Active Deep Learning In LAMOST DR2 MegaSpectra Archive

The Crucial Role of VO in Modern ML

Petr Škoda
Astronomical Institute of the Czech Academy of Sciences Ondřejov
Faculty of Information Technology of the Czech Technical University in Prague

Ondřej Podsztavek
Faculty of Information Technology of the Czech Technical University in Prague

Supported by OP VVV, Research Center for Informatics,
CZ.02.1.01/0.0/0.0/16_019/0000765

IVOA Interoperability Meeting KDD IG
Paris Observatory, Paris, France
13-th May 2019

Thanks to

- Astronomical institute CAS- 2m
- LAMOST people (Dongwei Fan, Yue Wu)
- China-VO (Chenzhou Cui)

Be Stars

Gamma Cas (Padre Angello Sechhi 1866)

Vatican obs predecessor – visual spectrograph

Some have or have had emission in Balmer lines

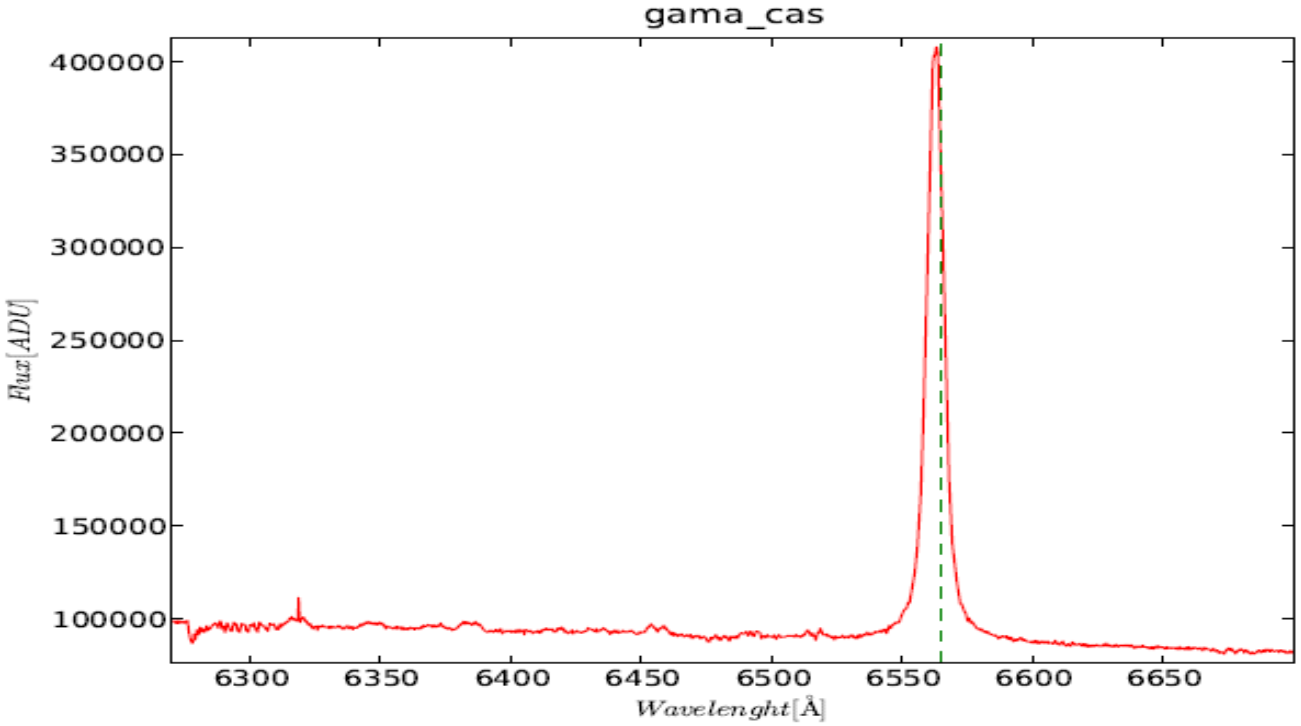
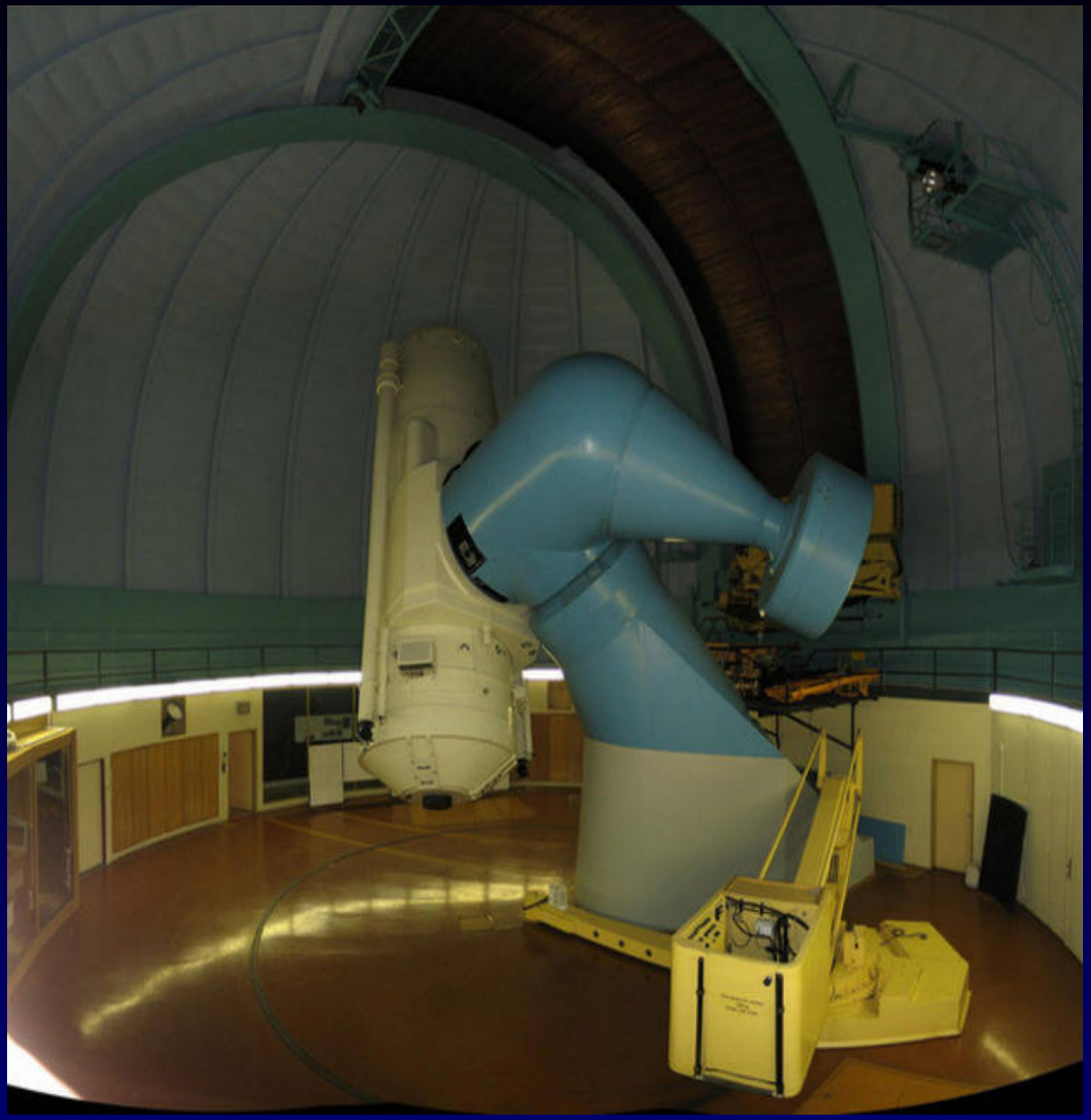


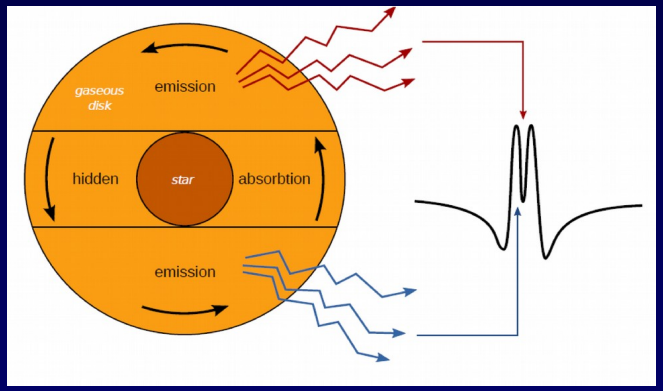
Figure 3.18: Spectrum of Gamma Cas. Be Star

Ondřejov 2m Perek Telescope (1967)

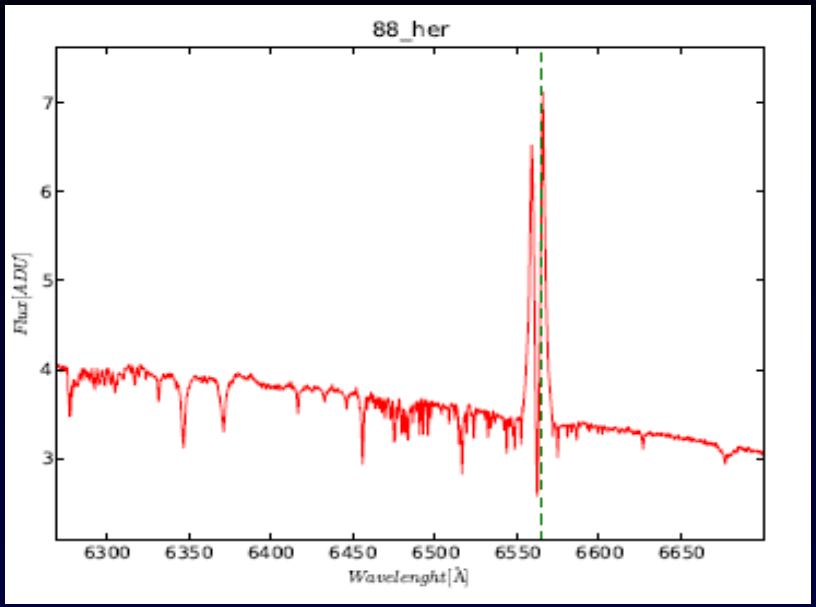


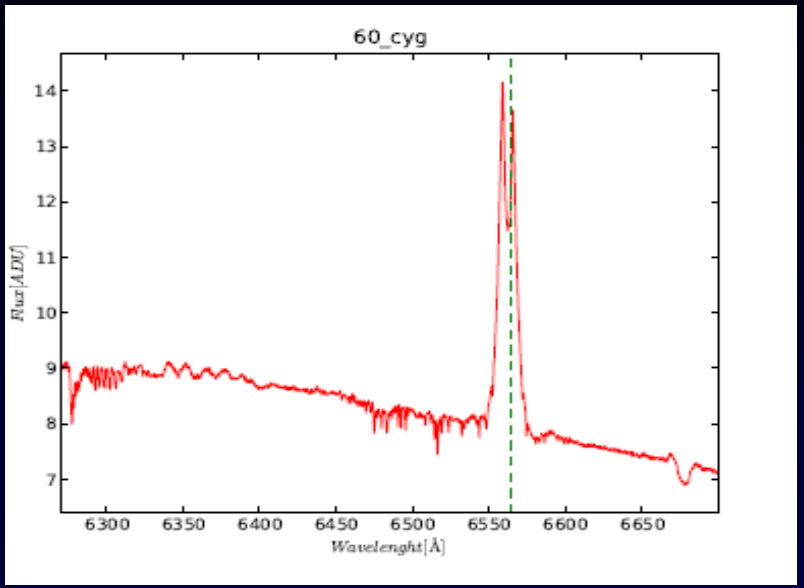
50 years tradition in Be stars
Archive over 20 000 spectra
13000 in Halpha region

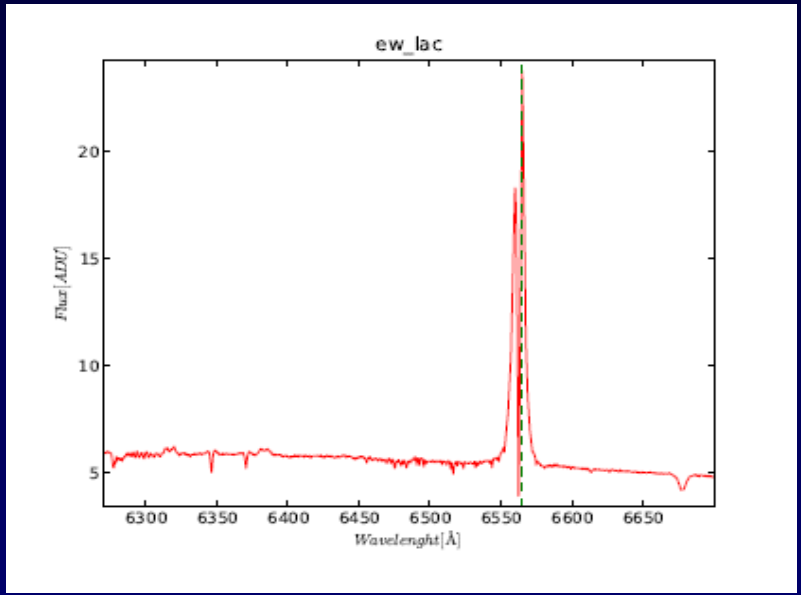
Be mysterious phenomenon
changes line profile
epizodes of emision
fast rotate
Hot (early B types)
disk or envelope

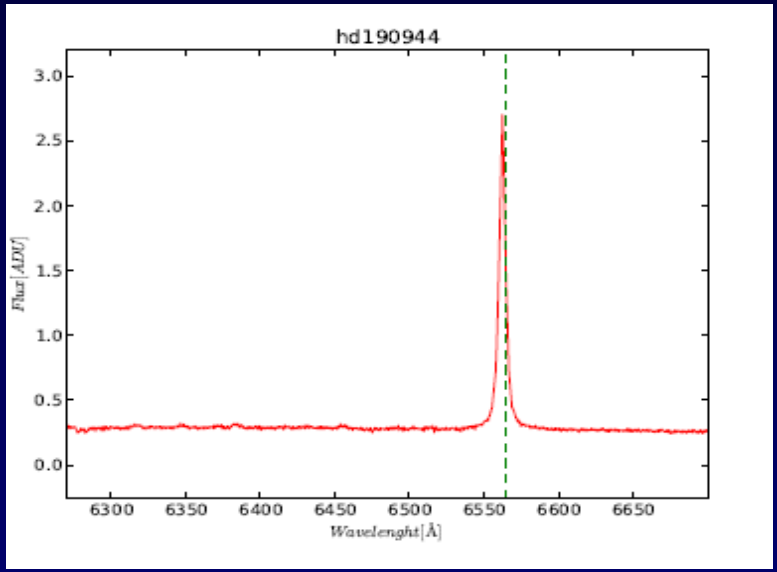


Be Stars : Shell lines vs. no absorption

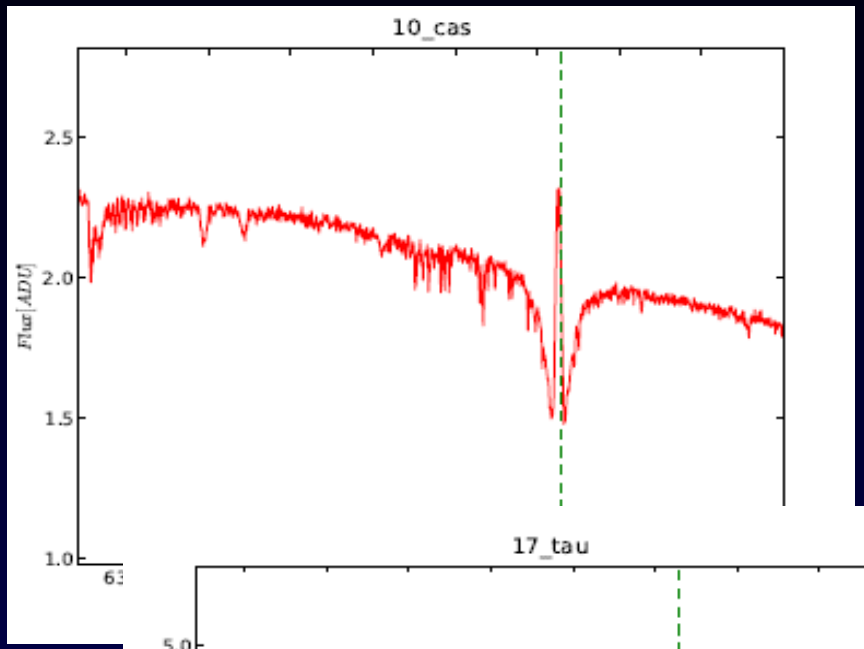


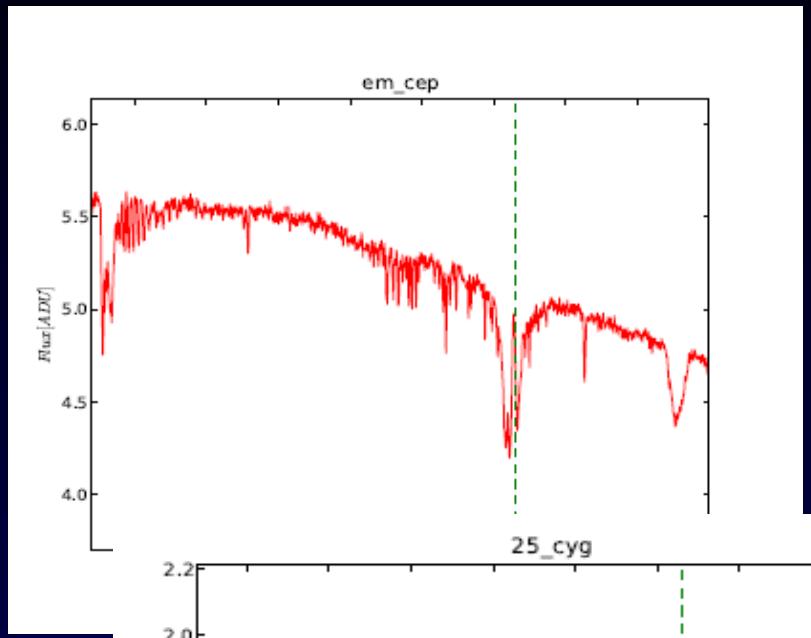


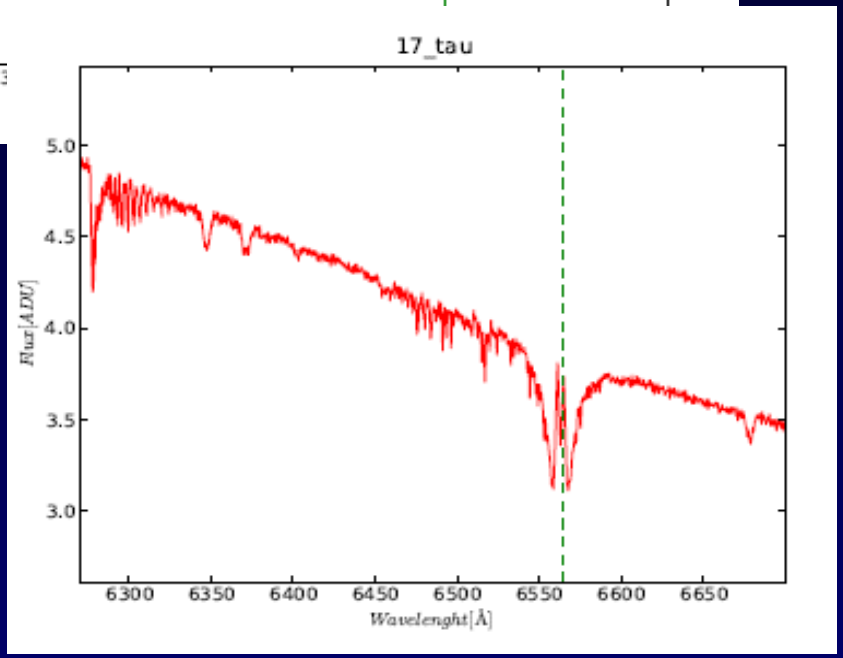


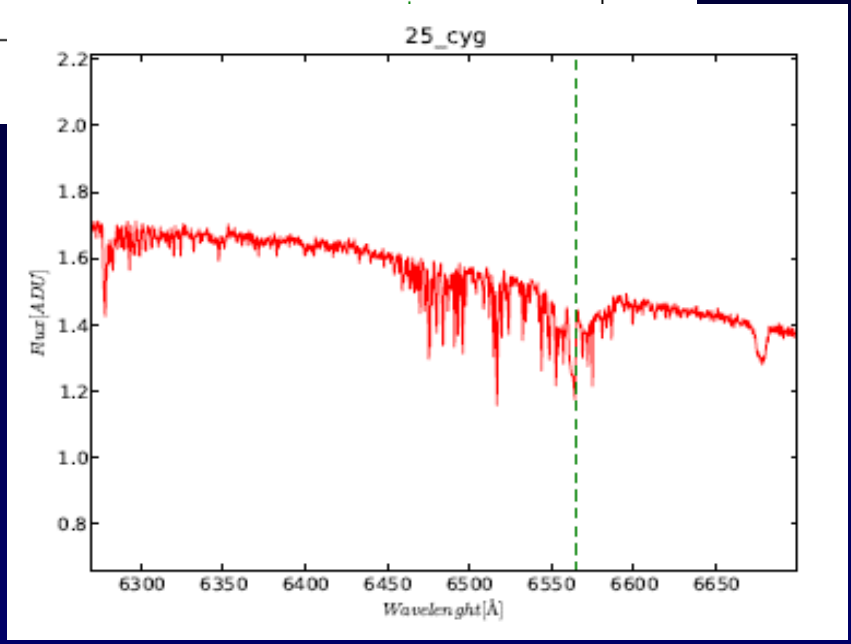


Be Stars : Emission in absorption



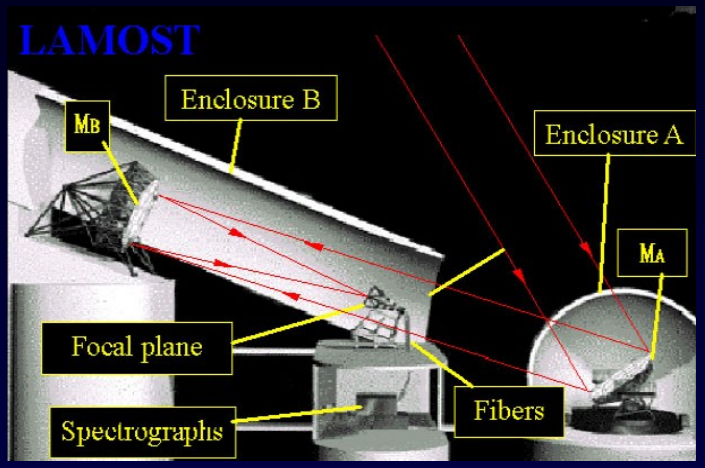


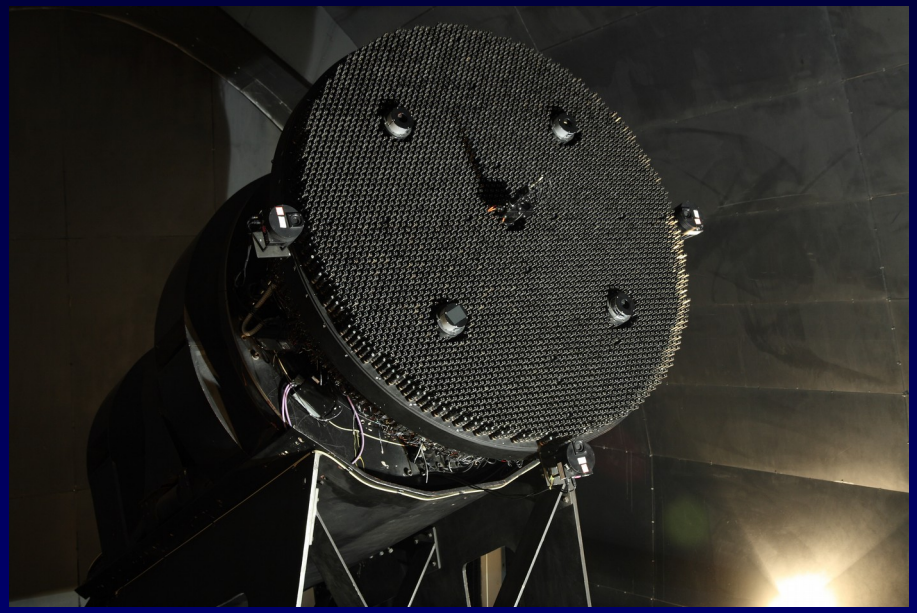




LAMOST (Guoshoujing)

Xinglong- China
4m mirror (30 deg meridian)
4000 fibers





LAMOST Spectral Surveys

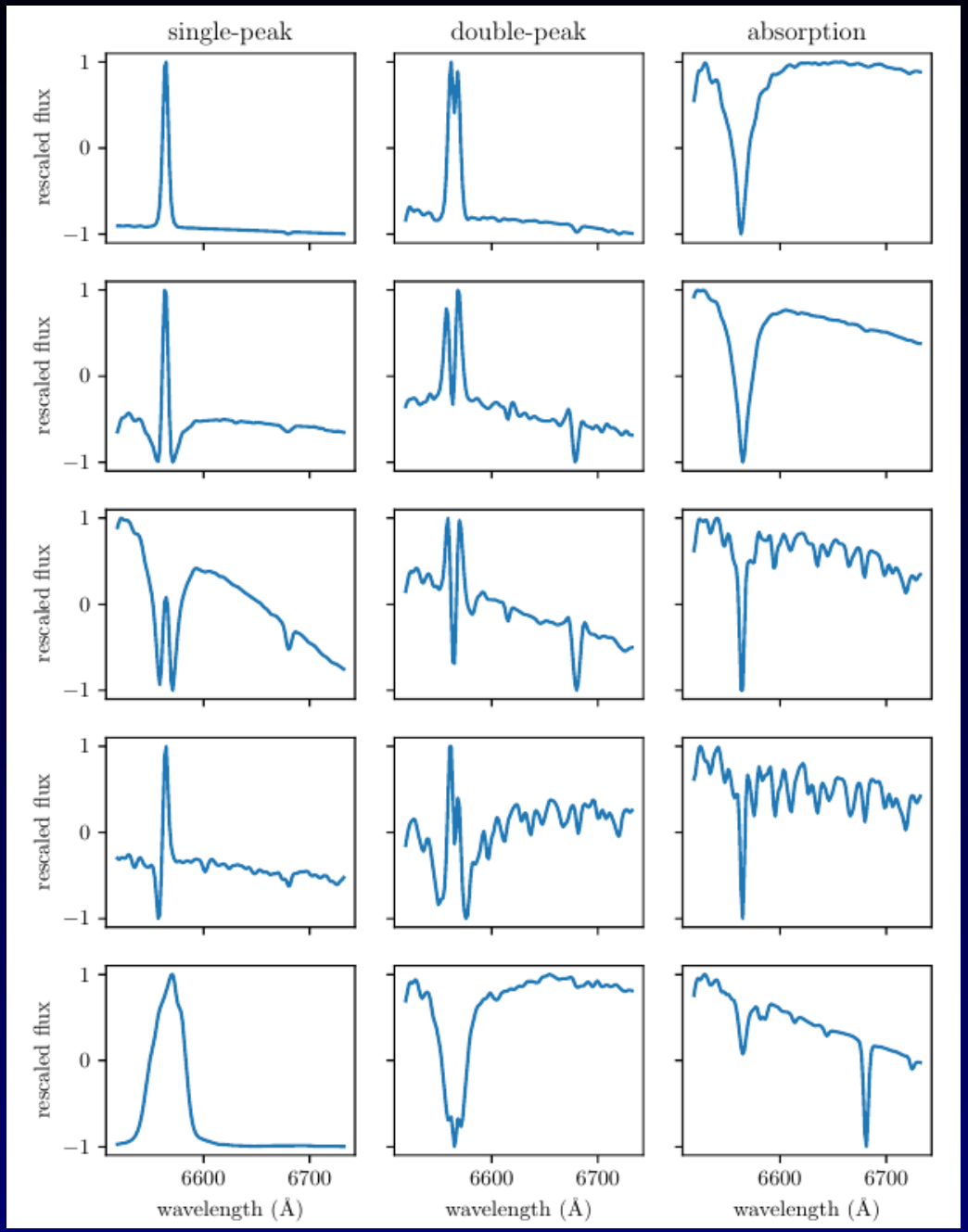
DR1 (end 2013) **2 204 860** spectra
1 085 404 stars classified by pipeline
DR2 (beg 2015) **4 132 782** spectra
3 779 674 stars
307 000 unknown !!!!!

DR5 (half 2017) **9 017 844** spectra
DR6 (half 2018) +739 006
+249 591 low res.
+ 3 508 695 mid res.

Each Fiber – 2 motors
double arm 33mm circle

Fibre collects light from
3.3 arcsec circle on sky

Ondřejov Data Classification



12936 spectra from CCD700

Our TARGET class only Be stars
=emission or double peak (2+1)

Still not enough labels for DL !

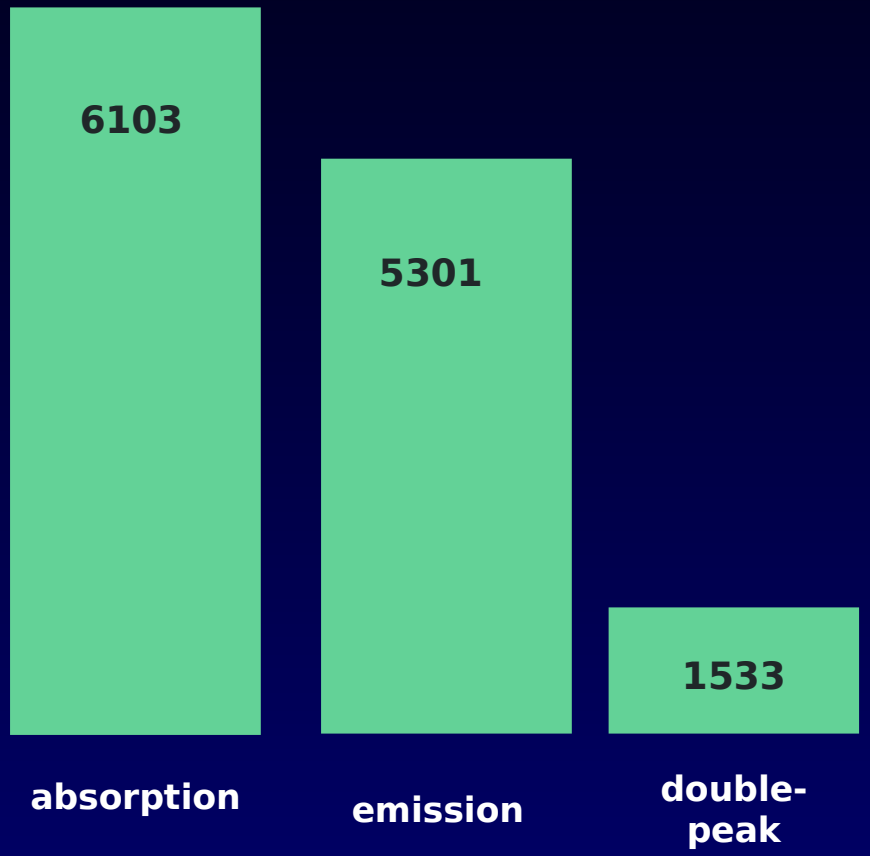
Ondrejov Data Set
https://zenodo.org/record/2640971

Balancing Classes in Training Set

Synthetic Minority Oversampling Technique

SMOTE

Bowyer et al. 2011



Common Spectra in Both Surveys

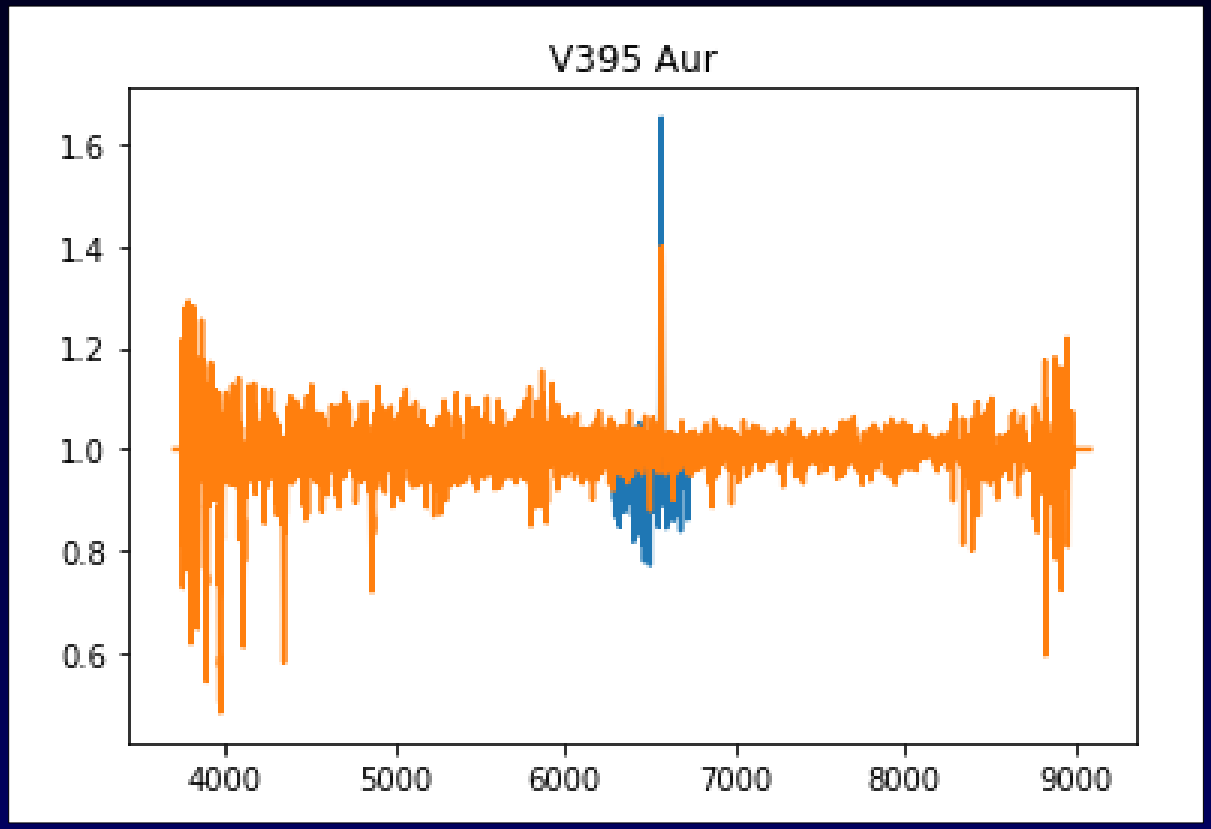
Needed cross-matching in VO (Topcat, SPLAT-VO)

LAMOST (4 mil) and Ondrejov (13000) on SSAP server (DaCHS) + DATALINK

ONLY 22 common (but different time of observation = changes in profile)

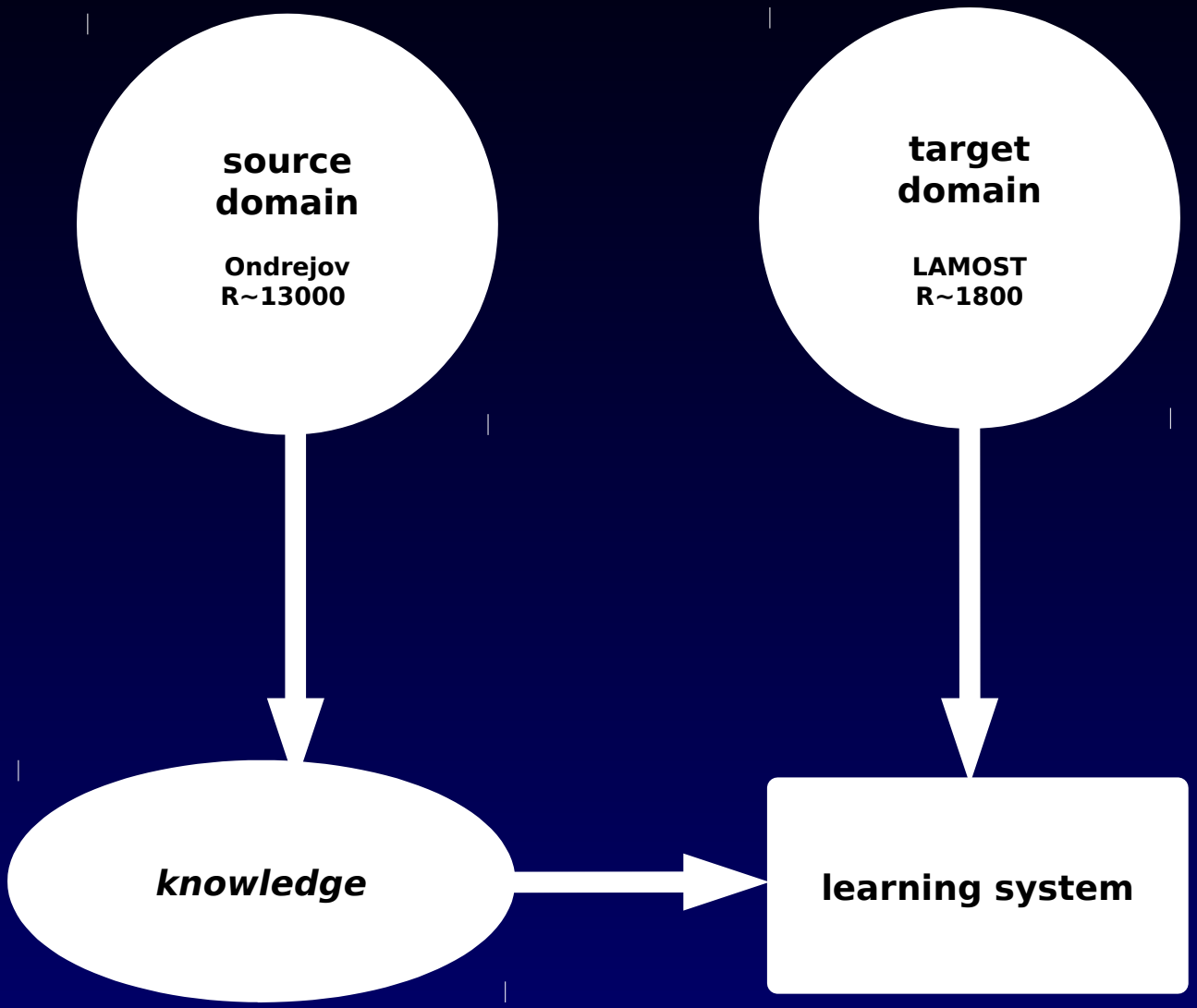
ONLY 4 emission line

No way to proper training set!

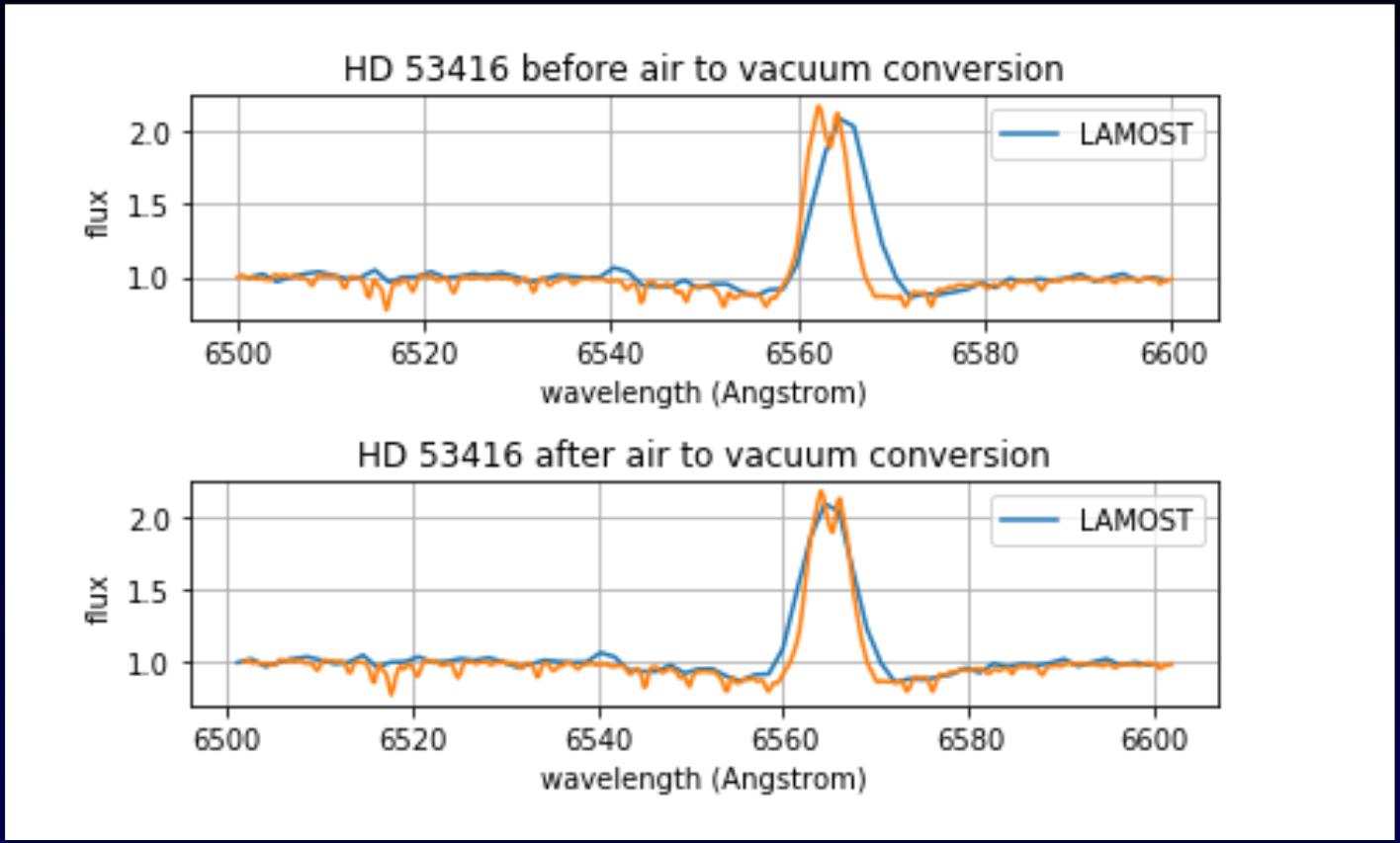


Domain Adaptation

Transfer Learning

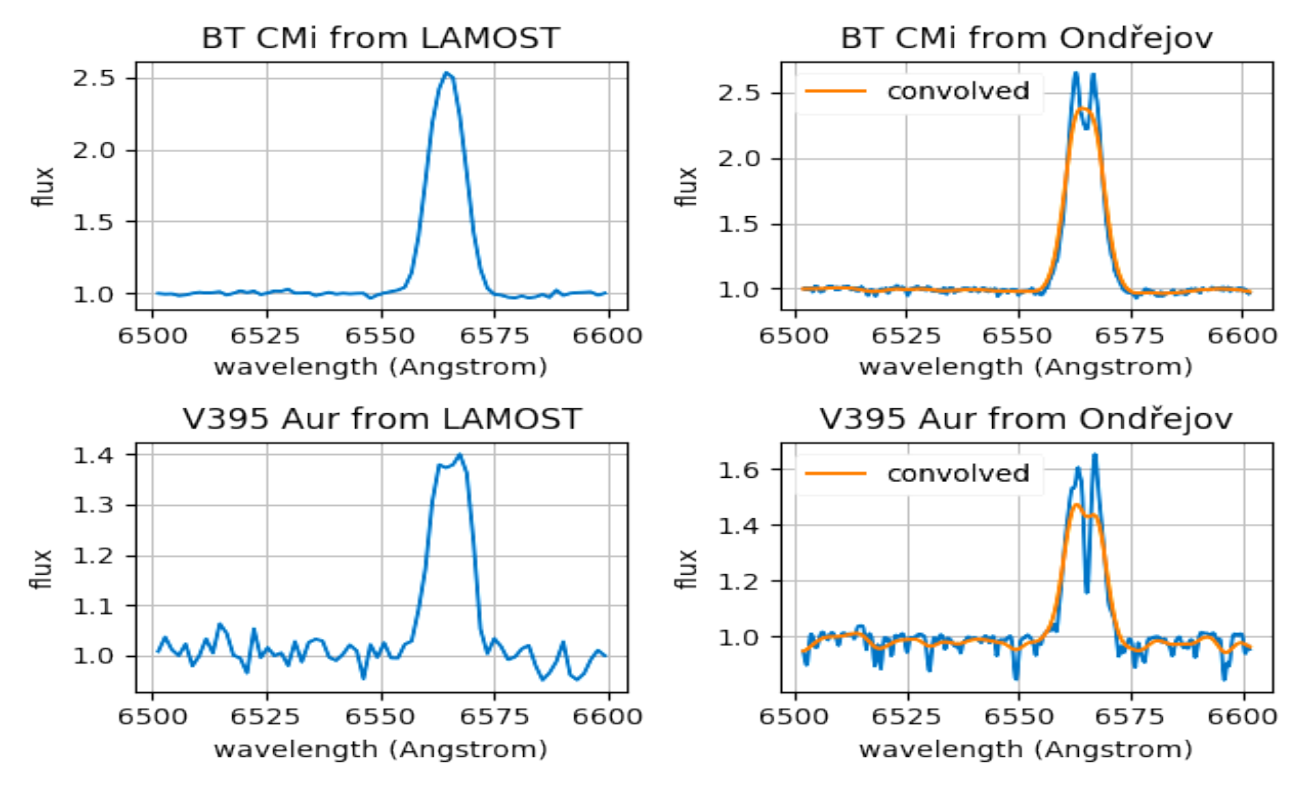


Domain Adaptation



Wavelength transform from air to vacuum

Domain Adaptation



Change of spectral resolution by Gaussian blur (13000 → 1800)
Rebinning to grid of 140 pixels in 6519 to 6732 A → FEATURE VECTOR

Normalisation od Data

Continuum normalisation – rectification (continuum on 1.0, limit at 0)

Ondřejov – automatic pipeline

experiments show that (at least for our Halpha region) the continuum normalisation is not needed (training on unrectified gives same validation accuracy 0.96)

Rescaling to zero mean, unit variance - common in ML

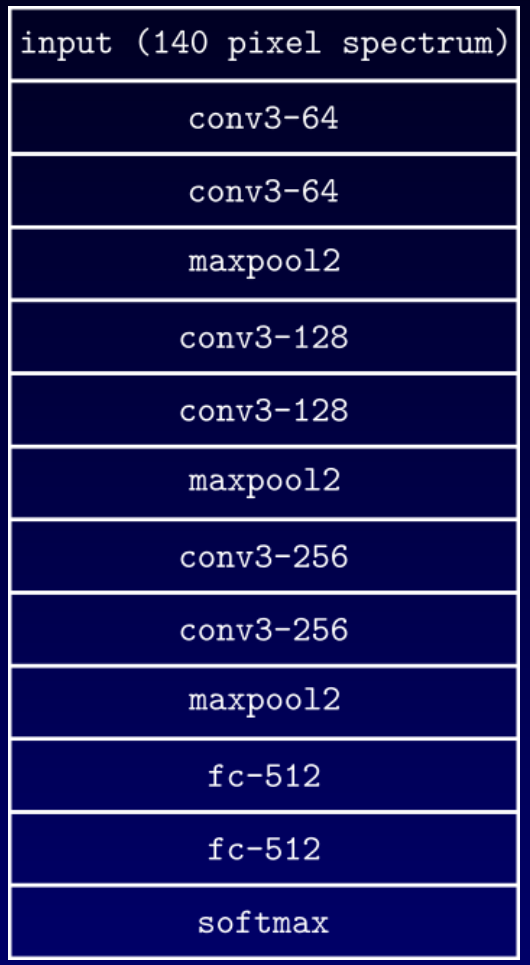
Deep Convolutional Net

deep network to have representation power
and **dropout** to reduce overfitting.

Inspired by **VGGNet**
adapted to 1D spectrum.

No feature extraction!

Training:
TENSORFLOW+KERAS on GTX980 GPU



Active Learning

Human (Oraculum) involved

Random/Uncertainty Sampling:

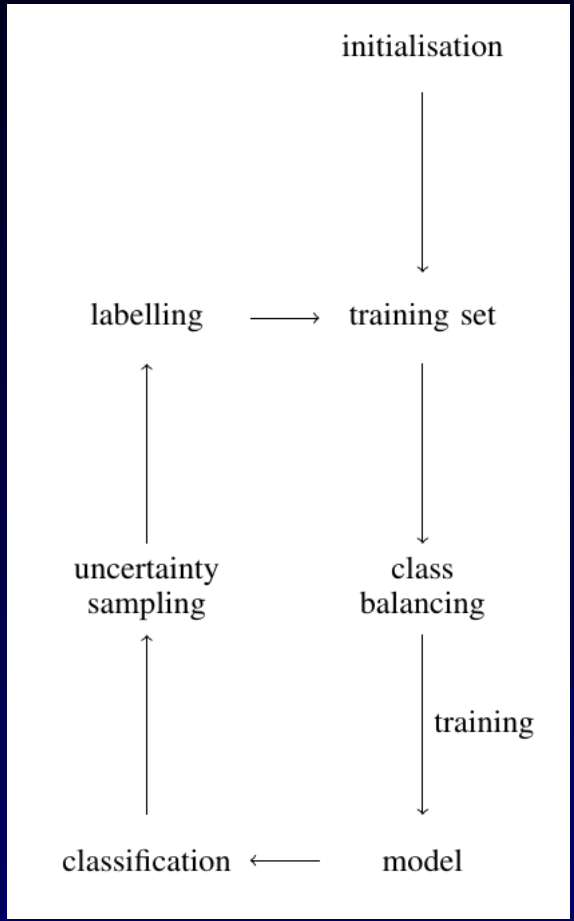
From predicted TARGET class (single or double peak) selected 100

randomly/with highest entropy

Visual check : re-classification (confirm, change, put in uninteresting)

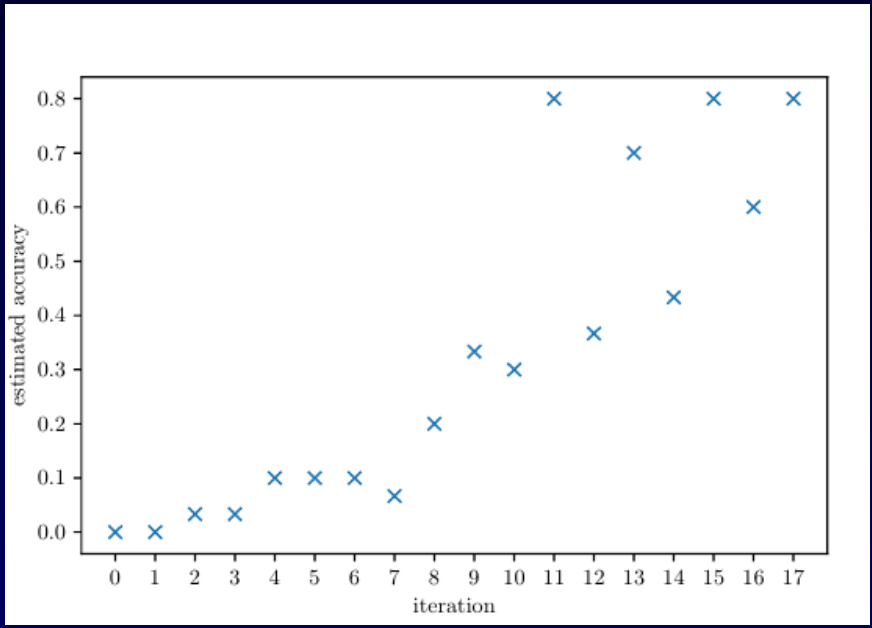
These data added to training set

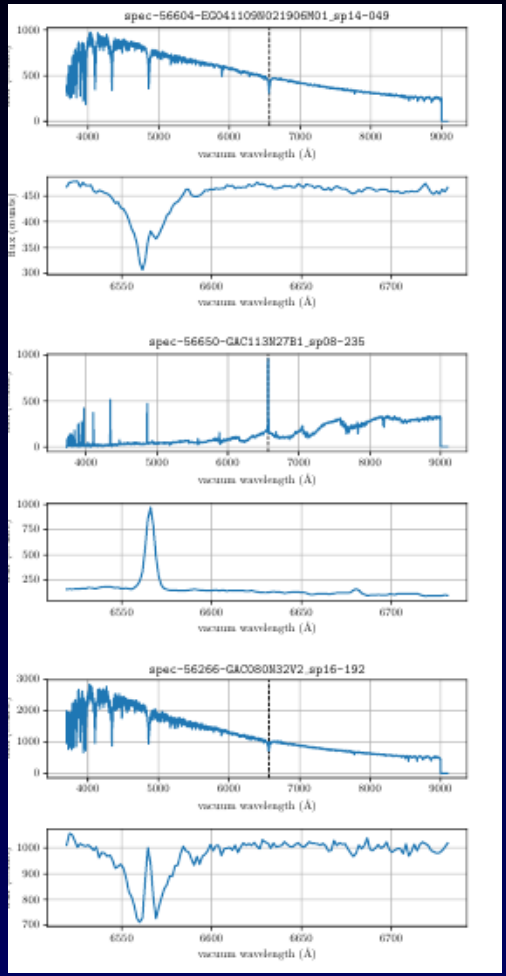
Repeat until few misclassifications (16 times)



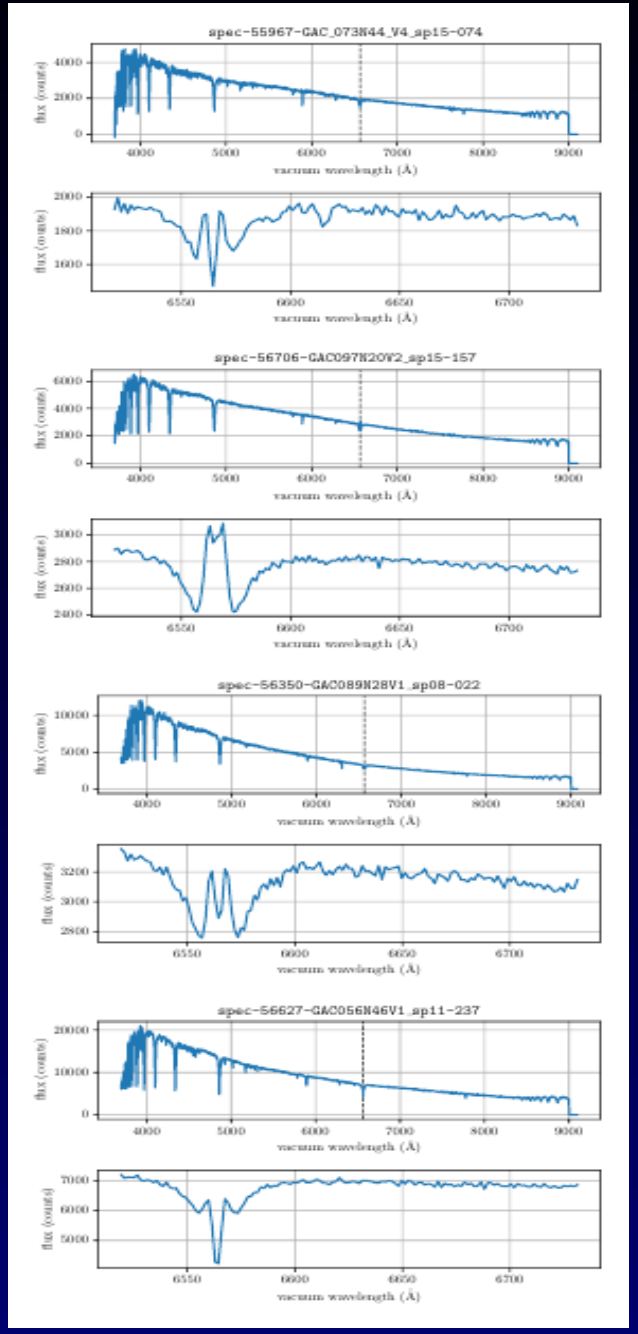
Active Learning (< 6 % error)

Predicted class	Actual class		
	single-peak	double-peak	uninteresting
single-peak	97.6% (3 641)	1.4% (53)	1.0% (37)
double-peak	2.8% (18)	94.0% (609)	3.2% (21)





Single peak



Double peak

Results

Number of SPECTRA (multiple)
4379 candidates (from 4 mil)
58 bad – but still interesting – e.g. LAMOST HVS-1

3731 single peak
648 double peak (but complex shapes)

New objects – 664 in SIMBAD (Be, CV, Seyfert Gal...)

1013 (948 objects) NEW , NOT known

Visual check , XMATCH in VO, DSS2 , SDSS in Aladin

Most are correct - ? physical origin (YSO, CV, M, Novae)
Unreliable classification from LAMOST 1D pipeline (F, A/B)

A lot of new Be stars
Many SUPRISES !!!

Comparison with Other Method

RAA 2016 Vol. **16** No. **9**, 138 (12pp) doi: 10.1088/1674–4527/16/9/138
http://www.raa-journal.org *http://iopscience.iop.org/raa*

Research in Astronomy and Astrophysics

A catalog of early-type emission-line stars and Hα line profiles from LAMOST DR2

Wen Hou[1,2], A-Li Luo[1,2], Jing-Yao Hu[1], Hai-Feng Yang[1,2,3], Chang-De Du[1,2], Chao Liu[1], Chien-De Lee[4], Chien-Cheng Lin[5], Yue-Fei Wang[6], Yong Zhang[6], Zi-Huang Cao[1] and Yong-Hui Hou[6]

[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; *lal@bao.ac.cn, whou@bao.ac.cn*
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
[4] Institute of Astronomy, National Central University, Jhongli
[5] Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
[6] Nanjing Institute of Astronomical Optics & Technology, National Astronomical Observatories, Chinese Academy of Sciences, Nanjing 210042, China

$$\sum_{i=-5}^{5} f_{\rm obs}[n_0 + i] \,/\, 11 > f_{\rm conti}[n_0]\,,$$

$$\sum_{i=-1}^{1} f[n_0+i]/3 > \sum_{i=-2}^{2} f[n_0+i]/5$$

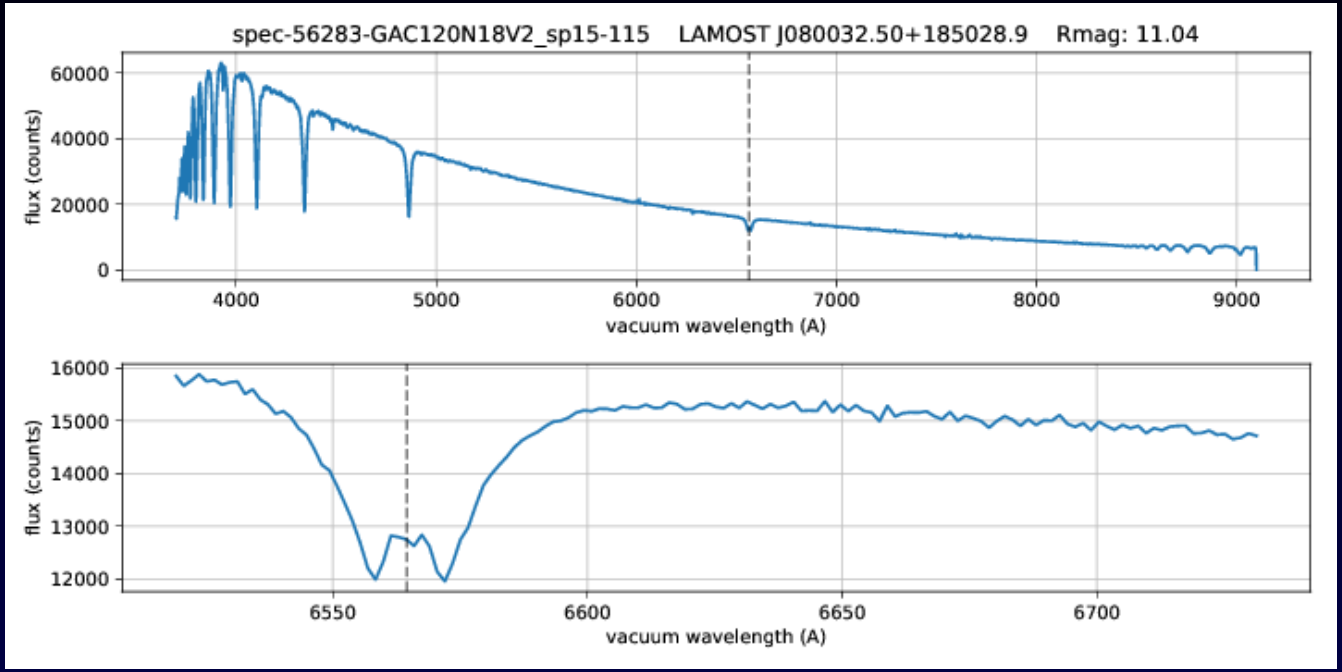
nd

$$\max\left(f_{\rm obs}[n_0-1:n_0+1]\right) \geqslant \max\left(f_{\rm obs}[n_0-2:n_0+2]\right)$$

Integral pixel statistics on different intervals around Hα

Hou (2016) DR2 - catalog of 11205 emission stars
we have **2644** of them - but 11k not well justified !
(VO xmatch, SPLAT visualization ….)

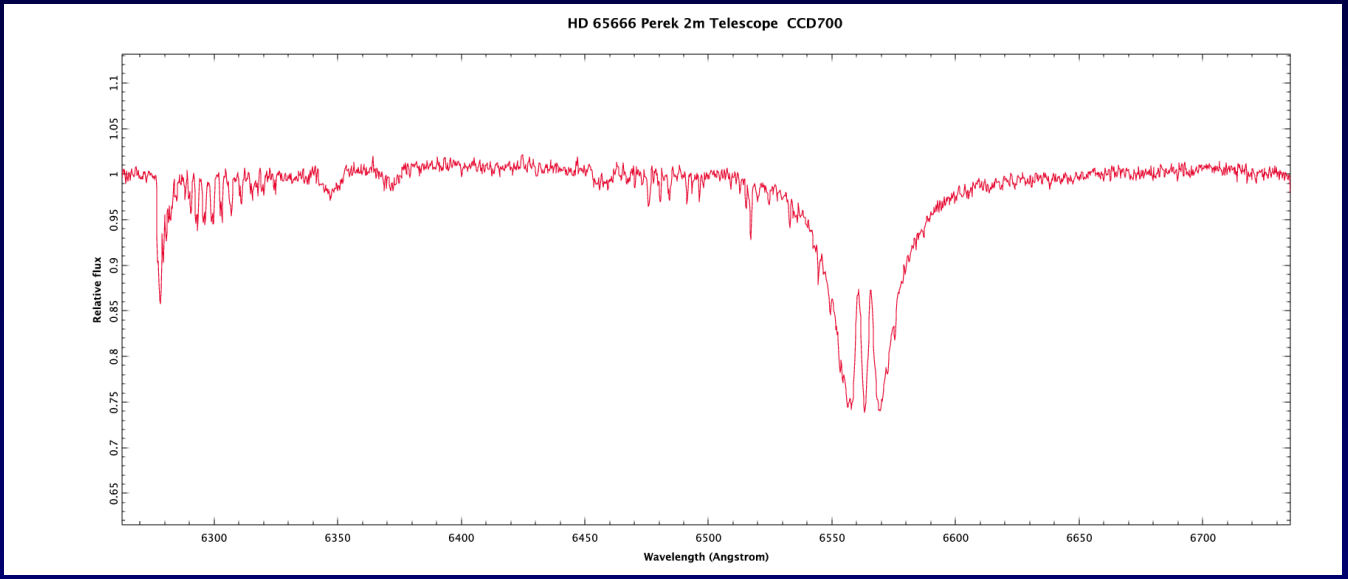
Results – Unknown Bright Be star



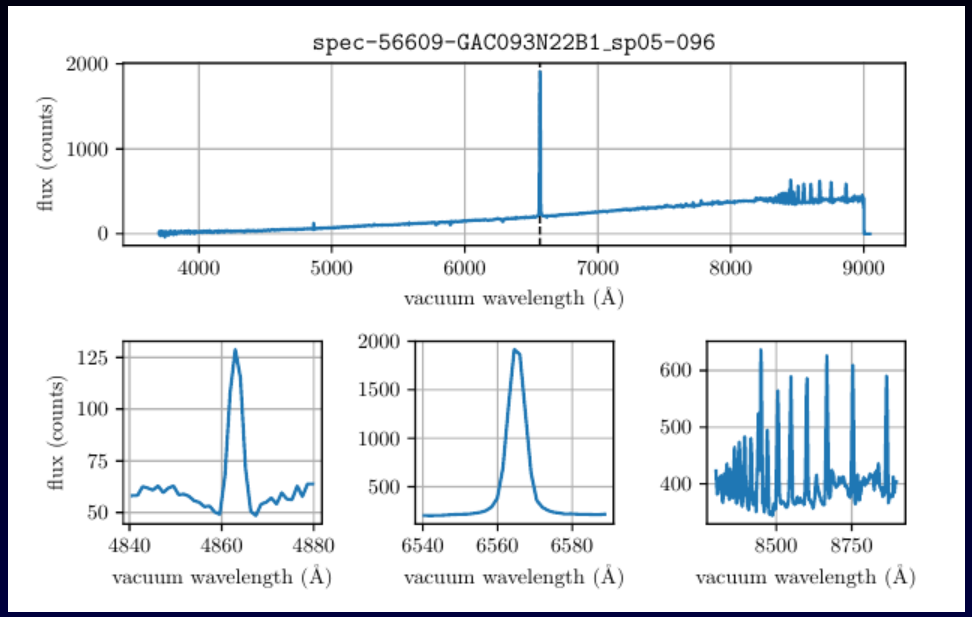
LAMOST

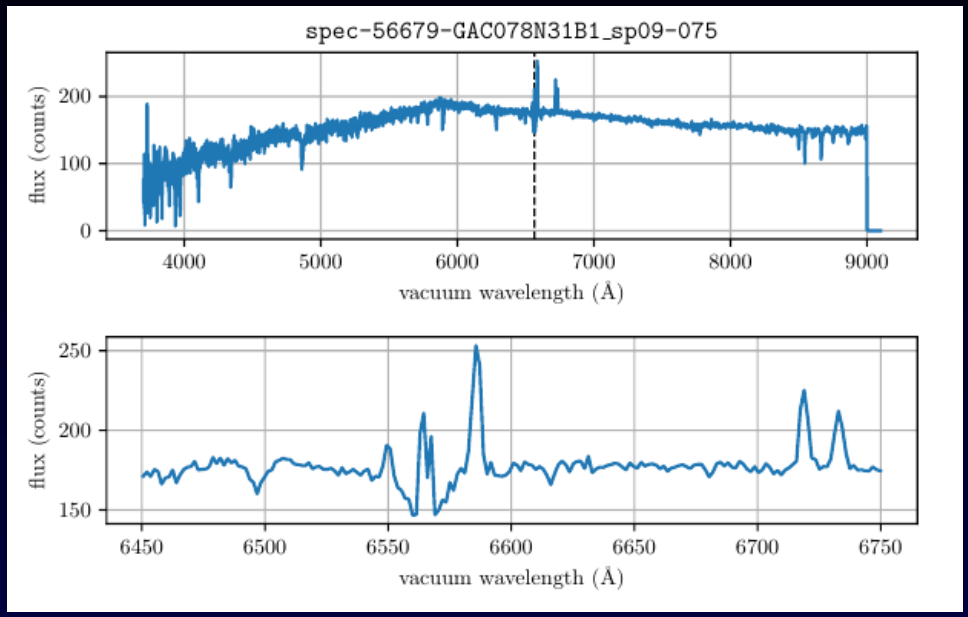
6.5 arcsec distant star 7 mag

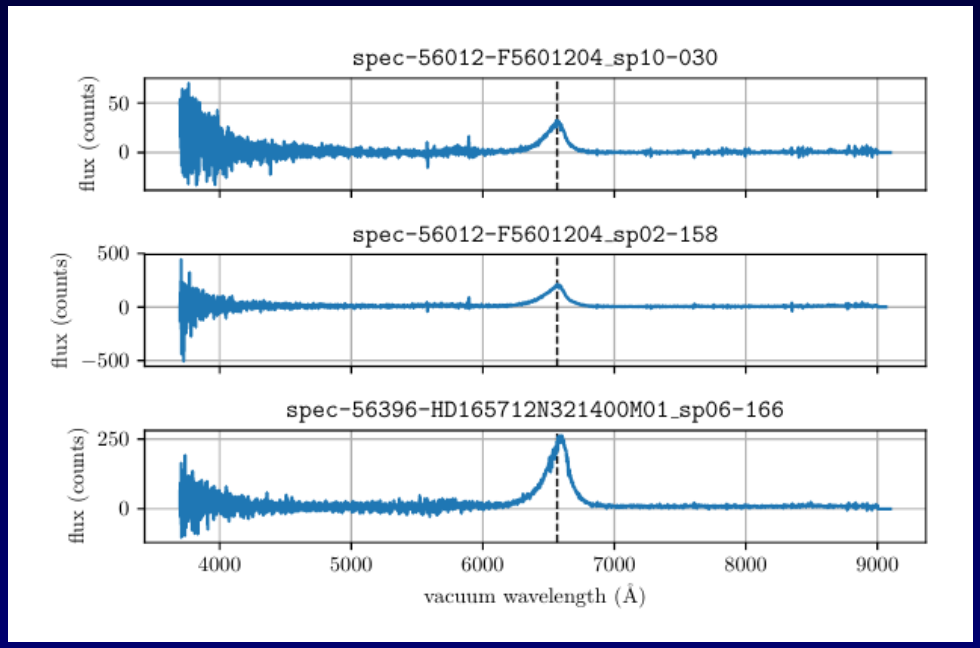
CCD700 OND



Results – Interesting







Supernovae ???

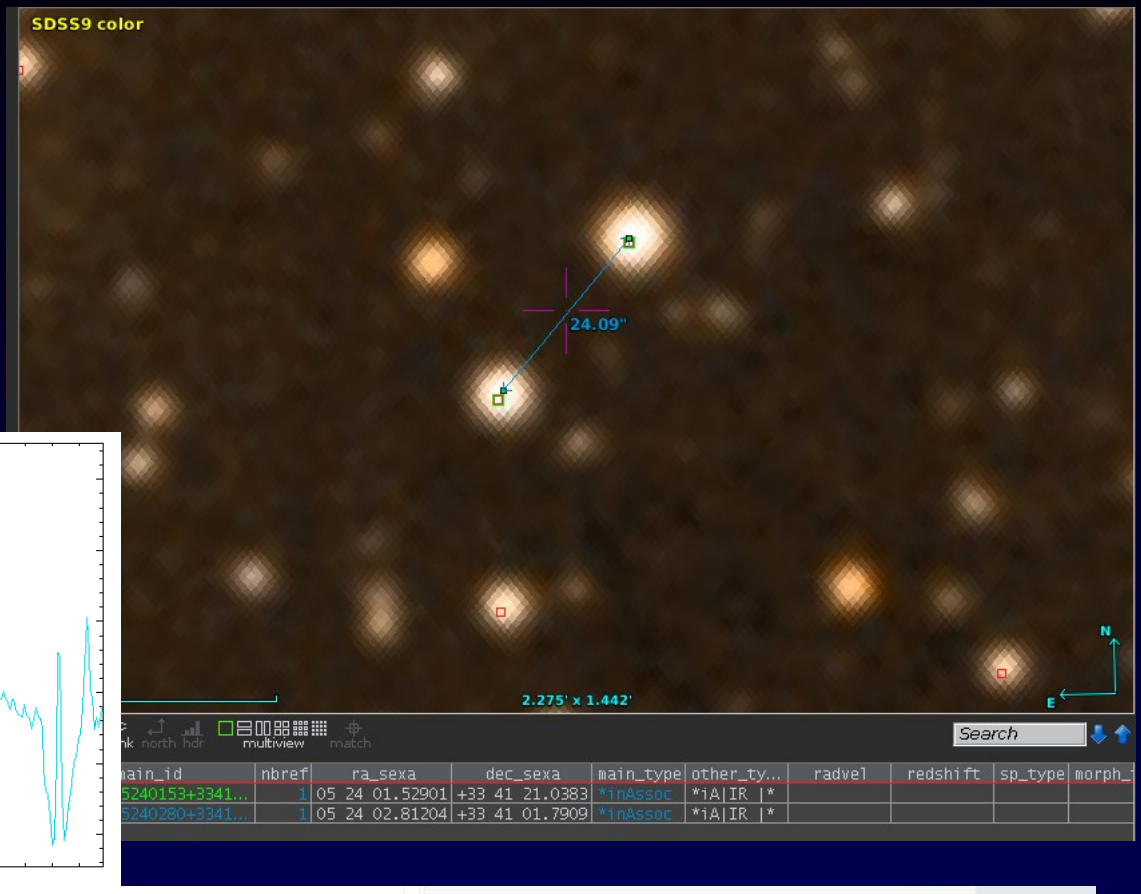
In DSS2 /SDSS there are galaxies in 10arcsec around ……

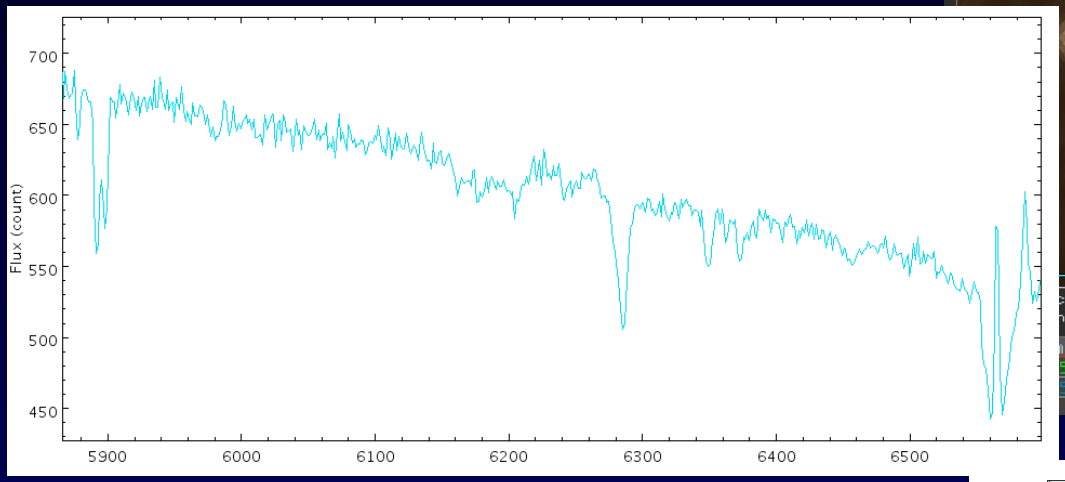
Confusion in Unique Identification

2 stars that are 24 arcsec apart

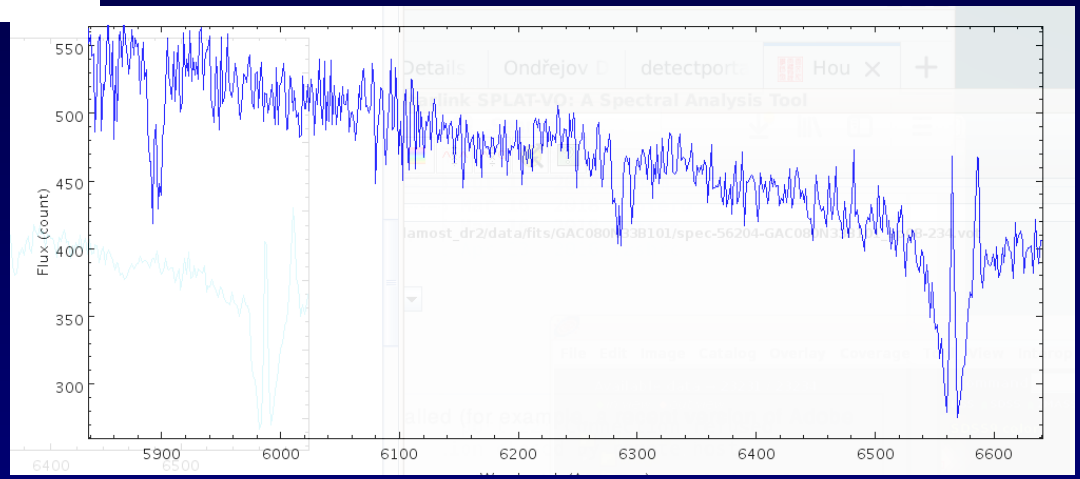
LAMOST J052402.81+334101.7
LAMOST J052401.53+334120.9

2MASS J05240280+3341017
2MASS J05240153+3341210





Is it the fiber light leakage ????
OR Nebula lines ??



LAMOST J034912.80+240820.0
Is Pleione 22 arcsec apart

Confusion in Unique Identification

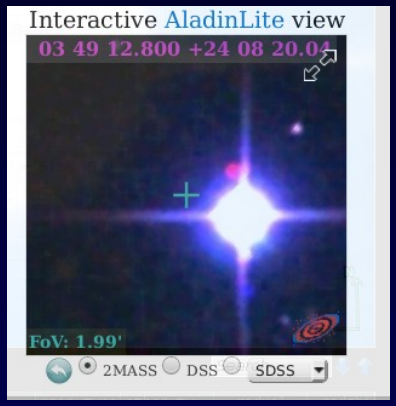
LAMOST J034912.80+240820.0

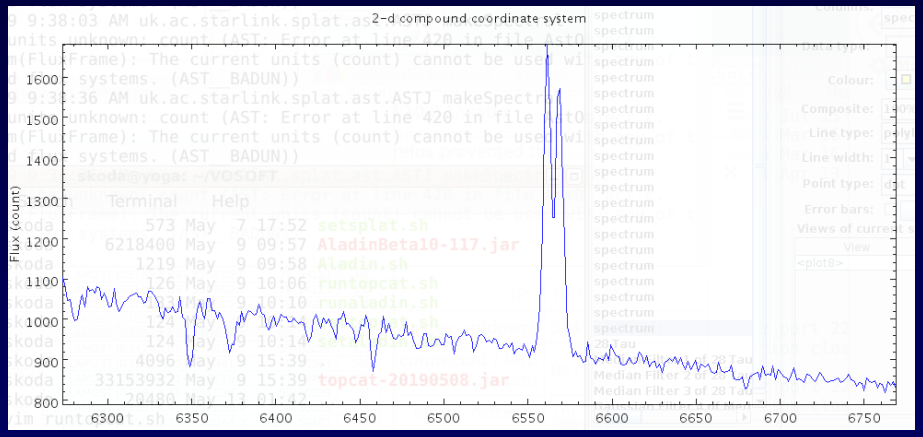
is Pleione (5mag) 22 arcsec apart

Basic data :

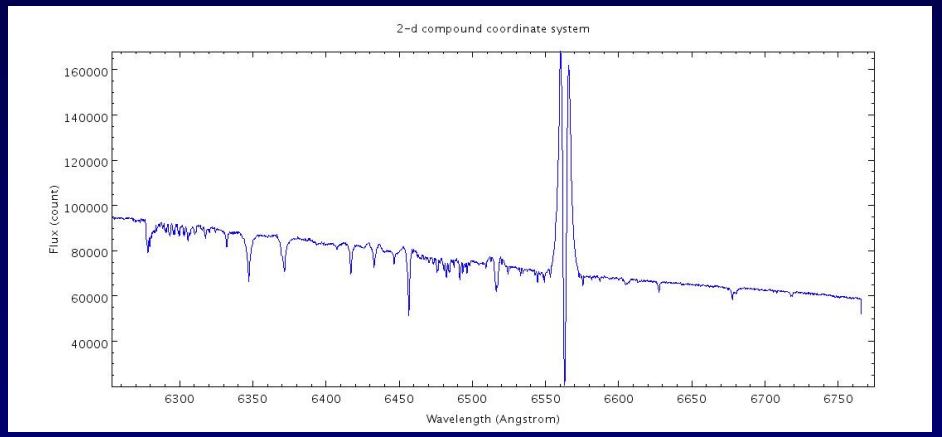
LAMOST J034912.80+240820.0 -- Peculiar Star

Other object types:	**Pe*** (Ref)
ICRS coord. *(ep=J2000) :*	**03 49 12.800 +24 08 20.04 (Optical) [ ] D 2015MNRAS.449.1401H**
FK4 coord. *(ep=B1950 eq=1950) :*	03 46 14.053 +23 59 13.23 [ ]
Gal coord. *(ep=J2000) :*	166.959918 -23.163713 [ ]
Spectral type:	**A1mF1 D 2015MNRAS.449.1401H**





LAMOST MJD 56295



Ondrejov at MJD 56153

Joining ML and VO

- Machine learning – supervised (labels)
- Classical approach – preprocessing , table (unique ID)
  - Interpretation ID to original data

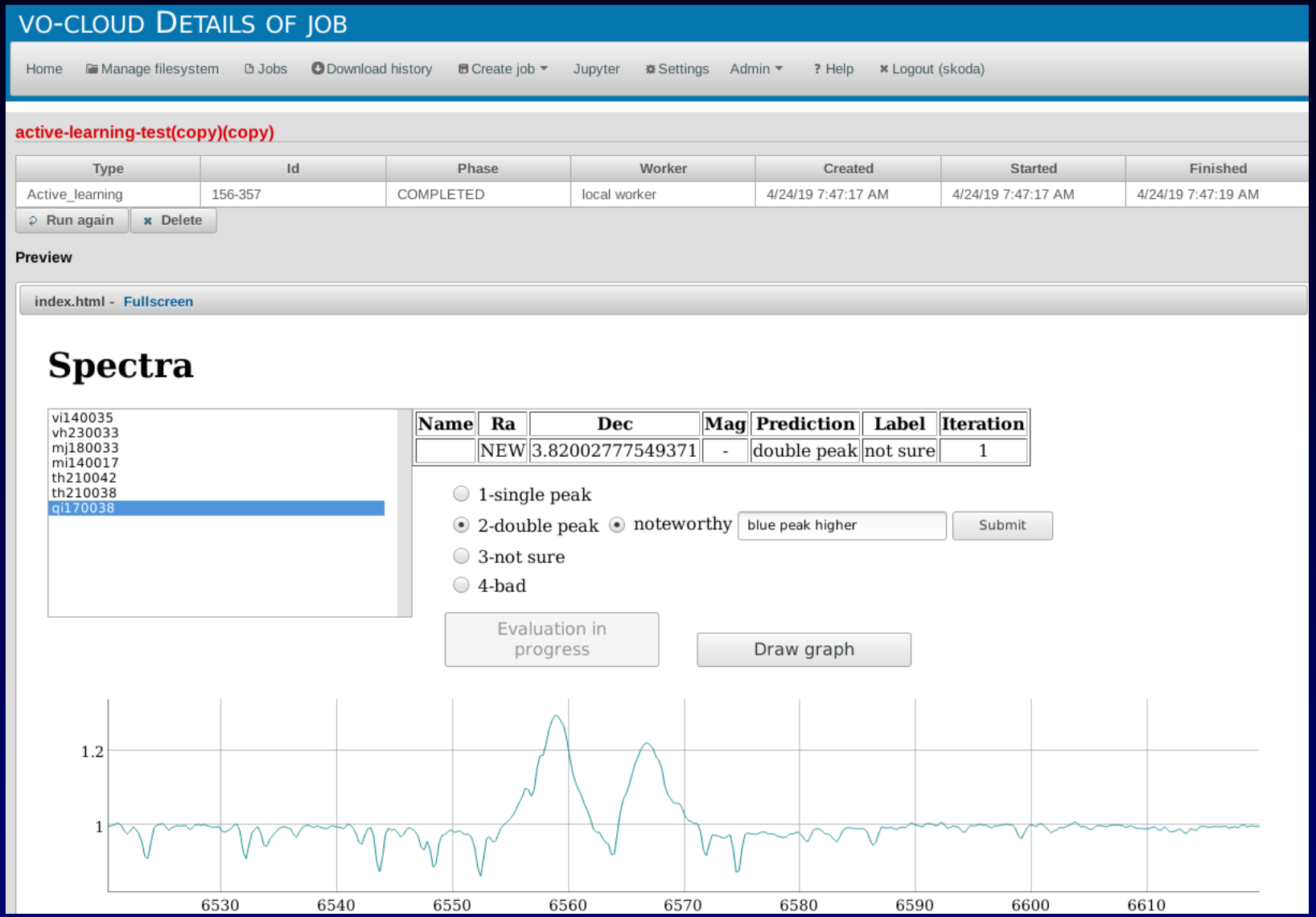
- Requires enough labels - in science difficult
- Deep learning – requires A LOT of data/labels

- Active learning - needs to decide ON THE FLY
- RECLASSIFICATION – interactive
- Oraculum – needs COMPLEX information to decide
- METADATA, comparable DATA

VO Interoperability – crucial part of process

Joining ML and VO

- ML is Big Data = Done in cloud – GPU cluster, Python
- Interactive visualization + classification GUI



Conclusions

- Active learning overcomes the lack of labeled data
- ML on big spectra archives may identify new interesting objects yet unknown/unexpected
- Domain adaptation gives LABELS on unlabeled set
- Crucial is interactive visualization of candidates …..
- Deep learning shows its strength

- Active Learning - ORACULUM requires VO to have
  METADATA to decide correctly
  OTHER DATA (global interoperability !)
- ML needs to visualize data as part of its process now!

THANK YOU