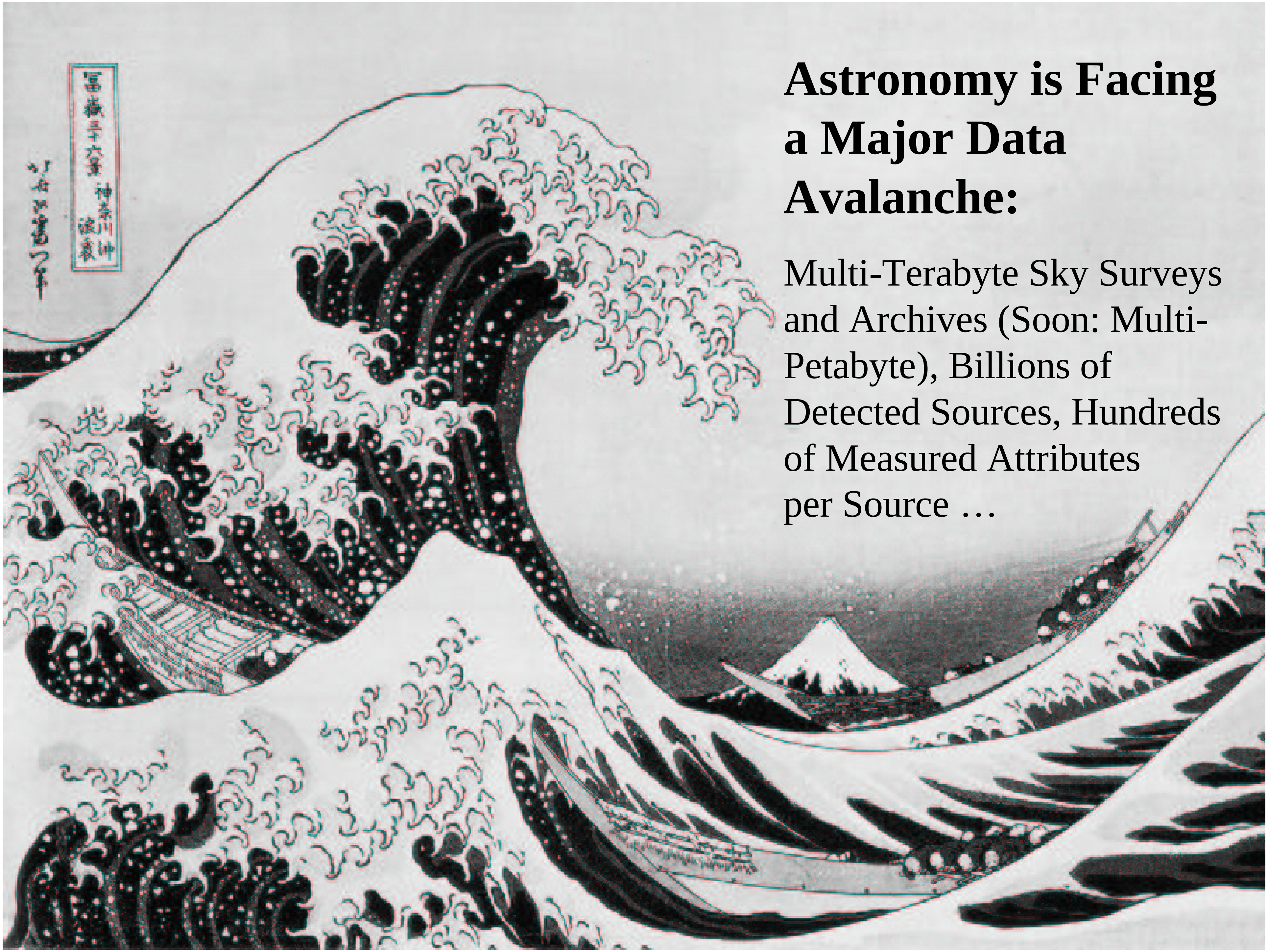
# New Astronomy With a Virtual Observatory

*S. G. Djorgovski (Caltech)*

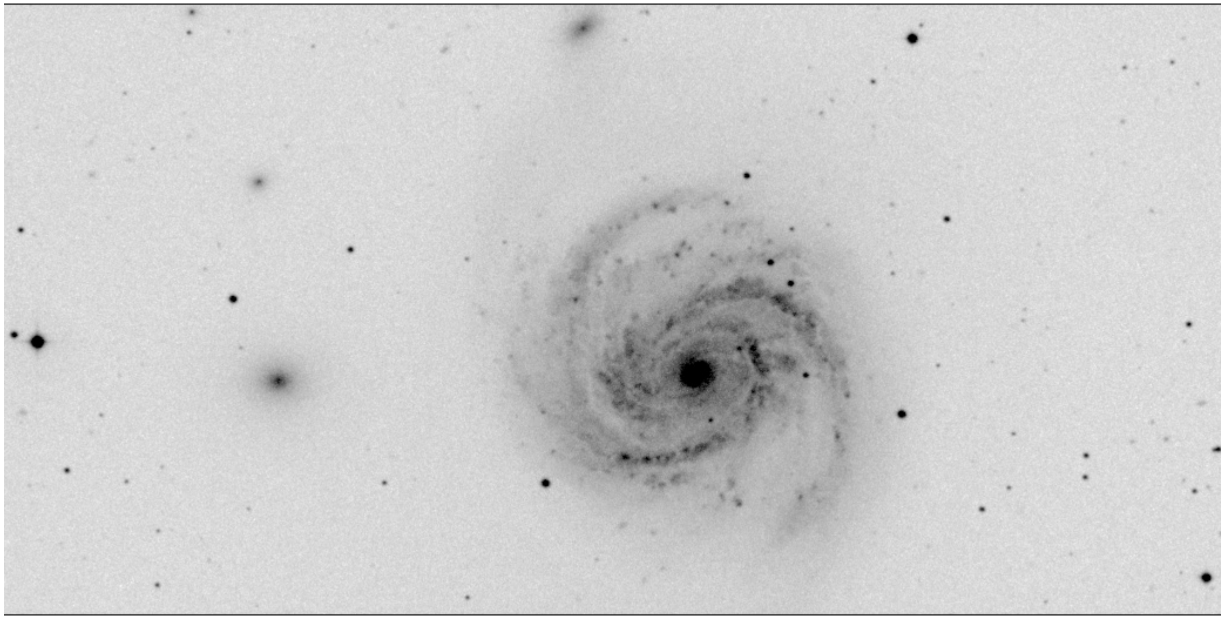
With special thanks to R. Brunner, A. Szalay, A. Mahabal, *et al*.

- Introduction: Astronomy in the Era of Information Abundance
- The Virtual Observatory Concept
- Science With a Virtual Observatory
- Some Ongoing Efforts
- Concluding Comments: What Next?

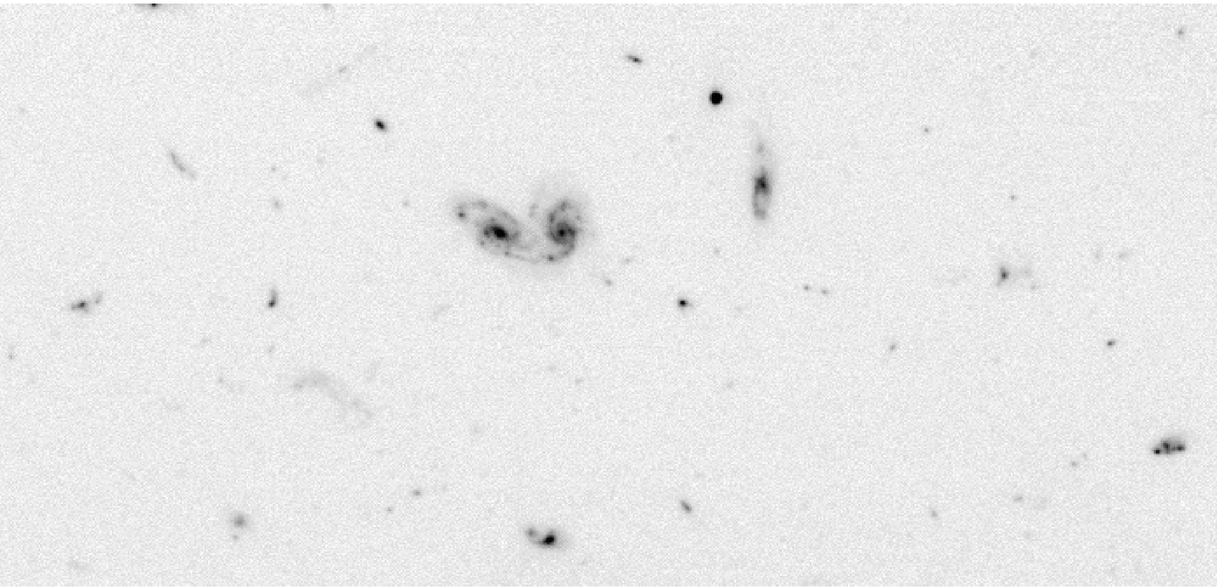


# Astronomy is Facing a Major Data Avalanche:

Multi-Terabyte Sky Surveys and Archives (Soon: Multi-Petabyte), Billions of Detected Sources, Hundreds of Measured Attributes per Source …

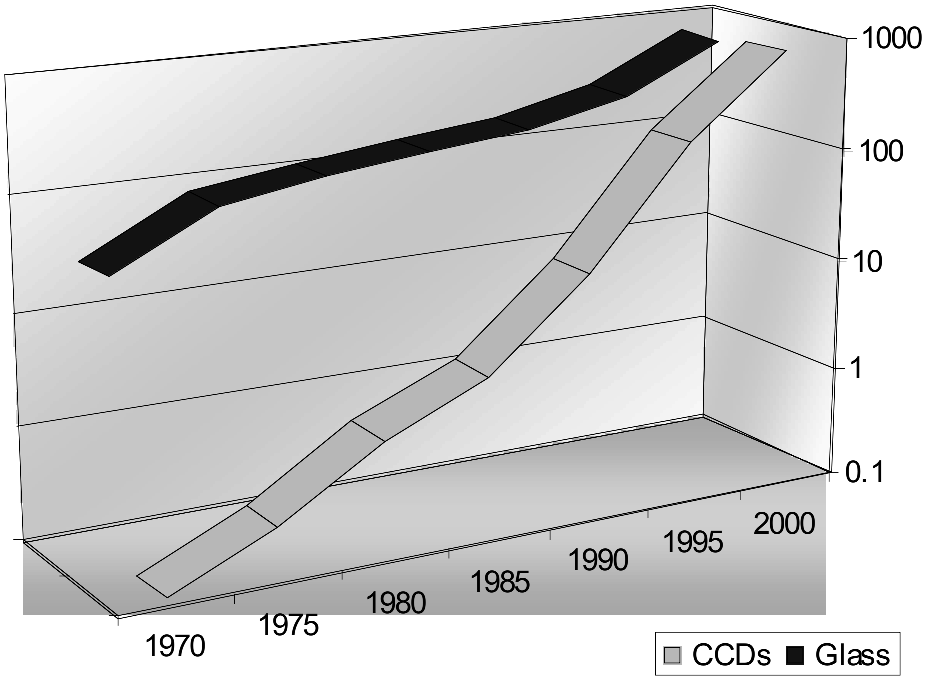


1 microSky
(DPOSS)



1 nanoSky
(HDF-S)

# The Exponential Growth of Information in Astronomy



*Total area of 3m+ telescopes in the world in m$^2$, total number of CCD pixels in Megapix, as a function of time. Growth over 25 years is a factor of 30 in glass, 3000 in pixels.*

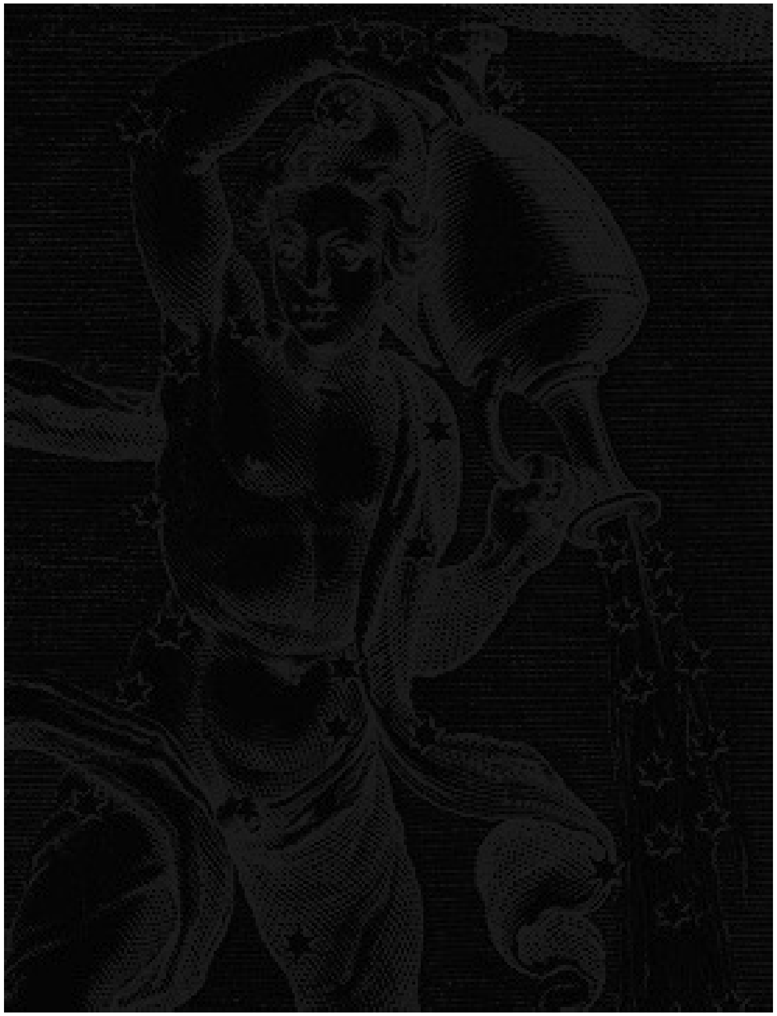
- Moore's Law growth in CCD capabilities/size
- Gigapixel arrays are on the horizon
- Improvements in computing and storage will track the growth in data volume
- Investment in software is critical, and growing

Data Volume
*and Complexity*
are Increasing!

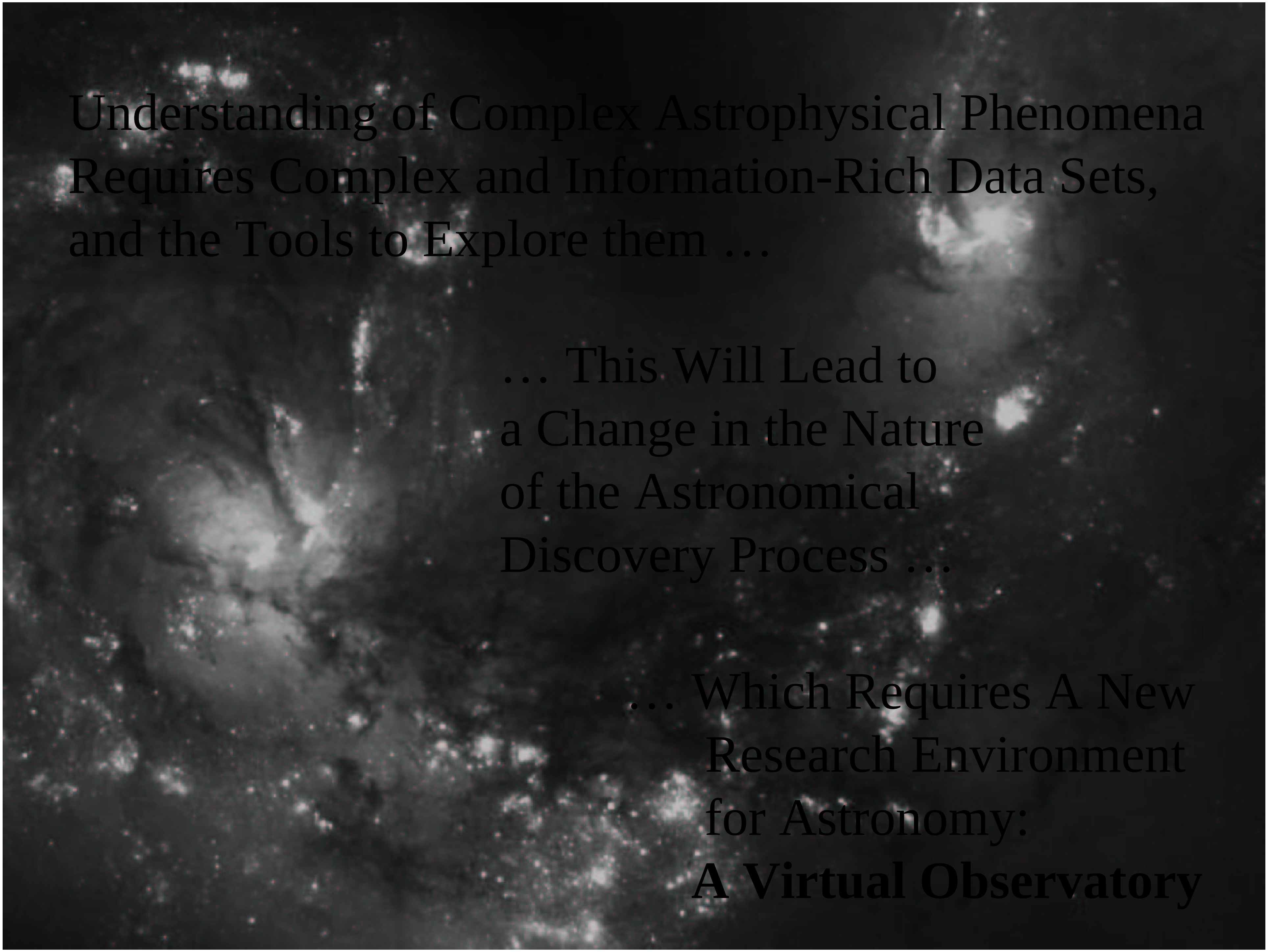
# The Changing Face of Observational Astronomy

- Large digital sky surveys are becoming the dominant source of data in astronomy: currently > 100 TB, and growing rapidly
  - Current examples:  SDSS, 2MASS, DPOSS, GSC, FIRST, NVSS, RASS, IRAS; CMBR experiments; Microlensing experiments; NEAT, LONEOS, and other searches for Solar system objects …
  - Digital libraries: ADS, astro-ph, NED, Simbad, NSSDC …
  - Observatory archives: HST, CXO, space and ground-based …
  - Future: QUEST2, LSST, and other synoptic surveys; GALEX, SIRTF, astrometric missions, GW detectors …
- Data sets orders of magnitude larger, more complex, and more homogeneous than in the past
- Roughly 1+ TB/Sky/band/epoch
  - NB: Human Genome is ~ 1 TB, Library of Congress ~ 20 TB

# This Quantitative Change in the Amount of Available Information Will Enable the **Science of a Qualitatively Different Nature:**



- **Statistical astronomy done right**
  - Precision cosmology, Galactic structure, stellar astrophysics …
  - Discovery of significant patterns and multivariate correlations
  - Poissonian errors unimportant
- **Systematic Exploration of the Observable Parameter Spaces**
  - Searches for rare and unknown types of objects and phenomena
  - Low surface brightness universe, the time domain …
- Confronting massive numerical simulations with massive data sets



Understanding of Complex Astrophysical Phenomena Requires Complex and Information-Rich Data Sets, and the Tools to Explore them …

… This Will Lead to a Change in the Nature of the Astronomical Discovery Process …

… Which Requires A New Research Environment for Astronomy:
**A Virtual Observatory**

# The Changing Style of Observational Astronomy

| **The Old Way:** | **Now:** | **Future:** |
|---|---|---|
| Pointed, heterogeneous observations (~ MB - GB) | Large, homogeneous sky surveys (multi-TB, ~ $10^6$ - $10^9$ sources) | Multiple, federated sky surveys and archives (~ PB)<br>↓ |
| Small samples of objects (~ $10^1$ - $10^3$) | Archives of pointed observations (~ TB) | **Virtual Observatory** |

# Technological Challenges for the VO:

1. **Data Handling**:
   - Efficient database architectures/query mechanisms
   - Archive interoperability, standards, metadata …
   - Survey federation (in the image and catalog domains)

   … etc.
2. **Data Analysis**:
   - Data mining / KDD tools and services (clustering analysis, anomaly and outlier searches, multivariate statistics…)
   - Visualization (image and catalog domains, high dimensionality parameter spaces)

   … etc.

**NB:** A typical (single survey) catalog may contain ~ $10^9$ data vectors in ~ $10^2$ dimensions ⇨ **Terascale** computing!

# Example: Today’s I/O rates

Reading a 1 TB data set:

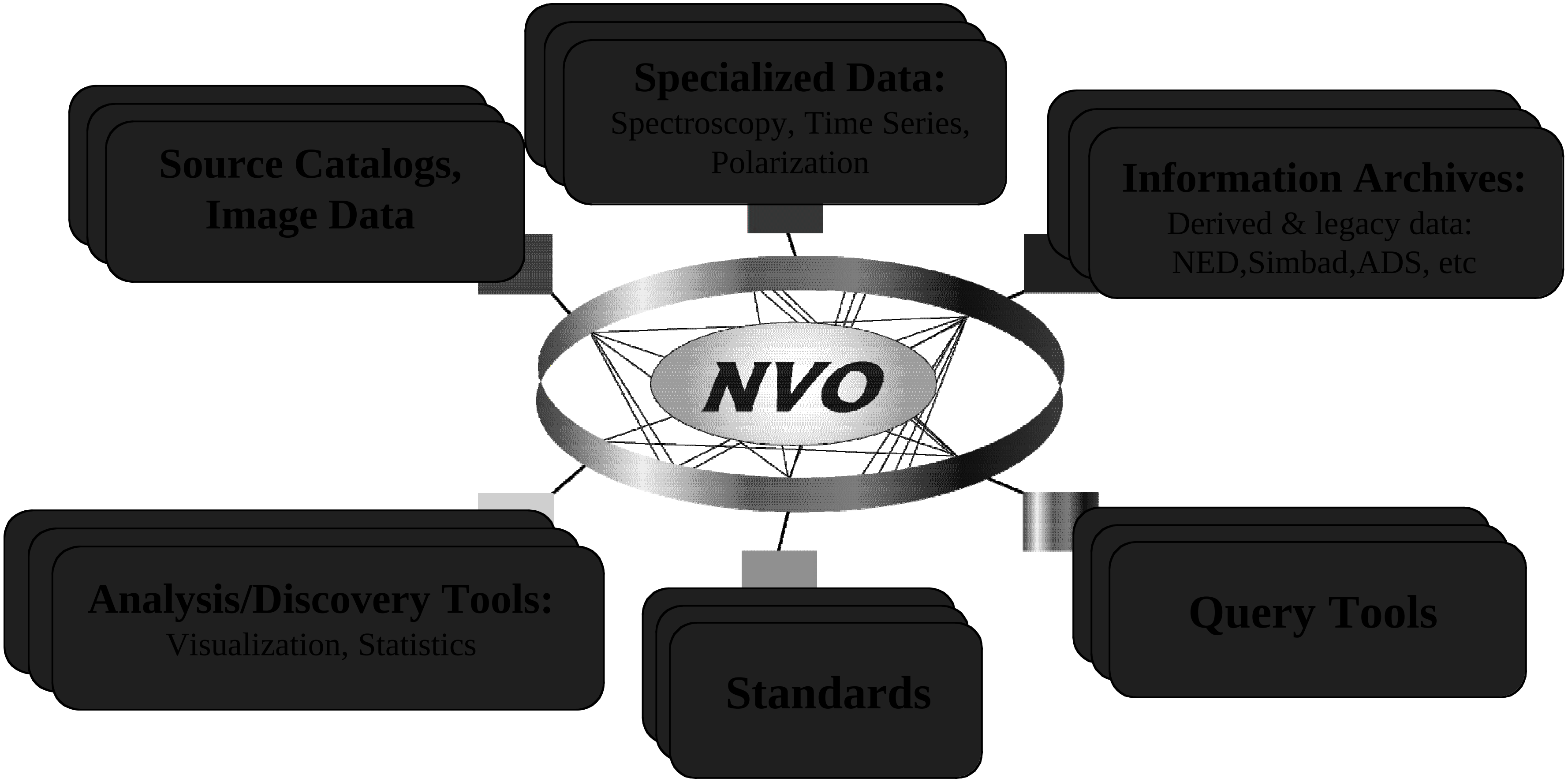
| *data access* | *speed* | *time [days]* |
|---|---|---|
| Fast database server | 50 MB/s | 0.23 |
| Local SCSI/Fast Ethernet | 10 MB/s | 1.2 |
| T1 | 0.5 MB/s | 23 |
| Typical ‘good’ www | 20 KB/s | 580 |

Most VO data would not be processed on a user’s desktop → Virtual Observatory environment

# The Virtual Observatory Concept

- A response of the astronomical community to the scientific and technological challenges posed by massive data sets
- A Virtual Observatory (VO) would federate the existing and forthcoming large digital sky surveys and archives, and provide the tools for their scientific exploitation
- A top recommendation of the NAS Decadal Survey, *Astronomy and Astrophysics in the New Millenium* (http://www.nap.edu/books/0309070317/html/) is the creation of the **National Virtual Observatory (NVO)**
- Combined with similar efforts in Europe, this will lead to a Global Virtual Observatory
- For details and the vision, see the **NVO White Paper,** http://www.arXiv.org/abs/astro-ph/0108115

# What is the NVO? - Content



**Specialized Data:**
Spectroscopy, Time Series,
Polarization

**Source Catalogs,**
**Image Data**

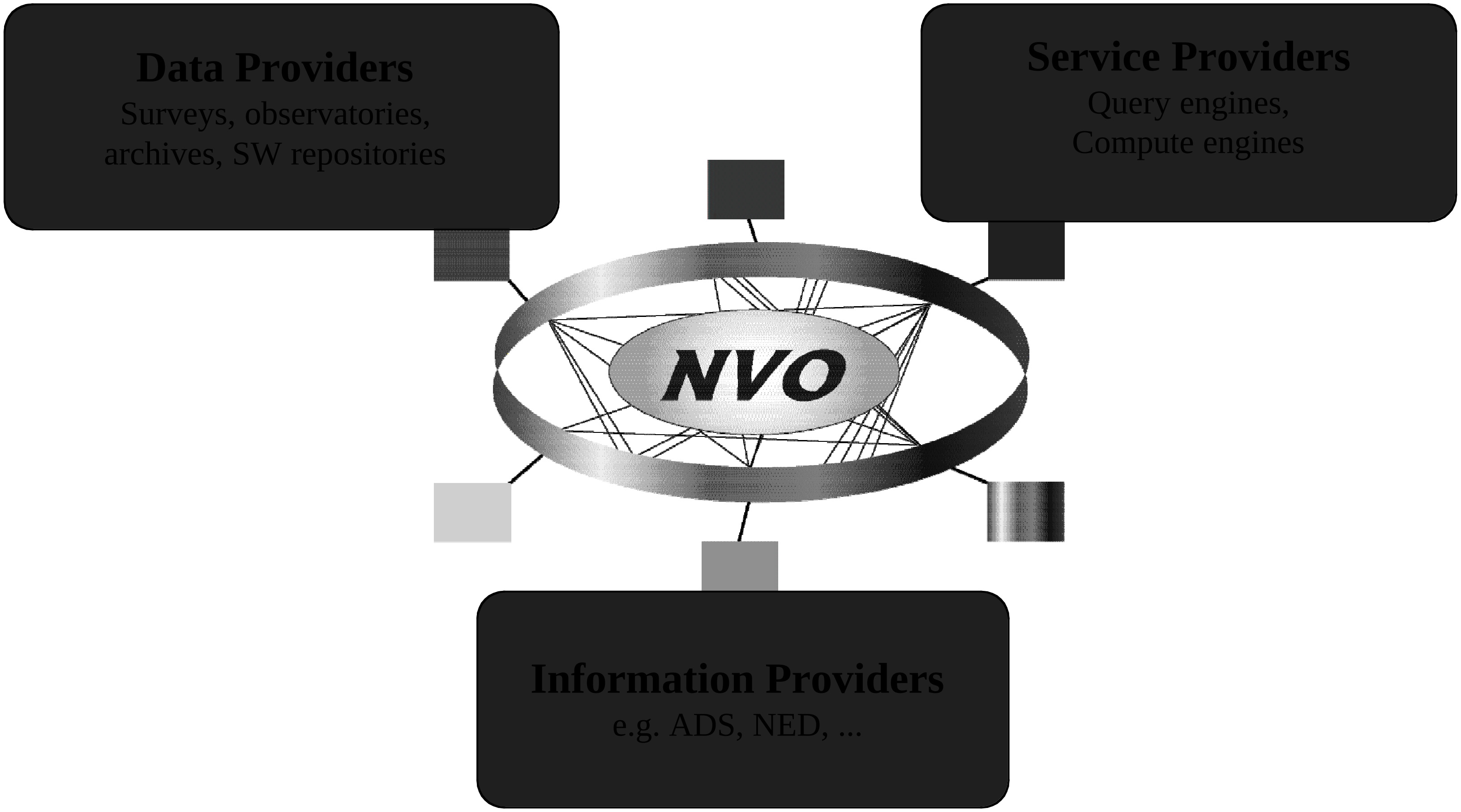
**Information Archives:**
Derived & legacy data:
NED,Simbad,ADS, etc

**Analysis/Discovery Tools:**
Visualization, Statistics

**Query Tools**

**Standards**

# What is the NVO? - Components



**Data Providers**

Surveys, observatories,
archives, SW repositories

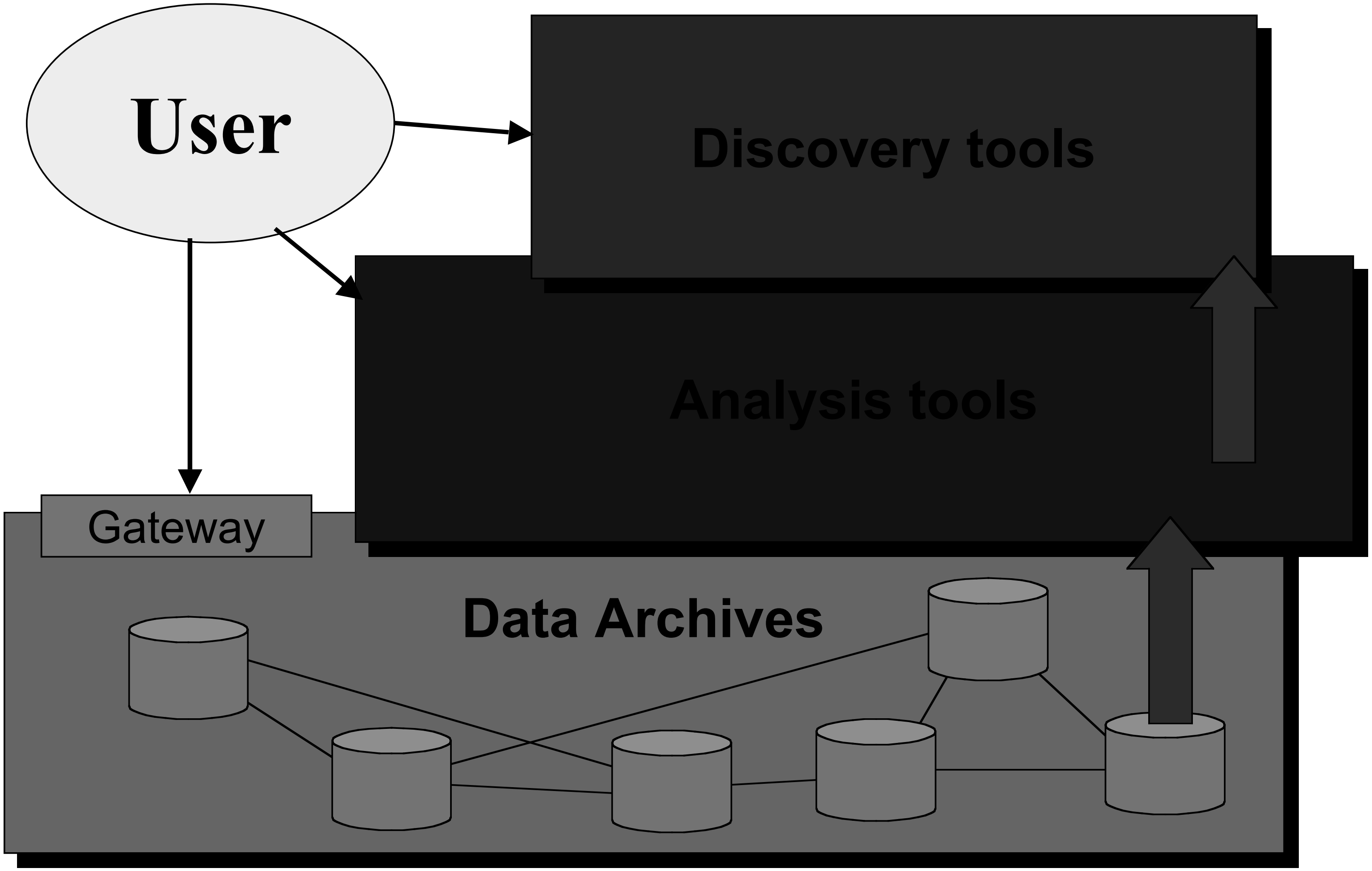
**Service Providers**

Query engines,
Compute engines

**Information Providers**

e.g. ADS, NED, ...

# Conceptual VO Architecture



# What We Have Now:

- Many separate archives, passively serving on demand small amounts of limited data
- Almost no data discovery capabilities
- No standards for metadata, data exchange protocols, formats
- No general tools or services for data fusion and analysis (data mining, KDD)

# What We Need:

- A dynamical, interactive, web-based environment for the new astronomy with massive data sets → **Virtual Observatory**

**VO Will Be Technology Enabled, But Science Driven**

# The Roles of a VO:

- **Facilitate science with massive data sets** (observations and theory/simulations)
- Provide **an added value from federated data sets** (e.g., multi-wavelength, multi-scale, multi-epoch …)
  - Historical examples: the discoveries of Quasars, ULIRGs, GRBs, radio or x-ray astronomy …
- **Enable and stimulate some *new* science** with massive data sets (not just old but bigger)
- **Optimize the use of expensive resources** (large ground-based telescopes, space missions)
  - Target selection from wide-field surveys
- **Science strategy and planning**: where are the gaps in our coverage of the observable parameter space?
- **Education and public outreach**

# Examples of Possible VO Projects:

- **A Panchromatic View of AGN and Their Evolution**
  - Cross-matching of surveys, radio to x-ray
  - Understanding of the selection effects
  - Obscuration, Type-2 AGN, a complete census

  → Evolution and net energetics, backgrounds
- **A Phase-Space Portrait of Our Galaxy**
  - Matching surveys: visible to NIR (stars), FIR to radio (ISM)
  - A 3-D picture of stars, gas, and dust, SFR …
  - Proper motions and gas velocities: a 6-D phase-space picture

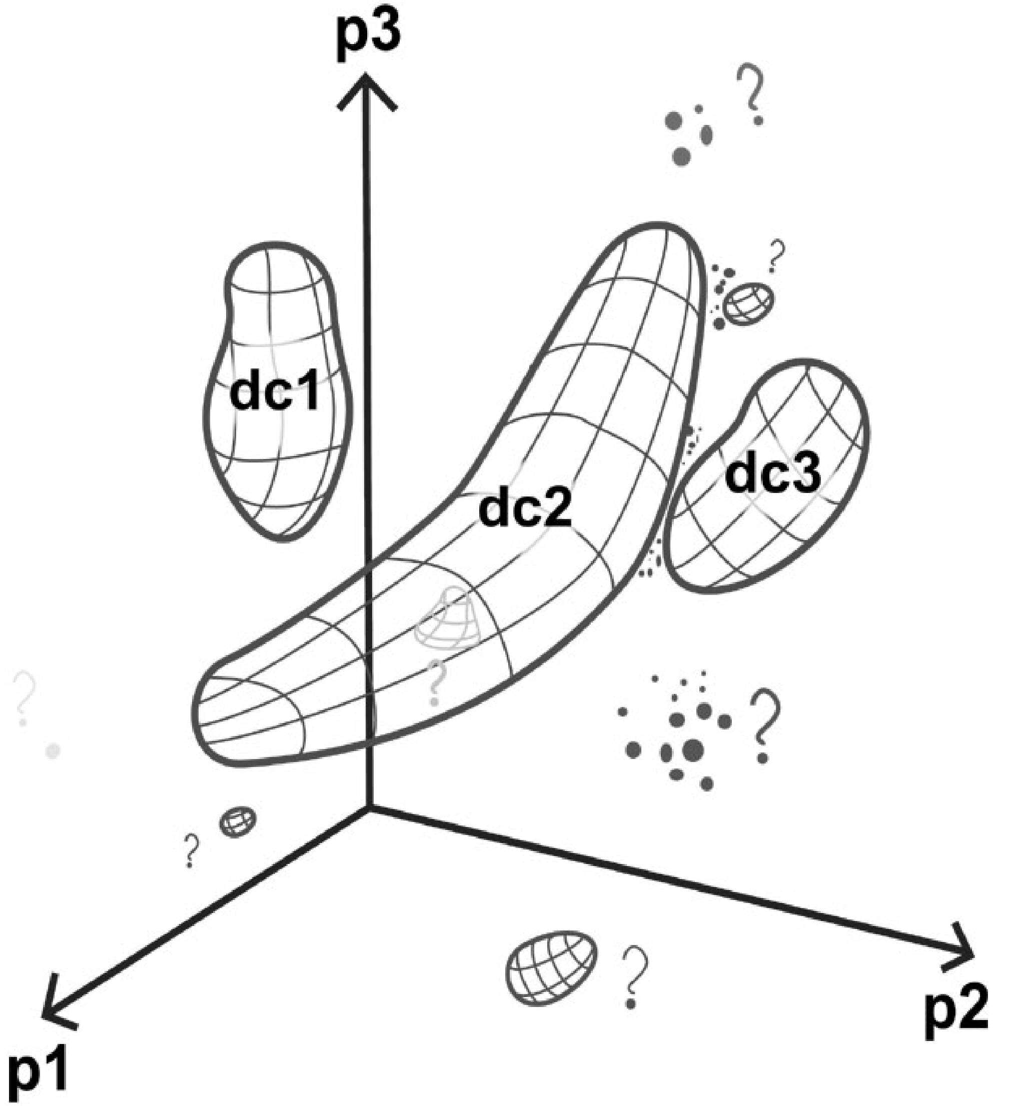
  → Structure, dynamics, and formation of the Galaxy
- **Galaxy Clusters as Probes of the LSS and its Evolution**
  - Cluster selection using a variety of methods: galaxy overdensity, x-rays, S-Z effect …
  - Understanding of the selection effects

  → Probing the evolution of the LSS, cosmology

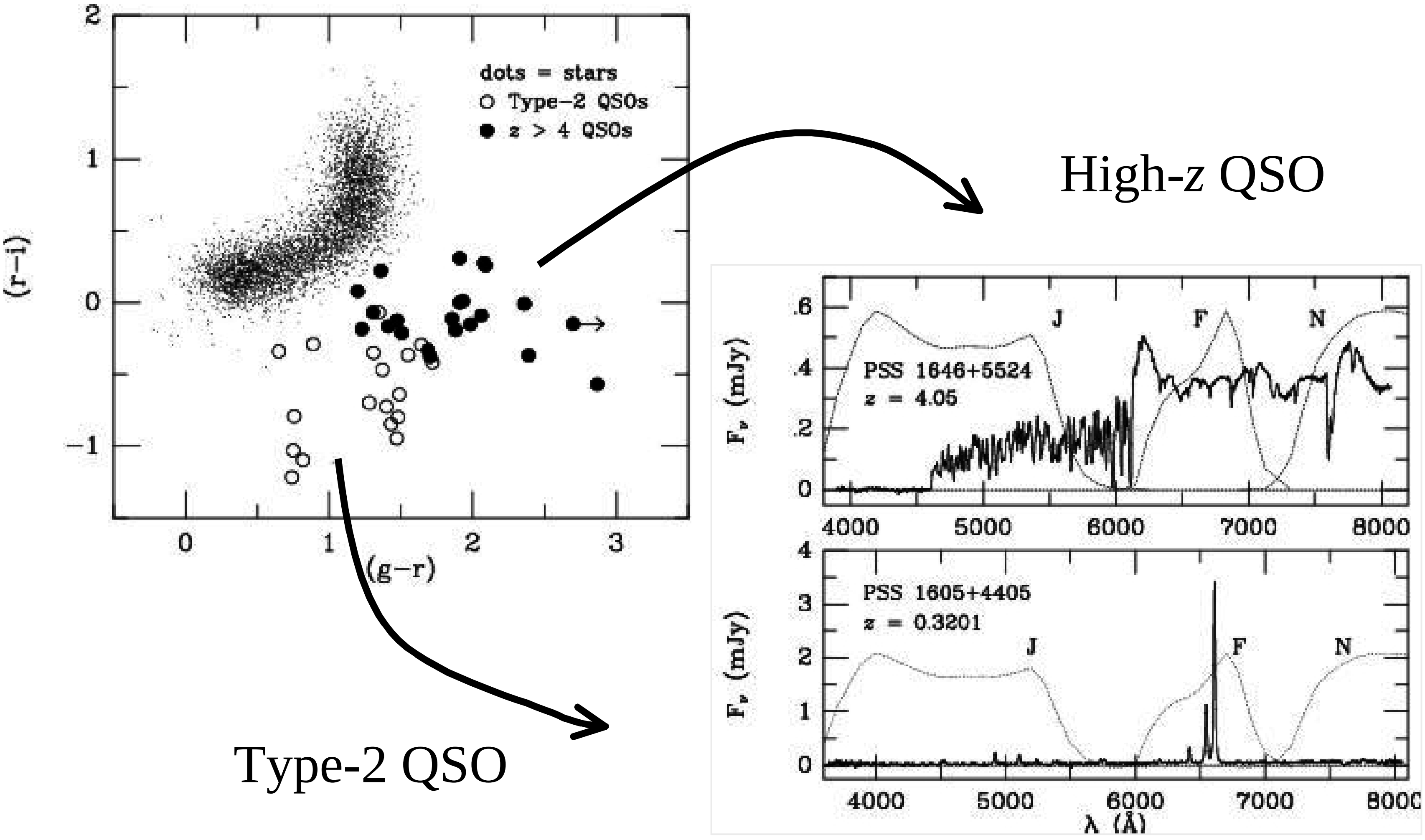
Exploration of parameter spaces of measured source attributes from federated sky surveys will be one of the principal techniques for the VO, e.g., in searches for rare or even new types of objects.

This will include supervised and unsupervised classification and clustering analysis techniques.

A Generic Machine-Assisted Discovery Problem: Data Mapping and a Search for Outliers



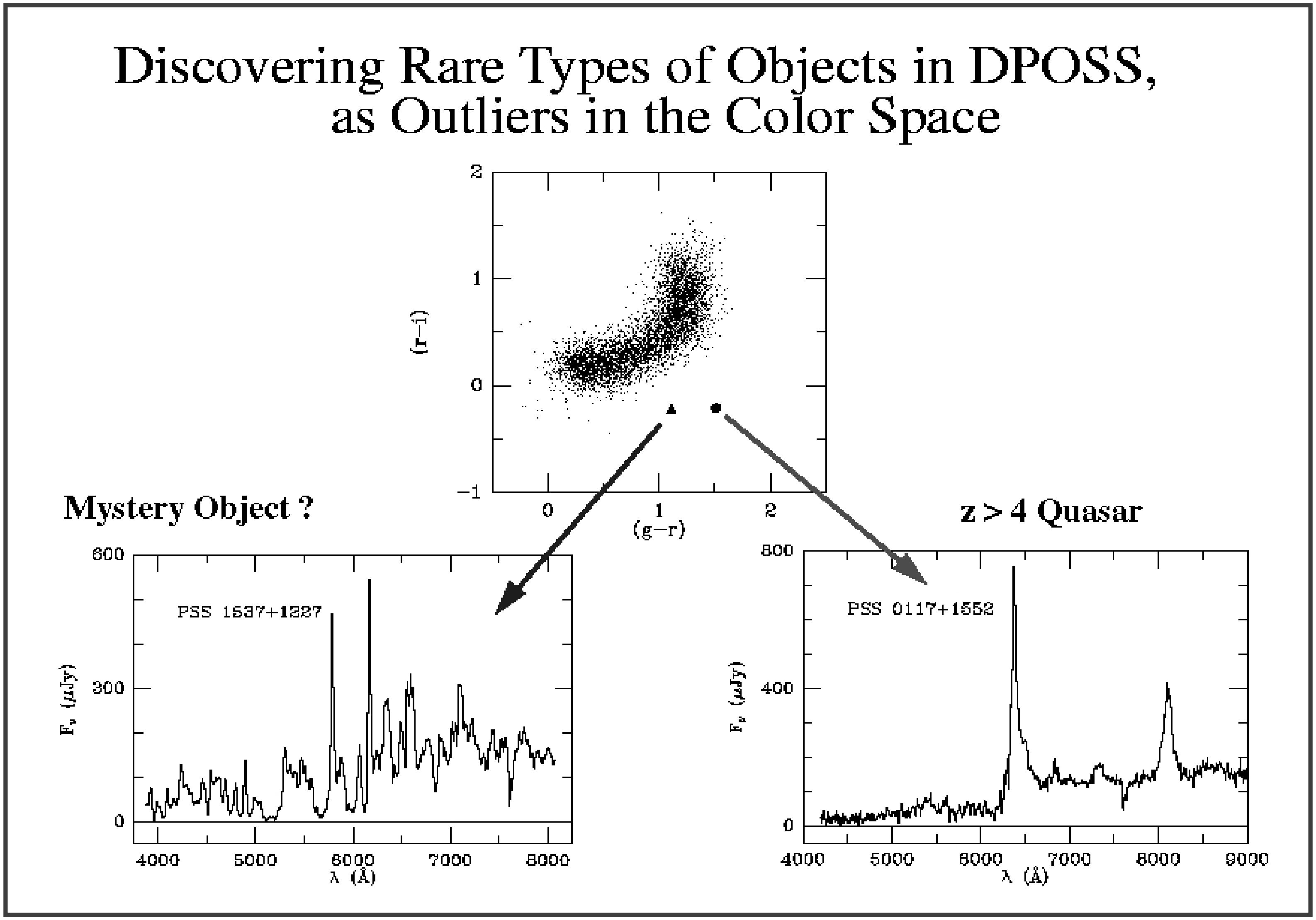
# An Example: Discoveries of High-Redshift Quasars and Type-2 Quasars in DPOSS



High-z QSO

Type-2 QSO

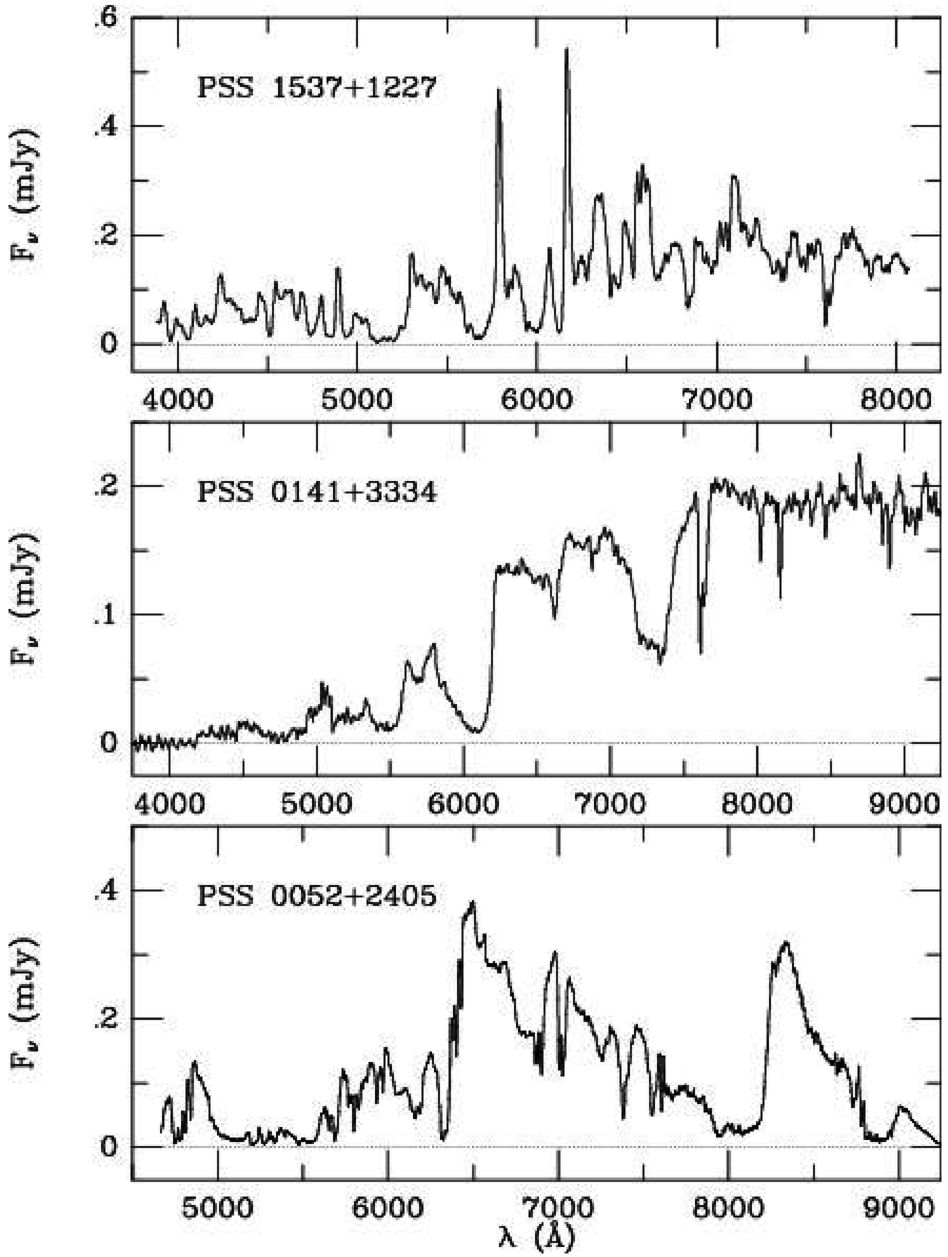
# But Sometimes You Find a Surprise…



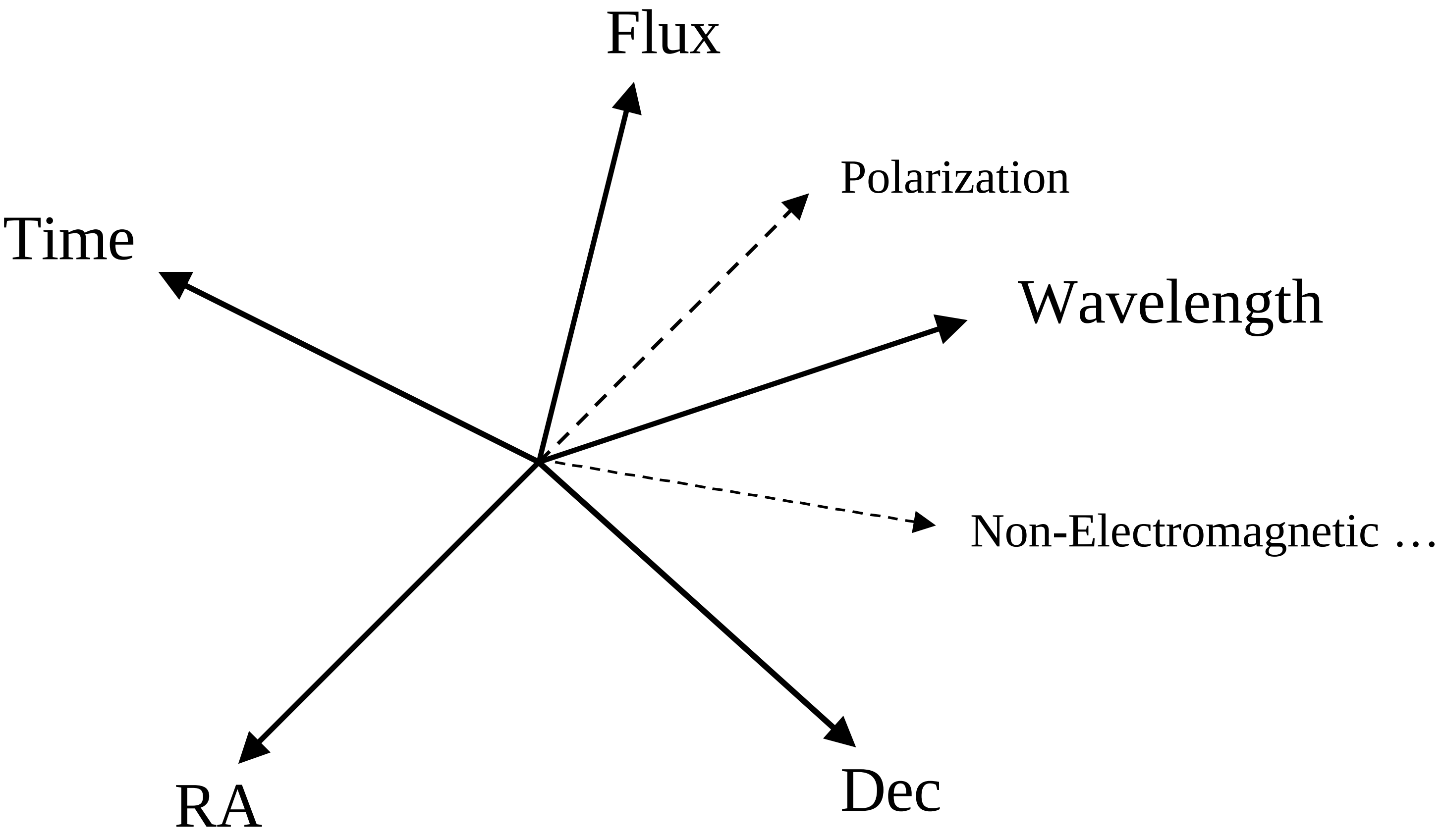
Discovering Rare Types of Objects in DPOSS, as Outliers in the Color Space

Spectra of Peculiar Lo-BAL (Fe) QSOs Discovered in DPOSS

(no longer a mystery, but a rare subspecies)



# Taking a Broader View: The Observable Parameter Space



Along each axis the measurements are characterized by the position, extent, sampling and resolution. All astronomical measurements span some volume in this parameter space. Some parts are better covered than others.

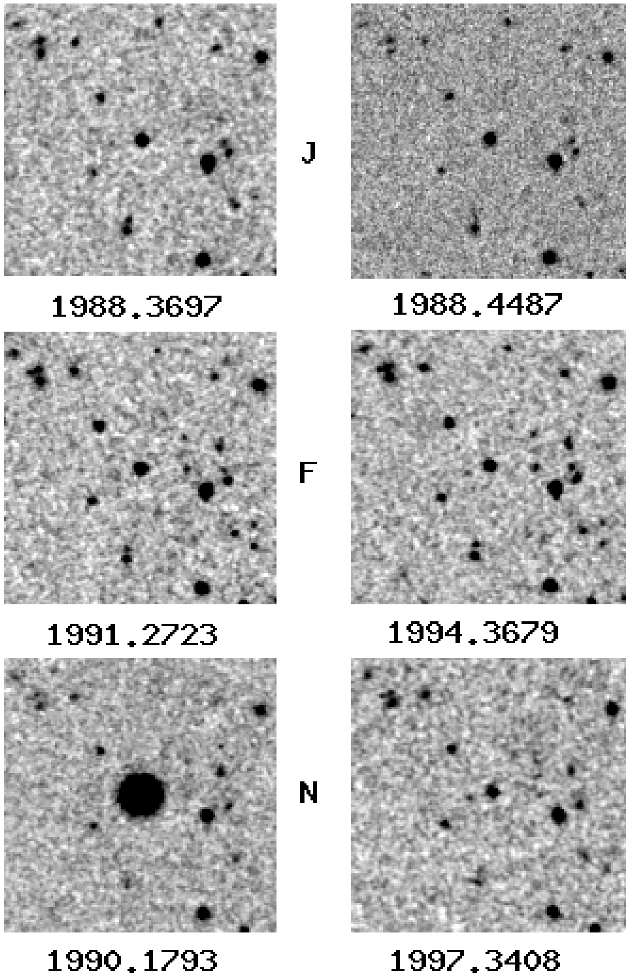
# Exploration of New Domains of the Observable Parameter Space

An example of a possible new type of a phenomenon, which can be discovered through a systematic exploration of the **Time Domain**:

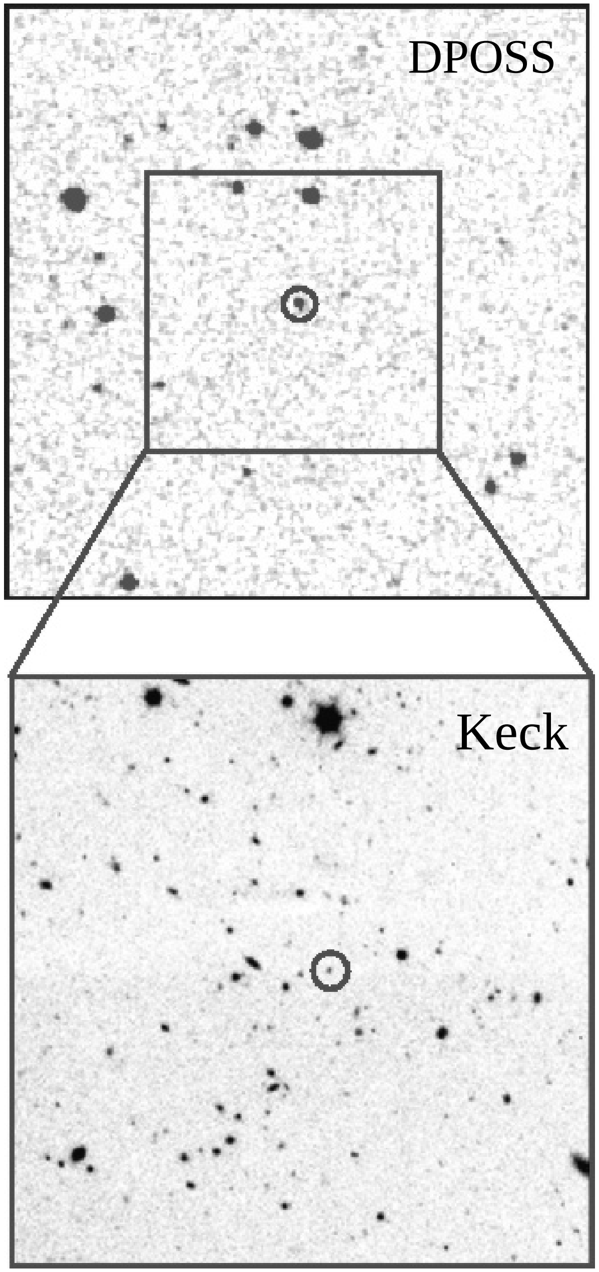
A normal, main-sequence star which underwent an outburst by a factor of > 300. There is some anecdotal evidence for such **megaflares** in normal stars.

The cause, duration, and frequency of these outbursts is currently **unknown**. Will our Sun do it?

A new generation of synoptic sky surveys may provide the answers -- and uncover other new kinds of objects or phenomena.



# Exploration of the Time Domain: Optical Transients

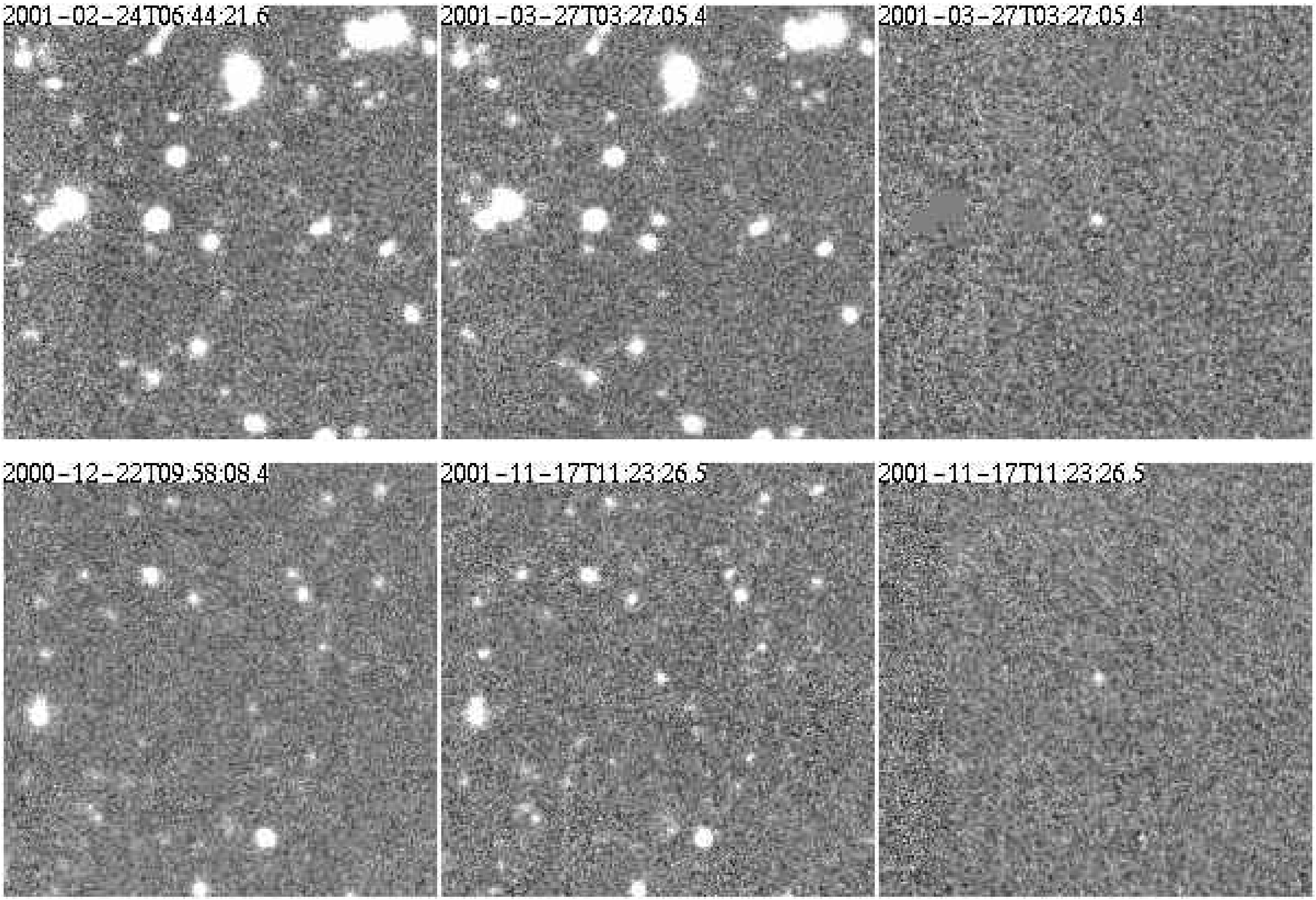


A Possible Example of an “Orphan Afterglow” (GRB?) discovered in DPOSS: an 18th mag transient associated with a 24.5 mag galaxy. At an estimated z ~ 1, the observed brightness is ~ 100 times that of a SN at the peak.

Or, is it something else, new?

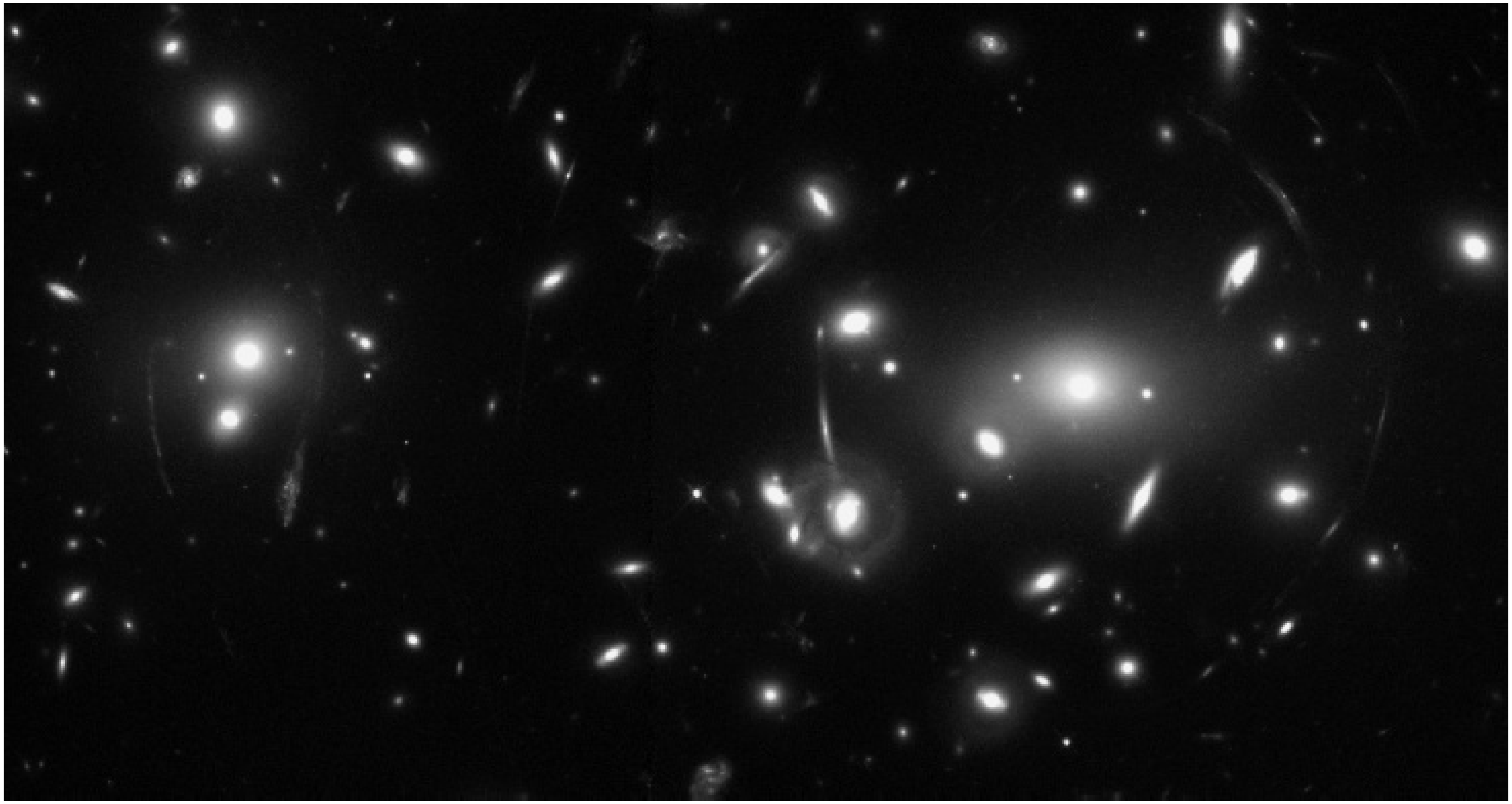
# Exploration of the Time Domain:

## Faint, Fast Transients (Tyson et al.)



**Data Mining in the Image Domain:** Can We Discover New Types of Phenomena Using Automated Pattern Recognition?

(Every object detection algorithm has its biases and limitations)



# The Ongoing VO Development Efforts

- The **NSF** ITR large project: Building the Framework for the NVO (plus a few smaller projects); also DTF
- **NASA** AISRP, HPCC, ADP programs -- addressing many technological aspects of the NVO
- Efforts at the NASA centers (IRSA, STScI, GSFC, CSC …)
- More modest efforts at the ground-based observatories (NOAO, NRAO, …)
- NSF Statistics FRG project, other small-scale efforts
- A good precursor project: NSF NPACI Digital Sky
- Many vigorous efforts in Europe (AVO, AstroGrid, CDS, AstroVirTel …), Canada (CADC), and elsewhere
- The NVO Science Definition Team: http://nvosdt.org (many links and relevant material)

# Astronomy and Other Fields

- Technical and methodological challenges facing the VO are **common to most data-intensive sciences** today, and beyond (commerce, industry, national security …)
- **Interdisciplinary exchanges** between different disciplines (e.g., astronomy, physics, biology, earth sciences …) are highly desirable
  - Avoid wasteful duplication of efforts and costs
  - Intellectual cross-fertilization
- **Partnerships and collaborations** with applied computer science and statistics are **essential**
- How is astronomy different?
  - An intermediate ground in information volume, heterogeneity, and complexity (cf. high-energy physics, genomics, finance …)

# Concluding Comments: Broad Issues

- We are not making the full use of the growing data abundance; we should!
- The old research methodologies, geared to deal with data sets many orders of magnitude smaller and simpler are no longer adequate: we have to learn some new tricks
- The necessary technology and know-how are available
- The key issues are **methodological:** we have to learn to ask **new kinds of questions,** enabled by the massive data sets and technology
- A key issue is the education of **the next generation of leaders** of science and technology, now students and postdocs
- (N)VO in particular offers **great possibilities for education** and **public outreach**, as it connects an appealing physical science with the computer science and information technology

# Concluding Comments: The Information-Rich Astronomy for the 21st Century

- Technological revolutions as the drivers/enablers of the bursts of scientific growth
- Historical examples in astronomy:
  - 1960’s: the advent of electronics and rocketry

    *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?***