol in an **open directory** (harvesting)
- If the data cannot be shared, it has to be justified

Using a data repository usually checks most boxes

Interoperable

Data are able to be operated / exchanged / compared between a variety of institutions, workflows, software, applications, systems, ...

- The (meta)data use a broadly compatible format (not proprietary if possible)
- The documentation is in English

Reusable

The data are **sufficiently described** and can be shared with as few restrictions as possible, as the ultimate goal is to optimise data reuse.

- The **licenses** are as open as possible.
- The format is as universal as possible
- The data is well documented

FAIR data

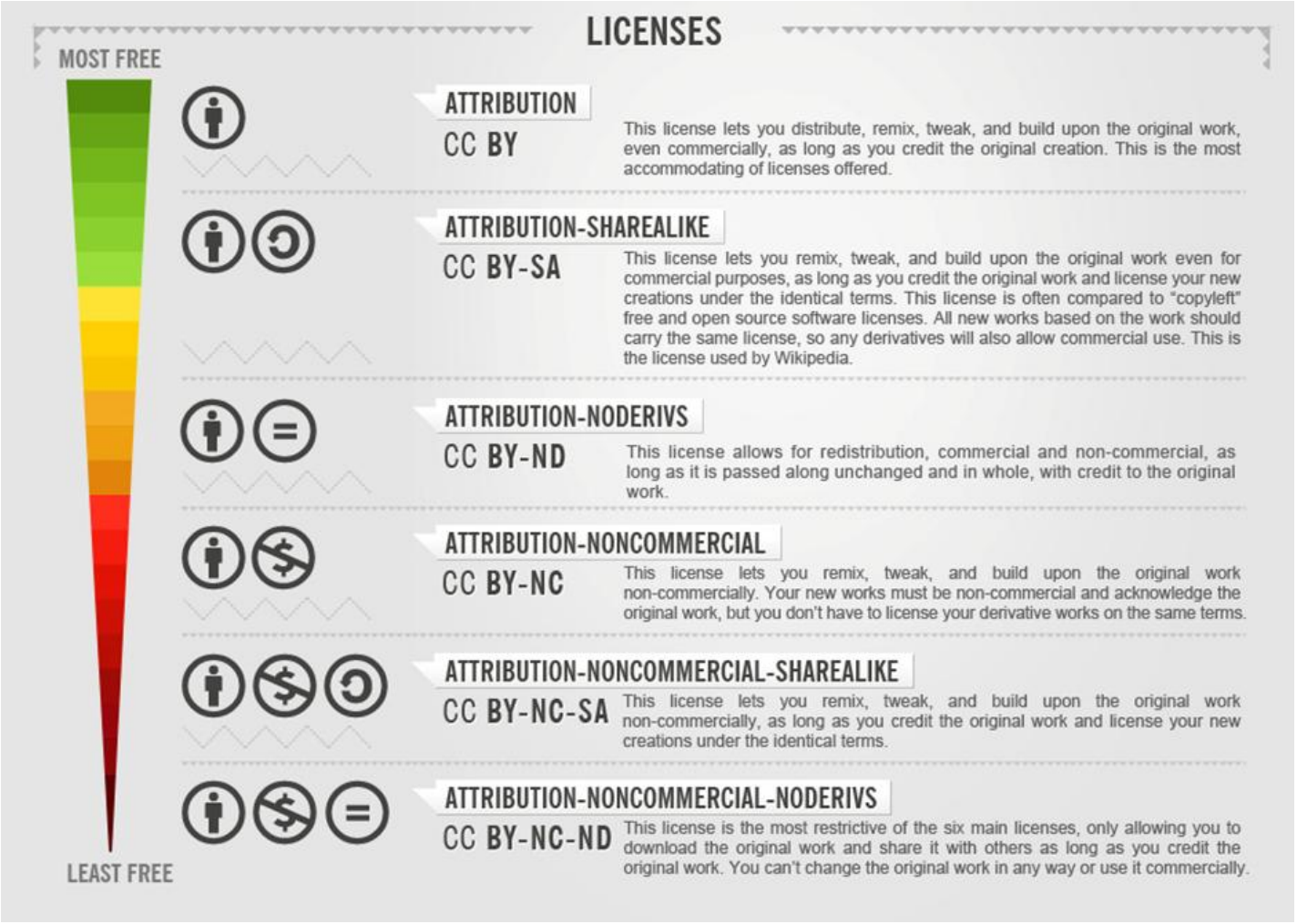
The data FAIRness spectrum

A license defines how to **reuse** the content:

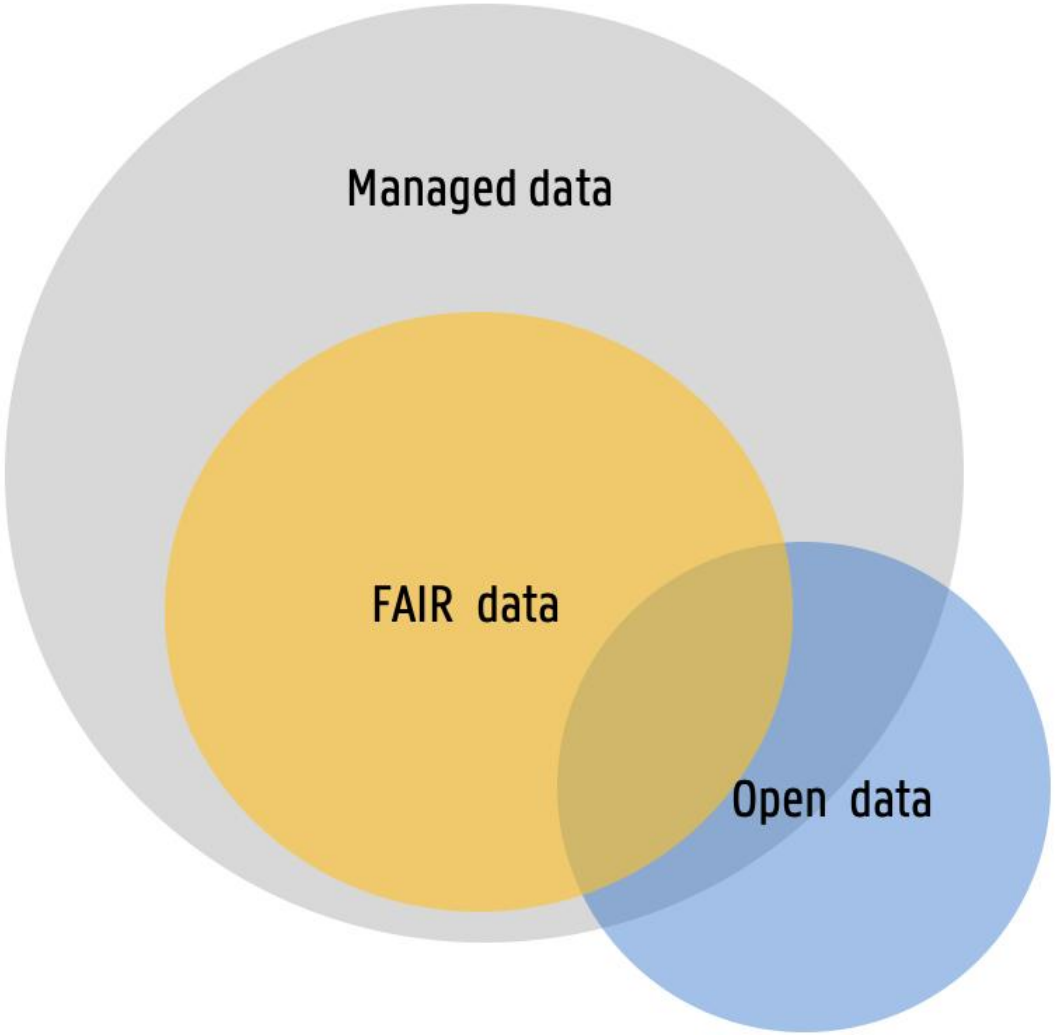
Rights to **reuse**, to **modification**, to **commercial** use, **obligation** to mention the **attribution** and to **share alike**

You can define your conditions, but most often:

- The license may come with the use of a **repository** (see example on Zenodo)
- The license may come with the publication through an **editor** (journal) -> be mindful of editors ☺



The data FAIRness spectrum

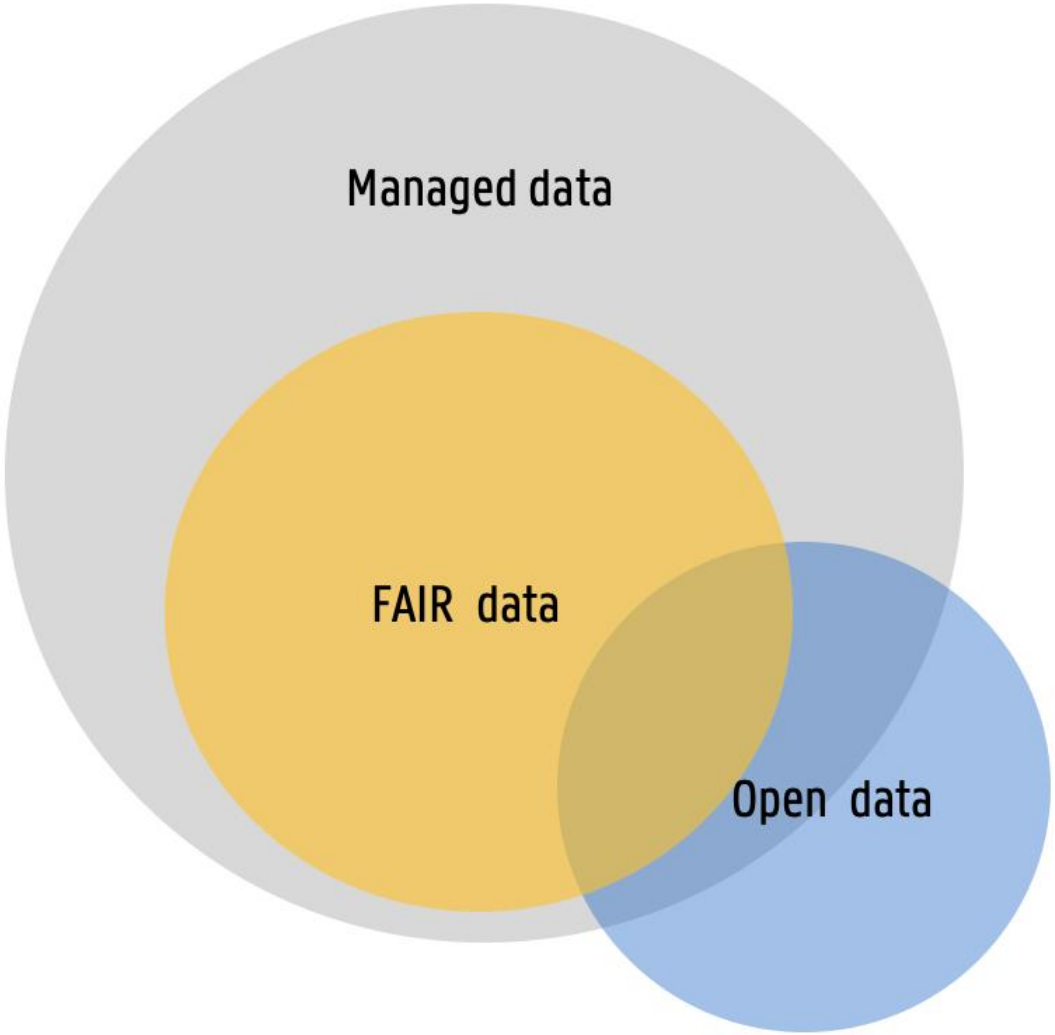


FAIR data is a bridge between individual good RDM habits and open data.

Making data FAIR is not only about the sharing step of the project, it starts at data creation:

- Storage
- Documentation
- Protection
- Traceability
- ...

The data FAIRness spectrum



FAIR data is a bridge between individual good RDM habits and open data.

Making data FAIR is not only about the sharing step of the project, it starts at data creation:

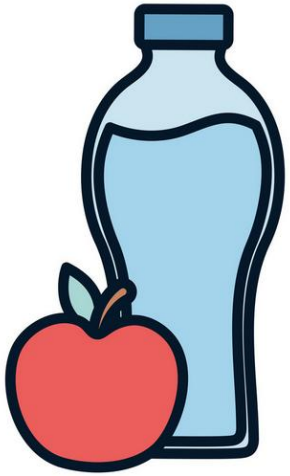
- Storage
- Documentation
- Protection
- Traceability
- ...

What habits can you take up early on to facilitate FAIR data sharing at the end of your PhD?

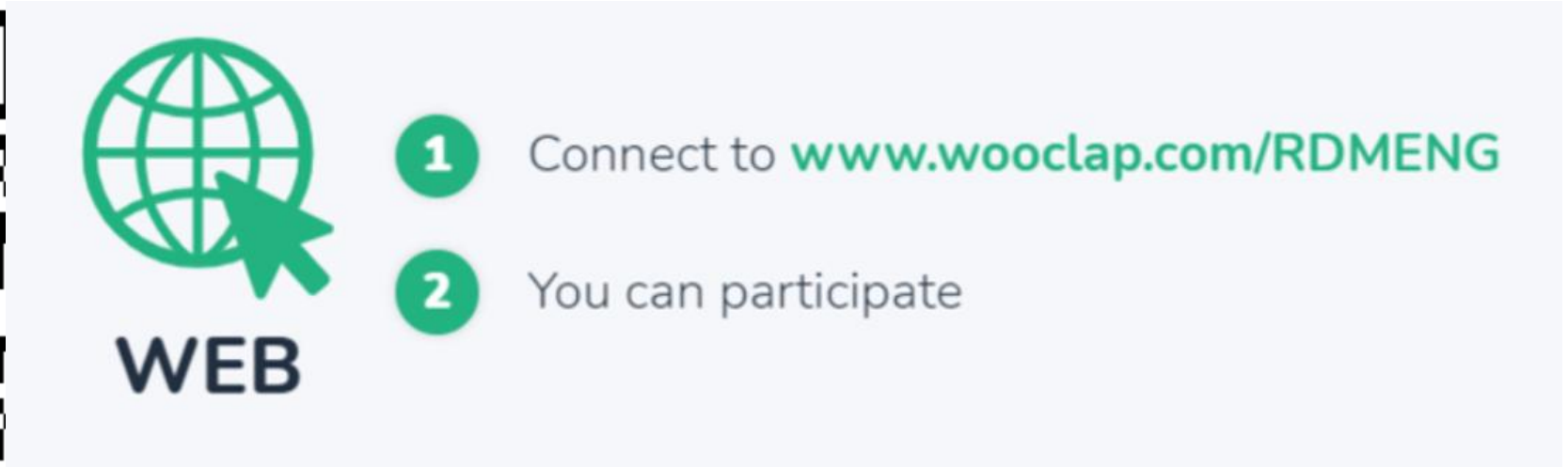
The data FAIRness spectrum

Questions break

15 minutes



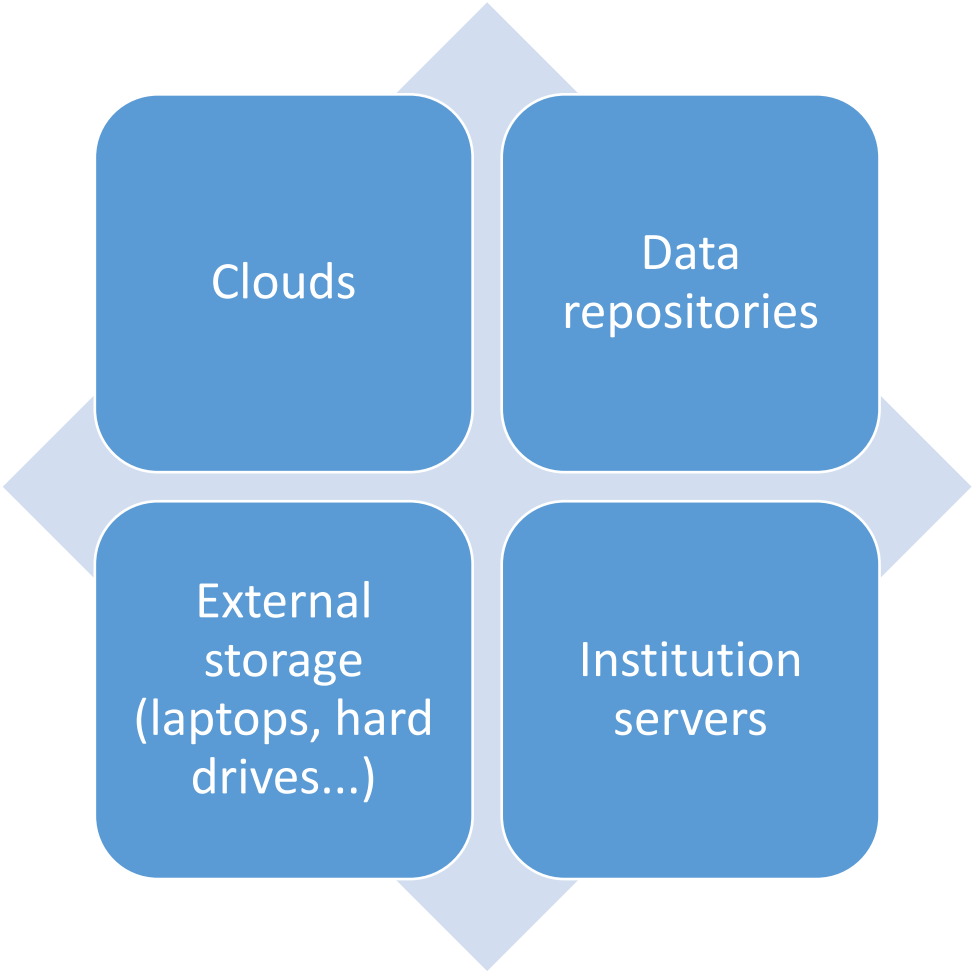
Data storage



Poll about your current data storage habits

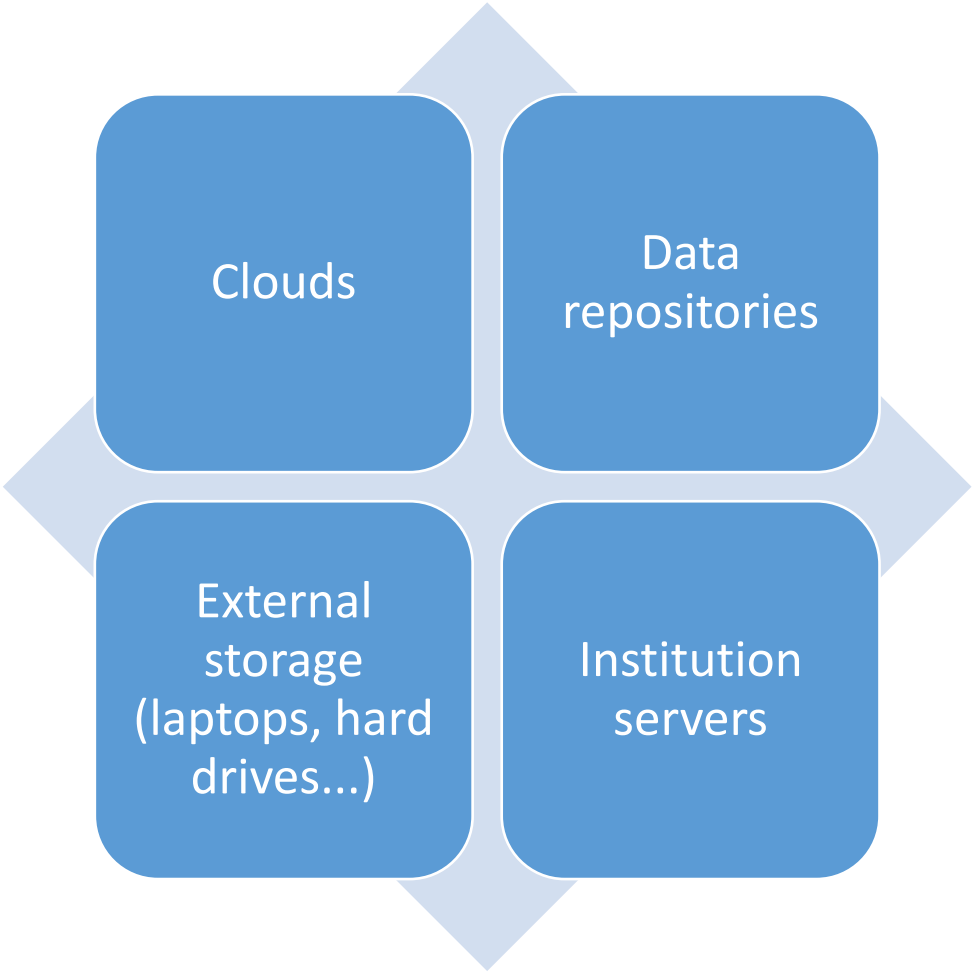
Data storage

There are four main families of storage solutions:



Data storage

How do I choose?



Data storage

How do I choose?

Documentation

Organisation

Security

Sustainability

Data storage

How do I choose?

Documentation

- Keep track as much as possible between raw data and results, even **inconclusive**
- **Never erase anything**
- The difference b/w data dredging and reproducibility is telling what you did)

Organisation

- Knowing that, how much volume do I need?
- **Tree structure?**
- Explicit filenames

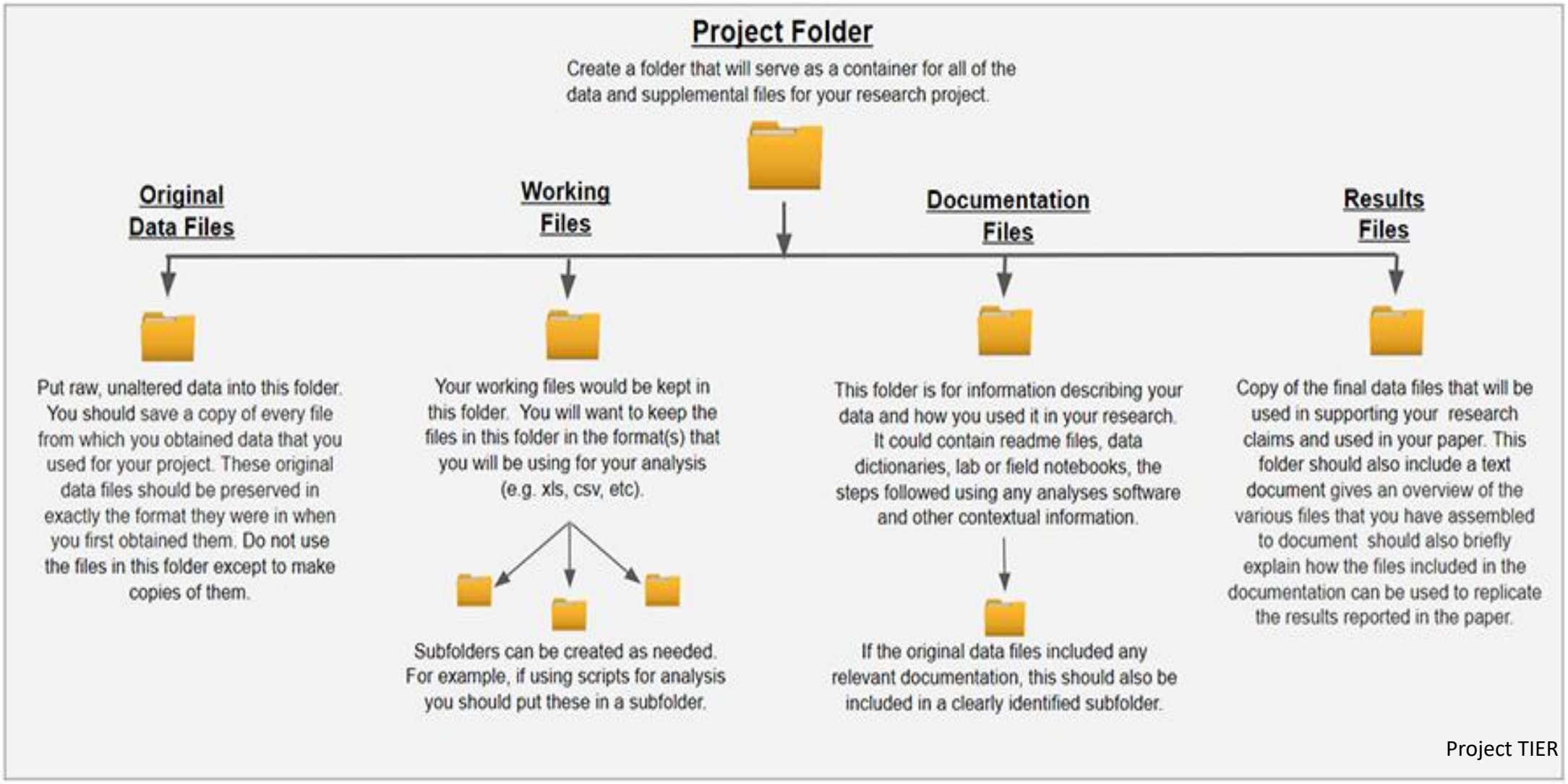
Security

- Which folders need to be protected from « leaking »? Which level of confidentiality?
- **Backup ?**
  3 copies
  2 different storage solutions
  1 off-site (if possible)

Sustainability

- Could someone reuse this easily in ten years?
  - Availability
  - Documents
  - Description
  - Format

Data storage

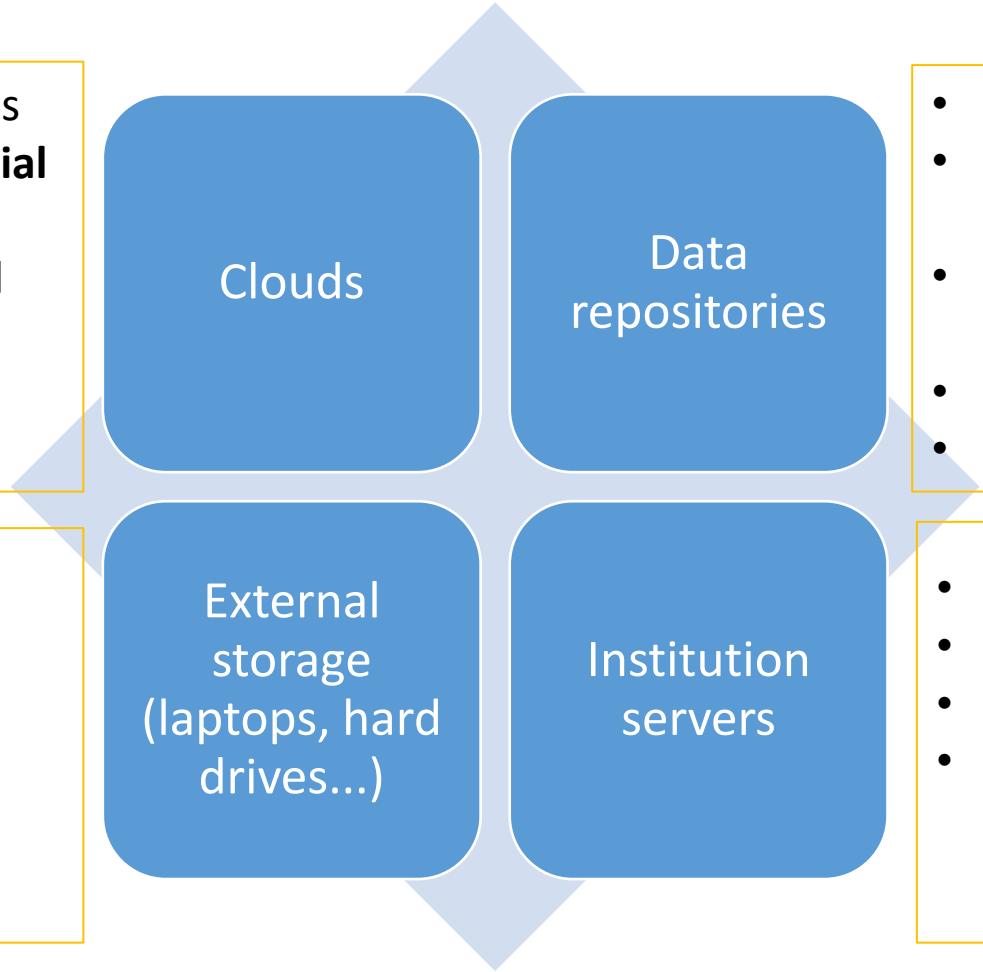


Project TIER

Data storage

How do I choose? -> according to my organisation & security choices or obligations:

- Google Drive, DropBox, ... Implies **outsourcing -> not for confidential data**
- If your institution has an internal cloud, much better
- Usually no metadata structure
- Be mindful of volume limits



- Zenodo, Figshare, Dataverse,...
- Made for open **sharing**, usually not for working data
- May imply outsourcing too -> keep a **local** copy
- Structured metadata
- Be mindful of volume limits

- Practical for working data but...
- Physical storage -> mobility **risks**
- **Password** protection is always a good idea

- **Internal** management = safer
- Usually automatically backed up
- Usually easily accessed remotely
- Most often for **working data**, since usually complicated to access from extra muros

Data storage

It may be a good idea to mix and match... e.g. my working data on a hard drive + a cloud, my sharing data on Zenodo -> no single recipe for success, but best practice

- Google Drive, DropBox, ... Implies **outsourcing -> not for confidential data**
- If your institution has an internal cloud, much better
- Usually no metadata structure
- Be mindful of volume limits

Clouds

Data repositories

- Zenodo, Figshare, Dataverse,...
- Made for open **sharing**, usually not for working data
- May imply outsourcing too -> keep a **local** copy
- Structured metadata
- Be mindful of volume limits

- Practical for working data but...
- Physical storage -> mobility **risks**
- **Password** protection is always a good idea

External storage (laptops, hard drives...)

Institution servers

- **Internal** management = safer
- Usually automatically backed up
- Usually easily accessed remotely
- Most often for **working data**, since usually complicated to access from extra muros

Data storage

A few extra tools:

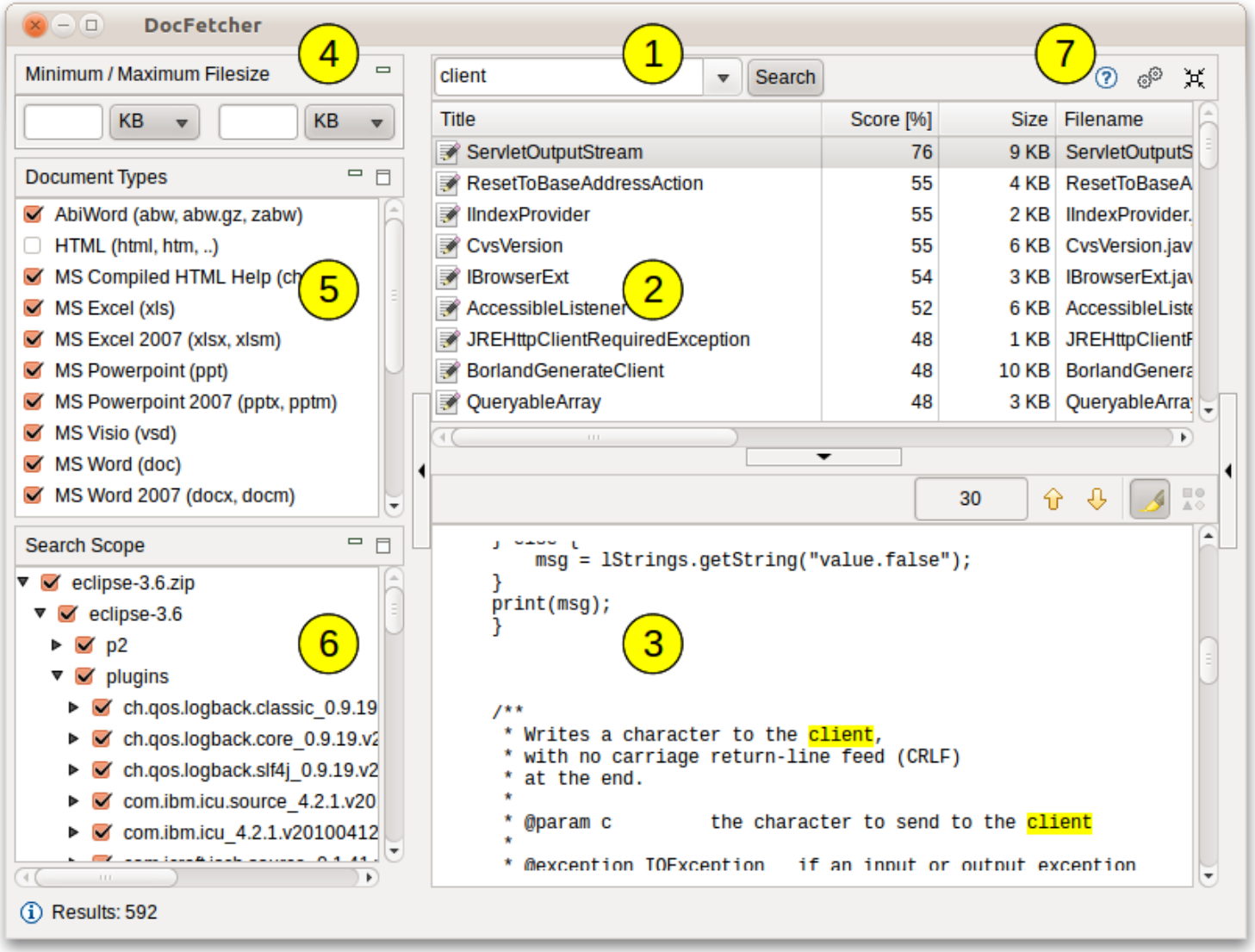
[Tips and examples] to organise, name and backup your data files

[Doc Fetcher]: indexing tool and tree structure visual

[Obsidian]: « work manager » to take notes, organise, version, index and mcreate link b/w files

[Jupyter]: python coding traceability tool

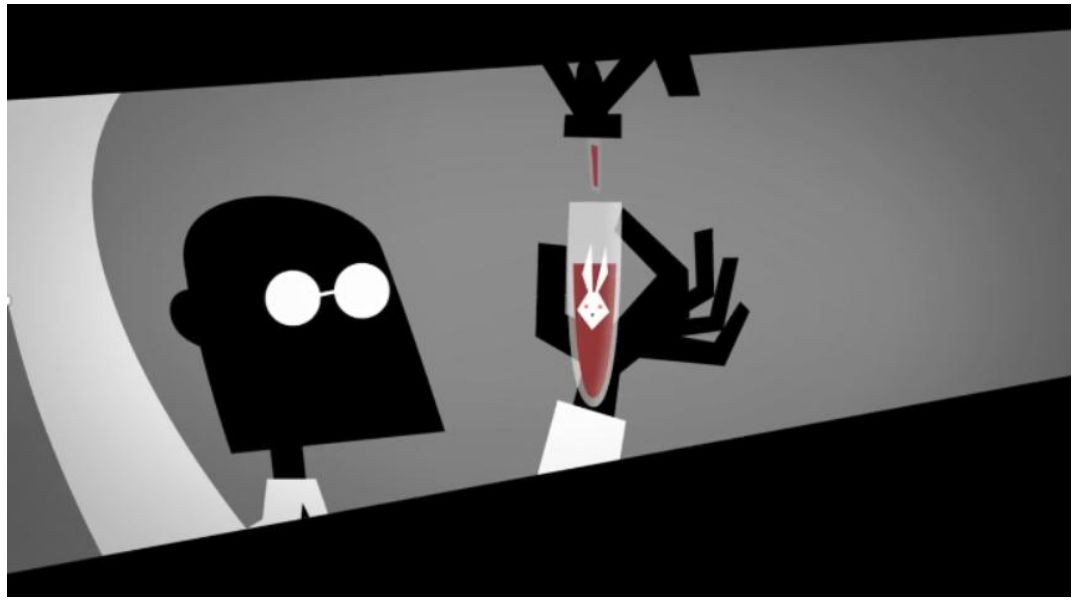
[Gitlab]: automated traceability tool for team work



Data storage

Questions break

Data management and ethics

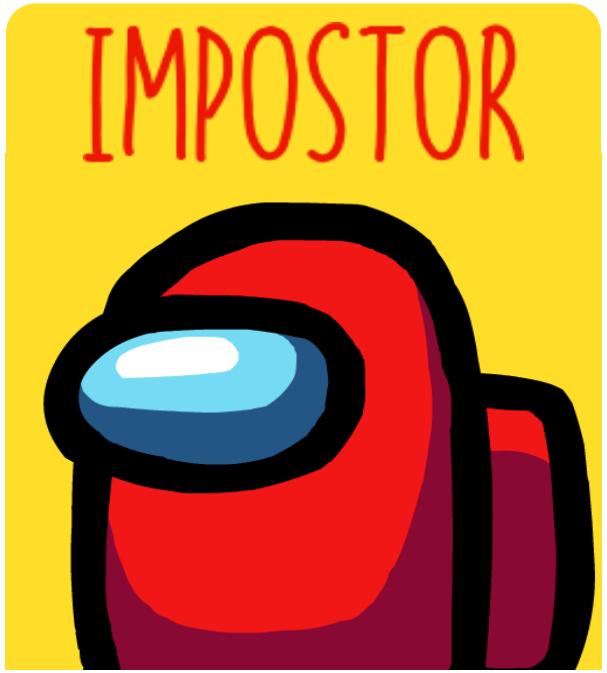


Research lives in a paradoxical context that may push us, even unconsciously, towards questionable practice

Again, there is a spectrum, from plain fraud to best practice, there are **grey areas** in which we must make the best choice possible to ensure reproducibility

Irreproducible science can be suspicious

Data management and ethics



Numerous famous cases:

- 2020 Retraction of a paper that held claims on hydroxychloroquine based on fabricated data. This had consequence on COVID-19 gov policies: LancetGate

https://retractionwatch.com/

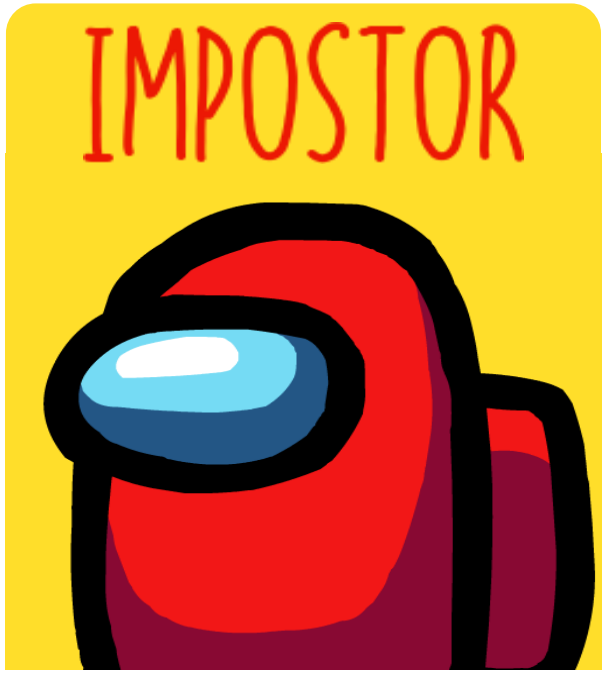
Research lives in a paradoxical context that may push us, even unconsciously, towards questionable practice

Again, there is a spectrum, from plain fraud to best practice, there are **grey areas** in which we must make the best choice possible to ensure reproducibility

Irreproducible science can be suspicious

Fraud = falsification, fabrication, plagiarism -> no tolerance

Data management and ethics



Numerous famous cases:

- 2020 Retraction of a paper that held claims on hydroxychloroquine based on fabricated data. This had consequence on COVID-19 gov policies: LancetGate

https://retractionwatch.com/

Research lives in a paradoxical context that may push us, even unconsciously, towards questionable practice

Again, there is a spectrum, from plain fraud to best practice, there are **grey areas** in which we must make the best choice possible to ensure reproducibility

Irreproducible science can be suspicious

Fraud = falsification, fabrication, plagiarism -> no tolerance

Then there are **shortcuts** (not fraud but not good practice)

Data management and ethics

What is data dredging, why does it happen, and what are its consequences?

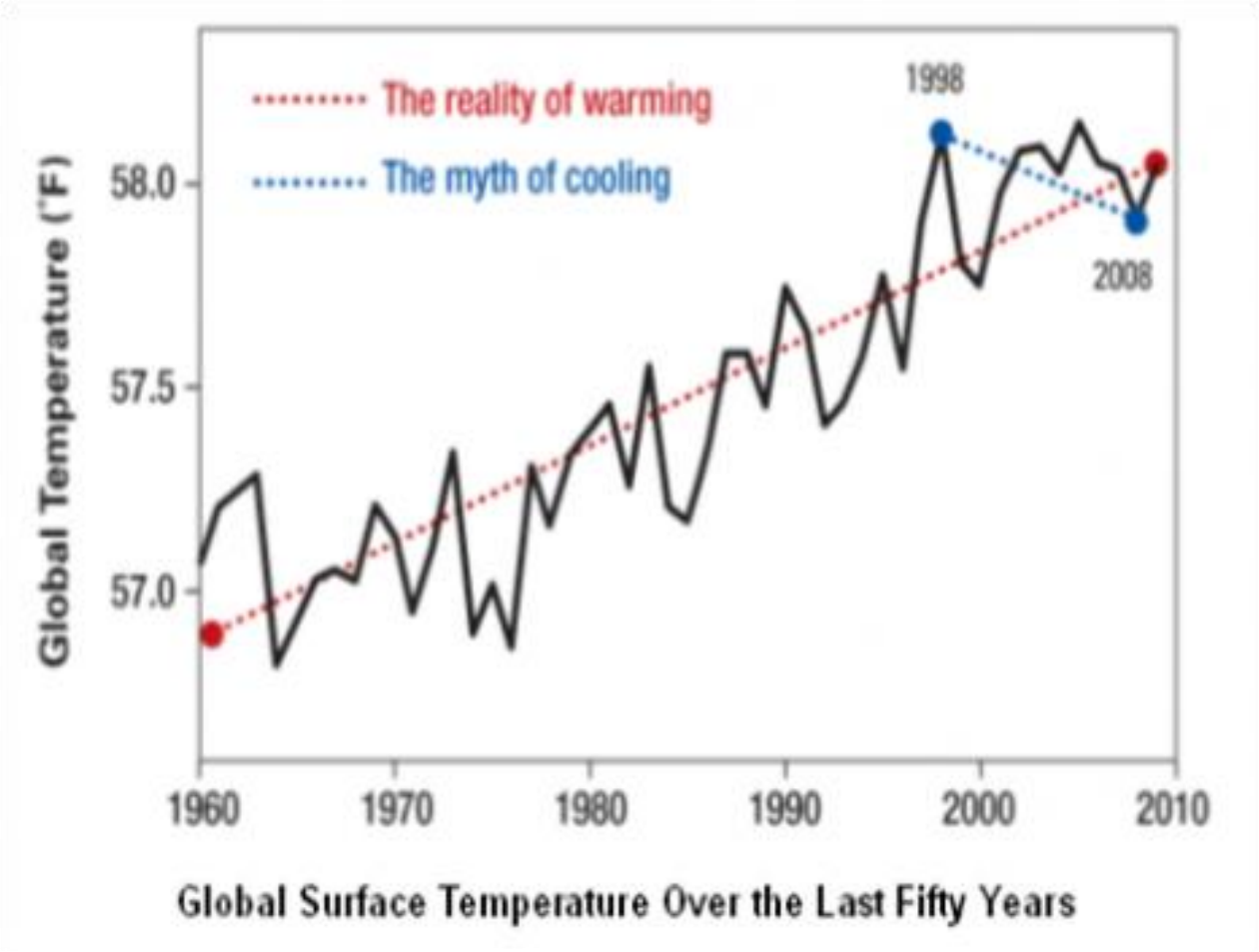
Data management and ethics

- **Pressure to publish** with tenure and funding on the line
- Pressure to find results that seem **new and striking**
- Numerous ways to **tweak** your study, **consciously or not**, until you get a result that counts as **statistically significant**, even though it is probably meaningless:
  - → Altering how long it lasts
  - → Play with the sample size
  - → P-hacking (collecting lots of variables and playing with data until finding counts as statistically significant)
- As a result: many studies that get media coverage seem to contradict each other, impeding the **trust** of society in (good) science

Rule of thumb : it is okay to play around with your data, but the difference b/w data dredging and exploring a dataset is telling about it in your publications

What is data dredging, why does it happen, and what are its consequences?

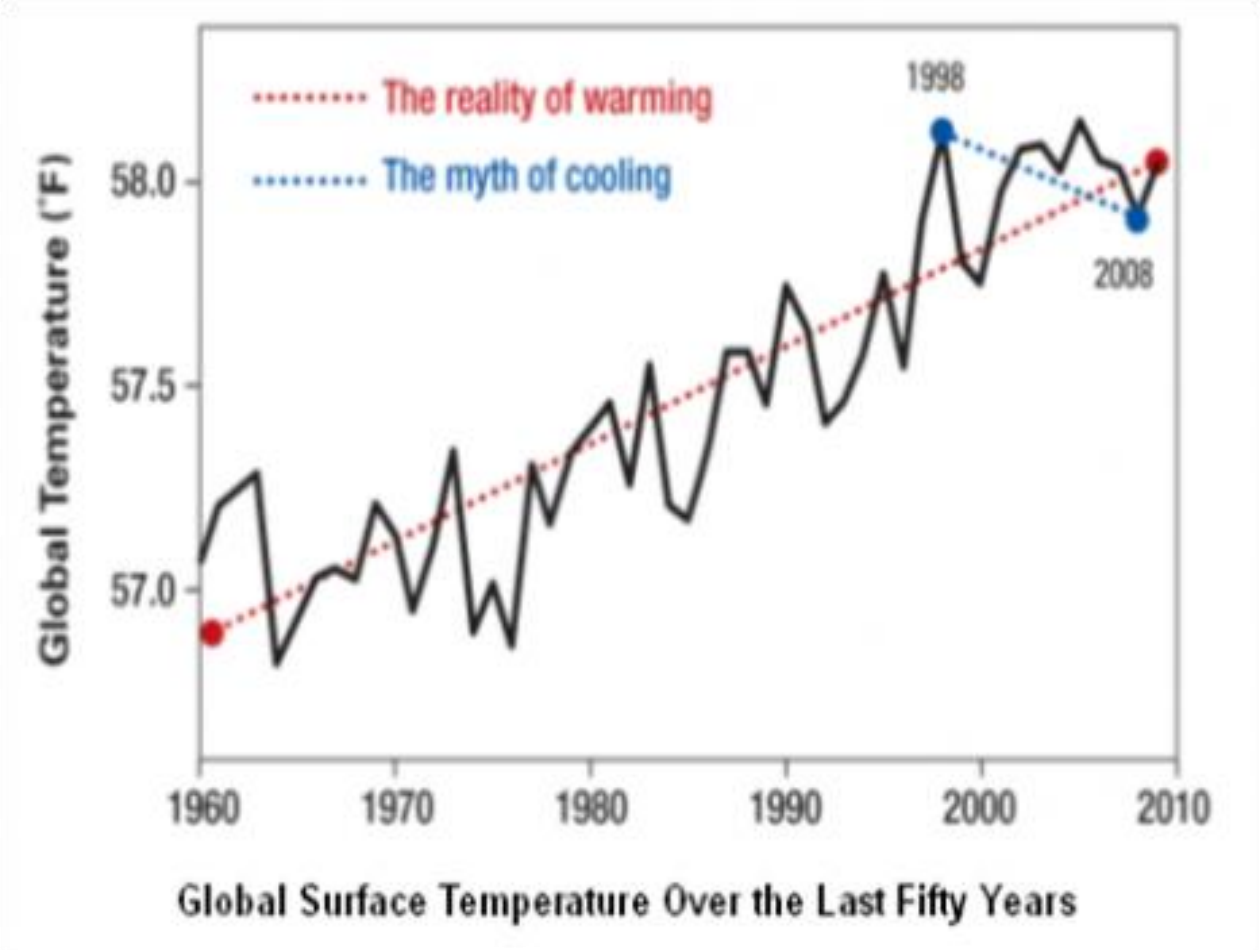
Data dredging



Cherry picking

Intentionally filtering out data that does not support the pet hypothesis

Data dredging



Cherry picking

Intentionally filtering out data that does not support the pet hypothesis

Filtering out is okay
Erasing and **neglecting to mention** is not

Data dredging

Outcome switching

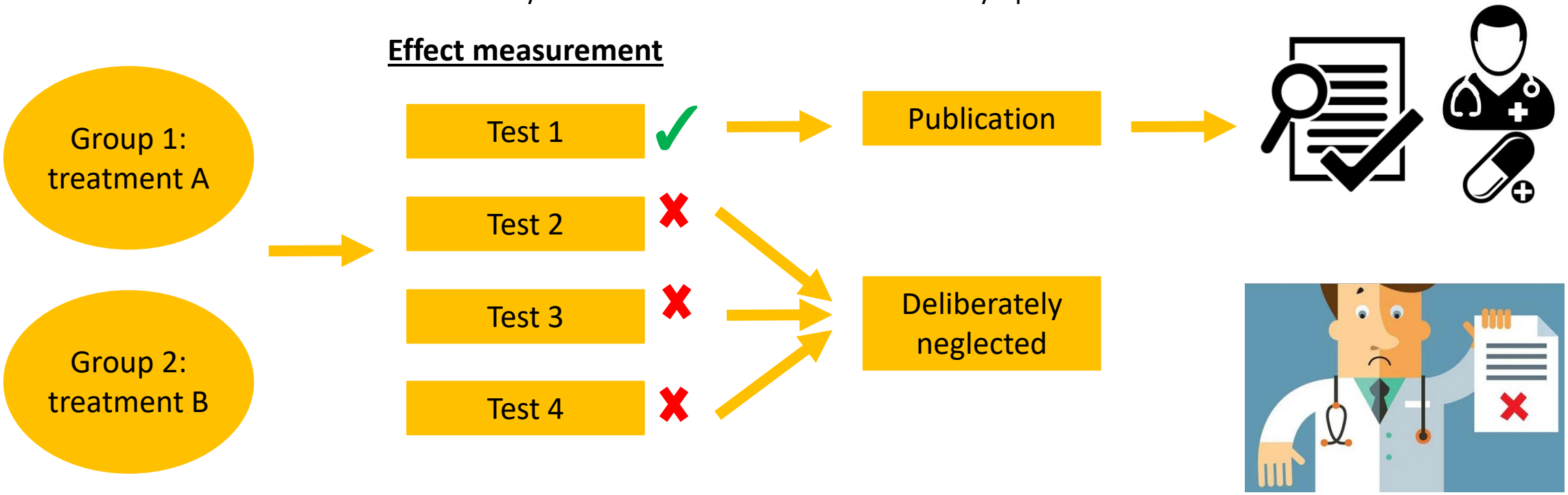
Changing the course of a study during its execution to eliminate inconclusive or negative results
Usually encountered in clinical trials

Data dredging

Outcome switching

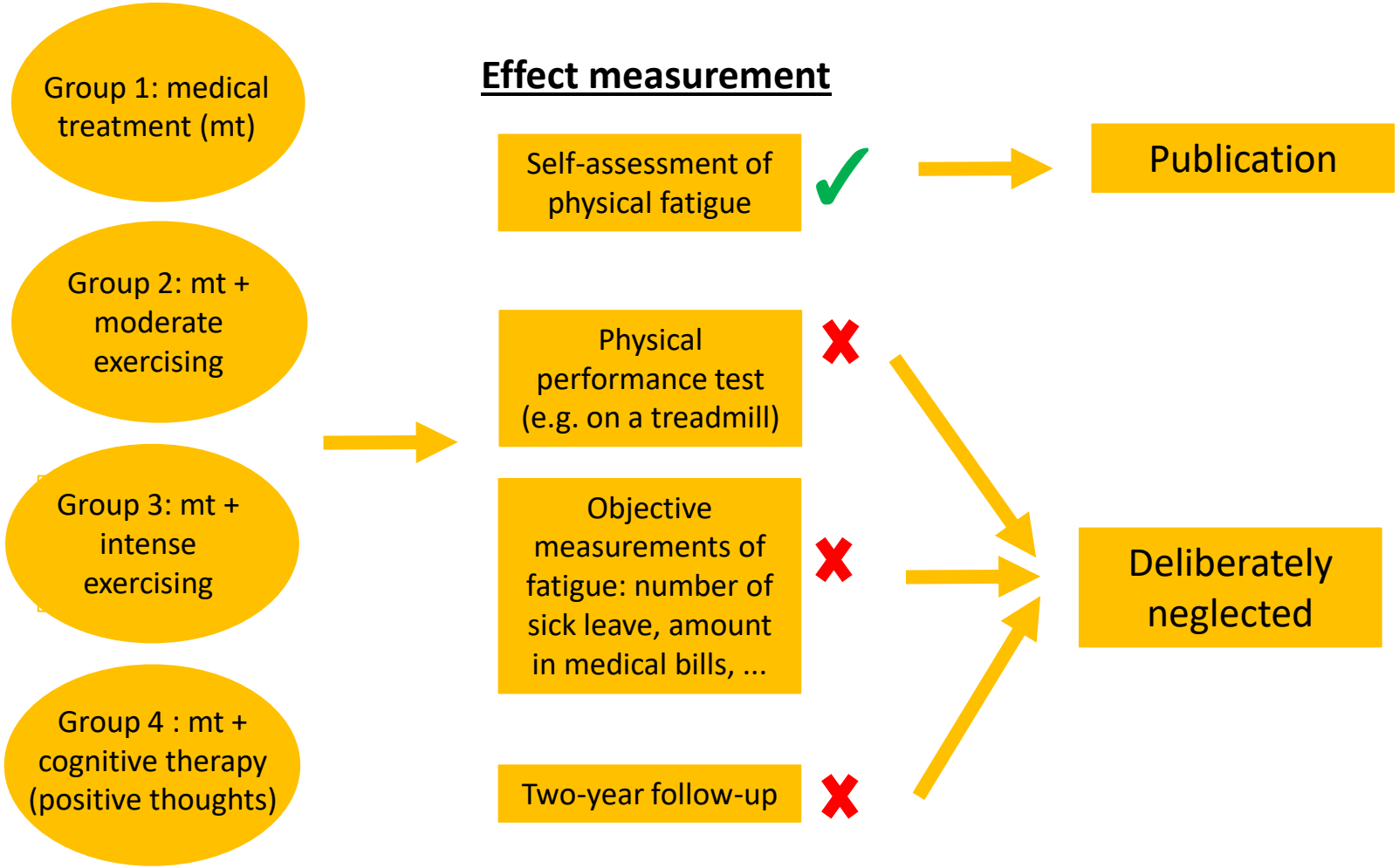
Changing the course of a study during its execution to eliminate inconclusive or negative results
Usually encountered in clinical trials

Do treatment A and treatment B have any effect on the considered disease or symptom?



Data dredging

Example: the « PACE trial »



B/w 2005 and 2010, the UK Medical Research Council conducted studies on the possible connection between exercising and decreasing the symptoms of myalgic encephalomyelitis

641 patients split randomly into 4 groups

Data dredging

Psychological Medicine (2013), **43**, 2227–2235. © Cambridge University Press 2013
doi:10.1017/S0033291713000020

ORIGINAL ARTICLE

Recovery from chronic fatigue syndrome after treatments given in the PACE trial

P. D. White[1]*, K. Goldsmith[2], A. L. Johnson[3,4], T. Chalder[5] and M. Sharpe[6]; PACE Trial Management Group†

[1] *Wolfson Institute of Preventive Medicine, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, UK*
[2] *Biostatistics Department, Institute of Psychiatry, King's College London, UK*
[3] *MRC Biostatistics Unit, Institute of Public Health, University of Cambridge, UK*
[4] *MRC Clinical Trials Unit, London, UK*
[5] *Academic Department of Psychological Medicine, King's College London, UK*
[6] *Department of Psychiatry, University of Oxford, UK*

Background. A multi-centre, four-arm trial (the PACE trial) found that rehabilitative cognitive behaviour therapy (CBT) and graded exercise therapy (GET) were more effective treatments for chronic fatigue syndrome (CFS) than specialist
(APT) wh

Method.
meeting
and (*b*) cl
used logi

Results. The percentages (number/total) meeting trial criteria for recovery were 22% (32/143) after CBT, 22% (32/143) after GET, 8% (12/149) after APT and 7% (11/150) after SMC. Similar proportions met criteria for clinical recovery. The odds ratio (OR) for trial recovery after CBT was 3.36 [95% confidence interval (CI) 1.64–6.88] and for GET 3.38 (95% CI 1.65–6.93), when compared to APT, and after CBT 3.69 (95% CI 1.77–7.69) and GET 3.71 (95% CI 1.78–7.74), when compared to SMC (p values $\leqslant 0.001$ for all comparisons). There was no significant difference between APT and SMC. Similar proportions recovered in trial subgroups meeting different definitions of the illness.

Conclusions. This study confirms that recovery from CFS is possible, and that CBT and GET are the therapies most likely to lead to recovery.

Received 16 August 2012; Revised 14 December 2012; Accepted 17 December 2012; First published online 31 January 2013

Conclusions. This study confirms that recovery from CFS is possible, and that CBT and GET are the therapies most likely to lead to recovery.

Chronic Fatigue Syndrome sufferers 'can overcome symptoms of ME with positive thinking and exercise'

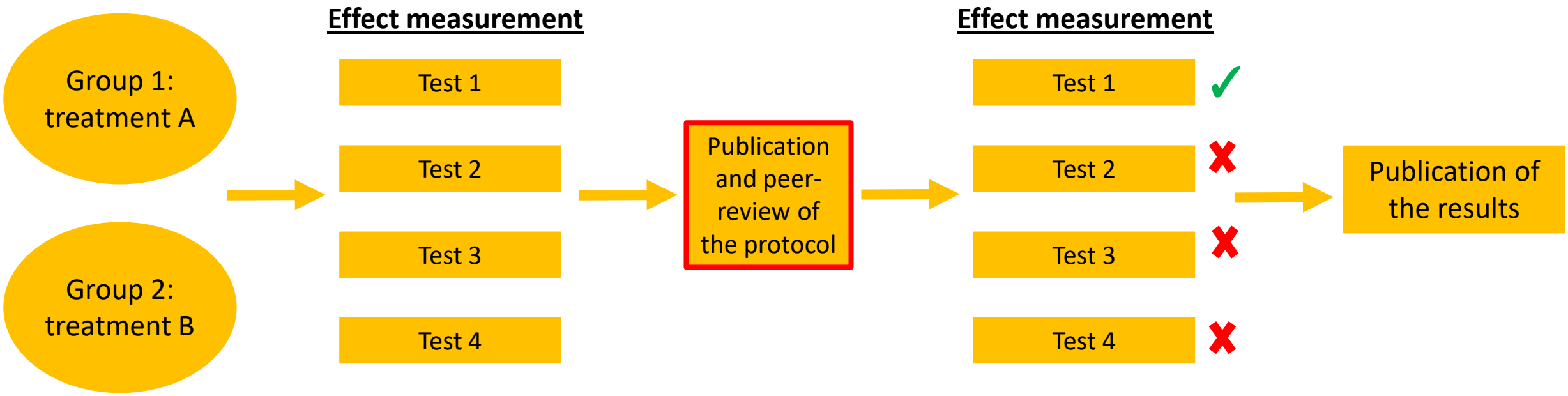
Oxford University has found ME is not actually a chronic illness

A study which followed hundreds of sufferers for two years found that those who were encouraged to be more active and alter their mind-set suffered less fatigue and were able to cope with daily life more easily Photo: Alamy

Data dredging

Outcome switching

Pre-registration of protocol?



Data dredging

P-hacking

Chopping up, testing, arranging, filtering, tweaking and/or tuning your dataset to obtain a **statistically significant result**

Even if it is random

Data dredging

I am testing **a hypothesis H**

ex: these diet pills do work

ex: this dice is loaded

I collect **relevant data**

ex: weight of a group of people before and after taking diet pills for a month

ex : number of times each face comes up after 50 dice rolls

Data dredging

I am testing **a hypothesis H**

ex: these diet pills do work

ex: this dice is loaded

I collect **relevant data**

ex: weight of a group of people before and after taking diet pills for a month

ex : number of times each face comes up after 50 dice rolls

I compute the **probability to obtain this same data even if my hypothesis H is wrong**

ex: if these pills do not work, what is the probability that these people would have lost weight anyway?

ex: if the dice is not loaded, what is the probability that face 6 only comes up 5 times out of 50?

Data dredging

I am testing **a hypothesis H** ex: these diet pills do work ex: this dice is loaded	I collect **relevant data** ex: weight of a group of people before and after taking diet pills for a month ex : number of times each face comes up after 50 dice rolls	I compute the **probability to obtain this same data even if my hypothesis H is wrong** ex: if these pills do not work, what is the probability that these people would have lost weight anyway? ex: if the dice is not loaded, what is the probability that face 6 only comes up 5 times out of 50?

The data drives the conclusion, not the opposite

Playing around with the p-value is fine, but **boiling down a complex scientific result to only one p-value** is not.

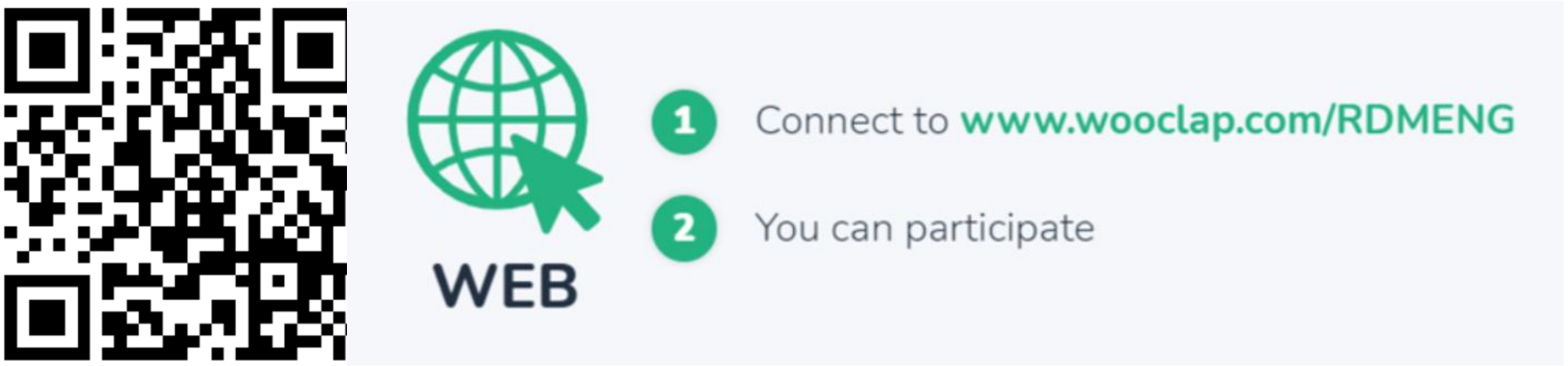
A small p-value is a **good indicator** that your hypothesis is correct, but is not enough:

→ **It does not prove H is true** (it only proves the opposite of H is improbable given this particular dataset)

→ It **does not prove** that the dataset is suitable for the test, or that the model is suitable for the hypothesis.

→ It **does not prove** the quality of the dataset (completeness, sample size, accuracy, ...)

Data dredging



P-hacking and nutrition

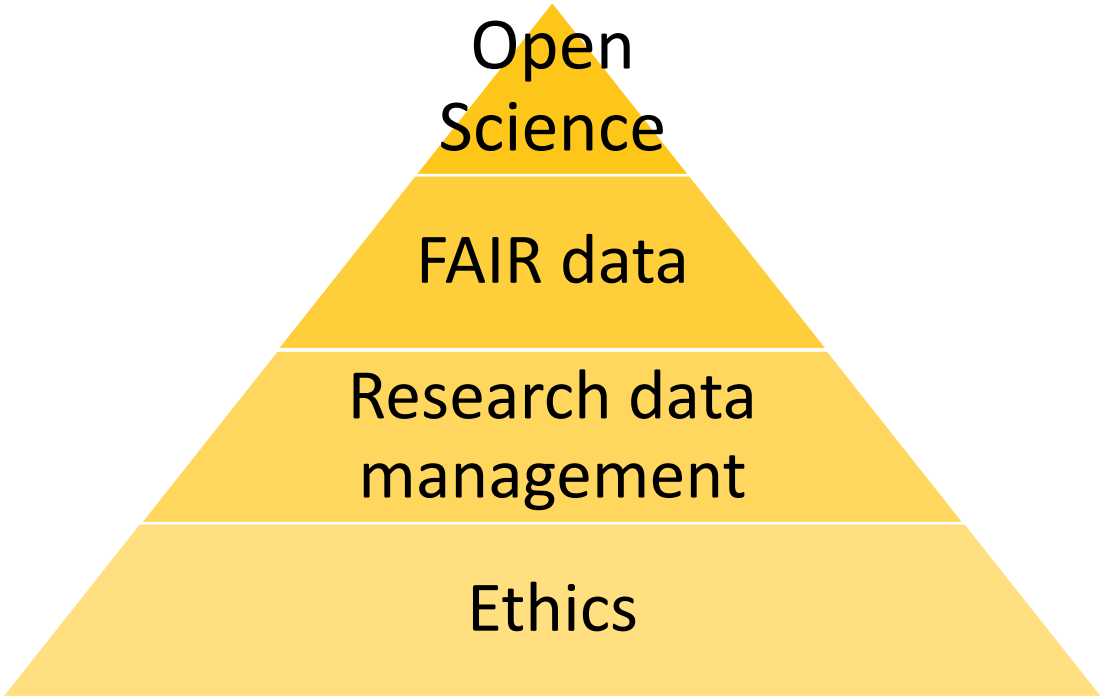
Data dredging

Questions break

The bigger picture

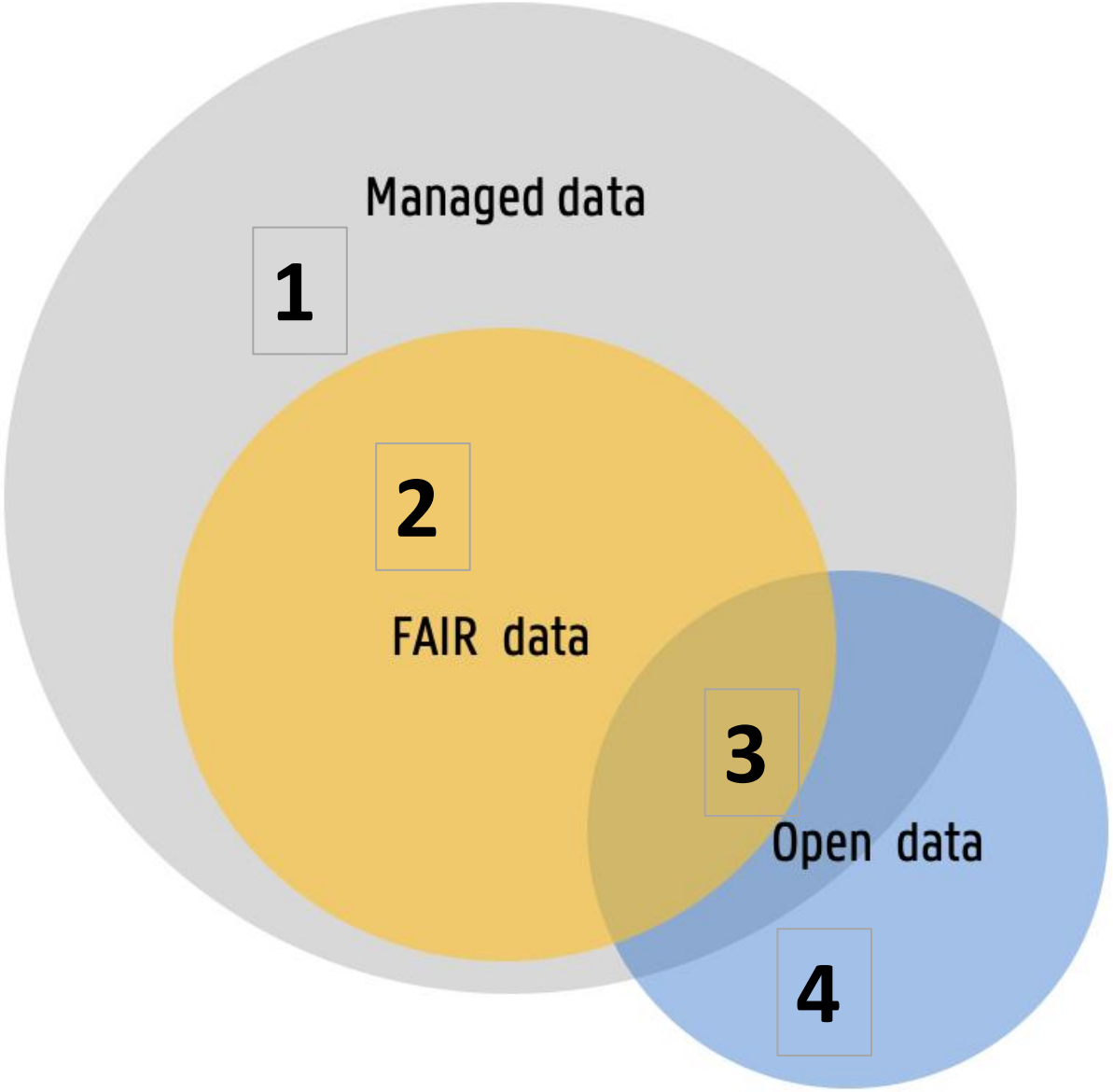
RDM is first and foremost a set of **best practices** that support scientific **reproducibility**

- Keeping track of every step of your data analysis and being transparent about it prevents any suspicion of **data dredging**
- Good habits in **data storage** enables this traceability and accountability
- It also ensures compliance **to rules and regulations**



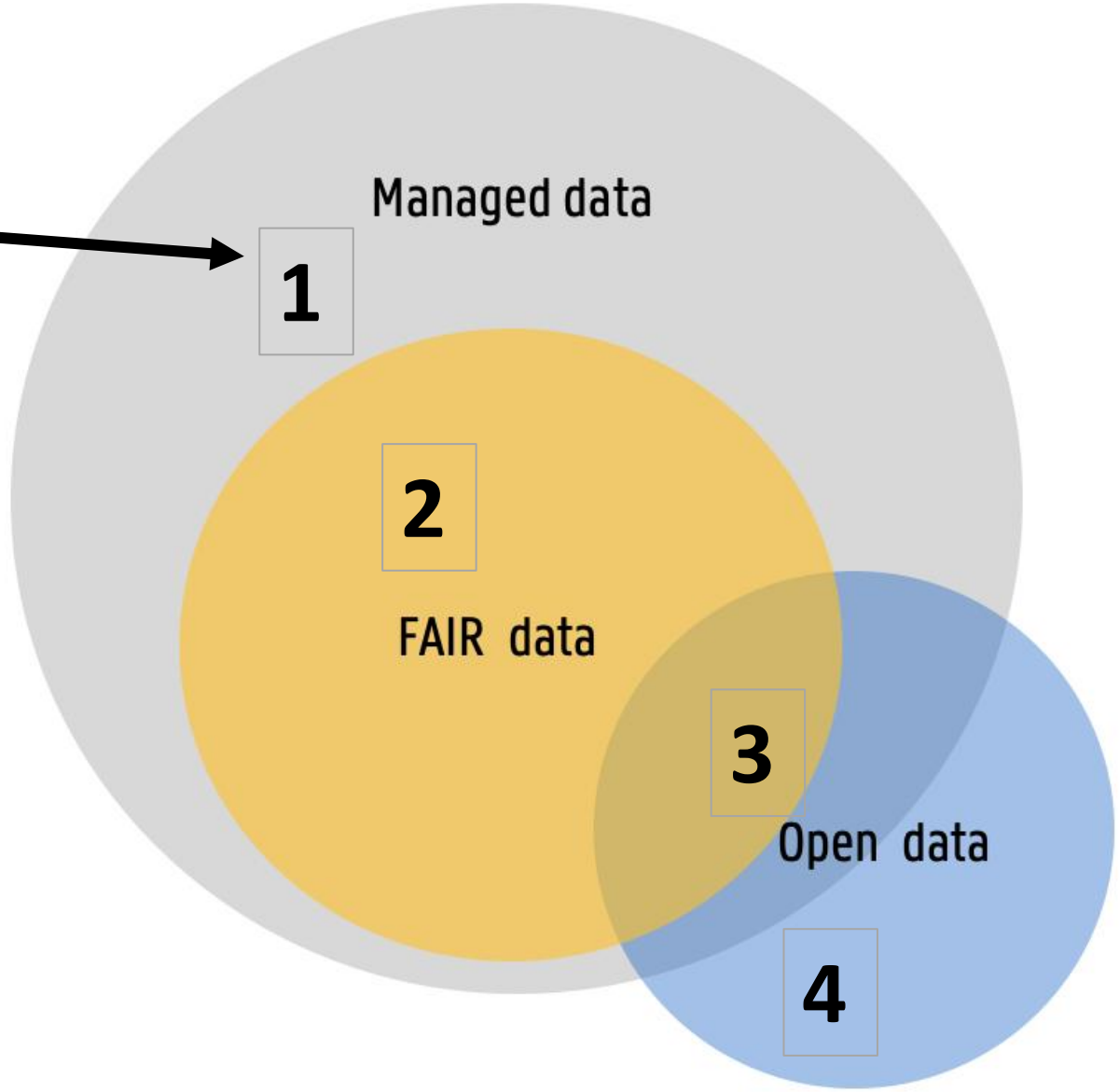
Good RDM habits facilitate **FAIR data sharing**, with **open science** being the cherry on top of the cake

The bigger picture



The bigger picture

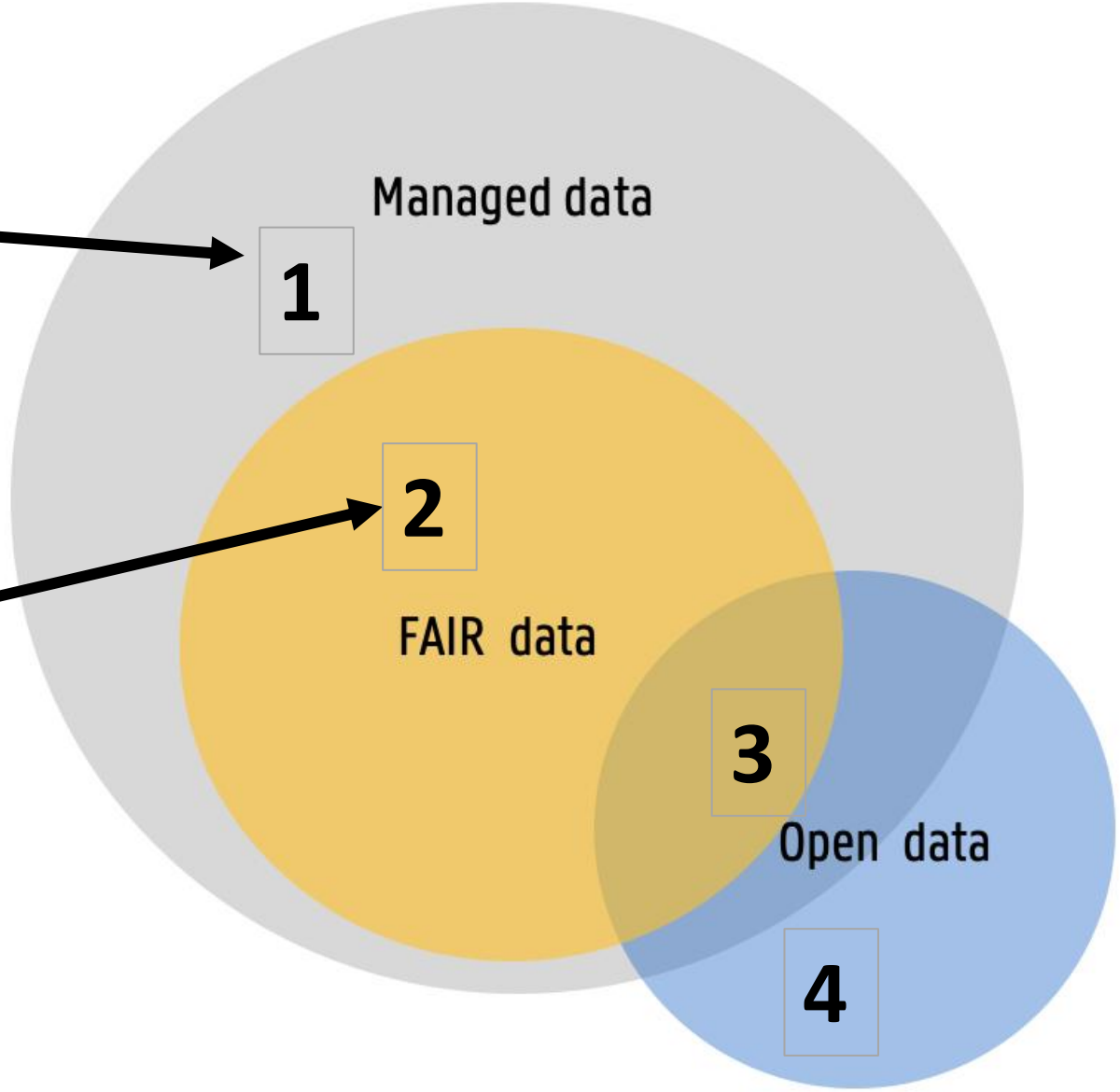
Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations...



The bigger picture

Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations...

... they are suitable for sharing, at least the metadata is, so they are made available on an open, structured repository...

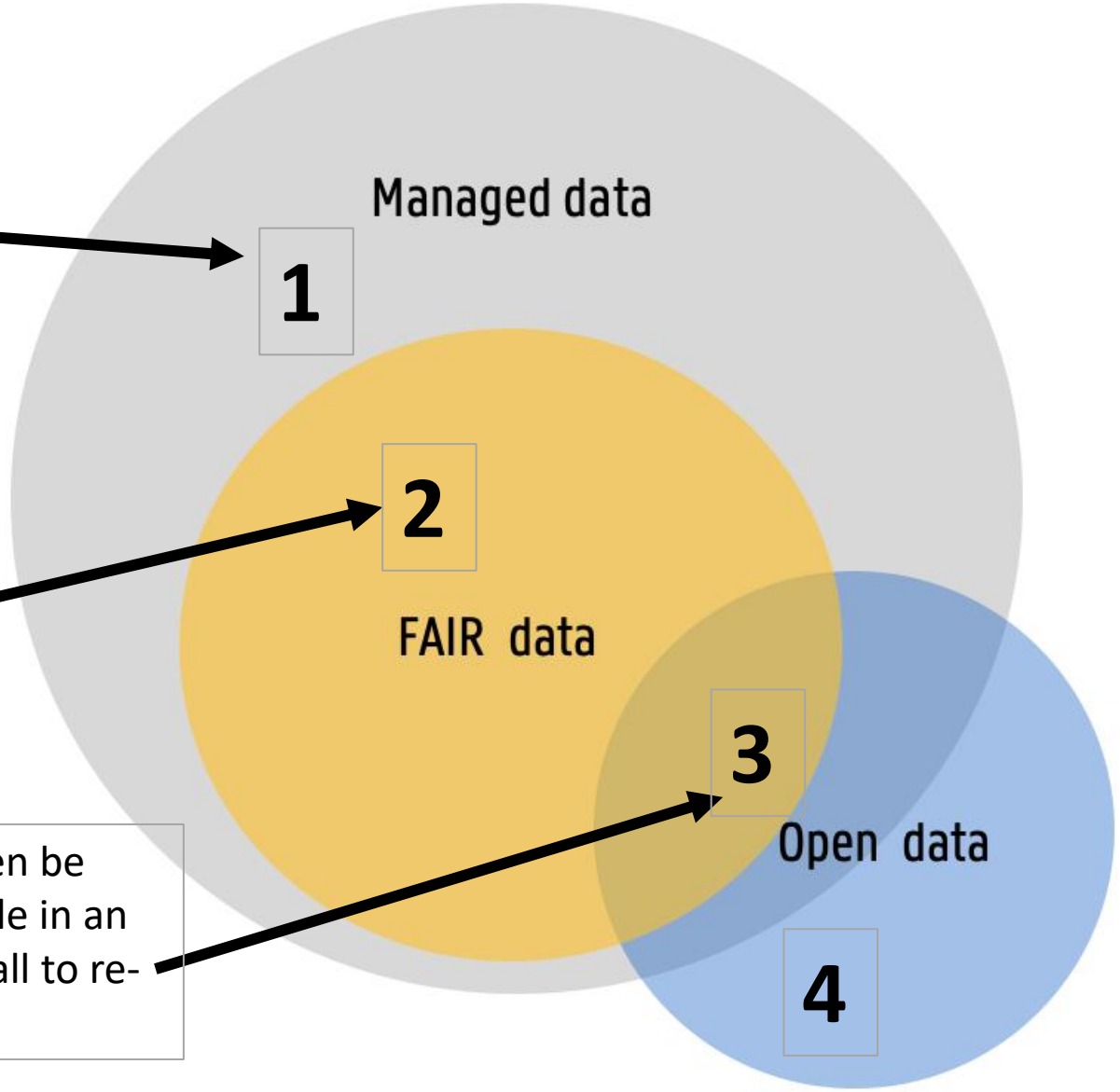


The bigger picture

Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations...

... they are suitable for sharing, at least the metadata is, so they are made available on an open, structured repository...

... they can even be made accessible in an open way, for all to re-use

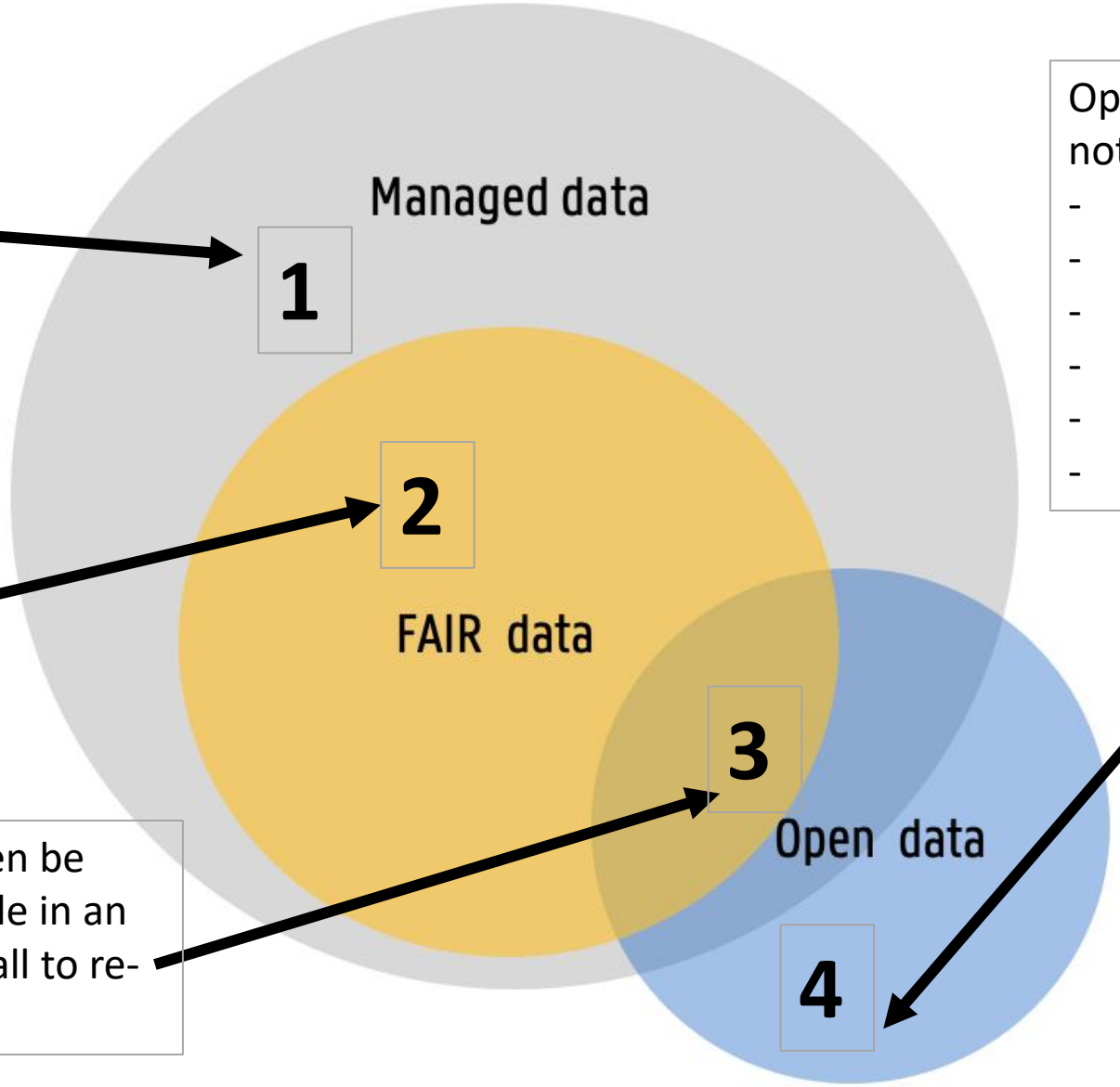


The bigger picture

Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations...

... they are suitable for sharing, at least the metadata is, so they are made available on an open, structured repository...

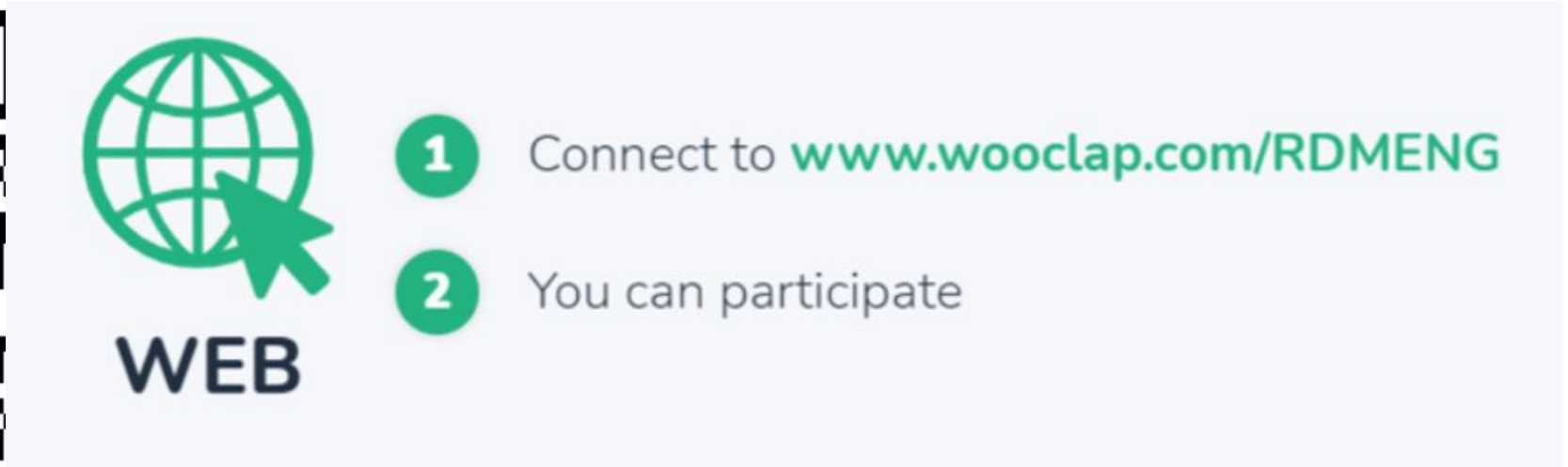
... they can even be made accessible in an open way, for all to re-use



Open data that is not FAIR or not well managed can be:

- Fraudulent
- Of poor quality
- Poorly described
- Lacking metadata
- In a proprietary format
- ...

Thank you!

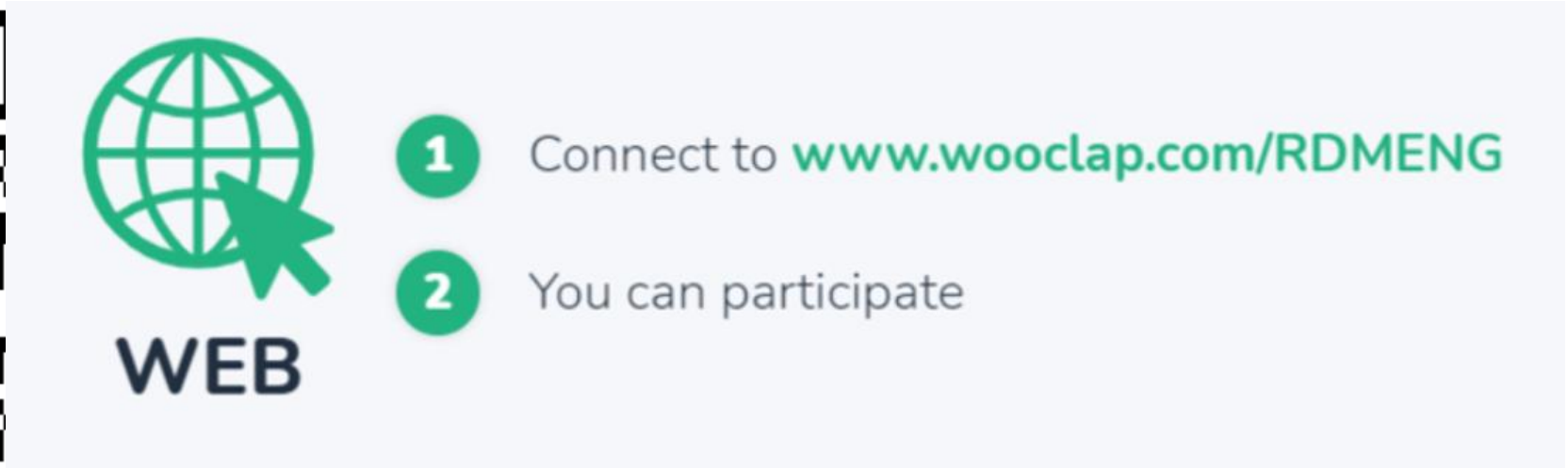


Quizz!

The bigger picture

Questions break

Thank you!



How are you feeling?

Sources

[*]
J. Ioannidis, 2005, Contradicted and Initially Stronger Effects in Highly Cited Clinical Research, JAMA. 2005;294(2):218-228. doi:10.1001/jama.294.2.218

Mark Otto Baerlocher et al., 2010, Data integrity, reliability and fraud in medical research, Elsevier European Journal of Internal Medicine 21 (2010) 40–45

Monya Baker, 1,500 scientists lift the lid on reproducibility, Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

Sources

“Three Camps, One Destination: The Intersections of Research Data Management, FAIR and Open”.
Higman, Rosie, Daniel Bangert, and Sarah Jones. 2019. *Insights* 32 (1): 18.
DOI: http://doi.org/10.1629/uksg.468

Formations courtes en ligne sur la gestion responsable des données, Macalester College, Minnesota, consulté le 18/09/20
https://libguides.macalester.edu/c.php?g=527786&p=3608657

What Is Data Quality and Why Is It Important? Aaron Moss, PhD, consulté le 18/09/20
https://www.cloudresearch.com/resources/guides/ultimate-guide-to-survey-data-quality/guide-data-quality-what-is-data-quality-why-important/

Research data management explained, University of Leeds, consulté le 13/09/20
https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

Fostering the practical implementation of Open Science in Horizon 2020 and beyond, consulté le 14/08/20
https://www.fosteropenscience.eu/node/1420

Ice Cream Sales Lead to Higher Homicide Rates: How Correlation Doesn’t Always Equal Causation, consulté le 18/09/20
https://www.egenerationmarketing.com/blog/causation-and-correlation-for-a-law-firm

How researchers dupe the public with a sneaky practice called "outcome switching“, consulté le 17/09/20
https://www.vox.com/2015/12/29/10654056/ben-goldacre-compare-trials

Quick Data Lessons: Data Dredging, consulté le 03/09/20
https://www.geckoboard.com/blog/quick-data-lessons-data-dredging/

Sources

You Can’t Trust What You Read About Nutrition , consulté le 16/09/20
https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/

Science Isn’t Broken, consulté le 16/09/20
https://fivethirtyeight.com/features/science-isnt-broken/#part1

Page « PACE trial » on me-pedia, consulté le 16/09/20
https://me-pedia.org/wiki/PACE_trial

For my next trick... Consulté le 17/09/20
https://www.economist.com/science-and-technology/2016/03/26/for-my-next-trick

Simmons, Nelson & Simonsohn, 2011
https://journals.sagepub.com/doi/pdf/10.1177/0956797611417632

American Statistical Association (ASA) Statement on Statistical Significance and P-Values, 2010
https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

Baerlocher et al., 2010
https://www.sciencedirect.com/science/article/pii/S0953620509002337

Sources

Manon Knockart et Thomas Tombal, « Quels droits sur les données », in *Actualités en droit du numérique*, Limal, Anthémis, 2019, p. 53 et suiv.

Thierry Léonard, Bojana Salovic, Olivia Guerguinov, « Protection des données : quel cadre juridique pour la recherche scientifique en Belgique ? », blog Droit et Technologies, 1er avril 2019 https://www.droit-technologie.org/wp-content/uploads/2019/04/v2.pdf (consulté le 24 février 2021)

Lionel Maurel, « A qui appartiennent les données de la recherche ? », Webinaire Tuto@Mate organisé par le Réseau Méthodes Analyses Terrains Enquêtes en SHS le14 septembre 2020 https://mate-shs.cnrs.fr/wp-content/uploads/2020/09/tuto25-mate-Données-de-recherche.pdf (consulté le 24 février 2021)

Anne-Laure Stérin, Camille Noûs, « Ouverture des données de la recherche : les mutations juridiques récentes », *Tracés. Revue de Sciences humaines* [En ligne], #19 | 2019, mis en ligne le 22 juillet 2020 http://journals.openedition.org/traces/10603 (consulté le 24 février 2021)

Questions juridiques liées aux données de recherche, interview de Lionel Maurel réalisée à l'occasion de la séquence de com' : La licence ouverte, à l'Inist-CNRS (Nancy) le 02 juillet 2019 https://doranum.fr/aspects-juridiques-ethiques/questions-juridiques-liees-aux-donnees-de-la-recherche/ (consulté le 24 février 2021)

Sources

Herbert Gruttemeier, Thérèse Hameau, « Accès aux données scientifiques et contraintes juridiques – une question d'équilibre », *I2D - Information, données & documents*, 2016/2 (Volume 53), p. 20-22 https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-20.htm (consulté le 24 février 2021)

Lionel Maurel, Données de la recherche et questions juridiques au sein des plans de gestion de données https://www.hisoma.mom.fr/sites/hisoma.mom.fr/files/docs/Recherche/quinquenal-2016-2020/axe-t/emmanuelle-morlock/seminaire_pgd_juridique_maurel.pdf (consulté le 24 février 2021)