Center for Data-Driven Discovery

Prof. S. George Djorgovski

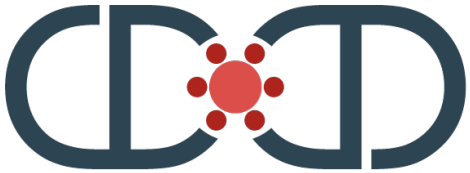
BDTF meeting,
JPL, Nov. 2017

Caltech

CENTER FOR DATA-DRIVEN DISCOVERY

Center for Data-Driven Discovery

- A part of the Caltech-JPL joint initiative for data science and technology
- Serves research efforts Institute-wide



- The goals: assist faculty in the **formulation and execution of data-intensive projects**, and facilitate **interdisciplinary sharing of methods, ideas**, novel projects, etc.

Key expertise:

- **Cyberinfrastructure development** and implementation (example: Virtual Observatory)
- **Knowledge discovery** tools (machine learning, statistics, innovative data visualization)
- Data science **methodology transfer** between different domains (e.g., astronomy to biology)

The Virtual Observatory Framework

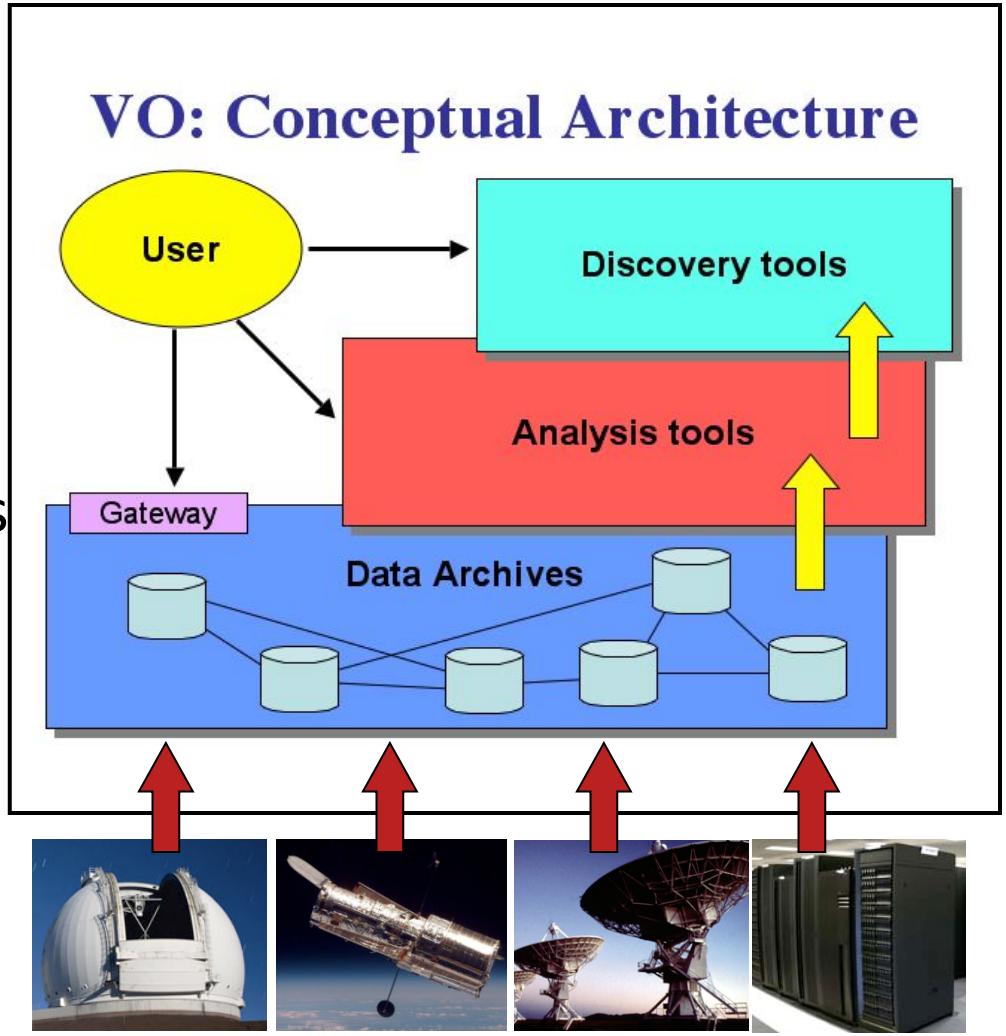
- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***

– Provide and federate content (data, metadata) services, standards, and analysis/compute services

– Develop and provide data exploration and discovery tools

– A successful example of a domain Cyber-Infrastructure

– Now a global data grid, coordinated by the Int'l VO Alliance (http://ivoa.net)



EarthCube: Software Architecture for Earth Science

Science Teams

Using the VO experience

EarthCube Cyberinfrastructure

Data Science Infrastructure (Data, Algorithms, Machines)

EarthCube Data Analytics Centers

EarthCube Repository

EarthCube Discovery

Satellite Information Data Systems

Airborne Data

Agency Earth Data Archives

Other Data Systems (In-Situ, University)

Data Providers

Research

Applications

Decision Support

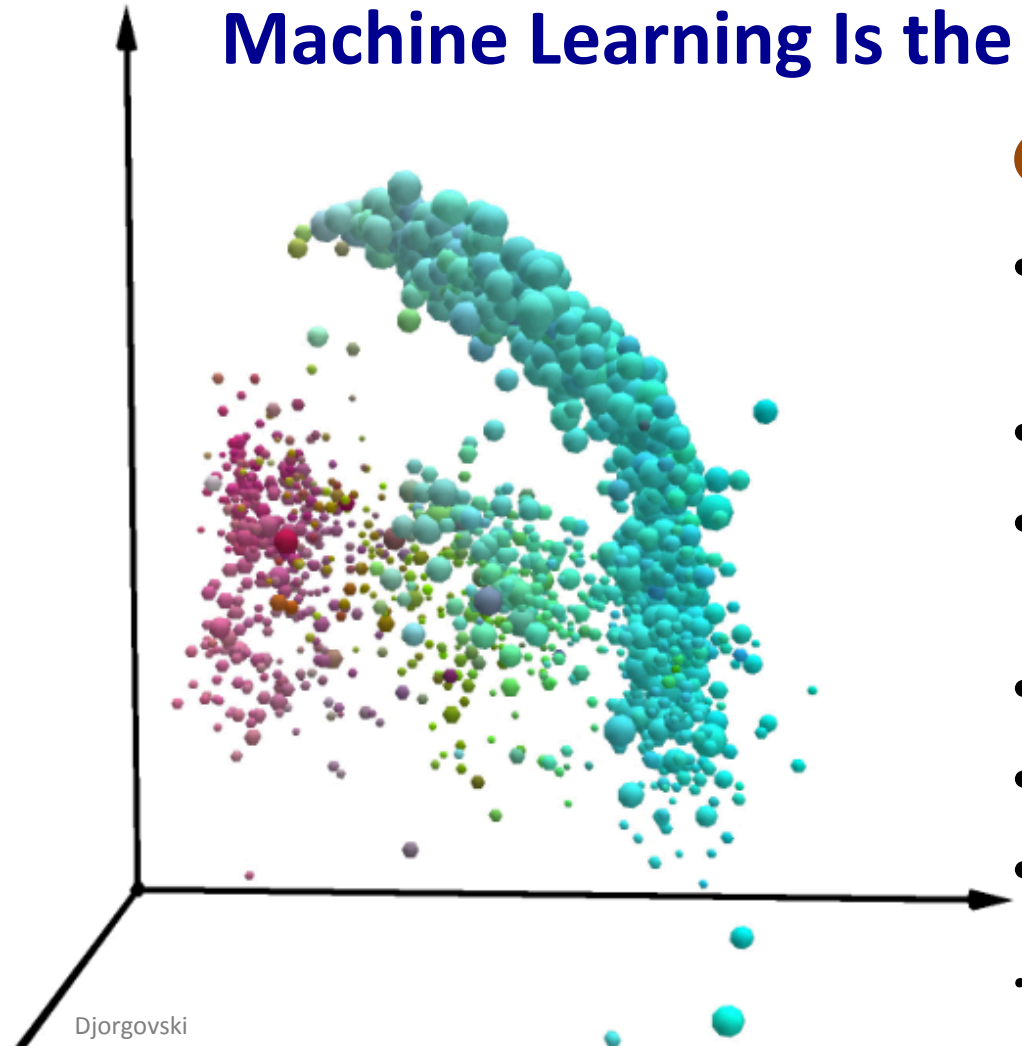
Applied Science

E.Law, D. Crichton (JPL)
A. Mahabal, SGD (Caltech)

Exploration of Parameter Spaces is the Central Problem of Data Science

Clustering, classification, correlation and outlier searches, …

Machine Learning Is the Key Methodology



Djorgovski

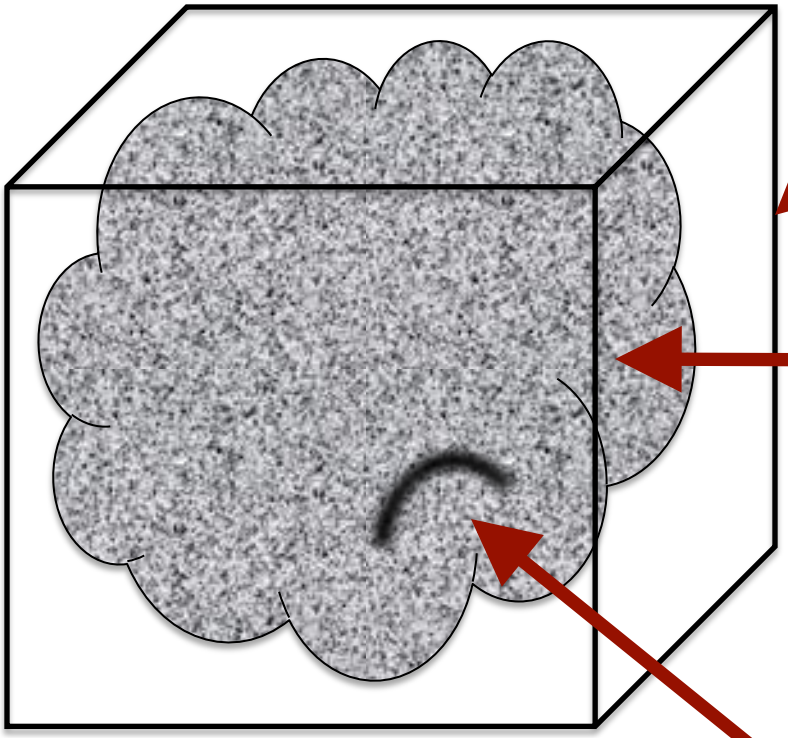
Challenges:

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

} Especially with the data dimensionality

… etc.

Pattern or structure (Correlations, Clustering, Outliers, etc.) Discovery in High-Dimensional Parameter Spaces



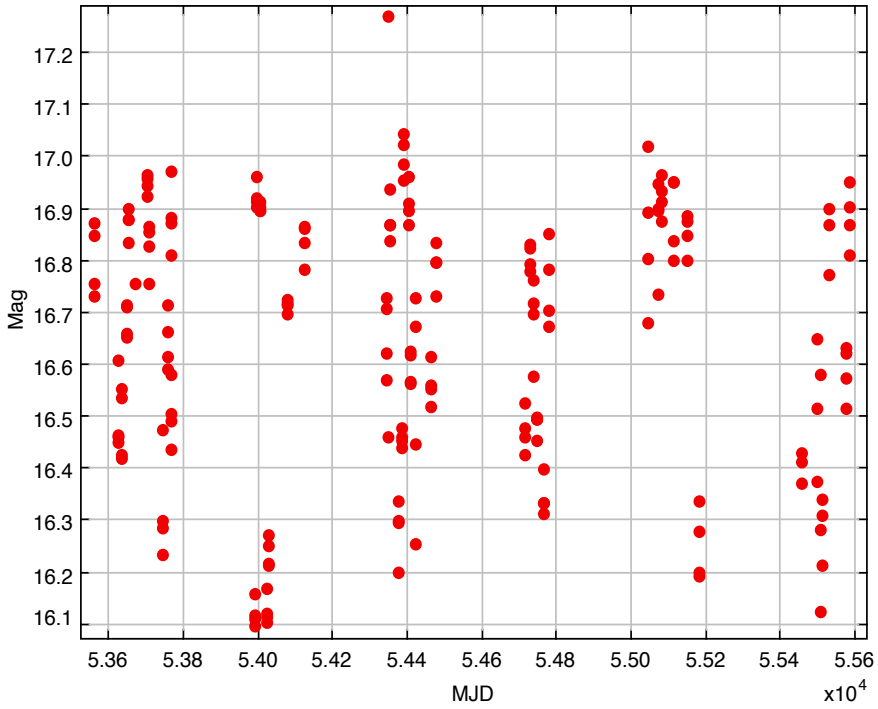
D >> 3 parameter space hypercube

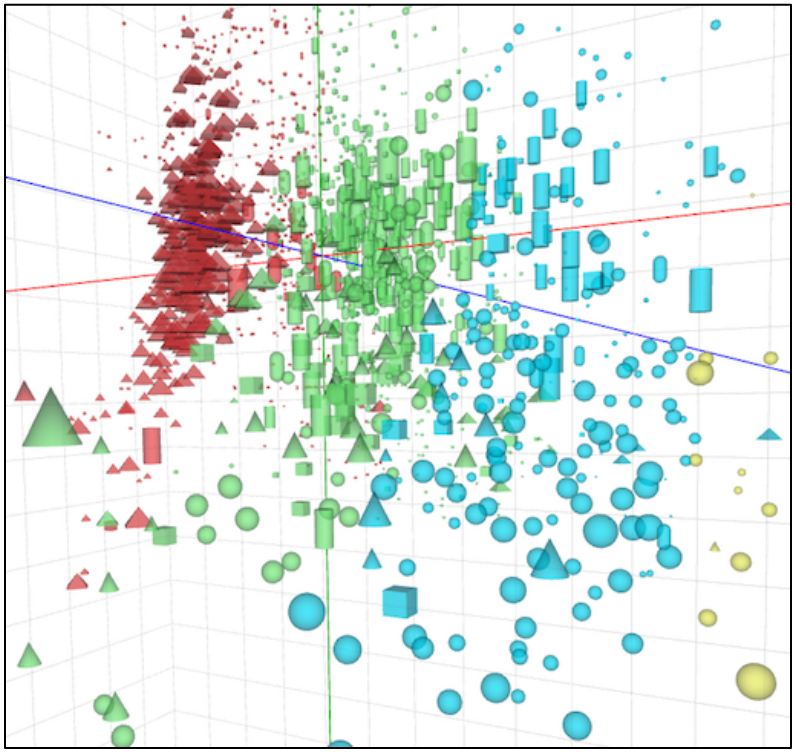
High-D data cloud: mostly noise, of an arbitrary distribution

But in some corner of some sub-D projection of this data space, there is ***something ≠ noise***

Dealing With Data Heterogeneity

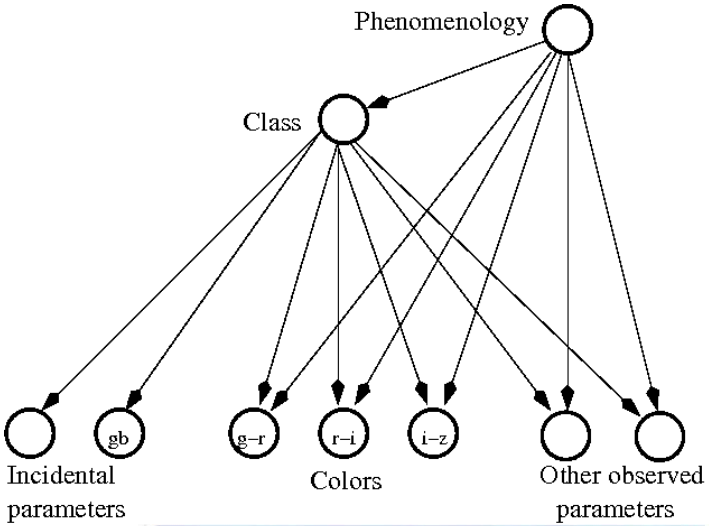
- Heterogeneous, irregularly sampled or missing data present a major obstacle for most ML methods
- Replacing the data with a comprehensive set of statistical descriptors turns **heterogeneous** data sets into **homogeneous** ***feature vectors*** in the parameter space

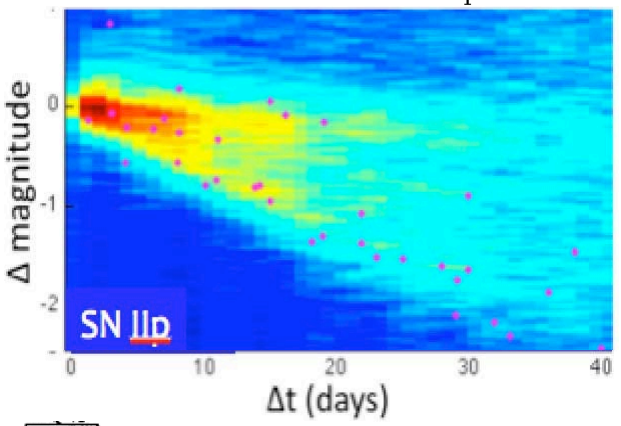


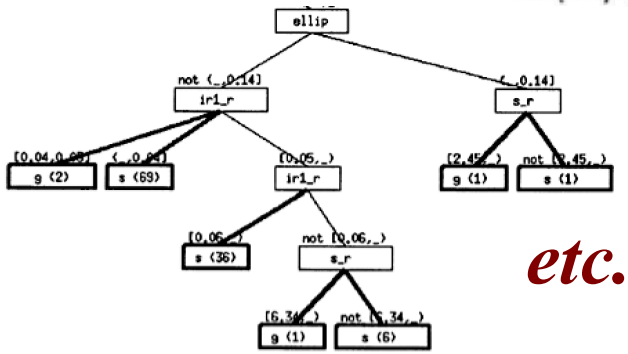


Automated Classification of Transients

- **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**



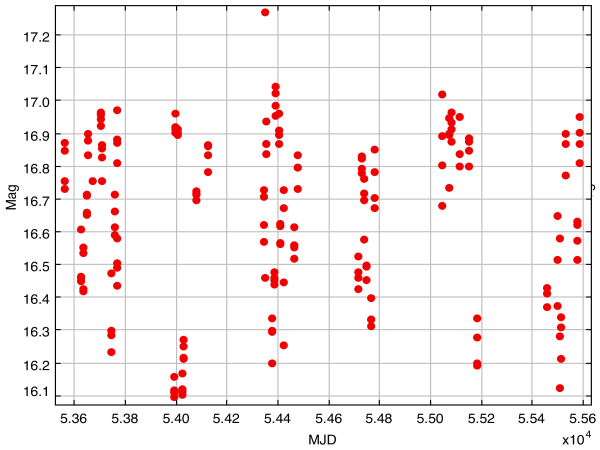




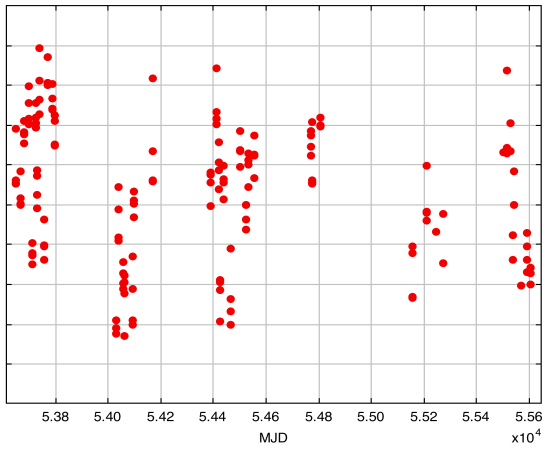
etc., etc.

Optimizing Feature Selection

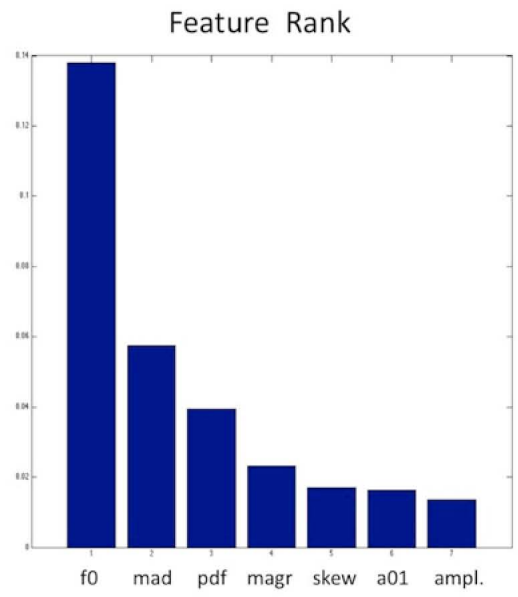
Rank features in the order of classification quality for a given classification problem, e.g., RR Lyrae vs. WUMa

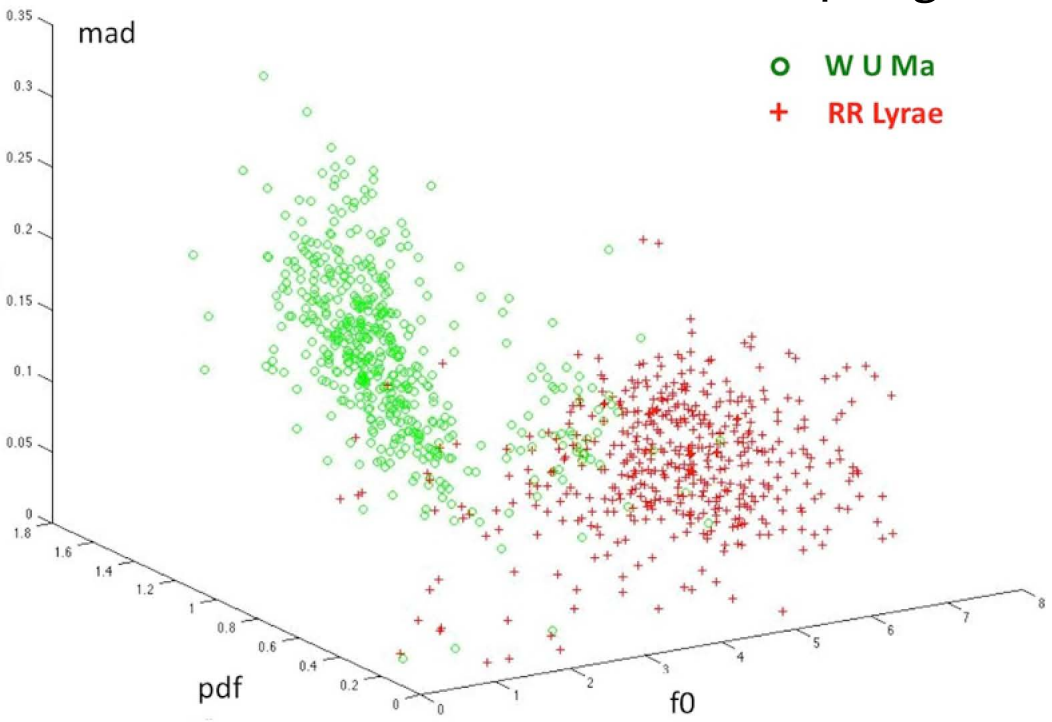


RR Lyrae



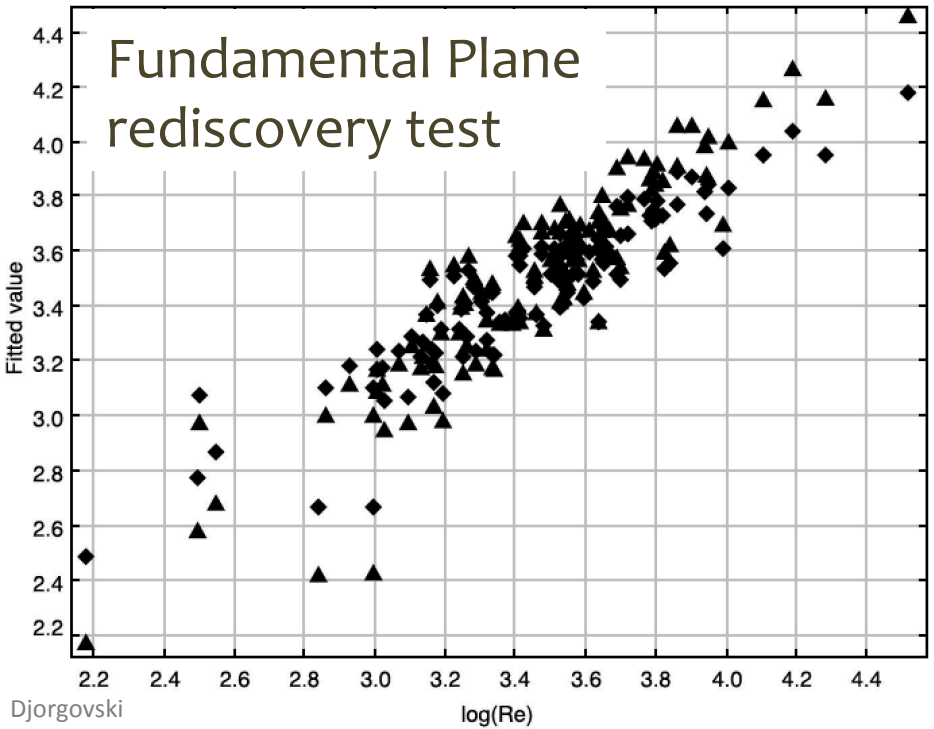
Eclipsing binary (W U Ma)

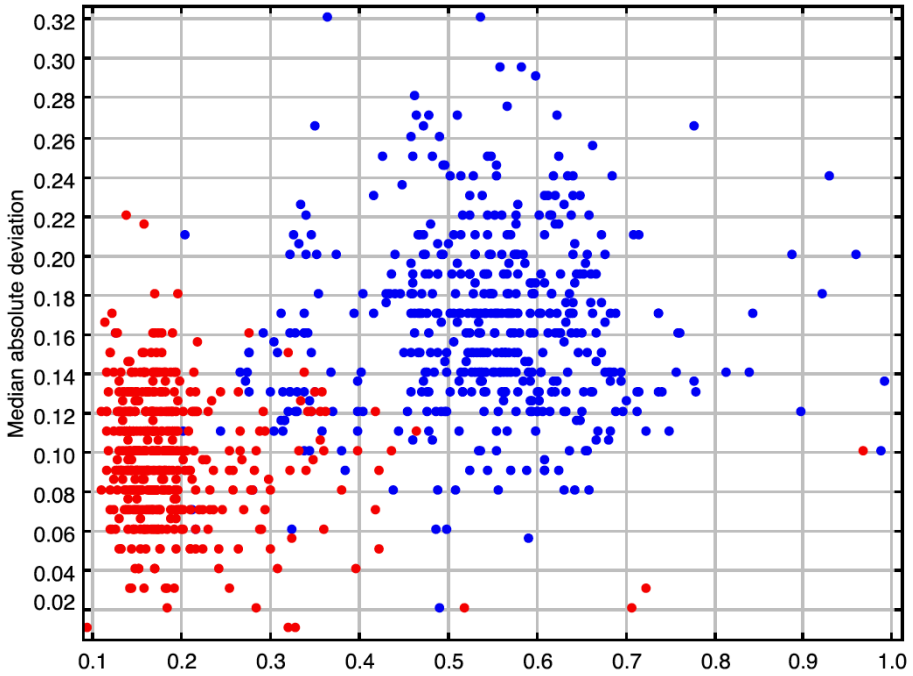




Machine Discovery Using *Eureqa*

- Employs symbolic regression and a genetic algorithm to determine best-fitting functional form to data and its parameters simultaneously
- Specify the building blocks to be used: algebraic operators, analytical functions, constants

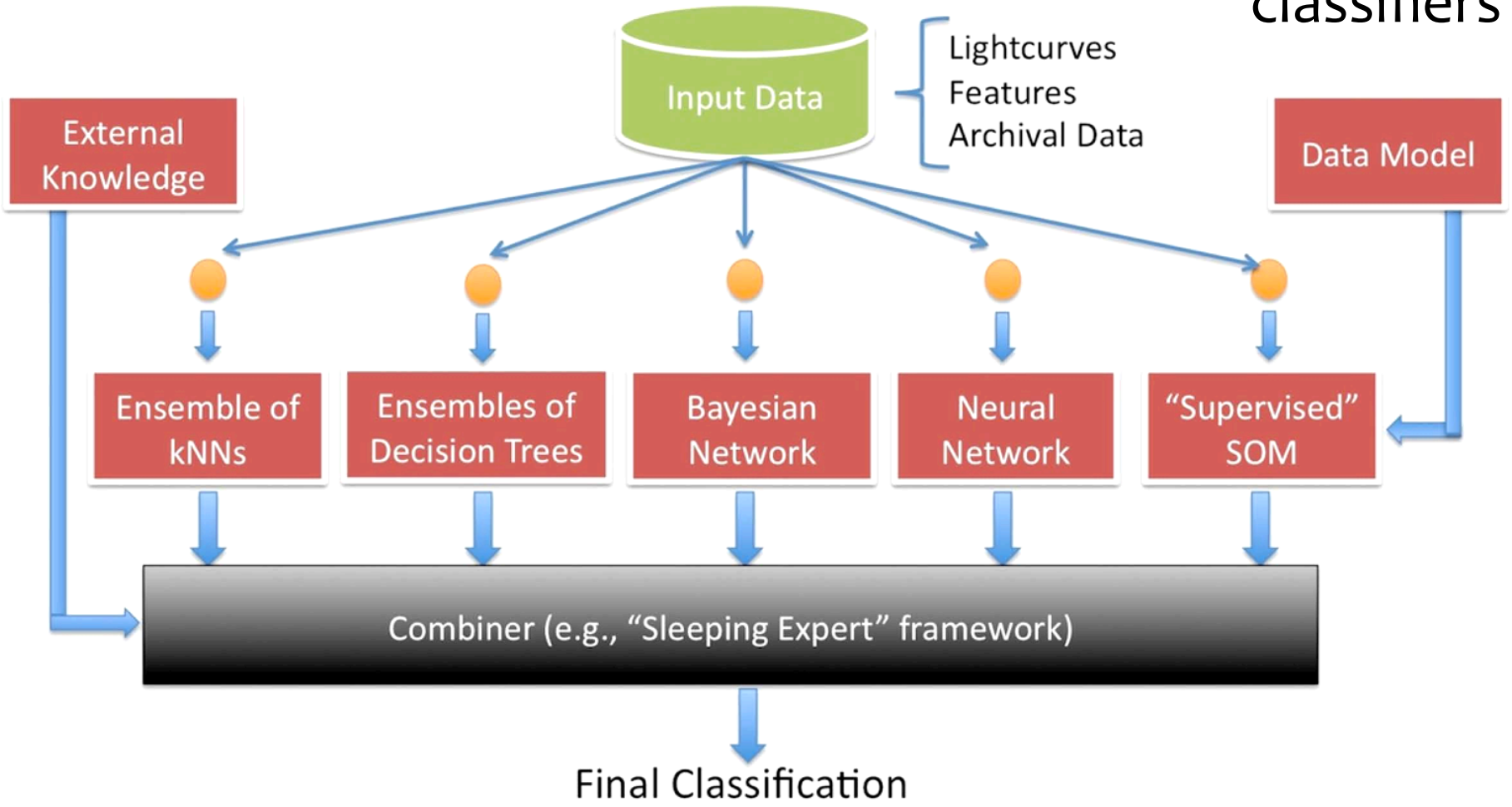




Classification of variable stars

Metaclassification:

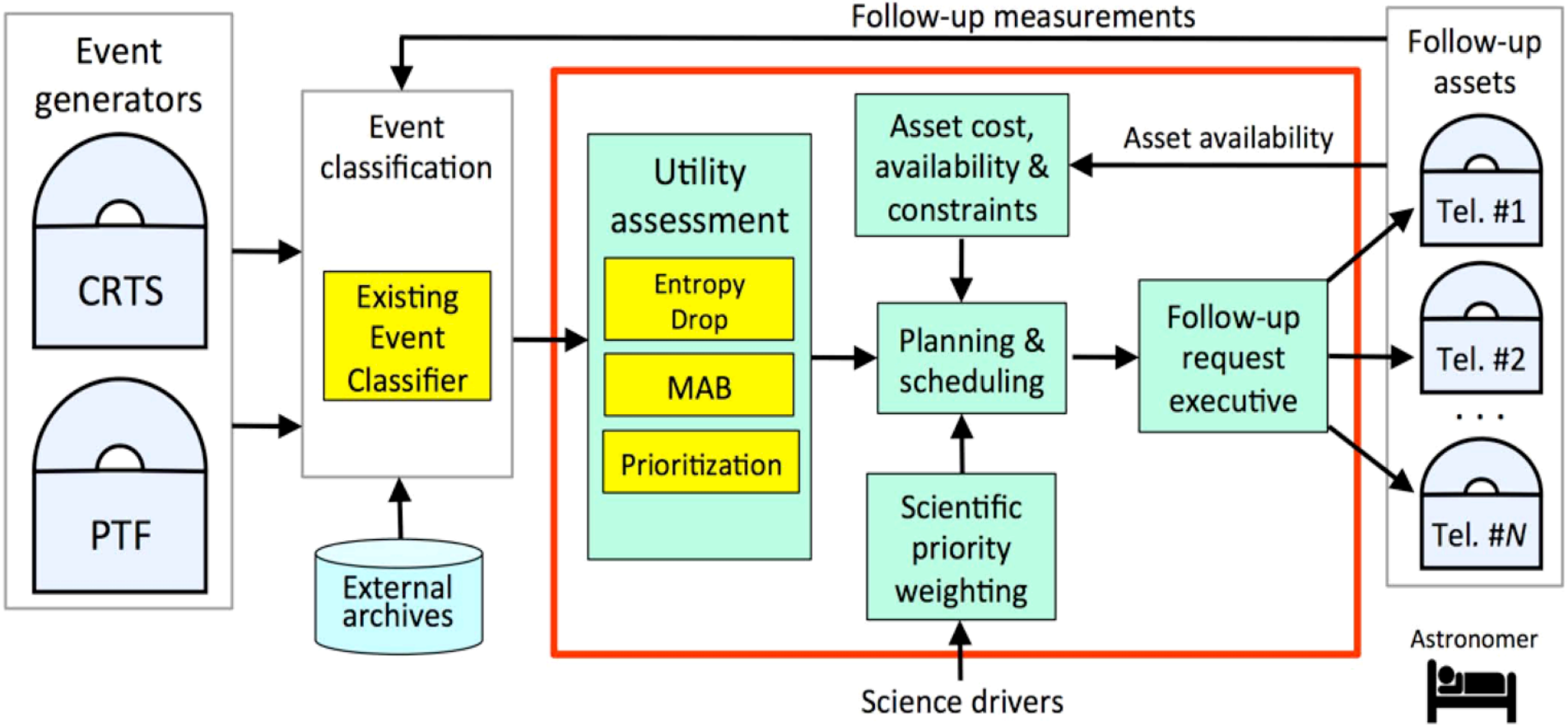
An optimal combining of classifiers



Exploring a variety of techniques for an optimal classification fusion: Markov Logic Networks, Diffusion Maps, Multi-Arm Bandit, Sleeping Expert…

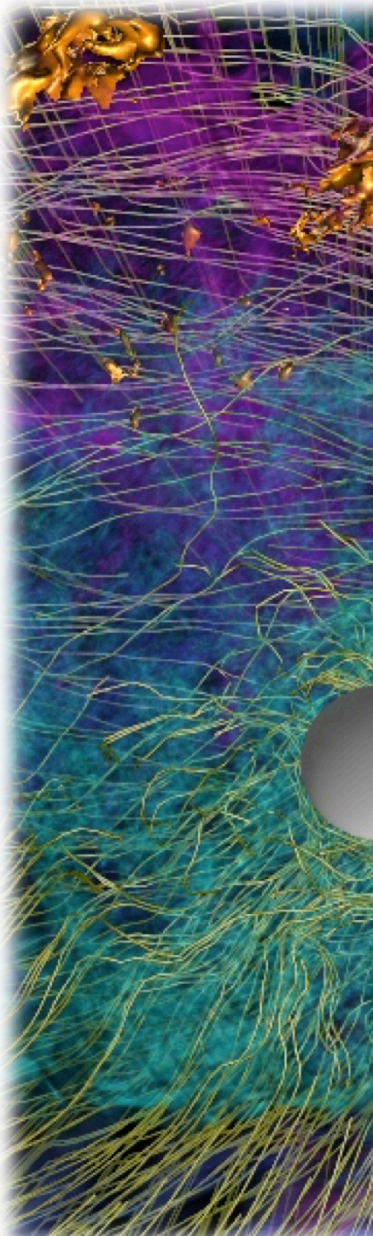
Automating the Optimal Follow-Up

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?

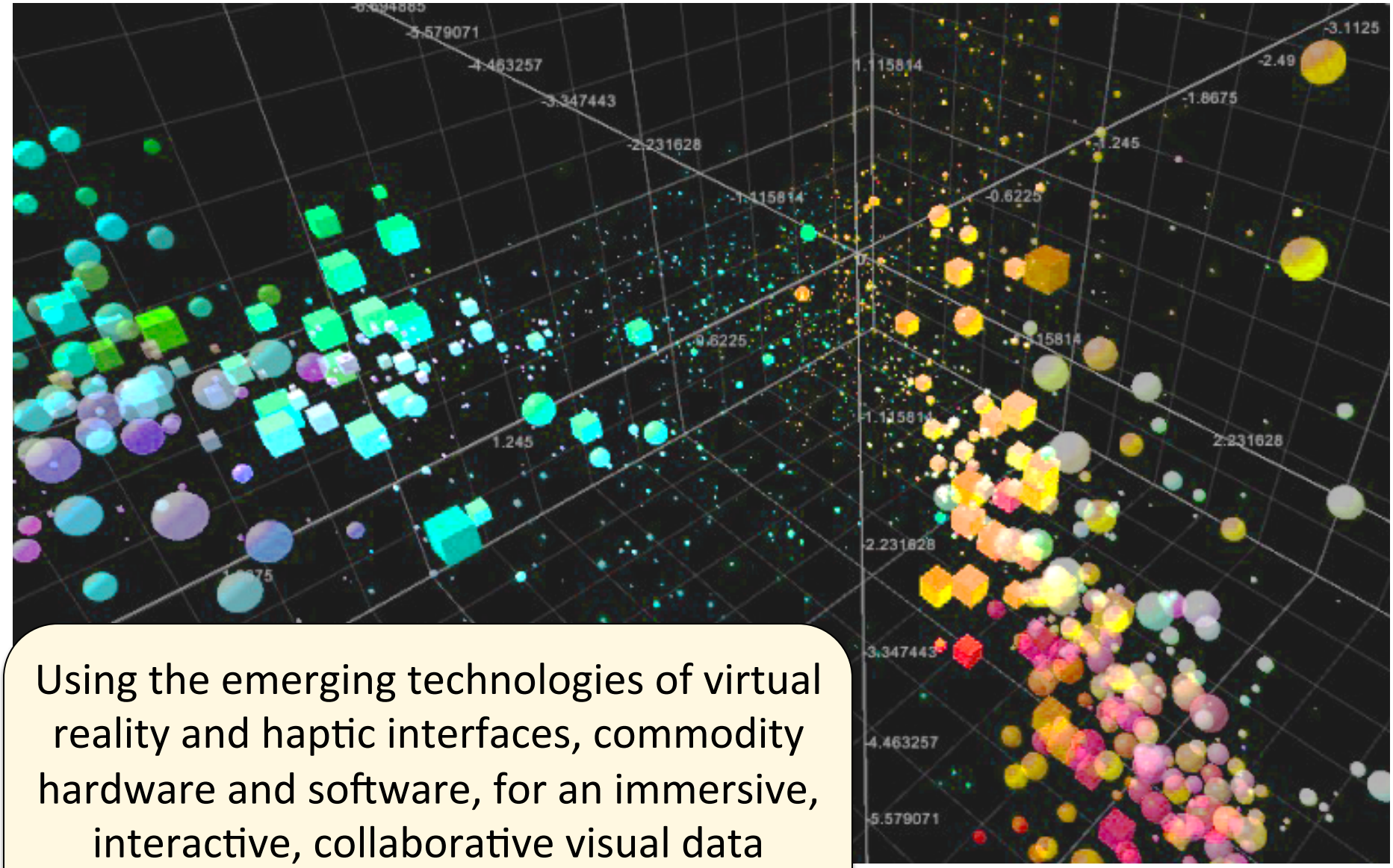


A Key Challenge: Visualizing Complexity

- Effective visualization is the bridge between quantitative information and human intuition
- What good are the data if we cannot effectively extract knowledge from them?
- Hyperdimensional structures (clusters, correlations, etc.) may be present in many complex data sets, with dimensionality in the range of $D \sim 10^2 - 10^4$, and will surely grow
- It is not only the matter of *data understanding*, but also of choosing the appropriate data mining algorithms, and interpreting the results
- We are biologically limited to perceiving ~ 3 - 12(?) dimensions



Innovative Data Visualization

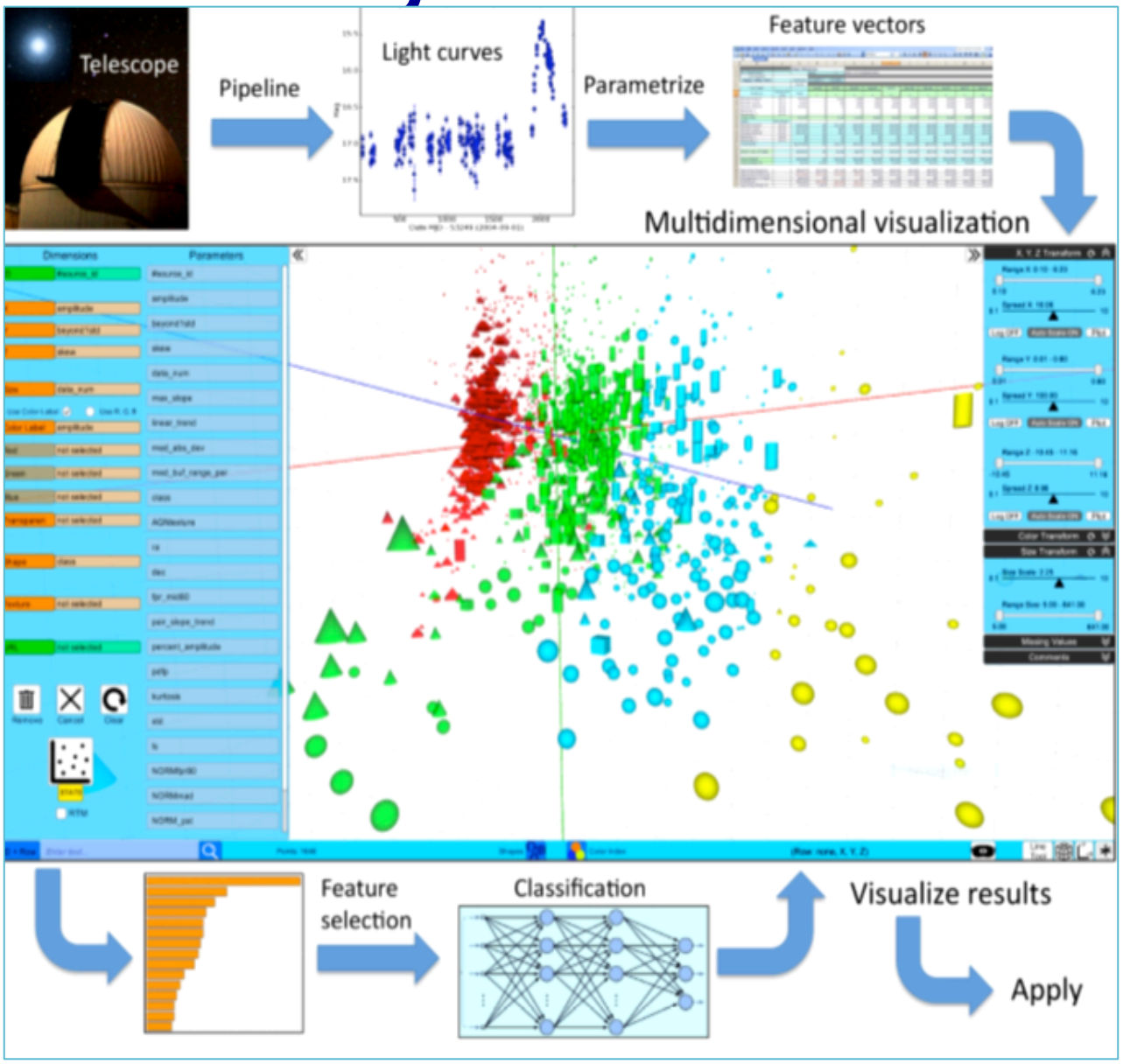


Using the emerging technologies of virtual reality and haptic interfaces, commodity hardware and software, for an immersive, interactive, collaborative visual data analytics and exploration

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

Data Visualization Using Virtual Reality

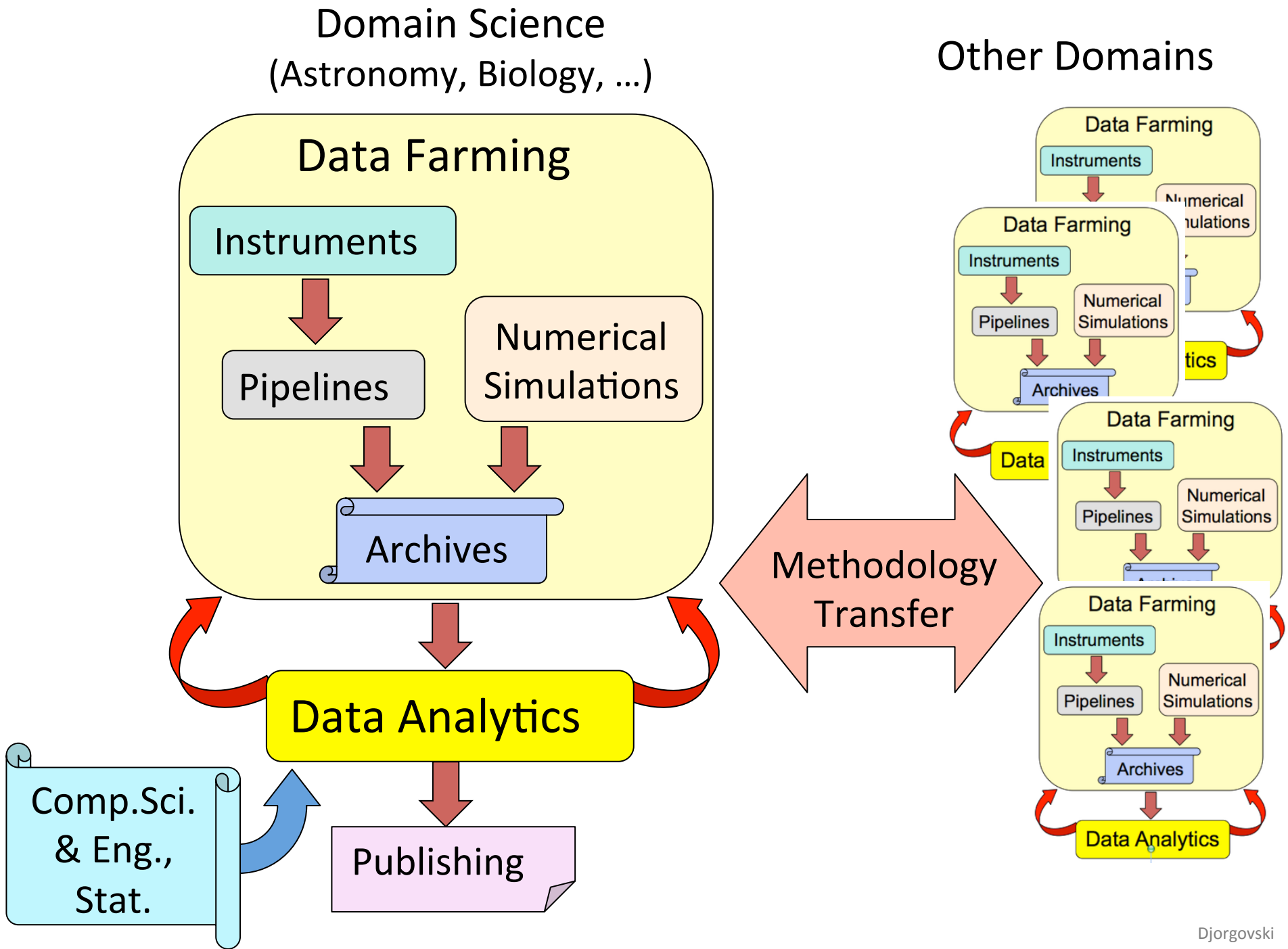
- Immersive, collaborative, easy to use tool for visual data analytics and exploration
- Use commodity hardware for immersive and augmented Virtual Reality



Data Science Methodology Transfer

There are common challenges and a common underlying methodology to much of the data science (computing, IT, ML, statistics…)

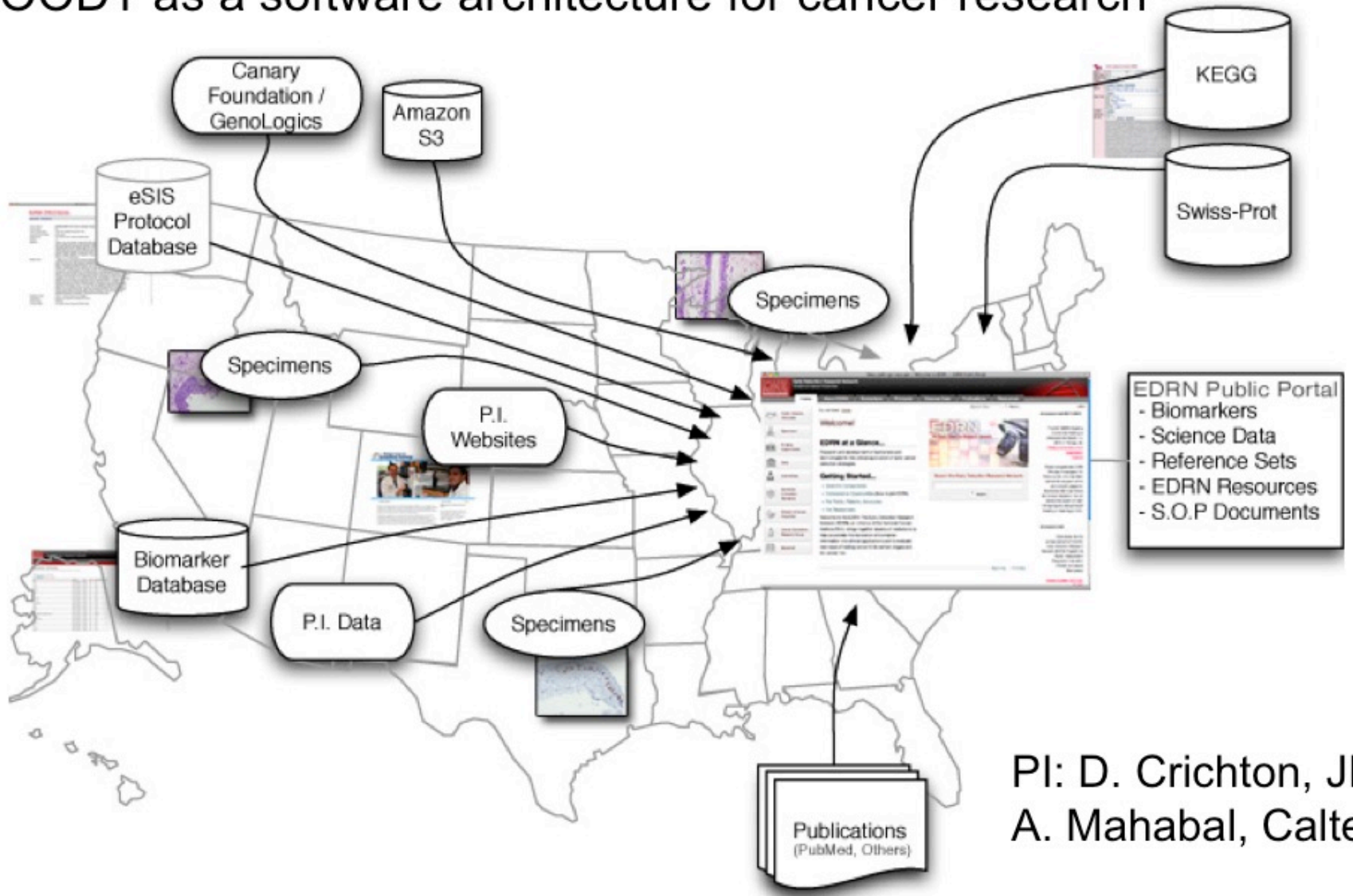
How can we transfer the cyberinfrastructure developments, experience, and solutions from one scientific domain to others?


Domain Science
(Astronomy, Biology, …)
Other Domains
Data Farming
Instruments
Pipelines
Numerical Simulations
Archives
Data Analytics
Publishing
Comp.Sci. & Eng., Stat.
Methodology Transfer
Data Farming
Instruments
Pipelines
Numerical Simulations
Archives
Data Analytics
Data
tics

Djorgovski

EDRN: A Virtual, National Integration Cancer Biomarkers Knowledge System

OODT as a software architecture for cancer research

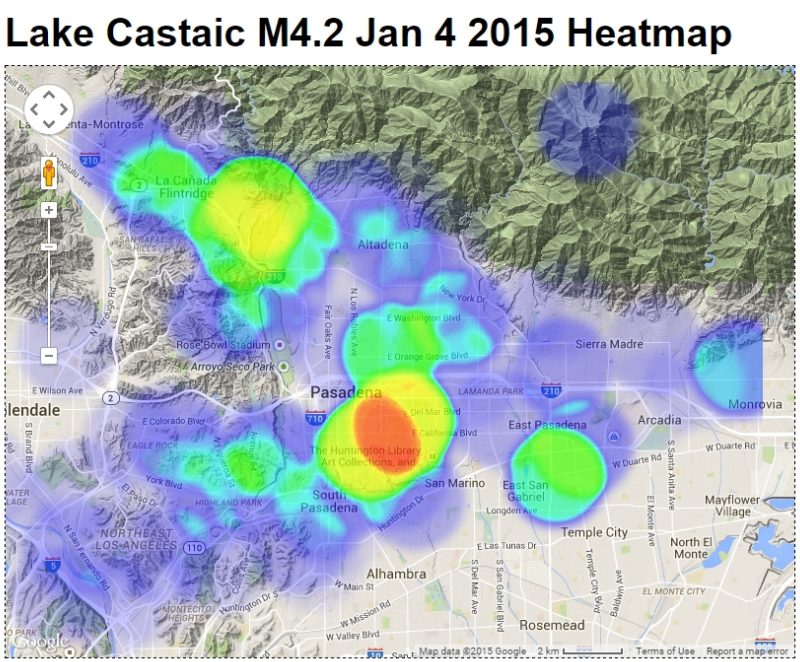


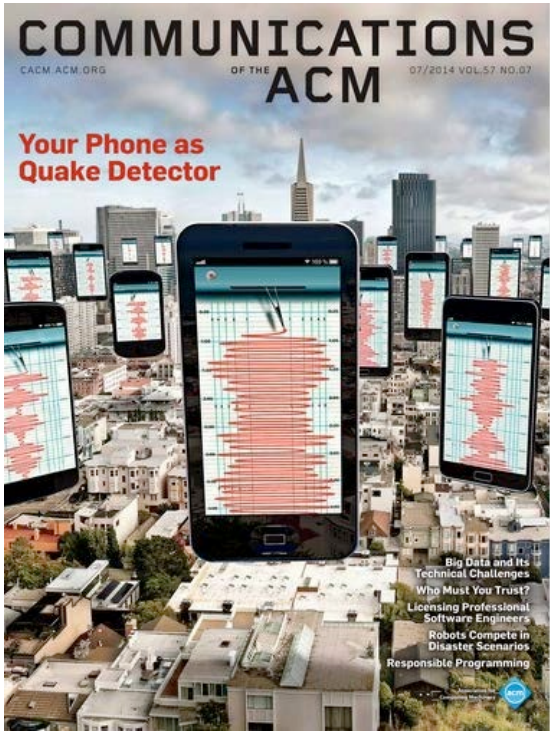
PI: D. Crichton, JPL
A. Mahabal, Caltech

Real Time Classification and Response

Seismology:
Cell phones as a sensor network

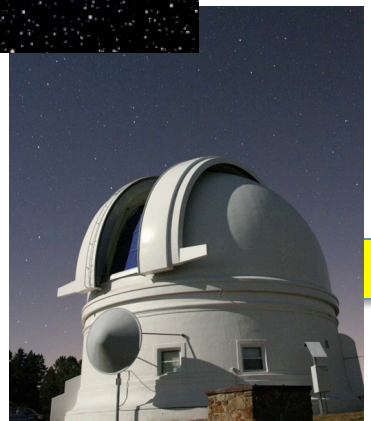
Time domain astronomy



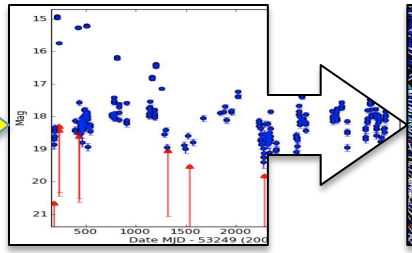


Event

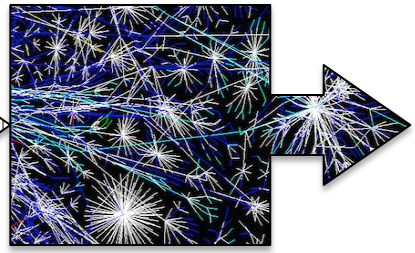
Detection



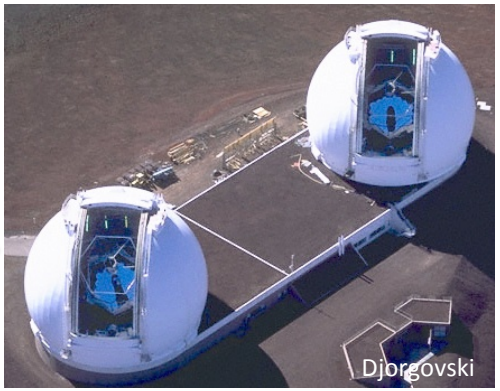
Classification



Decision making



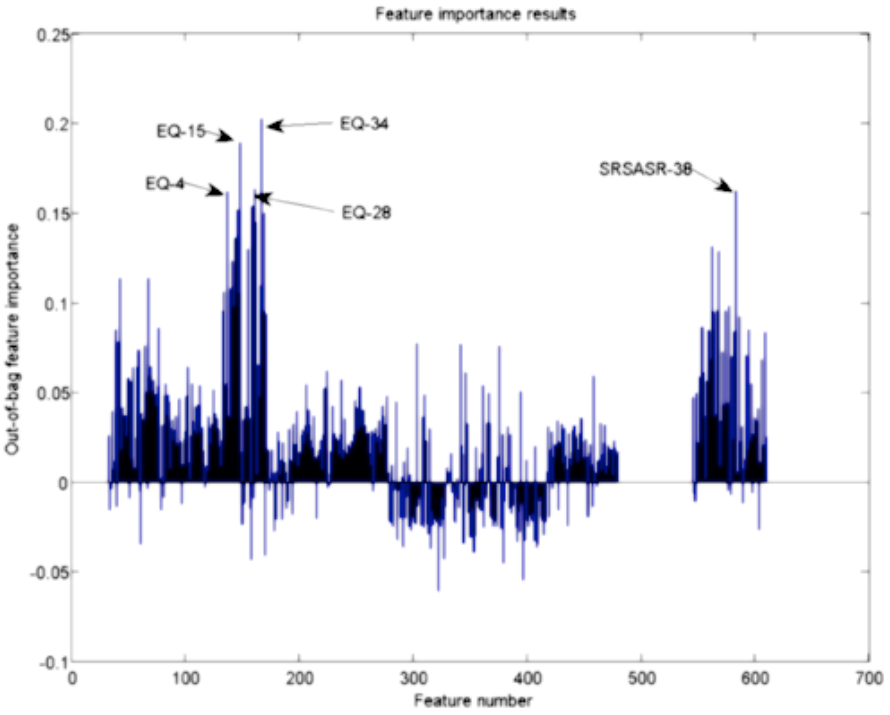
Follow-up



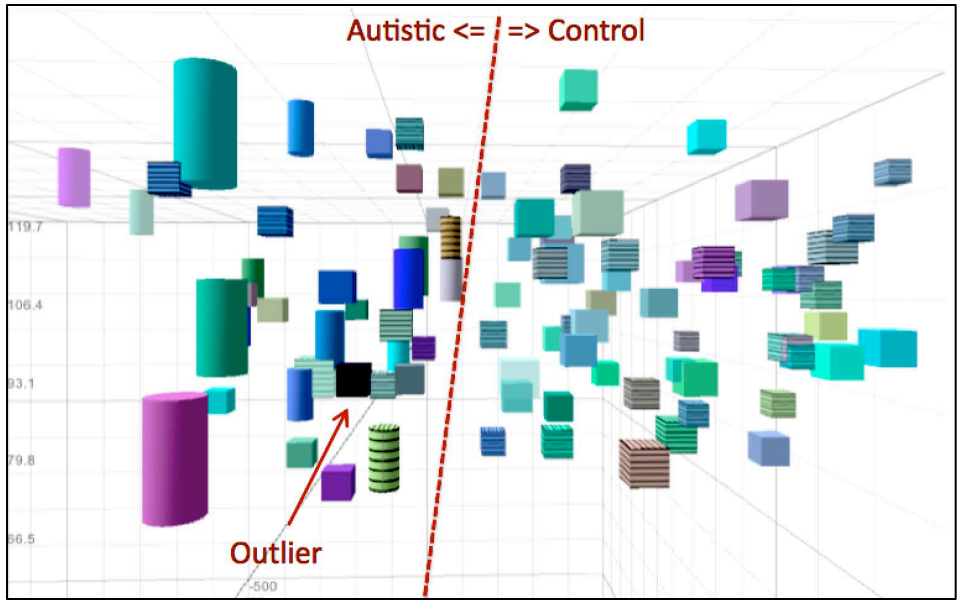
From Sky Surveys to Neurobiology

- Using the data analytics tools based on ML, developed for the analysis of sky surveys and from high-energy physics , to design a better diagnostics for autism

(with R. Adolphs et al.)



Feature importance using random forests =>



<= Classification in a multidimensional feature space

Next: machine learning applied to MRI scans

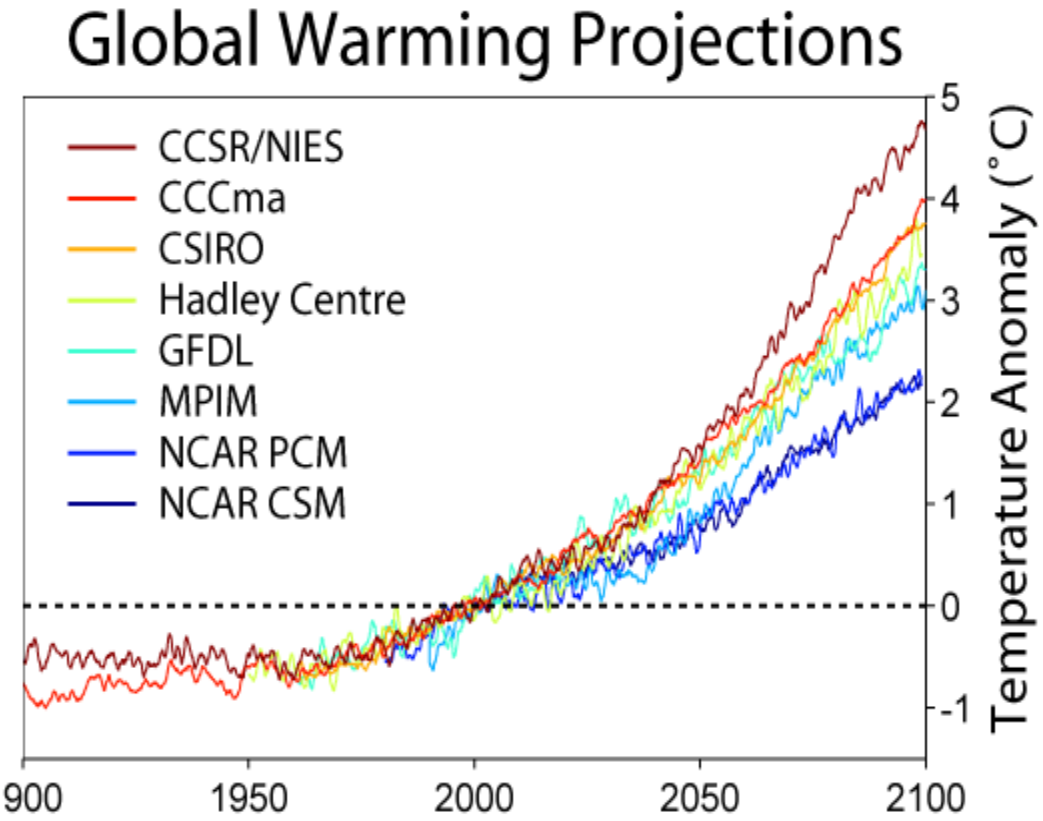
Quantifying a Model Uncertainty

… Whether the data come from measurements or from the output of numerical models and simulations

The sources of uncertainty:

- Measurement errors
- Numerical errors
- Sample sizes
- Processing algorithms
- Data representation
- Data mining choices and their implementations

… etc. etc.



Training the Next Generation

- We are developing a new curriculum about the tools and methods for computational, data-intensive research in the 21st century
- Making use of on-line educational technologies (e.g., MOOCs) to maximize our impact
- An example: ***the first virtual summer school on big data analytics***, a joint venture with JPL:

Summary: CD^3 Capabilities

- Creation of effective data commons (interoperability, middleware, standards, …)
- Analysis of high-dimensionality data/parameter spaces: machine learning, visualization using VR/AR
- Real-time characterization and classification of rare events in massive data streams
- Analysis and periodicity searches in irregular time series
- Data science methodology transfer between different domains
- On-line training in data science methodology
- Upcoming projects: a framework for uncertainty quantification; machine learning for cyber-security