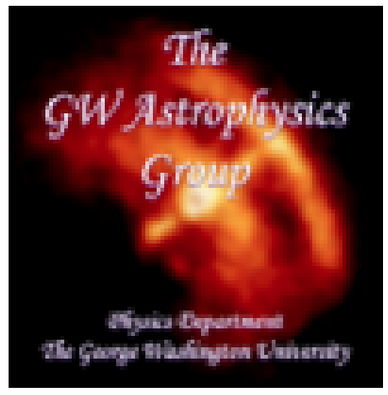
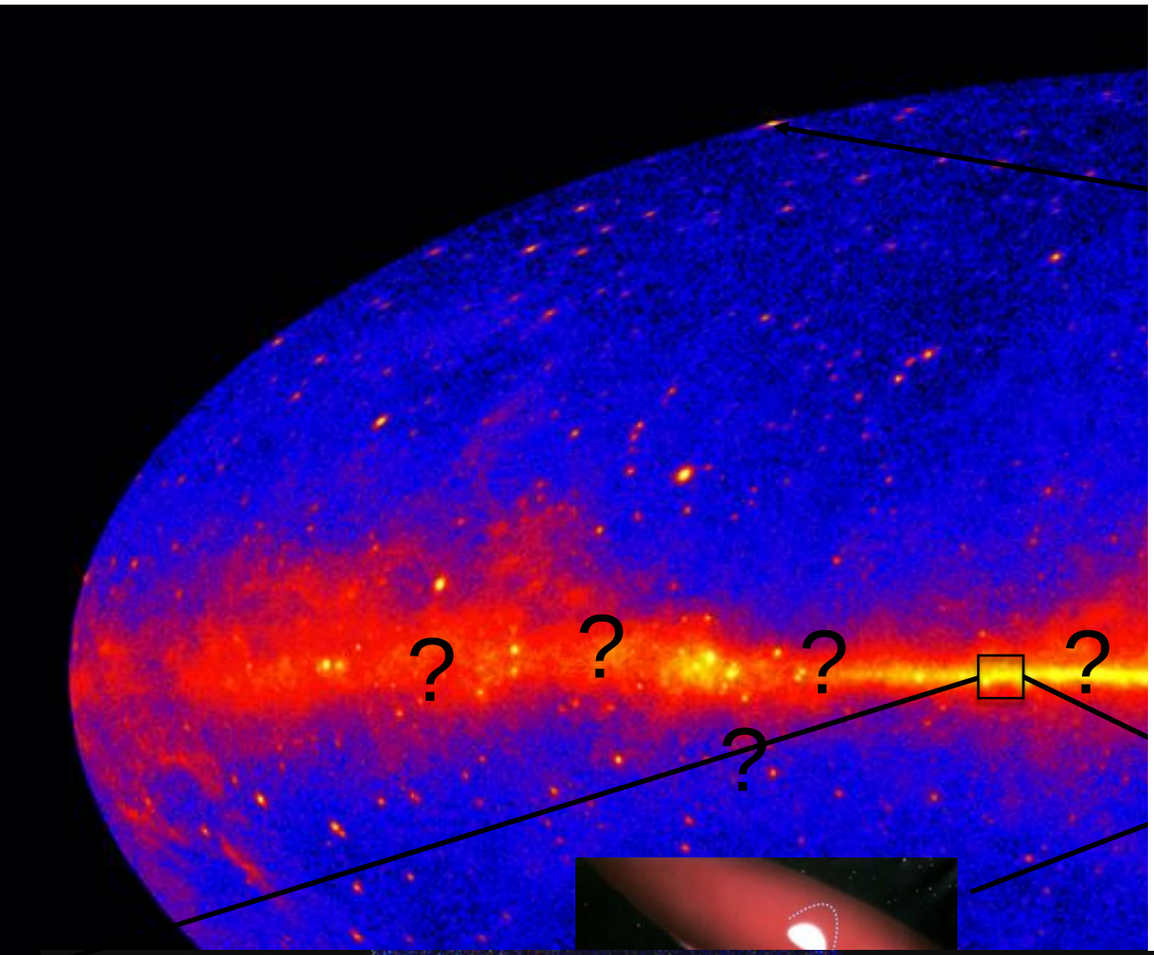
Application of Machine-Learning Techniques to Understand the Nature of X-ray and Gamma-ray Sources

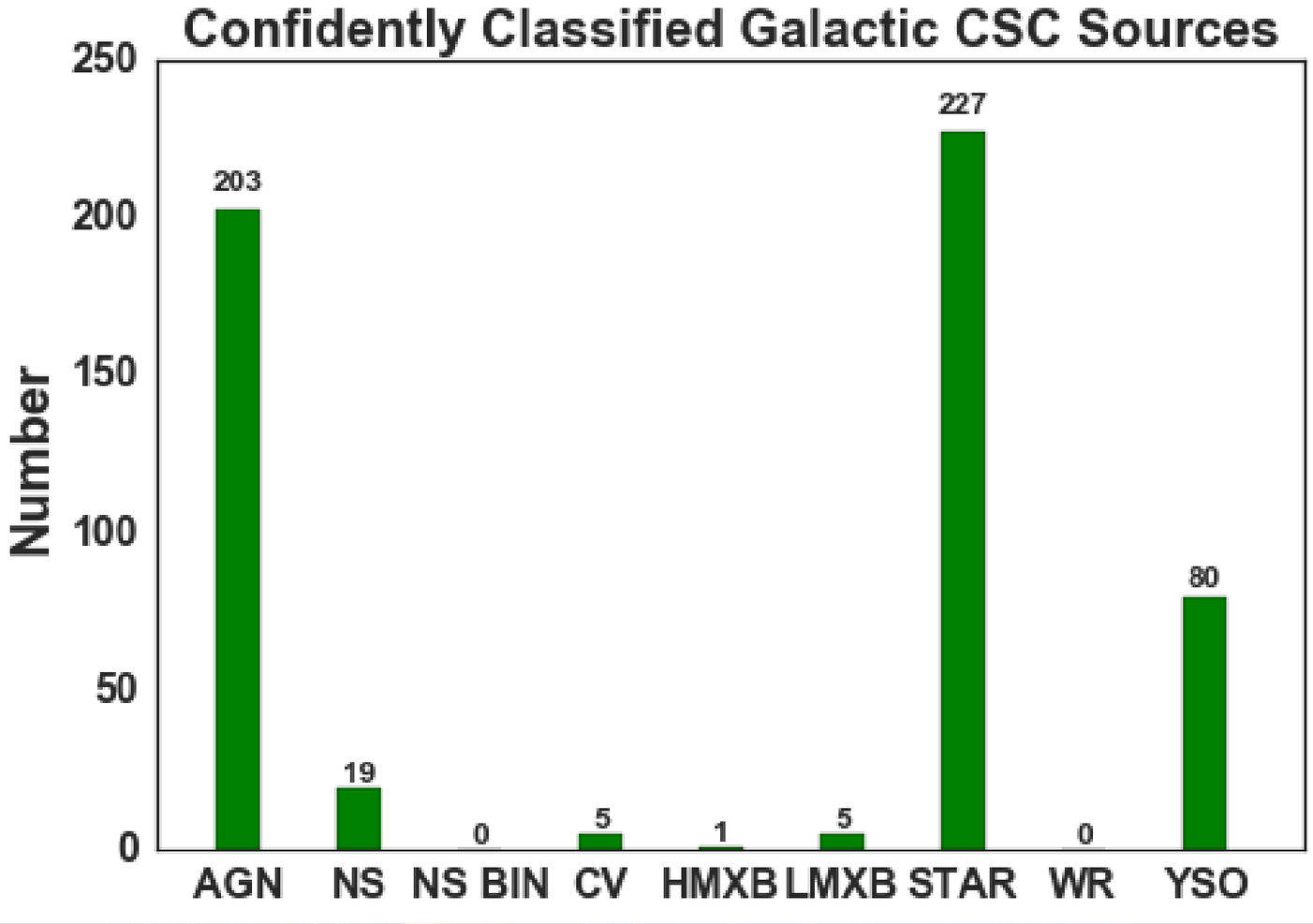
Jeremy Hare
Collaborators: Oleg Kargaltsev (GWU) George Pavlov (PSU), Blagoy Rangelov (TSU), Igor Volkov (UMD)

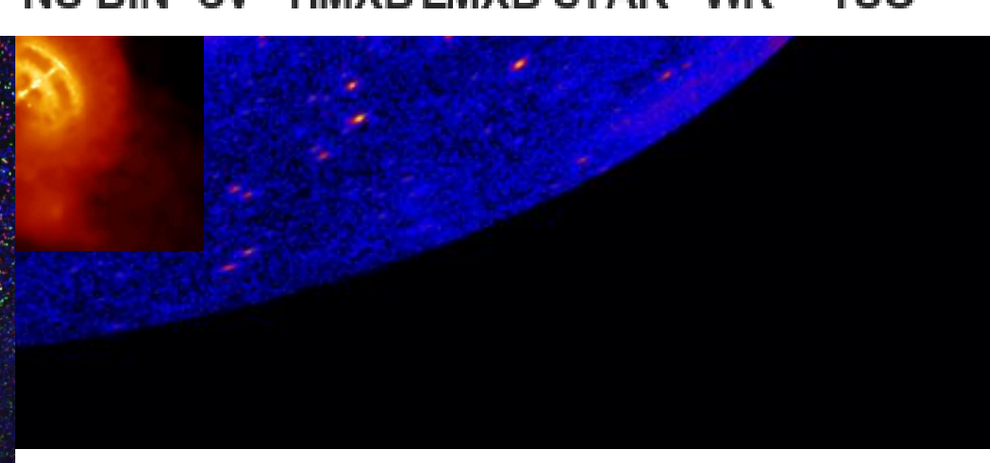
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC



IAU Symposium 325 Astroinformatics, October 20, 2016







-Ray Sky

997/39/image/a/

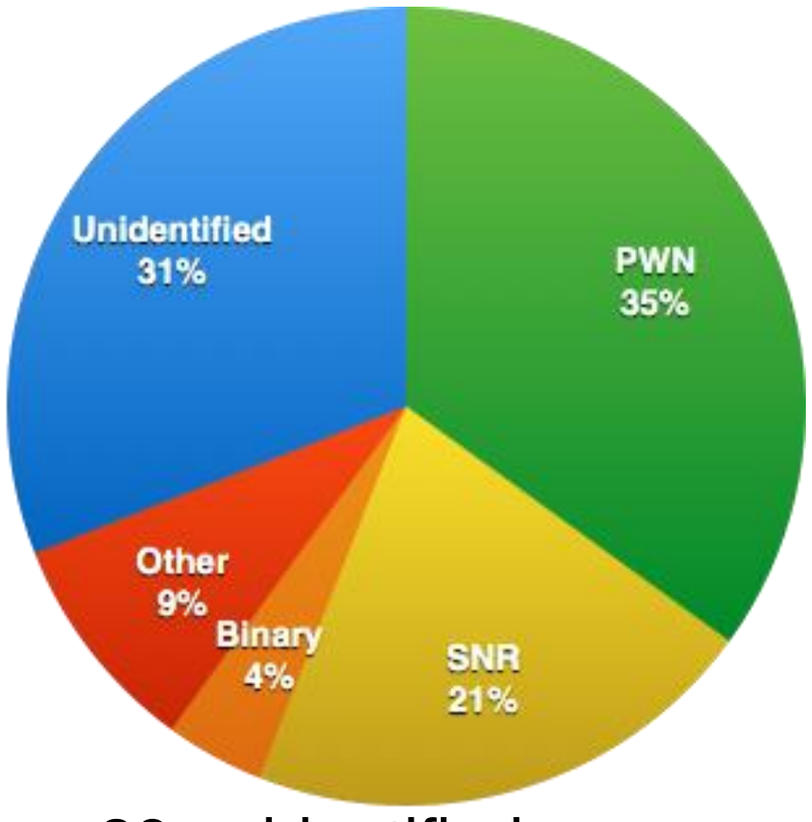
ations.html

boration

Defining the Problem

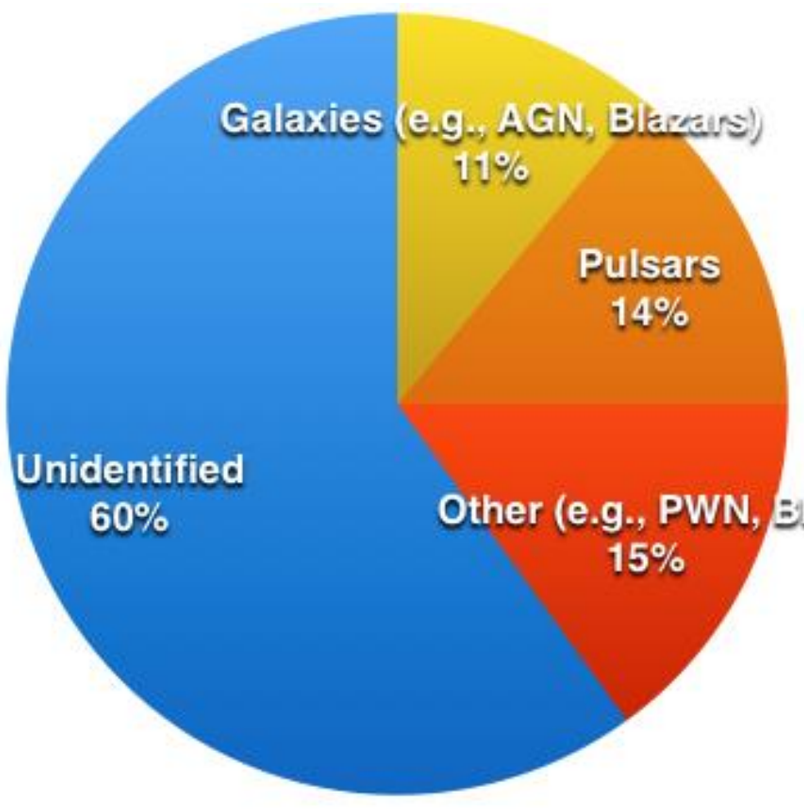
- Many unidentified very high energy sources

H.E.S.S. Galactic Source Breakdown



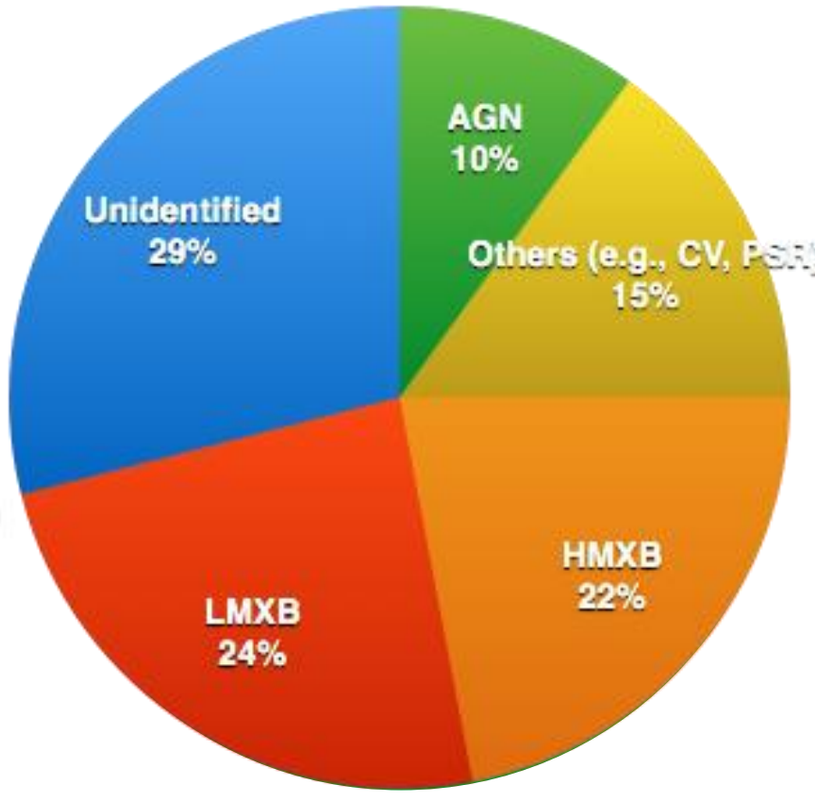
~20 unidentified sources

Fermi 3FGL Galactic Source Breakdown



~350 unidentified sources

Integral Galactic Source Breakdown

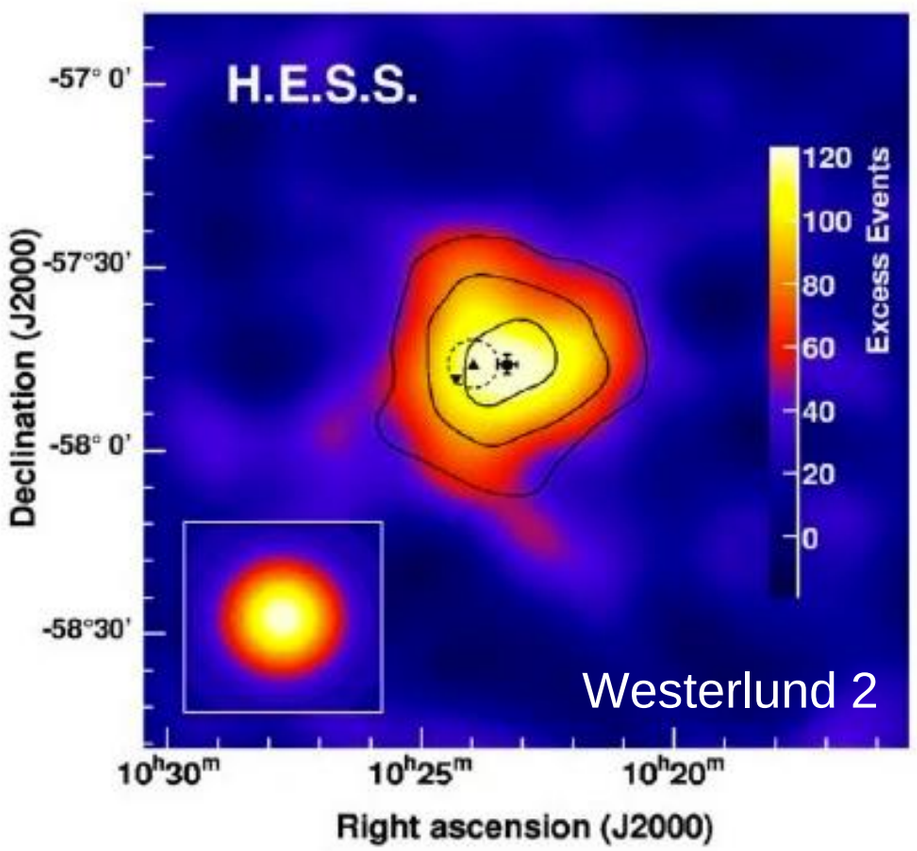


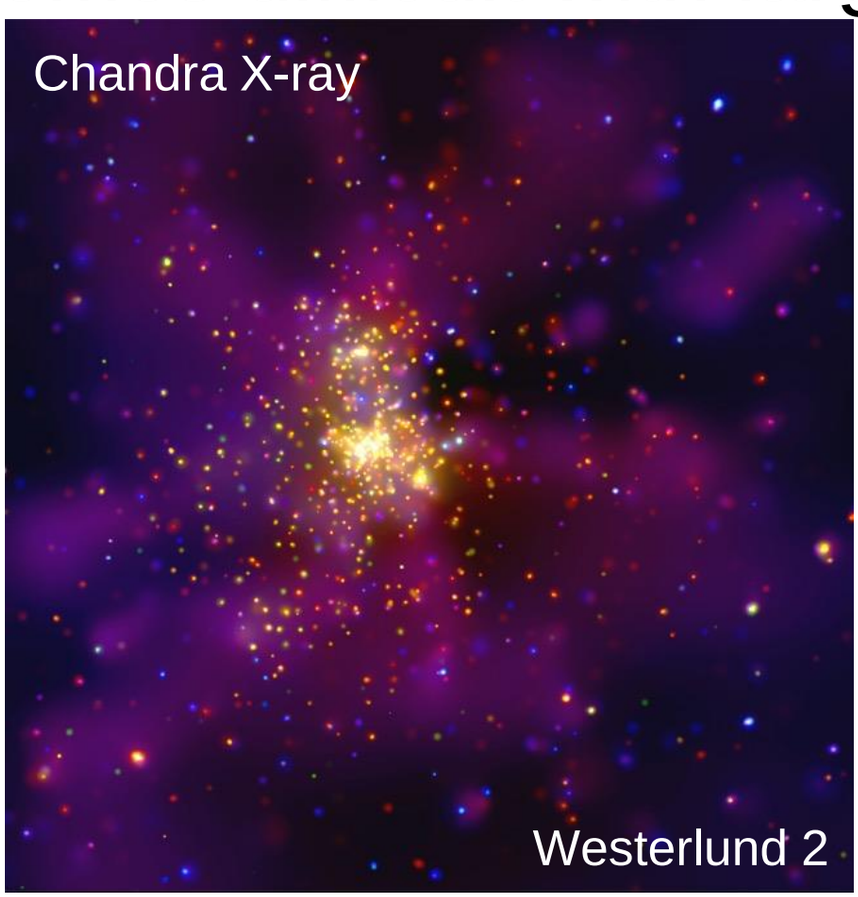
~50 unidentified sources

Defining the Problem

- MW counterparts can be used to classify these objects

- Identifying counterparts becomes difficult with larger error circles/ellipses





taken from https://www.mpi-hd.mpg.de/hfm/HESS/pages/home/som/2006/12/
Naze et al. (2008)
http://chandra.harvard.edu/photo/2008/wd2/

Defining the Problem

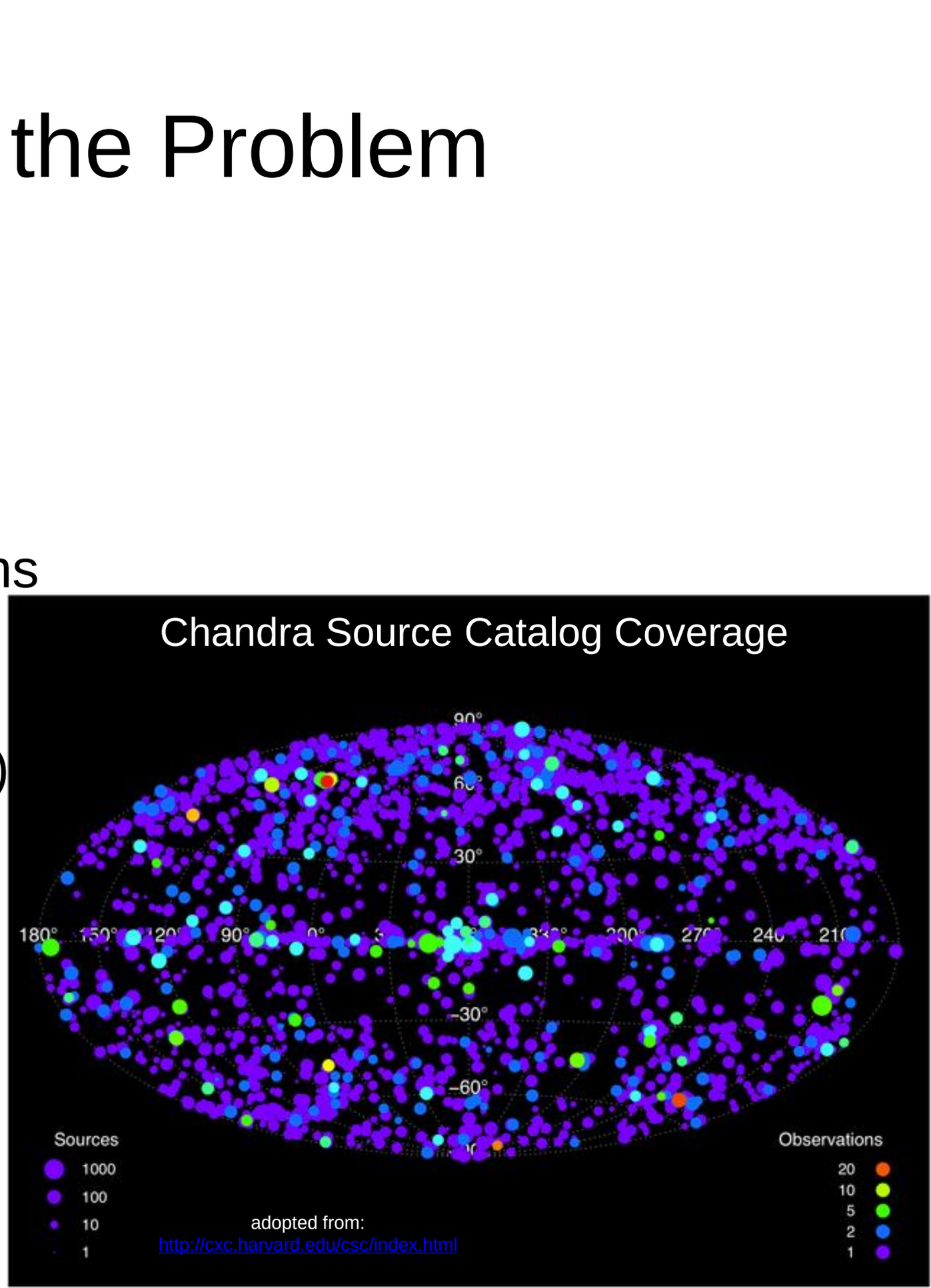
- Many X-ray archival images (many sources serendipitously observed >90% remain unidentified)

- First CSC data release contains ~95,000 sources

- 3XMM-DR6 (468,440 sources)

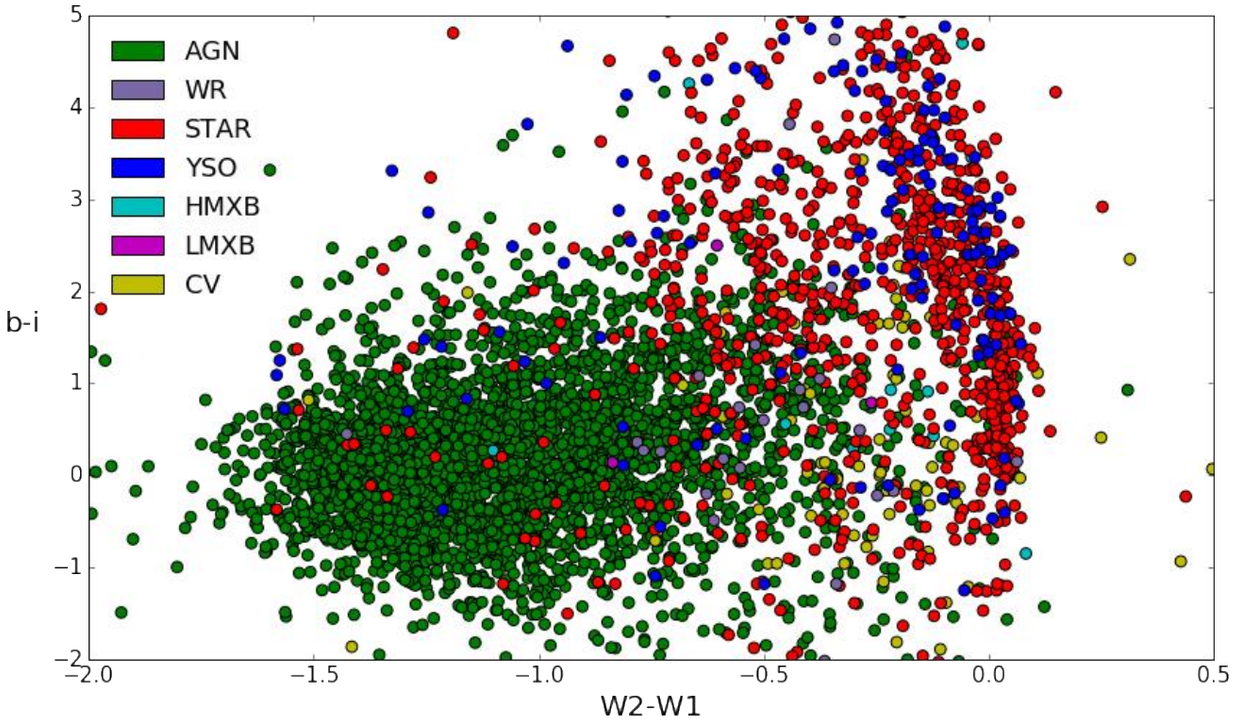
- CSCv2 ~400,000 sources (upcoming)

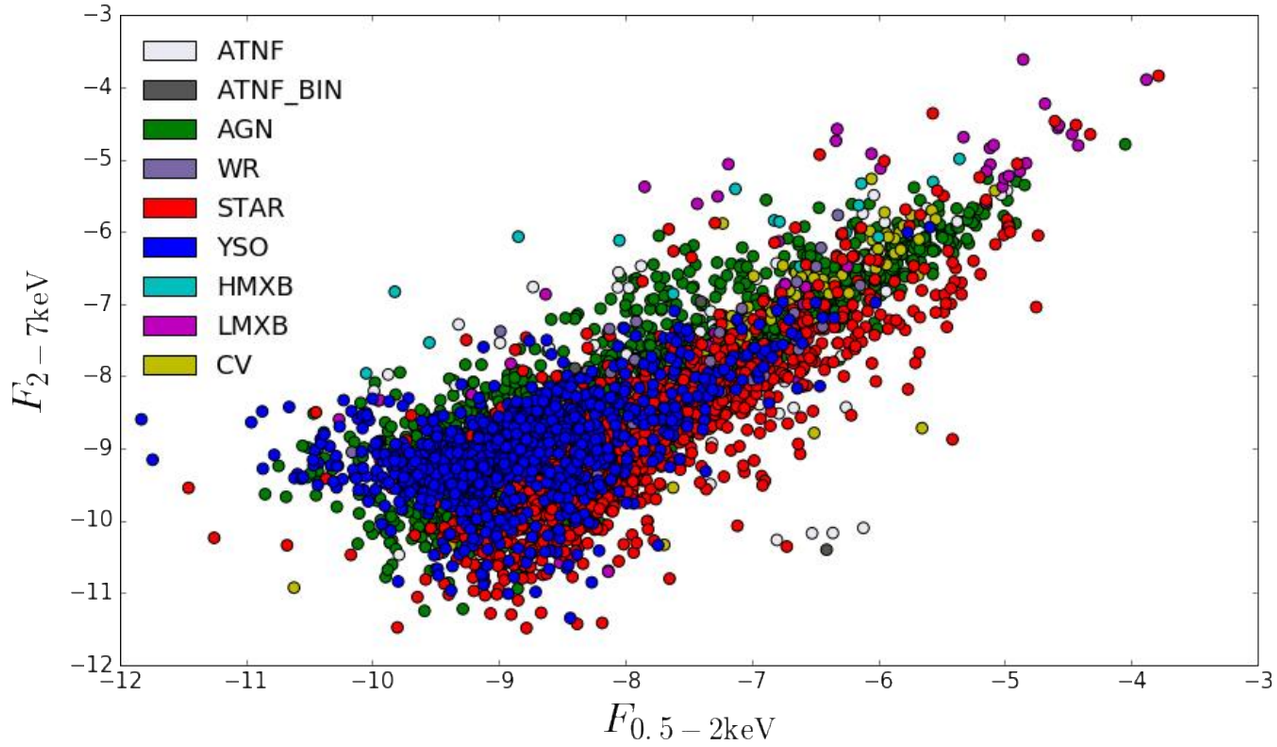
Color shows # of observations
Size shows # of sources

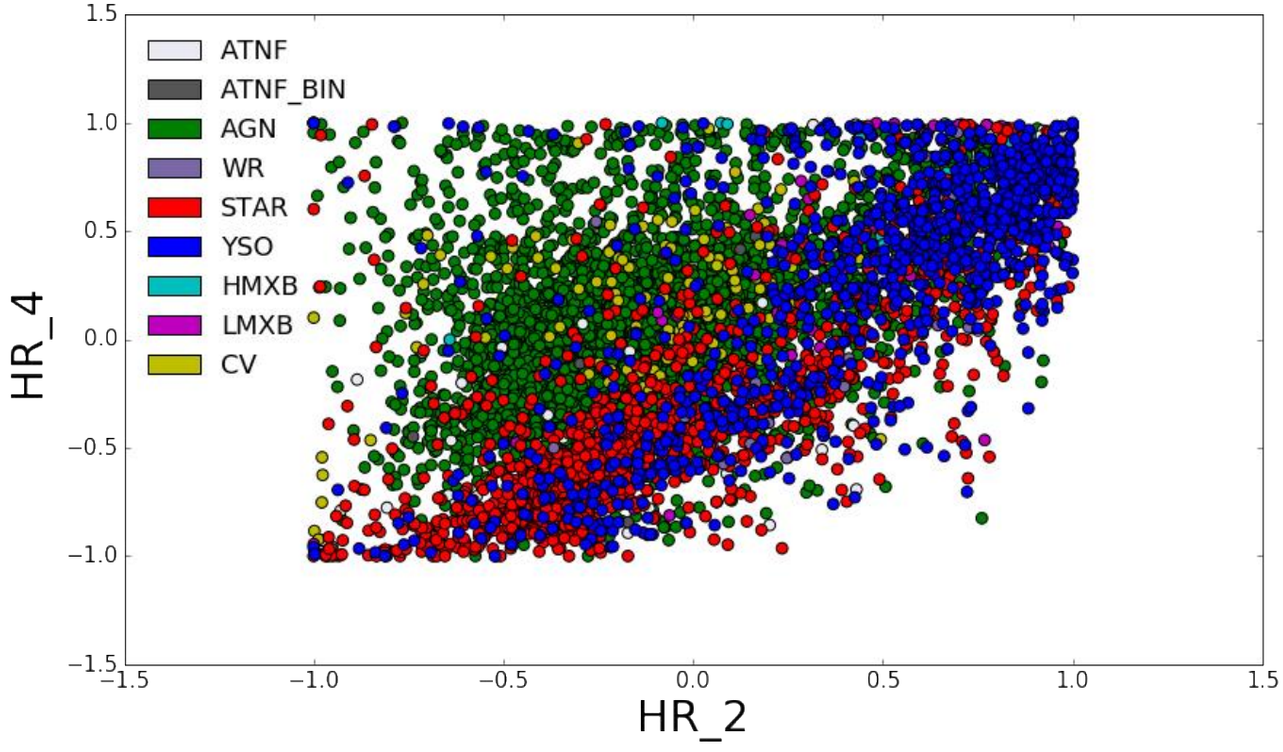


Defining the Problem

- Often X-ray data is not enough to classify sources
- Multi-wavelength data are needed
- Leads to high dimensionality problem


AGN
WR
STAR
YSO
HMXB
LMXB
CV
b-i
W2-W1


ATNF
ATNF_BIN
AGN
WR
STAR
YSO
HMXB
LMXB
CV
$F_{2-7\mathrm{keV}}$
$F_{0.5-2\mathrm{keV}}$


ATNF
ATNF_BIN
AGN
WR
STAR
YSO
HMXB
LMXB
CV
HR_4
HR_2

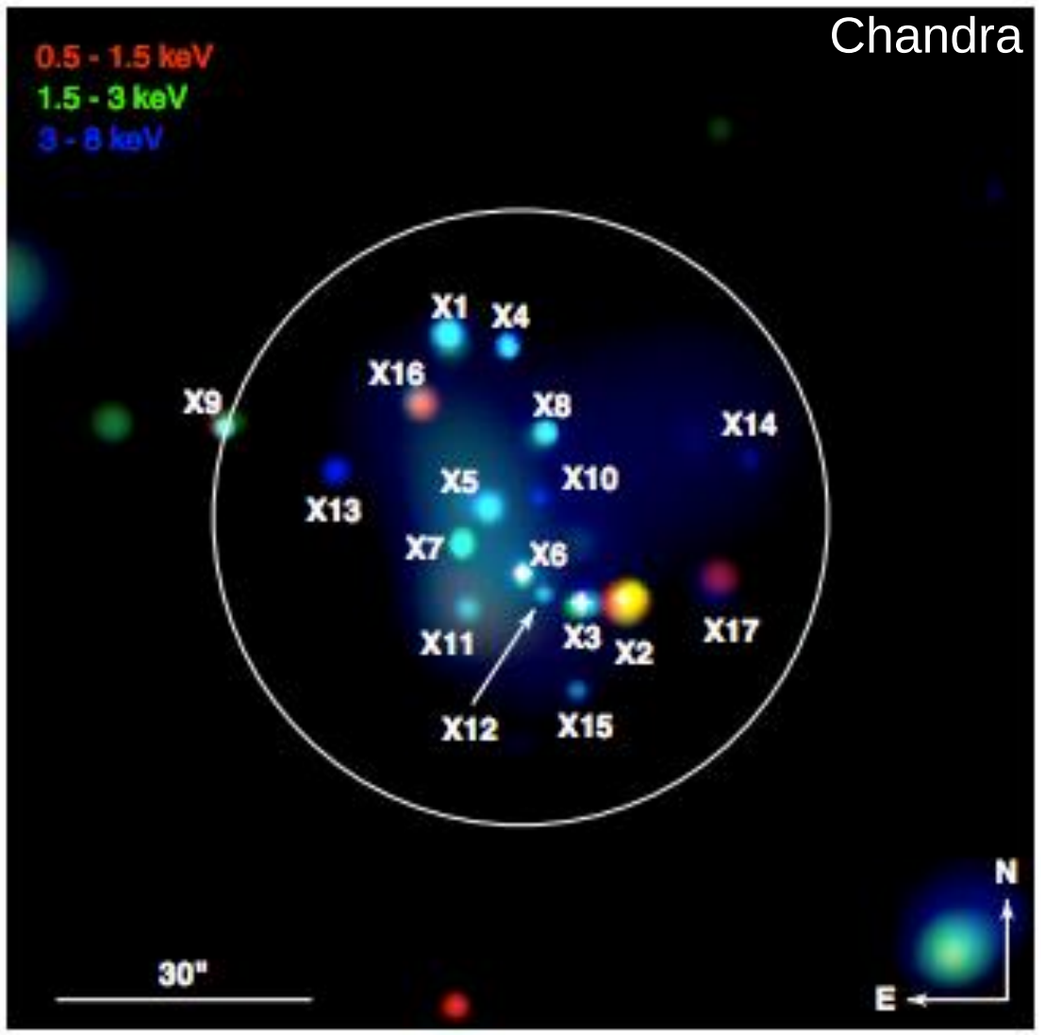
Astrophysical Importance: High Energy Sources

- Map out the population of GeV emitting compact objects in the galaxies (e.g., NS, BHs in binaries)
- Search for new and rare GeV and TeV emitting source classes (e.g., Colliding wind binaries)

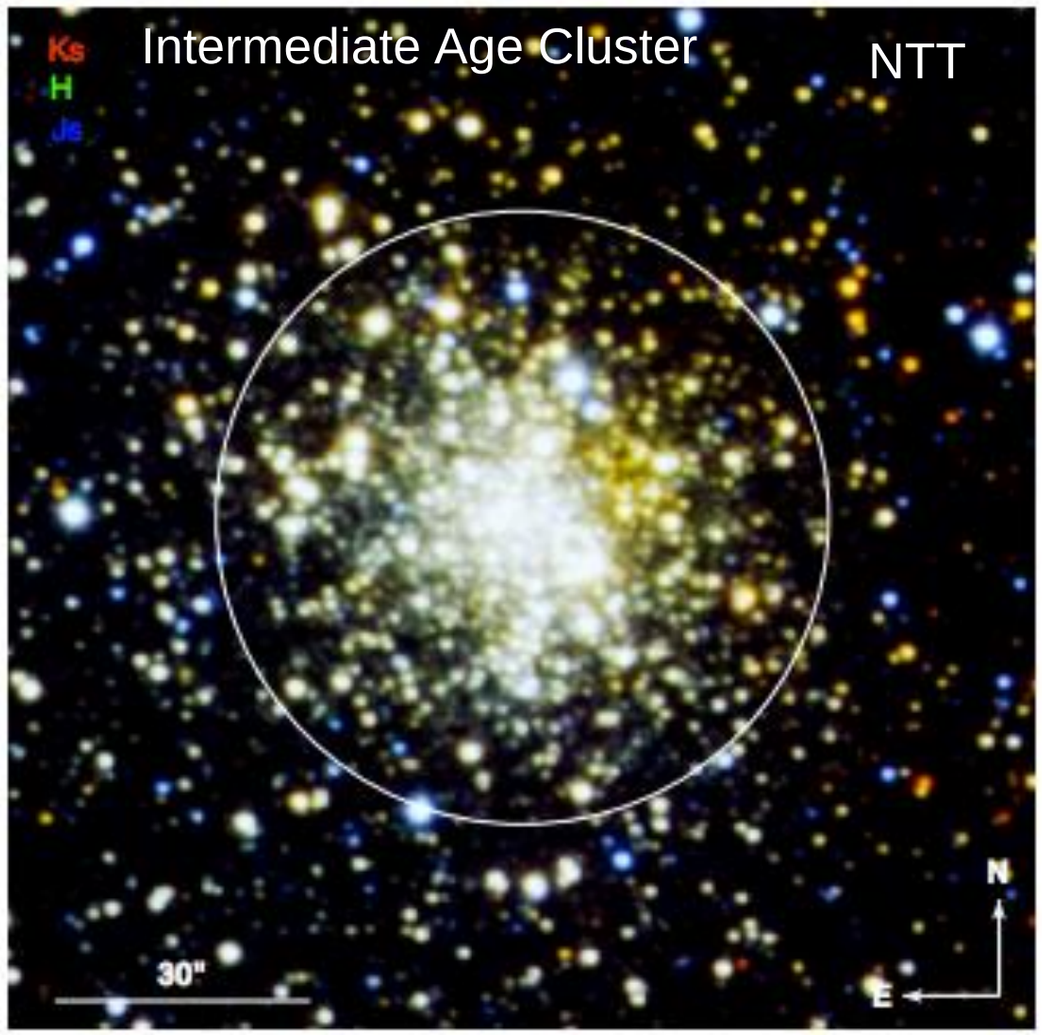
Astrophysical Importance: High Energy Sources

- Identifying compact objects in SNRs
- Understanding populations in galaxies/clusters

Pannuti et al. (2014)



Pooley et al. (2007)



Solution

- How can we efficiently classify all of these sources?
- Computers may eventually one day destroy us; however, until then we can take advantage of them!
- Machine learning can be used to handle large datasets and large numbers of parameters

Supervised Machine Learning

- Requires a training dataset
- Training dataset is used to teach the classifying algorithm how to make inferences about unclassified data

Training Dataset

- ~9,000 literature verified X-ray sources (from 8 catalogs, representing 9 source types)

Source Type	3XMM-DR6	CSC	Overlap
AGN	6526	1229	571
NS	88	33	32
NS BIN	10	5	4
CV	136	75	10
HMXB	21	8	5
LMXB	55	31	17
STAR	1380	514	291
WR	32	13	6
YSO	978	1152	221
Total	9226	3060	1157

Features=Measured Parameters

- Currently 19 features

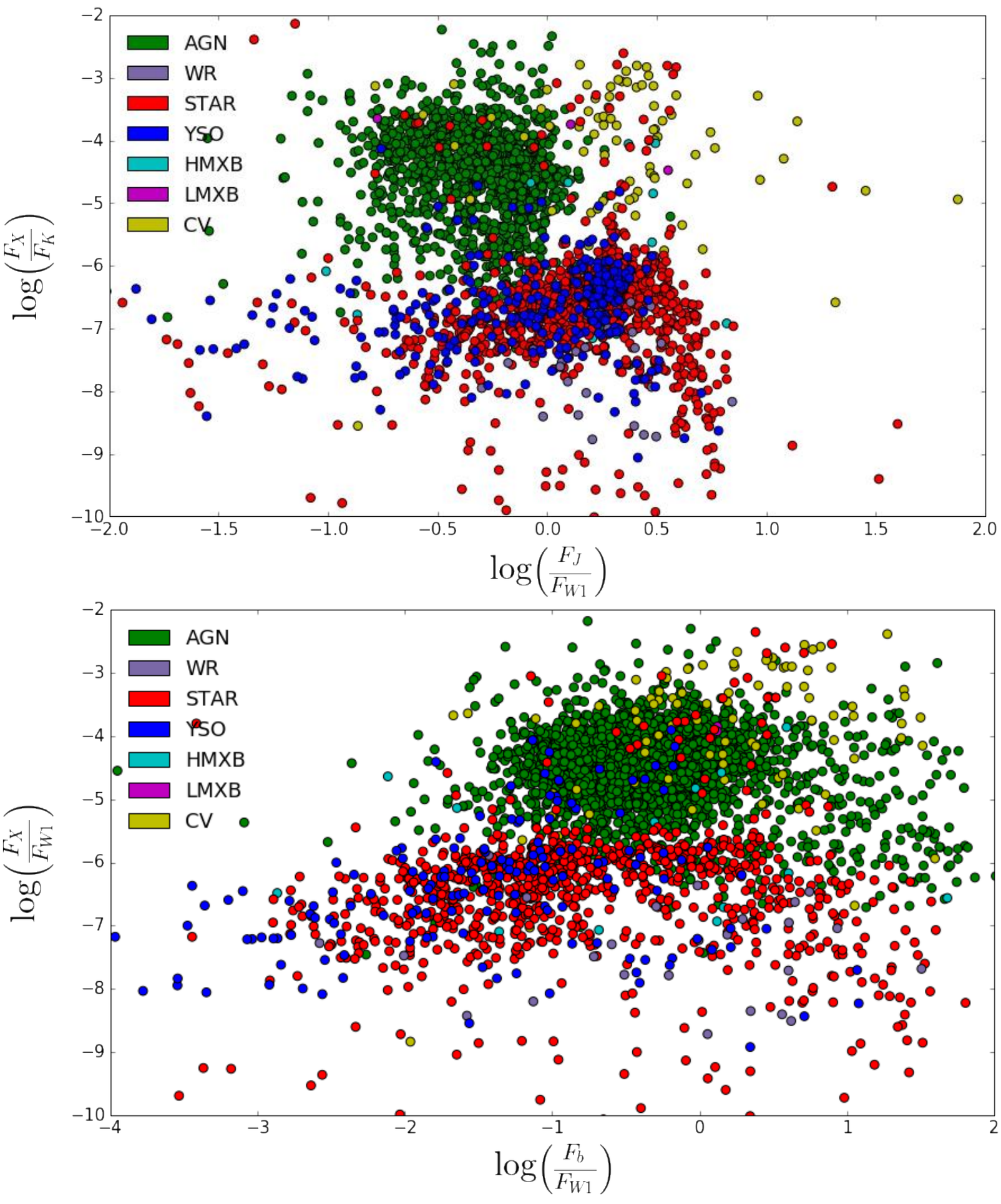
USNO-B

2MASS

WISE

All-Sky Catalogs

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
colw2j	Magnitude of W2-band minus the J-Band

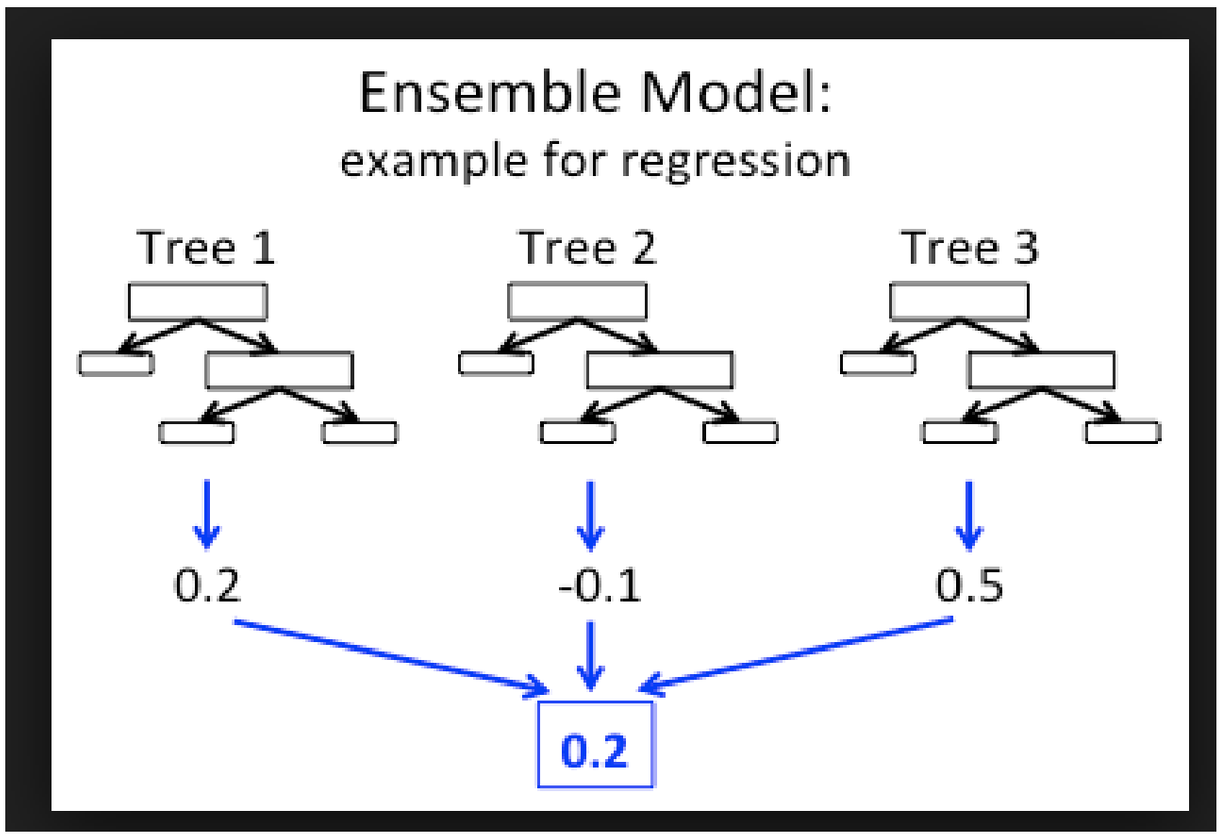

AGN
WR
STAR
YSO
HMXB
LMXB
CV
$\log\left(\frac{F_X}{F_K}\right)$
$\log\left(\frac{F_J}{F_{W1}}\right)$
AGN
WR
STAR
YSO
HMXB
LMXB
CV
$\log\left(\frac{F_X}{F_{W1}}\right)$
$\log\left(\frac{F_b}{F_{W1}}\right)$

Random Forest (RF)

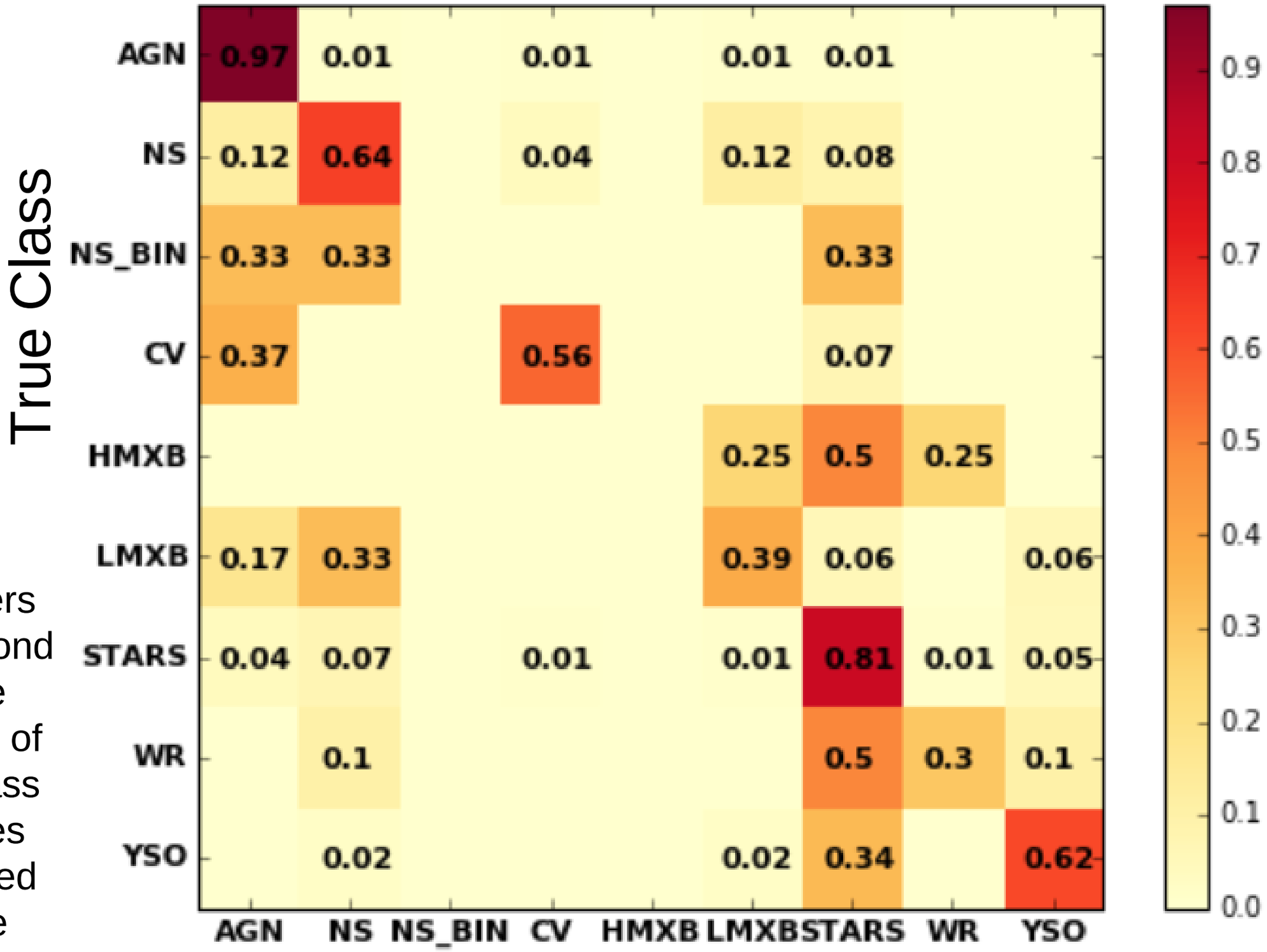
- Bootstraps training dataset
- Uses a random subset of features
- Reduces overfitting

Pedregosa *et al. (2011)*

- Not sensitive to uninformative features



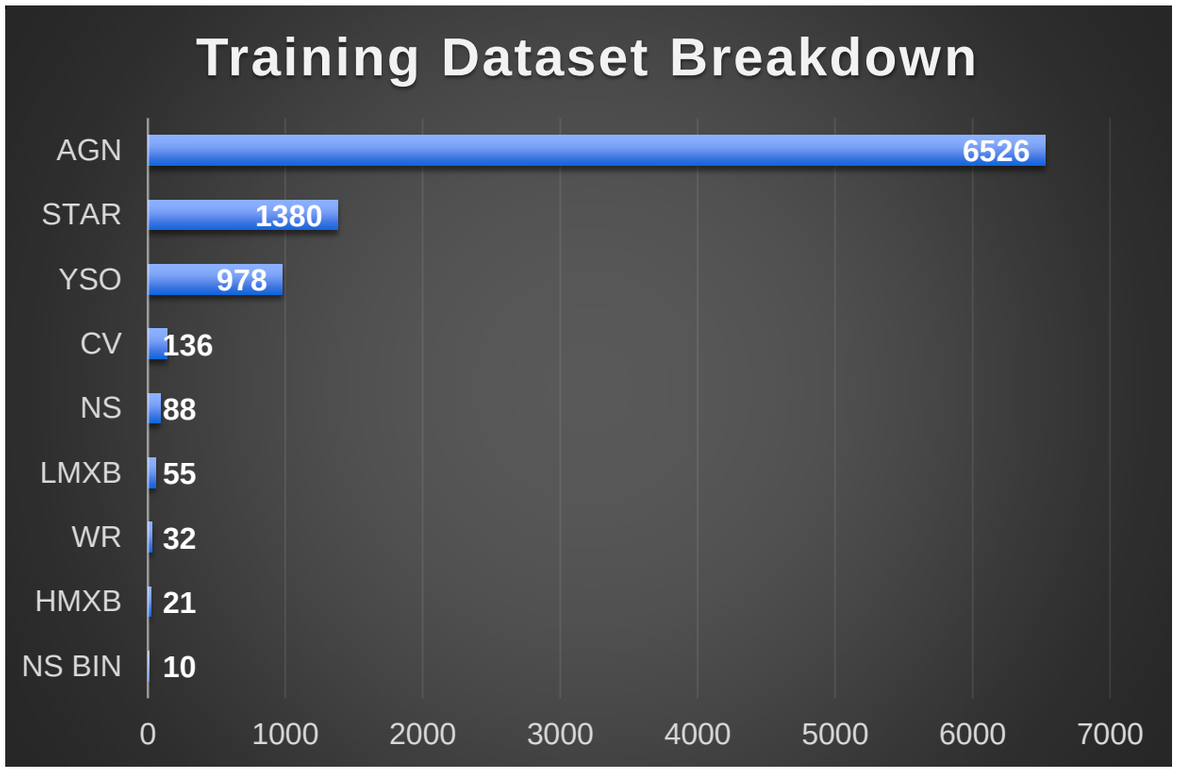
Confusion Matrix for Random Forest



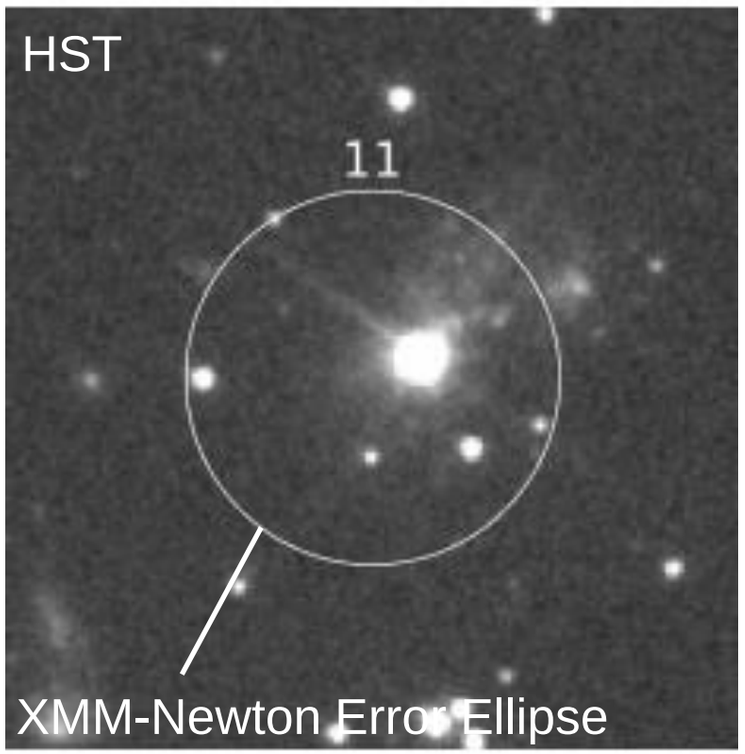
Numbers correspond to the fraction of true class sources classified as the predicted class

Current Issues

- Heavily Imbalanced Training Dataset
- Errors on features
- Confusion between sources and counterparts

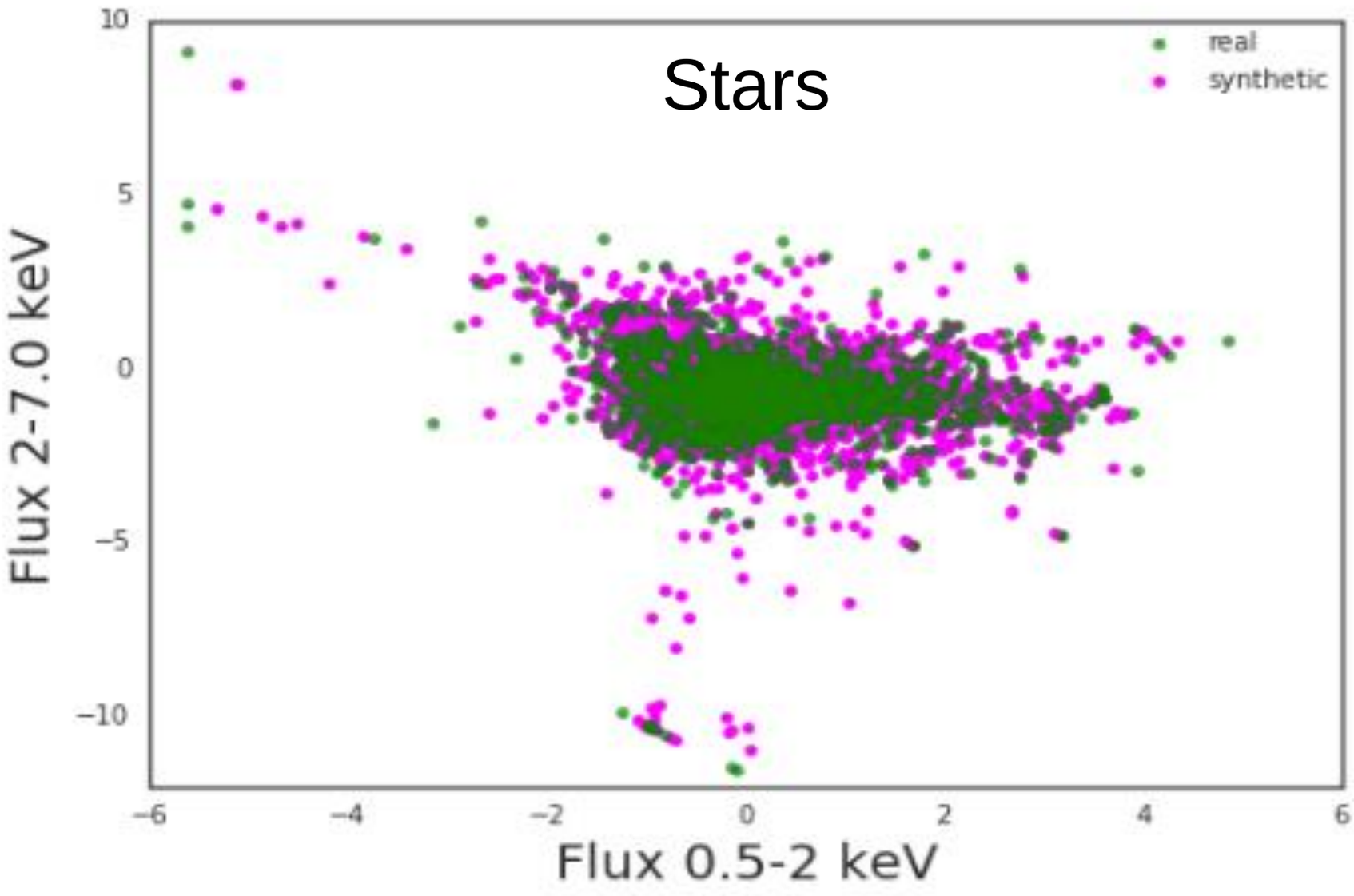


Sonbas et al. 2016



Imbalanced Training Data

- Currently we use the Synthetic Minority Over Sampling Technique (SMOTE; Chawla et al. 2011)

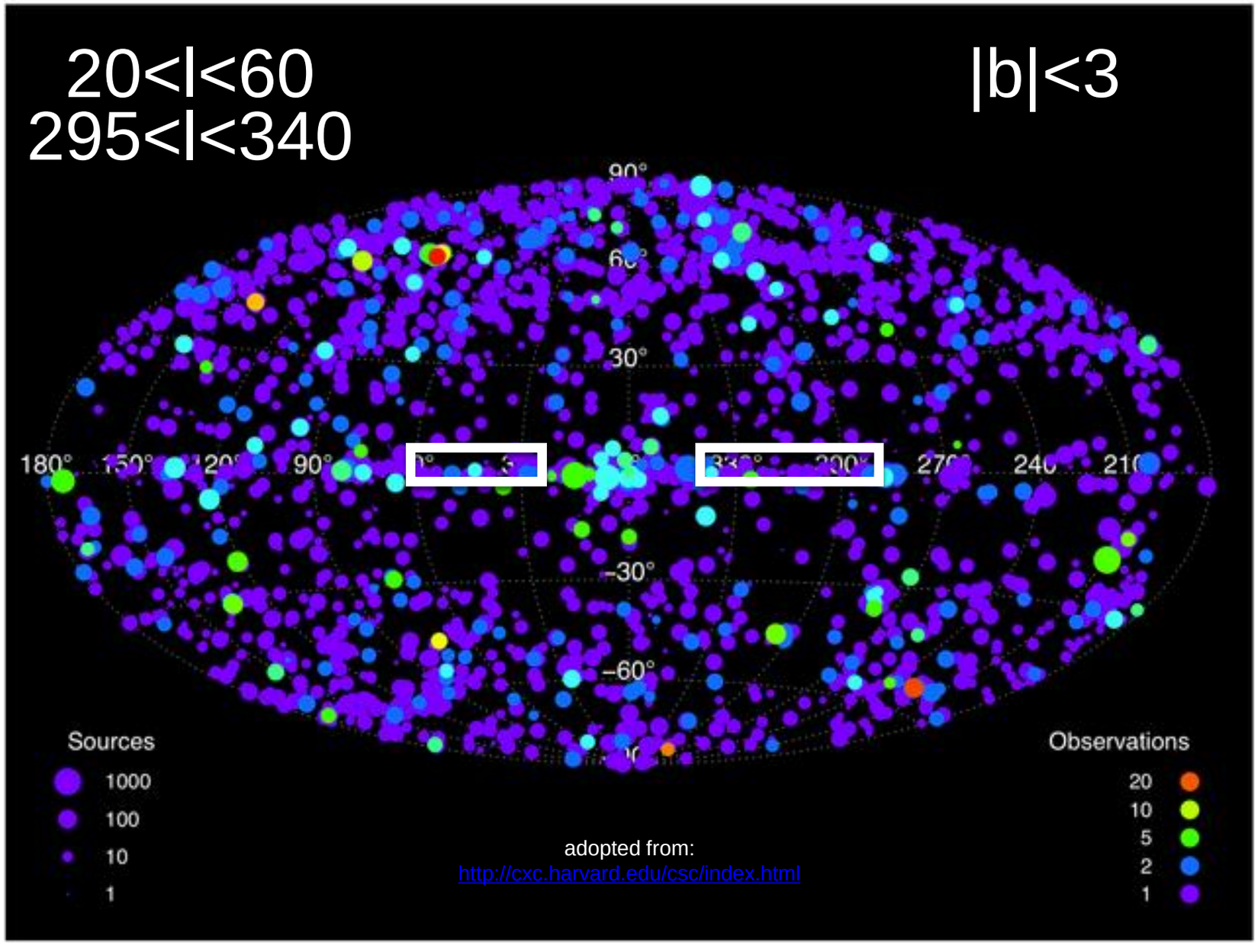


Biases and Missing Data

- Most of the classified AGN are off of the Galactic plane
- Must account for reddening through the plane
- Currently use a flag value for missing data

Chandra Source Catalog

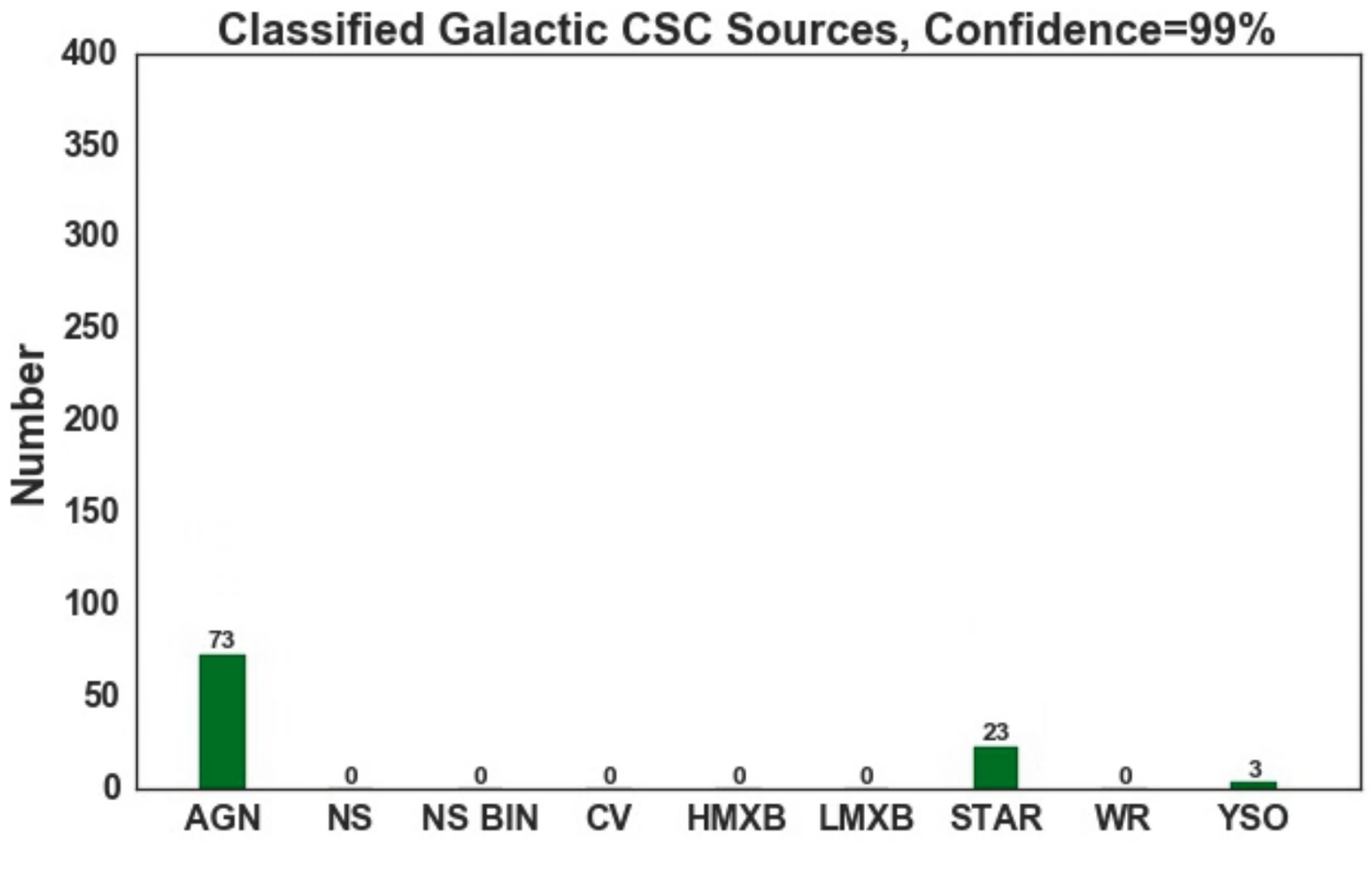
- Currently we are working to classify ~1,100 unclassified sources in the galactic plane



adopted from:
http://cxc.harvard.edu/csc/index.html

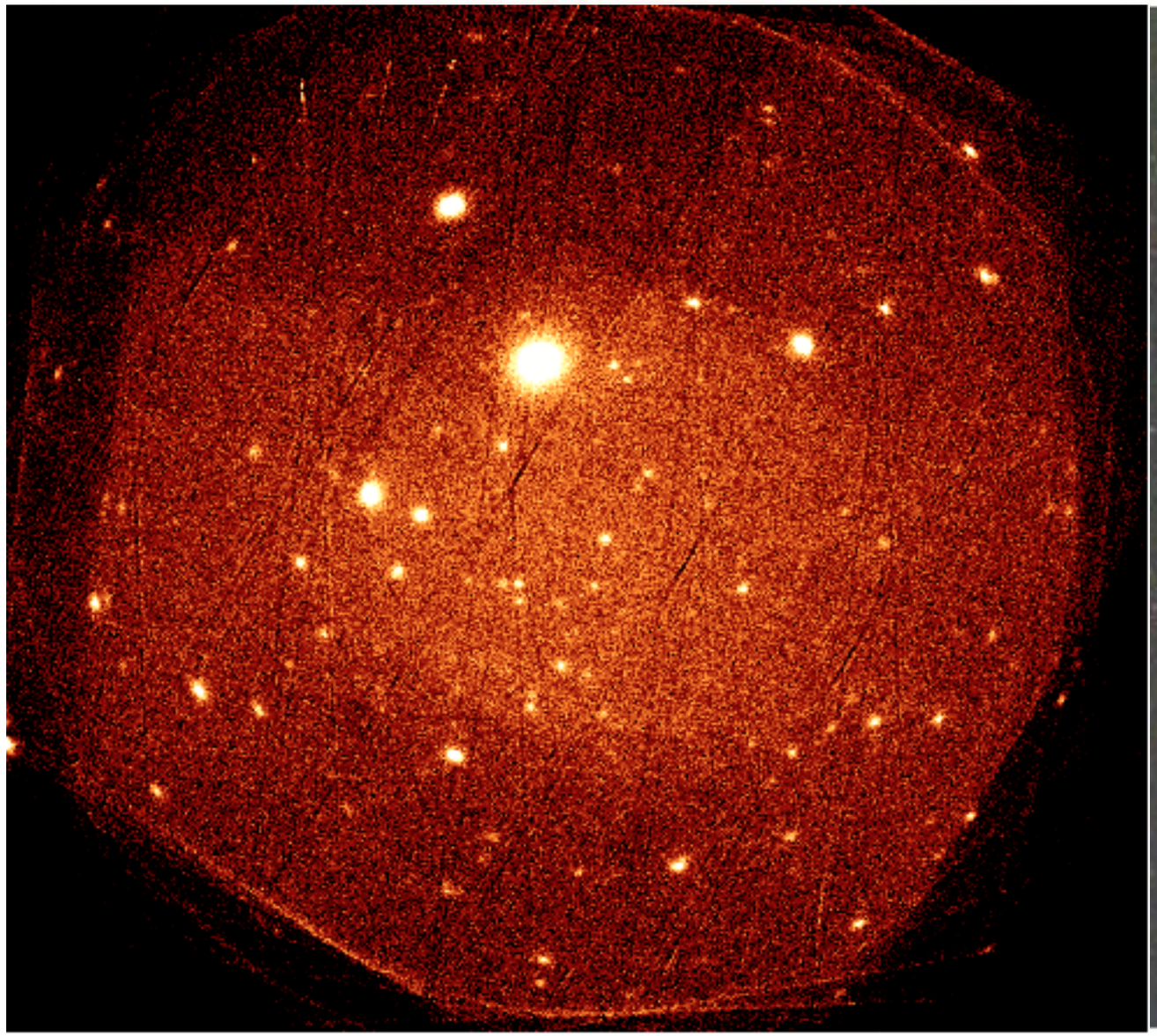
Preliminary Results

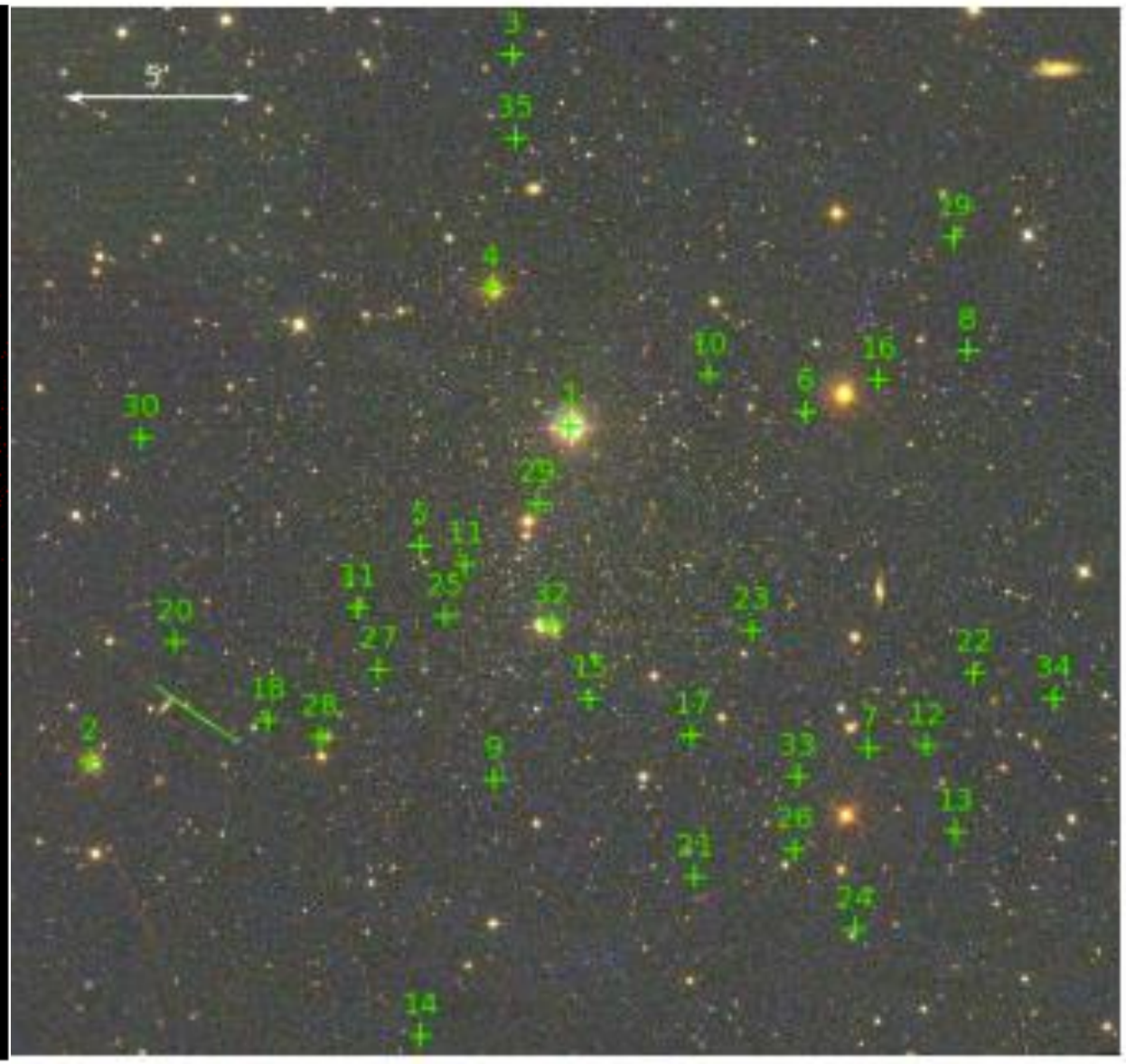
Class	P_{AGN}	P_{NS}	P_{NSBIN}	P_{CV}	P_{HMXB}	P_{LMXB}	P_{STAR}	P_{WR}	P_{YSO}	Source Name (CXO)	RA	DEC
STAR	0.0	0.0	0.0	0.0	0.02	0.0	0.89	0.06	0.03	CXO J185807.9+020411	284.53322509	2.06995874
STAR	0.03	0.0	0.0	0.12	0.0	0.03	0.82	0.0	0.0	CXO J184203.8-052331	280.51606568	-5.39194554
STAR	0.0	0.0	0.0	0.0	0.0	0.0	0.91	0.02	0.07	CXO J192318.7+140748	290.828027193	14.1302255582
AGN	0.97	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.02	CXO J163939.3-484513	249.91385475	-48.75383928
AGN	0.98	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	CXO J193938.9+213337	294.91229638	21.56052764
STAR	0.0	0.0	0.0	0.0	0.0	0.01	0.8	0.03	0.16	CXO J164044.3-485101	250.18497892	-48.8504673
AGN	0.99	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	CXO J185812.4+020049	284.55194378	2.01373768
CV	0.0	0.0	0.0	0.65	0.05	0.08	0.22	0.0	0.0	CXO J194015.9+213513	295.06635999	21.58709821
LMXB	0.09	0.01	0.0	0.0	0.0	0.89	0.01	0.0	0.0	CXO J183316.2-102341	278.317621267	-10.3949345091



Draco

Sonbas et al. 2016



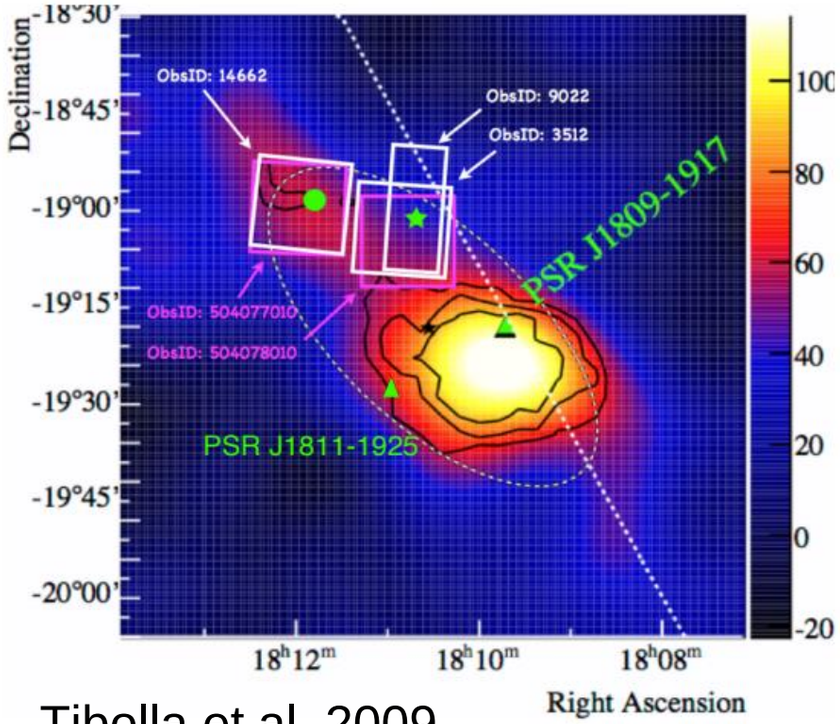


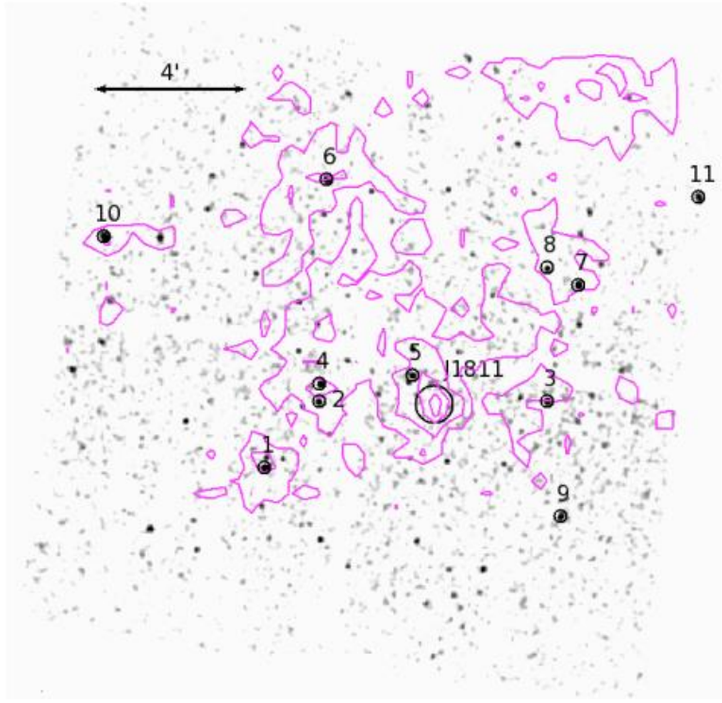
Found several possible qLMXB/CVs in Draco

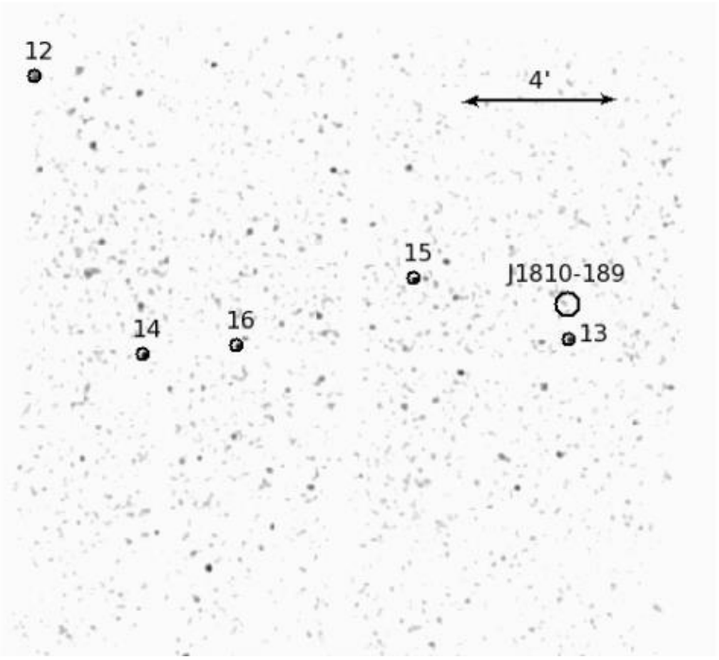
Classifications matched well with manual efforts by other teams

HESS J1809-193 & HESS J1741-302

Aharonian et al 2007

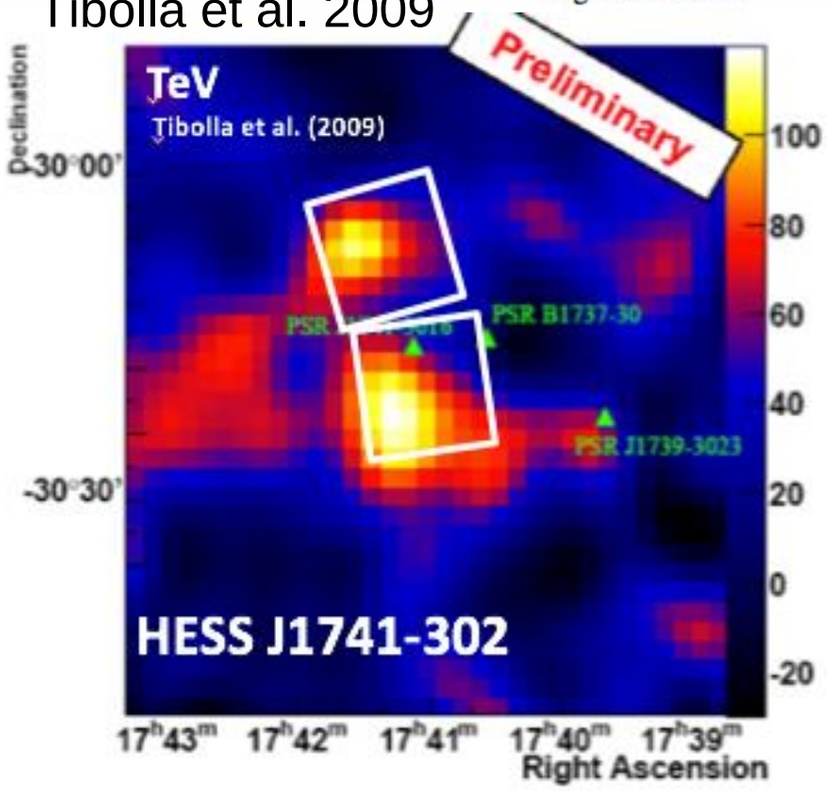




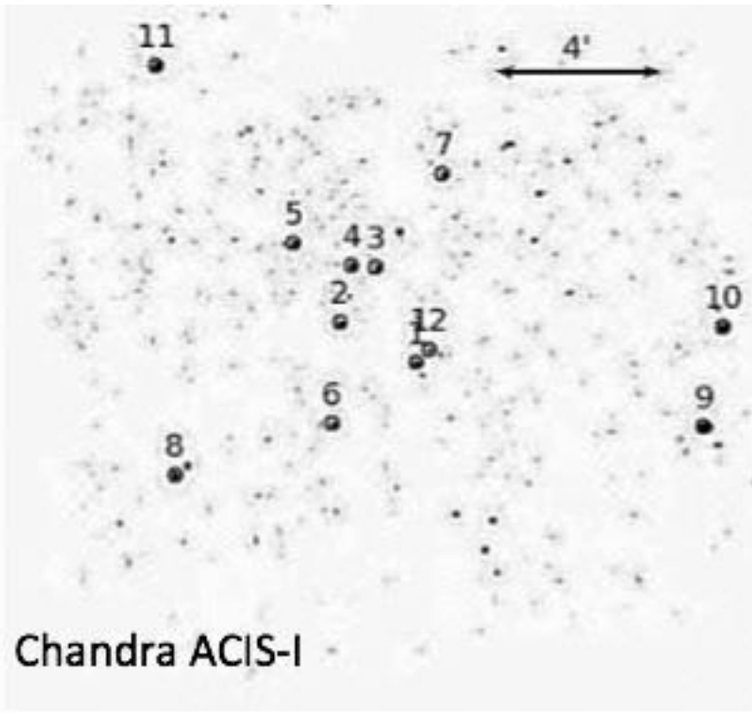


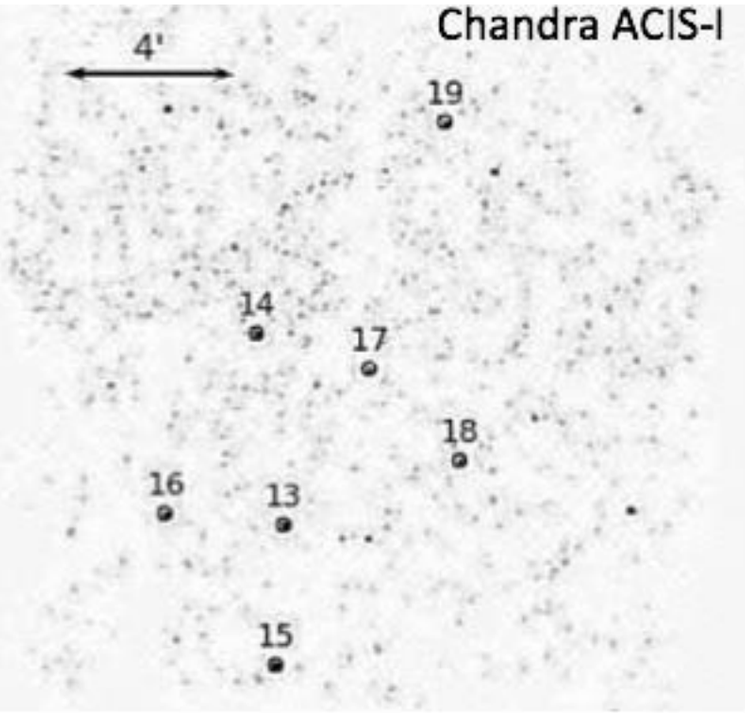
Rangelov et al. 2014

Tibolla et al. 2009



Hare et al. 2016

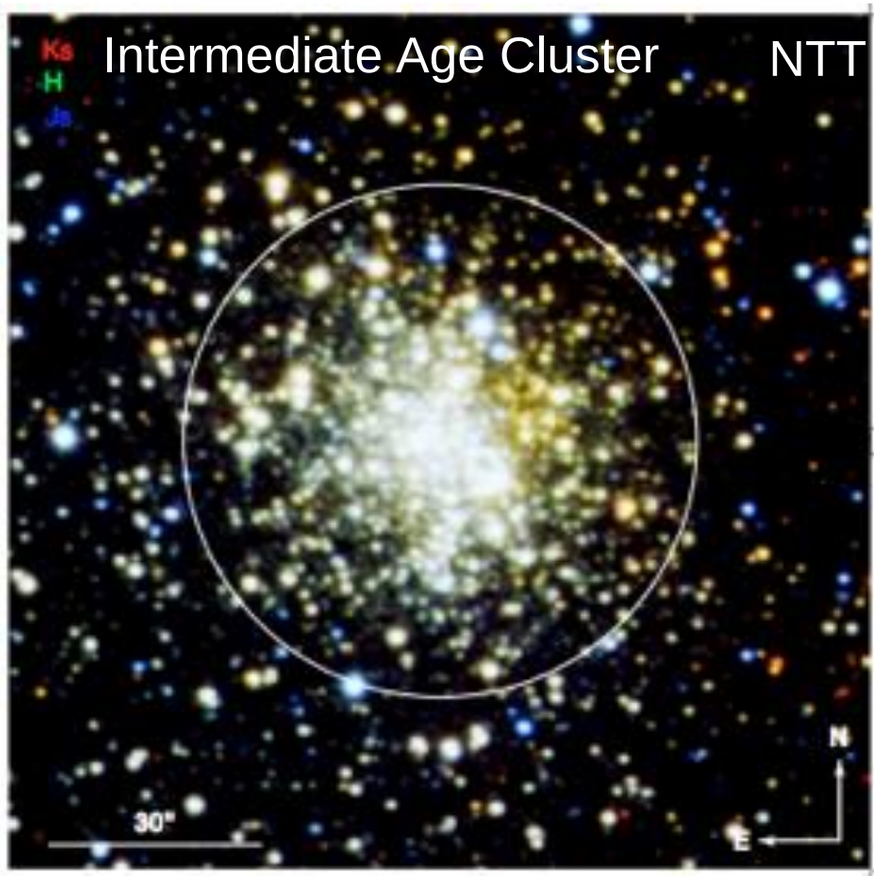




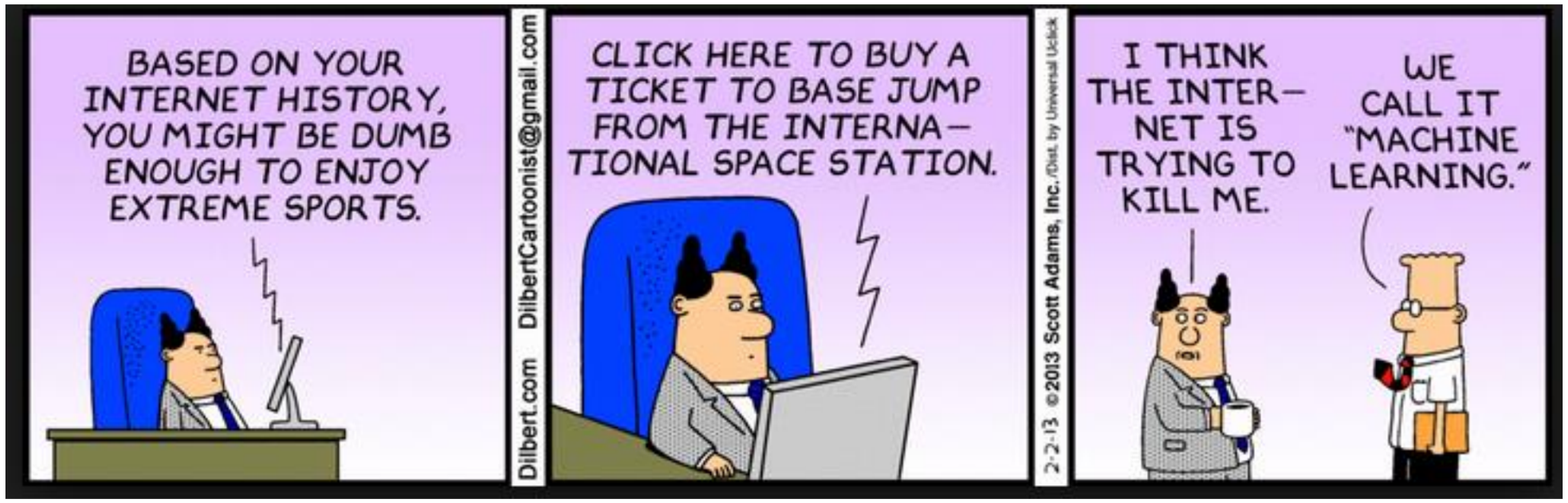
Future Work

- Add a more sophisticated cross-matching technique
- Include confusion into classification confidence in a robust way
- Include new catalogs to increase the number of underrepresented source types
- New MW features (e.g., radio, x-ray variability)
- Allow for use of catalogs that do not have full sky coverage (e.g., SDSS)

Pooley et al. (2007)



Thank You!


BASED ON YOUR INTERNET HISTORY, YOU MIGHT BE DUMB ENOUGH TO ENJOY EXTREME SPORTS.
CLICK HERE TO BUY A TICKET TO BASE JUMP FROM THE INTERNA— TIONAL SPACE STATION.
I THINK THE INTER— NET IS TRYING TO KILL ME.
WE CALL IT "MACHINE LEARNING."
DilbertCartoonist@gmail.com
Dilbert.com
2-2-13 ©2013 Scott Adams, Inc. /Dist. by Universal Uclick