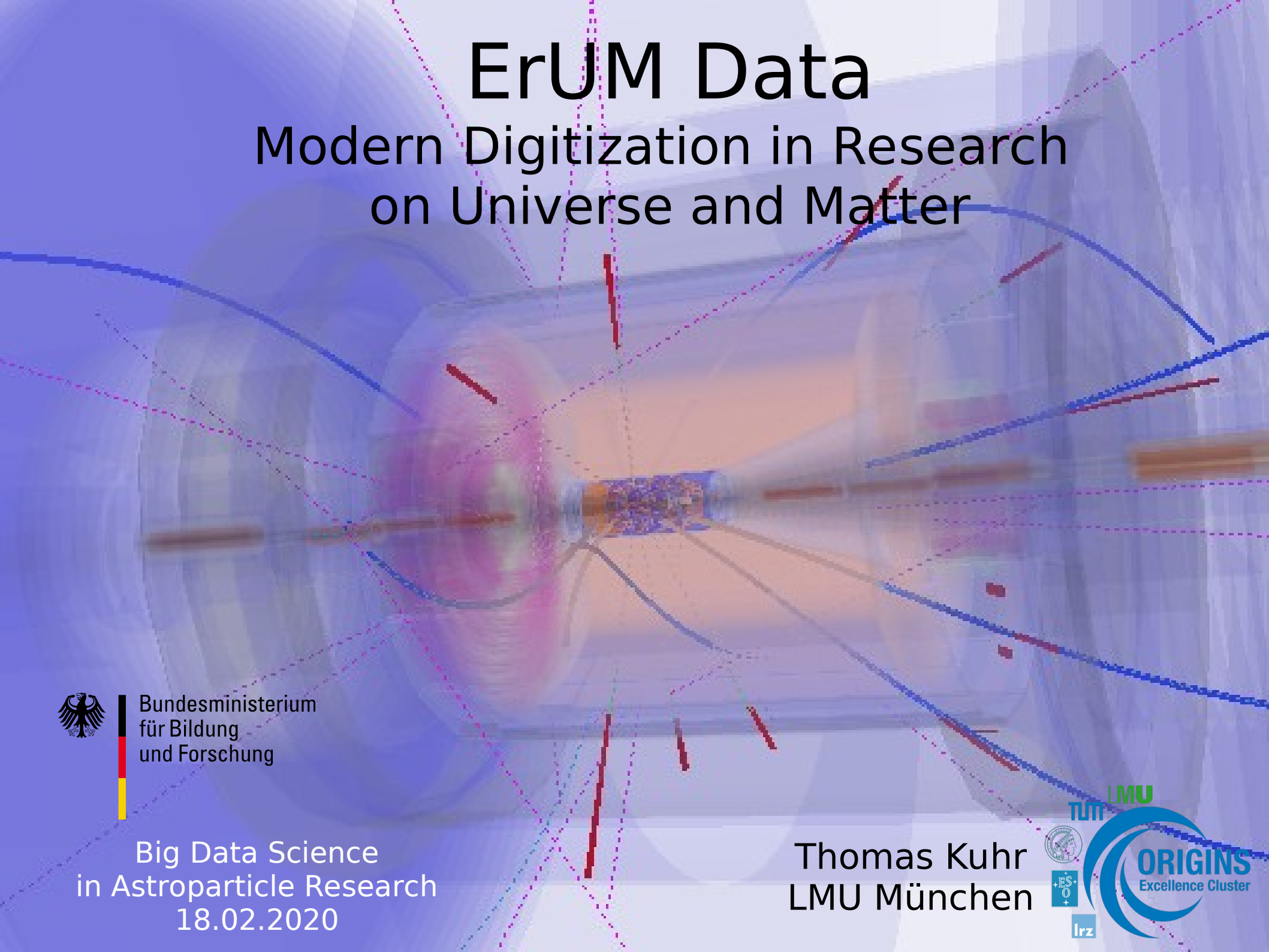
ErUM Data

Modern Digitization in Research on Universe and Matter



Bundesministerium
für Bildung
und Forschung

Big Data Science
in Astroparticle Research
18.02.2020

Thomas Kuhr
LMU München

LMU
TUM
ORIGINS
Excellence Cluster
ESO
lrz

Digital Agenda

Digitalisierung ist Chefsache

WIR GESTALTEN DIE DIGITALISIERUNG

Digital Agenda:

1. Digitalen Wandel in der Wissenschaft forcieren

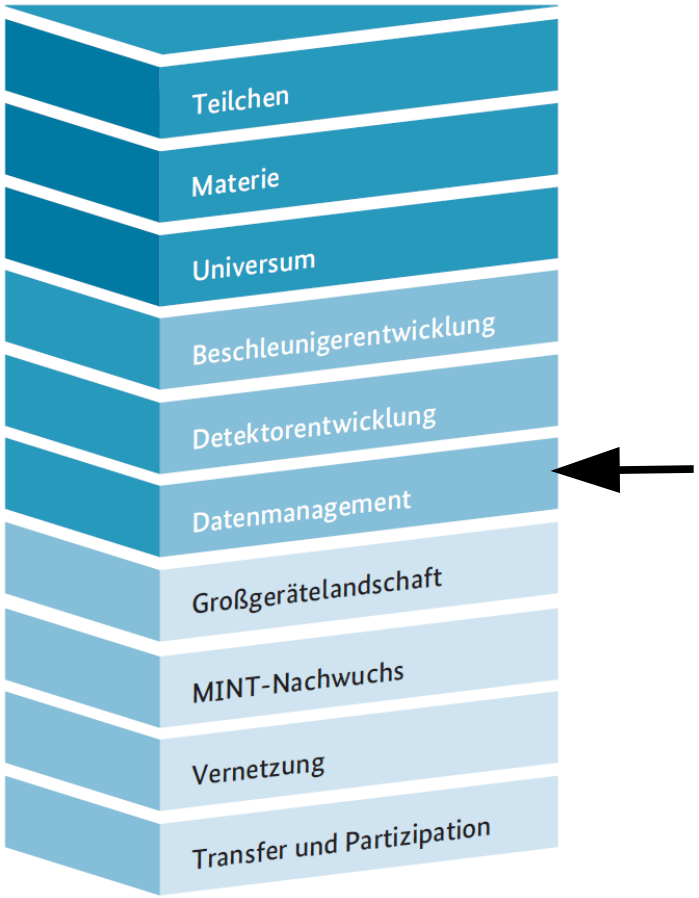
Um eine breite, disziplin- und organisationsübergreifende Zugänglichkeit und Nutzbarkeit von digitalen Informationen sicher zu stellen, werden die wissenschaftlichen Informationsinfrastrukturen gestärkt, ausgebaut und besser vernetzt.

1. Accelerate Digital Transformation in Science

To secure a broad, discipline and organization overarching access to and availability of digital information, the scientific information infrastructures are strengthened, extended, and better connected.

Erforschung von Universum und Materie – ErUM

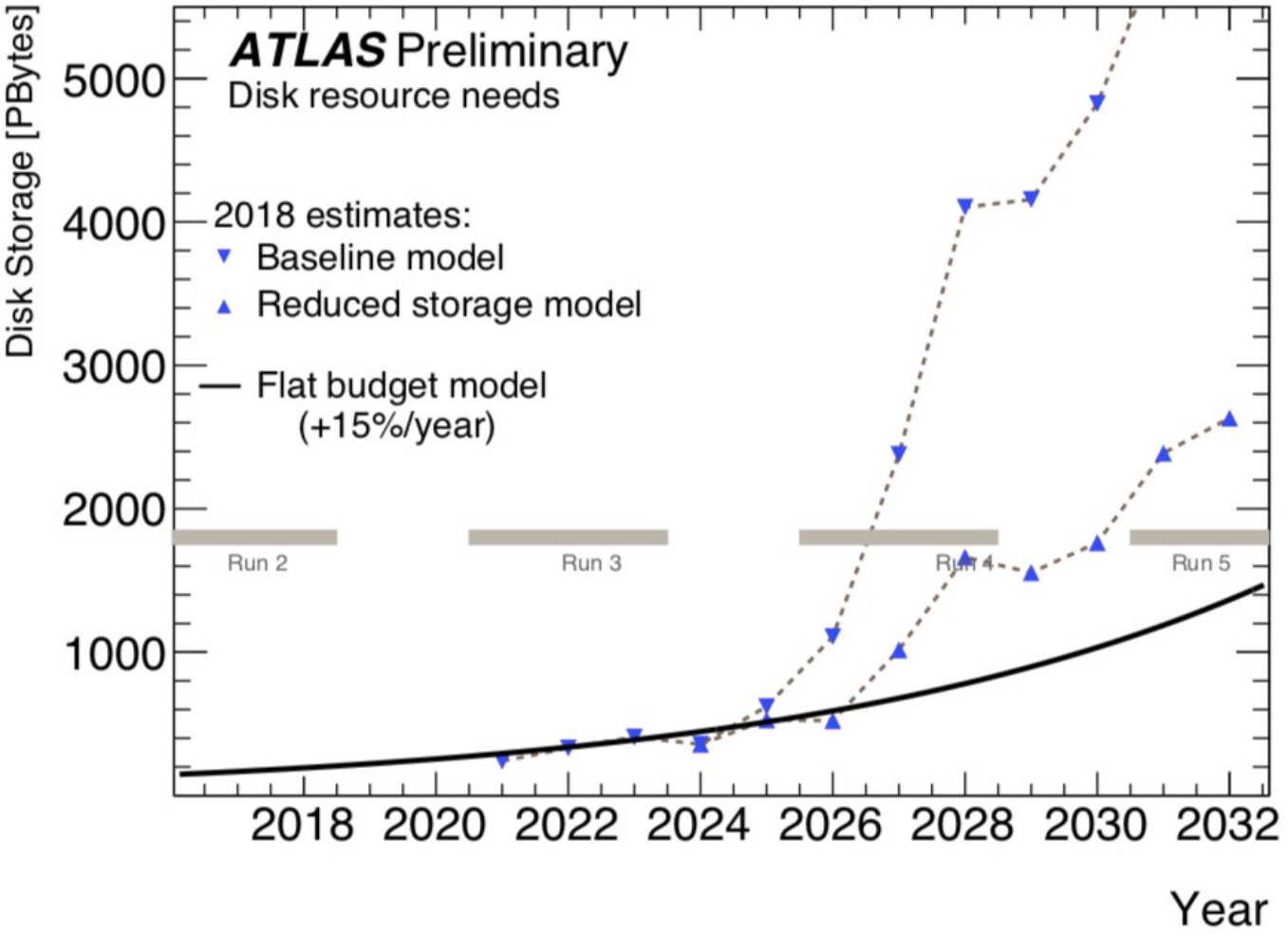
Rahmenprogramm des Bundesministeriums für Bildung und Forschung



Natural science fundamental research is a central area of application of new digital methods and techniques. It is a significant driver for further developments. Increasing computational effort and complex data management is addressed by site overarching work techniques and the elimination of technological bottlenecks. Open access and long term data management must continue to take into account the requirements and specifics of the different research infrastructures. Young scientists acquire a unique expertise in data management. New services and holistic solutions can arise in future based on the know-how in fundamental research.

Challenges: Data Volumes/Rates

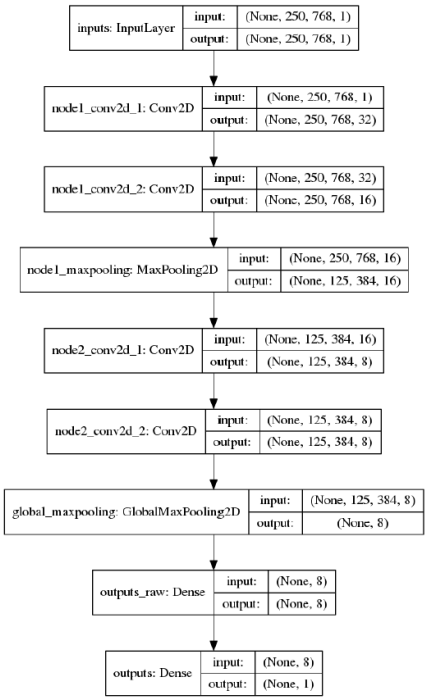
- HL-LHC resource estimates factors above flat budget scenario
- ALICE & LHCb: Triggerless readout in Run 3
- Belle II: 50x more data than Belle
- FAIR: 30 PB per year, 300.000 CPU cores
- CTA: several 10 PB per year

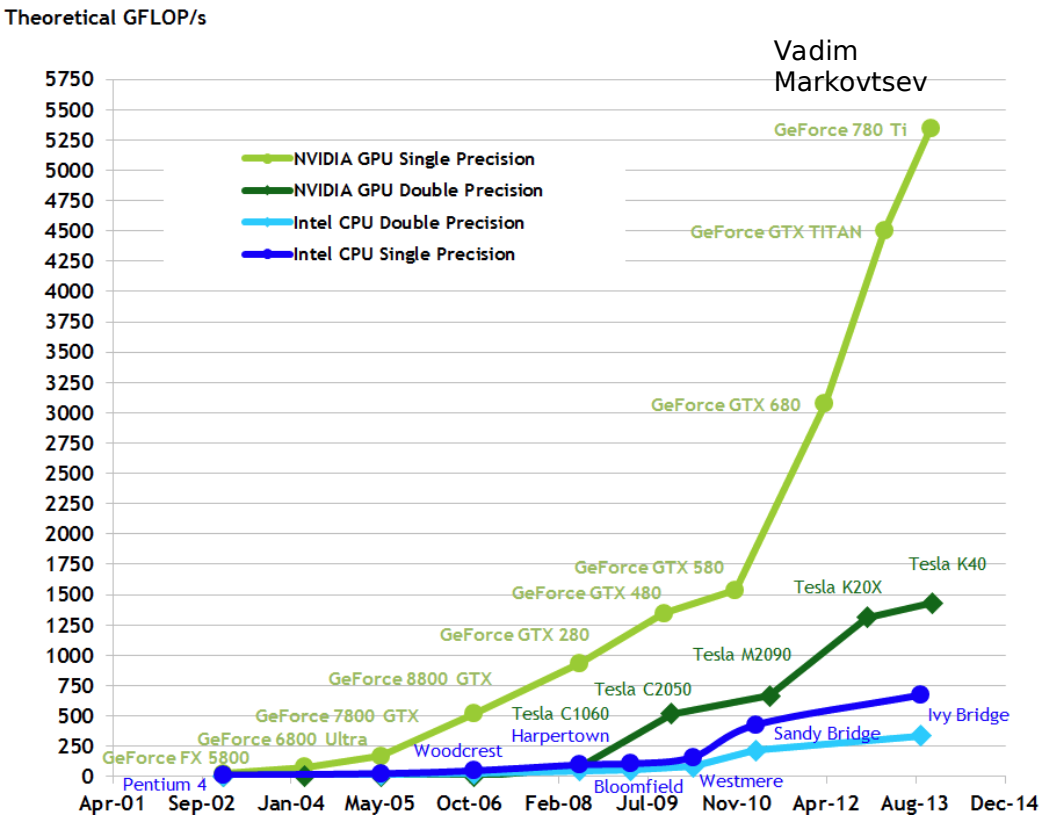


Challenges: Technological Evolution

And opportunities:

- Multicore machines
- GPUs
- SSDs
- Virtualization
- Machine Learning
- Artificial Intelligence
- Quantum computing
- ...

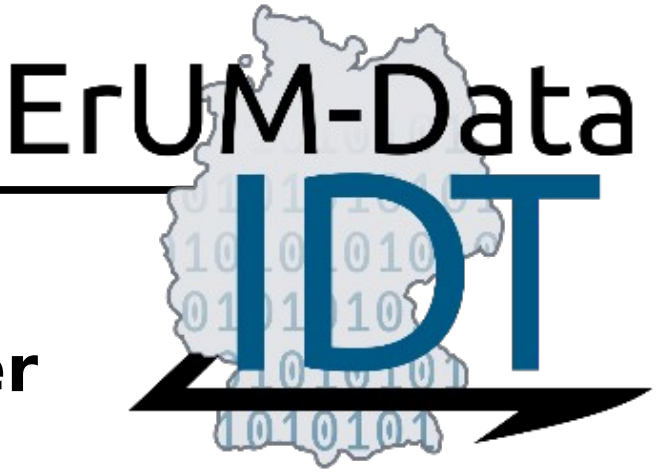




Approach:

- ➔ Find common solutions

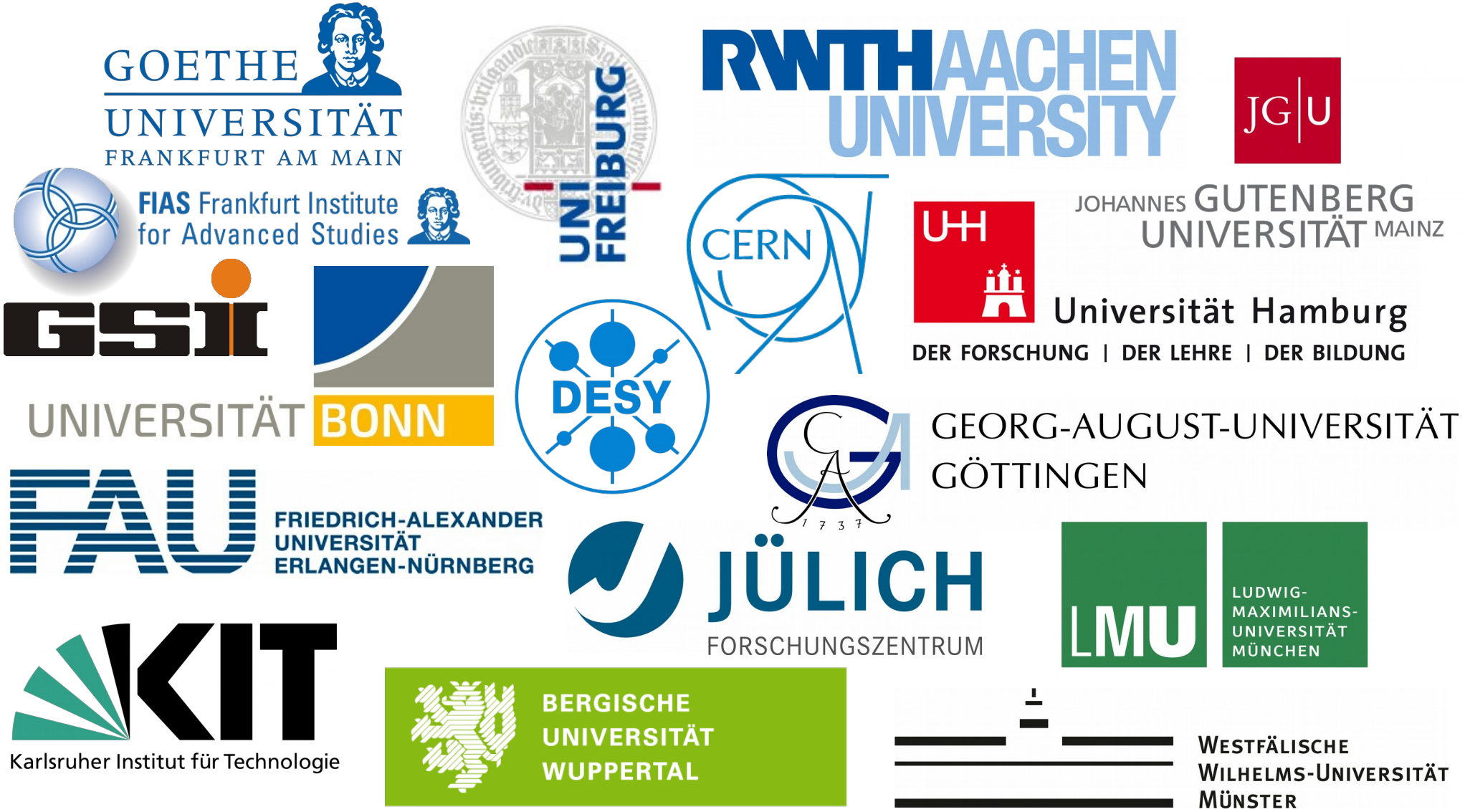
IDT-UM Project



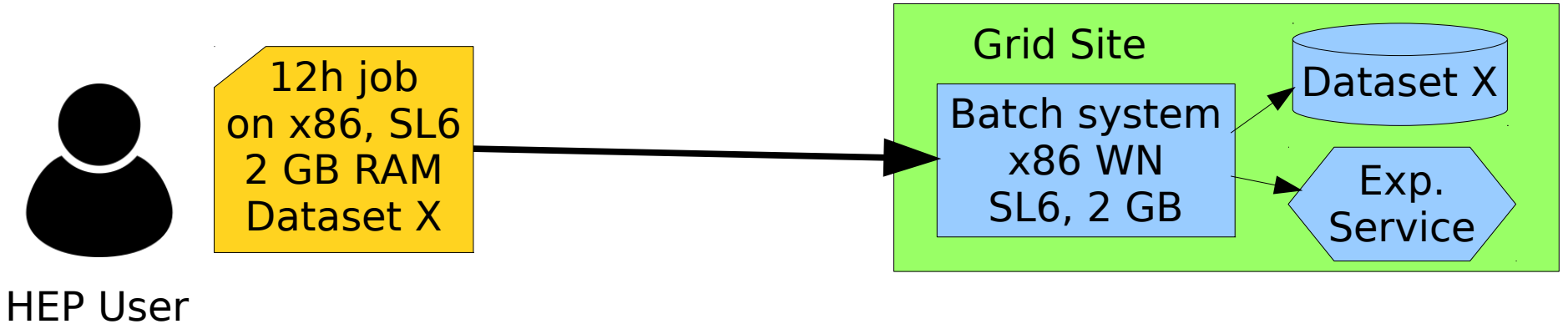
- **Innovative Digital Technologies for Research on Universe and Matter**
- Application of partners from
  - Particle Physics (ATLAS, Belle II, CMS)
  - Hadron and Nuclear Physics (ALICE, CBM, PANDA)
  - Astroparticle Physics (Auger, CTA, IceCube)
- → to develop experiment overarching solutions

- Evaluated by panel including computer scientists
- → Got 3.6 M€ for 3 years, started October 2018

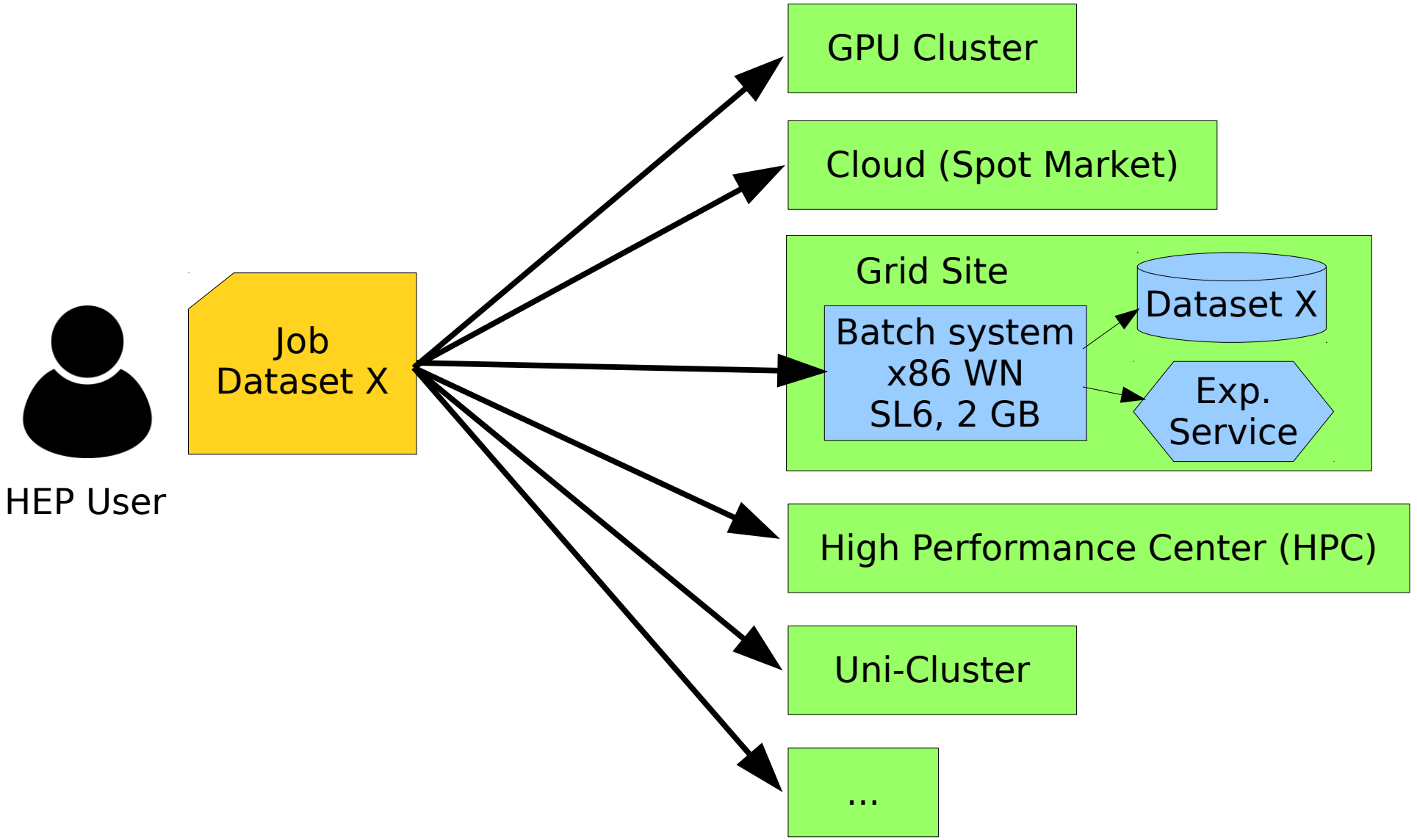
Project Partners



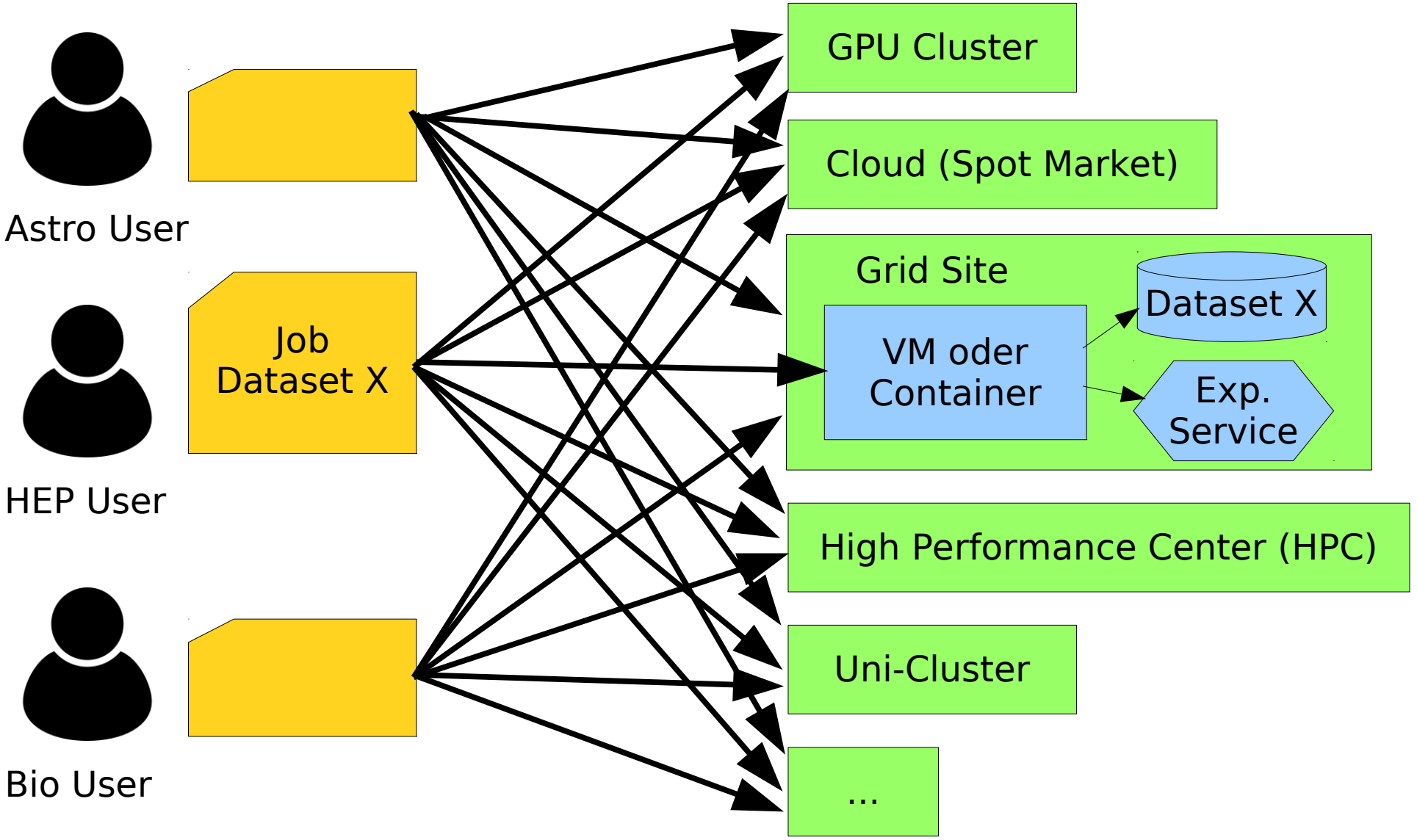
Scientific Computing Today



Scientific Computing Tomorrow



Scientific Computing Vision



Subject Area A

Coordination:
Manuel Giffels
Kilian Schwarz

- Developments for the provision of technologies for the use of heterogeneous computing resources

A1: Tools for integration • Scheduling of cloud jobs • Container technologies • Database access	**A2: Efficient Use** • Transient data caches • Transparent access to distributed data
A3: Workflow Control • Optimization with data mining	

Subject Area B

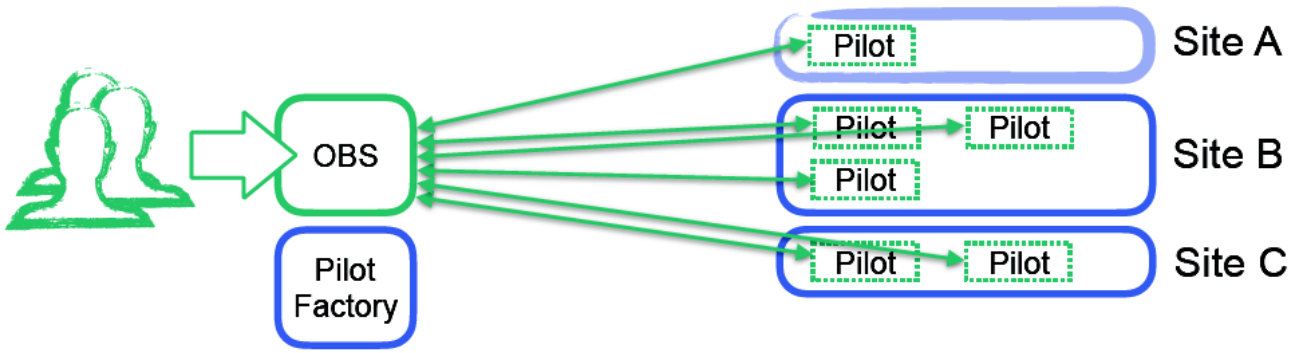
Coordination:
Christian Zeitnitz

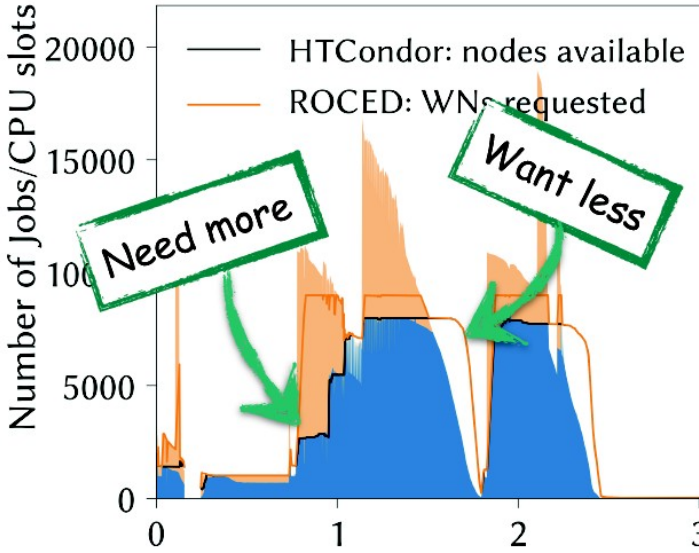
- Application and test of virtualized software components in the environment of heterogeneous computing resources

B1: Tests of Components Implementation and test on different platforms • Storage and caching solutions • Virtualized services (databases, monitoring, accounting)	**B2: Job and Resource Management** Job distribution and monitoring in a heterogeneous computing resource environment using container technologies
B3: Virtualization of User Jobs • Requirement capture • Determination and creation of run time environment • Creation of container and meta data	**B4: Combined Tests** Test of complete system on different platforms regarding • Installation and maintenance • Performance • Scalability • Robustness

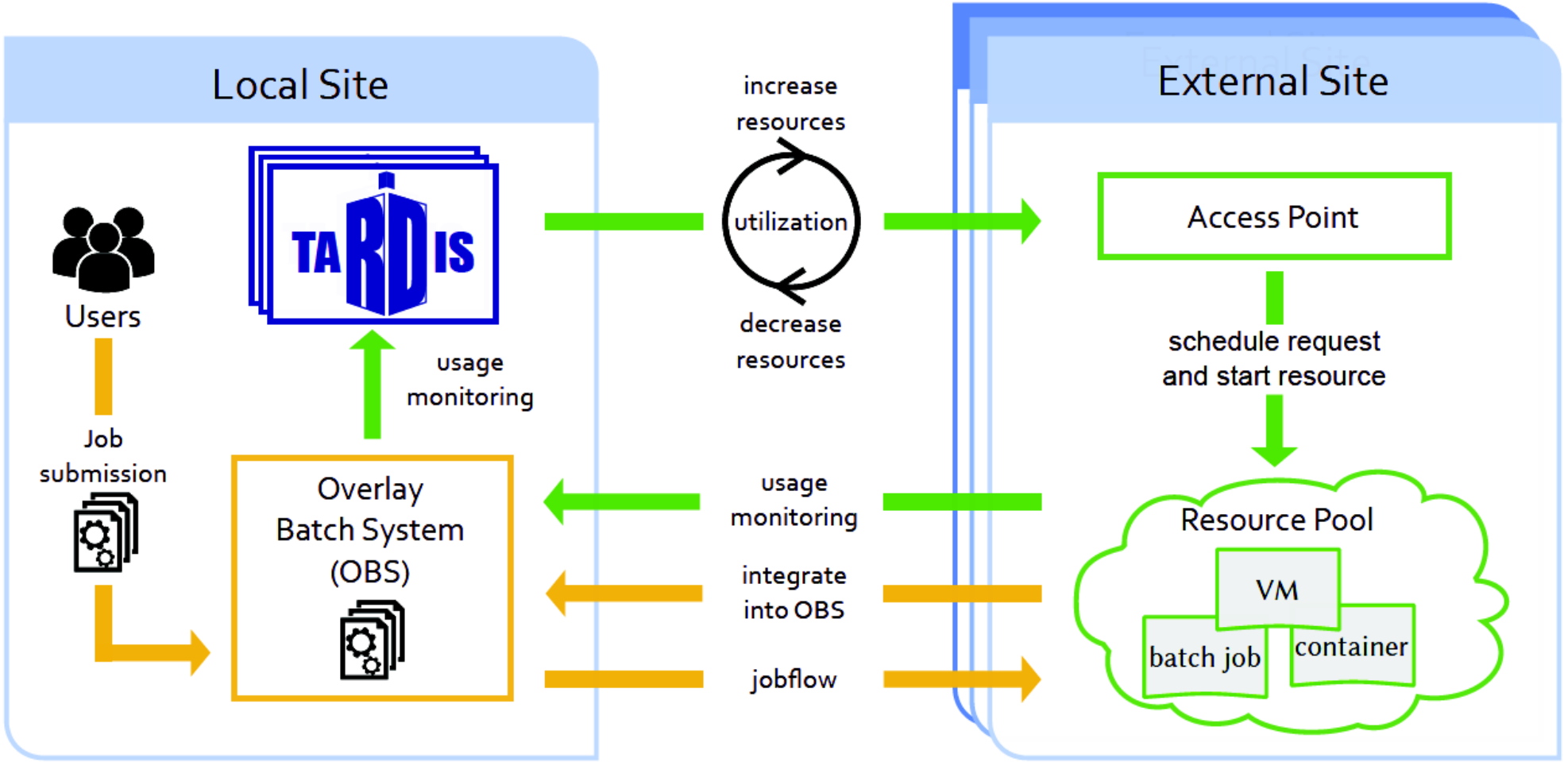
Example of a Common Solution

- COBalD (COBalD Opportunistic Balancing Daemon)
  - Overlay Batch System (OBS)
  - Pilot (resources) → Drone (resources and environment)
  - Adjustment of allocated resources to demands
- TARDIS (Transparent Adaptive Resource Dynamic Integration System)
  - Adapters for OpenStack, CloudStack, Moab, Slurm, HTCondor





Example of a Common Solution



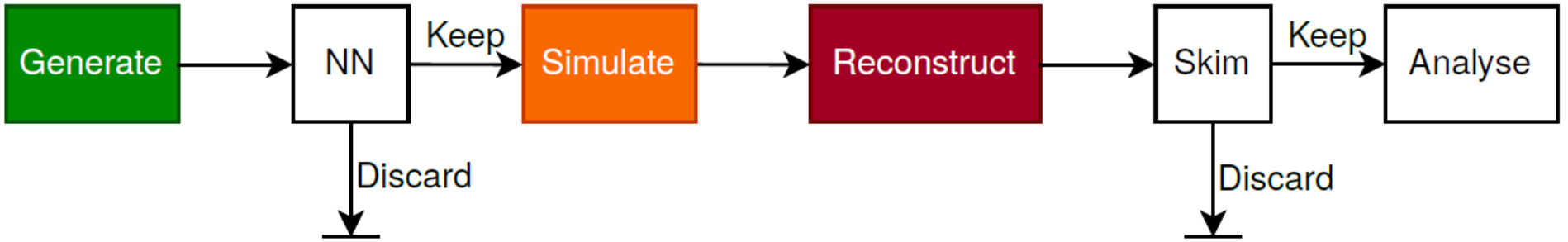
Subject Area C

Coordination:
Gregor Kasieczka

- Deep Learning, Gain of knowledge by substantiated data-driven methods

C1: Processing of Sensor Data • Signal filter, noise suppression • Processing of time dependent data	**C2: Object Reconstruction** • Track and cluster reconstruction, jet forming, event reconstruction • Questions of placement, order, assignment of data • Extraction of small signals in case of large backgrounds
C3: Network Accelerated Simulations • Generative adverserial networks, adjustment of simulation to data • Methods for the evaluation of the quality of network simulations	**C4: Quality of Network Predictions** • Reduction of experimental systematic uncertainties • Special learning strategies • Prediction relevant information • Uncertainty of predictions

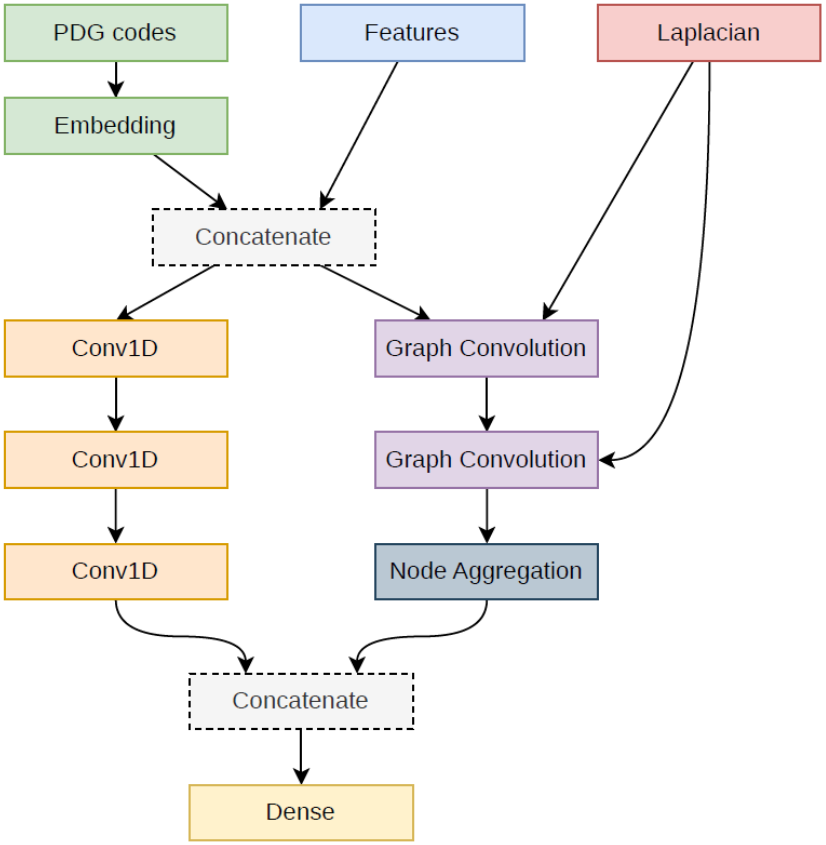
Example of a Common Solution



- Selection of (background) events on generator level to save simulation time

- Collection of problems and solution being worked on

VISPA



Subject Area D

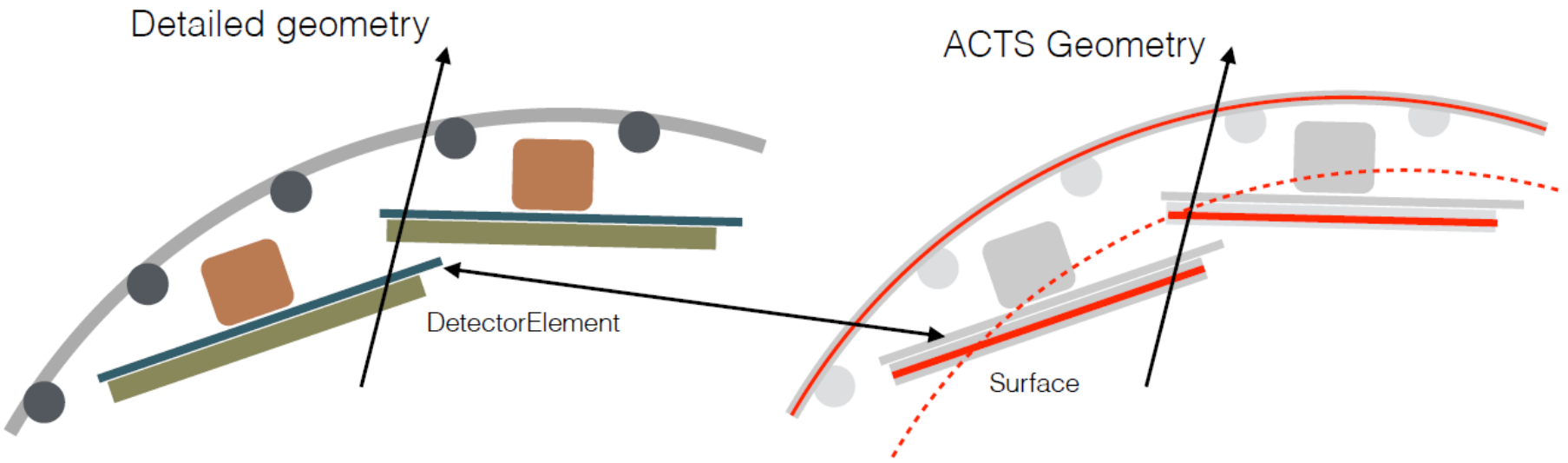
Coordination:
Florian Bernlochner

- Event reconstruction: Cost- and energy-efficient use of computing resources

D1: Track Finding	D2: Parameter Determination
• Alternative algorithms, e.g. cellular automata • Alternative architectures, e.g. GPUs	• Connection of GenFit2-ACTS

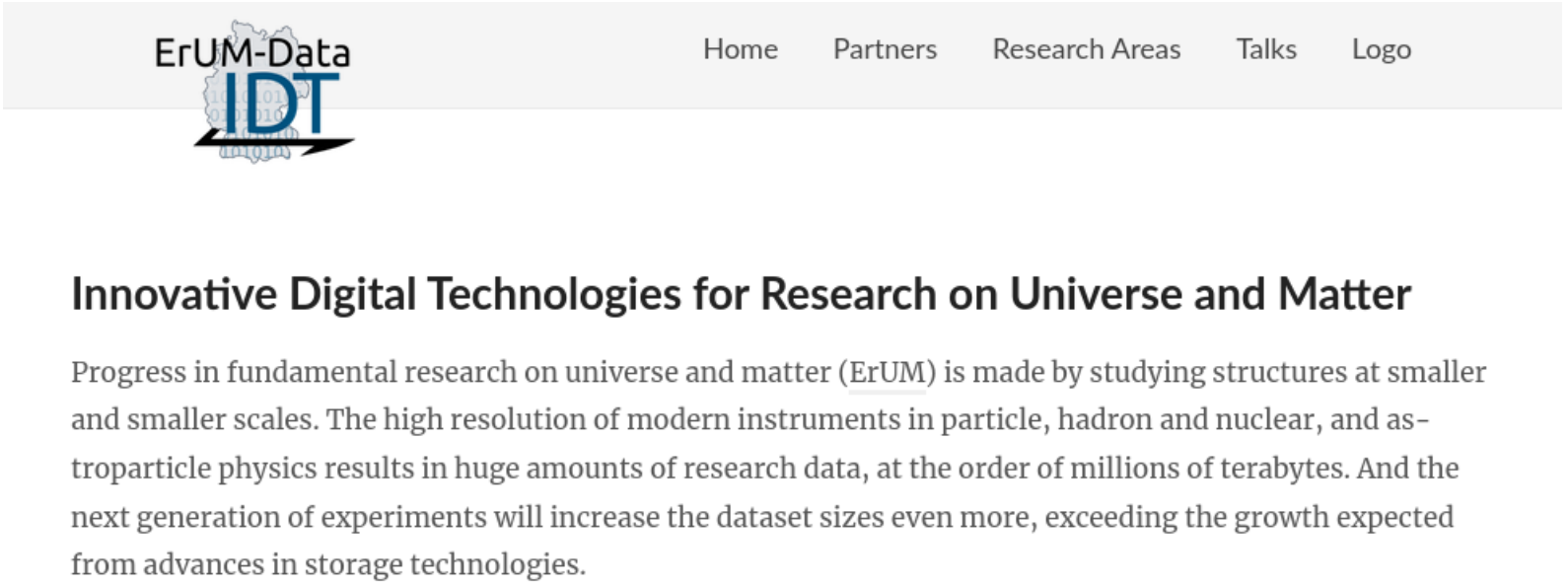
Example of a Common Solution

- A Common Tracking Software
- ACTS core, framework, simulation, data



IDT-UM Further Information

- Web page: https://www.erum-data-idt.de/
- Mailing list: computing-verbund@lists.lrz.de
- erum-data-idt organization on github
- Next collaboration meeting on Thursday, April 2nd at 15:00 in Bonn (during DPG conference)



Computing Strategy Workshop

- Purpose: Agree on a common computing infrastructure strategy to address in particular the HL-LHC needs in the context of a general KAT, KET, KHuK strategy
  - Roles of research centers and universities?
  - Community or workflow specific resources or science cloud?
  - Resource projections?
  - Required technological developments?
  - Required political boundary conditions?
  - Relations among communities, to international partners and funding agencies?
  - Long term sustainability?
- Tentatively planned for May at GridKa
- Open to all who are interested in the topic

BMBF News

25.10.2018 | FORSCHUNG

Wohin mit den gigantischen Datenmengen der Grundlagenforschung?

Experimente in der Grundlagenforschung sind Speicherfresser: 10 Millionen DVDs bräuchte man für die Daten, die jährlich am CERN anfallen. Mit innovativen Verarbeitungsmethoden wollen Forschende eines Computing-Verbundes dieser Datenflut Herr werden.

ErUM Data

- IDT-UM is a pilot project of ErUM Data
- ErUM Data action plan of BMBF expected this year
- Input from ErUM communities collected last year
  - ✔ Federated infrastructure
  - ✔ Big data analytics
  - ✔ Data management
- Communities:
  astro particle physics (KAT),
  particle physics (KET), astronomy (RDS),
  hadron and nuclear physics (KhuK),
  accelerator physics (KfB),
  research with neutrons (KFN) /
  synchrotron radiation (KFS) / ions (KFSI)

Scientists with doctoral degree

KFS	2,300
RDS	1,500
KHuK	1,500
KET	1,300
KFN	1,000
KAT	500
KfB	200
KFSI	100
	8,400

ErUM Data Community Input

- Workshop at BMBF 4./5.10.2018

5 Recommended measures and cost estimates

→ http://www.astroteilchenphysik.de/Offentlichkeitsarbeit_files/ErumData_DINA4_30.04.2019_Druck.pdf

- Strategy document of all ErUM communities given to BMBF on 2.5.2019

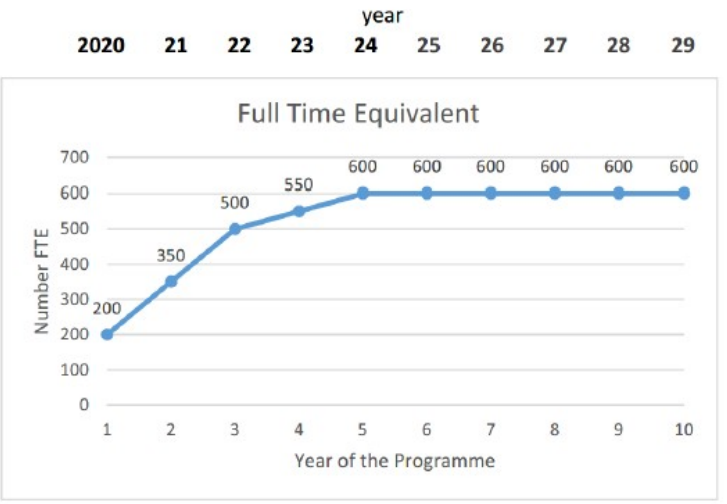
ErUM Data Cost Estimates

Full Time Equivalents

		MEuro/y /position in 2020
1. Workflows to exploit infrastructures	100	0.072
2. Management of research data	100	0.072
3. Big Data Analytics in physics research	200	0.072
4. Scientist's web working environment	100	0.072
5. Tenure track ErUM programme + 1 RA*	100	0.158
Total FTE	600	

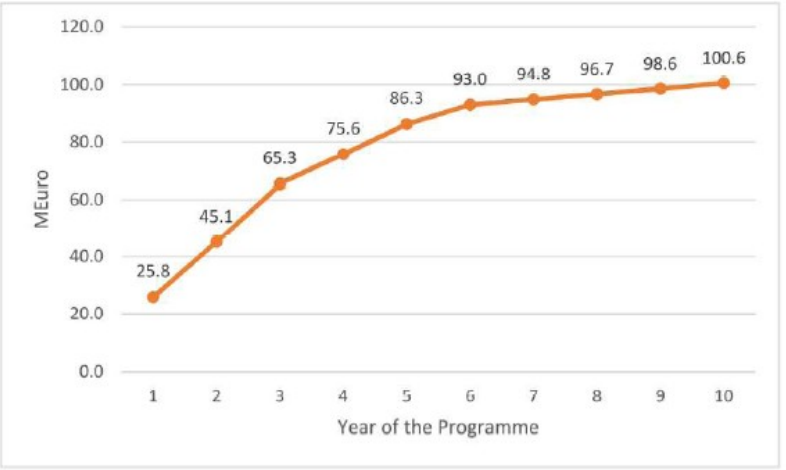
*RA=Research Associate

Due to the long-term nature of the responsibilities, these positions should ideally have long-term perspectives.



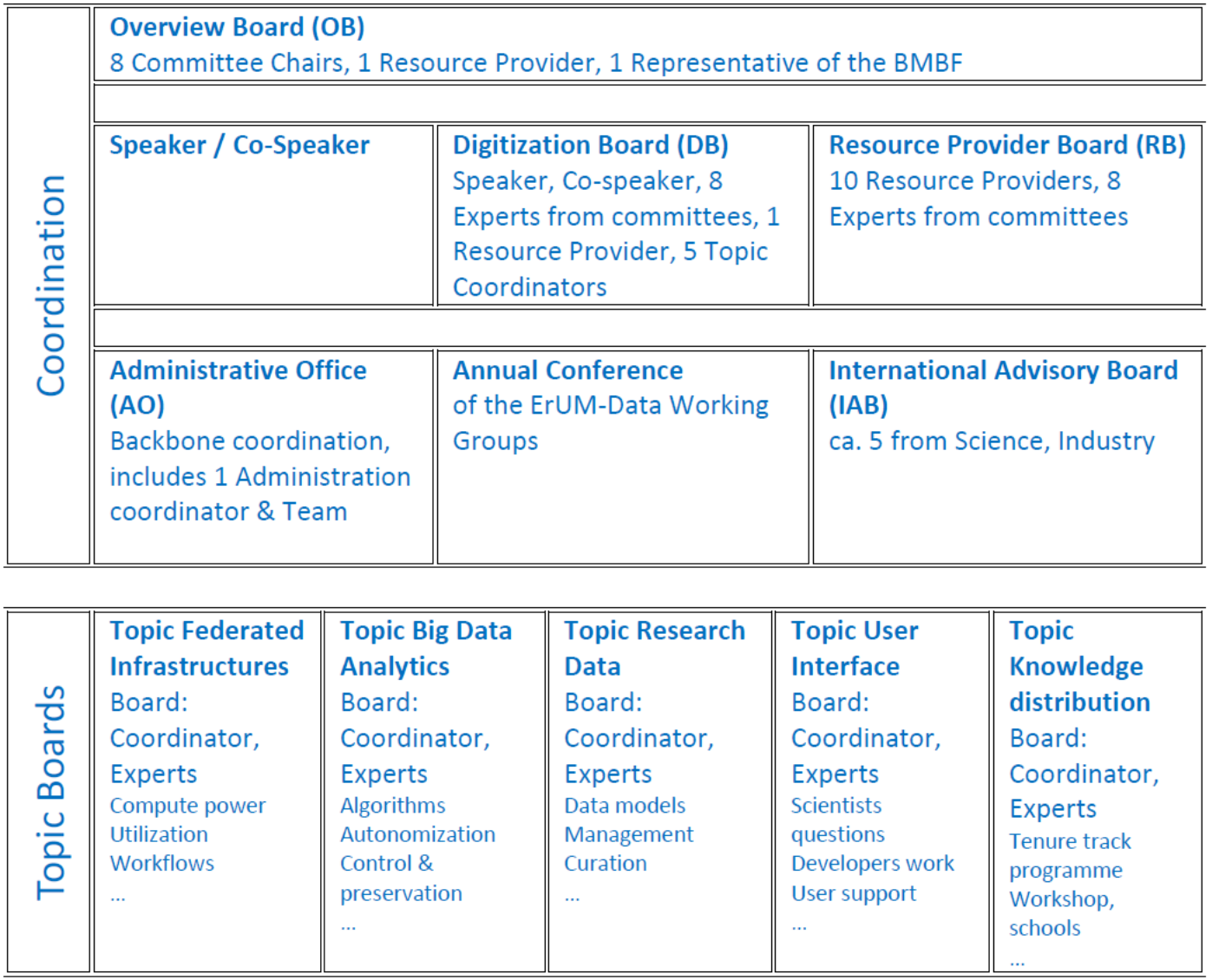
Cost estimate of recommended measures /MEuro

	year 2020	21	22	23	24	25	26	27	28	29	Meuro /topic over 10y
Full Time Equivalents	17.8	32.1	47.3	52.6	58.3	60.0	61.8	63.7	65.6	67.6	526.9
Large-scale federated infrastructures	5.0	10.0	15.0	20.0	25.0	30.0	30.0	30.0	30.0	30.0	225.0
Partnership for innovative digitization	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	30.0
Total Cost Estimate	25.8	45.1	65.3	75.6	86.3	93.0	94.8	96.7	98.6	100.6	781.9



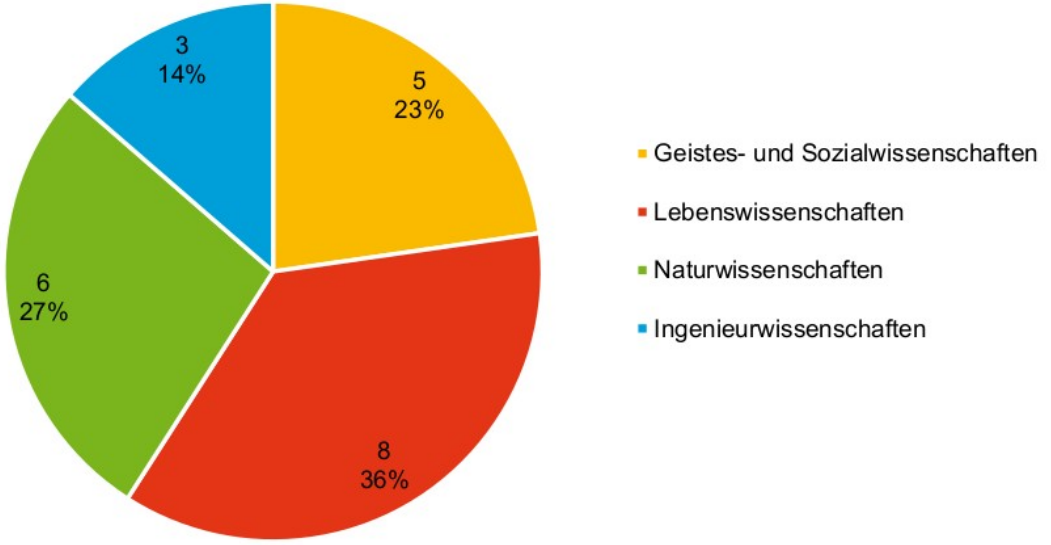
ErUM Data Organization

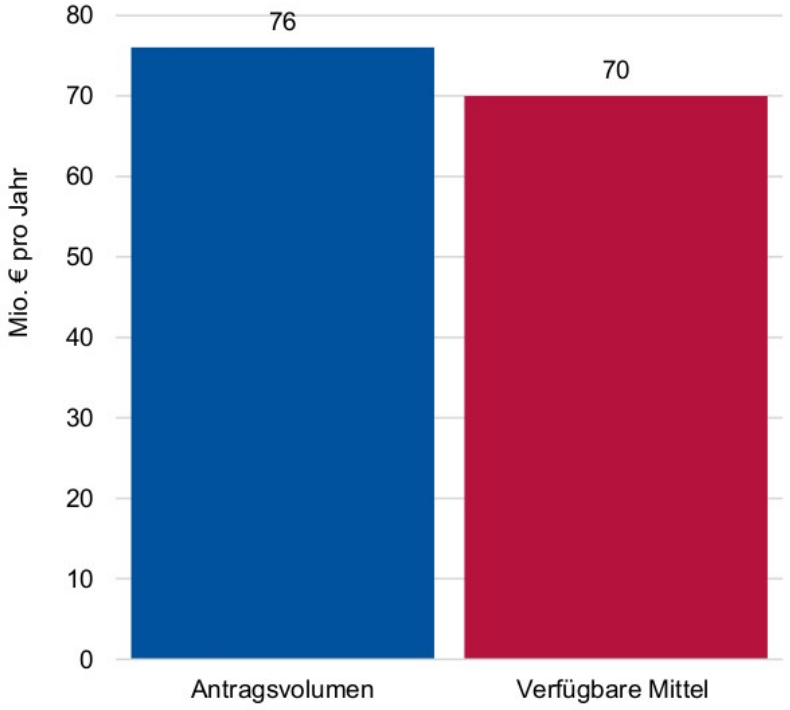
- Guidelines for organization across communities developed by digitization board



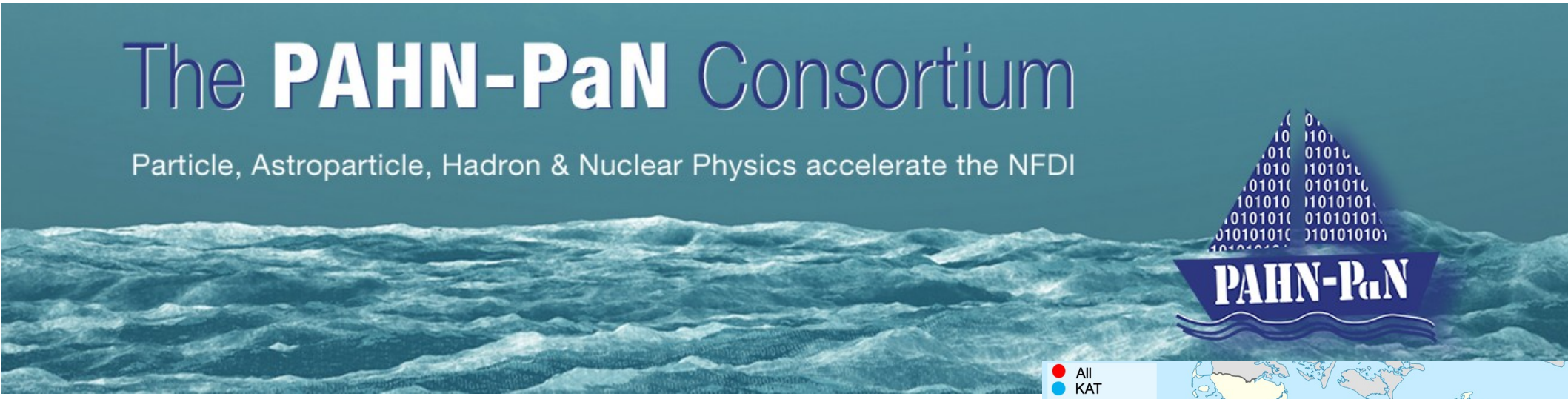
National Research Data Infrastructure

- The aim of the national research data infrastructure (NFDI) is to systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally. The NFDI will bring multiple stakeholders together in a coordinated network of consortia tasked with providing science-driven data services to research communities.





PAHN-PaN



Task area 1: Developing workflows and tools for data management

Task area 2: FAIR data lifecycle concepts and open data

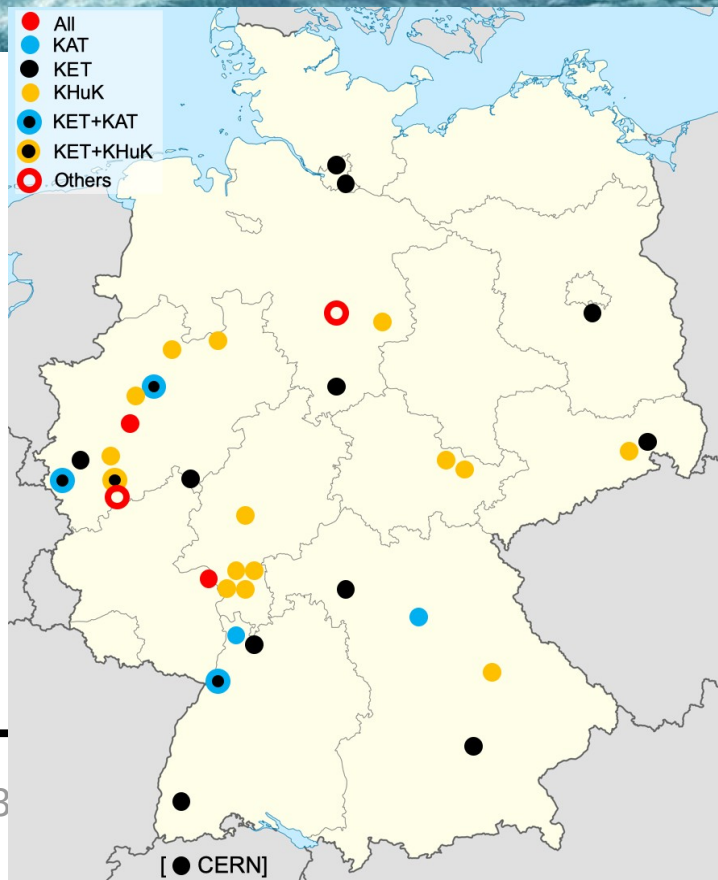
Task area 3: Data analysis procedures and services

Task area 4: Real-time data analysis and selection

Cross-cutting topic A: Synergies

Cross-cutting topic B: Services

Cross-cutting topic C: Professional training, education, outreach



Further National Context

- KAT Digitital Committee
- KET Computing and Software Panel
- KHuK Computing Committee

Arbeitskreis Physik,
moderne Informationstechnologie
und Künstliche Intelligenz

- DPG: AKPIK
- DPG: Physics and Information is one of four topics of 175th anniversary celebration
- ...

International Context

HEP Software Foundation

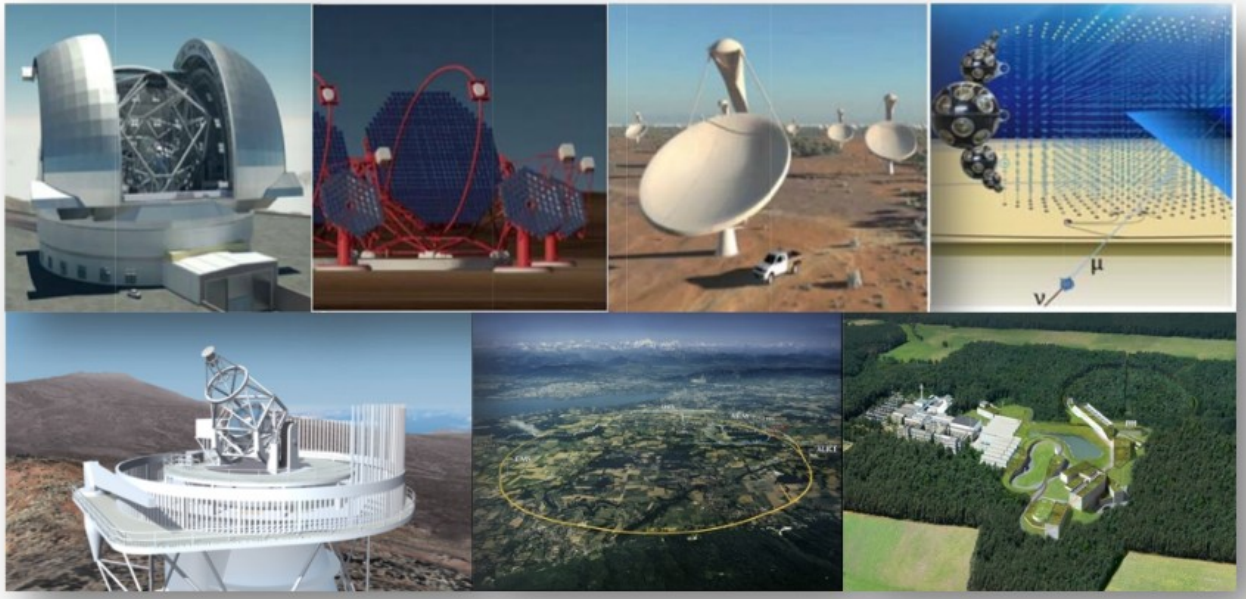
- International collaborations
- HEP Software Foundation: Community White Paper
  - Improvements in software efficiency, scalability and performance
  - Enable new approaches that can radically extend physics reach
  - Ensure the long-term sustainability of the software
- IRIS-HEP
- SIDIS: Software Institute for Data Intensive Science
- New journal: Computing and Software for Big Science
- EOSC: European Open Science Cloud
- …

ESCAPE

European Science Cluster of Astronomy & Particle physics
ESFRI research infrastructures

Horizon 2020
funded project

EUROPEAN OPEN SCIENCE CLOUD



Goals:

Prototype an infrastructure adapted to the Exabyte-scale needs of the large science projects.

Ensure the sciences drive the development of the EOSC

Address *FAIR* data management

Science Projects

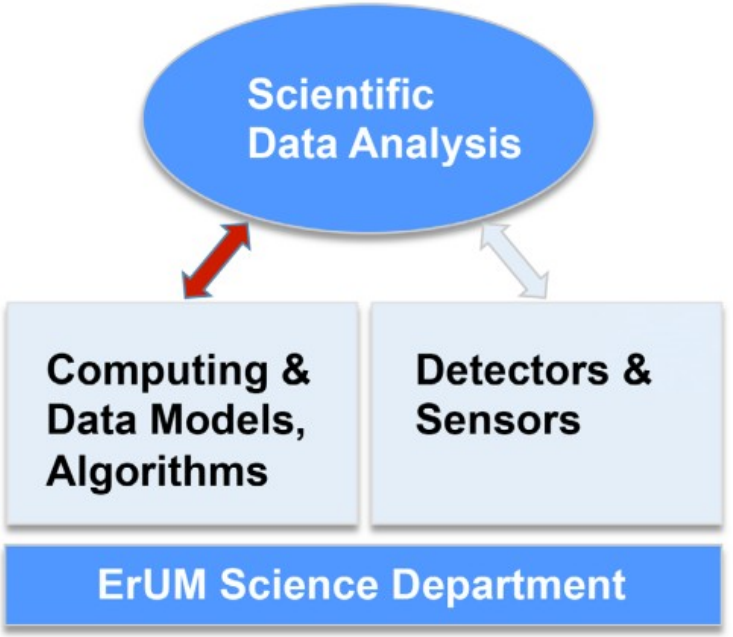
HL-LHC	SKA
FAIR	CTA
KM3Net	JIVE-ERIC
ELT	EST
EURO-VO (LSST)	EGO-VIRGO (CERN,ESO)

Data centres: CERN, INFN, DESY, GSI, Nikhef, SURFSara, RUG, CCIN2P3, PIC, LAPP, INAF

Simone Campana

Summary

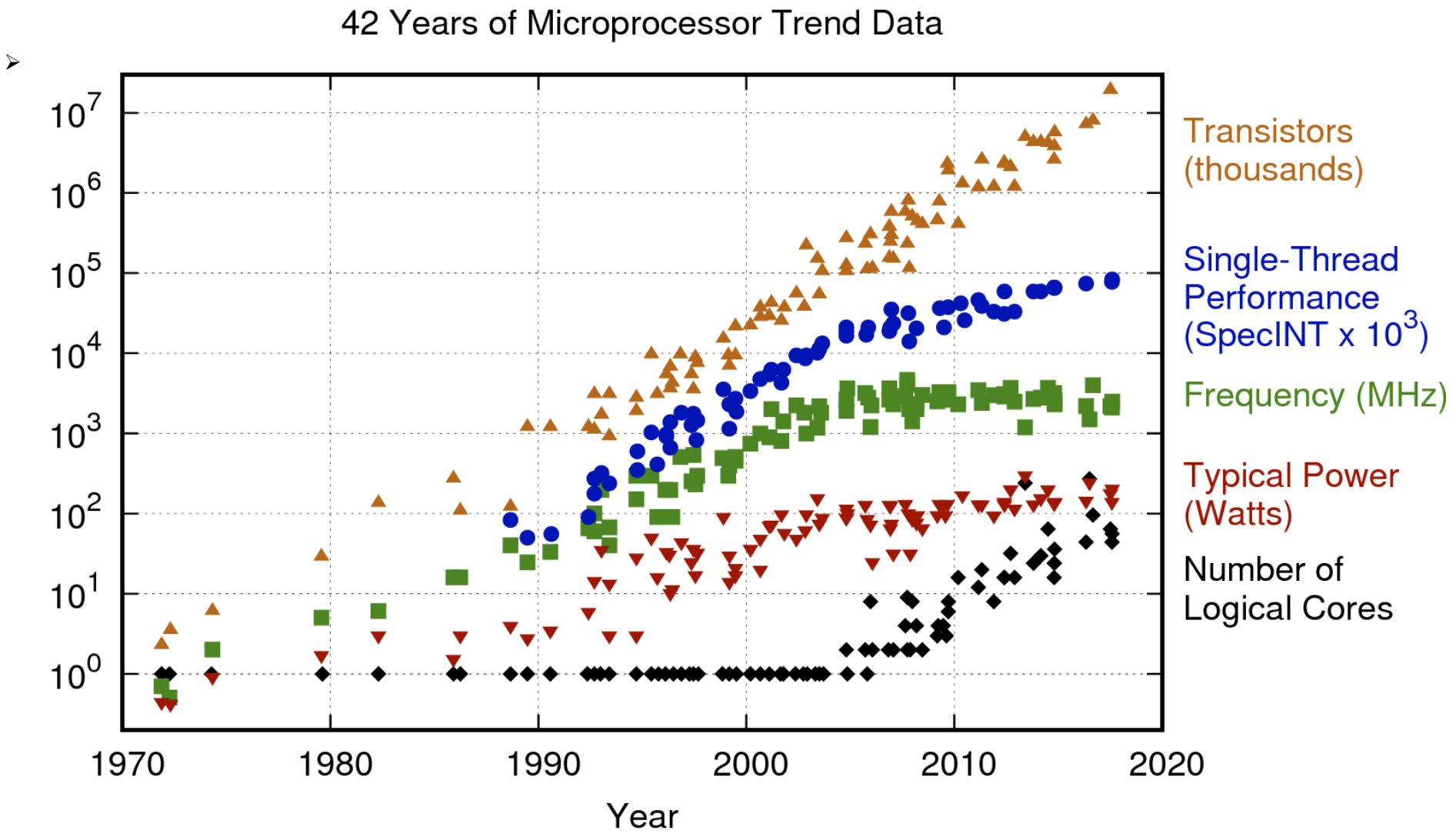
- Digitization offers opportunities to address the challenges of increasing data rates and volumes in fundamental research on universe and matter
- Development of common solutions encouraged by funding agencies
- → Pilot project with partners from particle, hadron and nuclear, and astro particle physics
- → ErUM Data can have high impact on our field of research



- A lot is currently happening in the field of digitization
- → You are part of this and can shape the future of science and society

Backup

Technological Evolution



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp