Earth Observation Data Science:
Learning from Peta-pixels and Giga-people
Mihai Datcu

The **volume and variety of valuable Earth Observation** (EO) images as well as non EO related data is **rapidly growing**. The **open free data access** becomes widespread and has an enormous **scientific and socio-economic relevance**. Particularly impacting is the new generation of very high resolution images, with resolutions ranging from 0.5 m to 20 m, due to their potential for observing detailed human activities in a global environmental context. Meanwhile, crowdsourcing and citizen science, become a huge user of EO information and producer of new observation data. The **barrier** lies in the extraction of meaningful information and understanding observations of large extended areas, over long periods of time, with a broad variety of EO imaging sensors in synergy with other related measurements and data as maps, in-situ measurements or **crowdsourcing data**.

EO images are "instrument" measurements, extending the observation beyond the visual information, gathering **physical parameters** of scenes in a broad electromagnetic spectrum. The EO main scope is the understanding of evolving processes, the **Satellite Image Time Series** (SITS) are of a particular importance. As a consequence, the meaningful information contained in the EO images and the **semantic aspects** are much broader and are rather difficult to be formalized and also need information from related data sources for interpretation. The **crowdsourcing data,** as inherently produced by people, has a strong semantic content.

The **challenge** addressed by this proposal is the elaboration of a specific Data Science paradigm for understanding of the EO images, for understanding evolving processes, with a focus on the synergy of SITS with other complementary data.

The overall **objectives** of this proposal addressing the issues listed above can be summarized as follows:
• Elaboration of a **new EO Data Science vision for semantic information** retrieval from multiple geo-spatial data sources and SITS
• Study, discuss and develop new paradigms for interactive **machine learning, and deep learning** specific techniques for multispectral, radar **EO Big Data and crowdsourcing data**.
• **Dissemination**, and demonstration of the developed concepts with the **academic communities and the public at large.**

The **methodology** to achieve the objectives of the proposal follows a stepwise transition and a transfer of the results obtained during our previous collaboration and state-of-the-art methods. The studies and developments to be performed are based on a theoretical formalization in the areas of statistics, signal processing, machine learning, computational intelligence, visualization, and related techniques.
The next 3 main directions will be treated in close synergy:

1. Kernel-based Active Learning for SITS and Related Data: The main goal of this activity is to merge information and to learn from combinations of EO images or from separate data sources, complementing the EO images (e.g., metadata, in-situ observations, or geospatial data). In particular, we will address machine learning techniques based on kernels in continuation of the Cascaded Active Learning method for Object Retrieval, CALOR, method developed jointly with CNAM, in synergy with probabilistic models of external related data. We want to study measures of probability applicable to distinct data sources and SITS in order to treat them commensurately. The expected outcome is learning SITS spatio-temporal evolution patterns by grouping together heterogeneous images and geospatial information. For instance, a new active learning based algorithm will have to combine, at pixel level, the unsupervised clustering results of different features, extracted from heterogeneous information resources, with user-defined semantic annotations in order to calculate the posterior probabilities that allow the final probabilistic searches in a database.

2. DNN for EO and crowdsourcing Feature Learning: This part of our activities shall handle rather innovative approaches using deep learning methods for EO images and heterogeneous data. In our case, deep learning shall serve as a strategy for the extraction of

characteristic patterns and physics- related discoveries from images and/or other data sources representing yet undiscovered content features. When we apply a deep multi-level neural network (DNN) then we can access the final results (that may not refer to a physical meaning) together with intermediate results at selected levels of the neural net (that may refer to typical target area characteristics). The challenges are in designing specialized DNNs and more specifically CNNs together with Deep Stacking and Recurrent Networks for EO images and SITS. The EO images are composed by complicated signatures, depending on the interdependencies among the various channels and strongly influenced by specific parameters as sensor resolution, acquisition modes, coherence conditions, look angle, and definitely on the scene structures. Our goal is to arrive at physically plausible features, able to describe semantically the observed scene and behaving invariance to the sensible parameters.

3. Detection and Learning with Adversarial Samples: Adversarial samples became popular in the area of deep learning, when dealing with large unknown data sets, where undesirable examples altered the classifier model. A typical example is the modification of 8 bit data to 32 bit data when adding noise to the LSBs. In the case of EO, our images are typically represented by 16 bits and multiband data (as multispectral images), or complex-valued observations with amplitude and phase (as SAR images). In all these cases, the visual selection of training samples can be affected by errors, thus including adversarial samples in the active learning process. The solutions proposed are two-fold. At image level, to avoid insertion of adversary samples, we will study transformation methods and information selection for a relevant RGB representation based on entropy measures for multispectral images, or SAR image transformations. Also, the design of the DNNs will be studied to achieve a given invariance. We will focus on contextually disambiguating the meaning in order to ensure a consistent training.

The study cases above suggest the huge potential, however very little explored and largely unknown to most scientists and users in EO fields, of using machine learning and deep learning tools for the purpose of the semantic and physically meaningful discovery at the data-exploration level. Where the most advanced methodologies are addressing the different nature of the data, i.e. sensor data, spatio-temporal nature, regular vs. irregular samples, incomplete, incommensurable, and noisy observations. A general, interactive, easy to use explorative tool that could more or less automatically tell the scientist what variables are the most important in explaining a given phenomenon, or automatically cluster a dataset into relevant classes could have an enormous potential in a diverse range of scientific fields – providing the researcher with very informative hints on how to explain a given phenomena directly from the data, before any modelling is done.

Organization of the Joint Research: in close collaboration and coordination with the CNAM, the following activities are planned:

- A Data Science for EO series of lectures, organized at CNAM monthly, addressing a broad community as the Universities, Research and Industry in Ile-de-France.
- Ph.D. students will be enrolled and co-directed at CNAM, as Ph.D. students enrolled at CNAM of invited for joint work with the team at CNAM.
- The topics of the research will be included in a series of lectures at graduate/post-graduate level at CNAM and potentially other Universities in Ile-de-France. Four specific lectures are planned.
- During the Chair period, will be done the best efforts to acquire third party projects jointly with CNAM, based on the obtained results and focusing at very interdisciplinary teams: EO, Computer Vision, Multimedia, Social sciences, or business analysts.

CNAM and the academic, research and industry communities involved will optimally exploit the complementarity of the developed resources.