An optimized, interpretable and automated source classification for large X-ray surveys

Hugo Tranin
2nd year PhD student
Supervisor : Natalie Webb
May 2021

I. Context

Classification of X-ray sources

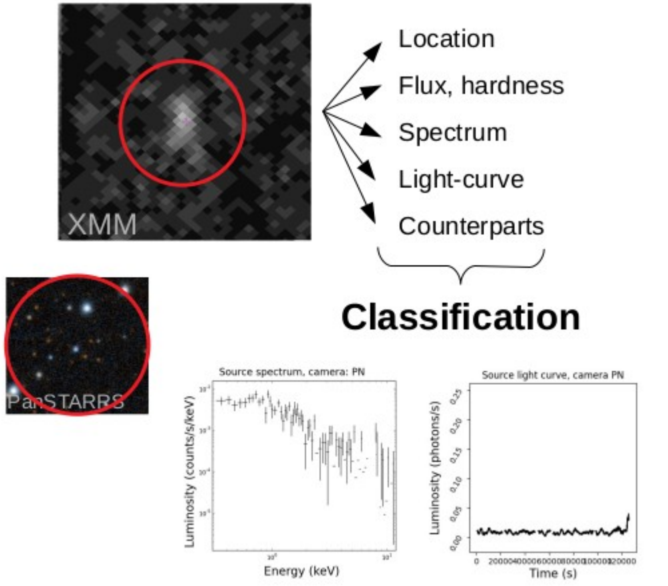
Citizen science

Context - X-ray catalogues

- 3 active telescopes (soft X-rays) : XMM-Newton, Swift and Chandra
- Today : about 1 million sources, most of them unknown !
- Tomorrow : eROSITA, > 3 million sources
- Need of a robust classification

Context - Typical example of X-ray classification

- Illustration on 1 source



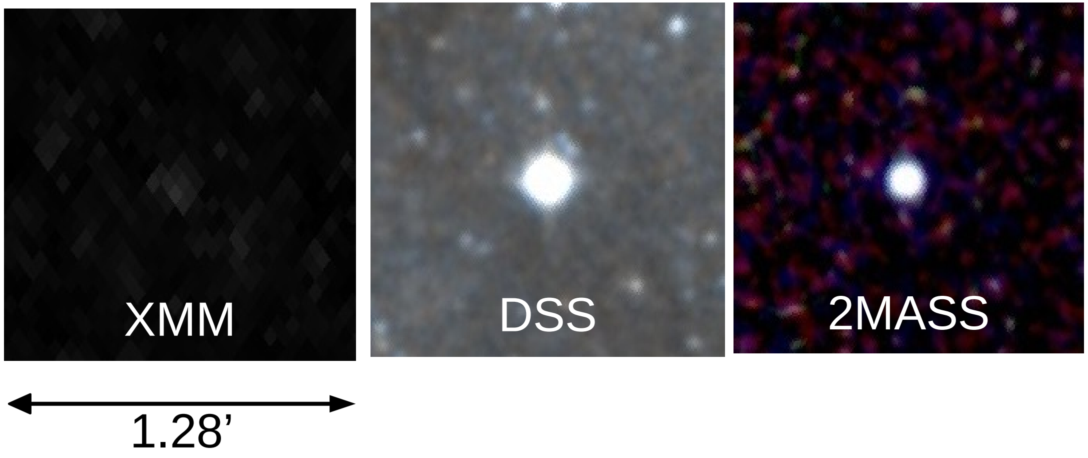
Context - Typical example of X-ray classification

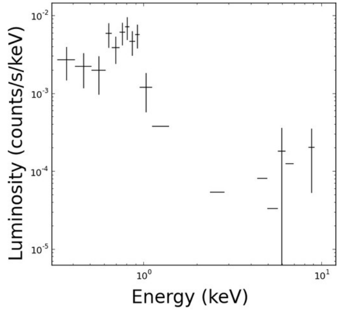
- Illustration on 1 source (star)

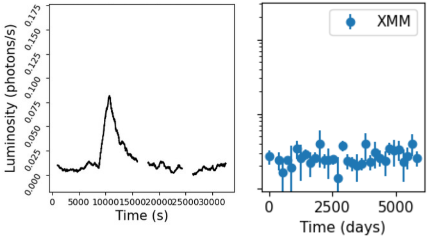
Location
Flux, hardness
Spectrum
Light-curve
Counterparts

Classification

b=-22.2°

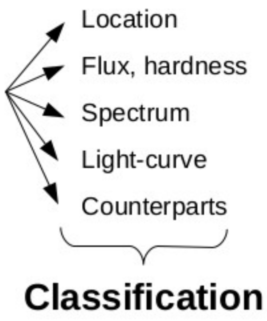




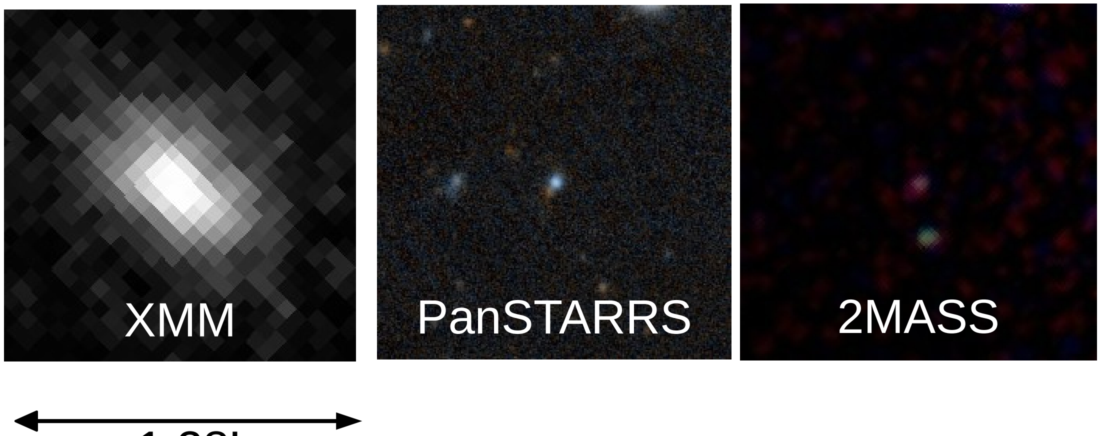


Context - Typical example of X-ray classification

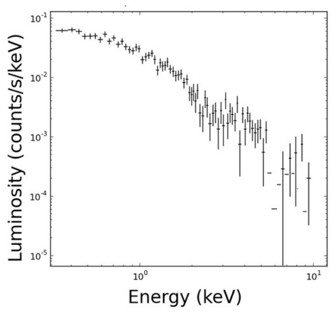
- Illustration on 1 source (AGN)

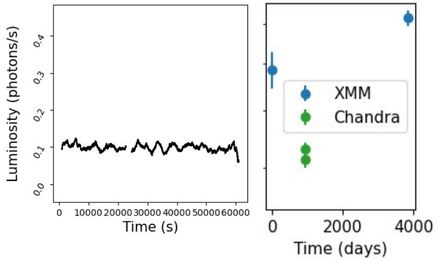


b=-47.1°



1.28’





Context - Typical example of X-ray classification

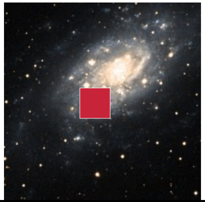
- Illustration on 1 source (XRB)

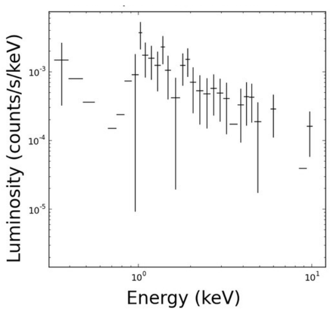
Location
Flux, hardness
Spectrum
Light-curve
Counterparts

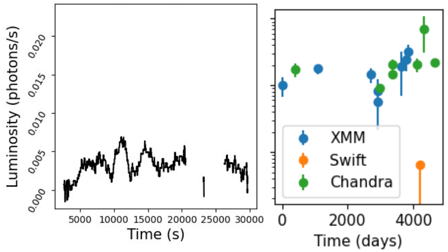
Classification

b=29.2°

38"







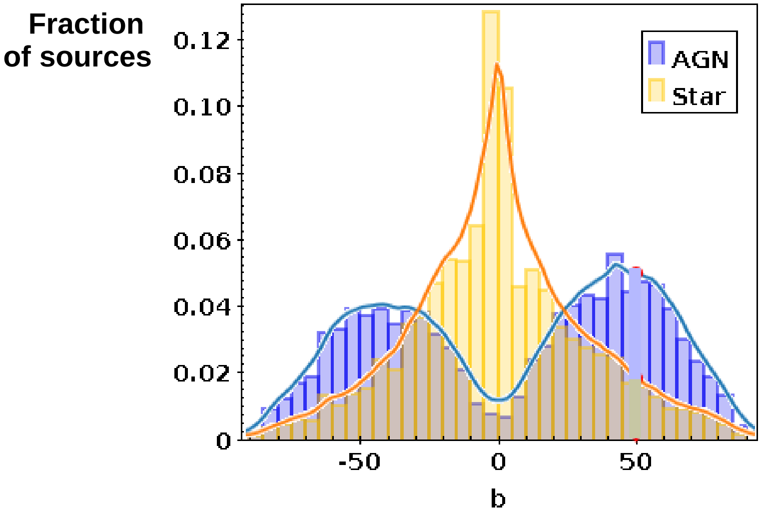
Context - hard and fast rules

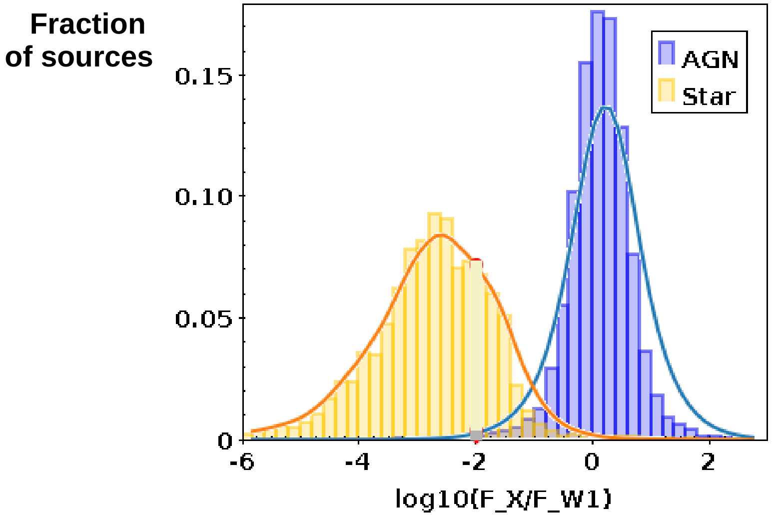
	AGN	Stars	X-ray binaries
Galactic latitude	abs(b)>20°	abs(b)<20°	abs(b)<20° OR extragalactic
X-ray / Optical flux ratio	>0.1, <10	<0.01	depends
X-ray variability (short-term)	low	<10	high
X-ray variability (long-term)	high but <100	low	high
X-ray spectrum	straight line	peak	depends

Inaccurate

Context - X-ray classification

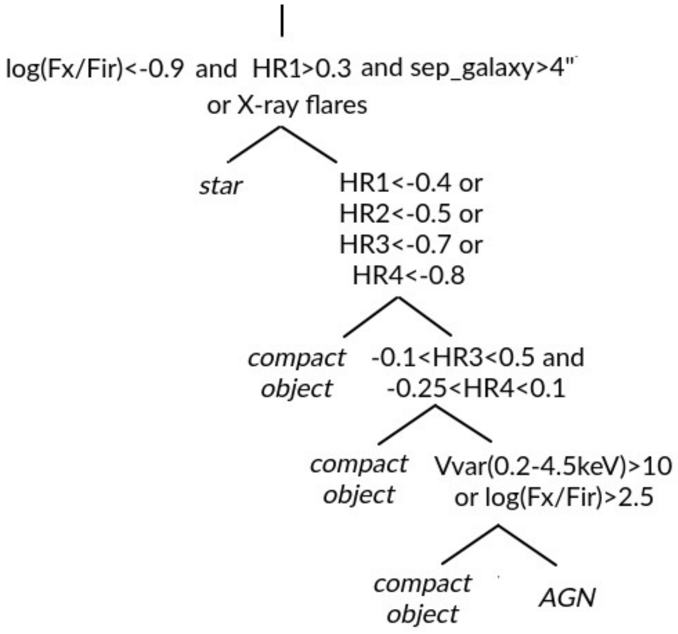
Statistical distributions of galactic latitude and F_X/F_{IR}





Context - X-ray classification

Decision tree of Lin+2012 for classification of 2XMMi-DR3 bright sources (~4000 sources)



Inaccurate

How to automatically classify X-ray sources?

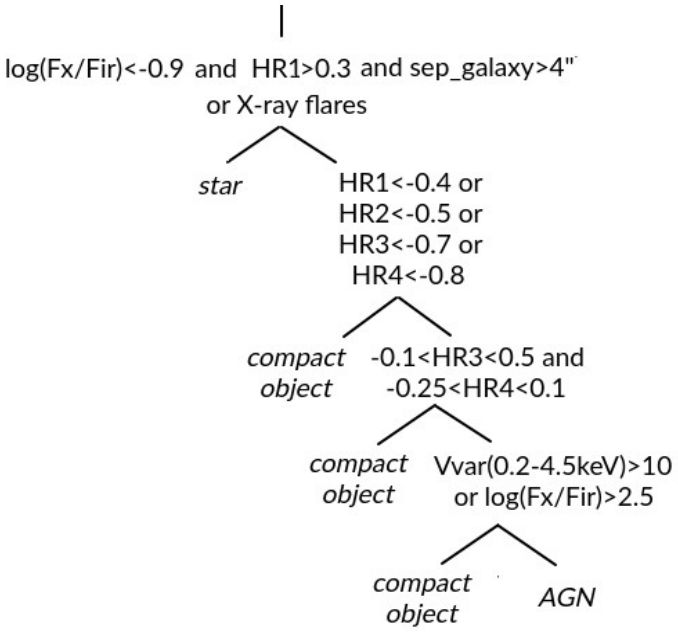
Machine learning (Arnason+ 2020)

- Naive Bayes (Tranin et al. submitted to A&A)
- Random Forest (e.g. Farrell+2015)

⇒ Based on a training sample

Example: in 4XMM-DR9

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



Decision tree of Lin+2012
for classification of 2XMMi-DR3
bright sources

Current classification works & issues

- Few are probabilistic
- Almost all are on small samples
- Summary : manual / decision tree / machine-learning
- Suboptimal catalogue enhancement
- Samples of known XRB, CV, TDE… are small

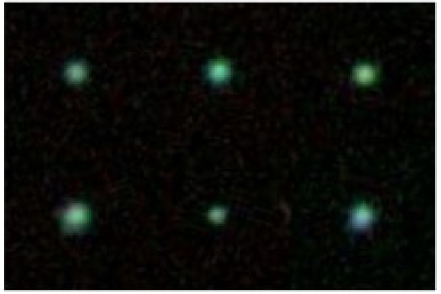
What could be done to improve them

- Favour probabilistic classifications
- Adapt them to data mining: large samples
- Find a trade-off between accuracy and interpretability
- Enhance catalogues properly
- Enlarge training samples

Context - citizen science

- Booming field since 20 years
- Wisdom of crowds
- Galaxy Zoo revolution
- Few projects in X-ray astronomy

→1 million galaxies classified, ~50/galaxy
→ 39 peer-reviewed publications in 2019

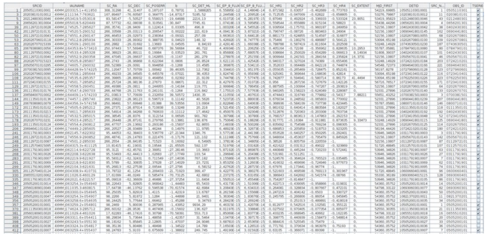


II. My work

Automatic classification

Citizen science experiment

Catalogue enhancement

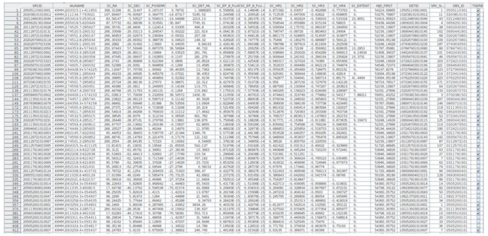


2SXPS / 4XMM-DR10
catalogue

Identification of known sources

⊗ SDSS QSO, HIPPARCOS...

Catalogue enhancement



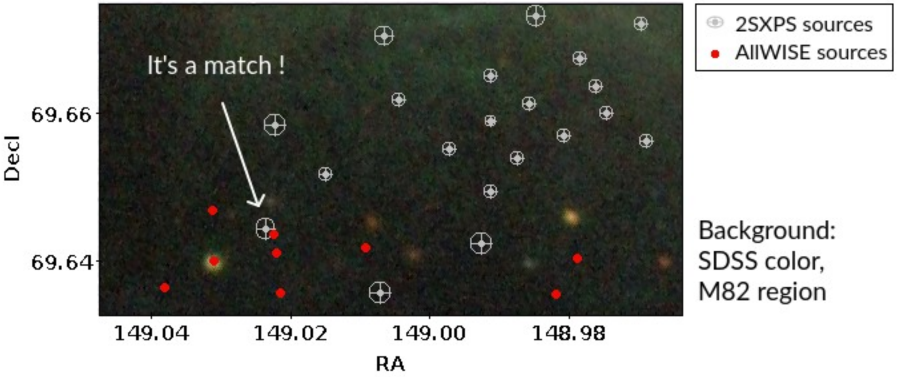
2SXPS / 4XMM-DR10 catalogue

Identification of known sources

⊗ SDSS QSO, HIPPARCOS...

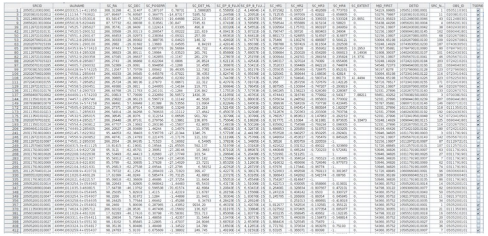
Multiwavelength associations

⊗ Gaia, AllWISE...



Background:
SDSS color,
M82 region

Catalogue enhancement



2SXPS / 4XMM-DR10
catalogue

Identification of known sources

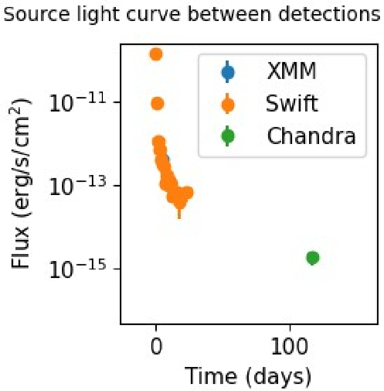
⊗ SDSS QSO, HIPPARCOS…

Multiwavelength associations

⊗ Gaia, AllWISE…

Long-term light curves
with all X-ray detections

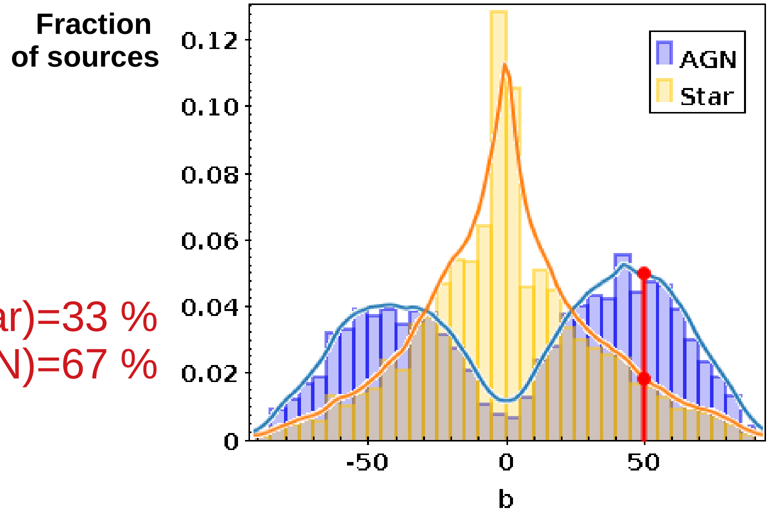
⊗ Swift, Chandra…



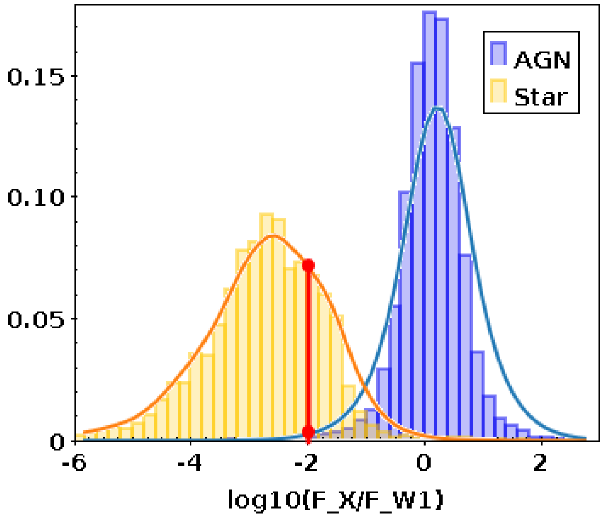
X-ray source classification

Unknown source
b=50°
$\log(F_X/F_{W1})$=-2
AGN or star ?

- Naive Bayes Classification principle



P(b|star)=33 %
P(b|AGN)=67 %



$P(F_X/F_{W1}|Star)$=93%
$P(F_X/F_{W1}|AGN)$=7%

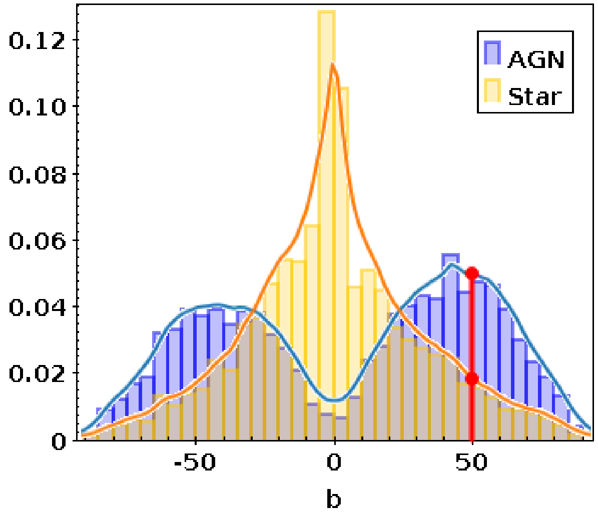
X-ray source classification

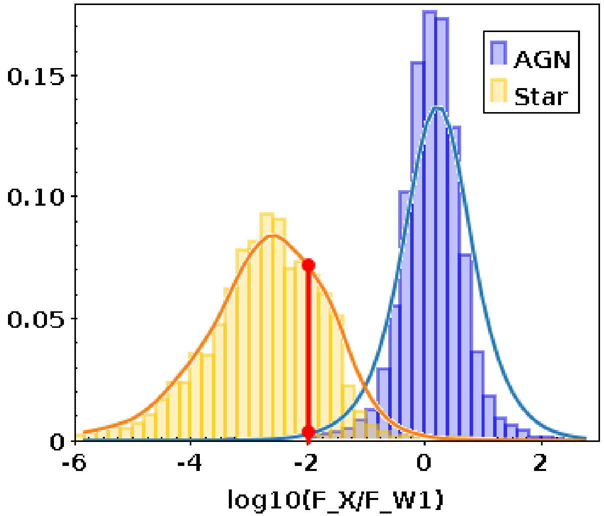
Unknown source
b=50°
$\log(F_X/F_{W1})$=-2
AGN or star ?

- Naive Bayes Classification principle

Fraction of sources

P(b|star)=33 %
P(b|AGN)=67 %





$P(F_X/F_{W1}|Star)=93\%$
$P(F_X/F_{W1}|AGN)=7\%$

Prior

$$\mathbb{P}(\mathrm{AGN}|D) = \frac{\mathcal{P}(\mathrm{AGN})\mathcal{L}(\mathrm{AGN}|D)}{\mathcal{P}(\mathrm{AGN})\mathcal{L}(\mathrm{AGN}|D) + \mathcal{P}(Star)\mathcal{L}(Star|D)}$$ = 19% here

X-ray source classification

- Naive Bayes Classification *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: $\alpha_{location}$, $\alpha_{spectrum}$, $\alpha_{variability}$, $\alpha_{counterparts}$

⇒ CLAXBOI code, Tranin et al. submitted to A&A

Classification results / reference sample

2SXPS

Truth →	AGN	Star	XRB	CV	Total cl.
→AGN	19737	87	77	238	20139
→Star	141	4954	19	4	5118
→XRB	193	37	248	41	519
→CV	44	0	3	36	83
Total	20115	5078	347	319	Average
recall (%)	98.1	97.6	71.5	11.3	69.6
precision (%)	94.9	96.6	82.3	52.4	81.6

Quantity to maximize :
f_1-score of XRB

$$f_1 = 2(recall^{-1} + precision^{-1})^{-1}$$

4XMM-DR10

Truth →	AGN	Star	XRB	CV	Total cl.
→AGN	18057	25	122	144	18348
→Star	55	6239	10	2	6306
→XRB	241	31	398	49	719
→CV	27	0	5	55	87
Total	18380	6295	535	250	Average
recall (%)	98.2	99.1	74.4	34.8	76.6
precision (%)	95.8	98.6	79.0	71.5	86.2

Tranin et al. submitted to A&A

Comparison to another machine learning method

- Optimized naive Bayes

2SXPS

Truth →	AGN	Star	XRB	CV	Total cl.
→AGN	19737	87	77	238	20139
→Star	141	4954	19	4	5118
→XRB	193	37	248	41	519
→CV	44	0	3	36	83
Total	20115	5078	347	319	Average
recall (%)	98.1	97.6	71.5	11.3	69.6
precision (%)	94.9	96.6	82.3	52.4	81.6

- Random Forest

2SXPS

Truth →	AGN	Star	XRB	CV	Total cl.
→AGN	5957	21	23	72	6073
→Star	19	1506	5	1	1531
→XRB	41	13	66	10	130
→CV	0	0	3	21	24
Total	6017	1540	97	104	Average
recall (%)	99.0	97.8	68.0	23.1	72.0
precision (%)	95.1	97.6	84.8	68.1	86.4

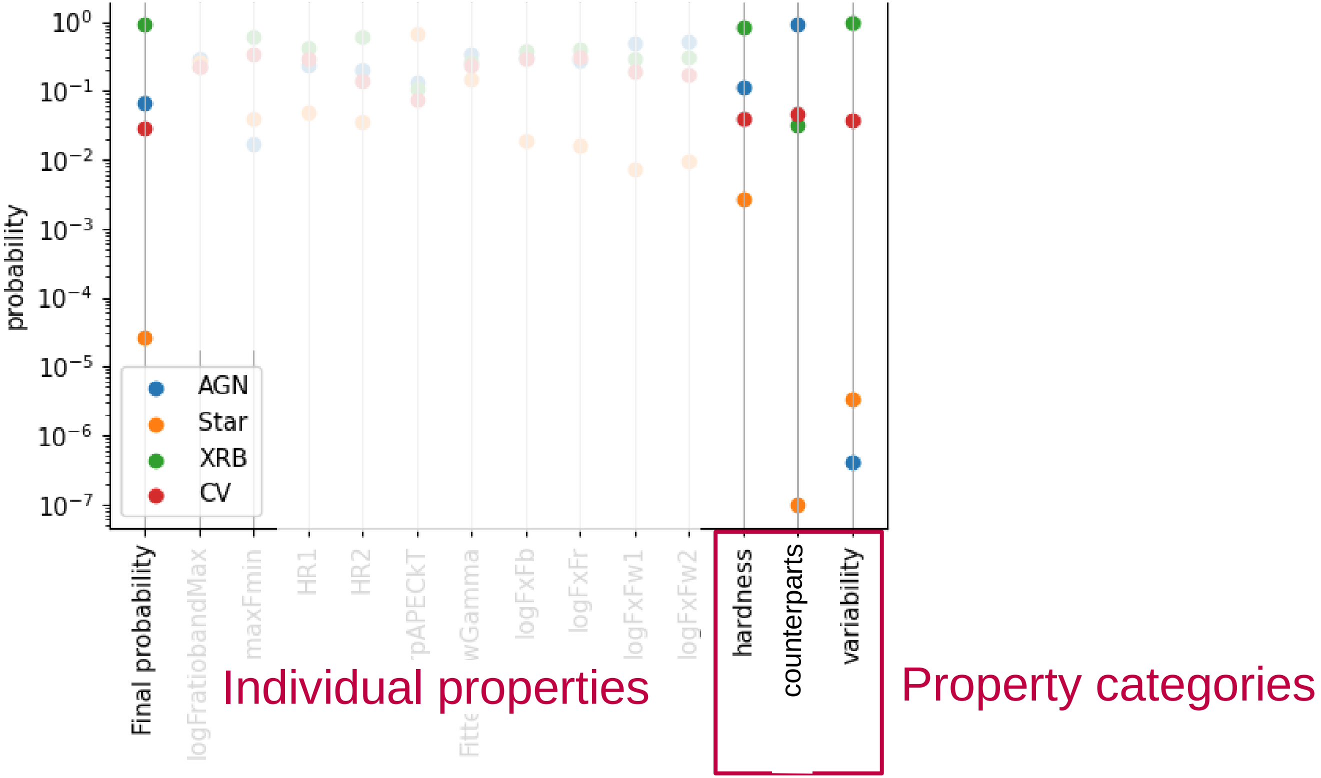
⇒ *better results on XRB + better interpretability*

Tranin et al. submitted to A&A

Interpretability of the results

- One classification product: the probability "track"

⇒ this object was classified as X-ray binary because of
  - its variability
  (- its hardness)



Tranin et al. submitted to A&A

Classification results / test sample

- What is the test sample?
  ⇒ sample that you could classify manually

  ⇒ Following at least 2 of the following:

  - Optical counterpart
  - Infrared counterpart
  - Spectrum acquired, or S/N>10
  - Several X-ray detections

- 55 % of each X-ray catalogue!

Classification results / test sample

- >200 sources analyzed manually (Swift+XMM)
- >90% accuracy on the sources classified as AGN and stars
- Between 30 % and 65 % accuracy on … as XRB

Tranin et al. submitted to A&A

Classification results / test sample

- >200 sources analyzed manually (Swift+XMM)
- >90% accuracy on the sources classified as AGN and stars
- Between 30 % and 65 % accuracy on … as XRB
  - → Issues in the multiwavelength matching algorithm
  - → Issues in the multi-instrument long-term light curves
  - → Lack of sufficient training sample

Tranin et al. submitted to A&A

Classification results / test sample

- Estimations (4XMM-DR10):
  - 180 000 new AGN
  - 30 000 new stars
  - 8 000 new XRB

(test sample having no match in Simbad, accuracy-corrected)

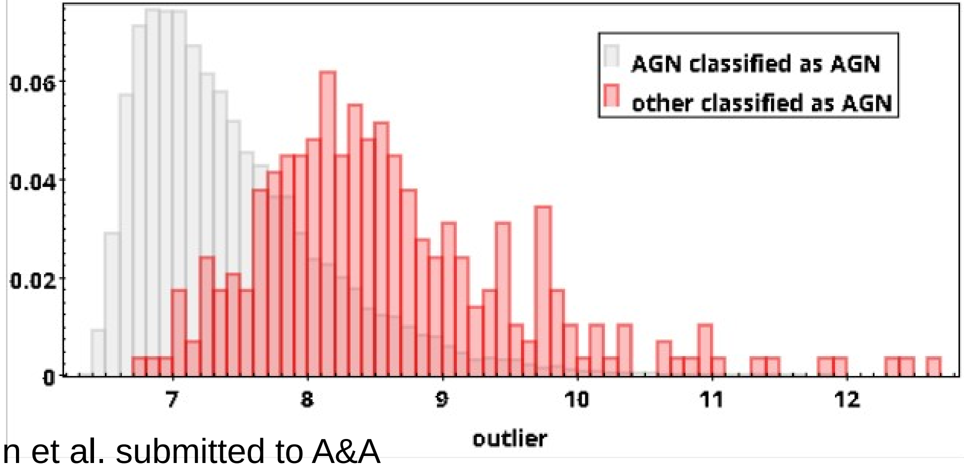
Tranin et al. submitted to A&A

Outlier measure

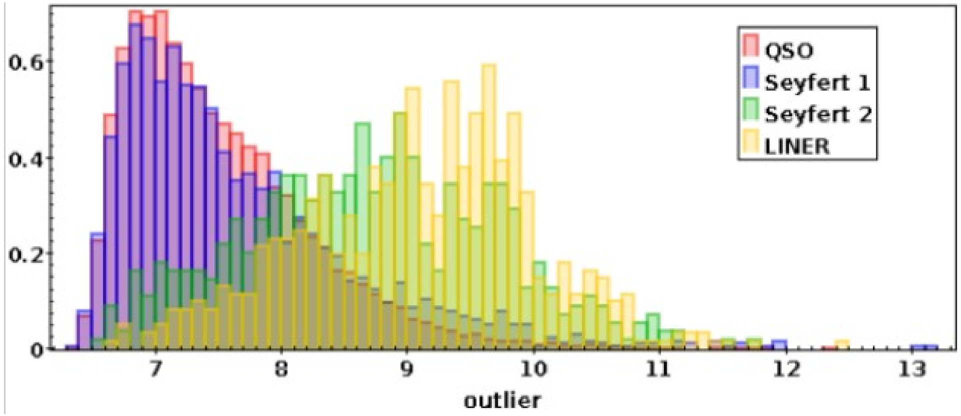
$$O.M. = -\log\left(\mathcal{P}(\mathrm{c}) \times \prod_{t\in\{\mathrm{cat}\}} \mathcal{L}(t|\mathrm{c})^{\alpha_t/\sum_{t\in\{\mathrm{cat}\}}\alpha_t}\right)$$

Interpretation: typical height of the class c distribution at this point of the parameter space

Fraction of sources

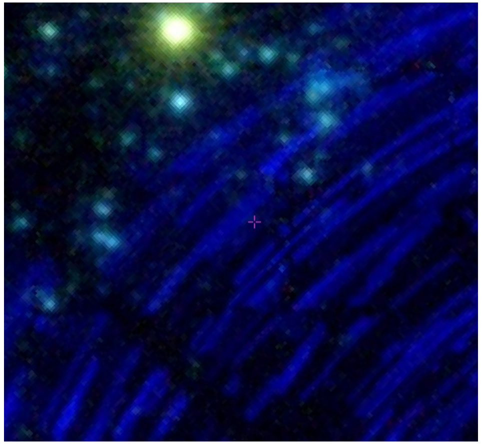


Fraction of sources



Tranin et al. submitted to A&A

Outlier measure



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

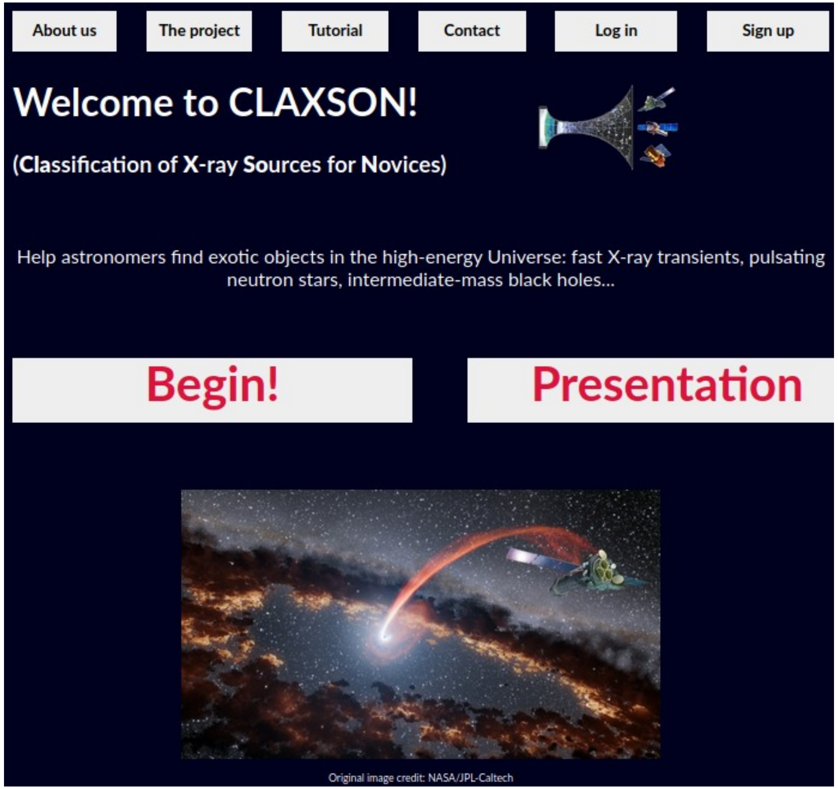
Tranin et al. submitted to A&A

Improvements of X-ray classification

- Favour probabilistic classifications ✓
- Adapt them to data mining: large samples ✓
- Find a trade-off between accuracy and interpretability ✓
- Enhance catalogues properly ✓
- Enlarge training samples ✗
  → citizen science

Enlarging the training samples

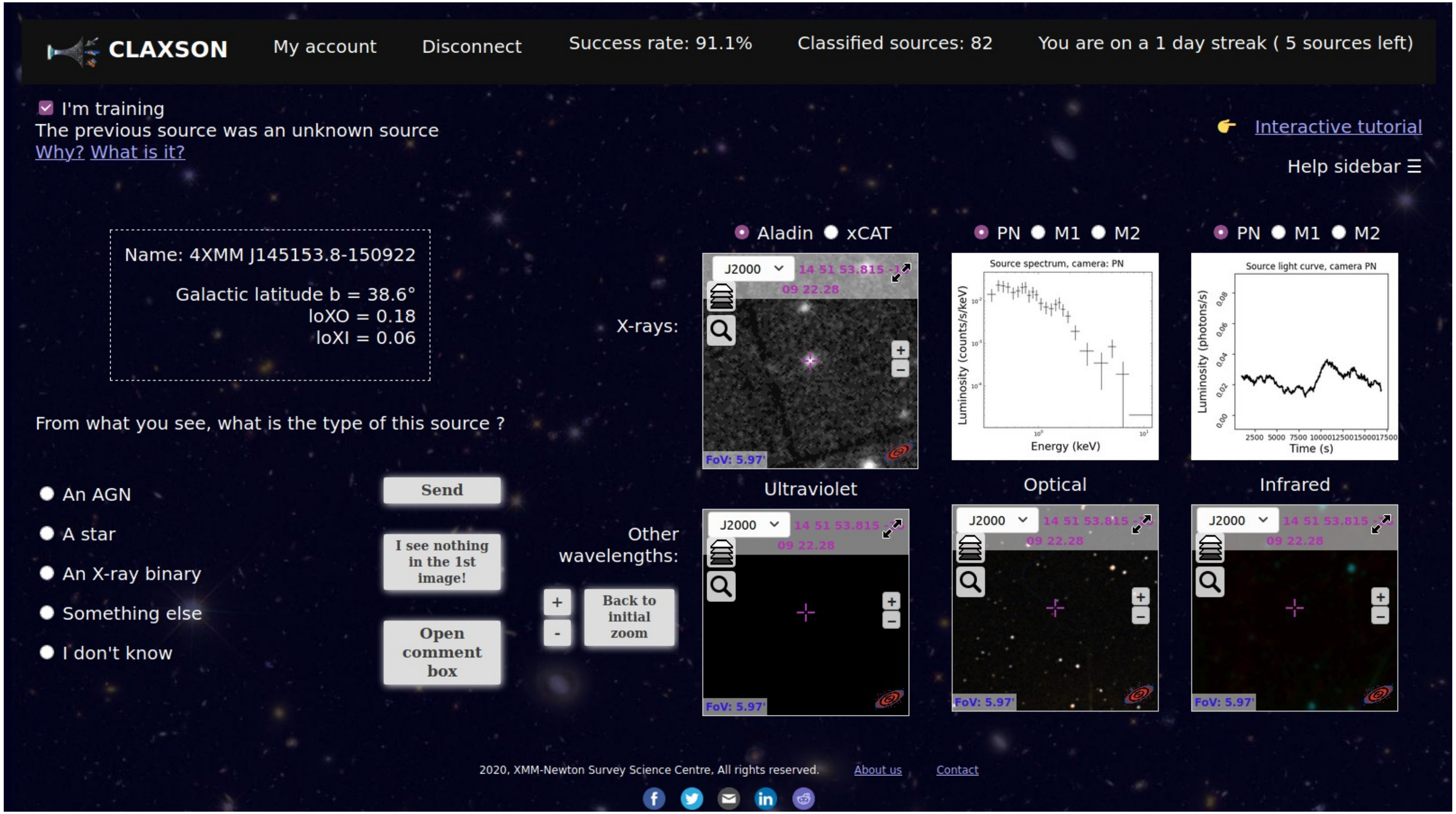
http://xmm-ssc.irap.omp.eu/claxson



- Alternative: Citizen Science

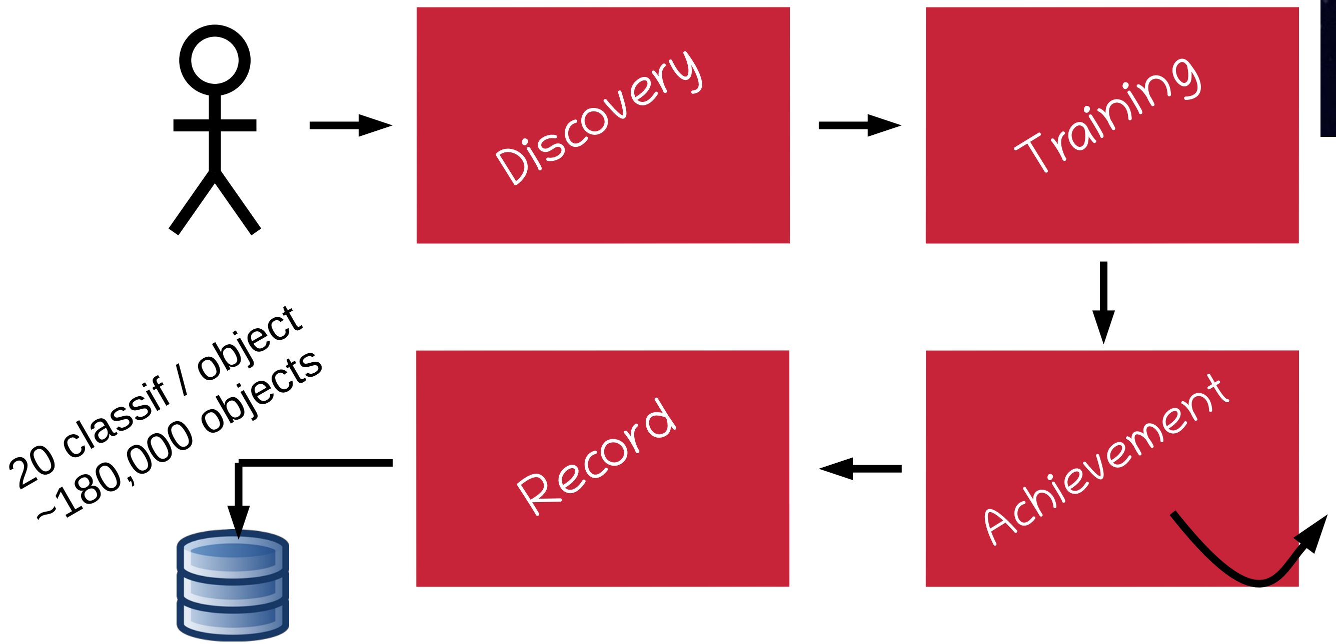
- Goals:
  - **Obtain a large training sample for ML**
  - + Serendipitous results?

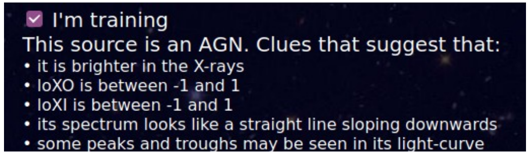
CLAXSON – quick look



CLAXSON

- How to ensure classification reliability?





⇒ trial and error

To reach level 2:
≥ 20 classif
75% success on the last 20

CLAXSON

~50 new X-ray binaries so far

- 972 unknown sources classified >10 times, by 46 users
- Mean success rate among trained users : 82 %
- *(preliminary)* 68 % of the end results agree with the output of the Naive Bayes classifier

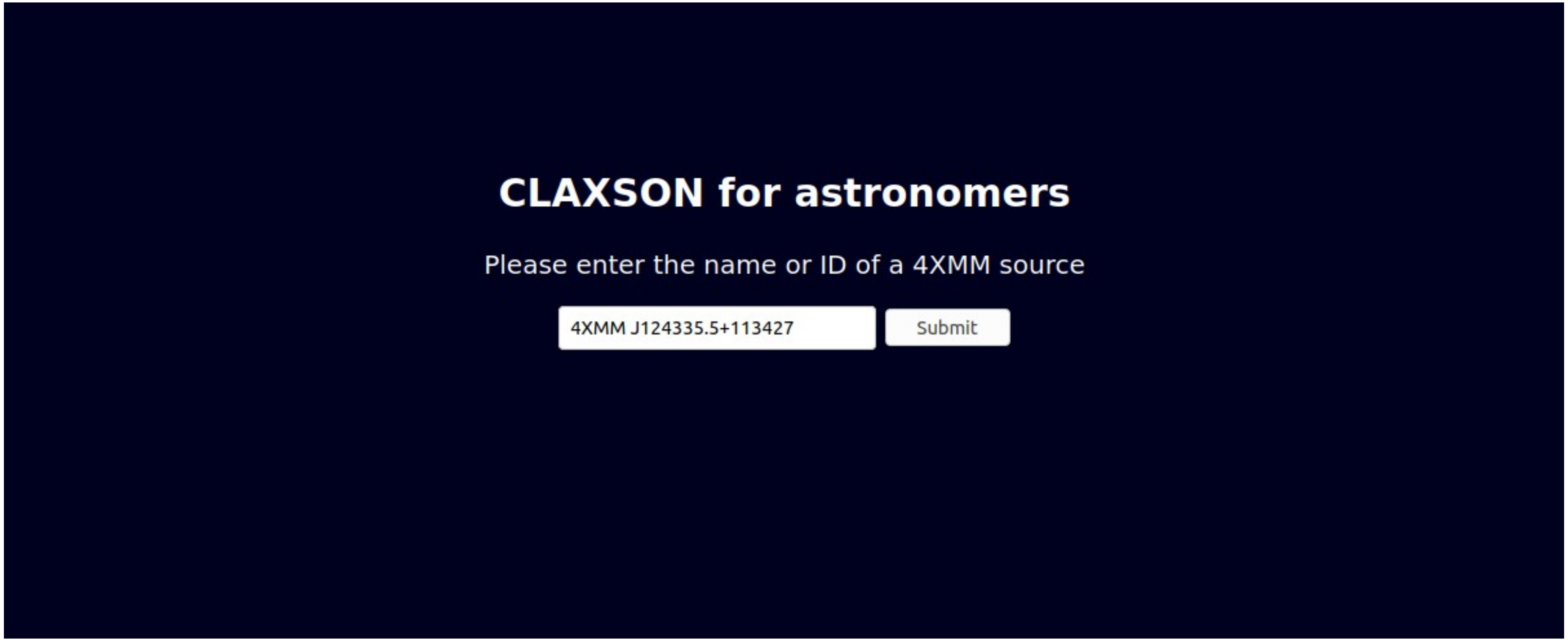
--- Ranking of level 2 users ---

Rank	User	Number of classifications	Since 1 week	Success rate
1	algol	9800	163	92.4
2	KrystianBykowski	4509	305	84.2
3	dani.gi	4466	6	86.5
4	Tsuki Eeen no	4032	0	64.6
5	SimonLeKlaxon	3847	317	91.8

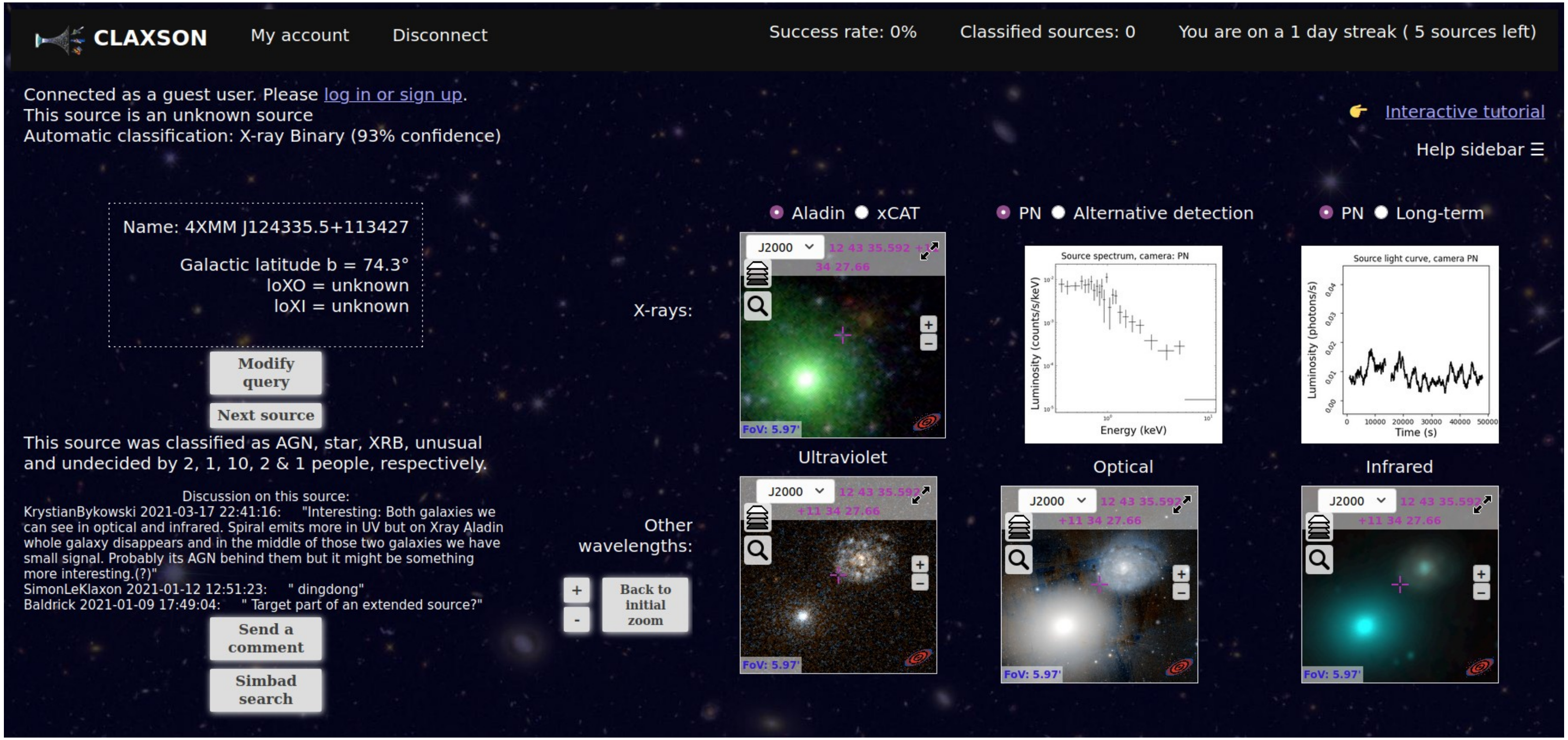
CLAXSON

- Outlook:
  - More communication
  - Translation (currently:)
  - More gamification
  - Interface ↔ researchers and volunteers

Versatile tool

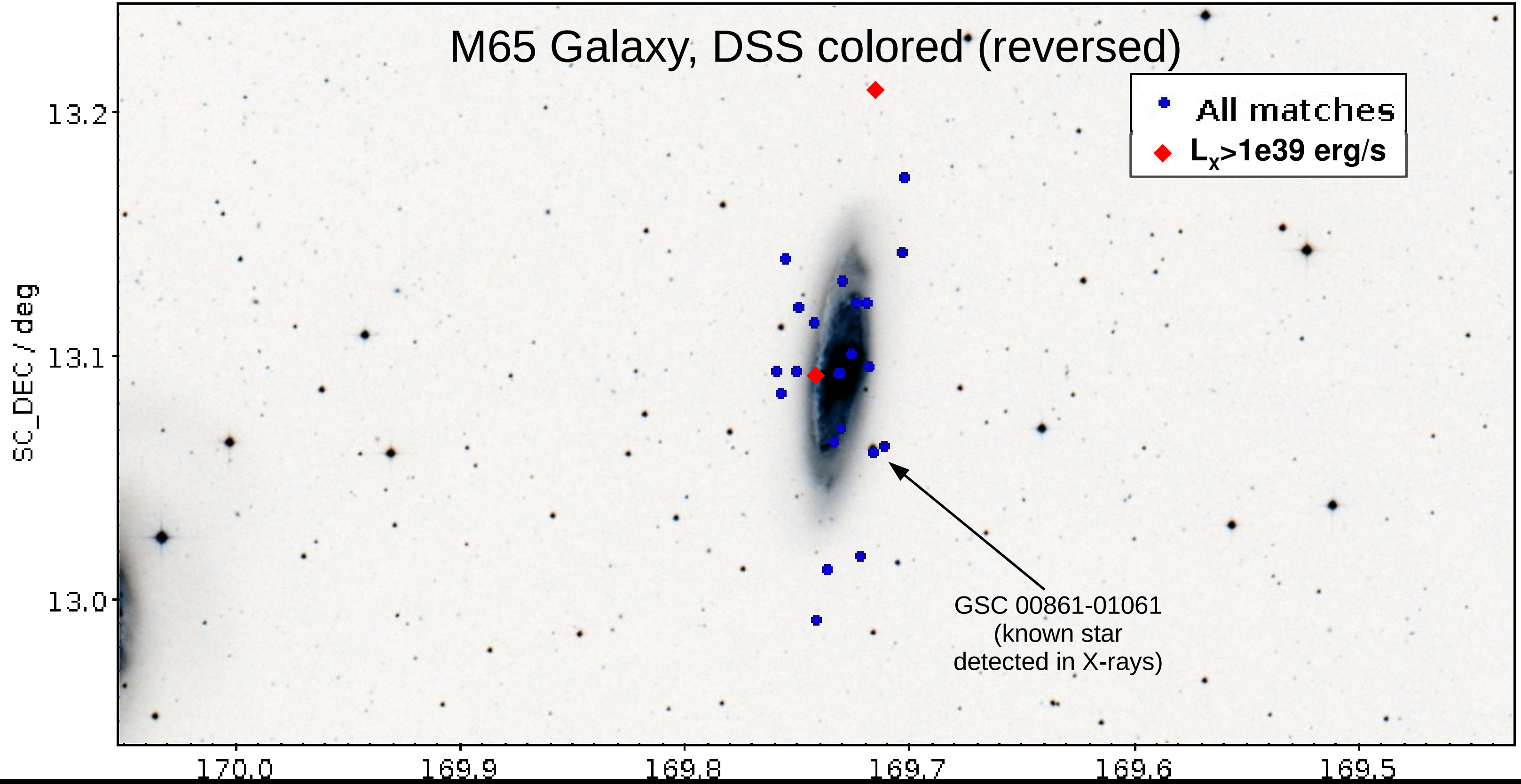


Versatile tool



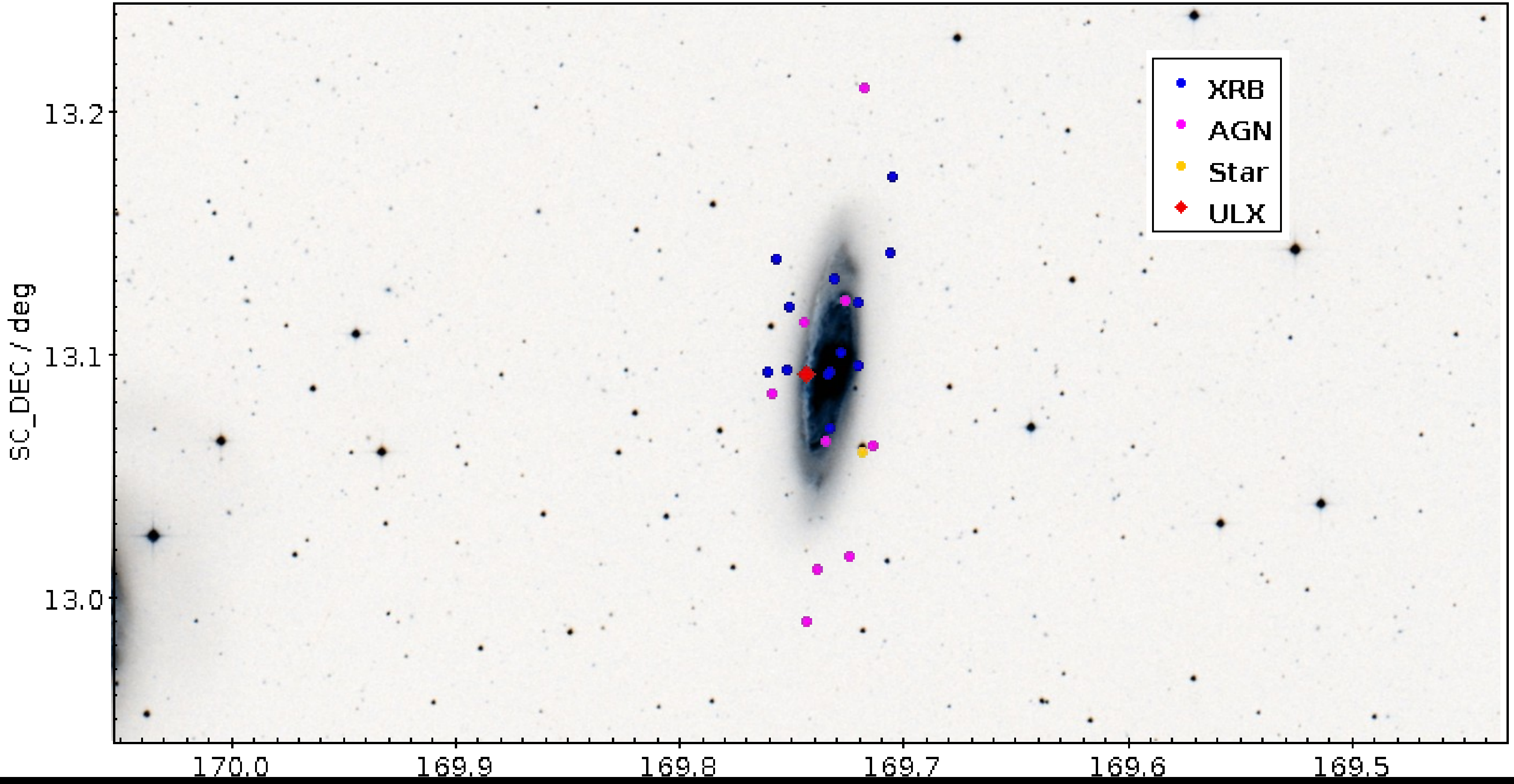
Back to classification results…
Applications!

First results - ULX

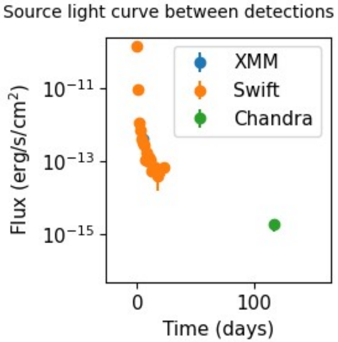


First results - ULX

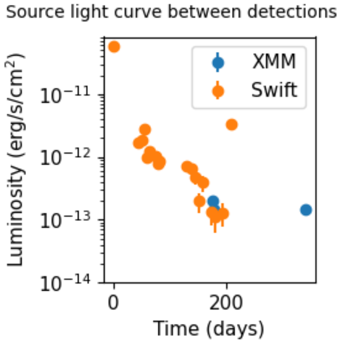
⇒ 1 confirmed ULX



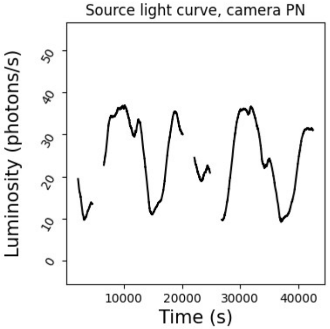
Crucial criterion = Variability !



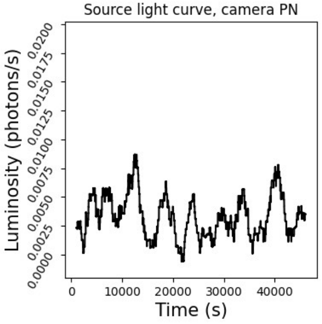
GRB



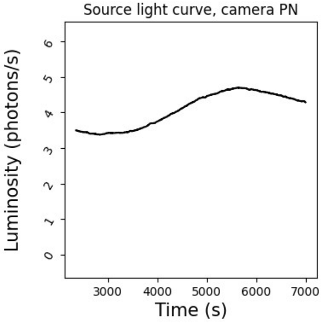
TDE



Cataclysmic variable



X-ray binary



Seyfert-1

Crucial criterion = Variability !

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

→ $\alpha_{variability}$ is the biggest coefficient after optimization

Thanks for your attention

Credit: Mark A. Garlick / Sheffield University / AFP Photo