

Datenintensive Wissenschaft in der Helmholtz-Gemeinschaft

Dr. Andreas Heiss
(Folien von Prof. Dr. Achim Streit)

Steinbuch Centre for Computing



Helmholtz-Gemeinschaft

- **Strategische Forschung für die “großen Herausforderungen”** in sechs Forschungsbereichen: Energie, Erde&Umwelt, Gesundheit, Luftfahrt, Raumfahrt&Verkehr, Materie, **Schlüsseltechnologien**
- Entwicklung und Betrieb komplexer und großer **Forschungsinfrastrukturen** für nationale und internationale Wissenschafts-Communities.



	Budget/ Billion€	Staff in FTE*	Centres/ Institutes
Helmholtz Association Use-inspired basic research with strategic programmes	3.29**	31,679	18
Max Planck Society Pure basic research	1.65	13.308	87
Fraunhofer Society Industry-oriented research and development	1.8	15.815	66
Leibniz Association Long-term research topics	1.3	13.230	86

Source: GWK Monitoring Report 2013 Joint Initiative for Innovation and Research
 *Staff in working hours (full-time equivalent)
 **excluding project sponsorships, project management agencies and other revenues

Programm “Supercomputing & Big Data”

Programmsprecher: Prof. Dr. Dr. Thomas Lippert (FZJ)

- **Supercomputing** und das **Management und die Analyse von “Big Data”** sind strategisch wichtige Technologien, die moderne Wissenschaft erst ermöglichen, neben Theorie und Experiment.
- **Entwicklung und Betrieb** von unverzichtbaren **Werkzeugen** und **Infrastrukturen** im Bereich HPC und Datenmanagement für das **deutsche** und **europäische** Wissenschaftssystem.



Topic 1 “Computational Science and Mathematical Methods”



Topic Speaker: Prof. Dr. Paul Gibbon (FZJ)



Topic 2 “Data-Intensive Science and Federated Computing”



Topic Speaker: Prof. Dr. Achim Streit (KIT)



Topic 3 “Supercomputer Facility”

Topic Speaker: Dr. Norbert Attig (FZJ)

Karlsruher Institut für Technology (KIT)

■ **Mission:** Die Forschungsuniversität in der Helmholtz-Gemeinschaft



■ **Zahlen und Fakten:**

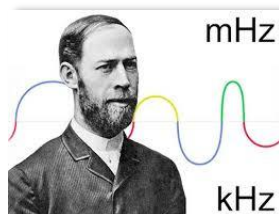
- > 850 M €
- > 25.000 Studenten
- > 150 Institute
- > 9.000 Angestellte
 - (> 5.800 in Forschung und Lehre , > 350 Professoren)



■ **Tradition and Vision**



Carl Benz



Heinrich Hertz



Ferdinand Braun



The first Faculty of Informatics in Germany



One of the largest energy research centers in Europe

@ The first E-Mail received in Germany

Wer sind wir?

Zentrale
wissenschaftliche
Einrichtung des KIT

Welche Anforderungen erfüllen wir?

- Computational Science & Engineering (CSE)
- Data-intensive Science (DIS)
- Anwender im KIT, BaWü, Deutschland und international

Was machen wir?

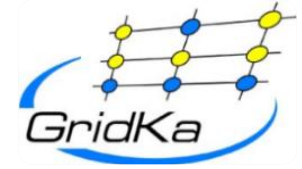
- Forschung, Ausbildung und Innovation in den Bereichen Supercomputing, Big Data und sichere IT-Föderationen
- Betrieb von großen IT-Infrastrukturen
- Betrieb der Basis-IT-Dienste

“Services for Science – Science for Services”

Daten-Intensive Wissenschaft (DIS)

■ Betrieb von **Datenzentren**

- GridKa – deutsches Tier1-Zentrum für LHC (Teilchenphysik)
 - 20 PB Onlinespeicher, 30 PB auf Bandspeicher, 16k CPU-Kerne
- Multi-disciplinary Large-Scale Data Facility
 - >6 PB Onlinespeicher für Klima, Energie, Systembiologie, ...



■ Betrieb von **HPC-Systemen**

- ForHLR mit direktem Anschluss an LSDF
 - > 35 k CPU-Kerne, 1.4 PetaFlop/s Peak, Energie-effiziente Warmwasser-Kühlung



■ **Gemeinsame Forschung** mit Wissenschafts-Communities

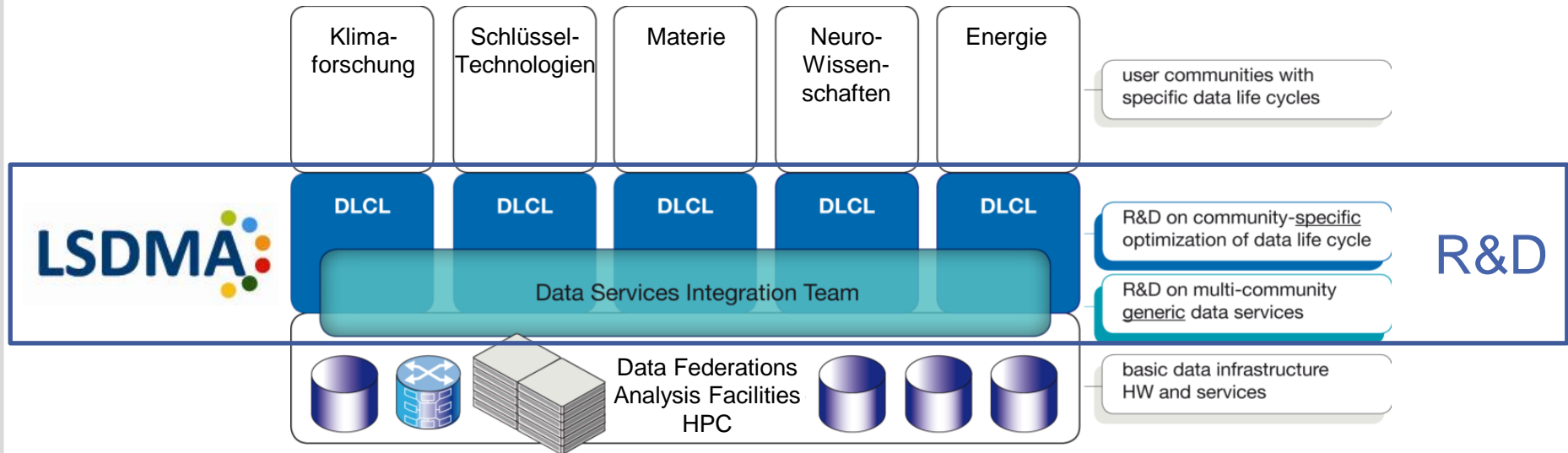
- Data Life Cycle Labs (DLCLs) und SimLabs
- Tools und Methoden für Datenmanagement und -analyse
- Performance Engineering für hoch-parallele Simulationen



■ **Innovations-Treiber** für KMUs and Industrie



Large Scale Data Management and Analysis



Data Life Cycle Labs

Gemeinsame Forschung mit den Anwendungswissenschaften

- Optimierung des Data Life Cycle
- Spezifische Datenanalyse-Tools, Methoden und Services

Data Services Integration Team

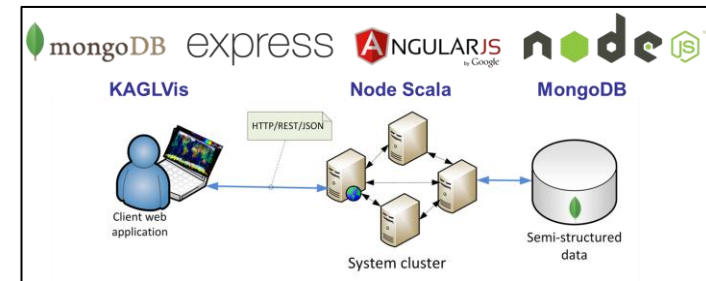
Generische Forschung und Entwicklung

- Schnittstelle zwischen föderierten Data-Infrastrukturen und DLCLs bzw. Communities
- Integration von Datenservices in die wissenschaftlichen Arbeitsprozesse

Ausgewählte Aktivitäten gemeinsamer F&E mit Anwendungswissenschaften

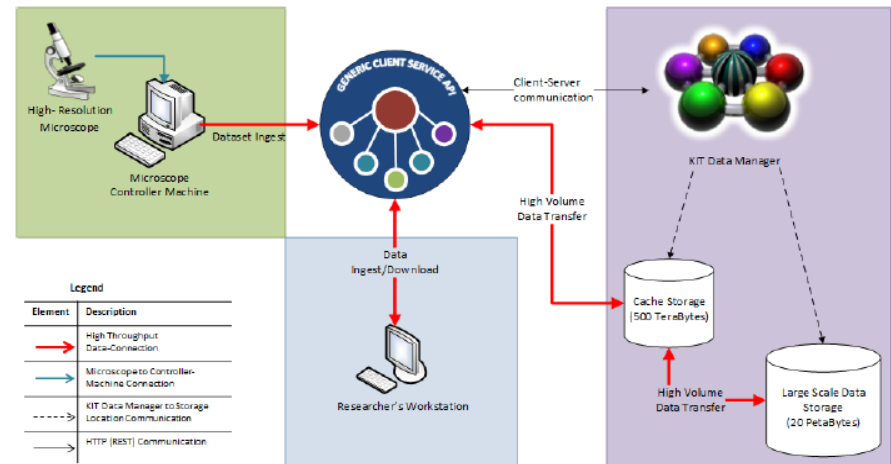
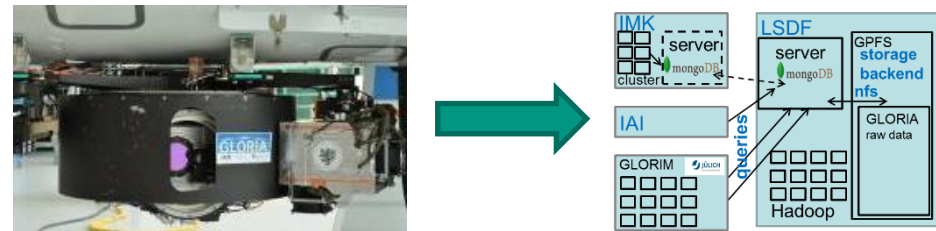
■ Klimaforschung

- Klima-Analyse mit dem “MEAN Stack”
- NoSQL DBs für GLORIA-Daten



■ Schlüsseltechnologien

- Automatische Prozessierung von Mikroskopdaten (Zebrafish-Imaging)
- Offenes Referenz-Daten-Repository; Management von O(100TB) Datensätzen



Ausgewählte Aktivitäten gemeinsamer F&E mit Anwendungswissenschaften

■ Materie

- Entwicklung von Tools und Methoden für die Forschung mit

- FAIR

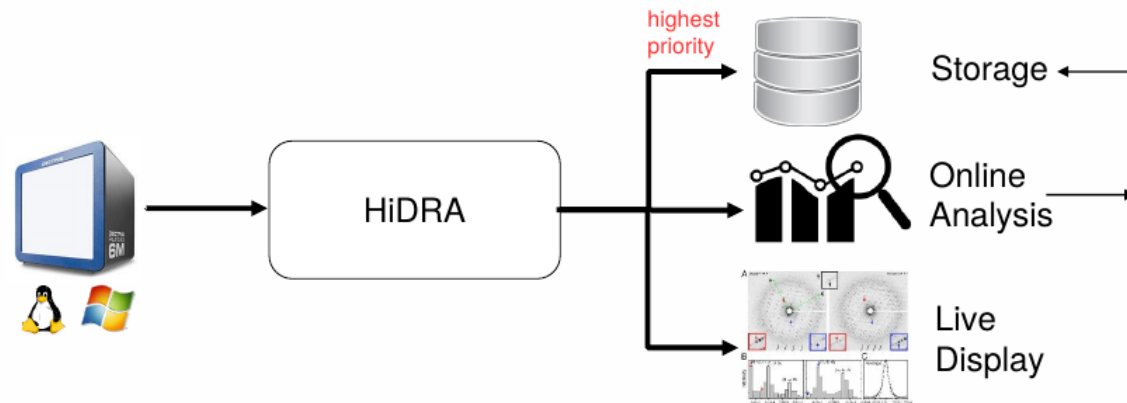


- Petra III



- European XFEL

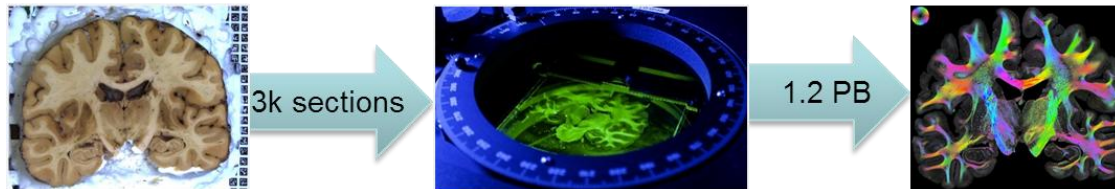
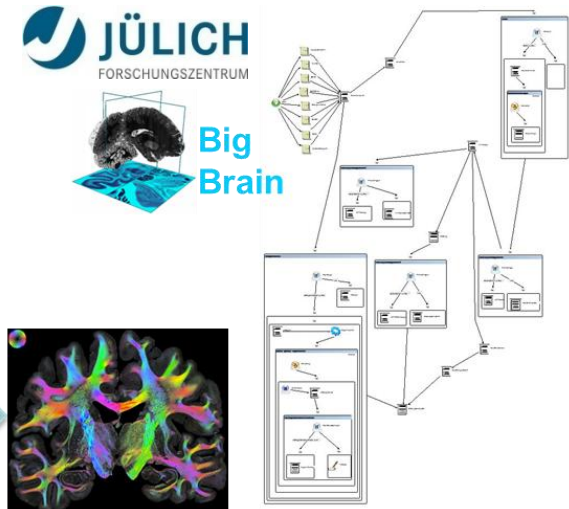
- Bsp: **HiDRA**: Generisches Toolset für high-performance Datenmultiplexing mit verschiedenen Service-Levels



Ausgewählte Aktivitäten gemeinsamer F&E mit Anwendungswissenschaften

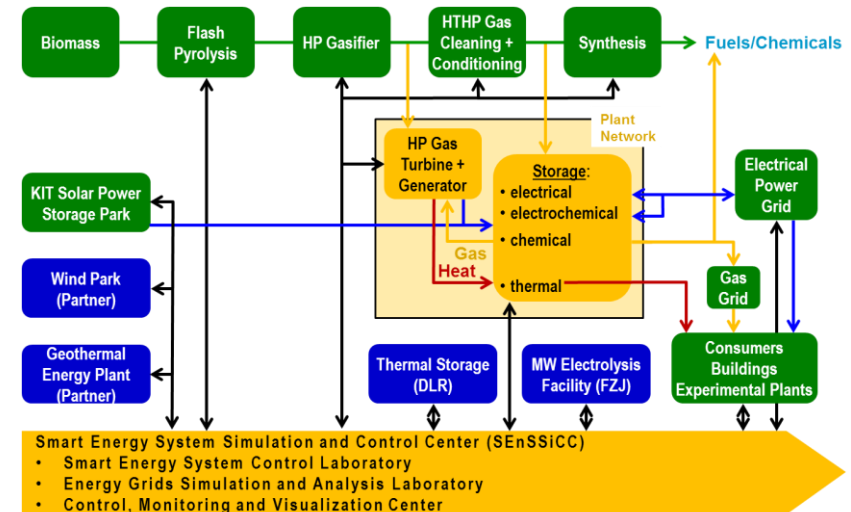
Neuro-Wissenschaften

- Analyse komplexer Hirn-Strukturen



Energie

- Sicheres Datenmanagement für techn. & ökonomische Analyse
- Analysierung, Modellierung und Simulation künftiger "Smart Power Grids" mit hohem Anteil erneuerbarer Energien.



Ausgewählte Aktivitäten der generischen F&E

■ AAI / Identitäts-Föderationen

- Integration von FedID (eduGAIN) und OpenID in verschiedenen Services
- SAML (Kommandozeilen-)Tools
- Interoperabilität von AAI, Sicherheitslevels und Identitäten

■ Archiv-Technologien

- Internationale Harmonisierung von QoS Policies für Speicher über RDA and SNIA/CDMI
- Benutzer-Registrierungs-Workflow für Langzeit-Daten-Archive
- Gemeinsamer Ansatz in EUDAT, INDIGO-DataCloud, RADAR, ...

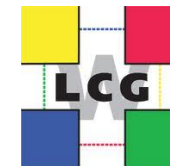
■ CDMI support in UNICORE und dCache

■ Repository System KIT-Data-Manager

■ Generische Ansätze für Metadaten für verschiedene Communities.

■ UNICORE-Erweiterungen

Kollaborationen, Projekte, Initiativen



Human Brain Project



Helmholtz Data Federation (HDF)

- Helmholtz entwickelt eine **föderierte Forschungsdateninfrastruktur** in Deutschland, offen für das gesamte deutsche Wissenschaftssystem.
 - Sichere Föderation bestehender Datenzentren und Verbindung zu EU/Intern.
 - Berücksichtigung community-spezifischer Anforderungen
 - Langzeit-Speicherung, “Curation” und Verfügbarkeit von Forschungsdaten
 - Multi-disziplinäres Arbeiten mit Daten
 - Datenhoheit bleibt erhalten
 - Koordiniert durch das KIT
- Nationaler Baustein der **European Open Science Cloud (EOSC)**
- Drei Elemente:
 - Innovative **Software** und **Services** für Datenmanagement
 - Exzellente **Benutzerunterstützung** und gemeinsame F&E
 - Neue und innovative Speicher und Analyse-**Hardware**

Helmholtz Data Federation (HDF)

- Start mit 6 Helmholtz-Zentren in 2017

- Gutachter-Aussagen:
 - “Sharing of infrastructure, services and datasets **opens the way for improved cross-disciplinary working and re-use of data.**”
 - “...**very strong potential to facilitate cutting-edge research.**”
 - “...**great potential to contribute to solving major challenges facing society** (including pure attainment of new insights).”
 - “...will be a **first example of a national shared research data infrastructure in Europe.**”



- “Think-Tank” mit ca. 40 Expertinnen und Experten aus allen Helmholtz-Zentren und namhaften, forschenden Unternehmen.
- Ziele:
 - Bündelung vorhandene Kompetenzen, Brücken zwischen Forschungsbereichen und Programmen bauen.
 - Identifizierung strategisch wichtige Forschungsfelder in Helmholtz
 - Zunächst Förderung von 5 Projekten in prioritären Forschungsfeldern. (17 Mio. € über 3 Jahre), u.a.
 - Helmholtz Analytics Framework (FZJ, KIT, DESY, DLR, DKFZ, HMGU):
 - Austausch von Methoden zwischen einzelnen wissenschaftlichen Bereichen
 - Standardisierung von Tools und Methoden

The Helmholtz Association is a key research organisation for enabling computational and data-intensive science in Europe

Danke!