Machine Learning
(for ISM astronomers)

Christophe MORISSET

IA – UNAM Ensenada

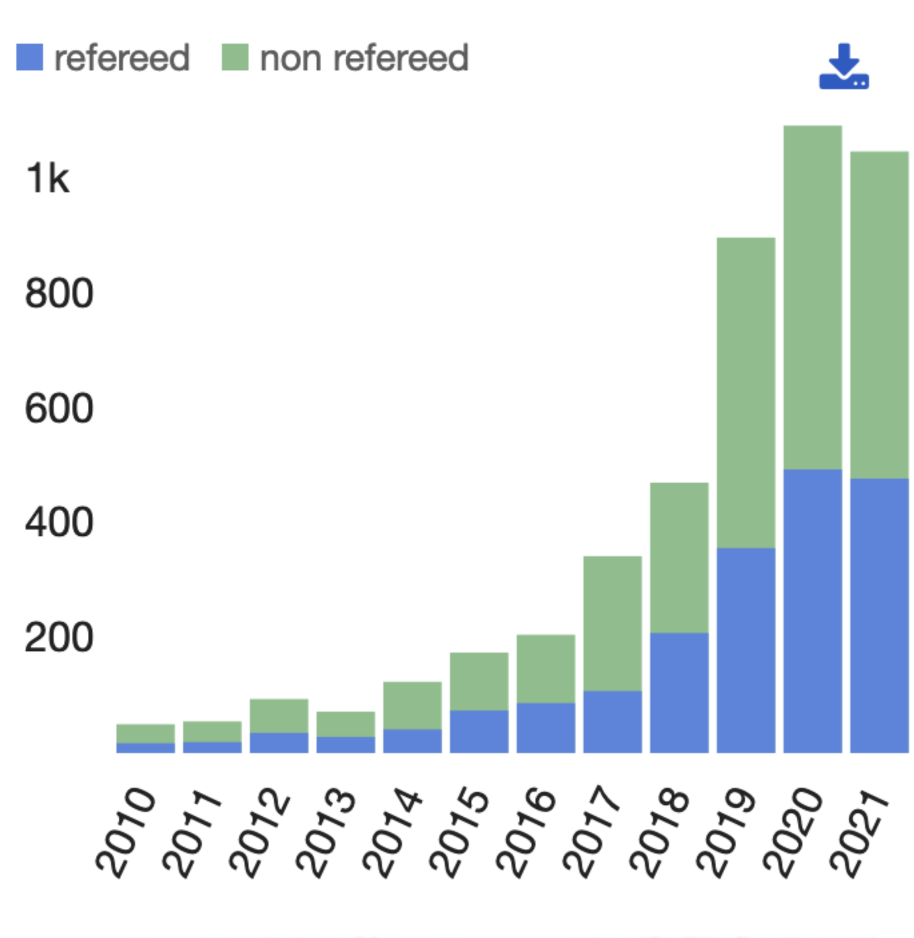
Sabbatical stage at IAP

Machine learning - Introduction

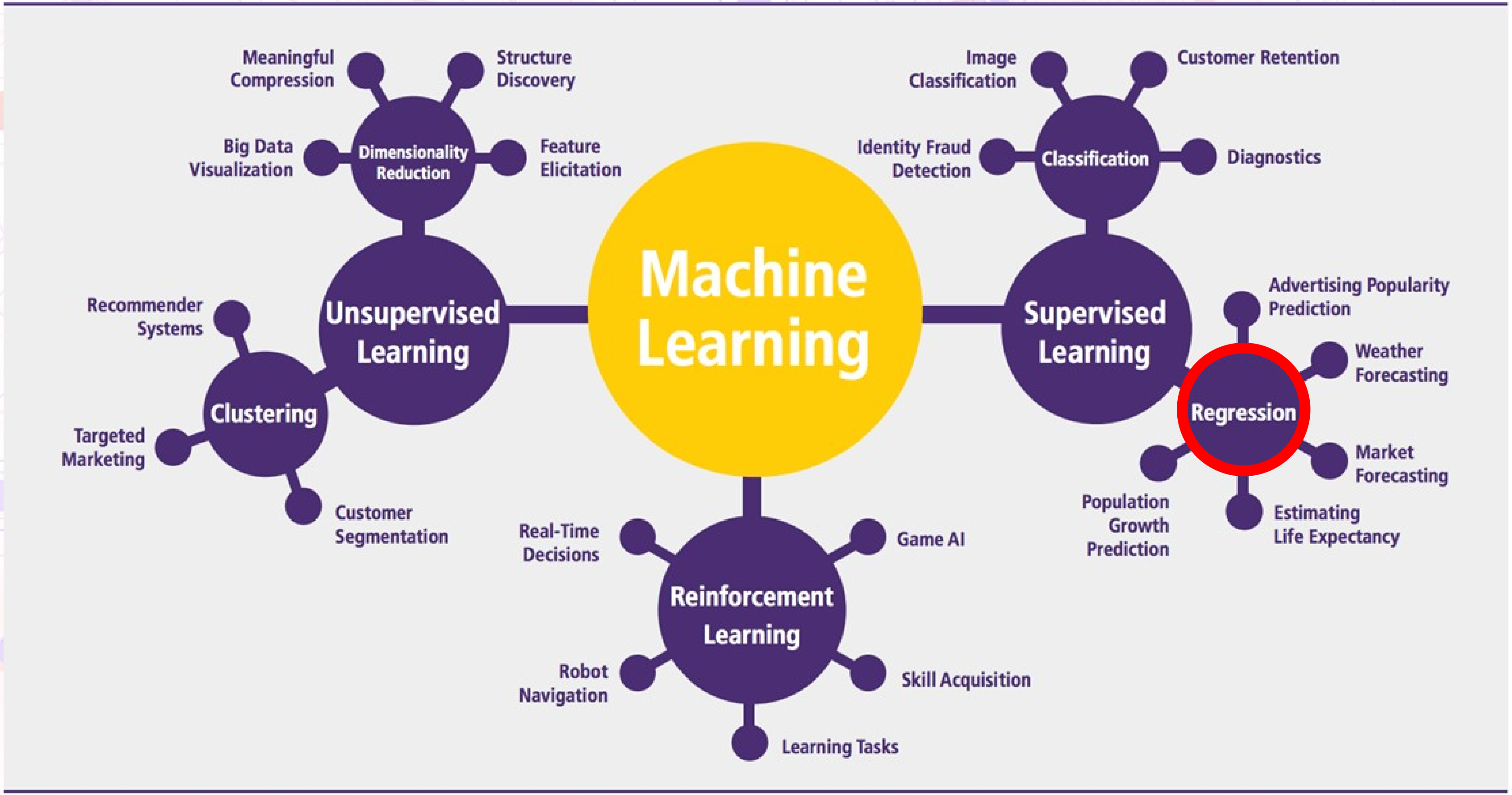
- Machine learning is a branch of **algorithmic** that manages models for a sample of data, based on a set of **examples**, to **predict** behavior of another set of data (training set and test set).
- It is part of “Artificial Intelligence”.
- Its performances recently increased due to improvements in hardware (GPU) and software (libraries).

ML use in astronomy

Astronomy papers in ADS containing "Artificial Intelligence" or "Machine learning" or "Deep learning" in the abstract.



Machine Learning : a whole world



ML used in astronomy

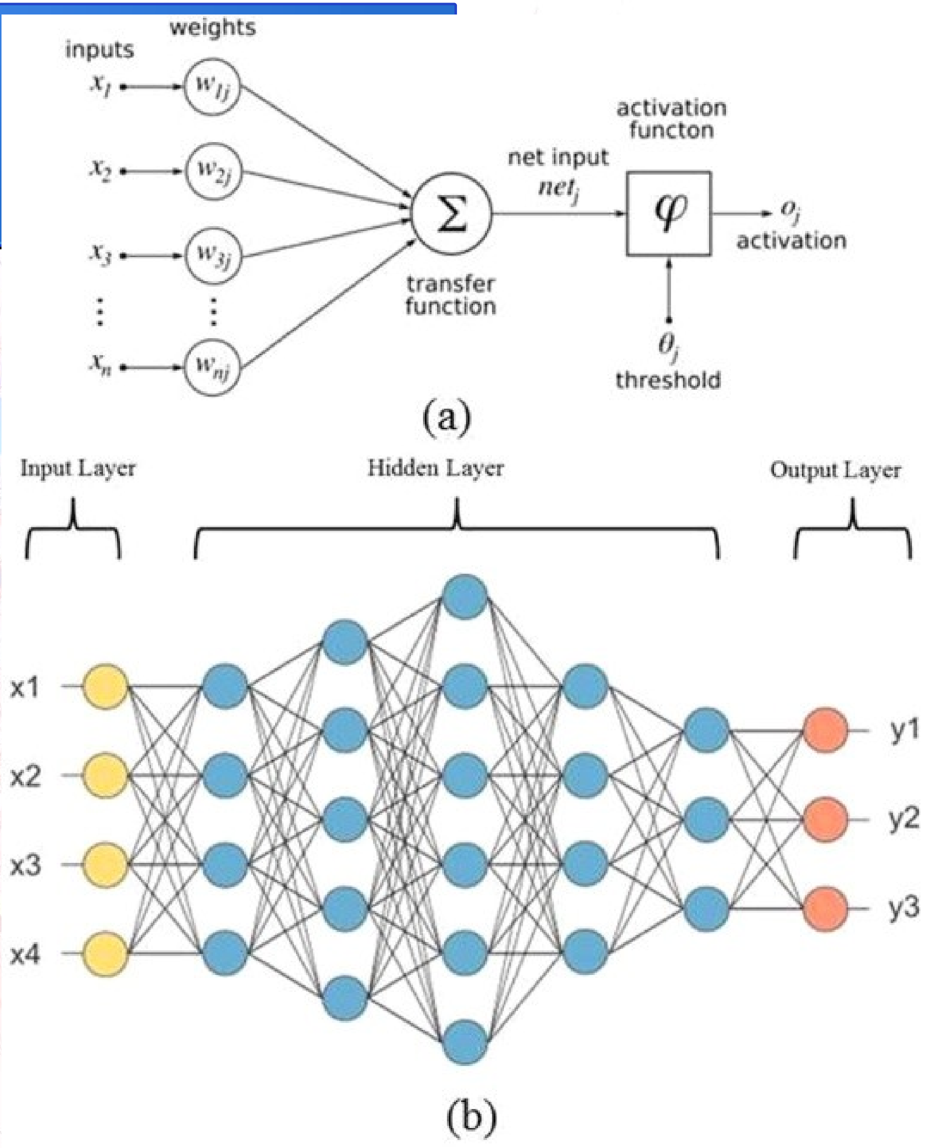
- Unsupervised Learning
  - Clustering: *Machine learning in APOGEE. Identification of stellar populations through chemical abundances*, Garcia-Dias+19
- Reinforcement Learning
  - *Deep reinforcement learning for smart calibration of radio telescopes* Yatawatta+21
- Supervised Learning
  - Classification: a lot e.g. *A diagnostic tool for the identification of supernova remnants* Kopsacheili+20
  - Regression: **THIS WORK**

Reviews:

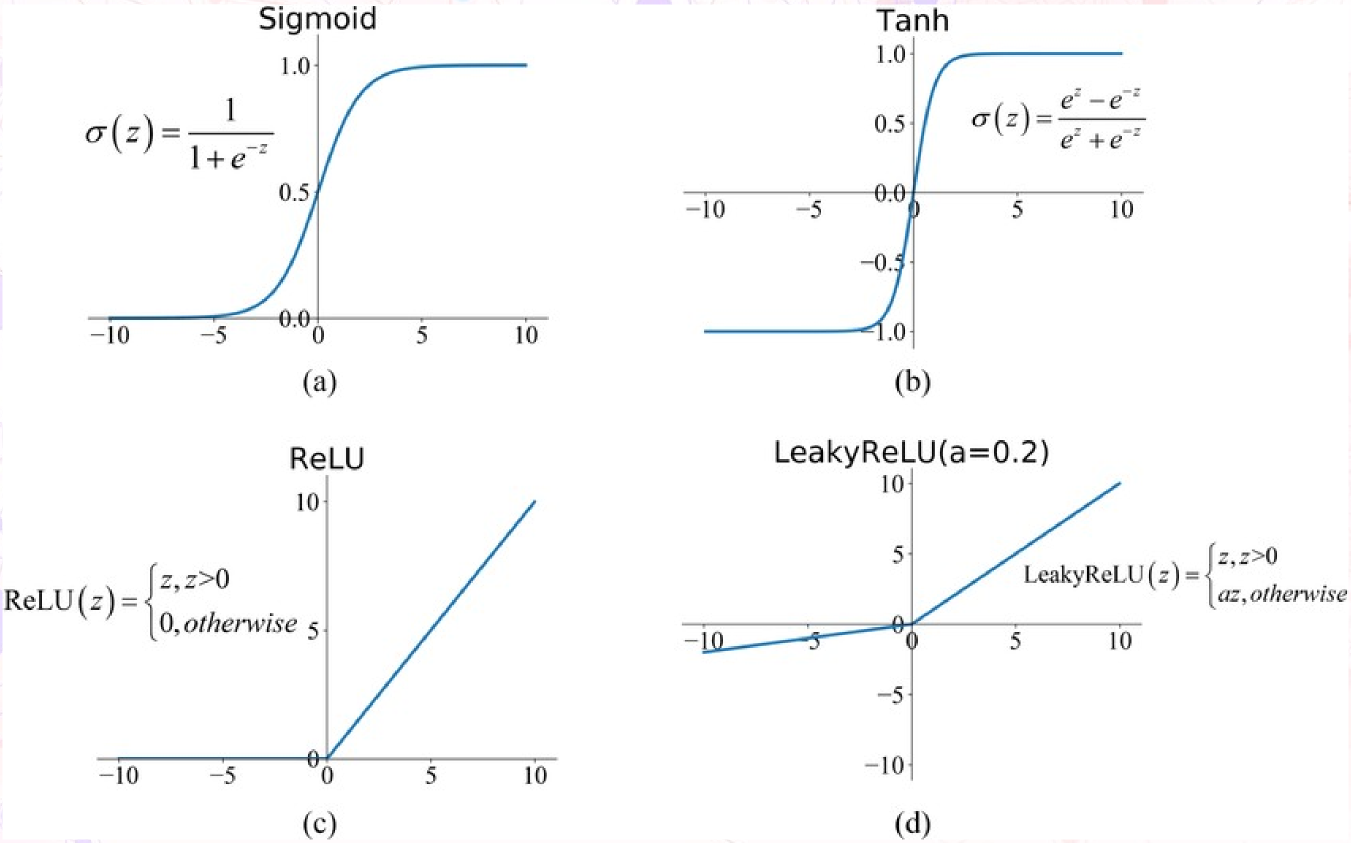
- *Surveying the reach and maturity of machine learning and artificial intelligence in astronomy,* Fluke & Jacobs 2020
- *Artificial Intelligence in Astrophysics,* book Zelinka+21

Artificial Neural Network

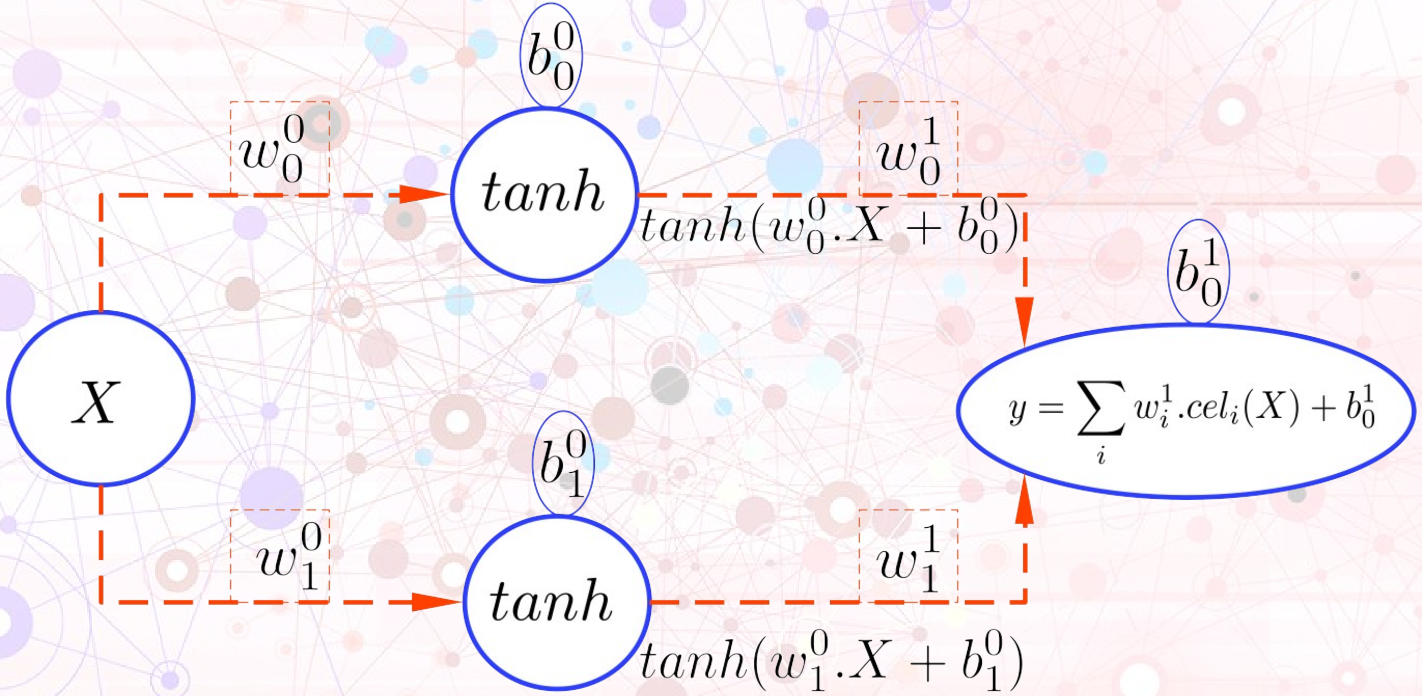
- Each neuron receives data (inputs) and produces a single output.
- The output is obtained by applying an activation function to the weighted sum of the inputs
- A constant term can also be added (bias).
- Neurons are grouped together by layers.



Activation functions



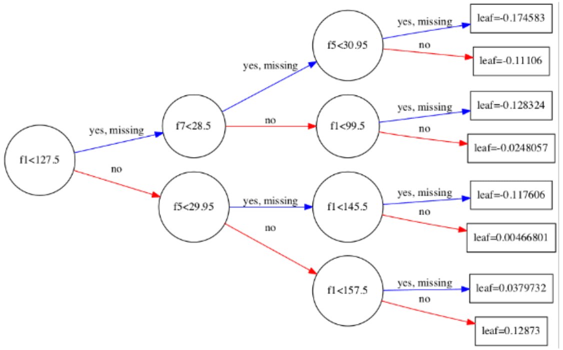
Simple example of a 2 neurons network

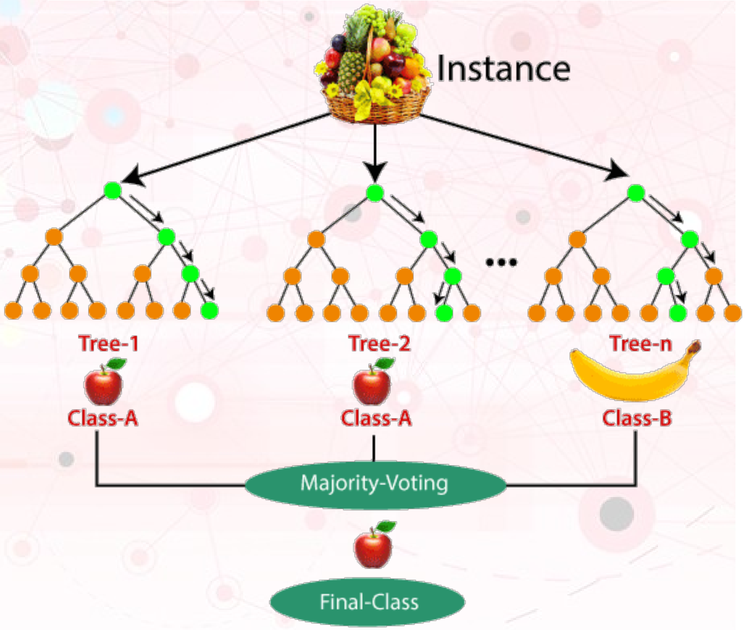


$$y(X) = w_0^1.tanh(w_0^0.X + b_0^0) + w_1^1.tanh(w_1^0.X + b_1^0) + b_0^1$$

Decision trees, random forest, gradient boosting

- Decision tree: sequential process, test-based, to determine a final value.
- Random forest: **majority of weak trees is strong!**
- Boosting is a method to increase strongness of weak trees.





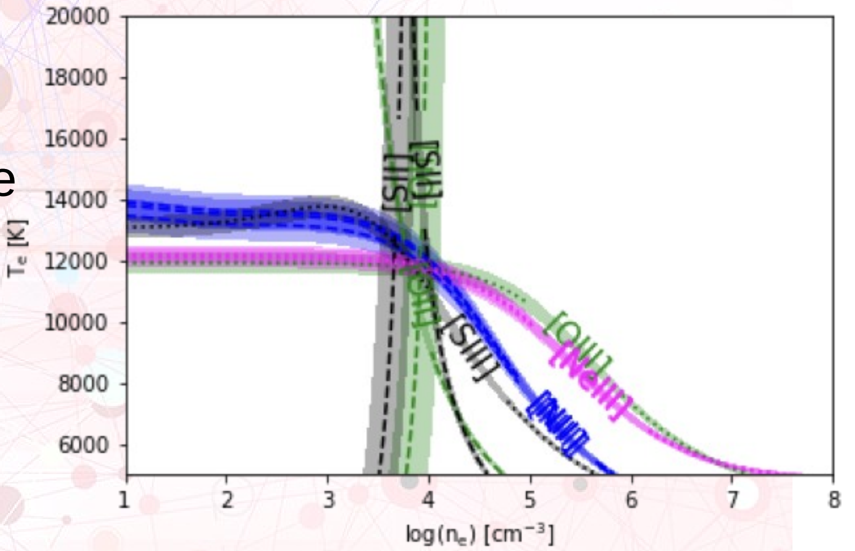
Summary:
My uses of ML in nebular studies

- Te-Ne : very fast determination
- ICFs : ad-hoc values
  - From other ionic fractions (Muse data)
  - From emission line ratios (PC-22)
- Exploring multiple solutions in O/H determination
  - (Direct)
  - Evolution models.

PyNeb.Diagnostics.getCrossTemDen

PyNeb.Diagnostics.getCrossTemDen:

- Obtain Te and Ne from a pair of diagnostic line ratios e.g. [OIII] 4363/5007 & [SII] 6716/6731.
- Starts to be slow when dealing with IFUs+MC data sets.
- SOLUTION:
  - Generate Diag1 & Diag2 from a grid of Te & Ne.
  - Train a **scikit-learn ANN** (10 secs, may be saved for future use) to predict reverse problem: gives Te & Ne from Diag1 & Diag2.
  - Use the ANN: **from 5 hours to 2 seconds**!



Need for fast solutions

- In case of MUSE observations: 200x200 spaxels.
- Monte-Carlo method to follow uncertainties through the whole pipeline (Reddening correction, Te-Ne, Xi/H+, X/H).
- → 200x200x150 = 6,000,000 “spectra” per object!
- Also used for T(Paschen Jump).
- Garcia-Rojas+21, subm:

Te maps for 3 PNe, under different hypothesis

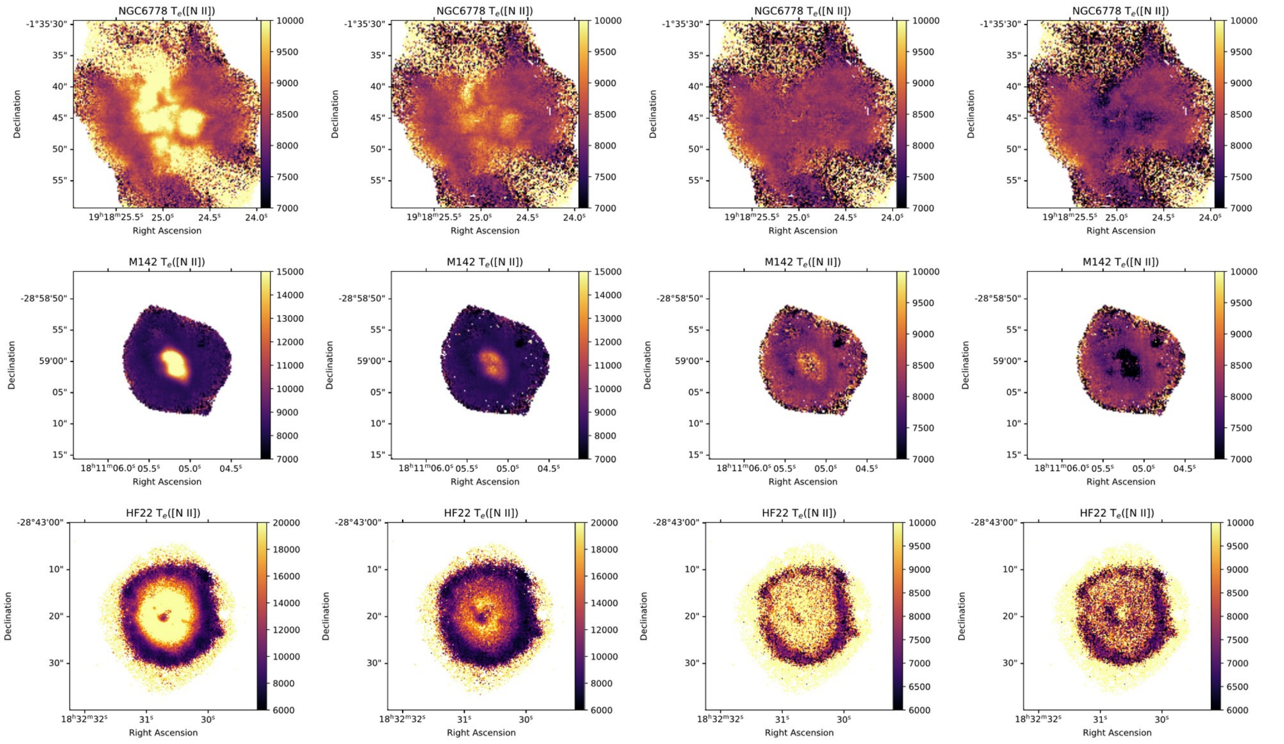


Figure 7. Variation of T_e([N II]) maps considering no recombination contribution correction (first column panels), and recombination contribution corrections assuming different temperatures for the recombination zone emission: T_e=1,000 K, 4,000 K, 8,000 K (second, third and last column panels, respectively, for our three PNe. The temperature scale is the same for the 4 cases in NGC 6778. In M 1-42 and Hf 2-2 we used wider T_e scales for the "no correction" and "1,000 K" cases given the large range in T_e observed in these cases.

- Easy to "play" with the data and to test the effect of recombination contribution correction at different temperature.
- Apparent **warm** gas in the central part is only due to not correctly taking into account this contribution, actually coming from a **cold** region (!).
- Stay tuned: Garcia-Rojas et al., submitted.

ICFs

To determine chemical abundances, one needs to take into account the presence of unseen ions, e.g.:

$$\frac{N}{H} = \frac{N^{+} + N^{++}}{H^{+}} = ICF.\frac{N^{+}}{H^{+}}$$

$$\frac{N}{O} = \frac{N^{+} + N^{++}}{O^{+} + O^{++}} = ICF(N/O).\frac{N+}{O+}$$

These ICF are determined using photoionization models (obtained for example running Cloudy).

Photoionization models

INPUTS:

- Ionizing SED:
  - Teff, log g, Z, Intensity
- Gas:
  - n_H(r), inner cavity
  - O/H, N/H, …
  - Dust

OUTPUTS:

ICFs

- Te
- H^+/H, N^+/N, O^+/O, O^{2+}/O, O^{3+}/O, ...
- Hβ, [NII] 6584, [OII] 3727, [OIII] 5007, …
- ...

CODE

3MdB

- Machine Learning techniques like to have A LOT of data to train with, to increase performances in prediction.
- 3MdB is a database of photoionization models, obtained with Cloudy (Ferland et al.), for PNe and HII regions.
- More than 2 million models, still growing.

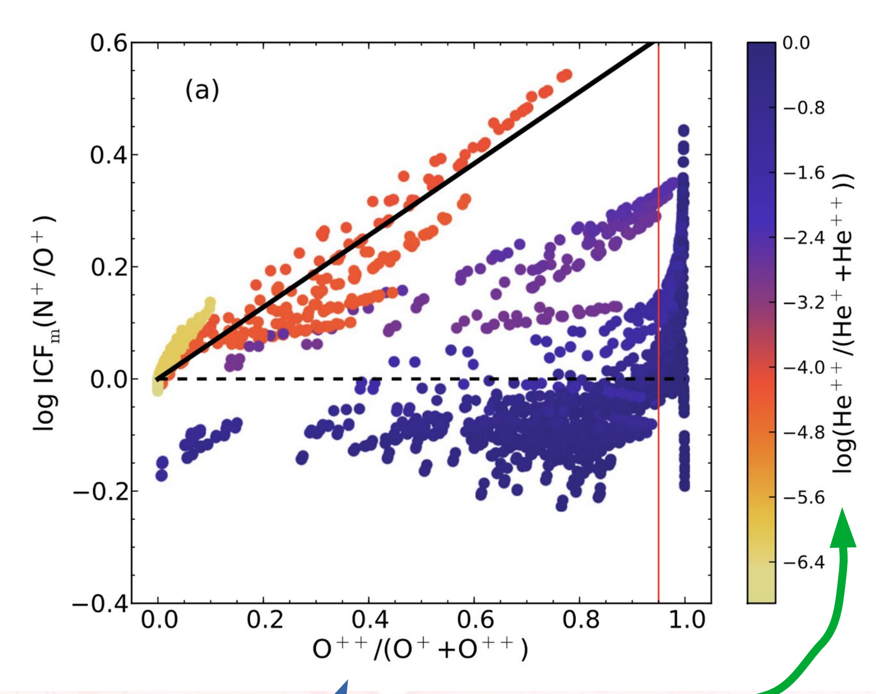
ICFs from Delgado-Inglada et al. 2015

- ICF(N+/O+) is commonly assumed to be 1.0:

$$\frac{N}{O} = \frac{N^+}{O^+}$$

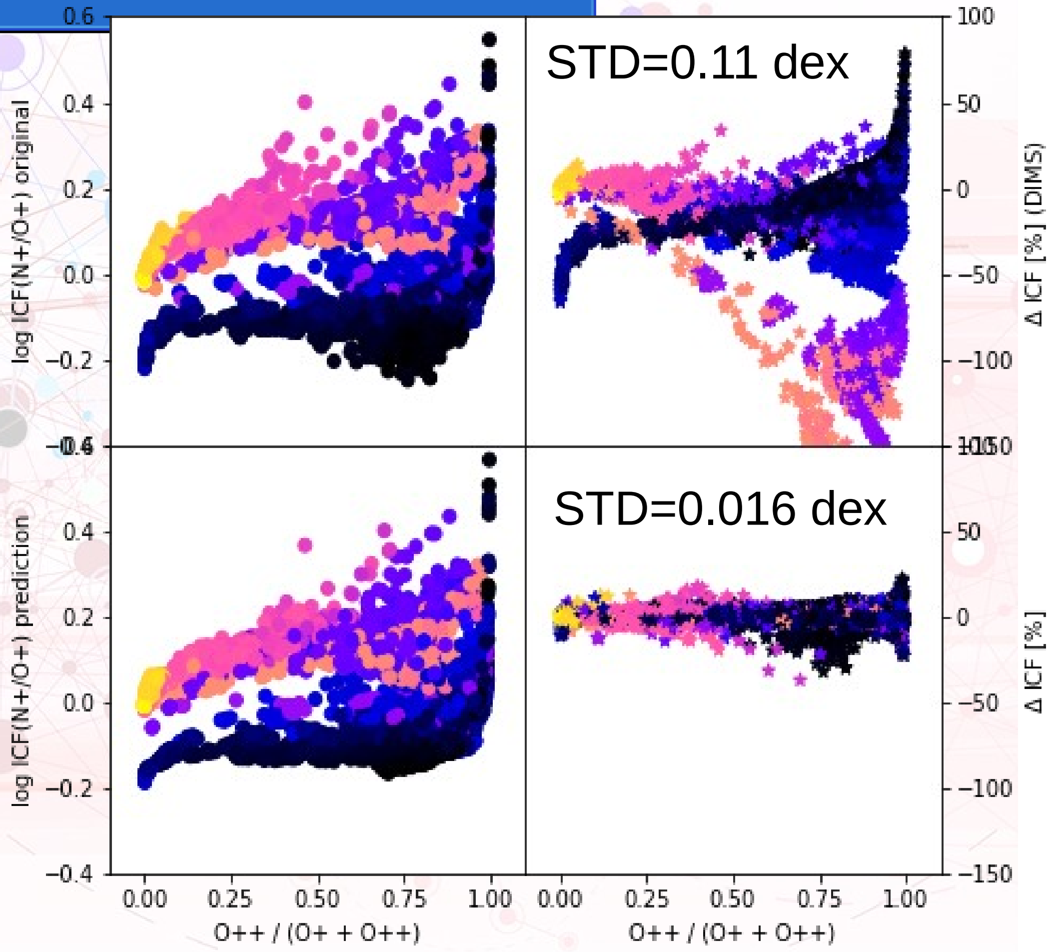
- Using grids of photoionization models, more complex ICFs can be determined (DIMS15):

log ICF(N+/O+)=−0.16ω(1+logυ).



New ICFs: example of N/O

- A neural network is trained with 35,000 models from 3MdB (50x50 neurons).
- **O^{++}/O, He^{++}/He and S^{++}/S^{+}** are used as inputs. Very hard to define an algebraic fit in a 3D space.
- ICFs obtained with ANN are closer to the expected values as determined from models.



Ad-hoc ICFs, for given object

- In the study of 3 PN observed by MUSE, we compute ICFs **adapted** to each PN to derive the elemental abundances for the collapsed spectra.
- We select models from 3MdB "close" to the given PN.
- We train an **XGBoost** Machine.
- Garcia-Rojas et al., submitted.

- Inputs for the ML
  - He^{2+}/He^{+}
  - O^{2+}/O^{+}
  - S^{2+}/S^{+}
  - Cl^{3+}/Cl^{2+}
  - Ar^{3+}/Ar^{2+}

- Predictions:
  - C/C^{+}
  - N/N^{+}
  - $(O^{+}/O).(N/N^{+})$
  - $O/(O^{+} + O^{2+})$
  - $S/(S^{+} + S^{2+})$
  - $Cl/(Cl^{2+} + Cl^{3+})$
  - $Ar/(Ar^{2+} + Ar^{3+})$

Feature importances

ICFs	Observed ionic fractions: He^{2+}/He^{+}	O^{2+}/O^{+}	S^{2+}/S^{+}	Cl^{3+}/Cl^{2+}	Ar^{3+}/Ar^{2+}
N^{+}	0.00	0.96	0.04	0.00	0.00
N^{+}/O^{+}	0.04	0.02	0.71	0.23	0.01
$O^{+} + O^{++}$	0.45	0.01	0.38	0.13	0.03
$S^{+} + S^{++}$	0.00	0.00	0.00	0.00	1.00
$Cl^{2+} + Cl^{3+}$	0.00	0.05	0.94	0.00	0.00
$Ar^{2+} + Ar^{3+}$	0.09	0.02	0.80	0.07	0.03

The importance of each ionic fraction is not the same for each ICF.

These values slightly change from one object to another.

ICFs using ML techniques

- In the case of the PN PC22, we determine 11 ICFs **from 6 line ratios,** using a ML method based on XGBoost.
- A Te-sensitive line ratio have been added to connect emissivities and abundancias.
- Sabin et al. submitted.

The input vector X is build from a 6D vector of the logarithmic values of the following line ratios:

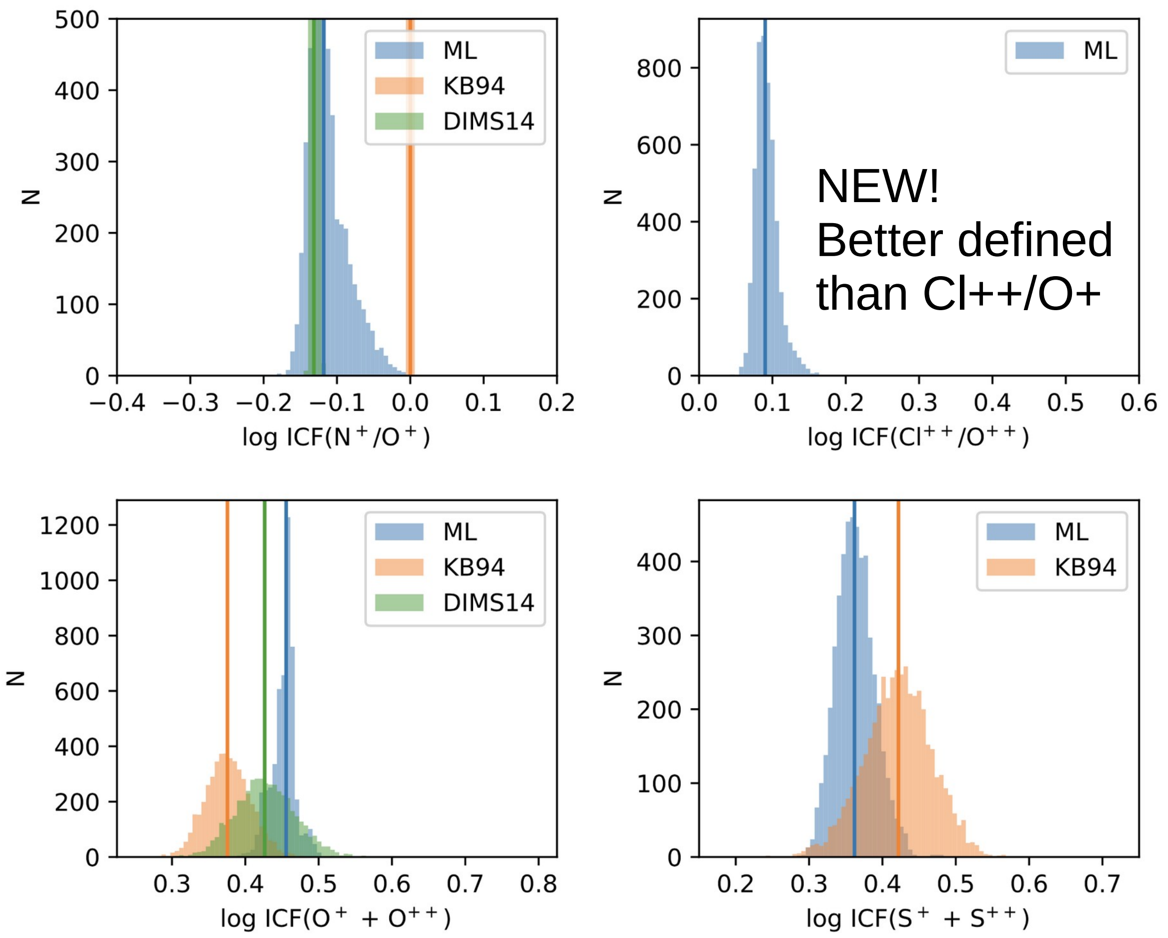
- He II $\lambda4686$ / He I $\lambda5876$
- [O III] $\lambda5007$ / [O II] $\lambda3727$
- [Ne V] $\lambda\lambda3426, 3346$ / [Ne IV] $\lambda4726$
- [Ne IV] $\lambda4726$ / [Ne III] $\lambda3869$
- [Ar V] $\lambda6435$ / [Ar IV] $\lambda\lambda4711, 4740$
- [O III] $\lambda\lambda4363/5007$

The output vector y is directly the set of the following ICFs (logarithmic values are used):

- $O / (O^{+} + O^{++})$
- $N/O \times O^{+} / N^{+}$
- $Ne / (Ne^{++} + Ne^{4+})$
- $Ne / (Ne^{++} + Ne^{3+} + Ne^{4+})$
- $Ne / O \times O^{++} / Ne^{++}$
- $S / (S^{+} + S^{++})$
- $S / O \times O^{+} / (S^{+} + S^{++})$
- $S / O \times O^{++} / (S^{+} + S^{++})$
- $Cl / O \times O^{+} / Cl^{++}$
- $Cl / O \times O^{++} / Cl^{++}$
- $Ar / (Ar^{3+} + Ar^{4+})$

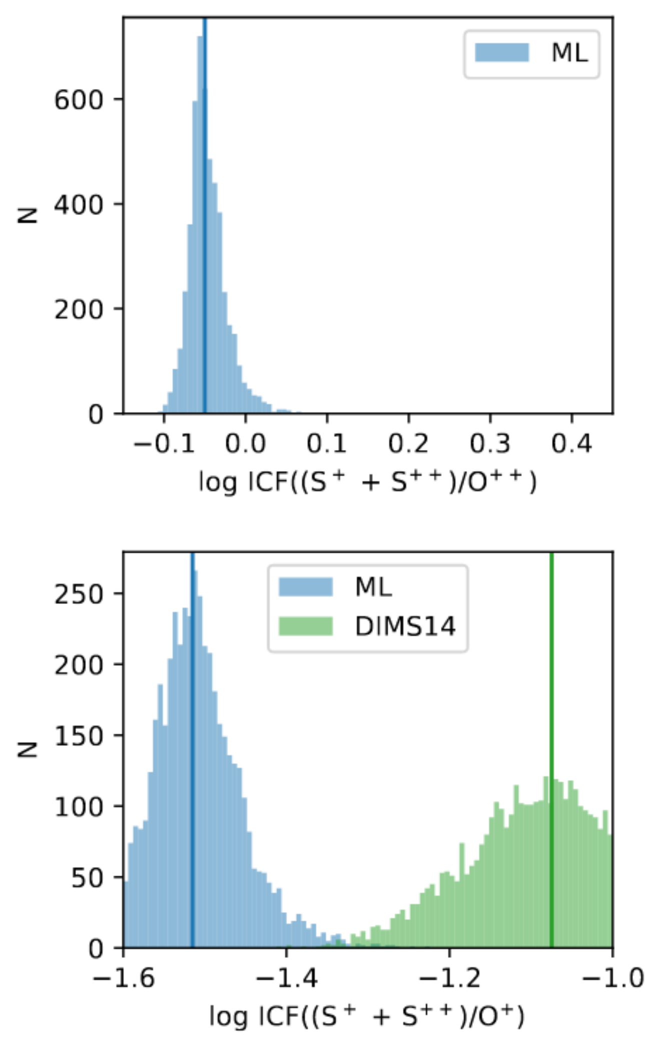
ICFs

- The ICFs we obtained can be compared to the classical ones from the literature.
- New ICFs have been obtained.
- Sabin et al. submitted.



Sulfur ICF

- We obtain a new ICF related to O^{++}.
- It is more reliable than based on residual ion O^{+}.
- Sabin et al. submitted.



Feature importance

<15%
>15%

Observed line ratios

ICFs

	[OIII]/[OII]	[NeV]/[NeIV]	[NeIV]/[NeIII]	[ArV]/[ArIV]	HeII/HeI	[OIII]5007/4363
O+ + O++	0.00	0.01	0.05	0.17	**0.75**	0.01
N+/O+	0.16	0.03	0.03	0.14	0.30	0.34
Ne2+ + Ne4+	0.06	0.09	0.01	0.03	**0.75**	0.07
Ne2+ + Ne3+ + Ne4+	0.08	0.05	0.02	0.60	0.02	0.22
Ne2+/O2+	0.10	0.09	0.03	0.06	0.49	0.23
S+ + S++/O++	0.18	0.06	0.03	0.10	0.17	0.46
Cl2+/O2+	0.24	0.05	0.03	0.09	0.14	0.46
S+ + S2+	0.12	0.10	0.02	0.18	0.39	0.18
Ar3+ + Ar4+	0.17	0.03	0.02	0.11	0.52	0.16
S+ + S2+/O+	0.08	0.01	0.00	0.12	**0.69**	0.09
Cl2+/O+	0.10	0.01	0.00	0.12	**0.66**	0.10

HeII/HeI and [OIII] 5007/4363 are the most helpful, but other line ratios also matter.
Sabin et al. submitted.

O/H from strong lines

- Ho 2019 already used a ML technique to determine O/H from strong lines.
- N/O(O/H) and U(O/H) relations have effect on the strong line method calibrators when models are used.
- We use the e-BOND models stored in 3MdB.
- We train an ANN regressor to **mimic the behavior of Cloudy, but very faster**
- We can now change the N/O(O/H) and U(O/H) relations and see the effects on the calibrations.

Photoionization models

INPUTS:

- Ionizing SED:
  - Teff, log g, Z, Intensity
- Gas:
  - n_H(r), inner cavity
  - O/H, N/H, …
  - Dust

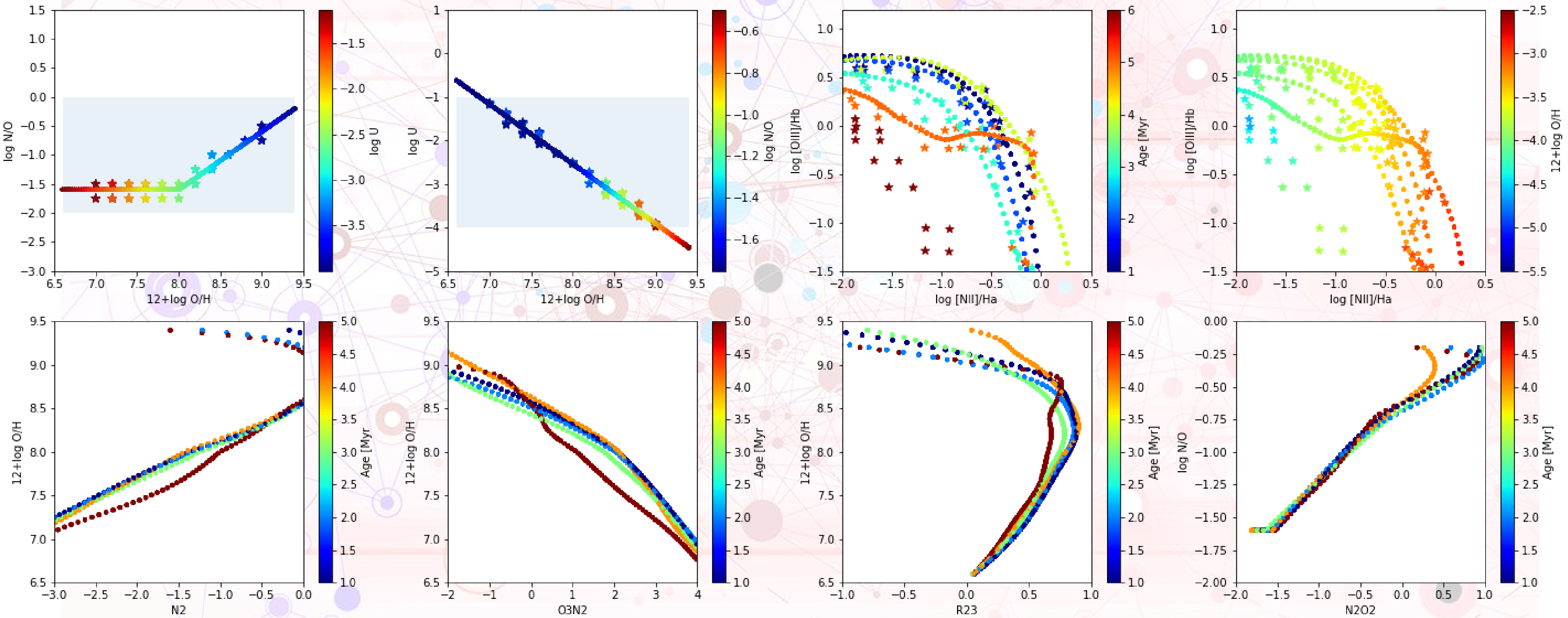
OUTPUTS:

- Te
- H^+/H, N^+/N, O^+/O, O^{2+}/O, O^{3+}/O, ...
- Hβ, [NII] 6584, [OII] 3727, [OIII] 5007, …
- ...

Cloudy

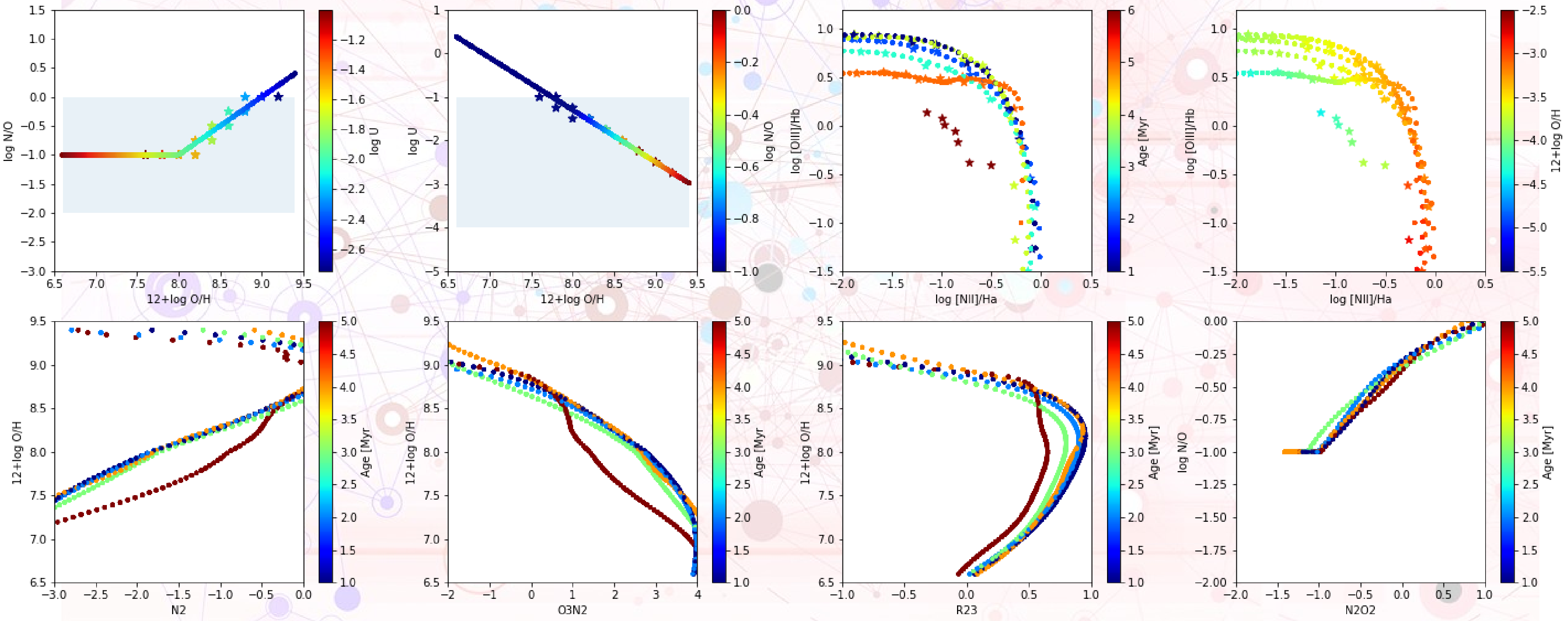
ML

Changing N/O y log U



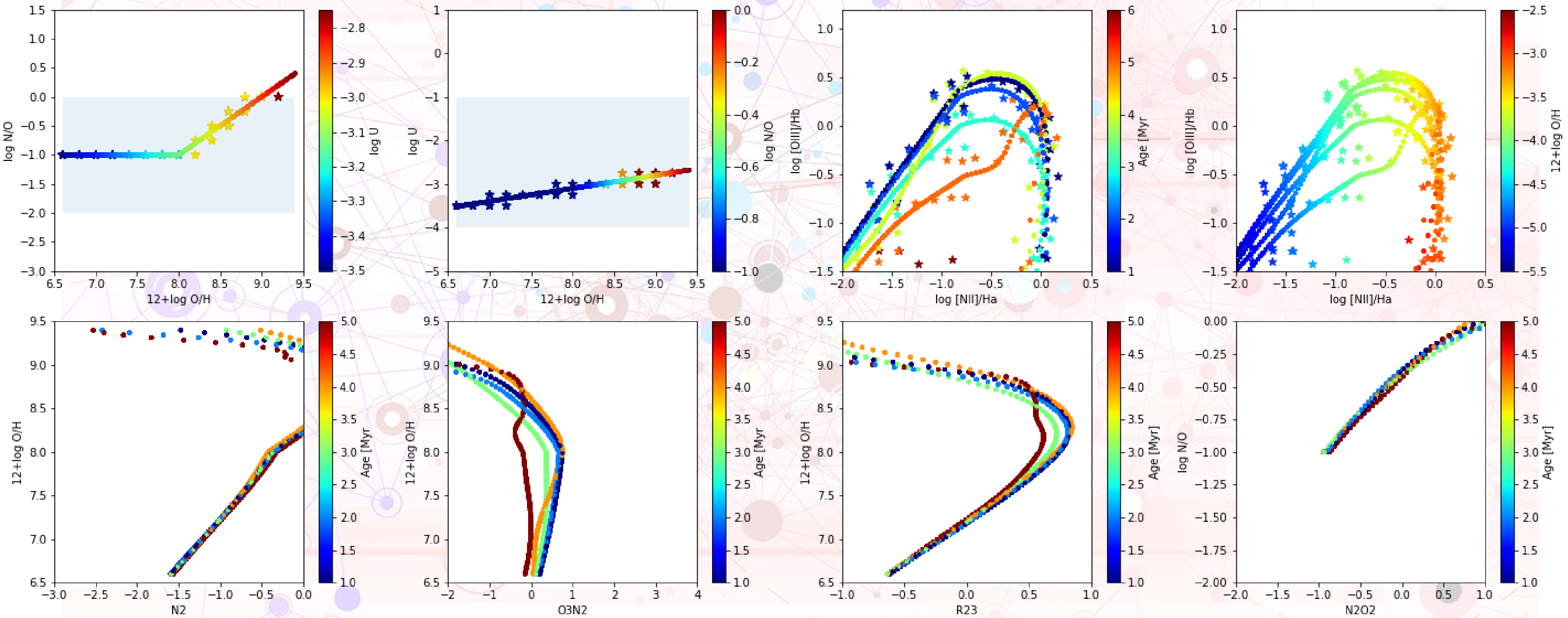
Espino-Ponce et al. In prep.

Changing N/O y log U



Espino-Ponce et al. In prep.

Changing N/O y log U



Espino-Ponce et al. In prep.

Photoionization models

INPUTS:

- Ionizing SED:
  - Teff, log g, Z, Intensity
- Gas:
  - n_H(r), inner cavity
  - O/H, N/H, …
  - Dust

OUTPUTS:

- Te
- H^+/H, N^+/N, O^+/O, O^{2+}/O, O^{3+}/O, ...
- Hβ, [NII] 6584, [OII] 3727, [OIII] 5007, …
- ...

Evolution Algo

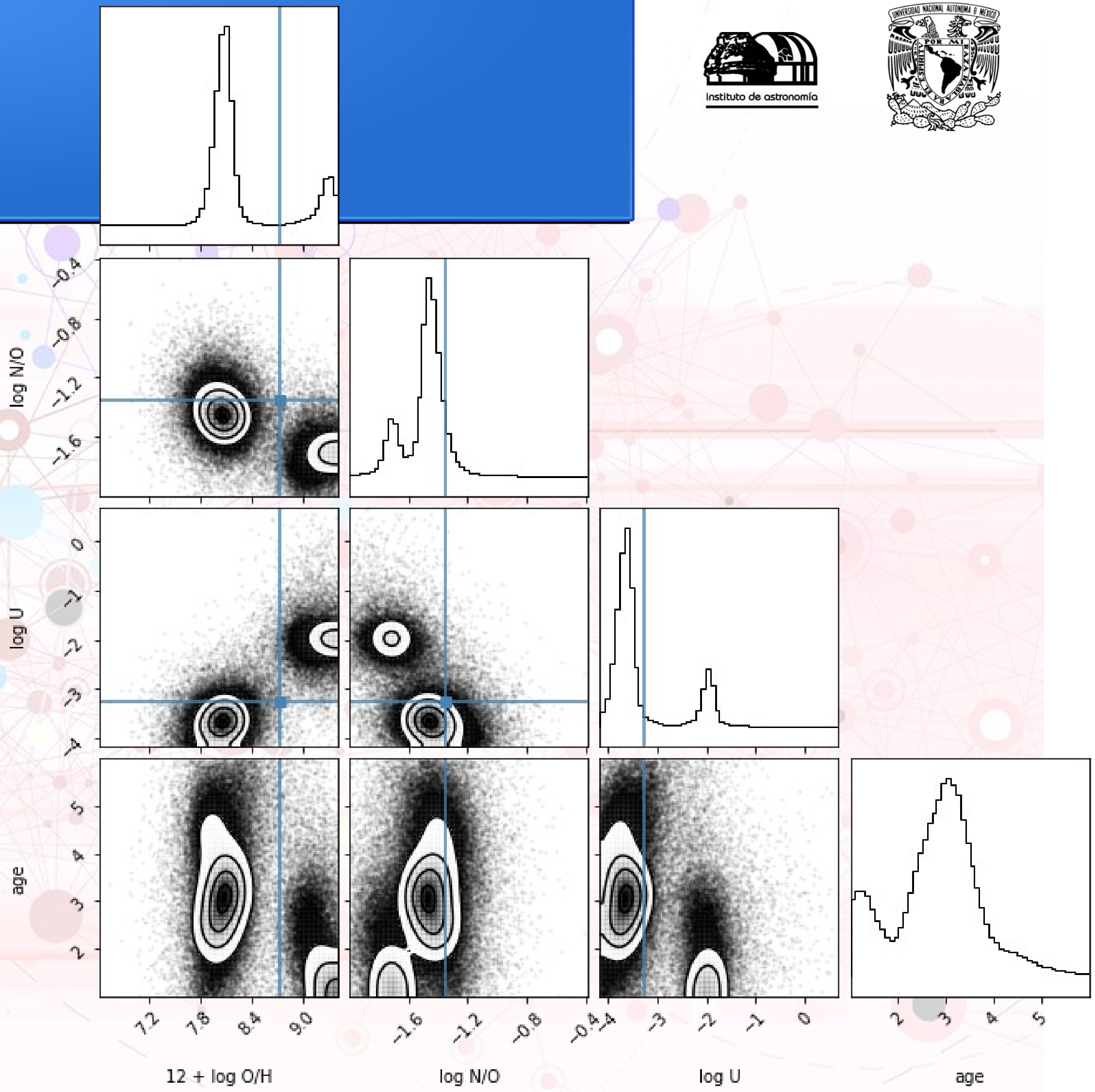
Cloudy

ML

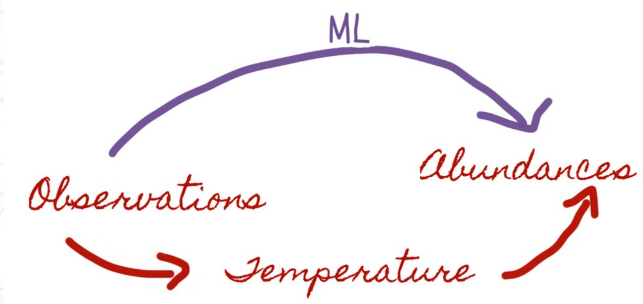
Looking for all the solutions

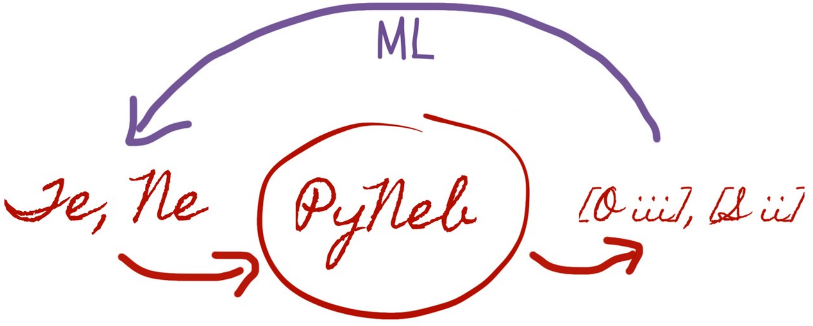
Strong lines

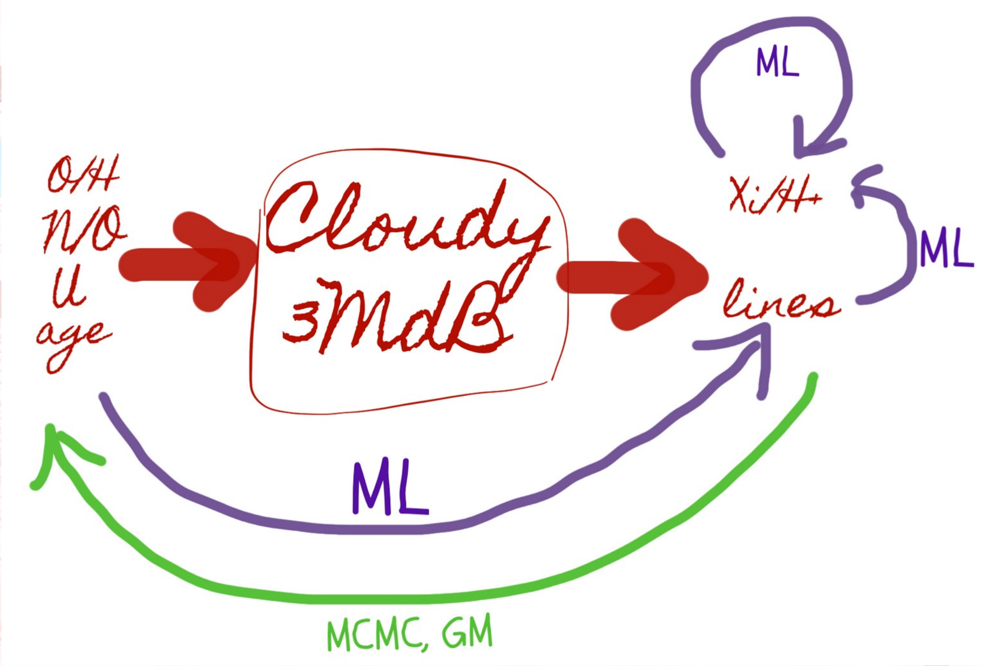
- Perez-Diaz+21 use [NII], [OIII] and [SII] to determine O/H running HII_CHI_m (Perez-Montero14)
- ANN is trained using e-BOND models from 3MdB to predict those lines, giving O/H, N/O, logU, and age.
- A **Genetic Evolution** model uses this ANN to look for the sets of parameters simultaneously fitting the observations of IC 2574. 370,000 calls to ANN.
- All the points in the contours correspond to values of parameters leading to reasonable fit to the observed data → degeneracy of O/H.
- **The “Best Model” is a meaningless concept.**
- **The “weighted mean value” is rather risky.**
- **Morisset et al. In Prep.**



Uses of ML







AI4Neb python library

Christophe Morisset
IA-UNAM

Thanks a lot!