



**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



# The Role of VO Technology in Astronomical Machine Learning

**P. Škoda<sup>1,2</sup>**

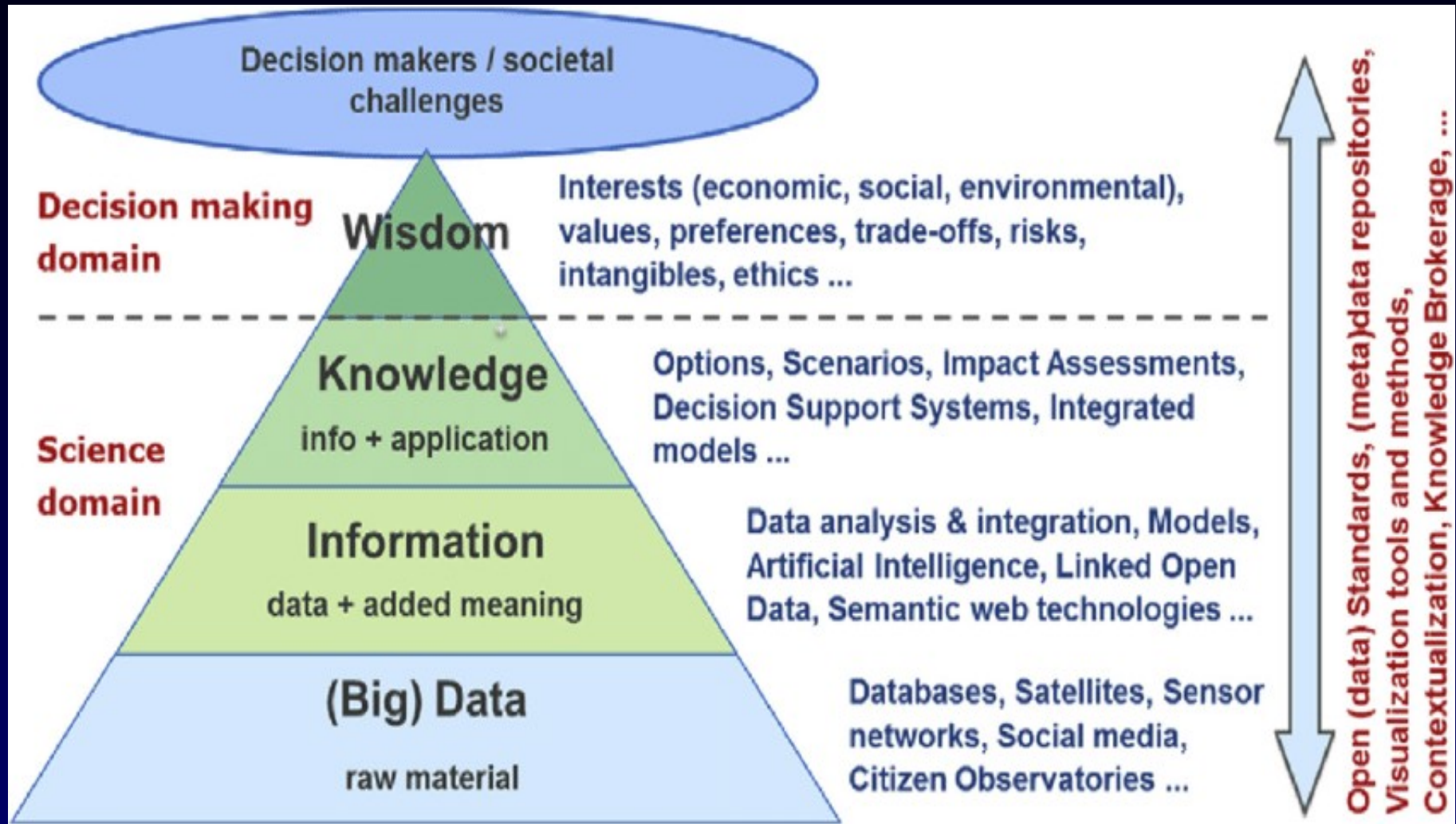
<sup>1</sup>Astronomical Institute of the Czech Academy of Sciences, Ondřejov

<sup>2</sup>Faculty of Information Technology, Czech Technical University in Prague

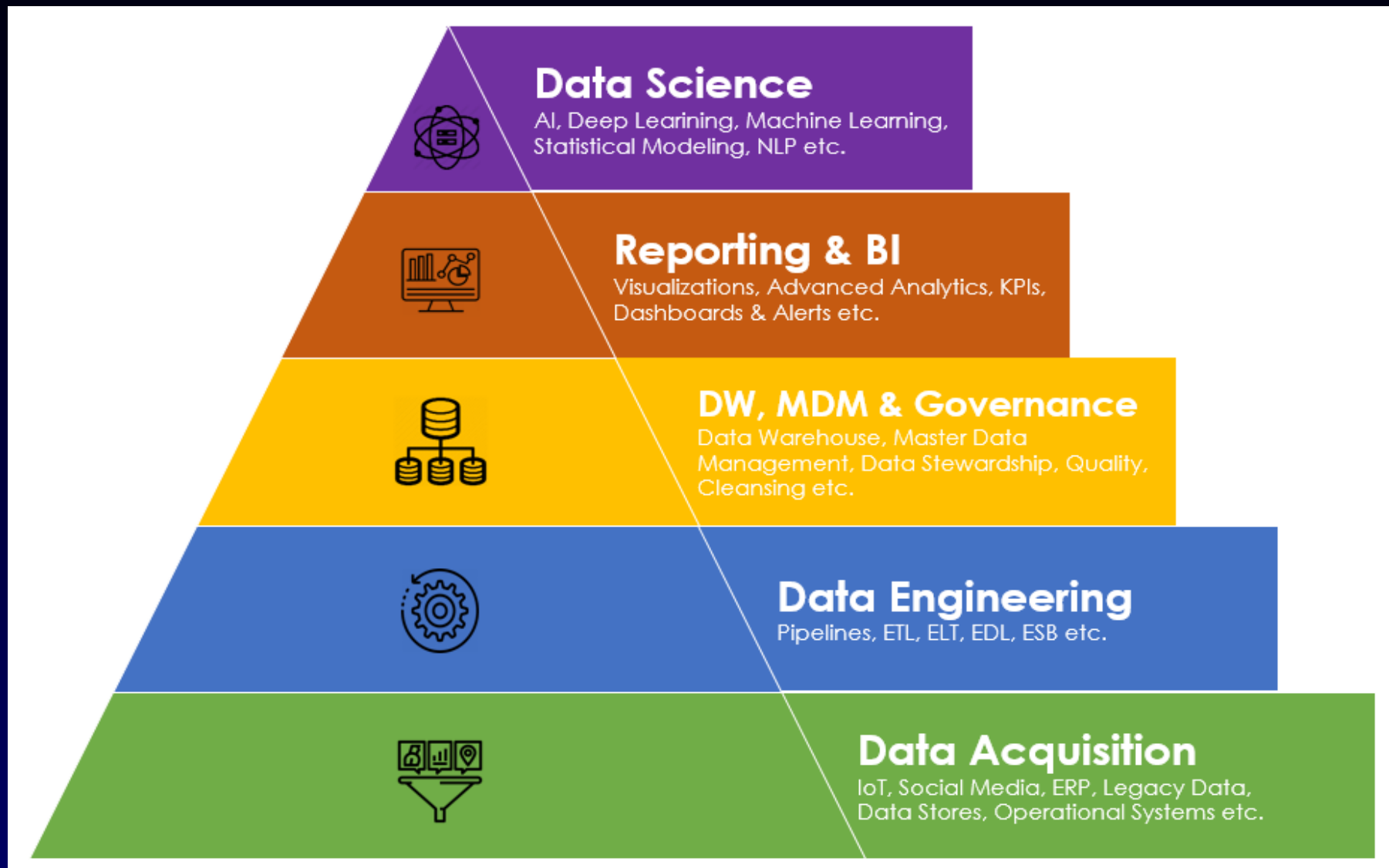
# KDD IG Charter (2009)

- Participating in the definition of new data preservation and exchange **formats** with respect to support **machine learning** algorithms.
- Introducing **uncertainties** and probabilistic description to **VO standards** and services.
- Presenting and collecting best practice examples of scientific data analytics in astronomy.
- Defining requirements for implementing and **adding machine learning capabilities to services**.
- **Coordinating** and unifying the access to **data visualization functionalities**.
- Discussing the aspect of **data provenance** with respect to data used to derive/train **models**.
- Introducing proper **statistical** scoring and evaluation **methods as services**.
- Contributing to the discussion on scripting and **orchestrating** the scientific discovery **workflow**.
- Supporting the development of **dedicated knowledge discovery applications**.

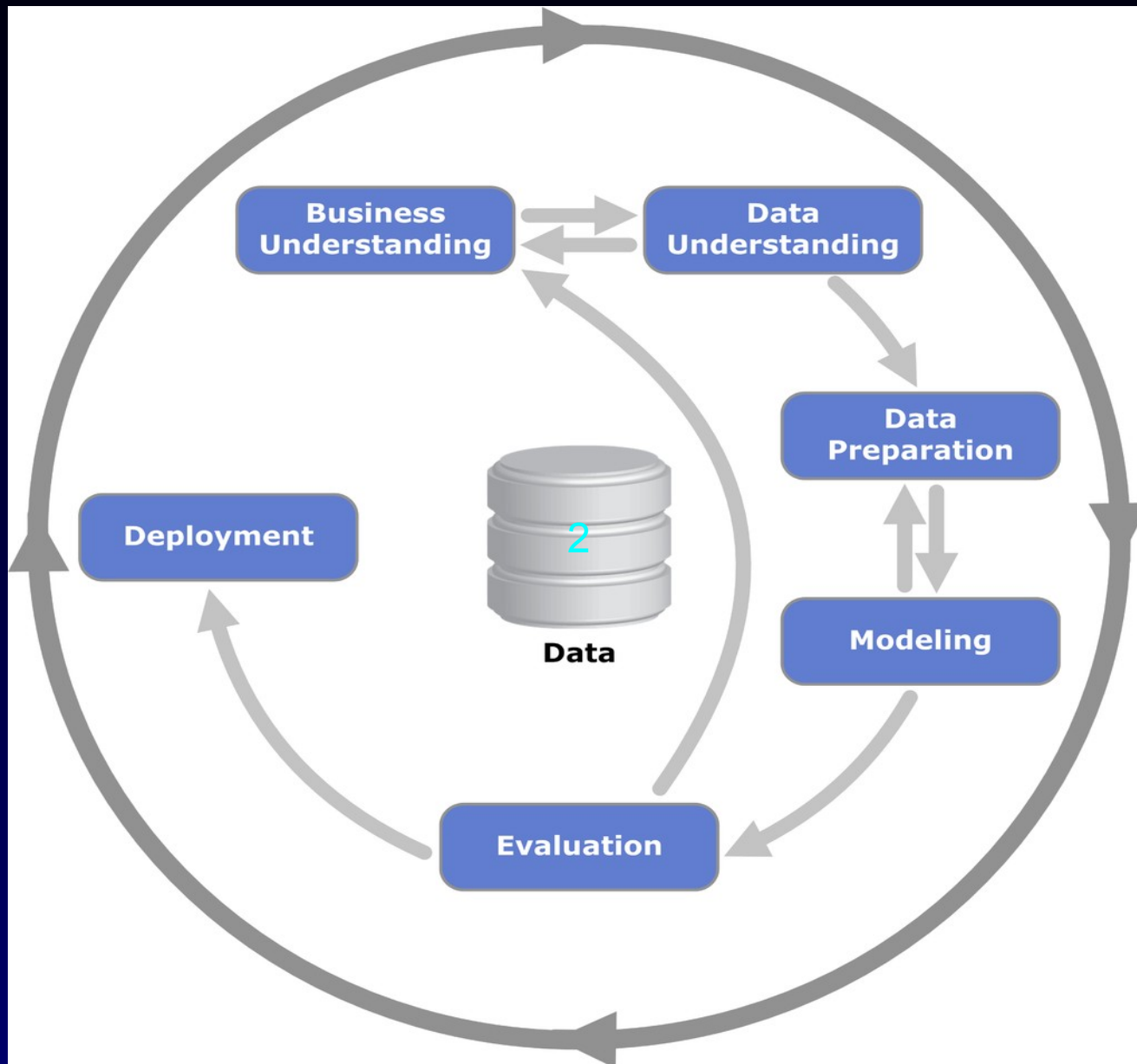
# DIKW Pyramid - Science



# DIKW Pyramid Data Science



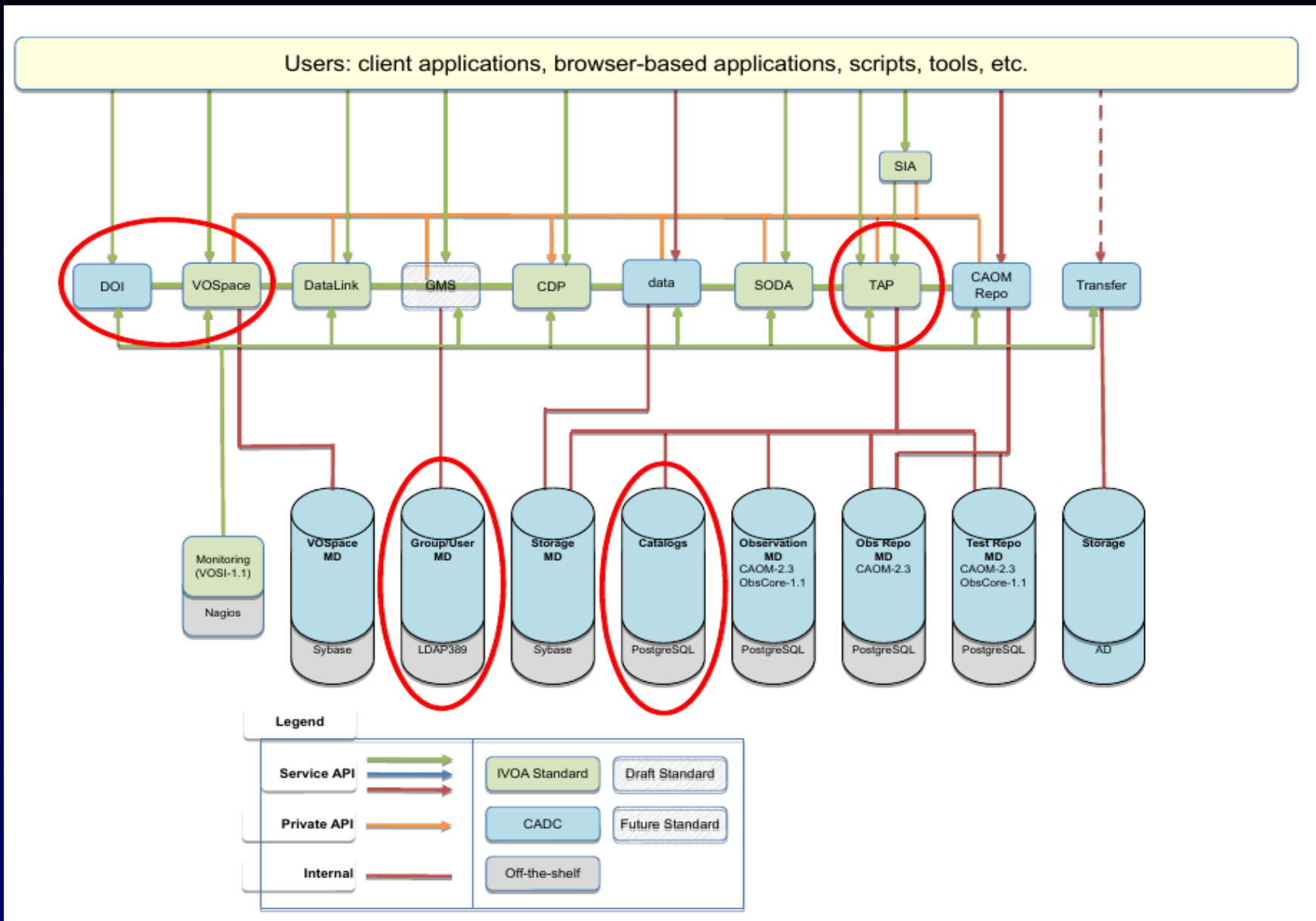
# CRISP-DM




Cross-Industry  
Standard  
Processes  
for Data Mining

Kind of ISO  
standard

# Science Platforms - CANFAR



# Science Platforms - SciServer

SciServer  Home Files Groups Science Domains user

## SciServer Dashboard

Data, Collaboration, Compute




### Your Activities

- Files**  
You have 0 Shared User Volumes.  
You have 2 Owned User Volumes.
- Groups**  
You have 0 Group Invitations.  
You have 0 Owned Groups.
- Compute Jobs**  
You have 0 Jobs Running.  
You have 0 Jobs Completed in 24 hours.
- Science Domains**  
You have joined 2 Science Domains.

### SciServer Apps

- CasJobs**  
Search online big relational databases collections, store the results online, and share them.
- Compute**  
Analyze data with Interactive Jupyter notebooks in Python, R and MATLAB.
- SkyServer**  
Access the Sloan Digital Sky Survey data, tutorials and educational materials.
- SkyQuery**  
A scalable database system for cross-matching astronomical source catalogs.

SciServer - 2.1.0 Dashboard - 2.1.2-134-g2cdfbf4

Powered by:   

# Simple KD with VO

## KD Schema

provide the necessary tools to access the data and transfer only the required parts

consider the usage of optimized hardware

provide the data for processing & schedule, optimize and distribute the workload

transfer / post-process data / collect, generate provenance

provide analysis capabilities

select training objects

extract training data

train model

use model

save results

analyze results

find/use/merge data  
annotate objects  
statistical analysis

SIA / SSA / HIPS ...

CPU / GPU  
IO / storage  
special purpose HW

data-access,  
compute, GWS

storage, VO-space,  
provenance, DM

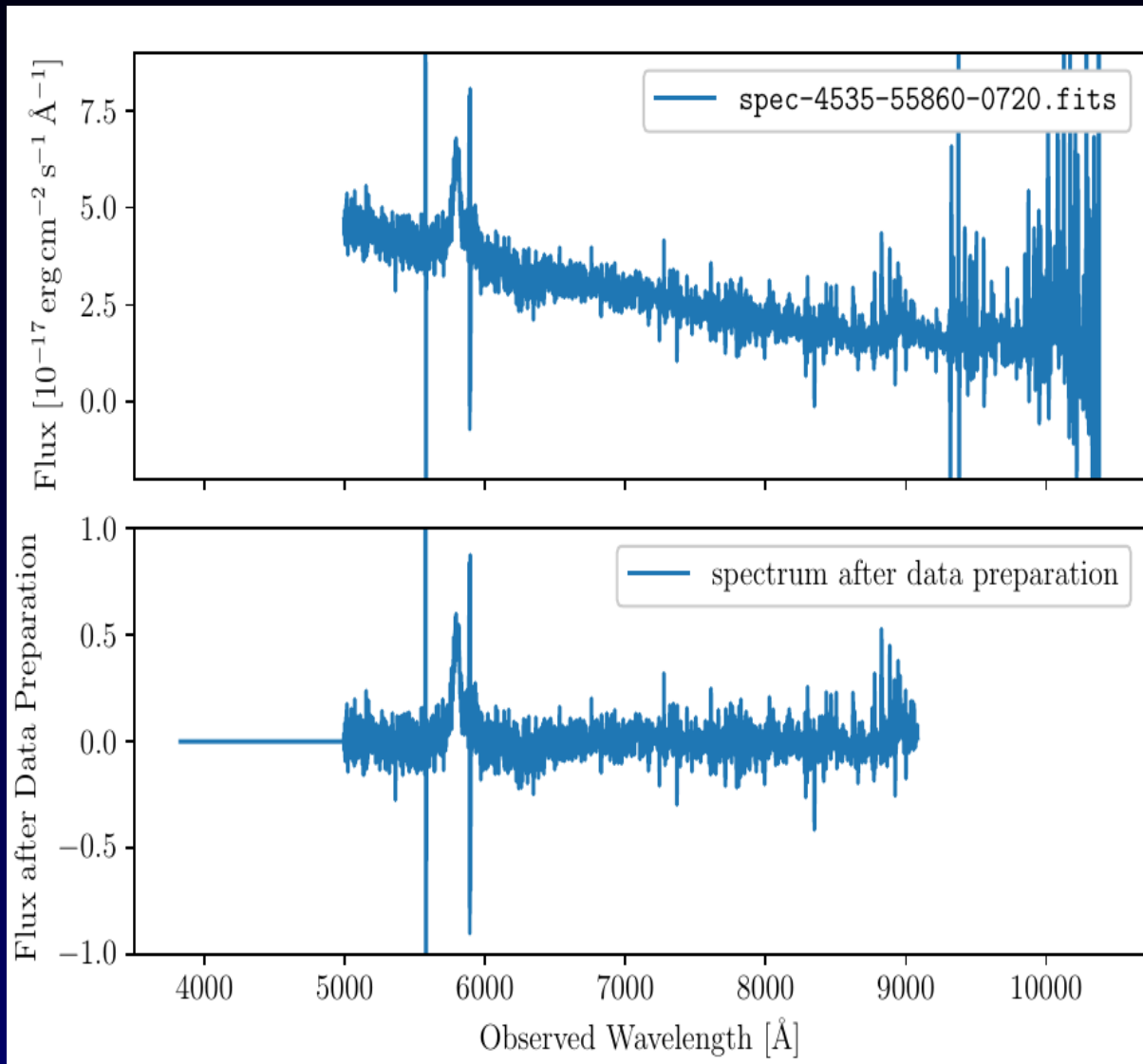
remote visualization  
statistical tools



# ML supported by VO

- Getting training (testing, validation) sets
  - Using Registry to identify
  - Query with filtering (good SNR, category, z range)
    - Using metadata
    - Previews
    - Cutouts
  - Transformation on the fly - feature vectors - SODA
  - Creating MATRIX of FVs (spectrum each row)
  - Download HDF5 (UWS)
- Applying ML model (VOspace, UWS, HDF5 ) > ActiveLearning
- Analysing results (Aladin,Splat,Cassis,Simbad)      ^^^^
- Visualization (Topcat, dedicated tools, web)

# Input Data Preprocessing (Spectra)



Pseudocontinuum normalization

Cutout to same range

Standardization (Z-score)

Zero-padding

Rebinning to the same grid

-----  
Vacuum  $\leftrightarrow$  air

Log wavelength  $\leftrightarrow$  wavelength

# Rescaling (Normalization)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min Max normalization  
Range [0,1]

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

Min max normalization  
Range [a,b]

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

Mean normalization

# Standardization

$$x' = \frac{x - \bar{x}}{\sigma}$$

Z-score normalization

Zero mean unit variance

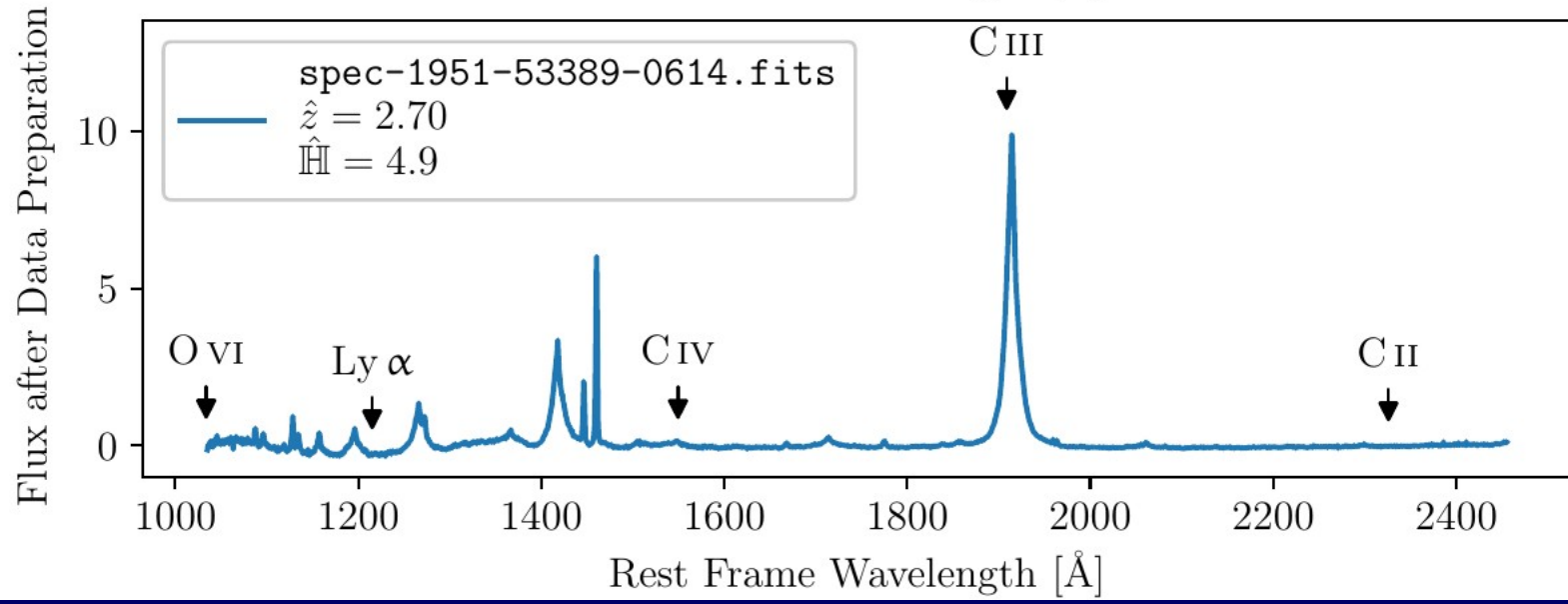
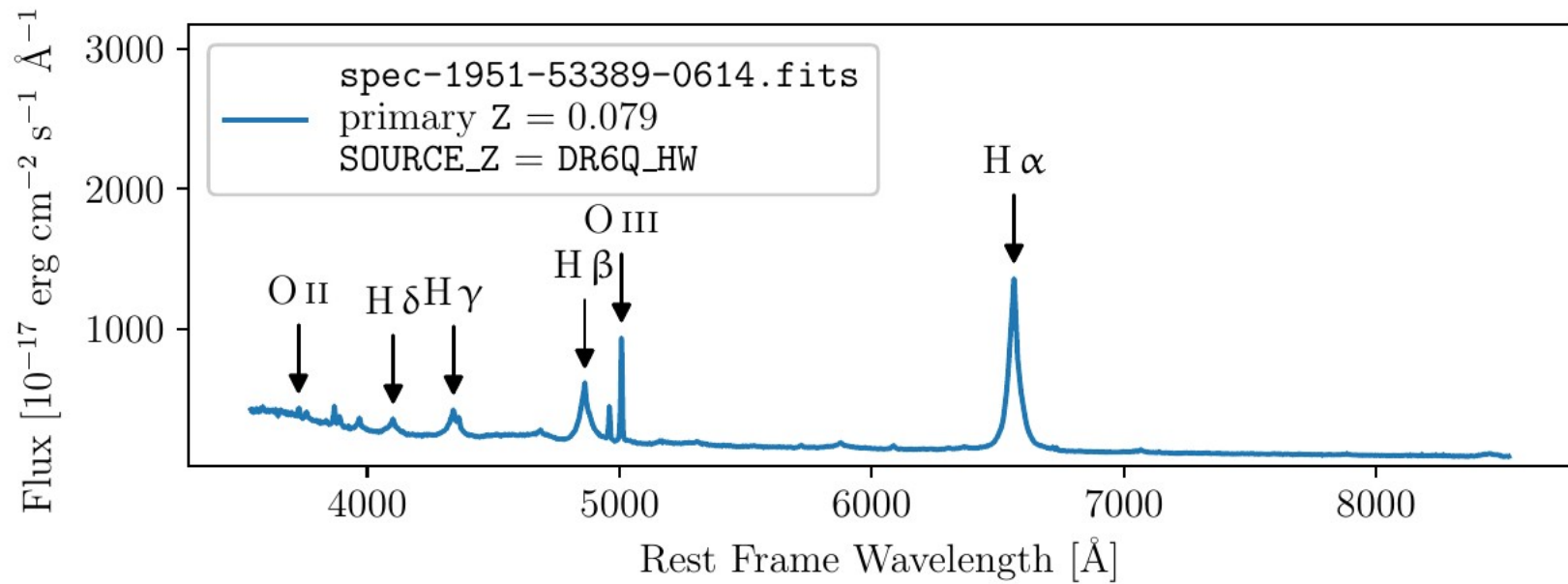
$$x' = \frac{x}{\|x\|}$$

$\|x\|$  Is the Euclidean length  
of the Feature Vector.

Scaling to unit length

Wavelength dependency

# Transformation Using the Redshift



# Advanced SODA, UWS

Multi Cut-outs Molinaro (DAL1)

## ViaLactea Visual Analytics

Directly consumes all VLKB services

Vitello & al. 2018



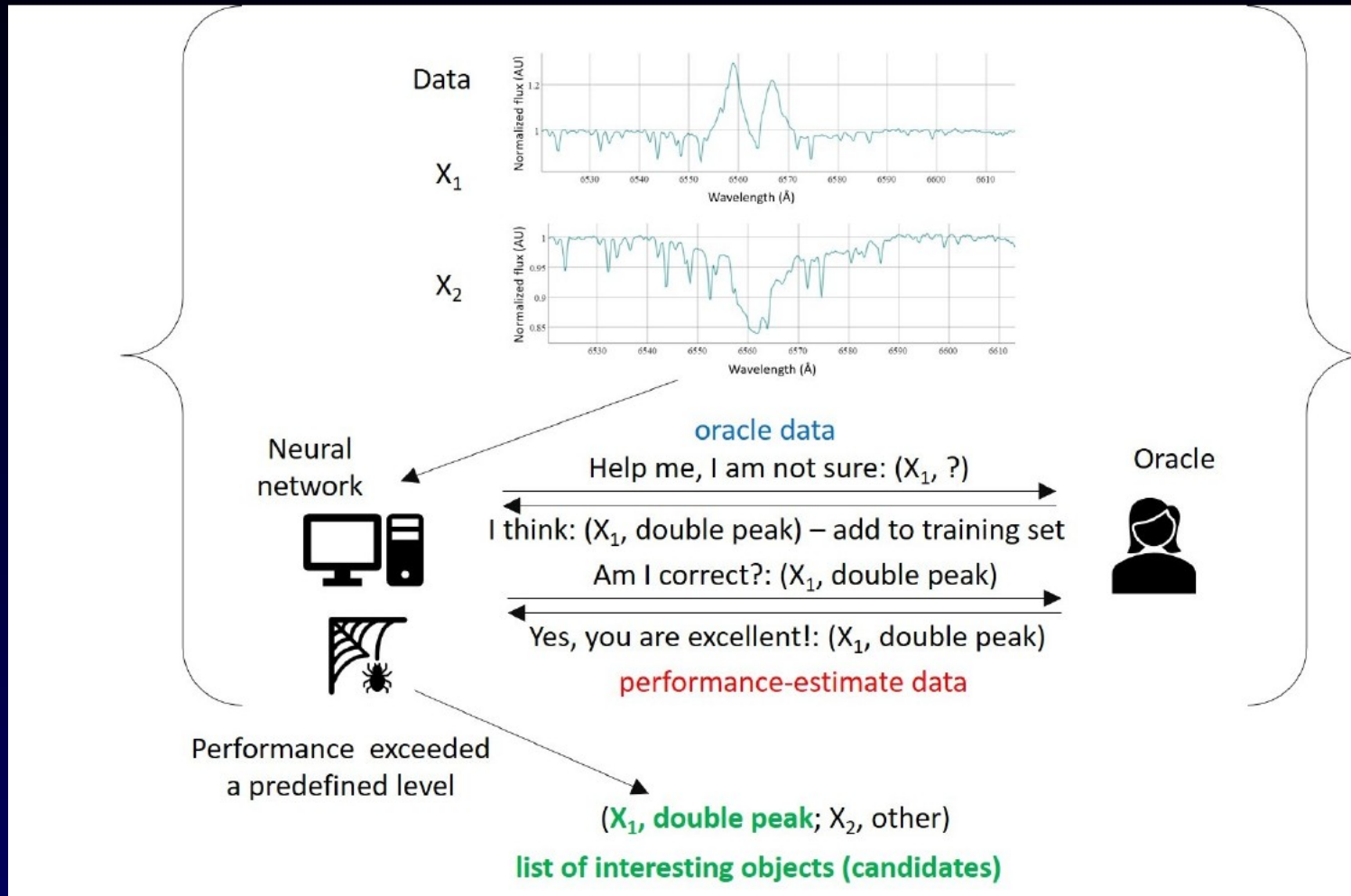
IVOA April 2022 Interoperability Meeting  
DAL – Tuesday 26 April 2022



VLKB VO Plans to Multi-Cutout  
M. Molinaro / R. Butora [INAF]

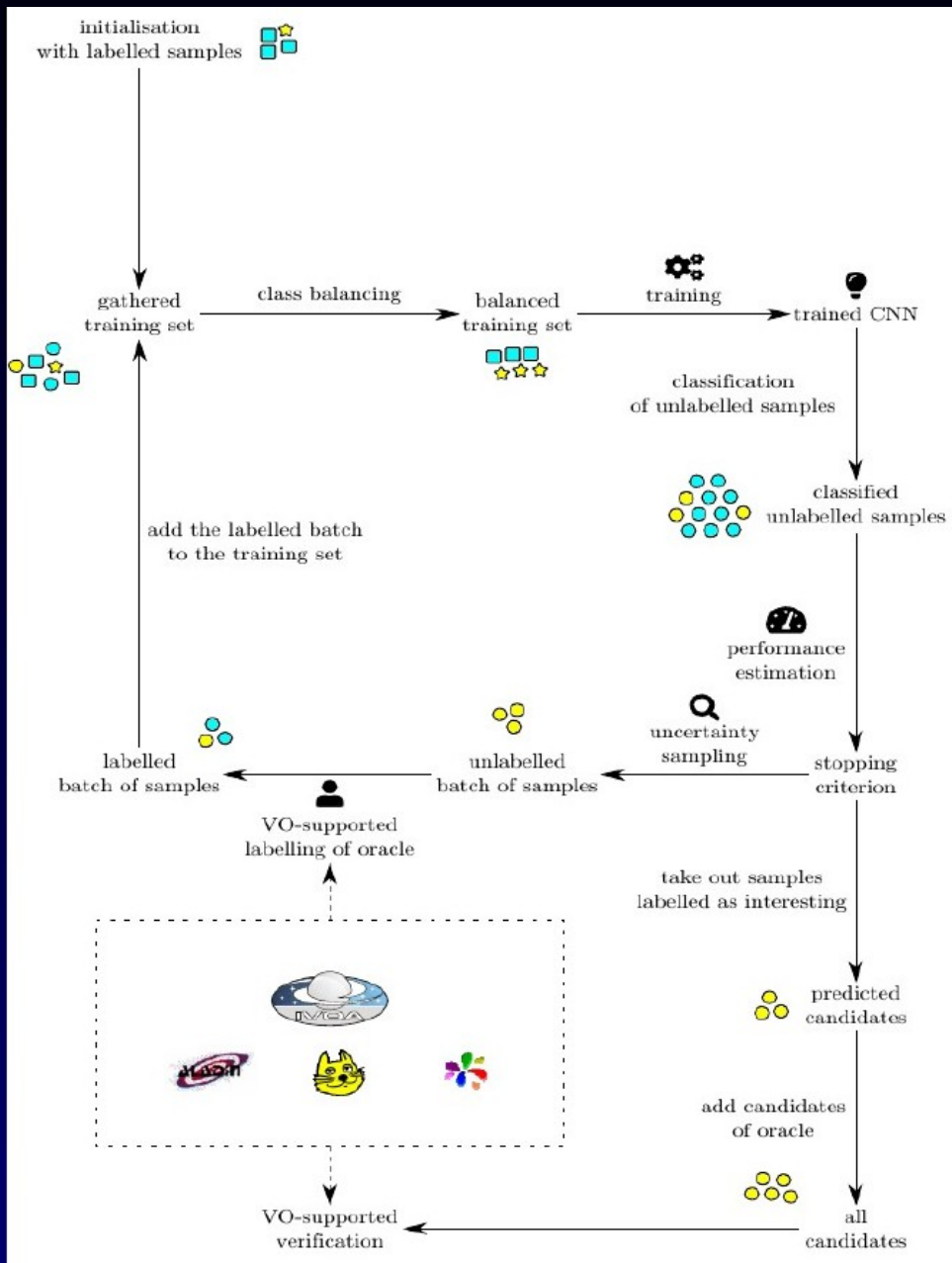


# Active Learning (insufficient labels)



Oracle : Human – Machine Interaction

# Active Learning



The Oracle (human annotator) decides :

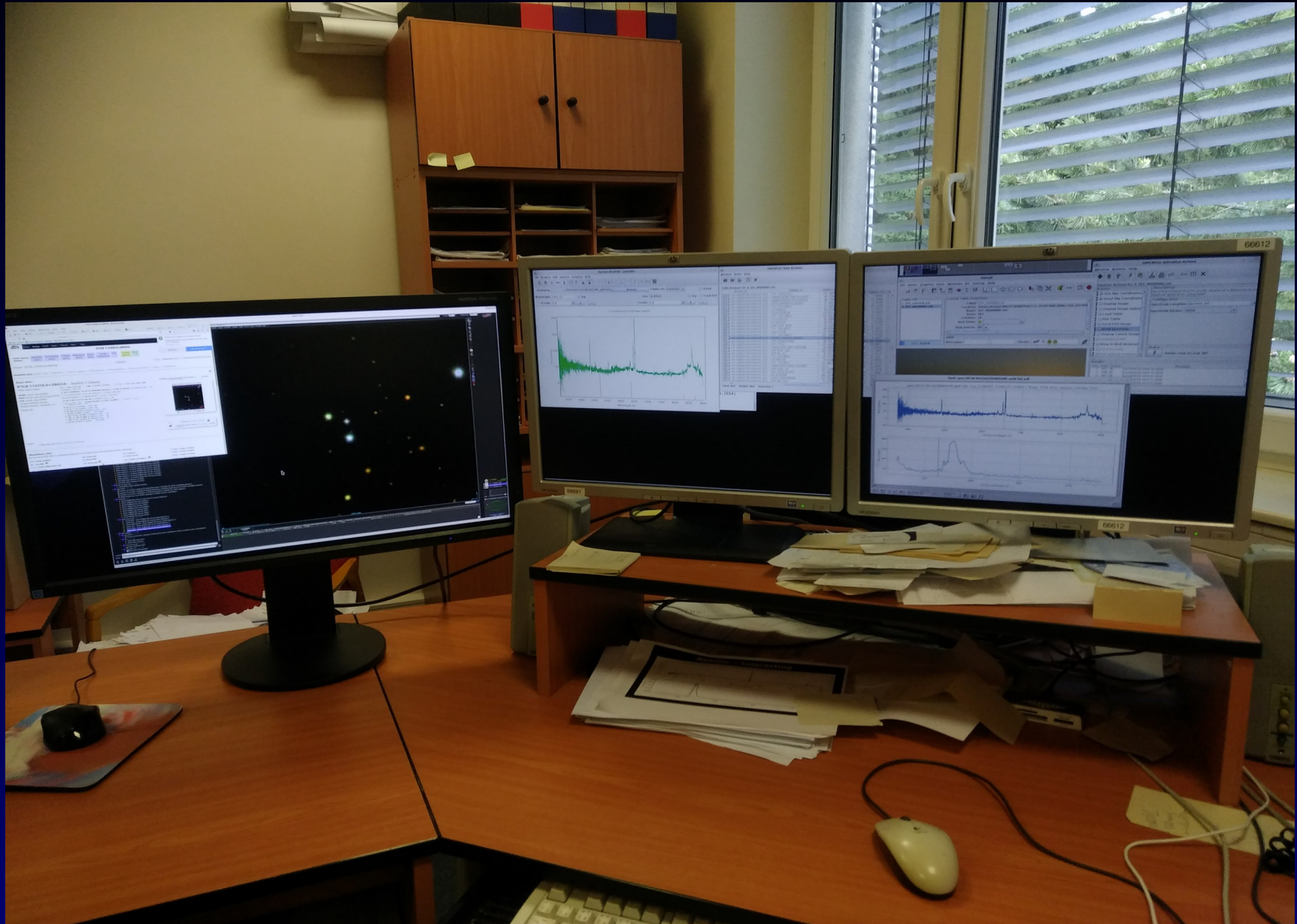
What sees the ML model

What is currently known

Metadata  
Images  
Spectra  
Catalogues  
Literature



# VO-Complex Workflows (SAMP)



# Uncertainties

*“Lack of knowledge about the truth”*

## Aleatoric :

- Due to the random nature of getting data (noise in measurements]
- Cannot be reduced by better understanding

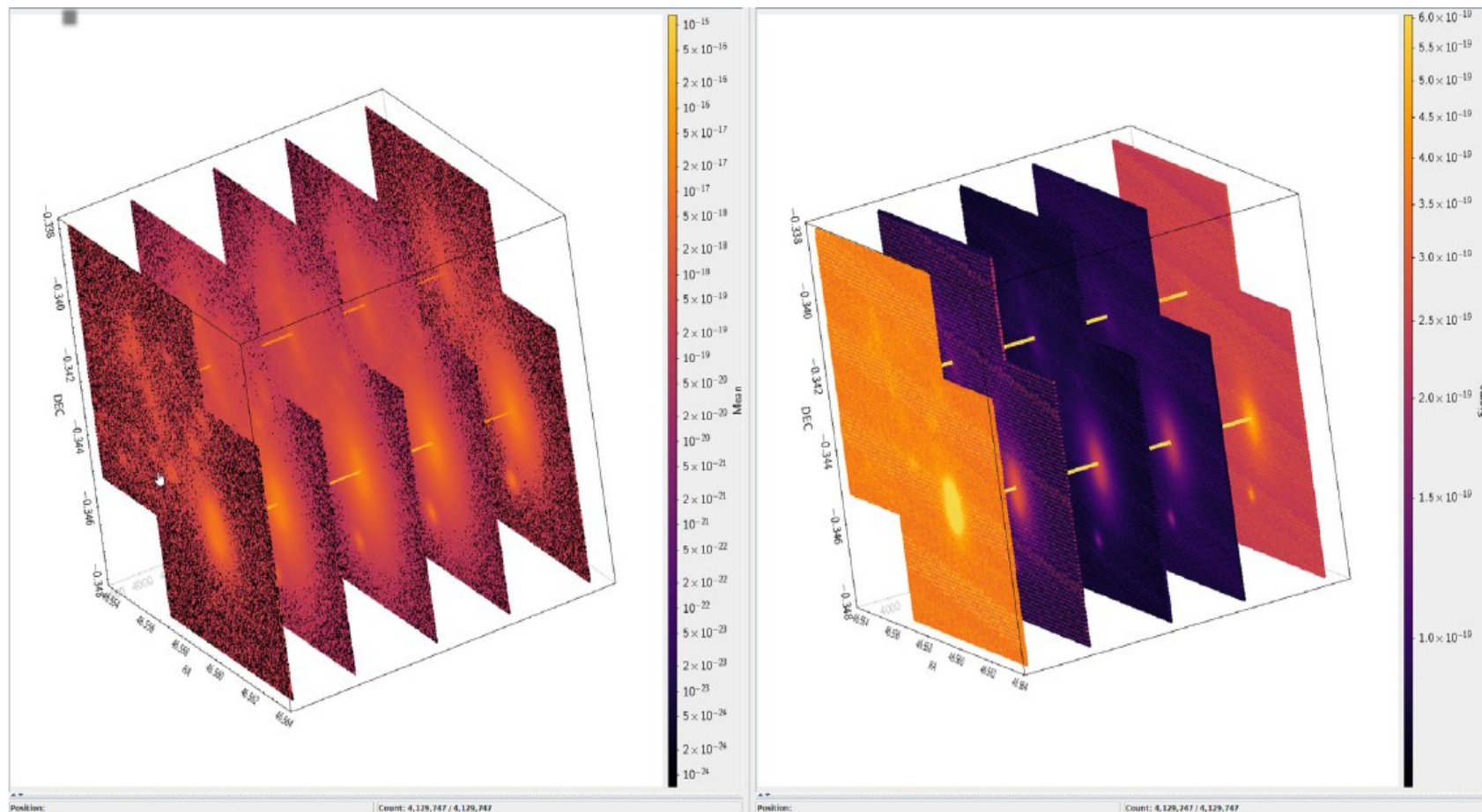
## Epistemic :

- Ignorance about the model that generated the data
- We can improve our knowledge by more experiments (e.g. different network architecture)
- Bayesian deep learning

# Uncertainties : HiSS-Cubes

J. Nádvořník, P. Škoda and P. Tvrđík

Astronomy and Computing 36 (2021) 100463



**Fig. 7.** Screenshot of Galaxy2 data set exported to VOTable and visualized in TOPCAT. The left-hand figure depicts the mean values and the right-hand figure presents the sigma values.

HDF5 structure named Hierarchical Semi-Sparse Cubes (images+spectra)

# Issues for Consideration

- Robustness (Big data)
- Consider the ML usage while proposing standards
- Collecting data ON REQUEST x Download ALL ?
- Hierchical Access (Kai Polsterer)
- Sampling data (parameter driven distributions – TOP ?)
- Quick search in metadata (select by z...)
- Datalink between various representations (untransformed, preprocessed)
- Creation of ML matrix (UWS)

Thank you !