Deep Learning in Space: The Big Data Problems or Big Problems with Data?

Jakub Nalepa, KP Labs, Poland (jnalepa@kplabs.pl)

Introduction

Hyperspectral images (HSIs) capture very detailed information about scanned objects (**Figure 1**) and, therefore, can be used to uncover various characteristics of the materials present in the analyzed scene.

However, due to a large amount of such data (i.e., its huge volume), HSIs are:

- **Difficult** to **visualize**,
- **Difficult to analyze**, **interpret**, and **label**,
- **Difficult and costly to transfer** (e.g., from an imaging satellite).

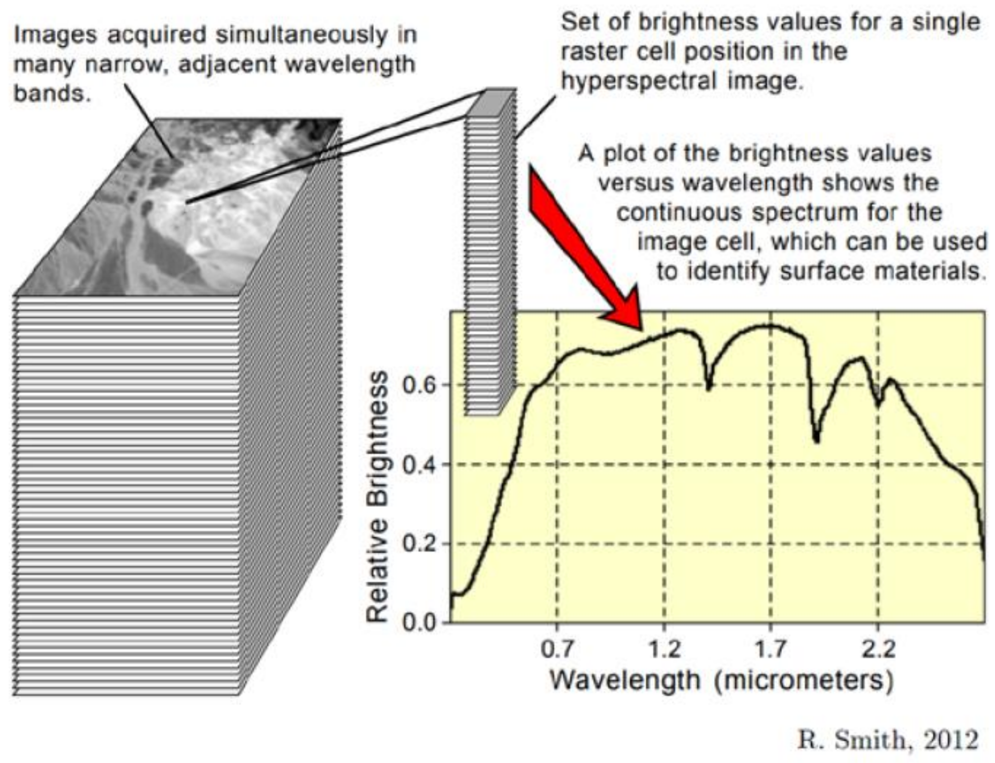


Figure 1: Hyperspectral images contain a large number of bands captured for a contiguous electromagnetic spectrum.

Therefore, we are lacking large and representative ground-truth datasets that could be used to train AI models for analyzing HSI in emerging EO applications.

Existing datasets for HSI classification

There exist state-of-the-art datasets for HSI classification, such as Indian Pines, Pavia University, Salinas Valley, or University of Houston (**Figure 2**), but **how can we use them to train large-capacity deep learning algorithms for other EO applications?**

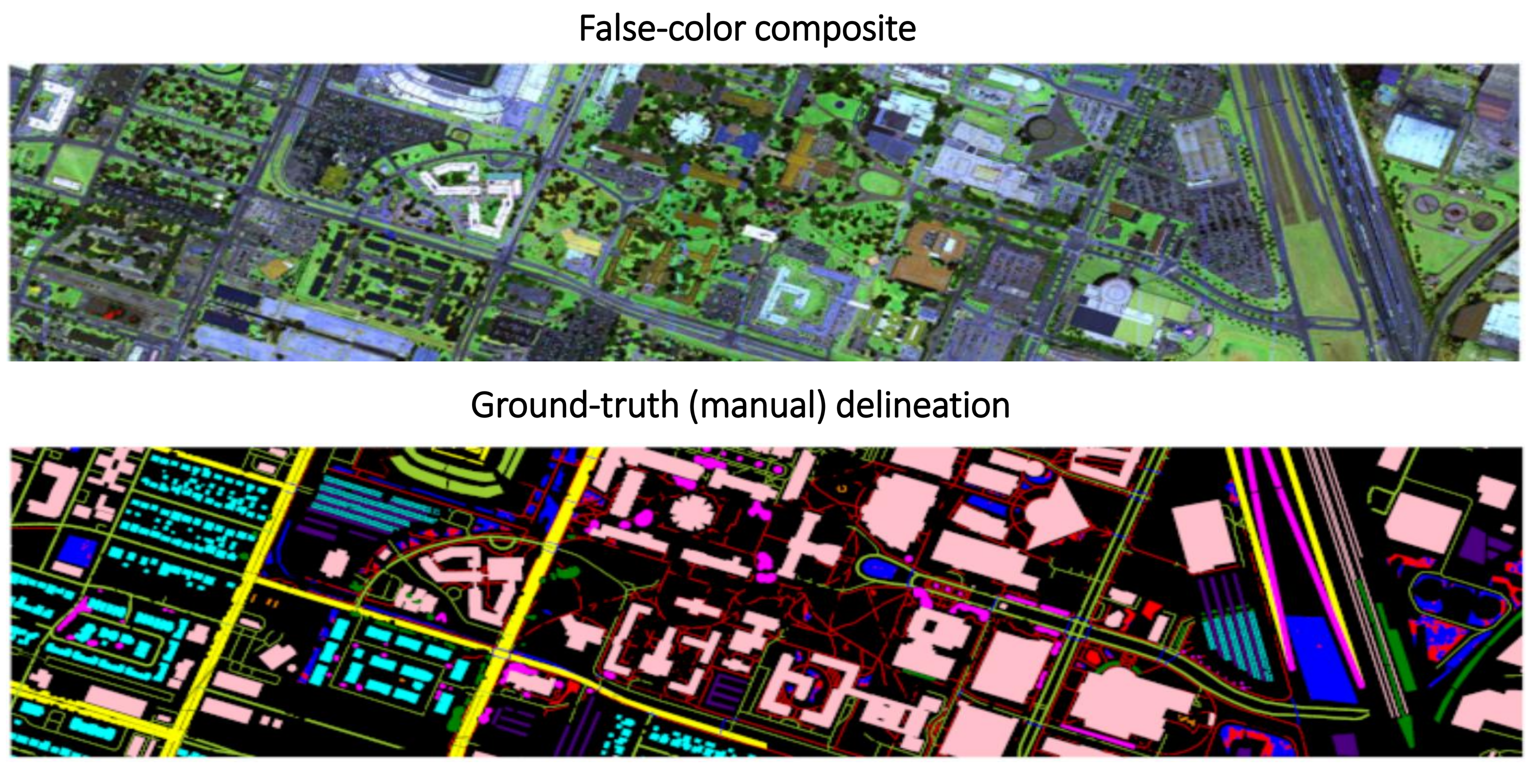


Figure 2: Example hyperspectral scene (University of Houston [7]), together with the corresponding ground truth delineating various objects of interest.

Also, is it **fair to quantify the performance of algorithms over the very same scene, e.g., using random sampling for splitting into training and test sets?** What if some pixels can land in both training and test subsamples (for models that exploit spectral and spatial information while classifying an incoming pixel, **Figure 3**) [1]?

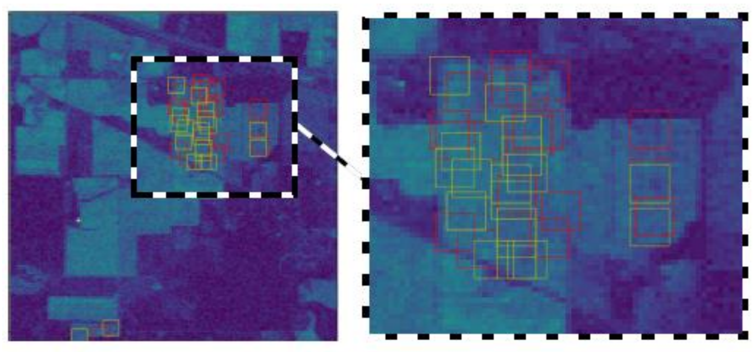
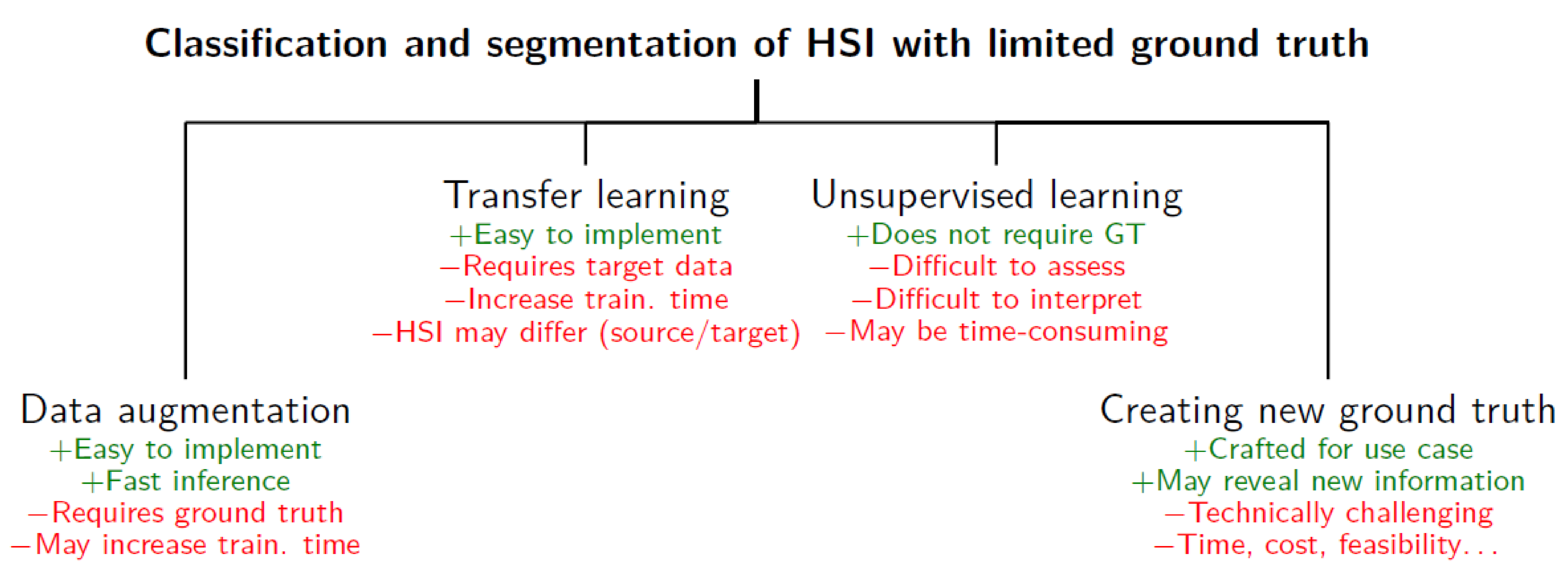


Figure 3: Random pixel selection from the Indian Pines dataset to the training and test sets (overlapping yellow and red squares) can cause a leak of information. Therefore, **we can have over-optimistic estimation of the classification performance of the investigated spectral-spatial algorithms**. This figure comes from [1].

Lacking ground truth: The remedies



Remedy 1: Data augmentation

In **data augmentation,** we synthesize artificial examples based on the available ground-truth data (hence, such ground-truth examples must exist). **For HSI,** we can exploit **generative adversarial networks (GANs), noise injection, guided noise injection,** and other augmentation techniques.

Data augmentation may be utilized **before the training (to increase the size of the training sample),** and **during the prediction, to benefit from the ensemble-like classification approach (Figure 4) [2].** Note that some techniques, e.g., GANs, cannot be used during the inference.

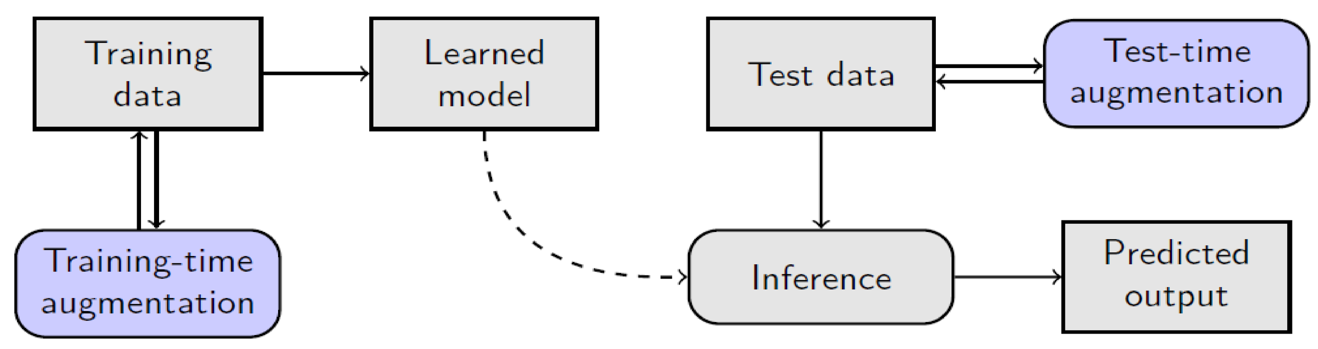


Figure 4: Synthesizing artificial examples through data augmentation helps us increase the size of ground-truth datasets based on the original data distribution, or benefit from ensemble-like approaches. This figure comes from [2].

Remedy 2: Transfer learning

The idea is to **train efficient feature extractors from large training data (the source data),** and later **fine-tune the classifier over the target data of interest (of a much lower size) [3].** As in data augmentation, **target data examples must exist.**

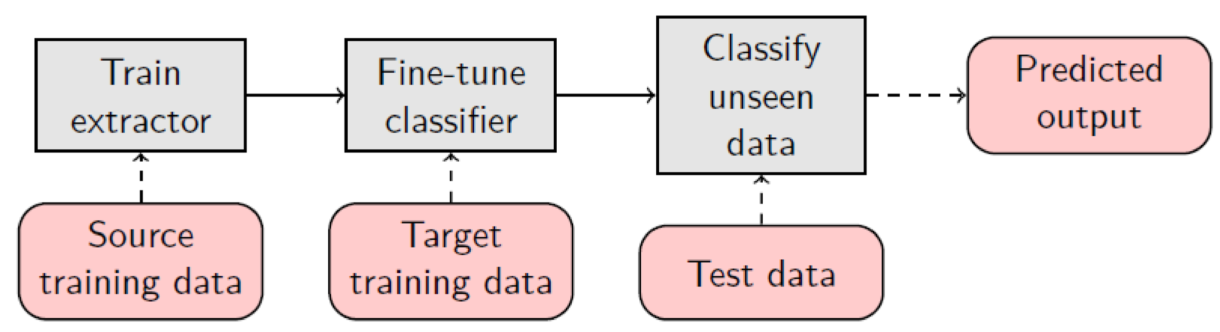
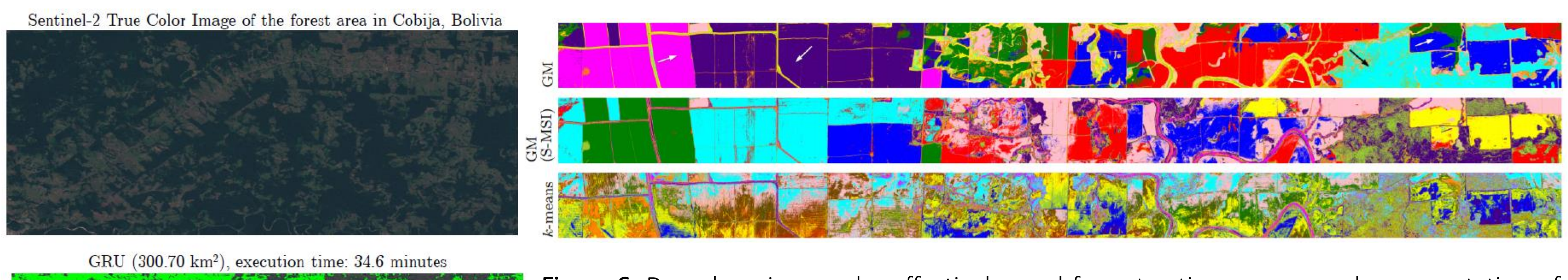


Figure 5: In transfer learning, we build efficient (deep) feature extractors over the source (larger) data, and fine-tune the classification part of a deep learning model over the target (much smaller) data of interest. This figure comes from [3].

Remedy 3: Unsupervised learning

The idea: **group the input data to find coherent regions of similar characteristics** (e.g., spectral) without the ground truth, and then **interpret the segmentation result (Figure 6).**



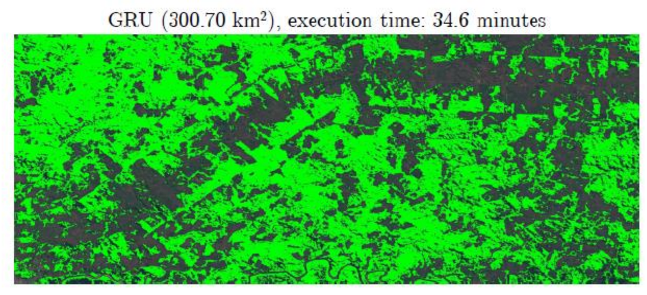


Figure 6: Deep learning can be effectively used for extracting compressed representation of the hyperspectral data (e.g., using various autoencoder-based deep architectures). The interpretation of the segmentation result may be fairly easy, e.g., if we expect just two classes of objects in a scene (on the left – forest vs. non-forest pixels), or quite challenging, if the number of classes is huge (the upper images, segmented using Gaussian modeling and k-means employed over original data and extracted features). This figure comes from [4].

Another advantage: unsupervised segmentation may be considered **a pre-segmentation step** which – if followed by manual assignment of class labels to coherent image regions – can **significantly accelerate the process of generating ground-truth data for emerging use cases**.

Remedy 4: Creating new ground truth

Creating new ground truth is always an option, but it is **costly, time-consuming, does not scale well, is user/area dependent…** But capturing even limited ground truth may be coupled with unsupervised segmentation (and intelligent data reduction and selection, e.g., through cloud detection for HSI captured on-board a satellite – do we need cloudy areas at all?), thus can make the manual assignment of class labels easier (**Figure 7**).

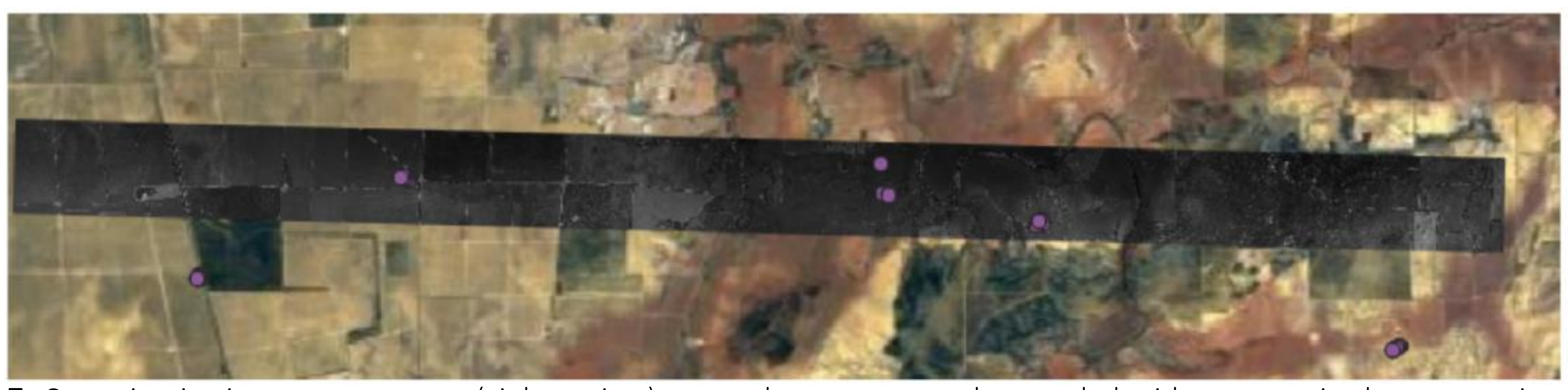
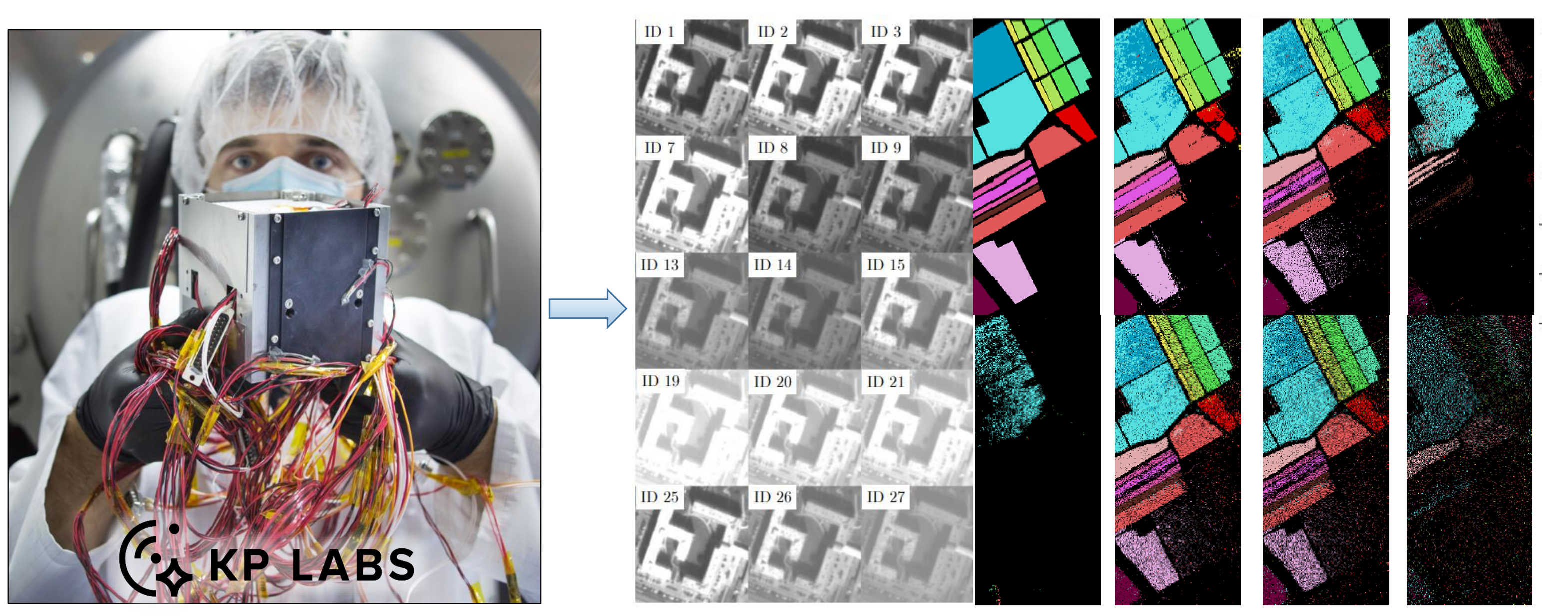
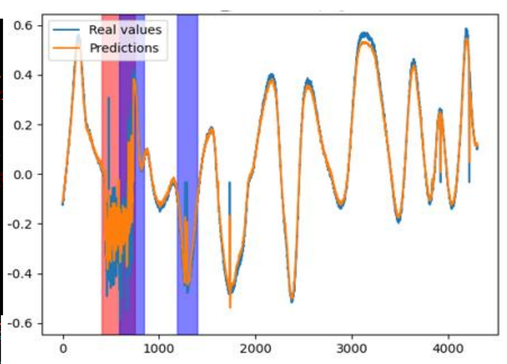


Figure 7: Capturing in-situ measurements (violet points) over a large area may be coupled with unsupervised segmentation (see Figure 6 in which the same area is pre-segmented). Why not to exploit two approaches in HSI analysis? This figure comes from [5].

Remedy 5: Towards digital twins (at the data level)





No ground truth? Why not to simulate it!

Simulators, that reflect the characteristics of a real piece of hardware, can give us lots of advantages (not only in HSI analysis, but also in e.g., detecting anomalies from telemetry):

- We can **simulate various acquisition scenarios** (e.g., atmospheric conditions) – capturing such real data would be extremely costly (or impossible) [6].
- We can **simulate noise of any characteristics**, hence we can **verify the robustness of the deep learning algorithms** (e.g., for HSI classification) against noise that can (and will) happen in space [6].
- We can **simulate lots of data with precise ground truth information** (imagine capturing the real life telemetry with all possible incorrect events for all hardware components…)

References

[1] J. Nalepa et al., IEEE Geosci. Remote. Sens. Lett. 16(8): 1264-1268 (2019)
[2] J. Nalepa, et al., IEEE Geosci. Remote. Sens. Lett. 17(2): 292-296 (2020)
[3] J. Nalepa et al., IEEE Geosci. Remote. Sens. Lett. 17(7): 1228-1232 (2020)
[4] L. Tulczyjew et al., IEEE Geosci. Remote Sens. Lett. 2020, DOI: 10.1109/LGRS.2020.3013205 (in press).
[5] J. Nalepa et al., IEEE Geosci. Remote. Sens. Lett. 17(11): 1948-1952 (2020)
[6] J. Nalepa et al., Remote. Sens. 13(8): 1532 (2021)
[7] Y. Xu et al., IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2019, 12, 1709–1724.