Hubble Heritage

PRC99-12 • Space Telescope Science Institute • Hubble Heritage Team (AURA/STScI/NASA)

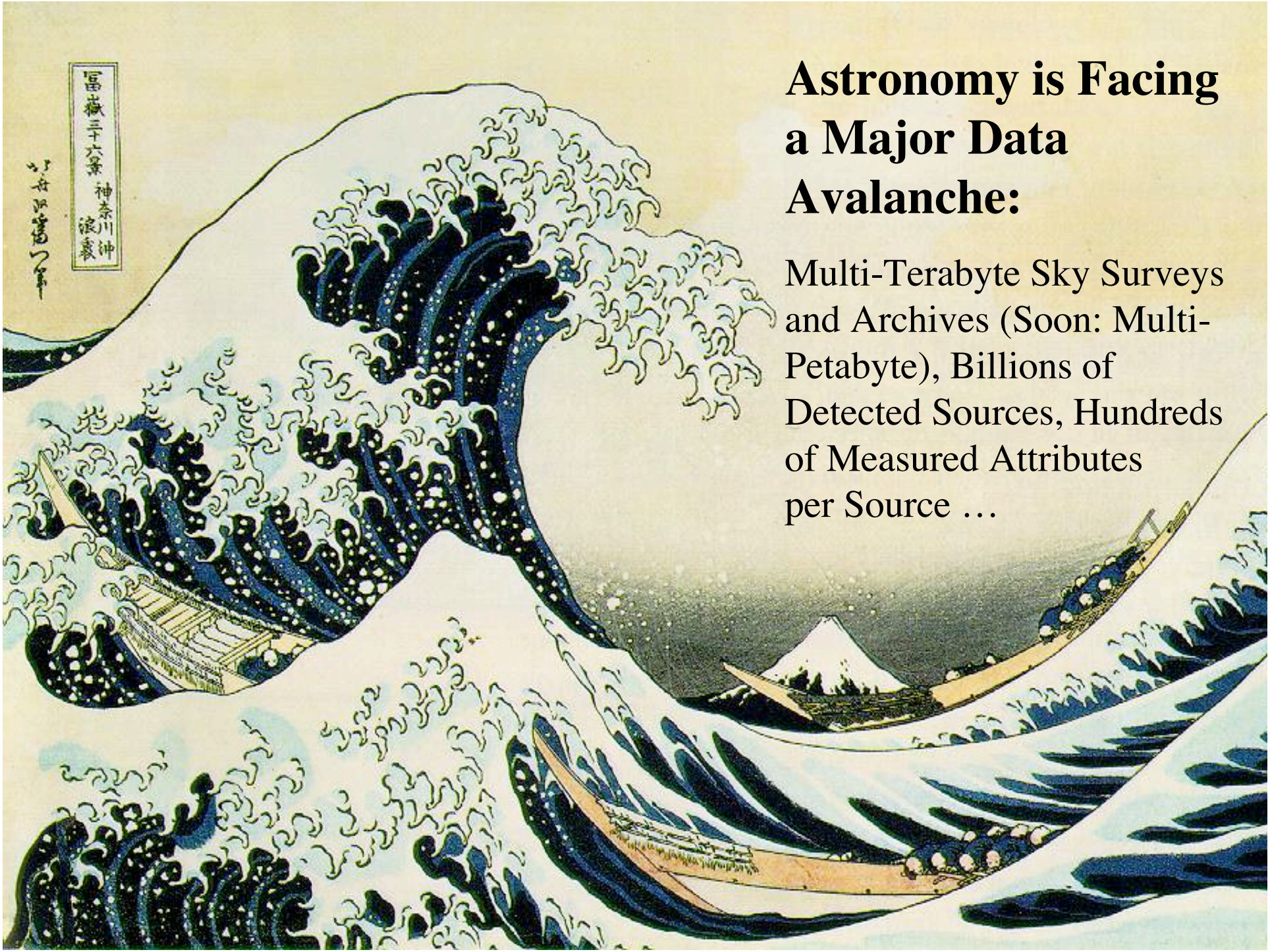
Data Mining Challenges and Opportunities in Astronomy

S. G. Djorgovski (Caltech)

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

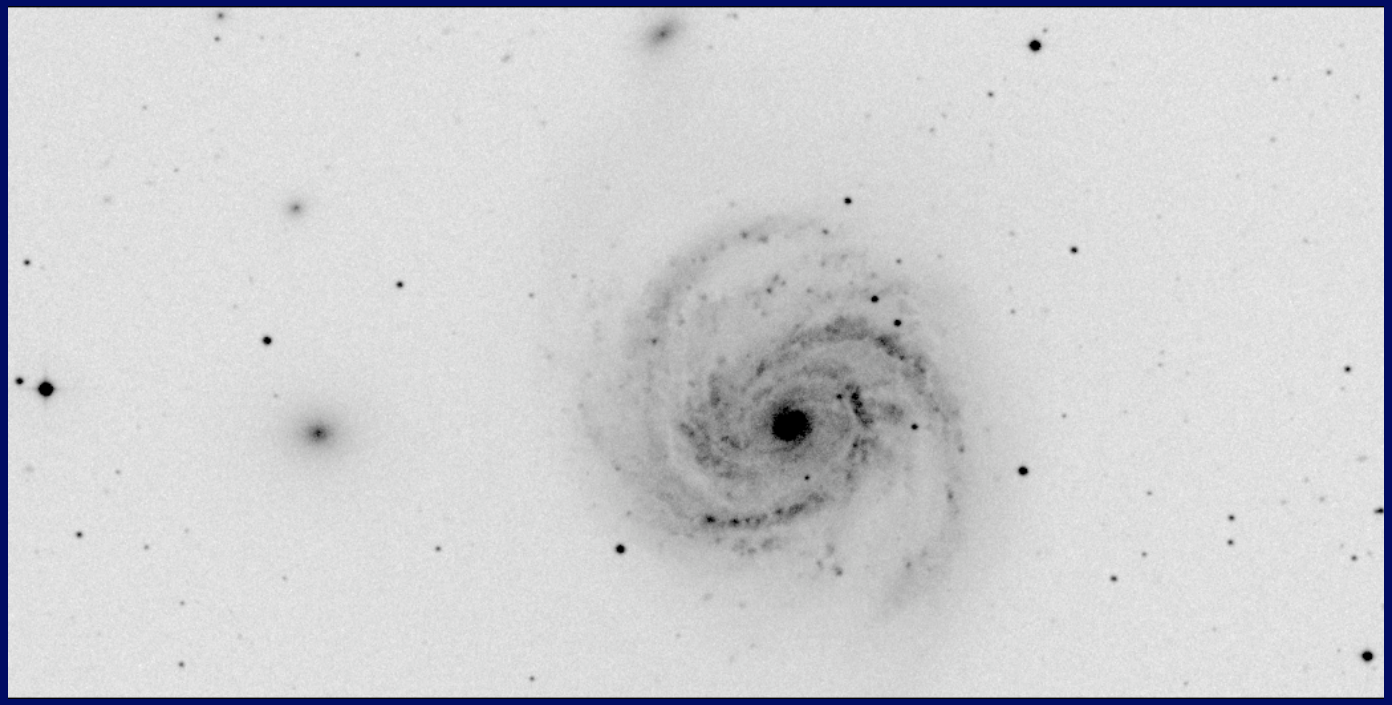
The Punchline:

- Astronomy has become an immensely data-rich field (and growing)
- There is a need for powerful DM/KDD tools
- There are excellent opportunities for inter-disciplinary collaborations/partnerships

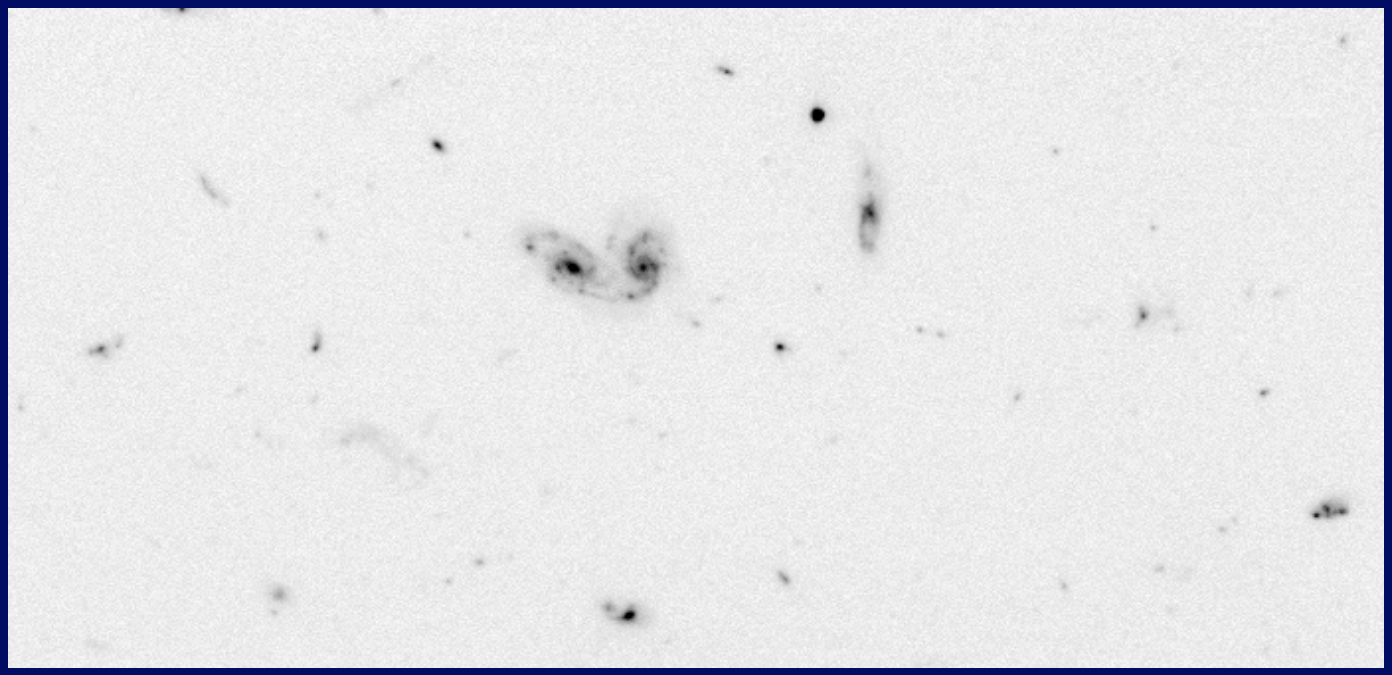


Astronomy is Facing a Major Data Avalanche:

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

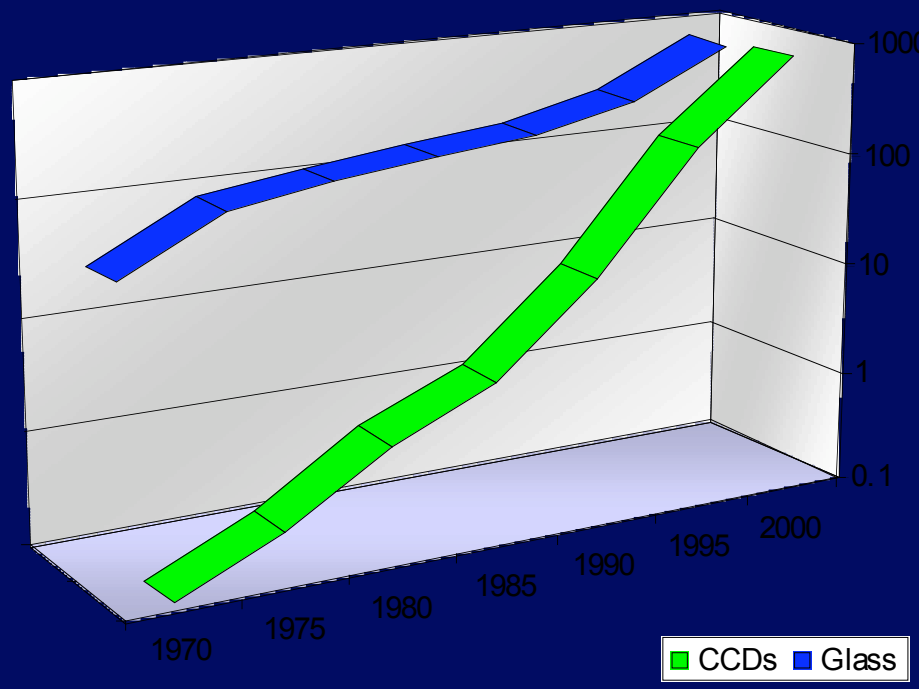


1 microSky
(DPOSS)



1 nanoSky
(HDF-S)

The Exponential Growth of Information in Astronomy



- 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

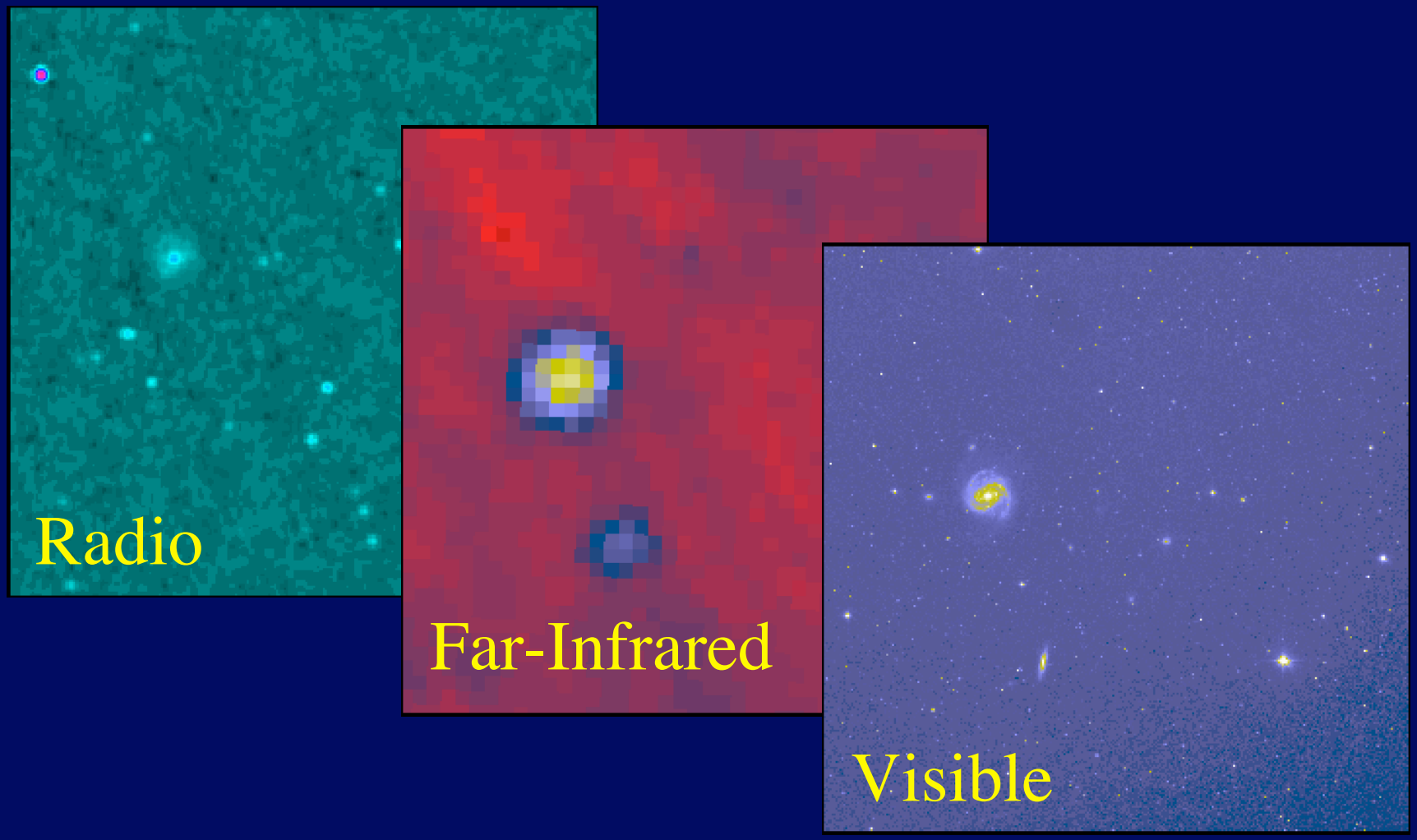
Total area of 3m+ telescopes in the world in m^2, total number of CCD pixels in Megapix, as a function of time. Growth over 25 years is a factor of 30 in glass, 3000 in pixels.

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 in major archives, and growing rapidly
- Typical sky survey today: ~ 10 TB of image data, ~ 10^9 detected sources, ~ 10^2 measured attributes per source
- Spanning the full range of wavelengths, radio through x-ray: a panchromatic, less biased view of the universe
- 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

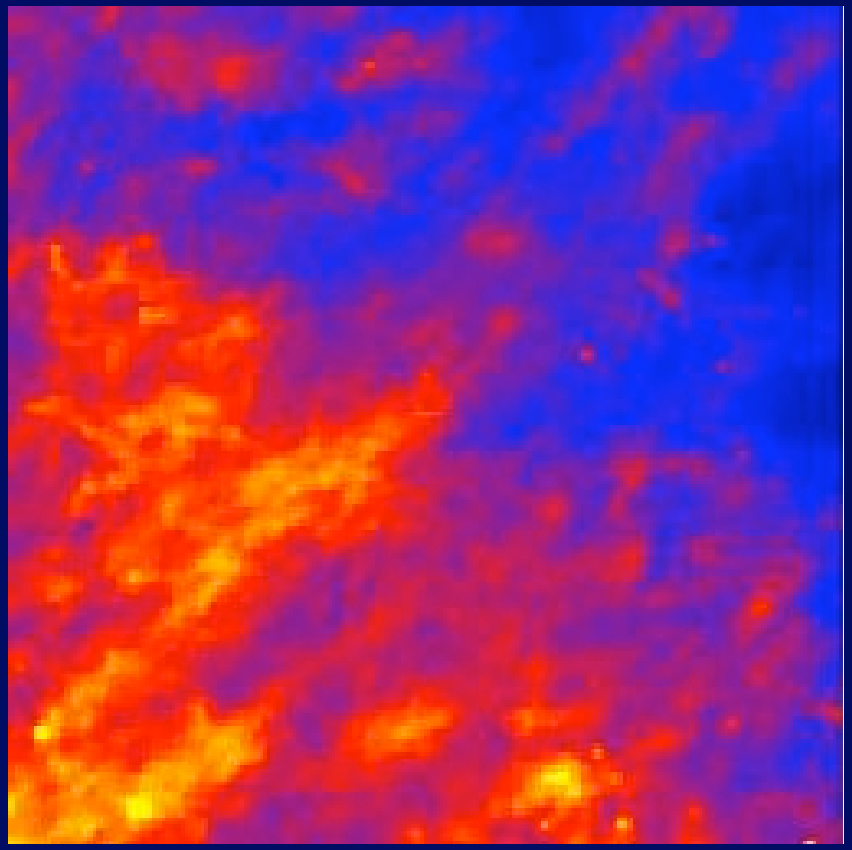
Panchromatic Views of the Universe



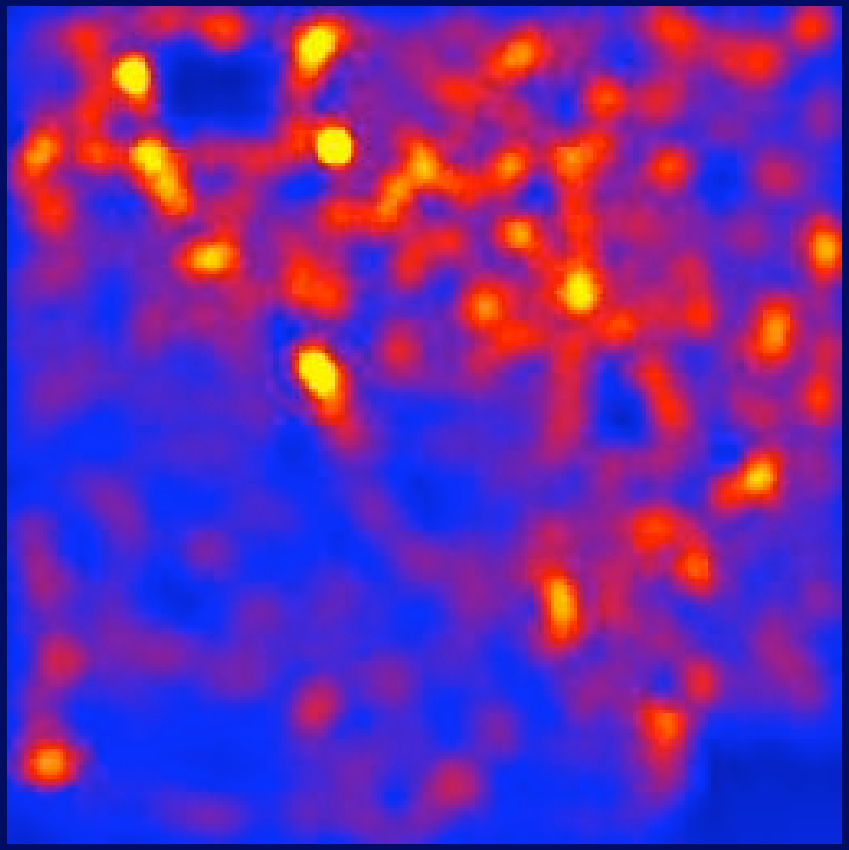
The Changing Style of Observational Astronomy

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)
Small samples of objects (~ 10^1 - 10^3)	Archives of pointed observations (~ TB)	**Virtual Observatory**

Multi-wavelength data paint a more complete (and a more complex!) picture of the universe



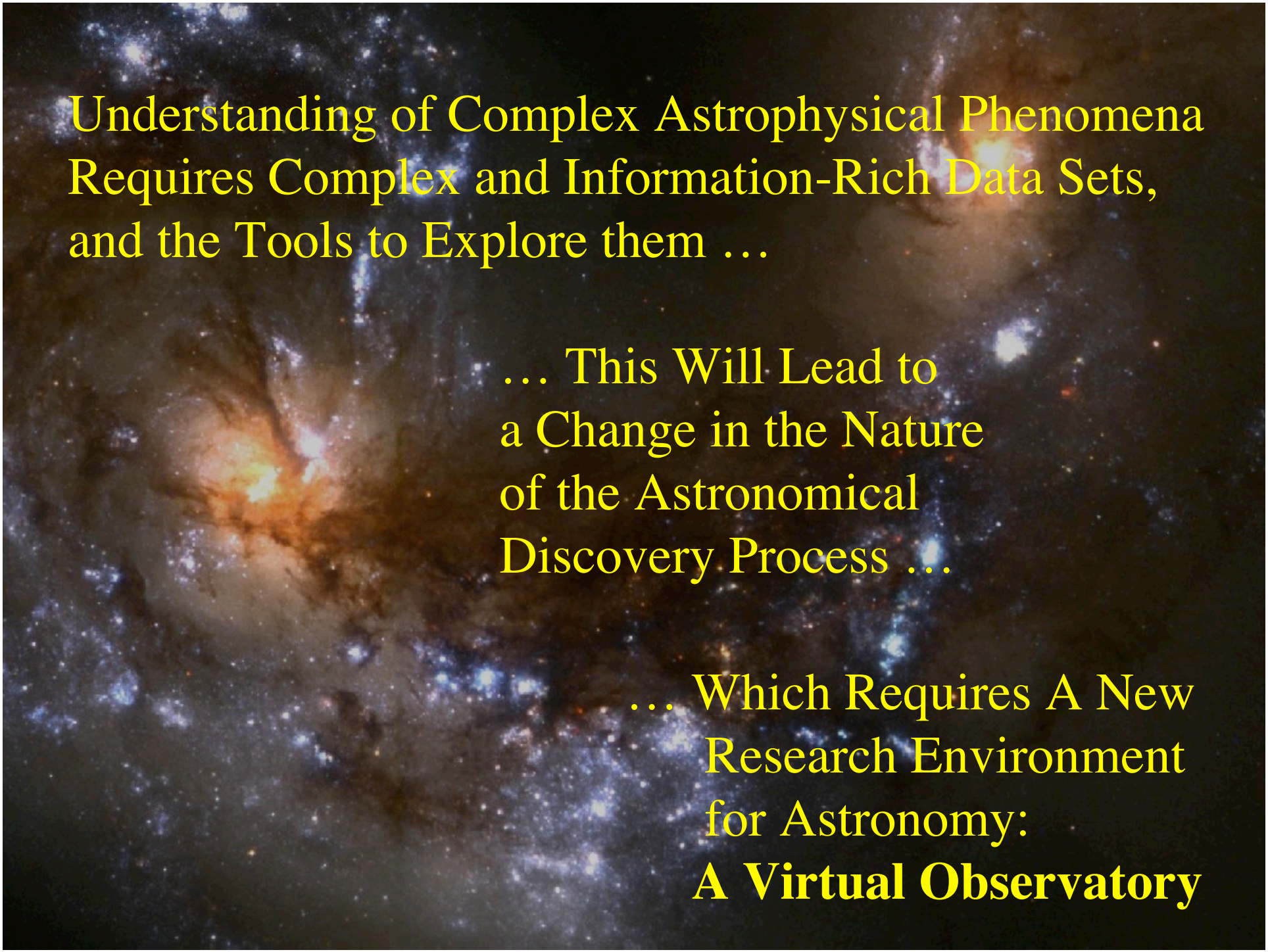
Infrared emission from interstellar dust



Smoothed galaxy density map

A panchromatic approach to the universe reveals a more complete physical picture

The resulting complexity of data translates into increased demands for data analysis, visualization, and understanding


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

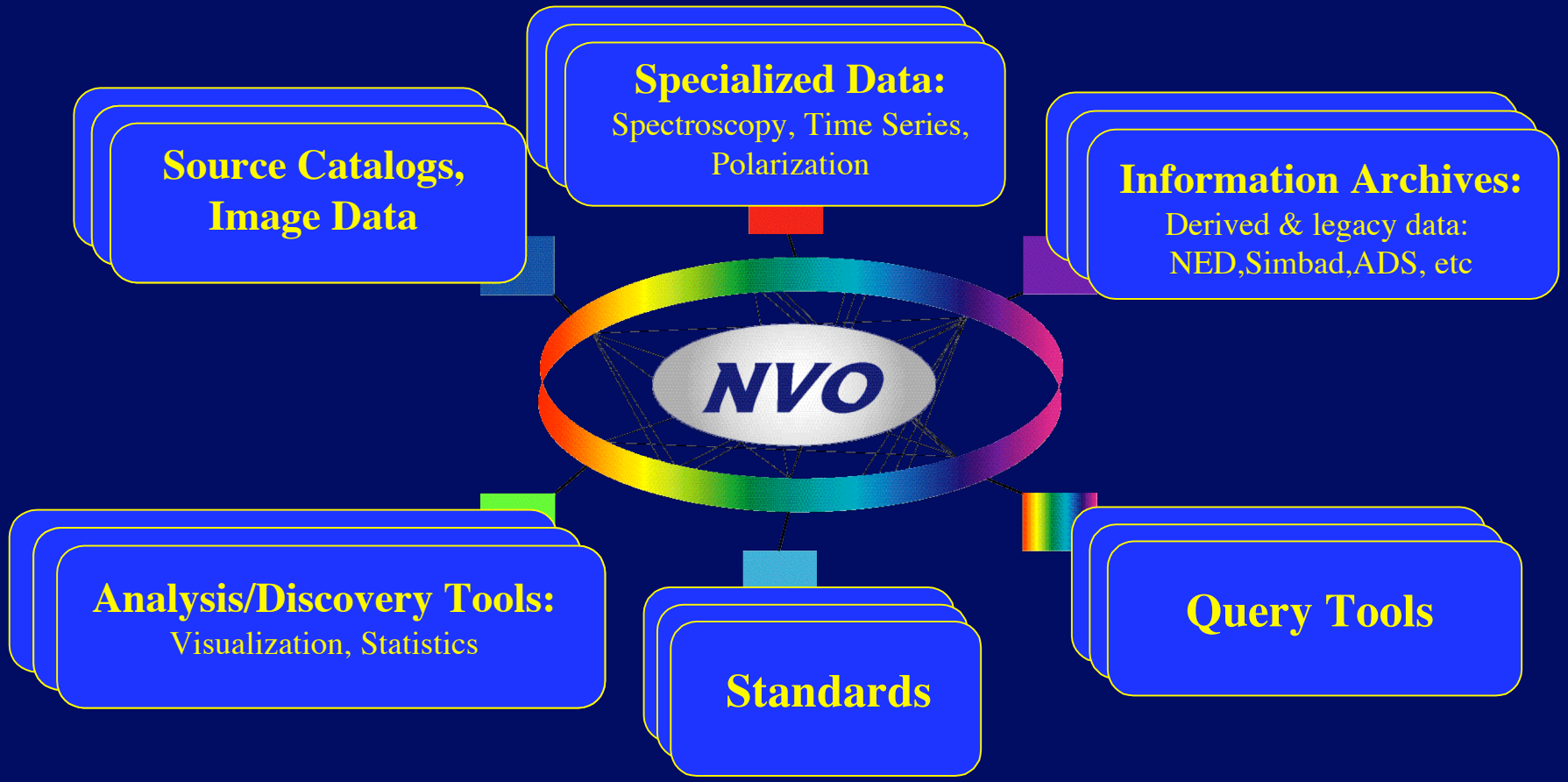
The Virtual Observatory (VO) Concept

- A response of the astronomical community to the scientific and technological challenges posed by massive data sets
- Federate the existing and forthcoming large digital sky surveys and archives, and provide the tools for their scientific exploitation
- A dynamical, interactive, web-based research environment for the new astronomy with massive data sets
- Technology-enabled, but science-driven

The Virtual Observatory Development

- A top recommendation of the NAS Decadal Survey, *Astronomy and Astrophysics in the New Millenium* is the creation of the **National Virtual Observatory (NVO)**
- Vigorous conceptual and technological design developments are under way
- Combined with similar efforts in Europe, this will lead to a Global Virtual Observatory
- For details and links, see http://www.astro.caltech.edu/~george/vo/

What is the NVO? - Content



What is the NVO? - Components

Data Providers
Surveys, observatories,
archives, SW repositories

Service Providers
Query engines,
Compute engines

Information Providers
e.g. ADS, NED, ...

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)

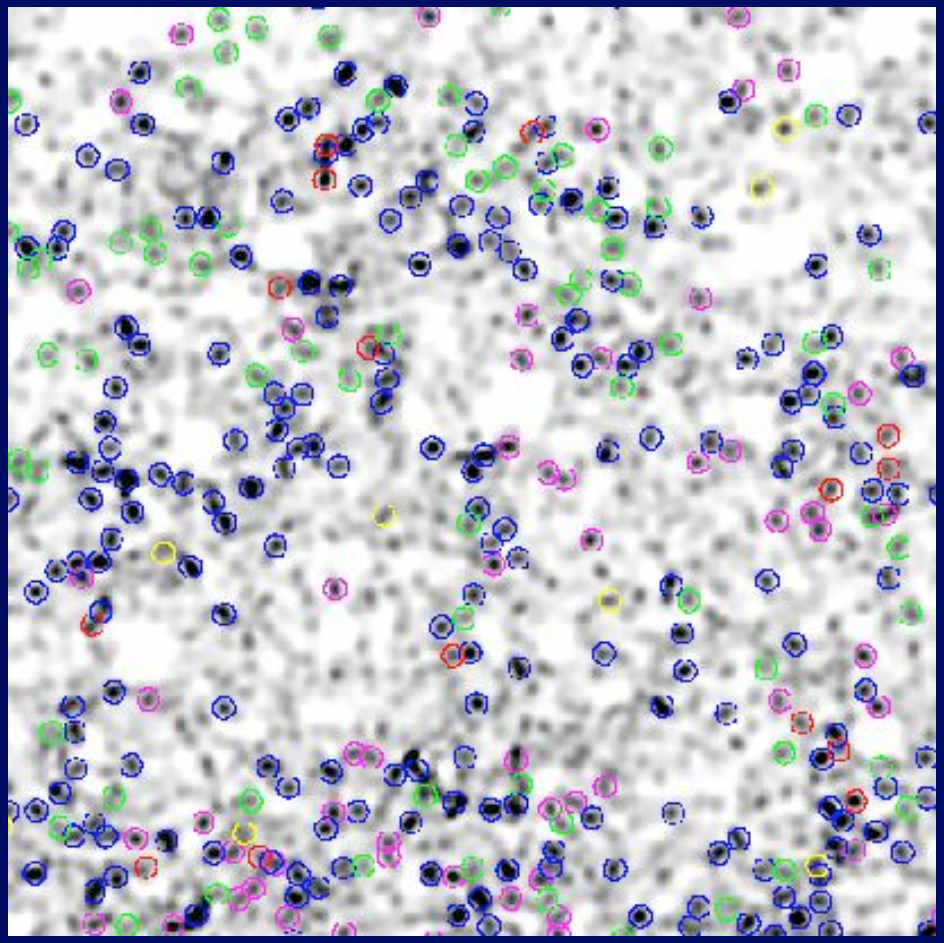
   … etc.

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

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**

A "classical" clustering problem in astronomy: galaxy clusters and clustering of galaxies in space



A problem:

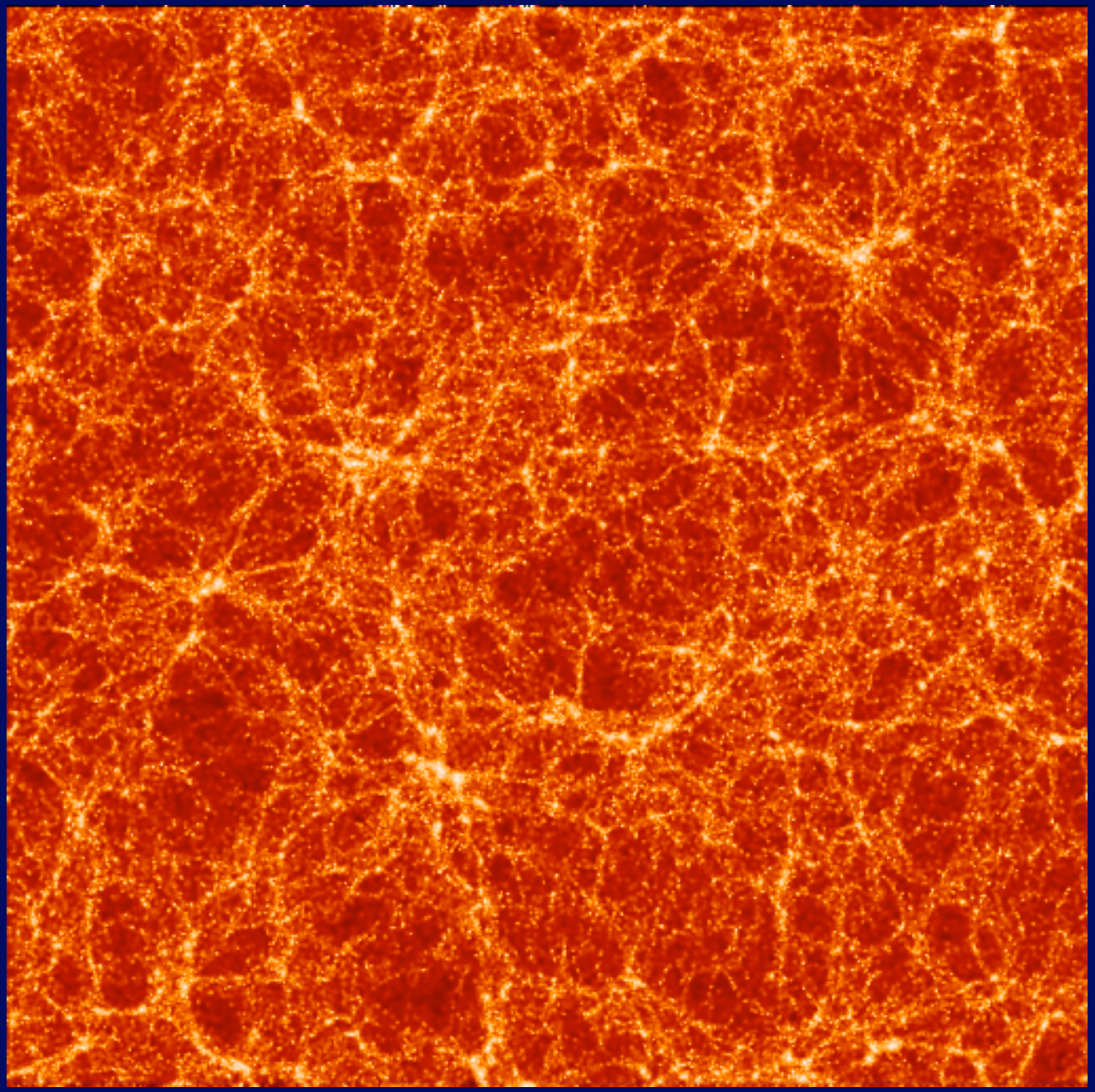
Physical clusters (virialized systems) are superposed on a *clustered background*

→ Clustering in the presence of a non-Gaussian, ~ 1/f noise …

A Problem:

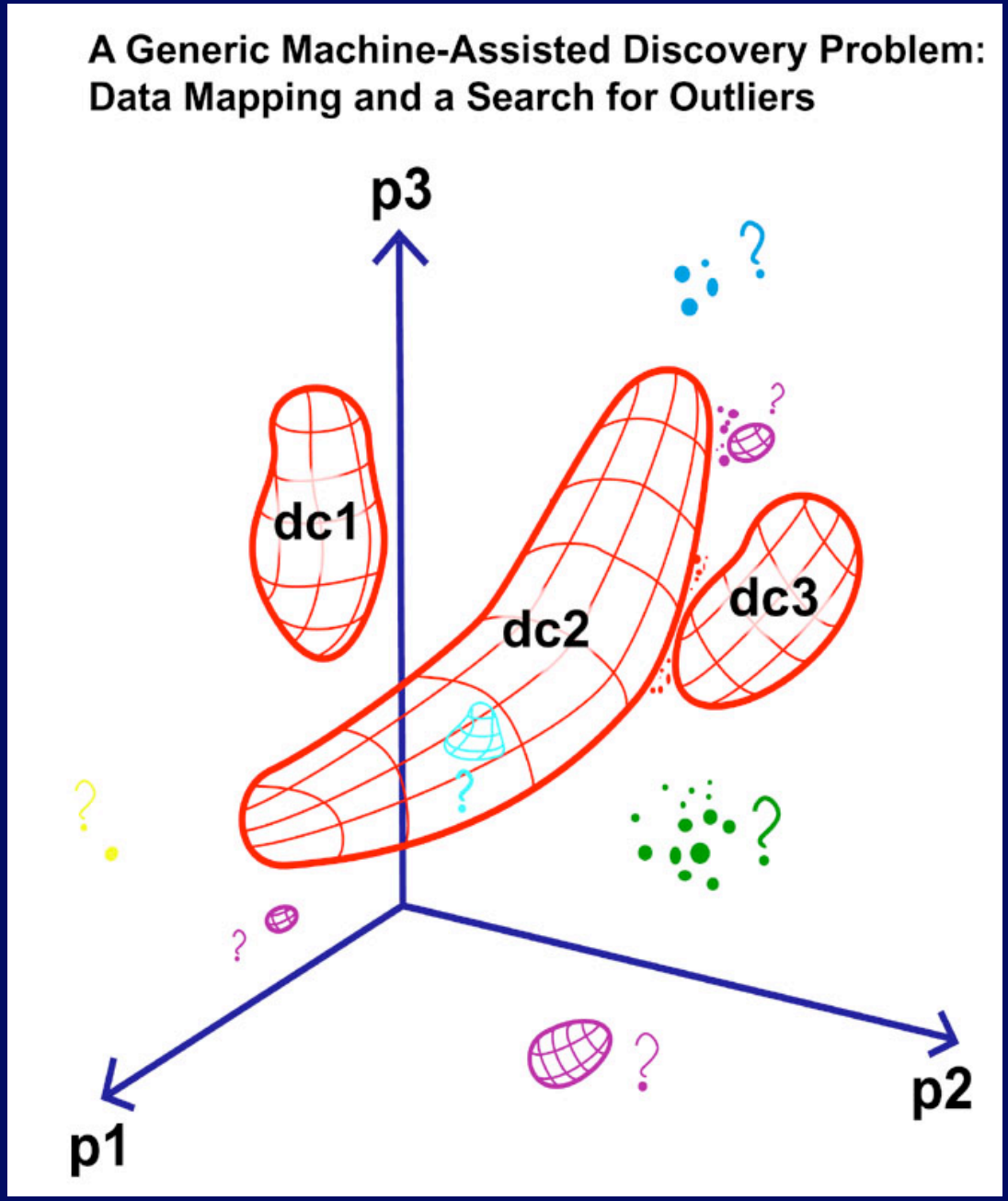
Non-trivial *topology* of clustering

(e.g., large-scale structure, but probably also in some other, parameter space applications)

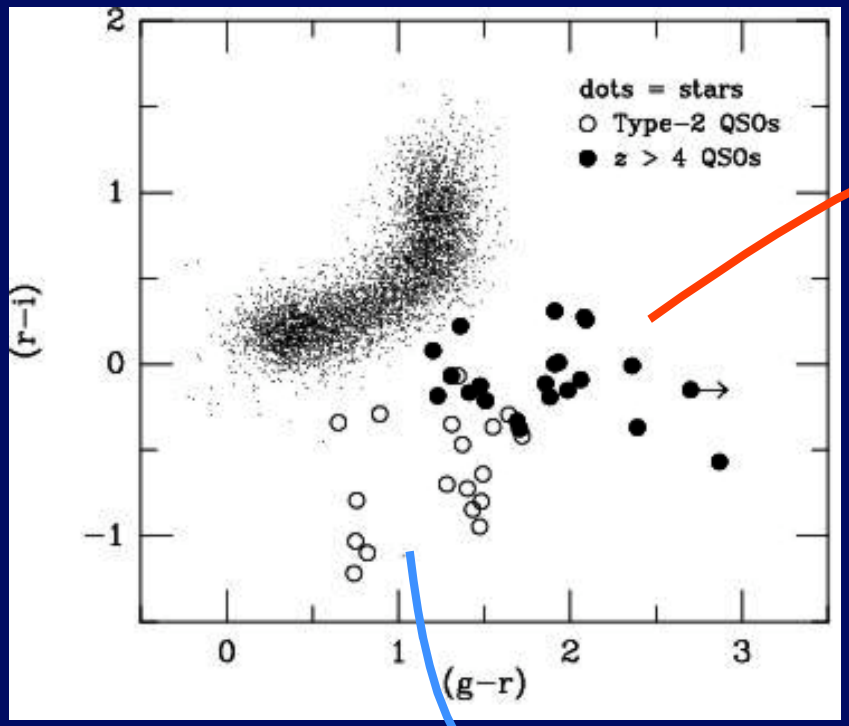


Exploration of parameter spaces of measured source attributes from federated sky surveys will be one of the principal techniques for the VO, e.g., in searches for rare or even new types of objects.

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

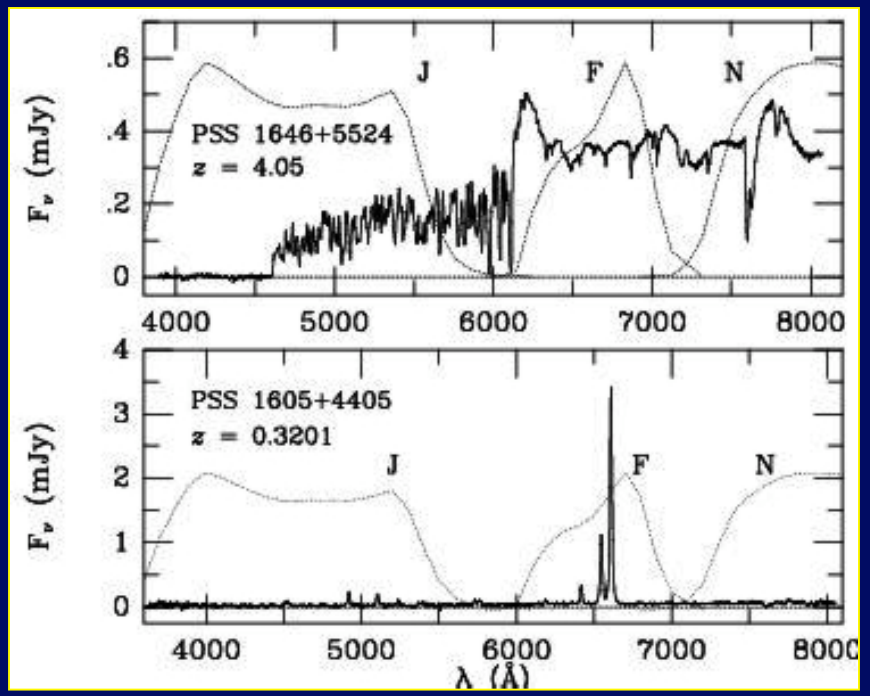


An Example: Discoveries of High-Redshift Quasars and Type-2 Quasars in DPOSS

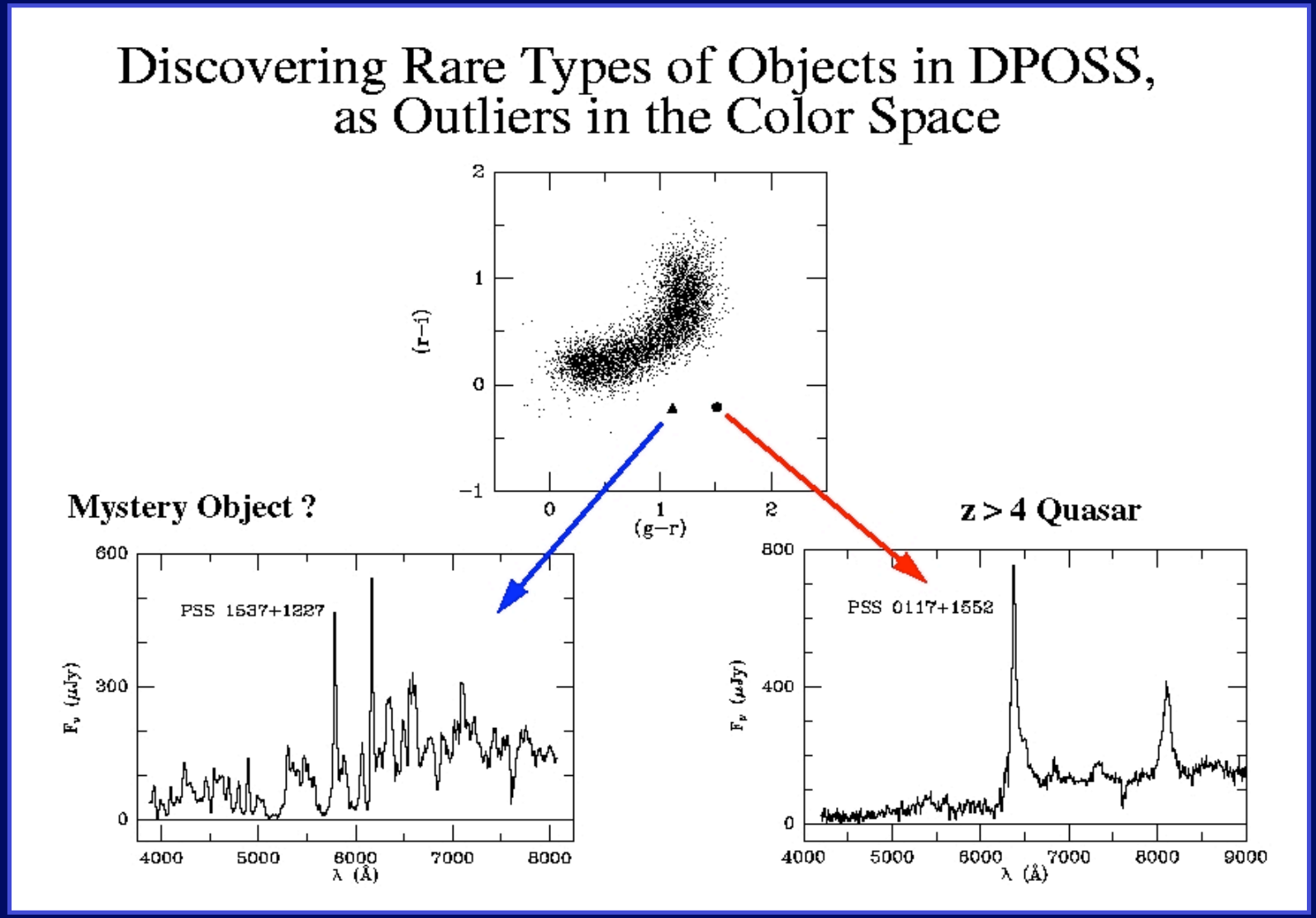


High-z QSO

Type-2 QSO

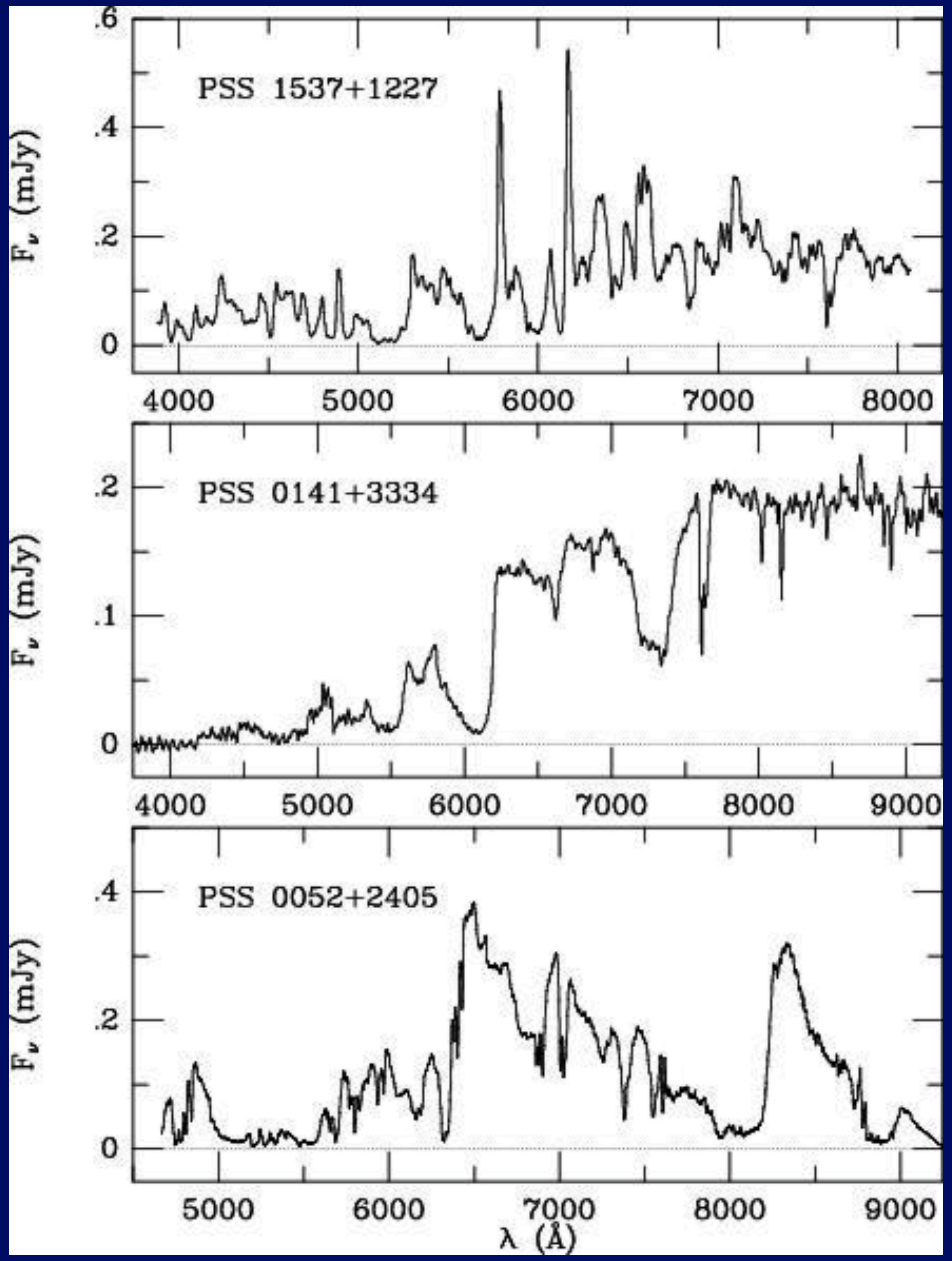


But Sometimes You Find a Surprise…



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

(no longer a mystery, but a rare subspecies)



Clustering Analysis of Astronomical Data Sets: Some Problems

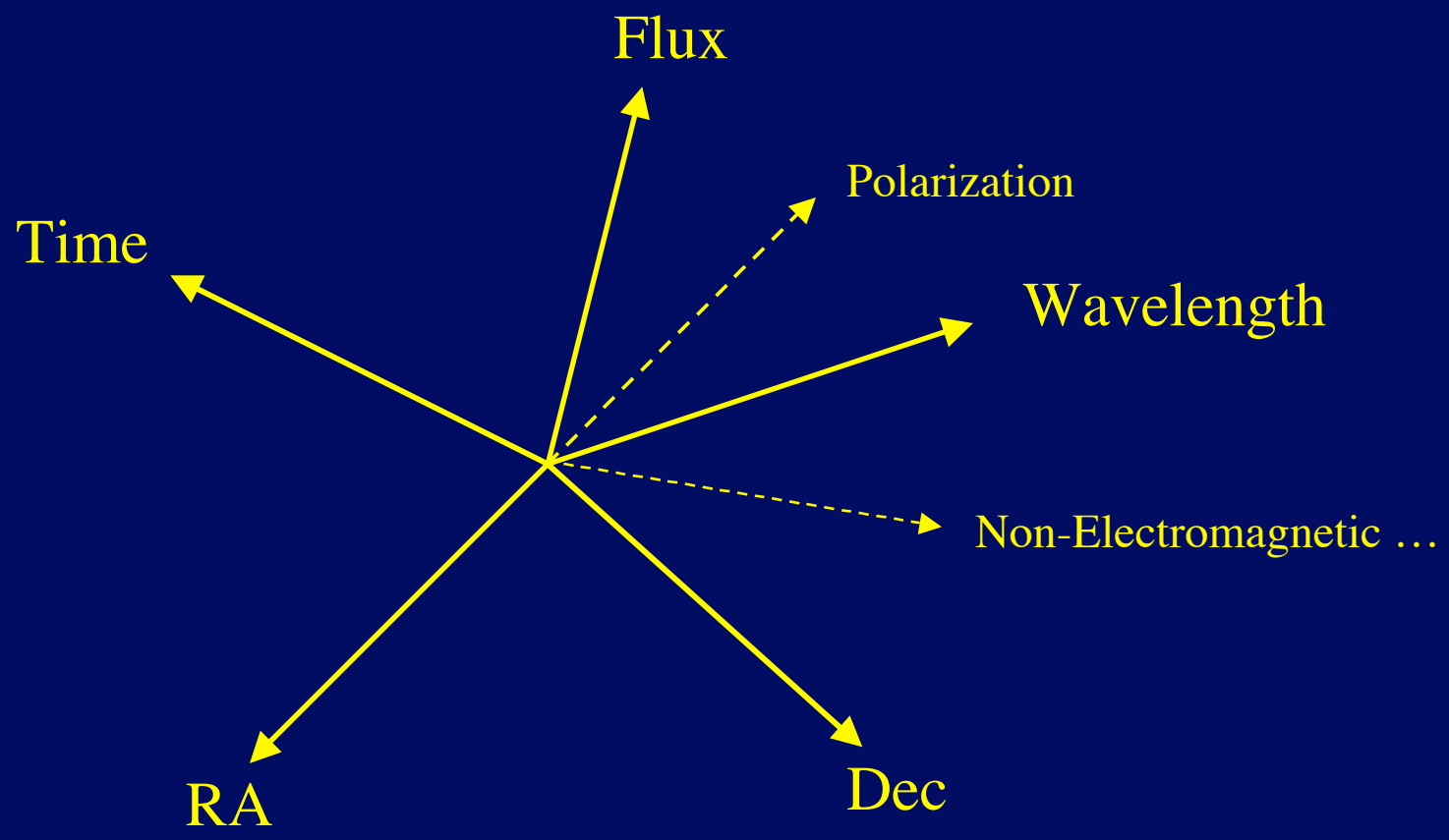
- Non-Gaussianity of clustering (*i.e.*, data modeling issues): power-law or exponential tails, clustering topology, etc.
  - — Essential in outlier/anomaly searches!
- Selection effects, data censoring, missing data, upper/lower limits, glitches …
- Data heterogeneity: variable types, very different error-bars and/or resolution

Another Challenge: **Effective Visualization** of Highly-Dimensional Parameter Spaces and Multivariate Correlations

- Your favorite graphics package is not enough
- A hybrid/interactive clustering+visualization approach?
- If it requires unusual equipment or long walks, forget it!
- A better use of dimensionality reduction techniques?

NASA and The Hubble Heritage Team (STScI) • Hubble Space Telescope W

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.

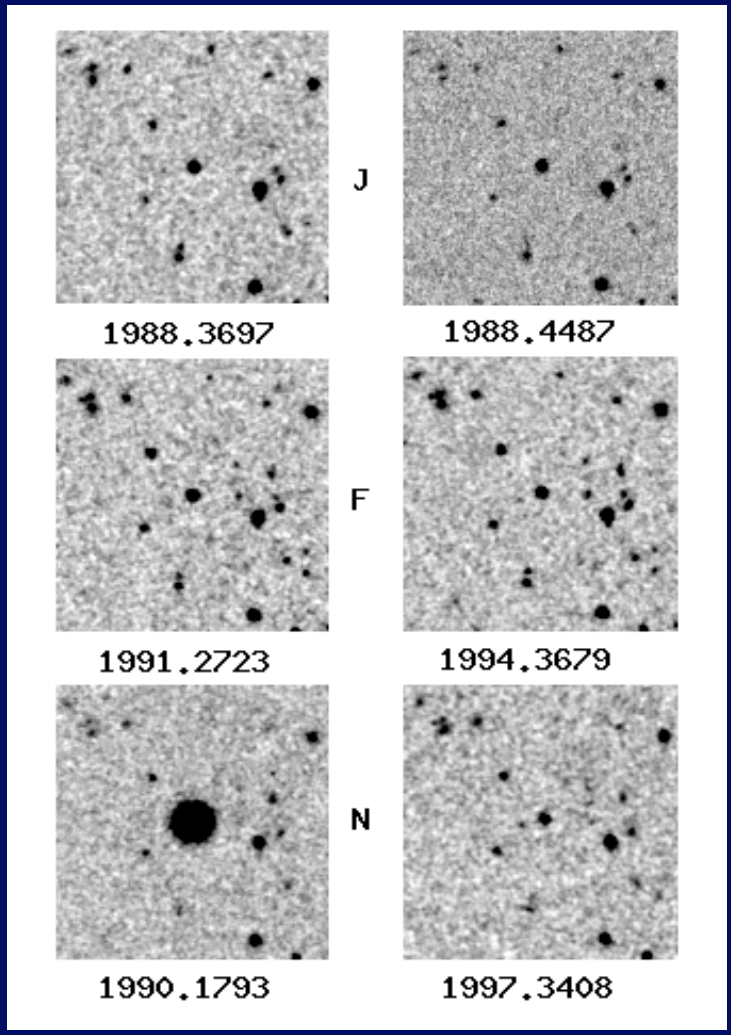
Exploration of New Domains of the Observable Parameter Space

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?

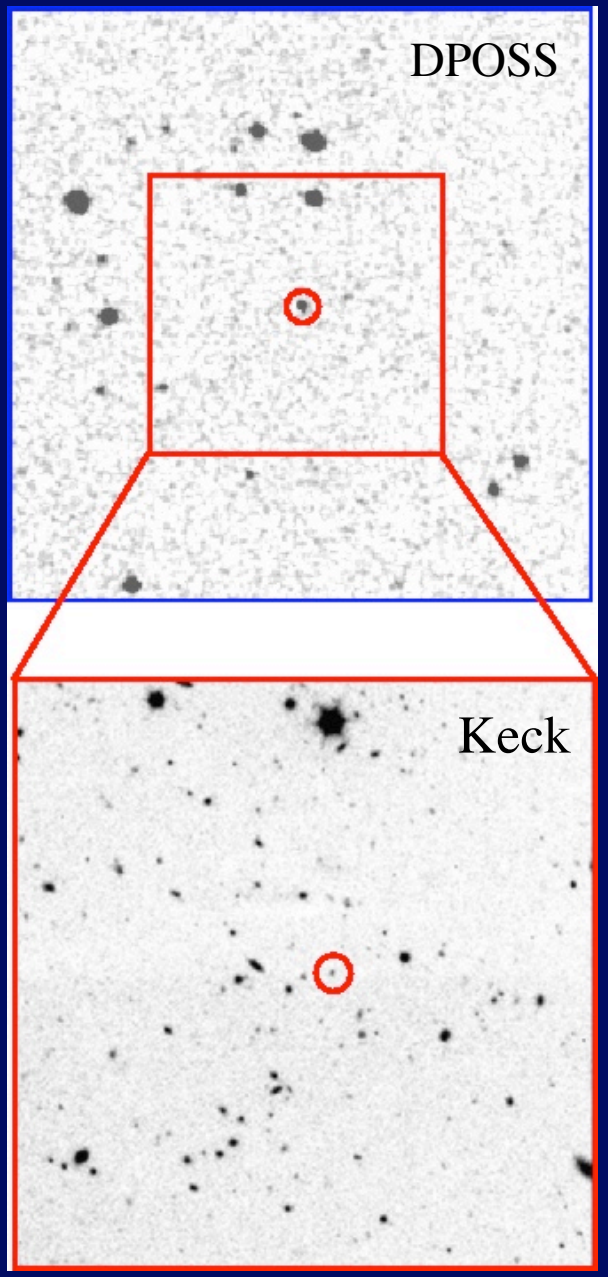
A new generation of synoptic sky surveys may provide the answers -- and uncover other new kinds of objects or phenomena.



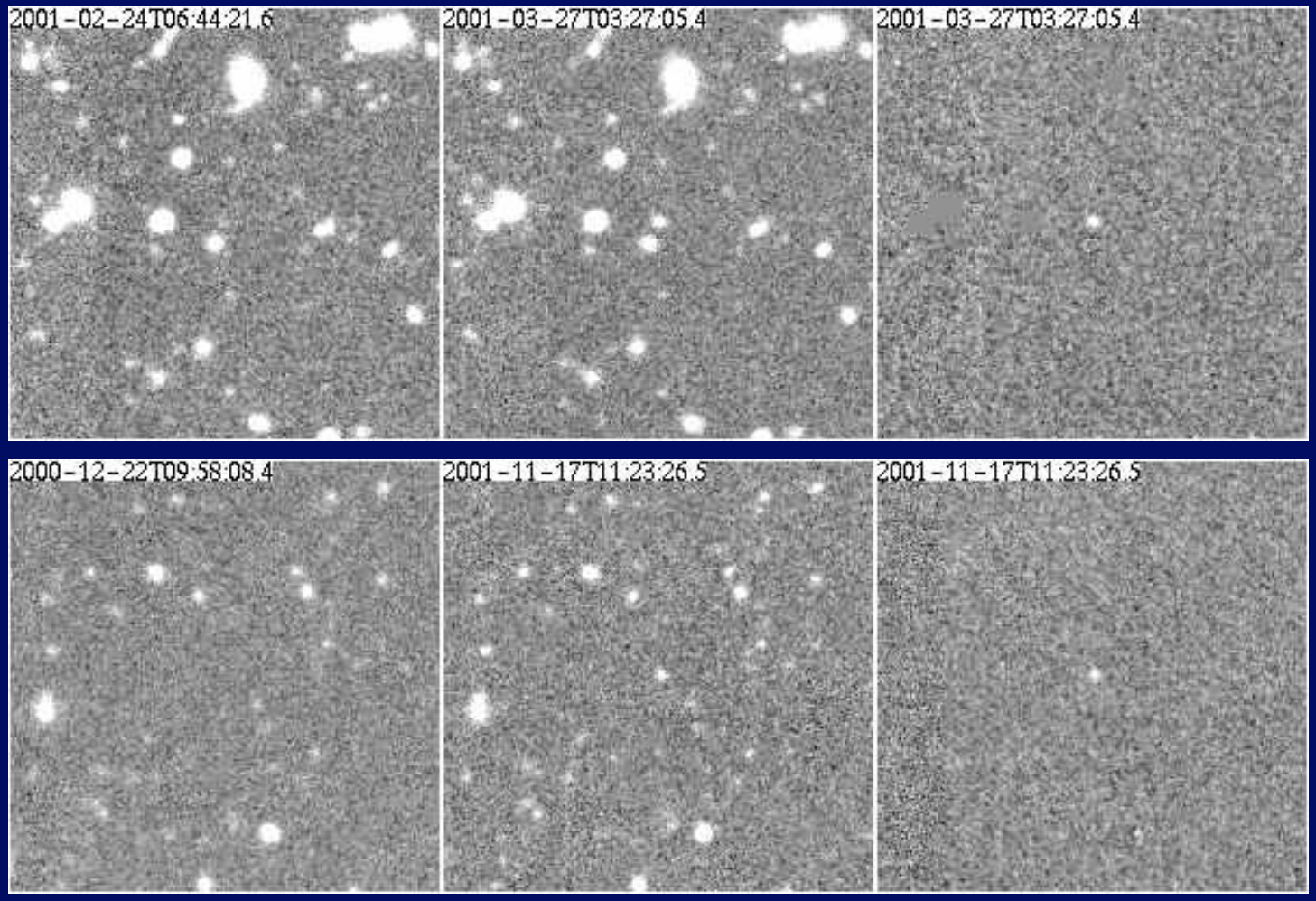
Exploration of the Time Domain: Optical Transients

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)

Astronomy and Other Fields

- Technical and methodological challenges facing the VO are **common to most data-intensive sciences** today, and beyond (commerce, industry, security …)
- **How is astronomy different?**
  - An intermediate ground in information volume, heterogeneity, and complexity (cf. high-energy physics, genomics, finance …)
- **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

Some Broad Issues:

- We are not making the full use of the growing data abundance in astronomy; we should!
- The old research methodologies, geared to deal with data sets many orders of magnitude smaller and simpler are no longer adequate
- The necessary technology and DM/KDD know-how are available, or can be developed
- The key issues are **methodological:** we have to learn to ask **new kinds of questions,** enabled by the massive data sets and technology

Sociological Issues:

Resisting the novelty of it …

But on the plus side:

- Enabling role! (professional outreach)
- Education and public outreach (astronomy *and* computer science)
- Training the new generation of scientific leaders

Towards 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

<u>The next golden age of discovery in astronomy?</u>

Concluding Comments:

- Astronomers need your help (and we know it)
- Great opportunities for collaborations and partnerships between astronomers, applied computer scientists, and statisticians
- Problems and challenges posed by the new astronomy may enrich and stimulate new CS/DM/KDD developments
- Interested? email george@astro.caltech.edu

visit http://www.astro.caltech.edu/~george/vo/