BIG data
big problems
big opportunities

ESRF

ERF

Rudolf Dimper

Head of Technical Infrastructure Division – ESRF

- ✓ 6 GeV, 850m circonference Storage Ring
- ✓ 42 public and CRG beamlines
- ✓ 6000+ user visits/y
- ✓ ~1000 experiments/y
- ✓ ~1.5-2 PB/y
- ✓ ~2000 publications/y

What is "Big data"?

"A volume of data that is impossible to process by simple means, hence requiring significant investments in IT infrastructure to capture, store, transfer, analyse and visualise datasets."

A Light for Science
ESRF news
Number 65 December 2013
Big data in focus
New lattice shapes light-source scene
Superconductor predicted from first principles
ESRF

"Big data" – one buzzword … different challenges

Industry

- Data mining
- Business intelligence
- Get additional information from (often) already existing data
- Data aggregation
- New field to make money

Science

- **Handling huge amounts of data**
  - Data transport
  - Distributed data sources and/or storage
  - (Global) data management
  - Data preservation
- **Analysing huge amounts of data**
  - Complexity of code
  - Parallel architectures
  - Different software environments
  - Scientific results must be verifiable

Data → BIG data

More photons

- ESRF upgrade → new lattice → lower emittance → more photons
- Shorter experiments

Optimised experiments

- More automation
- Multiple detectors

Better detectors

- Higher resolution
- Faster readout, less dead-time
- New experimental methods become possible
- Single experiments now **TBs and 100 000's of files**

Example: Computed Tomography

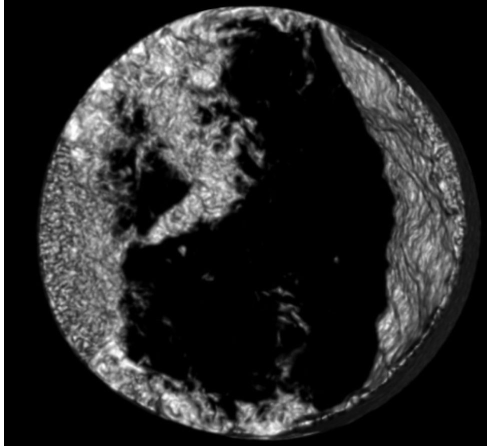
- X-ray computed tomography (CT) is an imagine technique to produce cross sectional images (previously also known as CAT scans (Computed Axial Tomography)
- Used of diagnostics and therapy purposes
- Many slices form a volume
- CT is known as a moderate- to high-radiation diagnostic technique

The Grenoble team in the control room of the medical research beamline. From left to right: Paola Coan, Alberto Bravin and Emmanuel Brun. Credit ESRF/Blascha Faust.

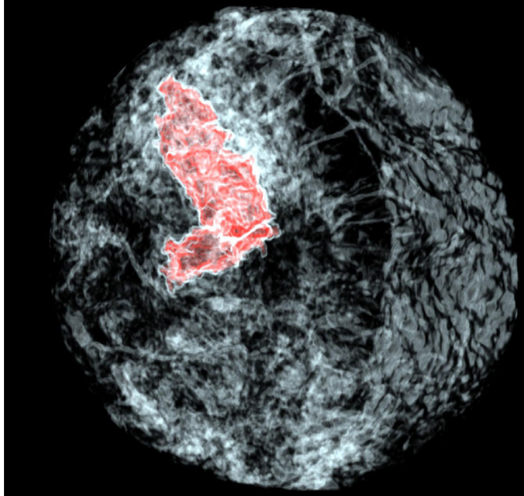
CT scans – Breast Cancer

- **3D – diagnostic computed tomography**
  - The typical dual view digital mammography is limited and does not detect 10-20% of breast tumors
  - Hospital CT scans can not be used – radiation dose too high
- **Synchrotron CT scans:**
  - High energy X-rays
  - Phase contrast imaging
  - Novel mathematical algorithms: “equally sloped tomography”
  - Spatial resolution 2-3 times higher than in a hospital
  - Radiation dose 25 times lower

In the US alone an estimated 40 000 persons/year die of breast cancer!



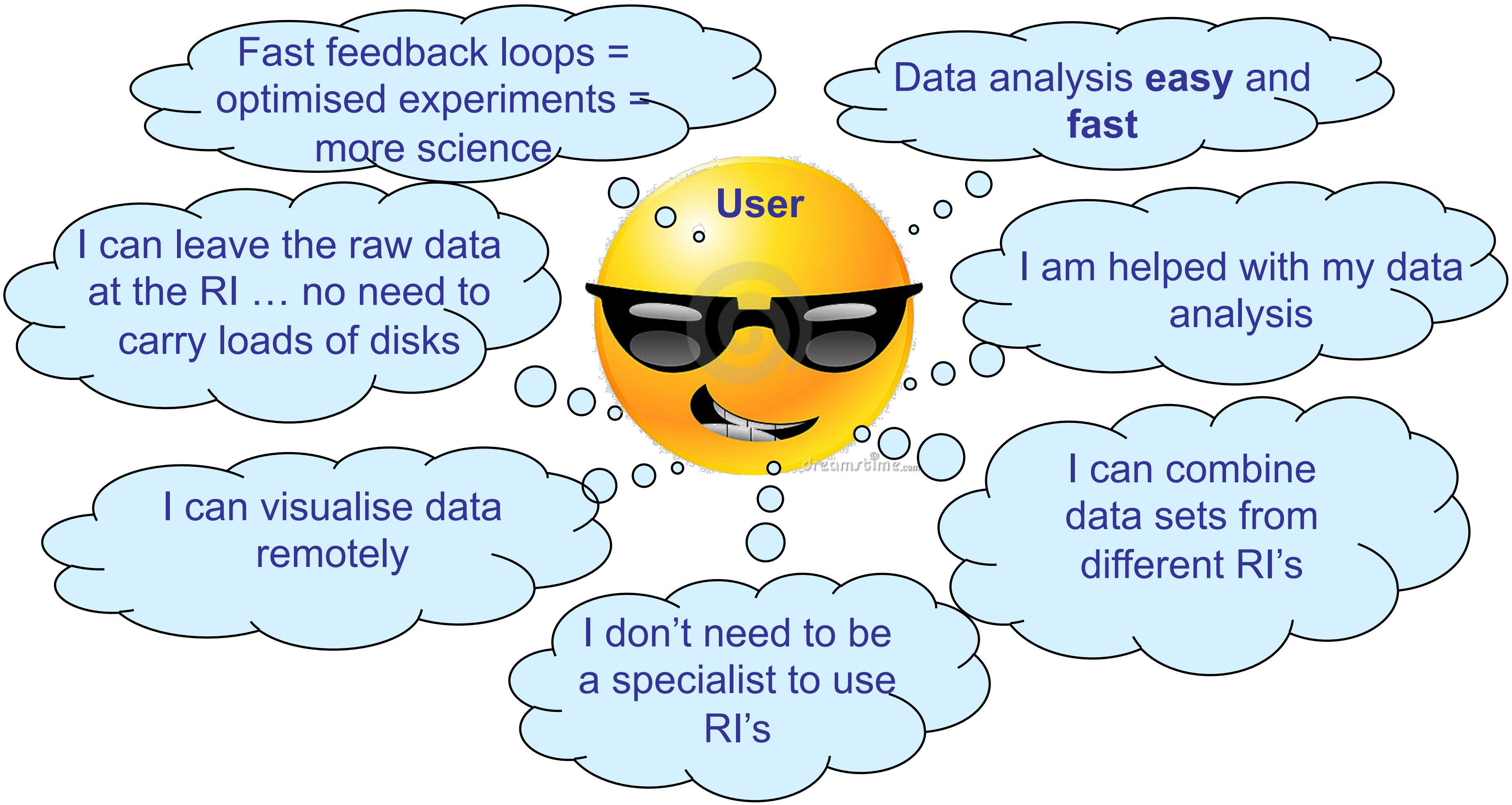
Conventional CT
Dose : 49±1 mGy



Phase Contrast CT
Dose : 2.0 ± 0.1 mGy

Wouldn’t it be cool if …

…we would have unlimited, easy to use IT in our facilities … →



Fast feedback loops = optimised experiments = more science

Data analysis **easy** and **fast**

I can leave the raw data at the RI … no need to carry loads of disks

I am helped with my data analysis

I can visualise data remotely

I can combine data sets from different RI’s

I don’t need to be a specialist to use RI’s

IT – a barrier?

- **Difficult to scale up**

- **Required bandwidth for detector read-out**
- **On-line or near on-line data analysis while taking data**

  Data analysis requires low latency IT
- **Automated metadata capture required**
- **Data management a challenge**

  Large number of individual files

IT – a barrier?

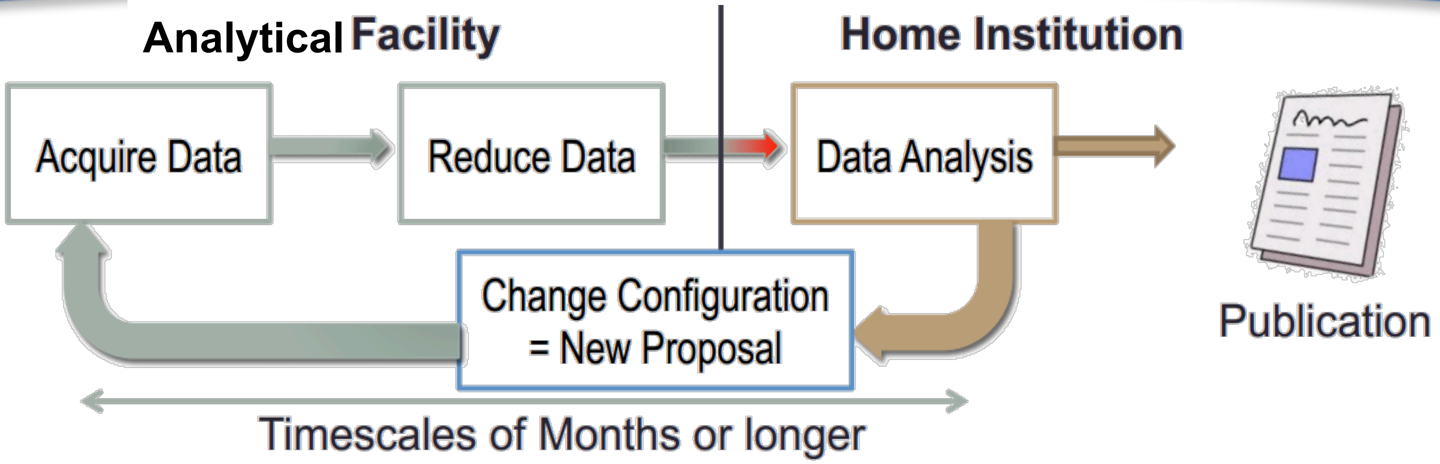
- **Difficult to scale up**

- **Data sets too big to take home**
- **Software environment a challenge**
  - Complexity, heterogeneity
  - Users usually not affiliated to the facility
  - Less IT literate scientists than e.g. in HEP
  - Users require support to install software and analyse data
- **Scientific collaborations require distributed infrastructures and networks**

Data Analysis THE barrier

- ✓ **We must match data analysis capabilities with advancements in detectors and sources**
- ✓ Since resources are scarce, we have **to do this collaboratively**
  - Create a homogeneous and compatible data analysis environment
  - Provide data analysis services to our users
  - Share best practices and solutions
  - Pool our resources between Research Infrastructures to create critical mass
  - Build sustainable solutions

Moving the barrier



Analytical Facility

Home Institution

Acquire Data

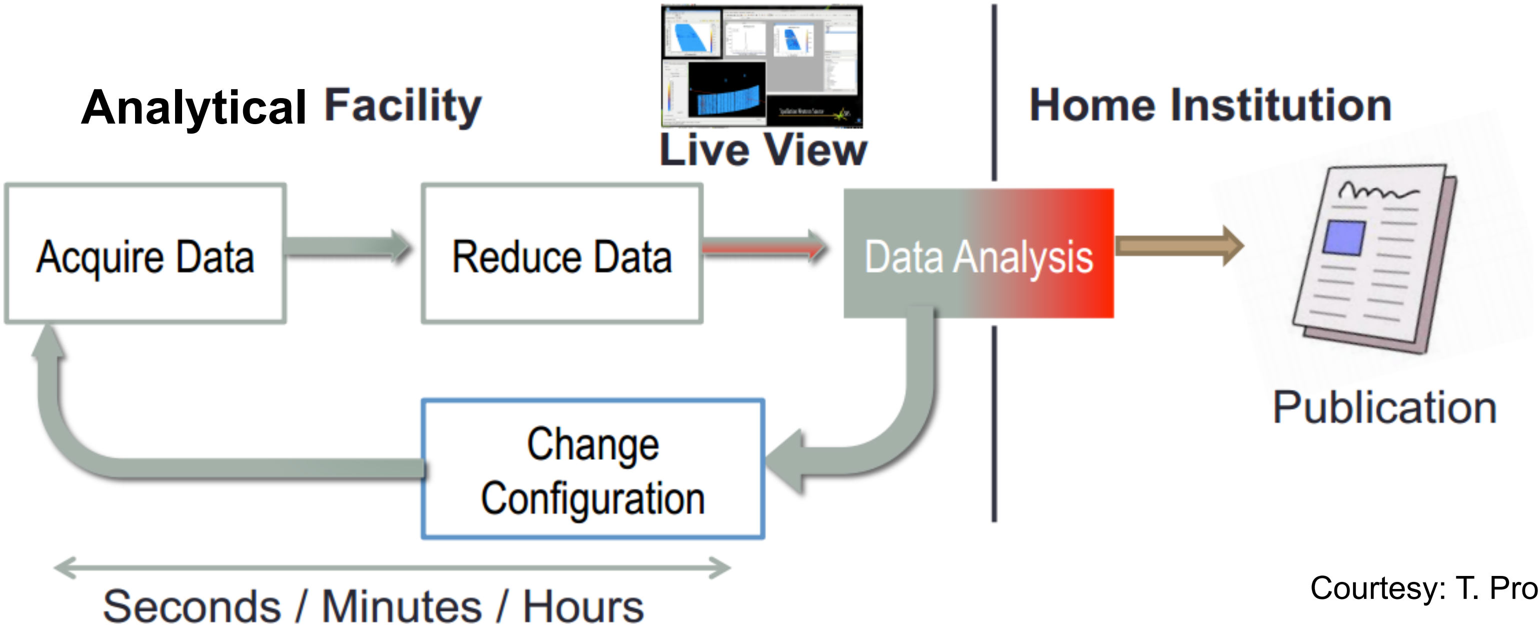
Reduce Data

Data Analysis

Publication

Change Configuration = New Proposal

Timescales of Months or longer



Analytical Facility

Live View

Home Institution

Acquire Data

Reduce Data

Data Analysis

Publication

Change Configuration

Seconds / Minutes / Hours

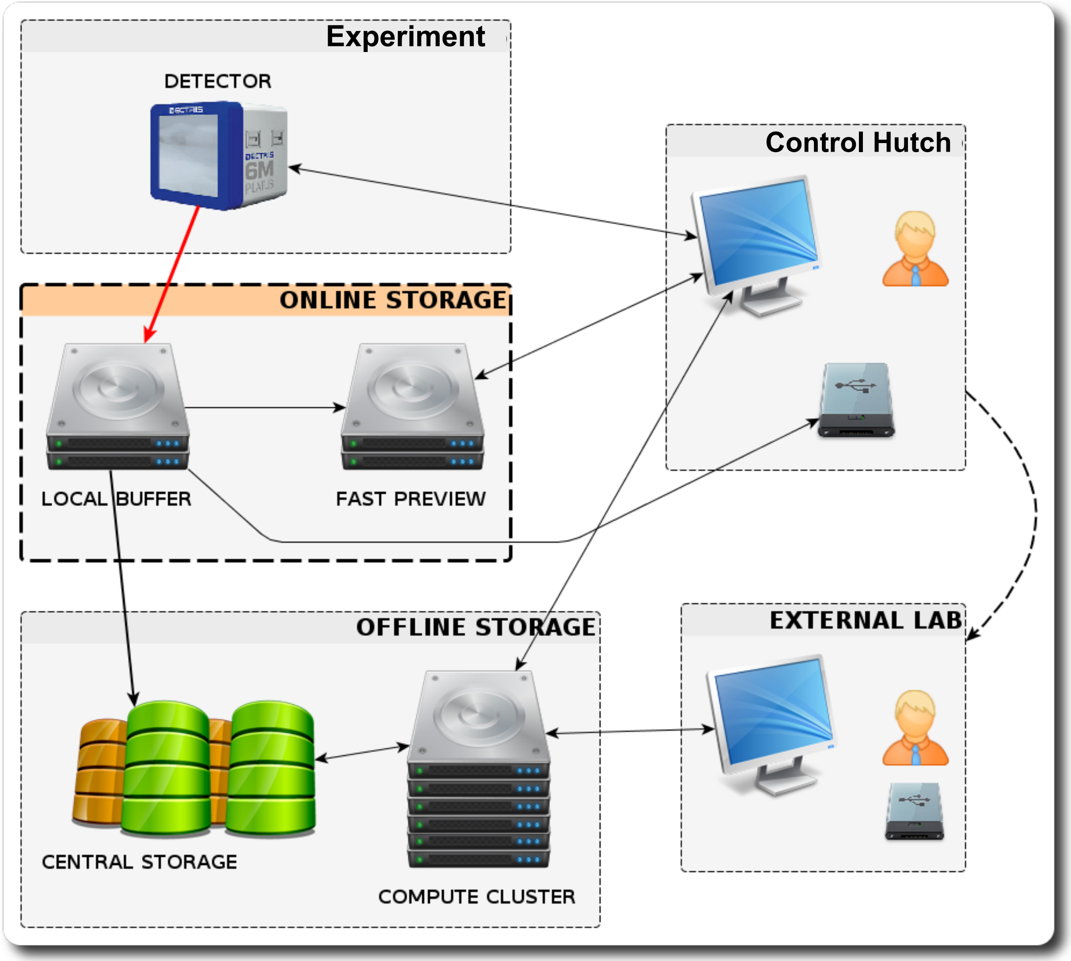
Courtesy: T. Proffen

Data acquisition

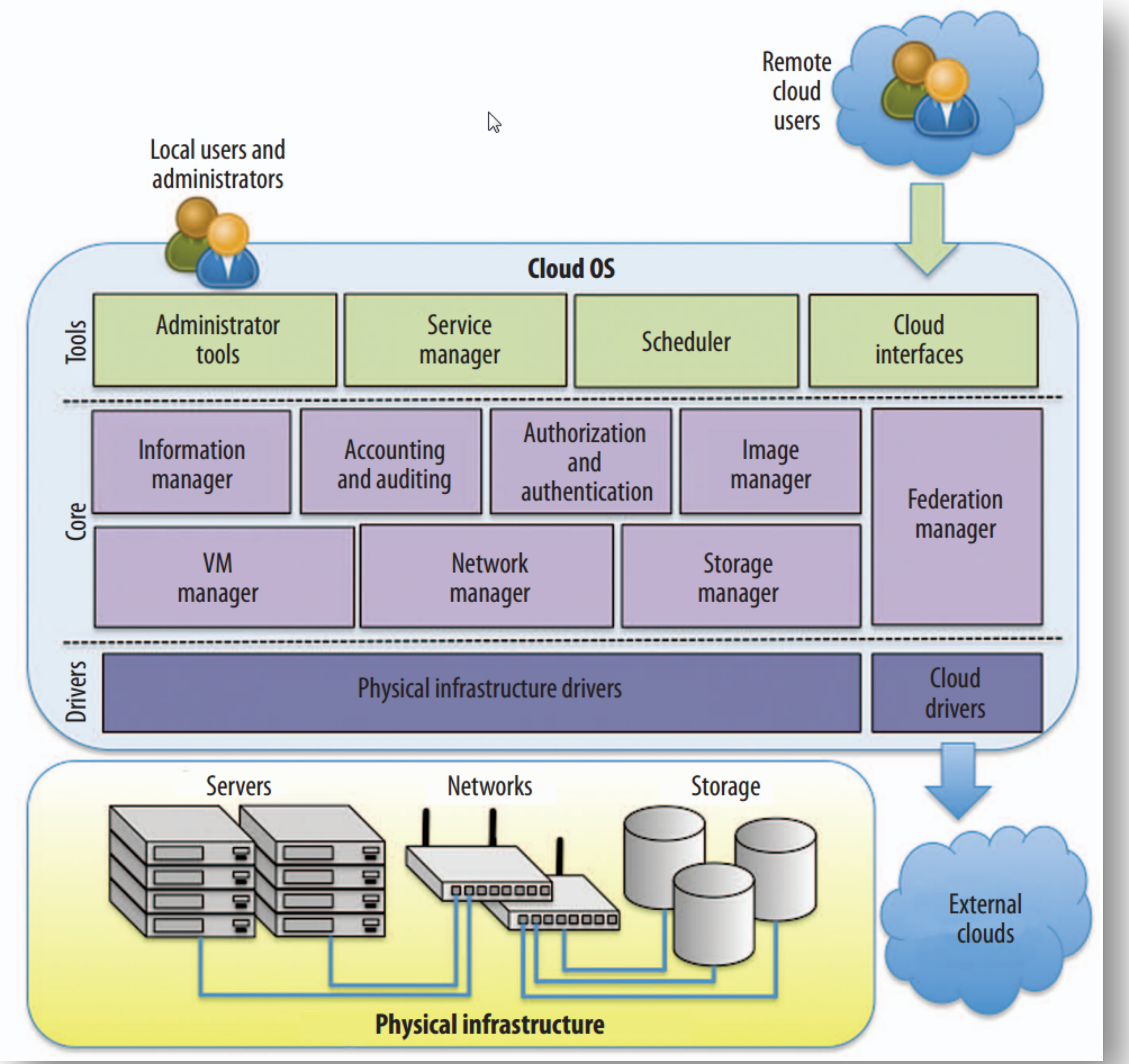
- ✓ Fight against latency
- ✓ Guarantee bandwidth

Example: Fast buffer storage:

- ✓ fast preview / on-line data analysis
- ✓ 2 days capacity
- ✓ central storage push
- ✓ multiple 10 Gbps
- ✓ NFS (V4+V3) and CIFS
- ✓ write >1 GB/s
- ✓ read ≥ write speed



Reaching for the clouds



Why Cloud?

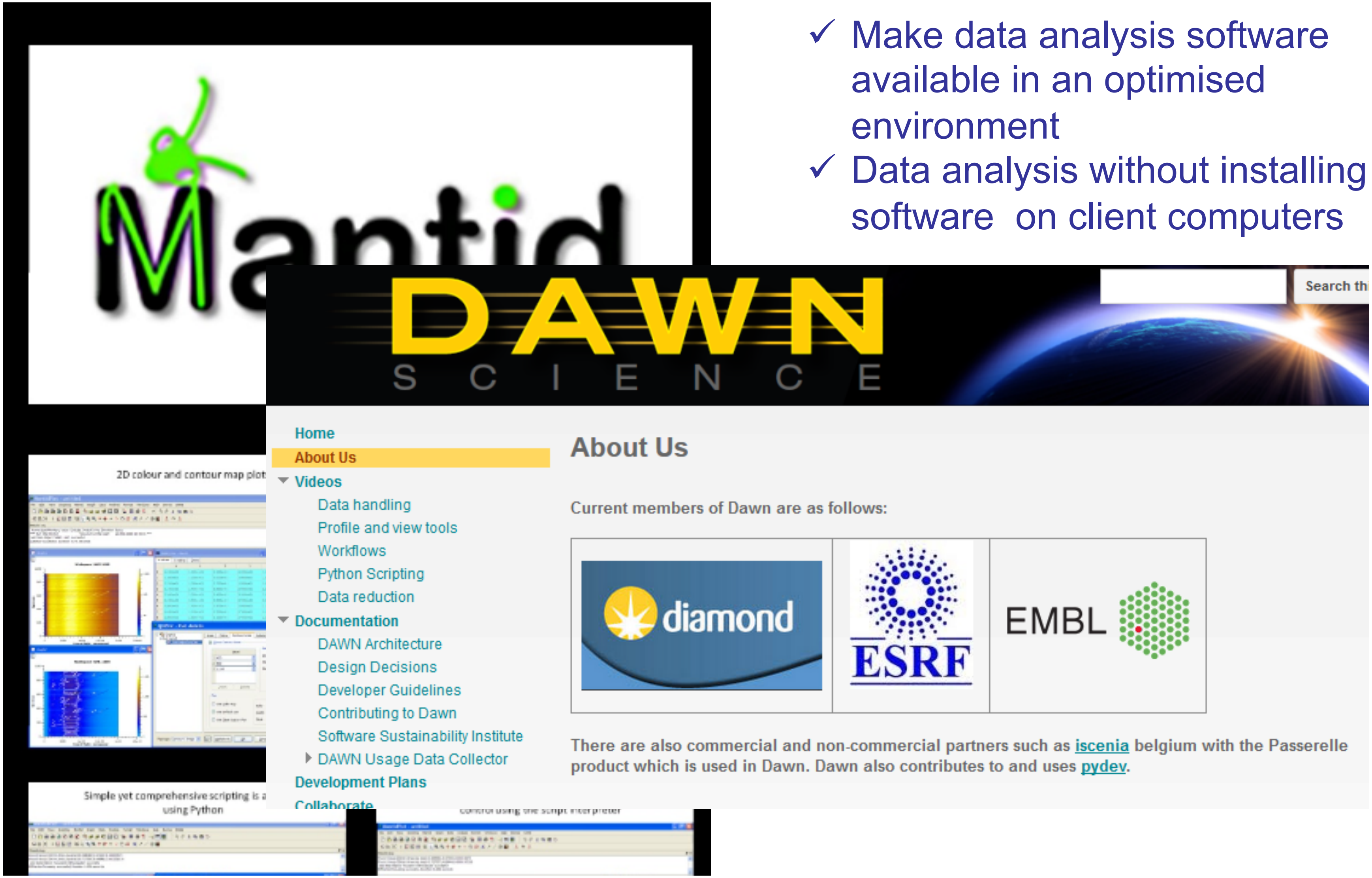
- ✓ Easy to use
- ✓ Economies of scale through standardisation
- ✓ Powerful software models
- ✓ Flexibile resource allocation

Complexity for simplicity

- ✓ Virtualisation
- ✓ Cloud platform to encapsulate services
- ✓ An all-in-one environment to implement “Data Analysis as a Service”
- ✓ Federation of local clouds to span RIs in Europe
- ✓ Open to public clouds, allowing to choose in the future

Integrate Data Analysis Code

- ✓ Make data analysis software available in an optimised environment
- ✓ Data analysis without installing software on client computers



Mantid
2D colour and contour map plot
Simple yet comprehensive scripting is a
using Python
control using the script interpreter
DAWN
SCIENCE
Search th

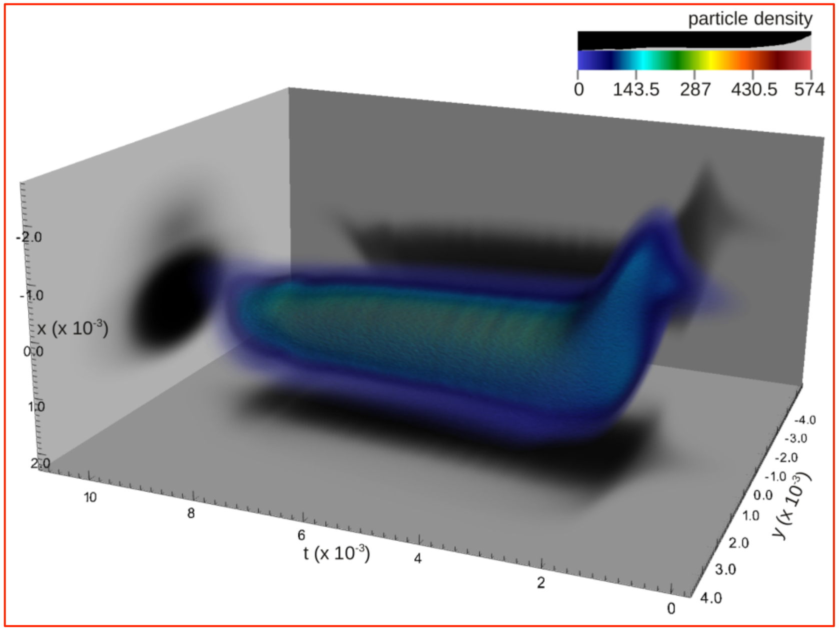
About Us

Current members of Dawn are as follows:

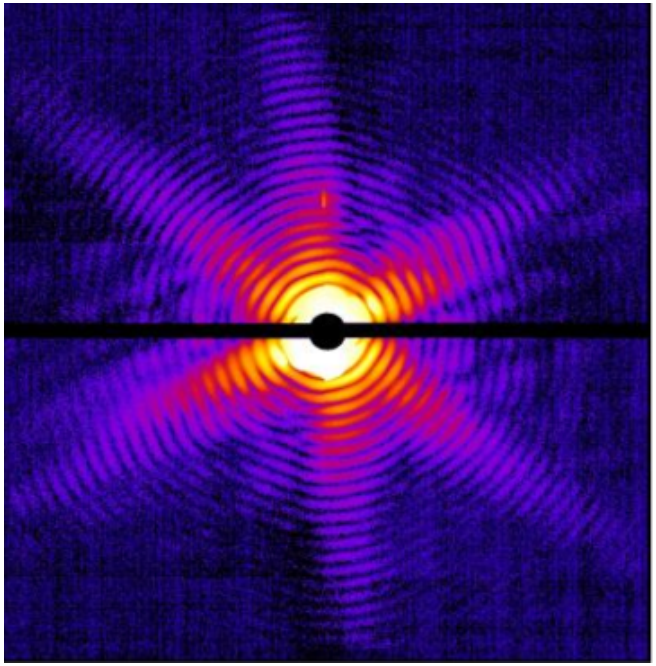
There are also commercial and non-commercial partners such as iscenia belgium with the Passerelle product which is used in Dawn. Dawn also contributes to and uses pydev.

Data visualisation

- Browsing and searching through meta-data
- Browsing through experimental data
- Multi dimensional web-based (remote) visualisation of large data sets



High-resolution simulations of beam dynamics in electron linacs, particle density in physical space. LBNL visualization group



X-ray diffraction pattern of a single Mimivirus particle, imaged at LCLS

Photon and Neutron Data Analysis As A Service

PaNDAAS

a
t
a

elettra

eli attosecond

ill NEUTRONS FOR SCIEN

DESY

SPALLATION NEUTRON

THE INS

ESRF The European Synchrotron

PAUL SCHERRER INSTITUT PSI

diamond

opean FEL

ISIS

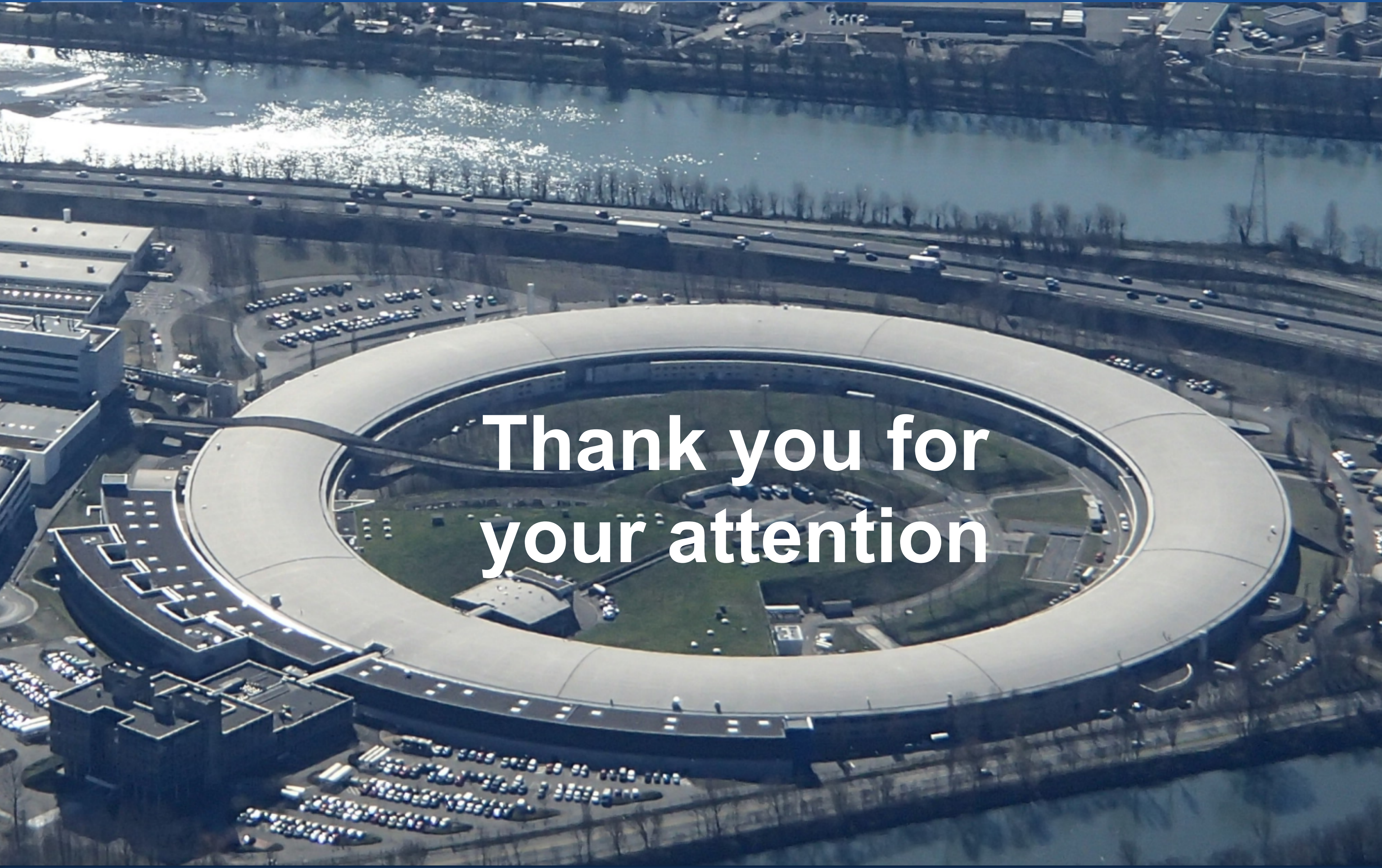
SOLEIL SYNCHROTRON

MAX IV LABORATORY

AME

ESFRI European Strategy Forum on Research Infrastructures

European Commission



Thank you for your attention