

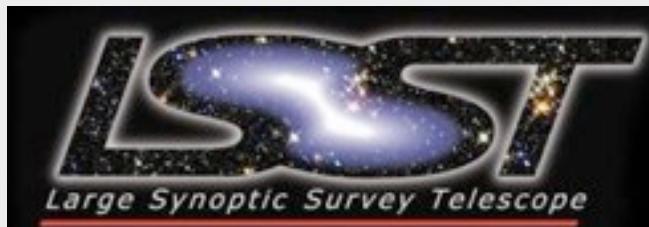


Reaching Out with Eventful Astronomy

Kirk Borne

George Mason University

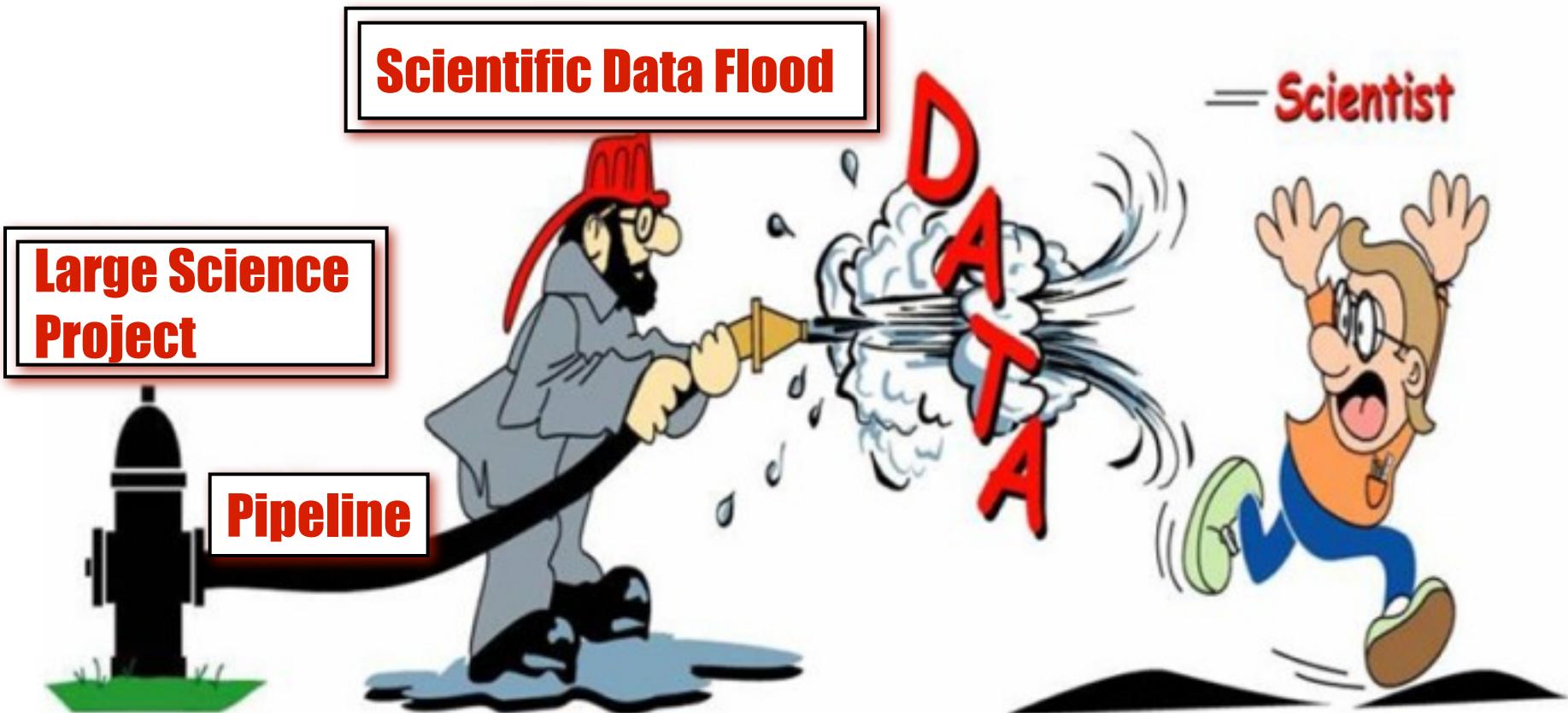




**The LSST will represent a
10K-100K times increase in
the VOEvent network traffic.
This poses significant
real-time classification
demands on the event stream:
from data to knowledge!
from sensors to sense!**

The Scientific Data Flood

Drinking from a FIREHOSE



How will we respond ?



We need something better ...

We need something better, Jim !



We need computers ...
but not the usual kind !



We need the classical kind
(which pre-dates computing
devices)

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

It takes a human to interpret a complex image



**It takes a human to interpret a complex image
... usually ...**



"It's black, and it looks like a hole.
I'd say it's a black hole."

Citizen Science

- Exploits the cognitive abilities of **Human Computation!**
- Novel mode of data collection:
 - Citizen Science! = Volunteer Science = Participatory Science
 - e.g., VGI = Volunteer Geographic Information (Goodchild '07)
 - e.g., Galaxy Zoo @ <http://www.galaxyzoo.org/>
- Citizen science refers to the involvement of volunteer non-professionals in the research enterprise.
- The Citizen Science experience ...
 - must be engaging,
 - must work with real scientific data/information,
 - must not be busy-work (all clicks must count),
 - **must address authentic science research questions** that are beyond the capacity of science teams and enterprises, and
 - must involve the scientists.

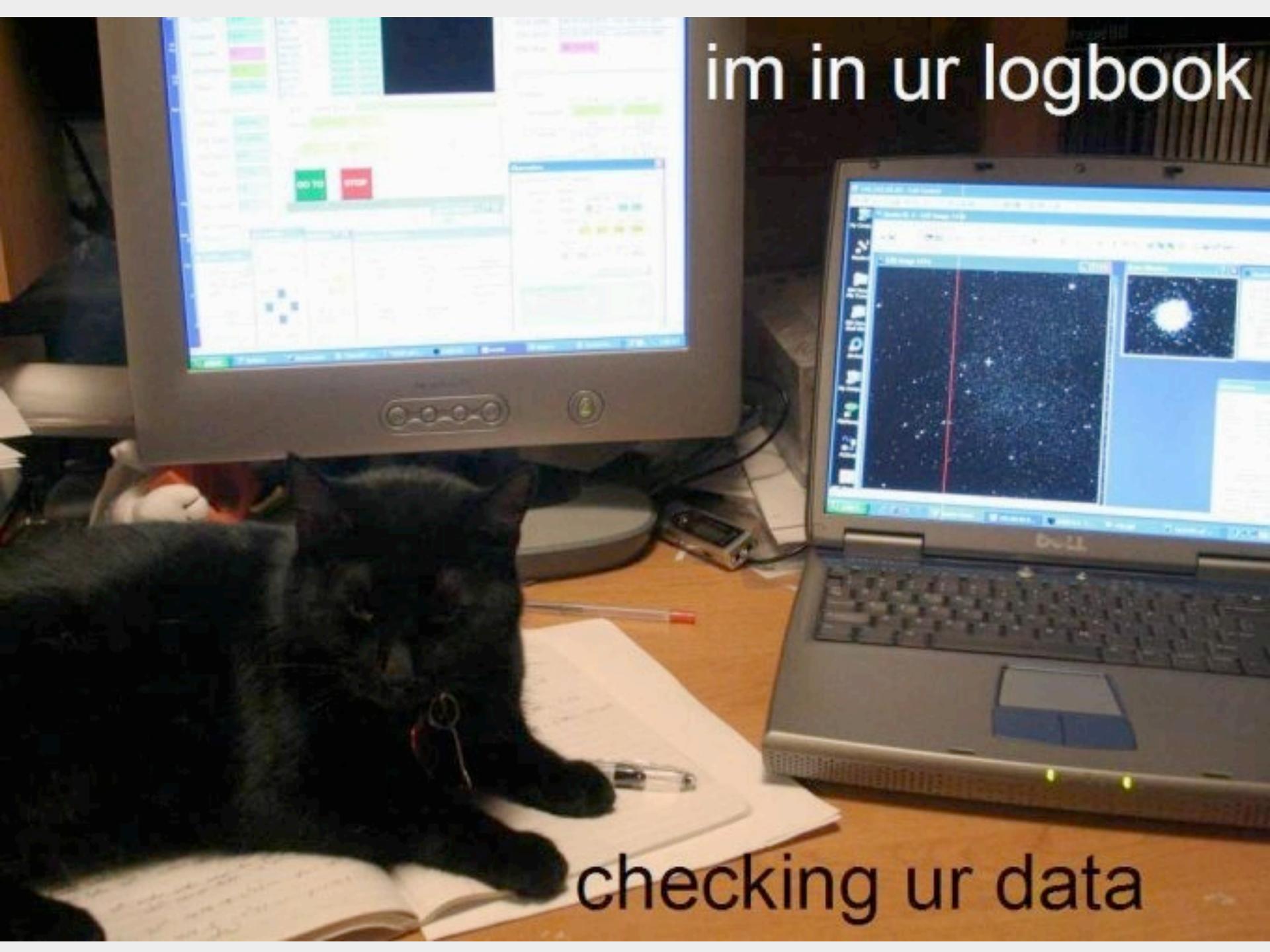
Examples of Volunteer Science

- AAVSO (Amer. Assoc. of Variable Star Observers)
- Audubon Bird Counts
- Project Budburst
- Stardust@Home
- VGI (Volunteer Geographic Information)
- CoCoRaHS (Community Collaborative Rain, Hail and Snow network)
- Galaxy Zoo (**~20 refereed pubs so far...**)
- Zooniverse (buffet of Zoos)
- U-Science (semantic science 2.0) [ref: Borne 2009]
 - includes Biodas.org, Wikiproteins, HPKB, AstroDAS
 - **Ubiquitous, User-oriented, User-led, Universal, Untethered, You-centric Science**

Anybody can participate and contribute to the science...



"On the Internet, nobody knows you're a dog"

A black cat is sitting on a wooden desk, facing a computer setup. On the left is a large monitor displaying a complex software interface with multiple windows, including a map and various data tables. To the right is a silver Dell laptop showing a star map with a red vertical line. A small white notepad with a pen lies on the desk in front of the cat. The background shows a dark room.

im in ur logbook

checking ur data

Galaxy Zoo helps scientists by engaging the public (hundreds of thousands of us) to classify millions of galaxies:



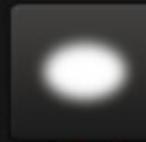
- Galaxy Zoo project:
 - ~260,000 participants (*and growing*)
 - ~1 million galaxies have been labeled (classified)
 - ~180 million classifications have been collected

[INVERT GALAXY IMAGE](#)[ADD TO MY FAVOURITES](#)

Classify Galaxies

Answer the question below using the buttons provided.

Is the galaxy simply smooth and rounded, with no sign of a disk?



Smooth



Features or disk



Star or artifact

[Need help? ?](#)

[INVERT GALAXY IMAGE](#)[ADD TO MY FAVOURITES](#)

Classify Galaxies

Answer the question below using the buttons provided.

Is the galaxy simply smooth and rounded, with no sign of a disk?



Smooth



Features or disk



Star or artifact

[Need help? ?](#)

True color picture of Hanny's Voorwerp:
Hanny's Object – the green blob is probably a light echo
from an old Quasar that burned out 100,000 years ago



The Zooniverse* : Advancing Science through User-Guided Learning in Massive Data Streams



* NSF CDI funded program @ <http://zooniverse.org>

The Zooniverse

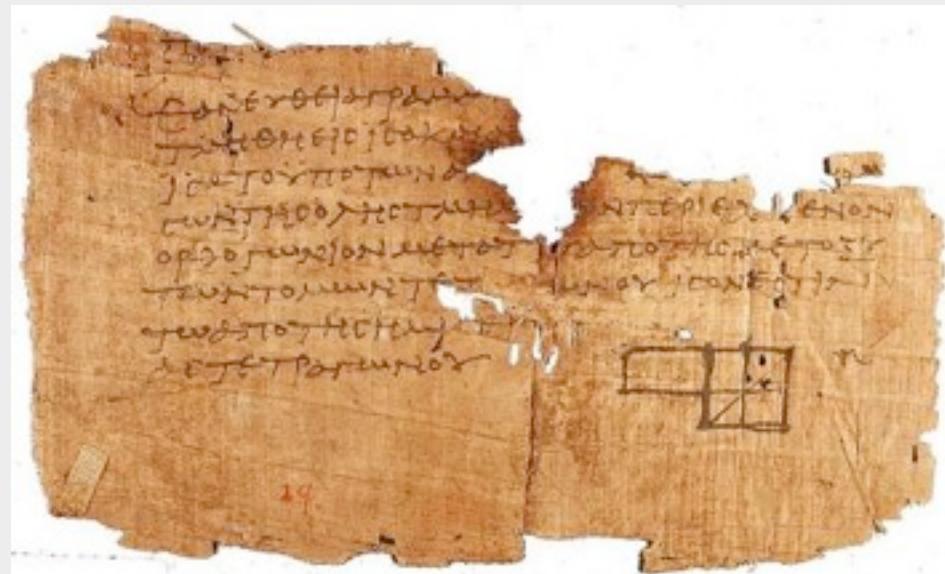
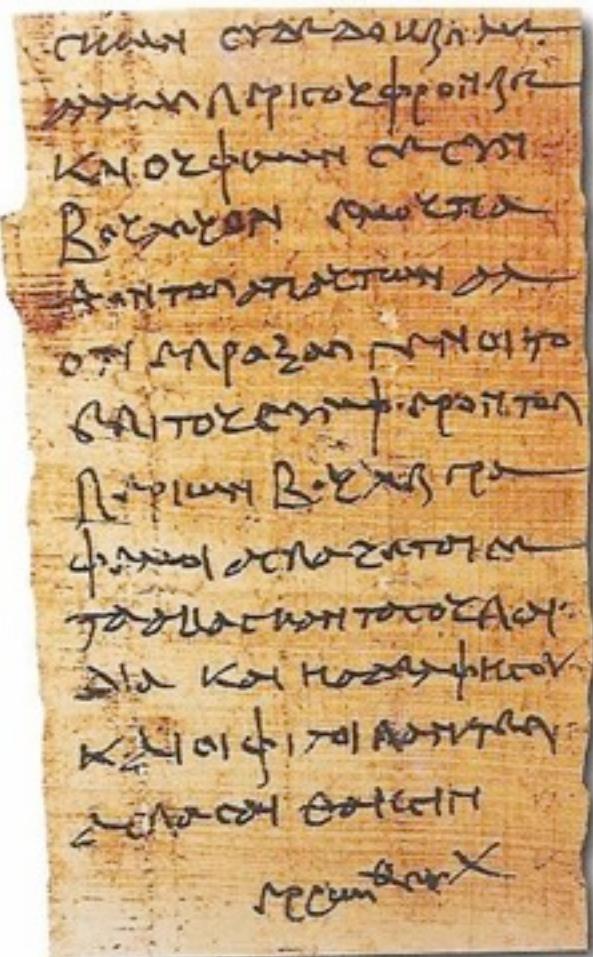
<http://zooniverse.org/>

- New funded NSF CDI grant (PI: L.Fortson, Adler Planetarium; co-PI J. Wallin & collaborator K.Borne, GMU; & collaborators at Oxford U)
- Building a framework for new Citizen Science projects, including user-based research tools
- Science domains:
 - Astronomy (Galaxy Merger Zoo)
 - The Moon (Lunar Reconnaissance Orbiter)
 - The Sun (STEREO dual spacecraft)
 - Egyptology (the Papyri Project)
 - and more (... accepting proposals from community)



Egyptology (the Papyri Project)

Oxyrhynchus Papyri Project @ <http://www.papyrology.ox.ac.uk/>



The Zooniverse: a Buffet of Zoos

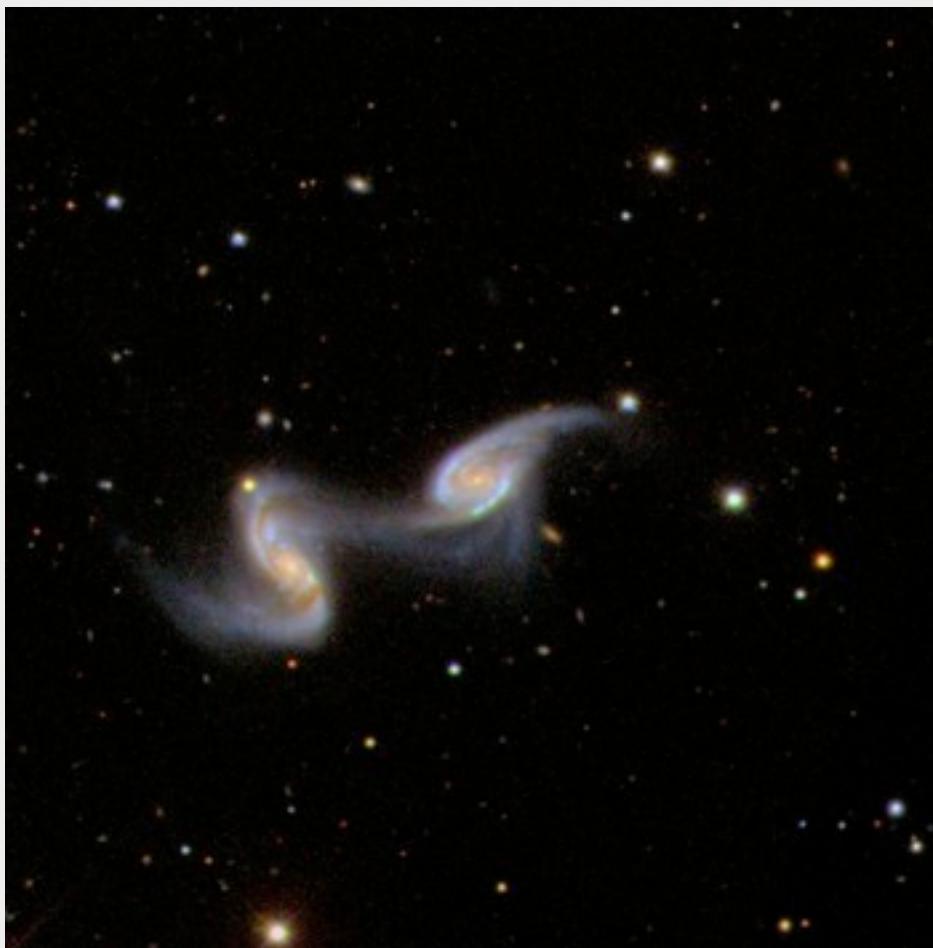
<http://zooniverse.org/>

- 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



Sloan image

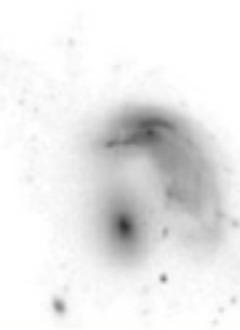


SDSS 587722984435351614

Galaxy Mergers Zoo Gallery

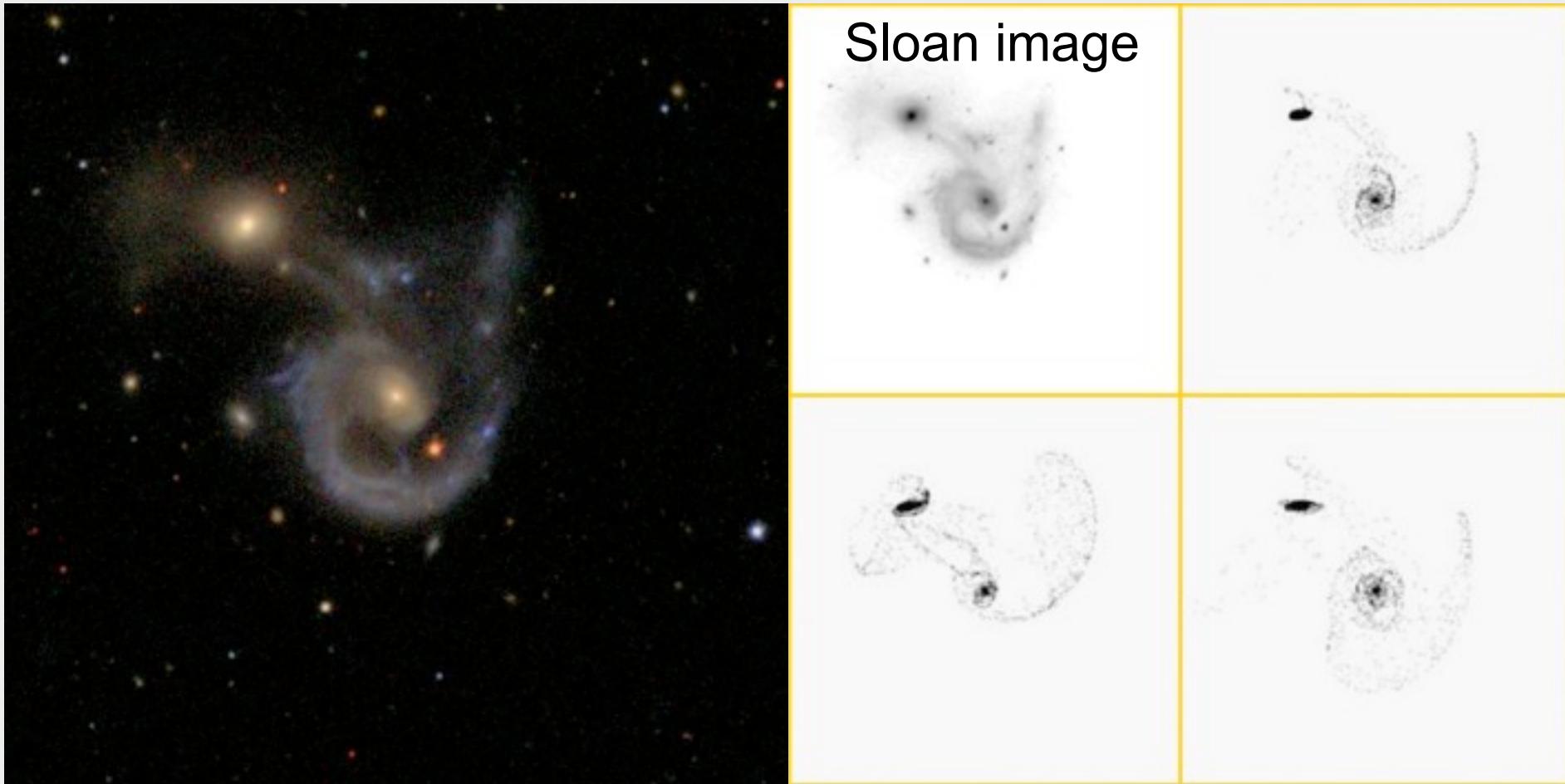


Sloan image



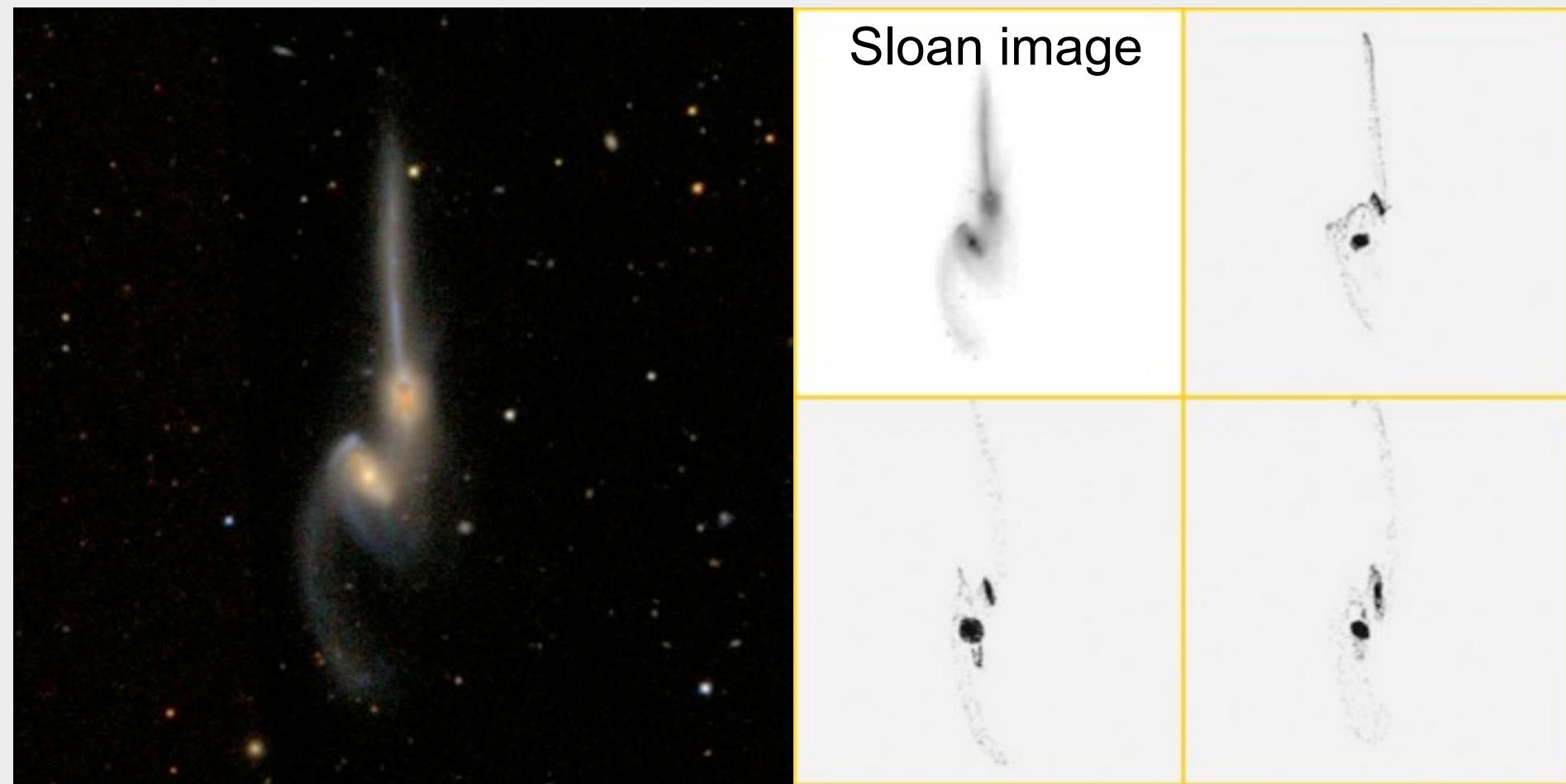
SDSS 587726033843585146

Galaxy Mergers Zoo Gallery



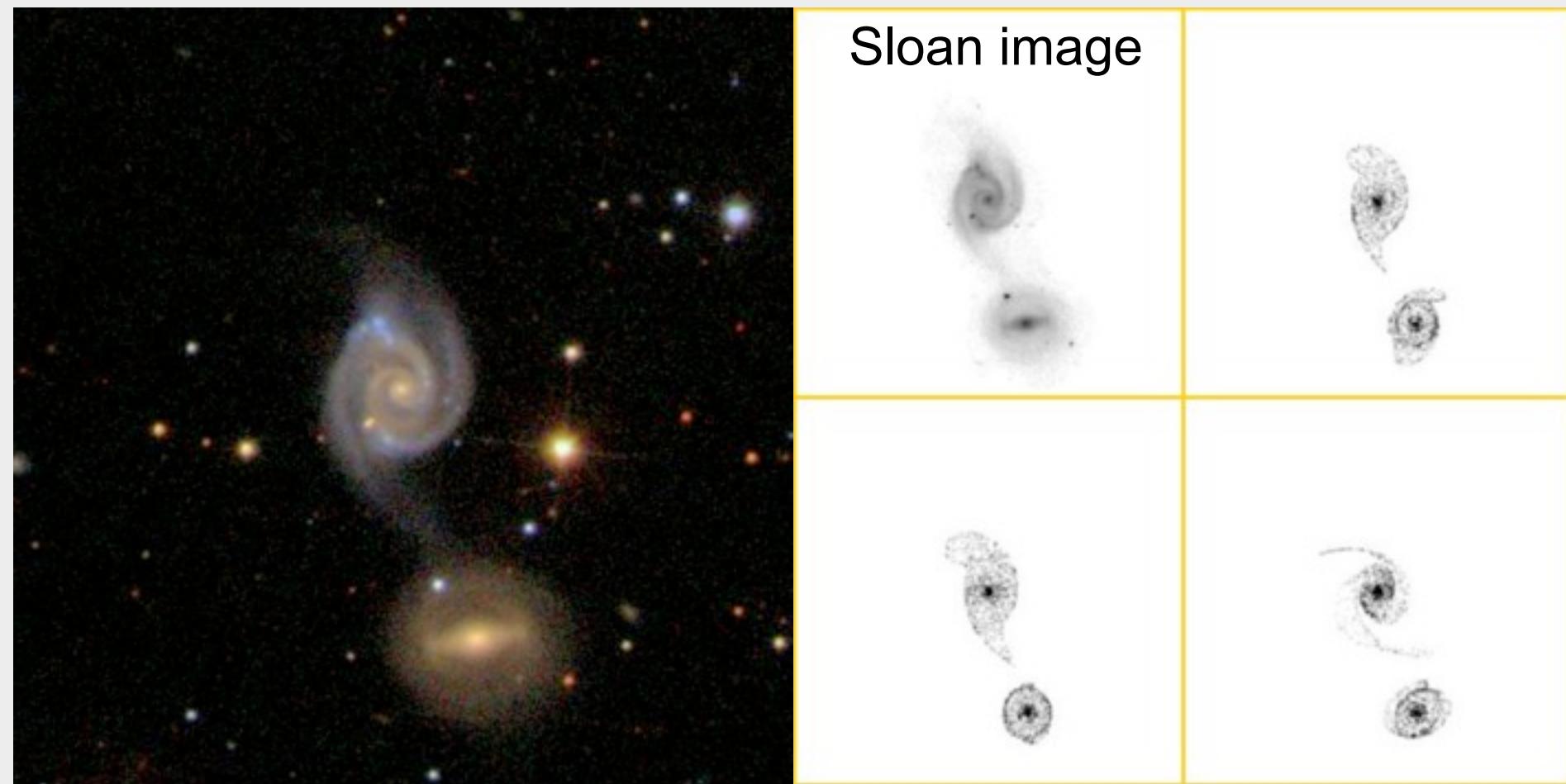
SDSS 587739646743412797

Galaxy Mergers Zoo Gallery



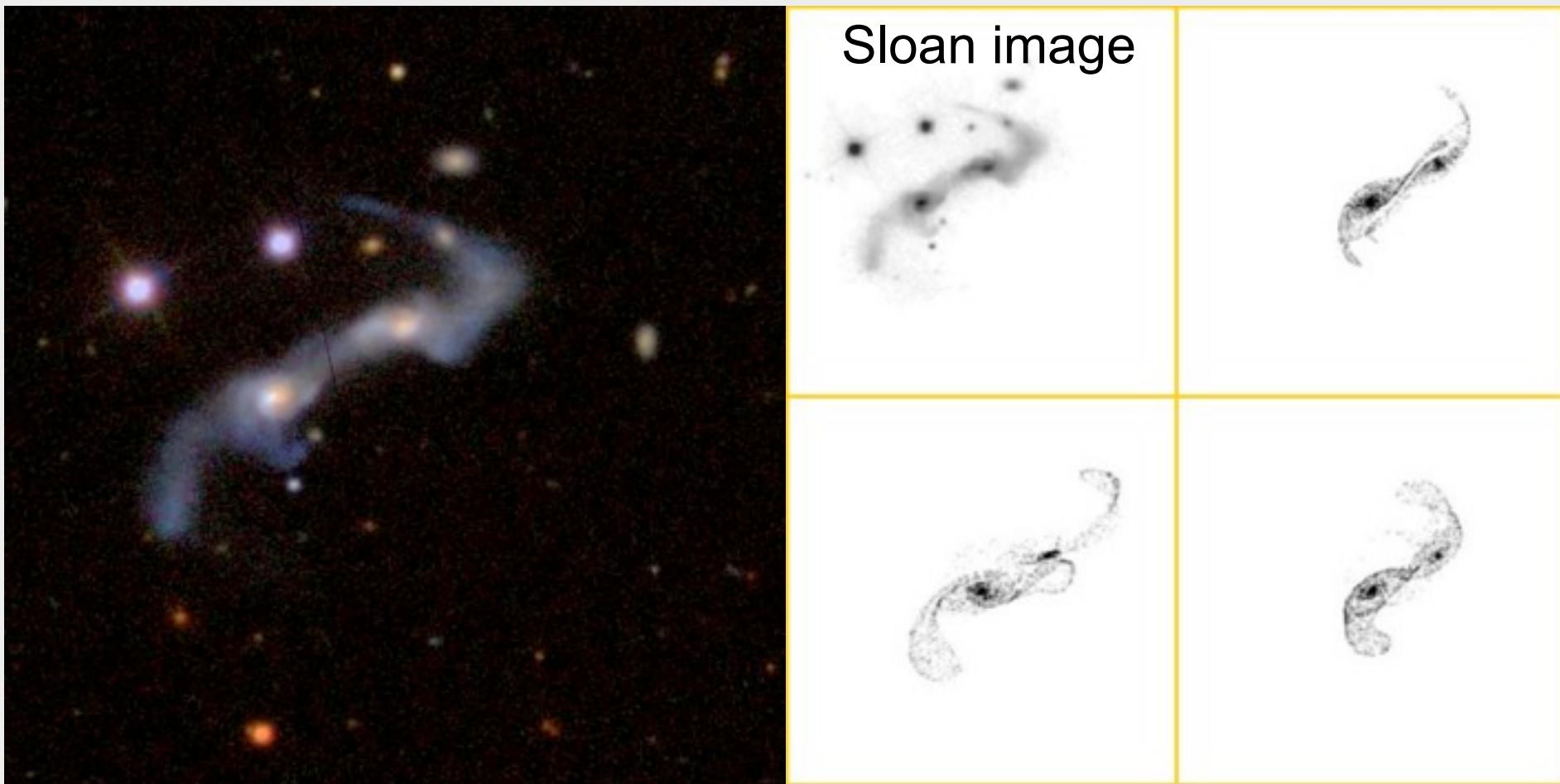
SDSS 587739721900163101

Galaxy Mergers Zoo Gallery



SDSS 587727222471131318

Galaxy Mergers Zoo Gallery



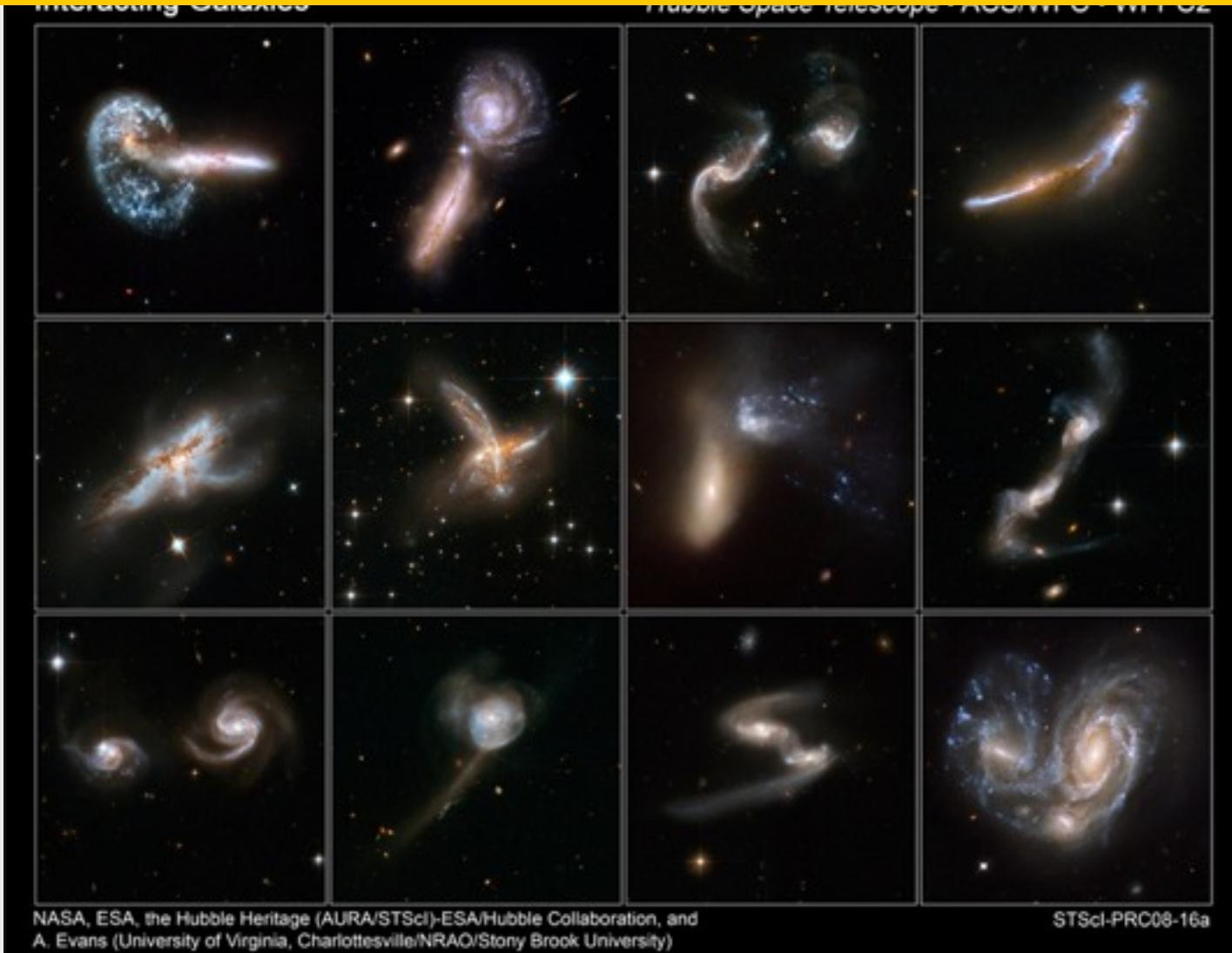
SDSS 58801124116422756

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



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



NASA, ESA, the Hubble Heritage (AURA/STScI)-ESA/Hubble Collaboration, and
A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University)

STScI-PRC06-16a

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

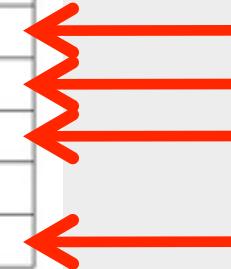


Sloan Science Database Attributes tested

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
$\ln LExp_g$	0.10118
$texture_g$	0.07335
$\ln LDeV_g$	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
$isoAGrad_u * z$	$isoAGrad_u * z$
$petroRad_u * z$	$petroRad_u * z$
$texture_u$	$texture_u$
$isoA_z * z$	$isoA_z * z$
$\ln LExp_u$	$\ln LExp_u$
$\ln LExp_g$	$\ln LExp_g$
$isoA_u * z$	$petroRad_u * z, isoB_z * z,$ $isoBGrad_u * z, \ln LExp_g$
$isoB_z * z$	$isoAGrad_u * z, \ln LExp_g$
$isoBGrad_u * z$	$petroRad_u * z, isoA_u * z, isoB_z * z,$ $\ln LExp_g$
$isoAGrad_z * z$	$isoAGrad_u * z, isoBGrad_u * z,$ $\ln LExp_g$



Sloan Science Database Attributes found !!

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
$\ln LExp_g$	0.10118
$texture_g$	0.07335
$\ln LDeV_g$	0.06864
$petroMag_{gr}$	0.06626
$isoAGrad_u * z$	0.05729

Correlation Zoo !

Results of cluster separation analysis

Best Separation in Single Dimension	Best Separation Among 1014 Combinations
$isoAGrad_u * z$	$isoAGrad_u * z$
$petroRad_u * z$	$petroRad_u * z$
$texture_u$	$texture_u$
$isoA_z * z$	$isoA_z * z$
$\ln LExp_u$	$\ln LExp_u$
$\ln LExp_g$	$\ln LExp_g$
$isoA_u * z$	$petroRad_u * z, isoB_z * z, isoBGrad_u * z, \ln LExp_g$
$isoB_z * z$	$isoAGrad_u * z, \ln LExp_g$
$isoBGrad_u * z$	$petroRad_u * z, isoA_u * z, isoB_z * z, \ln LExp_g$
$isoAGrad_z * z$	$isoAGrad_u * z, isoBGrad_u * z, \ln LExp_g$



Combinatorial Explosion !!

Challenge Problems

- Zooniverse Data Mining (Machine Learning) Challenge Problems (2011-2013)



Research Awards

Other similar examples:

- KDD cups
- Netflix Prize (#1 and #2)
- GREAT08 Challenge
- Digging into Data Challenge 2009 (diggingintodata.org)
- Transportation challenge problems
- KD2u.org – knowledge discovery from challenge data sets
- Photometric redshift (photo-z) challenge
- Supernova Classification Challenge (ends May 1, 2010)

Challenge Problems

- Zooniverse Data Mining (Machine Learning) Challenge Problems (2011-2013)



Other similar examples:

- KDD cups
- Netflix Prize (#1 and #~~2~~)
- GREAT08 Challenge
- Digging into Data Challenge 2009 (diggingintodata.org)
- Transportation challenge problems
- KD2u.org – knowledge discovery from challenge data sets
- Photometric redshift (photo-z) challenge
- Supernova Classification Challenge (ends May 1, 2010)

Next in the queue: *Light Curve Zoo (LCZ)*

- LCZ
 - development test project (2010-2012) for LSST
- What?
 - Explore the effectiveness of Citizen Scientists to characterize light curves (photometric time series)
- Eventual application: LSST light curves
- Initial implementation: MACHO light curves
- When?
 - Design and implementation – next 6-12 months
 - Deployment – 2011

Light Curve Zoo (LCZ)

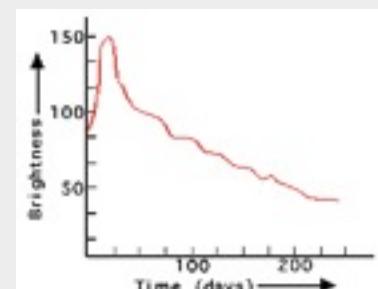
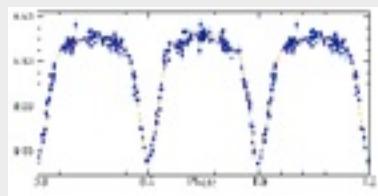
- User experience:
 - Similar to Galaxy Zoo 2: user-directed decision tree
 - Periodic or non-periodic?

- Periodic:

- Select trial periods, amplitudes, phasing zero-points
- Find best-fit light curve:
 - » Compare with sample variability classes, and/or
 - » Using visual inspection, and/or
 - » Plots of residuals

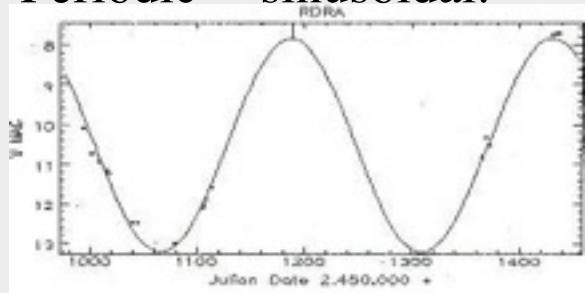
- Non-periodic:

- Select characterizations that describe the light curve:
 - » amplitude, shape, color, rise time, decay time, duty cycle
 - » These will be fed to scientists and to classifiers for classification.

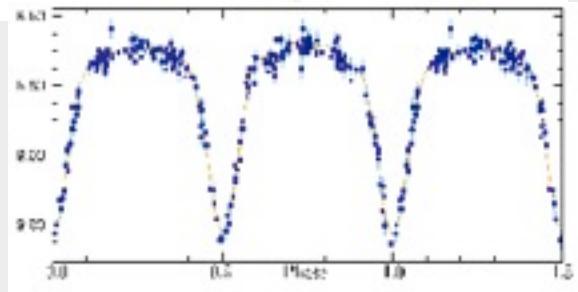


Sample training set of light curves for Categorization of Time Series Behavior

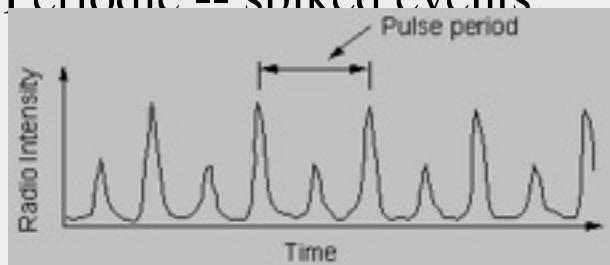
- Periodic -- sinusoidal:



- Periodic -- smooth non-sine:

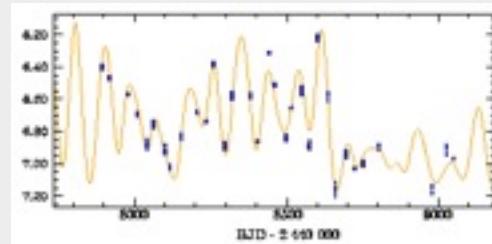


- Periodic -- spiked events:

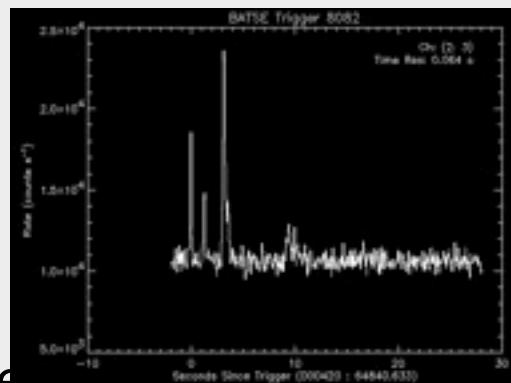


(Chirp)

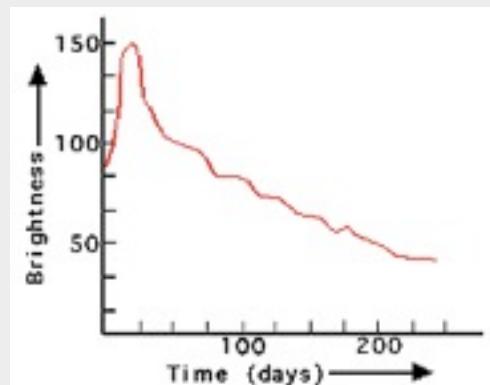
- Aperiodic events (noise?):



- Single spiked events:

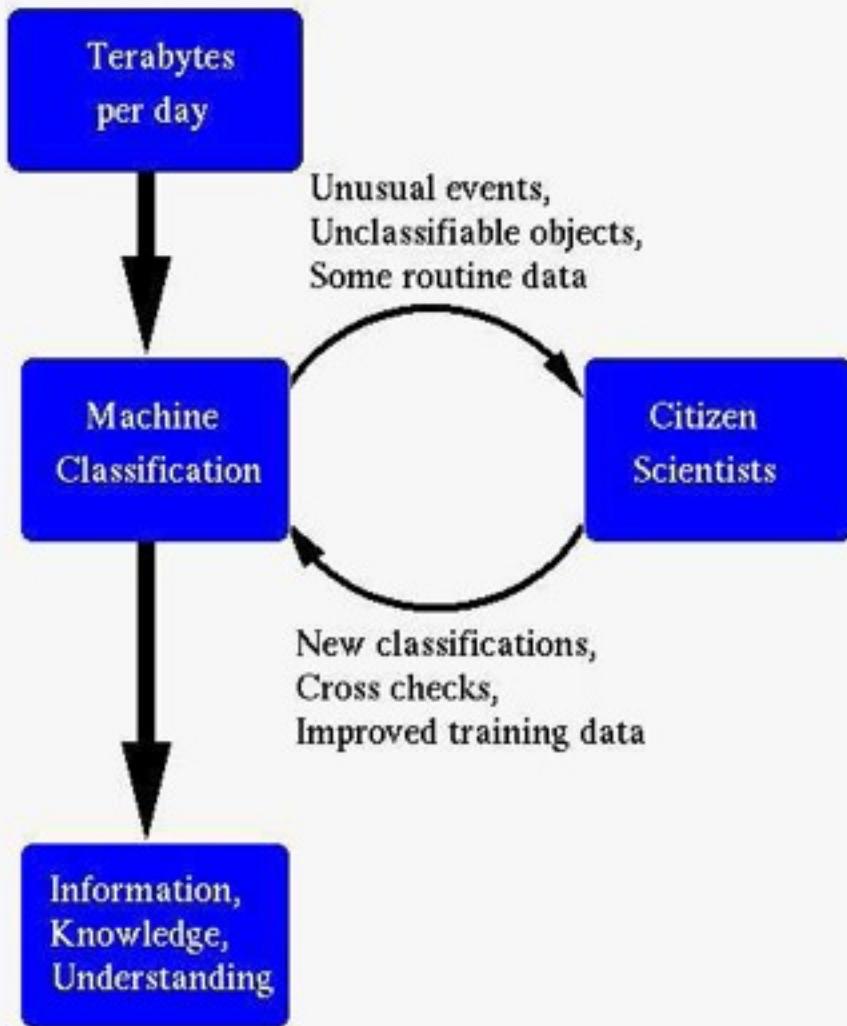


- Single long duration events:



Challenge Areas and The Future Man-Machine Partnership

- Data volumes
- Scalability
- Real-time analytics
- One-pass data stream
- Trust



Related References

- Borne (2009): “***U-Science***”, <http://essi.gsfc.nasa.gov/pdf/Borne2.pdf>
- Borne, Jacoby, ..., Wallin (2009): “***The Revolution in Astronomy Education: Data Science for the Masses***”, <http://arxiv.org/abs/0909.3895>
- Borne (2009): “***Astroinformatics: A 21st Century Approach to Astronomy***”, <http://arxiv.org/abs/0909.3892>
- Dutta, Zhu, Mahule, Kargupta, Borne, Lauth, Holz, & Heyer (2009): “***TagLearner: A P2P Classifier Learning System from Collaboratively Tagged Text Documents***”, accepted paper for ICDM-2009.
- M. F. Goodchild (2007): “***Citizens as Sensors: the World of Volunteered Geography***”, GeoJournal, 69, pp. 211-221.
- Lintott et al. (2009): “***Galaxy Zoo: 'Hanny's Voorwerp', a quasar light echo?***”, <http://arxiv.org/abs/0906.5304>
- Raddick et al. (2009): “***Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers***”, <http://arxiv.org/abs/0909.2925>
- Raddick, Bracey, Carney, Gyuk, Borne, Wallin, & Jacoby (2009): “***Citizen Science: Status and Research Directions for the Coming Decade***”, <http://www8.nationalacademies.org/astro2010/DetailFileDisplay.aspx?id=454>