The National Virtual Observatory Science Definition Team Report

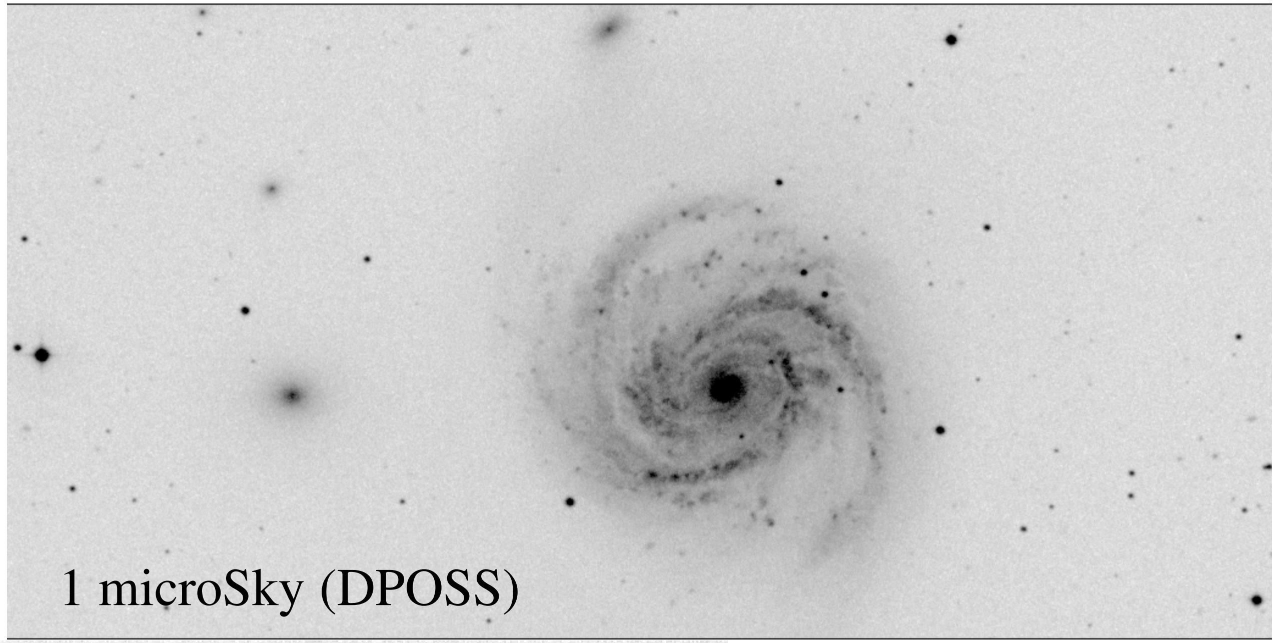
http://nvosdt.org

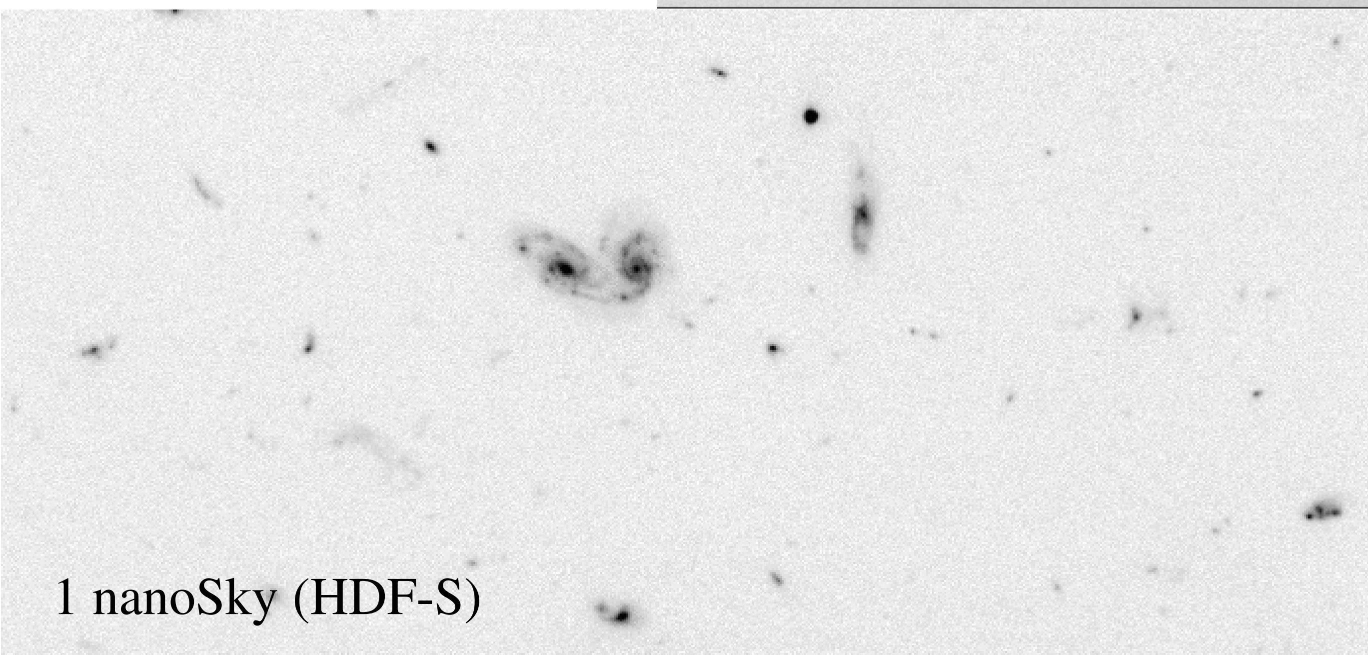
AAS Meeting, June 6, 2002

- Creation of the **National Virtual Observatory** (NVO): the top priority of the NAS Decadal Survey in the “small” (< $ 100 M) category
- In response, the NSF and NASA formed the NVO **Science Definition Team** (SDT) to:
  - Refine and formulate a joint NVO initiative, the scientific goals, and the technical requirements
  - Gather the input from the community and serve as a liaison to the space science, CS/IT, and statistics communities, and international VO efforts
  - Provide recommendations for proceeding
- The SDT delivered its report to the agencies on April 11. It is available at **http://nvosdt.org**

Astronomy is Facing a Major Data Avalanche:

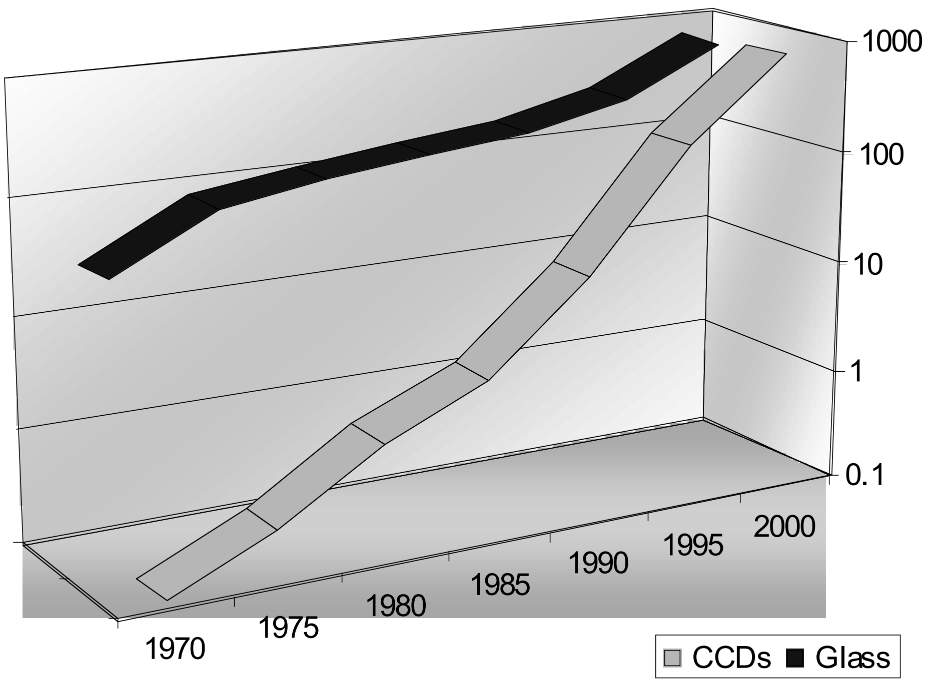
Multi-Terabyte (soon: multi-Petabyte) sky surveys and archives over a broad range of wavelengths …


1 microSky (DPOSS)


1 nanoSky (HDF-S)

Billions of detected sources, hundreds of measured attributes per source …

The Exponential Growth of Information in Astronomy



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.

- Computing technology drives the data volume and quality (through detectors and other hardware)
- Both data volume *and Complexity* are increasing
- Efficient data utilization requires novel information technology (IT)
- Harnessing the modern IT can drive a new revolution in astronomy

The Changing Face of Observational Astronomy

- **Large digital sky surveys** are becoming the dominant source of data in astronomy: > 100 TB, growing rapidly
  - Current examples: SDSS, 2MASS, DPOSS, GSC, FIRST, NVSS, RASS, IRAS; CMBR experiments; Microlensing experiments; NEAT, LONEOS, and other searches for Solar system objects …
  - Digital libraries: ADS, astro-ph, NED, CDS, NSSDC …
  - Observatory archives: HST, CXO, space and ground-based …
  - Future: QUEST2, LSST, and other synoptic surveys; GALEX, SIRTF, astrometric missions, GW detectors …
- **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 GB, Library of Congress ~ 20 TB

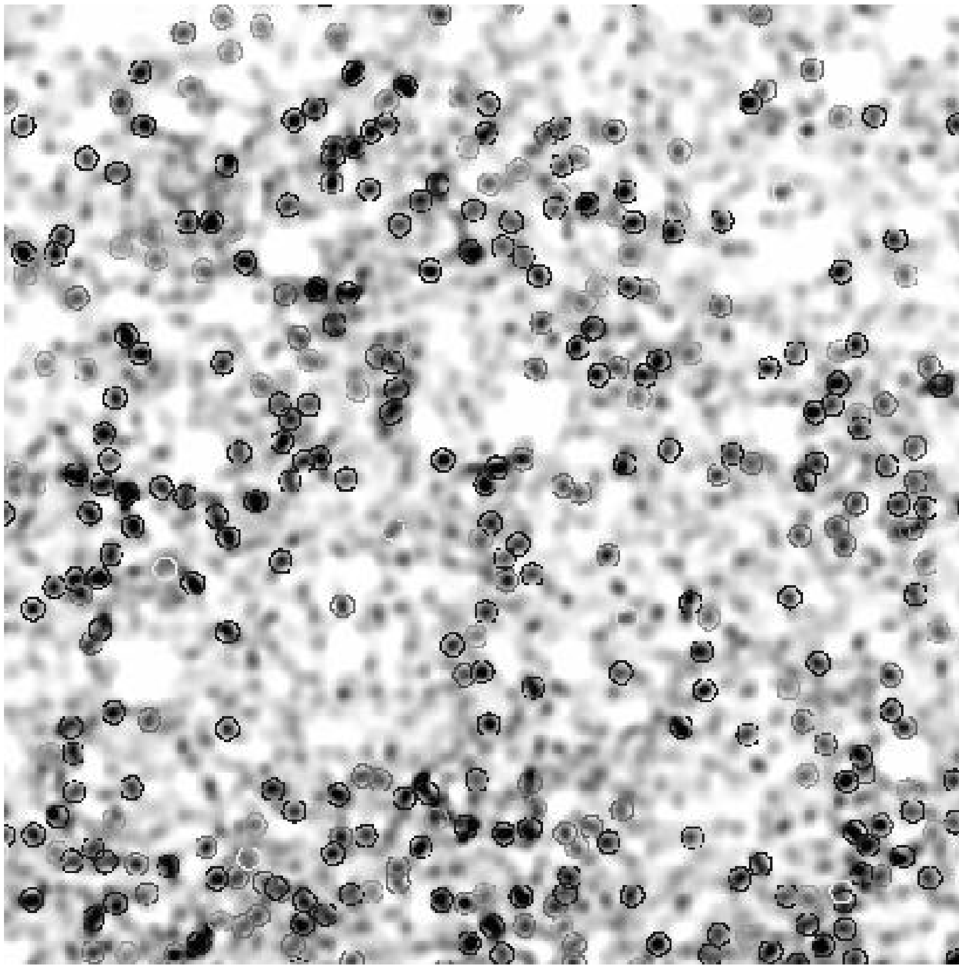
This quantitative change in the information volume and complexity will enable the

Science of a Qualitatively Different Nature:

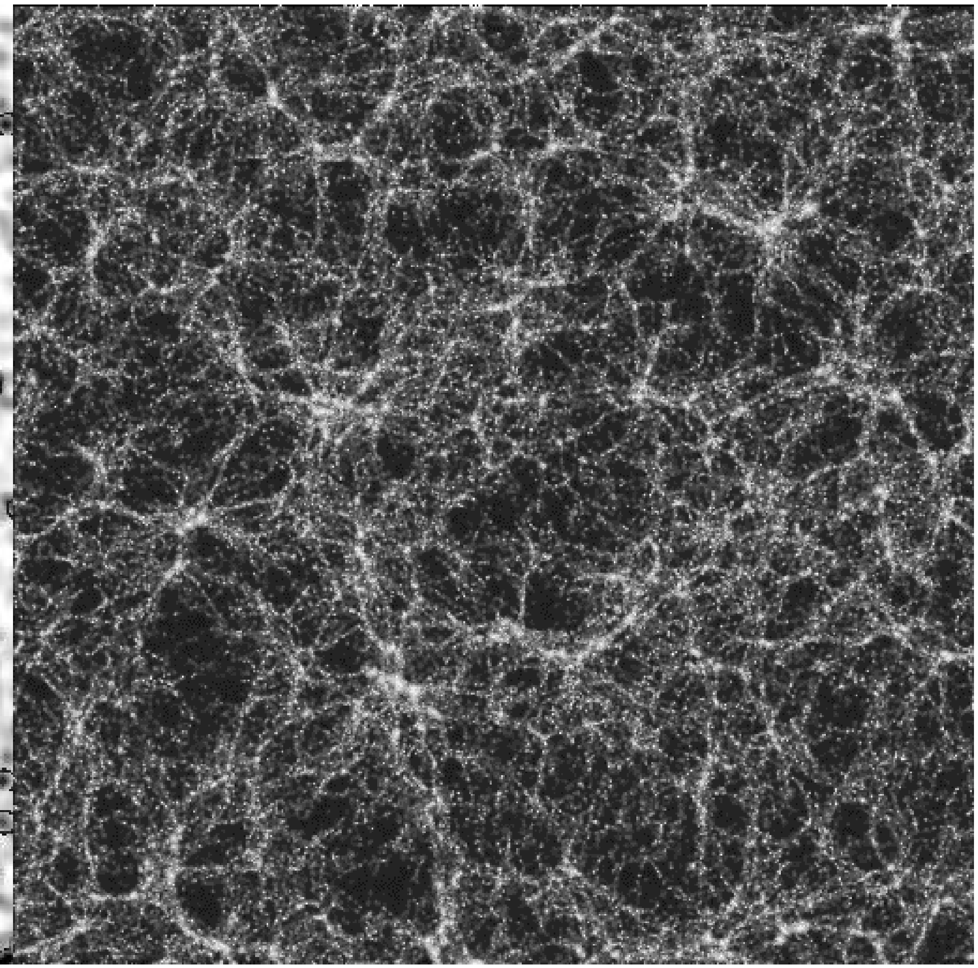
- **Statistical astronomy done right**
  - Precision cosmology, Galactic structure, stellar astrophysics …
  - Discovery of significant patterns and multivariate correlations
  - Poissonian errors unimportant
- **Systematic exploration of the observable parameter spaces**
  - Searches for rare or unknown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

Precision Cosmology …

… and a better marriage of theory and observations

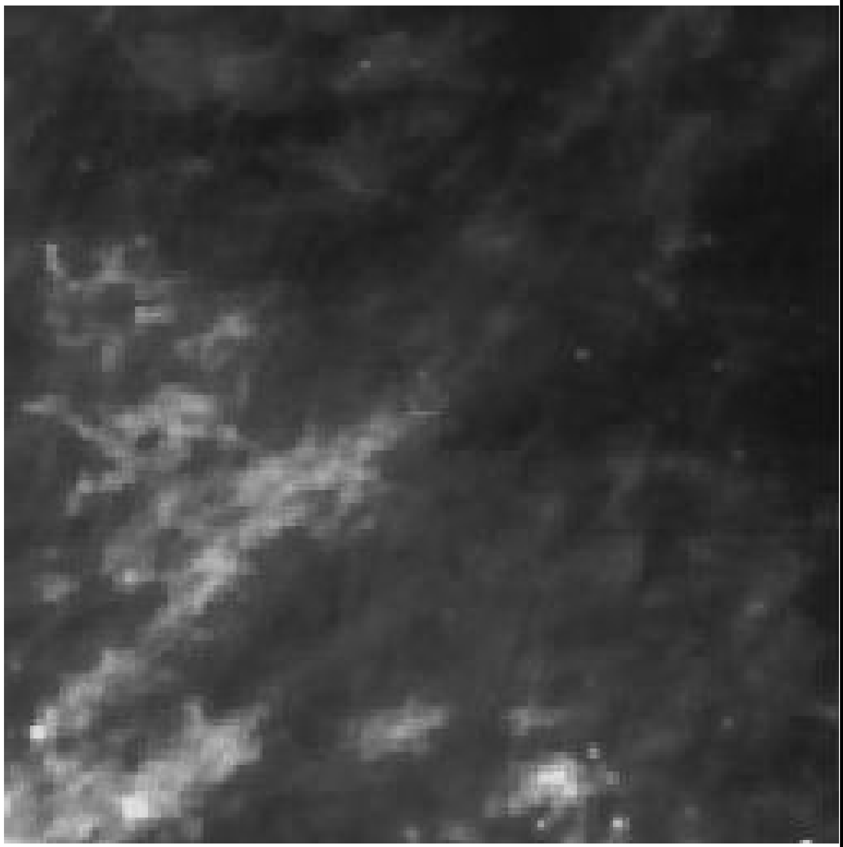


DPOSS Clusters

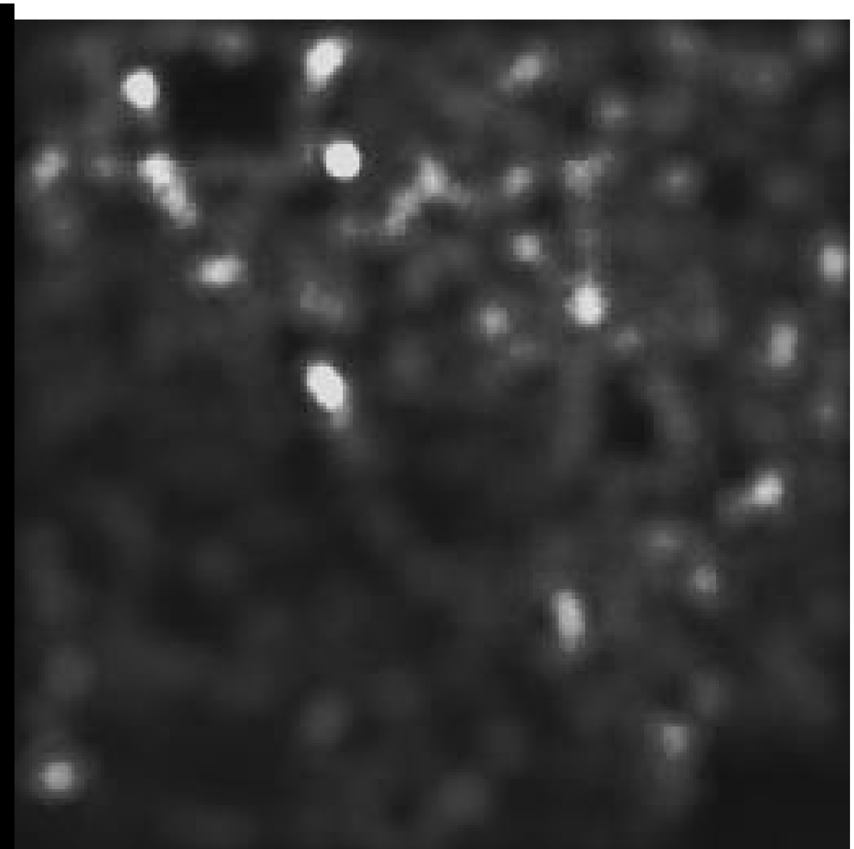


LSS Numerical Simulation

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



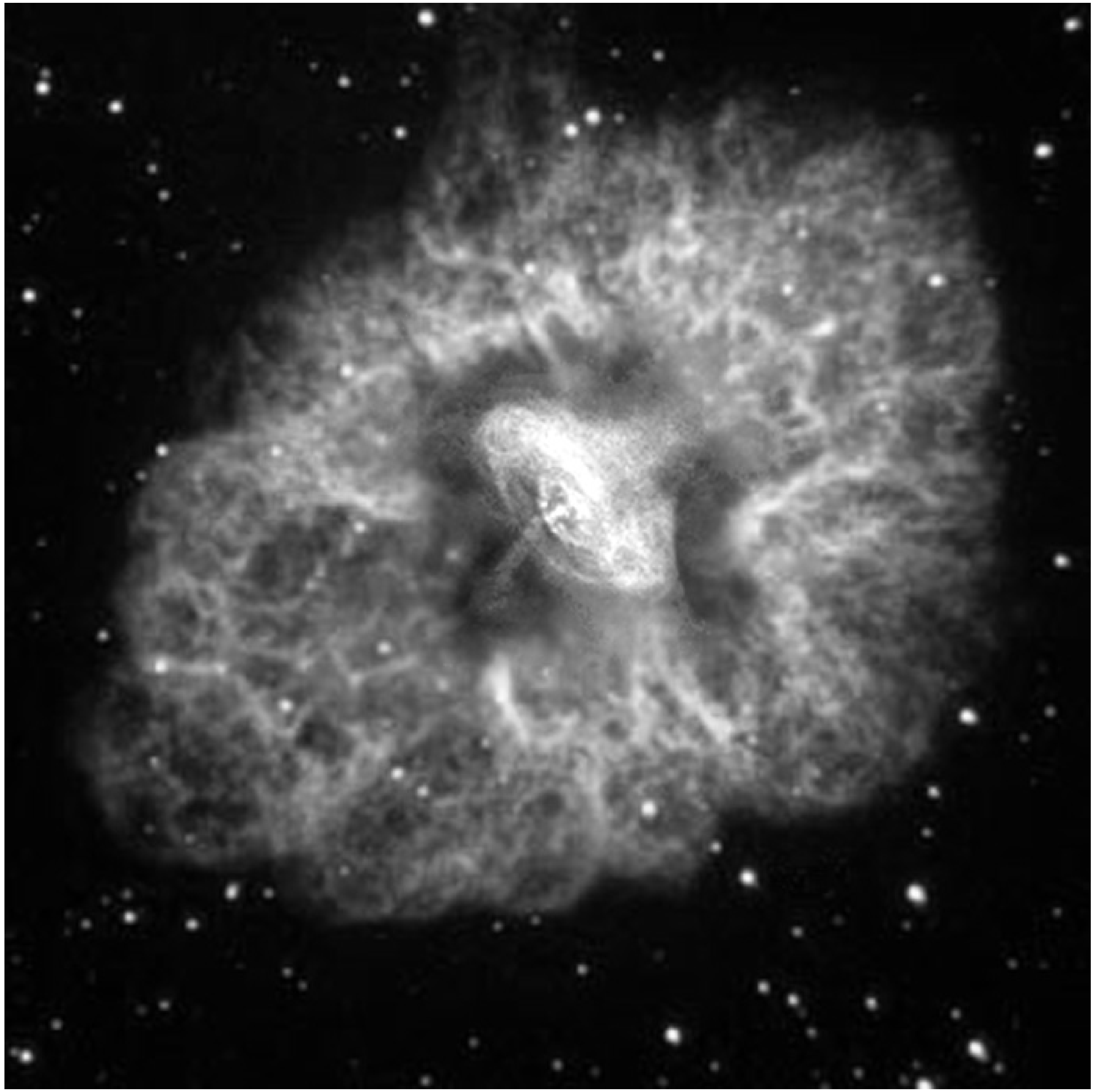
Infrared emission from interstellar dust



Smoothed galaxy density map

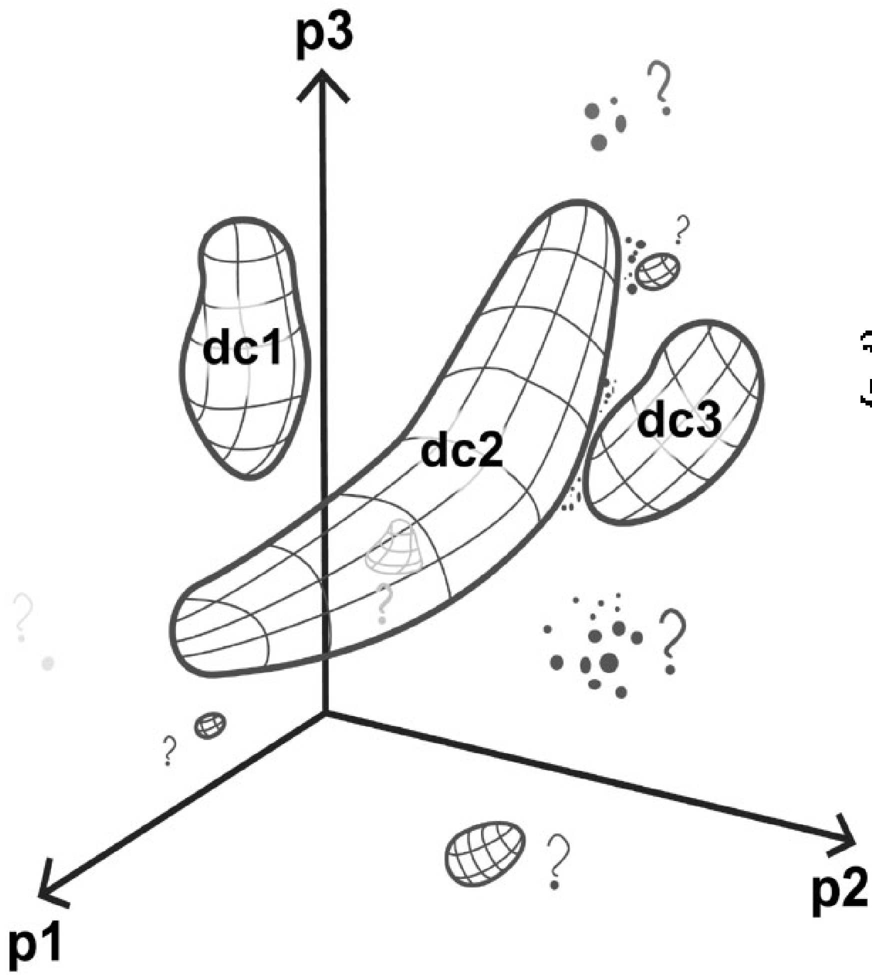
A **panchromatic** approach to the universe reveals **a more complete physical picture**

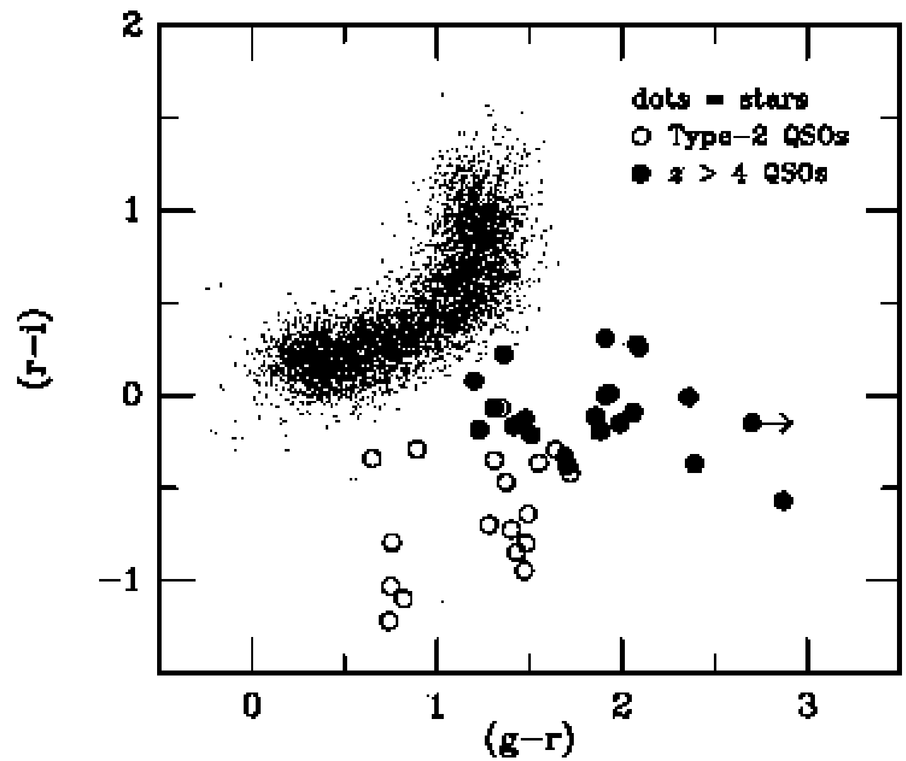
The resulting complexity of data translates into increased demands for data analysis, visualization, and understanding



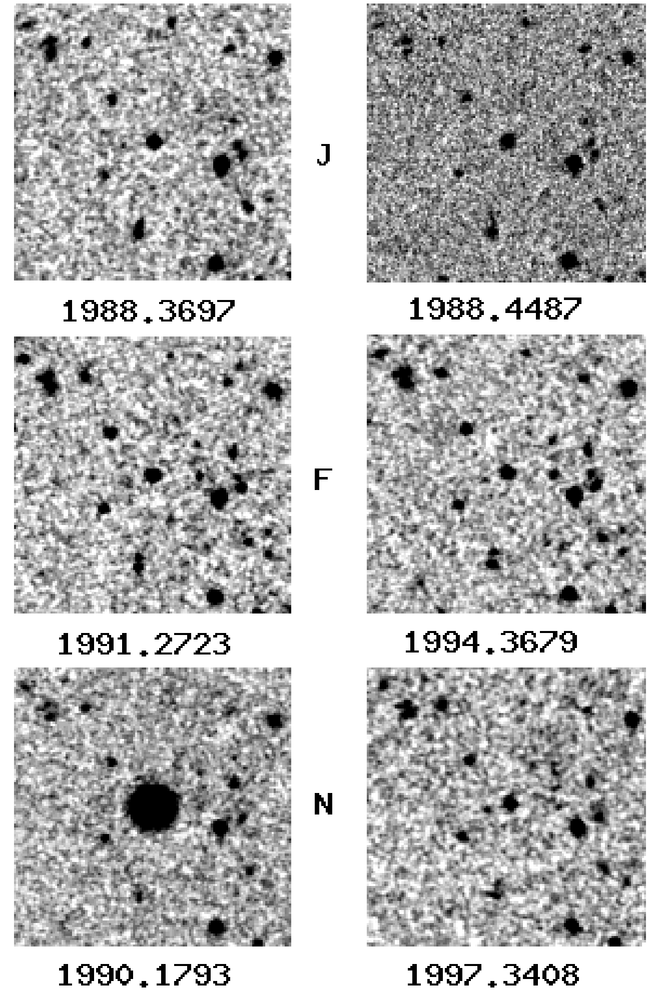
Exploration of observable parameter spaces and <u>searches for rare or new types of objects</u>

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



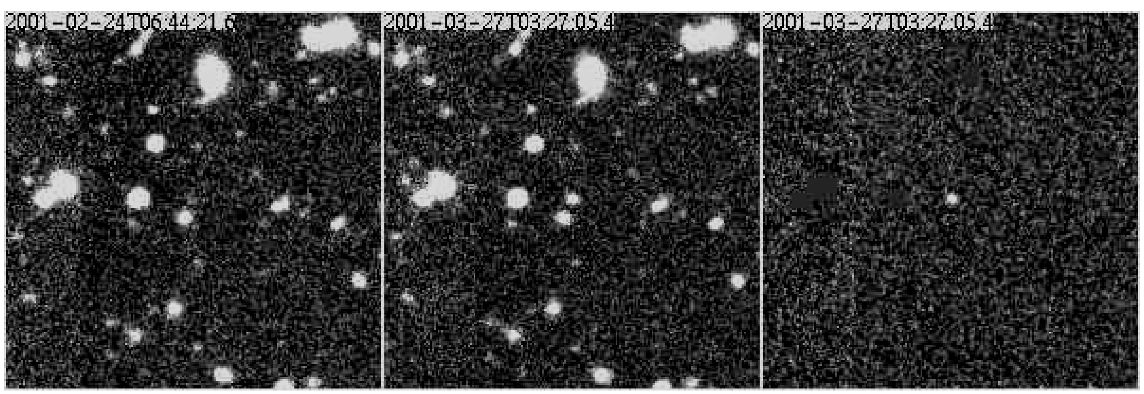


Exploration of new domains of the observable parameter space: **the Time Domain**



Megaflares on normal MS stars (DPOSS)

Faint, Fast Transients (Tyson et al.)



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*

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
- Federate the existing and forthcoming large digital sky surveys and archives, and provide the tools for their scientific exploitation
- A complete, dynamical, interactive, web-based **research environment for the new astronomy with massive data sets**
- *Technology-enabled, but science-driven*

What We Have Now:

- Many separate archives, passively serving on demand small amounts of limited data
- Almost no data discovery capabilities
- No standards for metadata, data exchange protocols, formats
- No general tools or services for data fusion, data mining, analysis, and visualization

What We Need:

- **A new, distributed research environment which will remedy all of these deficiencies**, and in that process:
  - Harness the power of the modern information technology
  - Enable the complete, efficient scientific exploitation of the massive data sets in astronomy
  - Organized transition to the new, data-rich astronomy

The Roles of the NVO: Scientific

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

The Roles of the NVO: Empowering

- **Professional:**
  - Scientists and students anywhere with an internet connection would be able to do a first-rate science ⇒ A broadening of the talent pool in astronomy, democratization of the field
- **Science Strategy and Planning:**
  - Where are the gaps in our coverage of the observable parameter space? Which new missions, instruments, experiments are likely to have the largest impact?
- **Education and Public Outreach:**
  - Unprecedented opportunities in terms of the content, broad geographical and societal range, for all educational levels
  - Engaging and broadening the amateur astronomy community
  - Astronomy as a magnet for CS/IT education
  - Creating a new generation of science and technology leaders

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 $\Longrightarrow$ **Terascale** computing!

Data-Rich Astronomy and Other Fields

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

High-energy physics $\rightarrow$ **WWW !**

High-energy physics, Astronomy (NVO) $\rightarrow$ **The Grid**

Astronomy (NVO) $\rightarrow$ **?**

The NVO-Enabled, 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 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

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

The NVO Implementation: Organizational Issues

- The NVO has to be:
  - **Distributed:** the expertise and the data are broadly spread across the country
  - **Evolutionary:** responding to the changing scientific needs and the changes in the enabling technologies
  - **Responsive** to the needs and constraints of all of its constituents
- The NVO has to communicate/coordinate with:
  - The funding agencies
  - The astronomical community as a whole
  - The existing data centers, archives, etc.
  - The international efforts
  - Other disciplines, especially CS/IT

The NVO Implementation: General Issues

- The NVO ***transcends the traditional boundaries*** between different wavelength regimes, agency domains (e.g., space vs. ground based), etc. (An ideal opportunity for inter-agency cooperation in the spirit of the COMRAA report)
- It rests on the foundations of highly successful data centers and information services, but goes beyond them
- The NVO has to be ***science-driven***, and some early investments in the NVO-based science will produce great returns and feedback for its development
- The **EPO** should be treated as an *essential part* of the NVO from the start, not just as an add-on

The NVO Organization and Management

- The NVO is **not** yet another data center, archive, mission, or a traditional project $\Rightarrow$ ***It does not fit into any of the usual structures today***
  - It has an unusually broad range of constituents and interfaces, and is inherently distributed
  - It requires a good inter-agency cooperation, and a long-term stability of structure and funding
- The NVO represents *a **novel type** of a scientific organization* for the era of information abundance
- Designing the NVO organizational/management structure is thus *a **creative challenge** in itself*

An Outline of the Development Process

- **Phase I:** Conceptual design, organizational structure, refined goals and capabilities, early development work, using programs already in place (CY ‘02-’03)
- **Phase II:** Definition of the organizational/mgmt. structure, detailed implementation plan, increase in capabilities, ground-based archives, NVO science funding (CY’02 - ‘05)
- **Phase III:** Full NVO implementation, international connections, major NVO science programs, routine operations (CY ‘06 onwards)

Sample NVO Development Budget (10 yrs)

NVO Budget Prototype (March 2002)

Item	2002 FTE	2002 $	2003 FTE	2003 $	2004 FTE	2004 $	2005 FTE	2005 $	2006 FTE	2006 $	2007 FTE	2007 $	2008 FTE	2008 $	2009 FTE	2009 $	2010 FTE	2010 $	2011 FTE	2011 $	Totals
Bring ground-based archives on-line																					
o Development	6	$900,000	6	$900,000	3	$450,000	1	$150,000													$2,400,000
o Maintenance					1	$150,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	$2,250,000
o Computer hardware		$200,000		$150,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000	$750,000
o Computer software		$20,000		$20,000		$10,000		$10,000		$10,000		$10,000		$10,000		$10,000		$10,000		$10,000	$120,000
o Operations			1	$150,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	$2,550,000
Subtotal, ground-based archives		$1,120,000		$1,220,000		$960,000		$810,000		$660,000		$660,000		$660,000		$660,000		$660,000		$660,000	$8,070,000
NVO Development																					
o NSF ITR project	13	$2,000,000	18	$2,750,000	12	$1,840,000	12	$1,750,000	11	$1,660,000											$10,000,000
o Follow-on development					4	$600,000	4	$600,000	6	$900,000	6	$900,000	6	$900,000	6	$900,000	6	$900,000	6	$900,000	$6,600,000
Subtotal, development		$2,000,000		$2,750,000		$2,440,000		$2,350,000		$2,560,000		$900,000		$900,000		$900,000		$900,000		$900,000	$16,600,000
Software and systems maintenance							4	$600,000	4	$600,000	8	$1,200,000	8	$1,200,000	8	$1,200,000	8	$1,200,000	8	$1,200,000	$7,200,000
Operations																					
o Support staff					2	$180,000	4	$360,000	6	$540,000	6	$540,000	6	$540,000	6	$540,000	6	$540,000	6	$540,000	$3,780,000
o Network services						$200,000		$350,000		$500,000		$500,000		$500,000		$500,000		$500,000		$500,000	$3,550,000
o Computer hardware						$250,000		$200,000		$150,000		$150,000		$150,000		$150,000		$150,000		$150,000	$1,350,000
o Computer software						$100,000		$100,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000	$500,000
Subtotal, NVO operations						$730,000		$1,010,000		$1,240,000		$1,240,000		$1,240,000		$1,240,000		$1,240,000		$1,240,000	$9,180,000
Surveys and Data Access Grants Program		$300,000		$500,000		$750,000		$750,000		$750,000		$750,000		$750,000		$750,000		$750,000		$750,000	$6,800,000
Research Grants				$500,000		$800,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000	$11,800,000
Fellowship Programs				$300,000		$500,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000	$7,800,000
Education and Public Outreach			1	$150,000	1	$150,000	2	$300,000	3	$450,000	4	$600,000	4	$600,000	4	$600,000	4	$600,000	4	$600,000	$4,050,000
Management					2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	$3,520,000
Administrative					1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	$600,000
Totals	19	$3,420,000	26	$5,420,000	28	$6,845,000	34	$8,835,000	37	$9,275,000	31	$8,365,000	31	$8,365,000	31	$8,365,000	31	$8,365,000	31	$8,365,000	$75,620,000
Totals, Inflation Adjusted		$3,420,000		$5,610,000		$7,333,000		$9,796,000		$10,643,000		$9,935,000		$10,283,000		$10,643,000		$11,015,000		$11,401,000	$90,079,000
Integral Cost		$3,420,000		$9,030,000		$16,363,000		$26,159,000		$36,802,000		$46,737,000		$57,020,000		$67,663,000		$78,678,000		$90,079,000	
Year		2002		2003		2004		2005		2006		2007		2008		2009		2010		2011	
Rates and assumptions:																					
Admin support labor rate, fully burdened		$75,000																			
Operations staff labor rate, fully burdened		$90,000																			
Technical labor rate, fully burdened		$150,000																			
Management labor rate, fully burdened		$220,000																			
Inflation rate		1.035																			

The SDT Recommendations:

1. Form a task force to define the organizational and management structure of the NVO
2. Form a successor to the SDT to maintain a vigorous development process until the actual NVO exists
3. Designate a portion of the funding from the existing programs to be NVO-specific, while seeking a more extensive, dedicated NVO funding for the future
4. Establishment of the NVO Fellowships (postdoc, graduate, undergraduate)
5. EPO component built in from the start in a way which would optimize the involvement, results

In Conclusion …

- The (N)VO is well on its way, with vigorous efforts worldwide, but it needs **your input**
- It responds to a **strong and growing need**
- It is a stage for the forthcoming **revolution in astronomy,** enabled by information technology
- It is a new type of a scientific research organization
- It will **empower** scientists and students everywhere
- It is an **unprecedented venue for EPO** in astronomy and CS/IT, with great societal benefits
- May lead to **significant advances in the applied CS and IT**, with far-reaching benefits

The SDT Membership:

Chairman:
George Djorgovski (Caltech)

Regular Members:
Charles Alcock (U. Penn.)
Piero Benvenuti (ESO)
Roger Brissenden (CfA)
Derek Buzasi (USAF Academy)
Dave DeYoung (NOAO)
Isabel Hawkins (UC Berkeley)
George Helou (Caltech/JPL/IPAC)
Frank Hill (NSO)
Stephen Kent (Fermilab)
Paul Messina (Caltech)
Andrew Moore (CMU)
Jim Schombert (U. Oregon)
Alex Szalay (JHU)
Meg Urry (Yale)
Nicholas White (NASA GSFC)

Ex Officio:
Joe Bredekamp (NASA HQ)
Wayne Van Citters (NSF AST)
Eileen Friel (NSF AST)

Consultative Members:
Robert Brunner (Caltech)
Pepi Fabbiano (CfA ADCCC)
Eric Feigelson (Penn State U.)
Francoise Genova (CDS)
Jim Gray (Microsoft Research)
Jon Hakkila (College of Charleston)
Bob Hanisch (STScI)
Sally Heap (NASA GSFC)
Roberta Humphreys (U. Minn.)
Barry Madore (IPAC/NED)
Roger Malina (UC Berkeley)
Janet Mattei (AAVSO)
Tom McGlynn (NASA GSFC USRA)
Robert Nichol (CMU)
Ethan Schreier (STScI)
Mark Sykes (U. Arizona)