Data Discovery in and Science Results by means of VOs

Masatoshi OHISHI （大石雅寿）
Director, Astronomy Data Center,
National Astronomical Observatory of Japan
&
President of Commission 5, IAU
masatoshi.ohishi@nao.ac.jp
Structure of my Lecture

- Era of Data Intensive Sciences
  - toward “4\textsuperscript{th} paradigm”
- Data Discovery in Astronomy
  - How to find necessary data for our research
- Towards Standardization
  - Differences can be overcome
- Examples of Science Results
  - Just a flavor
- Data Science in Other Fields
  - We share the same problems
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
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
Two Major Categories

<table>
<thead>
<tr>
<th>Pointing Obs.</th>
<th>Surveys</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ALMA</td>
<td>• LSST</td>
</tr>
<tr>
<td>• JWST</td>
<td>• Pan-STARRs</td>
</tr>
</tbody>
</table>

- Large collecting area
- High resolution
- Whole sky
- Time-domain astronomy

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
Science Paradigms

- Thousand years ago: science was **empirical** -- observations / experiments
- Last few hundred years: **theoretical** studies
- Last few decades: **simulations**
- Today: **data exploration** (e-Science) unify theory, experiment, and simulation
  - High-speed network
  - Computers, storages, databases
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 (e-Science) 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
Are we prepared for such a new era?
### Requirements in the Data Intensive Science Era

<table>
<thead>
<tr>
<th>Data producer side</th>
<th>Data center side</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Definition of data quality index, and establishment</td>
<td>• Establishment of data handling environment</td>
</tr>
<tr>
<td>Quality assurance of data (from obs. to data analyses)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>incl. data mining, knowledge discovery, statistics, event discovery</td>
</tr>
<tr>
<td></td>
<td>– High-speed network</td>
</tr>
</tbody>
</table>
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 $N \log N$
• Must accept approximate answers
• New algorithms
• Requires combination of
  – statistics &
  – computer science
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!**
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.
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
mystery outliers

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

[1] S.G. Djorgovski et al., 2005

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.

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
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
Resource Metadata

• Resource Identification:
  – Title, ShortName, Identifier

• Curation:
  – Publisher, PublisherID, Creator, Creator.Logo, Contributor, Date, Version, Contact.Name, Contact.Email

• General content:
  – Subject, Description, Source, ReferenceURL, Type, ContentLevel, Relationship, RelationshipID
Exchange of Meta Data: OAI-PMH
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 …

<table>
<thead>
<tr>
<th>Select</th>
<th>imageURL, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>naoj:imageData</td>
</tr>
<tr>
<td>Where</td>
<td>pos=Point(24,5) and size=0.2 and format='VOTable'</td>
</tr>
</tbody>
</table>
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**

**Applications**

**Web portal**
over 10,000 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
International Endorsements

- IAU XXVth GA Res. (2003 Jul.)
- OECD Rec. (‘04 Aug)
  - place archives that may be accessible via internet
  - provide adequate funding as long-term issues
- IAU DivXII (union-wide activity)
  - VO WG to endorse IVOA Recs since 2006
Examples of Science Results
VO Science

VO enables researchers

1) to find a small particular data subset from a large collection of catalog and observation data

2) to retrieve and use large amount data in an automated way

VO science papers

✓ http://www.ivoa.net/newsletter/ 150 refereed papers
✓ Over 1600 related publications
✓ More and more VO science papers are coming
✓ However, most of them are type-1 science case
✓ type-2 science case?
Refereed Papers by Euro-VO

http://www.euro-vo.org/pub/fc/papers.html

43 papers

- Scale Lengths of Disk Galaxies Fathi K., Allen M., Boch Th., Hatziiminaoglou E., Peletier R., MNFRAS, in press
- SDSS150654.27+013351.6: the second compact elliptical galaxy in the NGC876 group Chilingarian I & Bergond G., MNFRAS Letters, in press
- VANTOS-Integrated Tools and Services for Large-Scale Astrophysical Visualization Becciani et al., 2010, PASP, 122, 119
- The shaping of planetary nebula Sh2-188 through interaction with the interstellar medium Komossa S., 2007, AN, 328, 356
- Initial data release from the INT Photometric H SelectCal van Dokkum et al., 2007, PASP, 119, 898
- Scale Lengths of Disk Galaxies
- The shaping of planetary nebula Sh2-188 through interaction with the interstellar medium W., 2006, A&A, 456, 789
- Young stars and brown dwarfs surrounding Orionis Young Stars as Observed with ROSAT Bayo A., 2007, AN, 328, 356
- VisIVO Scale Lengths of Disk Galaxies
- Initial data release from the INT Photometric H SelectCal van Dokkum et al., 2007, PASP, 119, 898
- sQSO algorithms, data centre and visualization tools Auden E., Toutain T., Zharkov S., 2007, AN, 328, 356
- The shaping of planetary nebula Sh2-188 through interaction with the interstellar medium W., 2006, A&A, 456, 789
- The construction of the large quasar astronomical catalogue (LQAC) M., 2007, AN, 328, 356
- Ultracompact Dwarf Galaxies
- Young stars and brown dwarfs surrounding Orionis Young Stars as Observed with ROSAT Bayo A., 2007, AN, 328, 356
- VisIVO Scale Lengths of Disk Galaxies
- Initial data release from the INT Photometric H SelectCal van Dokkum et al., 2007, PASP, 119, 898
- sQSO algorithms, data centre and visualization tools Auden E., Toutain T., Zharkov S., 2007, AN, 328, 356
- The shaping of planetary nebula Sh2-188 through interaction with the interstellar medium W., 2006, A&A, 456, 789
Classification on VO papers

- Database (11)
- AGN/QSO (7)
- Star/White Dwarf (9)
- Cluster of galaxy (3)
- Sun (2)
- Galaxy (4)
- Others (1)

Distribution of VO Apps Usage

- TOPCAT (8)
- VOSpec (1)
- GAVO Xmatch (1)
- Montage (1)
- DataScope (1)
- STILTS (1)
- SkyView (2)
- AstroGrid Desktop (3)
- OpenSkyQuery (4)
- VOPlot (4)
- Aladin (11)
Use Cases in Using VOs (1)

✓ Specify data services; similar to the traditional ways
  – Find suitable data services, and access them one by one
    ➢ VOs provide user interfaces to access DBs
  – Simple & homogeneous query IFs
    ➢ Region queries, Get all data, etc.
    ➢ Possible to use a single IF in accessing different data services
  – Web portals vs applications
    ➢ Web portals : no need to install IFs
    ➢ Apps : provides with high-level GUIs
  – Save results in the VOTable format
    ➢ Readable in various VO apps

- Exploration of Fossil groups by means of VOs
- Fossil group:
  - A system with an isolated bright, giant elliptical galaxy at its center
  - Mass & X-ray brightness ~ local group
  - Possibly a final status after merges and/or coagulations

- Method
  - Use OpenSkyQuery (NVO portal)
  - Cross-matching SDSS LRG & Rosat All Sky catalogues
  - Find elliptical galaxies with extended X-ray emissions
  - Associated galaxies within $0.5 \, h^{-1}_{70} \, \text{Mpc}$ are dimmer by more than 2 mag

- 34 candidates (only 15 before)
Fossil Group

Chandra X-ray observations of the giant elliptical galaxy NGC 6482

DSS image of NGC 6482
Meeting of the Minds 3 August 2011

Ex. 2: “ALBUS 1: A Very Bright White Dwarf Candidate”

- Discovery of (candidate of) a bright white dwarf ($B_T = 11.8$)
- Accidental discovery while investigating regions around Alnilam ($\varepsilon$ Ori) and Mintaka ($\delta$ Ori)
- X-match Tycho-2 and 2MASS catalogues by Aladin
- Extremely blue star on a color-mag diagram
- 12th brightest white dwarf (isolated)

Fig. 1.—False-color composite image, 5.6 x 5.6 arcmin$^2$ wide, centered on Albus 1, North is up, east is left.
Color–Mag diagram

Fig. 2.—$V_T$ vs. $V_T - K_s$ color-magnitude diagram from the data in J. A. Caballero & E. Solano (2007, in preparation). Tycho-2/2MASS sources with proper motions larger and smaller than 15 mas yr$^{-1}$ are shown with crosses and dots, respectively. Albus 1 is highlighted with a big filled star.

SED of the candidate (Albus 1)

Fig. 4.—Spectral energy distributions of Albus 1, the DA1 white dwarf G191-B2B, and the B2 Vp star σ Ori E (shifted to a heliocentric distance of 0.5 kpc). The seven passbands ($B, V, R, I, J, H, K_s$) are indicated.

* Confirmed to be a B subdwarf by spectrum observations

“CPD -20 1123 (Albus 1) Is a Bright He-B Subdwarf”
Use Cases in Using VOs (2)

✓ **Cross-Query to multiple datasets**
  - Query multi-wavelength data in a single query within (an) interested sky region(s)
  - Easier data discoveries
  - Easier to utilize multiple DBs

Possible to conduct automated data reductions of huge data

→ makes it possible to conduct (effectively) impossible research in the past
JVO portal
http://jvo.nao.ac.jp/portal

- 10,551 Data Resources
  - 7,397 Catalogs
  - 208 Image Services
  - 84 Spectrum Services
  - ...

- Reduced Subaru Data
  - Suprime-Cam
  - HDS
JVO Subaru archive

✓ Suprime-Cam data reduction system
  - Data archive and parallel computing system are connected with a dedicated network (128Gbps)
  - The whole data can be processed in two weeks (using 48 CPU cores)

✓ VO access as well as a dedicated GUI
  - Data retrieval is “programmable”
  - Possible to retrieve cutout image for specified region
  - No need to download all the data (~10TB), data can be retrieved on demand
# GUI for Suprime-Cam archive

## JVO Sky

**Object Name** | Suprime-Cam/Subaru | HDS/Subaru | Suzaku
---|---|---|---

**Coordinates or ObjectName:**

- Suzaku: 3/9, hds: 3/3, spcam: 38/30, searchtime: 48ms

**Image Details:**
- Title: 12h24m13s14d05m03 (VIRGO_FIELD_2 W-C-RC)
- Link: [Image Link](#)
- Type: Image
- Center: 186.06584603 14.19427885
- Band name: W-C-RC

## Object Table

<table>
<thead>
<tr>
<th>Object Name</th>
<th>B-J</th>
<th>B-V</th>
<th>B-C-RC</th>
<th>B-C-IC</th>
<th>V-S-1a</th>
<th>V-S-2a</th>
<th>V-J-J</th>
<th>V-K-K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subname1</td>
<td>0.5</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>Subname2</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Subname3</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Subname4</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

**Update:**

- Total Number: 71

---

**Image with Diagrams:**

- Diagram 1: Visual representation of the Suprime-Cam archive
- Diagram 2: Detailed view of a specific object
- Diagram 3: Analysis of spectral data

---

**Additional Information:**

- CSP 2011
- 2011 Sep 28
“Early Science Result from the Japanese Virtual Observatory: AGN and Galaxy Clustering at z = 0.3 to 3.0”

✓ Measurement of AGN-Galaxy cross-correlation
  – Fueling mechanism of AGN
  – Co-evolution of galaxy and black hole
  – Use all the data of Suprime-Cam archive (nobody have done !)

✓ Previous works
  – Redshift measurement → 3D cross-correlation
  – z < 0.6 : Good statistics using SDSS data
  – z > 0.6 : Relatively poor stat. (several tens).
    • Hard to improve statistics:
    • Statistics at small scale (~1Mpc) is extremely poor
    • Affected by Cosmic Variance ← small number of samples
    • Biased to red galaxies in spectroscopic target selection
Dataset

AGN samples

✓ Veron QSO/AGN Catalog (12\textsuperscript{th} ed)
✓ SDSS DR–5 QSO Catalog (4\textsuperscript{th} ed)

Galaxy samples

✓ JVO Suprime–Cam Archive (B, V, R, I, i’, z’ bands)
✓ UKIDSS DR–2 Catalog (K band)
✓ Deepest observation data was used for each AGN

<table>
<thead>
<tr>
<th>Suprime–Cam</th>
<th>484</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKIDSS</td>
<td>1325</td>
</tr>
<tr>
<td>Total</td>
<td>1809</td>
</tr>
</tbody>
</table>
Search Suprime-Cam image around AGN

SELECT qso.*, img.*
FROM ivo://jvo/vizier/VII/235:qso_veron_2006 AS qso,
     ivo://jvo/subaru/spcam:image_cutout AS img
WHERE qso.z >= 1.0 and qso.z < 1.1
AND img.region = Circle(qso.raj2000, qso.dej2000, 0.14)

Similar to SQL (Structured Query Language)
Search Result

Only a part of data is displayed

No way to download all the images at once

Download coordinates in CSV format

Create a list of AGNs observed with Suprime-Cam

2011 Sep 28 CSP 2011
Automate using a script

✓ Download and analysis for ~12,000 AGNs
  – hard or impossible to do manually

✓ Make a script (e.g. shell script)
  – Script to create a catalog from retrieved image data ...
  – Execute this script for each AGN
  – 12,000 AGNs ➞ 40 parallels. Completed in one day.

✓ Access to VO
  – Use command line access tool.
  – Useful for repeating the same query by changing a query condition.
AGN redshift & absolute mag

Fig. 5. K-corrected V band absolute magnitude vs redshift of the AGNs used in this work. Open circles represent AGN samples for which the galaxy sample is derived from the UKIDSS data, and the crosses represent AGN samples for which the galaxy sample is derived from the Suprime-Cam data.
Analysis

Density of galaxies within $r_p : n(r_p); \gamma = 1.8$

$$< n(r_p) > = r_p \left( \frac{r_0}{r_p} \right)^\gamma \frac{\Gamma(1/2)\Gamma((\gamma - 1)/2)}{\Gamma(\gamma/2)} < \rho_0 > + < n_{bg} >$$

**z1-D**
- $z = 0.3 \sim 0.6$
- Dim AGNs (651)
- $r_0 = 5 \pm 2 \ h^{-1}\text{Mpc}$

**z4-B**
- $z = 1.3 \sim 1.8$
- Bright AGNs (142)
- $r_0 = 11 \pm 6 \ h^{-1}\text{Mpc}$

Analysis

✓ Projected correlation function: \( \omega(r_p) \)

\[
\omega(r_p) = 2 \int_0^\infty \xi(r_p, \pi) d\pi = r_p \left( \frac{r_0}{r_p} \right)^\gamma \frac{\Gamma(1/2)\Gamma((\gamma - 1)/2)}{\Gamma(\gamma/2)}
\]

\[
\xi(r) = (r_0/r)^\gamma.
\]

Transverse comoving distance (Mpc)

<table>
<thead>
<tr>
<th>z1-D</th>
<th>z = 0.3~0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim AGNs (651)</td>
<td></td>
</tr>
<tr>
<td>( r_0 = 5 \pm 2 ) h(^{-1})Mpc</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>z4-B</th>
<th>z = 1.3~1.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bright AGNs (142)</td>
<td></td>
</tr>
<tr>
<td>( r_0 = 11 \pm 6 ) h(^{-1})Mpc</td>
<td></td>
</tr>
</tbody>
</table>

Transverse comoving distance (Mpc)
Result

$\sigma_8$: rms of correlation function at < 8 Mpc

Galaxy-Galaxy
  - Top: Ks selected, Middle: r-band selected, Bottom: b_j-band selected
- LRG-LRG (Wake et al. 2008)
  - upper: radio-detected LRG
  - lower: radio-quiet LRG

This work (AGN-Galaxies)
- △ Dim AGN
- ○ Bright AGN

QSO-QSO
- ♦ SDSS (Shen et al. 2007)
- ■ SDSS (Ross et al. 2009)

Clustering increases for higher redshift

AGN-Galaxies
- ○ Bornancini et al. 2007
- △ Coil et al. 2007
- □ Norman et al. 2009
- ♦ Hickox et al. 2009
  - Radio, X-ray, IR selected AGN from top to bottom
Data Science in Other Fields
Virtual Observatories in Planetary Sciences

- ease discovery, access and use of planetary data
- NASA, ESA, JAXA and others
- Refers to the IVOA standard protocols, w/ some modifications

http://planetarydata.org/
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)
World Data System

• “Virtual Observatories” in a variety of science fields = advanced interconnections between data management components for disciplinary and multidisciplinary applications

• Organized under the ICSU (Int’l Council for Sciences)
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
References


[6] Y. Shirasaki et al., 2011: Early Science Result from the Japanese Virtual Observatory: AGN and Galaxy Clustering at z = 0.3 to 3.0, 63, PASJ, No.SP2, 469-491
References

JVO (Japan Vatial Observatory)
http://jvo.nao.ac.jp/portal

IVOA (International Virtual Observatory Alliance)
http://www.ivoa.net/

IPDA (International Planetary DATA Alliance)
http://planetarydata.org/

The fourth Paradigm: Data-Intensive Scientific Discovery