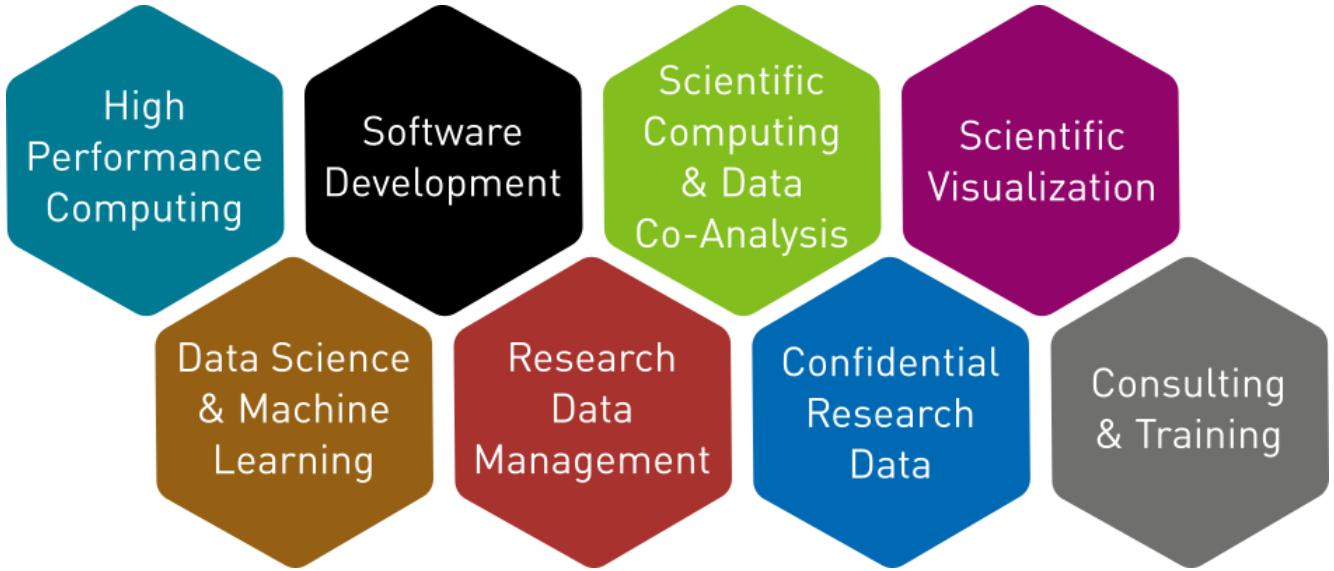
ETH zürich

IT Services

Introduction to Active Research Data Management

Caterina Barillari, Andrei Plamada
Scientific IT Services, ETH Zurich

Who is Scientific IT Services?



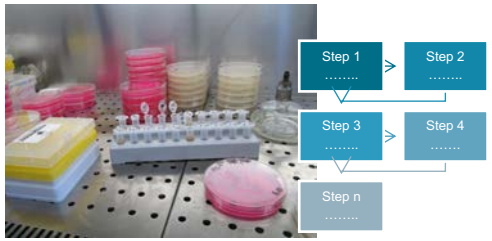
- A section of ETHZ IT Services
- About 40 experts in various areas of scientific computing
- With a background in different areas of science

Tell us a bit about yourselves

- Your affiliation
- Your research topic

Overview of today's workshop

ARDM intro



Sample & protocol management

Management of Data & Metadata

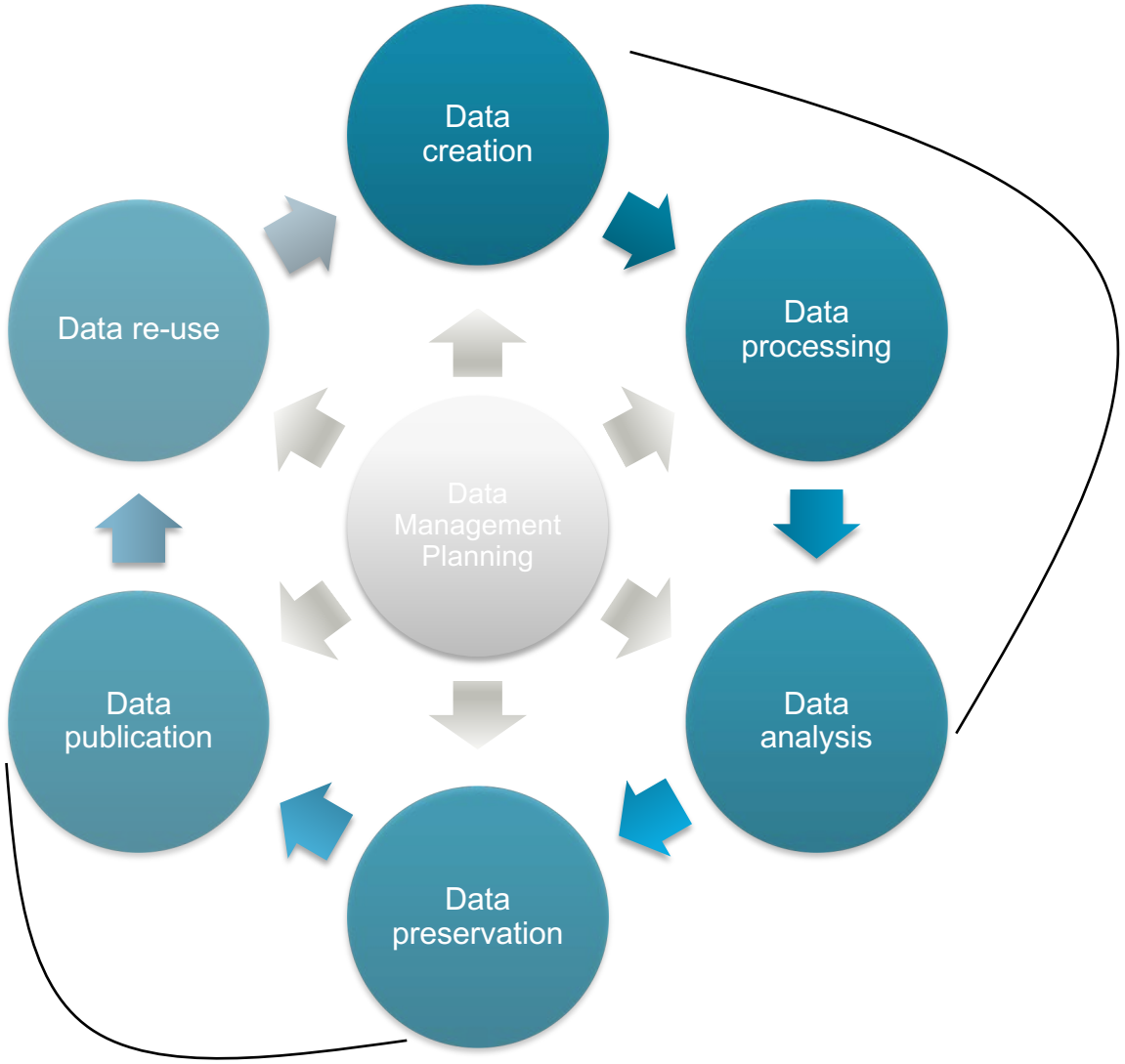
ELNs

SIS's RDM solution

Reproducible data analysis

Overview of active research data management

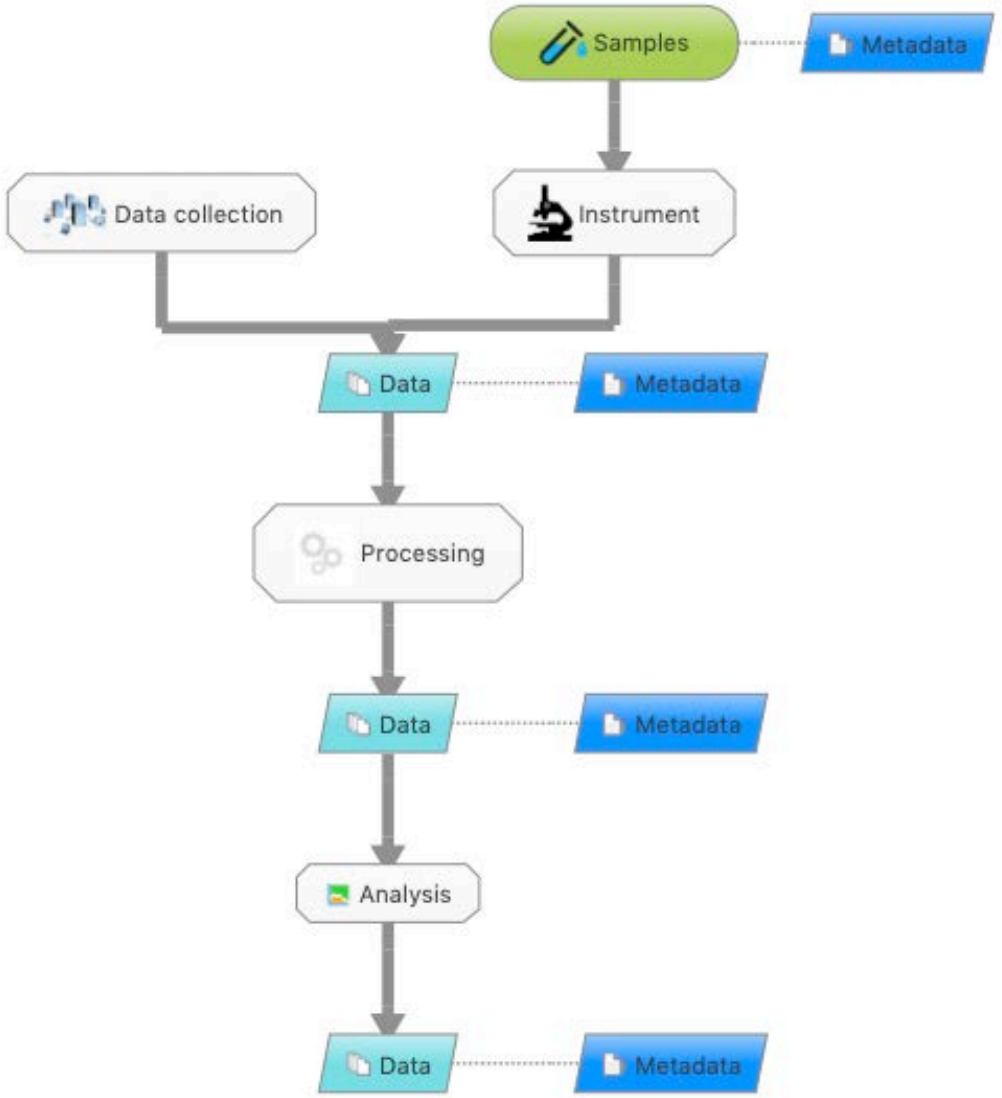
The data life cycle



Active data management:
Annotate, store, backup data while it is produced

Long term preservation:
Annotate, store, backup data at the end of a project or after publication

Research workflow in experimental & computational labs



FAIR Data Management

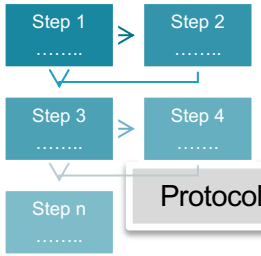
- Funding agencies and journals increasingly demand that data is published according to the **FAIR**[1] data principles (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)

- Lots of data are generated during a research project.

- Appropriate data and information management from the start of the research process can avoid “data drowning”

1. **The FAIR Guiding Principles for scientific data management and stewardship**, *Scientific Data, Issue 3, 2016. 10.1038/sdata.2016.18.*

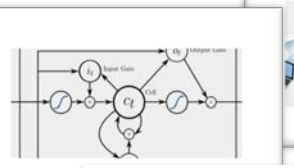
The "*data spread*": a common scenario in academic institutions



Protocols/ SOPs

Materials/samples

Code



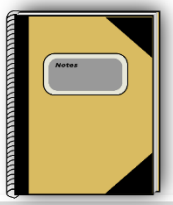
Model

Raw data

Processed data

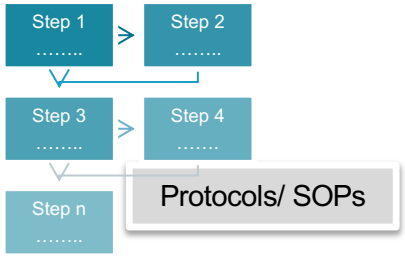
Results

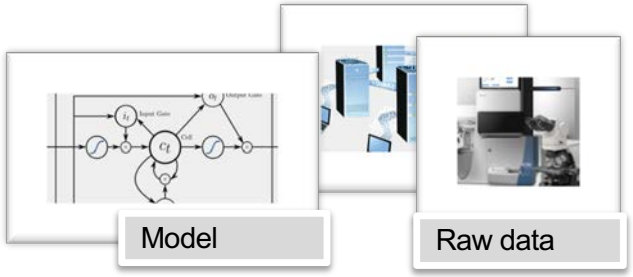
Analysis notebooks

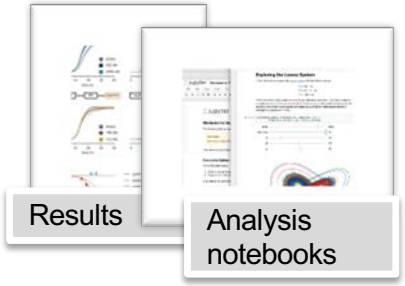


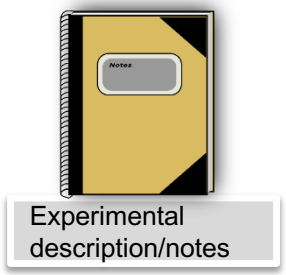
Experimental description/notes

The "*data spread*": a common scenario in academic institutions

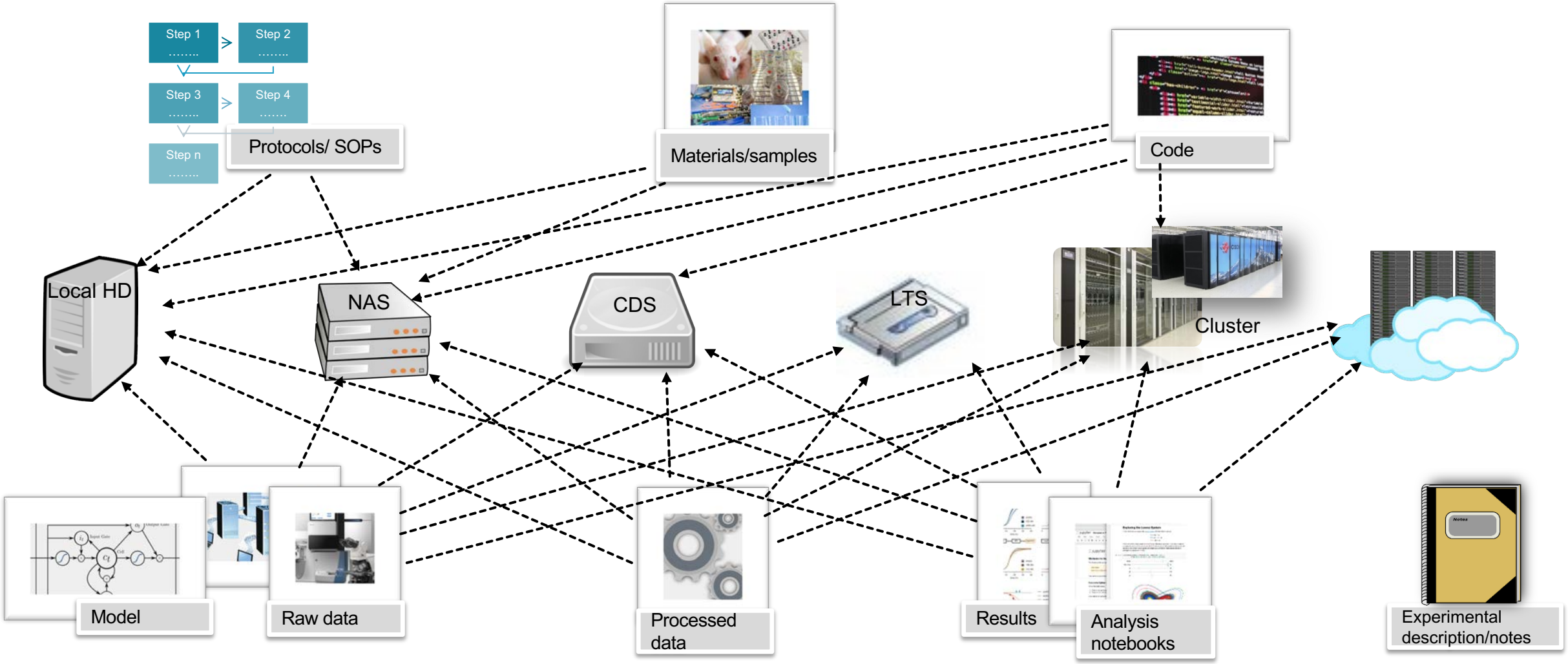








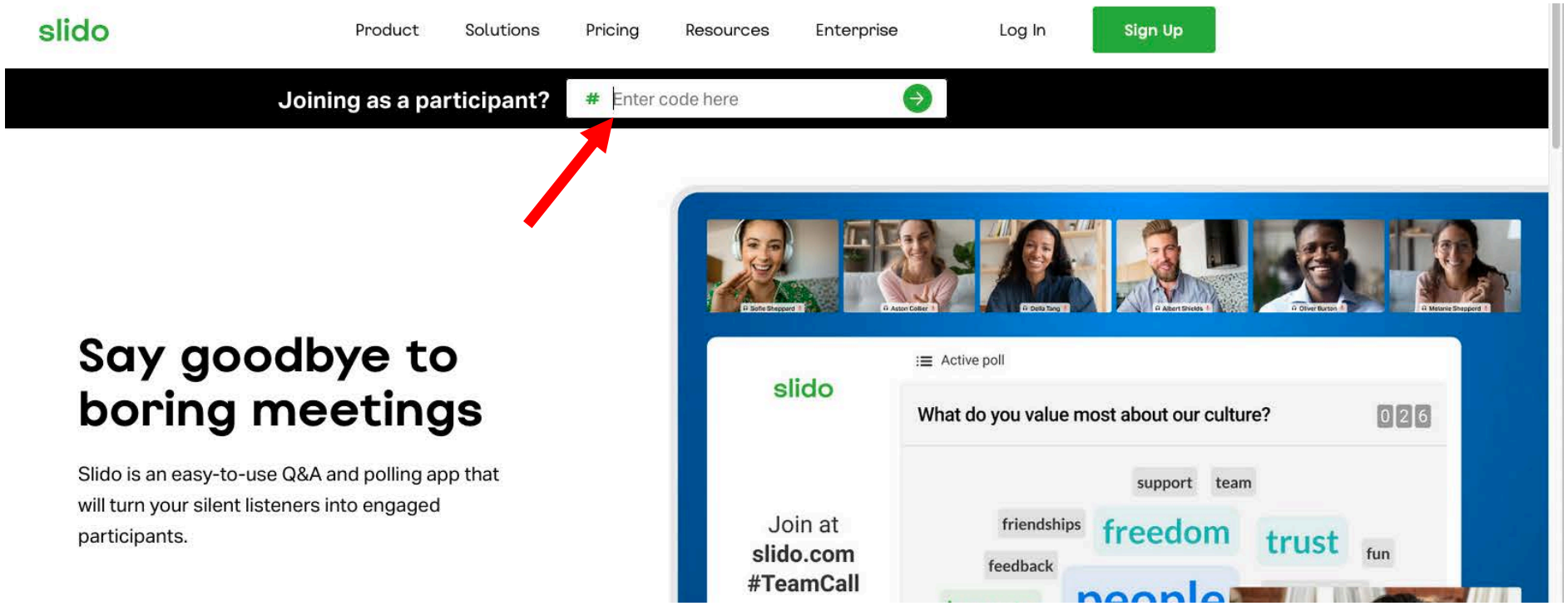
The "*data spread*": a common scenario in academic institutions

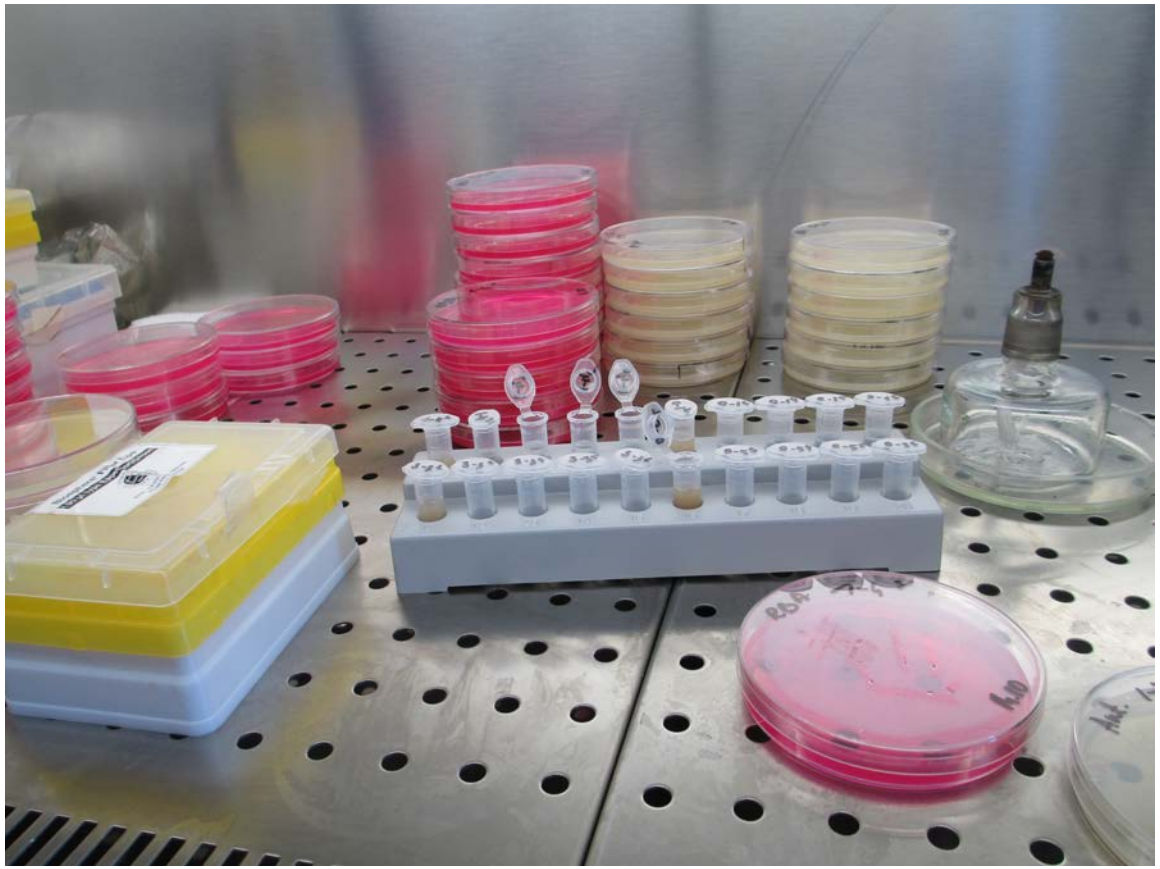


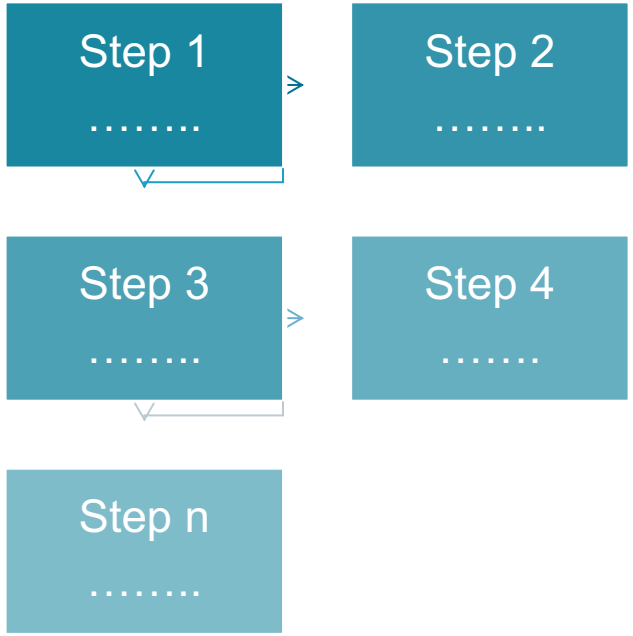
How can we take care of the individual components and how can we bring things together?

Discussion Time

- Go to www.slido.com and enter the event code ***#ETHRDM***







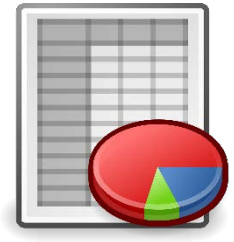
Management of samples and protocols

Management of materials and samples

What?

- Biological samples
- Chemical samples
- Materials
- Devices
- ….

How?



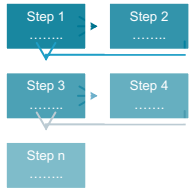
Spreadsheets

- *Not scalable*
- *No sharing*
- *No efficient search*
- *Easy to use*

Database/ LIMS

- *Scalable*
- *Sharing*
- *Search functionality*
- *Require time for set up and maintenance*

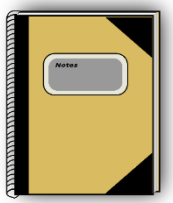
Management of protocols



What?

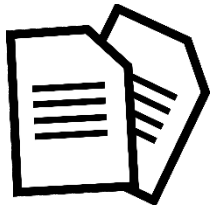
- Step by step description of procedure
- Experimental/computational parameters (e.g. temperature, time, etc.)
- Machine used (experimental)
- OS, program, version, etc. (calculation)

How?



Paper notebook

- *Not scalable*
- *No sharing*
- *No search*
- *Easy to use*



Text files

- *Not scalable*
- *No sharing*
- *No efficient search*
- *Easy to use*

WIKI

- *Scalable*
- *Sharing*
- *Search functionality*
- *Versioning*

Database/ LIMS

- *Scalable*
- *Sharing*
- *Search functionality*
- *Require time for set up and maintenance*

Laboratory Information Management System (LIMS)

- LIMS are software for managing laboratory operations:
  - **sample tracking** (*info about samples and about their storage*)
  - **sample data tracking** (*upload of data measured from samples, e.g. sequencing data, NMR, MS, etc.*)
  - **protocol management** (*info about sample preparation/handling*)

- LIMS were first used in companies for tracking the growing number of samples.

- LIMS were originally stand-alone solutions, which had to be integrated with other solutions such as ELNs. Nowadays **ELNs and LIMS are often combined** in one platform.

Wikis at ETH

- ETH ITS provides wiki services:

Confluence

https://ethz.ch/services/en/it-services/catalogue/web-application-hosting/wiki.html

GitLab

https://gitlab.ethz.ch

Size of Confluence Space/ git repository	Yearly price
<2 GB	free
<10 GB	300 CHF
<50 GB	1000 CHF

- Many departments and/or institutes host their own wiki. Always contact your ISG for info.

Example of confluence wiki use in SIS

ETH zürich Spaces People Calendars Create ··· Search

ID SIS

Calendars

PAGE TREE

- administration
- customers
- How-To's
- organisation
- outreach
- Personalized Health Data Services (PHdS)
- processes
- projects
- services
- SIS wiki organisation proposal
- trainings
  - openBIS trainings
    - 2019.09.24
    - 2019.10.09
    - **2020.01.23**
  - Organisational aspects of courses

Space tools

Dashboard / ... / openBIS trainings Edit Save for later Watching Share ···

2020.01.23

Created by Barillari Caterina (ID SIS), last modified less than a minute ago

Location	Zurich, HG F33.1
Time	09:00-13:00
Program	1. introduction to openBIS 2. How to manage the lab inventory of materials and samples 3. How to manage lab protocols 4. How to record experiments in the Electronic Lab Notebook & upload data 5. How to analyse data stored in openBIS using Jupyter notebooks and MATLAB
Participants	
Training material	https://gitlab.ethz.ch/sis-rdm-training/openbis-training

Like Be the first to like this No labels

Write a comment...

Powered by Atlassian Confluence 6.15.9 · Report a bug · Atlassian News

Example of versioning in wiki

Pages / ... / 2020

View Page Save for later ...

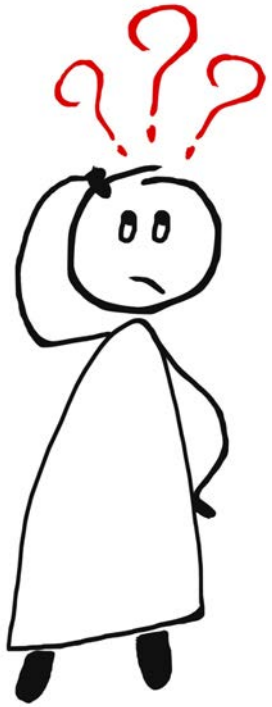
Page History

Compare selected versions

	Version	Published	Changed By	Comment	Actions
☐	**CURRENT (v. 17)**	**Jun 11, 2020 16:33**	Luetcke Henry (ID SIS)		
☐	v. 16	Jun 11, 2020 15:44	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 15	Jun 09, 2020 08:17	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 14	Jun 08, 2020 12:06	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 13	Jun 08, 2020 11:51	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 12	Jun 08, 2020 11:48	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 11	Jun 08, 2020 11:46	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 10	Jun 05, 2020 14:56	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 9	May 22, 2020 10:58	Plamada Andrei Valentin (ID SIS)	email update	Restore
☐	v. 8	May 22, 2020 10:58	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 7	May 22, 2020 10:46	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 6	May 18, 2020 14:29	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 5	May 18, 2020 14:28	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 4	May 15, 2020 14:29	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 3	Feb 14, 2020 14:34	Barillari Caterina (ID SIS)		Restore
☐	v. 2	Feb 14, 2020 14:34	Barillari Caterina (ID SIS)		Restore
☐	v. 1	Feb 14, 2020 14:32	Barillari Caterina (ID SIS)		Restore

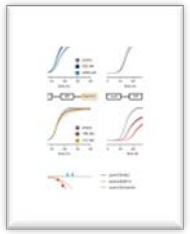
Return to Page Information

Questions on Management of Samples and Protocols?

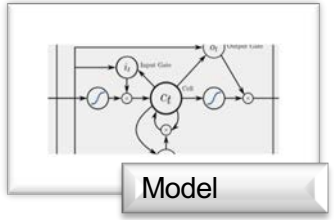


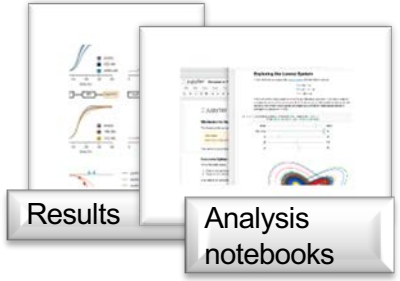
Data files management

Management of research data files

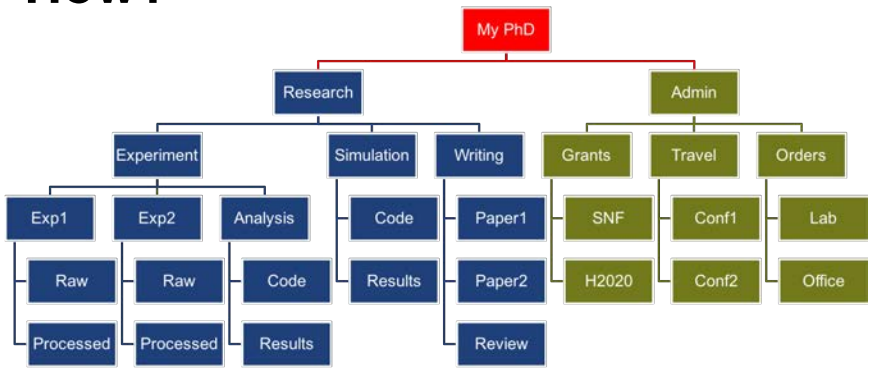


What?

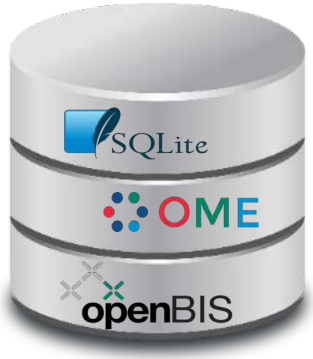




How?

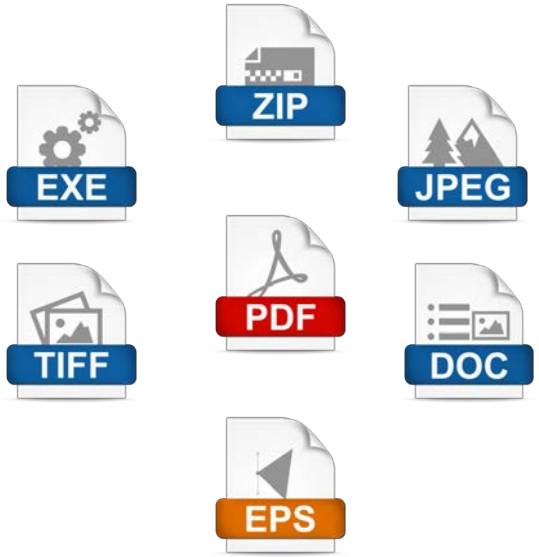


Files / folders hierarchy



Data management platform

File types and formats



- File format: a **convention for encoding information** in a computer file
- Extensions typically *indicate* a specific file format
- Some file formats are preferable to others
- Stick to **non-proprietary** and **widely used** formats!
- Several general-purpose **scientific data formats** exist (e.g. HDF5, netCDF, FITS)

-	+
Binary	Text-based
Proprietary	Open
New kid on the block	Old as the hills
Compressed/encrypted	Uncompressed/unencrypted
Platform dependent	Interoperable
Complex	Simple

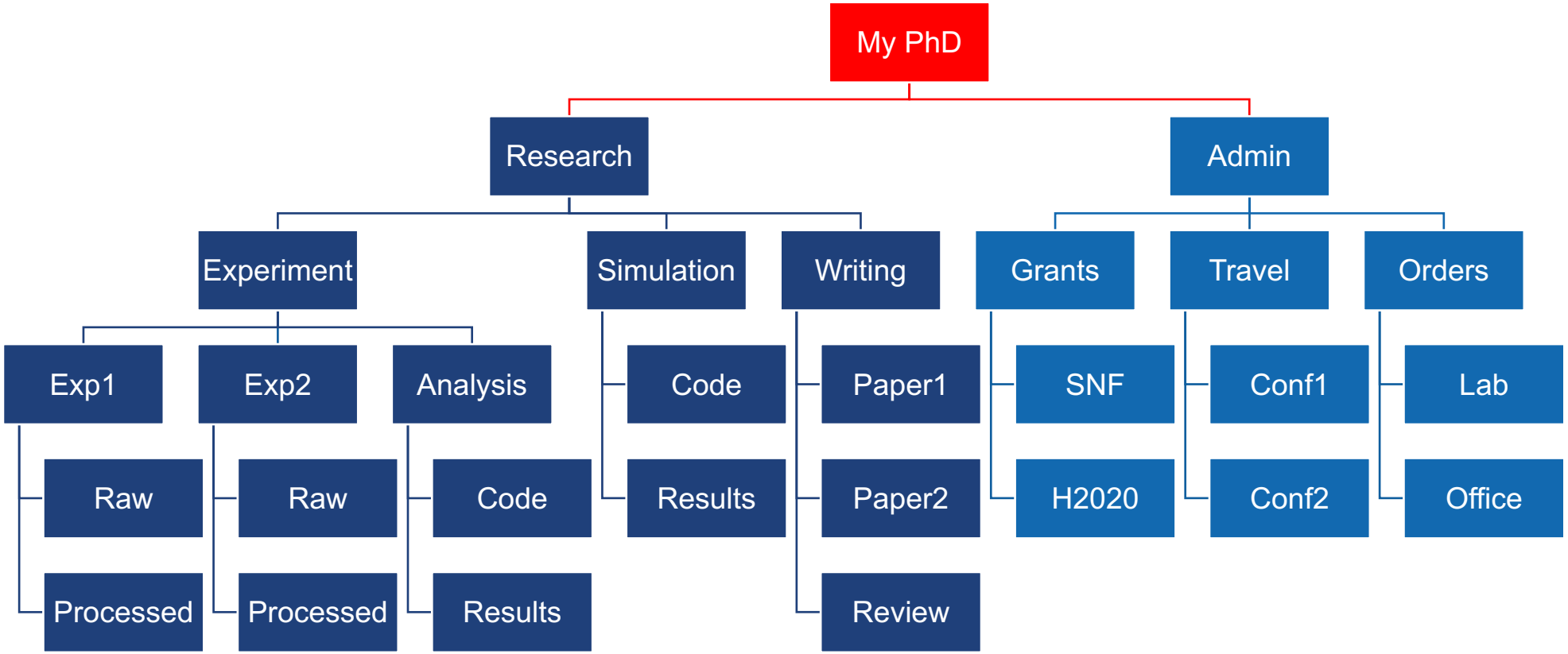
File & folder organisation

- **Goals** of efficient file / folder organization:
  - Easy to **find** something in the future (you, others)
  - Easy to **file** something
  - Save **disk space** (avoid multiple copies of files)
  - **Reusable** components
  - Avoid problems on different **operating systems**

- Planning a good **folder hierarchy**
  - **Who?** Individual, research group
  - **Where?** Local disk, shared network drive
  - Which operating system?
  - What information are you going to **search** for?
  - Avoid non-descriptive file and folder names (*figure_02_summary_stats.png* and not *stats.png*)
  - Add **descriptive text files** to folders (→ Metadata)
  - **Document** your hierarchy and file naming convention

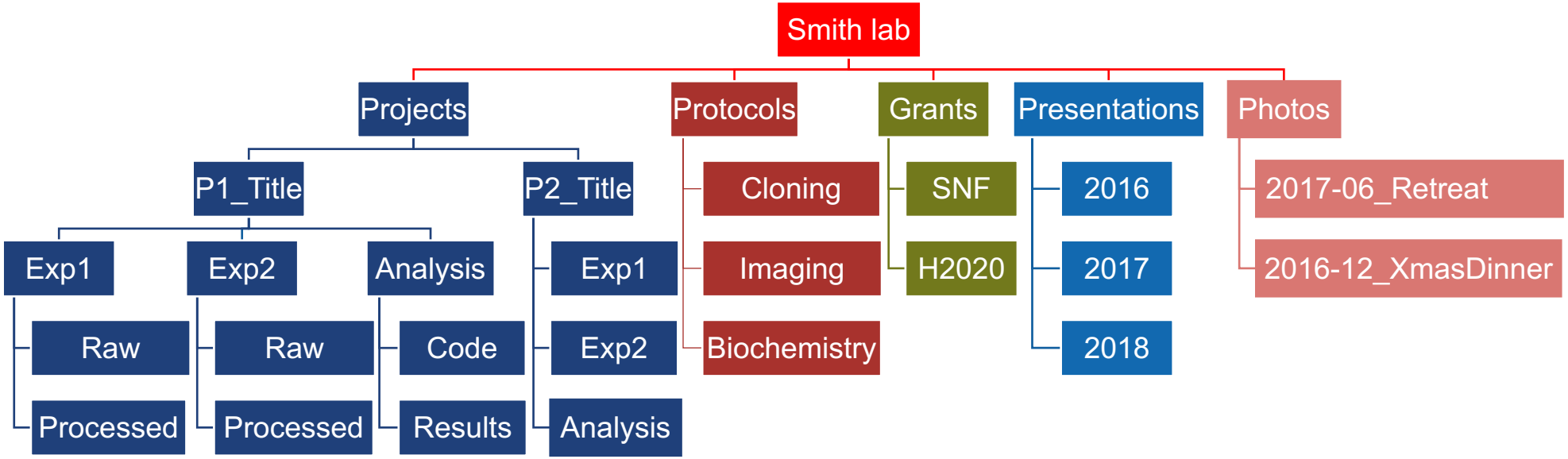
File & folder organisation

Example hierarchy for a PhD project



File & folder organisation

Example hierarchy for a research group



File & folder organisation

The project directory (for a computational project)

```
project
|- doc/                 documentation for the study
|
|- data/                raw and primary data, essentially all input files, never edit!
|   |- raw_external/
|   |- raw_internal/
|   |- meta/
|
|- code/                all code needed to go from input files to final results
|- notebooks/           notebooks that document your day-to-day work
|
|- intermediate/        output files from different analysis steps, can be deleted
|- scratch/             temporary files that can be safely deleted or lost
|- logs/                logs from the different analysis steps
|
|- results/             output from workflows and analyses
|   |- figures/
|   |- tables/
|   |- reports/
|
|- Snakefile            project workflow, carries out analysis contained in code/
|- config.yml           configuration of the project workflow
|- environment.yml      software dependencies list, used to create a project environment
|- Dockerfile           recipe to create a project container
```

Working in projects

RStudio

ATOM

Sublime Text

Visual Studio Code

git

Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424

File & folder organisation

A project in ATOM

Dockerfile — ~/Documents/projects/project_template

Project | Remote | Dockerfile

- project_template
  - code
  - data
  - doc
  - intermediate
  - logs
  - notebooks
  - results
  - scratch
  - ftpconfig
  - .gitignore
  - config.yml
  - Dockerfile
  - environment.yml
  - README.md
  - Snakefile

```
FROM ubuntu:16.04
LABEL description = "Lightweight image with Conda, Jupyter Notebook and Snakemake"

# Install Miniconda3 and prerequisites
RUN apt-get update && \
    apt-get install -y --no-install-recommends bzip2 curl ca-certificates
RUN curl https://repo.continuum.io/miniconda/Miniconda3-4.5.11-Linux-x86_64.sh -O && \
    bash Miniconda3-4.5.11-Linux-x86_64.sh -bf -p /opt/miniconda3/ && \
    rm Miniconda3-4.5.11-Linux-x86_64.sh

# Add Conda to PATH
ENV PATH="/opt/miniconda3/bin:${PATH}"

# Use bash as shell
SHELL ["/bin/bash", "-c"]

# Set up the Conda environment
COPY environment.yml .
RUN conda env update -n root -f environment.yml && \
    conda clean --all

# Install Jupyter Notebook and set default user to UID 1000
RUN pip install --no-cache-dir notebook==5.*
ENV NB_USER nbuser
ENV NB_UID 1000

RUN adduser --disabled-password --no-create-home \
    --gecos "Default user" \
    --uid ${NB_UID} \
```

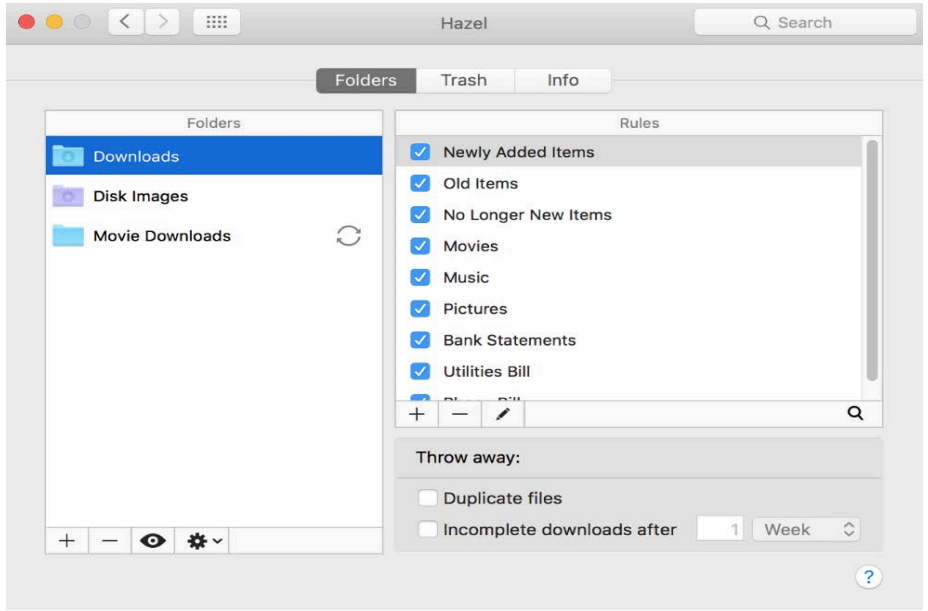
```
liv433l:project_template arasmus$ pwd
/Users/arasmus/Documents/projects/project_template
liv433l:project_template arasmus$
```

Dockerfile 22:46 | LF UTF-8 Dockerfile git+ master Fetch 1 file

File & folder organisation

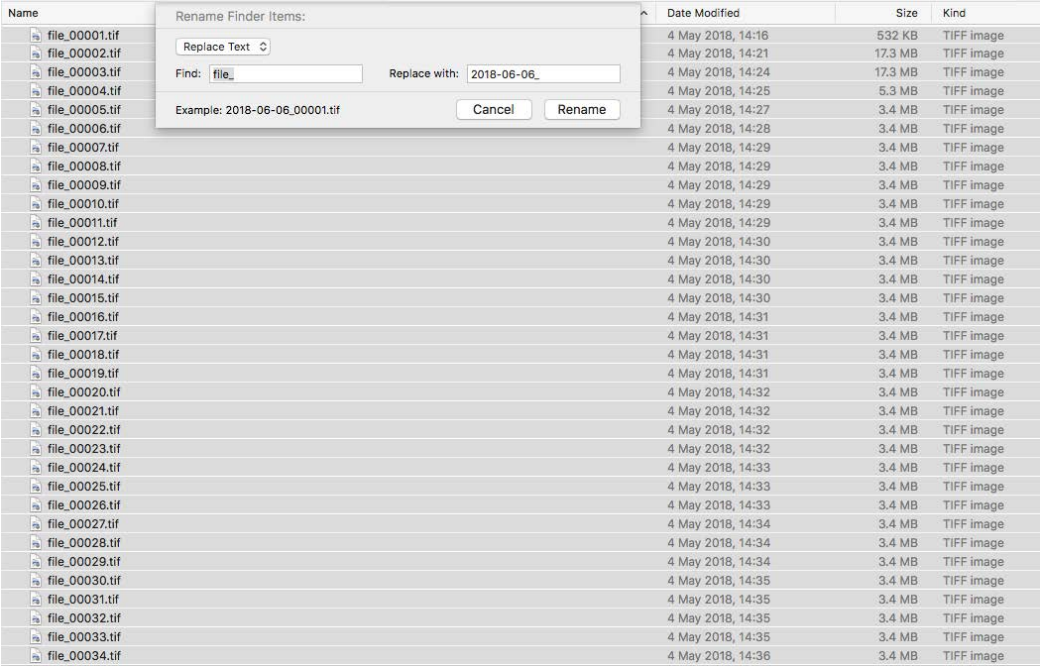
- Keep path names short (< 256 characters)
- Recommendation for file names:
  - Unique, reflect content (if possible)
  - Use only ASCII characters
  - Not include spaces
  - Be aware of case sensitivity
- Bad examples:
  - data%20management%20plan.docx
  - sup figure 2.png
  - lab meeting 19.10.2019.pptx
- Good examples:
  - Data_management_plan_SNF.docx
  - sup_figure_02_summary_stats.png
  - lab_meeting_2019-10-19.pptx

- Use links / shortcuts to avoid duplications
- Use tags for orthogonal classifications
- Create template folders
- Some tools for automated file organization:
  - Mac: Hazel (https://www.noodlesoft.com/)
  - PC: DropIt (http://www.dropitproject.com/)



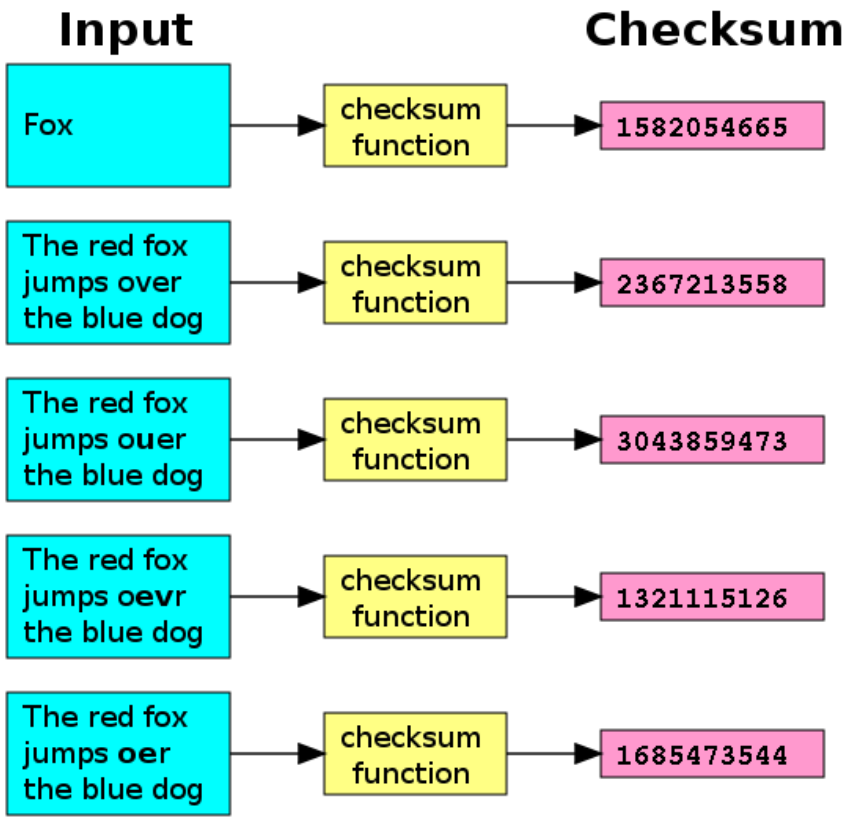
Batch renaming of files

- **Windows**
  - Bulk Rename Utility (www.bulkrenameutility.co.uk)
  - Advanced Renamer (www.advancedrenamer.com)
  - Command prompt / PowerShell scripts
- **macOS**
  - Finder rename functionality
  - Automator
  - Command line / scripts
- **Linux**
  - Command line utility `rename`
  - Métamorphose (http://file-folder-ren.sourceforge.net/)
  - pyRenamer (https://launchpad.net/pyrenamer)



File validation and compression

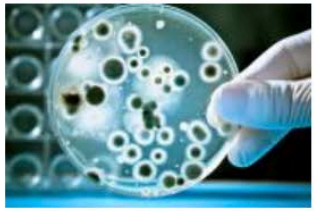
- **Checksum** algorithms are useful to **verify data integrity**
  - For datasets stored over long time periods
  - When transferring from one storage to another
  - Routinely performed by repositories
  - Examples: *cksum*, *md5sum*, *sha1sum*
- **Compression** algorithms encode information in fewer bits than the original (bit-rate reduction)
  - Lossless vs. lossy compression
  - Lossy compression may lead to distortion / loss in quality but compression rates are typically much higher
  - Trade-off between processing time and disk space
  - Examples: zip, gzip, specific formats (TIFF, HDF5)



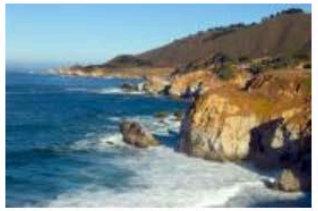
Metadata

- **Metadata** is the *data about your data* (a.k.a. **data model**)
- Use of structured metadata **facilitates data organization** and searches
- **Metadata** is a key element of the **FAIR data** principles
- Existing **metadata schemas** are preferred (can be extended, if necessary)

Search by Discipline

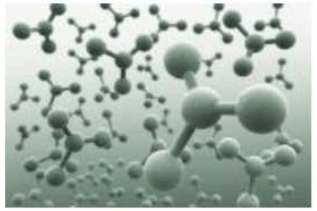


Biology



Earth Science

General Research Data

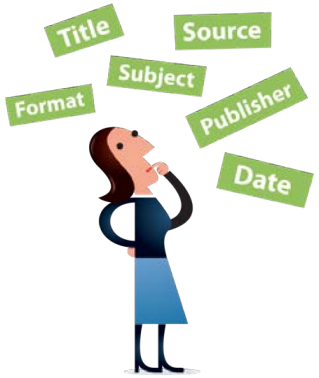


Physical Science

Social Science & Humanities

https://www.dcc.ac.uk/guidance/standards/metadata

Metadata



www.digitalbevaring.dk

Example for general research data: **DataCite Metadata schema**

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

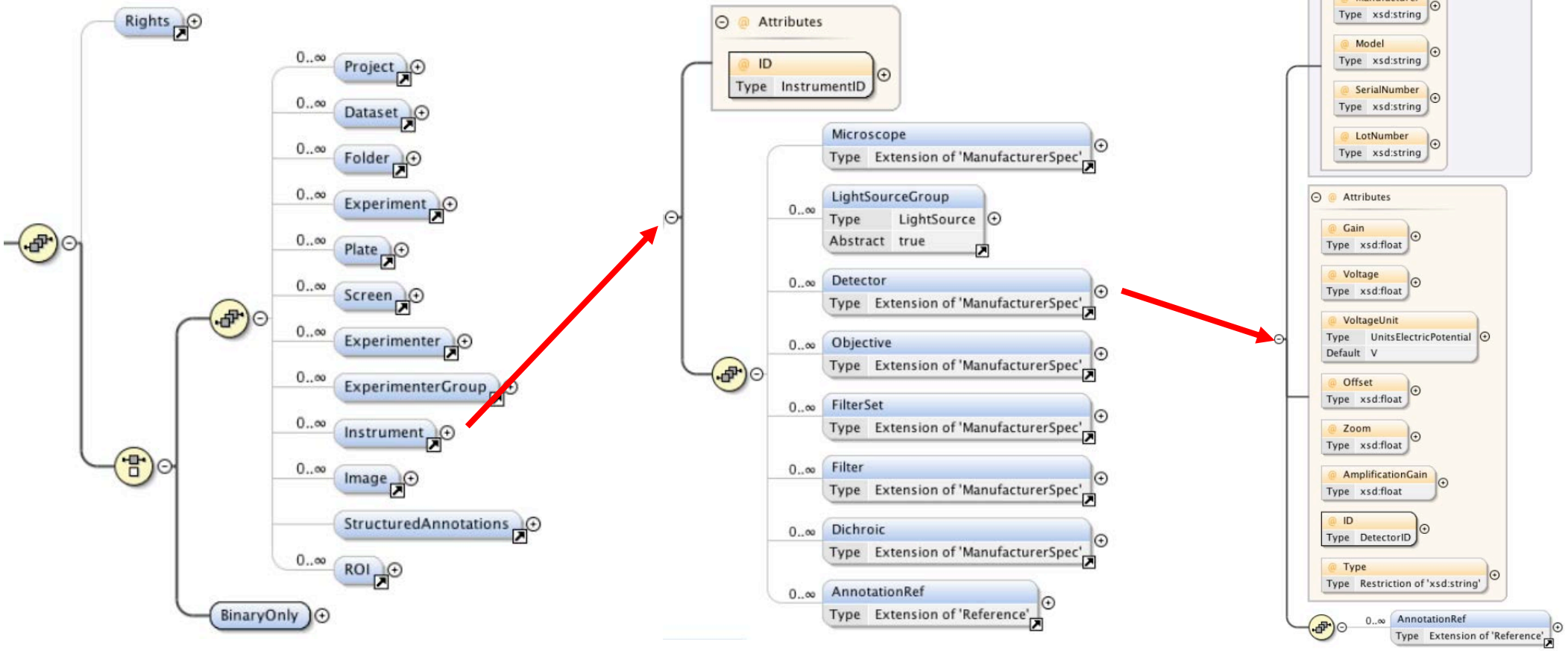
https://schema.datacite.org/

Table 2: DataCite Recommended and Optional Properties

ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O

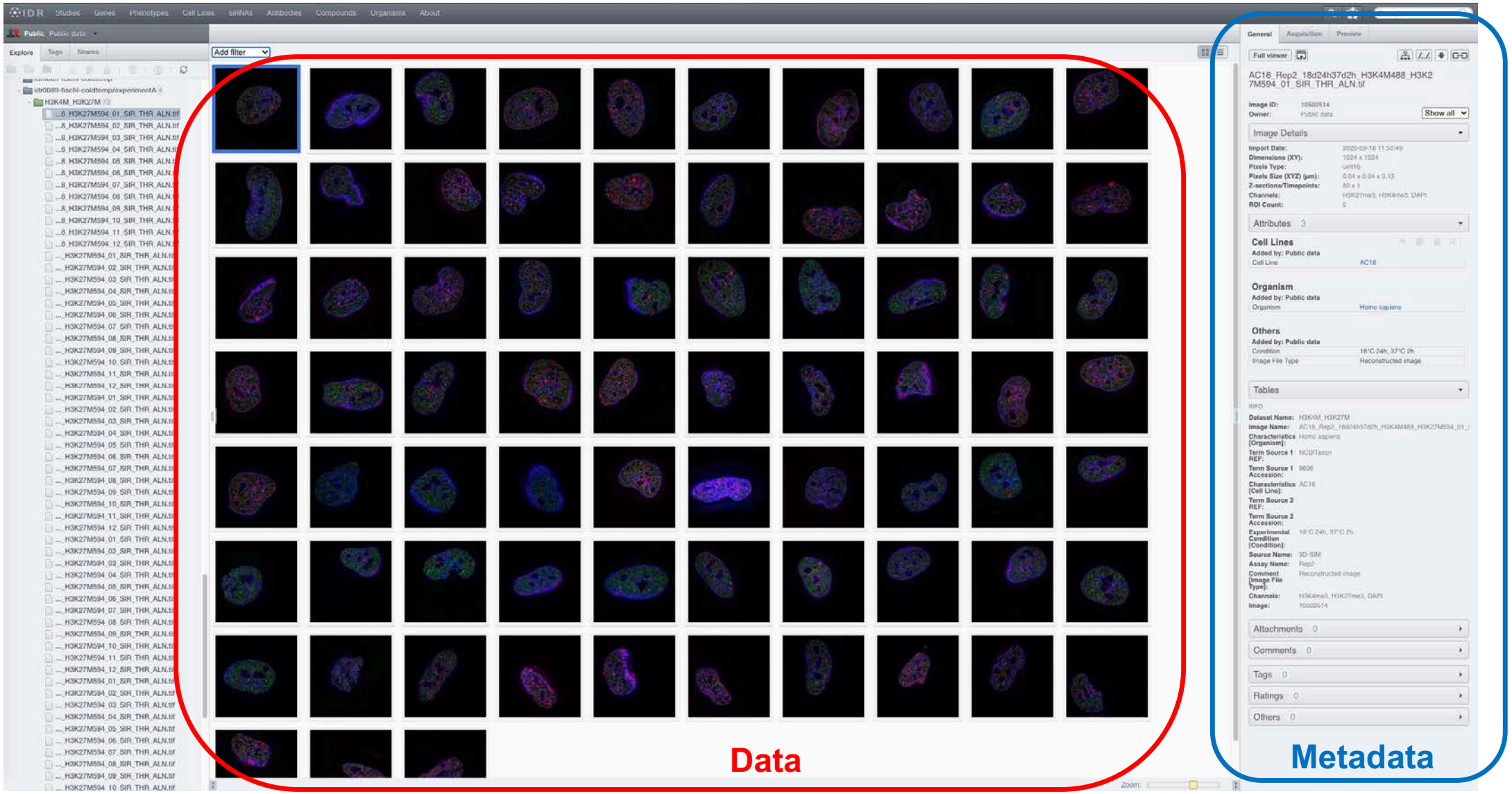
Metadata

Example for discipline-specific data: **OME Metadata schema**



https://www.openmicroscopy.org/Schemas/

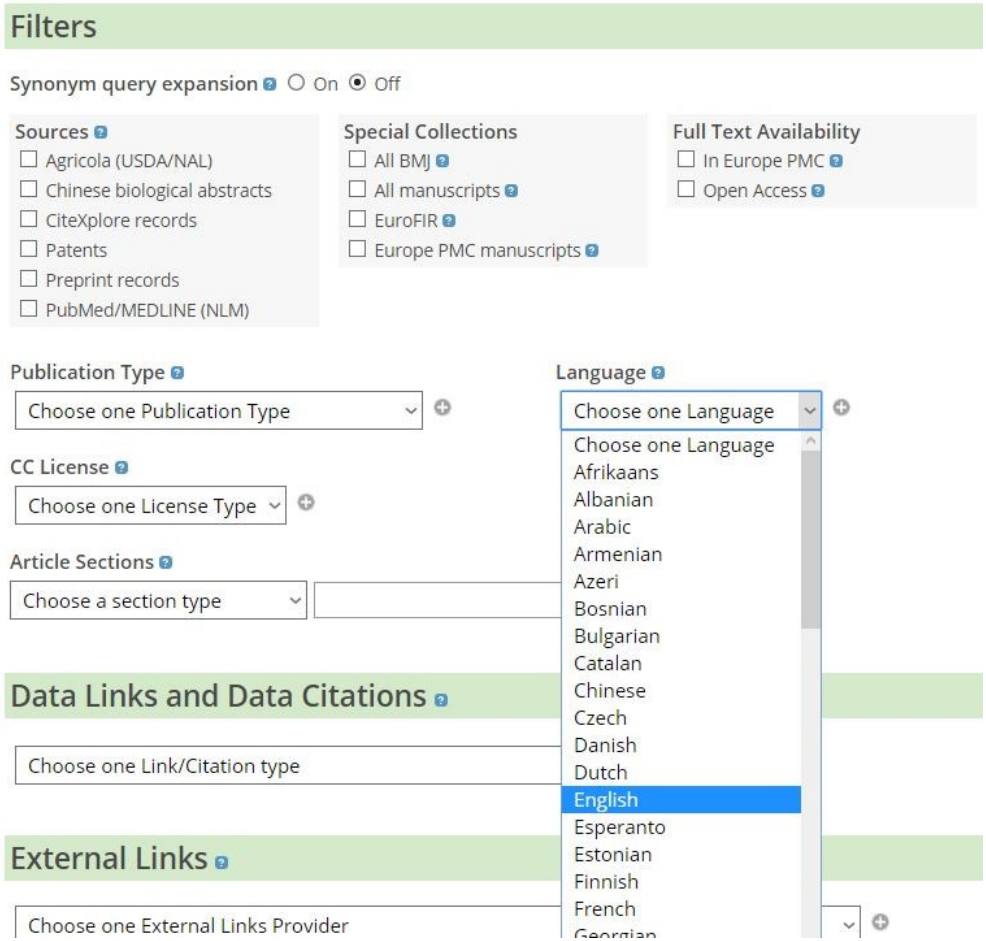
Metadata



https://idr.openmicroscopy.org/webclient/?show=project-1304

Metadata

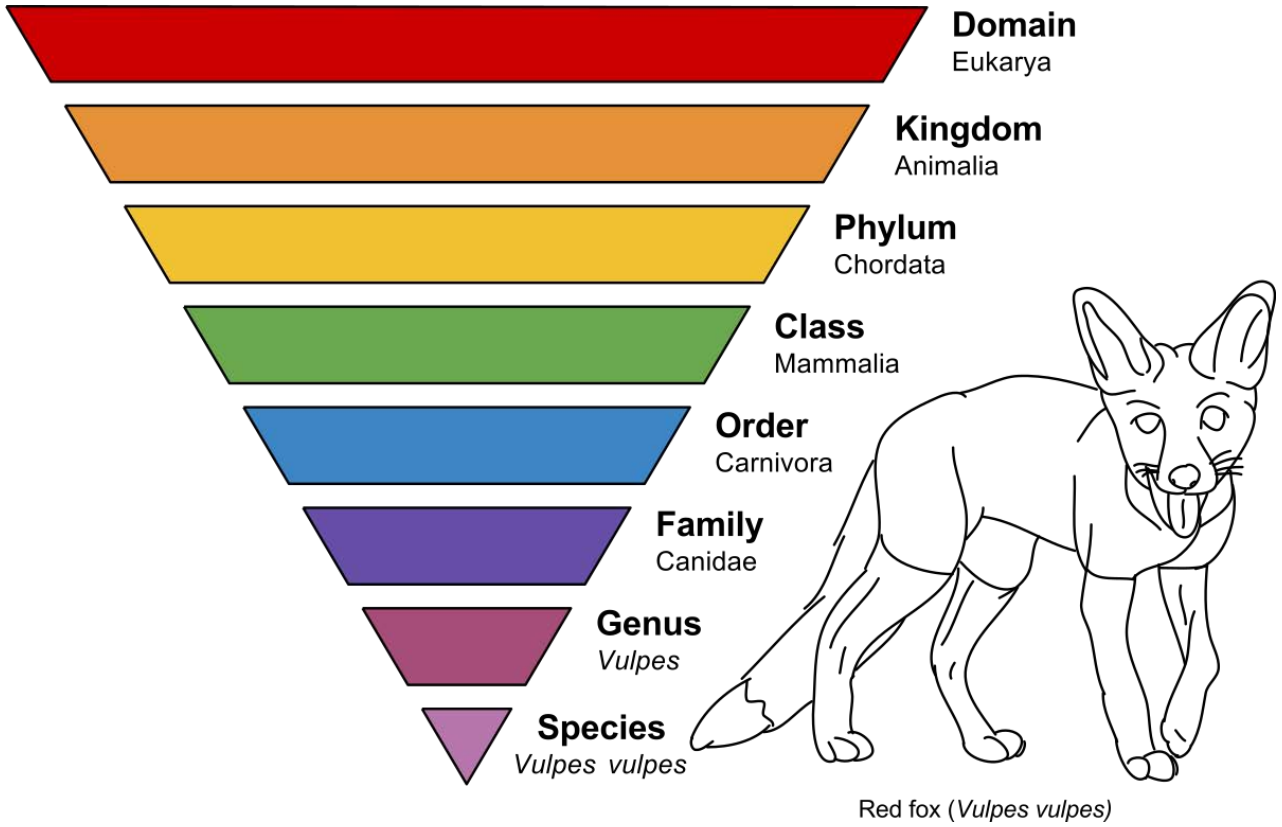
- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies



https://europepmc.org/advancesearch

Metadata

- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies



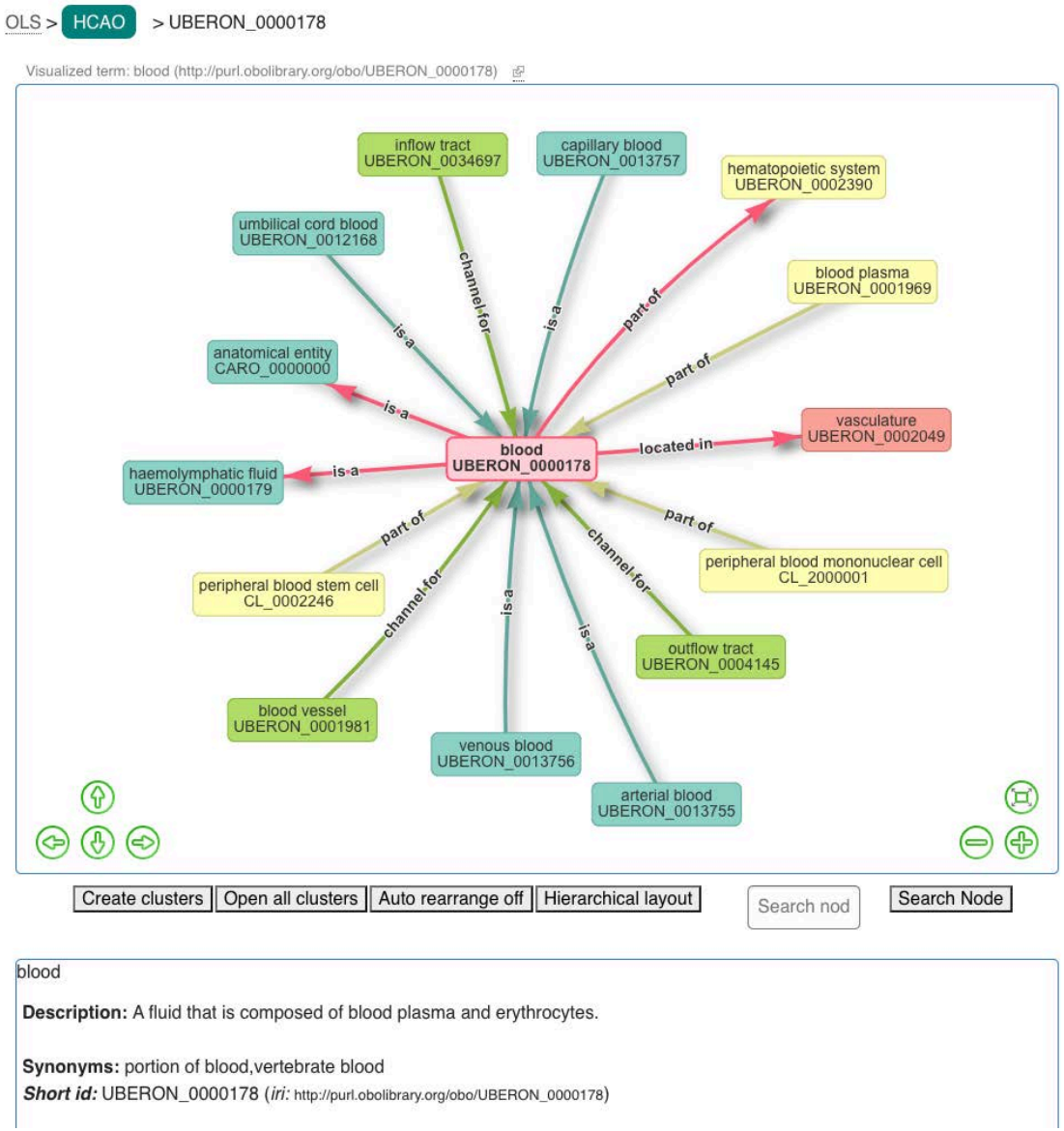
https://en.wikipedia.org/wiki/Domain_%28biology%29

Metadata

- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies
  - Standardized metadata are the basis for ontologies (knowledge representations)

  [Basic Register of Thesauri, Ontologies & Classifications]

  [EMBL-EBI Ontology Lookup Service]



Knowledge graph for *'blood'* from the [Human Cell Atlas Ontology]

Metadata

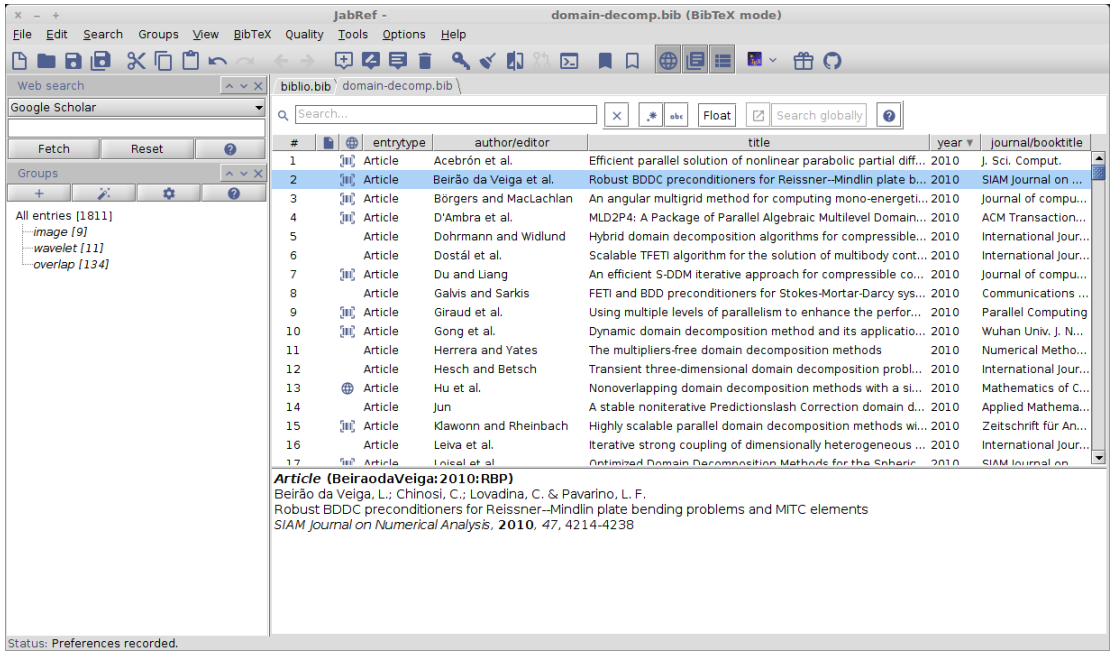
- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies
  - Standardized metadata are the basis for ontologies (knowledge representations)

    Basic Register of Thesauri, Ontologies & Classifications

    EMBL-EBI Ontology Lookup Service
- **How to apply metadata?**
  - File header (e.g. TIFF, netCDF)
  - Sidecar files (e.g. XML, JSON)
  - As columns in a database
  - README text files
- **Apply metadata early** in your workflow (ideally during acquisition)

Data management software

- System that allows structured organization of data
- Data is described by metadata
- Searchable, scalable, flexible
- Allows user rights management
- Back up procedures are easily implemented



Data management software

- System that allows structured organization of data
- Data is described by metadata
- Searchable, scalable, flexible
- Allows user rights management
- Back up procedures are easily implemented
- Examples

Generic

PostgreSQL SQLite FileMaker An Apple Subsidiary Microsoft Access

Scientific

openBIS LabKey PostGIS Spatial PostgreSQL

OME rasdaman raster data manager

Data management software

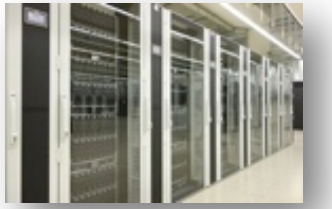
Generic database workflow

Measure / Simulate

Automated or manual

Import & Annotation

- **Search**
- **List**
- **Curate**
- **Share**



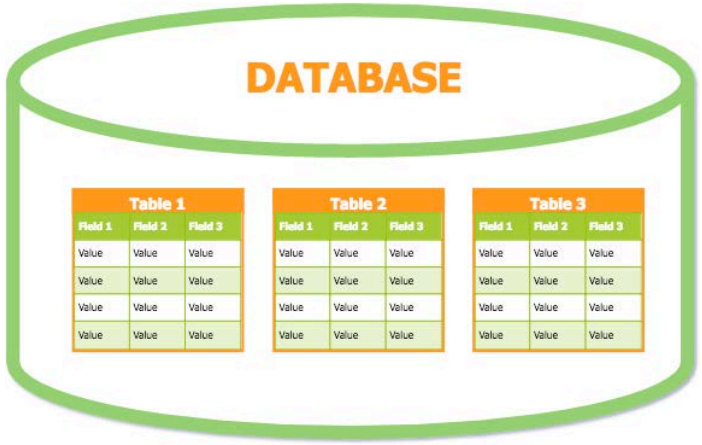
- **Process**
- **Analyze**
- **Visualize**

- **Backup**
- **Archive**

Data management software

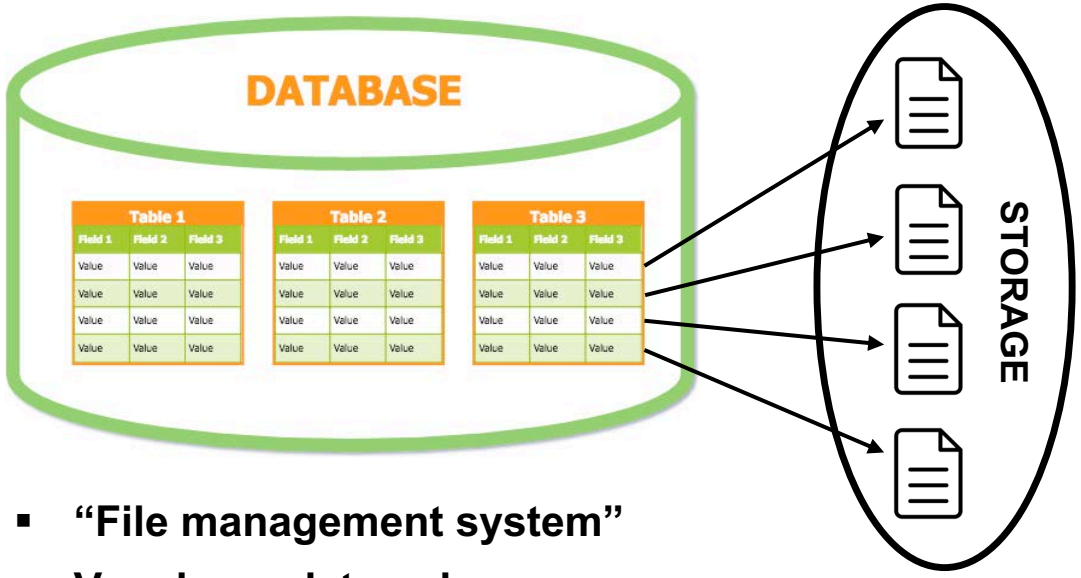
Where is my data???

Metadata + Data



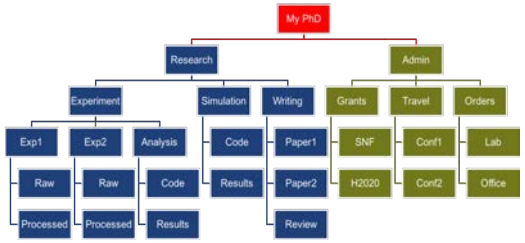
- Structured data required
- DB-specific data (array DBs)
- No conventional file access

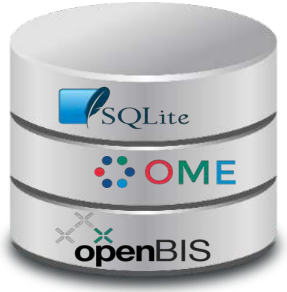
Metadata + Link to Data Files



- “File management system”
- Very large data volumes
- Conventional file access

Comparison: folder hierarchy vs. databases

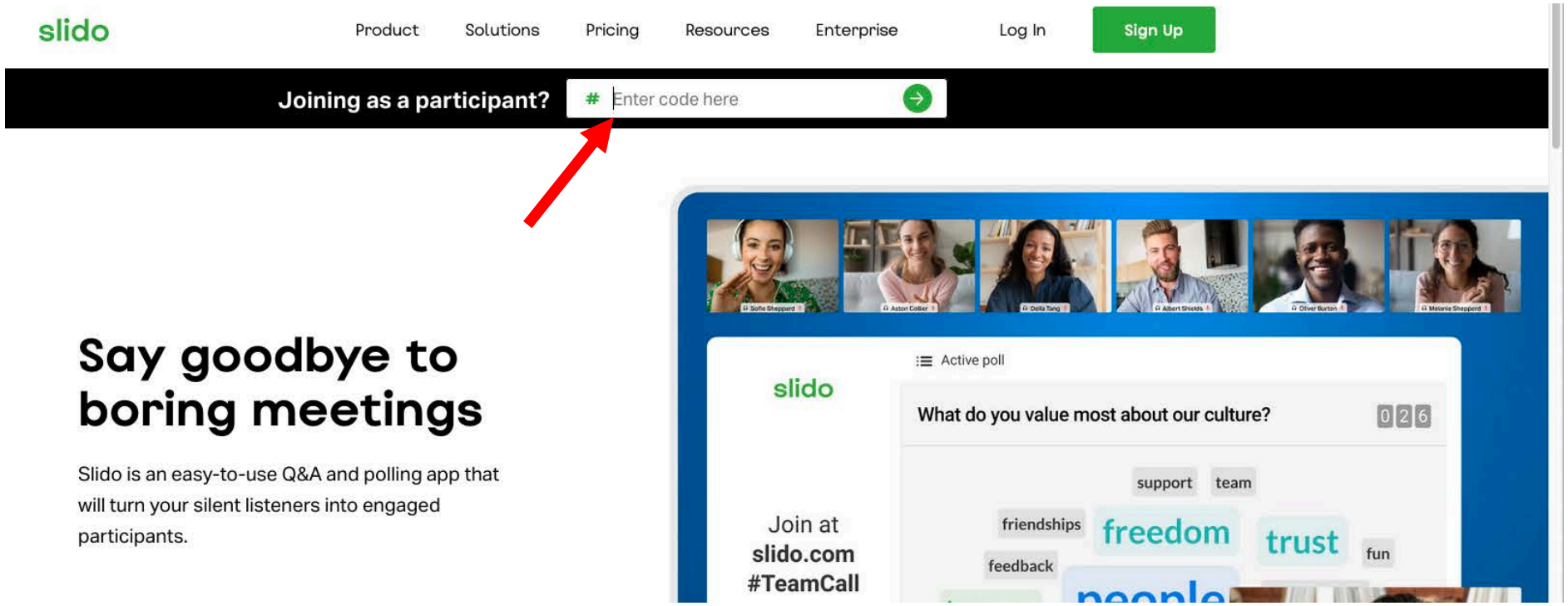




	File / folder Hierarchy	Database System
Easy to use	Yes	Initial learning curve
Flexibility	High	Limited
Scalability (data volume)	Limited	Yes
Scalability (users)	Limited	Yes
Versioning included	No	Yes
Backup included	No	Yes

Data storage – Your options …

- Go to www.slido.com and enter the event code ***#ETHRDM***



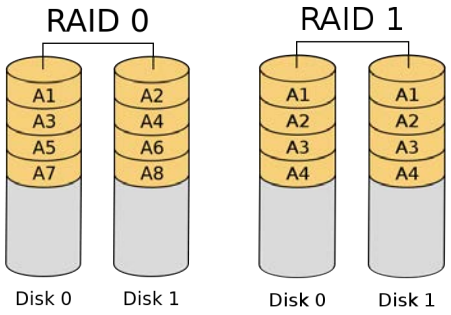
Data storage – Your options …

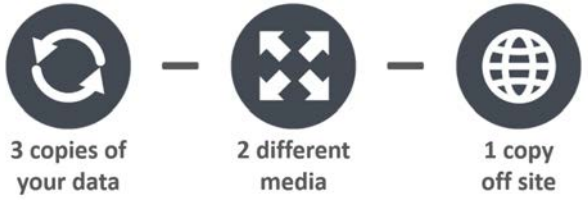
- External hard disks are cheap but unreliable and don’t scale!

Data storage – Your options …

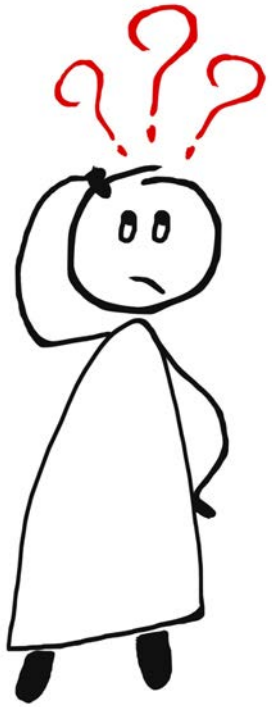
- External hard disks are cheap but unreliable and don't scale!
- Polybox / SwitchDrive
- Network Attached Storage (NAS)
  - NAS offering from IT Services – always check with your ISG first
- Cost Defined Storage (CDS)
  - For very large storage requirements (>100 TB)
- Long-term Storage (LTS)
  - Tape storage for data worthy of preservation ("Datenendlager")
  - Data should be compressed and sized between 10 - 200 GB
- Also consider backups (ideally follow 3-2-1 rule)
  - RAID is not a backup!
- Details: https://ethz.ch/services/en/it-services/catalogue/storage.html
- ETH news on cloud storage regulations: https://bit.ly/3BoQQD1

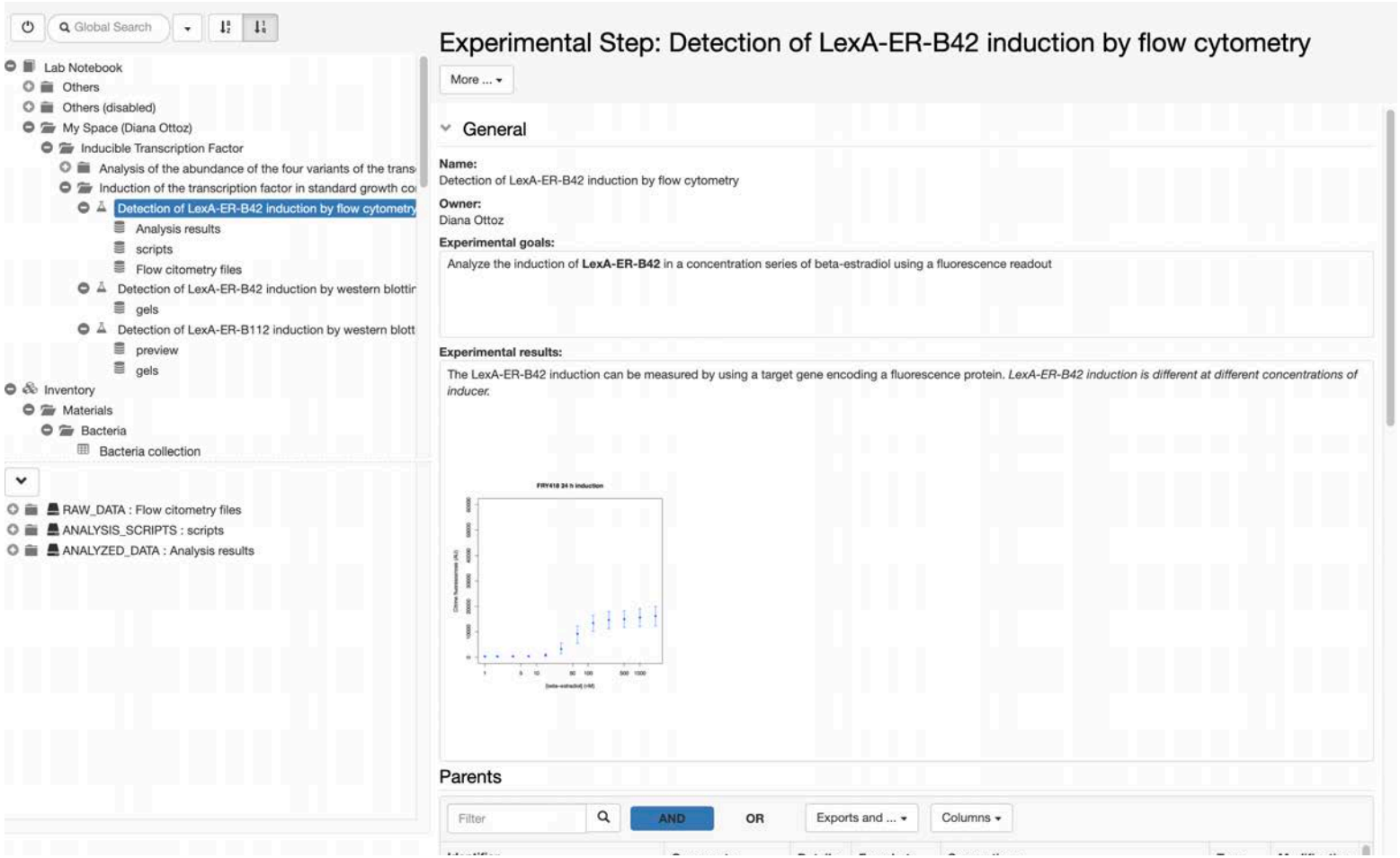
Redundant
Array of
Independent
Disks





Questions on Management of Data Files?





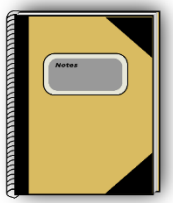
Electronic Laboratory Notebooks

Experimental description / notes

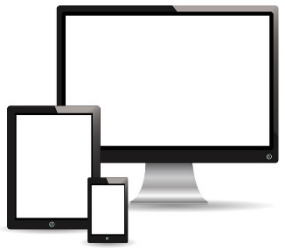
What to document?

- Goals
- Materials
- Methods
  - Experimental/computational procedure
  - Analysis procedures
- Results
- Links to data

How?



Paper laboratory notebook



Electronic laboratory notebook (ELN)

Definition of ELN & requirements

*An **electronic laboratory notebook** (also known as **electronic lab notebook** or **ELN**) is a software program or package designed to replace more traditional paper laboratory notebook . Laboratory notebooks in general are used by scientists and technicians to document, store, retrieve, and share fully electronic laboratory records in ways that meet all legal, regulatory, technical and scientific requirements.*

Legal requirements

Electronic lab notebooks used for development or research in regulated industries, such as medical devices or pharmaceuticals, are expected to comply with the **21 CRF Part 11 FDA** regulations: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?cfrpart=11

- To our knowledge, in academia, there is no such requirement, but this can vary from one Institution to another.

Source: https://www.limswiki.org/index.php/Electronic_laboratory_notebook

ELNs vs. paper notebook

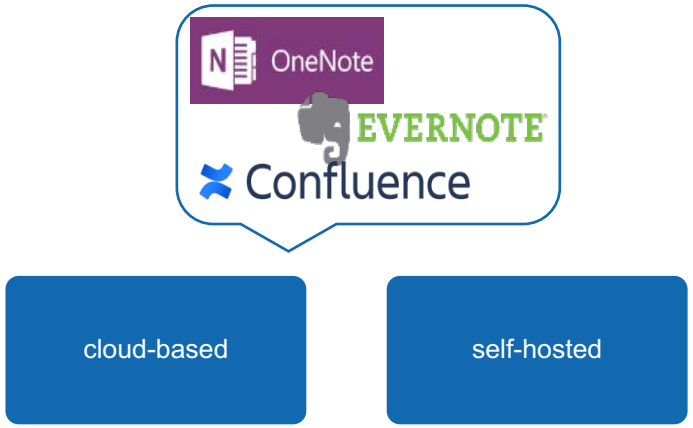
Advantages of ELNs over paper notebooks:

1. Sharing
2. Most ELNs have rights management
3. Most ELNs keep track of changes
4. Searching
5. Easier to link digital data
6. No issues with handwriting
7. Can be backed up

Disadvantages of ELNs over paper notebooks:

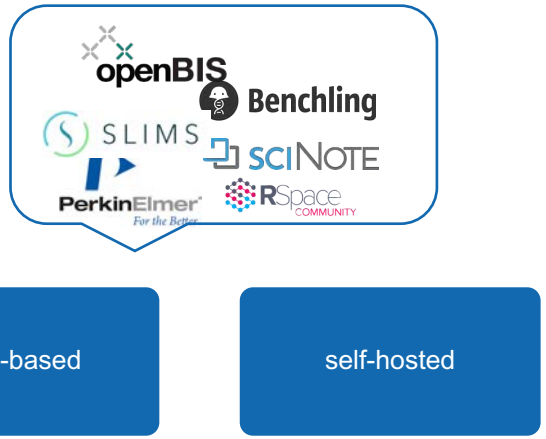
1. Require change in working mode
2. Have a learning curve

Note-keeping applications



- Most solutions are moving towards cloud-based services
- Straight replacement of paper notebooks with some added values (e.g. sharing, searching)
- Popular in academia due to ease of use
- Do not provide a solution for data management
- Do not comply with **21 CRF Part 11 FDA** regulations

ELNs with database back end



- Structured ELNs compared to note-keeping applications
- Additional functionalities compared to note-keeping applications (e.g. workflow management, chemical structures drawing, etc)
- Can be discipline-specific or cross-disciplines
- Some systems offer an all-in-one solution for RDM
- Some systems comply with **21 CRF Part 11 FDA** regulations

Which ELN to choose?

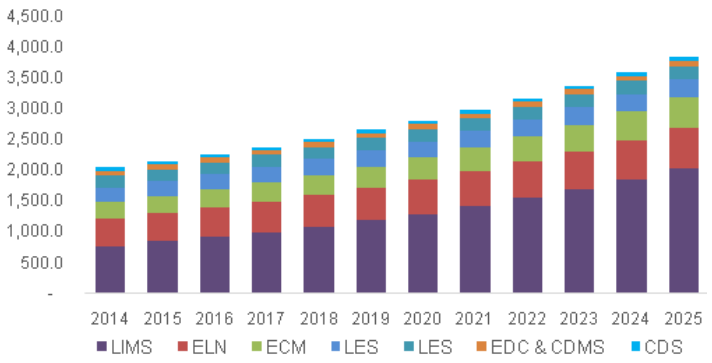
- The ELNs offer on the market is very large. Answering some of these questions might help you restricting the choice:

  1. Is it for personal use or group use?
  2. Can I/we use a cloud-based solution?
  3. Do I/we need specific features?
  4. What do I/we want to do with the ELN? (e.g. only write experimental descriptions, manage samples, manage data – how big?, etc.)
  5. Commercial v. open-source
  6. Budget?
  7. Can I export my data?

- *How to pick an Electronic Laboratory Notebook:* https://www.nature.com/articles/d41586-018-05895-3
- *Harvard University Comparison Grid:* https://datamanagement.hms.harvard.edu/electronic-lab-notebooks

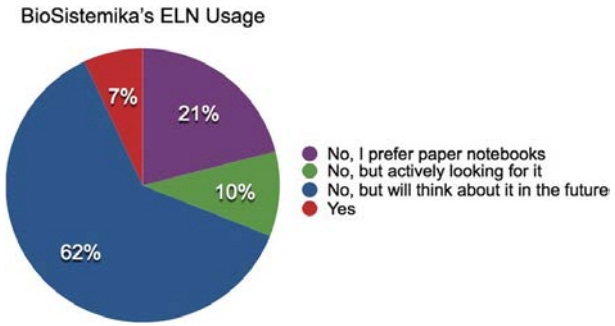
Current state of the art

- ELNs and LIMS are widely adopted in industry and continuous growth is forecasted for the next years.



Source: Grand View Research (http://www.grandviewresearch.com/industry-analysis/laboratory-informatics-market)

- ELNs and LIMS are not yet widely adopted in academia, but interest in them is starting to grow.



Source: Biosistemika's Webinars: Are you using ELNs on your daily lab routine?. J. Cheminform (2017) 9:31.

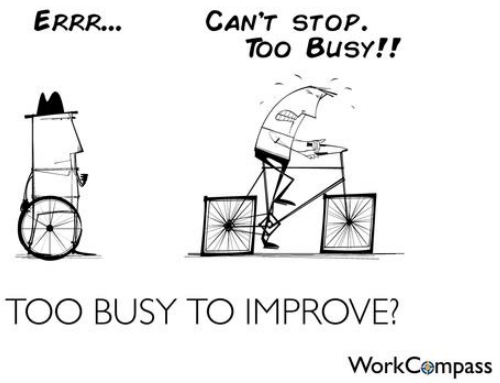
Reasons for slow adoption of ELNs/LIMS in academia

The Barriers	Research lab	Diagnostic lab
Limited budget	352	23
Time needed for implementation	235	21
Changes to existing working habits	266	25
Concerns about ELN being a system in the cloud	215	22
Contains features we do not need	130	7
Software in English only	76	10
No need for improvement	112	7

Source: Potential uses of ELNs in Academia survey (University of Southampton). J. Cheminform (2017) 9:31.

In our experience at ETH, the main reasons are:

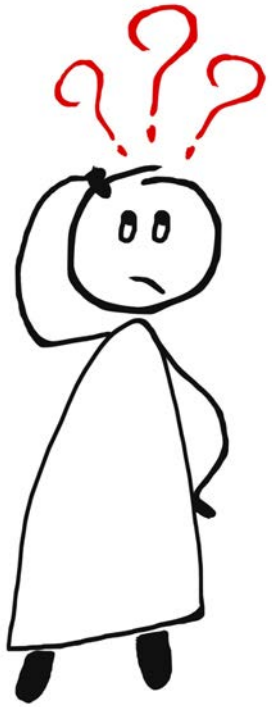
1. Change in working habits needed.
2. Time needed for introduction in a lab.
3. Concerns about data retrieval when leaving the lab.

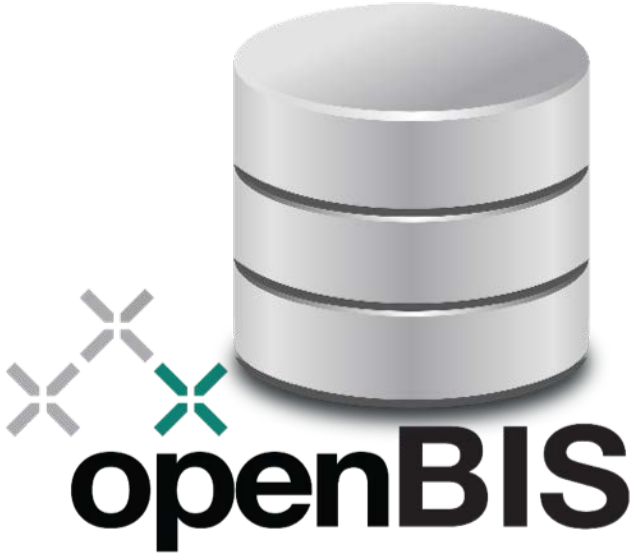


The top-down approach…

- Since 2018 we have seen a steady increase in demand for DM services, due to the new SNFS requirements..

Questions on Electronic Lab Notebooks?

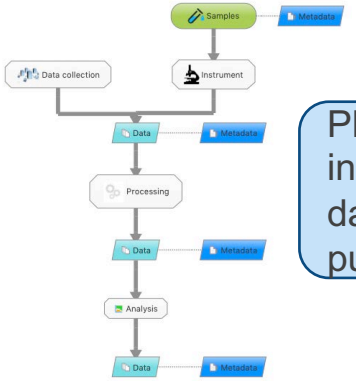




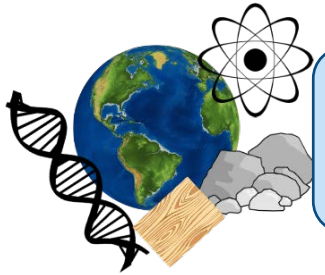
The ETH Scientific IT Services data management solution for research groups

openBIS facts

- Developed at ETHZ since 2007
- Open source software distributed under Apache v2.0 license



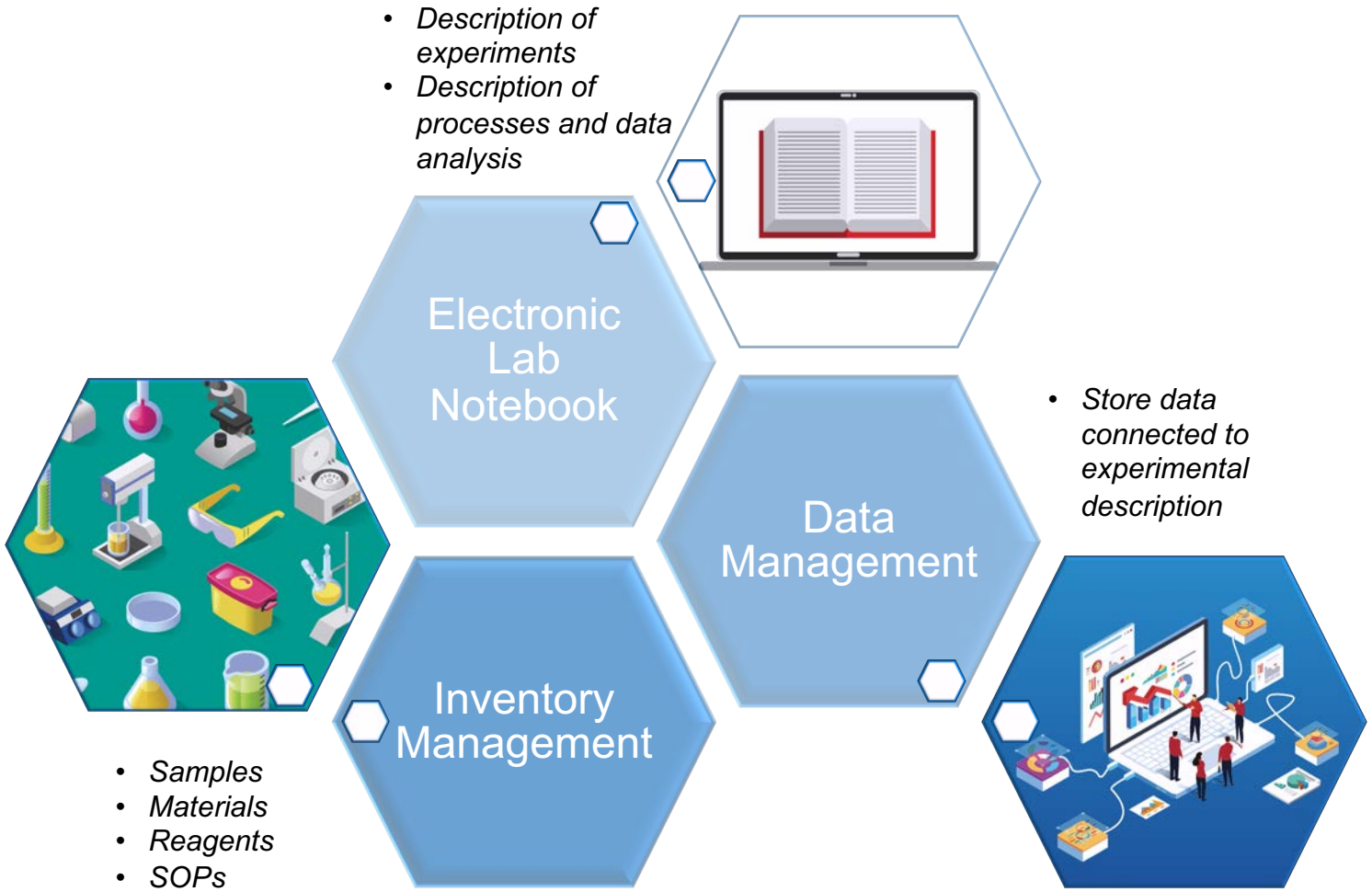
Platform for managing scientific information and supporting research data workflows from “bench” to publication



Can be used in most quantitative science fields (*e.g. life sciences, physics, env. sciences, material sciences, etc*)

Used by research groups and facilities @ ETHZ, Swiss & European Universities, a few companies

openBIS: a complete solution for FAIR data management

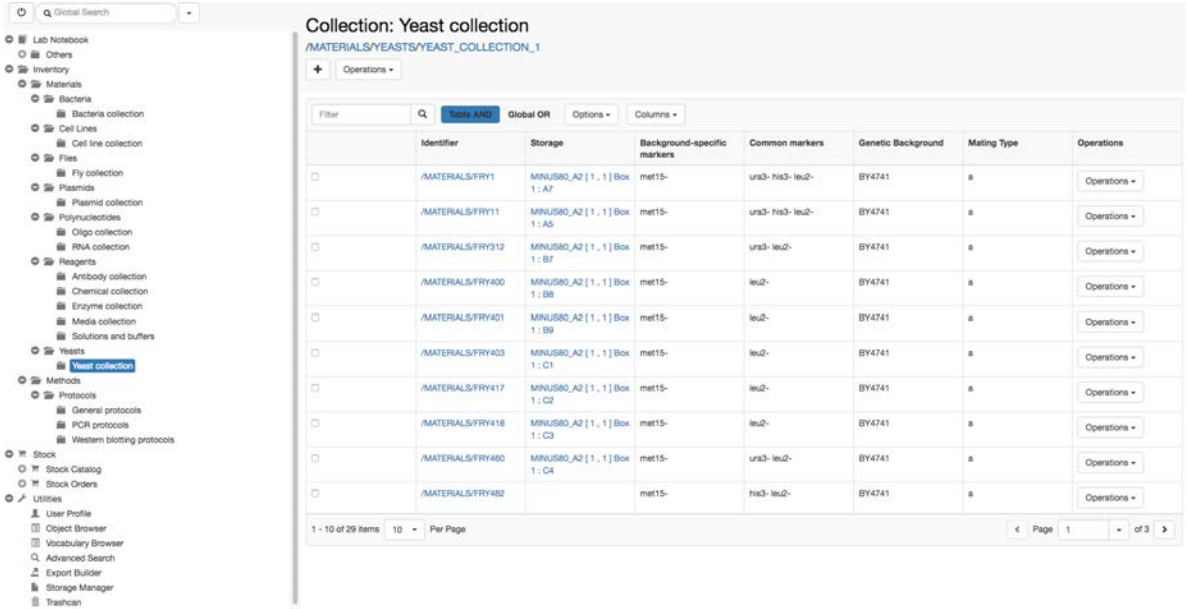


https://openbis.ch

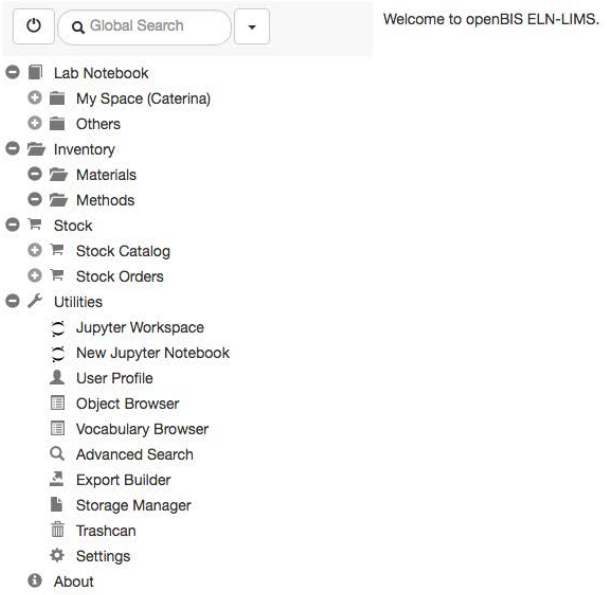
openBIS solutions

openBIS ELN-LIMS is available in two flavors:

For life sciences: customizable predefined types and fields suitable for most biological labs



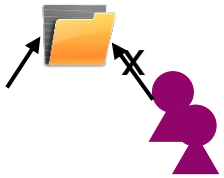
Generic: only basic generic types predefined. To be fully customized by users.



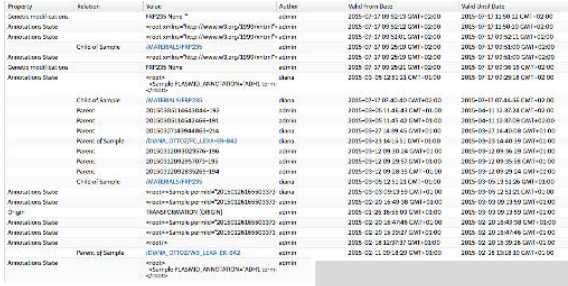
openBIS features

Relationships

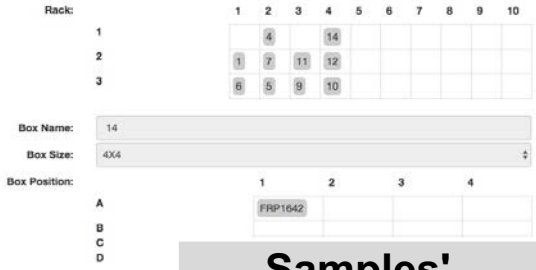
Import/Export



User rights management



Audit trail



Samples' storage manager

Barcode reader

zenodo

ETH zürich Research Collection

Integration with Data Repositories

Data immutability

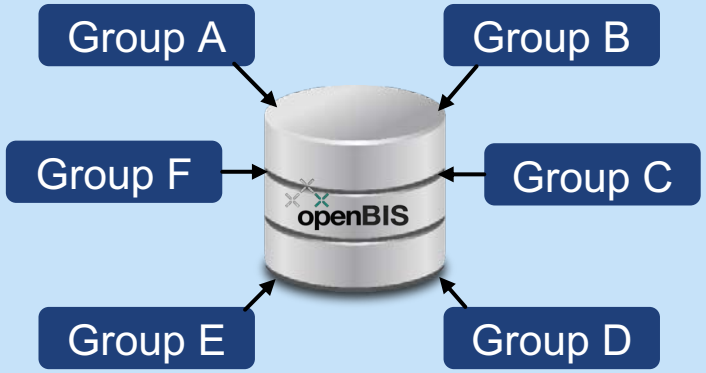
Archive to tape (ETHZ)

https://openbis.ch

RDM services at ETHZ

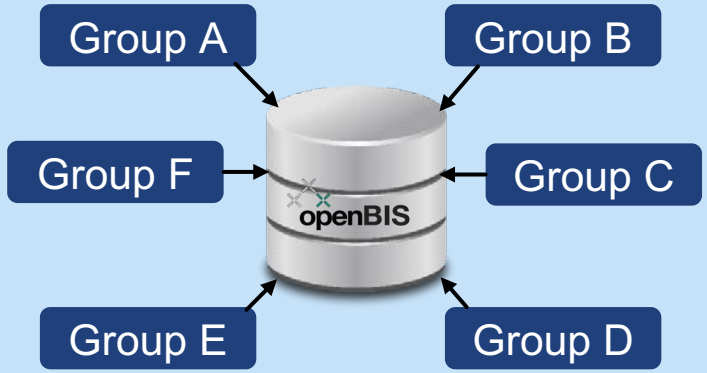
https://ethz.ch/services/en/it-services/catalogue/software-business-applications/research-data-management.html

Research Data Hub



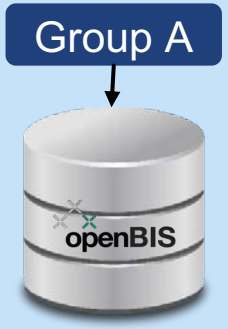
- Centrally managed
- Shared resource
- Limited Customization
- Only storage costs

Departmental Data Hub



- Centrally managed
- Shared resource
- Department customization
- Service fees + infrastructure costs

Research Data Node



- Managed by single group
- Dedicated resource
- Individual group customization
- Service fees + infrastructure costs

+Training & consulting

ETH Research Data Hub (ETH RDH)

- A **central openBIS instance**, ETH RDH, is available to **ETH research groups**
- Only **storage costs** have to be covered by research groups:
  - **First 100GB free**
  - Up to 1 TB: reduced rate
  - LTS (i.e. tapes): free
- ETH RDH is suitable for use by groups that:
  - Do not require much customization
  - Do not need to store sensitive data (e.g. patient data)
  - Do not have a high data volume (<50.000 objects/datasets)

ETH Research Data Hub (ETH RDH)

- Access can be requested via the **IT shop (**https://itshop.ethz.ch)
  - Request must be approved by a fund owner (usually PI).
  - An admin must be nominated in the lab. The admin will be able to do some minimal customization.
- Trainings are provided by SIS throughout the year

ETH Departmental Data Hub

- Currently some institutes/departments have their own openBIS instance, available to groups of the institute/department
- Interested institutes/departments can contact our helpdesk sis.helpdesk@ethz.ch for consulting
- Similar in functionality to ETH RDH, but subject to service fees

ETH Research Data Nodes

- Private instances can be requested by email to sis.helpdesk@ethz.ch
- Additional services available for private group instances (on demand)
  - Database customization
  - Migration of existing databases (subject to evaluation by developers)
  - Instrument integration for direct data upload
- Additional JupyterHub server
- Service is charged for groups in departments with no SIS subscription that cover these costs

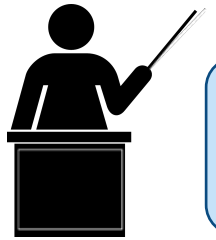
A national RDM service for the academic community

openRDM.swiss

- Service establishment funded by a swissunsiversities P5 project

Cloud-hosted openBIS
- Virtual servers per research group, institute or institution
- Optionally with JupyterHub server for analytics

Self-hosted openBIS
Support for set up on local IT infrastructure



Training & 'best effort' user support

Empa
University of Zurich
zh aw
UNIVERSITÄT BERN

Current customers

https://openrdm.swiss/

A European RDM service for the academic community

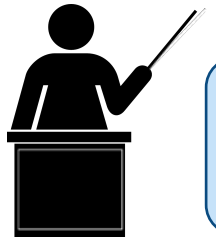
openRDM.eu

- Project funded by EGI-ACE, in the Horizon 2020 research and innovation program framework

SWITCHengines

Cloud-hosted openBIS
- Virtual servers per research group
- Optionally with JupyterHub server for analytics

Self-hosted openBIS
Support for set up on local IT infrastructure



Training & 'best effort' user support

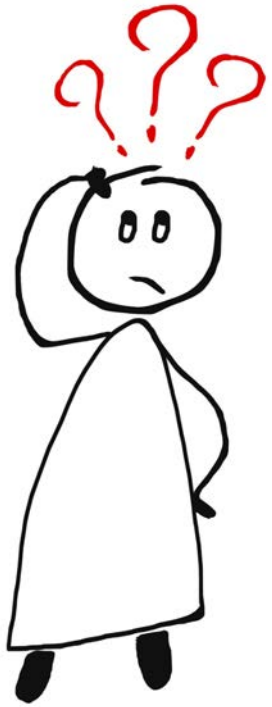
BAM Bundesanstalt für Materialforschung und -prüfung

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

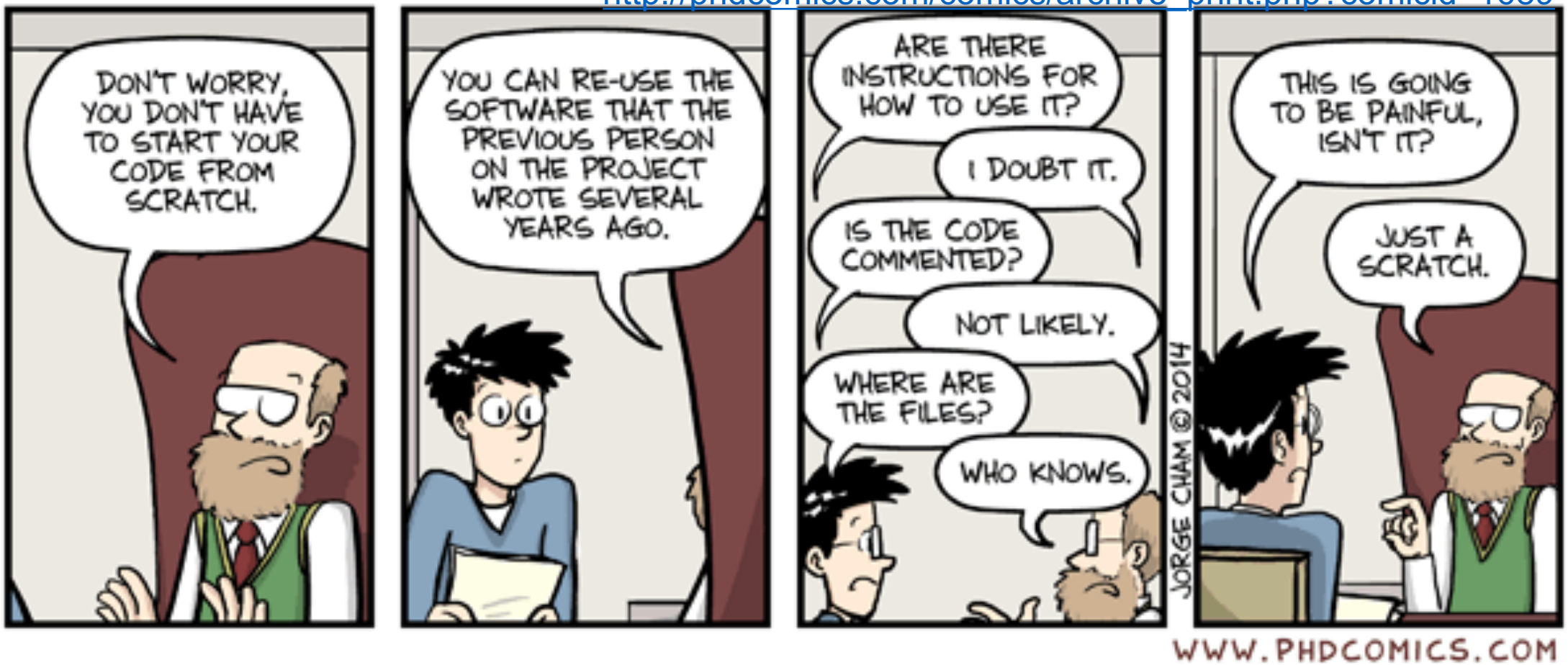
Current customers

https://openrdm-eu/

Questions on openBIS and RDM services provided by SIS?



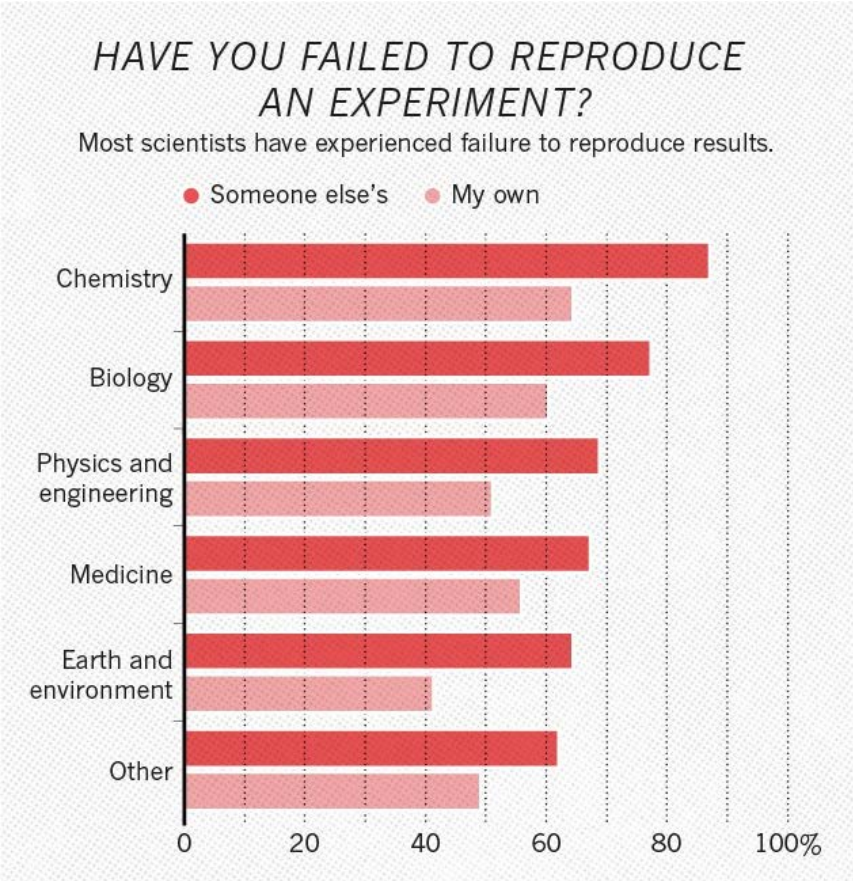
http://phdcomics.com/comics/archive_print.php?comicid=1689

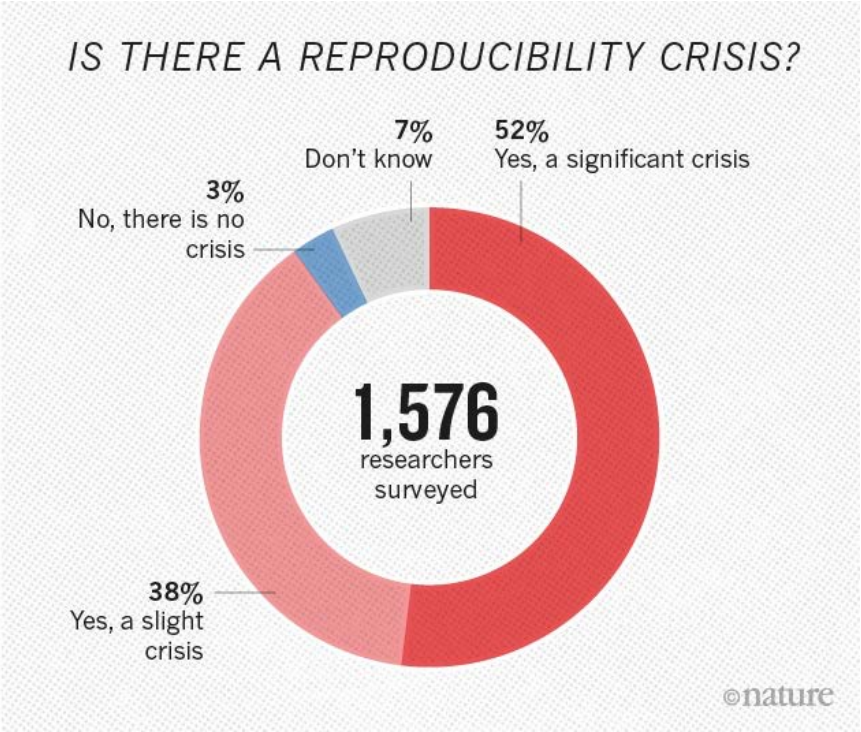


Reproducible Data Analysis

Why talk about Reproducibility?

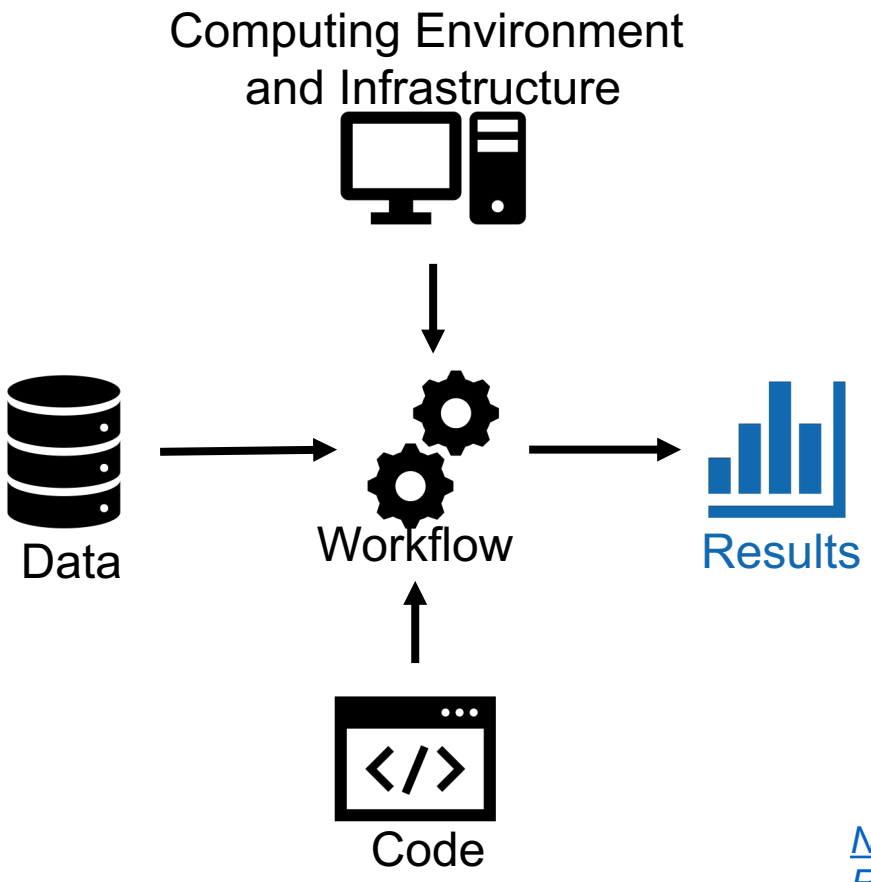
Nature survey on reproducibility across all scientific domains





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

What do we mean by Reproducibility?

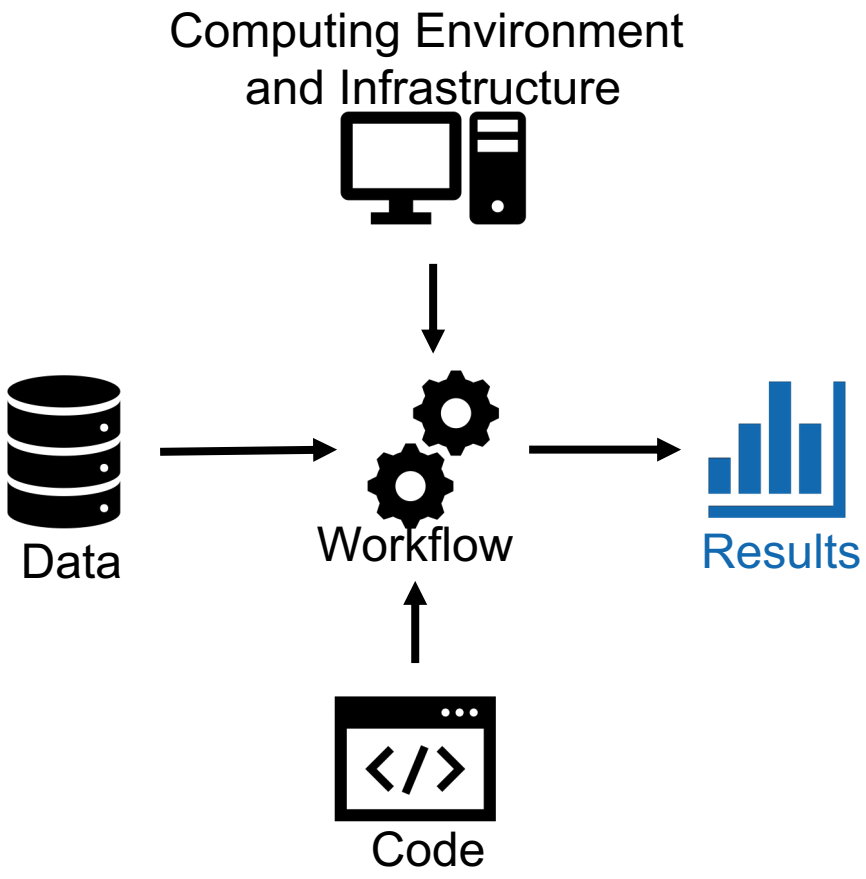


- different meanings across different domains

- **Recent effort of standardization:**
  - « ***Reproducibility*** is **obtaining consistent results** using the same input data; computational steps, methods, and code; and conditions of analysis. This definition is synonymous with "computational reproducibility"… »

National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. https://doi.org/10.17226/25303.

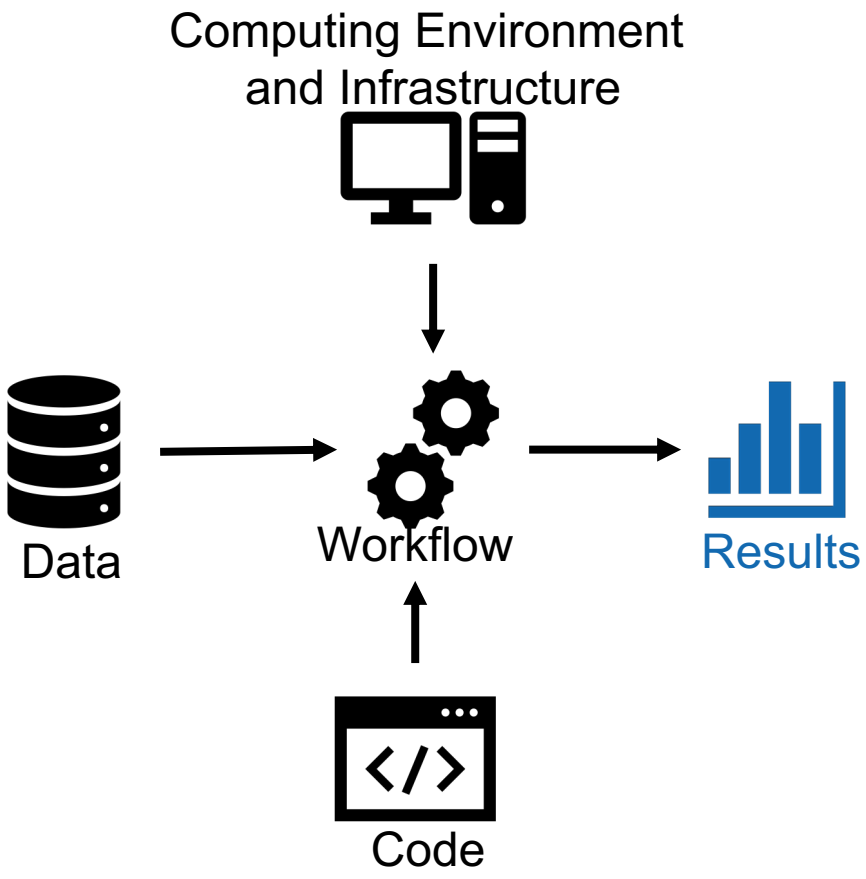
Components



All components have to be reproducible!

1. Code (your code)
2. Data
3. Computing environment
4. Infrastructure
5. Workflow (how to glue everything together)

Components



All components have to be reproducible!

1. **Code (your code)**
2. Data
3. Computing environment
4. Infrastructure
5. Workflow (how to glue everything together)

Code Management

- Proper code management is essential to ensure **reproducible results**
- Many journals require **code availability** after publication and **during review** (see *Nature* 555, 142)

Code Management
Version Control Systems: git

- Software tools specialized on managing and documenting changes to source code over time
- Used for managing large code bases
- They are the standard in professional software development
- Tools: **git**

- Git-Platforms for Collaborations:
  - GitLab:
    - https://about.gitlab.com/
    - https://gitlab.ethz.ch/
  - GitHub:
    - https://github.com/

GitLab

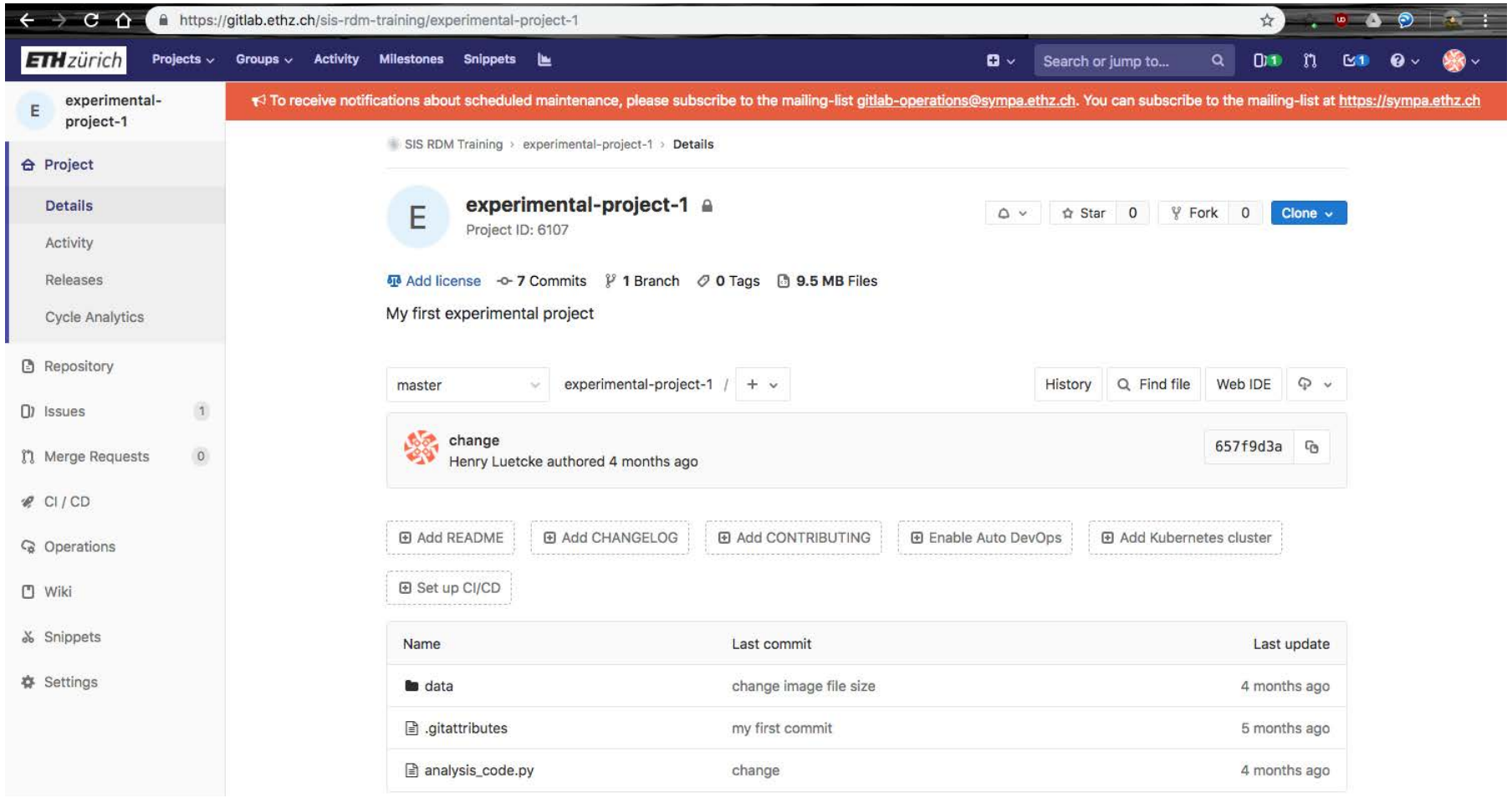
https://gitlab.ethz.ch

GitHub

https://github.com

Code Management

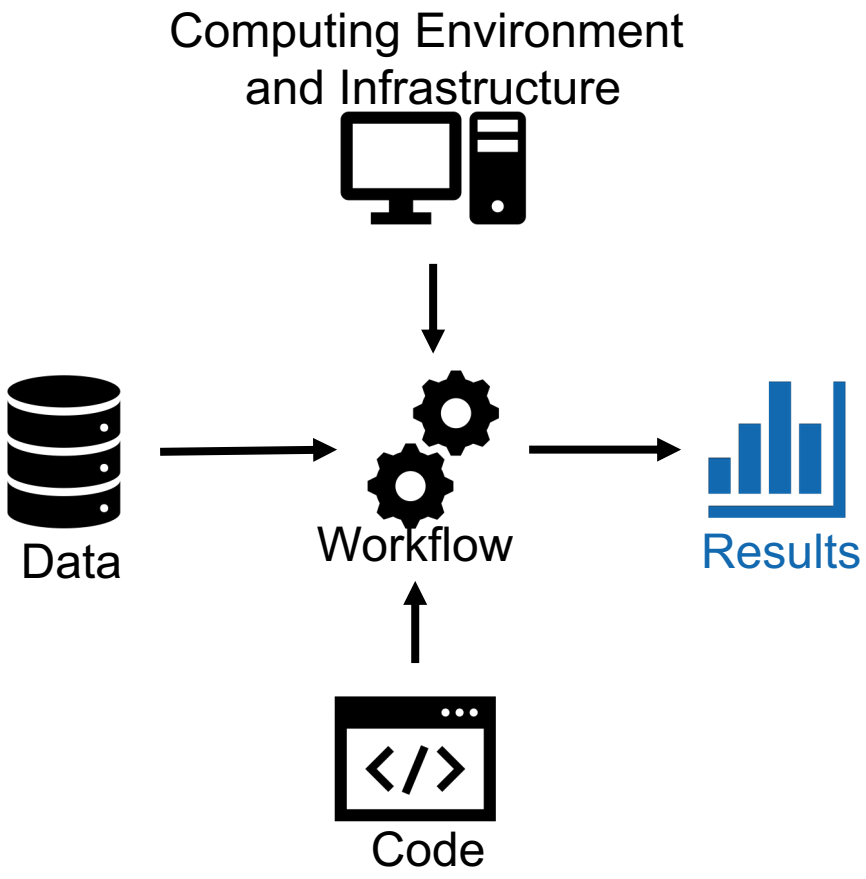
GitLab @ ETHZ – https://gitlab.ethz.ch



Advantages of using the ETH Gitlab Service

- Integrated file, task and documentation management for individuals and / or groups
- Built-in light-weight Wiki (protocols, list of materials etc.)
- Keep track of version history for everything
- Free for small repositories (< 2GB), otherwise yearly price of 250 CHF / TB
- Local and remote copies (off-site backup)
- Easily change permissions from private to public (e.g. after publication)
- Data can be exported (e.g. to Github)
- Container registry

Components



All components have to be reproducible!

1. Code (your code)
2. Data
3. Computing environment
4. Infrastructure
5. **Workflow (how to glue everything together)**

Interactive Notebooks

- Applications that combine documentation, code, input and output generated by the code, e.g. graphs, plots (*Nature 515, 151–152*)
- Useful for exploratory data analysis and reproducibility

RStudio

- Open source + commercial edition
- Integrated development environment for R

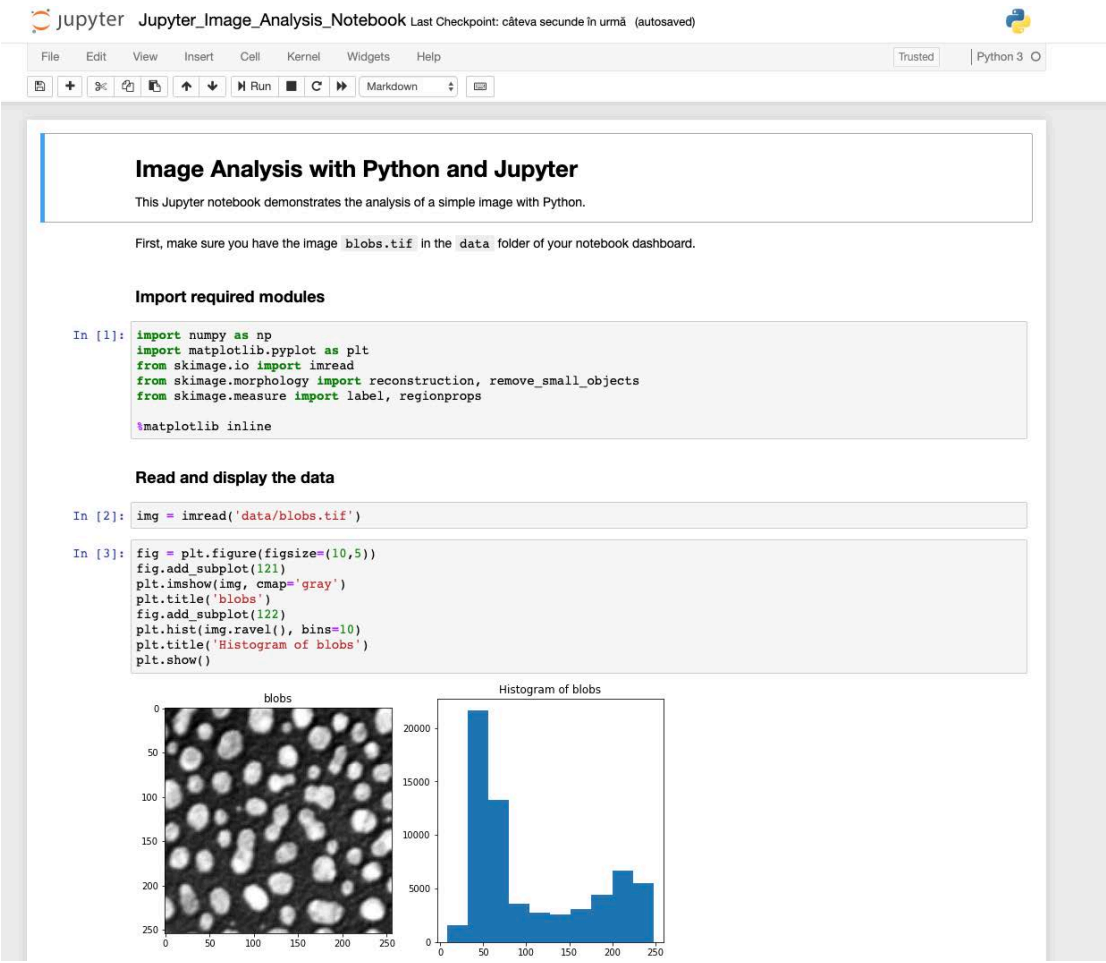
jupyter

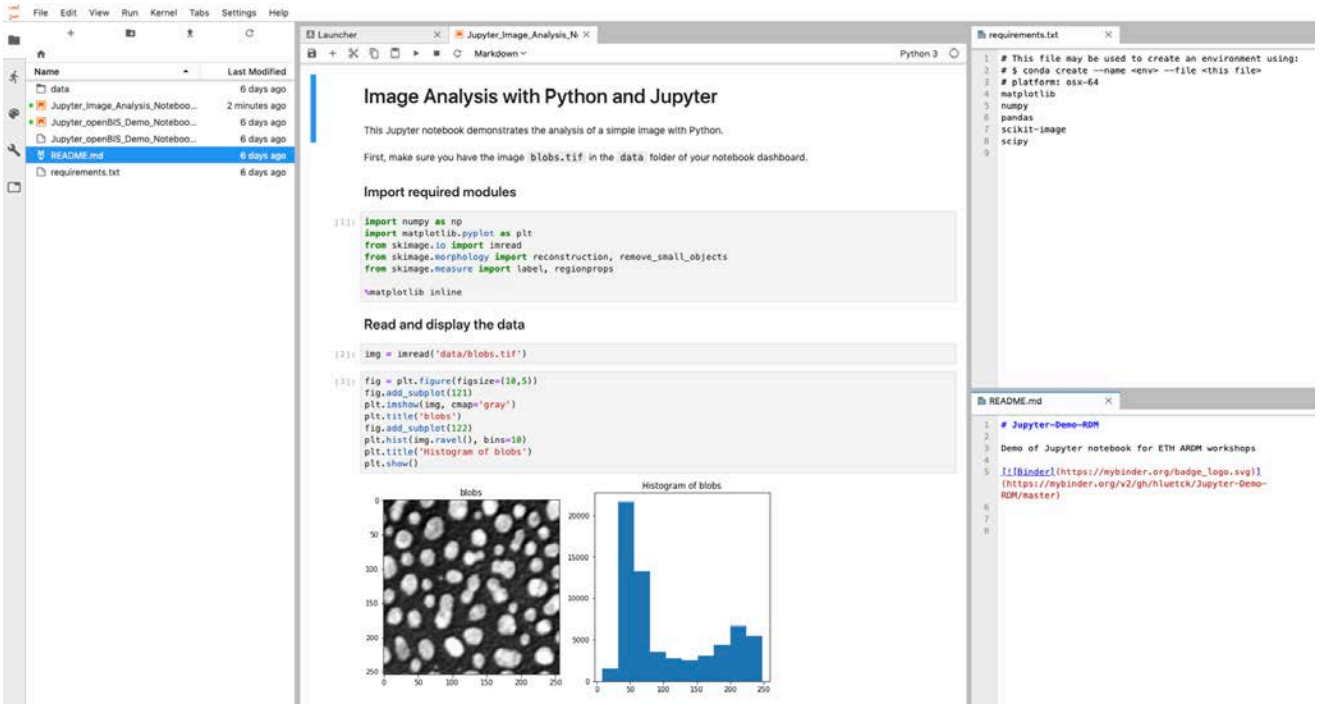
- Open source
- > 40 languages supported (Python, R, Julia, Matlab, IDL, etc.)

MATLAB

- Commercial
- Used in scientific, engineering, mathematical fields

Jupyter notebooks / JupyterLab





Jupyter notebooks / JupyterLab

- **Jupyter notebook:** web-based interactive computational environment
- **JupyterLab:** next-generation for Jupyter notebooks (and more)
- Dozens of programming languages supported (core: Julia, Python, R)
- Notebook export in various formats (HTML, PDF, Python …)
- Publication of interactive notebooks on mybinder.org
- Integration with ETH scientific computing infrastructure (see https://gitlab.ethz.ch/sfux/Jupyter-on-Euler-or-Leonhard-Open)
- **JupyterHub:** multi-user version of the notebook for research labs

Interactive Notebooks – what can go wrong?

- **Versioning**
    - Version control of even moderately complex NBs is challenging
    - Tracking NB history is harder than for traditional source code
    - Some tools may help (e.g. *nbdime*)

```
$ diff a.ipynb b.ipynb
76,77d75
<     "plt.rc('axes', grid=False)\n",
<     "plt.rc('axes', facecolor='white')\n",
90c88
<       "image/png": "iVBORw0KGgoAAAANSUhEUgAABLkAAAMQCAYAAADLj7dlAAAABHNCSVQICAgIfAhki
AAAAAlwSFlz\nAAAWJQAAFiUBSVIk8AAAIABJREFUeJzsvXeYZFd57b12h0maPNJII2lGOaCAkEBCFgozIxkBAp
lY\n1waDyDZg8MX+zMU2F4Mx1x8PwWAwxmBjg4yNi2BfQMa20iiAQFkIjXKWRtJIE3tSz3TXuX+8vV2n\nqyucv
N+9z/o9zzynprvq1D6nqqtqr1prbRNFEQghhBBCCCGEEEII8Zkh1wMghBBCCCGEEEIIISQv\nFLkIIYQQQgghhB
BCiPdQ5CKEEEIIIYQQQggh3kORixBCCCGEEEIIIYR4D0UuQgghhBBCCCGEEOI9\nFLkIIYQQQgghhBBCiPdQ5CK
EEEIIIYQQQggh3kORixBCCCGEEEIIIYR4D0UuQgghhBBCCCGEEOI9\nFLkIIYQQQgghhBBCiPdQ5CKEEEIIIYQQ
Qggh3kORixBCCCGEEEIIIYR4D0UuQgghhBBCCCGEEOI9\nFLkIIYQQQjzEGHOJMaZljPmo67EkZWq8D7keByGEE
ELChCIXIYQQQirDGPOmKaFj3BhzkMNx/H/G\nmG3GmP/pagwFEbkeQJUYY75gjNlijHmD67EQQgghRB8UuQghhB
BSJe+DCDMjAH7L4Tje4AmA+gLc5\nHEMRGNcDqJi3AVgI4DddD4DDD4QQQggh+qDIRQghhJBKMMacCuBMAFsg4sy7jTH
DjobzZwBuBvBxR/dP\nsvERADcC+LTrgRBCCCFEHxS5CCGEEFIVH4C4uP4SIlQcBOD1LgYSRVEzivjIXR1H0fRf3
```

Chattopadhyay et al. *(2020). What's Wrong with Computational Notebooks?*
doi:10.1145/3313831.3376729

Interactive Notebooks – what can go wrong?

- **Versioning**
  - Version control of even moderately complex NBs is challenging
  - Tracking NB history is harder than for traditional source code
  - Some tools may help (e.g. *nbdime*)
- **Scalability**
  - Scaling to large datasets is challenging (due to browser limitations)
- **Reproducibility**
  - Interactive working mode can result in hard-to-reproduce notebooks
  - Discipline is needed! Regular pruning & refactoring; *"Restart kernel & Run all"* is your friend
- **Collaboration**
  - Collaborative editing not fully supported
- **Security**
  - Data confidentiality & access controls may be problematic

Chattopadhyay et al. *(2020). What's Wrong with Computational Notebooks?* doi:10.1145/3313831.3376729

A Zoo of Workflow Management Systems

- An incomplete list of **286** Computational Data Analysis Workflow Systems
  - https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems
- A curated list of **109** Awesome Pipeline frameworks & libraries + **30** Workflow platforms
  - https://github.com/pditommaso/awesome-pipeline
- Some examples:

nextflow

Snakemake

Apache Taverna

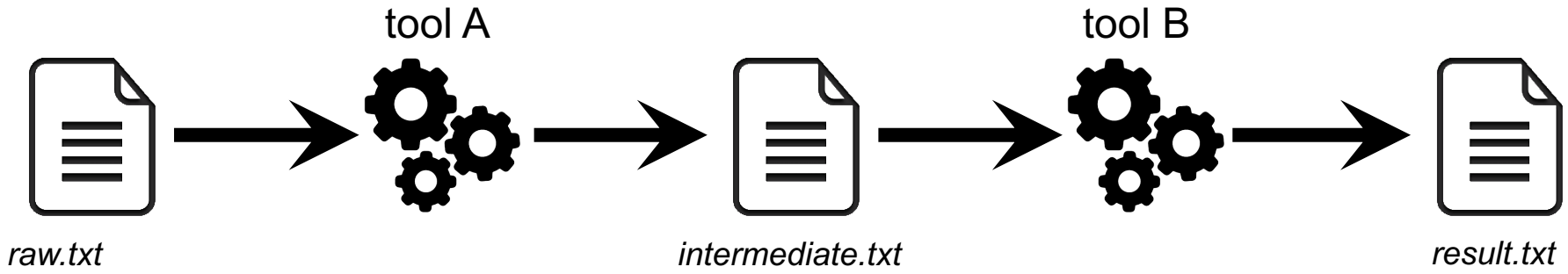
Galaxy PROJECT

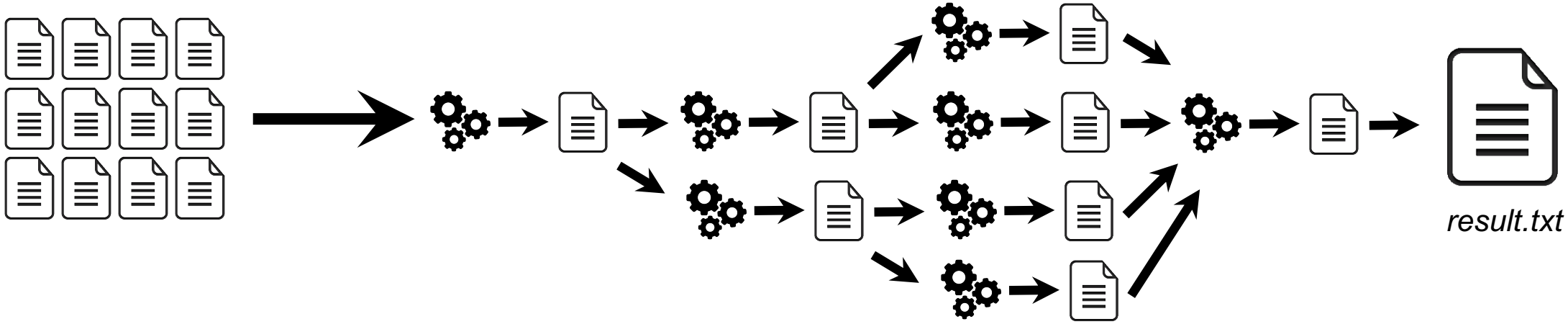
Open for Innovation KNIME

COMMON WORKFLOW LANGUAGE

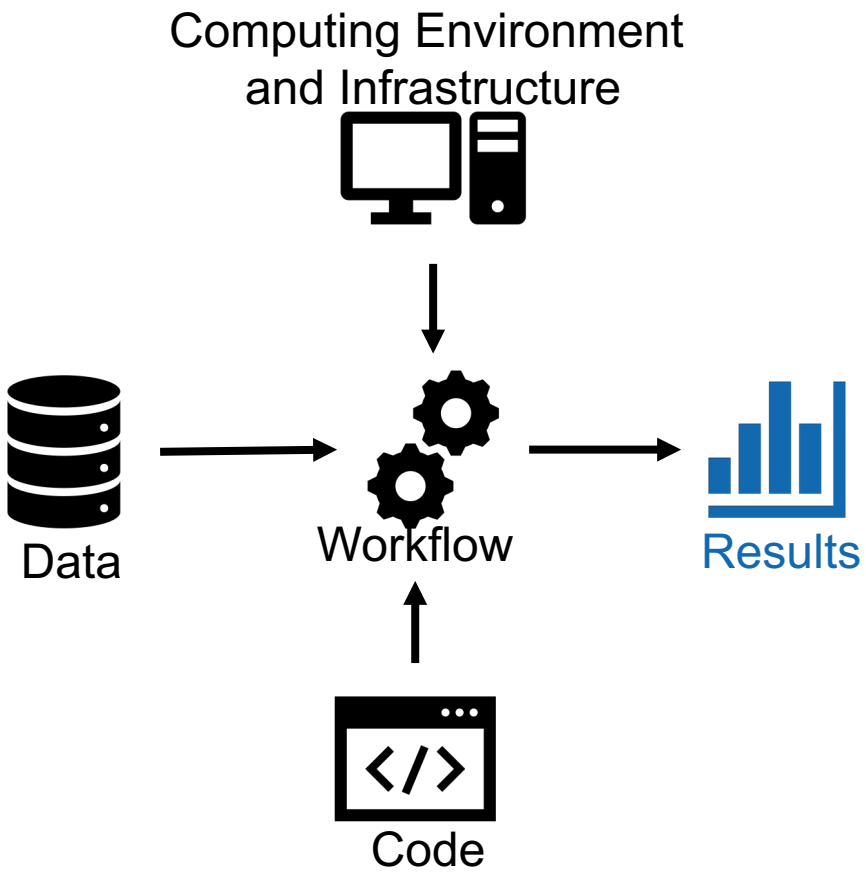
Airflow

Reproducible Workflow
Automate and Scale





Components



All components have to be reproducible!

1. Code (your code)
2. Data
3. **Computing environment**
4. Infrastructure
5. Workflow (how to glue everything together)

Reproducible Environment

Problem:
Full reproducibility requires the possibility to recreate the system that was originally used to generate the results.

Solution:

Bundle your application and all dependencies = Environment Isolation + Dependency Management

- Environment and Package Management

Tools:

- Virtual Machine (VM): VirtualBox, Vmware
- Container – lightweight VM: Docker, Singularity
- Application/Software only:
  - Python: venv, virtualenv, pip, Conda
  - R: renv, Conda

Virtual machines (VMs)

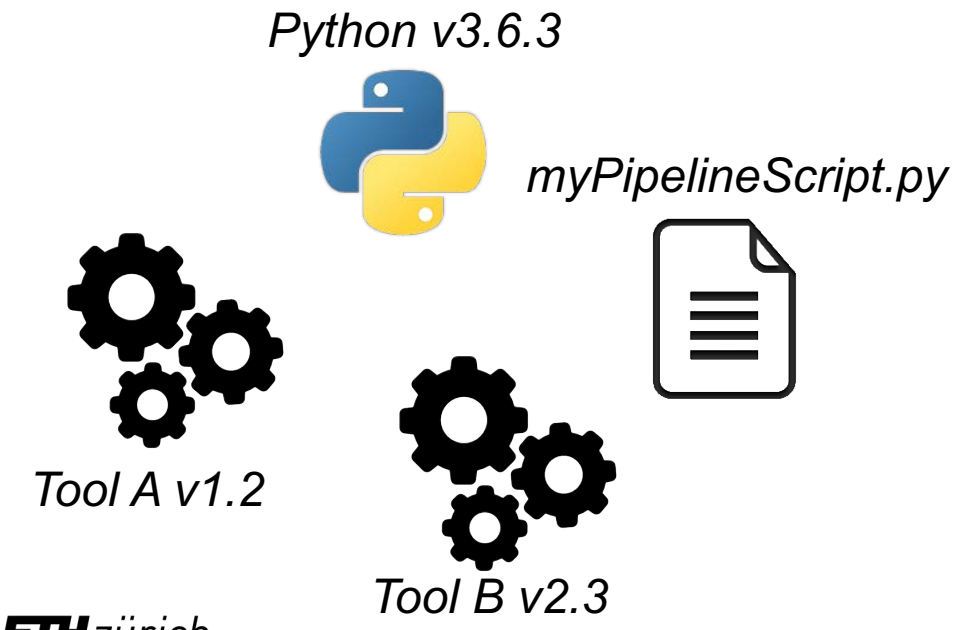
- A virtual machine (VM) is an operating system (“guest”) that runs inside another computing environment (“host”).

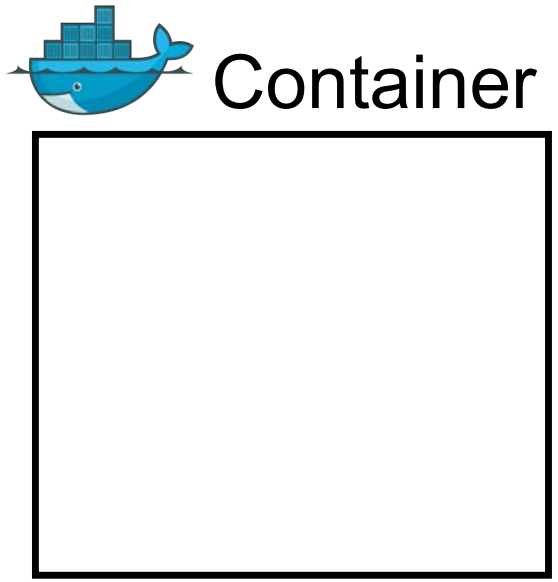
- **Advantages**:
  - Allows multiple OS environments on a single physical computer
  - VMs are widely available and are easy to manage, maintain and distribute
  - Offers application provisioning and disaster recovery options
- **Drawbacks**:
  - They are not as efficient as a physical computer because the hardware resources are distributed in an indirect way.
  - Multiple VMs running on a single physical machine can deliver unstable performance

https://searchservervirtualization.techtarget.com/definition/virtual-machine

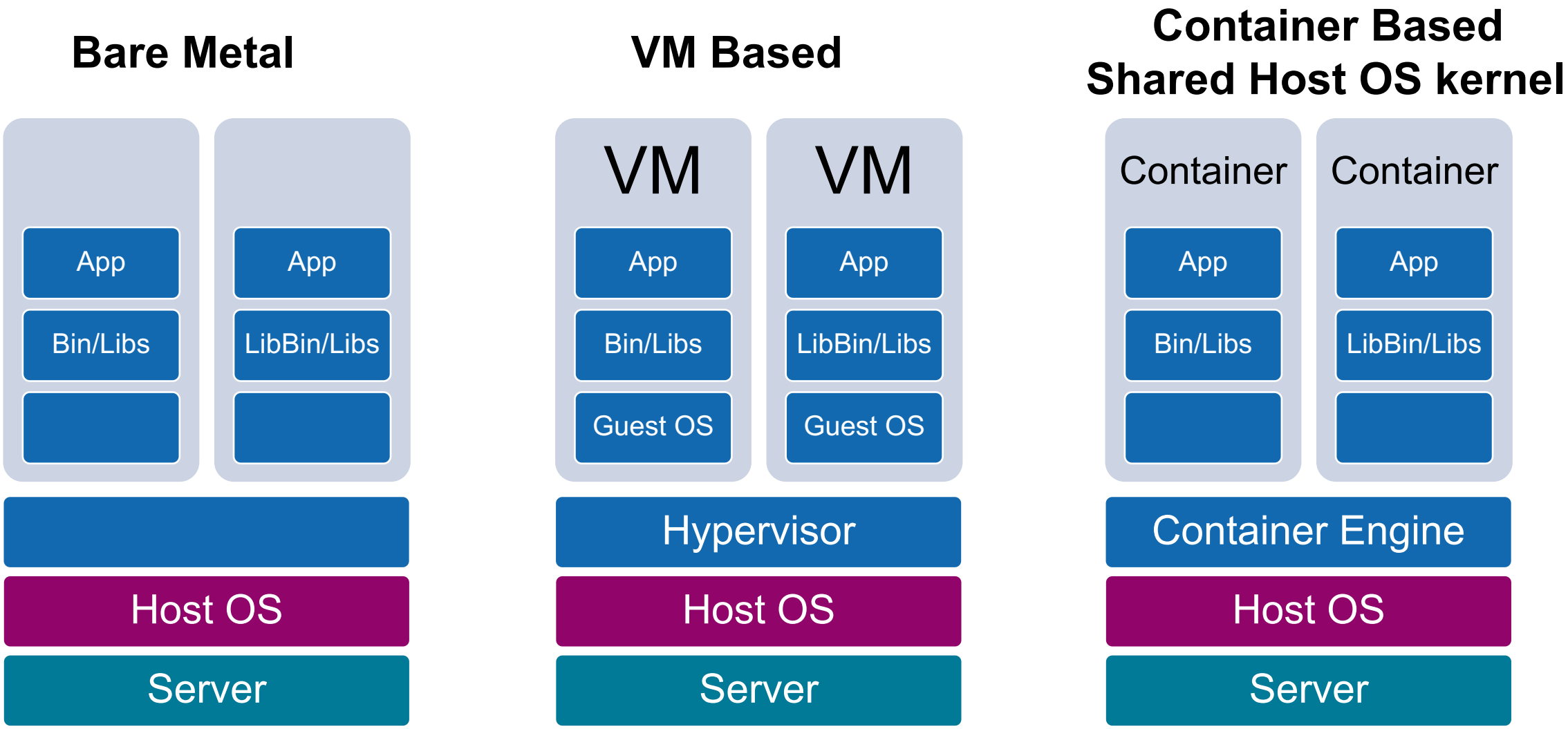
Containers

- Container: Operating system level **virtualization met**hod for running software without launching an entire virtual machine
- In simpler words: containers allow you to **package** your software / pipeline with the **dependencies** inside a **reproducible**, easy to **share**, runnable tool
- Tools: Docker





Bare Metal, Virtual Machine (VM) and Container (Docker)



Virtual Machines vs Containers

	VMs (Virtual Box)	Containers (Docker)
Use case	Complex Apps (GUI, …)	Data Analysis Scripts, Simple Apps, Microservices, Continuous Integration
Virtualisation	Hardware-level	OS-level
Size	GB	MB
Startup time	Minutes	Seconds
Guest OS	Windows, macOS, Linux	Primarily Linux-based
Host OS	Windows, macOS, Linux	Linux, Windows 10 and macOS with a hypervisor
Overhead (RAM, CPU)	High - reduced performance	Low - close to native performance
Security	Better (fully isolated)	Poorer (shared kernel)
How to use	Easy if you know to install OS	New things to learn

Containers - What can go wrong?

- The base image is updated - same tag different content: e.g. centos:latest
- The base image is deleted by the owner
- The image is not compatible with your machine (GPUs, High Performance Computing, ARM)
- You don’t have privileged permission (High Performance Computing)

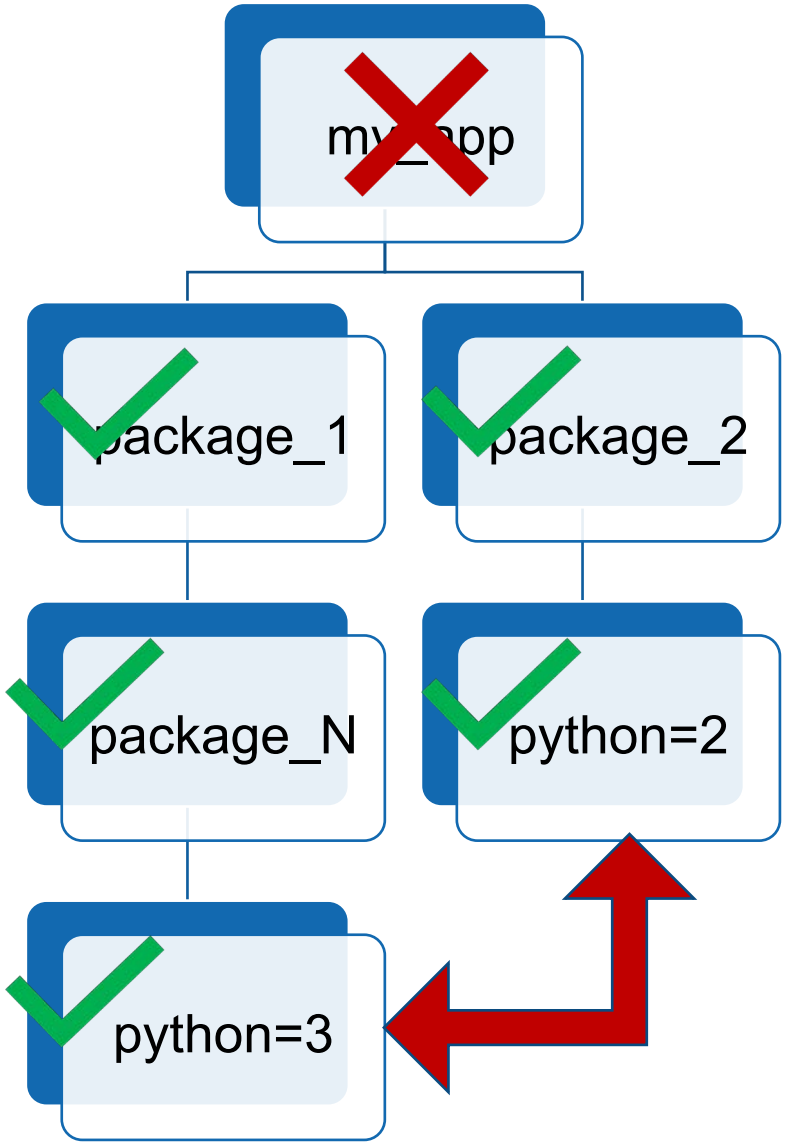
Reproducible Environment for R and Python

CONDA

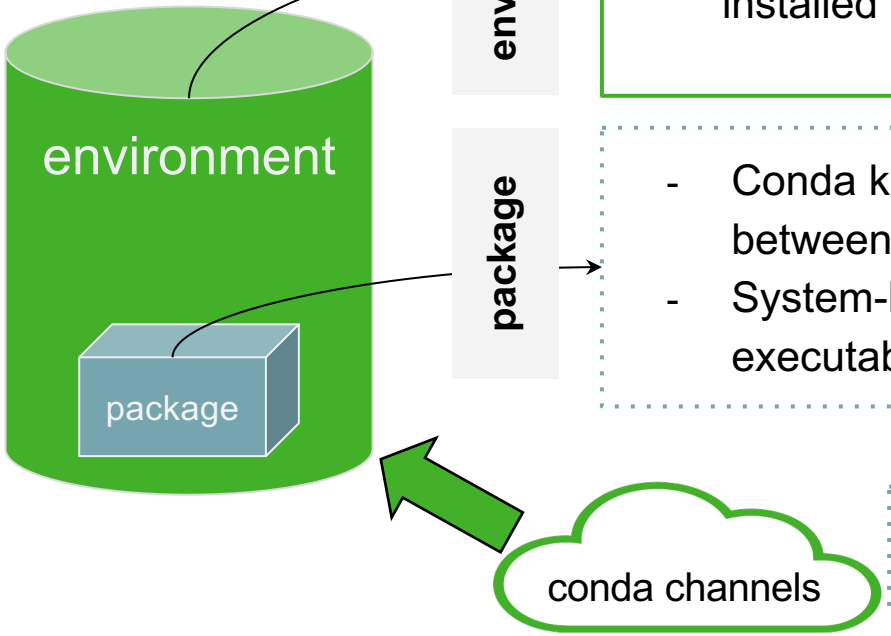
- Open source: Anaconda and Miniconda
- Commercial support: Anaconda Enterprise
- Multi-platform:
  - Windows, macOS, Linux

- Environment Management System

- Package Management System:
  - Supported programming Languages: Python, R, …
  - System libraries shipped in binary format
  - Dependency resolver (top level + first wins in pip<20.3)



Conda in a Nutshell



environment

- Conda environments are isolated
- Directory containing specific collection of installed Conda packages

package

- Conda keeps track of the dependencies between packages
- System-level libraries, Python or other modules, executable programs, other components

- Remote directories with packages

environment.yml

```
channels:
- defaults
- conda-forge
dependencies:
- python=3.8
- jupyterlab
```

Conda automatically creates an environment file with packages and dependencies

Slide credit: L. Wigge, R. Ågren & J. Sundh (NBIS & SciLifeLab, Sweden)

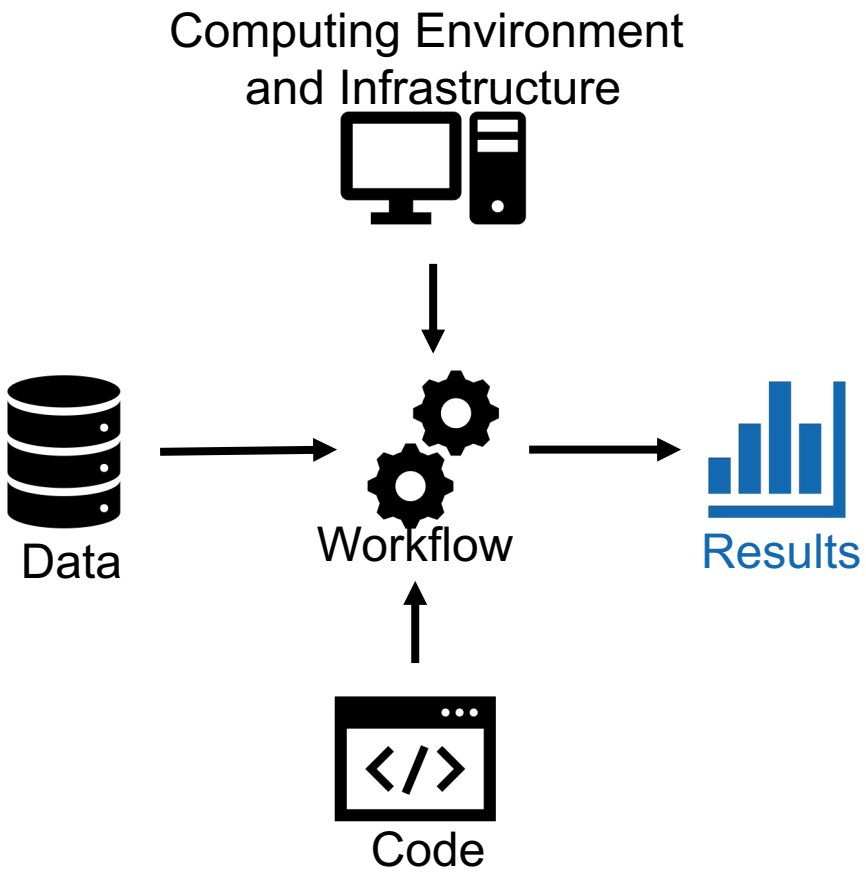
Conda - What can go wrong?

- The package metadata (dependency list) is update (not very likely)
- The package is deleted by the owner
- Python: you mix pip and conda and later do a conda update or conda install
- The package is not available under another platform
- There is no conda package for what are you looking for

Environment and Package Management Systems

Programming Language	Environment Management System	Package Management System	Comments
Python 2 (not supported)	virtualenv, conda	pip, conda	
Python 3	venv, virtualenv, pipenv poetry, conda	pip, pipenv, poetry, conda	only conda can install different Python versions (pyenv can be used)
R	packrat (soft-deprecated), renv, conda	packrat (soft-deprecated), renv, conda	only conda can install different R versions
Julia	Pkg, conda	Pkg, conda	conda provides outdated Julia versions
Matlab	N/A	Add-on manager, Matlab Package Manager (unofficial)	Matlab's search path determines dependencies

Components



All components have to be reproducible!

1. Code (your code)
2. Data
3. Computing environment
4. Infrastructure
5. Workflow (how to glue everything together)

Reproducible Platforms

binder (beta)

Turn a GitHub repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Build and launch a repository

GitHub repository name or URL

Git branch, tag, or commit

Path to a notebook file (optional)

Copy the URL below and share your Binder with others:

Fill in the fields to see a URL for sharing your Binder.

Copy the text below, then paste into your README to show a binder badge: launch binder

renku

Renku Projects

New project

Your Projects Starred Projects All Projects

virginiafriedrich/form-refactoring-tests-2-copy

Only description for this project

virginiafriedrich/form-refactoring-test-3

Description for this project.

virginiafriedrich/form-refactoring-tests-2

Only description for this project

virginiafriedrich/test-datasets-in-protected-branch

virginiafriedrich/project-11

CODE OCEAN

Discover & Run Scientific Code

Code Ocean is a cloud-based **computational reproducibility** platform

UPLOAD YOUR CODE

Search keyword, research field, title, author, DOI, etc.

SOCIAL SCIENCES

ENGINEERING

MEDICAL SCIENCES

SOCIAL SCIENCES

On Writing Reproducible and Interactive Papers

Exploratory subgroup analysis of the SEAQUAMAT trial using Random Forests:...

Mapping the Universe of Registered Reports

Continuous Improvements Towards Perfection

WORKING LIFE

By **Lorena A. Barba**

The hard road to reproducibility

Early in my Ph.D. studies, my supervisor assigned me the task of running computer code written by a previous student who was graduated and gone. It was hell. I had to sort through many different versions of the code, saved in folders with a mysterious numbering scheme. There was no documentation and scarcely an explanatory comment in the code itself. It took me at least a year to run the code reliably, and more to get results that reproduced those in my predecessor's thesis. Now that I run my own lab, I make sure that my students don't have to go through that.

In 2012, I wrote a manifesto in which I committed to best practices for reproducibility. Today, a new student arriving in my group finds all of our research code in tidy repositories, where every change is recorded automatically. Version control is our essential technology for record keeping and collaboration. Whenever we publish a paper, we create a "reproducibility package," deposited online, which includes the data sets and all the code that is needed to recreate the analyses and figures. These are the practices that work for us as computational scientists, but the principles behind them apply regardless of discipline.

It takes new students some time to learn how to work to these standards, but we have documentation and training materials to make it as painless as possible. My students don't resent investing their time in this. They know that practices like ours are crucial for the integrity of the scientific endeavor. They also appreciate that our approach will help them show potential future employers that they are careful, conscientious researchers.

I am pleased when our group is recognized for our high standards in other people's writings, and when we are invited to speak about these practices at meetings. But we've

"My students and I continuously discuss and perfect our standards."

additional details were documented in the supplementary materials. It was the very definition of reproducible research.

Three years of work and hundreds of runs with four different codes taught us just how many ways there are to go wrong! Failing to record the version of any piece of software or hardware, overlooking a single parameter, or glossing over a restriction on how to use another researcher's code can lead you astray.

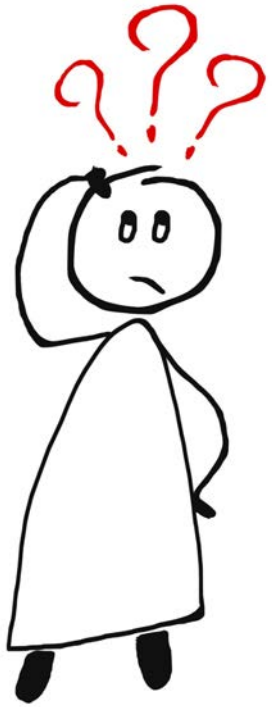
We've found that we can only achieve the necessary level of reliability and transparency by automating every step. Manual actions are replaced by scripts or logged into files. Plots are made only via code, not with a graphical user interface. Every result, including those from failed experiments, is documented. Every step of the way, we want to anticipate what another researcher might need to either reproduce our results (run our code with our data) or replicate them (independently arrive at the same findings).

About 150 years ago, Louis Pasteur demonstrated how experiments can be conducted reproducibly—and the value of doing so. His research had many skeptics at first, but they were persuaded by his claims after they reproduced his

"We've found that we can only achieve the necessary level of reliability and transparency by automating every step."

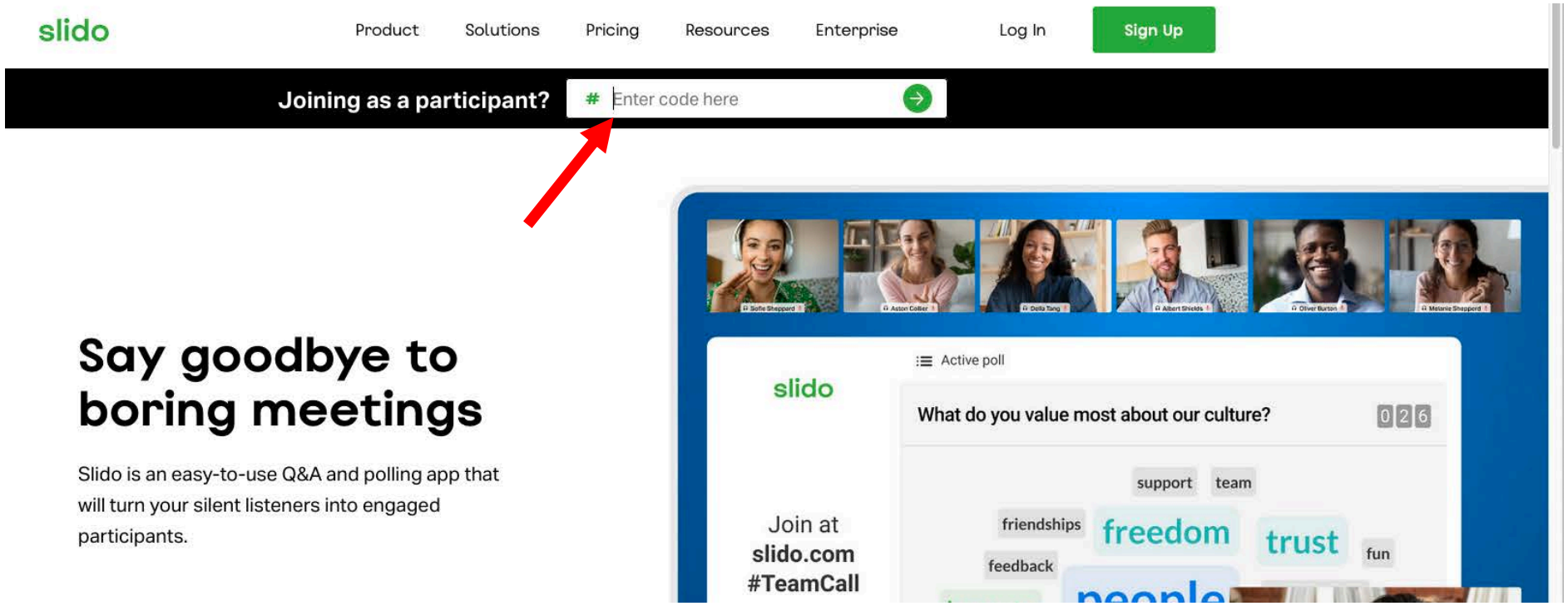
https://doi.org/10.1126/science.354.6308.142

Questions on Reproducible Data Analysis?



Which RDM practices & tools are you considering in the future?

- Go to www.slido.com and enter the event code #ETHRDM



What’s in it for me?

At the start of the project

- Forced to think about scope and limitations.
- Improved structure and organization.

During the project

- Easier to rerun experiments and analysis
- Closer interaction between collaborators.
- Much of the manuscript "writes itself".

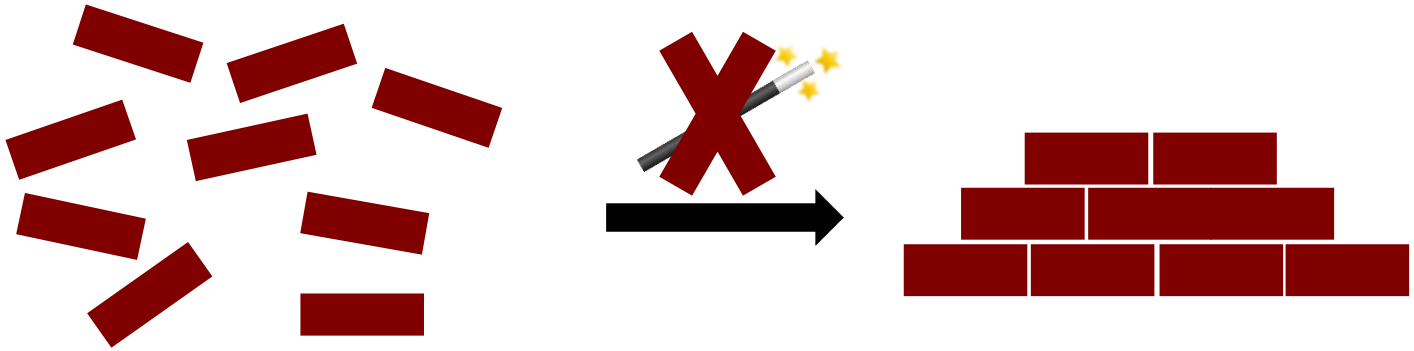
After the end of the project

- Faster resumption of research by others (or your future self), thereby increasing the impact of your work.
- Increased visibility in the scientific community.

Slide credit: L. Wigge, R. Ågren & J. Sundh (NBIS & SciLifeLab, Sweden)

What’s in it for me?

- Aim for improvement, not perfection!
- Try to establish standard procedures with colleagues (if they don’t exist yet)
- RDM requires **WORK** & **TIME**, but the time spent on this is an **investment** for the future!



Contact us for consultations / trainings on: data management, version control, reproducible computational workflows or data science support

sis.helpdesk@ethz.ch

Contacts

Caterina Barillari

caterina.barillari@id.ethz.ch

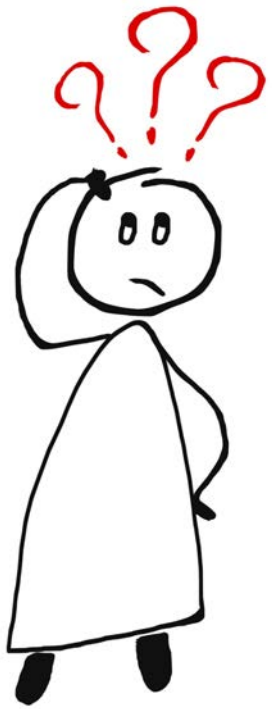
Andrei Plamada

andrei.plamada@id.ethz.ch

sis.helpdesk@ethz.ch
https://sis.id.ethz.ch/

https://twitter.com/ETH_SIS?lang=en

Any final questions on what we have seen today?



Feedback: https://www.umfrageonline.ch/s/a13b937