Classification of Serendipitous Chandra Source Catalog Sources Using a Multiwavelength Machine Learning Approach

Jeremy Hare (NASA GSFC/CUA/CRESST II)

25 years of Science with Chandra Symposium

Dec. 5, 2024

Collaborators

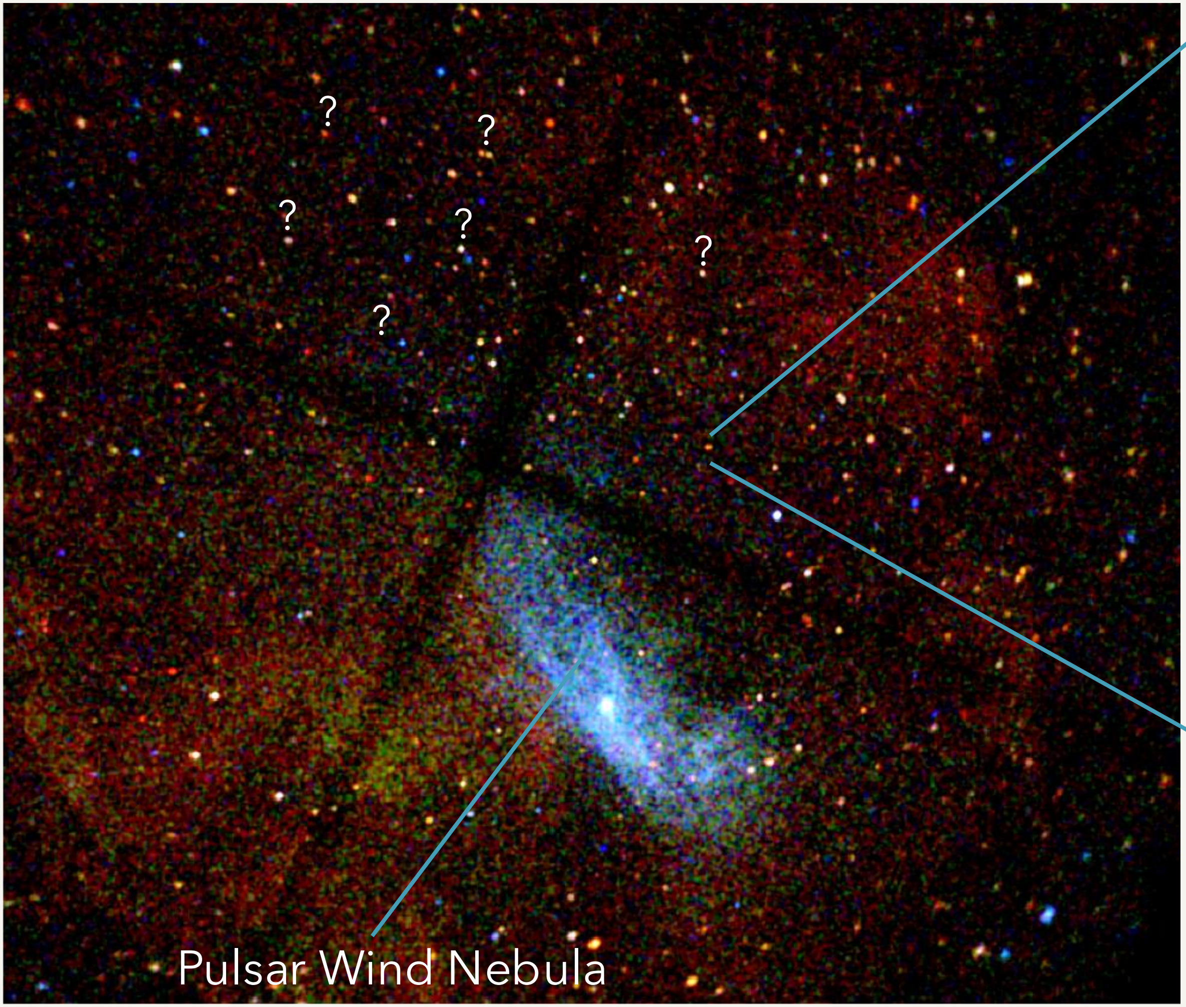
- Oleg Kargaltsev (GWU)
- Hui Yang (GWU/IRAP)
- Steven Chen (GWU)
- Yichao Lin (GWU)
- Igor Volkov (GWU)

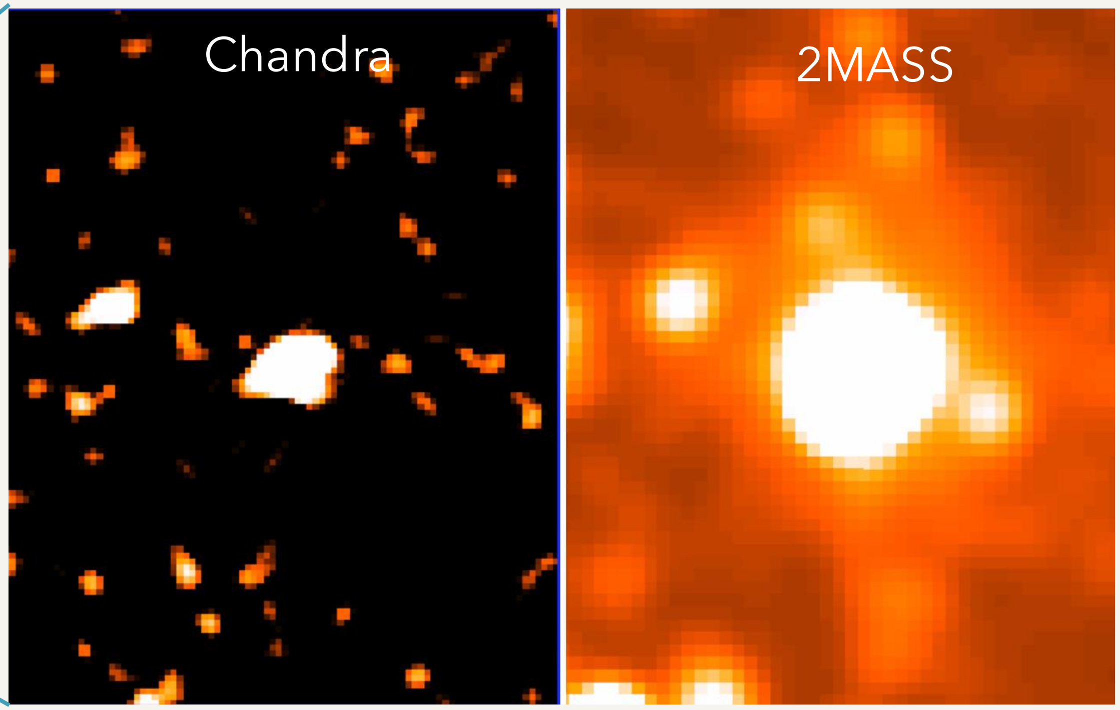
THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Motivation: Serendipitous Sources

How can we find more rare Galactic objects in X-ray catalogs?

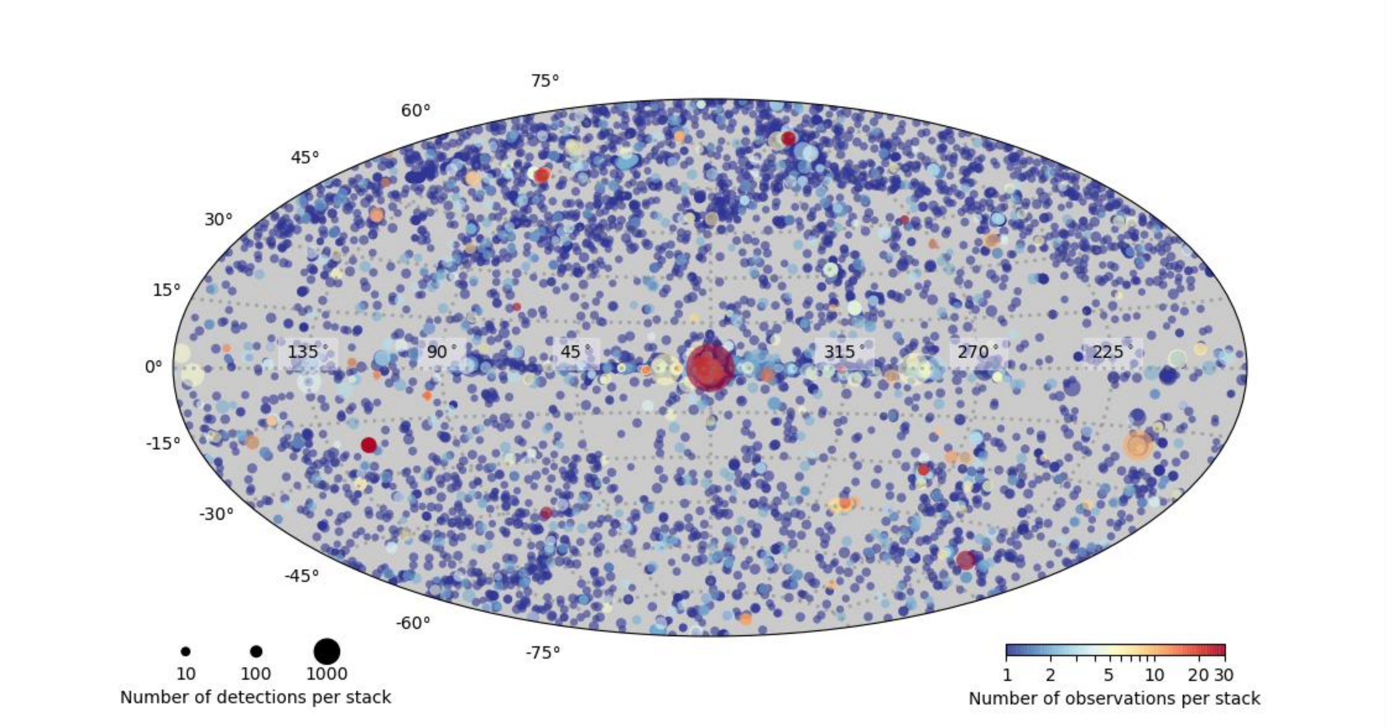




Has a multi-wavelength counterpart
Classified as a star

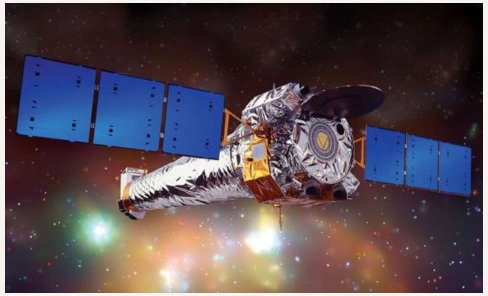
Motivation: Source Catalogs

- Most sources in the images of archived X-ray observations are serendipitously observed and their nature remains unknown
- Chandra Source Catalog version 2.1 contains about 410,000 unique sources (data up to end of 2021)
- 4XMM-DR13 catalog contains almost 660,000 unique sources



- How can we locate the most interesting sources?

XMM-Newton
0.5-10 keV

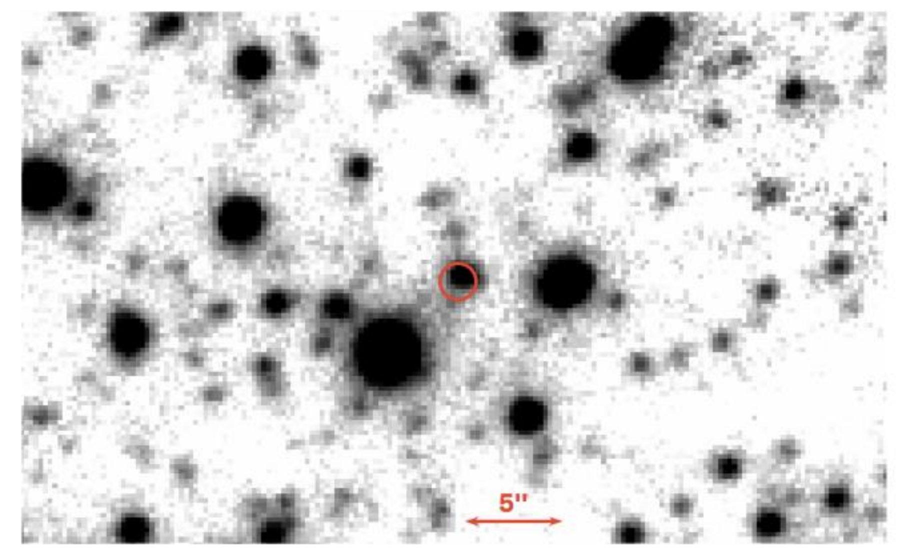


Chandra
0.5-8 keV

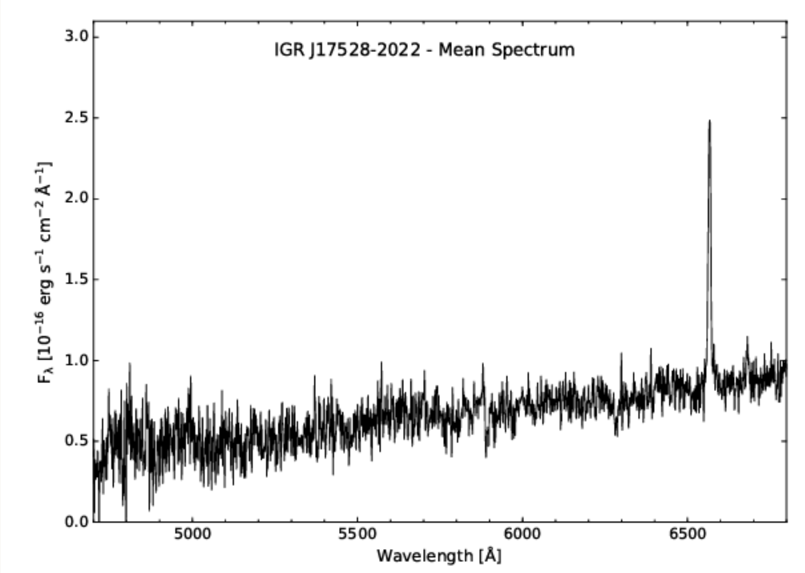
Motivation: Classifying bright X-ray sources

- Often need to locate the correct multi-wavelength counterpart
- Gather data from various multi-wavelength observing campaigns (e.g., 2MASS, VPHAS+, Gaia)
- Obtain optical/NIR spectra to identify nature of source
- Costs time and resources

Pan-STARRs y-band

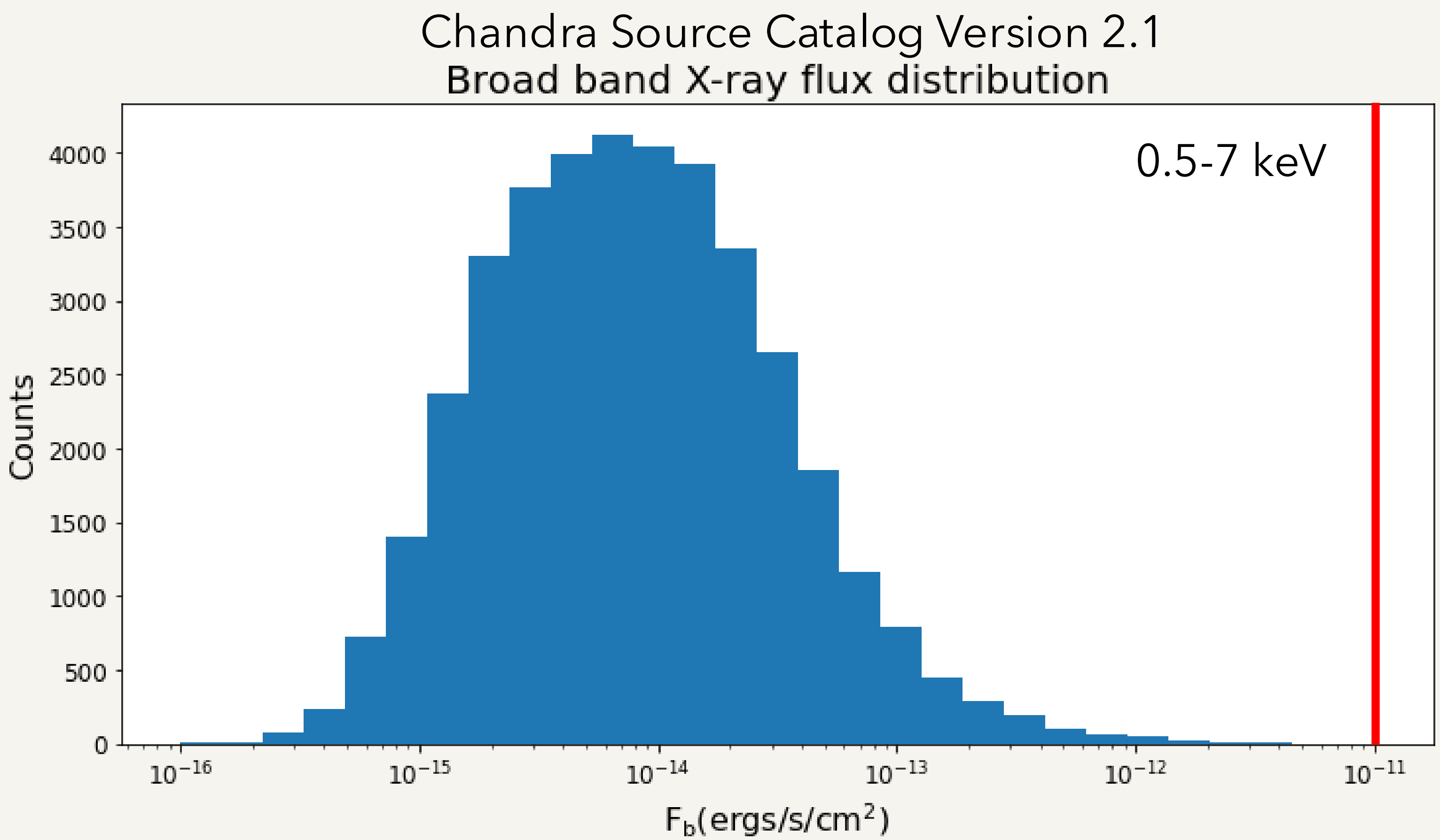


MDM optical spectrum



Hare et al. (2021)

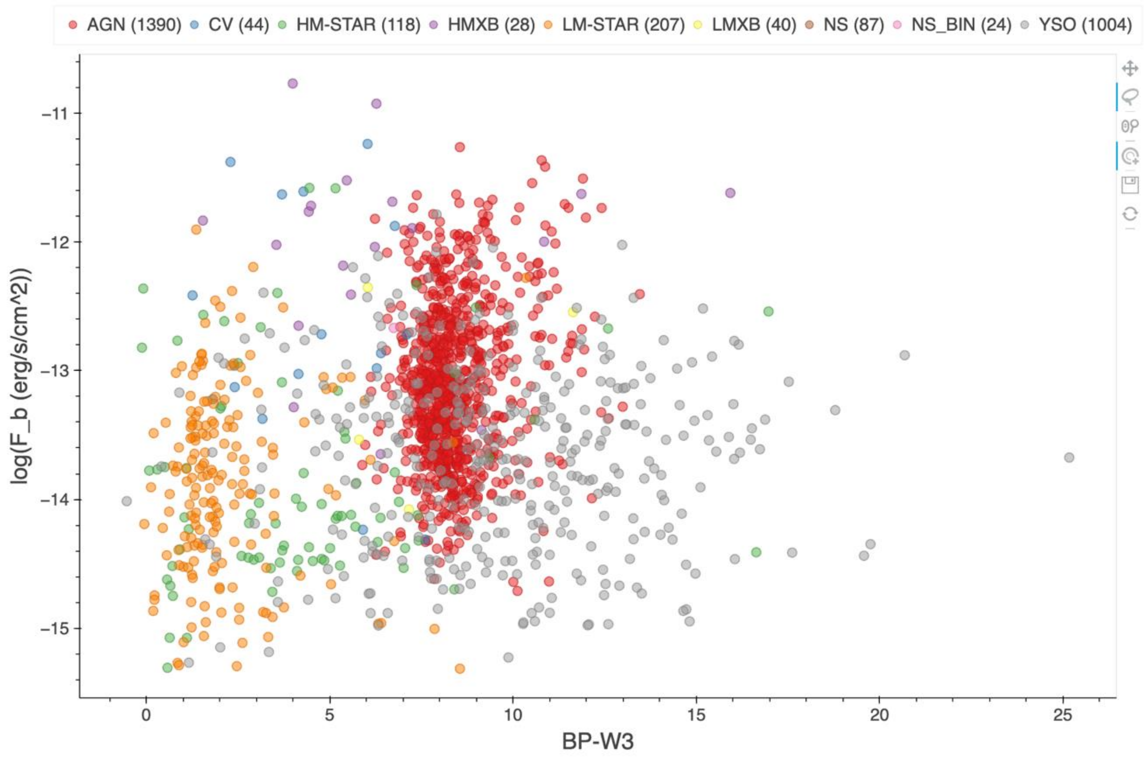
Motivation: Most sources are faint!

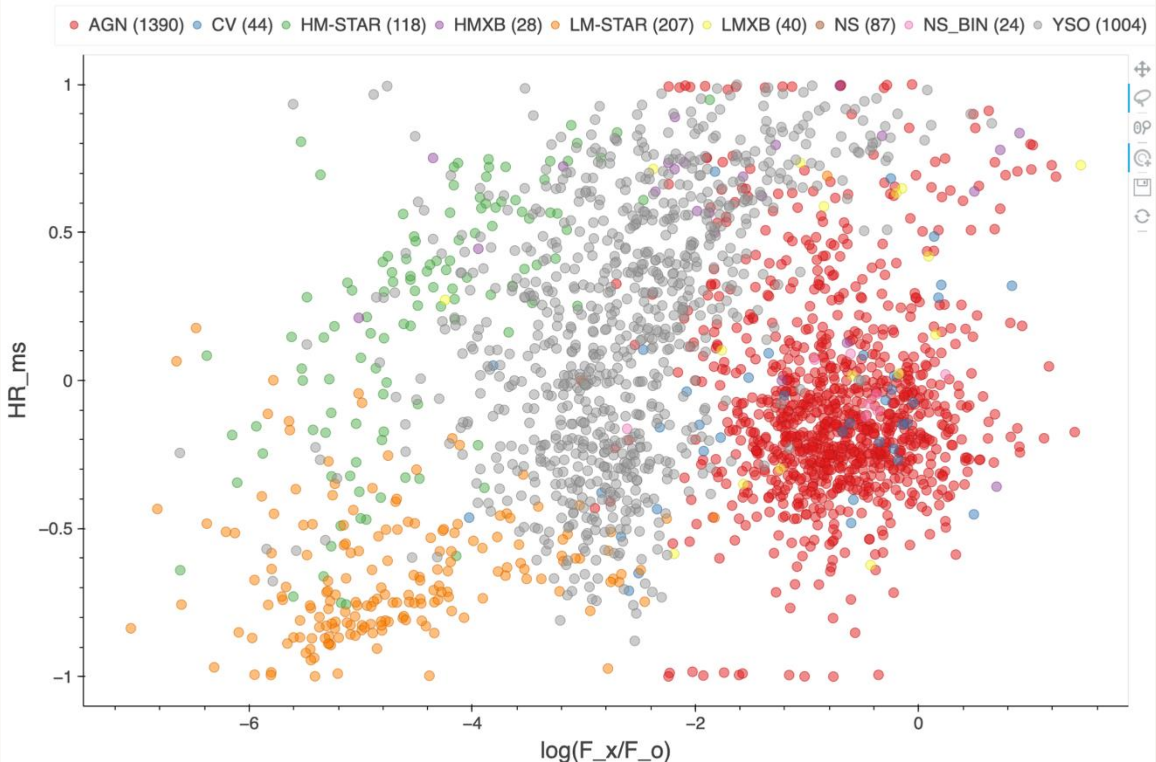


Motivation: Faint X-ray sources

- Multi-wavelength data is critical for classifying faint sources
- Manual classification often consists of looking at various 2D parameter plots to separate classes
- Limited to 2-3 dimensions and no rigorous way to assign a confidence to a given classification

Plots made using XCLASS; see Yang et al. (2021)



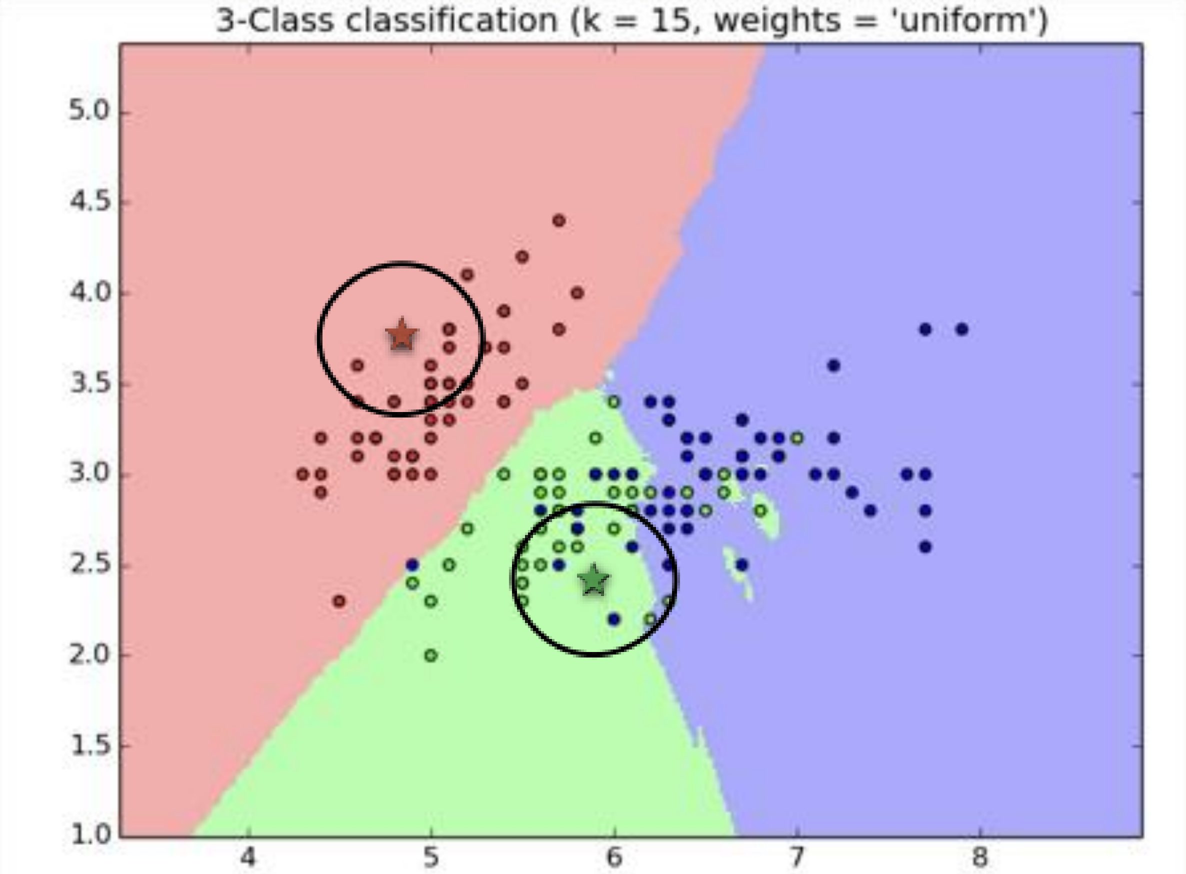


Solution: Machine Learning

- Artificial intelligence may eventually one day destroy us; however, until then we can take advantage of it!
- Machine learning can be used to handle large datasets and a large number of parameters (features)
- Several previous works in this area include McGlynn et al. (2004), Broos et al. (2013), Lo et al. (2014), Farrell et al. (2015), Kerby et al. (2021), Perez-Diaz et al. (2024)

Machine Learning Overview: Training Dataset

- Training dataset shown by colored circles representing sources of **KNOWN** types
- Unknown sources are assigned a type based on training data in vicinity



MUWCLASS: Training Dataset

Source Type	CSCv2
active galactic nuclei (AGN)	1390
cataclysmic variables (CV)	44
high mass stars (HM-STAR)	118
high mass X-ray binaries (HMXB)	26
low mass stars (LM-STAR)	207
low mass X-ray binaries (LMXB)	65
pulsars and isolated neutron stars (NS)	87
young stellar objects (YSO)	1004
Total	2941

Yang, Hare, et al. 2022

Red back and black widow MSPs in LMXB class

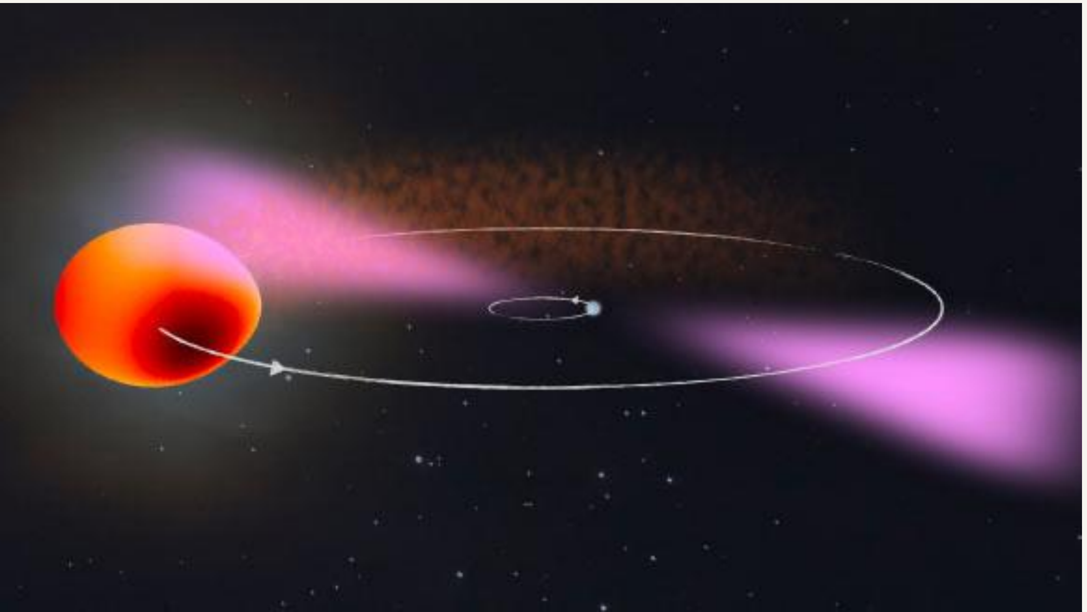


Image credit: B. Knispel / C.J. Clark / Max Planck Institute for Gravitational Physics / NASA's Goddard Space Flight Center.

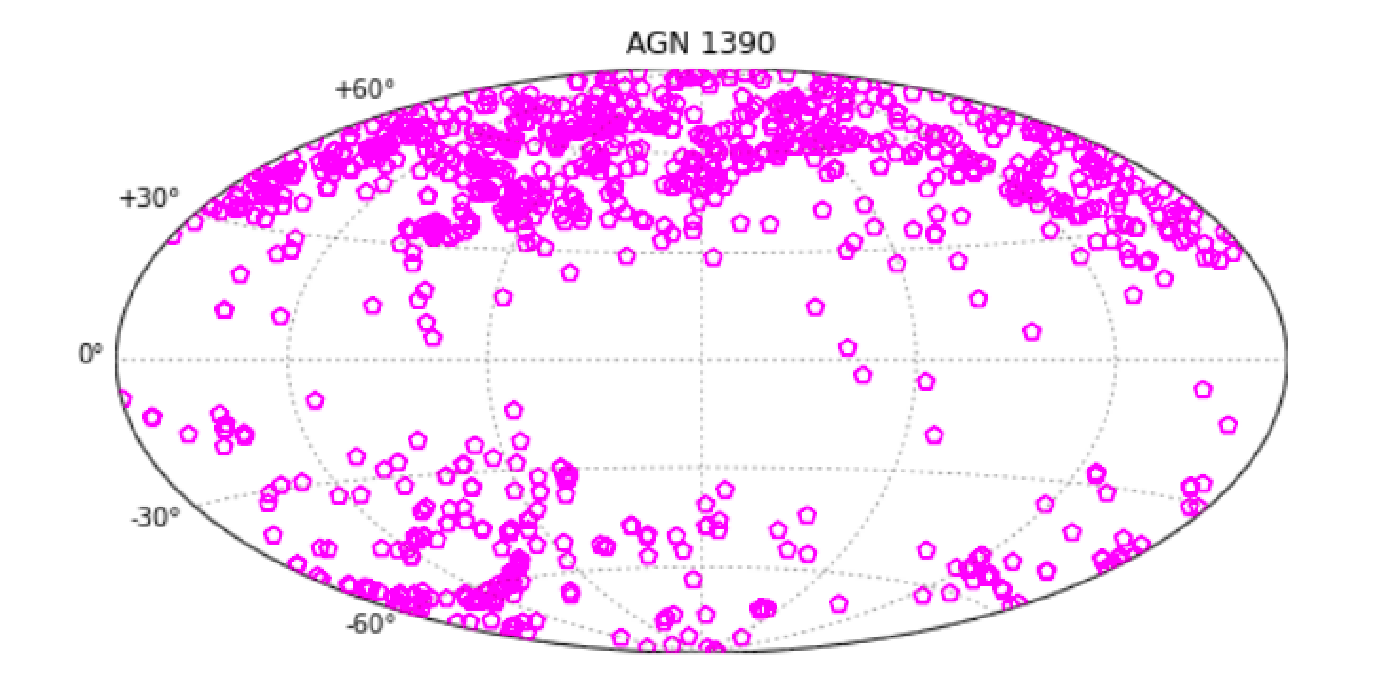
MUWCLASS: Observational Biases

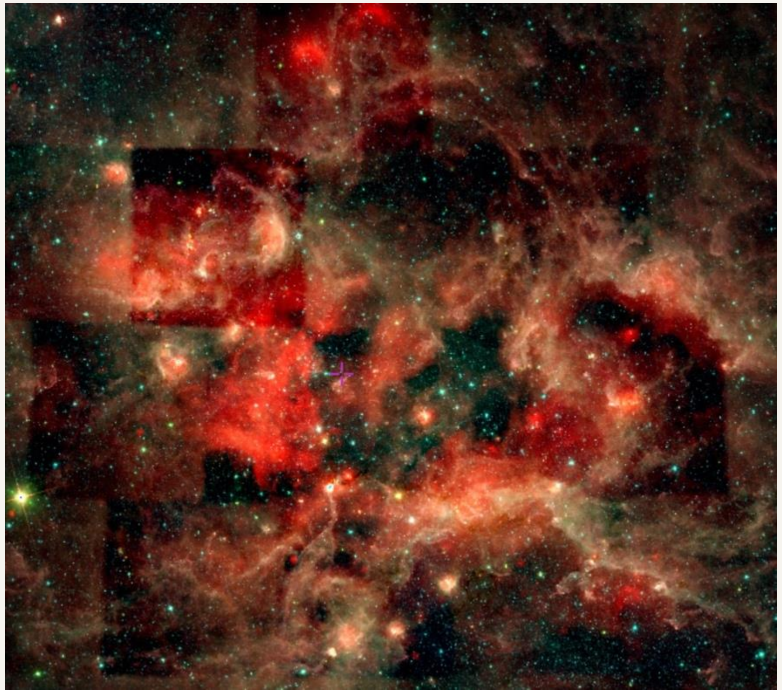
- Heavily Imbalanced Training Dataset

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

- Observational Biases

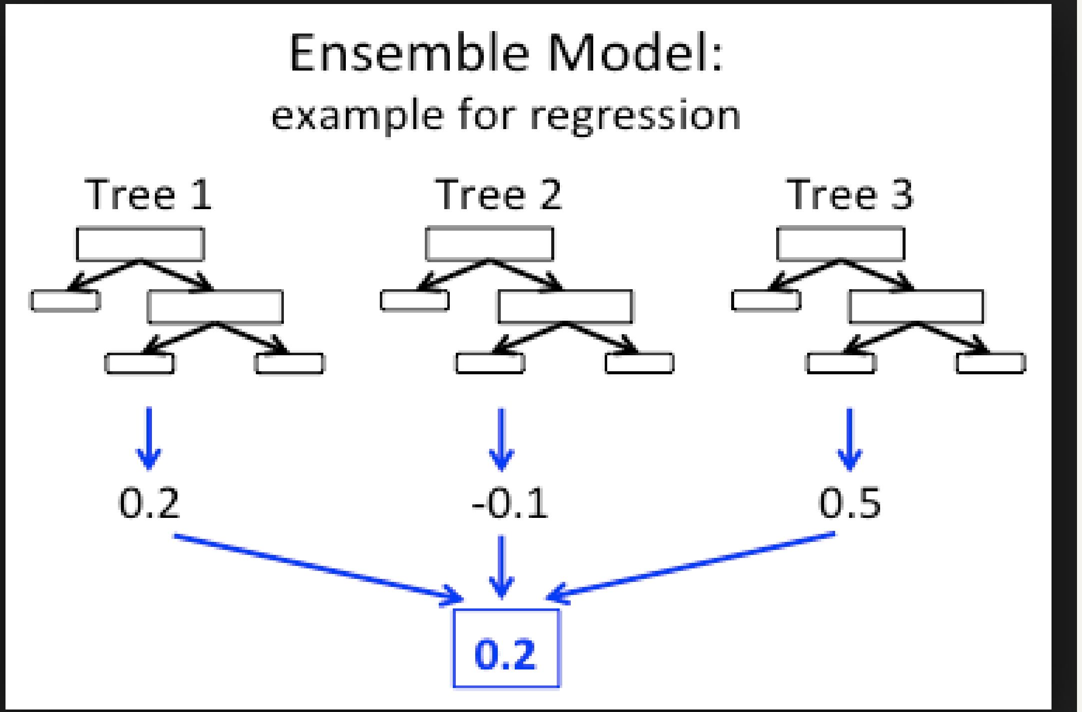
  Most AGN in training dataset are located far off the Galactic plane. We correct this bias by applying extinction to AGN based on location of sources being classified in the plane.





MUWCLASS: Random Forest

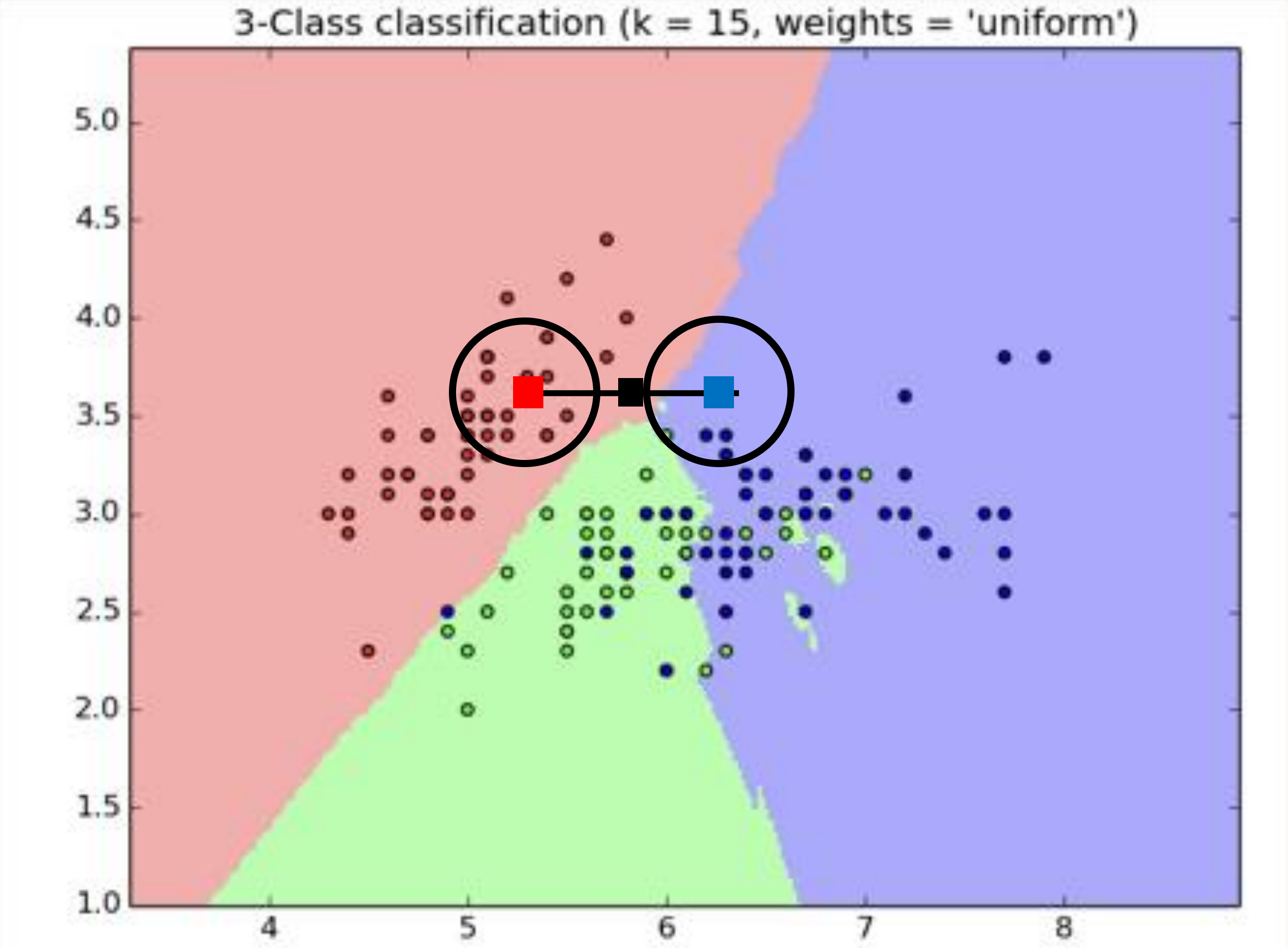
Pedregosa *et al. (2011)*



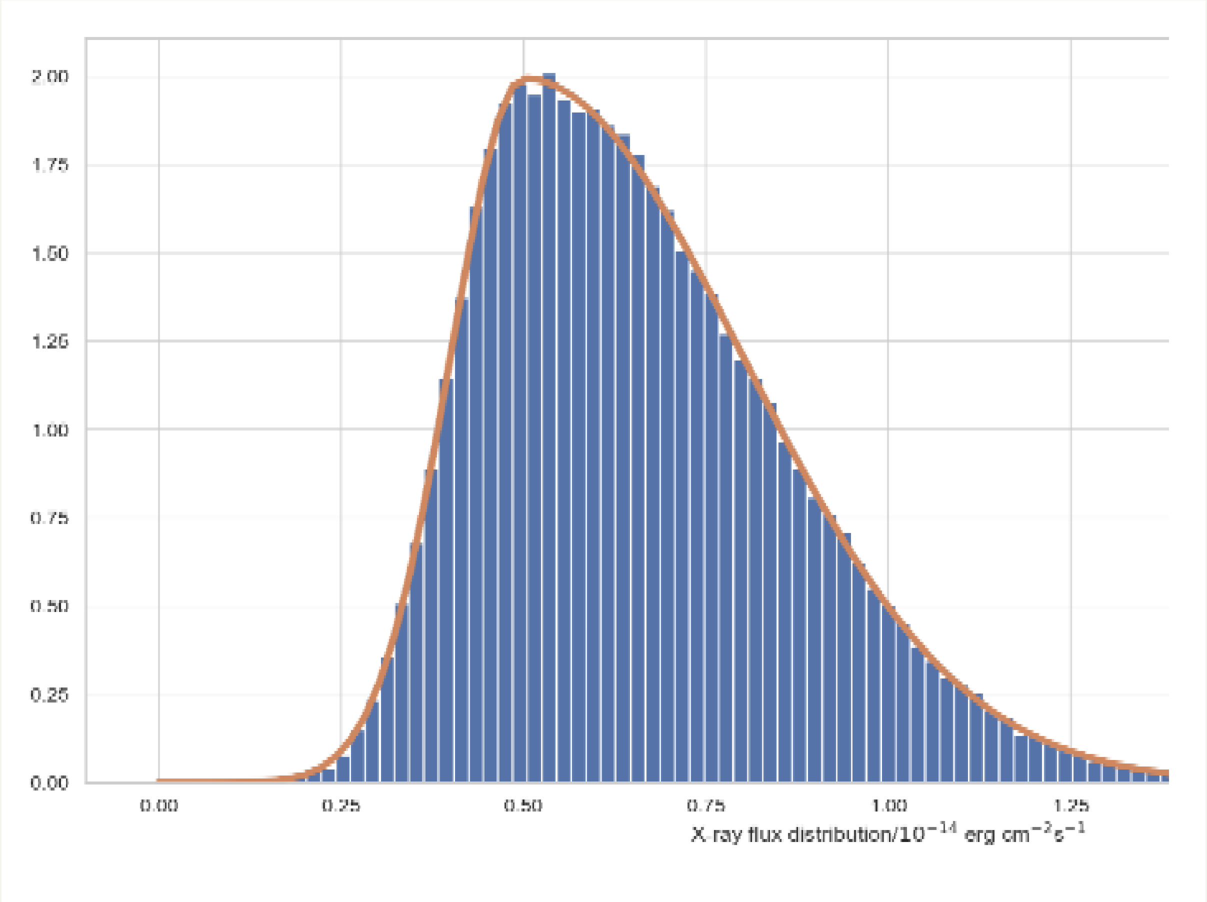
- Bootstraps training dataset
- Uses a random subset of feature at each split
- Helps reduce overfitting
- Pipeline is modular, so any algorithm from scikit learn can be used (assuming right normalization is applied)

MUWCLASS: Uncertainties on Fluxes/ Magnitudes

Red	Blue	Green
1	0	0
0	0.75	0.25



MUWCLASS: Uncertainties on Fluxes/ Magnitudes



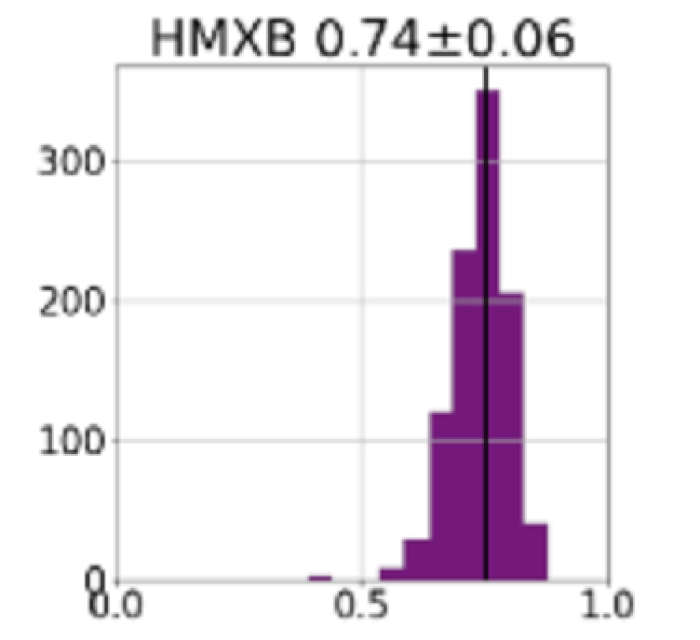
- Monte Carlo method to account for measurement uncertainties of features
- Resample all sources (both training dataset and unclassified sources) many times
- Recalculate features (e.g., Hardness ratios, colors) based on resampled features before passing to Machine Learning Algorithm

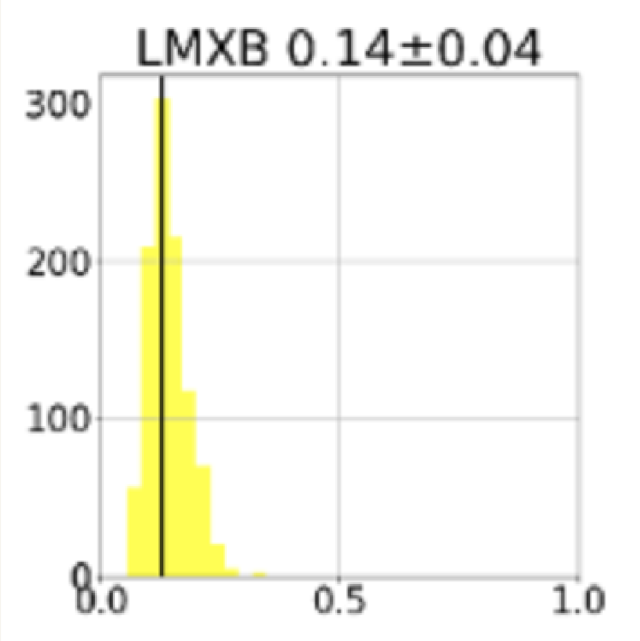
See also Probabilistic Random Forest: Reis et al. (2018)

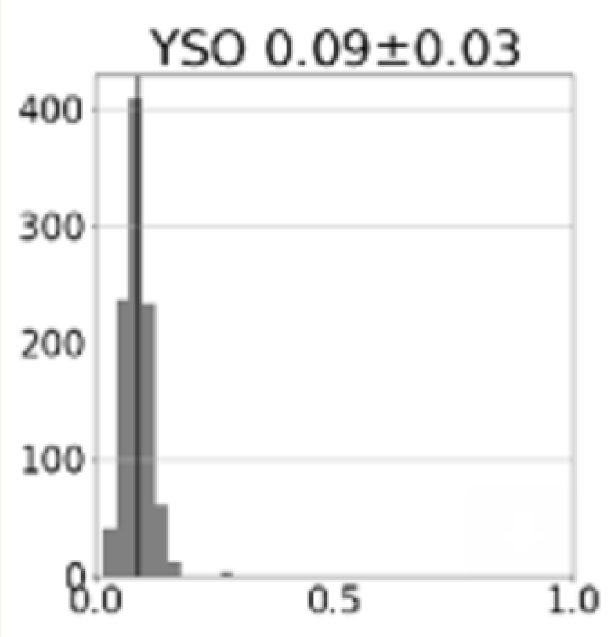
MUWCLASS: Classification distributions

- Probability distributions, instead of vectors, for classes assigned to sources

Source 1





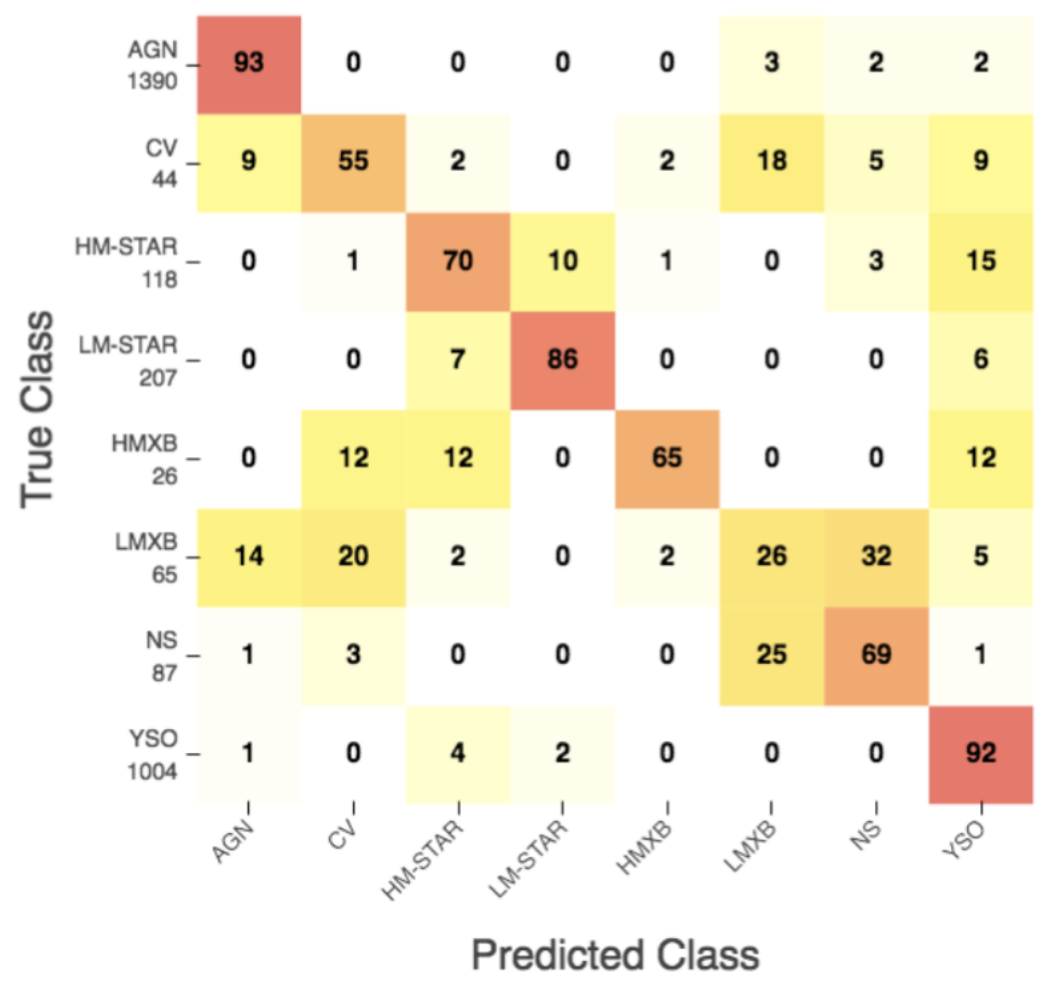


MUWCLASS: Performance Recall

Overall accuracy: 88%

Class Averaged accuracy: 70%

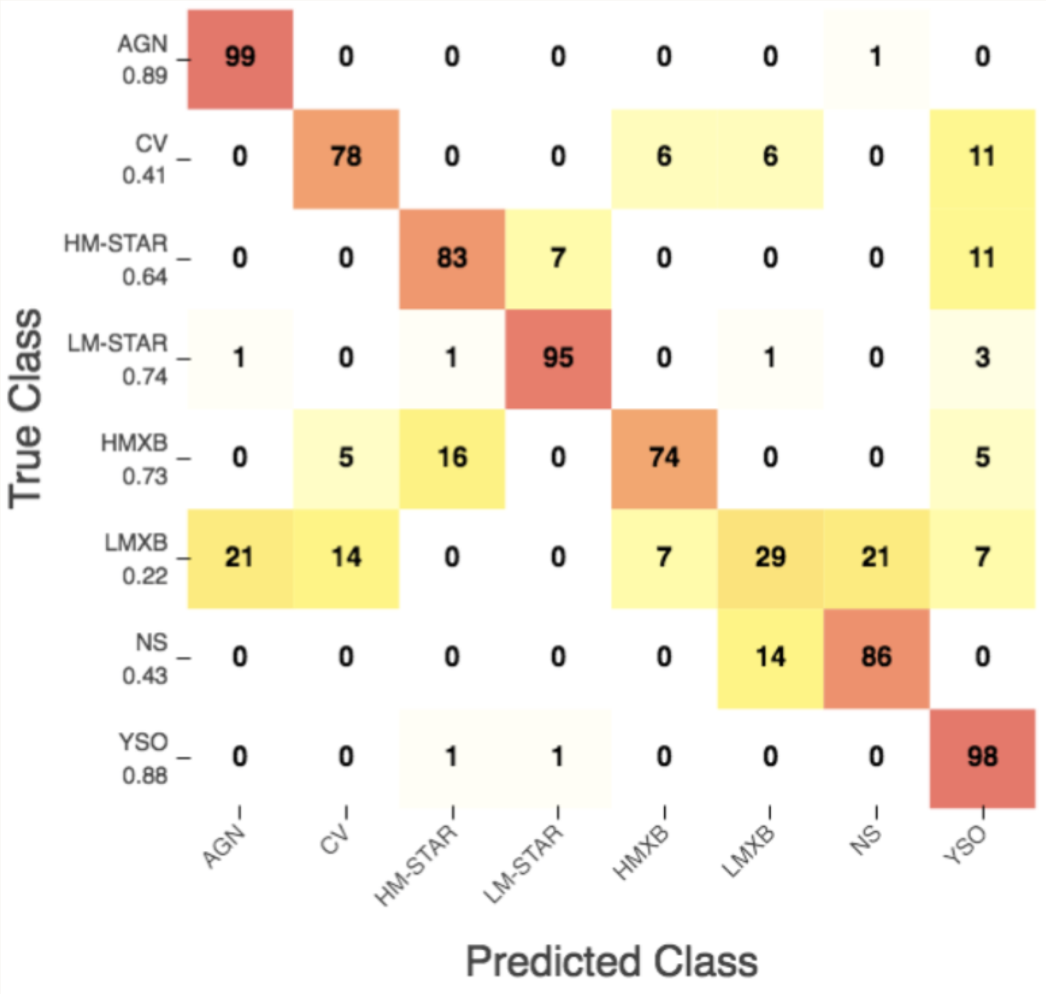
All



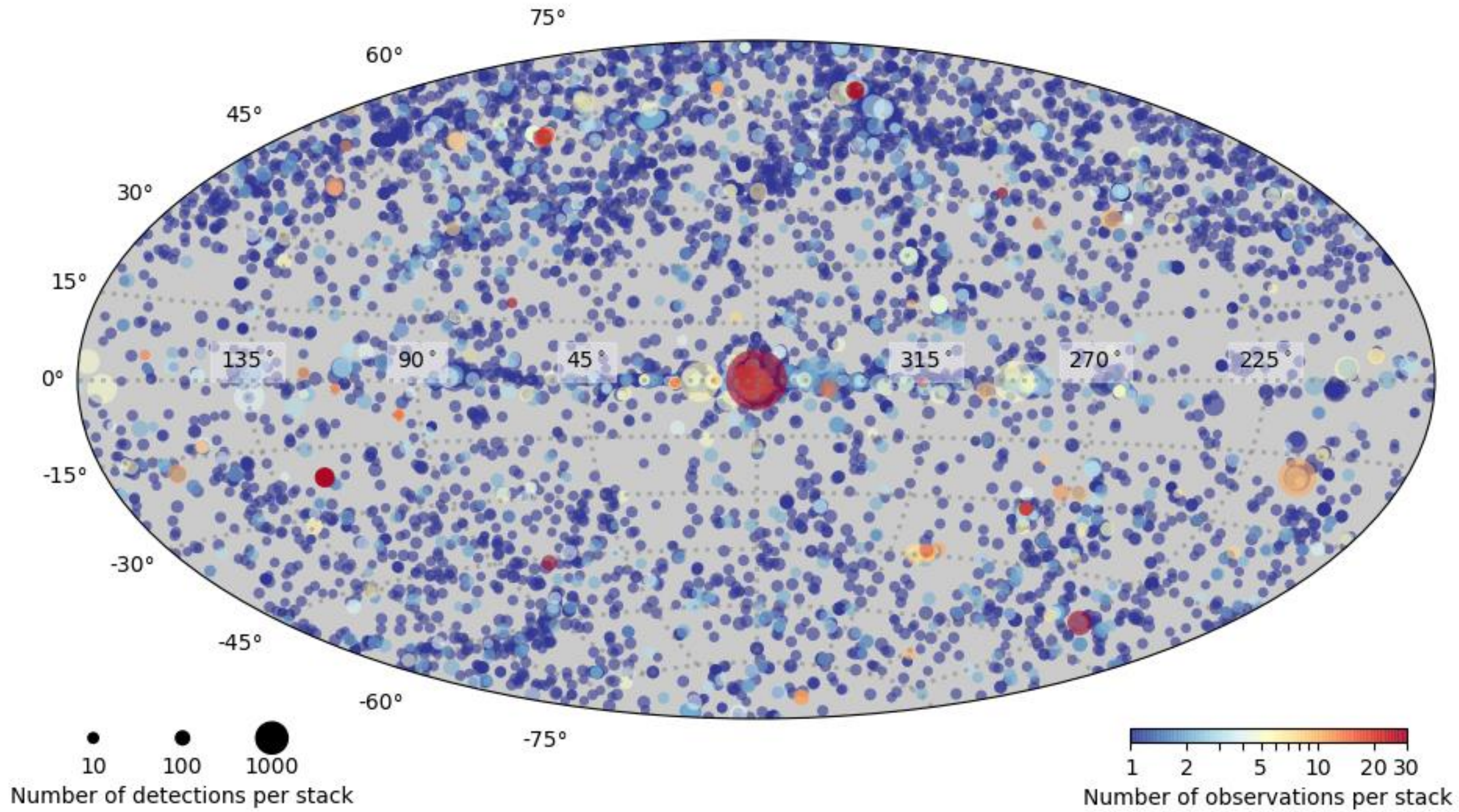
Overall accuracy: 97%

Class Averaged accuracy: 80%

Confident (~81% total)

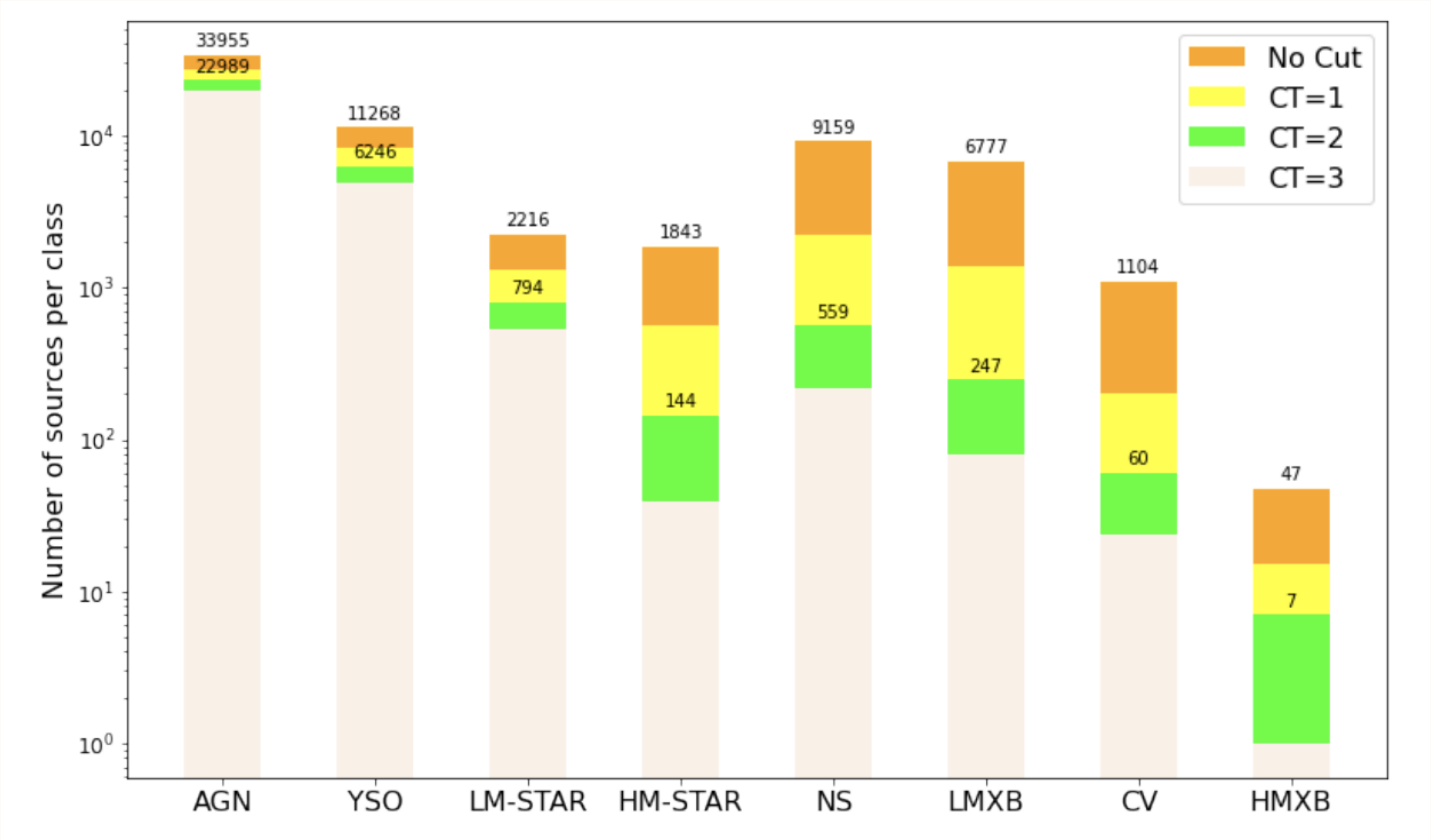


Classification of CSCv2 sources



Classification of CSCv2 sources

- Removed sources with positional uncertainties >1″ to limit source confusion
- Also removed sources sources with various CSC flags (extended/confused)
- Left with ~66,000 sources

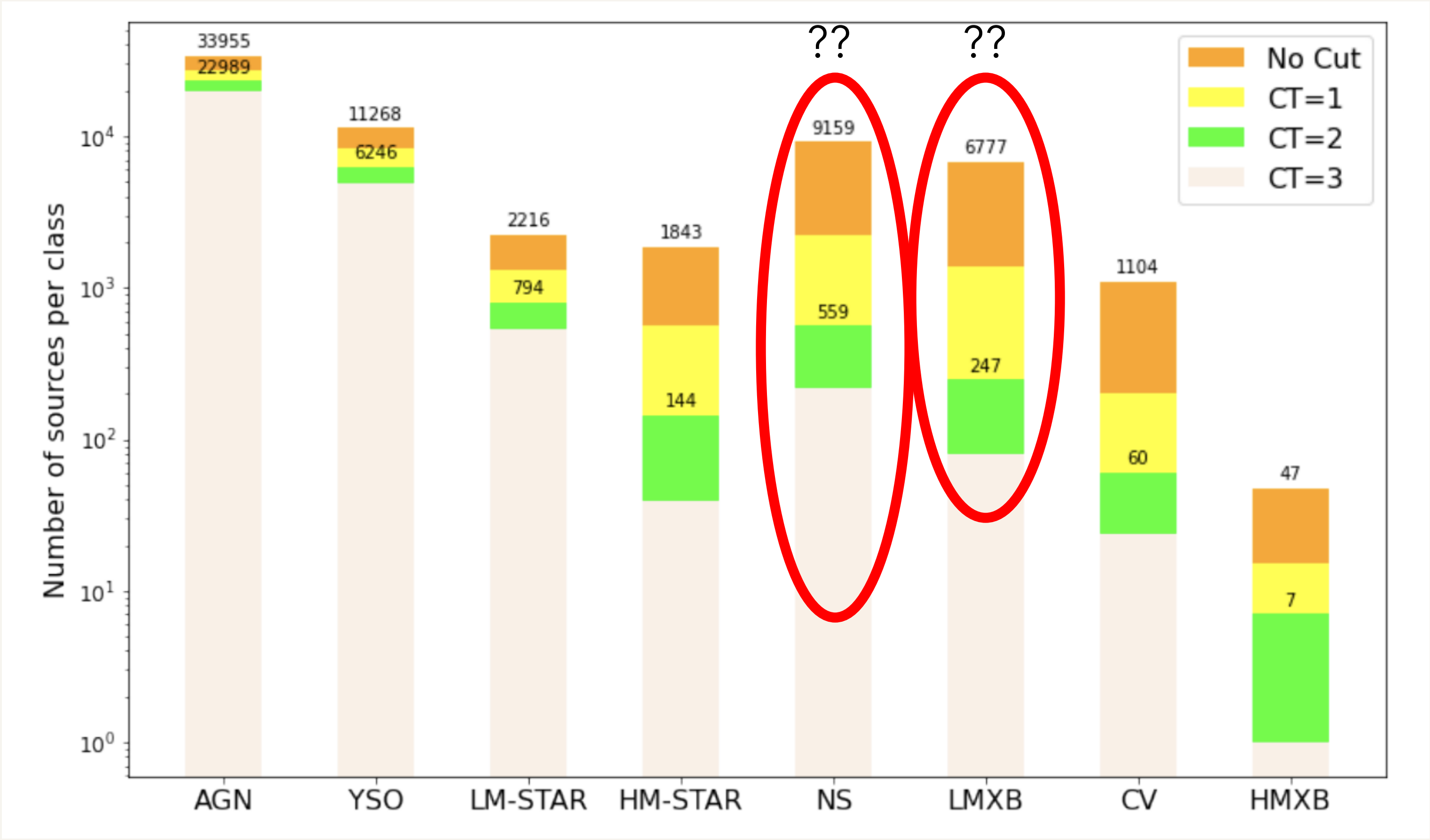


Classification of CSCv2: Simple Checks HMXBs

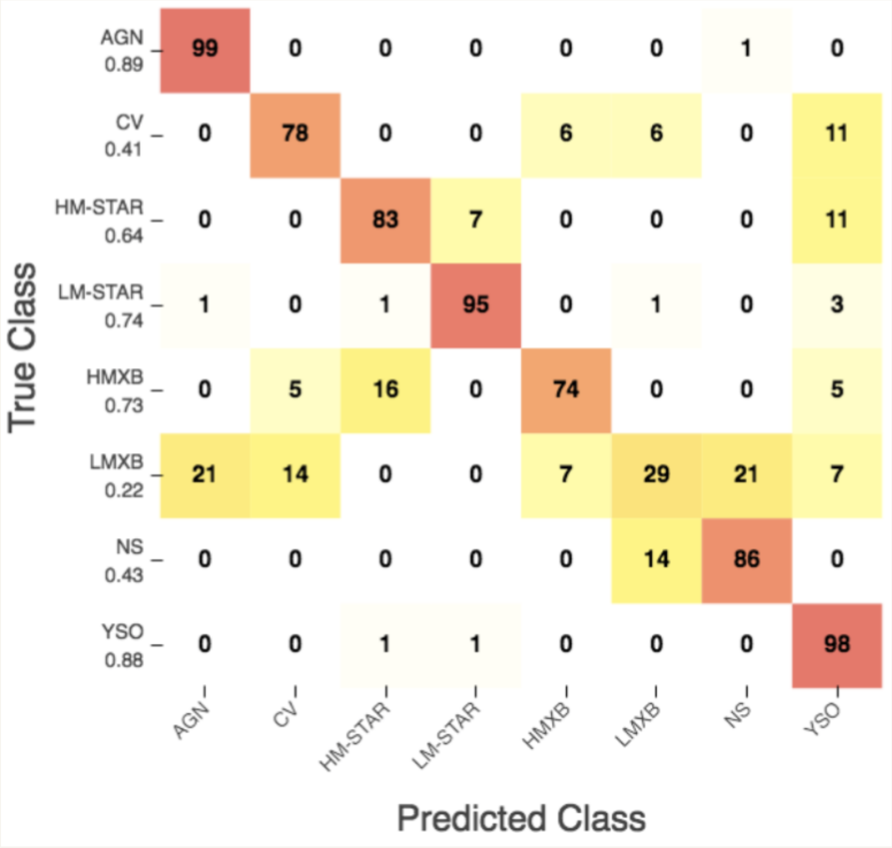
- HMXB 4U 1416-62 had a catalog position >5″ offset from the Chandra position, so was not included in our training dataset
- It was our most confidently classified HMXB with a probability of ~80%

Classification of CSCv2: Issues and Biases

- These classes can **not** be trusted



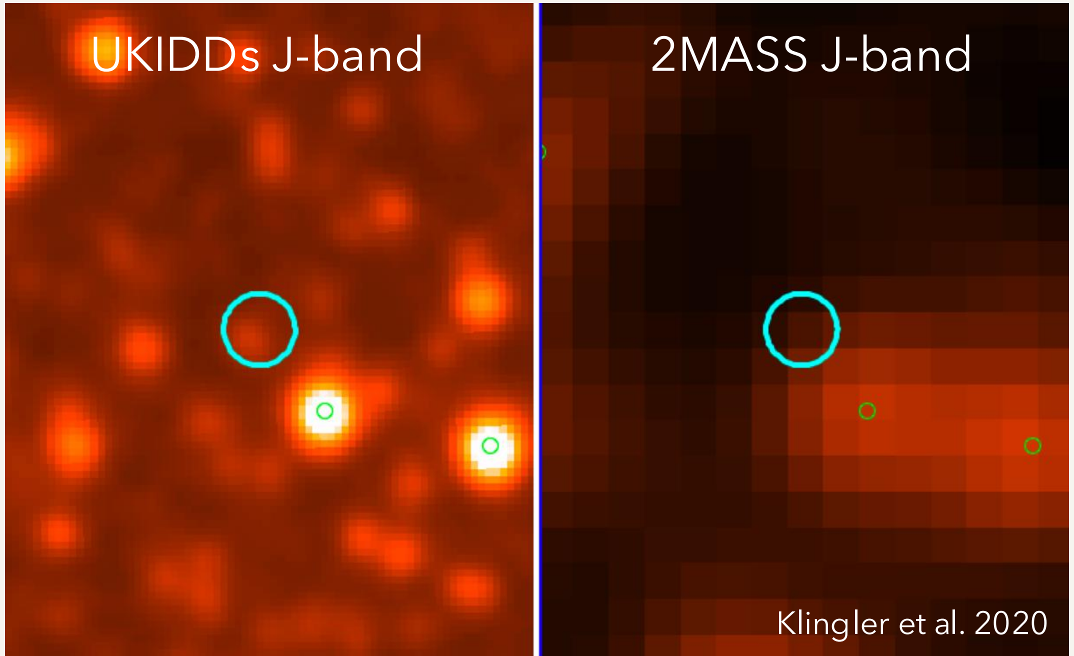
Confident



True \ Predicted	AGN	CV	HM-STAR	LM-STAR	HMXB	LMXB	NS	YSO
AGN 0.89	99	0	0	0	0	0	1	0
CV 0.41	0	78	0	0	6	6	0	11
HM-STAR 0.64	0	0	83	7	0	0	0	11
LM-STAR 0.74	1	0	1	95	0	1	0	3
HMXB 0.73	0	5	16	0	74	0	0	5
LMXB 0.22	21	14	0	0	7	29	21	7
NS 0.43	0	0	0	0	0	14	86	0
YSO 0.88	0	0	1	1	0	0	0	98

Classification of CSCv2: Issues and Biases

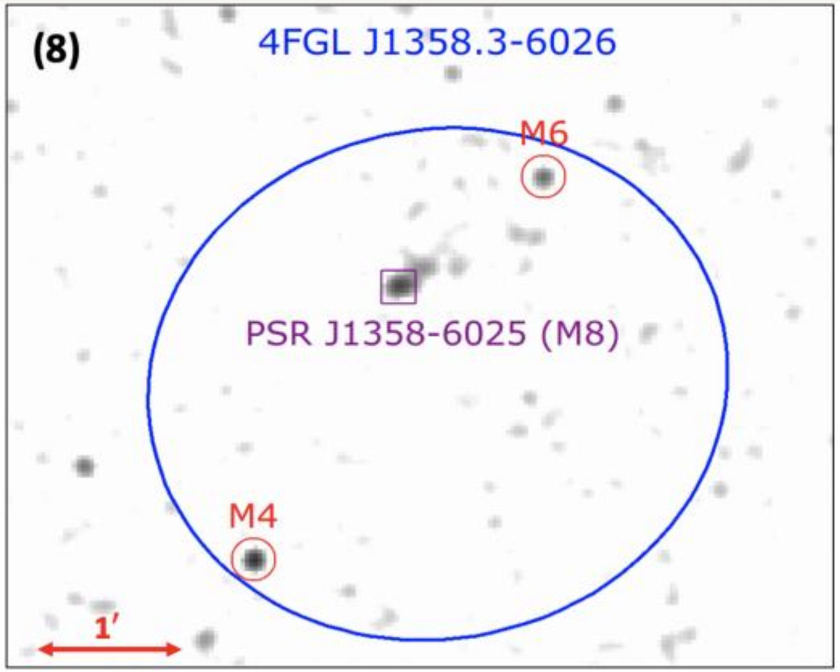
- NS virtually all too faint to be detected by multi-wavelength surveys used in training dataset
- Many LMXBs also have counterparts too faint to be detected by multi-wavelength surveys used in training dataset, hence they become confused with the NS class
- Sources too faint to be detected by these MW surveys (e.g., M-dwarfs, absorbed AGN) will be preferentially classified as NS/LMXBs



Example use cases

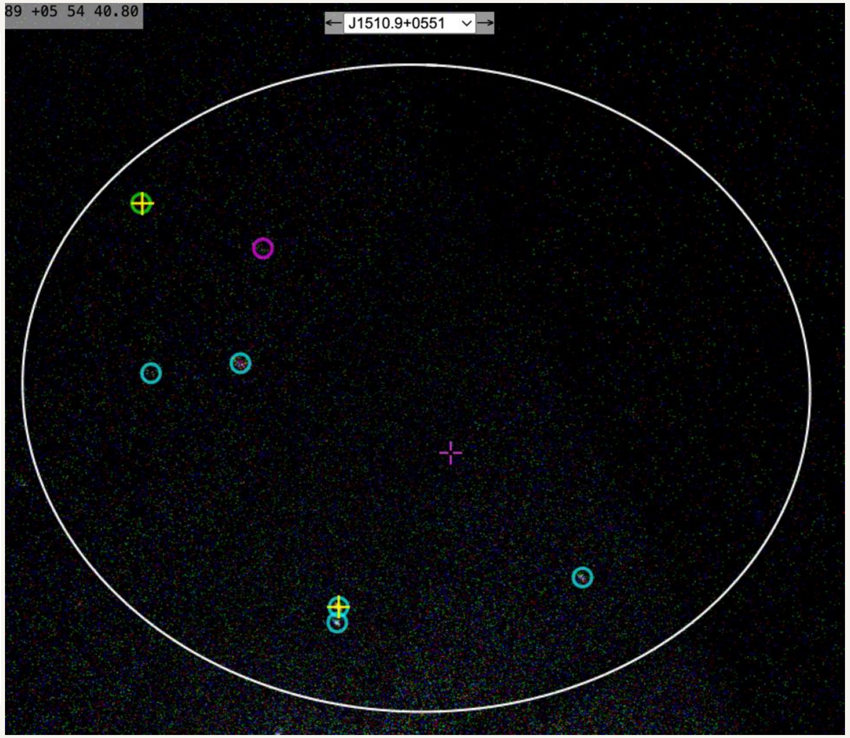
- Searching for X-ray counterparts to GeV Sources
- Classifying sources in Globular Clusters (See poster #5 by Steven Chen today)

Rangelov et al. 2024



Correctly identified a NS

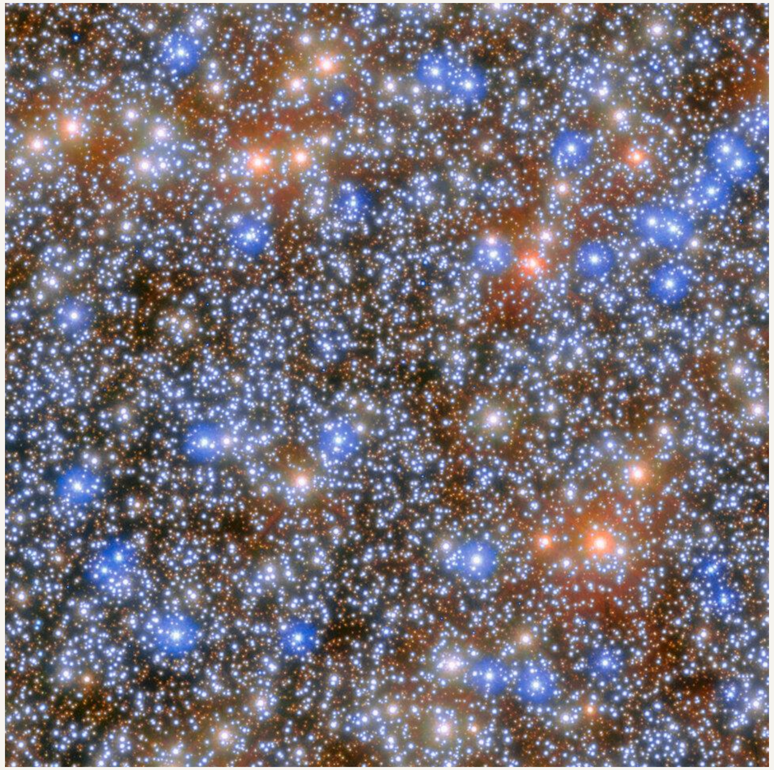
Yang et al. 2024



Identified new AGN counterparts

https://muwclass.github.io/MUWCLASS_4FGL-DR4/

Chen et al. in prep



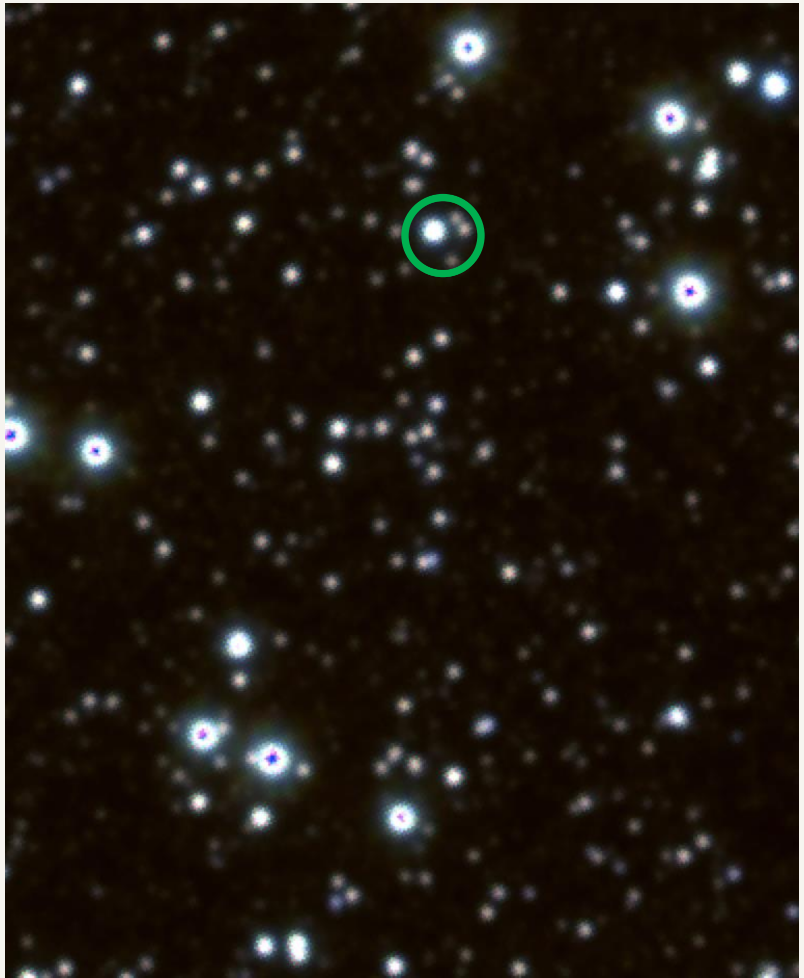
Omega Cen HST

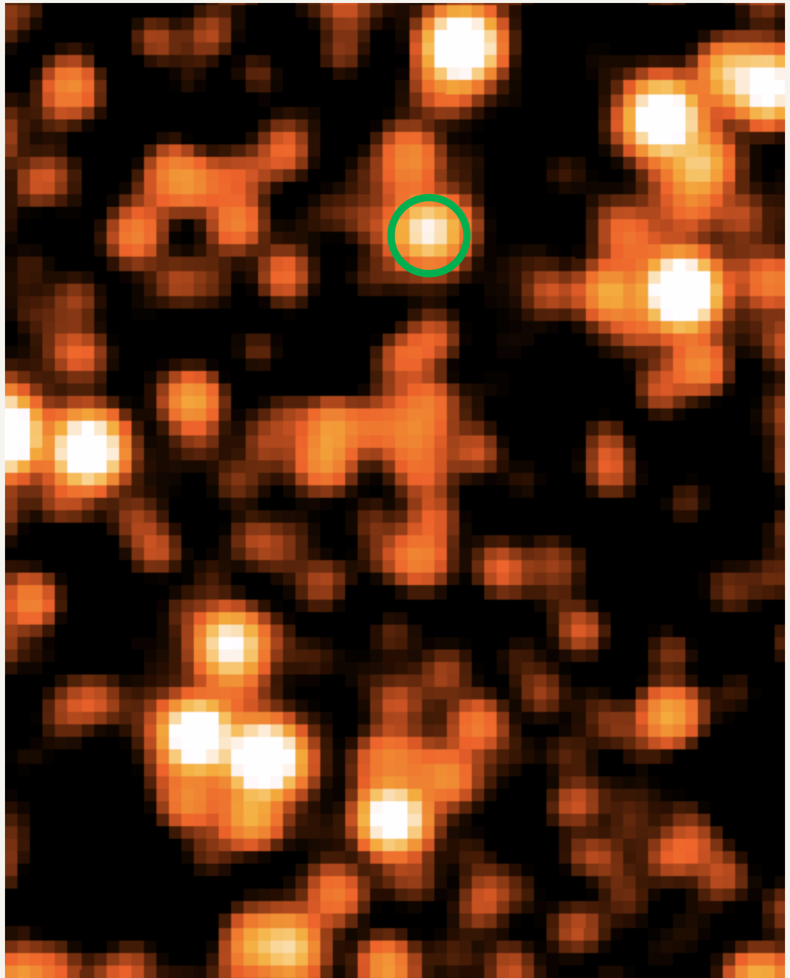
Klingler et al. 2020

Future Improvements: Deeper Surveys

- Update to more sensitive surveys (e.g., Pan-STARRs, DECaps, Vista VVV)

Vista VVV





2MASS

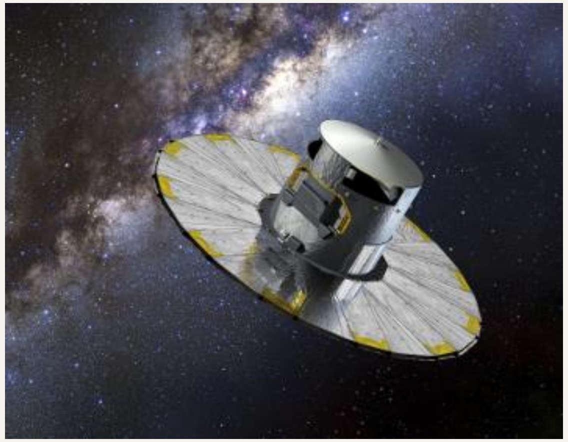
Future Improvements: New Multi-wavelength Features

ASKAP

https://www.atnf.csiro.au/projects/askap/index.html

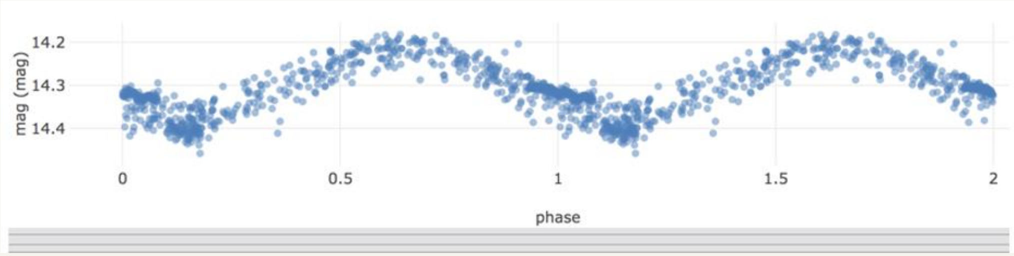
- Inclusion of new radio surveys:
- Australian SKA Pathfinder Telescope (ASKAP)
- VLA All-sky Survey (VLASS)
- MeerKAT source catalog

Gaia



https://upload.wikimedia.org/wikipedia/en/0/01/Gaia_spacecraft.jpg

- Distances and proper motions from Gaia eDR3



ZTF

https://www.ipac.caltech.edu/project/ztf

- Large field of view optical time domain surveys:
- ZTF
- TESS
- VCRO

Creating Training Datasets

ATNF Pulsar Catalogue

Catalogue Version: 1.65

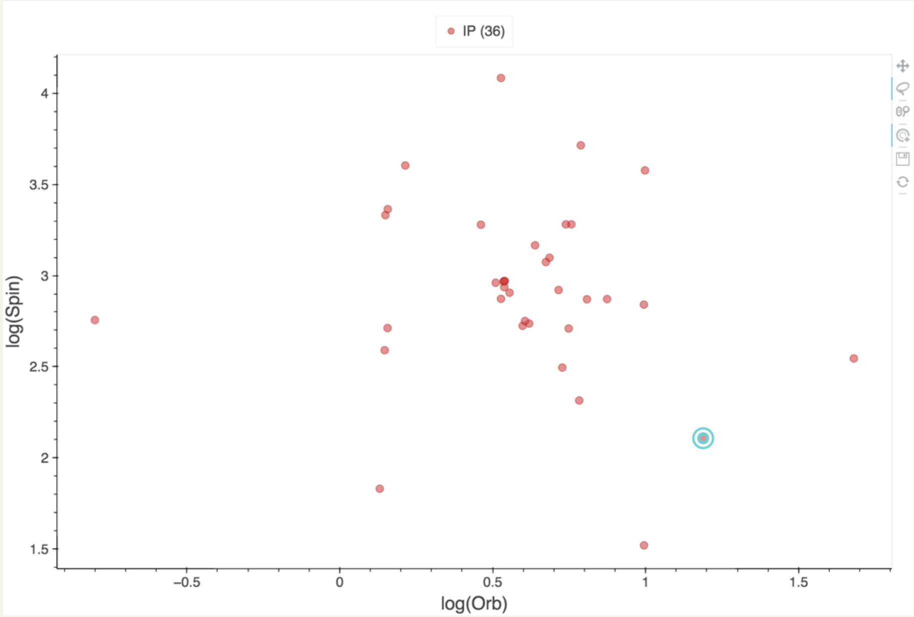
#	NAME	PSRJ	P0 (s)	P1	DM (cm^-3 pc)	BINARY (type)	DIST (kpc)	AGE (Yr)	BSURF (G)	EDOT (ergs/s)
1	J0002+6216	J0002+6216	0.115364	5.97e-15	218.60	*	6.357	3.06e+05	8.40e+11	1.53e+35
2	J0006+1834	J0006+1834	0.693748	2.10e-15	11.41	*	0.860	5.24e+06	1.22e+12	2.48e+32
3	J0007+7303	J0007+7303	0.315873	3.60e-13	*	*	1.400	1.39e+04	1.08e+13	4.51e+35
4	J0011+08	J0011+08	2.552870	*	24.90	*	5.399	*	*	*
5	B0011+47	J0014+4746	1.240699	5.64e-16	30.41	*	1.776	3.48e+07	8.47e+11	1.17e+31
6	J0021-0909	J0021-0909	2.314131	1.04e-15	25.20	*	25.000	3.53e+07	1.57e+12	3.31e+30
7	J0023+0923	J0023+0923	0.003050	1.14e-20	14.33	ELL1	1.111	4.23e+09	1.89e+08	1.59e+34
8	J0024-7204aa	J0024-7204aa	0.001840	*	24.97	*	2.688	*	*	*
9	J0024-7204ab	J0024-7204ab	0.003705	9.82e-21	24.37	*	2.540	5.98e+09	1.93e+08	7.62e+33
10	B0021-72C	J0024-7204C	0.005757	-4.99e-20	24.60	*	4.690	*	*	*
11	B0021-72D	J0024-7204D	0.005358	-3.42e-21	24.74	*	4.690	*	*	*
12	B0021-72E	J0024-7204E	0.003536	9.85e-20	24.24	DD	4.690	5.69e+08	5.97e+08	8.79e+34
13	B0021-72F	J0024-7204F	0.002624	6.45e-20	24.38	*	4.690	6.44e+08	4.16e+08	1.41e+35
14	B0021-72G	J0024-7204G	0.004040	-4.22e-20	24.44	*	4.690	*	*	*
15	B0021-72H	J0024-7204H	0.003210	-1.83e-21	24.37	DD	4.690	*	*	*

Intermediate Polar Catalog

No.	Var. Name	Alt. Name(s)	RA	Dec	P_o (h)	P_s (s)	Level
001	V1033 Cas	IGR J00234+6141 1RXS J002258.3+614111	00 22 57.63	+61 41 07.8	4.033	563.5	*****
002	V709 Cas	RX J0028.8+5917	00 28 48.9	+59 17 21.6	5.341	312.78	*****
003	V515 And	XSS J00564+4548 1RXS J005528.0+461143	00 55 20.0	+46 12 57	2.731086	465.48493	****
004		1RXS J015317.9+744641 RX J0153.3+7446	01 53 20.76	+74 46 22.2	3.9396	1974?	***
—	TT Ari		02 06 53.08	+15 17 41.8	3.3012		*
—	HP Cet	SDSS J023322.61+005059.5	02 33 22.61	+00 50 59.5	1.6013		*
005	XY Ari	H0253+193	02 56 08.15	+19 26 33.8	6.0648	206.3	*****
006	GK Per	Nova Persei 1901	03 31 12.0	+43 54 17	47.9233	351	*****
—	AH Eri		04 22 38.10	-13 21 30.2	5.7384	2520??	*
007		IGR J04571+4527 1RXS J045707.4+452751	04 57 08.32	+45 27 50.0	6.19?	1218.7	***
008	V1062 Tau	H0459+246	05 02 27.59	+24 45 22.1	9.952	3780	*****
009	UU Col	RX J0512.2-3241	05 12 13.22	-32 41 39.8	3.45	863.5	*****

https://asd.gsfc.nasa.gov/Koji.Mukai/iphome/catalog/alpha.html

- Tedious process, must be careful to select correct counterpart in crowded regions of Galactic plane.
- This often requires a review of the literature.
- Many source catalogs are not up to date
- ADAP funded project to create living catalogs based on CSC v2.1, 4XMM, eROSITA, and Swift source catalogs



Conclusion

- We have developed an automated machine learning pipeline to efficiently classify X-ray sources based on their X-ray and various multi-wavelength properties
- We have also included a framework to account for uncertainties and source confusion
- Overall performance on most common source types (e.g., AGN, YSO, Low mass) stars is very good and has been used in various environments (GeV sources, globular clusters, etc.)
- Need to build the infrastructure (e.g., catalogs) to handle and keep track of newly discovered sources and their multi-wavelength datasets.

Pipeline and training dataset publicly available at:
https://github.com/huiyang-astro/MUWCLASS_CSCv2

Link can be found in Yang, Hare, et al. 2022