The Role of VO Technology in Astronomical Machine Learning

P. Škoda[1,2]

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

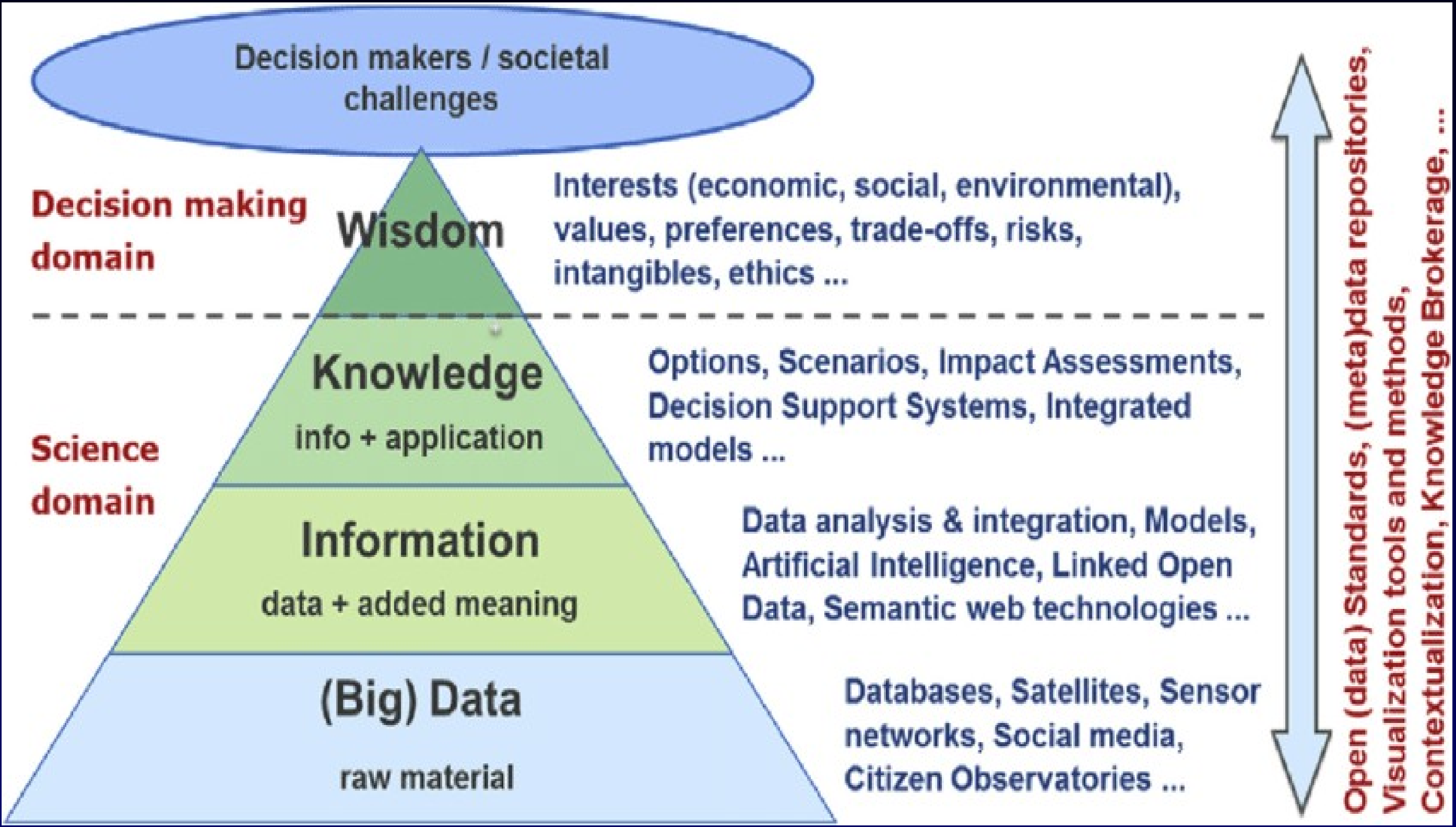
IVOA Interoperability Meeting, KDD session,
Worldwide , 28th April 2022

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.

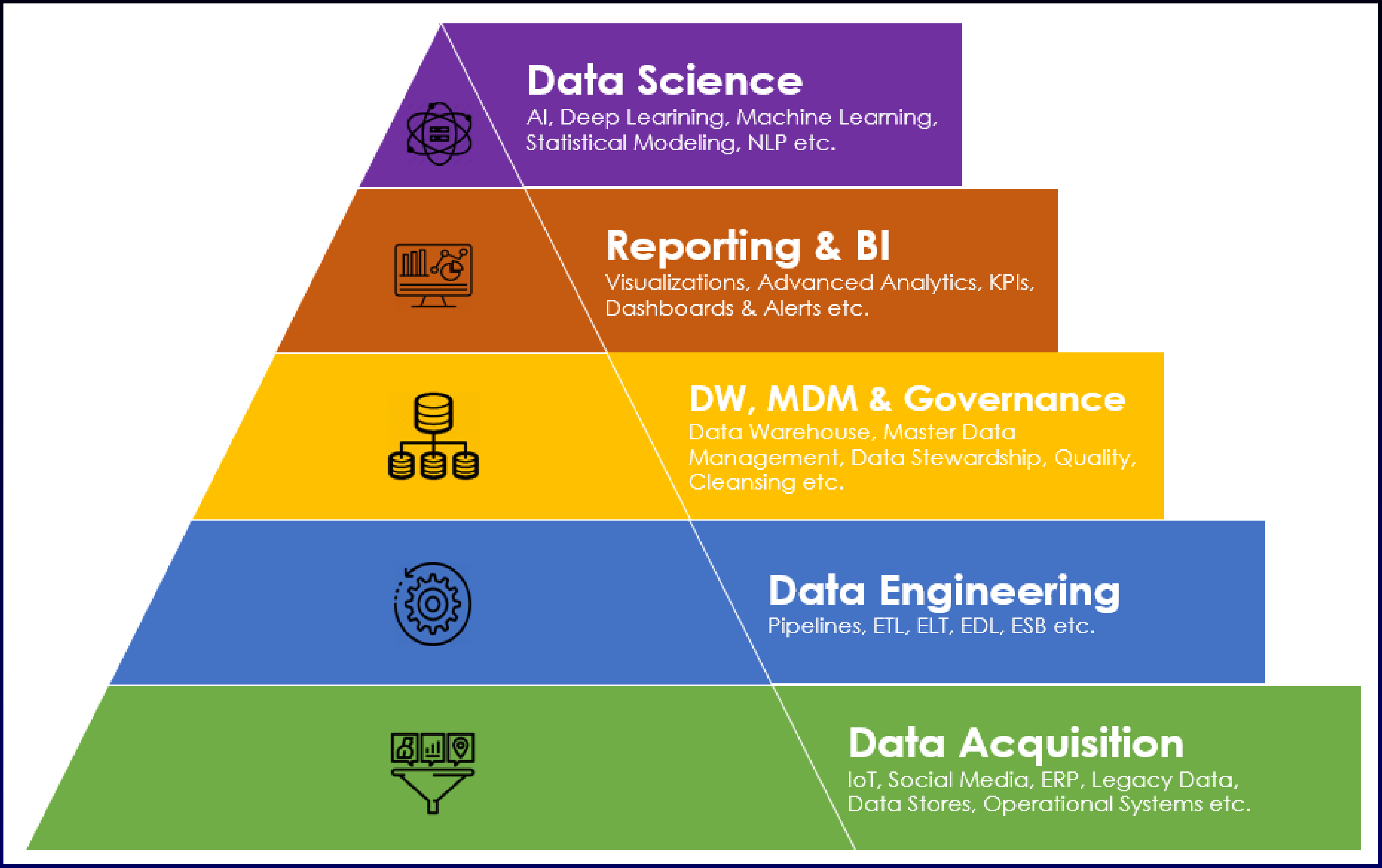
https://wiki.ivoa.net/twiki/bin/view/IVOA/IvoaKDD

DIKW Pyramid - Science



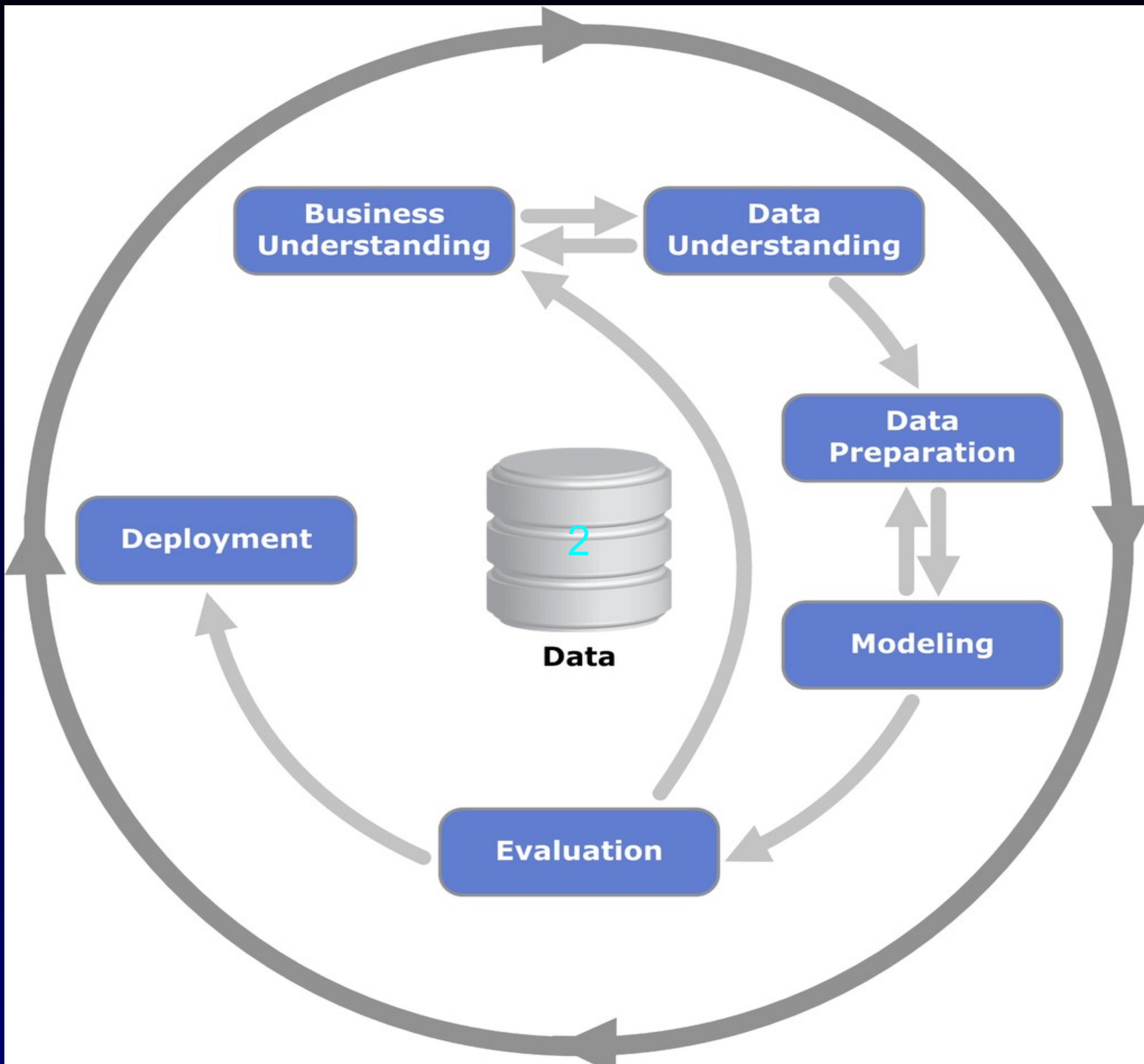
Lockers et al. 2016

DIKW Pyramid Data Science



Hussain 2022 , Medium

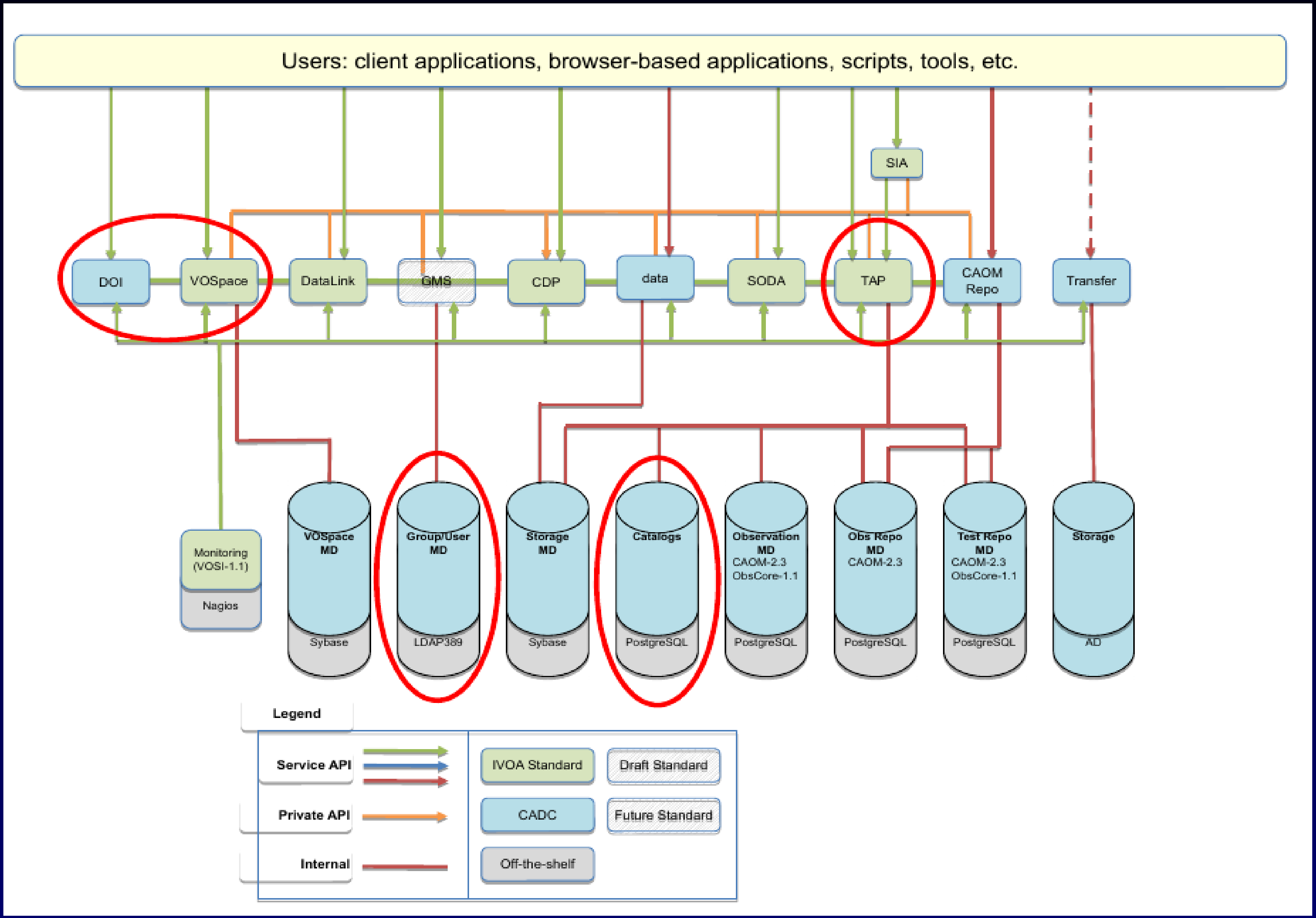
CRISP-DM



Cross-Industry Standard Processes for Data Mining

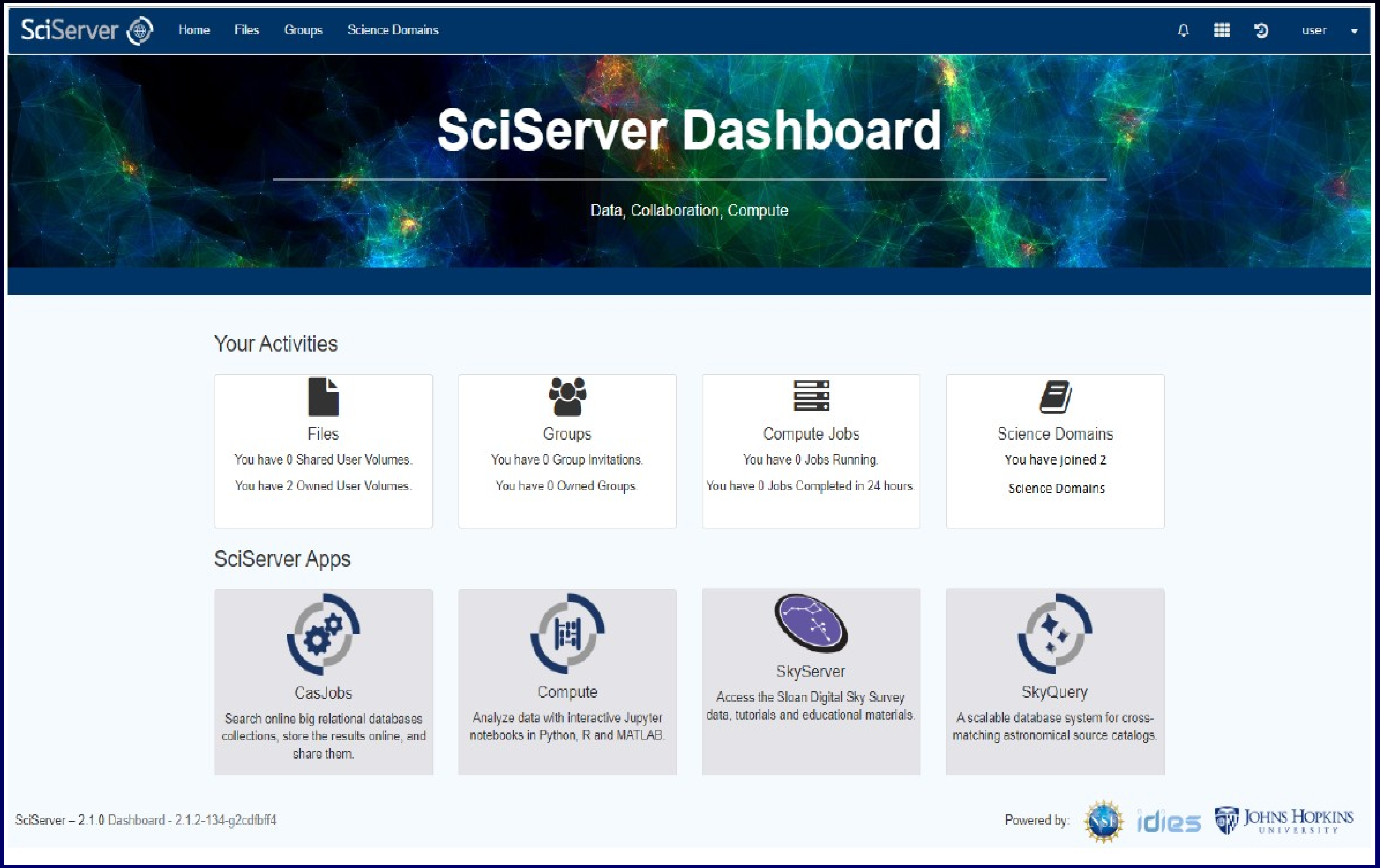
Kind of ISO standard

Science Platforms - CANFAR

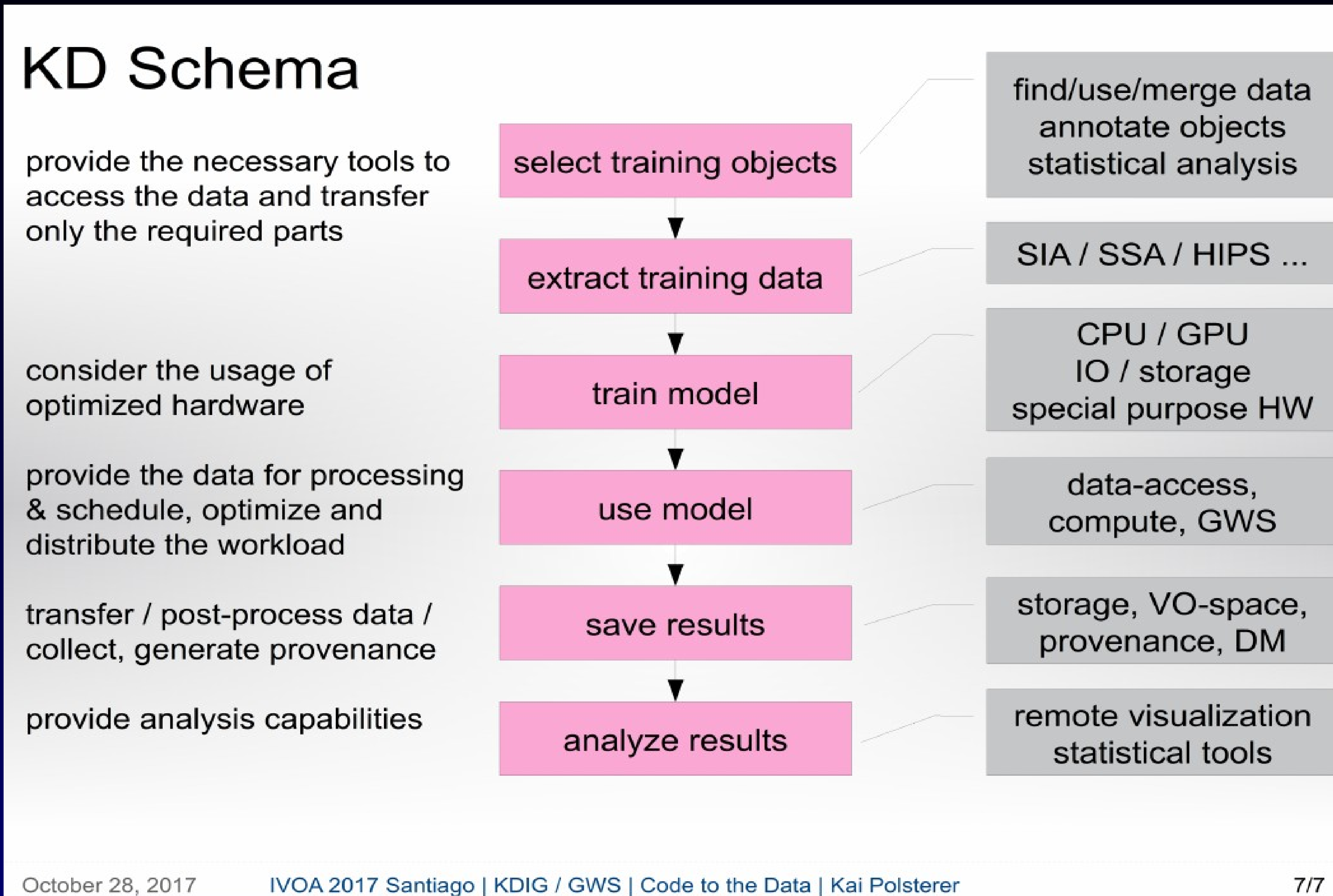


Kavelaars 2019

Science Platforms - SciServer



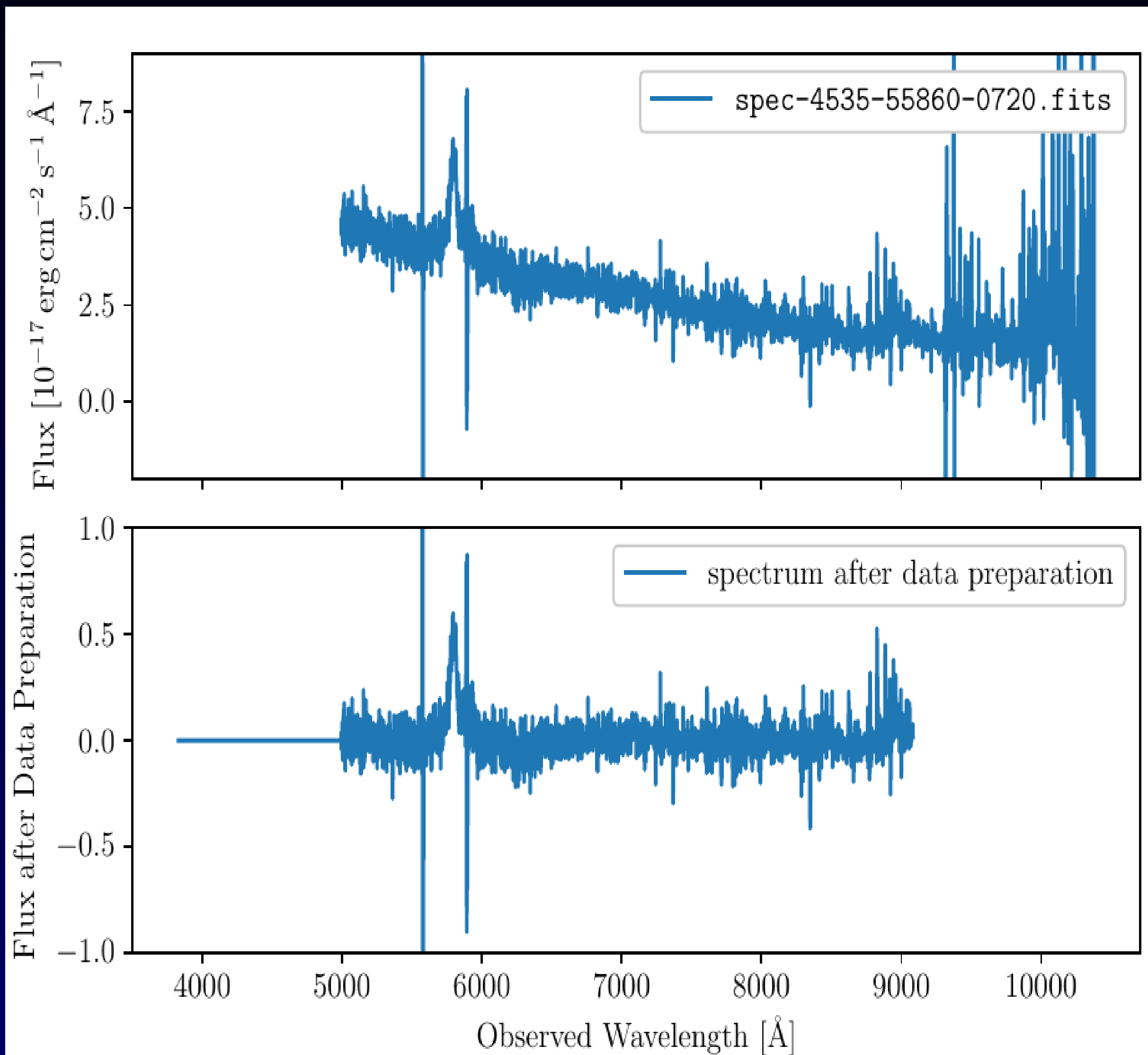
Simple KD with VO



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 ↔ air

Log wavelength ↔ 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

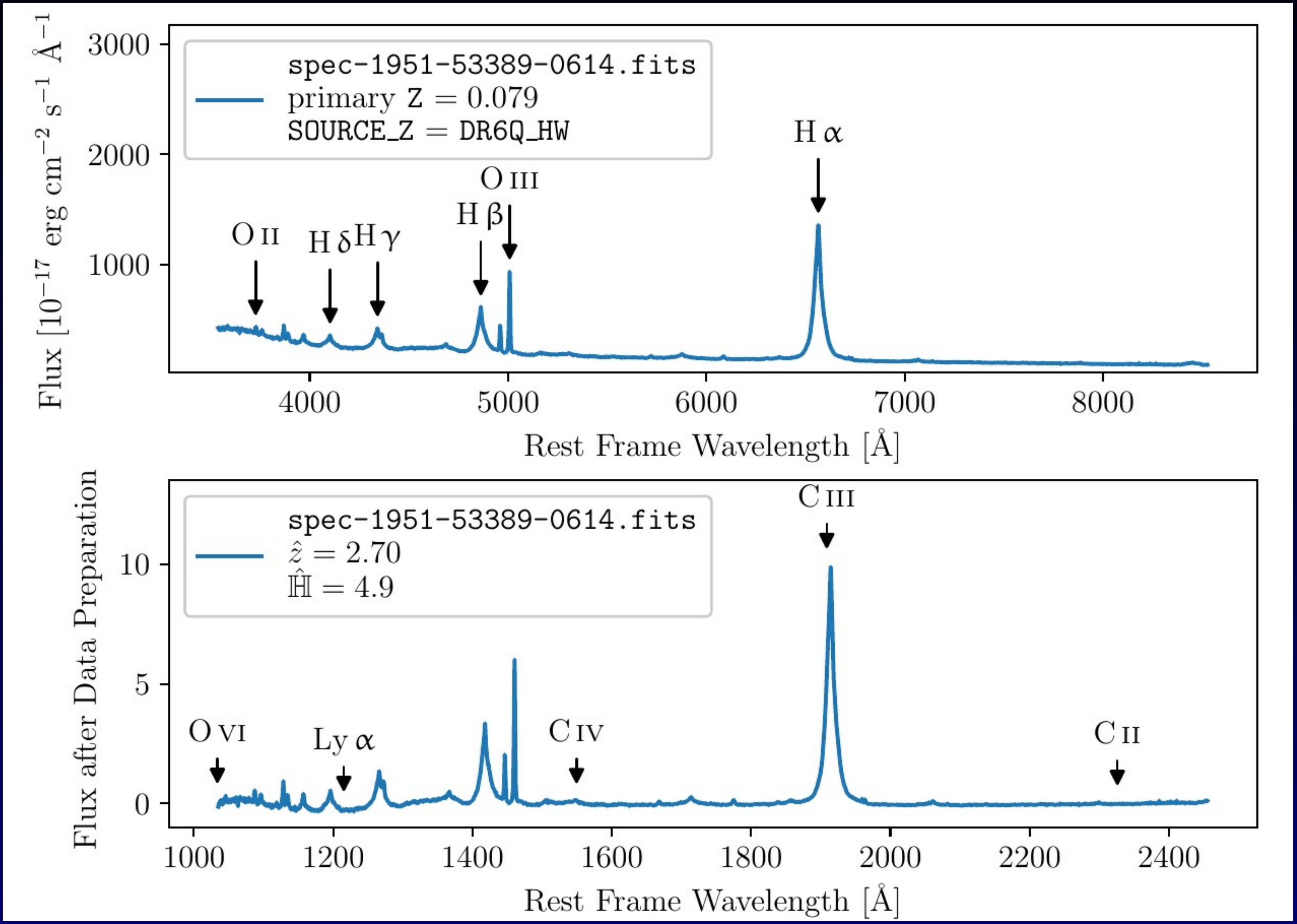
$$x' = \frac{x}{||x||}$$

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

Scaling to unit length

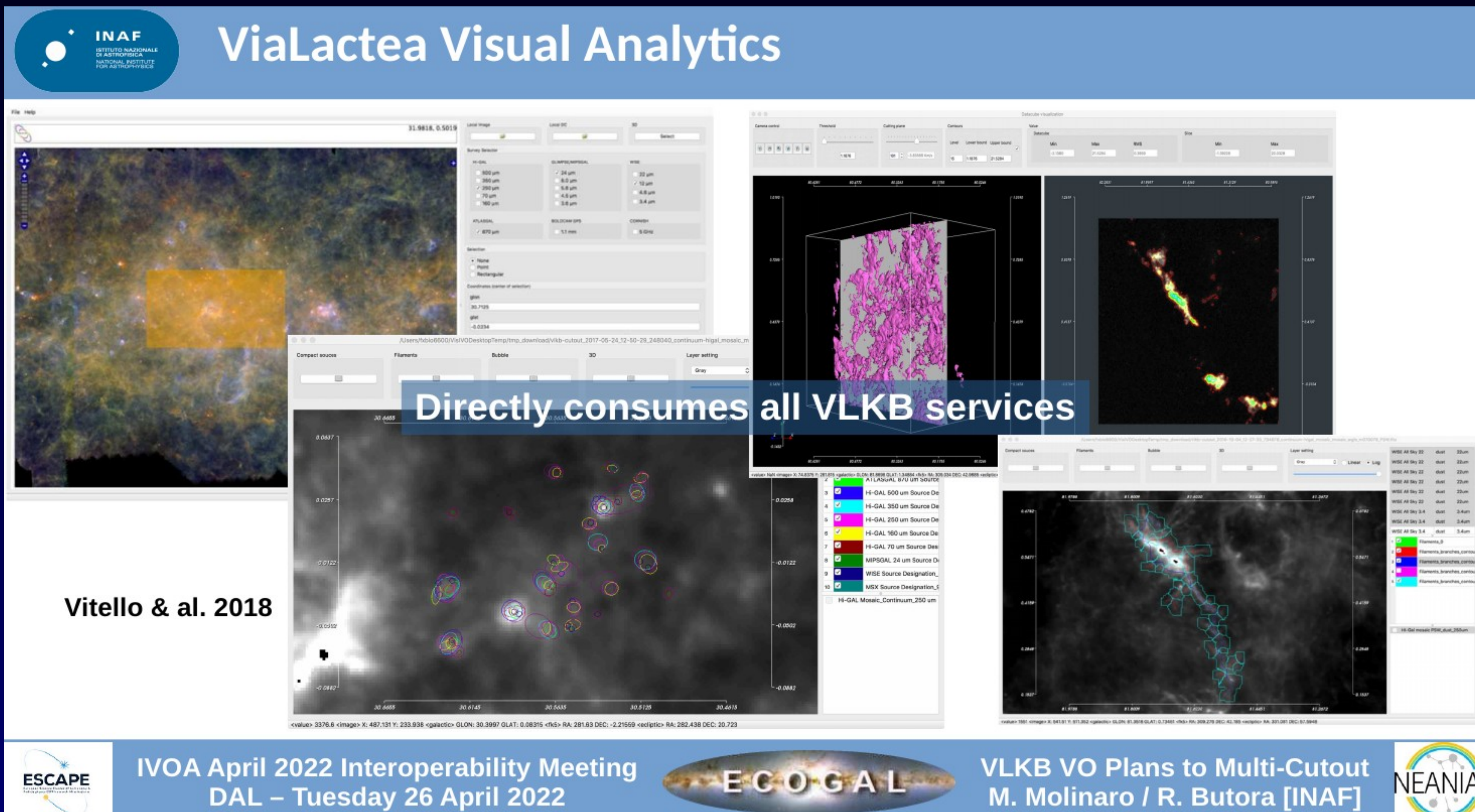
Wavelength dependency

Transformation Using the Redshift

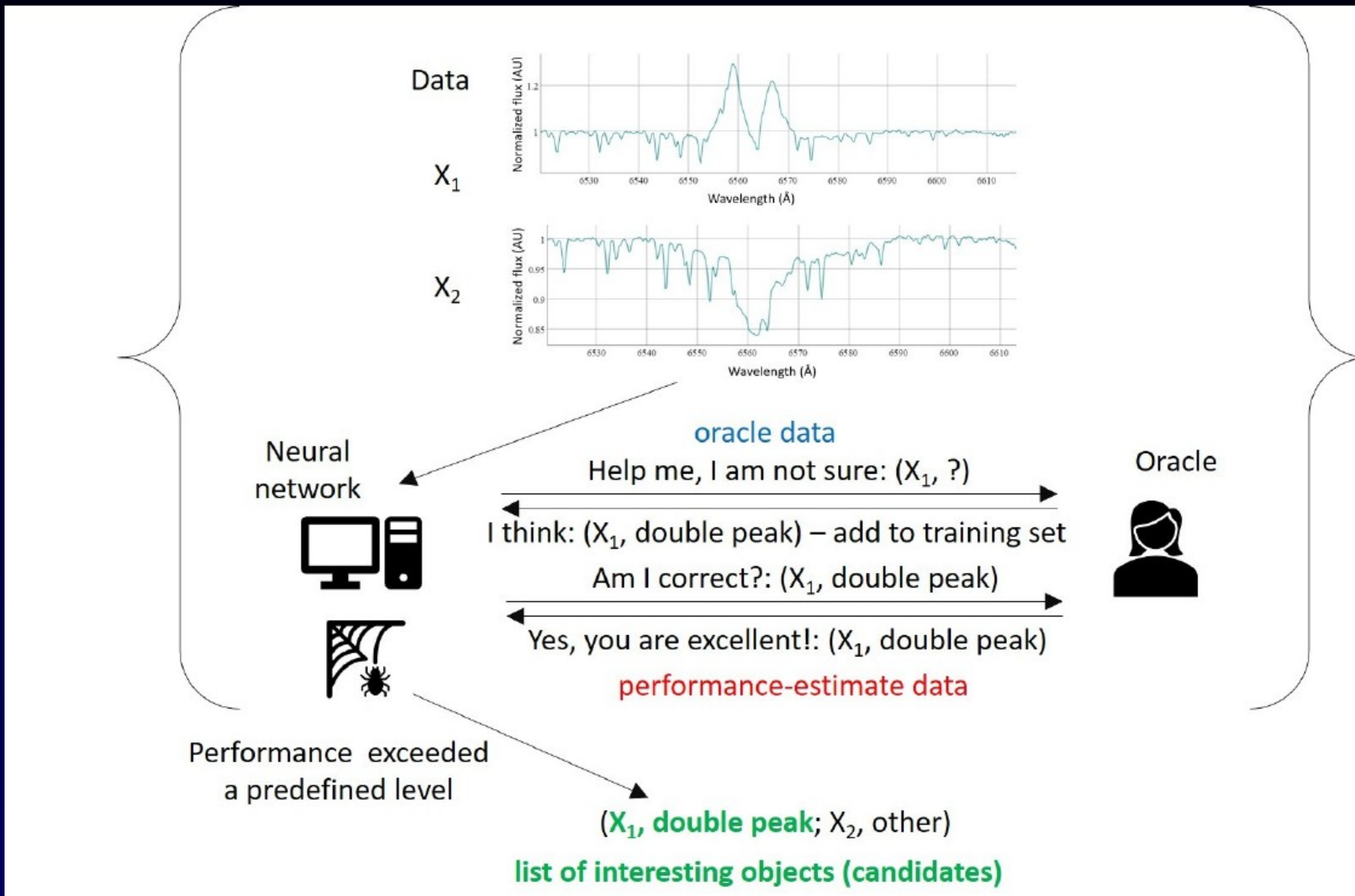


Advanced SODA, UWS

Multi Cut-outs Molinaro (DAL1)

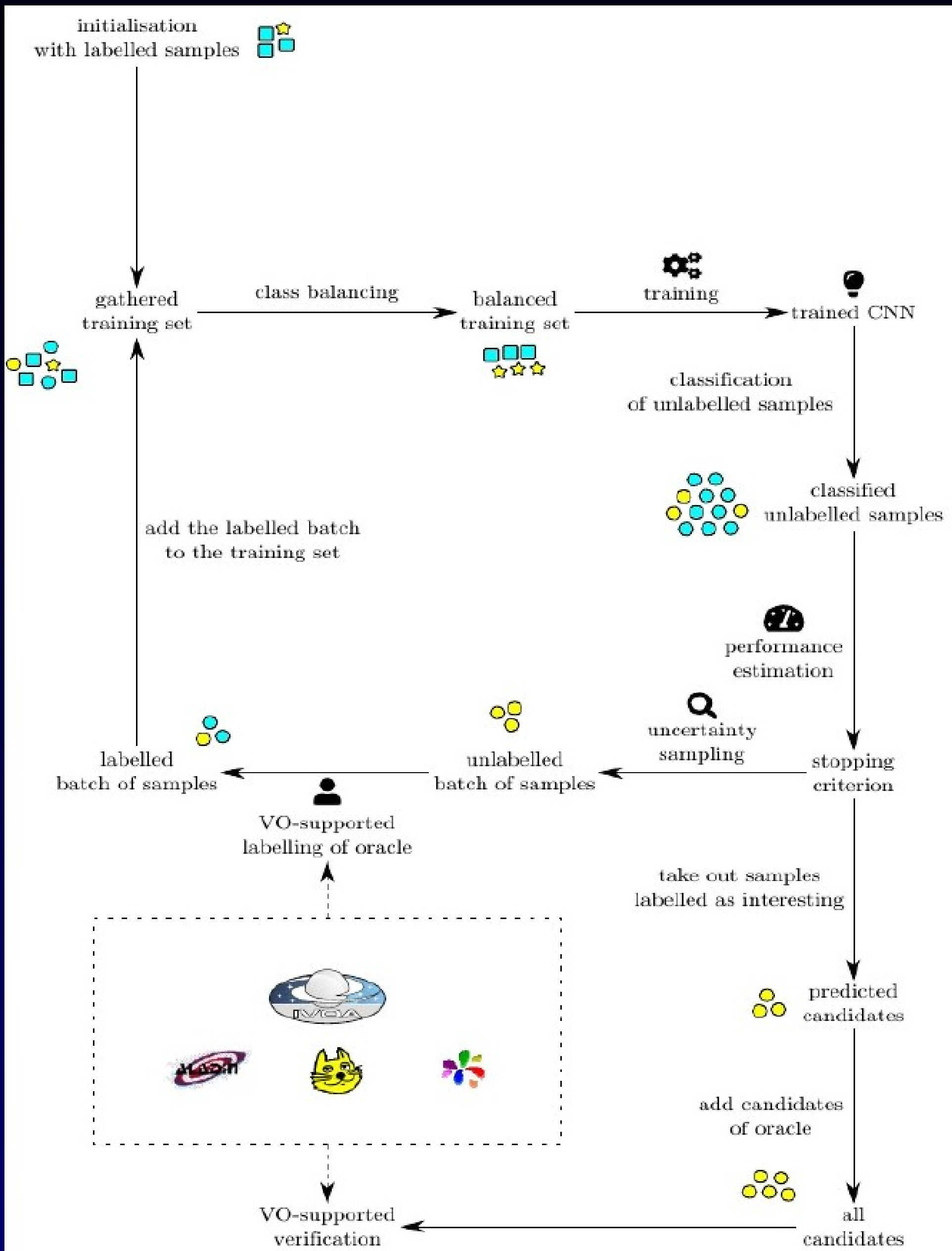


Active Learning (insufficient labels)



Oracle : Human – Machine Interaction

Active Learning



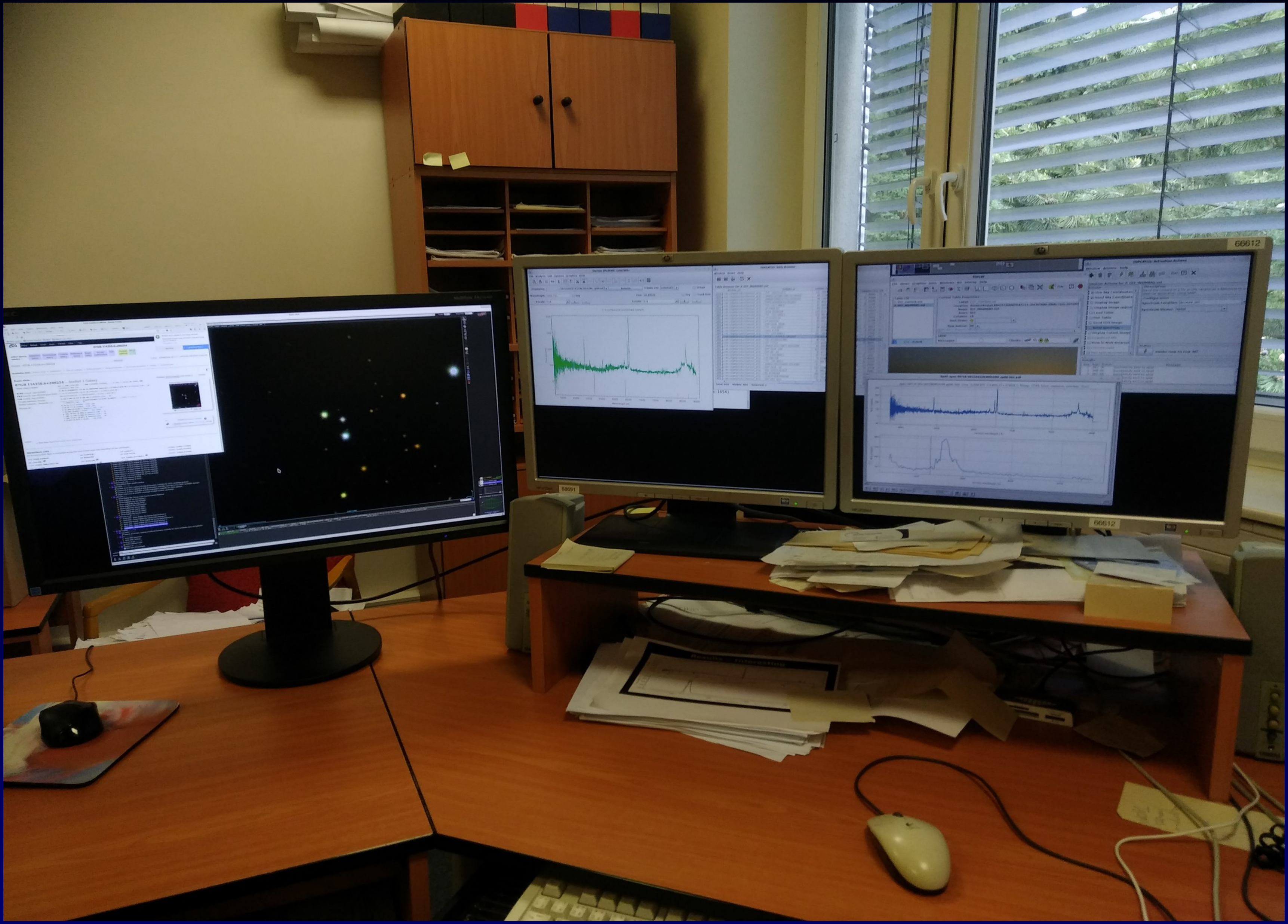
The Oracle (human annotator) decides :

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 he 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. Nádvorník, P. Škoda and P. Tvrdík *Astronomy and Computing 36 (2021) 100463*

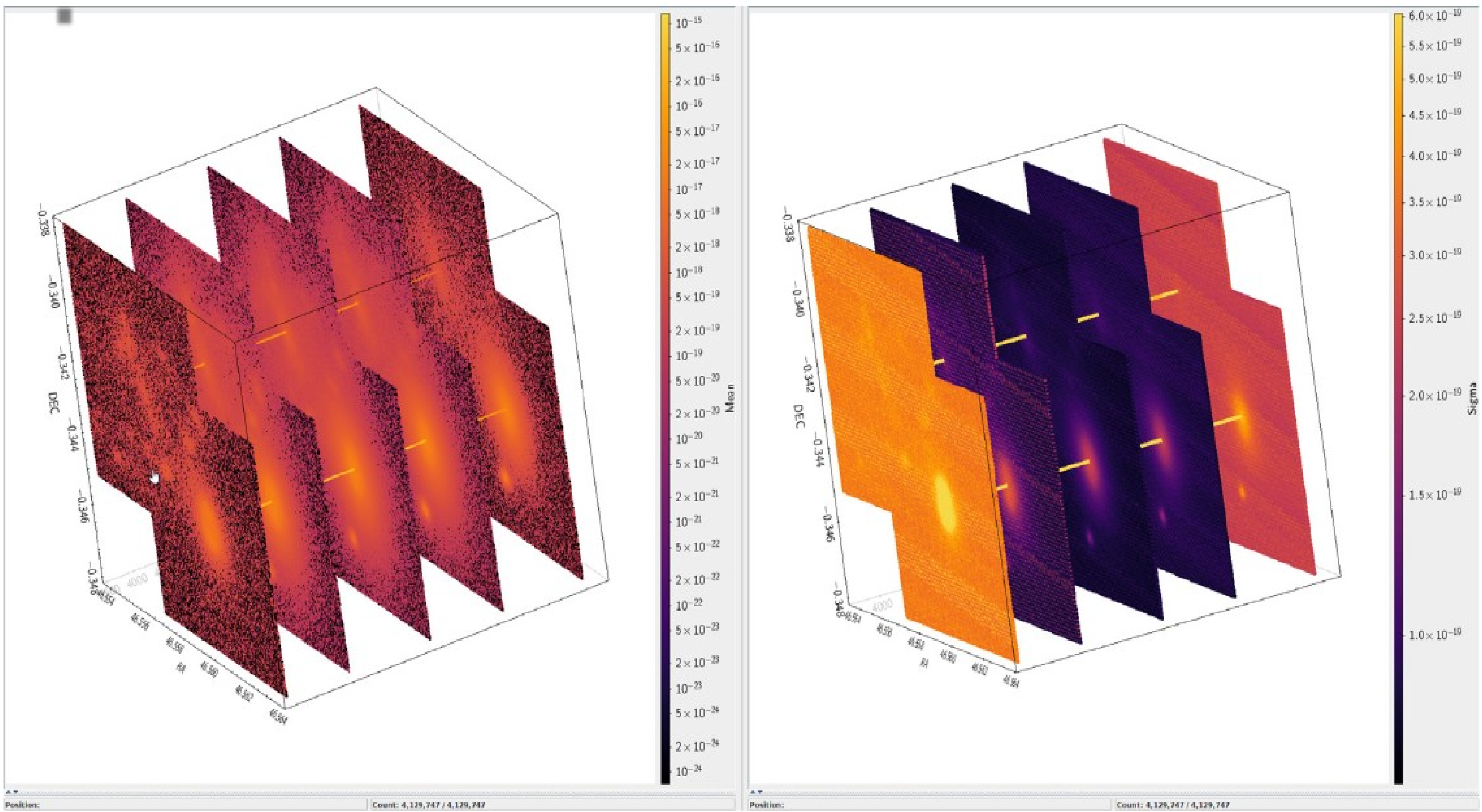


Fig. 7. Screenshot of Galaxy2 data set exported to VOTable and visualized in TOPCAT. The left-hand figure depicts the mean values and the right-hand figure presents the sigma values.

HDF5 structure named Hiearchical 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 !