Mining for Novel Information in Large and Complex Datasets

Dalya Baron, Tel Aviv University

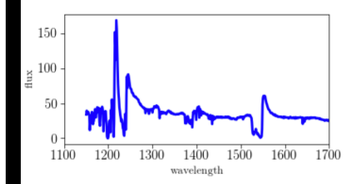
Artificial Intelligence in Astronomy workshop

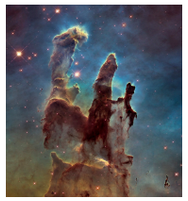
ESO, 2019

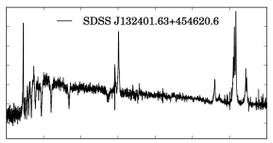
Data Landscape

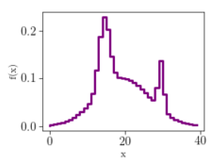
Description length

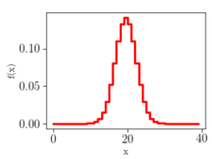
can't describe

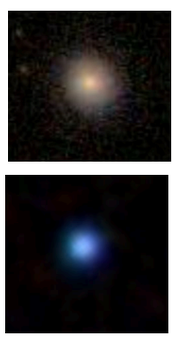


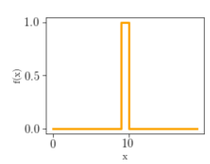


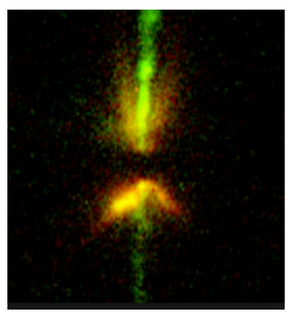


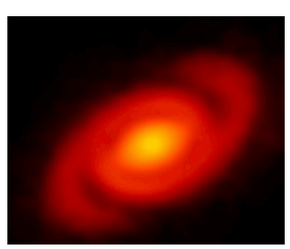






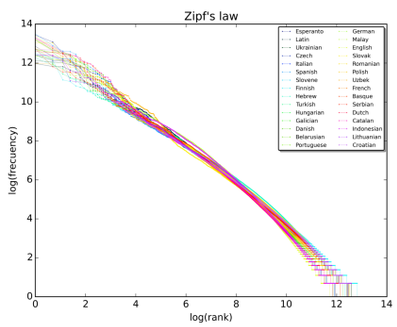




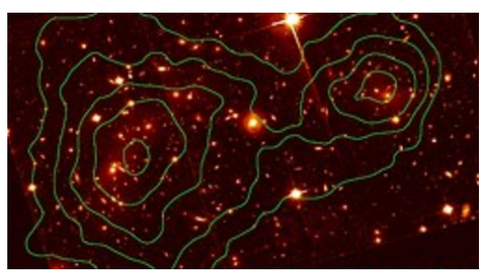


outliers

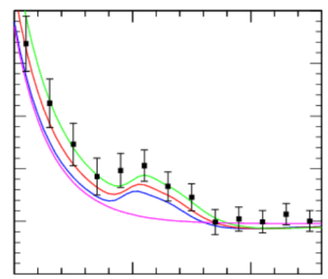
Word frequency in language



Weak lensing



BAO signal



can describe

High SNR

Low SNR

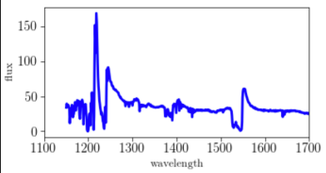
Signal to noise ratio

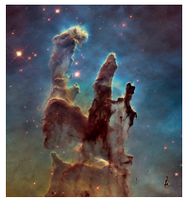
Data Landscape

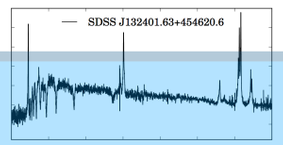
Description length

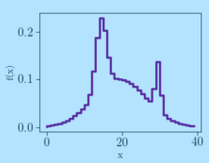
can't describe

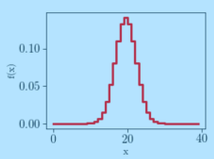
outliers

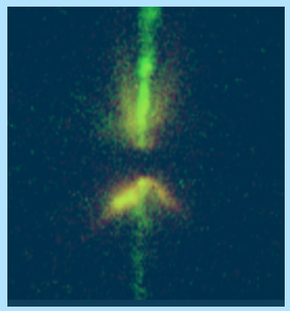


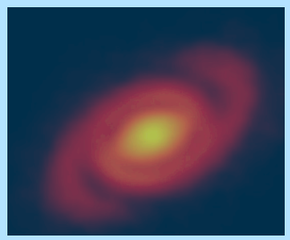




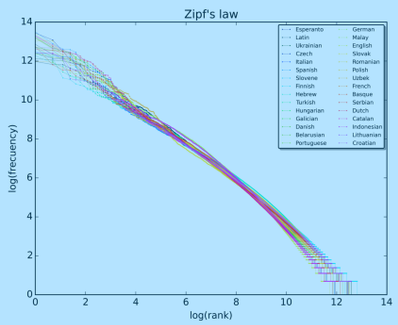




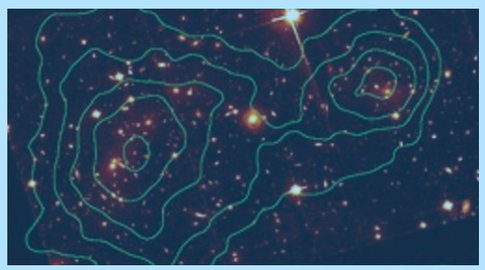




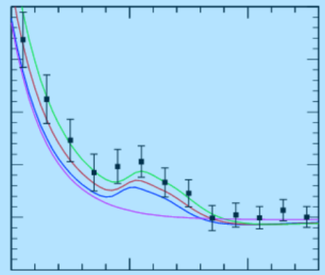
Word frequency in language



Weak lensing



BAO signal



Astronomy

can describe

High SNR

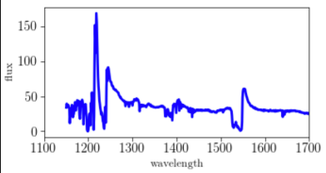
Low SNR

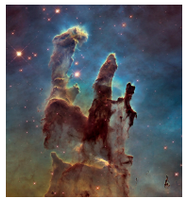
Signal to noise ratio

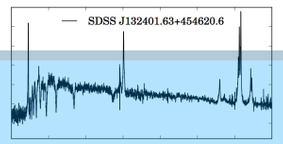
Data Landscape

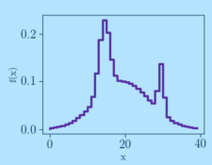
Description length

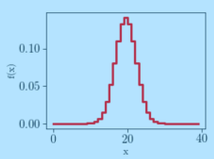
can't describe

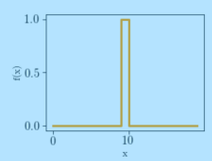


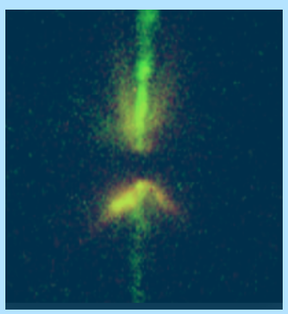


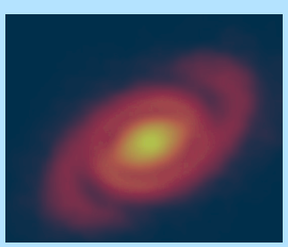












outliers

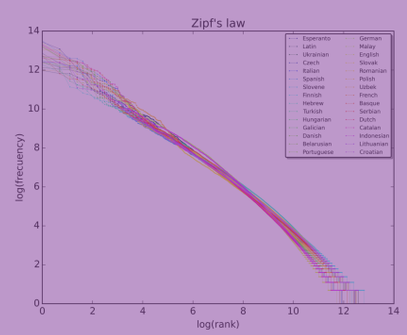
Astronomy

Statistical tools

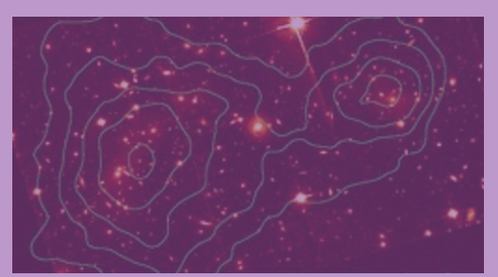
can describe

High SNR

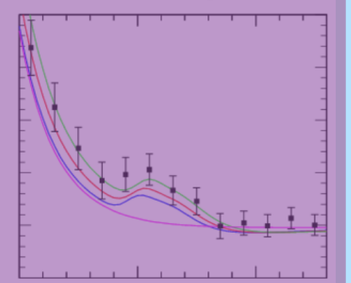
Word frequency in language



Weak lensing



BAO signal



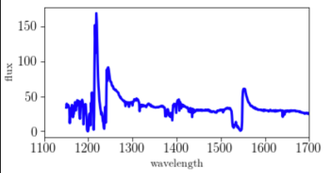
Low SNR

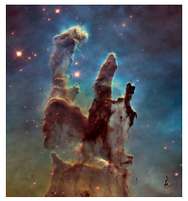
Signal to noise ratio

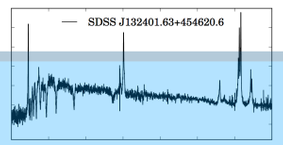
Data Landscape

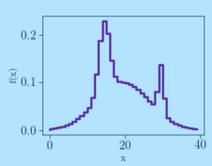
Description length

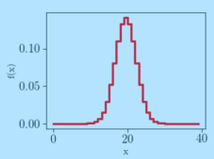
can't describe

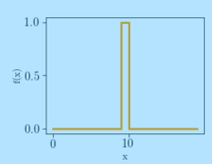


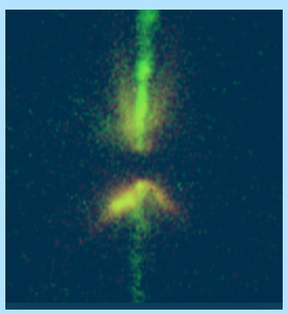


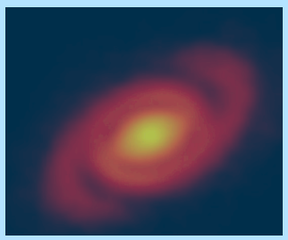












?

outliers

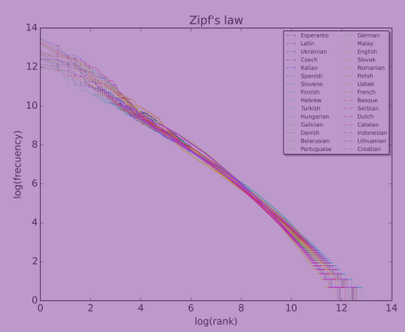
Astronomy

Statistical tools

can describe

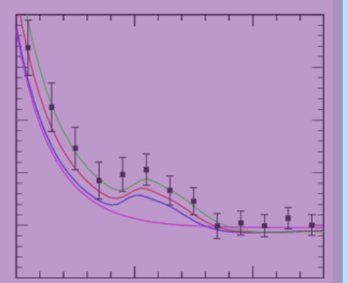
High SNR

Word frequency in language



Weak lensing

BAO signal



Low SNR

Signal to noise ratio

How can we extract novel information from large and complex datasets?

How can we extract novel information from large and complex datasets?

In this talk: using outlier detection and dimensionality reduction algorithms!

Machine Learning in Astronomy: a practical overview

Baron, Dalya

Astronomy is experiencing a rapid growth in data size and complexity. This change fosters the development of data-driven science as a useful companion to the common model-driven data analysis paradigm, where astronomers develop automatic tools to mine datasets and extract novel information from them. In recent years, machine learning algorithms have become increasingly popular among astronomers, and are now used for a wide variety of tasks. In light of these developments, and the promise and challenges associated with them, the IAC Winter School 2018 focused on big data in Astronomy, with a particular emphasis on machine learning and deep learning techniques. This document summarizes the topics of supervised and unsupervised learning algorithms presented during the school, and provides practical information on the application of such tools to astronomical datasets. In this document I cover basic topics in supervised machine learning, including selection and preprocessing of the input dataset, evaluation methods, and three popular supervised learning algorithms, Support Vector Machines, Random Forests, and shallow Artificial Neural Networks. My main focus is on unsupervised machine learning algorithms, that are used to perform cluster analysis, dimensionality reduction, visualization, and outlier detection. Unsupervised learning algorithms are of particular importance to scientific research, since they can be used to extract new knowledge from existing datasets, and can facilitate new discoveries.

Publication: eprint arXiv:1904.07248

Pub Date: April 2019

What are outliers?

"Bad" object: artifacts, cosmic rays, bad reduction.
Misclassified object: star classified as QSO, variable star classified as SN.
Tail of a distribution: most luminous SN, fastest accreting BH.
Unknown unknowns: completely new objects we did not know we should be looking for.

In astronomy: processes which happen on shorter time scales.

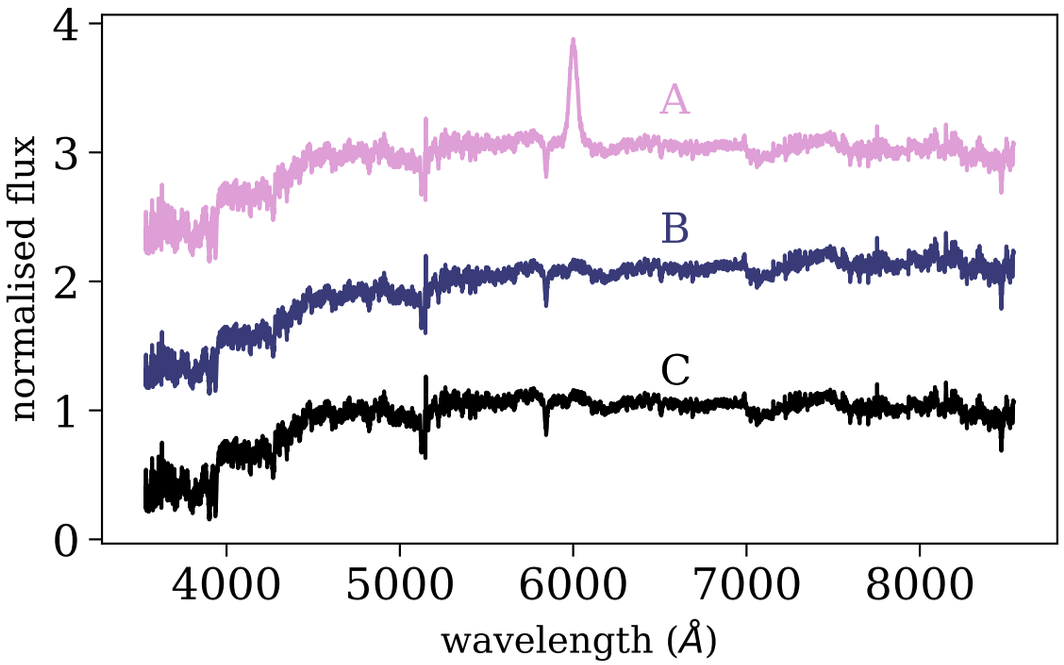
How can we find outliers?

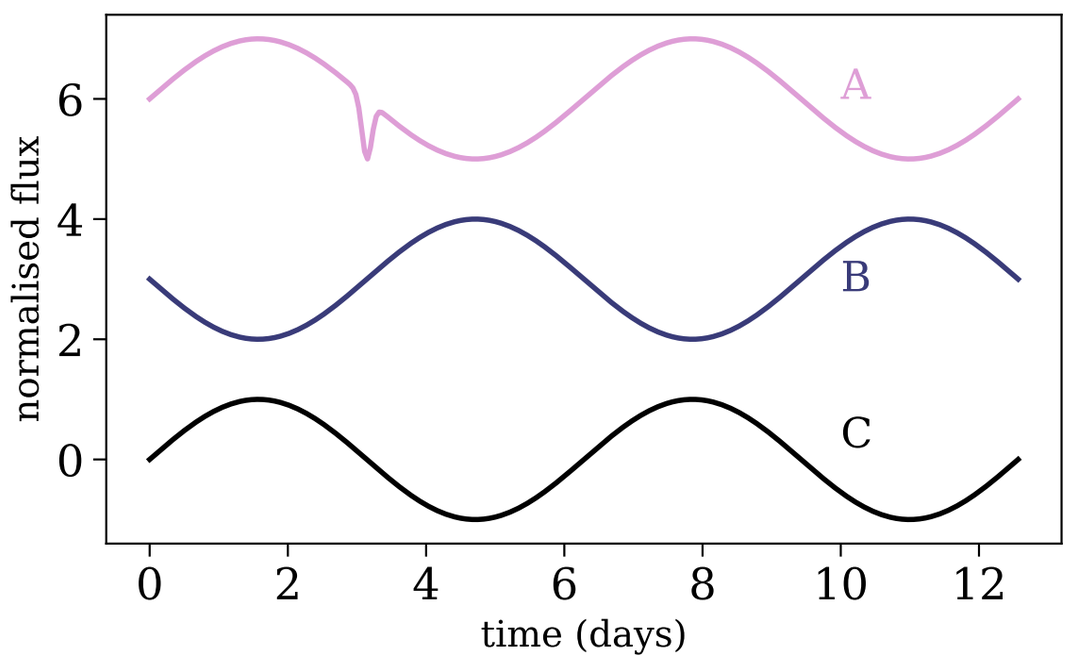
Serendipitously: an expert going through their data and finding unexpected objects. -> Usually not applicable for large datasets.

Using supervised learning: objects which have low probability to belong to any of the known classes will be considered outliers (or one-class SVM). -> Usually find the outliers that “shout the loudest”.

Using unsupervised learning: Isolation Forests, unsupervised distance assignment, and using dimensionality reduction algorithms. -> Strongly depend on the distance metric used.

see Baron & Poznanski (2017) and Baron (2019)

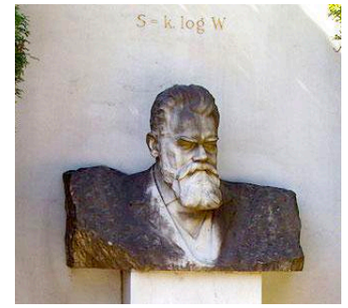




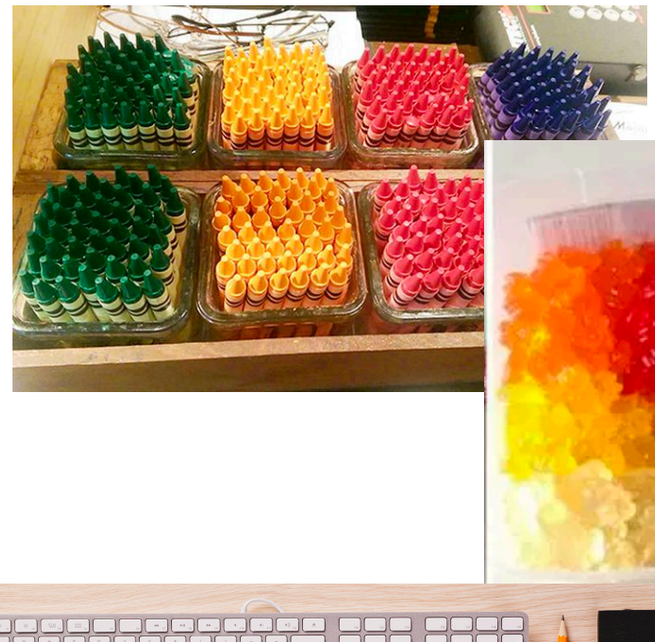
Humans

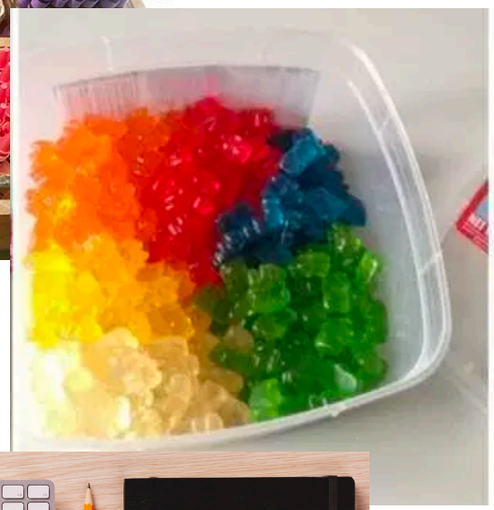
VS.

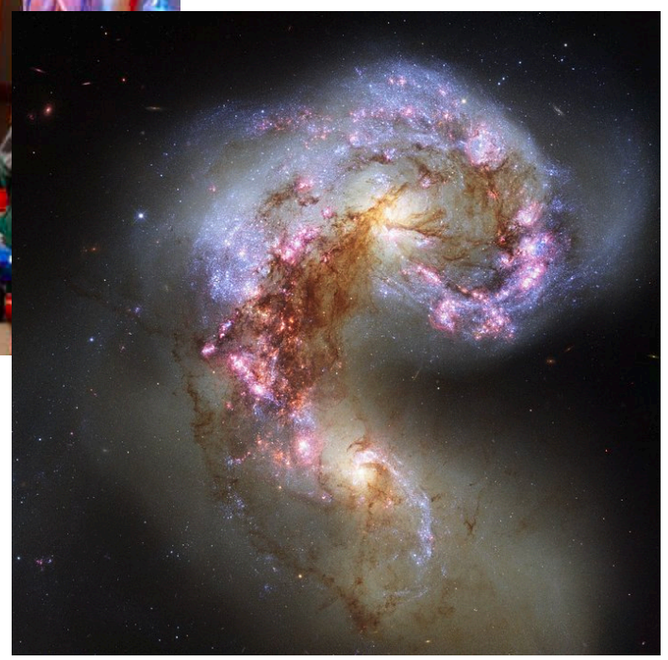
The second law of thermodynamics



$$S = k \log W$$





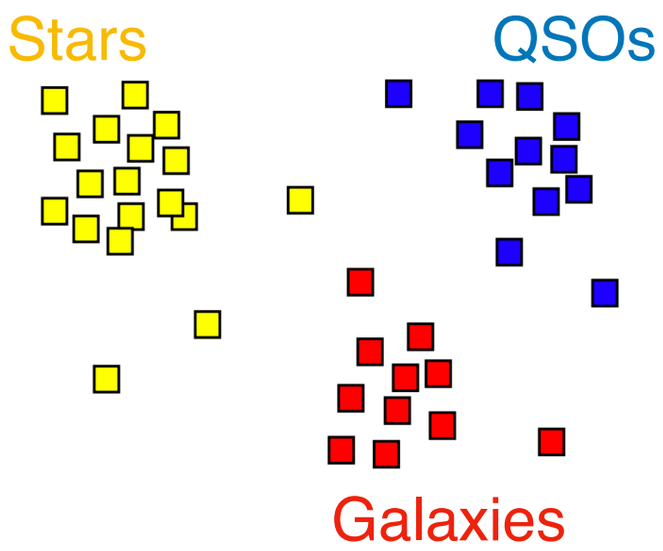


Science is about compression.
Obtaining knowledge = decreasing entropy.

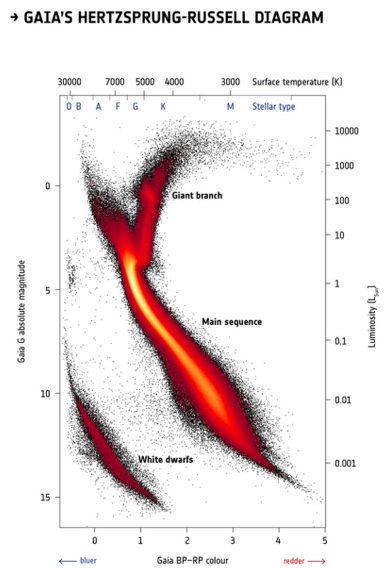
Exercise: ask Emille about a type Ia supernova.

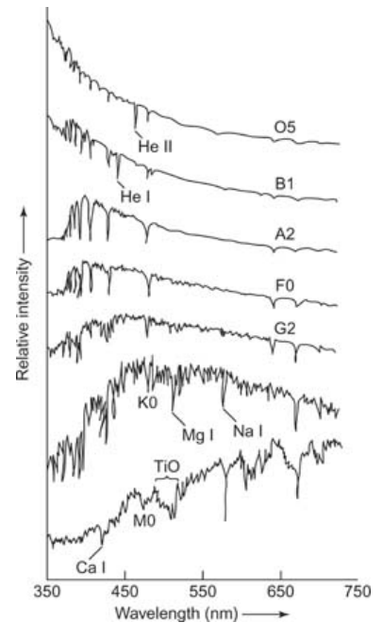
$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Clustering

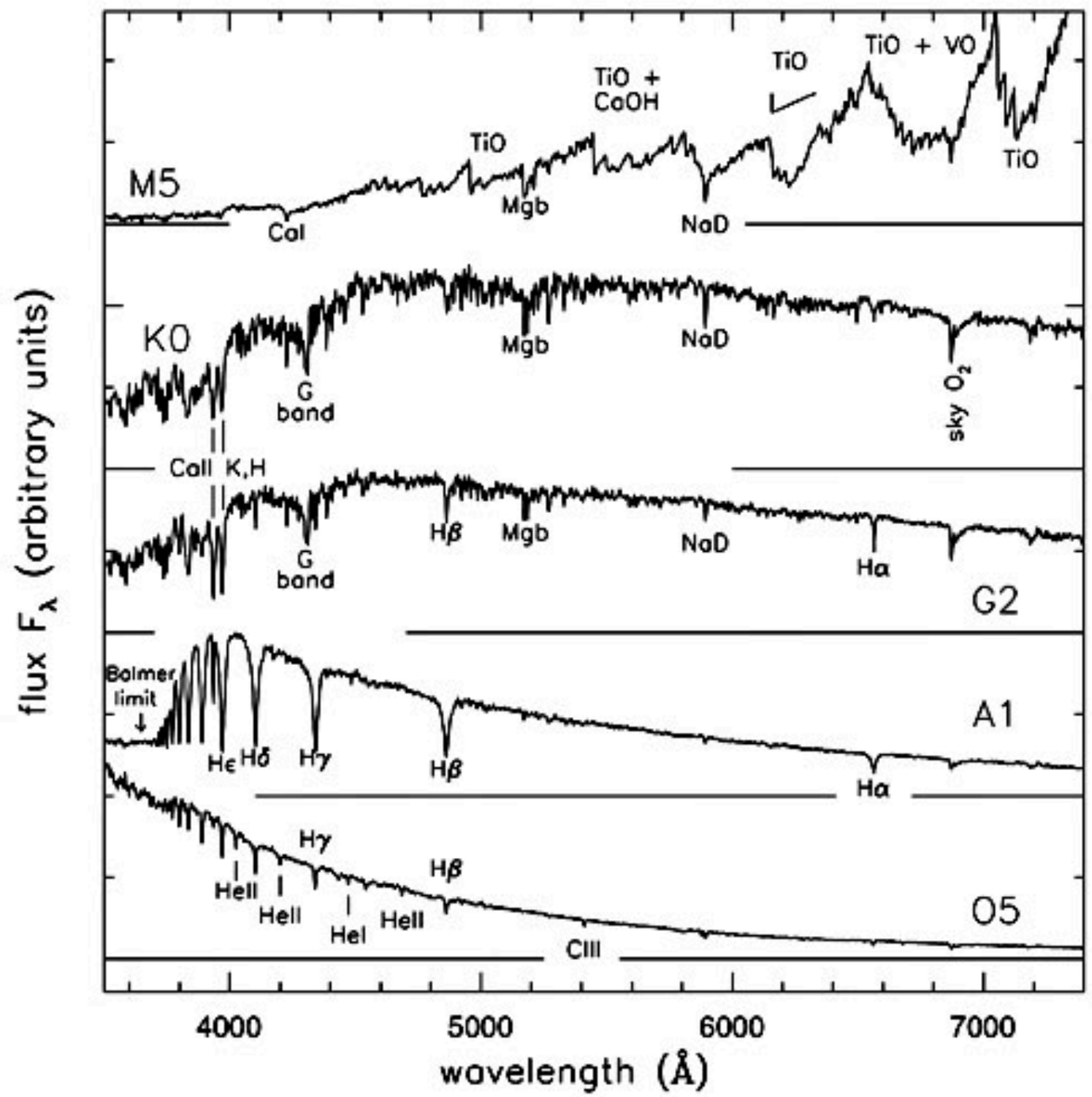


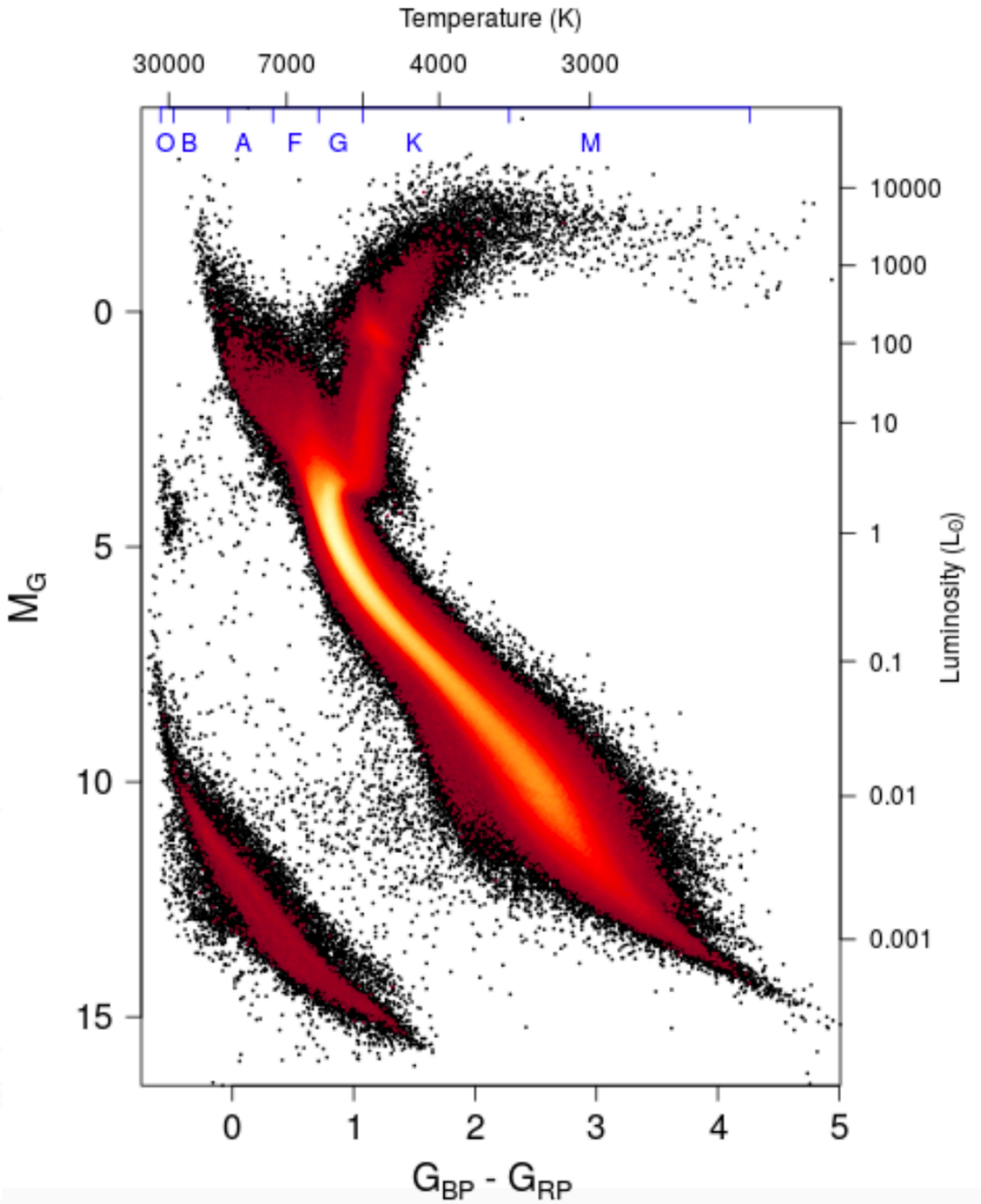
Dimensionality Reduction





Dimensionality reduction



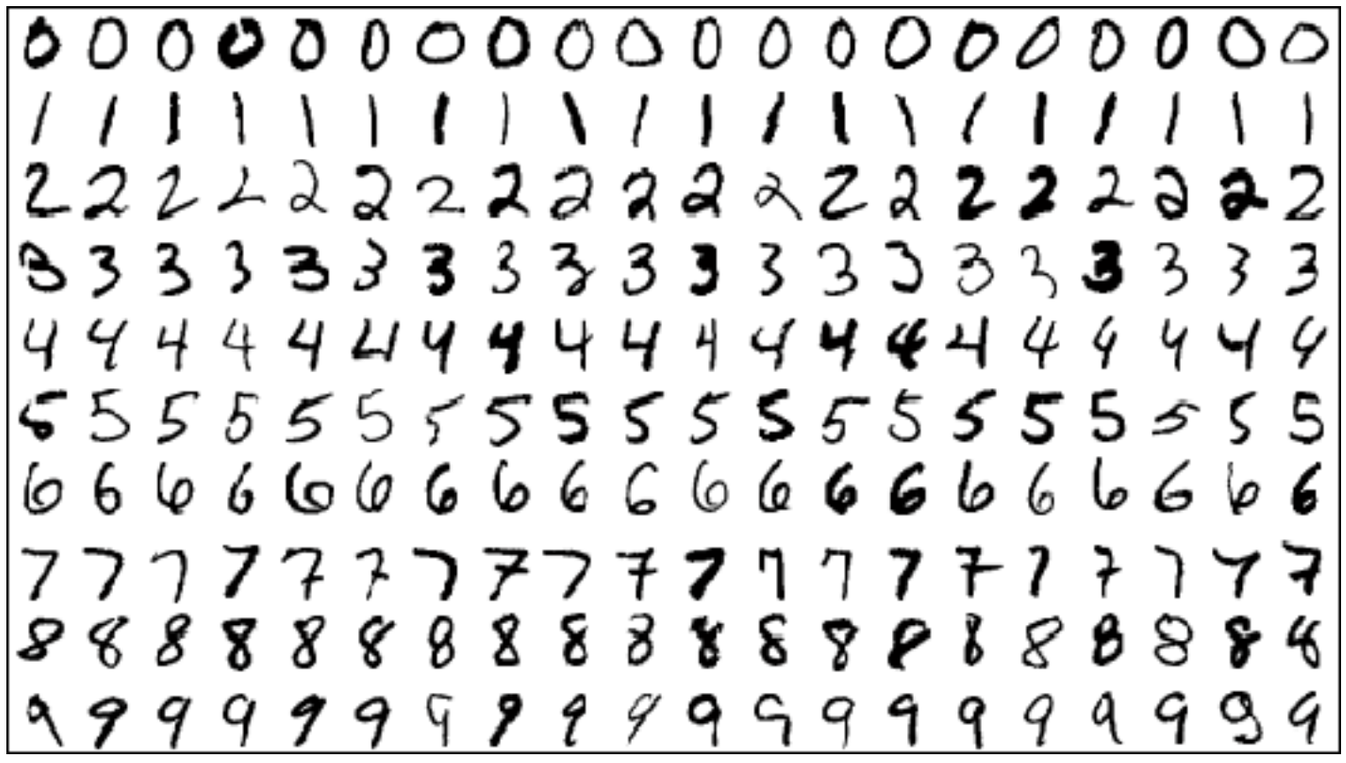


From: Gaia Collaboration et al. 2018

tSNE and UMAP

Dimensionality reduction algorithms used for embedding of high-dimensional data into a low dimensional space (typically 2D or 3D).

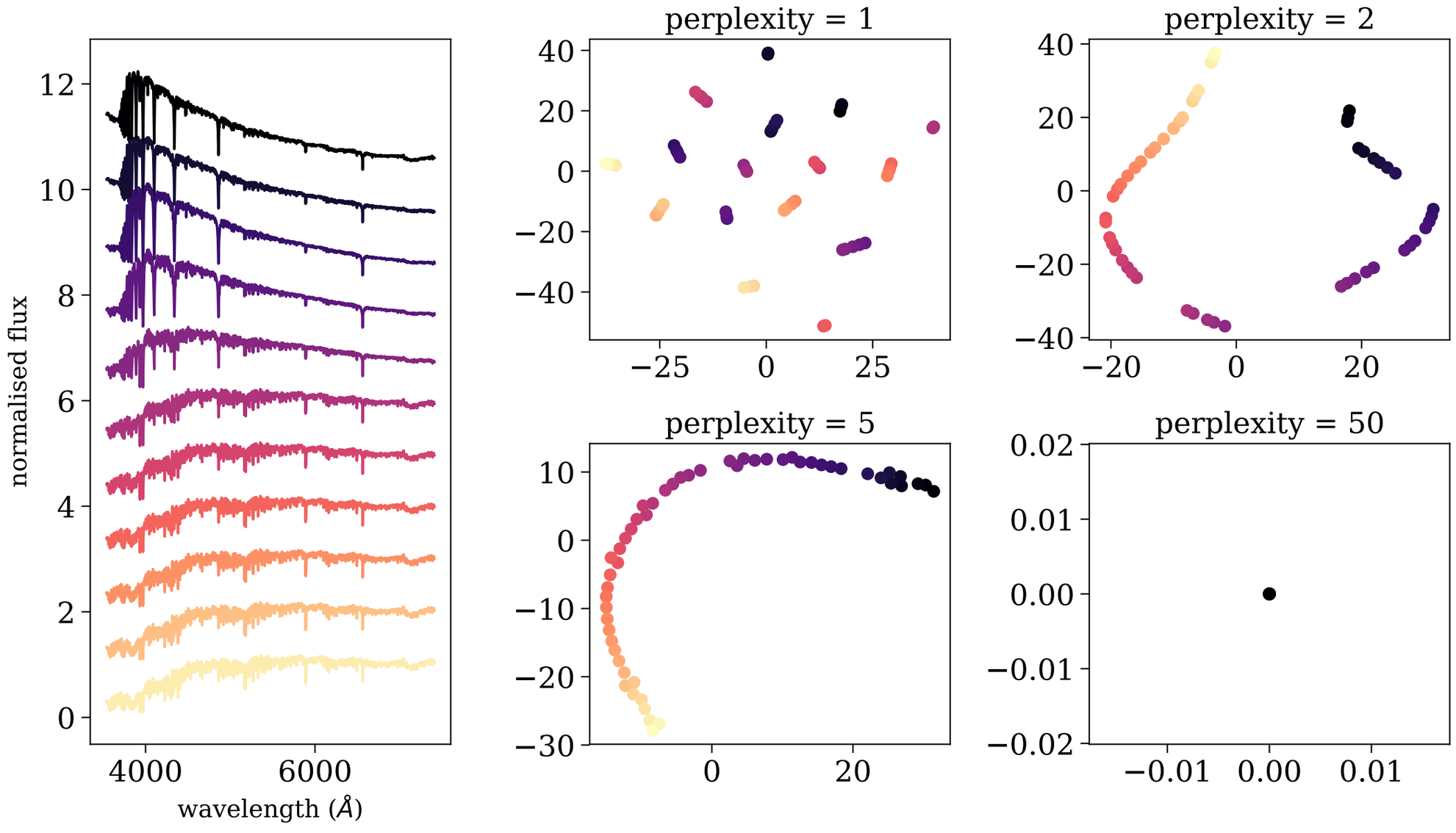
MNIST dataset: 28x28 features per image



tSNE: L.J.P. van der Maaten and G.E. Hinton (2008), https://lvdmaaten.github.io/tsne/
UMAP: Leland McInnes, John Healy, and James Melville (2018), https://umap-learn.readthedocs.io/en/latest/

tSNE and UMAP

The resulting embedding depends on several choices (e.g., the distance metric) and on several hyper-parameters.

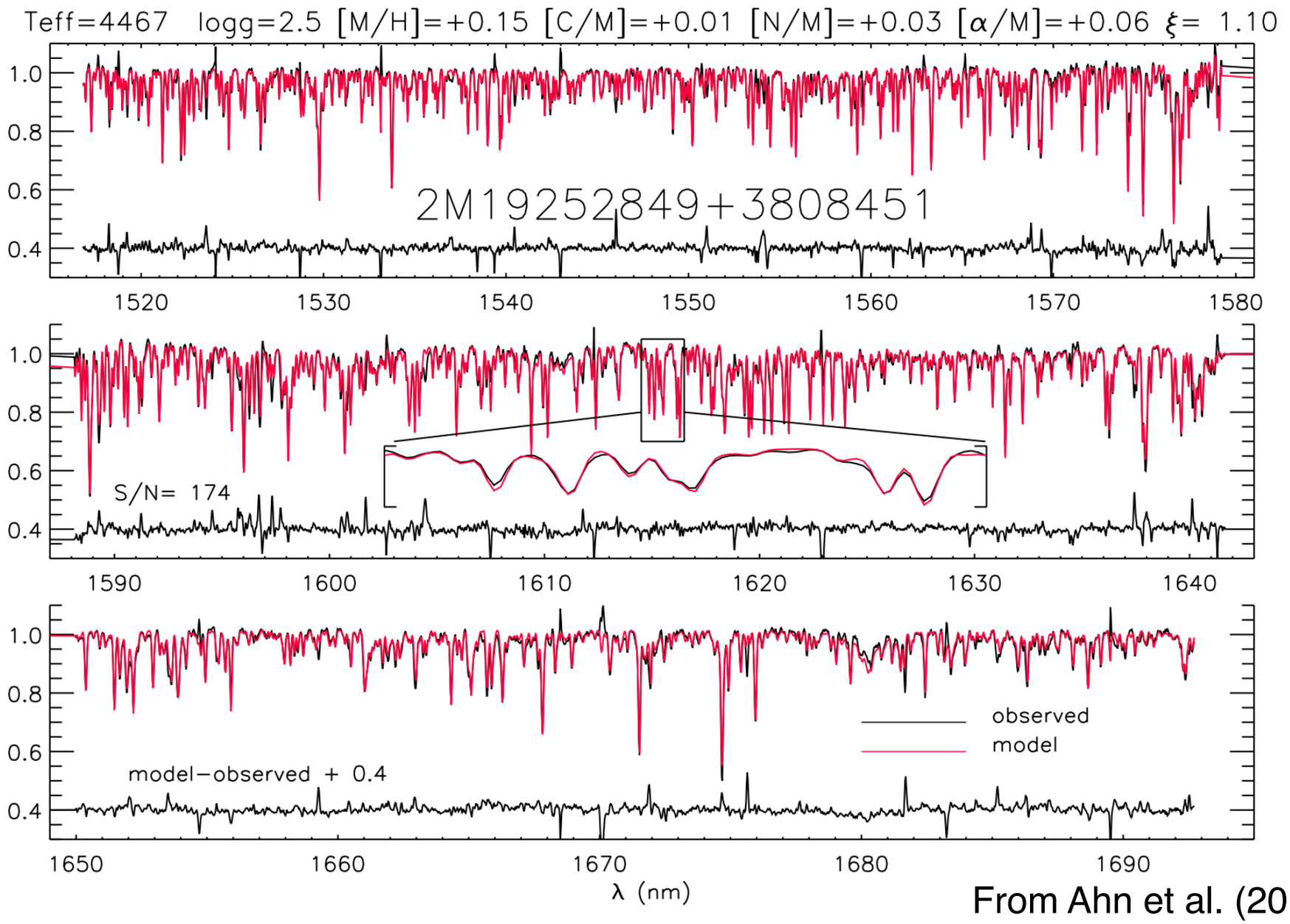


tSNE: L.J.P. van der Maaten and G.E. Hinton (2008), https://lvdmaaten.github.io/tsne/

UMAP: Leland McInnes, John Healy, and James Melville (2018), https://umap-learn.readthedocs.io/en/latest/

tSNE: example with APOGEE stars

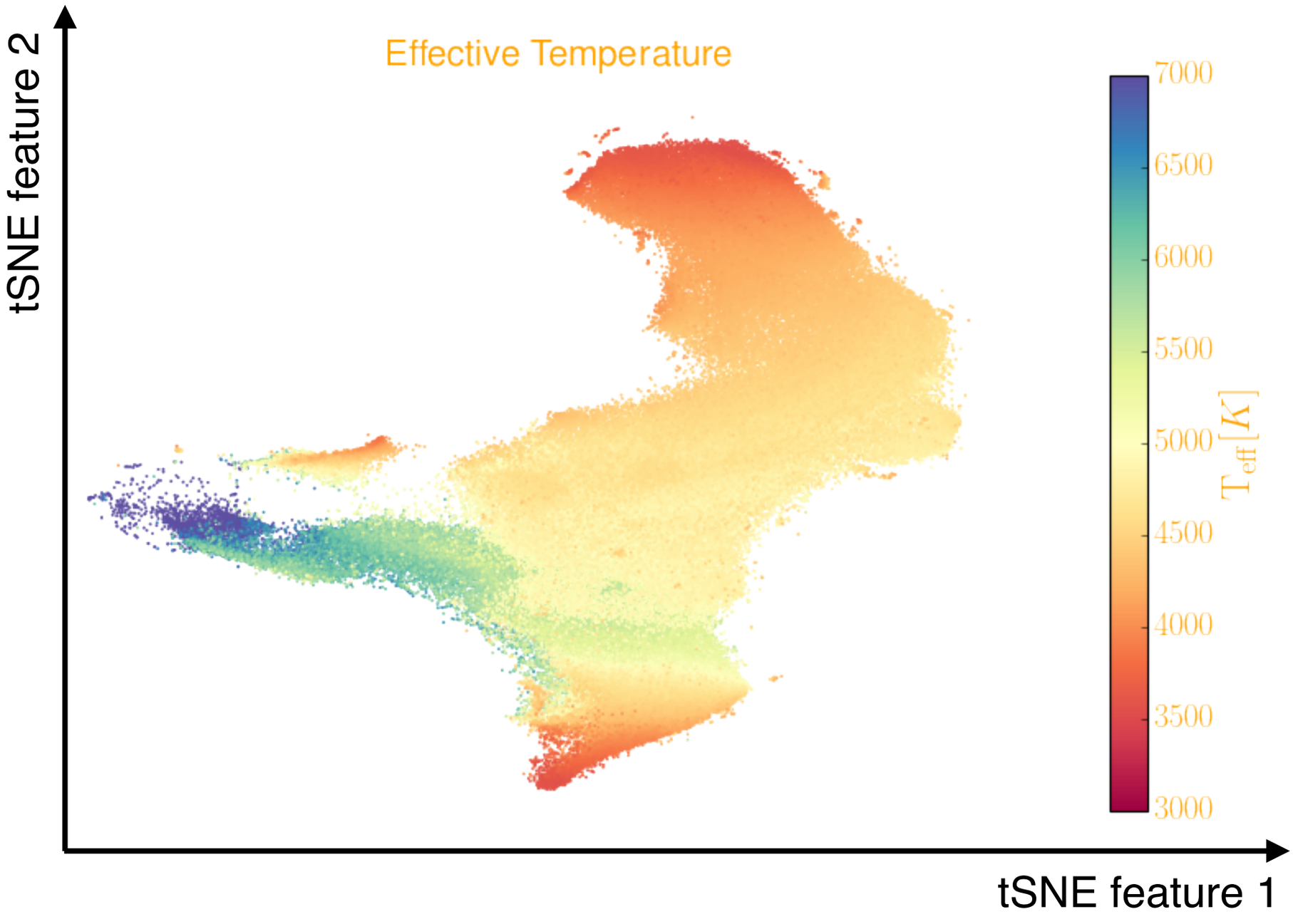
APOGEE: high resolution infrared spectra of more than 100,000 stars in the Milky Way.
The survey provides the processed infrared spectra, and catalogs of radial velocities, stellar parameters, and abundances derived from these spectra.



From Ahn et al. (2014)

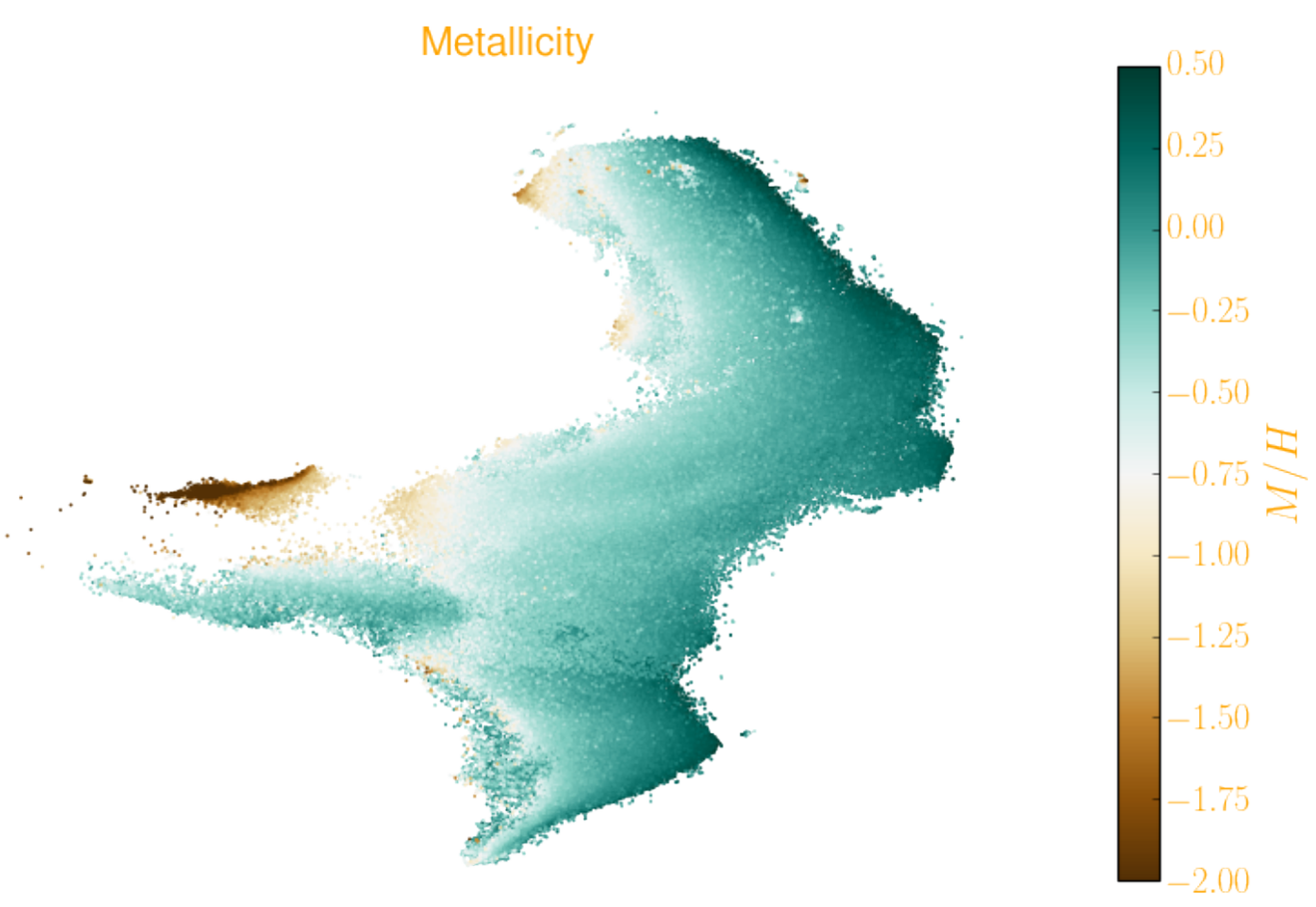
tSNE: example with APOGEE stars

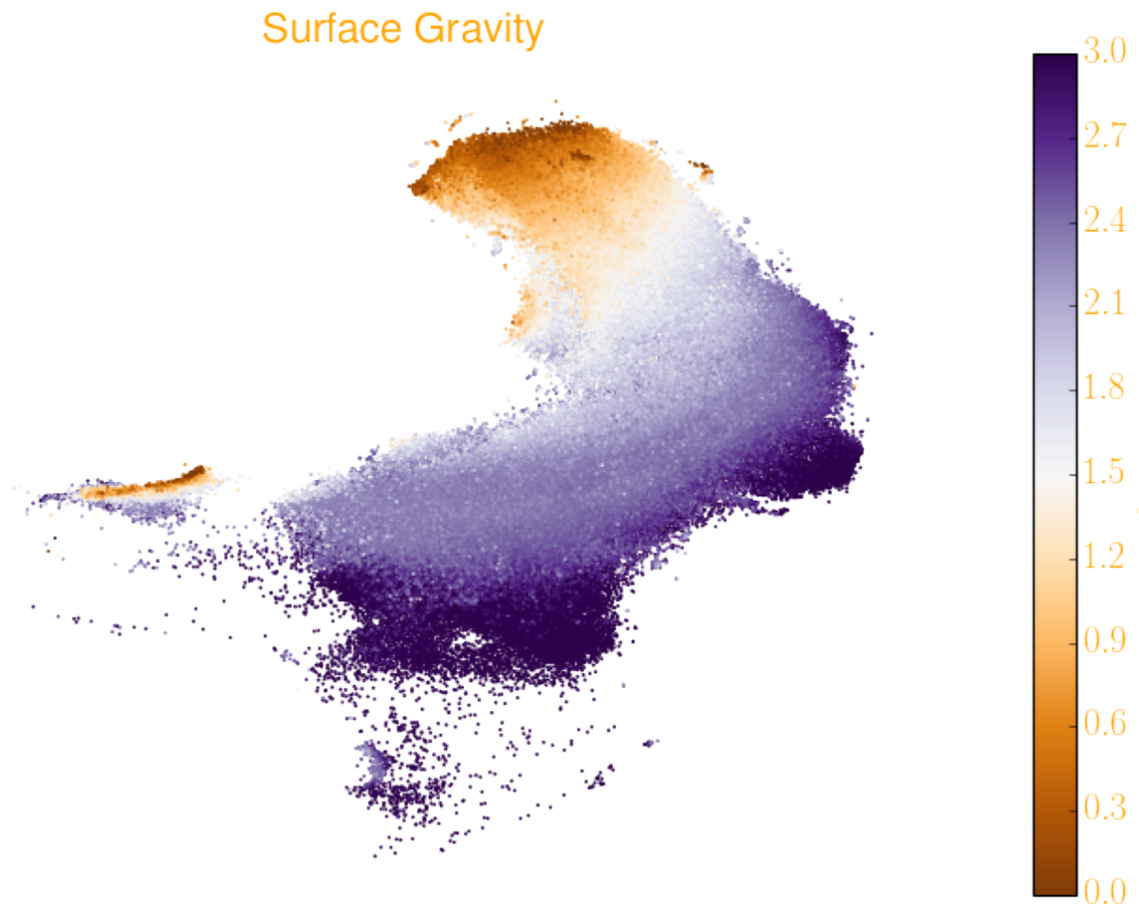
Dimensionality reduction of the APOGEE dataset: we assigned distances between the objects in the sample using unsupervised Random Forest (see Baron & Poznanski 2017), and applied tSNE for dimensionality reduction. The resulting embedding was then colored according to derived parameters from the public catalog (see Reis et al. 2018).



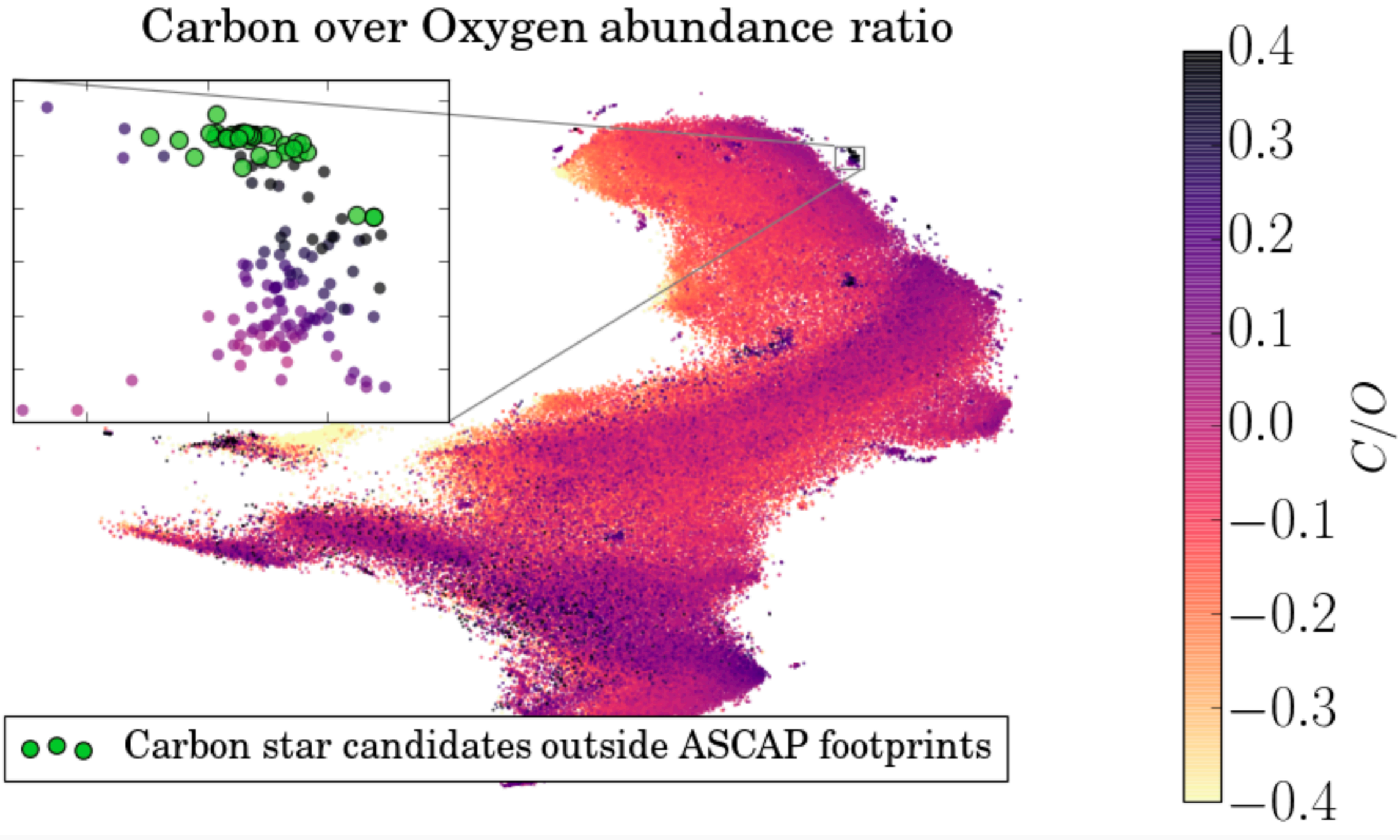
tSNE: example with APOGEE stars

Dimensionality reduction of the APOGEE dataset: we assigned distances between the objects in the sample using unsupervised Random Forest (see Baron & Poznanski 2017), and applied tSNE for dimensionality reduction. The resulting embedding was then colored according to derived parameters from the public catalog (see Reis et al. 2018).

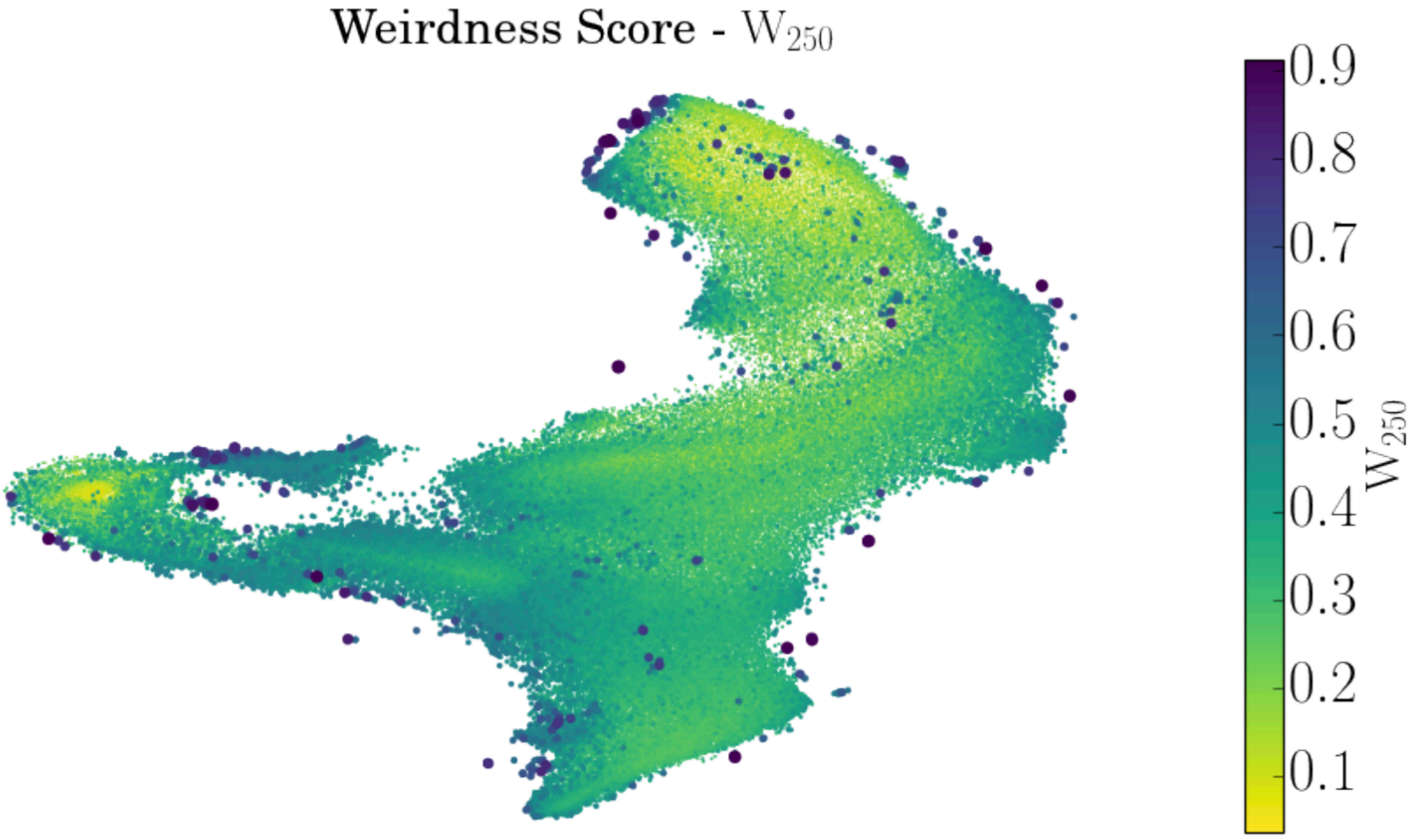




tSNE: example with APOGEE stars

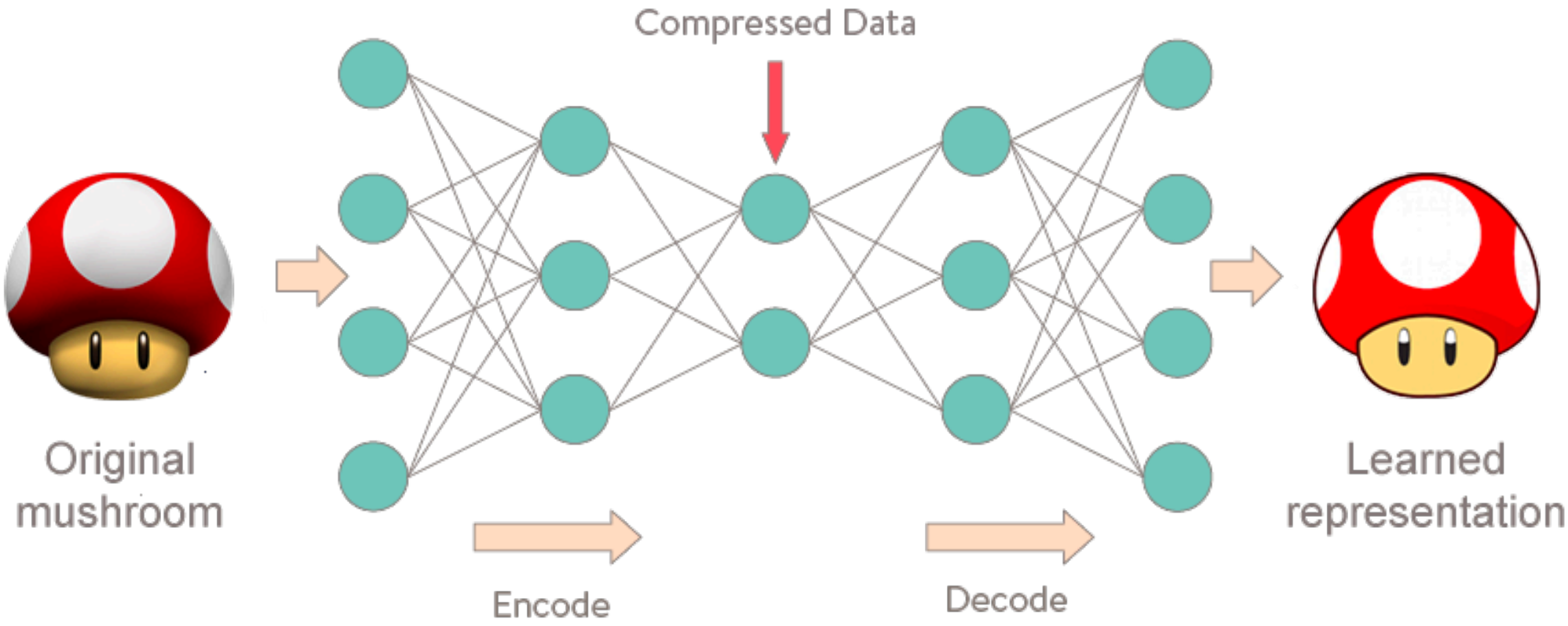


tSNE: outliers in the APOGEE dataset



Autoencoders

A neural network used to learn an efficient low-dimensional representation of the input dataset, and can be used for compression, dimensionality reduction, and visualization.

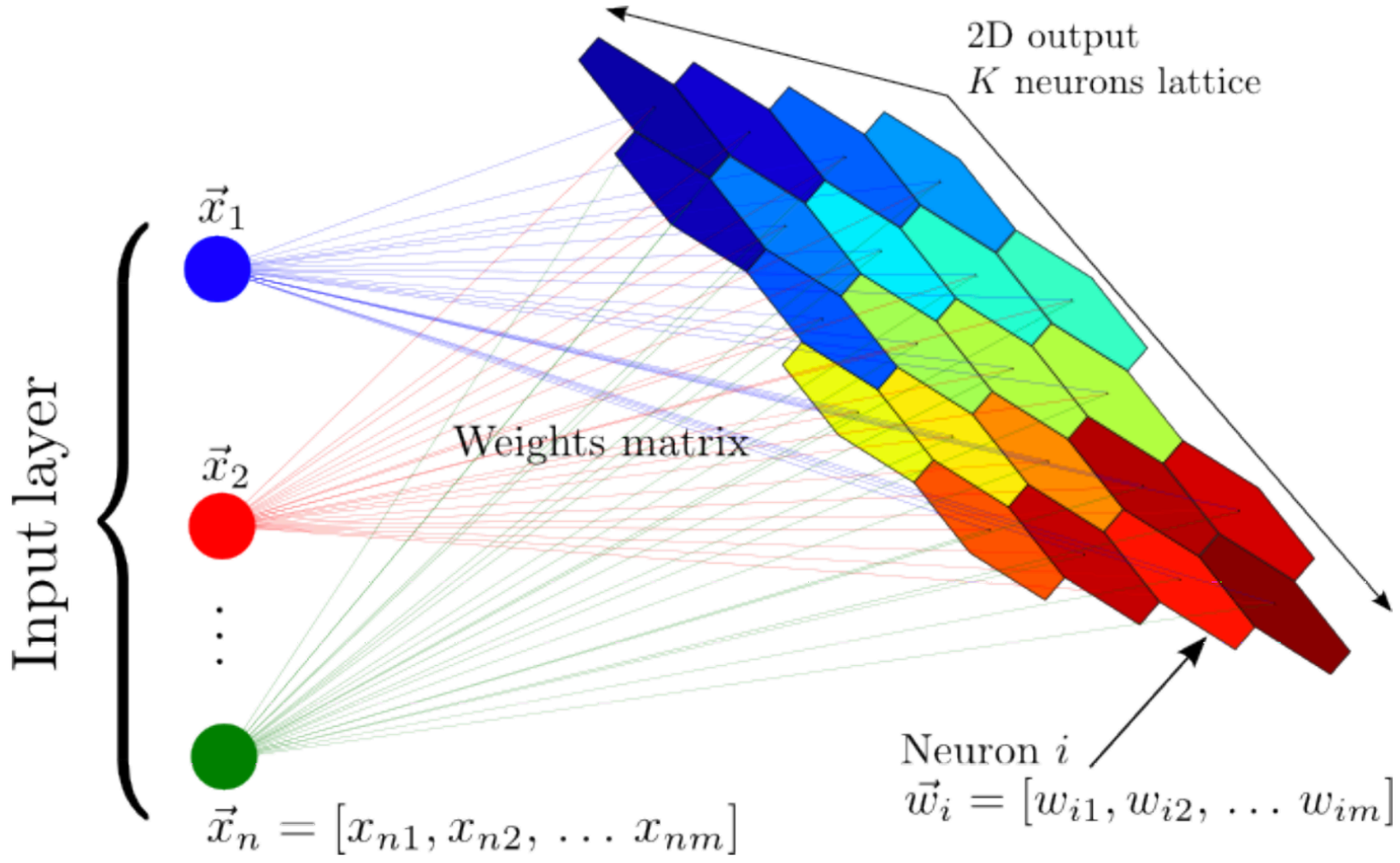


$$\text{loss function} = \left(\text{Original mushroom} - \text{Learned representation} \right)^2$$

Examples in astronomy include: Gianniotis et al. (2015); Yang & Li (2015); Gianniotis et al. (2016); Ma et al. (2018b); Schawinski et al. (2018); Ralph et al. (2019).

Self-Organizing Maps

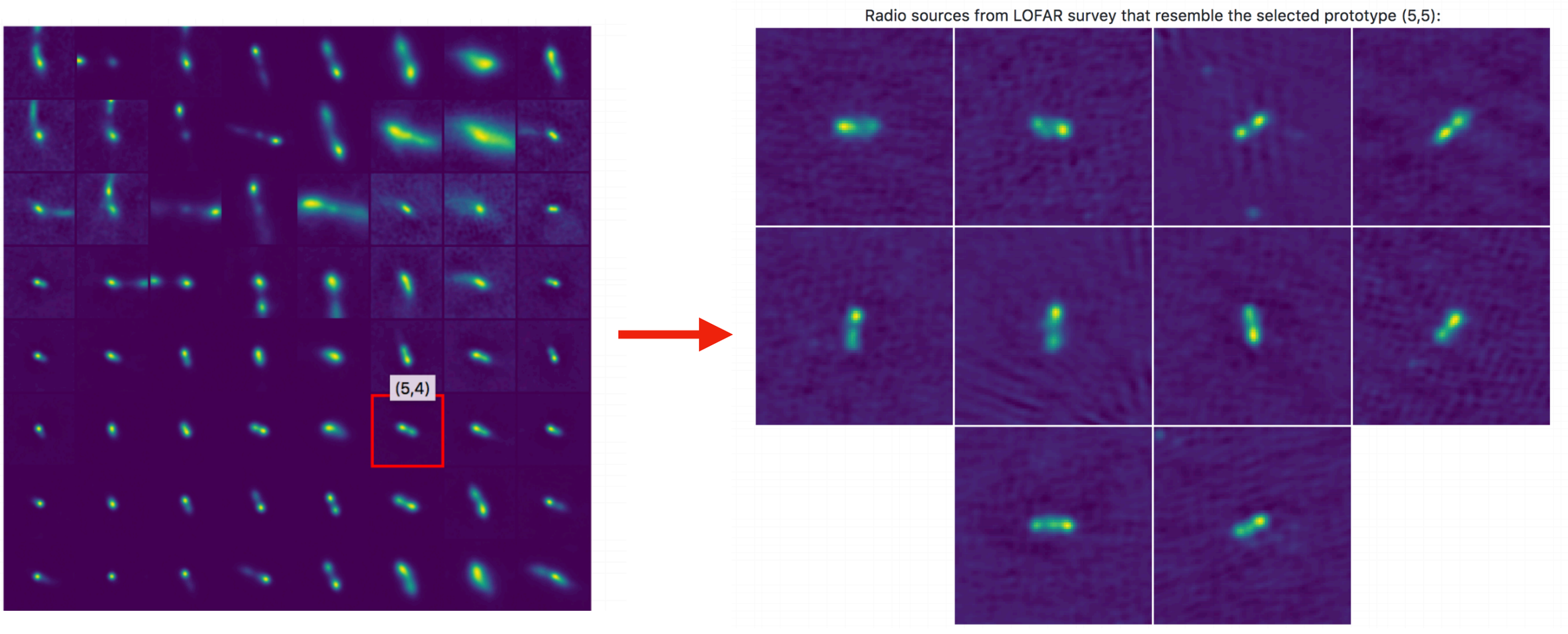
An unsupervised neural network used to produce a low dimensional representation of the input dataset using a set of prototypes. The prototypes are built during training to match as closely as possible the input data.



Examples in astronomy include: Fustes et al. (2013); Carrasco Kind & Brunner (2014); Armstrong et al. (2016); Polsterer et al. (2016); Armstrong et al. (2017); Meusinger et al. (2017); Rahmani et al. (2018); Galvin et al. (2019); Ralph et al. (2019).

Self-Organizing Maps

SOM prototypes allow a fast and efficient exploration of large datasets. The distance from the prototypes can be used to retrieve similar objects and to search for outliers.

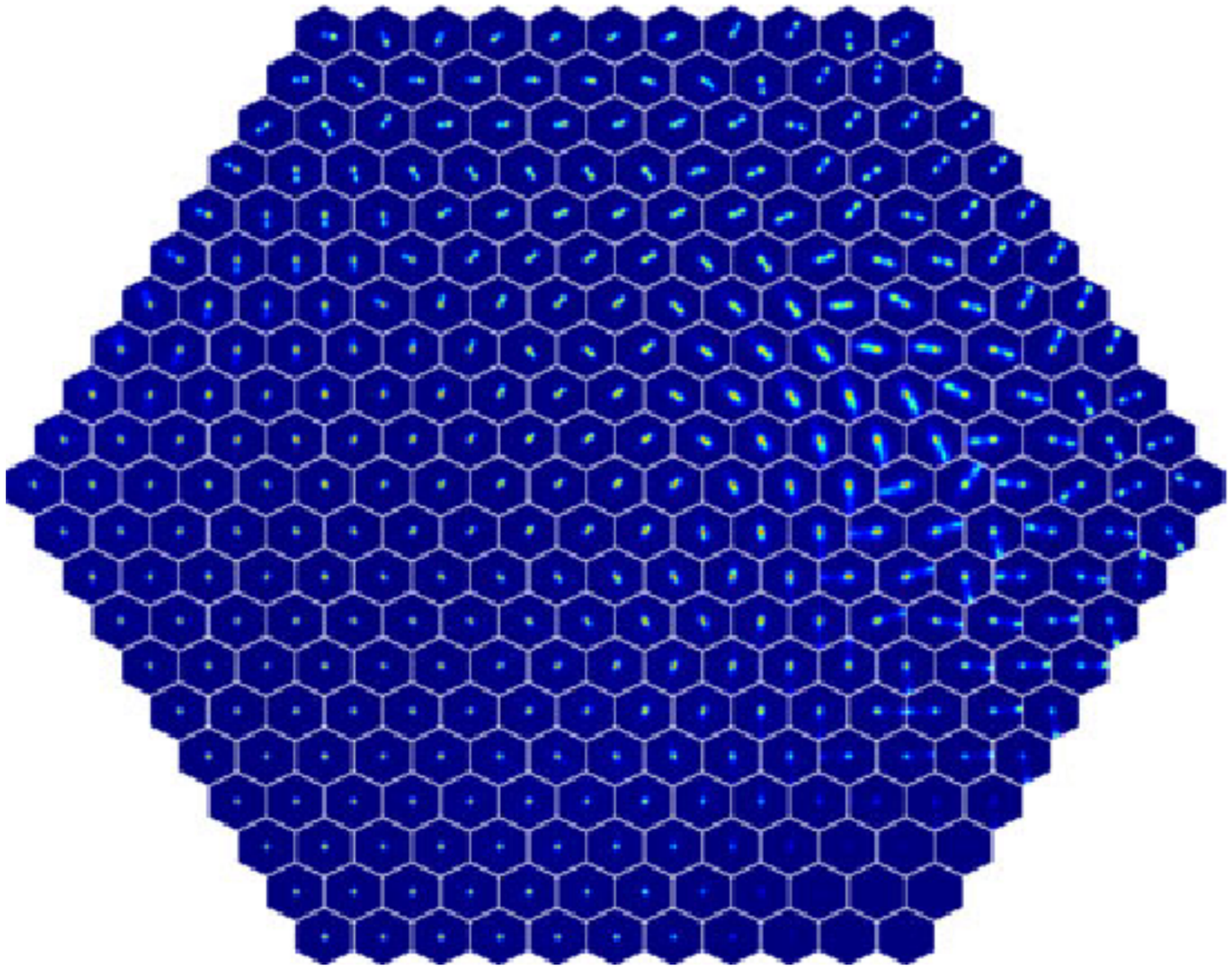


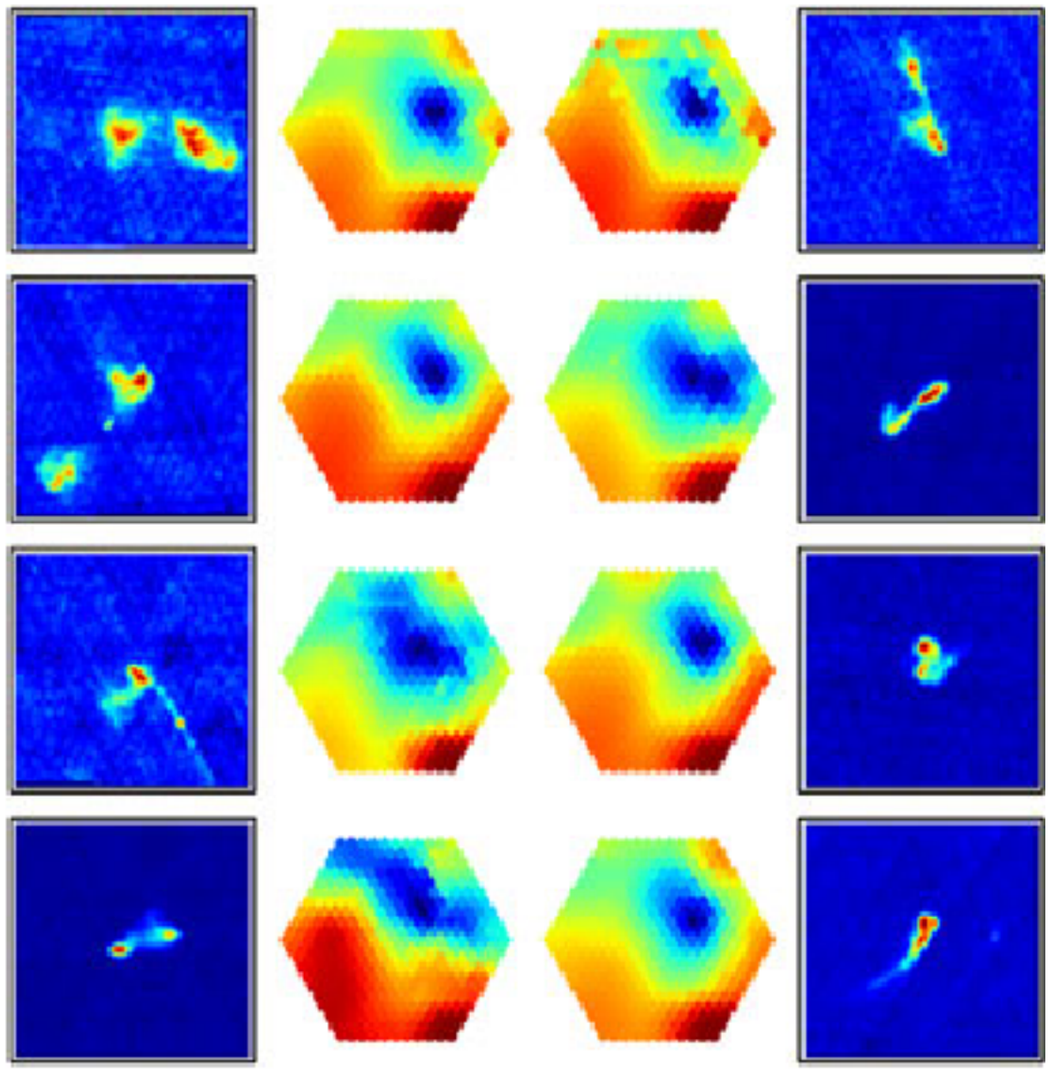
Taken from J. Harwood presentation

Examples in astronomy include: Fustes et al. (2013); Carrasco Kind & Brunner (2014); Armstrong et al. (2016); Polsterer et al. (2016); Armstrong et al. (2017); Meusinger et al. (2017); Rahmani et al. (2018); Galvin et al. (2019); Ralph et al. (2019).

Self-Organizing Maps: PINK

The low dimensional representation and the resulting prototypes depend on internal choices (e.g., distance assignment). Thus, they are not invariant under rotations and flips.



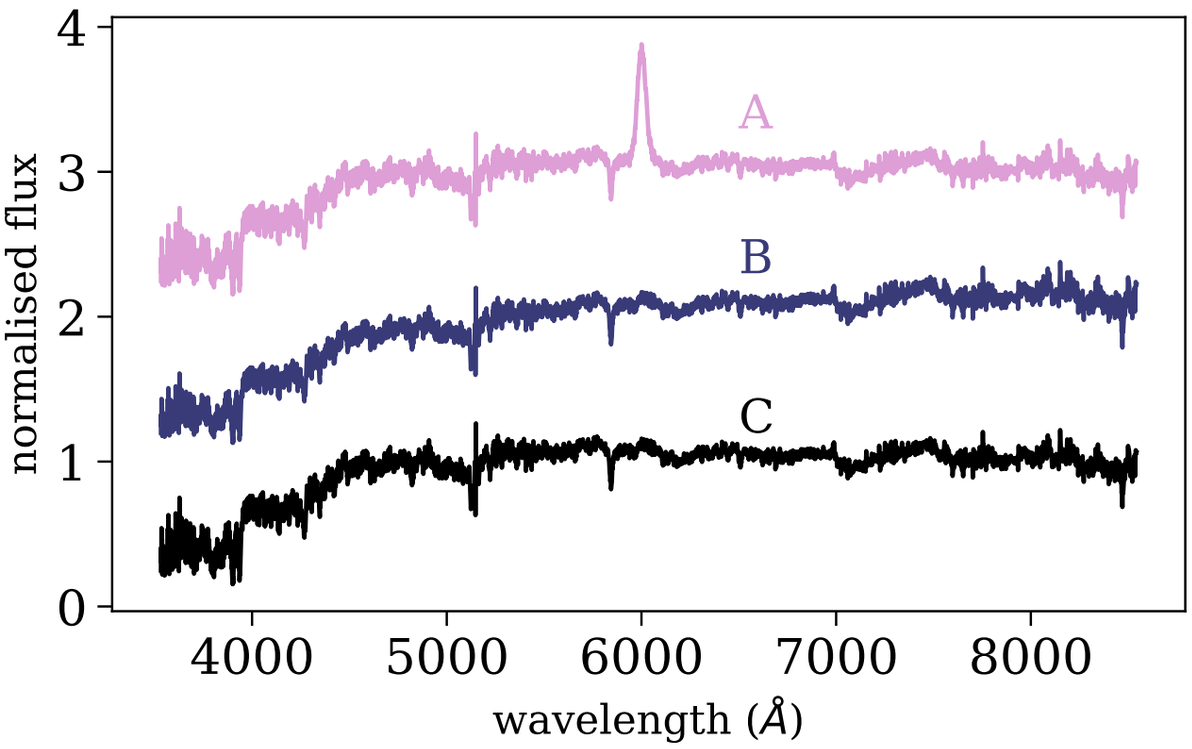


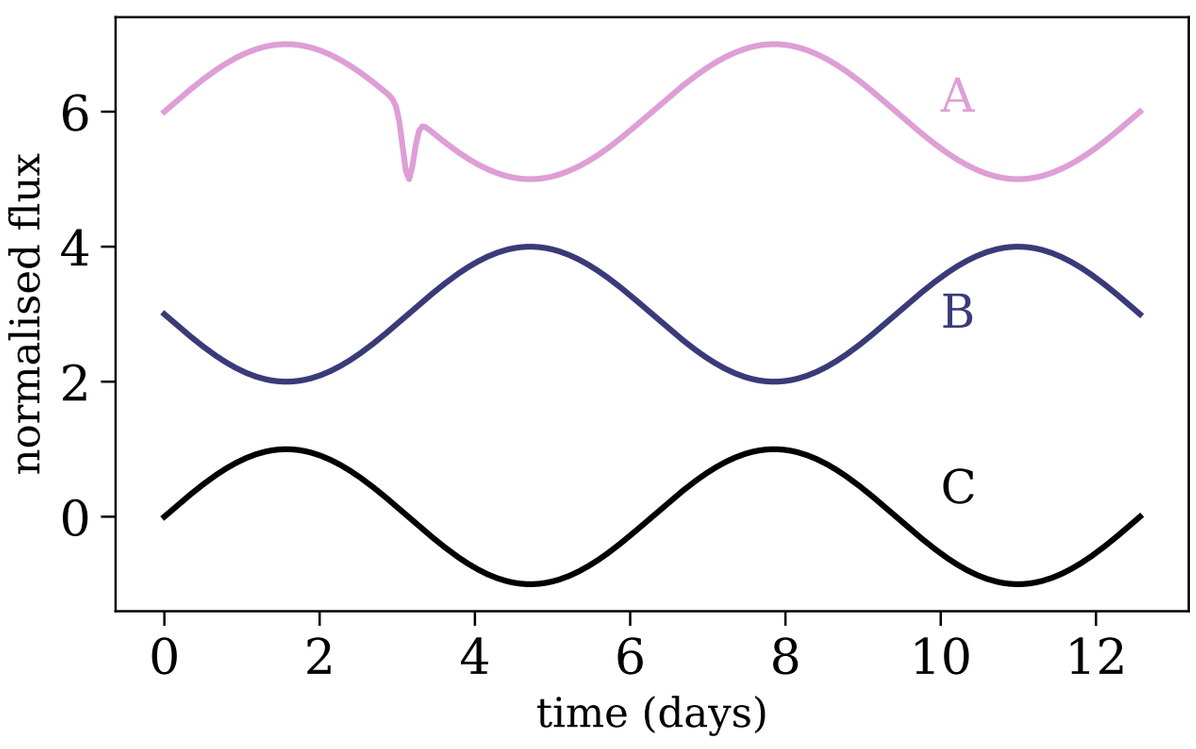
Examples in astronomy include: Fustes et al. (2013); Carrasco Kind & Brunner (2014); Armstrong et al. (2016); Polsterer et al. (2016); Armstrong et al. (2017); Meusinger et al. (2017); Rahmani et al. (2018); Galvin et al. (2019); Ralph et al. (2019).

Current Challenges

(*) The resulting dimensionality reduction depends on the algorithm’s hyper-parameters. How do we choose the “correct" hyper-parameter values?

(*) Most of the algorithms measure distances between the objects in the sample. Which distance metric is appropriate for the dataset at hand?

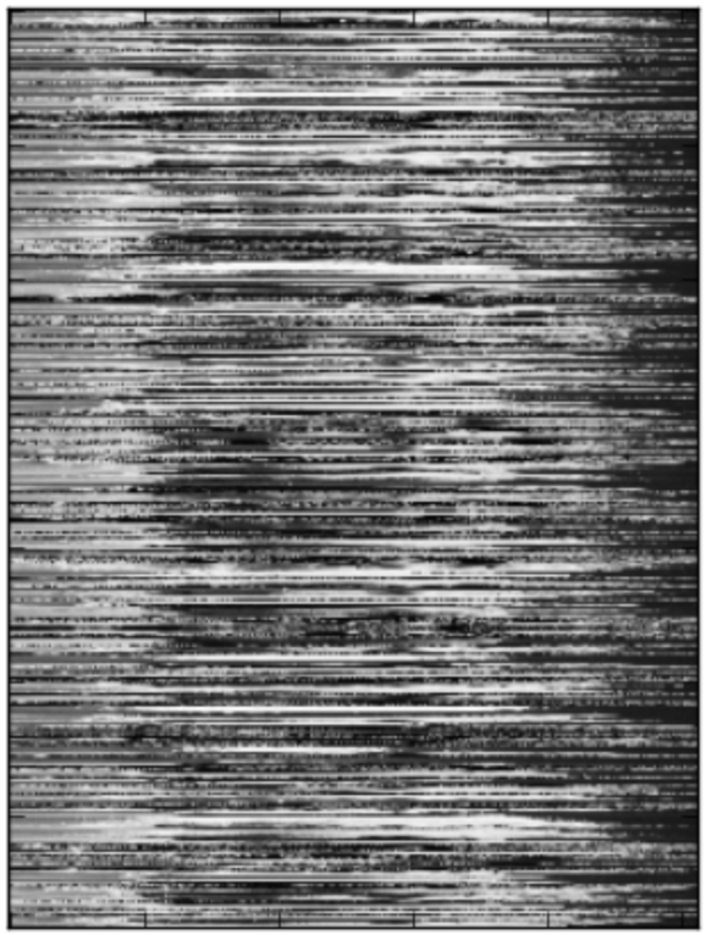




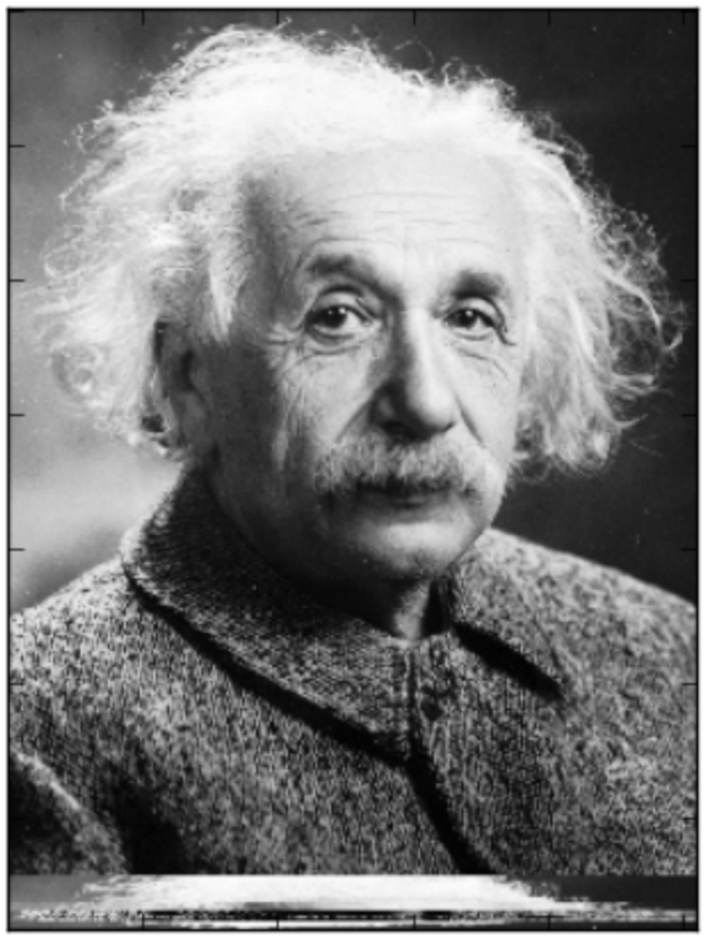
The Sequencer

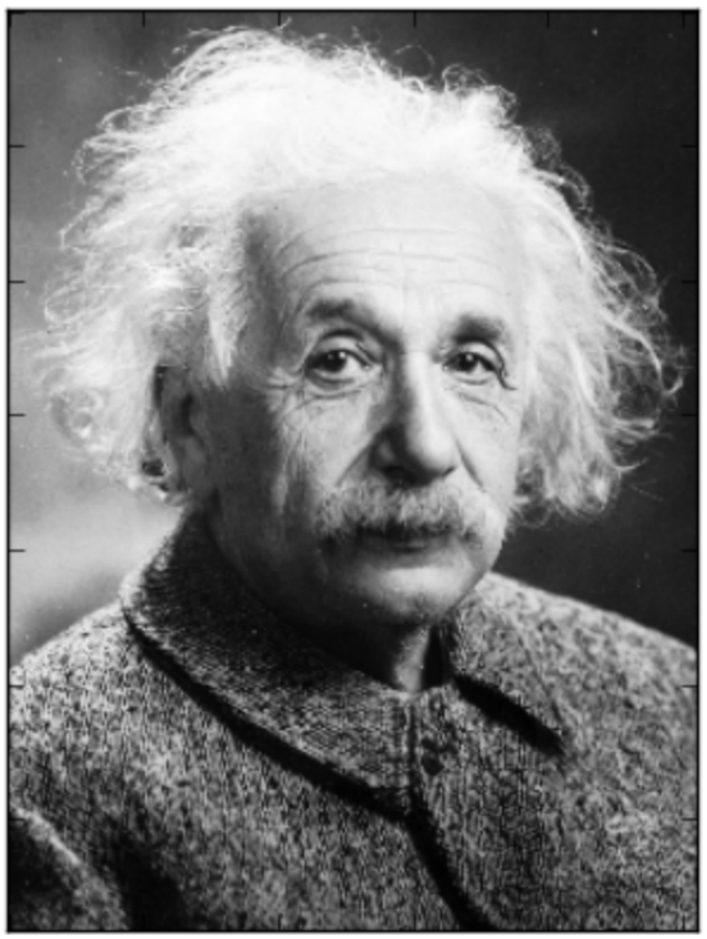
Baron & Ménard in prep.

Input



Output



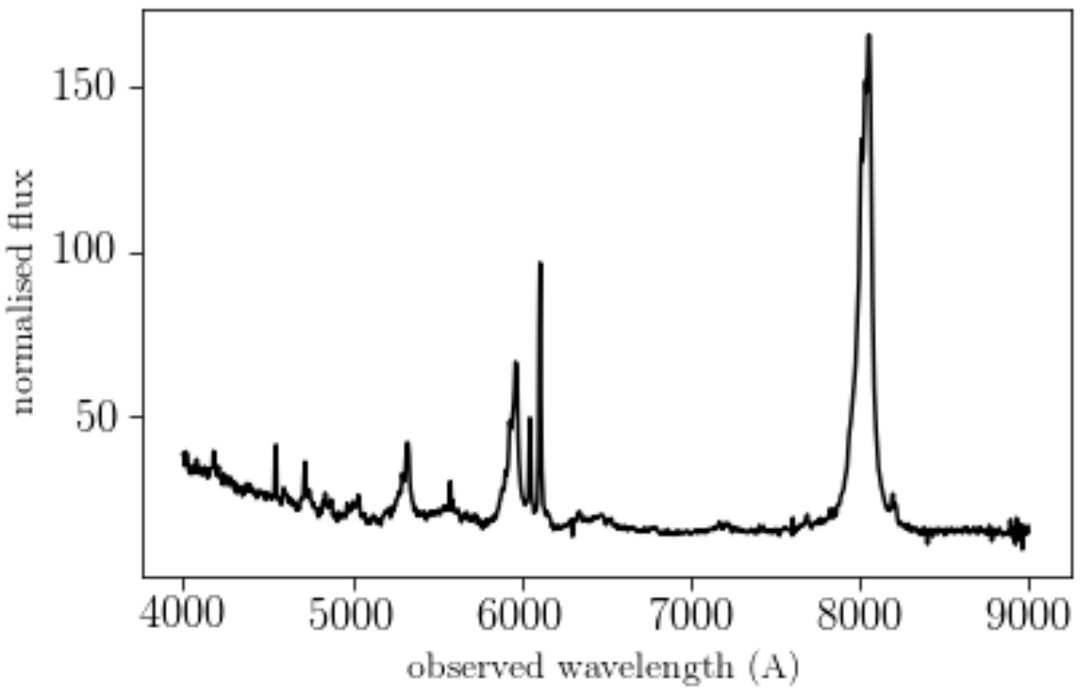


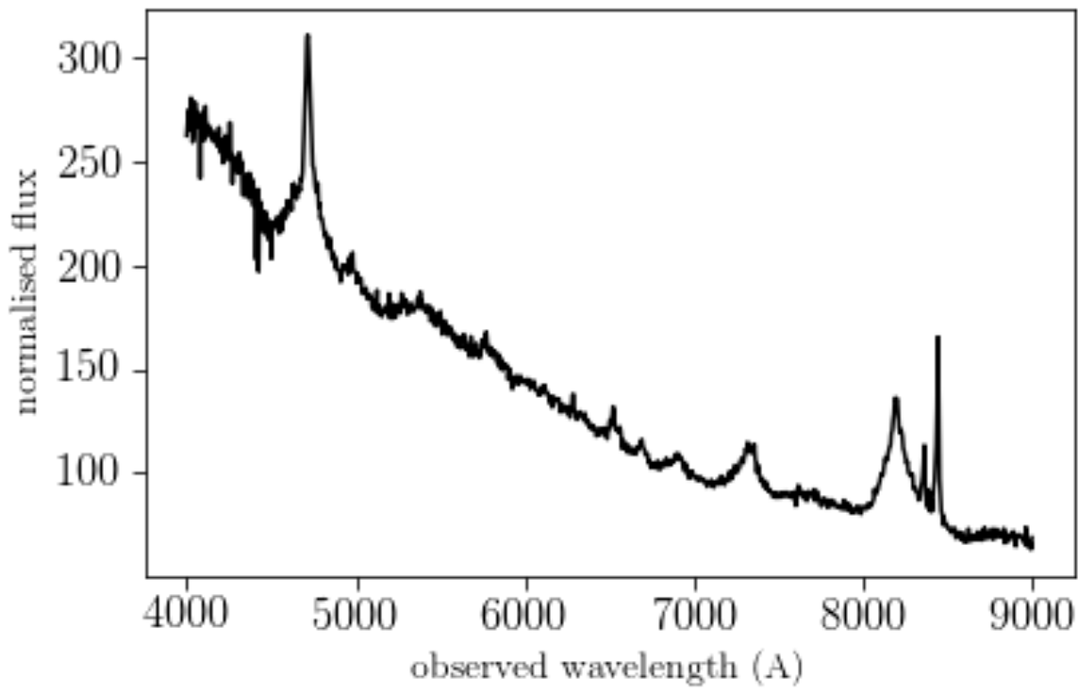
The algorithm reorders the data according to a detected sequence:

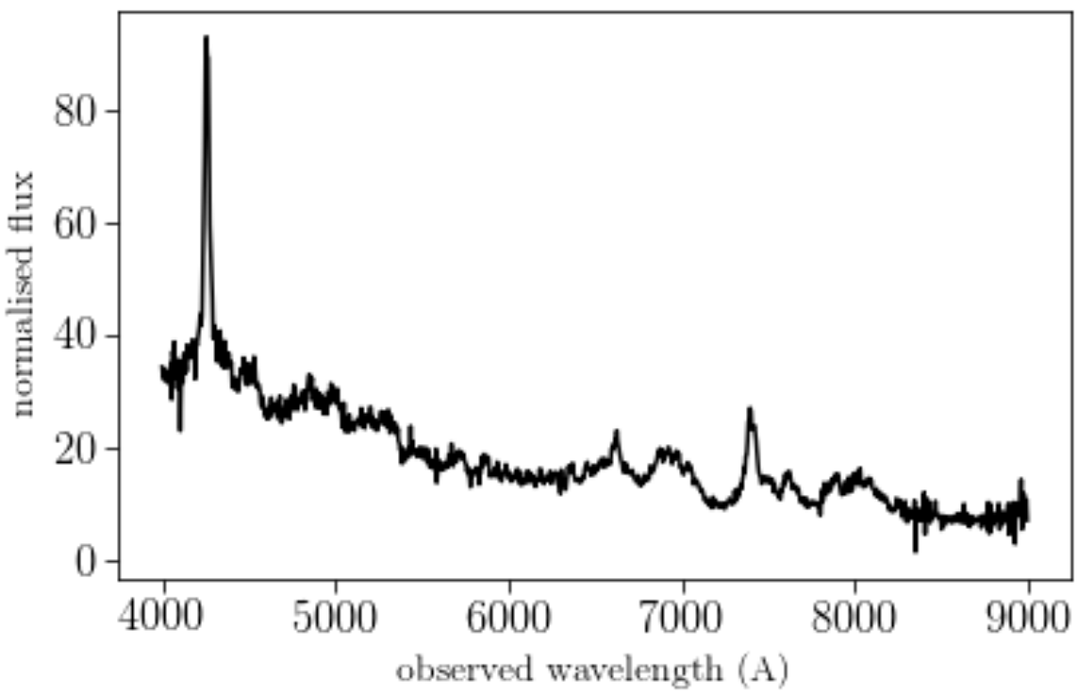
- Based on pure statistics - no training, no randomness, result is always the same.
- Provides a score.
- Algorithm hyper-parameters and distance assignments are optimized using the score. So, result does not depend on hyper-parameters or the distance metric.

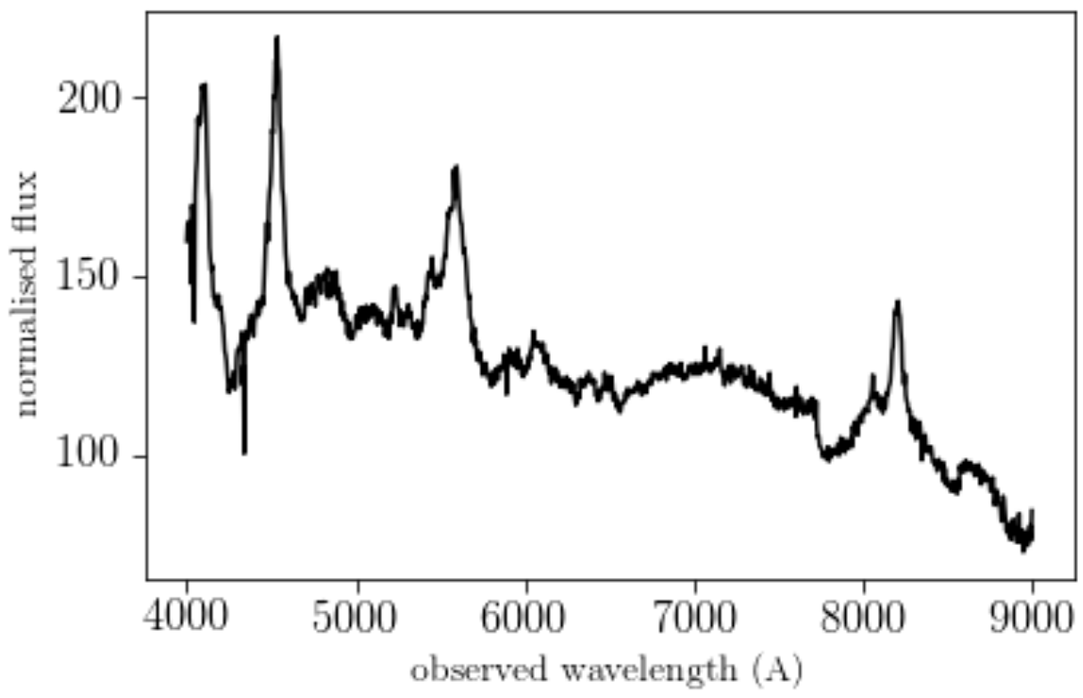
The Sequencer - the quasar case

Baron & Ménard in prep.



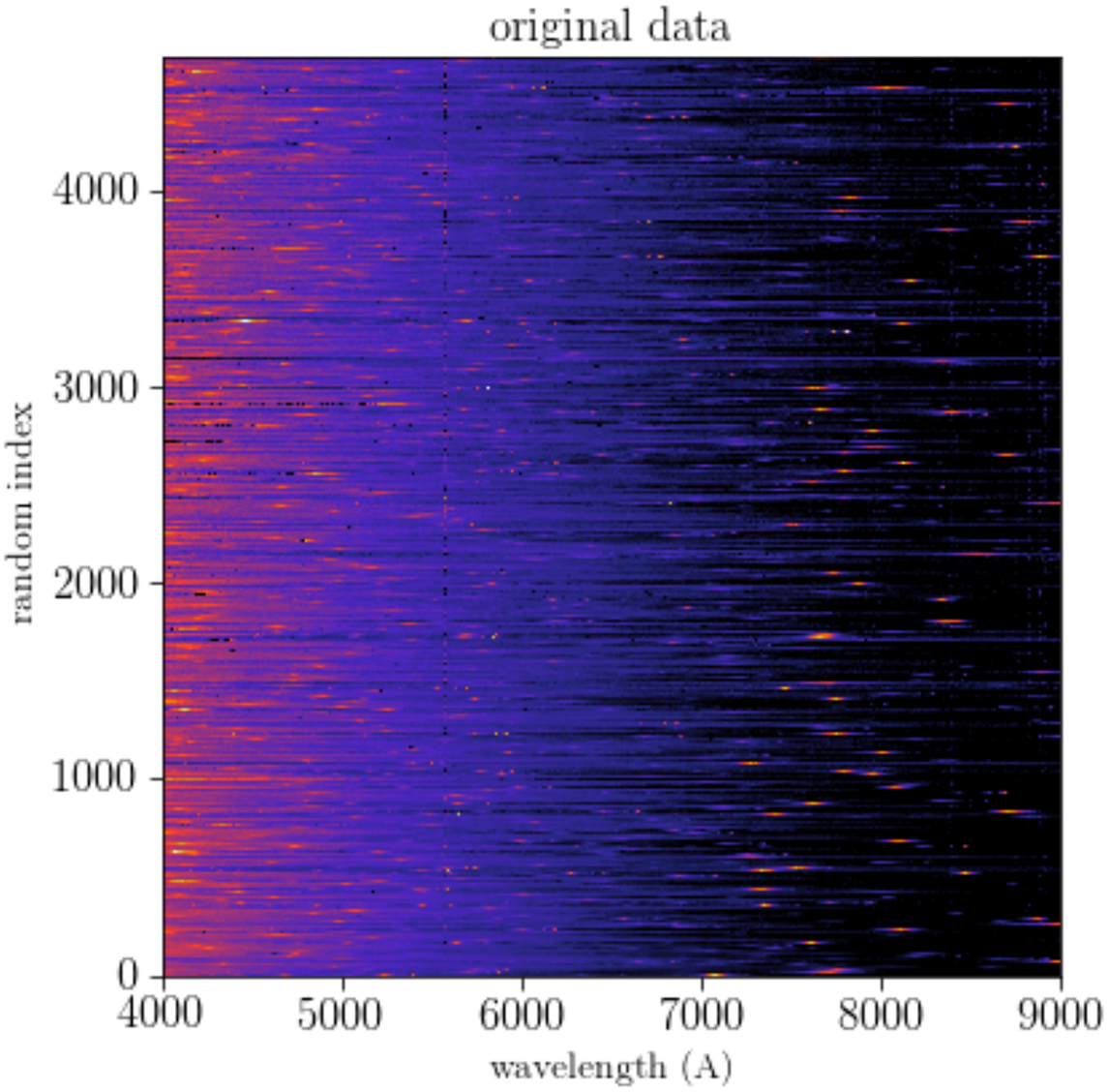


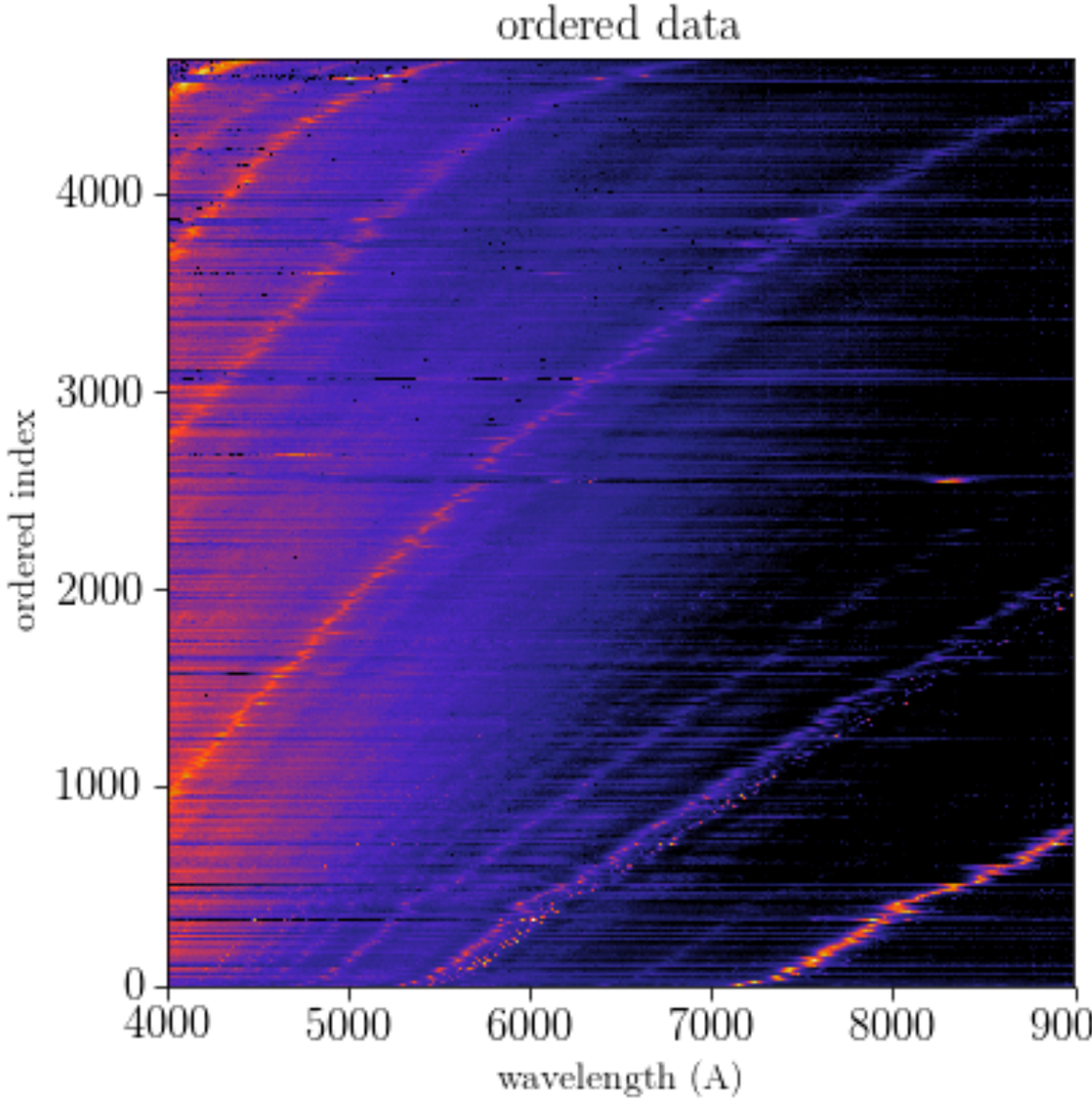


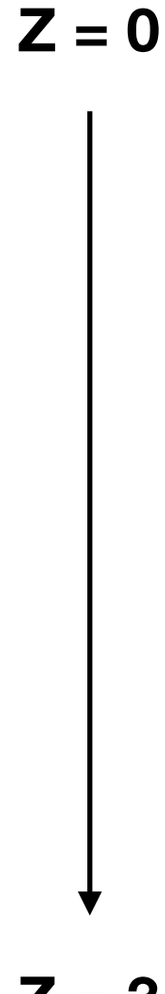


The Sequencer - the quasar case

Baron & Ménard in prep.

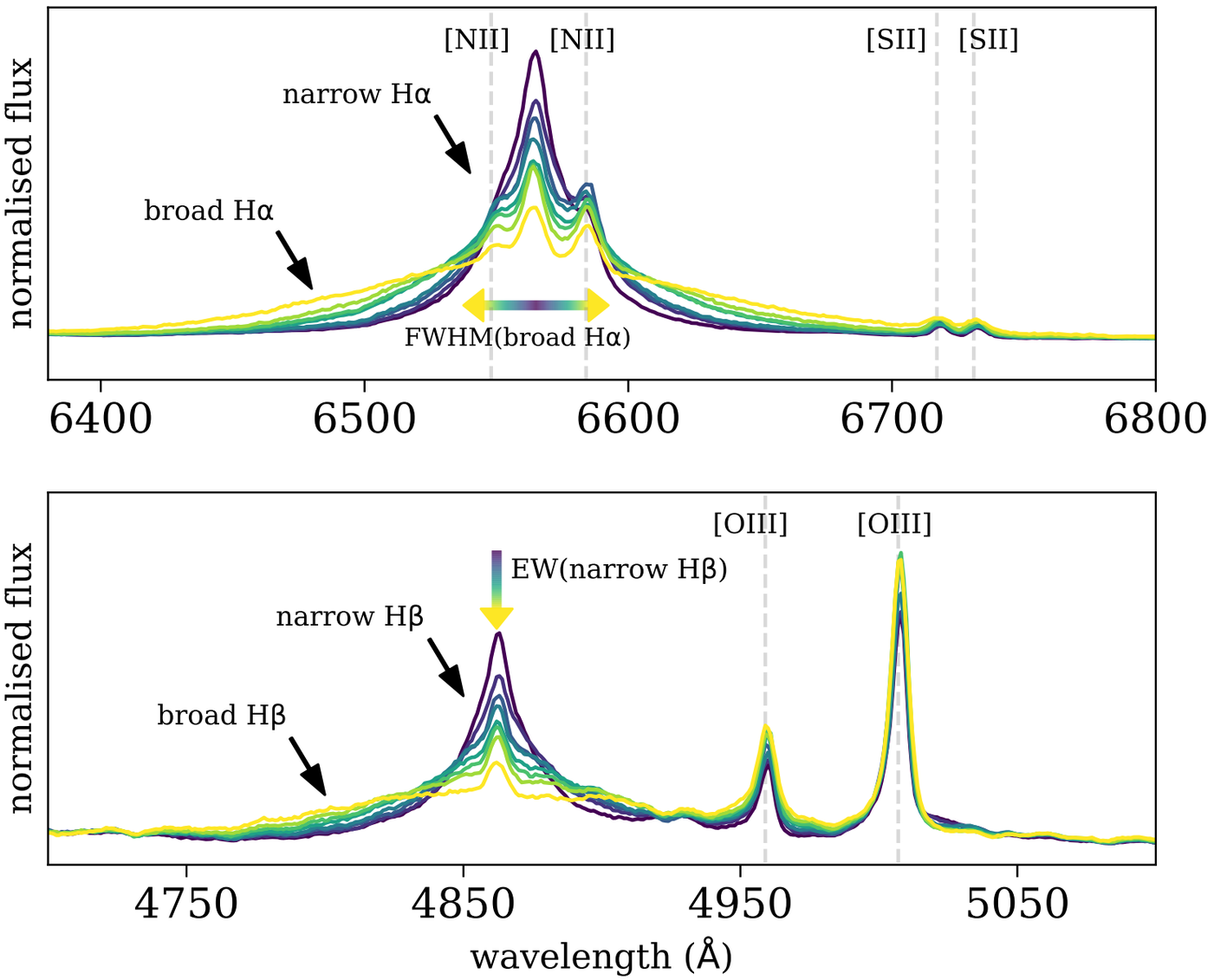


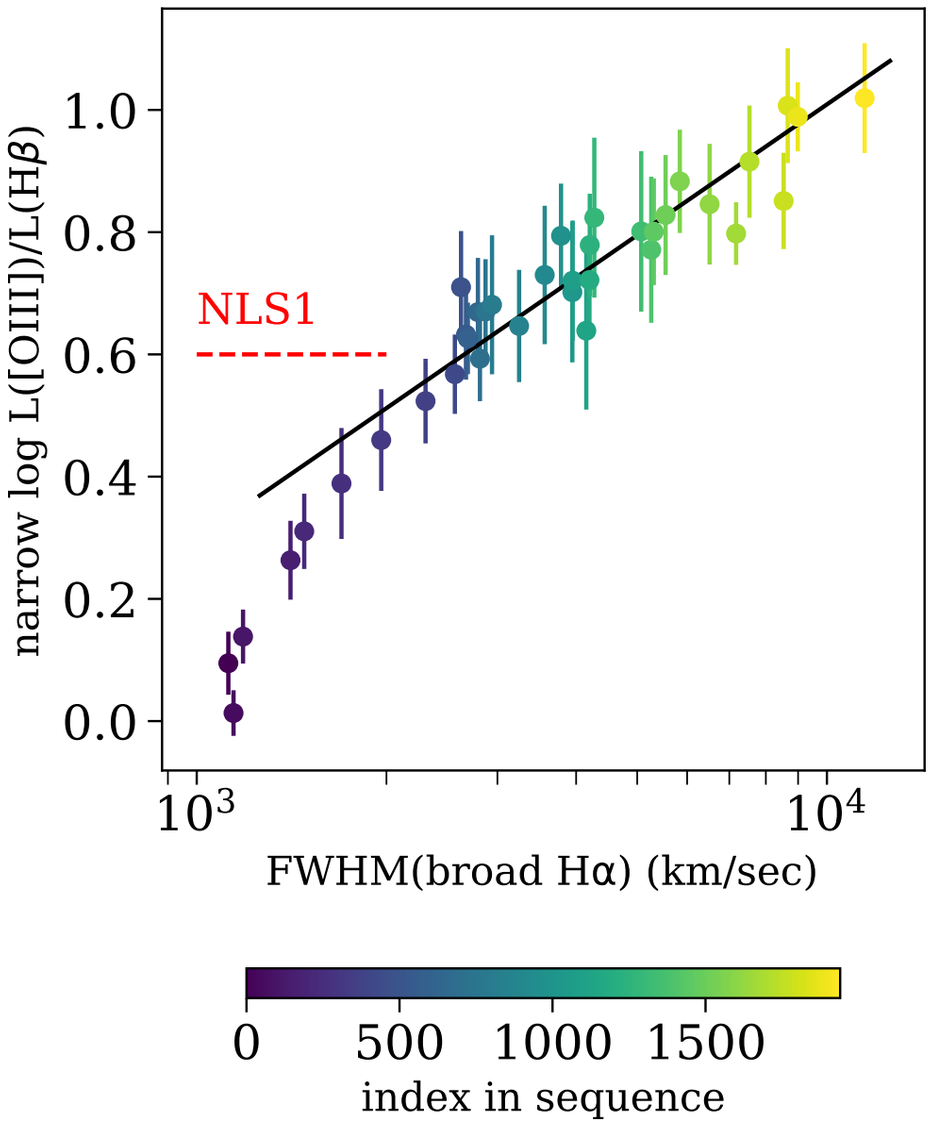




The Sequencer: a new correlation discovered in Active Galactic Nuclei.

Baron & Ménard (2019)

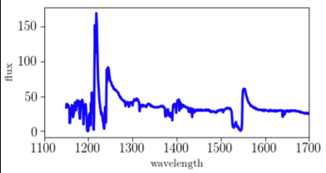


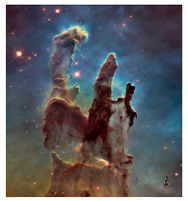


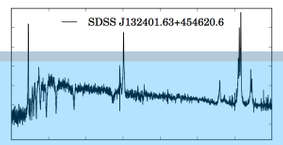
Data Landscape

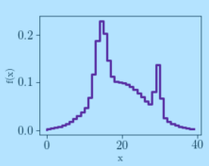
Description length

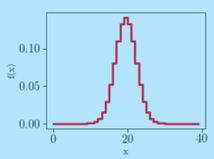
can't describe

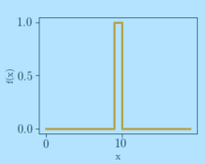


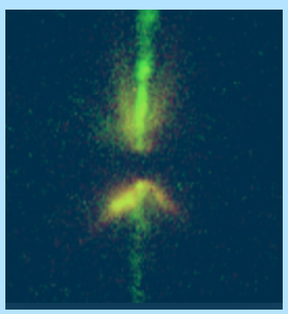


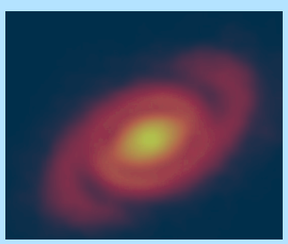












Dimensionality Reduction Algorithms

Sequencer

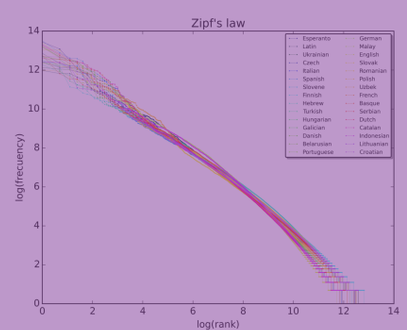
(spectroscopic) outliers

Astronomy

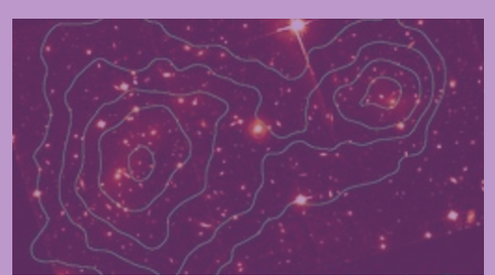
Statistical tools

can describe

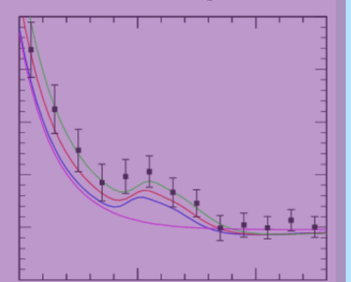
Word frequency in language



Weak lensing



BAO signal



High SNR

Low SNR

Signal to noise ratio

Thanks! :)