Classifying most of *XMM-Newton* sources: challenge accepted

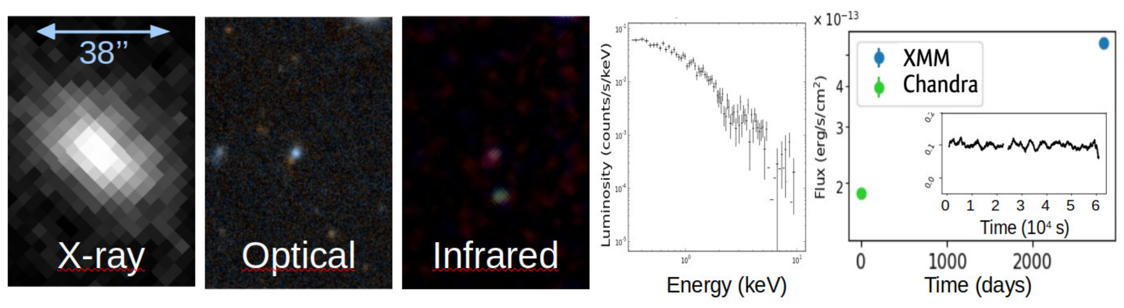
Hugo Tranin
hugo.tranin@irap.omp.eu
IRAP, Université de Toulouse, CNRS, CNES, 9 avenue du Colonel Roche, 31028 Toulouse, France

1. But why? Why would you do this?

X-ray observatories like *Swift*, *XMM-Newton* and *Chandra* observed about 1 million sources in the past 20 years. While most of them are still unstudied, constraining their nature is fundamental to find larger samples of exotic sources (such as tidal disruption events, changing-look AGN, ultraluminous X-ray sources, intermediate mass black holes...). Developing an automatic classification adapted to this data mining task will be crucial with the development of surveys of unprecedented size, such as the Vera Rubin observatory, SKA and Athena, and the search for counterparts of multi-messenger events.

2. The work of others before me

An X-ray source can be classified manually by using its location, the shape of its spectrum and light-curves (either intra or inter-observations) and the presence and magnitude of its multi-wavelength counterparts. You can use this approach to infer hard and fast rules, however the resulting classification will be inaccurate, because the property distributions of different classes overlap (Figure 2) (case of the decision tree in Lin et al. 2012). Other works rely on machine learning techniques such as Random Forest (e.g. Farrell et al. 2015, Arnason et al. 2020) but their results are hardly interpretable: **no classification is both accurate and easily interpretable**. Another caveat is that **they are all applied on small samples** of a few ~1000 sources (having the best quality), and that the reference samples of known sources are small for some classes (X-ray binaries, cataclysmic variables...). Sometimes their X-ray samples are also poorly enhanced, e.g. when they do not include the detections from other X-ray observatories in the long-term light curves, or when they do not search for counterparts in deep optical/infrared catalogues.

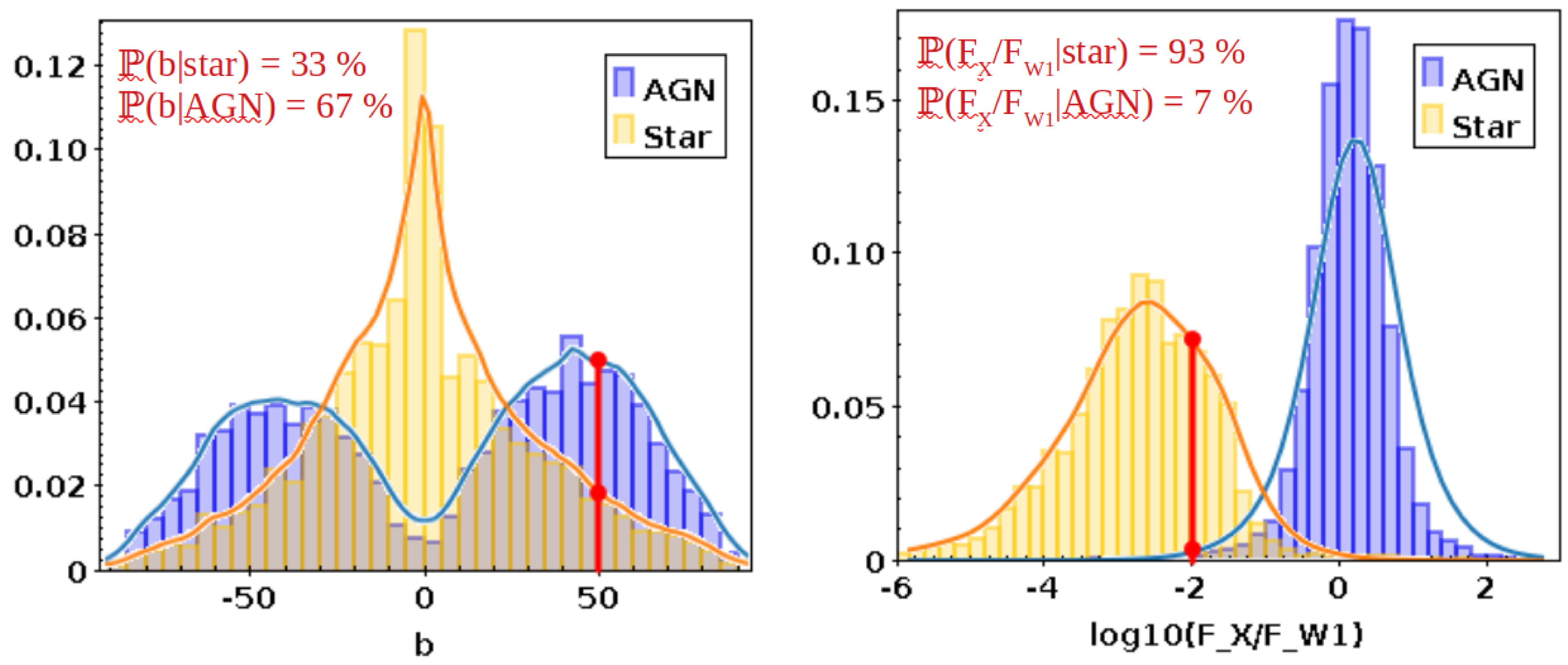


Multiwavelength images, X-ray spectrum and light curves of typical AGN (4XMM J214041.4-234718).

3. Everyone loves Bayesianism

Besides Random Forest, other machine learning techniques can be used to classify X-ray sources. One of them is the so-called "Naive Bayes Classifier" (Murphy et al., 2006), which is intuitive, probabilistic, highly interpretable and adapted to small reference samples (Table 1). Say that you want to classify an unknown source as "Star" or "AGN", and you known its galactic latitude $b = 50°$ and its X-ray to infrared flux ratio $F_X/F_{W1} = 0.01$. According to the distribution of b and $\log(F_X/F_{W1})$, with prior proportions $\mathcal{P}(\mathrm{AGN}) = 0.75$ and $\mathcal{P}(\mathrm{Star}) = 0.25$, we obtain the posterior probability:

$$\mathbb{P}(\mathrm{Star}|data) = \frac{\mathcal{P}(\mathrm{Star})\mathcal{L}(\mathrm{Star}|data)}{\mathcal{P}(\mathrm{AGN})\mathcal{L}(\mathrm{AGN}|data) + \mathcal{P}(\mathrm{Star})\mathcal{L}(\mathrm{Star}|data)} \approx 81\% \qquad (1)$$



Densities of 2 properties (Galactic latitude and X-ray to infrared flux ratio) in the sample of known AGN and stars.

In practice, we classified XMM sources as AGN, star, X-ray binary (XRB) or cataclysmic variable (CV). We used 13 of their properties, related to 4 categories weighted by a coefficient: $\alpha_{\mathrm{location}}, \alpha_{\mathrm{spectrum}}, \alpha_{\mathrm{variability}}, \alpha_{\mathrm{multiwavelength}}$, fine-tuned to optimize the classification results, i.e. **maximizing the *recall* and *precision* of the XRB class** (next panel). Equation (1) becomes:

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

4. Some cool results

After cross-correlating the 4XMM-DR10 catalogue (Webb et al., 2020) with many, many others – covering known AGN, stars, XRB, CV, plus X-ray, optical and infrared sources – and following the method described in the panel you just read, we obtained the results detailed in Table 1: high *recall* (fraction of this class successfully retrieved) and *precision* (fraction of true positives among sources with this classification) for AGN and stars, and a quite good performance for XRB as well (Tranin et al. submitted to A&A). The test sample, chosen to be all sources which could be classified manually – i.e. having at least 2 of these: **(a,b)** an optical/infrared counterpart, **(c)** a measured spectrum or $S/N > 10$, **(d)** several X-ray detections – represents **55% of the catalogue** (315573 sources)!

Classified as ↓	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	All
recall (%)	98.2	99.1	**74.4**	34.8	95.5
precision (%)	95.8	98.6	**79.0**	71.5	94.8

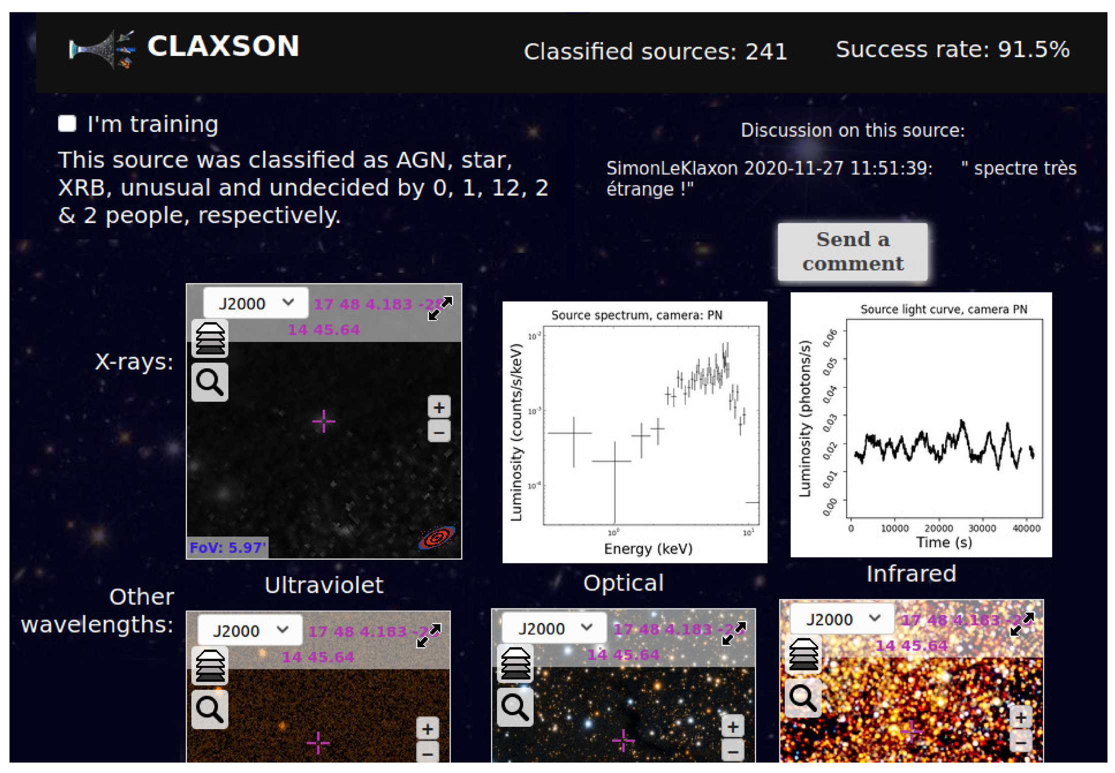
	Total classifications	*precision* (A)	Total outliers(B)
→AGN	120061	>90%	7119
→Star	19159	>90%	3878
→XRB	47516	30–65%(C)	7114
→CV	2484	~65%	1256
Total	315573	~82%	19367

(A) Manual estimation on a sample of >200 sources.
(B) Nice sources having an outlier measure > 10, not defined here.
(C) 65% when spurious multiwavelength correlations are removed.

Number counts and metrics of the classification applied to the reference sample (left) and test sample (right) of XMM.

5. Exploitation of citizen scientists

To improve the classification of XRB and CV, we want to enlarge their reference samples by using citizen science. We launched the platform CLAXSON (`http://xmm-ssc.irap.omp.eu/claxson`), on which volunteers can learn how to classify XMM sources manually (trial and error on known objects) and then classify unknown sources. Each object is given to several volunteers to obtain reliable classifications. So far, **46 volunteers performed 40000 classifications** of unknown sources, with a mean success rate of 82%. They found ~50 new XRB.



Glimpse of CLAXSON feedback on an unknown source

6. References

R. M. Arnason, P. Barmby, and N. Vulic. Identifying new X-ray binary candidates in M31 using random forest classification. MNRAS, 492(4):5075–5088, 2020.

S. A. Farrell, T. Murphy, and K. K. Lo. Autoclassification of the Variable 3XMM Sources Using the Random Forest Machine Learning Algorithm. ApJ, 813(1):28, 2015.

D. Lin, N. A. Webb, and D. Barret. Classification of X-Ray Sources in the XMM-Newton Serendipitous Source Catalog. ApJ, 756(1):27, 2012.

K. P. Murphy et al. Naive bayes classifiers. University of British Columbia, 18:60, 2006.

N. A. Webb, M. Coriat, and I. Traulsen el al. The XMM-Newton serendipitous survey. IX. The fourth XMM-Newton serendipitous source catalogue. A&A, 641:A136, 2020.