Machine learning techniques meet binaries

Gregor Traven

Mathieu Van der Swaelmen
Thibault Merle
Sofia Feltzing
Ross Church
Pablo Navarro
Klemen Čotar
Galah team

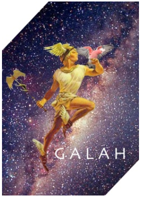
LUND UNIVERSITY

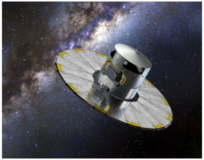
University *of Ljubljana*
Faculty *of Mathematics and Physics*

Lyon, June 2019


BIG
DATA

Science of large samples





THOUSANDS -> MILLIONS of binary stars

Science of large samples

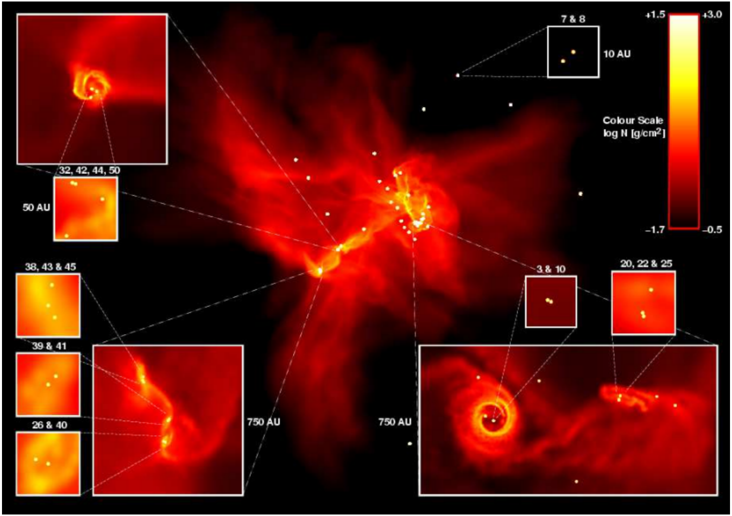


Figure: Bate et al. 2002

WHAT is machine learning

HOW can it help

Decisions

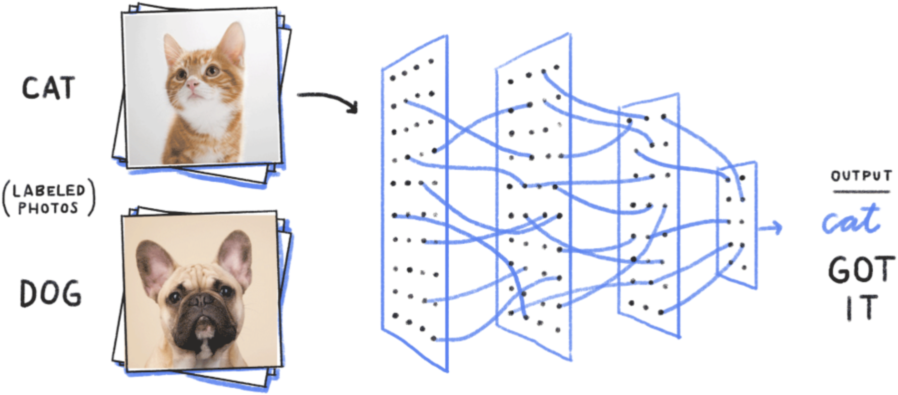


Figure: taken from www.becominghuman.ai

Decisions

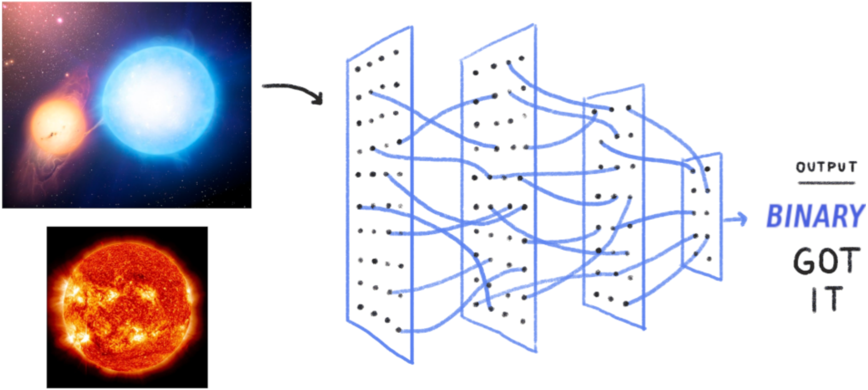
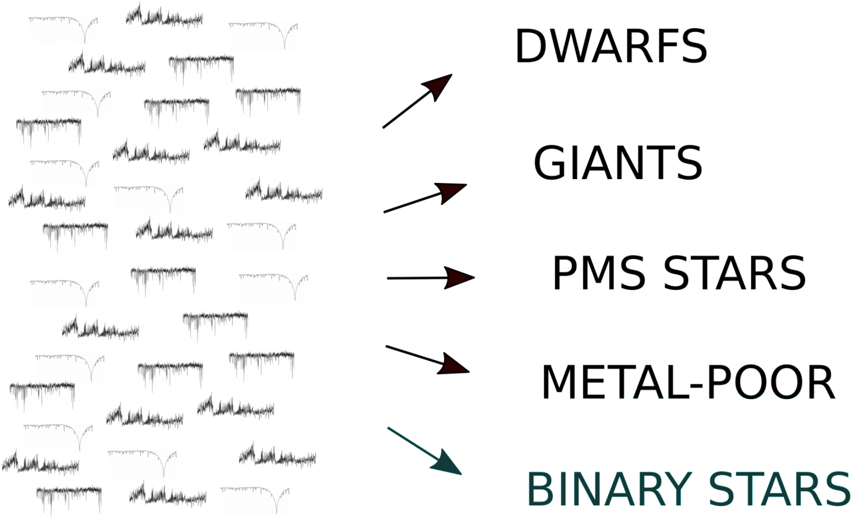


Figure: adapted from www.becominghuman.ai

Classification



Prediction

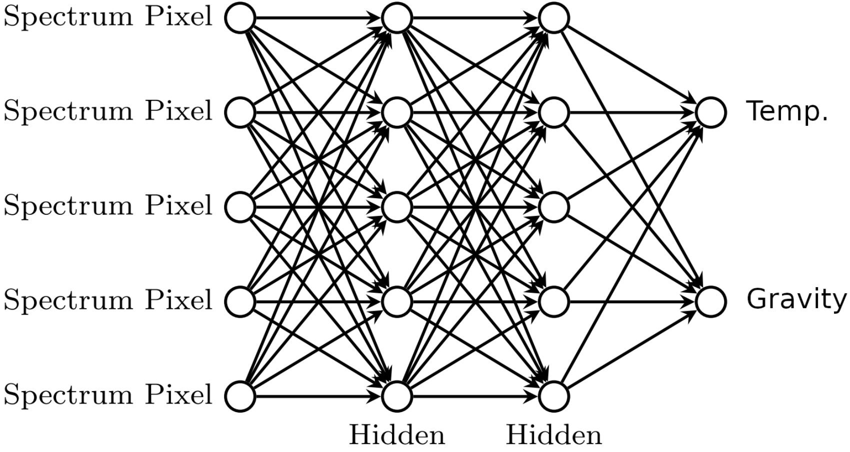


Figure: adapted from Leung & Bovy 2019

Machine learning scope

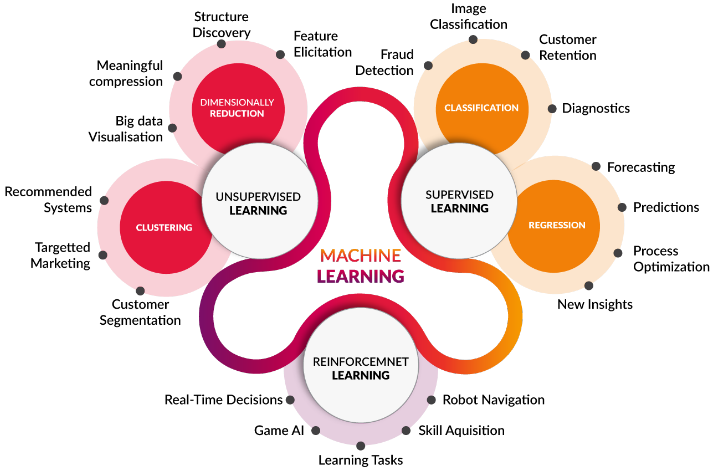


Figure: taken from www.cookieegroup.com

What is machine learning ?

Algorithm

What is machine learning ?

Algorithm

$+$

(training) Data

What is machine learning ?

Algorithm

+

(training) Data

=

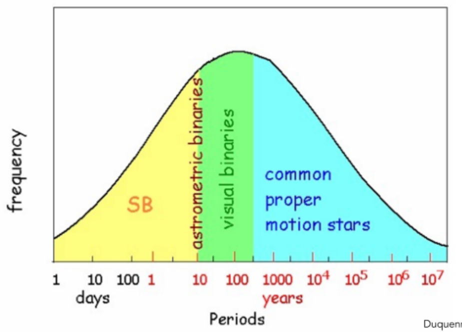
Decision maker, classifier, generative model, ...

WHAT is machine learning

HOW can it help

Machine learning for detection

F7-K dwarfs



Log-Normal distribution from 1 day to 10 million years

$\langle \log P_{days} \rangle = 4.8$

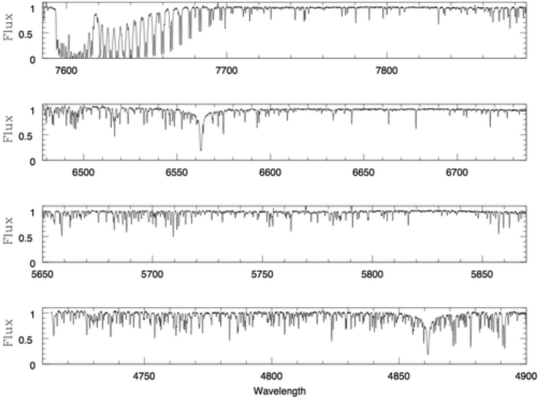
$\sigma_{\log P} = 2.3$

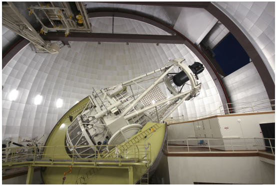
Duquennoy & Mayor 91
Halbwachs+ 10

common proper motio	6D phase space, chemical abundances
visual, resolved	imaging
astrometric	epoch astrometry (positions)
spectroscopic	doppler shift of spectral lines
photometric, eclipsing	variability in the light curve, eclipses

GALAH survey

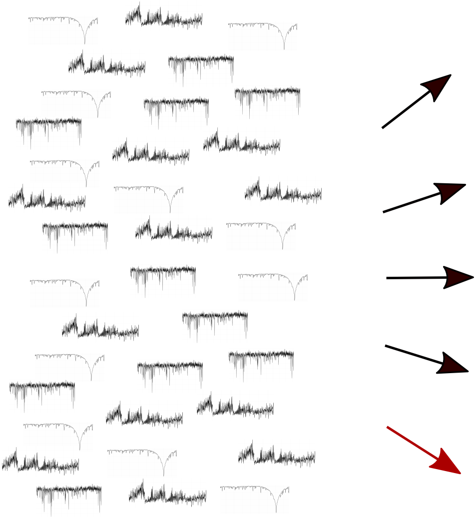
- Mission: Galactic archaeology
- 10^6 stars, magnitude range $12 < V < 14$
- $\sim$ 32 elemental abundances
- 471–490 565–587 648–674 758–789
- R $\sim$ 28 000 ($\Delta v_r \sim$ 15 km/s), SNR $\sim$ 100





Anglo-Australian Telescope (Coonabarabran, AU)

General classification - machine learning approach



DWARFS

GIANTS

PMS STARS

METAL-POOR

BINARY STARS

Dimensionality reduction $\Rightarrow$ Classification

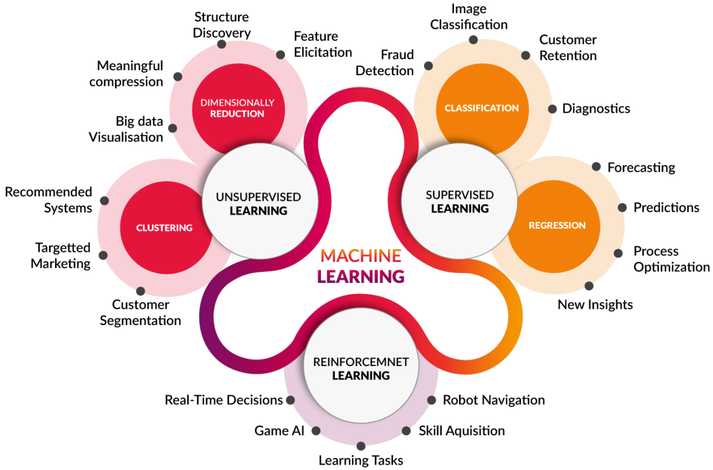


Figure: taken from www.cookieegroup.com

Dimensionality reduction - autoencoder (ANN)

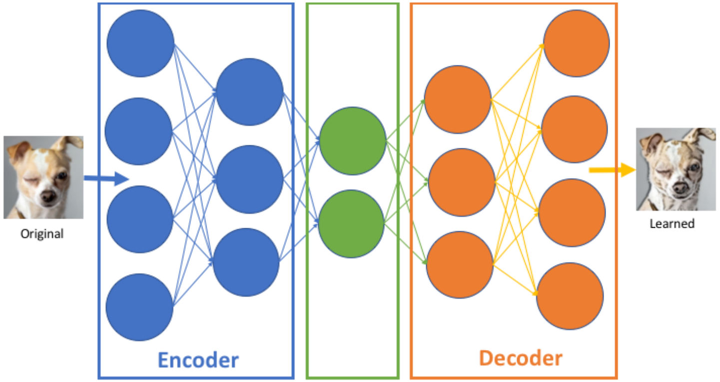


Figure: taken from www.blog.goodaudience.com

Dimensionality reduction - autoencoder (ANN)

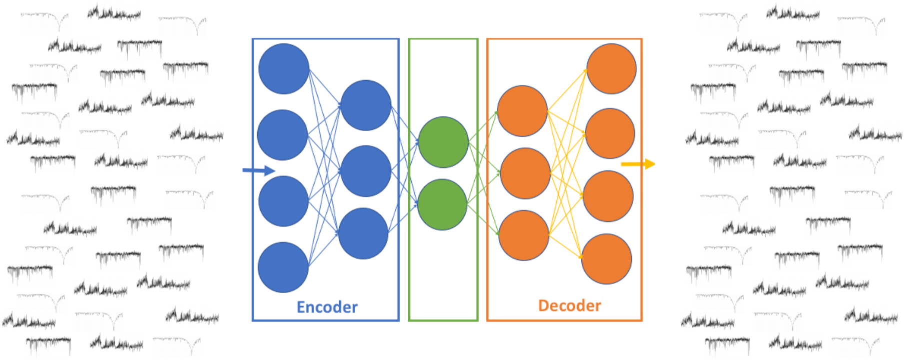


Figure: adapted from www.blog.goodaudience.com

Dimensionality reduction - autoencoder (ANN)

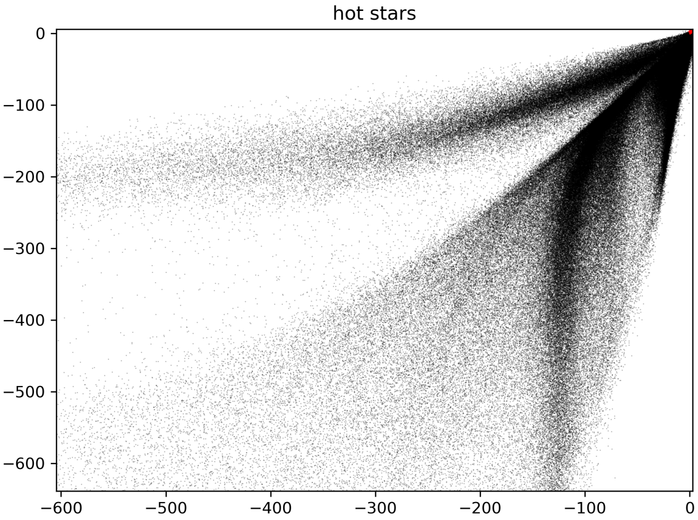


Figure: Autoencoder on GALAH spectra, provided by Klemen Čotar

Dimensionality reduction - autoencoder (ANN)

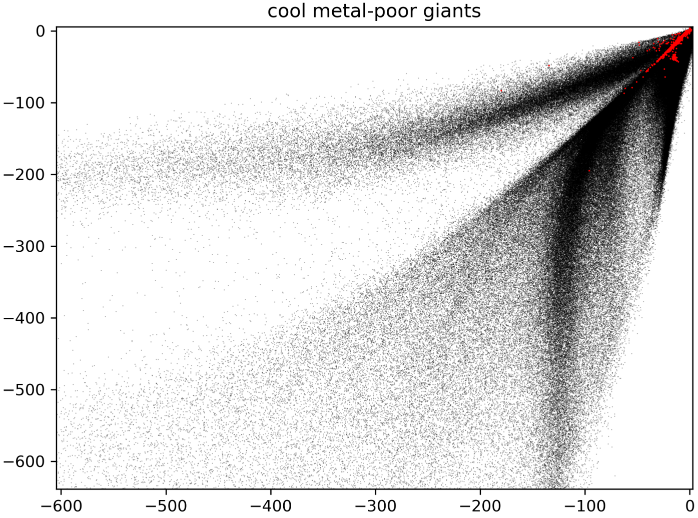


Figure: Autoencoder on GALAH spectra, provided by Klemen Čotar

Dimensionality reduction - autoencoder (ANN)

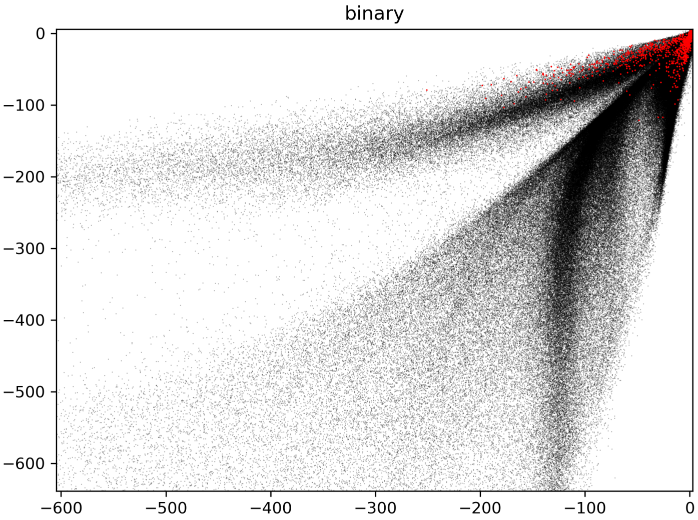


Figure: Autoencoder on GALAH spectra, provided by Klemen Čotar

Dimensionality reduction - autoencoder (ANN)

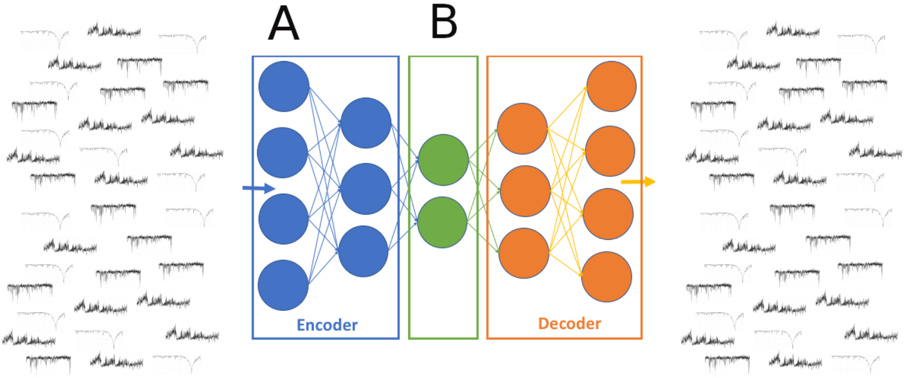
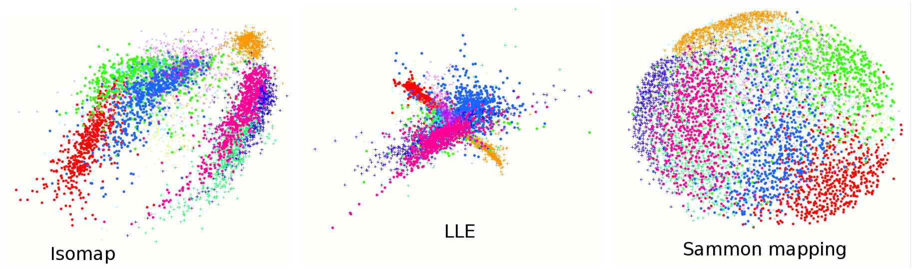
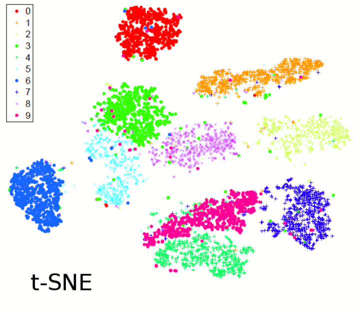
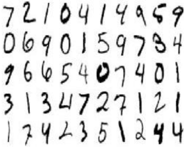


Figure: adapted from www.blog.goodaudience.com

t–distributed Stochastic Neighbour Embedding

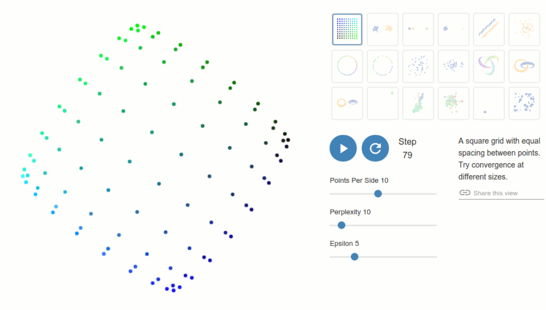






Learning about t–SNE

https://distill.pub/2016/misread-tsne

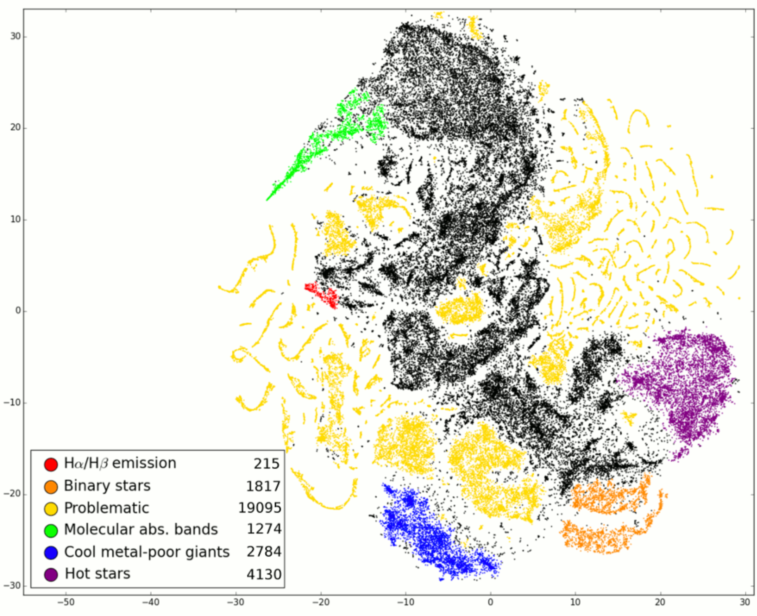


SEE ALSO

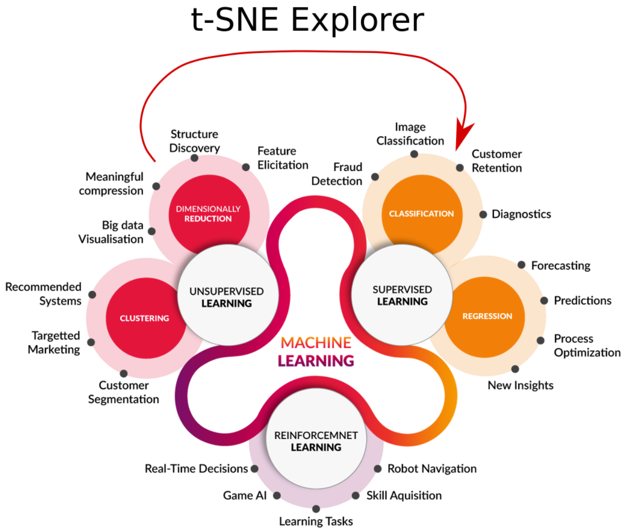
Kos et al. 2018

and the original publications by **Van Der Matten et al.**

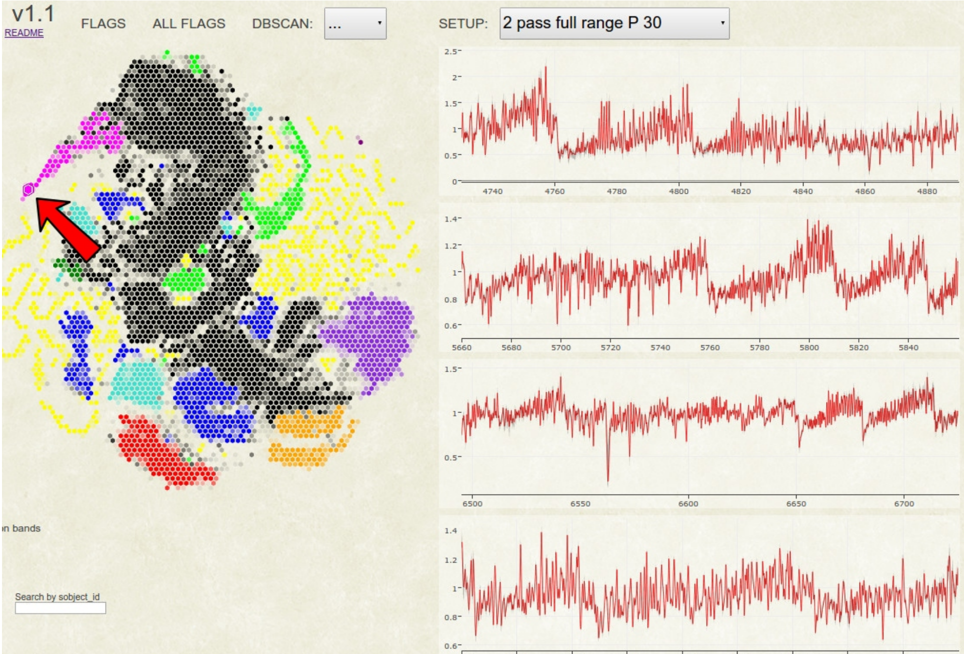
t-SNE map of ~ 80k GALAH spectra



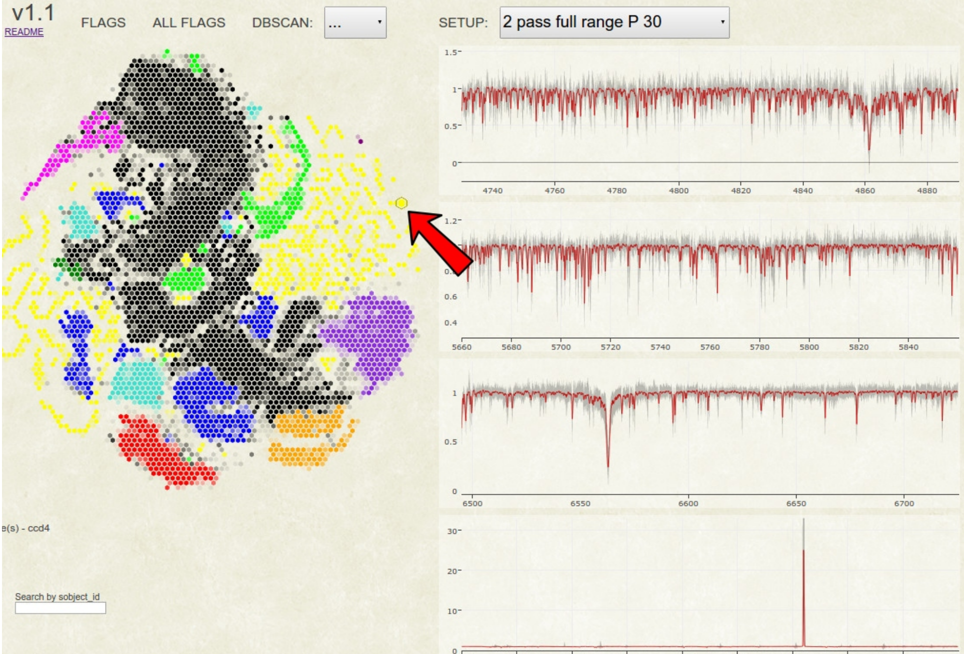
Dimensionality reduction ⇒ Classification



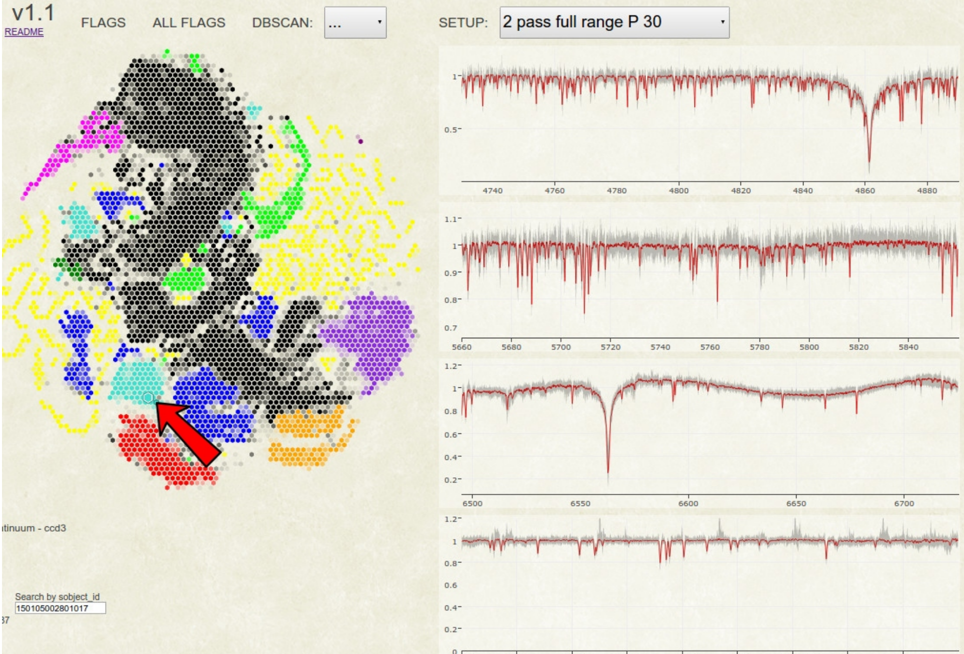
Molecular absorption bands



Problematic spectra

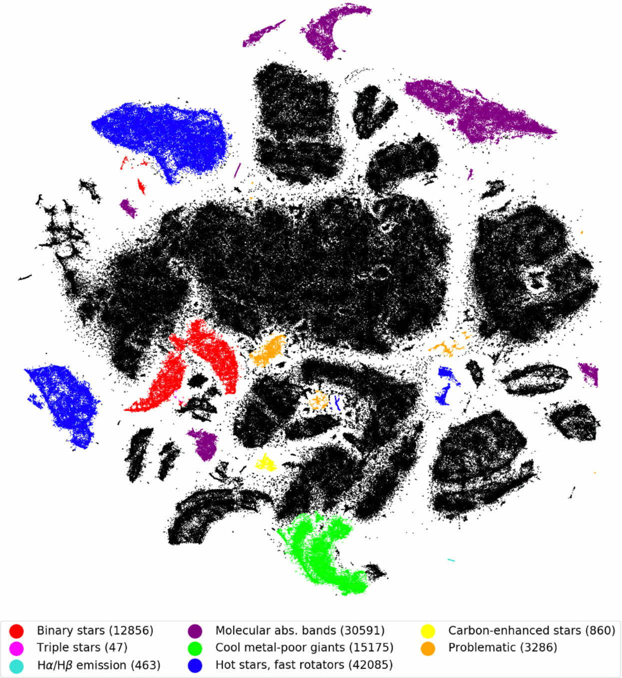


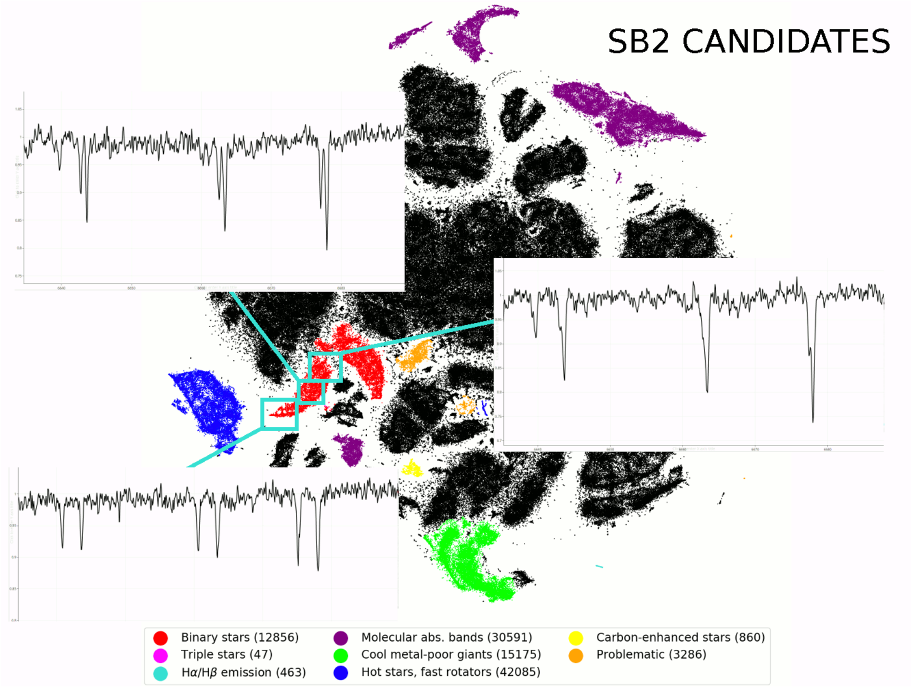
Oscillating continuum



t-SNE classification Galah DR2

Buder et al. 2018




SB2 CANDIDATES
Binary stars (12856)
Triple stars (47)
Hα/Hβ emission (463)
Molecular abs. bands (30591)
Cool metal-poor giants (15175)
Hot stars, fast rotators (42085)
Carbon-enhanced stars (860)
Problematic (3286)

Finding SB2(3,4) - conventional approach

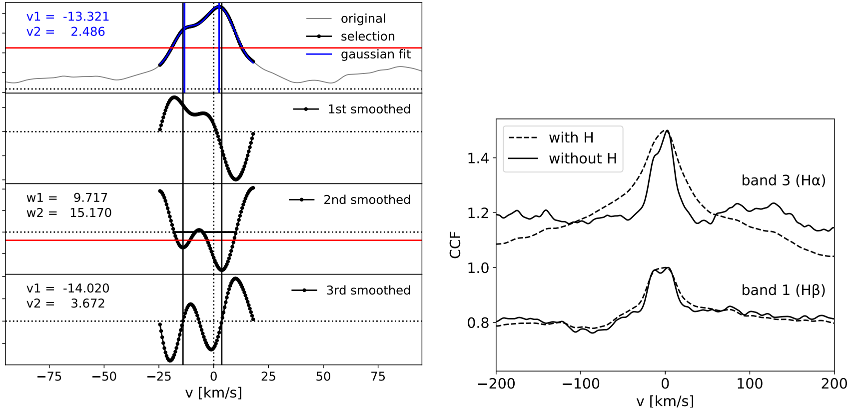


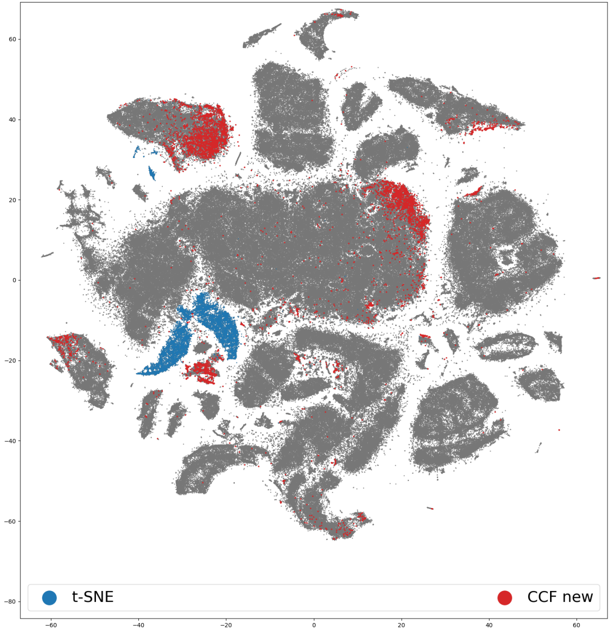
Figure: CCF detection method (Merle et al. 2017)

SB2 detection: t–SNE vs CCF

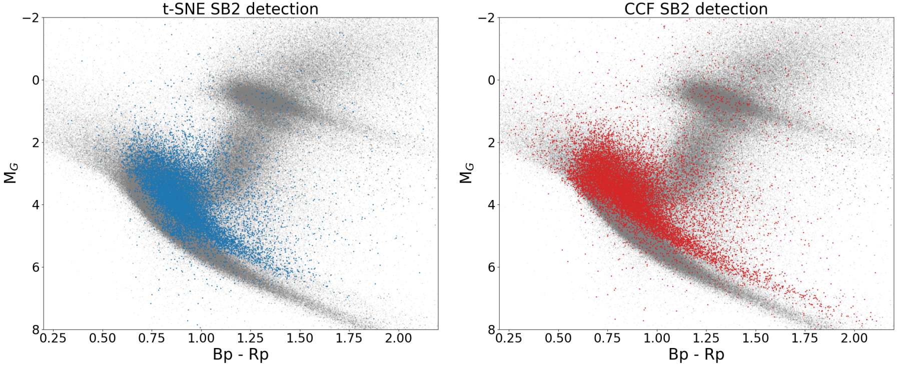
SB2 candidates from different detection methods

1. method: **t-SNE**

2. method: **CCF** (Cross Correlation Function; Merle et al. 2017)



SB2 detection: t–SNE vs CCF



Limits of t–SNE detection

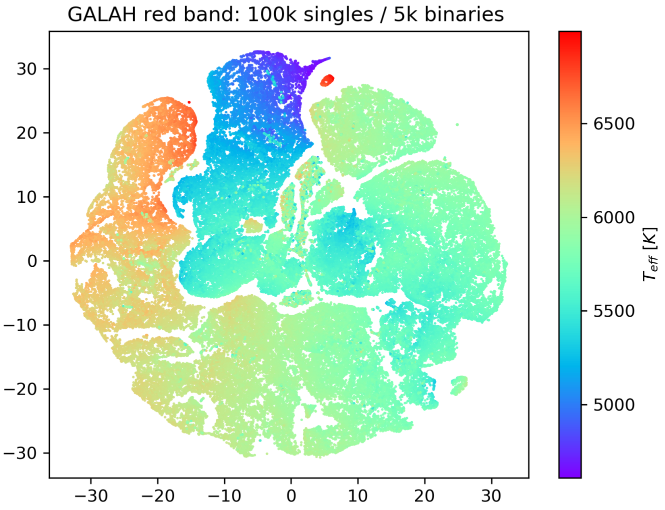


Figure: Synthetic single + binary stars based on GALAH parameters for dwarfs, provided by Pablo Navarro Barrachina

Limits of t–SNE detection

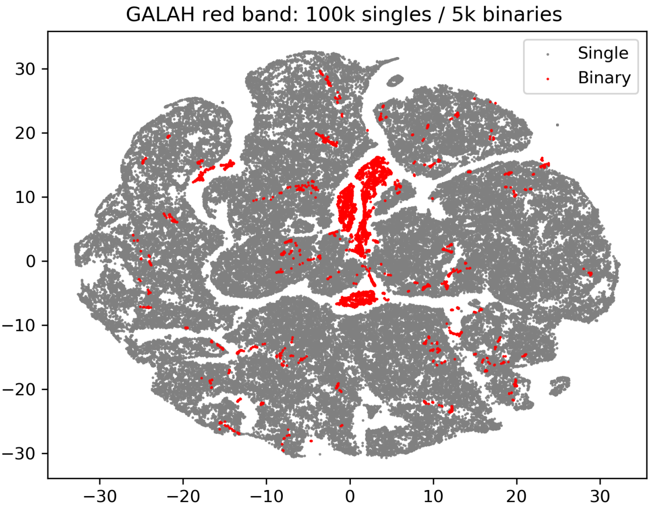


Figure: Synthetic single + binary stars based on GALAH parameters for dwarfs, provided by Pablo Navarro Barrachina

Limits of t–SNE detection

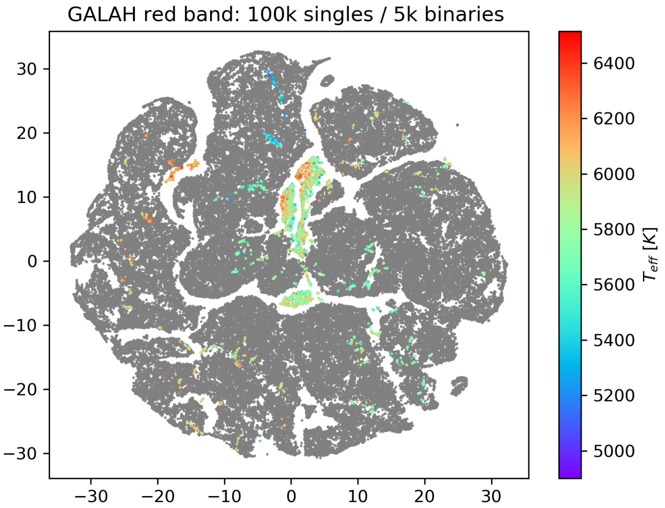


Figure: Synthetic single + binary stars based on GALAH parameters for dwarfs, provided by Pablo Navarro Barrachina

Analysis ⇒ detection (using a generative model)

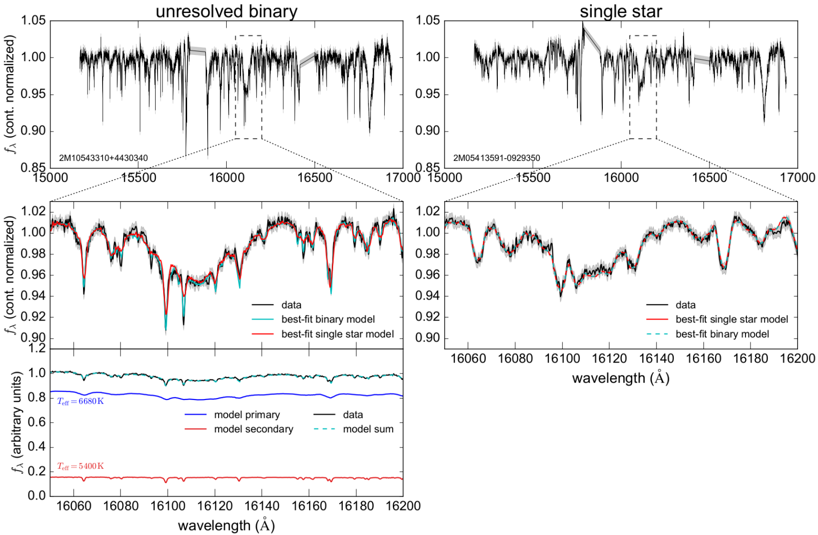


Figure: Binary star detection in APOGEE, El-Badry et al. 2018

Summary

- Population statistics of multiple stars are in high demand

Science of large samples

Astro2020 Science White Paper (Breivik+ 2019)

Stellar multiplicity: an interdisciplinary nexus

Thematic Areas: ☑ Planetary Systems ☑ Star and Planet Formation
☑ Formation and Evolution of Compact Objects ☑ Cosmology and Fundamental Physics
☑ Stars and Stellar Evolution ☑ Resolved Stellar Populations and their Environments
☑ Galaxy Evolution ☑ Multi-Messenger Astronomy and Astrophysics

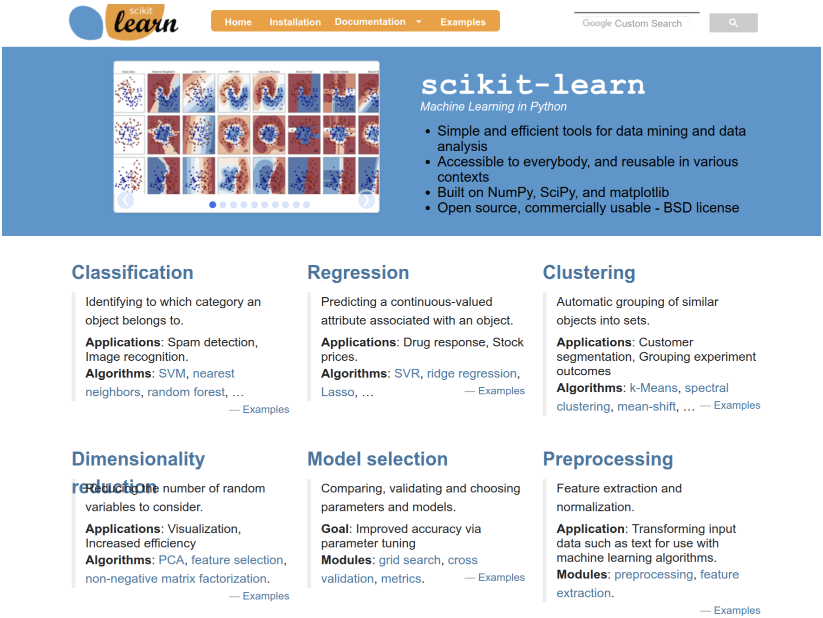
We need to understand the population statistics of stellar multiplicity and their variations with stellar type, chemistry, and dynamical environment"

- stellar multiplicity - direct outcome of star formation
- stellar populations - consequence of stellar and binary evolution
- high-redshift galaxy radiation and reionization - binary-dependent stellar physics
- multi-messenger astronomy and compact objects – the outcomes of binary evolution
- Hubble constant (Ia supernovae, GW mergers) – binary star progenitors
- dark-matter substructure masses – distorted by binary populations
- exoplanet experiments - unknown multiple star contamination

Summary

- Population statistics of multiple stars are in high demand
- ML for various tasks (e.g. detection, generative models, feature extraction)

Try it out !


scikit learn
Home
Installation
Documentation
Examples
Google Custom Search
scikit-learn
Machine Learning in Python
• Simple and efficient tools for data mining and data analysis
• Accessible to everybody, and reusable in various contexts
• Built on NumPy, SciPy, and matplotlib
• Open source, commercially usable - BSD license
Classification
Identifying to which category an object belongs to.
Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, …
— Examples
Regression
Predicting a continuous-valued attribute associated with an object.
Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, …
— Examples
Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, …
— Examples
Dimensionality reduction
Reducing the number of random variables to consider.
Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization.
— Examples
Model selection
Comparing, validating and choosing parameters and models.
Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics.
— Examples
Preprocessing
Feature extraction and normalization.
Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction.
— Examples

Try it out !

Home Installation Documentation Examples

Google Custom Search

Previous
Glossary of C...

Next
Compact estim...

scikit-learn v0.21.2
Other versions

Please **cite us** if you use the software.

Examples
Miscellaneous examples
Examples based on real world datasets
Biclustering
Calibration
Classification
Clustering
Pipelines and composite estimators
Covariance estimation
Cross decomposition
Dataset examples
Decomposition
Ensemble methods
Tutorial exercises
Feature Selection
Gaussian Process for Machine Learning
Missing Value Imputation
Inspection
Generalized Linear Models
Manifold learning
Gaussian Mixture Models

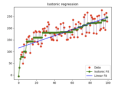
«

Examples

Miscellaneous examples

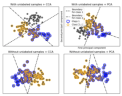
Miscellaneous and introductory examples for scikit-learn.

Compact estimator representations

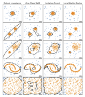


Isotonic Regression

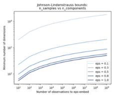
Face completion with a multi-output estimators



Multilabel classification



Comparing anomaly detection algorithms for outlier detection



The Johnson-Lindenstrauss bound for embedding with

Try it out !

astroML | Home | User Guide | Book Figures | Examples Plots | Google Custom Search

AstroML: Machine Learning and Data Mining for Astronomy

News

January 2014: the textbook accompanying astroML is now available! View it on Amazon.

November 2013: astroML 0.2 has been released! Get the source on Github

Our Introduction to astroML paper received the CIDU 2012 best paper award.

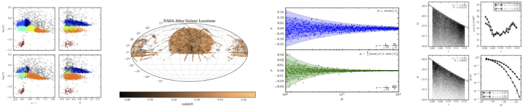
Links

astroML Mailing List

GitHub Issue Tracker

Videos

Scipy 2012 (15 minute



AstroML is a Python module for machine learning and data mining built on numpy, scipy, scikit-learn, matplotlib, and astropy, and distributed under the 3-clause BSD license. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in Python, loaders for several open astronomical datasets, and a large suite of examples of analyzing and visualizing astronomical datasets.

Downloads

- Released Versions: Python Package Index
- Bleeding-edge Source: github

The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics, to provide a uniform and easy-to-use interface to freely available astronomical datasets. We hope this package will be useful to researchers and students of astronomy. If you have an example you'd like to share, we are happy to accept a contribution via a GitHub Pull Request: the code repository can be found at http://github.com/astroML/astroML.

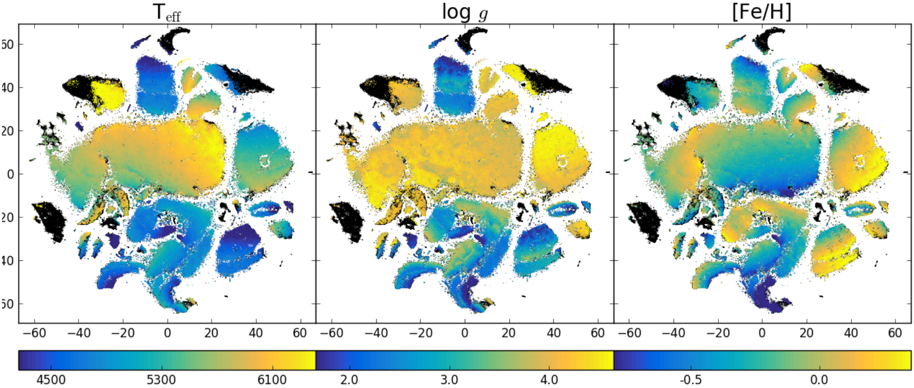
Summary

- Population statistics of multiple stars are in high demand
- ML for various tasks (e.g. detection, generative models, feature extraction)
- Smart combination of conventional and ML techniques

Summary

- Population statistics of multiple stars are in high demand
- ML for various tasks (e.g. detection, generative models, feature extraction)
- Smart combination of conventional and ML techniques
- Currently far from A.I., human interaction with ML essential

New Galah dataset ($\sim$ 600k spectra)



local and **global** structure of the data in a single map

Finding SB1 - conventional approach

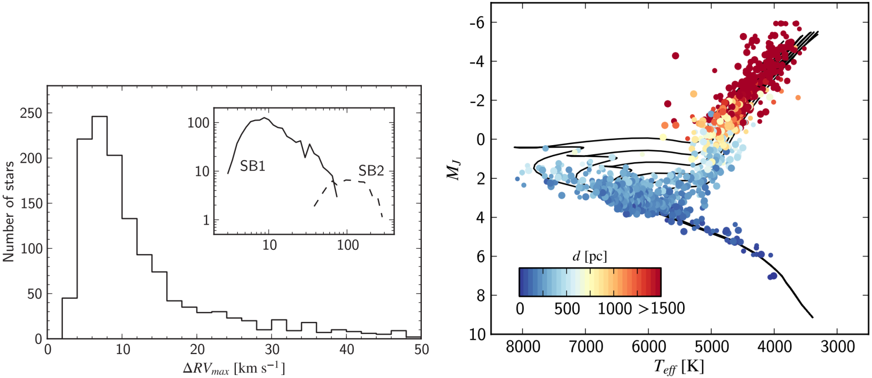


Figure: Detection of SB1s through RV variability (Matijevič et al. 2011)

Machine learning for analysis

$T_1,\ T_2, \log g_1, \log g_2, [\mathrm{Fe/H}], \triangle v_r, \ldots$

ML for the spectroscopic model - $\mathcal{M}_{spec,i}(\theta)$

Model atmospheres + spectral synthesis = synthetic templates

ML for the spectroscopic model - $\mathcal{M}_{spec,i}(\theta)$

Model atmospheres + spectral synthesis = synthetic templates

OR

Interpolation of observed spectra = data-driven generative model

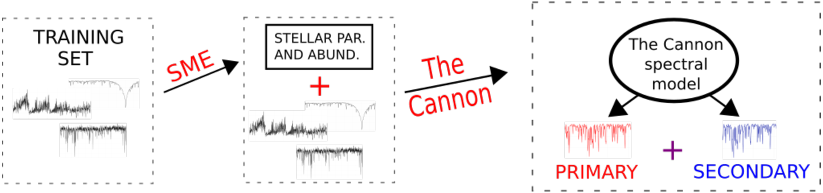
ML for the spectroscopic model - $\mathcal{M}_{spec,i}(\theta)$

Model atmospheres + spectral synthesis = synthetic templates

OR

Interpolation of observed spectra = data-driven generative model

(majority of spectral lines accounted for, effects of the instrument embedded automatically, identical resolution, directly determine e.g. mass, age - unknown how they affect spectra)



SME - Spectroscopy Made Easy by Piskunov & Valenti (2016)

The Cannon by Ness et al. (2015)

$\mathcal{M}_{spec}(\theta)$ by The Cannon

$$\mathit{flux}_{n,\lambda} = \Omega_\lambda^T \cdot l_n + \mathrm{noise}$$

$$l_n = f(\theta) = (T_{eff,n}, \log g_n, [\mathrm{Fe/H}]_n, T_{eff,n}^2, ...)$$

$\mathcal{M}_{spec}(\theta)$ by The Cannon

$$flux_{n,\lambda} = \Omega_\lambda^T \cdot l_n + \mathrm{noise}$$

$$l_n = f(\theta) = (T_{eff,n}, \log g_n, [\mathrm{Fe/H}]_n, T_{eff,n}^2, ...)$$

generative model for observed stellar spectra:

$$\mathcal{M}_{spec,\lambda}(\theta) = \Omega_\lambda^T \cdot l_n$$

$\mathcal{M}_{spec}(\theta)$ by The Payne

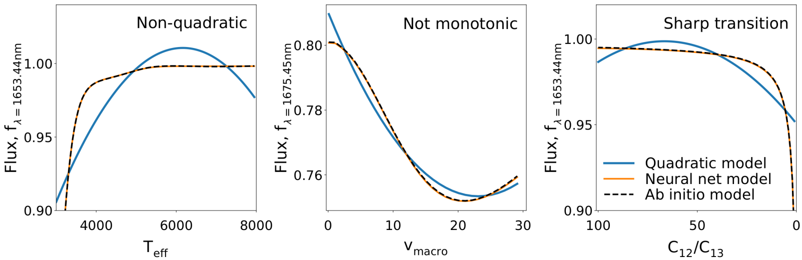


Figure: taken from Ting+ 2018

$\mathcal{M}_{spec}(\theta)$ by The Payne

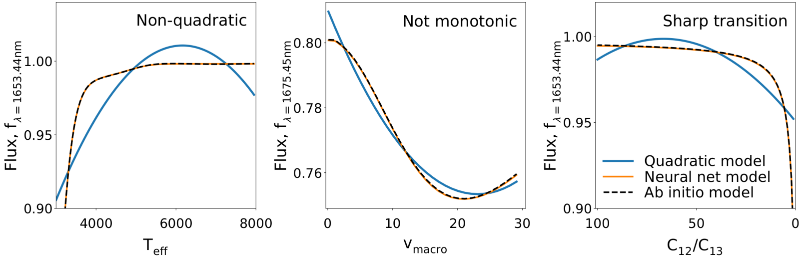


Figure: taken from Ting+ 2018

generative model for observed stellar spectra:

$$\mathcal{M}_{spec,\lambda}(\theta) = F_\lambda(\textit{coefficients}, \theta)$$

$\mathcal{M}_{spec}(\theta)$ by neural networks

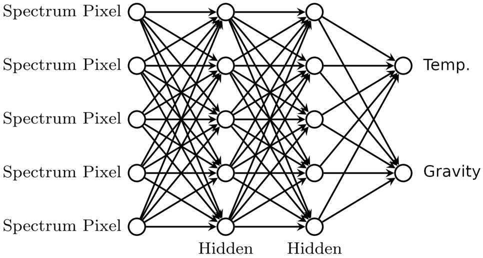


Figure: adapted from Leung & Bovy 2019

$\mathcal{M}_{spec}(\theta)$ by neural networks

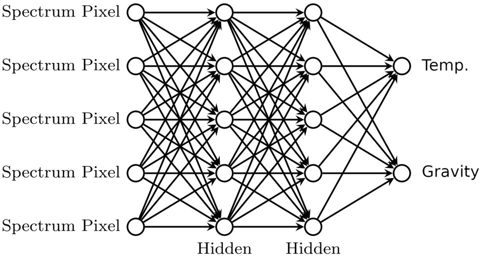


Figure: adapted from Leung & Bovy 2019

generative model for observed stellar spectra:

$$\mathcal{M}_{spec,\lambda}(\theta) = ?$$

$\mathcal{M}_{spec}(\theta)$ by Neural networks

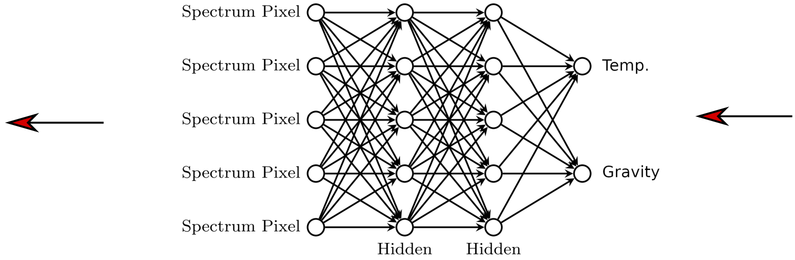
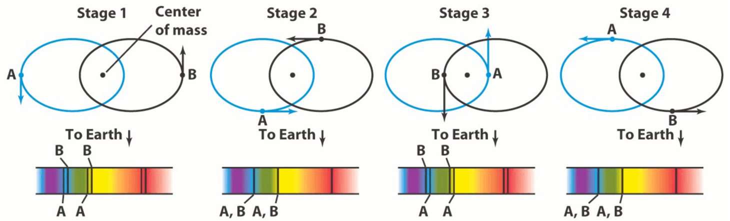


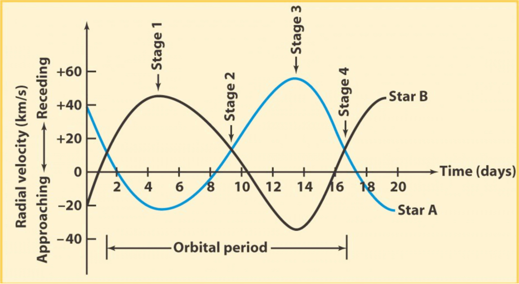
Figure: adapted from Leung & Bovy 2019

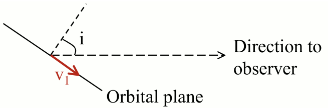
generative model for observed stellar spectra:

$$\mathcal{M}_{spec,\lambda}(\theta) = ?$$

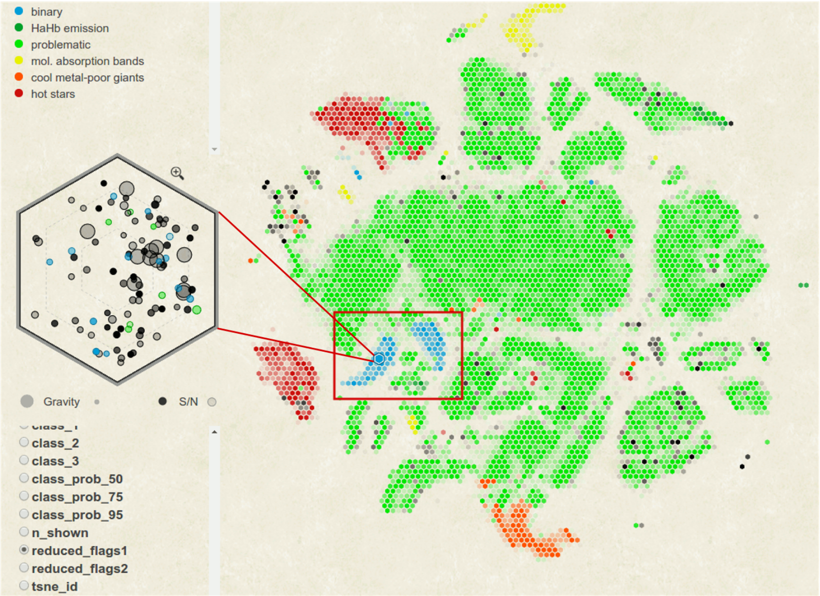
Finding binary stars in data







Updating classification



Dimensionality reduction - t–SNE

t-SNE objective: minimize divergence between pairwise similarities $p_{j|i}$ and q_{ij} of data points in original space A and in projection space B

1. Euclidean distances in original space A $\Rightarrow$ pairwise similarities ($p_{j|i}$)

$$p_{j|i} = \frac{\exp(-\|a_i - a_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|a_i - a_k\|^2/2\sigma_i^2)}, \qquad p_{i|i} = 0, \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

$$\sigma_i \to P_i, \quad Perp(P_i) = 2^{H(P_i)}, \quad H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

2. Pairwise similarities in projection space B with heavy-tailed Student-t

$$q_{ij} = \frac{(1 + \|b_i - b_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|b_k - b_l\|^2)^{-1}}$$

3. Minimize the Kullback-Leibler divergence between the two distributions

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Generative model for spectra

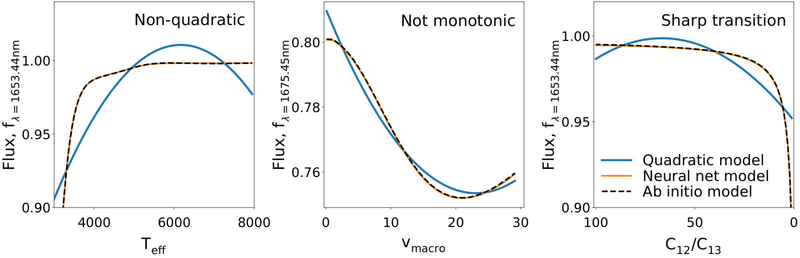


Figure: taken from Ting et al. 2018

The Cannon (Ness et al. 2015)

The Payne (Ting et al. 2018)

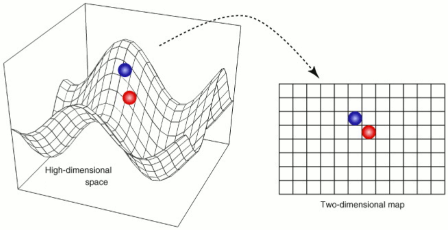
...

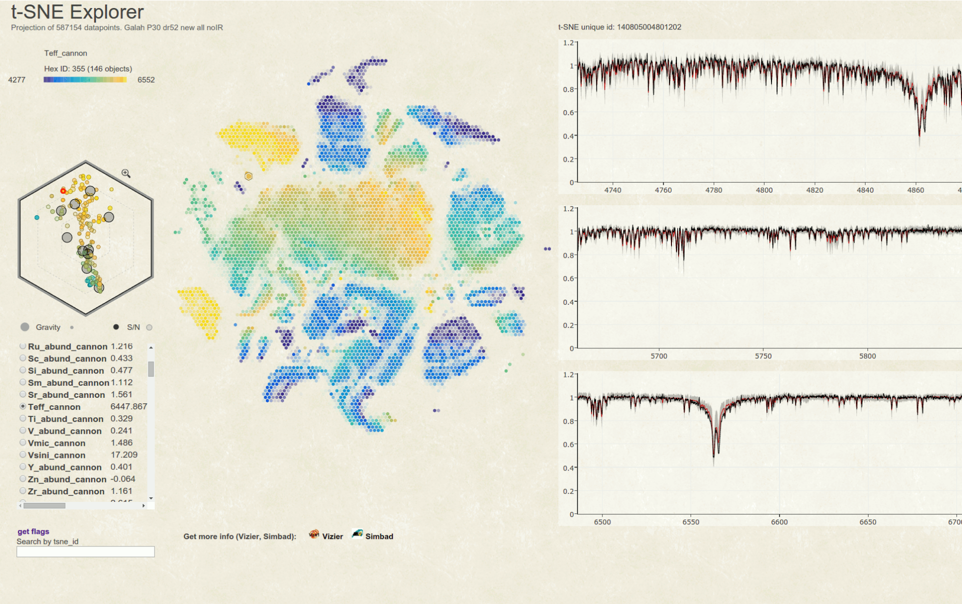
Classification - **dimensionality reduction**

"Essentially, all models are wrong but some are useful"

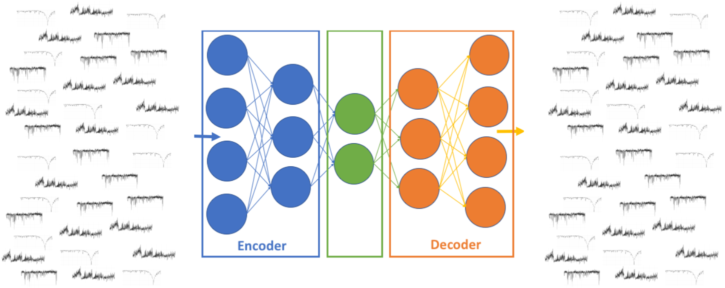
George E.P. Box

- High dimensional (pixel) space A $\Rightarrow$ low dimensional space B (map)
- Dim. reduction $\Rightarrow$ **information loss**
- Preserve **important information** $\Rightarrow$ intrinsic dimensionality of the spectra (Teff, elemental abundances, chromospheric emission, etc.)
- Projection should retain **structure** of some low-D manifold on which our datapoints lie




t-SNE Explorer
Projection of 587154 datapoints. Galah P30 dr52 new all noIR
Teff_cannon
Hex ID: 355 (146 objects)
4277
6552
Gravity
S/N
Ru_abund_cannon 1.216
Sc_abund_cannon 0.433
Si_abund_cannon 0.477
Sm_abund_cannon 1.112
Sr_abund_cannon 1.561
Teff_cannon 6447.867
Ti_abund_cannon 0.329
V_abund_cannon 0.241
Vmic_cannon 1.486
Vsini_cannon 17.209
Y_abund_cannon 0.401
Zn_abund_cannon -0.064
Zr_abund_cannon 1.161
get flags
Search by tsne_id
Get more info (Vizier, Simbad):
Vizier
Simbad
t-SNE unique id: 140805004801202
4740
4760
4780
4800
4820
4840
4860
5700
5750
5800
6500
6550
6600
6650
1.2
1
0.8
0.6
0.4
0.2
0

autoencoder (feature extraction) + t–SNE



+

t–SNE

autoencoder (feature extraction) + t–SNE

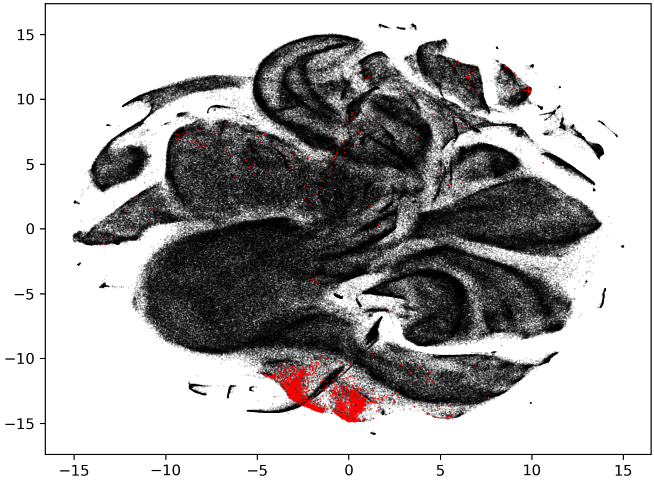
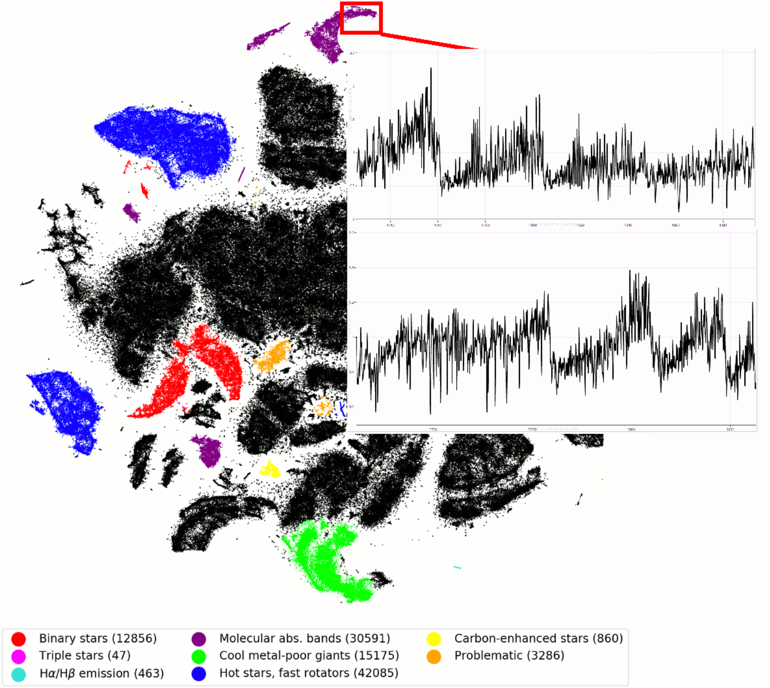
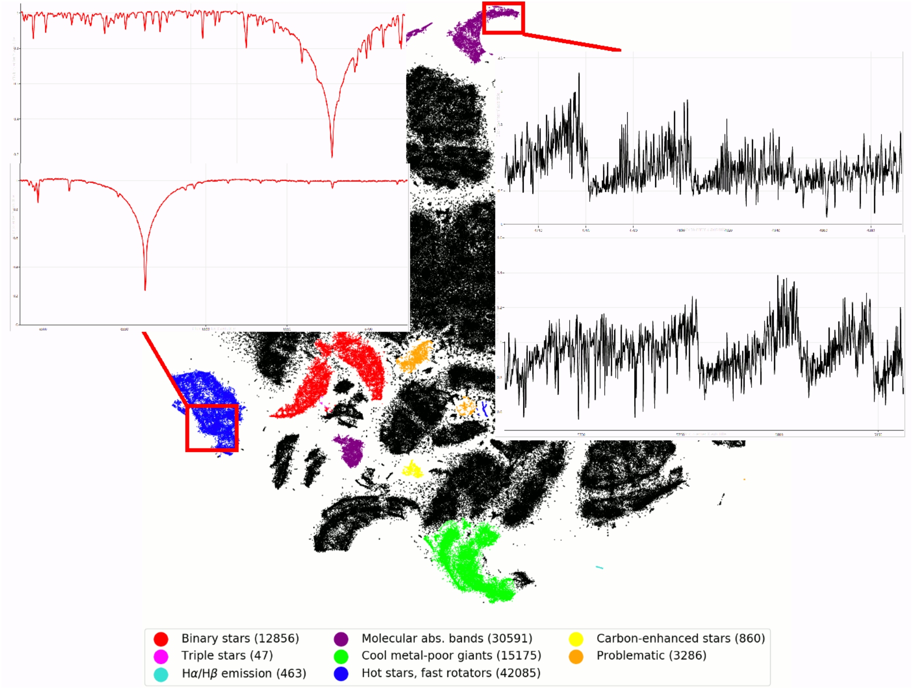
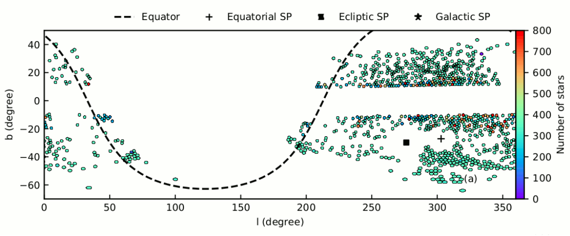


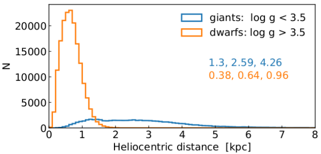
Figure: Autoencoder (100 neurons middle layer) + t–SNE on GALAH spectra, binary stars in red, provided by Klemen Čotar

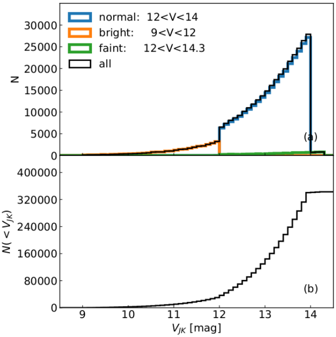

Binary stars (12856)
Triple stars (47)
Hα/Hβ emission (463)
Molecular abs. bands (30591)
Cool metal-poor giants (15175)
Hot stars, fast rotators (42085)
Carbon-enhanced stars (860)
Problematic (3286)

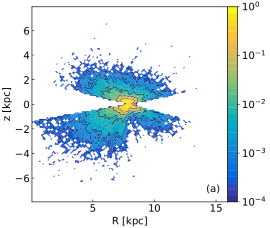

Binary stars (12856)
Triple stars (47)
Hα/Hβ emission (463)
Molecular abs. bands (30591)
Cool metal-poor giants (15175)
Hot stars, fast rotators (42085)
Carbon-enhanced stars (860)
Problematic (3286)

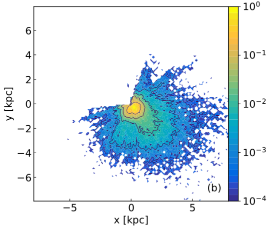
Galah survey - DR2 (Buder et al. 2018)





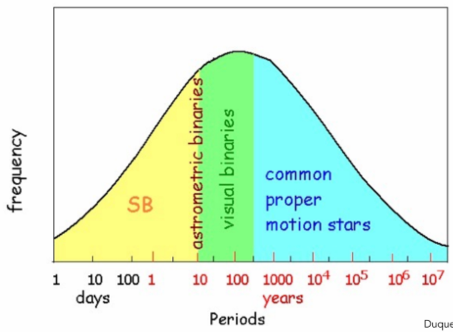






Machine learning for detection

F7-K dwarfs



Log-Normal distribution from 1 day to 10 million years

$\langle \log P_{days} \rangle = 4.8$

$\sigma_{\log P} = 2.3$

Duquennoy & Mayor 91
Halbwachs+ 10

visual, resolved	imaging
common proper motion	6D phase space, chemical abundances
astrometric	epoch astrometry (positions)
spectroscopic	doppler shift of spectral lines
photometric, eclipsing	variability in the light curve, eclipses

Science of large samples

The binary fraction (f_b, $f_{b,0}$)

Science of large samples

The binary fraction (f_b, $f_{b,0}$)

Initial distributions of

P (period), q (M_2/M_1), e (eccentricity), Age

Science of large samples

The binary fraction (f_b, $f_{b,0}$)

Initial distributions of

P (period), q (M_2/M_1), e (eccentricity), Age

$\Uparrow$

Observed properties of a binary population

(T_1, T_2, R_1, R_2, $\log g_1$, $\log g_2$, [Fe/H], $\triangle v_r$)

Science of large samples

The binary fraction (f_b, $f_{b,0}$)

Initial distributions of

P (period), q (M_2/M_1), e (eccentricity), Age

$\Uparrow$

Observed properties of a binary population
(T_1, T_2, R_1, R_2, $\log g_1$, $\log g_2$, [Fe/H], $\triangle v_r$)

But first: DETECTION and ANALYSIS

Example applications

A&A 602, A110 (2017)

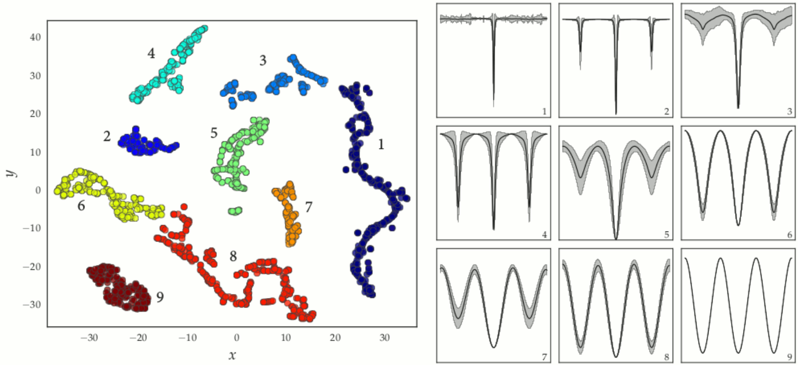


Fig. 8. *Left panel*: *t*-SNE+DBSCAN of filtered *Gaia*-sampled *Kepler* data set fitted with the *two-Gaussian* model. *Right panel*: mean of all normalized light curves in each DBSCAN class; gray shading indicates the region $[-\sigma, \sigma]$ around the computed mean at a given orbital phase.

Figure: Kochoska et al. 2017

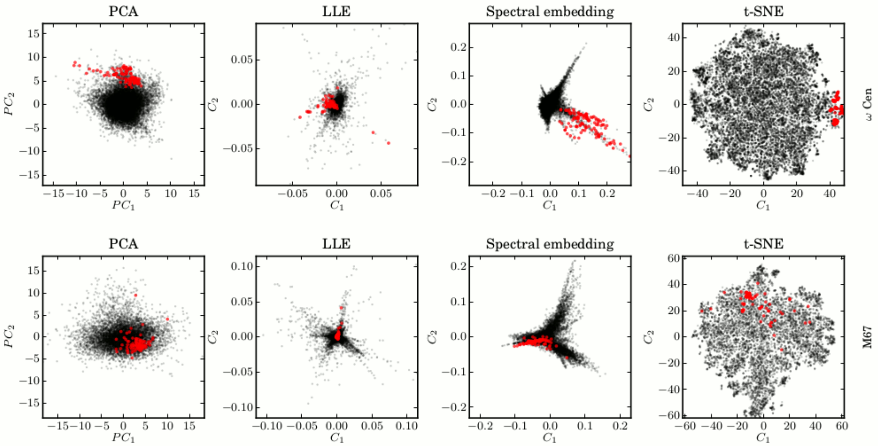


Figure A2. Comparison of the performance of PCA, Locally linear embedding (LLE), Spectral embedding and t-SNE. Methods follow from the most to the least linear (left to right). In the top row we compare the four methods on the case of ω Cen (an easy case) and in the bottom row for M67 (a harder case). Known cluster members are marked in red. Notice how efficiently t-SNE covers the plane and how many more distinct groups one can see.

Figure: Kos et al. 2018