ITINERIS

Developing Open Science on Cloud usingJupyter notebooks

About the course

Zhiming Zhao

University of Amsterdam,

LifeWatch Virtual Lab & Innovation Center (VLIC)

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System (D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Dr. Zhiming Zhao (z.zhao@uva.nl)

Associate Professor, Chair of MultiScale Networked Systems, UvA,
Technical manager, LifeWatch ERIC Virtual Lab and Innovation Center

Research areas

- Cloud computing and software-defined infrastructure
- Time-critical cloud application and infrastructure optimization
- Big data management, scientific workflow management and data-intensive systems
- Blockchain, Decentralized marketplaces

Teaching courses

- Distributed and Parallel Programming (DPP)
- DevOps and Cloud-based Software Development (DevOps)

Recent research projects

- LTER-LIFE, BioDT, VRE4EIC, BlueCloud, BlueCloud 2026: Digital twin, Virtual Research Environment
- ENVRI-plus, ENVRI-FAIR, ENVRI-HUB Next: Big data management
- SWITCH, EVERSE: Software engineering, time - critical cloud applications, DevOps, and infrastructure automation
- ARTICONF, CLARIFY, LIFEWATCH: Blockchain, decentralized marketplace

Dr. Gabriel Pelouze

VRE DevOps engineer, LifeWatch ERIC Virtual Lab and Innovation Center

Research areas

- Data management
- Cloud virtual research environment

Teaching courses

- DevOps and Cloud-based Software Development (DevOps)
- VRE training in EGU, EGI and LifeWatch

Recent research projects

- LTER-LIFE, BioDT,
- ENVRI-HUB Next: Big data management

Dr. Spiros Koulouzis

VRE DevOps engineer, LifeWatch ERIC Virtual Lab and Innovation Center

🌐Research areas

- Data management
- Cloud virtual research environment

🌐Teaching courses

- DevOps and Cloud-based Software Development (DevOps)
- VRE training in EGU, EGI and LifeWatch

🌐Recent research projects

- LTER-LIFE, BioDT,
- ENVRI-HUB Next: Big data management

What is this course about?

- A course (24 hours) aims to train junior data scientists, such as PhD students and new Postdocs, to learn technologies and practices for conducting research activities in a Virtual Research Environment.

- The course will be delivered using a project-based teaching method.

- The course will use LifeWatch Notebook-as-a-VRE (NaaVRE) for lab assignments and course projects.

Learning objectives

1. **Understand** the basic concepts of Virtual research environment, Research Infrastructure, Scientific workflow, and cloud computing;
2. **Understand** the basic techniques behind the Virtual research environment;
3. Able to **load** external data into the Jupyter environment and develop data-intensive applications;
4. Able to **scale** notebooks out as cloud workflows;
5. Able to **apply** basic research software quality control practices;
6. Able to **develop** a small-size research project using data management, cloud computing, and workflow technologies

Structure of the course

- Module 1: Open science on Cloud
- Module 2: Enhance research activities using Virtual Research Environment
- Module 3: Open science project

Teaching methods

- Lectures
- Tutorial
- Group projects
- Assessment

Lectures

- **Day 1:**
  - **Lecture**: Introduction to Open Science, Jupyter and Virtual research environment
  - **Tutorial**: NaaVRE
  - **Group project**: project plan
- **Day 2:**
  - **Group project pitch**
  - **Lecture**: Open science technologies: search, workflow, cloud computing
  - **Group project**: development
  - **Student presentation: progress and results**
- **Day 3:**
  - **Lecture**: Open science technologies: Digital Twin and AI
  - **Group project:** towards digital twin
  - **Exam and final presentation**

Group project

- A project group of 2-3 students
- Project basic part
  - Should be data science and workflow related;
  - Can be from your own research context
- Project advanced part
  - Design a Digital Twin

Assessment

- A light exam on concepts (40%)
- Project presentations (30% *2)

Discussions

- Questions?

To know more about you

- https://forms.gle/z5CN86Robmn3LHFh8

ITINERIS

Open science challenges and practices

Zhiming Zhao

University of Amsterdam,

LifeWatch Virtual Lab & Innovation Center (VLIC)

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Day 1: outline

ITINERIS

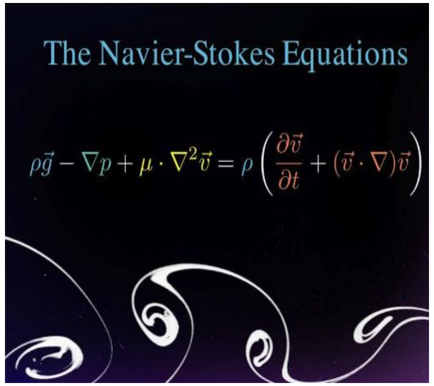
- Why Open Science?
- Data science project
- Data science technologies

1. Science paradigms

Science paradigms

1st paradigm:
empirical: observing and describing nature

2nd paradigm:
theoretical: using models and generalization


The Navier-Stokes Equations
$\rho\vec{g} - \nabla p + \mu \cdot \nabla^2 \vec{v} = \rho\left(\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \nabla)\vec{v}\right)$

3rd paradigm:

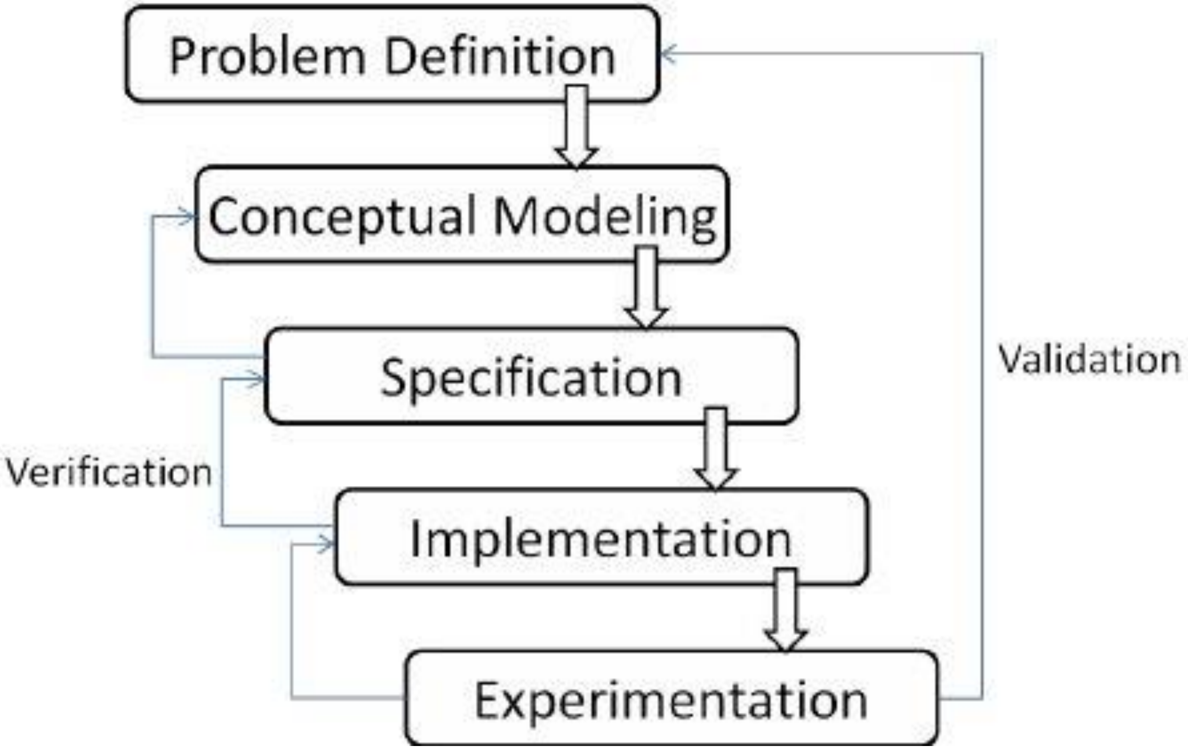
Computational: modelling and simulating complex phenomena

4th paradigm:

Data-intensive: big data, machine learning ...

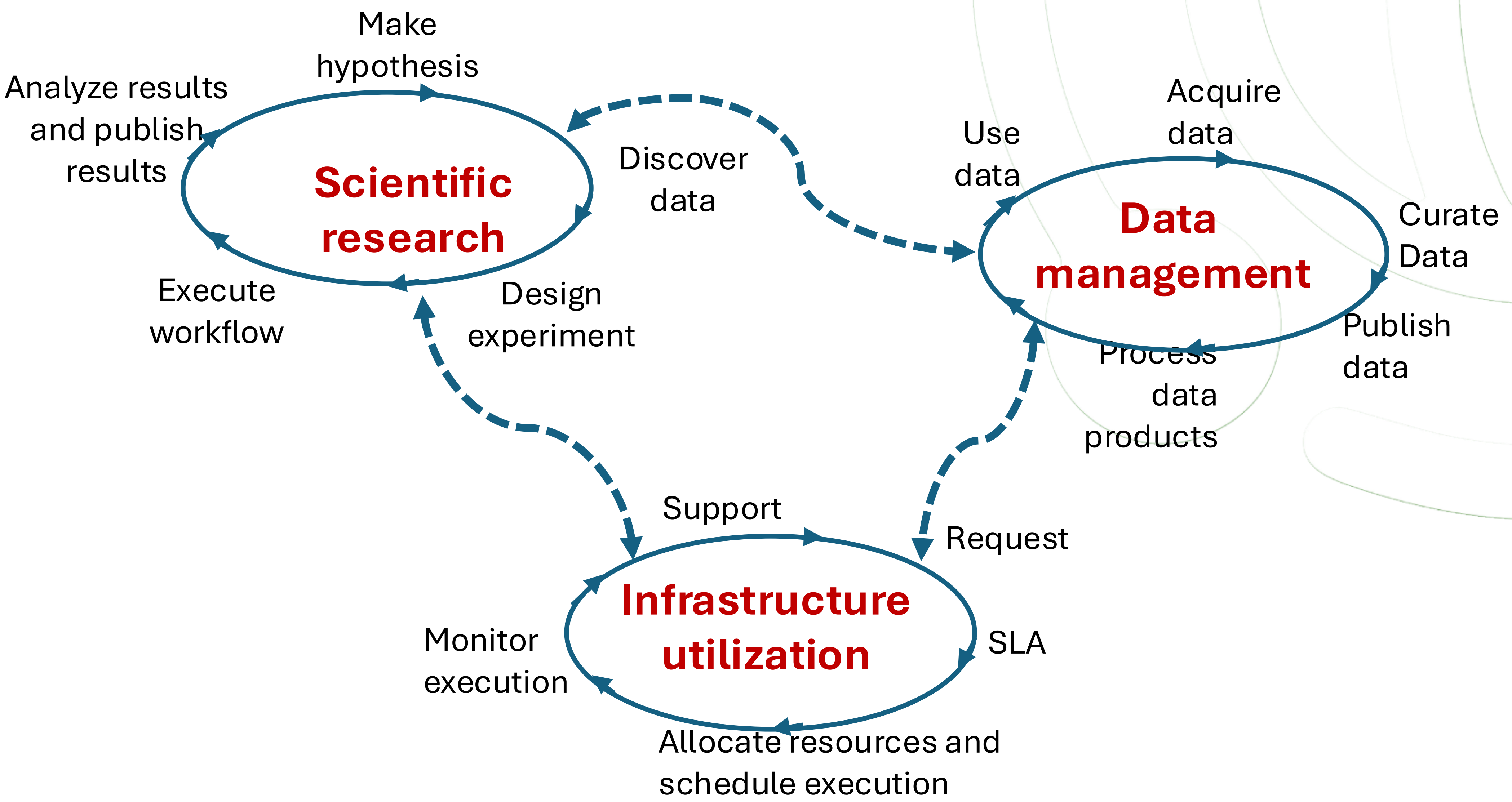
3rd paradigm: modelling and simulation

- Modelling and simulation lifecycle

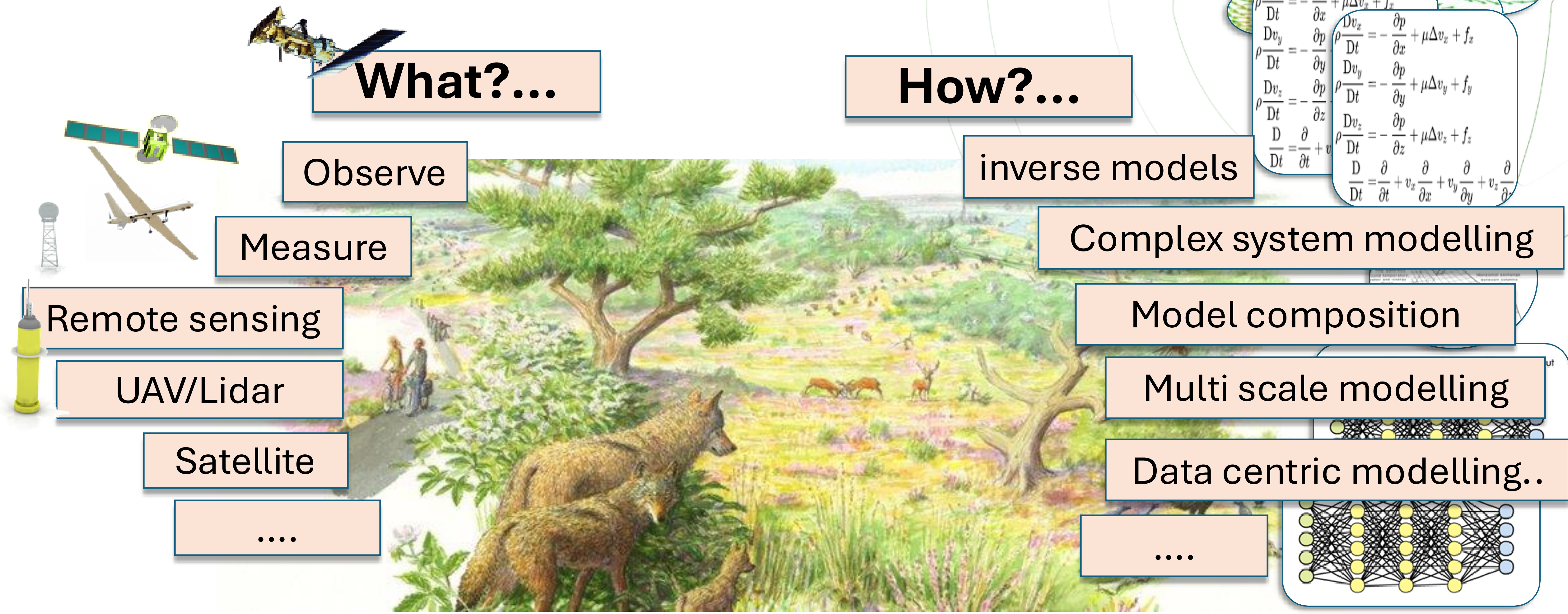


https://doi.org/10.4324/9780203886816

4th paradigm: data centric research activities



Ecosystem and data science



What?...

Observe

Measure

Remote sensing

UAV/Lidar

Satellite

....

How?...

inverse models

Complex system modelling

Model composition

Multi scale modelling

Data centric modelling..

....

Why?...

What-if

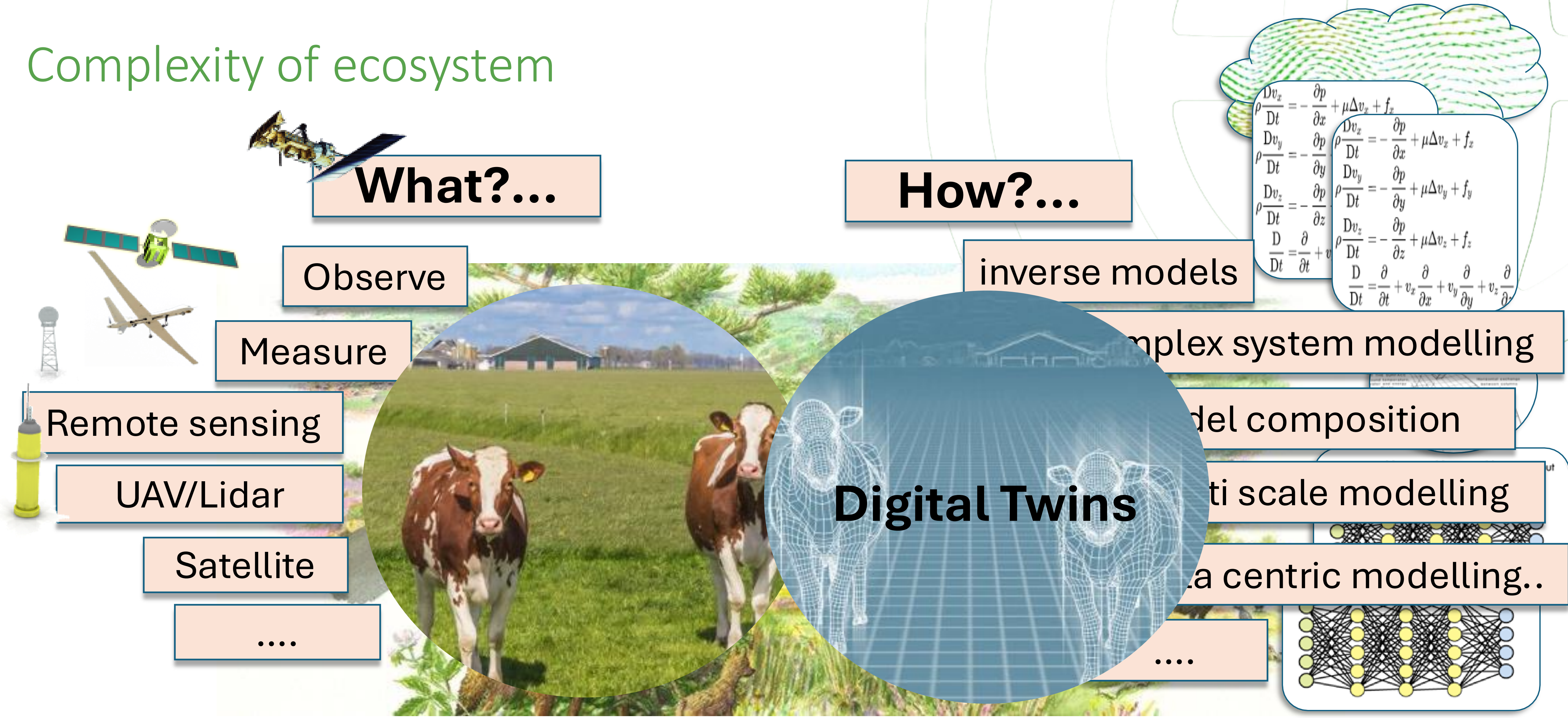
Scenarios

Workflows

Digital twins

...

Complexity of ecosystem



What?...

Observe

Measure

Remote sensing

UAV/Lidar

Satellite

....

How?...

inverse models

mplex system modelling

del composition

ti scale modelling

a centric modelling..

....

Digital Twins

Why?...

What-if

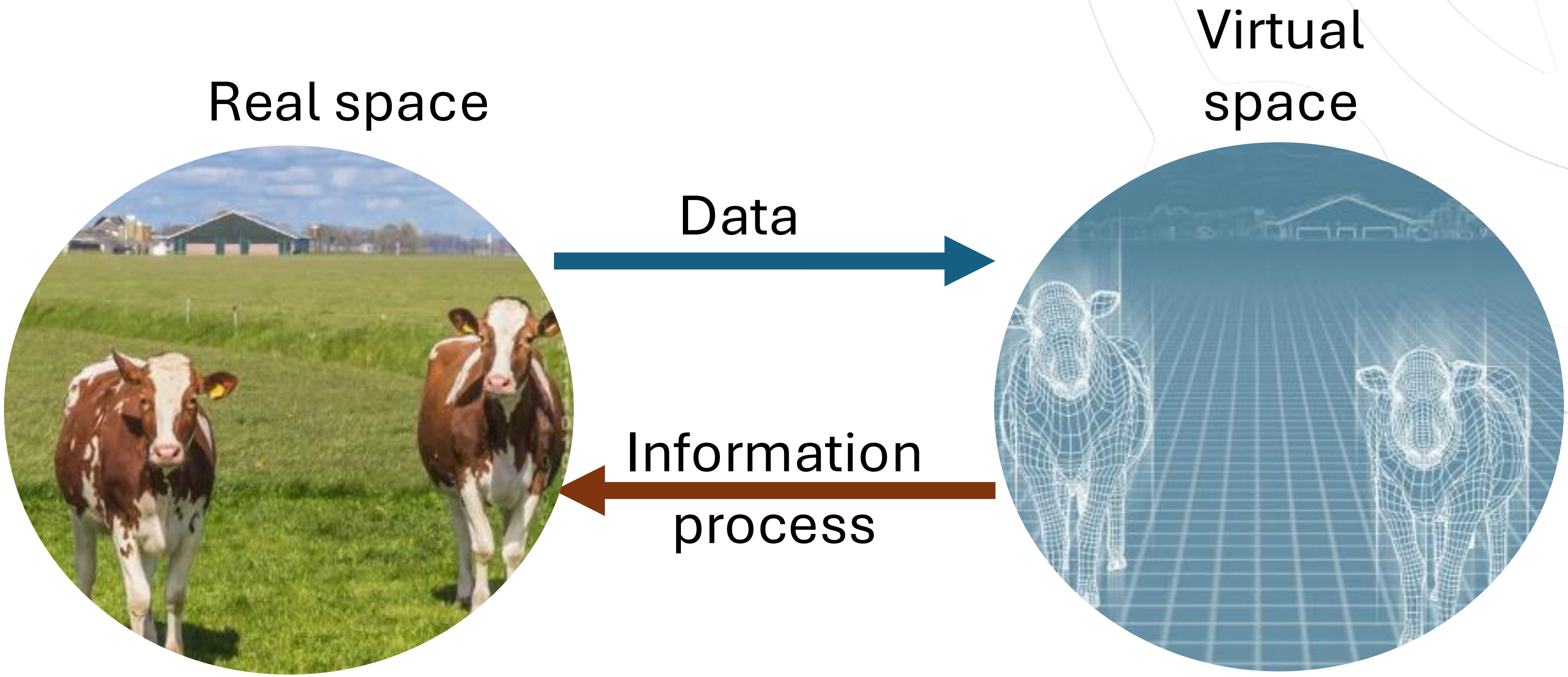
Scenarios

Workflows

Digital twins

...

1. Data between real and virtual spaces



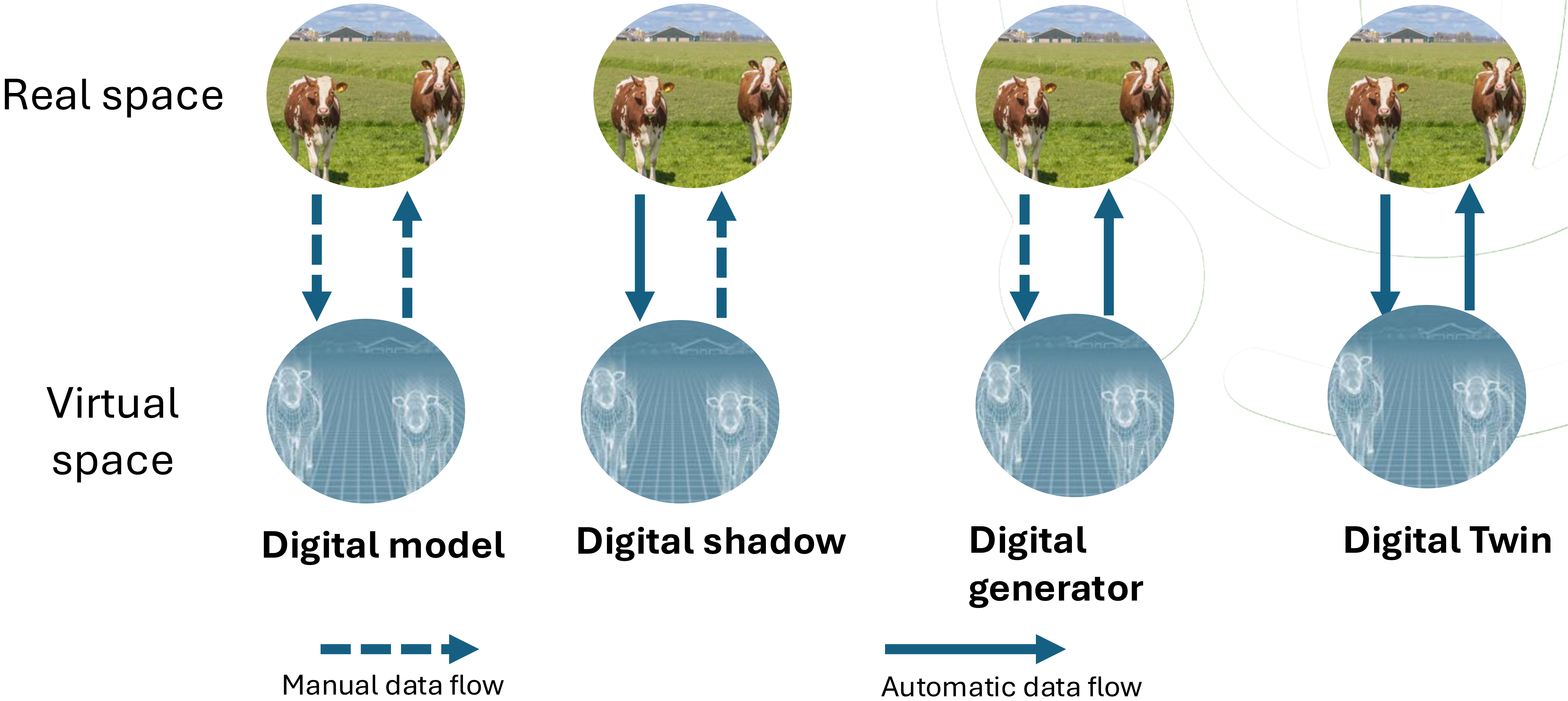
Real space

Virtual space

Data

Information process

Digital twining



Real space

Virtual space

Digital model

Digital shadow

Digital generator

Digital Twin

Manual data flow

Automatic data flow

Tekinerdogan & Verdouw 2020, https://doi.org/10.3390/s20185103

Why Digital Twins in environmental and earth sciences

- Understanding ecosystems

- Scenario studies

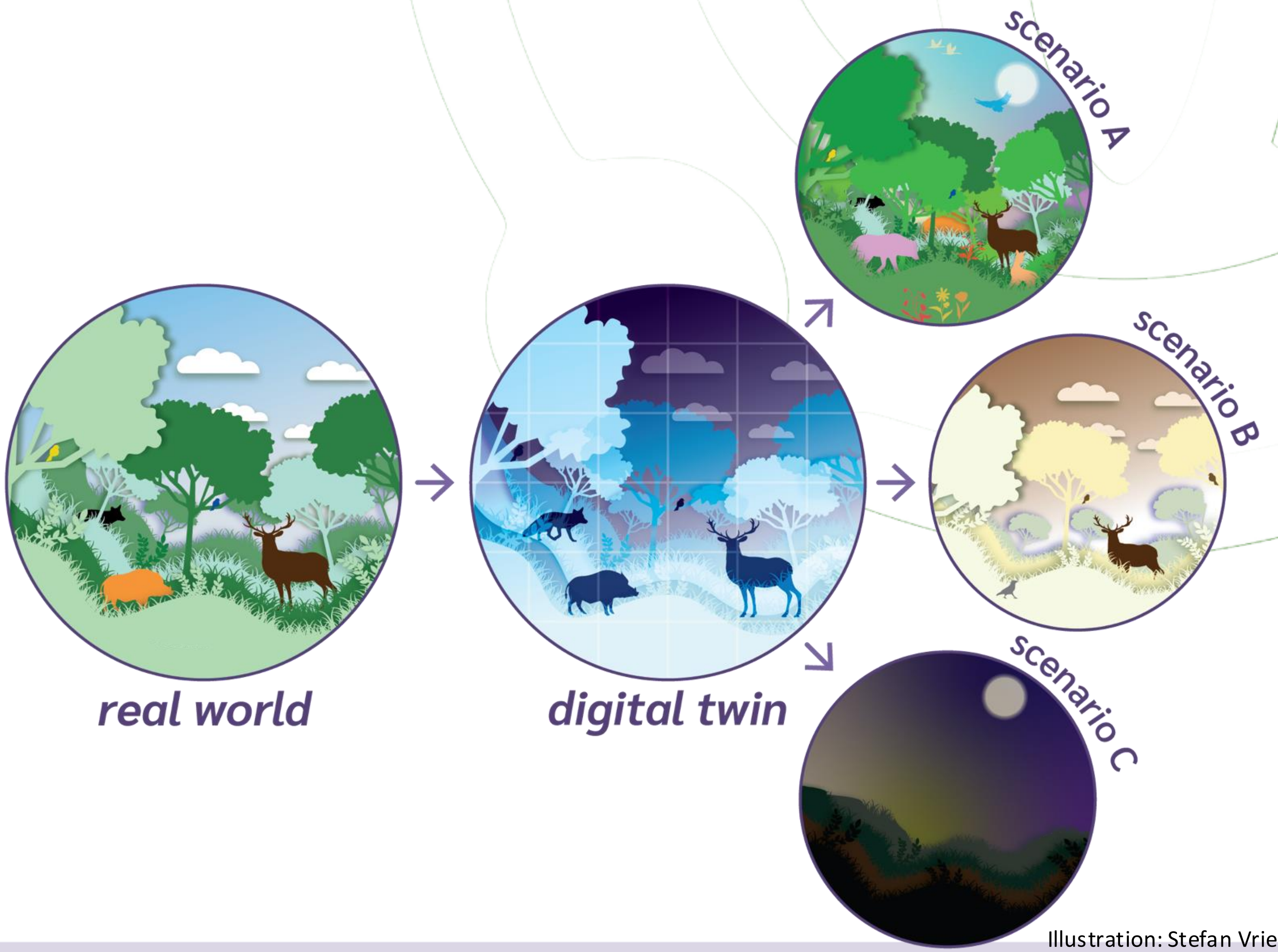
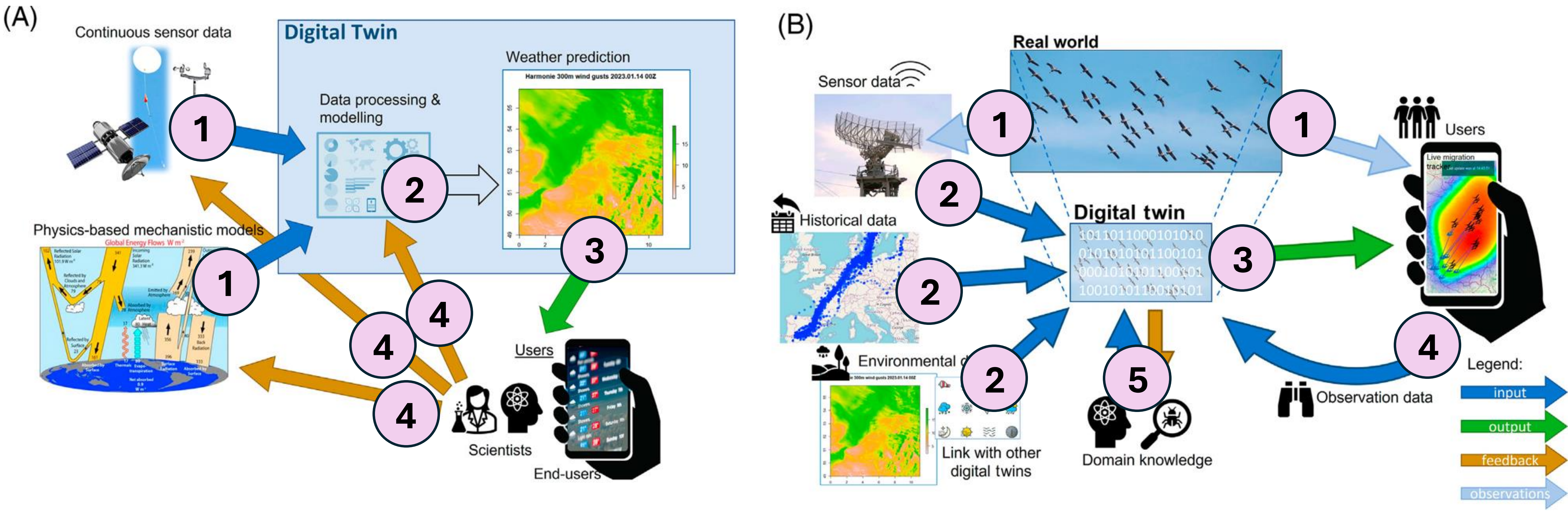


Illustration: Stefan Vriend

Destination Earth

*Analyse the **past**, monitor the **present**, predict the **future***

Digital twin in ecosystem research



https://doi.org/10.1016/j.tree.2023.04.010

Digital twin for ecological research?

Discussion: can we make a digital twin?

2. Data science project

Discussion: How do you do teamwork in a project?

Outline

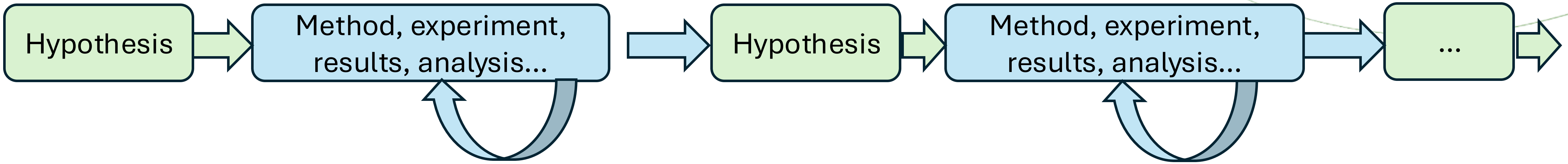
- Collaboration patterns
- Agile
- Version control
- Discussion

Collaboration patterns in project teamwork

- Leader intitialized collaboration
- Multi group based initialized collboration
- Brain storm collaboration
- Agile collaboration
- Water fall collaboration

Agile for science

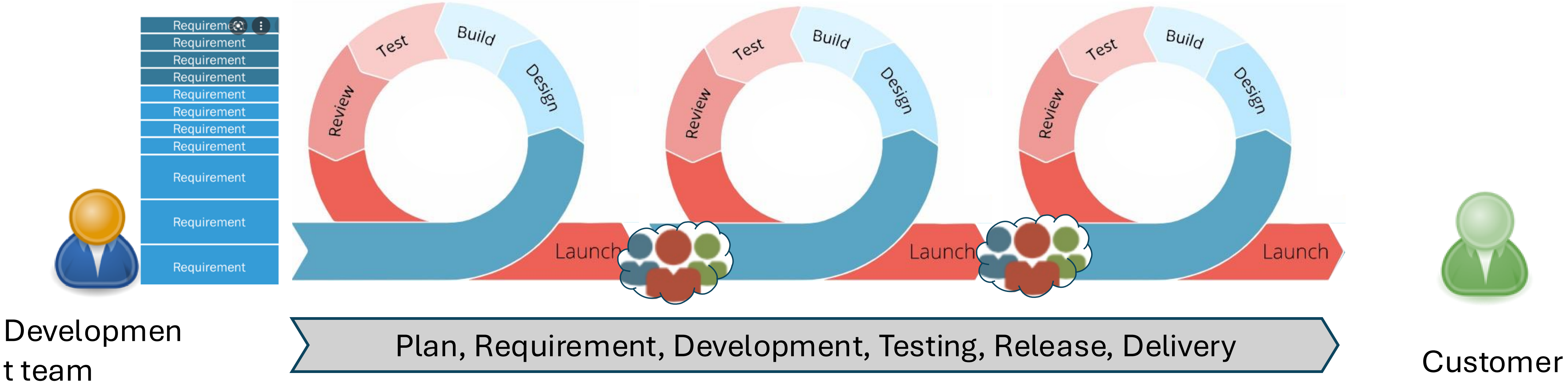
- Research is often a trajectory of hypothesis- study
- It has a high uncertainty journal
- Managing the process sequentially has a high risk
- Iterative development can detect risk


Hypothesis
Method, experiment, results, analysis…
Hypothesis
Method, experiment, results, analysis…
…

https://www.cell.com/matter/fulltext/S2590-2385(23)00510-6

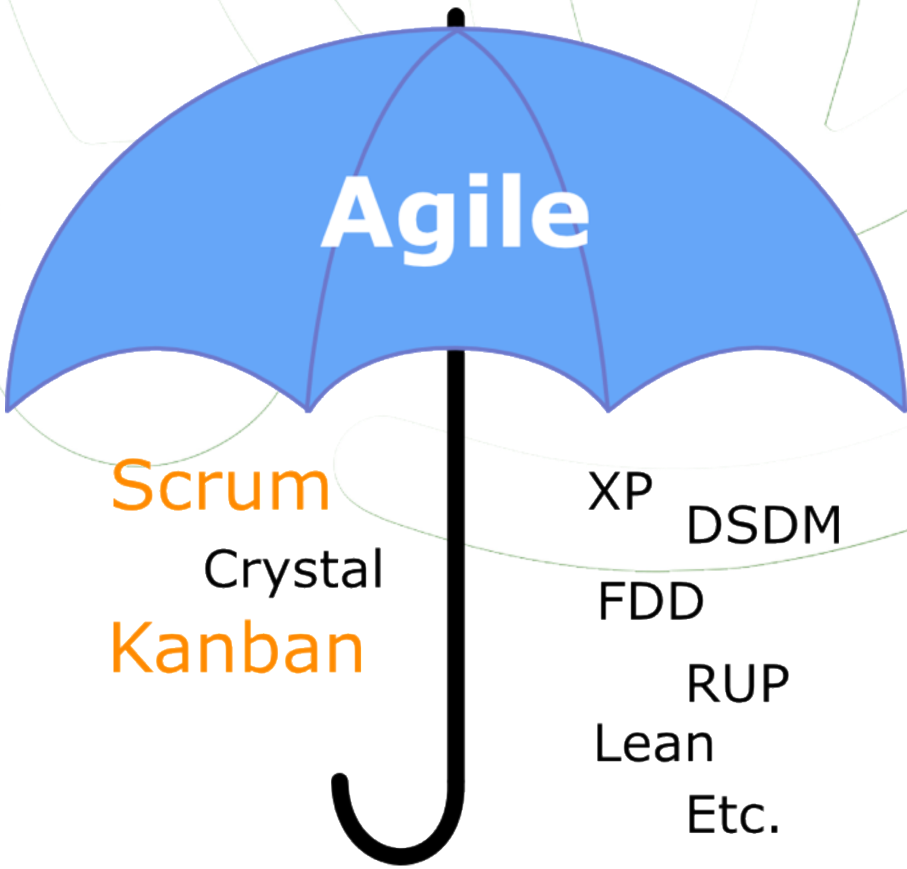
Agile model

- **Reduce** the **waiting time** of customers by increasing the **delivery frequency**
- **Improve delivery efficiency** by flexibly planning and scheduling activities
- **Reduce the development risks** by improving the review and adaptation **cycle**
- ..



Under the umbrella of **Agile**

- Feature Driven Development (FDD)
- Dynamic System Development Method (DSDM)
- Behavior Driven Development (BDD):
- Extreme program (XP)
- Kanban
- Crystal
- Lean
- Test Driven Development (TDD)
- Scrum
- …



Ref. Malek Al-Zewairi et al.: Agile Software Development Methodologies: Survey of Surveys, JCC, V5 (5)
https://m.scirp.org/papers/75114

Discussion

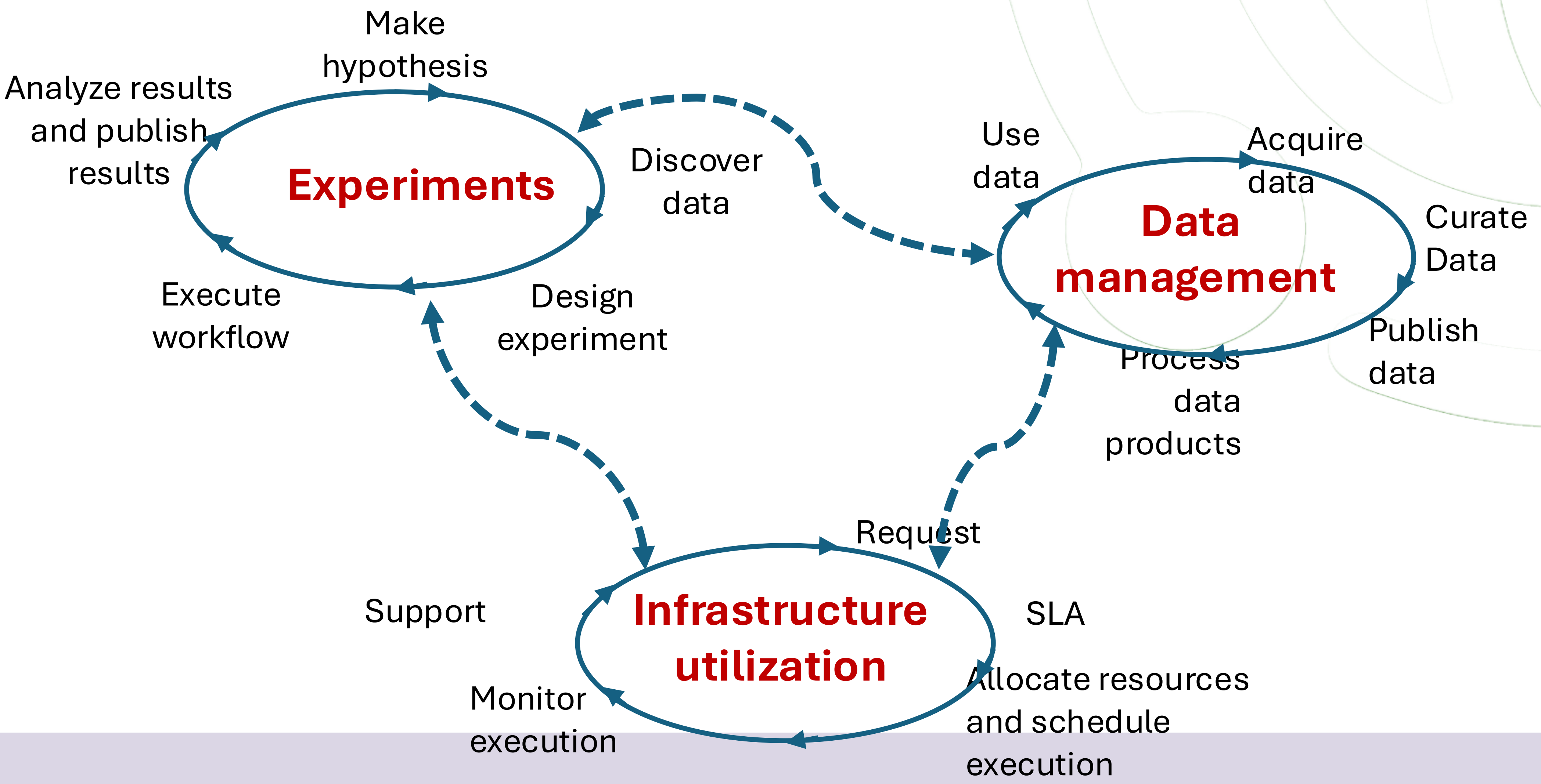
- How will you manage your team work?

3. Research support systems

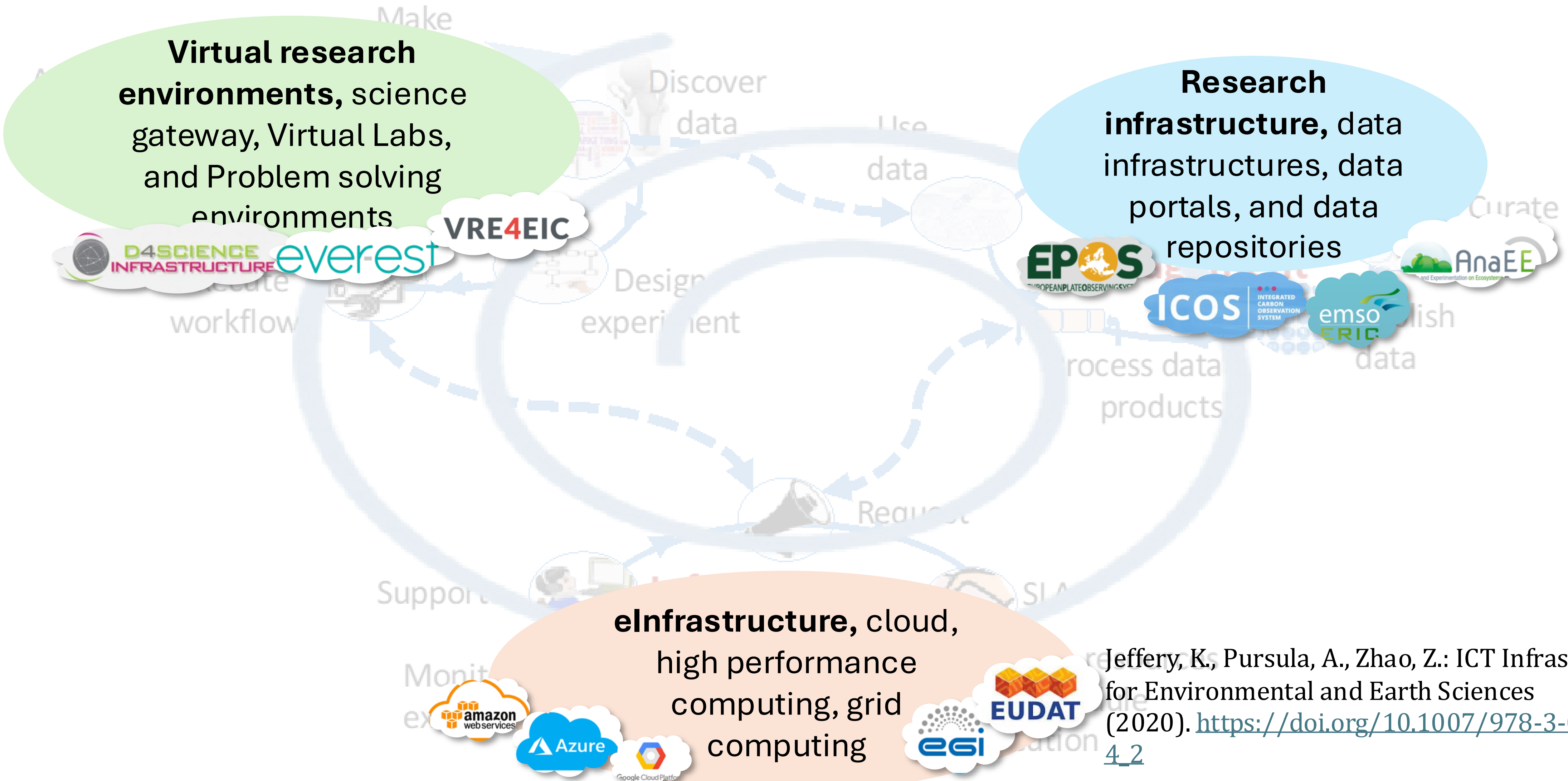
About the lecture

- Introduce the basic concepts and technologies in research support systems

Research activities



Research support systems



Virtual research environments, science gateway, Virtual Labs, and Problem solving environments

Research infrastructure, data infrastructures, data portals, and data repositories

eInfrastructure, cloud, high performance computing, grid computing

Jeffery, K., Pursula, A., Zhao, Z.: ICT Infrastructures for Environmental and Earth Sciences (2020). https://doi.org/10.1007/978-3-030-52829-4_2

VRE adoptions

The adoption of a VRE/VL depends on

- How close is it to the daily practice of a researcher?
- How effective can it solve the “pain points” of the research activities?
- How popular is it used by the community of the researcher?
- How many data, models, and other assets can it access?
- How sustainable is it?
- …

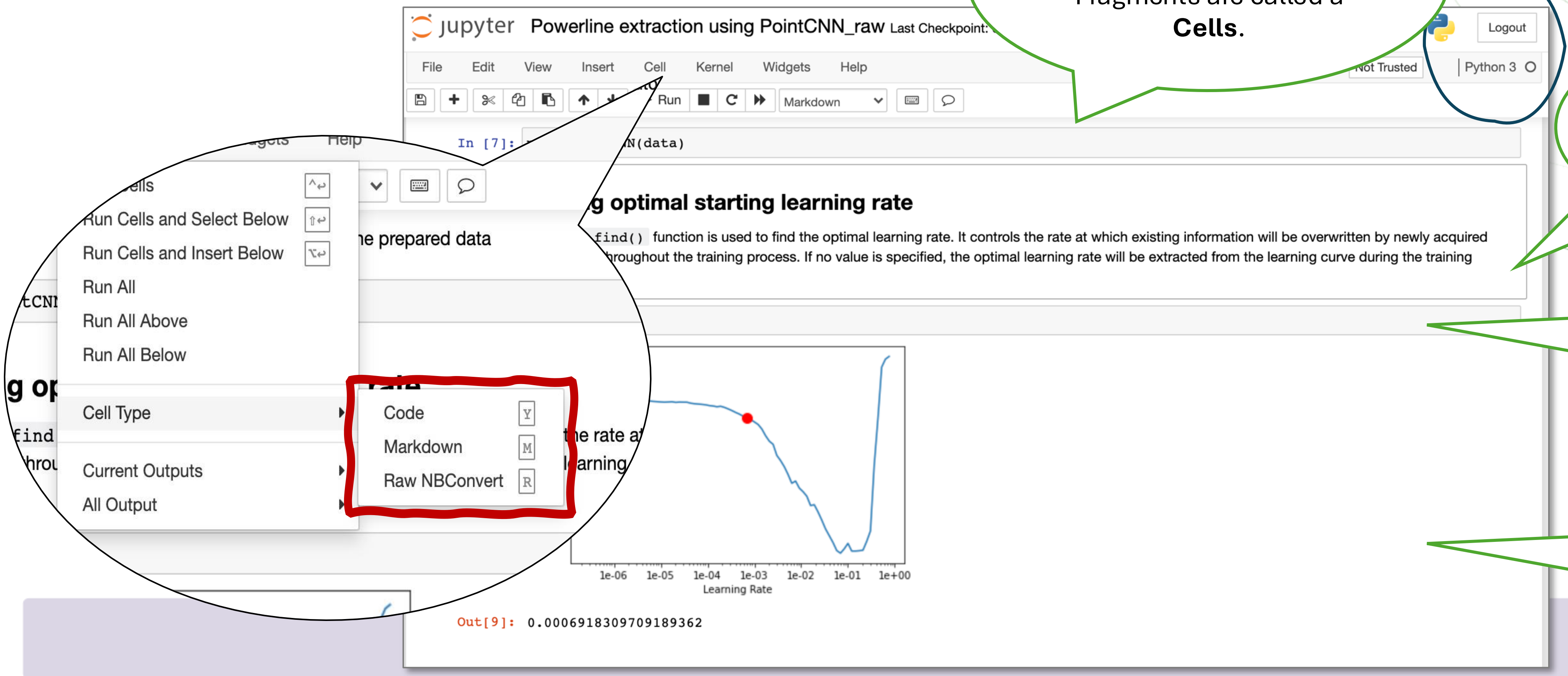
Notebook as a VRE: basic idea

Develop VRE functionality based on the Jupyter platform

- Discover and access research assets
- Design and automate experimental workflows
- Analyze and reproduce experimental results
- Collaborate with the community
- …

Jupyter.

- **Open source project,**
- **web applications**
- **Jupyter**: **Jul**ia, **Pyt**hon and **R**, and many m



g optimal starting learning rate

find() function is used to find the optimal learning rate. It controls the rate at which existing information will be overwritten by newly acquired hroughout the training process. If no value is specified, the optimal learning rate will be extracted from the learning curve during the training

```
Out[9]: 0.0006918309709189362
```

Notebook: Document developed in Jupyter. Fragments are called a **Cells**.

Markdown cell: rendered with format

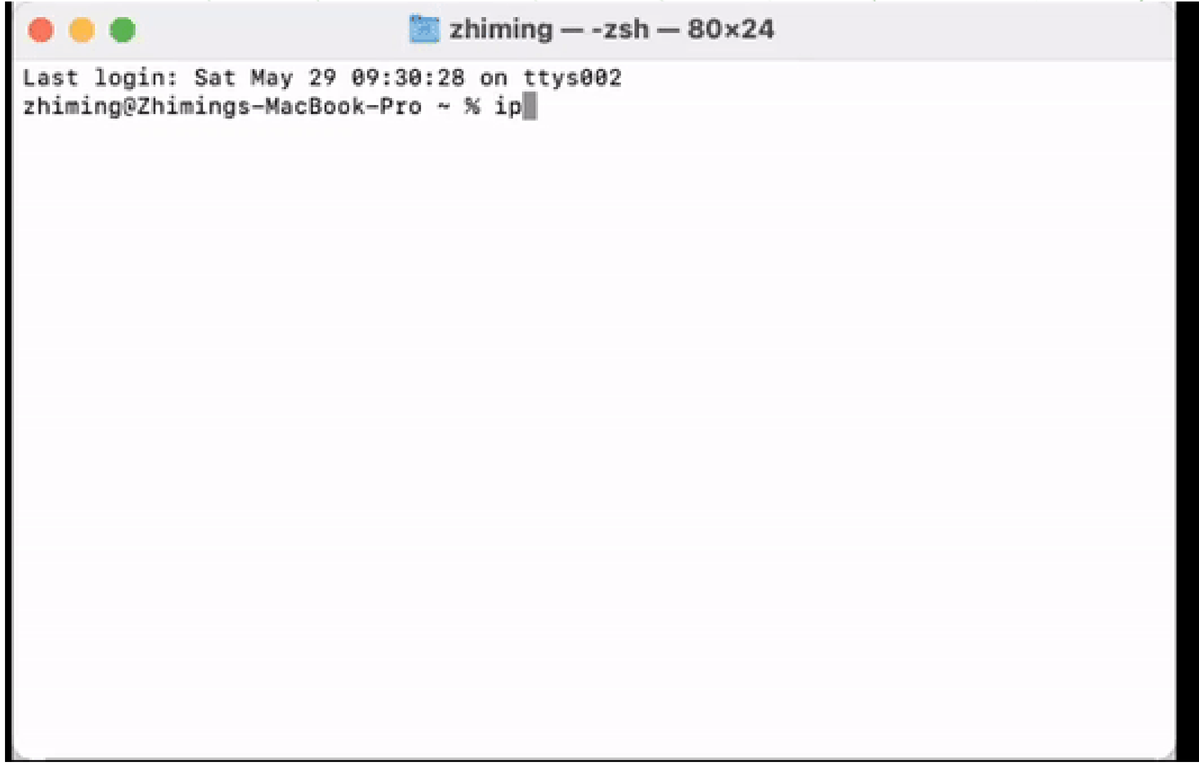
Code Cell: to be executed by the kernel of Jupyter

Output: the results after the execution of a code cell.

Why Jupyter: Interactive computing

ITINERIS

- A computing paradigm relies on input and output between a computer system and the user [1]
- Read-Eval-Print-Loop (REPL), e.g., a shell for a programming language
- E.g., Interactive Python shell (IPython)



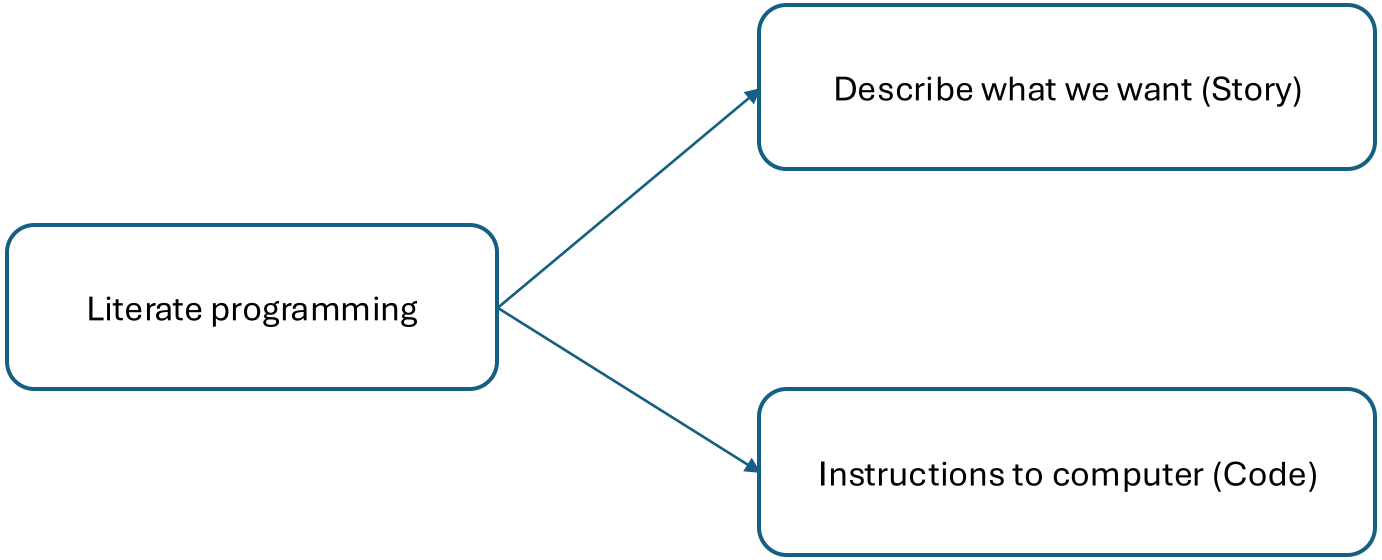
```
Last login: Sat May 29 09:30:28 on ttys002
zhiming@Zhimings-MacBook-Pro ~ % ip
```

IP[y]: IPython
Interactive Computing

[1] DOI: 10.4018/978-1-60566-368-5.ch001

Why Jupyter: literate programming

> Instead of imagining that our main task is to instruct a *computer* what to do, let us concentrate rather on explaining to *human beings* what we want a computer to do.
>
> *Literate programming (1984)*
> DONALD KNUTH



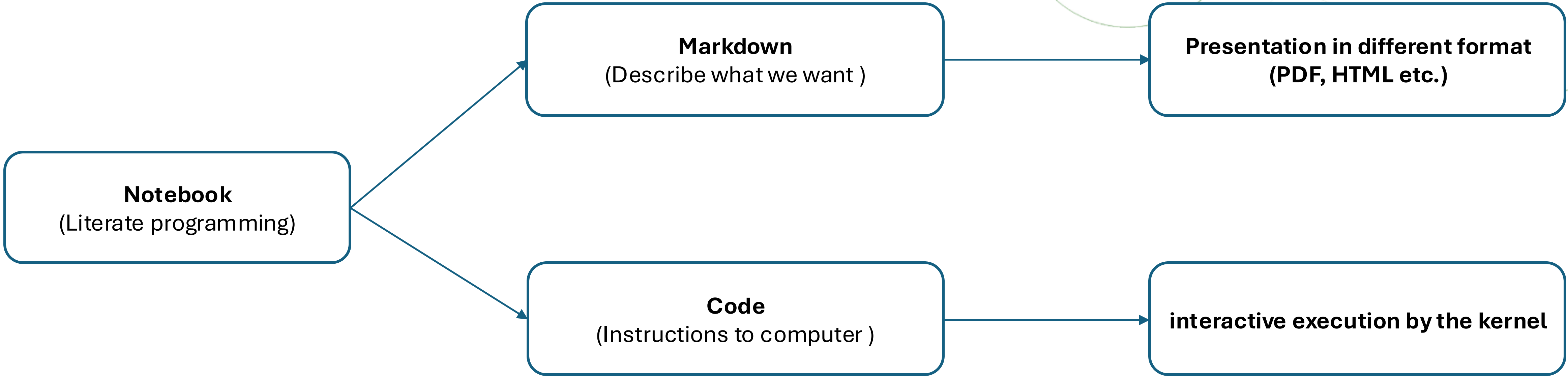
Literate programming

Describe what we want (Story)

Instructions to computer (Code)

Why Jupyter: literate programming

- Markdown texts: can be broader than comments inside the code, and have more options for layout and style;



Notebook
(Literate programming)

Markdown
(Describe what we want)

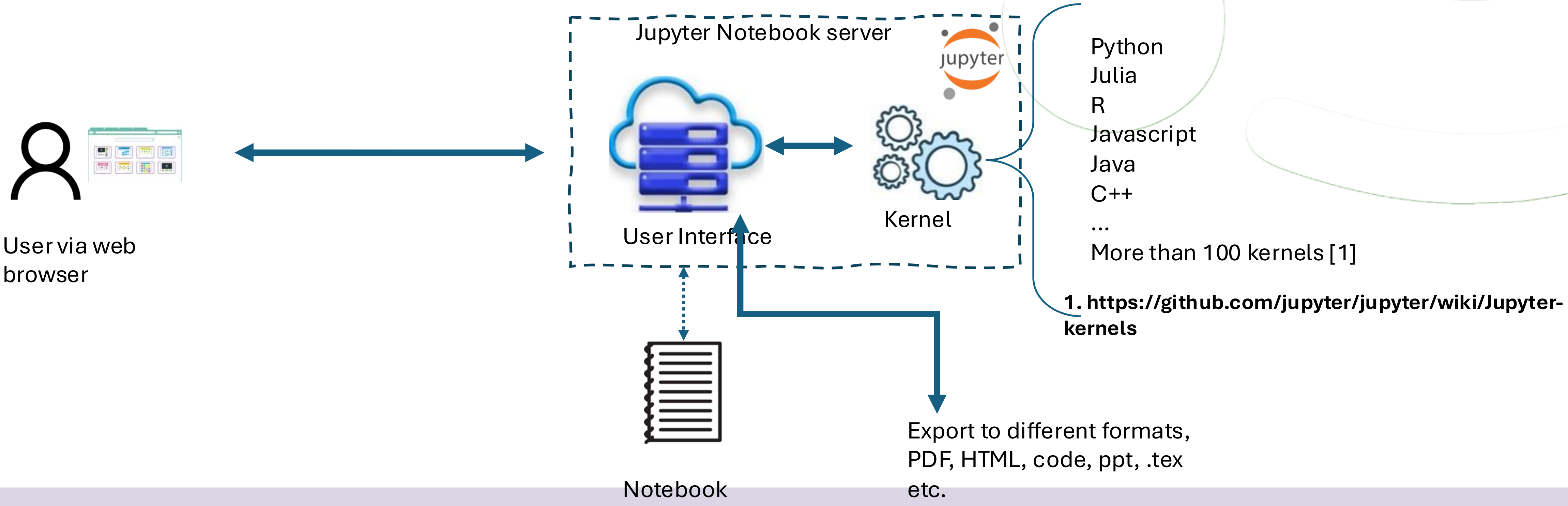
Presentation in different format
(PDF, HTML etc.)

Code
(Instructions to computer)

interactive execution by the kernel

How does Jupyter work?

- Client-server architecture



User via web browser

Jupyter Notebook server

User Interface

Kernel

Python
Julia
R
Javascript
Java
C++
...
More than 100 kernels [1]

1. https://github.com/jupyter/jupyter/wiki/Jupyter-kernels

Notebook

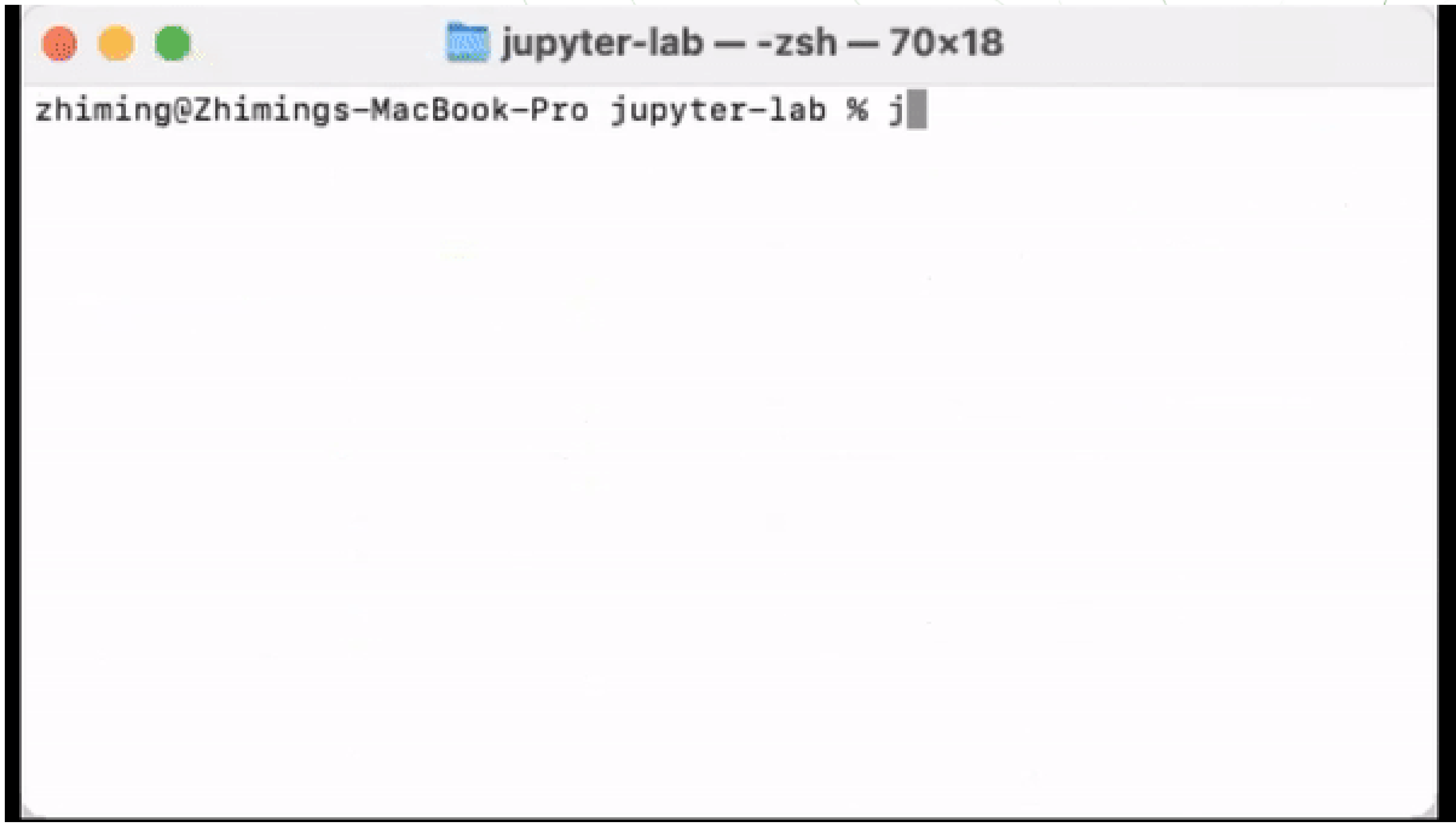
Export to different formats, PDF, HTML, code, ppt, .tex etc.

Different Jupyter working environment

- Via **Jupyter notebook** (single user, single notebook)
- Via **Jupyter Lab** (single user, multi notebooks)
- Via **Jupyter Hub** (multi users, each user has an independent Juter notebook/lab instance)

Jupyter Lab

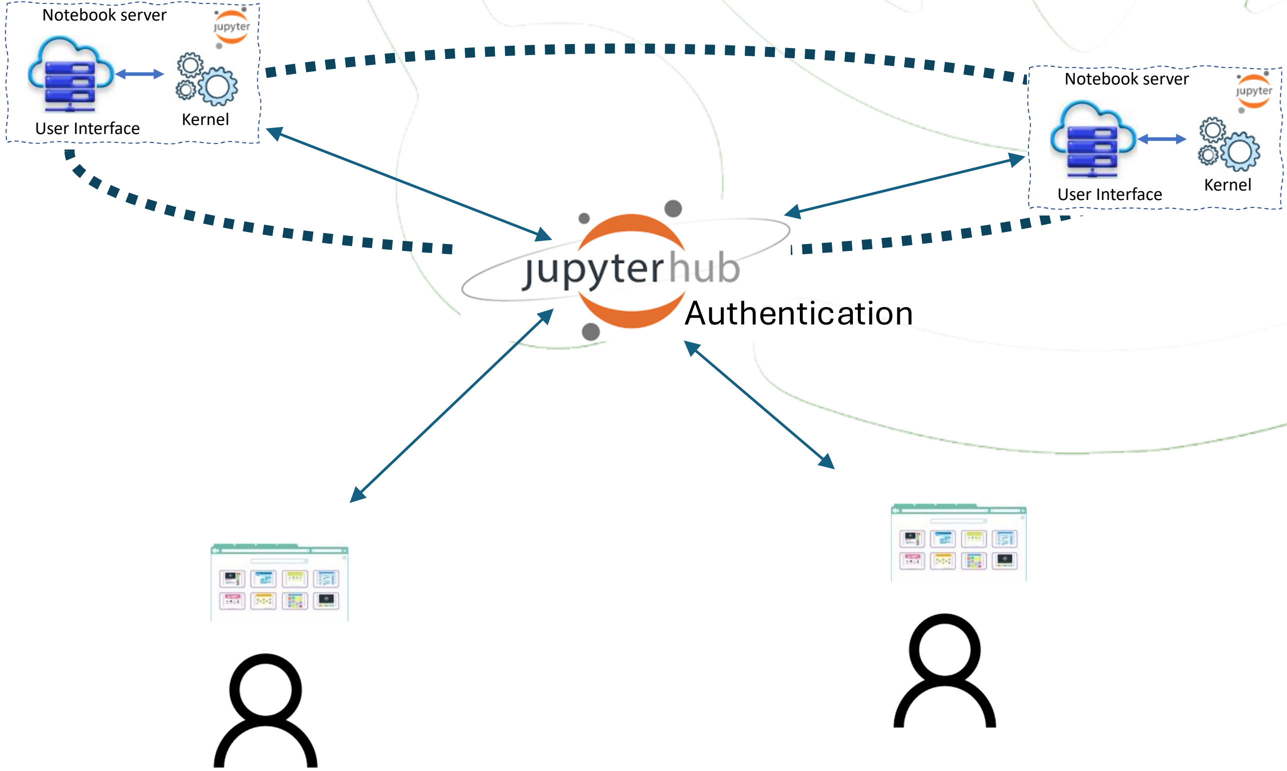
- Next generation web interface of Jupyter project
- Can open multi notebooks



```
zhiming@Zhimings-MacBook-Pro jupyter-lab % j
```

Jupyter hub

- Multiuser-shared Jupyter environment
- Each user gets an independent instance of Jupyter notebook/ Jupyter lab
- Share the pool of the computing resources
- Each instance uses the capacity of the machine where it is deployed



Jupyter in scientific applications

- Interactive programming for rapid prototyping
- Exploratory workflow for data and scientific experiments
- “Speak my language”, via more than 100 available Jupyter kernels
- Easy to share via version control system
- Self-contained notebooks include both narrative texts and code

https://www.nature.com/articles/d41586-018-07196-1

Limits of Jupyter

- Jupyter is *flexible* for developing scientific code, but also “*encourage*” poor coding practices;
- Jupyter allows a user to interact with the notebook in a “*non-linear*” fashion; it gives the user great power for *exploration*, but also require high responsibility for maintaining the *quality of the code;*
- A notebook can be shared among communities via version control systems or other repositories; however, the *reusability* of a notebook at the *cell* level is limited.

4. Why open science?

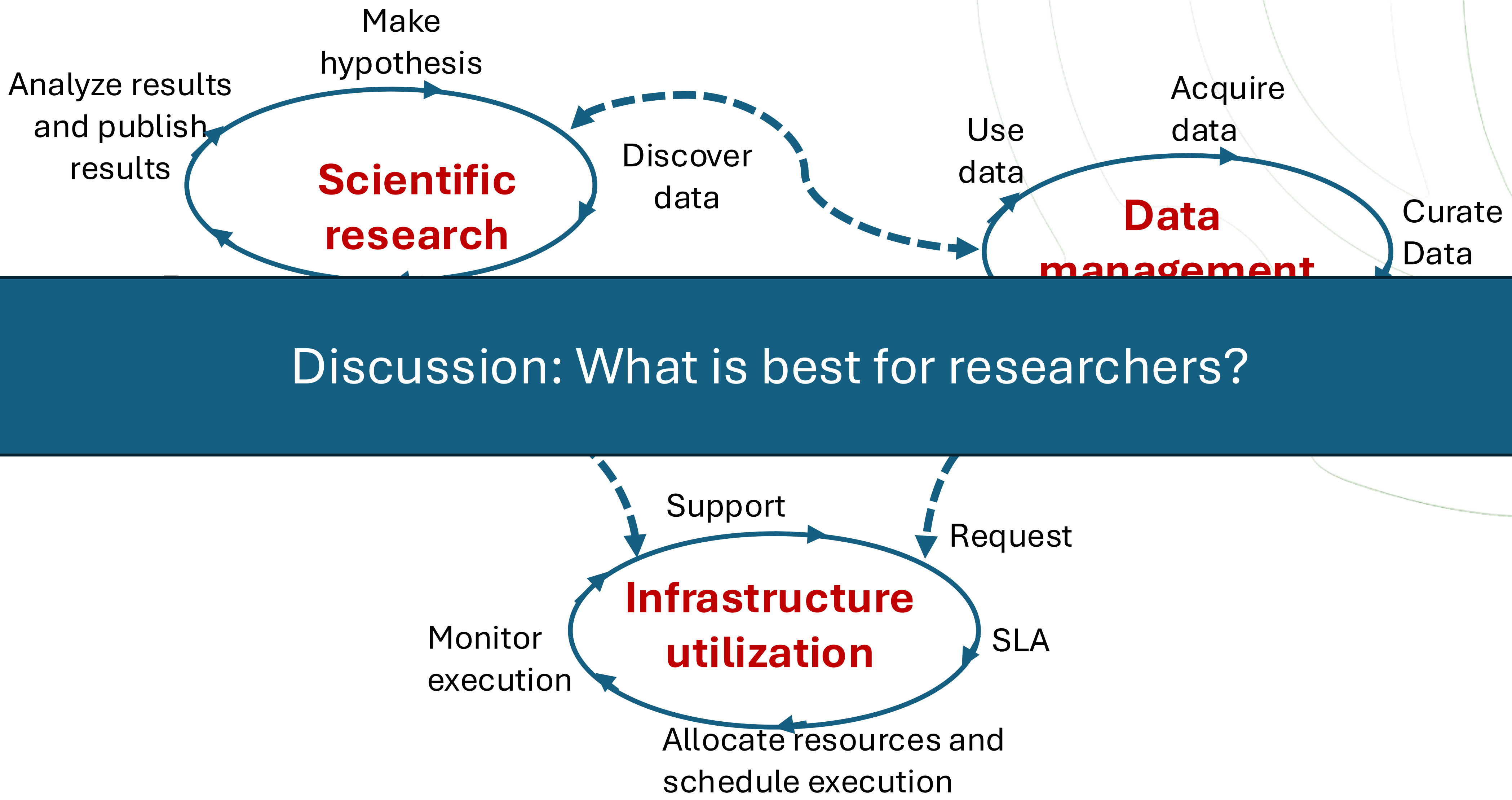
Discussion

- Think of a previous research project you attended, and ask yourself:
  - What activities did you conduct?
  - What data/computing tools have you used?
  - What software have you produced?
- What data and computing challenges have you experienced?

Outline

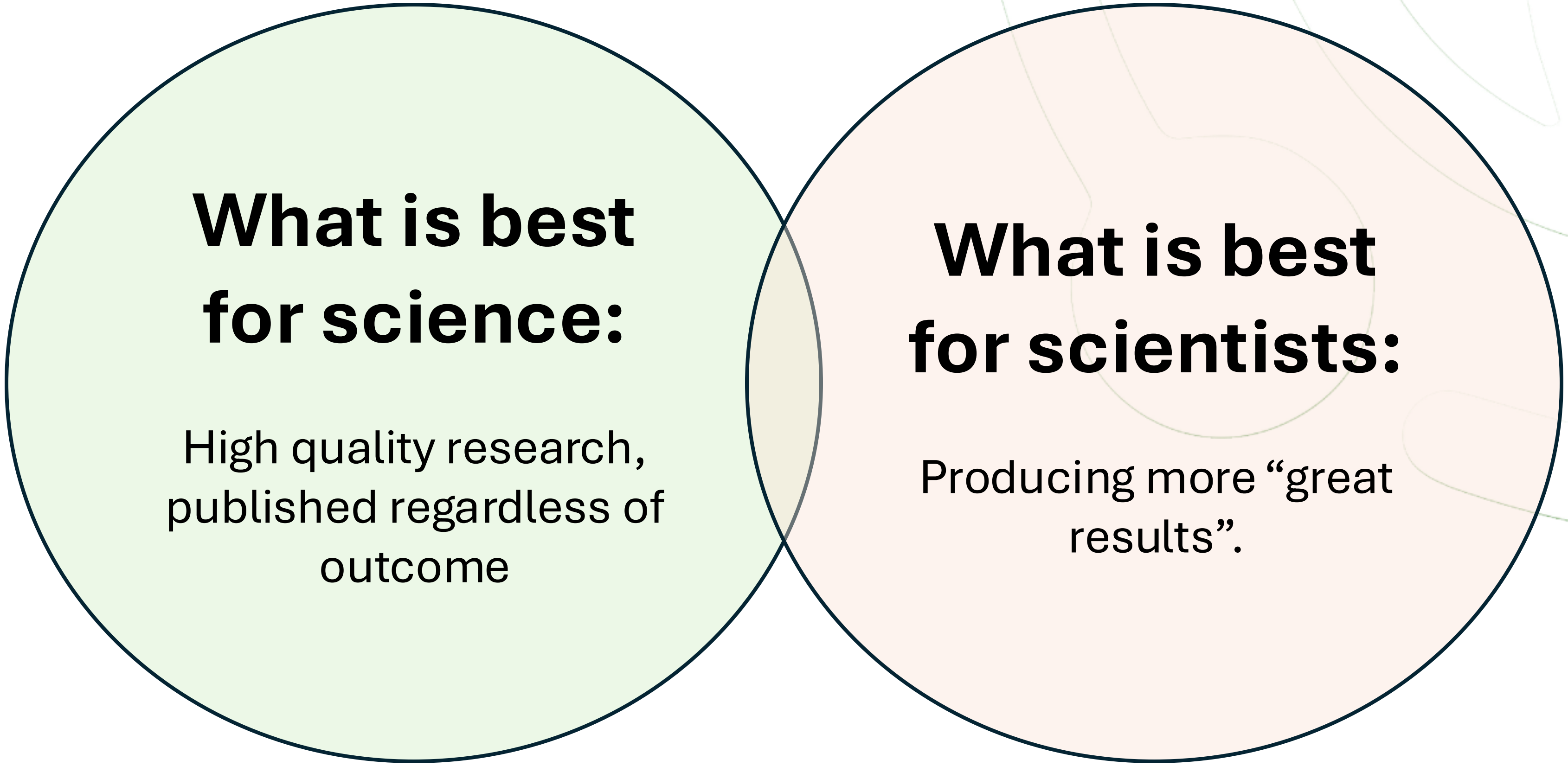
- Science paradigms
- Reproducibility crisis
- Open science
- Discussion

4th paradigm: data centric research activities



Discussion: What is best for researchers?

Quality and success in “Results-driven culture”



What is best for science:

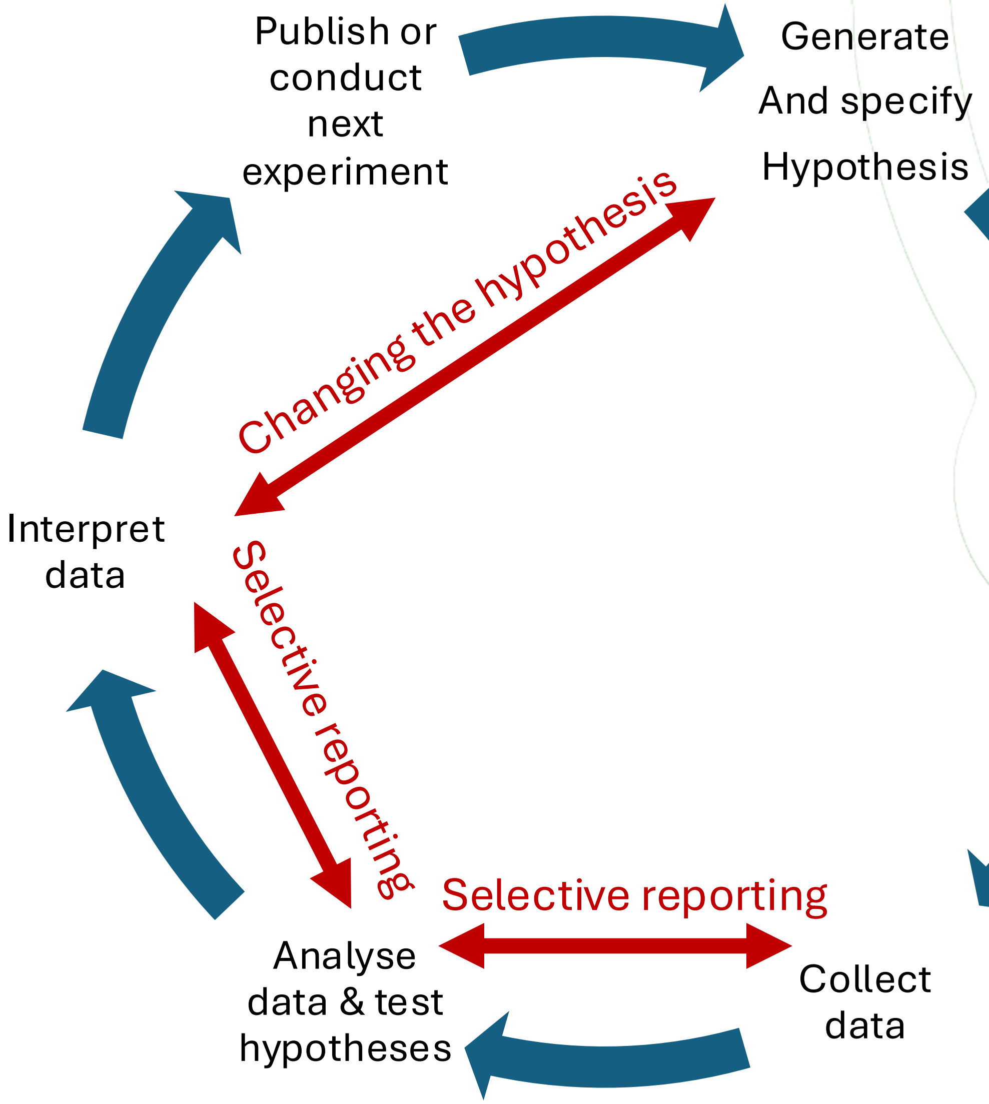
High quality research, published regardless of outcome

What is best for scientists:

Producing more “great results”.

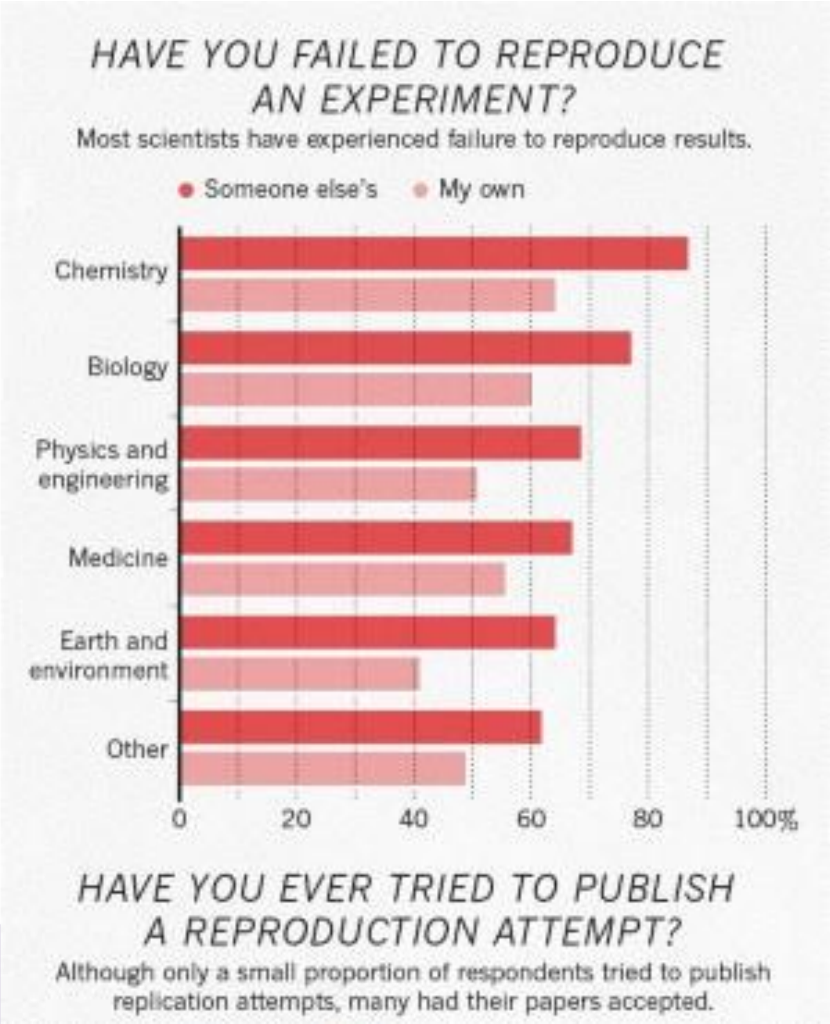
https://doi.org/10.1177/1745691612459058

When researchers under pressure to get "get results"



Reproducibility crisis?

- Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016). https://doi.org/10.1038/533452a

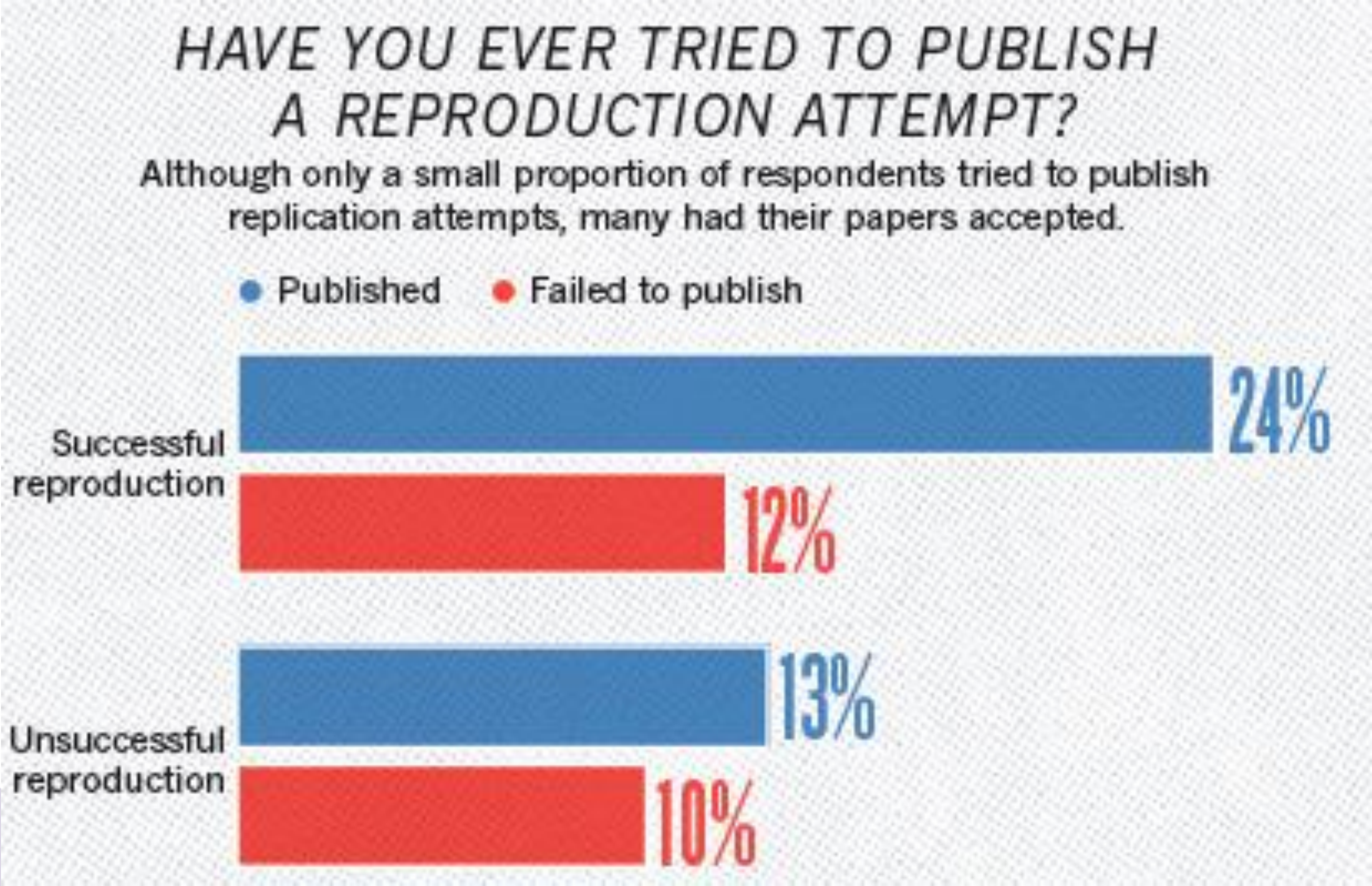


HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.

HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.

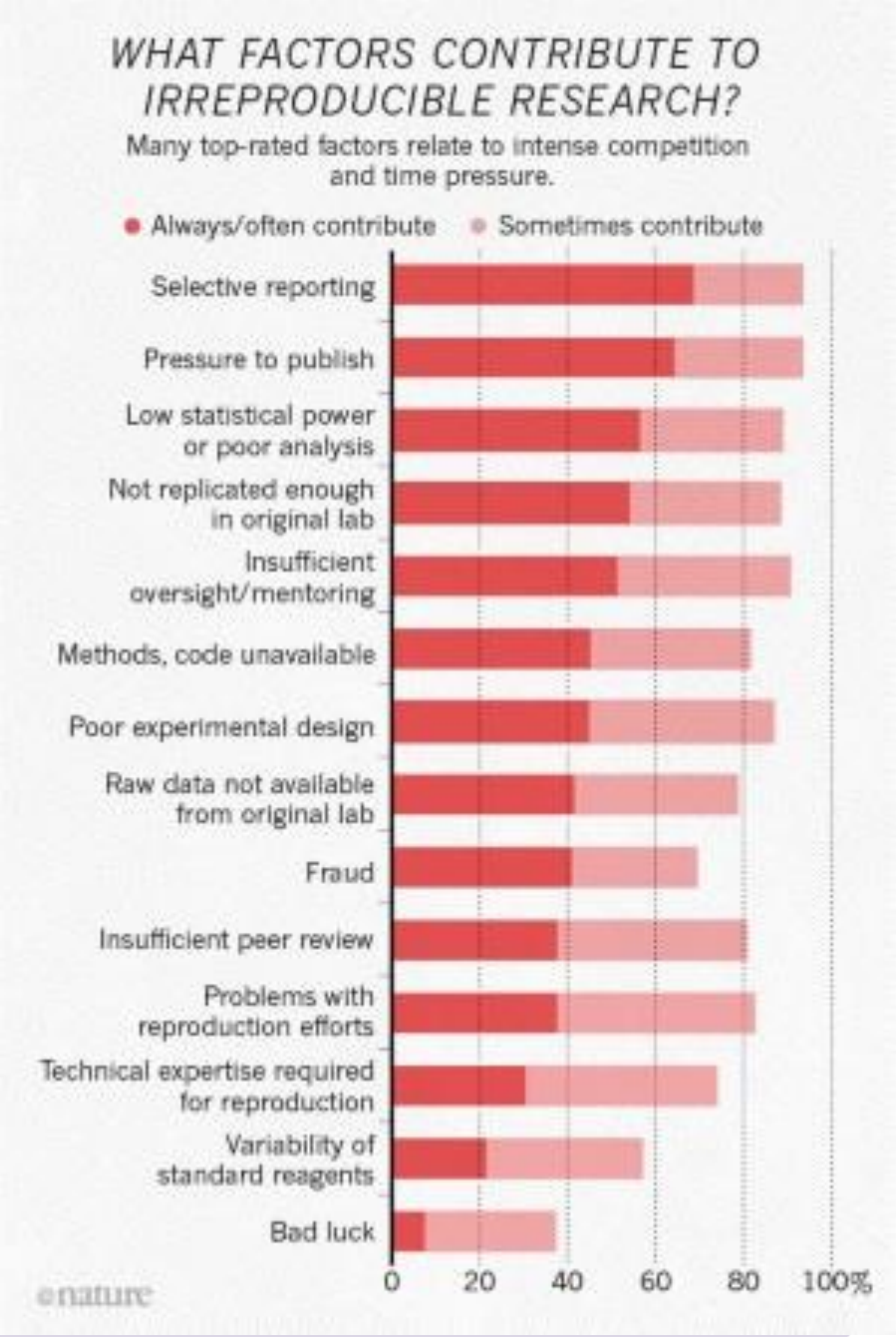


HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.

Factors

- Selective reporting
- Pressure to publish
- low statistical power or poor analysis
- Not replicated enough in the original lab
- Insufficient oversight/monitoring
- Poor experimental design
- Raw data not available
- Fraud
- Insufficient peer review



Open Science

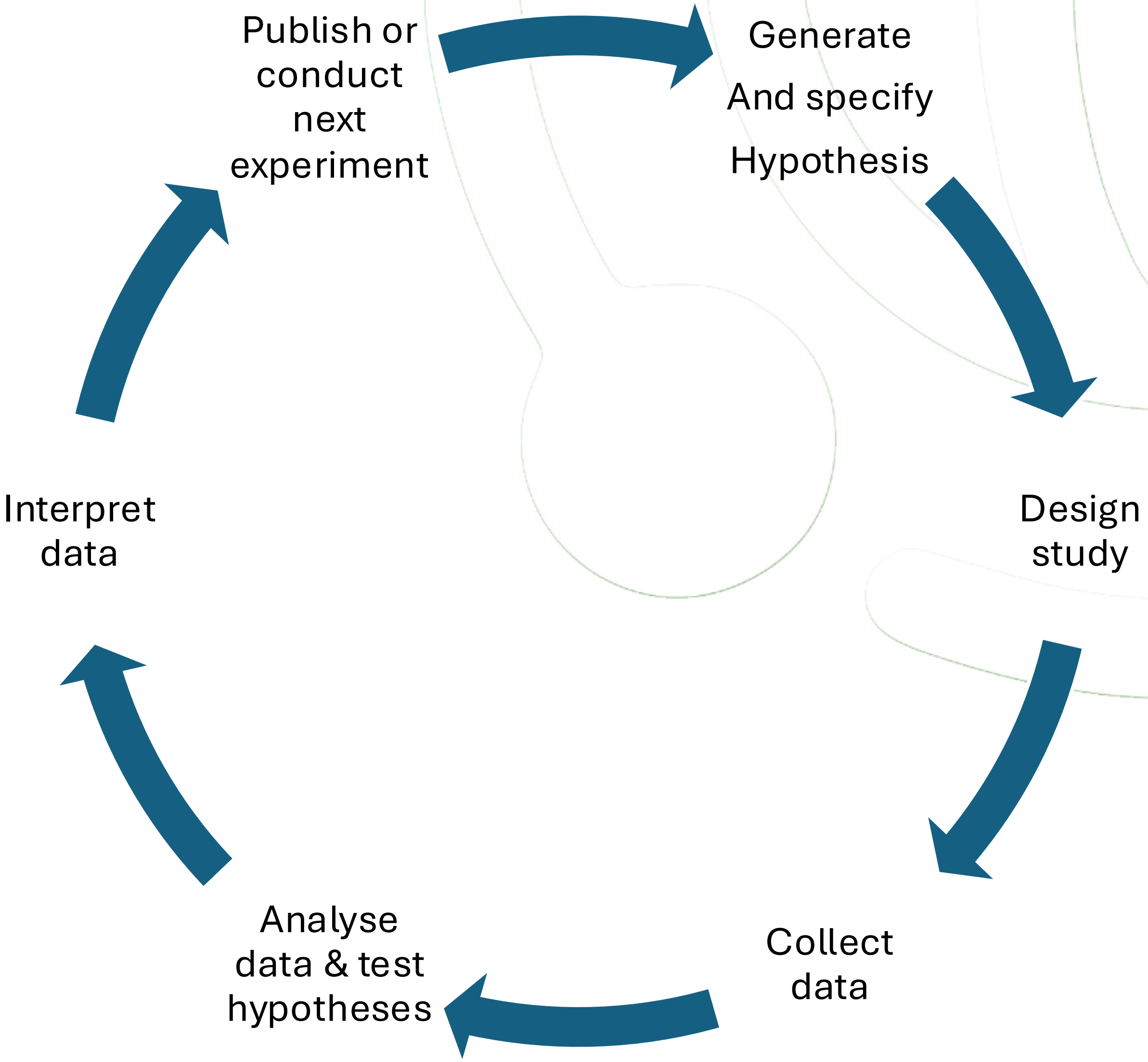
- Open science is a movement that aims for more open and collaborative research practices in which publications, data, software, and other types of academic output are shared as soon as possible and made available for reuse.
- Open Science is defined by UNESCO (2021) as an inclusive construct that combines various movements and practices aiming:
  - to make multilingual scientific knowledge openly available, accessible and reusable for everyone;
  - to increase scientific collaborations and sharing of information for the benefit of scienceand society;
  - and to open the processes of scientific knowledge creation, evaluation, and communication to societal actors beyond the traditional scientific community.

Why open science?

- Public funding supported research
- Reproducibility
- Collaboration
- …

Open Science

- Open data
- Open source
- Open education resources
- FAIR (Findable, Accessible, Interoperable, and Reusable)
- Open access,
- Open review
- Scientific social networks
- Transparent
- Reproducible
- …


Publish or conduct next experiment
Generate And specify Hypothesis
Design study
Collect data
Analyse data & test hypotheses
Interpret data

Discussion

- What are your open science practices?
- What open science challenges did you experience?

ITINERIS

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani