Deep learning for the selection of YSO candidates from IR surveys

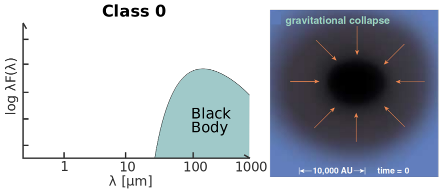
David Cornu, PhD Student

Supervised by: J. Montillaud & A. Robin

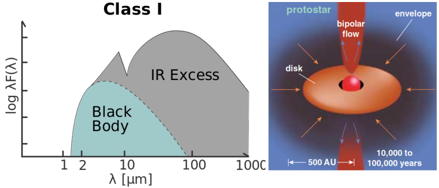
Institut UTINAM, Univ. Bourgogne Franche-Comté, OSU THETA, Besançon, France

Artificial Intelligence in Astronomy - 2019

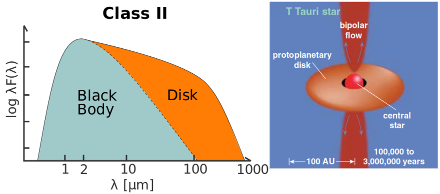
Young Stellar Objects YSOs → characterize star-forming regions.



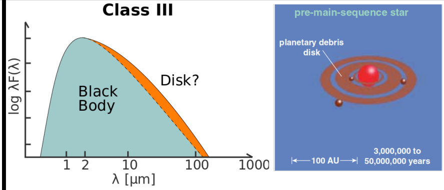
Faint protostar deeply embedded



Protostar with infalling envelop



Pre main sequence star with thick disk

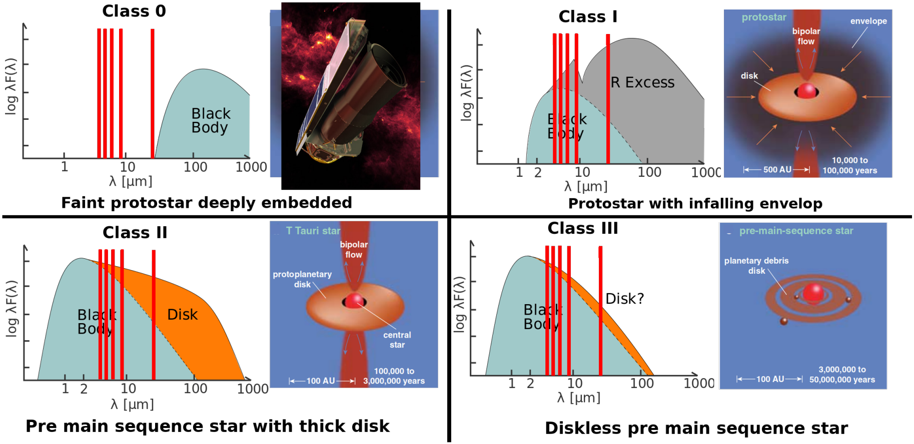


Diskless pre main sequence star

Classified by evolutionary steps using their infrared SEDs.

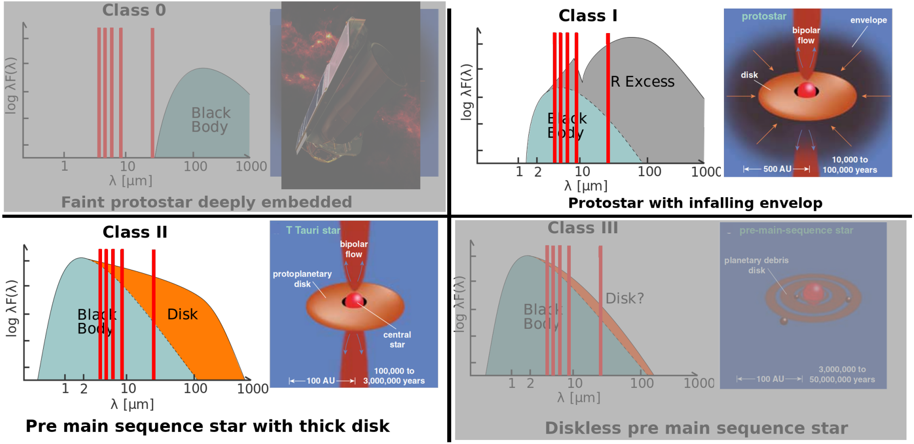
Young Stellar Objects

Young Stellar Objects YSOs $\rightarrow$ characterize star-forming regions.



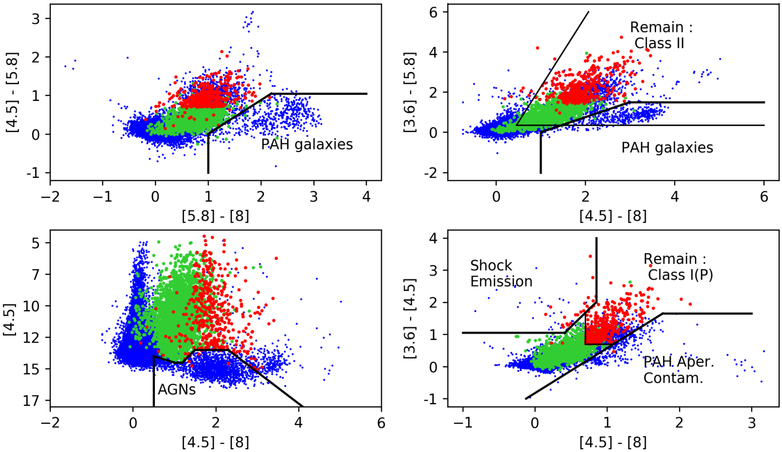
Classified by evolutionary steps using their infrared SEDs.

Young Stellar Objects YSOs $\rightarrow$ characterize star-forming regions.



Classified by evolutionary steps using their infrared SEDs.

Commonly used classification scheme



Adapted from Gutermuth et al. (2009) method (G09) using IRAC at 3.6, 4.5, 5.8, 8.0 μm and MIPS at 24 μm. Class I in red and Class II in green, and Other in blue.

Limitation: Arbitrariness remain in the placement of the cuts, objects near the cuts are less robustly classified, but it is difficult to quantify.

Machine Learning

⇒ **Core concept:** extract statistical information about a dataset and adapt the response accordingly

Supervised

- A training set with the expected targets provided

Unsupervised

- Dataset without targets.

Machine Learning

$\Rightarrow$ **Core concept:** extract statistical information about a dataset and adapt the response accordingly

Supervised

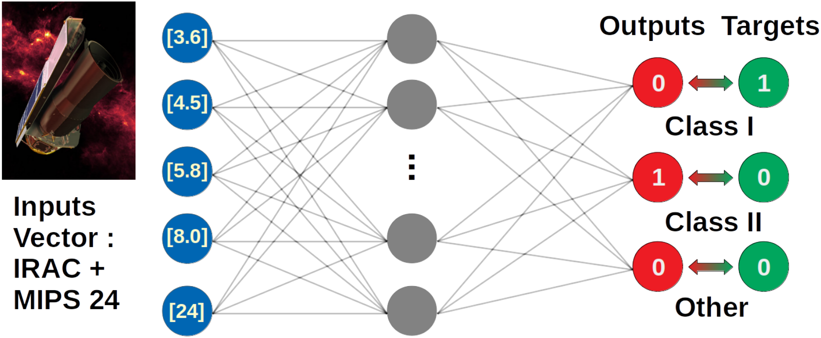
- A training set with the expected targets provided

Unsupervised

- Dataset without targets.

Main objective: Replacing straight cuts in YSO selection with non-linear and statistically learned splittings

YSO classification with MLP

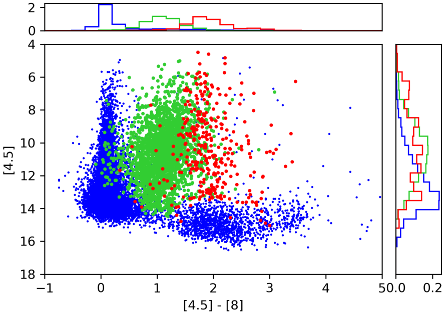


Network dimensions:

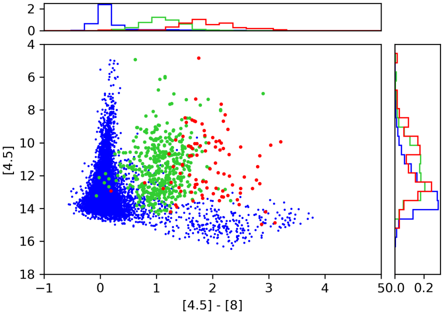
- Number of input nodes: number of dimensions of the problem.
  → 10 nodes ($5\,mag$ + 5 σ_{mag})
- Number of hidden layers: no impact on results
  → 1 hidden layer is enough
- Number of hidden neurons: $\propto$ difficulty of the problem
  → 1 neuron ≈ 1 hyper-plane in the input parameter space
- Number of output neurons: choose an encoding method.
  → Classification, one neuron per class ⇒ **SOFT-MAX activation**

Ensure diversity

Different star-forming regions ⇒ **cover different parts of the input parameter space.**



Orion



NGC 2264

Data selection and preparation

Spitzer datasets used:

1. Orion survey from Megeath et al. (2012)
2. NGC 2264 / Mon OB1 survey from Rapson et al. (2014)
3. Near 1kpc clouds from Gutermuth et al. (2009)

Use adapted G09 method $\Rightarrow$ define a labeled dataset.

Data selection and preparation

Spitzer datasets used:

1. Orion survey from Megeath et al. (2012)
2. NGC 2264 / Mon OB1 survey from Rapson et al. (2014)
3. Near 1kpc clouds from Gutermuth et al. (2009)

Use adapted G09 method $\Rightarrow$ define a labeled dataset.

Separate the labeled datasets into:

a training set (majority)	**a test set** (small part)
Used to perform training	Saved away to assess results quality

Data selection and preparation

Spitzer datasets used:

1. Orion survey from Megeath et al. (2012)
2. NGC 2264 / Mon OB1 survey from Rapson et al. (2014)
3. Near 1kpc clouds from Gutermuth et al. (2009)

Use adapted G09 method ⇒ define a labeled dataset.

Separate the labeled datasets into:

a training set (majority)	**a test set** (small part)
Used to perform training	Saved away to assess results quality

Labeled dataset : 414 CI, 2659 CII and 23830 Others

⇒ **Strong Imbalance**

Imbalance in results

Labeled dataset : 414 CI, 2659 CII and 23830 Others

Machine learning algorithms are made to work on balanced datasets.

Imbalance in results

Labeled dataset : 414 CI, 2659 CII and 23830 Others

Machine learning algorithms are made to work on balanced datasets.

Example on disease detection, the majority of the tested persons are not sick AND the cure presents risks $\Rightarrow$ **must avoid to give unnecessary medication**

Imbalance in results

Labeled dataset : 414 CI, 2659 CII and 23830 Others

Machine learning algorithms are made to work on balanced datasets.

Example on disease detection, the majority of the tested persons are not sick AND the cure presents risks $\Rightarrow$ **must avoid to give unnecessary medication**

Predicted

Actual	Class	Unhealthy	Healthy	Recall
	Unhealthy	8	2	80%
	Healthy	7	93	93%
	Precision	53.3%	97.9%	91.8%

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$TP \equiv$ True Positive $\quad TN \equiv$ True Negative

$FP \equiv$ False Positive $\quad FN \equiv$ False Negative

Imbalance in results

Labeled dataset : 414 CI, 2659 CII and 23830 Others

Machine learning algorithms are made to work on balanced datasets.

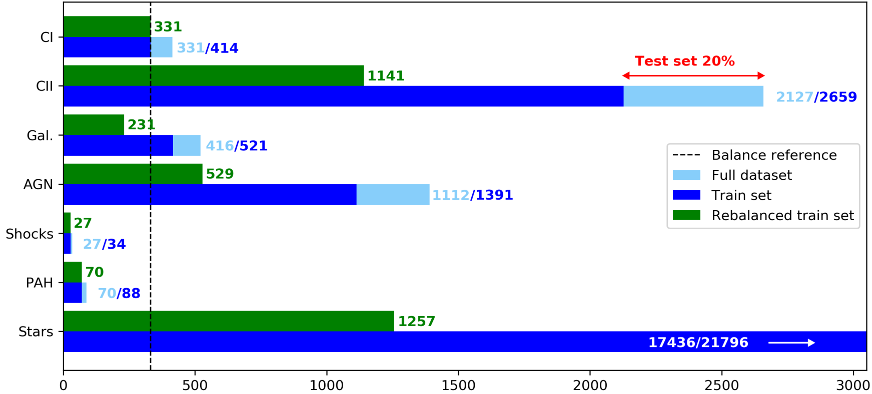
Example on disease detection, the majority of the tested persons are not sick AND the cure presents risks $\Rightarrow$ **must avoid to give unnecessary medication**

		Predicted		
	Class	Unhealthy	Healthy	Recall
Actual	Unhealthy	8	2	80%
Actual	Healthy	7	93	93%
	Precision	53.3%	97.9%	91.8%

Results of a classification must be tested on true use case scenario using "Observational proportions"

Imbalanced learning difficulty

Various methods can be applied to re-balance (mock data, weighting, ...).



- Control the impact of each class in the training set
- Must be $\propto$ input parameter space coverage
- Must keep enough objects apart in Obs. prop. for the test set (Saturation)

Summary of Imbalance precautions

Precautions regarding the size and balance of the dataset for classification:

	Large sample	**Small sample**
Balanced	No issue	Param. space coverage
Imbalanced	Obs. Proportions	Param. space coverage Obs. Proportions Must avoid dilution

Overall, having a large sample mitigates the difficulties caused by imbalanced datasets.

All clouds training results

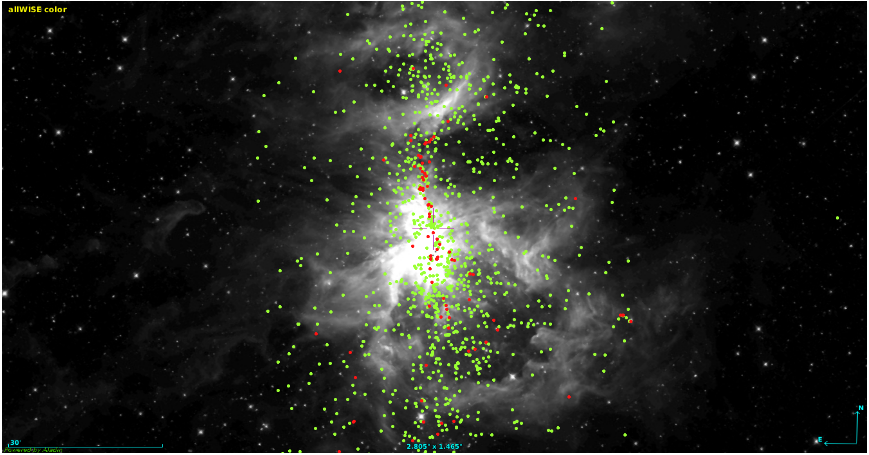
Results of the training from the near 1kpc dataset described before and using proper training proportions.

Predicted

	Class	YSO CI	YSO CII	Other	Recall
Actual	YSO CI	75	3	4	91.5%
	YSO CII	6	515	8	97.0%
	Other	8	42	4714	99.0%
	Precision	84.3%	92.0%	99.7%	98.6%

Test set size: 20% of the combined Orion and 2264 labeled dataset, using averaged observational proportions.

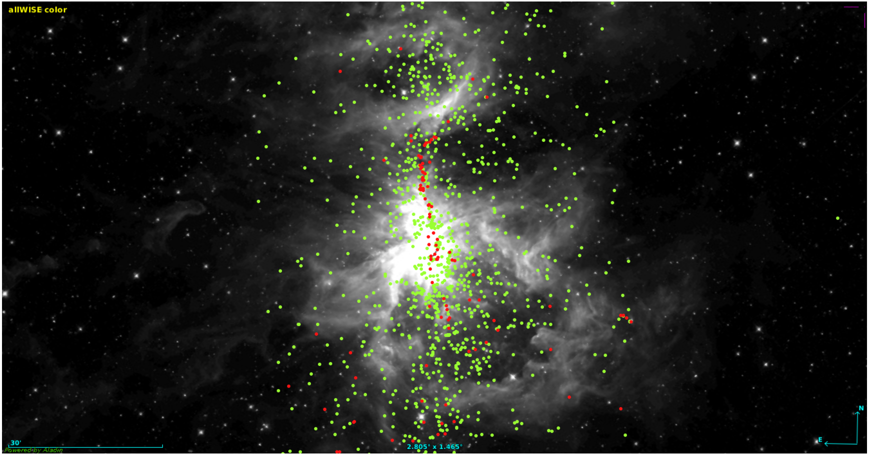
Classification comparison: Orion



Class II in green, and Class I in red
- Gutermuth classification

Wise 3.6 μm background image

Classification comparison: Orion



Class II in green, and Class I in red
- Learned with MLP

Wise 3.6 μm background image

Conclusion / Take home messages

- Usually ML methods they need large dataset to learn from
- Precautions must be taken in imbalanced cases (Obs. Proportions, Re-balance training set)
- ANN are able to balance some of the limitations of the usual YSO classification, providing efficient candidates catalogs.

On going work:

- Try more recent semi-supervised learning methods
- Use simulated YSOs as our training sample to avoid the G09 classification
- Extend to large survey (GLIMPSE) to provide wide candidates catalog

Other area and combined results

Adding the region NGC 2264 from Rapson+ 2014

Training: Orion; Forward: NGC 2264

Class	YSO CI	YSO CII	Other	Recall
YSO CI	74	2	14	82.2%
YSO CII	6	402	27	92.4%
Other	9	52	7203	99.2%
Precision	83.2%	88.2%	99.4%	98.6%

Training: NGC2264; Forward: Orion

Class	YSO CI	YSO CII	Other	Recall
YSO CI	285	33	6	88.0%
YSO CII	54	1967	203	88.4%
Other	98	293	16175	97.6%
Precision	65.2%	85.8%	98.7%	96.4%

Other area and combined results

Adding the region NGC 2264 from Rapson+ 2014

Training: Orion; Forward: NGC 2264

Class	YSO CI	YSO CII	Other	Recall
YSO CI	74	2	14	82.2%
YSO CII	6	402	27	92.4%
Other	9	52	7203	99.2%
Precision	83.2%	88.2%	99.4%	98.6%

Training: NGC2264; Forward: Orion

Class	YSO CI	YSO CII	Other	Recall
YSO CI	285	33	6	88.0%
YSO CII	54	1967	203	88.4%
Other	98	293	16175	97.6%
Precision	65.2%	85.8%	98.7%	96.4%

Confusion matrix for the Merged training set, forwarded on the corresponding test set.

Class	YSO CI	YSO CII	Other	Recall
YSO CI	77	2	3	93.9%
YSO CII	9	514	8	96.8%
Other	9	49	4706	98.8%
Precision	81.1%	91.0%	99.8%	98.5%

SOFTMAX output filter

Result on the full datasets:

Predicted

	Class	YSO CI	YSO CII	Other	Recall
Actual	YSO CI	391	13	10	94.4%
	YSO CII	37	2590	32	97.4%
	Other	46	210	23574	98.9%
	Precision	82.5%	92.1%	99.8%	98.7%

No filter, no object lost

SOFTMAX output filter

Result on the full datasets:

Predicted

	Class	YSO CI	YSO CII	Other	Recall
Actual	YSO CI	318	1	8	97.2%
	YSO CII	10	2443	14	99.0%
	Other	23	92	23383	99.5%
	Precision	90.6%	96.3%	99.9%	99.4%

0.9 filter, 611 lost (87 CI, 192 CII, 332 Other)