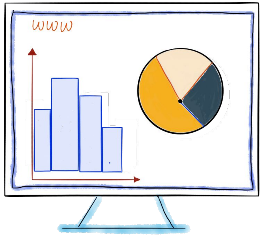
Reproducible Science in practice

tools and ideas



Arturo Sánchez Pineda (LAPP)

June 10, 2021 - ESCAPE (online) School

Arturo Sánchez Pineda

arturos@cern.ch

Linkedin /arturo-sanchez-pineda/

@artfisica

https://twitter.com/Arturo_RSP

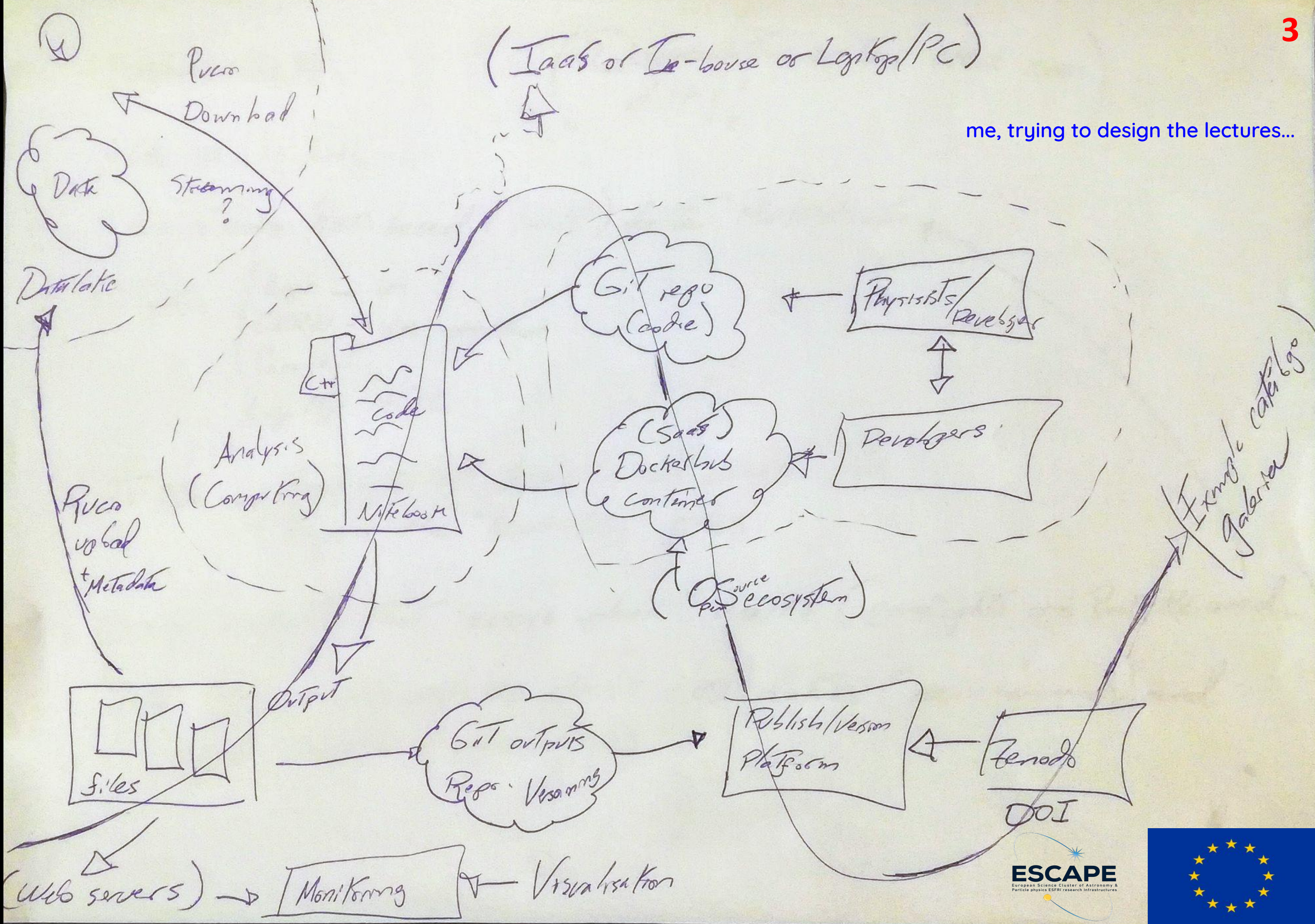
I am post-doctoral fellow at LAPP-CNRS, France. Member of the ESCAPE and ATLAS groups.

I studied Fundamental Physics and System Engineering in the Universidad de Los Andes, Venezuela, with a PhD in Fundamental and Applied Physics from Università di Napoli "Federico II", Italy.

I was previously a postdoctoral fellow at Physics Department at Università di Udine and an Associate at INFN, Italy. Also, an ATLAS TDAQ System Administrator at CERN, Switzerland, and Research Associate at the High Energy, Cosmology and Astroparticle Section at ICTP, Italy.

← **And I do a lot of outreach :)**

Quick reminder: please, don't forget the power of **pen & paper**



me, trying to design the lectures...

In my view, **reproducibility** refers to a series of principles, techniques, tools and practical considerations that allow the documentation, recording and preservation of data analysis pipelines — enhancing the possibilities of collaborations across borders and increasing the probabilities of replicating results by others (and yourself) in the future.

Reproducibility involves using standard and well-established protocols to ensure that your code will survive outside your computer, the passing of time and that others will be able to use it as a starting point for new analysis.

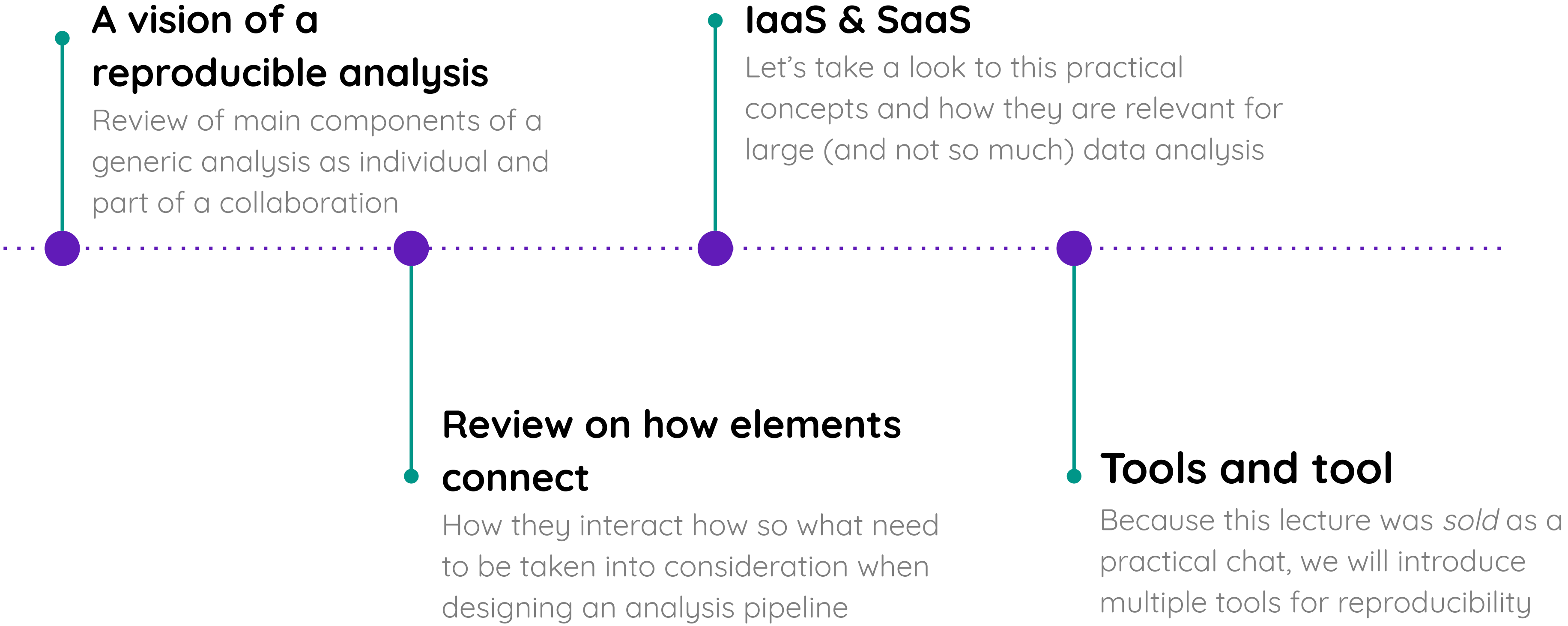
Another important observation

Generally, there is more than one way to perform an operation, create an object, deal with an issue, solve a problem,...

So, keep that in mind while in this school and during all your professional development :)

Overview of the **lectures**

Main ideas



A vision of a reproducible analysis

Review of main components of a generic analysis as individual and part of a collaboration

Review on how elements connect

How they interact how so what need to be taken into consideration when designing an analysis pipeline

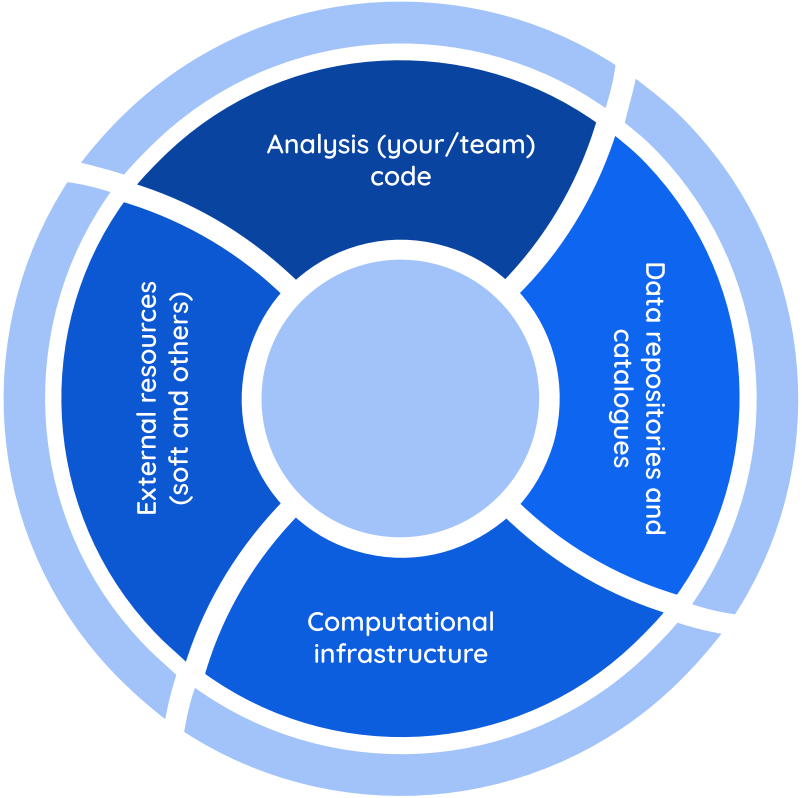
IaaS & SaaS

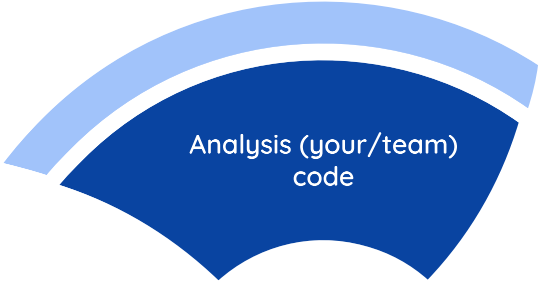
Let’s take a look to this practical concepts and how they are relevant for large (and not so much) data analysis

Tools and tool

Because this lecture was *sold* as a practical chat, we will introduce multiple tools for reproducibility

A vision of a reproducible **analysis**


Analysis (your/team) code
Data repositories and catalogues
Computational infrastructure
External resources (soft and others)



Analysis Code

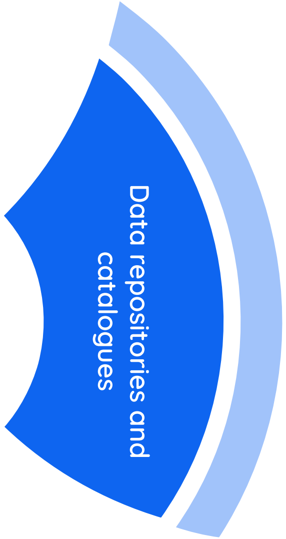
Let’s define this one as the code that you write. And the collaboration code too (in the case of an institutional or multinational experiment, for example)

The set of dedicated/custom software, macros and others that as an analyser you co-develop to interact with the data

Data repositories and catalogues

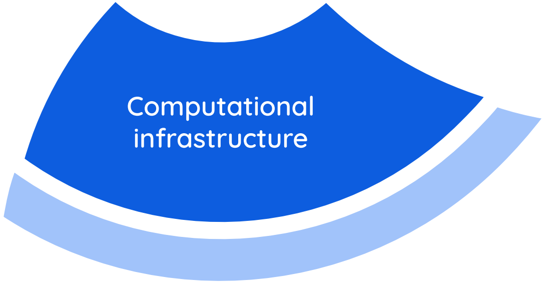
This refers to the storage of the data and the metadata associated Sometimes such pieces are hosted together. In other cases, they are separated (but connected) in different repositories and even facilities

This is the data that your analysis code reads and use to produce a meaningful result, generate new information and conclusions



Computational infrastructure

It refers to the computer power that is used for your code in order to explore and analyse such data
It can be your laptop, but also the remote computational infrastructure that can be used on-demand thanks to being part of an institution or collaboration. Also, when you pay for it (e.g. commercial clouds)

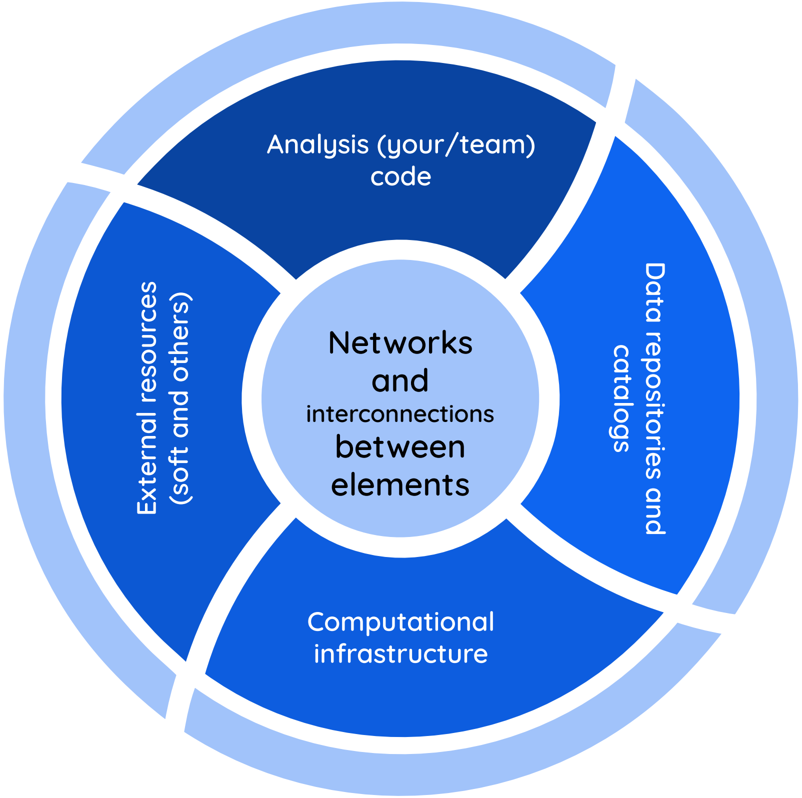


External resources (soft and others)

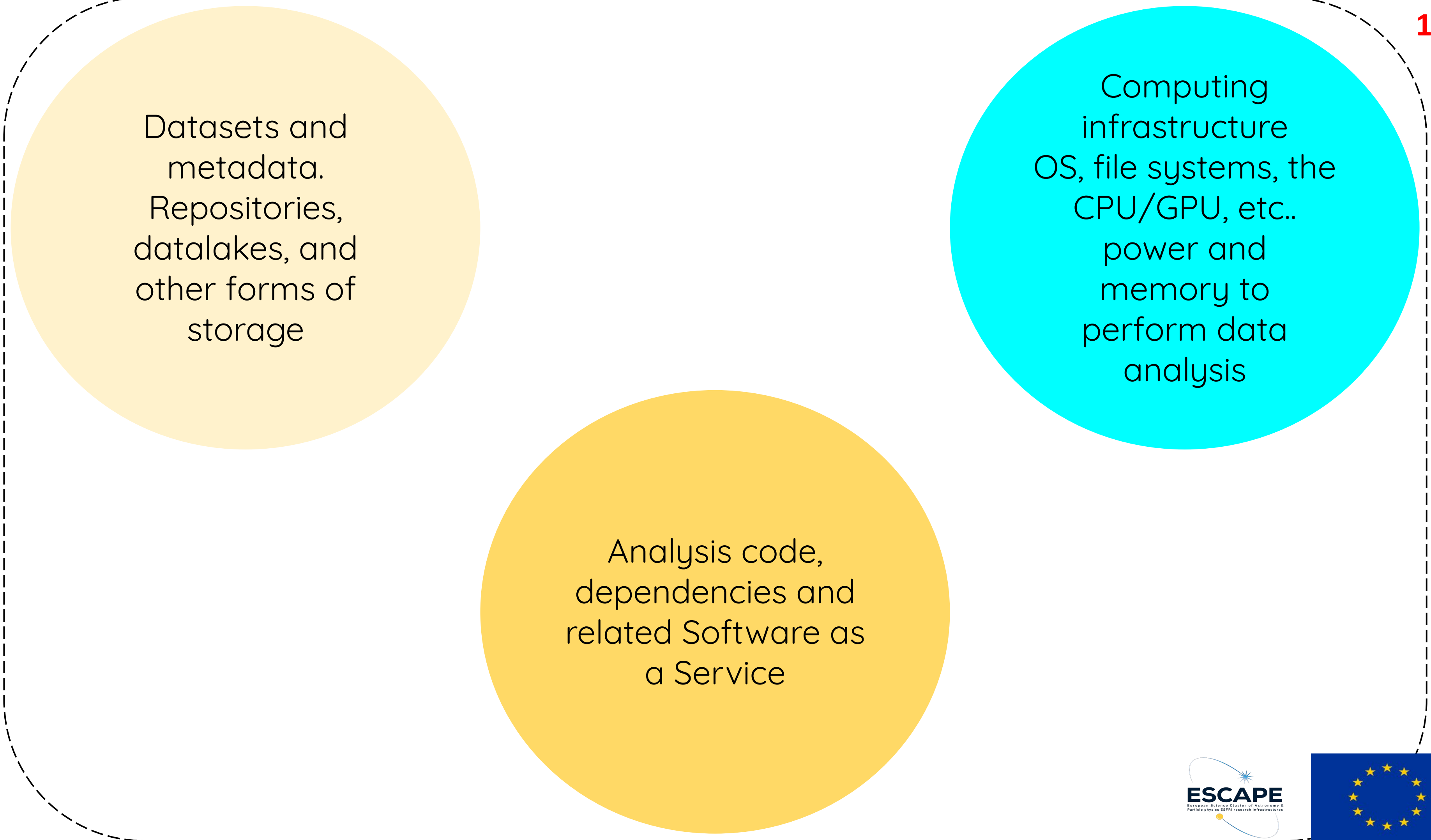
External resources (soft and others)

I can encapsulate those as the software that you import into your code. All the different libraries and external tools that allow your code to perform the needed operations and do the corresponding versioning (i.e. Git)

Also can refer to the platforms to develop soft, like the already mention Jupyter notebooks, online editors, among others


Analysis (your/team) code
Data repositories and catalogs
Computational infrastructure
External resources (soft and others)
Networks and interconnections between elements

Looking for the **elements**



Datasets and metadata. Repositories, datalakes, and other forms of storage

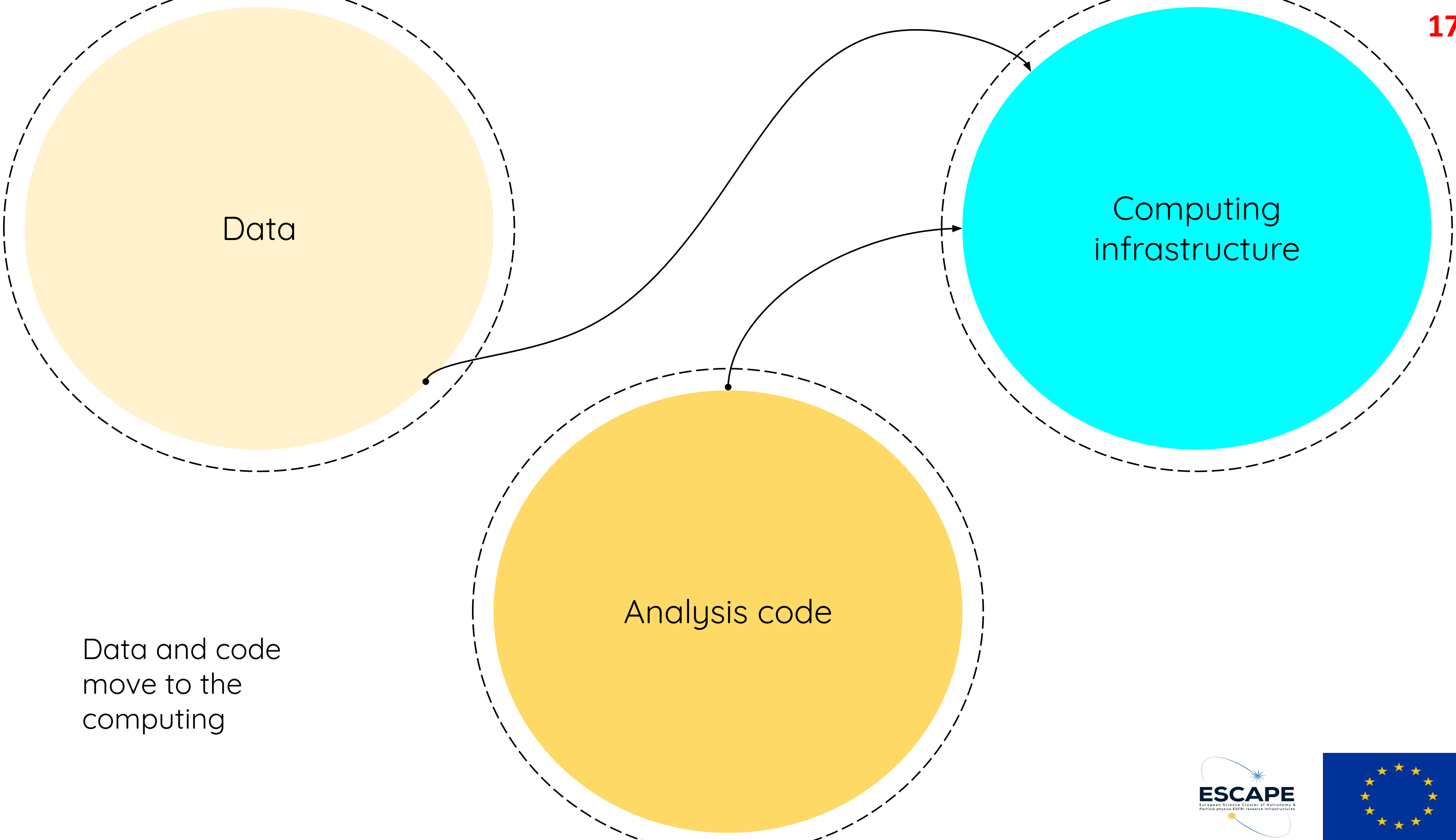
Computing infrastructure OS, file systems, the CPU/GPU, etc.. power and memory to perform data analysis

Analysis code, dependencies and related Software as a Service

Datasets and metadata. Repositories, datalakes, and other forms of storage

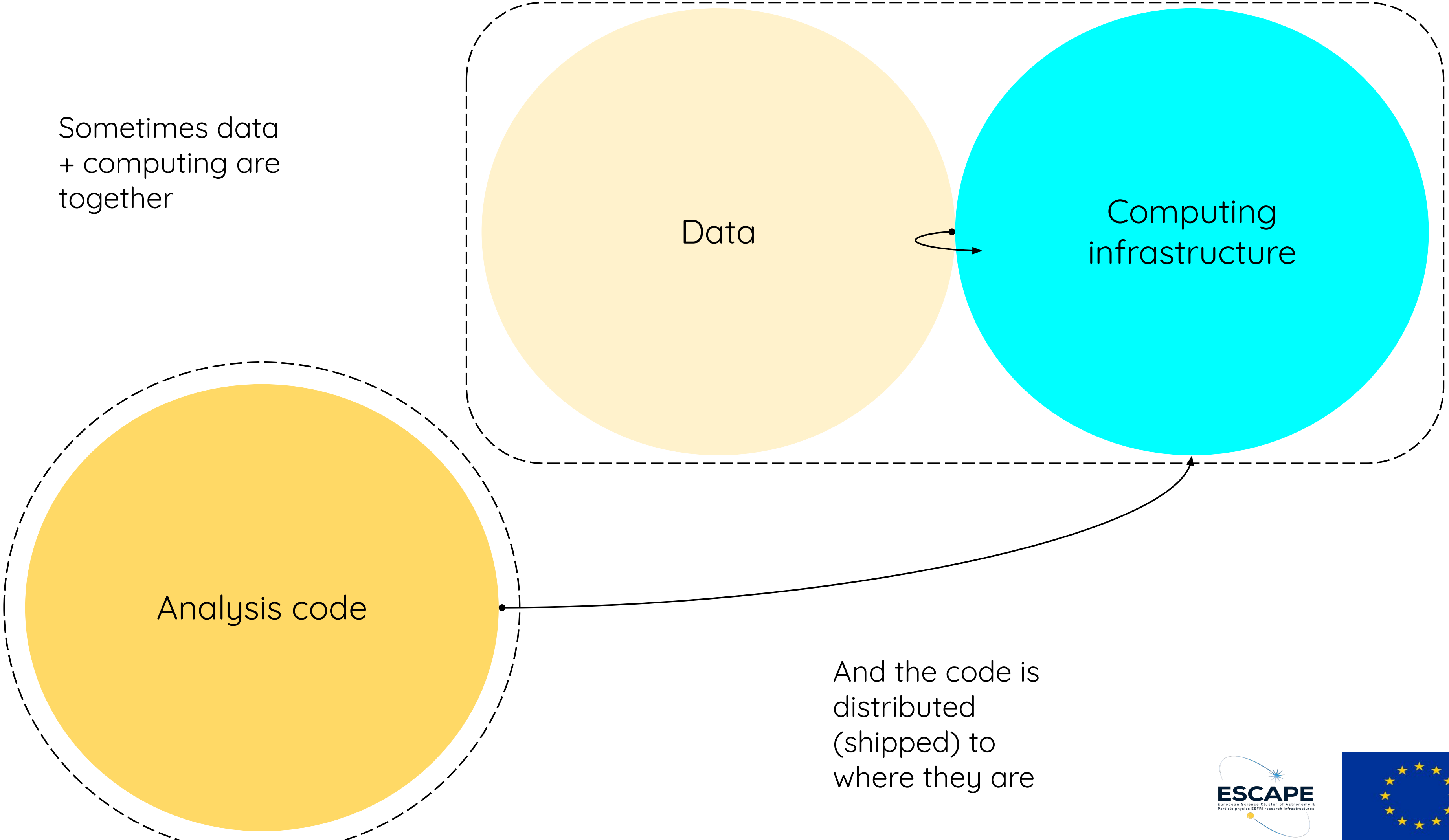
Computing infrastructure
OS, file systems, the CPU/GPU, etc.. power and memory to perform data analysis

Analysis code, dependencies and related Software as a Service

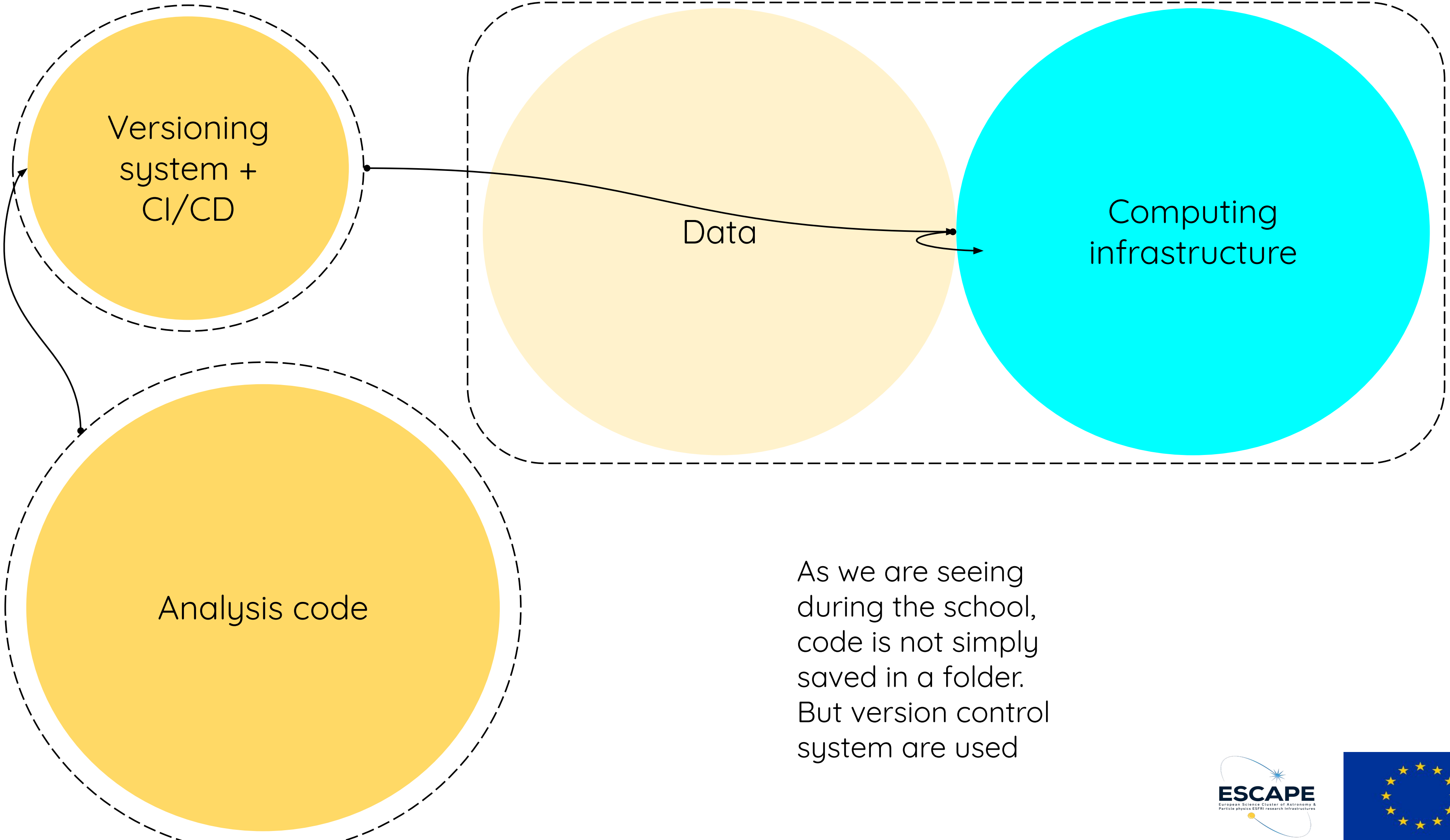


Data and code move to the computing

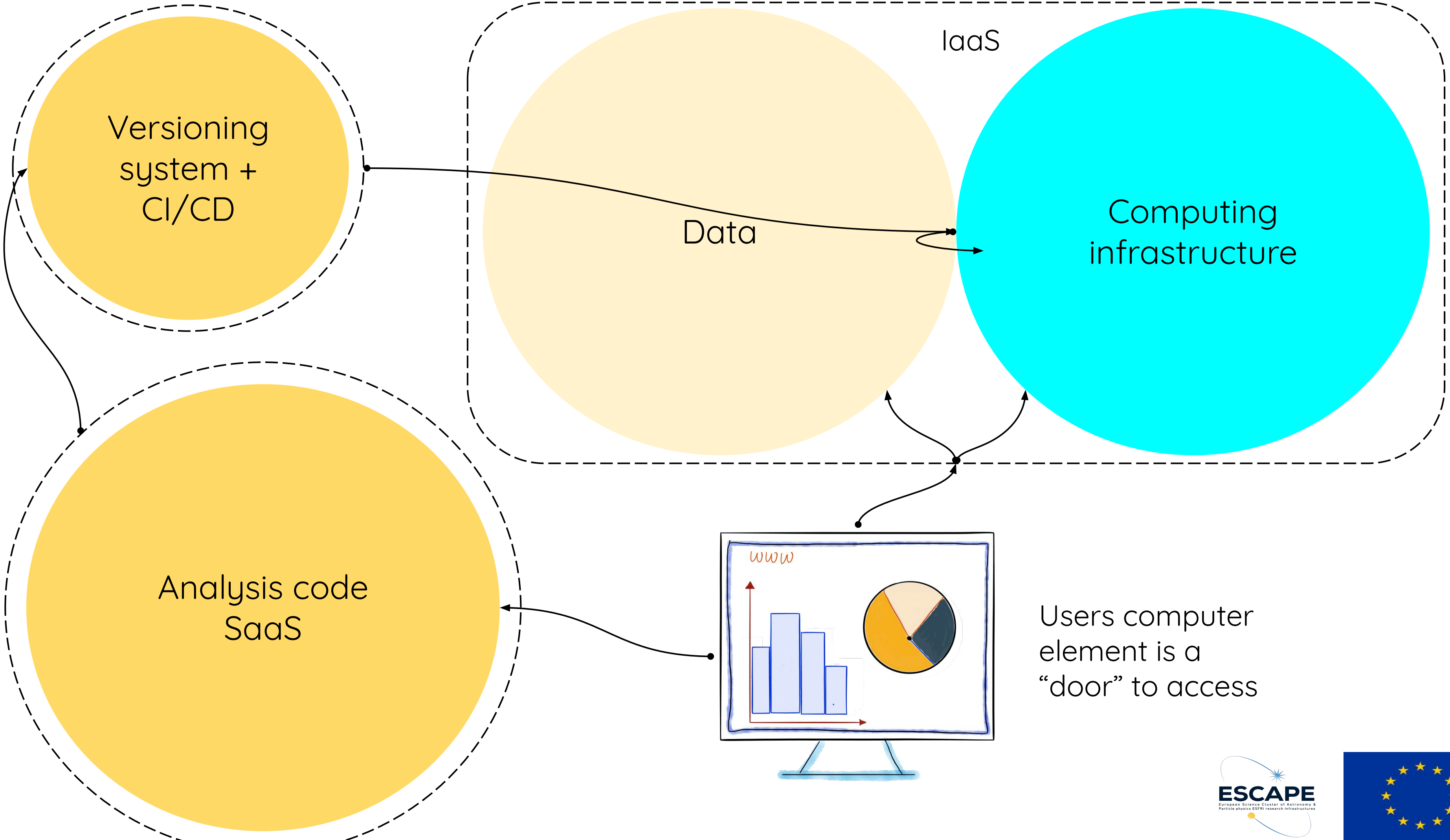
Sometimes data + computing are together



And the code is distributed (shipped) to where they are



As we are seeing during the school, code is not simply saved in a folder. But version control system are used



IaaS

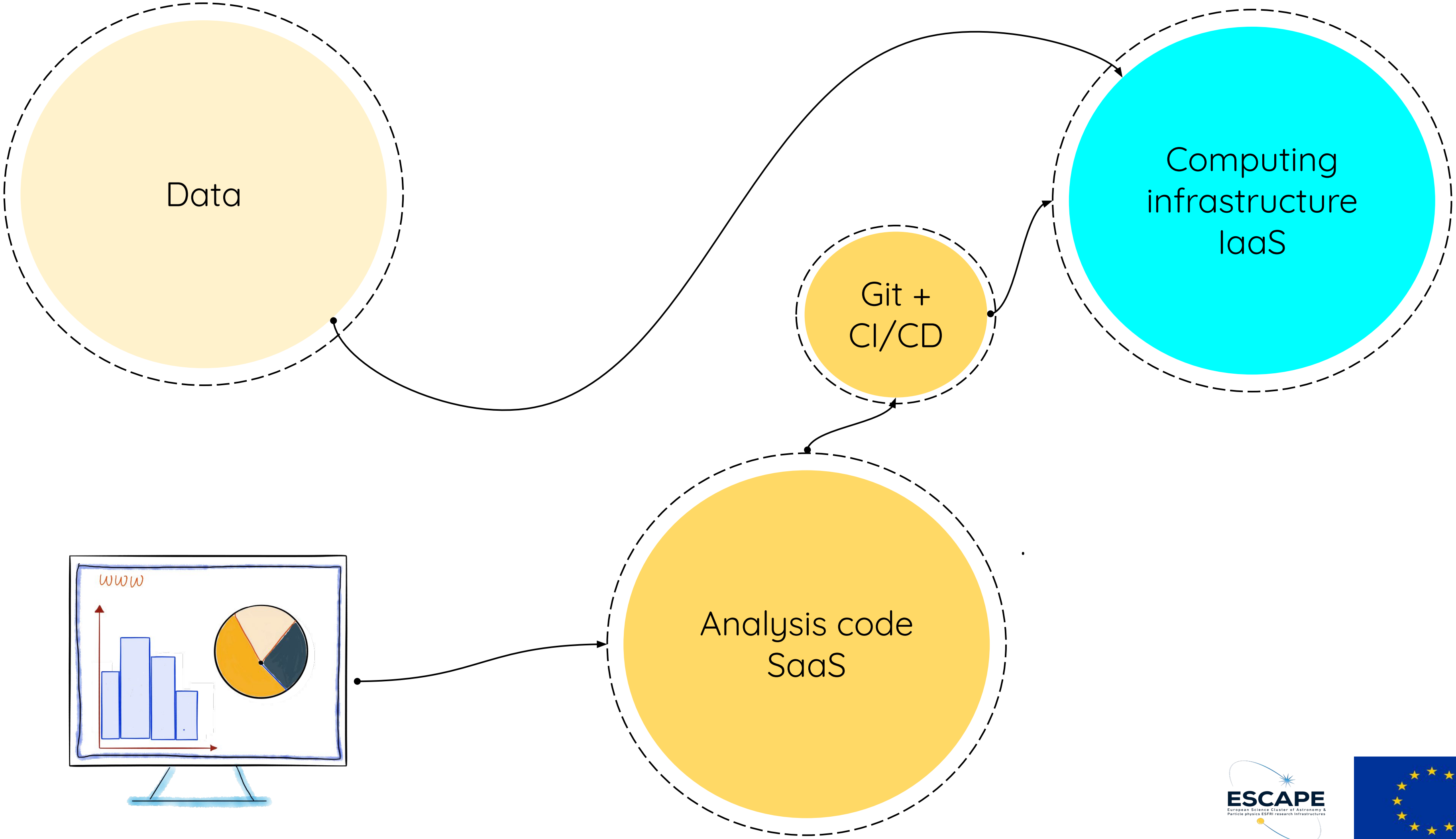
Versioning system + CI/CD

Data

Computing infrastructure

Analysis code SaaS

Users computer element is a “door” to access



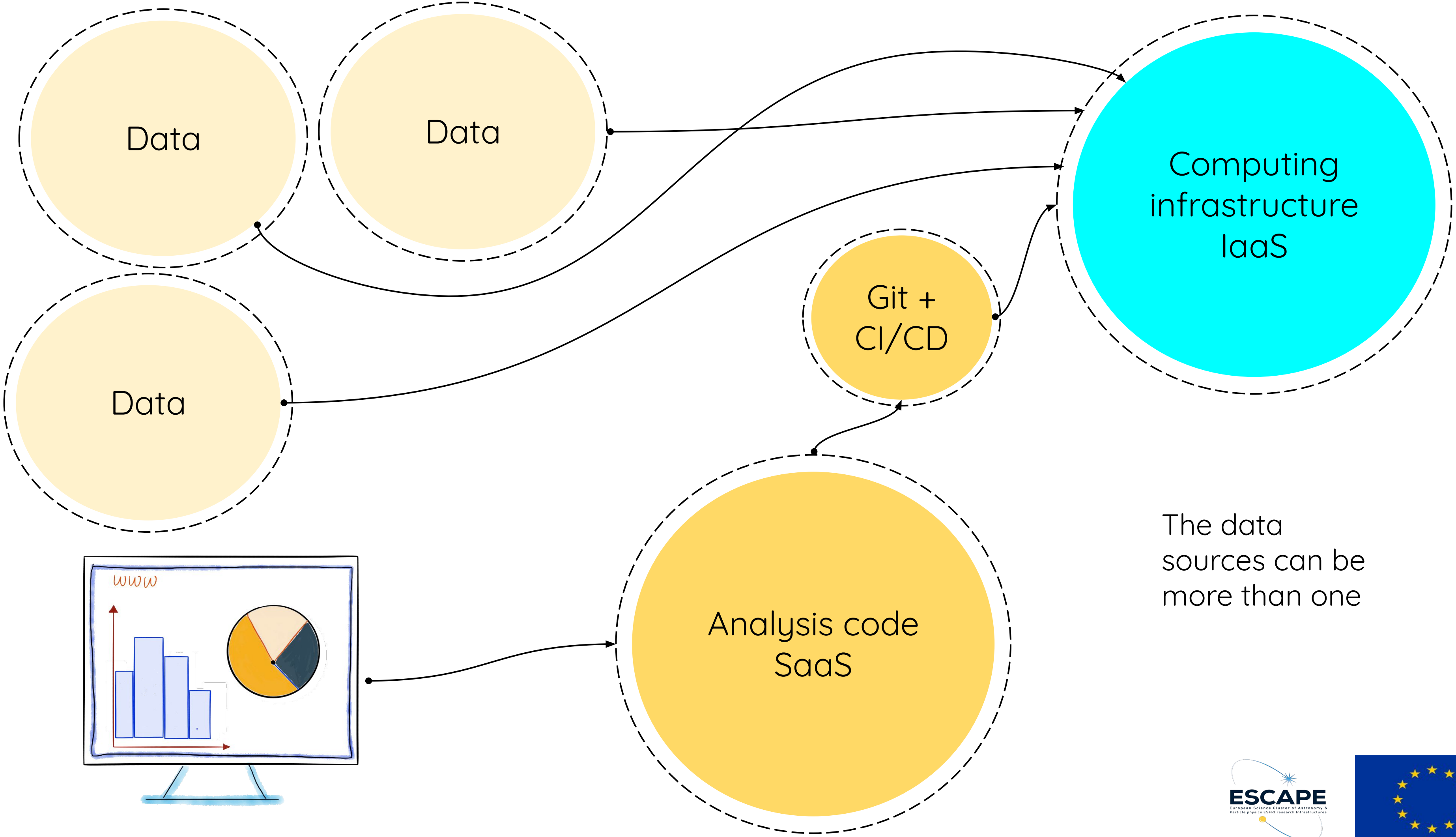
Data

Computing infrastructure IaaS

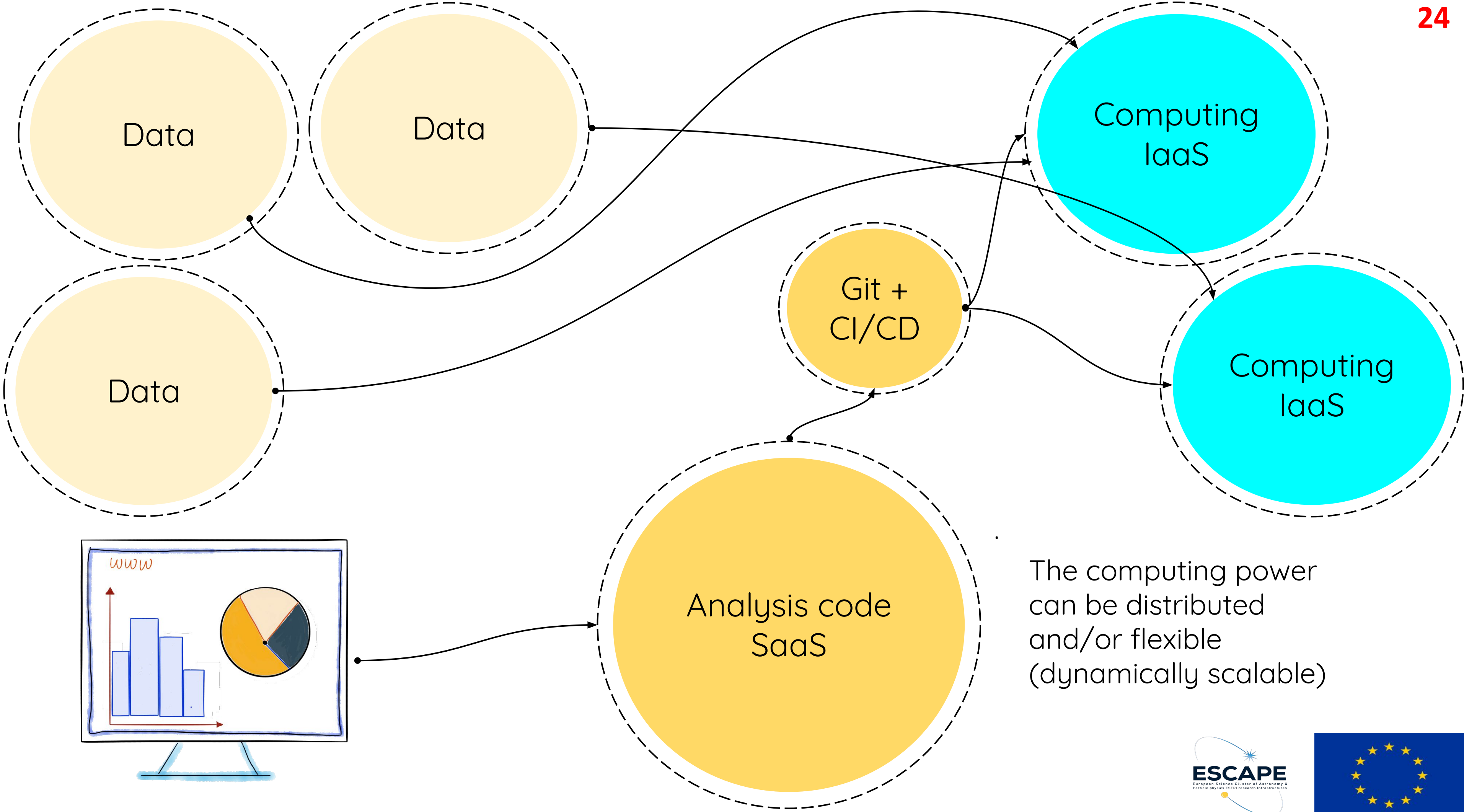
Git + CI/CD

Analysis code SaaS

Elements and **examples**



The data sources can be more than one



Data

Data

Data

Computing IaaS

Computing IaaS

Git + CI/CD

Analysis code SaaS

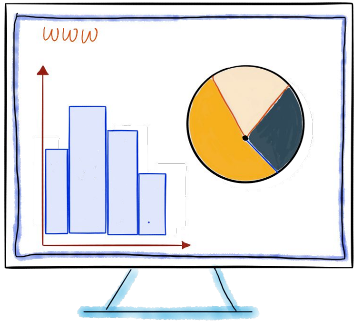
The computing power can be distributed and/or flexible (dynamically scalable)

Data

Analysis code
SaaS

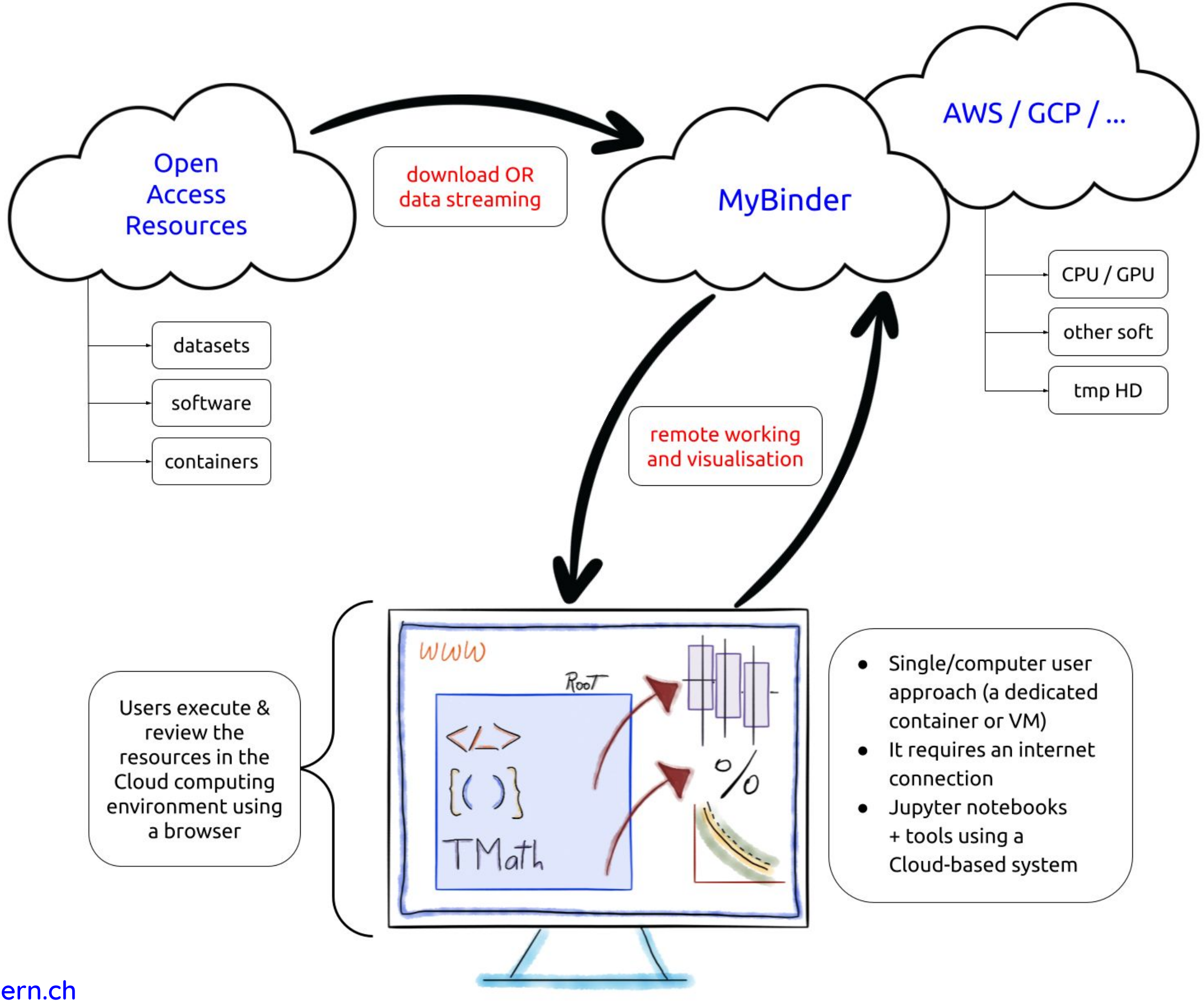
Git +
CI/CD

Computing
IaaS



Of course, we can identify more pieces and also increase the granularity, but let's evaluate this group

Software & Infrastructure as a **Service**



Open Access Resources

- datasets
- software
- containers

download OR data streaming

MyBinder

AWS / GCP / ...

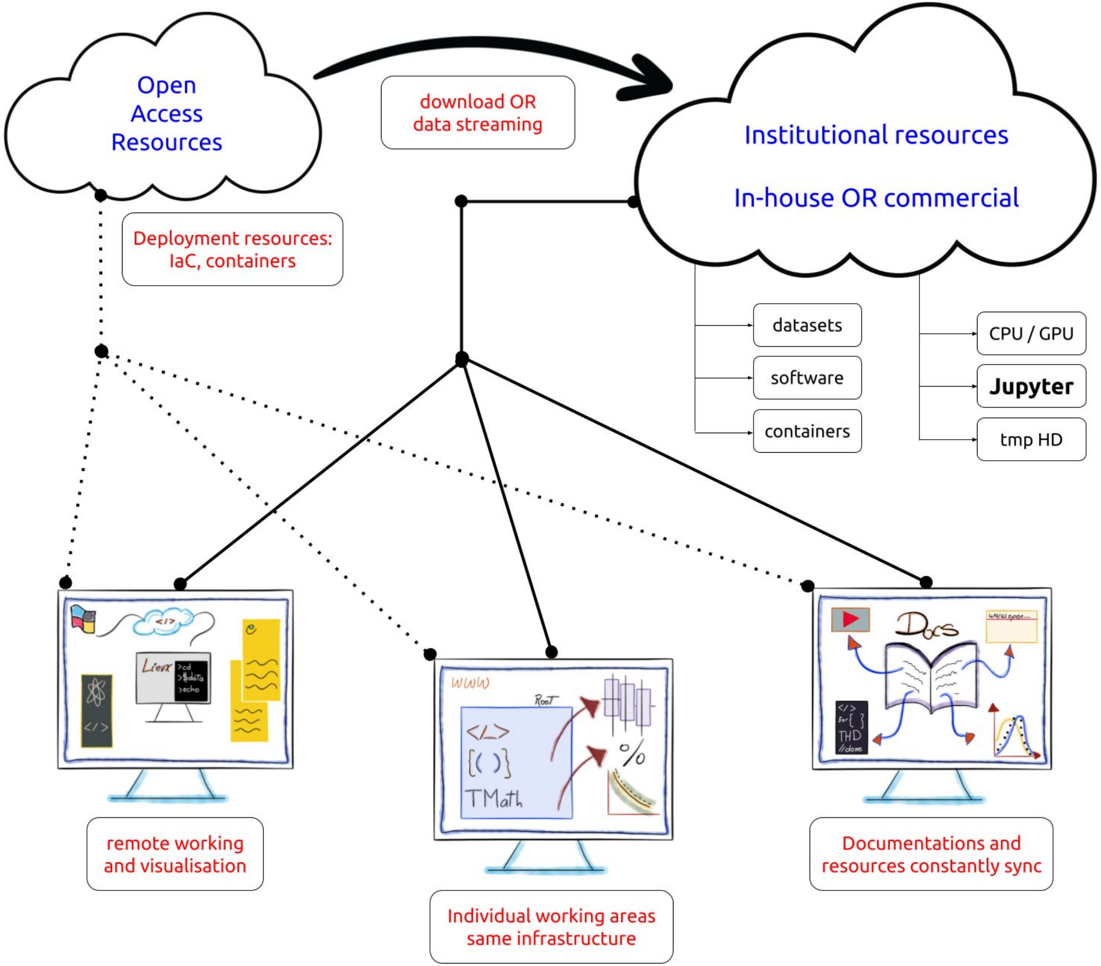
- CPU / GPU
- other soft
- tmp HD

remote working and visualisation

Users execute & review the resources in the Cloud computing environment using a browser

- Single/computer user approach (a dedicated container or VM)
- It requires an internet connection
- Jupyter notebooks + tools using a Cloud-based system

A pictorial description of the usage of resources in an offline environment. Tools like the VM and analysis frameworks provide a self-contained setup.



Open Access Resources

download OR data streaming

Institutional resources

In-house OR commercial

Deployment resources: IaC, containers

datasets

software

containers

CPU / GPU

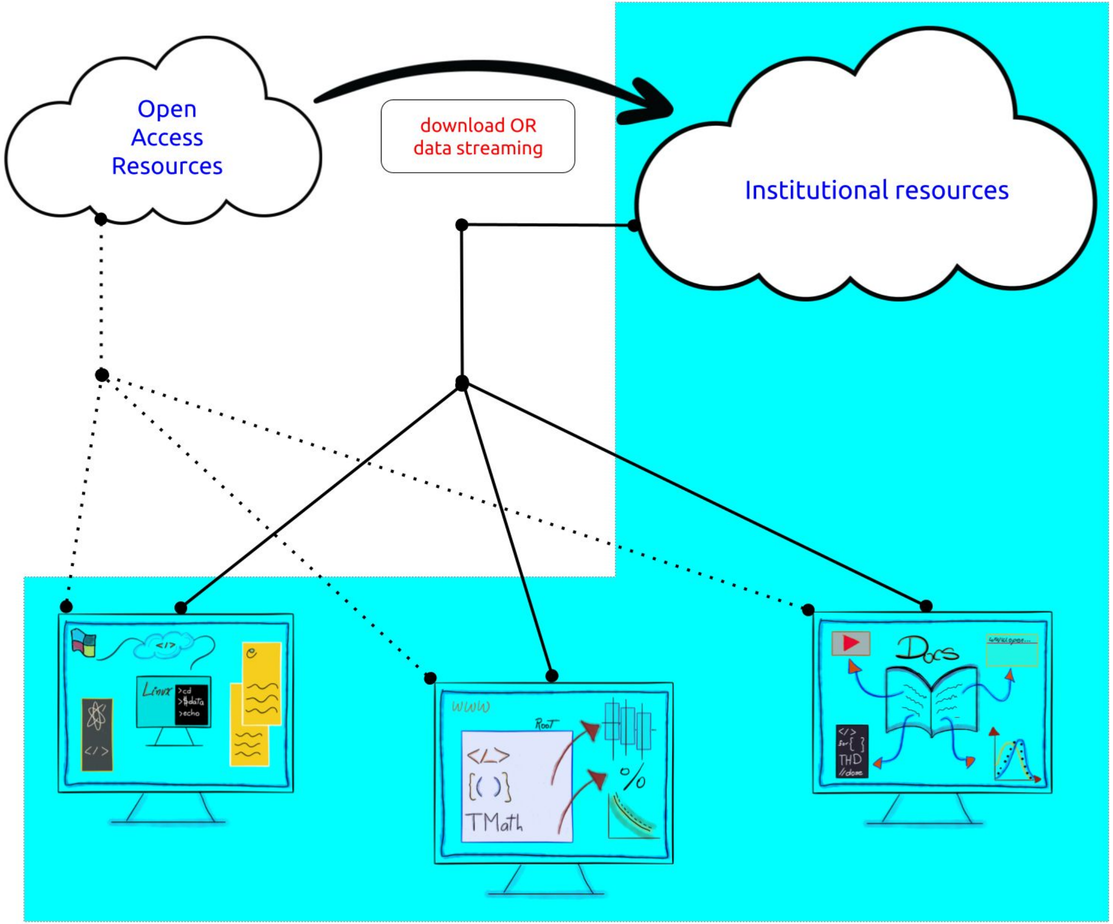
Jupyter

tmp HD

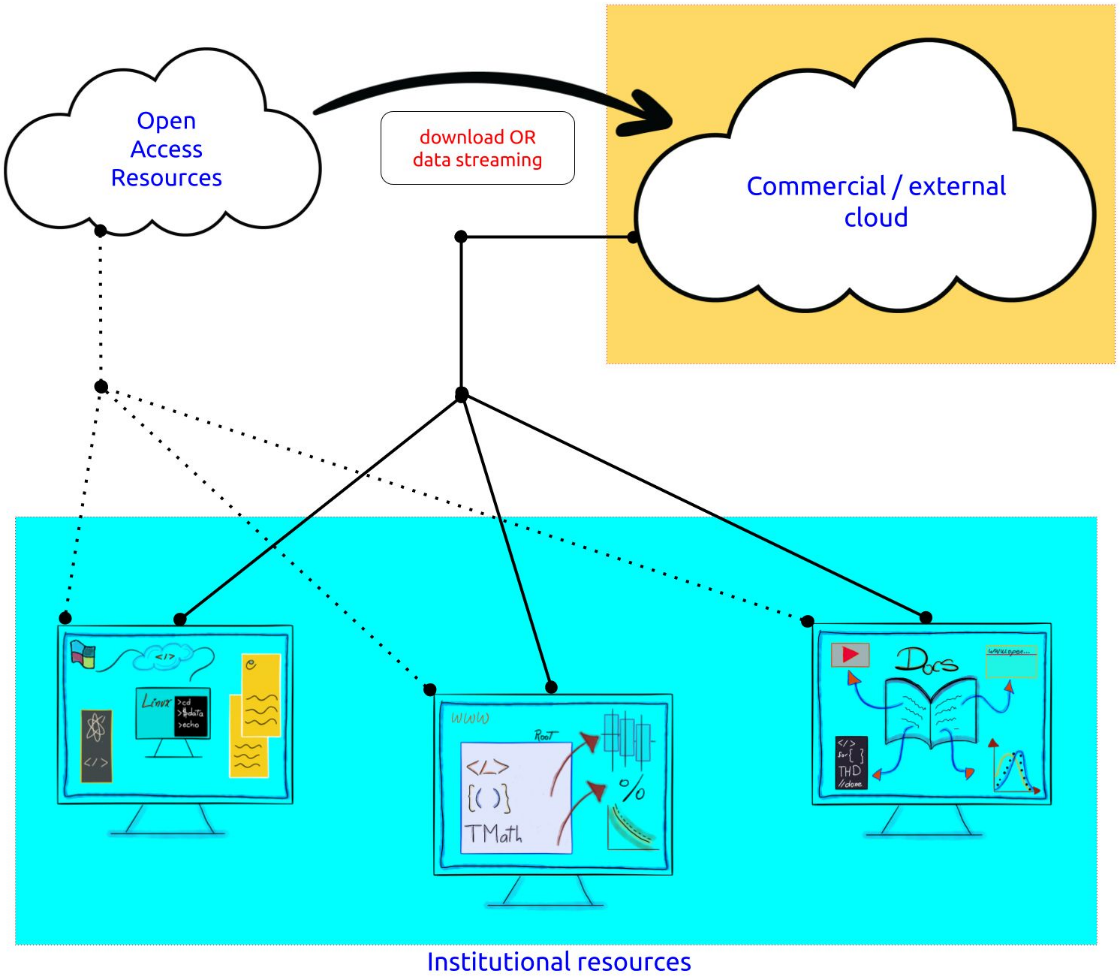
remote working and visualisation

Individual working areas same infrastructure

Documentations and resources constantly sync



Institutional resources

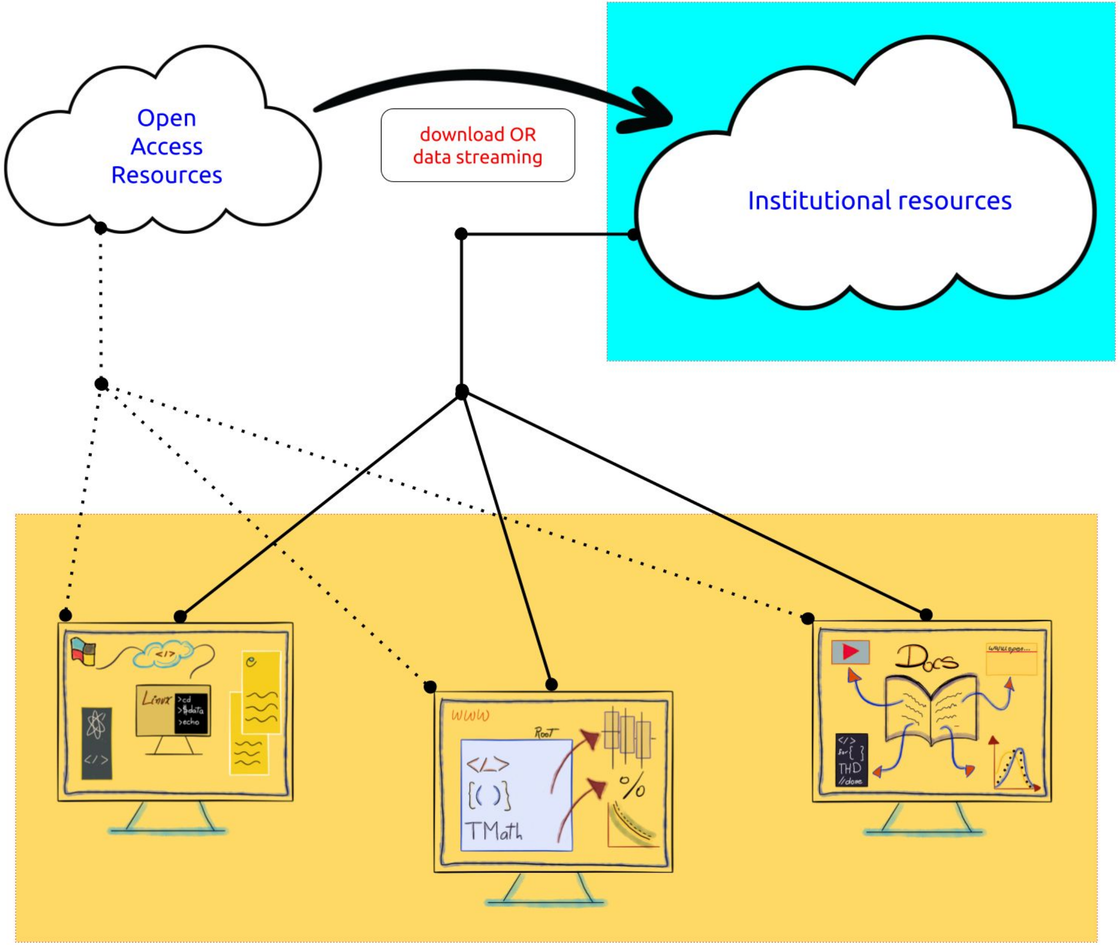


Open Access Resources

download OR data streaming

Commercial / external cloud

Institutional resources



Open Access Resources

download OR data streaming

Institutional resources

External users / public