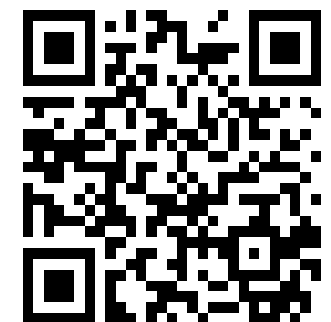# EDUCATING AND TRAINING FOR A **FAIR** FUTURE

Adina Wagner

🐦 @AdinaKrik

Psychoinformatics lab,
Institute of Neuroscience and Medicine, Brain & Behavior (INM-7)
Research Center Jülich
ReproNim/INCF fellow

JÜLICH
Forschungszentrum

Slides: DOI 10.5281/zenodo.4541323 (Scan the QR code)
Sources: github.com/datalad-handbook/datalad-course

TRAINING FOR A FAIR FUTURE IN NEUROSCIENCE

Neuroscience strives from interdisciplinarity
and collaboration

TRAINING FOR A FAIR FUTURE IN NEUROSCIENCE

Neuroscience strives from interdisciplinarity and collaboration

... but training groups with a diversity of skill sets is difficult

RESEARCH DATA MANAGEMENT (RDM)

... TYPICAL DIFFICULTIES

Lack of formal training
    Rarely included in primary/graduate/post-graduate studies. Rather: Learning "on the job" and inheriting project management from the previous person

"Science alone is hard already".
    RDM comes on top of everything else, in usually tight, competitive graduate programs or jobs

Too much to know and too little guidance
    Rather than a motivational problem, young researchers face the difficulty of finding out which tools exist and are helpful

Late RDM = little benefit
    If RDM is only incentivized, required, or tackled as the very last step, researchers can not benefit (fully) from RDM

How can we overcome this?

PERSONAL INSIGHTS FROM RDM TRAINING

RDM overhaul in an institute

User-focused software documentation

```
--- /data/BnB1/DATA/download_data/eNKI --------------
                      /..
    5.2 TiB [##########] /eNKI_unzipped
    3.3 TiB [######    ] /eNKI_redownload
    3.2 TiB [######    ] /eNKI_BIDSdownload
  724.2 GiB [#         ] /eNKI_20180806
  218.8 GiB [          ] /eNKI_aus_Raw_Data
```
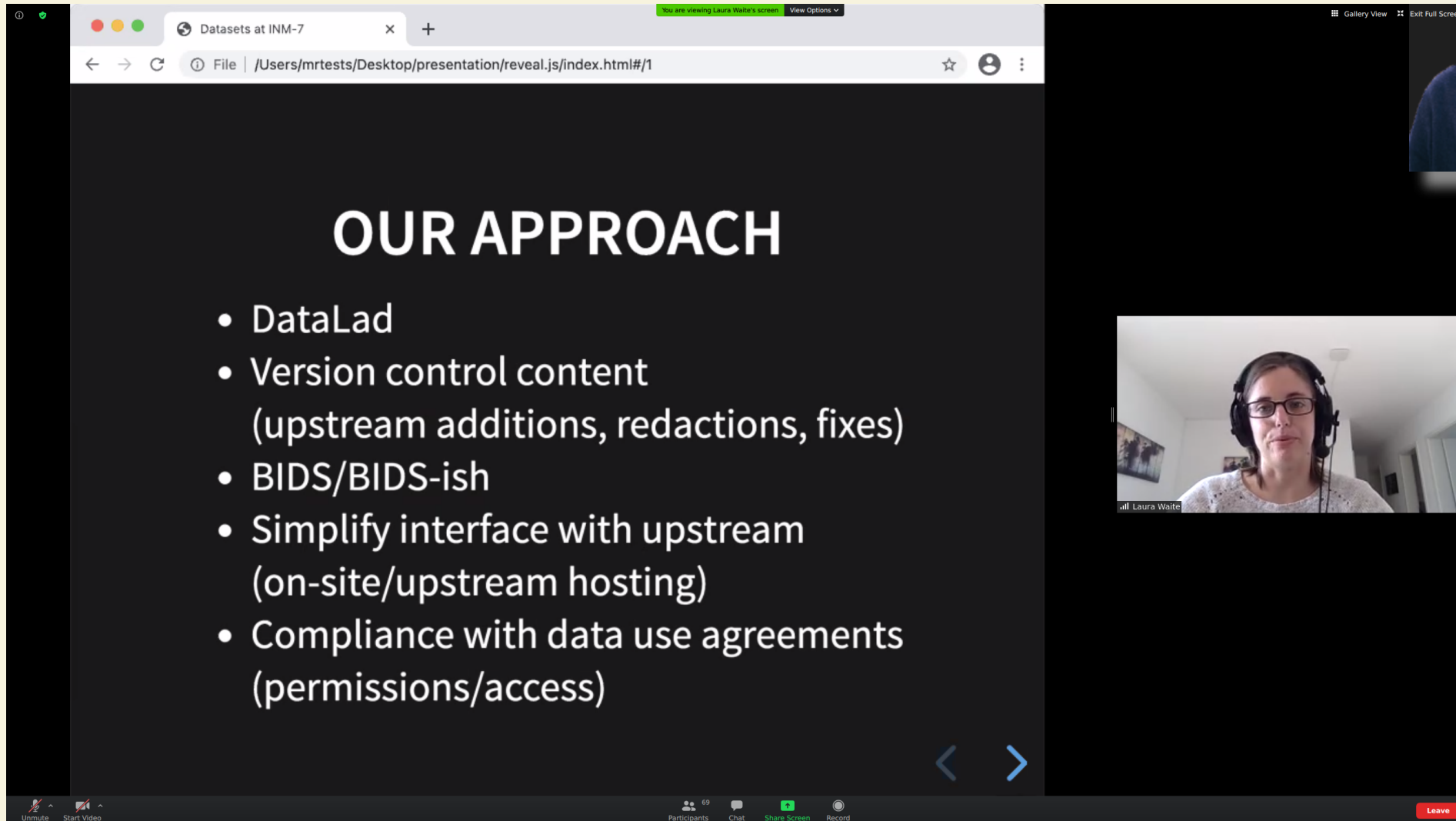
```
In [1]: from datalad.api import create

In [2]: create?
Signature:
create(
    path=None,
    initopts=None,
    force=False,
    description=None,
    dataset=None,
    no_annex=<class 'datalad.core.local.create._NoAnnexDefault'>,
    annex=True,
    fake_dates=False,
    cfg_proc=None,
)
Docstring:
Create a new dataset from scratch.

This command initializes a new dataset at a given location, or the
current directory. The new dataset can optionally be registered in an
existing superdataset (the new dataset's path needs to be located
within the superdataset for that, and the superdataset needs to be given
```

Aim: Researchers turn to RDM because it improves their science and work, not because they are forced to

RDM OVERHAUL IN AN INSTITUTE

- System-wide DataLad deployment
- All large datasets are managed and accessed via DataLad (to avoid duplication, for provenance-tracked, reproducible transformations)
- Effort led by Laura Waite

TRAINING AN INSTITUTE

- In-person and virtual training
- On-demand user documentation at handbook.datalad.org

THE DATALAD HANDBOOK

Its structure reflects different needs of different stakeholders in science:

Intro-
duction

Basics

Advanced

Use
cases

Starting must be easy
- High-level function/command overviews, Installation, Configuration, Cheatsheet

Tutorials for everyone
- Narrative-based code-along course
- Independent on background/skill level, suitable for data management novices

Tutorials for experts
- Complex functionality and workflows
- Examples: Computationally reproducible analysis on big data

**Overviews for decision makers**
- Step-by-step solutions to common data management problems, like how to make a reproducible paper

TAKE HOME MESSAGES

- Complex software tools need accessible documentation
- Way to adoption of tools or principles via requirements/force/external incentives, or through immediate, personal benefits of good RDM (the latter keeps trainers sane)

A few Handbook metrics:

- Three Handbook releases
- 36 coauthors
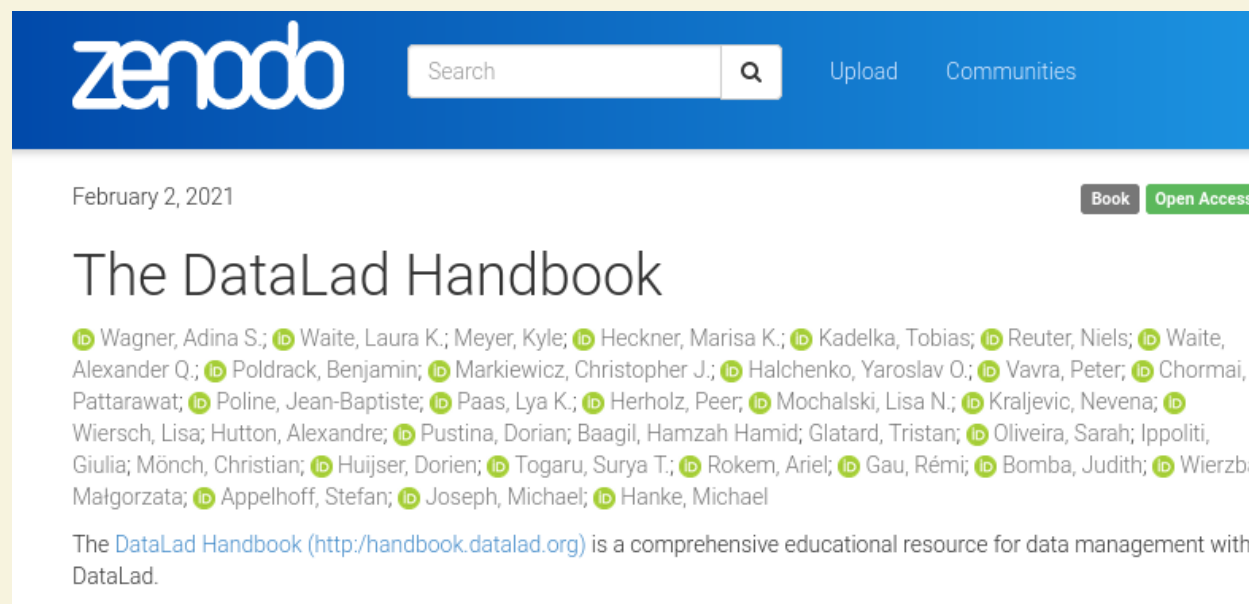
- 500 pages of content
- dozens of openly shared slides, code lists, and other training materials
- about 250 unique views per day

ACKNOWLEDGEMENTS