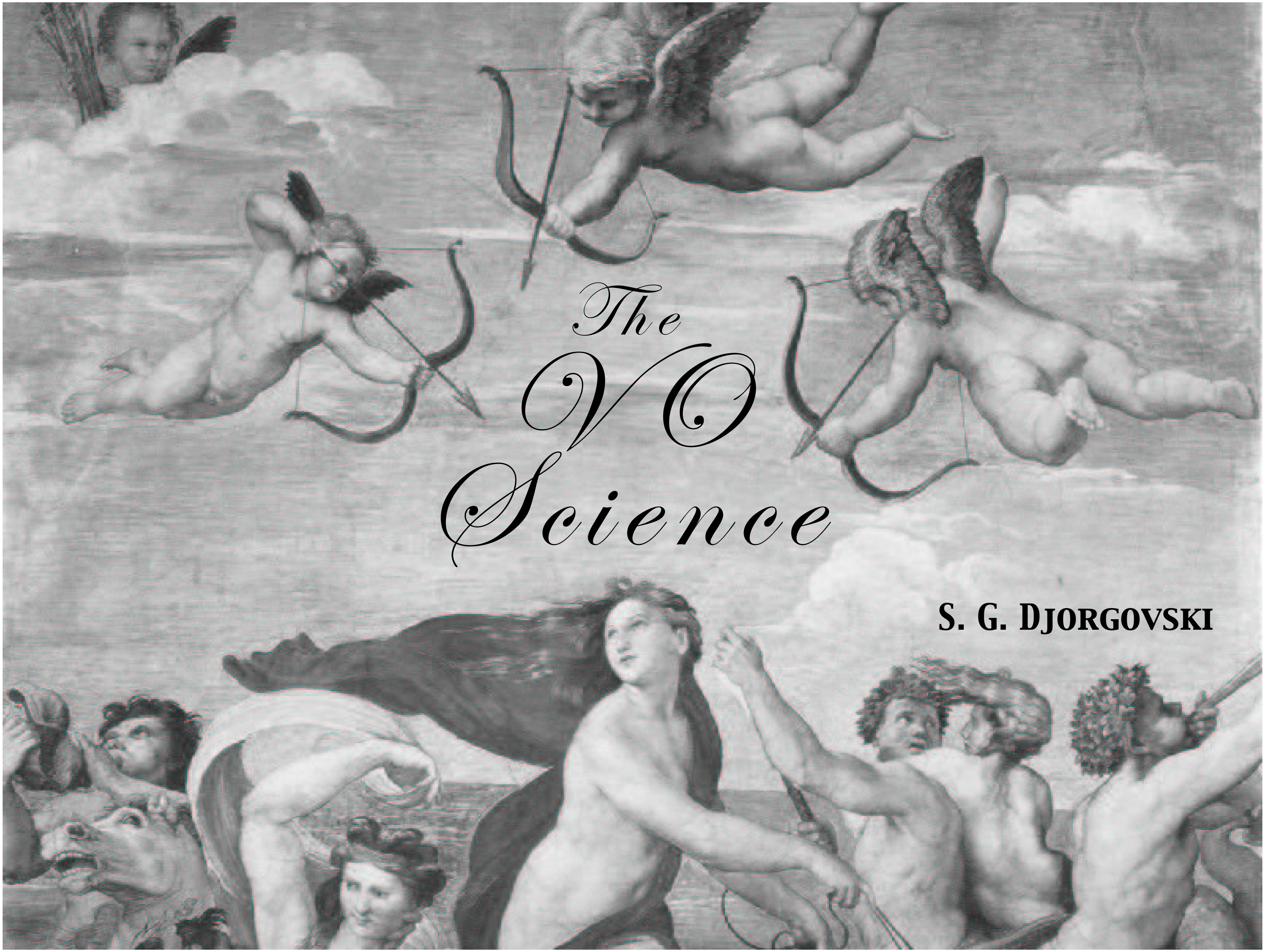
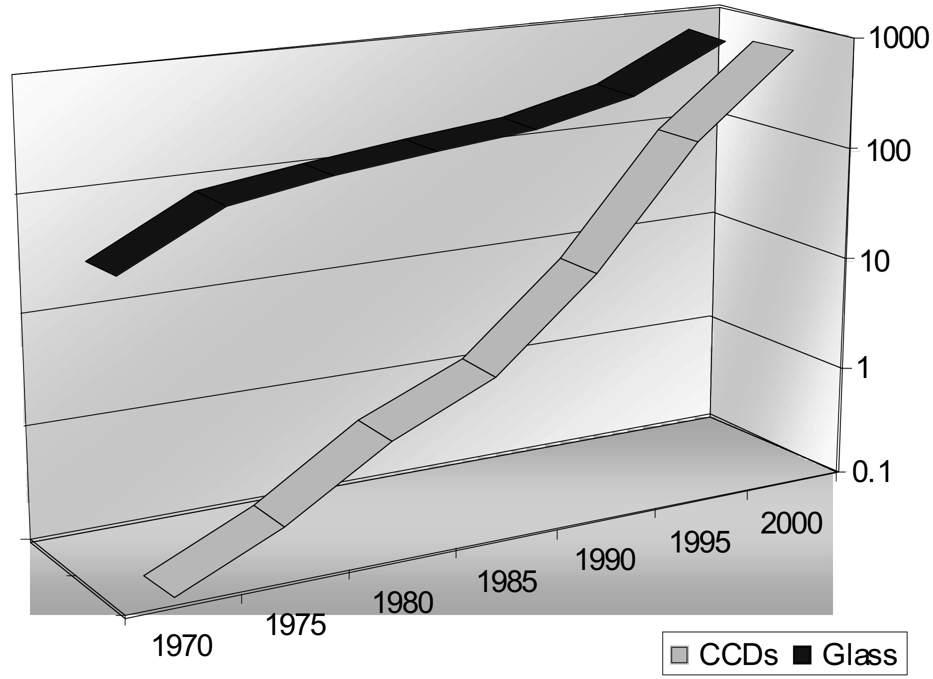


The VO Science

S. G. Djorgovski

Data ⇨ Knowledge ?

The exponential growth of data volume (and complexity, quality) driven by the exponential growth in information technology …



… But our understanding of the universe increases much more slowly -- ***Why?***

- ✦ Methodological bottleneck ➞ VO is the answer
- ✦ Human wetware limitations …

➞ AI-assisted discovery ➞ NGVO?

How and Where are Discoveries Made?

- **Conceptual Discoveries:** e.g., Relativity, QM, Brane World, Inflation … *Theoretical, may be inspired by observations*
- **Phenomenological Discoveries:** e.g., Dark Matter, QSOs, GRBs, CMBR, Extrasolar Planets, Obscured Universe … *Empirical, inspire theories, can be motivated by them*

New Technical Capabilities ⇒ (IT/VO) Observational Discoveries ⇔ (VO) Theory

Phenomenological Discoveries:

- Pushing along some parameter space axis ⇐ VO useful
- Making new connections (e.g., multi-λ) ⇐ **VO critical!**

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

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

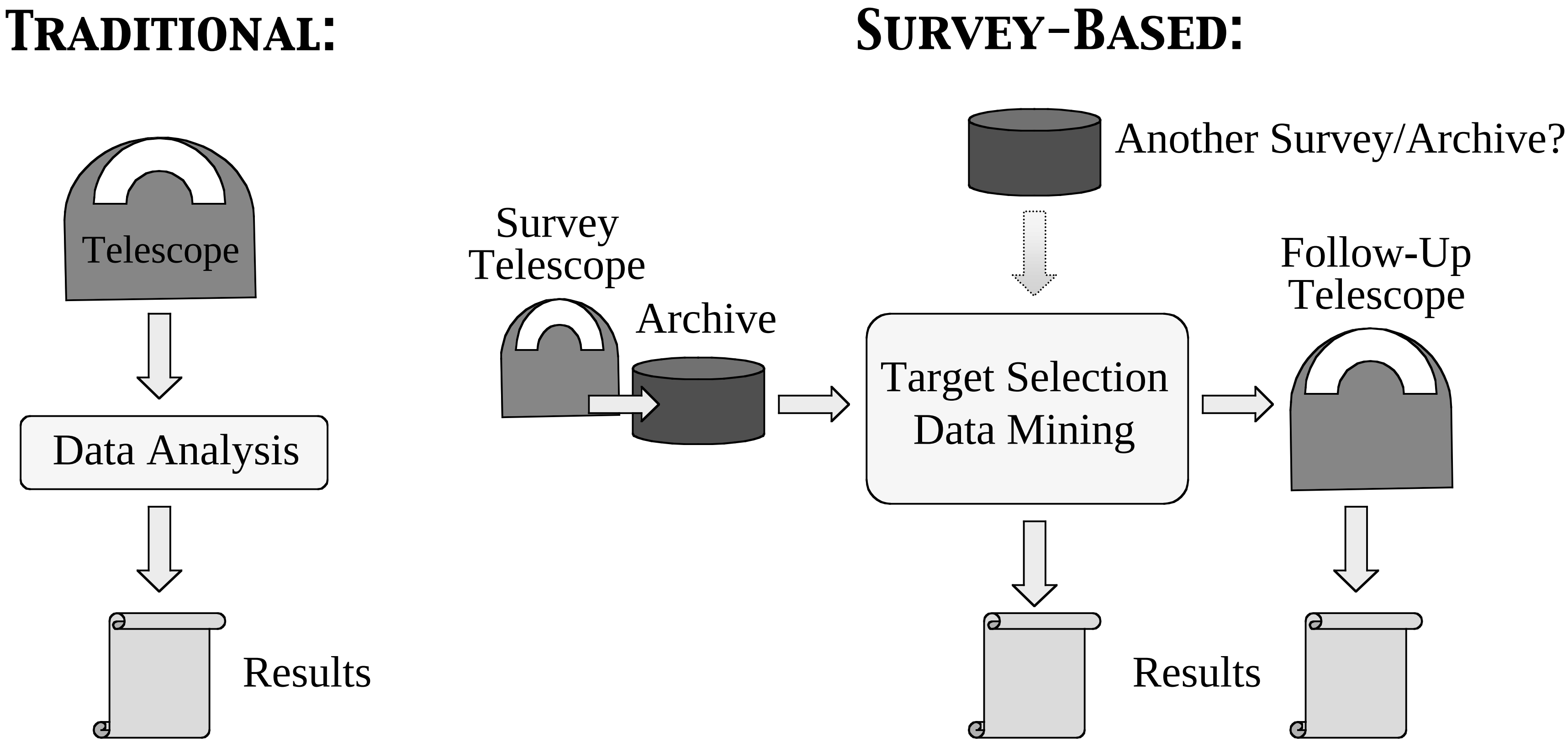
    Quasars, CMBR, x-ray astronomy, pulsars, GRBs, …
  - 1980's - 1990's: computers, digital detectors (CCDs etc.)

    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

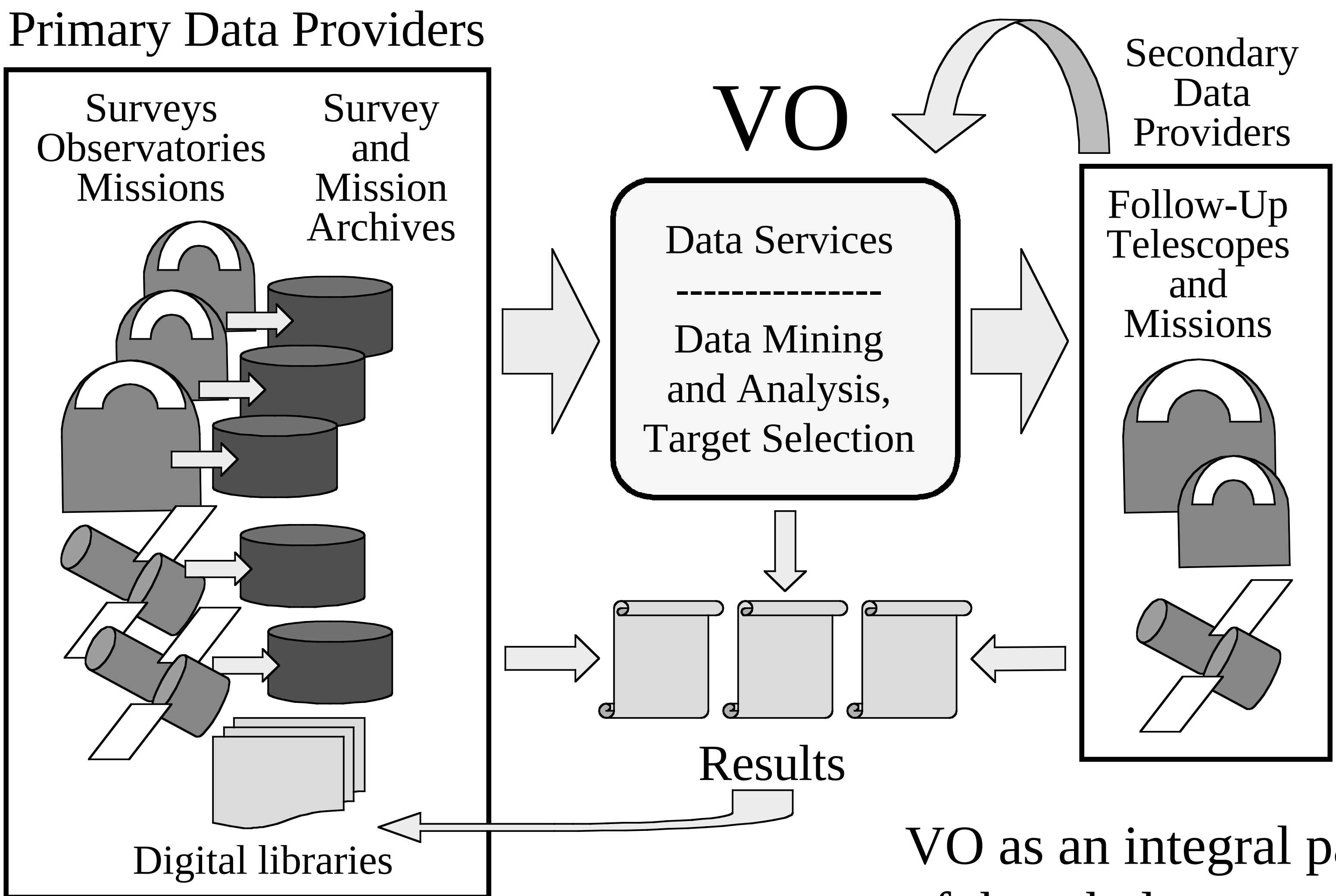
From Traditional to Survey to VO-Based Science



Highly successful and increasingly prominent, but inherently limited by the information content of individual surveys …

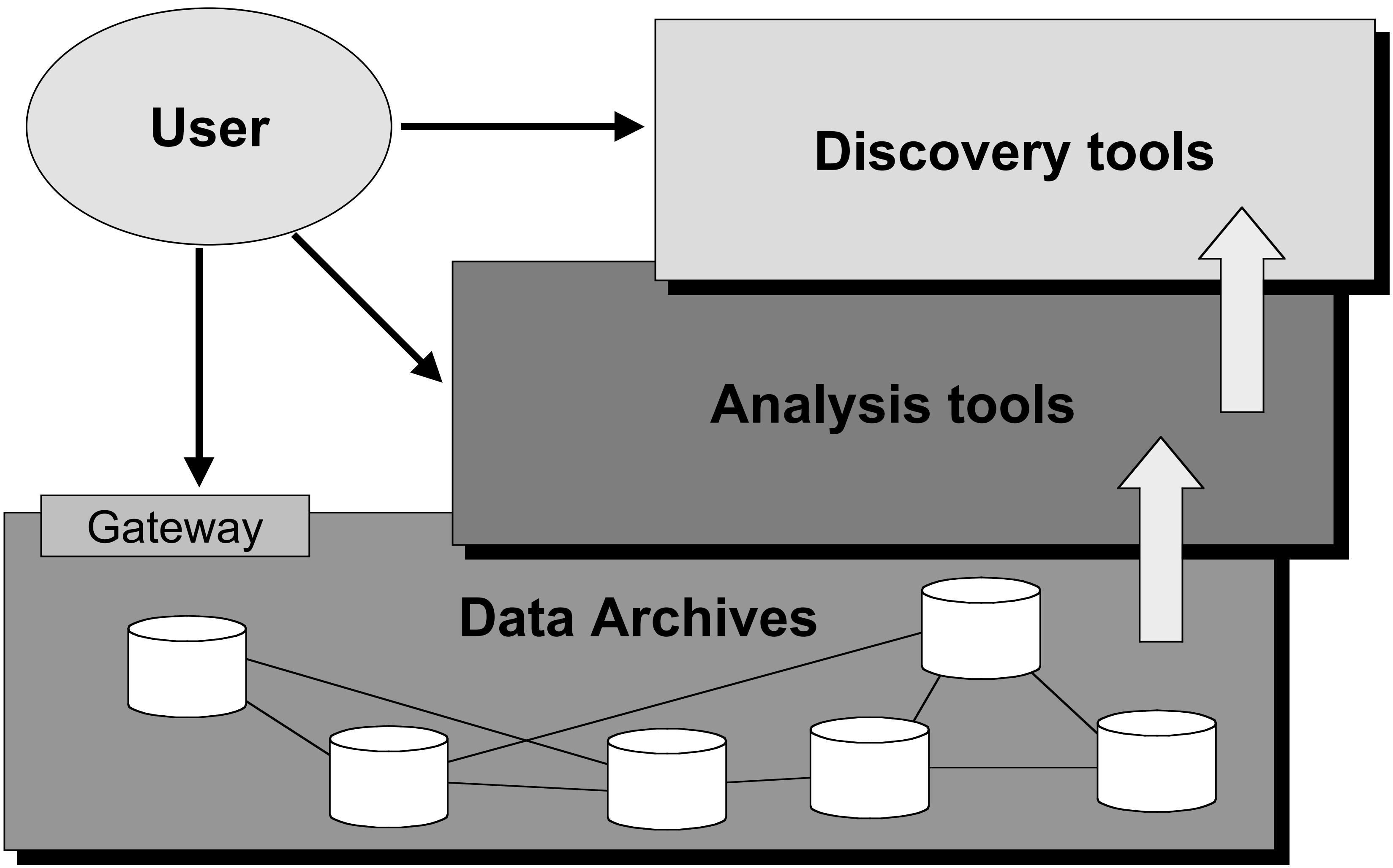
*What comes next, beyond survey science is the **VO science***

A Schematic Illustration of the VO-Based Science



VO as an integral part of the whole system …

VO: Conceptual Architecture



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 (~ 10^0 - 10^3)	Archives of pointed observations (~ TB)	**Virtual Observatory**

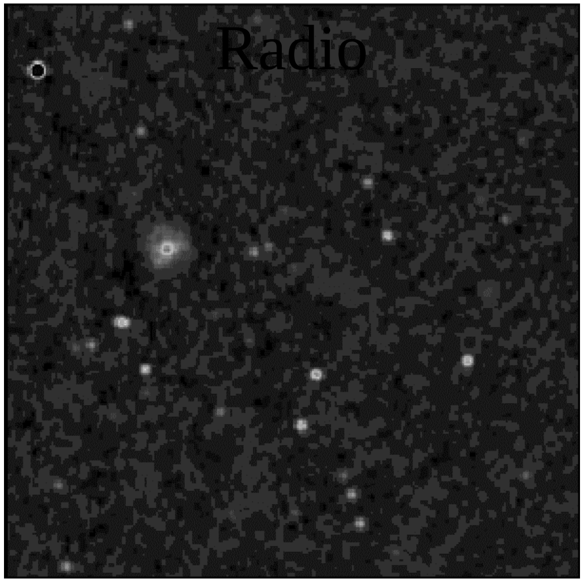
This quantitative change in the information volume and complexity will enable the

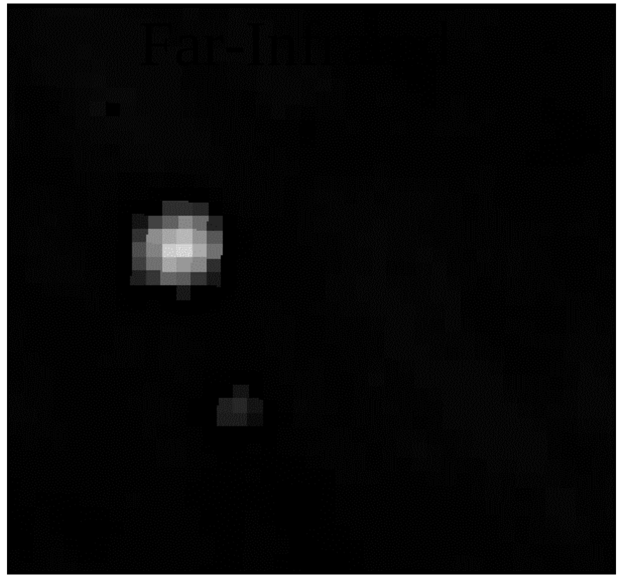
Science of a Qualitatively Different Nature:

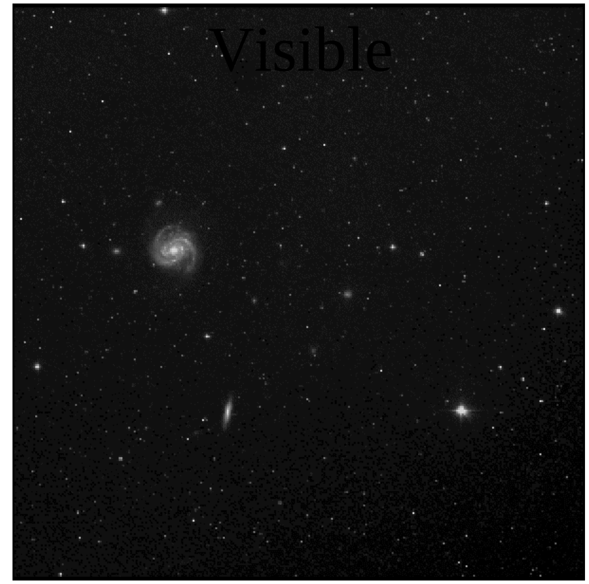
- **Statistical astronomy done right**
  - Precision cosmology, Galactic structure, stellar astrophysics …
  - Discovery of significant patterns and multivariate correlations
  - Poissonian errors unimportant
- **Systematic exploration of the observable parameter spaces** (NB: Energy content $\neq$ Information content)
  - Searches for rare or unknown types of objects and phenomena
  - Low surface brightness universe, the time domain …
- **Confronting massive numerical simulations with massive data sets**

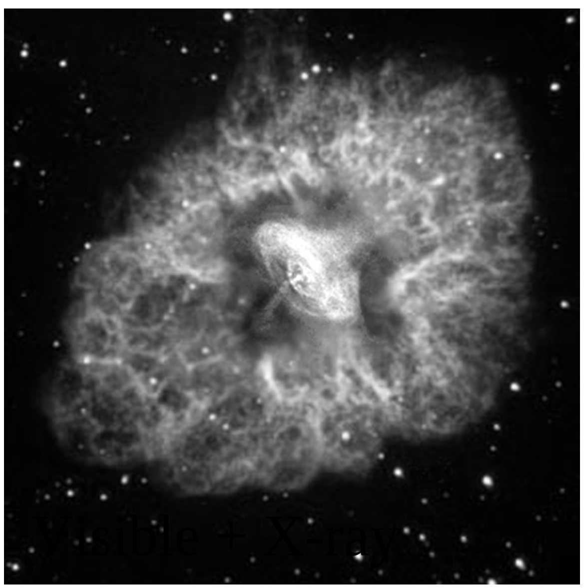
Panchromatic Views of the Universe:

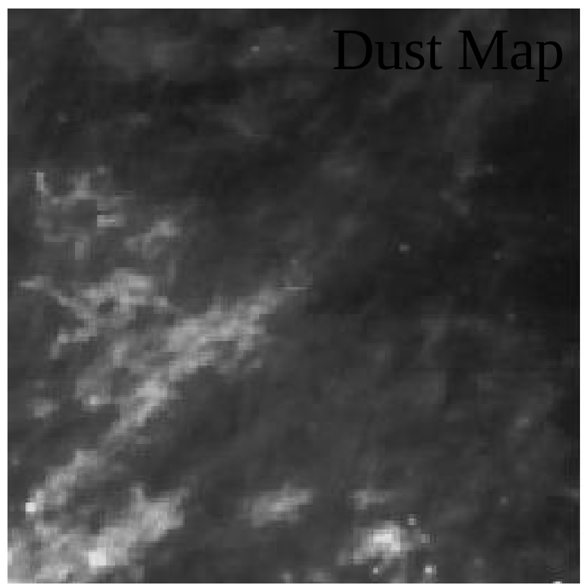
A More Complete, Less Biased Picture

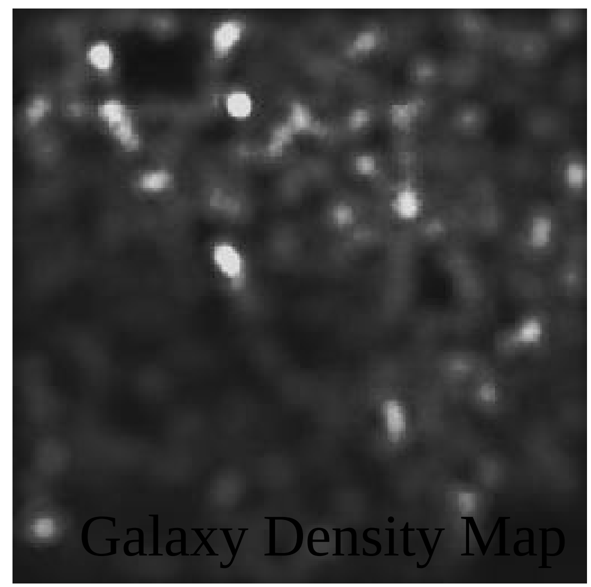










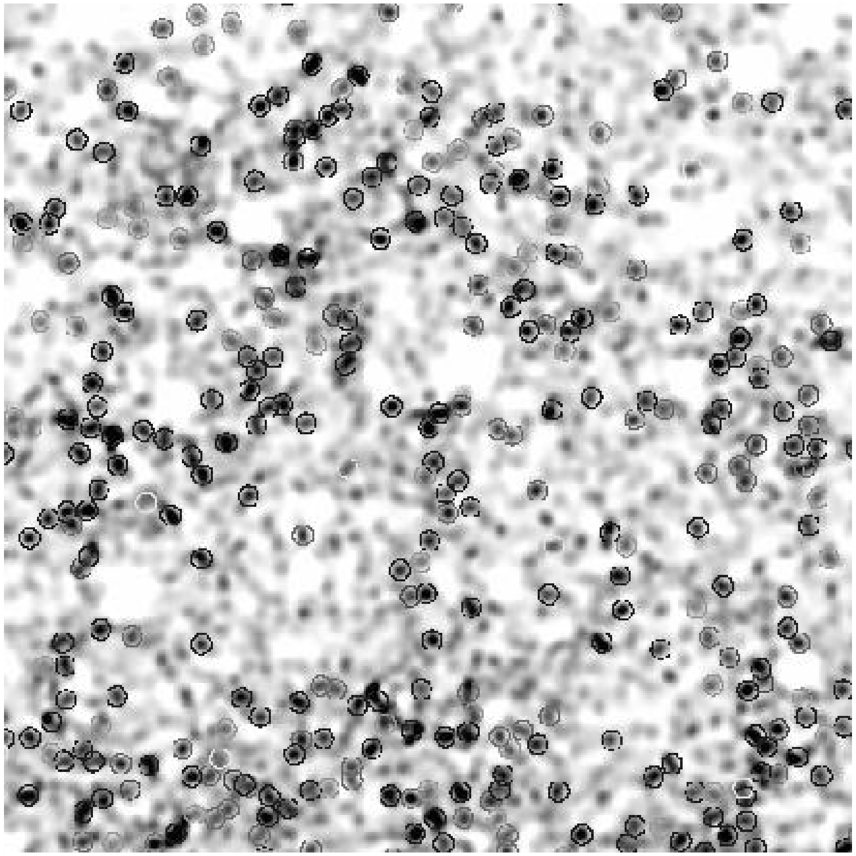


Examples of Possible VO Projects:

- **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, diffuse 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*

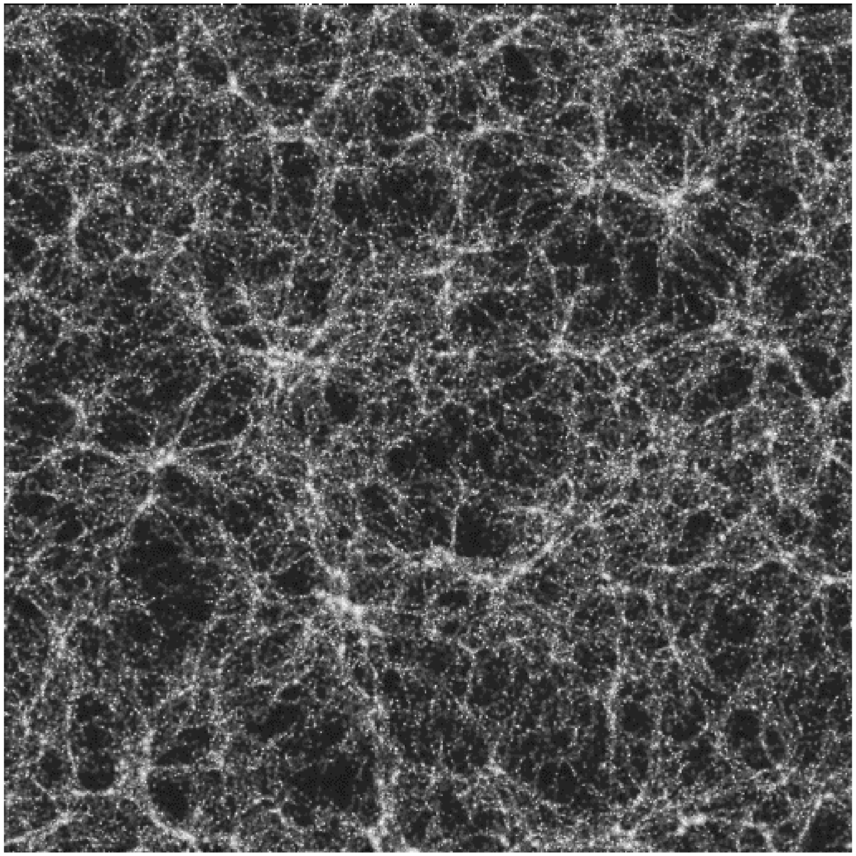
- ☆ Precision Cosmology and LSS
- ☆ Better matching of theory and observations

Clustering on a clustered background



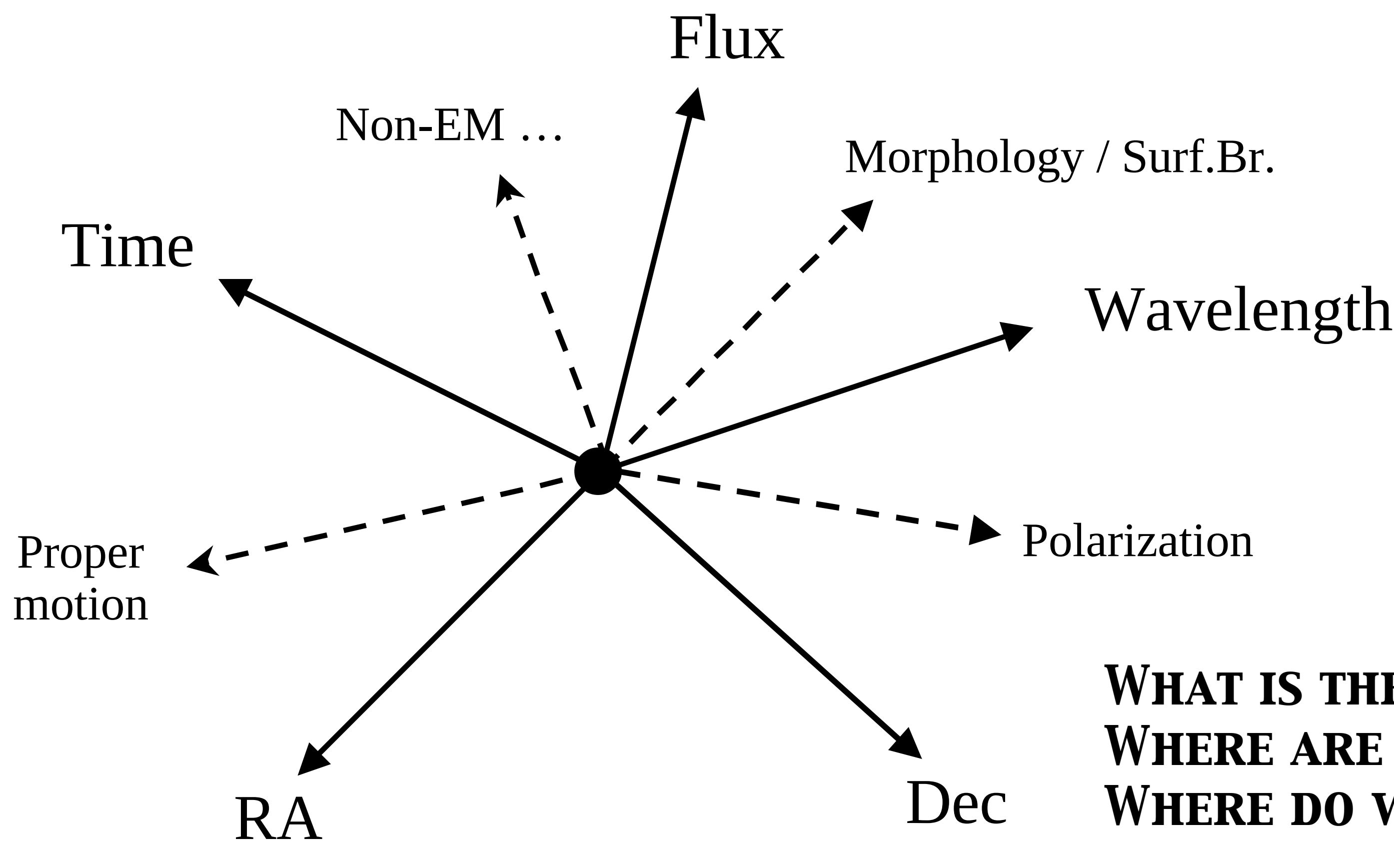
DPOSS Clusters (Gal et al.)

Clustering with a nontrivial topology



LSS Numerical Simulation (VIRGO)

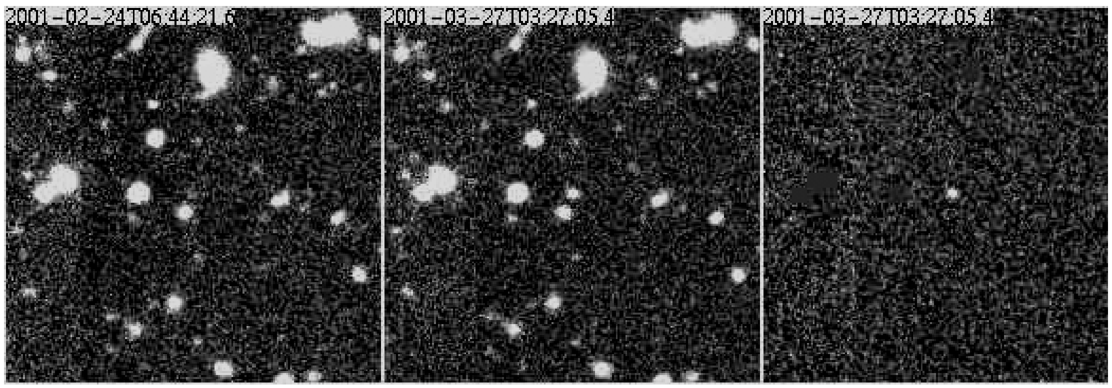
Taking a Broader View: The Observable Parameter Space



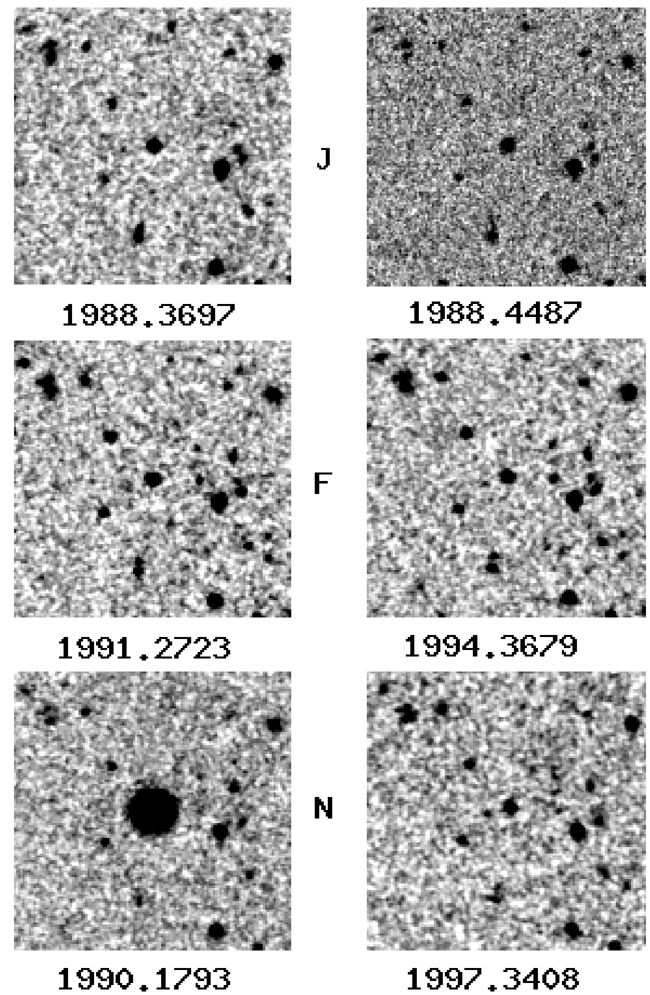
WHAT IS THE COVERAGE?
WHERE ARE THE GAPS?
WHERE DO WE GO NEXT?

Along each axis the measurements are characterized by the **position, extent, sampling and resolution.** All astronomical measurements span some volume in this parameter space.

Exploration of new domains of the observable parameter space: **the Time Domain**



Faint, Fast Transients (Tyson et al.)



Megaflares on normal MS stars (DPOSS)

Existing and Forthcoming surveys:

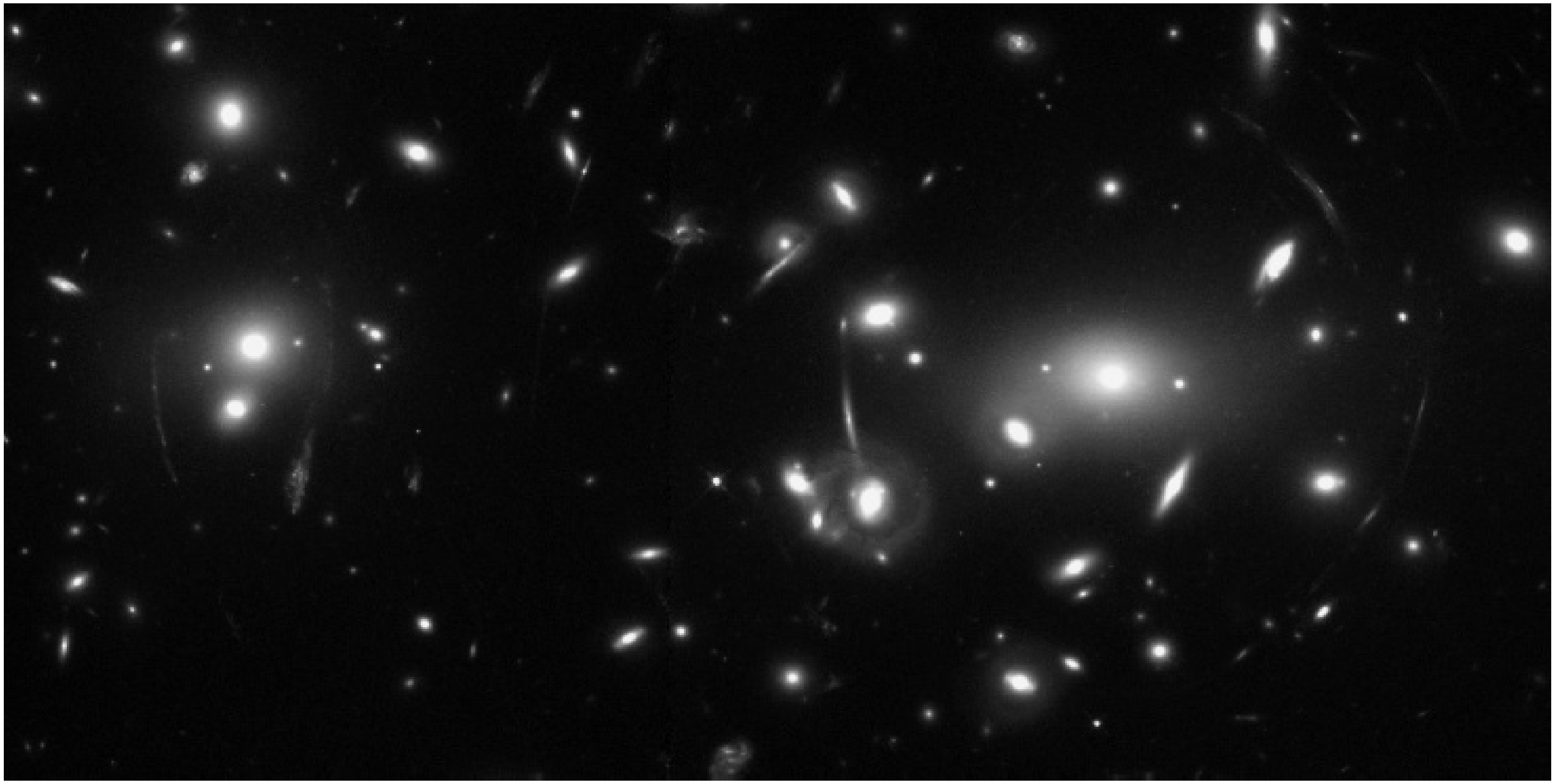
- Microlensing experiments (OGLE, MACHO …)
- Solar System patrols, GRB patrols …
- DPOSS plate overlaps (Mahabal et al.)
- QUEST-2 and NEAT at Palomar

… and many, many others …

- The future: **LSST** ?

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)

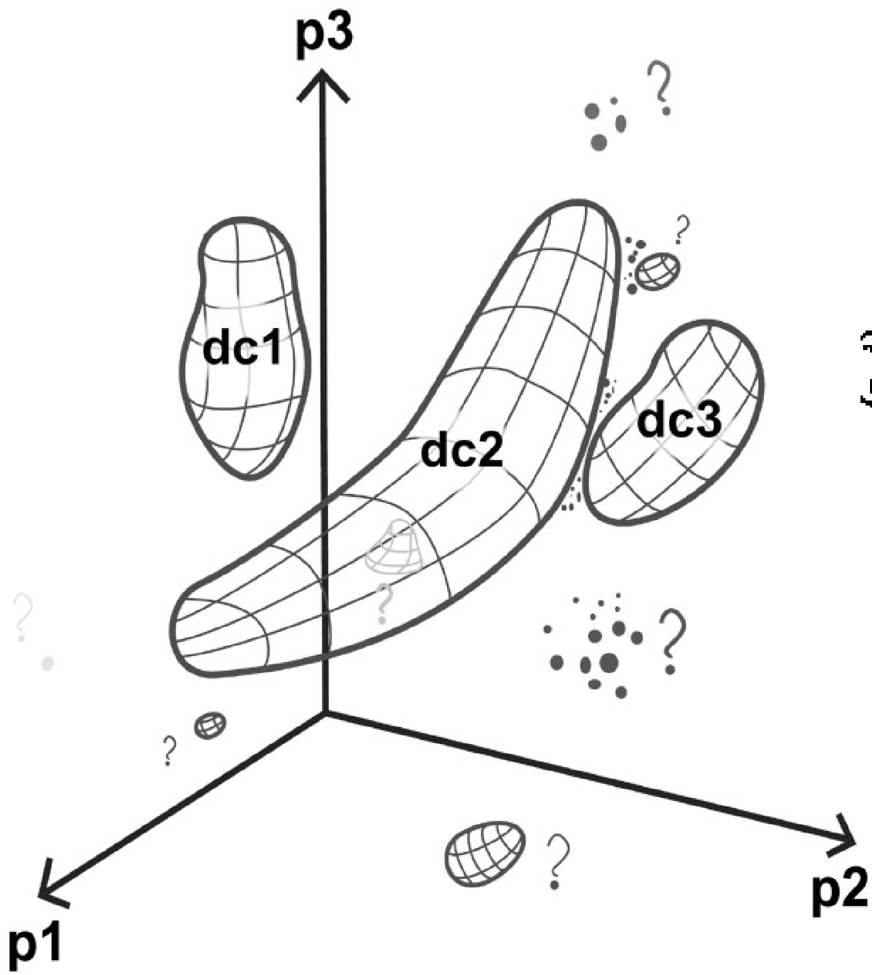


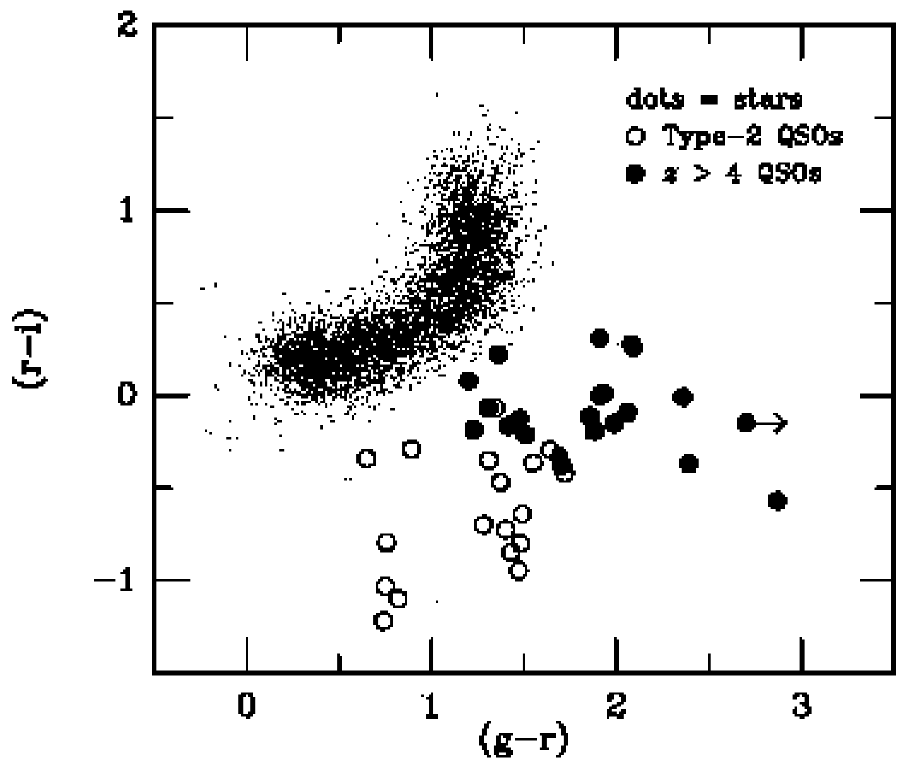
- Effective parametrization of source morphologies and environments
- Multiscale analysis

(Also: in the time/lightcurve domain)

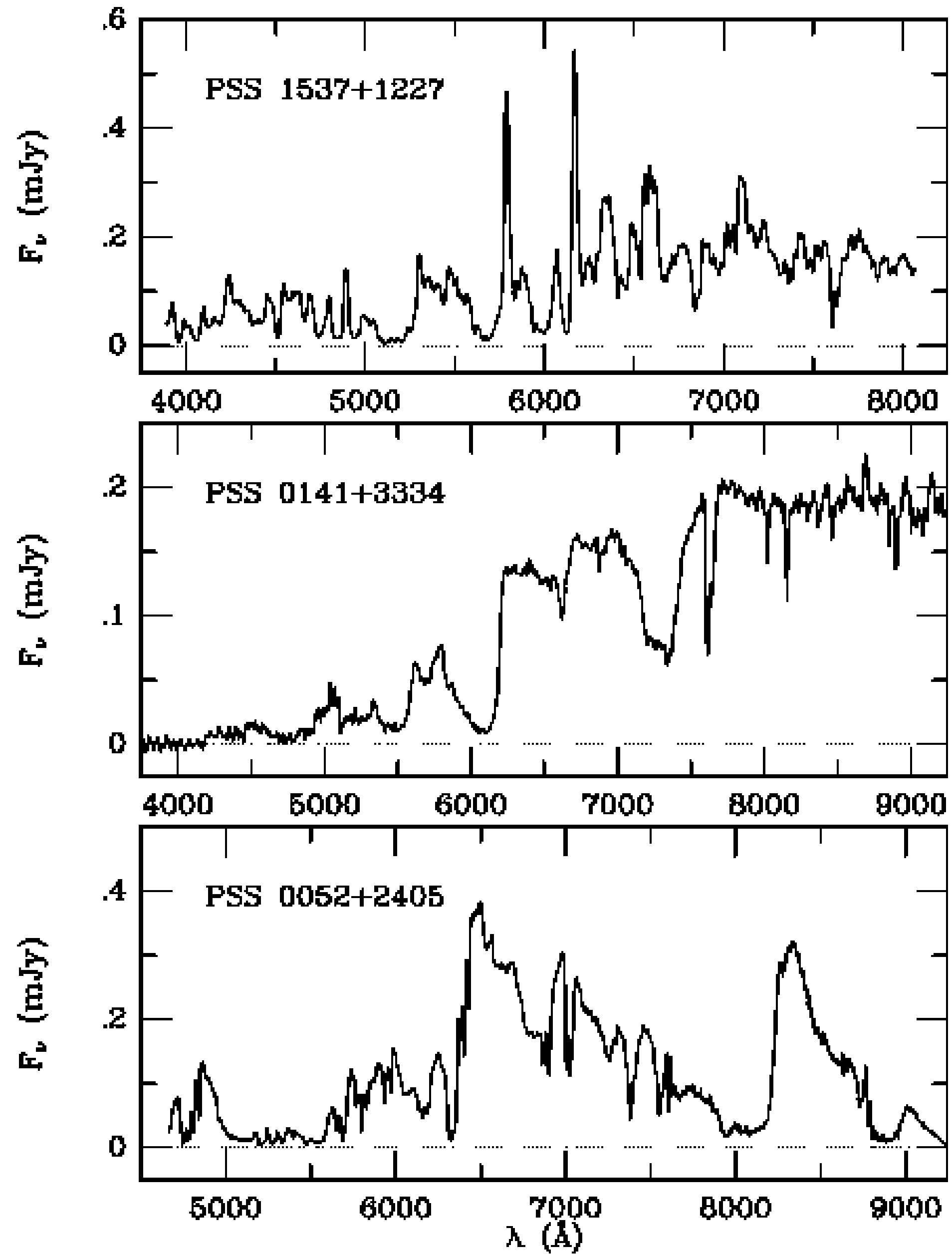
Exploration of observable parameter spaces and searches for rare or new types of objects

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





An example of a not quite new, but a rare subspecies of objects



Spectra of peculiar Lo-BAL (Fe) QSOs discovered in DPOSS (also in FIRST, SDSS)

Exploration of Parameter Spaces in the Catalog Domain (Source Attributes)

- Clustering Analysis (supervised and unsupervised):
  - How many different types of objects are there?
  - Are there any rare or new types, outliers?
- Multivariate Correlation Search:
  - Are there significant, nontrivial correlations present in the data?

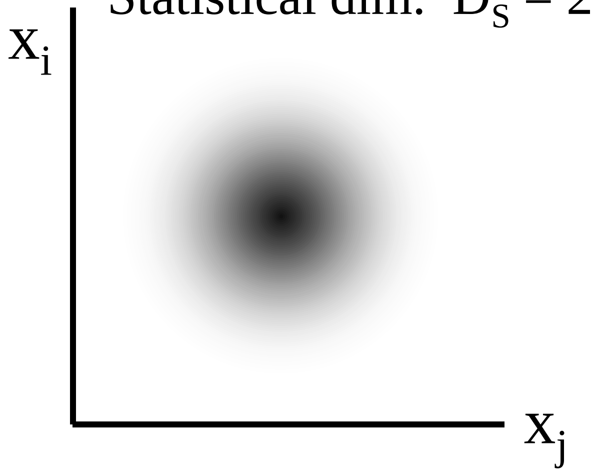
Clusters vs. Correlations:

Astrophysics ➞ Correlations

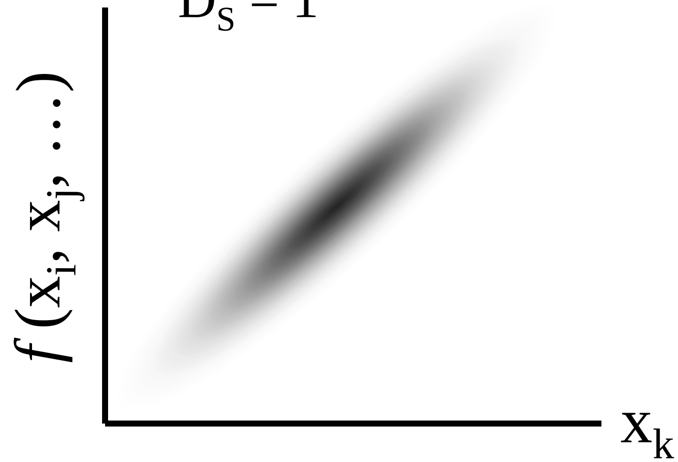
Correlations ➞ reduction of the statistical dimensionality

New Astrophysics from Multivariate Correlations

Data dimension $D_D = 2$
Statistical dim. $D_S = 2$

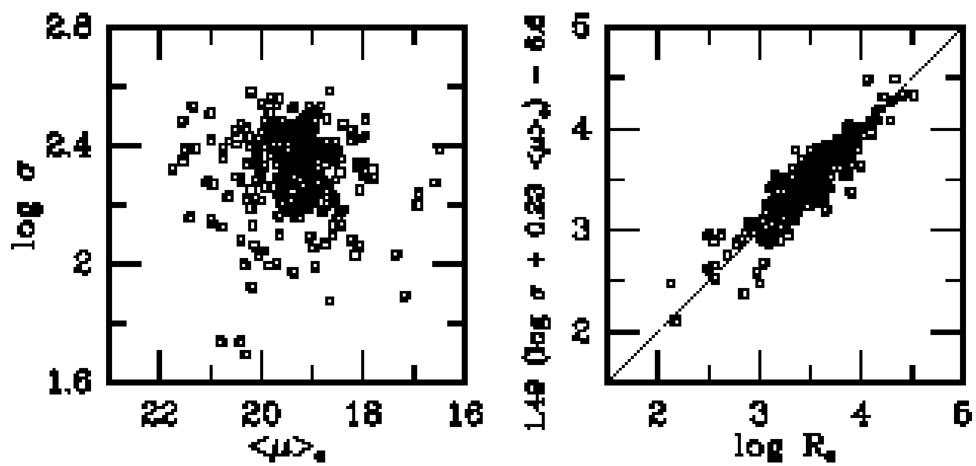


$D_D = 2$
$D_S = 1$



If $D_S < D_D$, then MV correlations are present

Fundamental Plane of E-galaxies:



Correlations objectively define types of objects, e.g.,
TFR ➔ normal spirals,
FP ➔ normal ellipticals
… and can lead to some **new insights**

In VO data sets: $D_D >> 1$, $D_S >> 1$

Data Complexity ➙ Multidimensionality ➙ Discoveries

But the bad news is …

The Curse of Hyperdimensionality, Part I:

The computational cost of clustering analysis:

K-means: $K \times N \times I \times \mathbf{D}$

Expectation Maximisation: $K \times N \times I \times \mathbf{D^2}$

Monte Carlo Cross-Validation: $M \times K_{max}{}^2 \times N \times I \times \mathbf{D^2}$

N = no. of data vectors, D = no. of data dimensions

K = no. of clusters chosen, K_{max} = max no. of clusters tried

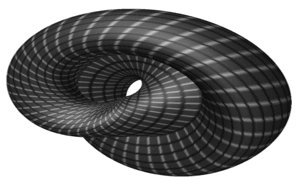
I = no. of iterations, M = no. of Monte Carlo trials/partitions

⇨ *Terascale (Petascale?) computing and/or better algorithms*

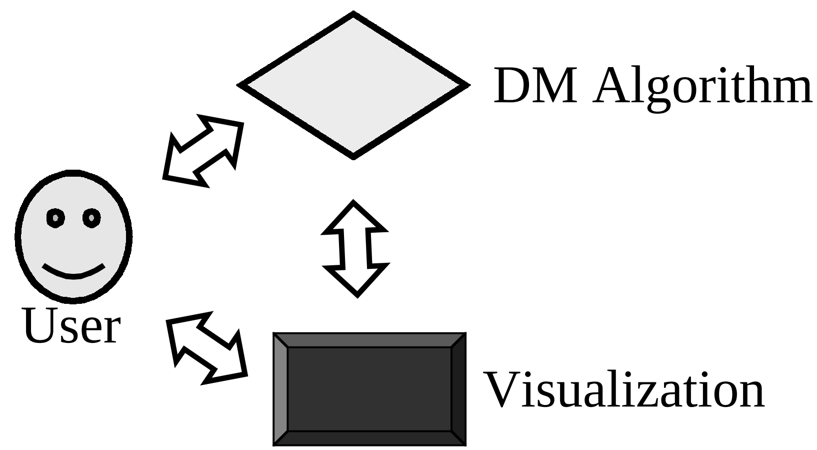
Some dimensionality reduction methods do exist (e.g., PCA, class prototypes, hierarchical methods, etc.), but more work is needed

The Curse of Hyperdimensionality, Part II:

- **Visualization!**



- A fundamental limitation of the human perception: D_{MAX} = 3? 5? (NB: We can certainly understand mathematically much higher dimensionalities, but cannot really visualize them; our own Neural Nets are powerful pattern recognition tools)
- Interactive visualization as a key part of the data mining process:



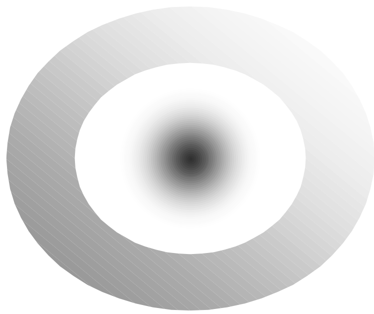
- Some methodology exists, but much more is needed

Some Problems and Issues in Multivariate Analysis of VO Data Sets:

- Data heterogeneity, biases, selection effects …
- Non-Gaussianity of clusters (data models)

Outlier population, or
a non-Gaussian tail?

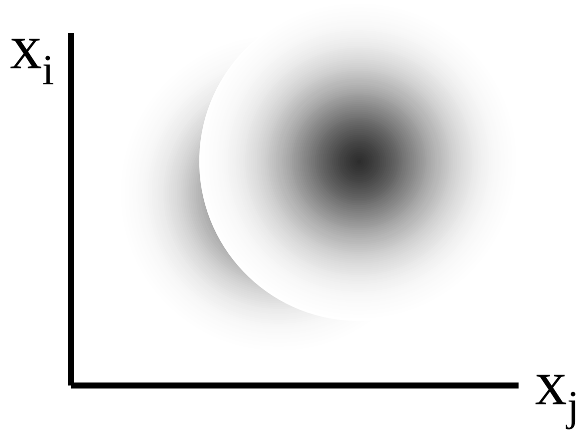
- Non-trivial topology of clustering

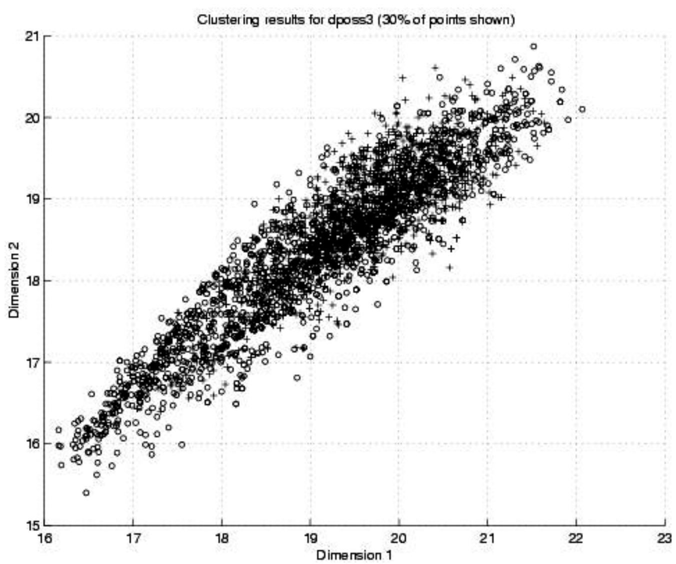


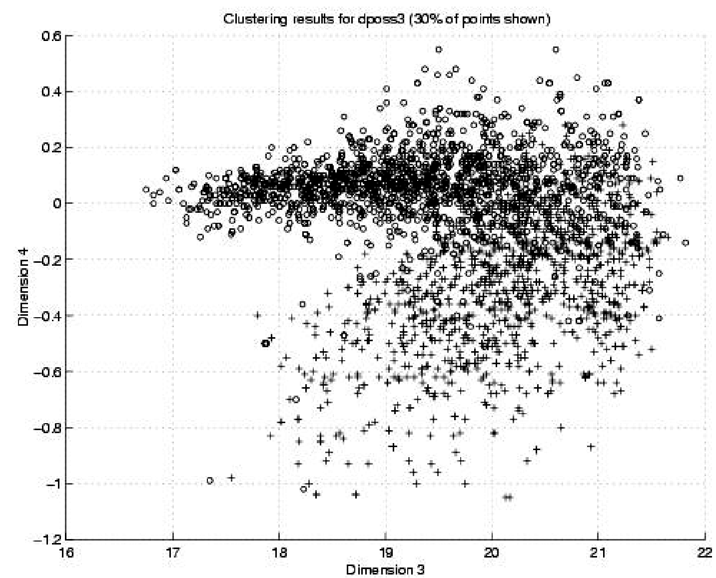
- Useful vs. “useless” parameters …

Useful *vs.* “Useless” Parameters:

Clusters (classes) and correlations may exist/separate in some parameter subspaces, but not in others







Scientific Roles of the VO

- **Efficiency amplifier:** Facilitate science with massive data sets (observations and theory/simulations)
- **Added value** from federated data sets (multi-wavelength, multi-scale, multi-epoch …)
  - Historical examples: the discoveries of Quasars, ULIRGs, GRBs, radio or x-ray astronomy …
- **Enable some *new* science** with massive data sets (not just old but bigger -- different!)
- **Optimize** the use of expensive resources: space missions and large ground-based telescopes)
- Provide R&D drivers, application testbeds, and stimulus to the **partnering disciplines** (CS/ITR, statistics …)

Data-Rich Astronomy and Other Fields

- Technical and methodological challenges facing the VO are **common to most data-intensive sciences** today, and beyond (commerce, industry, finance, etc.)
- **Interdisciplinary exchanges** (e.g., with physics, biology, earth sciences, etc.) → intellectual cross-fertilization, avoid wasteful duplication of efforts
- **Partnerships and collaborations** with applied CS/IT are **essential,** may lead to significant technological advances

High-energy physics → **WWW !**
High-energy physics → **The Grid**
Astronomy (VO) → **The Grid**
Astronomy (VO) → **???**

Sociological Issues …

Concluding Comments:
The VO-Enabled, Information-Rich Astronomy for the 21st Century

- The most promising (initial?) prospects for the new, VO-enabled science:
  - Unsupervised clustering analysis of large parameter spaces
  - Searches for new MV correlations ➙ **New Astrophysics?**
  - Systematic exploration of new domains of the observable parameter space (e.g., the time domain, multiscale universe)

➙ **Discoveries of new types of objects and phenomena?**

- We need:
  - More emphasis on data mining and visualization
  - Better clustering algorithms and exploratory statistical tools

  ➙ ***Stronger partnerships with CS/IT, statistics***

Scaling the VO Mountain