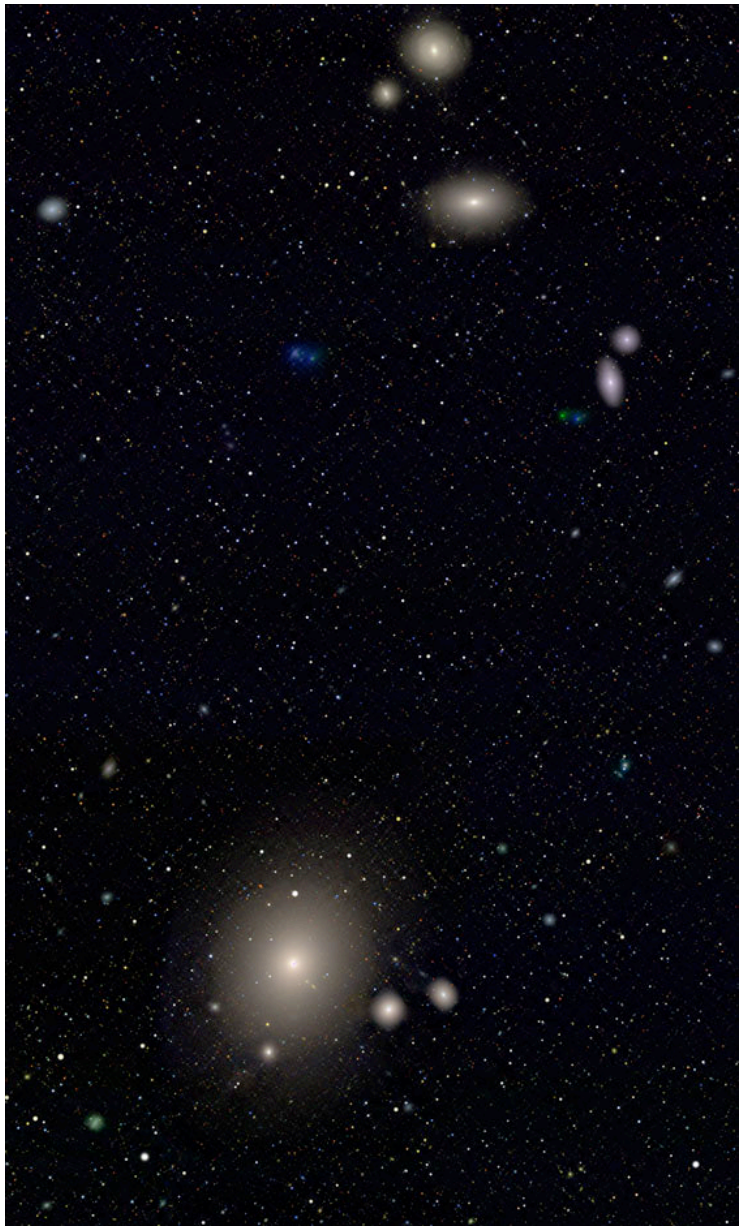
# Objects Classification in Synoptic Sky Surveys: contextual and external information

Ciro Donalek

Kiss Workshop

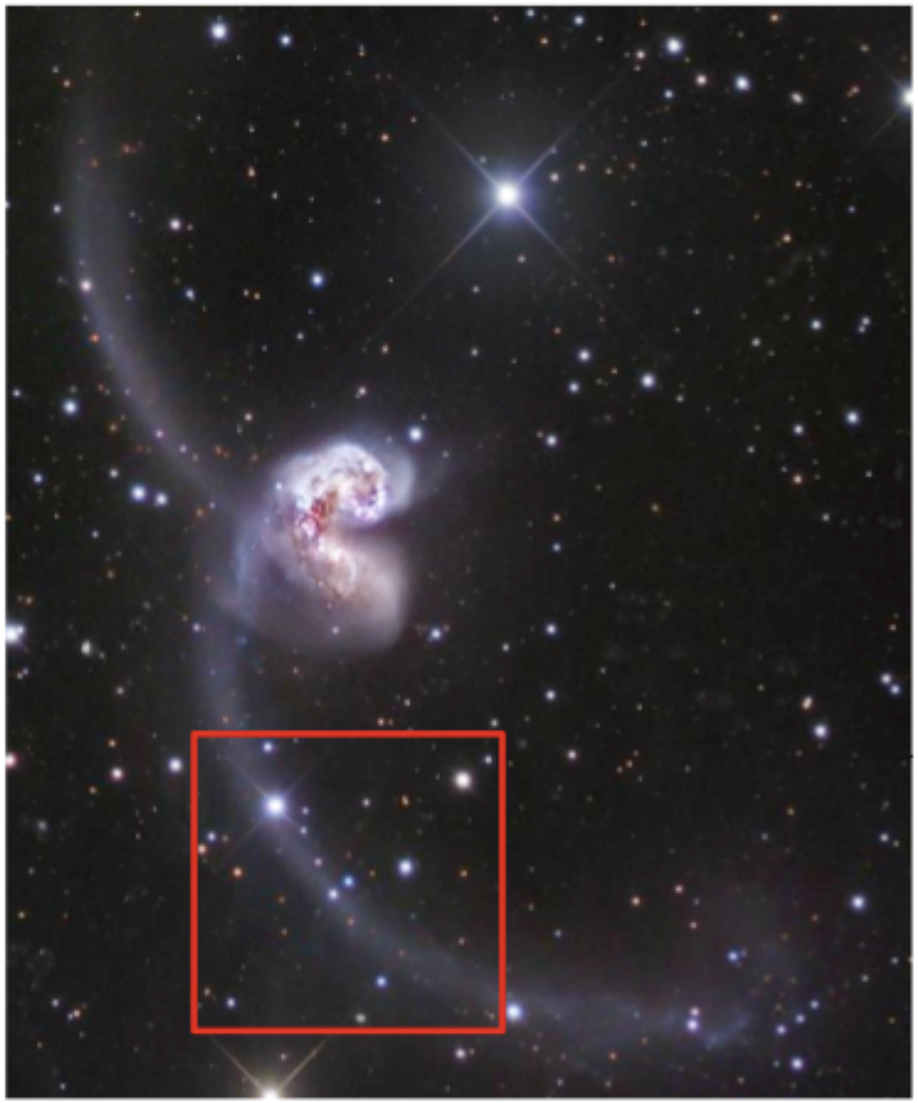
# Summary

- Using external and contextual knowledge
  - Citizen Science
  - Artifact Filter in PQ
  - Improve the S/G classification in multipass surveys



# Contextual Information and CitSci

- Some of the relevant information is contextual, and easily recognizable by humans looking at images, but it is very hard to encode in the data pipelines.
- Crowdsourcing (aka Citizen Science) provides one possible way to gather such information.
- **But…**



# Humans and Machine Working Together

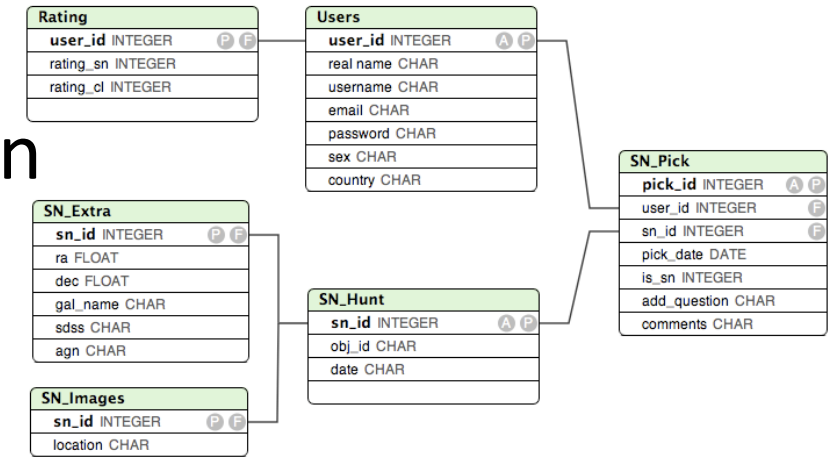
- **…**the scale of the datasets that can be attacked using citizen science today will soon grow far beyond all available human time and attention (LSST ~ $10^5$ candidate transients/night).

- **The goal**: use the work and decision process of human participants to train well-defined machine learning algorithms to be used in automating such data analysis in the future.

# SkyDiscovery.org

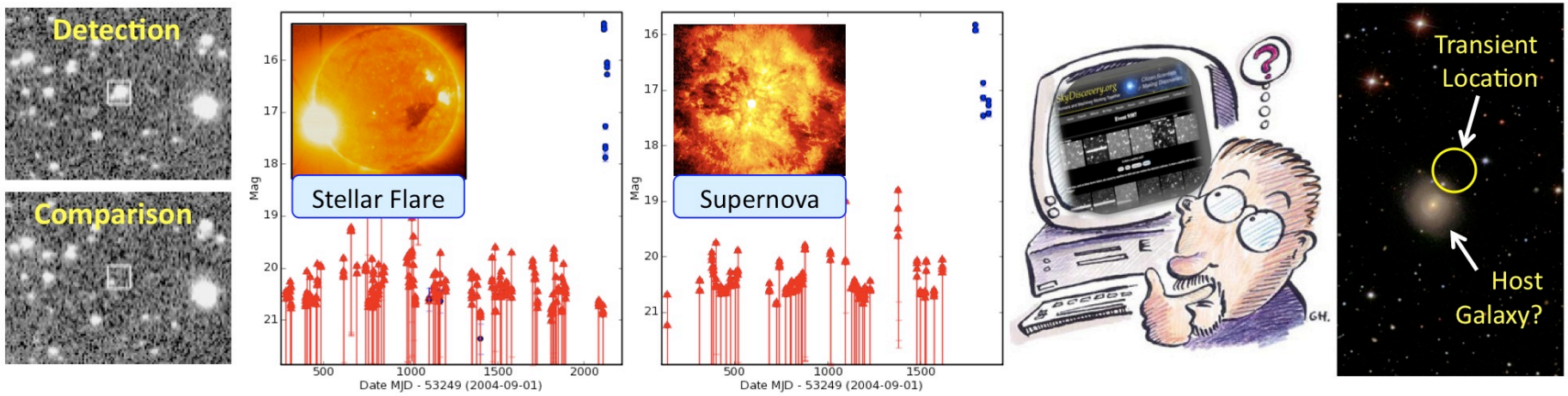
- SkyDiscovery.org is a website that allows experts and citizen science enthusiasts to work together and share information in a collaborative scientific discovery environment.

- Building the Community
- Modularity: new projects can be easily plugged in
- Multilevel Approach



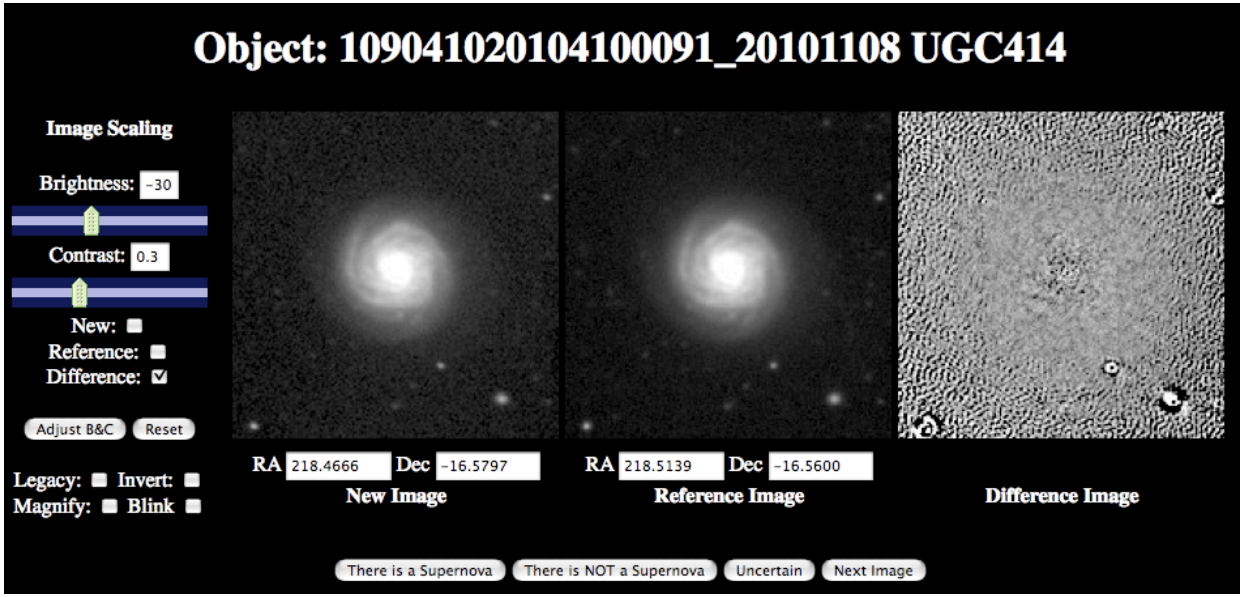
# Multilevel Approach

- A **multi level approach** allows the complexity of the interface to be tailored to the expertise level of the user.
- An entry level user can just review images and validate events as being real, while a more advanced user would be able to interact with the data associated to an event.



# First Project: The Great Supernova Hunt

- Users are requested to look for new objects appearing on images of galaxies taken by CRTS, in order to find all the supernovae occurring in nearby bright galaxies.
- Images served alongside with other tools that can help the discovery.
- **Reward system: users are listed as official discoverer of any supernovae that they report**, provided that we can confirm that they are real and they not already known.



# Results from CRTS

**New SNe from CRTS**

A.J. Drake, S.G. Djorgovski, A.A. Mahabal, M.J. Graham, R. Williams, C. Donalek (Caltech); J. Prieto (Carnegie Obs); M. Catelan (PUC); E. Christensen (Gemini Obs); E.C. Beshore, S.M. Larson (LPL/UA); R.H. McNaught (ANU).

subjects: Optical, Transients, Supernova.

Further to ATel#3340, we report the CRTS discovery of 27 new supernova candidates found between 2011-05-07 and 2011-06-08 UT:

| CRTS ID | Disc. Date | RA | Dec | Disc. Mag |
|---|---|---|---|---|
| CSS110608:214804+043448 | 2011-06-08 | 21:48:03.56 | 04:34:47.7 | 18.5 |
| MLS110607:110336+093819 | 2011-06-07 | 11:03:35.64 | 09:38:19.5 | 20.5 |
| MLS110607:140351-113108 | 2011-06-07 | 14:03:50.74 | -11:31:07.8 | 20.6 |
| CSS110607:113123-075009 | 2011-06-07 | 11:31:22.59 | -07:50:08.9 | 19.4 |
| CSS110607:111706-020617 | 2011-06-07 | 11:17:05.54 | -02:06:16.7 | 18.6 |
| CSS110606:140915-011055 | 2011-06-06 | 14:09:15.29 | -01:10:54.7 | 18.5 |
| MLS110605:105619+104444 | 2011-06-05 | 10:56:19.18 | 10:44:44.4 | 20.2 |
| CSS110604:155919-074418 | 2011-06-04 | 15:59:19.32 | -07:44:18.0 | 17.5 |
| CSS110604:130707-011044 | 2011-06-04 | 13:07:06.69 | -01:10:44.0 | 15.7☆ |
| MLS110604:095908+115727 | 2011-06-04 | 09:59:07.86 | 11:57:26.9 | 19.5 |
| MLS110603:135215-001421 | 2011-06-03 | 13:52:15.26 | -00:14:20.7 | 20.7 |
| CSS110602:153001+245229 | 2011-06-02 | 15:30:01.40 | 24:52:29.1 | 18.9 |
| CSS110527:135427+305803 | 2011-05-27 | 13:54:27.05 | 30:58:02.7 | 18.8 |
| MLS110526:094440+135936 | 2011-05-26 | 09:44:39.54 | 13:59:35.7 | 20.1 |
| MLS110526:093516+131102 | 2011-05-26 | 09:35:16.17 | 13:11:01.8 | 20.1 |
| MLS110526:104421+082012 | 2011-05-26 | 10:44:21.07 | 08:20:11.7 | 20.5 |
| MLS110526:153245-131910 | 2011-05-26 | 15:32:45.32 | -13:19:10.3 | 20.0 |
| CSS110525:150213+231553 | 2011-05-25 | 15:02:13.26 | 23:15:53.1 | 19.5 |
| CSS110525:154951+144930 | 2011-05-25 | 15:49:50.65 | 14:49:30.1 | 18.9 |
| CSS110525:140118-112359 | 2011-05-25 | 14:01:18.22 | -11:23:59.5 | 18.1 |
| CSS110525:205007-035021 | 2011-05-25 | 20:50:06.92 | -03:50:21.3 | 18.0 |
| MLS110525:134153+011557 | 2011-05-25 | 13:41:53.34 | 01:15:56.8 | 20.2 |
| MLS110525:134316+004749 | 2011-05-25 | 13:43:15.88 | 00:47:49.3 | 20.2 |
| CSS110512:141256+290215 | 2011-05-12 | 14:12:55.82 | 29:02:15.4 | 18.9 |
| MLS110512:151647-072202 | 2011-05-12 | 15:16:46.80 | -07:22:02.1 | 18.9☆ |
| MLS110508:083817+150321 | 2011-05-08 | 08:38:16.56 | 15:03:21.1 | 20.2 |
| CSS110507:131750-025717 | 2011-05-07 | 13:17:50.03 | -02:57:16.7 | 18.9 |

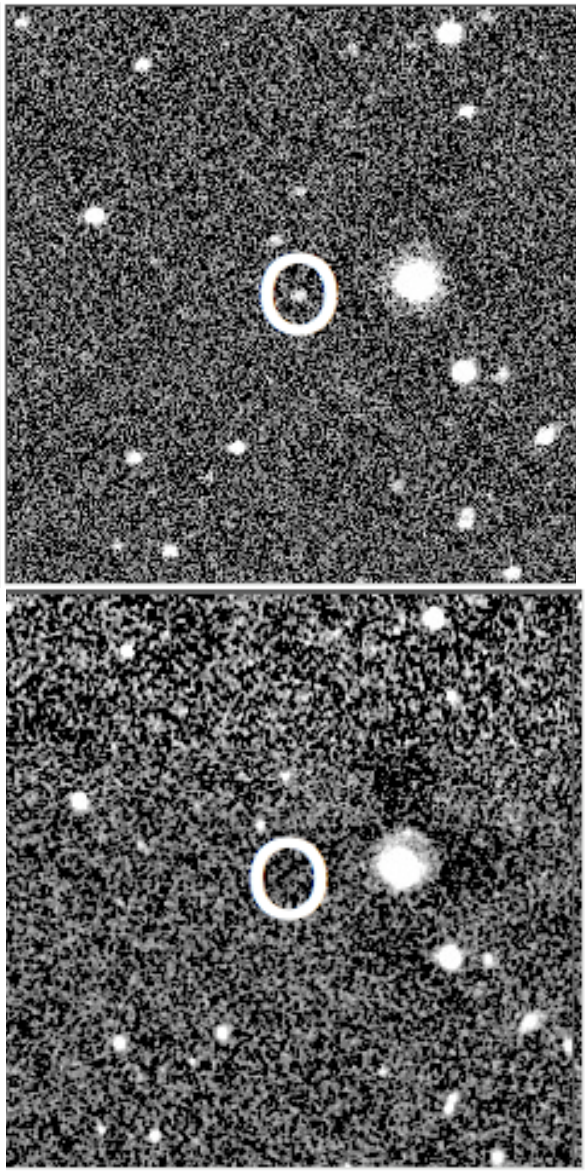
Notes:
☆Events announced on CBAT TOCP.
Finding charts for these events can be found at
http://voeventnet.caltech.edu/feeds/ATEL/CRTS

During this period eleven additional supernova candidates were discovered in the CRTS SN Hunt by S. Howerton. Additionally, SN 2011db and SN 2011dc were discovered by CRTS and confirmed by Tomasella et al. (2011, CBET#2725; CBET#2726).

**A.J. Drake, 6/8/2011 (Caltech)**

# An application: artifact removal

To perform a reliable real time object classification, there is also a need for an effective classification and removal of spurious objects that pop up as false positives.

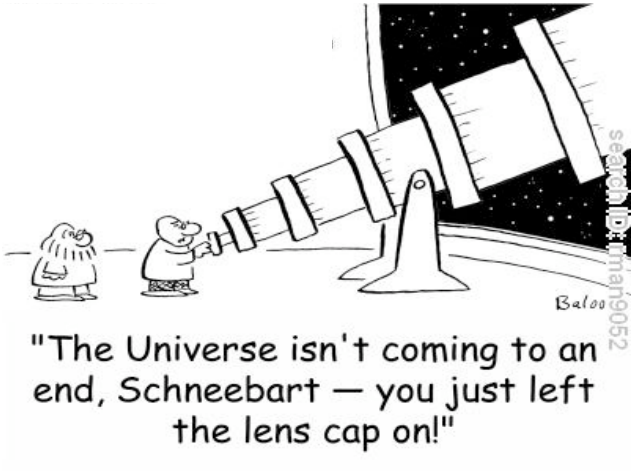


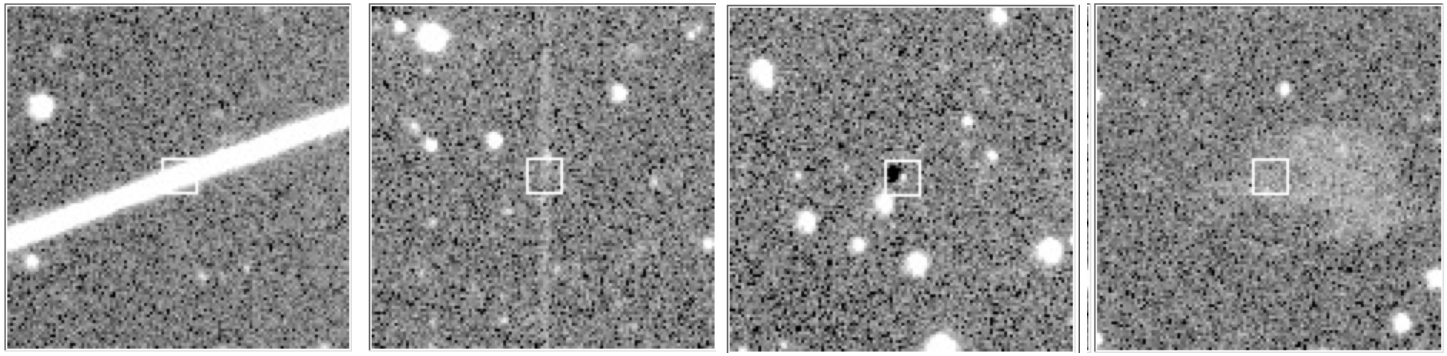
Transient search:

- compare new images with the baseline
- minimize false positives
  - remove asteroids, cosmic rays...
  - **remove data artifacts**
    - artifacts vs real objects classification
- transient classification

# Types of artifacts

Artifacts can appear in the images for many reasons: saturations, edge-effects, reflections, problems with the electronic...



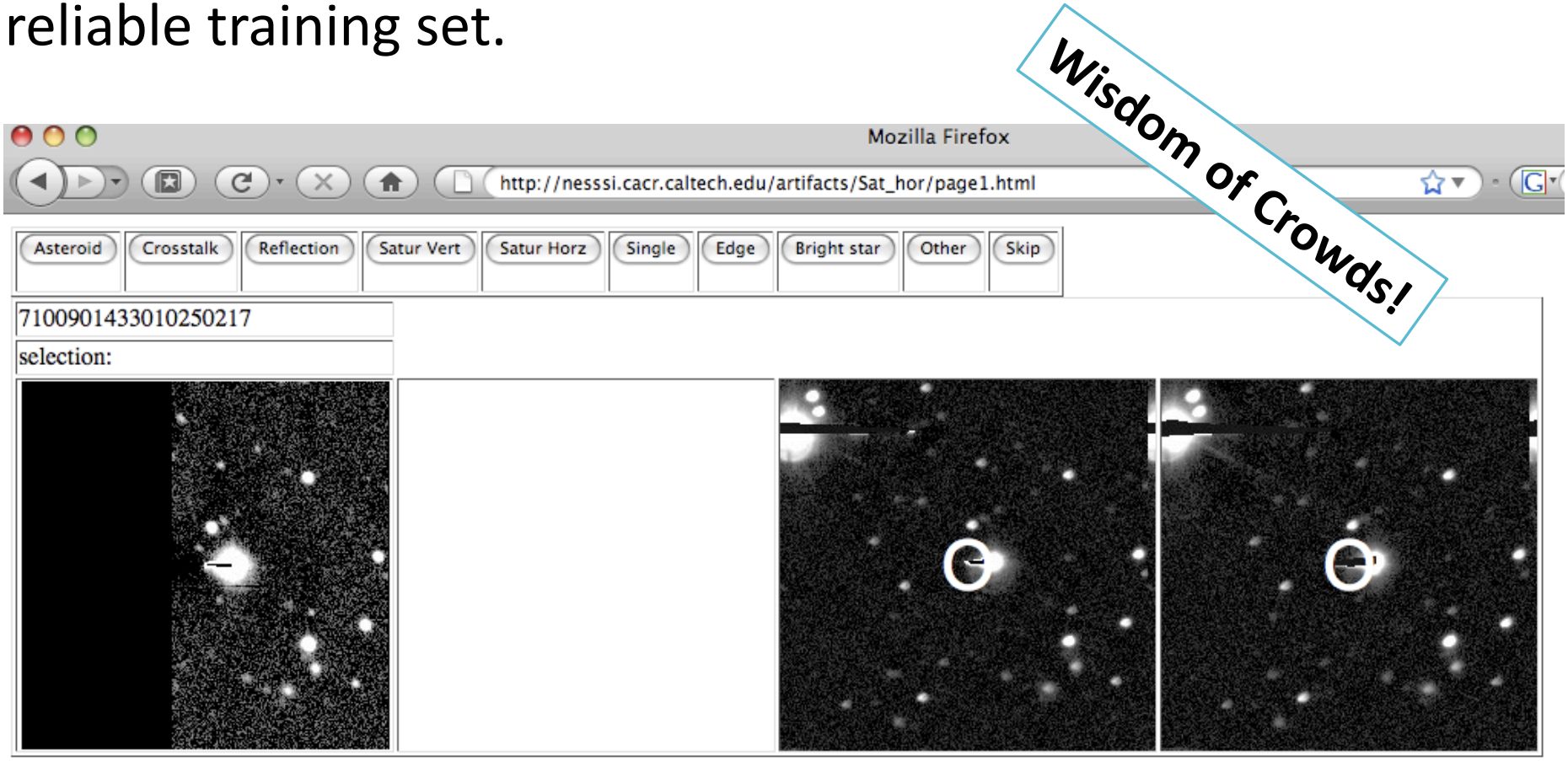



In the PQ Survey this problem has been addressed using a **supervised learning approach** in order to build a classifier that discriminates between artificial and real objects.

# Building the BoK

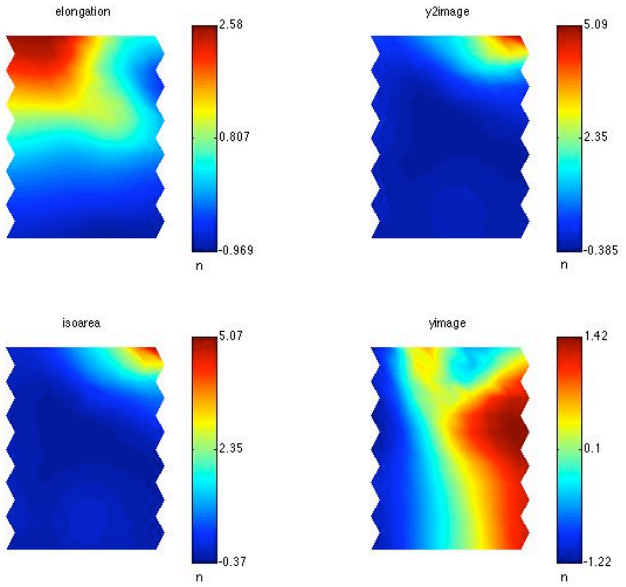
Setup a project with images of the artifacts found in our previous scans.

Visual classification of all the candidates in order to build a reliable training set.

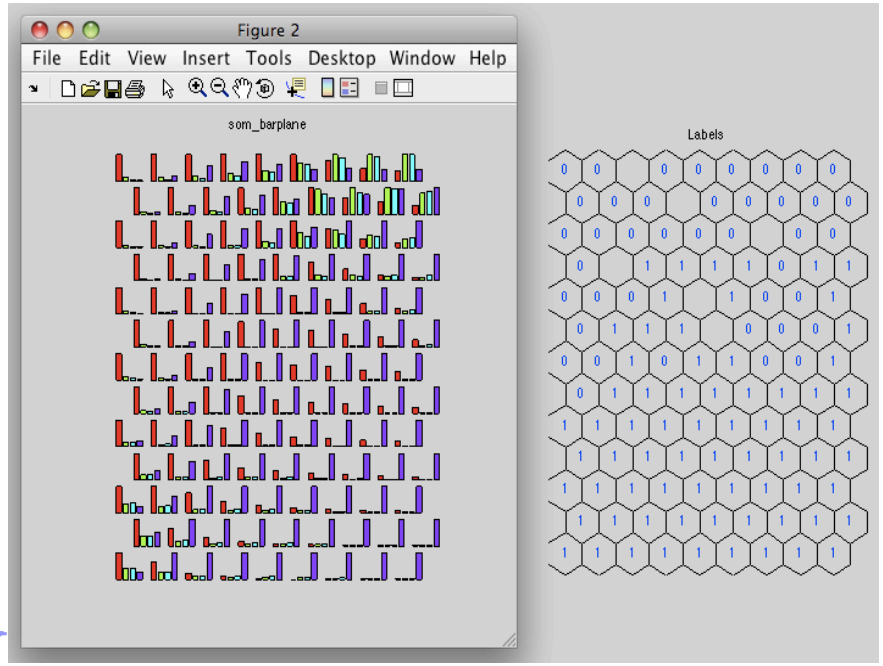


# Exploring the parameter space

List of parameters available through the pipeline
# 00 OBJID ID of detection
# 000 FRAMEID ID of frame
# 1 MAG_APER Fixed aperture magnitude vector [mag]
# 6 MAGERR_APER RMS error vector for fixed aperture mag. [mag]
# 11 MAG_BEST Best of MAG_AUTO and MAG_ISOCOR [mag]
# 12 MAGERR_BEST RMS error for MAG_BEST [mag]
# 13 MAG_ISO Isophotal magnitude [mag]
# 14 MAGERR_ISO RMS error for isophotal magnitude [mag]
# 15 MAG_ISOCOR Corrected isophotal magnitude [mag]
# 16 MAGERR_ISOCOR RMS error for corrected isophotal magnitude [mag]
# 17 MAG_AUTO Kron-like elliptical aperture magnitude [mag]
# 18 MAGERR_AUTO RMS error for AUTO magnitude [mag]
# 19 FWHM_IMAGE FWHM assuming a gaussian core [pixel]
# 20 FLAGS Extraction flags
# 21 FLUX_MAX Peak flux above background [count]
**# 22 ELONGATION A_IMAGE/B_IMAGE**
# 23 CLASS_STAR S/G classifier output
# 24 X_IMAGE Object position along x [pixel]
**# 25 Y_IMAGE Object position along y [pixel]**
**# 26 X2_IMAGE Variance along x [pixel**2]**
# 27 ERRX2_IMAGE Variance of position along x [pixel**2]
**# 28 Y2_IMAGE Variance along y [pixel**2]**
# 29 ERRY2_IMAGE Variance of position along y [pixel**2]
# 30 XY_IMAGE Covariance between x and y [pixel**2]
# 31 ERRXY_IMAGE Covariance of position between x and y [pixel**2]
**# 32 ISOAREA_IMAGE Isophotal area above Analysis threshold [pixel**2]**
# 33 THETA_IMAGE Position angle (CCW/x) [deg]
# 34 CONCENTRATION Abraham concentration parameter
# 35 BACKGROUND Background at centroid position [count]
# 36 NIMAFLAGS_ISO # of flagged pixels entering IMAFLAGS_ISO
# 37 IMAFLAGS_ISO FLAG-image flags OR'ed over the iso. profile
# 38 NEW_RA RA determined from new WCS [deg]
# 39 NEW_DEC DEC determined from new WCS [deg]
# 40 NEW_MAG Mag from match with USNO stars [mag]
# 41 HTM_ID HTM Id for RA and Dec location
# 42 CAR_X X coord for RA and Dec location
# 43 CAR_Y Y coord for RA and Dec location
# 44 CAR_Z Z coord for RA and Dec location



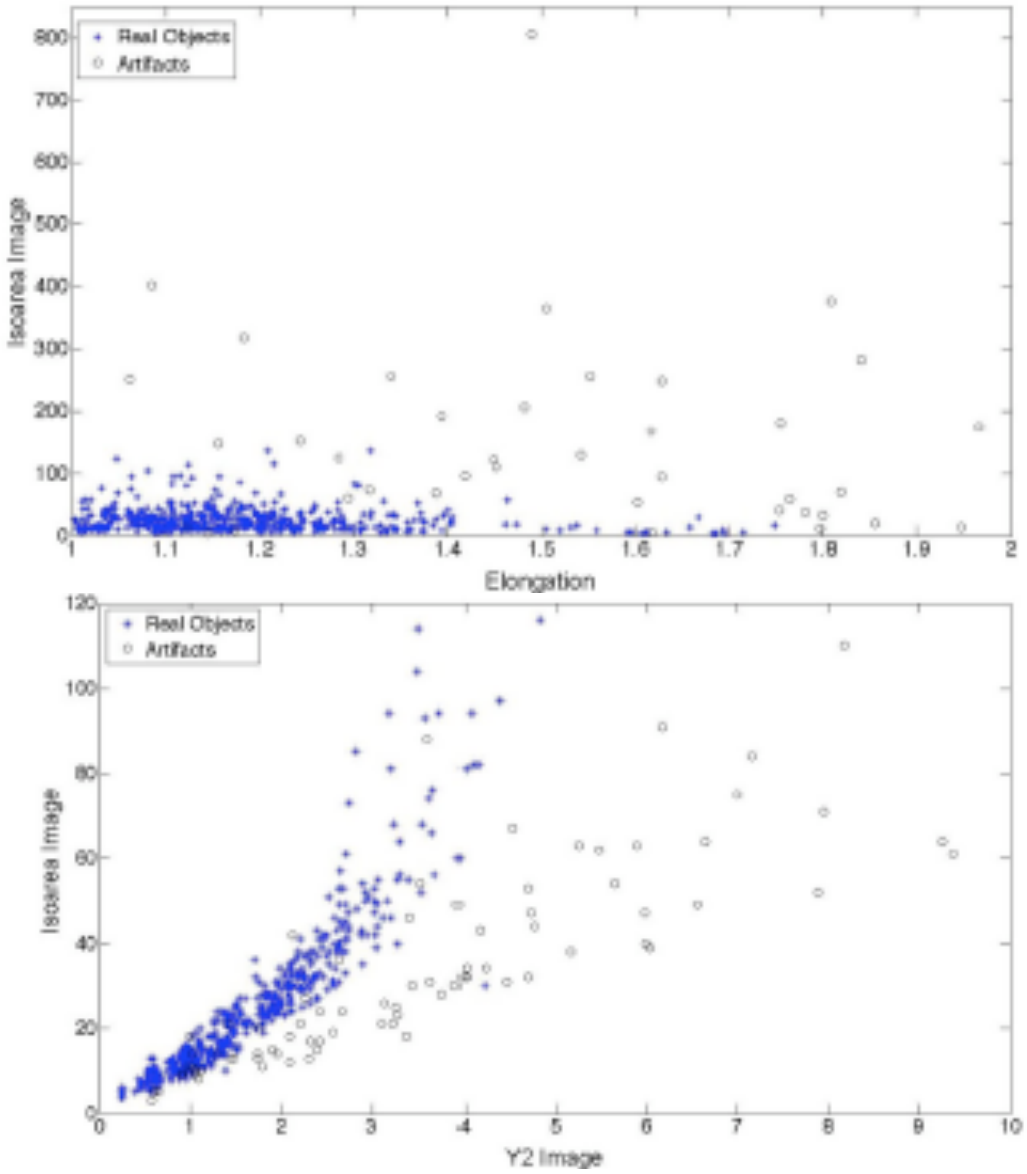

Hunting for correlations:
- SOM: N plots
- SP: N*(N-1)/2
- not very accurate
- easy to select interesting combos



# Artifacts and real objects

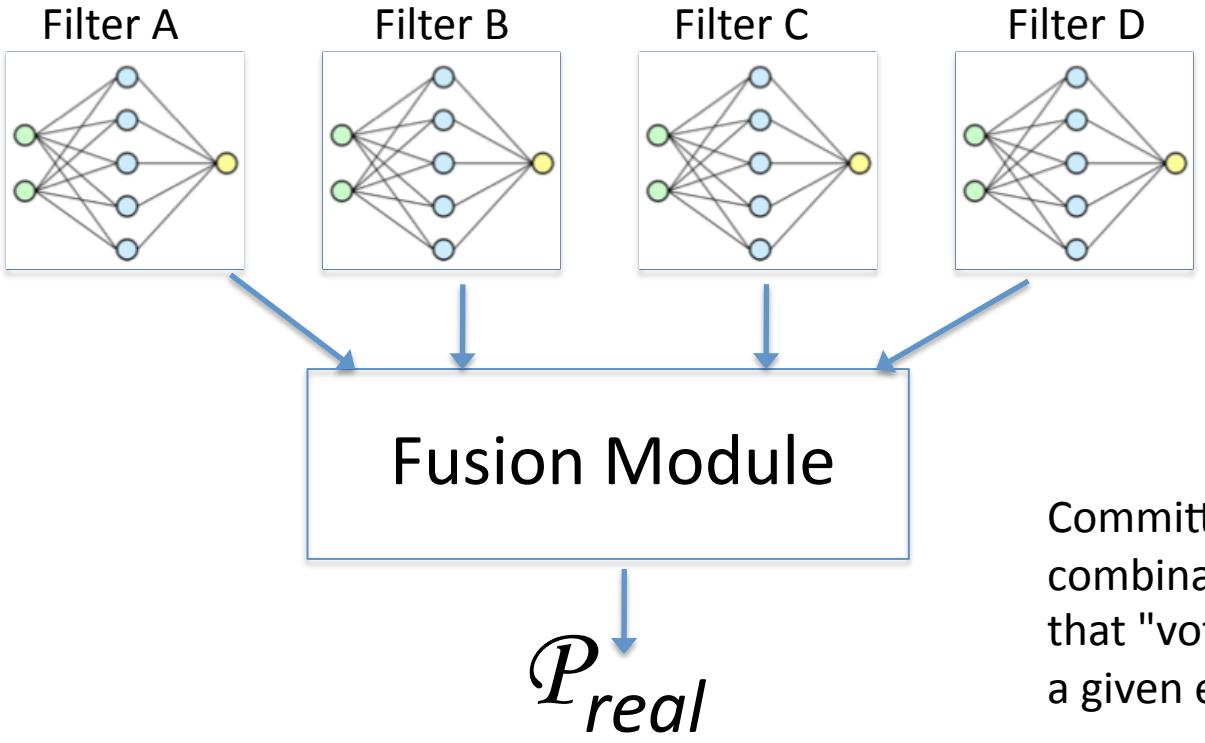
- The two plots show a couple of morphological parameter space projections, used to train the network, in which artifacts separate well from genuine objects.

- For each object, the classifier takes as input a set of parameters and returns the probability of it being a real object.



# How it works

- For each potential candidate we have up to 4 detections, one per filter
- They are fed separately to the NN and the outputs are combined to have the final classification.
- Each output can be interpreted as the conditional probabilities of each object to belong to the True-Objects Class or to the Artifacts class.
- We can use thresholds and decide to be more or less aggressive discarding or keeping objects (cost analysis).



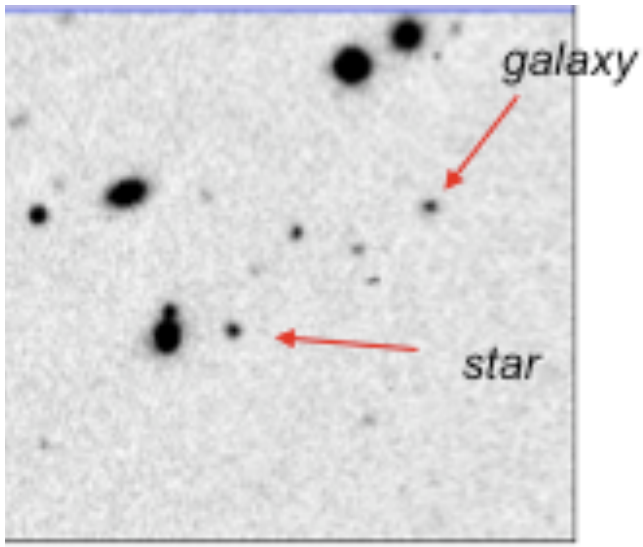

Committee Machines: combination of experts that "vote" together on a given example.

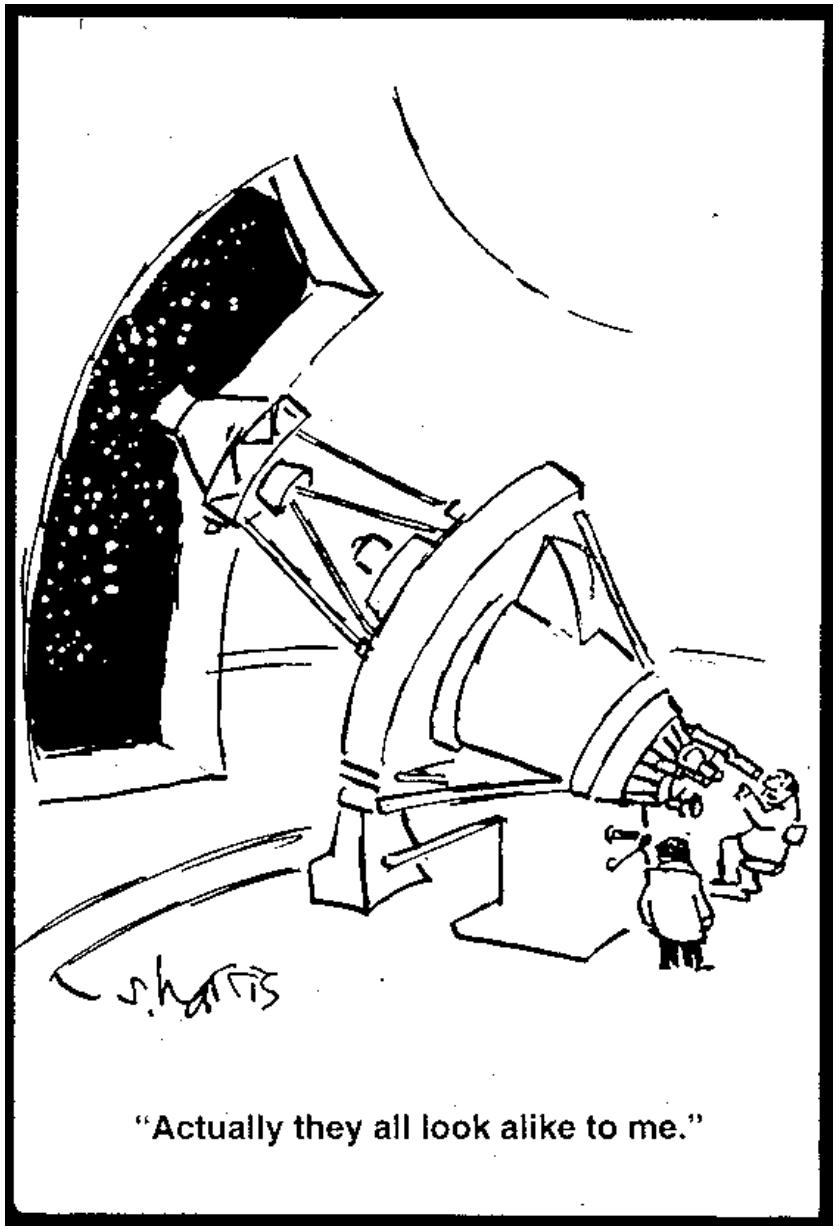
# Artifact Filter: conclusions and results

- It is an ANN-based classifier which separates real transient sources from a variety of spurious candidates.
- Despite the relatively low number of training cases for many kinds of artifacts, the overall artifact classification rate is around 90%, with no genuine transients misclassified during our real-time scans.
- **BoK built using crowdsourcing**.

# External Knowledge: an example

The second classification problem is related to the S/G classification that is a classical and crucial problem in the analysis of astronomical sky surveys.

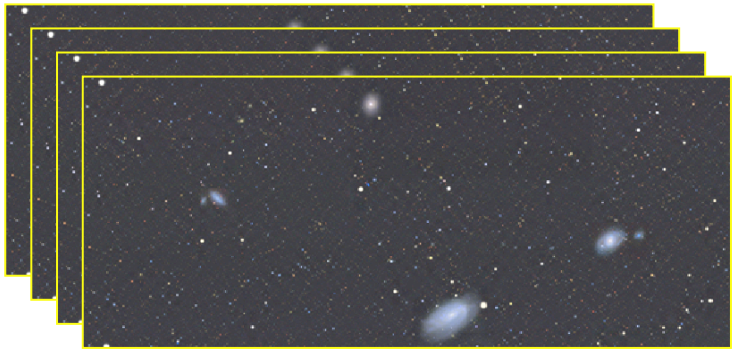


# Star-Galaxy classification in multipass survey

"*How do we assign an optimal star-galaxy classification in a multi-pass survey, where seeing and other external conditions change between different epochs, potentially leading to inconsistent classifications for the same object?*"

**Multiple imaging data sets**



Individually derived classifications $C_i$, $C_i$, ...

**Optimally combined imagery**

Classification $\langle C \rangle$

**Dataset dependent constraints**

**Optimal Classification**

**Context dependent constraints**

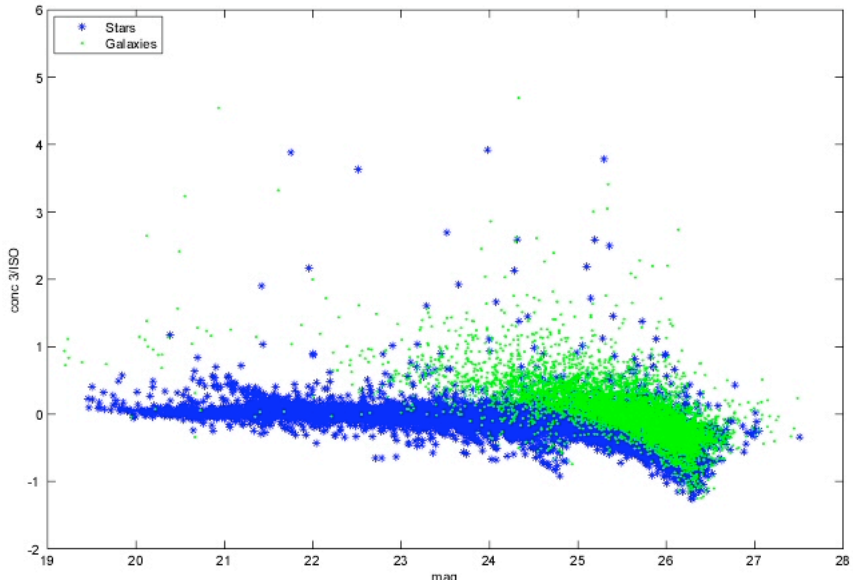
# The dataset

The catalog is built using parameters extracted from the images processed by the PQ pipeline. Objects detected in more than one pass are used to probe the correlation between the seeing and the goodness of the classification.

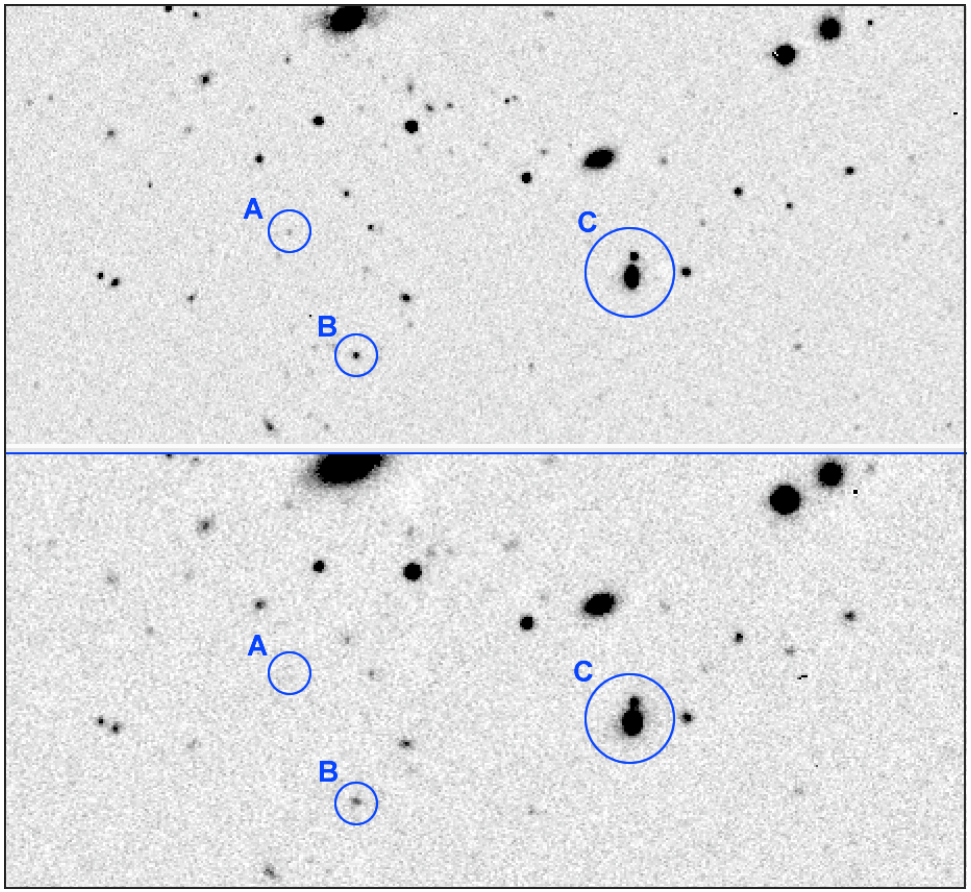
Overlap with SLOAN: gives the "true" class of each objects.

**Input**
**1. ra: right ascension**
**2. dec: declination**
**3. seeing**
**4. mag: magnitude**
**5. conc 2/4: 2" aperture magnitude – 4" aperture magnitude**
**6. conc 2/6**
**7. conc 2/iso: 2" aperture magnitude - isophotal.**
**8. conc 3/4**
**9. conc 3/6:**
**10. conc3/iso**
**11. objid num: unique object identifier**
**12. index arc: unique catalog identifier**

**Output**
**1. stellarity: probability of an object to be a star**



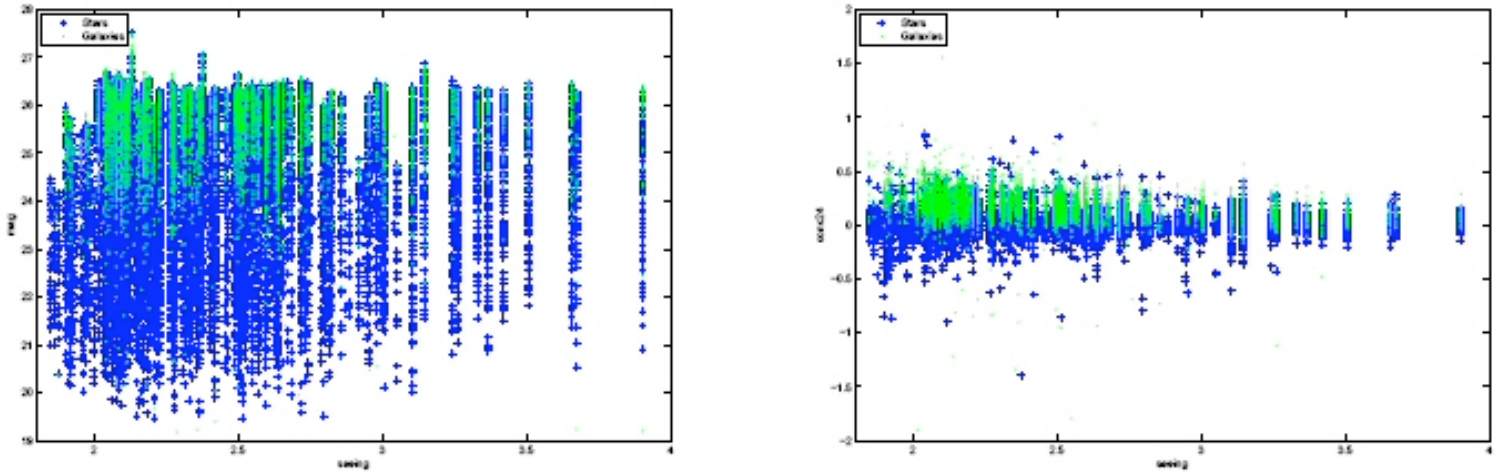
# External information: seeing




The main factor affecting the S/G classification in ground based optical data is the PSF of the images which is dominated by the **atmospheric seeing**. Upper figure: good seeing; lower figure: mediocre seeing. In the lower part many objects seem to be fuzzy, and thus potentially misclassified as galaxies (e.g. B and C) or vanishing (e.g. A).

# Including external knowledge among the parameters

Introducing the seeing directly as input to the network does not improve the network performance.



Seeing versus Magnitudes and Concentrations.

# Need of a priori information

In the specific case of S/G classification, the main a priori knowledge is the seeing.

| Set | σ < 2.1 | σ > 3 |
|---|---|---|
| σ < 2.1 | 94.1% | 83% |
| σ > 3 | 91.2% | 89% |

Classification rates for data with "good" (σ < 2.1) and "bad" (σ > 3) seeing.
If I train with good data and classify bad data I get bad results.

| Stars | Stellarity | Seeing |
|---|---|---|
| objid=1185 | 99.25% | 2.21 |
| objid=1185 | 97.98% | 2.28 |
| objid=1185 | 95.44% | 2.67 |
| objid=1185 | 82.79% | 3.67 |

How the same star is classified with different seeing conditions using a classifier without a priori knowledge.

The seeing actually affects the performances of the classifiers.

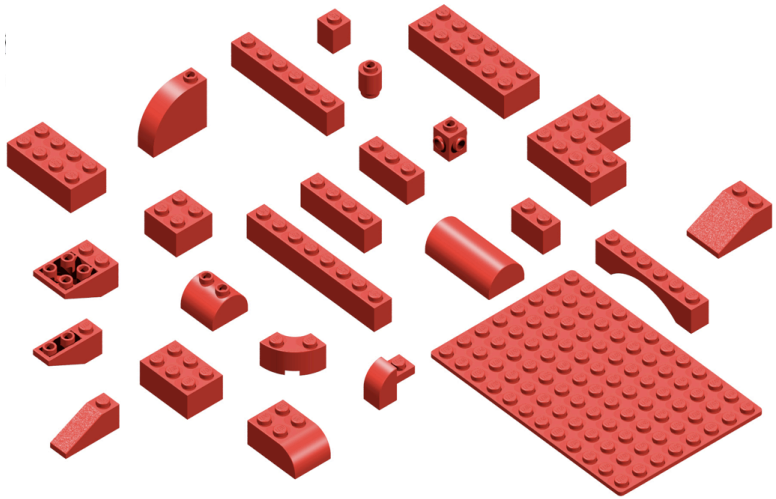
# Combining blocks: two approaches

It is often found that improved performance can be obtained by combining models together in some way, instead of using a single model in isolation.

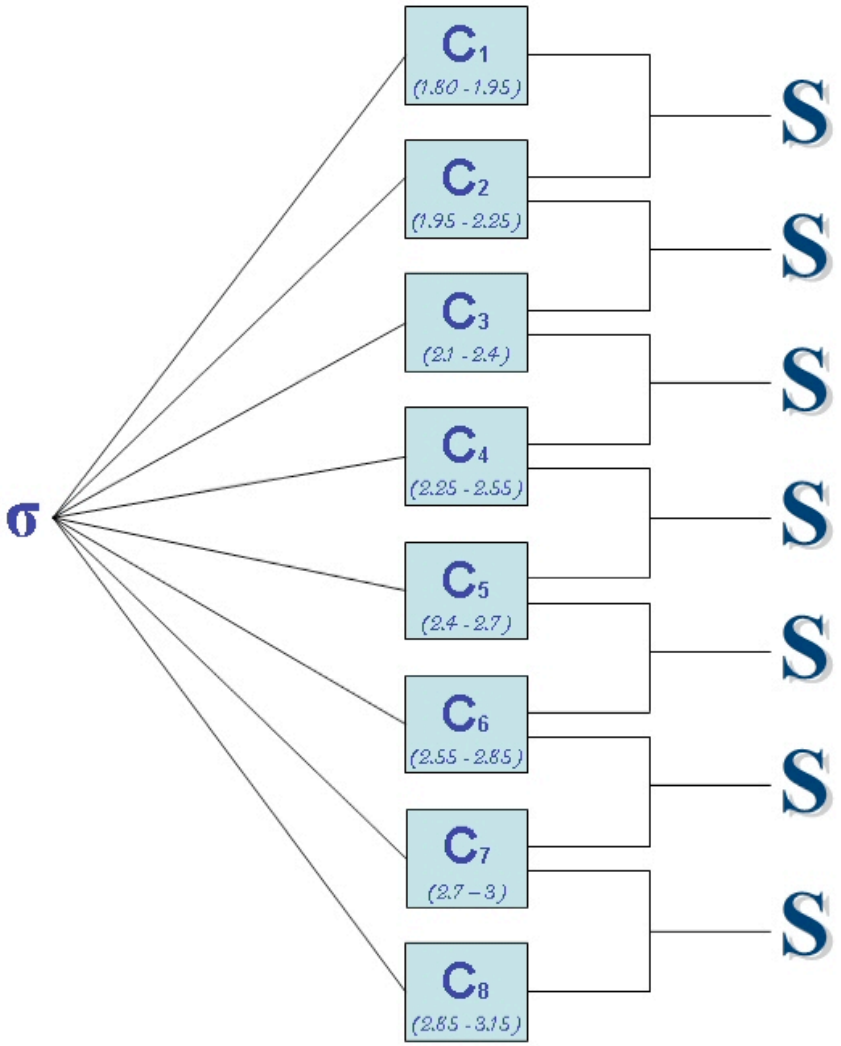
In this way, individual classifiers may be optimized or trained differently.

An alternative of model combinations is to select one of the models to make the classification, and let the choice of the model be driven by an input parameter or by an “external” knowledge.

In this way different models become responsible for making predictions in different regions of the input space.



# CT with overlap




Simple case: predictions are made using the average of the predictions made by the two classifiers in which each sample falls.

# CT with overlap: results

Using a classification tree with overlap and a stellarity threshold > 0.90, the star contamination is always very low, even at higher seeing.

At lower values of seeing ( < 2.7), the contamination is low for all the models, but we note an higher completeness for the classification trees.

**Table 5.8**: Star completeness and contamination with stellarity threshold = 0.90.

| | C1 | CT | CT ov. | C1 | CT | CT ov. |
|---|---|---|---|---|---|---|
| seeing | Compl. | Compl. | Compl. | Cont. | Cont. | Cont. |
| $1.80 < \sigma < 1.95$ | 82.16% | 89.58% | 89.58% | 1.95% | 1.79% | 1.79% |
| $1.95 < \sigma < 2.10$ | 82.68% | 87.83% | 86.46% | 1.45% | 1.50% | 1.50% |
| $2.10 < \sigma < 2.25$ | 86.83% | 88.49% | 90.54% | 1.10% | 1.14% | 1.12% |
| $2.25 < \sigma < 2.40$ | 80.10% | 82.10% | 83.83% | 1.46% | 1.55% | 1.33% |
| $2.40 < \sigma < 2.55$ | 80.86% | 82.62% | 84.80% | 1.53% | 1.42% | 1.40% |
| $2.55 < \sigma < 2.70$ | 81.50% | 82.00% | 82.50% | 2.36% | 2.27% | 2.10% |
| $2.70 < \sigma < 2.85$ | 81.60% | 81.00% | 81.50% | 2.47% | 2.11% | 1.91% |
| $2.85 < \sigma < 3.00$ | 82.50% | 81.14% | 77.44% | 3.50% | 3.20% | 1.70% |
| $3.00 < \sigma < 3.15$ | 71.50% | 64.33% | 63.36% | 6.4% | 2.38% | 2.02% |

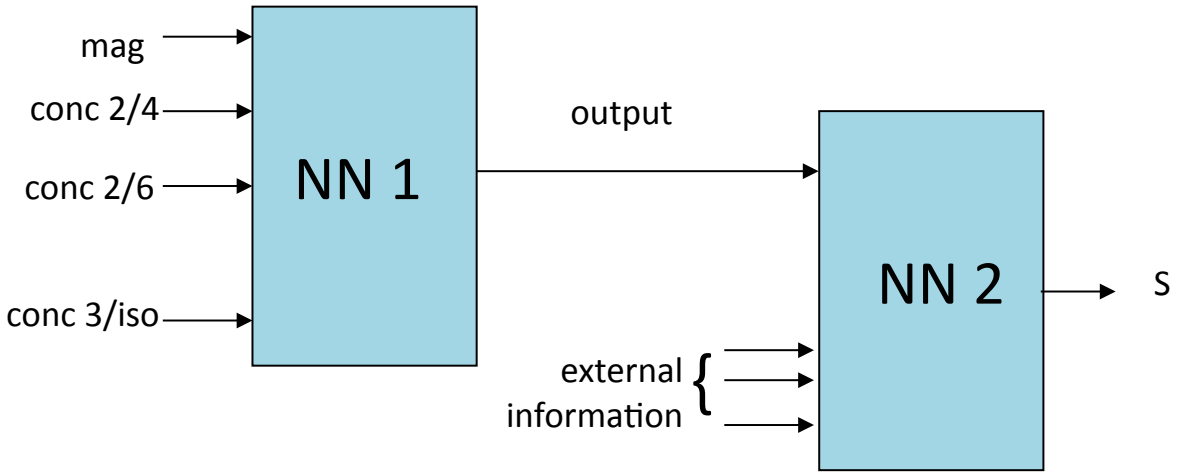
# CT with overlap: results

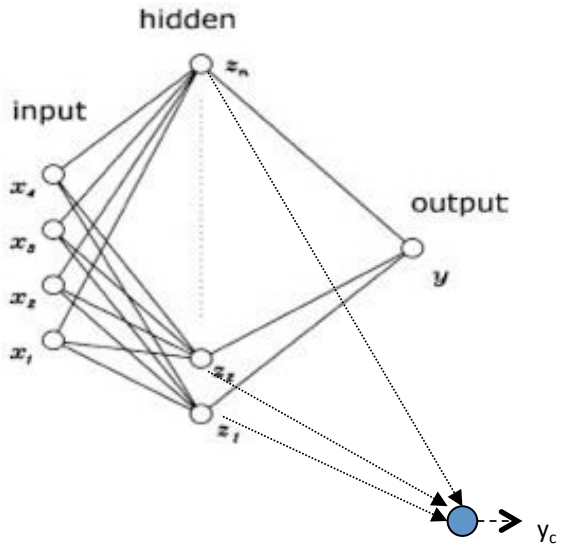
**Table 5.7**: How the same stars, detected in multiple passes, are classified using a classifier without a priori knowledge (C) and the classification tree with overlap (CTO).

| object | C | CTO | $\sigma$ |
|---|---|---|---|
| Star (objid=1185) | 99.25% | 99.67% | 2.21 |
| Star (objid=1185) | 97.98% | 98.80% | 2.28 |
| Star (objid=1185) | 95.44% | 95.78% | 2.67 |
| Star (objid=1185) | 82.79% | 90.38% | 3.67 |
| Star (objid=1722) | 89.00% | 90.50% | 2.34 |
| Star (objid=1722) | 96.14% | 94.38% | 2.50 |
| Star (objid=1722) | 78.20% | 90.10% | 3.27 |

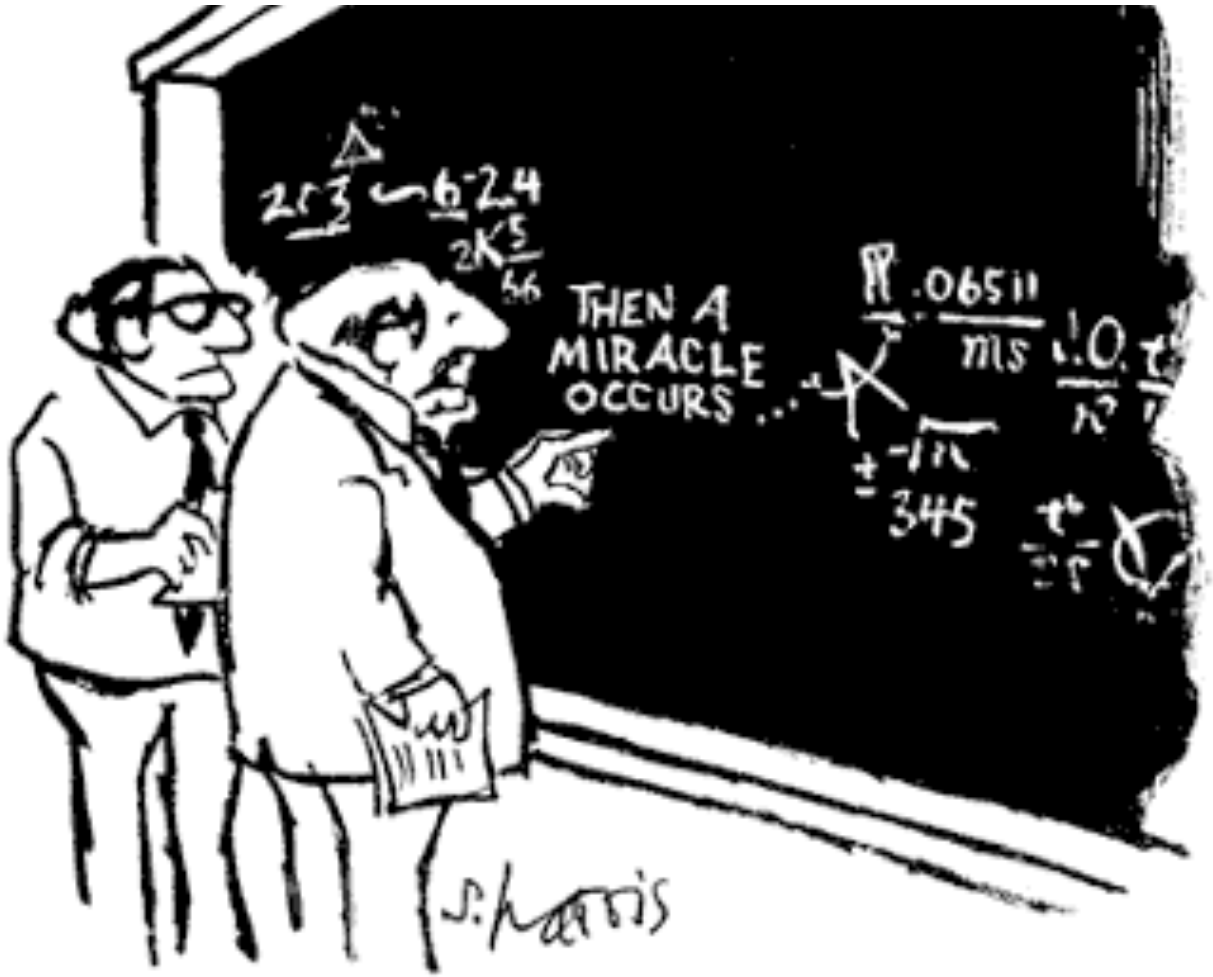
# Conclusions

- the introduction of external knowledge is still an open problem;
- this field of research is stimulated by the need to implement effective classification in synoptic digital sky surveys;
- requires that additional information not contained in the data themselves need to be taken into account;
- investigate more models, adding new parameters.



# Comments and Questions…




"I think you should be more explicit here in step two."

# Simple Cost Analysis: Stellarity

Probabilistic approach: the network output can be viewed as the ”stellarity”, how much a given object can be considered a star.

**Table 5.3**: Stars

| Stellarity threshold | Completeness | Contamination | Rejection Rate |
|---|---|---|---|
| 0.99 | 38% | 0.5% | 62% |
| 0.95 | 70% | 0.9% | 30% |
| 0.90 | 82% | 1.5% | 18% |
| 0.80 | 89% | 2.5% | 11% |
| 0.70 | 92% | 3.5% | 8% |
| 0.60 | 94% | 4% | 6% |
| 0.50 | 95% | 5% | 5% |
| 0.40 | 96% | 6% | 4% |
| 0.30 | 96% | 7% | 3% |
| 0.20 | 97% | 9% | 2% |
| 0.10 | 98% | 13% | 1% |
| 0.05 | 99% | 19% | 1% |

Results as a function of the stellarity threshold: the higher the threshold, the lower is the contamination but the higher is the rejection rate.

This kind of analysis is useful when some mistakes can be more costly than others. In order to minimize the cost, we can move the classification boundary and the threshold.

In the quasar search process the candidates must then be observed spectroscopically in order to be definitively accepted as quasars. A list of candidates contaminated by a large fraction of spurious objects causes a waste of precious observing time and man-power.