

# VO-Cloud

## A Cloud-based Science Platform for *Active* Machine Learning

P. Škoda<sup>1,2</sup>, O. Podsztavek<sup>1</sup>, J.Koza and T.Mazel<sup>1</sup>

<sup>1</sup>Faculty of Information Technology, Czech Technical University in Prague

<sup>2</sup>Astronomical Institute of the Czech Academy of Sciences, Ondřejov

IVOA GWS Session, Groningen, 11<sup>th</sup> October 2019



**RESEARCH  
CENTER FOR  
INFORMATICS**

[rci.cvut.cz](http://rci.cvut.cz)



EUROPEAN UNION  
European Structural and Investment Funds  
Operational Programme Research,  
Development and Education



MINISTRY OF EDUCATION,  
YOUTH AND SPORTS

CZECH TECHNICAL UNIVERSITY IN PRAGUE  
FACULTY OF INFORMATION TECHNOLOGY  
DEPARTMENT OF SOFTWARE ENGINEERING



Bachelor's thesis

## VO-KOREL, server for astronomical cloud computing

*Lumír Mrkva*

Supervisor: RNDr. Petr Škoda, CSc.

18th May 2012

CZECH TECHNICAL UNIVERSITY IN PRAGUE  
FACULTY OF INFORMATION TECHNOLOGY  
DEPARTMENT OF SOFTWARE ENGINEERING



Bachelor's thesis

## Design and implementation of a distributed platform for data mining of astronomical spectra archives

*Jakub Koza*

Supervisor: RNDr. Petr Škoda, CSc.

12th May 2015

CZECH TECHNICAL UNIVERSITY IN PRAGUE  
FACULTY OF INFORMATION TECHNOLOGY  
DEPARTMENT OF SOFTWARE ENGINEERING



Master's thesis

## Interactive Cloud-Based Platform for Parallelized Machine Learning of Astronomical Big Data

*Bc. Jakub Koza*

Supervisor: RNDr. Petr Škoda, CSc.

9th May 2017

# LAMOST Spectral Surveys

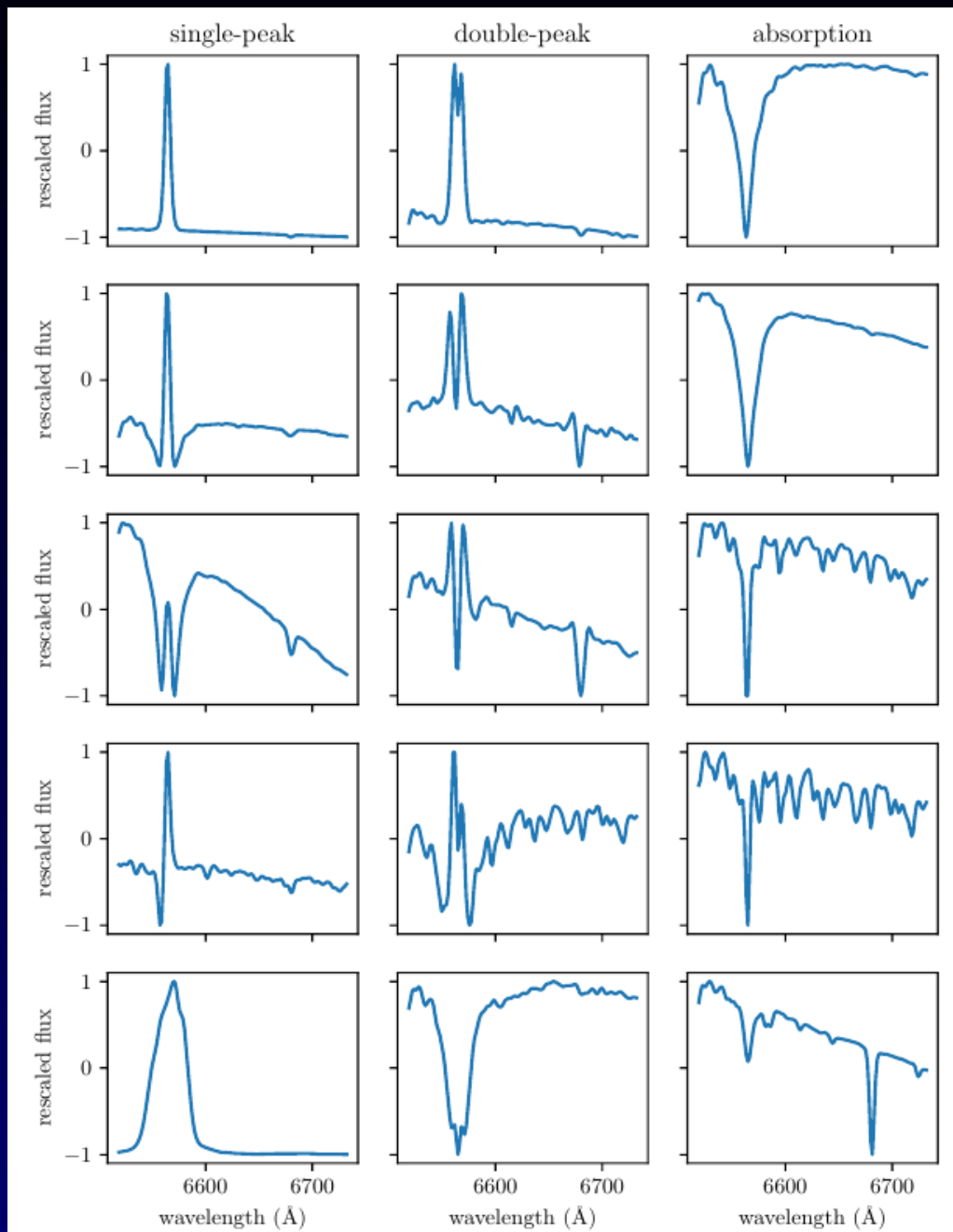
DR1 (end 2013)	<b>2 204 860</b> spectra	1 085 404 stars classified by pipeline
DR2 ( beg 2015)	<b>4 132 782</b> spectra	3 779 674 stars 307 000 unknown!
DR5 (half 2017)	<b>9 017 844</b> spectra	
DR6 (half 2018)	<b>+ 739 006</b> <b>+ 249 591 low res.</b> <b>+ 3 508 695 mid res.</b>	

Each fibre – 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 single-peak or double-peak (2+1)
- Still not enough labels for DL!

Ondřejov Dataset  
<https://zenodo.org/record/2640971>

# Concept of scientific „CLOUD“

ITERATIVE REPEATING of SAME computation (workflow)

Machine Learning (of emission line profiles of LAMOST)

LARGE **stable** INPUT data + small **changing** PARAMS

Many runs on SAME data (tuning required)

Graphics **visualization** from postprocessed output (text) files

Using WWW **browser** - supercomputing in PDA/mobil

# VO-CLOUD Architecture

Distributed engine - control by UWS 1.0

## **MASTER** (frontend)

Database of users and their experiments

Visualization

Scheduling

Load balancing

**SHARED DATA STORAGE** - controlled access (Big Data)

## **WORKERS** (backend)

Computation

# Sources of Spectra

## Getting spectra + store

(restricted access – big files)

### Files

UPLOAD from given local directory (recursive)

DOWNLOAD by http + index, FTP (recursive)

### VOTable

UPLOAD VOTable (e.g. prepared in TOPCAT - meta)

REMOTE VOTable

SSAP query + Accref

+ DataLink + SODA

SAMP control - send to SPLAT - **https ???**

# VO-CLOUD spectra visualisation

VO-CLOUD MANAGE FILESYSTEM

Home Manage filesystem Jobs Download history Create job Jupyter Settings Admin Help Logout (admin)

New Folder Append new files Delete items Download data

DATA > allond700

Name	Operation
vb040037.fits	Download Rename View content
ue210040.fits	Download Rename View content
sh180024.fits	Download Rename View content
rd260041.fits	Download Rename View content
vd040029.fits	Download Rename View content
a201503070034.fits	Download Rename View content
a201503240017.fits	Download Rename View content
rg080029.fits	Download Rename View content
th010022.fits	Download Rename View content
a201503040022.fits	Download Rename View content
a201502150025.fits	Download Rename View content
a201503080034.fits	Download Rename View content
ti060011.fits	Download Rename View content
va270017.fits	Download Rename View content
a201502200048.fits	Download Rename View content
sh150027.fits	Download Rename View content
a201504060004.fits	Download Rename View content
ue250024.fits	Download Rename View content

Spectra plotter

Figure 140706134745720

sh180024.fits: Altair  
th010022.fits: HD190603  
sh150027.fits: HD190603  
ti060011.fits: HD164353

png zoom rect

43.2 kB Apr 26, 2017 1:05:41 PM

43.2 kB Apr 26, 2017 1:05:40 PM

43.2 kB Apr 26, 2017 1:05:41 PM

43.2 kB Apr 26, 2017 1:05:40 PM



# Create job

vo-cloud Create new SOM job - Iceweasel

vo-cloud Create new SO... x

vocloud-dev.asu.cas.cz/vocloud/jobs/index.xhtml

Google

Most Visited Getting Started Connecting...

## VO-CLOUD CREATE NEW SOM JOB

Home Jobs Create Settings Admin Help Logout (skoda)

**Project label:**

Description:

Email me results

**Edit config.json**

```
{
  "Name": "Stellar_spectra",
  "Algorithm": {
    "Bmu": "normal",
    "Threads": 1
  },
  "Data": {
    "Path": ["spectra.1863.4"],
    "File_type": "csv",
  }
}
```

**Upload parameters**

Please attach data with config.json file.

(c) mrq 2014 - [feedback](#)

# Job is running

Jobs - Mozilla Firefox

Jobs

https://vocloud-dev.asu.cas.cz/vocloud-betelgeuse/jobs/index.x... 90%

Most Visited Getting Started

## VO-CLOUD JOBS

Home Manage filesystem Jobs Download history Create job Jupyter Settings Admin Help Logout (skoda)

Success New Active\_learning job was successfully enqueued

Show jobs of all users

Type	ID	Job label	Created	Duration	Phase	Action	Delete	Details
Active_learning	10-716	active-learning-demo	10/9/19 12:53:05 AM	0 sec	EXECUTING	Abort	x	📄
Active_learning	10-710	active-learning(copy)	10/8/19 4:51:30 PM	3 sec	COMPLETED		x	📄
Active_learning	10-709	lamost sample	10/8/19 4:38:14 PM	4 sec	ERROR		x	📄
SOM	10-708	AllSpecOndSOM(copy)(copy)(copy)(copy)(copy)	10/6/19 11:39:15 PM	5 sec	COMPLETED		x	📄
SOM	10-706	AllSpecOndSOM(copy)(copy)(copy)(copy)	10/6/19 11:34:16 PM	8 sec	COMPLETED		x	📄
Active_learning	10-705	my1	10/6/19 11:30:23 PM	2 sec	COMPLETED		x	📄
Active_learning	10-512	active-learning-test(copy)(copy)(copy)(copy)(copy)(cc	9/21/19 7:14:48 PM	2 sec	COMPLETED		x	📄
Active_learning	10-362	active-learning-test(copy)(copy)(copy)(copy)(copy)(cc	5/13/19 11:23:09 AM	2 sec	COMPLETED		x	📄
Active_learning	10-361	active-learning-test(copy)(copy)(copy)(copy)(copy)(cc	5/6/19 1:30:53 PM	2 sec	COMPLETED		x	📄
SOM	10-159	AllSpecOndSOM(copy)(copy)(copy)	3/5/19 6:15:33 PM	6 sec	COMPLETED		x	📄
Preprocessing	10-157	vocloud2 test(copy)	2/28/19 5:54:03 PM	9 sec	ERROR		x	📄
SOM	10-155	AllSpecOndSOM(copy)(copy)	10/24/18 6:46:21 PM	9 sec	COMPLETED		x	📄

# SOM Worker example

VO-CLOUD DETAILS OF JOB

Home Manage filesystem Jobs Download history Create job Jupyter Settings Admin Help Logout (admin)

**AllSpecOndSOM**

Type	Id	Phase	Worker	Created	Started	Finished	Executing time
SOM	1-92	COMPLETED	local worker	2/14/17 11:32:47 PM	2/14/17 11:32:47 PM	2/14/17 11:32:59 PM	11 sec

Run again Delete

Preview

index.html - Fullscreen

Files

VO-CLOUD DETAILS OF JOB

Home Manage filesystem Jobs Download history Create job Jupyter Settings Admin Help Logout (admin)

**AllSpecOndSOM**

Type	Id	Phase	Worker	Created	Started	Finished	Executing time
SOM	1-92	COMPLETED	local worker	2/14/17 11:32:47 PM	2/14/17 11:32:47 PM	2/14/17 11:32:59 PM	11 sec

Run again Delete

Preview

index.html - Fullscreen

Neuron x: 18 y: 1

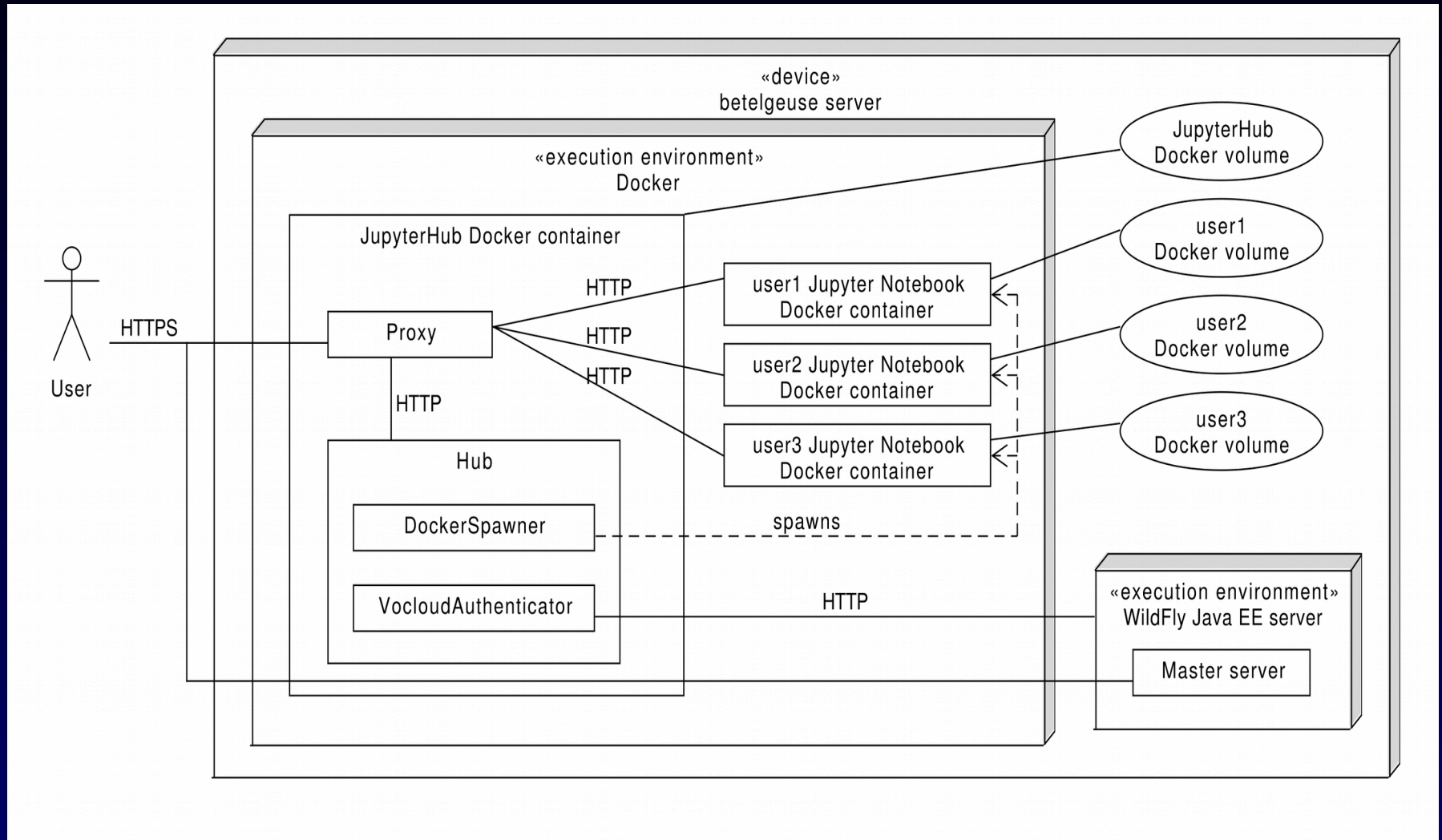
All Associated Spectra

- Y+
- X- HOME X+
- Y-
- Display reference vector
- Display all spectra

1. no name class: x
2. no name class: x
3. no name class: x
4. no name class: x
5. no name class: x
6. no name class: x
7. no name class: x
8. no name class: x
9. no name class: x

Files

# JupyterHub deployment



# JupyterHub example

jupyter example\_plotter Last Checkpoint: 05/06/2017 (unsaved changes) Control Panel Logout Python 3

File Edit View Insert Cell Kernel Widgets Help

Code CellToolbar

```
In [7]: path='filesystem/DATA/allond700/'
spectra=['sh180024.fits','th010022.fits','ti060011.fits','sh150027.fits']
files=[path + i for i in spectra]
files
```

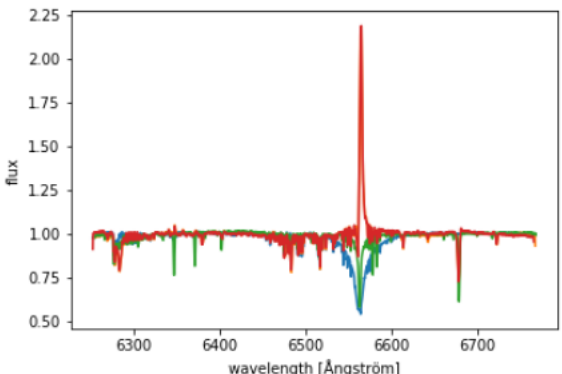
```
Out[7]: ['filesystem/DATA/allond700/sh180024.fits',
'filesystem/DATA/allond700/th010022.fits',
'filesystem/DATA/allond700/ti060011.fits',
'filesystem/DATA/allond700/sh150027.fits']
```

```
In [13]: parsed = [parse_spectrum_file(i) for i in files]
parsed[0]
```

```
Out[13]: {'flux': array([ 0.97623893,  0.97816423,  0.98200884, ...,  0.99071508,
 0.99049042,  0.98766227]),
'name': 'Altair',
'wave': array([ 6252.48405443,  6252.74072204,  6252.99738965, ...,  6764.27926559,
 6764.53593319,  6764.79260008 ])}
```

```
In [12]: for i in parsed:
plt.plot(i['wave'], i['flux'])
plt.xlabel('wavelength [Ångström]')
plt.ylabel('flux')
```

```
Out[12]: <matplotlib.text.Text at 0x7f7f108b8550>
```



# Deep Convolutional Neural Networks

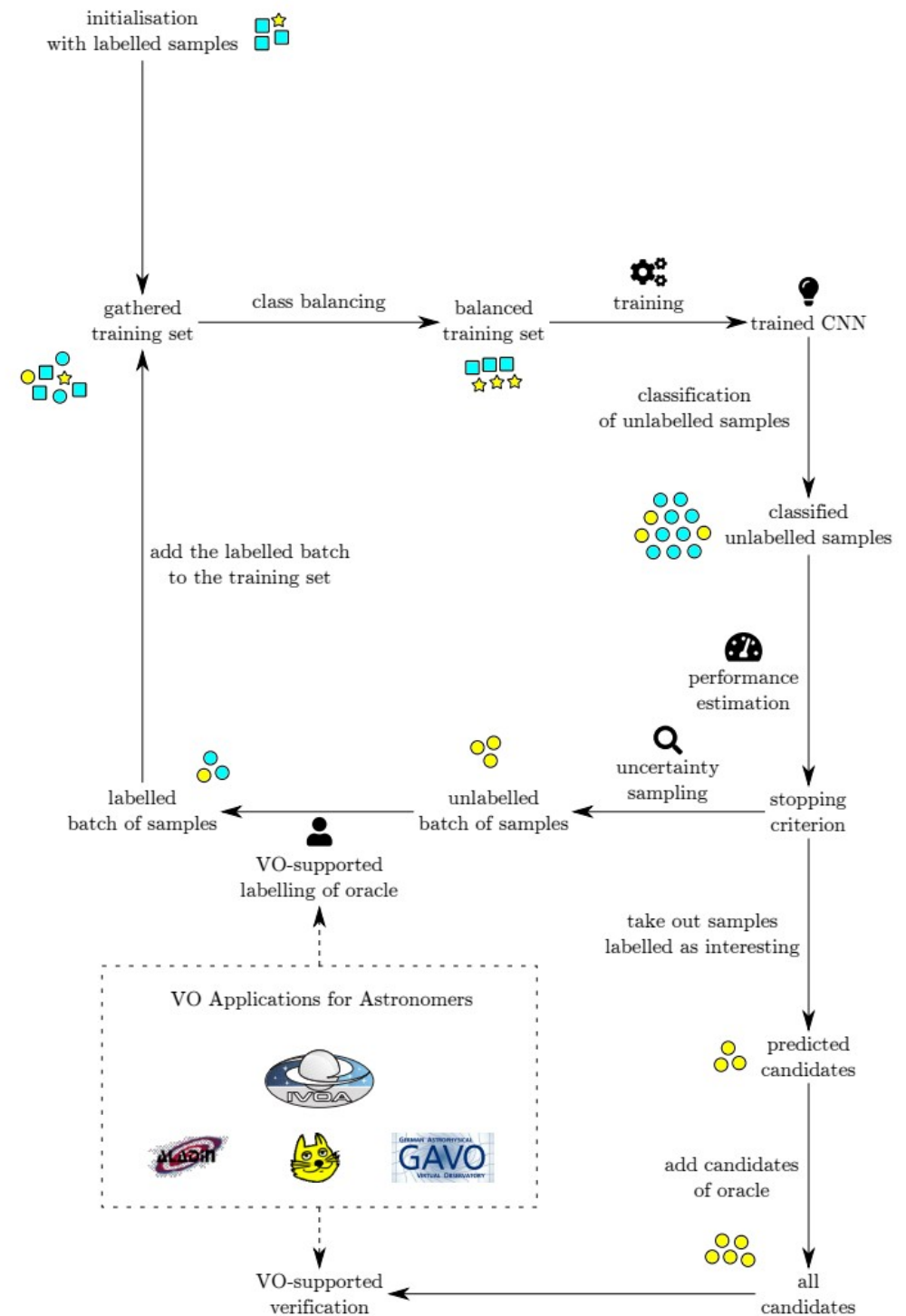
- Representational learning technique  
no feature extraction
- State-of-the-art in object recognition
- Needs huge labelled training set
- Good representativeness
- Never satisfied in science !

input (140 pixel spectrum)
conv3-64
conv3-64
maxpool2
conv3-128
conv3-128
maxpool2
conv3-256
conv3-256
maxpool2
fc-512
fc-512
softmax

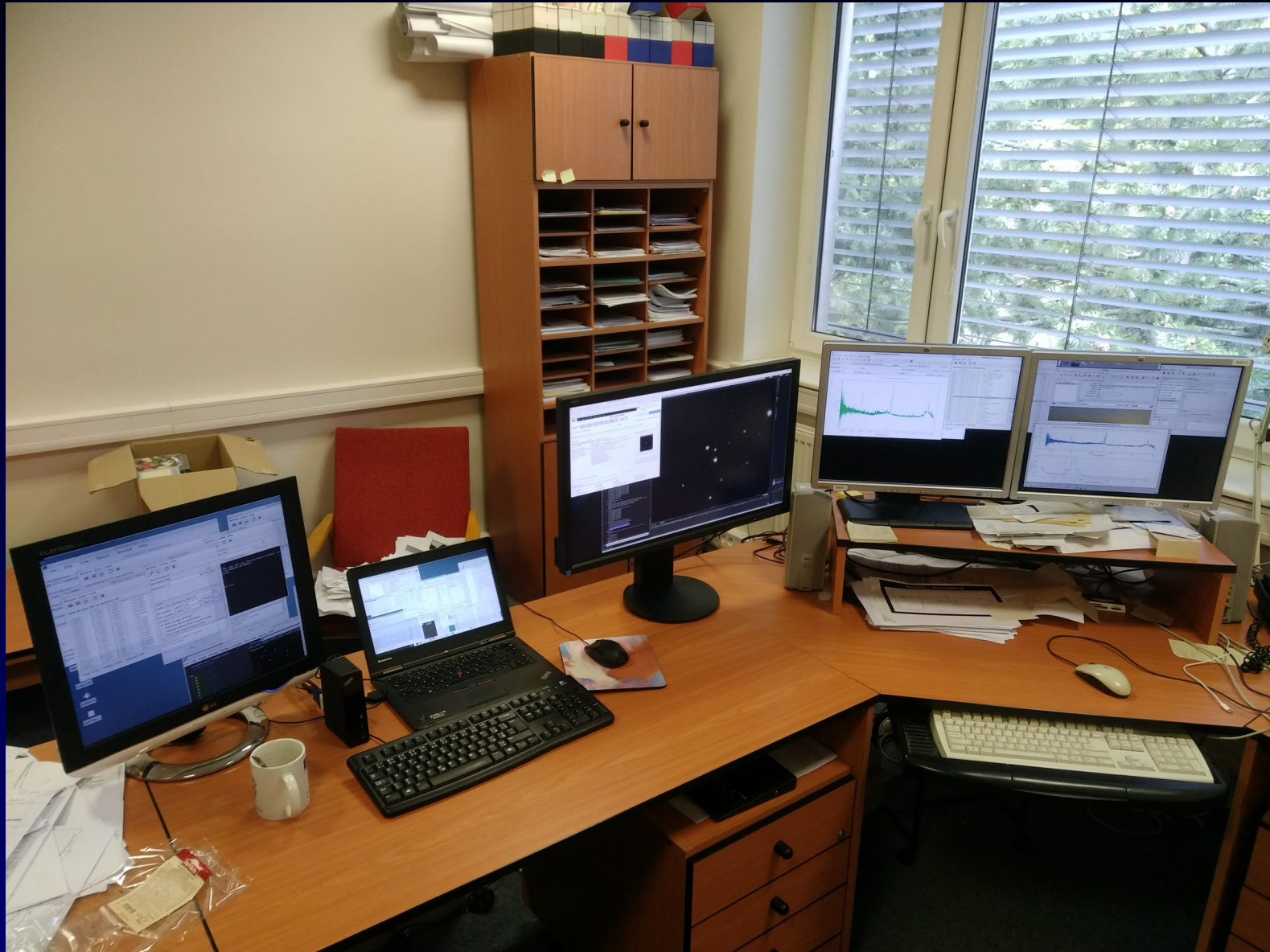
# Active Deep Learning

## CNN Chooses Data for Its Training

- Oracle classification (domain expert knowledge)
- Uncertainty sampling (entropy)
- From predicted target classes selected batch (n)
- Batch added to training set



# Analysis





# Confusion in Unique Identification

LAMOST J034912.80+240820.0

is Pleione (5mag) 22 arcsec apart

**Basic data :**

**LAMOST J034912.80+240820.0 -- Peculiar Star**

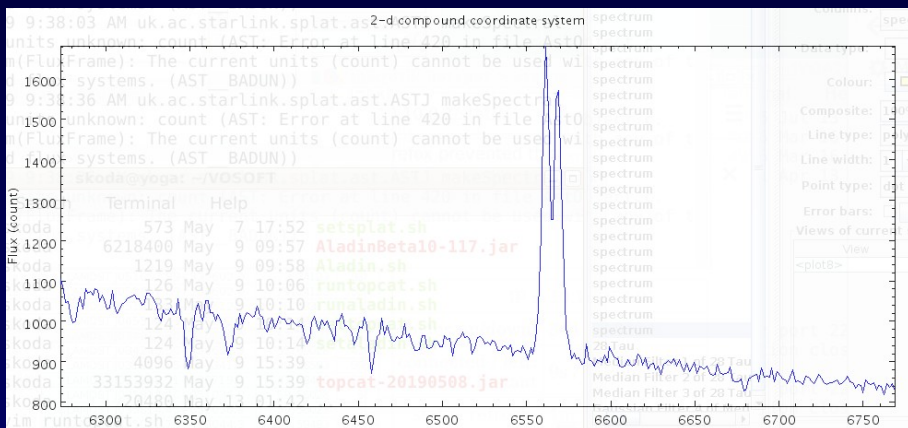
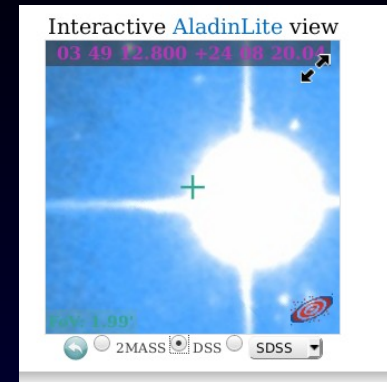
Other object types: Pe\* (Ref)

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 [ ]

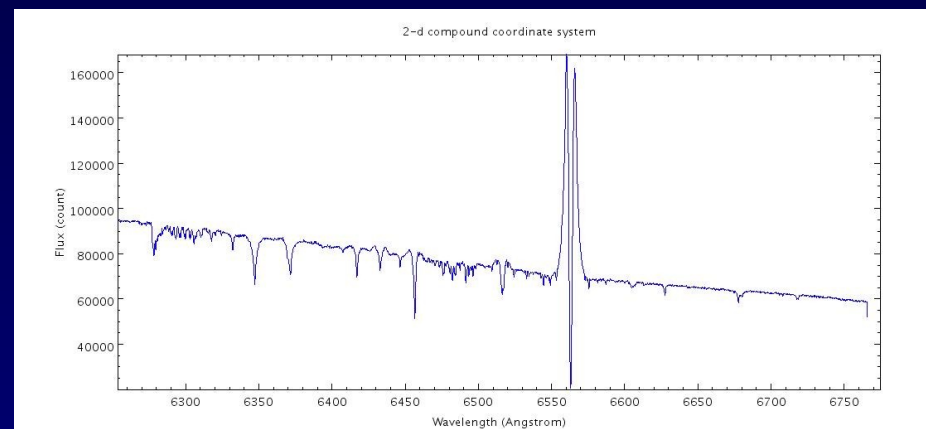
Gal coord. (ep=J2000) : 166.959918 -23.163713 [ ]

Spectral type: A1mF1 D 2015MNRAS.449.1401H



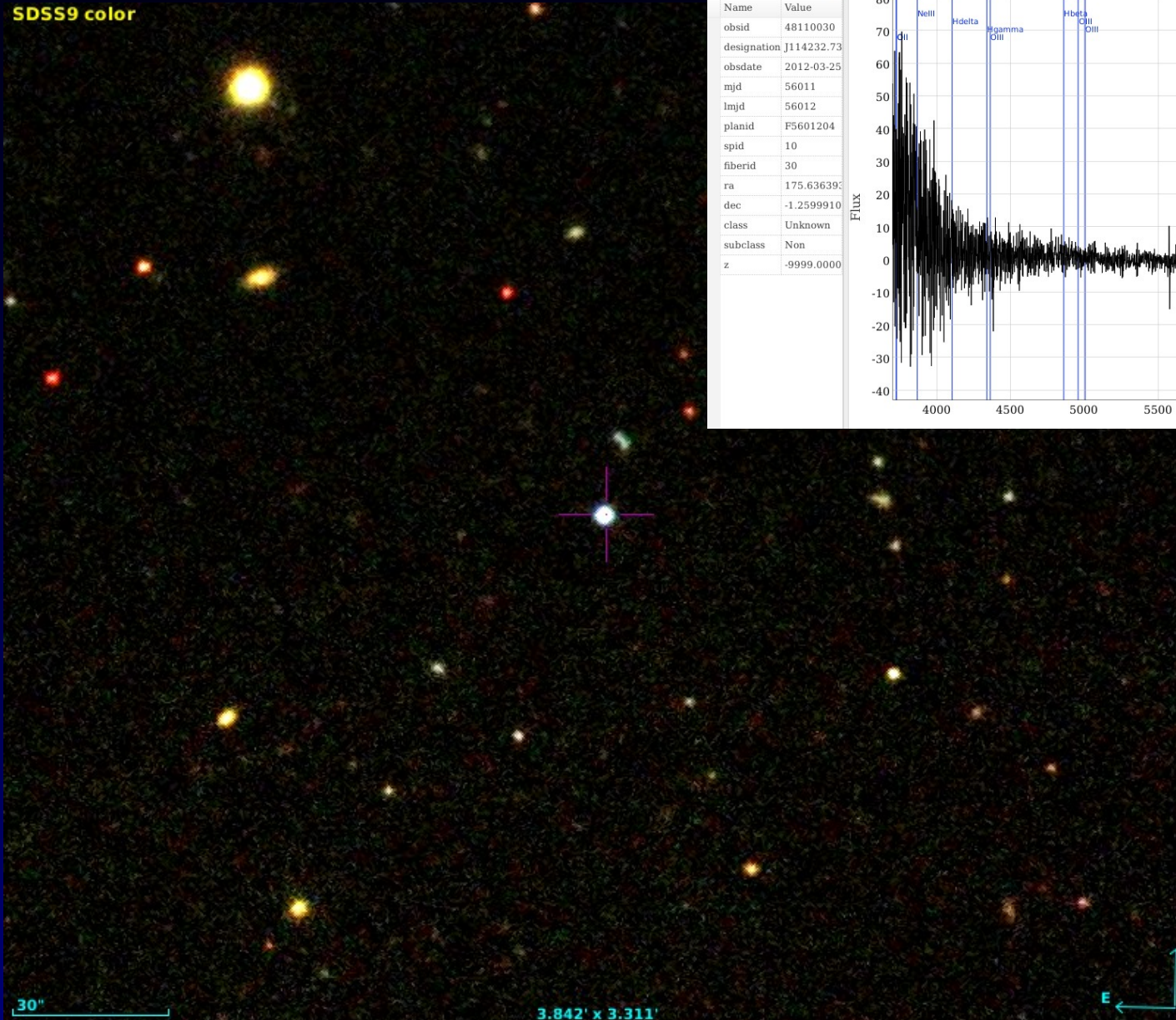
LAMOST MJD 56295

Ondrejov at MJD 56153



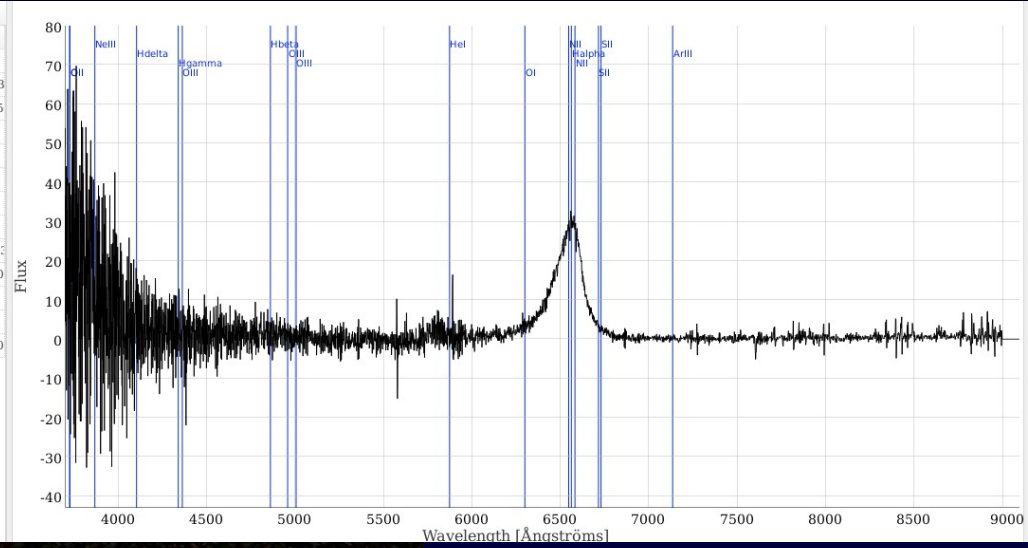
# Probable Supernova?

SDSS9 color



Information

Name	Value
obsid	48110030
designation	J114232.73
obsdate	2012-03-25
mjd	56011
lmjd	56012
planid	F5601204
spid	10
fiberid	30
ra	175.636390
dec	-1.2599910
class	Unknown
subclass	Non
z	-9999.0000



LAMOST J114232.73-011535.9.

No information in surveys

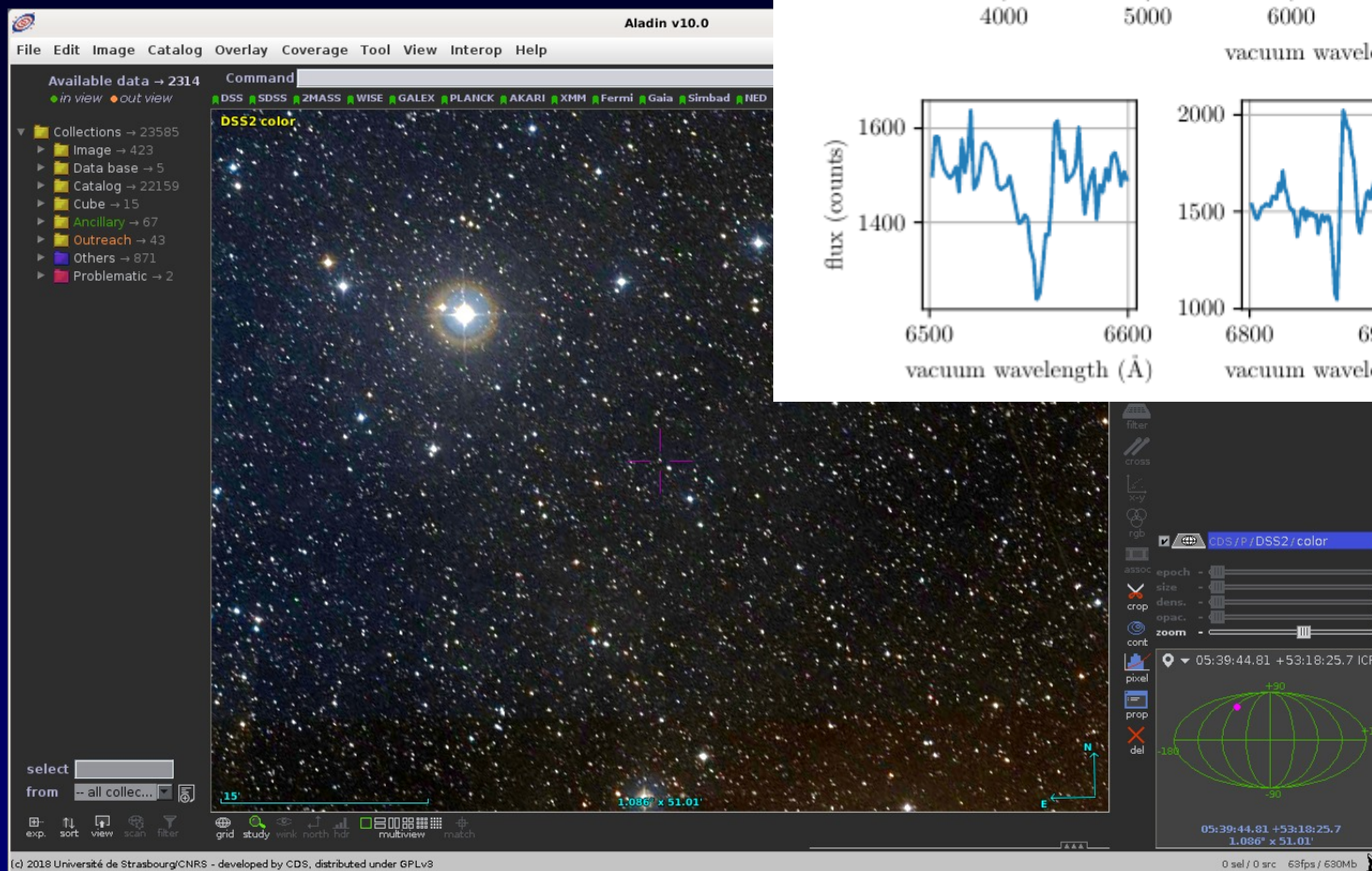
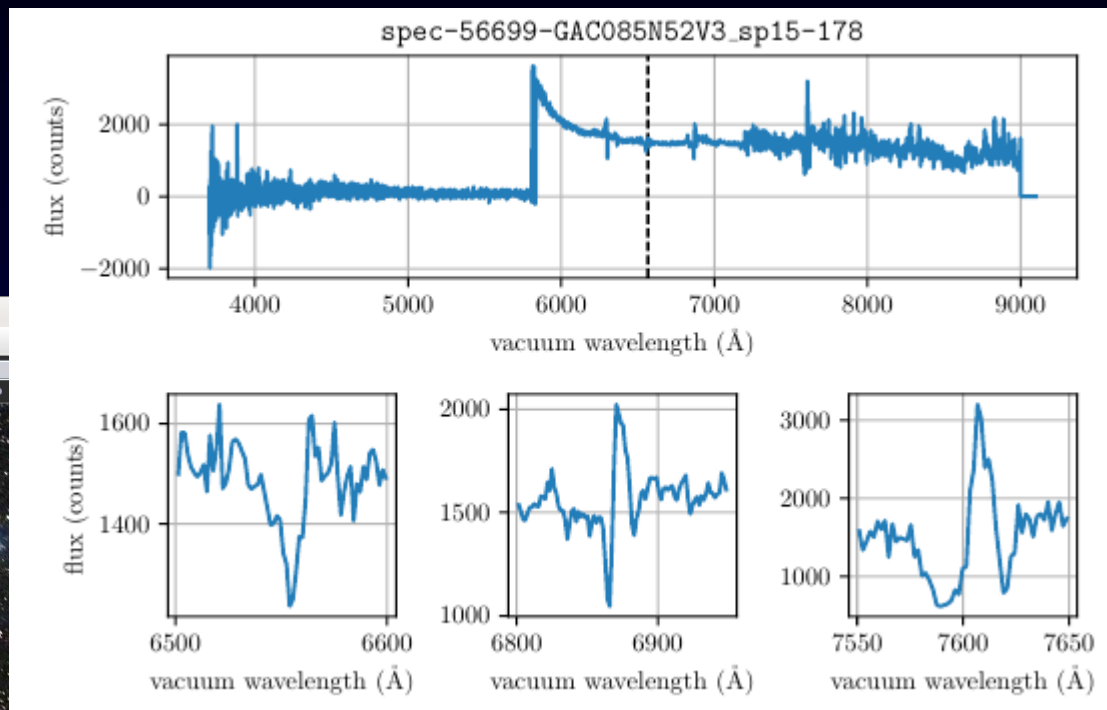
Surrounded by galaxies

30"

3.842' x 3.311'



# Visual Verification of Object with Disk



# Surprise!



- Circumstellar disk structure
- Possible exoplanet hosting star?
- Accretion disk?
- Artefact ???

# Active Learning – in preparation

VO-CLOUD DETAILS OF JOB

Home Manage filesystem Jobs Download history Create job Jupyter Settings Admin Help Logout (tomasmazel)

**active-learning**

Type	Id	Phase	Worker	Created	Started	Finished	Executing time
Active_learning	156-623	COMPLETED	local worker	10/3/19 9:17:31 AM	10/3/19 9:17:31 AM	10/3/19 9:17:33 AM	2 sec

Run again Delete

**Preview**

index.html - Fullscreen

### Spectra

Name	Ra	Dec	Prediction	Label	Iteration
vi140035	345.9691958351259	3.820027775493711	single peak	single peak	1

1-single peak  
 2-double peak  noteworthy    
 3-not sure  
 4-bad

6561.069576287892: vi140035: 3.29

- ElasticSearch
- HDF5 – Pandas
- Python scripts from jupyter nb
- Aladin Lite ?

# Conclusions

- Active learning overcomes the lack of labeled data
- A new kind of science platform needed
- ML needs to **visualise** data as part of its process now!
- Oracle requires **VO** to have
  - METADATA to decide correctly**
  - OTHER DATA (global interoperability)**
- Crucial is interactive visualization of candidates

# Thank You



**RESEARCH  
CENTER FOR  
INFORMATICS**

[rci.cvut.cz](http://rci.cvut.cz)



EUROPEAN UNION  
European Structural and Investment Funds  
Operational Programme Research,  
Development and Education



MINISTRY OF EDUCATION,  
YOUTH AND SPORTS