Digitale forskningsdata i et nasjonalt perspektiv

NARMA vårkonferanse

28. mars 2017

Gunnar Boe,
Daglig leder

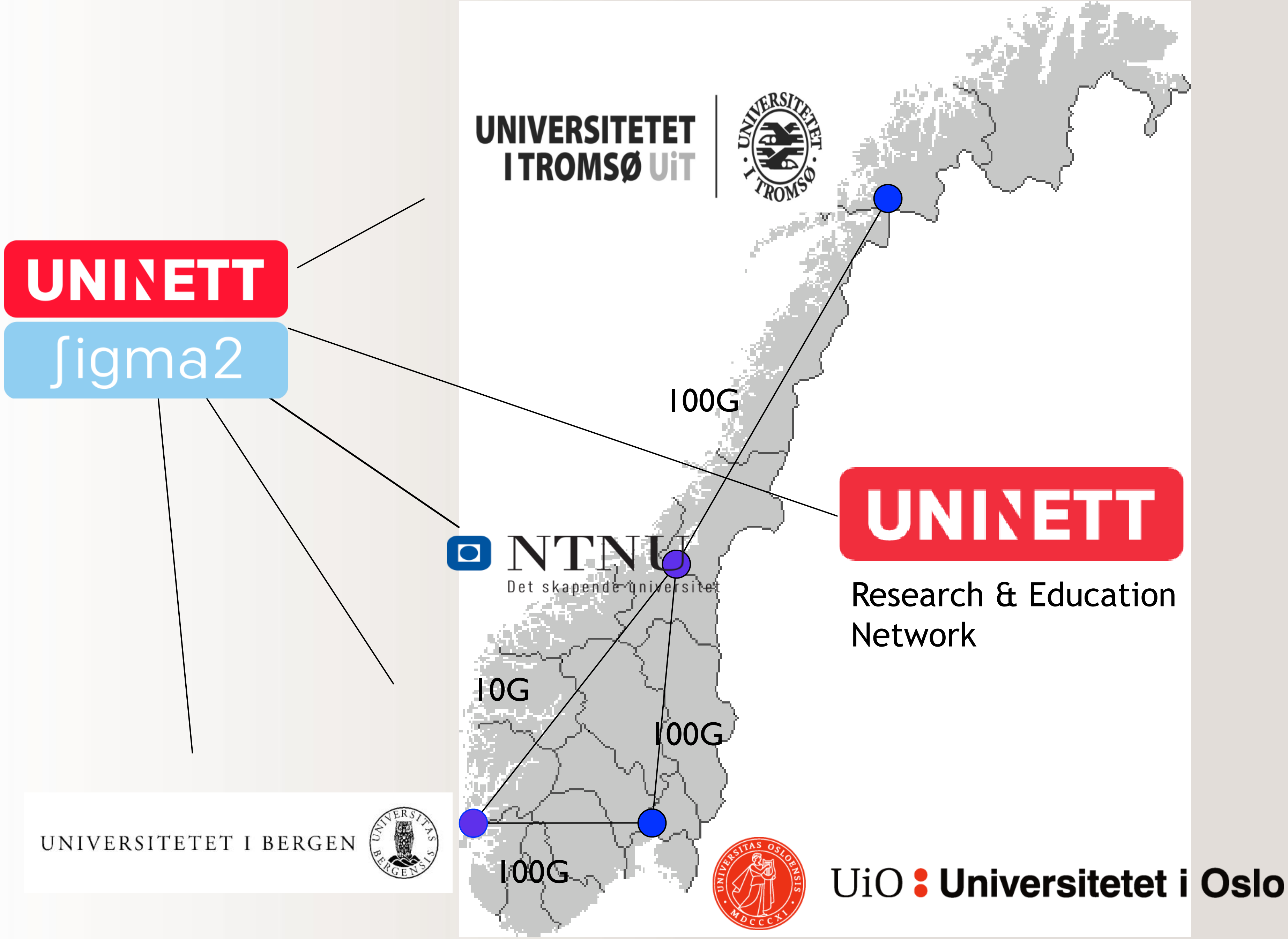
UNINETT ʃigma2

Agenda

- Om UNINETT Sigma2
- Om forskningsdata
- Om oppgavene (nasjonalt vs lokalt)
- Om e-infrastrukturen

About UNINETT Sigma2

- Established in December 2014 based on a decision from the 4 oldest universities and the Research Council of Norway
- A long-term model with 5+5 years and evaluation of the company after 5 years. (i.e. minimum 10 year lifetime for the company)
- Part of the UNINETT corporation, separate company
- Collaboration agreement with the 4 oldest universities incl. 50 MNOK yearly funding
- Contract with the Norwegian Research Council incl. 25 MNOK yearly funding
- Granted infrastructure funding (75.7 MNOK investment 2016-2017) from the Norwegian Research Council
- Operation and support contract with the 4 oldest universities
- Frame agreement with the universities for project work

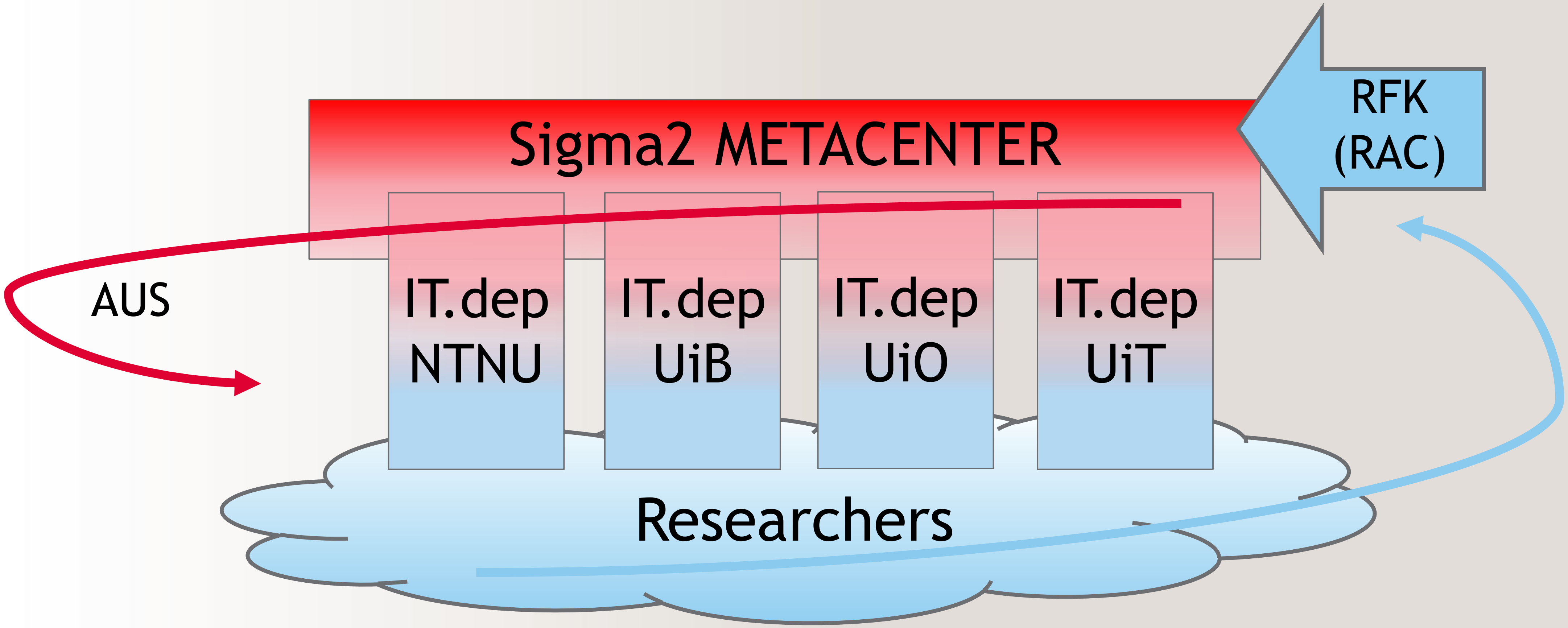


UNINETT

Research & Education Network

The Metacenter

- National coordination and shared, consolidated resources have cost and efficiency advantages but creates a “distance” to the end-users (researchers)
- This is avoided by keeping the support staff and competence near where the research is going on, at the universities
- Combined with a data-centric architecture for the e-infrastructure, this model combines the advantages of the centralized model and the local model



High level objectives

- Procure, operate and develop a critical national e-infrastructure
- Promote e-infrastructure to new research communities
- Lead and coordinate participation in international cooperation for e-infrastructure
- Provide an attractive and sustainable e-infrastructure for all research communities, with the following characteristics:
  - High reliability and availability
  - Cost effectiveness
  - Predictable access
  - Interoperability within the national e-infrastructure and between national and international infrastructures (e.g. PRACE, EUDAT)
- Provide services for data analytics of large datasets (Big Data)

The summary

- Provide services that researchers need today, e.g. advanced user support, training, data services such as data management and analytics of large datasets (Big Data), and of course high performance computing (HPC).

Research data

Research Council Policy Objectives

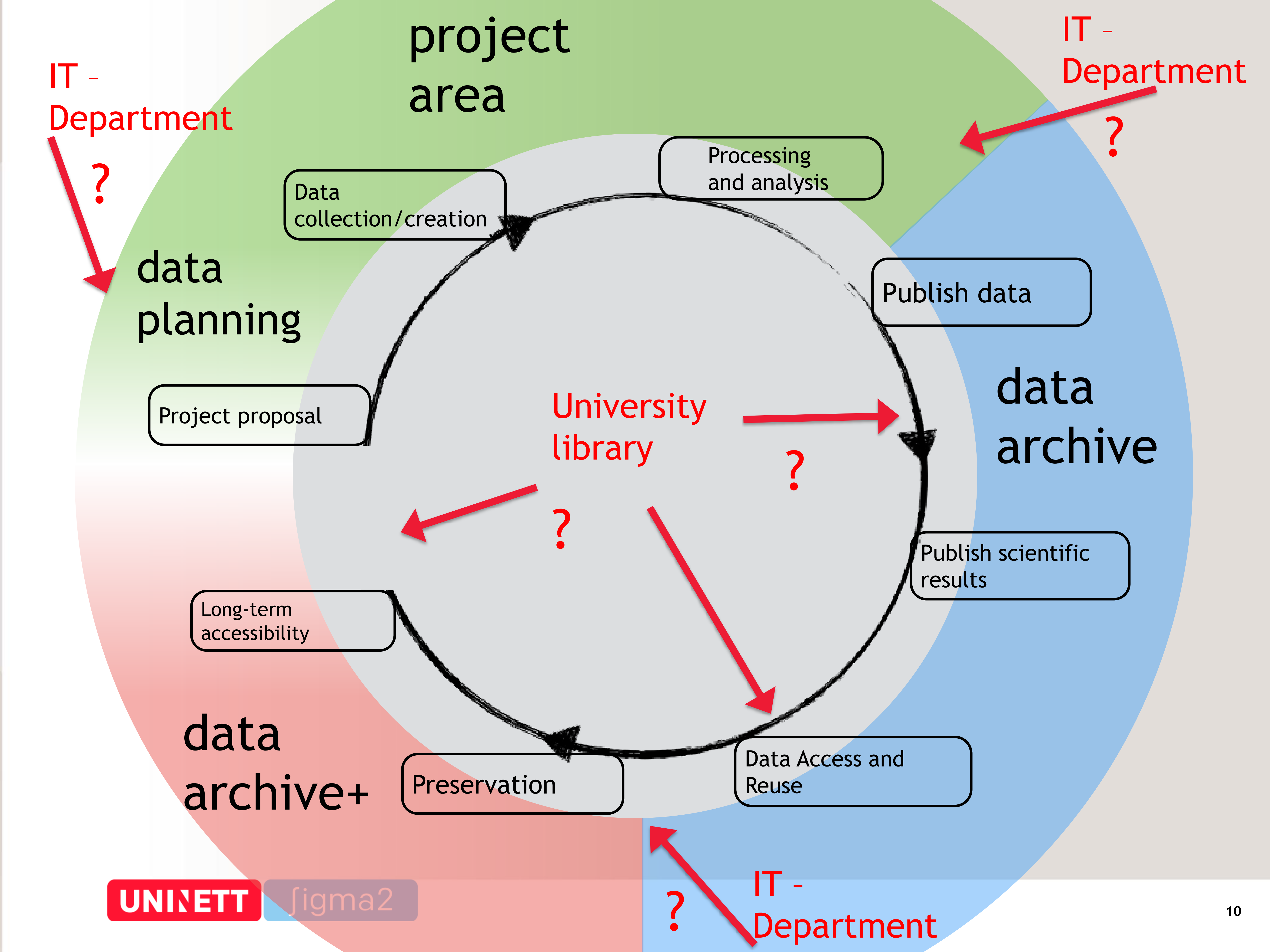
- Improve quality in research through better opportunities to use previous work and combine data in new ways
- Transparency in research process and better opportunities to verify scientific results
- Increased collaboration and less duplication of reaserch
- Increase innovation in business and public sector
- Efficiency improvement and better use of public funding

orskningsrådet

Tilgjengeliggjøring a
forskningsdat

Policy f
Norges forskningsrå

Forskningsrådet. Tilgjengeliggjøring av forskningsdata
- Policy for Norges forskningsråd.
Norges forskningsråd; 2014



project area

data planning

data archive

data archive+

Data collection/creation

Processing and analysis

Publish data

Publish scientific results

Data Access and Reuse

Preservation

Long-term accessibility

Project proposal

University library

?

?

IT – Department

?

IT – Department

?

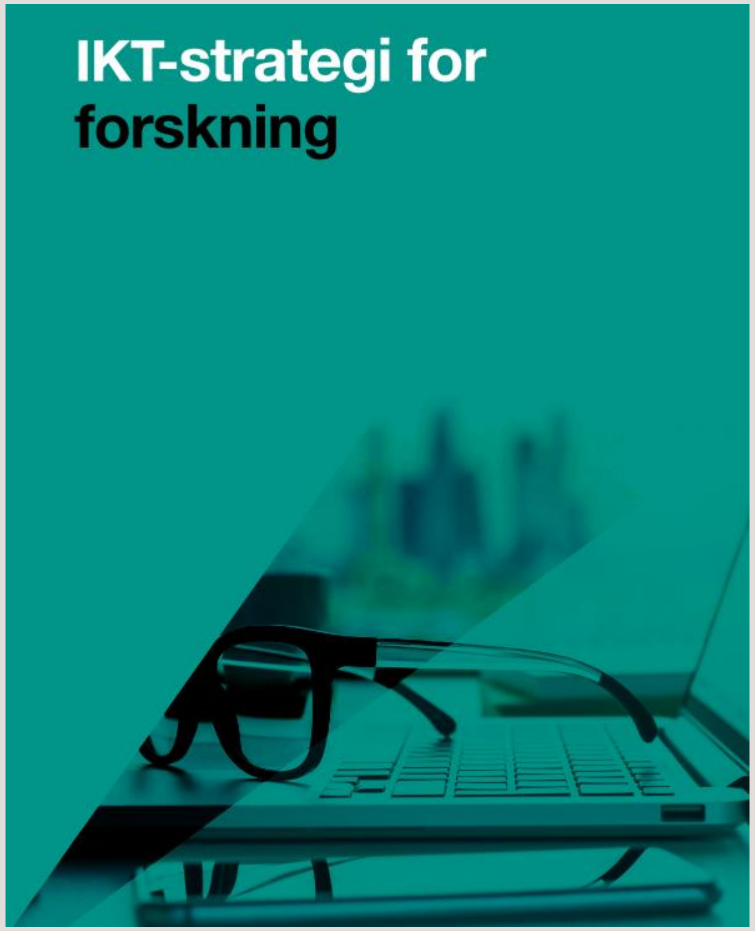
IT – Department

?

UNINETT Sigma2

The actors… who provides what

- International level
- National level
- University/institutional level
- Deparments / Faculties
- Institute or research group


IKT-strategi for forskning

IKT-strategi for norsk universitets- og høgskolesektor

IKT-strategi og helhetlige løsninger i norsk universitets- og høgskolesektor

National e-infrastructure level

- The global view, Interfacing with international services/e-infrastructures
- Generic services shared by many
- Economy of scale
- Providing services for publicly funded (RCN) research and enabling interaction between various stakeholders
- Competence

University/institutional level

- Special local needs, Specific for the university
- Integration with local services
- Connect and promote data to higher level repositories

- Data curation best done locally?

Services

Sigma2 e-infrastructure services 1/2

- Computation
  - Compute cycles for computational research
- Storage
  - Data management planning
  - Data storage, including Sensitive data
  - (Visualization, Data-analytics)
- Basic user support
  - Basic tech support through a ticket-based support service
  - Training
- Advanced user support

Advanced User Support (AUS)

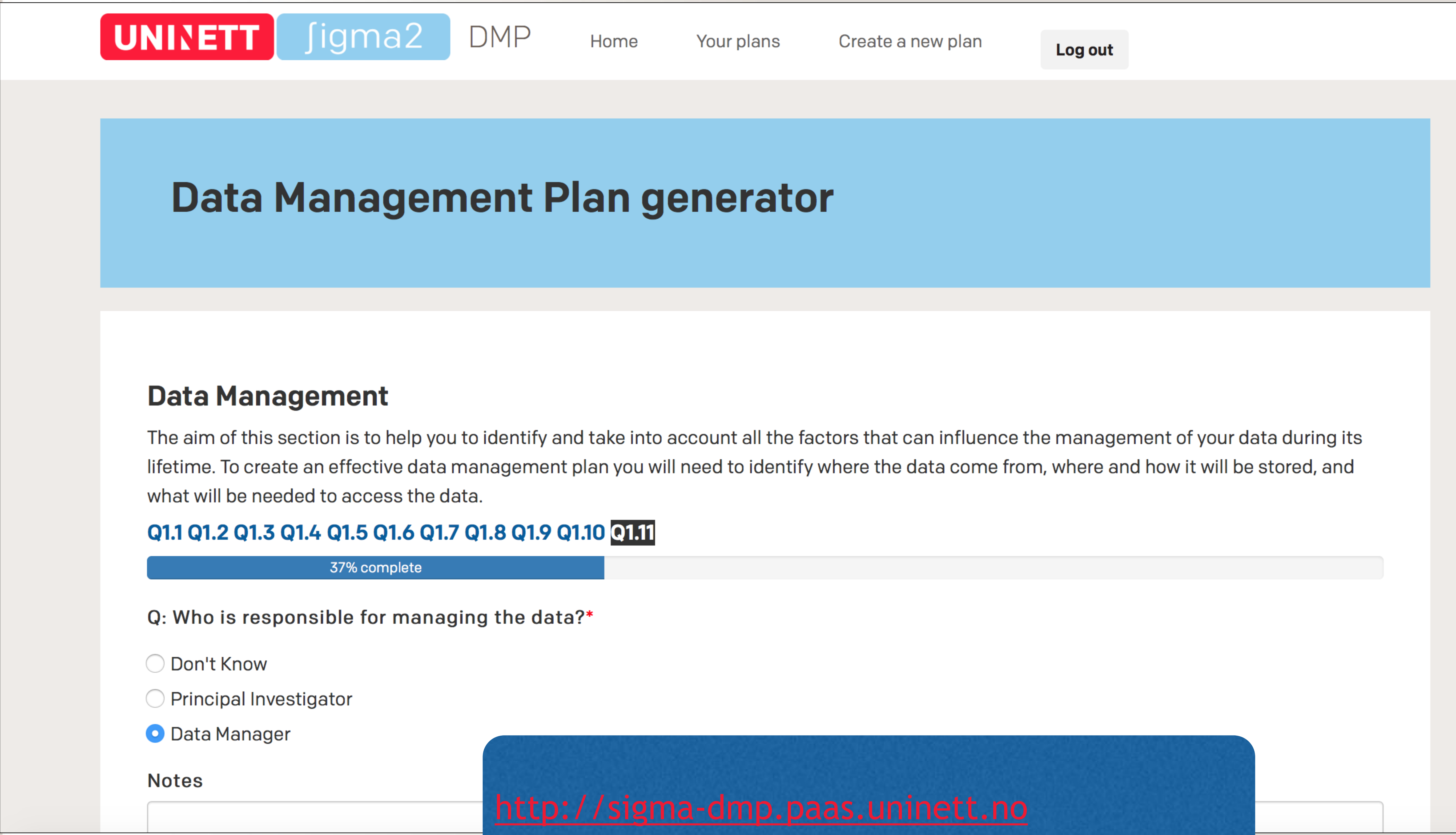
- 1) Project based AUS:
  - can be the sole initiative of a researcher or a science area
  - granted by RFK with 2-3 PMs spent over a maximum of 6 months.
- 2) Discipline specific AUS
  - initiated by Sigma2 in cooperation with a science discipline
  - can have allocations of more than 12 PMs spent over a maximum for 2 years
  - joint funding

Advanced User Support (AUS)

For the storage services, project based advanced user support aims at:

- assist researchers to create data plans
- implementing best practices for collecting and handling data
- identifying or defining meta-data schema
- identifying suitable storage formats
- identifying dedicated or specialised tools to help access or visualize data, utilise the facilities better

Data ‘policy’ for Research data



Data Management Plan generator

Data Management

The aim of this section is to help you to identify and take into account all the factors that can influence the management of your data during its lifetime. To create an effective data management plan you will need to identify where the data come from, where and how it will be stored, and what will be needed to access the data.

Q: Who is responsible for managing the data?*

○ Don't Know

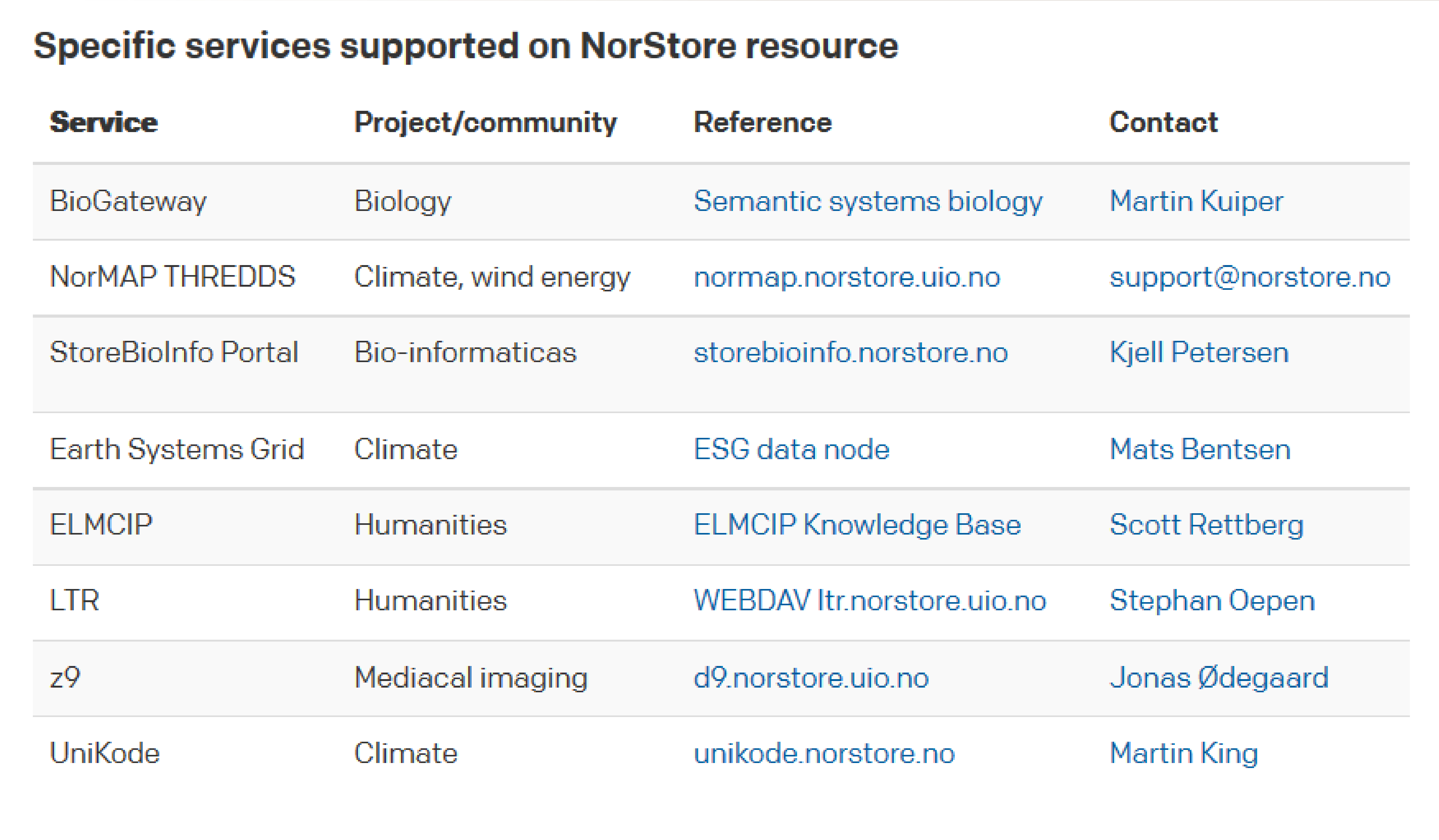
○ Principal Investigator

● Data Manager

Notes

http://sigma-dmp.paas.uninett.no

Sigma2 e-infrastructure services 2/2

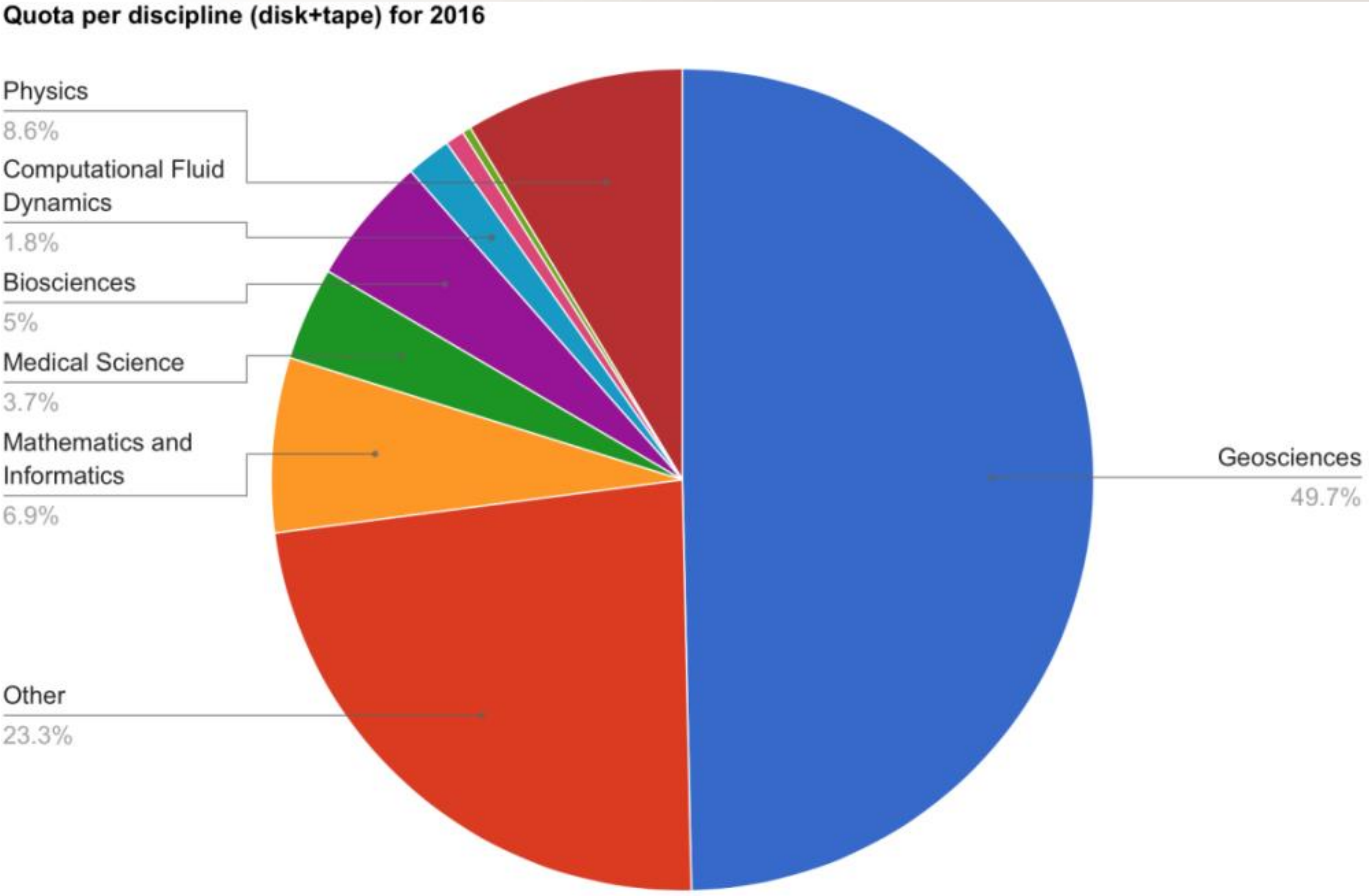


Specific services supported on NorStore resource

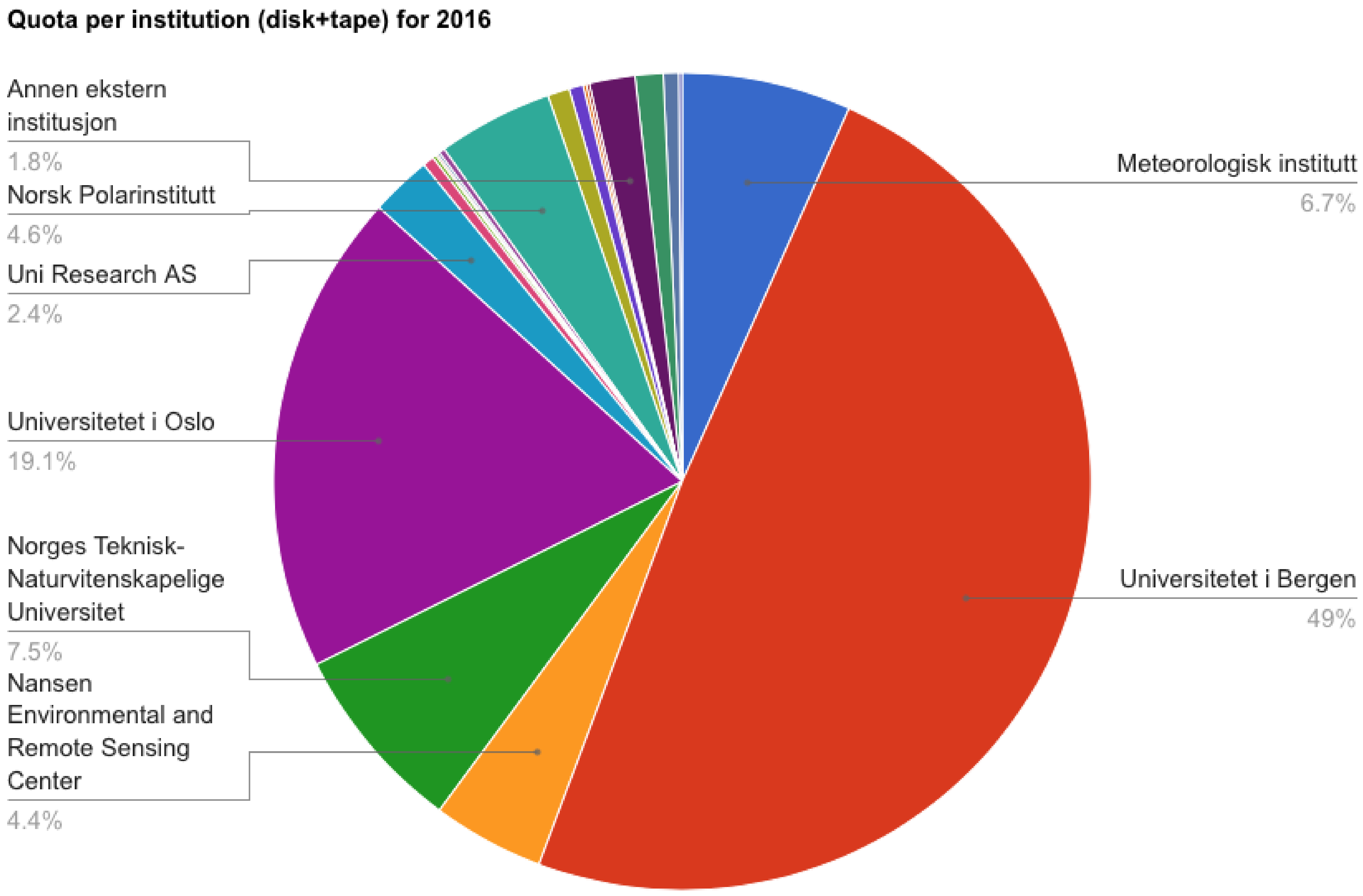
Service	Project/community	Reference	Contact
BioGateway	Biology	Semantic systems biology	Martin Kuiper
NorMAP THREDDS	Climate, wind energy	normap.norstore.uio.no	support@norstore.no
StoreBioInfo Portal	Bio-informaticas	storebioinfo.norstore.no	Kjell Petersen
Earth Systems Grid	Climate	ESG data node	Mats Bentsen
ELMCIP	Humanities	ELMCIP Knowledge Base	Scott Rettberg
LTR	Humanities	WEBDAV ltr.norstore.uio.no	Stephan Oepen
z9	Mediacal imaging	d9.norstore.uio.no	Jonas Ødegaard
UniKode	Climate	unikode.norstore.no	Martin King

- Organised in the Metacentre

UNINETT ʃigma2



Quota per discipline (disk+tape) for 2016



UNINETT ʃigma2

Data intensive Science Disiciplines

- Climate (IPCC production, ESGC data node, HPC intensive data)
  - large datasets, avoid moving data, scalability, data longevity and integrity

- Neuroscience (HumanBrain, Kavli Inst., INCF)
  - sensitive data, raw sensor data, data mgmt tool

- ELIXIR.NO (next generation sequencing, analysis/processing, sharing/archiving, data product delivery)
  - portals, AAI, work flow mgmt, access to tools

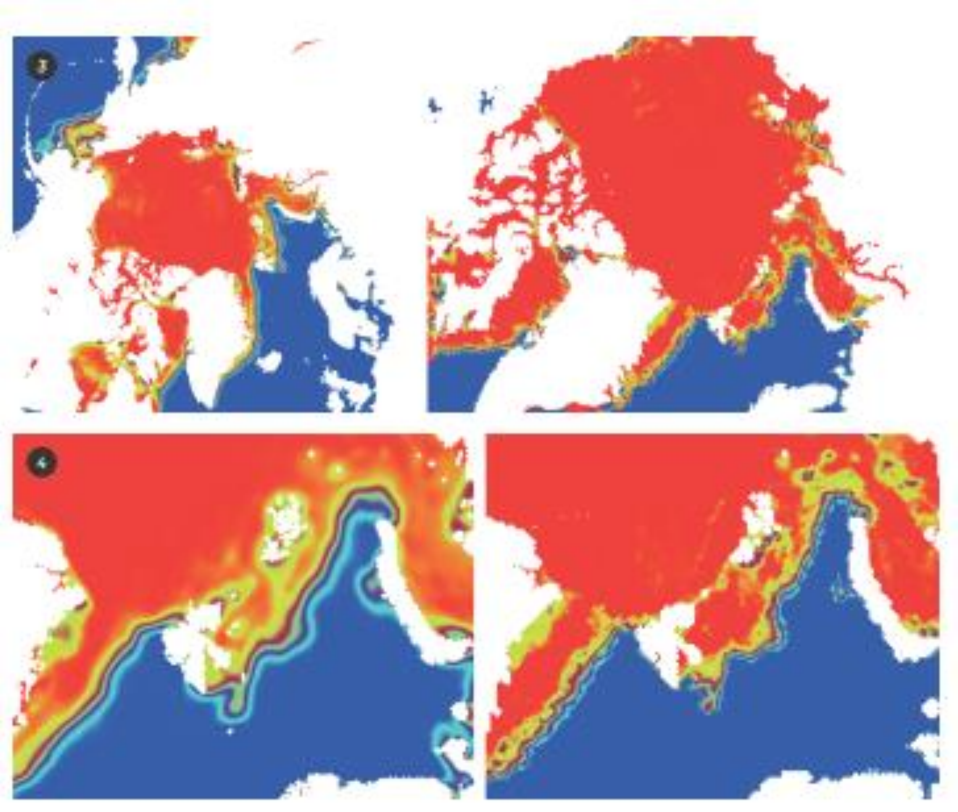
- CLARINO (structured data, corpus)
  - AAI, data access, DOIs, centralising HPC+data

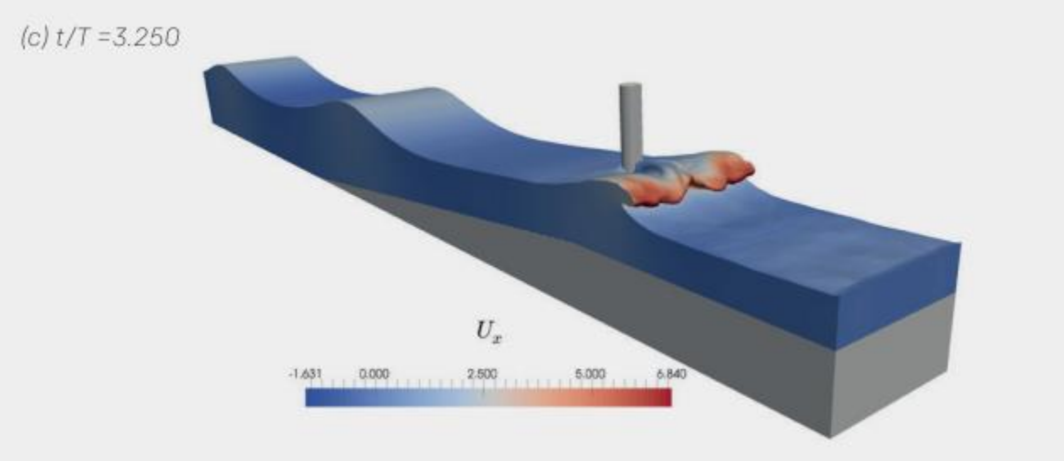
- Biodiversity (GBIF, LifeWatch)
  - portals, access/sharing, metadata, own PIDs, Biobanks)

- Marine environment (sensor collection, basic service needs) …

- EPOS (implementation phase, sensor collection) …

Examples of projects





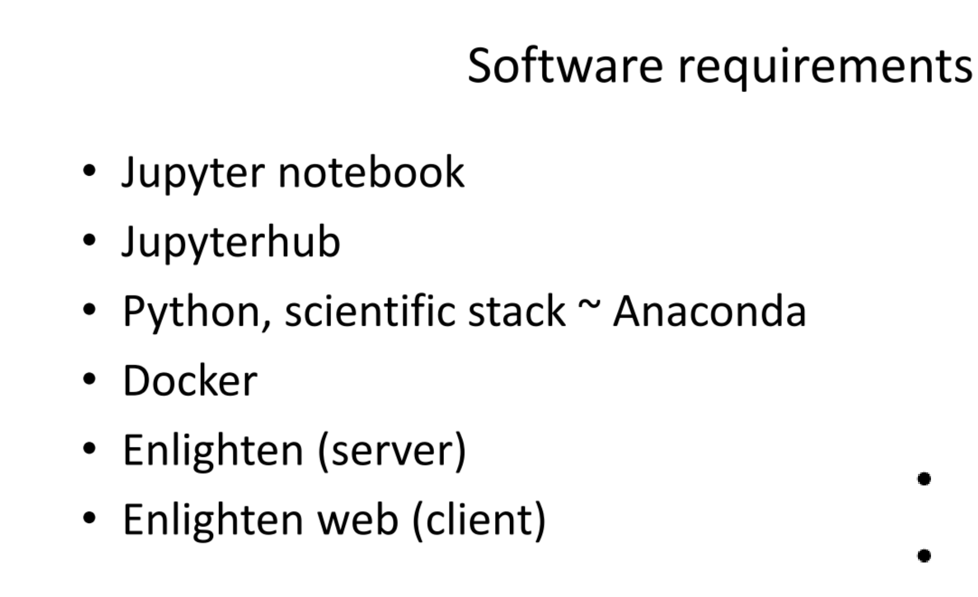
Pictures from

META 1/2015

The infrastructure: A new architecture

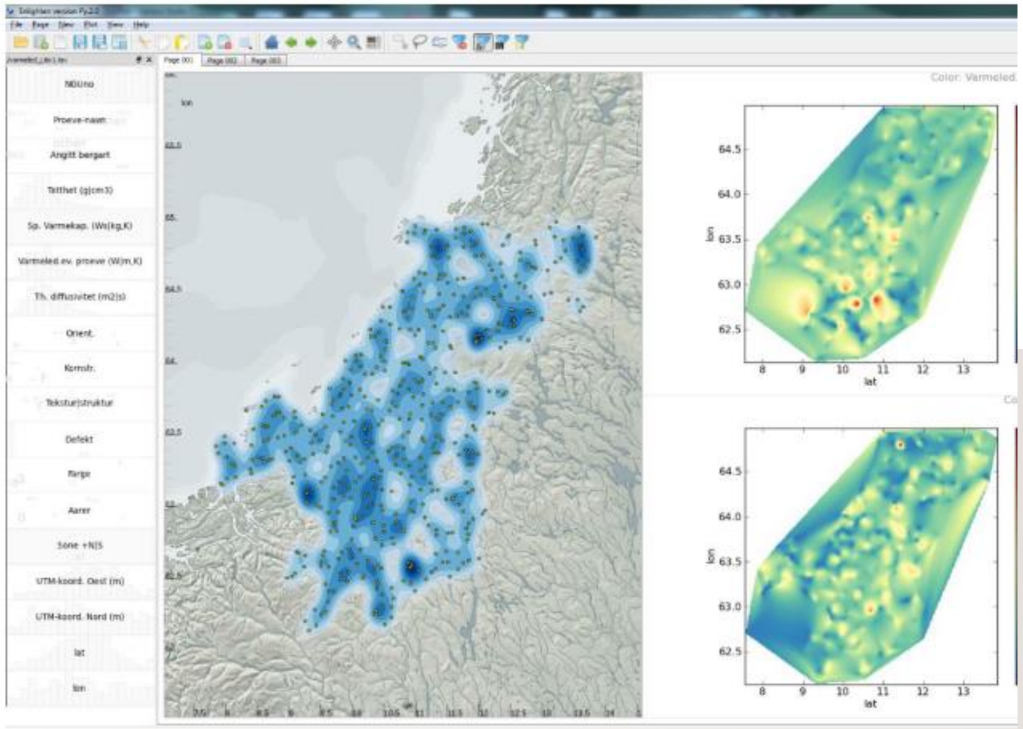
- National Infrastructure for Research Data (NIRD)

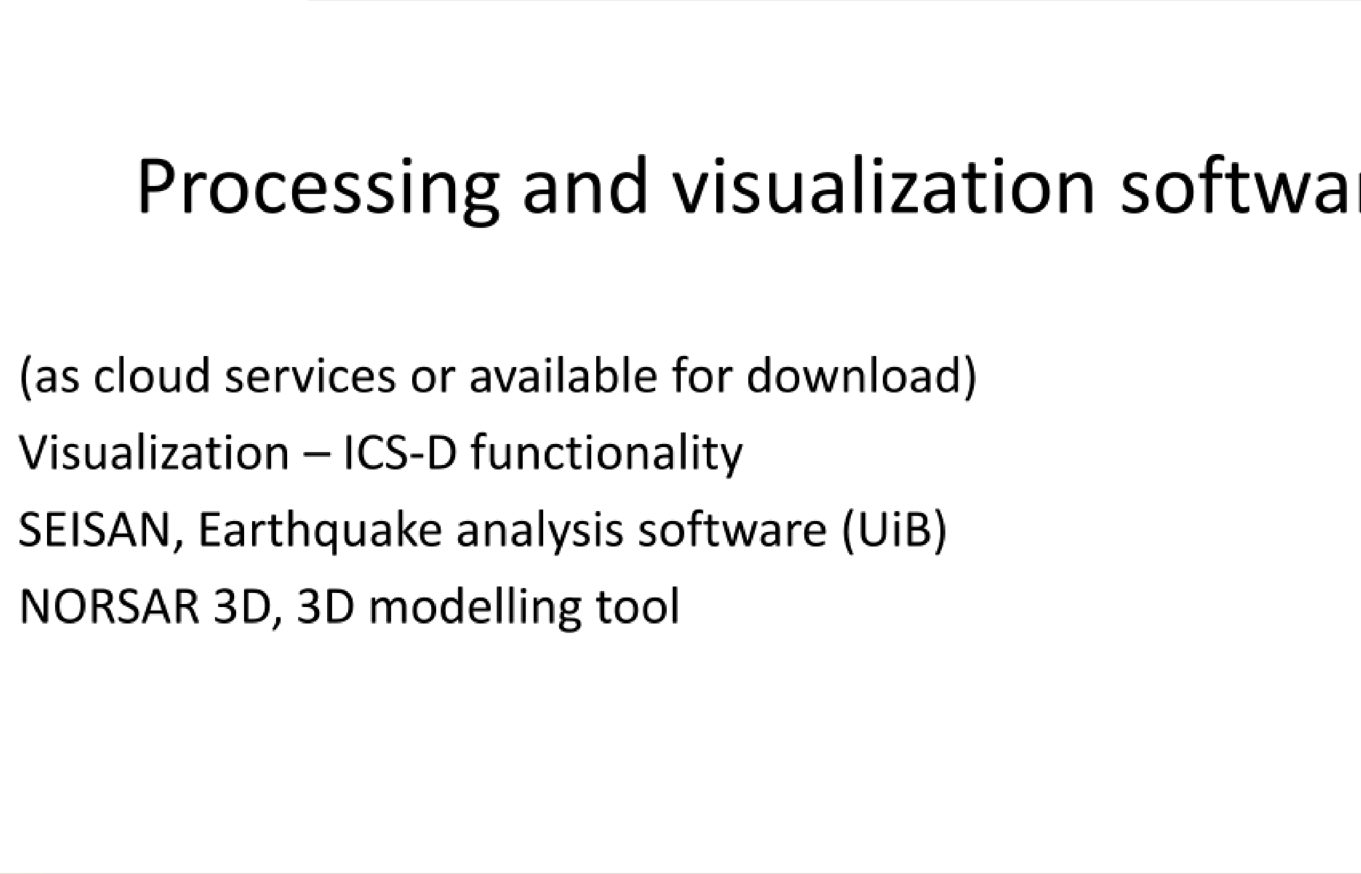
What researchers requests:



Software requirements

- Jupyter notebook
- Jupyterhub
- Python, scientific stack ~ Anaconda
- Docker
- Enlighten (server)
- Enlighten web (client)





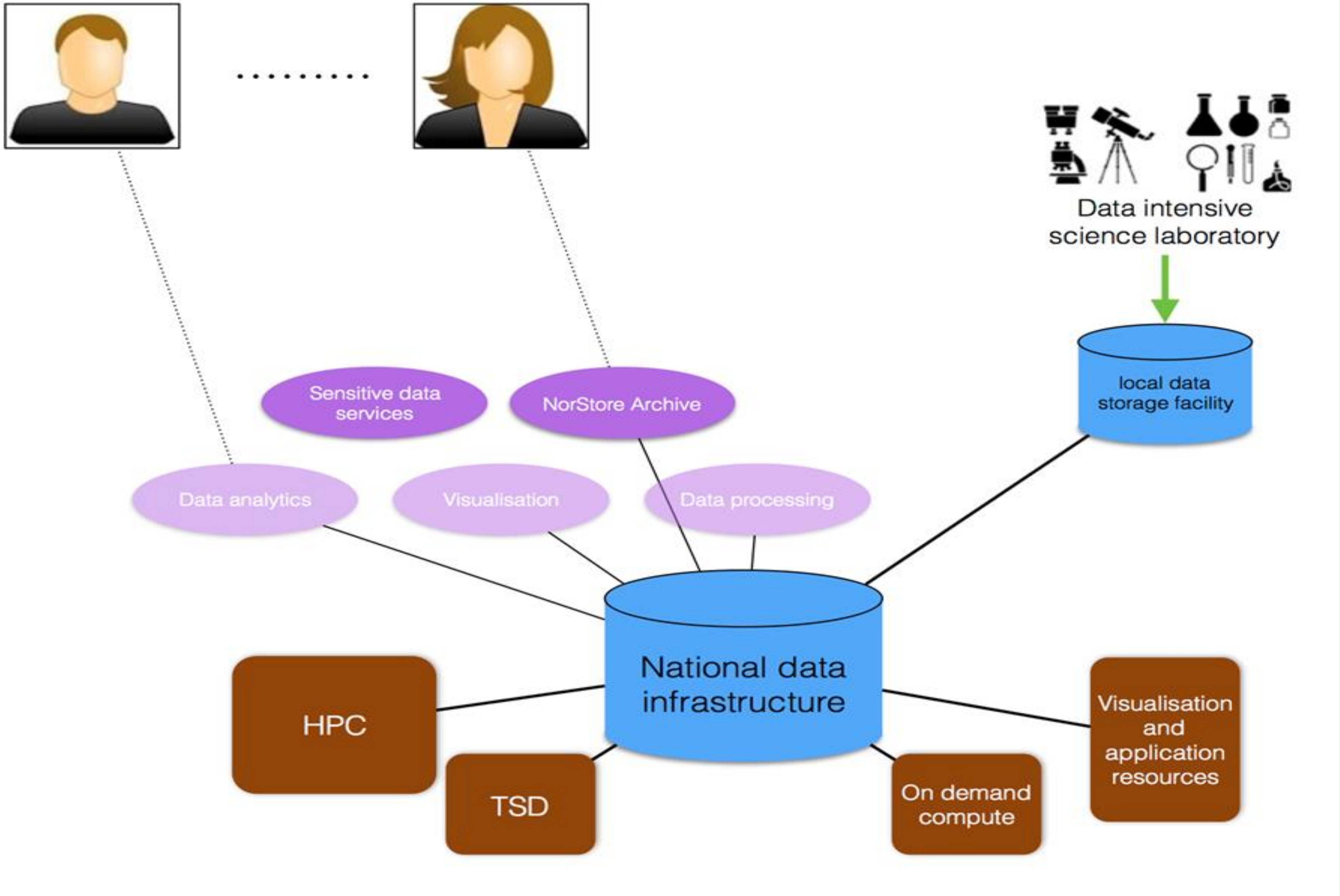
Processing and visualization softwa

- (as cloud services or available for download)
- Visualization – ICS-D functionality

SEISAN, Earthquake analysis software (UiB)

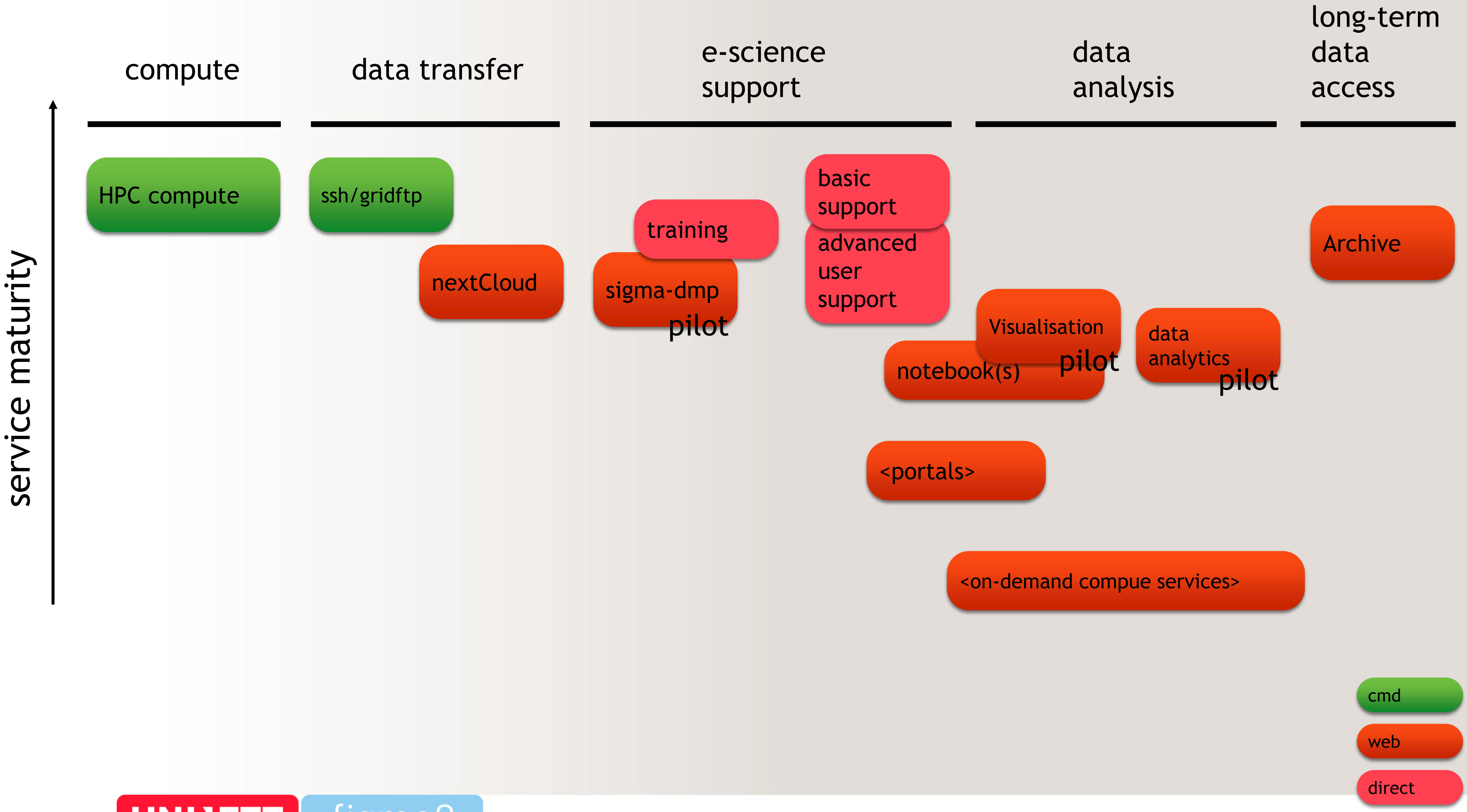
NORSAR 3D, 3D modelling tool

Data-centric architecture



Data intensive science laboratory
local data storage facility
Sensitive data services
NorStore Archive
Data analytics
Visualisation
Data processing
National data infrastructure
HPC
TSD
On demand compute
Visualisation and application resources

Services



compute

data transfer

e-science support

data analysis

long-term data access

service maturity

HPC compute

ssh/gridftp

nextCloud

training

sigma-dmp pilot

basic support

advanced user support

Visualisation pilot

notebook(s)

data analytics pilot

Archive

<portals>

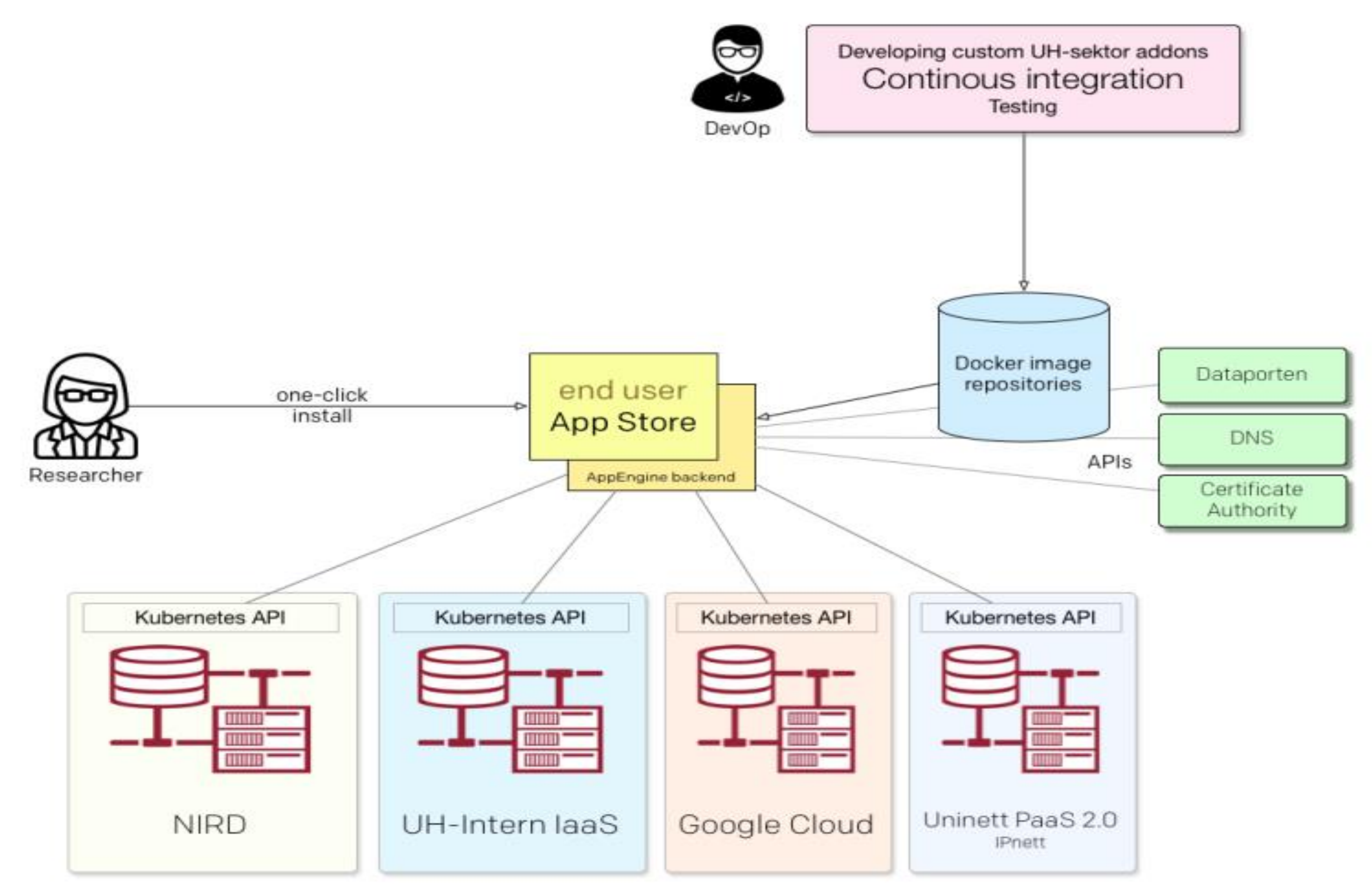
<on-demand compue services>

cmd

web

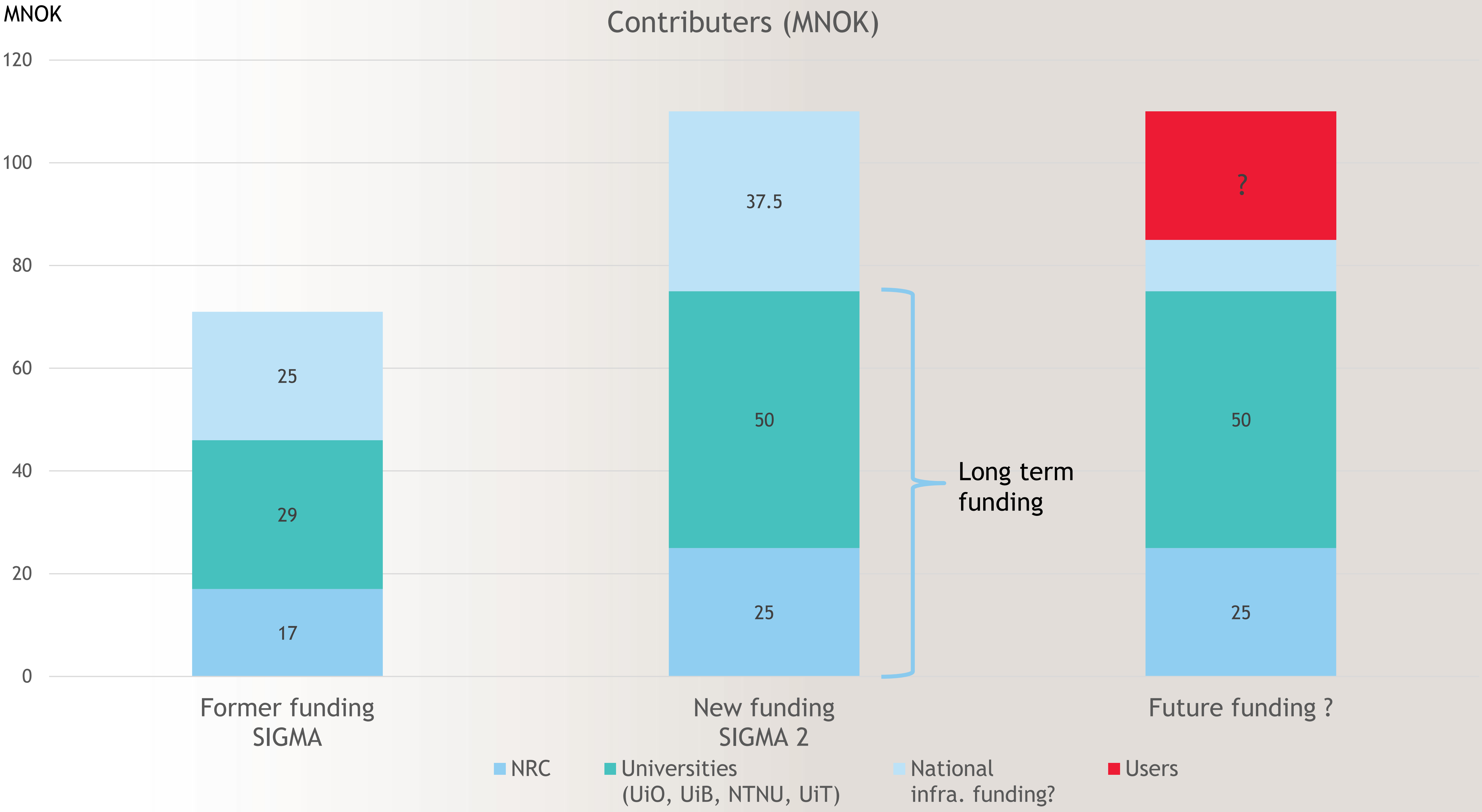
direct

A future common architecture?



www.sigma2.no

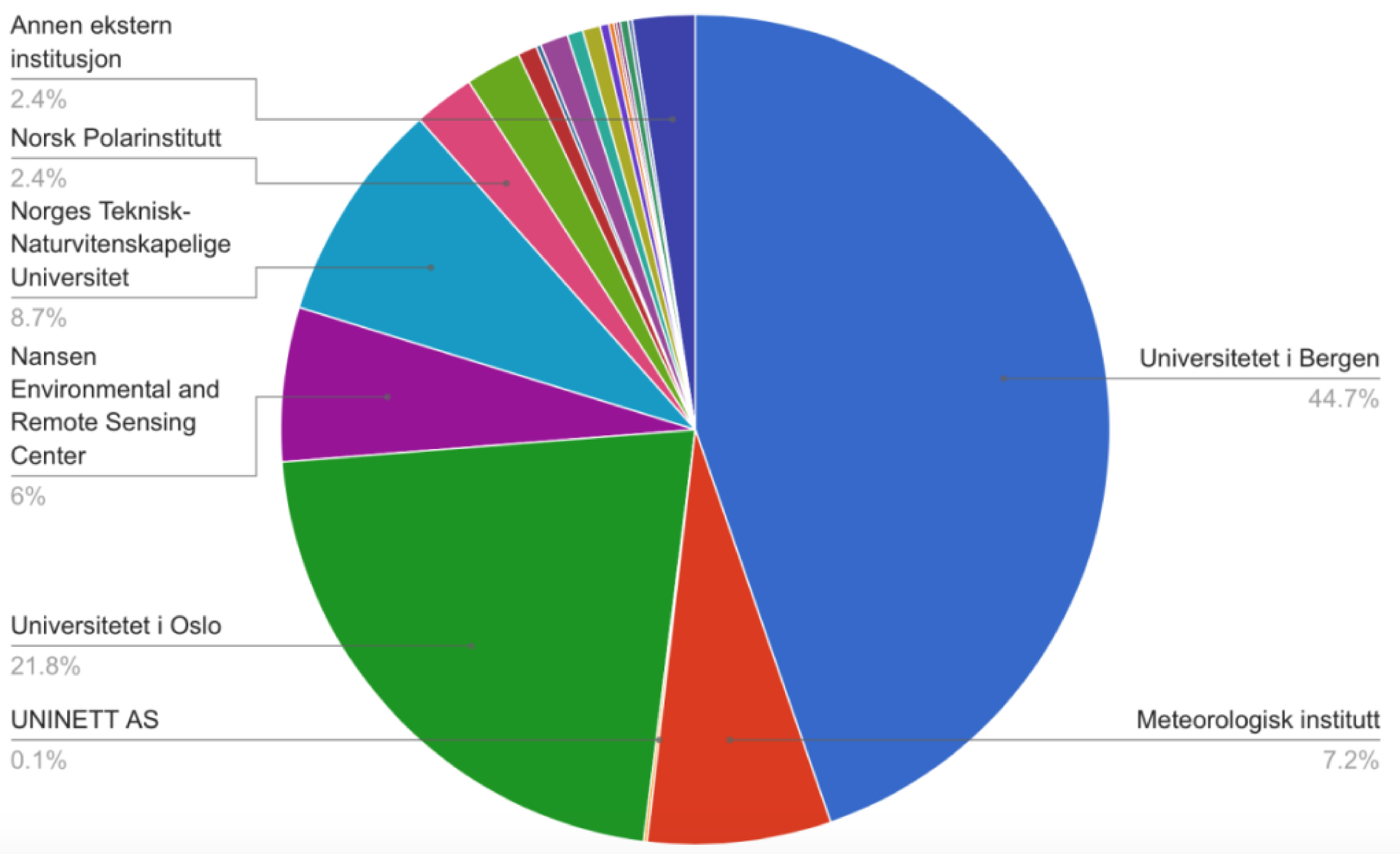
Yearly funding



High level objectives

- Procure, operate and develop a critical national e-infrastructure for researchers
- Promote e-infrastructure to new research communities
- Lead and coordinate participation in international cooperation for e-infrastructure
- Provide an attractive and sustainable e-infrastructure for all research communities, with the following characteristics:
  - High reliability and availability
  - Cost effectiveness
  - Predictable access
  - Interoperability within the national e-infrastructure (Notur/NorStore) and between national and international infrastructures (e.g. PRACE, EUDAT)
- Provide services for data analytics of large datasets (Big Data)

Quota per institution (disk+tape) for 2016



Compute - Technical details

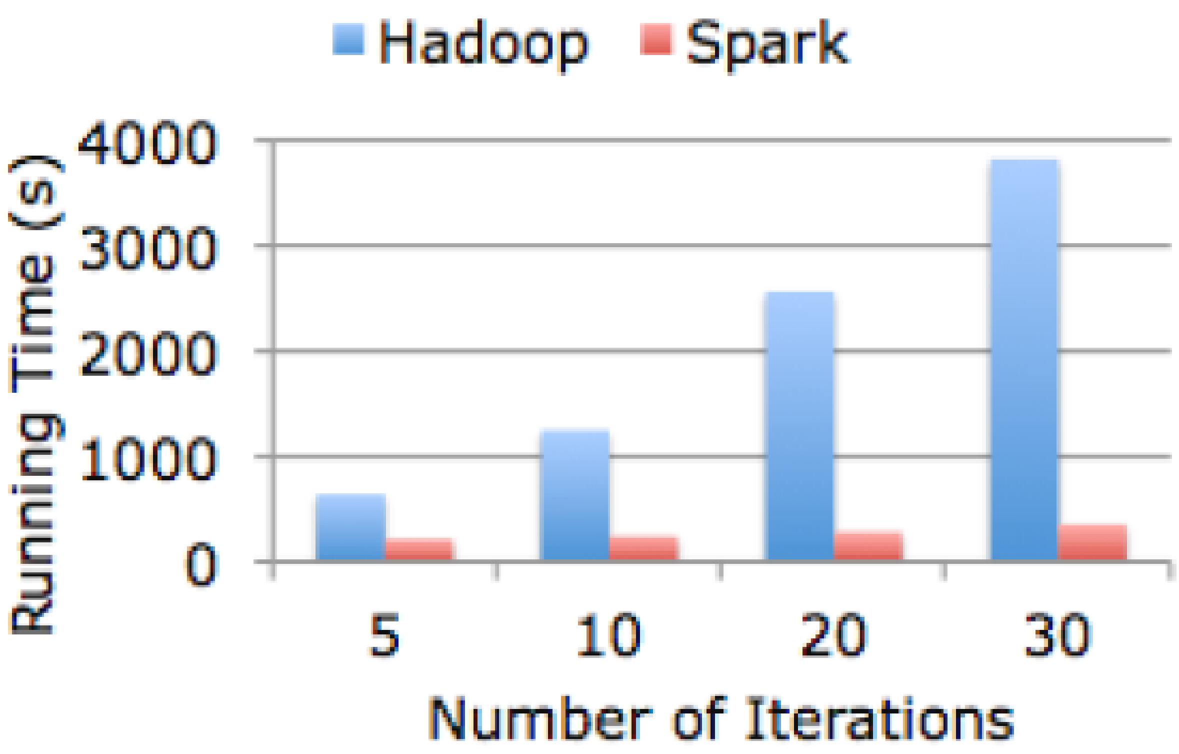
The current national hardware infrastructure is served by 4 sites:

- Abel@UiO: 8723 cores, 16-core nodes w/4 GiB/core except for 8 bigmem nodes w/32 GiB/core Intel Sandy Bridge
- Hexagon@UiB: 11736 cores (of 22272) w/1GiB/core AMD Opteron
- Stallo@UiT: 8896 cores (of 14116), w/2 GiB/core except for 32 bigmem nodes w/8 GiB/cores, 4864 Sandy Bridge and 4132 Ivy Bridge
- Vilje@NTNU: 12901 cores (of 22464) w/2 GiB/core Sandy Bridge
- 16 nodes w/ 2x Nvidia K20x , total 32 GPUs (Abel)
- 4 nodes w/ 2x Intel Xeon Phi 5110P, total 8 MICs (Abel)

Data-analytics (Big data)

- Low demand so far
- Technology used in other services from UNINETT
- Testing (Spark, replacing Hadoop) in cooperation with
  - St.Olav hospital/NTNU (Protein and Genomic analysis)
- Other use cases:
  - Computational Linguistics (common Crawl dataset (500 TB))
  - Fish genomics
  - EISCAT data
- Requirements related to a possible service will be assed

Spark*



Resilient Distributed Datasets:

distributed memory abstraction
fault-tolerance

Functional beauty such as:

lazy evaluation

REPL: Scala, Python, R(coming:)

Uninett has it and wants us to use!

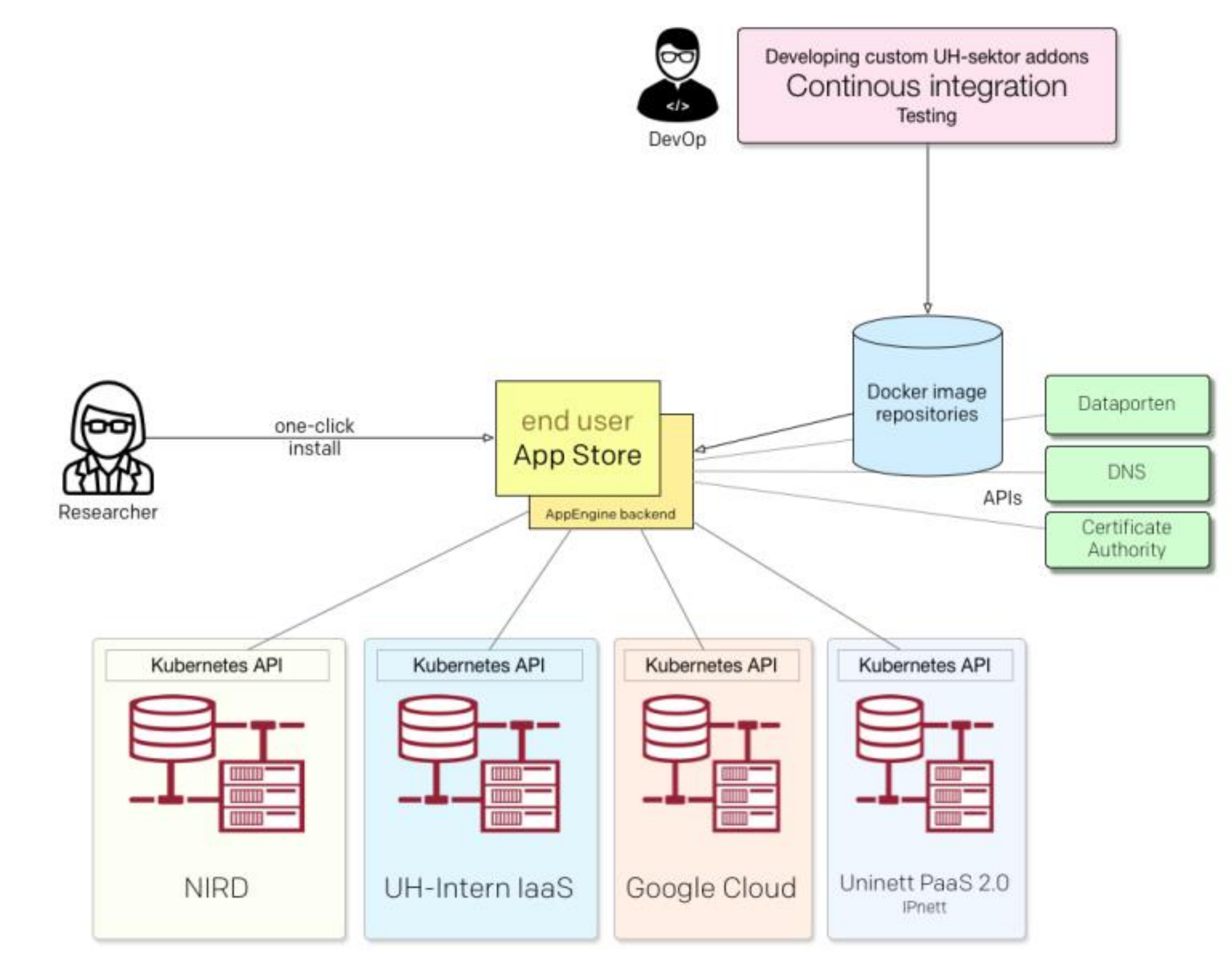
3M: 2 HDFS,1Mesos
15W: 4core@1.87GHz, 20GB, 6TB 7200RPM SATA
https://scm.uninett.no/gurvinder.singh/spark_apps

*https://amplab.cs.berkeley.edu/projects/spark-lightning-fast-cluster-computing/

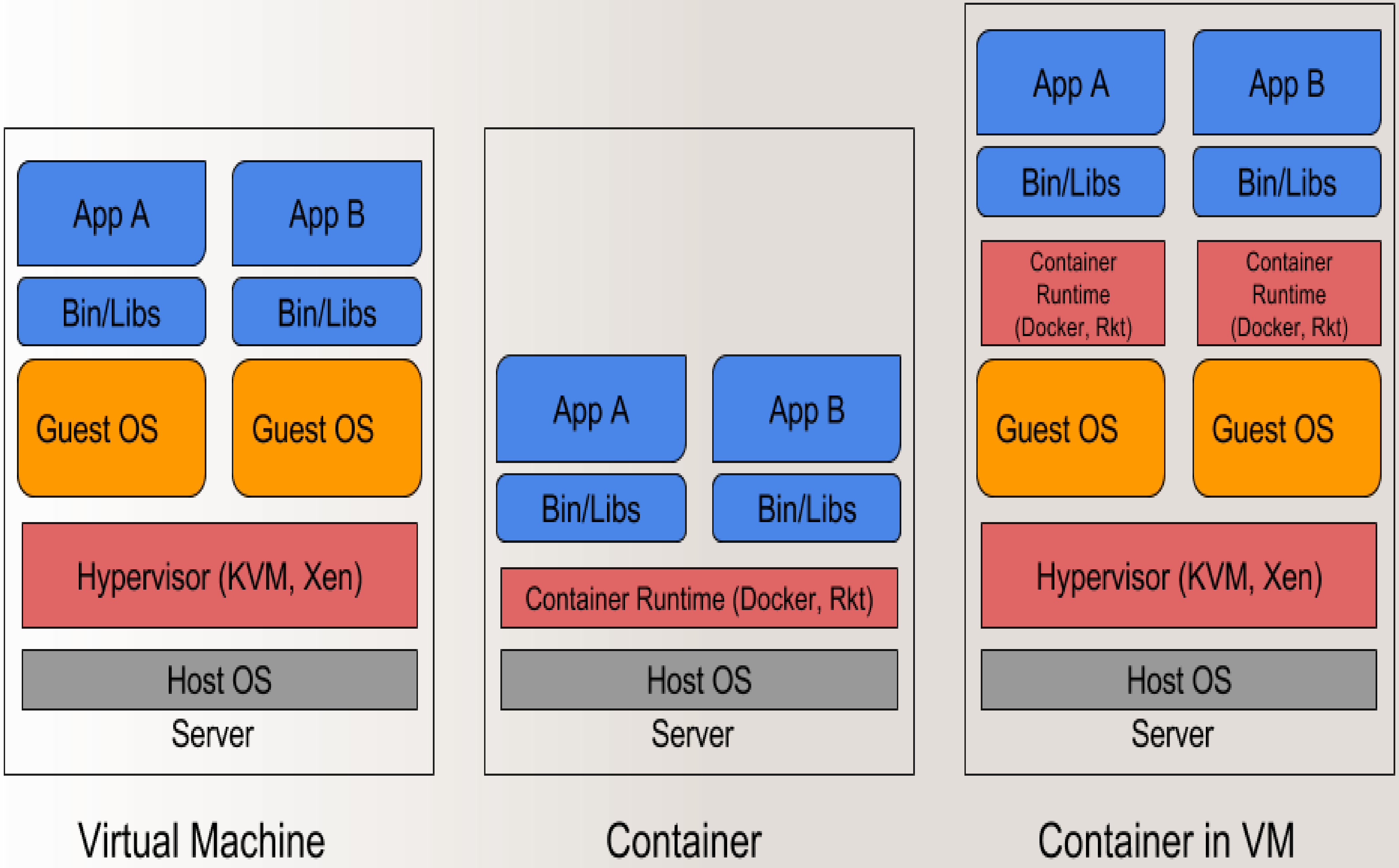
Picture from presentation by
Animesh.Sharma@ntnu.no,
researcher at St.Olav hospital

What data where?

A future common architecture?



Possible implementations



Extra...

Cloud services?

- *«A **cloud service** is any resource that is provided over the Internet.»* V

National Institute of Standards and Technology:

- Self service? X
- On demand? X
- Elasticity (rapid)? X
- Resource pooling? X
- Network Access? V

The Norwegian e-infrastructure

System	Type	Capacity (cpu core hours)	Performance(Tflops)	CPU Cores
Abel (UiO)	Capacity	76 413 480	181,6	8 723
Hexagon (UiB)	Capability	102 807 360	107,9	11 736
Vilje (NTNU)	Capability	113 012 760	268,3	12 901
Stallo (UiT)	Capacity	78 804 960	195,7	8 896
A1 (UiT)	Capacity	262 800 000	1 100	31 150
Total i 2017		418 mill CPU core hours		

NorStore Block storage 3 260 TB + 4 PB tape