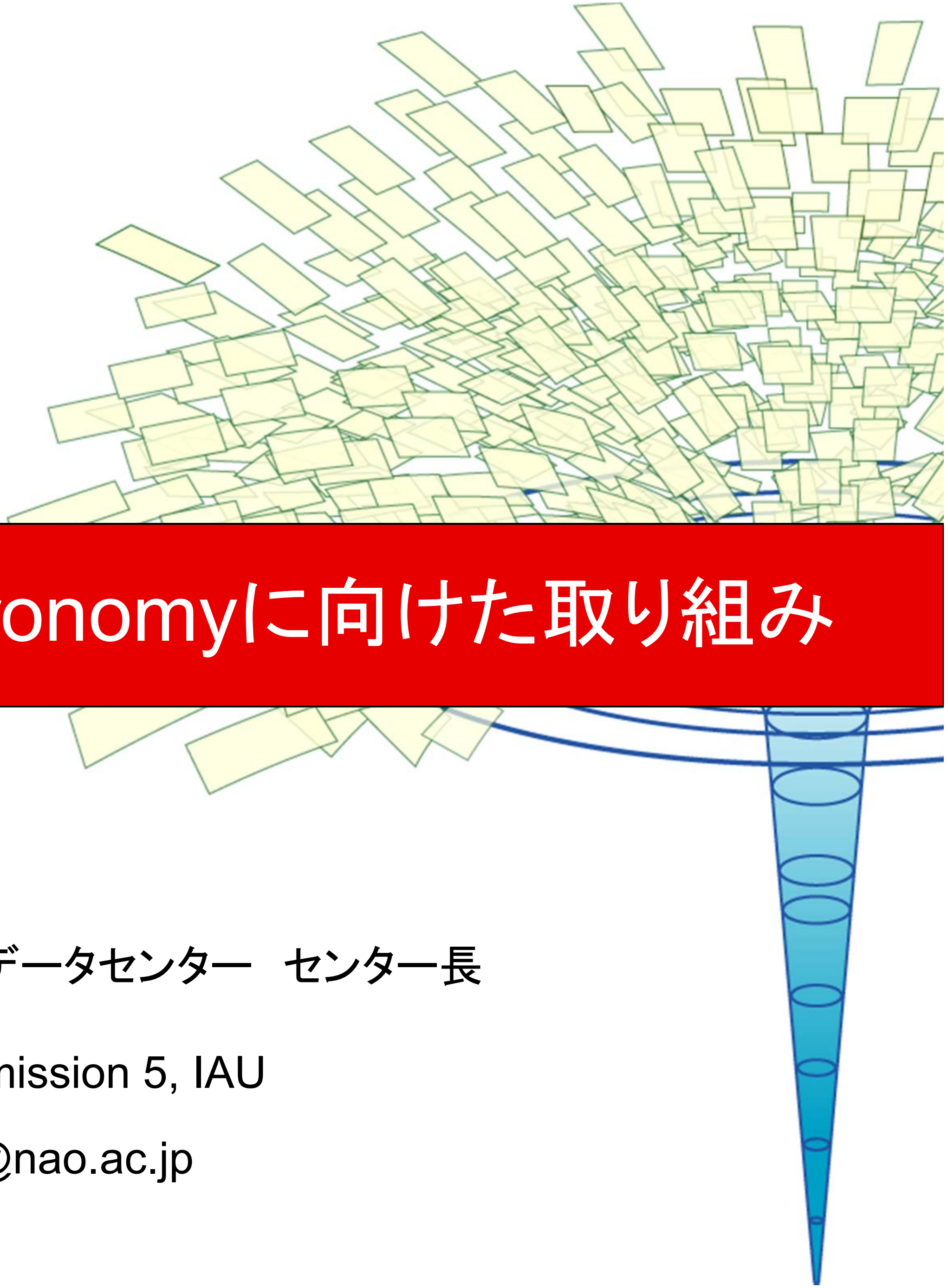


Data Intensive Astronomyに向けた取り組み

大石雅寿

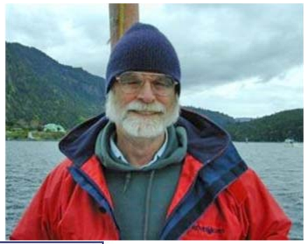
国立天文台　天文データセンター　センター長
&
President of Commission 5, IAU

masatoshi.ohishi@nao.ac.jp

Structure of my Talk

- Era of Data Intensive Sciences
  - toward “4th paradigm”
- Data Discovery in Astronomy
  - How to find necessary data for our research
- Towards Standardization
  - Differences can be overcome
- How do we manage data ?
  - ALMA, SKA
- Summary

Science Paradigms

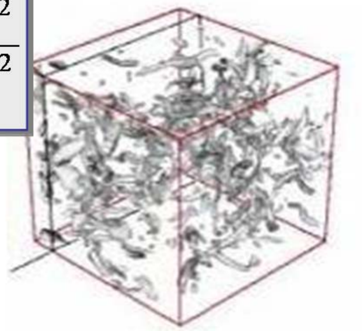


- Thousand years ago:
  science was **empirical**
  -- observations / experiments

- Last few hundred years:
  theoretical studies

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

- Last few decades:
  simulations



- Today:
  data exploration (e-Science)
  unify theory, experiment, and simulation
  – High-speed network
  – Computers, storages, databases

Era of Data Intensive Sciences

Accelerating Discoveries

- Issues, Planning
- Observation
- Data Reduction
  - Calib., Select, Combine

    , , ,
- Data Analysis
  - Physical Parameters
  - Thinking
  - Solution
- Publish

Data

↓

Information

↓

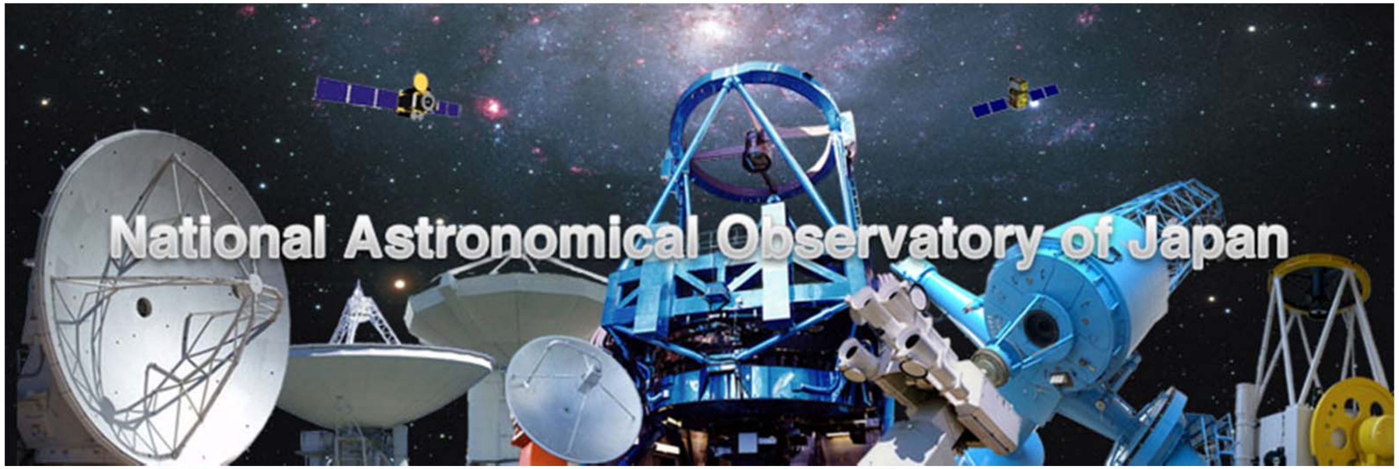
Knowledge

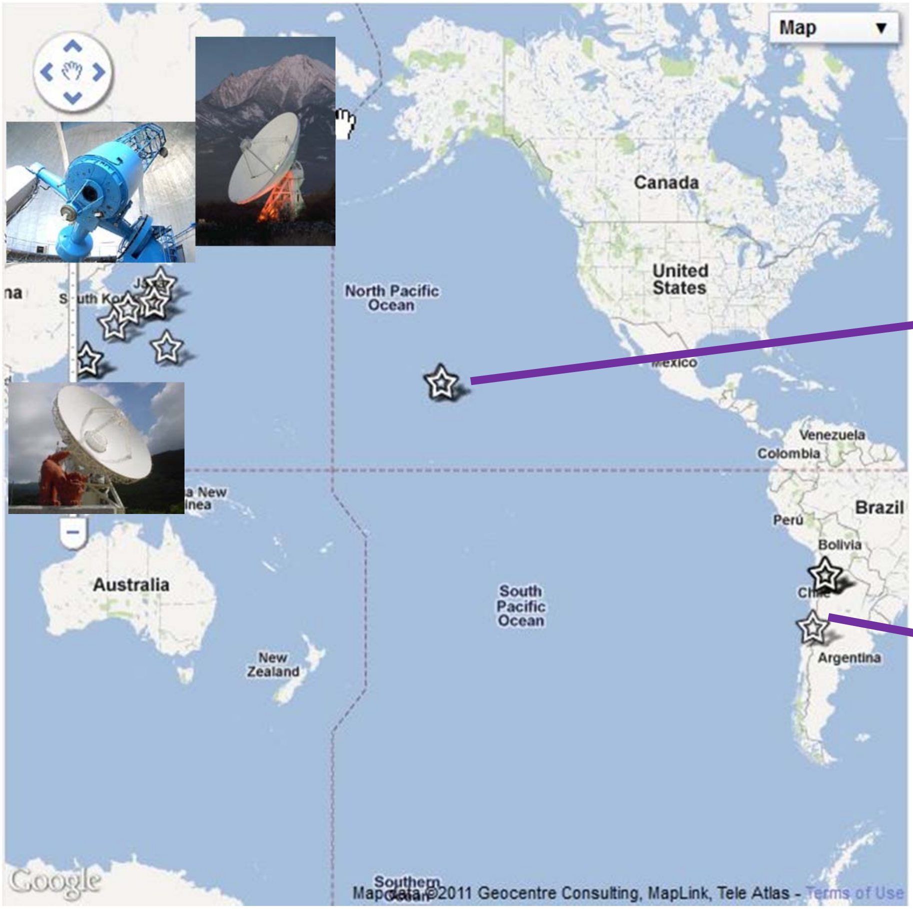
↓

Understanding

↓

Wisdom

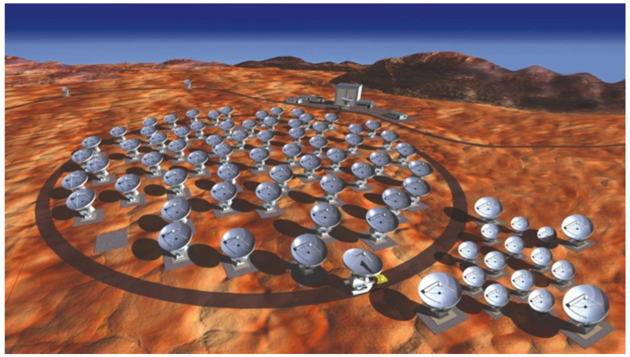




Map data ©2011 Geocentre Consulting, MapLink, Tele Atlas - Terms of Use

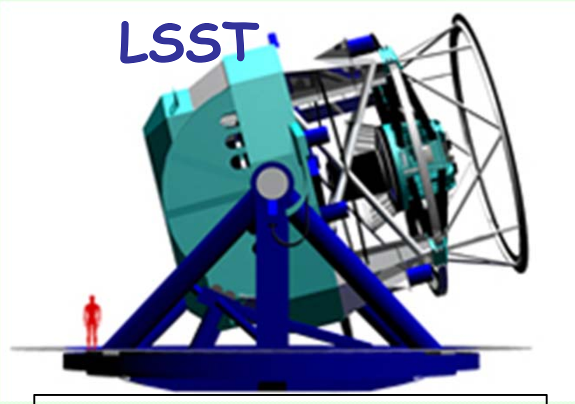
Subaru

ALMA

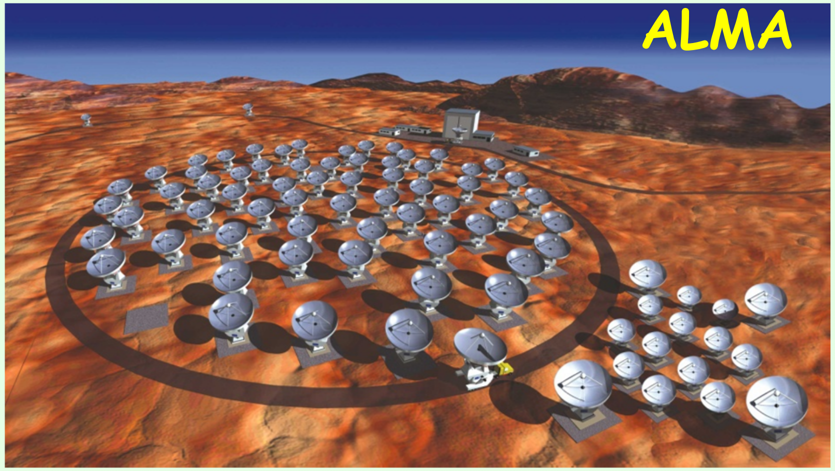


Planned Future Astronomy Projects

- ALMA
- JWST
- LSST
- LOFAR
- SKA
- TMT
- Pan-STARRs
- E-ELT

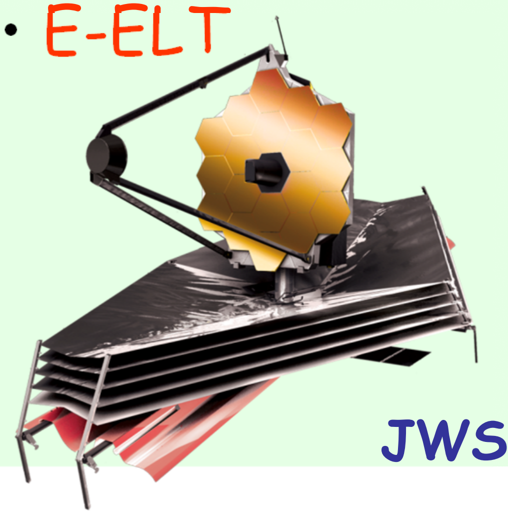


30 PB/yr x 6 yr ~ 200 PB

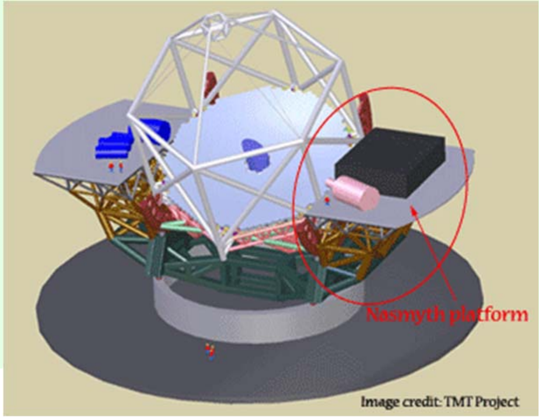


~ a few PB/yr

~ a few TB/night , only object params stored



JWST



TMT

Two Major Categories

Pointing Obs.

- ALMA
- JWST
- TMT
- E-ELT
- GMT

Large collecting area

High resolution

Surveys

- LSST
- Pan-STARRs
- SDSS2
- SKA ?

Whole sky

Time-domain astronomy

Two Major Categories

Pointing Obs.

- ALMA
- JWST

Surveys

- LSST
- Pan-STARRs

cosmology, the large-scale structure of the Universe, formation of galaxies, star formation, variable stars, transient phenomena such as the Gamma-ray bursts, small bodies in the solar system, extrasolar planets, life in the Universe, dark matter and dark energy, and others

Large collecting area

High resolution

Whole sky

Time-domain astronomy

Requirements in the Data Intensive Science Era

Data producer side

- Definition of data quality index, and establishment of quality assessment methodologies
- Quality assurance of data (from obs. to data analyses)

Data center side

- Establishment of data handling environment
  - Distributed CPUs
  - Distributed storage
  - Distributed data analysis software (pipeline) incl. data mining, knowledge discovery, statistics, event discovery
  - High-speed network

Requirements in the Data Intensive Science Era

Data producer side

- Definition of data quality index, and establishment

Quality assurance of data (from obs. to data analyses)

Data center side

- Establishment of data handling environment

incl. data mining, knowledge discovery, statistics, event discovery

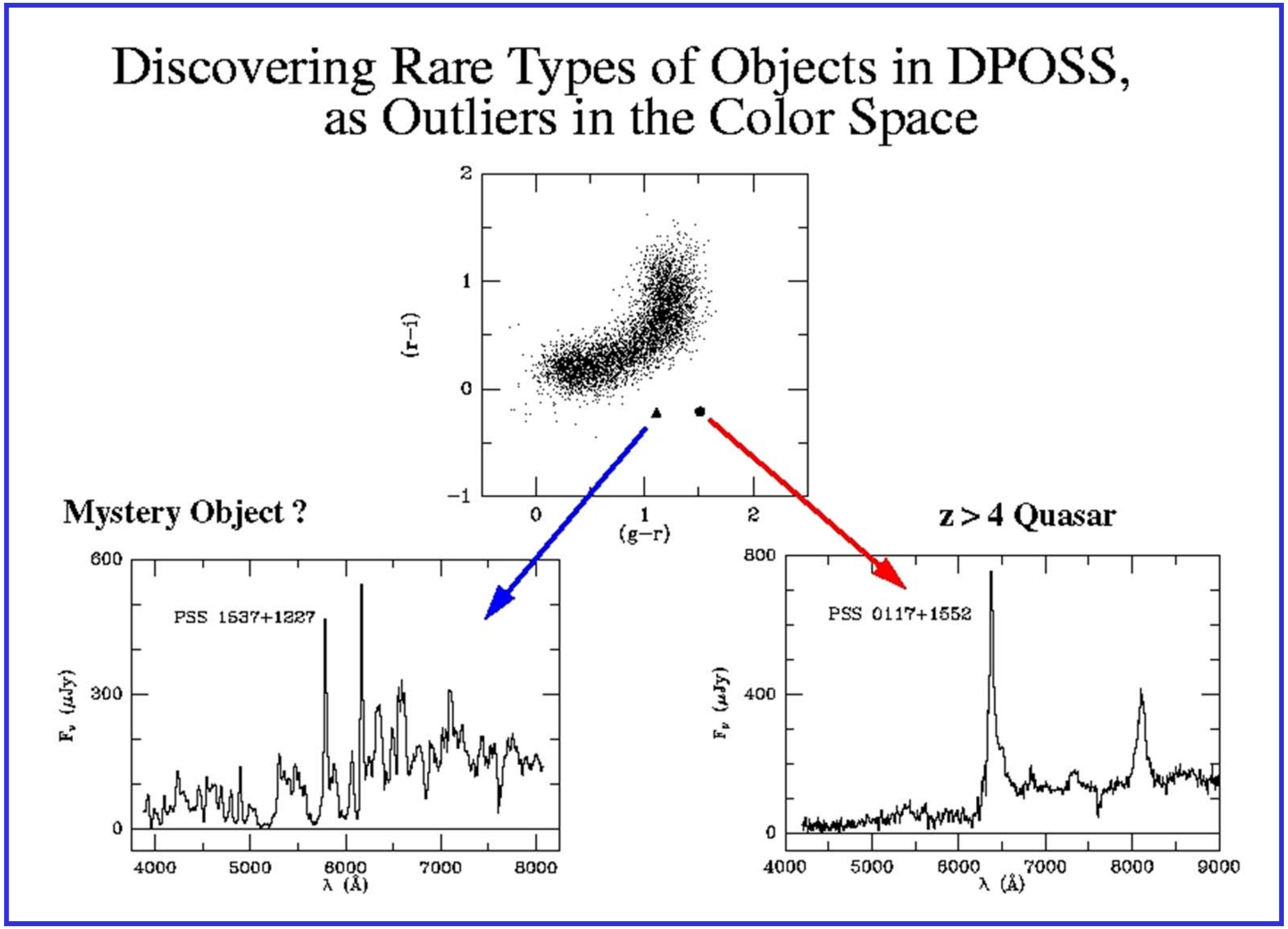
  - High-speed network

Data management / analysis cost will become a major issue

Getting Knowledge

- Approaches on Data analyses: mathematical statistics and/or taxonomy
- With scientific working hypothesis – what do we want to know from the deluge of data ?
  - We need to have a sensitive antenna
  - Serendipitous discoveries might be possible, but…
- Data publication as early as possible
- Data Scientists in exploring the deluge of data

mystery outliers



Discovering Rare Types of Objects in DPOSS, as Outliers in the Color Space

Mystery Object ?

z > 4 Quasar

graphics from US NVO project

Data Discovery in Astronomy

VO– New Research Infrastructure in the 21st Century

A collection of integrated astronomical data archives and software tools that utilize computer networks to create an environment in which research can be conducted.

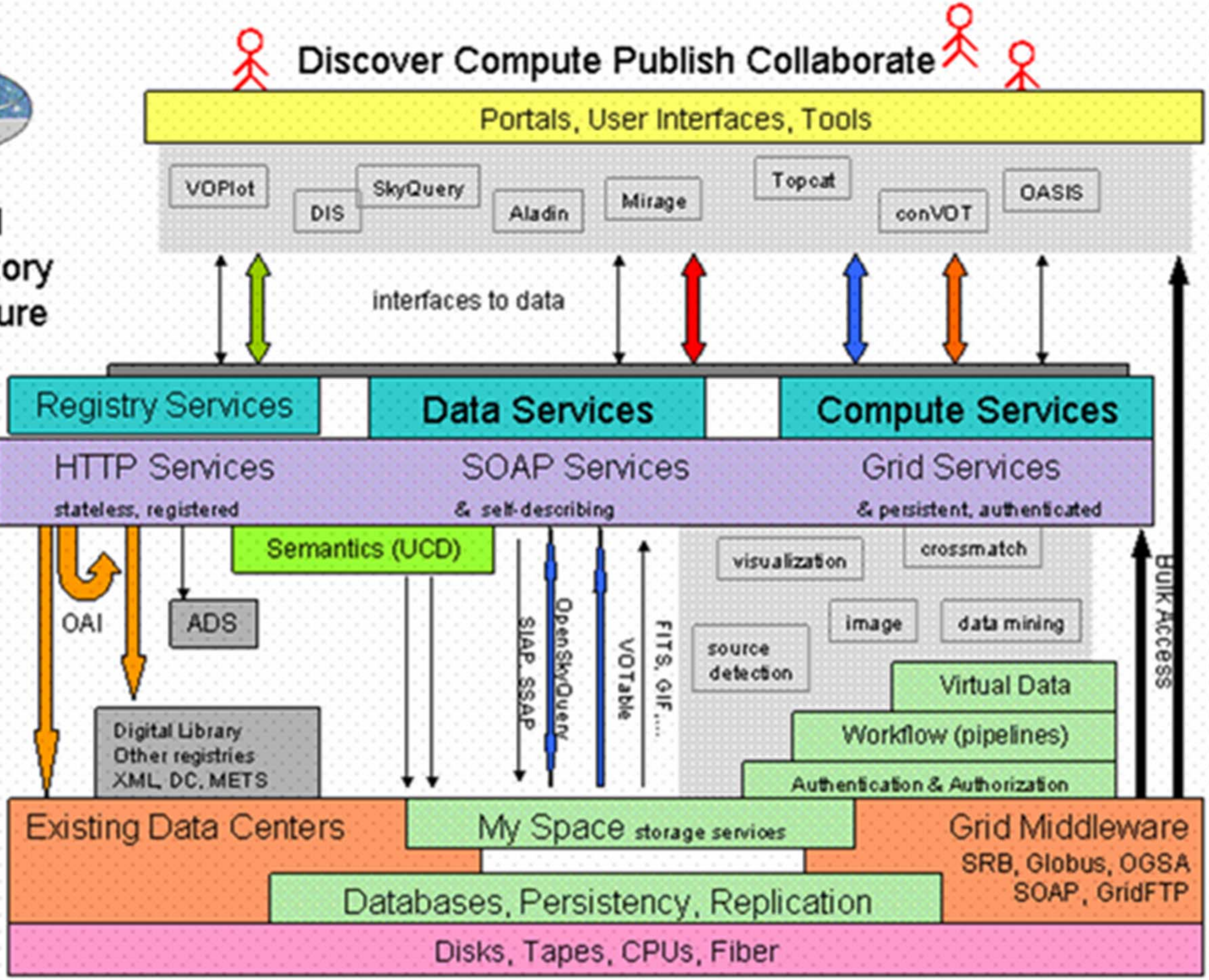
http://www.encyclopedia.com/html/v1/virtobserv.asp

VO Projects in the world

- 18 members worldwide
- International Virtual Observatory Alliance (IVOA – http://www.ivoa.net/)
  → Standards to interoperate VOs

- No center (good-will), No shared project funding

Virtual Observatory Architecture

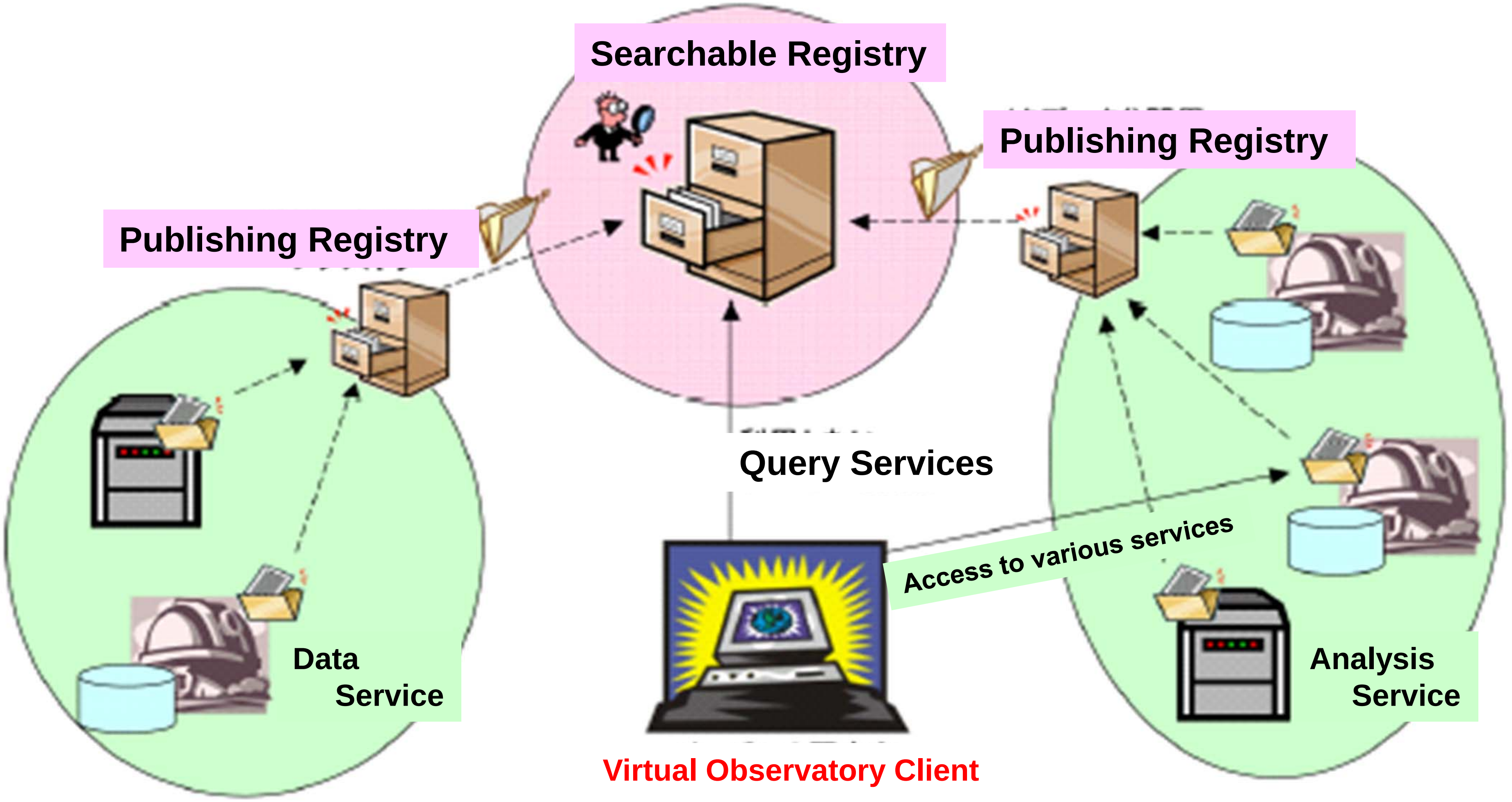


Standardization in IVOA

- Meta-data
  - Contents & access protocol
- Access Images, Spectra, Catalogues
  - TAP, SIAP, SSAP, STC, etc.
- Query Language to Federated DBs (ADQL)
- Unified Attribute Names
  - UCD (Unified Contents Descriptions)
- Output format : VOTable (in XML)
  - FITS

Exchange of Meta Data: OAI-PMH

2011 Oct 19



Data Access Protocols

- Parameter query in terms of the HTTP

```
http://jvo.nao.ac.jp/imageData?Pos=24,5&Size=0.2&format=VOTable
```

□Simple Image Access Protocol (SIAP)

□Simple Spectrum Access Protocol (SSAP)

□Table Access Protocol (TAP)

etc.

- Unified query language (JVOQL) for both the catalog and observation data such as image data, spectrum, 3D-cube, photon list …

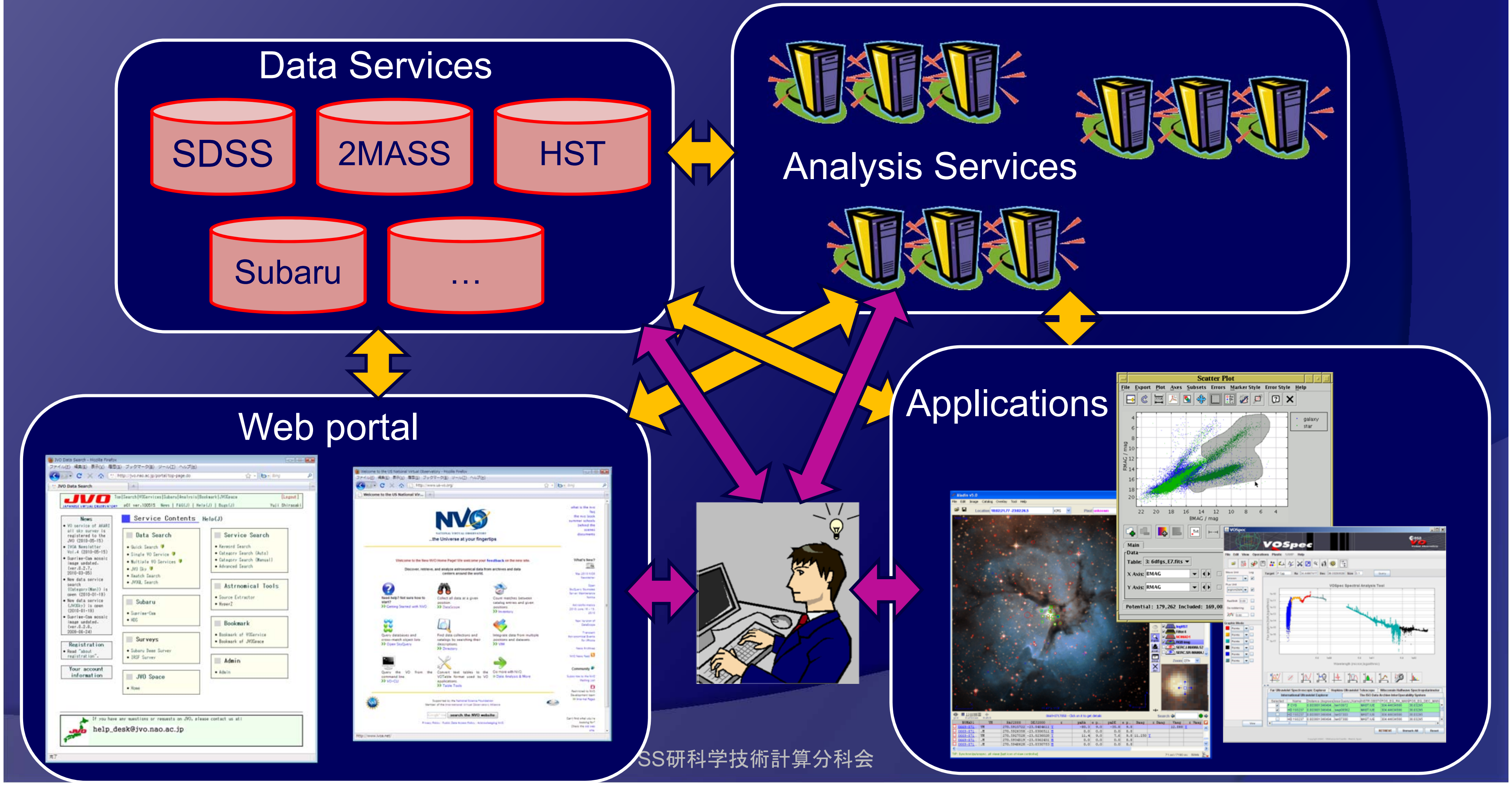
```
Select      imageURL, …
From        naoj:imageData
Where       pos=Point(24,5) and size=0.2 and format='VOTable'
```

File Formats

- Flexible Image Transfer System (FITS)
  - standardized in early 80's to exchange observed data
  - 1 record = 2880 bytes
  - (Header, Data)(Header, Data)･･･
  - IAU has the FITS WG to maintain its specification
- VOTable
  - used in Virtual Observatories as an output format
  - described in XML, and standardized in IVOA
  - can inline FITS files / contain a link to FITS files

Virtual Observatory

- ✓ Infrastructure for efficient research environment
- ✓ International standards for data publication & access
- ✓ Sharing data worldwide, Maximize scientific return



Data Services

SDSS

2MASS

HST

Subaru

…

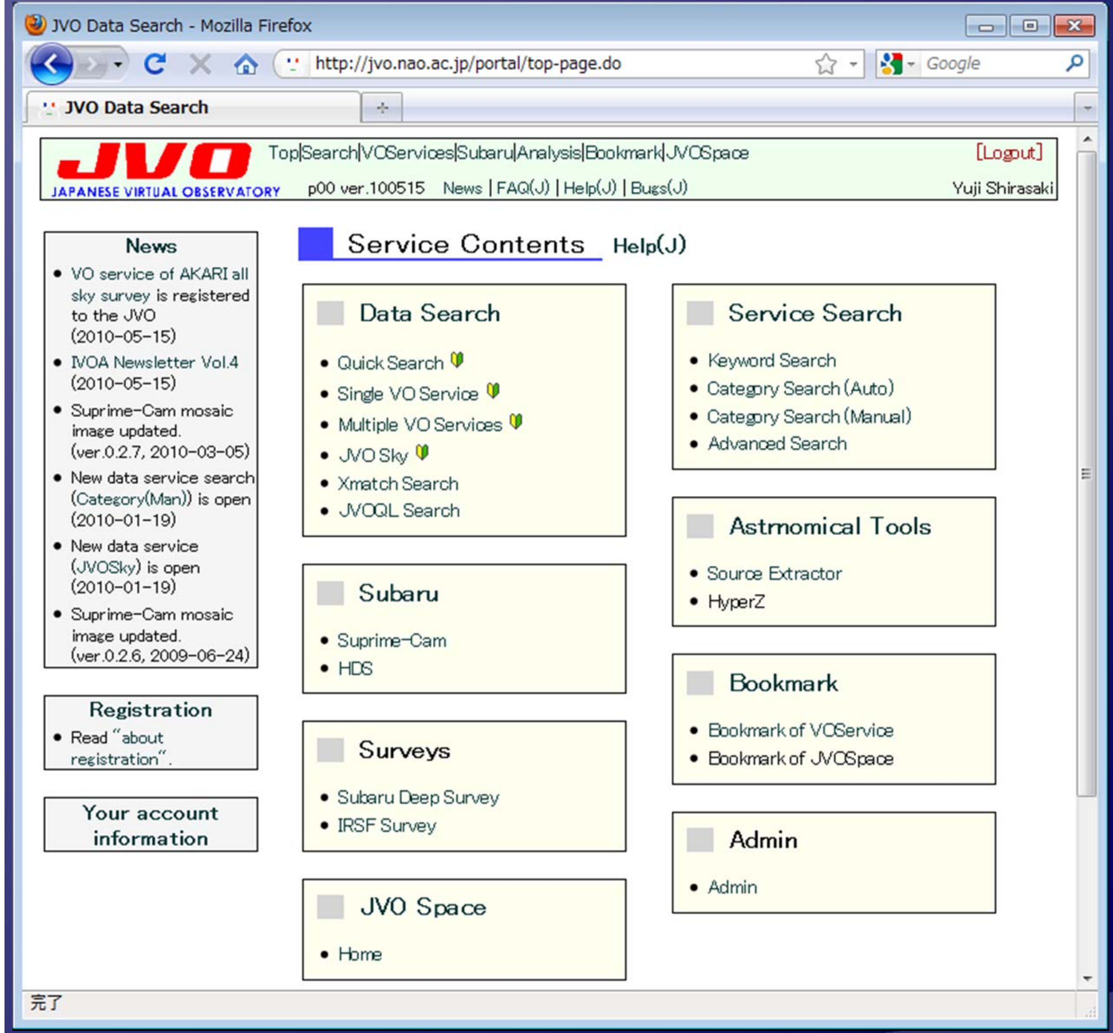
Analysis Services

Web portal

Applications

JVO portal

http://jvo.nao.ac.jp/portal



JVO Data Search - Mozilla Firefox

http://jvo.nao.ac.jp/portal/top-page.do

JVO Data Search

JVO JAPANESE VIRTUAL OBSERVATORY

Top|Search|VOServices|Subaru|Analysis|Bookmark|JVOSpace [Logout]

p00 ver.100515 News | FAQ(J) | Help(J) | Bugs(J) Yuji Shirasaki

News

- VO service of AKARI all sky survey is registered to the JVO (2010-05-15)
- IVOA Newsletter Vol.4 (2010-05-15)
- Suprime-Cam mosaic image updated. (ver.0.2.7, 2010-03-05)
- New data service search (Category(Man)) is open (2010-01-19)
- New data service (JVOSky) is open (2010-01-19)
- Suprime-Cam mosaic image updated. (ver.0.2.6, 2009-06-24)

Registration

- Read "about registration".

Your account information

Service Contents Help(J)

Data Search

- Quick Search
- Single VO Service
- Multiple VO Services
- JVO Sky
- Xmatch Search
- JVOQL Search

Subaru

- Suprime-Cam
- HDS

Surveys

- Subaru Deep Survey
- IRSF Survey

JVO Space

- Home

Service Search

- Keyword Search
- Category Search (Auto)
- Category Search (Manual)
- Advanced Search

Astrnomical Tools

- Source Extractor
- HyperZ

Bookmark

- Bookmark of VOService
- Bookmark of JVOSpace

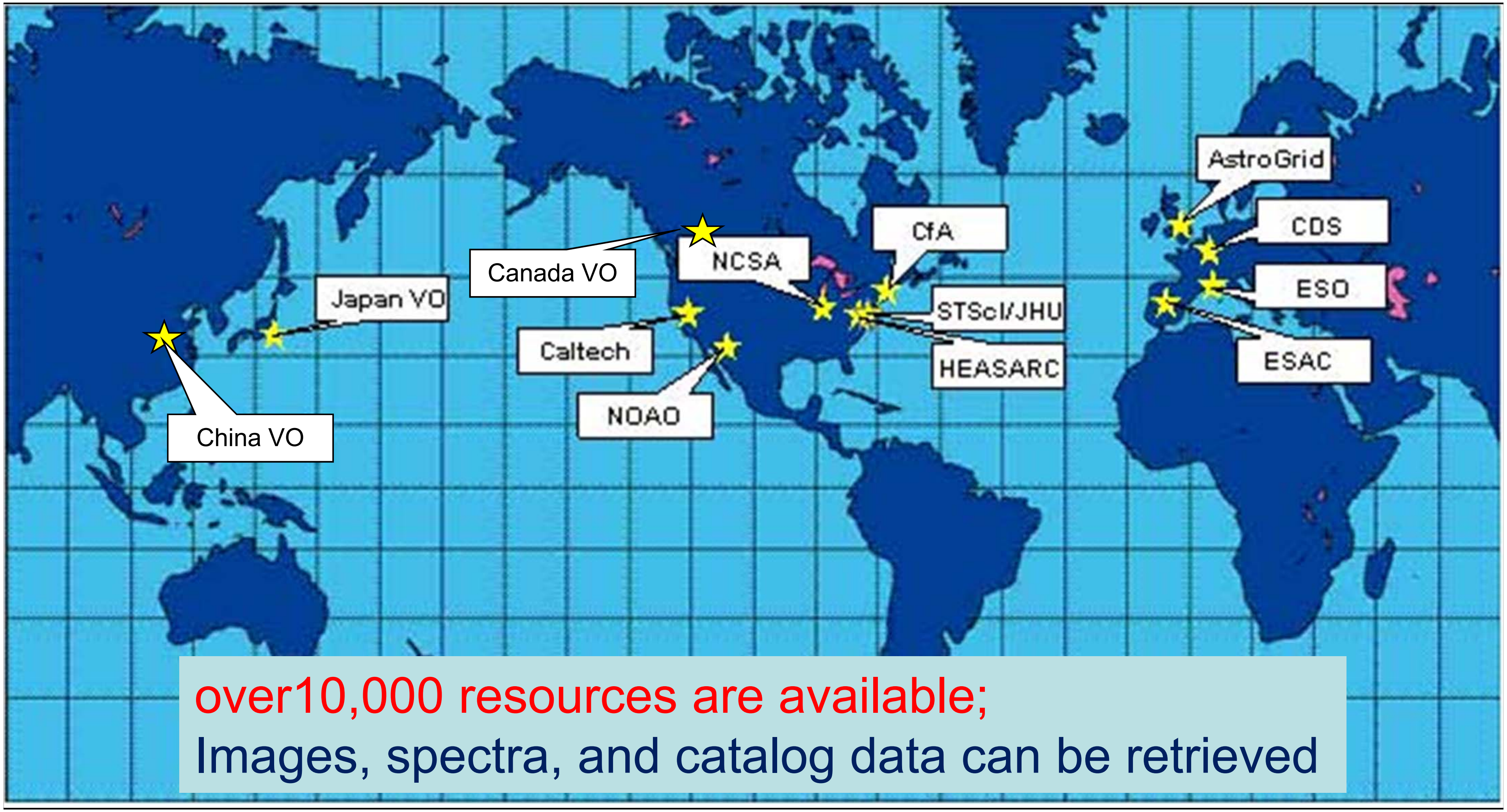
Admin

- Admin

完了

- ✓ 10,551 Data Resources
  - 7,397 Catalogs
  - 208 Image Services
  - 84 Spectrum Services
  - …
- ✓ Reduced Subaru Data
  - Suprime-Cam
  - HDS

Astronomical Virtual Observatories
~ Data Grid ~



over10,000 resources are available;
Images, spectra, and catalog data can be retrieved

Towards Standardization

Establishing Standards

- Standards are quite effective
  - Access protocols, data format, etc.
  - Interoperability → wider dissemination and application
  - Endorsement by the IAU (VO WG)

- Painful process
  - Philosophy, intention, life time of project,,,
  - Compromise, patience
  - Establishment of relationship : respect to each other
  - Coffee/tea breaks and lunch/dinner talks are crucial

IVOA Interoperability meetings

- Twice a year, since 2003
- Discussions toward standardization
- Human network as a basis for cyber network (Layer 0)

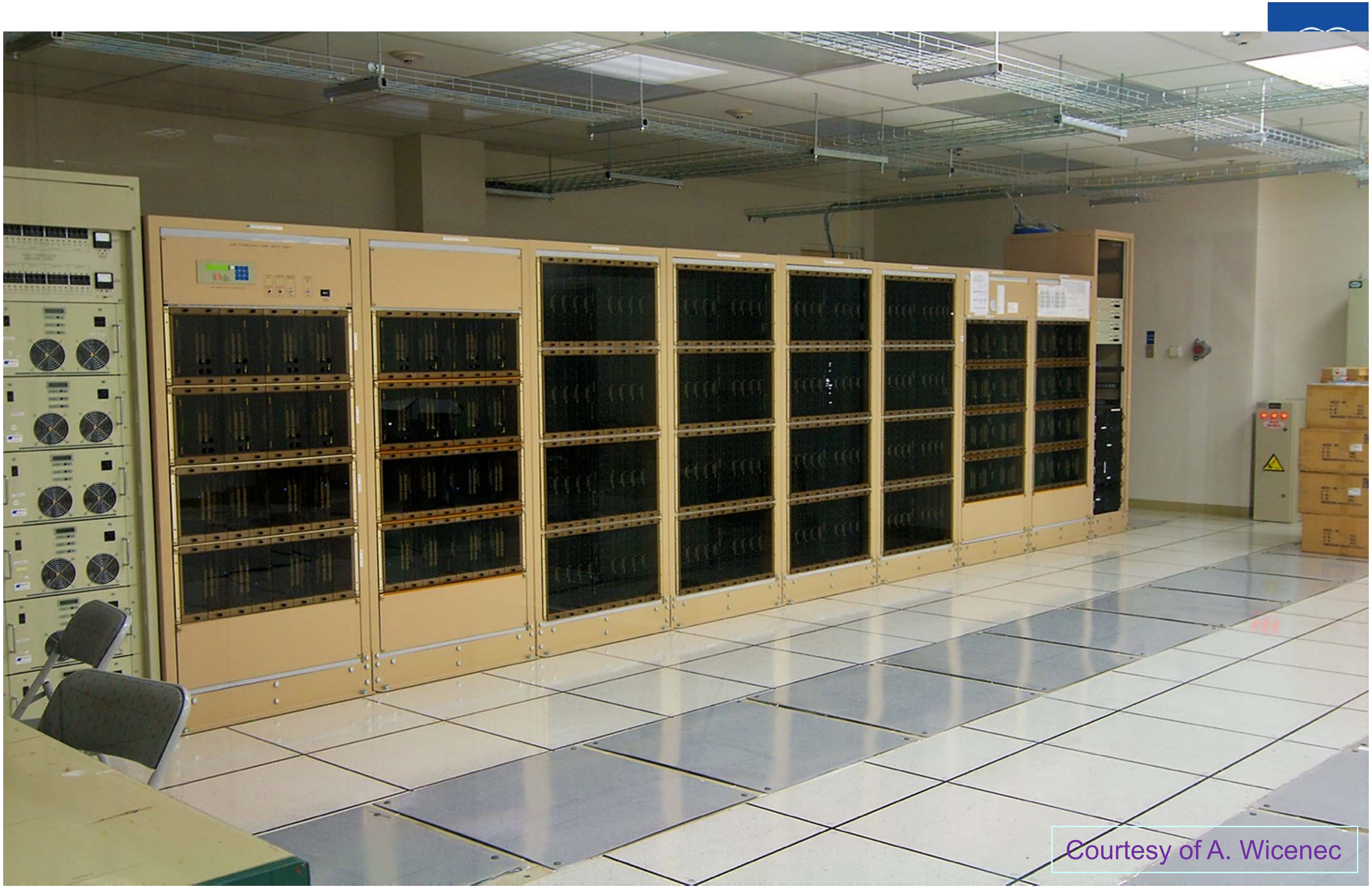
Nara, 2010 December

How do we manage data ?

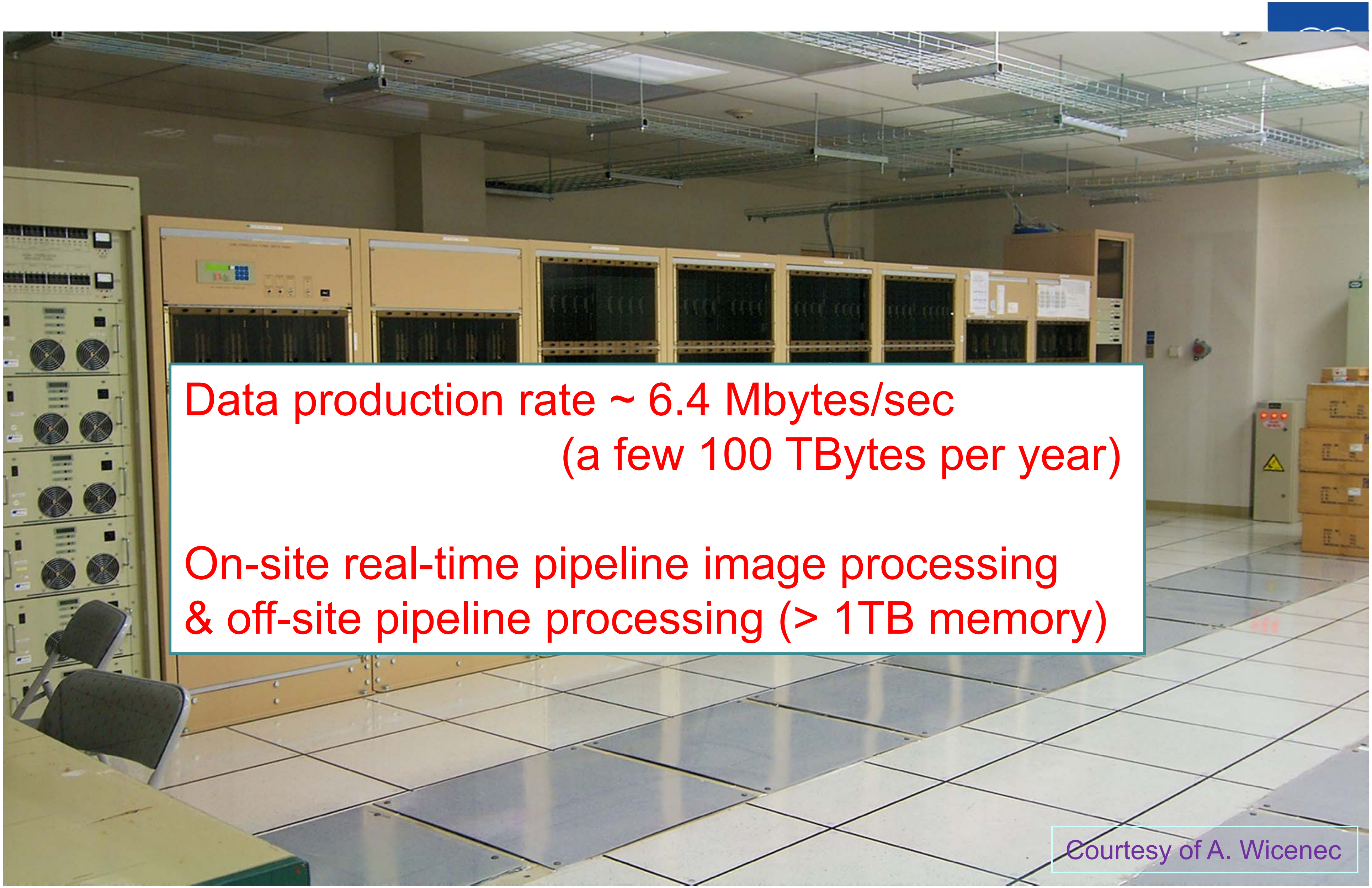
ALMA telescope in Chile

Scientific Systems

@ ESO/NAOJ/NRAO


Courtesy of A. Wicenec

FPGA based correlator: Highly customized HPC system directly attached to antenna output. Image shows one quadrant of ALMA correlator installed on the Chajnantor plateau (Chile) in 5000m elevation. Image: A. Wicenec



Data production rate ~ 6.4 Mbytes/sec
(a few 100 TBytes per year)

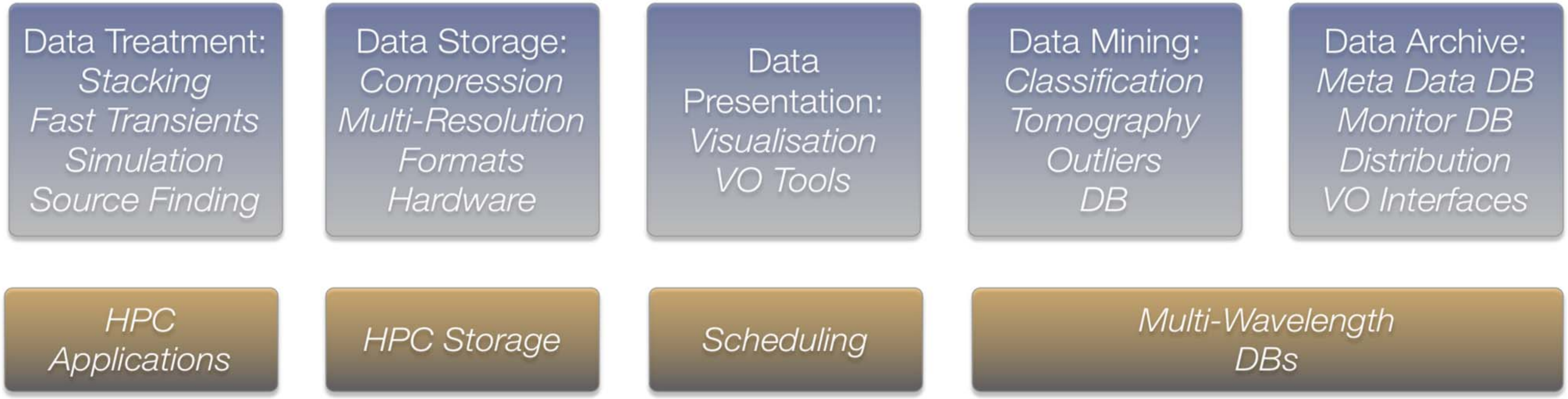
On-site real-time pipeline image processing
& off-site pipeline processing (> 1TB memory)

Courtesy of A. Wicenec

FPGA based correlator: Highly customized HPC system directly attached to antenna output. Image shows one quadrant of ALMA correlator installed on the Chajnantor plateau (Chile) in 5000m elevation. Image: A. Wicenec

Accessing Data

- **Data will not be used and can thus be deleted if it is not presented in a useful way.**
- **If there is too much data to move around,**
  take the analysis to the data! (by Jim Gray)
- **If all data is manipulated in databases, automatic parallelism is guaranteed; easy data management**
- **Scalable to Peta-byte scale**



Data Treatment: *Stacking* *Fast Transients* *Simulation* *Source Finding*

Data Storage: *Compression* *Multi-Resolution* *Formats* *Hardware*

Data Presentation: *Visualisation* *VO Tools*

Data Mining: *Classification* *Tomography* *Outliers* *DB*

Data Archive: *Meta Data DB* *Monitor DB* *Distribution* *VO Interfaces*

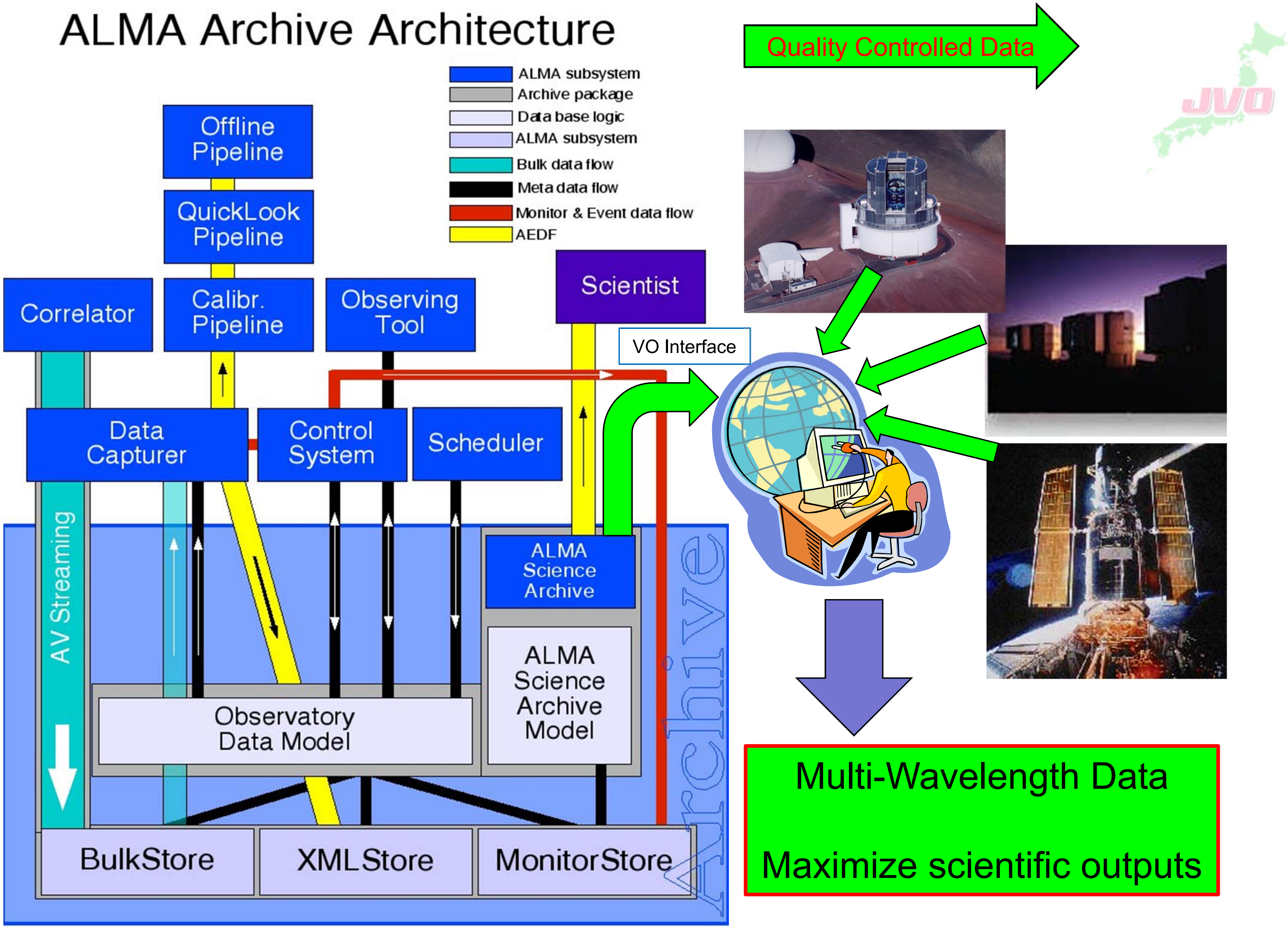
HPC Applications

HPC Storage

Scheduling

Multi-Wavelength DBs

ALMA Archive Architecture

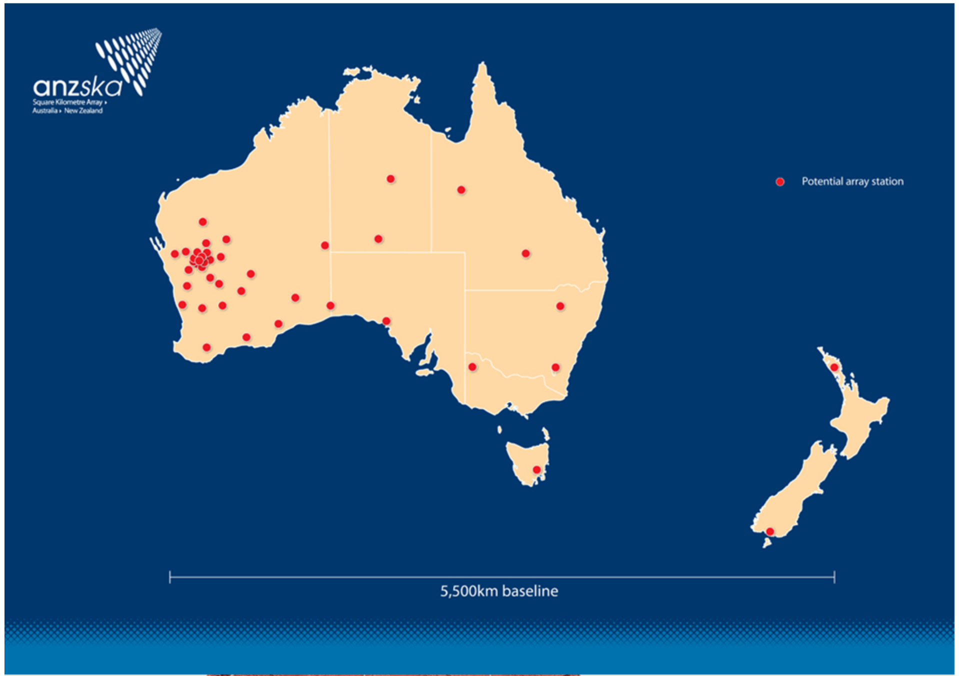


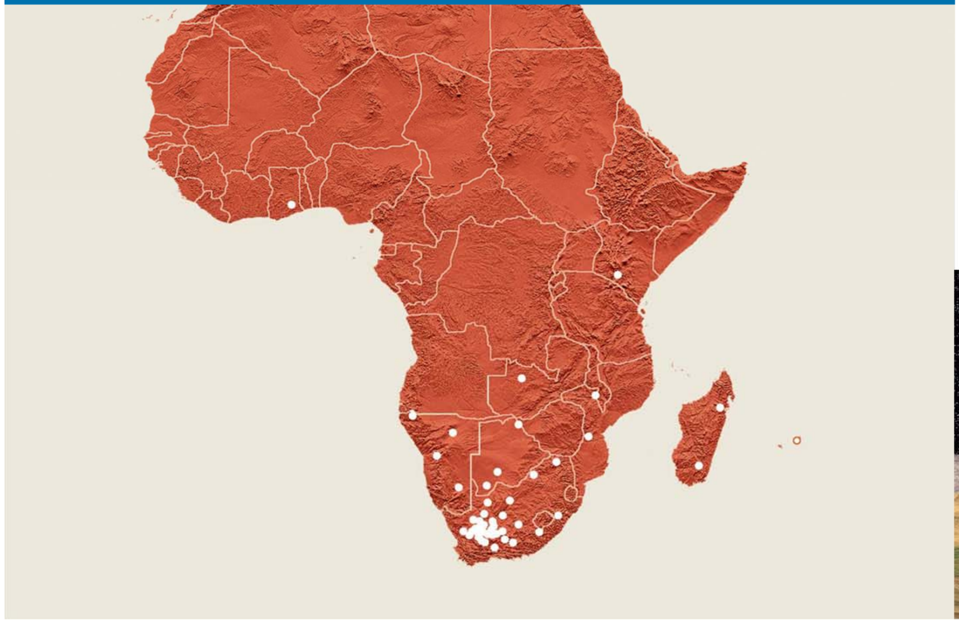
Multi-Wavelength Data

Maximize scientific outputs

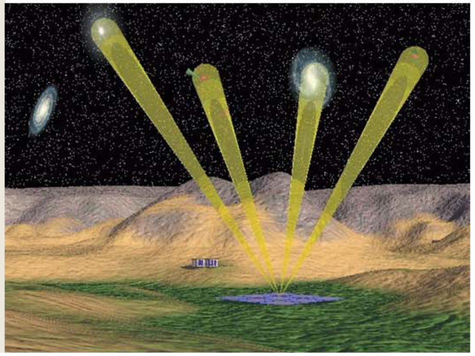
A. Wicenec 2004-05-31

Square Kilometre Array (SKA)

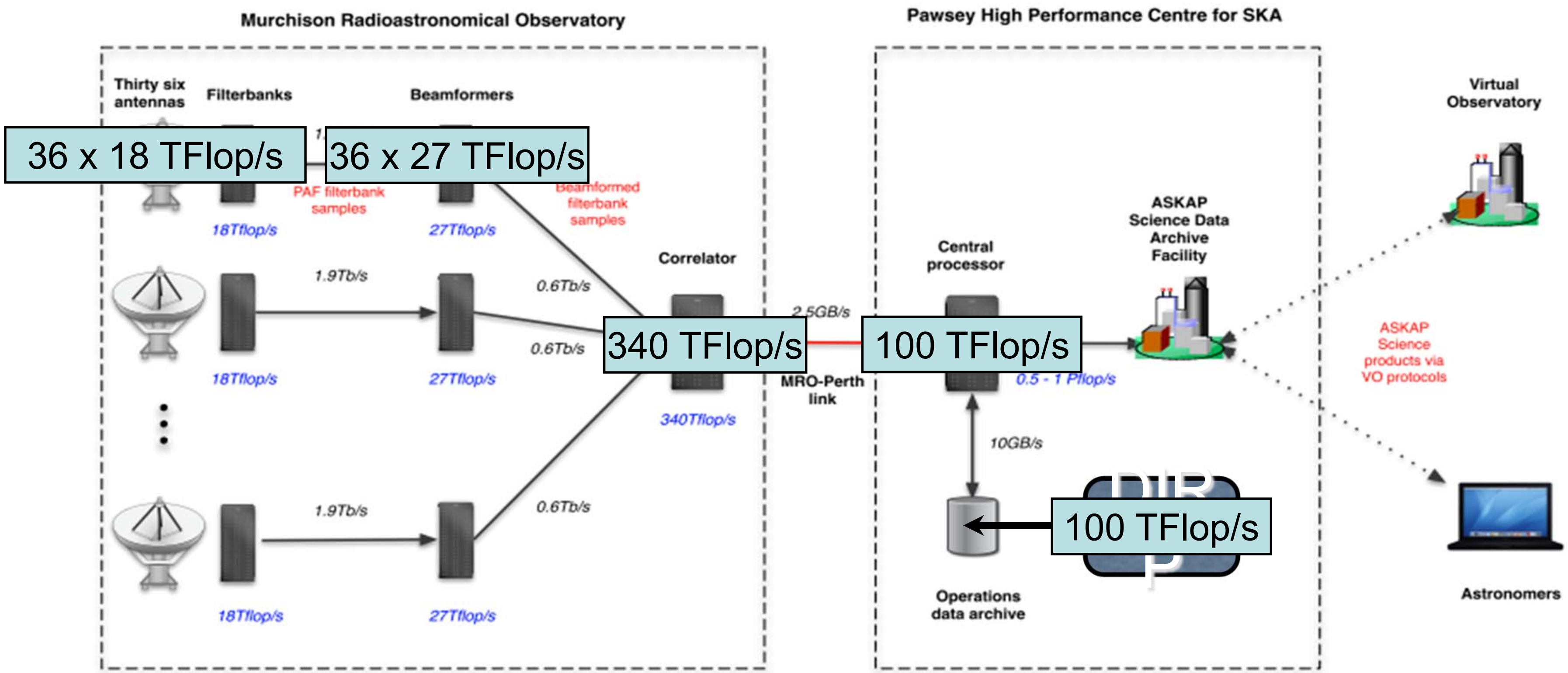




- 1km^2 collecting area
- Aperture synthesis radio telescope : 2D inverse FFT
- ~2020 ??
- Aus vs S. Africa
  - – ASKAP vs MeerKAT
  - – “1% SKA” prototypes



Schematic ASKAP Data Flow



36 x 18 TFlop/s

36 x 27 TFlop/s

340 TFlop/s

100 TFlop/s

100 TFlop/s

T. Cornwell, July 9 2010
with additions by A.Wicenec

Total: 2160 TFlop/s

Courtesy of A. Wicenec

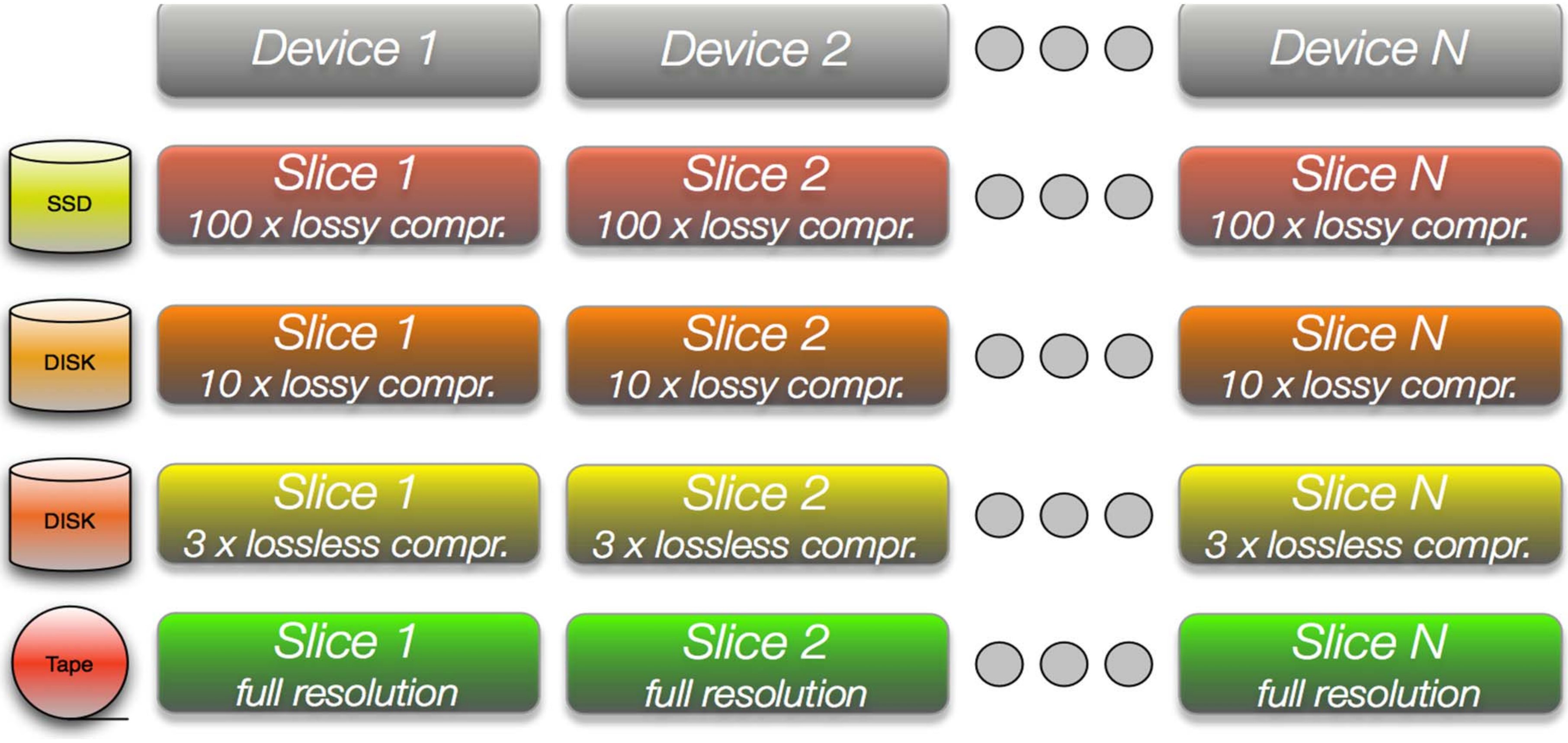
Data Storage

- TB size datasets require ‘smart’ storage, else risk of data graves.
- Problem arises from unproportional rise in capacity vs. transfer rate and random access speed:
  T2 = 1.5/4/10 years.
- Magnetic disks are degenerating to serial devices.
- Expensive solution: SSDs, but still have write degradation problem.
- Tapes don’t allow easy access to parts of data sets; problem enhanced by current access software.
- Data transfer stack requires careful planning to avoid bottle necks.

Data Storage:
Smart Storage, Smart Archive

- Evaluation and implemention of data life-cycle.
- Research on advanced, astronomy optimized *storage* formats (e.g. HDF5).
- Research on smart data distribution directly supported by storage format: Horizontal distribution.
- Research on smart data retrieval directly supported by storage format: Vertical distribution.
- Research on storage hardware supporting implementation of data aware storage and retrieval algorithms: Optimized, transparent access.
- Research on lossless and lossy compression and multi-resolution.

Data Storage:
Vertical and Horizontal Distribution



	Device 1	Device 2	○○○	Device N
SSD	Slice 1 100 x lossy compr.	Slice 2 100 x lossy compr.	○○○	Slice N 100 x lossy compr.
DISK	Slice 1 10 x lossy compr.	Slice 2 10 x lossy compr.	○○○	Slice N 10 x lossy compr.
DISK	Slice 1 3 x lossless compr.	Slice 2 3 x lossless compr.	○○○	Slice N 3 x lossless compr.
Tape	Slice 1 full resolution	Slice 2 full resolution	○○○	Slice N full resolution

Courtesy of A. Wicenec

Special Challenges

- HPC in real-time data reduction chain.
- High volume data streaming through top 100 supercomputer.
- Very big data sets. Data life cycle undefined.
  - ALMA data can (may) be manageable

- Towards SKA: Solutions should scale from ASKAP (1%) to SKA1 (10%) and SKA2 (100%)
- Algorithms are still mostly serial, or don’t scale to hundreds of thousands of cores.
- Budget is constrained, and power consumption has to come down by factor 10-100.

High level Data Analysis

- Looking for “new rules, insights” through huge dataset
  - Needles in haystacks – the Higgs particle
  - Haystacks: Dark matter, Dark energy

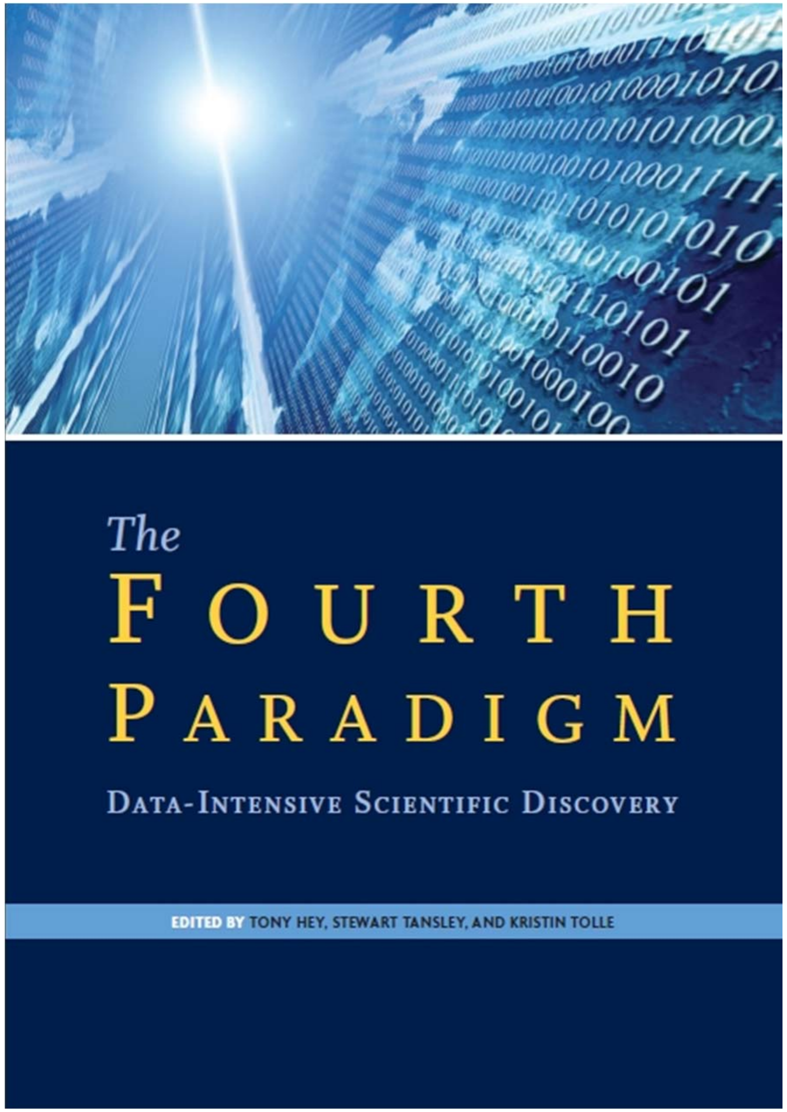
- Global statistics have poor scaling
  - Correlation functions are N^2, likelihood techniques N^3
  - We can only do $N \, logN$

- Must accept approximate answers
  New algorithms – Data Mining (KDD)

- Requires collaboration with
  - statisticians & computer scientists

Data Intensive Science

- Data deluge
  - Huge data size
  - Wide variety
  - Transient data
  - time-domain
- New paradigm in scientific research by introducing data management and advanced data analysis



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE