Training Set Triage

Umaa Rebbapragada, Ph.D.
Machine Learning and Instrument Autonomy Group

LSSTC Data Science Fellowship Program
Thursday, November 8, 2018
Northwestern University

Research described in this presentation was carried out at the Jet Propulsion Laboratory under a Research and Technology Development Grant, under contract with the National Aeronautics and Space Administration.

Jet Propulsion Laboratory
California Institute of Technology

Something is Not Right

- Performance isn’t acceptable or what you’d hoped for
- Let’s assume that you’ve experimented with different classifiers and you’re using the best performing one.
- How do you debug your machine learning performance?

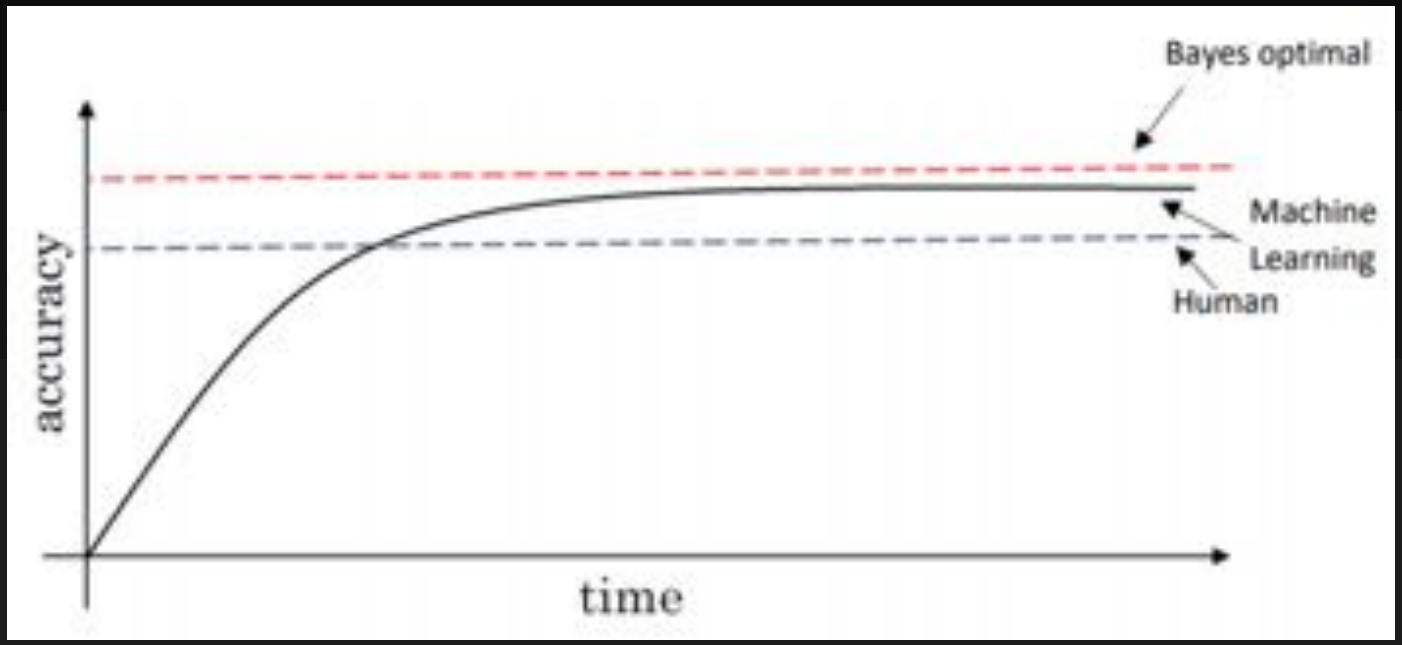
Outline

- Examine all the ingredients:
  - Metrics
  - Features
  - Examples, including Labels
- What debugging looks like at ZTF

Metrics

Metrics

- Bayes Error: best theoretical performance that your classifier can achieve



- How do you know when you've achieved it?

Source: Andrew Ng's Structuring Machine Learning Projects, Coursera

Metrics

- You can't! This guy thinks you can use human performance as a proxy for the Bayes error
- Alternatively, if human performance is better than your ML performance, then you have some hope of improving.

Looking at Train / Test Errors

- Consider this data about your classifier

	Classification error (%)	
	Scenario A	Scenario B
Humans	1	7.5
Training error	8	8
Development error	10	10

- On Scenario A, your training error is much worse than human error. You may have an **avoidable bias**
- On Scenario B, your training error is about the same

Looking at Train / Test Errors

	Classification error (%)	
	Scenario A	Scenario B
Humans	1	7.5
Training error	8	8
Development error	10	10

- In both cases, the test error is 2% more than training error. This is a sign that you have overfitting.

Looking at Train / Test Errors

	Classification error (%)	
	Scenario A	Scenario B
Humans	1	7.5
Training error	8	8
Development error	10	10

- Bias avoidance strategies: change classifiers
- Variance avoidance strategies: tune hyperparameters, use regularization

Features

Features

- Extreme values
  - are those valid or garbage examples?
- Sentinel values - usually these are stand-ins for NaN
  - Invalid values – could be indicative of a problem
- Check with science teams, pipeline people
  - Just because it can be a feature, doesn’t mean it should
  - Toss out features that are known to be problematic

Examples

Examples

- Are there examples left over from engineering or science validation phases of the survey?
- Are your classes balanced? Do you have an extreme minority class?
- **Are your labels contaminated?**

Examples

- Are there examples left over from engineering or science validation phases of the survey?
  - Remove any problematic examples that are “stale”
  - Remove extreme feature values that may be indicative of some type of problem
- Are your classes balanced? Do you have an extreme minority class?
  - Oversample your minority class
  - Undersample your majority class

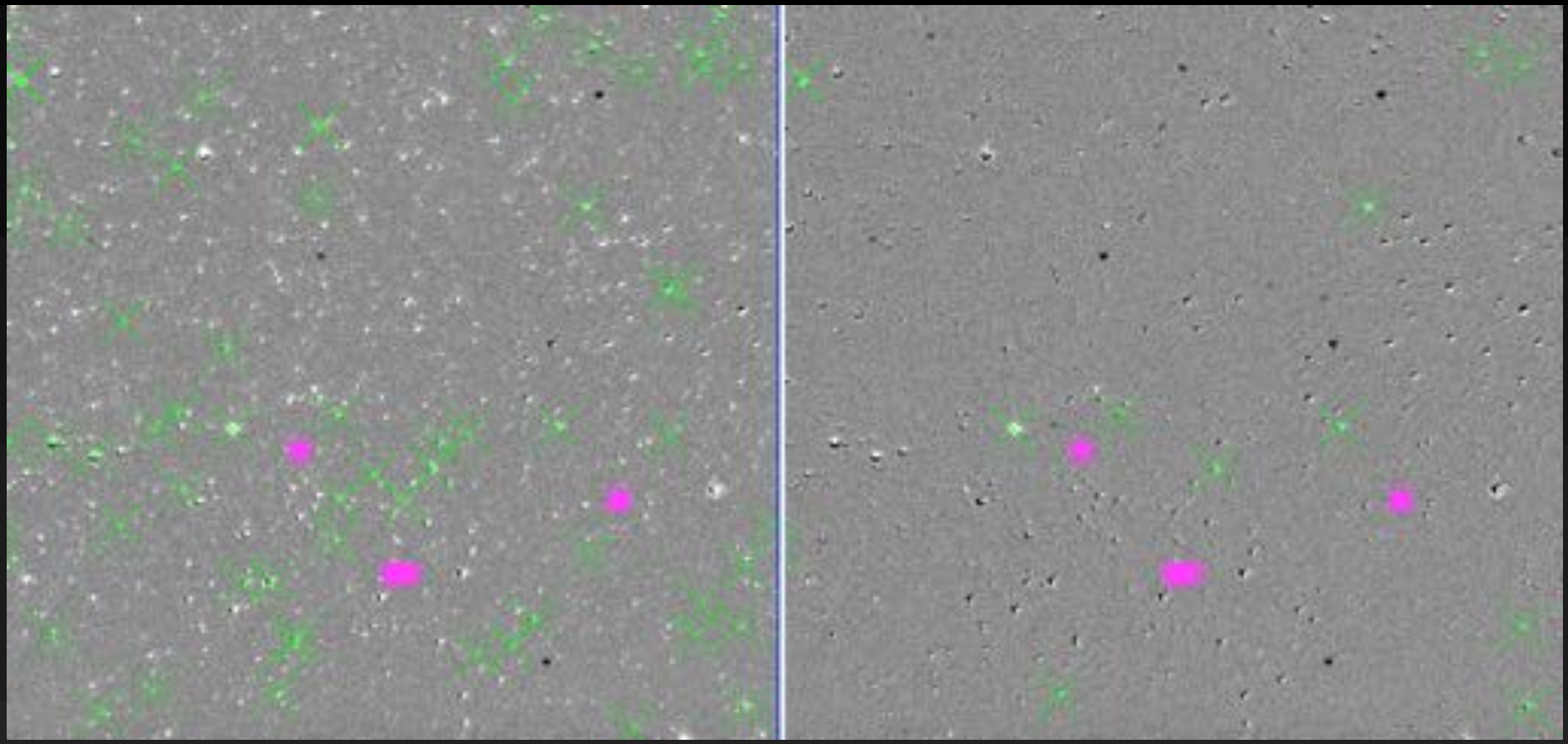
Concept Drift

- When your test distribution starts drifting away from your training distribution
- Why would this happen?

Stuff Happens

- Pipeline upgrades
  - Reference image upgrades
  - Image subtraction changes
- Telescope changes/repairs
- Survey priorities change (e.g., asteroid and GP surveys)
- No one tells the ML team

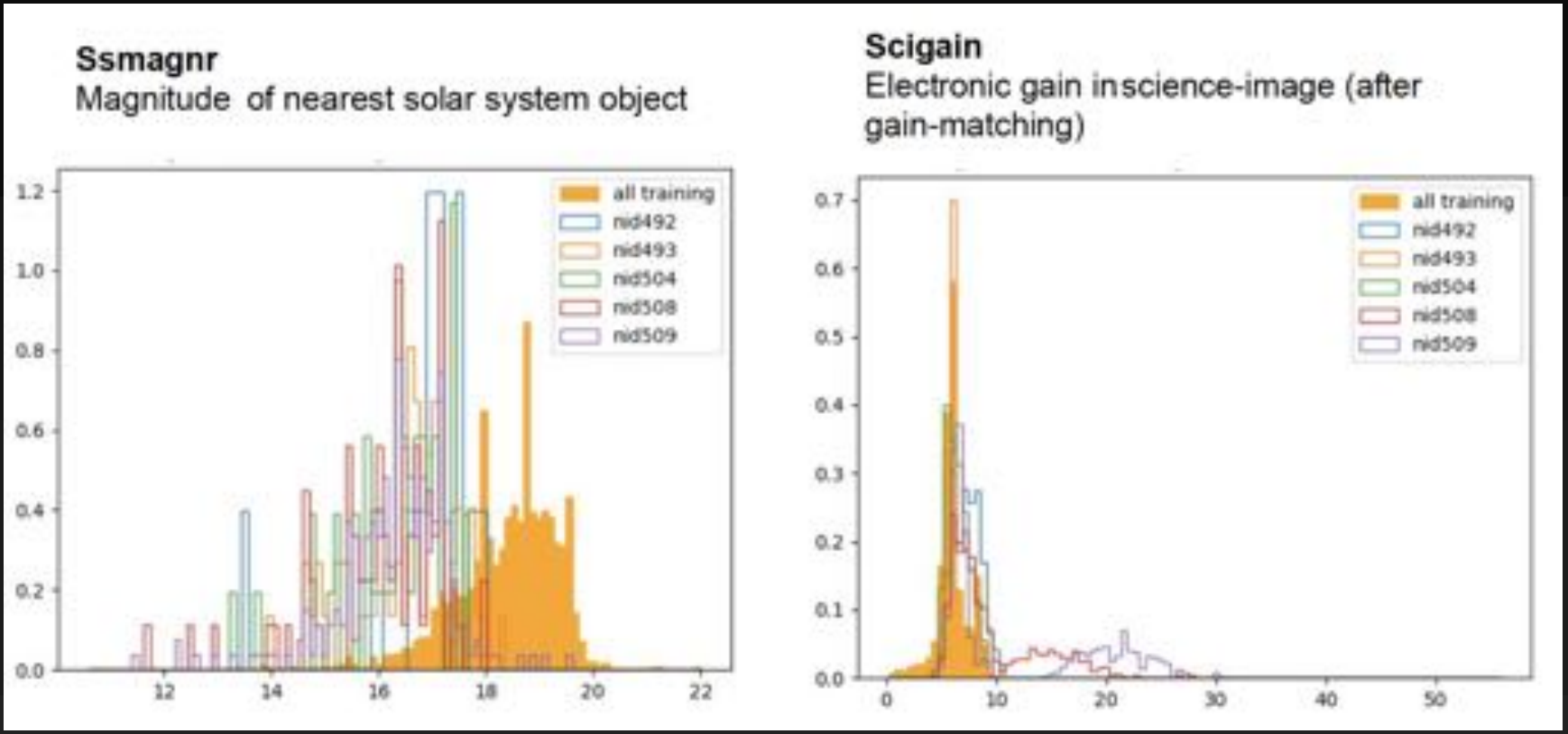
Image Subtraction Upgrade



Before

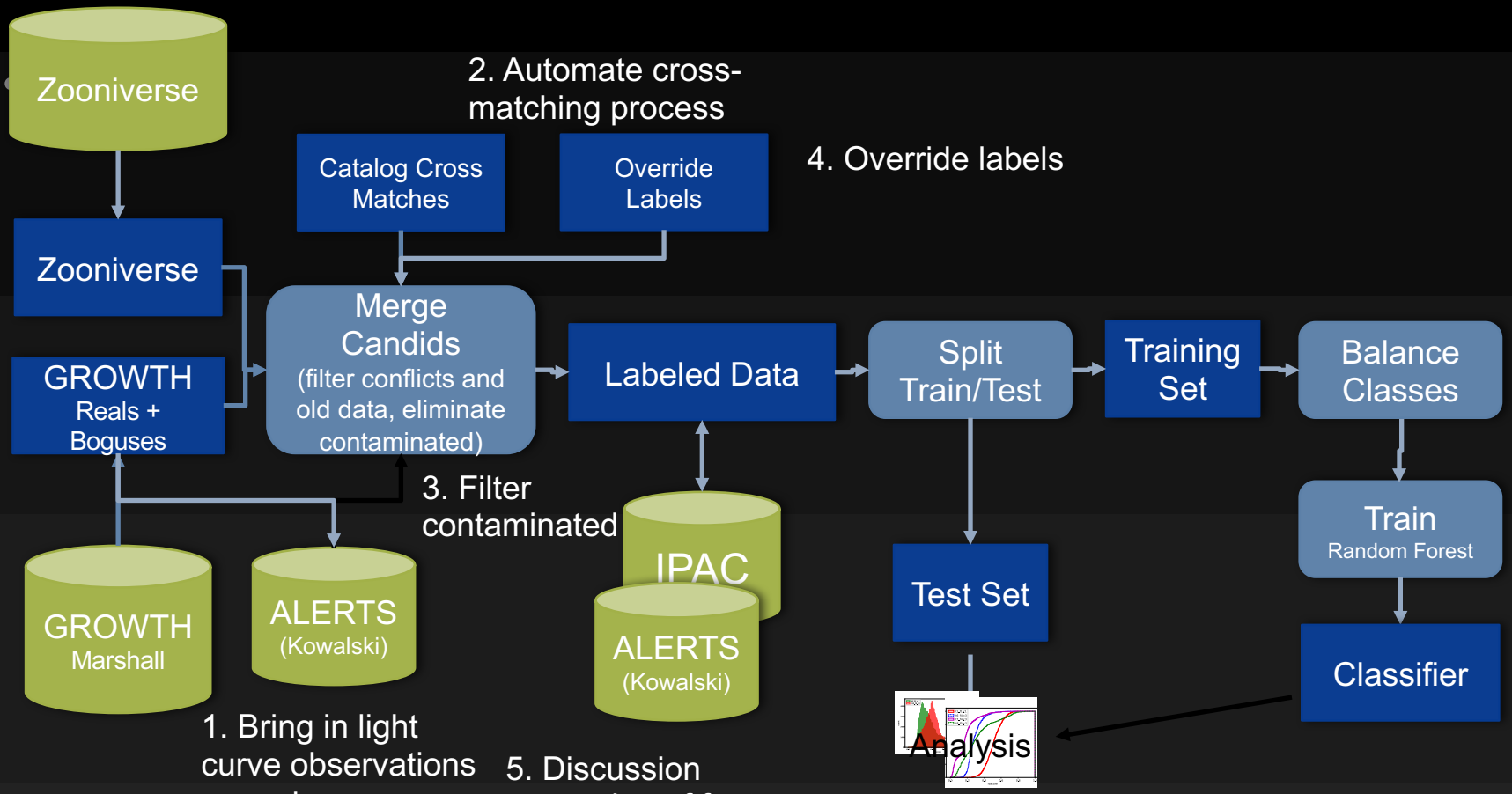
After

Checking Sample Bias

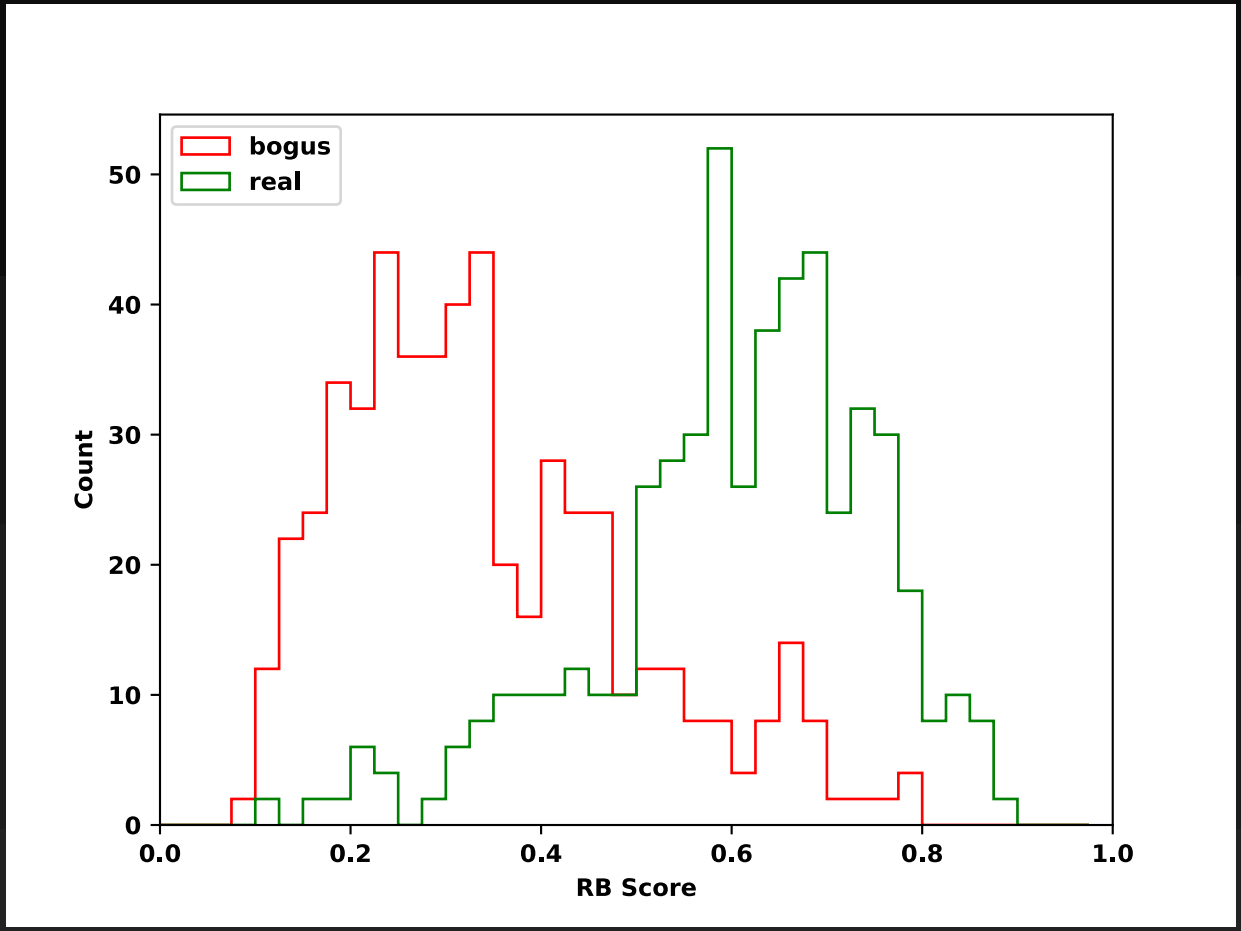


Debugging ZTF

RB Workflow

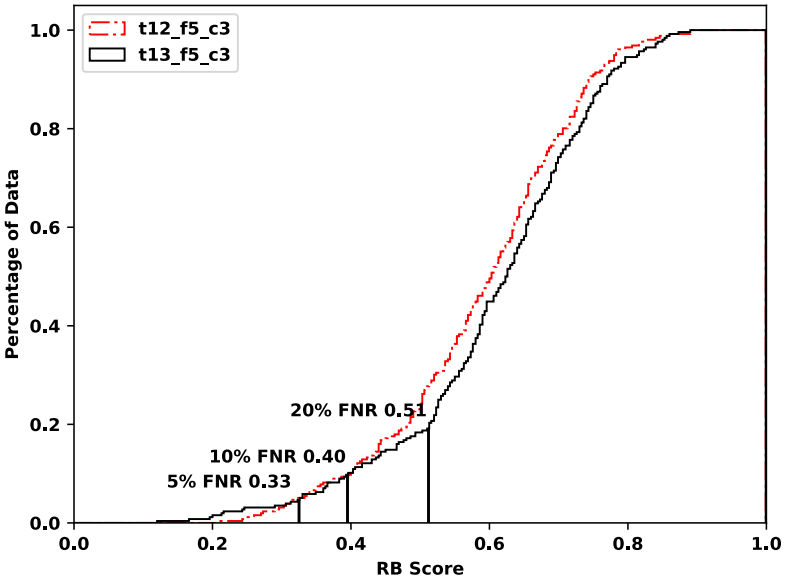

Zooniverse
2. Automate cross-matching process
Catalog Cross Matches
Override Labels
4. Override labels
Zooniverse
GROWTH
Reals + Boguses
Merge Candids
(filter conflicts and old data, eliminate contaminated)
Labeled Data
Split Train/Test
Training Set
Balance Classes
3. Filter contaminated
IPAC
ALERTS
(Kowalski)
Test Set
Train
Random Forest
GROWTH
Marshall
ALERTS
(Kowalski)
Classifier
Analysis
1. Bring in light curve observations as real
5. Discussion expansion of feature space to include alert meta-dataa

Test Set Score Distribution

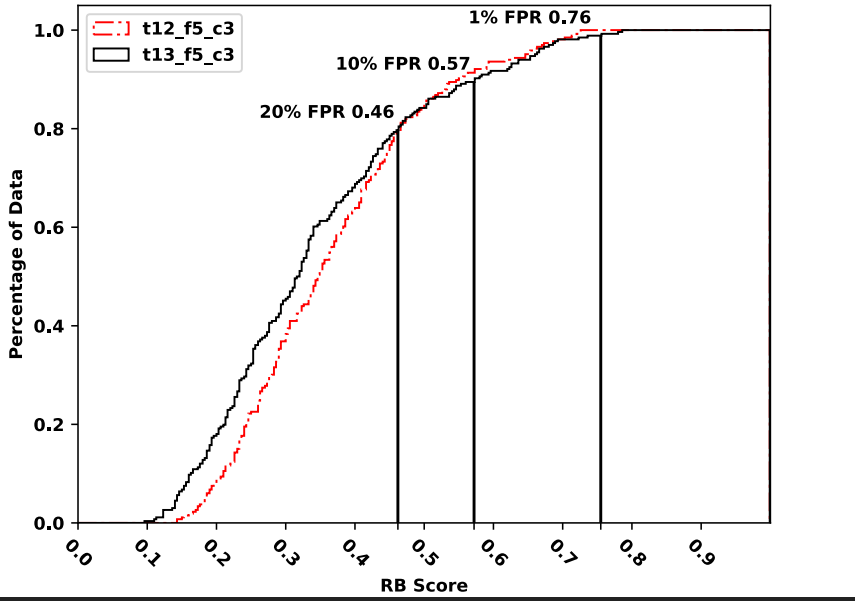


Current ZTF Performance

Bogus

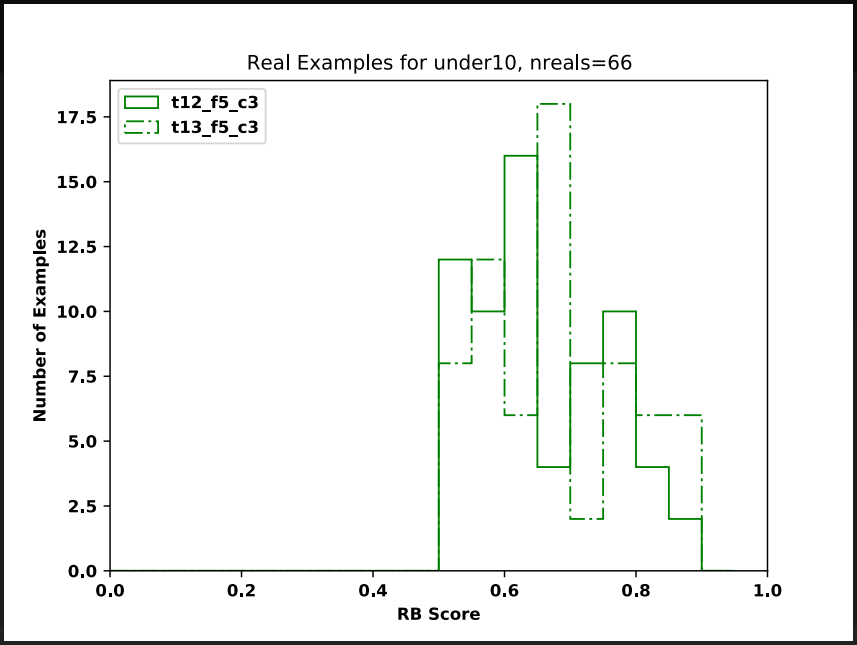


Reals

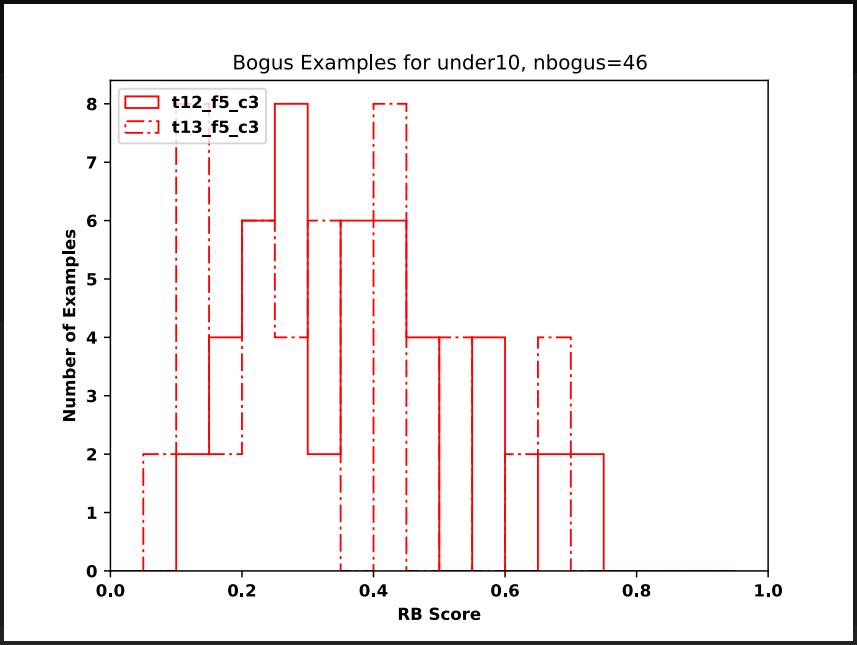


Test on Specific Science Cases – Galactic Plane

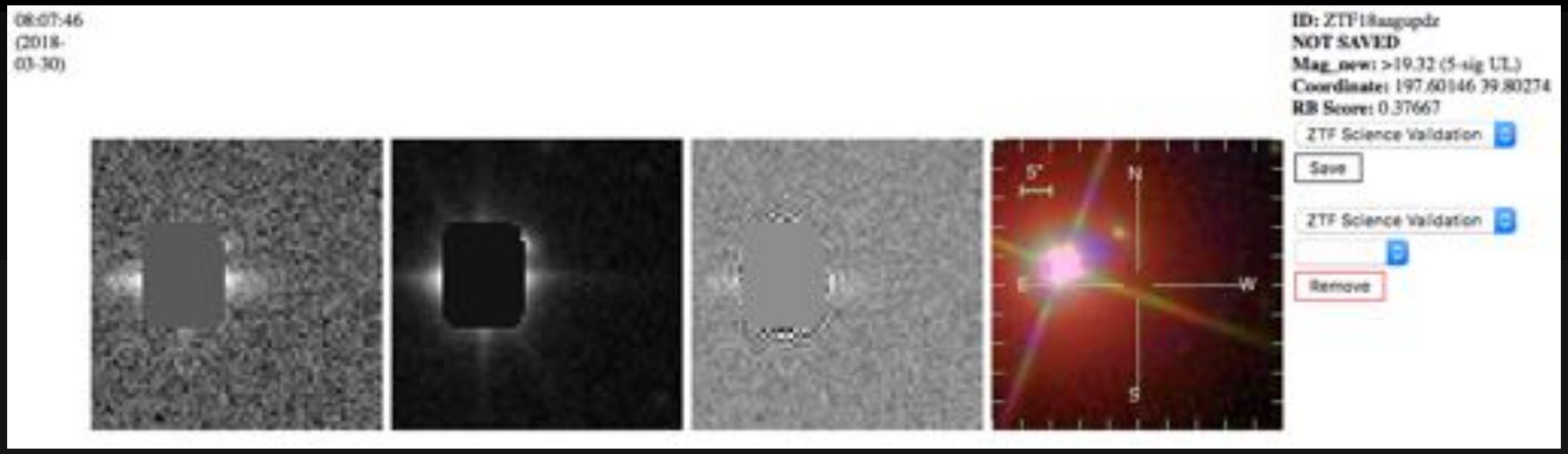
Reals



Bogus



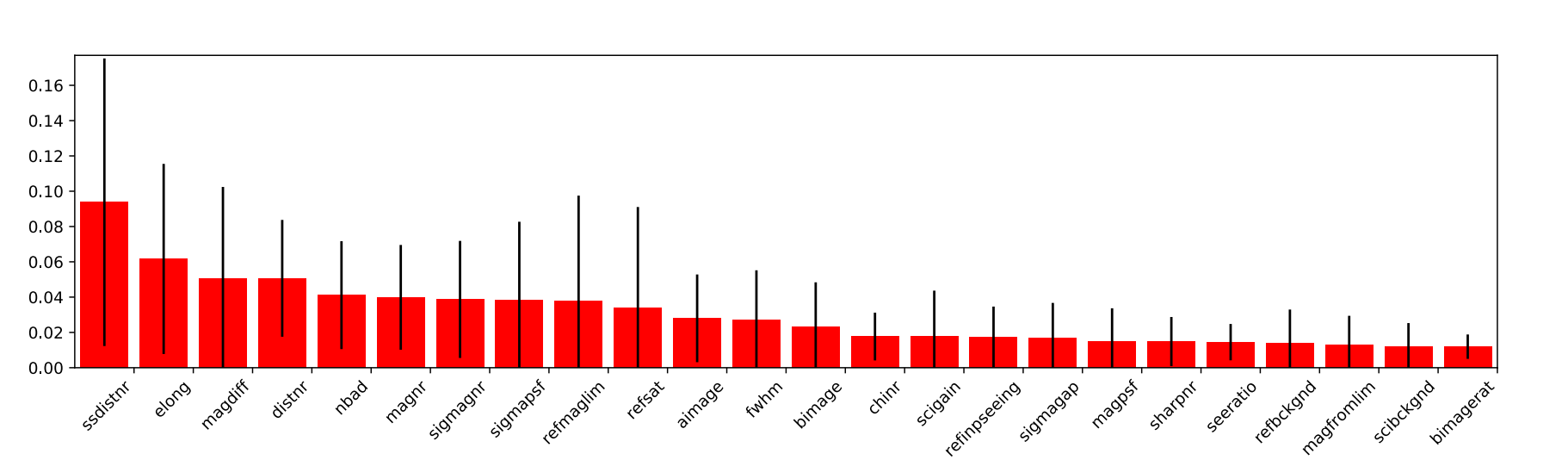
Bright Star Artifacts



candid	name	old rb	new rb
45339966001501500 2	ZTF18aagvjmk	0.433333	0.09333333
45351263581501500 8	ZTF18aagxpee	0.423333	0.13333333
45351974581501500 3	ZTF18aagxrfy	0.39	0.09333333

Feature Importance Diagrams

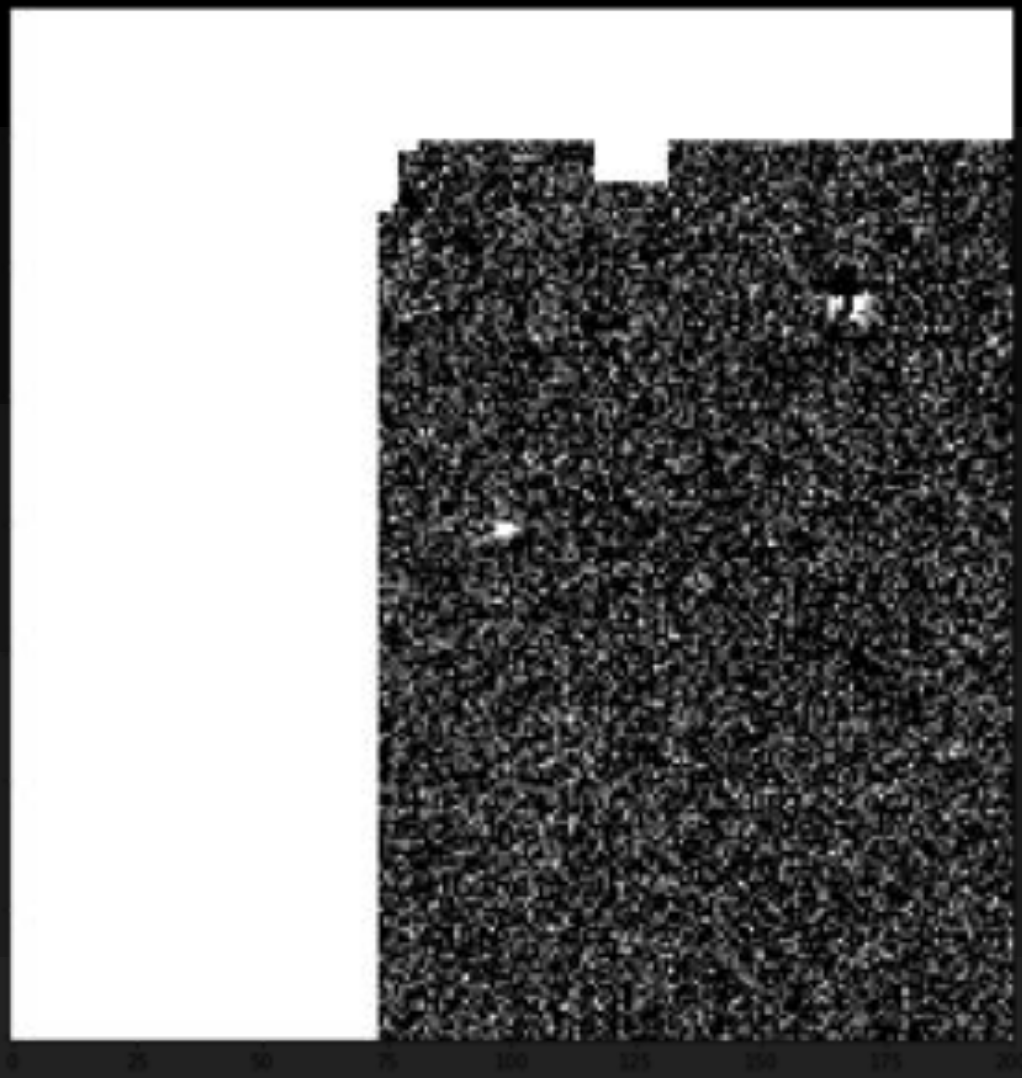
Great for developing intuition about your classifier, but doesn't help with individual examples



```
Prediction 0.703333333333
Bias (trainset prior) 0.812355967078
feature contribution pre-interp_value post-
chinr           -0.0494  4.693     4.693
distnr          -0.0481  2.7167    2.7167
ranr            -0.0413  162.5398  162.5398
magnr           0.0411   19.548    19.548
sharpnr         -0.0395  0.275     0.275
decnr           -0.0257  29.6238   29.6238
refsat          -0.0254  55231.6   55231.6
sigmagapbig     0.0245   0.2076    0.2076
zpmaginpsci     -0.0188  26.1393   26.1393
scigain         -0.0168  5.5119    5.5119
aimage          0.0163   0.833     0.833
magpsf          0.0153   19.2127   19.2127
chipsf          0.0145   4.1436    4.1436
magapbig        0.0124   19.4741   19.4741
mindtoedge      -0.0117  38.1865   38.1865
bimage          0.0113   0.685     0.685
fwhm            0.0108   2.7       2.7
difnumnoisepix  -0.0106  37.7078   37.7078
ncandrefmsciraw 0.0102   3.0       3.0
magap           0.0095   19.5536   19.5536
diffavgsqchg    -0.009   -93.828   -93.828
sigmagnr        0.0089   0.092     0.092
ncandscimrefraw 0.008    31.0      31.0
seeratio        0.0079   1.368     1.368
scisat          -0.0078  54724.7   54724.7
fluxrat         -0.0077  1.1248    1.1248
refmaglim       -0.0069  22.35     22.35
classtar        0.0067   0.911     0.911
magfromlim      0.0065   0.4911    0.4911
difffwhm        -0.0065  3.6935    3.6935
aimagerat       0.0062   0.3085    0.3085
ssdistnr        -0.006   nan       1.5
```

- RB score broken down into sum of bias (proportion of labeled examples that are positive) and contributions of each feature.
- Chinr, distnr and ranr decreased the RB score the most.
- Magnr, sigmagapbig and aimage increased the RB score the most.
- No dominant feature.
- The bias was very high, making it difficult for the features to decrease the RB score enough.

ZTFaabasre (bogus candidate classified as real)



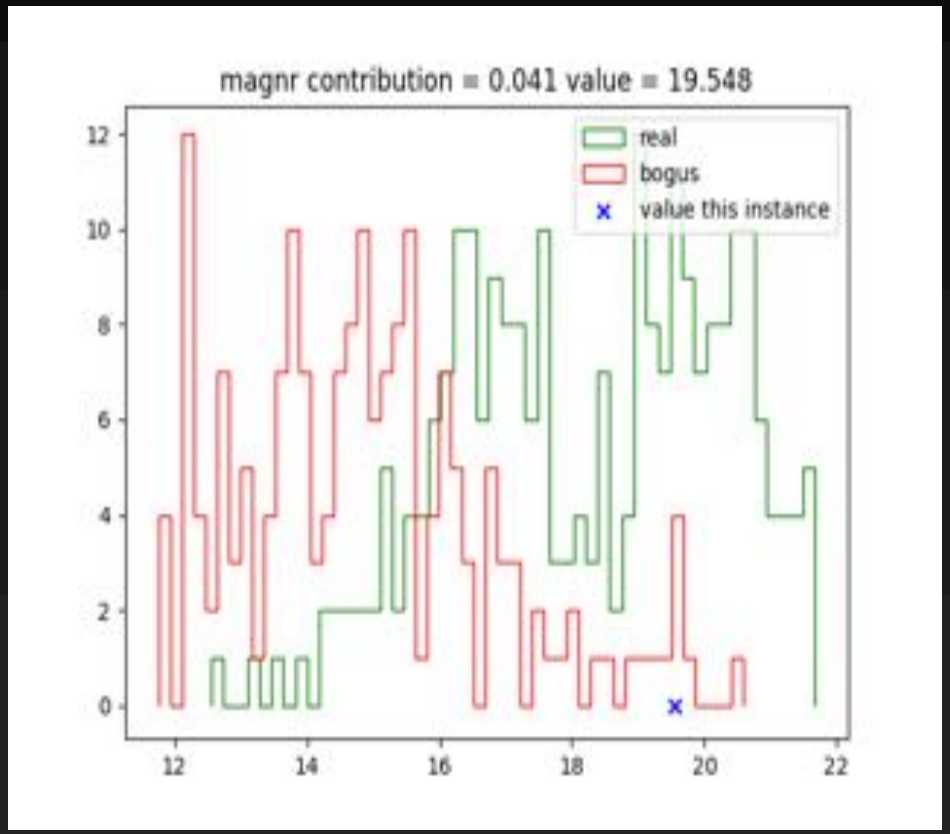
Biggest contributors to real classification by the model:

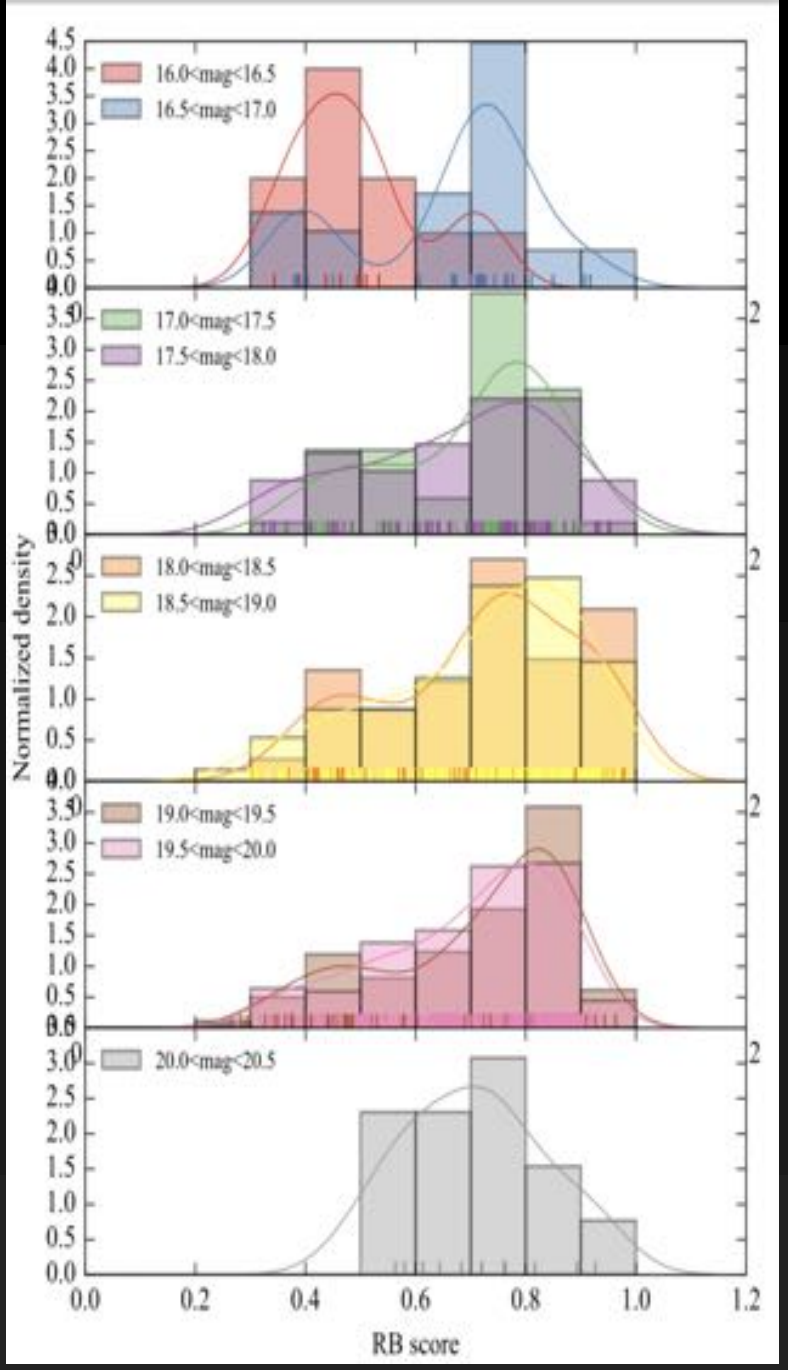
- **Magnr:** Magnitude of nearest reference image extraction
- **Sigmaapbig:** 1-sigma uncertainty in magapbig [mag] (which is magnitude from “big” aperture photometry)
- **Aimage:** Windowed RMS along major axis of source profile (pixels)
- **Magpsf:** Magnitude from PSF fit (mag)
- **Chipsf:** Chi of candidate

Magnr is often a big contributor to the RB scores.

It has a high importance value.

We see that fainter objects are given higher RB scores →





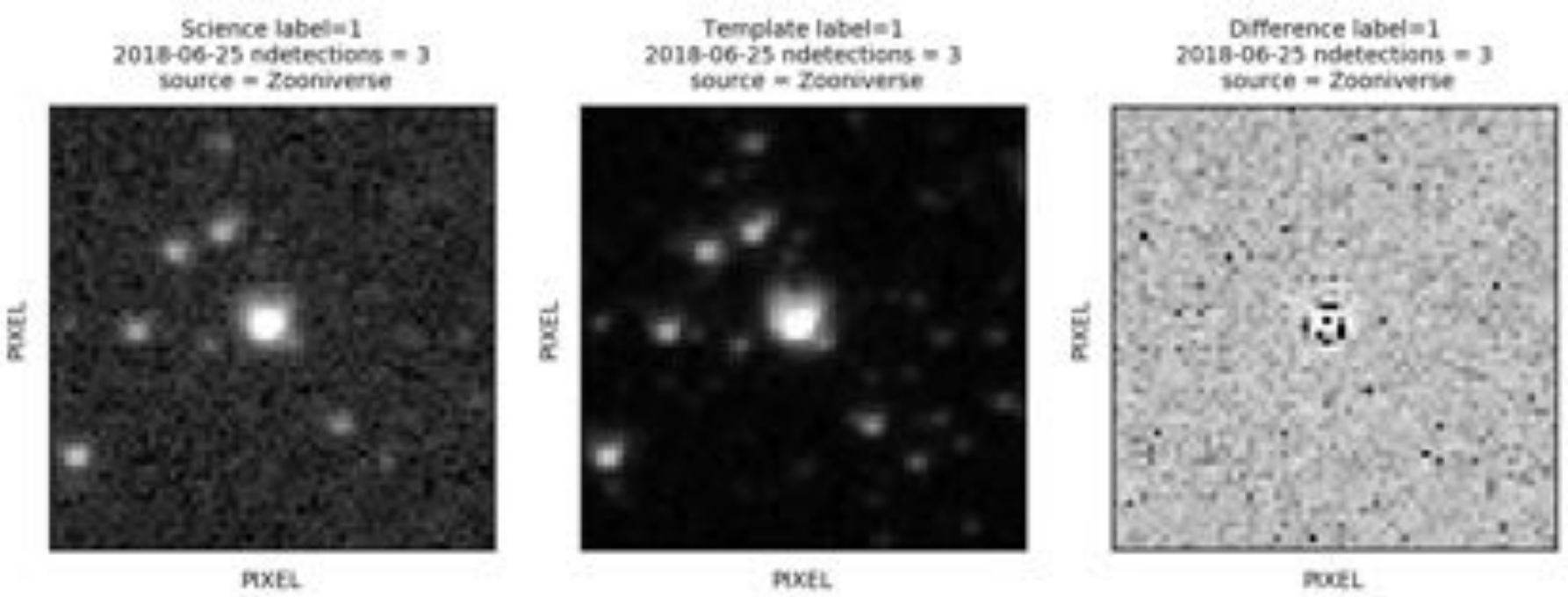
Finding Mislabeled Examples

- Of 386 randomly selected training set examples classified as **real**, 76 (**20%**) were misclassified.
- **56%** of these misclassified sources were from the marshal

- Of 575 randomly selected training set examples classified as **bogus**, 32 (**5.6%**) were misclassified.
- **45%** of these misclassified sources were from the **marshal**

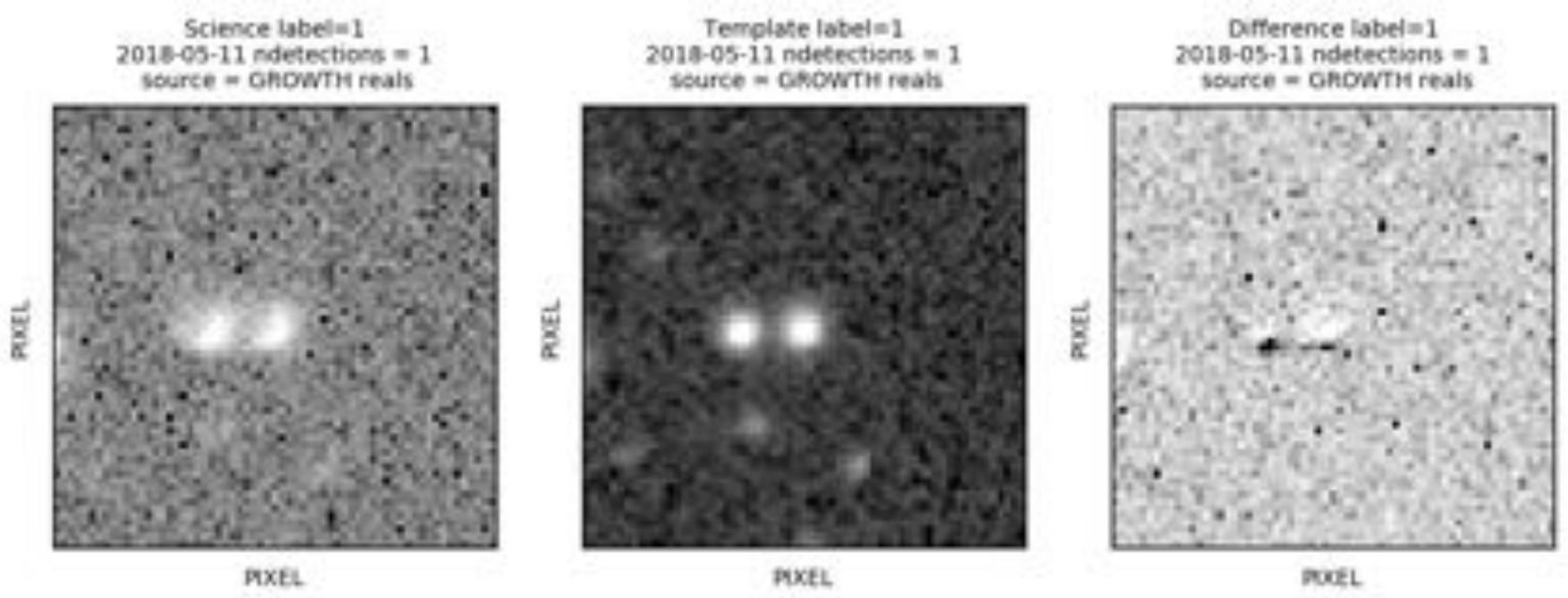
Finding Mislabeled Examples

- Examples of bogus that were classified as real: bad subtractions



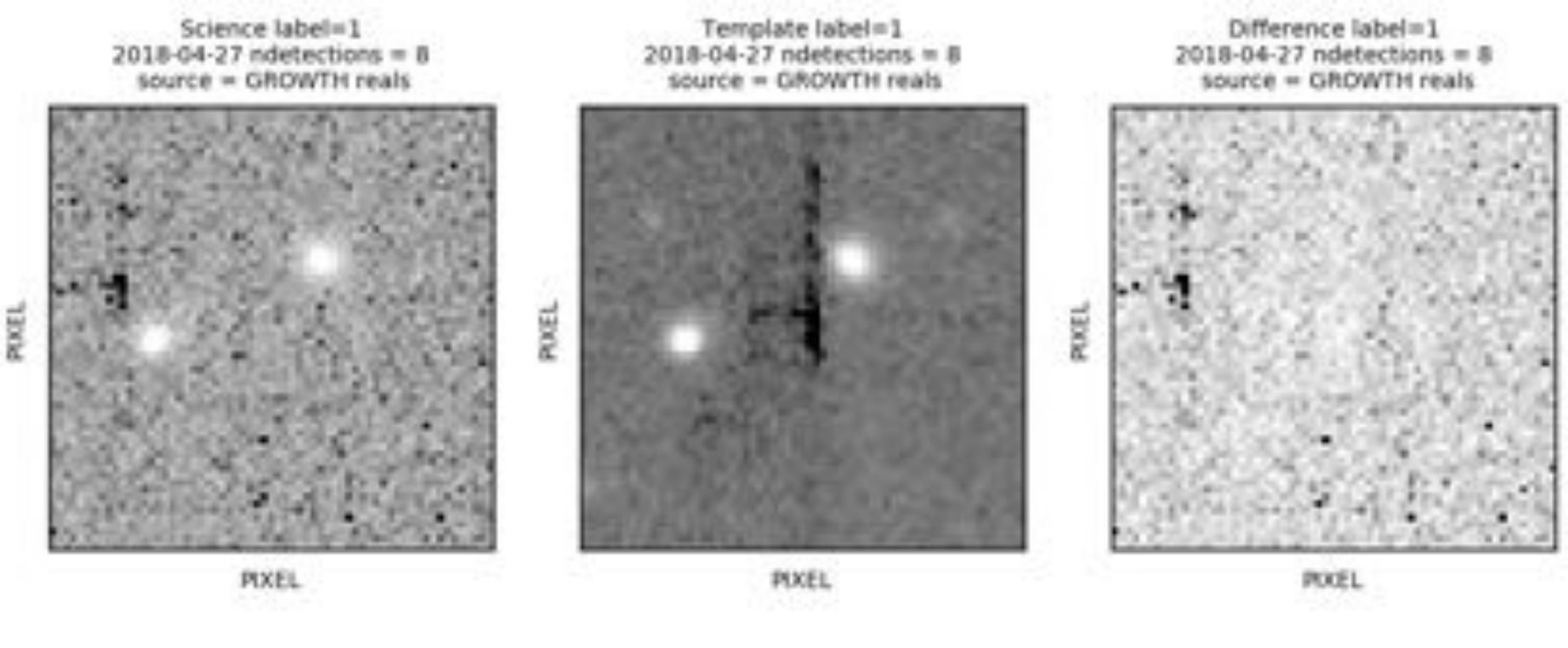
Finding Mislabeled Examples

- Examples of bogus that were classified as real: PSF differences causing apparent transients



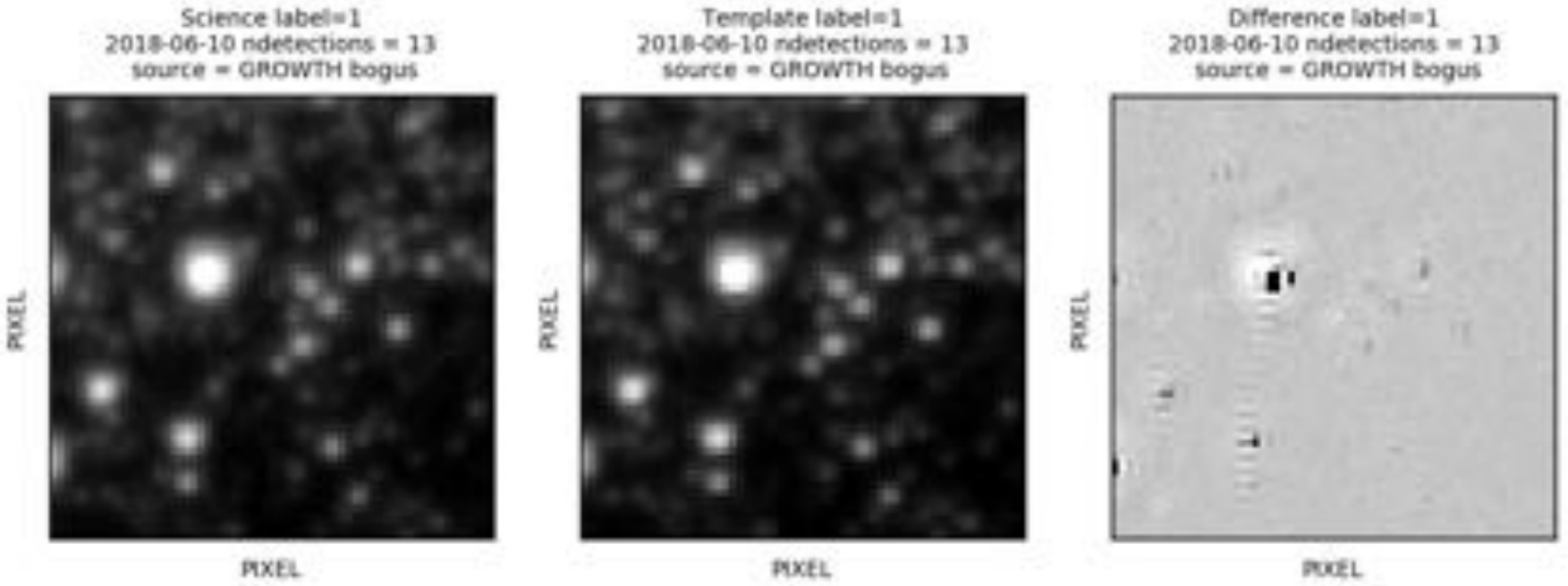
Finding Mislabeled Examples

- Examples of bogus that were classified as real: ghost-like artifacts



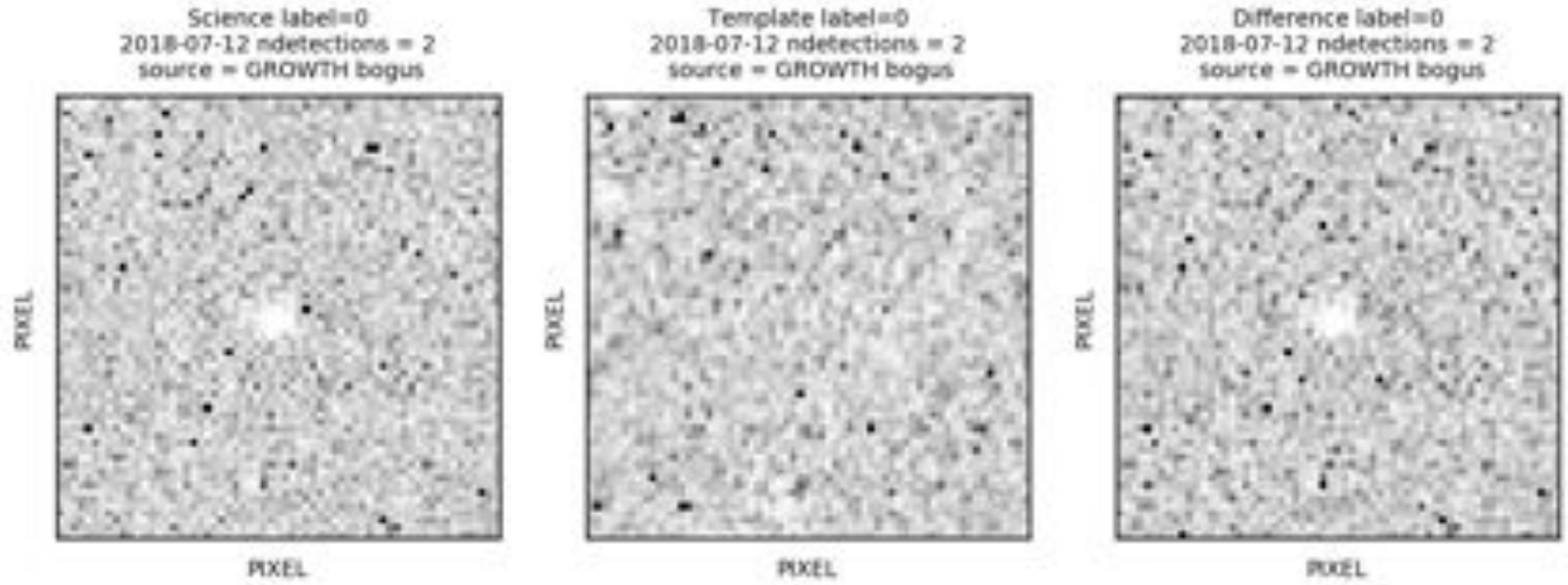
Finding Mislabeled Examples

- Examples of bogus that were classified as real: bad galactic plane subtractions:



Finding Mislabeled Examples

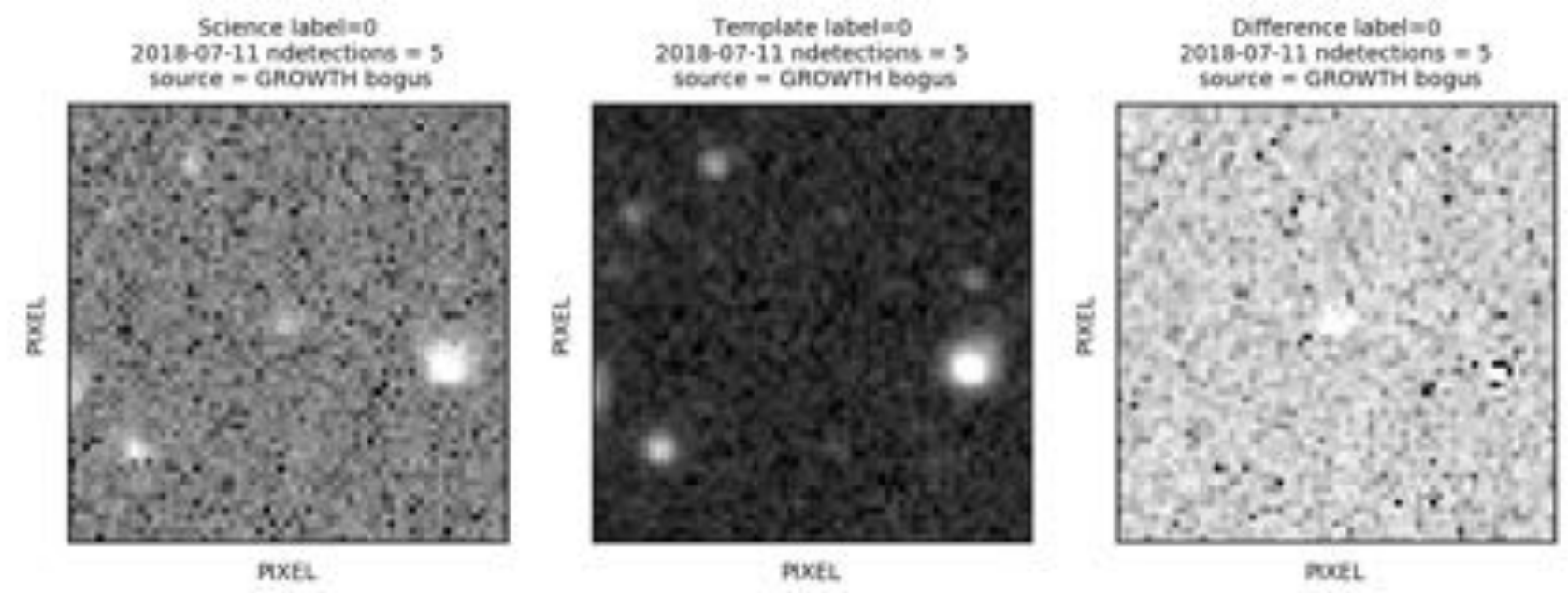
- Example of real that was classified as bogus:



Finding Mislabeled Examples

Analysis of repeatedly detected candidates to find reals incorrectly classified as bogus:

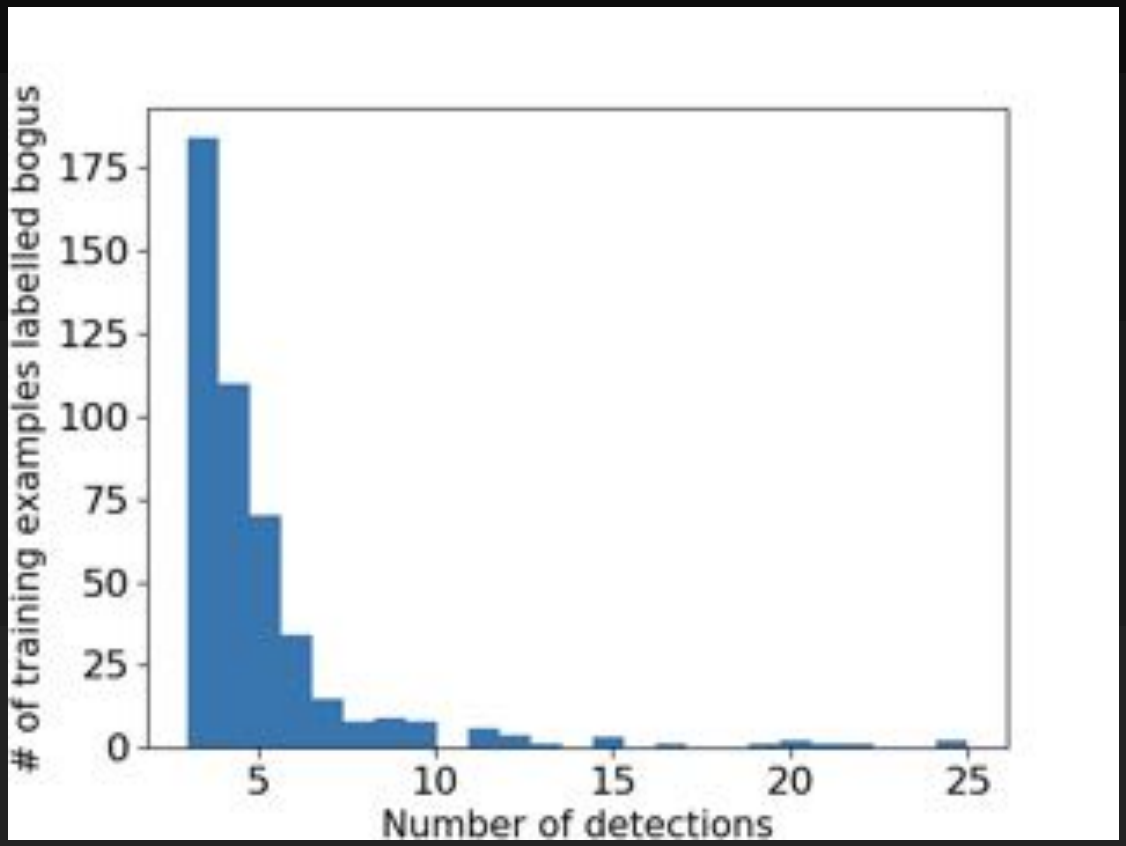
- A training set subset of 1200 examples labelled as bogus was examined.
- 32.5% of this subset have more than 2 observations.
- Approximately half of these which have more than 2 observations look to be real, e.g:



Finding Mislabeled Examples

Analysis of repeatedly detected candidates to find incorrectly classified variable stars:

- The histogram of the number of detections for sources labelled as bogus with n>2:



Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov