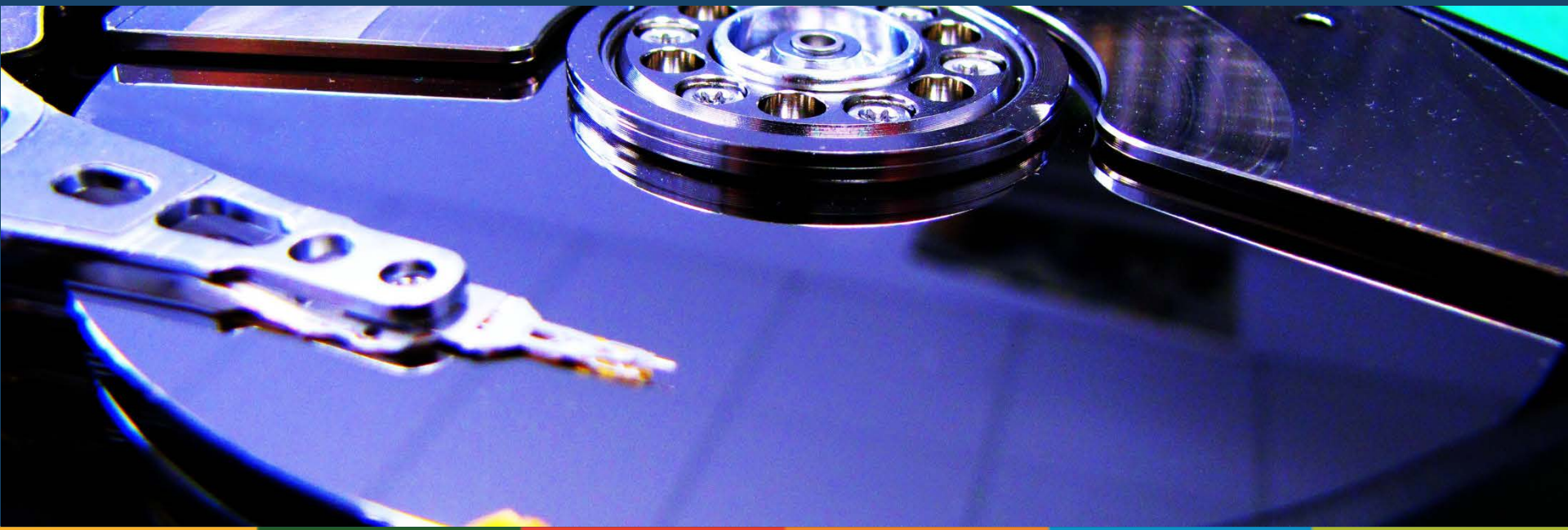


# Optimization of data life cycles

M. Gasthuber, DESY; A. Giesler, FZ Jülich; K. Schwarz, GSI;  
M. Hardt, C. Jung, J. Meyer, F. Rigoll, R. Stotzka, A. Streit,  
KIT





# (Big) Data in Science



Data life cycle as a central part of the scientific life cycle

## Big Data

- Volume
- Variety
- Velocity
- (Veracity)

## Data in modern science:

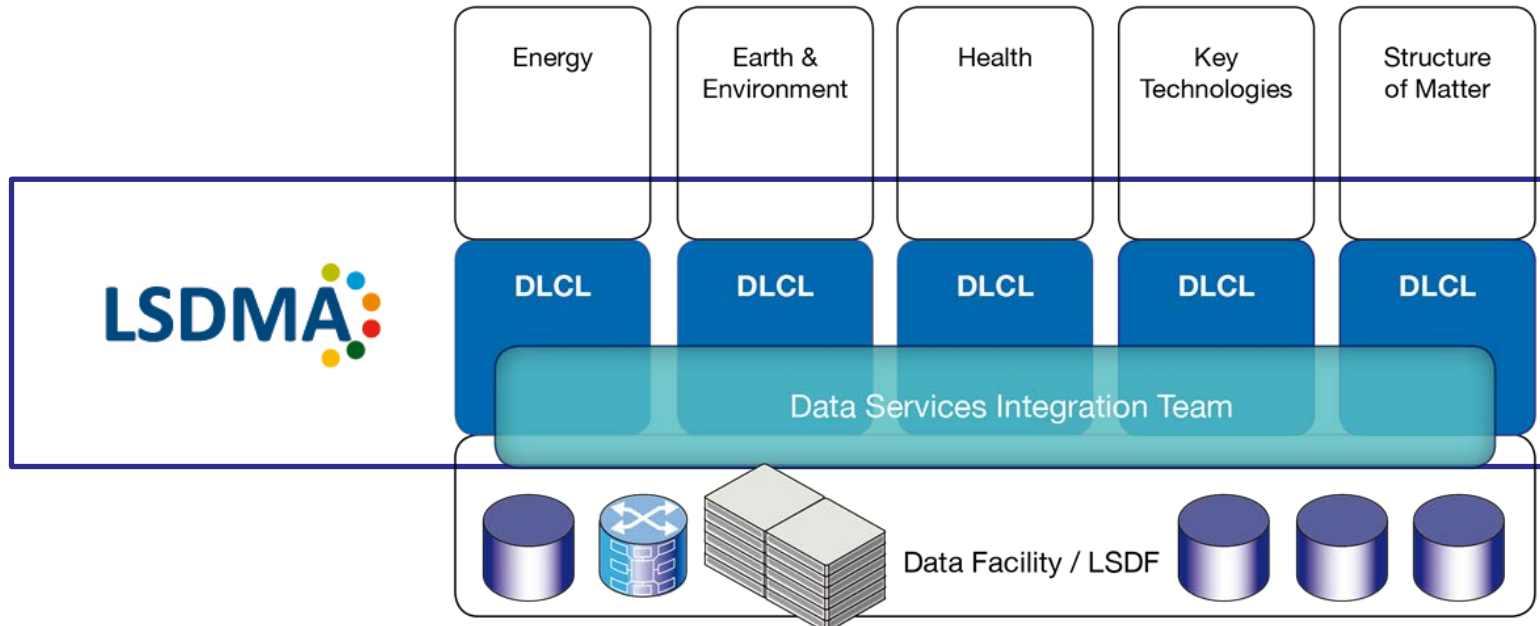
- Valuable good
- Data deluge
- Data exploration as 4<sup>th</sup> pillar

# Aspects of Data Management

- Access:
  - Authentication and Authorization Infrastructure
  - Persistent identifiers
  - Open Access/Data
- Archival, Preservation, Curation
- Federations
- Provenance
- Metadata
- Security and Privacy
- Safety
- Hierarchical Storage
- ...

Scientific communities usually only care about some aspects of data management

# LSDMA: Dual Approach



## Data Life Cycle Labs

Joint R&D with scientific user communities

- Optimization of the data life cycle
- Community-specific data analysis tools and services

## Data Services Integration Team

Generic methods R&D

- Data analysis tools and services common to several DLCLs
- Interface between federated data infrastructures and DLCLs/communities

# LSDMA Facts & Figures



- Initial duration: 2012-2016
  - Project is a Helmholtz portfolio extension → inclusion of activities into Helmholtz program-oriented funding in 2015, cross-program initiative
- Partners:
  - Helmholtz Association: KIT, DESY, FZJ, GSI
  - External: DKRZ, U-Heidelberg, U-Ulm, TU-Dresden, U-Hamburg, HTW-Berlin, U-Frankfurt
- Coordination: KIT



# Events and Collaborations

- Events:
  - Annual symposium “The Challenge of Big Data in Science”
  - Annual Community Forum
  - Planned: Technical Forum
  - Cooperation w/GridKa School
- LSDMA members involved in international projects and ESFRI items:
  - Human Brain Project
  - EUDAT
  - Research Data Alliance
  - DARIAH
  - European XFEL
  - FAIR



## Six work packages

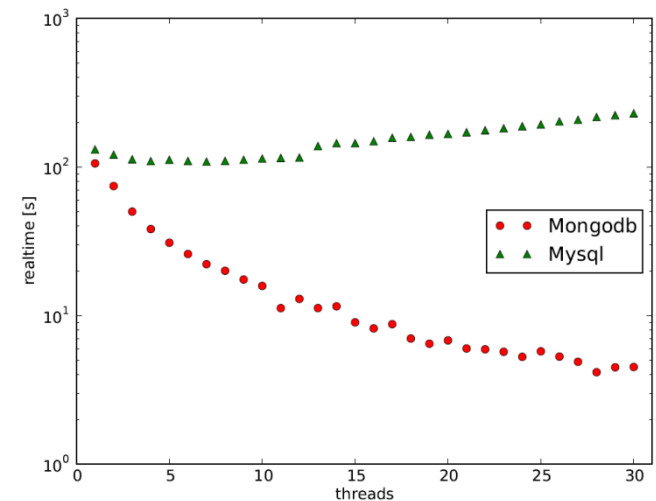
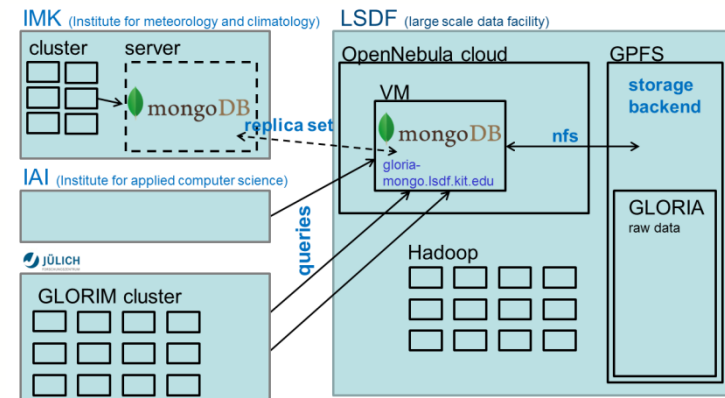
- Federated Identity Management
  - Credential translation between SAML and other authentication services (X.509, OpenID, OAuth2)
- Federated Data Access
  - Performance tuning of services, e.g. dCache and other scalable object stores
  - Simplification of access to data for users: http/WebDAV, NFS backend, Globus Online, KIT Data Manager
  - WAN-like control mechanisms for federated data access and metropolitan area networks
- Metadata Catalogs and Repositories
  - Search over an arbitrary number of UNICORE metadata instances
  - Design and implementation of 'Open Archives Initiative Protocol for Metadata Harvesting' (OAI-PMH) infrastructures



- Archive Service
  - Creation of policies and service levels: where, whether and how long datasets are to be archived
  - Bitstream preservation: evaluation of different archival middlewares
- Monitoring, Modeling, Optimization
  - Monitoring the I/O performance for analysis of performance loss between subsystems using SIOX framework and visualizing with VAMPIR
  - Monitoring information within LUSTRE for automatic detection of performance degradation
- Data-intensive Computing
  - Automatic and policy based triggering of analysis workflows, e.g. metadata extraction after data ingest
  - LAMBDA execution framework for large scale applications

# DLCL Earth and Environment

- Data sources:
  - Instruments, e.g. GLORIA, MIPAS
  - Simulation
- Various data formats: HDF5, NETCDF, ASCII
- Goals:
  - Speedup and simplification of analyses by
    - new database allowing fast parallel access
    - matching of geo coordinates inside DB
    - import of all satellite meta data
  - Simplification of analysis workflows
  - Full safe replication for ENES (w/FZJ, DKRZ and CSC within EUDAT; based on iRODS) being set up

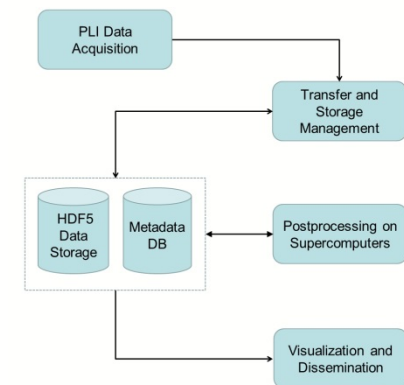
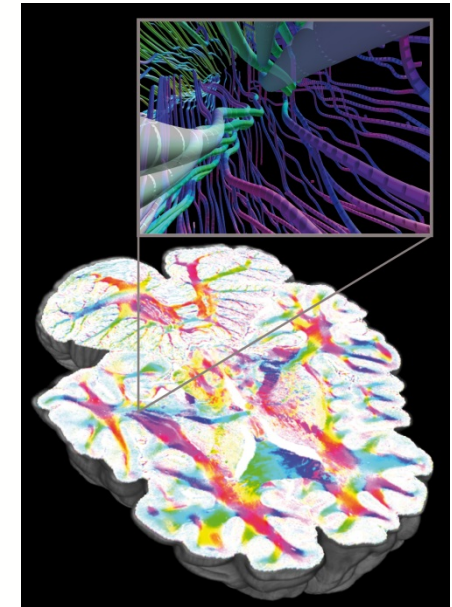


- SISKA: analyzing stereoscopic satellite images for estimating the efficiency of solar energy
  - Complex workflow w/several languages used
  - Porting to the cluster of the Large Scale Data Facility (LSDF) at KIT
- Most energy data concerns privacy (e-cars, at work, ...)
  - Standard anonymization or pseudonymization techniques often insufficient
  - Affects data usability
  - New focus of DLCL activities



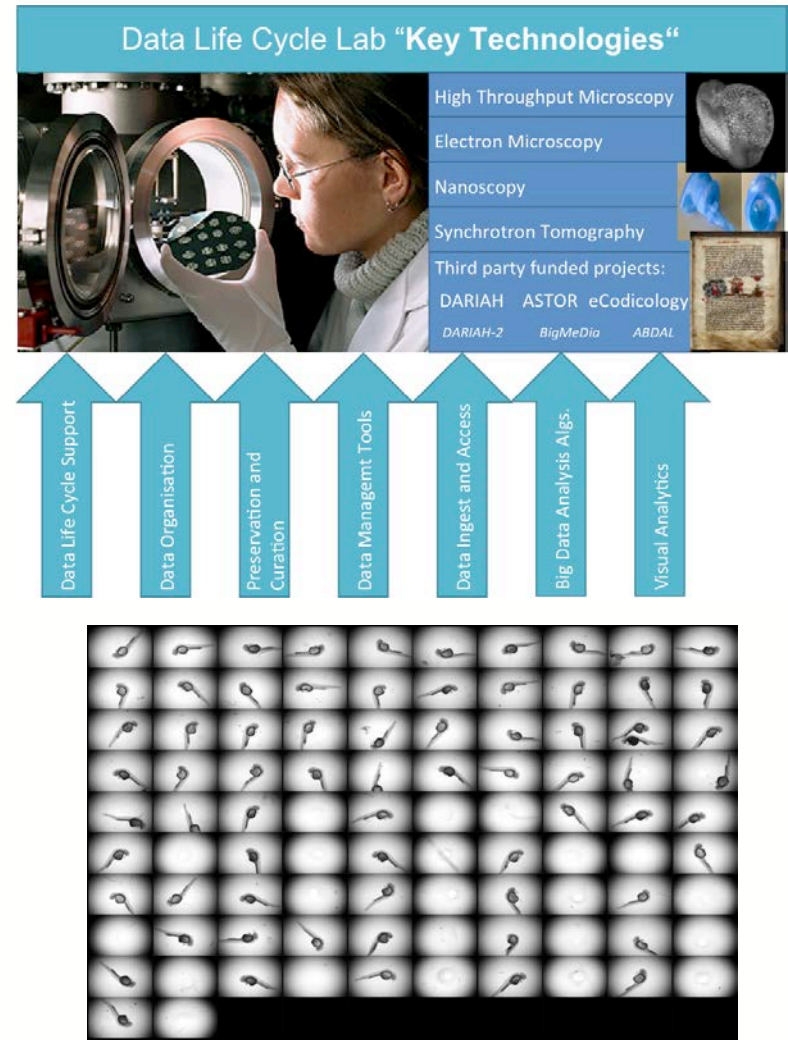
courtesy of SISKA

- anatomical connections of cortical areas and subcortical nuclei
  - Polarized light imaging (PLI) in human postmortem brains with resolution of  $\mu\text{m}$  at FZ Jülich
  - Secure, performant transport mechanism with UFTP (up to 4 times faster than SCP)
  - Parallelized post-processing on HPC
  - Long-term storage
- Brain Big Data project in collaboration w/U-Düsseldorf: high resolution scans of postmortem brain slices
  - Central storage at FZ Jülich
  - Policy based replication w/iRODS



# DLCL Key Technologies

- Ultra-fast imaging with spatio-temporal observations (even in living species)
  - Algorithms on GPGPU systems allow quick preview of recorded volumes
  - New algorithm for high quality images using fewer projections
- High throughput microscopy for zebrafish embryo development
  - 3D volumes of embryos
  - Automatic data transfer and ingest into LSDF w/KIT Data Manager



# DLCL Structure of Matter

- Photon Science has several similarities with HEP, differs in analysis, re-use and collaboration sizes
  - I/O tracing for detailed studies of storage systems used
  - Baseline support for HDF5
  - Parallel execution and programming
- Heavy ion research at FAIR, triggerless detector
  - Perform high speed (online) data processing in real time
  - ZeroMQ integrated into FairRoot
  - Metropolitan area network via fibre link
  - Optimization of data rates between U-Frankfurt and GSI



## Communities differ in

- Previous knowledge
- Level of specification of the data life cycle
- Tools, services, formats used
- Size

## Needs driven by

- '3 Vs'
- Cooperation between groups
- Policies
  - Open Access/Data
  - Long-term preservation
  - Data privacy

## Lessons learned

- Communities:
  - Focus on data analysis
  - Evolution, not revolution in data management
  - Often have very specific needs
  - Visualization important
  - Profit from 'consulting'
- Interoperable AAI crucial
- Data privacy very challenging, both legally and technically
- Automatic workflows and metadata important
- Funding of data archival often unclear

# Summary and Outlook



- LSDMA's dual approach
  - DLCLs
  - DSIT
- Project's R&D driven by the communities' need → diverse activities
  - AAI interoperability
  - Automatic Metadata Extraction
  - Data Privacy
  - Speedup and simplification of analyses
  - ...
- Inclusion of activities into Helmholtz program-oriented funding in 2015, cross-program initiative
- Plans for additional Data Life Cycle Labs