A Machine Learning Classification Broker for Petascale Mining of Large-scale Astronomy Sky Survey Databases

Kirk Borne

George Mason University

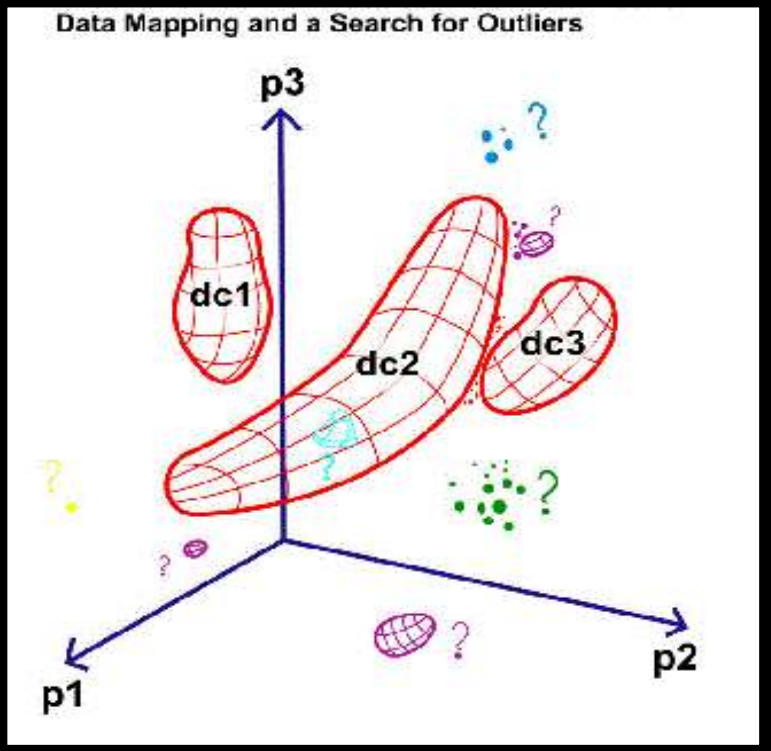
Astronomers have been doing Data Mining for centuries

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

- Seriously ...
- Astronomers love to classify things ... (Supervised Learning. eg, classification)
- And we love to discover new things ... (Unsupervised Learning. eg, outlier detection)

This sums it up

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



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

The Changing Landscape of Astronomical Research

- Astronomy is now a data-intensive science
- Astronomy will become even more data-intensive in the coming decade
- **Astroinformatics** (data-intensive astronomical research) will become a stand-alone research discipline (similar to Bioinformatics, Geoinformatics)
- Astronomical data are now accessible from distributed heterogeneous sources through standalone interfaces = the Virtual Observatory (e.g., the US NVO: National Virtual Observatory)

The Astronomical VO = Virtual Observatory

[www.ivoa.net]

- Middleware for discovery, access, integration, mining, and analysis of distributed information sources *("it's the middleware")*
- Based on Web Services (e-Science) paradigm
- Includes XML-based data-sharing protocols and data models for images, catalogs, time series, new events, spectra
- Links together multiple astronomical object catalogs, Sky Surveys, any data anywhere

Mining the VO

- Data Mining the VO can take two forms:
  - Distributed Data Mining (1)
  - Distributed Data Mining (2)
- In other words ...
  - Distributed Mining of data (1)
  - mining of Distributed Data (2)
- (1) mines the data *in situ* with distributed DDM algorithms
- (2) mines the data with standard DM at a central site where the data are collected

DDM Results

- Team: H.Kargupta, H.Dutta (PhD thesis), C.Giannella, R.Wolff, K.Borne
- Distributed PCA for fundamental plane (hyperplane) discovery – rediscovered fundamental plane for Elliptical Galaxies
- Distributed Outlier Detection through analysis of last PCA component – identified potential outliers for analysis as new astronomical discoveries
- See papers in SIAM DM proceedings ’06, ’07

DDM Scalability Issues

- Results so far are based on small numbers (10-100K tuples, a few attributes)
- What happens when you have …
  - hundreds of attributes?
  - millions of tuples?
  - billions of tuples?
  - trillions of tuples?
  - Petabyte-scale databases?

- This is coming soon, in the form of **LSST** = Large Synoptic Sky Survey [www.lsst.org]

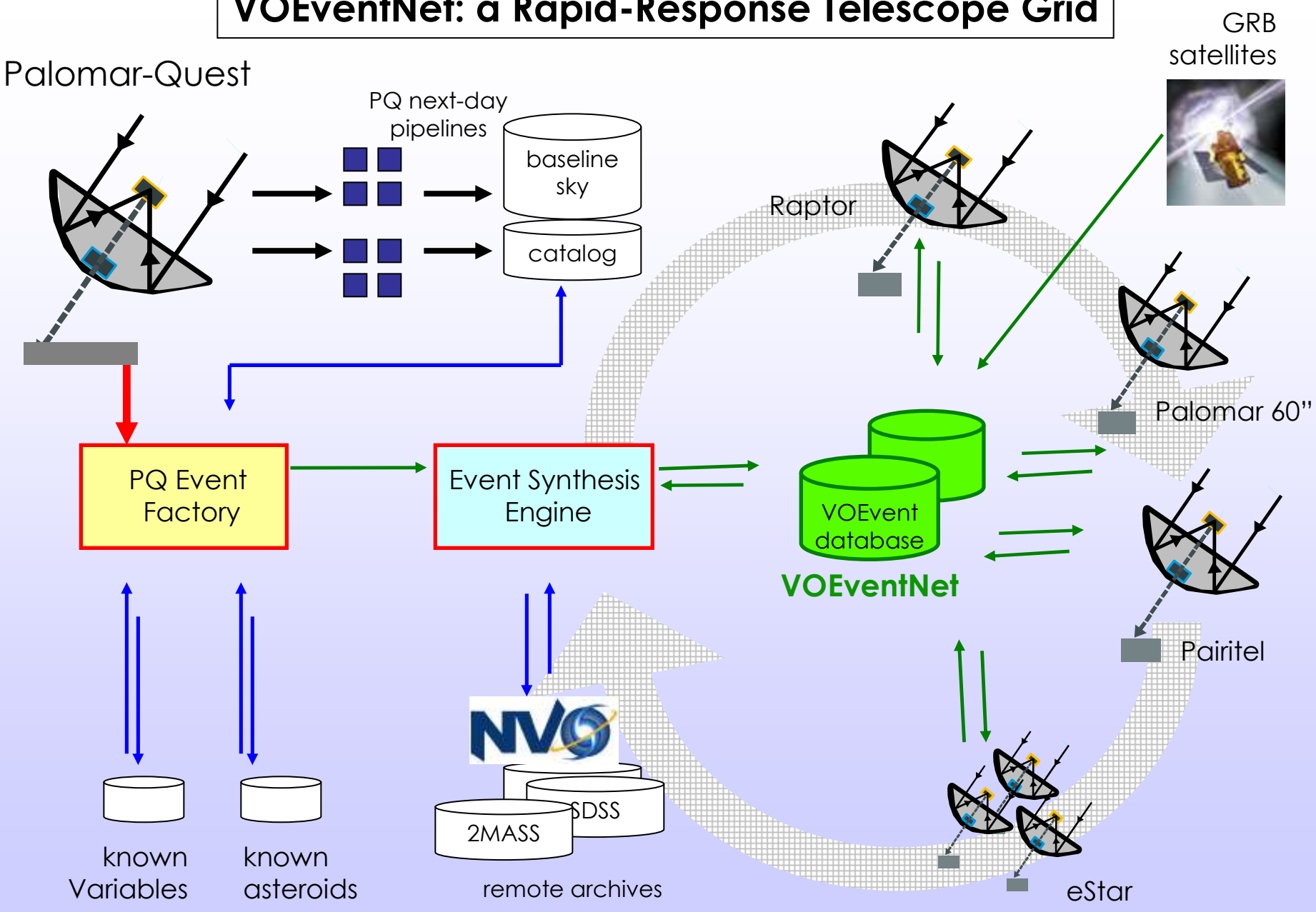
Terminology

- LSST = Large Synoptic Survey Telescope
- HTN = Heterogeneous Telescopes Network: a 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

VOEventNet: a Rapid-Response Telescope Grid

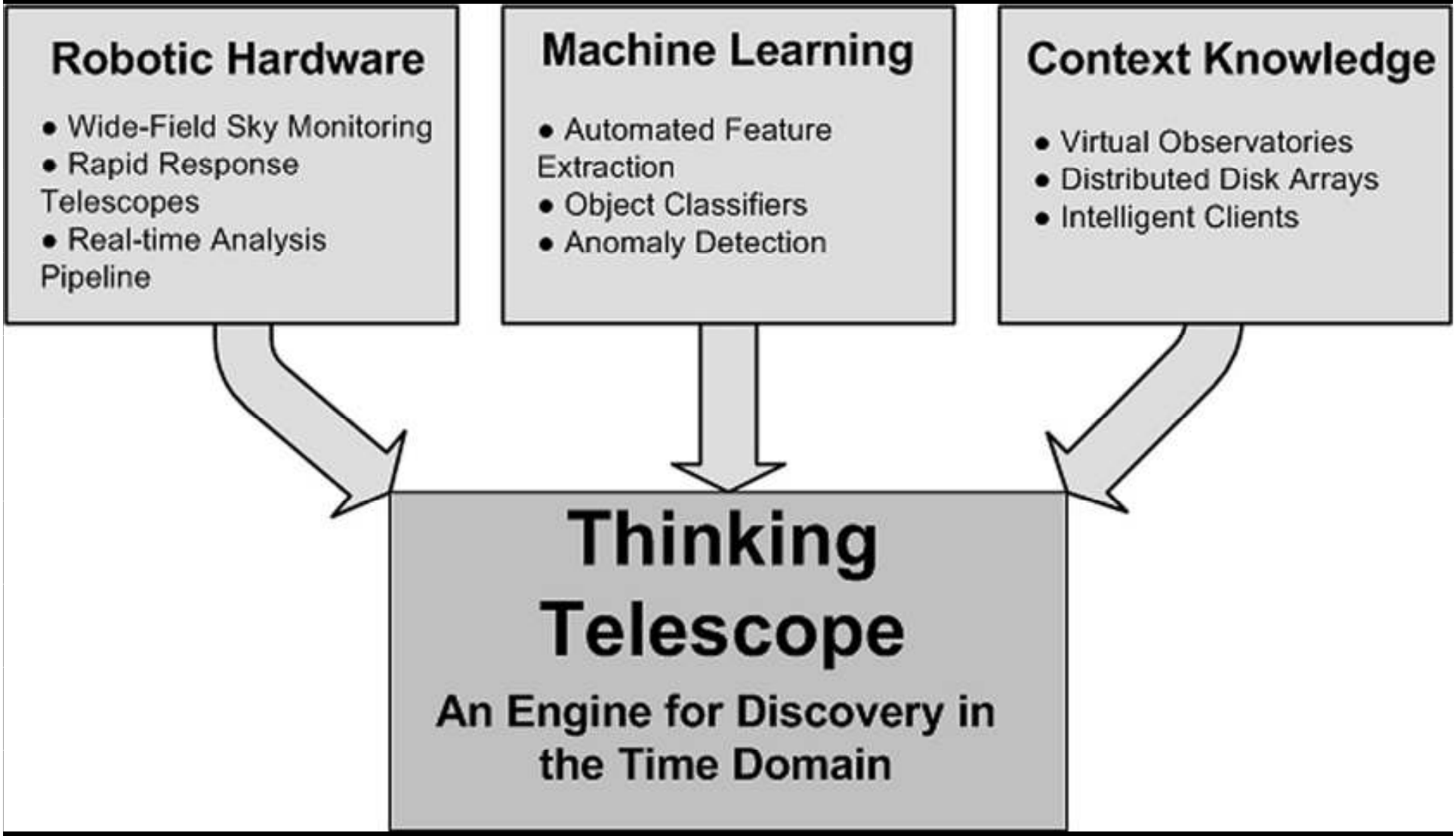


Reference: http://voeventnet.caltech.edu/

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

Example: The Thinking Telescope



Reference: http://www.thinkingtelescopes.lanl.gov/

Machine Learning Classification Broker

Robotic Hardware

- Wide-Field Sky Monitoring
- Rapid Response Telescopes
- Real-time Analysis Pipeline

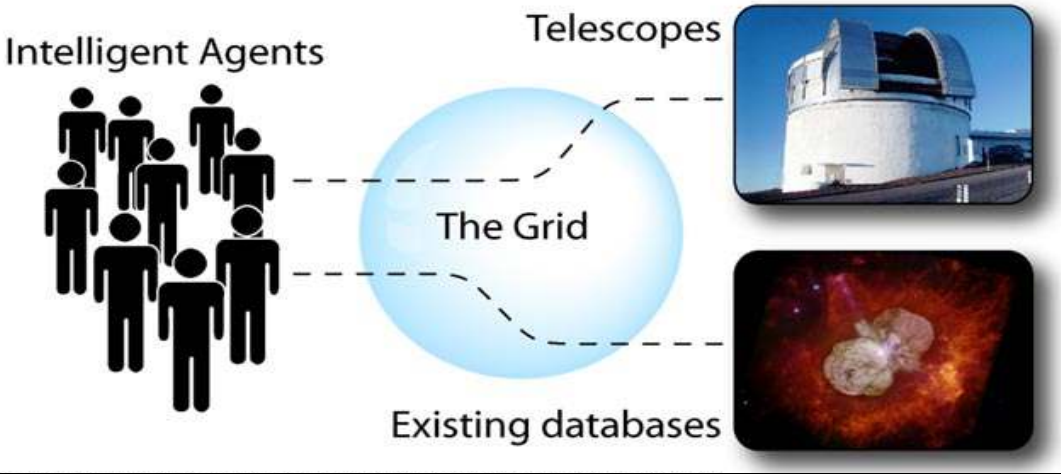
Machine Learning

- Automated Feature Extraction
- Object Classifiers
- Anomaly Detection

Context Knowledge

- Virtual Observatories
- Distributed Disk Arrays
- Intelligent Clients

From Data to Knowledge

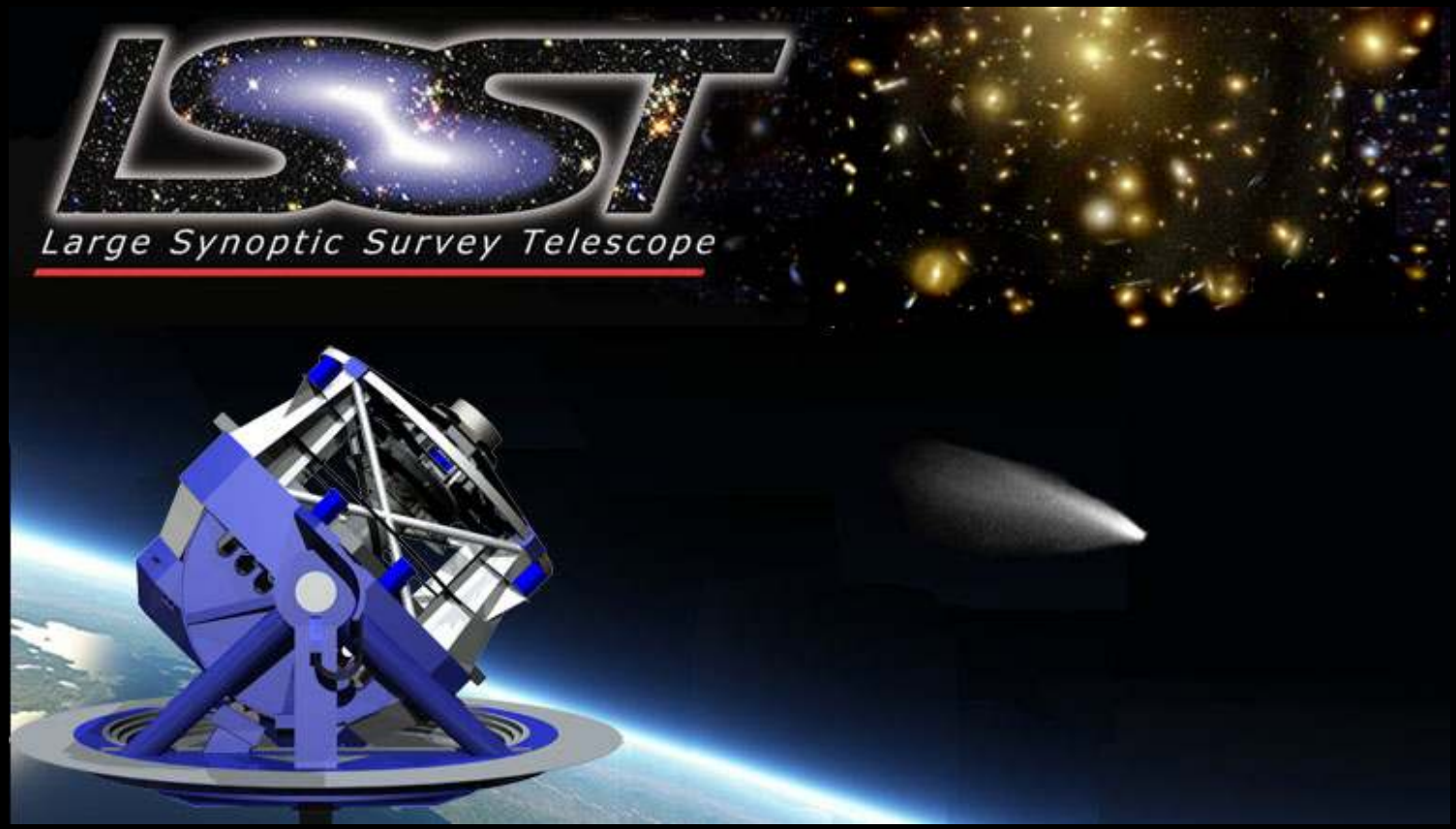


Reference: http://www.estar.org.uk/

The LSST will represent a 100,000-fold increase in the VOEvent network traffic and in the machine learning demands!!

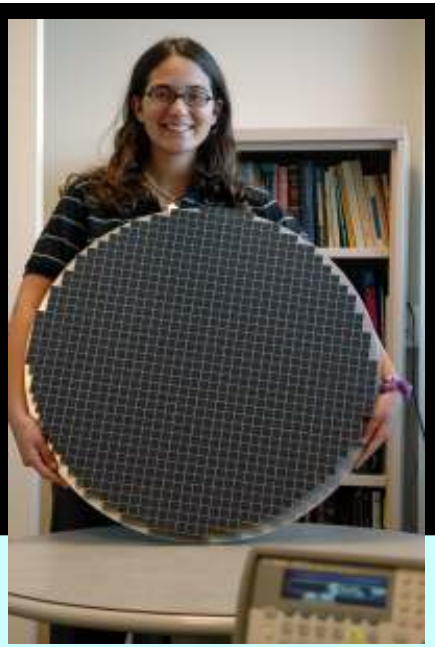
- **LSST (Large Synoptic Survey Telescope):**
  - Ten-year time series imaging of the night sky
  - **100,000 events each night** – *anything that goes bump in the night!*
  - ***Cosmic Cinematography! The New Sky! @ http://www.lsst.org/***

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 (2014-2024), with time domain sampling in log(time) intervals (to capture dynamic range of transients).



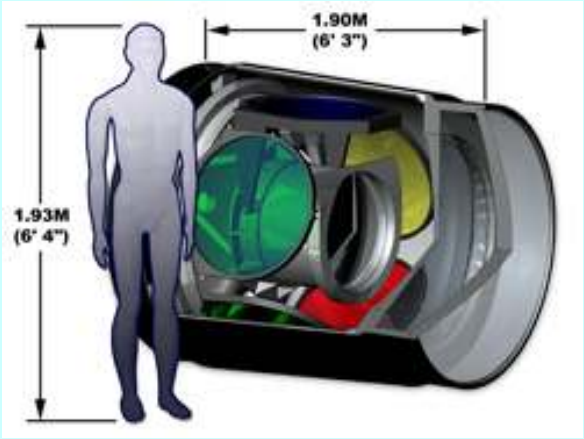
The LSST focal plane array

Camera Specs: 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 every 15 seconds
- Process 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
- Obtain 2000 images per night
- Produce ~30 Terabytes per night
- Move the data from South America to SDSC daily
- Repeat this every day for 10 years (2014-2024)
- Provide rapid DB access to worldwide community:
  - **60-100 Petabyte image archive**
  - **10-20 Petabyte database catalog**



The LSST Data Flood and Event Flood

Machine Learning Classification Broker for Petascale Mining

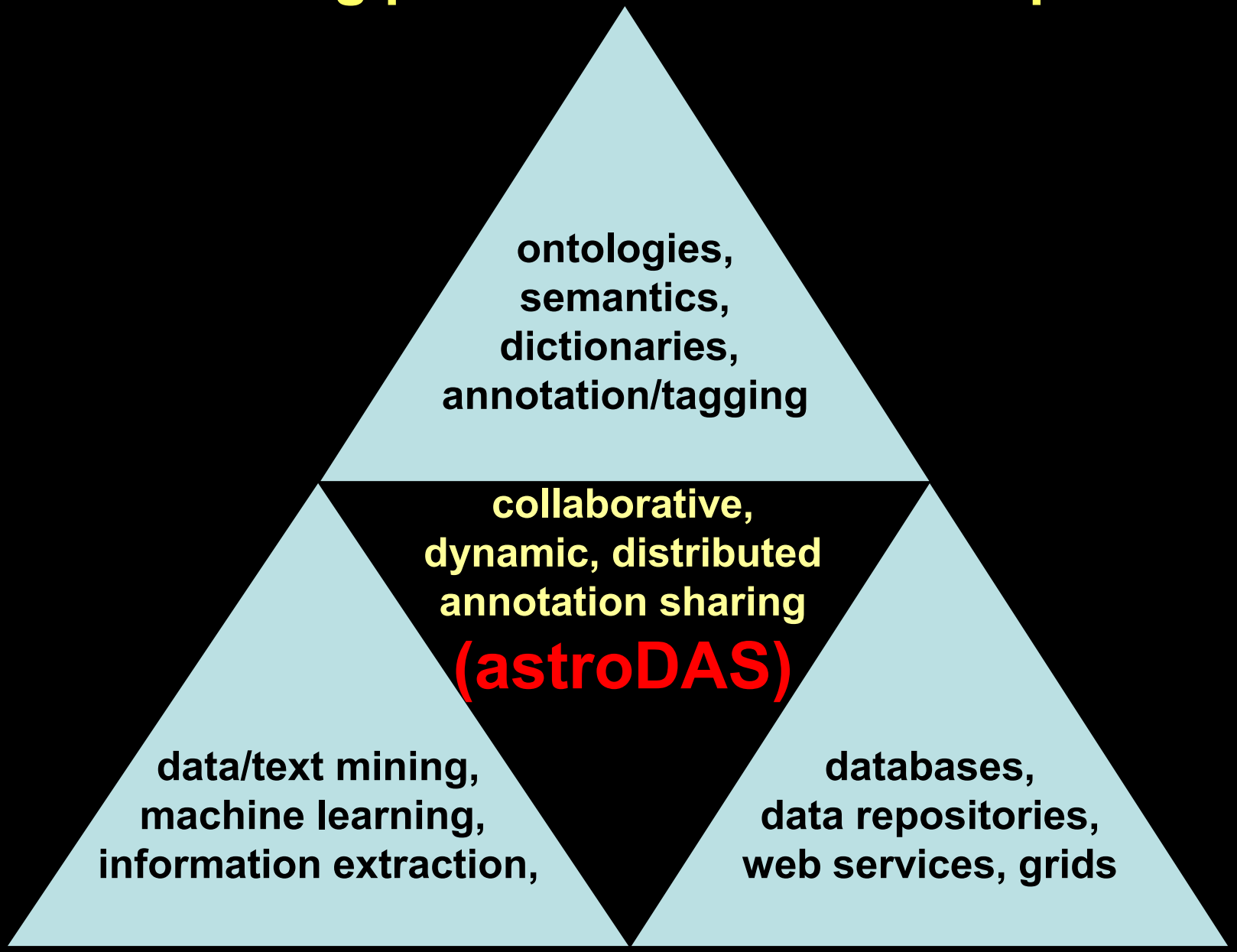
- LSST will produce 100,000 events per night
- Astronomers worldwide will want to know what these are (classifications)
- HTN will need to know what priority to assign (probabilistic classifications)
- DDM: Information, data, catalog attributes from the VO will be integrated with real-time information from telescopes and humans to classify and characterize new events
- DAS provides one implementation

What is DAS? (Distributed Annotation System)

- Annotation is *"Extra information associated with a particular point in a document or in a program or in the sky!"*
- Annotation provides user feedback (follow-up) on existing content. (compare with biodas.org, Wikiproteins)

  The LSST Challenge = 10^5 events/night, for 10 years!!!!
- DAS provides reporting mechanism for scientists (pro-am-EPO) to mark up scientific databases
- astroDAS: DAS for astronomy
  - Applicable to VOEvents or any other astronomical database (e.g., large sky survey object catalog)
  - Mechanism for reporting follow-up observations
  - Provides Classification Broker for Petascale Mining of Large Astronomical Sky Surveys
  - Think ... " Web 2.0 Mashups ".

astroDAS: big picture of the main components



LSST: Bringing the Dynamic Universe to the Public = **Data Mining for Everyone**

- Strong public interest in the dynamic sky:
  - Asteroids: motion, variability, colors, incoming!
  - Comets
  - Novae & Supernovae
  - Variable stars
  - Anything that "goes bump in the night"
- Strong desire to be involved by:
  - Amateurs (AAVSO)
  - Teachers / Students / Classrooms:
    - "Using Data in the Classroom" initiative
  - Citizen scientists (museums, planetaria, @home)

Forget about ...

- CSI Las Vegas
- CSI Miami
- CSI New York

- It’s time for data mining to go public

- Introducing ....

CSI Astronomy: "No doubt about it ... an asteroid killed the dinosaurs!"