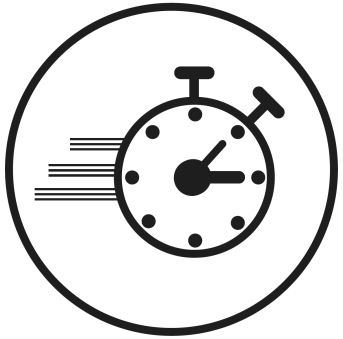
Task Area 5

Data Irreversibility



PUNCH Kickoff Workshop

15/10/2021

Prof. Dr. Michael Kramer

PD Dr. Andreas Redelbach

Motivation

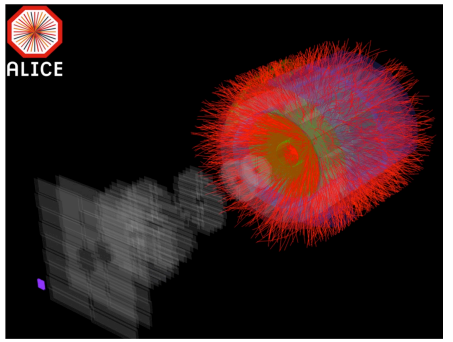
Rapid increase in both data rates and data complexity
→ **Challenges for many experiments** in the fields of Particle/Nuclear/Hadron/Astro(particle) Physics
→ Data rates of SKA will exceed the global internet traffic by a factor of a few

Decisions in real-time without human intervention, which information to keep and how to compress it.

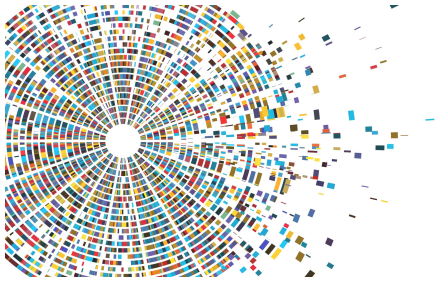
Loss will be inevitable and **mostly irreversible**

Taskarea 5 of PUNCH4NFDI:
- Focus on different aspects of data irreversibility
- Addressed in TA5-related work packages
- Develop methods to record, characterise, and quantify the degree of information loss

Challenges of reducing data streams in real-time:
Similar challenges for consortia well removed from physical sciences in near future
→ Developing tools and techniques applicable **in many data-intensive disciplines**

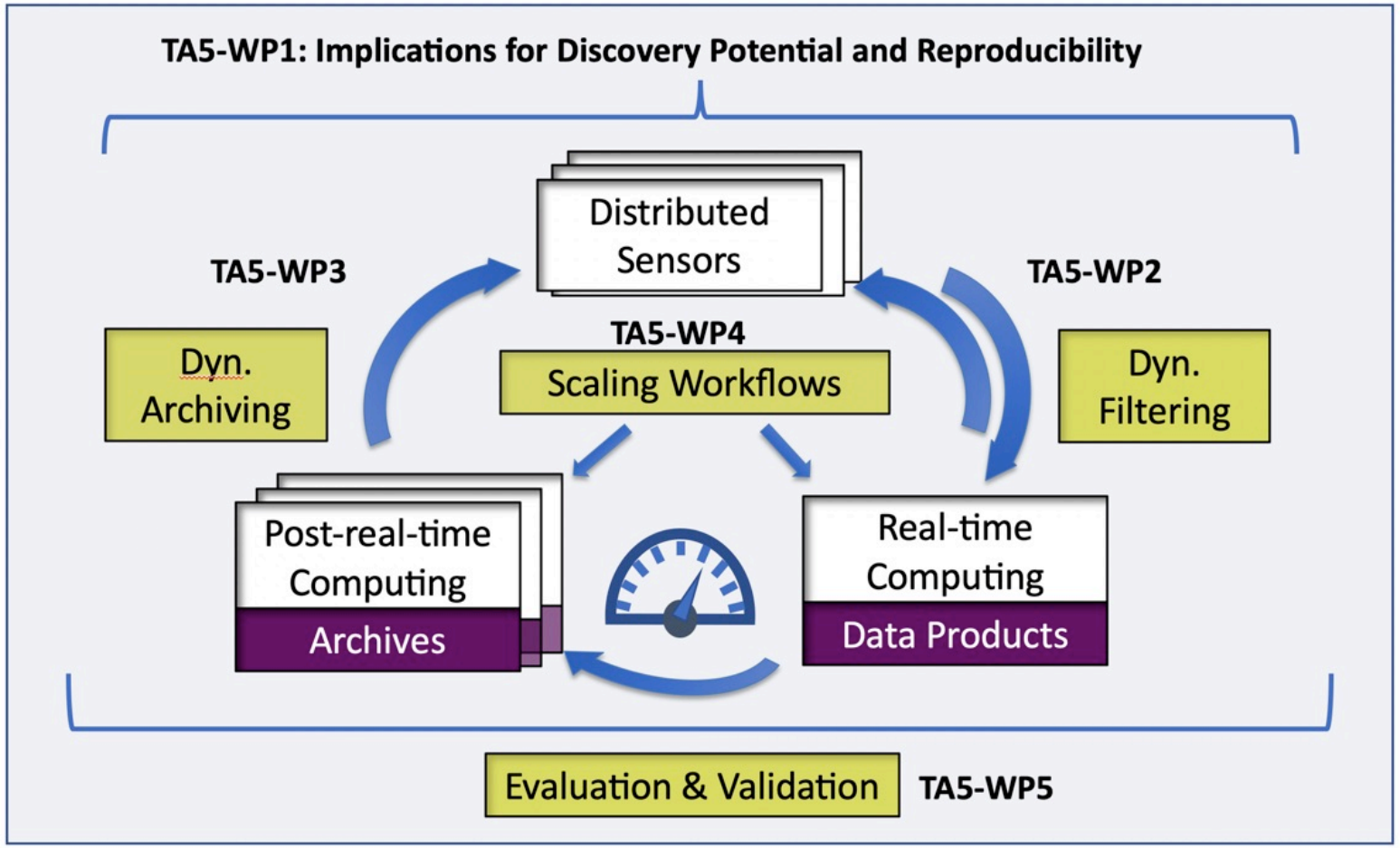




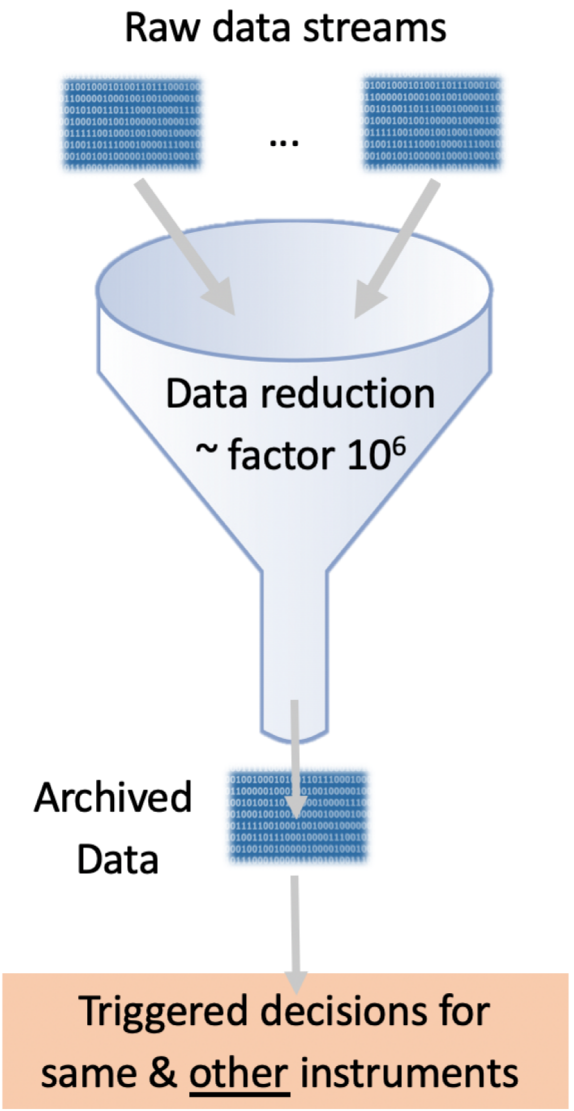


Overview

Institutes:
DESY
FIAS
FZJ
HTW
JGU
MPIfR
TUDO
TUDD
UBi
UHD



Dynamic Life Cycle Model



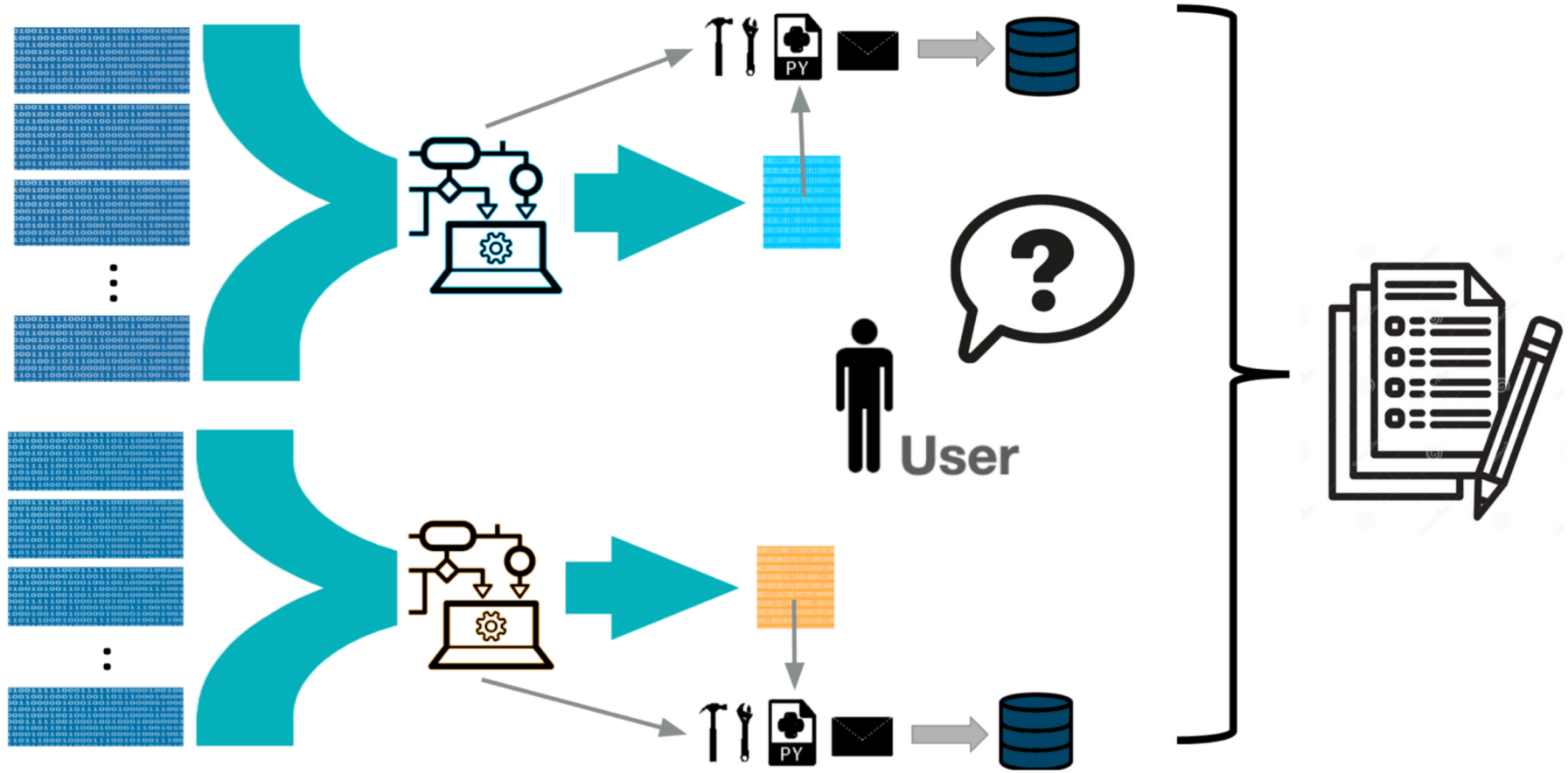
TA5 – WP1: Implications for discovery potential & reproducibility

Data from upcoming research facilities like HL-LHC and SKA will be massively filtered and refined. This process is irreversible and data are incomplete.

Explore tension and interplay between **reproducibility** of filtered and refined data and the **potential for unexpected discoveries.**

Identify and evaluate **methods and tools that can recover unexpected science (the unknown ”unknowns”) from incomplete data**

.



TA5 – WP1: Deliverables

D-TA5-WP1-1 (30 Sep 2023): Report on impact of on-line filtering on discovery potential

D-TA5-WP1-2 (30 Sep 2025): Report on impact of on-line filtering on FAIR principles

D-TA5-WP1-3 (30 Sep 2026): Concepts towards a general protocol on capturing the decisions made by, and status of, real-time sensors, as a basis for a future demonstator

Position @ UBi to be announced in 2022

WP1 leads:
Dominik Schwarz (UBi)
Stefan Wagner (UHD)

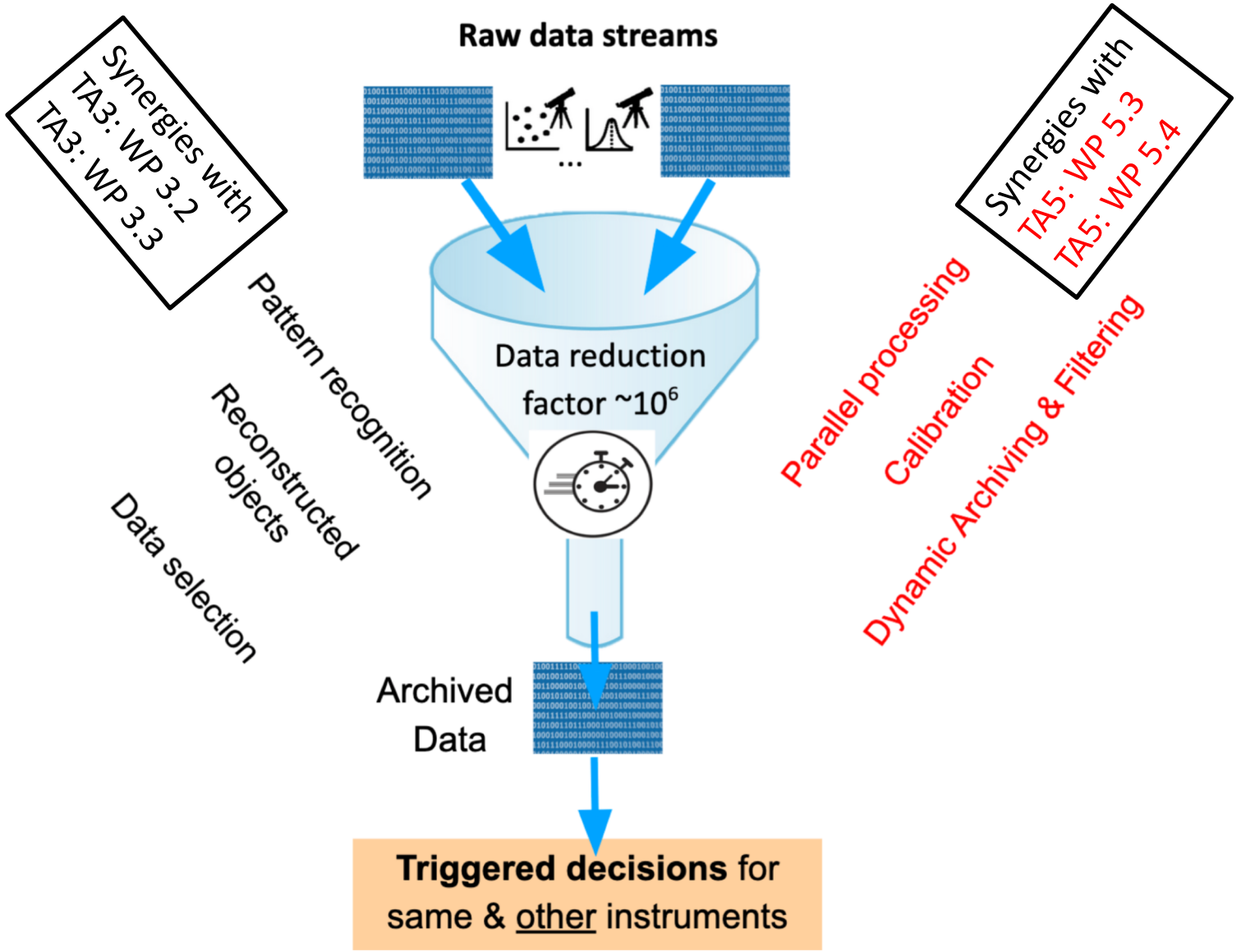
TA5 – WP2: Dynamic filtering

Addressing: Discarding irrelevant information with minimal time budget using filtering of various forms of noise and irrelevant backgrounds

Delivering: Solutions for real-time selection of (anomalous) signals, including a description of decisions in metadata

Looking at **diverse technology:**

CPU / GPU / FPGA



TA5 – WP2: Deliverables

D-TA5-WP2-1 (31 May 2022): Curation & metadata schemes for dynamic filtering.
D-TA5-WP2-2 (30 Sep 2022): Strategy concept for identifying highly complex (multi-parametric) signals in huge data streams.
D-TA5-WP2-3 (31 Dec 2023): Test environment for identifying highly complex (multi-parametric) signals in huge data streams using MeerKAT data.
D-TA5-WP2-4 (30 Sep 2024): Generic tool to convert trained neural networks into efficient HLS/VHDL FPGA firmware optimised for a real-time, low-latency environment.
D-TA5-WP2-5 (01 Mar 2026): Algorithms for massively parallel real-time sorting, clustering and pattern recognition on specialised hardware.
D-TA5-WP2-6 (01 Mar 2026): Algorithms and Machine Learning methods for filtering and selecting relevant transient/anomalous signals.
D-TA5-WP2-7 (30 Sep 2026): Pipeline for anomalous signal detection with low false-alarm probability for multi-messenger follow-up.

WP2 leads:
Johannes Albrecht (TUDO)
Michael Kramer(MPIfR)

"Run this metric of incompleteness"

WP1

"This is a filter which captures this kind of anomalous objects someone found in an archive"

WP2

"We stumbled on this new kind of object X. Did this happen before?"

WP2

TA5 – WP3: Dynamic archiving

Foundation for managing Data Irreversibility. Needs (e.g.):

- Methods by which a facility encodes what data it would be sensitive to (energy/frequency/messenger etc).
- **Common real-time protocols** through which facilities can record what observations were made, why they were executed, which candidates were detected/retained and a reference to the decision process used when saving/triggering on candidates.
- Database middle-layer which can view a set of static databases as one ordered real-time stream, following real-time protocols.

Once combined, these tools will enable the **creation of complex DB queries** that can both be used to **quantify data incompleteness** as well as be transformed into a **live dynamic filter**.

"Could facility F have seen object Y at time T?"

TA4

"What would this kind of filter do?"

WP2

"Why did we not trigger on Y?"

TA4

WP4

"Can you execute this workflow/ML model on the archive?"

TA2

"Here is how to access a set of distributed, traditional archives."

TA2

"Archive this data in an efficient manner for us."

WP5

"These are simulated events. Stick them into an archive and make it look real."

TA5 – WP3: Deliverables

D-TA5-WP3-1 (31 Mar 2023): Specifying the concept of a dynamic archive: Requirements in relationship to other WPs (information loss, dynamic filters, scalable workflows and simulated catalogs) as well as to information present in traditional archives (other TAs).

D-TA5-WP3-2 (31 Mar 2025): Present a framework in which queries to dynamic archives can be transformed into a dynamic filter (as used by some combination of sensors), and vice versa.

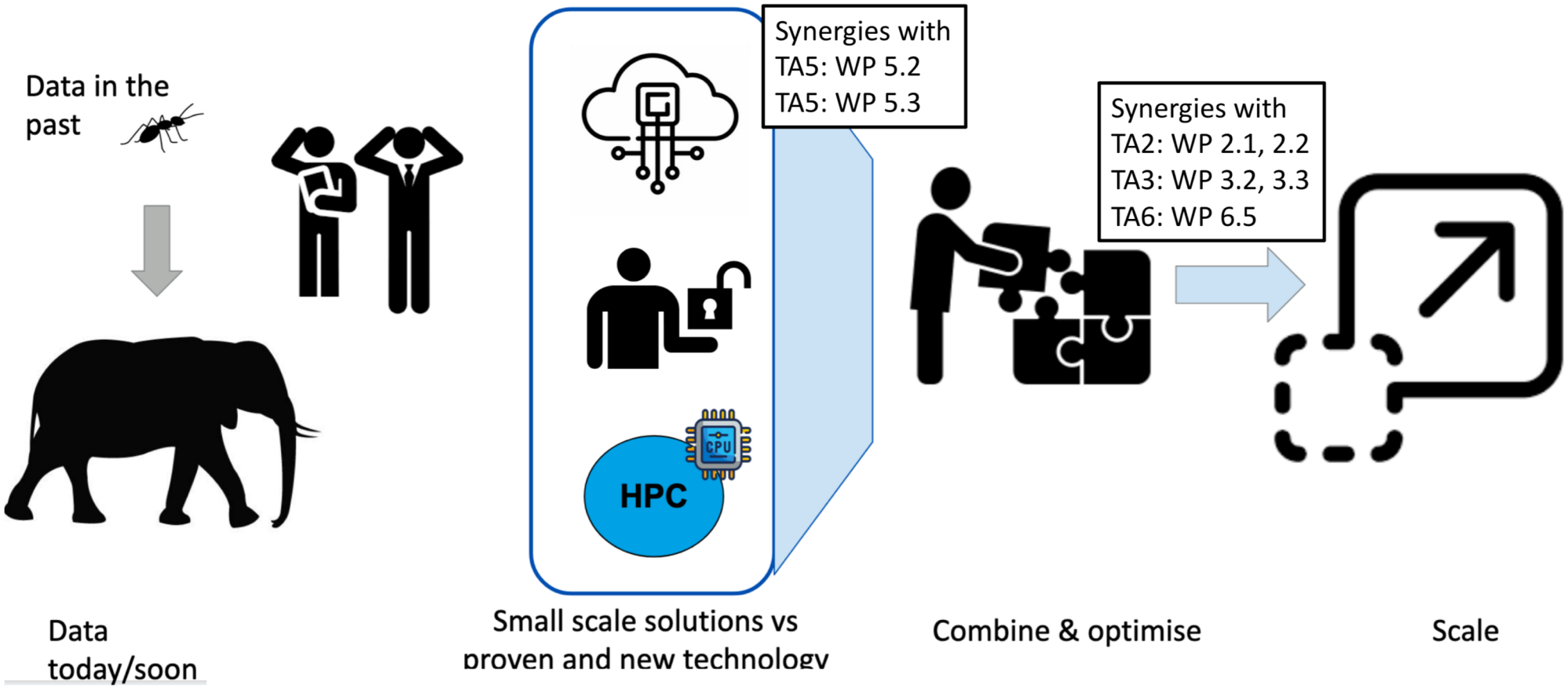
D-TA5-WP3-3 (30 Sep 2026): Present methods by which queries to dynamic archives also return an estimate on the potential of information loss, i.e. how well the archive response can be assumed to approximate the response of a real-time sensor.

WP3 leads:
Jakob Nordin (DESY)
Laura Spitler (MPIfR)

TA5 – WP4: Scaling Workflows

Addressing: optimal use of hardware and software resources particularly when analysing single huge data sets

Delivering: report on technology solutions for scaling the „online“ and „offline“ workflows



TA5 – WP4: Deliverables

- **D-TA5-WP4-1** (31 Dec 2024): Porting common off-line packages (e.g. CASA) to a memory-based computing prototype to prepare analysis of “data monster”
- **D-TA5-WP4-2** (31 Mar 2025): Standard software (e.g. CASA) compatible with Gen–Z.
- **D-TA5-WP4-3** (30 Jun 2025): Caching strategies for processing a set of benchmark files with the evaluated efficiencies and latencies.
- **D-TA5-WP4-4** (30 Jun 2026): Definition and initial implementation of an efficient real-time data processing framework
- **D-TA5-WP4-5** (30 Sep 2026): Scaled feedback interfaces between off-line software (e.g. CASA) and selected real-time processes using MeerKAT data

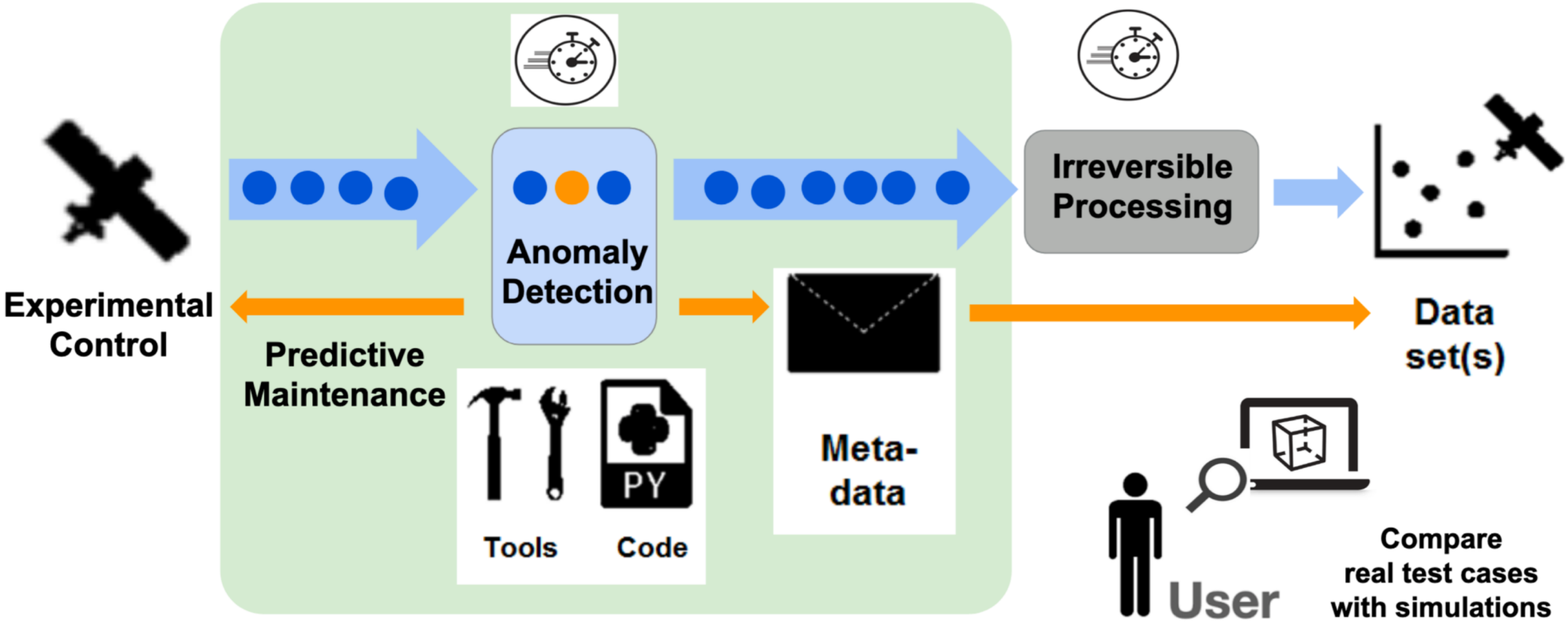
WP4 leads:
Hermann Heßling (HTW)
Andreas Redelbach (FIAS)

Position @ HTW to be announced in 2022

TA5-WP5: Evaluation & validation of instrument response & characteristics

Addressing: data quality assurance for systems with irreversible data processing in presence of variable, unpredictable background

Delivering: anomaly detection and predictive maintenance for real-time systems exploiting machine learning solutions with on-line feedback to archive metadata and study impact on noise



Primary **synergies** with TA3 (WP 3.1 and 3.3) and TA5 (WP 5.1, 5.2, 5.3 and 5.4)

TA5 – WP5: Deliverables

D-TA5-WP5-1 (30 Sep 2024): Development of machine learning prototypes for anomaly detection and predictive maintenance

D-TA5-WP5-2 (30 Sep 2024): Interference recognition and mitigation schemes for transient discovery leading to a robust triggering system

D-TA5-WP5-3 (30 Sep2026): Expansion of the concept to a generalized toolkit for predictive maintenance and anomaly detection

D-TA5-WP5-4 (30 Sep 2026): Evaluation of the machine learning approaches by analyzing false-alarm rates and online feedback

WP5 leads:
Bernhard Spaan (TUDO)
Laura Spitler (MPIfR)

TA5 status and outlook:

Re-worked work program and deliverables

Internal documentation:
https://gitlab-p4n.aip.de/punch/intra-docs-content/-/blob/master/docs/TA5/overview.md

Establishing regular meetings at the level of TA and WPs
→ Get in contact with leads of TA5 / WPs

Use cases of general interest (see also PUNCHLunch **Data Irreversibility** on 25/03/2021):

- Online event selection of signatures of displaced vertices
- Anomaly-based triggers for efficient online selection at future collider experiments
- Comparing simulations with (lossy) data

→ Concepts for Green IT by efficient and dynamic selection of data

→ Blueprint for other fields of data intensive science

BACKUP

Use Cases

5.1 HEP/HuK: Real-time processing and trigger for the High-Luminosity era of the LHC Experiments

5.2 Astro/HuK: Comparing simulations with (lossy) data

5.3 Astro: Radio Transient Discovery and Identification

5.4 HEP/HuK: Online event selection of signatures of displaced vertices

5.5 HEP/HuK/Astro: Identifying anomalous signals and the hunt for dark matter

5.6 Astro: High-cadence pulsar observations with individual LOFAR stations

5.7 Astro: S-Band Survey with the SKA-MPG telescope

5.8 Astro: Joint processing of IceCube, LIGO/VIRGO, CTA, Roman Observatory and Ultrasat real-time data

5.9 Astro: Data Monster