The PUNCH4NFDI Consortium

Particles, Universe, NuClei and Hadrons for the NFDI

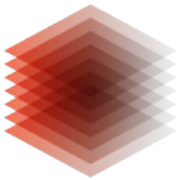
Christiane Schneide (DESY) for the PUNCH4NFDI Consortium

Big Data Analytics workshop, 24.02.2023

Who We Are

Universities, Helmholtz, Max Planck, Leibniz





Who We Are

43 Partners
20 Co-applicants
23 Participants

Currently more than 100 people involved in PUNCH4NFDI

Representing close to 10000 scientists in Germany (KAT, KET, KHuK, RdS)

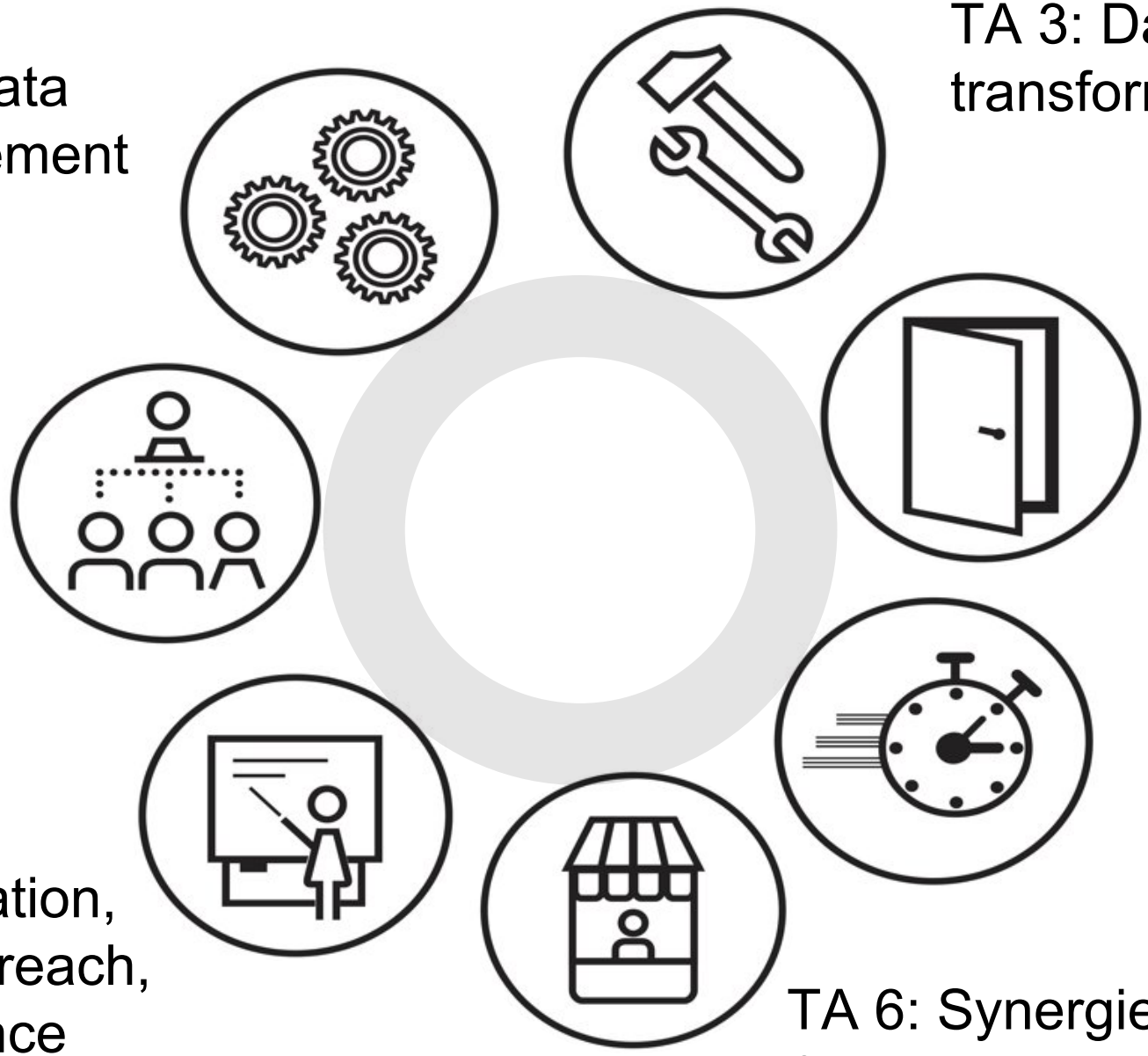
KAT, KET, KHuK and RdS representatives in our User Committee

Task Areas

TA 2: Data management

TA 3: Data transformations

TA 1: Management and governance

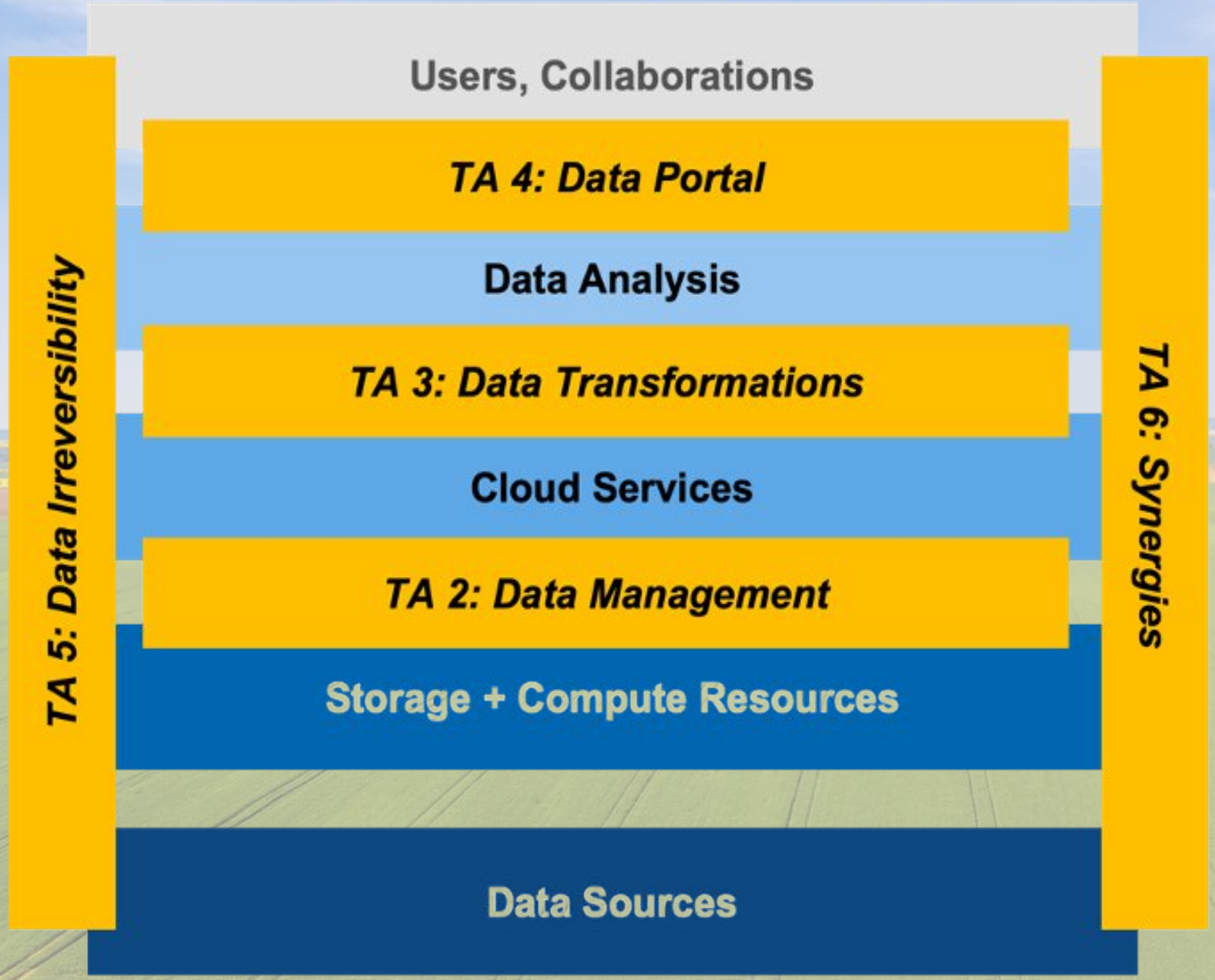


TA 4: Data portal

TA 5: Data Irreversibility

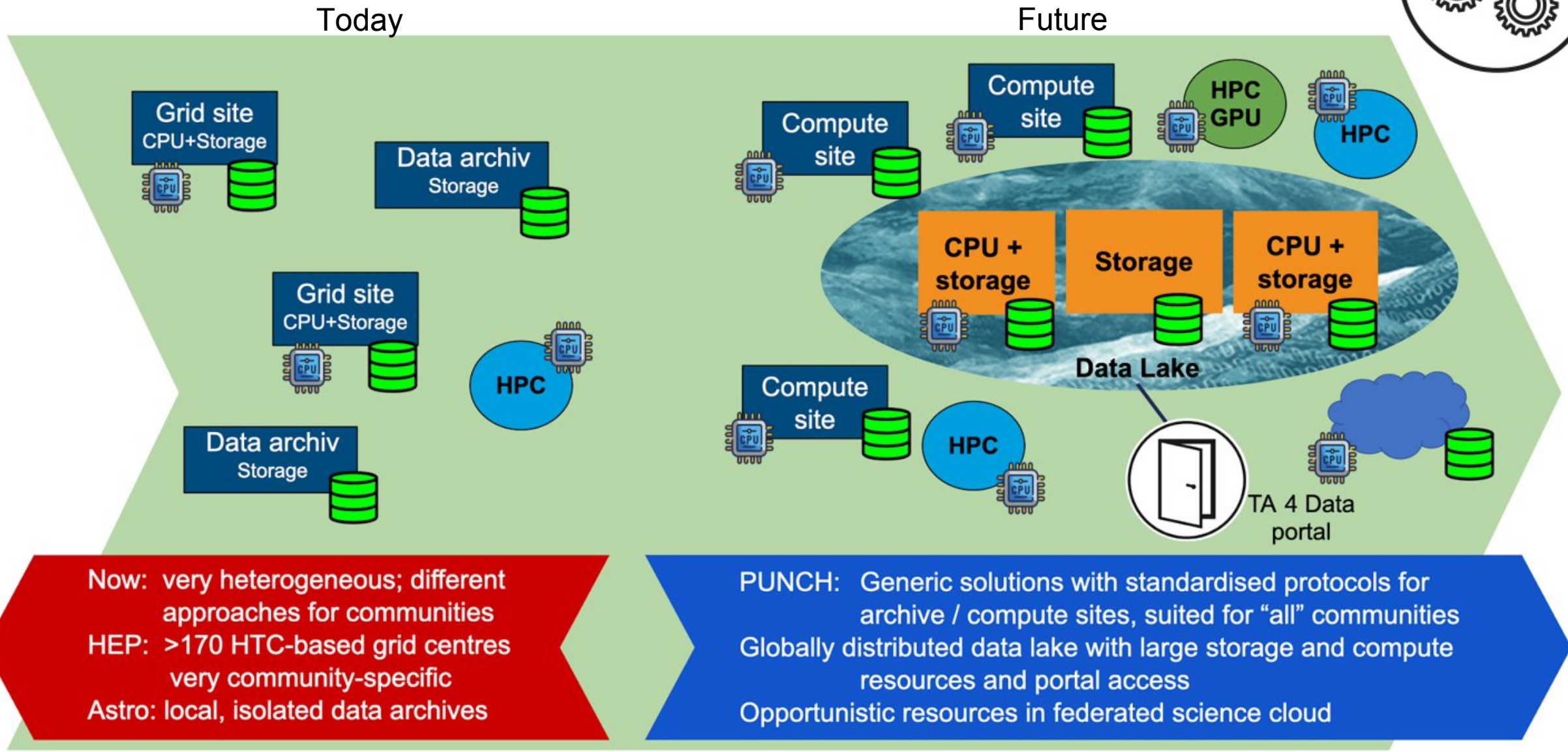
TA 7: Education, training, outreach, citizen science

TA 6: Synergies & services


Users, Collaborations
TA 4: Data Portal
Data Analysis
TA 3: Data Transformations
Cloud Services
TA 2: Data Management
Storage + Compute Resources
Data Sources
TA 5: Data Irreversibility
TA 6: Synergies

TA 2: Data management

Access to data, federated computing, automation, data lake prototype

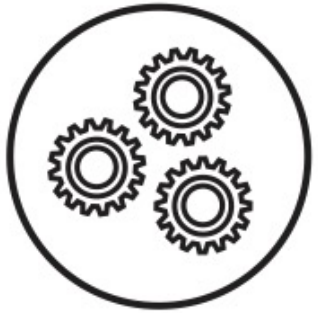


Now: very heterogeneous; different approaches for communities
HEP: >170 HTC-based grid centres very community-specific
Astro: local, isolated data archives

PUNCH: Generic solutions with standardised protocols for archive / compute sites, suited for "all" communities
Globally distributed data lake with large storage and compute resources and portal access
Opportunistic resources in federated science cloud

TA 2: Data management

Main topics & Achievements



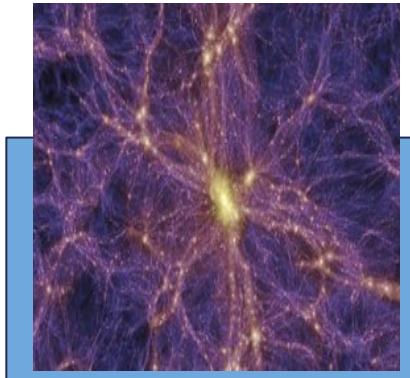
- Storage4PUNCH prototype
  - dCache based system at DESY
  - XRootD based system at U Bonn
  - Another system is being prepared

- Prototype of federated Compute4PUNCH
  - Dynamic integration of two compute sites
  - Two more compute sites will follow soon

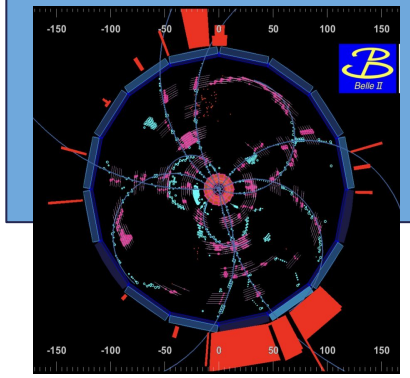
- Login node available to all PUNCH members
- Container registry available
- Test/ demo of Metadata Catalogue at two sites
- Prototype of Data lake monitoring infrastructure

TA 3: Data transformations

Integration of common tools into a data infrastructure based on code-to-data principle
Provision of tools for parallel processing of huge data sets on heterogeneous resources



Data and metadata

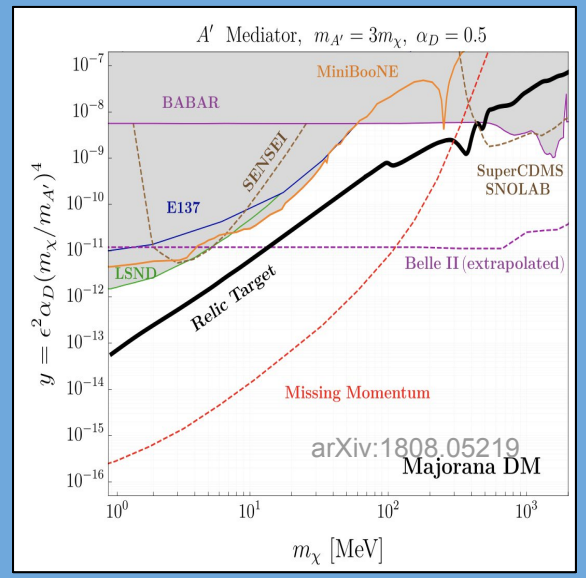


User
Selection of tools / workflows via portal

Image processing	Particle selection	Data modelling

Statistical methods	Numerical methods + simulations	Machine learning	Across-dataset methods

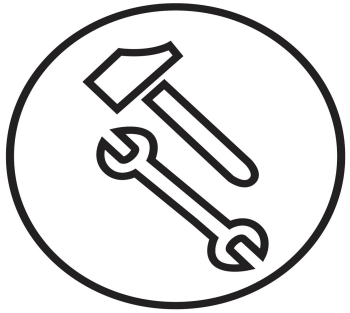
Scientific result



Tools common to many science fields

TA 3: Data transformations

Main topics & Achievements



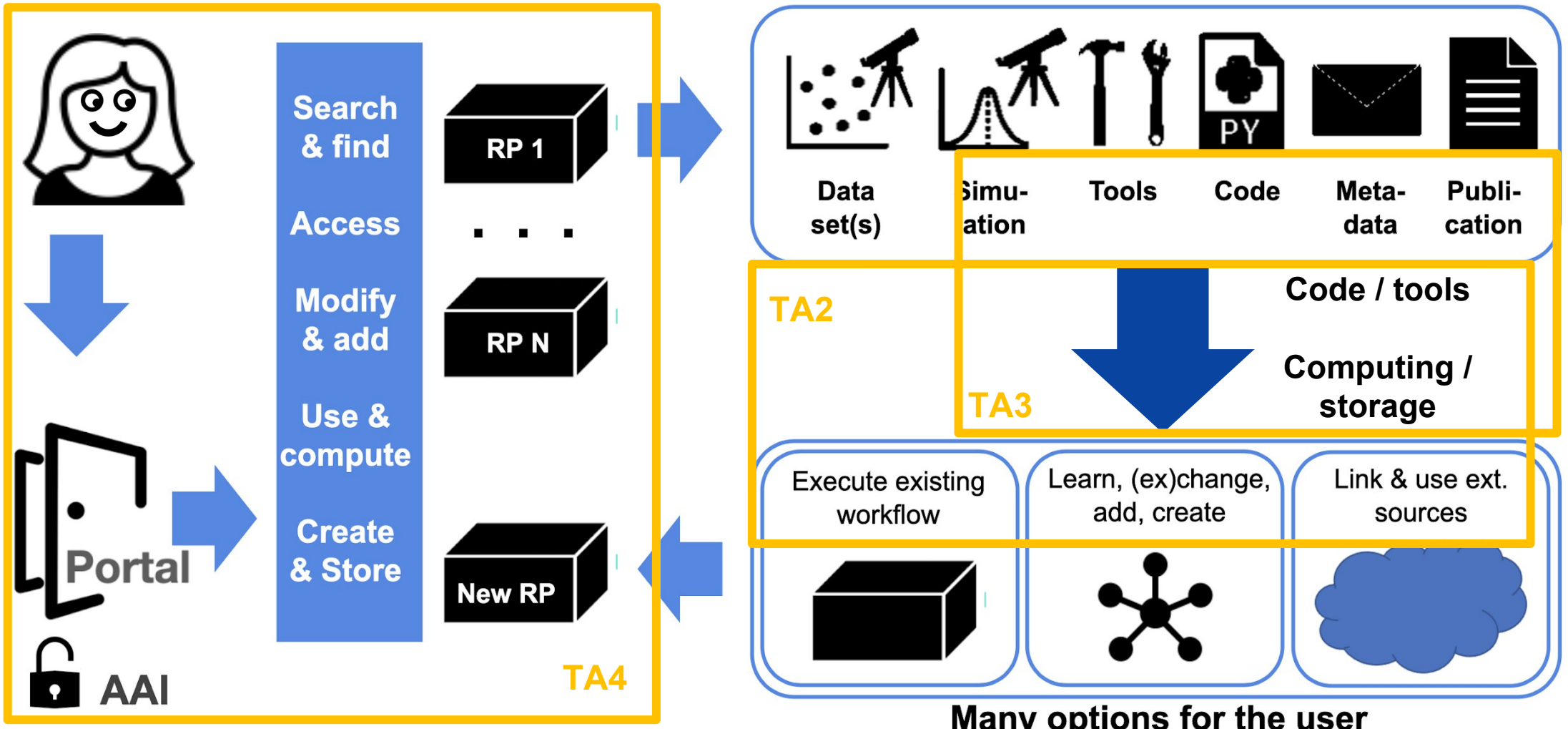
- Ongoing development of BAT.jl
  - BAT.jl to python interface (batty) https://github.com/bat/batty
  - Implementation of ML driven space transformations for MCMC sampling in BAT.jl

- Simulation codes in Astrophysics
  - User survey (short summary here)
  - List of commonly used codes is available
  - Deployment, workflows and (scaling) benchmarks of important codes at HPC centers

- Similar work ongoing for lattice QCD codes
- Evaluation of workflow management systems https://arxiv.org/abs/2212.01422
- Implementation of workflows on Compute4PUNCH

TA 4: Data portal

PUNCH-SDP



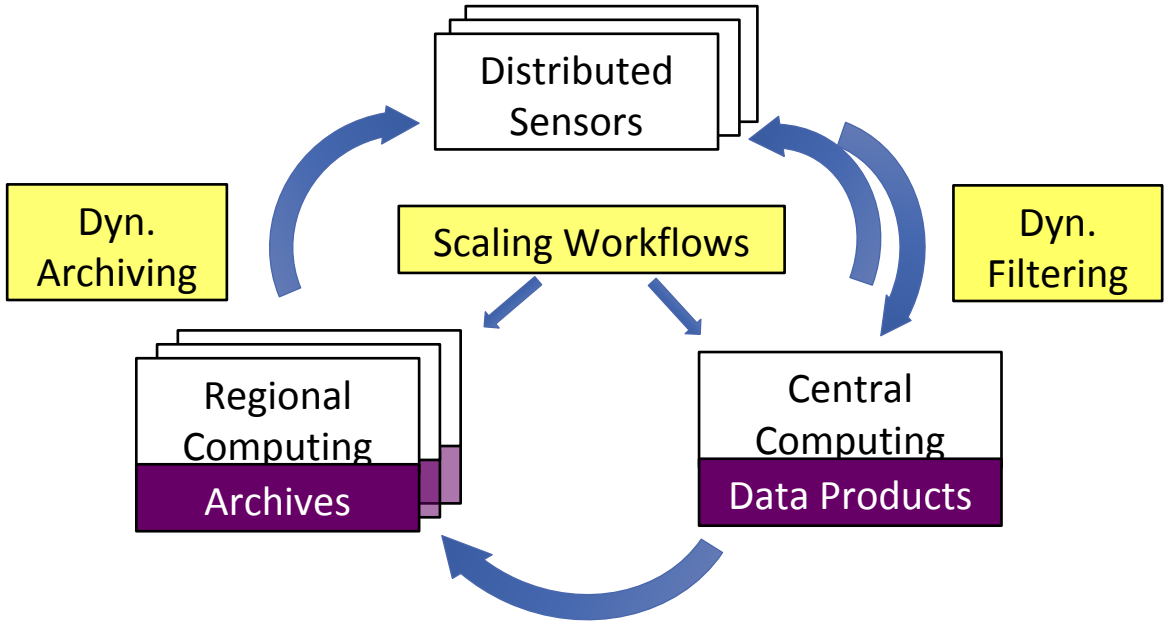
TA 4: Data portal

Main topics & Achievements

- Prototype metadata describing the interaction of software and data
- Collection of elements of DRPs (digital research products)
  - Functional diagram for a DRP registry/ database
  - Metadata schemas from different communities
  - PID for cross community identification of (public) data
- Docker and Kubernetes infrastructures
- Gitlab + Continuous Integration
  - Code Repository
  - Code Registry
  - Package Registry
- REANA (with support of Jupyter notebook)
- Intranet & results page

TA 5: Data irreversibility

IT problems of tomorrow ... solved today with high-energy physics and astrophysics



dynamic data life cycle

Nowadays: (most) data are stored and re-analysed over and over again.

Soon: only a small fraction of data can be stored long term
⇒ irreversible loss of information

Solutions:

- dynamic filtering: extraction of relevant information from huge data streams in real time (without human assistance, e.g. machine learning algorithms).
- dynamic archiving: feedback from offline analysis to sensor controls.
- scaling: increasing collection of information by sensors leads to huge individual data objects. For the analysis of this kind of data we need a paradigm shift:
  from process oriented to storage oriented computing.
- reproducability: reconstruction of how and why specific decisions were made in real time. Simulations are critical for validation and understanding.

TA 5: Data irreversibility

Main topics & Achievements

- Metadata concept
  - Hierarchic and dynamic/ flexible metadata
  - Capturing of filtering process & complex workflows
  - Include chain of algorithms in pipeline/ trigger process
  - ➢ Enable reproducibility of results at different levels of processing

- Machine Learning on FPGAs
  - Dedicated meetings
  - Workshop in November 2022
  - Identification of useful tools for ML on FPGAs

- Development of a new scalable pipeline for pulsar analysis

TA 6: Synergies & services

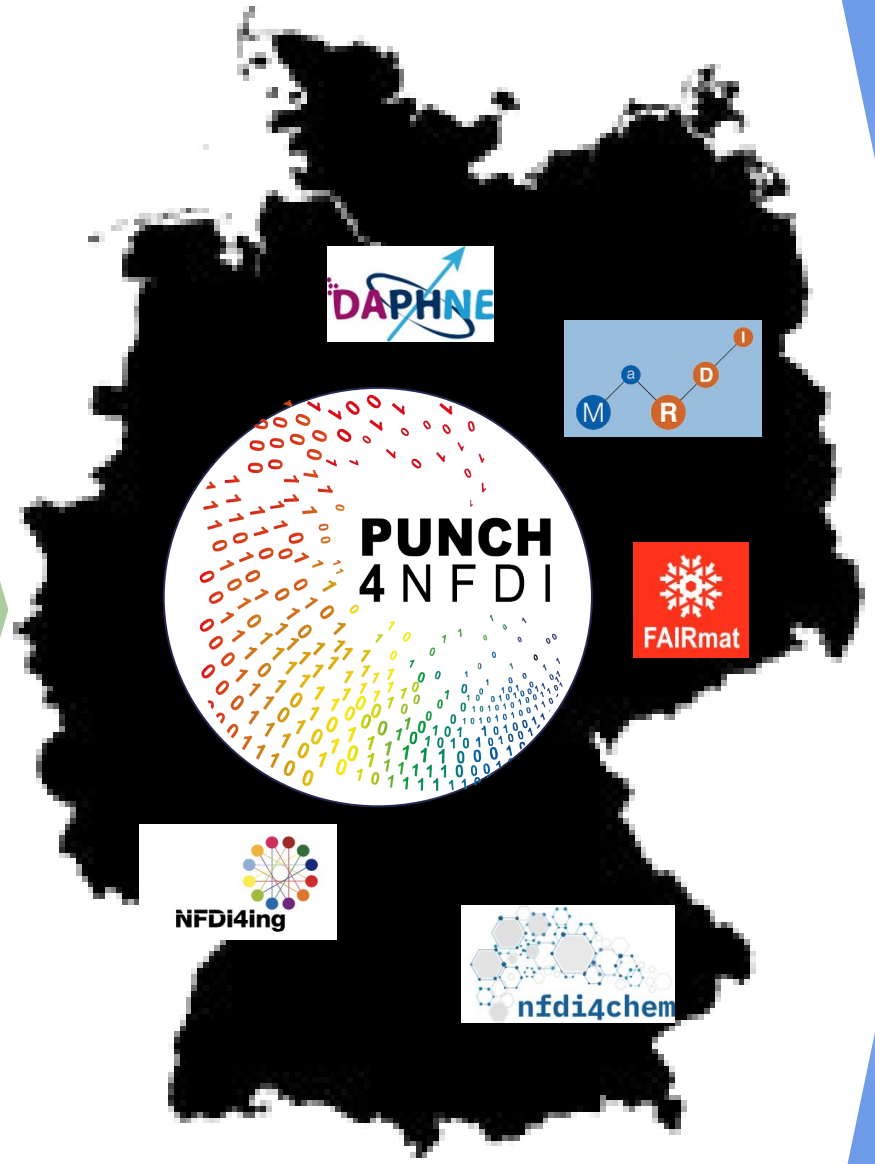
Section Common Infrastructures

Section Training and Education

Section Ethical, legal and social aspects

Section (Meta)data, Terminologies, Provenance

DAPHNE4NFDI & FAIRMat & DPG: address physics-specific aspects



Marketplace, PUNCH-SDP, data portal

Knowledge fabric, digital research products, metadata services

AAI infrastructure, dynamic disk caching

Big data management & data storage services

Machine learning services and real-time applications

IT resources via Compute4PUNCH interactive analysis interface

Cloud-based testbed

Teaching and education

(Non-exclusive examples)

TA 6: Synergies & services

- PUNCHLunch seminars, Thu 12:30-13:30 https://indico.desy.de/category/897/
- arXiv software repositories and arXiv study on software products used in PUNCH sciences
  - Evaluation is still ongoing

- List of open source analysis tools
- List of statistical methods used in PUNCH
- Topics on PUNCH-AAI, close contact to Base4NFDI IAM group
- Management/ tracking of involvement in NFDI sections and working groups
- Planning cooperation with other consortia & projects
  - Common booth at ISC with ErUM FSP and FIDIUM

TA 7: Education, training, outreach and citizen science

For	Training	Education	Outreach	Citizen Science
	Experts	Universities: lecturers and students	Scientists, media, schools, public	Amateur, public
Goal	foster expertise and career prospects	focused education and career promotion	communicate, foster young talents, strengthen schools training of communication, school-academy-network, events, ressources	foster commitment and deeper understanding, democratise science
Means	provisioning of data and educational ressources	educational and data ressources, on-site and online seminars		online projects and campaigns

TA 7: Education, training, outreach and citizen science

- PUNCH Young Academy
  - One-to-one mentoring
  - Soft skill workshops
  - Technical workshops

- Toolbox for Science Communication Zenodo
- Exhibit at MS Wissenschaft in 2023
- Data literacy workshop at jDPG seminar
- Machine Learning Masterclass (in collaboration with Netzwerk Teilchenwelt) – test, evaluation and preparation of material & documents
- Einstein@Home https://einsteinathome.org/

Thank you!

The PUNCH4NFDI Consortium

Spokesperson:
PD Dr. Thomas Schörner (DESY, thomas.schoerner@desy.de)
DESY, Notkestr. 85, D-22607 Hamburg

Contact:
Mail: punch4nfdi@desy.de
Web: www.punch4nfdi.de
Twitter: @punch4nfdi