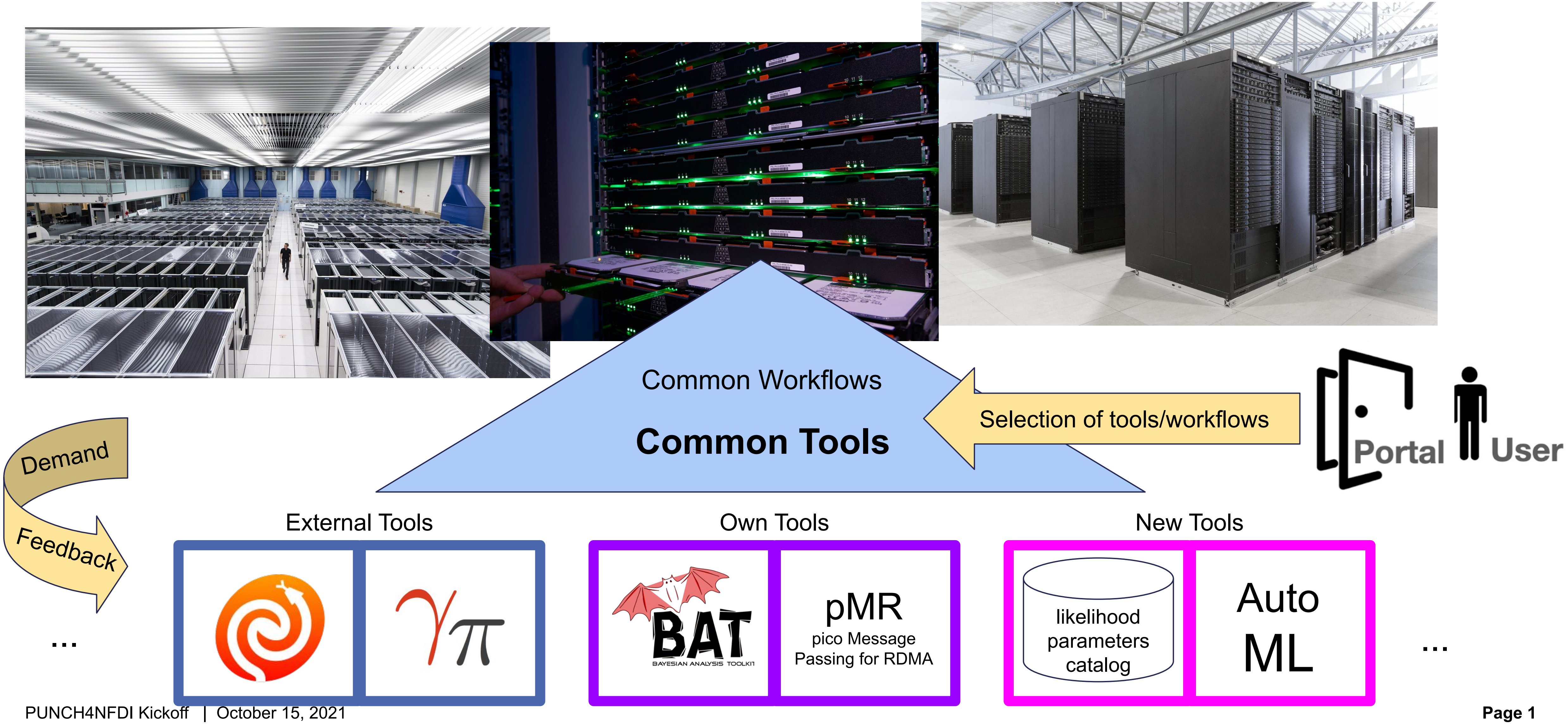
Task Area 3: Data Transformations

Provision of tools for parallel processing of huge datasets on heterogeneous computing resources



Common Workflows

Common Tools

Selection of tools/workflows

Portal

User

Demand

Feedback

…

External Tools

Own Tools

pMR
pico Message Passing for RDMA

New Tools

likelihood parameters catalog

Auto ML

…

TA 3: Common Tools

Commonalities among multiple science fields

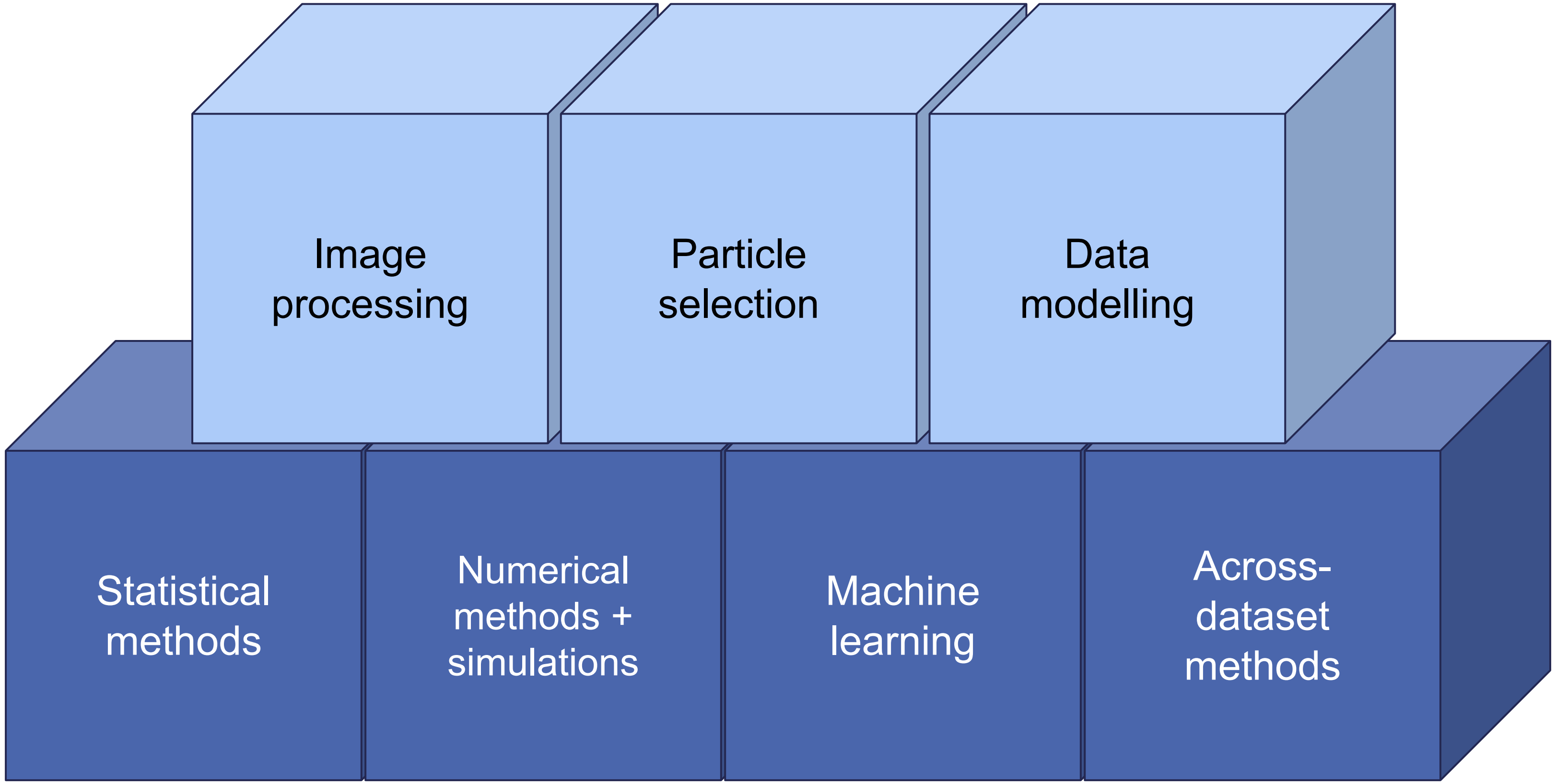


Image processing

Particle selection

Data modelling

Statistical methods

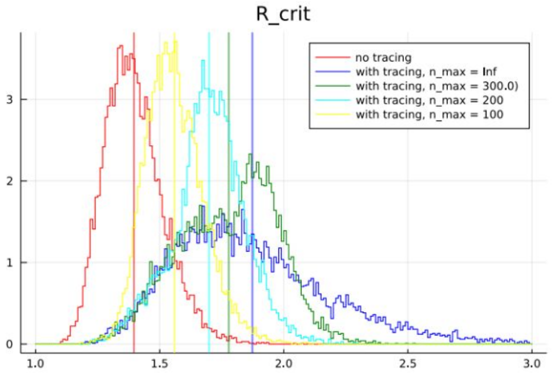
Numerical methods + simulations

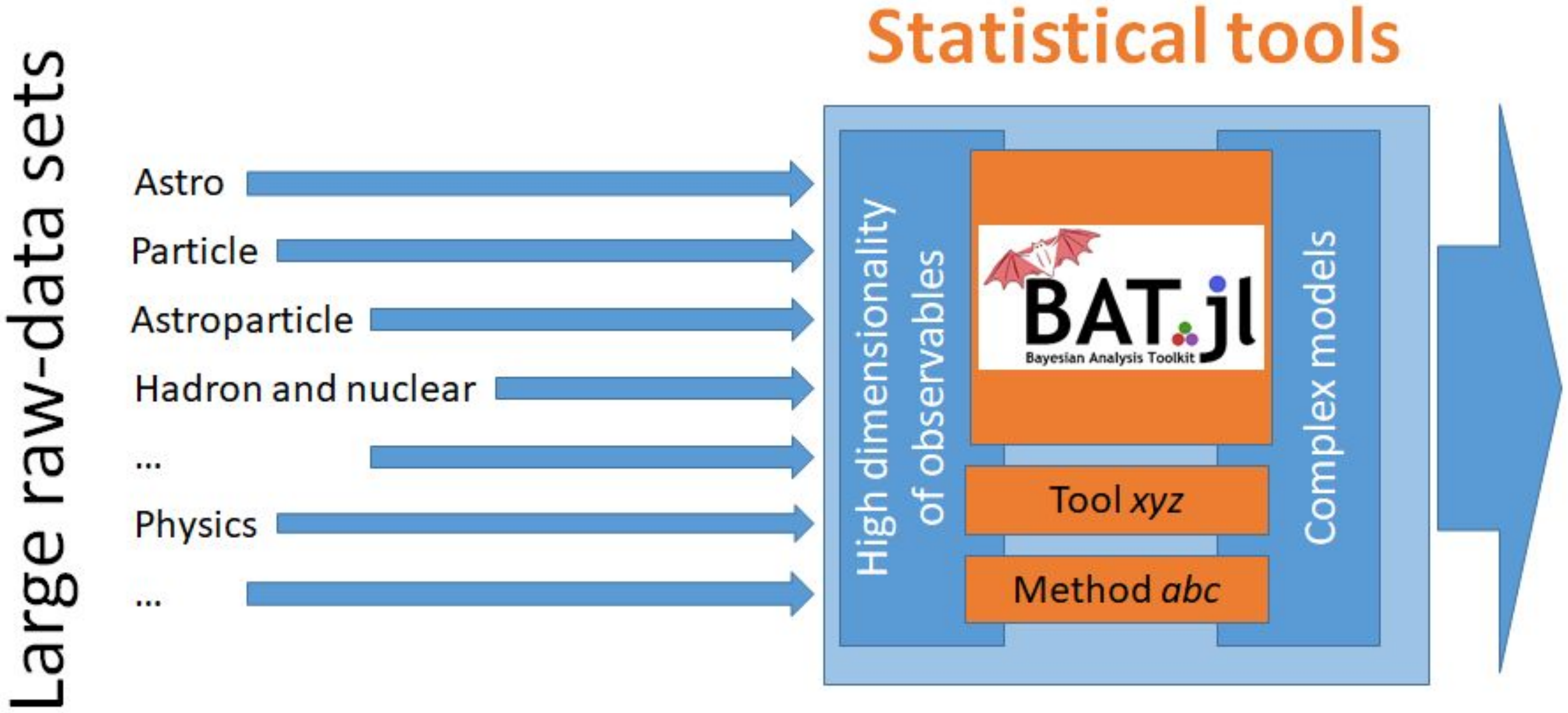
Machine learning

Across-dataset methods

TA 3 / WP1: Statistical methods

Methods and tools for analyzing large data sets and complex models





Interpretable and visualized results

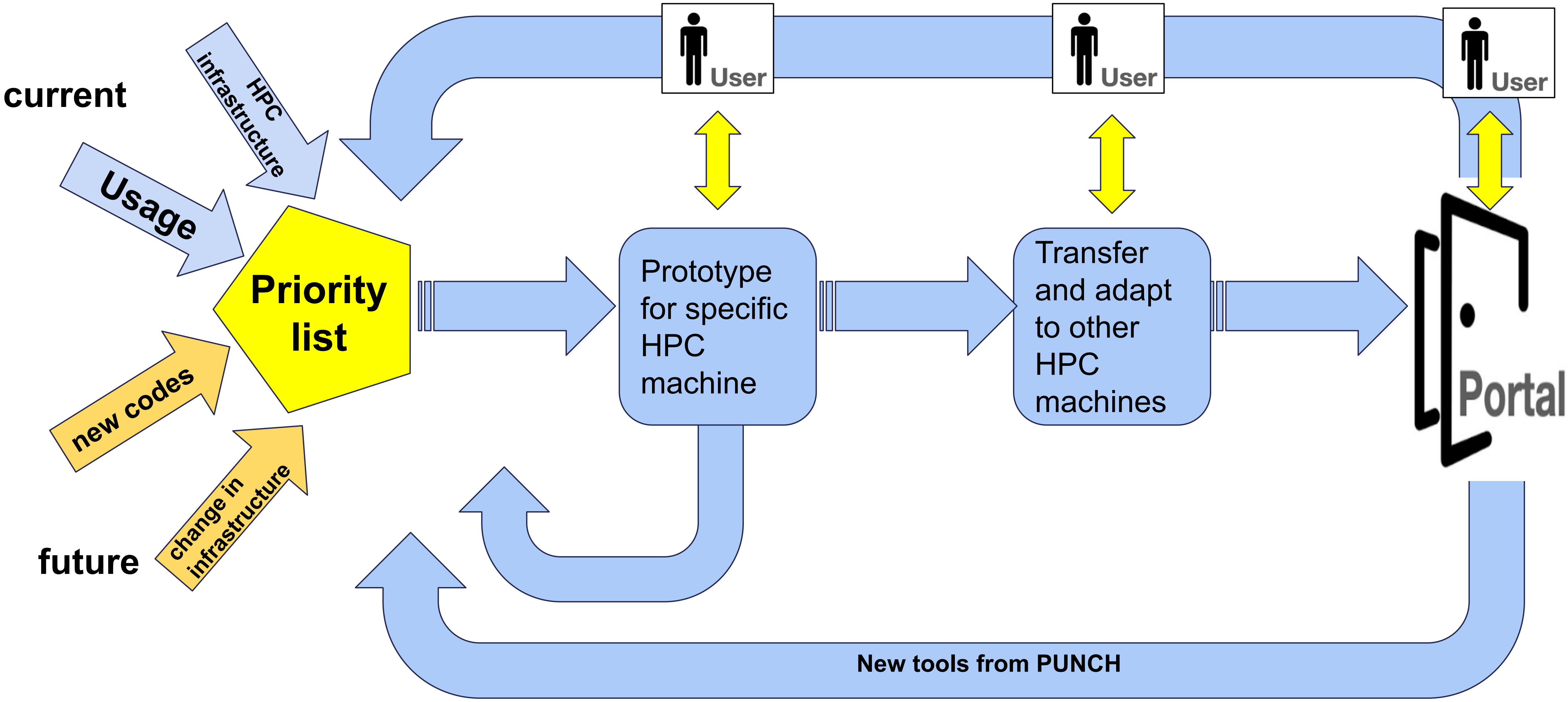
Deliverables:

– **D-TA3-WP1-1 (30 Sep 2026):** Statistical inference in the limit of large datasets and highly parallel computing.

– **D-TA3-WP1-2 (30 Sep 2026):** Integration of a broad set of statistical methods; further development of a subset of methods into a common set of cross-community tools.

TA 3 / WP2: Numerical Methods and Simulations

Provision of tools optimized for simulations on heterogeneous computing resources



TA 3 / WP3: Machine Learning

Automated tools for machine learning on large datasets

Machine learning is a transformative technology, showing and promising gains for many aspects of PUNCH science

WP3 will focus on two aspects:

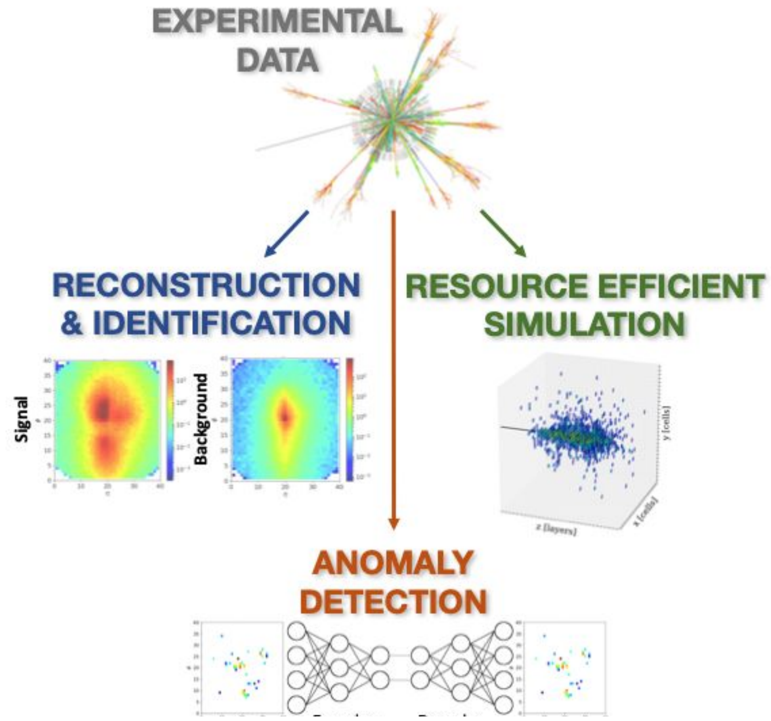
WP3.1: *Framework for AutoML on scientific data based on the PUNCH domain:*

- **Fully automate pre-processing and training workflows to develop machine learning as a service for scientific data:**
  - **Find and benchmark algorithms on datasets from different domains**
  - **Transfer learning based on successful architectures**
  - **Automated model selection and hyperparameter selection**
- Previous work: Physics Data for Machine Learning (pd4ml) https://github.com/erum-data-idt/pd4ml/, arXiv:2107.00656

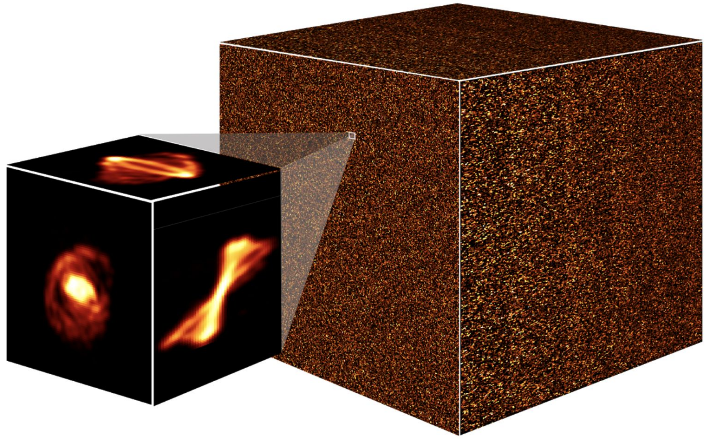
AutoML.org

WP3.2: *Tools and solutions for distributed learning using very large datasets:*

- **Scalable solutions for very large datasets:**
  - **efficient parallel training on partitions depending on distribution of data**
  - **combination (ensembling) of classifiers**



EXPERIMENTAL DATA
RECONSTRUCTION & IDENTIFICATION
RESOURCE EFFICIENT SIMULATION
Signal
Background
ANOMALY DETECTION



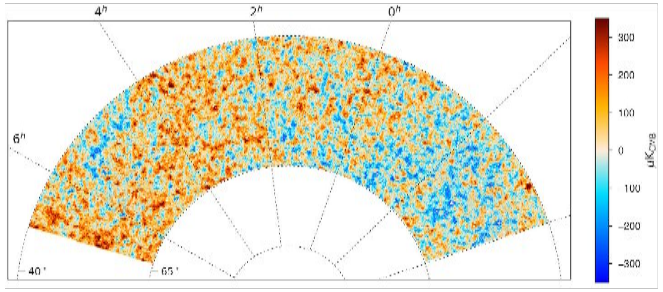
https://sdc2.astronomers.skatelescope.org/

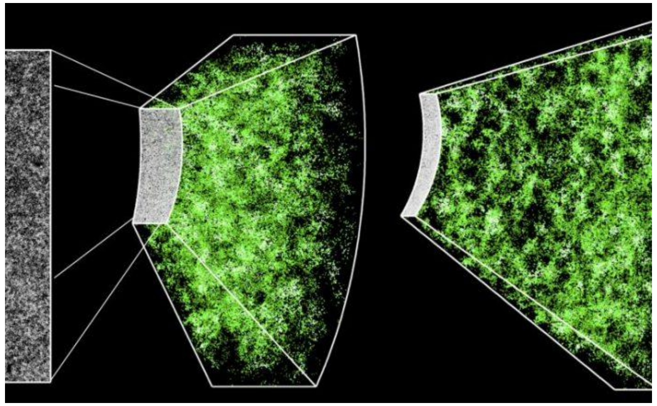
The simulated datacube, before noise and instrumental effects are added. Covering a sky area of 20 square degrees and featuring nearly a quarter of a million galaxies, the cube represents an SKA observation of neutral hydrogen — or "HI" — emission.

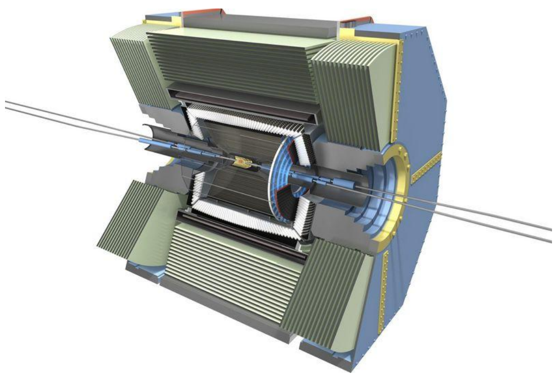
Illustration of ongoing machine learning developments

TA 3 / WP4: Analyses Across Datasets

Methods for exploiting the full potential of data from multiple sources





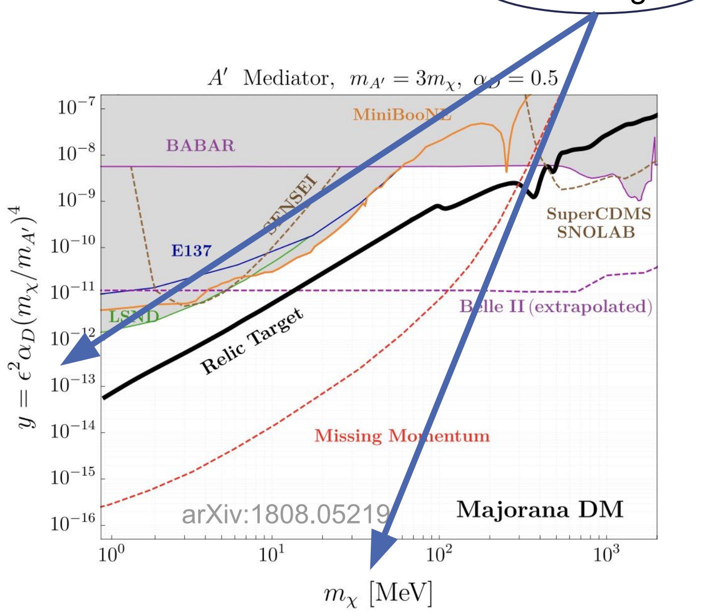


Enable joint analyses at pixel/event level

Transformation from low to high level data

Common definition of parameters

likelihood parameters catalog



Common analysis framework

- Data format converters
- Tools to support workflows executing across multiple archive nodes
- Library of workflow templates

Common likelihood interface to implementations as

- Function
- Histogram
- MCMC
- Calculated on demand from data

TA 3: Organization

TA3:

- *Monthly meetings on first Tuesday at 9:00*
- punch4nfdi-ta3@desy.de
- Contact: Thomas.Kuhr@lmu.de, mbrueggen@hs.uni-hamburg.de

WP1 (statistical methods):

- punch4nfdi-ta3-wp1@desy.de
- Contact: kevin.kroeninger@tu-dortmund.de, Joseph.Mohr@physik.lmu.de

WP2 (numerical methods and simulations):

- punch4nfdi-ta3-wp2@desy.de
- Contact: s.pfalzner@fz-juelich.de, tilo.wettig@ur.de

WP3 (machine learning):

- punch4nfdi-ta3-wp3@desy.de
- Contact: gregor.kasieczka@cern.ch, mbrueggen@hs.uni-hamburg.de

WP4 (methods for analyses across datasets):

- punch4nfdi-ta3-wp4@desy.de
- Contact: Joseph.Mohr@Physik.lmu.de, Thomas.Kuhr@lmu.de

Subscribe to lists at
https://lists.desy.de

or by sending mail with subject “subscribe listname FirstName LastName” to sympa@desy.de