



JURECA
T-Platforms Supercomputer
@ FZ-Jülich



Thomas Lippert
Helmholtz-Programme
“Supercomputing & Big Data”
Juelich Supercomputing Centre



Research Data Infrastructures

in the
Helmholtz-Association,
in
Germany
and in
Europe

Science in the „era of data“

Digitization ...

- ... is epitome of the „**Era of Data**“
- ... covers **all fields** of science and research
- ... enters rapidly industry and science (e.g.: **autonomous driving**)
- ... places highest requirements with regard to
 - ★ data volumes, data rates, heterogeneity
 - ★ compute and storage requirements
 - ★ interoperability, standardization, data federation
 - ★ data curation, data security, authentication, reproducibility
 - ★ education, training

Create Knowledge from Data...

- ... is the **primary challenge** for science
- ... is reliant on **scientific analysis of big data**
 - ★ generation, transport, storage, processing, management
 - ★ integration, assimilation in simulation, hybrid modeling
 - ★ Machine Learning, Deep Learning
 - ★ Today: HPC / HPDA – Tomorrow: Quantum and Neuromorphic Computing

Situation in
GERMANY

Research Data in Germany

- Steady **growth** in terms of **quantity** / **heterogeneity**
- Large needs as to **funding** and **qualified staff**
- Too many obstacles **slow down** progress
 - heterogeneous funding schemes and poor coordination
 - focused on technology instead of process orientation
 - high diversity of disciplines
 - lack of quality assurance
 - diffuse mechanisms of reputation gain
- Lack of enforcement of policies

Council for Information Infrastructures (RfII)

Challenges

- Lack of **basic supply** of research data
- Inefficiency due to **fragmentation**, poor coordination
- Not aligned strategically → **broadness of application** limited
- **Loss of expertise** through staff turnover

Recommendations

- Position paper „**Leistung aus Vielfalt**“
- Recommendation: **National Research Data Infrastructure** (NFDI) as collaborative, distributed network
 - Sustainability
 - data responsibility of the scientists
 - development in an international context
 - human resource development on all levels



Engagement of
HELMHOLTZ ASSOCIATION

Helmholtz position paper

Take added value from the resource ,information‘!

Goals

- Improve the potential of information by reinforcing ,Digital Science‘, **towards knowledge**
- strengthen **cross-cutting competences** of all research fields

Actions

- focused research in the area of information technology
- process and give access to research data through research data infrastructures (→ Helmholtz Data Federation - HDF)
- participate in national and international initiatives
- strengthen **human resources development**

Helmholtz Data Federation – HDF

- HGF provides a **comprehensive and supportive** system for data intense science on the **Exascale level**
- **Cross-domain** challenges are linked and coordinated with **domain-specific** requirements
- **Open to** the entire German scientific landscape
- **Secure federation** of data centres, seamless **interaction** with other German and **European** data infrastructures and integration in the HPC-ecosystem
[Gauß-Allianz (D), PRACE/HBP /EOSC (EU)]

HDF: 6 centres, 5 fields of research

- Federation and expansion of **multi-thematic data centres**
- Data- Management- Solutions become **cross-sectorial**
- User support by **DataLabs / SimLabs**
- Ready to support the planned **National Data Infrastructure**



Achieve **Scalability** on Many Levels

- Cope with **massively-heterogeneous** data sets
- Enable processing of **extreme** amounts of Big Data
- Integrate **all** Helmholtz centres and groups
- Integrate data formats of **different** research fields
- Guarantee **Interoperability** of all metadata formats
- Create scalable support to **annotate metadata**
- **Federate** with other data infrastructures in Germany and Europe
- Enable fast processing by **scalable HPC technology**
- Integrate other scalable **data networks** (HGF, D, EU, world)
- Ensure **data security**
- Guarantee **long-term availability** of data

Joint task:

Push Information and Data Science forward

Helmholtz-Inkubator „Information & Data Sciences“

- Think Tank with scientists from **all** research fields

Goals

- **networking** of the brightest minds within the HGF
- Form a basis for **innovative networks**
- Identify most interesting **topics and pilot projects**
- Draw up a common, **long-term strategy**
- Develop a common **education platform**

Funding

- Impuls und Vernetzungsfonds

Interaction with HDF

- Inkubator projects will promote the **effective usage of HDF**
- HDF as a **service platform** for Inkubator projects (IaaS)

Application example:

COHORT STUDY IN NEUROSCIENCE

1000 Brains Study (I)

Population based data collection in the framework of 'National Cohort'

FZJ

INM1, Prof. Amunts

INM1, Prof. Eickhoff

INM1, Prof. Caspers

INM4, Prof. Shah

INM8, Prof. Sturma

National Cohort

- 2014-2024
- Sample 200.000 persons, base protocol
- Full-body MRT 30.000 persons, extended protocol

1000-Brains Study

- Detailed neuropsychological / MRT examination of same samples

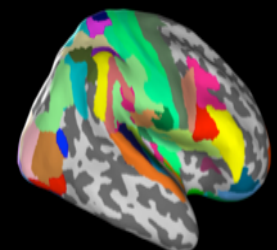
Heterogeneous data

- 250 TB for 1000 brains only

GEFÖRDELT VOM



Bundesministerium
für Bildung
und Forschung



1000 Brains Study (II)

Population based data collection within the framework of 'National Cohort'

FZJ

INM1, Prof. Amunts

INM1, Prof. Eickhoff

INM1, Prof. Caspers

INM4, Prof. Shah

INM8, Prof. Sturma

Examination of the brain's natural ageing process

Variability of structure and function under influence of several factors

Scientific relevance

- **Only way to „weak“ factors** like environment, nutrition, profession
- these factors interact with another and the **genetic Disposition**
→ phenotype like blood pressure or # of plaques
- environmental factors are able to retroact on genes = **Epigenetic**

GEFÖRDERT VOM



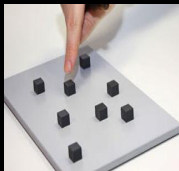
Bundesministerium
für Bildung
und Forschung

Economic relevance

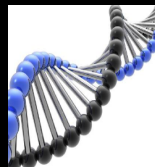
- identify **carrier of the genes**, who **are responsive to specific medication**
- individualized medicine



MRI



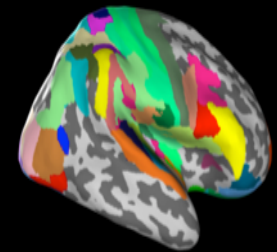
Neuropsycho-tests



Genetic Analysis



Motor function tests



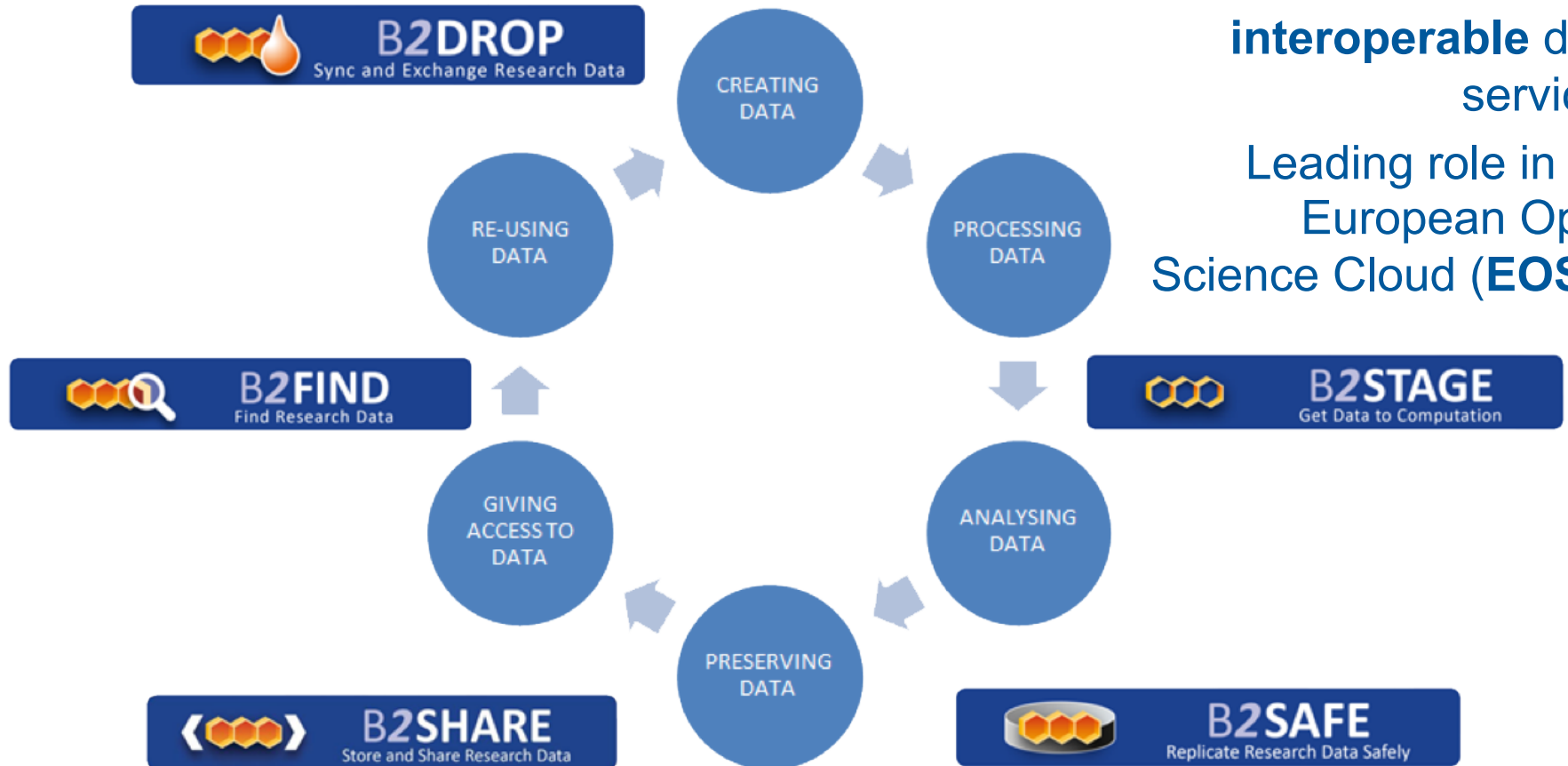
Examples of

EUROPEAN DATA INFRASTRUCTURES

1. EUDAT – Support of the Data Life-Cycle

European data
infrastructure for
interoperable data
services

Leading role in the
European Open
Science Cloud (**EOSC**)



2. FENIX – Data Concept of the Human Brain Project

RESEARCH INFRASTRUCTURE

DATA ANALYSIS & SIMULATION

- Federated EuropeanN Information Exchange
- Federating multiple data resources enables data replication at different sites
- FENIX IaaS connects archive data and “active storage” in close proximity with supercomputers

ETHICS & SOCIETY

Decoding the Human Brain

Multi-Scale System in space and time, multimodal

NEUROSCIENCE
EXPERIMENT & THEORY

Amunts et al., The Human Brain Project: Towards a European Infrastructure to Decode the Human Brain; Neuron, 2016

Summary

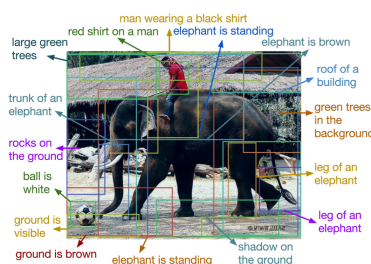
HGF has recognized the needs of today :

- Helmholtz **Data Federation**
- Helmholtz-**Inkubator** “Information und Data Sciences”
- Ready to support the **National Data Infrastructure**
- Interoperability and **claim to leadership** at the European level

We are aware of the convergence of HPC and Data analytics



**Mass-
Data-Analysis**



**Deep
Learning**



Interactivity



HPDA on Eascale level

JURECA
T-Platforms Supercomputer
@ FZ-Jülich

CREMLIN WP2 Big Data Management Workshop



- Within CREMLIN WP2 (“Exchange platform”): European-Russian Workshop on “Big data Management”
- Date and location: 15-16 February 2017, at the NRC “Kurchatov Institute”
- 30 European and 30 Russian experts will discuss challenges and possible solutions on Big data management in the context of large-scale RI /megascience facilities
- *Task 2.4: to “support the integration of the Russian megascience projects into the European and global e-infrastructure initiatives”.*
- Exchanging with key stakeholders from EOSC (European Open Science Cloud), GÉANT, EGI (European Grid Initiative), PRACE aisbl, RDA (Research data Alliance)