Distributed data mining of spectra archives using VO standards

Petr Škoda

Astronomical Institute, Czech Academy of Sciences Ondřejov

Jakub Koza, Lukáš Lopatovský, Andrej Palička, Lumír Mrkva, Tomáš Peterka

Faculty of Informatics, Czech Technical University, Prague

Supported by grant GAČR 13-08195S

IVOA Interoperability meeting , Apps session 4
Sesto/Sexten Italy, 18th June 2015

Concept of scientific „CLOUD"

ITERATIVE REPEATING of SAME computation (workflow)

Global non-linear optimization (spectra disentangling)

Synthetic spectra (various elements, wavelength-ranges)

Machine Learning (almost all methods)

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

VO-CLOUD (former VO-KOREL)

Distributed engine

MASTER (frontend)

- Database of users and their experiments
- Visualization
- Scheduling
- Load balancing

WORKERS (backend)

- Computation [+ output for visualization]

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 big astronomical spectra archives

Jakub Koza

Supervisor: RNDr. Petr Škoda, CSc.

12th May 2015

VO-CLOUD Design Details

Master controls more workers using UWS

UWS for interactive work (abort returns data, users isolation)

Visualization – now workers (xhtml) but wil be centralized
(driven by some recipe particular to the worker type)

Worker type – automatic registration describe capability
Currently Preprocessing, SOM, RDF (?DEEP-LEARN)
JSON, XSD

Machine Learning of Spectra
SW view

ML does not produce new data – same spectra in groups

Results the same size as input (+ small overhead)

RDF – supervised – need classes (by eye)

Solf-Organizing maps – finding outliers

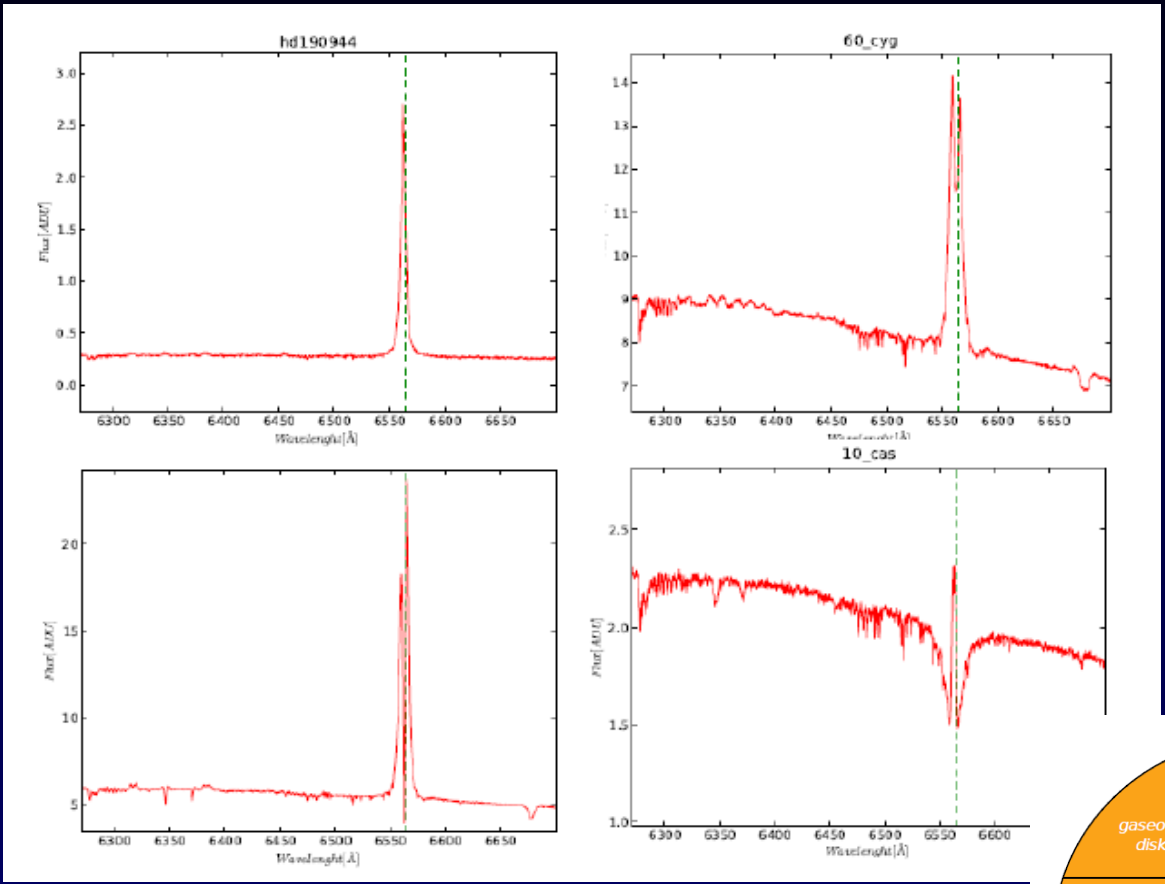
Easy trace shape from results - clickable maps

Visualisation of many spectra in web

- after rebinnig (+ Dim Reduced – PCA...)
- data obtained (normalized, cutout of sp. lines)
- original spectra (whole size, just extracted...)

Machine Learning of Spectra

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

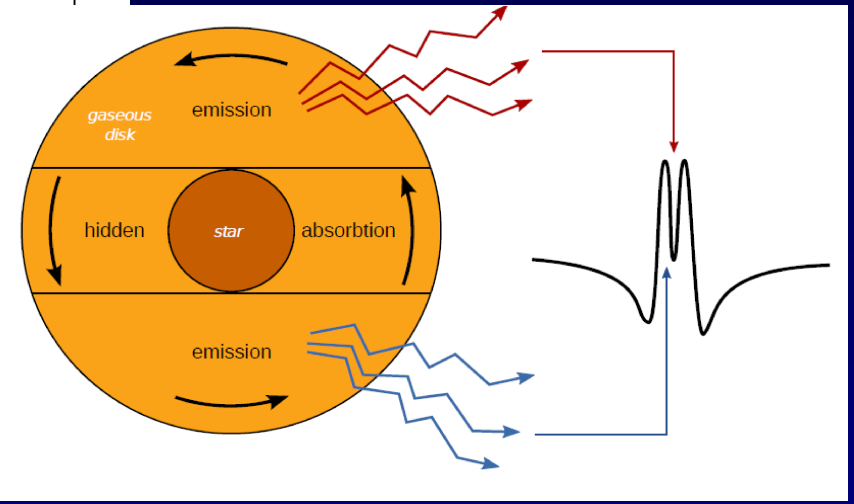


Be stars

Disk or envelope

Rotates, Hot

Origin ?????



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 (PUBDID + mime)

SAMP control - send to SPLAT

Machine Learning of Spectra Science case

Ondřejov 2m Perek Telescope – 1700/10 000 spectra

PRE-PROCESSING

Normalization to continuum, Cutout (SSAP+DL)

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

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

Produces feature vectors in CSV (same length, dimensions)

MACHINE-LEARNING

Unified wrapper running multiple applications - same call

Name-of-wrapper + parameters (json) – method as param

VISUALIZATION

JavaScript (dygraph, HighCharts)

Czech Technical University in Prague
Faculty of Information Technology
Department of Computer Science

Bachelor's thesis

Application of Random Decision Forests in Astroinformatics

Andrej Palička

Supervisor: RNDr. Petr Škoda, CSc.

12th May 2014

Czech Technical University in Prague
Faculty of Information Technology
Department of Theoretical Computer Science

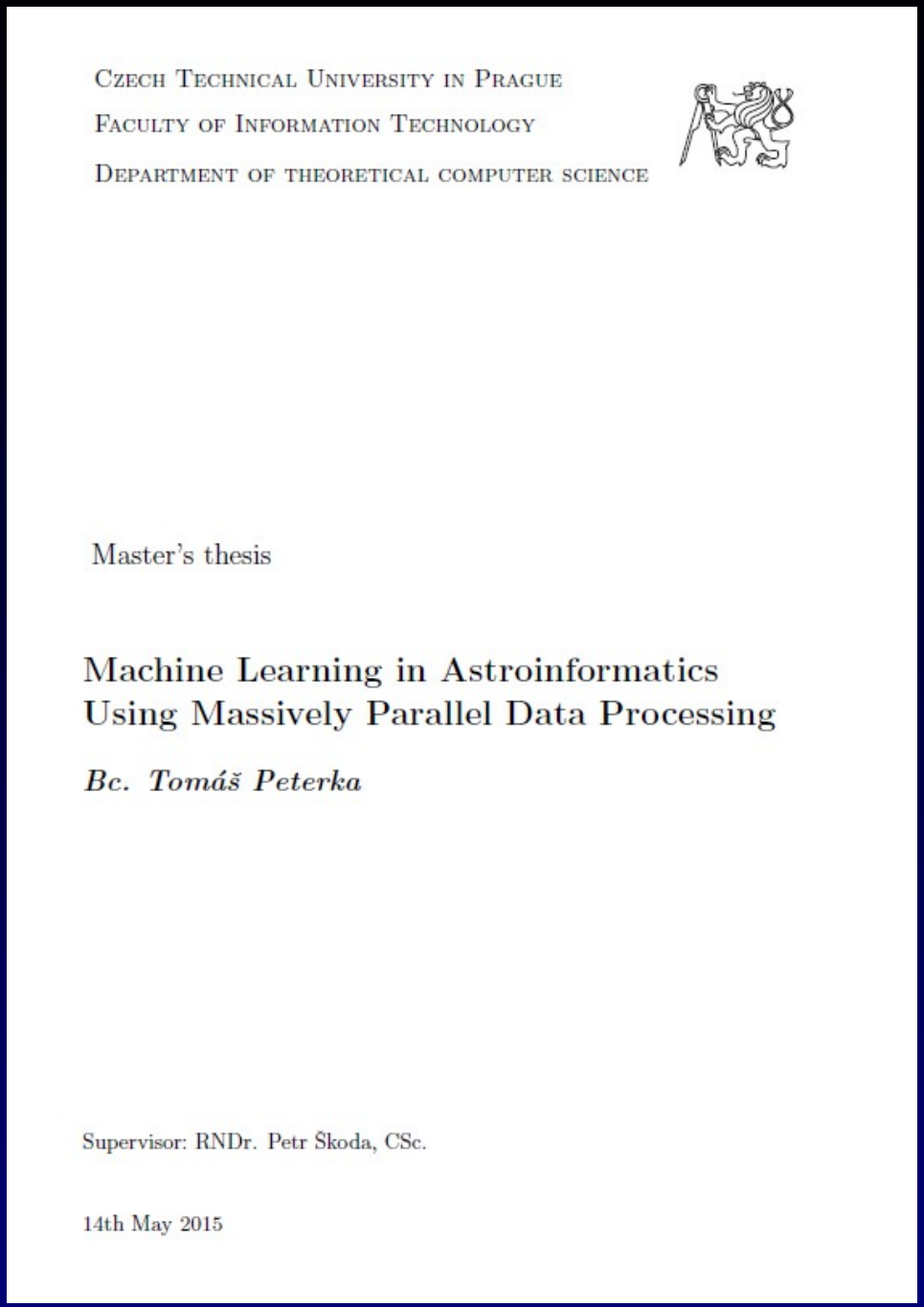
Bachelor's thesis

Application of Self-Organizing Maps in Astroinformatics

Lopatovský Lukáš

Supervisor: RNDr. Petr Škoda, Csc.

14th May 2014



Czech Technical University in Prague
Faculty of Information Technology
Department of theoretical computer science

Master's thesis

Machine Learning in Astroinformatics Using Massively Parallel Data Processing

Bc. Tomáš Peterka

Supervisor: RNDr. Petr Škoda, CSc.

14th May 2015

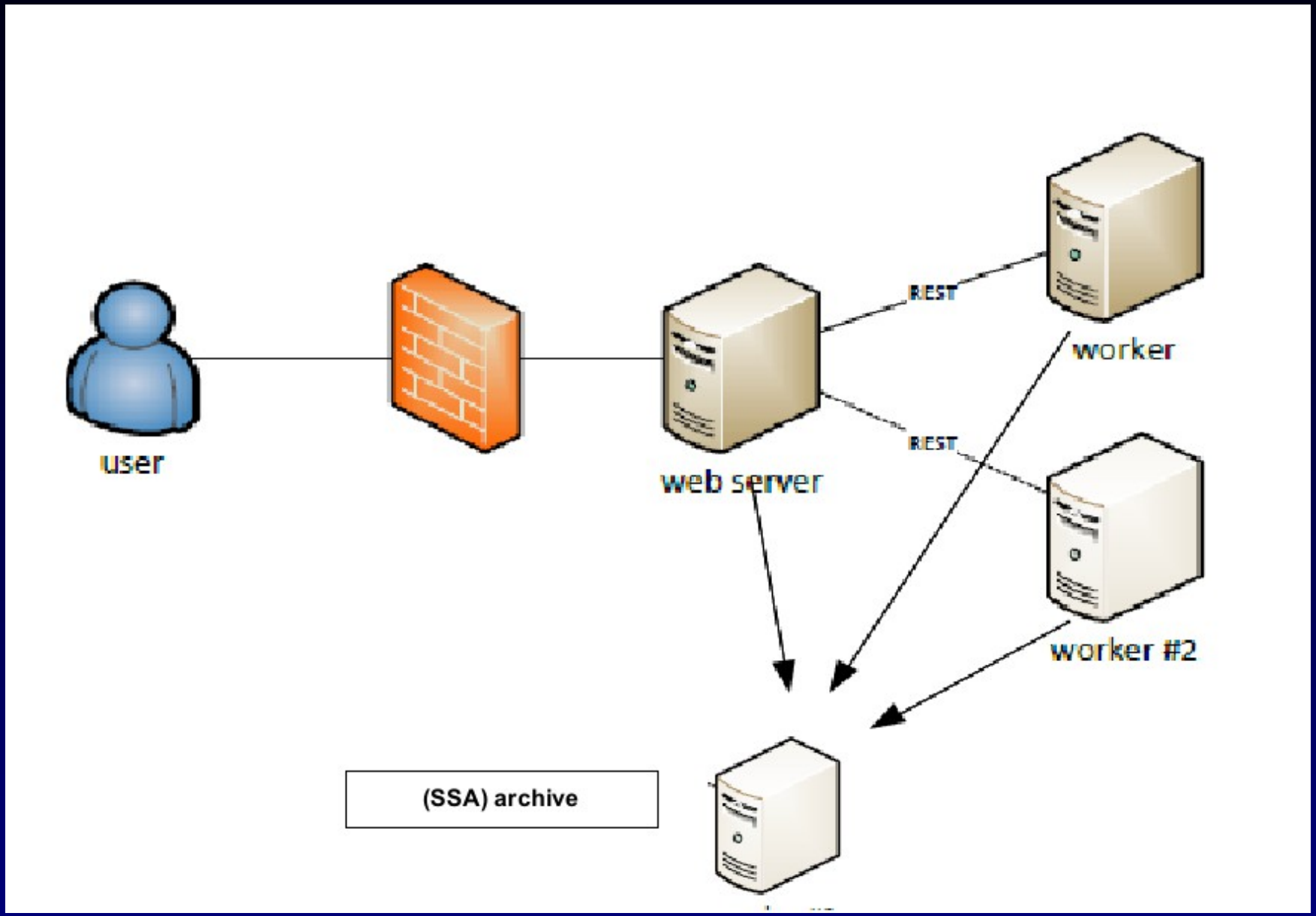
Deep Learning

Caffe + Big Data Layer

GPU /CPU switch

Will be part of VO-CLOUD soon

Machine Learning of BIG Archive



Machine Learning of BIG Archive?

Idea – 2.2 mil of LAMOST spectra (3.3 mil. in SDSS)

NOT Upload data by user (VO compatible archive)

Driven by SPECTRA LIST (votable obtained by TAP ?)

Workers on same hi-speed network as archive

Calling SSAP + DL always (client on GRID worker ?)

Pre-cache ?

Compute feature vectors – store for whole experiment ?

PERSISTENT STORAGE - network FS ?

Visualisation - needs input data (spectrum), lists from class

Source Code

https://github.com/vodev/vocloud

https://github.com/vodev/vocloud-preprocessing

https://github.com/vodev/vocloud-som

https://github.com/vodev/vocloud-RDF

https://github.com/vodev/vocloud-deeplearning

DEMO

http://vocloud-dev.asu.cas.cz/vocloud2