On-Line Resources, Archives, Tools, Virtual Observatory, etc.

The Key Points

- Chances are that the data you need already exist, at least for a pilot project (no proposals, no waiting, no clouds…)
- The existing data sets are so information-rich (and getting more so), that it is practically certain that there are potential new discoveries waiting to be made
- This is the wave of the future: computationally-enabled, data-rich science for the 21st century
  - This applies to all fields of science, not just astronomy
  - You need some new research skills
  - "The computer is the new telescope" (and the database is "the new sky") – you can make first-rate observational discoveries without ever going observing
    - ✧ Good bye, Caltech astronomy dominance model!
  - Often the best leverage for new observations is to combine them with the existing archival data

Astronomy Has Become Very Data-Rich

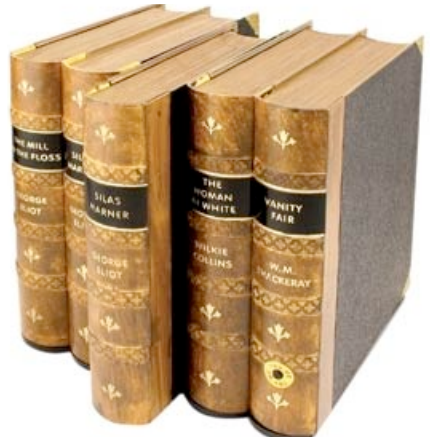
- Typical digital sky survey generated ~ 10 - 100 TB each, plus a comparable amount of derived data products
  - PB-scale data sets are imminent
- Astronomy today has ~ a few PB of archived data, and generates ~ 10 TB/day
  - Both data volumes and data rates grow exponentially, with a ***doubling time ~ 1.5 years***
  - Even more important is the growth of ***data complexity***
- For comparison:

  Human Genome < 1 GB

  Human Memory < 1 GB (?)

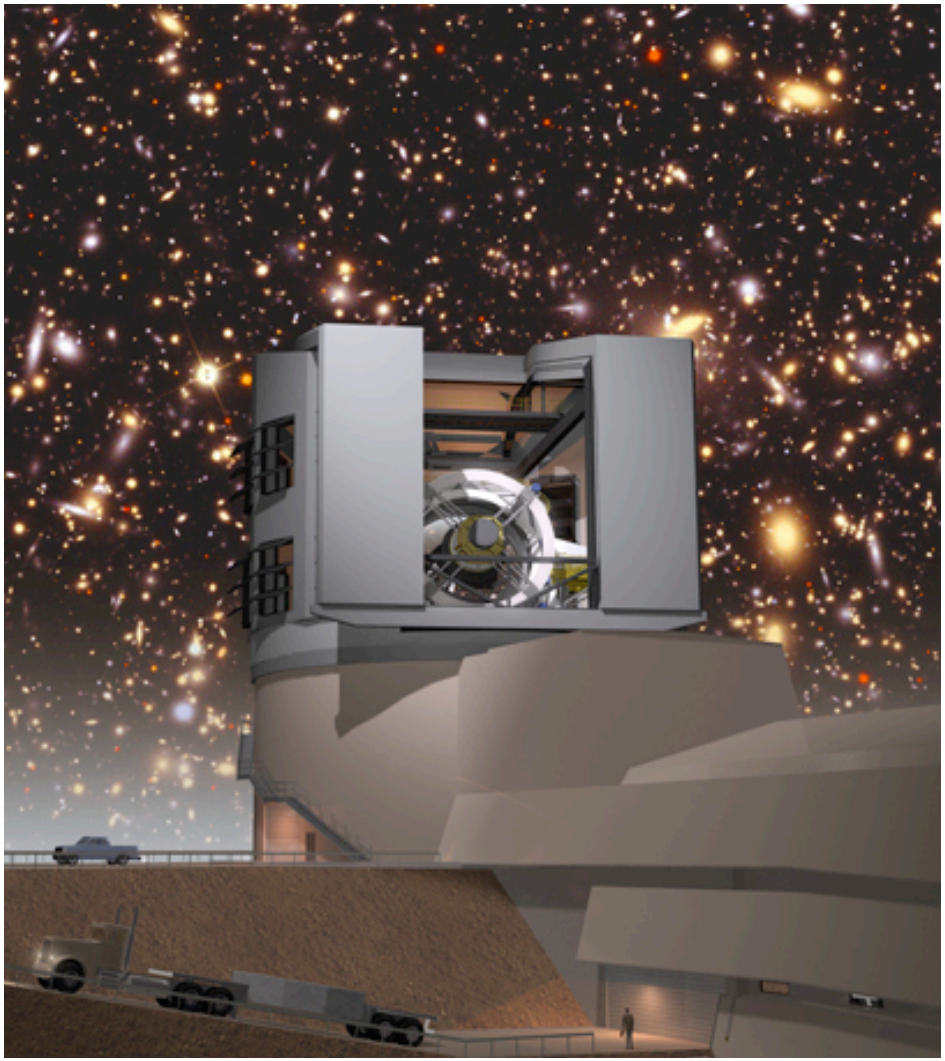
  1 TB ~ 2 million books

  Human Bandwidth ~ 1 TB / year (±)

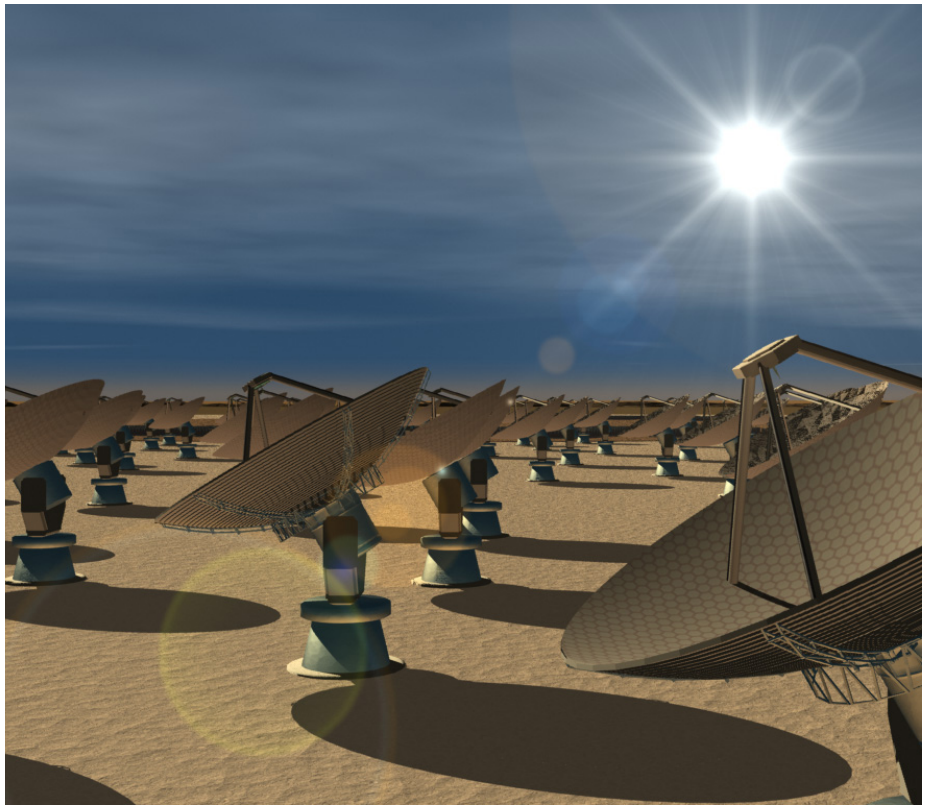


… And It Will Get Much More So

Large Synoptic Survey Telescope
(LSST) ~ 30 TB / night



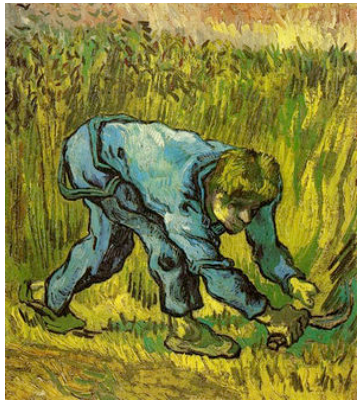
Square Kilometer Array (SKA)
~ 1 EB / second (raw data)
(EB = 1,000,000 TB)



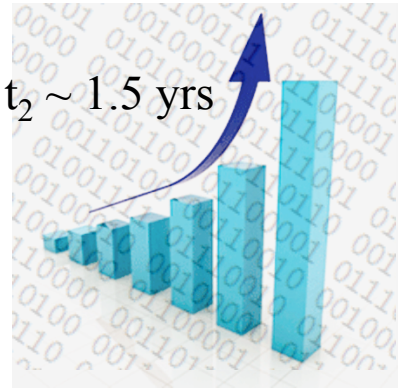
The Evolving Data-Rich Astronomy

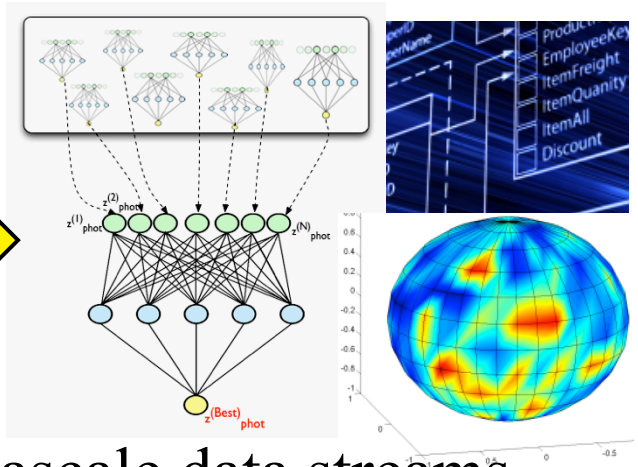
From "arts & crafts" to industry

From data subsistence to an exponential overabundance



Astronomy is now driven by the progress in information technology



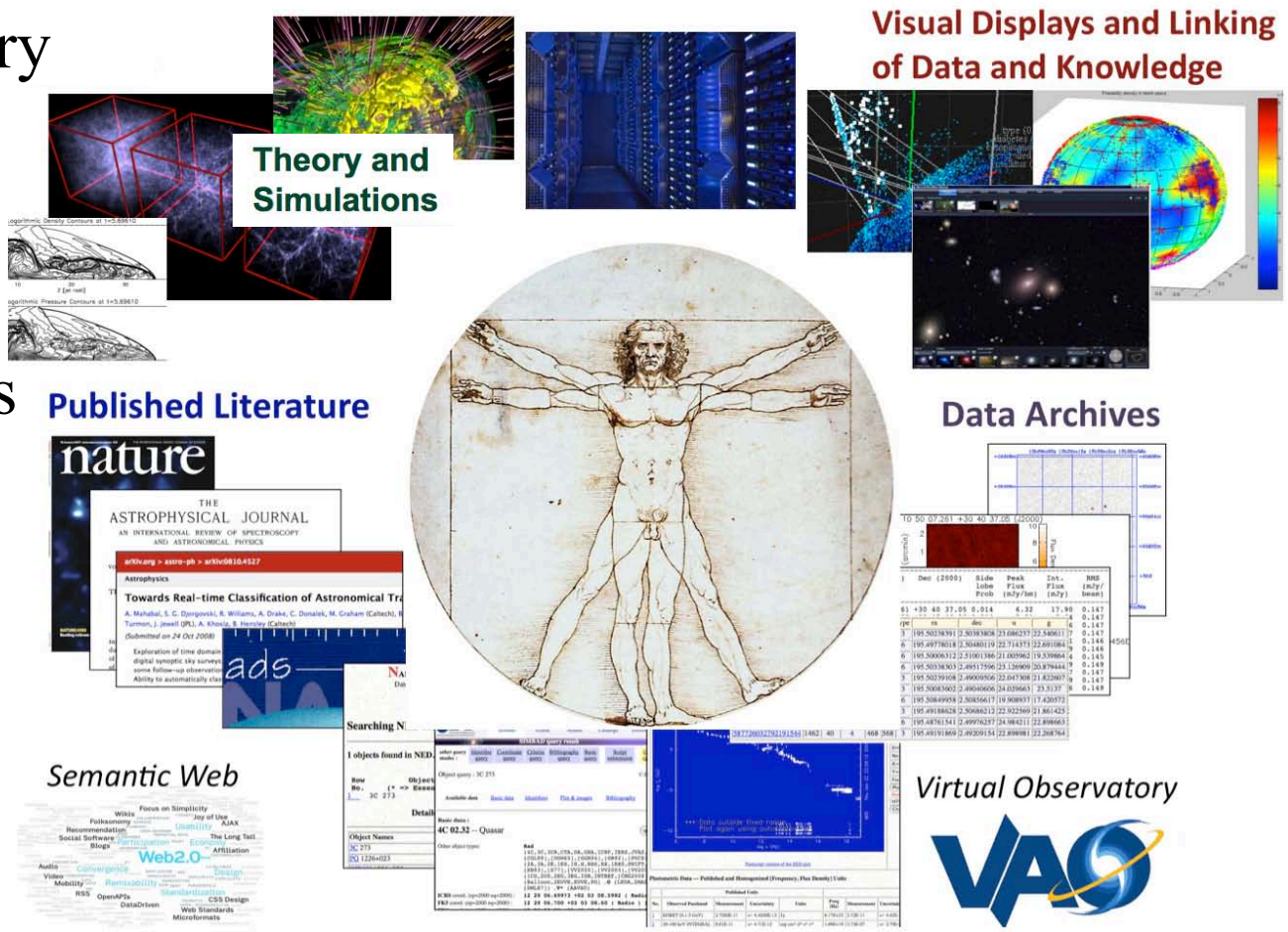


Synoptic sky surveys: from Terascale to Petascale data streams

Telescope+instrument are "just" a front end to data systems, where the real action is

What is Available On-Line:

- Data (various archives, surveys, derived data products…)
- Tools for data exploration (data mining, statistics, visualization)
- Literature (arXiv, e-journals, novel forms of publishing)
- Virtual Observatory framework (and AstroInformatics)
- Interaction and collaboration tools (email to telepresence)
- Education and public outreach



Data Archives

- Space missions (and wavelength-domain specific archives)
- Large digital sky surveys (e.g., SDSS, 2MASS, NVSS, …)
- Observatory archives (still sparse; ESO is probably the best)
- Derived data products services (NED, SIMBAD, etc.)
- Some modest data sets available through e-journals
- Numerical simulation output (e.g., the Millenium run)
- All of these are in principle connected through the Virtual Observatory (VO) framework
- Note:
  - Newly obtained data are usually a subject to proprietary periods, typically 12 – 18 mos.
  - Archival research is sometimes specifically funded by NASA missions, and has a high impact

Some Popular Archives

- Sloan Digital Sky Survey (SDSS); DR9 is the latest: http://skyserver.sdss3.org/dr9/en/
- Multi-mission Archive for Space Telescopes (MAST); includes the HST, GALEX, DSS, various UV missions: http://archive.stsci.edu/
- Infra-Red Science Archive (IRSA); all NASA IR missions (IRAS, Spitzer, WISE, etc.) and 2MASS: http://irsa.ipac.caltech.edu/
- High Energy Astrophysics Science Archive Research Center (HEASARC); all NASA (and some ESA, JAXA) X-ray and γ-ray missions (Chandra, Fermi, Rosat, SWIFT, etc.), other: http://heasarc.gsfc.nasa.gov/
- Canadian Astronomy Data Centre (CADC): various ground and space-based: http://www3.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/cadc/

Literature and High-Level Data Products

- NASA/IPAC Extragalactic Database (NED); major catalogs and published data for anything extragalactic: http://ned.ipac.caltech.edu/
- Centre de Donées astronomiques de Strasbourg (CDS); major catalogs, published data, SIMBAD, etc.: http://cdsweb.u-strasbg.fr/
- The SAO/NASA Astrophysics Data System (ADS); all of the astronomical literature: http://www.adsabs.harvard.edu/
- And of course http://arxiv.org/

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 …

- We need to create ***a new scientific methodology*** on the basis of applied CS and IT
- Important for practical applications beyond science

The Virtual Observatory Concept

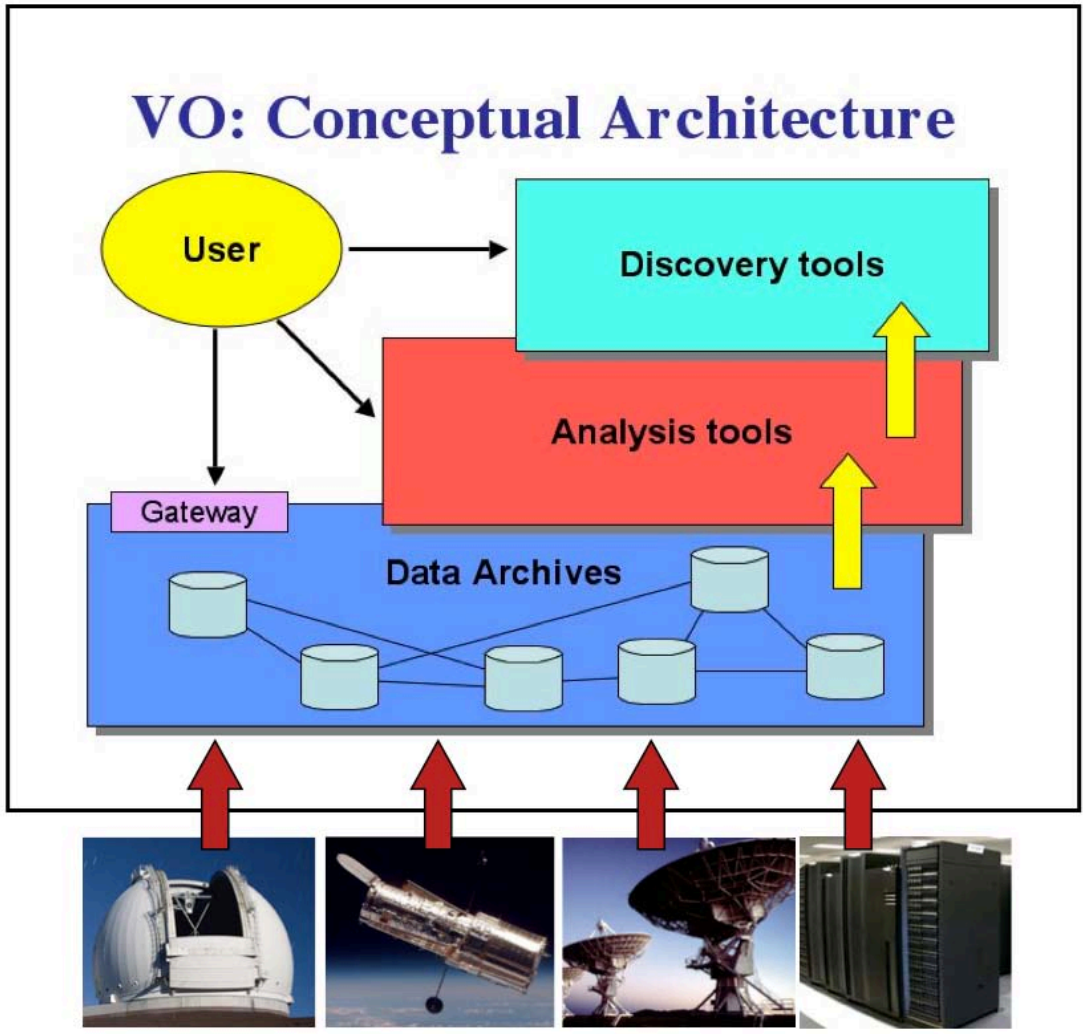
- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***

– Provide and federate content (data, metadata) services, standards, and analysis/compute services

– Develop and provide data exploration and discovery tools

– Harness the IT revolution in the service of astronomy

– A part of the broader e-Science /Cyber-Infrastructure



Virtual Observatory Is Real!

VIRTUAL ASTRONOMICAL OBSERVATORY

http://us-vo.org

Discover, retrieve, and analyze astronomical data from archives and data centers around the world.

http:// ivoa.net

http://www.euro-vo.org

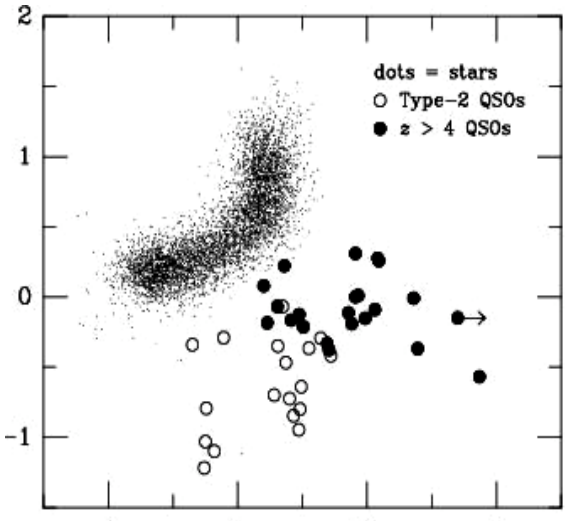
Virtual Observatory Portals

- The U.S. Virtual Astronomical Observatory (VAO): http://www.usvao.org/
- The European Virtual Observatory (EuroVO): http://www.euro-vo.org/pub/
- International Virtual Observatory Alliance (IVOA): http://ivoa.net/

Virtual Observatory Science Examples

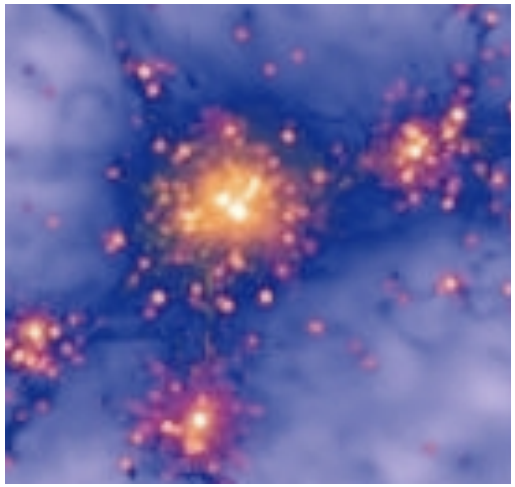
Combine the data from multi-TB, billion-object surveys in the optical, IR, radio, X-ray, etc.

- Precision large scale structure in the universe
- Precision structure of our Galaxy



Discover rare and unusual (one-in-a-million or one-in-a-billion) types of sources

- E.g., extremely distant or unusual quasars, new types, etc.



Match Peta-scale numerical simulations of star or galaxy formation with equally large and complex observations

... etc., etc.

VO Functionality Today

What we did so far:

- Lots of progress on interoperability, standards, etc.
- An incipient ***data grid of astronomy***
- Some useful web services
- Community training, EPO

What we did not do (yet):

- Significant data exploration and mining tools

That is where the science will come from!

Thus, little VO-enabled science so far

Thus, a slow community buy-in

➔ Development of powerful knowledge discovery tools should be a key priority

What Do You Need To Know

- What are databases, how to access them, and probably the Structured Query Language (SQL)
  - More powerful than most “canned” user interfaces
- Finding stuff on the Web!
- Programming in at least one or two modern languages, e.g., Python, Java; also C and/or Fortran (legacy codes)
  - But use the available packages as much as you can, do not try to reinvent the wheel!
- Data mining basics
  - Ditto; see the links on the website
- Statistics: the more the better, and certainly some Bayesian
  - Ditto; see the links on the website

From Data to Knowledge

- Data are incidental to our real goal, knowledge discovery
- This is not so easy due to:
  - ✓ Exponentially growing data volumes (TB, PB, EB …)
  - ✓ Increasing data complexity, e.g., dimensionality
- You need some data mining and knowledge discovery skills
  - ✓ Check out our "Methods of computational science" class:

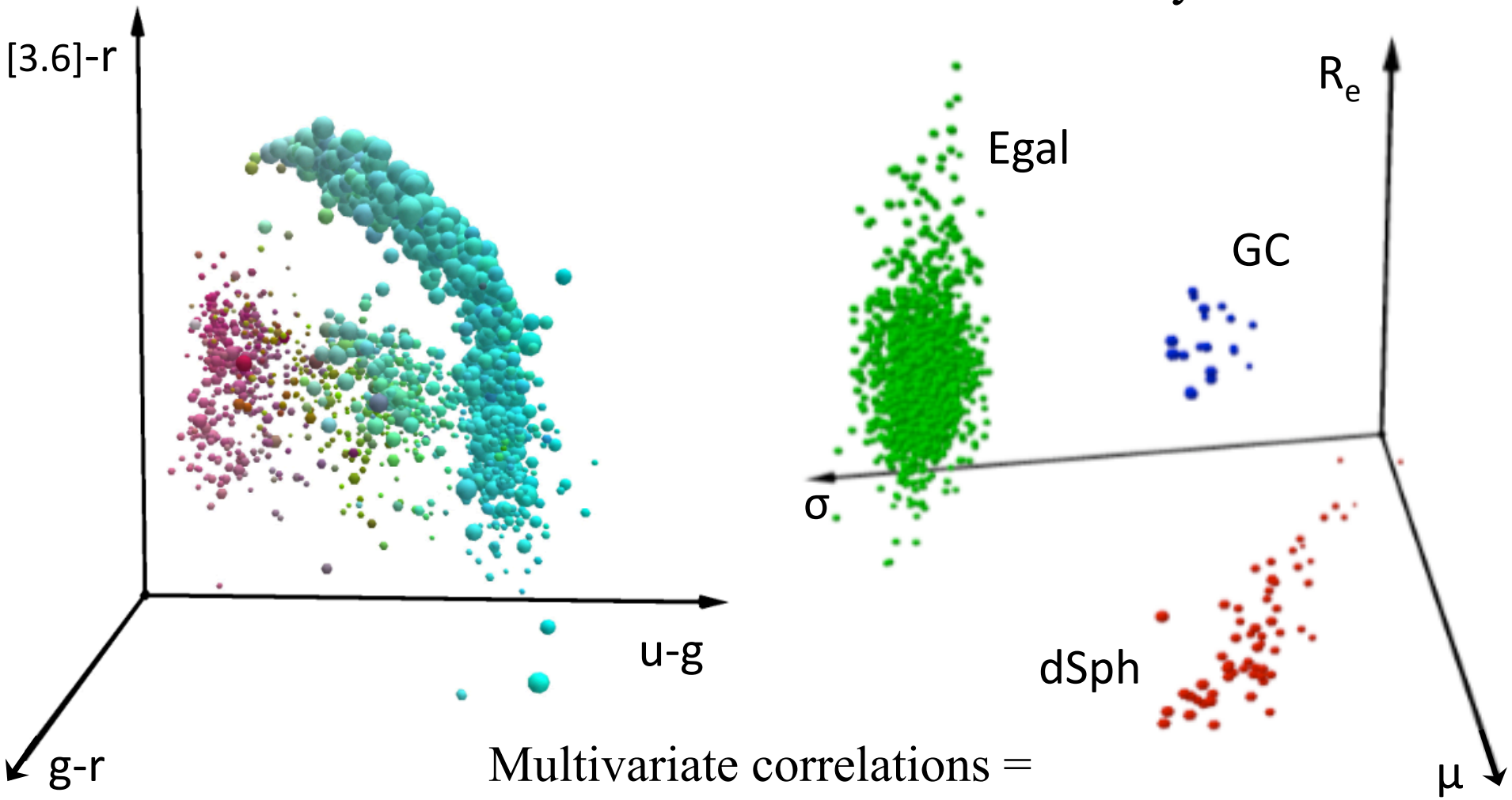
    Ay 119: https://hesperus.caltech.edu/wiki/projects/ay119/Ay_119.html

    Ay/Bi 199: http://www.astro.caltech.edu/~george/aybi199/
- Some challenges actually require a new applied CS research

Clustering in Parameter Spaces

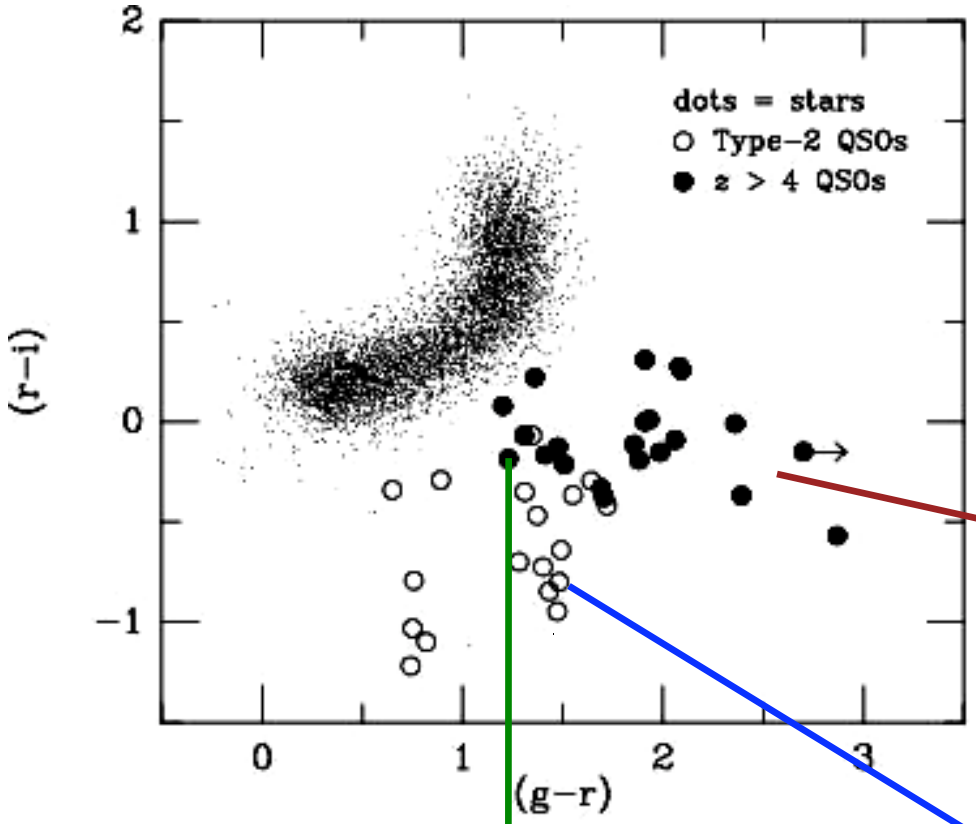
Colors of stars and
active galactic nuclei

Families of dynamically
hot stellar systems



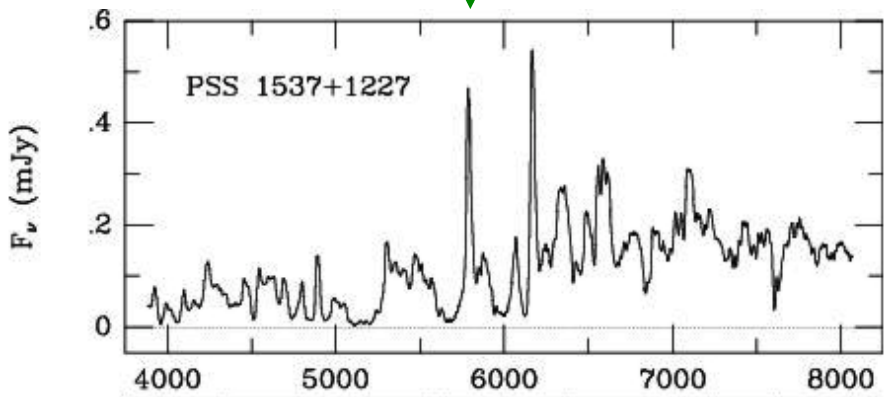
Multivariate correlations =
clusters with a reduction of the statistical dimensionality

Quasar Selection in Color Parameter Space

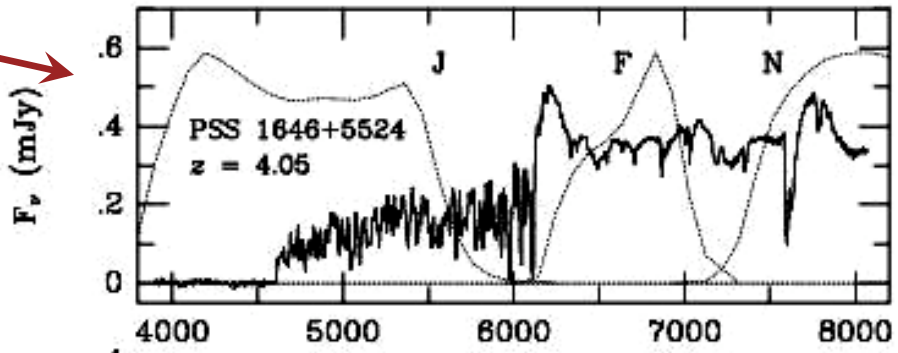


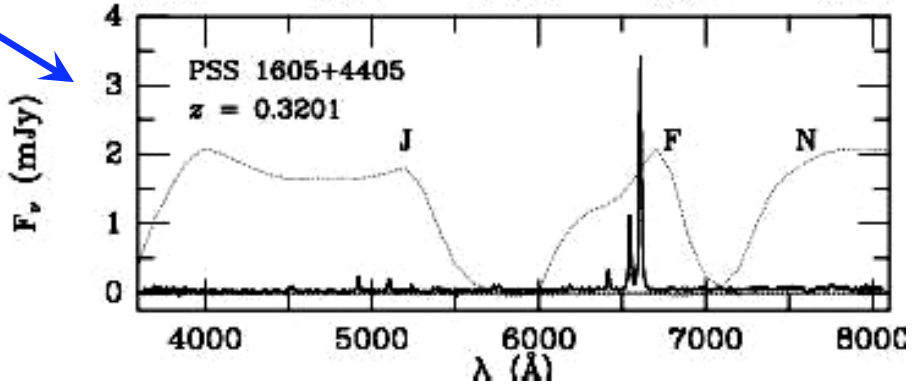
Some outliers belong to the known, but rare species, but some may be new discoveries

Something new



High-redshift QSO





Type-2 QSO

The Key Challenge: Data Complexity

Or: The Curse of Hyper-Dimensionality

1. **Data mining algorithms scale very poorly:**

   N = data vectors, ~ 10^8 - 10^9, D = dimension, ~ 10^2 - 10^3

   – Clustering ~ N log N → N^2, ~ D^2

   – Correlations ~ N log N → N^2, ~ D^k ($k \geq 1$)

   – Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)

2. **Visualization in >> 3 dimensions**

- The complexity of data sets and interesting, meaningful constructs in them is *exceeding the cognitive capacity of the human brain*
- We are biologically limited to perceiving D ~ 3 - 10(?)
- Visualization must be a component of the data mining / exploration process
- It is the bridge between the quantitative content of data and human understanding