Identification of Interesting Objects in Large Spectral Surveys Using Highly Parallelized Machine Learning

Petr Škoda

Astronomical Institute, Czech Academy of Sciences Ondřejov

Andrej Palička, Jakub Koza, Ksenia Shakurova

Faculty of Information Technology, Czech Technical University, Prague

Supported by grant LD-15113 of the Czech Ministry of Education, Youth and Sports

Using computer power of MetaCentrum
Supported by CESNET (LM2015042)
and CERIT-Scientific Cloud (LM2015085) projects of the same ministry

Presented at IAU Sympozium 325 on Astroinformatics
21st October 2016, Sorrento, Italy

Be Stars - Introduction

Gamma Cas (Padre Angello Secchi 1866)

(Pontifical Gregorian univ. - Roman College – Gregor XIII) – visual spectrograph

Be* have or have had emission in Balmer lines

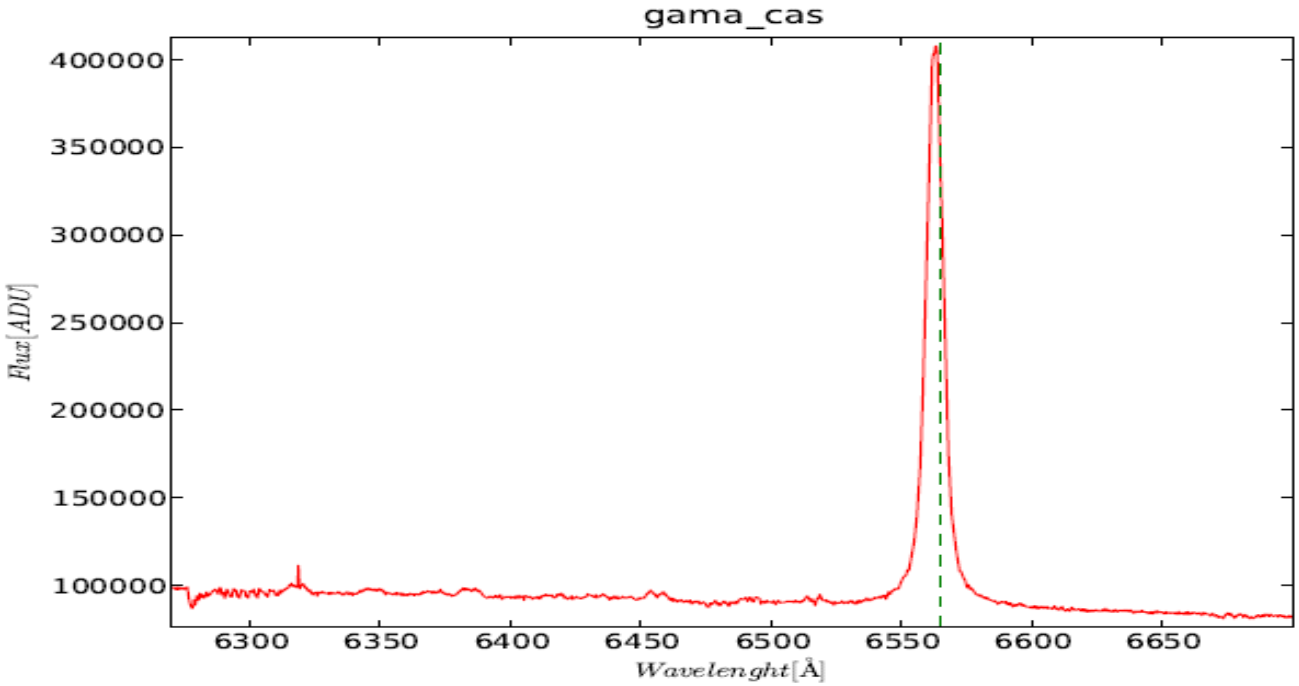
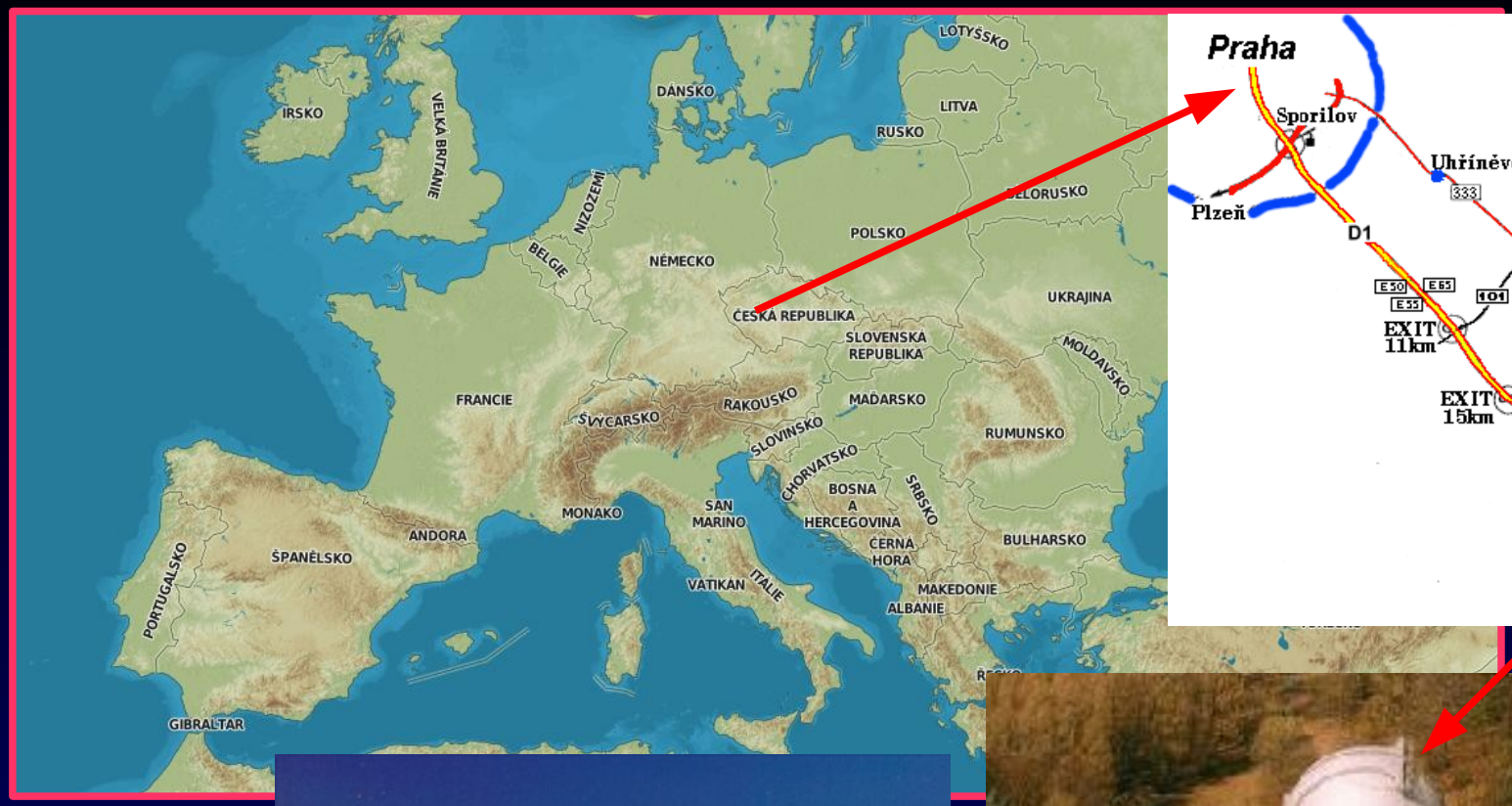
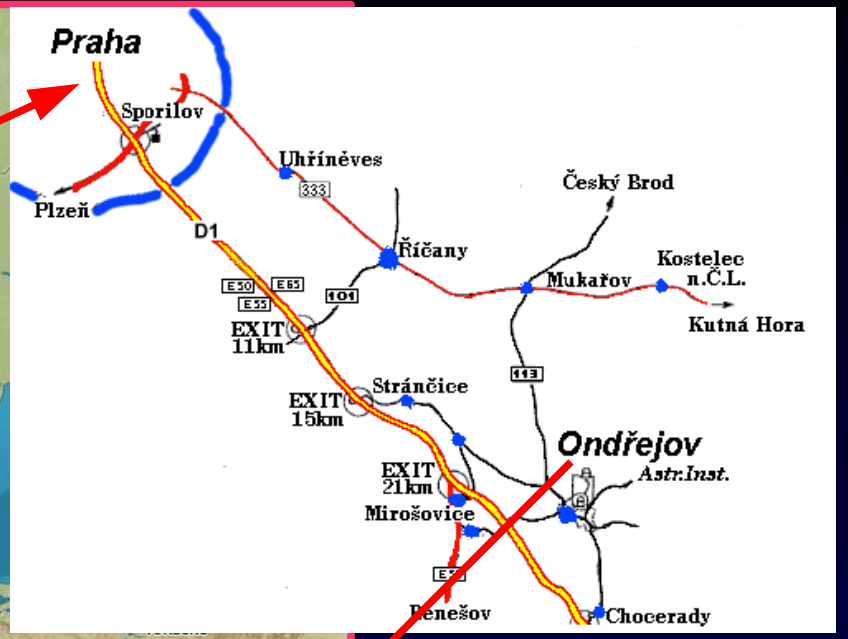


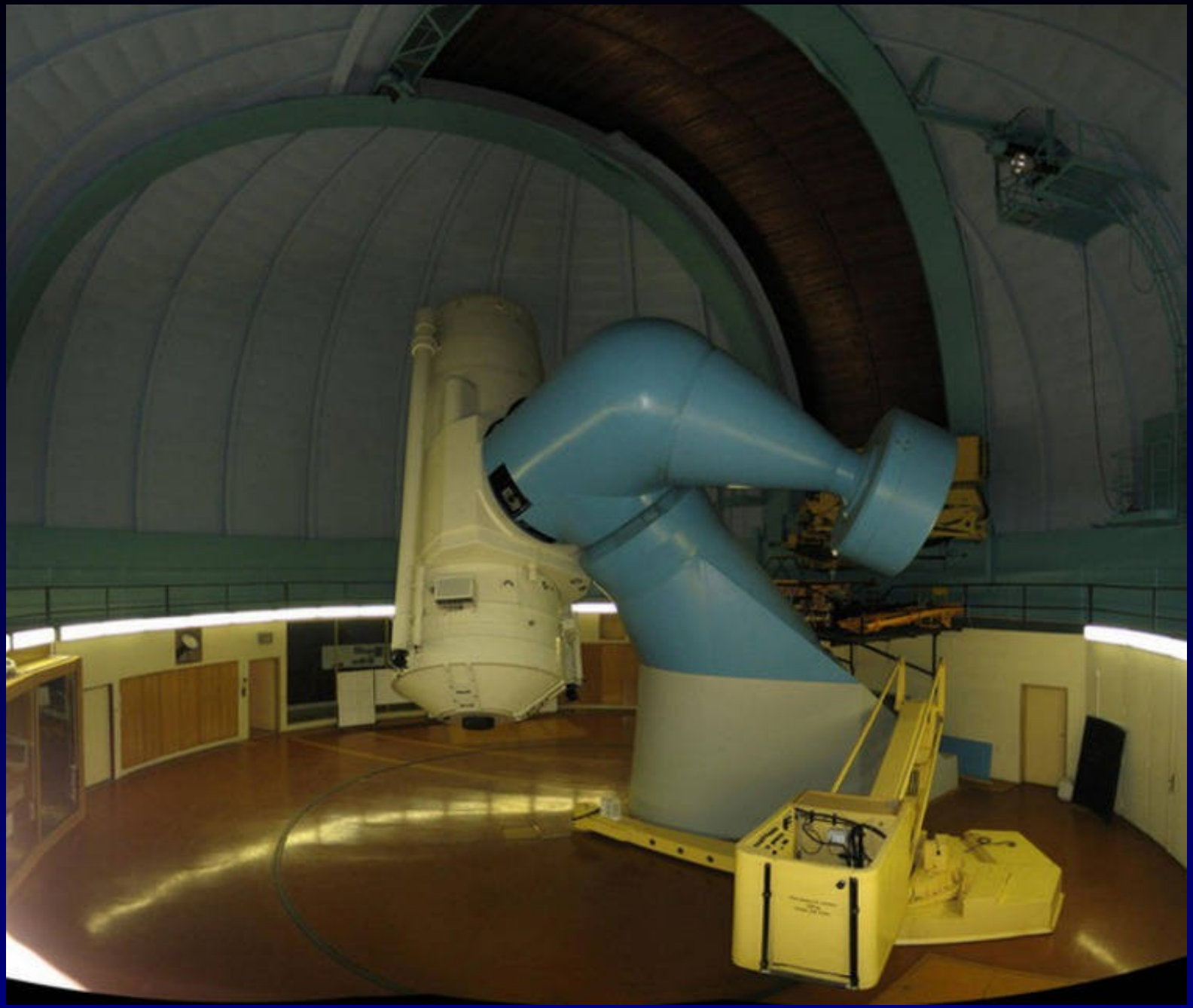
Figure 3.18: Spectrum of Gamma Cas. Be Star

Ondřejov observatory





Ondřejov 2m Perek Telescope (1967)



Motivations for Be Stars

The ~40 years of research in Ondřejov (data)

Be stars are mysterious (after >100 years)

Different time scales, quasi periodicity (not sure)

emission epizodes, can look normal (20% of B = Be!)

Zoo of line profiles (winebottle, abs+em, high em)

Transitions:

shell phase → emission, single → double peak

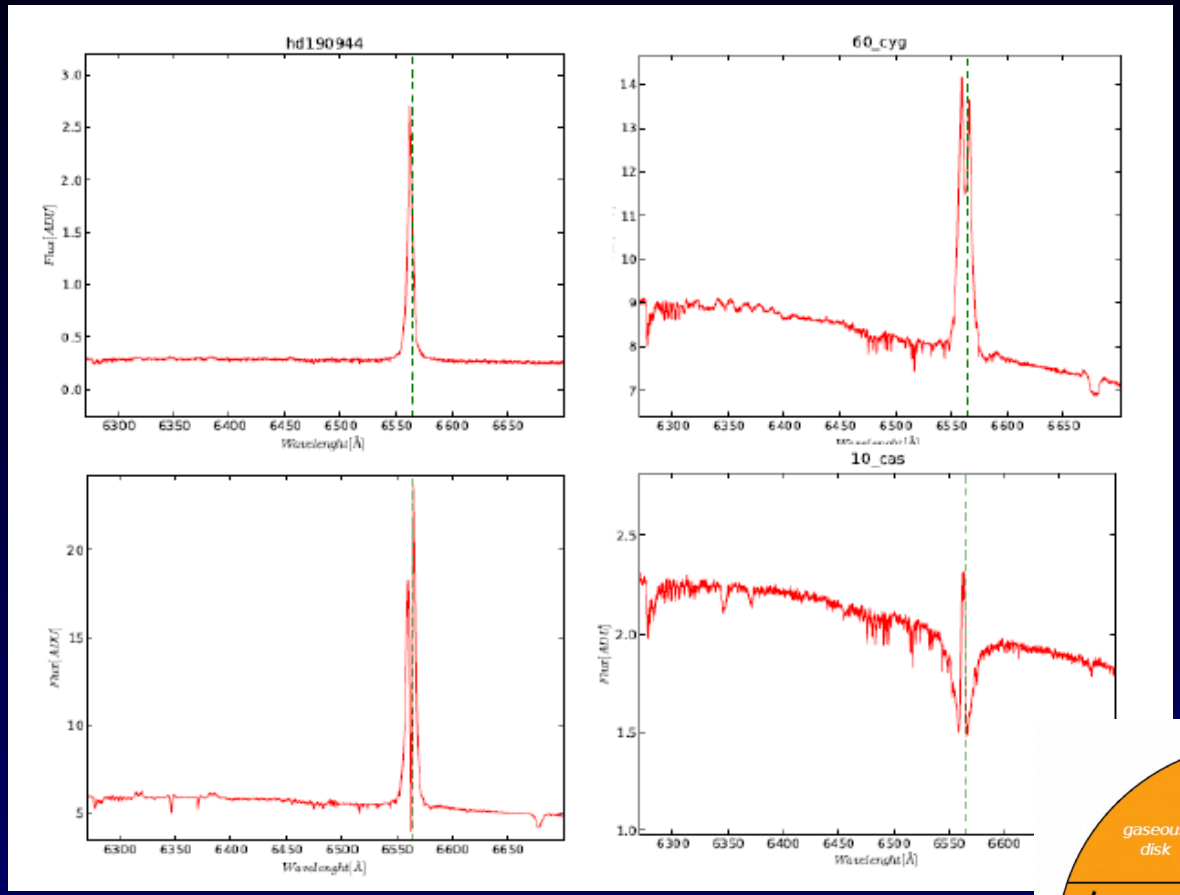
V/R changes (1/3 of double peak), some stable V=R

transition V/R var to stable (1-10s years)

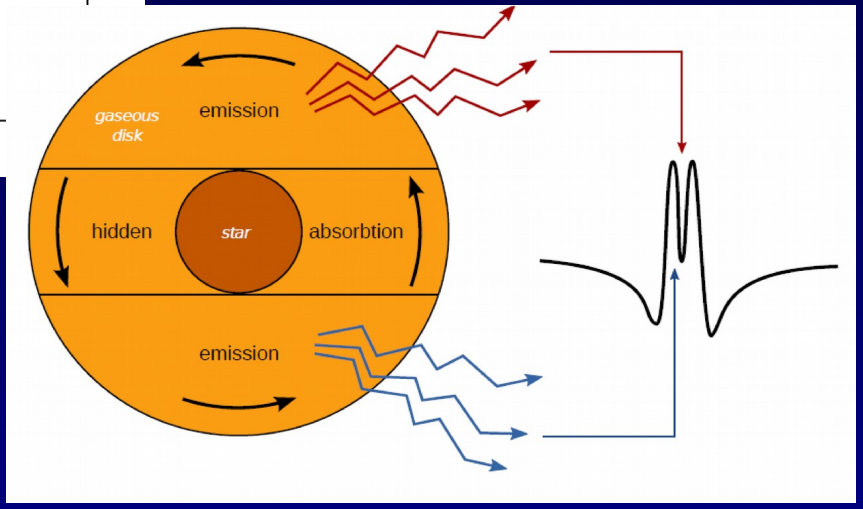
Chalenge for Machine Learning (and theory !)

Machine Learning of Spectra

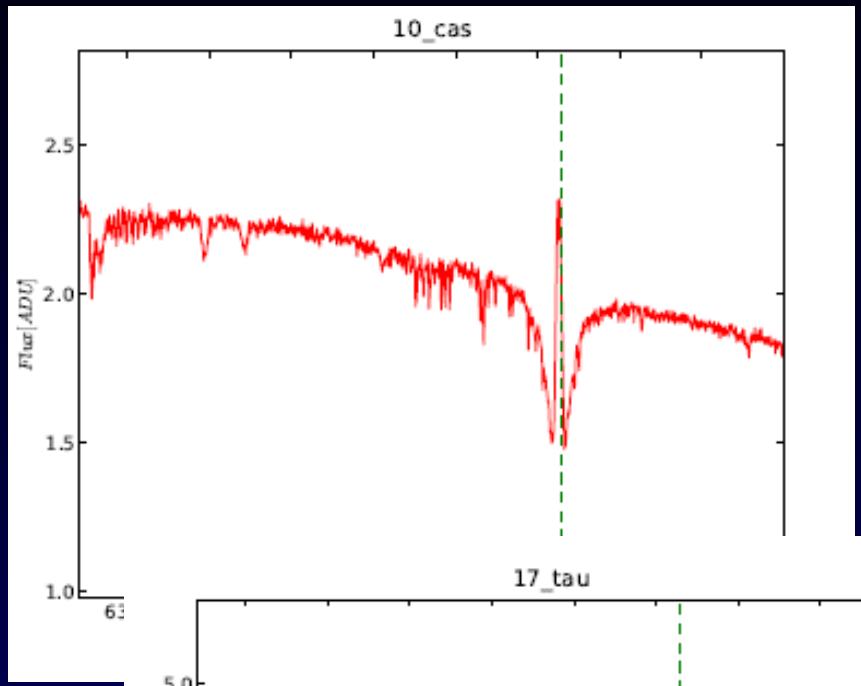
Use case: ML of spectra profile of Halpha line (Be stars)

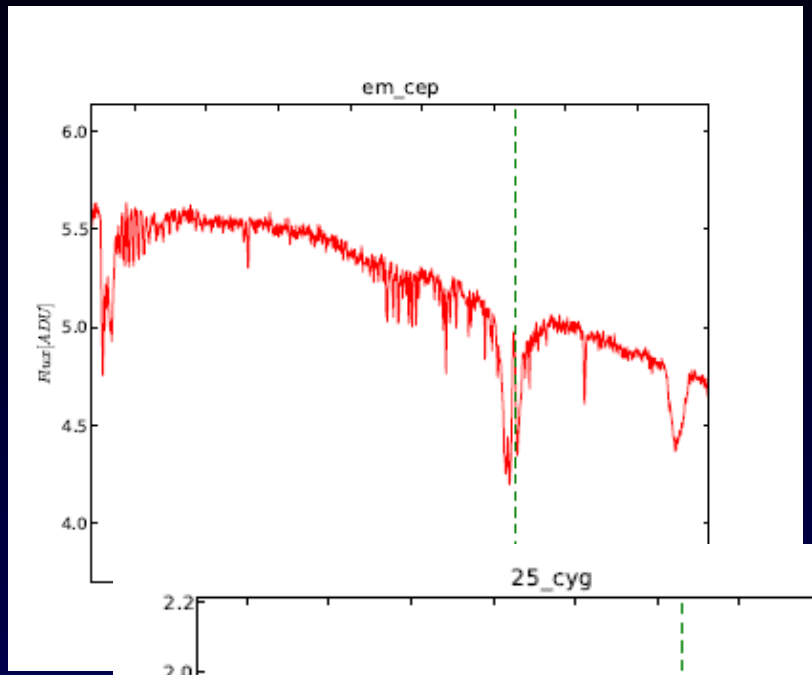


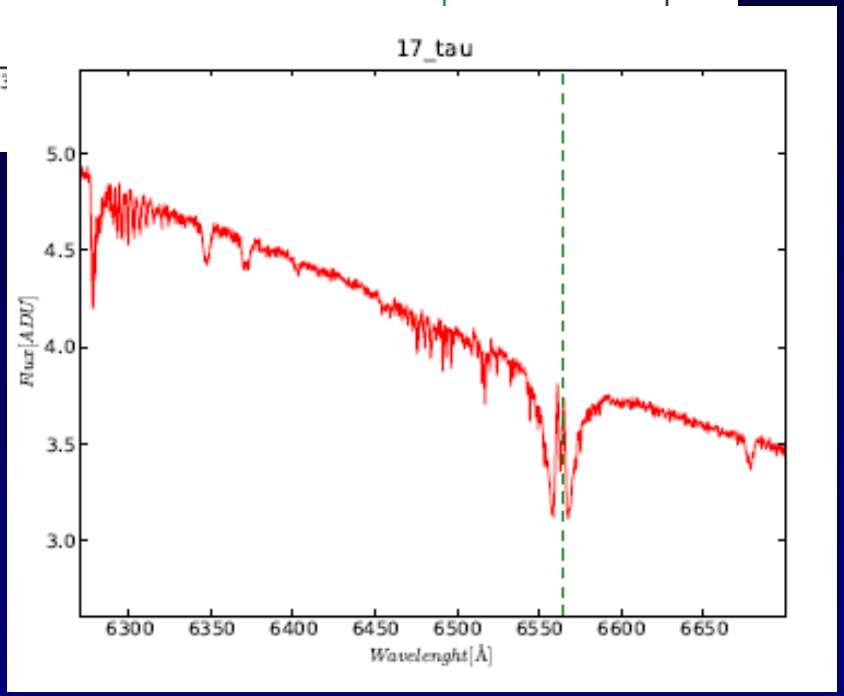
- Be stars
- Disk or envelope
- Rotates, Hot
- Origin ?????

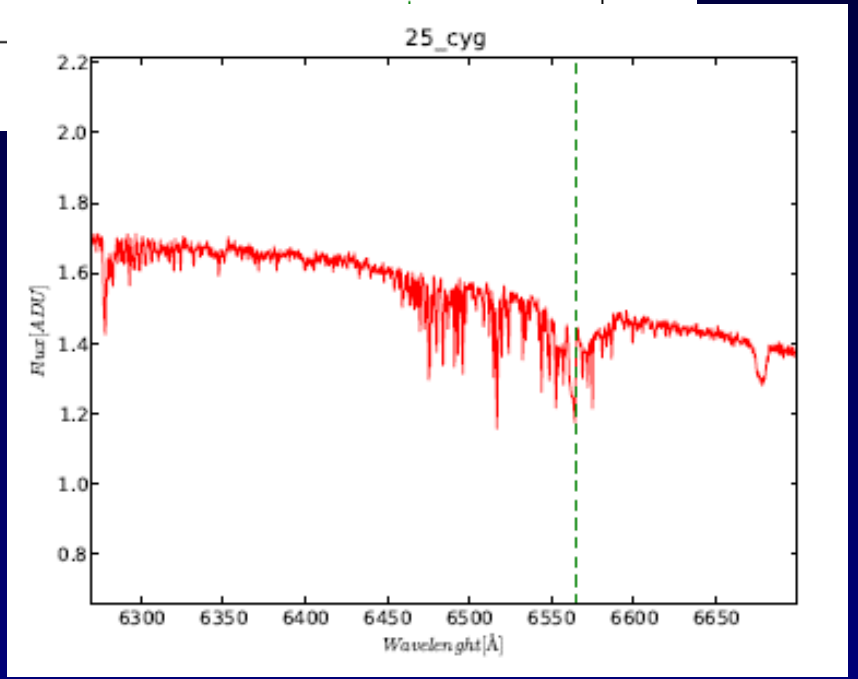


Be Stars : Emission in absorption

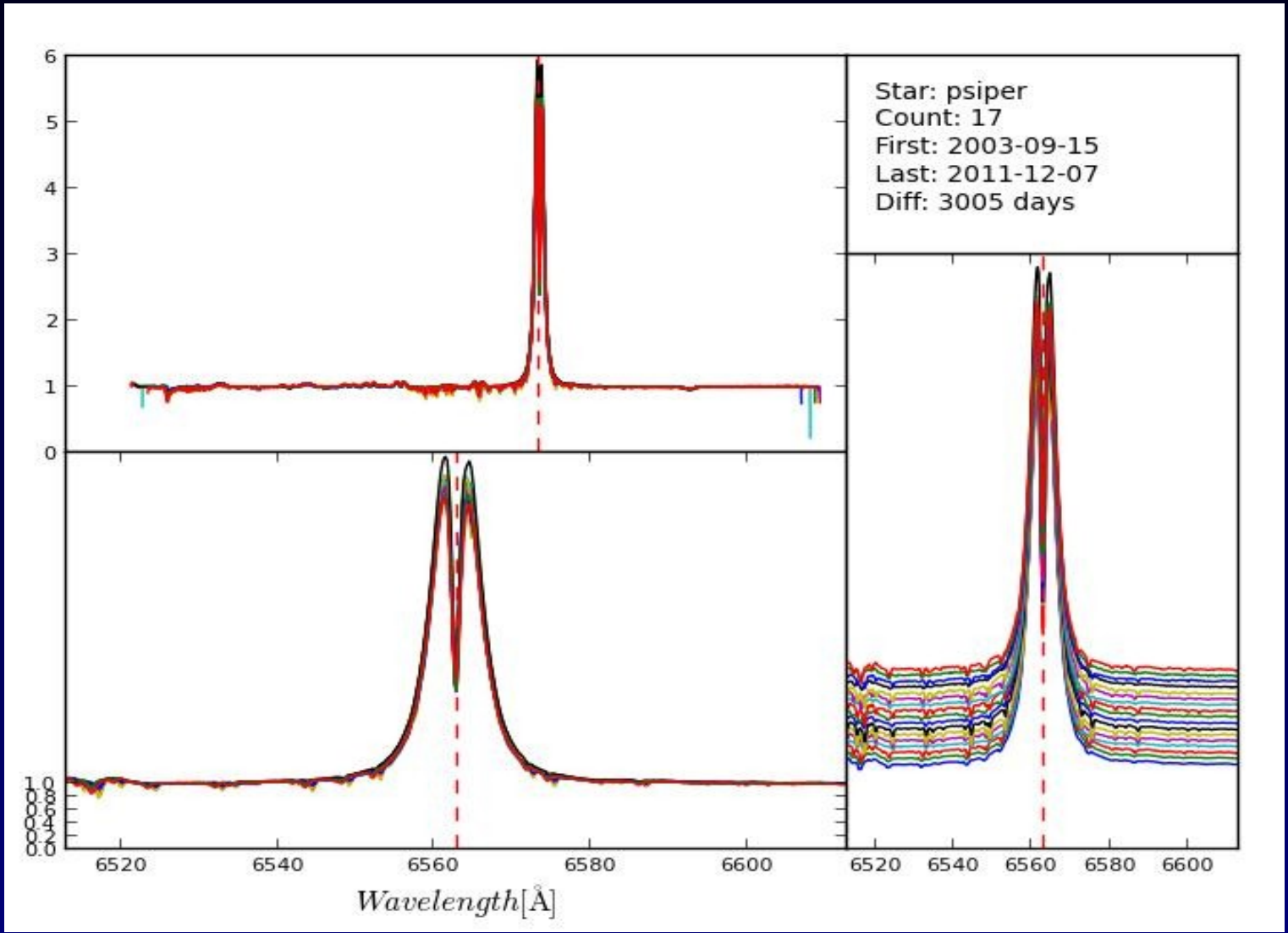








Ond2m Archive - Stable Emission



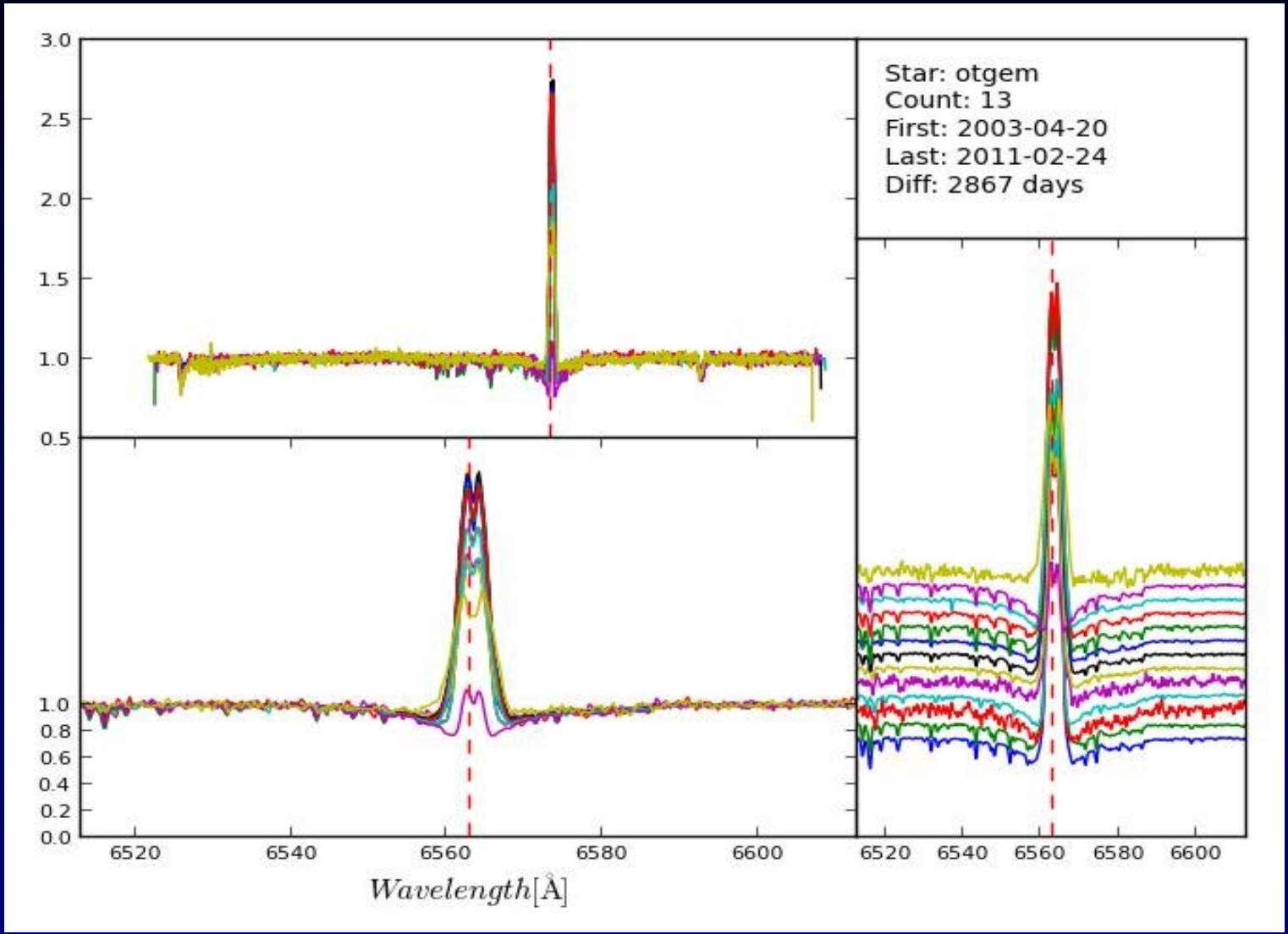
R=13000

12000 spectra of >1000 stars

Stable samples needed for ML

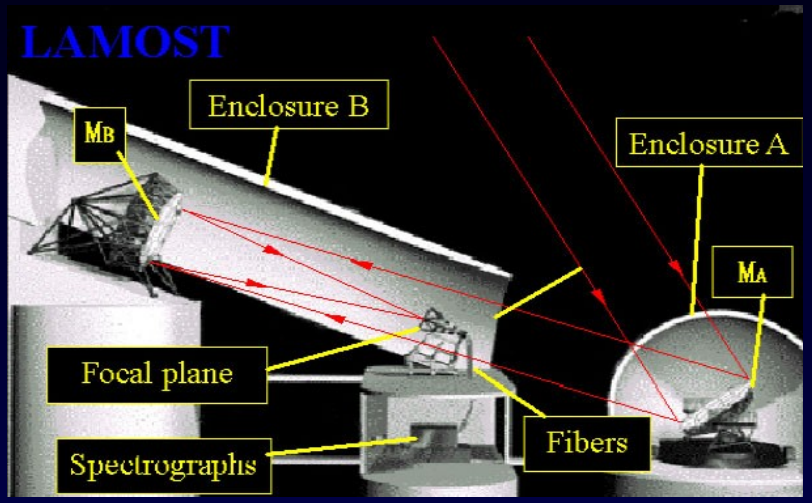
Selected 1600

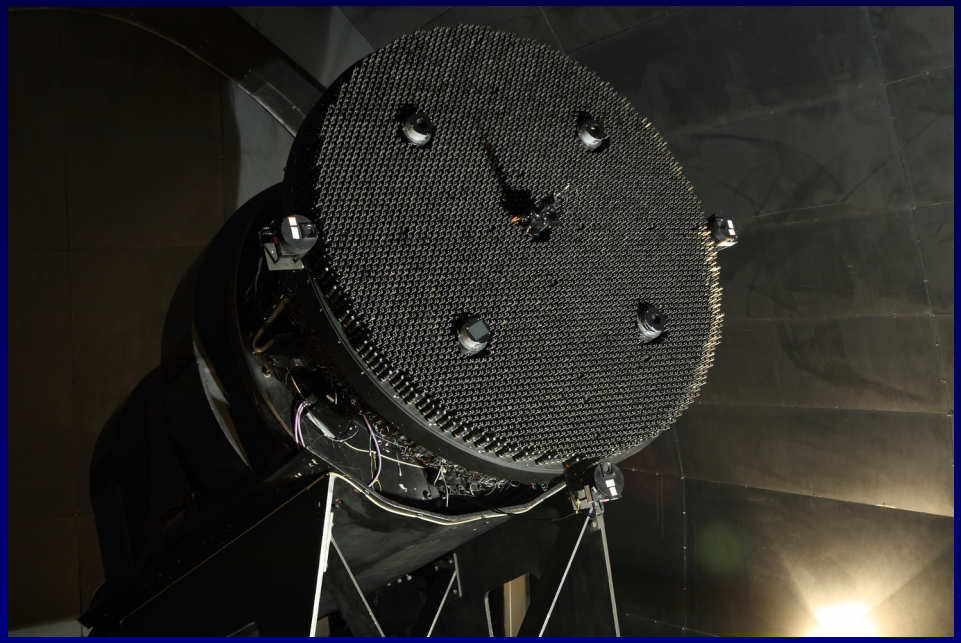
Ond2m Archive – Unstable Emission

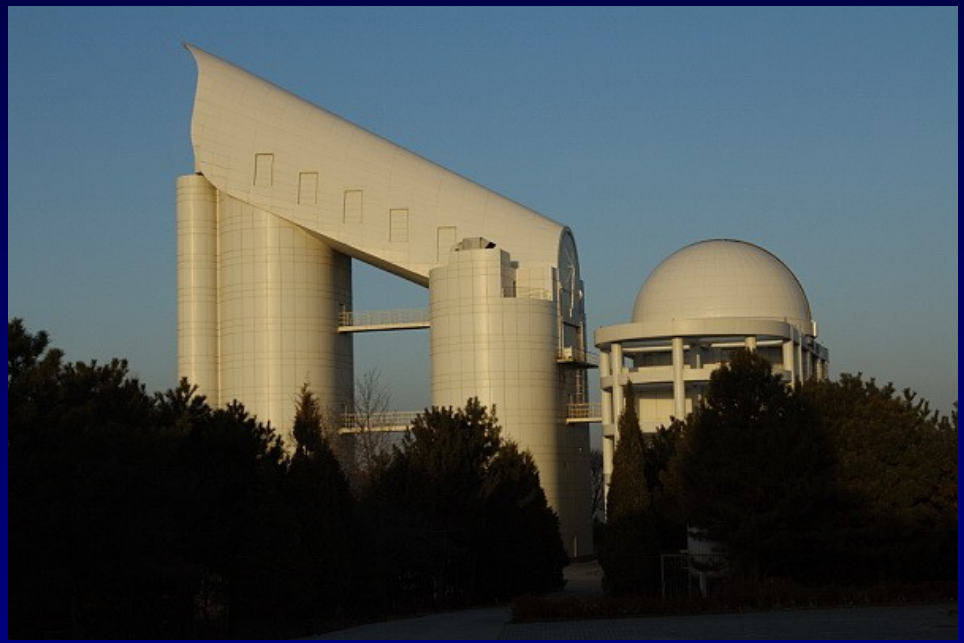


LAMOST (Guoshoujing)

Xinglong- China
4m mirror (30 deg meridian)
4000 fibers/16 spectrogr.
10 mil spectra / 5 yr
Automatic RV-z







LAMOST Spectral Surveys

DR3 (half 2015) **5 755 126** spectra
DR4 (Feb 2016) **+ 741 522**

3700-9000A R~500-5000
Limiting mag 19-20 for single exp.

Each Fiber – 2 motors
double arm 33mm circle

Fibre collects light from
3.3 arcsec circle on sky

LAMOST Spectral Surveys

DR1 (end 2013)

2 204 860 spectra including 717660 PDR

1 085 404 stars with estimated physical parameters

LEGAS extragalactic

LEGUE - galactic

3700-9000A **R~1800**

Machine Learning of Spectra

PRE-PROCESSING of LAMOST

Normalization to continuum (another FITS extension in DR1)

Cutout to Ond 2m archive range (6250-6700A)

Rebinning (same wavelegth points) + Renormalization [-1,+1]

(Reduction of dimensionality (wavelets, PCA, LLE...))

Produces feature vectors in CSV (same length, dimensions)

Machine Learning of Spectra

PRE-PROCESSING of OND 2m archive

Normalization to continuum (automatic algorithm)

Cutout to common smallest range (6250-6700A)

Convolution to LAMOST resolution (12000->1000 SRP)

Rebinning (same wavelegth points) + Renormalization [-1,+1]

(Reduction of dimensionality (wavelets, PCA, LLE...))

Produces feature vectors in CSV (same length, dimensions)

AND LABELS (1,0) – has interesting emission

Domain Adaptation

Convolve(OND2m) → LAMOST Simulated

Domain Adaptation - not Supervised Training

Learn from labelled D1- the model learnt applied on D2

In our case WE KNOW THE MODEL - spectrograph SRP

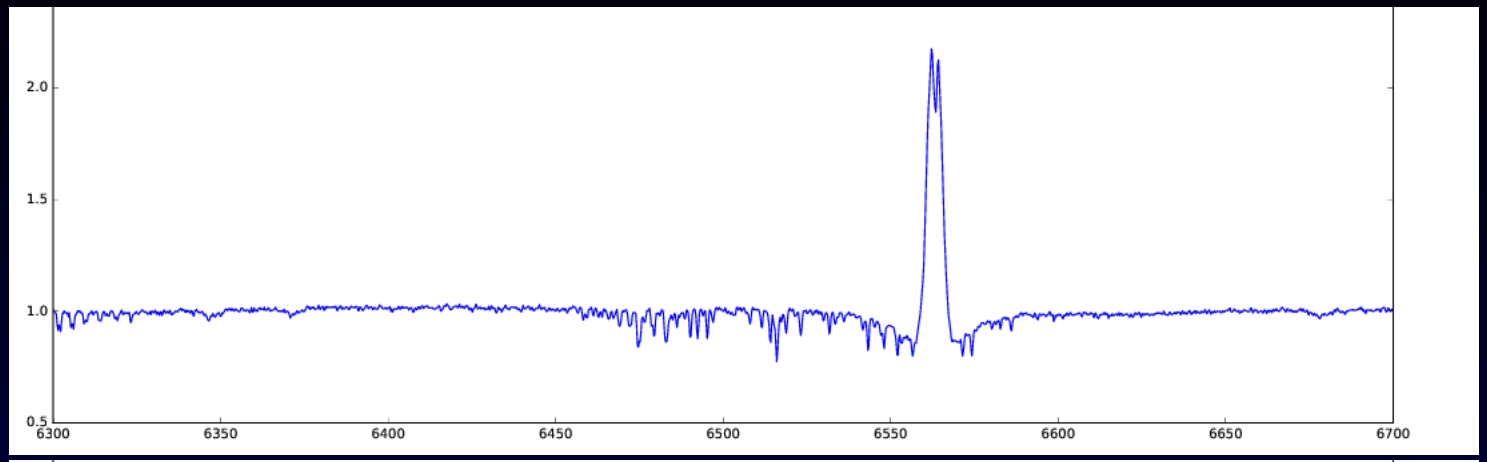
Degradation of resolution

(simply Gaussian profile convolution ~ 6 pix)

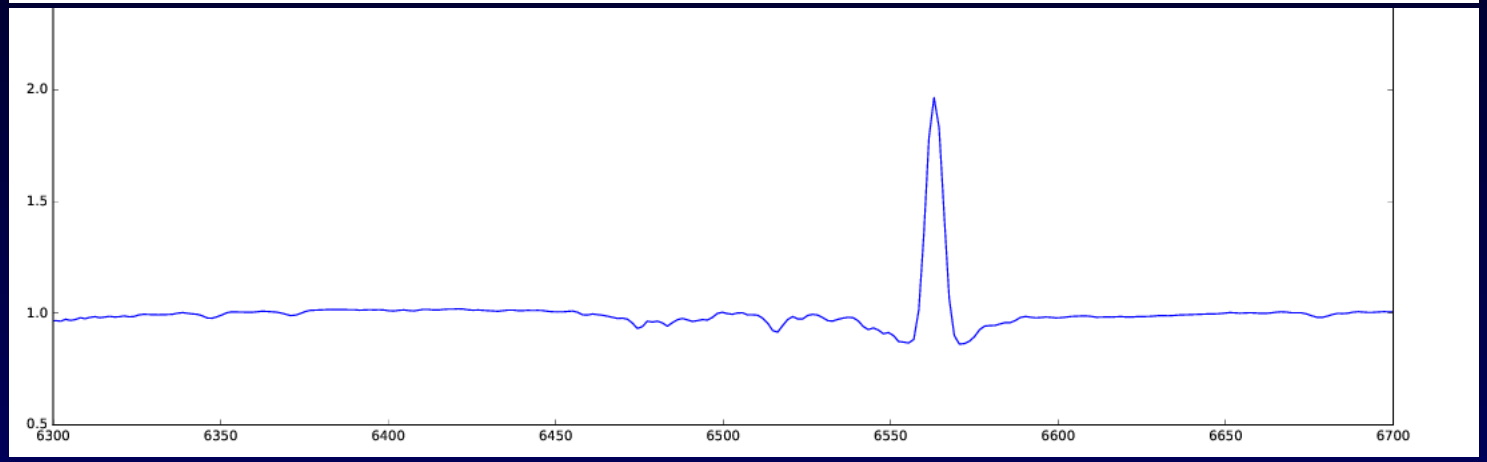
Test: Cross-matching Ond2m with LAMOST (using VO)

Found several stars observed in both archives

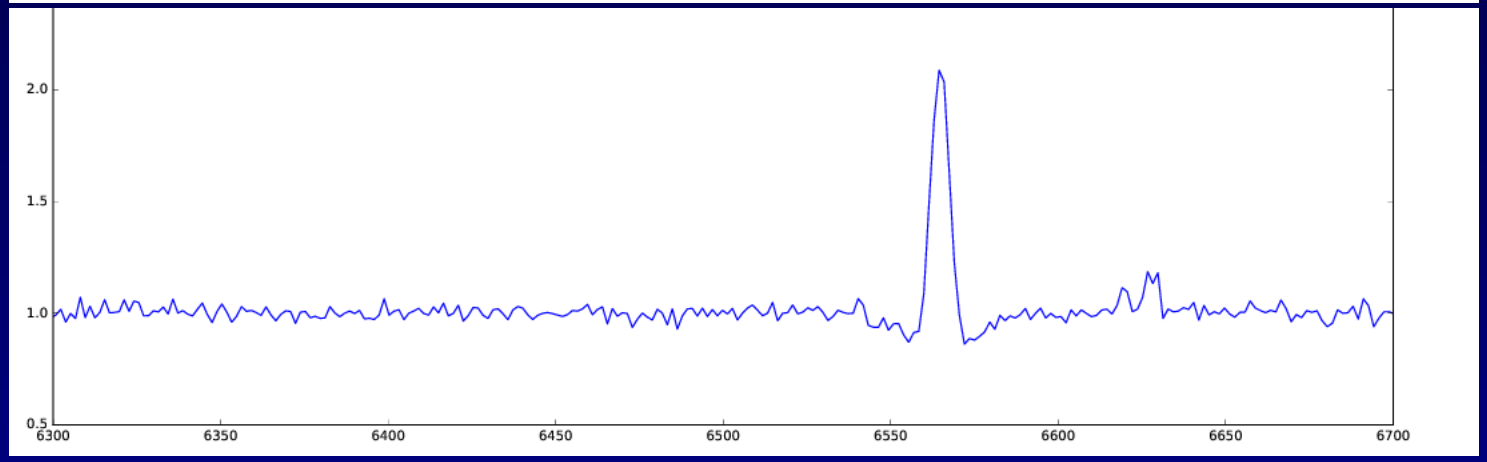
Resolution Degradation



OND R=13000

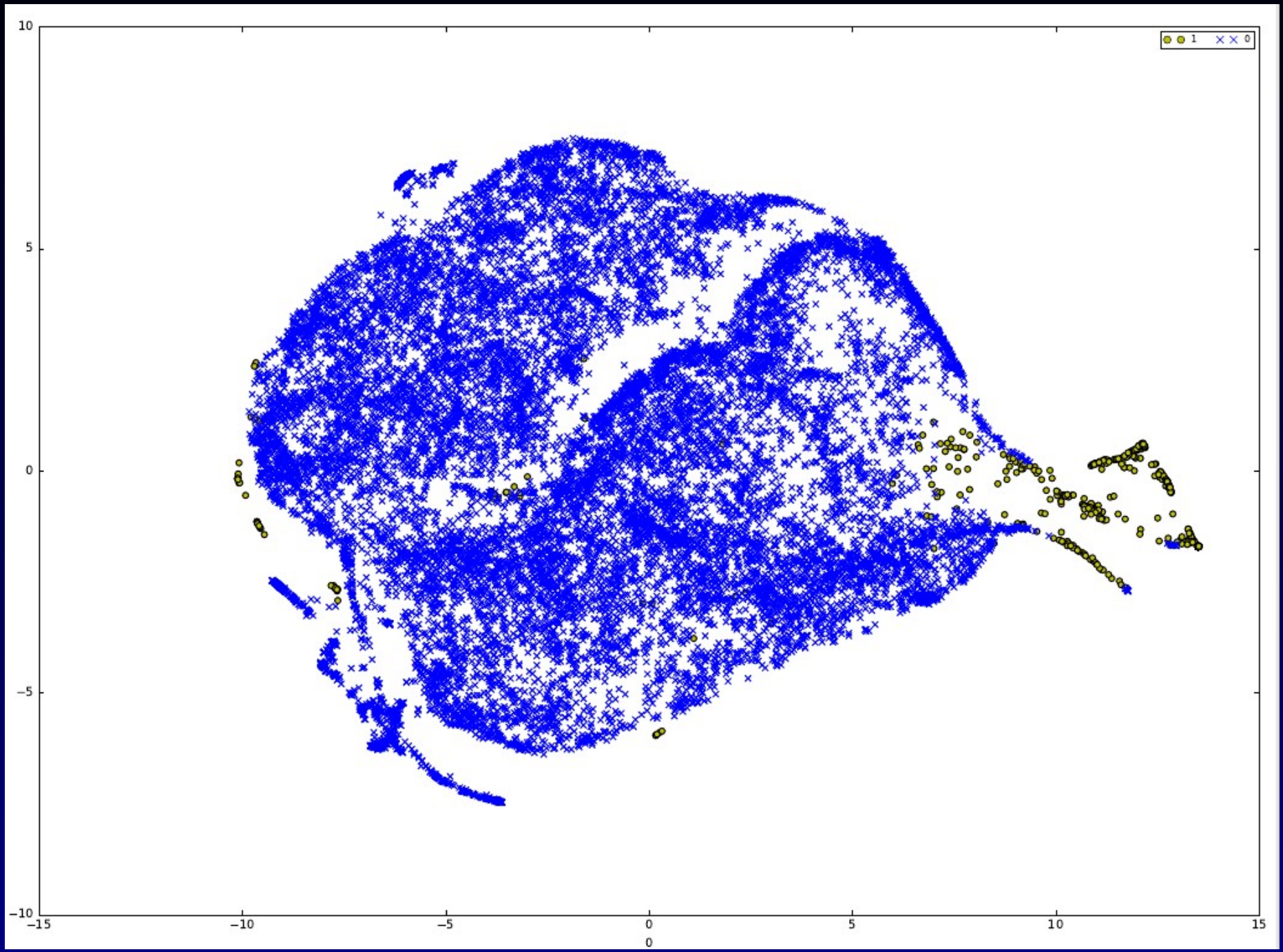


OND R=1800



LAMOST

TSNE Structure



Semi-Supervised Training

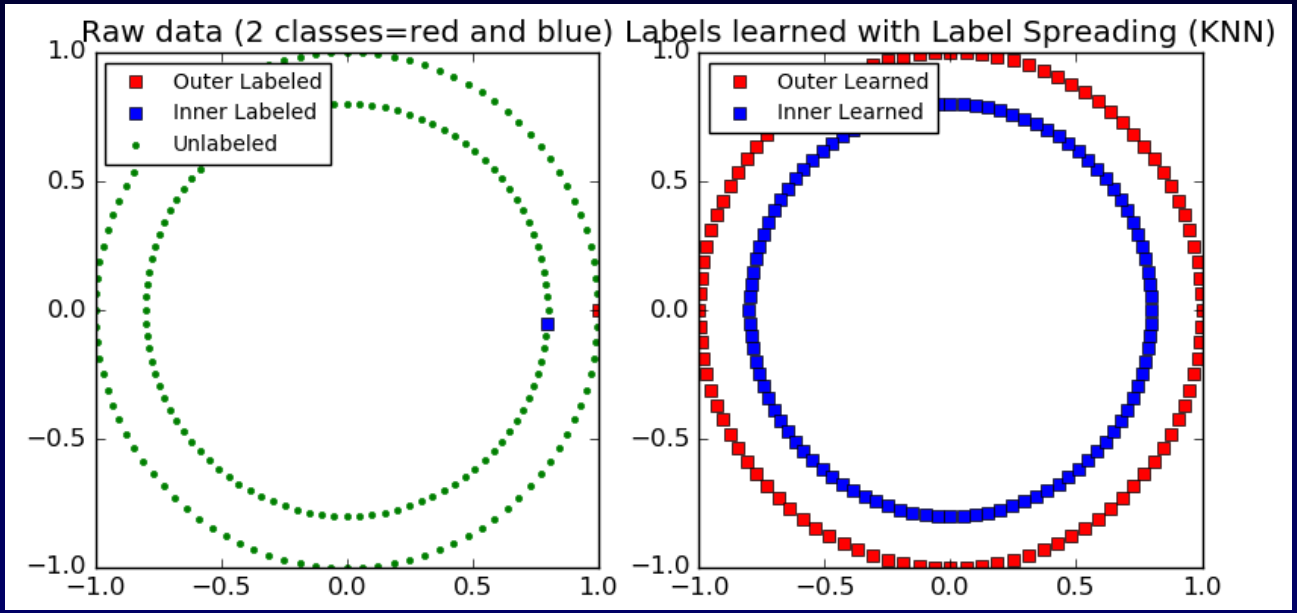
Not supervised (even if not Domain Adaptation)

- sample of labelled data - about 1600
- unlabelled (LAMOST) - HDFS limit to 1,048,576 (2^20)

Graph methods:

Label spreading

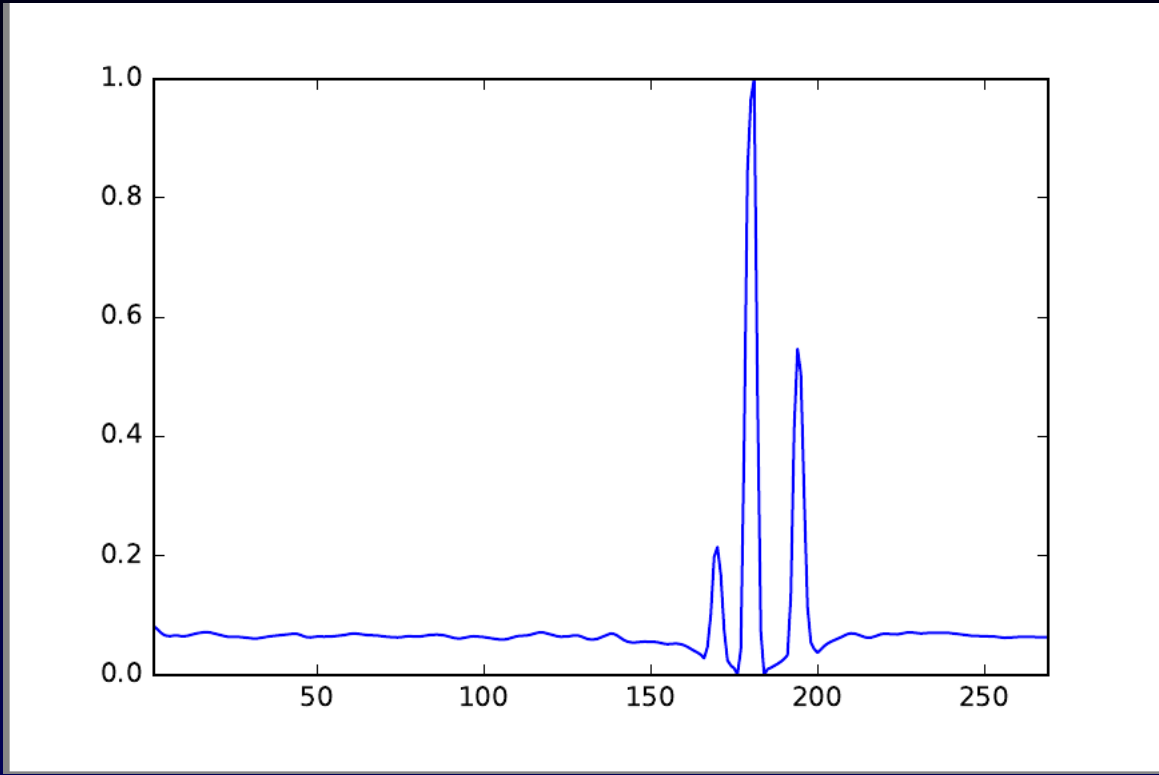
Label propagation

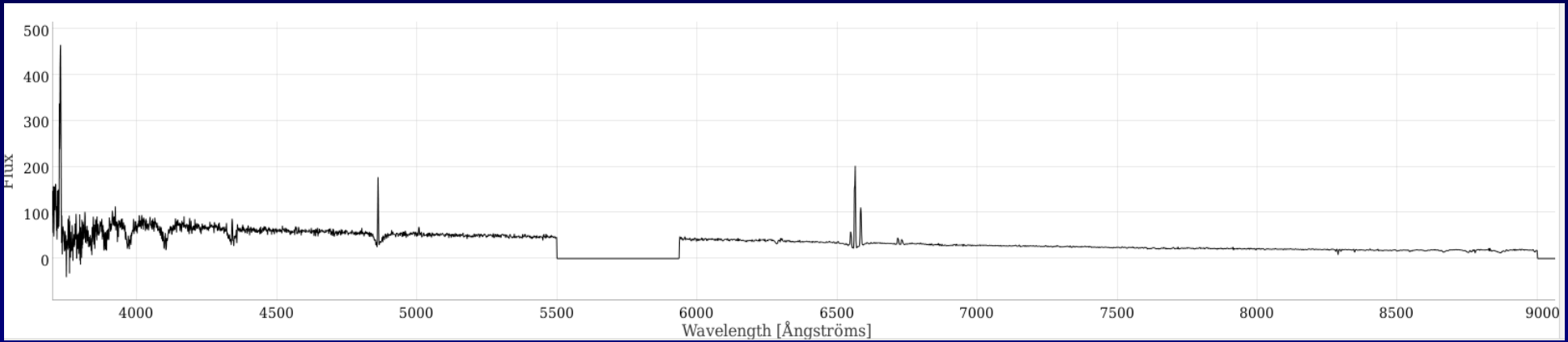


Spark on HDFS - National cloud MetaCentrum

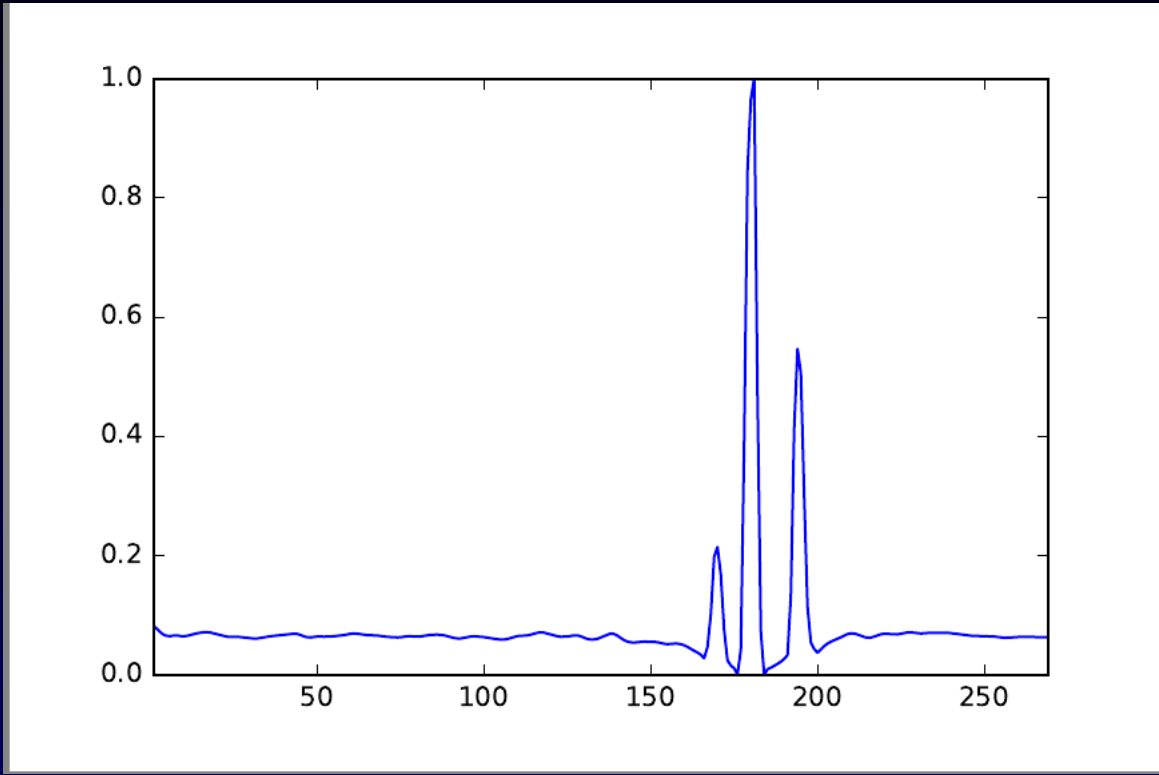
24 x16-core nodes ~ 380 nodes (real load dependent)

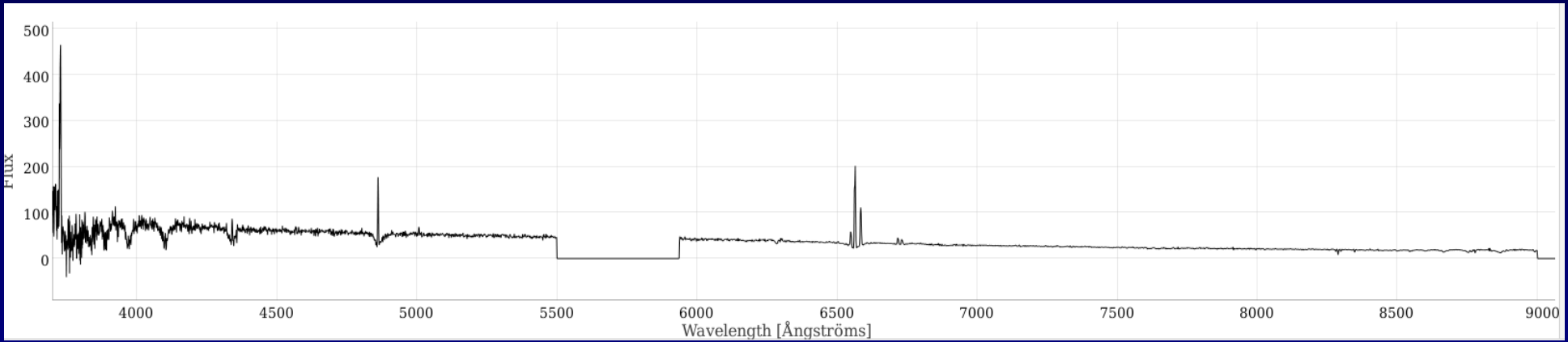
Be Candidates Foud



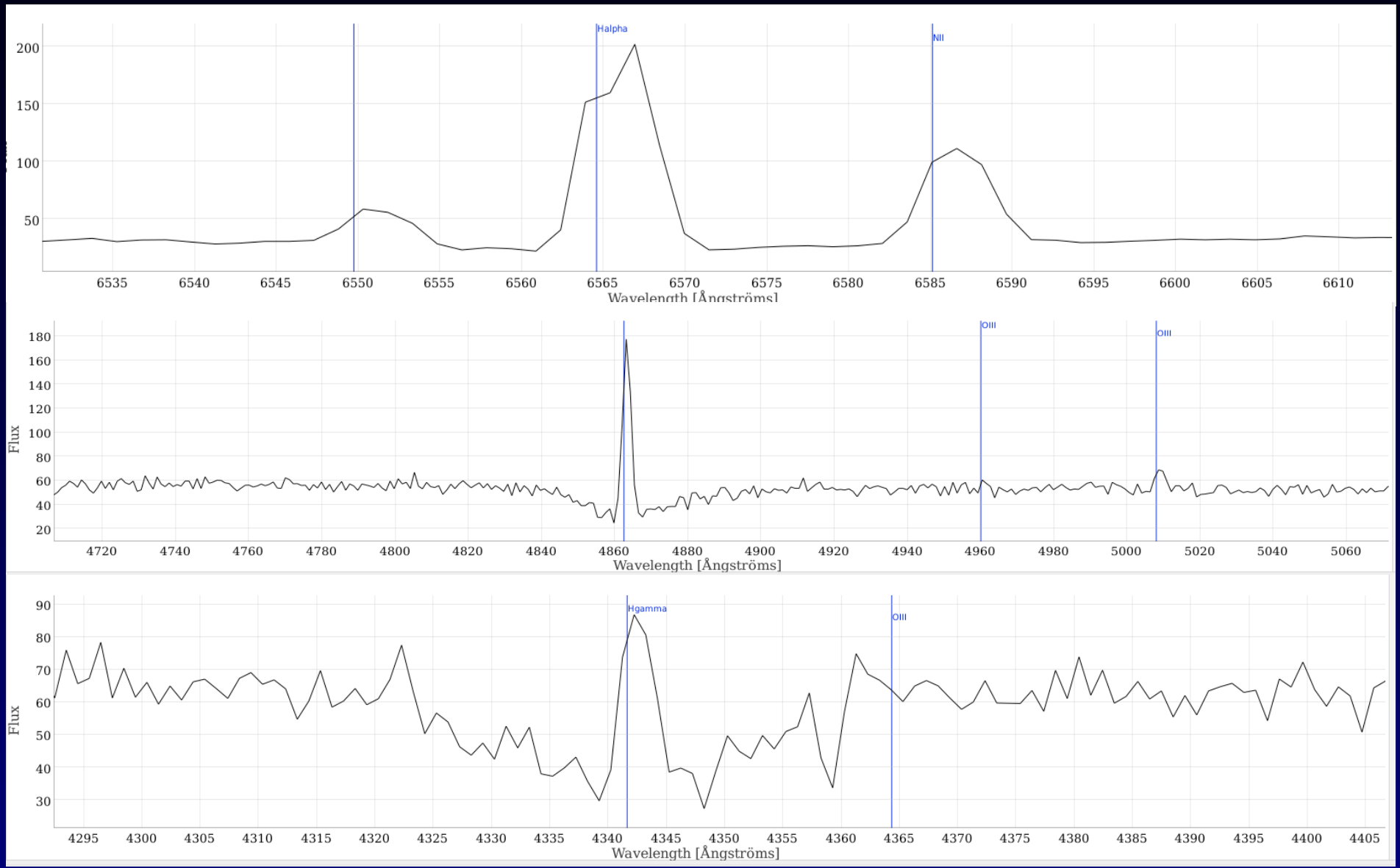


Be Candidates Foud

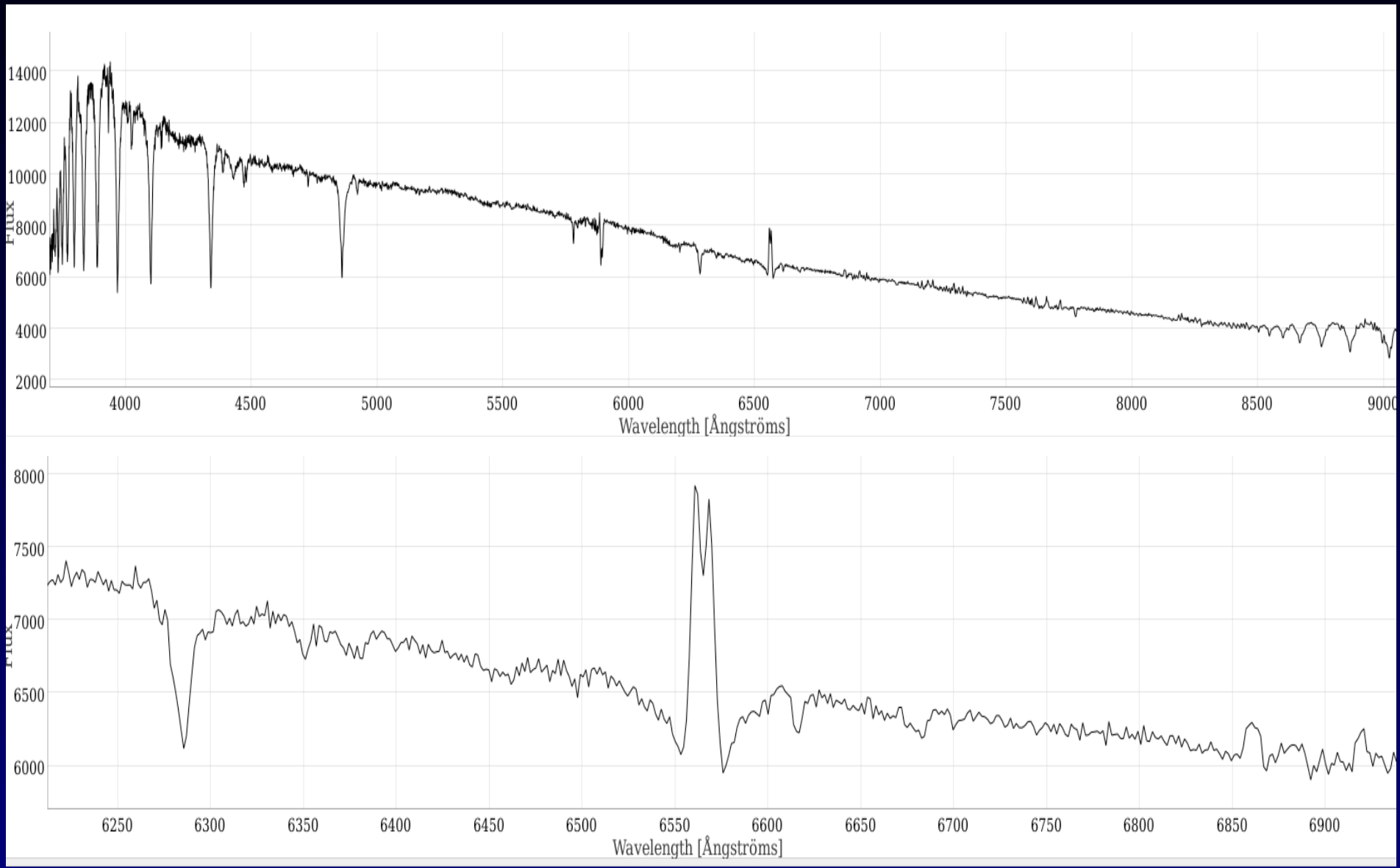




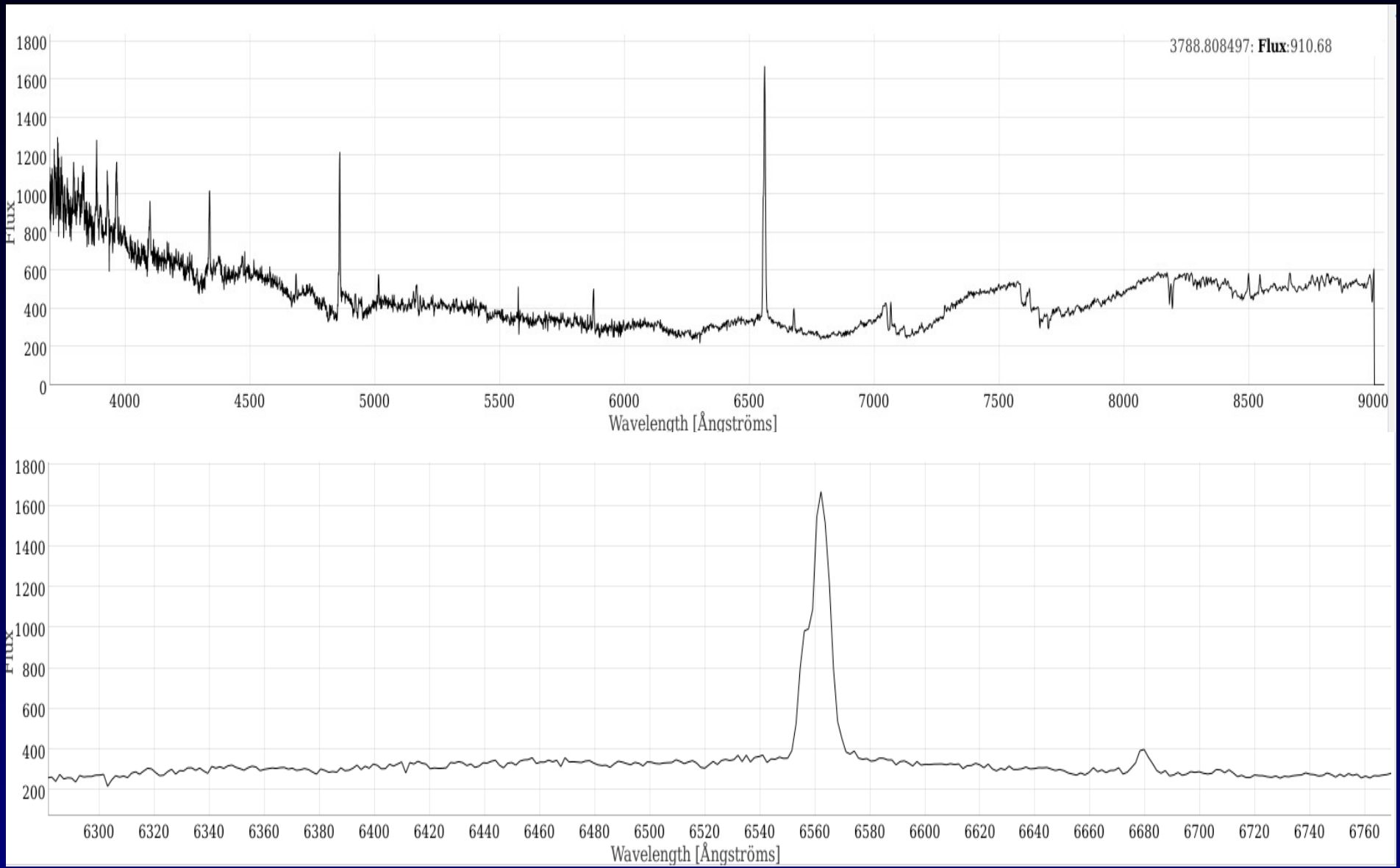
Be Candidates Found



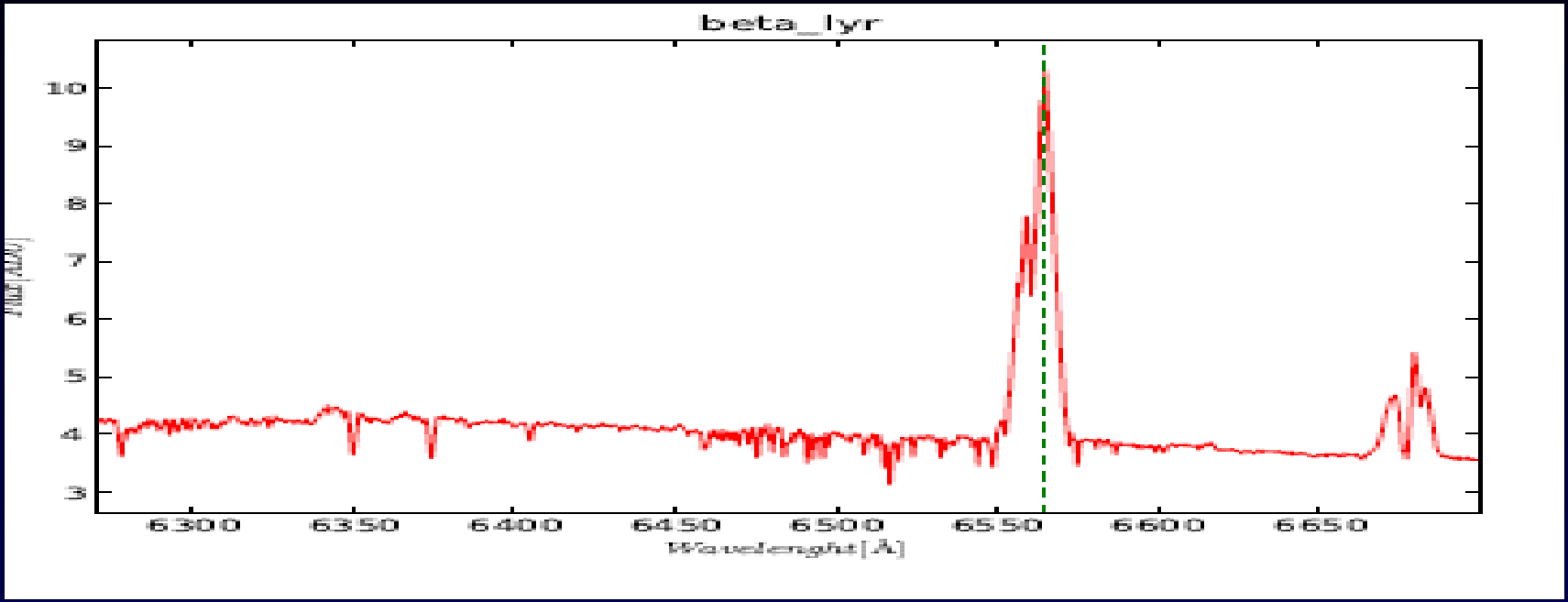
Be Candidates Found

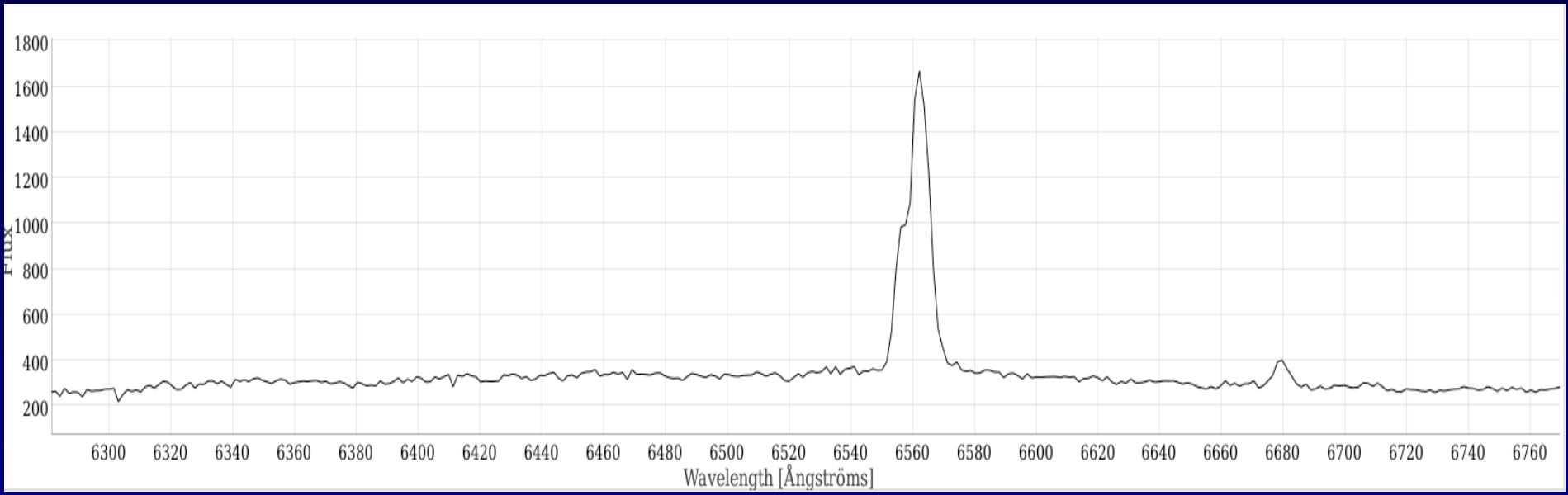


Be Candidates Found

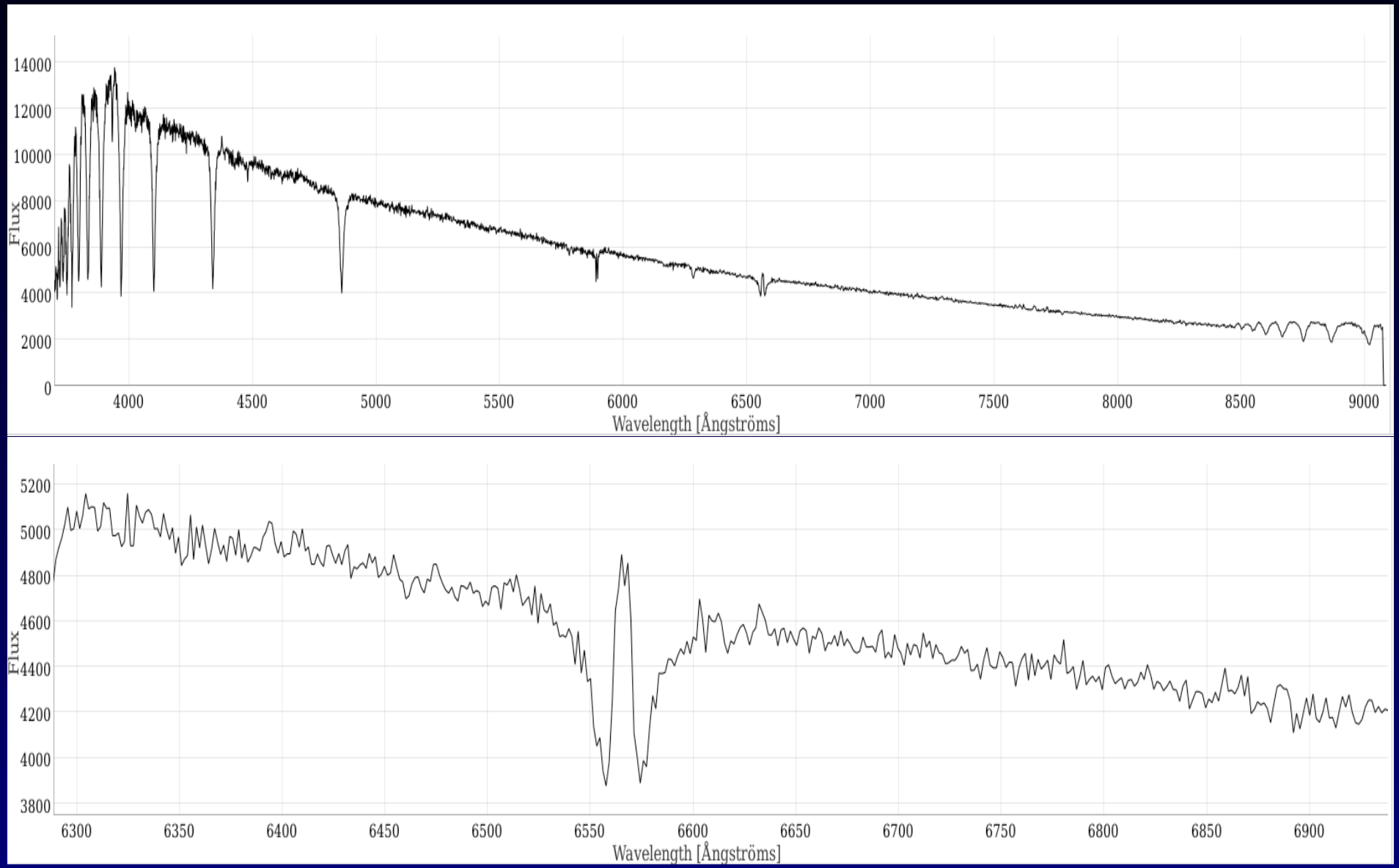


Be Candidates Found

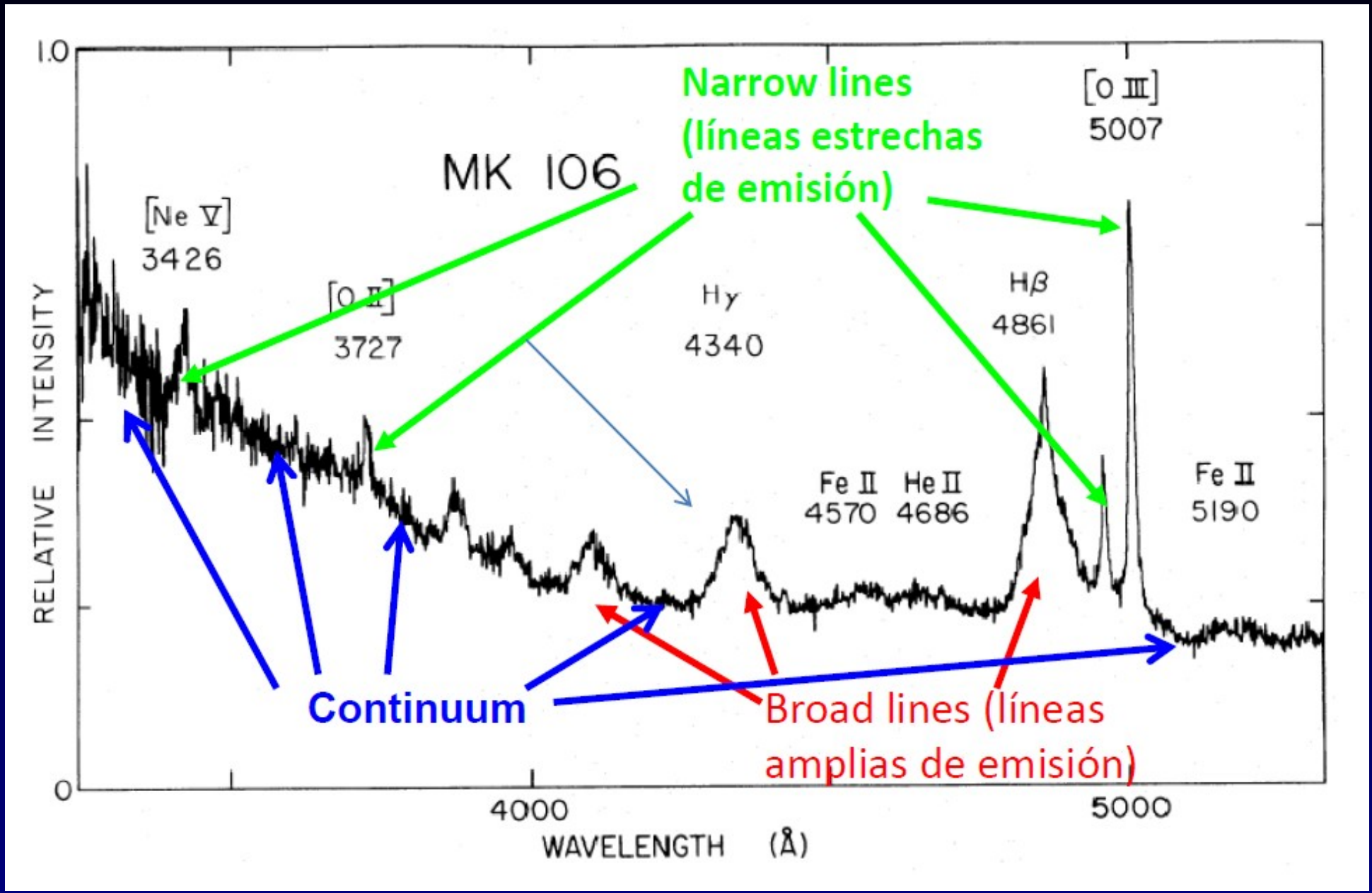




Yet Unknown Be Star

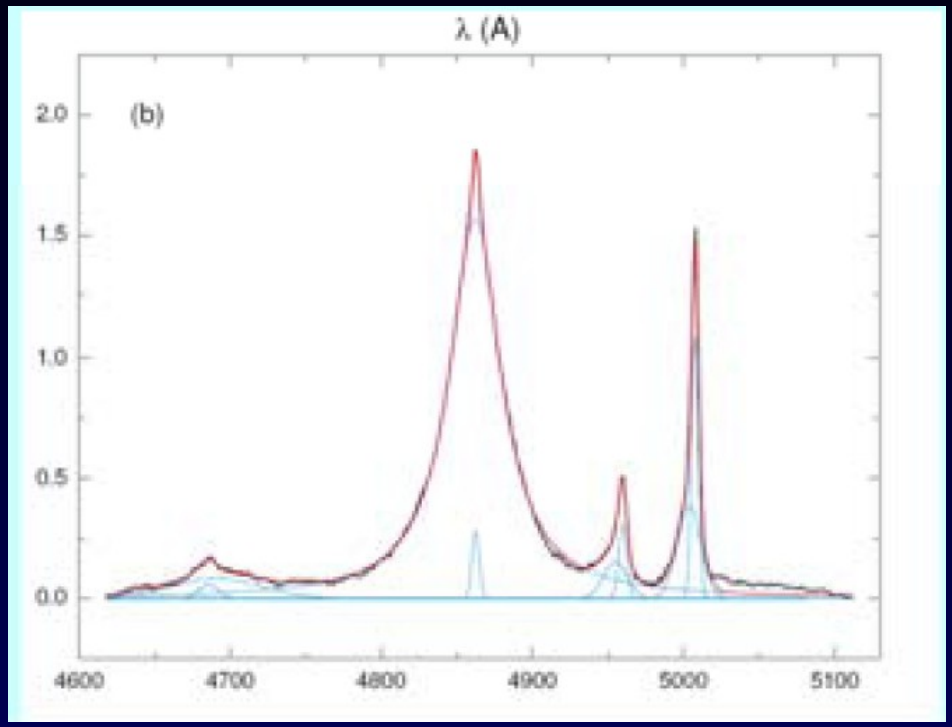


AGN Spectrum

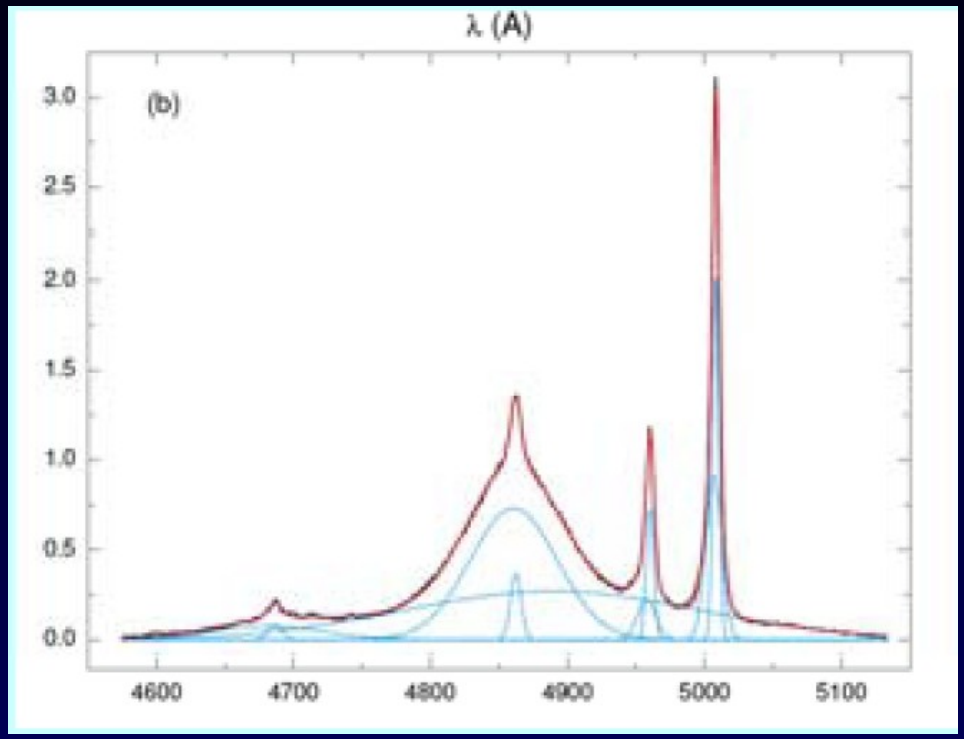


Gaskell 2009

AGN Populations



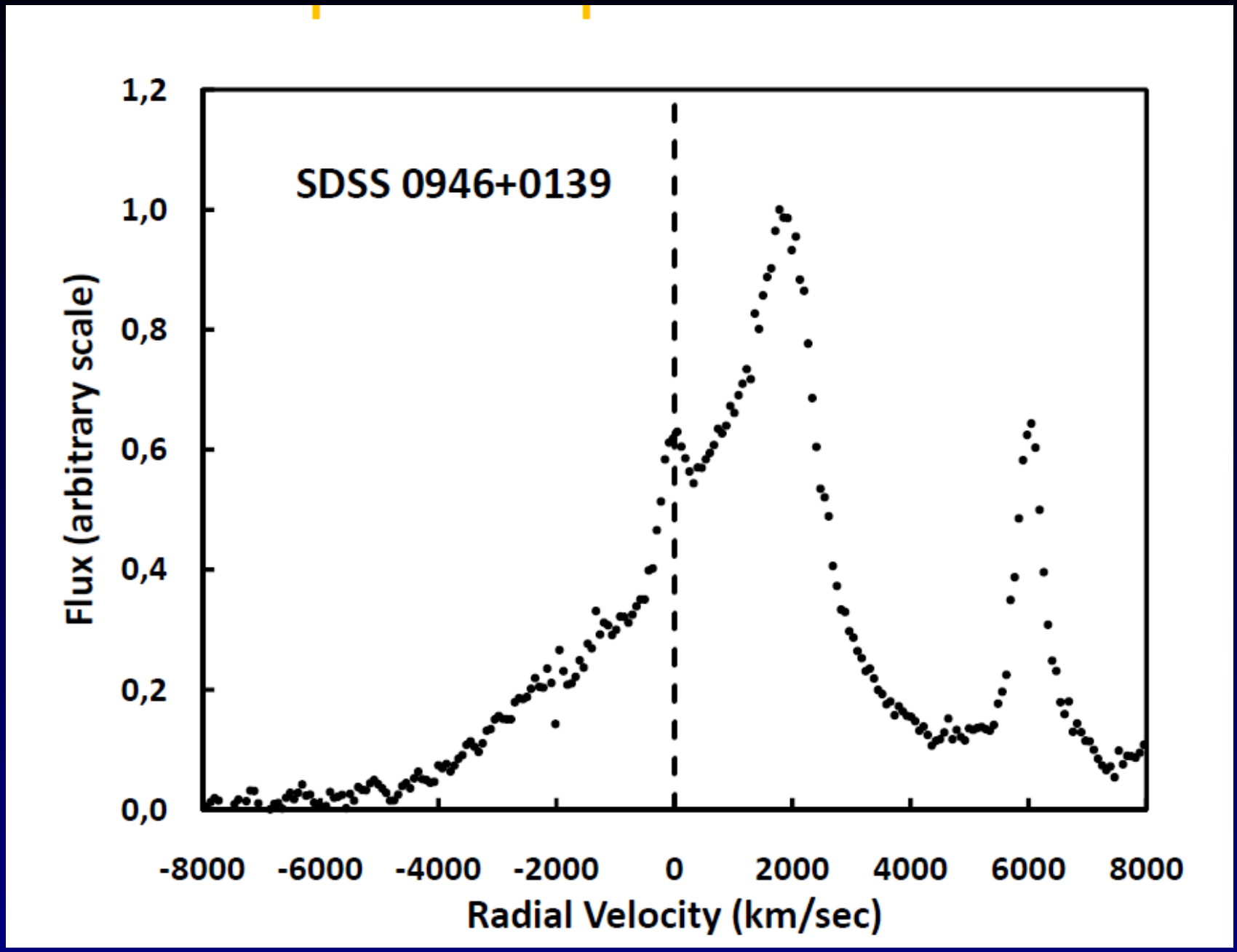
Population A



Population B

Sulentic et al. 2002

Extreme AGN Spectra



Virtual Observatory inside

- OND 2m archive on SSAP protocol (spectra access)
- LAMOST DR1 on SSAP (using DaCHS)
- Preprocessing (rectify, cutout) – DataLink on server
- SAMP (send spectra to SPLAT-VO - view details)
- Visualization of results

- VO-CLOUD – cloud engine based on UWS REST jobs
- Cross-matching (ADQL, TAP, TOPCAT, TAPhandle, pyVO, Vizier)

Conclusions

- Machine learning on big spectra archives may identify new interesting objects yet unknown
- Crucial is interactive visualization of candidates
- VO technology helps in every step

- Future vision – New VO-Cloud
  - Maintains search, acquisition and processing parameters
  - Distributes work on Spark /HDFS
  - Visualizes candidates + original spectra
  - VO server using metadata in CSV