THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Applying Machine-learning to Understand the Nature of X-ray and Gamma-ray Sources

The GW Astrophysics Group
Physics Department
The George Washington University

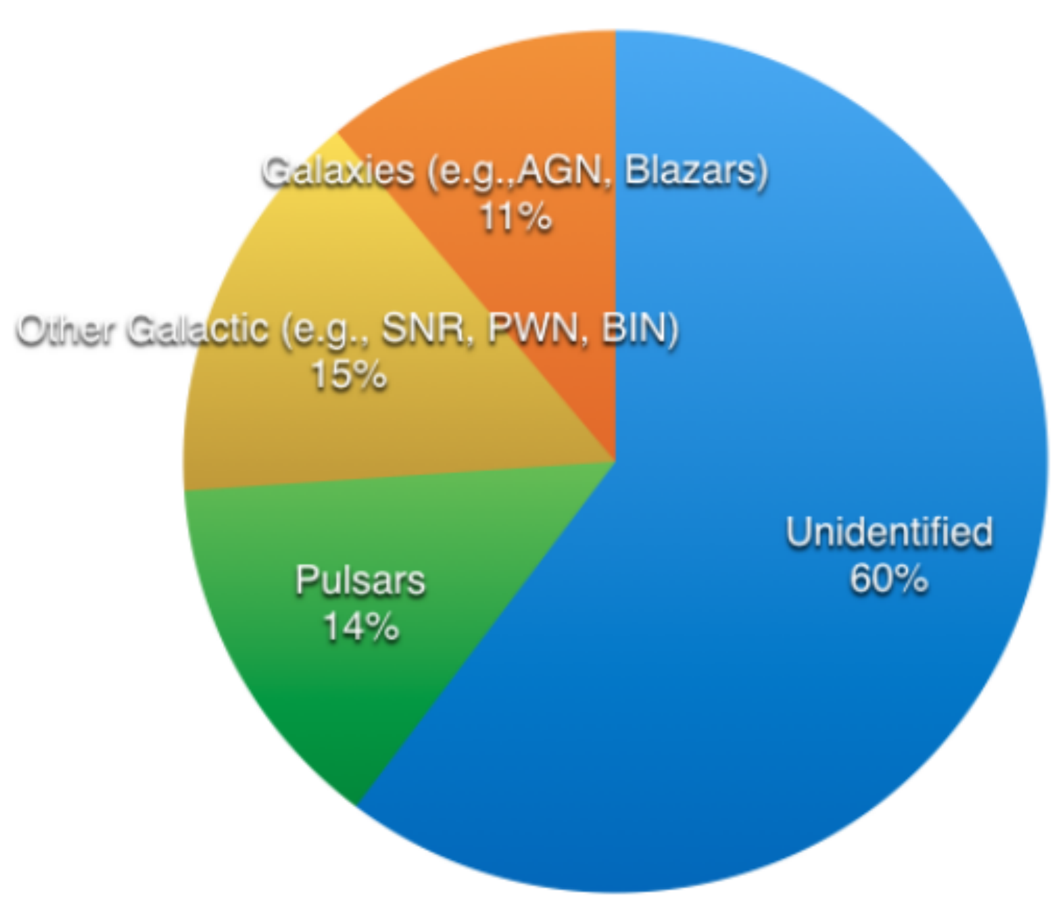
Jeremy Hare[1], Blagoy Rangelov[1], Oleg Kargaltsev[1], Igor Volkov[1,2], George Pavlov[3]

1. The George Washington University 2. University of Maryland College Park 3. Pennsylvania State University

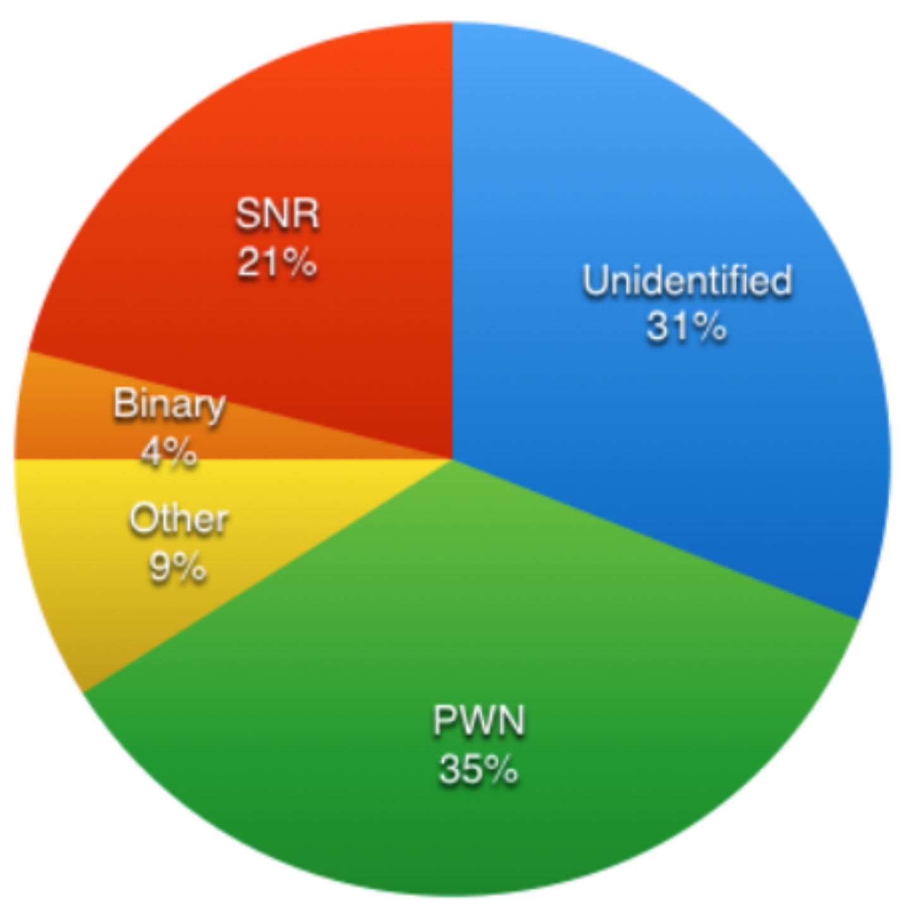
Introduction

High-energy astrophysics is currently in a golden era with observatories such as Chandra, Fermi-LAT, and H.E.S.S. viewing the sky over a multitude of wavelengths. Large amounts of data are continuously being produced and this amount will continue to grow as new, more sensitive observatories are brought online (e.g., CTA, Astro-H, eROSITA, Athena). This is leading to a rapid increase in the number of discovered high-energy sources, many of which have uncertain classifications or remain entirely unidentified. One example of this problem are the ~3,000 sources discovered by the Fermi-LAT, of which about 1/3 are unidentified (according to the 3FGL catalog produced by the Fermi-LAT Collaboration 2015). Since the gamma-ray positions are very uncertain, one promising approach is to classify all X-ray sources within the gamma-ray source fields in order to find the X-ray counterpart of the gamma-ray source. Classifying GeV and TeV γ-ray sources enables population studies (e.g., evolution, spatial distribution) and offers an opportunity to find remarkable outliers which may represent new classes of high-energy objects. In order to maximize the scientific return and observing power of current (and future) instruments, automated and accurate methods of classification must be developed and tested.

Fermi Galactic Source Breakdown

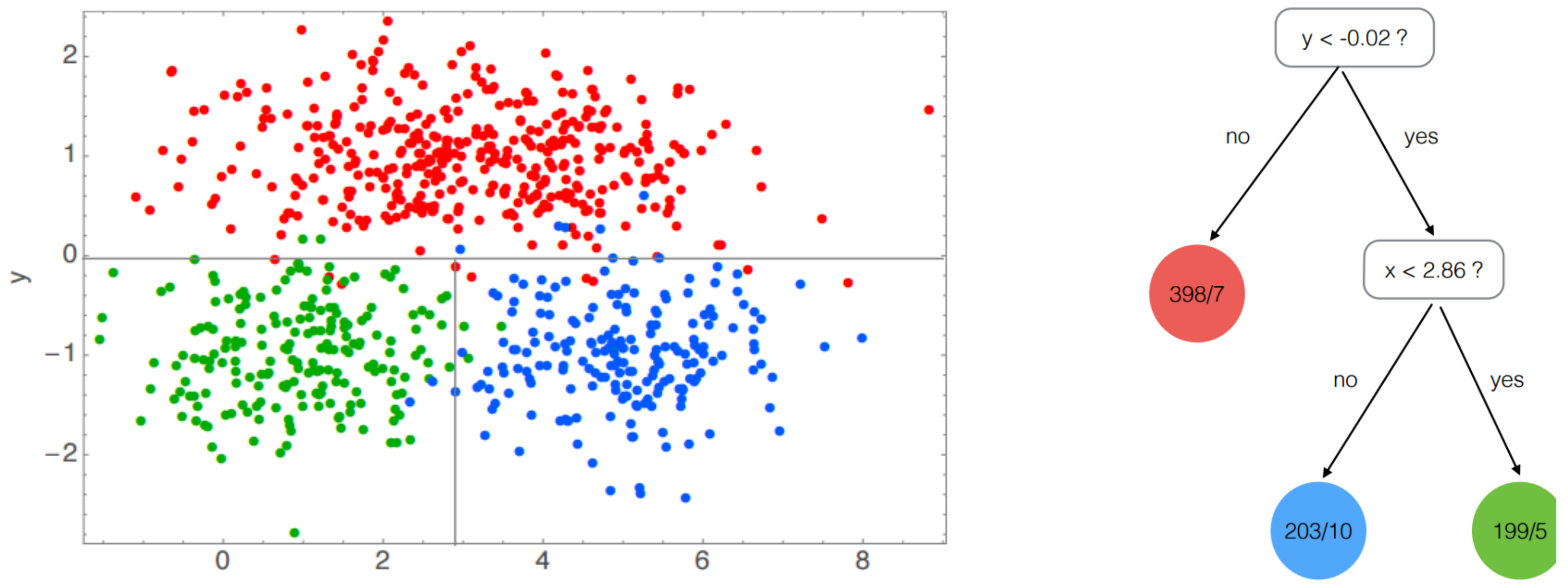


H.E.S.S. Galactic Source Breakdown



Machine Learning

To classify these objects, ML algorithms using multi-wavelength (MW) data can be applied. We use the Random Forest (RF) algorithm, which is similar in structure to decision trees (e.g., C4.5, CART), but has considerable gains. Specifically, this algorithm works by taking a sample of the training dataset with replacement (i.e., bootstrapping), building a decision tree by randomly selecting a small subset of parameters to maximize on, and then repeating the process. By doing this, the RF algorithm builds an ensemble of decision trees, each one different from the rest. This method is much less prone to overfitting when compared to a single decision tree (Breiman et al. 2001) and is generally more robust due to the larger number of decision trees.



Training Dataset

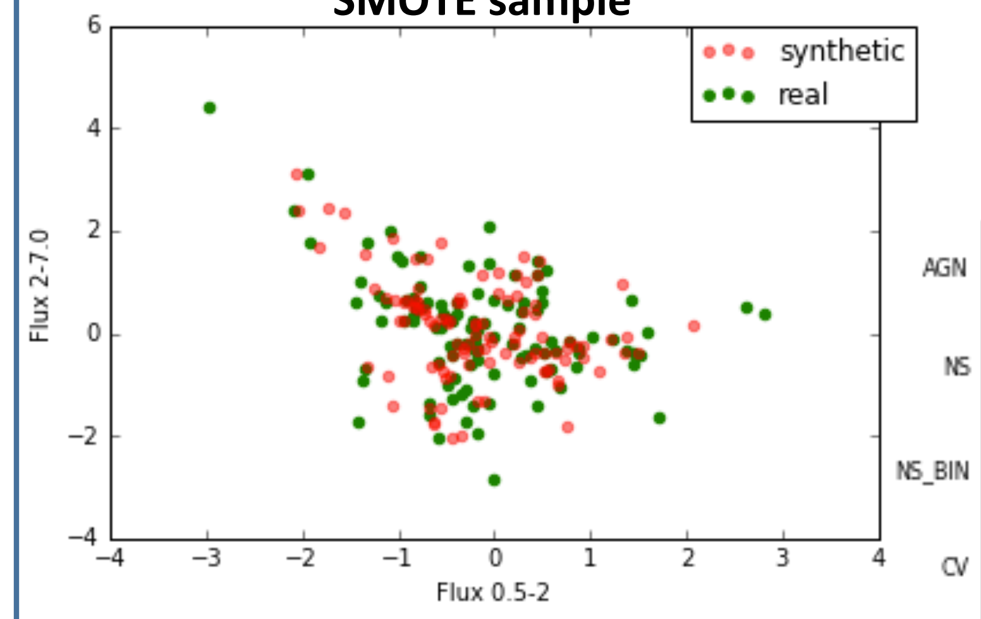
The training data set is built and used to evaluate the features of objects from known classes and to build a decision tree. The current training dataset consists of ~7,700 confidently classified X-ray sources with X-ray and MW features (up to 18 listed in the Table in the next section) extracted from 14 catalogs provided by Vizier. The current classification scheme uses 9 different object classes: Active Galactic Nuclei (AGN), Neutron Stars (NS), Neutron Star Binaries (NS_BIN), Low Mass X-ray Binaries (LMXB), High Mass X-ray Binaries (HMXB), Cataclysmic Variables (CV), Stars, Wolf-Rayet Stars (WR), and Young Stellar Objects (YSO).

	AGN	NS	NS_BIN	CV	HMXB	LMXB	STAR	WR	YSO
Number	5794	85	10	134	21	54	1313	27	212

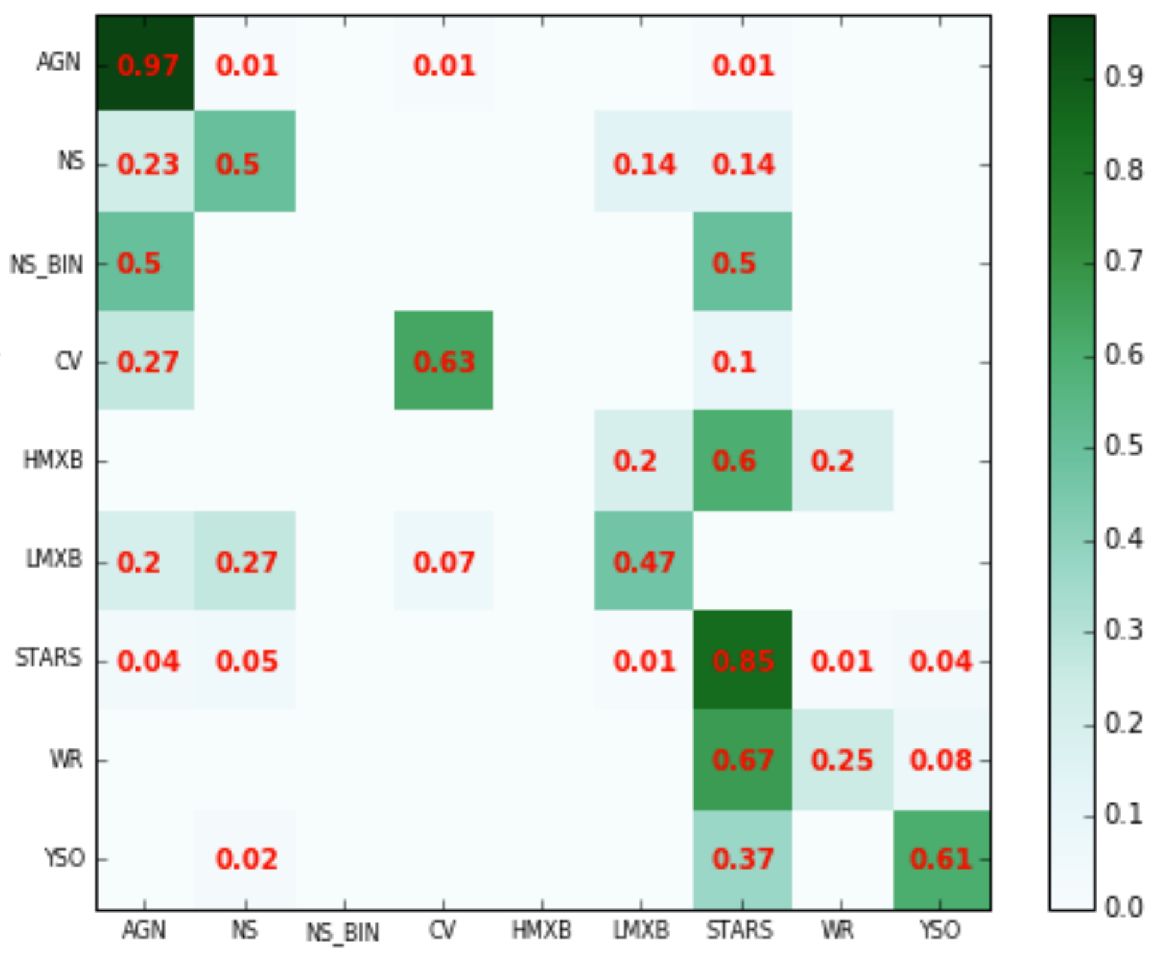
Imbalanced Data

Our current training dataset is heavily imbalanced, which can lead to skewed classifications if not accounted for. We account for this imbalance in two ways. The first is by weighting the RF algorithm based on the number of sources in a particular class. This means that the algorithm assigns a higher cost to the rarer source classes. The second method we use is the Synthetic Minority Over-sampling Technique (SMOTE; Chawla et. al 2011). This methods find the 5 nearest class neighbors to a minority class and selects one of these 5 at random. Then the algorithm creates a new synthetic source that is scaled to lie between the two originally selected sources (see Figure below). This process is repeated until all sources have the same number of samples.

SMOTE sample



Confident Confusion Matrix
(50% trained / 50% test)



Features used for Classification

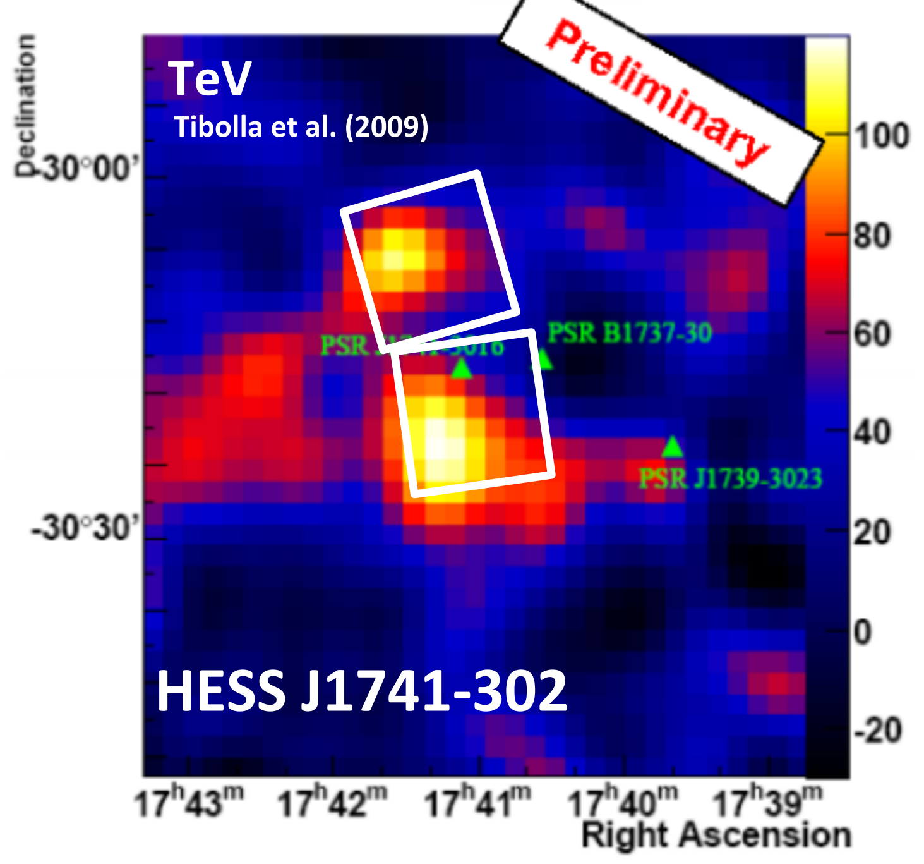
Feature	Description
EP052Flux	X-ray flux in the 0.5−2 keV band
EP27Flux	X-ray flux in the 2−7 keV band
HR2	Soft band hardness ratio defined in Section 2.2
HR4	Hard band hardness ratio defined in Section 2.2
Bmag	Magnitude of counterpart in B-band
Rmag	Magnitude of counterpart in R-band
Imag	Magnitude of counterpart in I-band
Jmag	Magnitude of counterpart in J-band
Hmag	Magnitude of counterpart in H-band
Kmag	Magnitude of counterpart in K-band
W1mag	Magnitude of counterpart in W1-band
W2mag	Magnitude of counterpart in W2-band
W3mag	Magnitude of counterpart in W3-band
coljk	Magnitude of J-band minus K-band
coljh	Magnitude of J-band minus H-band
colri	Magnitude of R-band minus I-band
colw12	Magnitude of W1-band minus W2-band
colw13	Magnitude of W1-band minus W3-band

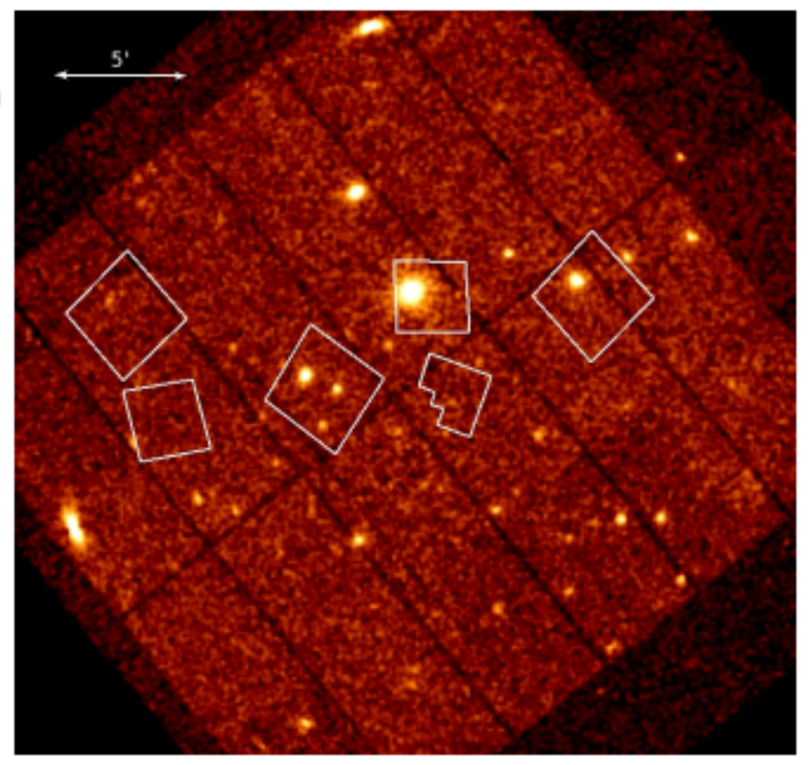
Cross-validation

We have used 10 fold cross-validation with a weighted RF algorithm and have gotten an overall accuracy of 91%. The weighted RF algorithm provides a superior overall classification confidence when compared with the SMOTE'd training dataset (86%). However, the SMOTE'd training dataset is more accurate at predicting minority classes (i.e., NS, LMXB, YSO, WR). Above you can see the SMOTE'd confusion matrix for confidently (>70% classification confidence) classified sources our training dataset. In total, out of the sources that had confident classifications, 93% were confidently classified as the correct class.

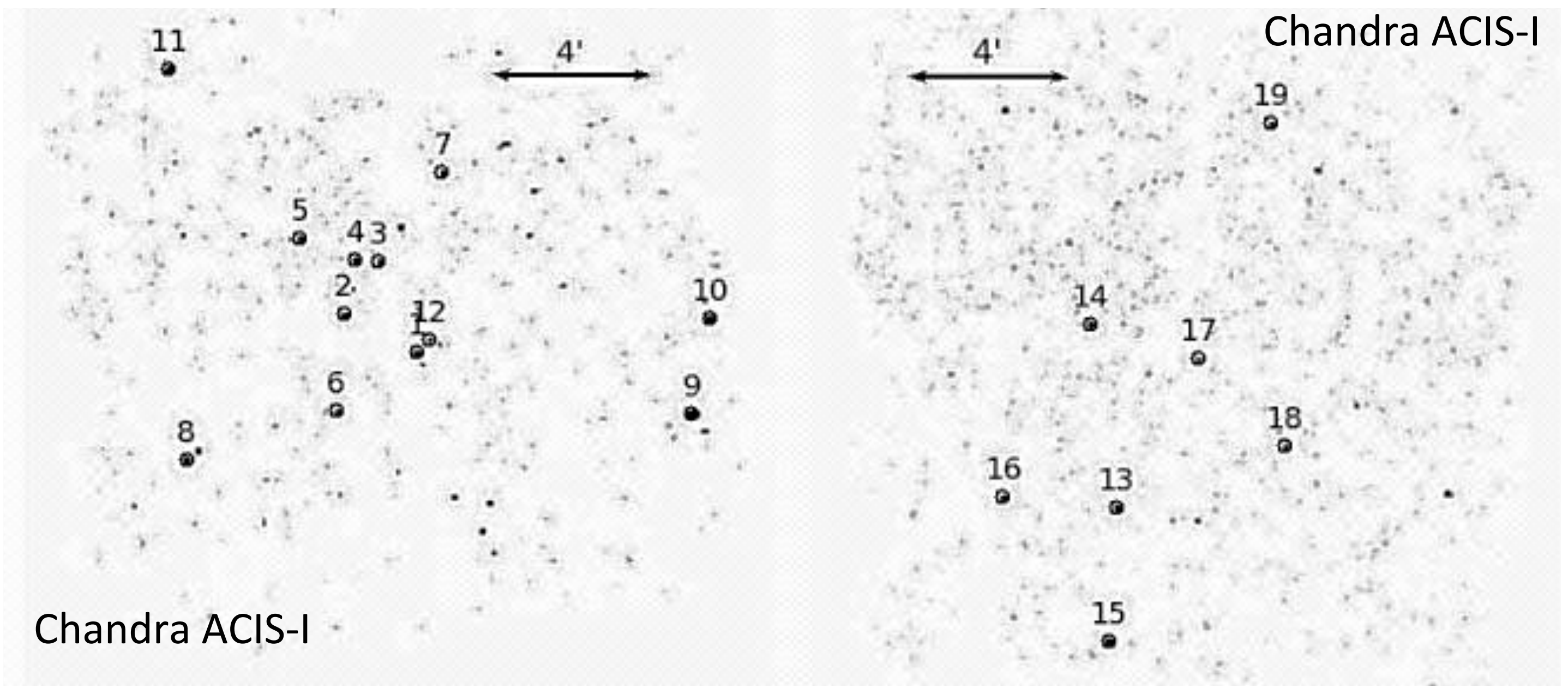
Applications: HESS J1741-302 & Draco

- Unidentified galactic VHE sources observed with Chandra X-ray Observatory;
- 19 sources were found in the Chandra ACIS-I fields of view marked by white squares on top of the TeV images shown below.
- We used our automated classification pipeline to look for potential X-ray counterparts of the TeV sources and to classify all X-ray sources in the ACIS-I fields.





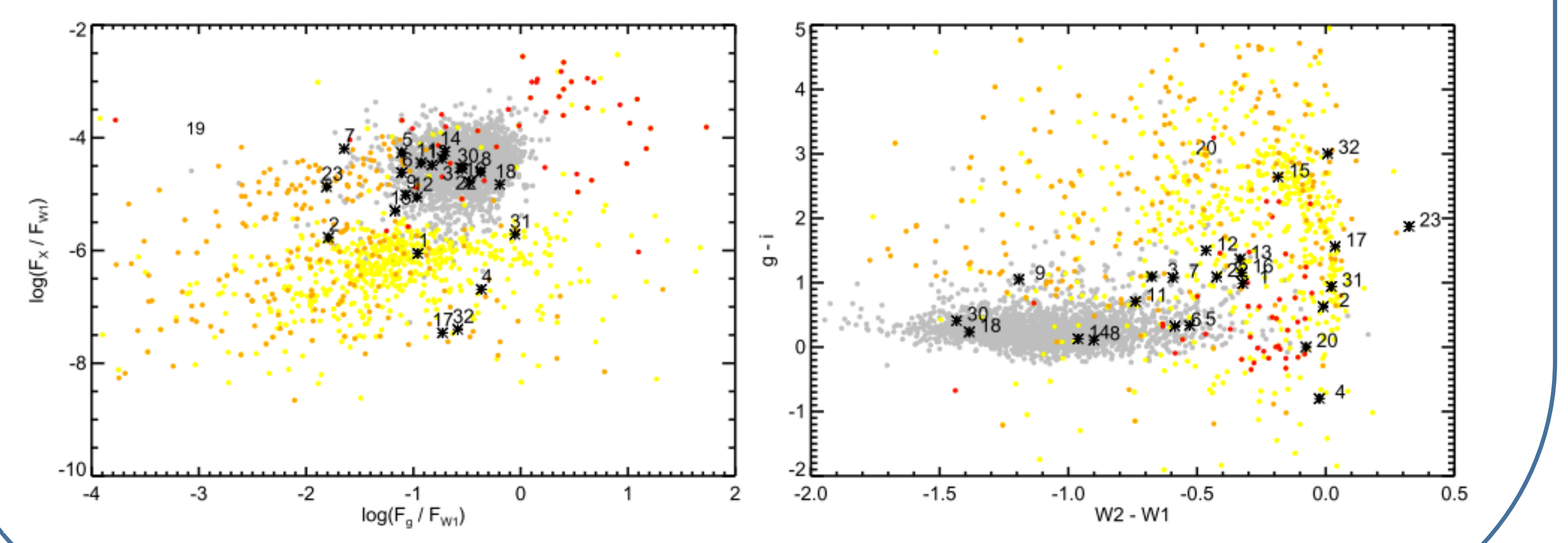
HESS J1741-302



- Classifications from our pipeline (which reliably detects Stars and AGN) allowed us to rule out many X-ray sources in the fields
- HESS J1741 field contains an likely WR binary system, which have been hypothesized to produce TeV emission (Aliu et al. 2008). The only other potential counterpart is the offset 20-kyr-old pulsar B1737-30.
- See Hare et al.(2015) for the detailed analysis description from this HESS source.

Draco

XMM-Newton observations of Draco were fed through our pipeline in order to probe the X-ray source population of this dwarf galaxy. There was a high level of similarity between our algorithms classifications and classifications done manually (see Manni et al. 2015). See Sonbas et al. (2016). The two parameter plots can be seen below.



Two parameter plots showing separation of classes. AGN: Grey, YSO: Orange, Star: Yellow, Draco Unidentified: Black

Outlook

With the growing number of all sky and wide-field observatories coming online in the near future (e.g. LSST, eRosita, Athena) developing and understanding these methods is crucial. These methods will also allow us to get a larger return on the data from current telescopes as well (e.g., Fermi-LAT, Chandra, H.E.S.S., VERITAS). By automating this procedure we can classify sources in near real-time in the data streams from future observatories , which will allow for the identification and multi-wavelength follow-up of interesting sources.

References

Acero, F., et al. "Fermi Large Area Telescope Third Source Catalog." *The Astrophysical Journal Supplement Series* 218.2 (2015): 23.
Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.
Hare, Jeremy, et al. "Multi-wavelength study of HESS J1741-302." *arXiv preprint arXiv:1506.05736* (2015).
Tibolla, Omar, et al. "New unidentified HESS Galactic sources." *arXiv preprint arXiv:0907.0574* (2009).
Manni, L., et al. "A XMM-Newton observation of a sample of four close dwarf spheroidal galaxies." *Monthly Notices of the Royal Astronomical Society* 451.3 (2015): 2735-2749.
Sonbas, E., et al. "X-ray Sources in the Dwarf Spheroidal Galaxy Draco." *arXiv preprint arXiv:1505.00216* (2015).
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2011, arXiv:1106.1813

Acknowledgements

I would personally like to thank the co-authors of this poster for their useful comments and feedback. Support for this work was provided by NASA through the Chandra X-ray Observatory award AR3-14017X.