Image credit:
NASA/JPL-Caltech

Mining the high-energy Universe: a probabilistic, interpretable classification of X-ray sources for large X-ray surveys

– The power of CLAXBOI

Hugo TRANIN, Postdoc, ICCUB, University of Barcelona 28 Feb 2024

Outline

1) Data preparation

2) Classification and interpretation

3) Applications

XMM2ATHENA

X-ray catalogs grow larger and larger



	Observations period, Coverage	PSF, Median Sensitivity	Number of sources
XMM-Newton **4XMM-DR13** (Webb+2020)	2000-2022 1328 deg²	6'' 1e-14 $erg/cm^2/s$	657k
Chandra **CSC2** (Evans+2019)	2000-2014 560 deg²	0.5'' on-axis 4e-15 $erg/cm^2/s$	317k
Swift-XRT **2SXPS** (Evans+2020)	2005-2018 3790 deg²	6'' 8e-14 $erg/cm^2/s$	206k

XMM2ATHENA

Focus of this talk

	Observations period, Coverage	PSF, Median Sensitivity	Number of sources
XMM-Newton **4XMM-DR13** (Webb+2020)	2000-2022 1328 deg²	6'' 1e-14 erg/cm^2/s	657k

→ Expected content: AGN, stars, XRB, CV, galaxy clusters…
How to find them? ⇒ **automatic source classification**

XMM2ATHENA

1) Data preparation

"Prepare for battle" – Gandalf

XMM2ATHENA

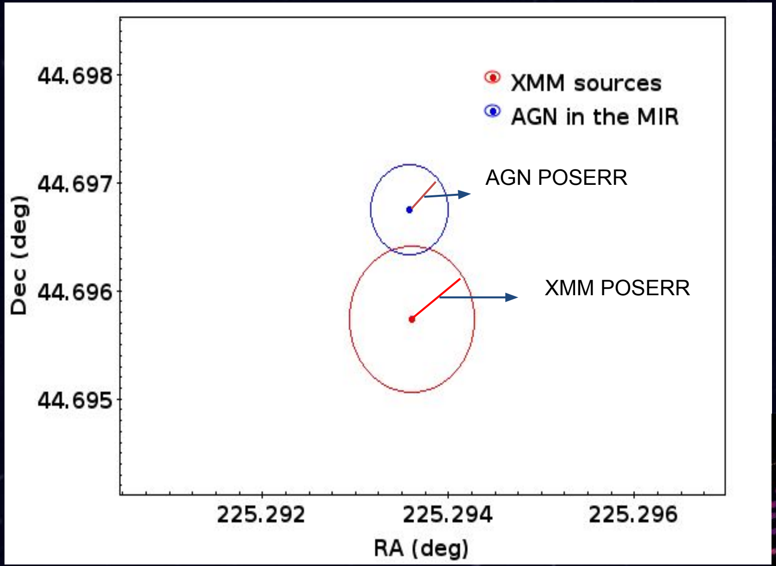
Preparing the dataset for classification

1) Identification of known sources

X-ray samples ⊗ Catalogs of AGN (e.g. Secrest+2015)
Catalogs of stars (e.g. Kharchenko+2009)
Catalogs of XRB & CV (e.g. Ritter+2014)

TOPCAT software (Taylor+2005)
Sky with errors

(Simplistic crossmatch)



Ex. training sample of 4XMM-DR10

AGN	Star	XRB	CV
19,000	6,000	730	260

Tranin et al. A&A 2022

XMM2ATHENA

Preparing the dataset for classification

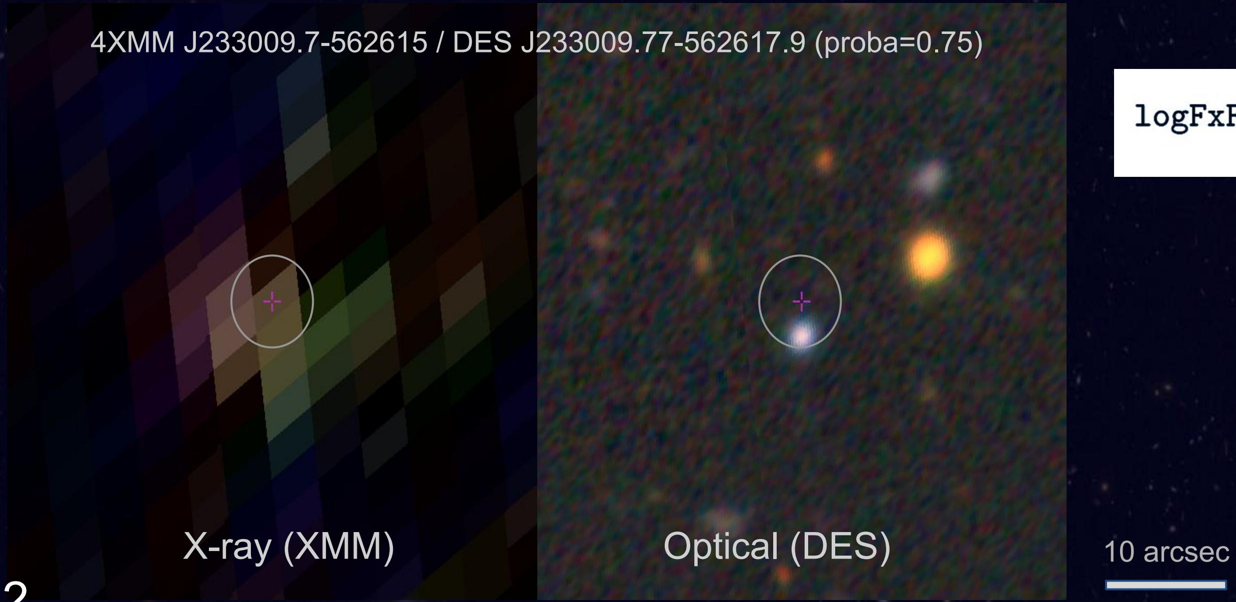
2) Identification of counterparts

X-ray samples ⊗ optical / IR surveys (Gaia, 2MASS…) ⇒ Multiwavelength associations

high sky density → probabilistic treatment

Flux ratios

$$\mathtt{logFxFr} = \log_{10}\left(\frac{F_X}{F_{R\ (\mathrm{Gaia})}}\right)$$



Tranin et al. A&A 2022

XMM2ATHENA

Preparing the dataset for classification

3) Distance estimate

X-ray samples

⊗ Gaia distances (Bailer-Jones+2021)

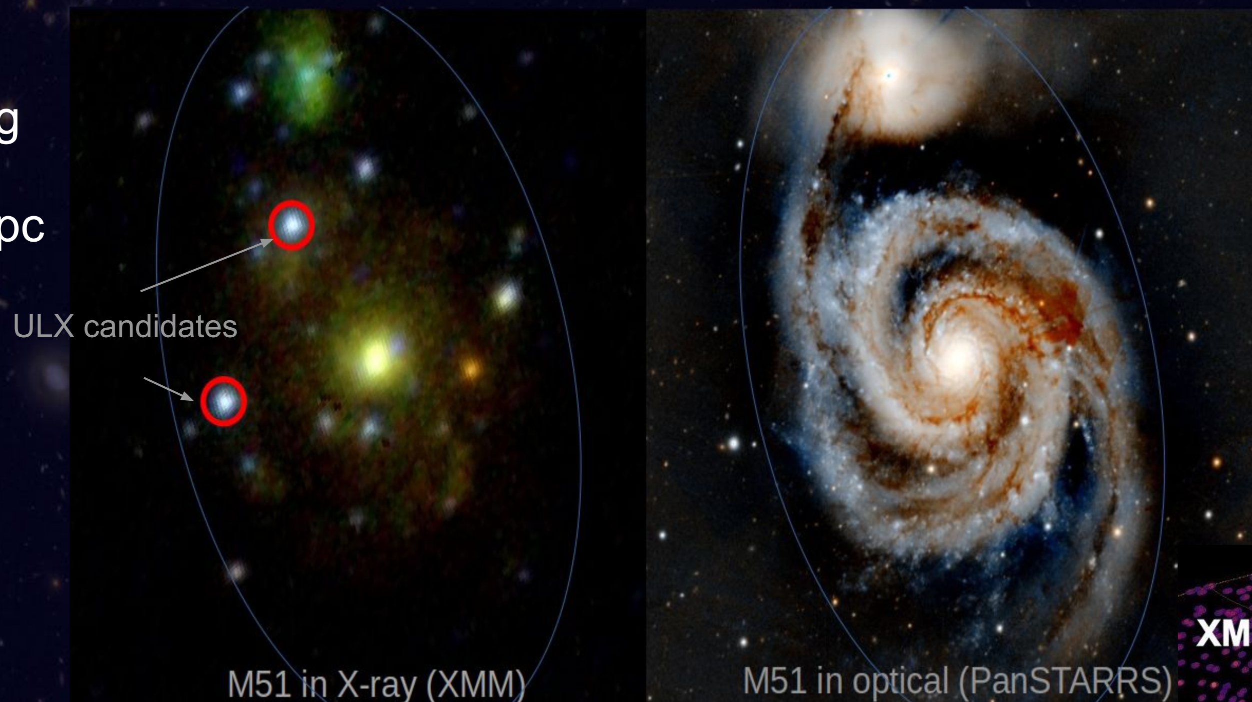
⊗ GLADE (Dalya+2016)
TOPCAT Sky Ellipses Match

⇒ source distance & luminosity

$$L_X = 4\pi D^2 \times F_X$$

GLADE = all-sky highly complete galaxy catalog

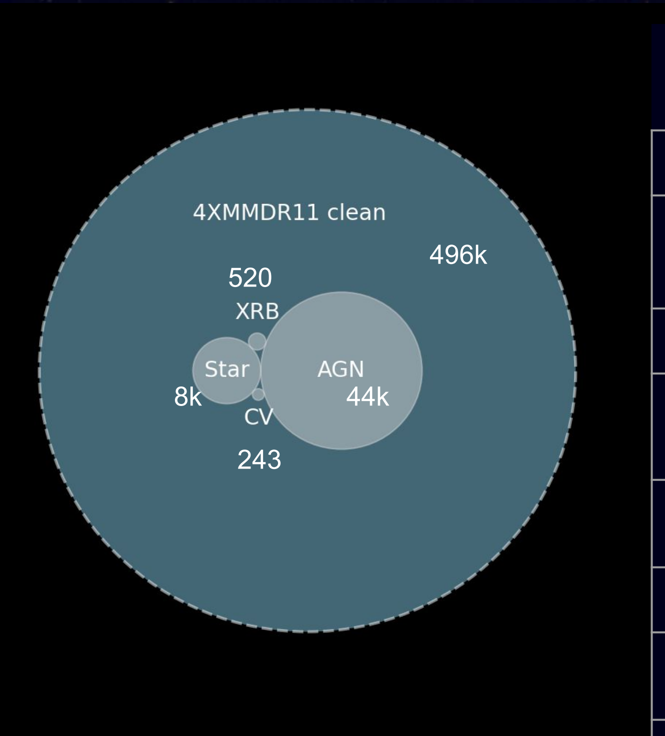
>1M galaxies at D<500Mpc



Tranin et al. A&A 2022

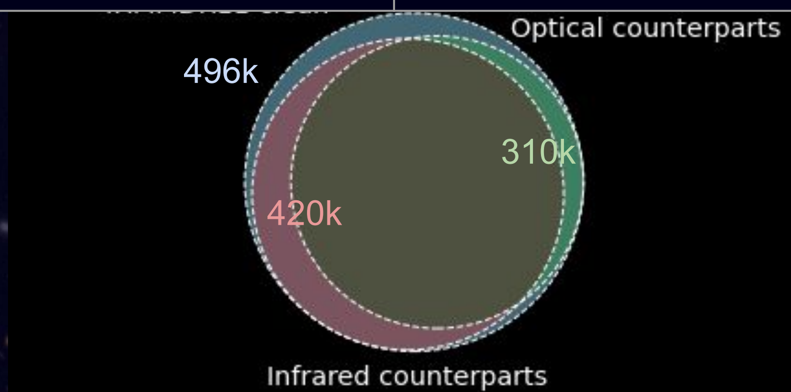
XMM2ATHENA

Multiwavelength dataset ready for classification



	Name / Reference	in 4XMM-DR11
X-ray samples	-	496k
Optical sources	Gaia EDR3, PanSTARRS, DES	310k
Infrared sources	2MASS, AllWISE, UnWISE	420k
Matches with galaxies	GLADE (Dalya+2016)	16k
Identified AGN	Véron-Cetty+2010, Secrest+2015, Simbad	44k
Identified Stars	ASCC (Kharchenko+2009)	8k
Identified XRB	Liu Q. Z.+2006, 2007, Humphrey+2008, Mineo+2012…	520
Identified CV	Downes+2006, Ritter+2014	243

small samples



Tranin et al. A&A 2022

XMM2ATHENA

Features used by the classifier

Name	Category
Galactic latitude	Location
Gaia proper motion	Location
Relative distance to the host center	Location
X-ray luminosity	Location
X-ray over optical (b,r) flux ratio	Counterparts
X-ray over infrared (W1,W2) flux ratio	Counterparts
X-ray max to min flux ratio	Variability
X-ray lower max to higher min flux ratio	Variability
X-ray hardness ratio HR1, HR2, HR3...	Hardness
Power law index fitted to X-ray spectrum	Hardness

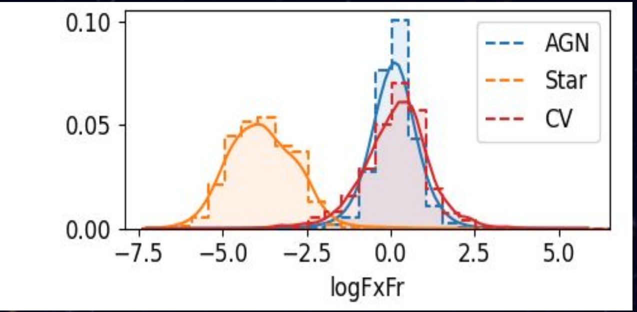
Tranin et al. A&A 2022

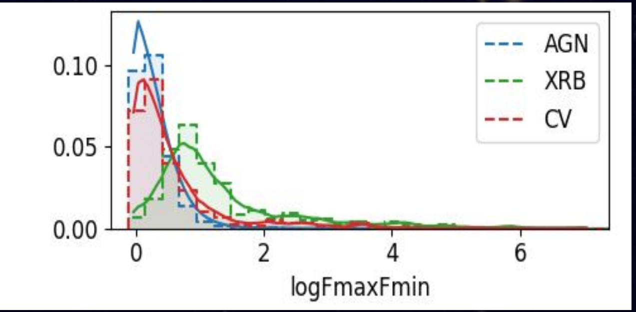
XMM2ATHENA

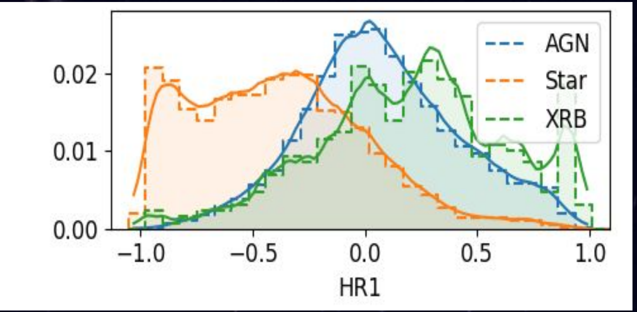
Probability densities of the training samples

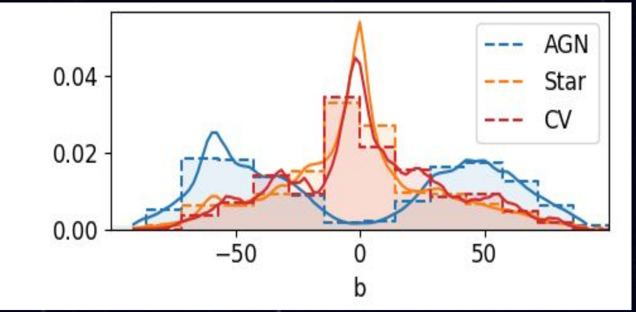
Physical properties:

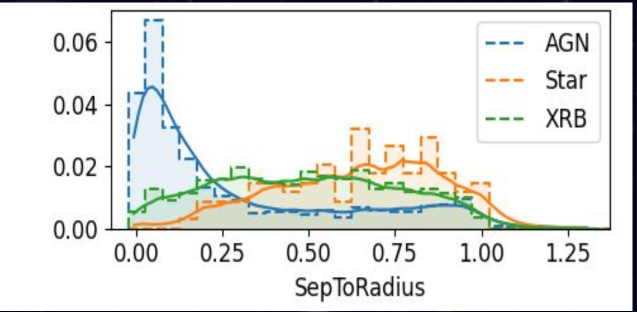
- logFxFr *(counterpart)*
  - logFmaxFmin *(variability)*
- HR1 *(spectrum)*
  - b *(location)*
- sep *(location)*
  - L_X *(spectrum)*

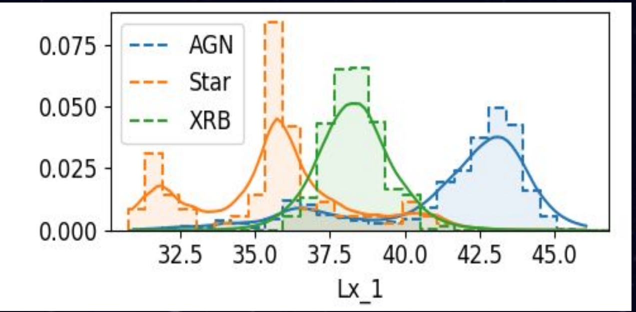












2) Probabilistic classification (CLAXBOI) and interpretation

"You're a wizard, Harry" – Hagrid

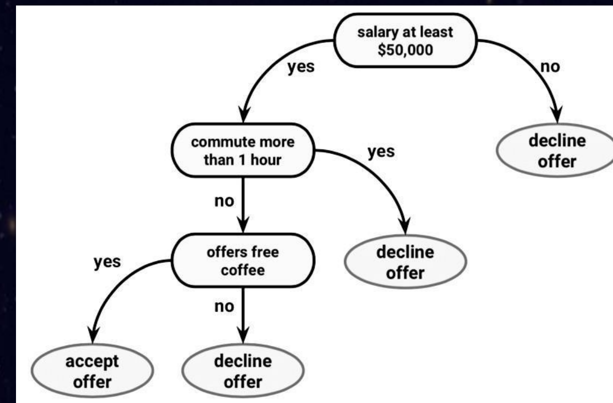
XMM2ATHENA

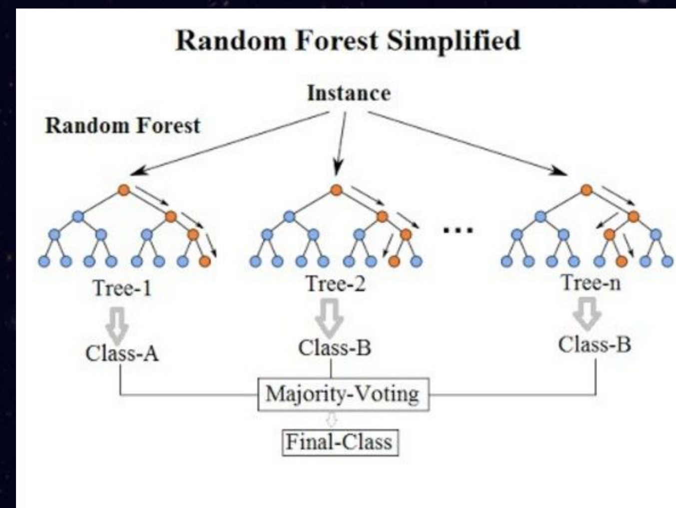
Methods for automatic source classification

Before 2022, in X-ray astronomy:

- Decision tree (e.g. Lin+2012) → *poor performance*
- Random forest (e.g. Farrell+2015, Arnason+2020) → *poor interpretability*
- Other machine learning algorithm (nearest neighbors, naive Bayes…) (e.g. Pineau+2017, Arnason+2020)

CLAXBOI: **probabilistic** classification, good **interpretability** and **reliability**





XMM2ATHENA

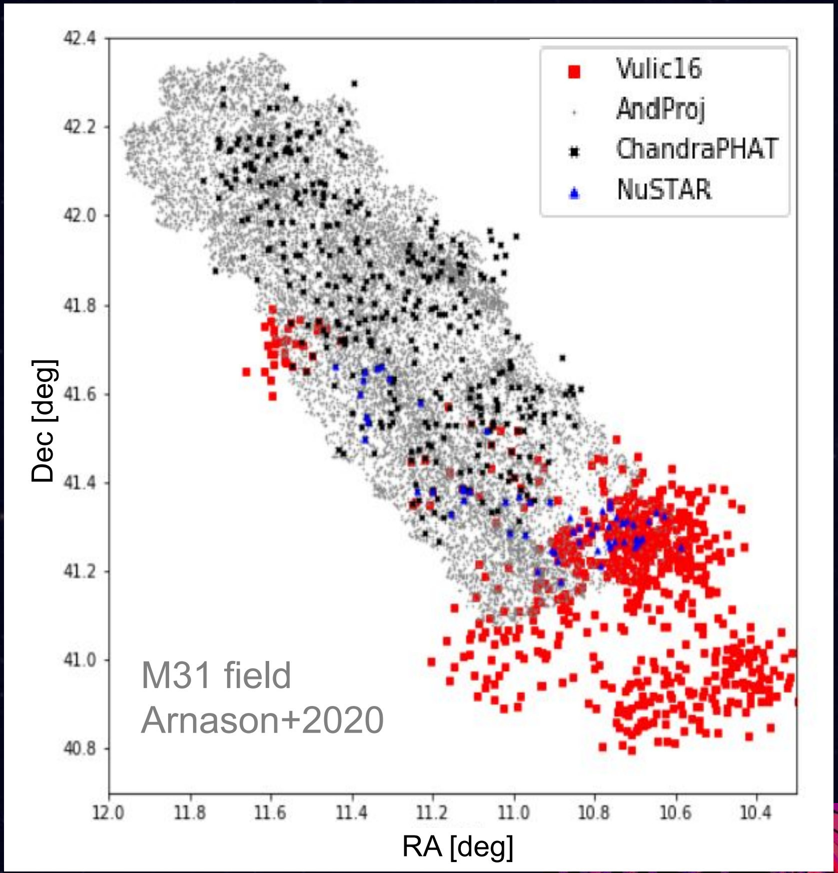
Previous studies

Previously classified samples (before 2022)

Small! ~ $10^{3\text{-}4}$ sources instead of 10^6 detected

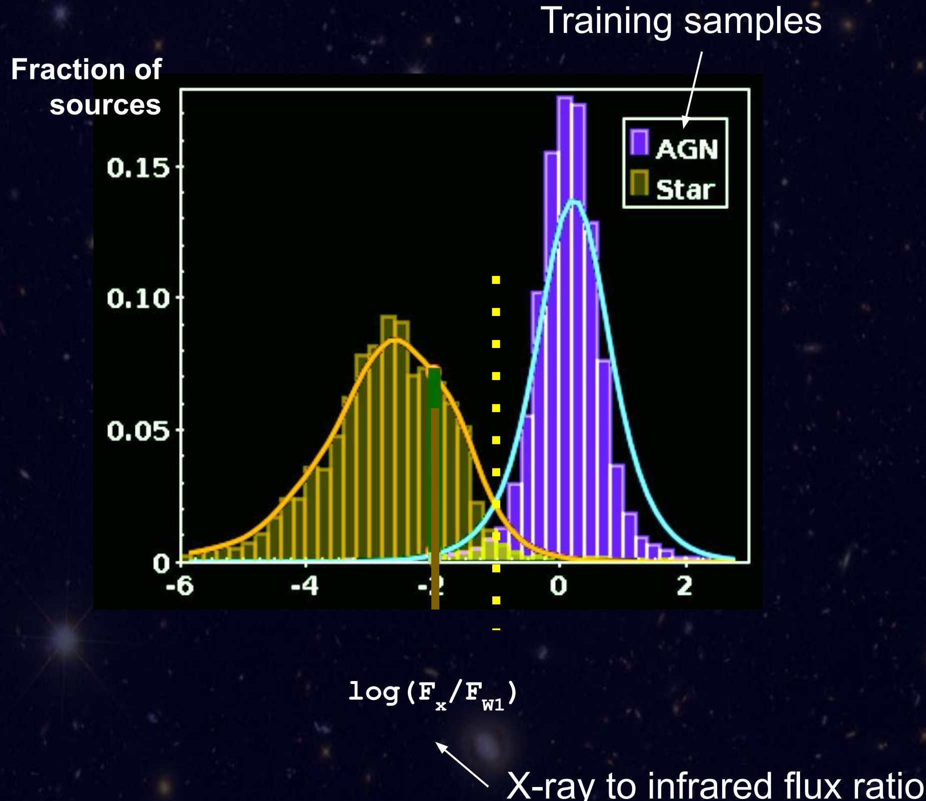
- Only bright sources (e.g. Lin+2012)
- Only variable sources (e.g. Farrell+2015)
- Only specific fields (e.g. Arnason+2020)

CLAXBOI: classification of **most of well-detected point-like sources**



XMM2ATHENA

Naive Bayes Classifier (2 classes)



Possible criterion:

$\log(F_X/F_{W1}) < -1 \Rightarrow$ star

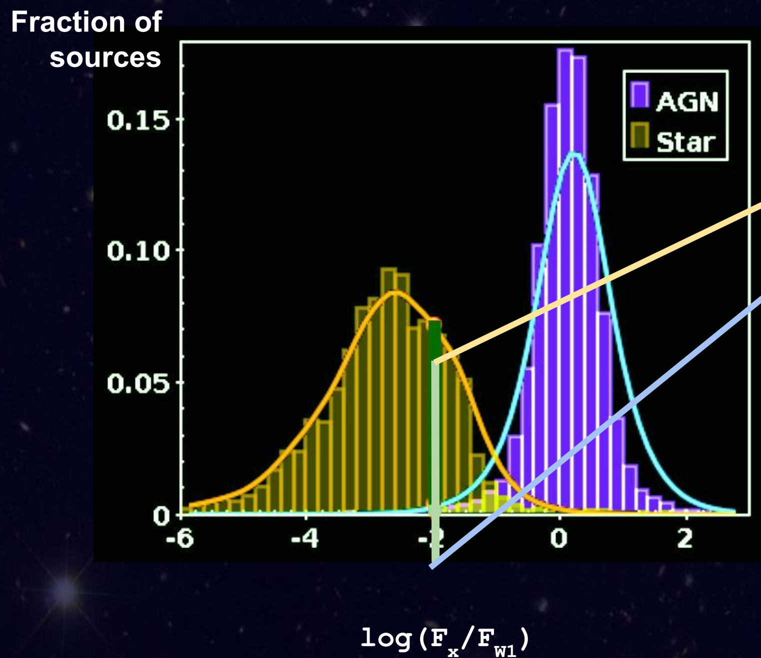
else $\Rightarrow$ AGN

… but overlap

Tranin et al. A&A 2022

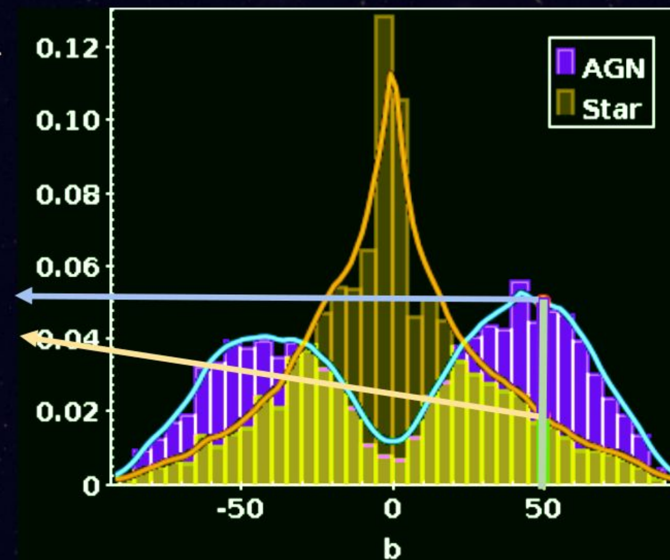
XMM2ATHENA

Naive Bayes Classifier (2 classes)



If F_x/F_{W1} = 0.01
$\mathcal{L}_{b,\,Star}$ = 93%
$\mathcal{L}_{b,\,AGN}$ = 7%

If b = 50°
$\mathcal{L}_{b,\,AGN}$ = 67%
$\mathcal{L}_{b,\,Star}$ = 33%



$$\mathbb{P}(\mathrm{AGN}|D) = \frac{\mathcal{P}(\mathrm{AGN})\mathcal{L}(\mathrm{AGN}|D)}{\mathcal{P}(\mathrm{AGN})\mathcal{L}(\mathrm{AGN}|D) + \mathcal{P}(Star)\mathcal{L}(Star|D)}$$ = 31% here

(with priors $\mathcal{P}$(AGN)=0.75, $\mathcal{P}$ (Star)=0.25)

Combine the 18 features ⇒ **Naive Bayes classification**

Maximising the classification performance

- Trade-off between recall and precision
- Optimization : fine-tuning the α_t

$$\mathbb{P}(\mathrm{c}|data) = \frac{\mathcal{P}(\mathrm{c}) \times \left(\prod_{t\in\{\mathrm{cat}\}} \mathcal{L}(t|\mathrm{c})^{\alpha_t}\right)^{1/\sum_{t\in\{\mathrm{cat}\}}\alpha_t}}{\sum_{C\in\{\mathrm{classes}\}} \mathcal{P}(C) \times \left(\prod_{t\in\{\mathrm{cat}\}} \mathcal{L}(t|C)^{\alpha_t}\right)^{1/\sum_{t\in\{\mathrm{cat}\}}\alpha_t}}$$

One α_t per category of properties: α_{location}, α_{spectrum}, $\alpha_{\text{variability}}$, $\alpha_{\text{counterparts}}$

Optimized to maximize the f_1-score of XRB ($f_1 = (\text{recall}^{-1}+\text{precision}^{-1})^{-1}$)

XMM2ATHENA

Results (Confusion matrix)

on 4XMM training sample (because no overfitting + few XRB and CV)

	AGN	Star	XRB	CV
→AGN	18373	25	46	149
→Star	15	6197	10	12
→XRB	80	12	479	10
→CV	4	0	8	81
recall (%)	99.5	99.4	88.2	32.1
precision (%)	98.9	97.2	93.7	84.6
f_1-score	0.992	0.983	0.909	0.465

on 2SXPS

Truth →	AGN	Star	XRB	CV	Total cl.
→AGN	19515	82	25	191	19813
→Star	44	4628	3	27	4702
→XRB	140	18	326	17	501
→CV	9	9	2	124	144
Total	19708	4737	356	359	Average
recall (%)	99.0	97.7	91.6	34.5	80.7
precision (%)	97.0	98.6	90.7	85.5	92.3

Random Forest on 2SXPS

Truth →	AGN	Star	XRB	CV	Total cl.
→AGN	5889	7	20	39	5955
→Star	6	1404	1	3	1414
→XRB	9	5	83	5	102
→CV	7	1	1	68	77
Total	5911	1417	105	115	Average
recall (%)	99.6	99.1	79.0	59.1	84.2
precision (%)	96.8	99.2	95.2	87.9	95.2

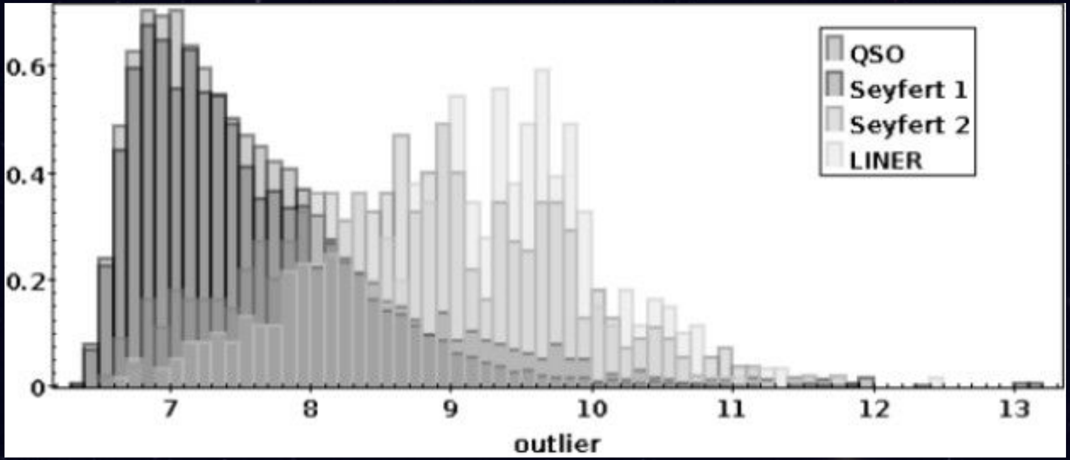
Tranin et al. A&A 2022

⇒ better results on XRB + better interpretability

Interpretation #1: Finding outliers

$$O.M. = -\log\left(\mathcal{P}(\mathrm{c}) \times \prod_{t\in\{\mathrm{cat}\}} \mathcal{L}(t|\mathrm{c})^{\alpha_t / \sum_{t\in\{\mathrm{cat}\}} \alpha_t}\right)$$

~ scarcity of the training sample at the location of the source in the parameter space
Depends on the output class *c*
⇒ way to nuance the classification



Tranin et al. A&A 2022

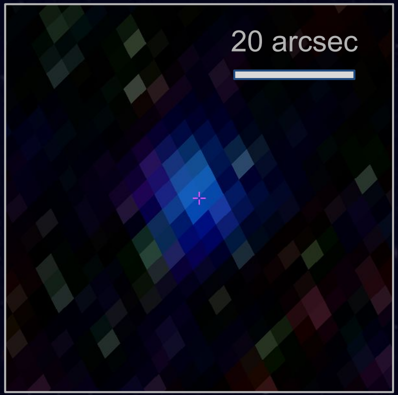
Outliers = one of these:

- Spurious sources
- Spurious identifications
- If classified as star/AGN : special types of star/AGN
- If classified as XRB : rare & variable objects such as TDE, GRB, supernovae...

XMM2ATHENA

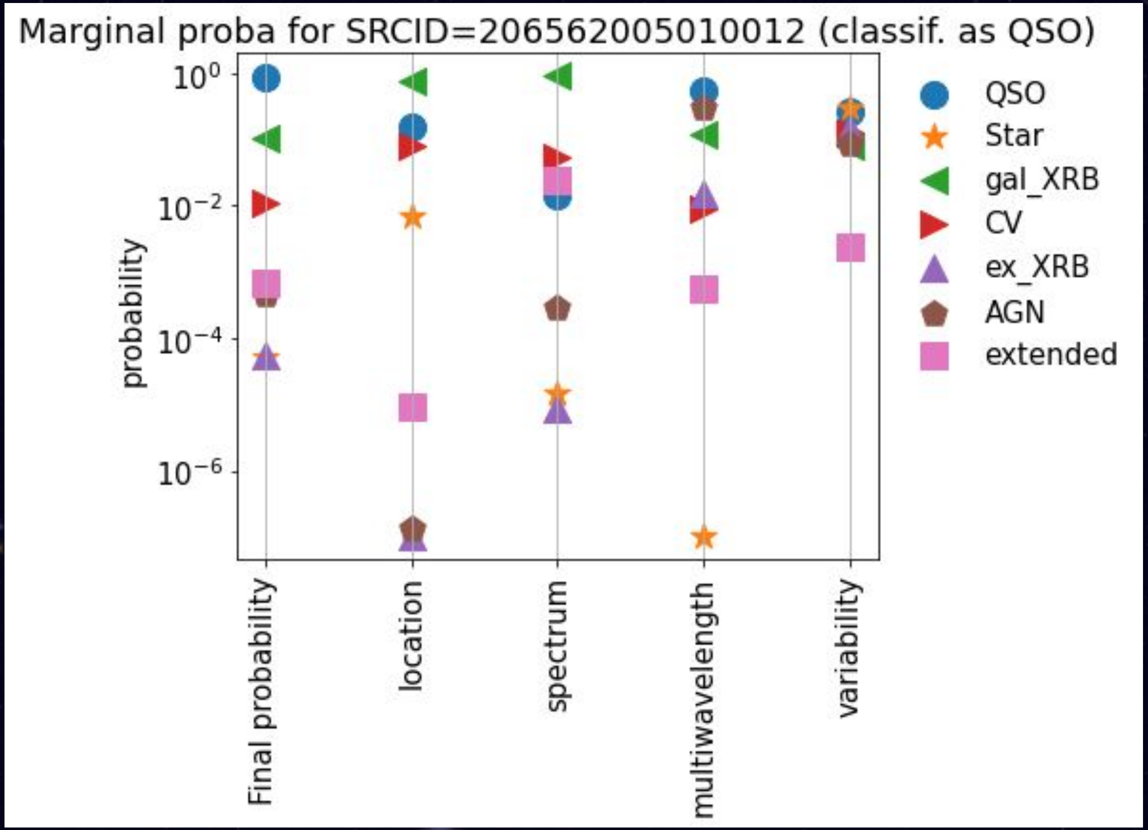
Interpretation #2: marginal probabilities

Sources are classified based on their location, spectrum, counterparts and variability
⇒ find the discriminant properties thanks to marginal probabilities



XMM-Newton

Legacy Survey



P_{AGN} = 88%

Source inspection:
- Hard source
- No optical c. found
- little data

Marginal proba:
- spec and loc suggests Galactic XRB
- other+prior suggest AGN

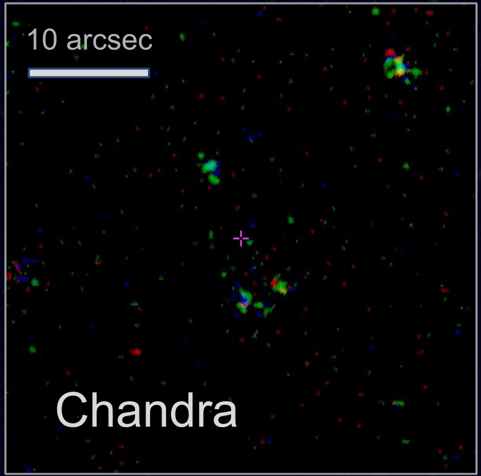
⇒ classification as AGN is explained

Interpretation #3: alternative classifications

Sources are classified based on their
- location,
- spectrum,
- counterparts,
- variability

What if we ignore a category of properties? ⇒ **Alternative classification**

Ex. previous source: no alternative classification
this blended source: alternative classification without location = Galactic XRB



$P_{extended}$ = 92%

XMM extent 42”
Blends 3 Chandra sources
No opt or IR counterpart
Low Galactic latitude b=1°

3) Applications

"This is a beautiful tool but it still needs an active brain to use it"
– Mara Salvato

XMM2ATHENA

Classification of a whole catalog

- 4XMM-DR12 fully classified (XMM2ATHENA deliverables) Published in April 2023: http://xmm-ssc.irap.omp.eu/xmm2athena/catalogues/

[redacted]

7 classes

Priors: 0.55,0.20,0.03,0.02, 0.05,0.05,0.10

truth →	AGN	Star	gal_XRB	CV	AGN_2	ex_XRB	extended
→AGN	23770	26	55	151	0	0	1097
→Star	8	8246	2	6	0	3	597
→gal_XRB	15	2	79	30	0	0	12
→CV	1	2	3	78	0	0	1
→AGN_2	7	3	0	1	958	27	313
→ex_XRB	1	2	1	5	55	510	559
→extended	0	0	0	0	0	0	61438
recall (%)	99.9	99.6	56.4	28.8	94.6	94.4	95.9
precision (%)	95.5	98.9	86.6	88.9	93.3	91.7	100

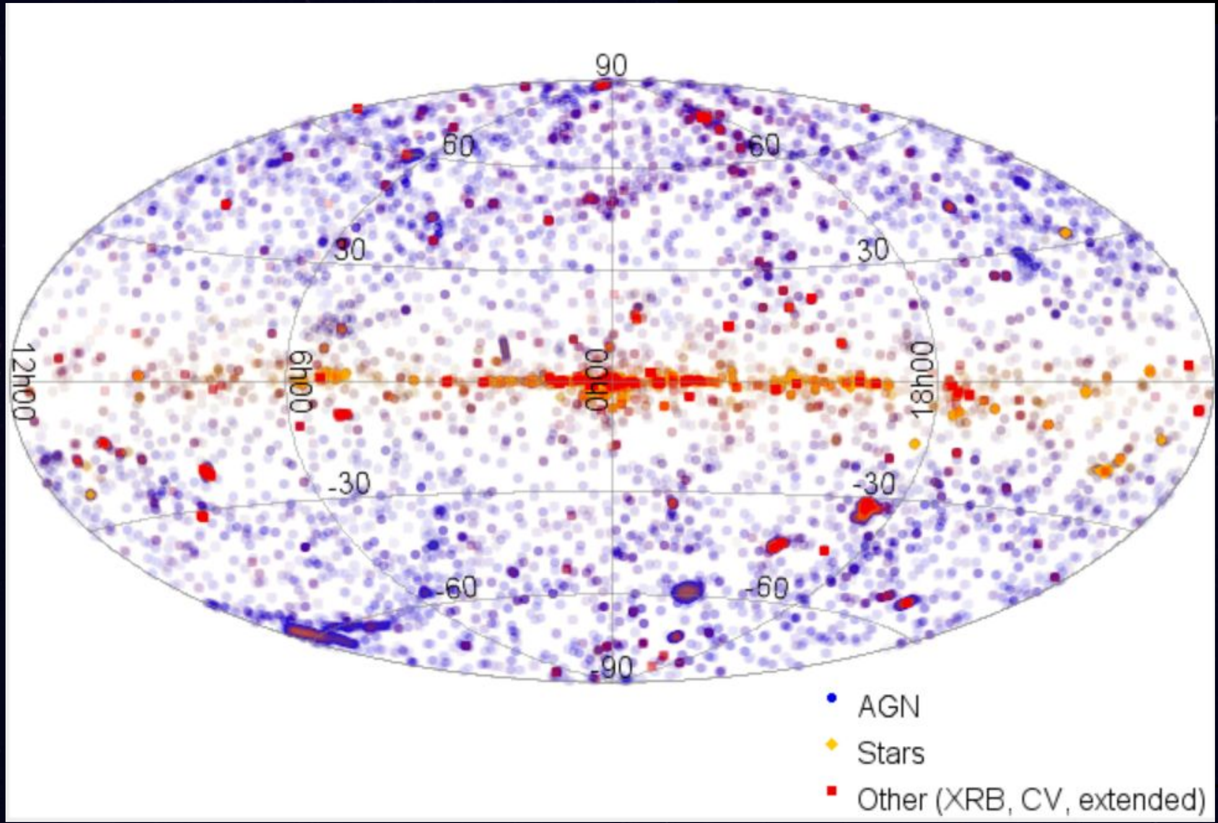
Classification of a whole catalog

- 4XMM-DR12 fully classified (XMM2ATHENA deliverables)
  Published in April 2023:
  http://xmm-ssc.irap.omp.eu/xmm2athena/catalogues/

- **Content**
  430,941 AGN
  75,160 stars
  42,810 Galactic XRB
  8,889 extragalactic XRB
  920 Cataclysmic Variables
  71,627 extended sources

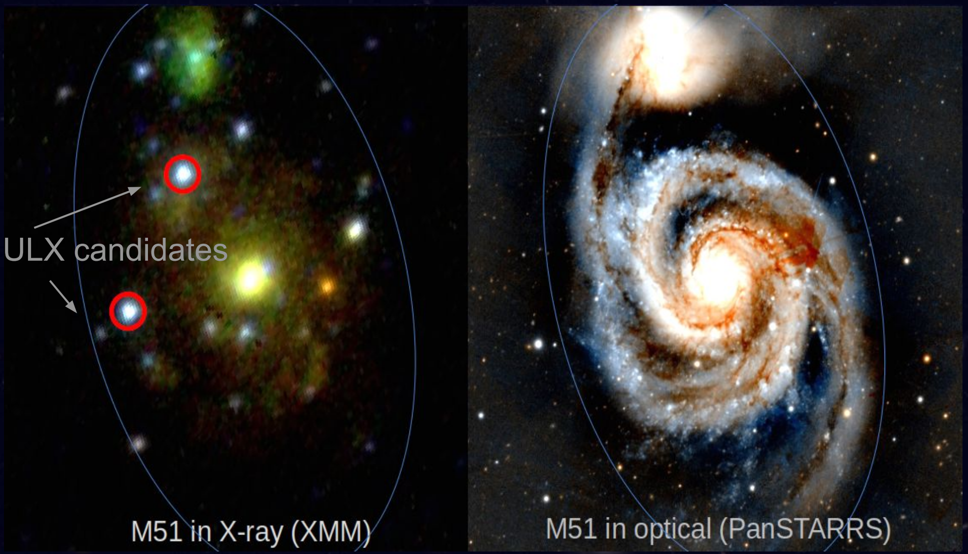
Priors: 0.55,0.20,0.03,0.02,0.05,0.05,0.10

Beware of spurious sources + crowded regions



Specialisation of the classification

X-ray samples ⊗ GLADE (44k sources)



CLAXBOI

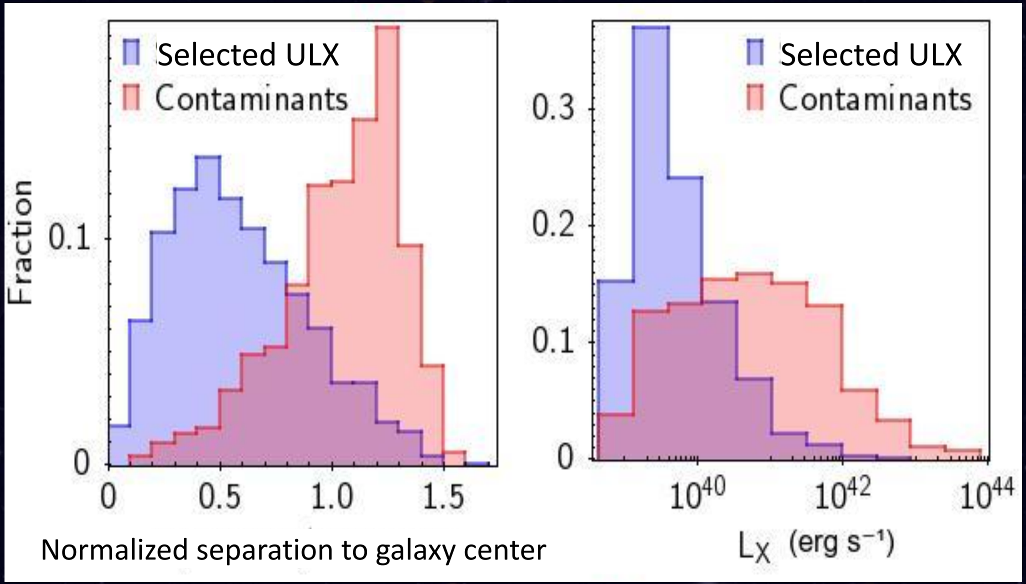
	AGN (background sources)	Soft source (foreground sources, SNR)	XRB
recall	95.2	50.9	89.7
precision	95.8	68.9	80.4

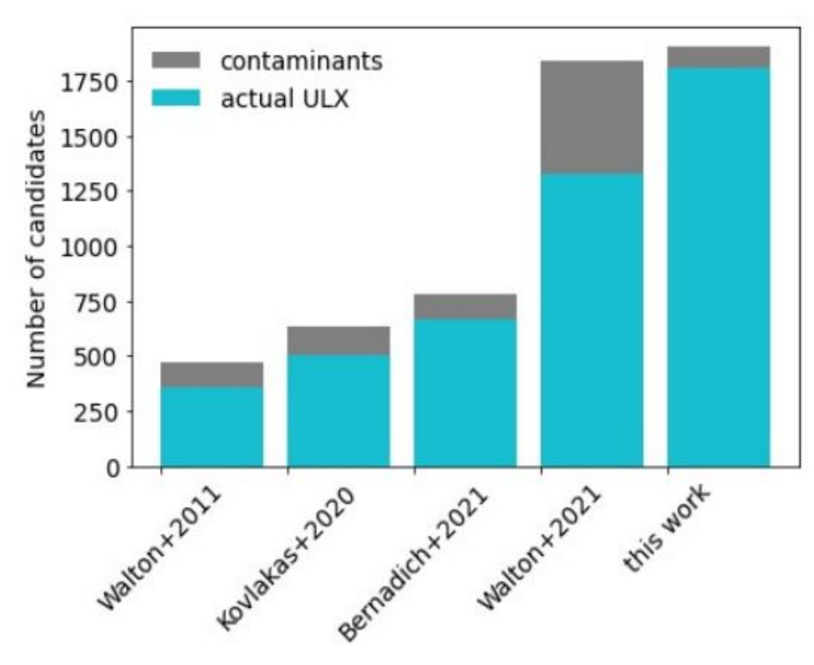
Goal: properly identify ULX

XMM2ATHENA

Identifying ULX in nearby galaxies

- A lot of interlopers remain here if we trust the maximum probability
- We need a physical prior and compare it with P_{XRB}
- Selection criterion : $P_{XRB} > f_{contaminant}$, frequency of background AGN from logN-logS





For the full population study check Tranin et al 2024, A&A 681 A16

[your science case here!]

README

A probabilistic classification of X-ray sources

Classification of X-ray sources using Naive Bayes Optimized Inference (CLAXBOI)

This folder contains all the necessary to run the CLAXBOI code (Tranin et al. 2022, A&A 657, 138) to augment and classify your X-ray catalog. Requirements on the system:

1. Python >3.6
2. Ubuntu (for a few os system commands)

CLAXBOI is public, documented and accessible via github (updated this week): https://github.com/htranin/classificationXray

Feel free to use it for your science cases and reach me in case of questions!

XMM2ATHENA

Complementarity with citizen science

- CLAXBOI includes data preparation and value-adding ✓
- Fully probabilistic classification ✓
- Well-behaved on catalog-sized samples ✓
- Both reliable and interpretable ✓
- Samples of known XRB, CV, TDE… are still small ✗

⇒ to enlarge traning samples and find anomalies, use citizen science.

⇒ **Tomorrow's talk on CLAXSON**

XMM2ATHENA

Conclusion

- **CLAXBOI** is a **versatile, open-source and straightforward code** to make the most of one's X-ray catalog
- It can be easily tuned to identify X-ray sources in **both general** (entire catalogs) and **specific** (population study) **frameworks**
- It has been **successfully applied to 4XMM-DR12** (DR14 coming soon) but also CSC2, 2SXPS
- It provides **highly interpretable classifications**, helping scientific exploitation
- Automatic and Human-based source classification are complementary → see tomorrow's talk about CLAXSON **citizen science project**

github link:


Image credit:
NASA/JPL-Caltech
THANKS FOR
YOUR
ATTENTION
XMM2ATHENA