Machine Learning from End-user Database Annotations

Kirk Borne

George Mason University

19 June 2010

Outline

Talk will have 2 Practical Semantic Astronomy parts:

1)Astroinformatics for Eventful Astronomy

2)Machine Learning in the Galaxy Zoo Database

Part 1

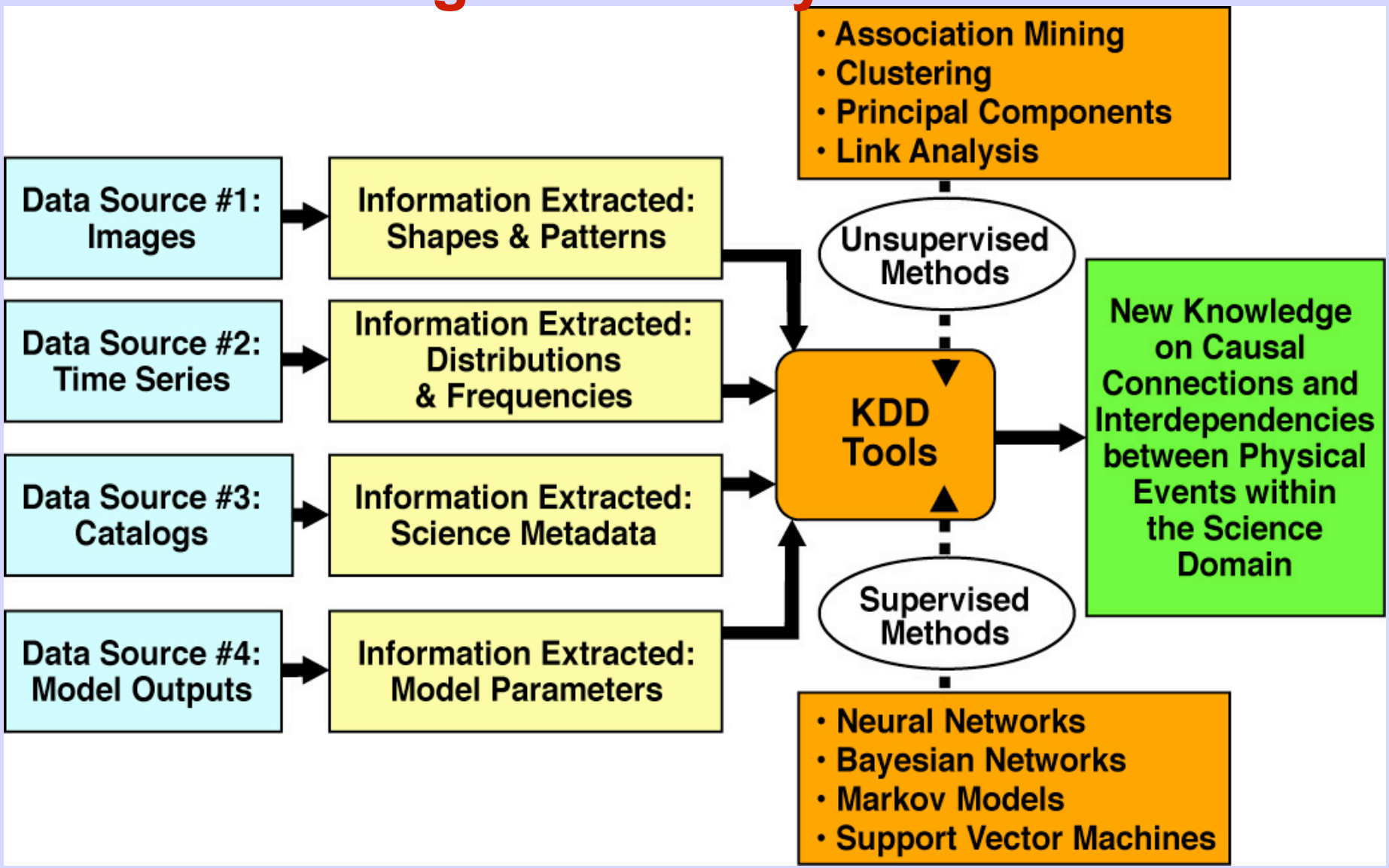
Astroinformatics for Eventful Astronomy

The LSST will represent a 10K-100K times increase in the VOEvent network traffic. This poses significant real-time classification demands on the event stream:

from data to knowledge!

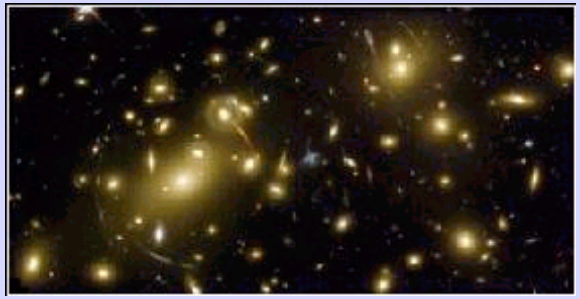
<u>from sensors to sense!</u>

Scientific KDD: Knowledge Discovery in Databases

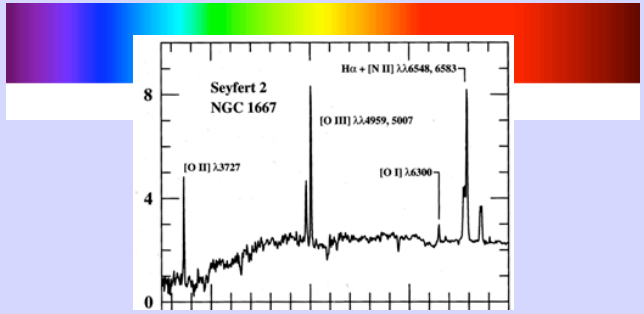


Astronomy Example

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 ➡

Redshift = Distance

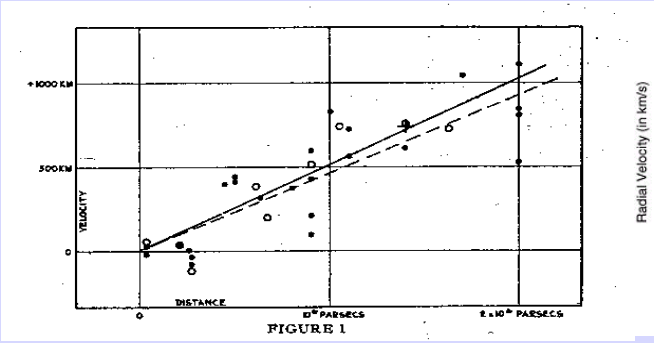
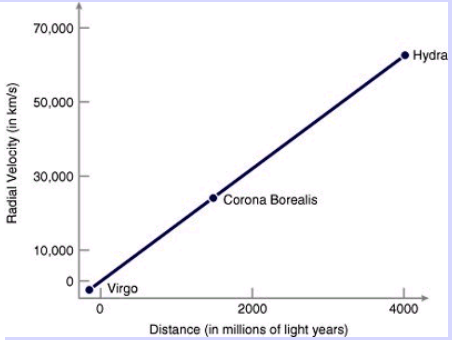
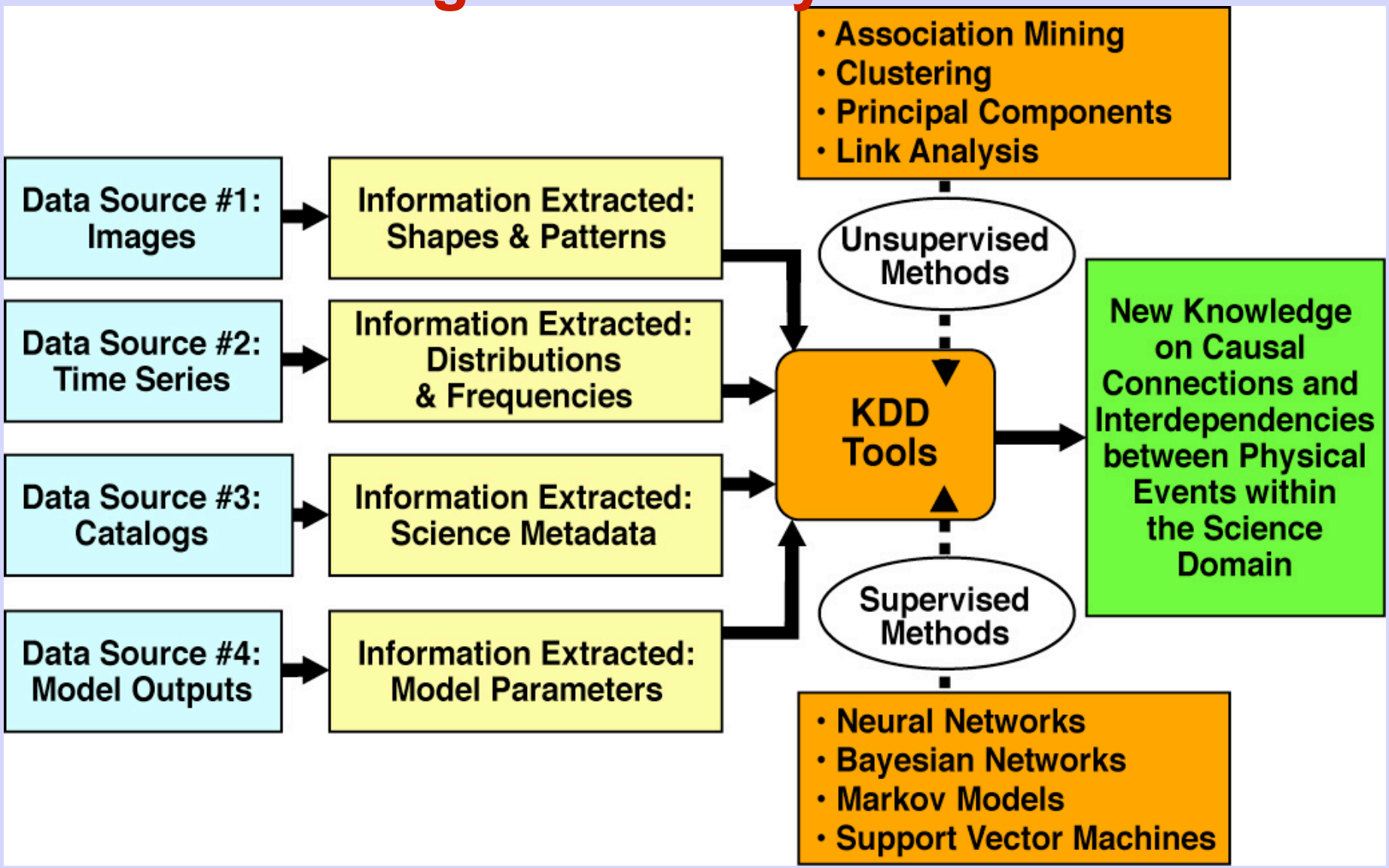


FIGURE 1

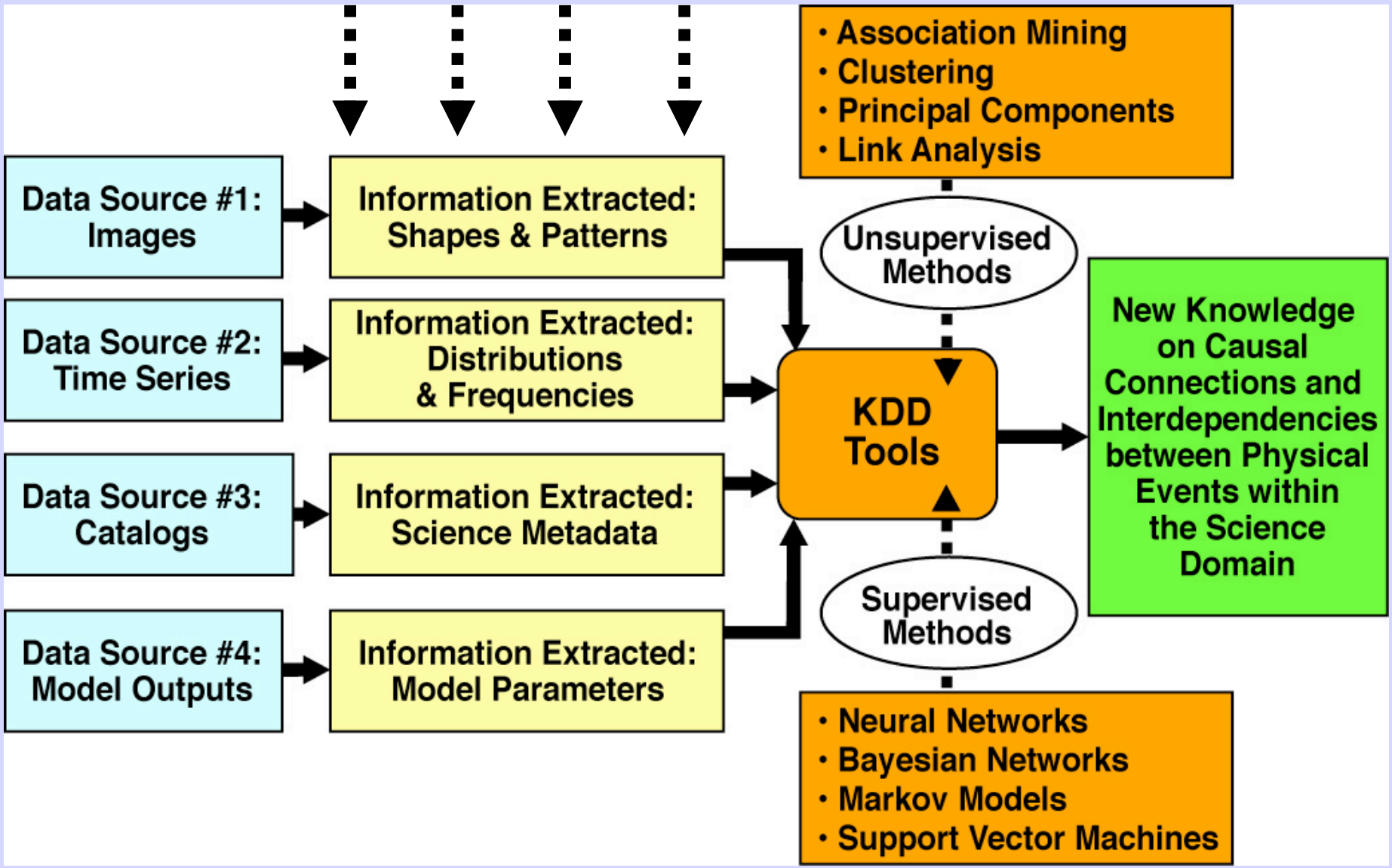


Understanding: the Universe is expanding!!

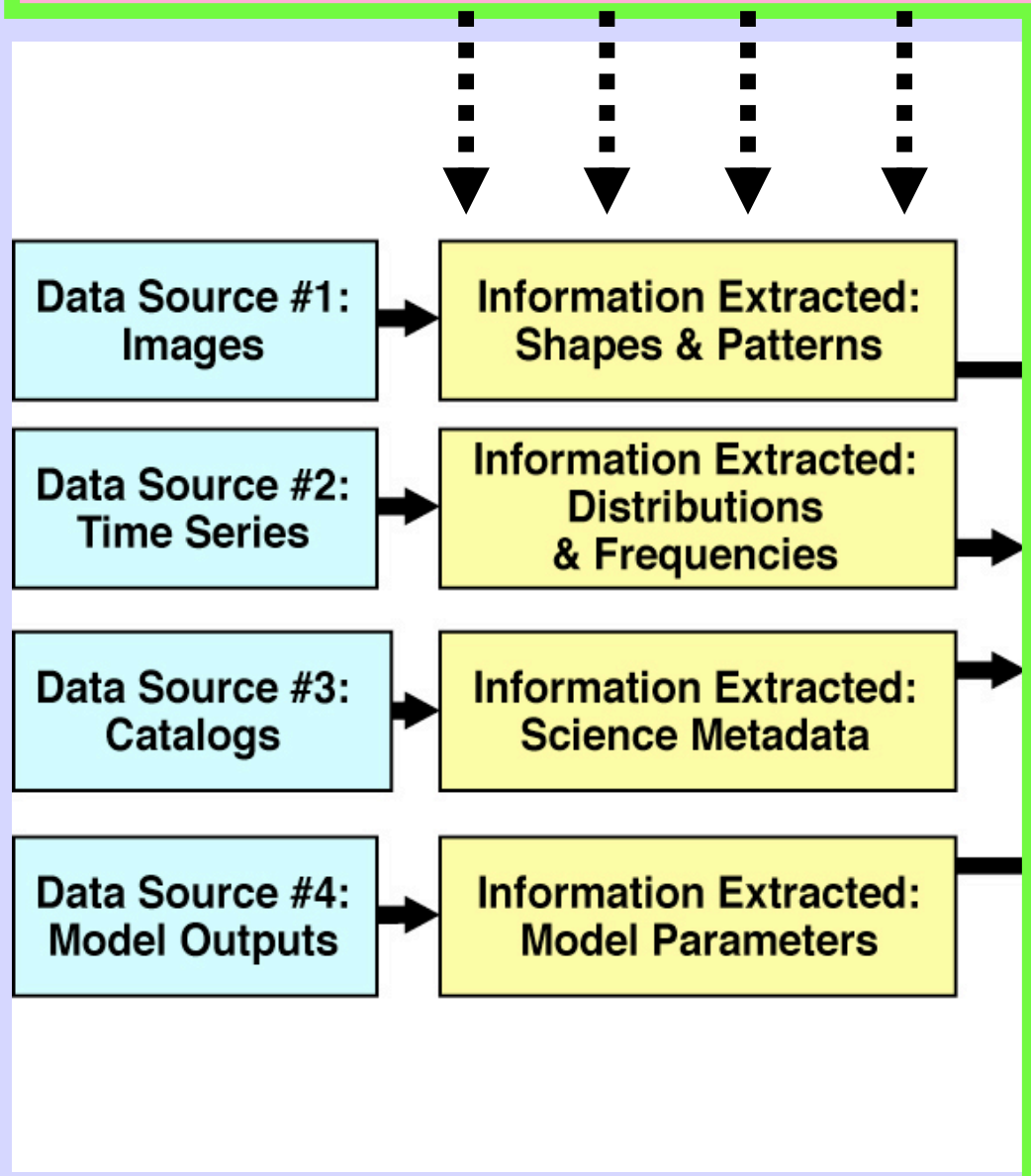
Scientific KDD: Knowledge Discovery in Databases



This is the Informatics Layer



This is the Informatics Layer



Informatics Layer:

- Provides standardized representations of the "information extracted" – for use in the KDD (data mining) layer.
- Standardization is not required (nor feasible) at the "data source" layer.
- The informatics is discipline-specific.
- Informatics enables KDD across large distributed heterogeneous scientific data repositories.

Astroinformatics for Eventful Astronomy

- Report discoveries back to the science database for community reuse
- Basic astronomical objects (informatics granules) are annotated ...
  - with follow-up observations of any kind
  - with new knowledge discovered
  - with common knowledge
  - with inter-relationships between objects and their properties
  - with concepts
  - with context
  - With assertions (e.g., classifications, concepts, quality flags, relationships, references, observational parameters, common knowledge, inter-connectivity with other objects)

    Semantics!
  - with experimental parameters
  - with observer / observatory descriptors

    Provenance (traceback)
- Enables knowledge-sharing and reuse

Astroinformatics for Eventful Astronomy

- In order to facilitate filtering and prioritization of events for rapid follow-up observations, a near real-time **characterization** provider of tags (end-user annotations) for each object and event is needed.

- The semantic integration of real-time survey data products with federated VO-accessible archival information resources will facilitate the sharing of knowledge-rich quantifiable astronomical features (event characterizations) to the research community.

- An astroinformatics-enabled characterization service for large sky surveys provides uniform tags, metadata, labels, terminology.

- Use cases of the characterization service include knowledge capture, annotation, data mining, & queries of distributed knowledgebases.

- The addition of human-provided annotations and semantic tagging, in structured form, will enhance and improve eventful astronomy research and worldwide astronomical knowledge.

Part 2

Machine Learning in the Galaxy Zoo Database

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

ZOONIVERSE
REAL SCIENCE ONLINE

http://zooniverse.org/

Modes of Computing

- **Numerical Computation (*in silico*)**
  - Fast, efficient
  - Processing power is rapidly increasing
  - Model-dependent, subjective, only as good as your best hypothesis
- **Computational Intelligence**
  - Data-driven, objective (machine learning)
  - Often relies on human-generated training data
  - Often generated by a single investigator
  - Primitive algorithms
  - Not as good as humans on most tasks
- **Human Computation (*Carbon-based Computing*)**
  - Data-driven, objective (human cognition)
  - Creates training sets, Cross-checks machine results
  - Excellent at finding patterns, image classification
  - Capable of classifying anomalies that machines don't understand
  - Slow at numerical processing, low bandwidth, easily distracted

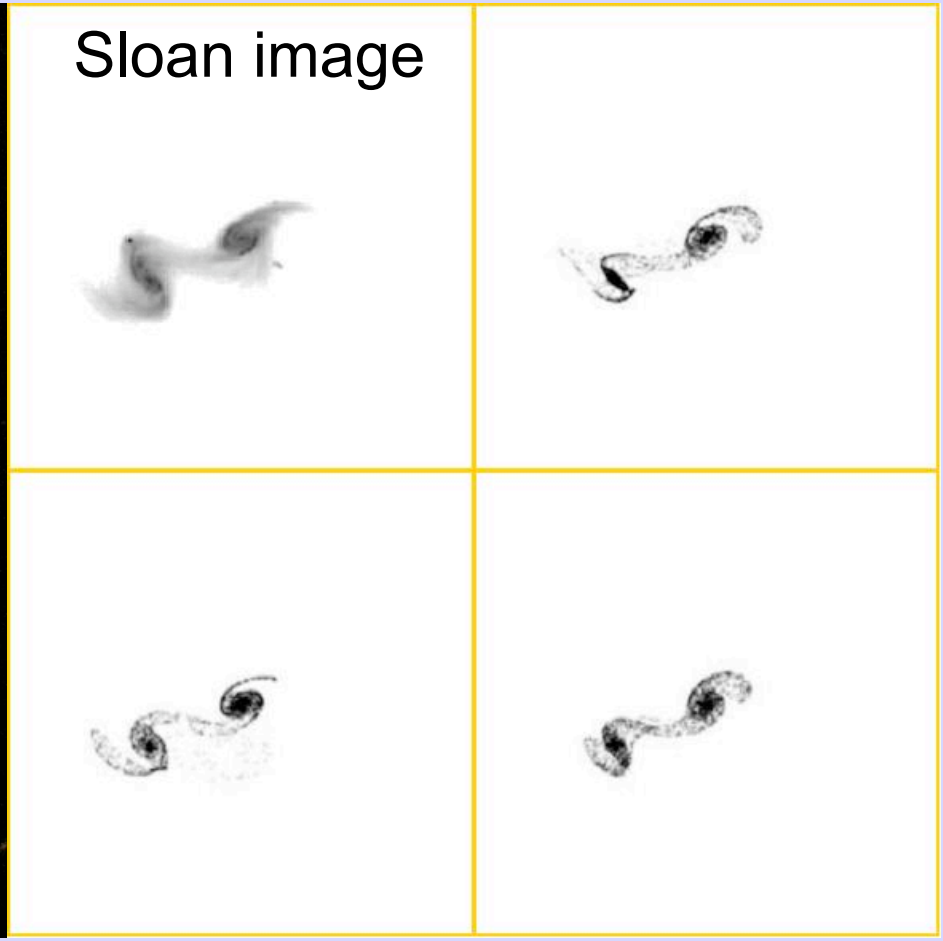
The Zooniverse: a Buffet of Zoos

http://zooniverse.org/

- Galaxy Zoo project (released July 2007):
  - http://www.galaxyzoo.org/
  - Classify galaxies (Spiral, Elliptical, Merger, or image artifact)
- Galaxy Merger Zoo (release November 2009)
  - http://mergers.galaxyzoo.org/
  - Run N-body simulations to find best model to match a real merger
  - One new merger every day
- The Hunt for Supernovae (released December 2009)
  - http://supernova.galaxyzoo.org/
  - Real-time event detection and classification
- Solar Storm Watch (released March 2010)
  - http://solarstormwatch.com/
  - Spot solar storms (CMEs) in near real-time

Merging/Colliding Galaxies are the building blocks of the Universe: **1 + 1 = 1**

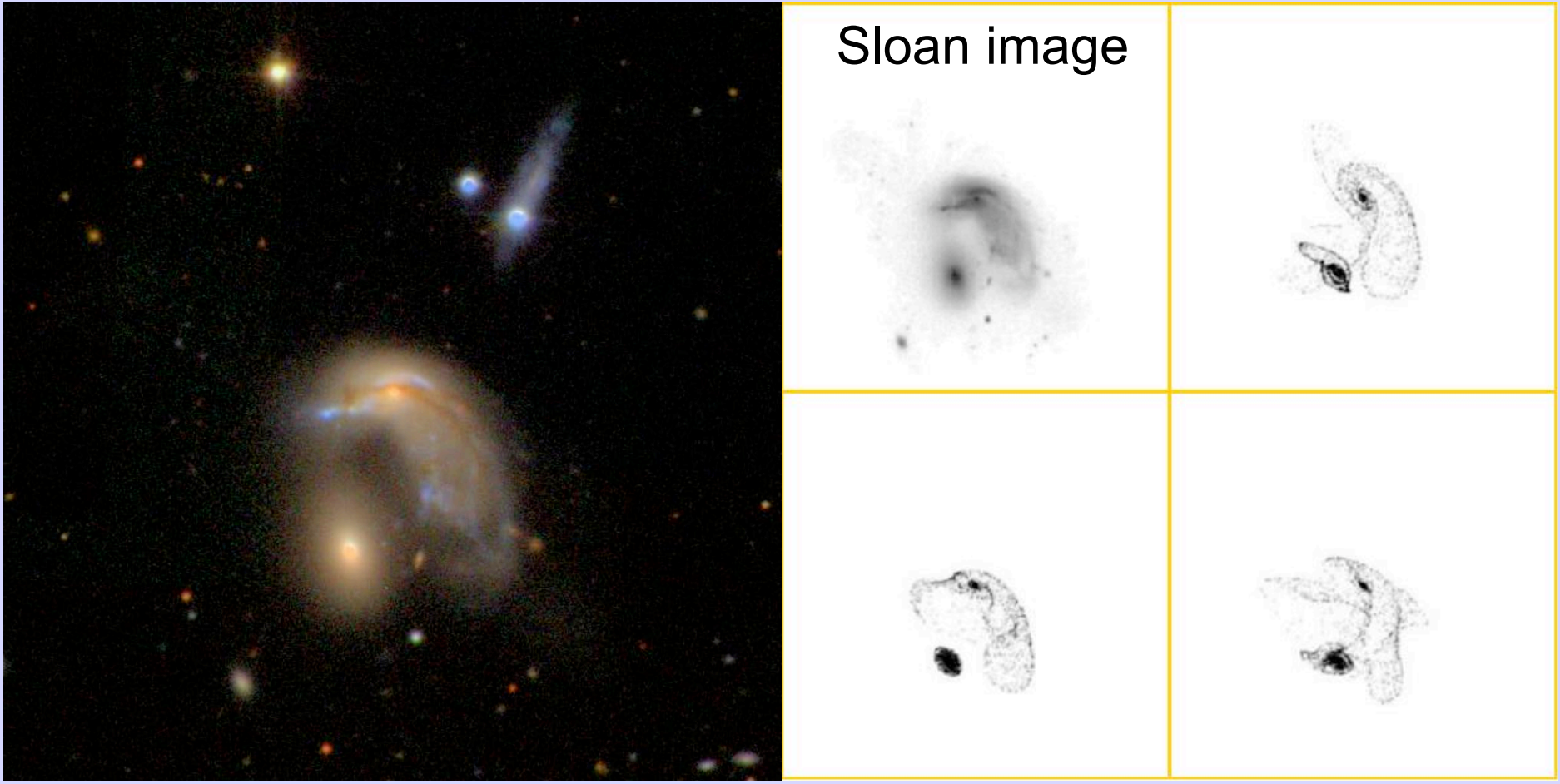
Galaxy Mergers Zoo Gallery



SDSS 587722984435351614

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

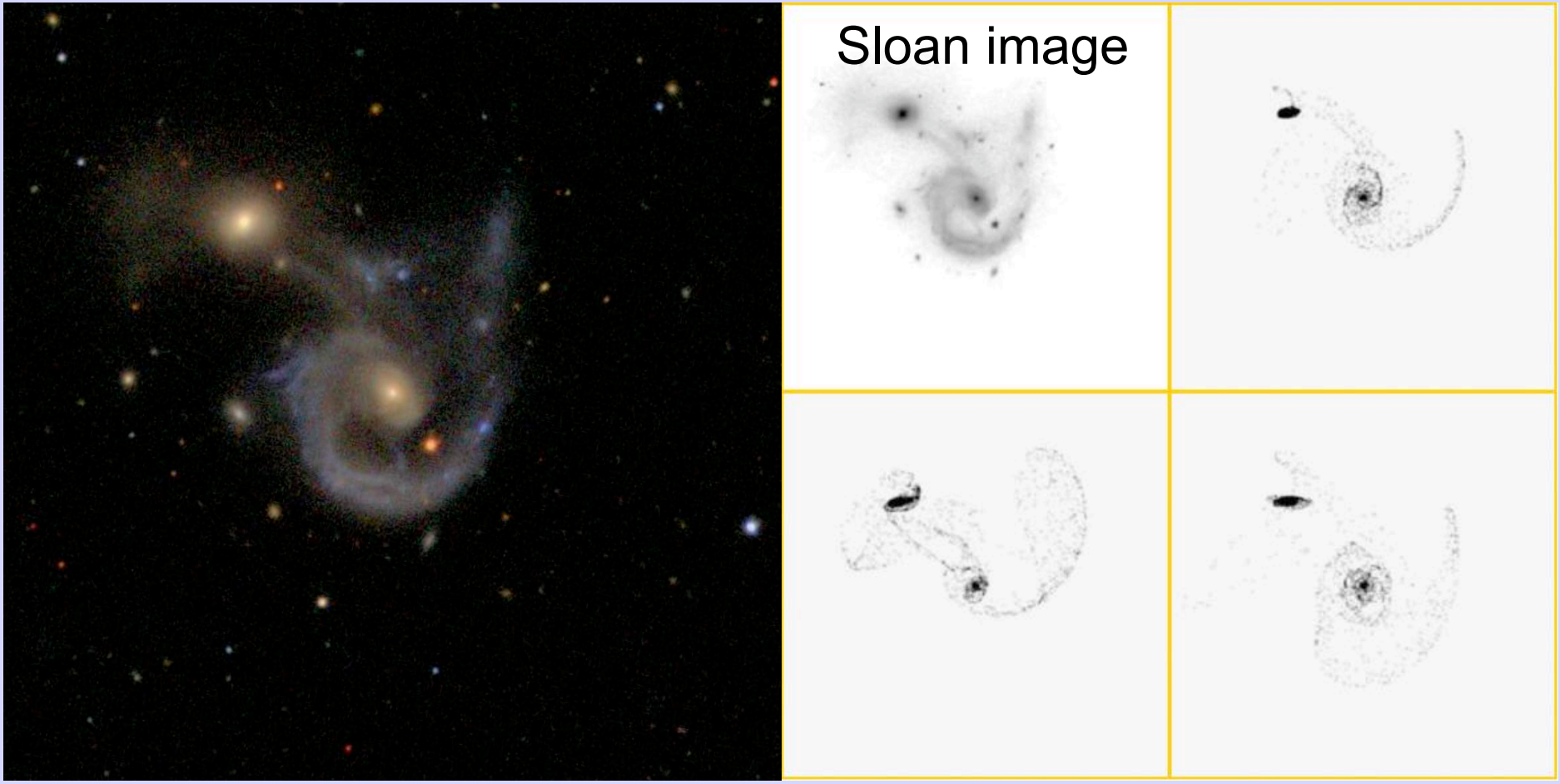
Galaxy Mergers Zoo Gallery



SDSS 587726033843585146

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

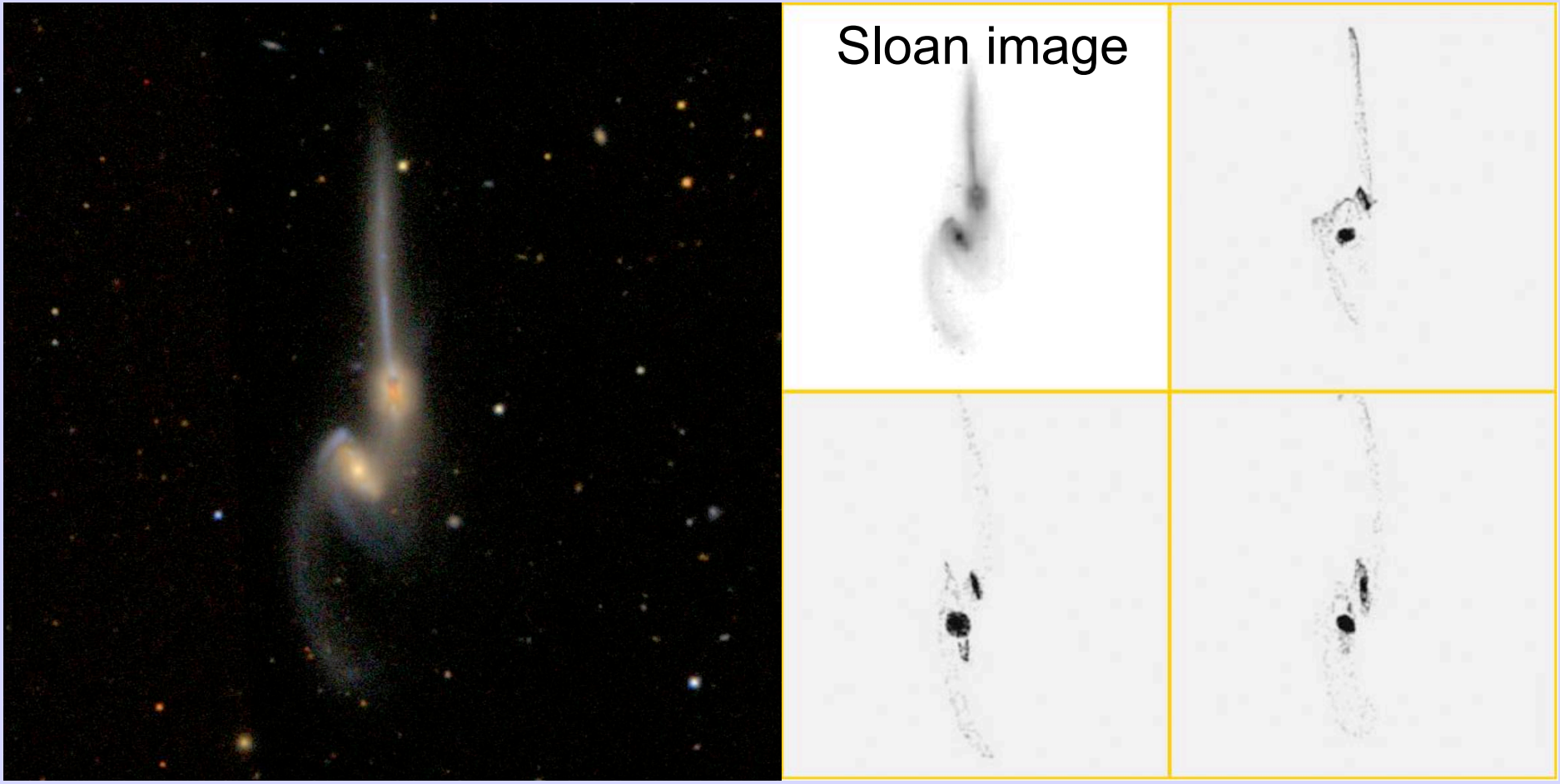
Galaxy Mergers Zoo Gallery



SDSS 587739646743412797

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

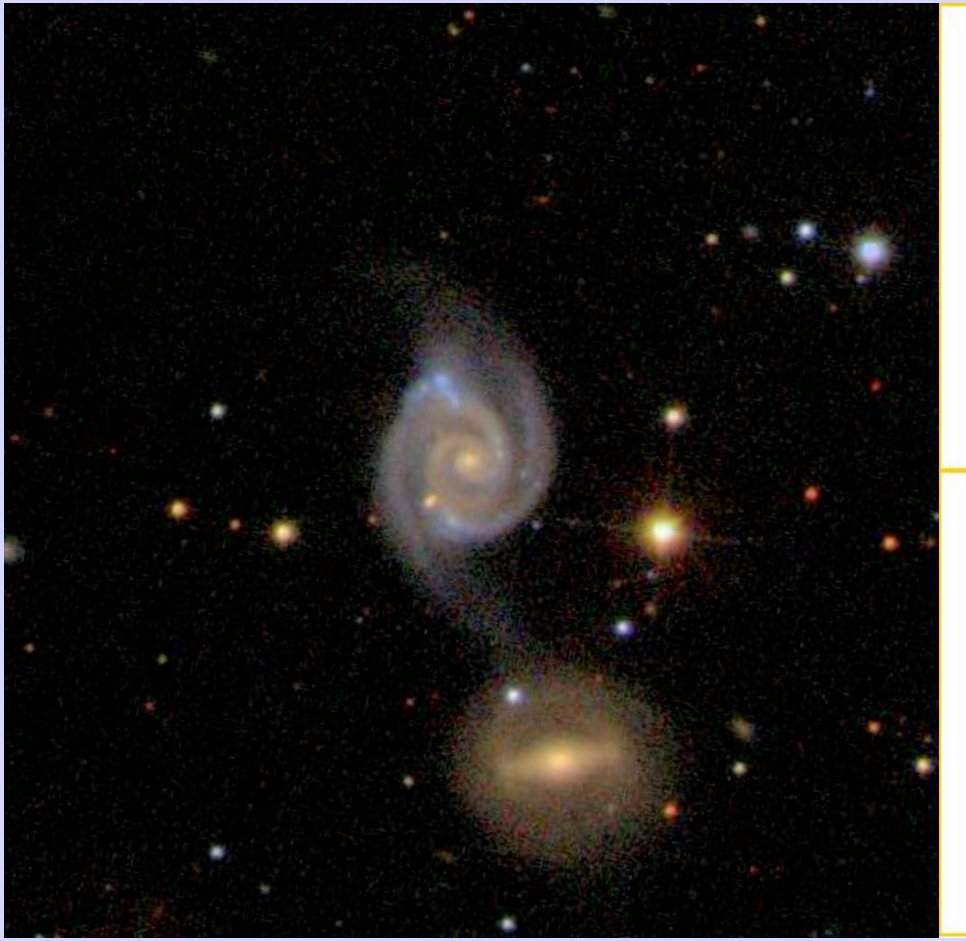
Galaxy Mergers Zoo Gallery

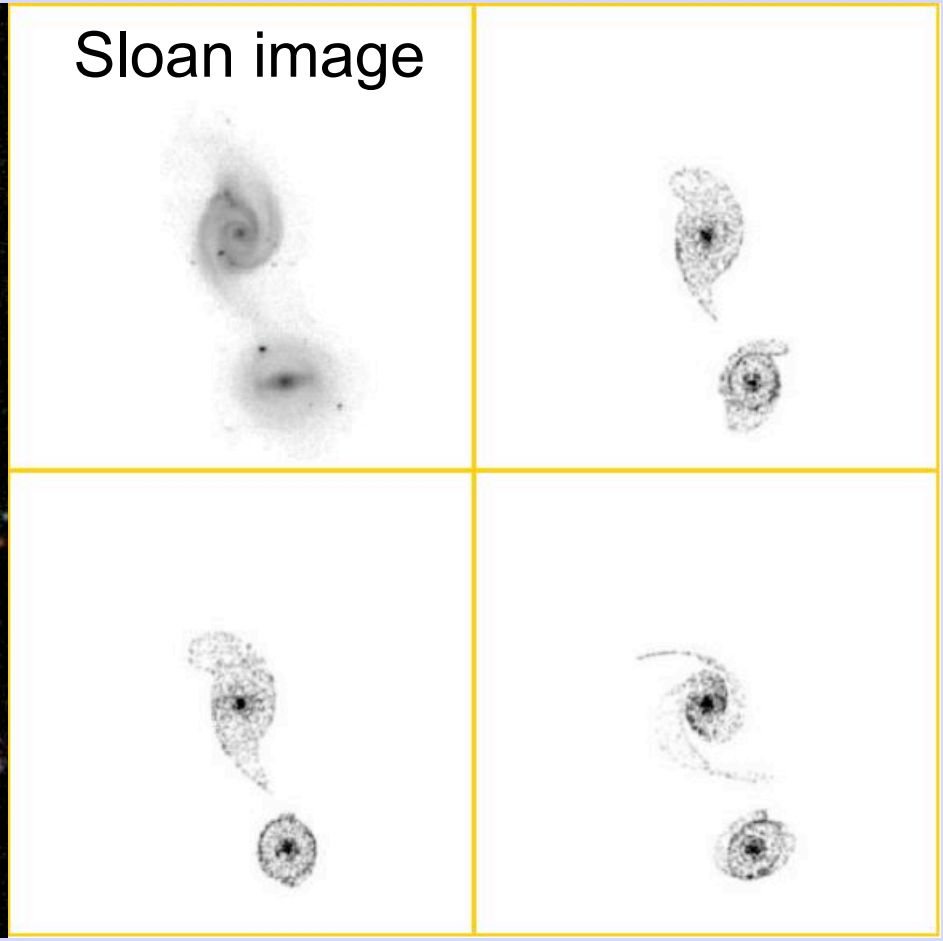


SDSS 587739721900163101

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery

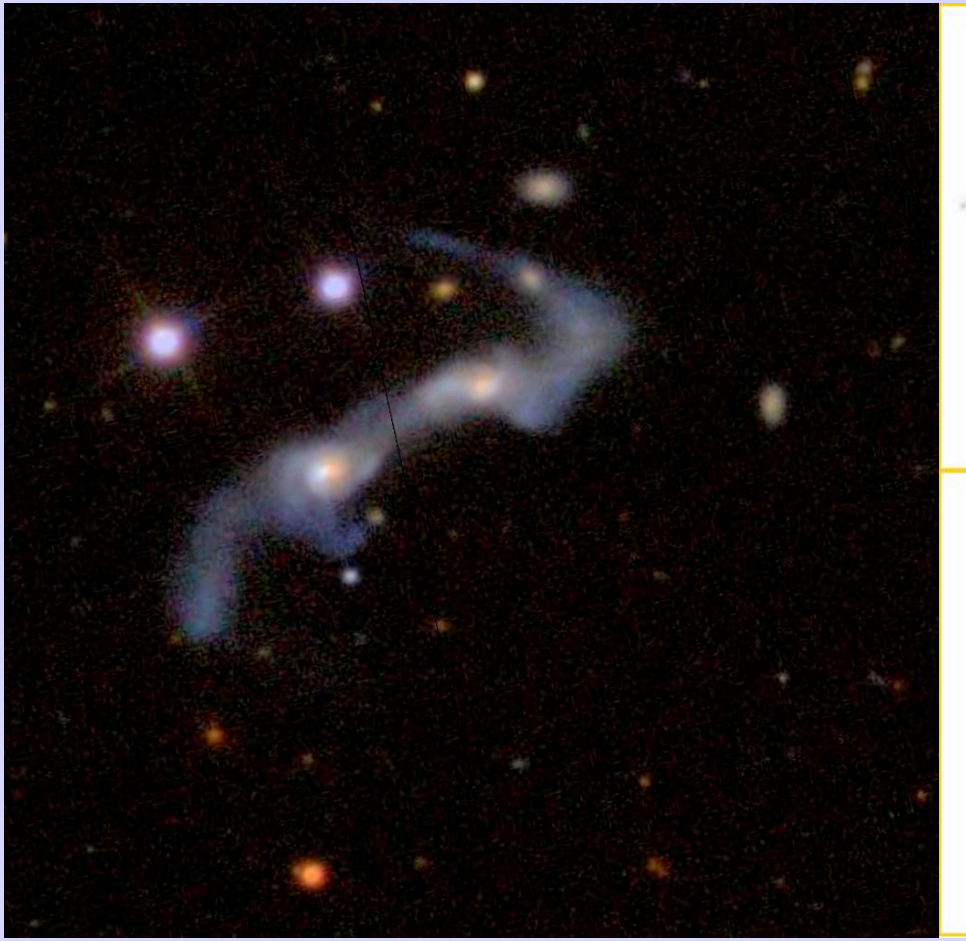


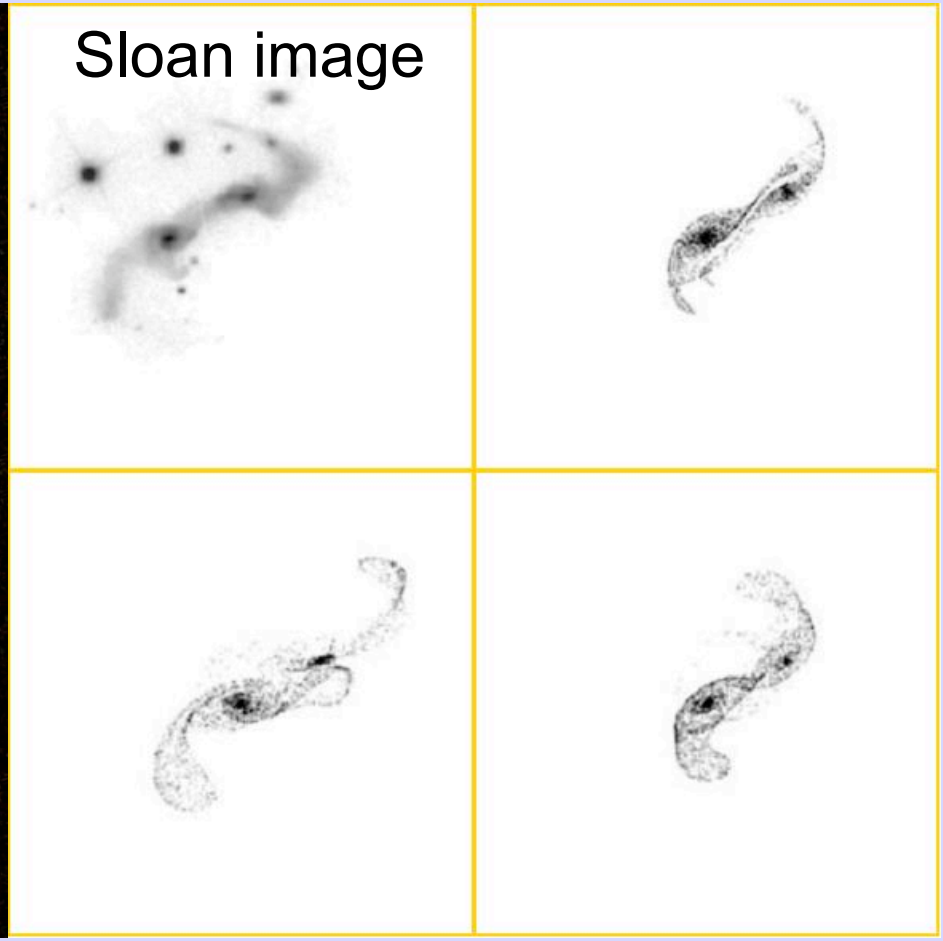


SDSS 587727222471131318

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery





SDSS 588011124116422756

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Tags produce a new data flood

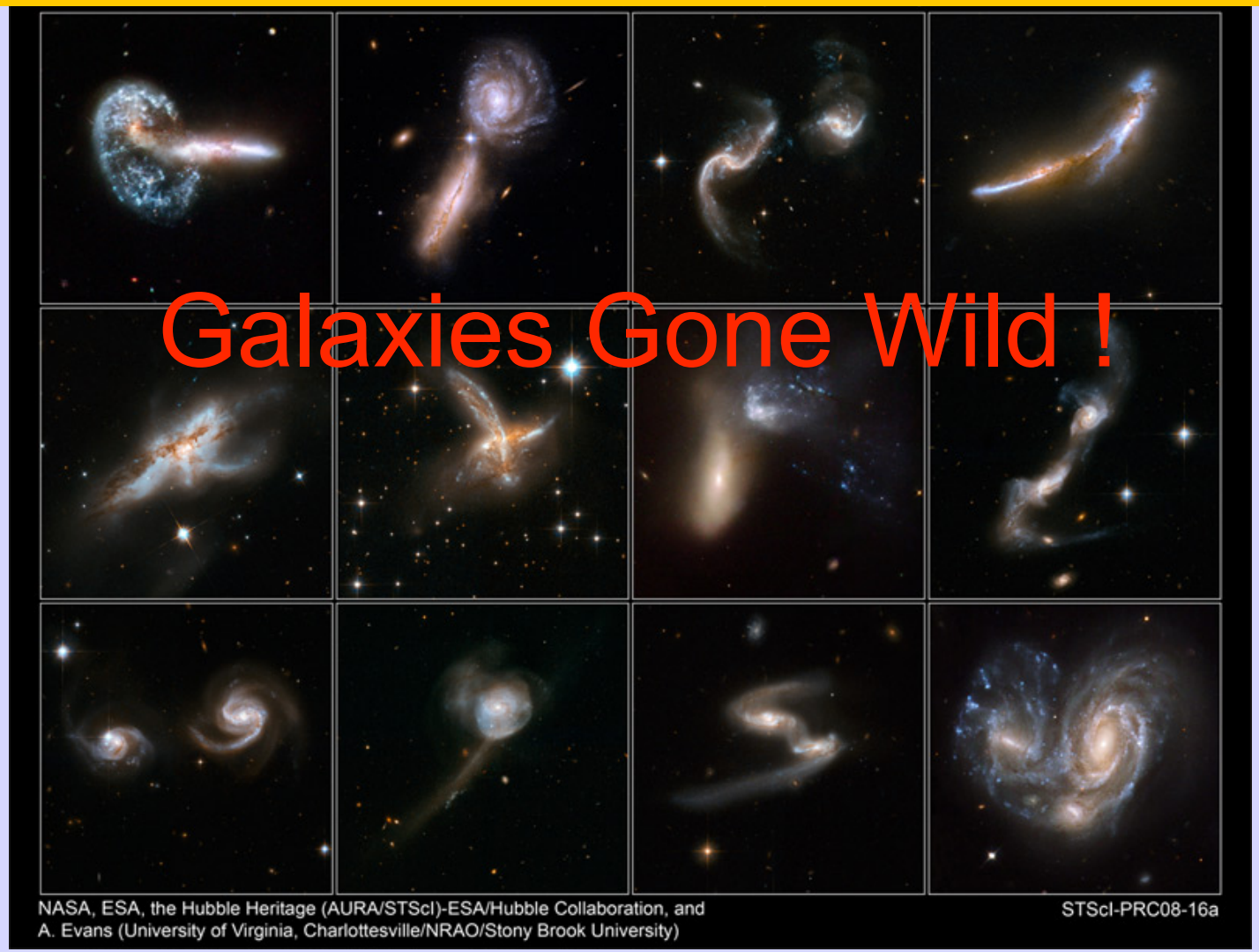
- **Tagging enables semantic data fusion & integration, and knowledge acquisition / representation / sharing.**
- User-contributed content adds **more** data to the data flood.
- Example – Galaxy Zoo project:
  - ~250,000 participants (*and growing*)
  - ~1 million galaxies have been labeled (classified)
  - ~200 million classifications have been collected
- Tagging is applicable to any data source, including document repositories – adding lightweight semantics to the data repository (taxonomies, folksonomies, annotations)
- Reference: *"TagLearner: A P2P Classifier Learning System from Collaboratively Tagged Text Documents",* Dutta, Zhu, Mahule, Kargupta, Borne, Lauth, Holz, & Heyer, 2009 ICDM paper.

Key Feature of Zooniverse:
Data mining from the volunteer-contributed labels

- Train the automated pipeline classifiers with:
  - Improved classification algorithms
  - Better identification of anomalies
  - Fewer classification errors
- Millions of training examples
- Hundreds of millions of class labels
- Statistics deluxe! …
  - Users (see paper: http://arxiv.org/abs/0909.2925)
  - Uncertainty quantification
  - Classification certainty vs. Classification dispersion

ZOONIVERSE
REAL SCIENCE ONLINE

First Case Study: test SDSS science catalog attributes to find which attributes correlate most strongly with user-classified mergers.



Sloan Science Database Attributes tested

Attribute	Description
$petroMag_{ug}$	Petrosian magnitude colors. A color was calculated for four independent pairs of bands in SDSS (u, g, r, i, z).
$petroRad_u * z$	Petrosian radius, transformed with redshift to be distance-independent.
$invConIndx_u$	Inverse concentration index. The ratio of the 50% Petrosian magnitude to the 90% Petrosian magnitude.
$isoRowcGrad_u * z$	Gradient of the isophotal row centroid, transformed with redshift to be distance-independent.
$isoColcGrad_u * z$	Gradient of the isophotal column centroid, transformed with redshift to be distance-independent.
$isoA_u * z$	Isophotal major axis, transformed with redshift to be distance-independent.
$isoB_u * z$	Isophotal minor axis, transformed with redshift to be distance-independent.
$isoAGrad_u * z$	Gradient of the isophotal major axis, transformed with redshift to be distance-independent.
$isoBGrad_u * z$	Gradient of the isophotal minor axis, transformed with redshift to be distance-independent.
$isoPhiGrad_u * z$	Gradient of the isophotal orientation, transformed with redshift to be distance-independent.
$texture_u$	Measurement of surface texture.
$lnLExp_u$	Log-likelihood of exponential profile fit.
$lnLDeV_u$	Log-likelihood of De Vaucouleurs profile fit.
$fracDev_u$	Fraction of the brightness profile explained by the De Vaucouleurs profile.

Results of Decision Tree Information Gain analysis

Attribute	Information Gain
$lnLExp_g$	0.10118
$texture_g$	0.07335
$lnLDeV_g$	0.06864
$petroMag_{gr}$	0.06626
$isoAGrad_u * z$	0.05729

Results of cluster separation analysis

Best Separation in Single Dimension	Best Separation Among 1014 Combinations
$isoAGrad_u * z$	$isoAGrad_u * z$
$petroRad_u * z$	$petroRad_u * z$
$texture_u$	$texture_u$
$isoA_z * z$	$isoA_z * z$
$lnLExp_u$	$lnLExp_u$
$lnLExp_g$	$lnLExp_g$
$isoA_u * z$	$petroRad_u * z$, $isoB_z * z$, $isoBGrad_u * z$, $lnLExp_g$
$isoB_z * z$	$isoAGrad_u * z$, $lnLExp_g$
$isoBGrad_u * z$	$petroRad_u * z$, $isoA_u * z$, $isoB_z * z$, $lnLExp_g$
$isoAGrad_z * z$	$isoAGrad_u * z$, $isoBGrad_u * z$, $lnLExp_g$

Summary Remarks

- These are just the first steps
- There is more to come
- We will seek involvement from Machine Learning community
- We will advise LSST project based on our results
- We anticipate using (requiring) lightweight semantics (folksonomies, taxonomies, …)
- Collecting annotations and tags in structured form (e.g., RDBMS) is essential for ease of retrieval and semantic data mining
- … and finally … my recommendation …

...

Party with your data !