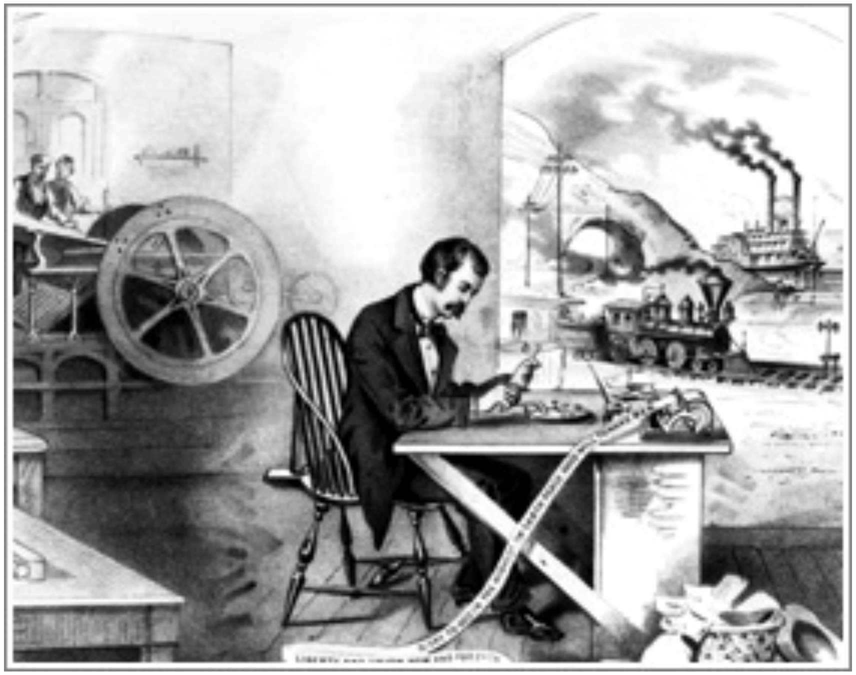
S212

Data and Computing in the Astronomy Community

Matthew J. Graham, CACR, Caltech
S.G. Djorgovski, Caltech

January 31, 2013

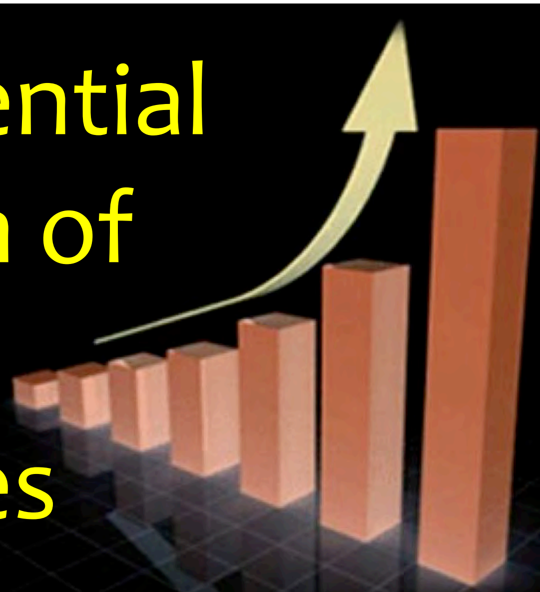


Information technology revolution is historically unprecedented - in its impact it is like the industrial revolution and the invention of printing combined

Science and scholarship are slowly adopting the new tools and technologies and there are great scientific and leadership opportunities in this arena

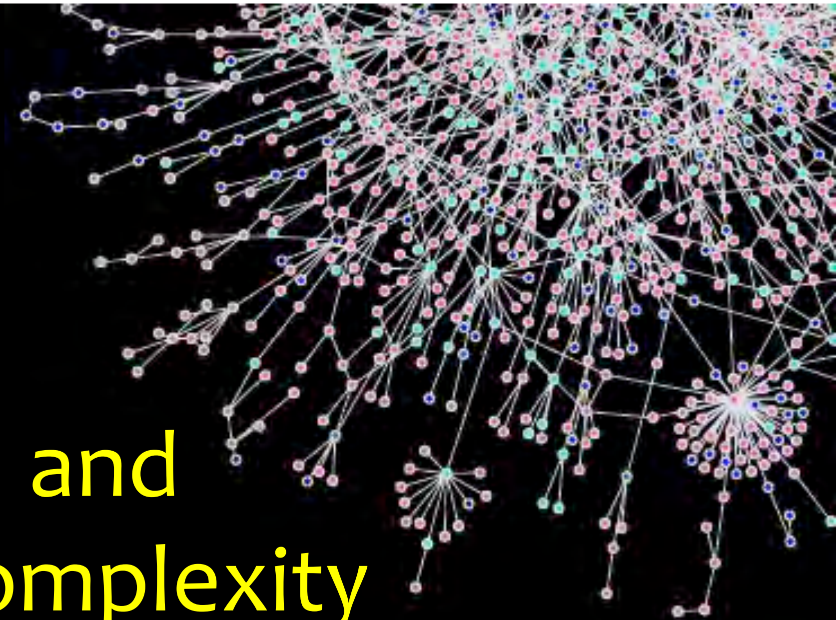
We are effectively developing a new methodology of science and scholarship for the 21st century

Exponential Growth of Data Volumes



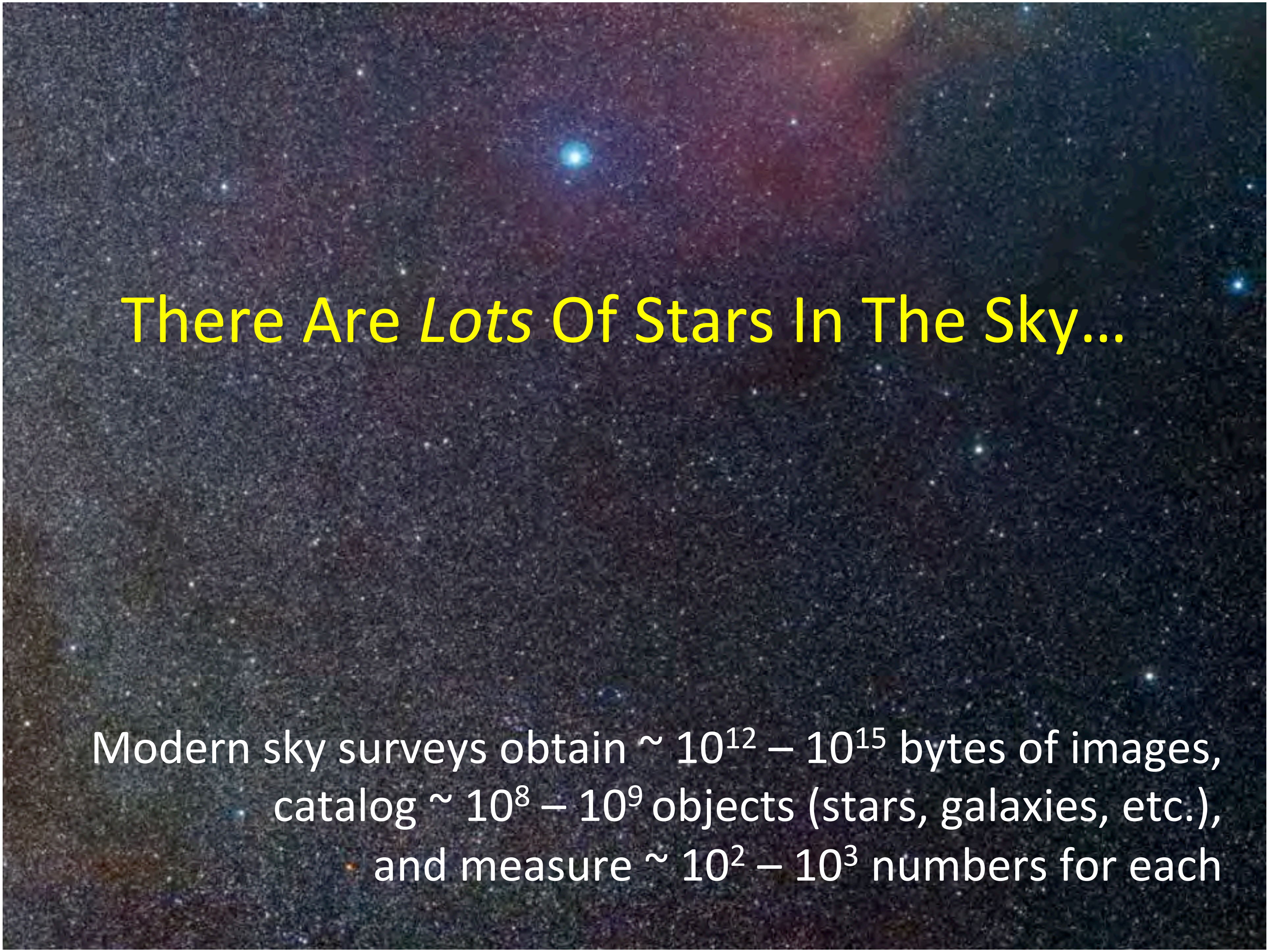
on Moore's law time scales

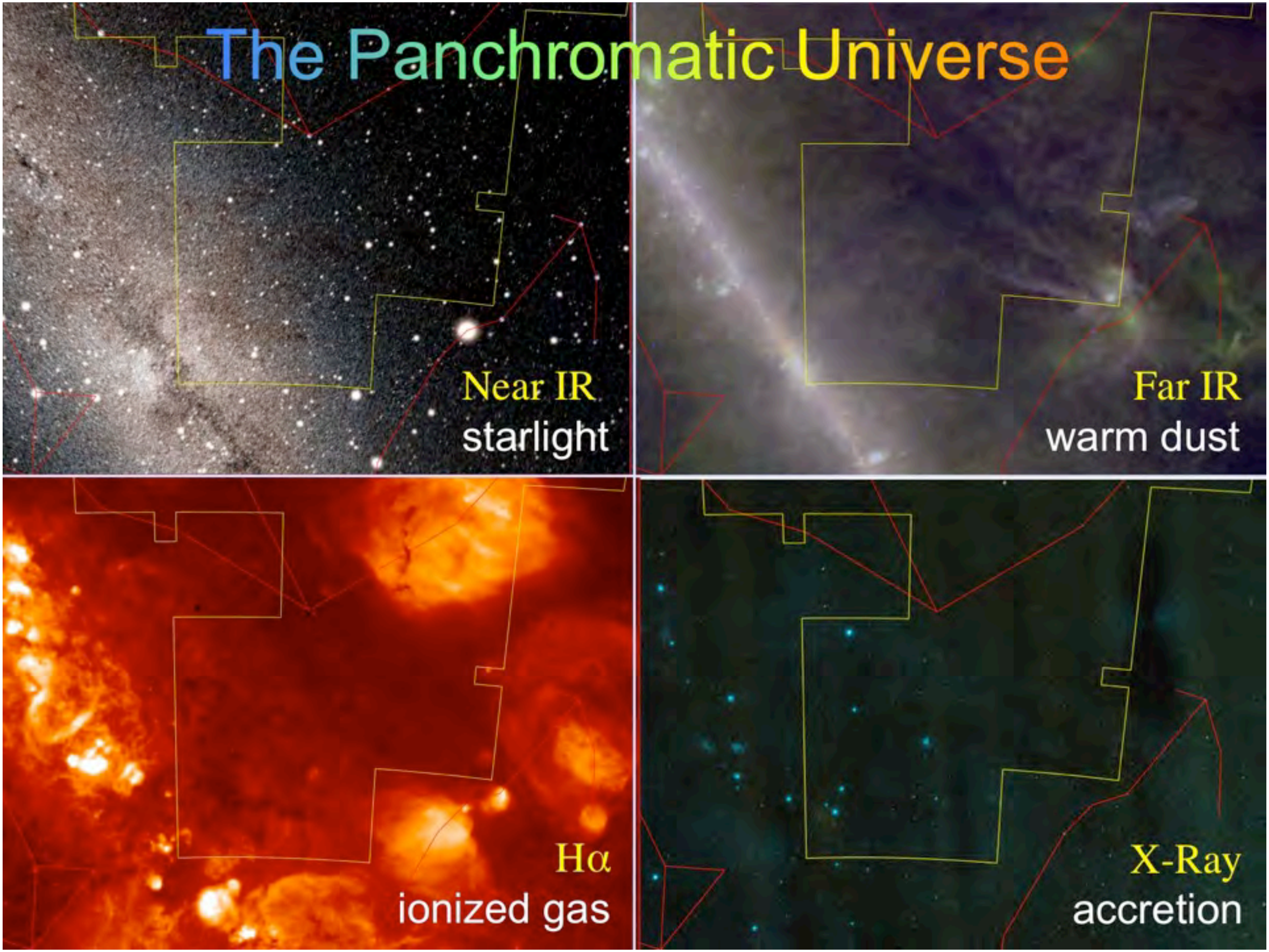
… and Complexity



Understanding of complex phenomena requires complex data!

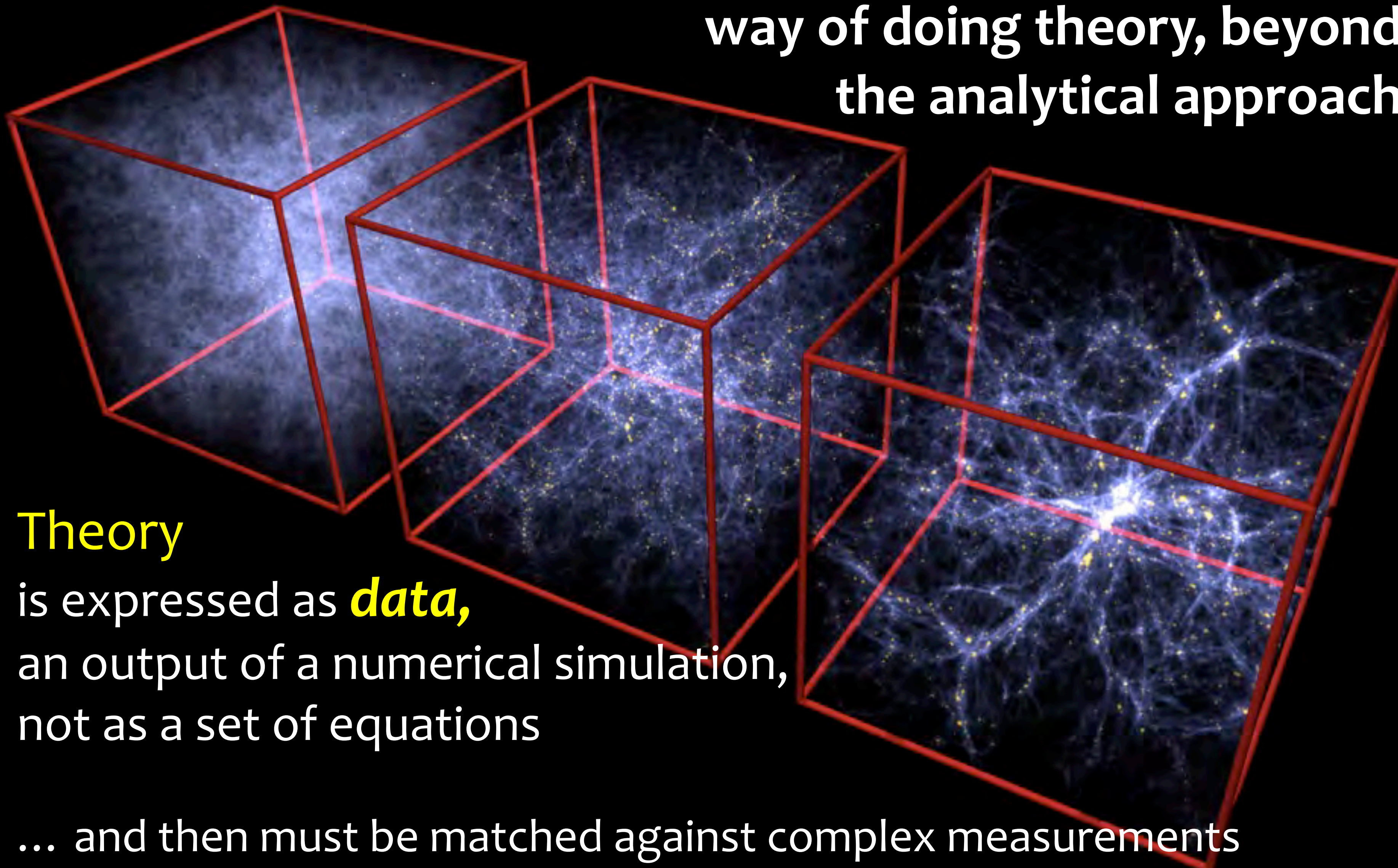
From data poverty to data glut
From data sets to data streams
Theory expressed as data
From static to dynamic, evolving data
From anytime to real-time analysis and discovery
From centralized to distributed resources
From ownership of data to ownership of expertise


There Are *Lots* Of Stars In The Sky…
Modern sky surveys obtain ~ 10^{12} – 10^{15} bytes of images,
catalog ~ 10^{8} – 10^{9} objects (stars, galaxies, etc.),
and measure ~ 10^{2} – 10^{3} numbers for each


The Panchromatic Universe
Near IR
starlight
Far IR
warm dust
Hα
ionized gas
X-Ray
accretion

Numerical Simulations:

A qualitatively different and necessary way of doing theory, beyond the analytical approach



Theory

is expressed as ***data,***

an output of a numerical simulation,

not as a set of equations

… and then must be matched against complex measurements

Astronomy Has Become Very Data-Rich

- Typical digital sky surveys generate ~ 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

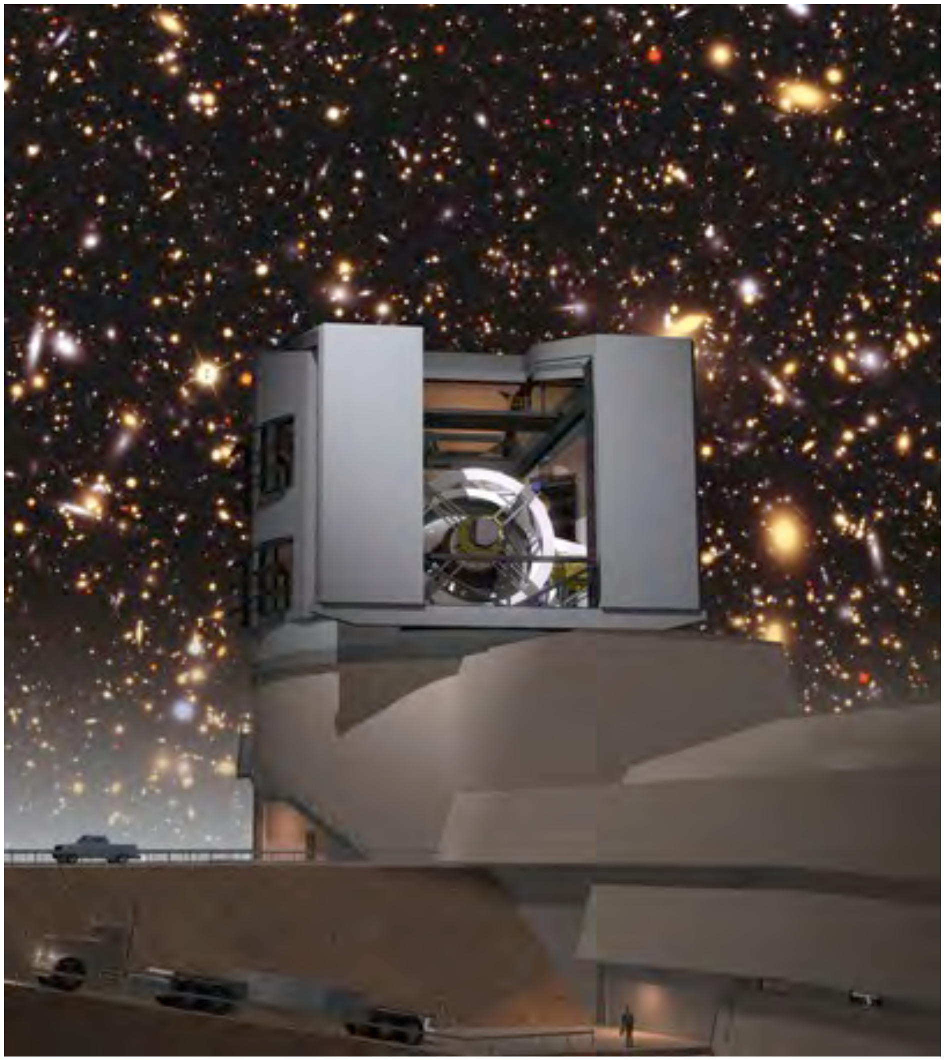
    Human Memory < 1 GB (?)

    1 TB ~ 2 million books

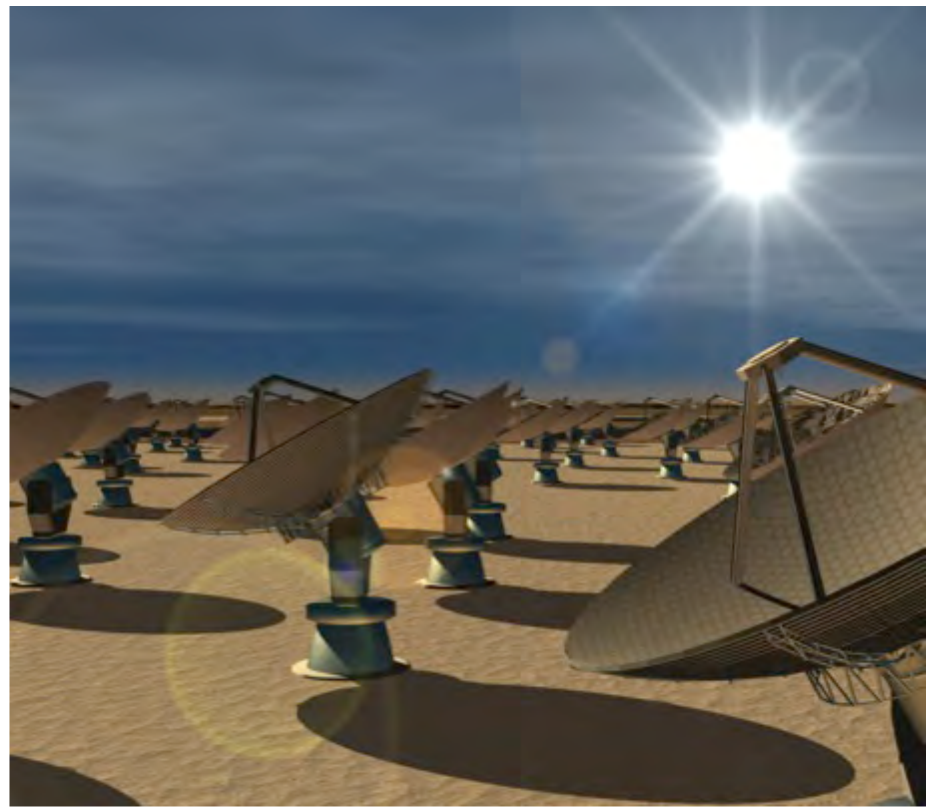
    Human Bandwidth ~ 1 TB / year (±)

… And It Will Get Much More So

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



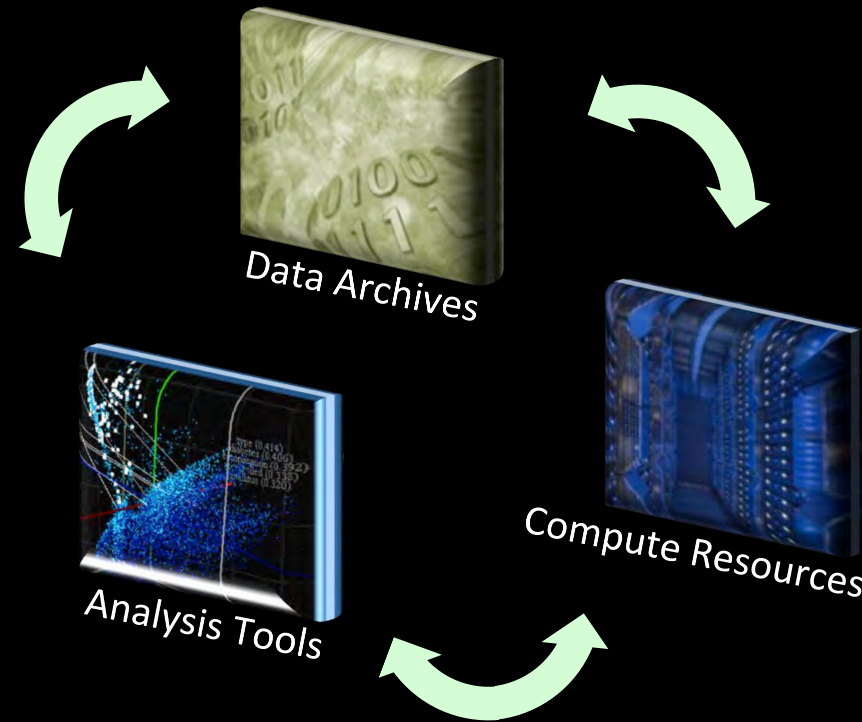
Square Kilometer Array (SKA)
~ 1 EB / second (raw data)
(EB = 1,000,000 TB)



The Era of Cosmic Cinematography

- Synoptic digital sky surveys are now becoming the dominant data producers in astronomy
  - From Terascale to Petascale data streams
- A major new growth area of astrophysics
  - Driven by the new generation of large digital synoptic sky surveys (CRTS, PTF, PS1, … *Fermi*), leading to LSST, SKA, etc.
- All the challenges of traditional sky surveys, plus the time dimension and time-critical analysis requirements
- A broader significance for an automated, real-time knowledge discovery in massive data streams

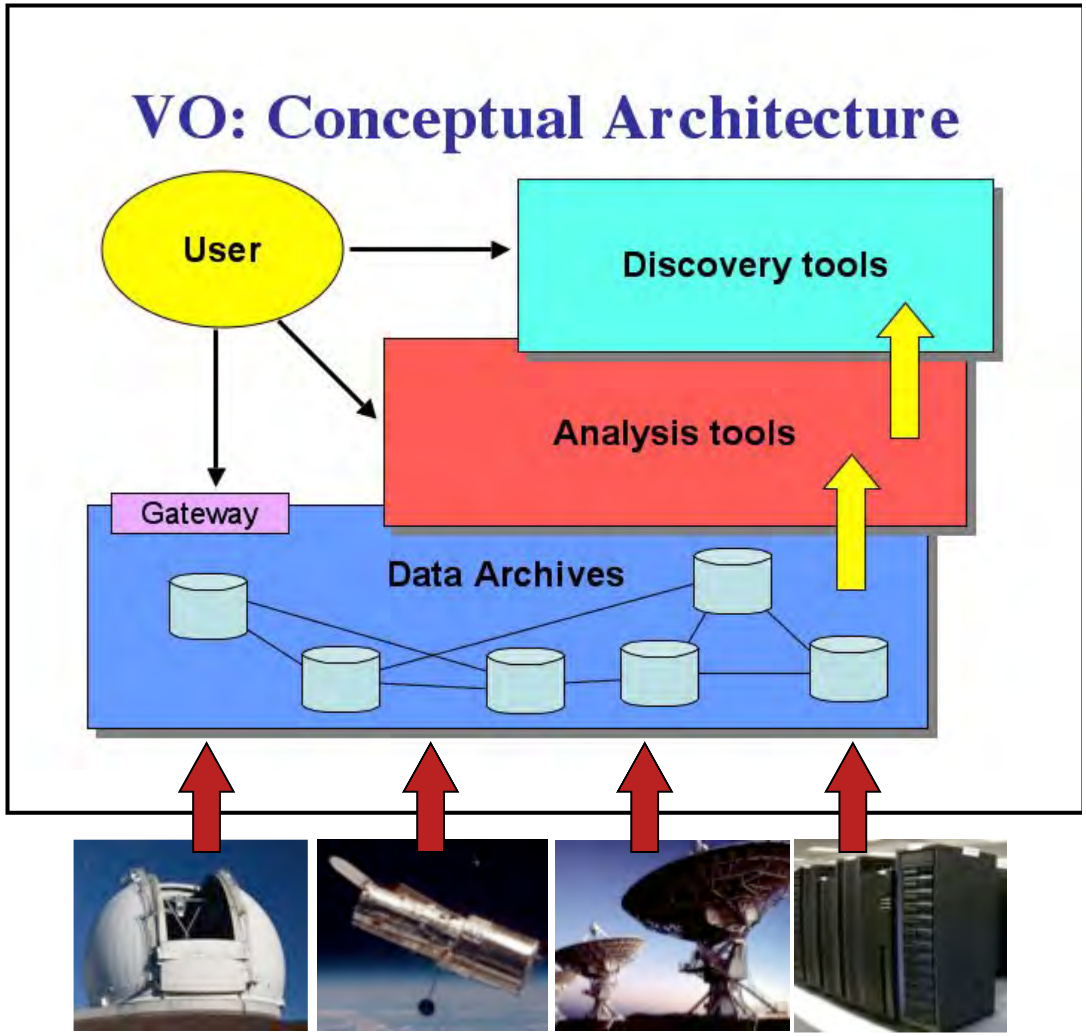
The Rise of Virtual Scientific Organizations



- A grassroots response of scientific communities to the challenges and opportunities brought by the data glut
- Domain-specific, not institution-based; inherently distributed
  - The human, data, and compute resources are distributed
  - A new type of a scientific organization, needing new management models
- Should VO's have a finite lifetime, as they fulfill their role?

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

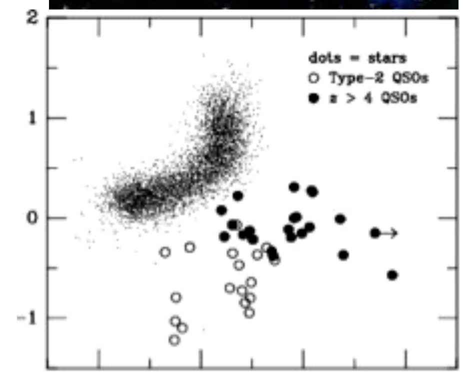
Scientific Roles and Benefits of a VO

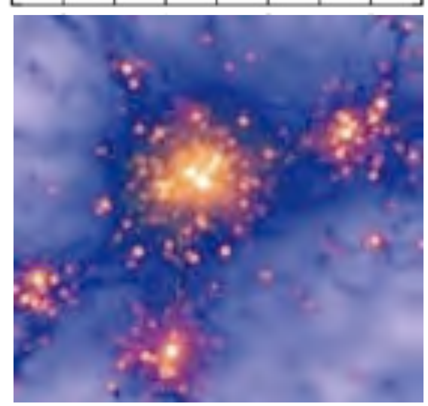
- **Facilitate science with massive data sets** (observations and theory/simulations) ⟹ **efficiency amplifier**
- Provide an **added value** from federated data sets (e.g., multi-wavelength, multi-scale, multi-epoch …)
  - Discover the knowledge which is present in the data, but can be uncovered *only* through data fusion
- **Enable and stimulate some *qualitatively new* science** with massive data sets (not just old-but-bigger)
- **Optimize the use of expensive resources** (e.g., space missions, large ground-based telescopes, computing …)
- Provide R&D drivers, application testbeds, and stimulus to the **partnering disciplines** (CS/IT, statistics …)

Virtual Observatory Science Examples

- Combine the data from multi-TB, billion-object surveys in the optical, IR, radio, X-ray, etc., for:
  - 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, brown dwarfs, new types, etc.
- Probe the evolution of quasars, galaxies, or clusters discovered using different techniques over the cosmic time
- Match Peta-scale numerical simulations of star or galaxy formation with equally large and complex observations

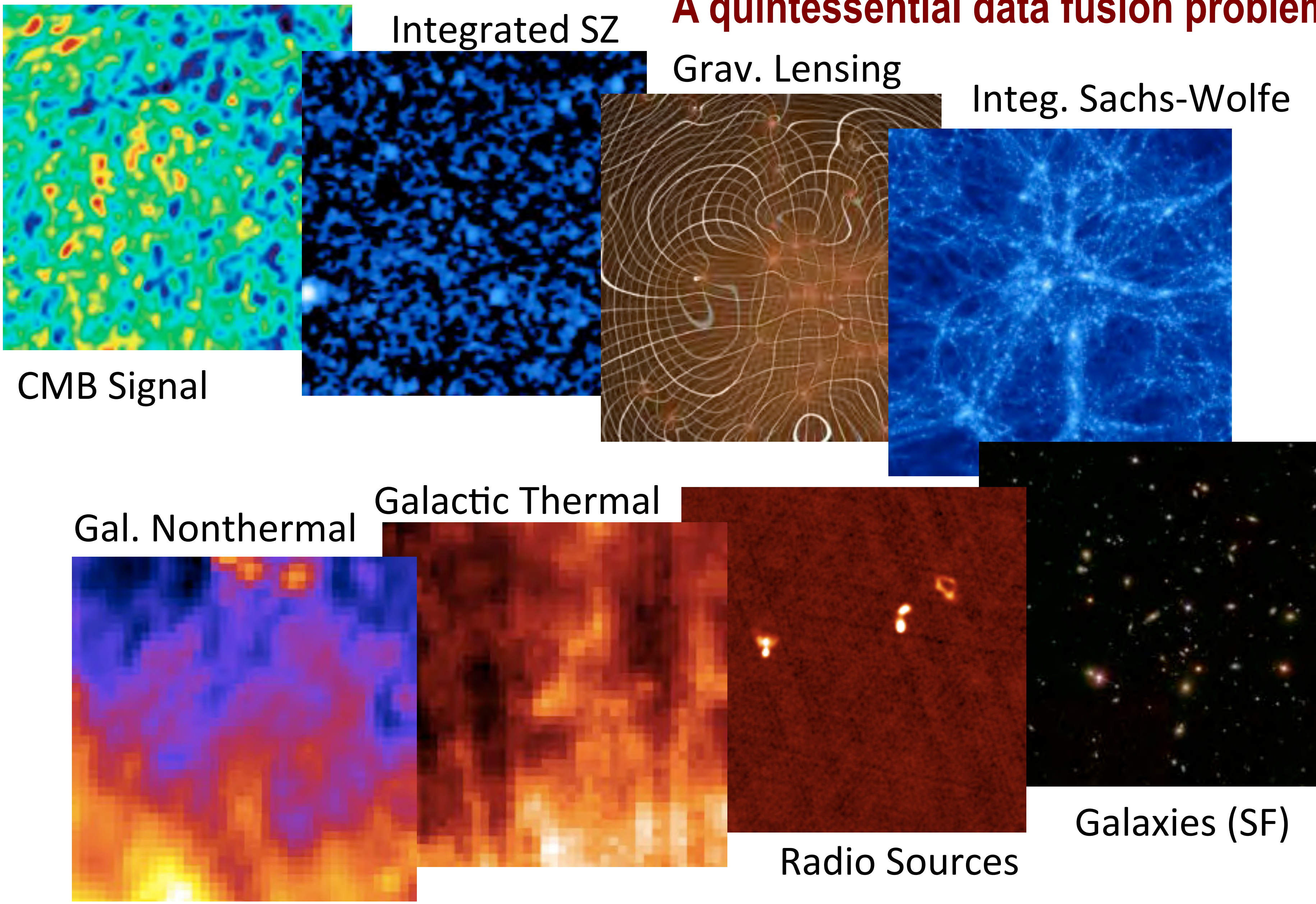
... etc., etc.





Understanding the CMBR Foregrounds

A quintessential data fusion problem



Virtual Observatory Is Real!

http://usvao.org

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

A Modern Scientific Discovery Process

Data Gathering (e.g., from sensor networks, telescopes…)

└→ **Data Farming:**

- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability

} Database Technologies

└→ **Data Mining** (or **K**nowledge **D**iscovery in **D**atabases):

Key Technical Challenges →

- Pattern or correlation search
- Clustering analysis, classification
- Outlier / anomaly searches
- Hyperdimensional visualization

└→ **Data Understanding**

← Key Methodological Challenges

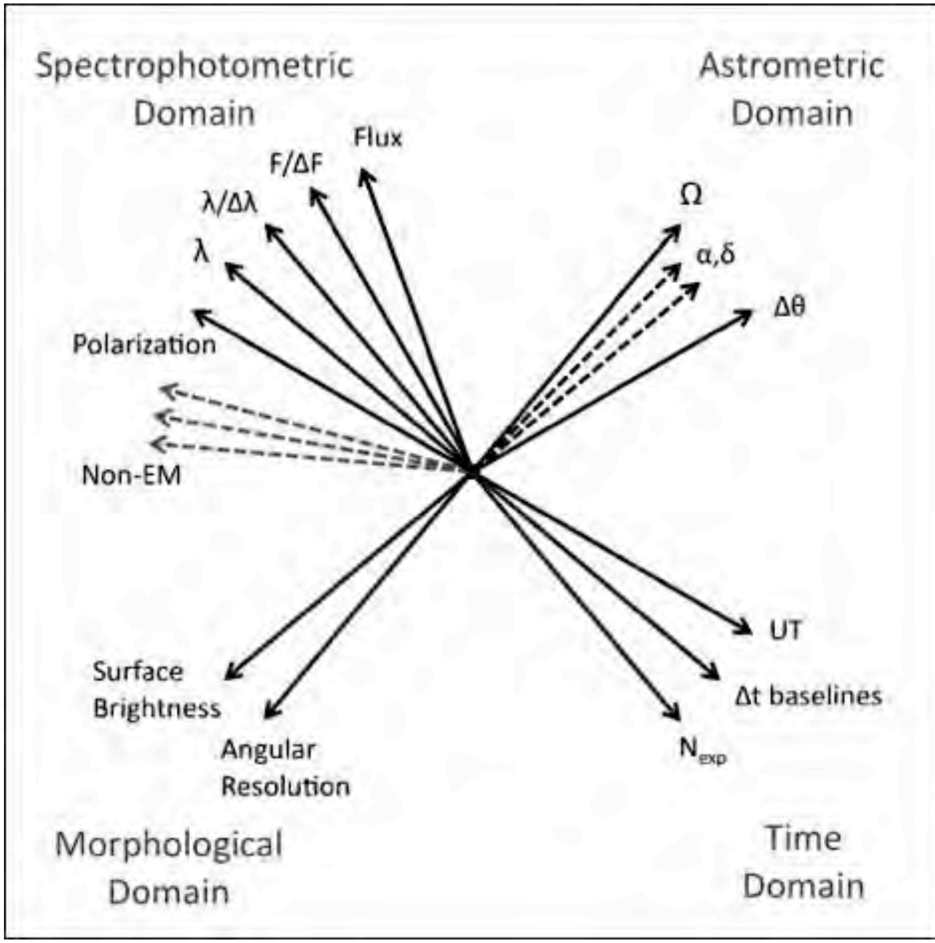
└→ **New Knowledge**

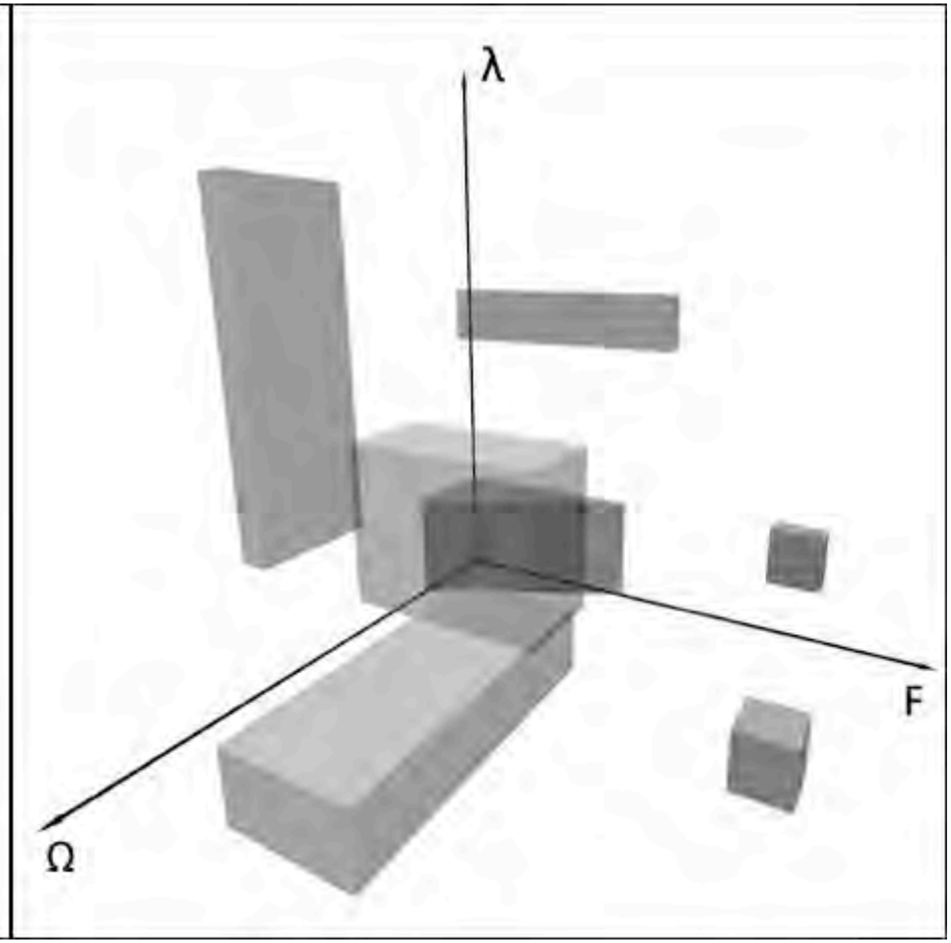
+feedback

Systematic Exploration of the Observable Parameter Space (OPS)

Its axes are defined by the observable quantities

Every observation, surveys included, carves out a hypervolume in the OPS





Technology opens new domains of the OPS → **New discoveries**

Astronomy in the Time Domain

- Rich phenomenology, from the Solar system to cosmology and extreme relativistic physics
  - Touches essentially every field of astronomy
- For some phenomena, time domain information is a key to the physical understanding
- A qualitative change:

Static ⇨ Dynamic sky

Sources ⇨ Events

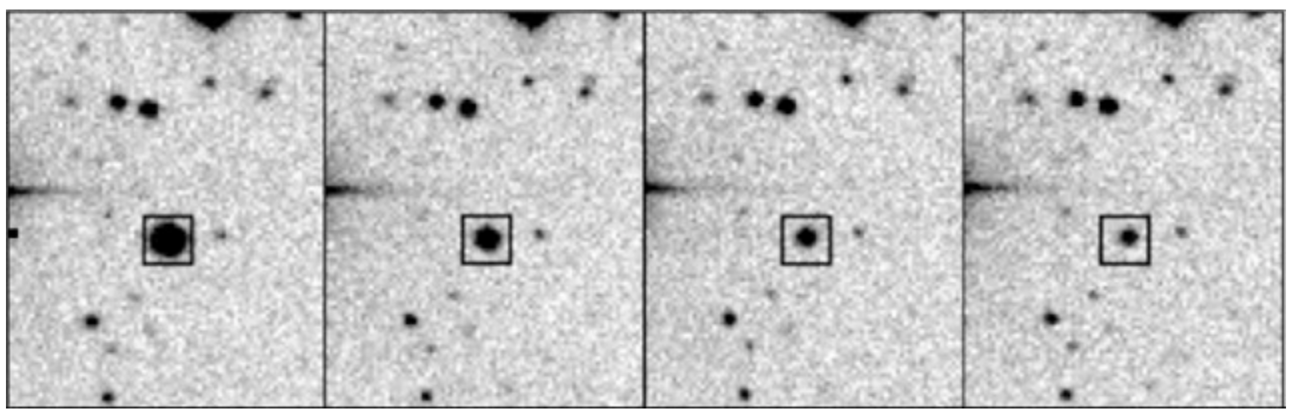
- Real-time discovery/reaction requirements pose new challenges for knowledge discovery

Synoptic, panoramic surveys ➔ event discovery

Rapid follow-up and multi-λ ➔ keys to understanding

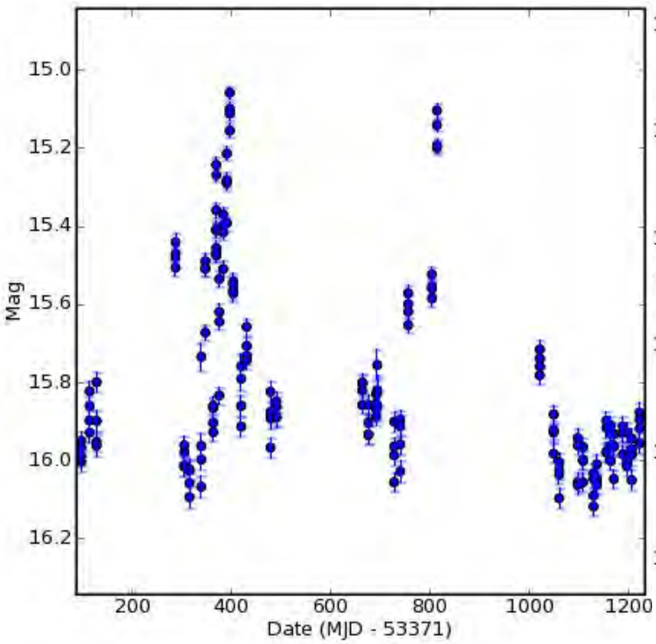
The Catalina Real-Time Transient Survey (CRTS)

- Collaboration with UAz/LPL search for NEA/PHA asteroids
- 3 small telescopes up to 2,500 deg^2/night with 4 exposures/ pointing, limiting mags ~ 19 – 21, several tens of passes per year, total area coverage ~ 33,000 deg^2 , time baselines from 10 min to years, ~ 7+ years coverage
- Real time processing and event discovery and publication
- **Open data policy:** ***all data are made public immediately***
- ~ 6,500 unique transients so far, a number of discoveries made

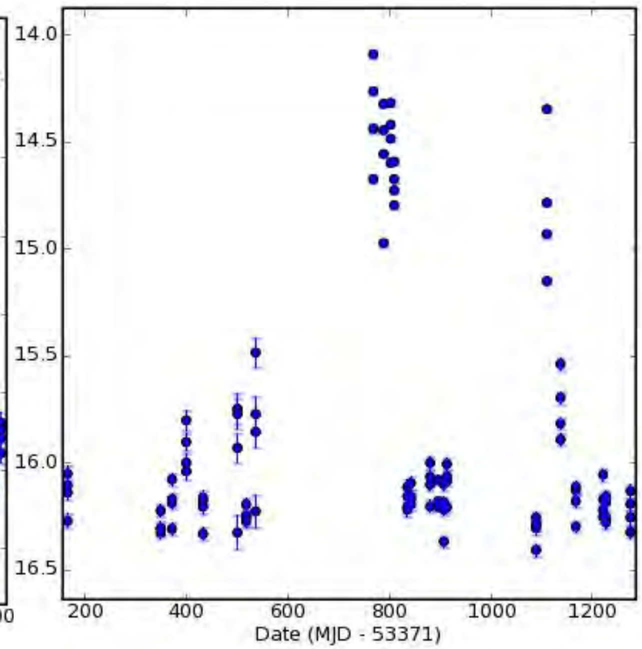


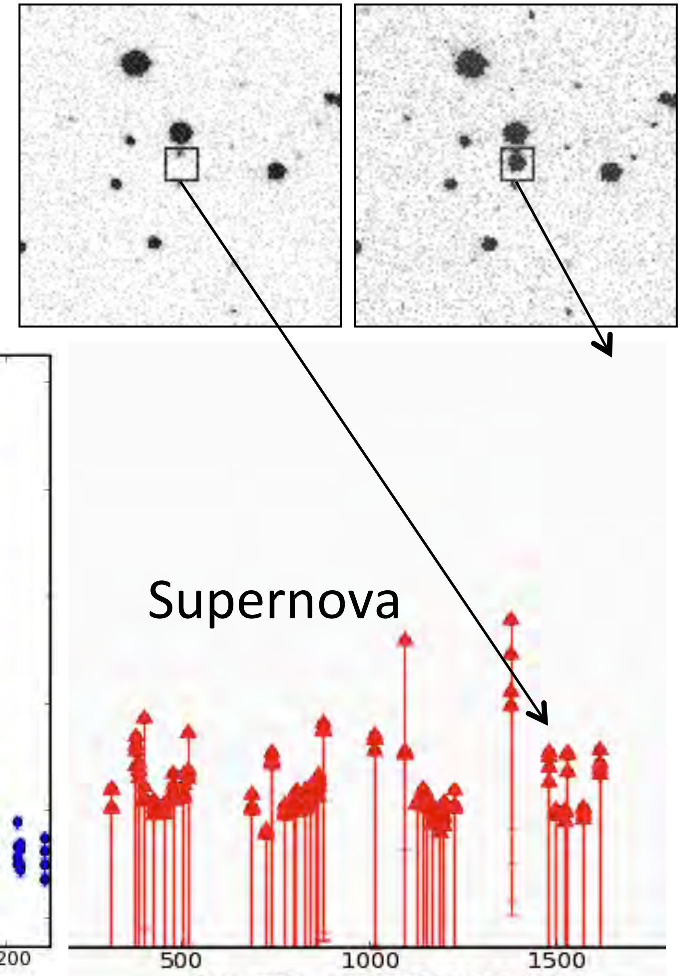
Sample Light Curves

Blazar PKS0823+033



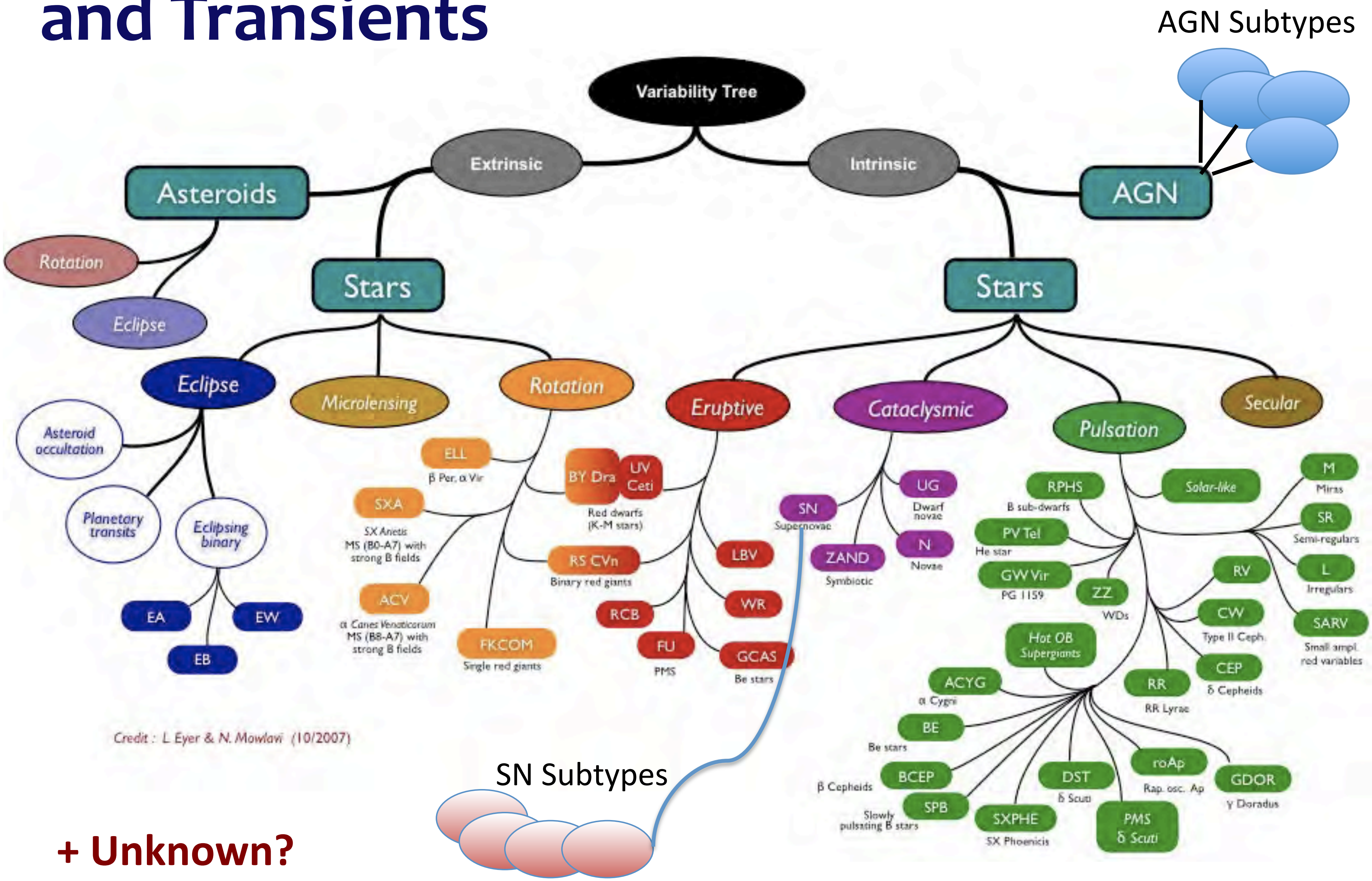
CV 111545+425822





- Large-amplitude transients published immediately, light curves accumulated for every source (~ 500 million)
- Transients are perishable – must be followed rapidly in order to get the science, but the follow-up is very limited

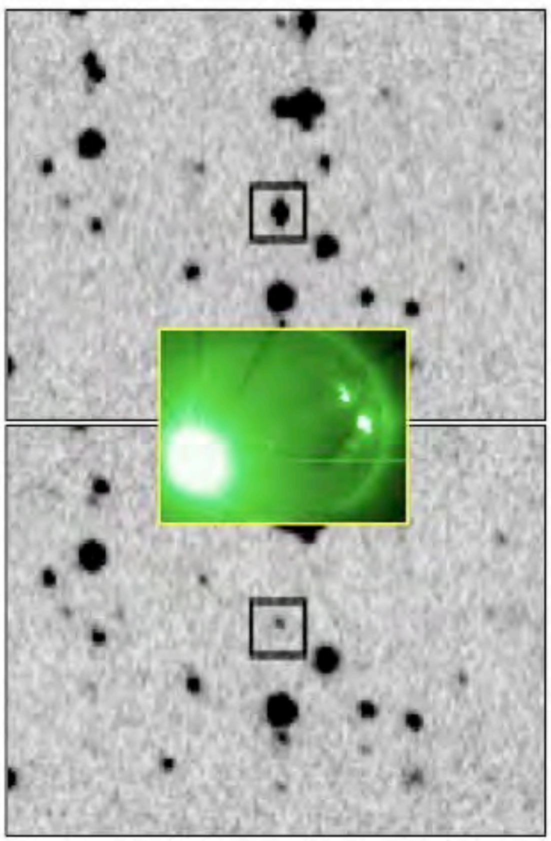
Semantic Tree of Astronomical Variables and Transients



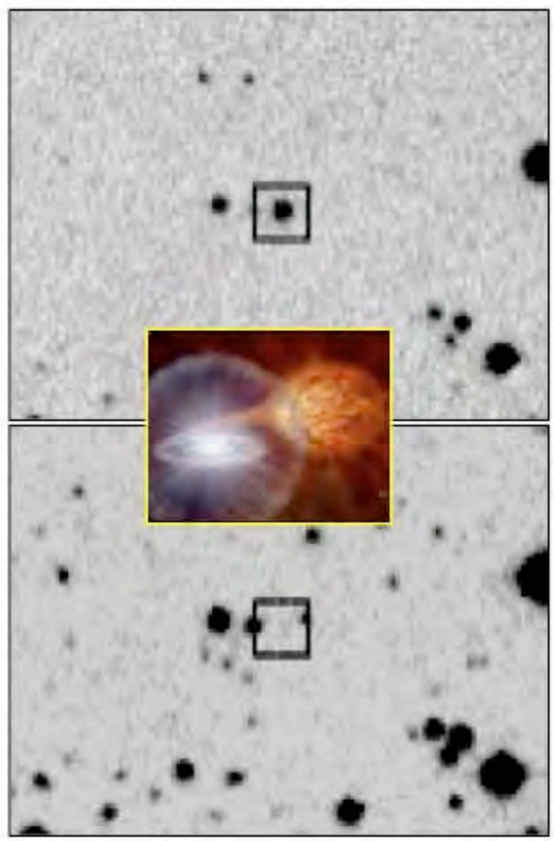
+ Unknown?

Automated Classification of Transients

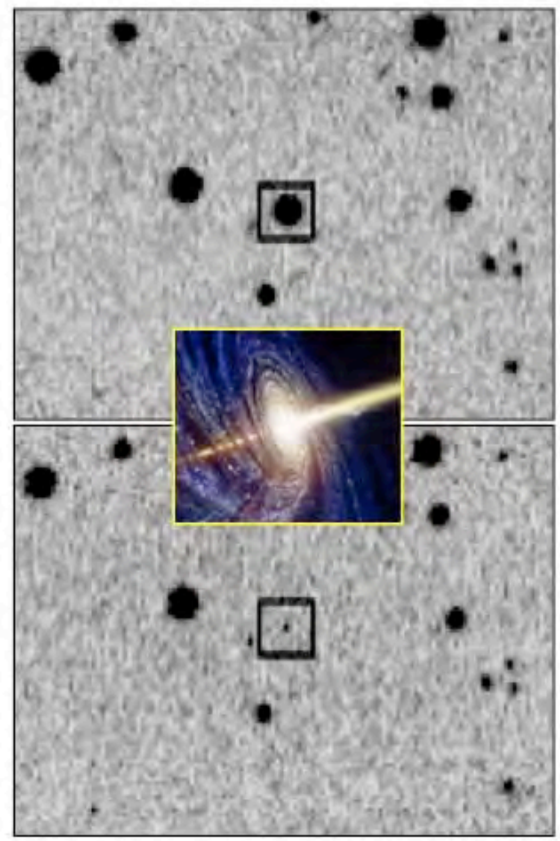
Flare star



Dwarf Nova



Blazar



Vastly different physical phenomena, yet they look the same!
Which ones are the most interesting and worthy of follow-up?

Rapid, automated transient classification is a critical need!

This is a Critical Problem

(and it will get a lot worse)

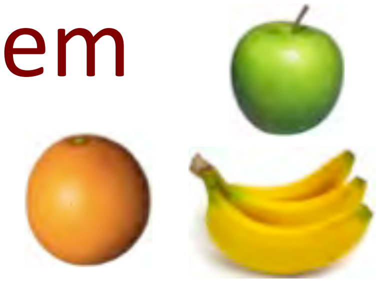
- Now: data streams of **~ 0.1 TB** / **night, ~ 10^2 transients** / **night** (CRTS, PTF, various SN surveys, microlensing, etc.)
  - ✧ We are already in the regime where we ***cannot follow them all***
  - ✧ Spectroscopy is the key bottleneck now, and it will get worse
- Forthcoming on a time scale ~ 1 - 5 years: **~ 1 TB / night, ~ 10^3 - 10^4 transients / night** (PanSTARRS, Skymapper, VISTA, VST, SKA precursors…)
- Forthcoming in ~ 8 – 10 (?) years: LSST, **~ 30 TB / night, ~ 10^5 - 10^7 transients / night**, SKA

} **A major, qualitative change!**

- So… which ones will you follow up?
- Follow-up resources will likely remain limited

} ***Transient classification is essential***

Event Classification is a Hard Problem

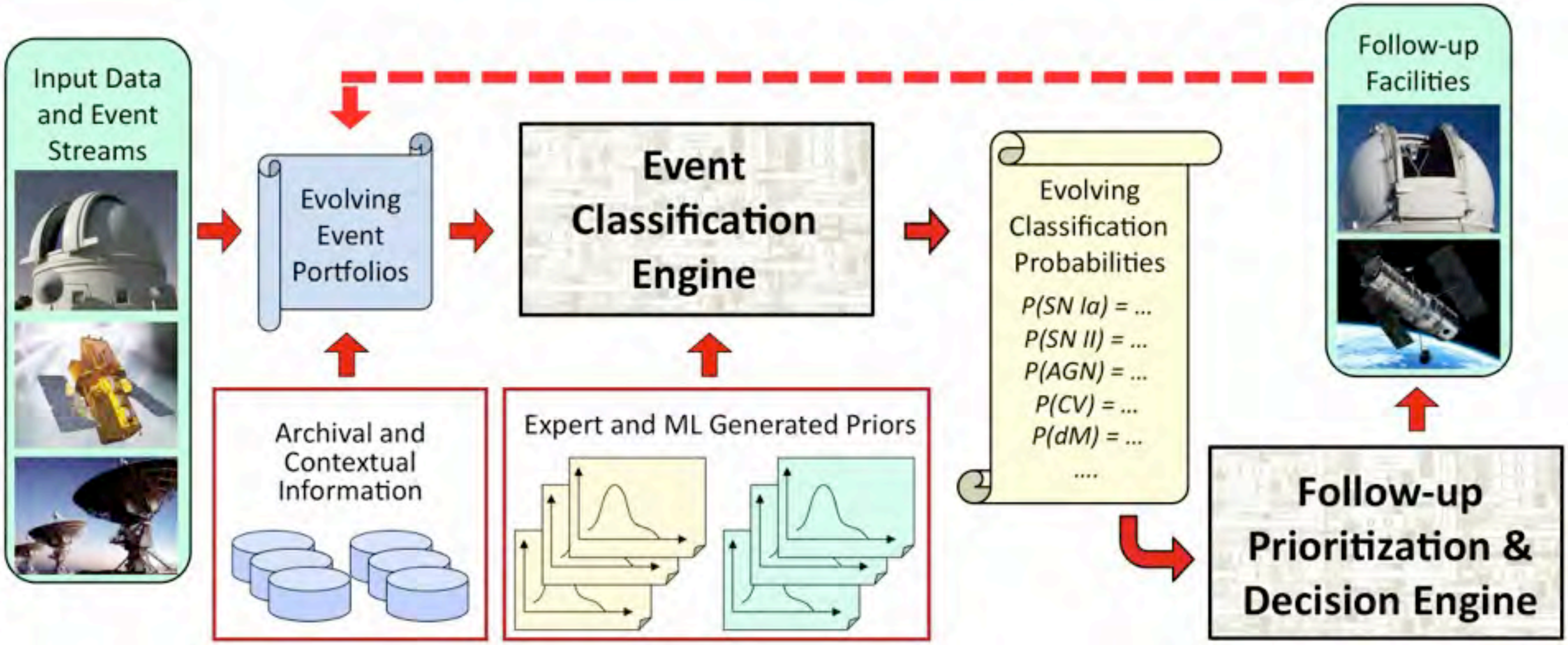


- Classification of transient events is essential for their astrophysical interpretation and uses
  - Must be done in real time and iterated dynamically

- Human classification is already unsustainable, and will not scale to the Petascale data streams
- This is hard:

  - Data are sparse and heterogeneous: feature vector approaches do not work; using Bayesian approach
  - Completeness vs. contamination ☯
  - Follow-up resources are expensive and/or limited: only the most interesting events
  - Iterate classifications dynamically as new data come in
- Traditional DP pipelines do not capture a lot of the relevant contextual information, prior/expert knowledge, etc.

Towards an Automated Event Classification



- Incorporation of the contextual information (archival, and from the data themselves) is essential
- Automated prioritization of follow-up observations, given the available resources and their cost
- A dynamical, iterative system

The Key Challenge: Data Complexity

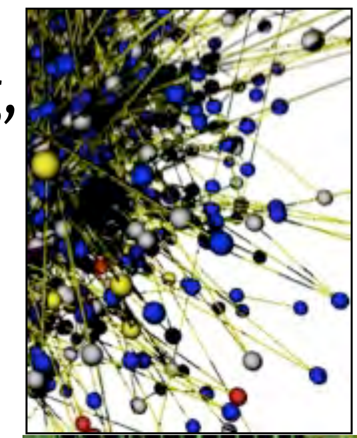
Or: The Curse of Hyper-Dimensionality

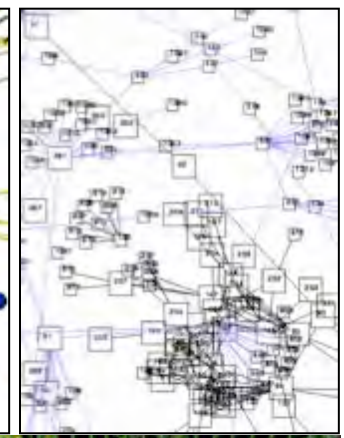
1. Data mining algorithms scale very poorly:

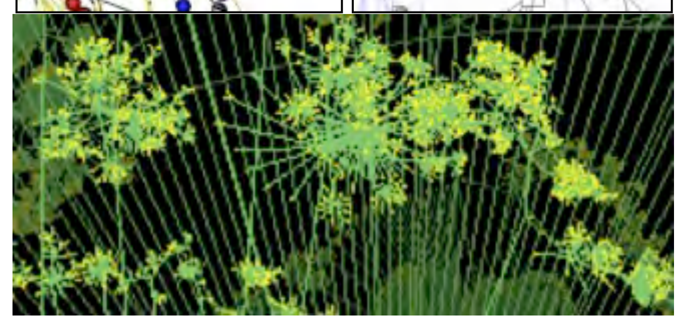
N = data vectors, ~ 10^8 - 10^9, D = dimension, ~ 10^2 - 10^3

- Clustering ~ $N \log N \rightarrow N^2$, ~ D^2
- Correlations ~ $N \log N \rightarrow N^2$, ~ D^k $(k \geq 1)$
- Likelihood, Bayesian ~ N^m $(m \geq 3)$, ~ D^k $(k \geq 1)$

2. Visualization in >> 3 dimensions



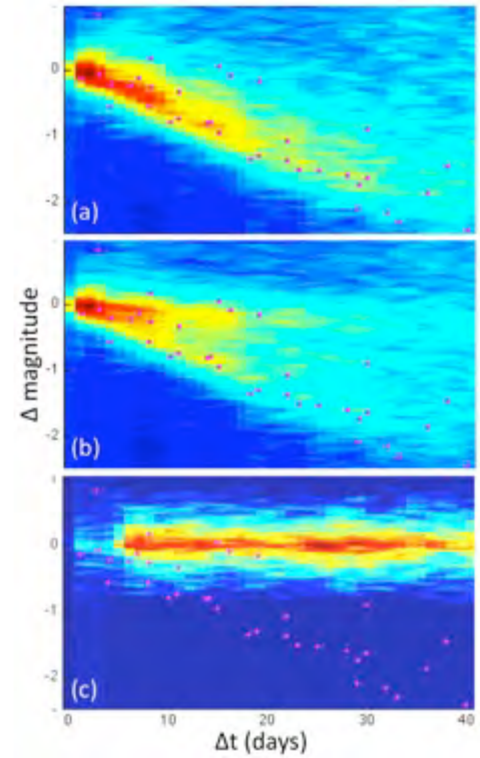


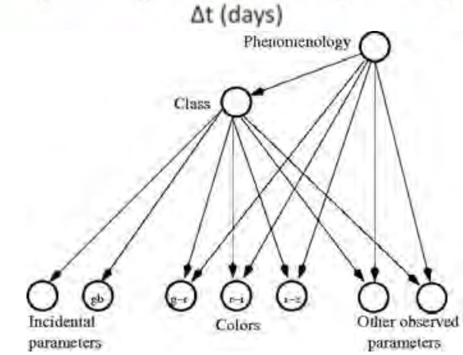


- The complexity of data sets and interesting, meaningful constructs in them is *exceeding the cognitive capacity of the human brain*
- We are biologically limited to perceiving D ~ 3 - 10(?) dimensions
- Visualization must be a component of the data mining / exploration process
- It is the bridge between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

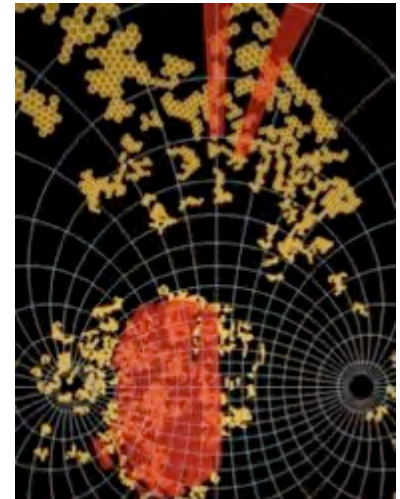
Look to new techniques (for astronomy)

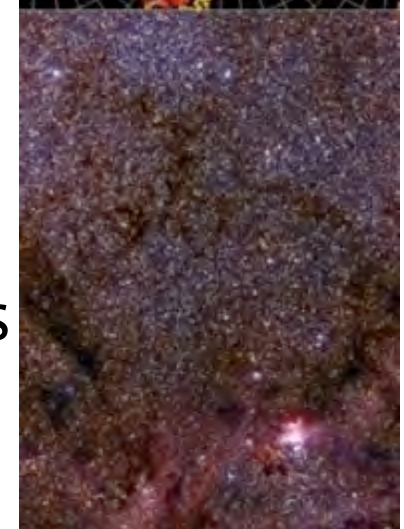
- Data are sparse and heterogeneous
- Light curve characterization/feature extraction
  - Thiel-San estimator
  - AR(1) time series
  - Bayesian blocks
  - Local regression
- Classification
  - Bayesian networks
  - Symbolic regression
  - Probabilistic structure functions
  - Knowledge-based (semantic)
  - Hierarchical approaches
  - Fusion modules

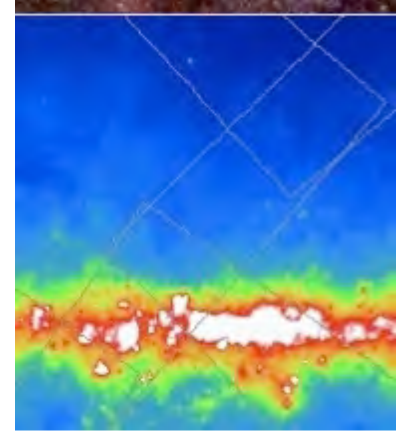




VO Functionality Today







What we did so far:

- 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!

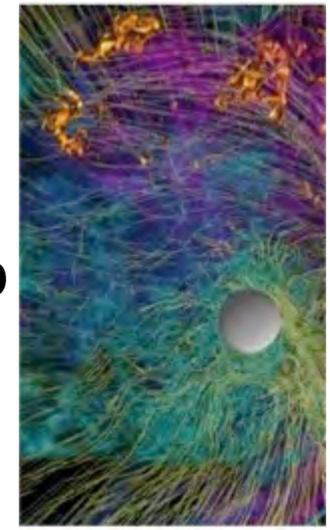
Thus, little VO-enabled science so far

Thus, a slow community buy-in

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

Some Thoughts About e-Science

- Comput***ational*** science ≠ Comput***er*** science
- Data-driven science is *not* about data, it is about ***knowledge extraction*** (the data are incidental to our real mission)
- Information and data are (relatively) cheap, but the expertise is expensive
  - Just like the hardware/software situation
- Computer science as the "new mathematics"
  - It plays the role in relation to other sciences which mathematics did in ~ 17^{th} - 20^{th} century
- Computation: an interdisciplinary glue/lubricant
  - Many important problems (e.g., climate change) are inherently inter/multi-disciplinary



The 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 ≠ 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**

+ *things we have not thought of yet …*

Beyond Virtual Scientific Organizations:

The Rise of X-Informatics (X = Astro, Bio, Geo, ..)

- Domain-specific amalgam fields (science + CS + ICT)
- A mechanism for a broader community inclusion (both as contributors and as consumers)
- A mechanism for interdisciplinary e-Science methodological sharing