CDCS: Center for Data and Computing Science

DESY Scientific Committee

Klaus Ehret

DESY, June 6th, 2018

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

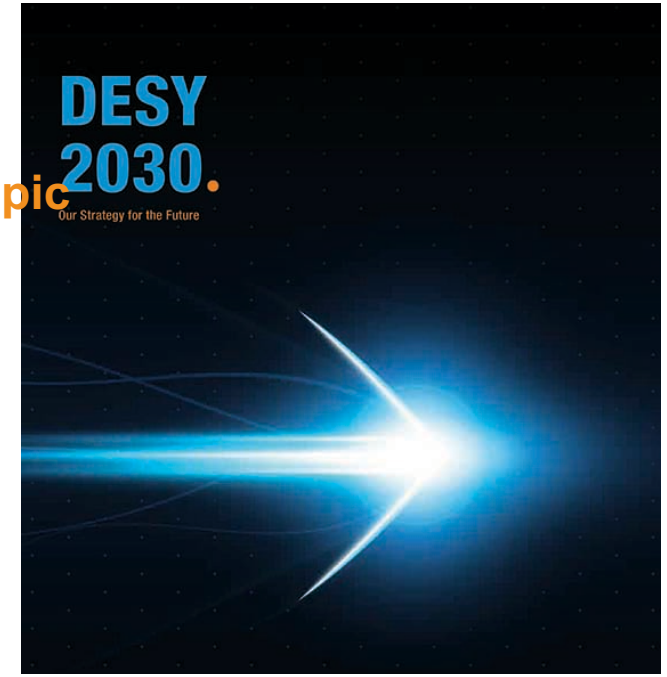
What is CDCS?

Variety of Expectations, Different View

- DESY 2030
- HGF:MT-DMA
- Hamburg
- Objectives
- Scheme
- Campus
- Project Structure
- …

CDCS & DESY 2030.

Interdisciplinary Data Science as New Research Topic



The key elements of DESY's strategy are:
…
The Centre for Data and Computing Science (CDCS) is being established at the Hamburg Campus, to meet the increasing demands made by data-intensive applications in research.
…

Central Features of CDCS (as described in DESY 2030)

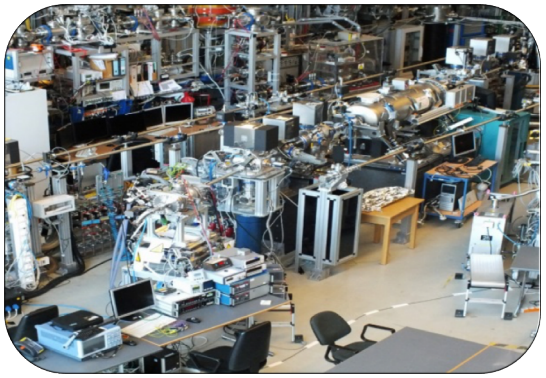
- **Interdisciplinary cooperation** across universities with **computing science** and applied mathematic departments
  - Create and strenghten necessary new competences
- Establish a **Data Science Graduate School**
  - Focus on knowledge an education („Brainware")
- Interdisciplinary R&D Center for **Applied** Information and Data Science
  - Focus on research for use cases / applications out of the Domain Science
  - Key Ressource on Campus, Science Park
- Support for **designing** efficient IT Infrastructure
  - But not the operation of IT Infrastructure

Information as Pillars of Excellent Science @ DESY.

Research Field Matter / Programme Matter & Technologie / Campus Bahrenfeld

Accelerators

MT ARD



Detectors

MT DTS

Information

MT DMA

Data Management and Analysis new Topic in Matter and Technology:

- Interdisciplinary Approach and Cross Divisional Activity
- Anchoring of CDCS activities in POF IV
- Foster and Secure Excellent Research
- Establish Scientific Computing and Data Science as Research Topic
- Hub for Knowledge and Expertise, „Brainware“

CDCS Status – A Look Back.

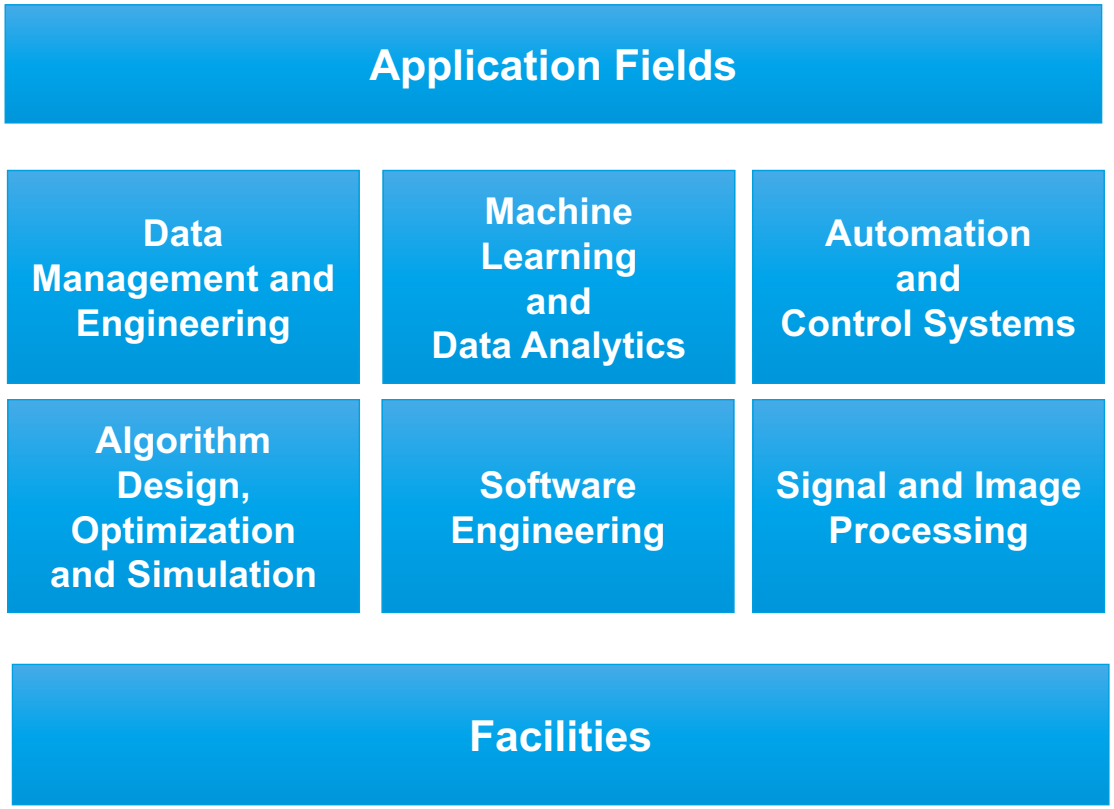
Flash on Various Activities in the Last Year

- ***DESY***: DIR, BR, Foundation Council, Round Table, Exchange and Discussions with Partners, CT-DMA
- ***Helmholtz***: High priority to „Information and Data Science“ activities (Incubator, new research field „Information“)
- ***Hamburg***: computer science plattform ahoi.digital established: data science – one pillar
- ***MINT Research Council***: Recommendations give high priority to „digitization“ and CDCS plans.

„Die Informatik an den Hochschulen ist weiterhin substantiell zu verstärken. Zudem wird empfohlen *cross-disciplinary Labs*, insbesondere ein einrichtungsübergreifendes ***Center for Data and Computing Science in Bahrenfeld***, zu etablieren und zu finanzieren.“

Data Science.

Excellence in Applications Requires Profound Data Science Knowledge



- > Identified six major Data Science topics (DASHH application)
- > Demand defines research focus of CDCS Labs, helpful structe

Applications

- Structural Biology
- Particle Physics
- Astroparticle Physics
- Material Science
- Accelerators
- Ultrafast X-ray Science

Data Science Topics

- Data Management and Engineering
- Machine Learning and Data Analytics
- Signal and Image Processing
- Algorithm Design, Optimization and Simulation
- Software Engineering
- Automation and Control Systems

Facilities

- > PETRA, FLASH
- > European XFEL
- > Interdisciplinary Data Analysis Facility

DMA: A new Topic in Matter and Technology.

DMA Measures at DESY – CDCS as Umbrella for LKI Research

- DESY DMA activities defines major DESY contributions to CDCS
  - Anchor in POF IV, Financial Planing
- Plan for three new research groups
  1. Research on Meta Data, Data Engineering and HPC
  2. Research for Imaging / AI
  3. Research on Control Systems (Machine, Experiment)

Data Management and Engineering	Machine Learning and Data Analytics	Automation and Control Systems

- CDCS / DMA arrange data science activities at DESY for interdisciplinary research cooperation with external partners

  - ***DMA***: underlines the role of *"Matter" inside Helmholtz* and beyond for scientific computing - DESY maintain key position in matter -> also in DMA@MT

  - ***CDCS***: interdisciplinary data science as research topic with *partners from regional universities* – DESY a strong partner and key player in MINT@ Metropolregion HH

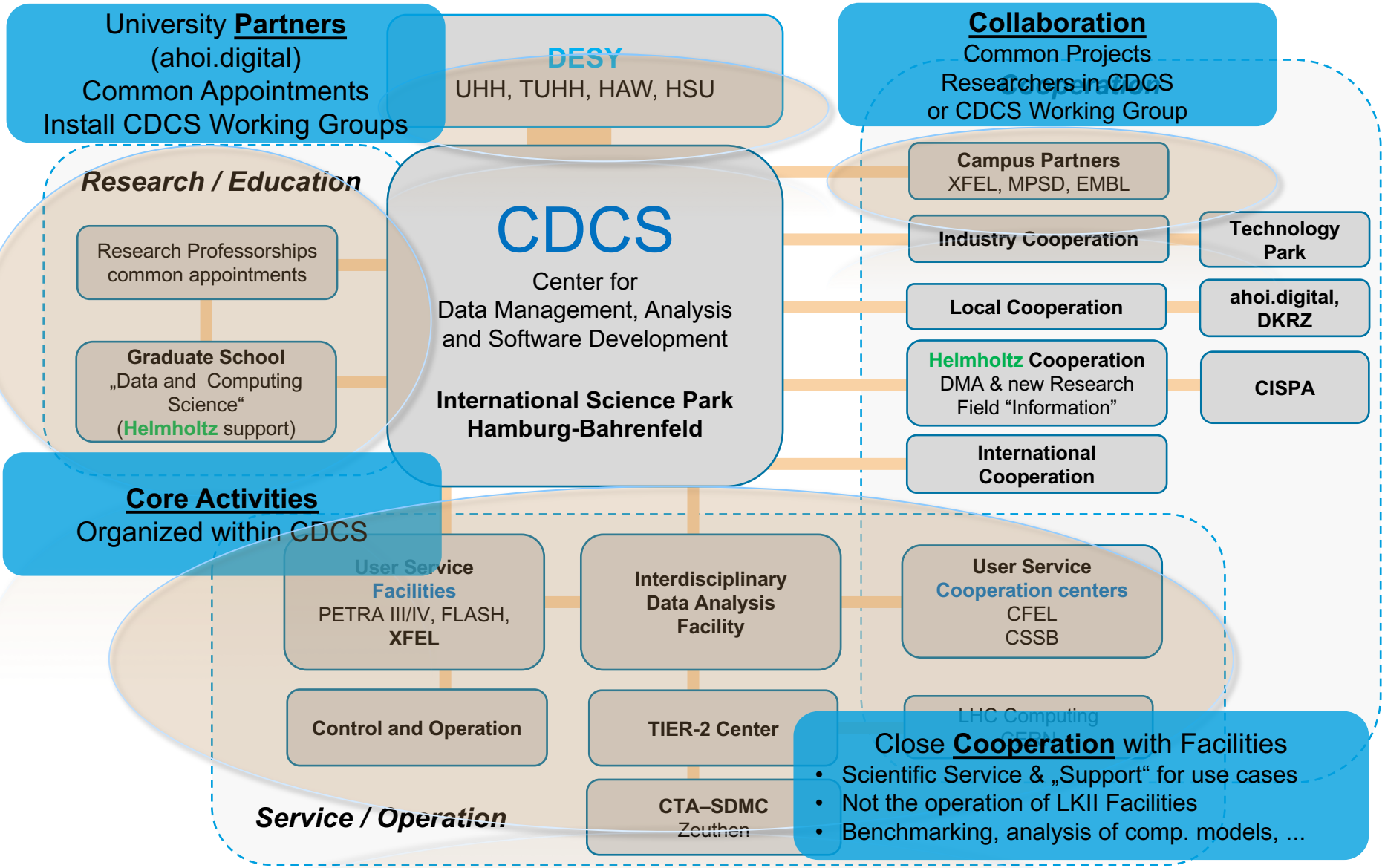
CDCS Objectives.

Secure Excellence by Innovative Interdisciplinary Collaboration

- Center for Data and Computing Science: close *cooperation* of „Domain Science“ (basic research on campus) *with computer science and applied mathematics*
- Utilize *competences, knowledge and methods* of advanced computer science, to solve „big data“ challenges
  - at DESY facilities (PETRA III, XFEL, IDAF, …)
  - in the domain science application (structural biology, particle physics, ultrafast X-ray science or material research).
- Information and Data Science as *research topic*
- Focus on knowledge and expertise (*Soft- and Brainware*, „Data Scientists“)
  - not the operation of a „computing center“, but advice for scientific computing infrastructure
- Setup of interdisciplinary *Data Science Research Schools*: structuring and identity forming for CDCS; (applied DASHH in HH, HEIBRIDS in Zeuthen selecting PhD stud.)
- Close cooperation with *MINT Partner on Campus Bahrenfeld* and HH computer science platform *ahoi.digital*

Proposal: Center for Data and Computing Science.

Cross Disciplinary Data Science Research Activity



CDCS: Key Resource on Campus Bahrenfeld.

Information and Data Science: Central Foundation for Excellent Science

- Scientific challenges of CDCS affect all parts of the research campus Bahrenfeld.
- An inspiring place for students and researchers.
- Think tank for innovative and disruptive ideas.
- Builds on existing strengths and unique features.
- Building on Trabrennbahn area, approx. 3.000 m^2

- 6 labs or working groups / 100 people
- State-of-the-art labs
- Student lab: information technology
- Start as virtual center, realization until 2025

- CSSB like government model anticipated:

- People stay employed at their home institut
- Scientific directorate, office, steering board and scientific advisory group

CDCS Science Case.

Examples of demands, challenges and opportunities

- First CDCS ideas - 2017
  - There are substantial computing needs in all DESYscience divisions.
  - These needs refer to both research-related topics (LK-I) and large-scale facilities (LK-II).
  - Moreover, current developments in photon science are leading to a paradigm shift in data science. This includes the need for rapid online data analysis.
  - In order to address the increased needs in scientific computing, we would like to combine resources and expertise from all DESY science divisions.
- ***DASHH Proposal***: 15 highlighted interdisciplinary project proposals out of more than 40 project ideas
- ***Application for a HYIG@M, together with TUHH***

  Data-Science for Diagnosis and Control of Large Complex Systems at the Example of Accelerators
  - ***Prototype CDCS activity***: strong link to research on campus, obvious demand and profit, new / increase of competences, interdisciplinary research (informatics), scientific service
  - Hopefully: ***first CDCS appointment*** in 2018

Center for Data and Computing Science - CDCS.

Partners, plans and next steps

- Partner - large interest and support plans: *DESY provides excellent environment*, competences and very *interesting use cases* for data science
- Prepare „*Round Table II*“ with potential partners
  - Setup a project structure with partners / LoIs until mid 2019
- Common Project Structure (Sc. comp, DASHH, DMA & ext. Partners)
  - DESY internal „*Scientific Computing Board*“: strategic recommendations to directorate
  - *Scientific review* - members from PRC, PSC, MAC and further experts: External data science advisory committee:
- Appointment of LS / Prof. (3 new research groups)
  - *Appointment of key professorship* -> major step, scientific representative of „Data Science@DESY“, director of virtual „CDCS“
- Pilot activities: data science workshop, school student lab
- DASHH Graduate school: assembly in autumn 2018

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

TUHH
Technische Universität Hamburg

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

European XFEL

Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

mpsd
Max-Planck-Institut für Struktur und Dynamik der Materie

Conclusions.

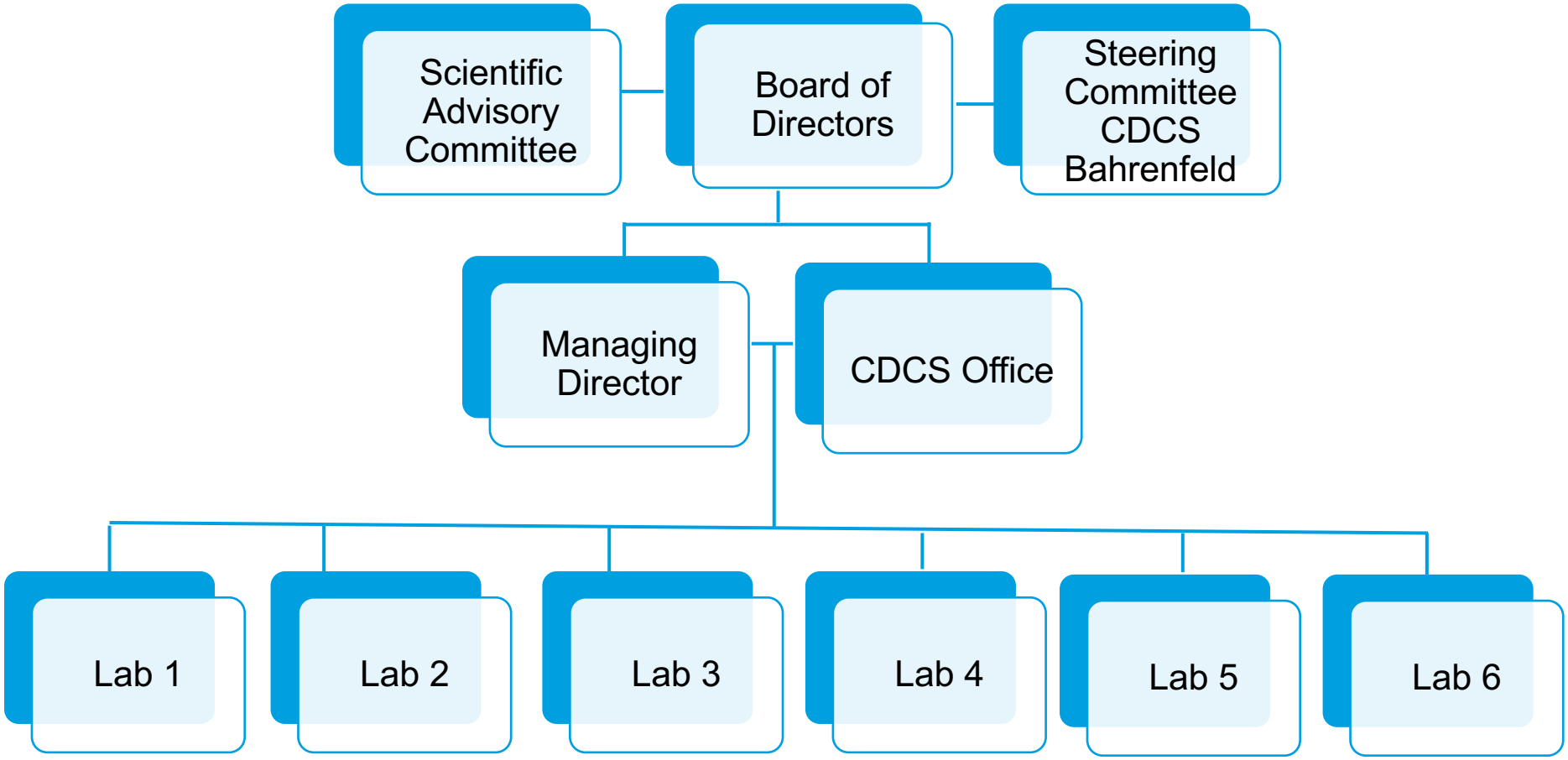
Scientific Computing & Data Science as Key Competence for DESY

- Essential to maintain top level research at DESY
- CDCS / DMA initiates a new research field
  - Complementary pillars of DESY 2030 strategy for the enormous demands in scientific computing and data science
- Built on the existing broad expertise at DESY with Big Data to solve the enormous Data Science challenges
- CDCS is key element of the DESY Strategy
- DASHH Application demonstrates large interest and high potential
- CDCS@2027 (end of PoV IV period):
  - CDCS building inaugurated, > 6 Labs
  - CDCS a renowned international center for data science in basic research
  - Prolonged graduate school
  - Established Collaborative Research Center on Data Science (DFG-SFB)

Back Up Slides

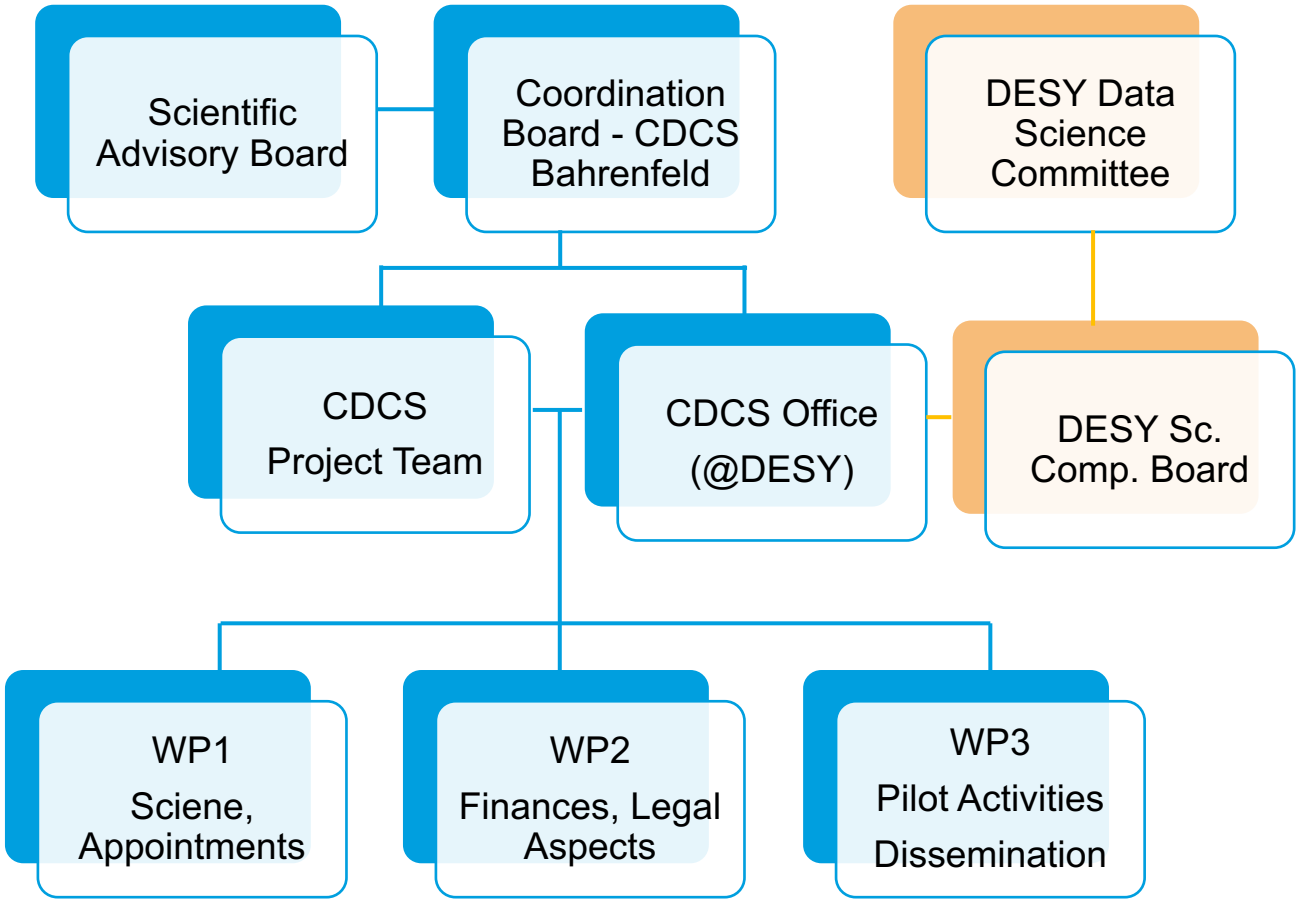
CDCS Government Structure - Proposal

CSSB like Model, Aligned with DASHH



CDCS Project Structure – Draft Version

With Partners, aligned with DASHH requirements and DESY internal setup



Leading scientist nominated by potential partners

V, Bau, V3, V5
BWFG, Campus

Scientist nominated by potential partners

CDCS Financial Planing

- Construction (building incl. equipment, Labs etc.): approx.12 Mio. €.
- Annual costs: personal and operation: approx. 9 Mio. €:
  - 50 FTE - 100.000 €/a: 5 Mio.€
  - 20 PhD Students - 50.000 €/a: 1 Mio.€,
  - Operation / Infrastructure: 1,5 Mio. €,
  - Investments: 1,5 Mio. €.
- Financing of CDCS with partners; rough sketch of contributions
  - 1/3 from DESY : ram up until 2026
    - 0,9 Mio. € - three research groups
    - 1,1 Mio. € - DASHH
    - Additional: DMA, PoF, …
  - 1/3: UHH, TUHH, HAW
  - 1/3: research institutions, third party money, other universities and partner from industry
- Seed funding: FHH and Bund
- Project Funds

Data Science Workshop

First Ideas -> to stimulate discussions, find interested people to develop this further

What is the aim, benefit – why should we do this?

- Pilot activity: beneficial for presentation of DASHH proposal
- Help to sharpen use cases and opportunities for CDCS
- Prefered configuration: interdisciplinary approach with DASHH partners and potential regional partners for CDCS-Campus Bahrenfeld
- HUB, platform, exchange -> learn from othe communities, understand their plans, potential and problems (3 P‘s)
- create new ideas

CDCS – an Interdisciplinary Research Center.

Establish a world leading data science and research location in Hamburh

a. Expertise and methods of computer science

b. Challenges of data-intensive scientific computing

- are combined in CDCS:

> Development and use of state-of-the-art technologies for processing and analyzing enormous amounts of data

> create innovative solutions

- in order to exploit the full scientific potential of the world‘s leading large-scale facilities at DESY and XFEL

> Think tank for innovative and disruptive ideas.

University of Hamburg: Department of Informatics

Degrees related to Computer Science (2017)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DEPARTMENT OF
INFORMATICS

DEGREE PROGRAMS

Bachelor

- Computer Science
- Business Informatics
- Software Systems Development
- Human-Computer-Interaction
- Computing in Science
- Education/Teaching

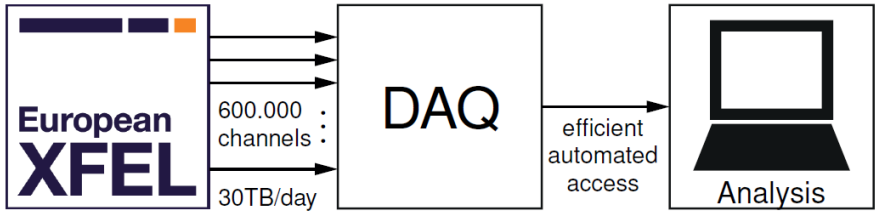
Master

- Computer Science
- Business Informatics
- IT Management & Consulting
- Intelligent Adaptive Systems (E)
- Bioinformatics
- Education/Teaching

WP1: Data Mining

Goal: identify key features and **critical components** for **reliability, availability and performance**

- **data-management**: efficient automatic interface with DAQ
- **pre-processing**: classification to check labeling, configuration and quality, data cleaning
- **information extraction** (prescriptive): clustering, association rule discovery (dependencies), sequence discovery
- **validation**: based on expert knowledge (and validation data)

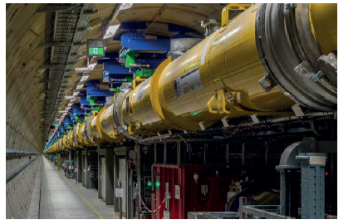


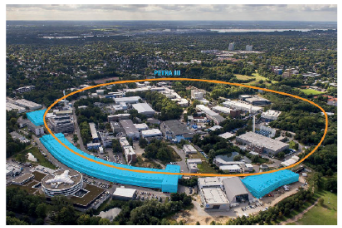
WP2: Feedback Control

Goal: improve performance by **advanced** and **distributed feedback concepts**

recent developments in embedded computation allow for:

- **advanced feedback control concepts:** optimization based control, model predictive control, adaptive or learning-based control
  example: FEL longitudinal feedback control
- **network control concepts:** distributed controller of interconnected subsystems
  example: PETRA IV photon-beam monitoring system





www.desy.de

WP3: Fault Diagnosis

Goal: fault detection, isolation and identification

- **data-driven approaches**
  - classification/ clustering
  - Bayesian inference
  - neural networks
  - support vector machines
  - extended PCA, PLS
- **model-based approaches**
  - observer
  - parity space
  - parameter estimation

	data-driven	model-based
detection speed	++	+
a priori data	-	+
ease of deployment	+	-
large complex	++	-
physical insight	–	++
adaptability	-	++

[Tidriri et al., 2016]

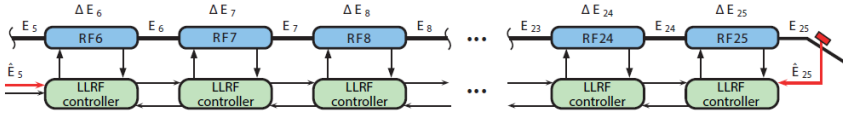
⇒ **hybrid approach**

first steps, PhD thesis MSK [Nawaz et al., 2018]

WP4: Supervisory Control

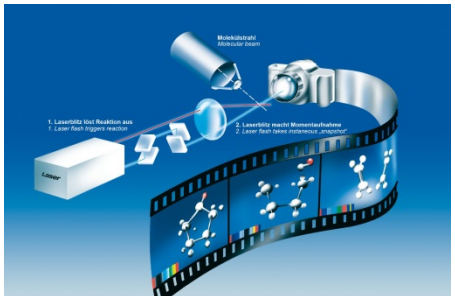
Goal: improve availability and reliability by **fault tolerance and predictive maintenance**

- **supervision:** predictive maintenance (early detection of anomalies), safe operation
- **management control layer:** scheduling, set overall specifications, hybrid system
- **optimization and coordination:** optimize interaction within subsystems, exploit redundancies for fault tolerance



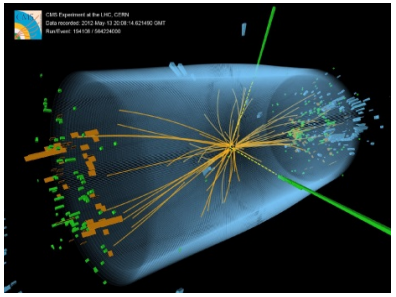
Data Challenges at DESY.

More and More Complex Data



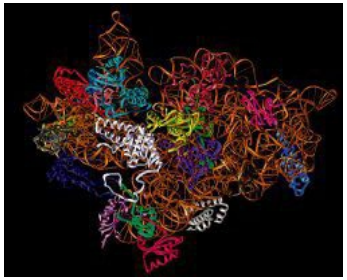
Recording of molecular dynamics
→ **very fast & high troughput data processing**

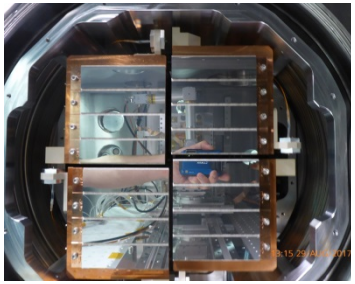
Operation of complex accellerators
→ **large scale supervisory control systems with > 1 Mio. channels**

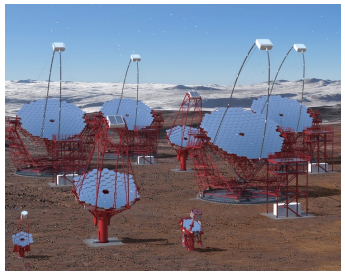


LHC Experiments: **hundreds of PByte highly complex data per year**

Simulation of very complex structures
→ essential for interpretation of measured data







Astroparticle Experiments
DESY as center for scientific data management

new detector technologies and powerful accelerators

→ **fast data acquisition (HW&SW) and processing of very large data sets.**