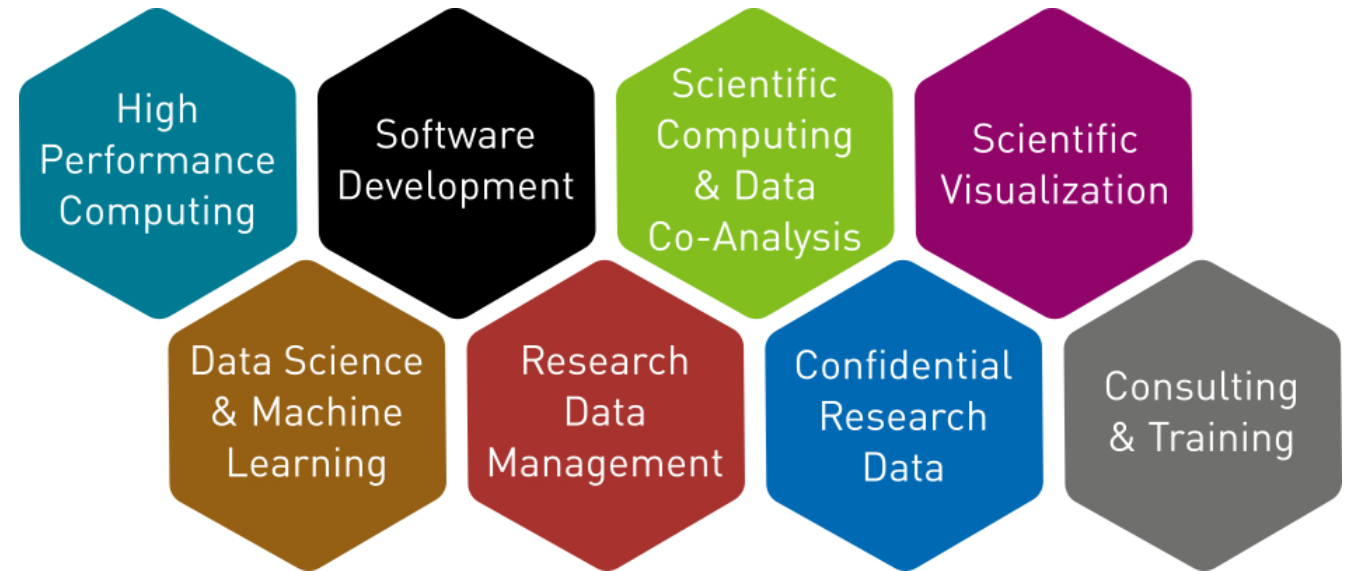


Introduction to Active Research Data Management

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

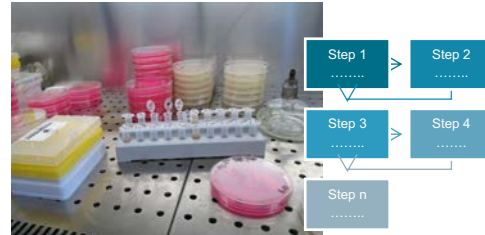
- Your affiliation
- Your research topic



Overview of today's workshop



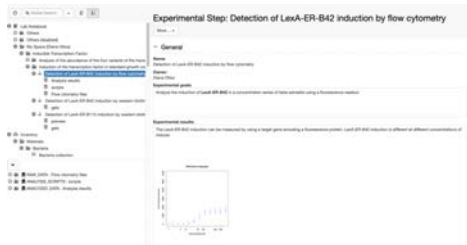
ARDM intro



Sample & protocol management



Management of Data & Metadata



ELNs



SIS's RDM solution



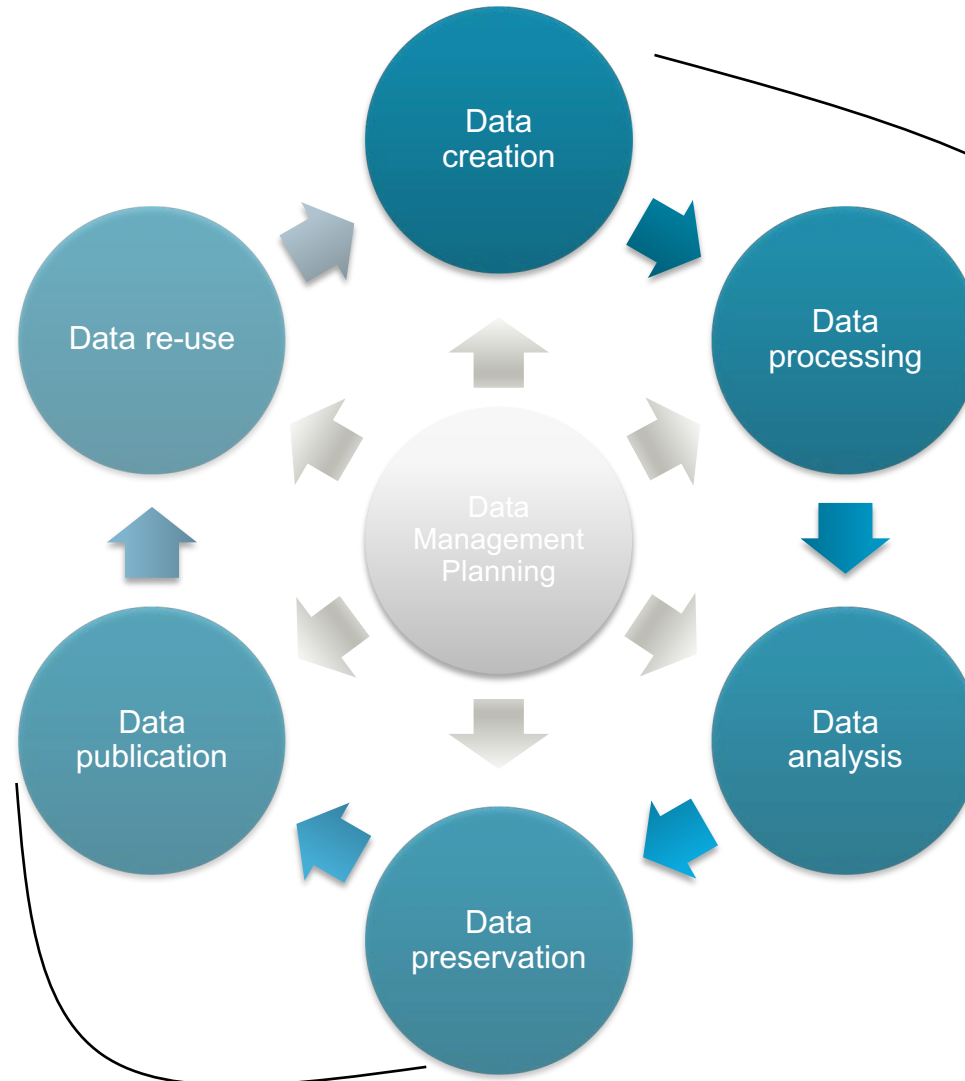
Reproducible data analysis



www.digitalbevaring.dk

Overview of active research data management

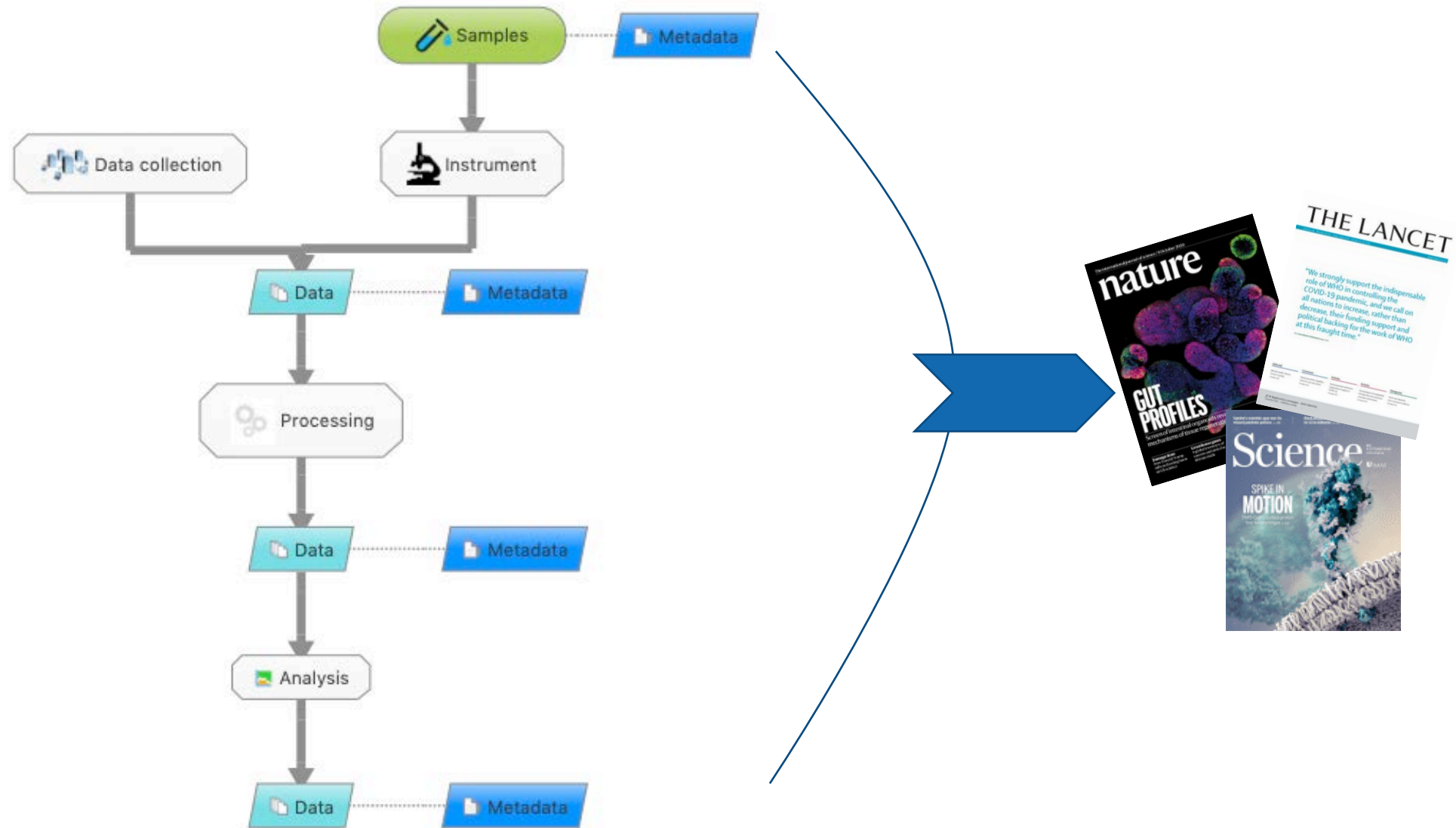
The data life cycle



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

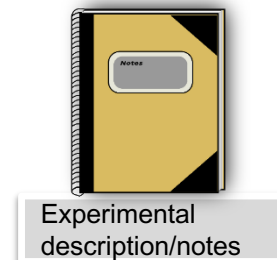
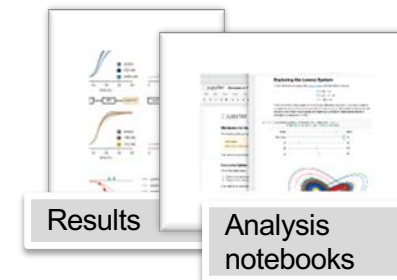
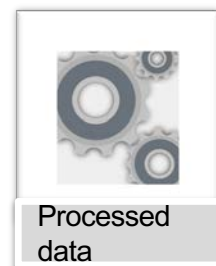
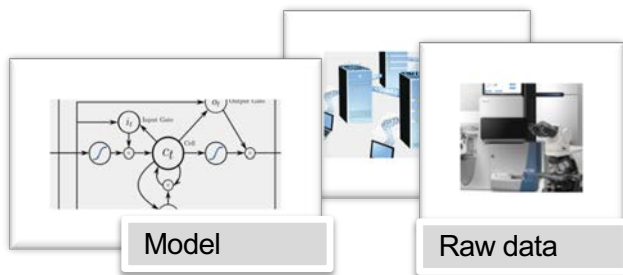
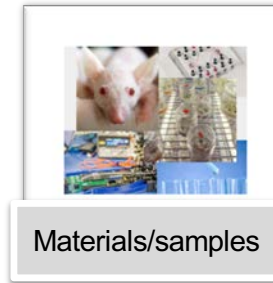
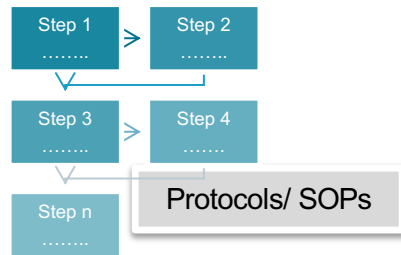
- ❑ Funding agencies and journals increasingly demand that data is published according to the **FAIR**¹ data principles (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)
- ❑ Lots of data are generated during a research project.



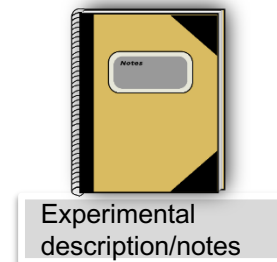
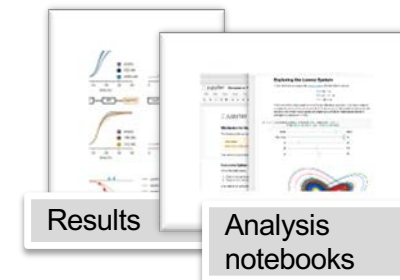
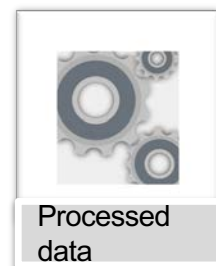
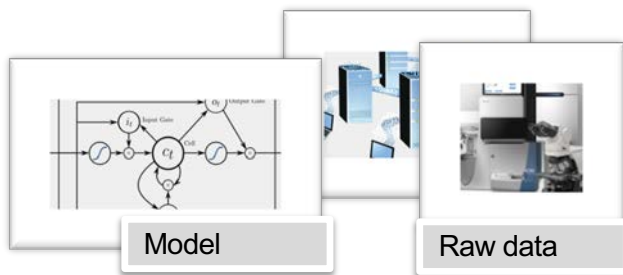
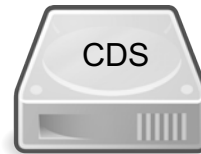
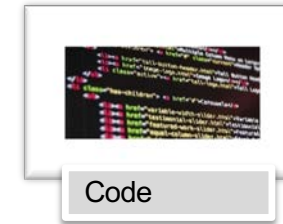
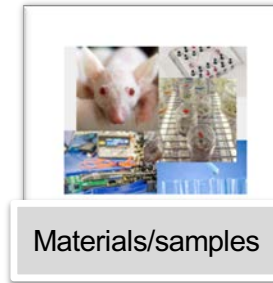
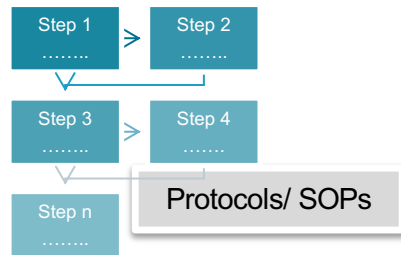
- ❑ Appropriate data and information management from the start of the research process can avoid “data drowning”

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

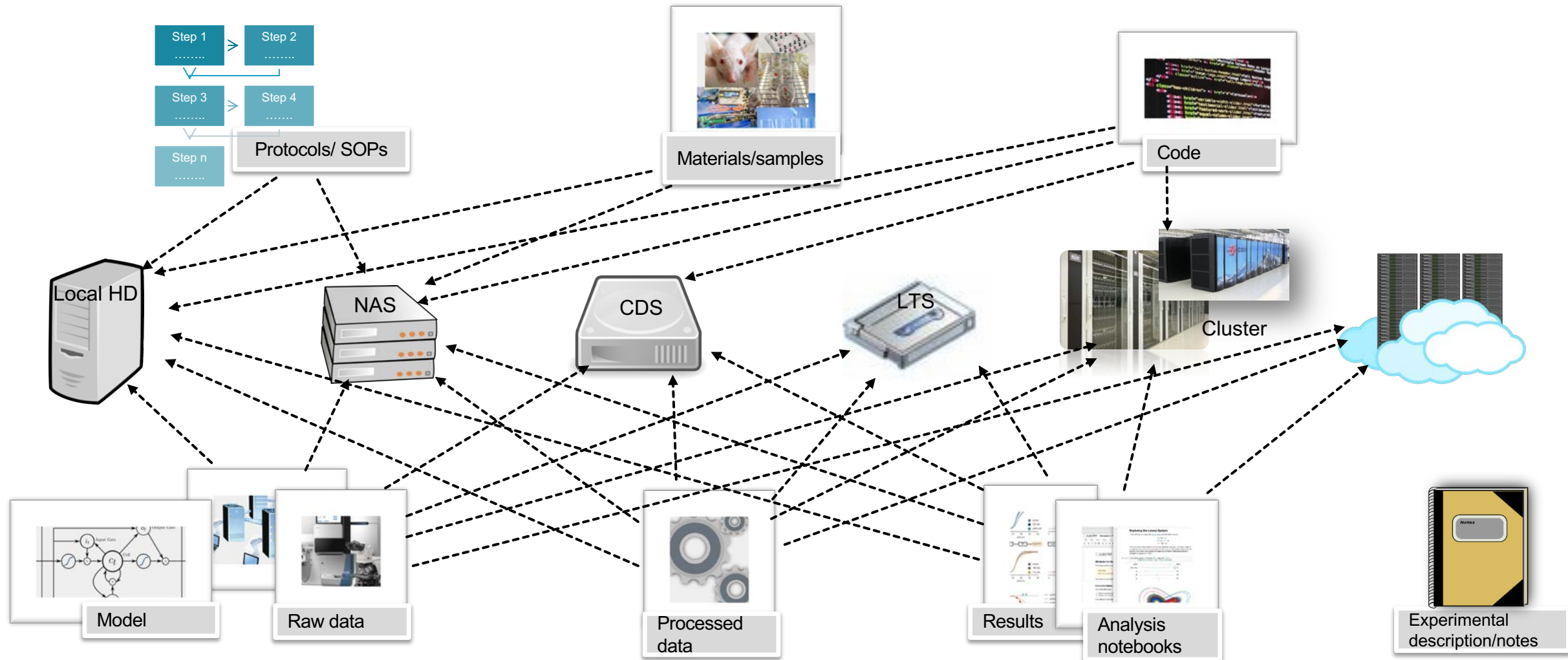
The “*data spread*”: a common scenario in academic institutions



The “*data spread*”: a common scenario in academic institutions



The “*data spread*”: a common scenario in academic institutions

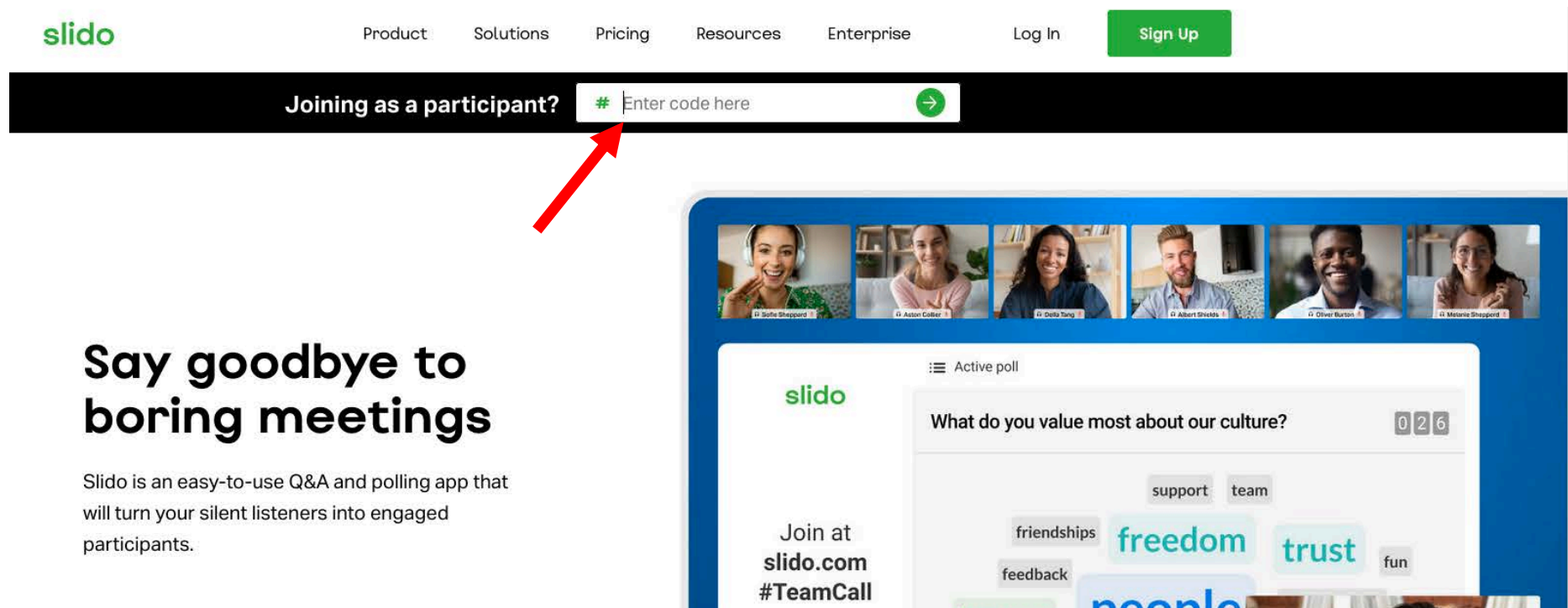


How can we take care of the individual components and how can we bring things together?



Discussion Time

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

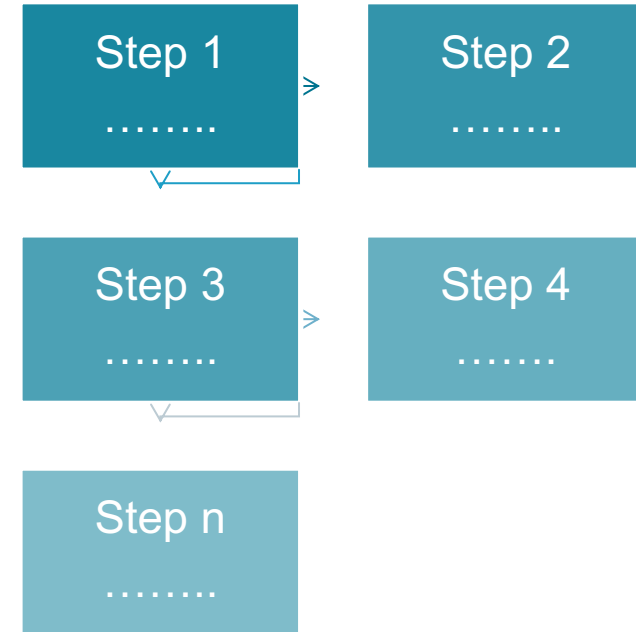
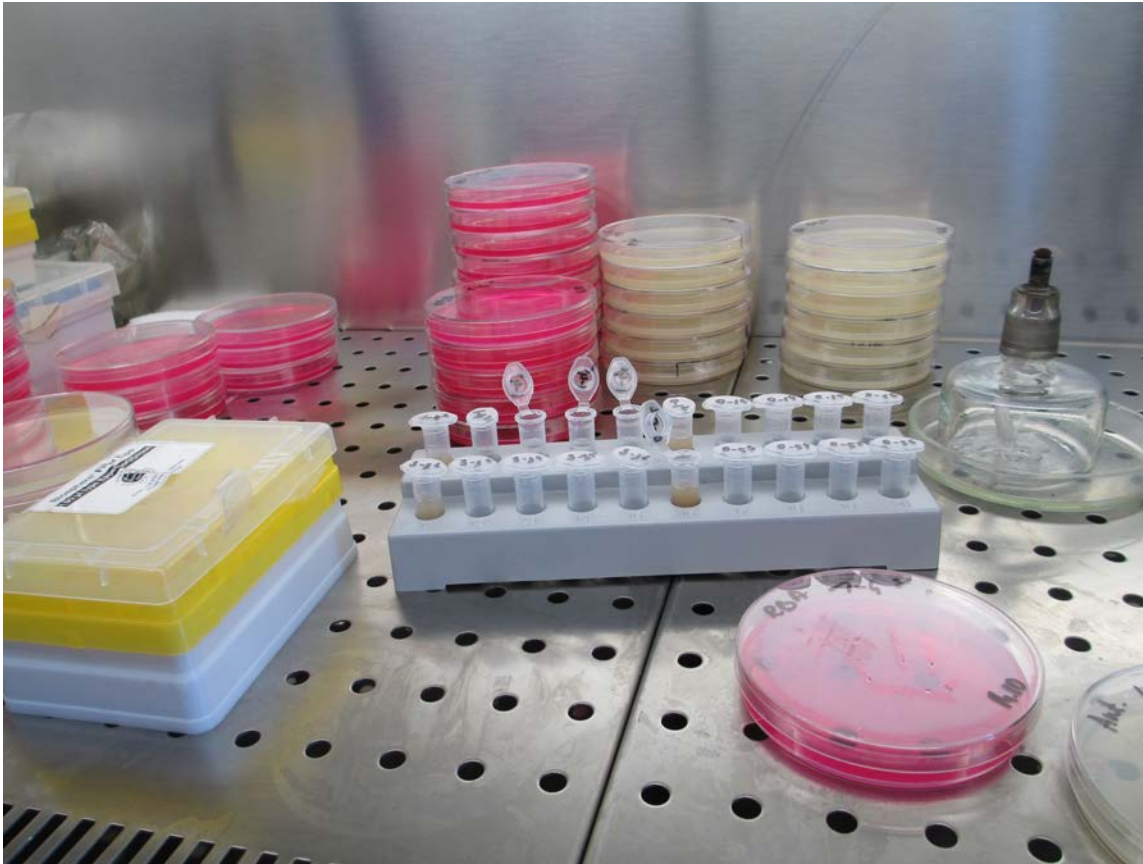


The screenshot displays the Slido website's interface. At the top, the Slido logo is on the left, and navigation links for Product, Solutions, Pricing, Resources, Enterprise, Log In, and a green Sign Up button are on the right. Below the navigation is a black banner with the text "Joining as a participant?" followed by a white input field containing "# Enter code here" and a green arrow button. A red arrow points to the input field. Below the banner, the main content area features a blue header with six video thumbnails of participants. Below the thumbnails is a white poll interface titled "Active poll" with the question "What do you value most about our culture?". The poll shows a count of 026 and several response tags: support, team, friendships, feedback, freedom, trust, fun, and people. To the left of the poll, there is text that says "Join at slido.com #TeamCall".

Say goodbye to boring meetings

Slido is an easy-to-use Q&A and polling app that will turn your silent listeners into engaged participants.

Join at [slido.com](https://www.slido.com)
#TeamCall



Management of samples and protocols

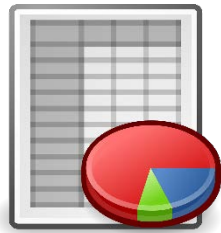
Management of materials and samples



What?

- Biological samples
- Chemical samples
- Materials
- Devices
-

How?



Spreadsheets

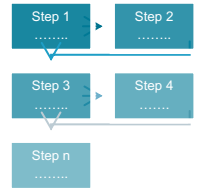
- *Not scalable*
- *No sharing*
- *No efficient search*
- *Easy to use*



Database/
LIMS

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

Management of protocols



What?

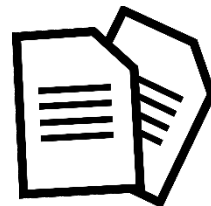
- Step by step description of procedure
- Experimental/computational parameters (e.g. temperature, time, etc.)
- Machine used (experimental)
- OS, program, version, etc. (calculation)

How?



Paper notebook

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



Text files

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



- *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:



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



<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

The screenshot shows a Confluence wiki page for a training session. The page is titled '2020.01.23' and is located within the 'openBIS trainings' space. The page content includes a table with details about the training, such as location, time, program, and training material. The program details are as follows:

Location	Zurich, HG F33.1
Time	09:00-13:00
Program	<ol style="list-style-type: none">1. introduction to openBIS2. How to manage the lab inventory of materials and samples3. How to manage lab protocols4. How to record experiments in the Electronic Lab Notebook & upload data5. How to analyse data stored in openBIS using Jupyter notebooks and MATLAB
Participants	
Training material	https://gitlab.ethz.ch/sis-rdm-training/openbis-training

The page also features a sidebar with a navigation menu, a top navigation bar with search and user options, and a footer with 'Powered by Atlassian Confluence 6.15.9'.

Example of versioning in wiki

Pages /... / 2020 ← View Page ☆ Save for later ...

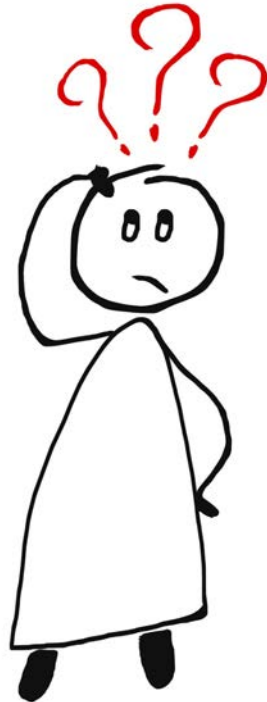
Page History

Compare selected versions

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

[Return to Page Information](#)

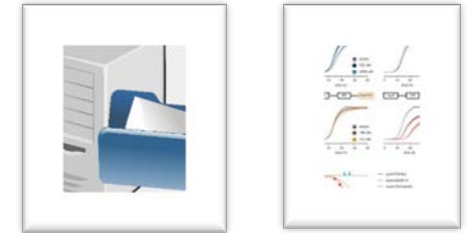
Questions on Management of Samples and Protocols?



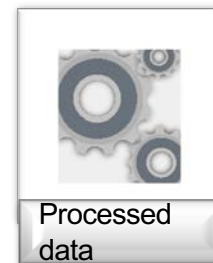
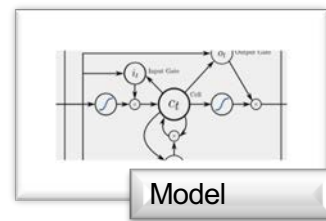


Data files management

Management of research data files



What?



How?



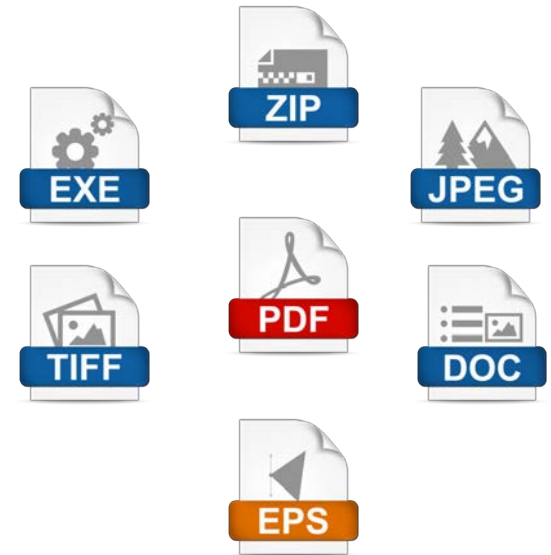
Files / folders hierarchy



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

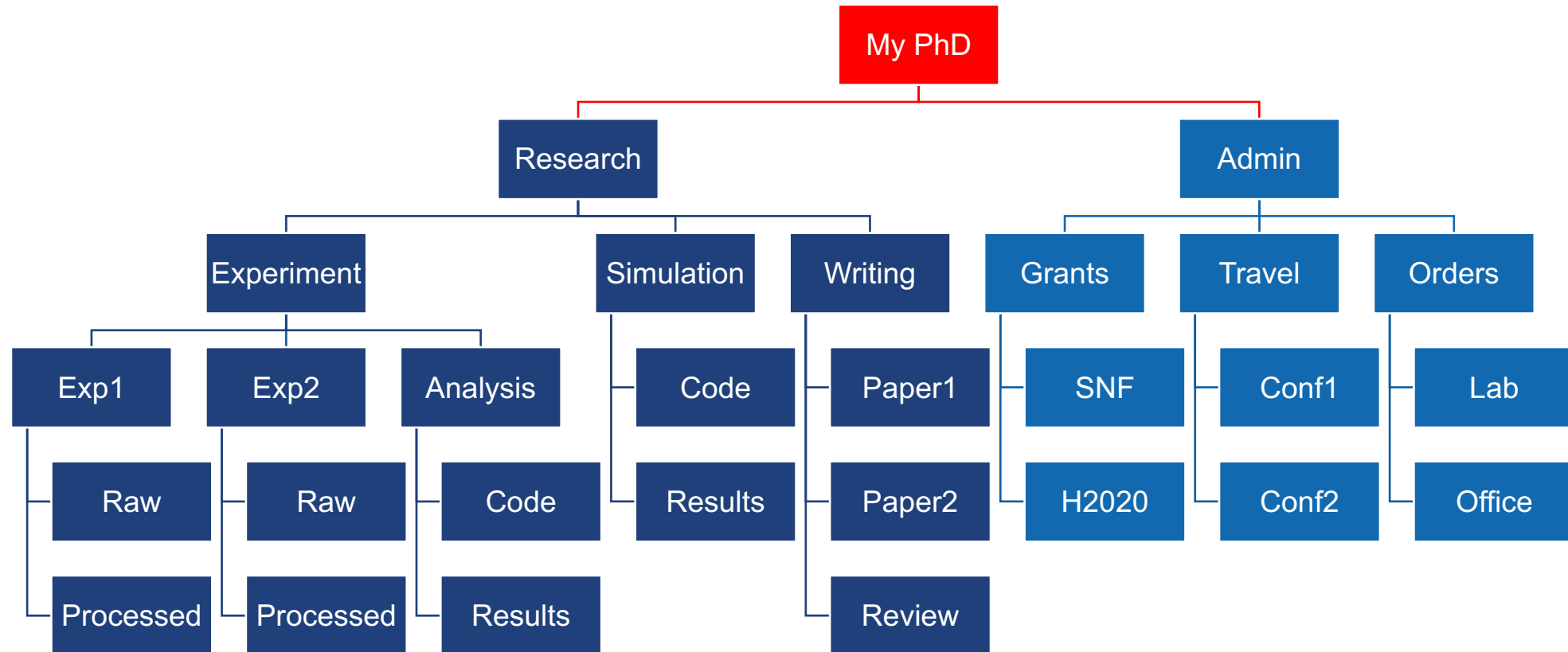
- **Goals** of efficient file / folder organization:
 - Easy to **find** something in the future (you, others)
 - Easy to **file** something
 - Save **disk space** (avoid multiple copies of files)
 - **Reusable** components
 - Avoid problems on different **operating systems**



- 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

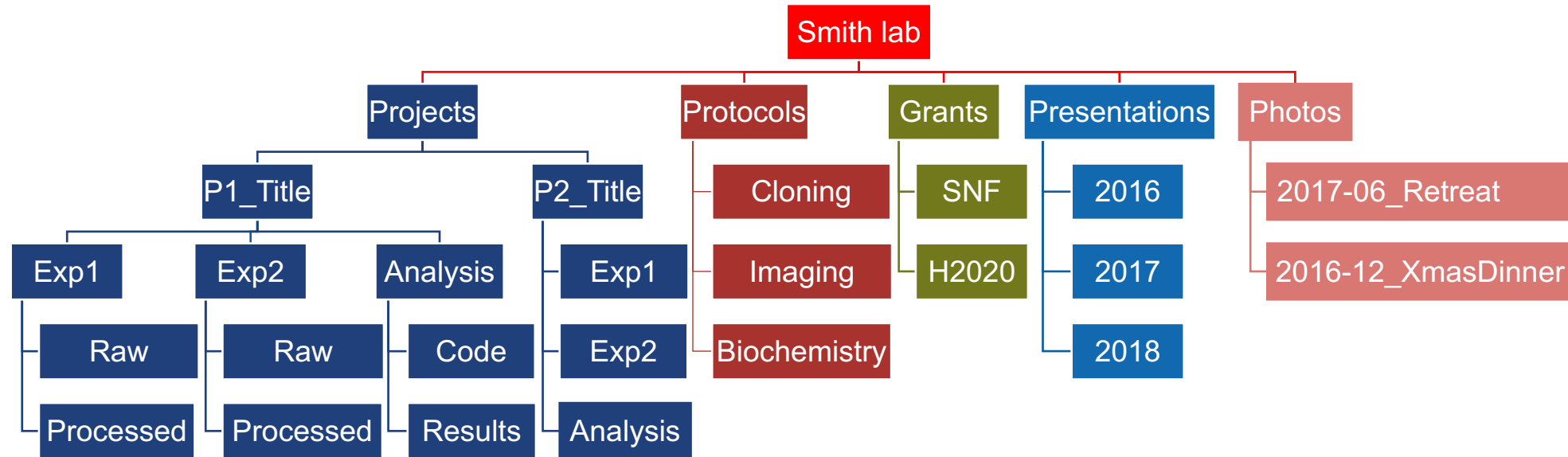
File & folder organisation

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)

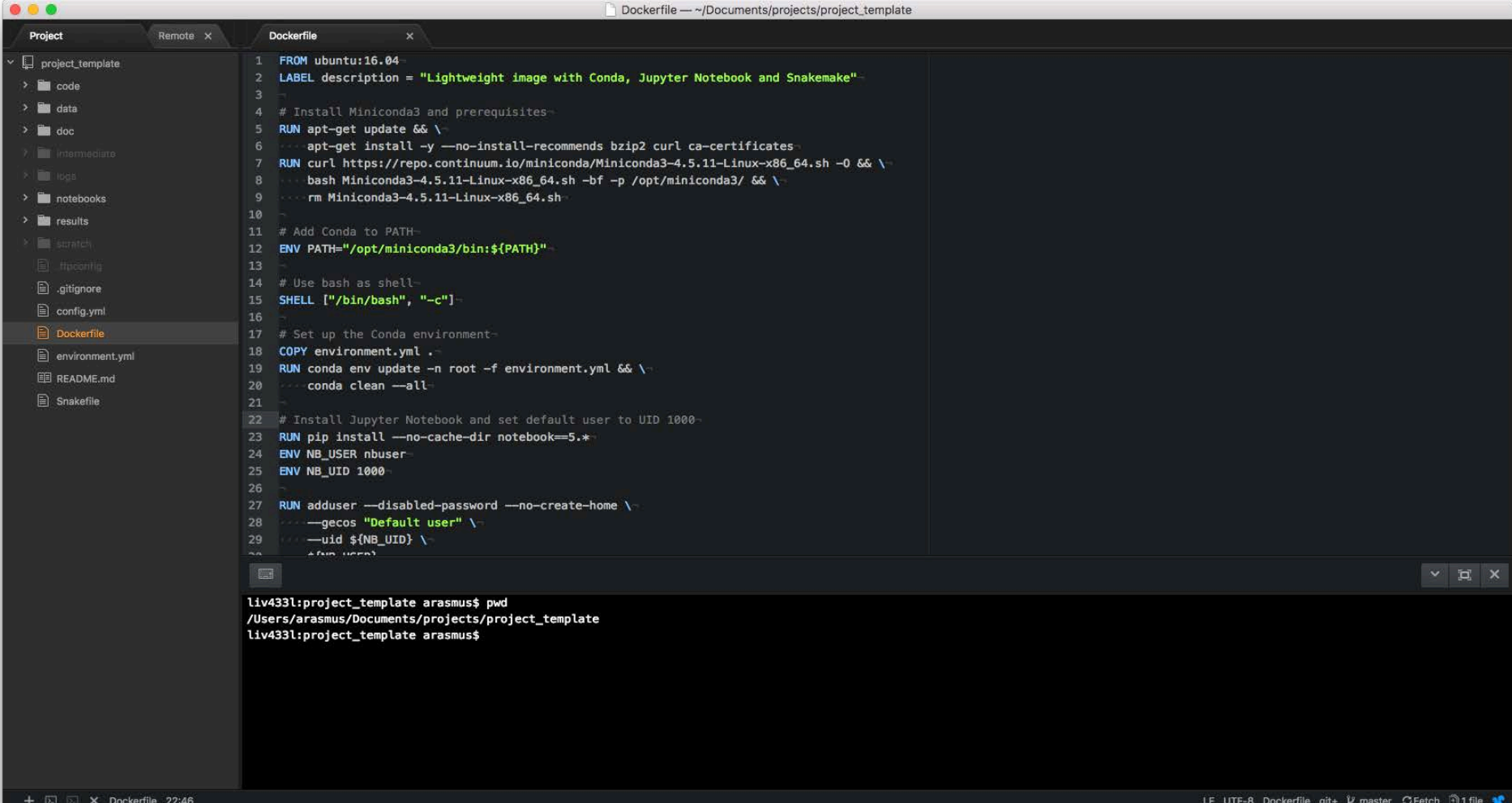
project	
- doc/	documentation for the study
- data/	raw and primary data, essentially all input files, never edit!
- raw_external /	
- raw_internal /	
- meta/	
- code/	all code needed to go from input files to final results
- notebooks/	notebooks that document your day-to-day work
- intermediate/	output 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



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

File & folder organisation

A project in ATOM



The screenshot shows the ATOM editor interface with a project named 'project_template'. The left sidebar displays a file tree with folders like 'code', 'data', 'doc', 'infrastructure', 'logs', 'notebooks', 'results', 'scratch', and files like 'environment.yml', 'README.md', and 'Snakefile'. The main editor area shows a Dockerfile with the following content:

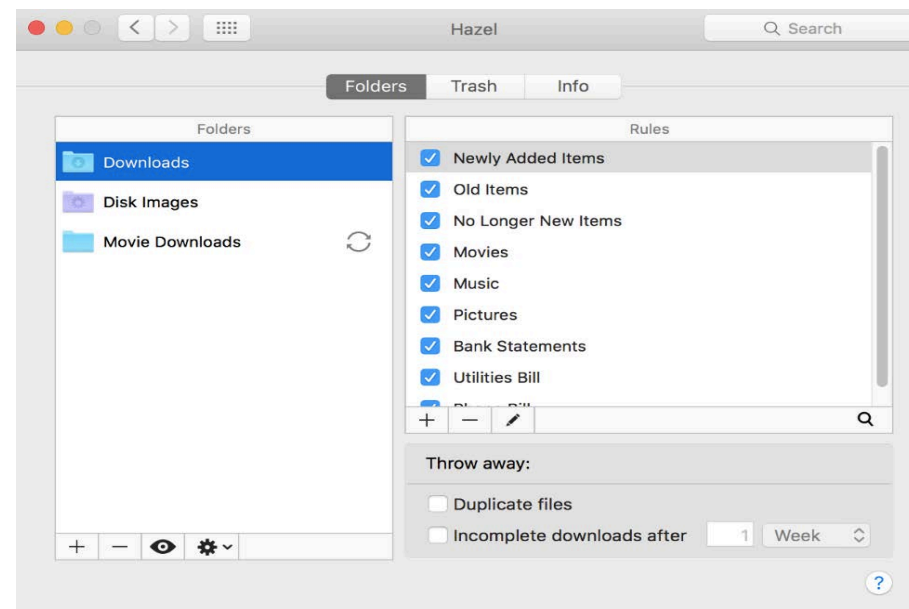
```
1 FROM ubuntu:16.04
2 LABEL description = "Lightweight image with Conda, Jupyter Notebook and Snakemake"
3
4 # Install Miniconda3 and prerequisites
5 RUN apt-get update && \
6     apt-get install -y --no-install-recommends bzip2 curl ca-certificates
7 RUN curl https://repo.continuum.io/miniconda/Miniconda3-4.5.11-Linux-x86_64.sh -O && \
8     bash Miniconda3-4.5.11-Linux-x86_64.sh -bf -p /opt/miniconda3/ && \
9     rm Miniconda3-4.5.11-Linux-x86_64.sh
10
11 # Add Conda to PATH
12 ENV PATH="/opt/miniconda3/bin:${PATH}"
13
14 # Use bash as shell
15 SHELL ["/bin/bash", "-c"]
16
17 # Set up the Conda environment
18 COPY environment.yml .
19 RUN conda env update -n root -f environment.yml && \
20     conda clean --all
21
22 # Install Jupyter Notebook and set default user to UID 1000
23 RUN pip install --no-cache-dir notebook==5.*
24 ENV NB_USER nbuser
25 ENV NB_UID 1000
26
27 RUN adduser --disabled-password --no-create-home \
28     --gecos "Default user" \
29     --uid ${NB_UID} \
30     $(for user in
```

Below the code editor, a terminal window shows the execution of the Dockerfile:

```
liv4331:project_template arasmus$ pwd
/Users/arasmus/Documents/projects/project_template
liv4331:project_template arasmus$
```

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>)



Name	Date Modified	Size	Kind
file_00001.tif	4 May 2018, 14:16	532 KB	TIFF image
file_00002.tif	4 May 2018, 14:21	17.3 MB	TIFF image
file_00003.tif	4 May 2018, 14:24	17.3 MB	TIFF image
file_00004.tif	4 May 2018, 14:25	5.3 MB	TIFF image
file_00005.tif	4 May 2018, 14:27	3.4 MB	TIFF image
file_00006.tif	4 May 2018, 14:28	3.4 MB	TIFF image
file_00007.tif	4 May 2018, 14:29	3.4 MB	TIFF image
file_00008.tif	4 May 2018, 14:29	3.4 MB	TIFF image
file_00009.tif	4 May 2018, 14:29	3.4 MB	TIFF image
file_00010.tif	4 May 2018, 14:29	3.4 MB	TIFF image
file_00011.tif	4 May 2018, 14:29	3.4 MB	TIFF image
file_00012.tif	4 May 2018, 14:30	3.4 MB	TIFF image
file_00013.tif	4 May 2018, 14:30	3.4 MB	TIFF image
file_00014.tif	4 May 2018, 14:30	3.4 MB	TIFF image
file_00015.tif	4 May 2018, 14:30	3.4 MB	TIFF image
file_00016.tif	4 May 2018, 14:31	3.4 MB	TIFF image
file_00017.tif	4 May 2018, 14:31	3.4 MB	TIFF image
file_00018.tif	4 May 2018, 14:31	3.4 MB	TIFF image
file_00019.tif	4 May 2018, 14:31	3.4 MB	TIFF image
file_00020.tif	4 May 2018, 14:32	3.4 MB	TIFF image
file_00021.tif	4 May 2018, 14:32	3.4 MB	TIFF image
file_00022.tif	4 May 2018, 14:32	3.4 MB	TIFF image
file_00023.tif	4 May 2018, 14:32	3.4 MB	TIFF image
file_00024.tif	4 May 2018, 14:33	3.4 MB	TIFF image
file_00025.tif	4 May 2018, 14:33	3.4 MB	TIFF image
file_00026.tif	4 May 2018, 14:33	3.4 MB	TIFF image
file_00027.tif	4 May 2018, 14:34	3.4 MB	TIFF image
file_00028.tif	4 May 2018, 14:34	3.4 MB	TIFF image
file_00029.tif	4 May 2018, 14:34	3.4 MB	TIFF image
file_00030.tif	4 May 2018, 14:35	3.4 MB	TIFF image
file_00031.tif	4 May 2018, 14:35	3.4 MB	TIFF image
file_00032.tif	4 May 2018, 14:35	3.4 MB	TIFF image
file_00033.tif	4 May 2018, 14:35	3.4 MB	TIFF image
file_00034.tif	4 May 2018, 14:36	3.4 MB	TIFF image

Rename Finder Items:

Replace Text

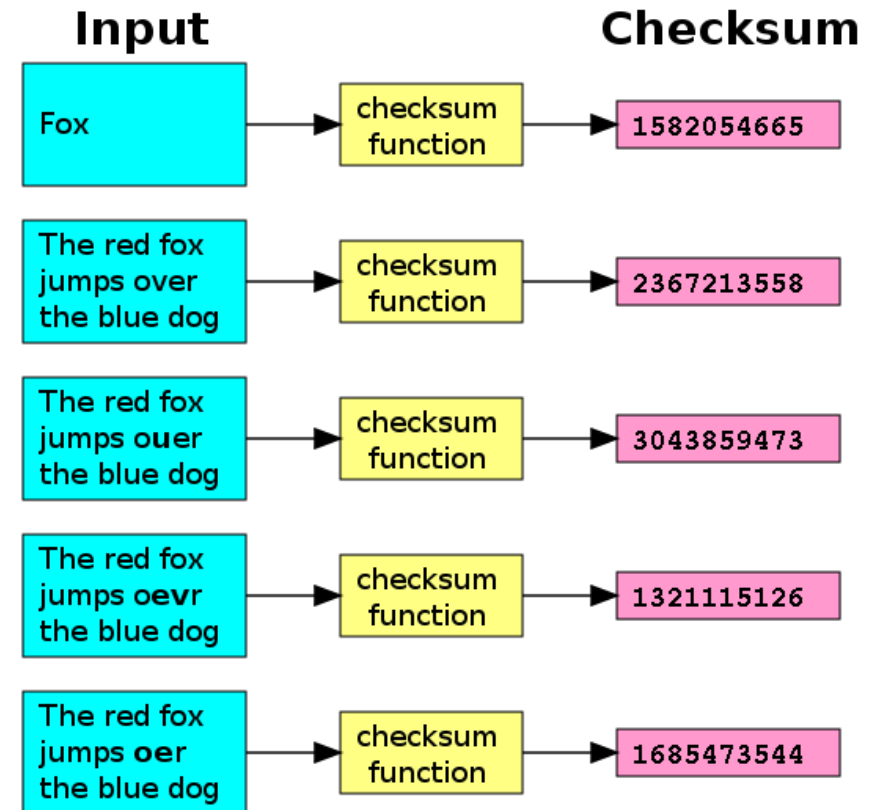
Find: file_ Replace with: 2018-06-06_

Example: 2018-06-06_00001.tif

Cancel Rename

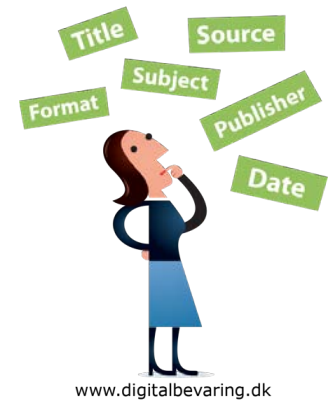
File validation and compression

- **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

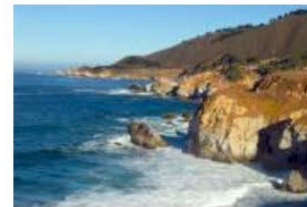
- **Metadata** is the *data about your data* (a.k.a. **data model**)
- Use of structured metadata **facilitates data organization** and searches
- **Metadata** is a key element of the **FAIR data** principles
- Existing **metadata schemas** are preferred (can be extended, if necessary)



Search by Discipline



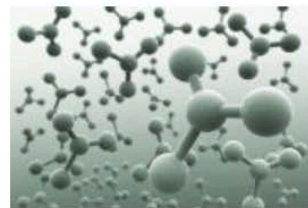
Biology



Earth Science



General Research Data



Physical Science



Social Science & Humanities

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

Metadata

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

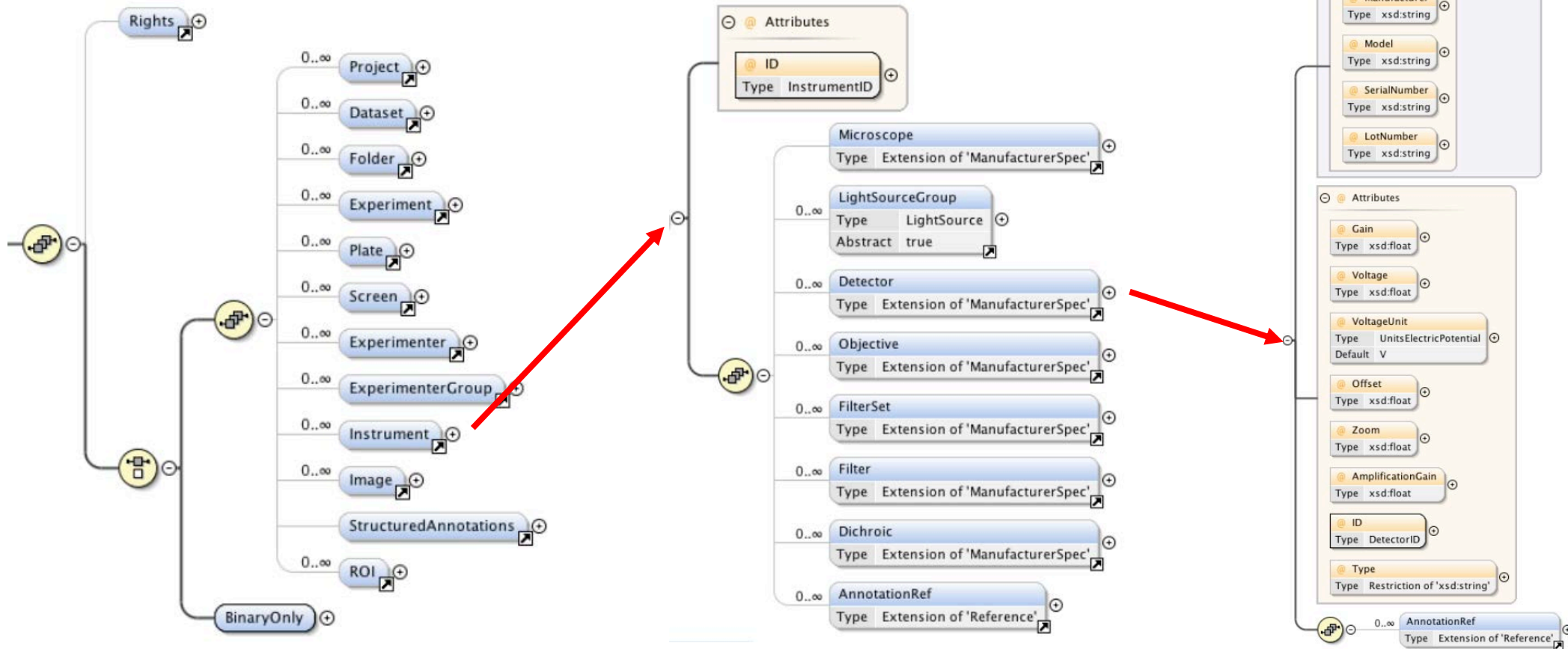
ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation 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



www.digitalbevaring.dk

Metadata

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



www.digitalbevaring.dk

Metadata

The screenshot displays the IDR web client interface. On the left, a file explorer shows a tree structure of data files. The main area is a grid of microscopy images, with the top-left image highlighted by a blue border. A red oval encircles the entire grid of images, and the word "Data" is written in red below it. On the right, a metadata panel is shown, with a blue oval around it and the word "Metadata" written in blue below it. The metadata panel includes sections for General, Acquisition, and Preview, with the following details:

- General:** Dataset Name: H3K4M_H3K27M; Image Name: AC16_Rep2_18d24h37d2h_H3K4M488_H3K27M594_01_SIR_THR_ALN.tif; Characteristics: Homo sapiens; Organism: Homo sapiens; Term Source 1: NCBI/Taxon; Term Source 2: REF.
- Acquisition:** Image ID: 10502514; Owner: Public data; Import Date: 2020-09-18 11:50:49; Dimensions (XY): 1024 x 1024; Pixels Type: uint16; Pixels Size (XZ) (µm): 0.04 x 0.04 x 0.12; Z-sections/Timepoints: 80 x 1; Channels: H3K27me3, H3K4me3, DAPI; ROI Count: 0.
- Preview:** Cell Lines: AC16; Organism: Homo sapiens; Condition: 18°C 24h, 37°C 2h; Image File Type: Reconstructed image.

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

Filters

Synonym query expansion On Off

Sources

- Agricola (USDA/NAL)
- Chinese biological abstracts
- CiteXplore records
- Patents
- Preprint records
- PubMed/MEDLINE (NLM)

Special Collections

- All BMJ
- All manuscripts
- EuroFIR
- Europe PMC manuscripts

Full Text Availability

- In Europe PMC
- Open Access

Publication Type

Choose one Publication Type

CC License

Choose one License Type

Article Sections

Choose a section type

Data Links and Data Citations

Choose one Link/Citation type

External Links

Choose one External Links Provider

Language

Choose one Language

- Choose one Language
- Afrikaans
- Albanian
- Arabic
- Armenian
- Azeri
- Bosnian
- Bulgarian
- Catalan
- Chinese
- Czech
- Danish
- Dutch
- English
- Esperanto
- Estonian
- Finnish
- French
- Georgian

<https://europepmc.org/advancesearch>

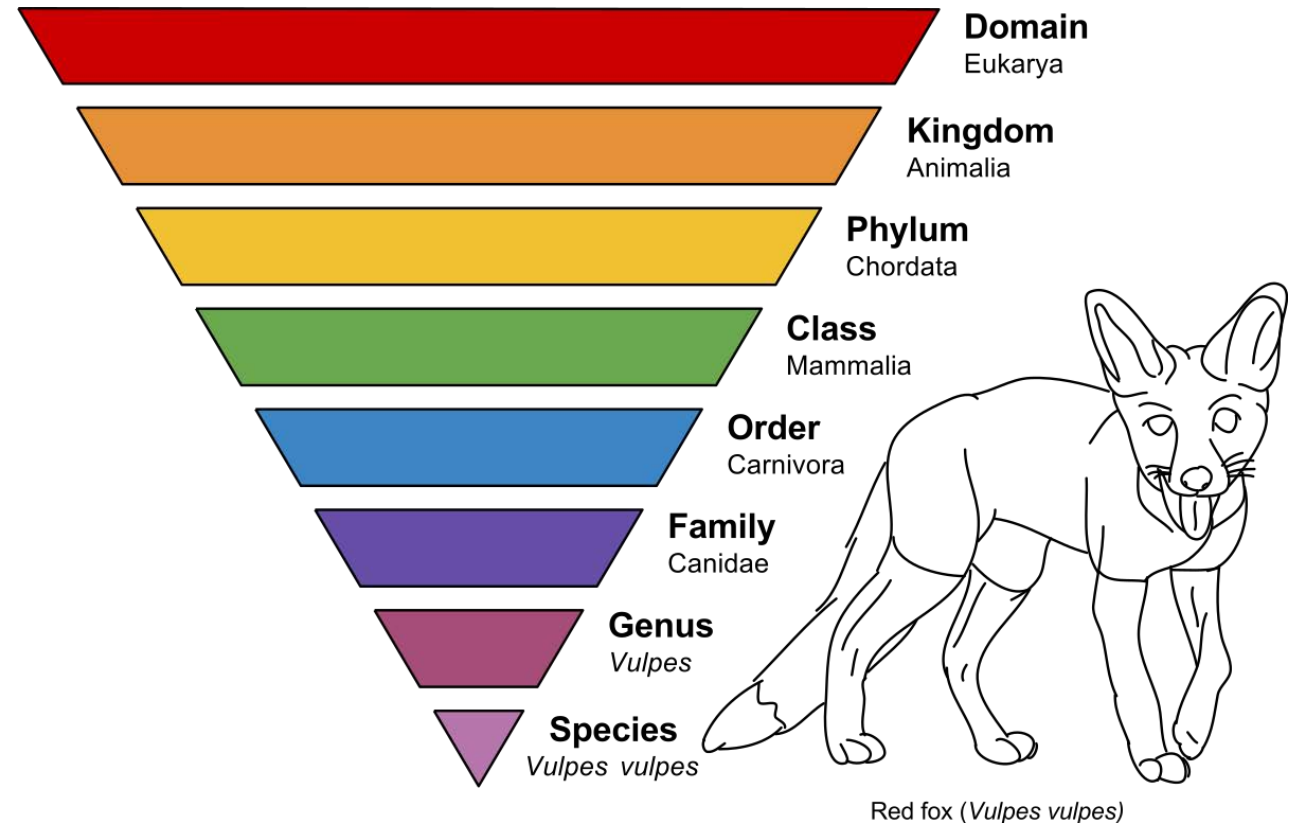
Metadata

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

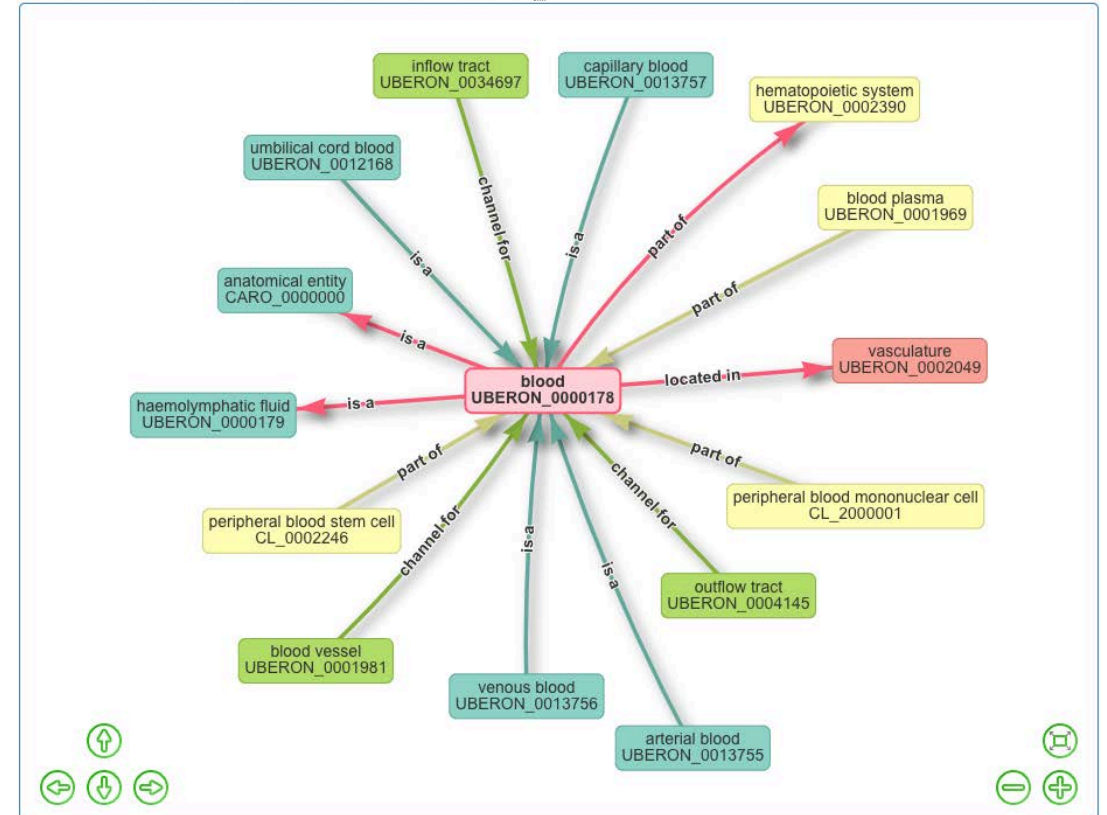
- Annotation based on common standards
- Controlled vocabularies, taxonomies
- Standardized metadata are the basis for ontologies (knowledge representations)

[Basic Register of Thesauri, Ontologies & Classifications](#)

[EMBL-EBI Ontology Lookup Service](#)

OLS > HCAO > UBERON_0000178

Visualized term: blood (http://purl.obolibrary.org/obo/UBERON_0000178)



Create clusters | Open all clusters | Auto rearrange off | Hierarchical layout | Search nod | Search Node

blood

Description: A fluid that is composed of blood plasma and erythrocytes.

Synonyms: portion of blood, vertebrate blood

Short id: UBERON_0000178 (*iri:* http://purl.obolibrary.org/obo/UBERON_0000178)

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)
- **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)

[Basic Register of Thesauri, Ontologies & Classifications](#)

[EMBL-EBI Ontology Lookup Service](#)

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

The screenshot shows the JabRef application window titled 'domain-decomp.bib (BibTeX mode)'. The interface includes a menu bar (File, Edit, Search, Groups, View, BibTeX, Quality, Tools, Options, Help), a toolbar, and a search bar. A 'Web search' panel on the left shows 'Google Scholar' as the selected source. The main area displays a table of bibliographic entries:

#	entrytype	author/editor	title	year	journal/booktitle
1	Article	Acebrón et al.	Efficient parallel solution of nonlinear parabolic partial diff...	2010	J. Sci. Comput.
2	Article	Beirão da Veiga et al.	Robust BDDC preconditioners for Reissner-Mindlin plate b...	2010	SIAM Journal on ...
3	Article	Börgers and Maclachlan	An angular multigrid method for computing mono-energeti...	2010	Journal of compu...
4	Article	D'Ambra et al.	MLD2P4: A Package of Parallel Algebraic Multilevel Domain...	2010	ACM Transaction...
5	Article	Dohrmann and Widlund	Hybrid domain decomposition algorithms for compressible...	2010	International Jour...
6	Article	Dostál et al.	Scalable TFETI algorithm for the solution of multibody cont...	2010	International Jour...
7	Article	Du and Liang	An efficient S-DDM iterative approach for compressible co...	2010	Journal of compu...
8	Article	Galvis and Sarkis	FETI and BDD preconditioners for Stokes-Mortar-Darcy sys...	2010	Communications ...
9	Article	Giraud et al.	Using multiple levels of parallelism to enhance the perfor...	2010	Parallel Computing
10	Article	Gong et al.	Dynamic domain decomposition method and its applicatio...	2010	Wuhan Univ. J. N...
11	Article	Herrera and Yates	The multipliers-free domain decomposition methods	2010	Numerical Metho...
12	Article	Hesch and Betsch	Transient three-dimensional domain decomposition probl...	2010	International Jour...
13	Article	Hu et al.	Nonoverlapping domain decomposition methods with a si...	2010	Mathematics of C...
14	Article	Jun	A stable noniterative Predictionslash Correction domain d...	2010	Applied Mathema...
15	Article	Klawonn and Rheinbach	Highly scalable parallel domain decomposition methods wi...	2010	Zeitschrift für An...
16	Article	Leiva et al.	Iterative strong coupling of dimensionally heterogeneous ...	2010	International Jour...
17	Article	Loiselet et al.	Optimized Domain Decomposition Methods for the Spheric...	2010	SIAM Journal on

Below the table, a detailed view of the selected entry (2) is shown:

Article (BeirãodaVeiga:2010:RBP)
Beirão da Veiga, L.; Chinosi, C.; Lovadina, C. & Pavarino, L. F.
Robust BDDC preconditioners for Reissner-Mindlin plate bending problems and MITC elements
SIAM Journal on Numerical Analysis, 2010, 47, 4214-4238

Status: Preferences recorded.

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

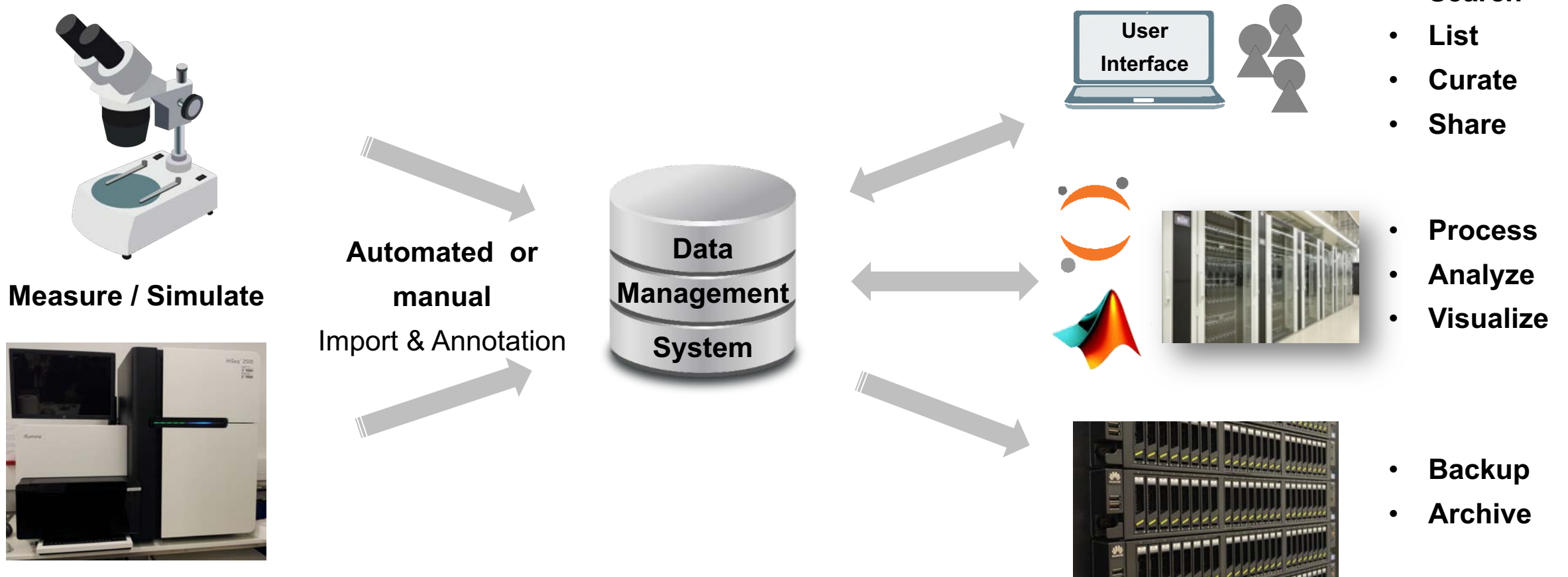


Scientific



Data management software

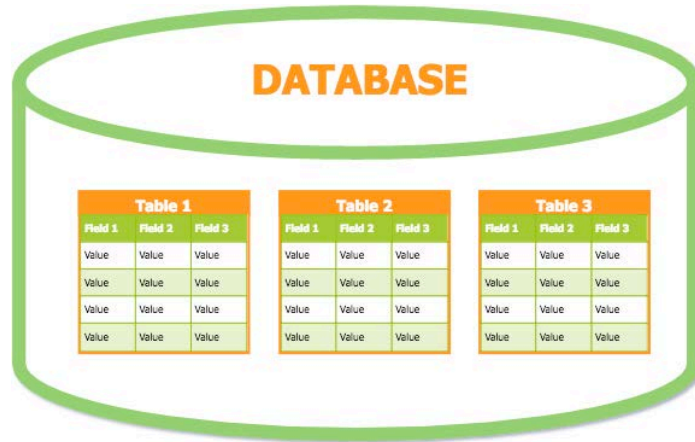
Generic database workflow



Data management software

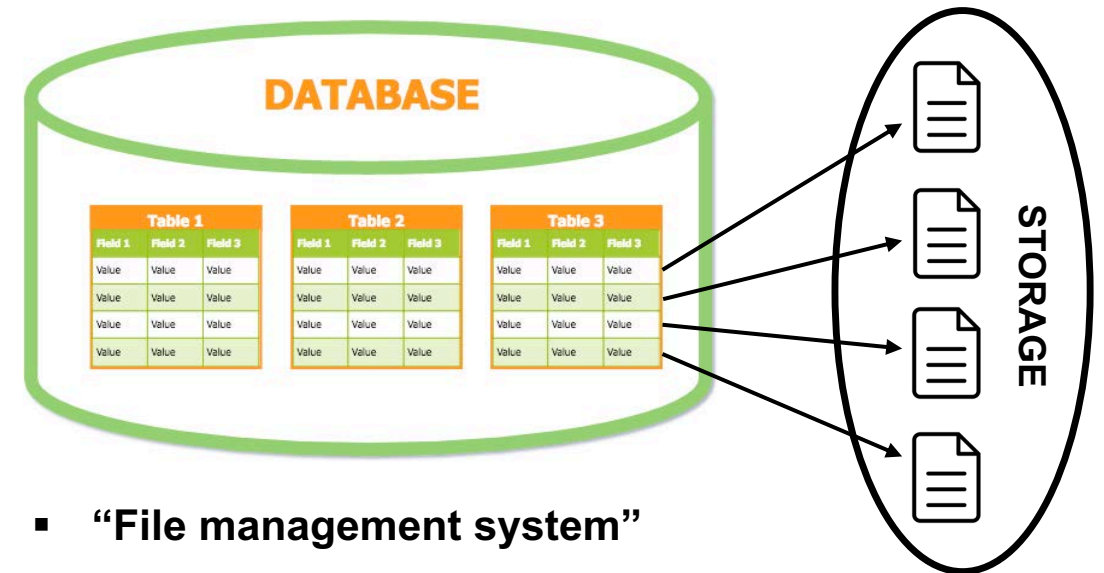
Where is my data???

Metadata + Data



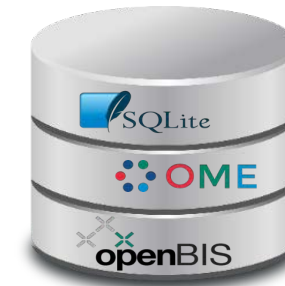
- Structured data required
- DB-specific data (array DBs)
- No conventional file access

Metadata + Link to Data Files



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

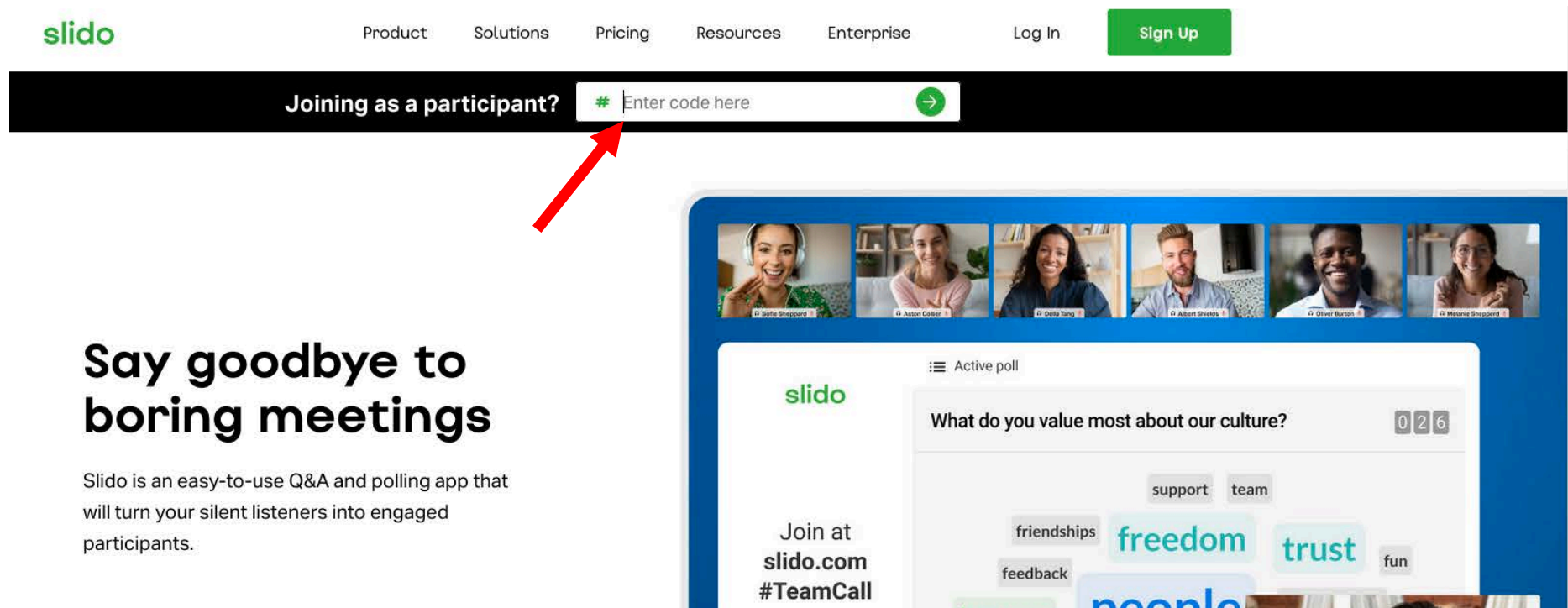
Comparison: folder hierarchy vs. databases



	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**



The screenshot shows the Slido website interface. At the top, there is a navigation bar with the Slido logo and links for Product, Solutions, Pricing, Resources, Enterprise, Log In, and a green Sign Up button. Below this is a black banner with the text "Joining as a participant?" followed by a white input field containing "# Enter code here" and a green arrow button. A red arrow points to the input field. Below the banner, there is a blue header for a meeting with six video thumbnails of participants. The main content area features a poll titled "What do you value most about our culture?" with a "0 2 6" counter. The poll results show words like "support", "team", "friendships", "freedom", "trust", "fun", and "people". To the left of the poll, there is text that says "Join at slido.com #TeamCall".

Say goodbye to boring meetings

Slido is an easy-to-use Q&A and polling app that will turn your silent listeners into engaged participants.

Join at
slido.com
#TeamCall

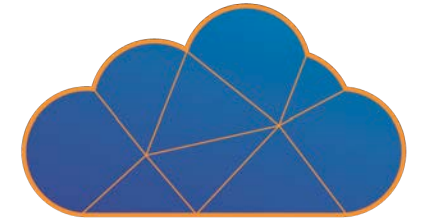
Data storage – Your options ...

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

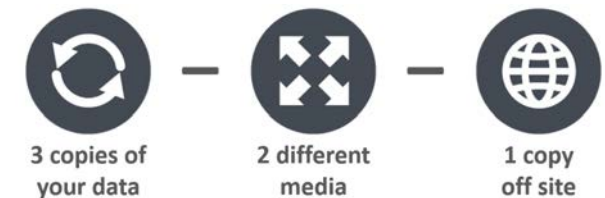
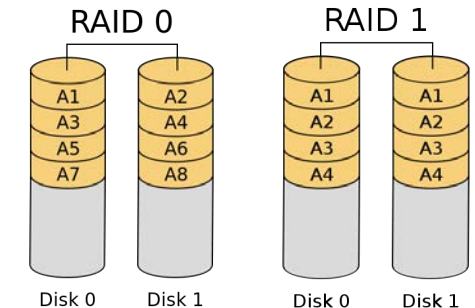


Data storage – Your options ...

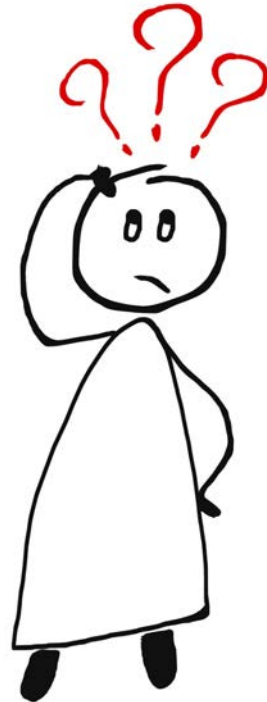
- External hard disks are cheap but unreliable and don't scale!
- Polybox / SwitchDrive
- Network Attached Storage (NAS)
 - NAS offering from IT Services – always check with your ISG first
- Cost Defined Storage (CDS)
 - For very large storage requirements (>100 TB)
- Long-term Storage (LTS)
 - Tape 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?





Global Search

- Lab Notebook
 - Others
 - Others (disabled)
 - My Space (Diana Ottoz)
 - Inducible Transcription Factor
 - Analysis of the abundance of the four variants of the trans
 - Induction of the transcription factor in standard growth co
 - Detection of LexA-ER-B42 induction by flow cytometry**
 - Analysis results
 - scripts
 - Flow cytometry files
 - Detection of LexA-ER-B42 induction by western blottir
 - gels
 - Detection of LexA-ER-B112 induction by western blott
 - preview
 - gels
- Inventory
 - Materials
 - Bacteria
 - Bacteria collection

RAW_DATA : Flow cytometry files
 ANALYSIS_SCRIPTS : scripts
 ANALYZED_DATA : Analysis results

Experimental Step: Detection of LexA-ER-B42 induction by flow cytometry

More ...

General

Name:
Detection of LexA-ER-B42 induction by flow cytometry

Owner:
Diana Ottoz

Experimental goals:
Analyze the induction of **LexA-ER-B42** in a concentration series of beta-estradiol using a fluorescence readout

Experimental results:
The LexA-ER-B42 induction can be measured by using a target gene encoding a fluorescence protein. *LexA-ER-B42 induction is different at different concentrations of inducer.*

Parents

Filter Exports and ... Columns

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

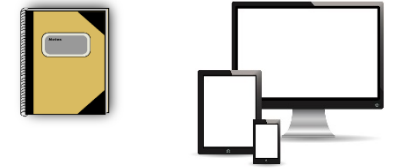
Electronic lab notebooks used for development or research in regulated industries, such as medical devices or pharmaceuticals, are expected to comply with the **21 CFR Part 11 FDA** regulations:

<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



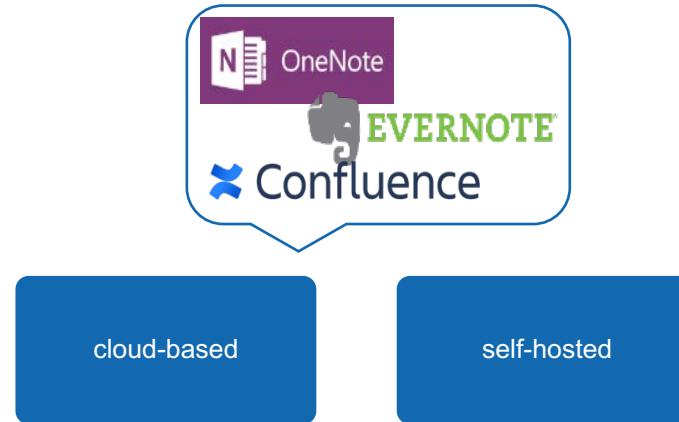
Advantages of ELNs over paper notebooks:

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 CFR Part 11 FDA** regulations

ELNs with database back end



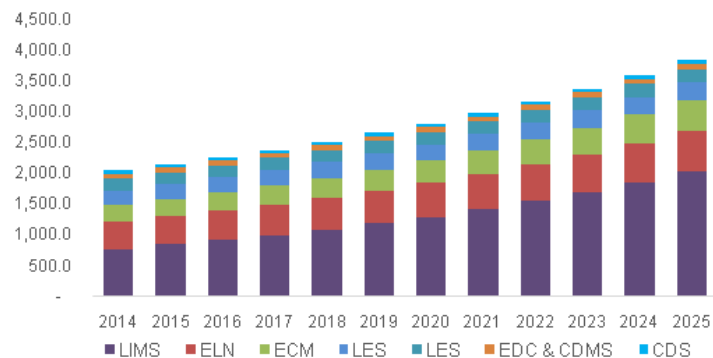
- Structured ELNs compared to note-keeping applications
- Additional functionalities compared to note-keeping applications (e.g. workflow management, chemical structures drawing, etc)
- Can be discipline-specific or cross-disciplines
- Some systems offer an all-in-one solution for RDM
- Some systems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

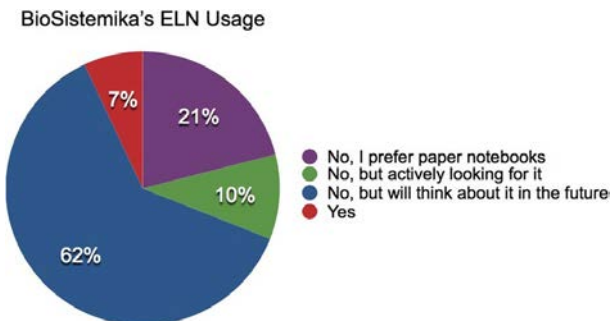
Current state of the art

- 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.

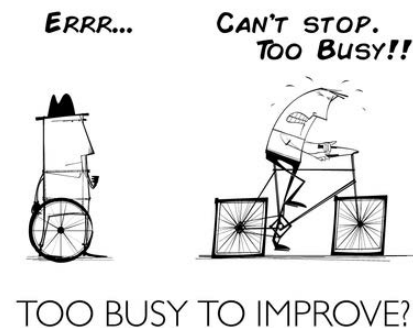
Reasons for slow adoption of ELNs/LIMS in academia

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.

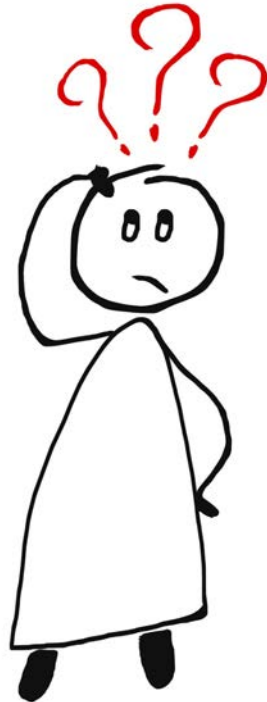


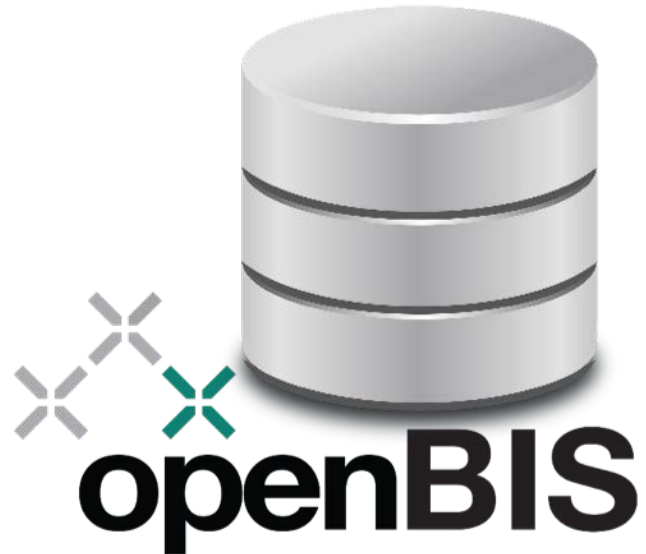
WorkCompass

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?

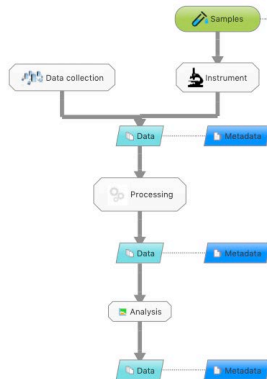




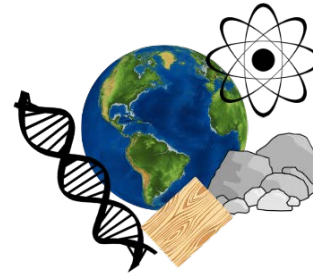
The ETH Scientific IT Services data management solution for research groups

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

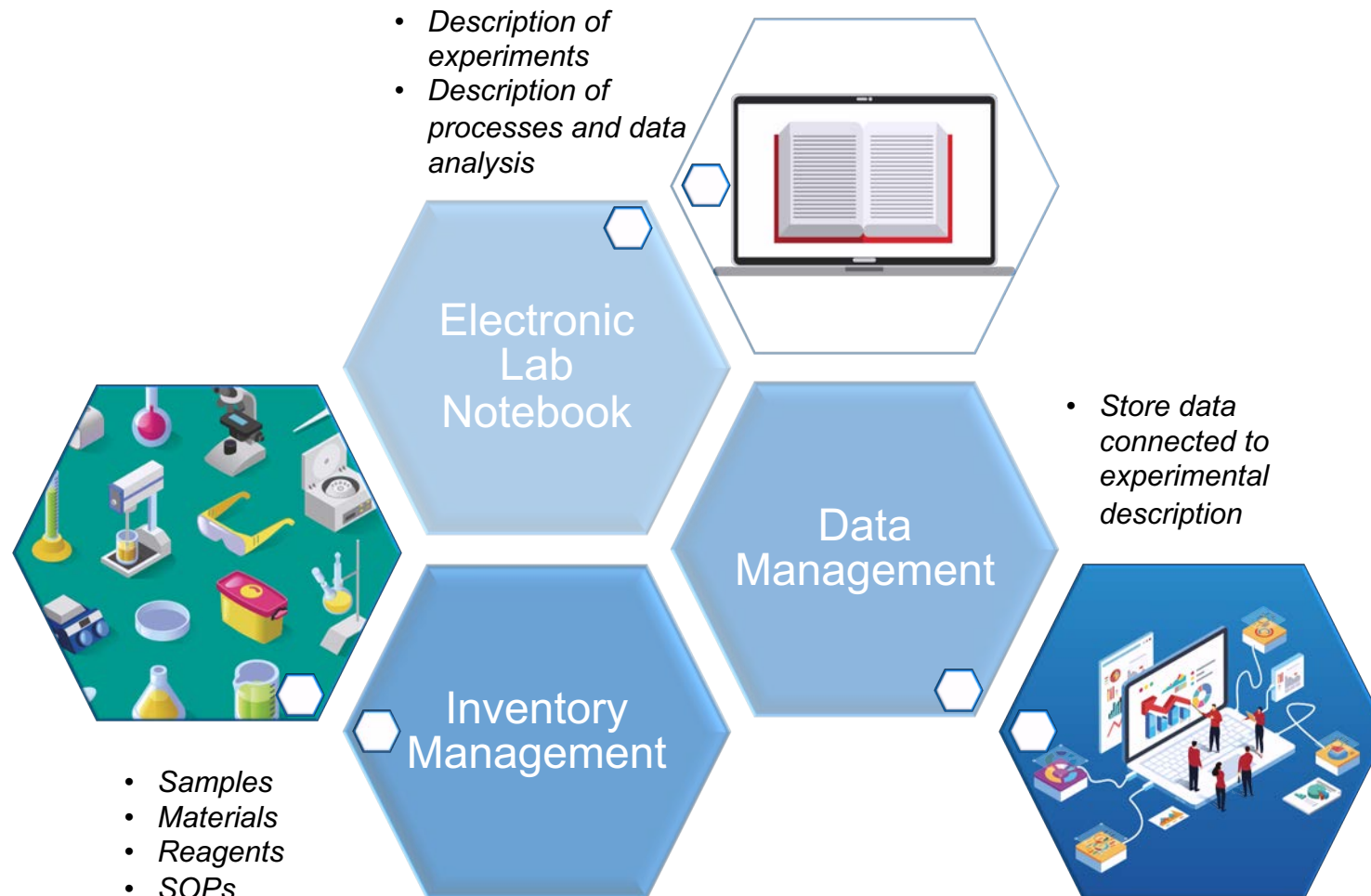


Can be used in most quantitative science fields (e.g. *life sciences, physics, env. sciences, material sciences, etc*)



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

openBIS: a complete solution for FAIR data management



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.

Collection: Yeast collection
/MATERIALS/YEASTS/YEAST_COLLECTION_1

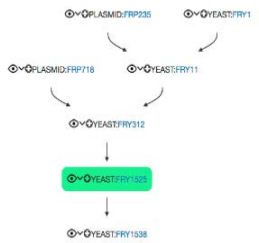
Filter	Q	Toggle AND	Global OR	Options	Columns						
Identifier	Storage	Background-specific markers	Common markers	Genetic Background	Mating Type	Operations					
<input type="checkbox"/> /MATERIALS/FRY1	MINUS80_A2 [1, 1] Box 1: A7	met15-	ura3- his3- leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY11	MINUS80_A2 [1, 1] Box 1: A5	met15-	ura3- his3- leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY312	MINUS80_A2 [1, 1] Box 1: B7	met15-	ura3- leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY400	MINUS80_A2 [1, 1] Box 1: B6	met15-	leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY401	MINUS80_A2 [1, 1] Box 1: B9	met15-	leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY403	MINUS80_A2 [1, 1] Box 1: C1	met15-	leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY417	MINUS80_A2 [1, 1] Box 1: C2	met15-	leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY418	MINUS80_A2 [1, 1] Box 1: C3	met15-	leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY460	MINUS80_A2 [1, 1] Box 1: C4	met15-	ura3- leu2-	BY4741	a	Operations					
<input type="checkbox"/> /MATERIALS/FRY482		met15-	his3- leu2-	BY4741	a	Operations					

1 - 10 of 29 items | 10 - Per Page | Page 1 of 3

Welcome to openBIS ELN-LIMS.

- Lab Notebook
 - My Space (Caterina)
 - Others
- Inventory
 - Materials
 - Methods
- Stock
 - Stock Catalog
 - Stock Orders
- Utilities
 - Jupyter Workspace
 - New Jupyter Notebook
 - User Profile
 - Object Browser
 - Vocabulary Browser
 - Advanced Search
 - Export Builder
 - Storage Manager
 - Trashcan
 - Settings
 - About

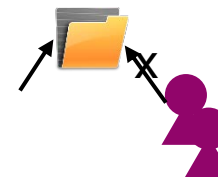
openBIS features



Relationships



Import/Export



User rights management

Principal	Role	Name	Author	Total Item Date	Total Item Date
Samuel Mott	Curator	FRP235_Nova *	adam	2015-01-17 09:52:23 GMT+02:00	2015-01-17 11:58:12 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:52:23 GMT+02:00	2015-01-17 09:52:23 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:52:23 GMT+02:00	2015-01-17 09:52:23 GMT+02:00
Child of Service		AVARIFA:FRP235	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:20 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:20 GMT+02:00
Samuel Mott	Curator	FRP235_Nova	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Child of Service		AVARIFA:FRP235	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Power		001-180-18484888-83	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Power		001-180-18484888-84	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Power		001-180-18484888-85	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Power of Sample		001-180-18484888-86	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Power		001-180-18484888-87	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Power		001-180-18484888-88	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Power		001-180-18484888-89	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Power		001-180-18484888-90	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Child of Service		AVARIFA:FRP235	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-01-17 09:28:28 GMT+02:00	2015-01-17 09:58:18 GMT+02:00
Power of Sample		001-180-18484888-91	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00
Annotations State		https://zenodo.org/record/399111	adam	2015-04-11 11:48:48 GMT+02:00	2015-04-11 11:48:48 GMT+02:00

Audit trail

Rack: 1 2 3 4 5 6 7 8 9 10

1	4	14							
2	1	7	11	12					
3	6	5	9	10					

Box Name: 14

Box Size: 4X4

Box Position: 1 2 3 4

A	FRP1642			
B				
C				
D				

Samples' storage manager



Barcode reader



Integration with Data Repositories



Data immutability

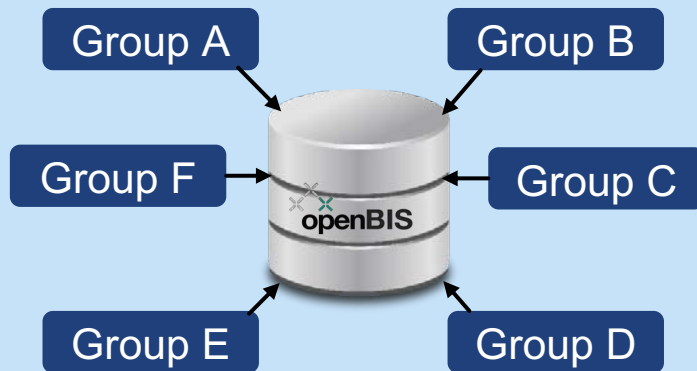


Archive to tape (ETHZ)

RDM services at ETHZ

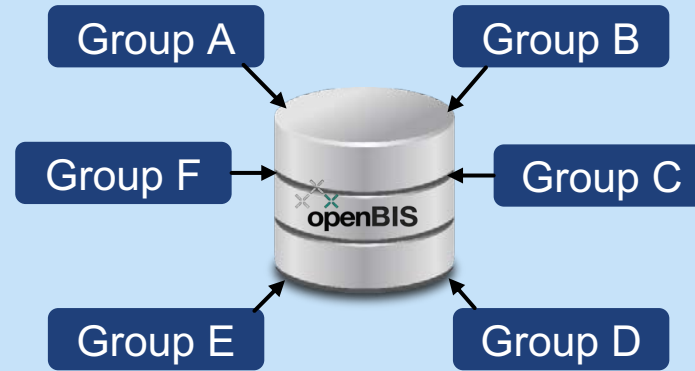
<https://ethz.ch/services/en/it-services/catalogue/software-business-applications/research-data-management.html>

Research Data Hub



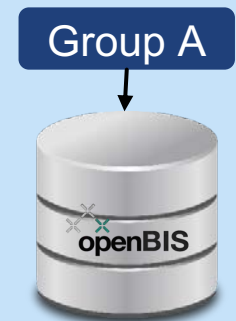
- Centrally managed
- Shared resource
- Limited Customization
- Only storage costs

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

- ❑ Service establishment funded by a swissuniversities 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



Current customers

A European RDM service for the academic community

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



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



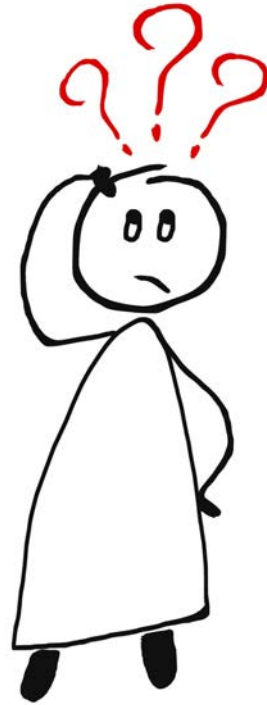
Training & 'best effort' user support

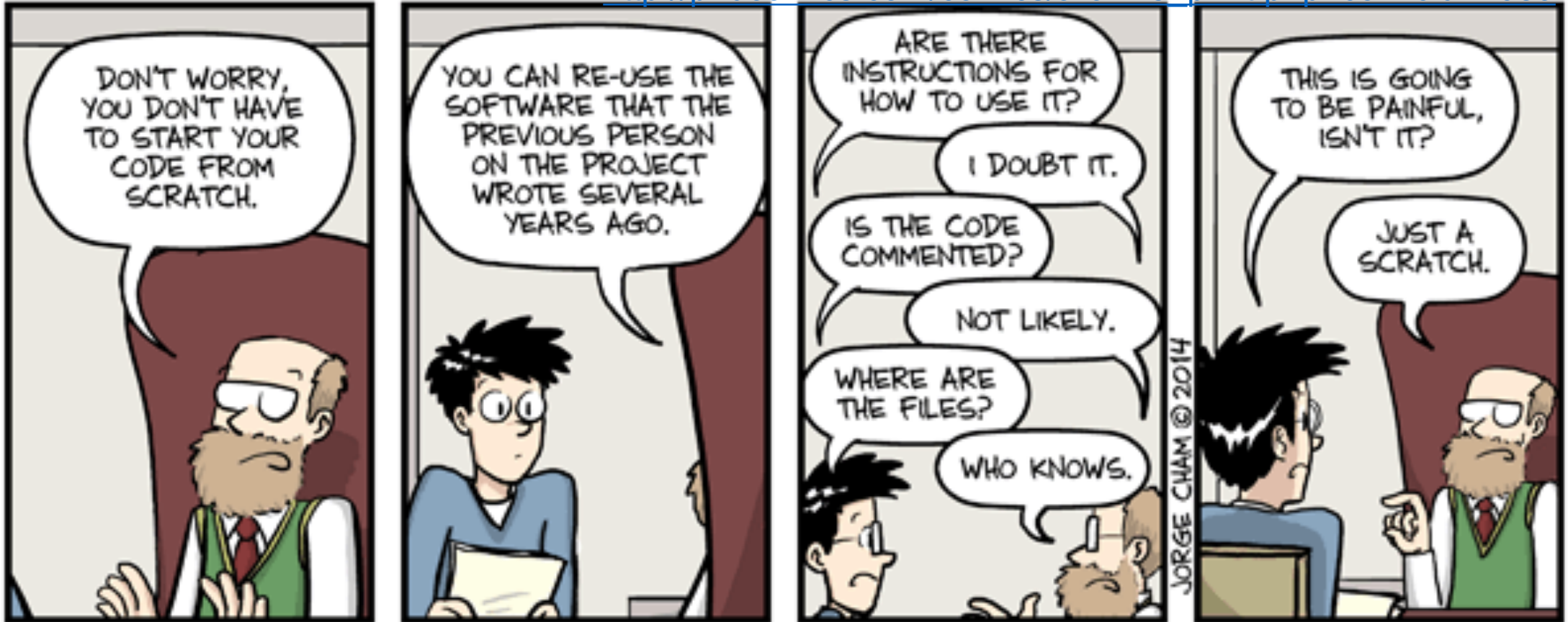


Current customers

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Questions on openBIS and RDM services provided by SIS?



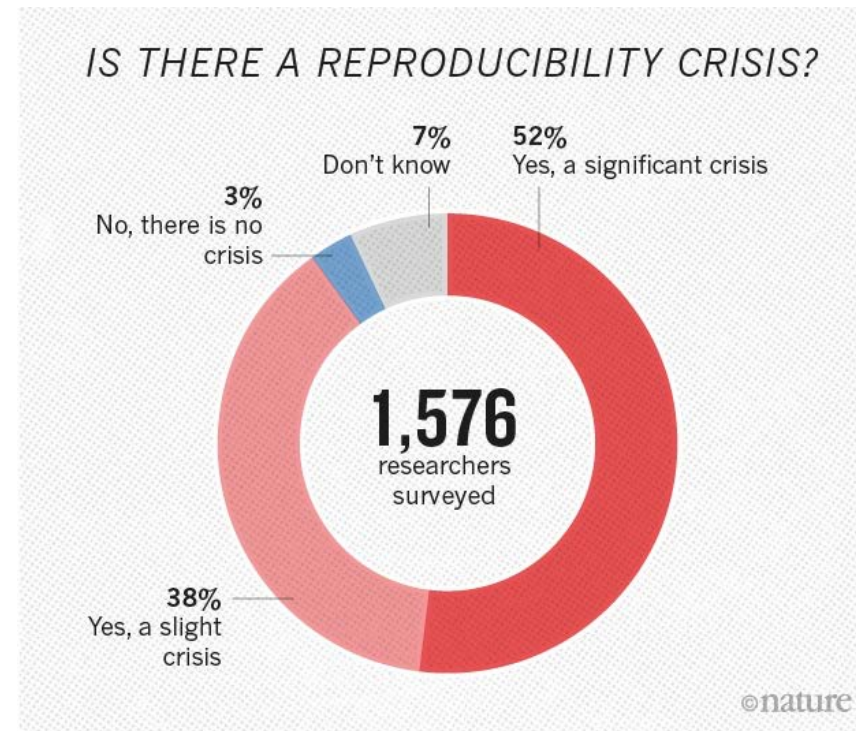
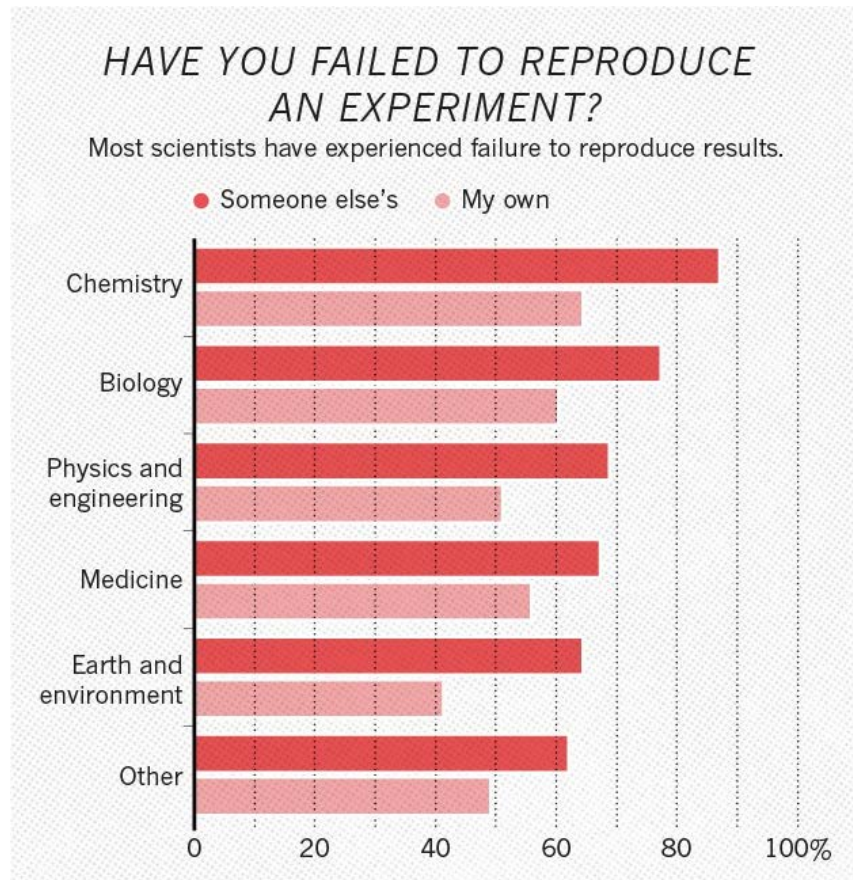


WWW.PHDCOMICS.COM

Reproducible Data Analysis

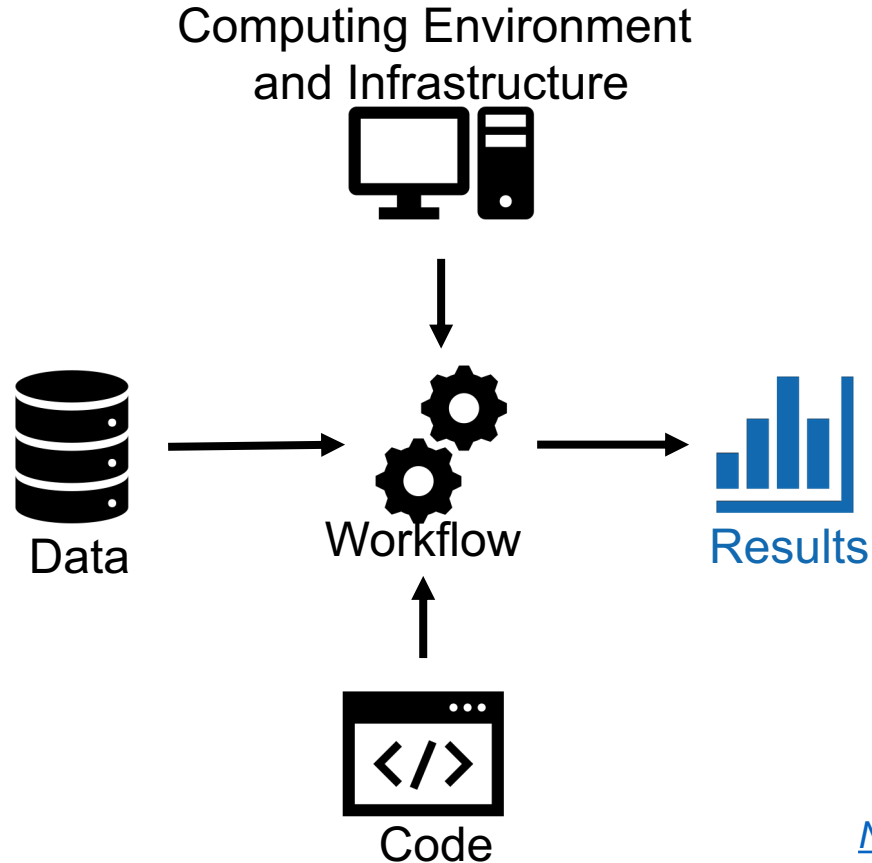
Why talk about Reproducibility?

Nature survey on reproducibility across all scientific domains



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

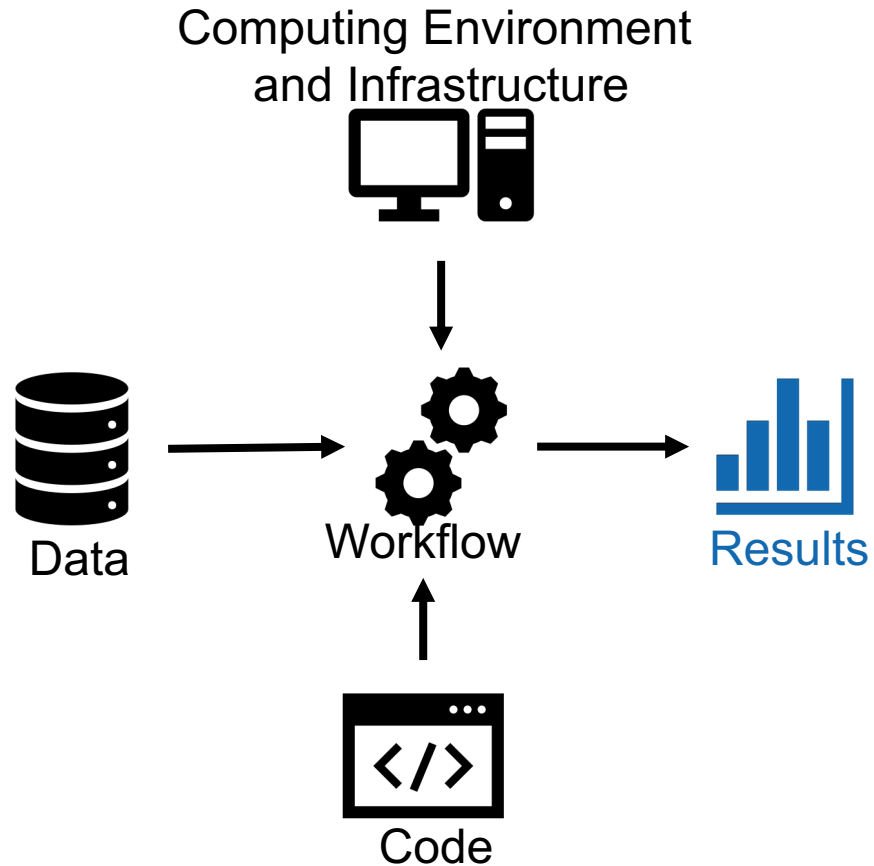
What do we mean by Reproducibility?



- 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>.

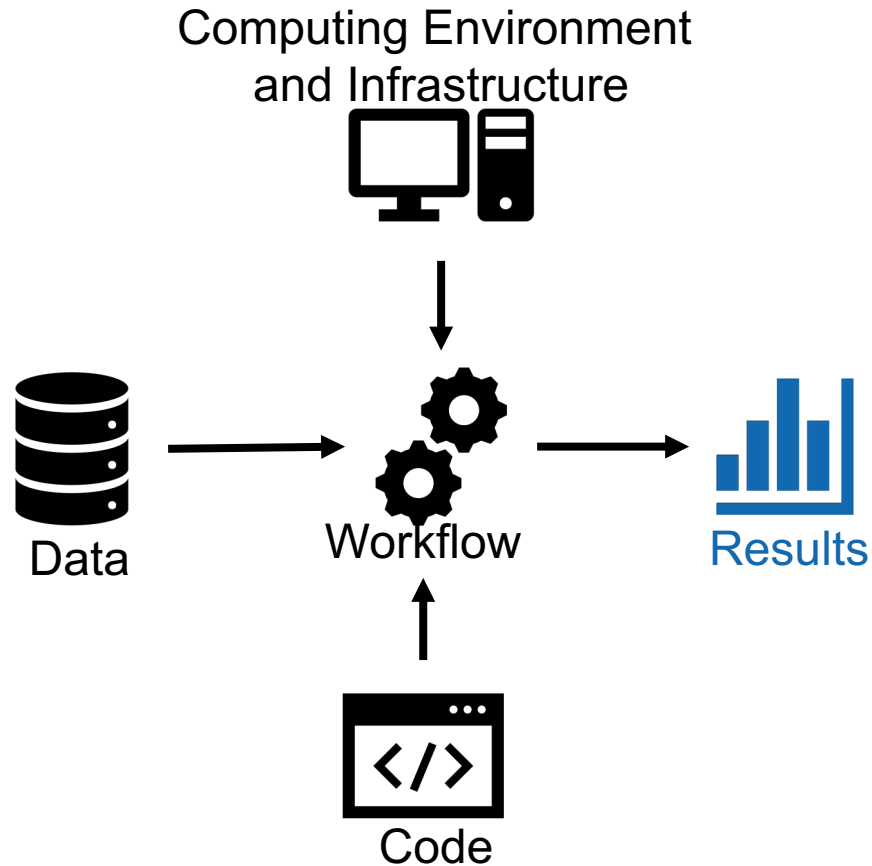
Components



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/>



<https://gitlab.ethz.ch>



<https://github.com>

Code Management

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

The screenshot shows the GitLab web interface for a project named 'experimental-project-1'. The browser address bar shows the URL <https://gitlab.ethz.ch/sis-rdm-training/experimental-project-1>. The page header includes the ETH zürich logo and navigation links for Projects, Groups, Activity, Milestones, and Snippets. A search bar is also present. A red banner at the top of the page contains a notification: 'To receive notifications about scheduled maintenance, please subscribe to the mailing-list gitlab-operations@sympa.ethz.ch. You can subscribe to the mailing-list at <https://sympa.ethz.ch>'.

The main content area shows the project details for 'experimental-project-1' (Project ID: 6107). It includes a 'Clone' button, 'Star' (0), and 'Fork' (0) options. Below this, it displays '7 Commits', '1 Branch', '0 Tags', and '9.5 MB Files'. The project description is 'My first experimental project'. The current branch is 'master'.

A commit history section shows a recent commit by Henry Luetcke, titled 'change', authored 4 months ago. The commit hash is 657f9d3a. Below the commit history, there are several buttons for adding project features: 'Add README', 'Add CHANGELOG', 'Add CONTRIBUTING', 'Enable Auto DevOps', and 'Add Kubernetes cluster'. There is also a 'Set up CI/CD' button.

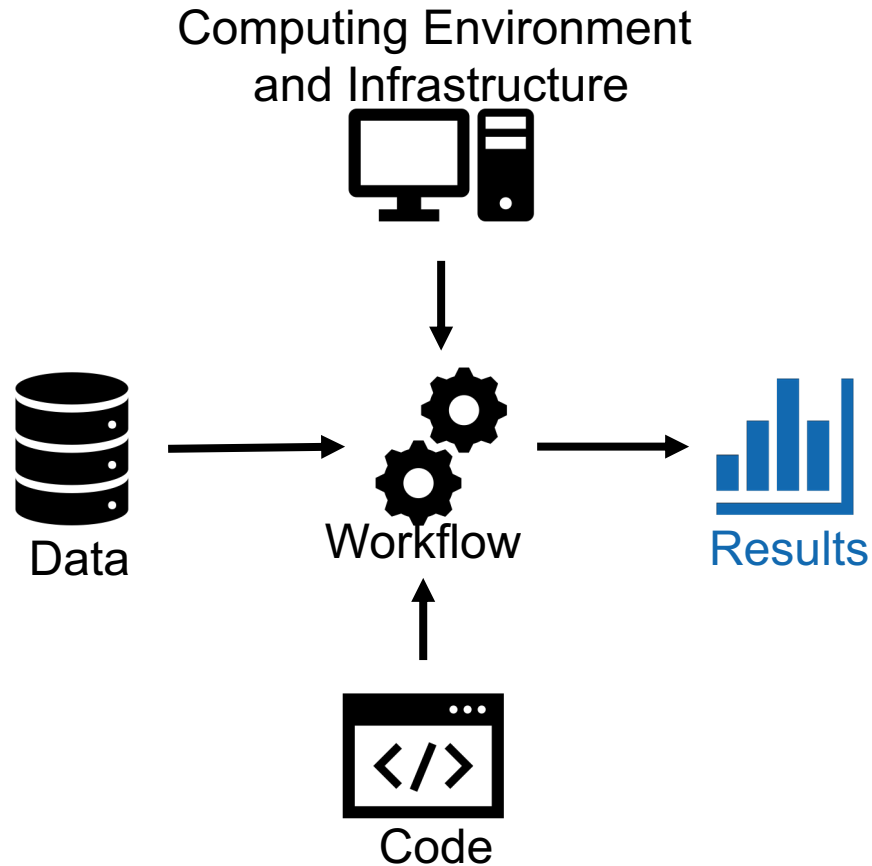
At the bottom, a table lists the files in the repository:

Name	Last commit	Last update
data	change image file size	4 months ago
.gitattributes	my first commit	5 months ago
analysis_code.py	change	4 months ago

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

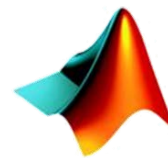
- Applications that combine documentation, code, input and output generated by the code, e.g. graphs, plots ([Nature 515, 151–152](#))
- Useful for exploratory data analysis and reproducibility



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



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



MATLAB

- Commercial
- Used in scientific, engineering, mathematical fields

Jupyter notebooks / JupyterLab

jupyter Jupyter_Image_Analysis_Notebook Last Checkpoint: câteva secunde în urmă (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Image Analysis with Python and Jupyter

This Jupyter notebook demonstrates the analysis of a simple image with Python.

First, make sure you have the image `blobs.tif` in the `data` folder of your notebook dashboard.

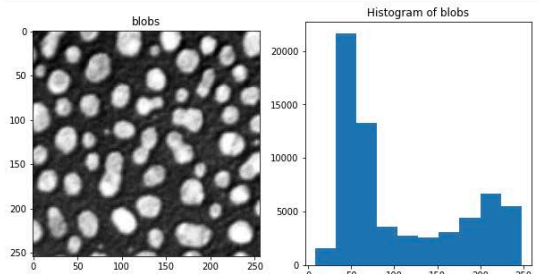
Import required modules

```
In [1]: import numpy as np
import matplotlib.pyplot as plt
from skimage.io import imread
from skimage.morphology import reconstruction, remove_small_objects
from skimage.measure import label, regionprops
%matplotlib inline
```

Read and display the data

```
In [2]: img = imread('data/blobs.tif')
```

```
In [3]: fig = plt.figure(figsize=(10,5))
fig.add_subplot(121)
plt.imshow(img, cmap='gray')
plt.title('blobs')
fig.add_subplot(122)
plt.hist(img.ravel(), bins=10)
plt.title('Histogram of blobs')
plt.show()
```



File Edit View Run Kernel Tabs Settings Help

Launcher Jupyter_Image_Analysis_N Python 3

Image Analysis with Python and Jupyter

This Jupyter notebook demonstrates the analysis of a simple image with Python.

First, make sure you have the image `blobs.tif` in the `data` folder of your notebook dashboard.

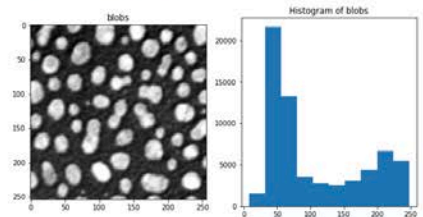
Import required modules

```
[1]: import numpy as np
import matplotlib.pyplot as plt
from skimage.io import imread
from skimage.morphology import reconstruction, remove_small_objects
from skimage.measure import label, regionprops
%matplotlib inline
```

Read and display the data

```
[2]: img = imread('data/blobs.tif')
```

```
[3]: fig = plt.figure(figsize=(10,5))
fig.add_subplot(121)
plt.imshow(img, cmap='gray')
plt.title('blobs')
fig.add_subplot(122)
plt.hist(img.ravel(), bins=10)
plt.title('Histogram of blobs')
plt.show()
```



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
AAAAA\wSFlz\nAAAWJQAAFiUBSVIk8AAAAIABJREFUeJzsvXeYZFd57b12h0maPNJII2lGOaCAKEBCFgozIxkBAp
lY\n1waDyDZg8MX+zMU2F4Mx1x8PwwAwxBjg4yNi2BfQMa20iiAQFkIjXKWRtJIE3tSz3TXuX+8vV2n\nnqyucv
N+9z/o9zzynprvq1D6nqqqr1prbRNFQgghhBCCCCGEEEEII8Zkh1wMghBCCCCGEEEEIIISQv\nnFLkIIYQQQgghhB
BCiPdQ5CKEEEEIIYQQQggh3kORixBCCCCGEEEEIIYR4D0UuQgghhBCCCCGEE0I9\nnFLkIIYQQQgghhBBCiPdQ5CK
EEEEIIYQQQggh3kORixBCCCCGEEEEIIYR4D0UuQgghhBCCCCGEE0I9\nnFLkIIYQQQgghhBBCiPdQ5CKEEEEIIYQQ
Qggh3kORixBCCCCGEEEEIIYR4D0UuQgghhBCCCCGEE0I9\nnFLkIIYQQQjzEGH0JMaZljPmo67EkZWq8D7keByGEE
ELChCIXIYQQQirDGP0mKaFj3BhzkMNx/H/G\nnmG3GmP/pagwFEbkeQJUY75gjNljHmD67EQQgghRB8UuQgghB
BSJe+DCDMjAH7L4TjeAmA+gLc5\nnHEMRGncDqJi3AVgI4DddD4QQQggh+qDIRQgghJBKMMacCuBMAFsg4sy7jTH
DjobzZwBuBvBxR/dP\nnsvERADcC+LTrgRBCCCFEHxS5CCGEEFIVH4C4uP4SI\lQcBOD1LgYSRVEz iqIXR1H0fRf3
T7IRRDfF\nnRlH0K1EUXe96LIQQQgjRB0UuQgghhJSOMWYpgP8BoAXg7wH8HcTN9Tsux0UIIYQQQsKBhchhBBC\
nqyBd4QYAuDyKocBfBvAlrBnGwQe73PlhBCCCCFkCChvEIIITTTtPIHdFGUitf0RyciK7nDMR3n65C\nnNychhBBC
```


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

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



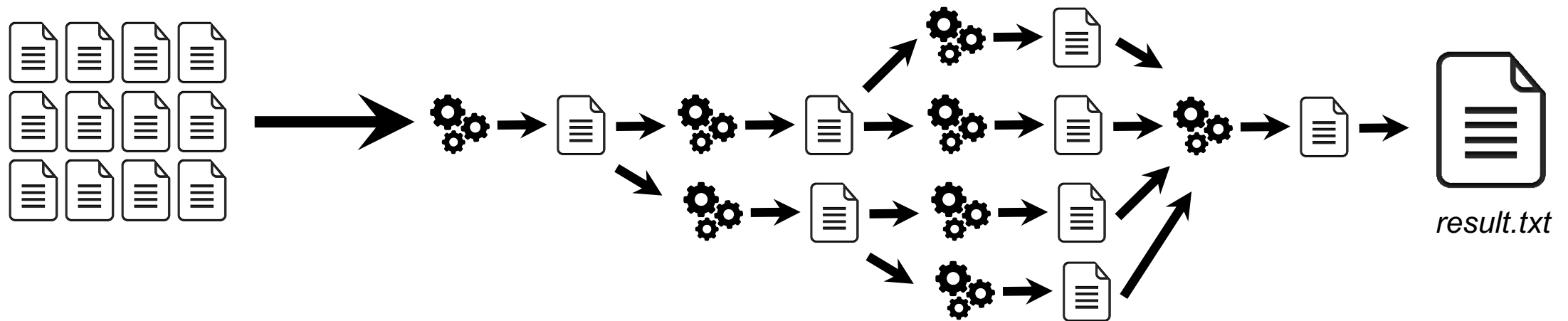
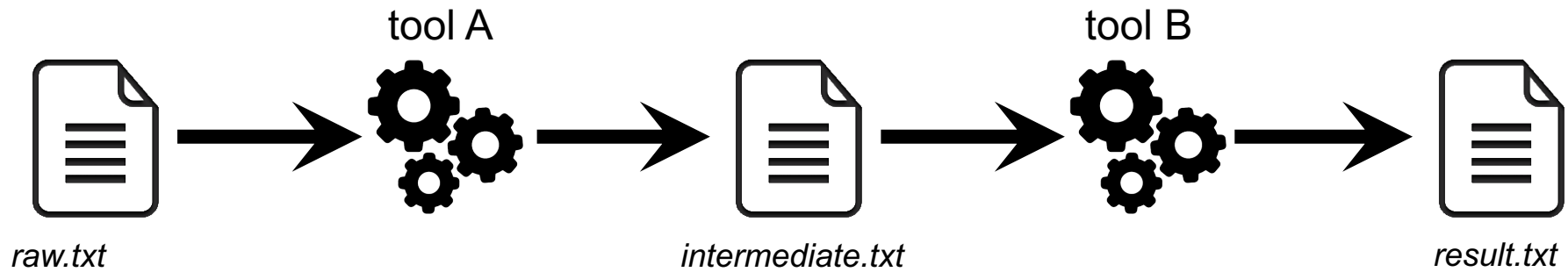
COMMON
WORKFLOW
LANGUAGE



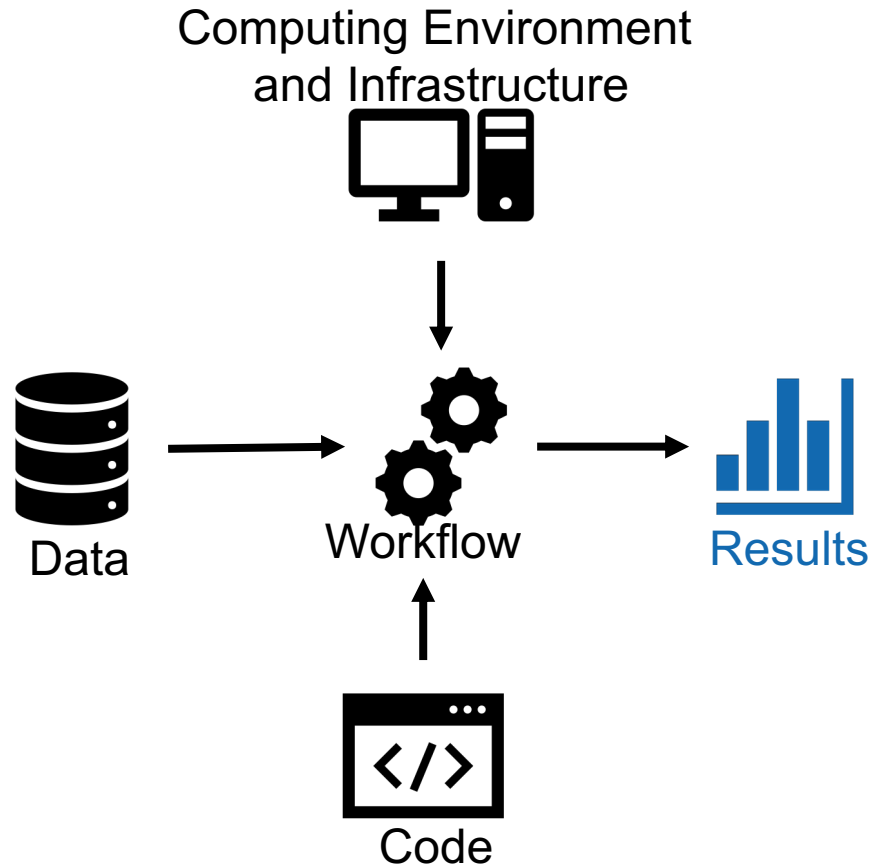
Airflow

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

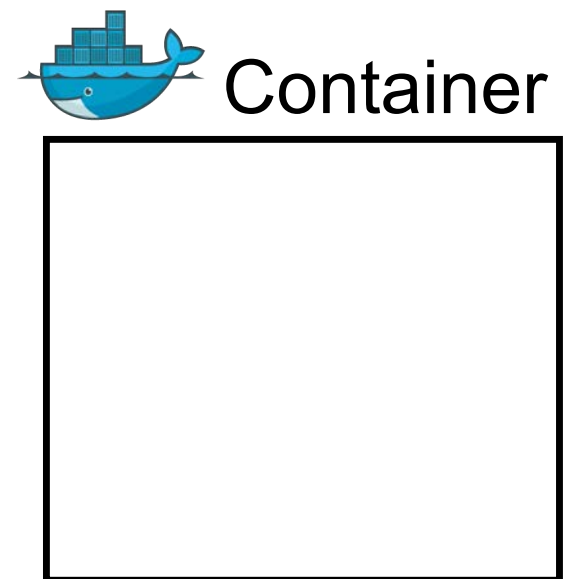
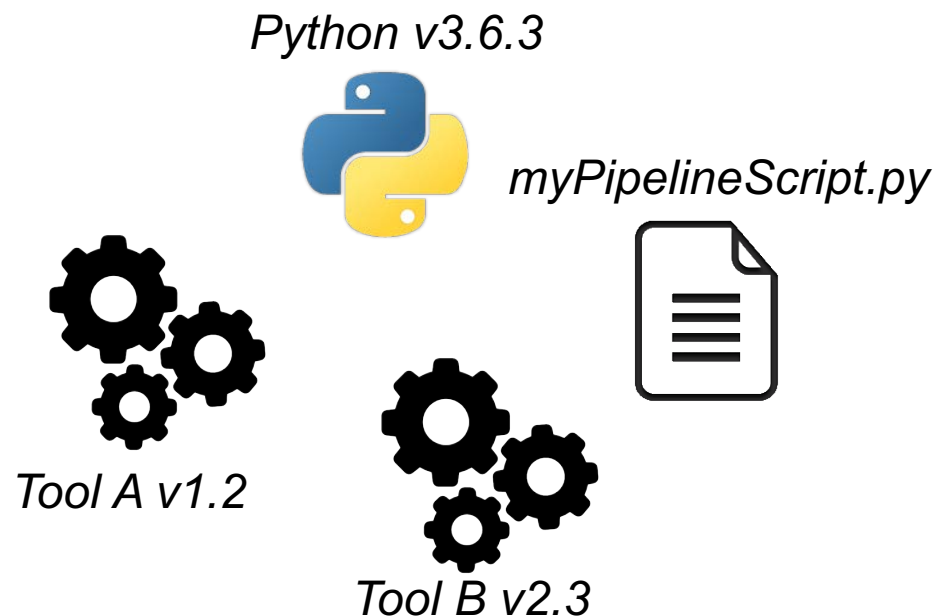
Virtual machines (VMs)

- 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>

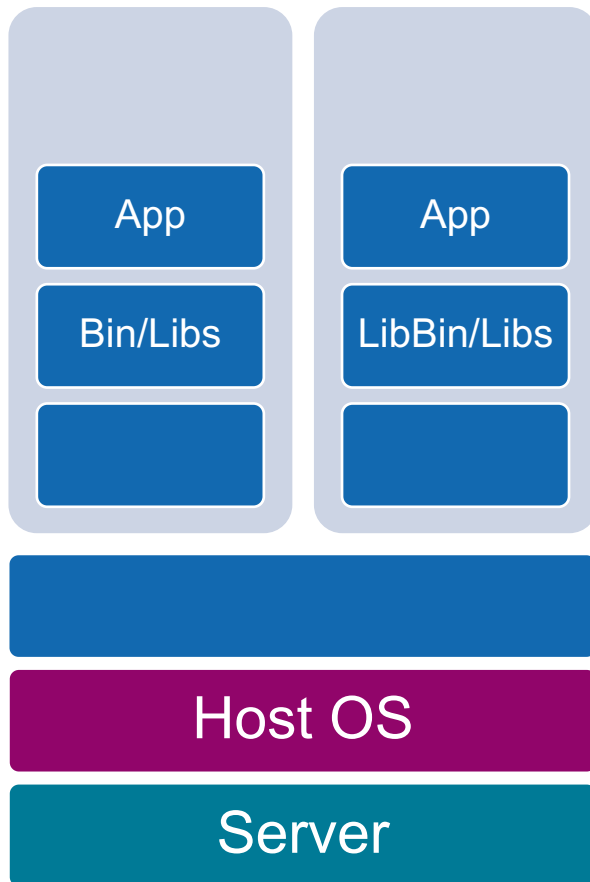
Containers

- Container: Operating system level **virtualization method** 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

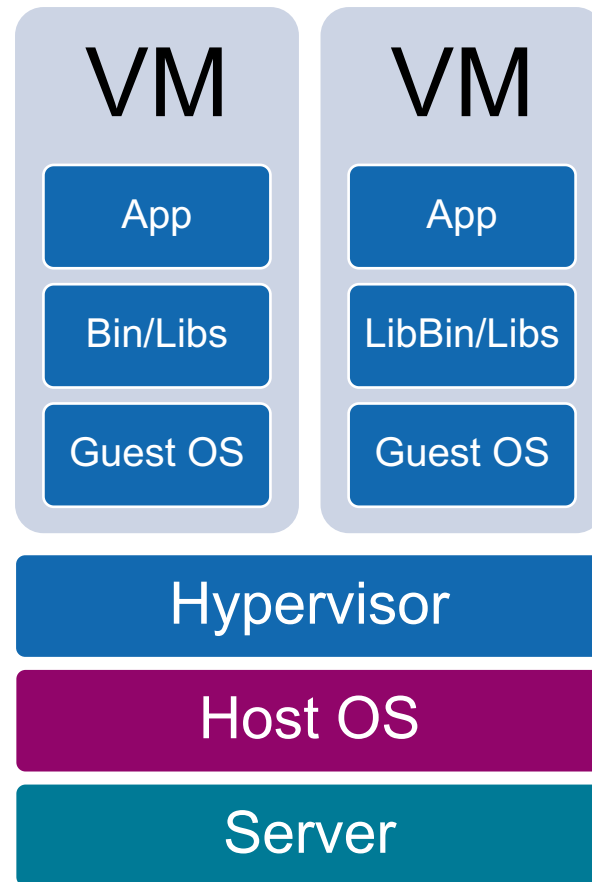


Bare Metal, Virtual Machine (VM) and Container (Docker)

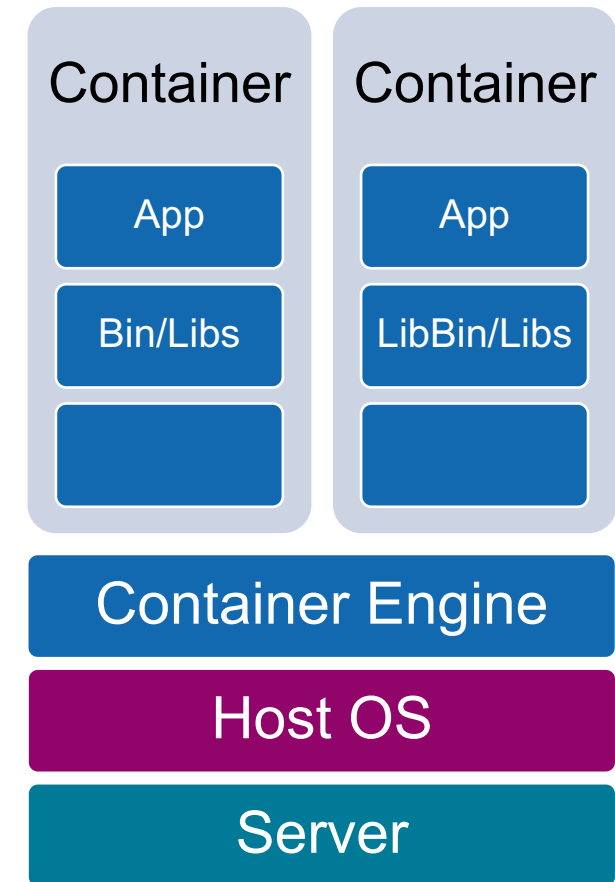
Bare Metal



VM Based



Container Based Shared Host OS kernel



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

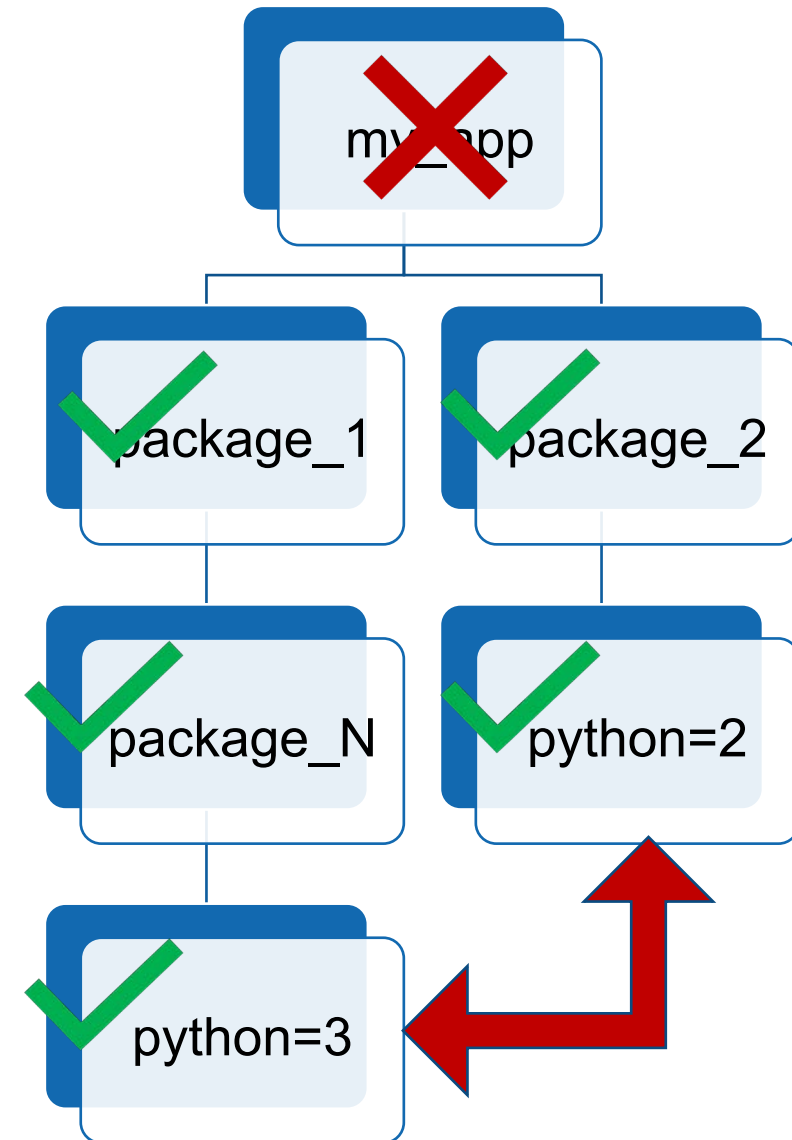
Containers - What can go wrong?

- 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)

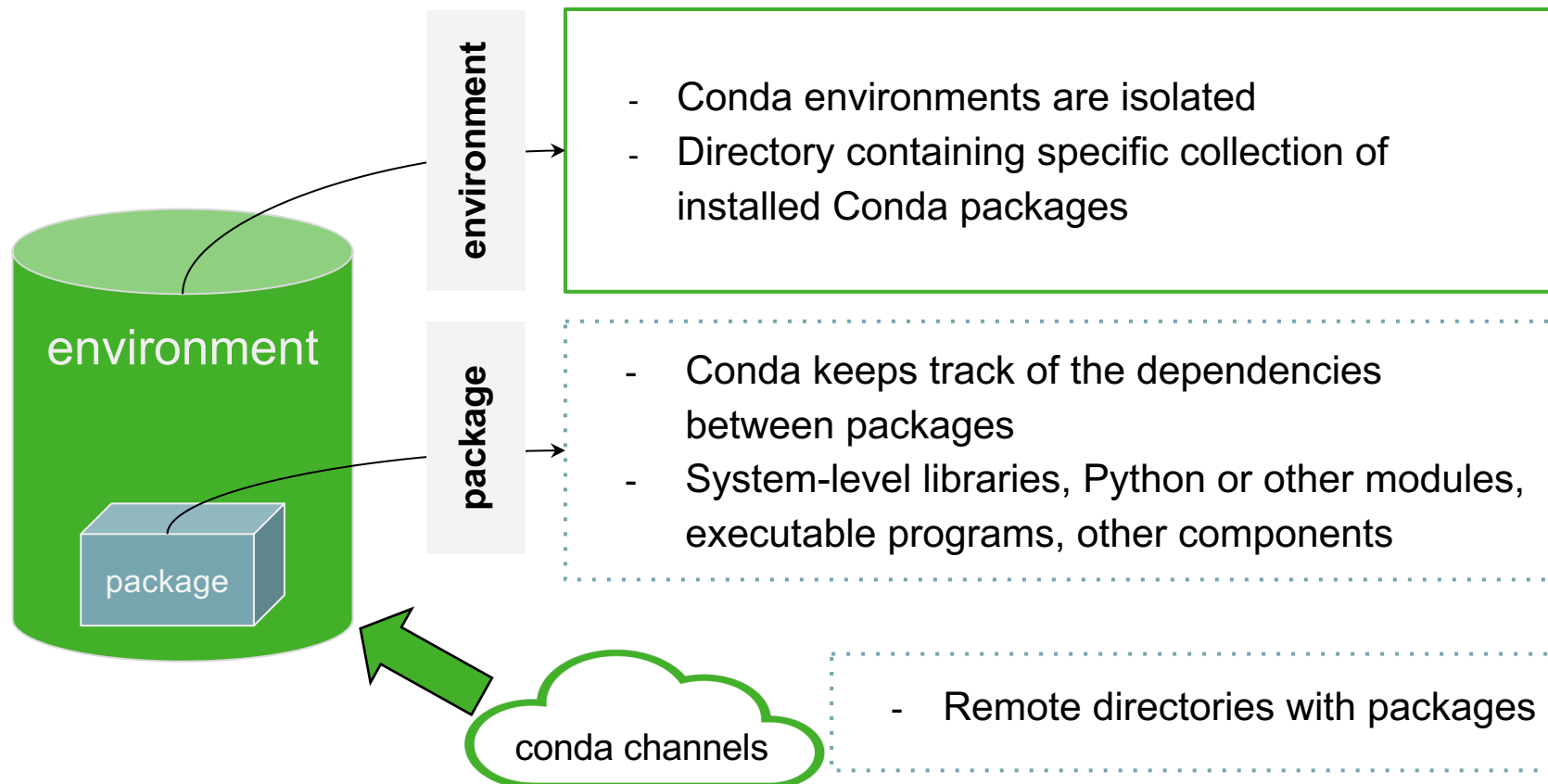
Reproducible Environment for R and Python



- 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.yml

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

Conda automatically creates an environment file with packages and dependencies

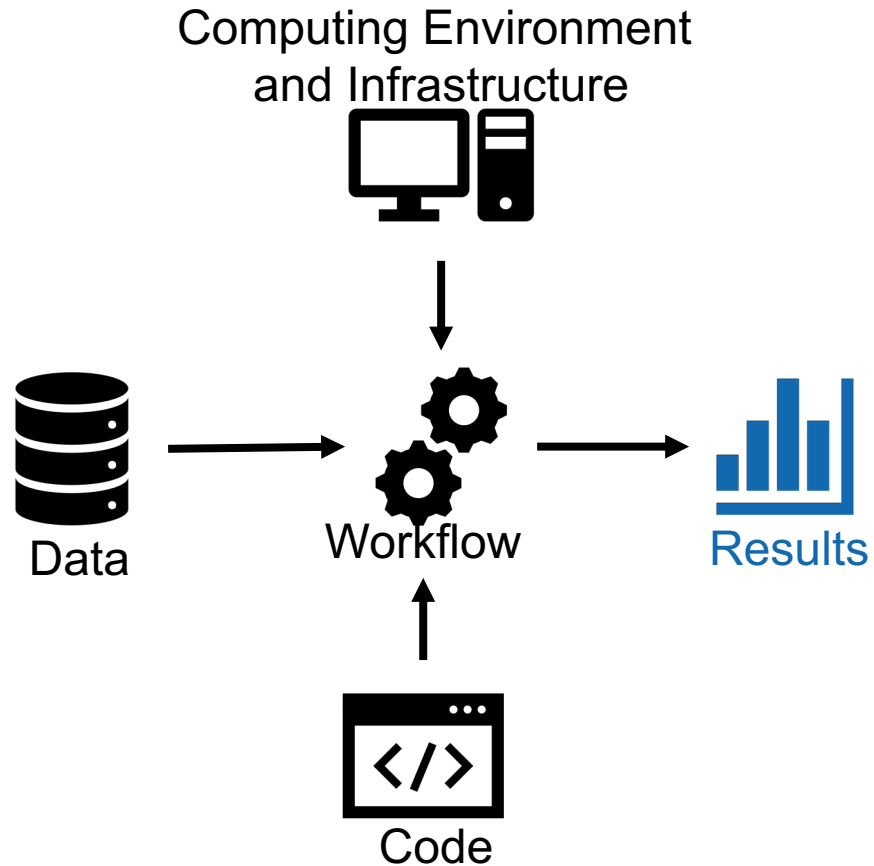
Conda - What can go wrong?

- 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



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:

Copy the text below, then paste into your README to show a binder badge: [launch binder](#)



Renku Projects

Search or jump to...

Projects Datasets Environments

virginia User Order by: Update date

- virginiafriedrich/form-refactoring-tests-2-copy
- virginiafriedrich/form-refactoring-test-3
- virginiafriedrich/form-refactoring-tests-2
- 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 Jun 2018
- ENGINEERING Jun 2018
- MEDICAL SCIENCES May 2018
- SOCIAL SCIENCES Aug 2018

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

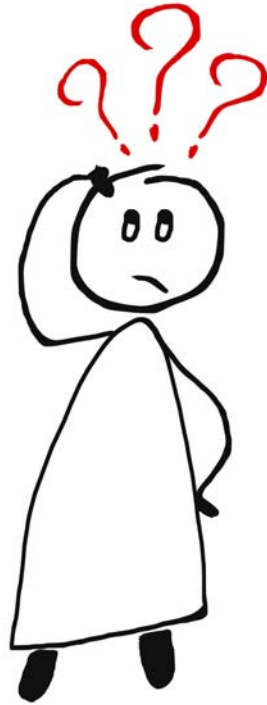
documented. Every step of the way, we want to anticipate what another researcher might need to either reproduce our results (run our code with our data) or replicate them (independently arrive at the same findings).

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.”

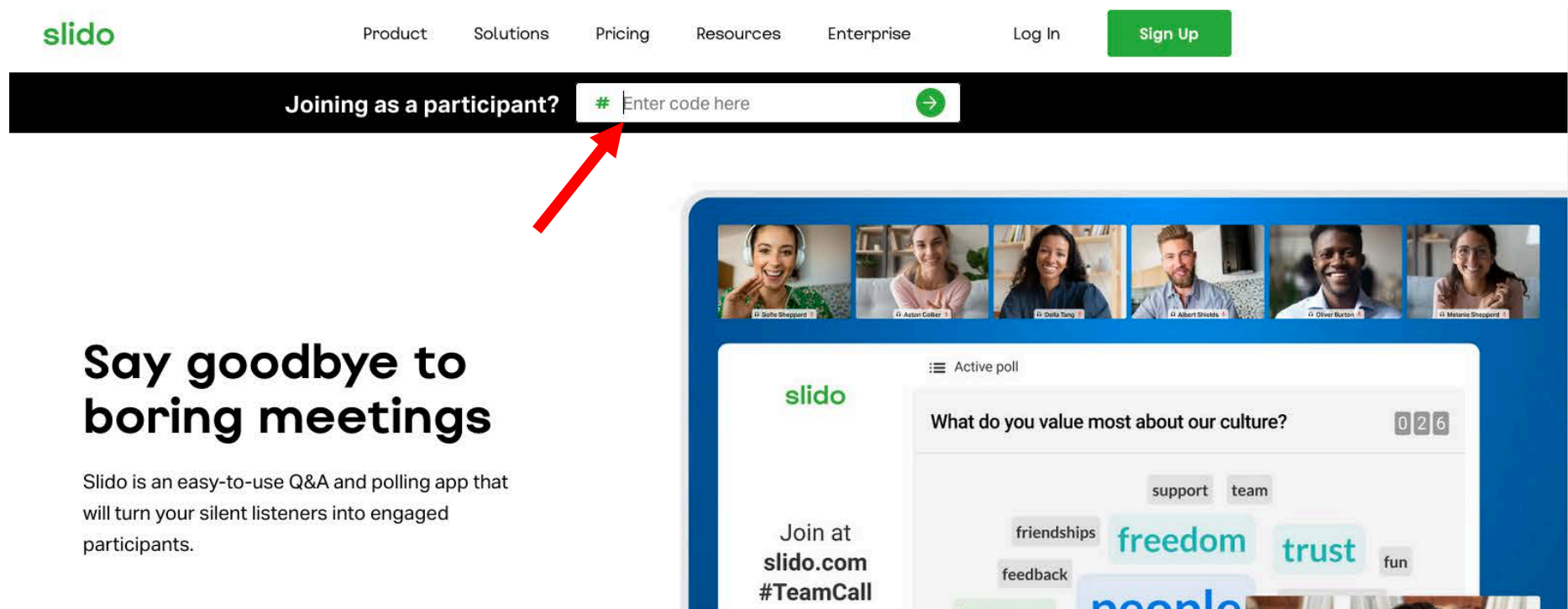
<https://doi.org/10.1126/science.354.6308.142>

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



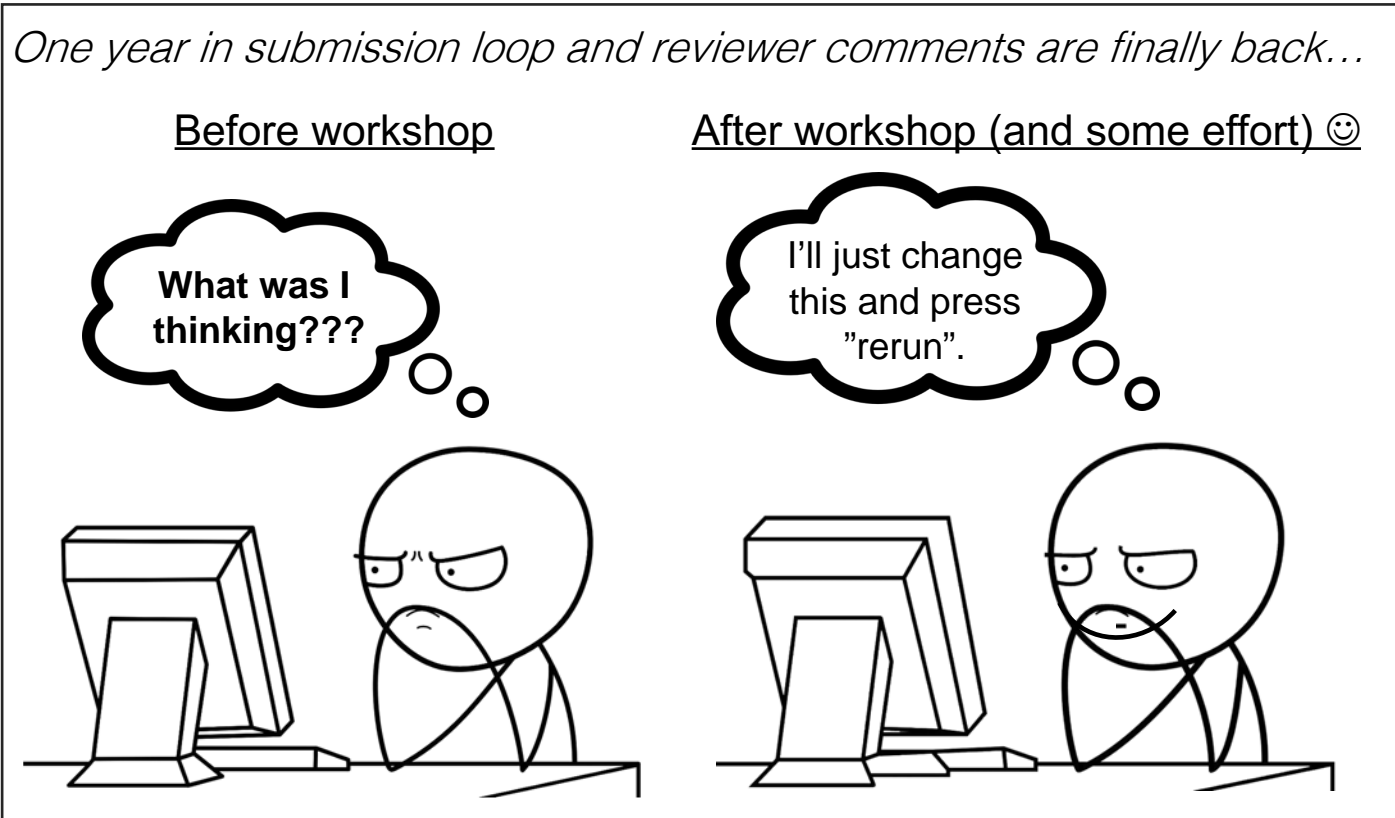
The screenshot shows the Slido website interface. At the top, there is a navigation bar with the Slido logo and links for Product, Solutions, Pricing, Resources, Enterprise, Log In, and a green Sign Up button. Below this is a black banner with the text "Joining as a participant?" and a white input field containing "# Enter code here" and a green arrow button. A red arrow points to the input field. Below the banner, there is a blue section with a grid of six video thumbnails showing participants. Below the thumbnails, there is a white box with the Slido logo and the text "Active poll". The poll question is "What do you value most about our culture?" with a counter of "0 2 6". Below the question, there are several word cloud items: "support", "team", "friendships", "freedom", "trust", "fun", "feedback", and "people". To the left of the poll, there is text that says "Join at slido.com #TeamCall".

Say goodbye to boring meetings

Slido is an easy-to-use Q&A and polling app that will turn your silent listeners into engaged participants.

Join at
slido.com
#TeamCall

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

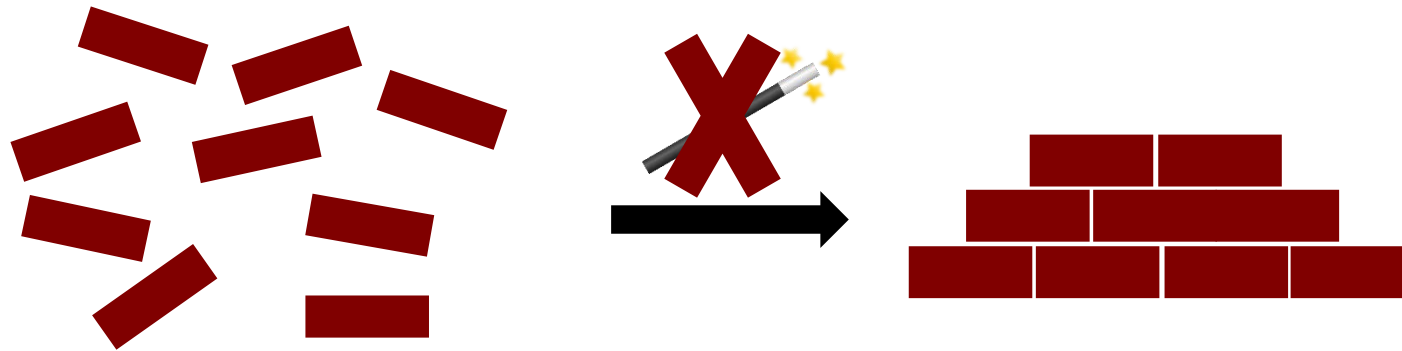
- Easier to rerun experiments and analysis
- Closer interaction between collaborators.
- Much of the manuscript "writes itself".

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.

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

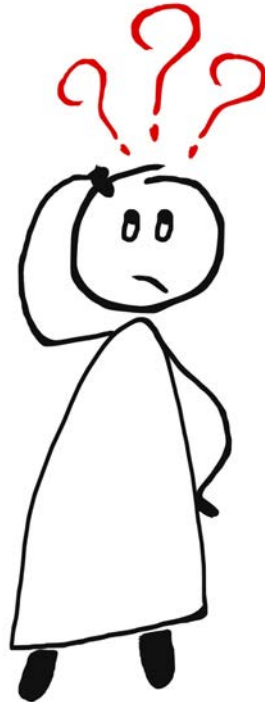
sis.helpdesk@ethz.ch

<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>