The PUNCH4NFDI Consortium

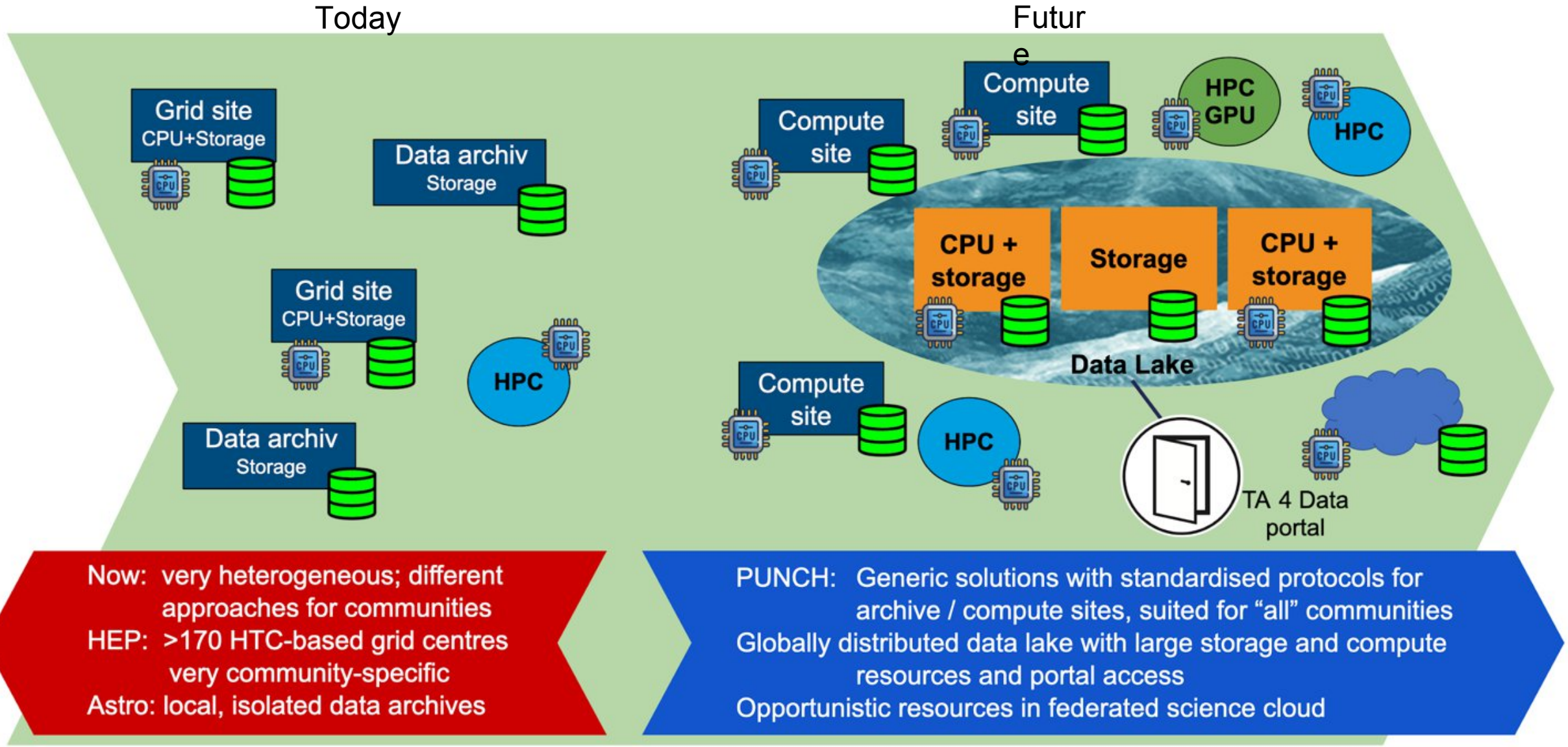
Particles, Universe, NuClei and Hadrons for the NFDI

Christiane Schneide (DESY) for the PUNCH4NFDI Consortium

FIDIUM Collaboration Meeting, 21.10.2022

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

Test systems in preparation

Storage4PUNCH

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

Compute4PUNCH

- Login noe at KIt using tokens (no local account required)
- CPU resources included located at KIT, Bonn, Münster – other locations being prepared

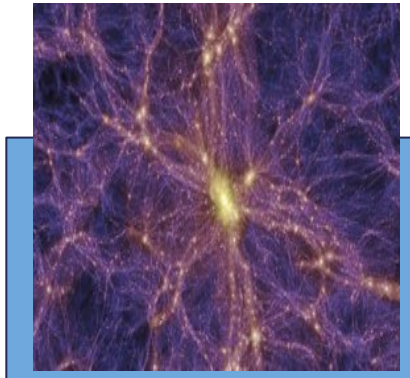
We will have some demonstrations of these systems on Thursday.

Metadata Catalogue

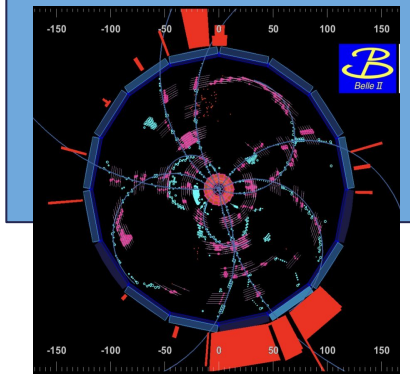
- Catalogue with flexible schema
- Initial focus on LQCD metadata and related applications
  - Development system now setup
  - Planning for other applications beyond LQCD

Task Area 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



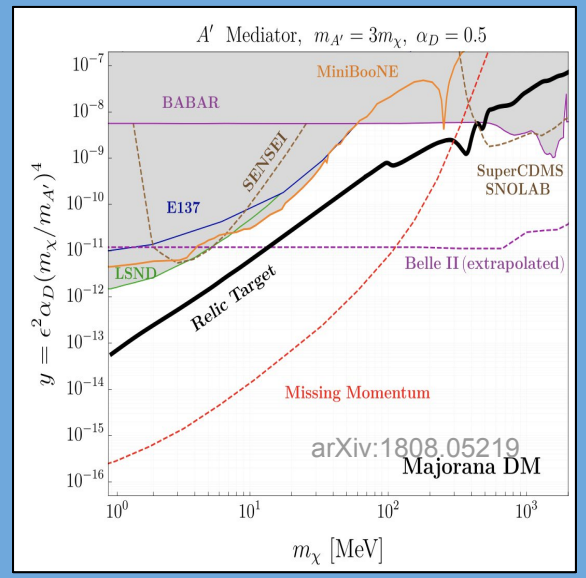
Data and metadata



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

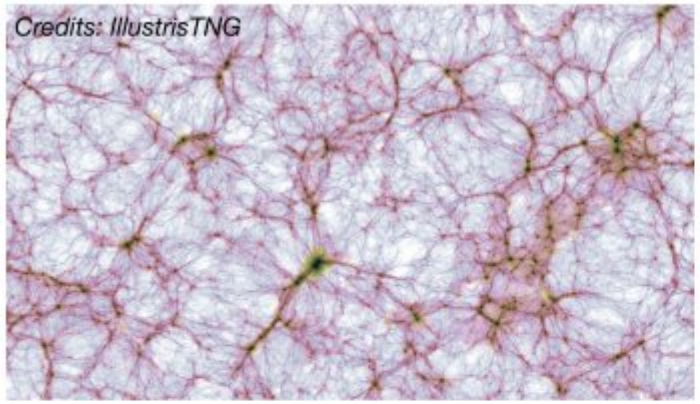


Tools common to many science fields

TA 3: Data transformations

The problem

Nicola Malavasi
(LMU München)

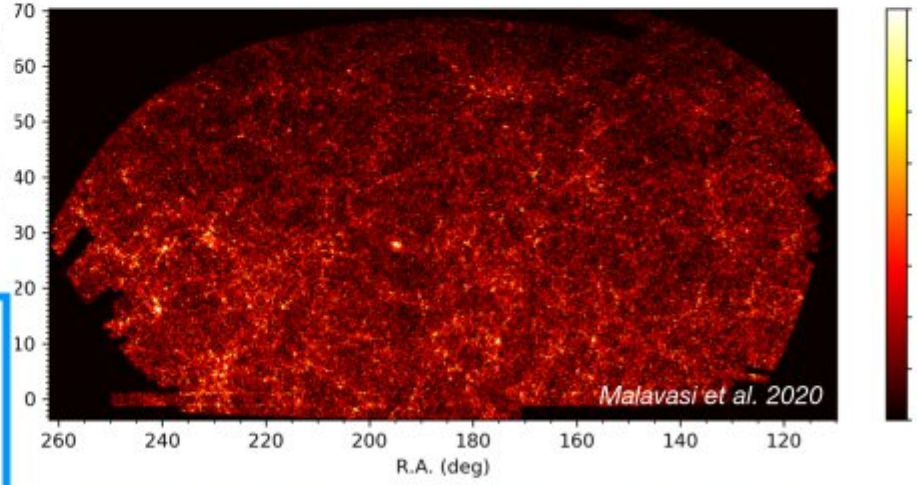


The cosmic web is a network of clusters and filaments made of galaxies, dark matter, and gas that fills the Universe.

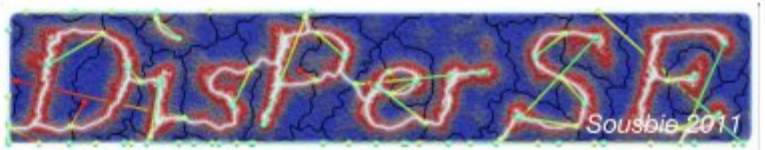
Detecting the cosmic web from the galaxy distribution for galaxy evolution and cosmology studies requires several ingredients.

Large galaxy surveys to sample vast regions of the Universe (e.g. SDSS).

Complex algorithms to trace anisotropic and multi-scale structures (e.g. DisPerSE).



How do we apply complex algorithms to large data sets in a way which is efficient, easily implementable, reproducible, and easy to modify and build upon (e.g. for future analyses)?

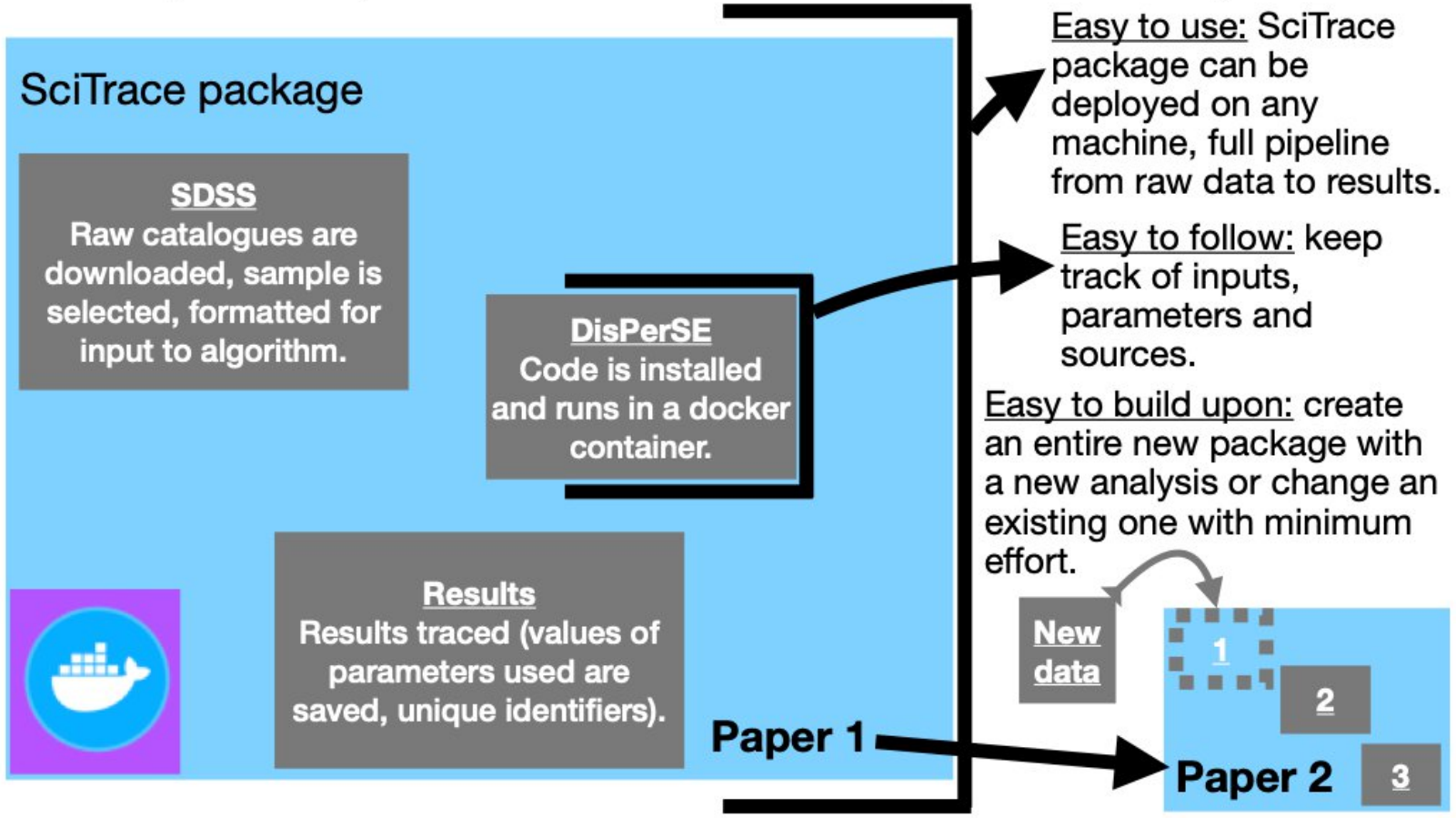


TA 3: Data transformations

Nicola Malavasi
(LMU München)

Our solution

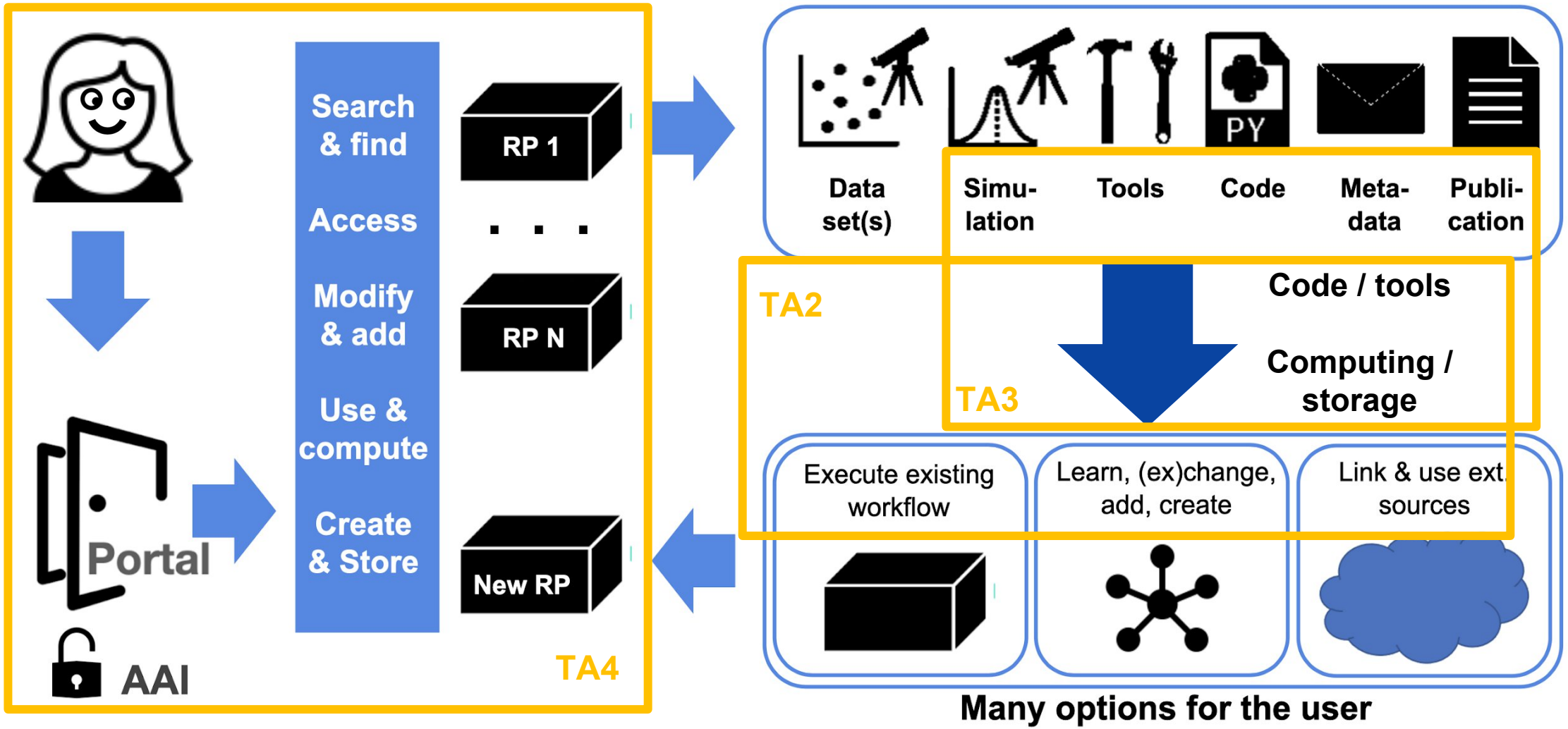
Using the SciTrace software (developed by Yori Fournier and Anastasia Galkin at AIP Potsdam) to apply the DisPerSE algorithm to the SDSS survey. Goal: reproduce Malavasi et al. 2020a, b and build upon that.



co-operation of TAs 2, 3 and 4

TA 4: Data portal

PUNCH-SDP



TA 4: Data portal

Collected elements of the DRP

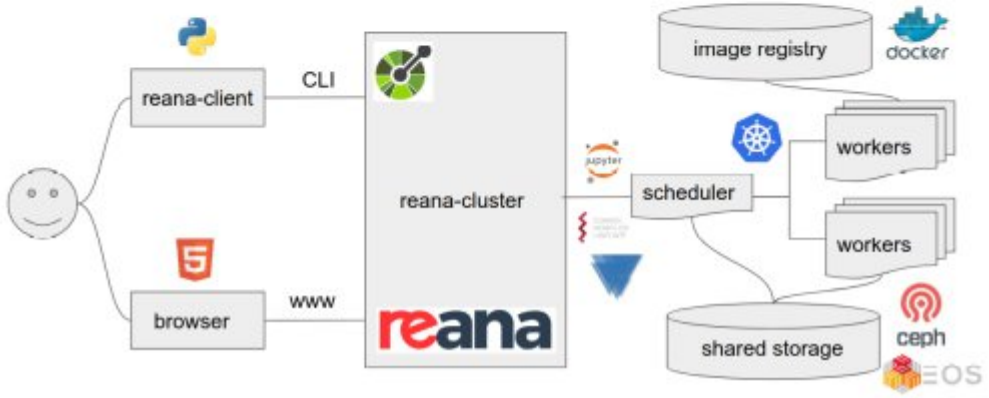
- Collected metadata schemas
  - IVOA (Astronomy)
  - ILDG (Lattice community)
  - CERN Open Data
  - Astro particles (?)

- Currently: “findability” aspect of FAIR
  - Evaluating usage of DOI and the Metadata Kernel of DataCity
  - Looking at registry concepts and implementations
  - OAI-PMH and other harvesting metadata protocols and APIs
  - Representation of xml and other MD formats

TA 4: Data portal

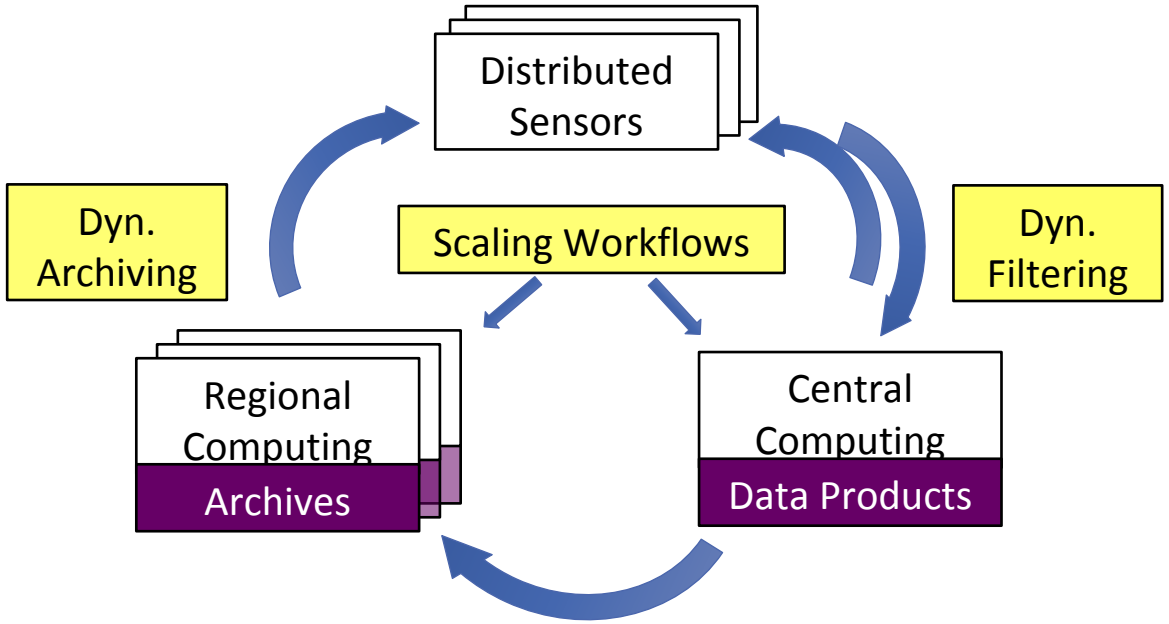
Collected elements of SDP

- PUNCH AAI
- Compute4PUNCH
- Storage4PUNCH
- S3 storage
- Docker and Kubernetes infrastructures
- Gitlab + Continuous Integration
  - Code Repository
  - Code Registry (in operation)
  - Package Registry (tbc)
  - Integration with
- REANA (with support of Jupyter notebook)
- Dashboard + Intranet (based on Gitlab)



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

- First deliverable „Metadata concept“ being worked on – input from all WPs required; experience and existing solutions from other (sub-)communities being considered and sought.
- First draft shared with other TAs for comments

- Highlight: good example for benefit of PUNCH for involved communities is the TA5 work on implementations of machine learning concepts on FPGA. Applications in both astronomy and hight-energy physics now being worked on. This link didn‘t exist before PUNCH.

TA 6: Synergies & services

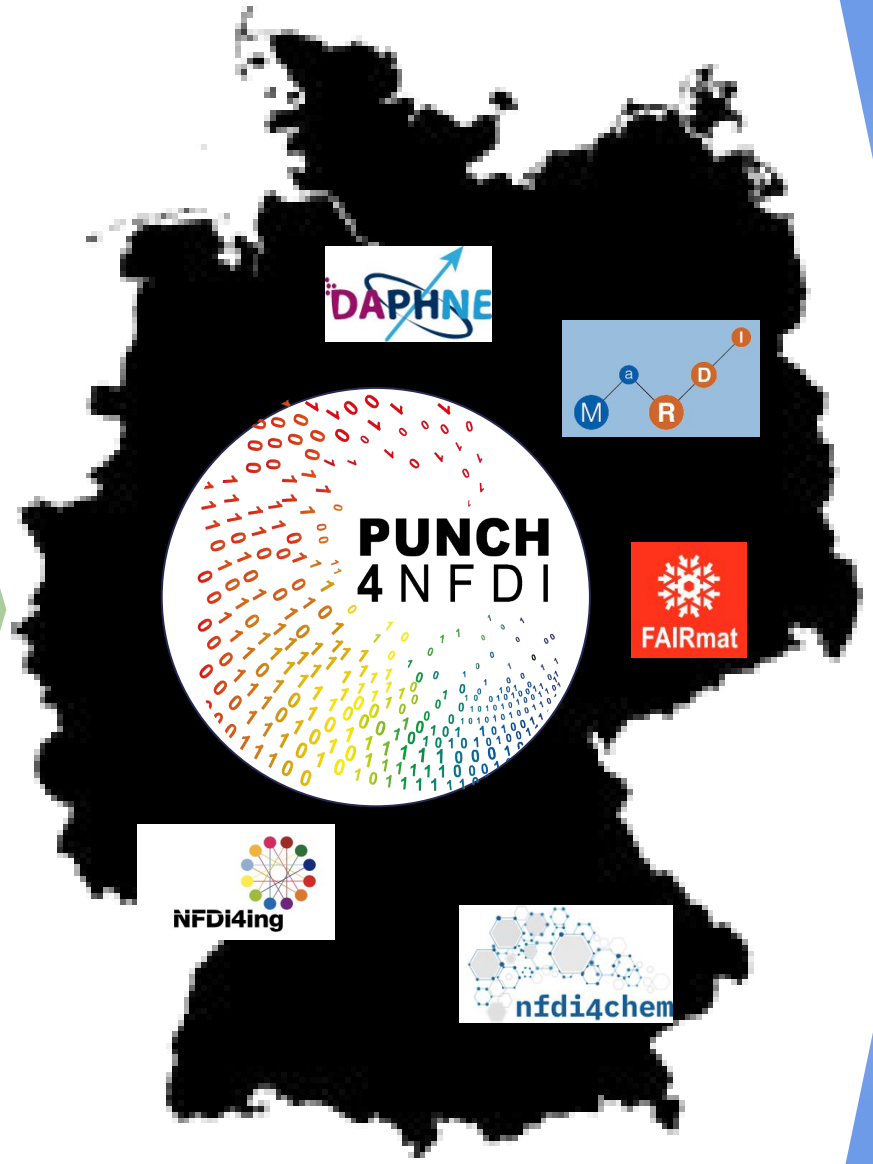
Section Common Infrastructures

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

WP1: Marketplace

- Booth at AG annual meeting (12-16 Sep 2022):
  posters, flyer, presentations
- Noticeboard set up (deliverable D-TA6-WP1-1)
- Query of archive repositories
- Archive access support for time domain data
- Interfacing with NFDI sections, NFDI software marketplace

WP2: AAI

- Deliverable D-TA6-WP2-3a „prototype group management“ is set up

WP3: „FAIRness“

- White paper on Metadata in advanced draft; interactions with data providers to be integrated

TA 7: Education, training, outreach and citizen science

	Training	Education	Outreach	Citizen Science
For	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

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