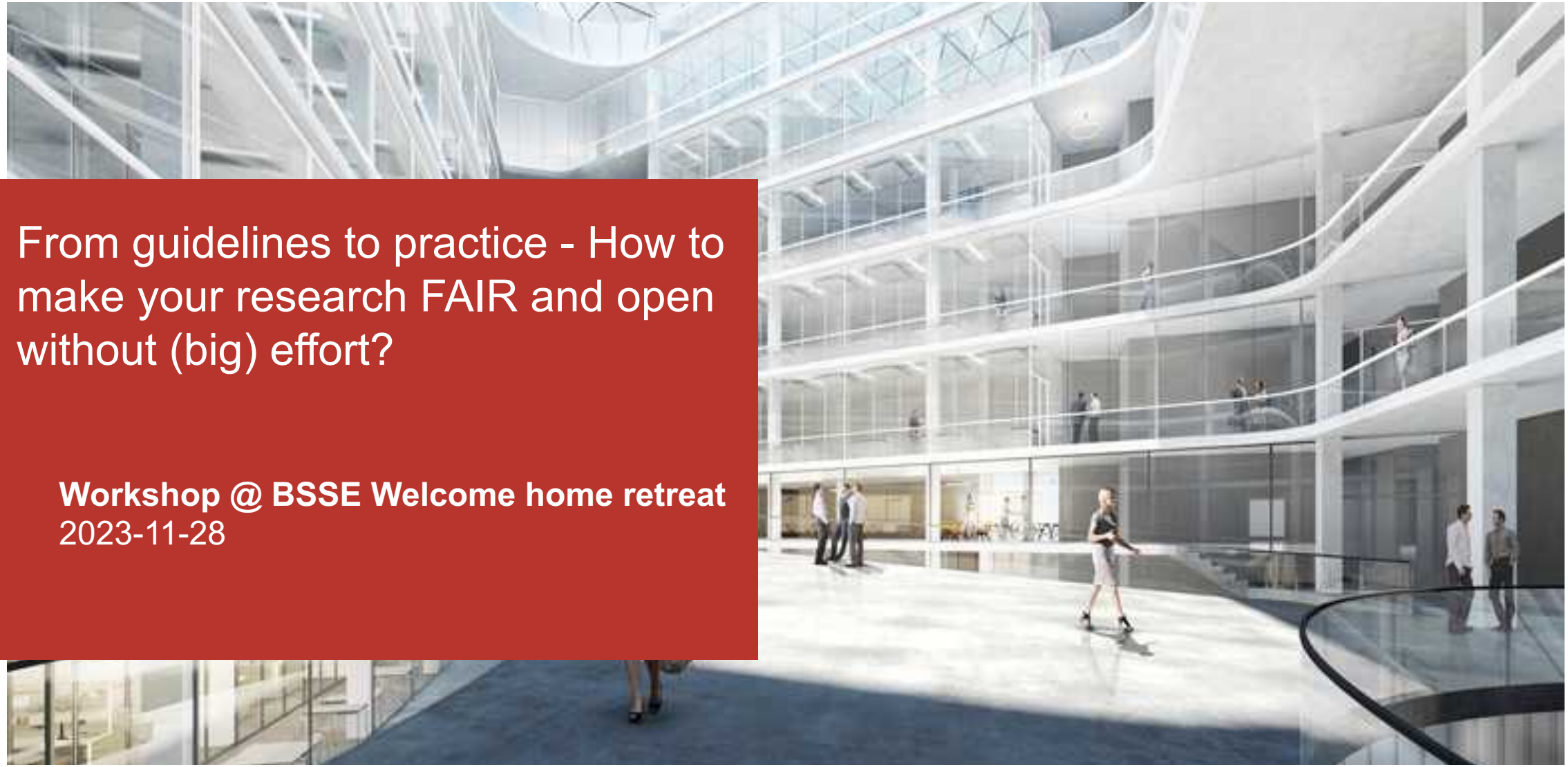
ETH zürich

nexus
Personalized Health Technologies



From guidelines to practice - How to make your research FAIR and open without (big) effort?

Workshop @ BSSE Welcome home retreat
2023-11-28

Workshop Agenda

1. Welcome and Intro
2. Overview ETH ORD, concept of FAIR data, and RDMS
3. Experiences and challenges at the D-BSSE
4. Small steps with a big effect: documentation, version control, docker
5. Wrap-up

Intro

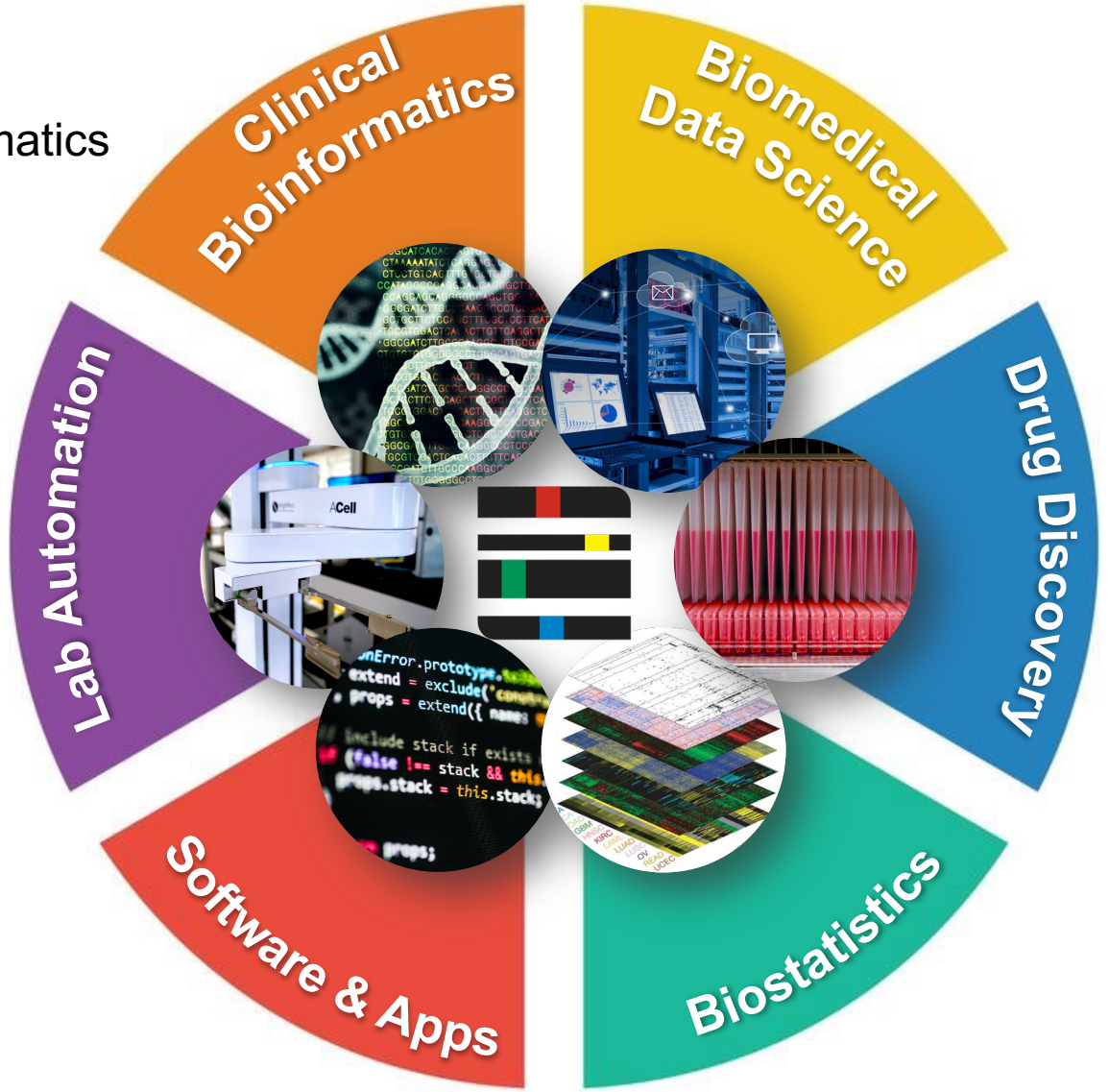
An ETH technology platform to enable and accelerate the execution of R&D projects for the personalized health community

nexus
Personalized Health Technologies

Franziska Singer
Deputy Director & Head Clinical Bioinformatics

Matteo Carrara
Senior Bioinformatics Scientist

David Meyer
Head Software Engineering



Workshop Agenda

1. Welcome and Intro
2. **Overview ETH ORD, concept of FAIR data, and RDMS**
3. Experiences and challenges at the D-BSSE
4. Small steps with a big effect: documentation, version control, docker
5. Wrap-up

Findable Accessible Interoperable Reusable

Guidelines to ensure the **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability of research data

Findable

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata
- Metadata clearly and explicitly include in the identifier of the data it describes
- (Meta)data are registered or indexed in a searchable resource

Accessible

- (Meta)data are retrievable by their identifier using a standardized protocol
- The protocol is open, free and universal
- The protocol allows for authentication and authorization, as needed
- Metadata are accessible, even when the data are no longer available

Interoperable

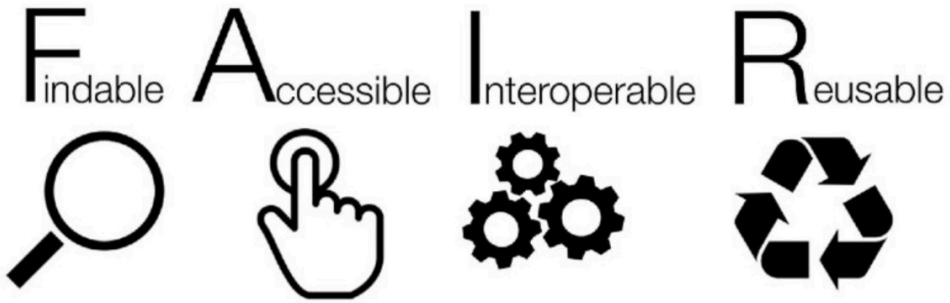
- (Meta)data use a formal, accessible, shared and broadly applicable language
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

Reusable

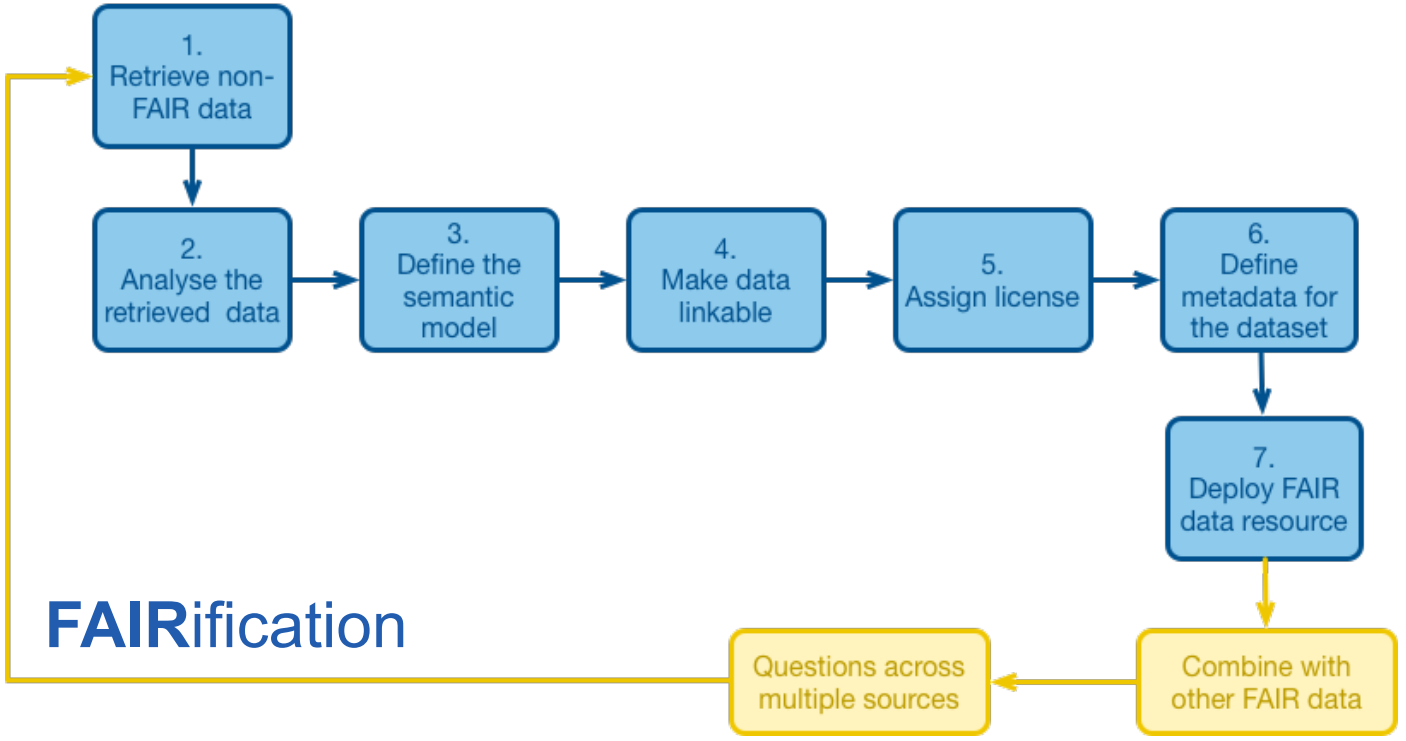
- (Meta)data are richly described with a plurality of accurate and relevant attributes
- (Meta)data are released with a clear and accessible data usage licence
- (Meta)data are associated with a detailed provenance
- (Meta)data meet domain-relevant community standards

Adapted from: https://www.ccdc.cam.ac.uk/solutions/about-the-csd/fair-data-principles

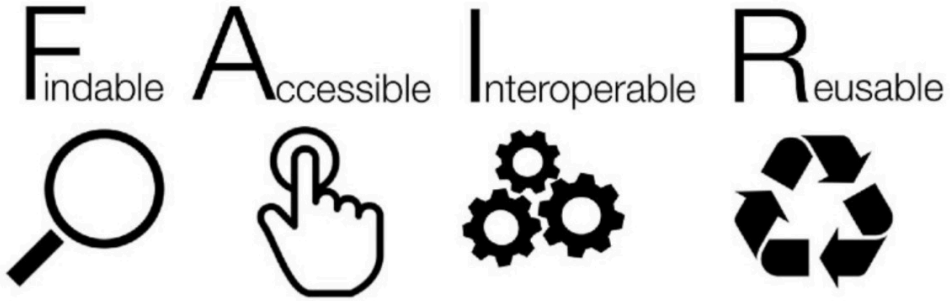
- Wilkinson et al., 2016, Scientific Data: “The FAIR Guiding Principles for scientific data management and stewardship”
- https://www.go-fair.org/fair-principles/



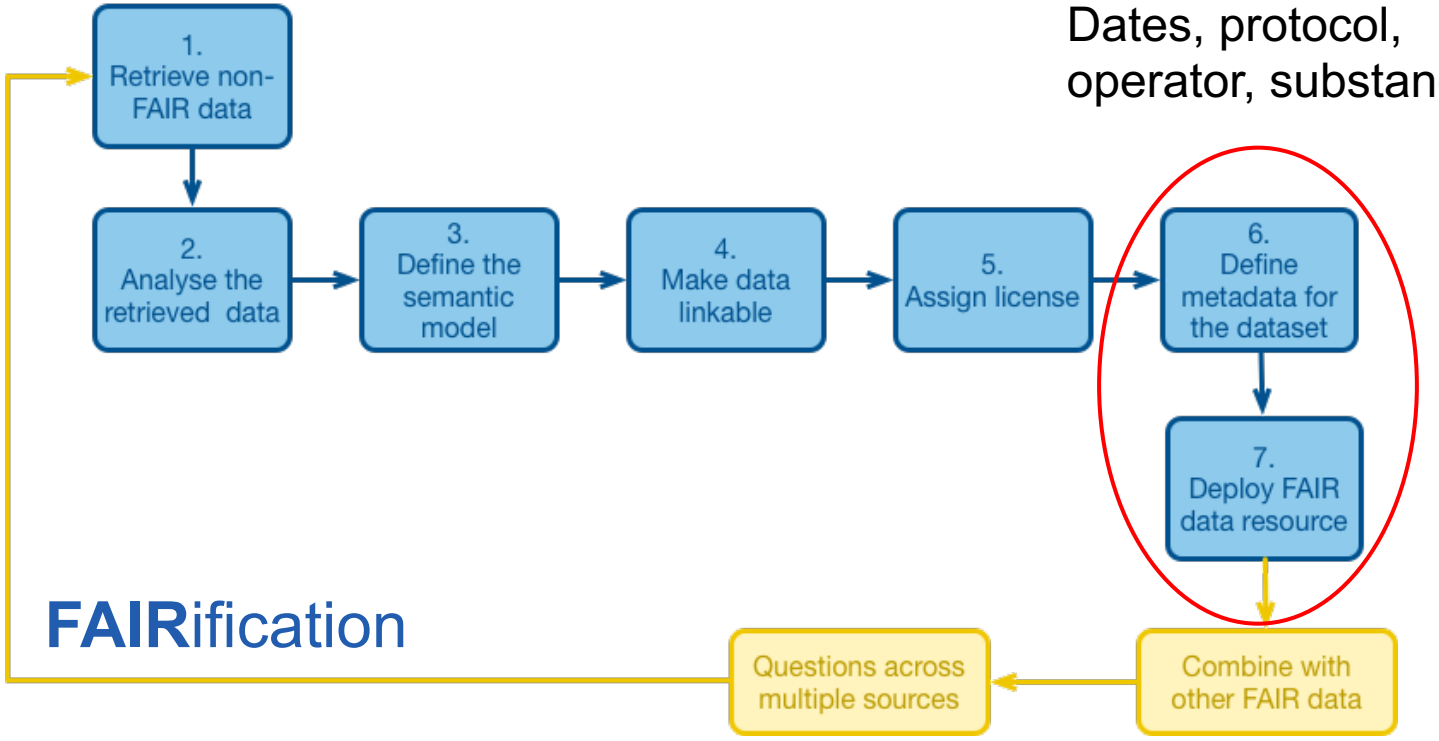
Guidelines to ensure the **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability of research data



- Wilkinson et al., 2016, Scientific Data: "The FAIR Guiding Principles for scientific data management and stewardship"
- https://www.go-fair.org/fair-principles/



Guidelines to ensure the **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability of research data



Dates, protocol, operator, substances,….

zenodo

ENA European Nucleotide Archive

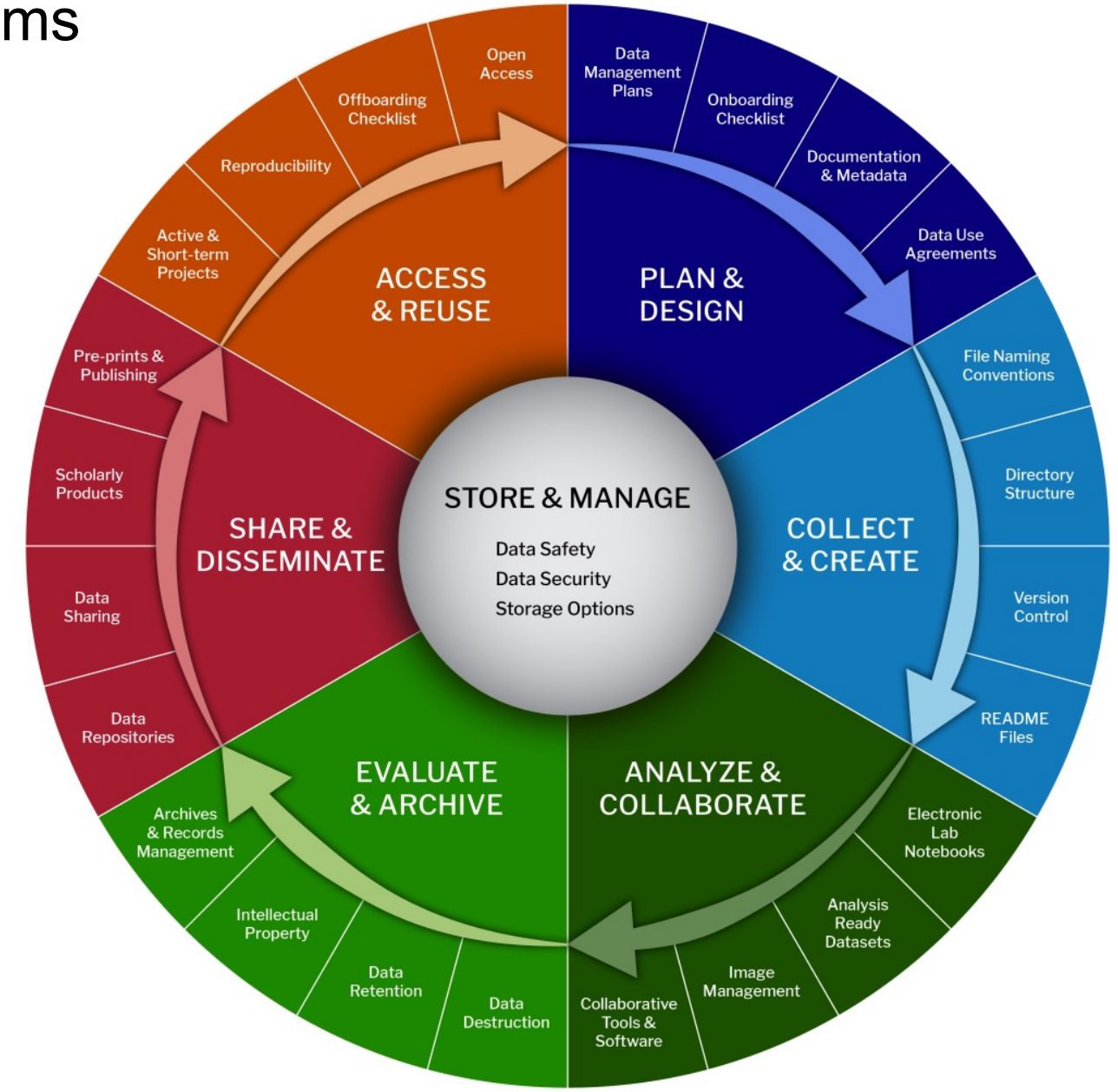
GitHub

- Wilkinson et al., 2016, Scientific Data: “The FAIR Guiding Principles for scientific data management and stewardship”
- https://www.go-fair.org/fair-principles/

Research **D**ata **M**anagement **S**ystems

Simplify:

- **Publication of work**: enabling straightforward long-term storage for data and methods
- **Writing of data management plans**: providing project-specific templates
- **Sharing data with established resources**: building appropriate interfaces
- **Confidentiality classification**: facilitating the required process of classification

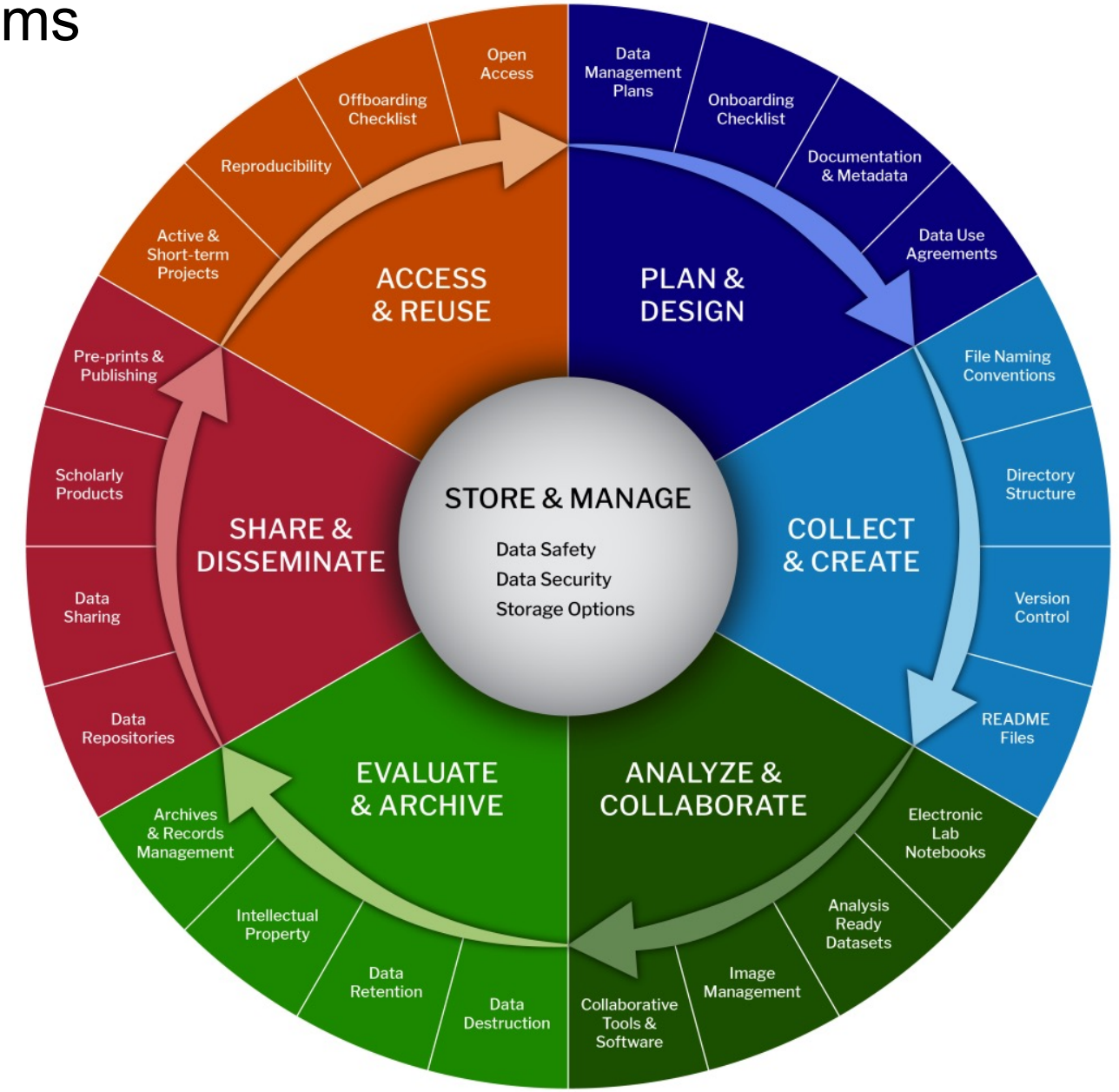


[https://uwyo.libguides.com/data-mgmt]

Research **D**ata **M**anagement **S**ystems

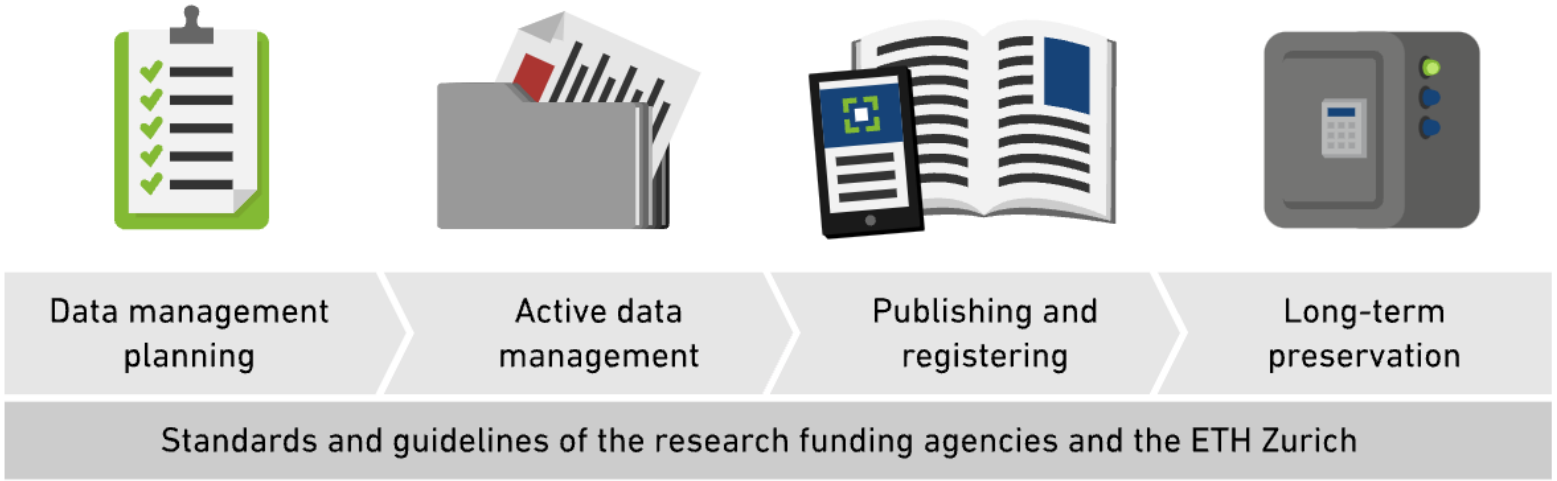
Ensure:

- **FAIR data**: making data findable and accessible, standardization ensures interoperability, and stringent documentation ensures reusability
- **Reproducible research**: allowing to put FAIR data into context of their workflows and methods
- **Data safety and security**: putting data in the right place with the right setup
- **Compliance with ETH / funding agency rules** on storing data from publications and on data management plans



[https://uwyo.libguides.com/data-mgmt]

Open Research Data (ORD) programme in the ETH Domain
Five measures to foster and improve ORD Practices



- ETH Domain mandated by the Federal Council to assume a leading role in respect of ORD[1].
  - Adopt emerging open science practices that allow research outputs –including publications, data, and software - to be disseminated and made accessible.
  - Research data management as an integral part of the research process[2].
- E.g., ETHZ research collection[3], Calls for field-specific actions[4]

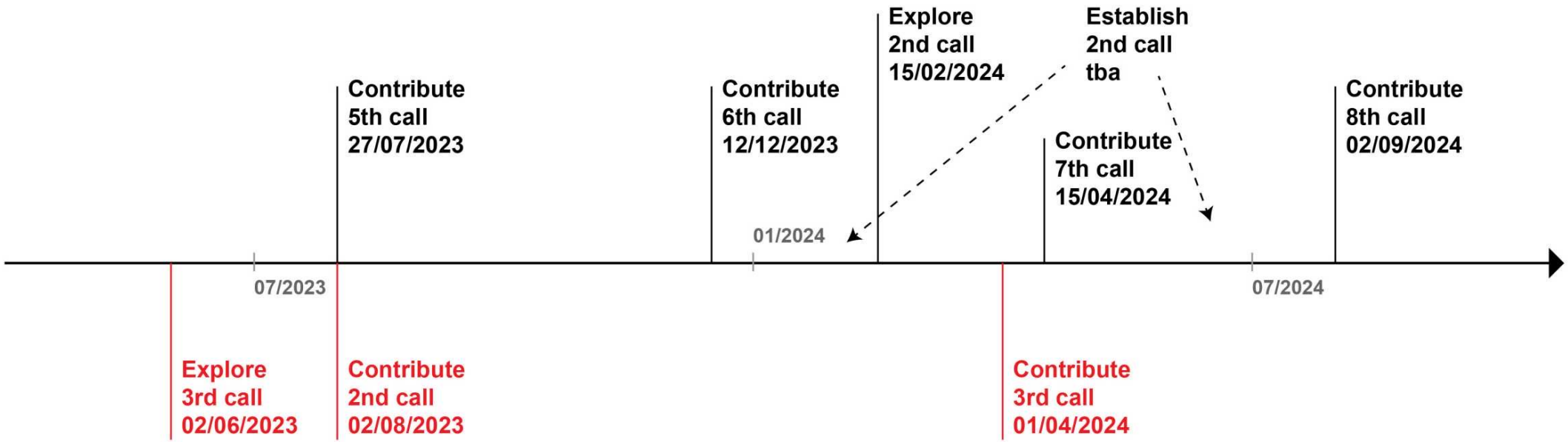
[1]https://ethrat.ch/en/eth-domain/open-research-data/

[2]https://library.ethz.ch/en/researching-and-publishing/data-management-and-policies/research-data-management.html

[4]https://www.research-collection.ethz.ch/

[5]https://ethrat.ch/en/measure-1-calls-for-field-specific-actions/

ORD Programme ETH Domain



ORD Programme swissuniversities

Workshop Agenda

1. Welcome and Intro
2. Overview ETH ORD, concept of FAIR data, and RDMS
3. **Experiences and challenges at the D-BSSE**
4. Small steps with a big effect: documentation, version control, docker
5. Wrap-up

What parts do we want to address?

What parts <u>can</u> we address?

What works already?

What are the challenges?

[https://uwyo.libguides.com/data-mgmt]

Workshop Agenda

1. Welcome and Intro
2. Overview ETH ORD, concept of FAIR data, and RDMS
3. Experiences and challenges at the D-BSSE
4. **Small steps with a big effect: documentation, version control, docker**
5. Wrap-up

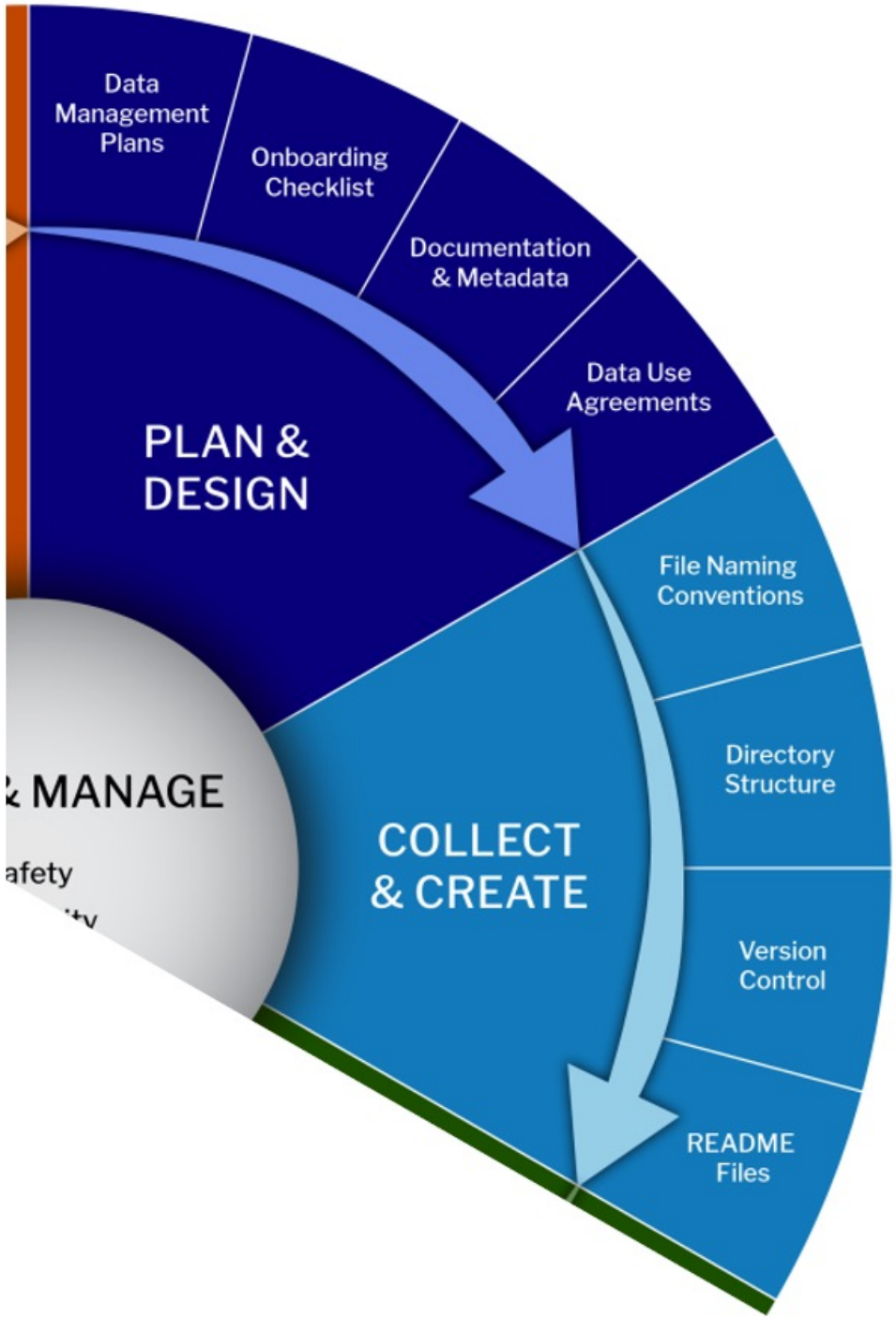
Documentation

File naming conventions & Directory Structure:

- At project initialization, all projects look the same
- (Empty files and folders are OK)
- Always add a README file with basic info (e.g., lead, point-of-contact, scope, resources…)

Version control:

- Resource mgmt.: shared directory for large resources (e.g., gene annotation files, reference sequences)
- Small resources (e.g., list of interesting genes): project-specific directory, or ideally git
- Document data retrieval (access date, webpage, …)
- Track code modifications (git or readme)
- Track data modification (lakefs, ...)



Version control

Every project should have a dedicated git repository that hosts e.g.:

- Code, analysis scripts, licenses
- Documentation file
- (Small) input files (e.g. list of genes of interest)

NOT to be stored in a git repository:

- Large files, e.g., sequencing data or image data

How To: Both command line as well as web-based access are possible.

Version control - HowTo

Checkout the workshop repository

```
➢ git clone https://github.com/ETH-NEXUS/Workshop_RDMS_BSSE_retreat.git .
```

Get basic info about the repository

```
➢ git status
```

Create your own branch (replace [yourName])

```
➢ git checkout –b [yourName]
```

Create a new file for testing and check the status of your repo again

```
➢ Touch newFile.txt
➢ git status
```

Add the new file to the next commit and then commit to git, including a commit message

```
➢ git add newFile.txt
➢ git commit -m “Added new file for testing.”
```

Push the changes to remote

```
➢ git push –u origin [yourName]
```

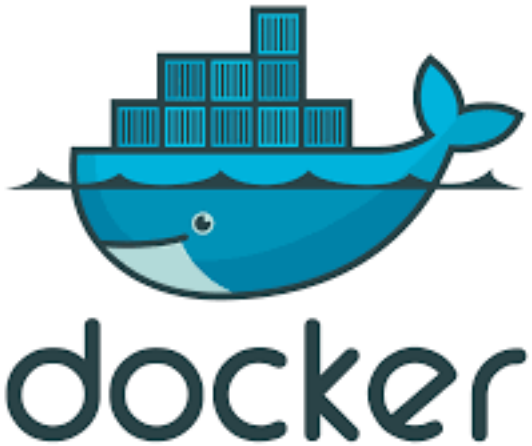
Containers

What are containers?

All software necessary for a certain analysis is packaged into a “file” that can be used to setup the exact compute environment in future analysis.

Why are they useful?

They help you and the community to easily **reproduce** an analysis. You activate the container and it creates the exact environment you need to perform the analysis again.
Port your software to many different systems and allow easy **sharing** of your software with others.



What is availble?
Depot.nexus.ethz.ch -> resource for downloading images

Containers

How to setup a container?

```
➢ Rscript myAnalysis.R --input input.txt --output plot.png

(with R version 4.3.2 and libraries ggplot2 v3.4.4, argparse v2.2.2)
```

Example

```
• FROM condaforge/miniforge3:23.3.1-1
• RUN mamba install -y r-base=4.3.2 r-ggplot2=3.4.4 r-argparse=2.2.2
• ENTRYPOINT ["Rscript", "/input/myAnalysis.R", "--input",
  "/input/input.tsv", "--output", "/output/plot.png" ]
```

Receipe

```
➢ docker build -t docker_example .
```

Create the image

How to use a container?

```
➢ docker run --volume=</path/to/output/dir>:/output --
  volume=</path/to/input/dir>:/input myAnalysis
```

Thank you!

Find us @ BSS G 41 & 45.1

Franziska Singer
singer@nexus.ethz.ch

www.nexus.ethz.ch