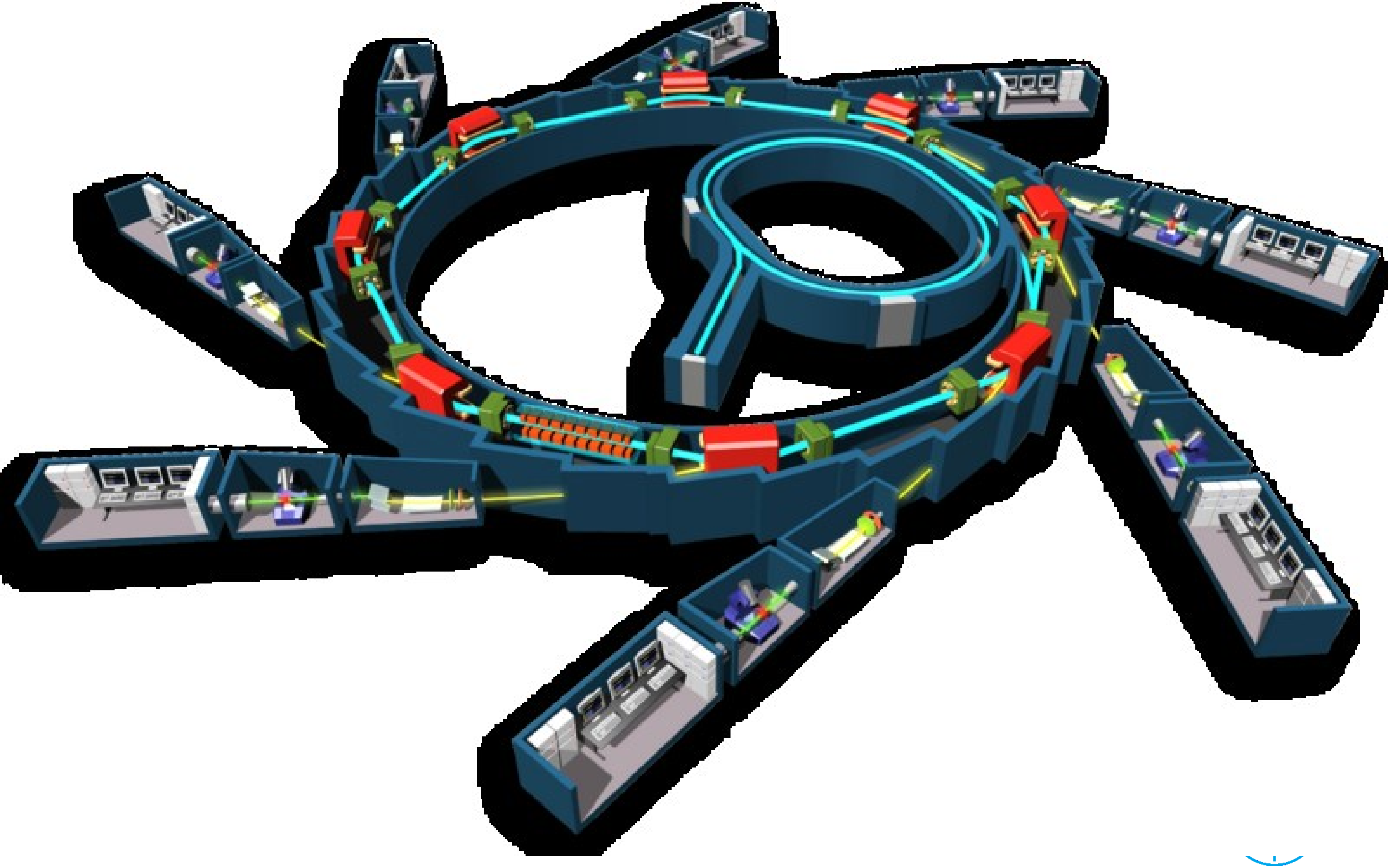
Big Data

For large scale facilities

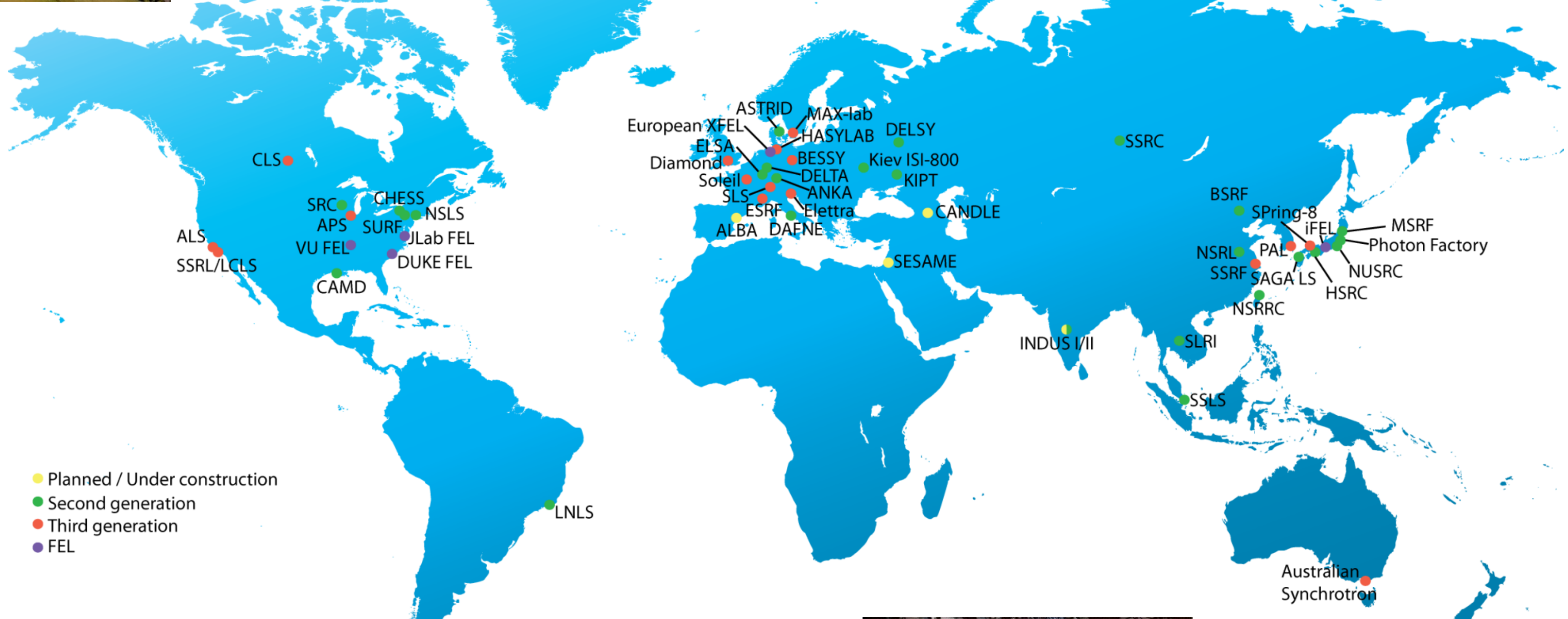
Volker Guelzow
DESY

Cremlin Workshop
Moscow, Feb. 15th, 2017

Synchrotrons Shedding New Light onto Sciences

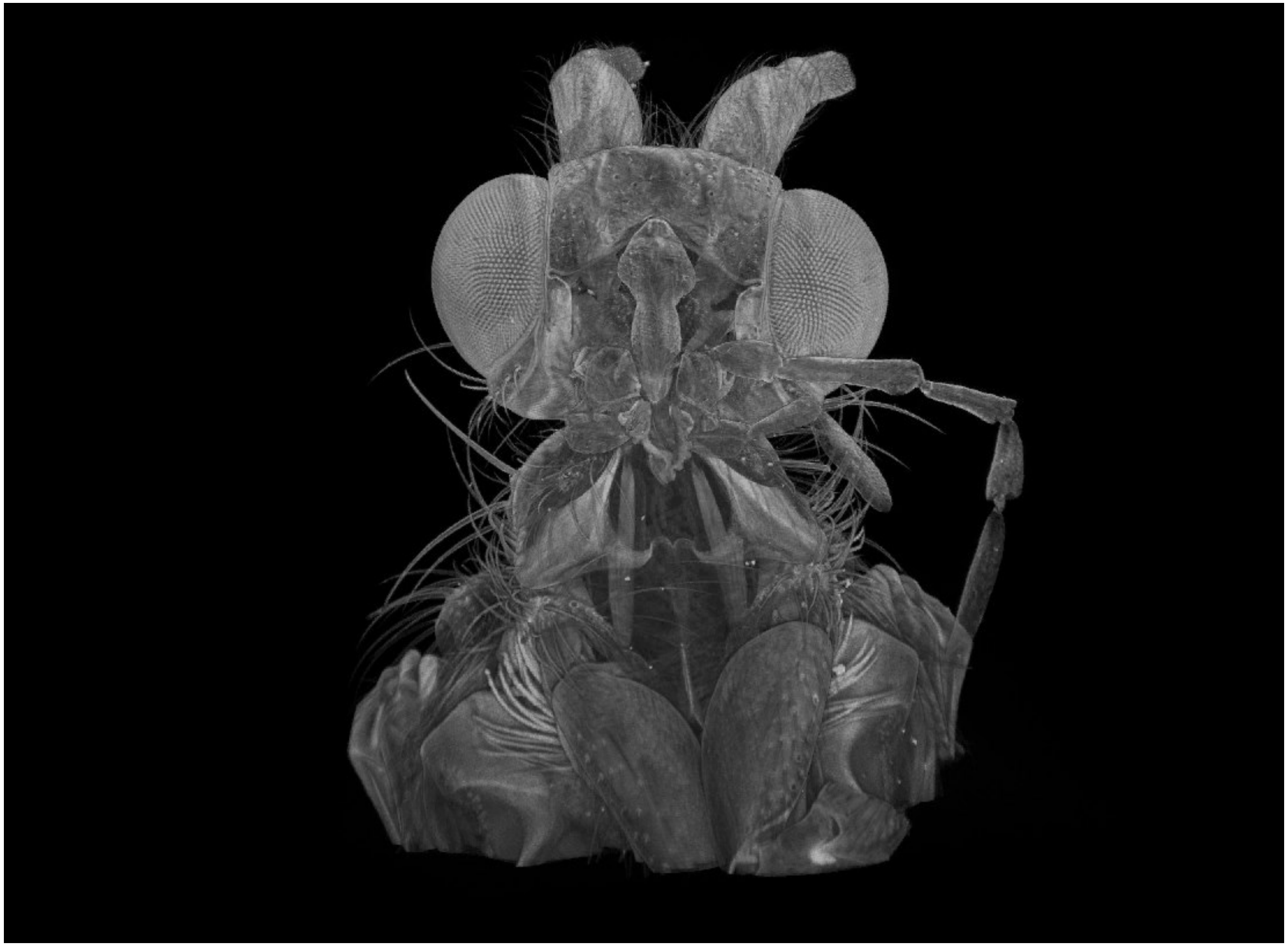


Synchrotrons Shedding New Light onto Sciences



- 🟨 Planned / Under construction
- 🟩 Second generation
- 🟥 Third generation
- 🟪 FEL

Köcherfliege (*Limnephilus flacivornis*) Kopf + Thorax

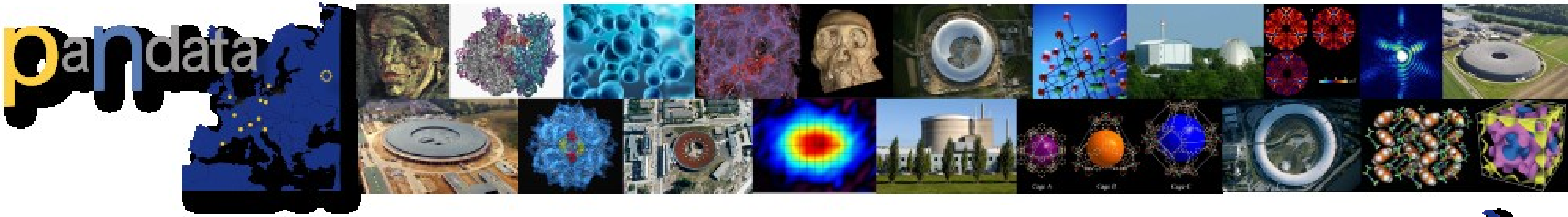


Courtesy:
Dr. F. Beckmann

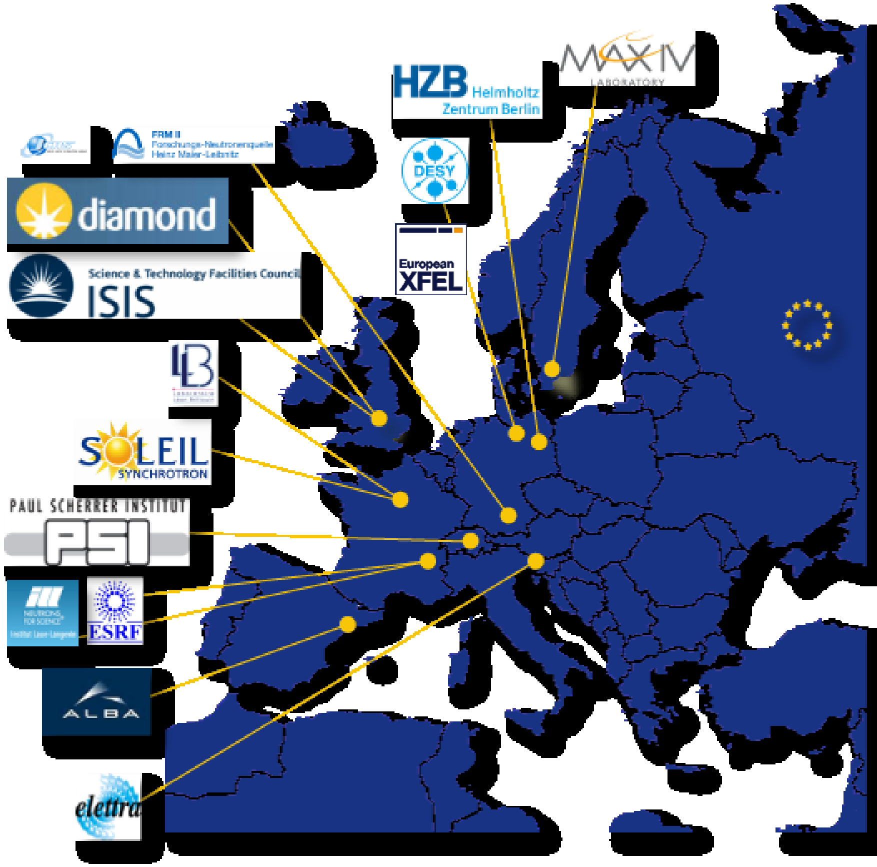
Helmholtz-Zentrum
Geesthacht
Centre for Materials and Coastal Research

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

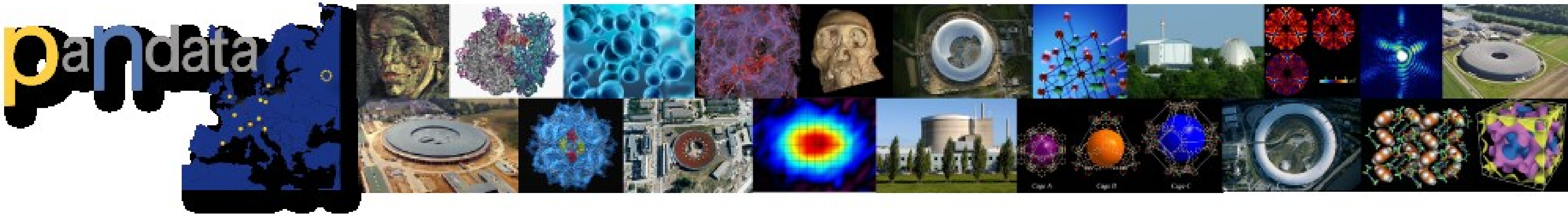
Strength in Numbers



- Photon and Neutron Data Infrastructure
- FP 7 Project established in 2007 starting with 4 facilities
- Combined Number of Unique Users more than 35000 in 2011
- Combines Scientific and IT staff from the collaborating facilities


HZB Helmholtz Zentrum Berlin
MAX IV LABORATORY
FRM II Forschungs-Neutronenquelle Heinz Maier-Leibnitz
DESY
diamond
Science & Technology Facilities Council ISIS
European XFEL
SOLEIL SYNCHROTRON
PAUL SCHERRER INSTITUT PSI
ESRF
ALBA
elettra

Aims of the Pandata Project



- Harmonize authentication and authorization - Shibbolleth
- Standardize data formats NeXus/HDF5 and annotation of data
- Allow transparent and secure remote access to data
- Establish sustainable and compatible distributed data catalogues
- Allow long term preservation of data
- Provide compatible data analysis software
- Promote data policies in laboratories

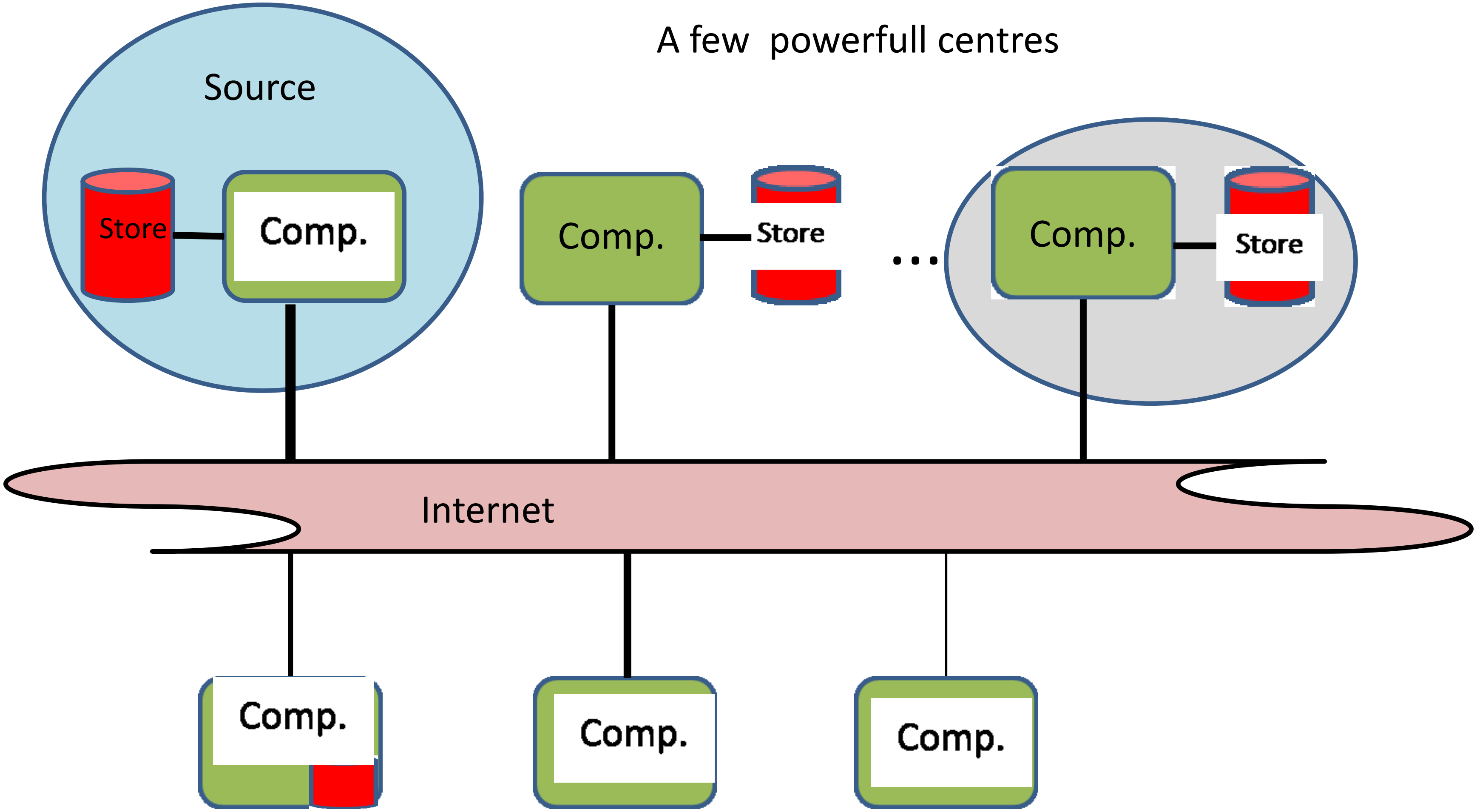
Some important topics:

- Data ingest: -> get the speed!
- „online control“ -> quick first analysis
- Find your data: -> Datamanagementsoftware, metadata
- Find your data in due time: -> professional software (+HW)
- Sophisticated analysis methods -> collaboration with others
- Analysis facilities: -> today Cloud style, commercial provider?
- Visualization:
- Portals, AAI:
- Networks!
- Don‘t forget about control systems

Topics to work with whom?

- Cross centre/ transnational access
  - Cloud initiatives -> European Open Science Cloud
  - Common Services, Interoperability, data management -> EGI, EOSC, INDIGO data cloud
- Cross communities
  - Metadata handling, open access -> RDA, EGI, EOSC, OpenAIRE
- Networks -> GEANT/NRENS
- Analysis facilities
  - Professional SW development, methods, visualization -> our competence, „Google",.
- Digitalization of our infrastructures
  - Control software-> cooperation with industry

A possible computing model for large facilities



Smaller centres, commercial providers

A key Initiative by the European Commision

- The European Open Science Cloud

  and follow ups

Why Europe is not fully tapping into the potential of data:

- **Data not always open** and **lack of incentives and rewards** for data sharing

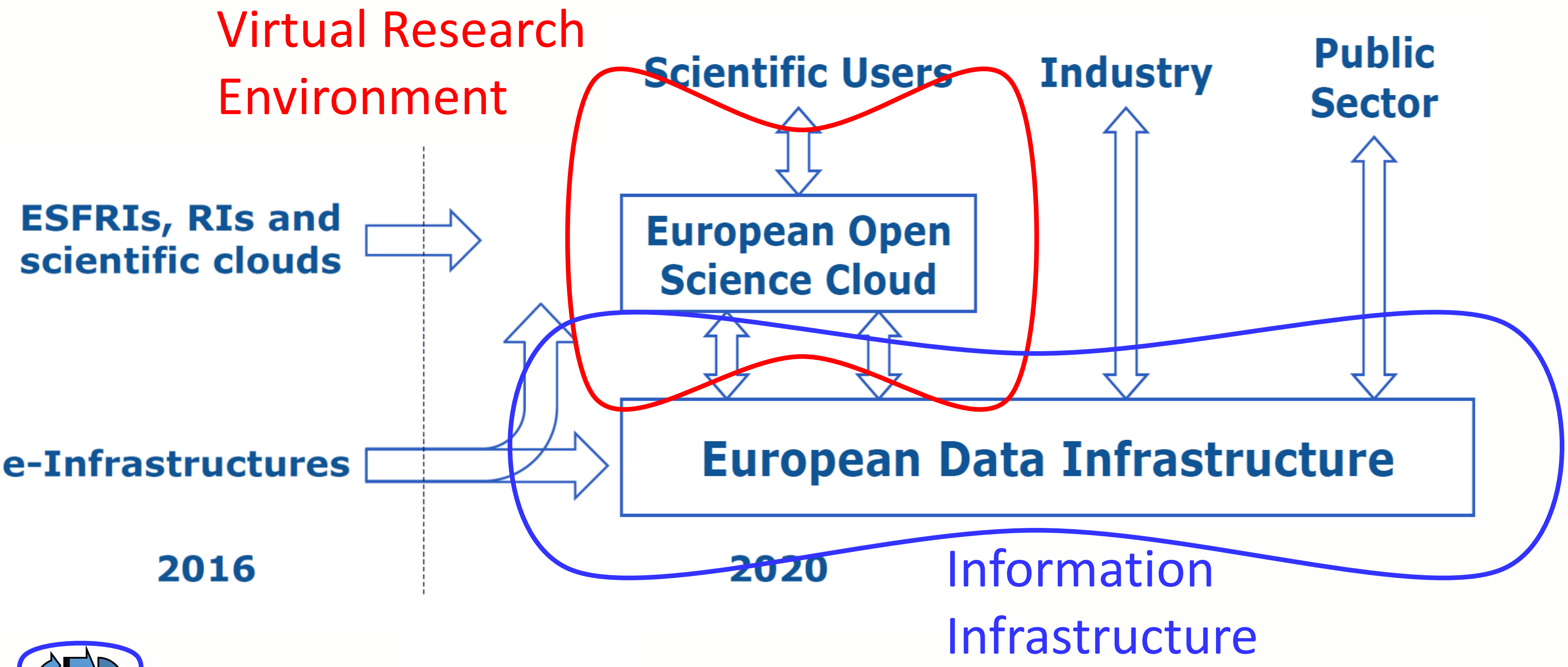
- **Lack of interoperability** required for data sharing … noting deep-rooted walls between disciplines.

- **Fragmentation between data infrastructures** that are split by scientific and economic domains, countries and governance models

- Surging demand for **High Performance Computing** at a scale above single member state resources

- **Data reuse employing advance analysis techniques** adequate protection of personal data considering forthcoming revision of Copyright legislation.

Evolution of infrastructure


Virtual Research Environment
Scientific Users
Industry
Public Sector
ESFRIs, RIs and scientific clouds
European Open Science Cloud
e-Infrastructures
European Data Infrastructure
2016
2020
Information Infrastructure

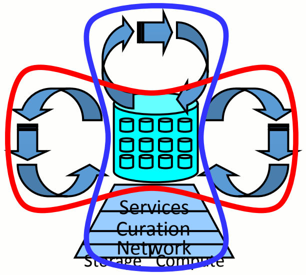


Diagram: Augusto BURGUEÑO ARJONA

EOSCpilot: High Level Aims

The *EOSCpilot* project will support the first phase in the development of the EOSC. It will

- **Establish the governance framework** for the EOSC and contribute to the development of European open science policy and best practice;
- **Develop a number of demonstrators** functioning as high-profile pilots that integrate services and infrastructures to show interoperability and its benefits in a number of scientific domains;
- **Engage with a broad range of stakeholders**, crossing borders and communities, to build the trust and skills required for adoption of an open approach to scientific research.

EOSCpilot Challenges

Three types of challenges addressed by the EOSCpilot:

Scientific Challenges are really *Opportunities*

Scientific Challenges: deploying the EOSC to deliver Open Science

Technical Challenges are *Barriers to overcome*

Technical Challenges: developing technical solutions that meet the scientific needs

Cultural Challenges are also *Barriers*

Cultural Challenges: adopting new, more open ways of working

Context for cooperations:

- National Cooperation programs like Helmholtz – RSF JRG
- European Calls: Russia ist third party country, cooperations are welcome, see (status 2016):

http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020_localsupp_russia_en.pdf

- And the programs from Ministry of Education and Science of the Russian Federation
- www.rfbr.ru
- www.rfh.ru
- www.rscf.ru
- www.fasie.ru

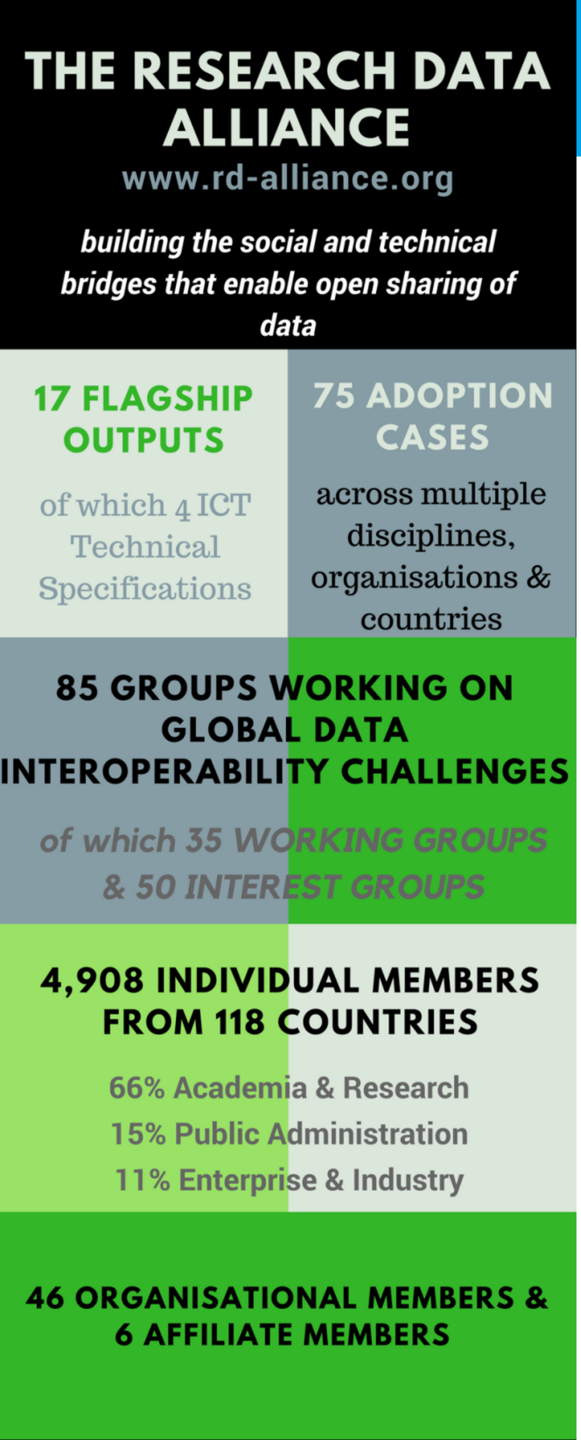
See: www.rda-alliance.org

What is RDA?

> RDA is an international **member based organization** focused on the development of infrastructure and community activities that reduce barriers to data sharing and exchange, and the acceleration of data driven innovation worldwide.

With more than 4,900 members globally representing 118 countries, RDA includes **researchers, scientists and data science professionals** working in multiple disciplines, domains and thematic fields and from different types of organisations across the globe.

RDA is building the social and technical bridges that enable open sharing of data to achieve its vision of researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.



THE RESEARCH DATA ALLIANCE

www.rd-alliance.org

building the social and technical bridges that enable open sharing of data

17 FLAGSHIP OUTPUTS

of which 4 ICT Technical Specifications

75 ADOPTION CASES

across multiple disciplines, organisations & countries

85 GROUPS WORKING ON GLOBAL DATA INTEROPERABILITY CHALLENGES

of which 35 WORKING GROUPS & 50 INTEREST GROUPS

4,908 INDIVIDUAL MEMBERS FROM 118 COUNTRIES

66% Academia & Research
15% Public Administration
11% Enterprise & Industry

46 ORGANISATIONAL MEMBERS & 6 AFFILIATE MEMBERS

See: www.rda-alliance.org

RESEARCH DATA ALLIANCE

Vision

Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.

Mission

RDA builds the **social and technical bridges** that **enable open sharing** of data.

RDA-member from the Russian Federation

Dr	Andrey	Ustyuzhanin	Yandex School of Data Analysis	Russian Federation
Mr	Petya	Kohts	kohts.com	Russian Federation
Dr	MARINE	MELKONYAN	THE NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY MISIS	Russian Federation
Mr	Nikolay	Skvortsov	IPI FRC CSC RAS	Russian Federation
Mr	Andrey	Shevel	National Research University of Information Technologies, Mechanics and Optics	Russian Federation
Prof	Teymur	Zulfugarzade	Plekhanov Russian University of Economics	Russian Federation
Mr	Vyacheslav	Popov	LabHUB.ru	Russian Federation

Summary

- Big data is NOT just volume
- … but data ingest is often a problem
- Big data at large facilities needs cooperation and openess
- The ease of use is a key feature
- Sophisticated data management SW is mandatory
- Distributed and federated computing models are the future
- Joint analysis SW development is needed
- Networks are a key element