Data Science Challenges from Distributed Petascale Astronomical Sky Surveys

Kirk Borne

George Mason University, Department of Computational & Data Sciences

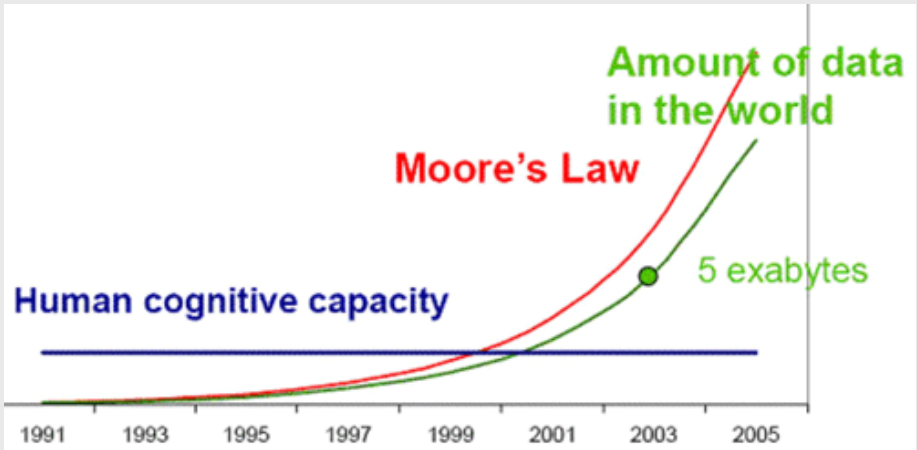
http://classweb.gmu.edu/kborne/

http://www.lsst.org/

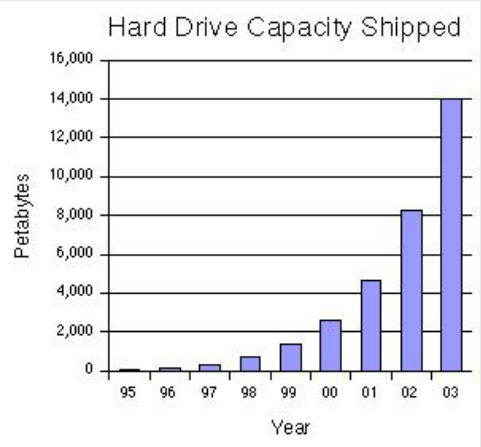
http://www.us-vo.org/

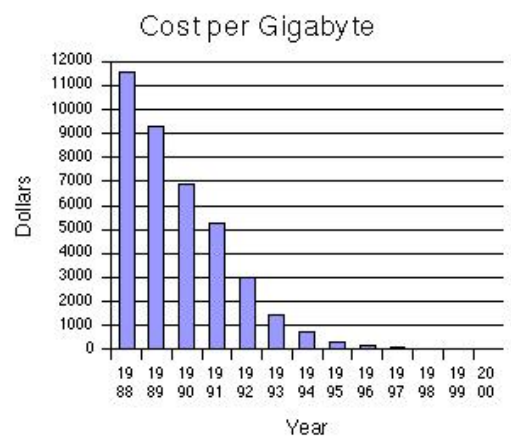
Data Doubles Every Year

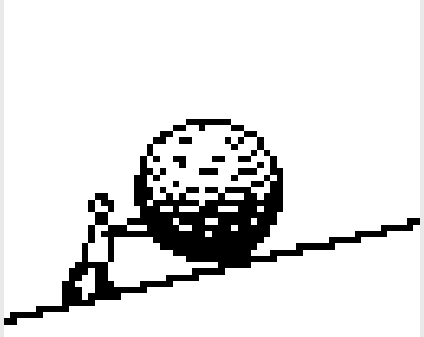
- Computing power doubles every 18 months (Moore's Law) …
  - That's **100x in 10 years**
- I/O bandwidth increases ~10% / year
  - That's **<3x in 10 years**.
- Data doubles every year …
  - That's **1000x in 10 years**, and 1,000,000x in 20 years.
  - NCSA Example:
    - First 19 years: 1 PB
    - Year 20 (2007): 2 PB
    - Year 21 (2008): 4 PB
    - By 2020 : ~20 Exabytes ?
  - *In the Year 2525* : 10^{156} PB ?????
- As our data volumes grow, especially in the sciences (where the scientific funding economics are not driven by data volumes), we will fall farther and farther behind in our ability to analyze, assimilate, and extract knowledge from our data collections … unless we develop and apply exponentially more powerful algorithms and methods.

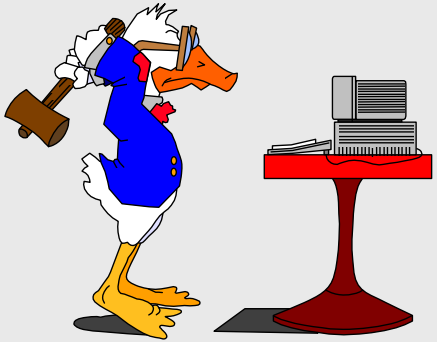


(Illustration adapted from a slide by J. Heer, PARC User Interface Research Group)

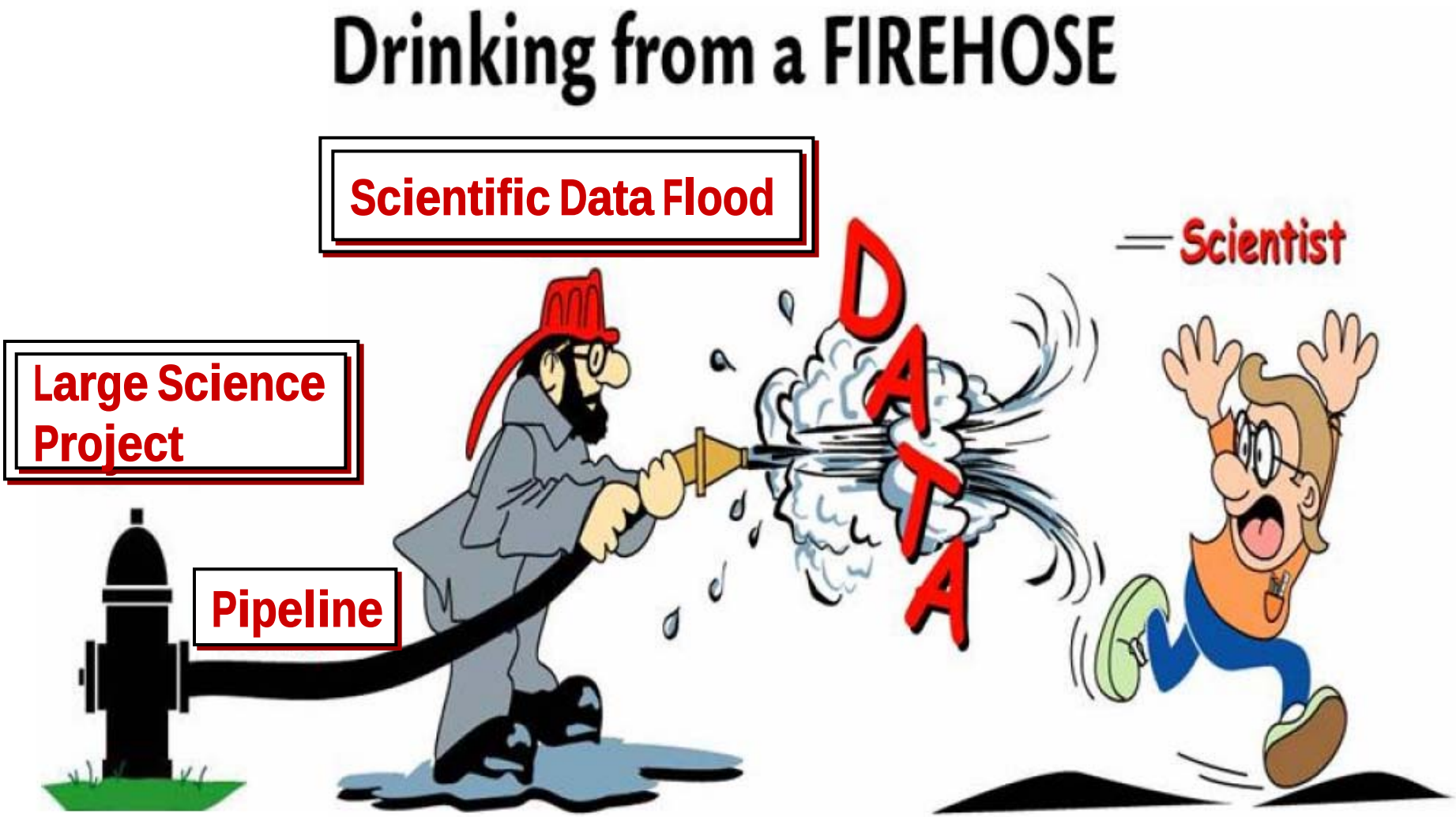








Large Scientific Database Projects



Astronomy is a Discovery-driven Science

- Discoveries are enabled by:
  - New questions
  - New ideas
  - New models
  - New theory
  - And most importantly ... New data!

Discoveries have shown that the astronomical zoo is rich and diverse ...

Black Holes

Quasars

Supernovae

Pulsars

Blazars

Tidal Streams

Colliding Galaxies

Magnetars

Gamma-ray bursts

Brown Dwarfs

Gravitational Lenses

Exo-planets

Serendipity !!

Incoming Killer Asteroid

Astronomy is a Discovery-driven Science

- Discoveries are enabled by:
  - New questions
  - New ideas
  - New models
  - New theory
  - And most importantly ... New data!
- **Discoveries lead to:**
  - **New questions**
  - **New ideas**
  - **New models**
  - **New theory**
  - **And most importantly ... <u>More</u> new data!**
- We need increasingly efficient mining & analysis algorithms

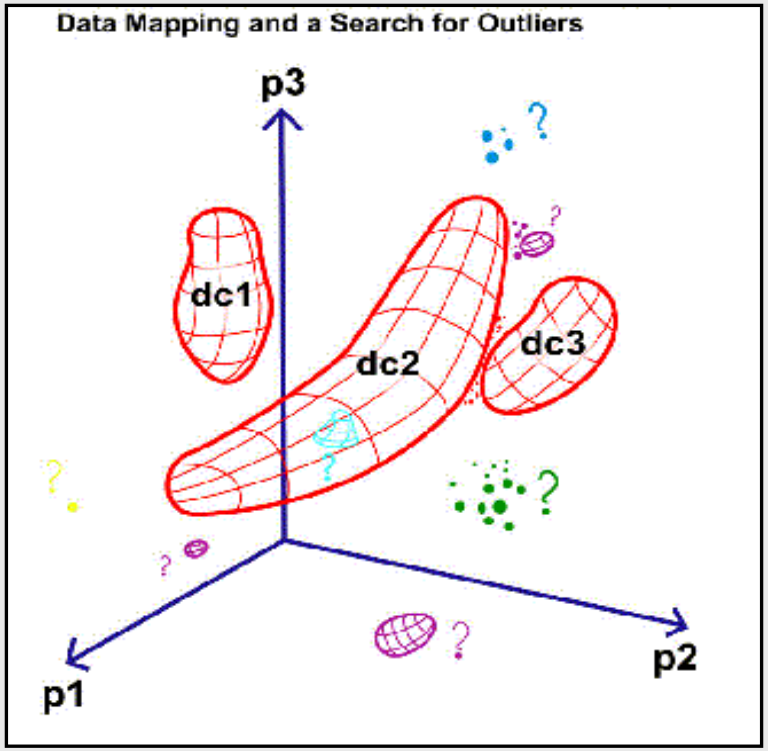
Astronomers have been doing Data Mining for centuries

"The data are mine, and you can't have them!"

- Seriously ...
- Astronomers love to classify things ...
  (Supervised Learning. e.g., classification)
- Astronomers love to characterize things ...
  (Unsupervised Learning. e.g., clustering)
- And we love to discover new things ...
  (Semi-supervised Learning. e.g., outlier detection)

This sums it up ...

- Characterize the known (clustering)
- Assign the new (classification)
- Discover the unknown (outlier detection)

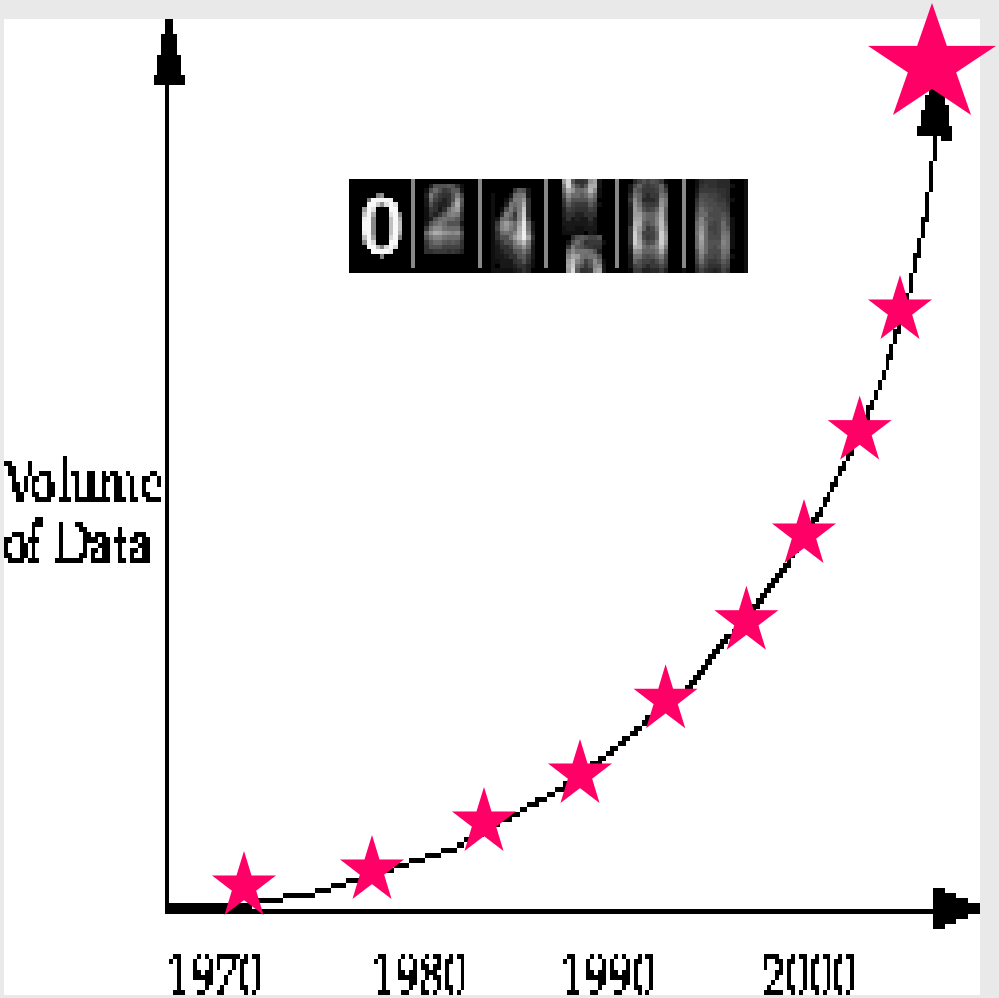


Graphic from S. G. Djorgovski

- Benefits of very large data sets within a scientific domain:
  - **best statistical analysis of "typical" events**
  - **automated search for "rare" events**

Astronomy data volumes are growing and growing astronomically

- **a few terabytes "yesterday" (10,000 CDROMs)**
- **tens of terabytes "today" (100,000 CDROMs)**
- **100's of petabytes "tomorrow" (within 10-20 years) (1,000,000,000 CDROMs)**



From Data-Driven to Data-Intensive

- Astronomy has always been a data-driven science
- It is now a data-intensive science: <u>Astroinformatics</u>
  - And it will become even more data-intensive in the coming decade(s)
- Some key data-driven questions for astronomers:
  - What is it?
  - Where is it?
  - What causes that behavior?
  - When did it form?
  - How did it form?
  - Why did it do that?
  - Who will let me use their telescope to get more data???

Knowledge !

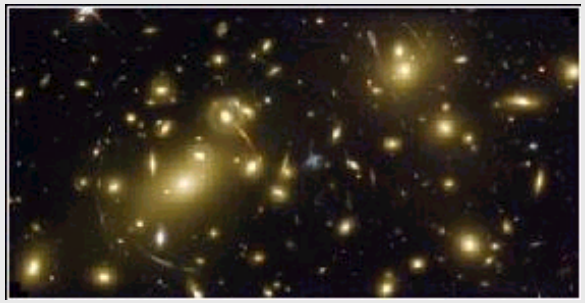
Science is Knowledge Work

Data ➔ Information ➔ Knowledge

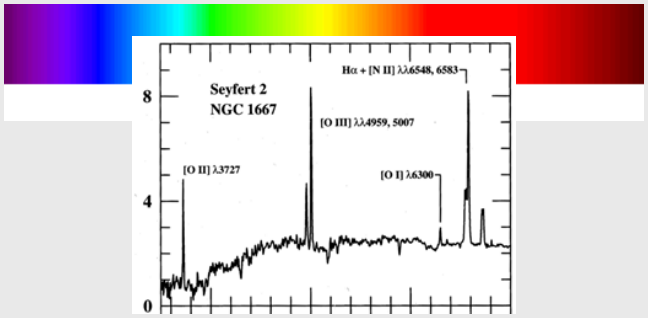
- Knowledge Discovery is the central theme of science.
- Knowledge Discovery in Databases (KDD) is the killer app for large scientific databases.
- Therefore, KDD (i.e., Data Mining) is an essential tool, since “big-data” science is here to stay (at petabytes and beyond).
- Example ...

Astronomy Example

Data:



(a) Imaging data (ones & zeroes)



(b) Spectral data (ones & zeroes)

Information (catalogs / databases):

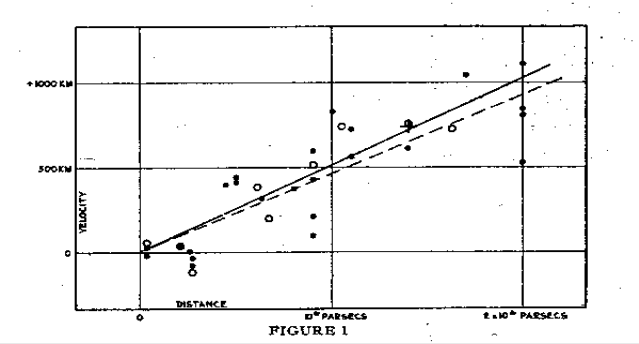
- Measure brightness of galaxies from image (e.g., 14.2 or 21.7)
- Measure redshift of galaxies from spectrum (e.g., 0.0167 or 0.346)

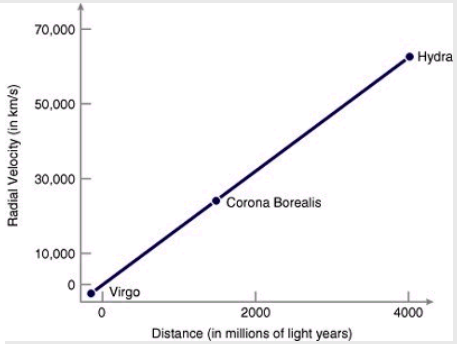
Knowledge:

Hubble Diagram →

Redshift-Brightness Correlation →

Doppler shift = Distance





Understanding: the Universe is expanding!!

Hubble's discovery

- For his discovery of that special correlation and for his remarkable insight regarding its meaning ... ***The Universe is Expanding*** ... we honor Edwin Hubble as one of the great observers (collectors and analyzers of astronomical data) of the 20th century, and... we named a certain famous telescope after him!
- *But what about all of the other correlations discovered and published every single day!?*

(1889-1953)

Statistics, Information Theory, and Data Mining

- We can mine our data for interesting correlations
- We can apply fabulous algorithms to discover these correlations
- We can do statistical tests on the significance of these correlations
- We can apply information theory to measure the information content (negative entropy) of the discovered patterns
- **But how do we know that the patterns and correlations are interesting, useful, scientifically meaningful, and a true Hubble-level discovery??**

Basic Astronomical Knowledge Problems – 1

- **The distance problem:**
  - Finding the distance to things on the “2-D” sky
  - We see everything in 2-D projection
  - But the Universe is deep in both space and time
  - We need distance to understand the physics and astrophysics of objects in space and time:
    - Space: Where are they? What are their neighbors?
    - Time: When did they form? How long do they live?
  - What observational parameters correlate with distance?
  - Are there combinations (linear or non-linear functions) of observed parameters that correlate more strongly with distance (i.e., what is the most accurate estimator)?
  - What is the most unbiased estimator for distance?

Basic Astronomical Knowledge Problems – 2

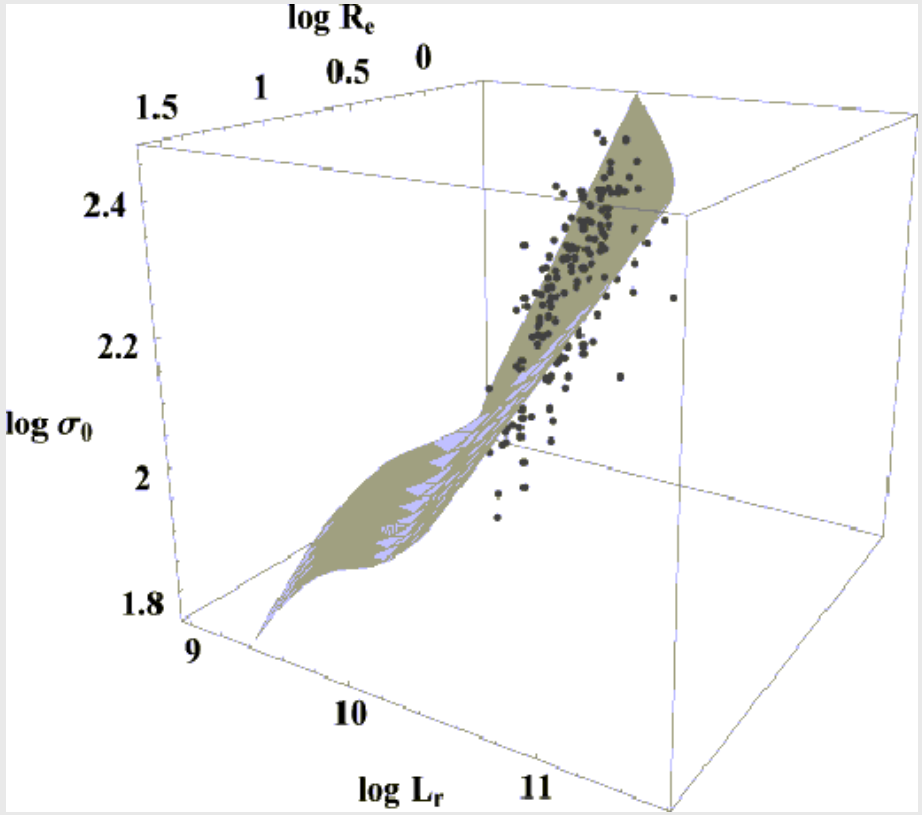
- **The clustering problem:**
  - Finding clusters of objects within a data set
  - What is the significance of the clusters (statistically and scientifically)?
  - What is the optimal algorithm for finding friends-of-friends or nearest neighbors?
    - N is $>10^{10}$, so what is the most efficient way to sort?
  - Are there pair-wise (2-point) or higher-order (N-way) correlations?
    - N is $>10^{10}$, so what is the most efficient way to do an N-point correlation?
      - algorithms that scale as $N^2 \log N$ won't get us there

Basic Astronomical Knowledge Problems – 3

- **Outlier detection: (unknown unknowns)**
  - Finding the objects and events that are outside the bounds of our expectations (outside known clusters)
  - These may be real scientific discoveries or garbage
  - Outlier detection is therefore useful for:
    - Novelty Discovery – *is my Nobel prize waiting?*
    - Anomaly Detection – *is the detector system working?*
    - Data Quality Assurance – *is the data pipeline working?*
  - How does one optimally find outliers in 10^3-D parameter space? or in interesting subspaces (in lower dimensions)?
  - How do we measure their "interestingness"?

Basic Astronomical Knowledge Problems – 4

- **The dimension reduction problem:**
  - Finding correlations and "fundamental planes" of parameters
  - Number of attributes can be hundreds or thousands
    - **The Curse of High Dimensionality !**
  - Are there combinations (linear or non-linear functions) of observational parameters that correlate strongly with one another?
  - Are there eigenvectors or condensed representations (e.g., basis sets) that represent the full set of properties?



Basic Astronomical Knowledge Problems – 5

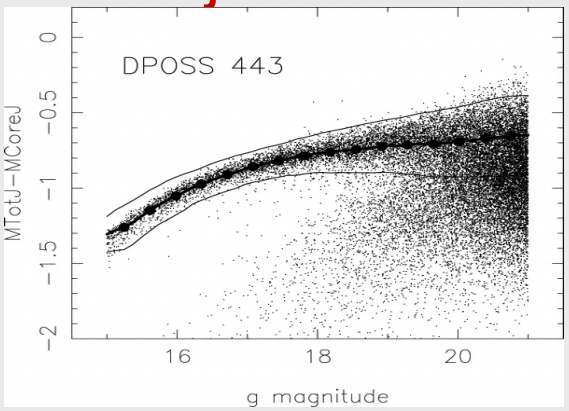
- **The cross-match problem:**
  - Matching objects in Catalog A to the corresponding objects in Catalog B
    - N is $>10^{10}$, so what is the most efficient way to proceed?
  - What is the likelihood function?
  - How do we include uncertainties in the scientific measurements?
  - How do include constraints from other information sources?
  - Objects are moving ... hundreds of them! ... Matching multiple observations of the same object is a challenge:
    - So what is the optimal solution (all objects cross-matched, maximizing the global likelihood in a massive data cube)?

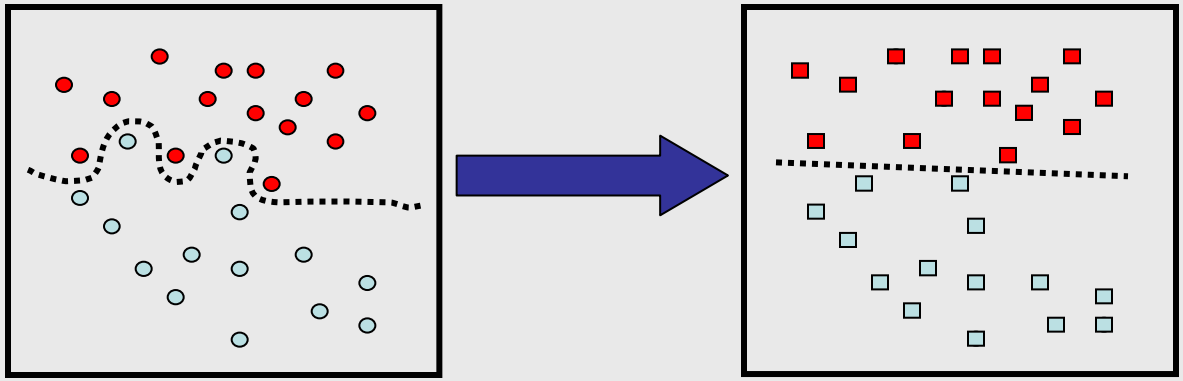
Basic Astronomical Knowledge Problems – 6

- **The classification problem:**
  - Classifying an object based upon observed attributes (using rules learned from the historical training data)
    - e.g., Star-Galaxy separation: very important problem !
  - There are dozens (hundreds?) of classification algorithms, so which algorithm is optimal when there are hundreds to thousands of attributes?
- The class discovery and sub-class discovery problem:
  - Are there new classes? Are there subclasses?
    - How do you discover them?
    - Which algorithms distinguish subclasses best? ...
      - SVM (Support Vector Machines), PCA (Principle Component Analysis), ICA (Independent Component Analysis), or ???

Basic Astronomical Knowledge Problems – 7

- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

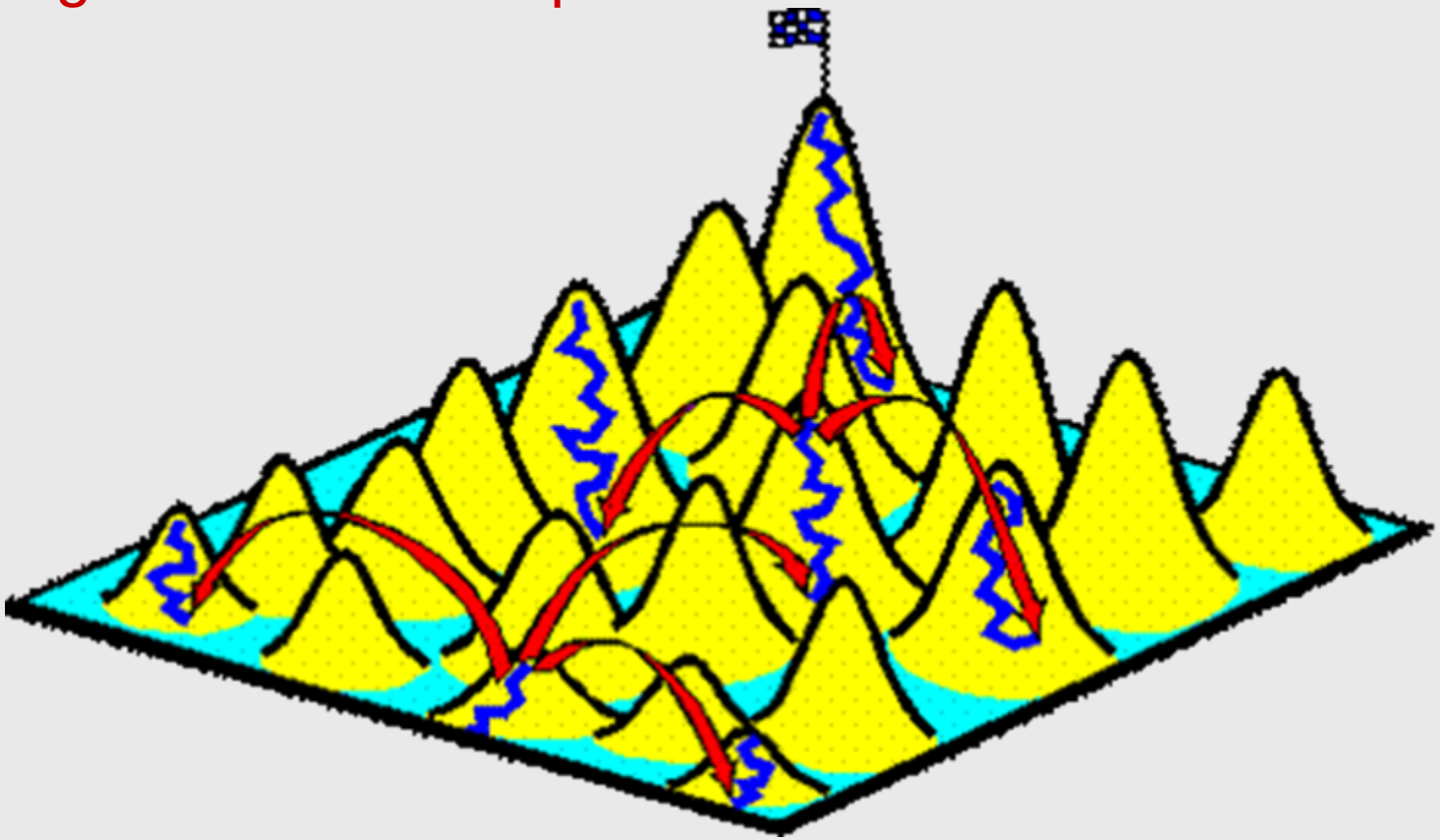




  - What if there are 10^{10} objects that overlap in a 10^{3}-D parameter space?
  - What is the optimal way to separate and extract the different unique classes of objects?
  - How are constraints applied (as in operations research or linear programming)?

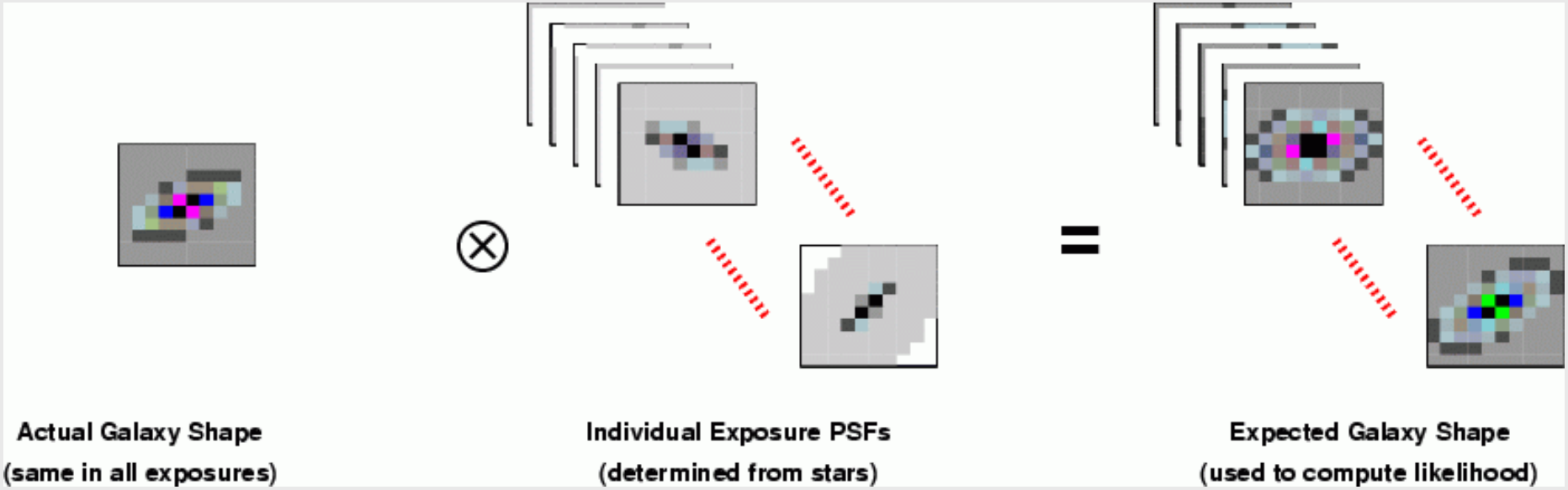
Basic Astronomical Knowledge Problems – 8

- **The optimization problem:**
  - Finding the optimal (best-fit, global maximum likelihood) solution to complex multi-variate functions over very high-dimensional spaces



Basic Astronomical Knowledge Problems – 8

- **Example of the optimization problem:**
  - Finding the optimal simultaneous solution for 100,000,000 objects' shapes across 2000 image planes, each of which has 201x4096x4096 pixels...10^{22} floating-point operations!
  - This illustrates an example for just one such object:



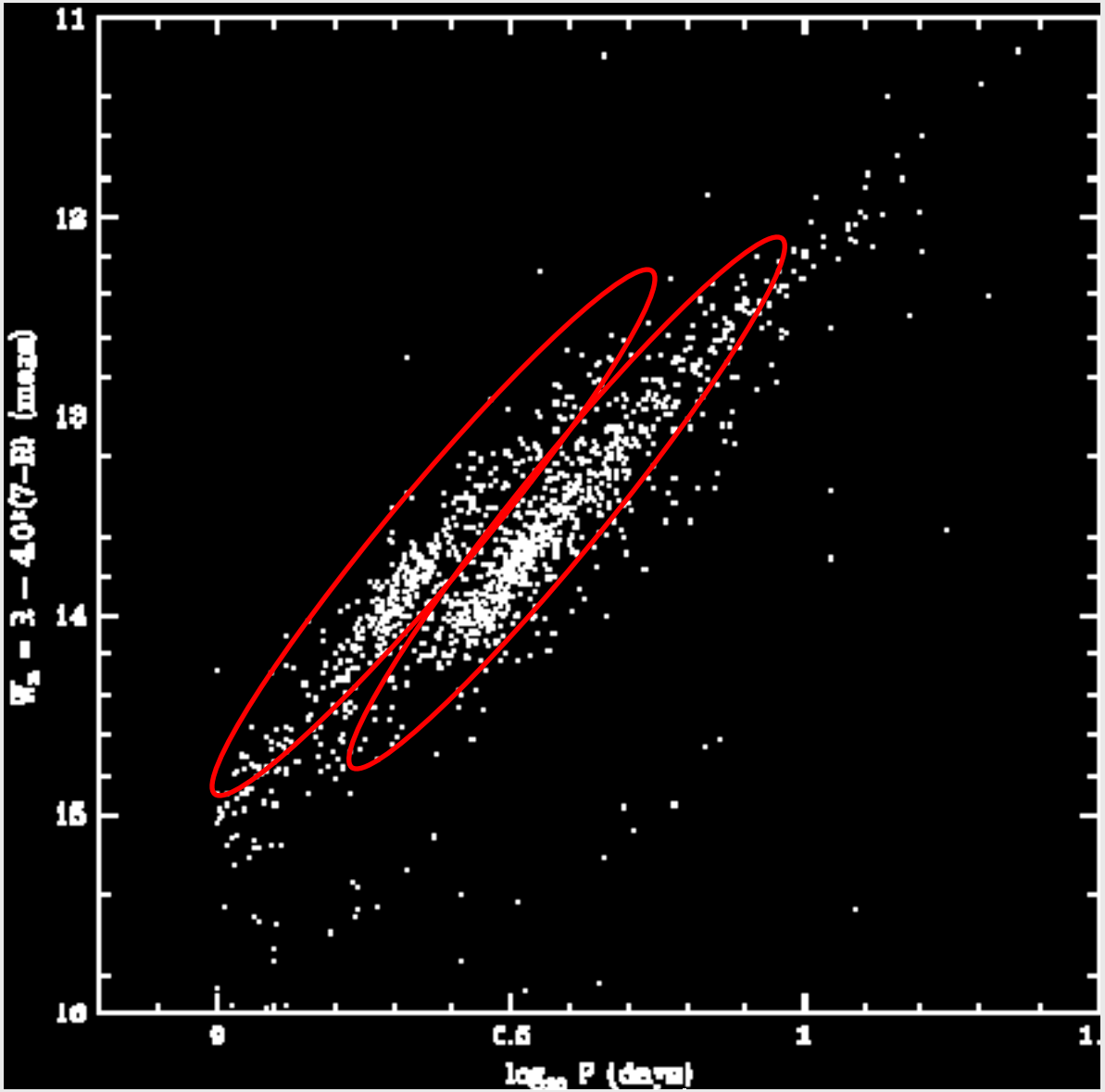
References:
http://universe.ucdavis.edu/docs/MultiFit-ADASS.pdf
http://code.google.com/p/multifit/

Example of Correlation, Classification, Clustering, Cross-matching, and Subclass Discovery

- Hubble found his correlation (and made his great discovery) by using Cepheid Variable Stars (= Cepheids)
- Cepheids are pulsating stars, varying in brightness – the class was discovered (and these stars are subsequently classified as Cepheids) because their brightness pulses in a very specific and predictable pattern, different from other variable stars.
- Cepheids show a period-luminosity (P-L) correlation: the bigger (and brighter) the star, then the longer the period of pulsation
- Typical Astrophysical Application: Get Distances to Galaxies ...

1. Find clusters of galaxies = at the same distance from us; they all have the same conversion factor between their apparent brightness and true absolute brightness
2. Find the variable stars in the galaxies – cross-match them with prior observations of the same stars – produce light curves, classify the variability (Cepheid or no?)
3. Apply P-L correlation to find absolute brightness → Distance !!
4. But what if the correlation is really 2 correlations? (2 subclasses!)

The problem with the prior analysis: there are 2 subclasses !



Cepheid Variables: Cosmic Yardsticks
-- One Correlation
-- Two Classes!

Consequently, Hubble got the wrong answers to the questions: What are the size and age of the Universe?

Sky Surveys: Partly the Solution and partly the Problem

As our chemistry friends say ....

- ***If you are not part of the solution, then you are part of the precipitate !***

Sky Surveys: Solution & Challenge

- To avoid biases caused by limited data, astronomers now study the sky systematically: Sky Surveys
- Surveys are used to measure and collect data from all objects that are contained in large regions of the sky, in a systematic, controlled, repeatable fashion.
- These surveys include (... this is just a subset):
  - Digitized Palomar Sky Survey: 3 Terabytes
  - 2MASS (2-Micron All-Sky Survey): 10 Terabytes
  - GALEX (ultraviolet all-sky survey): 30 Terabytes
  - Sloan Digital Sky Survey (1/4 of the sky): 40 Terabytes
  - and this one is just starting: Pan-STARRS: 40 **Peta**bytes!
- **Leading up to the big survey next decade:**
  - **LSST (Large Synoptic Survey Telescope): 200 Petabytes!**

LSST (Large Synoptic Survey Telescope)

http://www.lssto.org/

- Begin operations in 2015, with 3-Gigapixel camera
- One 6-Gigabyte image every 20 seconds
- 30 Terabytes every night for 10 years
- 200-Petabyte final image data archive anticipated – **all data are public!!!**
- **20-Petabyte final database catalog anticipated**
- Real-Time Event Mining: 10,000-100,000 events per night, every night, for 10 yrs
- Repeat images of the entire night sky every 3 nights: ***Celestial Cinematography***

Some Terminology

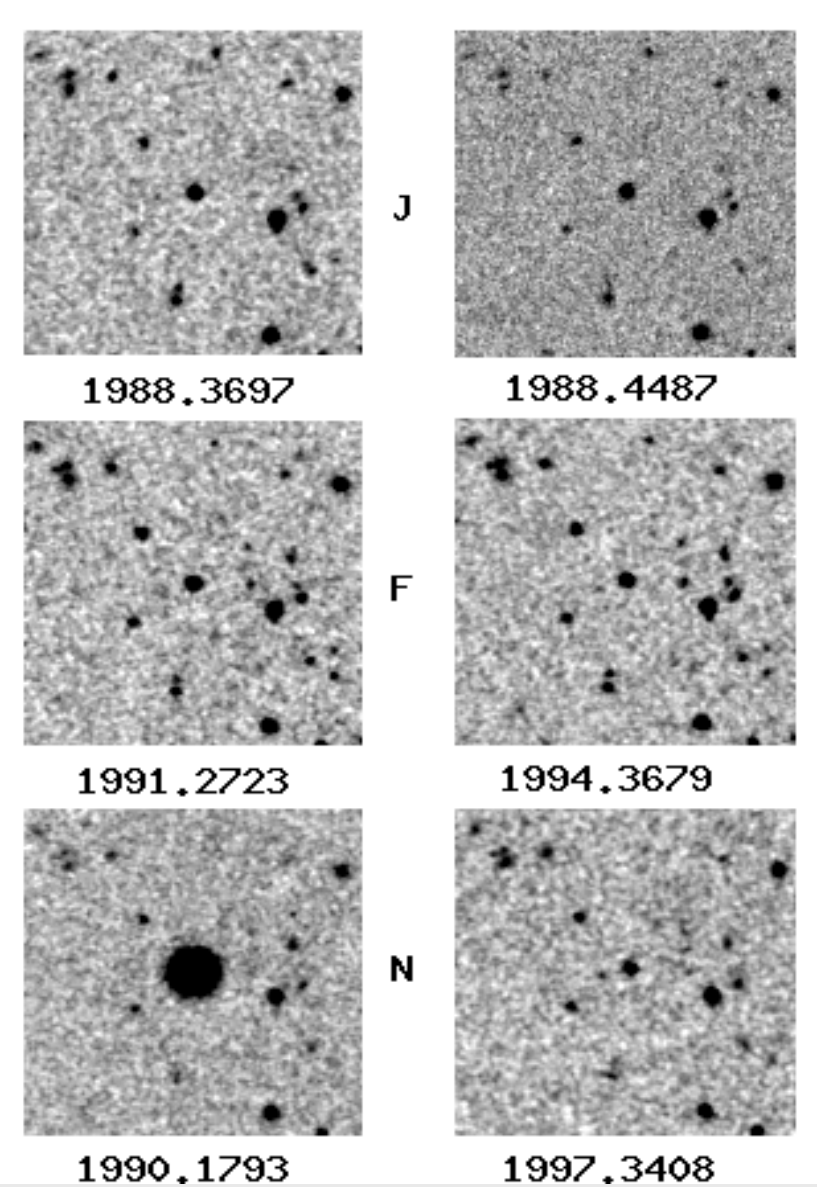
- LSST = Large Synoptic Survey Telescope
- HTN = Heterogeneous Telescopes Network: worldwide network of telescopes, most of which are robotic
- Event = an astronomical event discovered by any telescope anywhere
- VOEvent = a VO XML messaging protocol for alerting the world about the new event
- VOEventNet = a network of VOEvent providers and consumers (including HTN)

What is an Event?

- Anything that changes (motion or brightness)
- Variable stars of all kinds
- Optical transients: e.g., extra-solar planets
- Supernova
- Gamma-ray burst
- New comet
- New asteroid
- Incoming Killer Asteroid
- Anything that goes bump in the night

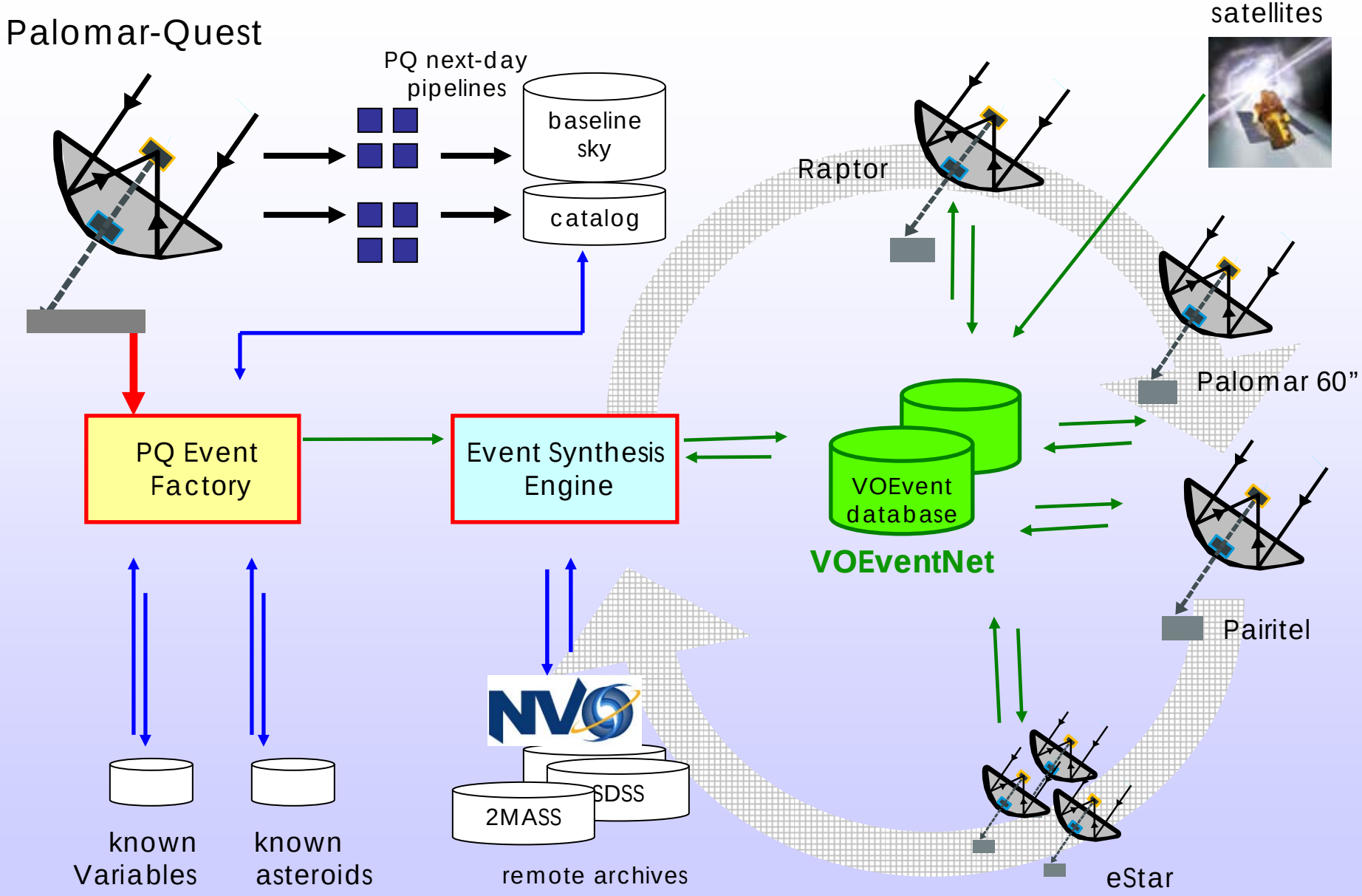
Here is one type of event ***

- **Optical Transient:** here today, gone tomorrow
- It is a normal dwarf star, similar to our sun, except …
- **it increased in brightness by 300x in one night** …
- and then returned to normal.



***Courtesy: Caltech / Palomar Quest Survey

VOEventNet: a Rapid-Response Telescope Grid



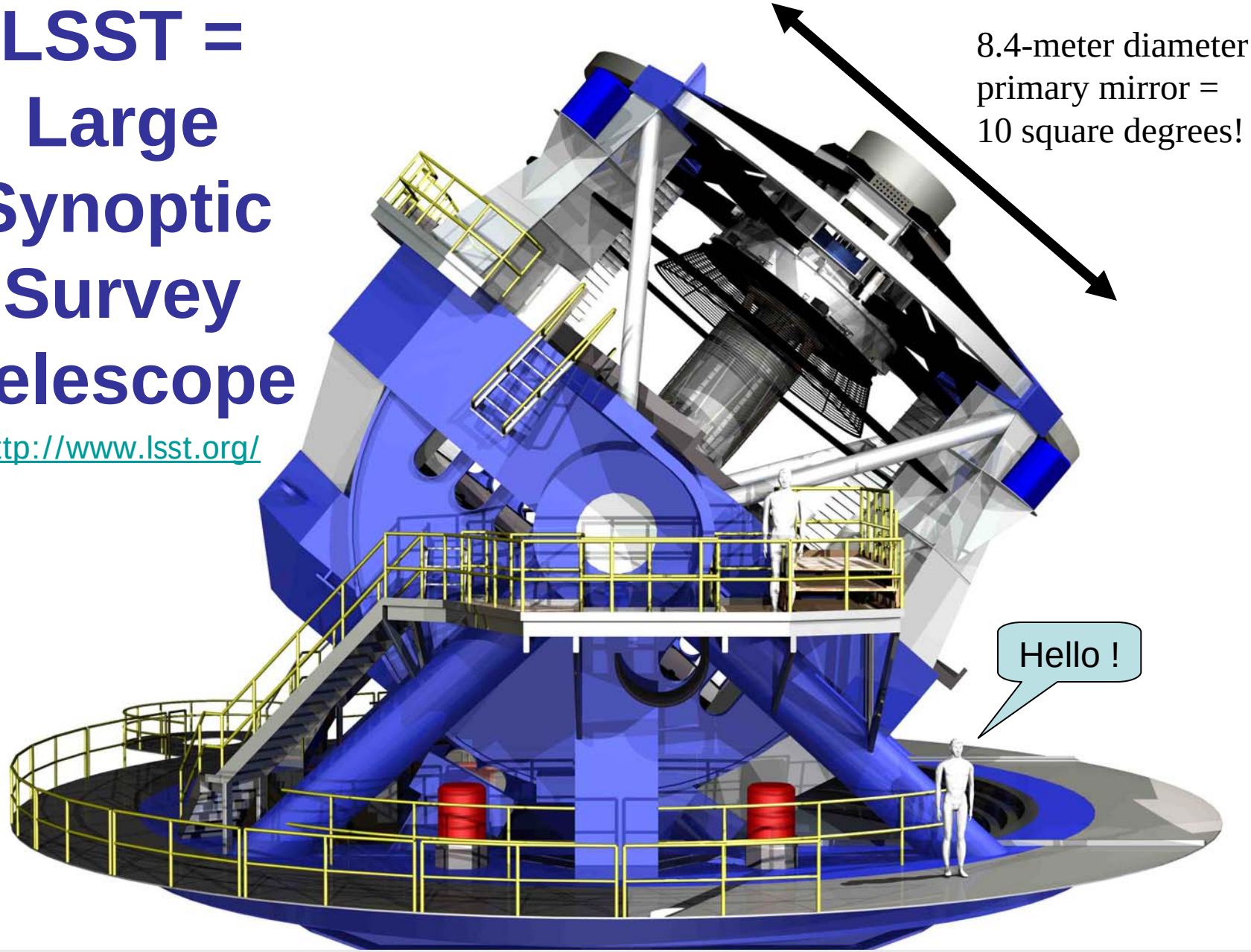
Reference: http://voeventnet.caltech.edu/

MIPS model for HTN / VOeventNet

- MIPS =
  - Measurement – Inference – Prediction – Steering
- HTN is a Global Network of Sensors:
  - Similar projects in NASA, Earth Science, DOE, NOAA, Homeland Security, NSF DDDAS (voeventnet)
- Machine Learning enables “IP” part of MIPS:
  - Autonomous (or semi-autonomous) Classification
  - Intelligent Data Understanding
  - Rule-based
  - Model-based
  - Neural Networks
  - Markov Models
  - Bayes Inference Engines

LSST = Large Synoptic Survey Telescope

http://www.lsst.org/



(design, construction, and operations of telescope, observatory, and data system: NSF) (camera: DOE)

LSST Key Science Drivers:

– Solar System Map (moving objects, NEOs, asteroids: census & tracking)
– Nature of Dark Energy (distant supernovae, weak lensing, cosmology)
– Optical transients (of all kinds, with alert notifications within 60 seconds)
– Galactic Structure (proper motions, stellar populations, star streams)

South America

Chile

Region de Coquimbo

Summit of Cerro Pachon -

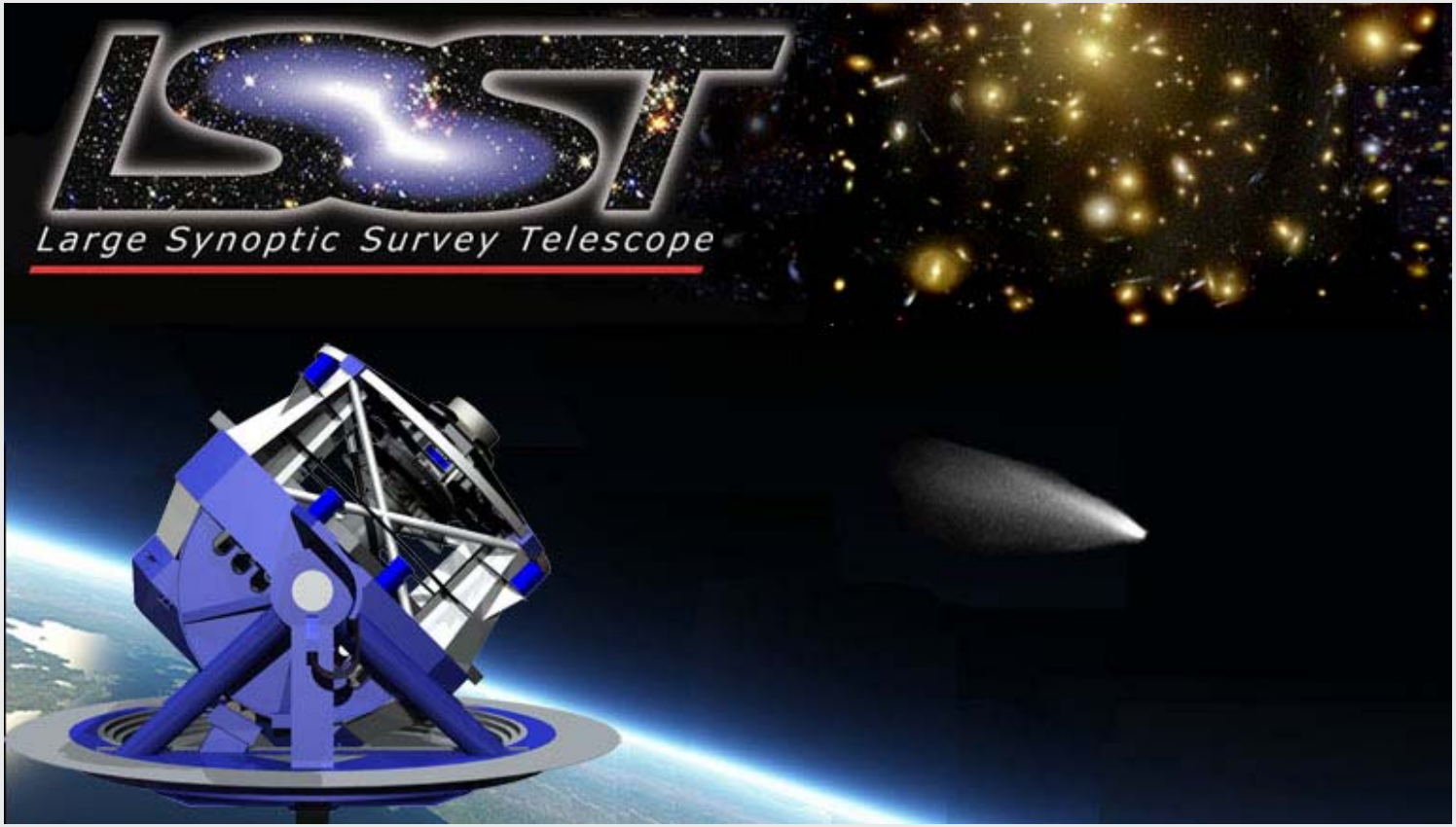
Model of LSST Observatory

<u>LSST in time and space:</u>

– When? 2015-2025
– Where? Cerro Pachon, Chile

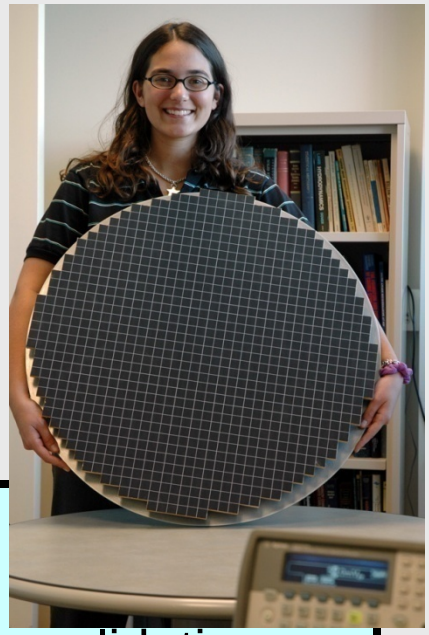
- **LSST (Large Synoptic Survey Telescope):**
  - Ten-year time series imaging of the night sky – mapping the Universe !
  - **100,000 events each night** – *anything that goes bump in the night !*
  - ***Cosmic Cinematography! The New Sky! @ http://www.lsst.org/***

Observing Strategy: One pair of images every 40 seconds for each spot on the sky, then continue across the sky continuously every night for 10 years (2015-2025), with time domain sampling in log(time) intervals (to capture dynamic range of transients).



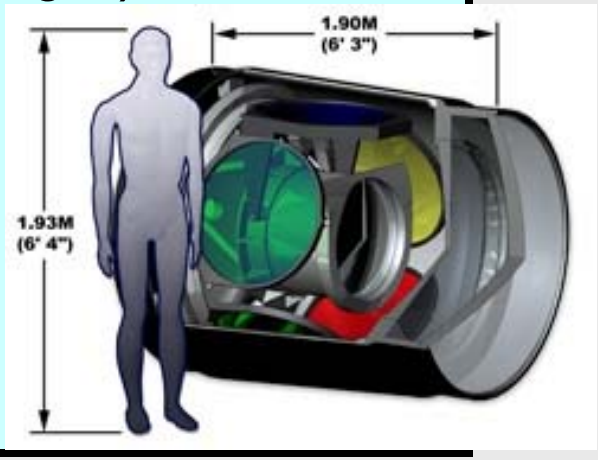
The LSST focal plane array

Camera Specs: **(pending funding from the DOE)**
201 CCDs @ 4096x4096 pixels each!
= 3 Gigapixels = 6 GB per image, covering 10 sq.degrees
= ~3000 times the area of one Hubble Telescope image



LSST Data Challenges

- Obtain one 6-GB sky image in 15 seconds
- Process that image in 5 seconds
- Obtain & process another co-located image for science validation within 20^s (= 15-second exposure + 5-second processing & slew)
- Process the 100 million sources in each image pair, catalog all sources, and generate worldwide alerts within 60 seconds (e.g., incoming killer asteroid)
- Generate 100,000 alerts per night (VOEvent messages)
- Obtain 2000 images per night
- Produce ~30 Terabytes per night
- Move the data from South America to US daily
- Repeat this every day for 10 years (2015-2025)
- Provide rapid DB access to worldwide community:
  - **100-200 Petabyte image archive**
  - **10-20 Petabyte database catalog**



The LSST will represent a 100,000-fold increase in the VOEvent network traffic and in the real-time classification demands:

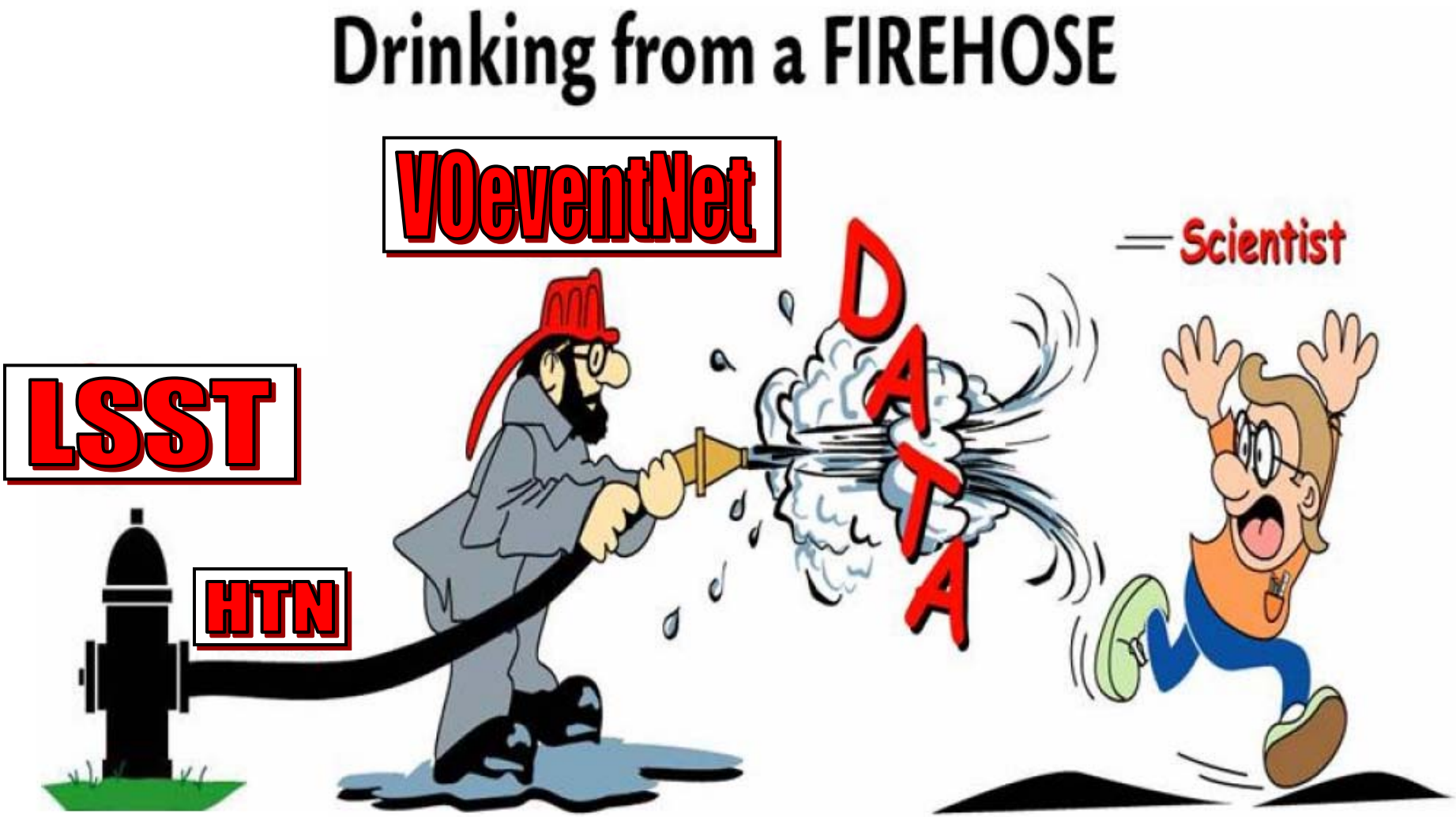
from data to knowledge!

<u>from sensors to sense!</u>

The LSST Data Challenge

- ~2 Terabytes per hour that must be mined in real time.
- More than 10 billion objects will be monitored for important variations in real time.
- Knowledge extraction in real time.

The LSST Data Flood and Event Flood



The LSST Petascale Challenges

(document is available on-line)

LSST Petascale Data R&D Challenges

Achieving scalability and reliability in LSST computing, storage, and network resources

The design of the DM system architecture is influenced by the technology. We expect to be available to implement it, starting with construction in 2011 – 2014 and continuing through the principal survey period until 2024. This technology includes not only more powerful components, but completely new system architectures and potentially disruptive technologies. Most computing throughput improvements will come not from increased CPU clock speeds as in the past, but from larger concentrations of CPUs/cores and advanced computing architectures. Solid state technology may change storage and the way we physically organize data. Hardware failures will be routine for the LSST data system due to the large number of CPUs and disk drives, and reliance on high-speed network connectivity. It is a challenge to create a system sufficiently robust to these failures. We need to predict the characteristics of CPU, network, storage hardware, and system software sufficiently well that our design is appropriate. Further, we need to insulate the design as much as possible from underlying platform dependencies.

Reliability and performance issues for very large databases

LSST's main data products from the 20,000 square degree survey with 2000 images over ten years per patch of sky are in the form of relational database tables. These tables are very large (50 billion rows in the Object table, 600 billion rows in the Source table). They must be extensible, and partitioned and indexed to facilitate high query performance, and replicated across multiple centers. Queries in the time domain (Source table) are likely to be of equal importance to those in the spatial domain. Since these are traditionally optimized by different database organizations, it is unclear what choices will perform best for LSST. Some intensive applications will involve n-point correlations of object attributes over all objects. All these factors suggest that database performance and reliability are risk areas.

Efficient automated data quality assessment

LSST will produce large volumes of science data. The Data Management System (DMS) produces derived products for scientific use both during observing (i.e. alerts and supporting image and source data) and in daily and periodic reprocessing. The periodic reprocessing also results in released science products. Analysis of the nightly data will also provide insight into the health of the telescope/camera system. An automated data quality assessment system must be developed, which efficiently searches for outliers in raw image data and unusual correlations. This will involve aspects of machine learning.

Operational control and monitoring of the DMS

The DMS will be a complex distributed system with enormous dataflows that operates 24/7. The DMS must be continuously monitored and controlled to ensure the proper functioning of all computing hardware, network connections, and software, including the data quality of the science pipelines. Most of the monitoring tasks, and some of the control tasks, must be highly automated, since the data volumes preclude human examination of all but a tiny fraction of the data.

Achieving acceptably low False Transient Alert Rate

The science mission places high demand on the LSST's ability to rapidly and accurately detect and classify varying and transient objects and to achieve a low false alarm rate. Given the very high data volume produced by the LSST, the corresponding large number of detections in each image (up to one million objects detected per image), as well as the likelihood of entirely new classes of transients, the LSST will not be able to rely on traditional labor-intensive validation of detections, classifications, and alerts. To achieve the levels of accuracy required, new algorithms for detection and classification must be created, as well as innovative automated techniques for alert filtering and validation.

Efficient detection and orbit determination for solar system objects

One of the LSST's science missions is to catalog the population of solar system objects, with a particular focus on potentially hazardous objects. Due to the depth of LSST's images, about 300 solar system objects per square degree will be detected near the ecliptic. The LSST cadence on the sky is not optimized solely for tracking solar system objects, so this dense swarm of objects must be reliably tracked through considerable gaps in time. Algorithms must be developed that are robust to possible mis-associations of detections at different epochs, and have acceptable computational scalability.

Achieving required photometric accuracy and precision

The LSST Science Requirements Document (SRD) requires a level of photometric (intensity data) accuracy and precision that may be difficult to achieve over the entire sky, particularly since the LSST will be operating in a wide variety of seeing, sky brightness, and atmospheric extinction. To achieve this requires a thoroughly tested calibration procedure and associated image processing pipeline. In addition to the point-source requirements in the SRD, accurate photometric redshifts require precision photometry for spatially extended objects.

Achieving required astrometric accuracy and precision

The LSST SRD requires a level of astrometric (position on the sky) accuracy and precision that is difficult to achieve over the entire sky. Achieving this astrometric performance requires a global, whole-sky, numerical solution for all per-frame astrometric quantities that minimizes a cost function. Considerable work will be required to develop an effective cost function.

Achieving optimal object detection and shape measurement from stacks of images

Most objects that will be used for dark matter and energy science are too faint to be usefully measured in a single LSST exposure. Instead, the LSST must detect and measure the properties of objects combining information from multiple exposures of the same region of sky (image stacks). Weak lensing galaxy shape measurements are particularly vulnerable to systematic effects introduced by errors in the local point-spread function (PSF) determination, and these systematic effects must be minimized. Exposures may vary significantly in their signal-to-noise and PSF quality, and defining how to optimally combine information from all of them is a research problem. See http://universe.ucdavis.edu/docs/MultiFit-ADASS.pdf for more information.

Need to develop a flexible approach that enables highly reliable classification of objects

Classification of astronomical objects is important and difficult. A wide variety of information must be assessed to reliably classify an object. This includes spatial morphology in multiple colors, photometry in multiple colors, time dependent behavior, and astrometric motion. Further, the best classifications will make use of surveys in other wavelength regimes and spectral information where available, not solely information from the LSST. Experience from many surveys has shown that no single algorithm can do a good job on all objects. Rather, good algorithms tend to be specialist, limited to particular objects classes, e.g. eclipsing binaries or supernovae. A successful system must allow the development and incorporation of a wide variety of algorithms in a flexible manner.

Adaptive retuning of algorithm behavior

Several key algorithms employed in the LSST application pipelines are complex, containing many data-dependent decisions and a large number of tuning parameters that affect their behavior. As observing conditions change, an algorithm may begin to fail for a particular choice of tuning parameters. LSST's extremely large data volume makes human intervention in such cases impractical, but it is essential that the pipelines continue to function successfully.

Need to verify scientific usefulness of the LSST database schema and its implementation against realistic queries

The LSST database schema must efficiently support queries of data that have many relationships between multiple locations on the sky, epochs of observation, and filters employed. A high performance implementation of this schema has many complexities that are addressed in the peta-scale database architecture and analysis task. The ultimate test of how well these tasks have been carried out is to perform science with the database. To do this usefully, we are simulating LSST data, using data from current surveys, and engaging the LSST Science Collaborations and scientific community.

http://universe.ucdavis.edu/docs/LSST_petascale_challenge.pdf

The LSST Data Challenges

MANAGING AND MINING THE LSST DATA SETS

Astronomy is undergoing an exciting revolution -- a revolution in the way we probe the universe and the way we answer fundamental questions. New technology enables this: novel detectors are opening new windows on the universe, creating unprecedented volumes of high quality data, and computing technology is keeping up with this explosion. In turn, this is driving a shift in the way science is produced in astronomy and astrophysics: huge surveys of the sky over wide wavelengths can be analyzed statistically for low-level correlations and inverse problems may be solved by statistical inversion, producing new understanding of the underlying physics.

This parallels progress in high energy physics. Decades ago, a handful of photographs of events sufficed for ground-breaking discoveries. This gave way to experiments in which the systematic measuring (scanning) of many bubble chamber pictures allowed the measurement of statistical properties, such as lifetimes. Current experiments extend the technique by recording all events electronically and subjecting Petabyte data sets to rigorous statistical analysis.

A key ingredient in mining our astronomical science from such huge databases, efficient algorithms for statistical analysis, has been under-emphasized in the rush to utilize new technology and get the data products out to the science community. Past data sets in astronomy (and indeed in most areas of science) have been small enough that one individual could visualize the data and discover unanticipated correlations. This is often how major discoveries have been made. Data sets are now becoming sufficiently large that this is less possible -- even prescribed processing of the data to test a hypothesis is becoming challenging. In the near future, analysis of Petabyte databases will require the solution of this problem.

New Horizons

It is worthwhile to briefly review this sea-change in the way astronomers produce science. A giant departure from the tradition of one astronomer and one modest data set per project has been the *Sloan Digital Sky Survey*: a 15TB imaging data set covering multiple wavelengths and up to 10,000 square degrees of the sky (http://www.sdss.org/). Nearly 100 Co-Is will mine these data in prescribed ways. Current plans do not include mining the 15TB. Rather, 1TB of catalogs of detected objects and another 2TB of their "cutout" pictures will be produced and mined. Nevertheless, this will surely result in new understanding of our universe. Imagine what might be discovered if the full 15TB could be explored efficiently! Another refreshing and very successful departure from tradition is the *2MASS* infrared survey of the sky (http://irsa.ipac.caltech.edu). This group has poured major effort into usability of the data products and efficient remote searching.

A New Collaboration

We see this research program attracting a broad range of mathematical, computer and physical scientists. In addition to the obvious connections to astronomy, statistics and large-scale computation, this program would also include probability, data visualization and data management. We would also seek to include representatives from the high energy physics community, who have faced somewhat different problems involving massive data sets and immense data streams for many years now. Some representation from theoretical cosmologists who simulate universes would add to the mix and allow the question of comparing simulated universes to the actual universe to be more profitably addressed.

It will be particularly useful to study the characteristics of spatial processes, since it nicely combines the central computational and statistical challenges. Very little work has been done to date in this area, although a recent paper by Moore et al. (2001) recognizes the importance of this problem and describes an algorithm for computing estimates of higher order correlation functions that, for sufficiently large data sets, is much more efficient than the obvious approach.

We need not simply a theoretical study of how massive astronomical data sets should be analyzed, but major efforts to analyze the most recently available data sets. Data from the Sloan Digital Sky Survey should be publicly available by 2003. It will be useful to work with this database in new ways, searching for low-level correlations. Deeper imaging surveys, such as the Deep Lens Survey, are producing imaging data and catalogs nearly to the depth that LSST will reach, but over a very small area of sky by comparison to a decade of LSST operations. Such surveys are precursors to LSST and their data products will prove to valuable sand boxes for development of new algorithms.

A common technique in modern high-energy physics experiments is the "mock data challenge." The data stream, from detector, through data acquisition and processing, to final science analysis, is simulated at the appropriate level of detail. This allows a final acceptance testing of all data systems to be completed along with the hardware, so that full-up science operations can begin on a much better schedule, with good diagnostics in place. For the science, these studies are just as important. Analysis teams combing for subtle effects can, in then end, compare their result (and error estimate) with the "true" values of parameters that were in the simulation. Often, a sample of "real" data is used to get the background distribution of events correct. Using catalogs from the *SDSS* and the Deep Lens Survey as a basis for the mock data challenge for the *LSST* will make it more effective.

http://universe.ucdavis.edu/docs/data-challenge.pdf

Astronomical Challenge Areas not addressed here:

Petascale Computational Models of Astrophysical Phenomena

- Numerical simulations will generate Petabytes of data output.
- In many cases, multiple generations of models will be simulated to test against telescopic observational data.
- This model-to-data (theory-to-observation) matching leads to model validation and new scientific understanding ...
  - but it also leads to enormous combinatorial challenges, as the number of input parameters and model output parameters continue to grow as our models are refined.
- Petascale computational challenges exist in the areas of Monte Carlo sampling, optimization, maximum likelihood estimation, and computational steering.
- As the numerical simulations grow in size and complexity, we also seek faster methods of numerical integration and differentiation over densely gridded data structures.

Summary: Data Science Research Challenge Areas in Astronomy over the next 10 years

- Scalability of statistical, computational, & data mining algorithms to multi-petabyte scales
- Algorithms for optimization of simultaneous multi-point fitting across massive multi-dimensional data cubes
- Multi-resolution, multi-pole, fractal, hierarchical methods and structures for exploration of condensed representations of petascale databases
- Petascale analytics for visual exploratory data analysis of massive databases (including feature detection, pattern recognition, correlation analysis, clustering, decomposition, eigenvector discovery, dimension reduction)
- Indexing and associative memory techniques (trees, graphs, networks) for highly-dimensional petabyte databases
- Rapid query and search algorithms for petabyte databases (cf. think of Google's *PageRank* statistic) (e.g., *Query By Example*)