VO-Cloud

A Cloud-based Platform for Machine Learning

P. Škoda[1, 2], L. Mrkva, J.Koza, T.Mazel[1]

[1]Faculty of Information Technology, Czech Technical University in Prague
[2]Astronomical Institute of the Czech Academy of Sciences, Ondřejov

ADASS BoF - Science Platforms , Groningen, 9th October 2019

RESEARCH CENTER FOR INFORMATICS

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

Concept of scientific „CLOUD"

ITERATIVE REPEATING of SAME computation (workflow)

Machine Learning (of emission line profiles of LAMOST)

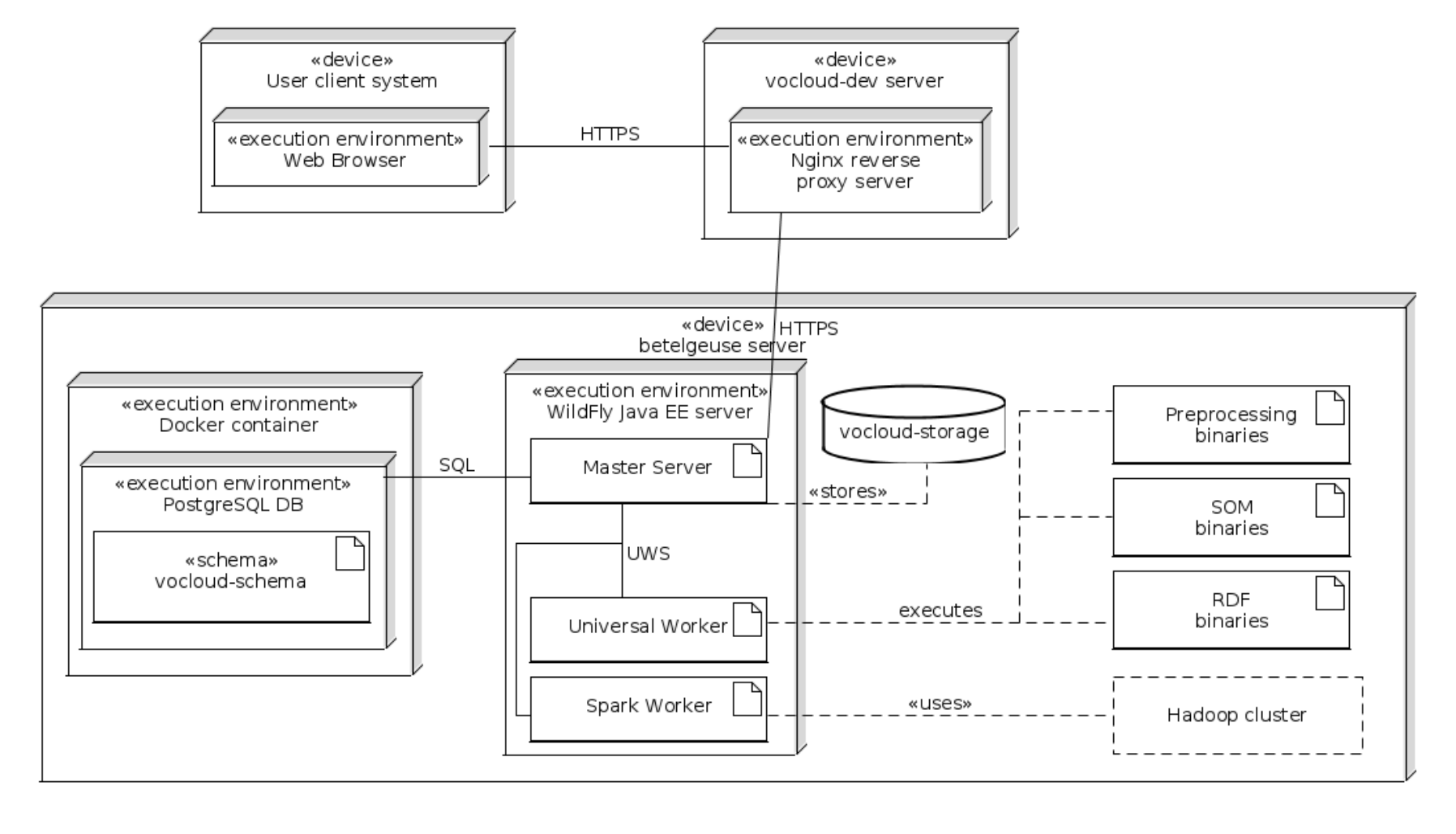
LARGE stable INPUT data + small changing PARAMS

Many runs on SAME data (tuning required)

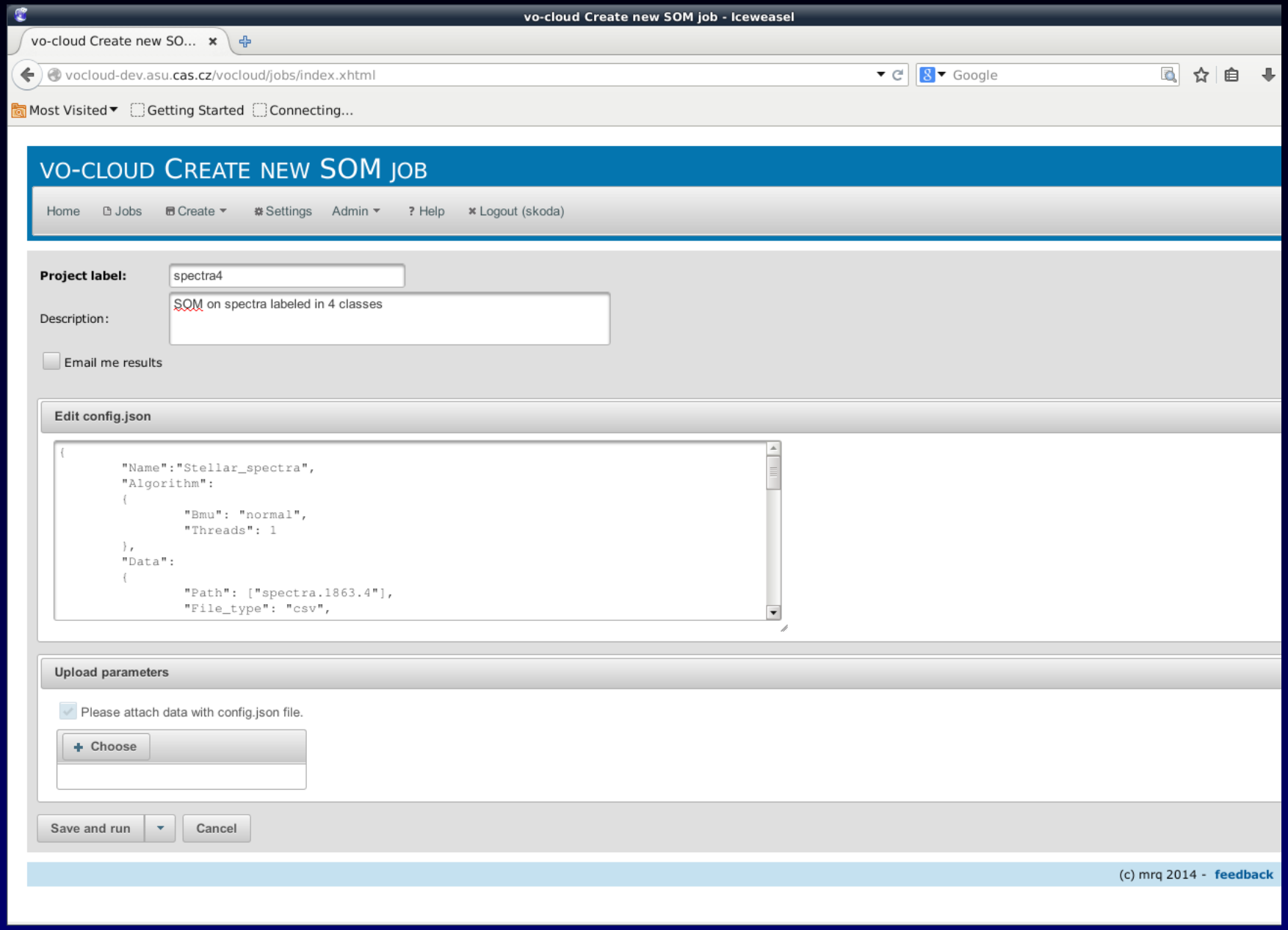
Graphics visualization from postprocessed output (text) files

Using WWW browser - supercomputing in PDA/mobil

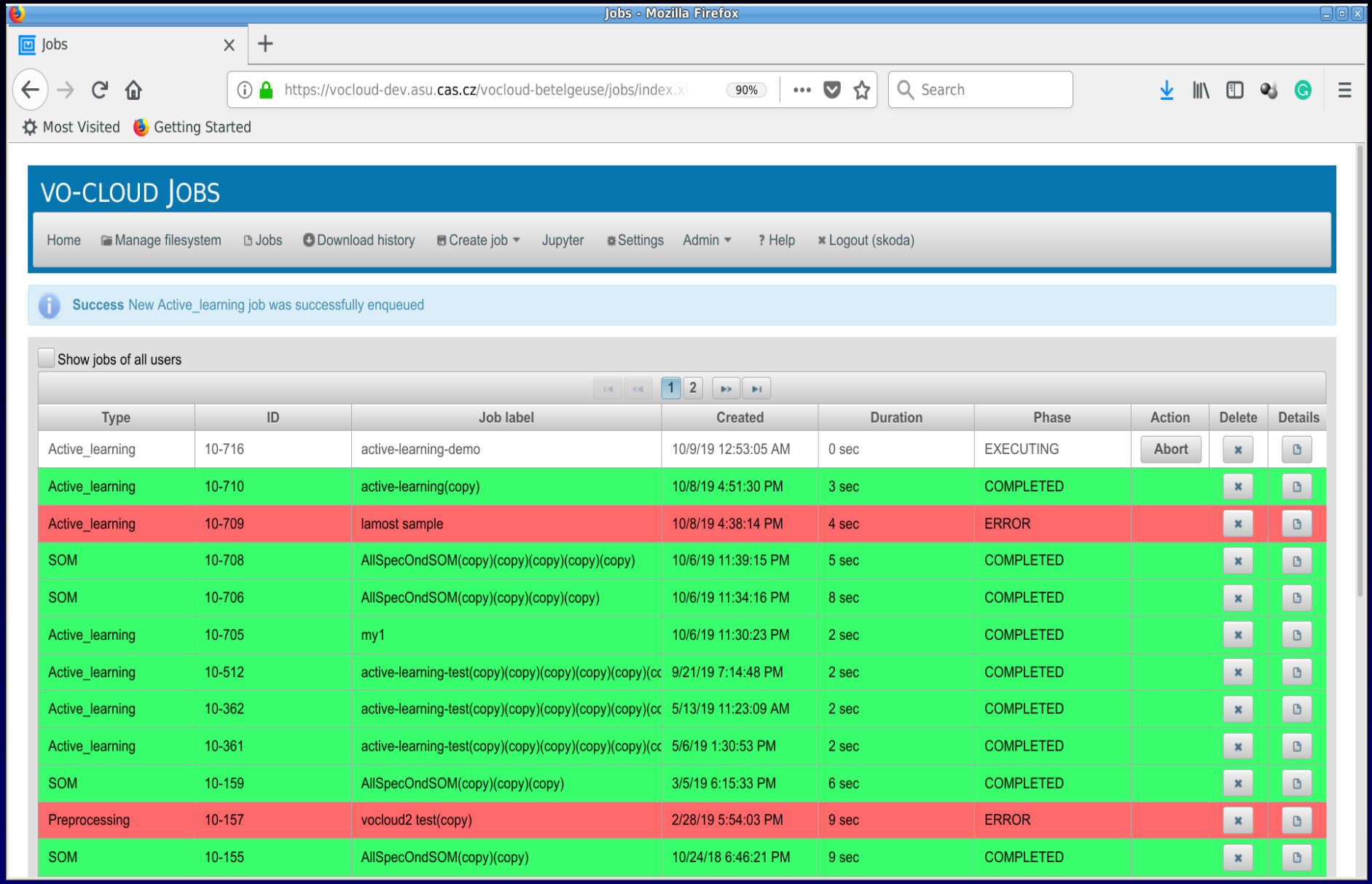
VO-CLOUD - UWS driven



Create Job


vo-cloud Create new SOM job - Iceweasel
vo-cloud Create new SO...
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:
spectra4
Description:
SOM on spectra labeled in 4 classes
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.
Choose
Save and run
Cancel
(c) mrq 2014 - feedback

Job is running



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	×	
Active_learning	10-710	active-learning(copy)	10/8/19 4:51:30 PM	3 sec	COMPLETED		×	
Active_learning	10-709	lamost sample	10/8/19 4:38:14 PM	4 sec	ERROR		×	
SOM	10-708	AllSpecOndSOM(copy)(copy)(copy)(copy)(copy)	10/6/19 11:39:15 PM	5 sec	COMPLETED		×	
SOM	10-706	AllSpecOndSOM(copy)(copy)(copy)(copy)	10/6/19 11:34:16 PM	8 sec	COMPLETED		×	
Active_learning	10-705	my1	10/6/19 11:30:23 PM	2 sec	COMPLETED		×	
Active_learning	10-512	active-learning-test(copy)(copy)(copy)(copy)(copy)(cc	9/21/19 7:14:48 PM	2 sec	COMPLETED		×	
Active_learning	10-362	active-learning-test(copy)(copy)(copy)(copy)(copy)(cc	5/13/19 11:23:09 AM	2 sec	COMPLETED		×	
Active_learning	10-361	active-learning-test(copy)(copy)(copy)(copy)(copy)(cc	5/6/19 1:30:53 PM	2 sec	COMPLETED		×	
SOM	10-159	AllSpecOndSOM(copy)(copy)(copy)	3/5/19 6:15:33 PM	6 sec	COMPLETED		×	
Preprocessing	10-157	vocloud2 test(copy)	2/28/19 5:54:03 PM	9 sec	ERROR		×	
SOM	10-155	AllSpecOndSOM(copy)(copy)	10/24/18 6:46:21 PM	9 sec	COMPLETED		×	

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)

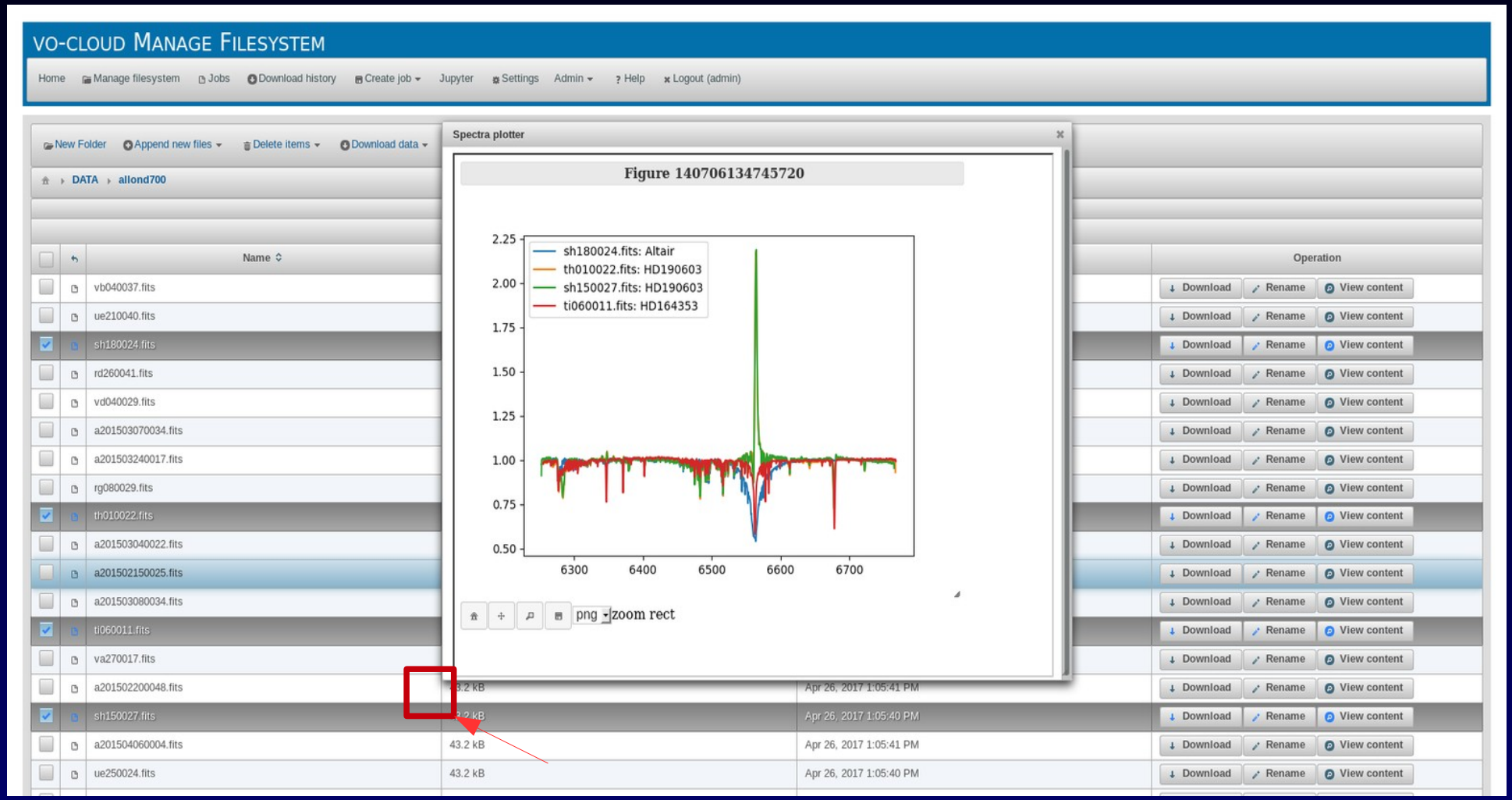
REMOTE VOTable

SSAP query + Accref

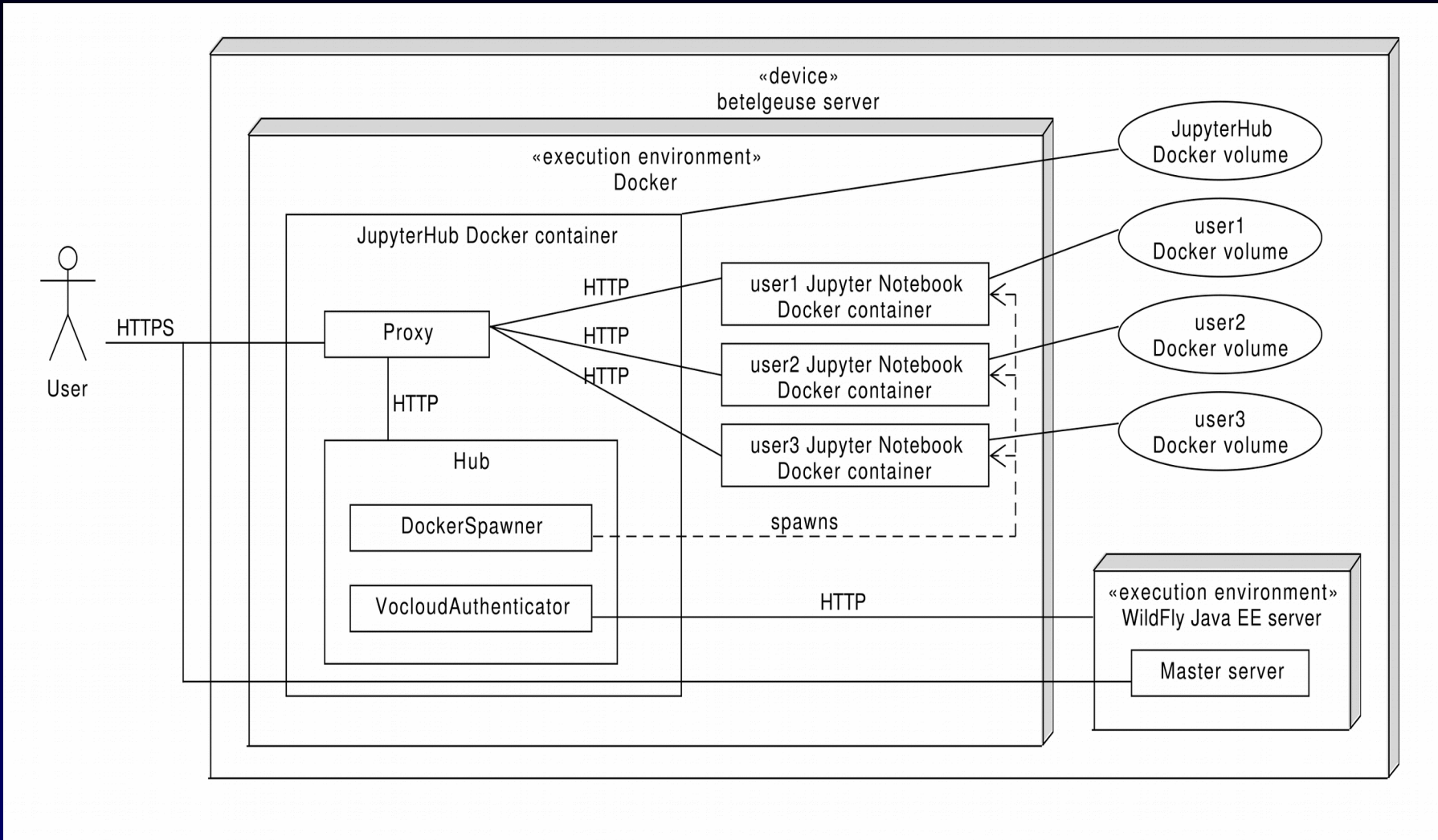
+ DataLink + SODA

~~SAMP control - send to SPLAT???~~

VO-CLOUD spectra visualisation



JupyterHub deployment



JupyterHub example

jupyter example_plotter Last Checkpoint: 05/06/2017 (unsaved changes)

Control Panel | Logout

File Edit View Insert Cell Kernel Widgets Help | Python 3

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
```

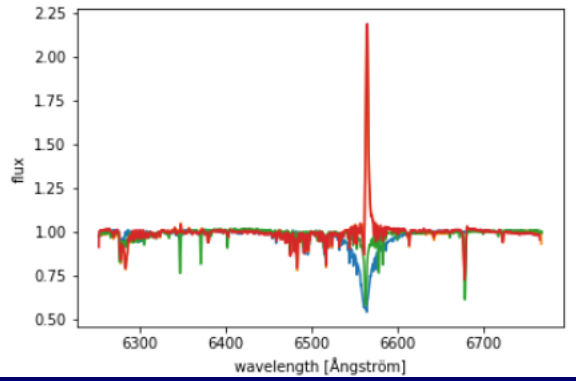
Out[7]:
```
['filesystem/DATA/allond700/sh180024.fits',
 'filesystem/DATA/allond700/th010022.fits',
 'filesystem/DATA/allond700/ti060011.fits',
 'filesystem/DATA/allond700/sh150027.fits']
```

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,  6764.7926008 ])}
```

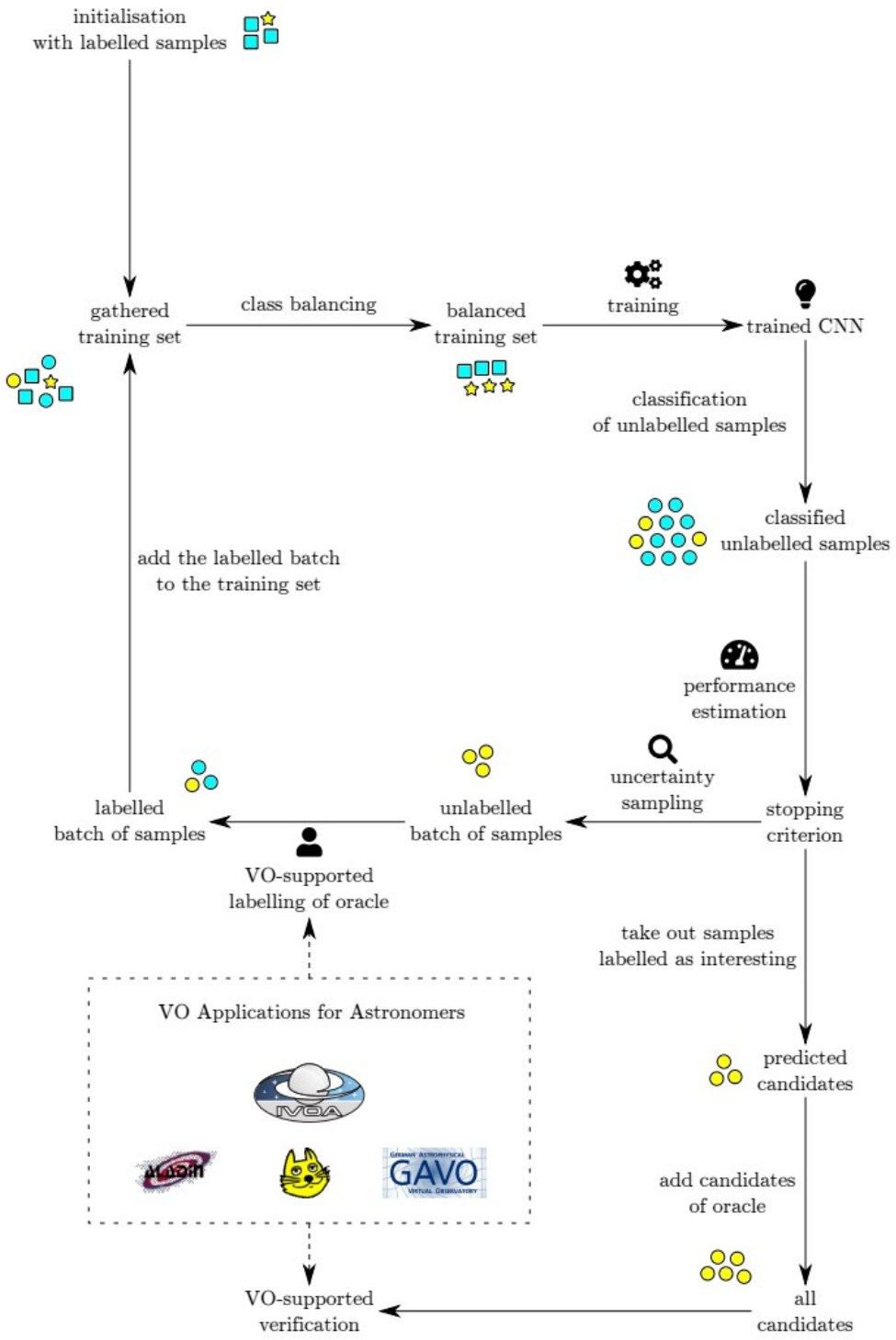
In [12]:
```
for i in parsed:
    plt.plot(i['wave'], i['flux'])
plt.xlabel('wavelength [Ångström]')
plt.ylabel('flux')
```

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



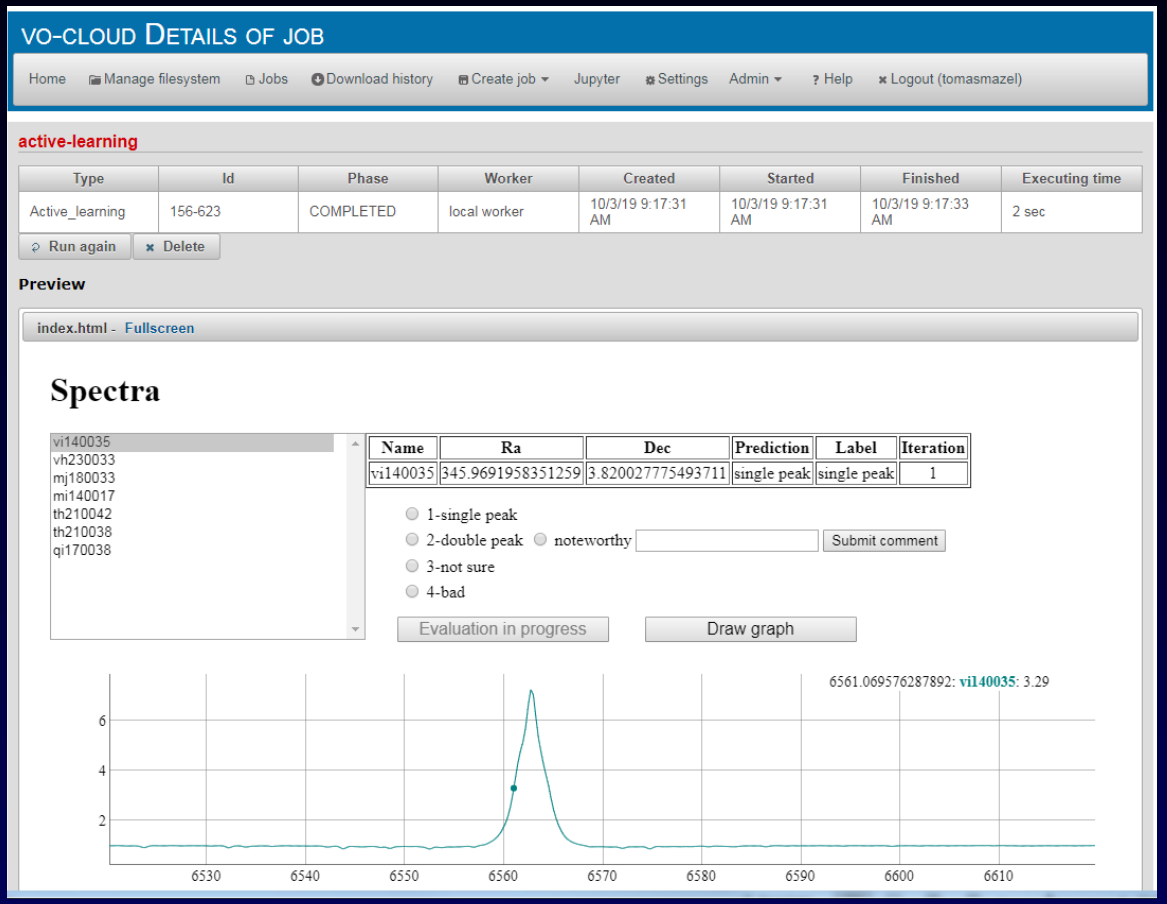
Active Deep Learning: CNN Chooses Data for Its Training

- Oracle classification (domain expert knowledge)
- Uncertainty sampling (entropy)
- From predicted target classes selected batch of 100
- Batch added to training set
- True positive rate estimation



Icons by Font Awesome are licensed under CC BY 4.0

Active Learning – in preparation



- ElasticSearch
- HDF5
- Python scripts/nb
- CUDA/TF/KERAS
- Aladin Lite ?

Thank You

RESEARCH
CENTER FOR
INFORMATICS

rci.cvut.cz

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