Cyber-Infrastructure for Astronomy

S. George Djorgovski

Professor of Astronomy, Caltech
PI, Digital Palomar Observatory Sky Survey
Chairman, National Virtual Observatory Science Definition Team

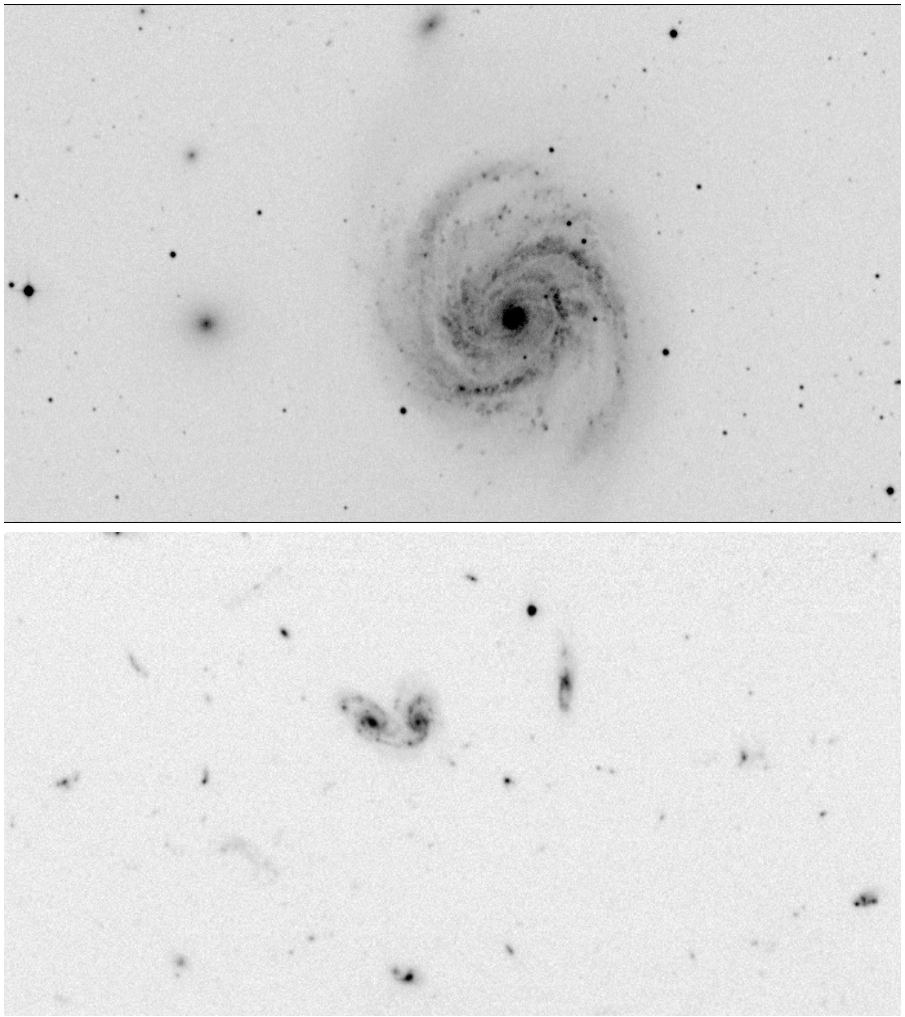
NSF Advisory Committee on Cyber-Infrastructure
November 30, 2001

http://www.astro.caltech.edu/~george/nsfcyber/

Information-Rich Astronomy for the 21st Century

- Like in most sciences, **the amount and complexity of data in astronomy is increasing exponentially**, roughly following the Moore's law
- Increasingly, most data are taken through **large, digital sky surveys** (multi-TB, soon multi-PB), over the entire available electromagnetic spectrum
- Current holdings in "organized" archives are **> 100 TB**, with much more data distributed and stored in a disorganized fashion
- This is **changing the way observational astronomy is done**: a pure survey science, an optimal target selection for large telescopes and space observatories, etc.
- A similar situation exists in the astrophysics theory/simulations

Digital Sky Survey Samples



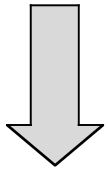
1 microSky
(DPOSS)

1 nanoSky
(HDF-S)

The New Astronomy, and the Old Astronomy Done Better

- The quantitative explosion in data volume will produce **a qualitative change in the way astronomy is done**. It will enable, reinvent, or even open new fields of inquiry
- There is now a lot of survey-based astronomy from individual massive data sets, and archival research from pointed observations (e.g., with the HST and other NASA missions)
- We also have a good system of digital libraries / depositories of data products, publications, etc. (e.g., NSSDC, NED, ADS, Simbad, etc.)
- What is coming is **multi-survey-based science**, where a combination of complementary data (e.g., optical, radio, infrared, x-ray, etc.) is providing an added value

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**

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 Will a Virtual Observatory Do?

- The VO represents both the core information (cyber?) **infrastructure** for the new astronomy, and a general **research environment** in the era of information abundance
- The VO will be **technology-enabled**, but **science-driven.** It poses both technical and scientific/methodological challenges
- Types of VO-based astronomy may include:
  - **Statistical astronomy done right** (e.g., precision cosmology or Galactic structure, with large numbers of sources making the Poissonian errors unimportant).
  - **Exploration of the new domains of the observable parameter space** (e.g., the time-variable universe, the low surface brightness universe, etc.)
  - Searches for **rare** or **new, previously unknown types of objects** or phenomena (e.g., brown dwarfs, high-redshift quasars, … ??)

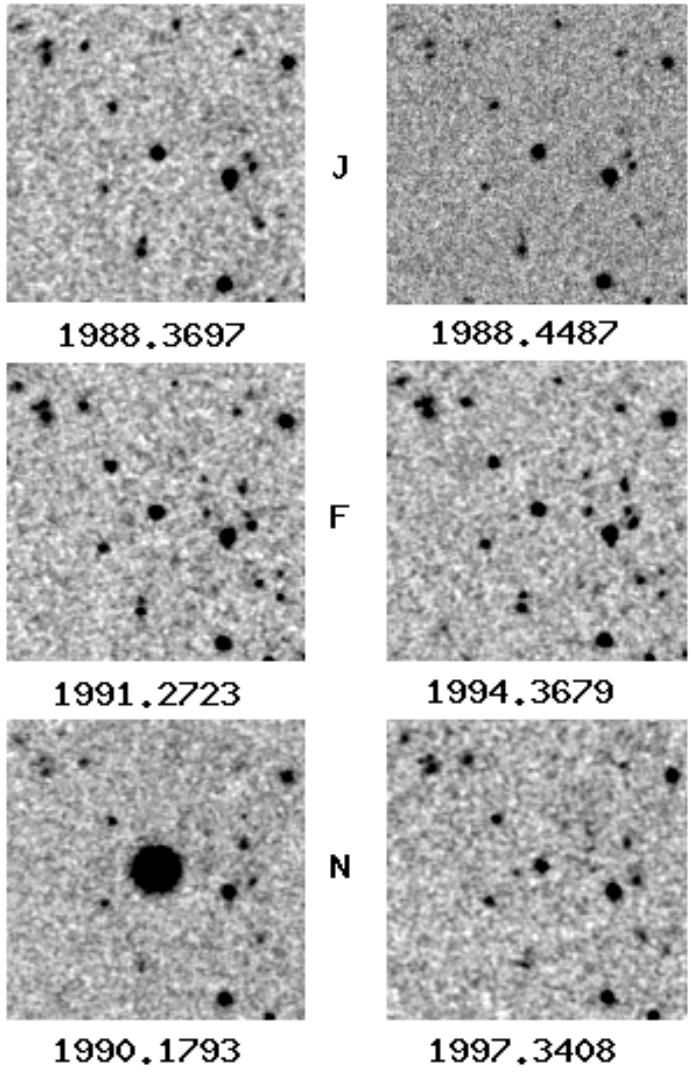
Exploration of New Domains of the Observable Parameter Space

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.



Technological/Infrastructure 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!

The Ongoing VO Development Efforts

- The **NSF** ITR large project: Building the Framework for the NVO (plus a few smaller projects)
- **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
- Promising possibilities: NSF DTF, other …?
- Many vigorous efforts in Europe (AVO, AstroGrid, CDS, AstroVirTel …), Canada (CADC), and elsewhere

Astronomy and Other Fields

- These challenges are common to most data-intensive sciences today, or indeed many fields or endeavor (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, markets…)
  - Data sets are freely available; the methodology is open to all

Towards a New Scientific Methodology

- Are we making the full use of the available wealth of information? (Recall the rapid exponential growth in the information volume.) -- ***No***
- Are we wasting the scientific opportunities offered by the data explosion and the information technology revolution? -- ***Yes***
- Where is the bottleneck?
  - Old *data handling and analysis methodologies*, developed for dealing with data sets many orders of magnitude smaller and simpler, are no longer adequate
  - An improved *information infrastructure* (computing, networks, software) is needed
  - We are not used to think in terms of problems and questions which can be addressed effectively with the wealth of data and the necessary (but available!) information technologies. *A failure of our imagination and creativity?*

The Path Forward

- Technological/Infrastructure Issues:
  - If you build it, they will come (?)
  - The right balance of the centralized vs. distributed assets
  - Data access and data analysis and understanding are the key issues
  - What are the good organizational/structural models? e.g., WWW
- Methodological/Scientific Issues:
  - Having the right tools (DM/KDD, statistics, visualization…)
  - Learning to ask new kinds of questions
  - Educating the new generation of IT-savvy scientists
- Political/Sociological Issues:
  - Data rights, proprietary periods, intellectual rights …
  - Giving the credit and the recognition where it is due

What Can/Should the Federal Agencies Do?

- Provide the tools for the new, information-rich science
  - Hardware and network infrastructure (computing and storage)
  - Data exploration tools (DM/KDD, applied statistics, visualization)
- Help foster interdisciplinary partnerships, collaborations
  - More focus on scientific discipline-driven projects, less on IT?
  - Conferences, fellowships, …
- Facilitate partnerships between academia, industry, gov't agencies, and private foundations
- Help generate mechanisms for giving the proper credit and recognition to people working in this area
- A smoother and more extensive inter-agency cooperation

A Broader Significance and the Benefits to the Nation

- These problems and issues are **common to most fields** of the modern science and economy, and the society in general, including national security
- The methods, tools, services, and perhaps the infrastructure itself will likely find many uses and applications outside their initial scientific context
- 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

Some General Comments and Musings

- 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 and 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?
- Applied CS and IT are becoming the "new mathematics", serving all sciences, providing the tools and the frameworks