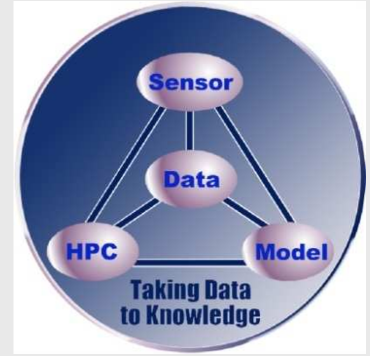


Conquering the Astronomical Data Flood through Machine Learning and Citizen Science

Kirk Borne

George Mason University

School of Physics, Astronomy, & Computational Sciences

http://spacs.gmu.edu/

The Problem: Big Data is a Big Challenge

LSST big data challenge #1

- **Each night** for 10 years LSST will obtain roughly the equivalent amount of data that was obtained by the entire Sloan Digital Sky Survey
- Our grad students will be asked to mine these data (~20 TB each night ≈ 40,000 CDs filled with data):
  - *A truckload of CDs each and every day for 10 yrs*
  - *Cumulatively, a football stadium full of 100 million CDs after 10 yrs*

LSST big data challenge #1

- **Each night** for 10 years LSST will obtain roughly the equivalent amount of data that was obtained by the entire Sloan Digital Sky Survey
- Our grad students will be asked to mine these data (~20 TB each night ≈ 40,000 CDs filled with data):
  - *A truckload of CDs each and every day for 10 yrs*
  - *Cumulatively, a football stadium full of 100 million CDs after 10 yrs*
- The challenge is to find the new, the novel, the interesting, and the surprises (the unknown unknowns) within all of these data.
- *Yes, more is most definitely different !*

LSST big data challenge # 2

- Approximately 2,000,000 times each night for 10 years LSST will detect a new sky event, and the astronomical community will be challenged with classifying these events. What will we do with all of these events?

LSST big data challenge # 2

- Approximately 2,000,000 times each night for 10 years LSST will detect a new sky event, and the astronomical community will be challenged with classifying these events. What will we do with all of these events?

Characterize first !
(Unsupervised Learning)

Classify later.

Characterization includes …

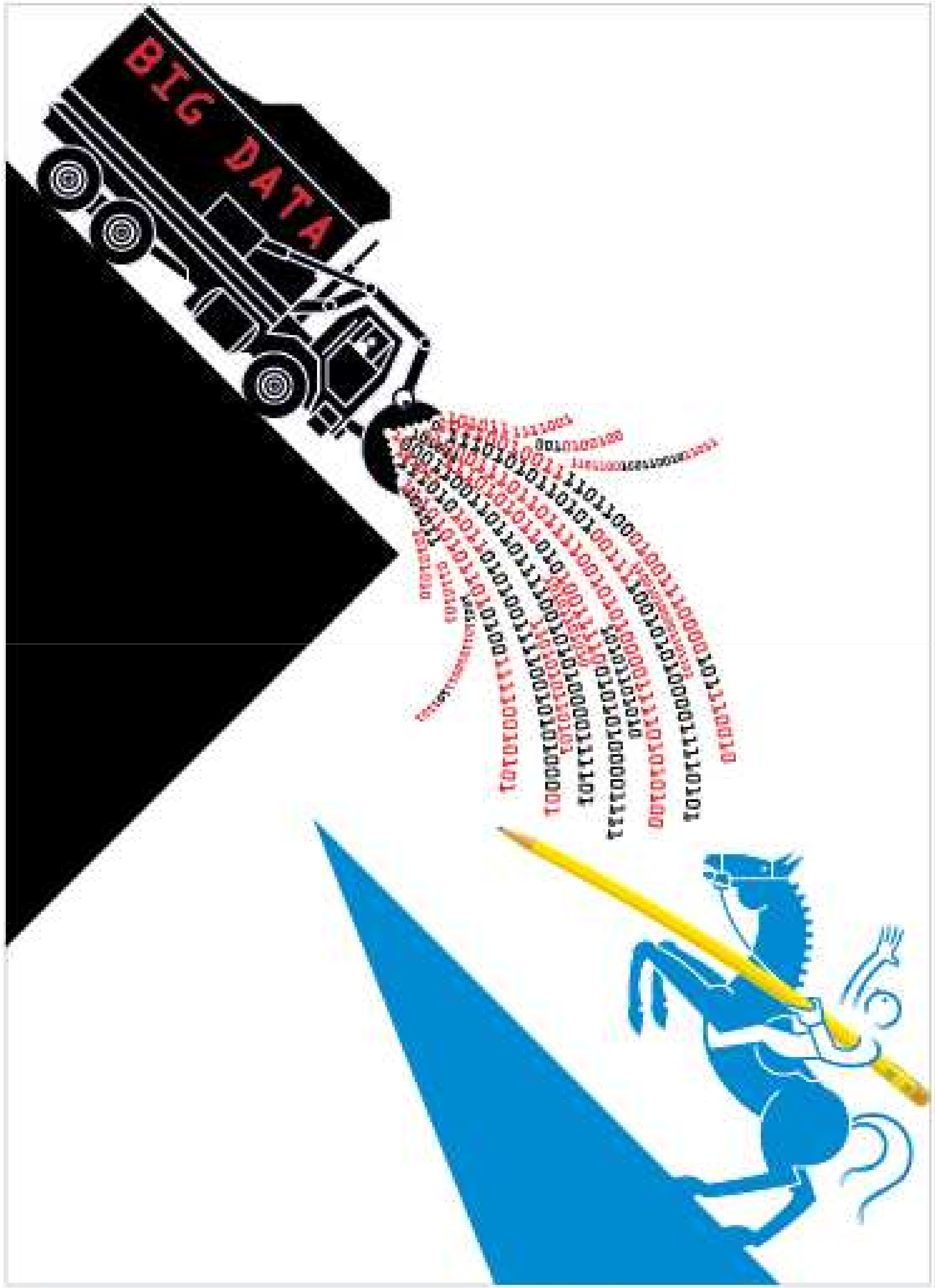
- Feature Detection and Extraction:
  - Identifying and describing features in the data
    - via machine algorithms or human inspection (including the potentially huge contributions from Citizen Science)
  - Extracting feature descriptors from the data
  - Curating these features for search, re-use, & discovery
  - Finding other parameters and features from other archives, other databases, other information sources – and using those to help characterize (ultimately classify) each new event.

Data-driven Discovery (Unsupervised Learning)
i.e., What can I do with characterizations?

1. Class Discovery – Clustering
2. Principal Component Analysis – Dimension Reduction
3. Outlier (Anomaly / Deviation / Novelty) Detection – Surprise Discovery
4. Link Analysis – Association Analysis – Network Analysis
5. and more.

The Promise: Big Data leads to Big Insights and New Discoveries

http://kdd2012.sigkdd.org/

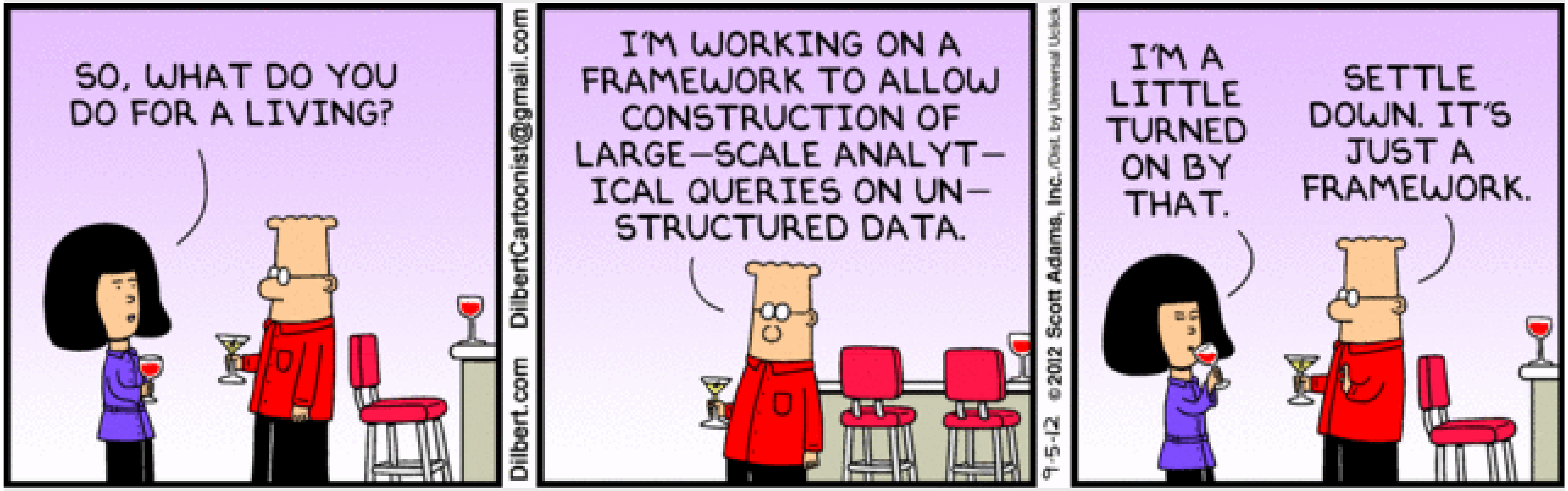


Scary News:

Big Data is taking us to a Tipping Point

http://bit.ly/HUqmu5

Good News: Big Data is Sexy

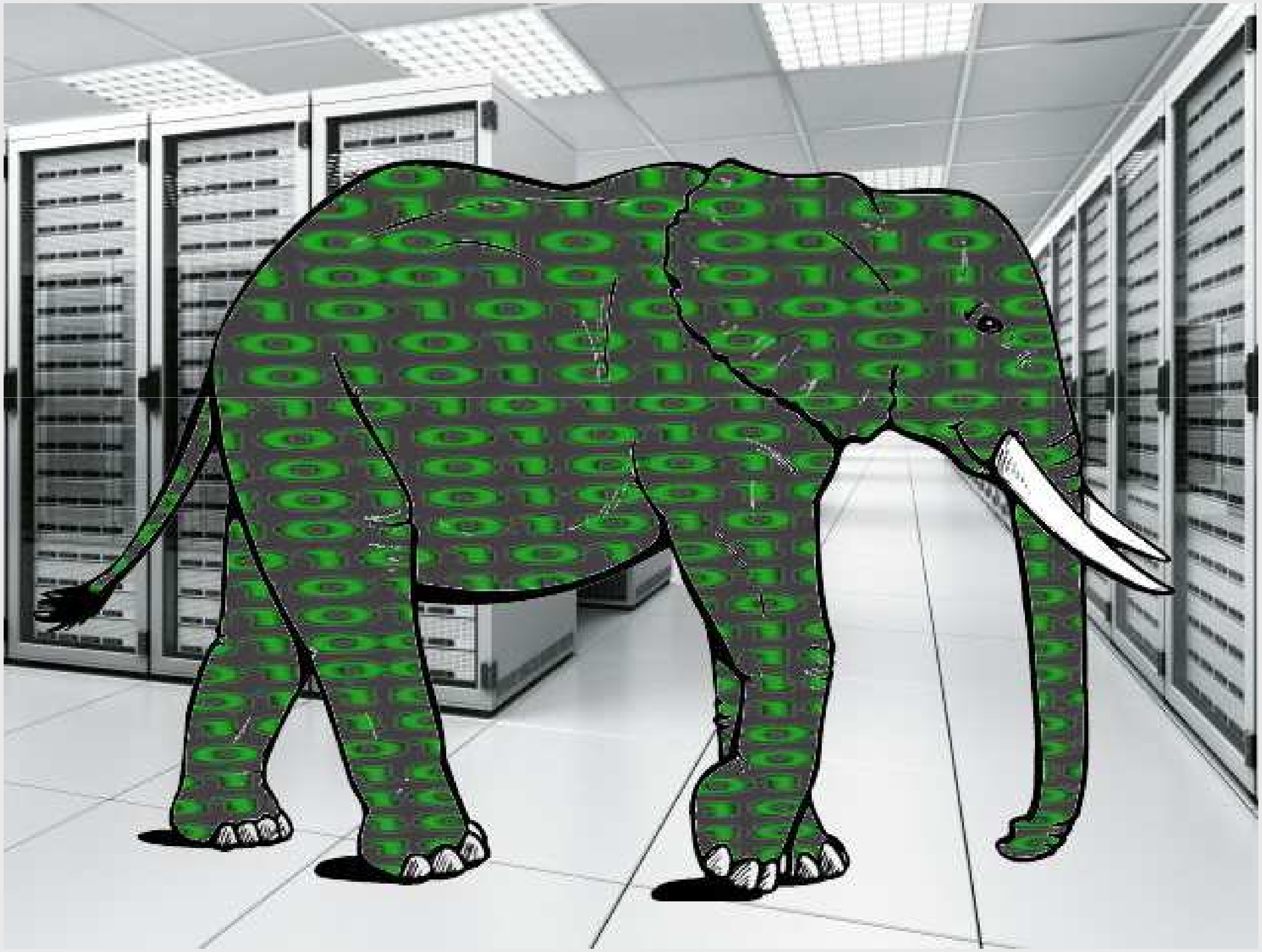


http://dilbert.com/strips/comic/2012-09-05/

There are many technologies associated with Big Data

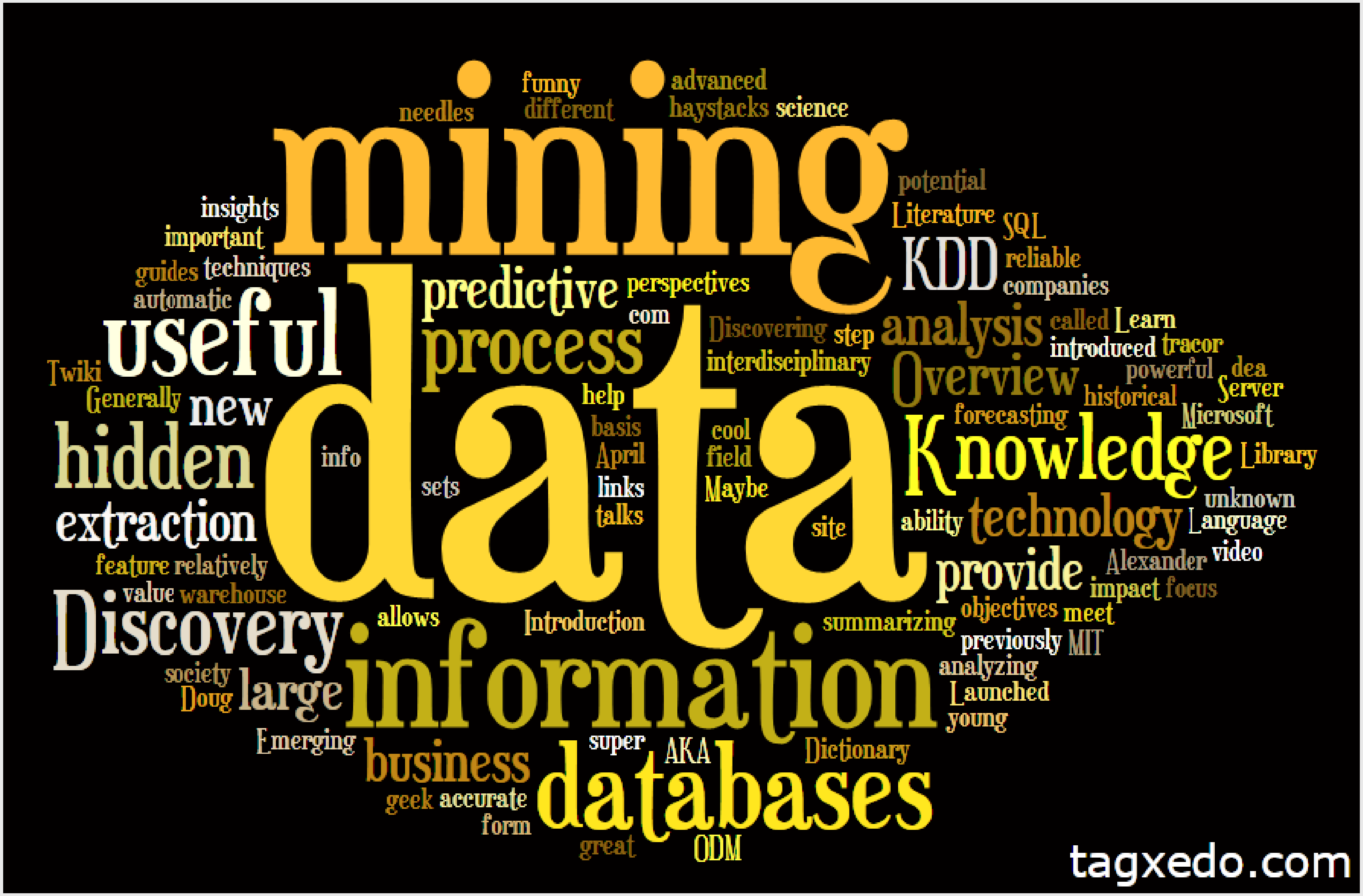
http://siliconangle.com/blog/2012/07/13/big-data-nightmares/

One approach to Big Data: Hadoop and Map/Reduce (Computational Science)

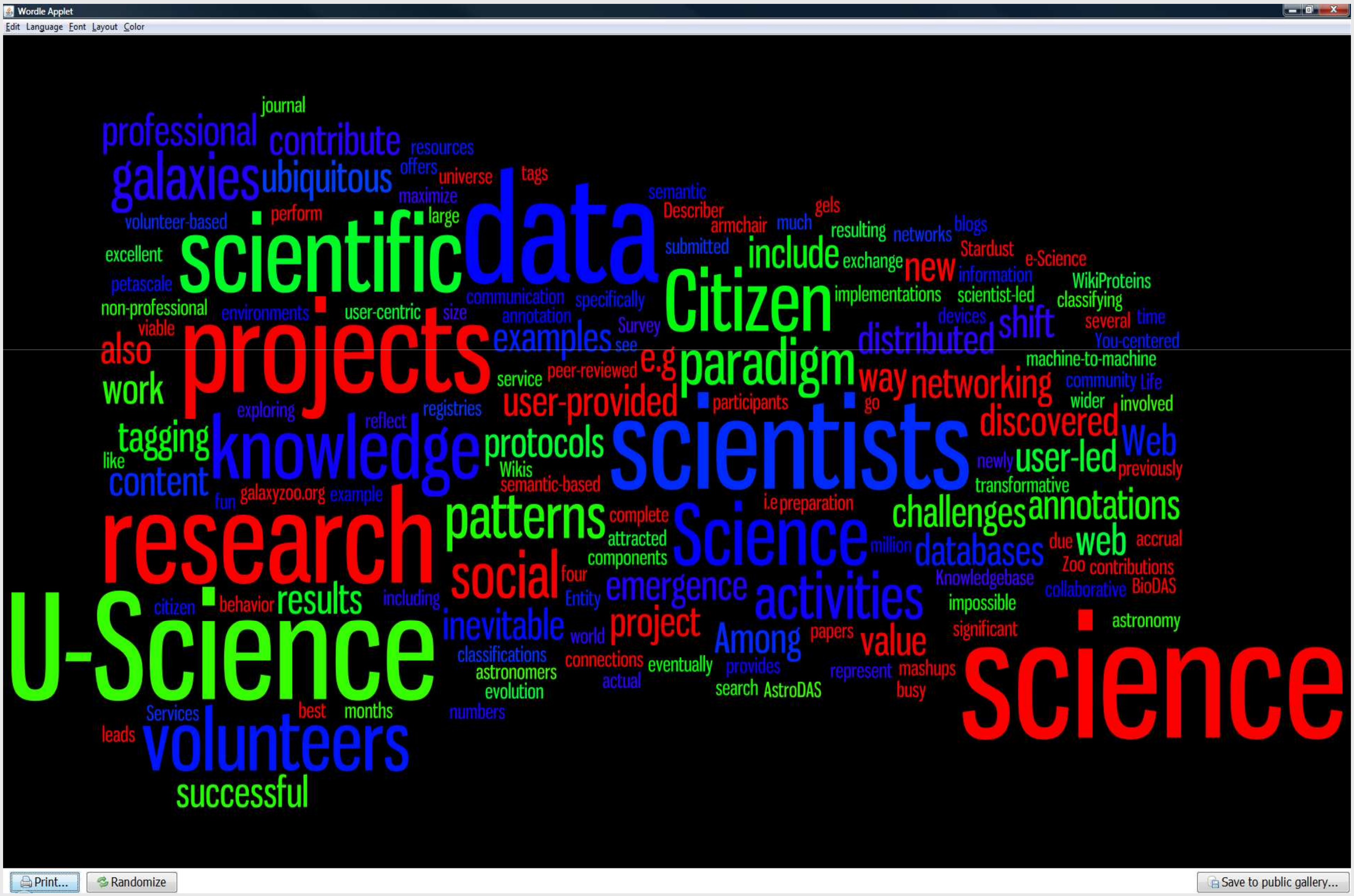


http://www.bigdatabytes.com/wp-content/uploads/2012/01/big-data.jpg

Another approach to Big Data:
Data Science (Informatics)



A third approach to Big Data:
Citizen Science (crowdsourcing)



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

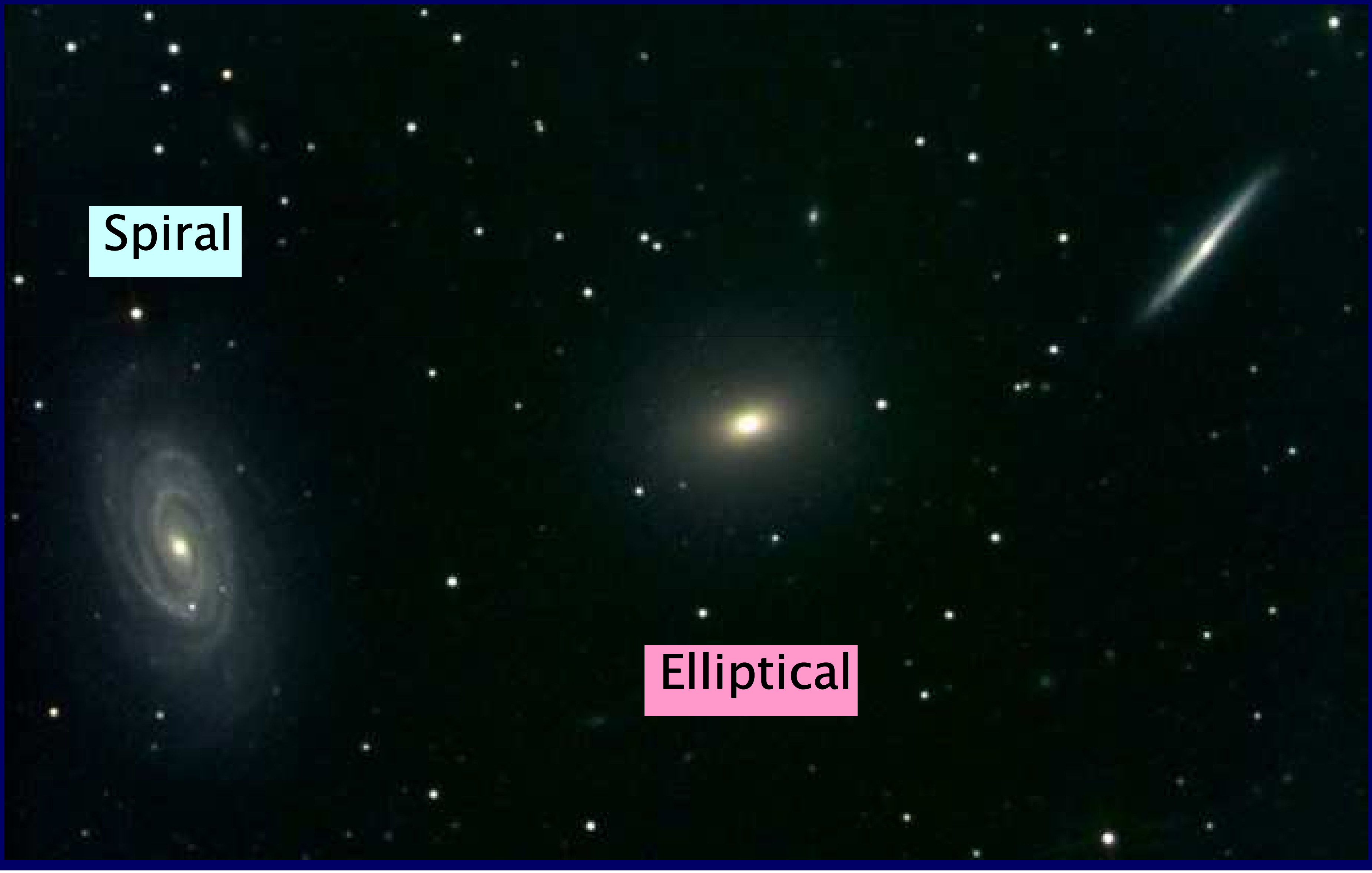
Galaxy Zoo: example of Citizen Science (crowdsourcing)

http://astrophysics.gsfc.nasa.gov/outreach/podcast/wordpress/index.php/2010/10/08/saras-blog-be-a-scientist/

Galaxy Zoo: example of Citizen Science (crowdsourcing)

http://astrophysics.gsfc.nasa.gov/outreach/podcast/wordpress/index.php/2010/10/08/saras-blog-be-a-scientist/

There are 2 main types of galaxies in the Universe: Spiral & Elliptical (plus there are some peculiar & irregular galaxies)

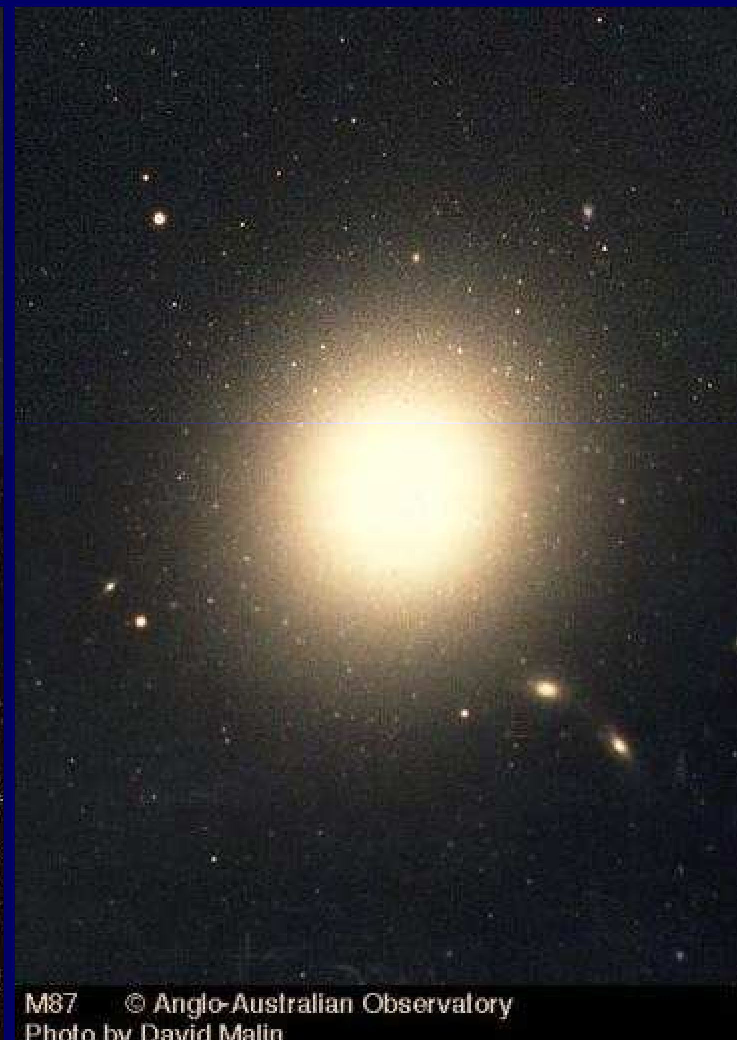


Gallery of Elliptical Galaxies

M32

M59

M87

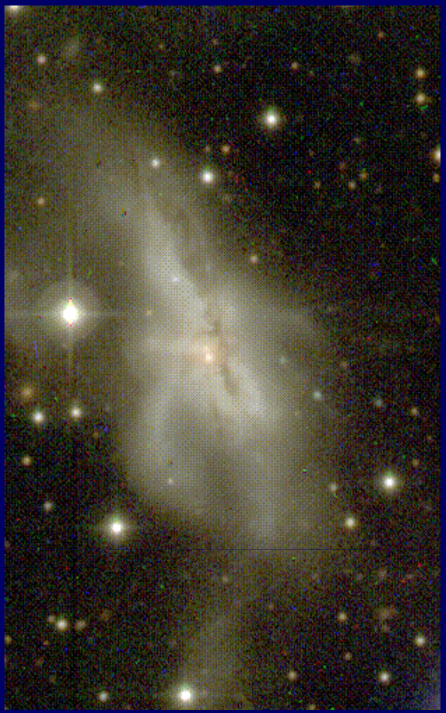


M105

M110=NGC205

Gallery of Face-on Spiral Galaxies

There are lots of Peculiar Galaxies also !



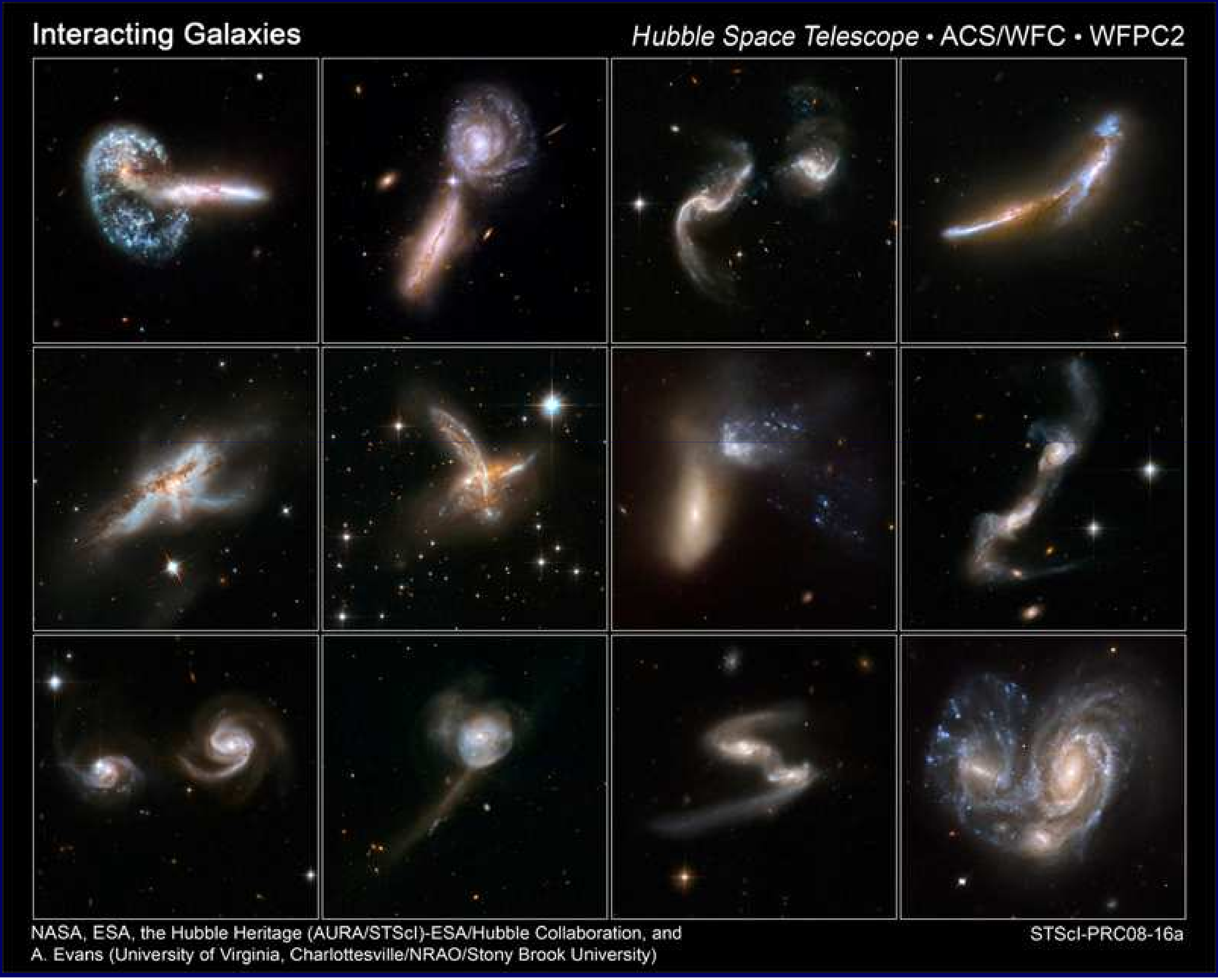
There are lots of things you can do with these peculiar galaxies …

- Spell your name in galaxies @
- http://writing.galaxyzoo.org

k i r k   b o r n e

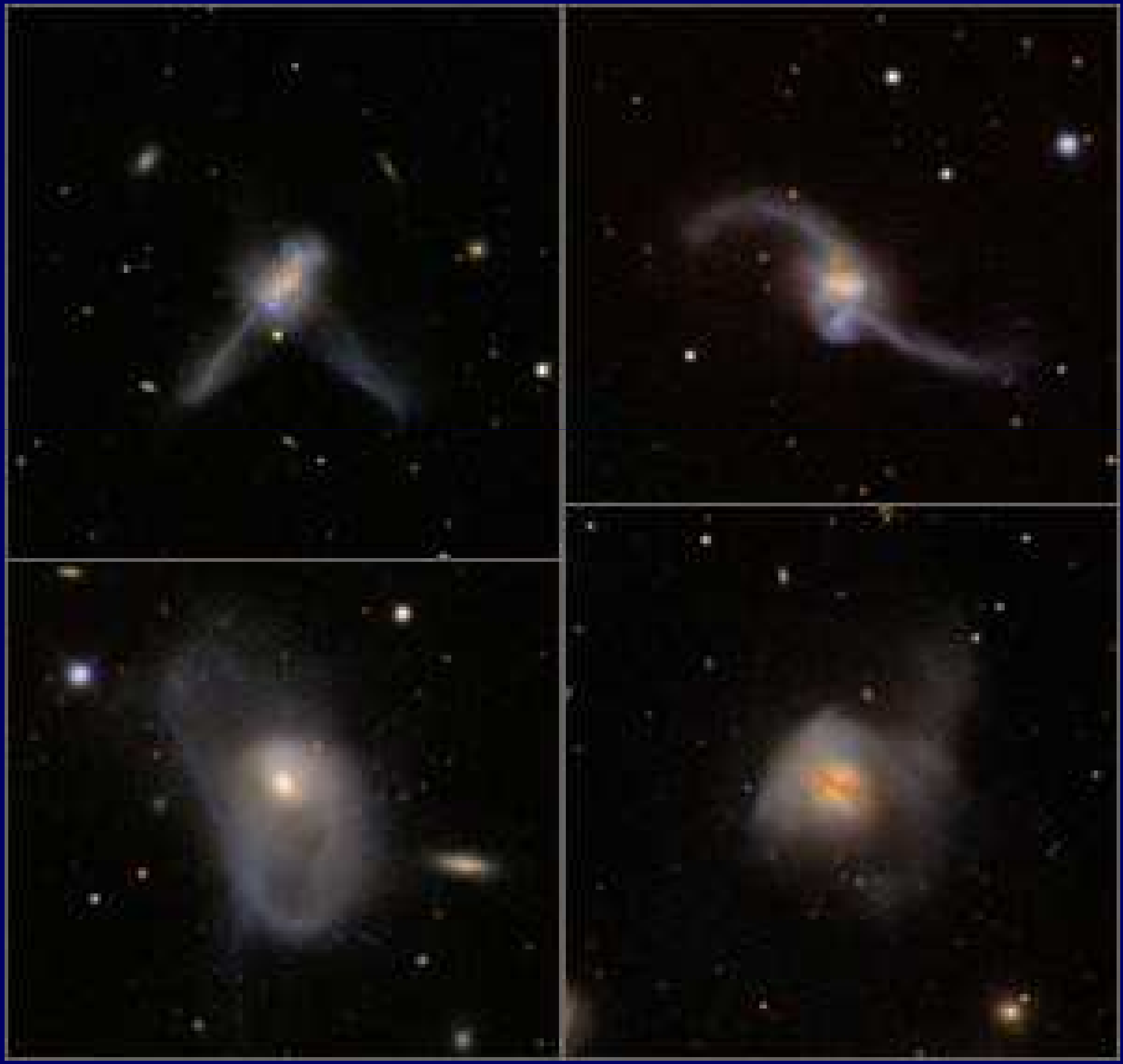
Galaxies Gone Wild !

Colliding and Merging Galaxies = Interacting Galaxies

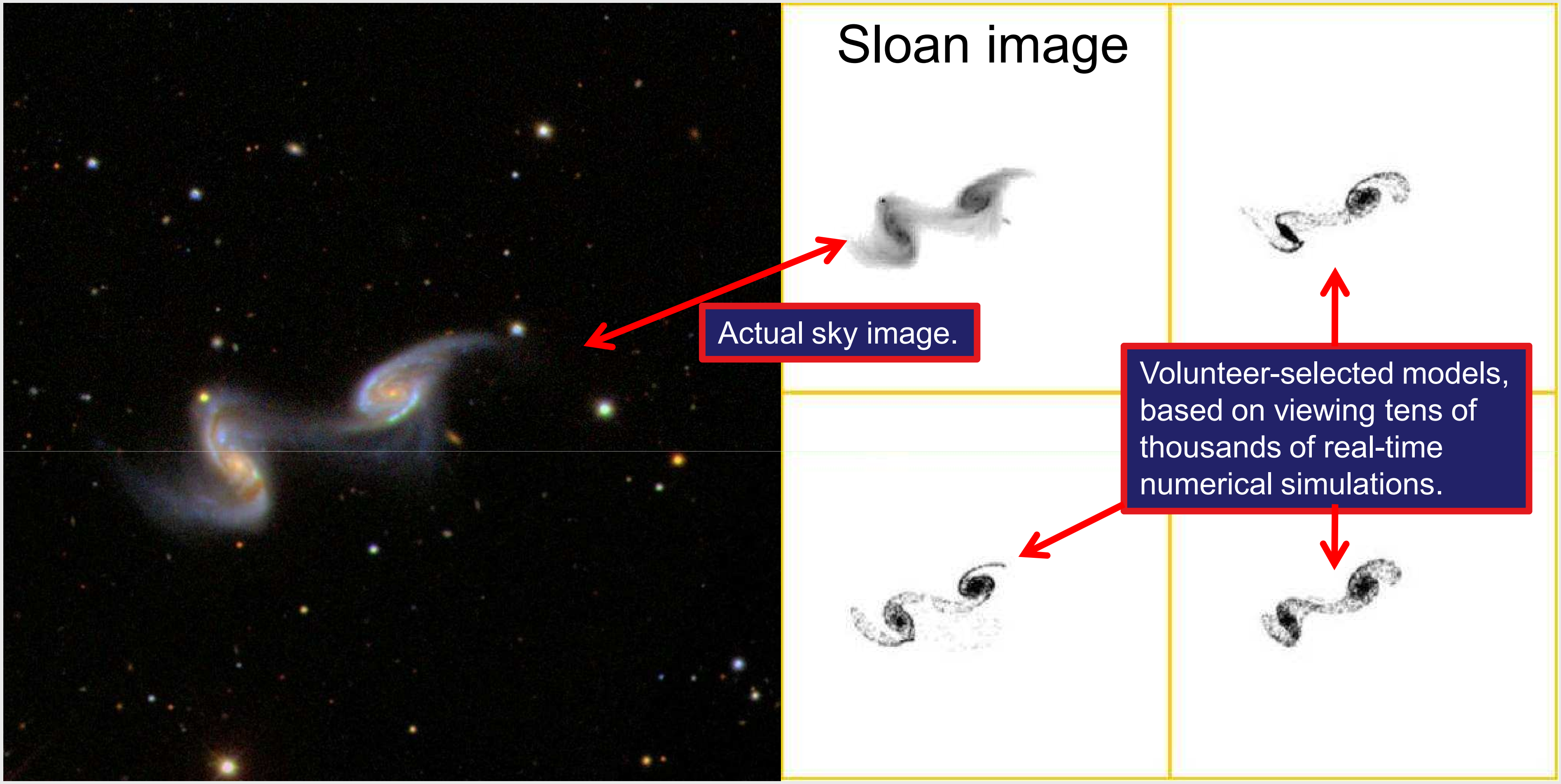


Colliding and Merging Galaxies = Interacting Galaxies

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



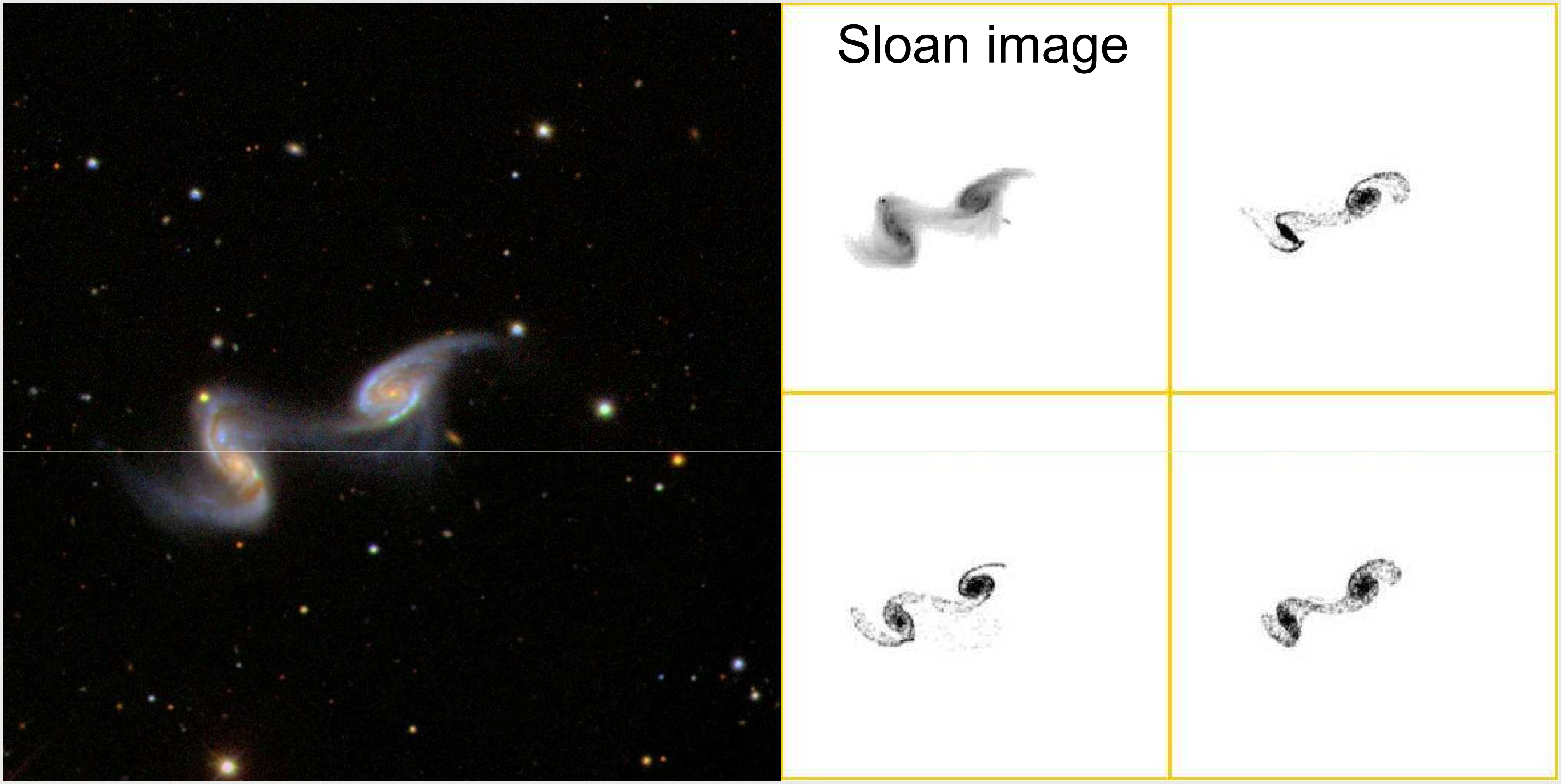
Galaxy Mergers Zoo Gallery



SDSS 587722984435351614

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

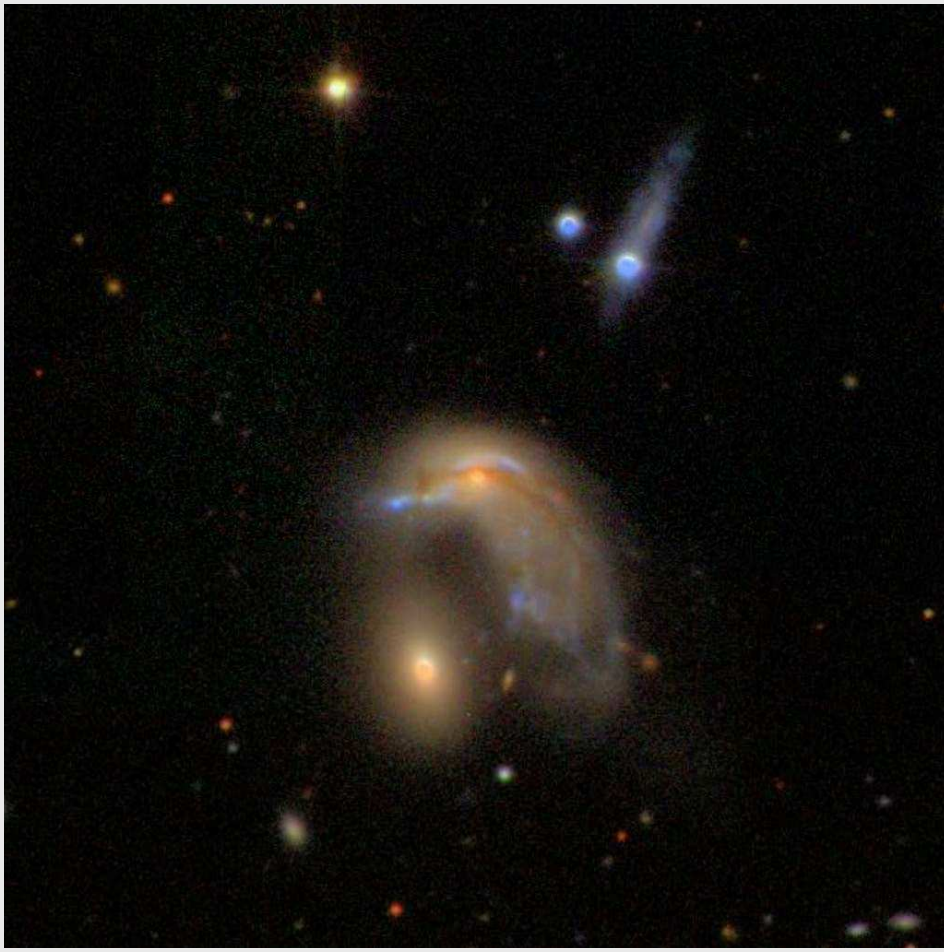
Galaxy Mergers Zoo Gallery

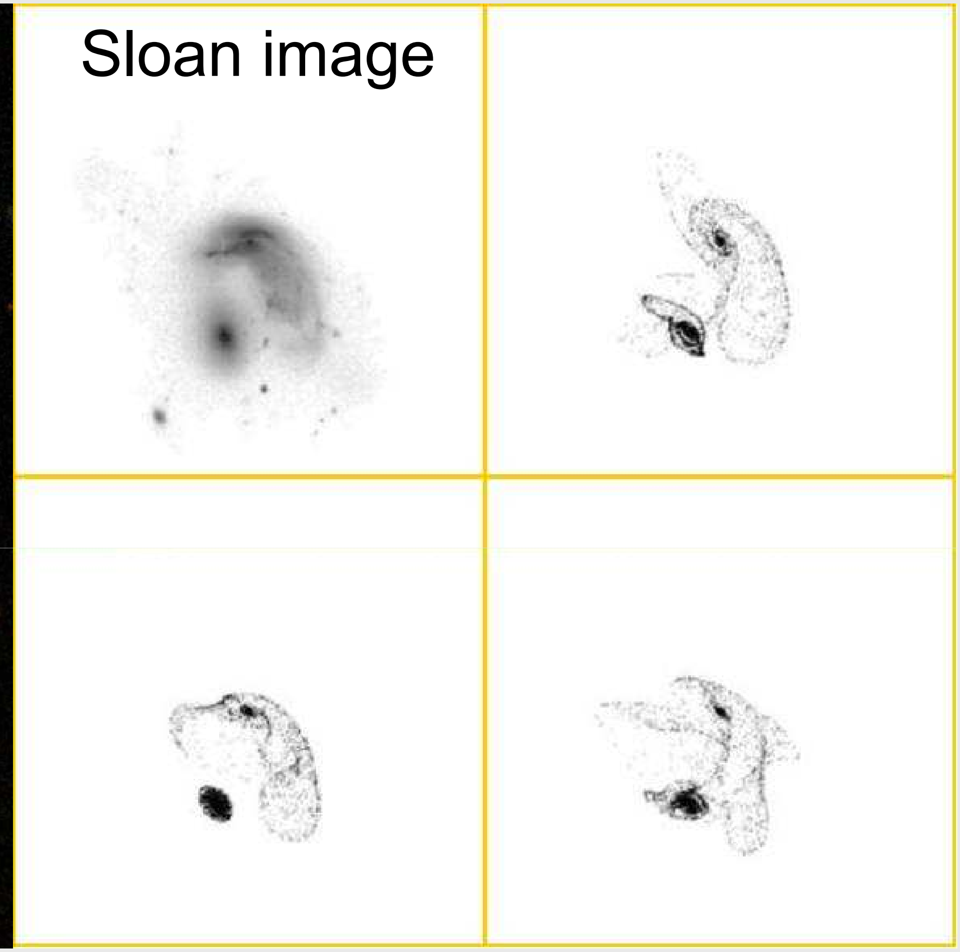


SDSS 587722984435351614

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery

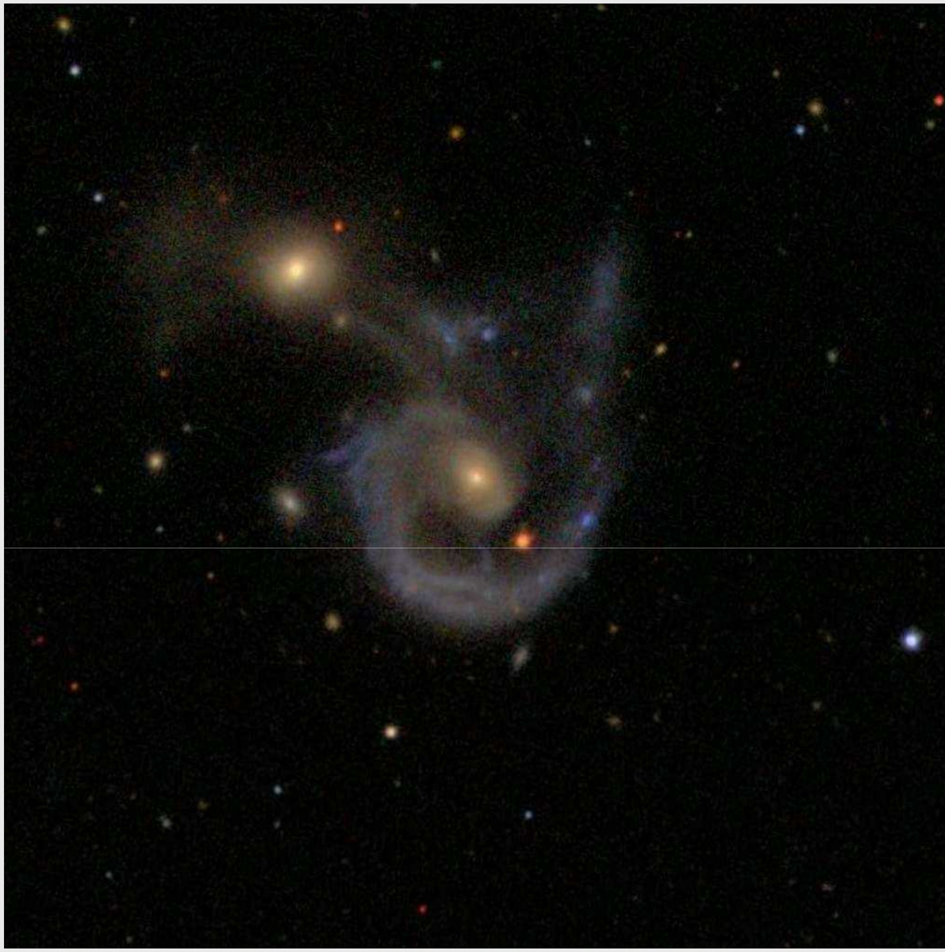


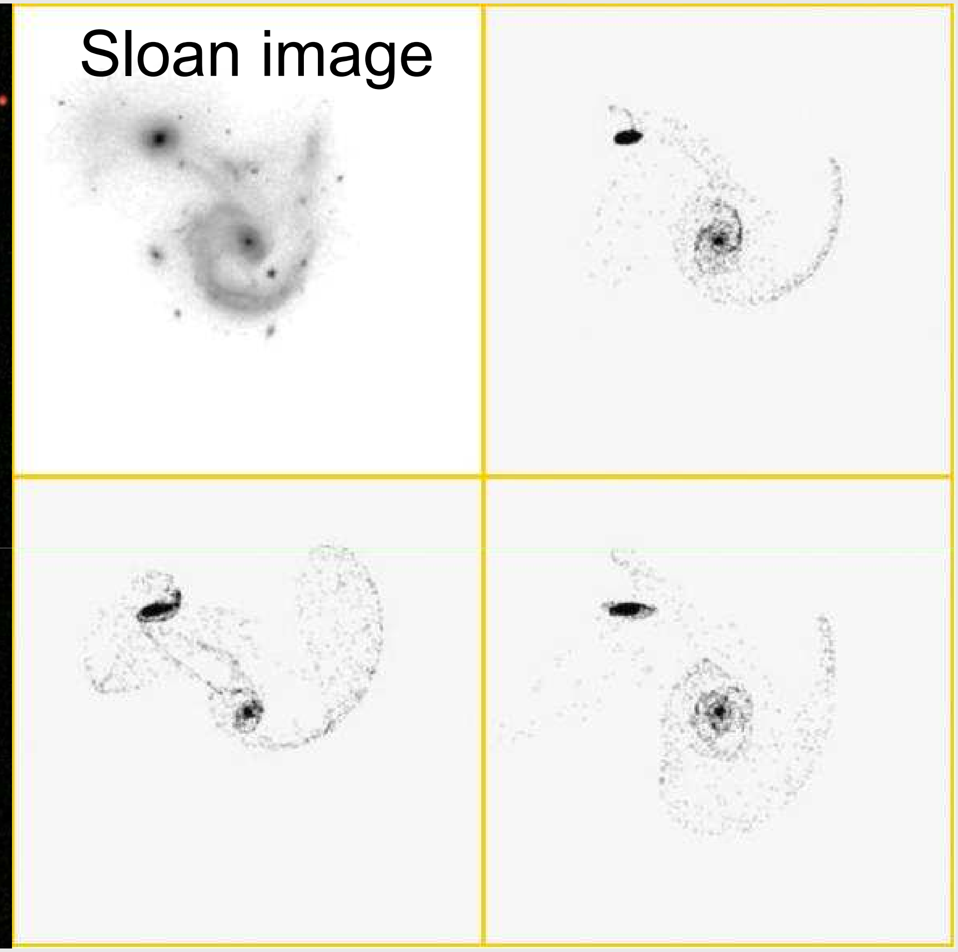


SDSS 587726033843585146

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery

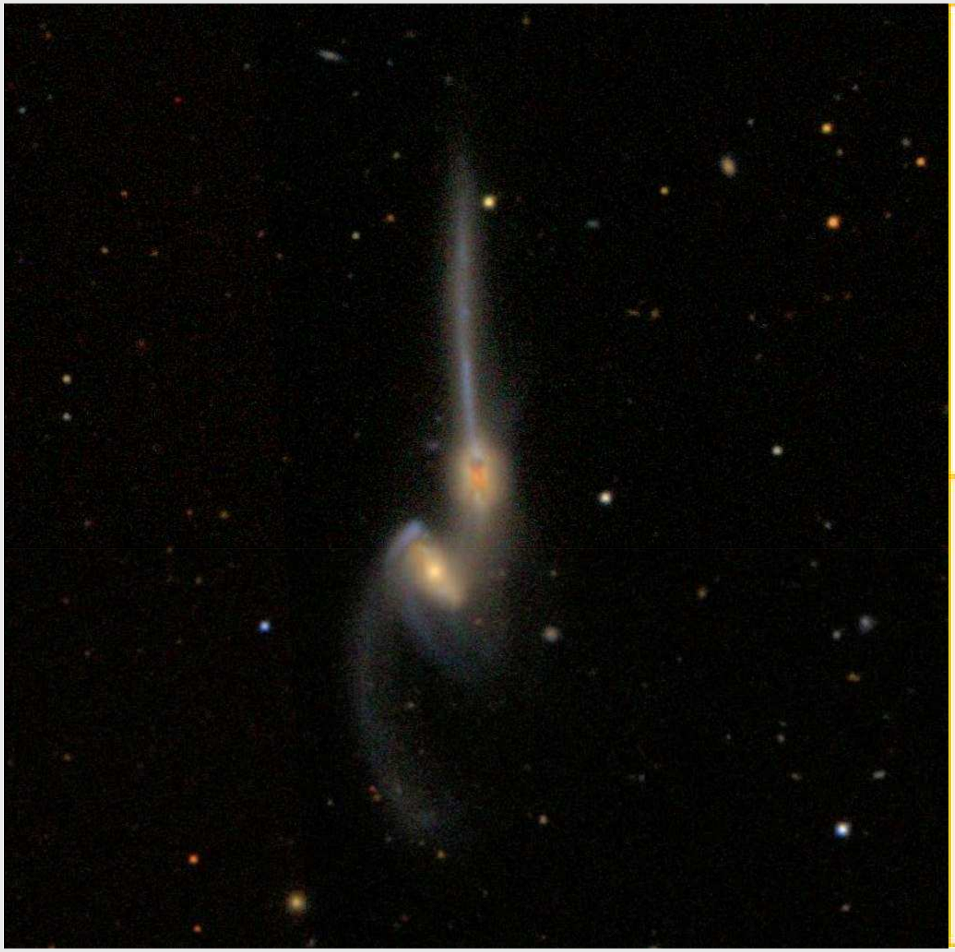


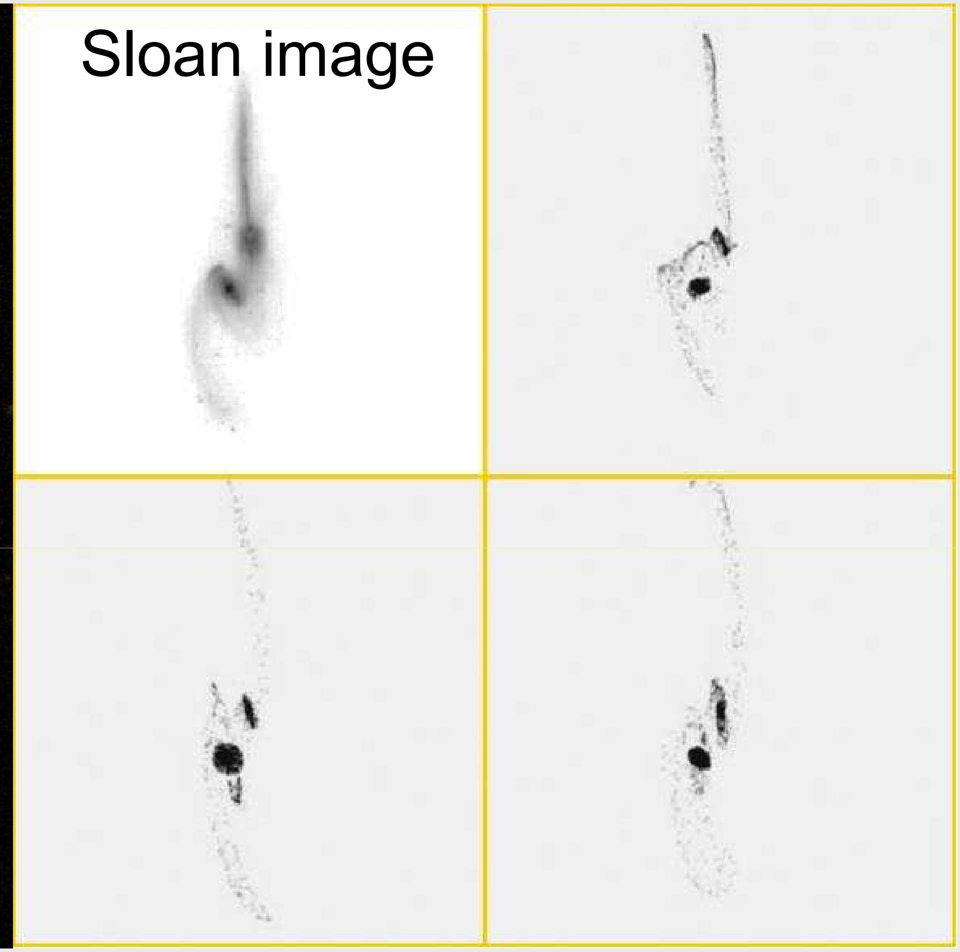


SDSS 587739646743412797

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery

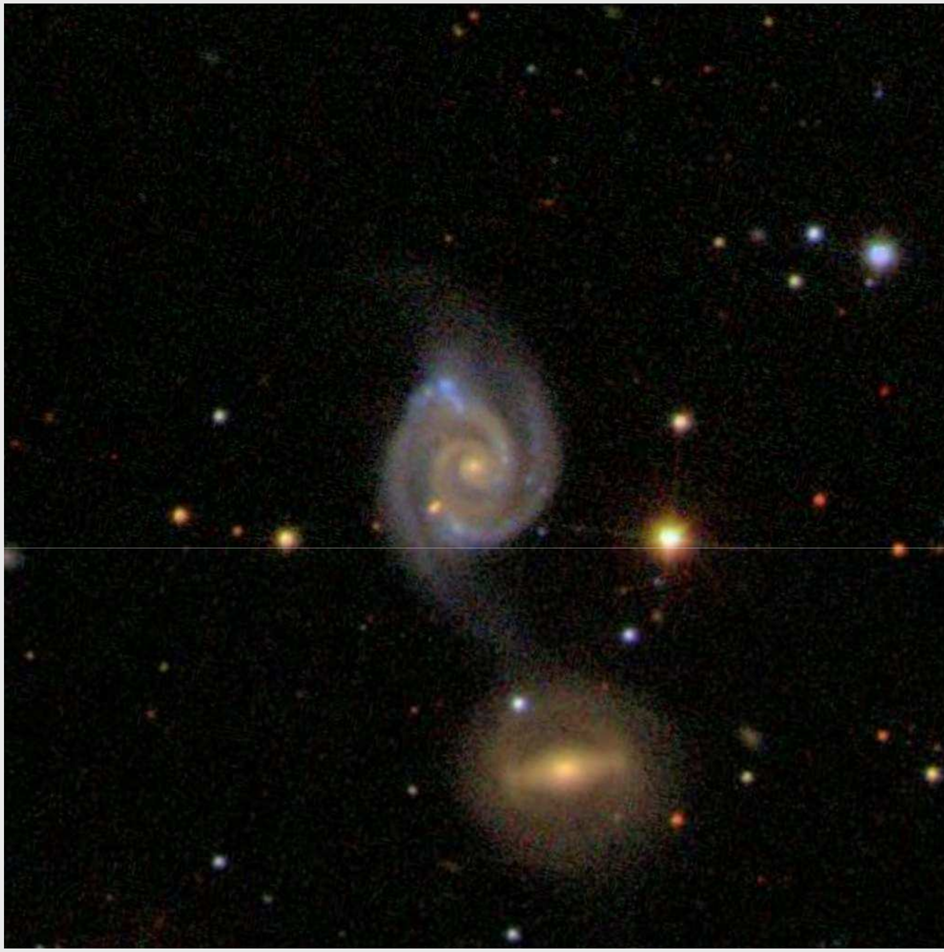


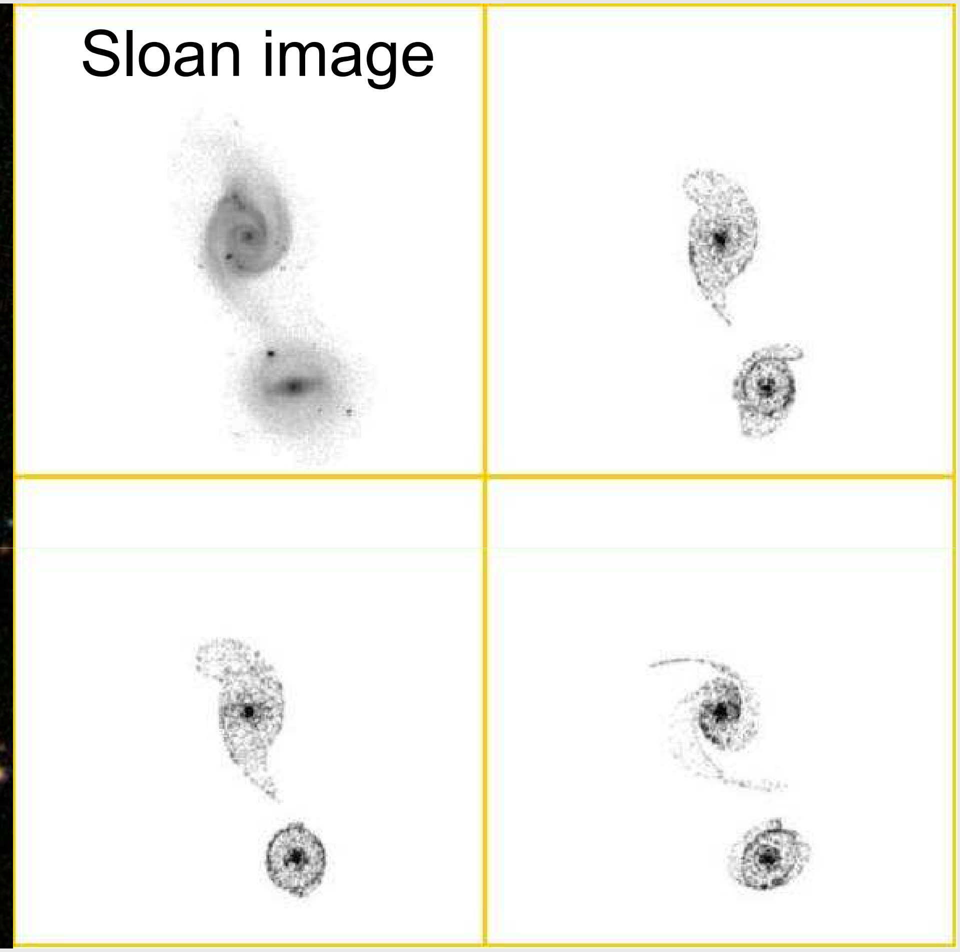


SDSS 587739721900163101

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery

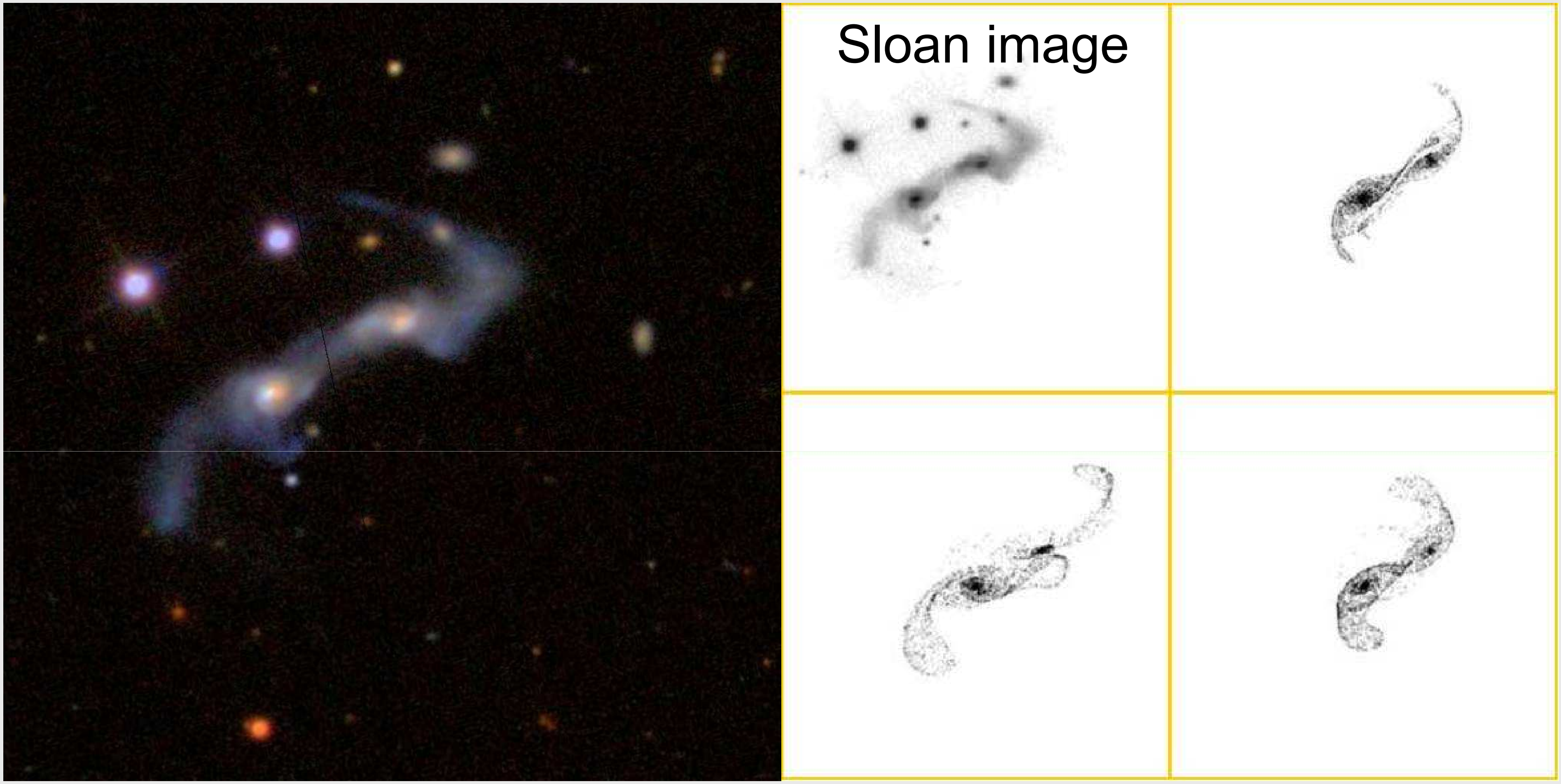




SDSS 587727222471131318

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery



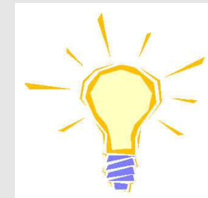
SDSS 588011124116422756

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 (V&V)
- Hundreds of millions of class labels
- Statistics deluxe! …
  - Users (see paper: http://arxiv.org/abs/0909.2925)
  - Uncertainty Quantification (UQ)
  - Classification certainty vs. Classification dispersion

ZOONIVERSE
REAL SCIENCE ONLINE

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)
  - with experimental parameters
  - with observer / observatory descriptors

Semantics!

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.

Responding to Big Data in Science

- ***X-Informatics*** (*e.g.,* X = Bio, Geo, Astro, …):
  - addresses the scientific data lifecycle challenges in the era of Big Data and data-intensive science …
  - via data science techniques for indexing, accessing, searching, fusing, integrating, mining, and analyzing massive data repositories.
  - Includes automatic (autonomous) tagging and annotation

- ***Citizen Science*** (user-guided, informatics-powered):
  - Human computation (*e.g.,* tagging, labeling, classification)
    - characterized by enormous cognitive capacity and pattern recognition efficiency (carbon-based computing)
  - Semantic e-Science **and** Volunteer Citizen Science
  - Tagging everything, everywhere: ***Analytics in the Cloud***

Astroinformatics and Citizen Science: Resistance is futile, you will be assimilated