technische universität
dortmund

lehrstuhl
physik e5

Next Generation Environment for Interoperable Data Analysis

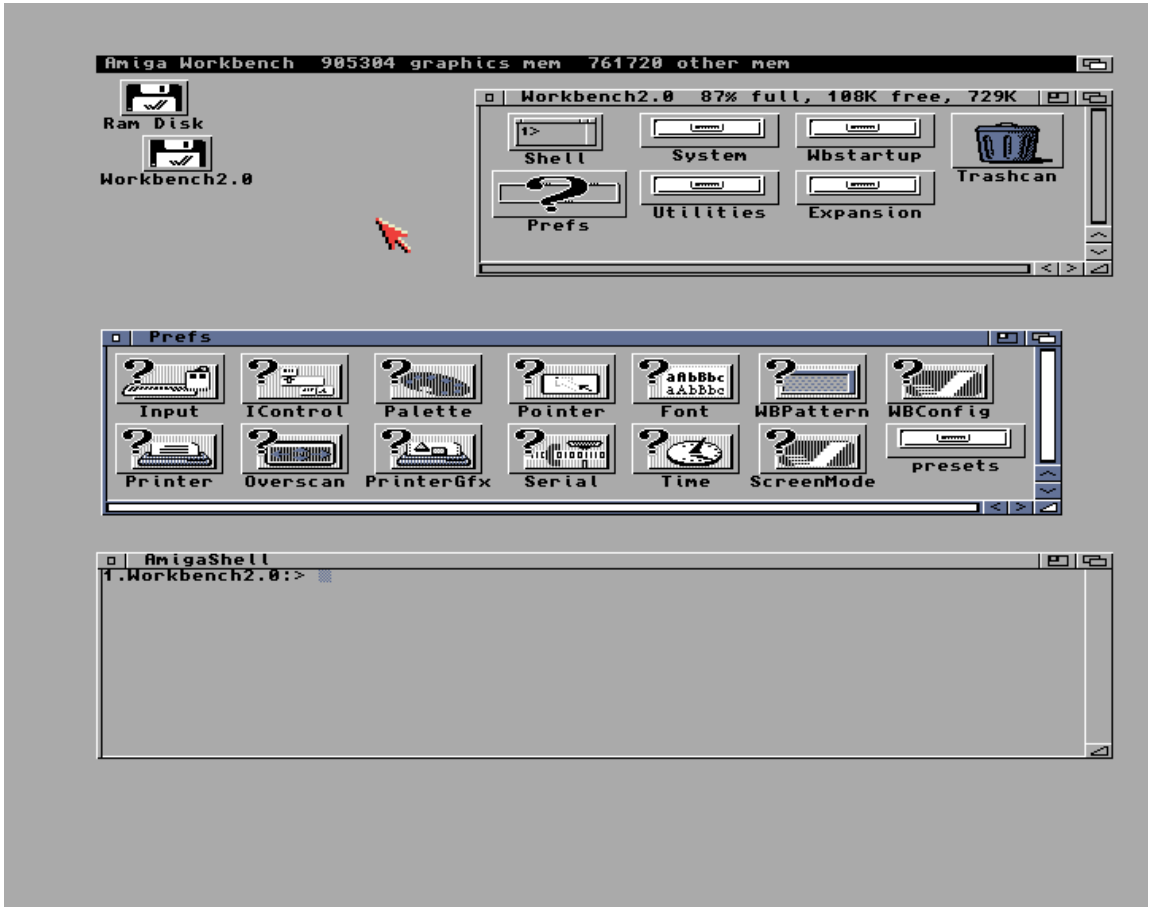
Tim Ruhe

Tim Ruhe, Next Generation Environment for Interoperable Data Analysis
Expert Workshop, HZB 2023

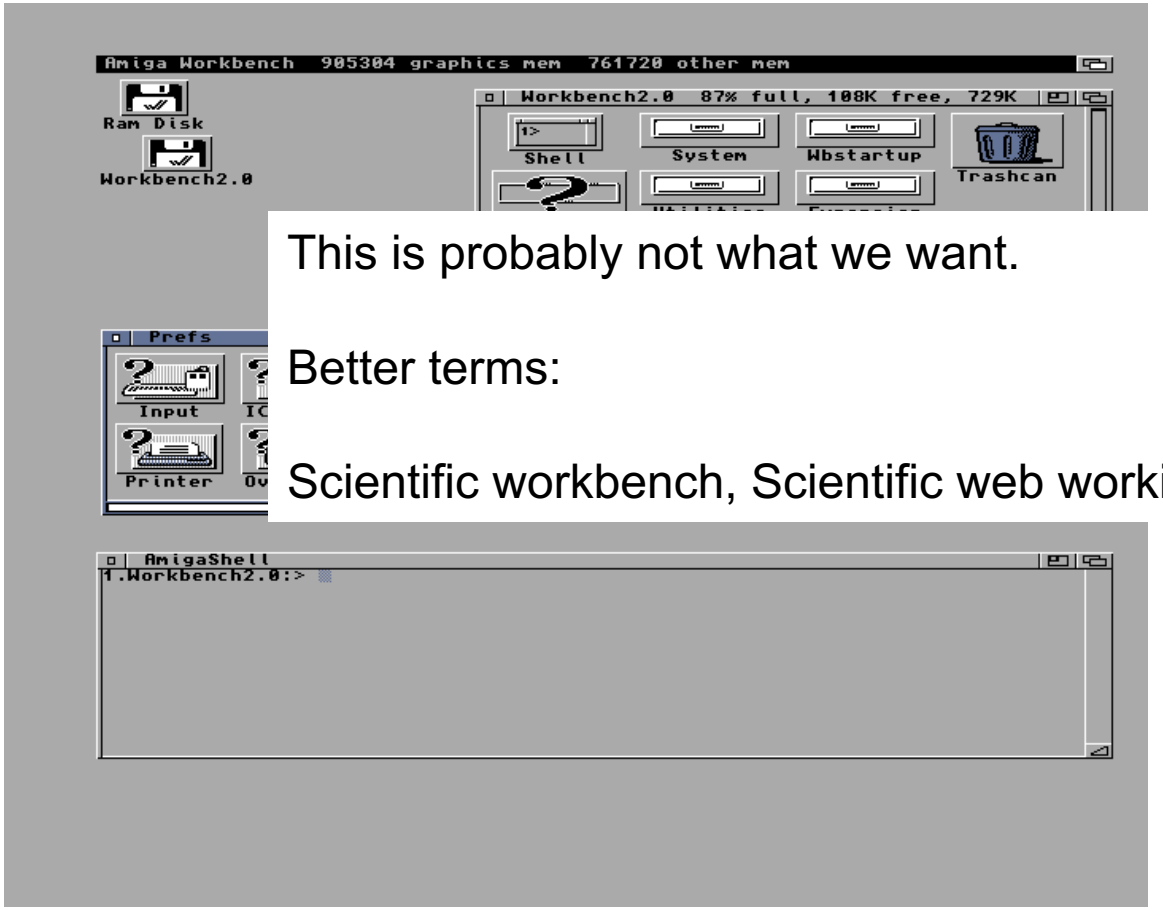
This introduction is about...

- The User Interface Topic Group
- User Interfaces
- This Workshop

A User Interface



A User Interface

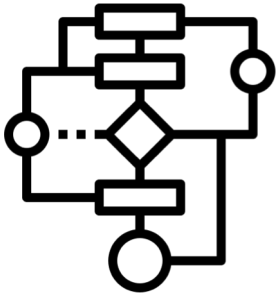


This is probably not what we want.

Better terms:

Scientific workbench, Scientific web working environment

Scientific Workflow



Develop own analysis software and algorithms

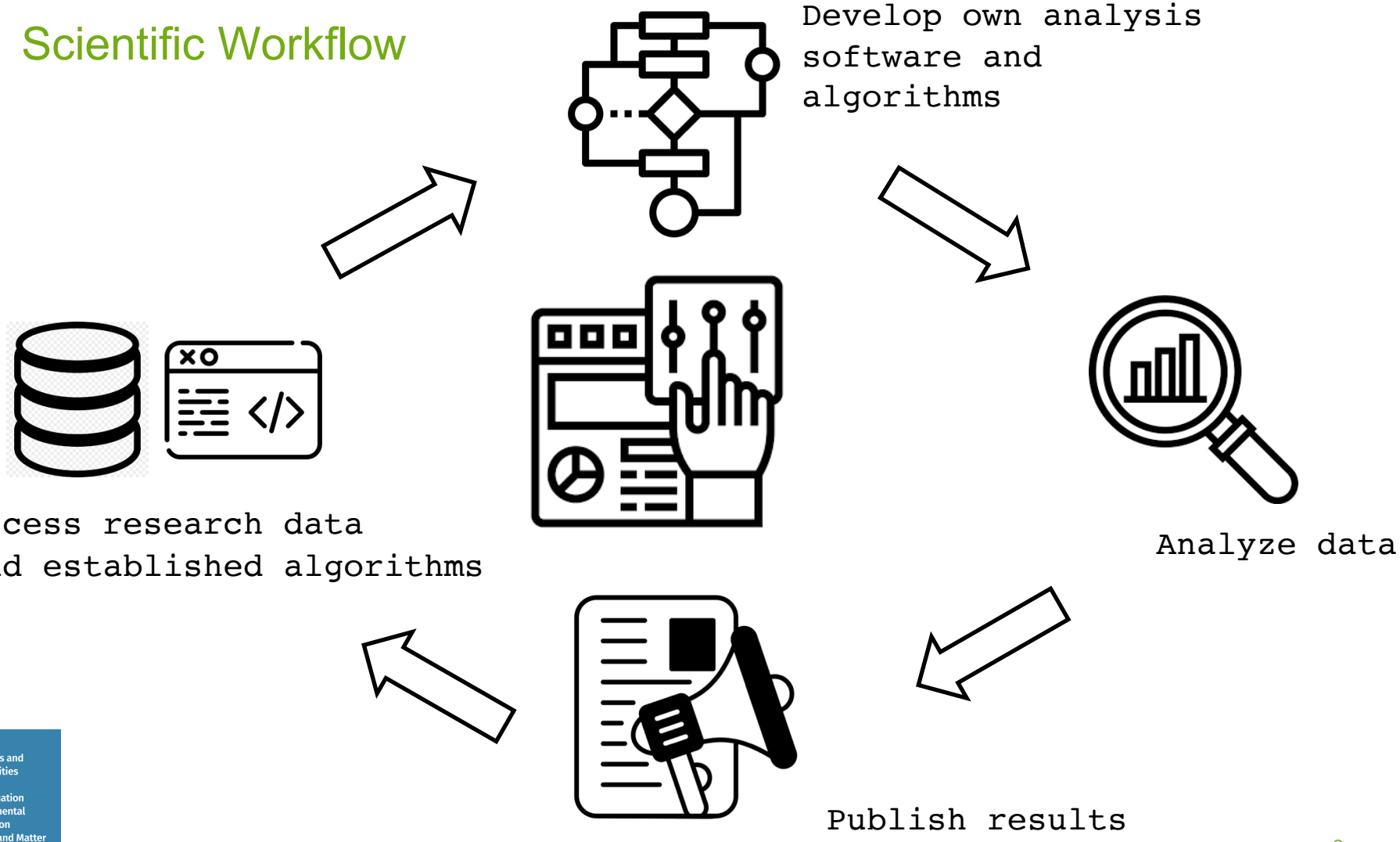
Analyze data

Publish results

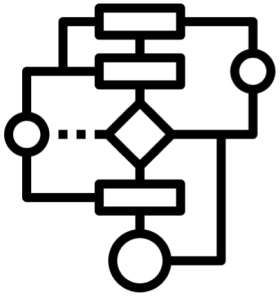
Access research data and established algorithms

Challenges and Opportunities of Digital Transformation in Fundamental Research on Universe and Matter

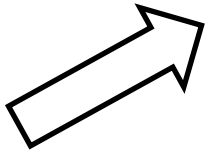
Scientific Workflow

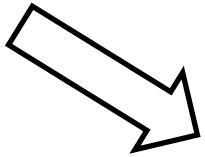


Scientific Workflow



Develop own analysis software and algorithms



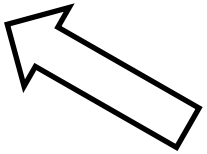


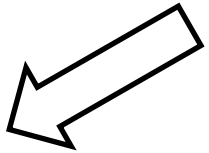
Let‘s test this workflow!

Let‘s plot some published data!

Access research data and established algorithms

Analyze data





Publish results

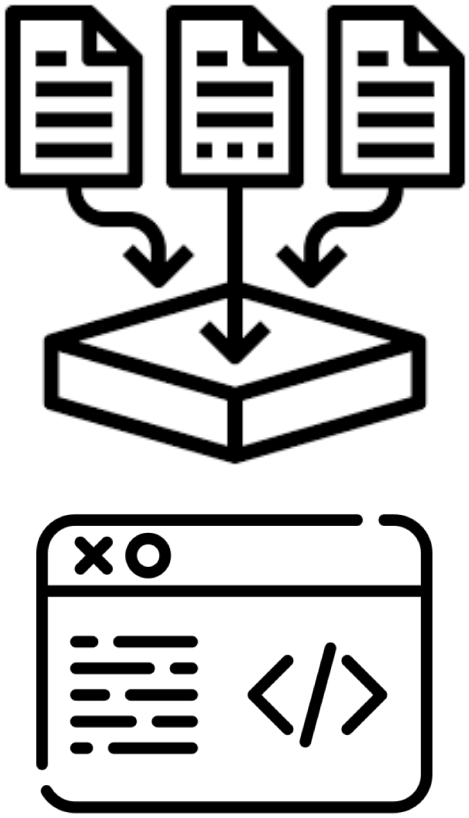
Plotting Published Data Part I: Accessing the Data

What I would like to plot:

- 3 different experimental results
- 1 theoretical model
- 1 direct simulation

*: Work that I did myself or work by collaborators that I was lucky enough to have on my computer.

Accessing the Data: Lessons Learned

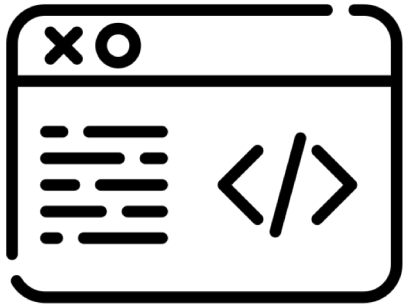


Multiple Sources for Data and Algorithms

- Collaboration
- Institution
- Working Group
- Publications
- Other Scientists
- ...

The work of people who left *a while ago.*

Accessing the Data: Lessons Learned



Multiple Sources for Data and Algorithms

Collaboration
Institution
Working Group
Publications
Other Scientists
...

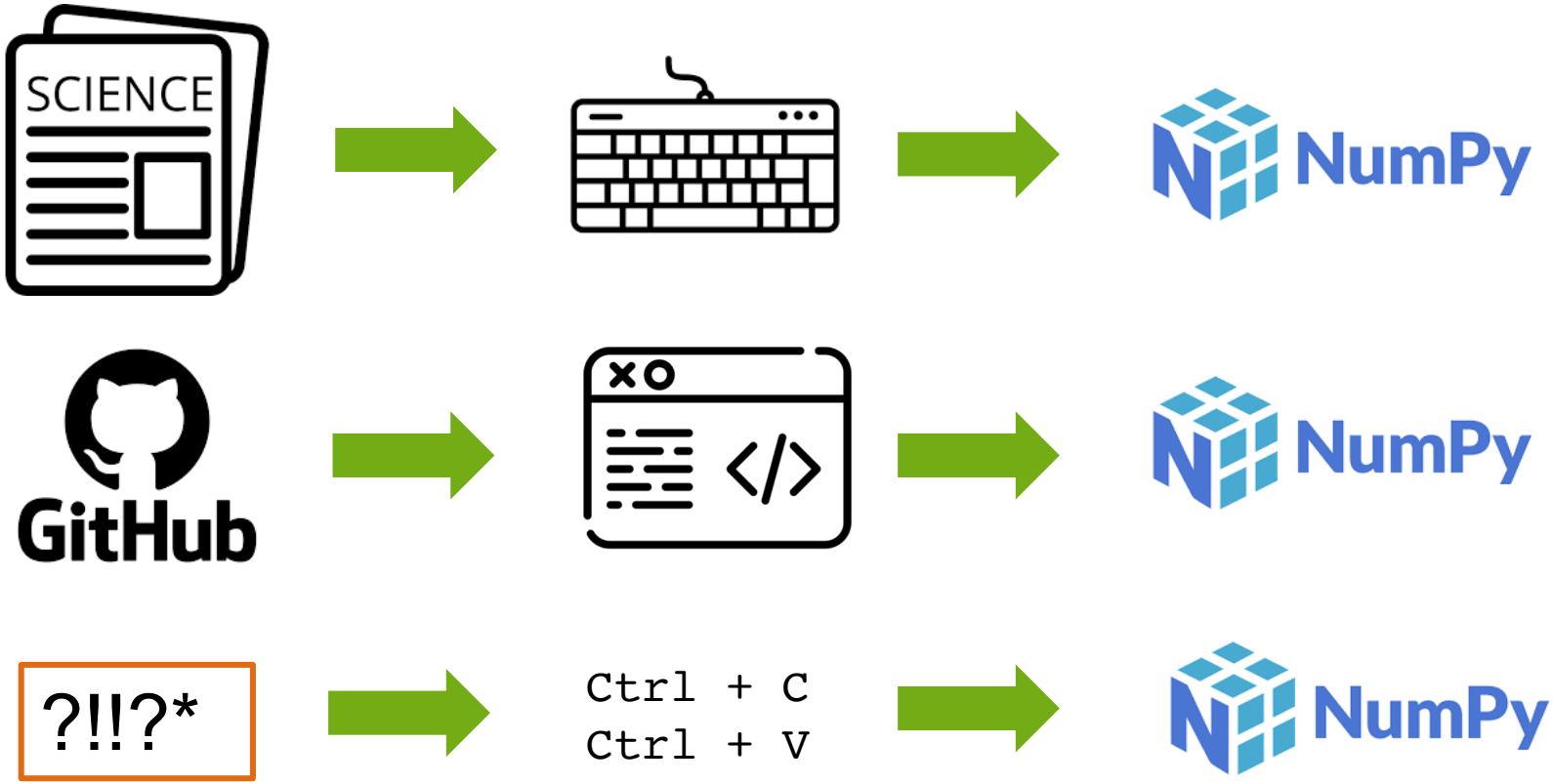
Hopefully relatively well documented.

The work of people who left *a while ago.*

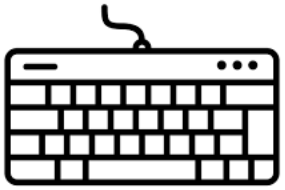
Maybe less well documented?!?

Data and algorithms generally not findable at the same place.

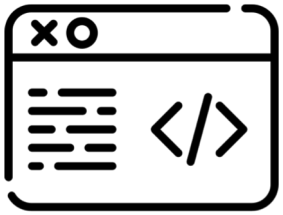
Plotting Published Data Part II: Writing the Script



Writing the Script: Lessons Learned



Error-prone!



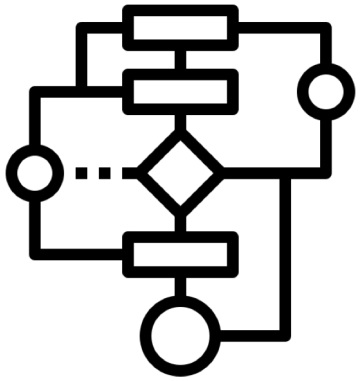
Duplication of work?

?!!?*

Ctrl + C
Ctrl + V

How do I know this is actually correct?

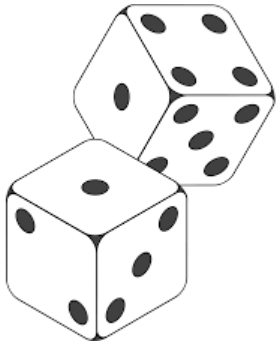
Plotting Published Data Part III: Making the Plot



Hardware/Software:

Where can I find the most efficient hardware for my algorithm?

How can this hardware be accessed?

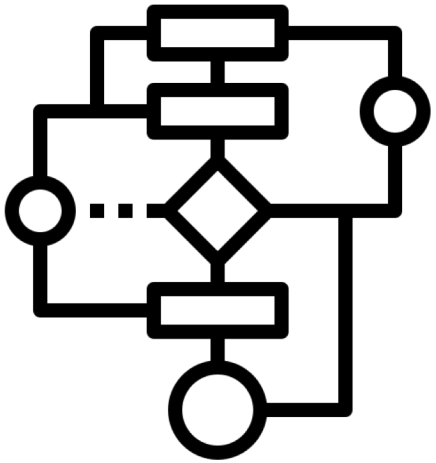


Using Simulations and numerical calculations:

Did someone else already run this type of simulation?

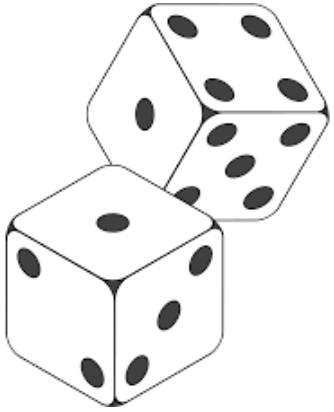
Do I really need to run all of it?

Plotting Published Data Part IV: Publishing the Results

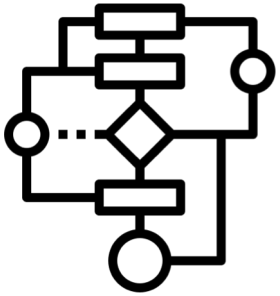


How to put all of this together to make it:

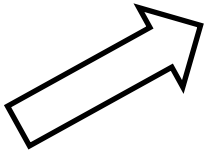
- Findable
- Accessible
- Interoperable
- Reproducible
- Reliable
- ...

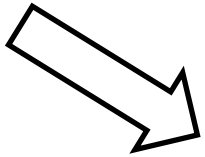


Scientific Workflow



Develop own analysis software and algorithms

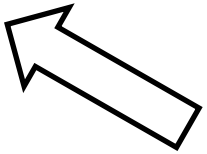


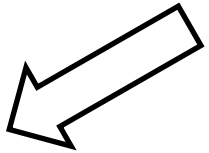


Making this work in practice requires a communities effort!

Access research data and established algorithms

Analyze data



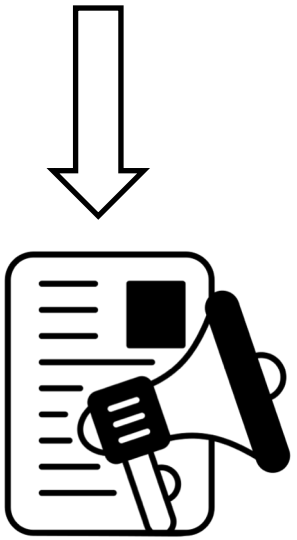


Publish results

Challenges and Opportunities of Digital Transformation in Fundamental Research on Universe and Matter

ErUM and the ErUM communities

Turn Big Data into Smart Data!

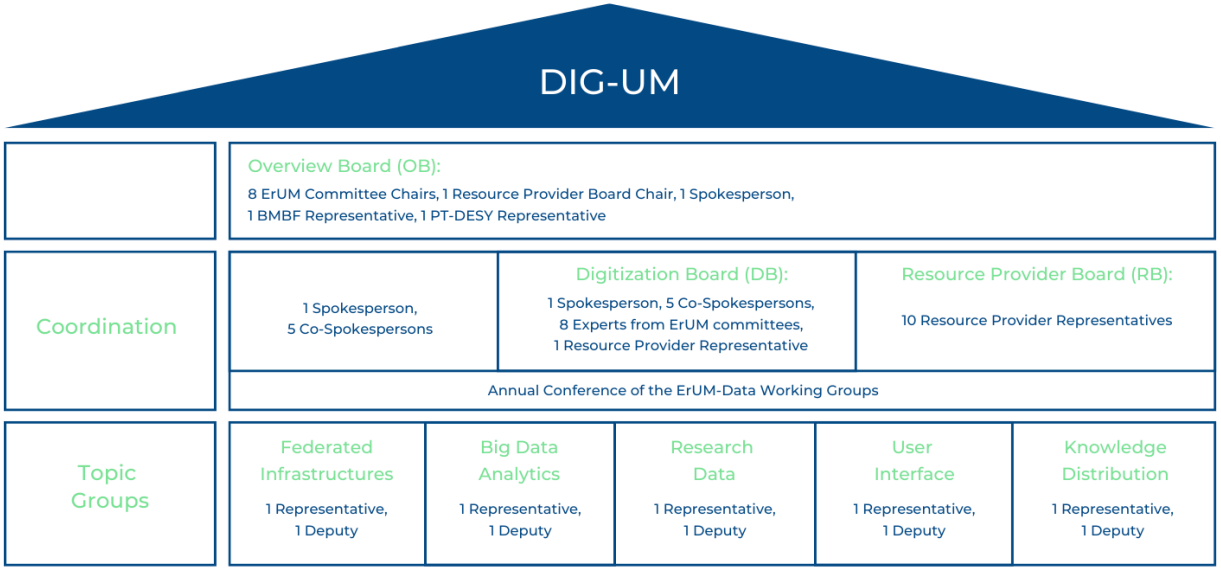


- Acronym for Research on Universe and Matter (Erforschung von Materie und Universum)
- Brings together approx. 10k scientists from different communities
- Upgrades in experimental facilities lead to an increased data rate (data avanlache)
- These data need to be stored, analyzed, processed and curated (by people)
- Big Data needs to be turned into Smart Data
- User-oriented digitization is mandatory to maintain a leading role in a highly competitive field

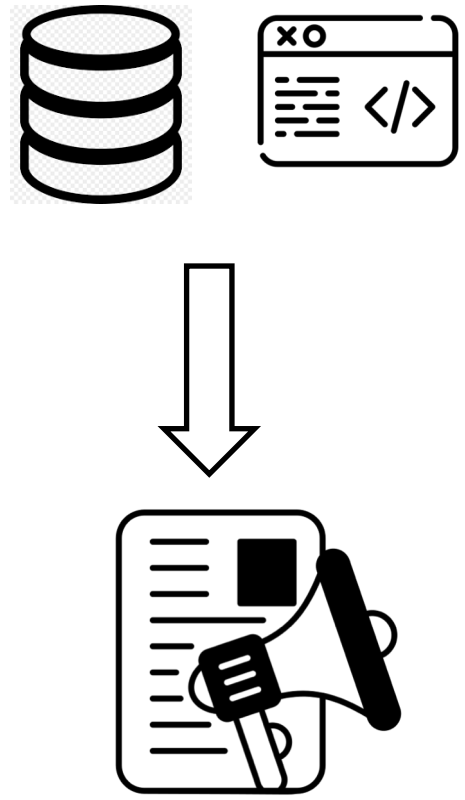
ERUM DATA HUB

DIGUM

- Digital Transformation in the Research on Universe and Matter
- Bottom-up approach of the ErUM scientists for synergetic work on the digital transformation

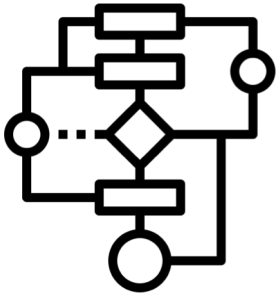


Turn Big Data into Smart Data!



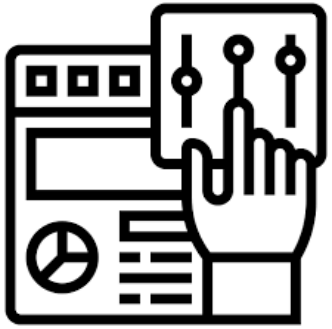
ERUM DATA HUB

The DIGUM Topic Groups



Big Data Analytics

Federated Infrastructures



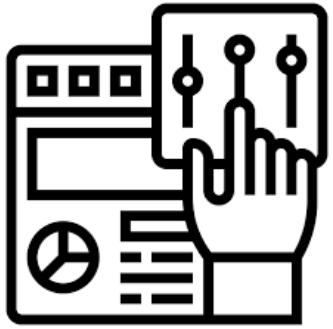
Knowledge Distribution

Big Data Analytics

Federated Infrastructures

Research Data Management

Topic Group User Interface: Goals

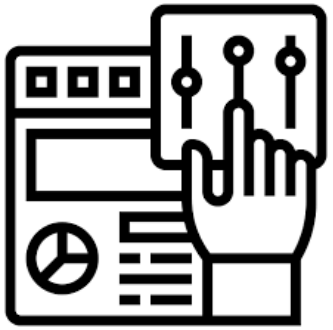


- Build and maintain a web-working environment for scientists
- Requires experienced scientists with knowlegde on research infrastructures and experience in analyzing research data
- Get people behind the mission
- Connect with NFDI and EOSC
- Act as a multiplier
- Bring people together a build a community (of communities)
- **Let scientists focus on science!**

What the User Interface should include

- Place where scientists can find the research data they are interested in
- Marketplace for algorithms
- Comprehensive integration of core functionalities
  - Software development and management
  - Integrated workflow management
  - Analysis preservation
  - Software Portals
  - Communication systems
- This needs the work and input from scientists experienced in working with research data
- Prototypes exist (also in individual collaborations) and people do science this way (every day)
- But we can probably do better!

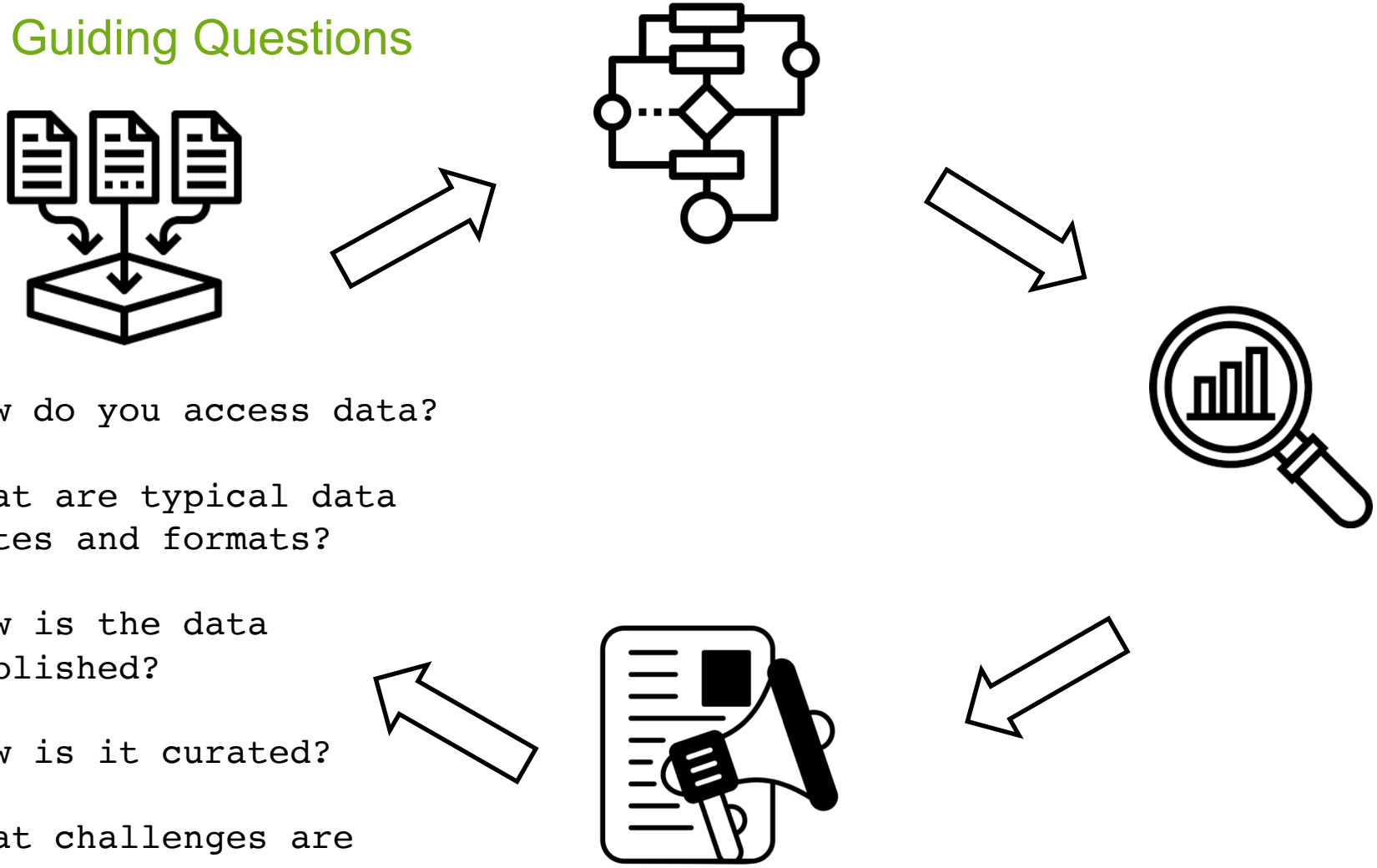
Topic Group User Interface: People



- Pierre Schnizer
- Judith Reindl
- Harry Enke
- Kay Graf

- Join our bi-weekly phone calls Tueday at 9.30
- Join our mailinglist!
  https://wiki.erumdatahub.de/de/mailing-lists

Guiding Questions



How do you access data?

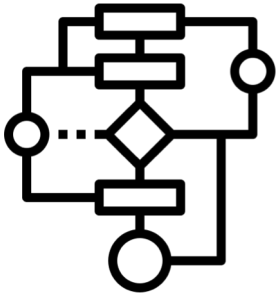
What are typical data rates and formats?

How is the data published?

How is it curated?

What challenges are anticipated for the future?

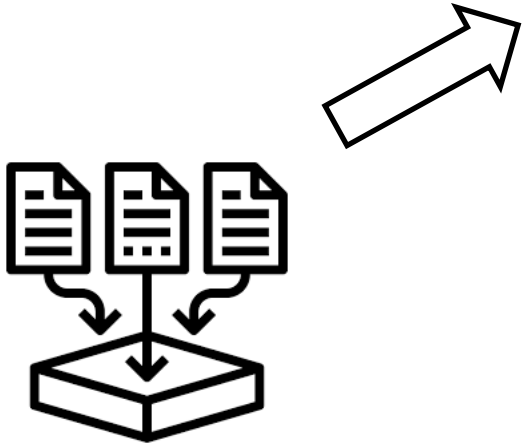
Guiding Questions



How are software and algorithms developed maintained and shared?

How can the reliability of the code be ensured?

How to avoid the duplication of work across different communities?



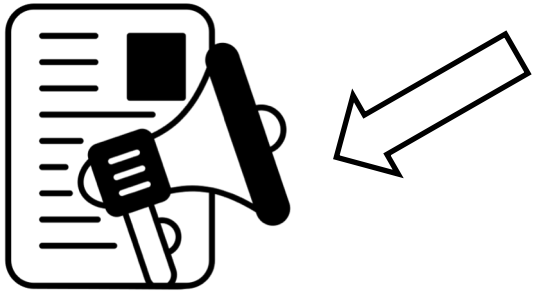
How do you access data?

What are typical data rates and formats?

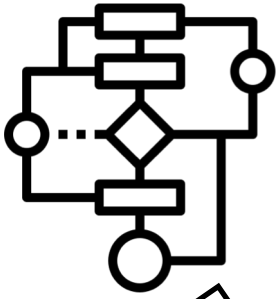
How is the data published?

How is it curated?

What challenges are anticipated for the future?



Guiding Questions



How are software and algorithms developed maintained and shared?

How can the reliability of the code be ensured?

How to avoid the duplication of work across different communities?

How do you access data?

What are typical data rates and formats?

How is the data published?

How is it curated?

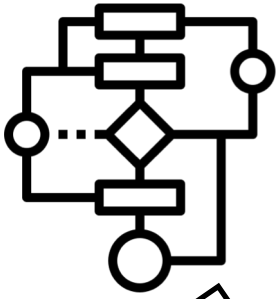
What challenges are anticipated for the future?

How to avoid duplicating simulation?

How to find and access the most efficient hardware?

How to avoid duplicating analyses?

Guiding Questions



How are software and algorithms developed maintained and shared?

How can the reliability of the code be ensured?

How to avoid the duplication of work across different communities?

How do you access data?

What are typical data rates and formats?

How is the data published?

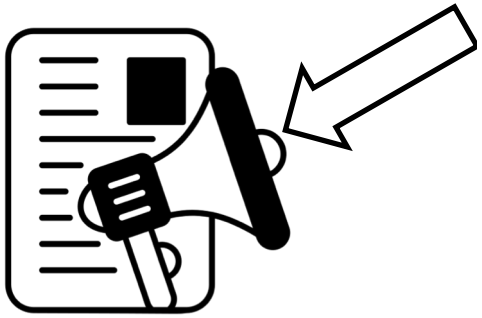
How is it curated?

What challenges are anticipated for the future?

How to avoid duplicating simulation?

How to find and access the most efficient hardware?

How to avoid duplicating analyses?



How to share data and software across different communities?

How to make data FAIR?

How to encourage people to actually do this?

This Workshop: World Cafè

Community Requirements:

What do the individual communities need?

Data Sources:

What are typical data rates, data volumes and catalogs?

Real World Experience:

Experience today and challenges tomorrow

Design Patterns:

Common cross-community patterns

This Workshop: Keynotes

- **Nicolas Eich**: VISPA: Data Analysis in the Web Browser powered by JupyterLab
- **Verena Kain**: CERN's control system approach for machine learning for accelerators
- **Mohammad Al-Turany:** ALFA
- **Kai Polsterer:** Unsupervised ML to explore data: lessons learned from learning machines

Workshop Goals

- Discuss and interact
- Identify common patterns and requirements, which should not be the smallest common denominator
- Share experiences, best practices and failures

- Community Building

Welcome.