i-explosion

Data Intensive Scienceのススメ

National Astronomical Observatory of Japan
President of Comm 5, IAU

Masatoshi OHISHI

masatoshi.ohishi@nao.ac.jp

Accelerating Discoveries

- Issues, Planning
- Observation
- Data Reduction
  - Calib., Select, Combine, , ,
- Data Analysis
  - Physical Parameters
  - Thinking
  - Solution
- Publish

Data

↓

Information

↓

Knowledge

↓

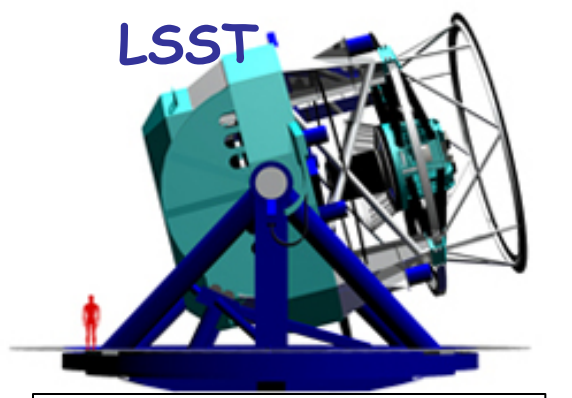
Understanding

↓

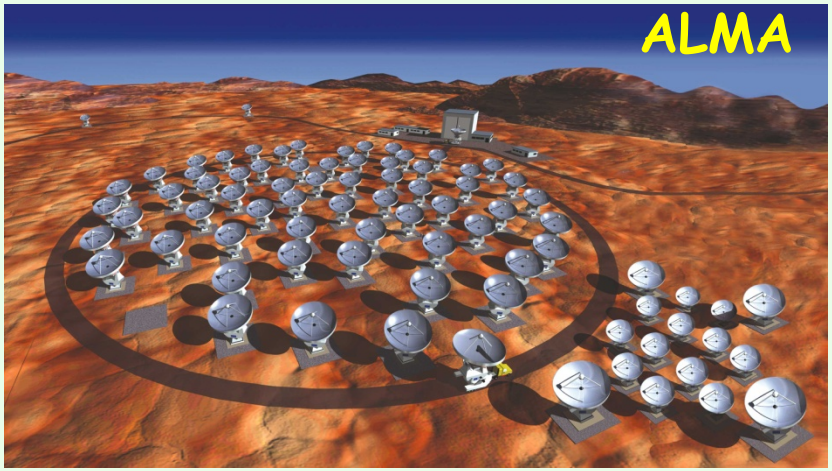
Wisdom

Planned Future Astronomy Projects

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

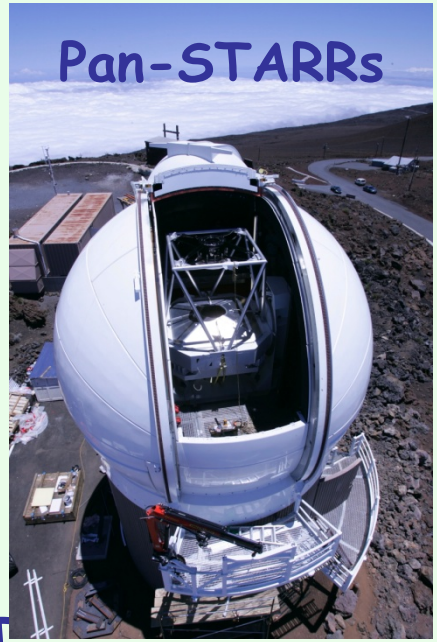


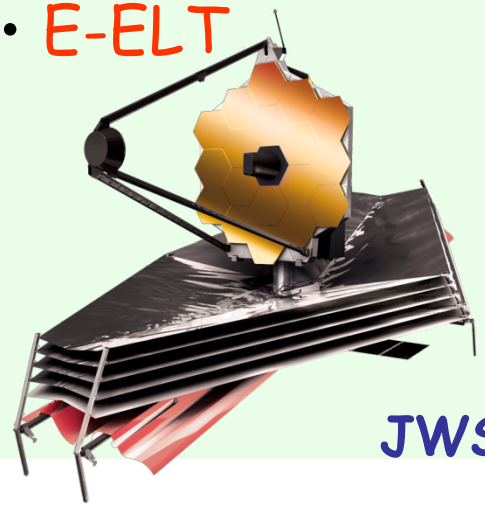
30 PB/yr x 6 yr ~ 200 PB



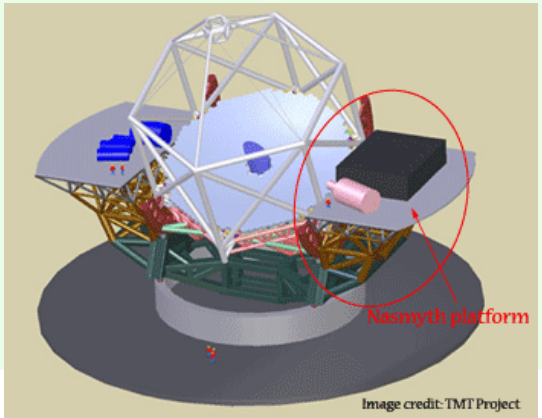
~ a few PB/yr

~ a few TB/night , only object params stored



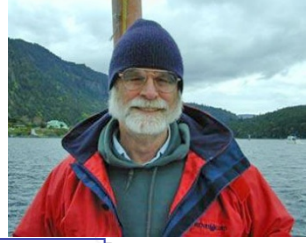


JWST



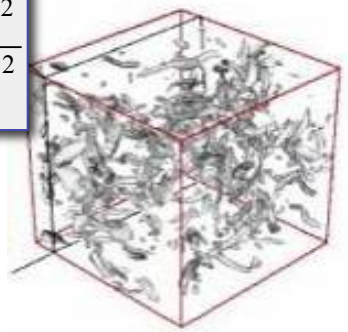
TMT

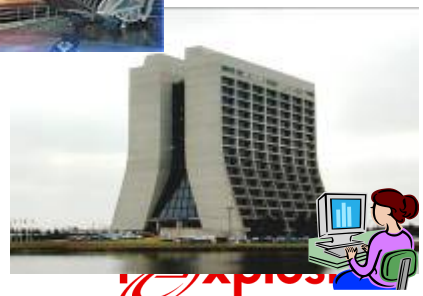
Science Paradigms



- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    Or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes database / files
    using data management and statistics

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

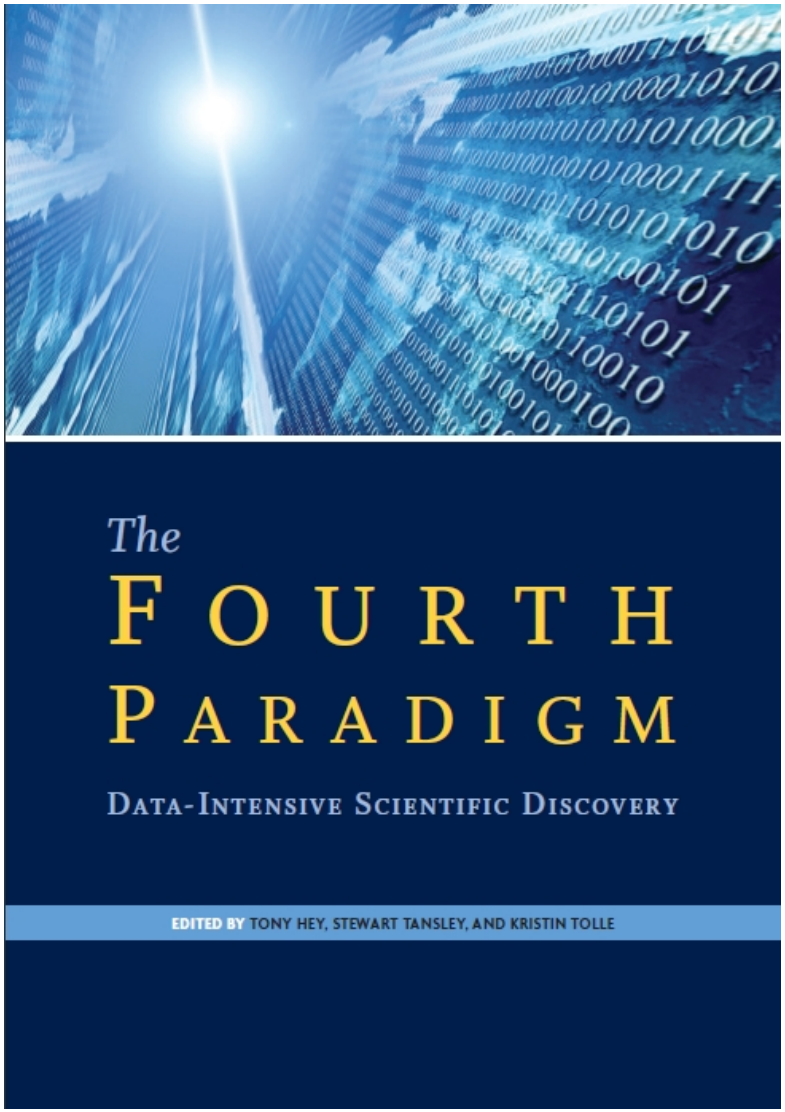




Data Intensive Science

- Data deluge
  - Huge data size
  - Wide variety
  - Transient data
  - time-domain

- New paradigm in scientific research by introducing data management and advanced data analysis



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)

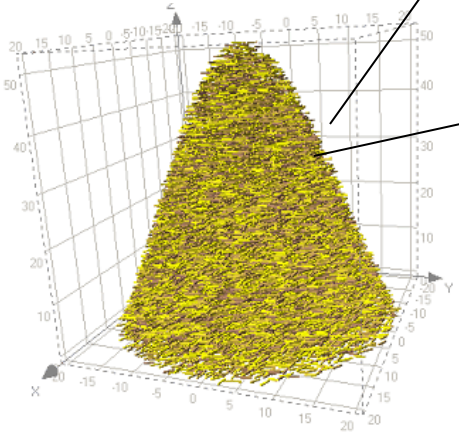
Data center side

- 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

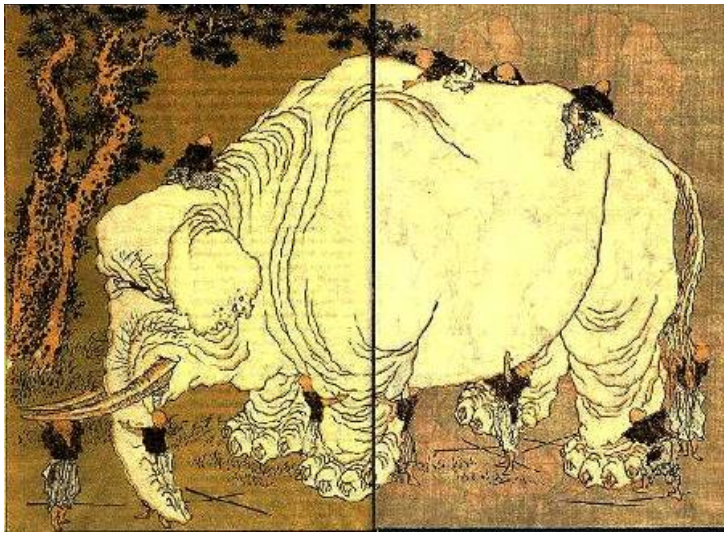
Data Analysis

- Looking for
  - Needles in haystacks – the Higgs particle
  - Haystacks: Dark matter, Dark energy
- Needles are easier than haystacks
- 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
- Requires combination of
  - statistics &
  - computer science



Analysis and Databases

- Much statistical analysis deals with
  - Creating uniform samples –
  - data filtering
  - Assembling relevant subsets
  - Estimating completeness
  - Censoring bad data
  - Counting and building histograms
  - Generating Monte-Carlo subsets
  - Likelihood calculations
  - Hypothesis testing



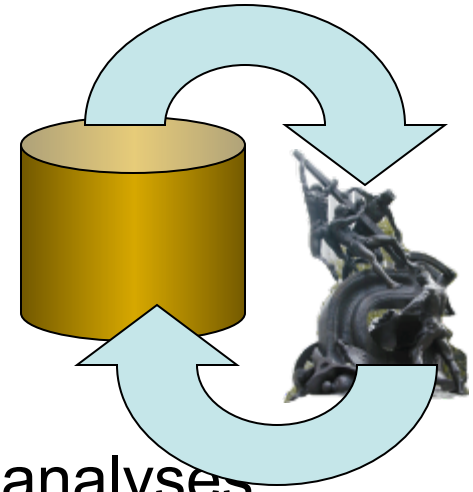
- Traditionally performed on files
- These tasks better done in structured store with
  - indexing,
  - aggregation,
  - parallelism
  - query, analysis,
  - visualization tools.

Accessing Data

- If there is too much data to move around,

  take the analysis to the data!

- Do all data manipulations at database
  - Build custom procedures and functions in the database
- Automatic parallelism guaranteed
- Easy to build-in custom functionality
  - Databases & Procedures being unified
  - Example temporal and spatial indexing
  - Pixel processing
- Easy to reorganize the data
  - Multiple views, each optimal for certain analyses
  - Building hierarchical summaries are trivial
- Scalable to Petabyte datasets



active databases!

Data Mining
(Knowledge Discovery in Database – KDD)

i-explosion

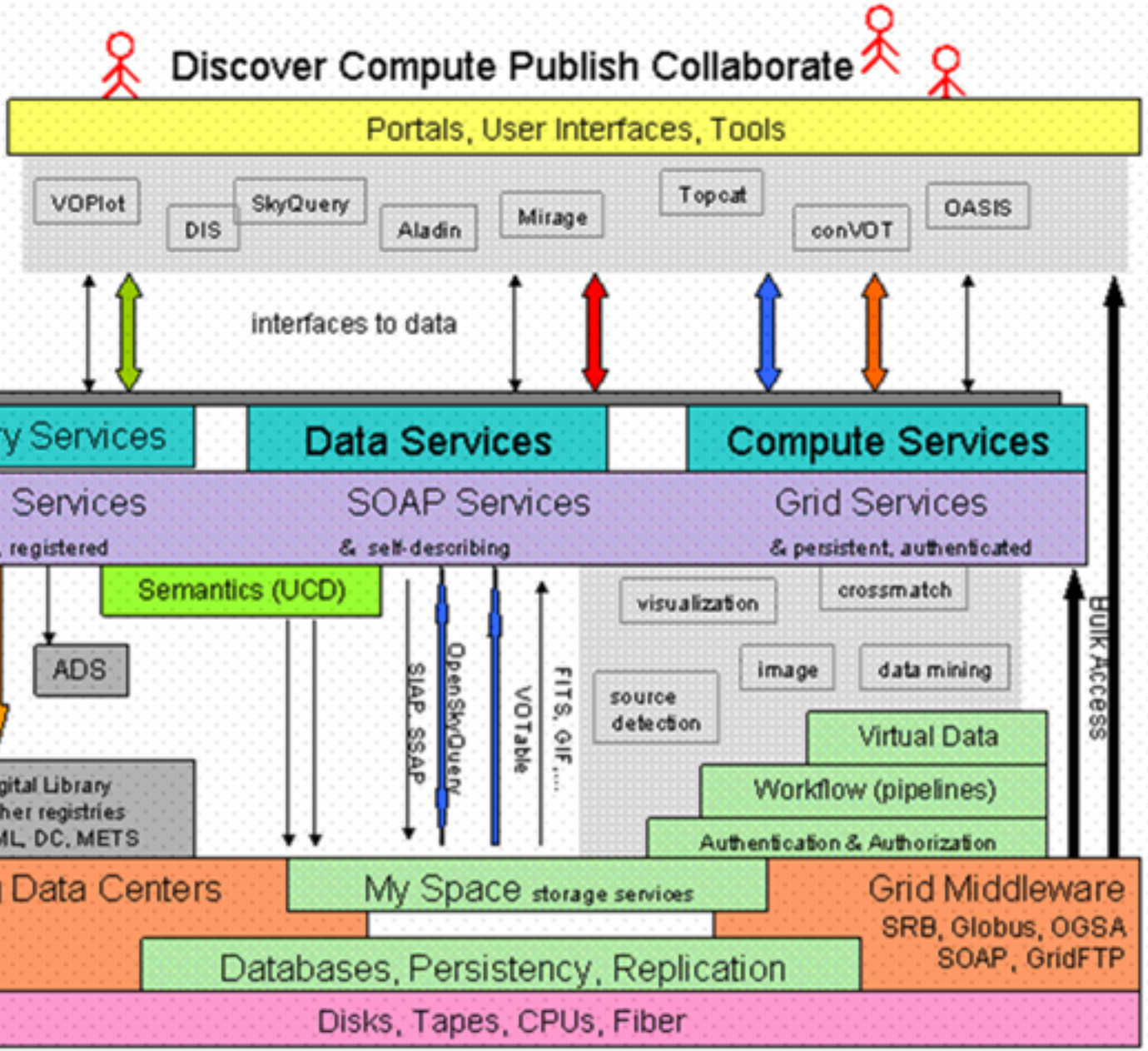
- the process of extracting patterns from data
- currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery

- DM tasks
  - Clustering
  - Classification
    - decision tree
    - nearest neighbor
    - neural networks
    - naïve Bayesian classification
    - support vector machines
  - Regression
  - Association rule learning

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
- Challenging researchers in exploring the deluge of data

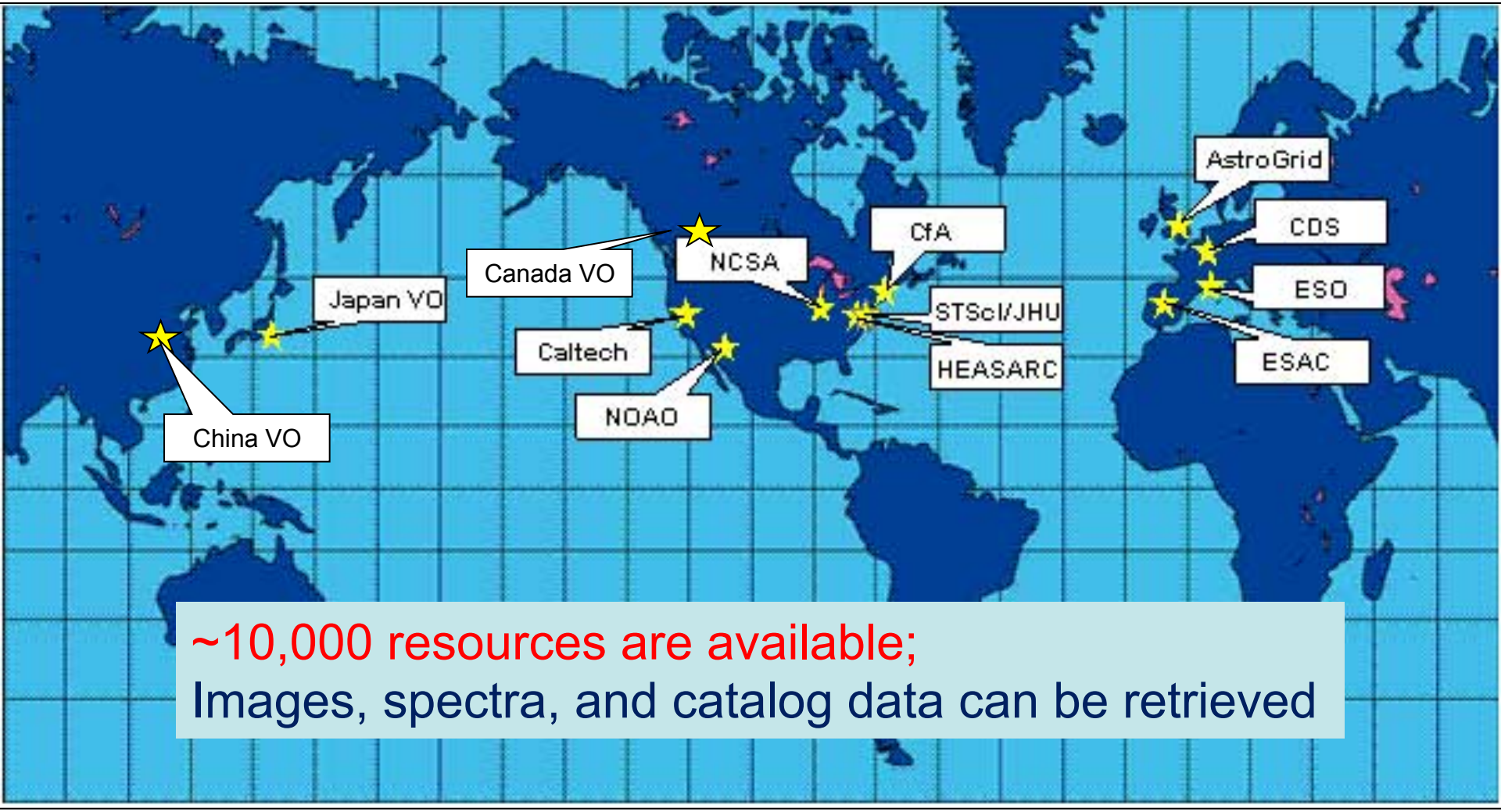
Virtual Observatory Architecture



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
- Storage Interface (VOSpace)
- Message passing among Apps. (SAMP)

Astronomical Virtual Observatories
~ Data Grid ~



~10,000 resources are available;
Images, spectra, and catalog data can be retrieved

Successful Models on Data Sharing

1. Protein Data Banks (PDB)
2. OneGeology/CGI model
3. Intergovernmental Panel of Climate Change (IPCC)
4. International Virtual Observatory (IVOA)

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

Issues to be considered

More Science-Driven

- Demonstrate scientific merit
  - Publish “product papers” by yourselves
- Select most commonly used functionalities
- Quality Index
  - Toward quality assurance, jointly with observatories
- Young researchers
  - Researchers are VERY conservative !
  - Young researchers tend to show interest to new ones

Users View Point

i-explosion

- Easiness to use
  - self-explanatory
  - Basic functionalities are sufficient
  - Others could be done by a local machine
- Market research
  - Science use cases
  - tutorials
- Novice vs Expert
  - GUI vs CUI
  - Almost no scientists know SQL

For Data providers

- Give credit to them
  - Hard and invisible to prepare science-ready data
- Easy implementation
  - tool kit
  - avoiding unwanted bugs as possible
- Validation tool prior to publication of data
  - Ensure reliability of the data product

Technology

i-explosion

- Not too early, not too late
- Stability, robustness
  - "doable or not" is the issue
  - Users do NOT care about invisible technology
- Sustainability, support
- Popularity
  - help desk around you
- Platform independency
  - for easy dissemination

Summary

- An era of Data Intensive Sciences has come or will come soon.
- Need to be ready in each domain through establishing domain-specific standards (metadata, data access, data format, and others)
- Statistical analysis would be a must for Data Intensive Sciences
- For sustainability, it is needed to get support from individual research community and funding agencies