Detecting M Giants in Space Using XGBoost

Dr. Zesheng Chen
Department of Computer Science
Purdue University Fort Wayne

PURDUE UNIVERSITY FORT WAYNE

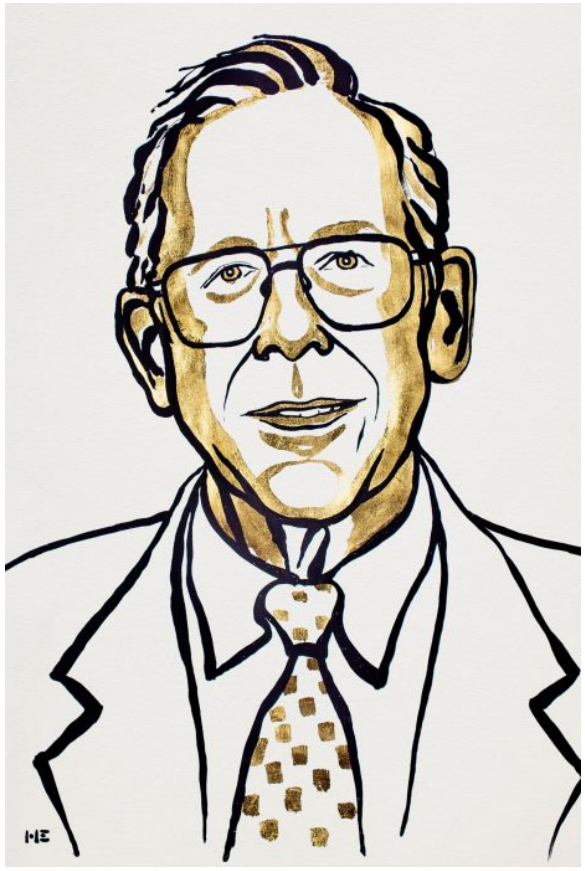
An Efficient Spectral Selection of M Giants Using XGBoost

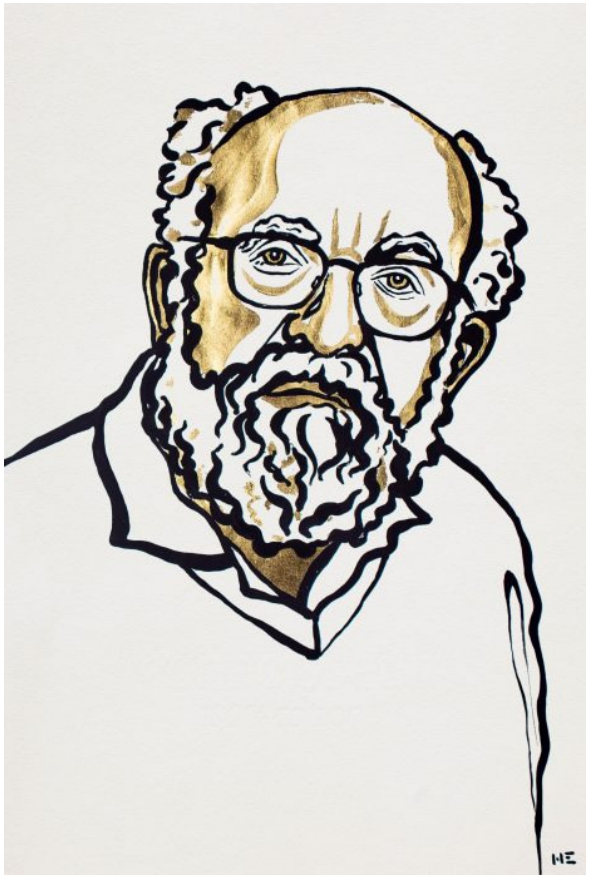
ZHENPING YI,[1] ZESHENG CHEN,[2] JINGCHANG PAN,[1] LILI YUE,[1] YUXIANG LU,[1] JIA LI,[1] AND A-LI LUO[3]

[1] *School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, 264209, Shandong, China*
[2] *Department of Computer Science, Purdue University Fort Wayne, Fort Wayne, IN 46805 , USA*
[3] *Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China*

PURDUE UNIVERSITY FORT WAYNE

The Nobel Prize in Physics 2019





PURDUE
UNIVERSITY
FORT WAYNE

PURDUE UNIVERSITY FORT WAYNE

M Giants

- Red giants with spectral type M
- Lower surface temperature (≤ 4000K)
- Extremely bright with typical luminosities of $10^3 L\odot$

- M giants provide a way for researchers to explore the substructures of the halo of the Milky Way

PURDU
UNIVERSI
FORT WAYN

Red giants map how the Milky W

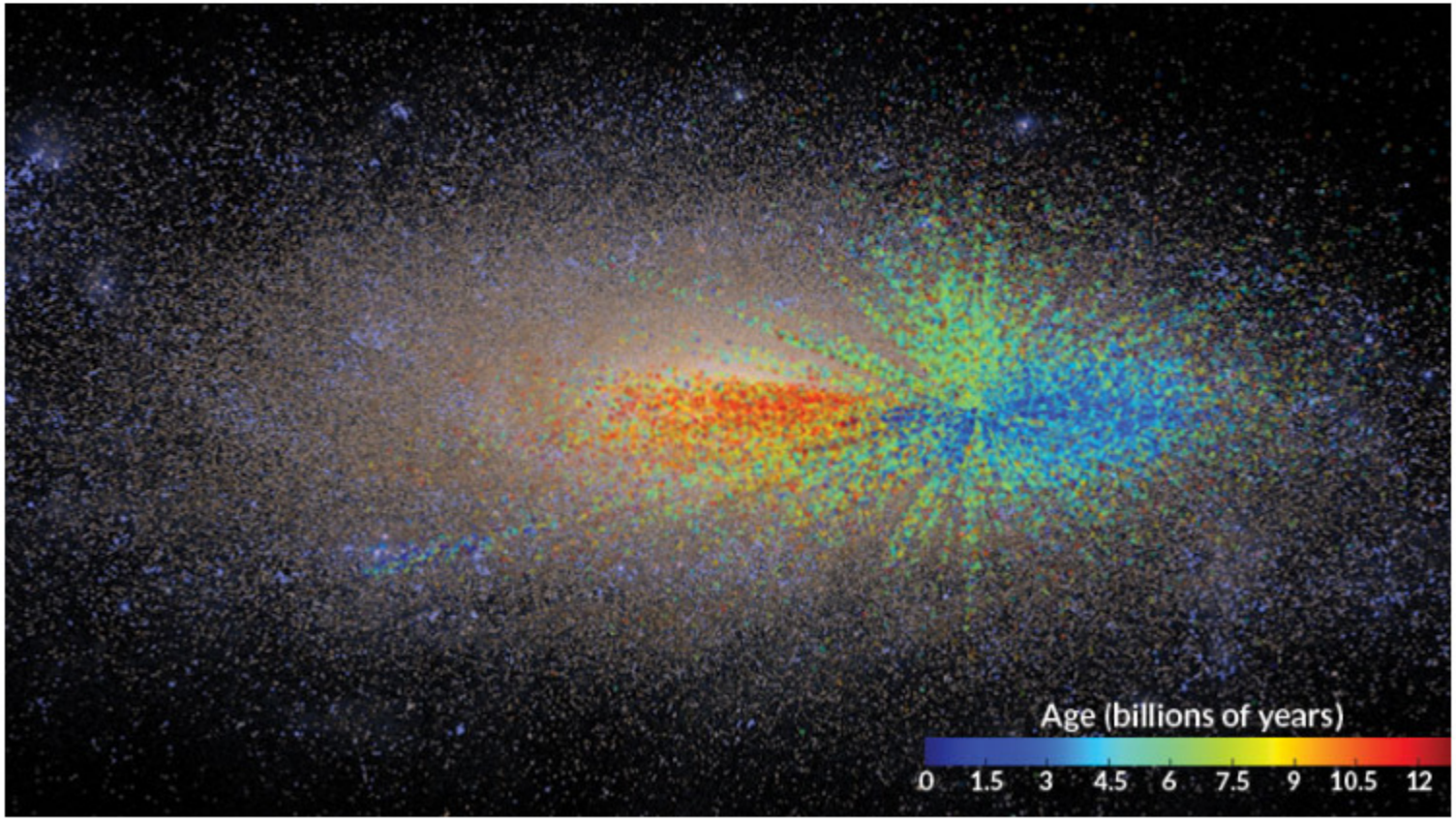
sciencenews.org/article/red-giants-map-how-milky-way-gr...

Apps | Project List (Fall 201... | Master List | Machine learning te... | Course schedule

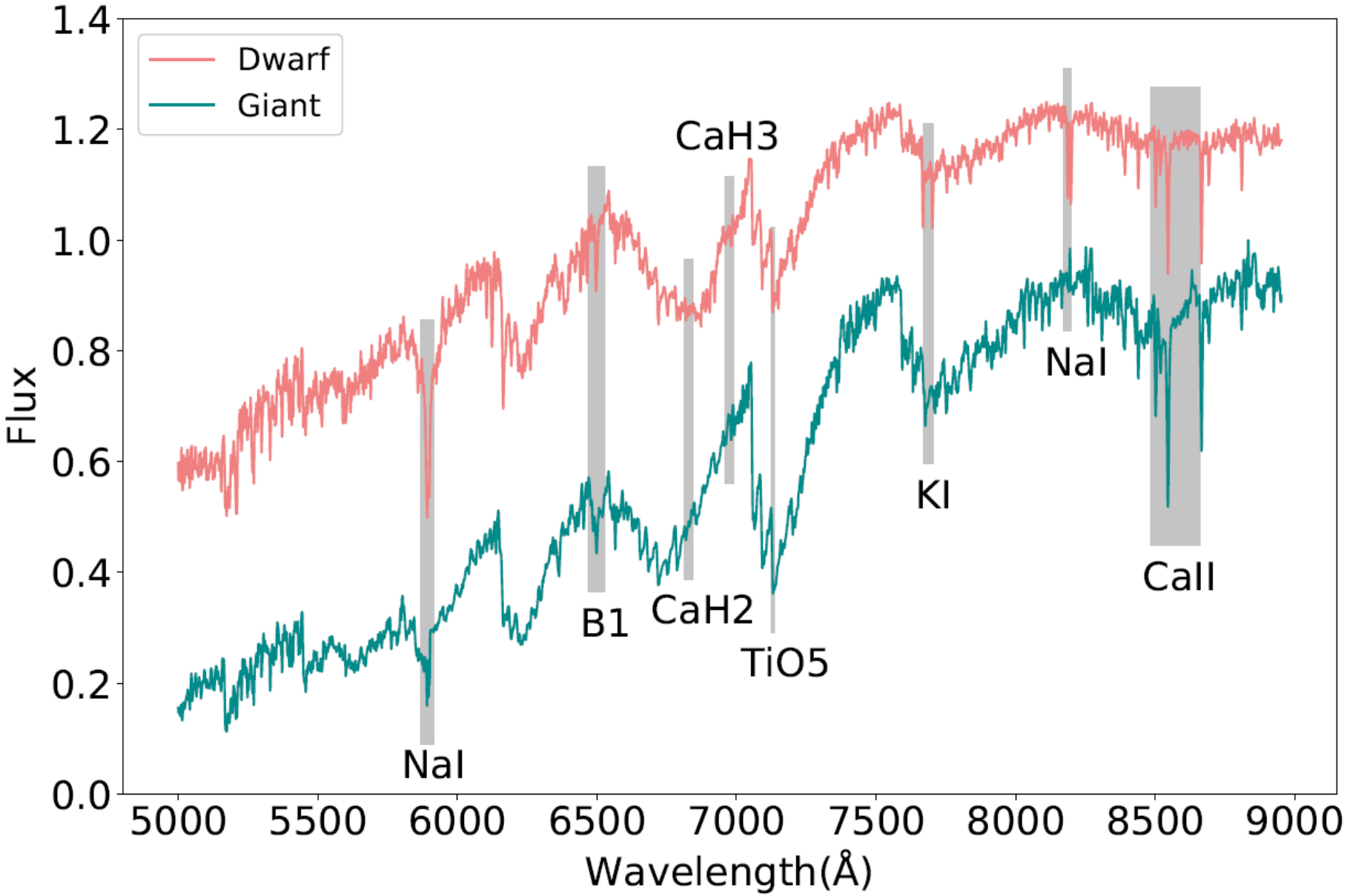
NEWS ASTRONOMY

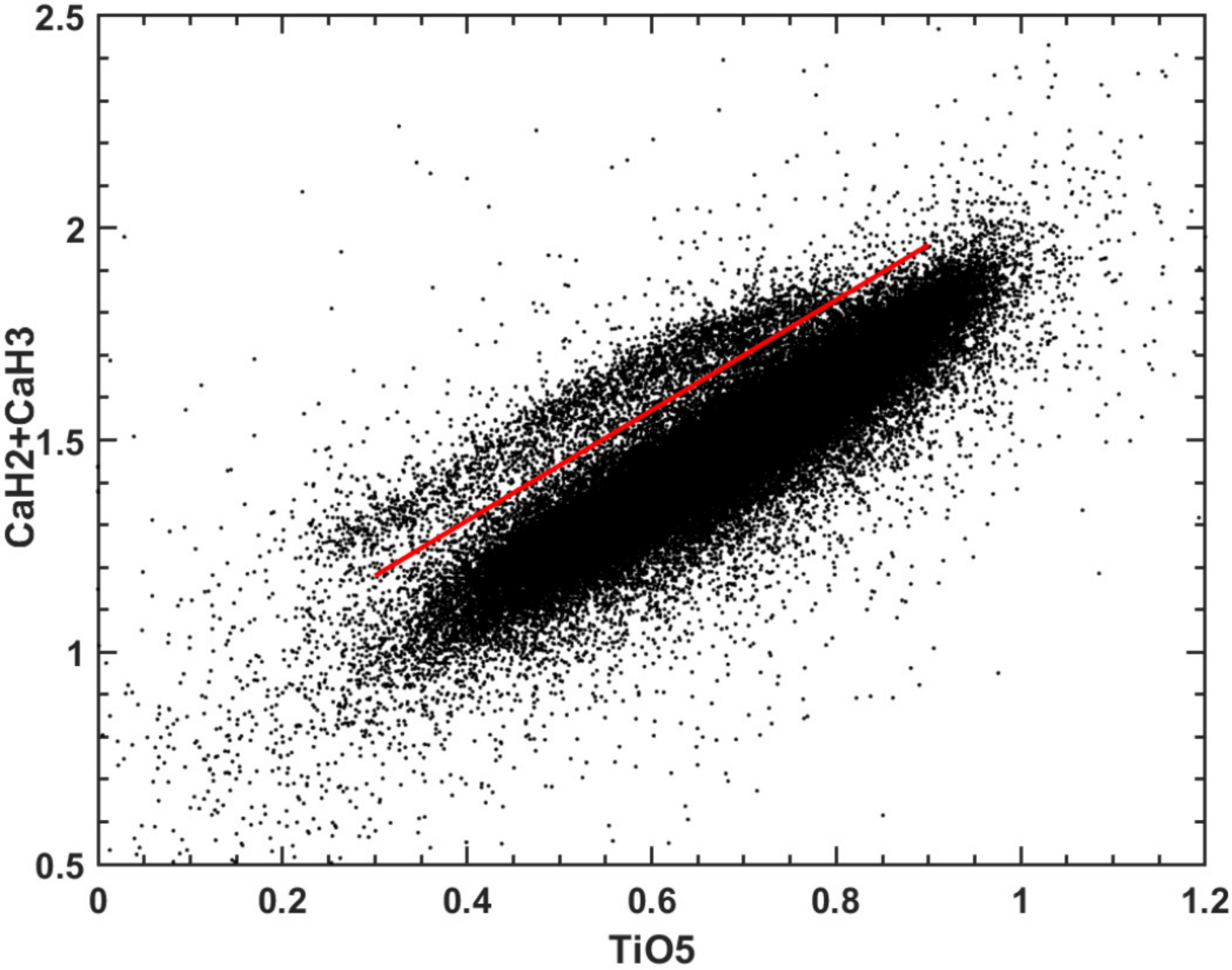
Red giants map how the Milky Way grew

Measuring the mass of 70,000 aging stars reveals older center, younger outskirts



AGE MAP The ages of tens of thousands of red giant stars are charted atop a map of the Milky Way. The oldest stars are in red, near the galactic center. The youngest stars are blue.


Dwarf
Giant
CaH3
NaI
B1
CaH2
TiO5
KI
NaI
CaII
Flux
Wavelength(Å)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
5000
5500
6000
6500
7000
7500
8000
8500
9000


2.5
2
1.5
1
0.5
CaH2+CaH3
0
0.2
0.4
0.6
0.8
1
1.2
TiO5

PURDUE UNIVERSITY FORT WAYNE

Outline

- Data
- XGBoost
- Results

PURDUE UNIVERSITY FORT WAYNE

Data

- LAMOST DR4 data
- LAMOST is a new type of wide-field telescopes with a large aperture and a large field of view
- Currently, LAMOST DR4 has released 7.68 million spectra
- We used 6,311 M giant spectra and 5,883 M dwarf spectra, with labels
- We randomly selected about 70% as the training data

PURDUE UNIVERSITY FORT WAYNE

XGBoost

- Extreme Gradient Boosting
- A scalable machine learning system for tree boosting
- An open source package
- Widely recognized in many machine learning and data mining challenges (e.g., Kaggle)

- Use slides from “Introduction to Boosted Trees” by Tianqi Chen

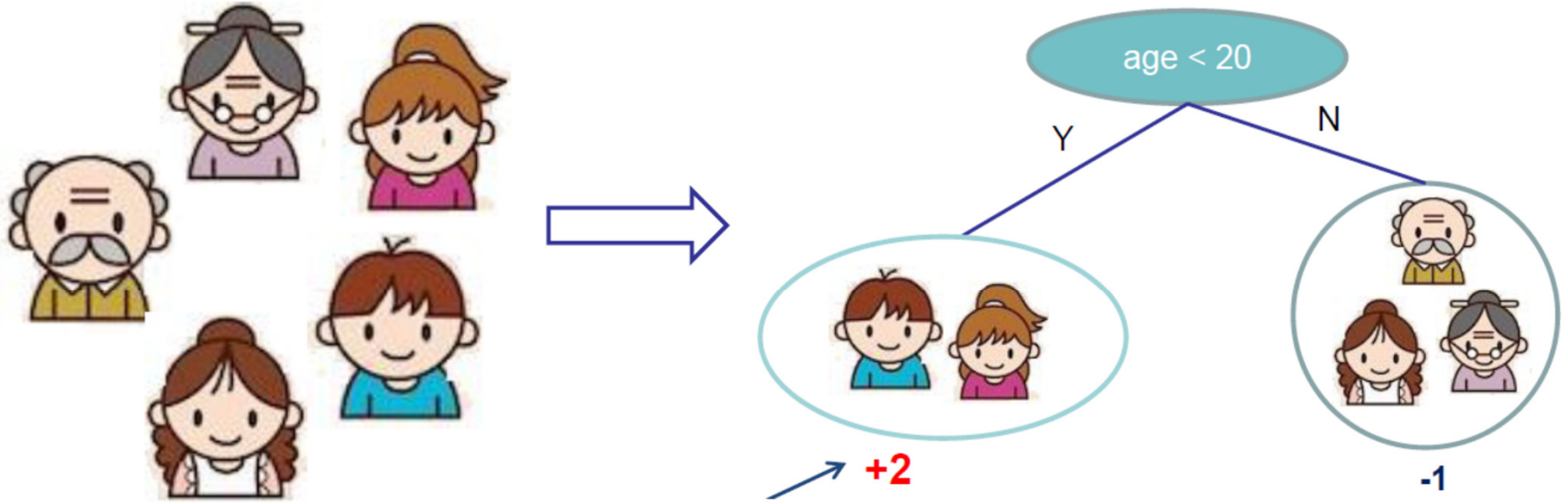
PURDUE UNIVERSITY FORT WAYNE

Classification and Regression Tree (CART)

- Decision rules same as in decision tree
- Contains one score in each leaf value

Input: age, gender, occupation, …

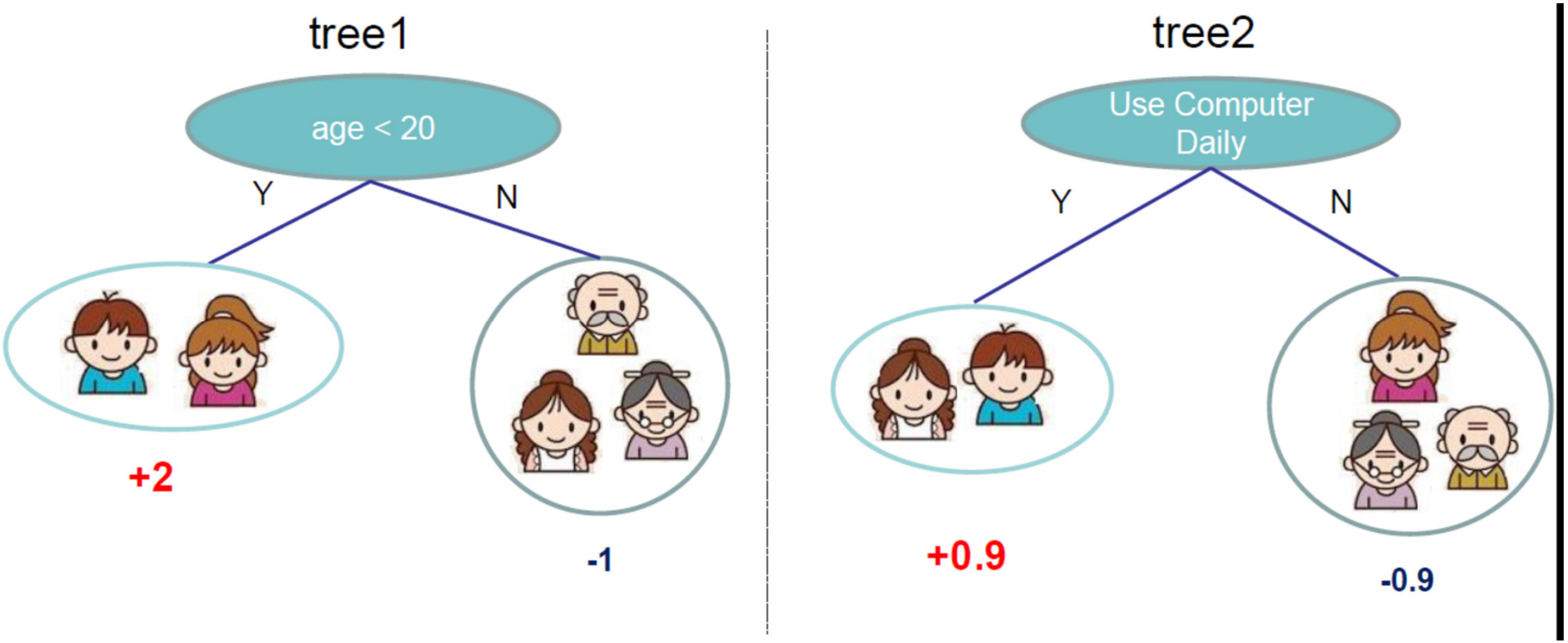
Like the computer game X



prediction score in each leaf

PURDUE UNIVERSITY FORT WAYNE

Regress Tree Ensemble



f() = 2 + 0.9= 2.9

f()= -1 - 0.9= -1.9

Prediction score is the sum of scores predicted by each of the tree.

PURDUE UNIVERSITY FORT WAYNE

Objective and Bias-Variance Trade-off

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

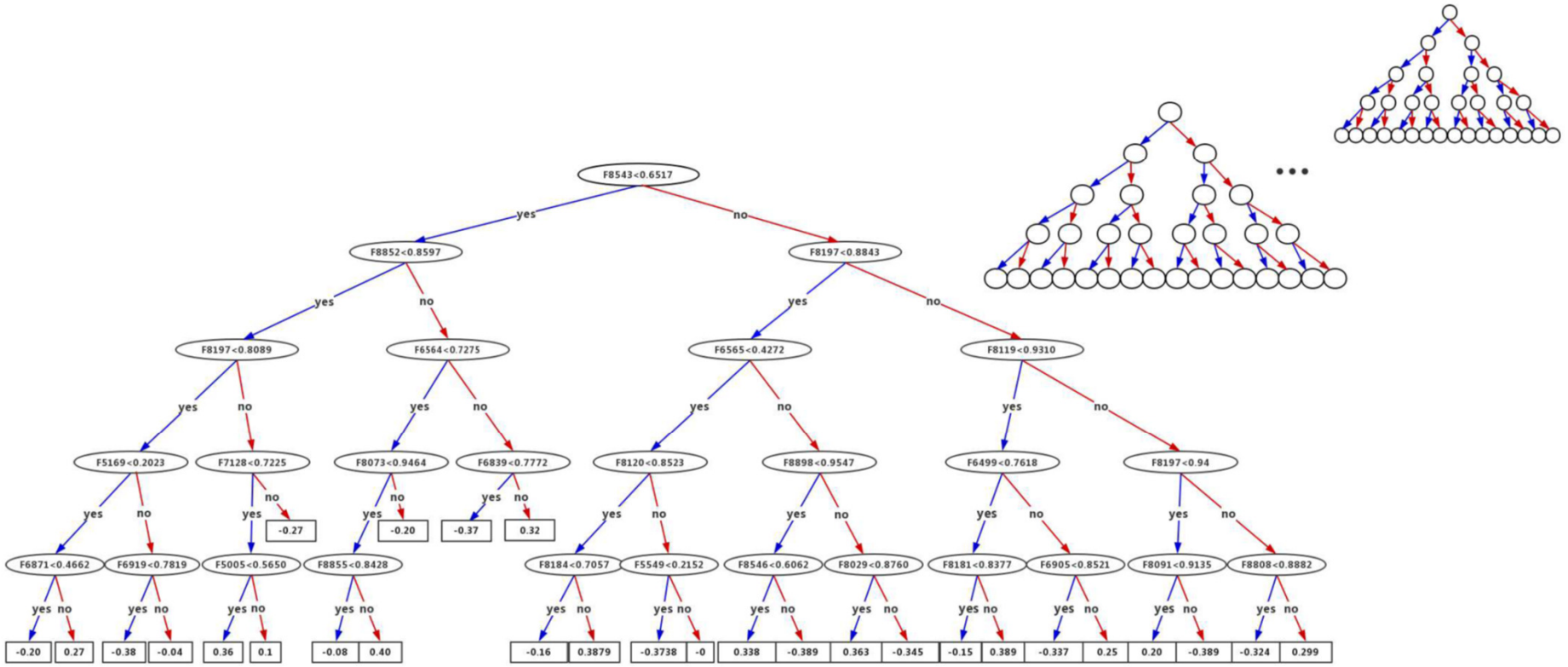
Training Loss measures how well model fit on training data

Regularization, measures complexity of model

- Why do we want to contain two components in the objective?
- Optimizing training loss encourages predictive models
  - Fitting well in training data at least get you close to training data which is hopefully close to the underlying distribution
- Optimizing regularization encourages simple models
  - Simpler models tends to have smaller variance in future predictions, making prediction stable

PURDUE UNIVERSITY FORT WAYNE

XGBoost Classifier



PURDUE UNIVERSITY FORT WAYNE

Shallow Learning vs. Deep Learning

- Shallow learning algorithms learn the parameters of a model directly from the features of training samples and build a structurally understandable model
- We focus on shallow learning to identify most important features to separate M giants from M dwarfs

PURDUE UNIVERSITY FORT WAYNE

Performance Comparison of Four Machine Learning Methods

Algorithm	Accuracy	Precision	Recall
XGBoost	99.79	96.87	98.93
SVM	99.53	92.08	98.94
Random forests	99.29	90.05	96.23
ELM	98.75	80.71	97.85

PURDUE UNIVERSITY FORT WAYNE

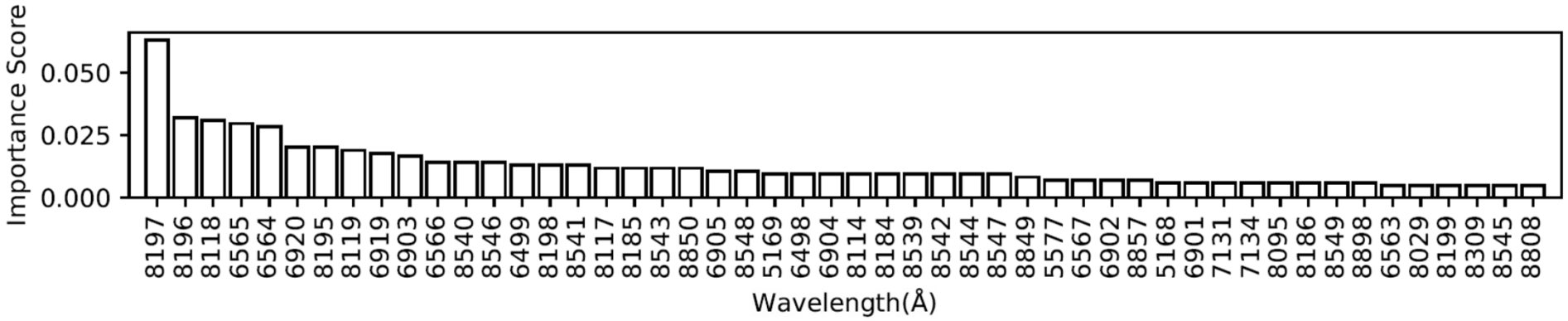
Important Features

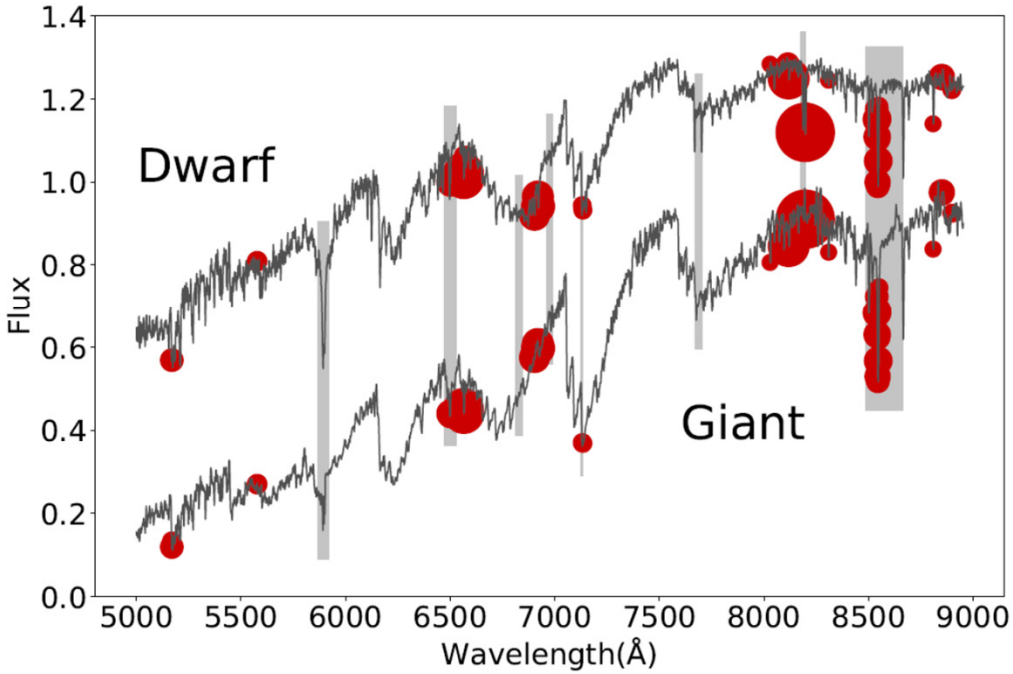
- We found that 287 features among 3,951 pixels of input data are used in XGBoost
- The more times a feature is used in XGBoost tree, the more important it is

$$IPS = N_{times}/SUM_{times}$$

PURDUE UNIVERSITY FORT WAYNE

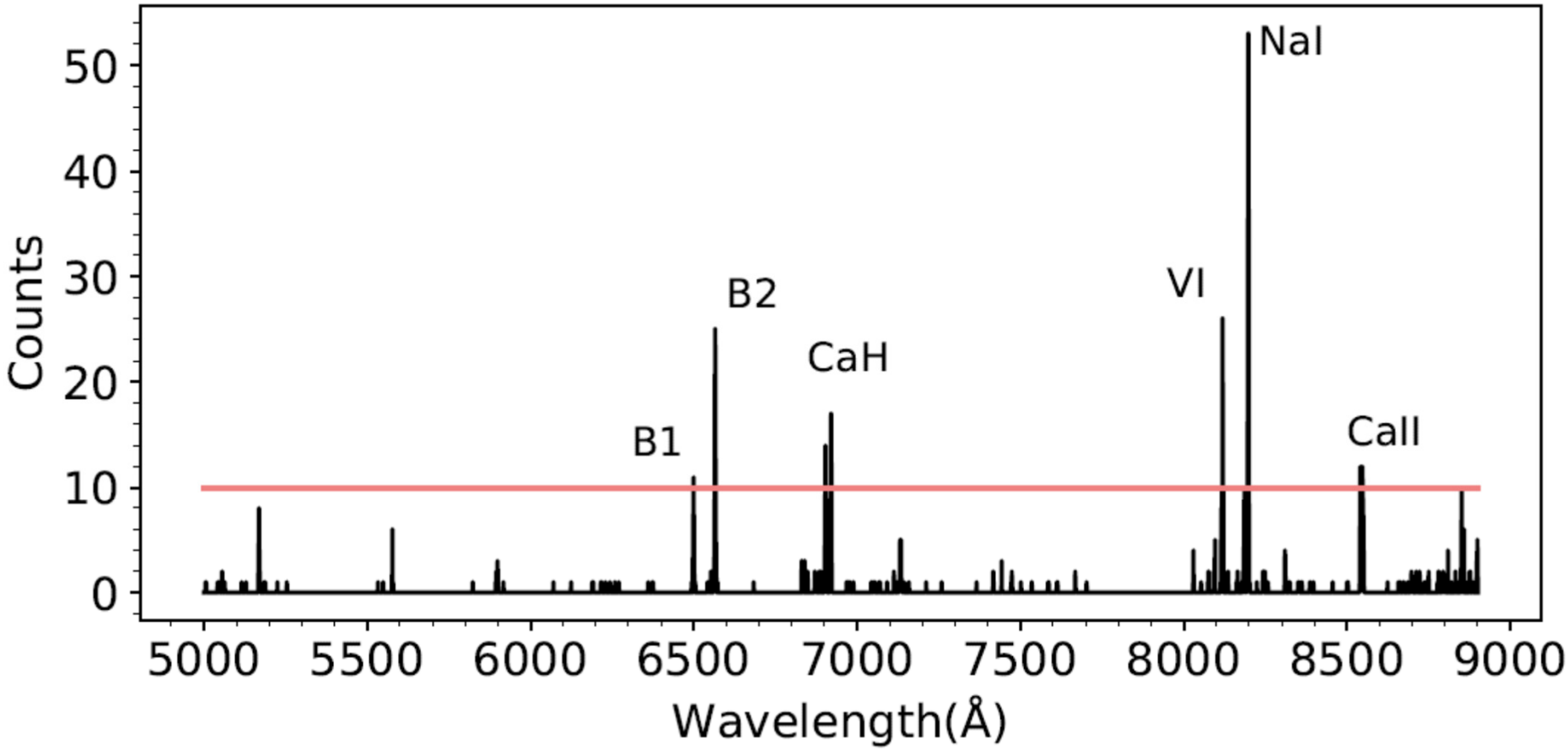
Important Features





PURDUE UNIVERSITY FORT WAYNE

Important Features



PURDUE UNIVERSITY FORT WAYNE

Conclusions

- XGBoost is used to discern M giants from M dwarfs for spectroscopic surveys
- The important feature bands for distinguishing between M giants and M dwarfs are accurately identified by the XGBoost method
- We think that our XGBoost classifier will perform effectively for other spectral surveys as well if the corresponding features wavelength bands are covered

PURDUE UNIVERSITY FORT WAYNE

Thanks For Your Attention