Machine Learning in Astronomy

Reza Monadi

UC Riverside

May 14, 2020

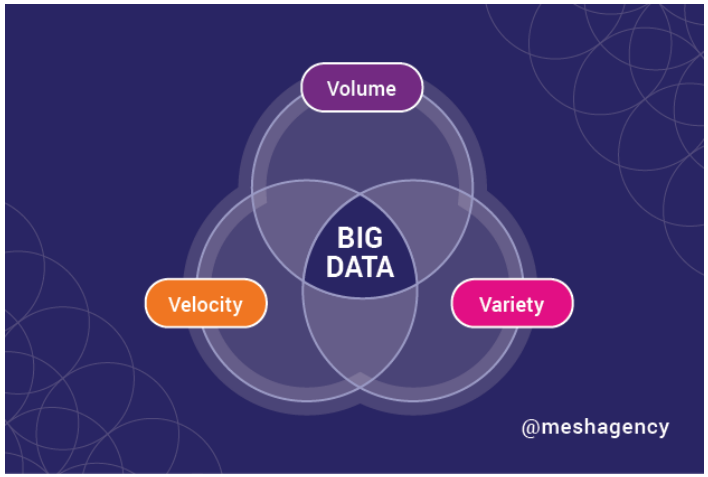
- Astronomy in **Big Data** era

- Astronomy in **Big Data** era
- Usage of Supervised **ML** in astronomy

- Astronomy in **Big Data** era
- Usage of Supervised **ML** in astronomy
- Implementation of unsupervised **ML** in astronomy

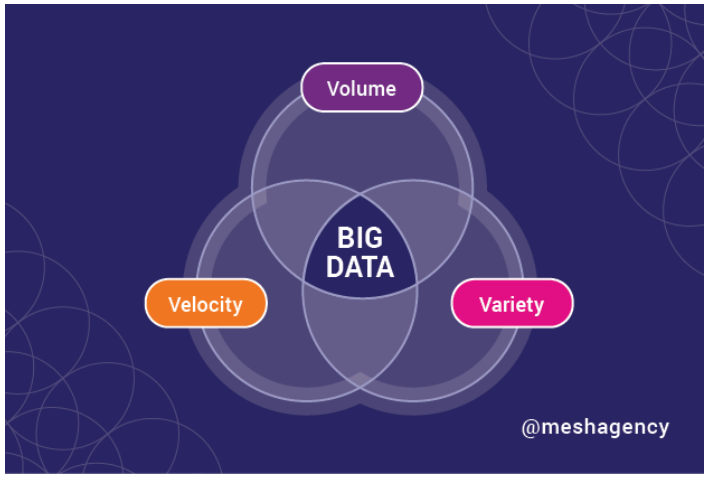
- Astronomy in **Big Data** era
- Usage of Supervised **ML** in astronomy
- Implementation of unsupervised **ML** in astronomy
- Limitations of **ML** in astronomy

3Vs in astronomy



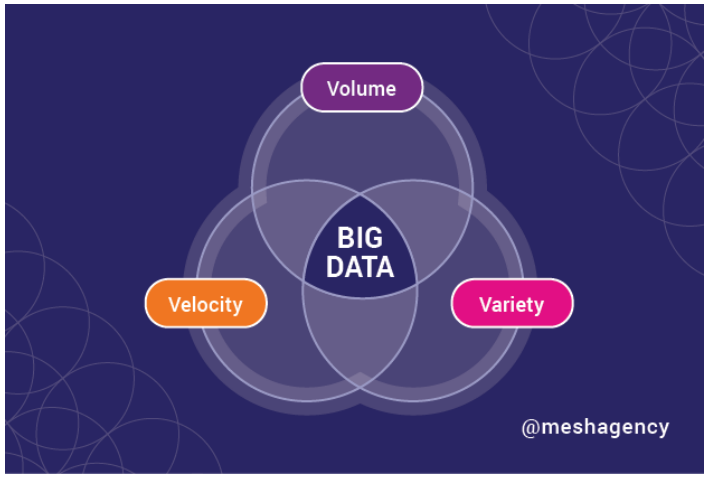
- Volume: larger quantities of data by the advent of better telescopes and surveys

3Vs in astronomy



- Volume: larger quantities of data by the advent of better telescopes and surveys
- Velocity: Higher speed of incoming observational data

3Vs in astronomy

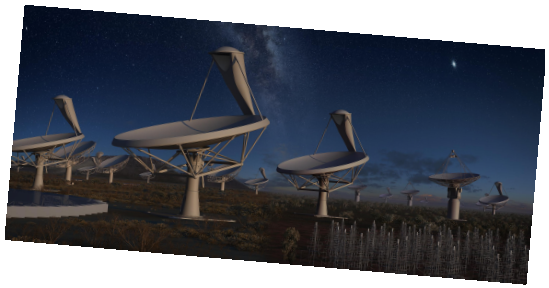


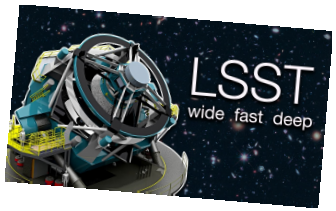
- Volume: larger quantities of data by the advent of better telescopes and surveys
- Velocity: Higher speed of incoming observational data
- Variety: multi-wavelength spectroscopic and photometeric data from versatile astronomical objects

Astronomical surveys are exponentially growing

Sky Server Prokects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	40 PB
LSST (The Large Synoptic Survey Telescope)	200 PB
SKA (The Square Kilometer Array)	4.6 EB expected

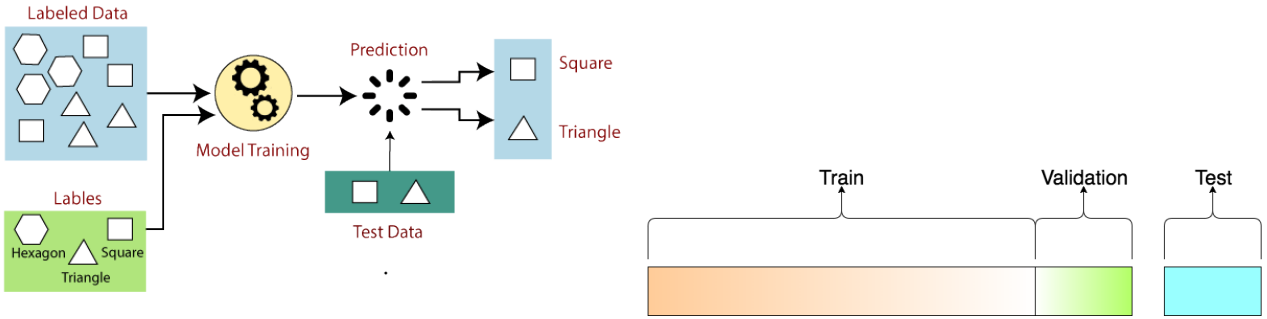
Zhang et. al 2015

SKA 

Rubin's Observatory 

Large surveys = Big Data
Big Data + ML = Astronomical Knowledge

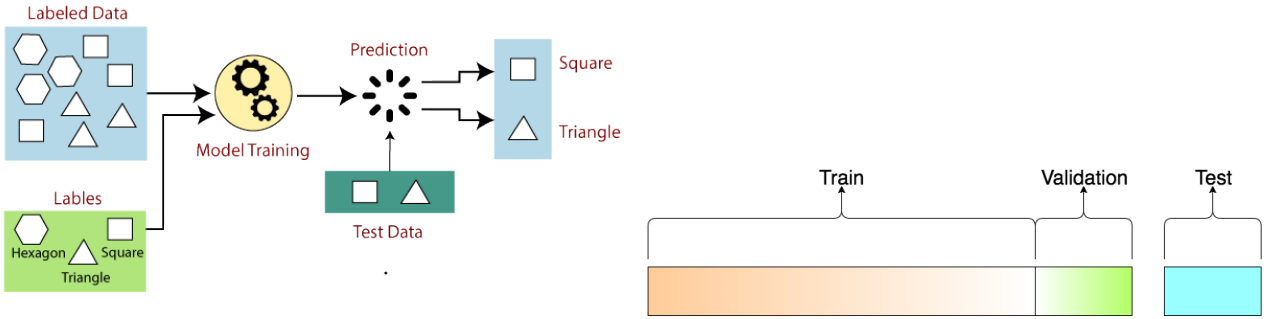
Stages of Supervised Learning



javatpoint.com

- Training:

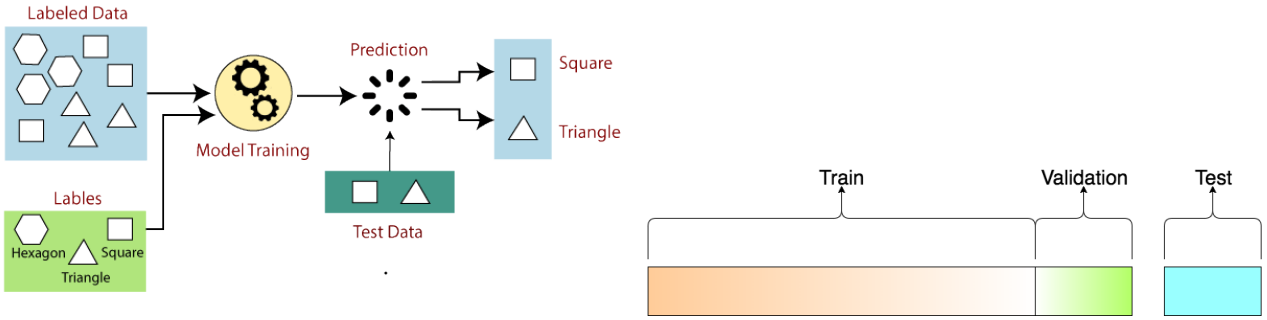
Stages of Supervised Learning



javatpoint.com

- Training:
  1. Select a model

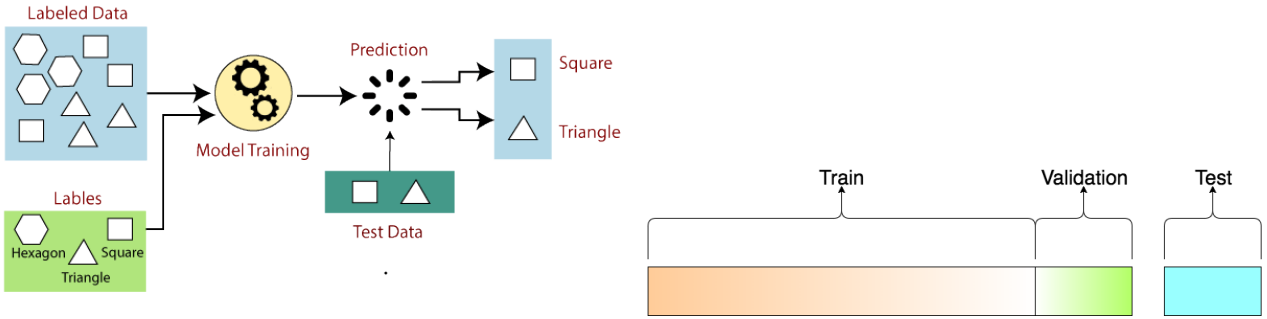
Stages of Supervised Learning



javatpoint.com

- Training:
  1. Select a model
  2. Train the model by the training set

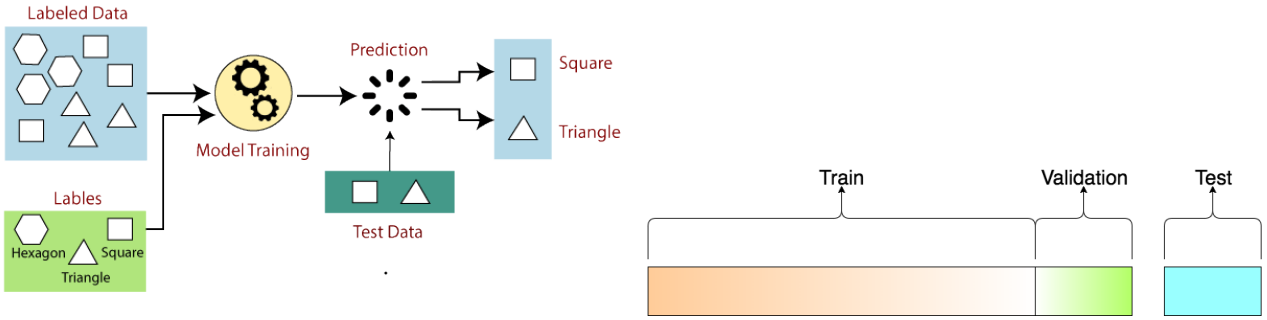
Stages of Supervised Learning



javatpoint.com

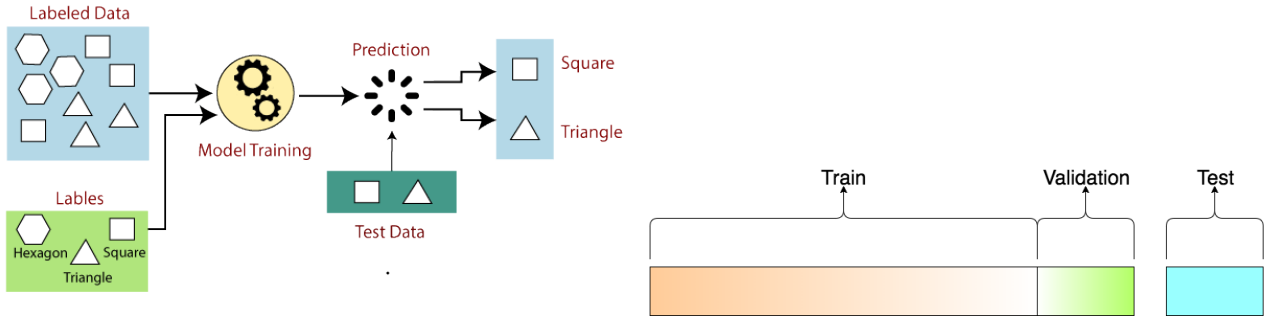
- Training:
  1. Select a model
  2. Train the model by the training set
  3. Validate the model by the validation set

Stages of Supervised Learning



- Training:
  1. Select a model
  2. Train the model by the training set
  3. Validate the model by the validation set
  4. Select the optimum model

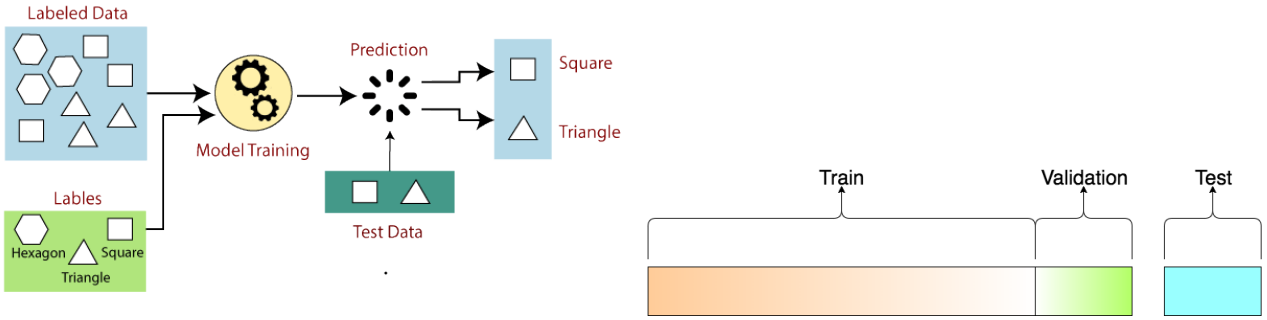
Stages of Supervised Learning



javatpoint.com

- Training:
  1. Select a model
  2. Train the model by the training set
  3. Validate the model by the validation set
  4. Select the optimum model
- Testing:

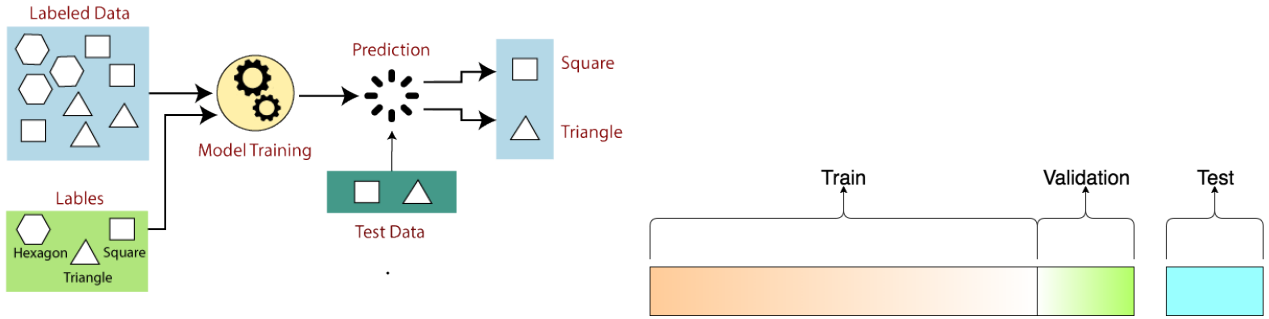
Stages of Supervised Learning



javatpoint.com

- Training:
  1. Select a model
  2. Train the model by the training set
  3. Validate the model by the validation set
  4. Select the optimum model
- Testing:
  1. Test learned model by an unseen part of the data-set.

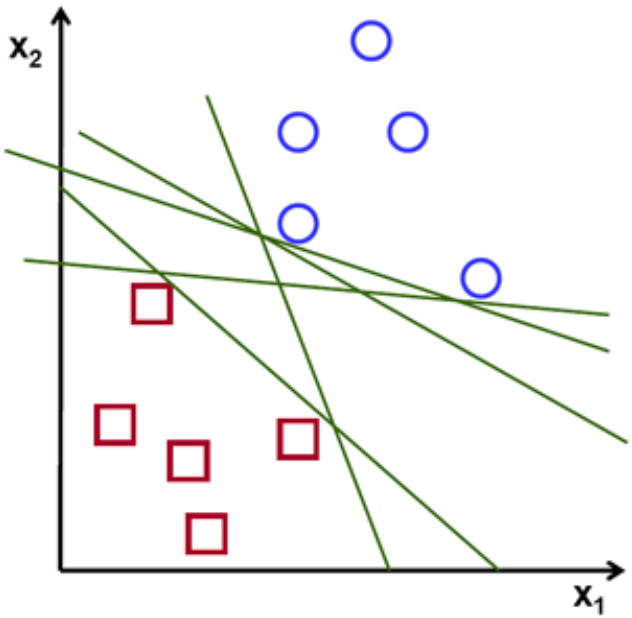
Stages of Supervised Learning



javatpoint.com

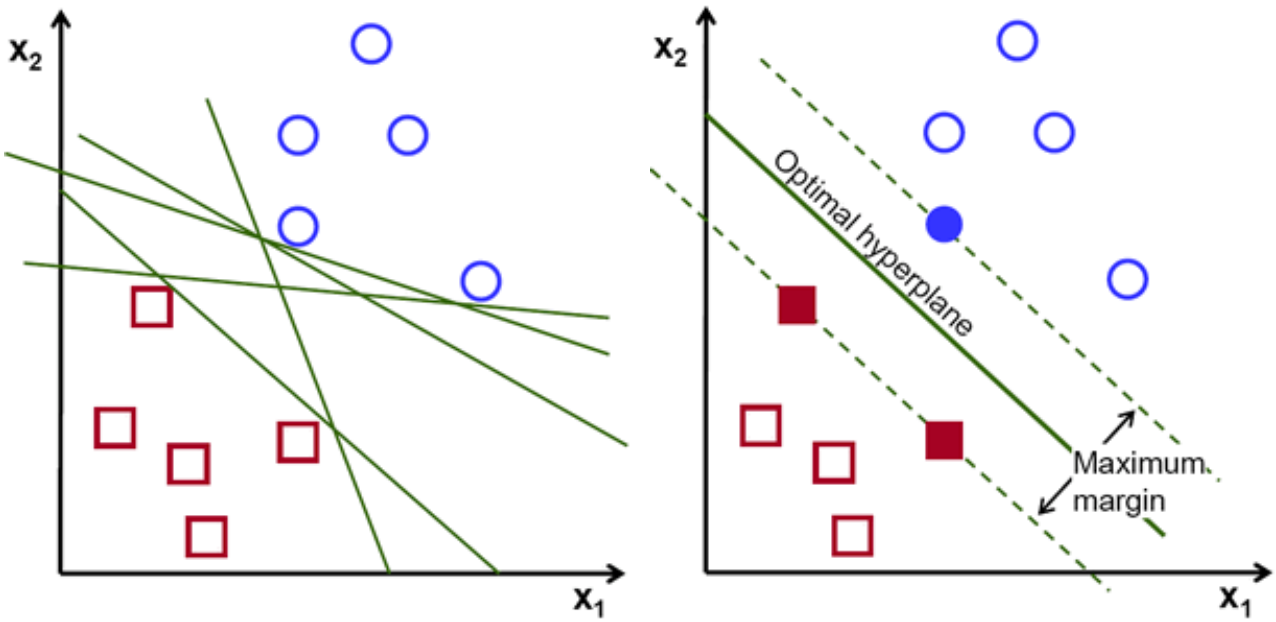
- Training:
  1. Select a model
  2. Train the model by the training set
  3. Validate the model by the validation set
  4. Select the optimum model
- Testing:
  1. Test learned model by an unseen part of the data-set.
  2. Use the best model for predictions about unseen observations.

How Support Vector Machine Works?



towardsdatascience.com

How Support Vector Machine Works?



towardsdatascience.com

Classifying Pre-Main-Sequence Stars using SVM

Hubble Tarantula Treasury Project – VI. Identification of Pre–Main-Sequence Stars using Machine Learning techniques

Victor F. Ksoll,[1,2,⋆] Dimitrios A. Gouliermis,[1,3,†] Ralf S. Klessen,[1] Eva K. Grebel,[4] Elena Sabbi,[5] Jay Anderson,[5] Daniel J. Lennon,[6] Michele Cignoni,[7] Guido de Marchi,[8] Linda J. Smith,[9] Monica Tosi,[10] and Roeland P. van der Marel[5]

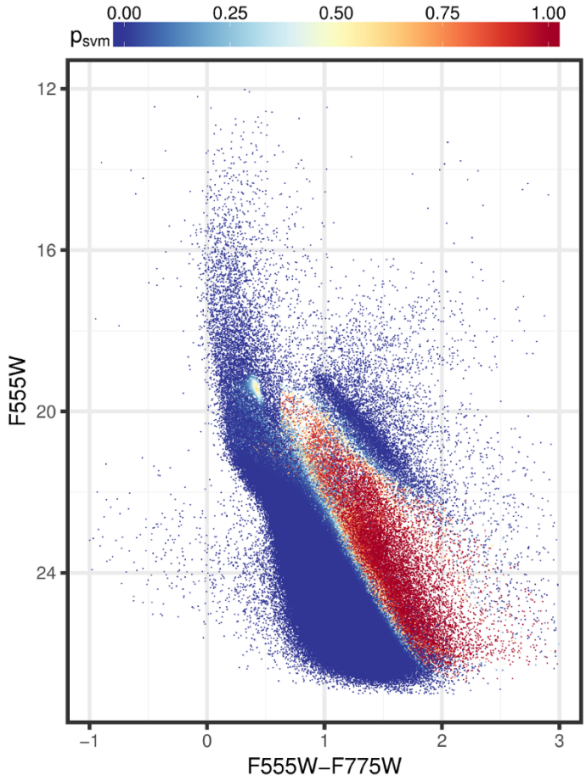
[1]*Institut für Theoretische Astrophysik, Zentrum für Astronomie der Universität Heidelberg, Albert-Ueberle-Str. 2, 69120 Heidelberg, Germany*
[2]*Interdisciplinary Center for Scientific Computing, University of Heidelberg, Mathematikon, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany*
[3]*Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany*
[4]*Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstr. 12-14, 69120 Heidelberg, Germany*
[5]*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*
[6]*ESA – European Space Astronomy Center, Apdo. de Correo 78, E-28691 Associate Villanueva de la Cañada, Madrid, Spain*
[7]*Department of Physics, University of Pisa, Largo Pontecorvo 3, I-56127 Pisa, Italy*
[8]*European Space Research and Technology Centre, Keplerlaan 1, 2200 AG Noordwijk, Netherlands*
[9]*European Space Agency and Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*
[10]*INAF-Osservatorio Astronomico di Bologna, Via Ranzani 1, I-40127 Bologna, Italy*

Draft version 21 May 2018

ABSTRACT

The Hubble Tarantula Treasury Project (HTTP) has provided an unprecedented photometric coverage of the entire star-burst region of 30 Doradus down to the half Solar mass limit. We use the deep stellar catalogue of HTTP to identify all the pre–main-sequence (PMS) stars of the region, i.e., stars that have not started their lives on the main-sequence yet. The photometric distinction of these stars from the more evolved populations is not a trivial task due to several factors that alter their colour-magnitude diagram positions. The identification of PMS stars requires, thus, sophisticated statistical methods. We employ Machine Learning Classification techniques on the HTTP survey of more than 800,000 sources to identify the PMS stellar content of the observed field. Our methodology consists of 1) carefully selecting the most probable low-mass PMS stellar population of the star-forming cluster NGC 2070, 2) using this sample to train classification algorithms to build a predictive model for PMS stars, and 3) applying this model in order to identify the most probable PMS content across the entire Tarantula Nebula. We employ Decision Tree, Random Forest and Support Vector Machine classifiers to categorise the stars as PMS and Non-PMS. The Random Forest and Support Vector Machine provided the most accurate models, predicting about 20,000 sources with a candidateship probability higher than 50 percent, and almost 10,000 PMS candidates with a probability higher than 95 percent. This is the richest and most accurate photometric catalogue of extragalactic PMS candidates across the extent of a whole star-forming complex.

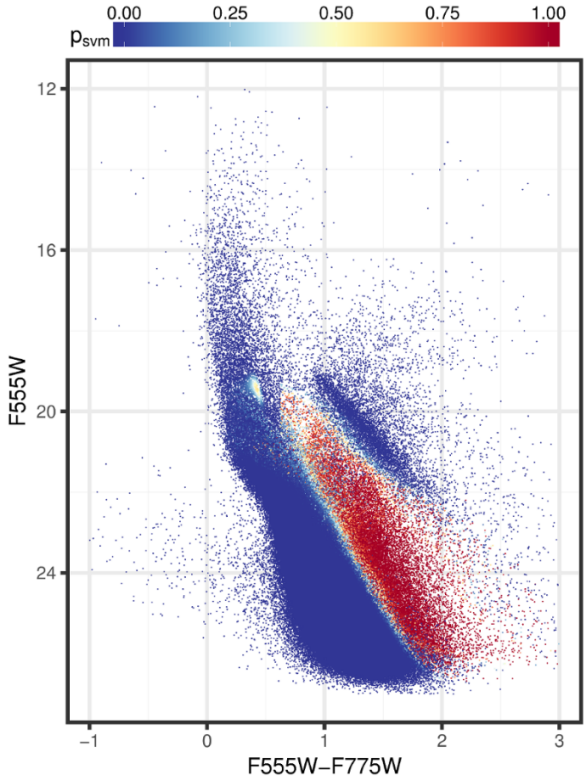
Classifying Pre-Main-Sequence Stars using SVM



Credit: Ksoll, Victor F., et al. 2018

1. Select the most certain PMSs as the training set

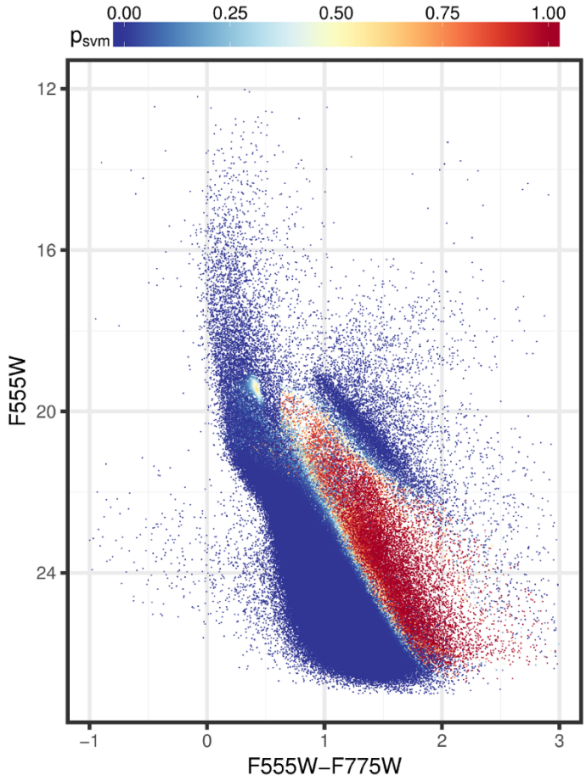
Classifying Pre-Main-Sequence Stars using SVM



Credit: Ksoll, Victor F., et al. 2018

1. Select the most certain PMSs as the training set
2. Train the SVM classifier

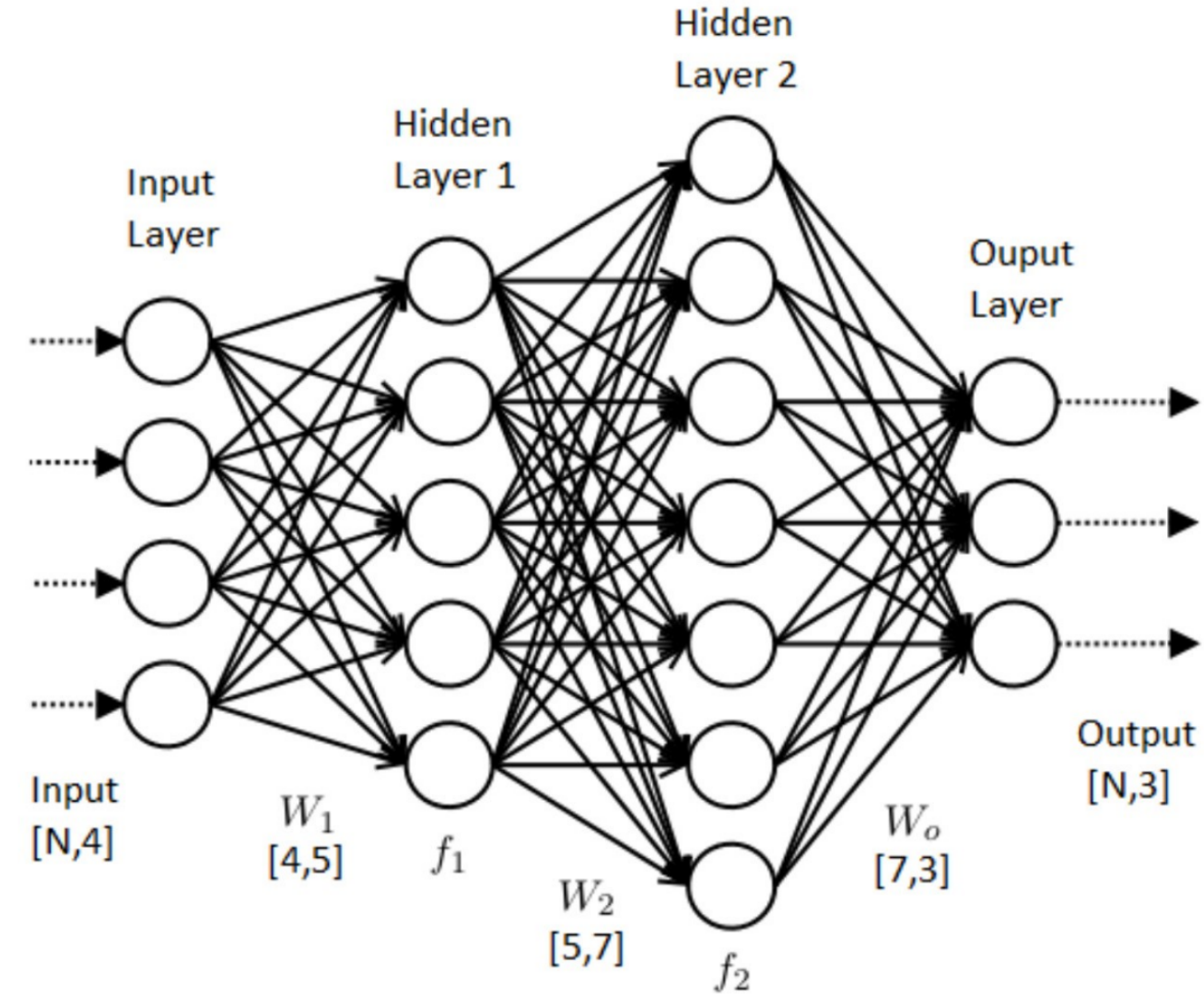
Classifying Pre-Main-Sequence Stars using SVM



Credit: Ksoll, Victor F., et al. 2018

1. Select the most certain PMSs as the training set
2. Train the SVM classifier
3. Use the trained classifier for the entire Tarantula Nebula

How Neural Network works?



Credit: Dalya 2019

Convolutional Neural Network and DLAs

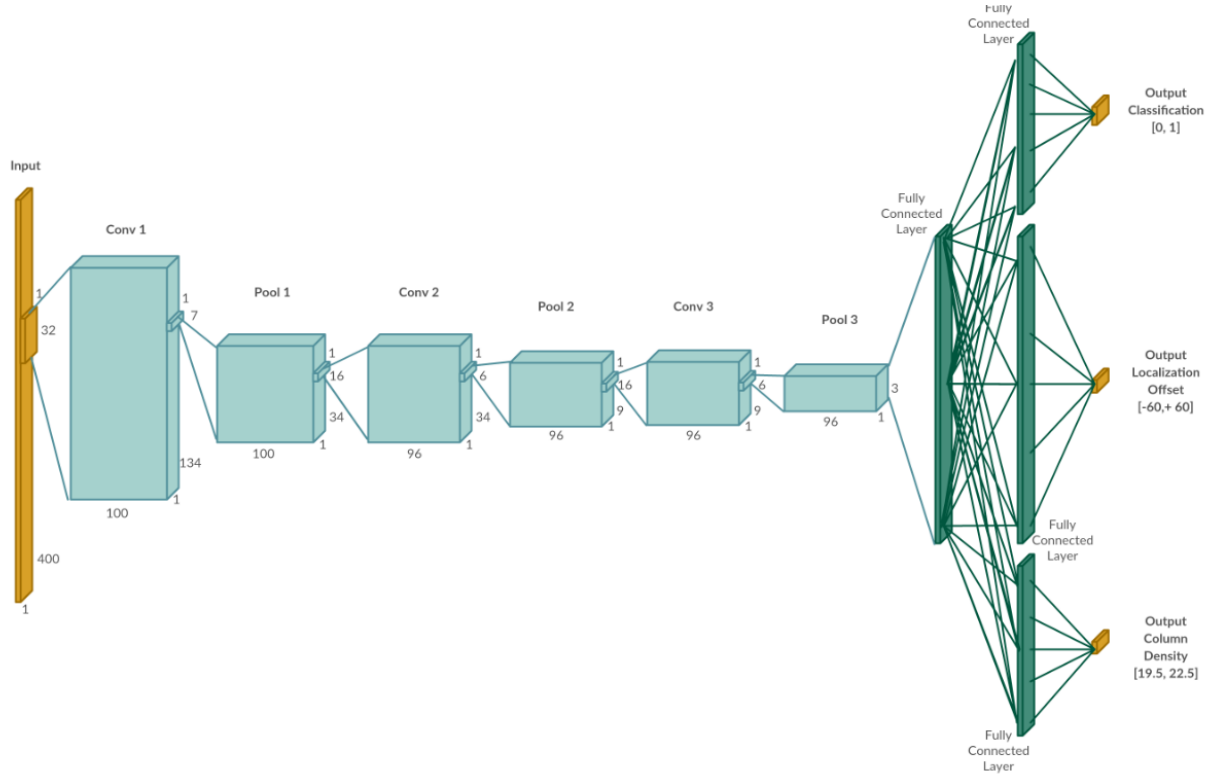
Deep Learning of Quasar Spectra to Discover and Characterize Damped Lyα Systems

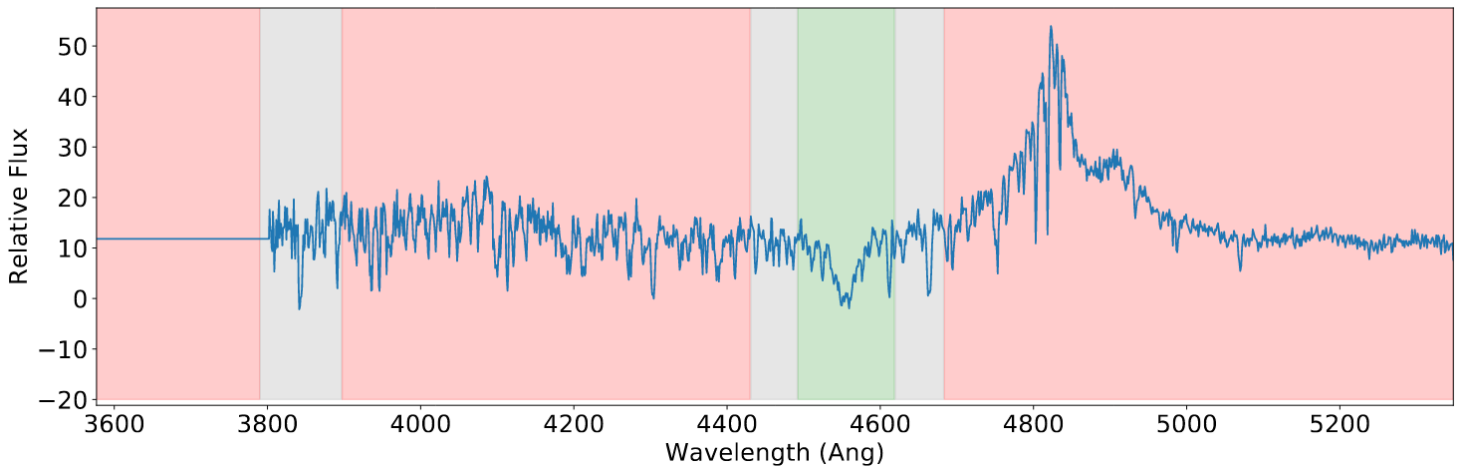
David Parks,[1] J. Xavier Prochaska,[2] Shawfeng Dong,[3] Zheng Cai,[2,4]
[1] *Computer Science, UC Santa Cruz, 1156 High St., Santa Cruz, CA 95064 USA — dfparks@ucsc.edu*
[2] *Astronomy & Astrophysics, UC Santa Cruz, 1156 High St., Santa Cruz, CA 95064 USA*
[3] *Applied Mathematics & Statistics, UC Santa Cruz, 1156 High St., Santa Cruz, CA 95064 USA*
[4] *Hubble Fellow*

Accepted XXX. Received YYY; in original form ZZZ

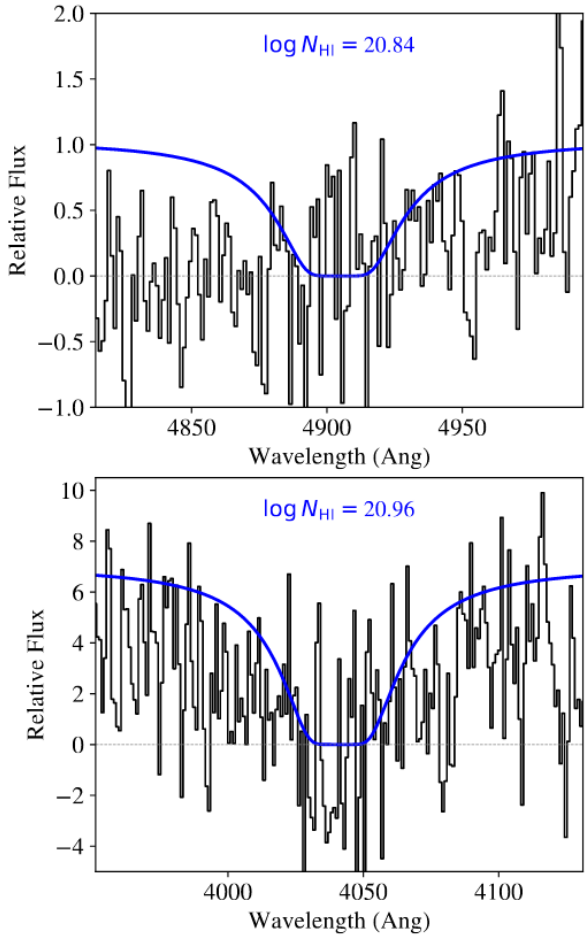
ABSTRACT
We have designed, developed, and applied a convolutional neural network (CNN) architecture using multi-task learning to search for and characterize strong HI Lyα absorption in quasar spectra. Without any explicit modeling of the quasar continuum nor application of the predicted line-profile for Lyα from quantum mechanics, our algorithm predicts the presence of strong HI absorption and estimates the corresponding redshift $z_{\rm abs}$ and HI column density $N_{\rm HI}$, with emphasis on damped Lyα systems (DLAs, absorbers with $N_{\rm HI} \geq 2 \times 10^{20}\,{\rm cm}^{-2}$). We tuned the CNN model using a custom training set of DLAs injected into DLA-free quasar spectra from the Sloan Digital Sky Survey (SDSS), data release 5 (DR5). Testing on a held-back validation set demonstrates a high incidence of DLAs recovered by the algorithm (97.4% as DLAs and 99% as an HI absorber with $N_{\rm HI} > 10^{19.5}\,{\rm cm}^{-2}$) and excellent estimates for $z_{\rm abs}$ and $N_{\rm HI}$. Similar results are obtained against a human-generated survey of the SDSS DR5 dataset. The algorithm yields a low incidence of false positives and negatives but is challenged by overlapping DLAs and/or very high $N_{\rm HI}$ systems. We have applied this CNN model to the quasar spectra of SDSS-DR7 and the Baryonic Oscillation Spectroscopic Survey (BOSS, data release 12) and provide catalogs of 4,913 and 50,969 DLAs respectively (including 1,659 and 9,230 high-confidence DLAs that were previously unpublished). This work validates the application of deep learning techniques to astronomical spectra for both classification and quantitative measurements.

Training DLA/noDLA models: 1D image recognition



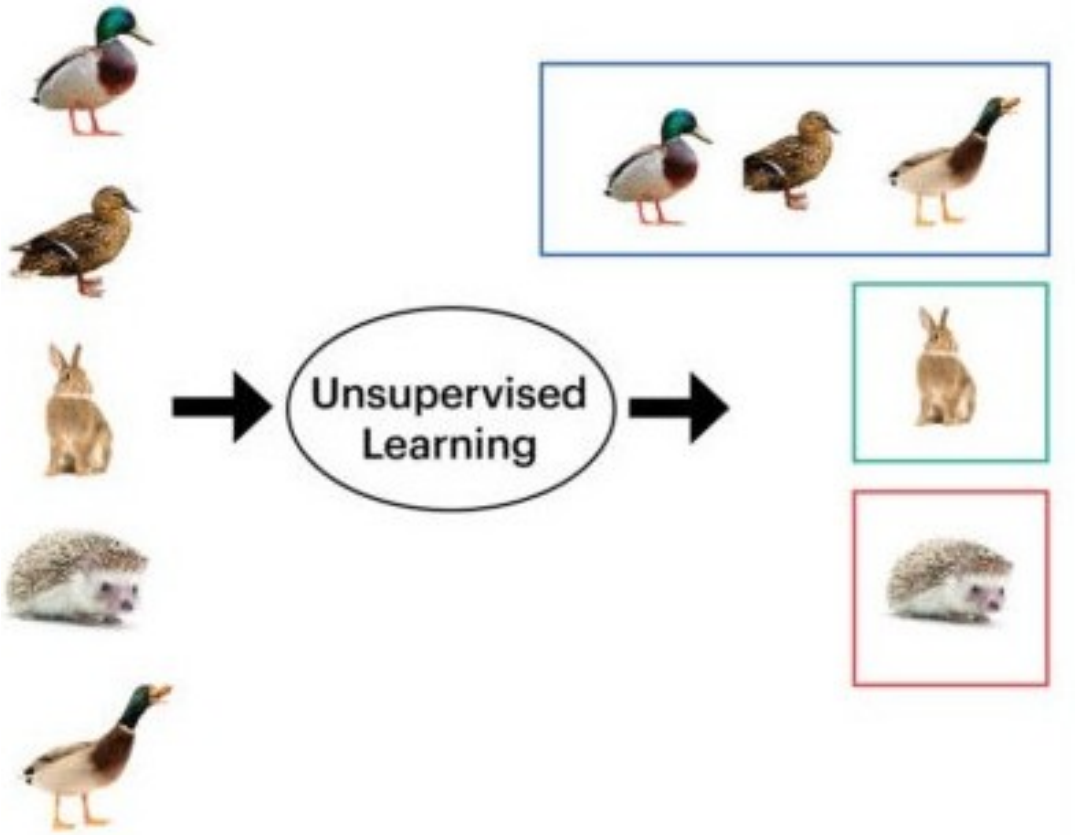


Multitask CNN: Location, Density and Class



Credit: Parks et. al. 2017

How unsupervised learning works?



Unsupervised Learning and knowledge discovery

- Clustering:
  Distinguishing similar objects and patterns in the data-set

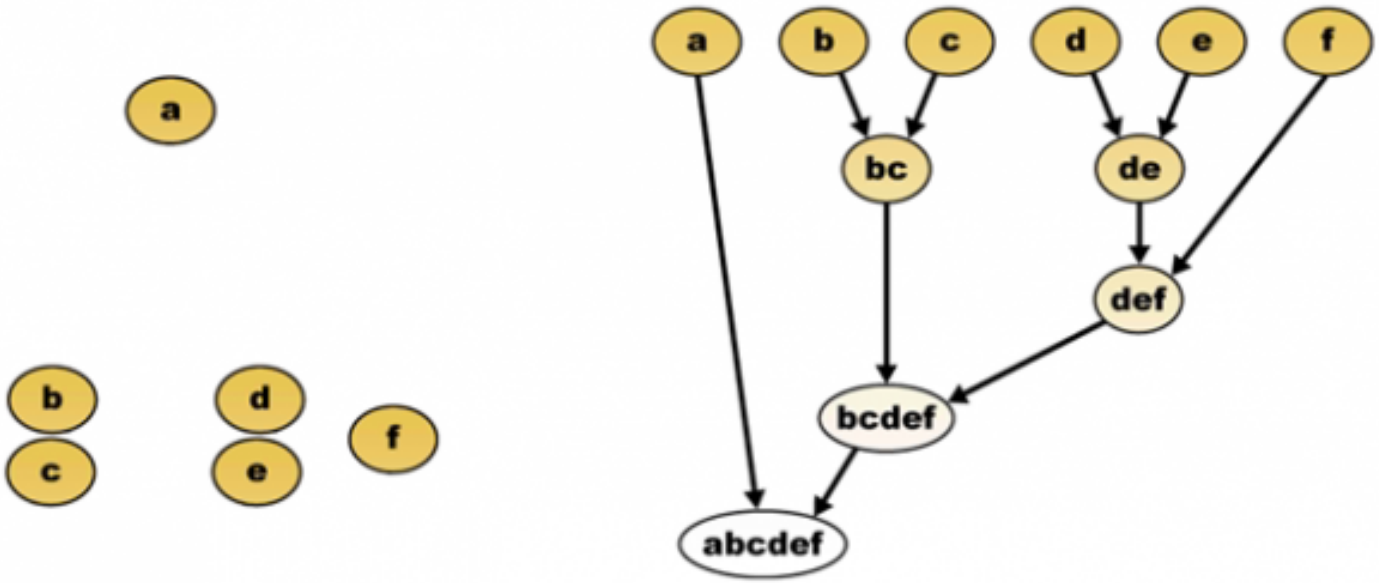
Unsupervised Learning and knowledge discovery

- Clustering:
  Distinguishing similar objects and patterns in the data-set
- Outlier detection:
  Recognizing odd objects which cannot be categorized and do not belong to any group in the data-set

Unsupervised Learning and knowledge discovery

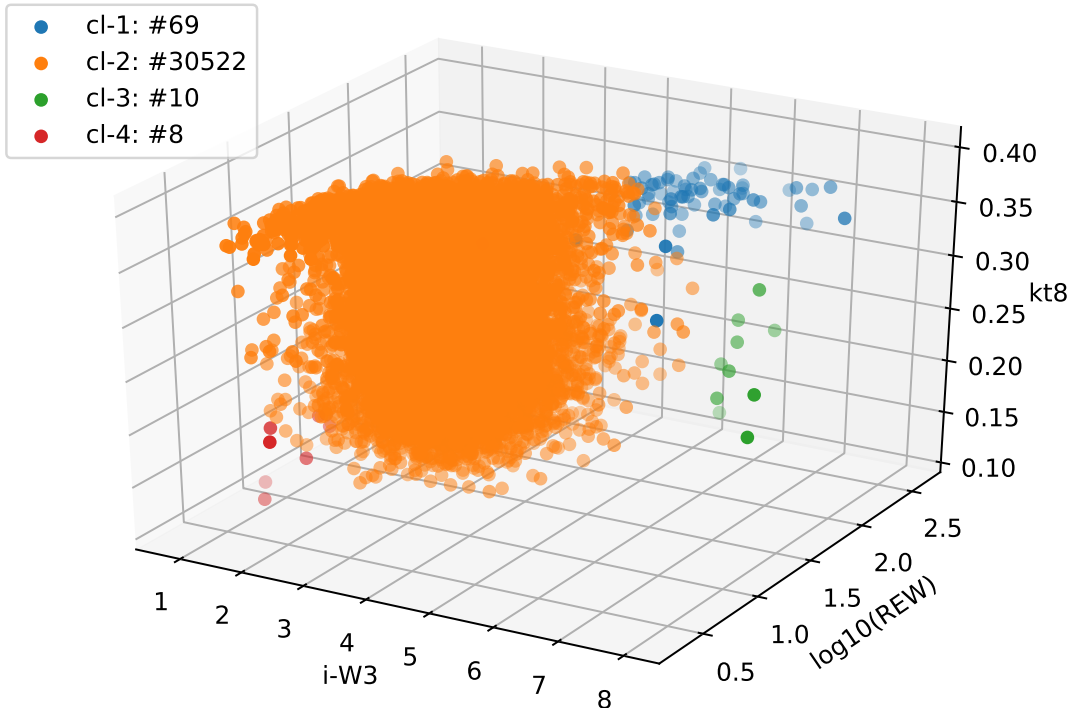
- Clustering:
  Distinguishing similar objects and patterns in the data-set
- Outlier detection:
  Recognizing odd objects which cannot be categorized and do not belong to any group in the data-set
- Visualization and dimentionality reduction:
  Mapping high dimensional data set to a 2D or 3D space

How Agglomerative clustering works?

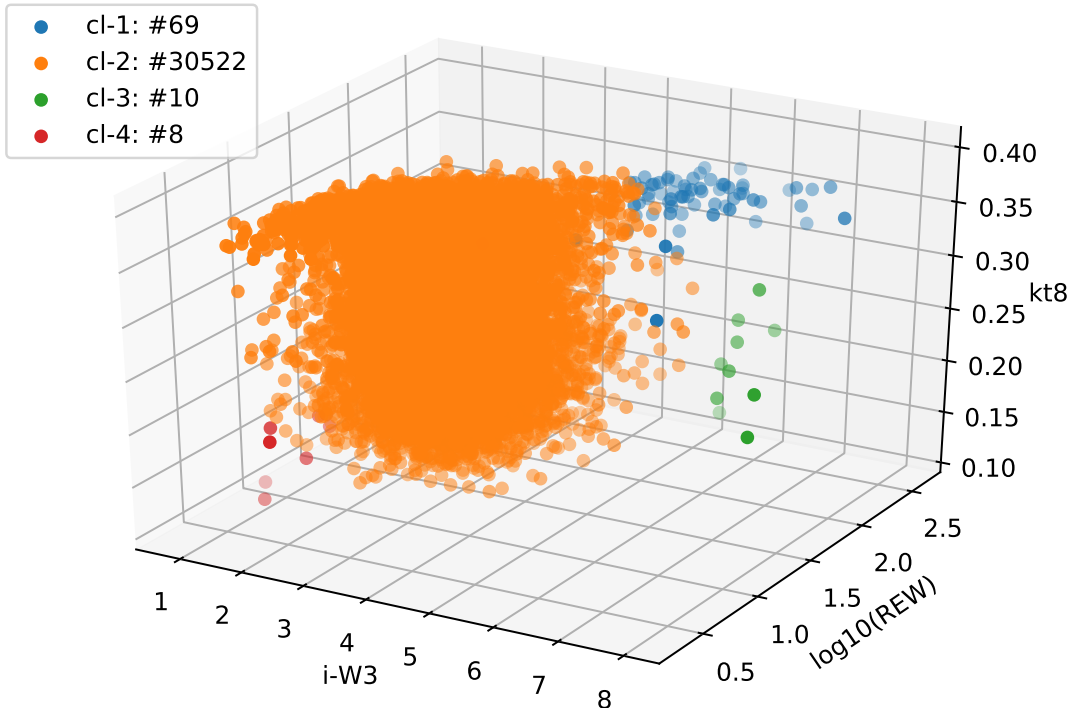


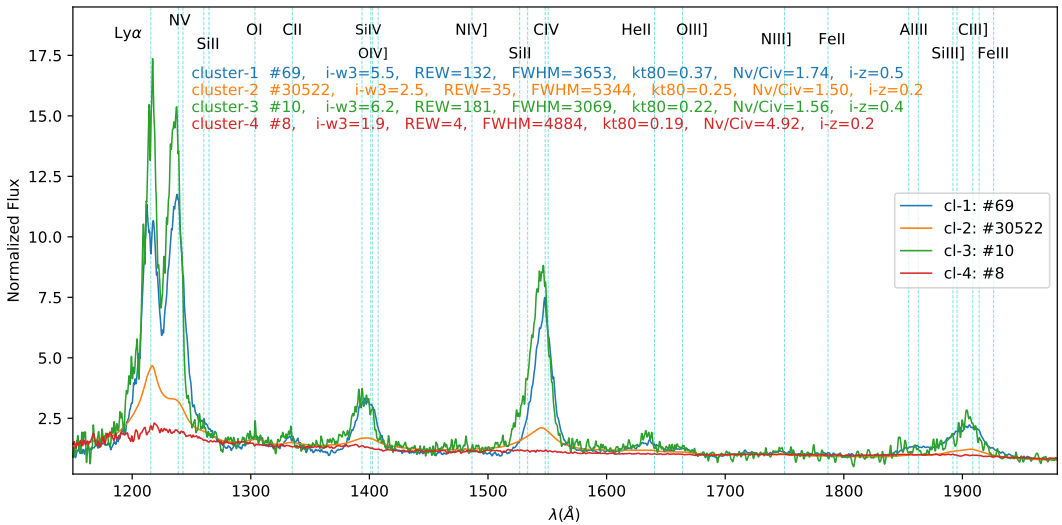
Credit: blog.minitab.com

Agglomerative clustering finds weird quasars?

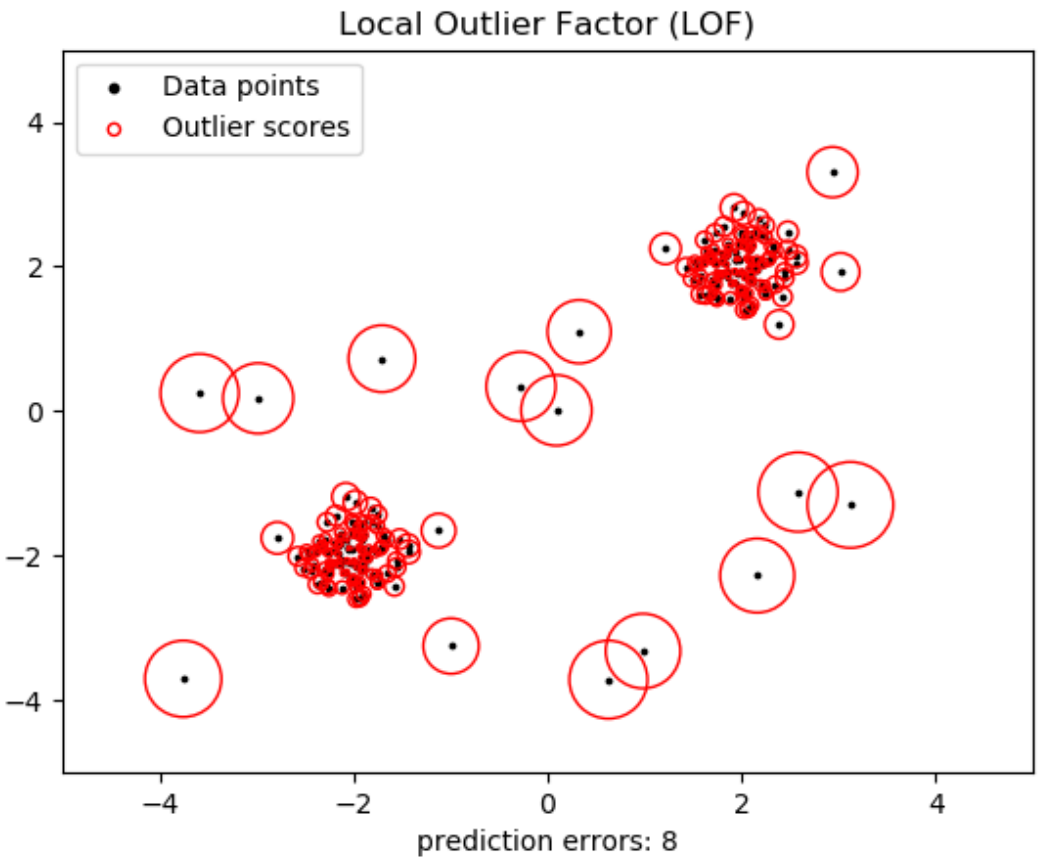


Agglomerative clustering finds weird quasars?



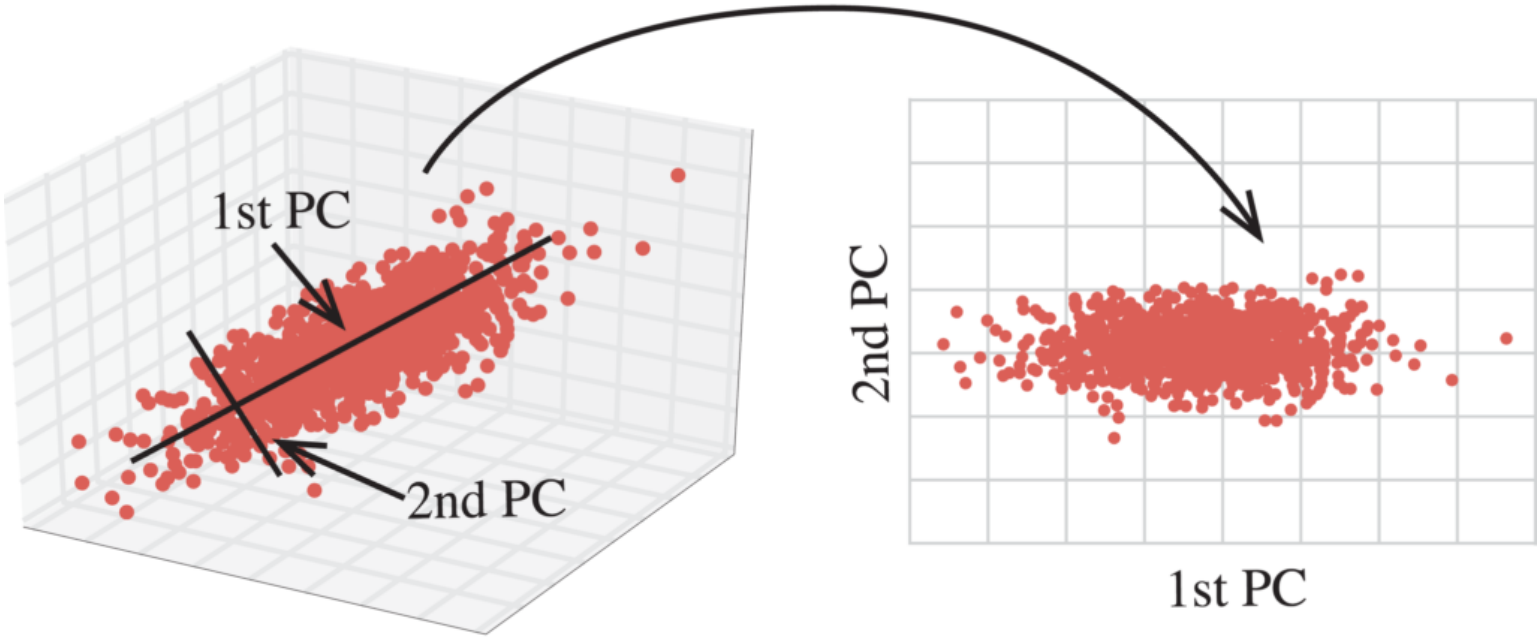


Local Outlier Factor finds outliers



Credit: scikit-learn.org

Dimensionality Reduction with Principal Component Analysis



Credit: medium.com

Limitation of ML in astronomy

- No pre-processing = misleading results

Limitation of ML in astronomy

- No pre-processing = misleading results
- Training on a biased sample = wrong predictions

Limitation of ML in astronomy

- No pre-processing = misleading results
- Training on a biased sample = wrong predictions
- ML is not plug and play, it needs **expertise**

Limitation of ML in astronomy

- No pre-processing = misleading results
- Training on a biased sample = wrong predictions
- ML is not plug and play, it needs **expertise**
- There are a variety of methods for each task: pre-knowledge is needed

Limitation of ML in astronomy

- No pre-processing = misleading results
- Training on a biased sample = wrong predictions
- ML is not plug and play, it needs **expertise**
- There are a variety of methods for each task: pre-knowledge is needed
- Some algorithms scarify interpretation for functionality: black boxes

Limitation of ML in astronomy

- No pre-processing = misleading results
- Training on a biased sample = wrong predictions
- ML is not plug and play, it needs **expertise**
- There are a variety of methods for each task: pre-knowledge is needed
- Some algorithms scarify interpretation for functionality: black boxes
- ML algorithms are mostly designed for industry (e.g. Google, Walmart, Uber, ...) not for astronomy.

Limitation of ML in astronomy

- No pre-processing = misleading results
- Training on a biased sample = wrong predictions
- ML is not plug and play, it needs **expertise**
- There are a variety of methods for each task: pre-knowledge is needed
- Some algorithms scarify interpretation for functionality: black boxes
- ML algorithms are mostly designed for industry (e.g. Google, Walmart, Uber, ...) not for astronomy.
- Collaboration of astronomers with computer scientists needs to be much stronger

References

1. **ML, General:** Baron, Dalya. "Machine learning in astronomy: A practical overview." arXiv preprint arXiv:1904.07248 (2019).
2. **Big Data:** Zhang, Yanxia, and Yongheng Zhao. "Astronomy in the big data era." Data Science Journal 14 (2015).
3. **SVM:** Ksoll, Victor F., et al. "Hubble Tarantula Treasury ProjectVI. Identification of pre-main-sequence stars using machine-learning techniques." Monthly Notices of the Royal Astronomical Society 479.2 (2018): 2389-2414.
4. **Decision Tree:** Vasconcellos, E. C., et al. "Decision tree classifiers for star/galaxy separation." The Astronomical Journal 141.6 (2011): 189.
5. **Neural Network:** Parks, David, et al. "Deep learning of quasar spectra to discover and characterize damped Ly systems." Monthly Notices of the Royal Astronomical Society 476.1 (2018): 1151-1168.
6. **Agglomerative clustering:** R. Monadi, F. Hamann, S. Bird, Precise Selection of Extremely Red Quasars, (in preparation)