Virtual Astronomy, Information technology, and the New Scientific Methodology

S. G. Djorgovski (Caltech)

ECURE ’05 Conference, ASU, 01 March 2005

We Are Entering the Second Phase of the Information Technology Revolution

First, the advent of cheap and ubiquitous computing and the rise of the WWW.
Now, the rise of the information-driven computing.

There is a great emerging synergy between the computationally enabled science, and the science-driven information technology.

An Overview:

• Astronomy in the era of information abundance
  The IT revolution, challenges and opportunities
• The Virtual Observatory concept
  An example of a new type of a scientific enterprise
• Virtual Observatory status
  Where are we now, where are we going
• From technology to science (and back)
  New tools for the science of 21st century
• Musings on cyber-science in general
  The changing nature of scientific inquiry
• The new roles of research libraries
  The changing nature of data and information needs

Facing the Data Tsunami

Astronomy, all sciences, and every other modern field of human endeavor (commerce, security, etc.) are facing a dramatic increase in the volume and complexity of data.
Astronomy is Now a Very Data-Rich Science

Multi-Terabyte (soon: multi-PB) 

**sky surveys and archives** over a broad range of wavelengths …

1 microSky (DPOSS)

Billions of detected sources, hundreds of measured attributes per source …

1 nanoSky (HDF-S)

- **Large digital sky surveys** are becoming the dominant source of data in astronomy: ~ 10-100 TB/survey (soon PB), ~ $10^6 - 10^9$ sources/survey, many wavelengths…

- **Data sets** many orders of magnitude larger, more complex, and more homogeneous than in the past

Data ➔ Knowledge?

The exponential growth of data volume (and also complexity, quality) driven by the exponential growth in detector and computing technology

... but our understanding of the universe increases much more slowly!

Panchromatic Views of the Universe:

Data Fusion ➔ A More Complete, Less Biased Picture

Radio
Far-Infrared
Visible
Visible + X-ray
Dust Map
Galaxy Density Map
Exploration of the Time Domain in Astronomy

The advent of **Synoptic Sky Surveys**: things that move, and things that go BANG! in the night…

This will generate multi-Petabyte data sets which must be analysed in a (near) real time.

**NEAT Sedna**

Theoretical Simulations Are Also Becoming More Complex and Generate TB’s of Data

Structure formation in the Universe

Supernova explosion instabilities

VO: Conceptual Architecture

- **Astronomy community response to the scientific and technological challenges posed by massive data sets**
  - Harness the modern information technology in service of astronomy, and partner with it

- **A complete, dynamical, distributed, open research environment for the new astronomy with massive and complex data sets**
  - Provide content (data, metadata) services, standards, and analysis/compute services
  - Federate the existing and forthcoming large digital sky surveys and archives, facilitate data inclusion and distribution
  - Develop and provide data exploration and discovery tools
  - *Technology-enabled, but science-driven*
Why is VO a Good Scientific Prospect?

• Technological revolutions as the drivers/enablers of the bursts of scientific growth

• Historical examples in astronomy:
  – 1960’s: the advent of electronics and access to space
      *Quasars, CMBR, x-ray astronomy, pulsars, GRBs, …*
  – 1980’s - 1990’s: computers, digital detectors (CCDs etc.)
      *Galaxy formation and evolution, extrasolar planets, CMBR fluctuations, dark matter and energy, GRBs, …*
  – 2000’s and beyond: information technology

*The next golden age of discovery in astronomy?*

VO is the mechanism to effect this process

A Systemic View of the NVO

Information Technology ➔ New Science

• The information volume grows exponentially
  
  *Most data will never be seen by humans!*

  ➔ The need for data storage, network, database-related technologies, standards, etc.

• Information complexity is also increasing greatly
  
  *Most data (and data constructs) cannot be comprehended by humans directly!*

  ➔ The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery …

• VO is the framework to effect this for astronomy

A Modern Scientific Discovery Process

**Data Gathering**

**Data Farming:**

- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability

**Data Mining** (or Knowledge Discovery in Databases):

- Pattern or correlation search
- Clustering analysis, automated classification
- Outlier / anomaly searches
- Hyperdimensional visualization

**Data Understanding**

**New Knowledge**
How and Where are Discoveries Made?

- **Conceptual Discoveries:** e.g., Relativity, Quantum Mechanics, Strings, Inflation… **Theoretical, may be inspired by observations**

- **Phenomenological Discoveries:** e.g., Dark Matter, Dark Energy, QSOs, GRBs, CMBR, Extrasolar Planets, Obscured Universe … **Empirical, inspire theories, can be motivated by them**

Phenomenological Discoveries:
- Pushing along some parameter space axis \( \text{VO useful} \)
- Making new connections (e.g., multi-\( \lambda \)) \( \text{VO critical!} \)

Understanding of complex astrophysical phenomena requires complex, information-rich data (and simulations?)

The Mixed Blessings of Data Richness

Modern digital sky surveys typically contain \( \sim 10 - 100 \) TB, detect \( N_{\text{obj}} \sim 10^8 - 10^9 \) sources, with \( D \sim 10^2 - 10^3 \) parameters measured for each one -- and PB data sets are on the horizon

Potential for discovery \( \{ \begin{align*} N_{\text{obj}} \text{ or data volume} & \quad \rightarrow \text{Big surveys!} \vspace{1em}\vspace{1em} N_{\text{surveys}}^2 \text{ (connections)} & \quad \rightarrow \text{Data federation} \end{align*} \)

Great! However …

It takes minutes to hours to search 1 TB (you’d like a few seconds to minutes); 1 PB will take a day to a few months!

We better do it right the first time …

Or do something more clever (db structuring, statistics?)

… And Moreover …

- **DM algorithms tend to scale very badly:**
  - Clustering \( \sim N \log N \rightarrow N^2, \sim D^2 \)
  - Correlations \( \sim N \log N \rightarrow N^2, \sim D^k \ (k \geq 1) \)
  - Likelihood, Bayesian \( \sim N^m \ (m \geq 3), \sim D^k \ (k \geq 1) \)

- **Visualization fails for \( D > 3 - 5 \)**
  - An inherent limitation of the human mind?

- We need better DM algorithms and some novel methods for dimensionality reduction (and some AI help?)

- Or, we learn to accept approximate results
  - Sometimes that is good enough, sometimes not
Scientific Roles and Benefits of a VO

- Facilitate science with massive data sets (observations and theory/simulations) **efficiency amplifier**
- Provide an **added value** from federated data sets (e.g., multi-wavelength, multi-scale, multi-epoch ...)
  - Discover the knowledge which is present in the data, but can be uncovered only through data fusion
- **Enable and stimulate some qualitatively new science** with massive data sets (not just old-but-bigger)
- **Optimize the use of expensive resources** (e.g., space missions, large ground-based telescopes, computing ...)
- Provide R&D drivers, application testbeds, and stimulus to the **partnering disciplines** (CS/IT, statistics ...)

VO Developments and Status

- The concept originated in 1990’s, developed and refined through several conferences and workshops
- Major blessing by the National Academy Report
- **In the US:** National Virtual Observatory (NVO)
  - Concept developed by the NVO Science Definition Team (SDT). See the report at [http://www.nvosdt.org](http://www.nvosdt.org)
  - NSF/ITR funded project: [http://us-vo.org](http://us-vo.org)
  - A number of other, smaller projects under way
- **Worldwide** efforts: International V.O. Alliance
- A good synergy of astronomy and CS/IT
- Good progress on data management issues, a little on data mining/analysis, first science demos forthcoming

NVO Workflow Components

**Discover Compute Publish Collaborate**

- **Portals, User Interfaces, Tools**
  - VOPlot
  - SkyQuery
  - Aladin
  - Topcat
  - conVOT
  - DADIS
  - OASIS
  - conVOT

- **HTTP Services**
  - web services, registered

- **SOAP Services** & self-describing
  - bulk access

- **Grid Services** & persistent, authenticated
  - Compute services
  - storage services

**NVO Registry Layer**

Existing Data Centers
- **Data Services**
  - semantisc (UCD)
  - SIAP, SSAP
  - VOTable
  - FITS, GIF...

- **Semantic (UCD)**
  - data mining
  - visualization
  - image
  - source detection

- **Workflow (pipelines)**
  - fitting
  - crossmatch

Digital Library
- **Other registries**
  - OAI
  - ADS

- **miscellaneous registries**
  - OAI
  - ADS
  - DADIS

My Space
- **storage services**
  - disks, tapes, CPUs, fiber

Grid Middleware
- **SRB, Globus, COSA**
  - SOAP, GridFTP

**NVO Registry Layer**

- **Existing Data Centers**
  - **Data Services**
    - semantisc (UCD)
    - SIAP, SSAP
    - VOTable
    - FITS, GIF...

- **Semantic (UCD)**
  - data mining
  - visualization
  - image
  - source detection

- **Workflow (pipelines)**
  - fitting
  - crossmatch

Digital Library
- **Other registries**
  - OAI
  - ADS

- **miscellaneous registries**
  - OAI
  - ADS
  - DADIS

My Space
- **storage services**
  - disks, tapes, CPUs, fiber

Grid Middleware
- **SRB, Globus, COSA**
  - SOAP, GridFTP

http://us-vo.org
“What data are available for some object or some region on the sky? Can I get them easily?”

Data Inventory Service

DIS user interface

DIS search results

Pipe them into a data analysis and visualization tool
Broader and Societal Benefits of a VO

- **Professional Empowerment:** Scientists and students anywhere with an internet connection would be able to do a first-rate science. A broadening of the talent pool in astronomy, democratization of the field.
- **Interdisciplinary Exchanges:**
  - The challenges facing the VO are common to most sciences and other fields of the modern human endeavor.
  - Intellectual cross-fertilization, feedback to IT/CS.
- **Education and Public Outreach:**
  - Unprecedented opportunities in terms of the content, broad geographical and societal range, at all levels.
  - Astronomy as a magnet for the CS/IT education.

“Weapons of Mass Instruction”

Do We Know How to Run a VO?

- The VO is **not** yet another data center, archive, mission, or a traditional project.
  - It does not fit into any of the usual structures today.
    - It is inherently **distributed**, and web-centric.
    - It is fundamentally based on a **rapidly developing technology** (IT/CS).
    - It transcends the traditional boundaries between different wavelength regimes, agency domains.
    - It has an unusually broad range of constituents and interfaces.
    - It is inherently **multidisciplinary**.
- The VO represents a **novel type of a scientific organization** for the era of information abundance.
Now Let’s Take A Look At Some Relevant Technology Trends …

Exponentially Declining Cost of Data Storage

Computing is Cheap …

Today (~2004), 1 $ buys:
- 1 day of CPU time
- 4 GB (fast) RAM for a day
- 1 GB of network bandwidth
- 1 GB of disk storage for 3 years
- 10 M database accesses
- 10 TB of disk access (sequential)
- 10 TB of LAN bandwidth (bulk)
- 10 KWh = 4 days of computer time

… Yet somehow computer companies make billions: you do want some toys, about $ 10^5 worth ≈ 1 postdoc year

… But People are Expensive!
People ~ Software, maintenance, expertise, creativity …
Moving Data is Slow!

How long does it take to move a Terabyte? (~Petabyte)

<table>
<thead>
<tr>
<th>Context</th>
<th>Speed</th>
<th>Rent $/month</th>
<th>$/Mbps</th>
<th>$/TB Sent</th>
<th>Time/TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home phone</td>
<td>0.04</td>
<td>40</td>
<td>1,000</td>
<td>3,086</td>
<td>6 years</td>
</tr>
<tr>
<td>Home DSL</td>
<td>0.6</td>
<td>50</td>
<td>117</td>
<td>360</td>
<td>5 months</td>
</tr>
<tr>
<td>T1</td>
<td>1.5</td>
<td>1,200</td>
<td>800</td>
<td>2,469</td>
<td>2 months</td>
</tr>
<tr>
<td>T3</td>
<td>43</td>
<td>28,000</td>
<td>651</td>
<td>2,010</td>
<td>2 days</td>
</tr>
<tr>
<td>OC3</td>
<td>155</td>
<td>49,000</td>
<td>316</td>
<td>976</td>
<td>14 hours</td>
</tr>
<tr>
<td>OC 192</td>
<td>9600</td>
<td>1,920,000</td>
<td>200</td>
<td>617</td>
<td>14 minutes</td>
</tr>
</tbody>
</table>

100 Mbps       | 100    | 1,920,000    | 200    | 617       | 1 day   |

| Gbps          | 1000   | 1,920,000    | 200    | 617       | 2.2 hours|

Solution: bring the computation to the data!

Disks are Cheap and Efficient

- Price/performance of disks is improving faster than the computing (Moore’s law): a factor of ~ 100 over 10 years!
  - Disks are now already cheaper than paper
- Network bandwidth used to grow even faster, but no longer does
  - And most telcos are bankrupt …
  - Sneakernet is faster than any network
- Disks make data preservation easier as the storage technology evolves
  - Can you still read your 10 (5?) year old tapes?

The Gospel According to Jim Gray:

- Store everything on disks, with a high redundancy (cheaper than the maintenance/recovery)
  - Curate data where the expertise is
- Do not move data over the network: bring the computation to data!
  - The Beowulf paradigm: Datawulf clusters, smart disks …
  - The Grid paradigm (done right): move only the questions and answers, and the flow control
- You will learn to use databases!
- And we need a better fusion of databases and data mining and exploration

An Early Disk for Information Storage

- Phaistos Disk: Minoan, 1700 BC
  - No one can read it 😊

(From Jim Gray)
Some Musings on CyberScience

- Enables a broad spectrum of users/contributors
  - From large teams to small teams to individuals
  - Data volume ~ Team size
  - Scientific returns ≠ f(team size)
  - Human talent is distributed very broadly geographically
- Transition from data-poor to data-rich science
  - Chaotic ➔ Organized … However, some chaos (or the lack of excessive regulation) is good, as it correlates with the creative freedom (recall the WWW)
- Computer science as the “new mathematics”
  - It plays the role in relation to other sciences which mathematics did in ~ 17th - 20th century
  (The frontiers of mathematics are now elsewhere…)

The Evolution of Science

- Astronomical data volume ca. 2004: a few × 10^2 TB (but PB’s are coming soon!)
- All recorded information in the world: a few × 10^7 TB (but most of it is video, i.e., junk)
- The data volume everywhere is growing exponentially, with e-folding times ~ 1.5 yrs (Moore’s law)
  - NB: the data rate is also growing exponentially!
- So, everybody needs efficient db techniques, DM (searches, trends & correlations, anomaly/outlier detection, clustering/classification, summarization, visualization, etc.)
- There is a real possibility of major advances which would change the world (remember the WWW!)

Some Musings on CyberScience

- Enables a broad spectrum of users/contributors
  - From large teams to small teams to individuals
  - Data volume ~ Team size
  - Scientific returns ≠ f(team size)
  - Human talent is distributed very broadly geographically
- Transition from data-poor to data-rich science
  - Chaotic ➔ Organized … However, some chaos (or the lack of excessive regulation) is good, as it correlates with the creative freedom (recall the WWW)
- Computer science as the “new mathematics”
  - It plays the role in relation to other sciences which mathematics did in ~ 17th - 20th century
  (The frontiers of mathematics are now elsewhere…)
The Concept of Data (and Scientific Results) is Becoming More Complex

- Actual data (preserved)
- Virtual data (recomputed as needed)

Primary Data
And Metadata

Produced and often archived by the primary data providers

Derived Data Products
And Results, Increasingly Distilled down

Produced and published by the domain experts

Information is cheap, but expertise and knowledge are expensive!

Scientific Publishing is Changing

- Journals (and books?) are obsolete formats; must evolve to accommodate data-intensive science
- Massive data sets can be only published as electronic archives - and should be curated by domain experts
- Peer review / quality control for data and algorithms?
- The rise of un-refereed archives (e.g., archiv.org): very effective and useful, but highly heterogeneous and unselective
- A low-cost entry to publish on the web
  - Who needs journals?
  - Will there be science blogs?
- Persistency and integrity of data (and pointers)
- Interoperability and metadata standards

Research Libraries for the 21st Century

- How should research libraries evolve in the era of information abundance and complexity?
- What should be their roles / functionality?
  - Data discovery services
  - Data provider federators
  - Primary and/or derived data archivers
  - How much domain expertise should be provided?
  - Quality control?
  - Relationship with web portals and search engines?
- Is this too much for a single type of an institution?
  - Are libraries obsolete (inadequate)?
  - Should they split into several types of institutions?

The Changing Nature of Scientific Data and Results:

Static ➔ Dynamic

- Recalibrations
  - Which versions to save?
- Intrinsically growing data sets
  - Which versions to save?
- Virtual data
  - Re-compute on demand, save just the algorithm, but operating on which input version?
  - What about improved algorithms?
- Domain expertise is necessary!
VO Summary

- National/International Virtual Observatory is an emerging framework to harness the power of IT for astronomy with massive and complex data sets
  - Enable data archiving, fusion, exploration, discovery
  - Cross the traditional boundaries (wavelength regimes, ground/space, theory/observation …)
  - Facilitate inclusion of major new data providers, surveys
  - Broad professional empowerment via the WWW
- It is inherently multidisciplinary: an excellent synergy with the applied CS/IT, statistics…and it can lead to new IT advances of a broad importance
- It is inherently distributed and web-based

But It Is More General Than That:

- Coping with the data flood and extracting knowledge from massive/complex data sets is a universal problem facing all sciences today:
  - Quantitative changes in data volumes + IT advances:
    → Qualitative changes in the way we do science
- (N)VO is an example of a new type of a scientific research environment / institution(?) in the era of information abundance
- This requires new types of scientific management and organization structures, a challenge in itself
- The real intellectual challenges are methodological: how do we formulate genuinely new types of scientific inquiries, enabled by this technological revolution?

… and the Evolution of Libraries

- Scientific / research libraries must evolve, in order to stay useful in the era of data-intensive, computation-based science
  - Database technologies are essential
  - Fusion with data exploration technologies will be next
  - A growing importance of domain expertise
  - Blending in the web, then semantic web?

For more details and links, please see http://www.astro.caltech.edu/~george/vo/