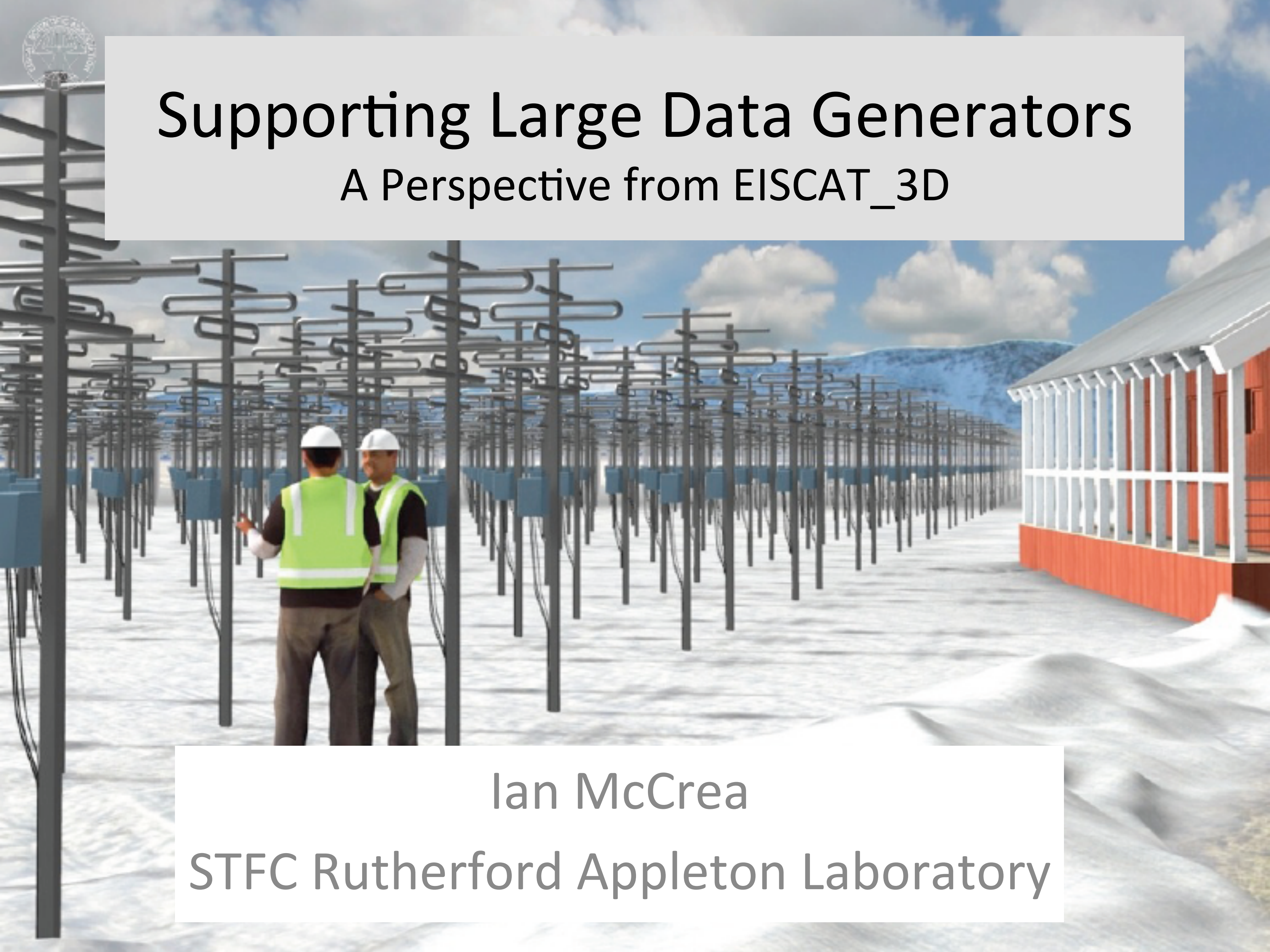


Supporting Large Data Generators

A Perspective from EISCAT_3D

Ian McCrea

STFC Rutherford Appleton Laboratory

EISCAT Scientific Association

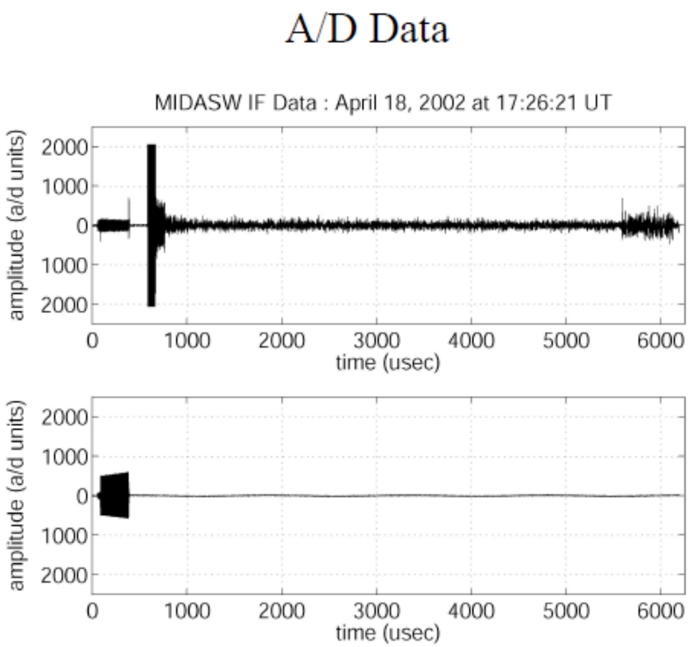
Associate countries and institutes

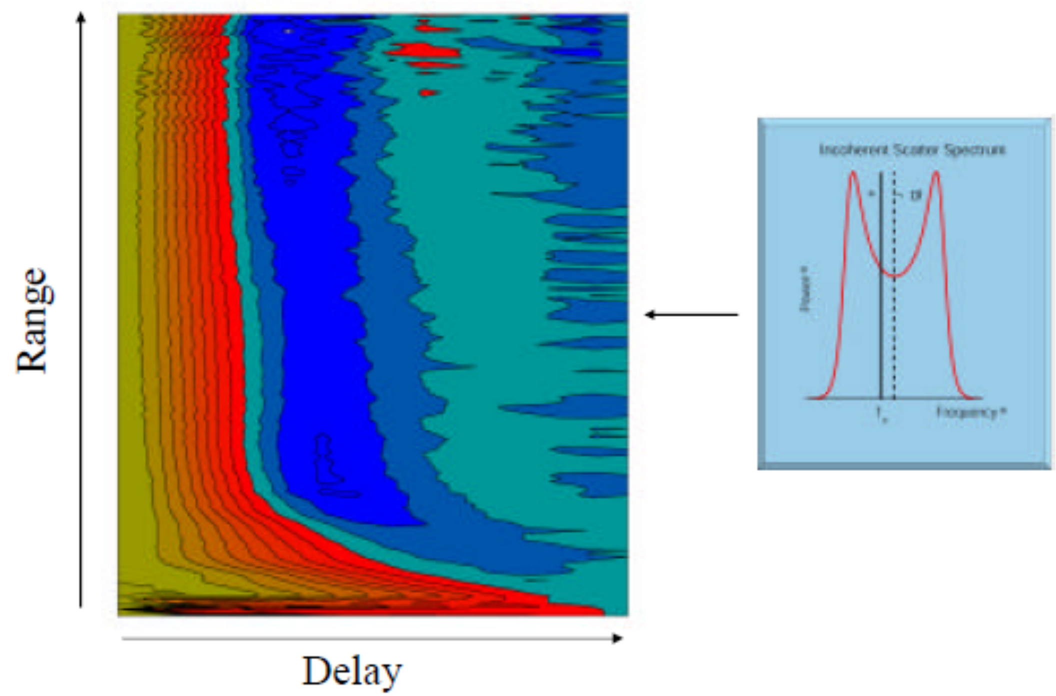
Contributing:

EISCAT Incoherent Scatter Radars

EISCAT Svalbard Radar

Data Types within EISCAT





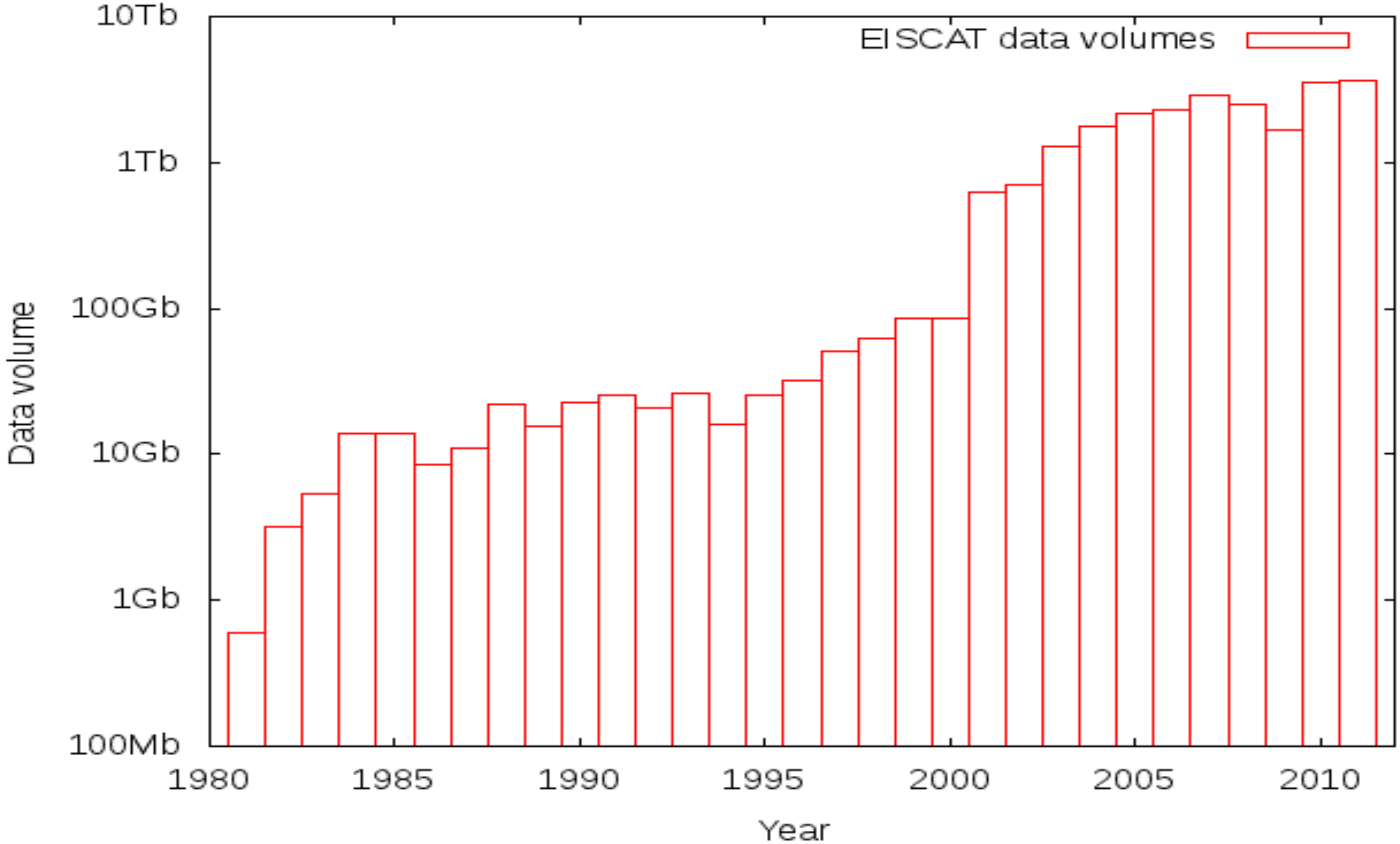
Voltage-level data. Two (I&Q) streams of 16-bit samples per sensor, separation ~ 1 μs, repetition rate ~ 1ms.

Semi-stochastic. Cannot be integrated.

Lag Profile Matrices formed by integrating lagged autocorrelation functions formed from the voltage level data.

Multiple heights, multiple lags, but can be time-integrated

The current EISCAT system is not a high-volume data generator…



EISCAT: 5-10 years from now

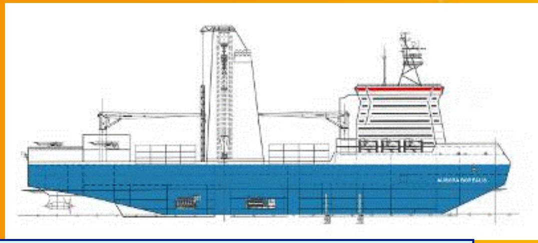
ESFRI Projects for Env. Sciences

EURO-ARGO

SIOS

IAGOS-ERI

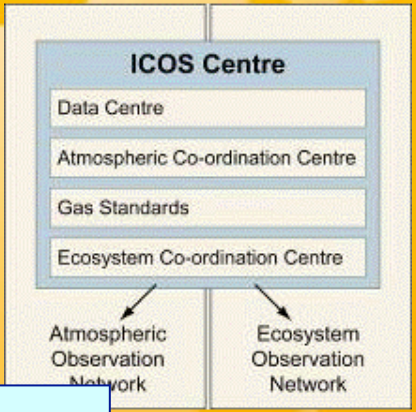
EUFAR-COPAL



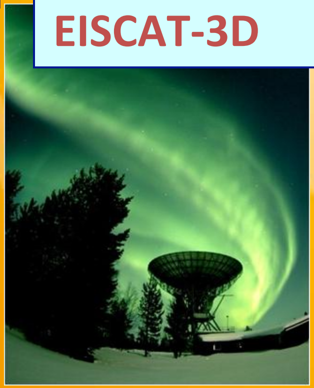
AURORA BOREALIS

Status 2009

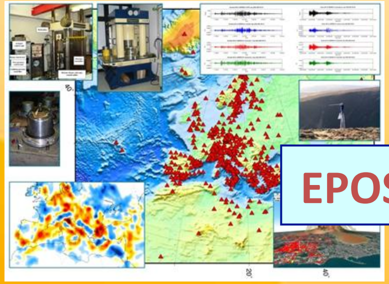
LIFEWATCH



EMSO



EISCAT-3D



EPOS

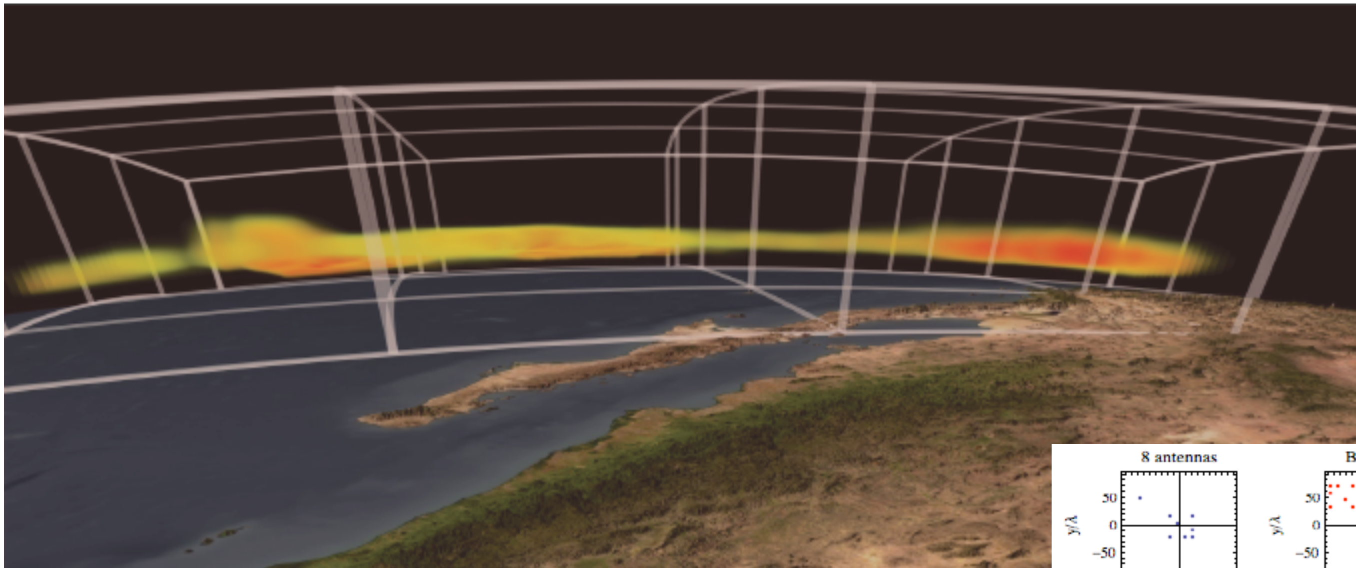
ICOS

EISCAT_3D: Key Capabilities

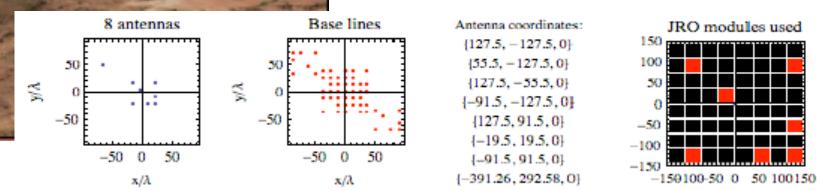
- Key capabilities:
  - Volumetric imaging and tracking
  - Aperture Synthesis imaging
  - Multistatic configuration
  - Greatly improved sensitivity
  - Transmitter flexibility
  - Inter-operable data with other facilities and models

- Similar (in some ways) to radio astronomy

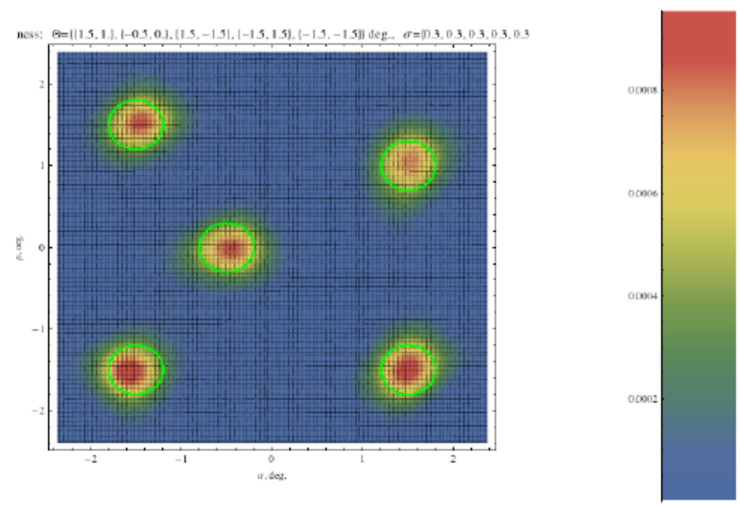
Two Kinds of Imaging



Volumetric

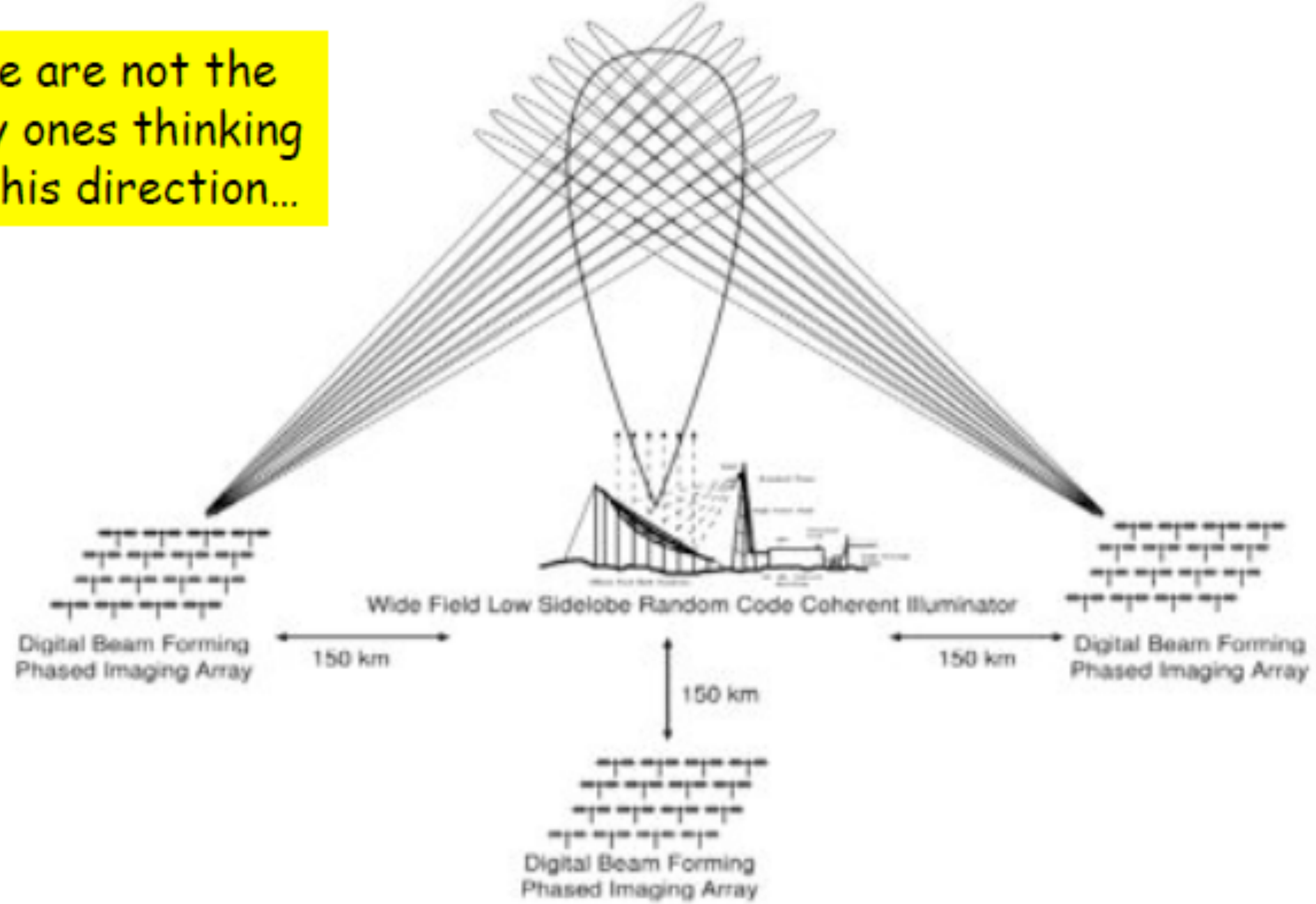


Interferometric



The Holographic Radar

We are not the only ones thinking in this direction…



Volumetric ionospheric imaging using a multi-static CW ISR design
Optimal ionospheric measurement of volume and vector quantities
Computationally intensive (petaops)

(Frank Lind, Millstone Hill)

"Adjusted Double Mercedes"

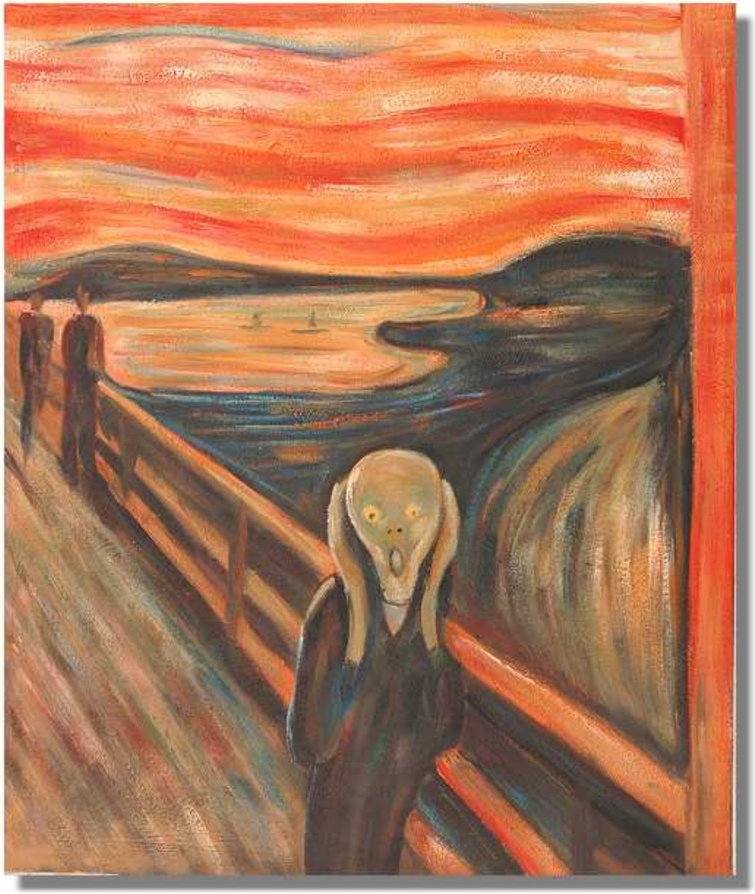
N-S drift **in E-region**

E-W drift **in F-region**


EISCAT SCIENTIFIC ASSOCIATION

Low-Level Data

- Voltage data (lowest level)
  - 80 MHz sampling, 16 bits
  - 2.56 Gb/s/element means 4×10^{13} b/s (!)
  - Combine by group (49 antennas)
  - Then into <10 beams
  - Each beam ~ 25 TB/day
- Beam-formed data: Central site
  - Only one (fast scanning) signal beam
  - Small volume calibration beam(s)
  - Approx 1 TB/hour (320 MB/s)
  - Voltage data, not band limited
- Beam-formed data: Remote sites
  - 5-10 beams, but intersection limited
  - Same order as central site
  - Identical short-term storage at all sites

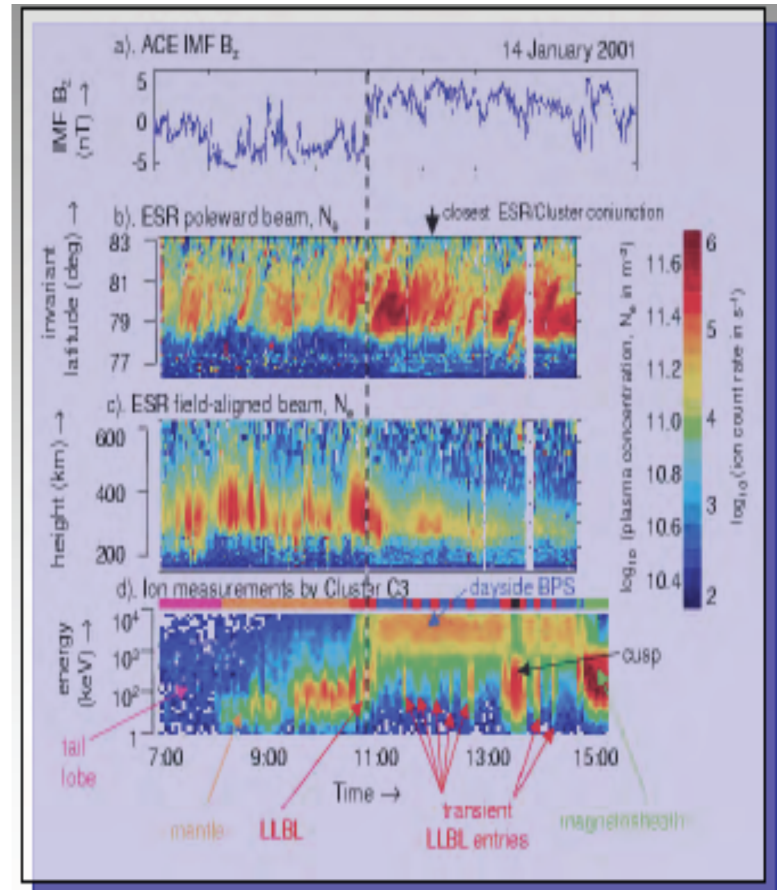


Higher-Level Data

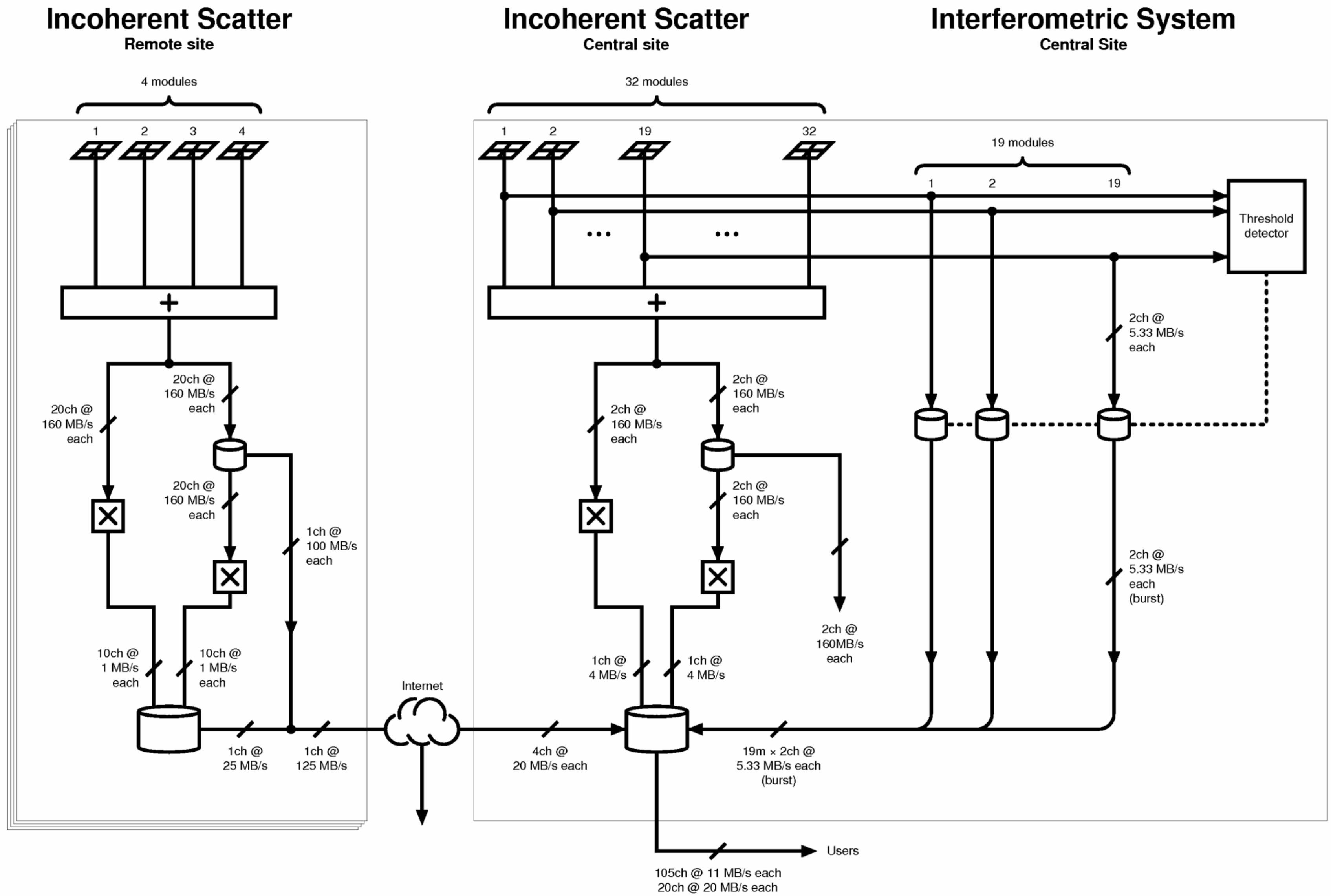
- **Interferometry Data**
  - 19 modules in use (202 MB/s, 17 TB/day)
  - But keep only 5% of samples above threshold
  - Lead-in and follow-on data (tens of GB)

- **Supporting Instruments**
  - Common data network for other diagnostics
  - Optical instruments, other radars
  - Estimated at 150 GB/day at central site
  - 30 GB/day for each remote station

- **Highest-Level Data**
  - Analysed data products (small)
  - Correlation functions ~200 TB/year
  - Maybe not needed...



EISCAT-3D Data System



Network Requirements

- Data transfer from the remotes
  - 1 beam is 320 MB/s, remotes have multiple beams
  - Supporting instruments add ~30% overhead
- Recover from interrupts quickly
  - Otherwise we may never catch up
  - Interrupts might last days/weeks
- Fast links already practical
  - Protocols for 10 GB/s links exist already
  - How should we factor network costs into our plan?
- Back-up if the network fails
  - Something to tell us if the site is alive
  - …and how cold it is……
  - Mobile phone, satellite, microwave link

Lumiora Fibre Network

Archiving Requirements

Ring Buffer

High volume (~100 TB) short duration (hours/days)

Data accumulate constantly , oldest over-written

Records interferometry when events detected

Latent archive data in event of network outage

Interferometry System

Small area (~100 GB), few minutes of data

Data accumulate constantly, threshold tested

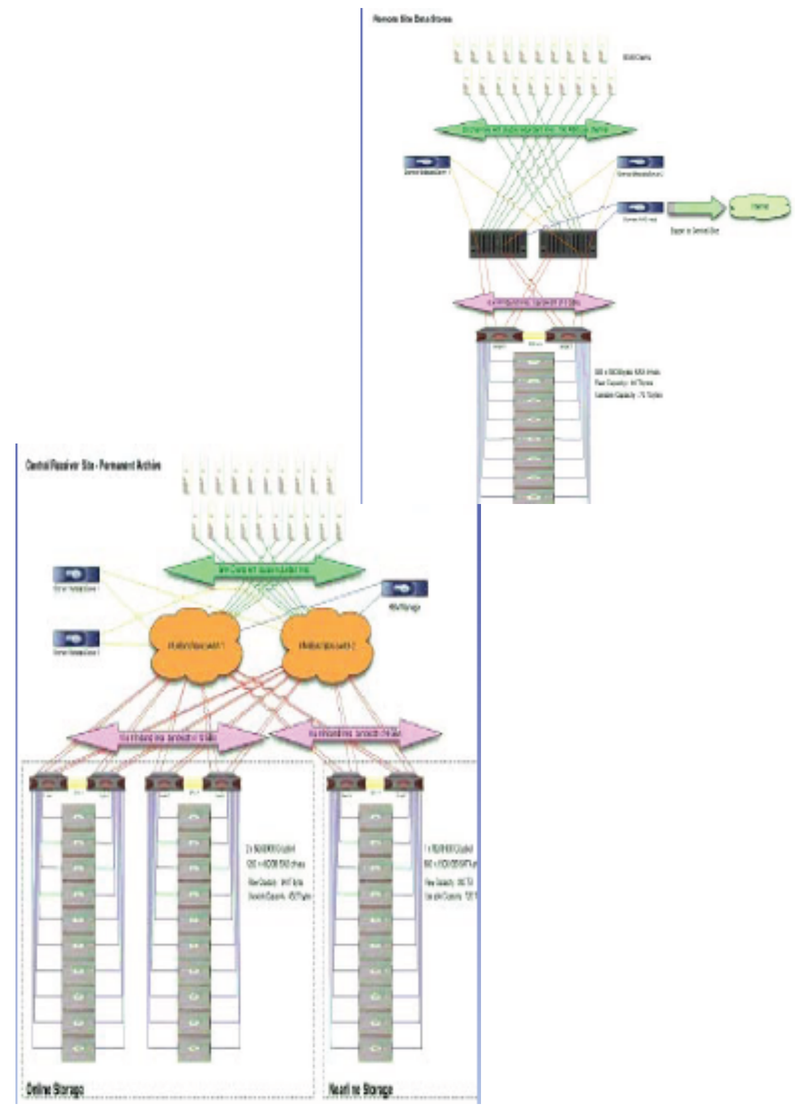
If event detected, divert data flow, otherwise delete

Permanent Archive

Large capacity (~1PB) permanent archive

Mid and high-level data @ 200 TB/year

Tiered storage, connected to multi-user computing



Data Centres and HPC

Cost-effective option is to integrate storage and computing provision into existing e-infrastructure

Commission is encouraging us in that direction (e.g. ENVRI, ESPAS, COOPEUUS)

However, network capacity means that we have to do some level of initial data storage, event detection and data decimation on site, in possibly remote locations

SNIC (HPC2N) is a partner in our Preparatory Phase, helping us to handle these issues

Collaboration also being established with EUDAT via Kimmo Koski (CSC Helsinki)

Preparatory Phase Issues

- EISCAT_3D Preparatory Phase (2010-2014) "Data handling and distribution"
- Networking to sites
- Use of existing fibre networks
- Existing infrastructures for storage and computing?
- Interaction with other ESFRI projects and non-ESFRI RIs with similar needs

_EISCAT_3D_

A European Three-Dimensional Imaging Radar for Atmospheric and Geospace Research

Application for Preparatory Phase Funding under the European 7th Framework

2020 The Geospace Instrument Array

Facility Scale Instruments



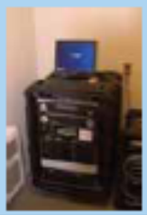

Major Geospace Facilities
Large Radio Telescopes

Medium Scale Instruments



SuperDARN Network
Low Cost IS Radars
Optics Arrays
(moderate numbers)

Small Scale Instruments




Software Radio Arrays
GPS Arrays
All Sky Camera Arrays
Magnetometer Arrays
(large numbers)

World Wide Web

Supercomputing Geospace Assimilation Grid

Geospace Search Engines	Virtual Observatories	Space Weather Models

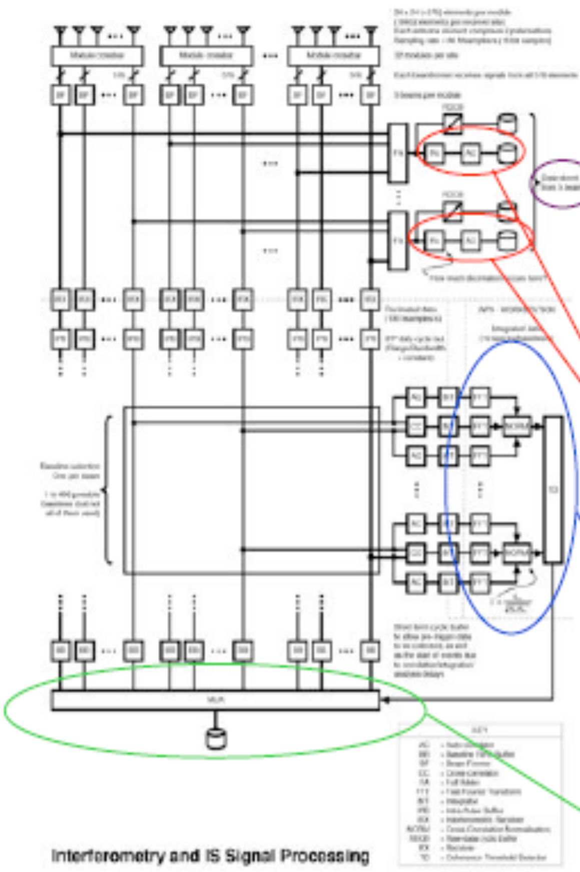
Scientists Educators Public

Summary

- EISCAT_3D has potential to transform ISR science in data handling and flexibility
- It also has the potential to be just like existing radars
- The difference is the supporting e-infrastructure
- Networking plays a big role in helping E3D achieve its potential
- Many similarities to radio astronomy. Would be interested in GEANT/EXPReS case study
- More engagement needed – maybe via EEF as the point of contact?

How much data could EISCAT_3D produce?

More 3D Data Products:



Correlated data

The first "permanent" data product:

Polarisations combined for max SNR,

Sample matrix inflated into lag profiles,

Time-windowing applied at remotes to match signal reception phase of each IPP,

Time integration applied to further reduce data vector size,

Data volumes manageable; e.g., 150 gates per profile @ 50 lags/gate generates about 150-200 MB/hour/beam.

Analysed data

A representative analysed data set will always be generated and stored

Each beam analysed separately

Standard pre-integration

Standard, well documented analysis procedure (GUISDAP),

Well defined analysis strategy

Long term storage (archive)

Volumes about n times now (since n simultaneous beams)

File-based data (easy to access particular dates/experiments)

Relational tables (easier event identification and searching).

Interferometry data

10-15 baseline pairs used in coherence detection,

Threshold logic monitors coherence levels and signals the storage system when predefined thresholds are exceeded,

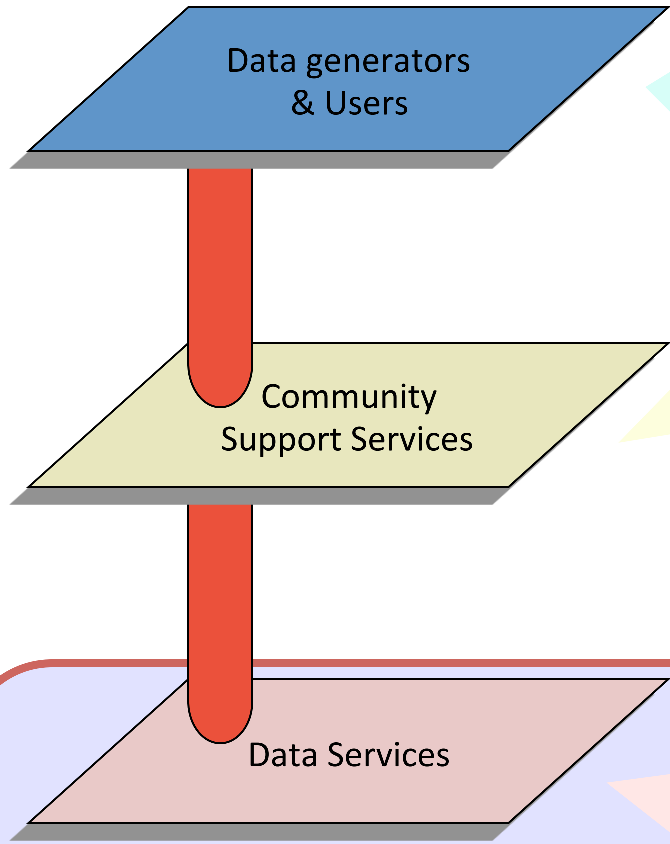
Beam-formed data from each array module used as an interferometer baseline endpoint (decimated to ion line BW) are then written to short-term (ring-buffer) storage,

Interferometry users are automatically alerted and asked to copy the data to their own storage.

Most data products available in near-real time via the Web!

Three layers

(slide by Peter Wittenburg and Wouter Los)



<u>Data providers & Users</u>
<u>Humans & Instruments</u>

<u>Roles</u>

Sensors	Curators	Researchers
Observers	Aggregators	Public

<u>Functionalities</u>

Virtual Environments & Collaborative organisations
Security & Protection

<u>Data discovery & Navigation</u>

(meta) data tagging tools
Data submission tools
Operational Semantic Interoperability

<u>Workflow Generator</u>

Data correlation
Knowledge management
Virtualisation

<u>Persistant storage capacity</u>

24/7 operation
Preservation & Sustainability

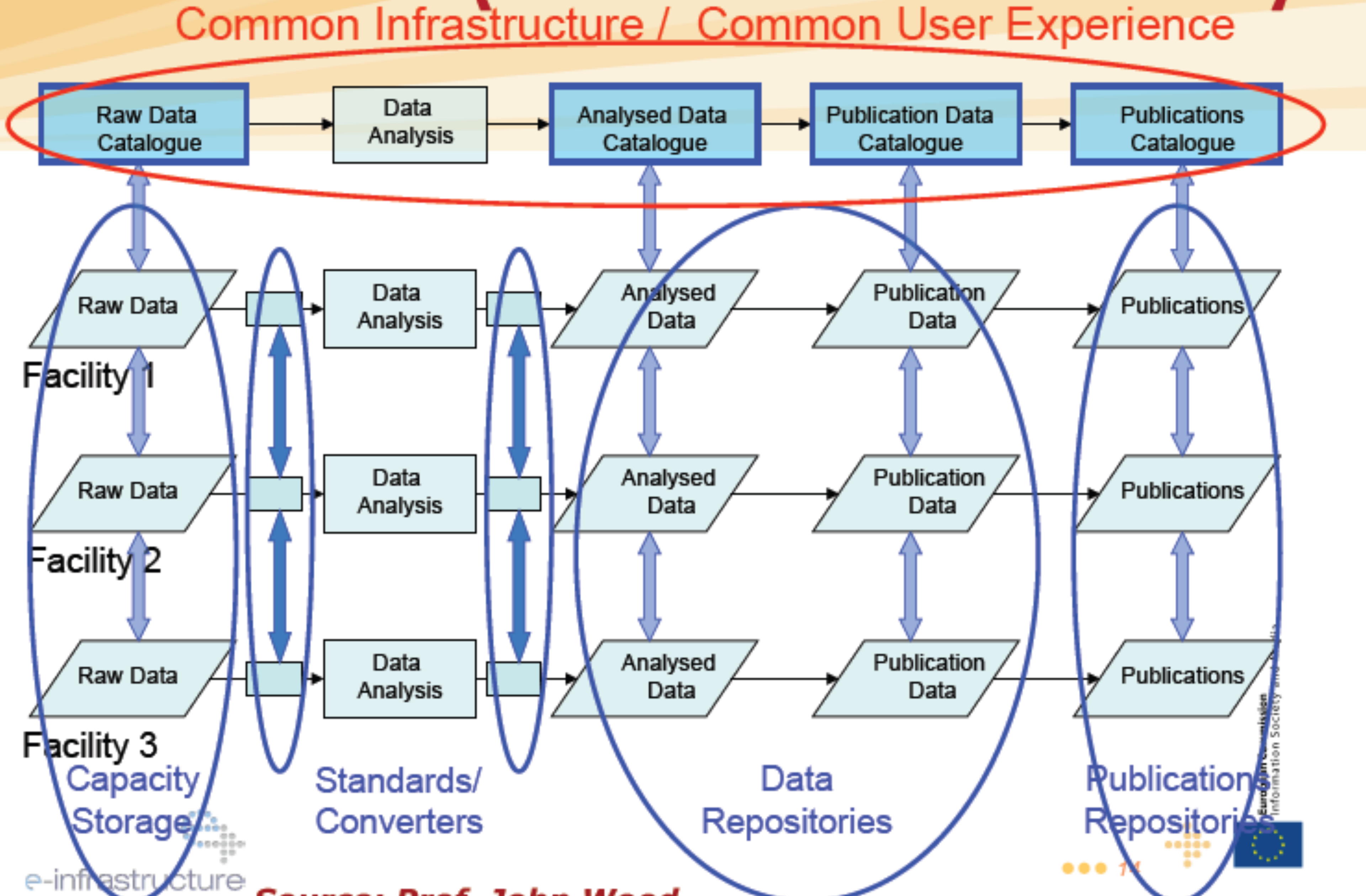
<u>Authenticity</u>

Certification & Integrity
GUIDs

<u>Generic interoperability</u>

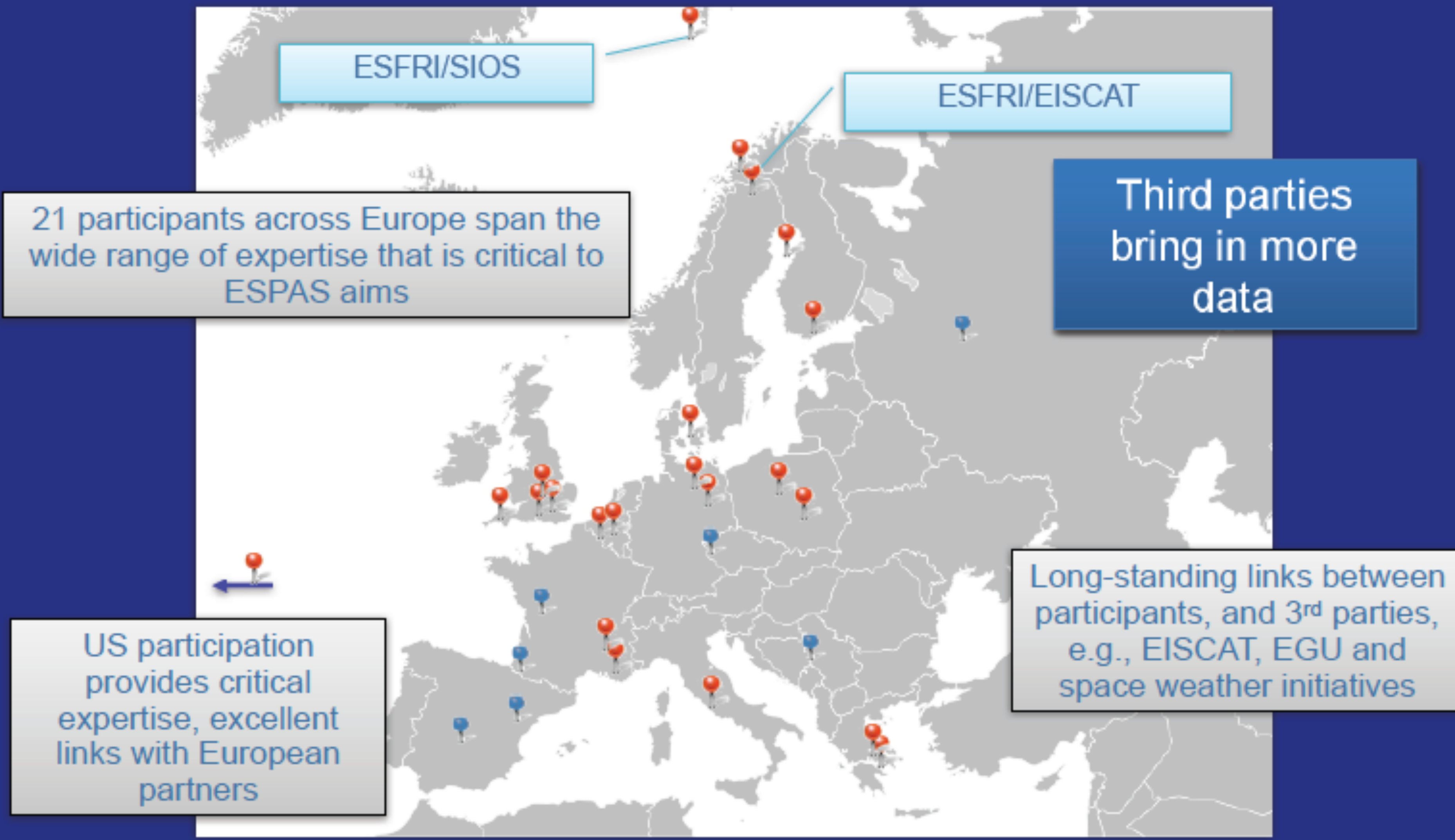
Technical
Legal
Semantic

Future view (e-Infrastructure enabled)



Source: Prof. John Wood

ESPAS: Near-Earth Space Data Infrastructure for e-Science *(proposal to EU FP7)*



ESPAS : technology (interoperability, efficiency) AND policies (quality, access)

US/EU Bilateral Initiatives

AGENDA

International Cooperation in Research Infrastructures

Symposium on transatlantic EU-U.S. cooperation in the field of large scale research infrastructures

1st October 2010

Rome, Italy
CNR - Piazzale Aldo Moro, 7, 00185 Rome

Implications for the E-Infra Community

How much data do we need?

E-Infrastructure Collaborations