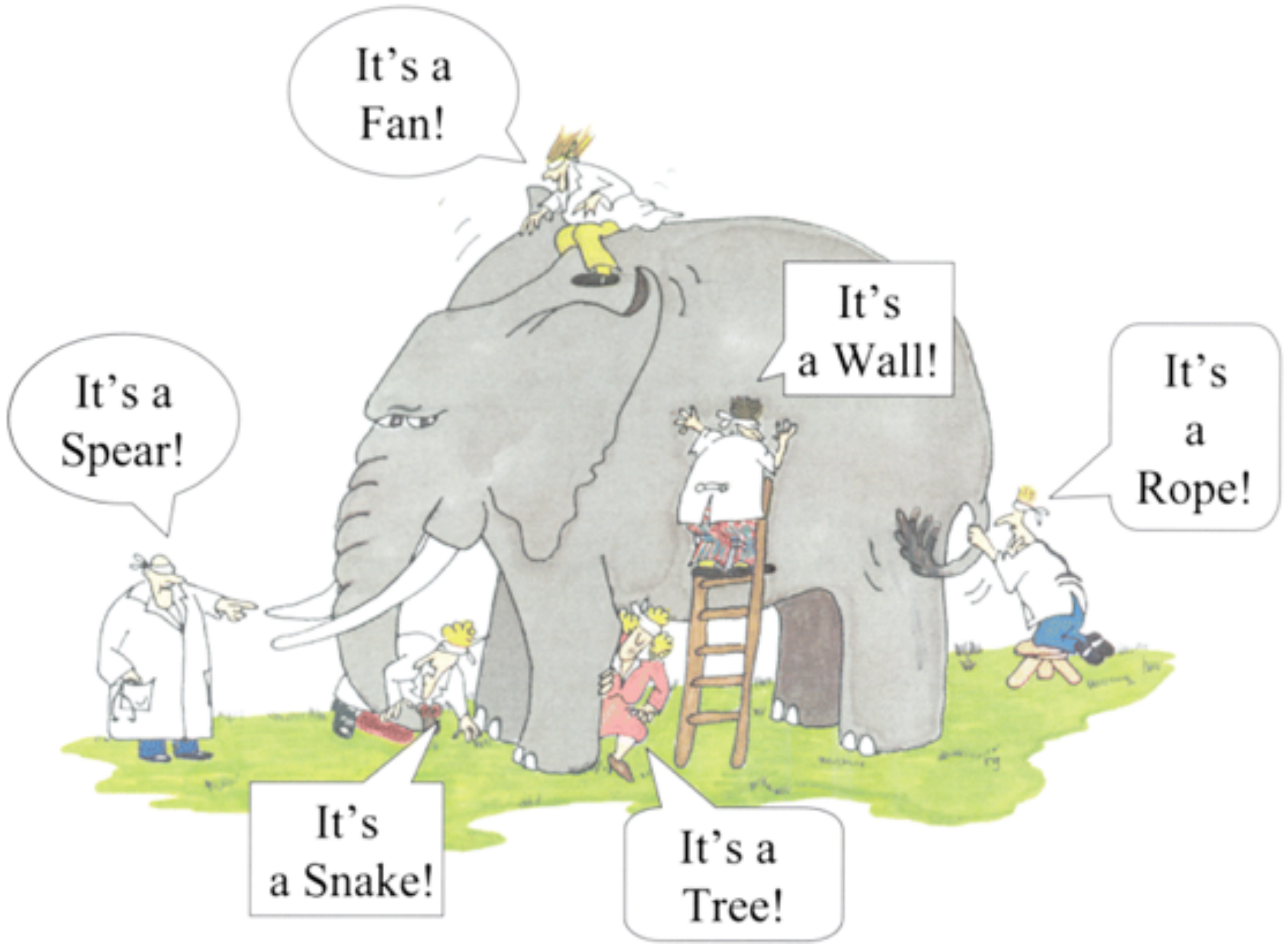
Classification/Characterization TVS subgroup report



The LSST Classification/Characterization Team

(with a heavy dose of editorialization by Adam Miller)

Argonne National Laboratory
24 Mar 2016

Direction

Science Drivers

- discovery of previously unknown transients + variables
- improved understanding of variability in rare classes
  - ➡ e.g. what SDSS S82 did for QSOs
- refinement of existing phenomenological classification scheme
  - ➡ e.g. can existing variable and transient classes be further subdivided? or combined?
- Optimization of follow-up resources
- ALL TVS science (?)
  - MAJOR DISCUSSION POINT: should this group be responsible for ALL T+V classification in LSST? downsides for having every group working separately?

Observational Challenges

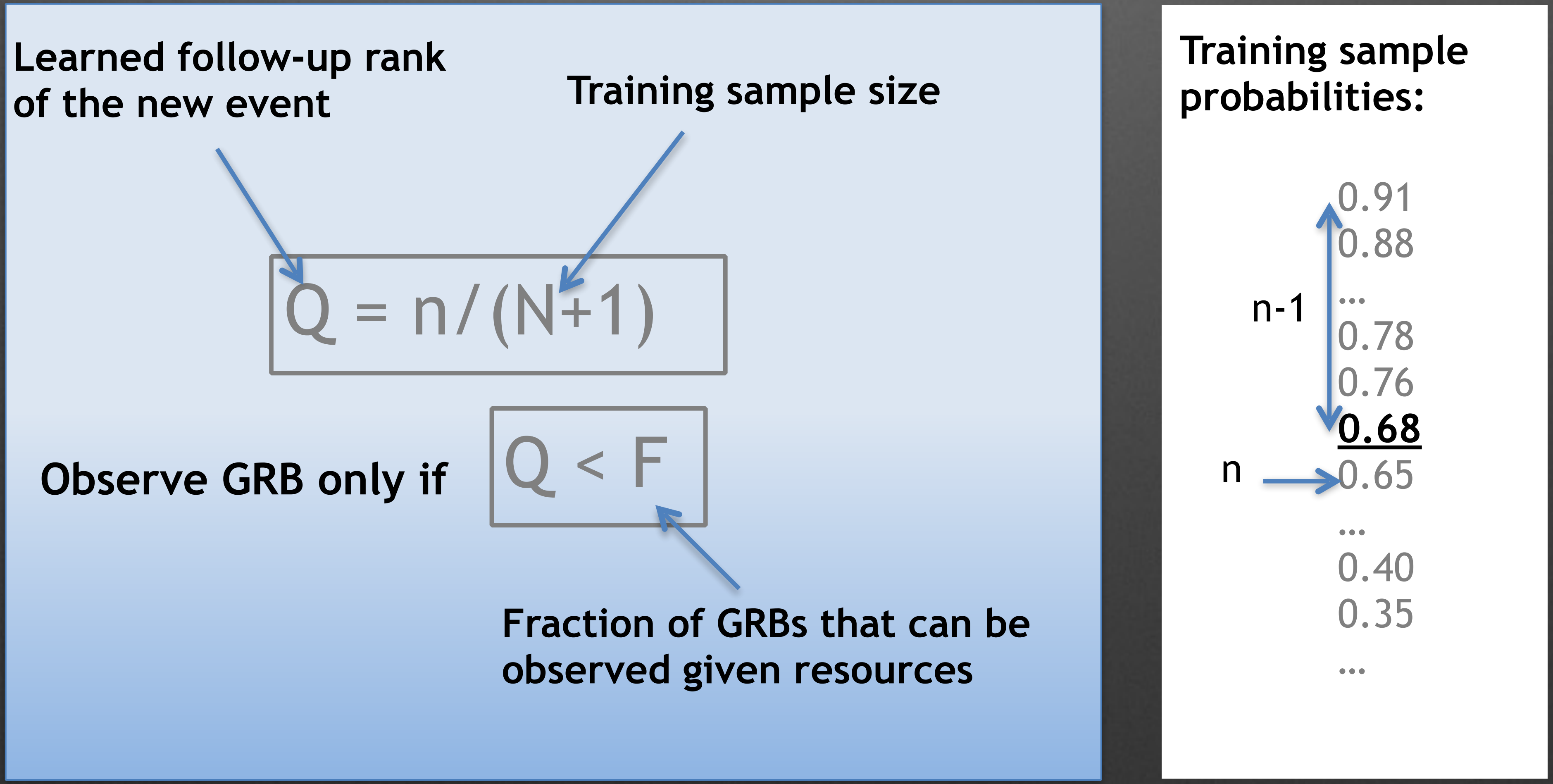
- Classification is strongly sensitive to cadence
  - ➡ insensitive to cadence: short period (<~ 7 d) variables, some flaring sources
  - ➡ highly sensitive to cadence: fast fading transients, SNe
- Lack of resources for faint transient follow-up [also - proper allocation of resources]
- Few existing surveys with comprehensive classification strategies (esp. for faint obj)
- Lack of photo-*z* (during early stages of survey)
- Machine Learning (real-bogus?) required for positive transient detection (?)

Key Questions for LSST

- Depth of LSST is unique, only survey capable of measuring rates for extremely rare events
  - ➡ e.g. orphan GRB afterglows

Direction: Follow-up allocation

- Machine learning can help to prioritize follow-up targets and balance resources
- Example: selection of high-z GRBs using promptly available Swift data products
- Ukwatta et al. 2015: high-z classifier and machine-z regressor using random forest
- Morgan et al. 2012: scoring method to schedule follow-up with limited resources



Learned follow-up rank of the new event

Training sample size

$$\boxed{Q = n/(N+1)}$$

Observe GRB only if $\boxed{Q < F}$

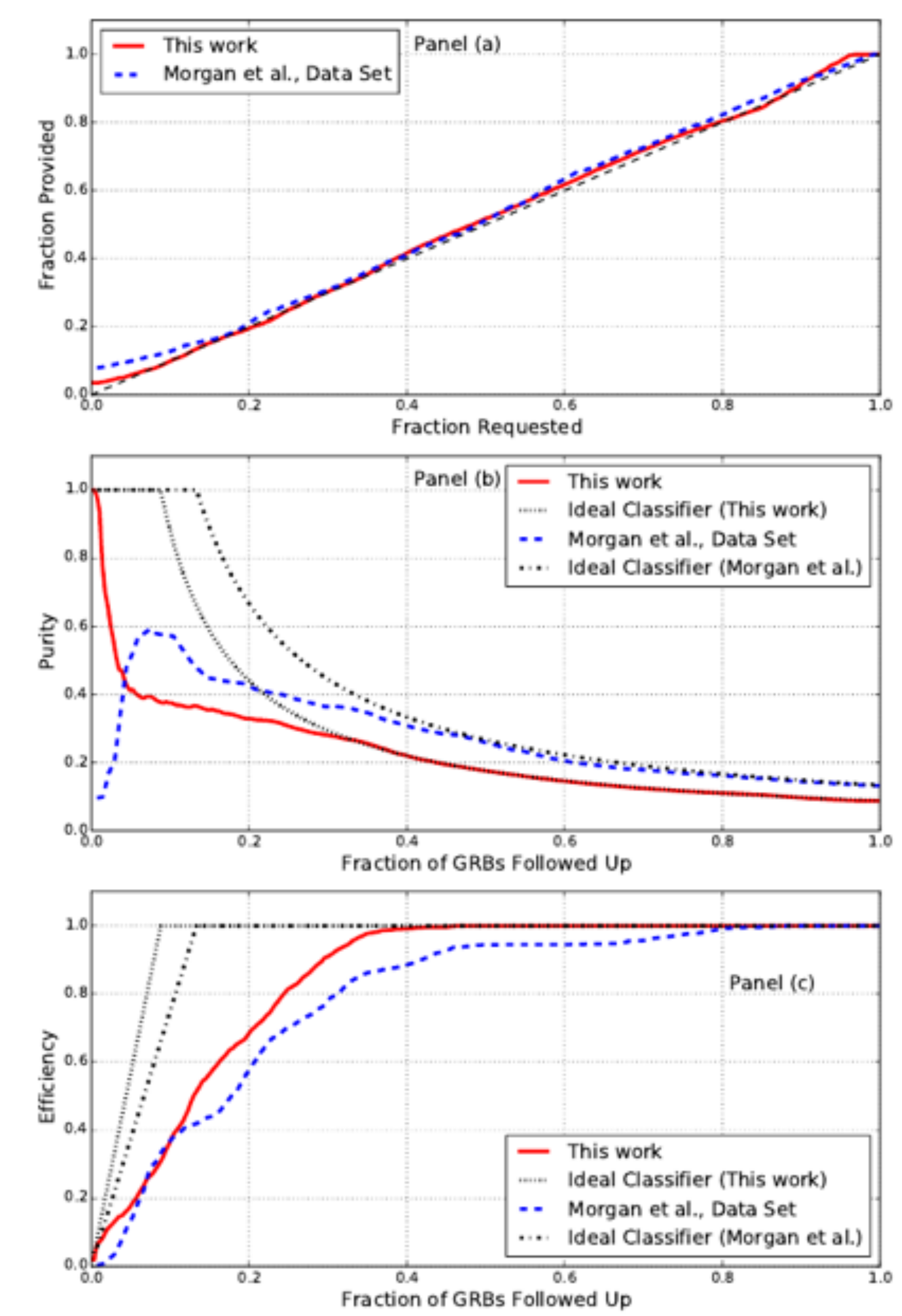
Fraction of GRBs that can be observed given resources

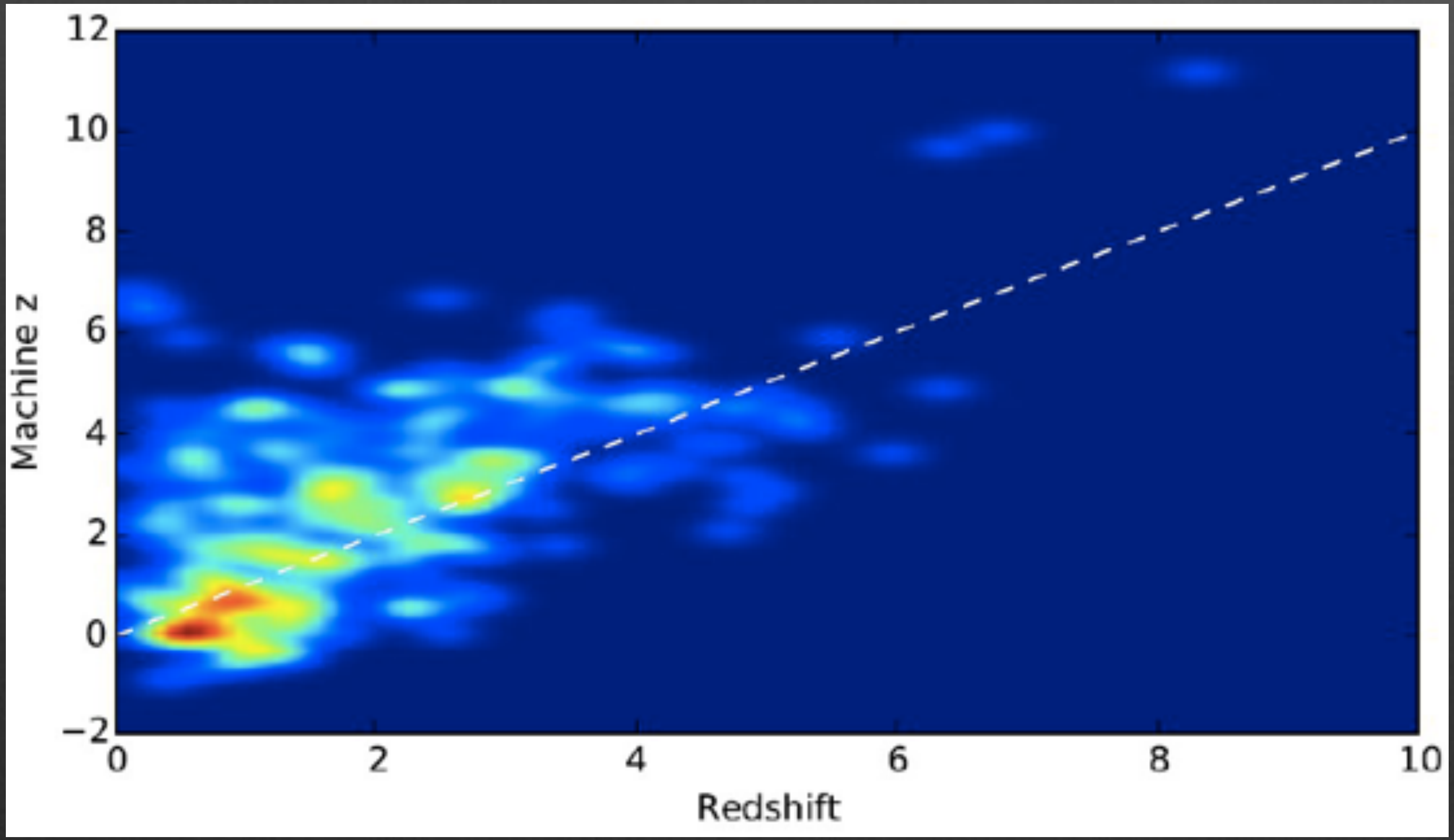
Training sample probabilities:

0.91
0.88
…
0.78
0.76
0.68
0.65
…
0.40
0.35
…

Direction: Follow-up allocation

- Sample of 284 Swift GRBs with spectroscopic *z*
- Cross-validation results:
  - 80% recall with 20% false positive rate
  - 100% recall with 40% false positives
- The redshift correlation coefficient is ~0.6
- Confirmed by new 2015 data





Direction: Class Distributions

Example: Eclipsing Binaries

Traditional approach: *discrete classes*

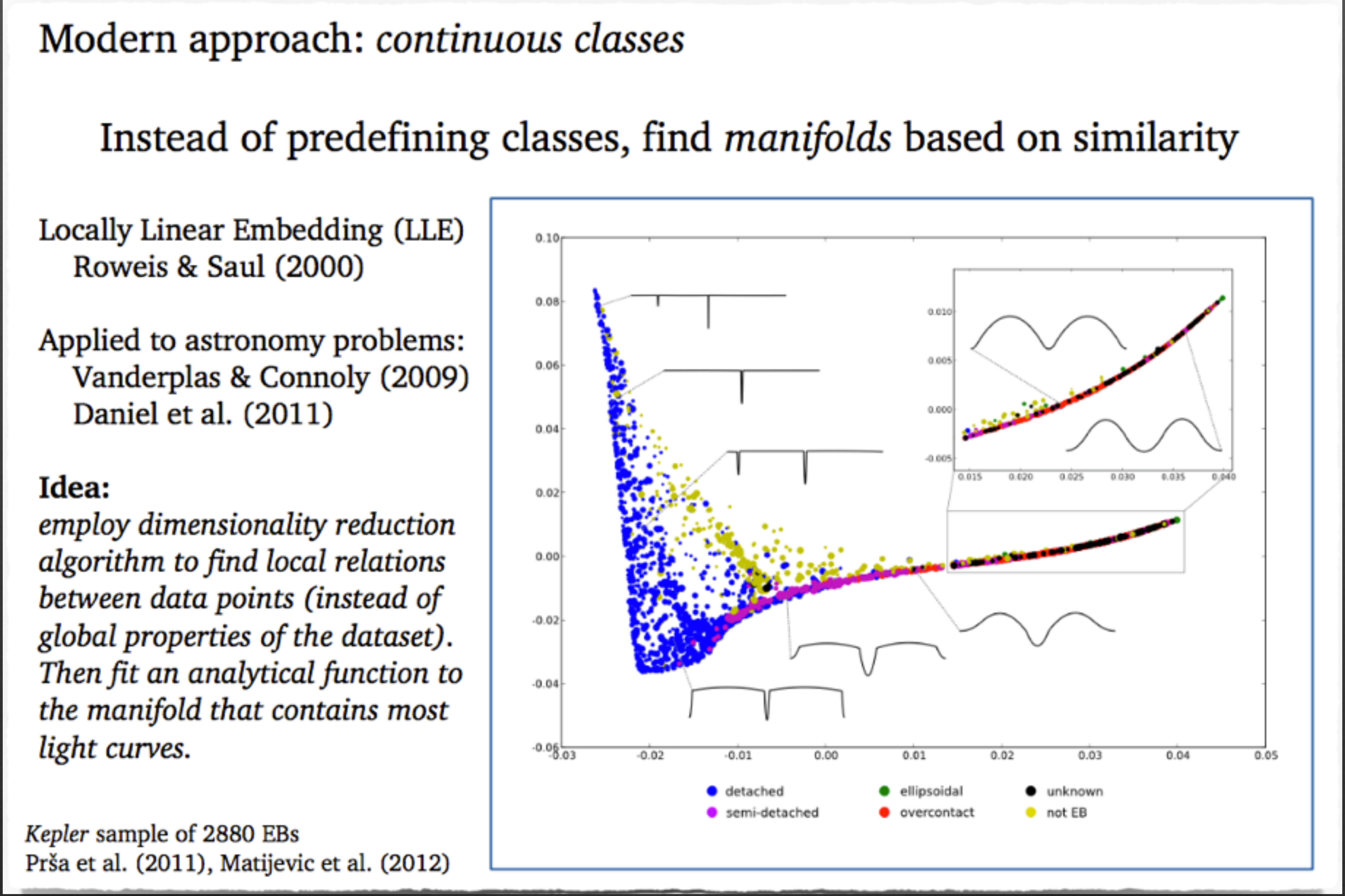
EA (detached), EB (semi-detached), EW (overcontact)
Visual inspection commonly used - inadequate for large data sets

Kepler data classification:
Eyer & Blake (2005)
Sarro et al. (2006)
Debosscher et al. (2007)
Blomme et al. (2011)
Dubath et al. (2011)
Richards et al. (2011)

Most of the methods share the common point of using training sets of known light curves to define the classes to which the unknown light curves are assigned to.

Direction: Class Distributions

Example: Eclipsing Binaries



Modern approach: *continuous classes*

Instead of predefining classes, find *manifolds* based on similarity

Locally Linear Embedding (LLE)
Roweis & Saul (2000)

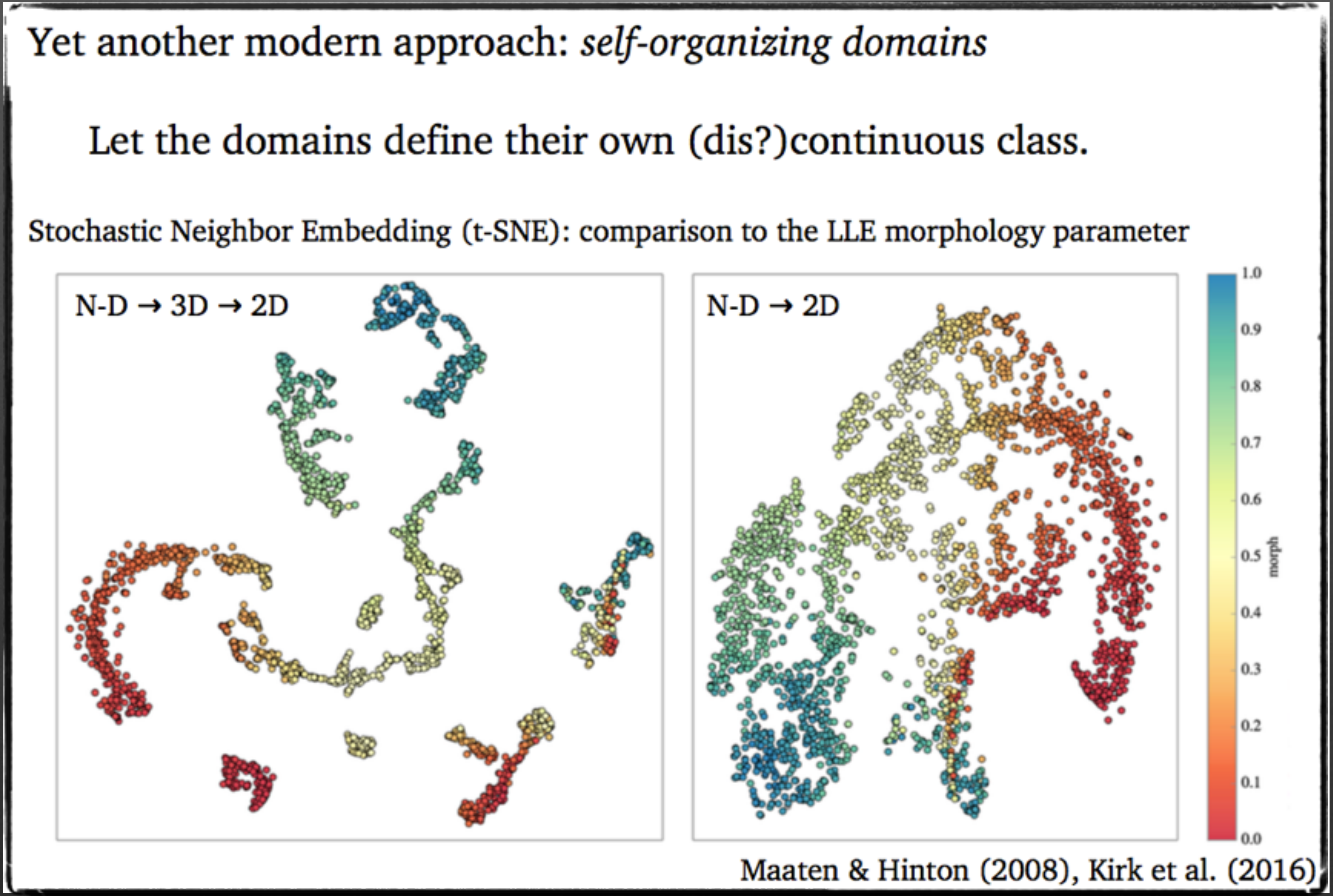
Applied to astronomy problems:
Vanderplas & Connoly (2009)
Daniel et al. (2011)

Idea:
employ dimensionality reduction algorithm to find local relations between data points (instead of global properties of the dataset). Then fit an analytical function to the manifold that contains most light curves.

Kepler sample of 2880 EBs
Prša et al. (2011), Matijevic et al. (2012)

Direction: Class Distributions

Example: Eclipsing Binaries



Yet another modern approach: *self-organizing domains*

Let the domains define their own (dis?)continuous class.

Stochastic Neighbor Embedding (t-SNE): comparison to the LLE morphology parameter

Maaten & Hinton (2008), Kirk et al. (2016)

Direction: Class Distributions

Example: Eclipsing Binaries

Conclusions:

There is a lot to be learned by allowing classes to exist on a manifold

Elimination of discrete subclasses works well for EBs

- possible this will work for other (all?) classes as well

Results are for *Kepler*, but methods extend to *Gaia*

Additional methods are currently being tested

Status

Cadences Explored and Tested

- Limited exploration of rolling cadence (waiting on further output from LSST OPSIM)
- Additional exploration is minimal (or anecdotal)

Current Collaborations and Connections

- Ridgway, Saha — NOAO ANTARES broker development project
- Bellm, Mahabal, Miller, Wozniak — PTF/ZTF members
- Wozniak, Ukwatta — Follow-up resource allocation (focused on GRBs)
- Others with connections to PS1, DES, LINEAR, etc.
  - ➡ overall group structure does not currently exist

Synergy w/ other Subgroups

- Some connection to Informatics & Statistics has been made
- All TVS subgroups rely on classification — each develops independent methods?

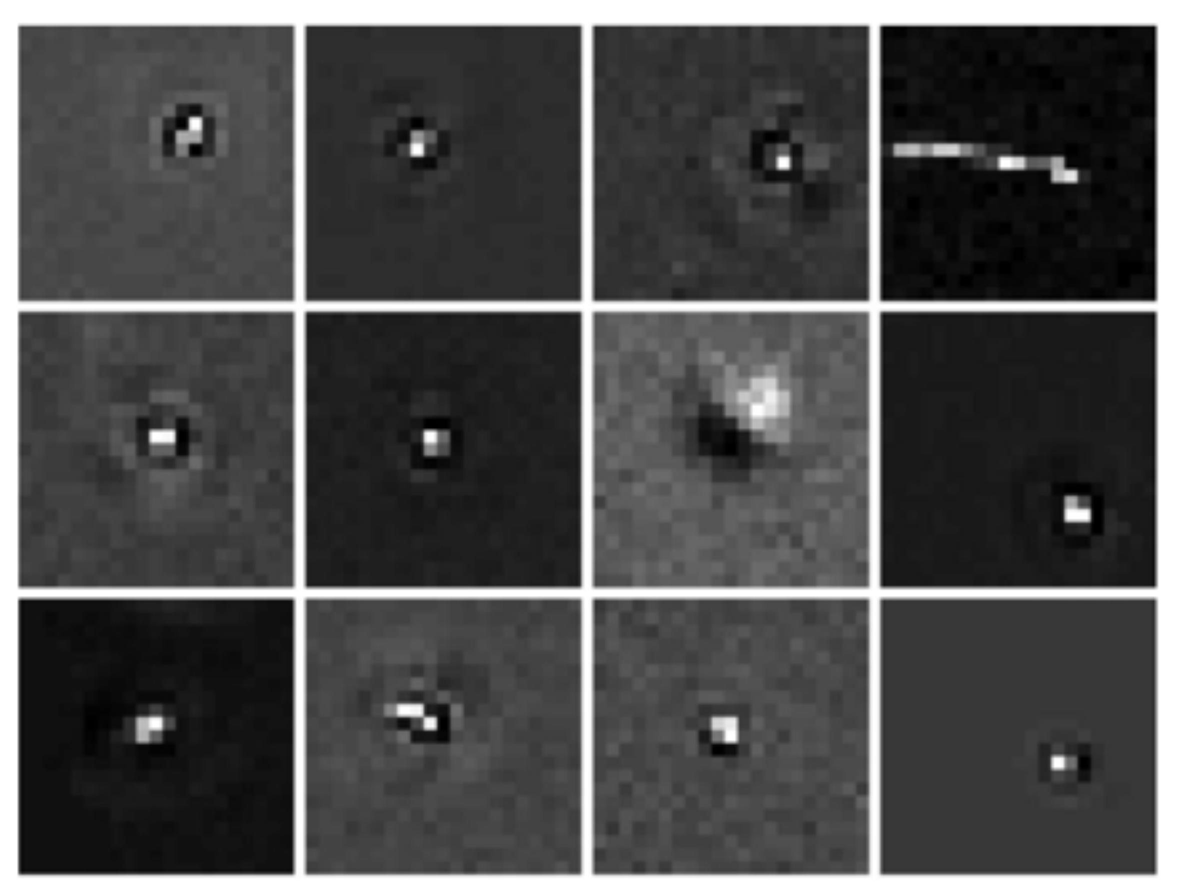
➡ [again] what is the precise role of the classification/characterization group

- classify everything for everyone?
- develop a few specific methodologies for outliers and follow-up?

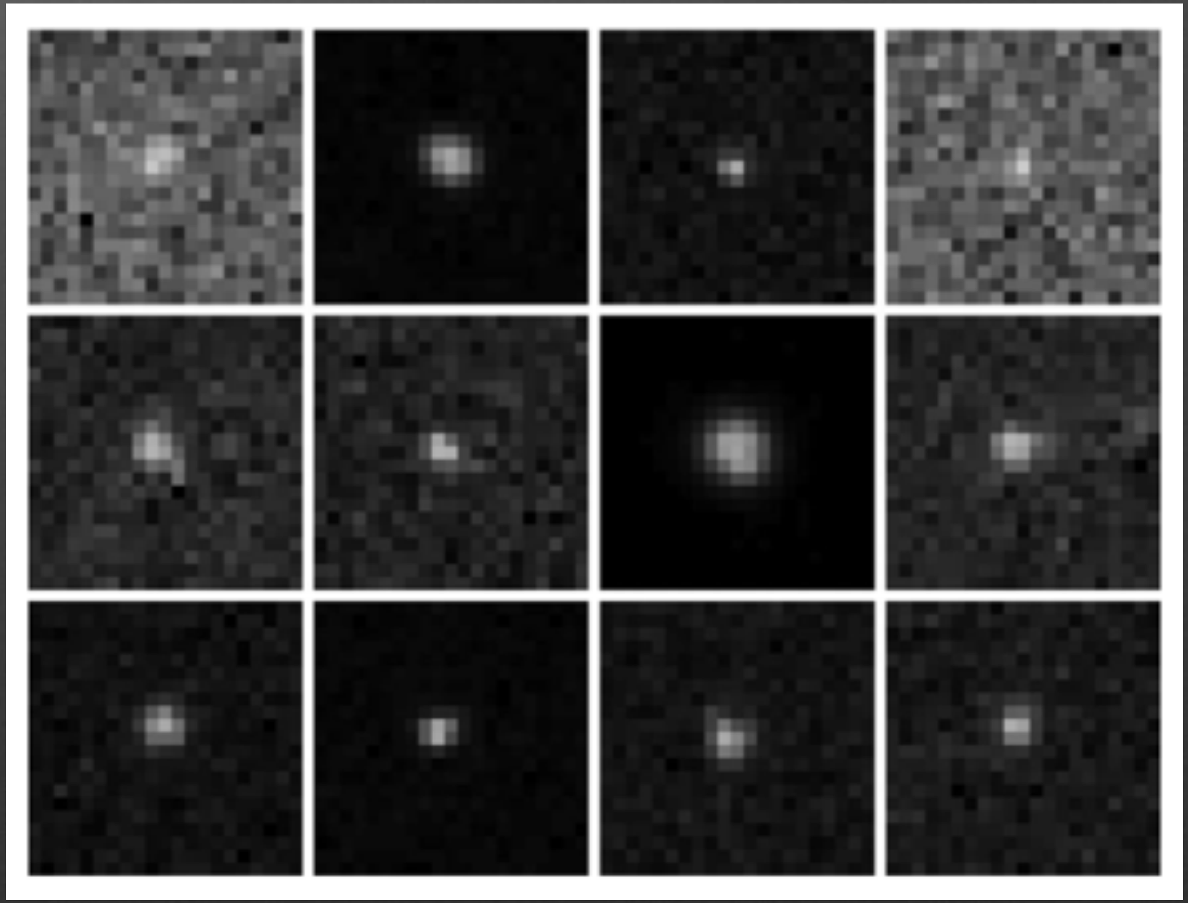
Status: Real-Bogus

- PTF - 1M transient candidates/night
- ~1k real variable sources/night
- Lots of human hours needed (wasted?) filtering candidates

1M of these



only ~1k of these



Bloom+12, Brink+13, Wozniak+13, Rebbapragada+15

Workshop Goals

Subgroups to discuss synergy and friction

- All TVS, Informatics, classification priors from specific subgroups

Specific questions explored and outlined

- What is the role of the Classification subgroup in relation to other subgroups
- What TVS science classes would be harmed by a SSO or SNe optimized cadence
- Can we develop a classification framework that works well independent of cadence

Accomplishments over different LSST phases

- commissioning — ?? extremely challenging to classify sources with ~0 observations
- 1-yr — domain adaptation to classify sources with early light curves?
- 3-yr — rapid identification of new transients/outliers following reference construction
- 10-yr — full classification catalog for all LSST variables?

Technical tools needed to achieve science goals

- Full historical light curves for all T + V — including image diff.
- Forced photometry (on diffs and new) at the location of discovered transients
- Postage stamps for T + V — most ML development currently in image domain
- Details about brokers, what will be delivered to community, and what add-ons are possible

Support needed from LSSTC and TVS chairs

- OPSIM simulations over wide range of cadence + MAF support

Support needed to facilitate collaboration

- TBD [report back in ~48 hr]