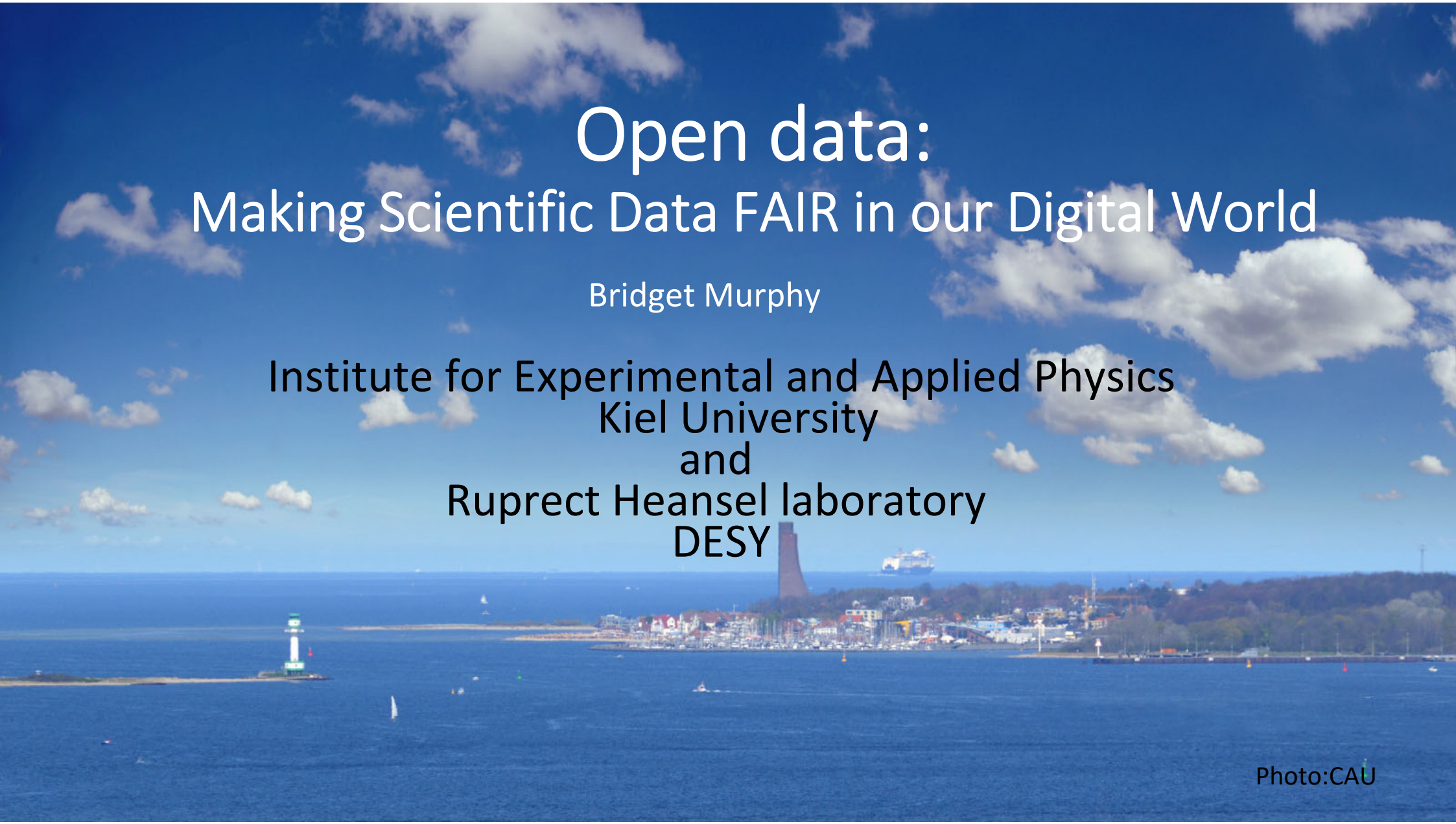


Open data:
Making Scientific Data FAIR in our Digital World

Bridget Murphy

Institute for Experimental and Applied Physics
Kiel University
and
Ruprect Heansel laboratory
DESY

Photo:CAU

How to approach the data challenge Sustanibility?

DAPHNE4NFDI

Data for Photon and Neutron Experiments

Consortium

Brings together 18 partners:

- University user groups
- Large scale facilities
- In addition: KFN + KFS
- > 60 participants (without funding)

Task Area Leader:

Bridget Murphy TA1, TA6 (CAU/DESY Speaker)
Astrid Schneidewind TA4 (FZJ, Deputy speaker)
Christian Gutt TA5 (U Siegen)
Wiebke Lohstroh TA1 (TUM)
Anton Barty (DESY) TA3

Sebastian Busch TA2 (Hereon)
Jan-Dierk Grunwaldt TA4 (KIT)
Frank Schreiber TA3 (U Tübingen)
Tobias Unruh TA2 (FAU)

Coordinator: lisa.amelung@desy.de

DAPHNE4NFDI.de

DAPHNE4NFDI brings the partners together

to address data, metadata management and high data rate challenges, and develop solutions for outstanding scientific experiments with the user community.

Currently 50 PB of data in a year collected in Germany

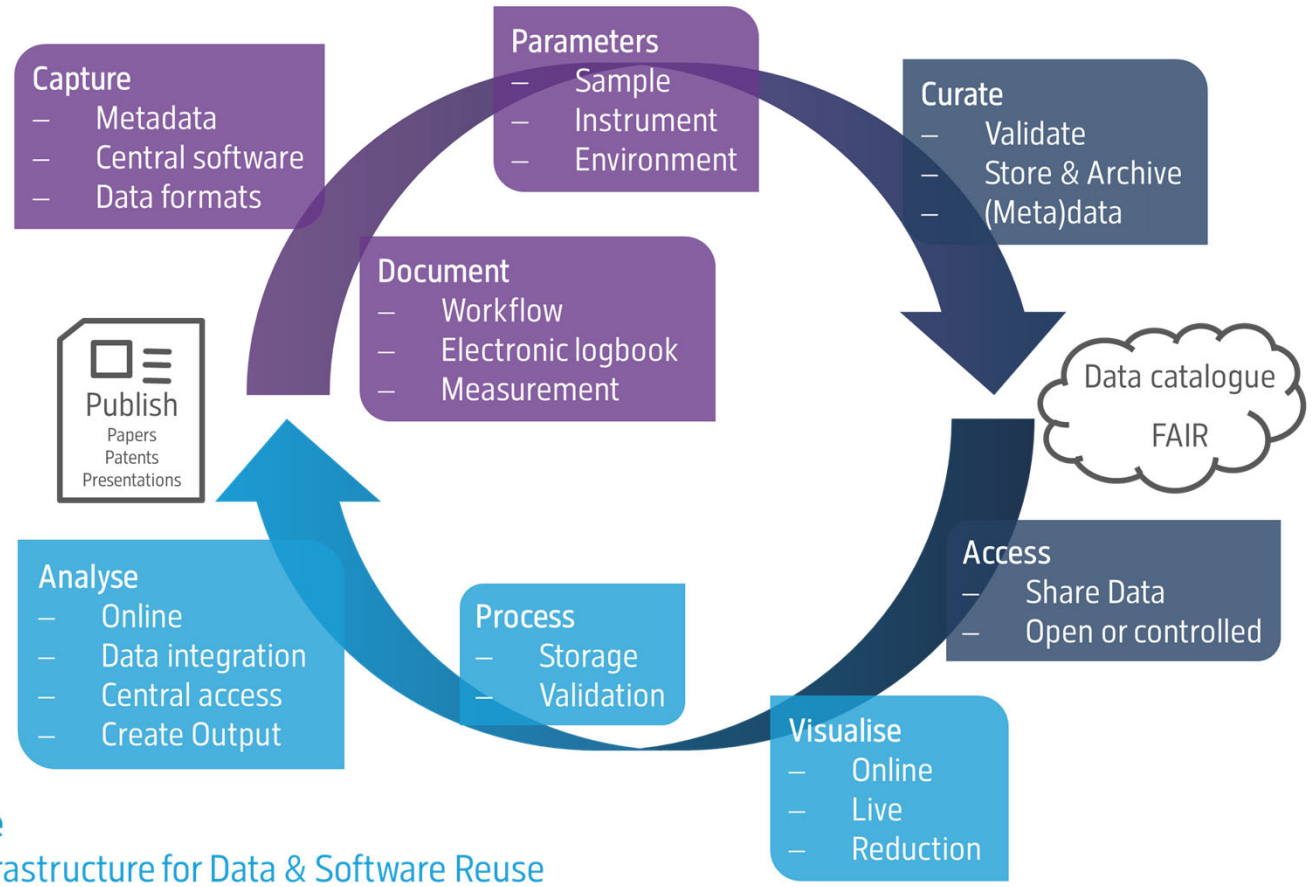
- **50 million smartphone users in Germany**
- **50 GB typical smartphone storage**

DAPHNE4NFDI aims

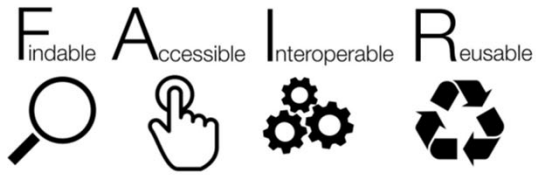
to make the growing volume of valuable measured data FAIR for the scientific community.

Collect
TA1: Managing Data Production

Store
TA2: (Meta)data Repositories & Catalogues



Evaluate
TA3: Infrastructure for Data & Software Reuse



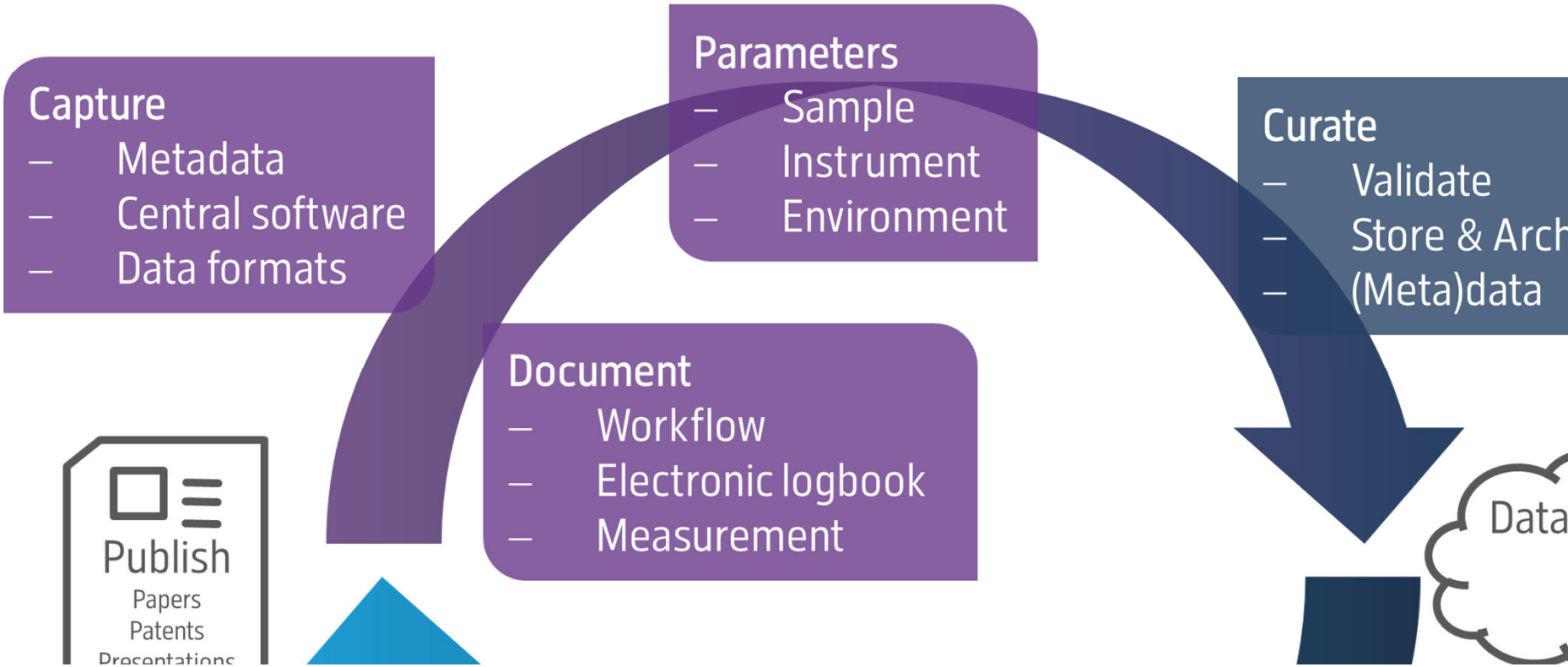
Outreach and Dissemination
TA4: Establish & enhance awareness of FAIR principles

External Communication and Policy
TA5: Common data policy & alignment with European partners

Project Management
TA6: Project coordination & Administration

DAPHNE4NFDI aims

Collect
TA1: Managing Data Production

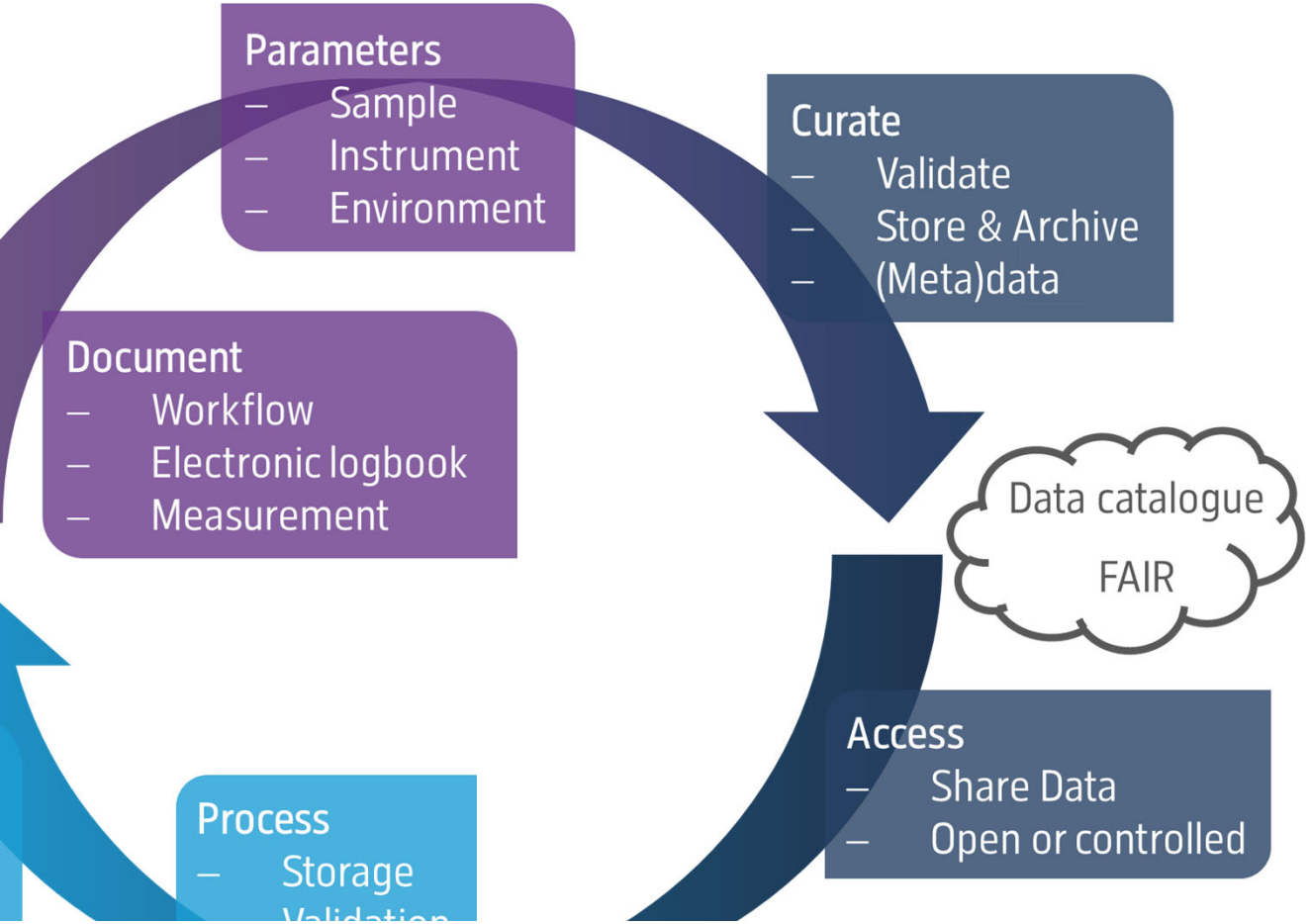


DAPHNE4NFDI aims

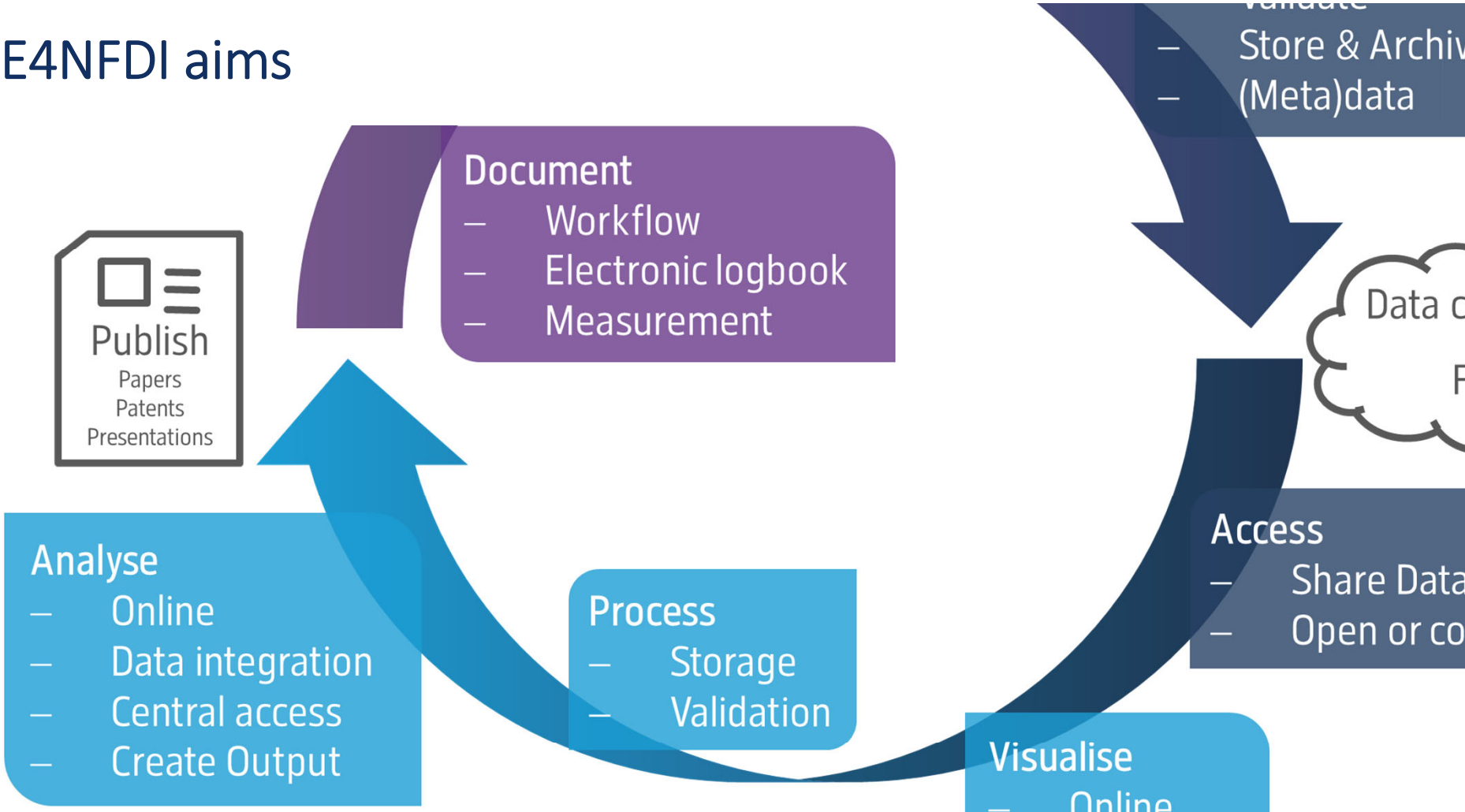
Findable Accessible Interoperable Reusable

Store

TA2: (Meta)data Repositories & Catalogues



DAPHNE4NFDI aims



Evaluate

TA3: Infrastructure for Data & Software Reuse

DAPHNE4NFDI Use Cases serve as flagship projects

Saving the ~~world~~ data step by step…

Biomaterials
X-ray imaging
LMU - Uni Göttingen

Energy and battery materials, catalysis
Tomography
TUM - MLZ - BAM - hereon - HZB - KIT

Correlated electron systems **Spectroscopy**
KIT - FZJ - MLZ

Soft matter and liquid interfaces
x-ray reflectivity
CAU - Uni Tübingen

Proteins & Food science
Diffraction (small and wide angle) Spectroscopy
FAU - Uni Tübingen - EMBL - CAU

Magnetic structures
Ultrafast / Magnetic x-ray scattering
DESY - Uni Siegen

Dynamics
Correlation spectroscopy - XPCS
Uni Siegen - EuXFEL

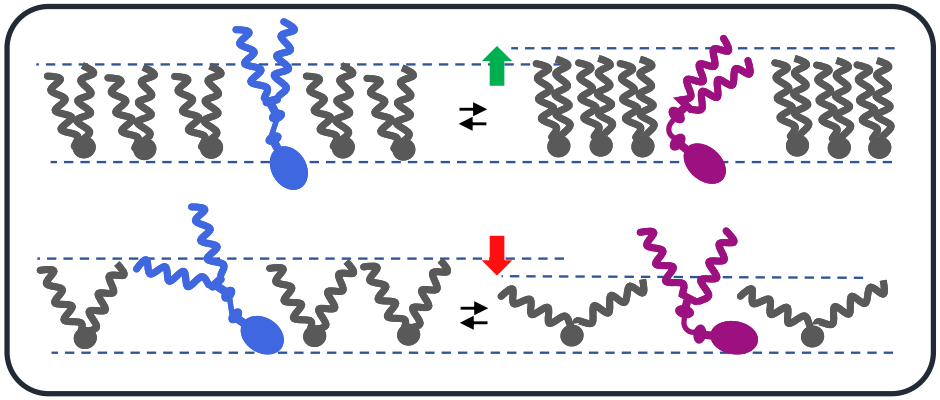
Amorphous materials for catalysis
x-ray absorption spectroscopy
KIT - TUB - Uni Wuppertal

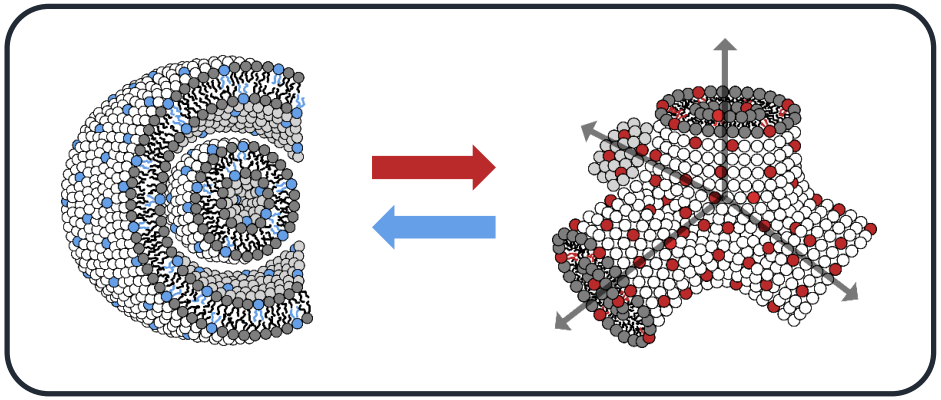
Electrochemistry & Catalysis
High energy x-ray diffraction
HZDR - CAU - DESY

Reusable powder refinement
Neutron TOF diffraction
FZJ - MLZ - ESS - RWTH

Chemical systems
x-ray emission spectra, RIXS etc.
KIT - ESRF - DESY

Dynamics in lipid membranes



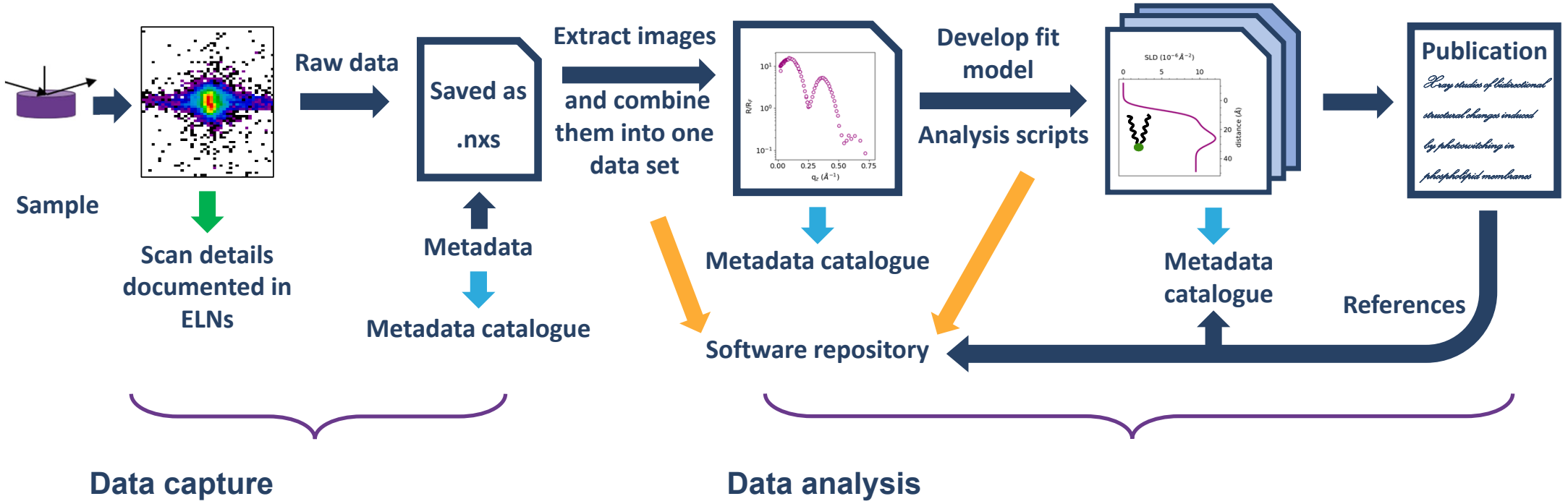


How is this related to data management?

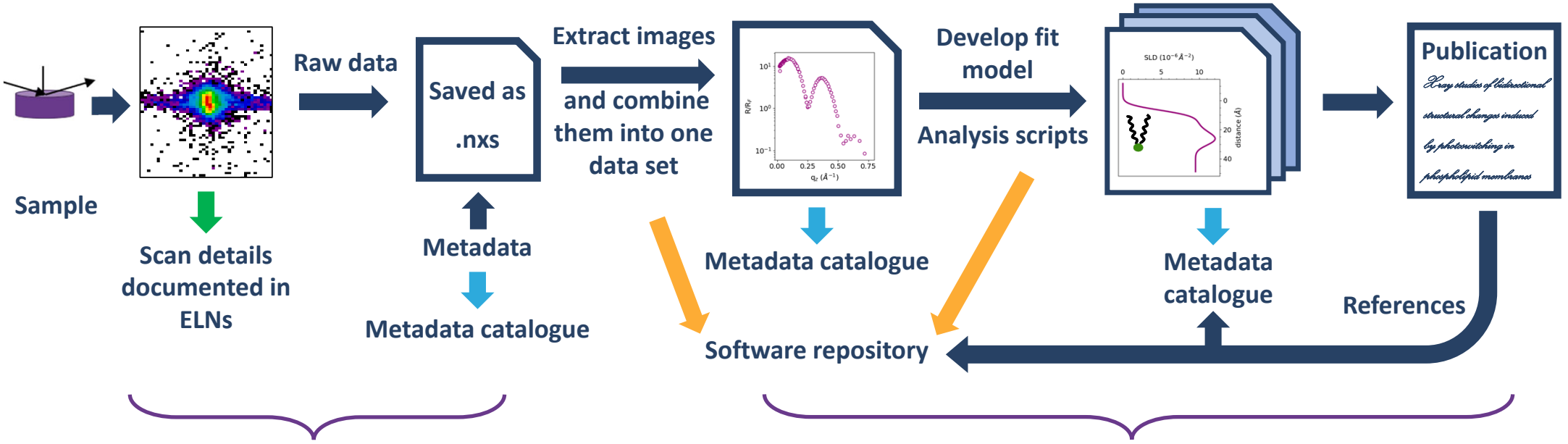
Data handling
- Collect
- Store/Archive
- Evaluate

FAIR principles

Research Data Management Workflow



Overview



Data capture

- FAIR data management
- Definition of metadata
- Documentation of the measurement and data

Data analysis

Data capture: Definition of metadata

What is metadata?

Metadata is "data that provides information about other data" [1], but not the content of the data such as the text of a message or the image itself:

- Descriptive
- Structural
- Administration
- Reference
- Statistical data
- Legal

Copyright: Wikipedia

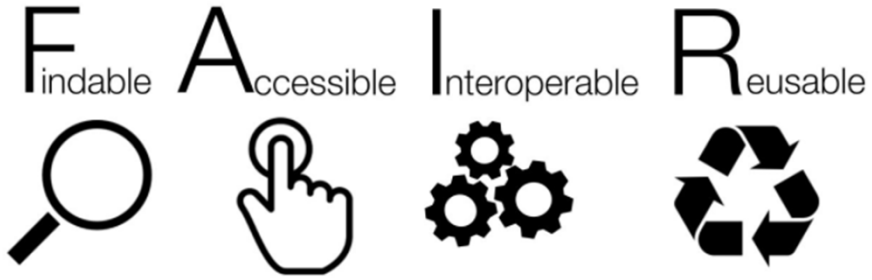
[1] National Information Standards Organization (NISO) (2001). Understanding Metadata (PDF). NISO Press. p. 1. ISBN 978-1-880124-62-8.
[2] Wikipedia

Data capture: FAIR Data Management

What is FAIR and how to be FAIR?

FAIR

- Findable
- Accessible
- Interoperable
- Reusable



The FAIR Guiding Principles for scientific data management and stewardship
https://www.nature.com/articles/sdata201618

Data capture : FAIR Data Management

Findable

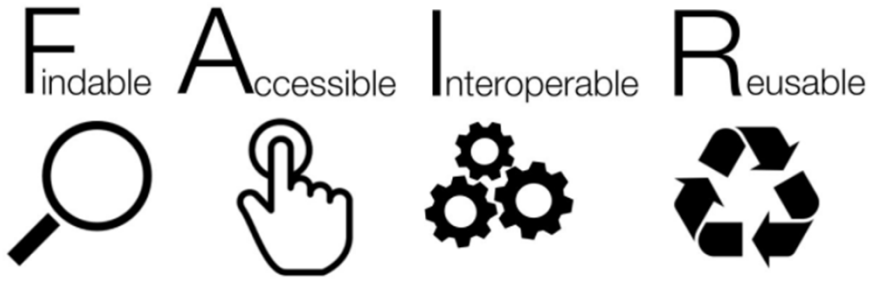
Step 1

Metadata and data should be easy for humans and computers to find. Machine-readable metadata is essential for the automatic discovery of records and services.

Accessible

Step 2

Once the user has found the data they want, they need to know how to access it, possibly including authentication and authorisation.



The FAIR Guiding Principles for scientific data management and stewardship
https://www.nature.com/articles/sdata201618

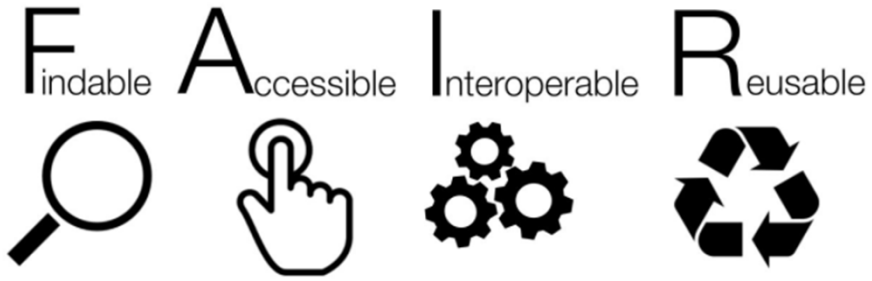
The FAIR Guiding Principles for scientific data management and stewardship https://www.nature.com/articles/sdata201618

Data capture : FAIR Data Management

Interoperable

Step 3

The data usually needs to be integrated with other data. In addition, the data must work with applications or workflows for analysis, storage and processing.



The FAIR Guiding Principles for scientific data management and stewardship
https://www.nature.com/articles/sdata201618

Reusable

Step 4

The goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well described so that they can be replicated and/or combined in different environments.

The FAIR Guiding Principles for scientific data management and stewardship https://www.nature.com/articles/sdata201618

Data Capture: Record the Experiment

Document the experimental **workflow**

- Written record of the experiment **(Electronic lab book ELN)**
- **Digital object identifier (**Sample preparation, calibration, other measurements)
- **Automatic capture** of: instrument, sample environment and sample data @experiment
- Save **data** (Data format - Nexus, OpenPMD)

Tabulate the data

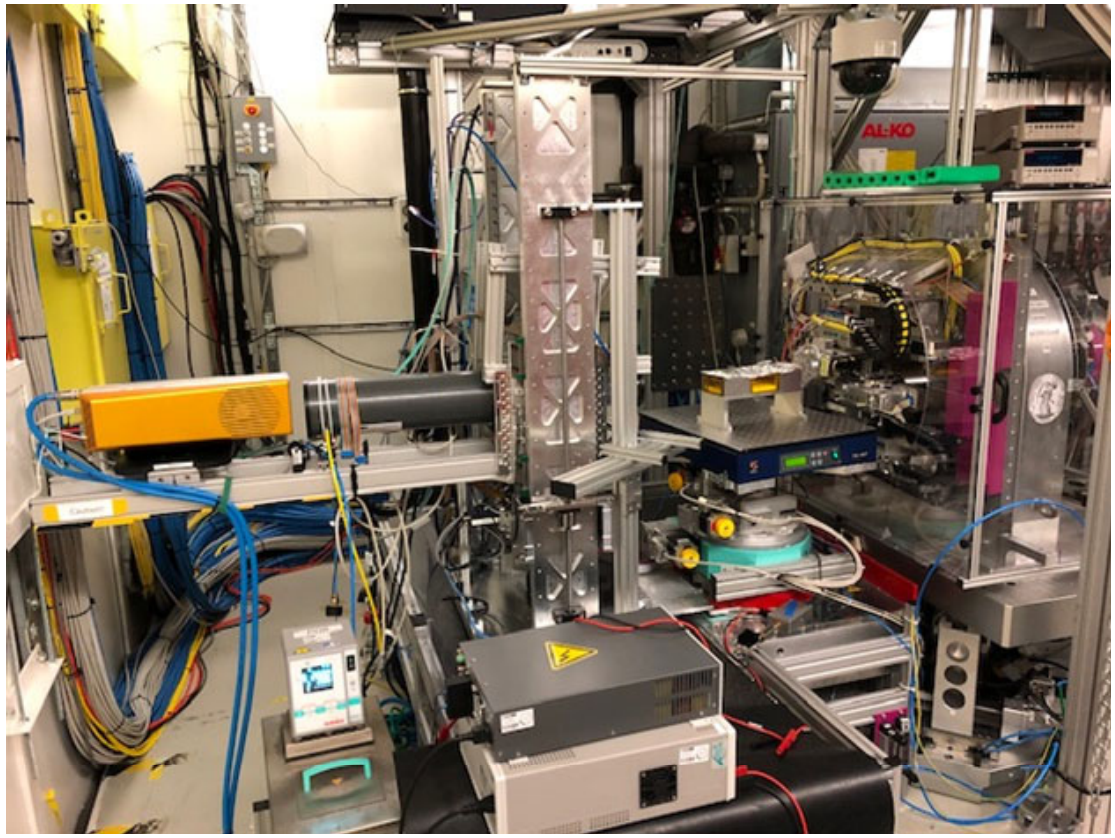
- **Independent** variable in the first column (set)
- **Dependent** variable in the second column (measured)
- **Control** data
  - Other parameters (point number, time, monitor signals)
  - **Metadata**
  - File header
  - File name

Data capture: Documentation of measurements and data

Recording of data and results

Tabular representation of the data using the example of X-ray reflections:

independent variable	dependent variable	Error
Angle of incidence q_z / Å	Intensity / counts	Intensity errors/ counts



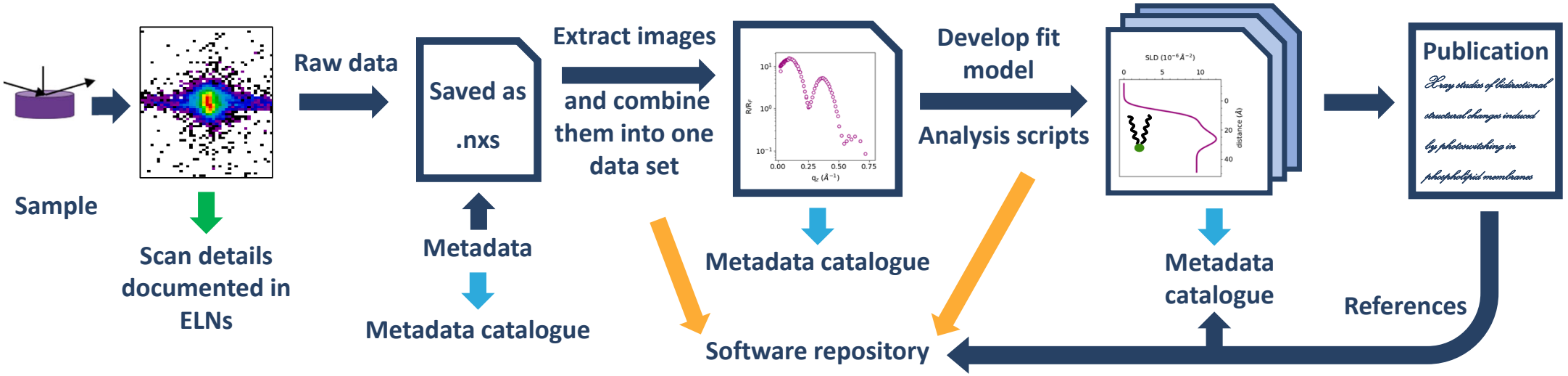
Data capture: Documentation of measurements and data

Recording of data and results

Tabular representation of the data using the example of X-ray reflectior

independent variable	dependent variable	Error	Control data		
Area/ per molecule	Surface pressure / mN/m	Surface pressure error/ mN/m	Temperature	O2 content	Gas pressure

Overview



Data capture

- FAIR data management
- Definition of metadata
- Documentation of the measurement and data

Data analysis

- Data preparation
- Fit curve, statistics, errors, correlations
- Model Function
- Artificial intelligence and machine learning techniques (convolutional neural network)

Kiel, Tübingen, ESRF, DESY (P08, FS-EC , FS-SC, central IT)) L Pithan *et al., JSR. (2023),* Hövelmann *et al., IUCrJ. (2024)*

Data Analysis:

Reliability

- How close are measured values when repeated measurements are taken? E.g. the intensity at an angle or a certain temperature when measured again.
  - ❑ Choice of equipment for precise setting of the measuring positions (e.g. motor with encoder)
  - ❑ Increase number or repetitions
  - ❑ Use averaging

Accuracy

- How close is the final result to the correct or acceptable value?
- Reference measurement with a known sample for
  - ❑ Instrument calibration
  - ❑ Improving the accuracy of individual measurements (e.g. counting time, detector resolution, position accuracy)

Validity

- Is the experimental method suitable for the objective?
  - ❑ Are the control variables constant?
  - ❑ Are the assumptions confirmed by the experiment? (e.g. mode of operation of the AFM)

Data analysis: Data preparation

Data reduction

2D detector image -> 1D data set

- Definition of an area of interest (ROI)
- Linear integration along an axe
- Conversion of the pixels into reciprocal space or energy

What data is needed?

Independent variable (first column/ x-axis)

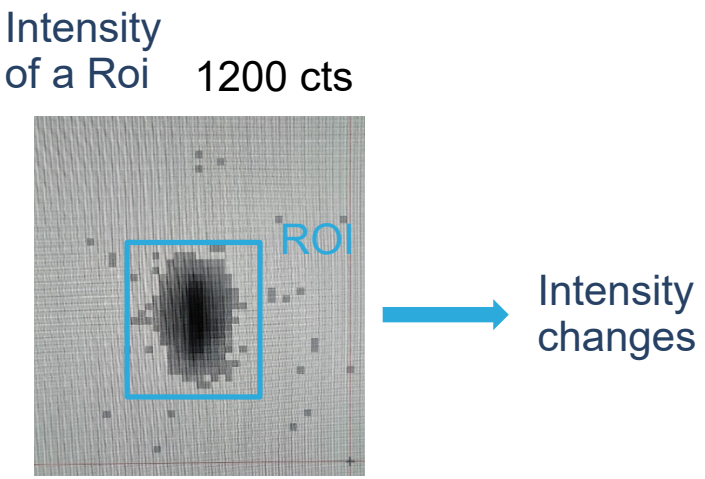
- Position of the detector, temperature, pressure, ...

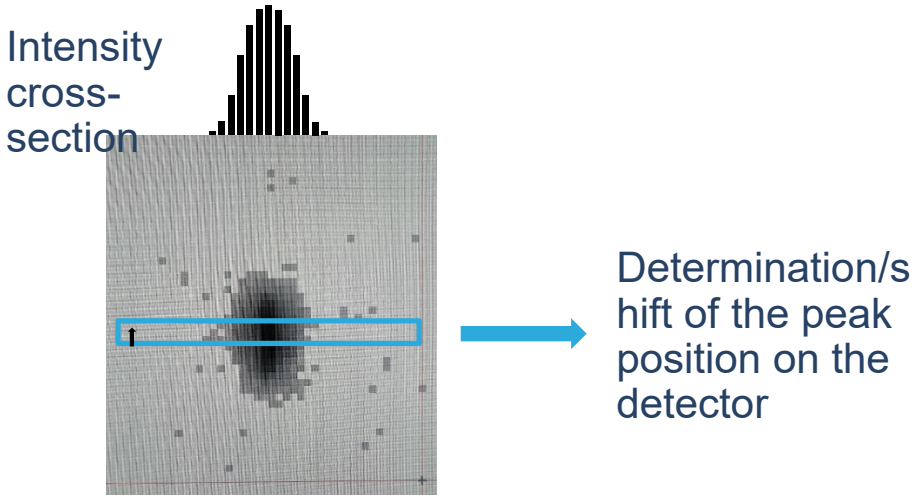
Dependent variable (second column/ y-axis)

- Intensity of the detector, the ROI, cross-section

Optional conversion into other units

- e.g. angle to q-space...



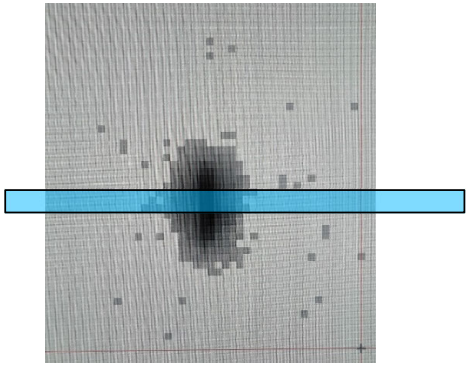


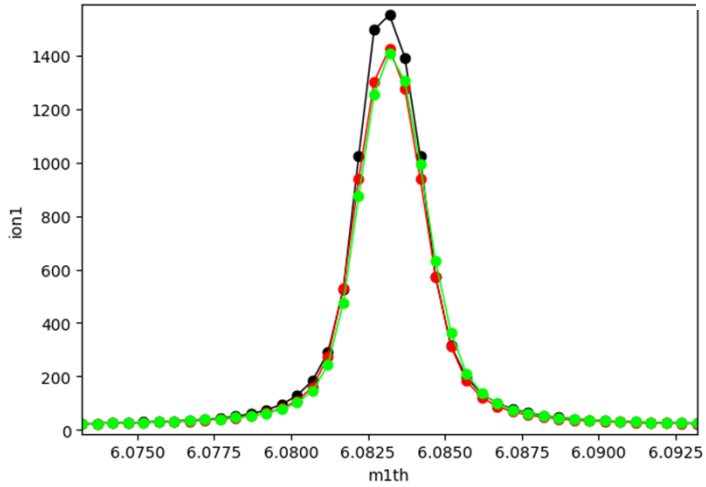
Preparing data for analysis

After or during data collection

Reduction

- Region of interest





- Linear integration

- 2 Detector – convert pixel to recipricoal space or to energy

What to do with the reduced data

Fit the data

- line
- Curve or number of curves
- Gradient

Fit a model

- Example X-ray reflectivity
- Recursive algrothim for coherent scattering

Infrastructure for data and software re-use

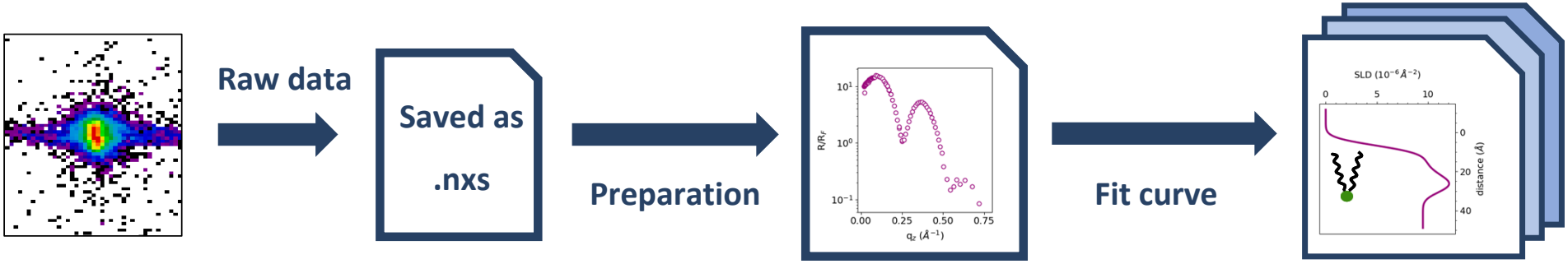
Community data analysis software and data mining strategies including machine learning

- **Share** analysis software
- **Teach** and practice sustainable research software development
- **Automatize** data analysis chains
- Adoption of common **data models** e.g. by using NeXus

Evaluate

V. Starostin et al., Synchr. Rad. News 35 (2022) 21, 10.1080/08940886.2022.2112499
A. Hinderhofer et al., J. Appl. Cryst. (2023, in press) 10.1107/S1600576722011566

Data analysis: Curve fitting



What does it mean to fit a curve? And why do you have to fit curves?

First step: Understand and select measurement methods.

- Which property should be measured? Can I do this with the method and what are the limits?

Second step: Data preparation and extraction

- e.g. 2D to 1D data set reduction

Third step: Fit data curve

- Which parameters can be fit and how are they related?

Data analysis: curve fitting - example X-ray reflectivity

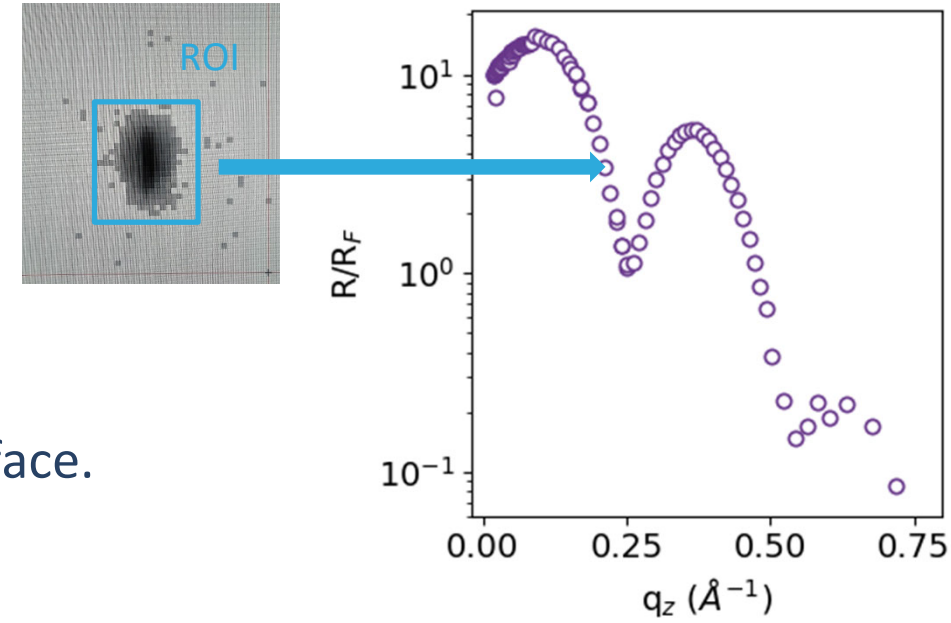
Second step: Data preparation and extraction

LISA@PETRA III

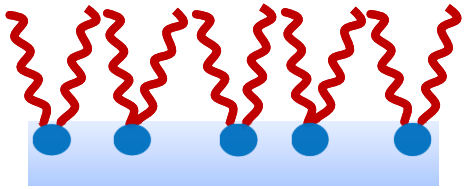
Preparation and extraction: X-ray reflectivity

- Application Flatfield and Hot Pixel Mask
- Choice of the area to be considered (ROIs)
- Integration of intensity
- Background correction

-> 1D data set of intensities against



DPPC monolayers at the water-air interface.



Data analysis: Curve fitting - example X-ray reflectivity

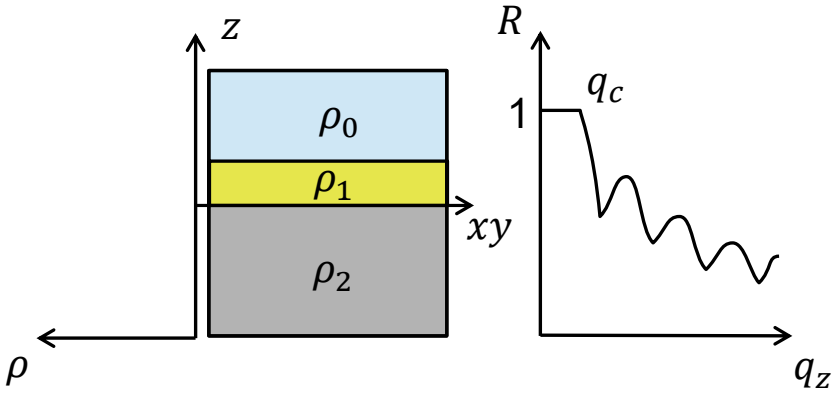
Third step: Fit data curve

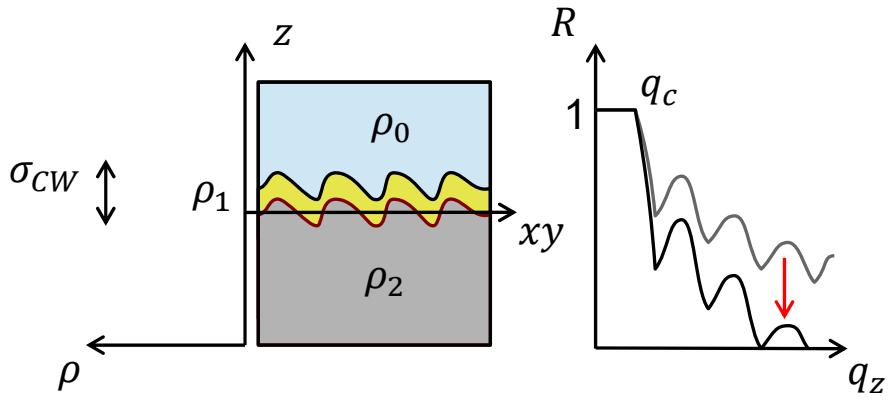
Method: X-ray reflectivity - How does it work?

- Fourier transformation of the electron density

$$R = R_F \left| \int_{-\infty}^{\infty} e^{iQz} \cdot \frac{df(z)}{dz} dz \right|^2 = R_F \left| \frac{1}{\rho_\infty} \int_{-\infty}^{\infty} e^{iQz} \cdot \frac{\partial \langle \rho \rangle_{x,y}}{\partial z} \right|^2$$

- Kissing fringes in multilayer systems (interfereance)



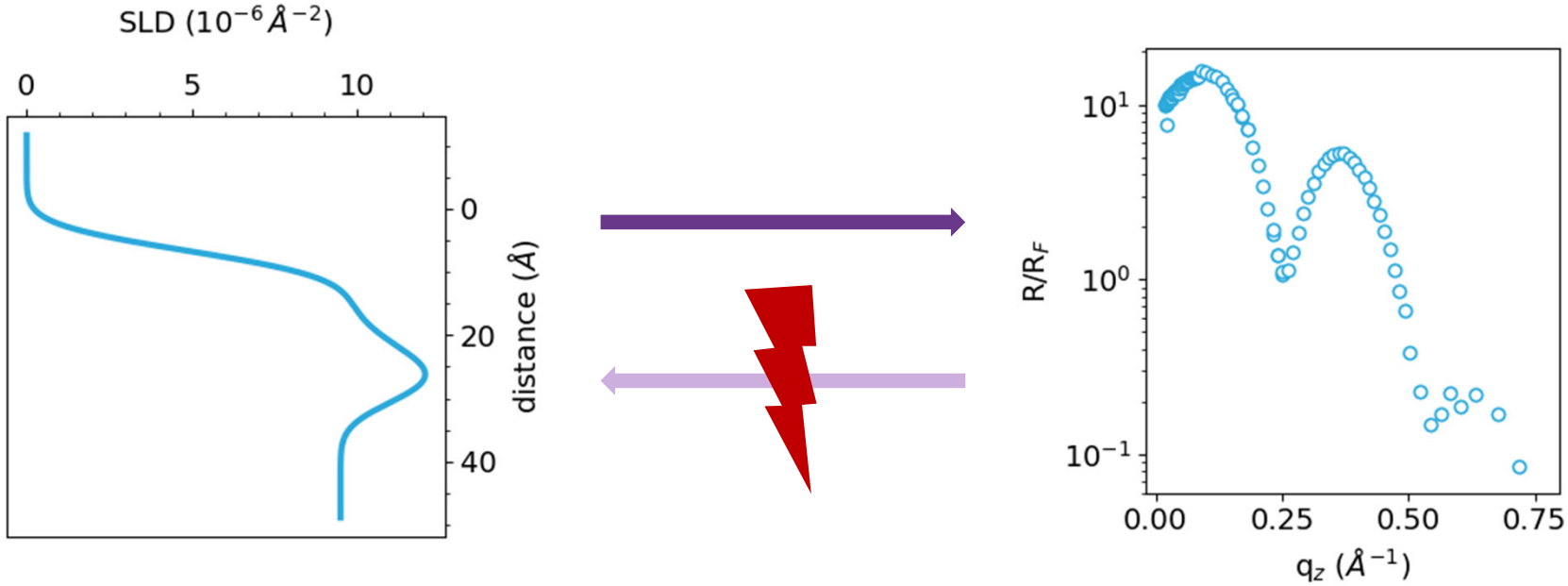


Data analysis: curve fitting - example X-ray reflectivity

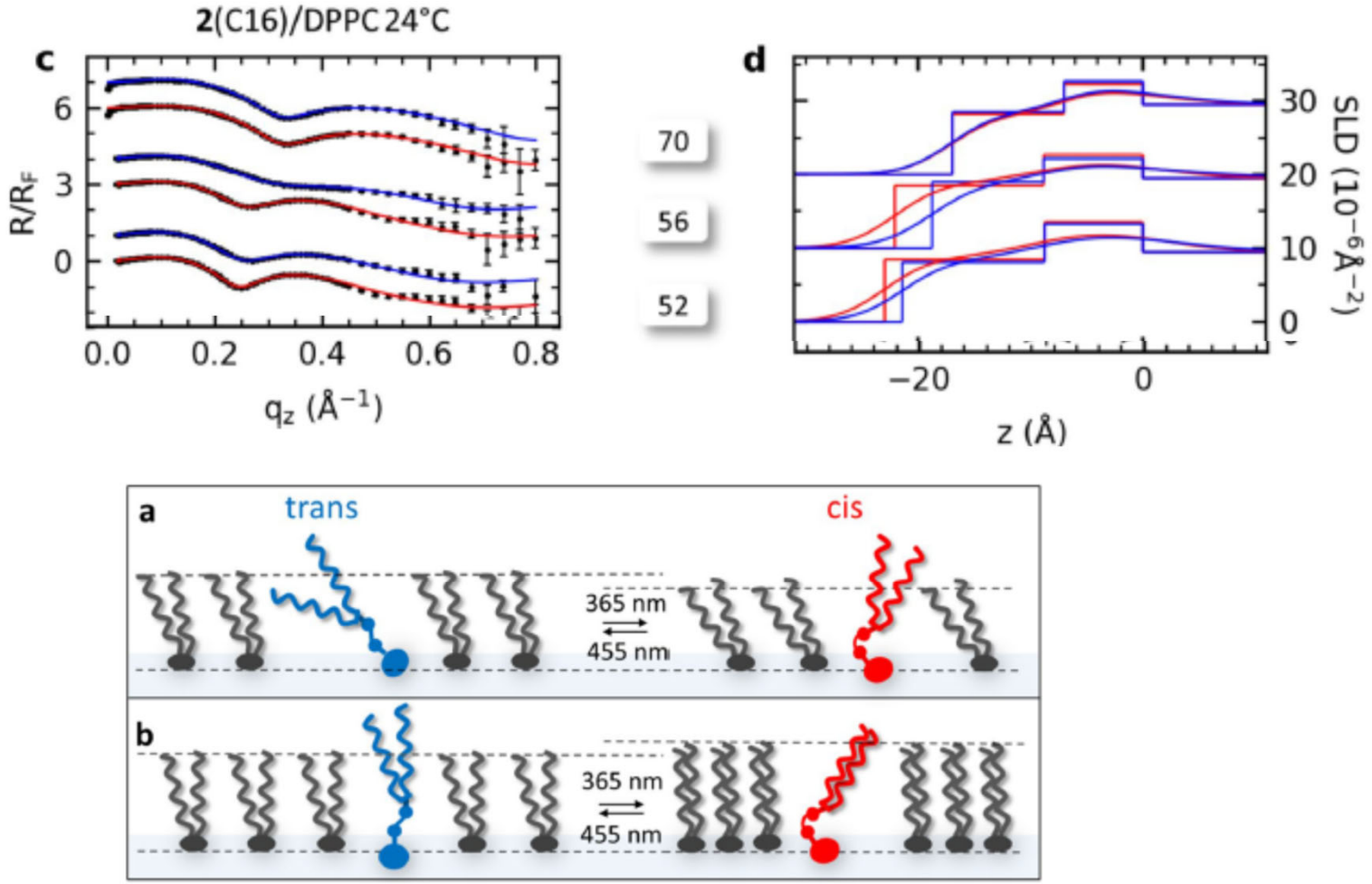
Third step: Fit data curve

Why do we need to fit the data?

- Data curve is the Fourier-transformed electron density and cannot be directly back-calculated



Photoswitichable lipids in a monolayer – Bidirectional switch

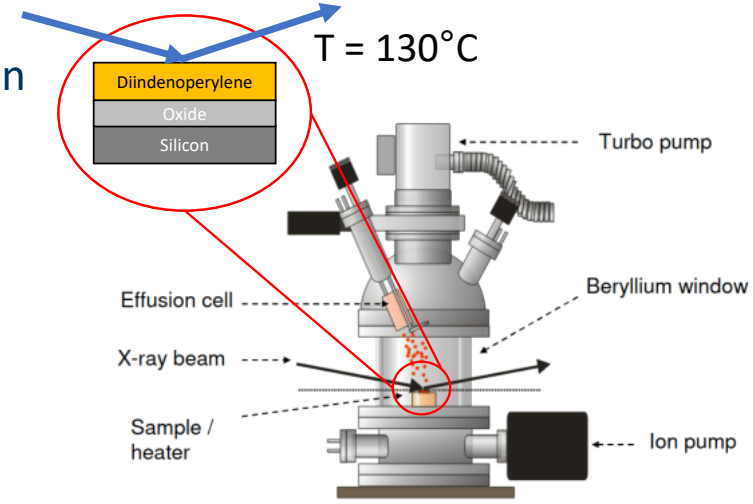


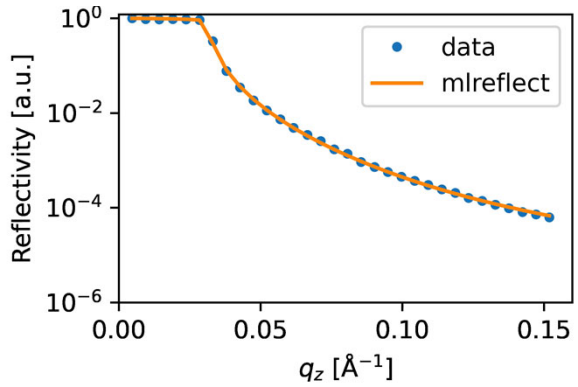
Warias, et al. " *Scientific Reports* 13.1 (2023): 11480.

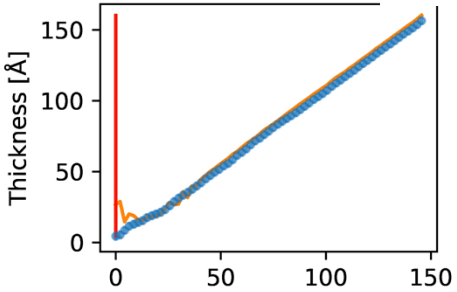
In situ applications of mlreflect

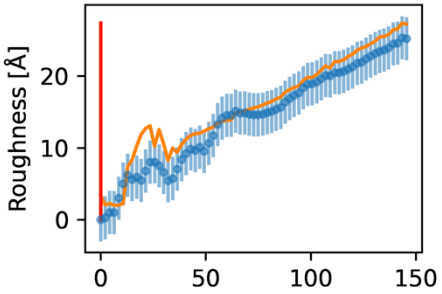
- Real-time parameter prediction useful for in situ experiments
- After training, no human input is necessary
- Results are obtained within <1s per curve
- Ideal for monitoring and feedback loops

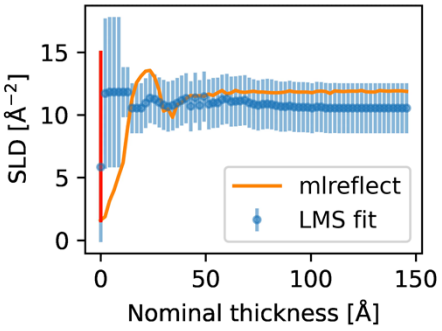
In situ XRR during film deposition







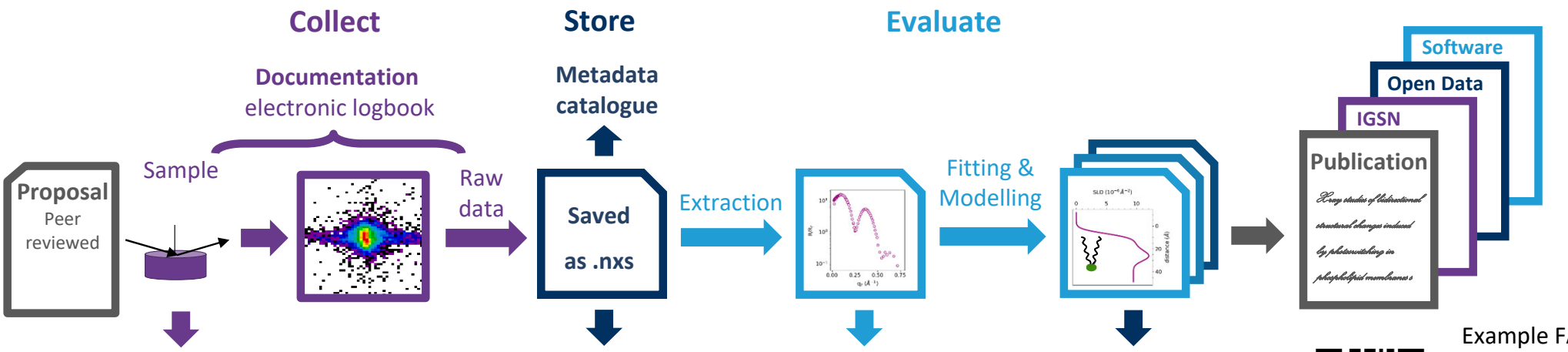




Hinderhofer *et al. Europhys. Lett.*, 2010, **91**, 56002
Kowarik *et al. Phys. Rev. Lett.*, 2006, **96**, 125504
Bommel *et al. Nat. Comm.*, 2014, **5**, 5388
Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

Pithan *et al. JSR. (2023),* **https://arxiv.org/abs/2306.11899**

Our effort to make data FAIR



www.igsn.uni-kiel.de

Service provided by
CAU central computing
Thorge Petersen
Axel Wedemeyer
Thilo Paul-Stüve

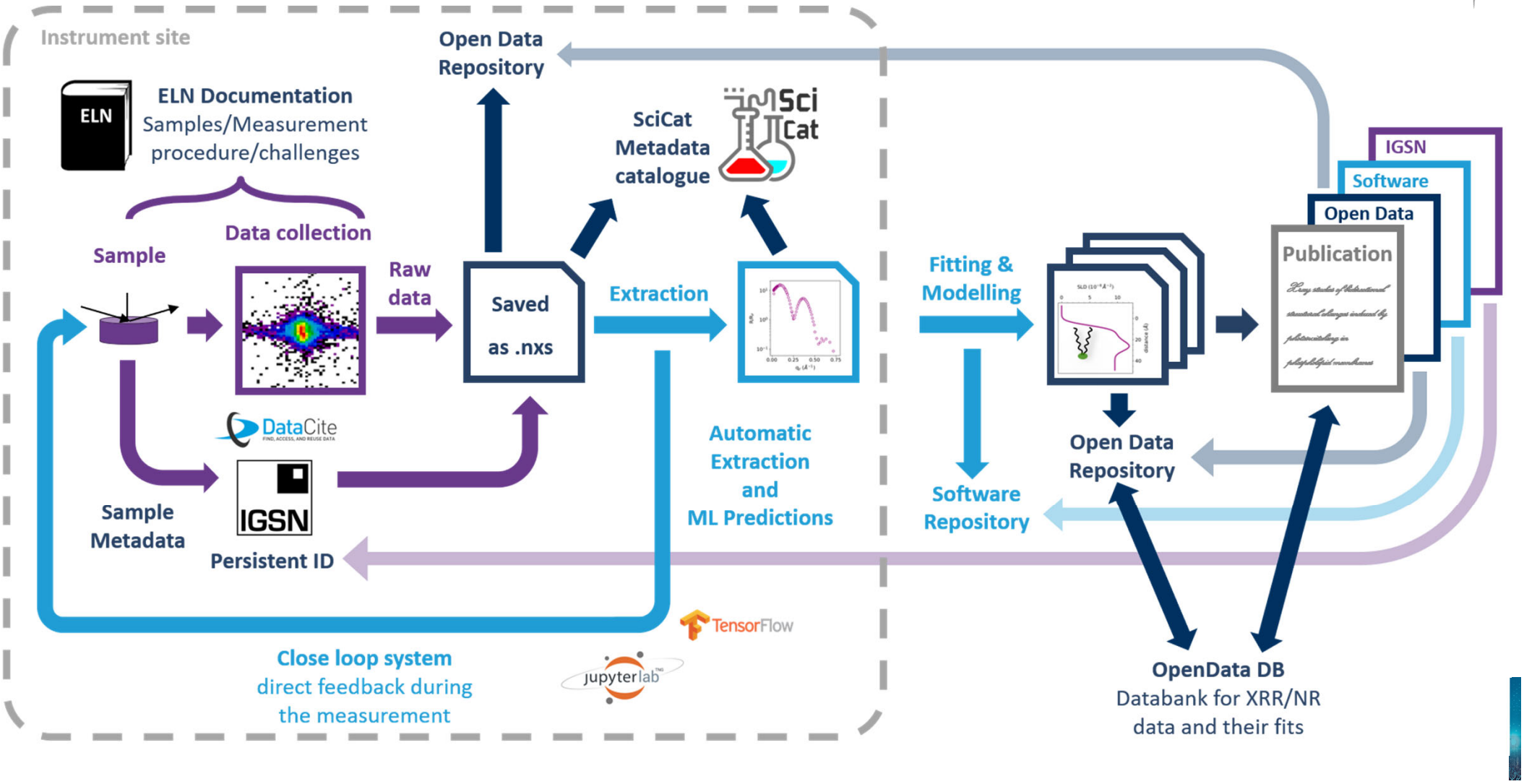
L. Pithan, J. *Synchr. Rad.* **30**, 1064 (2023)

Example FAIR data publication

S. Hövelmann, *IUCr J* **11**, 4 (2024)

Findable Accessible Interoperable Reusable

Best practice for FAIR data collection and storage



LISA/DAPHNE Team

Kiel University

Prashant Hitaishi	Svenja Hövelmann
Philipp Jordt	Julia Kobus
Nicolas Hayden	Otto Lipmann
Ali Ashtiani	Ella Dieball
Olaf Magnussen	Matthias Greve
Jonas Warias	Rajendra Giri
Lukas Petersdorf	Andrea Sartori
Sven Festersen	Benjamin Runge
Annika Elsen	Christoph Lemke
Klaas Loger	Hiromasa Fujii
Jochim Stettner	Björn Haushahn

DESY

Chen Shen	Regina Hinzmann	Lisa Amelung
Florian Bertram	Igor Khokhriakov	Partha Tirumalai

Funding: DFG; BMBF ErUM-Pro current **05K19FK2,** RAC **05K23FKA** DESY, SFB 677,1261