AMPATH2002, Valdivia, April 2002

Virtual Observatories

Alex Szalay

Department of Physics and Astronomy

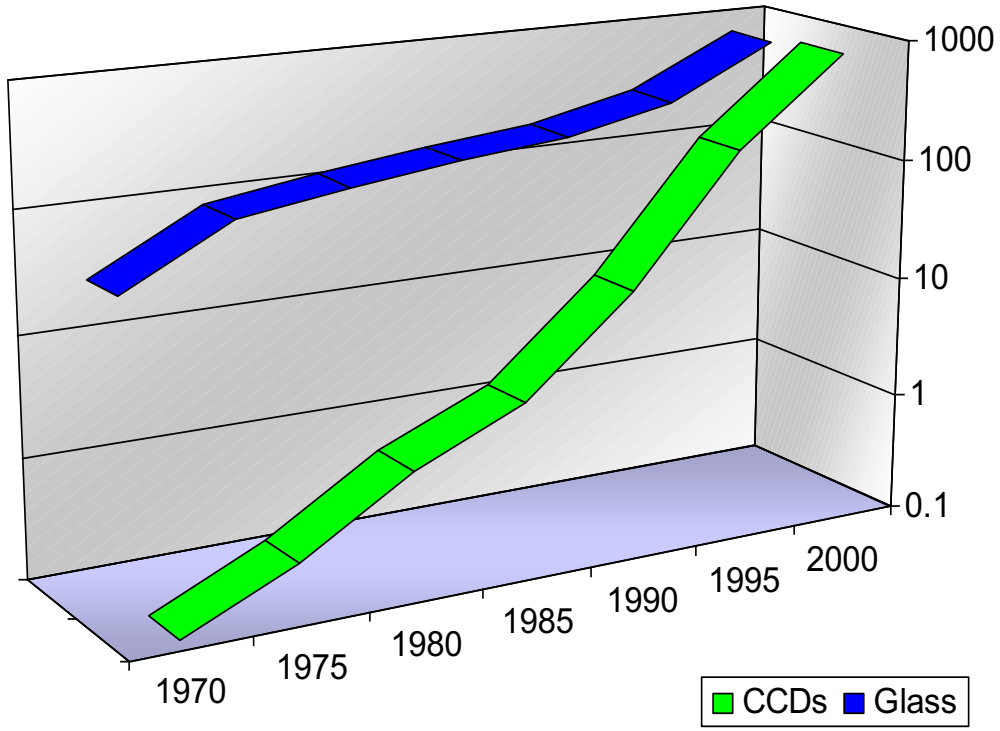
The Johns Hopkins University

Nature of Astronomical Data

- Theory, Observation, Simulations and Data Exploration
- Imaging
  - *2D map of the sky at multiple wavelengths*
- Derived catalogs
  - *subsequent processing of images*
  - *extracting object parameters (400+ per object)*
- Spectroscopic follow-up
  - *spectra: more detailed object properties*
  - *clues to physical state and formation history*
  - *lead to distances: 3D maps*
- Numerical simulations
- All inter-related!

Trends

Future dominated by detector improvements



- Moore's Law growth in CCD capabilities
- Gigapixel arrays on the horizon
- Improvements in computing and storage will track 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.

The Age of Mega-Surveys

- The next generation mega-surveys and archives will change astronomy, due to
  - *top-down design*
  - *large sky coverage*
  - *sound statistical plans*
  - *well controlled systematics*
- The technology to store and access the data is here

  we are riding Moore's law
- Data mining will lead to stunning new discoveries
- Integrating these archives is for the whole community

=> Virtual Observatory

Ongoing surveys

- Large number of new surveys
  - *multi-TB in size, 100 million objects or more*
  - *individual archives planned, or under way*
- Multi-wavelength view of the sky
  - *more than 13 wavelength coverage in 5 years*
- Impressive early discoveries
  - *finding exotic objects by unusual colors*
    - L,T dwarfs, high-z quasars
  - *finding objects by time variability*
    - gravitational microlensing
- Over 50 datasets with 100 TB today, doubling every year

MACHO
2MASS
DENIS
SDSS
GALEX
FIRST
DPOSS
GSC-II
COBE
MAP
NVSS
FIRST
ROSAT
OGLE
...

VO- The challenges

- Size of the archived data

  40,000 square degrees is 2 Trillion pixels

  – *One band* — *4 Terabytes*

  – *Multi-wavelength* — *10-100 Terabytes*

  – *Time dimension* — *10 Petabytes*
- Data sets extremely diverse
  – *New metadata standards needed*
- Current techniques inadequate
  – *new archival methods, tools*
- Hardware/networking requirements
  – *scalable solutions required*
- Transition to the new astronomy (sociological issues)

New Astronomy- Different!

- Data “Avalanche”
  - *the flood of Terabytes of data is already happening, whether we like it or not*
  - *our present techniques of handling these data do not scale well with data volume*
- Systematic data exploration
  - *will have a central role*
  - *statistical analysis of the “typical” objects*
  - *automated search for the “rare” events*
- Digital archives of the sky (example:SDSS)
  - *will be the main access to data*
  - *hundreds to thousands of queries per day*

Distributed Archives

- Astronomy data will never be centralized
  - *Scattered over the world, doubling every year*
- It should be easy to add new data sets
  - *Templates for archives, services, auto-discovery*
  - *Even static data reaches over many wavelengths*
  - *Currently over 10 different bands, soon 20+*
  - *Time domain experiments add more complexity*
  - *LSST : 4PB/yr by 2008, 10PB/yr by 2012*
  - *May need to use triggers on the detectors*
- Distributed cross-correlations over the system
  - *Needs to be fast, automated, dynamic*
  - *Lookups at object level, dynamic assemblies needed*

Relation to the HEP Problem

- Similarities
  - *need to handle large amounts of data*
  - *data is located at multiple sites*
  - *data should be highly clustered*
  - *substantial amounts of custom reprocessing*
  - *need for a hierarchical organization of resources*
  - *scalable solutions required*
- Differences of Astro from HEP
  - *data migration is in opposite direction*
  - *the role of small queries is more important*
  - *relations between separate data sets (same sky)*
  - *data size currently smaller, we can keep it all on disk*

The Virtual Observatory Effort

- Several coordinated efforts world-wide
- NVO (US), AVO(Europe), AstroGrid (UK)
- Agreements reached on common interoperability standards
- First activities involve metadata standards
  - *UCD: Universal Content Descriptor*
  - *VOTable: Compact XML representation of astro data*
- Prototype science scenarios under way
  - *Expect first public demos in Jan 2003*
- Will need dynamic object assembly services
- Close collaboration with iVDGL

Conclusions

www.voforum.org **www.us-vo.org**

- Databases became an essential part of astronomy: most data access will soon be via digital archives
- Data at separate locations, distributed worldwide, evolving in time: move queries not data!
- Computations in both processing and analysis will be substantial: need to create a `Virtual Data Grid'
- Problems similar to HEP, lot of commonalities, but data flow more complex, need lots of object services
- Interoperability of archives is essential: the Virtual Observatory is inevitable
- But: sociological challenges formidable