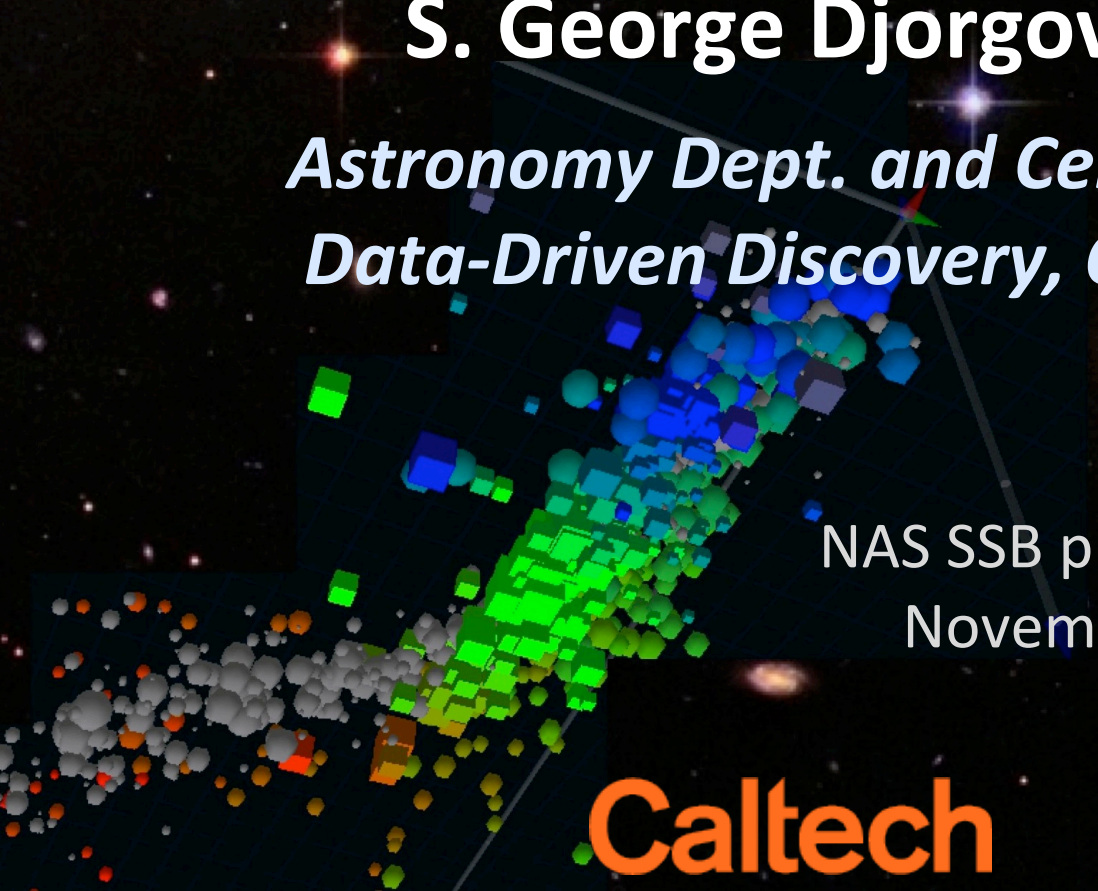
Machine Learning and Other Data Science Methodology in Astronomy

S. George Djorgovski

Astronomy Dept. and Center for Data-Driven Discovery, Caltech

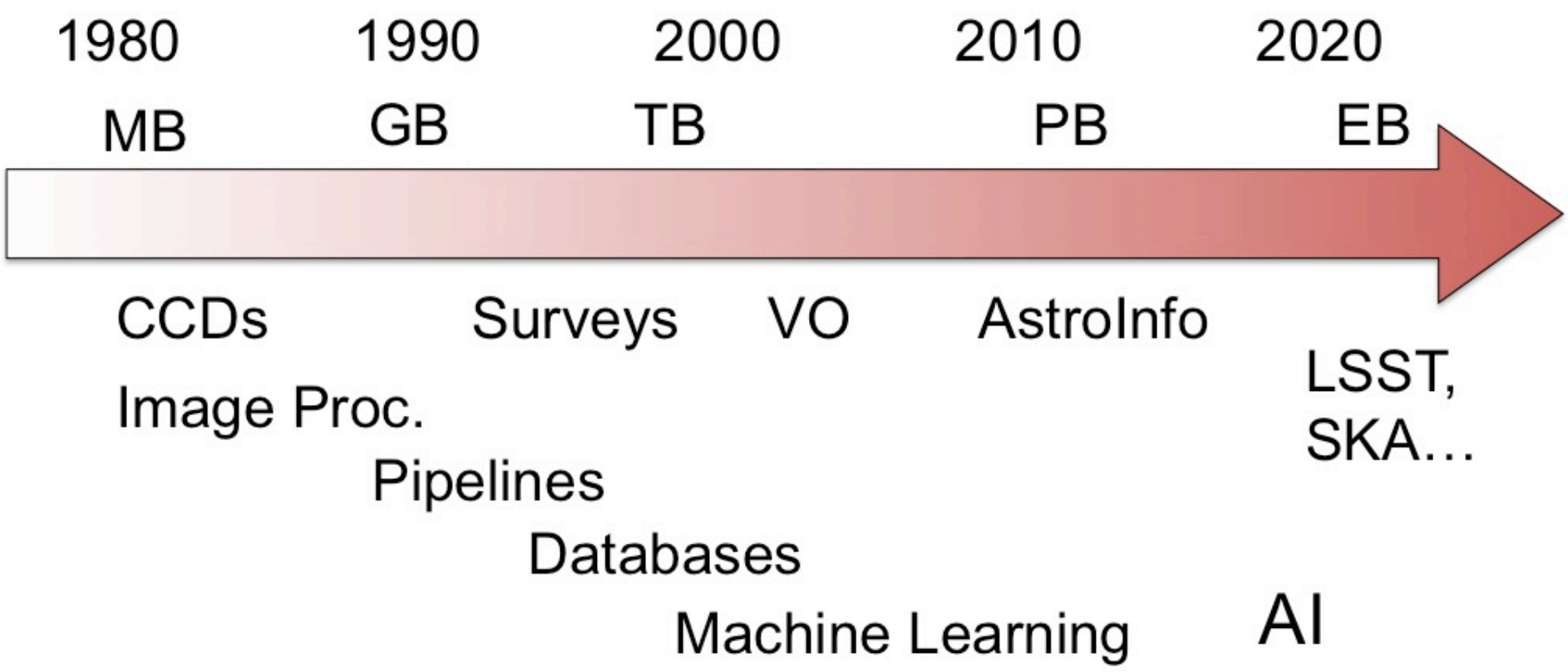


NAS SSB presentation,
November 2017

Caltech

The Evolving Data-Rich Astronomy

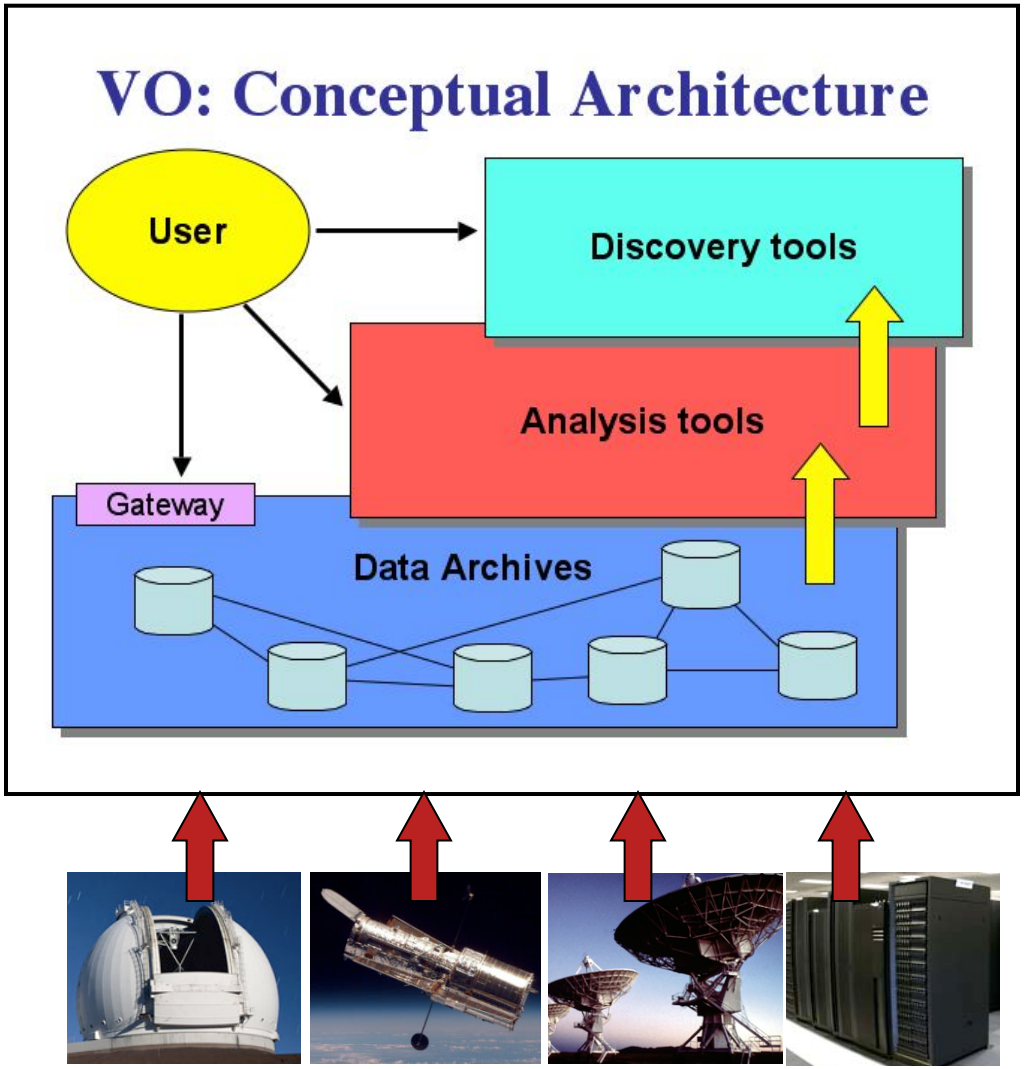
An example of a "Big Data" science driven by the advances in computing/information technology



Key challenges: data heterogeneity and complexity

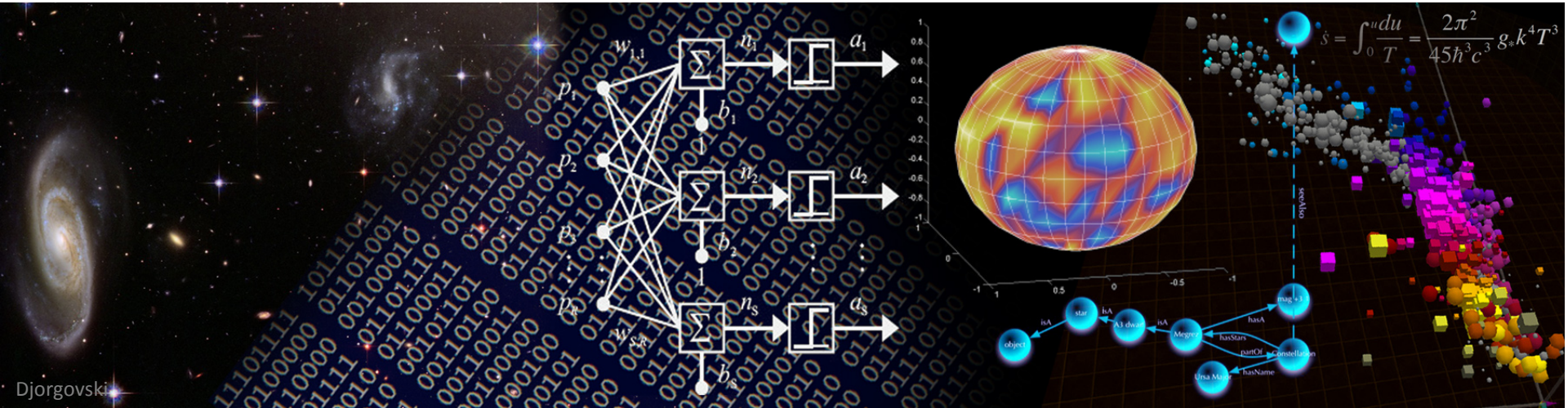
2000: The Virtual Observatory Concept

- 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
  - A grassroots response to the exponential data flood
  - Astro 2000: top "small projects" recommendation
  - Successful interagency (NSF, NASA) cooperation
  - Implementation: NVO, VAO, IVOA



2010: AstroInformatics

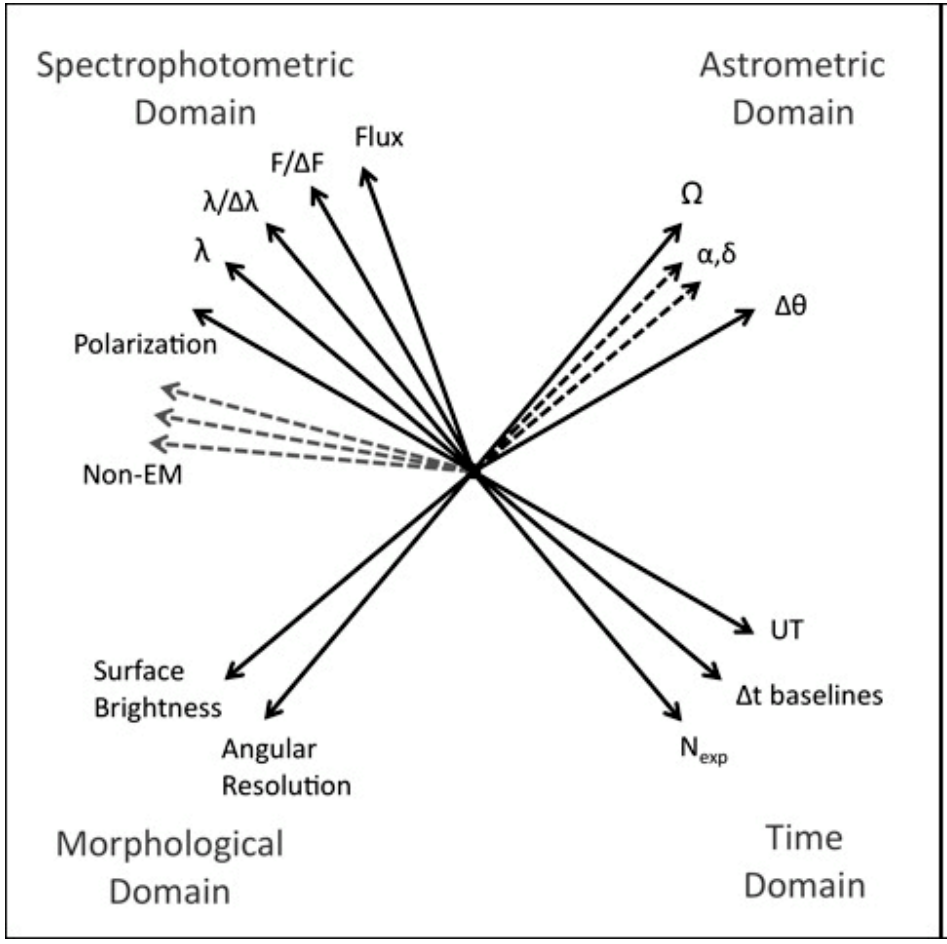
- One of the emerging X-informatics: domain-specific amalgam fields (domain science + CS + ICT)
- A mechanism for a broader community inclusion (both as contributors and as consumers)
- A mechanism for the multi/interdisciplinary data science methodology sharing
- Founding conference: 2010; Astroinformatics 2016 was an IAU Symposium
- Working groups within AAS, IAU

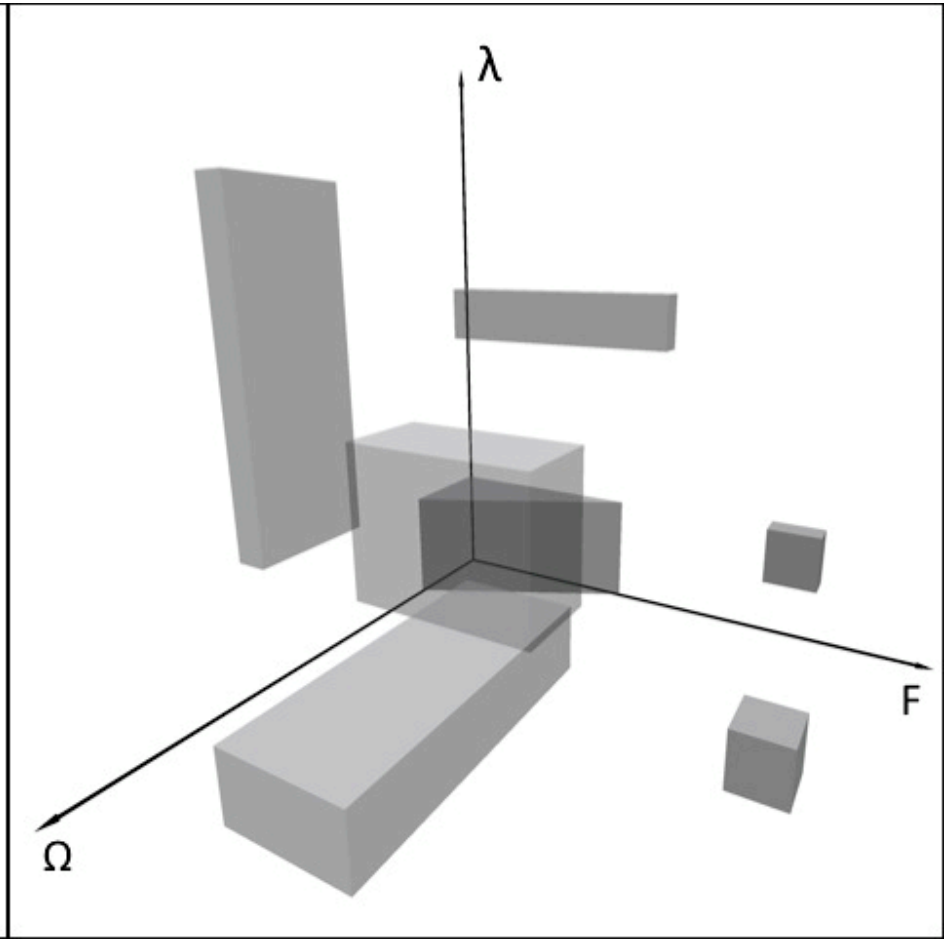


Systematic Exploration of the Observable Parameter Spaces (OPS)

Its axes are defined by the observable quantities

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

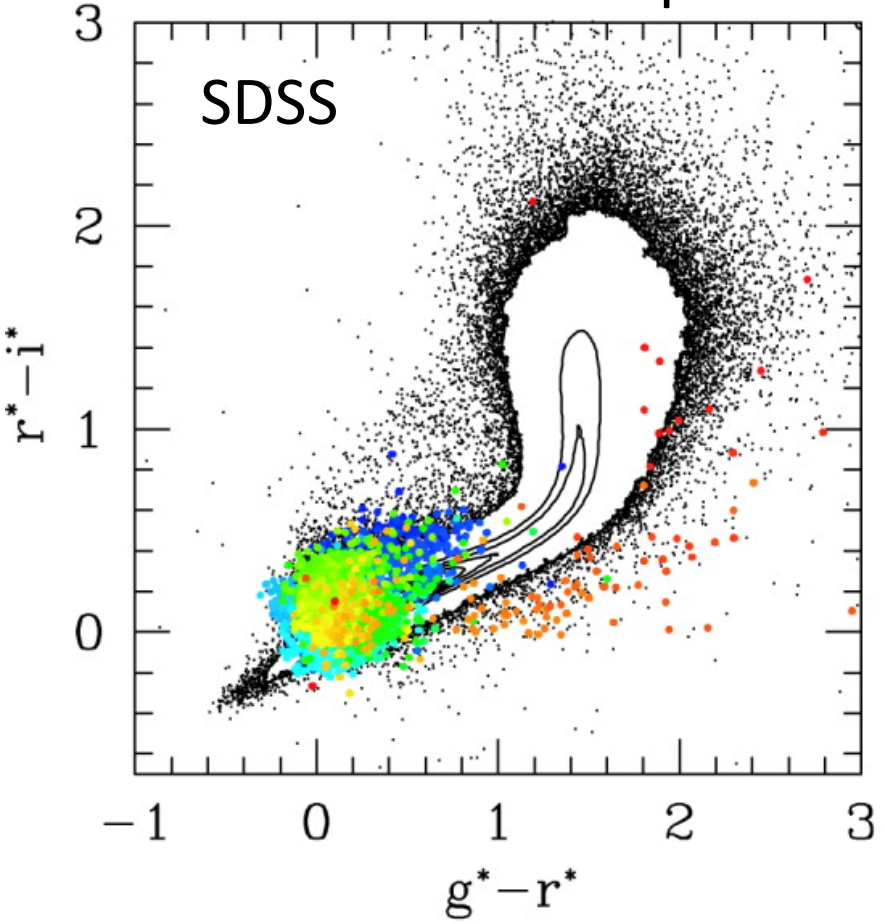




Technology opens new domains of the OPS → New discoveries

Measurements Parameter Space

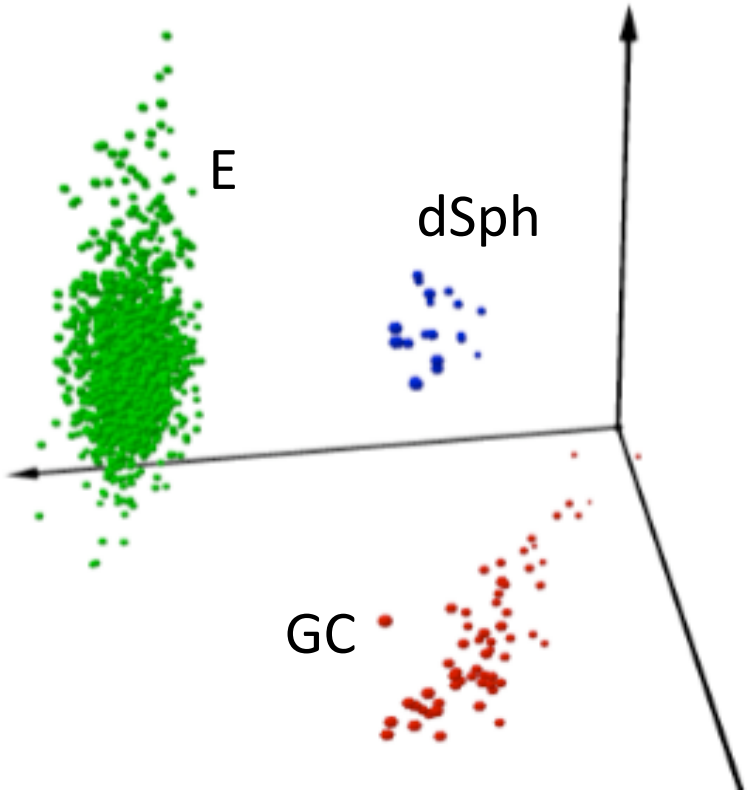
Colors of stars and quasars



Dimensionality ≤ the number of observed quantities

Physical Parameter Space

Fundamental Plane of hot stellar systems

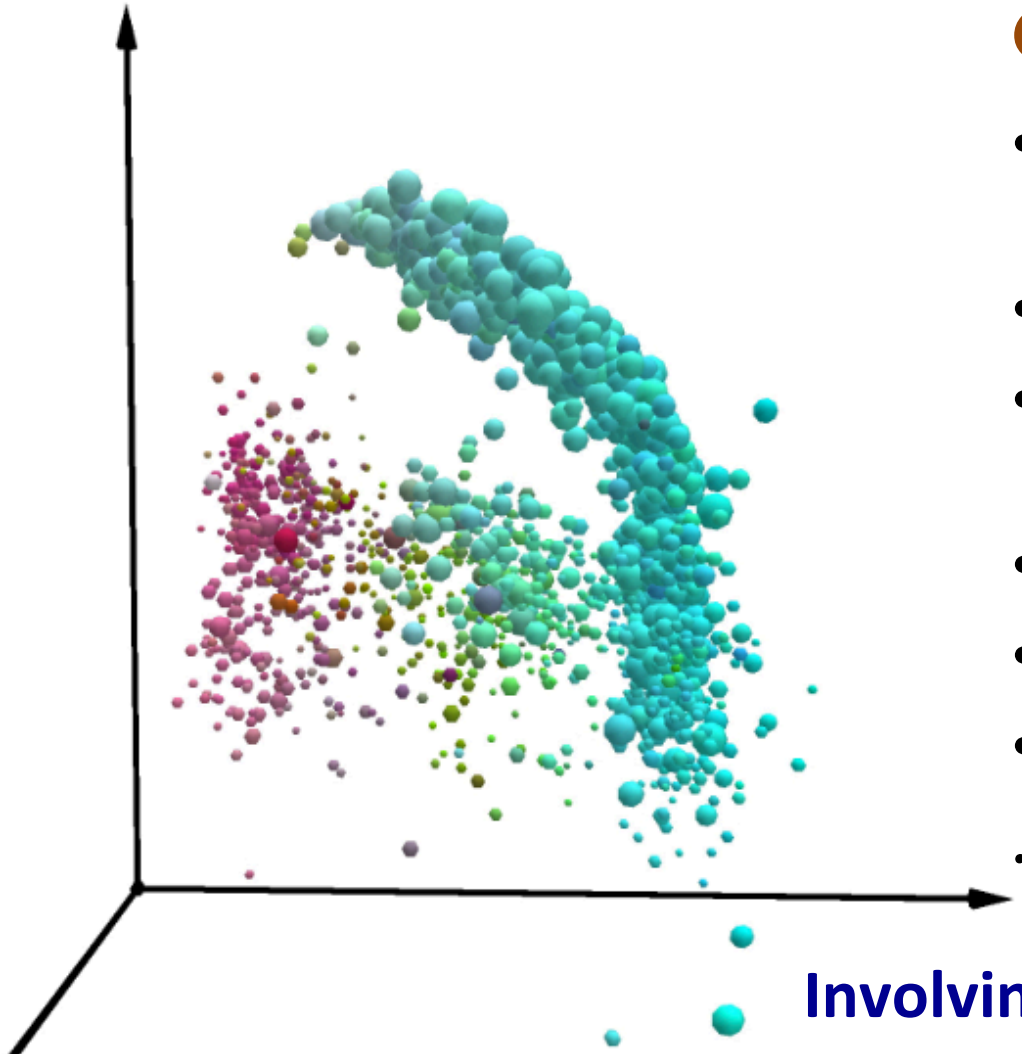


Both are populated by objects or events

Machine Learning Is the Key Methodology

Syrveys ⇨ Catalogs ⇨ Feature spaces ⇨ Phenomenology

Clustering, classification, correlation and outlier searches, …



Challenges:

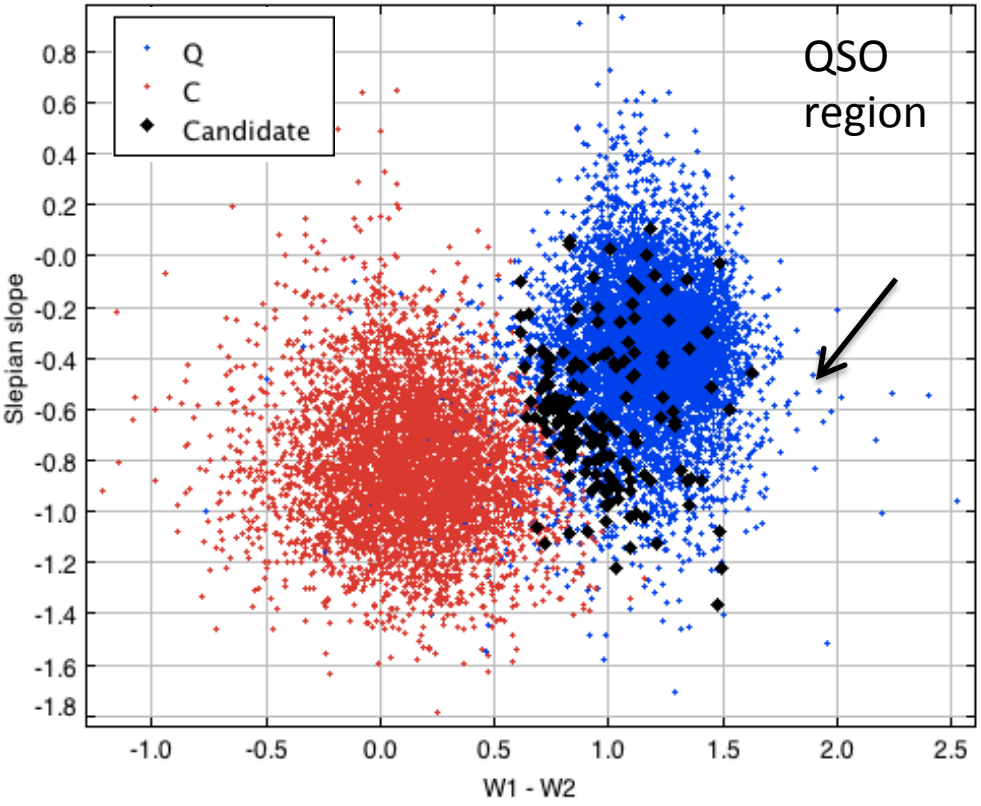
- Algorithm and data model choices
- Data incompleteness
- Feature selection and dimensionality reduction
- Uncertainty estimation
- Scalability
- Visualization

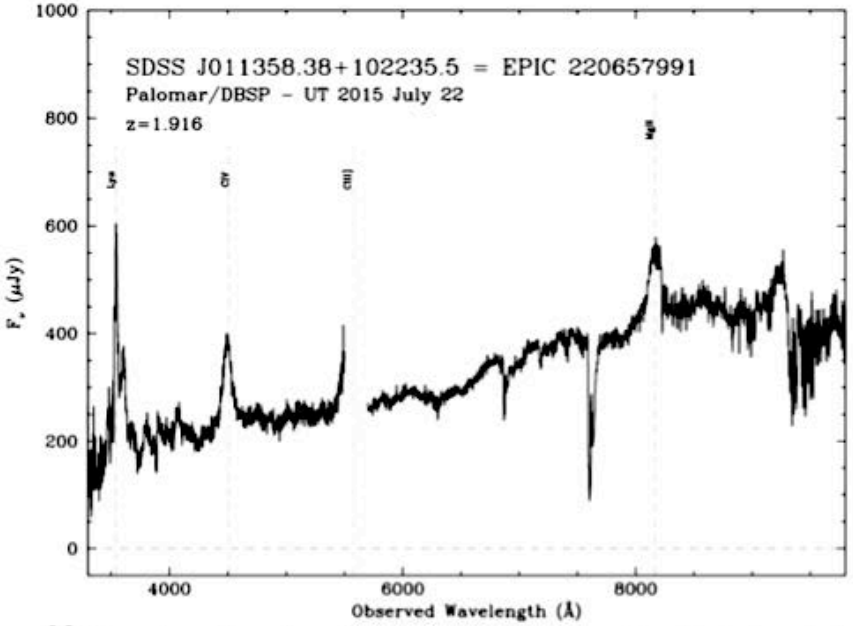
} Especially with the data dimensionality

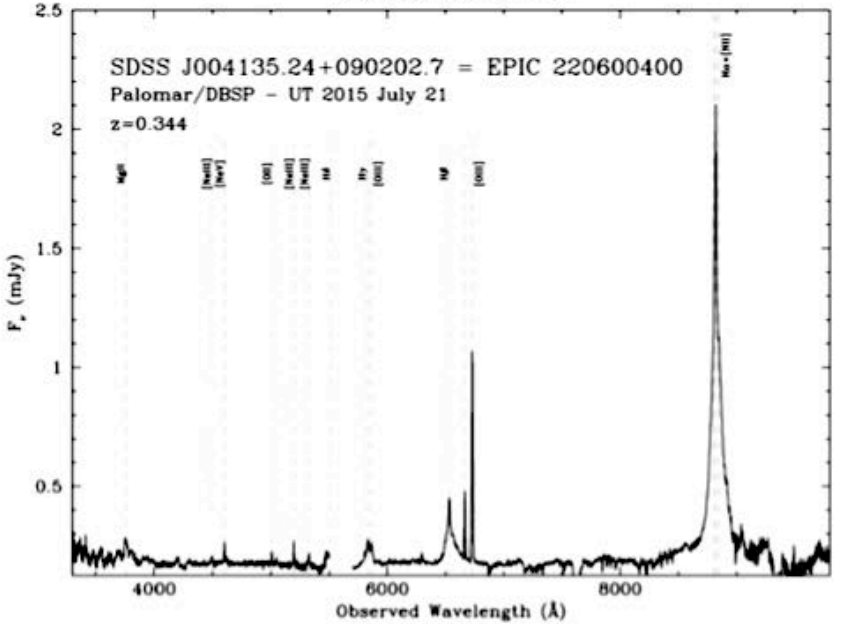
… etc.

Involving CS professionals is essential

Quasar Selection in a Combined Parameter Space of Variability and WISE Colors





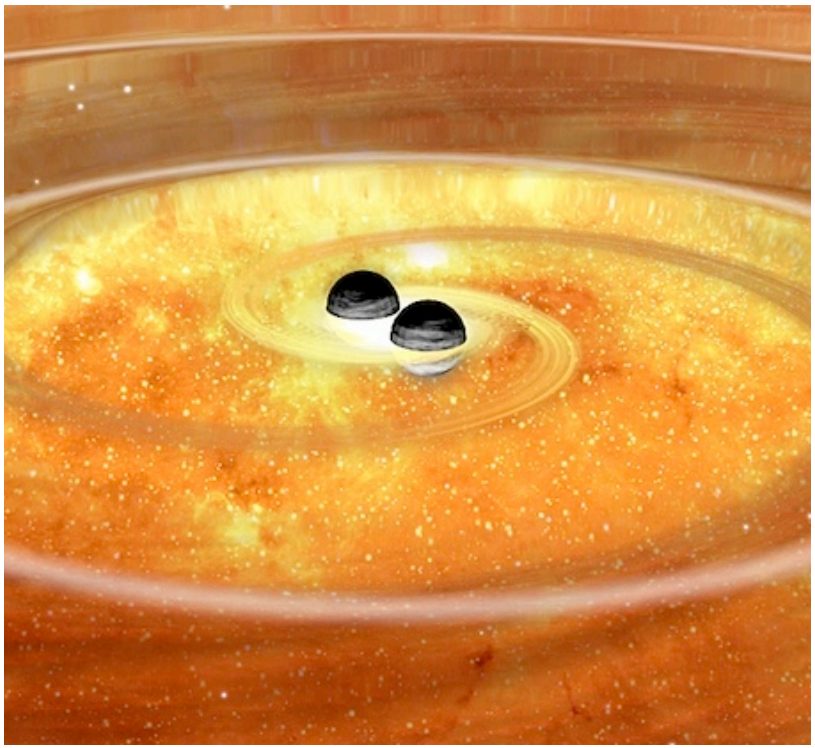


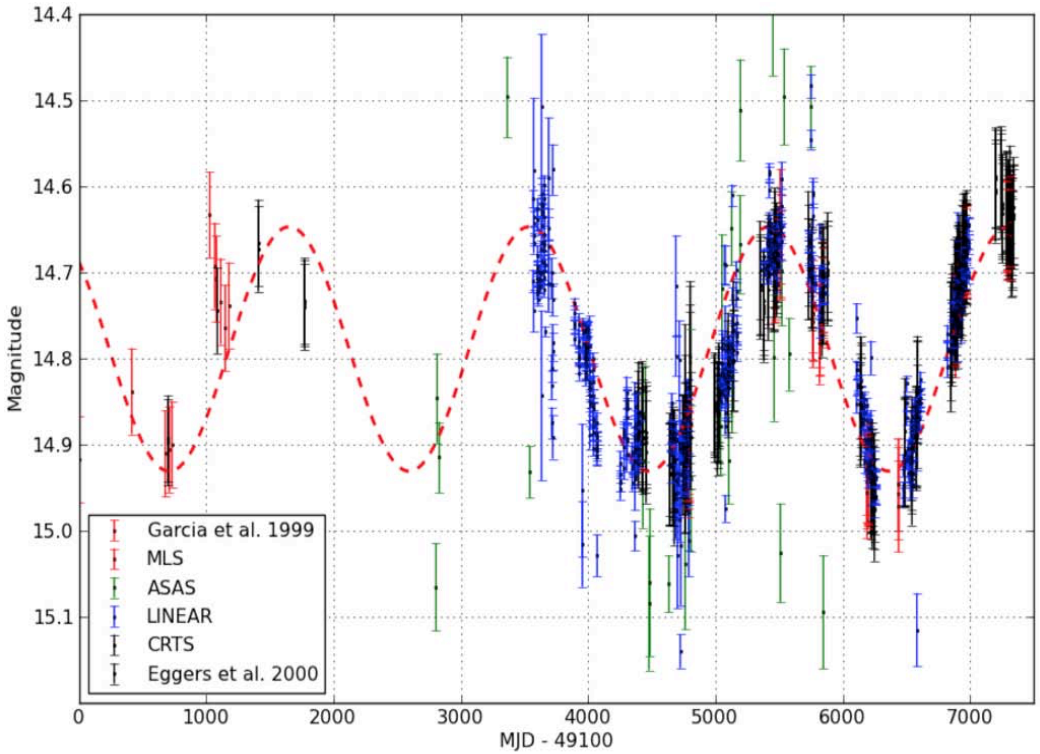
Initial results from the Kepler field: a 100% success rate!

Supermassive Black Hole Binaries

- Discovery of periodically variable quasars: a signature of binary supermassive black holes en route to a merger
  - Periodic signal superposed on a correlated (CAR1) noise, novel periodicity search algorithms, extensive Monte-Carlo modeling of false positives

(M. Graham et al.)

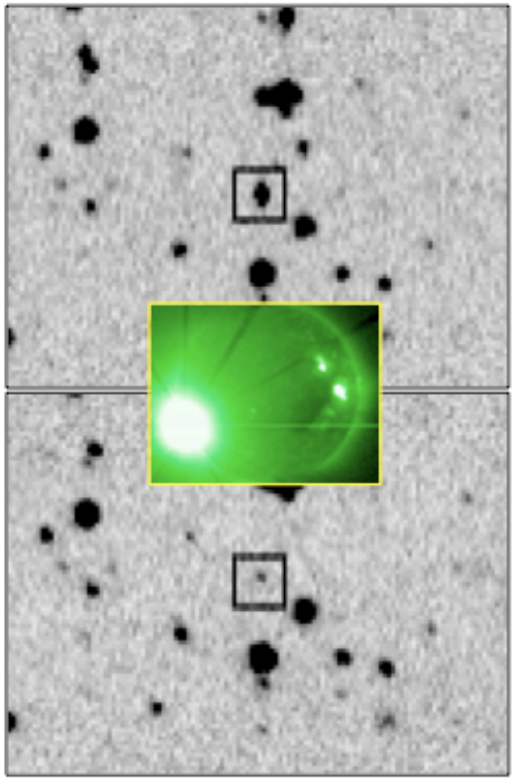




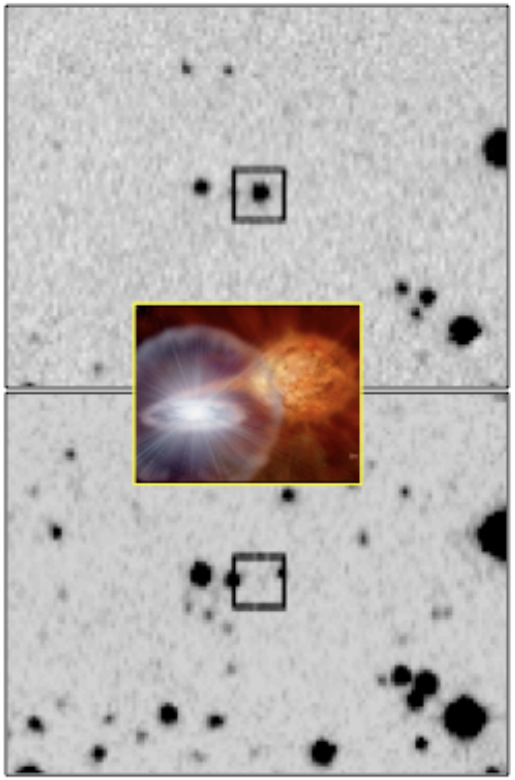
Automated Classification of Transient Events

In digital sky surveys: real-time mining of massive data streams

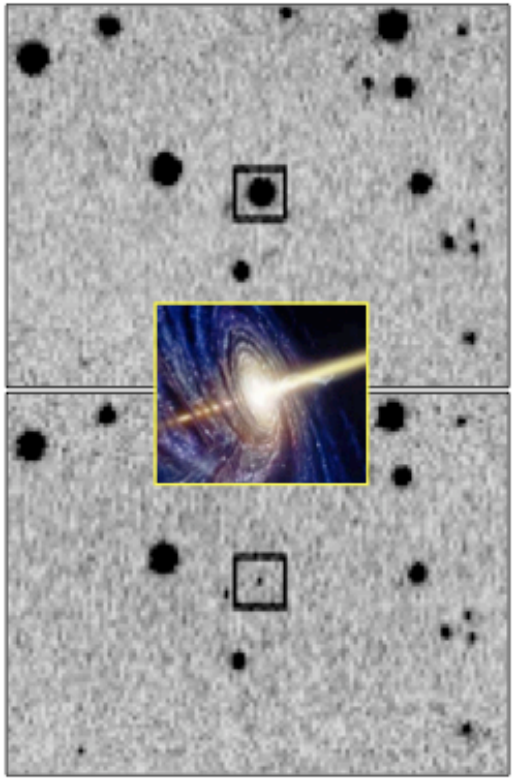
Flare star



Dwarf Nova



Blazar



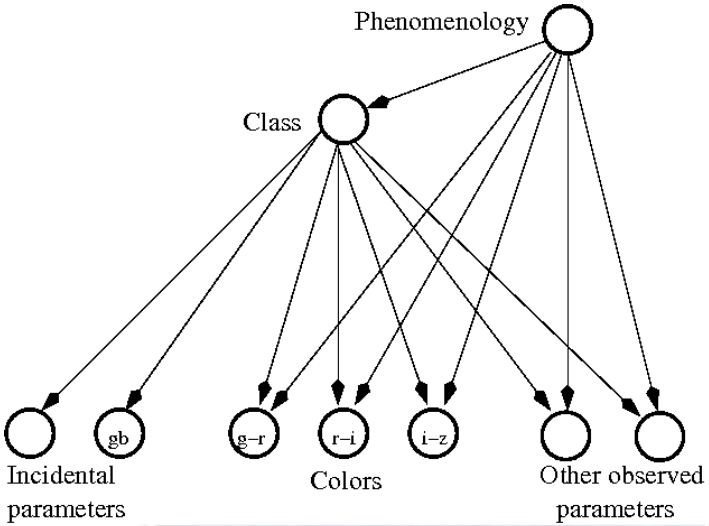
Vastly different physical phenomena, yet they look the same!
Which ones are the most interesting and worthy of follow-up?

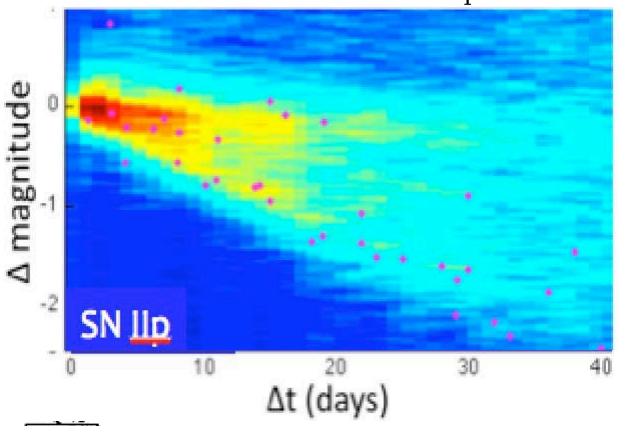
Rapid, automated transient classification is a critical need!

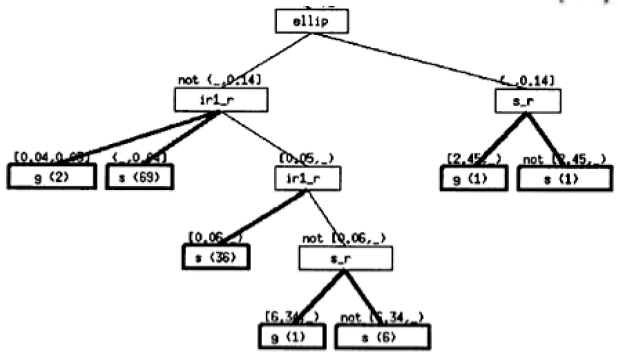
Automated Classification of Transients

- **Bayesian Networks**
  - Can incorporate heterogeneous and/ or missing data
  - Can incorporate contextual data, e.g., distance to the nearest star or galaxy
- **Probabilistic Structure Functions**
  - Based on 2D [Δt_1, Δm] distributions
- **Random Forests**
  - Ensembles of Decision Trees
- **Feature Selection Strategies**
  - Optimizing classifiers
- **Machine-Assisted Discovery**

etc., etc.

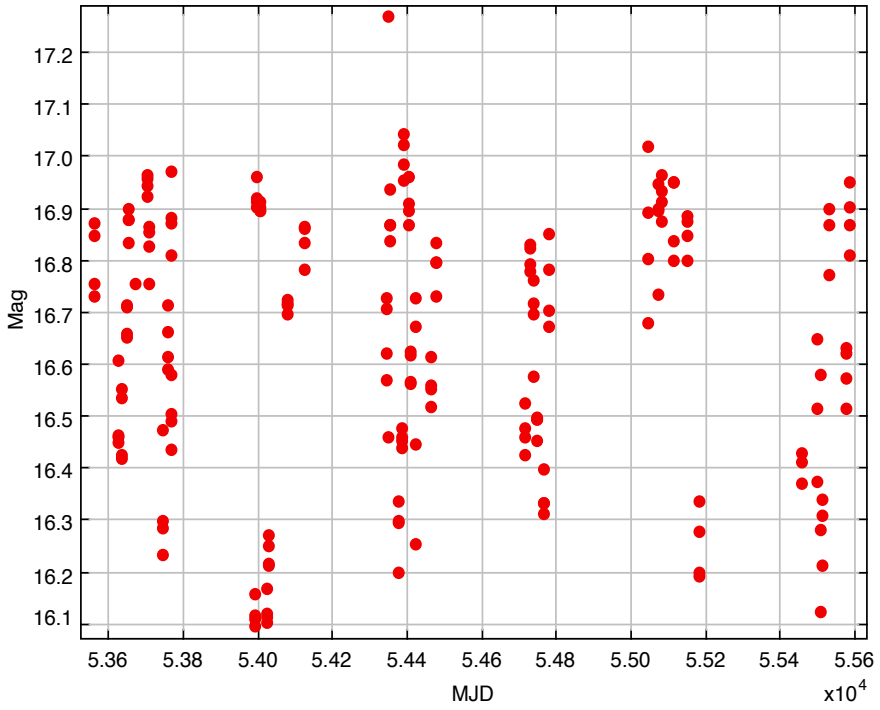


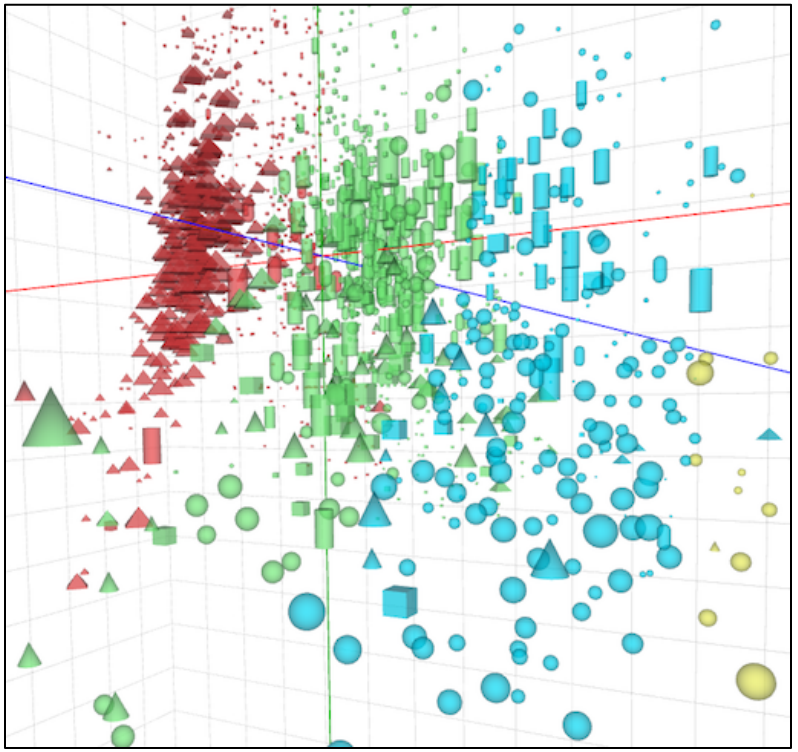




From Light Curves to Feature Vectors

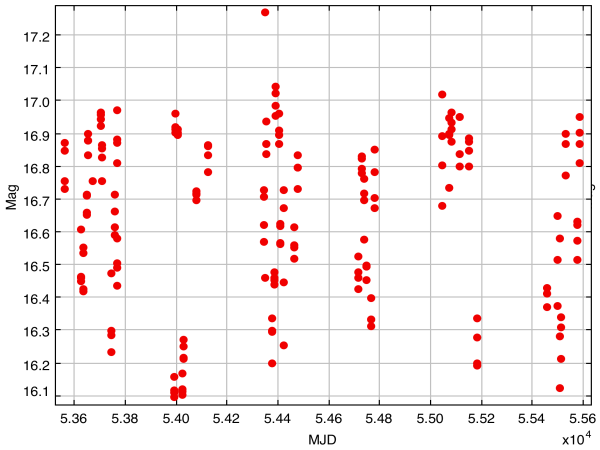
- We compute ~ 70 parameters and statistical measures for each light curve: amplitudes, moments, periodicity, etc.
- This turns **heterogeneous** light curves into **homogeneous** ***feature vectors*** in the parameter space
- Apply a variety of automated classification methods



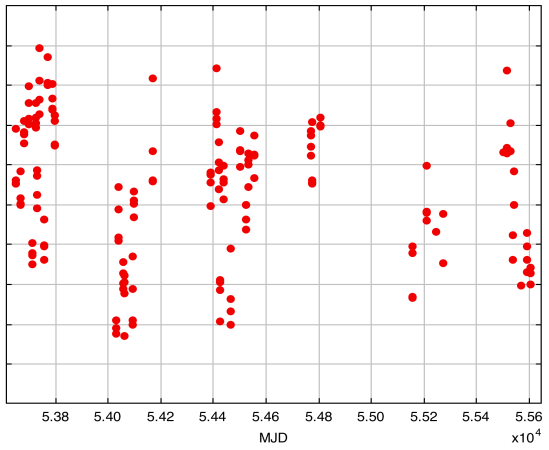


Optimizing Feature Selection

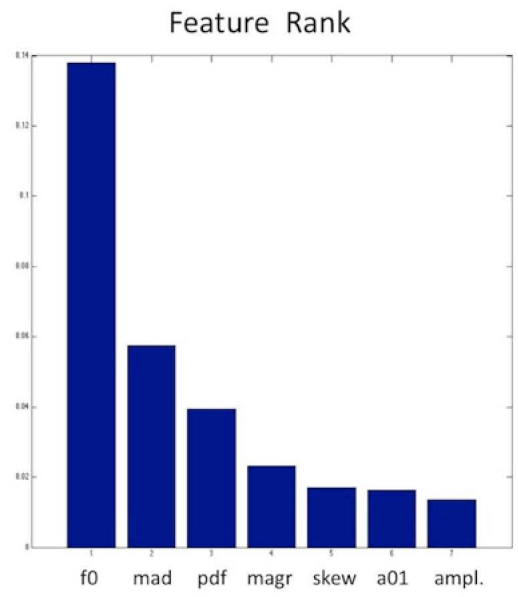
Rank features in the order of classification quality for a given classification problem, e.g., RR Lyrae vs. WUMa

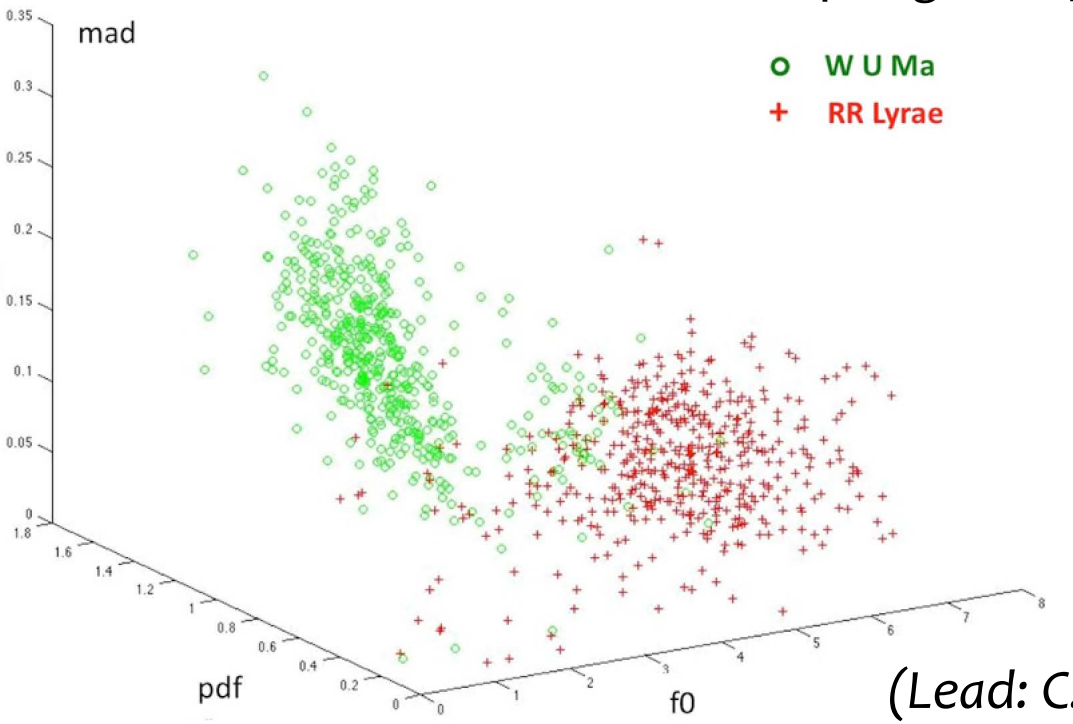


RR Lyrae



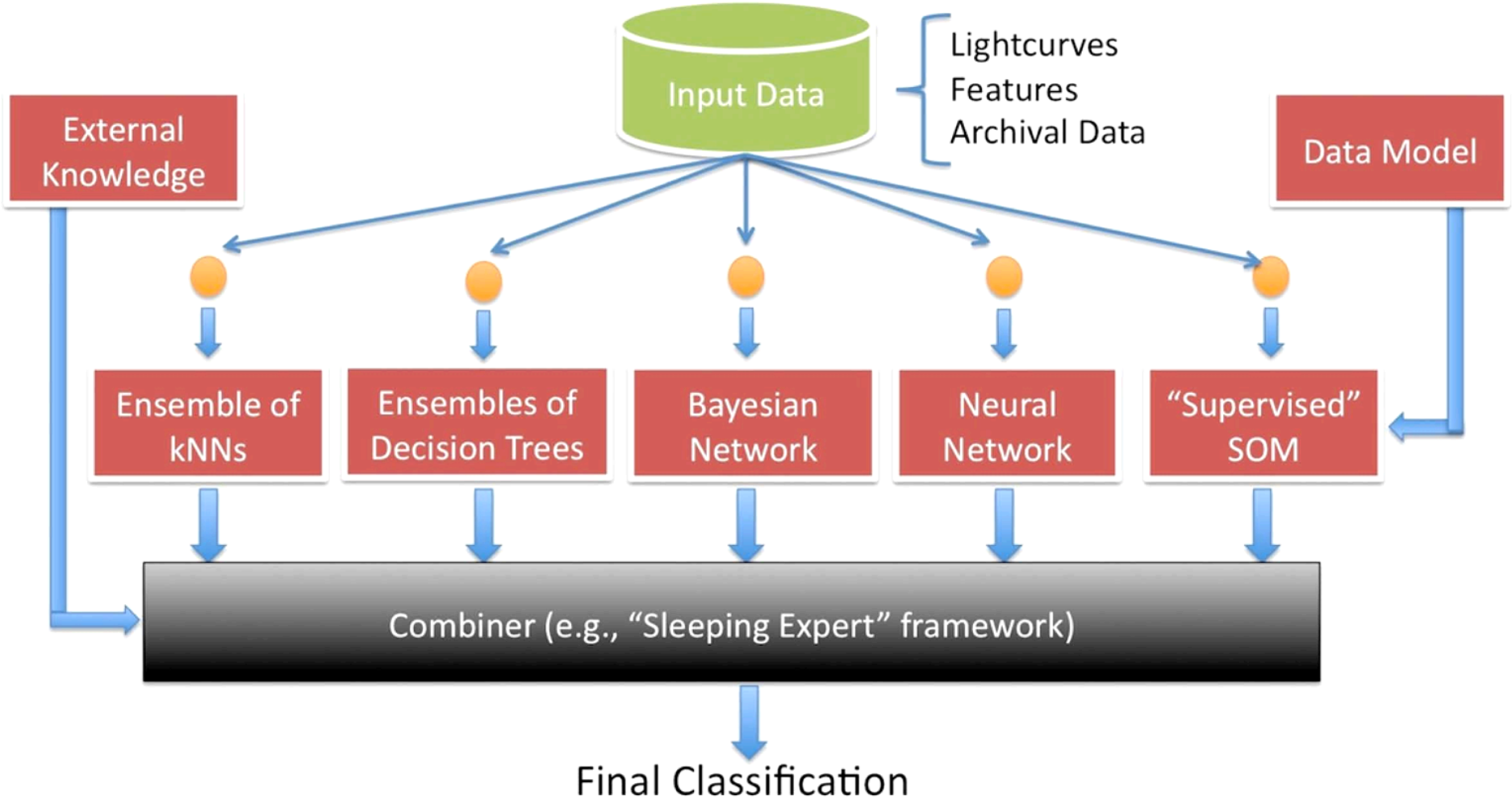
Eclipsing binary (W U Ma)





(Lead: C. Donalek)

Metaclassification: An optimal combining of classifiers

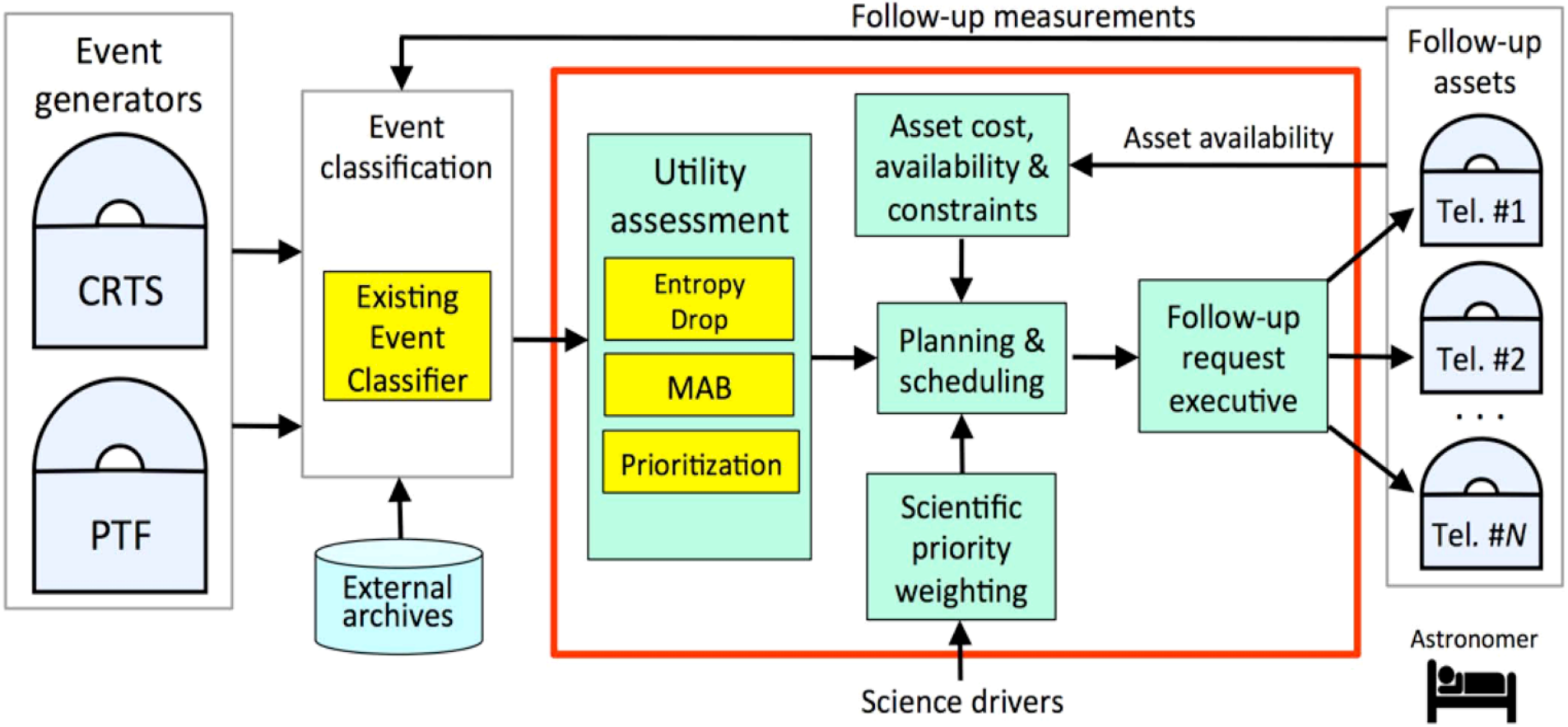


Exploring a variety of techniques for an optimal classification fusion:

Markov Logic Networks, Diffusion Maps, Multi-Arm Bandit, Sleeping Expert…

Automating an Optimal Follow-Up

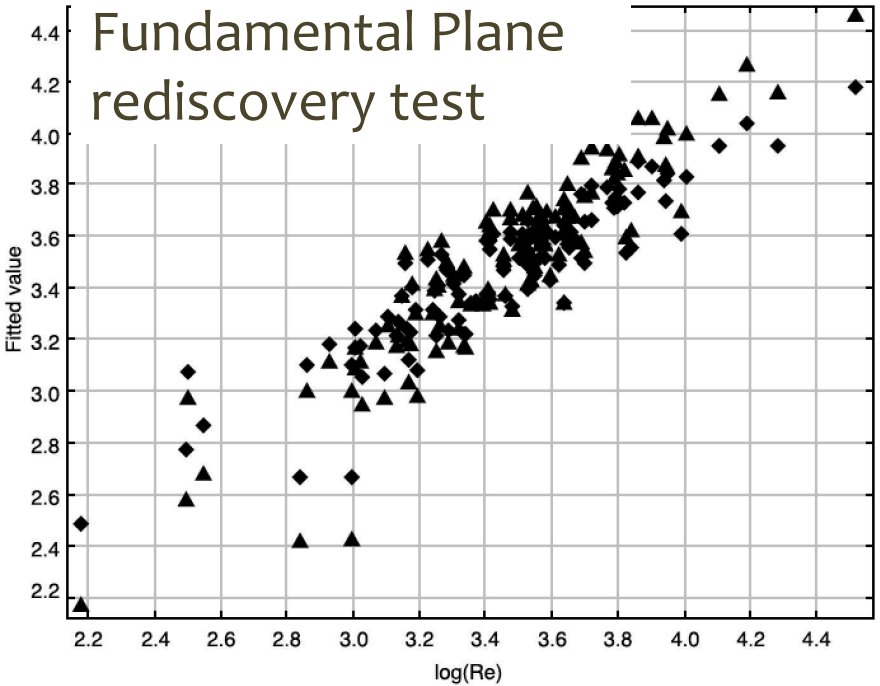
For the ***potentially most interesting events***, what type of follow-up observations has the greatest potential to discriminate among the competing event classes, given the available assets, and the potential scientific value?

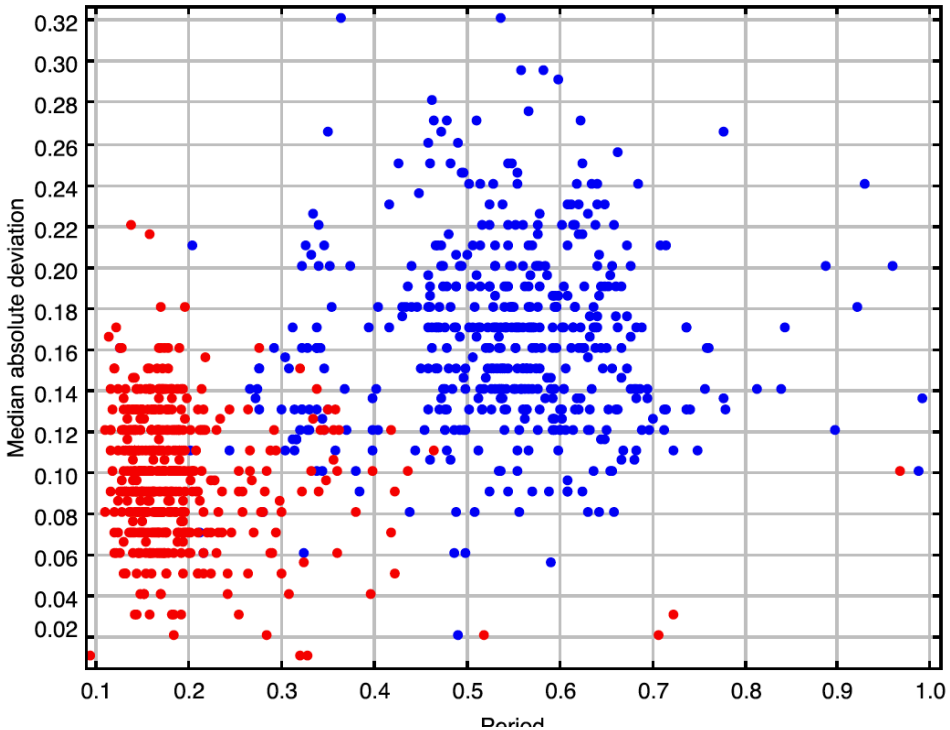


Machine Discovery Using *Eureqa*

Lipman et al., http://creativemachines.cornell.edu/eureqa

- Employs **symbolic regression** to determine best-fitting functional form to data and its parameters simultaneously
- Specify the building blocks to be used: algebraic operators, analytical functions, constants
- See Graham et al. (2013)





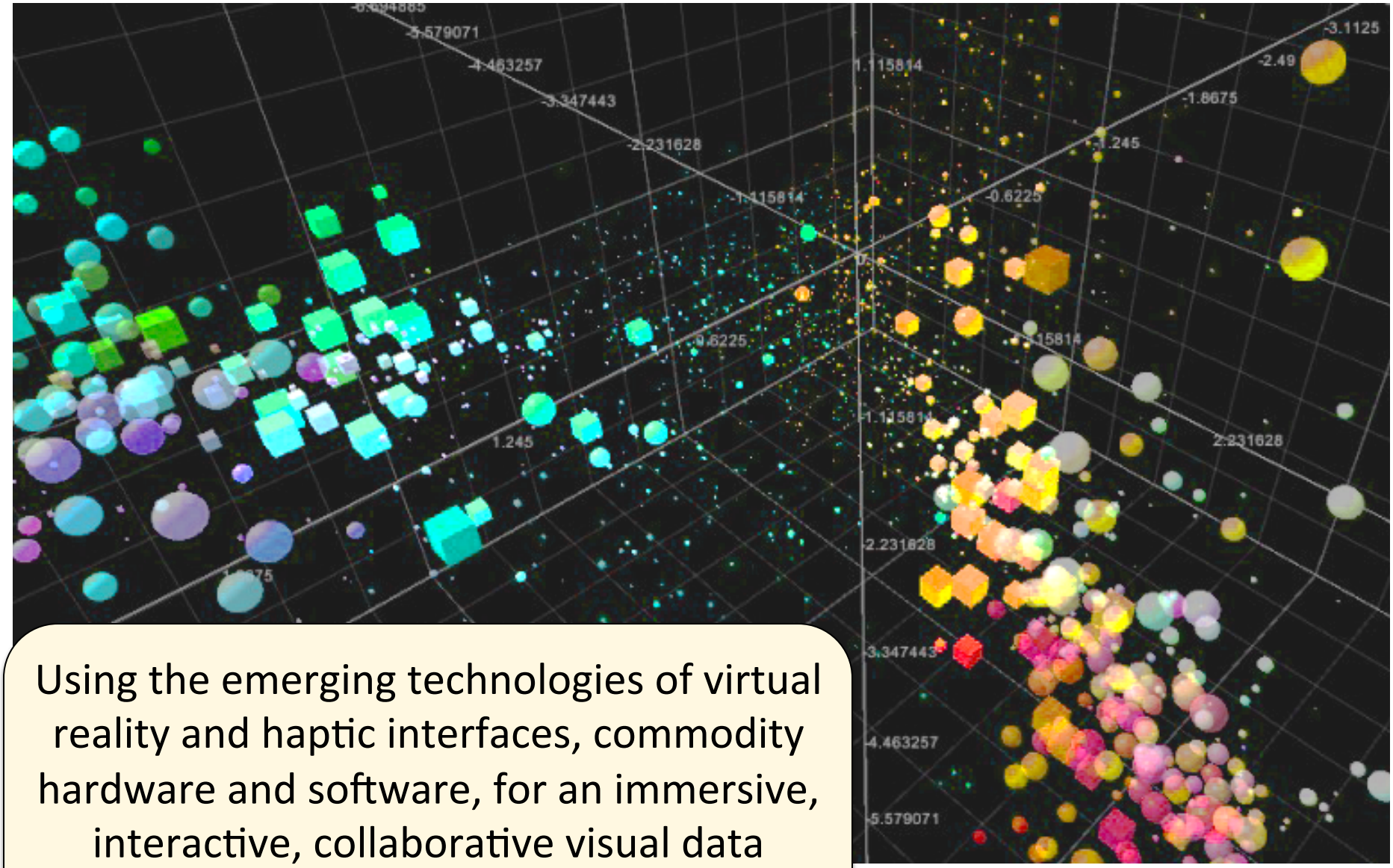
Classification of variable stars

The Uses of Machine Intelligence: Science on the Carbon-Silicon Interface

- **Data processing:**
  - Automated object / event classification, pattern recognition
  - Automated data quality control (anomaly/fault detection and repair)

- **Data mining, analysis, and understanding:**
  - Clustering, classification, outlier / anomaly detection
  - Pattern recognition, hidden correlation search
  - Assisted dimensionality reduction for visualization
  - Workflow control in Grid- or Cloud-based apps
- **Data farming and data discovery:** semantic web, etc.
- **Code design and implementation:** from art to science?

Innovative Data Visualization



Using the emerging technologies of virtual reality and haptic interfaces, commodity hardware and software, for an immersive, interactive, collaborative visual data analytics and exploration

C. Donalek, SGD (CD^3)
S. Davidoff (JPL)

Methodology Transfer in Action

With JPL's Center for Data Science and Technology (D. Crichton, R. Doyle, et al.)

Genome Biology

APACHE OODT

COMMENT

Astrogenomics: big data, old problems, old solutions?

National Cancer Institute

Early Detection Research Network

Biomarkers: the key to early detection

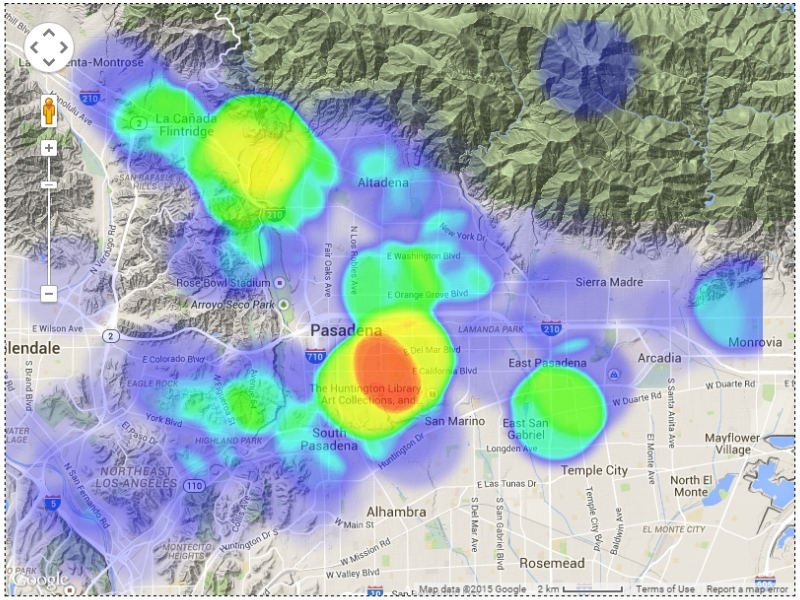
From genomics to cancer research, to geosciences, climate change, etc. etc.

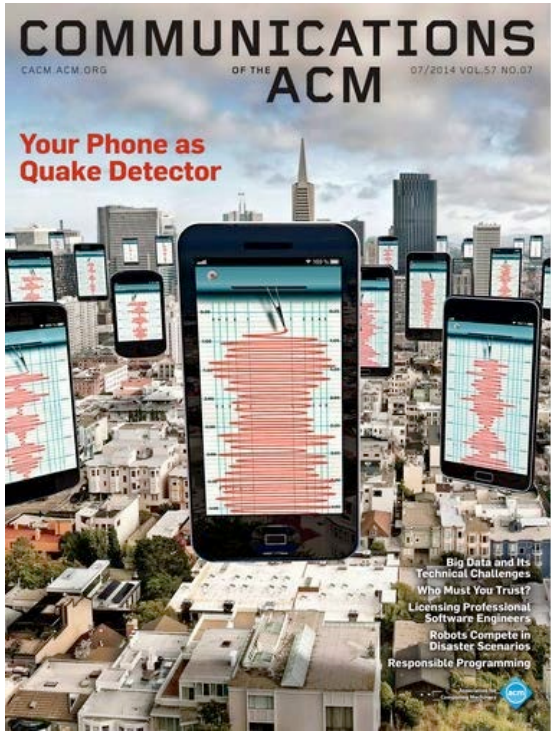
Real Time Classification and Response

Seismology:
Cell phones as a sensor network

Time domain astronomy

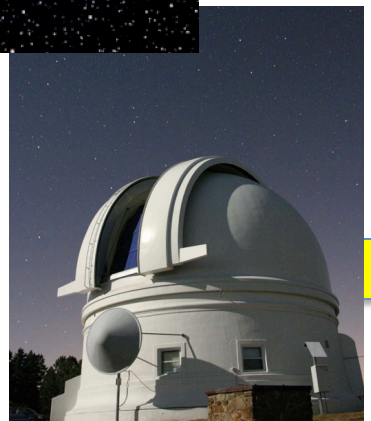
Lake Castaic M4.2 Jan 4 2015 Heatmap



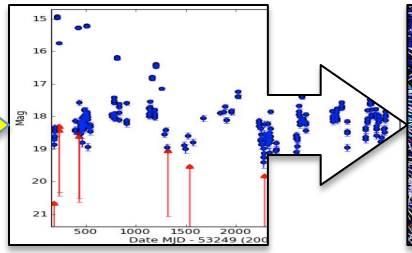


Event

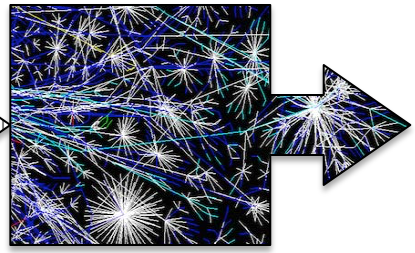
Detection



Classification



Decision making



Follow-up

Some Takeaway Points:

- Astronomy remains at the forefront of data science (VO, Astroinformatics, …)
- Machine learning and associated technologies are *essential* for the exploration of vast data spaces and knowledge discovery
  - They are *critical* for the Time Domain Astronomy
  - They have to become parts of a standard toolkit for the researchers in the 21st century – but our educational efforts are lagging
- Data science methodology transfer is possible, and there are some early success stories