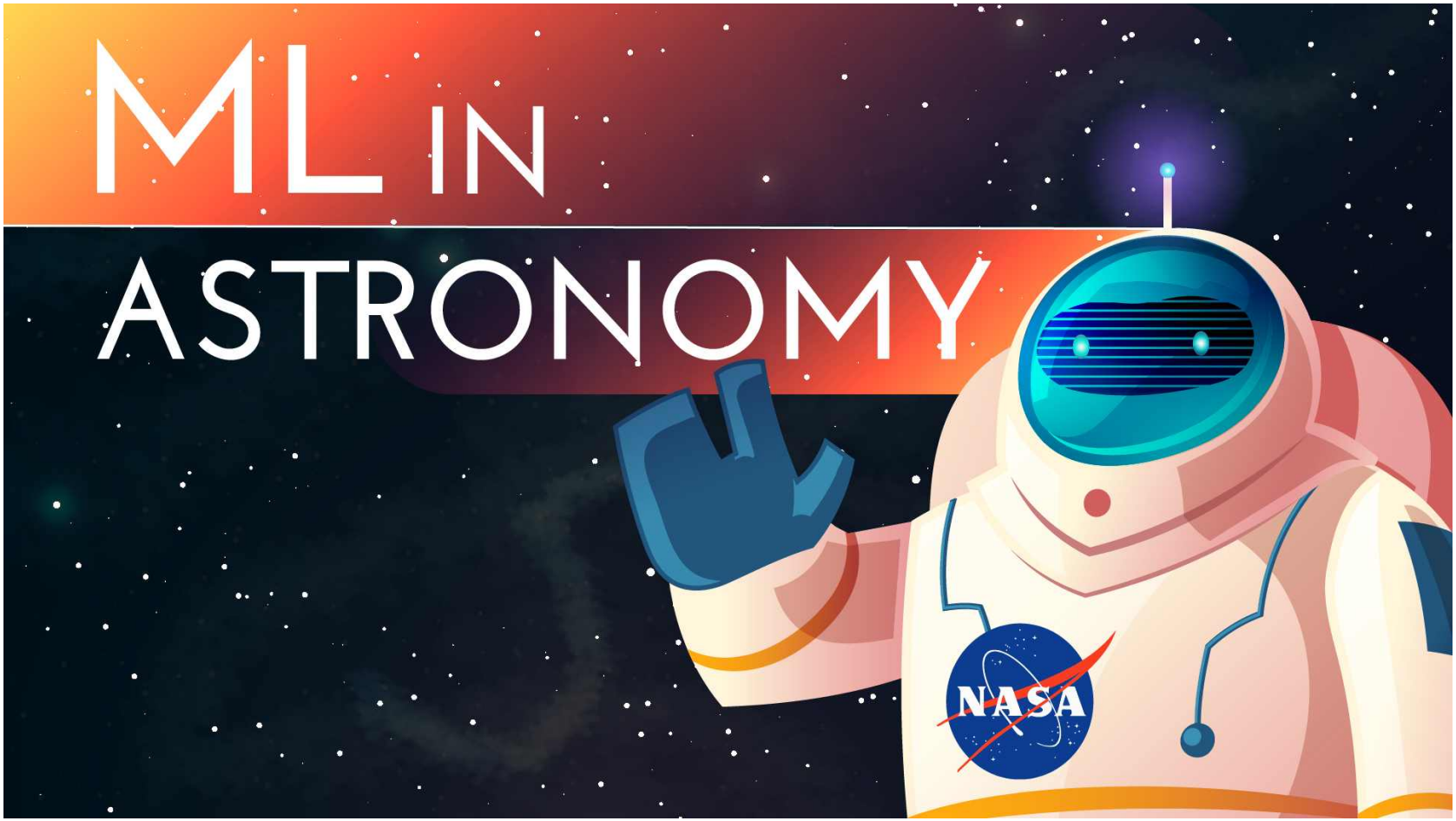
Machine Learning in Astronomy

Reza Monadi

UC Riverside

May 14, 2020



credit: 365datascience.com

- How astronomy is tied to **BIG DATA**?

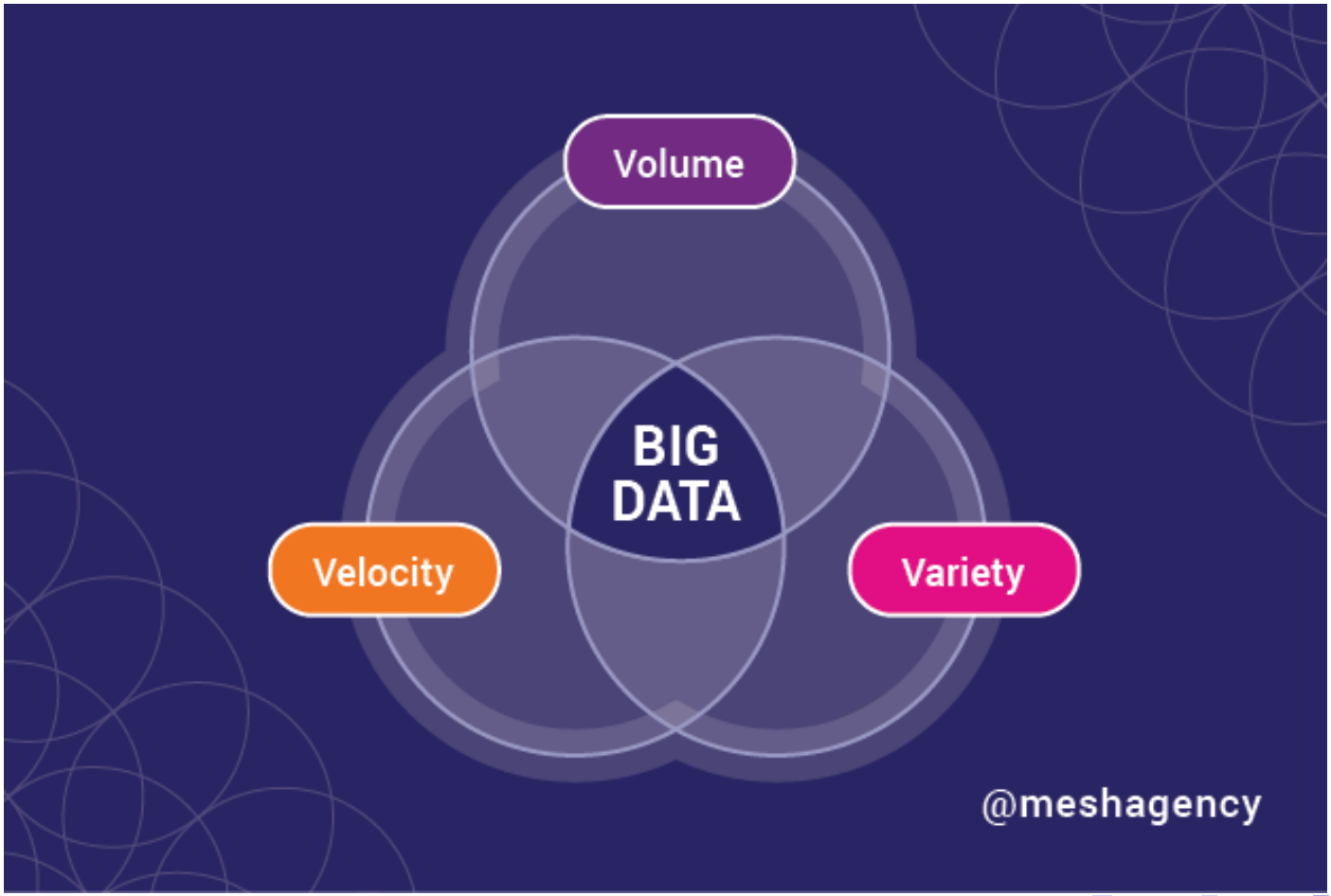
- How astronomy is tied to **BIG DATA**?
- What is supervised and unsupervised **ML**?

- How astronomy is tied to **BIG DATA**?
- What is supervised and unsupervised **ML**?
- How to implement **ML** algorithms in astronomy?

- How astronomy is tied to **BIG DATA**?
- What is supervised and unsupervised **ML**?
- How to implement **ML** algorithms in astronomy?
- How **ML** helps **SKA**?

- How astronomy is tied to **BIG DATA**?
- What is supervised and unsupervised **ML**?
- How to implement **ML** algorithms in astronomy?
- How **ML** helps **SKA**?
- What are the pitfalls of **ML** in astronomy?

What is BIG DATA?



VVV in astronomy

- Volume: larger quantities of data by better facilitates

VVV in astronomy

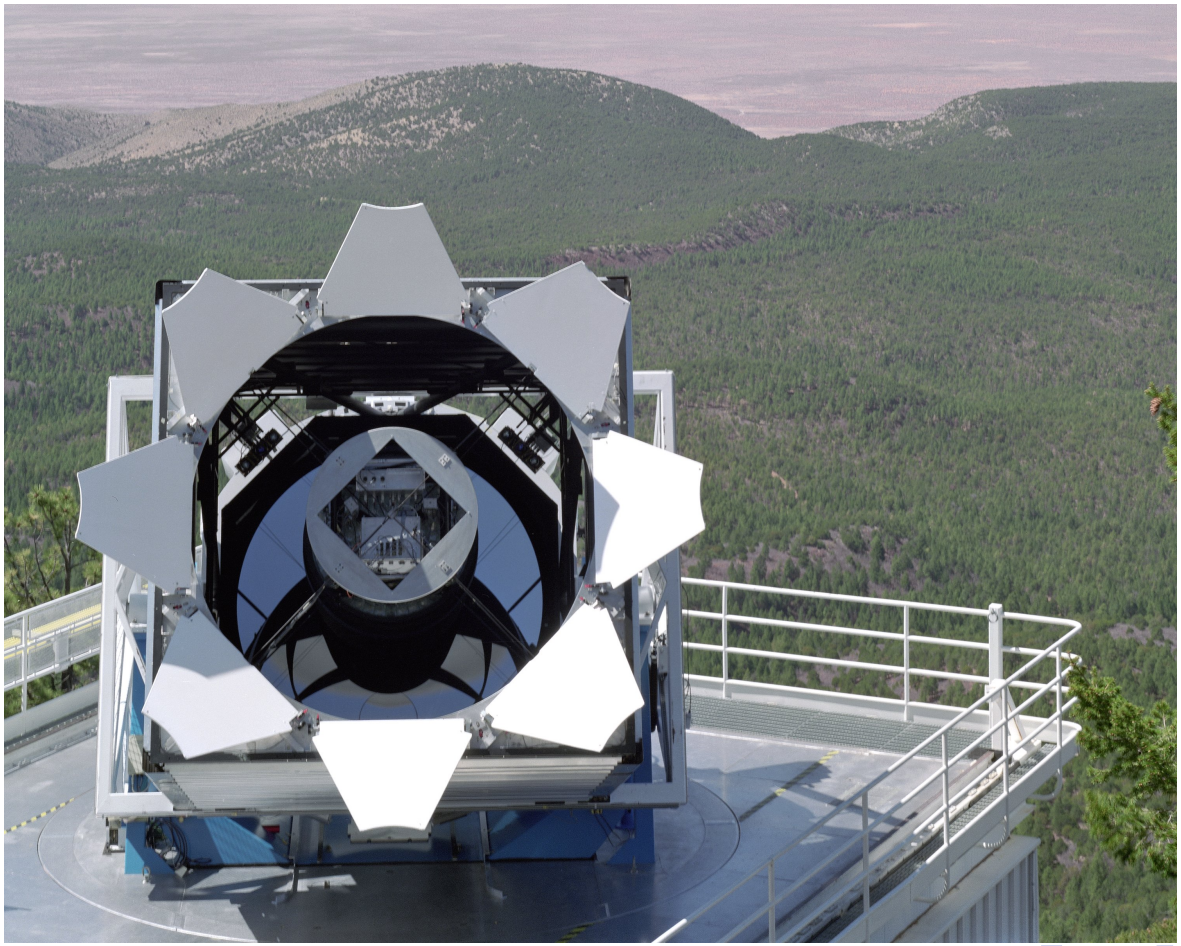
- Volume: larger quantities of data by better facilitates
- Velocity: Higher speed of getting data

VVV in astronomy

- Volume: larger quantities of data by better facilitates
- Velocity: Higher speed of getting data
- Verity: More complex structures of data

Astronomical surveys are really Astronomical.

Sloan Digital Sky Server ⇒ 40 TB



Sloan Digital Sky Server $\Rightarrow$ 40 TB

Imaging statistics

Total unique area covered	14,555 square degrees
Total area of imaging (including overlaps)	31,637 square degrees (excluding supernova runs)
Individual image field size	1361x2048 pixels (0.0337 square degrees)
Number of individual fields	938,046 (excluding supernova runs)
Number of catalog objects	1,231,051,050
Number of unique detections	932,891,133
Median PSF FWHM, *r*-band	1.3 arcsec
Pixel scale	0.396 arcsec
Exposure time per band	53.9 sec
Time difference between observations of each band	71.72 sec (in *riuzg* order)
Global astrometric precision	0.1 arcsec rms (absolute)

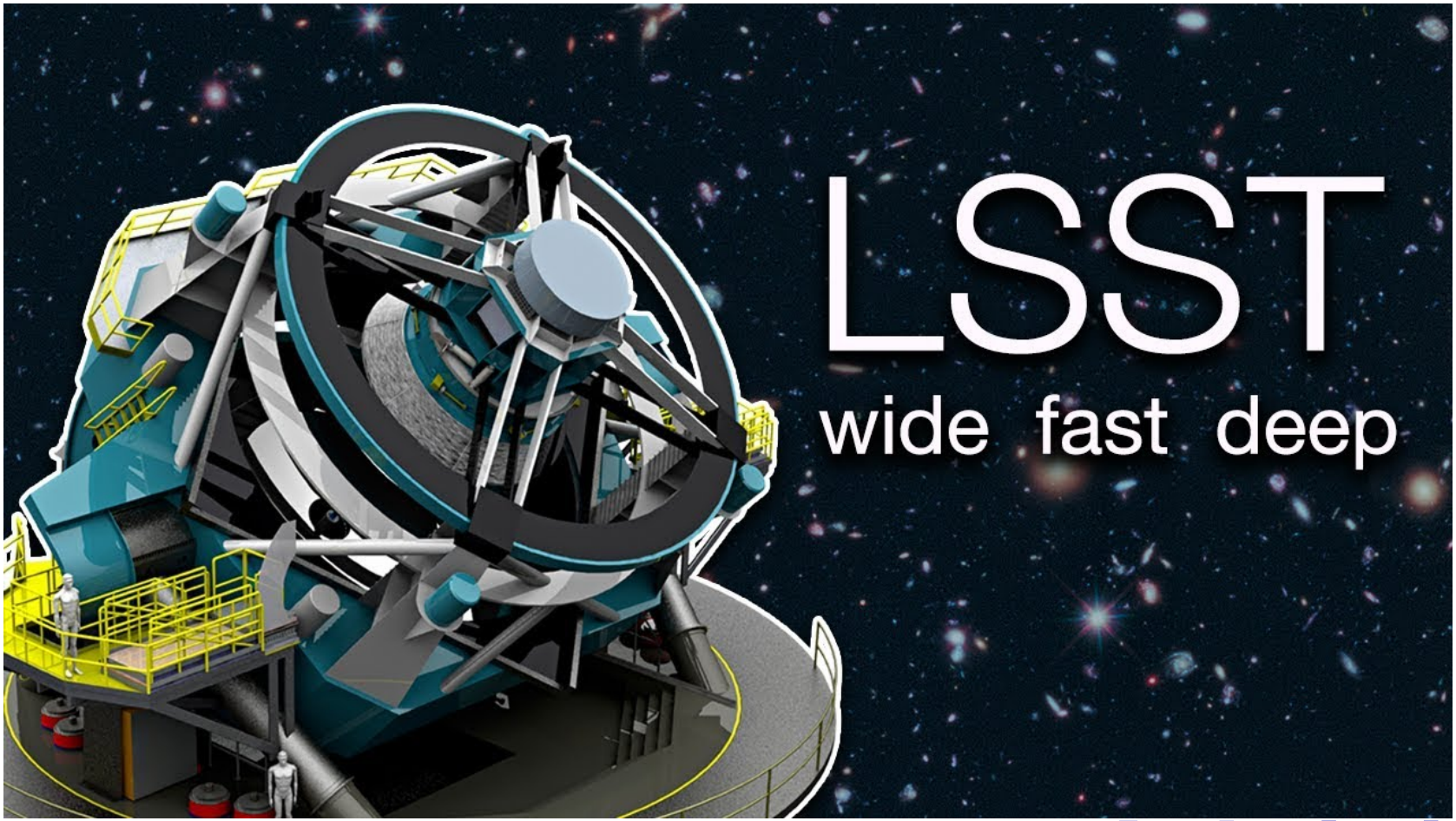
Sloan Digital Sky Server ⇒ 40 TB

Optical spectroscopy data statistics

All programs combined

Total spectra	5,789,200
Useful spectra	4,846,156
Galaxies	2,863,635
Quasars	960,678
Stars	1,021,843
Sky	475,531
Standards	108,603
Unknown	352,320

Large Synaptic Survey Telescope $\Rightarrow$ 200 PB



Large Synaptic Survey Telescope $\Rightarrow$ 200 PB

- 800+ panoramic images each night

Large Synaptic Survey Telescope ⇒ 200 PB

- 800+ panoramic images each night
- 3.2 billion-pixel camera

Large Synaptic Survey Telescope $\Rightarrow$ 200 PB

- 800+ panoramic images each night
- 3.2 billion-pixel camera
- Recording the entire visible sky twice each week

Large Synaptic Survey Telescope $\Rightarrow$ 200 PB

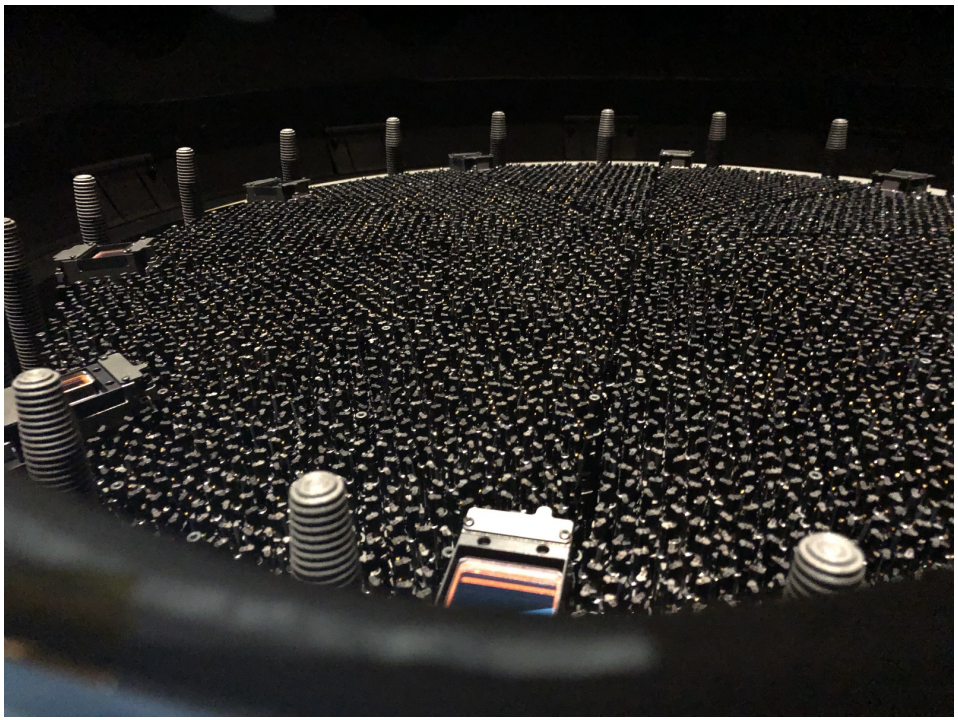
- 800+ panoramic images each night
- 3.2 billion-pixel camera
- Recording the entire visible sky twice each week
- Each patch of sky will be visited 1000 times

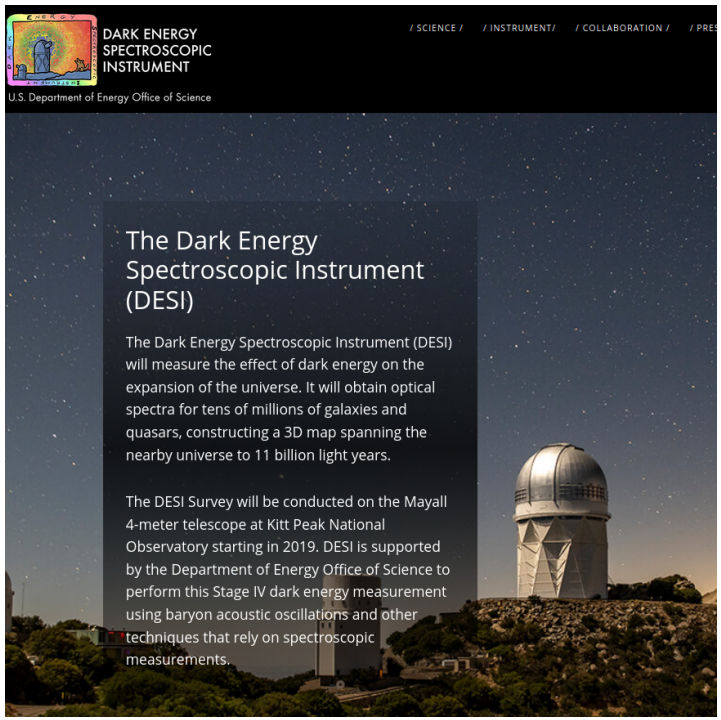
Zwicky Transient Facility

Zwicky Transient Facility

Filter(s)	#PSFcat-*sci* sources	#Aperturecat-*sci* sources	#PSFcat-*ref* sources	#Aperturecat-*ref* sources
g	34,799,157,939	22,008,868,967	2,053,507,697	656,772,487
r	64,698,158,346	40,249,465,415	2,954,247,127	1,013,085,916
i	868,042,655	495,178,759	557,831,277	163,240,538
g + *r* + *i*	100,365,358,940	62,753,513,141	5,565,586,101	1,833,098,941

Dark Energy Spectroscopic Instrument





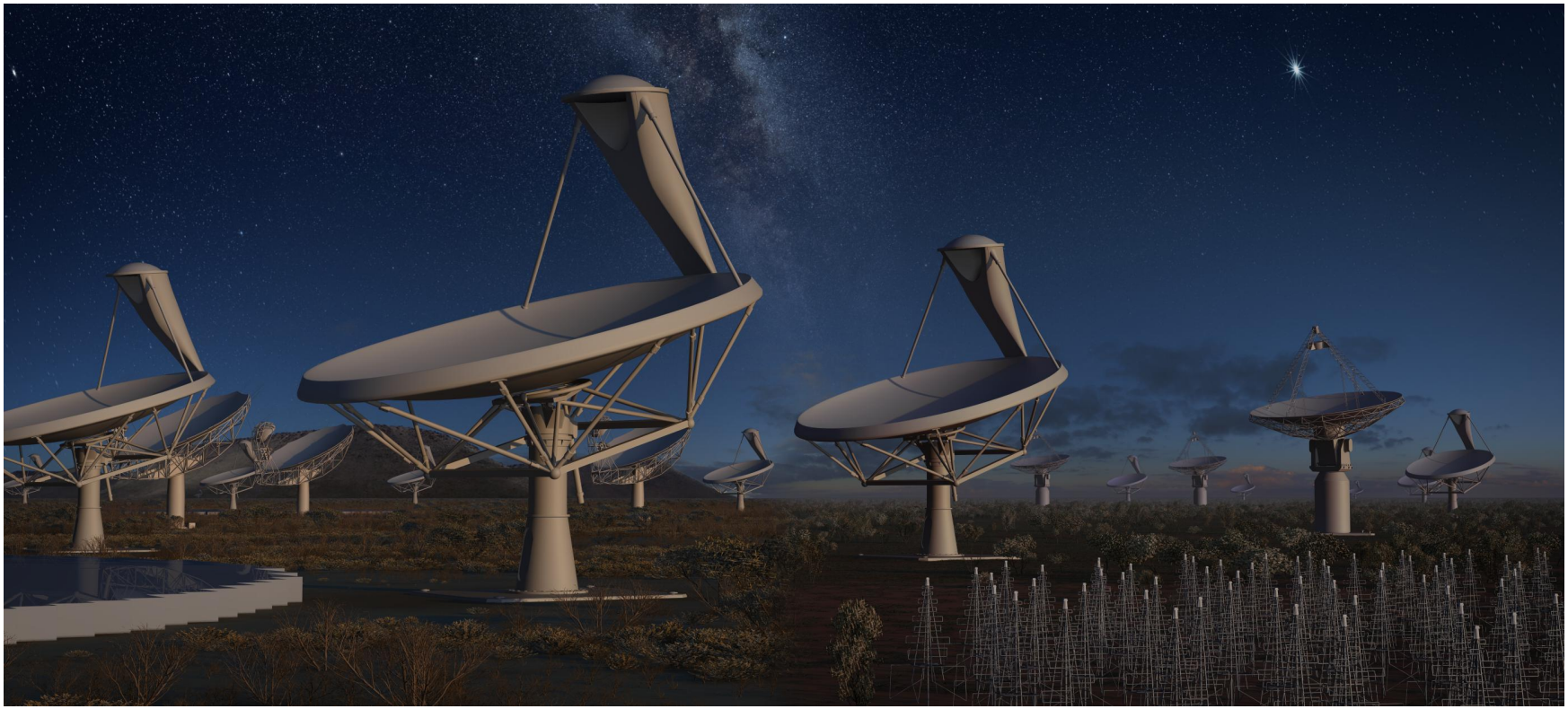
Dark Energy Spectroscopic Instrument

- Spectra of 25 M galxies, quasars, stars.

Dark Energy Spectroscopic Instrument

- Spectra of 25 M galxies, quasars, stars.
- 5000 spectra per exposure

Square Kilometer Array $\Rightarrow$ 4.6 EB



Large surveys

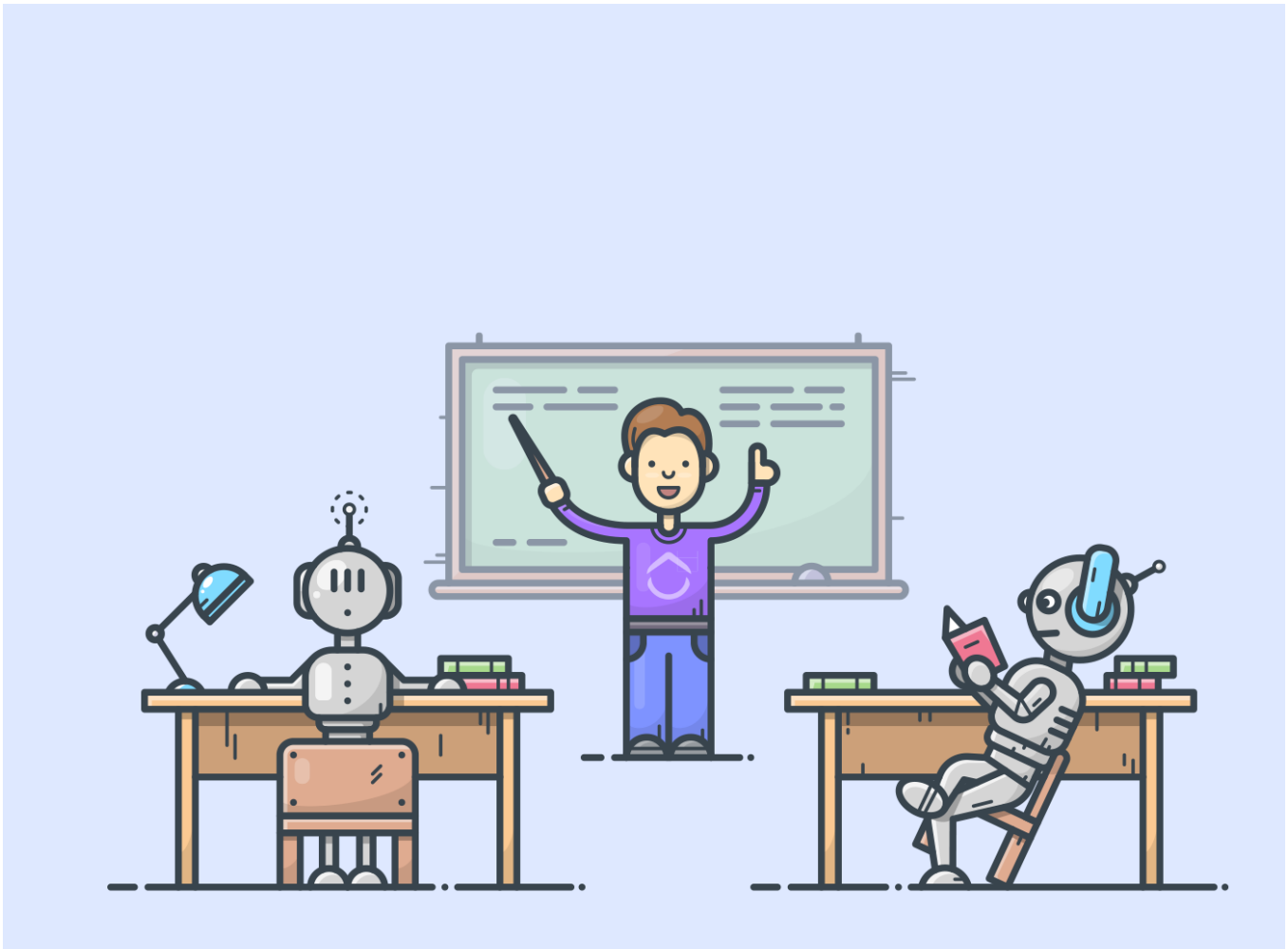
Large surveys $\Rightarrow$

Large surveys $\Rightarrow$ Big Data

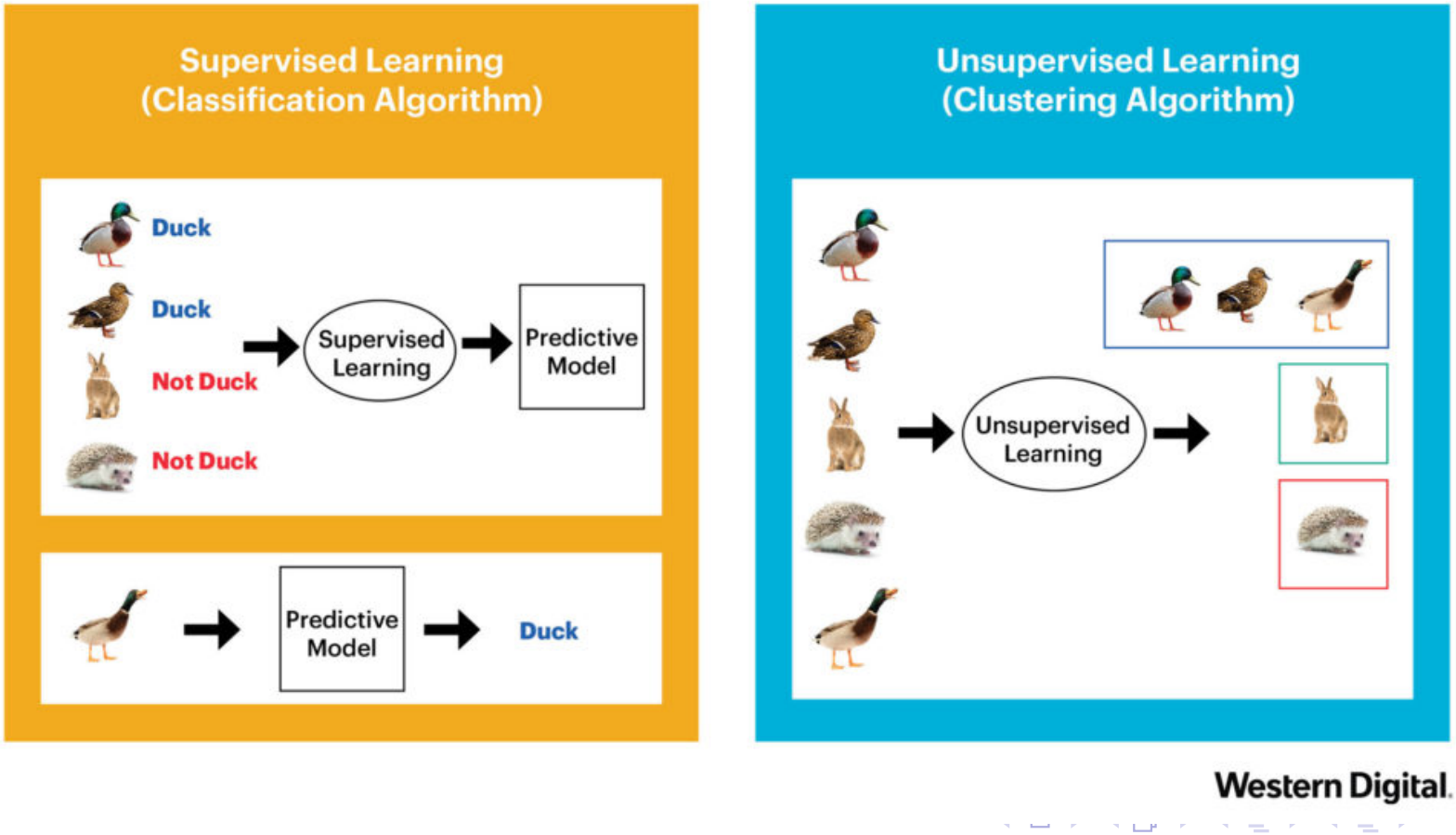
Large surveys $\Rightarrow$ Big Data $\overset{ML}{\Longrightarrow}$

$$\text{Large surveys} \Rightarrow \text{Big Data} \overset{ML}{\Longrightarrow} \text{Astronomy Knowledge}$$

Supervised ML vs. Unsupervised ML



Supervised ML vs. Unsupervised ML



Stages of Supervised Learning

- Training:

Stages of Supervised Learning

- Training:
  1. Select a model

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
- Validation:

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
- Validation:
  1. Change the hyper-parameters

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
- Validation:
  1. Change the hyper-parameters
  2. Select the optimum hyper-parameters

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
- Validation:
  1. Change the hyper-parameters
  2. Select the optimum hyper-parameters
- Testing:

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
- Validation:
  1. Change the hyper-parameters
  2. Select the optimum hyper-parameters
- Testing:
  1. Test learned model by an unseen part of the data-set.

Stages of Supervised Learning

- Training:
  1. Select a model
  2. Set up hyper-parameters of model
  3. Teach the machine by training set
- Validation:
  1. Change the hyper-parameters
  2. Select the optimum hyper-parameters
- Testing:
  1. Test learned model by an unseen part of the data-set.
  2. Select the best model and use it for predictions.

Supervised Learning vs Traditional Model Fitting ?

Similarities

Supervised Learning vs Traditional Model Fitting ?

Similarities

- Both need a set of labeled measurements and a model

Supervised Learning vs Traditional Model Fitting ?

Similarities

- Both need a set of labeled measurements and a model

Differences

Supervised Learning vs Traditional Model Fitting ?

Similarities

- Both need a set of labeled measurements and a model

Differences

- SML: model is constructed based on input data

Supervised Learning vs Traditional Model Fitting ?

Similarities

- Both need a set of labeled measurements and a model

Differences

- SML: model is constructed based on input data
- SML: model can be very nonlinear/complex

Supervised Learning vs Traditional Model Fitting ?

Similarities

- Both need a set of labeled measurements and a model

Differences

- SML: model is constructed based on input data
- SML: model can be very nonlinear/complex
- TMF: model is predefined and has limited adaptivity

Supervised Learning vs Traditional Model Fitting ?

Similarities

- Both need a set of labeled measurements and a model

Differences

- SML: model is constructed based on input data
- SML: model can be very nonlinear/complex
- TMF: model is predefined and has limited adaptivity
- SML: designed for predicting unseen data

Supervised Learning vs Traditional Model Fitting ?

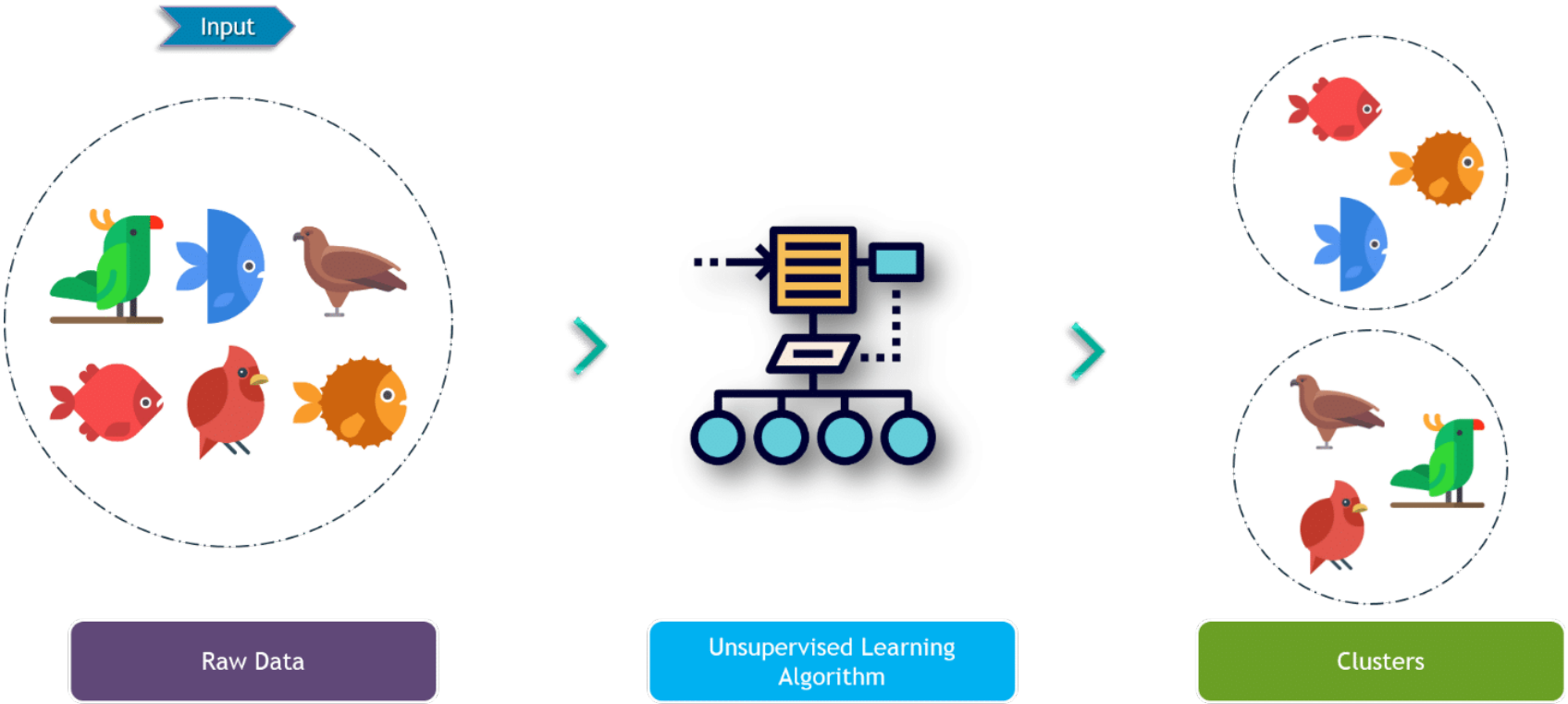
Similarities

- Both need a set of labeled measurements and a model

Differences

- SML: model is constructed based on input data
- SML: model can be very nonlinear/complex
- TMF: model is predefined and has limited adaptivity
- SML: designed for predicting unseen data
- TMF: infers relationships between features

How unsupervised learning works?



Unsupervised Learning and knowledge discovery

- Finding unknown patterns in the data set

Unsupervised Learning and knowledge discovery

- Finding unknown patterns in the data set
- Distinguishing similar and dissimilar objects

Unsupervised Learning and knowledge discovery

- Finding unknown patterns in the data set
- Distinguishing similar and dissimilar objects
- Helping for visualization in lower dimensions

Unsupervised Learning and knowledge discovery

- Finding unknown patterns in the data set
- Distinguishing similar and dissimilar objects
- Helping for visualization in lower dimensions
- Recognizing odd objects (unknown unknowns)

Unsupervised Learning and knowledge discovery

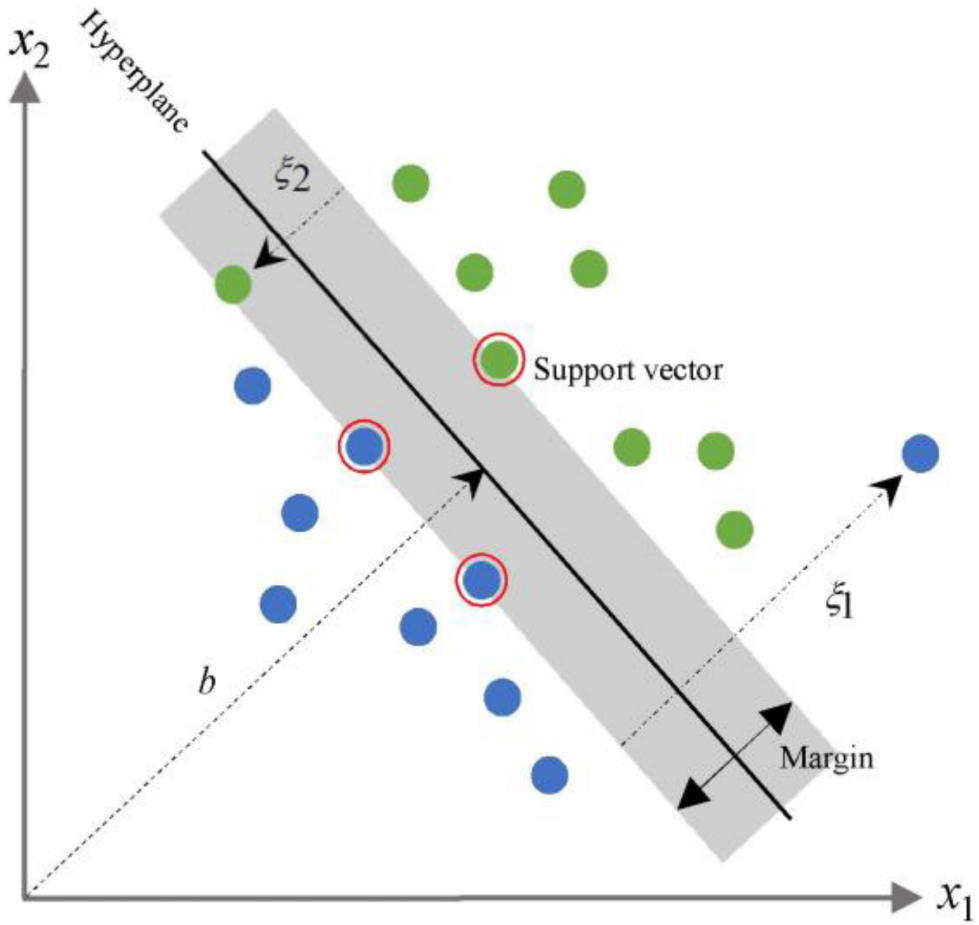
- Finding unknown patterns in the data set
- Distinguishing similar and dissimilar objects
- Helping for visualization in lower dimensions
- Recognizing odd objects (unknown unknowns)

$$\text{DATA} \xrightarrow{ML} \text{Knowledge}$$

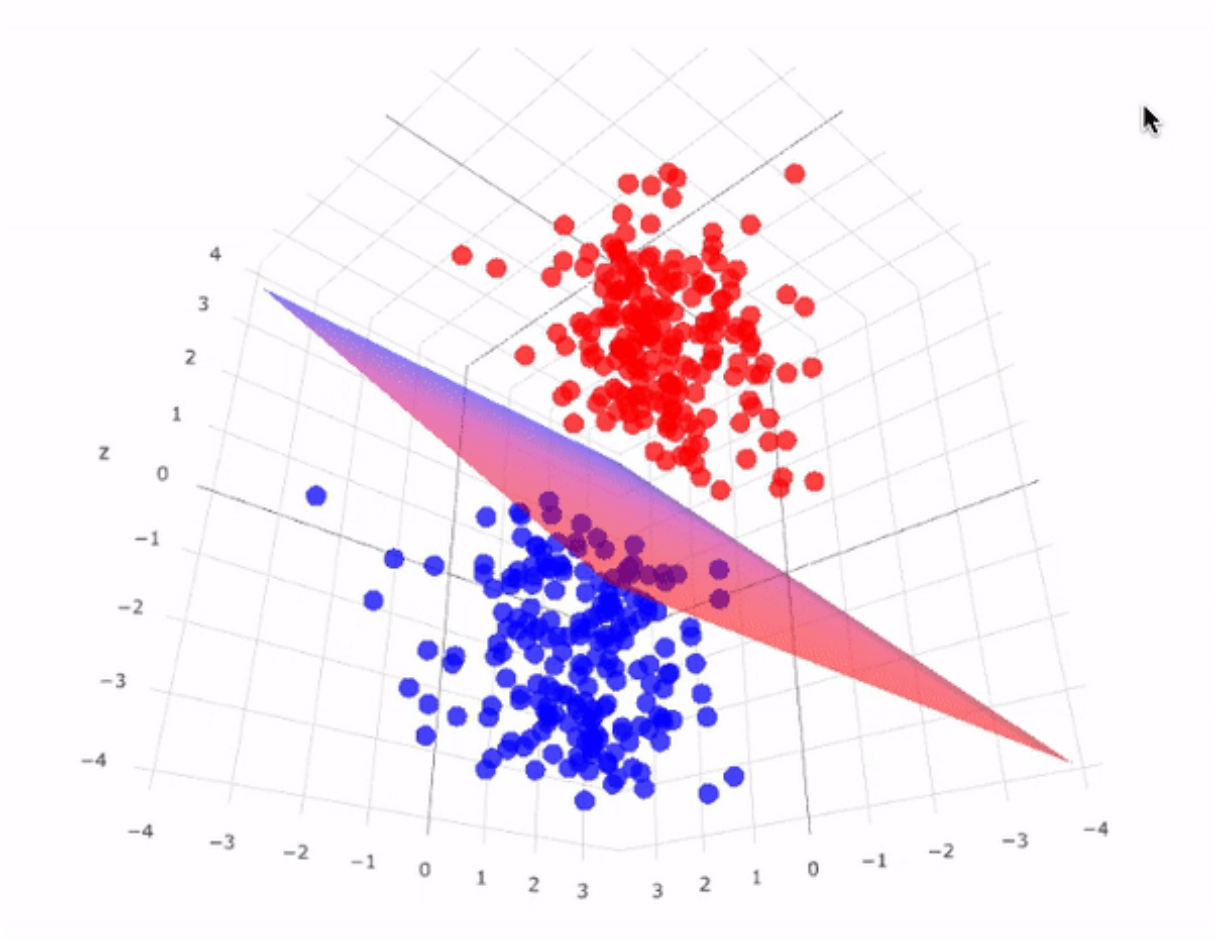
- Classification

- Classification
- Regression

How Support Vector Machine Works?



Hyper-plane in Support Vector Machine



Classifying Pre-Main-Sequence Stars using SVM

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

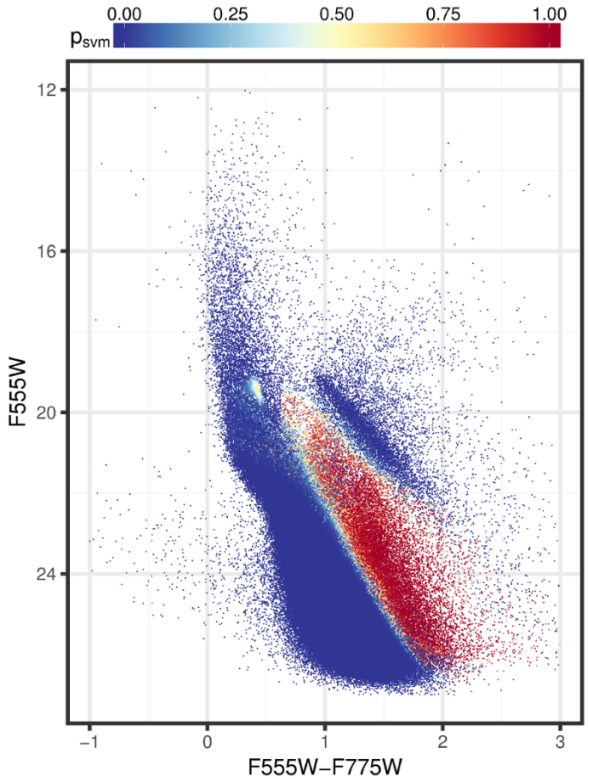
Victor F. Ksoll,[1,2,⋆] Dimitrios A. Gouliermis,[1,3,†] Ralf S. Klessen,[1] Eva K. Grebel,[4] Elena Sabbi,[5] Jay Anderson,[5] Daniel J. Lennon,[6] Michele Cignoni,[7] Guido de Marchi,[8] Linda J. Smith,[9] Monica Tosi,[10] and Roeland P. van der Marel[5]

[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*

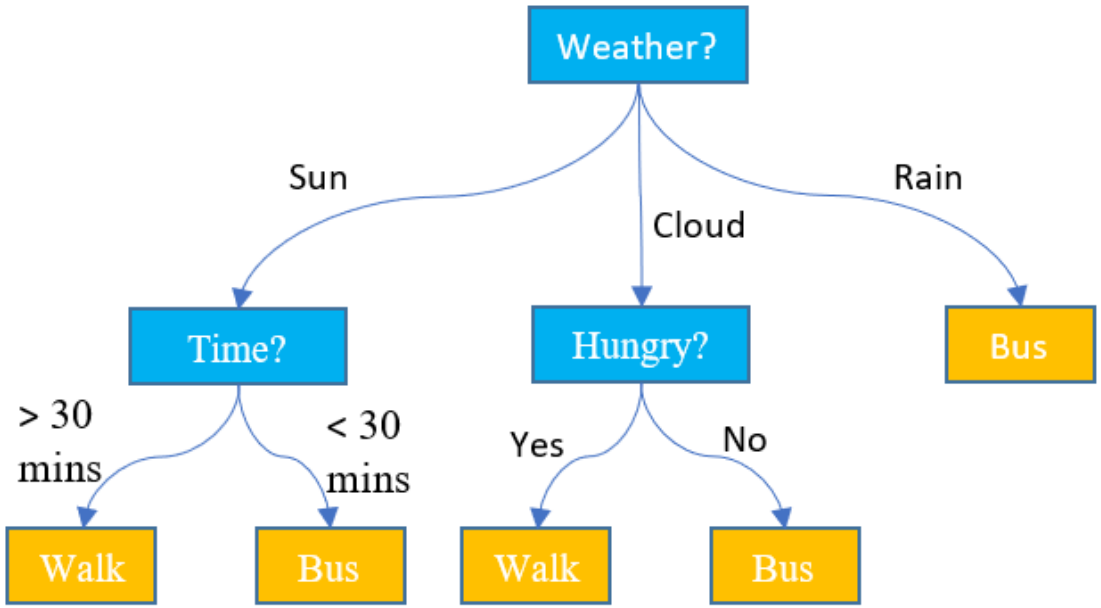
Draft version 21 May 2018

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



How Decision Tree works?



Using DT for classifying galaxies and stars in SDSS?

THE ASTRONOMICAL JOURNAL, 141:189 (12pp), 2011 June doi:10.1088/0004-6256/141/6/189

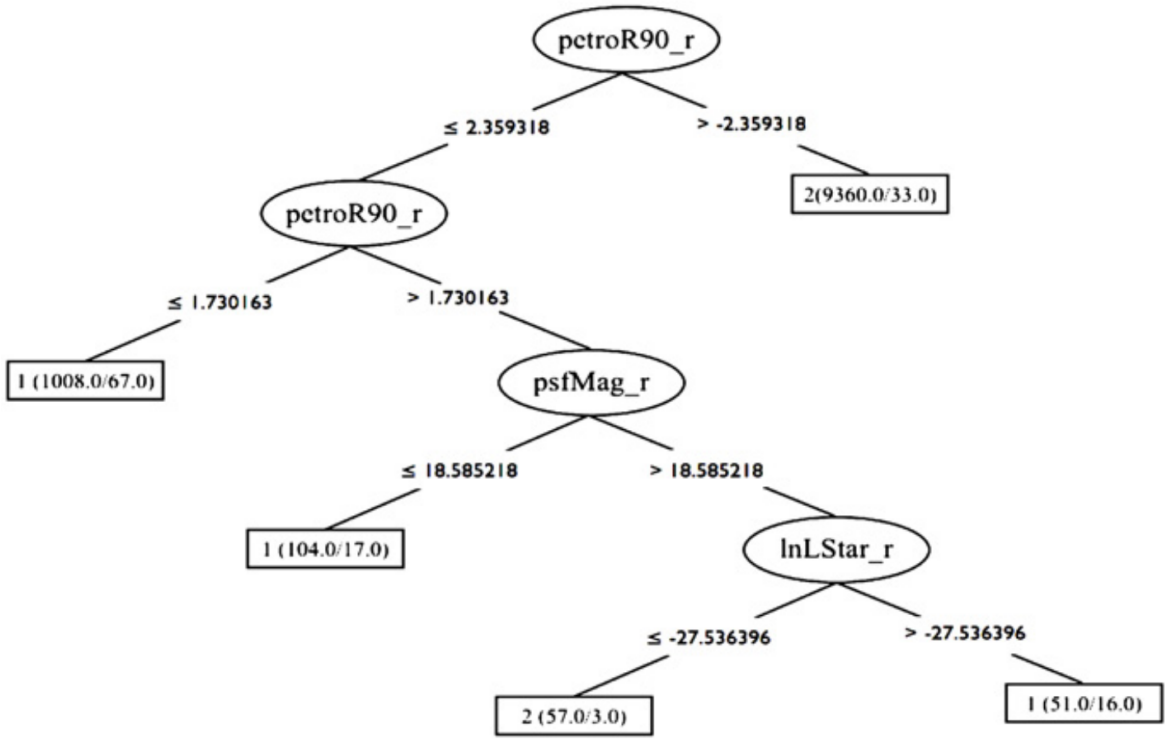
DECISION TREE CLASSIFIERS FOR STAR/GALAXY SEPARATION

E. C. VASCONCELLOS[1], R. R. DE CARVALHO[2], R. R. GAL[3], F. L. LABARBERA[4],
H. V. CAPELATO[2], H. FRAGO CAMPOS VELHO[5], M. TREVISAN[6], AND R. S. R. RUIZ[1]
[1] CAP, National Institute of Space Research, Av. dos Astronautas 1758, São José dos Campos 12227-010, Brazil
[2] DAS, National Institute of Space Research, Av. dos Astronautas 1758, São José dos Campos 12227-010, Brazil
[3] Institute for Astronomy, University of Hawaii, 2680 Woodlawn Dr., Honolulu, HI 96822, USA
[4] INAF-Osservatorio Astronomico di Capodimonte, via Moiariello 16, Napoli 80131, Italy
[5] LAC, National Institute of Space Research, Av. dos Astronautas 1758, São José dos Campos 12227-010, Brazil
[6] IAG, University of São Paulo, Rua do Matão 1226, São Paulo 05508-090, Brazil
Received 2010 October 27; accepted 2011 February 11; published 2011 May 9

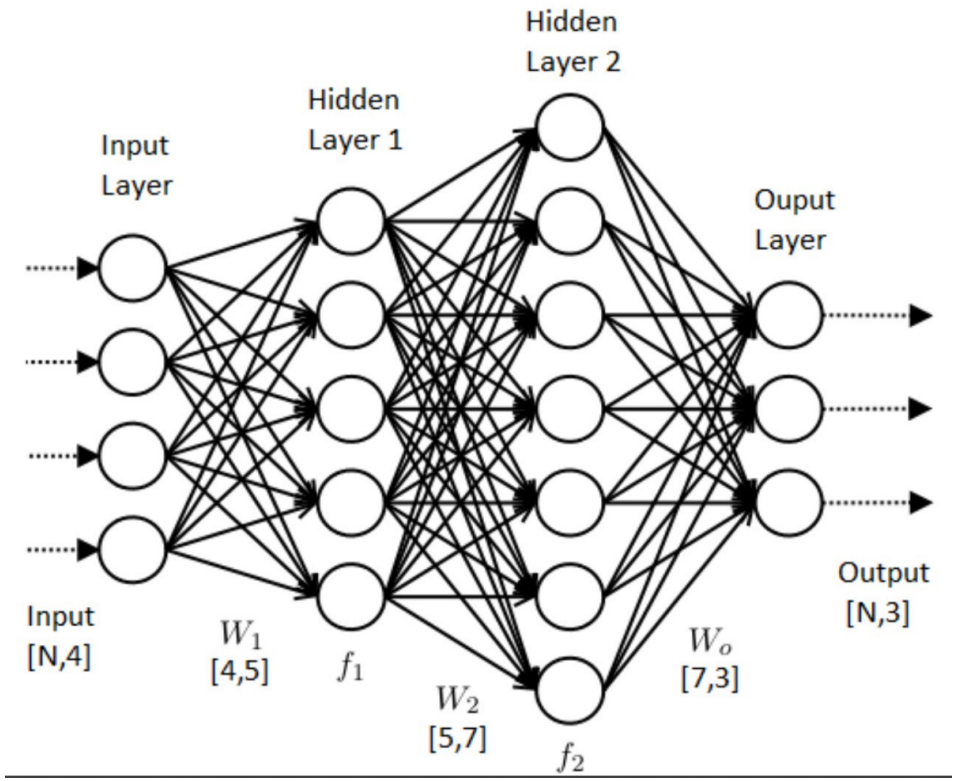
ABSTRACT

We study the star/galaxy classification efficiency of 13 different decision tree algorithms applied to photometric objects in the Sloan Digital Sky Survey Data Release Seven (SDSS-DR7). Each algorithm is defined by a set of parameters which, when varied, produce different final classification trees. We extensively explore the parameter space of each algorithm, using the set of 884,126 SDSS objects with spectroscopic data as the training set. The efficiency of star–galaxy separation is measured using the completeness function. We find that the Functional Tree algorithm (FT) yields the best results as measured by the mean completeness in two magnitude intervals: $14 \leqslant r \leqslant 21$ (85.2%) and $r \geqslant 19$ (82.1%). We compare the performance of the tree generated with the optimal FT configuration to the classifications provided by the SDSS parametric classifier, 2DPHOT, and Ball et al. We find that our FT classifier is comparable to or better in completeness over the full magnitude range $15 \leqslant r \leqslant 21$, with much lower contamination than all but the Ball et al. classifier. At the faintest magnitudes ($r > 19$), our classifier is the only one that maintains high completeness (>80%) while simultaneously achieving low contamination (∼2.5%). We also examine the SDSS parametric classifier (psfMag − modelMag) to see if the dividing line between stars and galaxies can be adjusted to improve the classifier. We find that currently stars in close pairs are often misclassified as galaxies, and suggest a new cut to improve the classifier. Finally, we apply our FT classifier to separate stars from galaxies in the full set of 69,545,326 SDSS photometric objects in the magnitude range $14 \leqslant r \leqslant 21$.

Using DT for classifying galaxies and stars in SDSS?



How Neural Network works?



Convolutional Neural Network and DLAs

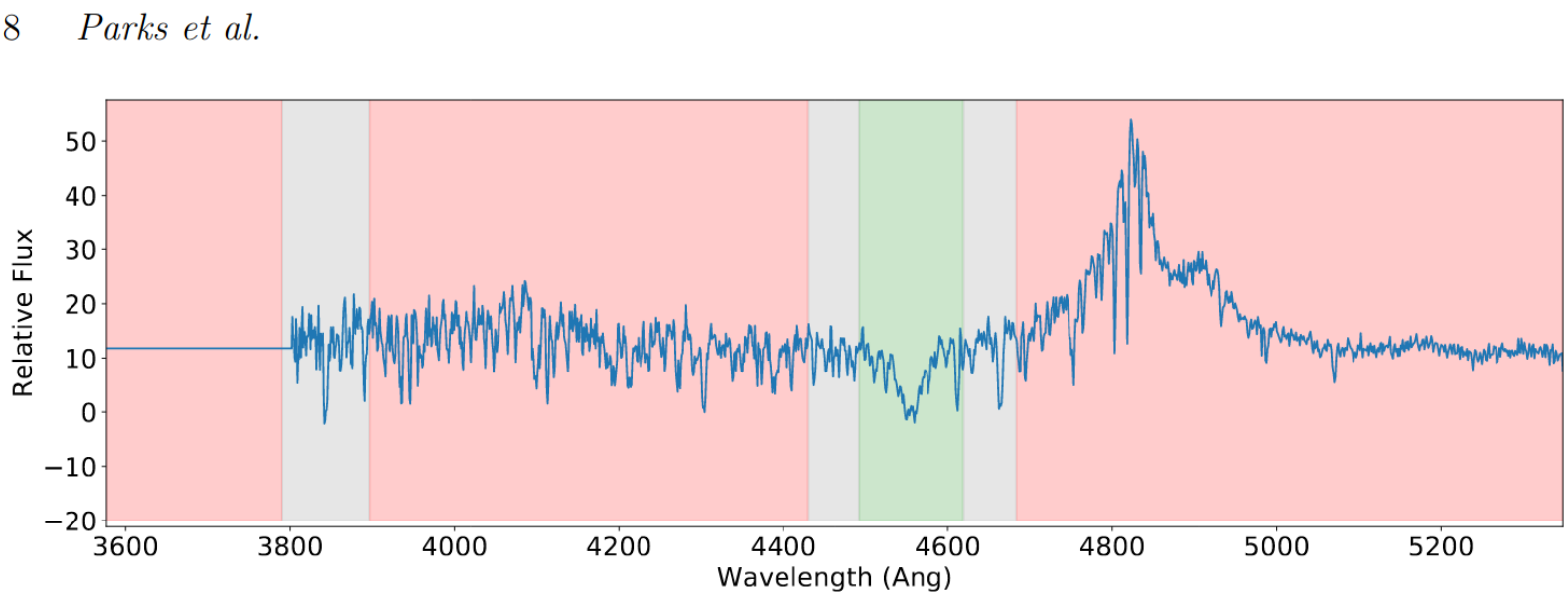
Deep Learning of Quasar Spectra to Discover and Characterize Damped Lyα Systems

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*

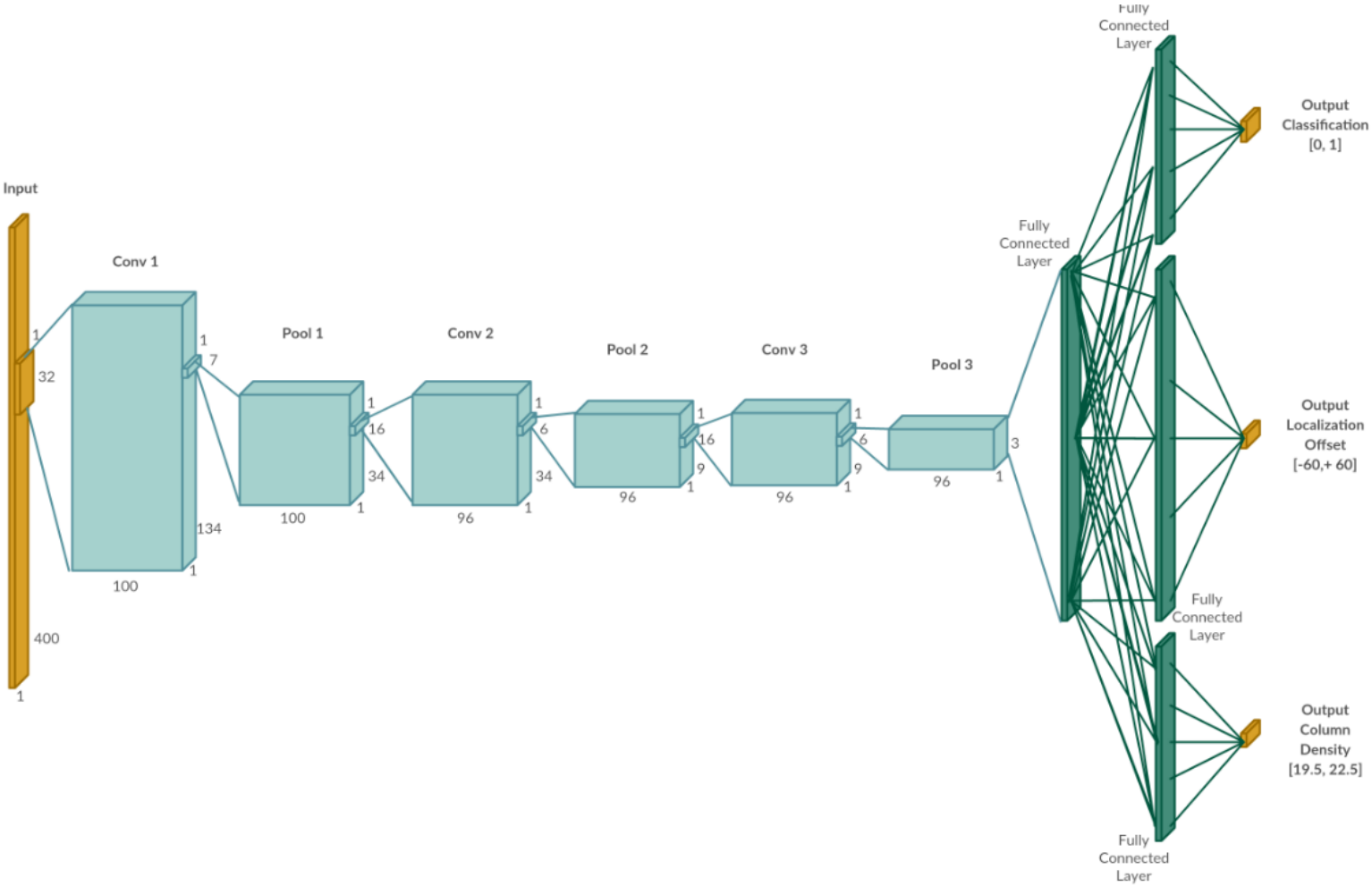
Accepted XXX. Received YYY; in original form ZZZ

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 DLAs/no DLAs



Treating DLA as a 1D image



text

- ML algorithms are not designed for astronomy but mostly for business.

- ML algorithms are not designed for astronomy but mostly for business.
- Training on a biased sample is very dangerous

- ML algorithms are not designed for astronomy but mostly for business.
- Training on a biased sample is very dangerous
- ML is the best for how only cares about results