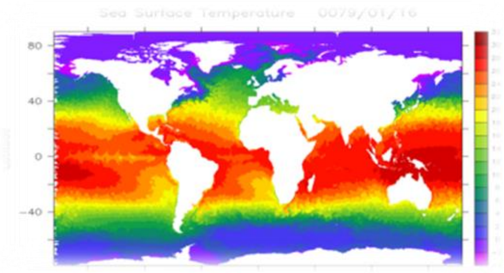
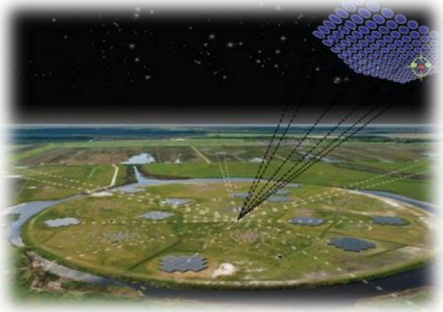
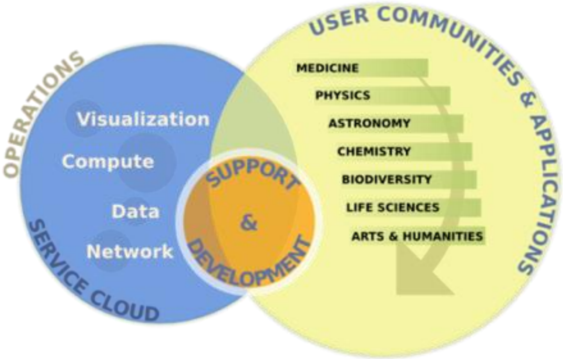
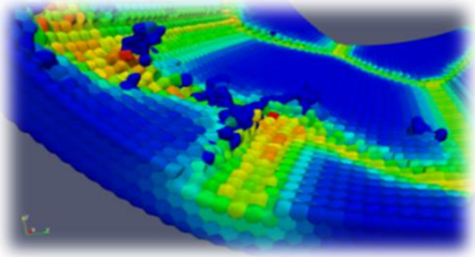
SURFsara Data Services

SUPPORTING DATA-INTENSIVE SCIENCES





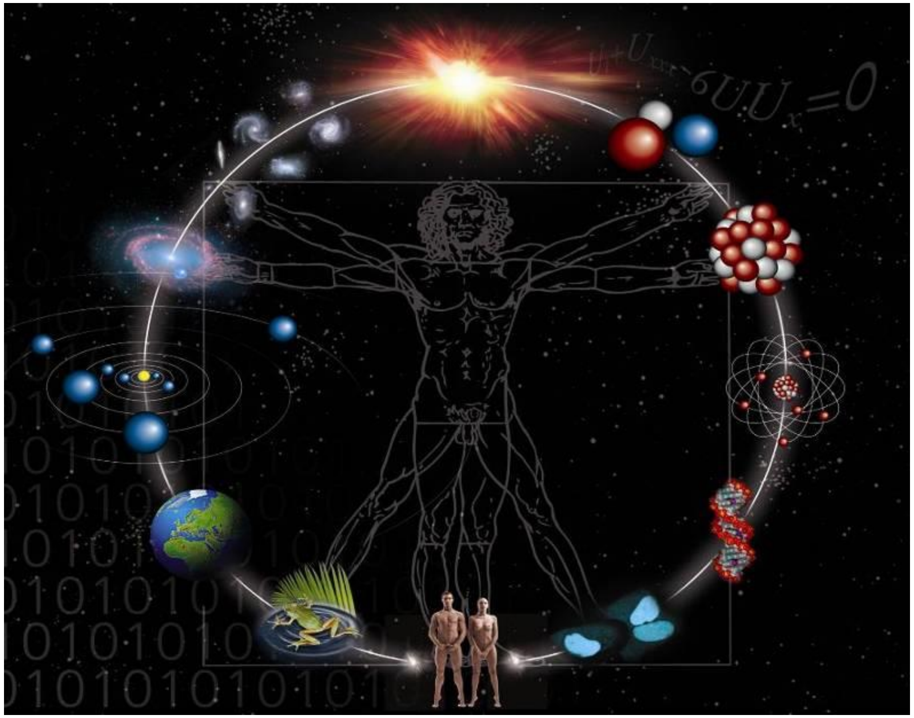




Mark van de Sanden

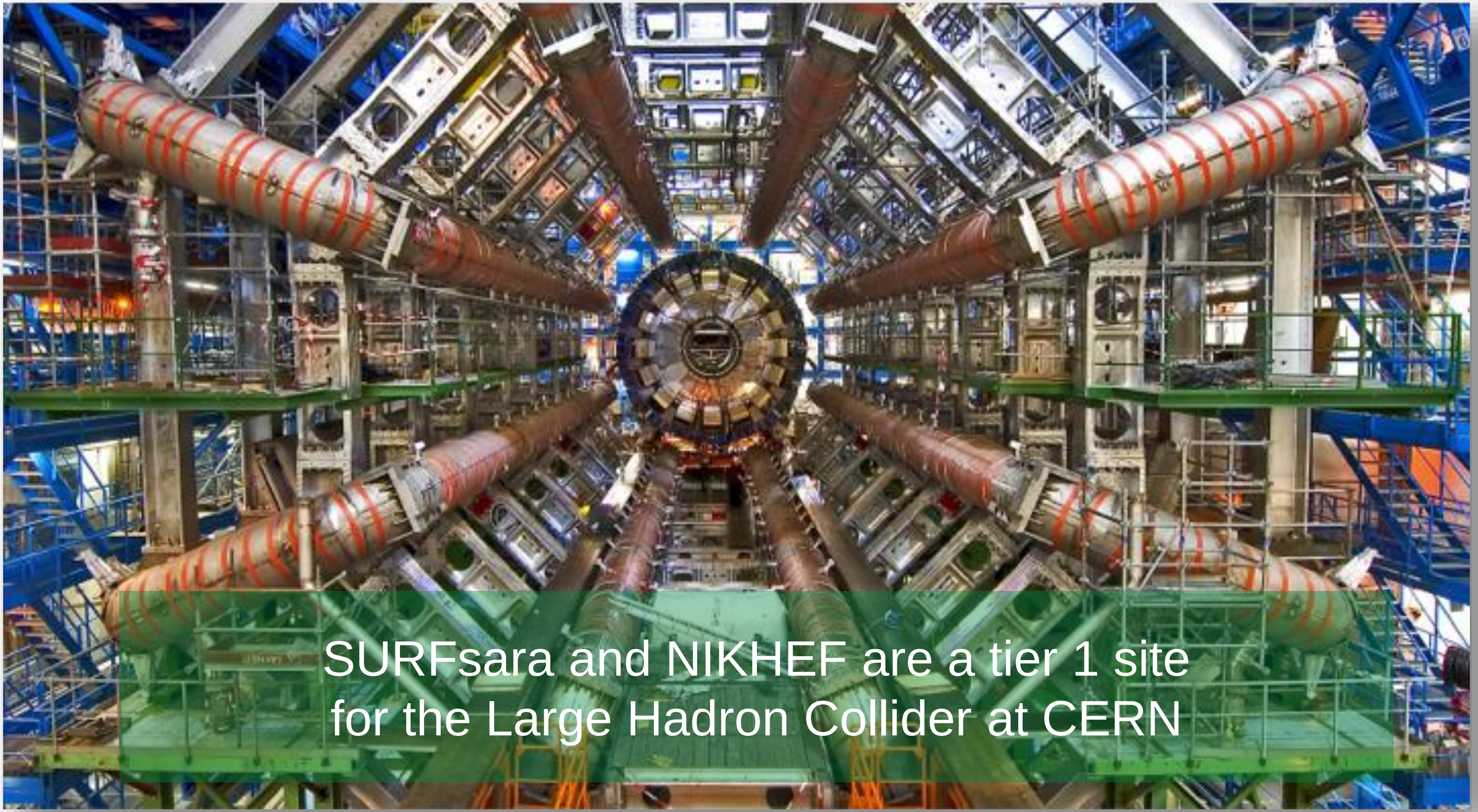
Dutch Scientific Challenges

- From High Energy Physics
- to atomic and molecular physics (DNA);
- Life sciences (cell biology);
- Human interaction (all human sciences from linguistics to even phobia studies);
- and from the big bang;
- to astronomy;
- science of the solar system;
- earth (climate and geophysics);
- into life and biodiversity.



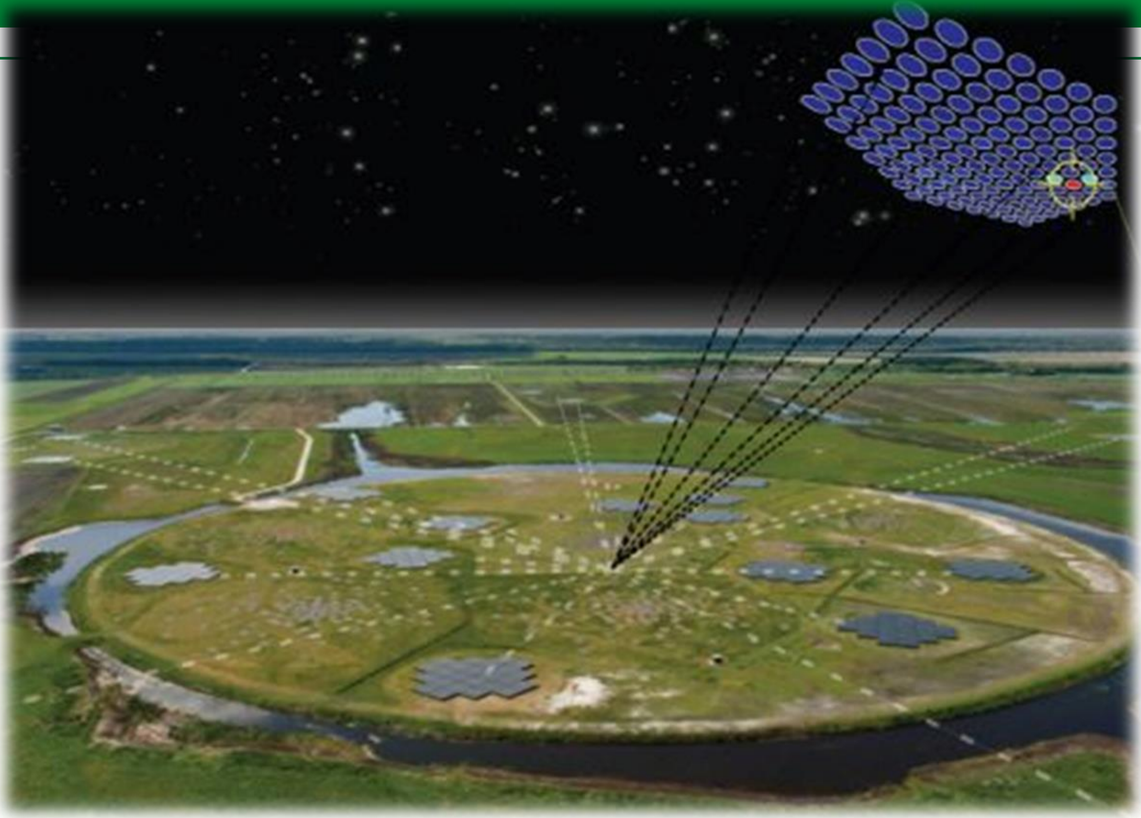
Picture courtesy of UvA FNWI

High Energy Physics, discovery of Higgs boson



SURFsara and NIKHEF are a tier 1 site for the Large Hadron Collider at CERN

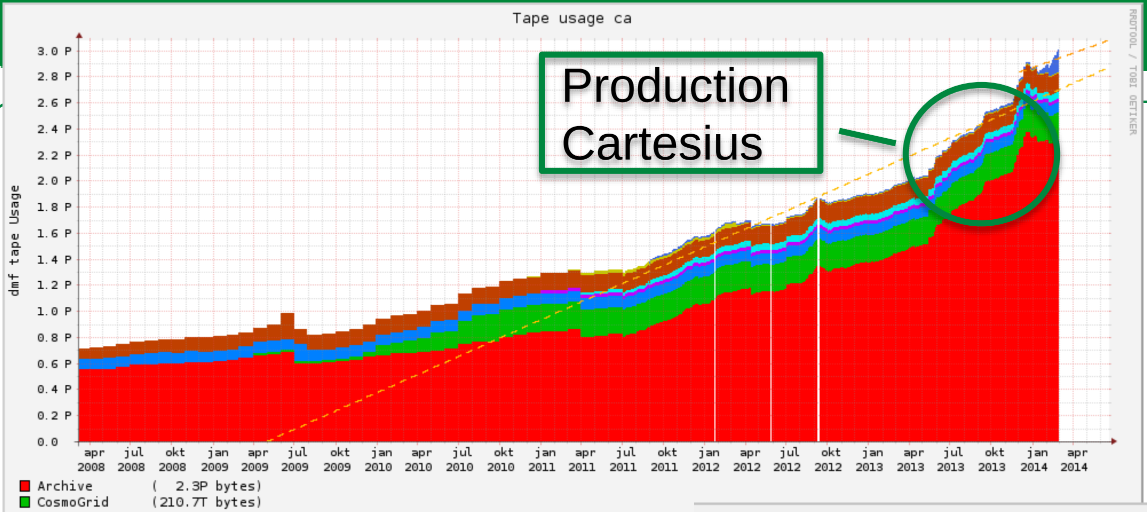
LOFAR, discovery of two new pulsars



Using computing resources provided by SURFsara, which is part of the Dutch and European Grid Infrastructure, Coenen and the team needed only a month to search through a set of 2010-2013 LOFAR images that would have occupied a single computer for more than a century.

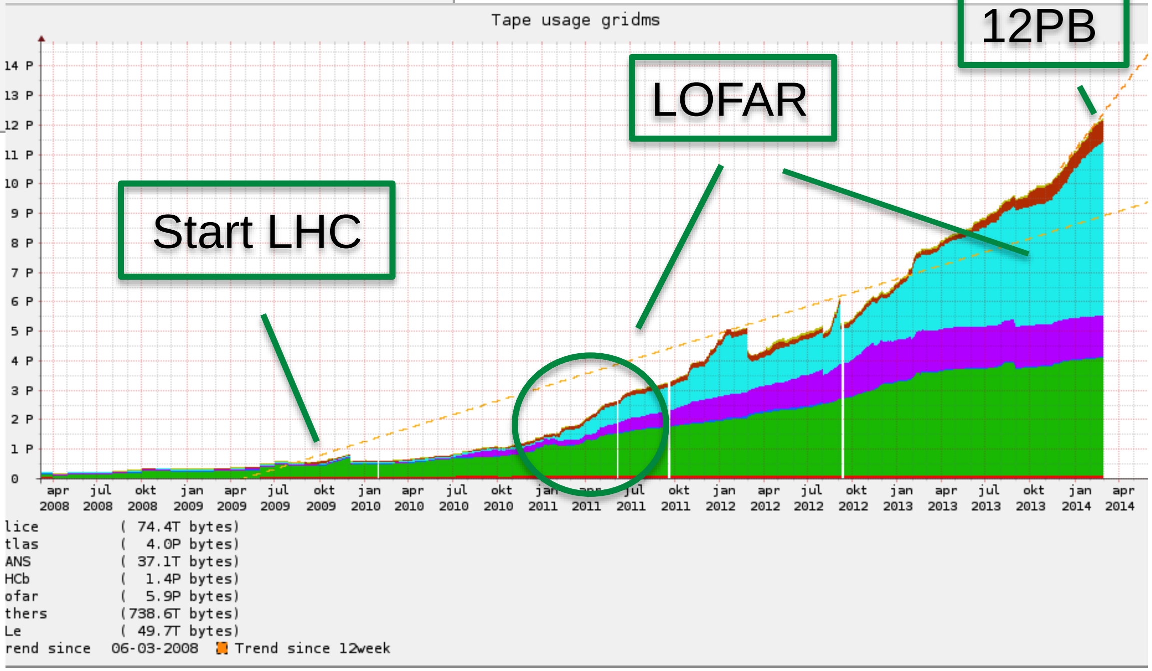
THE DATA EXPLOSION

Projects	Volume
COSMO GRID (2009)	105TB
RUMC (2013)	90TB
OWLS (2006)	70TB
ESSENCE (2006)	41TB
ENTRAIN (2011)	26TB
ITAMOC (2011)	25TB
EAGLE (2013)	2,5TB (600TB)
TITAN (2014)	(1PB)
HPC ARCHIVE (1993)	1172TB



Production Cartesius

Community	Volume
LOFAR	5,8PB
LHC/ATLAS	3.9PB
LHC/LHCb	1,4PB
BBMRI	115TB
MOLEPI	78TB
LHC/ALICE	74TB
ILDG	43TB
DANS	18TB
OTHERS	50TB



12PB

LOFAR

Start LHC

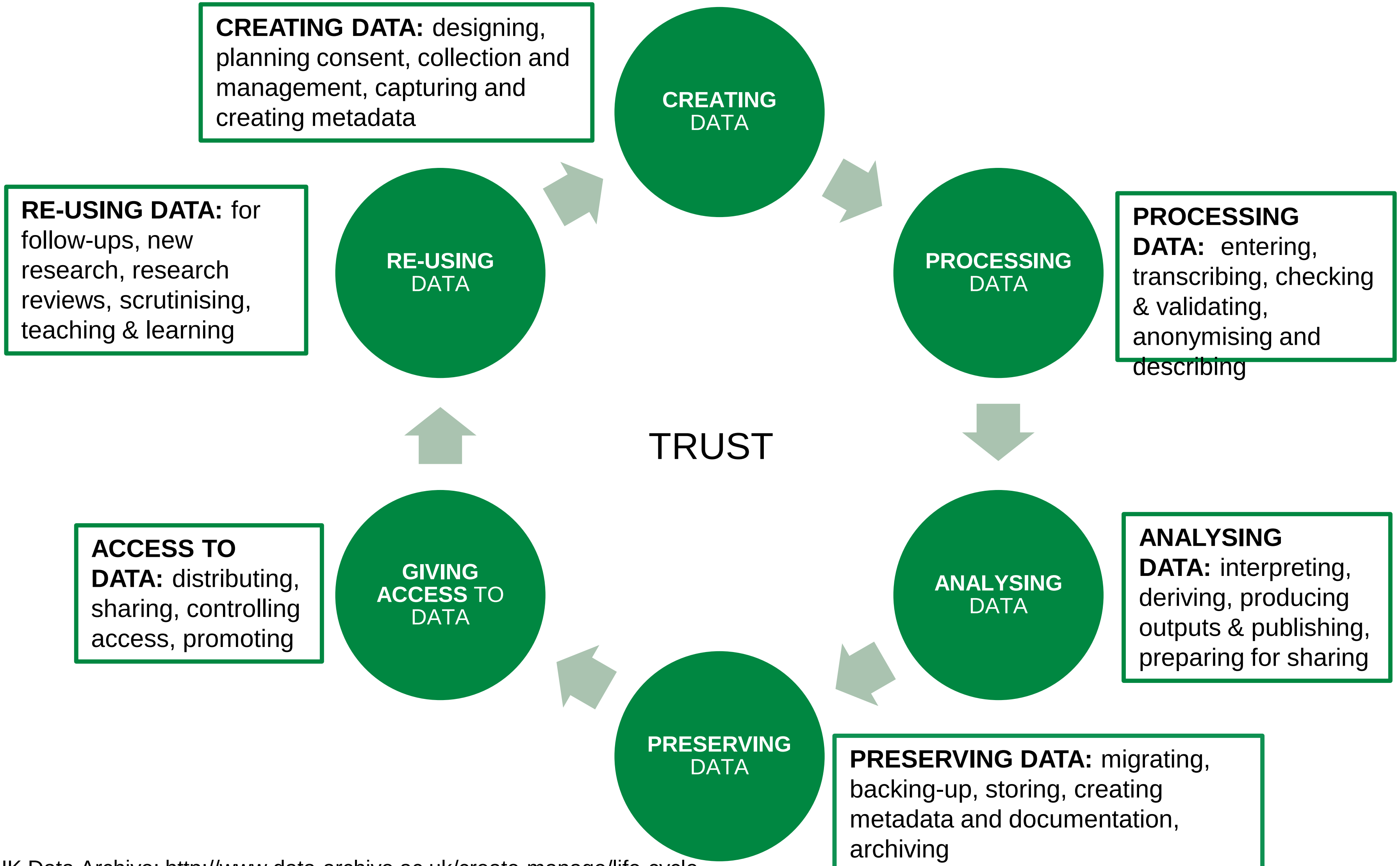
The world of the many

- **Many different users** (well organised (international) user communities, research groups, universities, research institutes, individual researchers, etc.)

- **Many different 'use cases'** (central data, distributed data, small files, large files, static data, dynamic data, etc.)

- **Many different user requirements** (storage, meta data, data searching, persistent identifiers, long-term preservation, privacy, visualisation, data analytics, data sharing, data re-use, semantic annotation, work flows etc. etc. etc. etc.)

- **Many different V's** of Big Data (varieties, volumes, velocities, veracities, validities, volatilities)



CREATING DATA: designing, planning consent, collection and management, capturing and creating metadata

CREATING DATA

PROCESSING DATA

PROCESSING DATA: entering, transcribing, checking & validating, anonymising and describing

RE-USING DATA

RE-USING DATA: for follow-ups, new research, research reviews, scrutinising, teaching & learning

TRUST

ANALYSING DATA

ANALYSING DATA: interpreting, deriving, producing outputs & publishing, preparing for sharing

GIVING ACCESS TO DATA

ACCESS TO DATA: distributing, sharing, controlling access, promoting

PRESERVING DATA

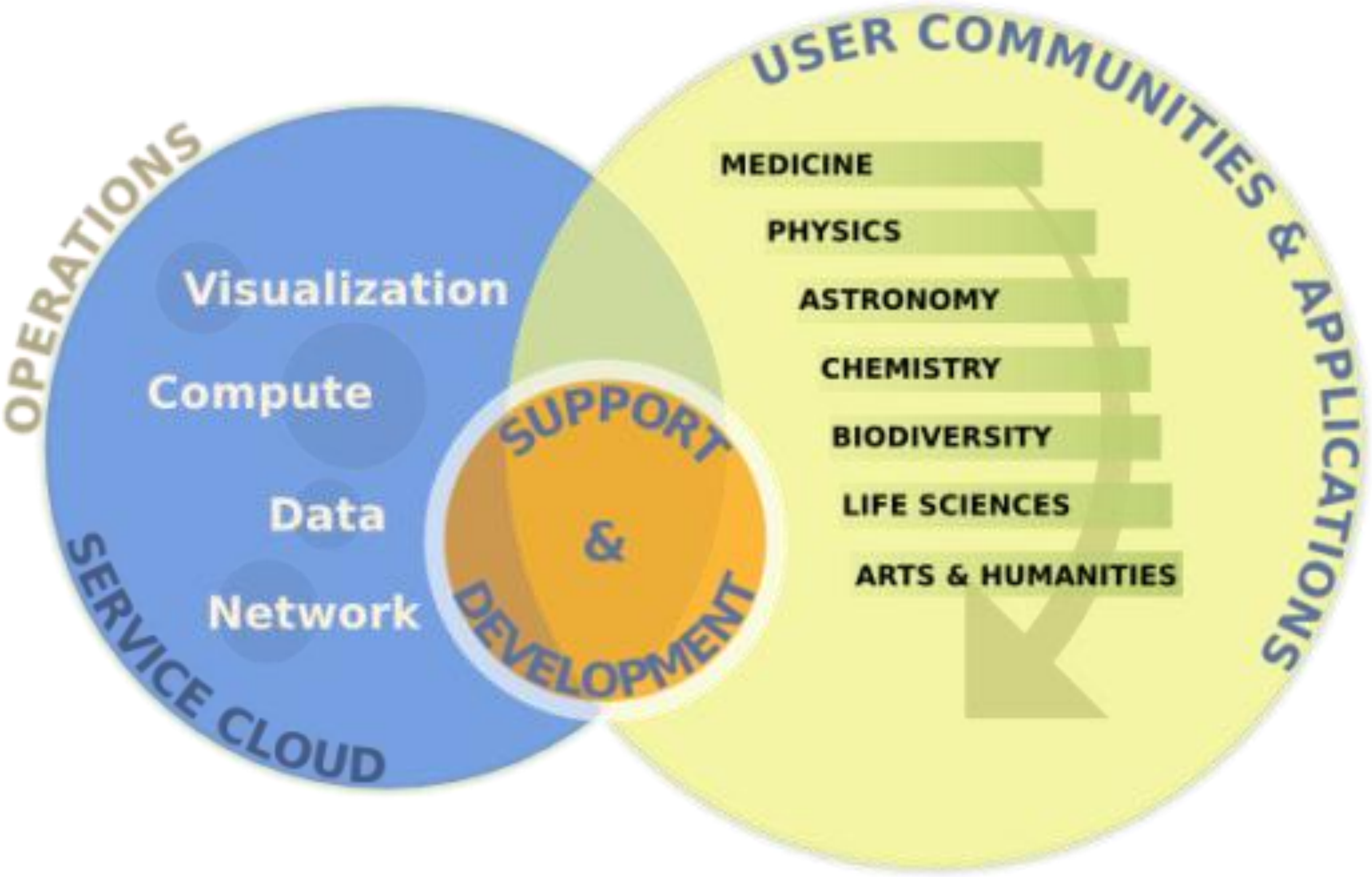
PRESERVING DATA: migrating, backing-up, storing, creating metadata and documentation, archiving

Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

Research Data Life Cycle

SURF SARA

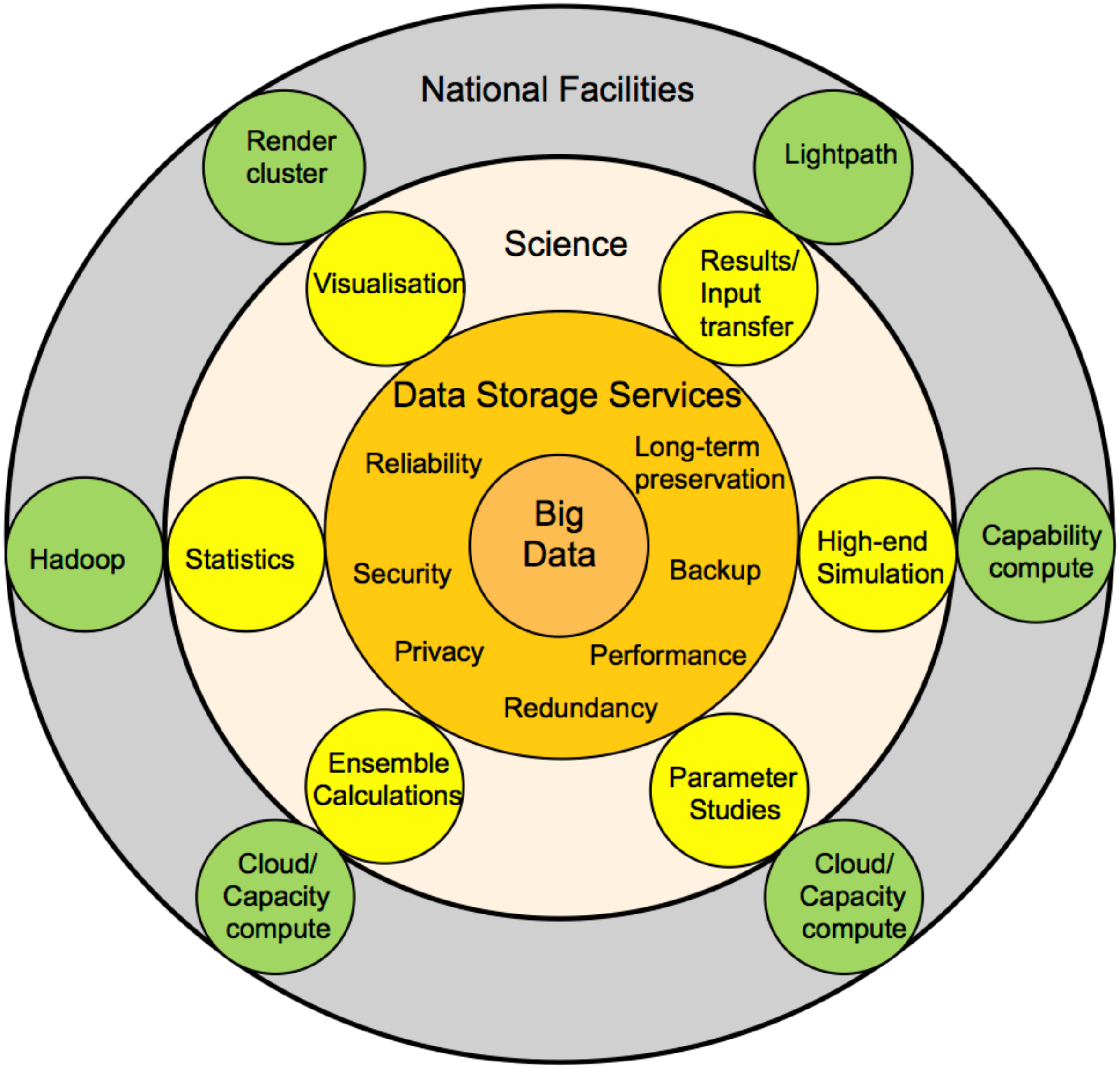
SURFsara services



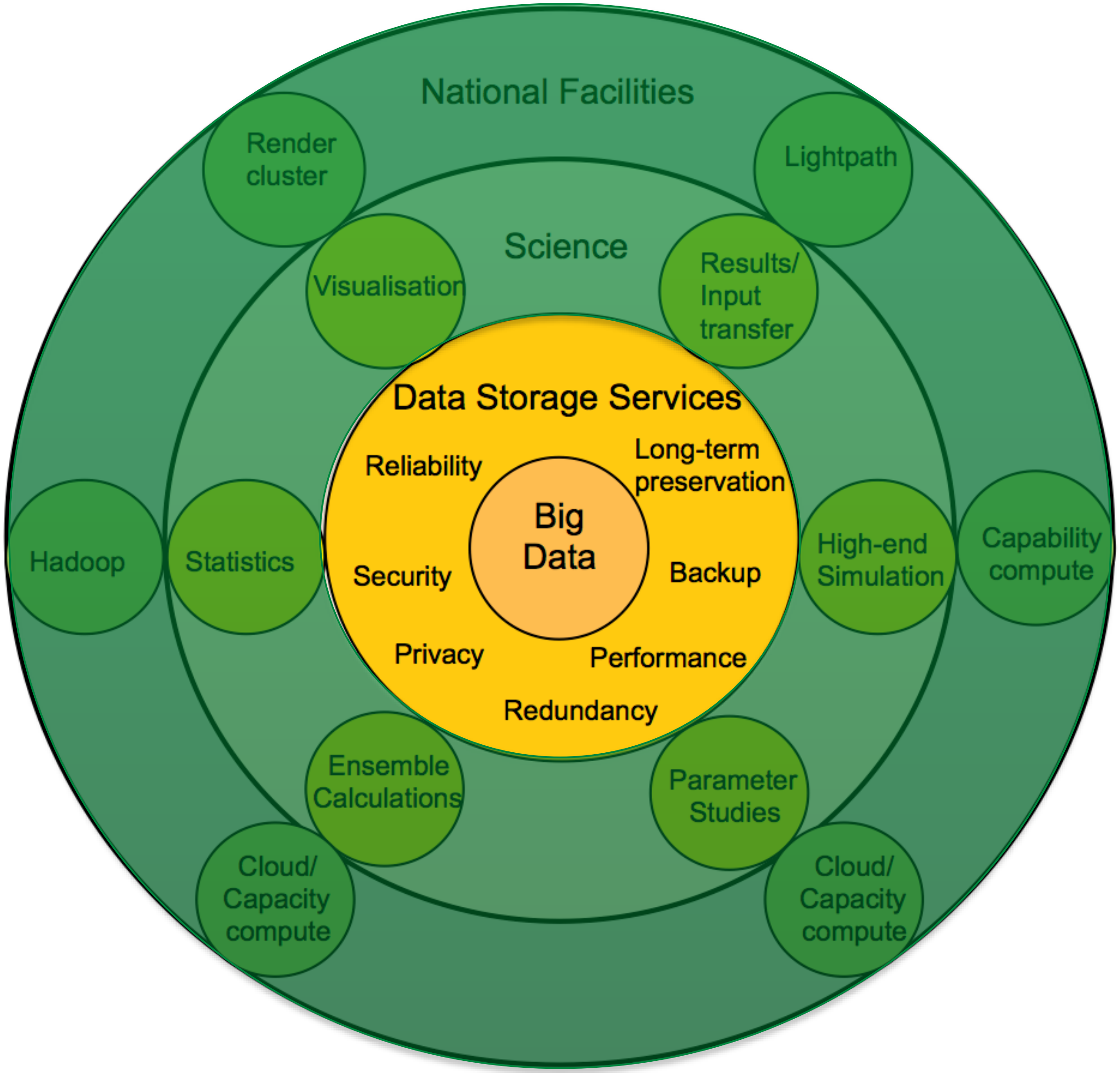
SURFsara data strategy

- SURFsara is providing **integrated ICT services** (e.g. computing, visualisation, data storage, data analysis) for the scientific research community inthe Netherlands
- SURFsara is going to provide a **Trusted Digital Repository** to:
  - to enable **long-term archiving** of research data at SURFsara
  - improve the quality of research data with **metadata** and **persistent identifiers** enrich access with **open standard protocols**
  - improve **search** and **discoverability** via harvesting
- SURFsara is going to serve the **Long Tail Data** (e.g. community clouds)
- SURFsara is providing services for **data analysis** (Hadoop, HPC Cloud, Grid, remote visualisation)
- Collaborate with Universities, Institutes and research groups on **data management** issues (e.g. community clouds) and data management plans
- Engage with funding agencies and data owners to come to transparent models on **preservation**, **policies, costs, security** and **privacy** of research data and services

SURFsara (Big) Data Centric infrastructure services



SURFsara (Big) Data Centric infrastructure services



SURFsara Data Archiving Facilities

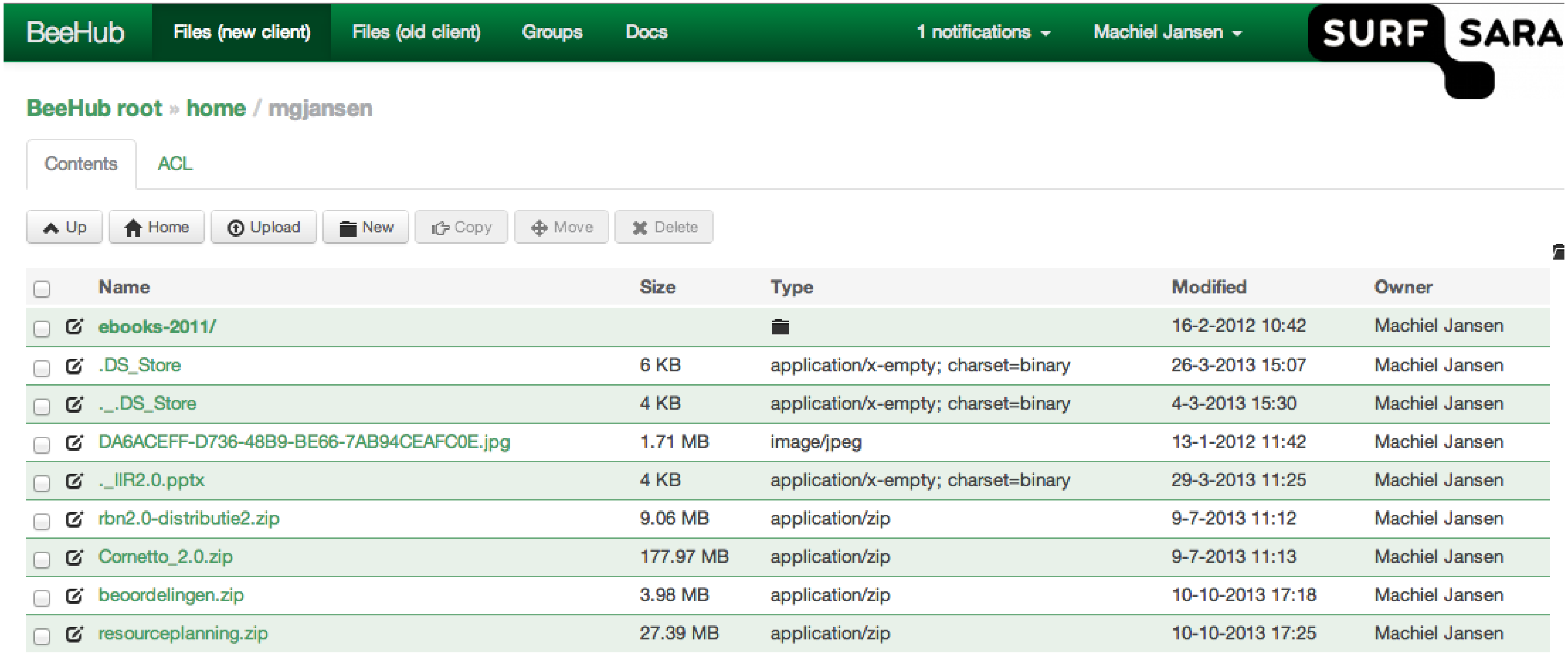
	Central Archive	GRID SE
Usage	HPC, External	GRID, HPC Cloud, Hadoop, Workflows, External
Preservation	Medium, Long term	Medium, Long term
Media	Disk + Tape	Disk + Tape
Capacity (current)	240TB + 3PB	5PB + 13PB
Bandwidth (aggregated)	1GB/s + 1GB/s	>16GB/s + 2GB/s
Latency	Direct + High (50s)	Low + High (100s)
Objects	Medium, Large (60M)	Large (30M)
Protocols	NFS, GridFTP, (hpn-) SCP, Rsync	SRM, GridFTP, HTTP, Webdav, NFS, Xrootd, dCap
Access	NWO Grant	SURF E-Infrastructure Grant

Personal cloud storage (SURFdrive)

- Trusted community cloud for personal storage
- Collaboration between SURFsara, SURFnet and Dutch universities
- Advantages:
  - Trusted community solution for personal cloud storage
  - Hosted, managed and served by community, data stored at SURFsara
  - Login through SURFconext
  - Specifications and service determined by end-users (universities)
- Initial capacity: 200 TB, 100 GB storage capacity per user
- Based on ownCloud
- Operational: April 1st, 2014

SURF DRIVE

Cloud storage for research (Beehub)



BeeHub root » home / mgjansen

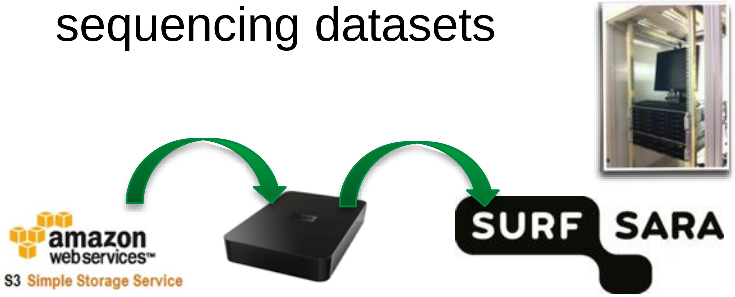
□	Name	Size	Type	Modified	Owner
□	ebooks-2011/			16-2-2012 10:42	Machiel Jansen
□	.DS_Store	6 KB	application/x-empty; charset=binary	26-3-2013 15:07	Machiel Jansen
□	._.DS_Store	4 KB	application/x-empty; charset=binary	4-3-2013 15:30	Machiel Jansen
□	DA6ACEFF-D736-48B9-BE66-7AB94CEAFC0E.jpg	1.71 MB	image/jpeg	13-1-2012 11:42	Machiel Jansen
□	._IIR2.0.pptx	4 KB	application/x-empty; charset=binary	29-3-2013 11:25	Machiel Jansen
□	rbn2.0-distributie2.zip	9.06 MB	application/zip	9-7-2013 11:12	Machiel Jansen
□	Cornetto_2.0.zip	177.97 MB	application/zip	9-7-2013 11:13	Machiel Jansen
□	beoordelingen.zip	3.98 MB	application/zip	10-10-2013 17:18	Machiel Jansen
□	resourceplanning.zip	27.39 MB	application/zip	10-10-2013 17:25	Machiel Jansen

Capacity: 80 TB
Nr of users: 544 users (69 use SURFconext)
Usage: 21 TB in user folders, 2 TB in group folders, 23 TB stored
Interface: webdav

SURFsara Data Services

Data Ingest Service

- Easy way to upload large data from disk onto SURFsara facilities
- Upload data from 45 disks in parallel
- Used for archiving large sequencing datasets



EPIC/Handle **PID service**

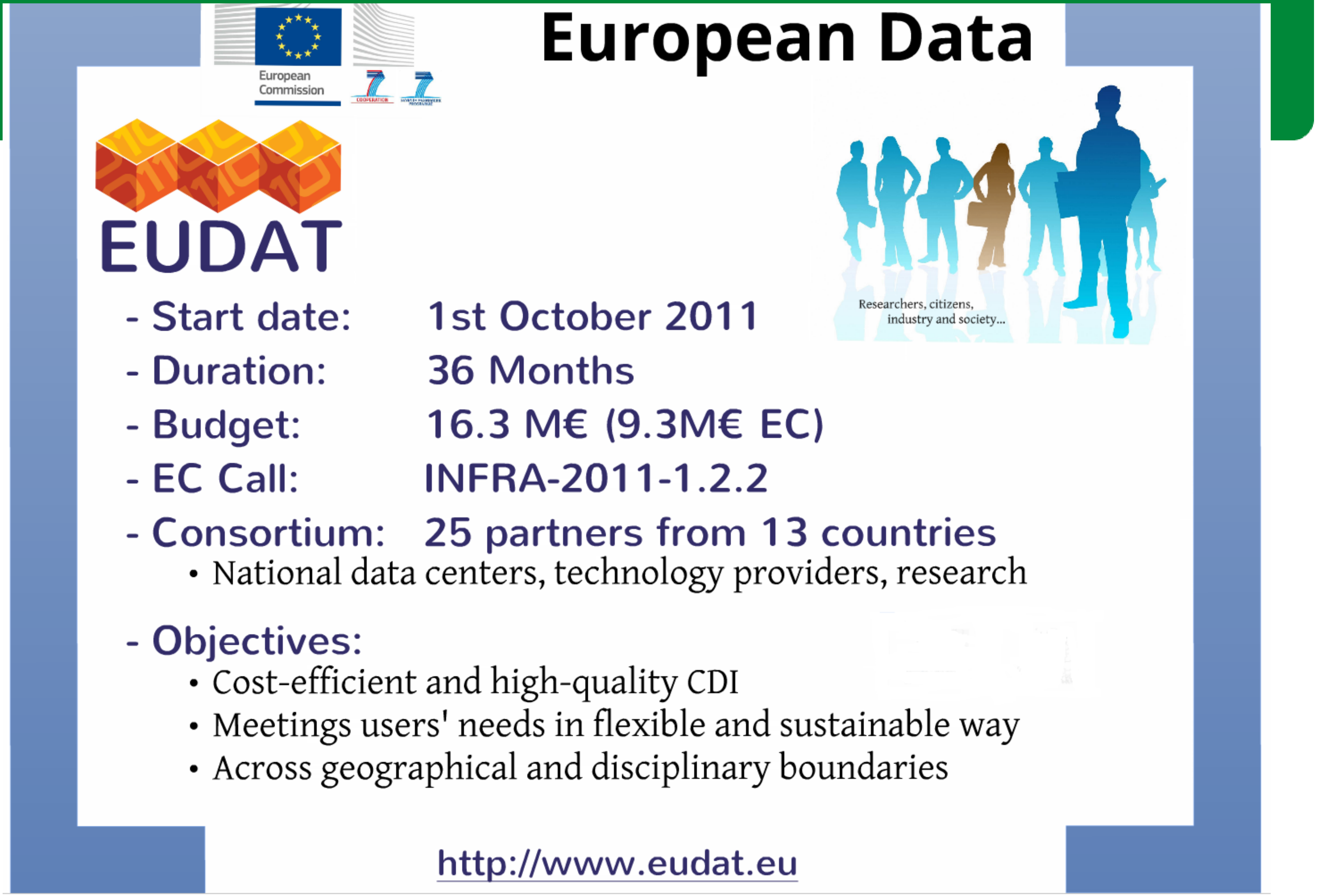
- Persistent references to physical locations to make data objects findable and citable
- PID is comparable to SBN of Books
- EPIC/Handle system is comparable to DNS for Data Objects
- EPIC consortium: Providing a sustainable service for storing an maintaining large volumes of PIDs

Trusted Digital Repository

- Data repository service to deposit data sets and objects
- Long-term preservation of research data
- Provides quality to data sets and objects via metadata descriptions
- Makes data sets and objects citable and findable in the future via Persistent Identifiers (EPIC PID service)
- Makes data sets discoverable via metadata harvesting
- Improves data access and re-usability of research data
- Ensures trust to researchers via regular auditing against standardized certifications (e.g. Data Seal of Approval, ISO16363, DIN 31644)
- Trusted Digital Repository service is currently under construction at SURFsara
- Collaboration with DANS and Universities to provide service to manage research data

SURFsara Research Data Storage

- Provides a central location for storing and archiving research data and files
- It is based on available technology at SURFsara providing a high reliable and available storage system
- Has good access to other SURFsara facilities
- Stores dual copies of data on tape at 2 geographical separated locations (Amsterdam and Almere), regular integrity checks are performed
- Provides different access protocols (e.g. SCP, SFTP, Rsync, GridFTP) to upload and download data
- SSH access is provided for easy management, individual users will get an user account
- It has high speed network connections (10GE) and high single transfer speeds between 100-250MB/s
- Paid service, can be used by SURF affiliated institutes



European Data

European Commission

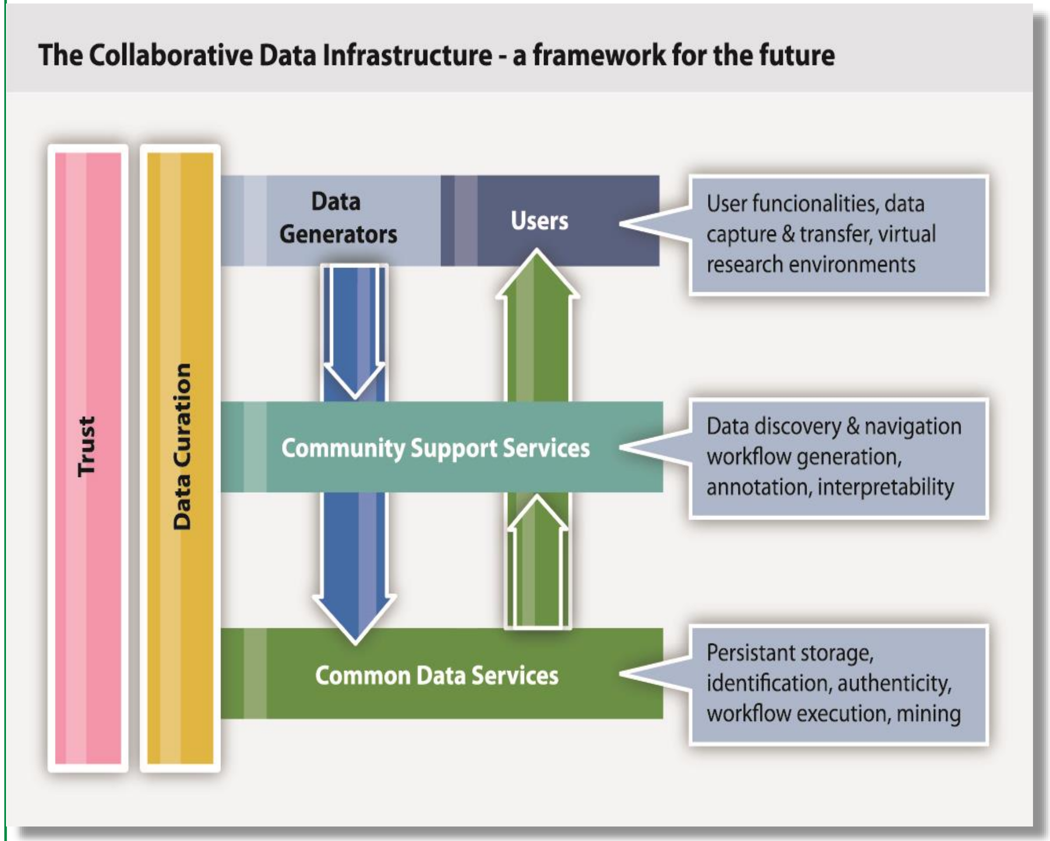
EUDAT

Researchers, citizens,
industry and society...

- Start date: 1st October 2011
- Duration: 36 Months
- Budget: 16.3 M€ (9.3M€ EC)
- EC Call: INFRA-2011-1.2.2
- Consortium: 25 partners from 13 countries
  - National data centers, technology providers, research
- Objectives:
  - Cost-efficient and high-quality CDI
  - Meetings users' needs in flexible and sustainable way
  - Across geographical and disciplinary boundaries

http://www.eudat.eu

EUDAT Mission

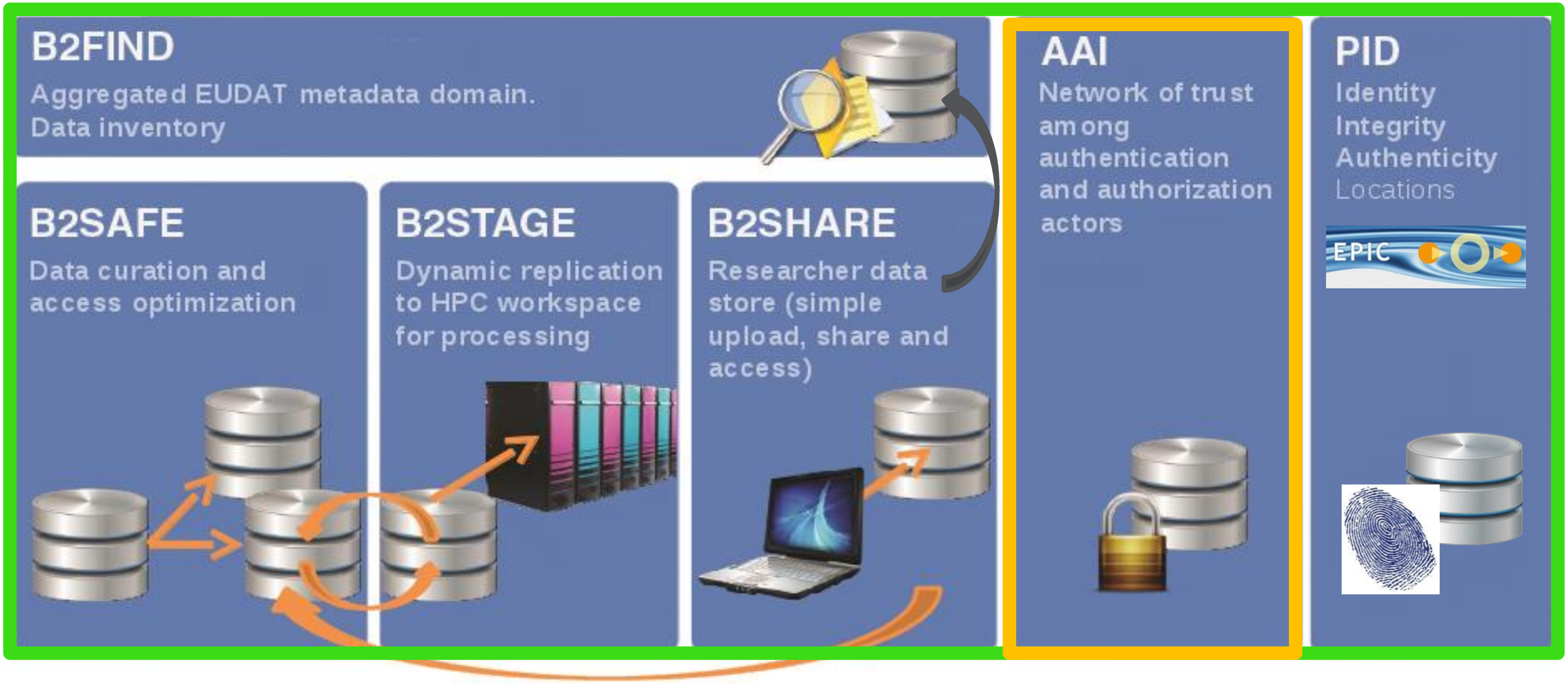

The Collaborative Data Infrastructure - a framework for the future
Trust
Data Curation
Data Generators
Users
User funcionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation workflow generation, annotation, interpretability
Common Data Services
Persistant storage, identification, authenticity, workflow execution, mining

- offer common data services in CDI to all European researchers
- services will address the needs of big data volumes as well as of long tail of data
- respect the communities' choices of data organizations
- achieve harmonization and efficiency in the long term

🟩 being offered

🟧 in progress

EUDAT Service Overview



B2FIND
Aggregated EUDAT metadata domain.
Data inventory

B2SAFE
Data curation and access optimization

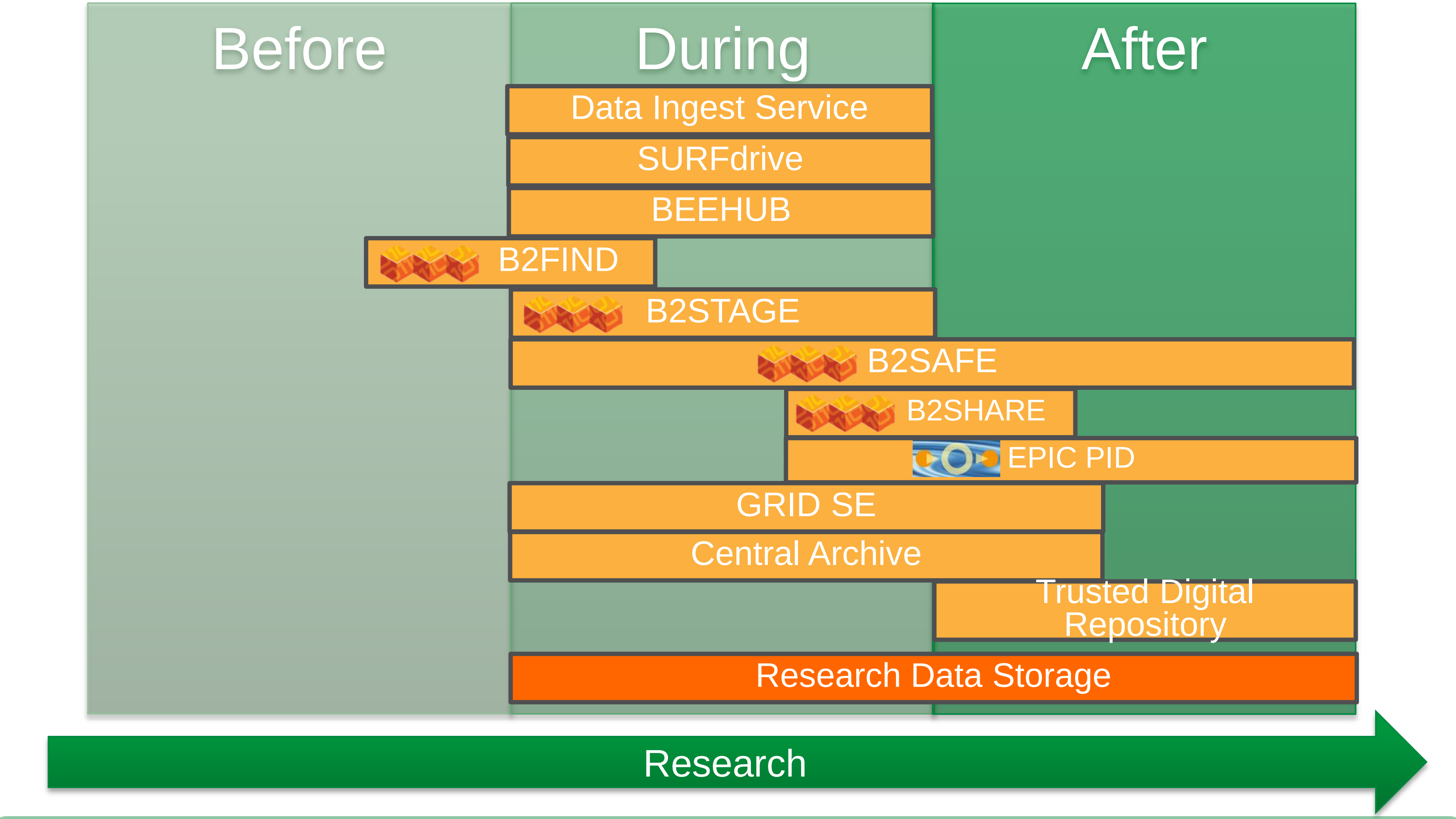
B2STAGE
Dynamic replication to HPC workspace for processing

B2SHARE
Researcher data store (simple upload, share and access)

AAI
Network of trust among authentication and authorization actors

PID
Identity
Integrity
Authenticity
Locations

Research Timeline



Research Data Alliance

- Vision is researchers and innovators openly sharing data across technologies, scientific disciplines and countries
- RDA builds the social and technical bridges that enables open sharing of data
- Brings researchers from different scientific disciplines, librarians and IT people together
- Young initiative, started in March 2013, still growing (now 1300+ people from 55 countries)
- Supported by the European Commission, US National Science Foundation and National institute of Standards and Technologies and Australian Government's Department of Innovation
- Organized via exploratory Interest Groups (now 29) and focused Working Groups (now 9)
- IS en WG are open to anyone who agrees with the RDA principles
- Subjects discussed: metadata, persistent identifiers, data citation, data registries, policies, certification, terminologies, data interoperability, domain specific issues
- Biannual plenary meetings are organized, next PM will be on the 22-24th of September in Amsterdam
- Website: http://www.rd-alliance.org

Questions?

Mark van de Sanden, mark.vandesanden@surfsara.nl