DEEPCUBE: EXPLAINABLE AI PIPELINES FOR BIG COPERNICUS DATA

Ioannis Papoutsis[1], Alkyoni Baglatzi[1], Souzana Touloumtzi[1], Markus Reichstein[2], Nuno Carvalhais[2], Fabian Gans[2], Gustau Camps-Valls[3], Maria Piles[3], Theofilos Kakantousis[4], Jim Dowling[4], Manolis Koubarakis[5], Dimitris Bilidas[5], Despina-Athanasia Pantazi[5], George Stamoulis[5],Christophe Demange[6], Léo-Gad Journel[6], Marco Bianchi[7], Chiara Gervasi[7], Alessio Rucci[7], Ioannis Tsampoulatidis[8], Eleni Kamateri[8], Tarek Habib[9], Alejandro Díaz Bolívar[9], Zisoula Ntasiou[10], Anastasios Paschalis[10]

[1]National Observatory of Athens, Institute for Astronomy, Astrophysics, Space Applications & Remote Sensing, [2]Max Planck Institute for Biogeochemistry, [3]University of Valencia, Image Processing Laboratory, [4]Logical Clocks AB, [5]National and Kapodistrian University of Athens, Department of Informatics and Telecommunications, [6]GAEL Systems, [7]TRE ALTAMIRA, CLS Group, [8]INFALIA [9]MURMURATION SAS - Flockeo.com, [10]Hellenic Fire Service

ABSTRACT

The H2020 DeepCube project leverages advances in the fields of Artificial Intelligence and Semantic Web to unlock the potential of Copernicus Big Data and contribute to the Digital Twin Earth initiative. DeepCube aims to address problems of high socio-environmental impact and enhance our understanding of Earth's processes correlated with Climate Change. To achieve this, the project employs novel technologies, such as the Earth System Data Cube, the Semantic Cube, the Hopsworks platform for distributed deep learning, and visual analytics tools, integrating them into an open, cloud-interoperable platform. DeepCube will develop Deep Learning architectures that extend to non-conventional data, apply hybrid modeling for data-driven AI models that respect physical laws, and open up the Deep Learning black box with Explainable Artificial Intelligence and Causality.

***Index Terms*—** Data cubes, Artificial Intelligence, semantic web, hybrid modeling, explainable AI, causality, climate change, Digital Twin Earth

1. INTRODUCTION

The Copernicus program is believed to be a game changer for both science and the industry. Free and open data available at this scale, frequency, and quality constitutes a fundamental paradigm change in Earth Observation (EO). However, the availability of the sheer volume of Copernicus data outstrips our capacity to extract meaningful information. The EO community needs technology enablers to propel the development of entirely new applications at scale.

Deep Learning (DL) has been one of the fastest-growing trends in big data analysis. It is only relatively recently that DL was introduced to the EO research community for information extraction from big satellite data. The majority of the applications that use DL though, seem to reiterate old EO problems, which now can be solved faster and provide incrementally higher accuracy with respect to conventional Machine Learning (ML) approaches.

Furthermore, DL leads to highly nonlinear, overparameterized models. They excel in prediction accuracy, but such complexity hampers interpretability and trustworthiness. Predictive accuracy is important but often insufficient, and interpreting what the models learned becomes important, especially in problems with economical, societal or environmental implications. The lack of interpretability, i.e. the degree to which a human can understand the cause of a decision has become a main barrier of DL in its wide-spread applications for geosciences.

Finally, EO data becomes useful only when analyzed together with other sources (e.g., geospatial & in-situ data) and turned into knowledge. Linked data is a data paradigm that studies how one can make Resource Description Framework (RDF) data available on the web and interconnect it with other data with the aim of increasing its value. Nevertheless, there are only a handful of applications that showcase the semantic integration of linked EO and non-EO products.

The H2020 DeepCube project (Jan. 2021 - Dec. 2023, https://deepcube-h2020.eu/) leverages advancements in the fields of AI and semantic web to unlock the potential of big Copernicus data. It aims to address problems that imply high environmental and societal impact, enhance our understanding of Earth's processes, correlated with the climate emergency, and feasibly generate high business value, in line with the Destination Earth and the Digital Twin Earth objectives. To achieve this, DeepCube integrates mature and new technologies into an open interoperable platform that can be deployed in cloud environments, DIAS included. The platform is then used to develop novel DL pipelines to extract value from big Copernicus data. DeepCube develops

DL architectures that extend to non-conventional data and problems, introduces a novel hybrid modeling paradigm for data-driven AI models that respect physical laws [1], and opens-up the DL black box through Explainable AI (XAI) and Causality. We showcase these in six applications.

2. TECHNOLOGIES

DeepCube makes use of mature technology enablers that have been developed in other European Commission and European Space Agency funded research. In DeepCube these enablers are integrated to an interoperable environment allowing EO and AI specialists to create value chains from a wide offer of raw EO and non-EO big data. This environment is the DeepCube platform (Fig. 1), which will scale to big Copernicus datasets, designed to share resources and to define dataflows in a coherent integrated solution. DeepCube platform will be deployed into more than one cloud environments, including Copernicus DIAS. Its individual components are briefly described next.

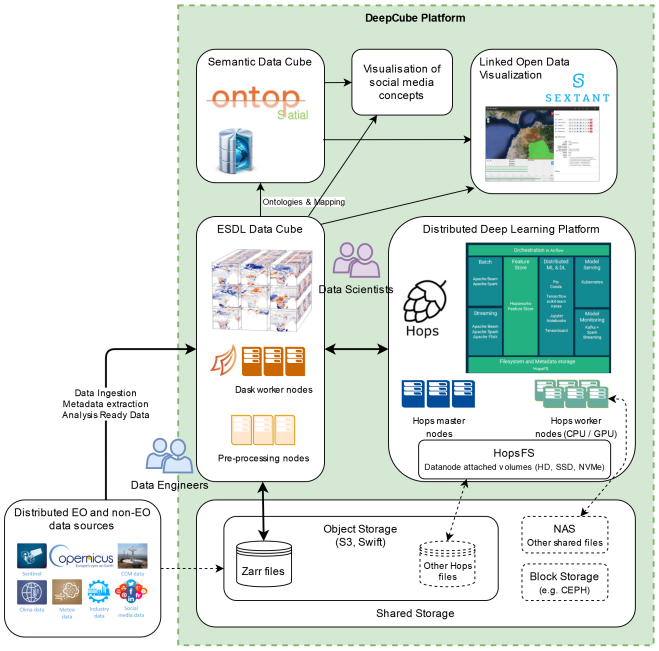

Fig. 1. High level architecture of the DeepCube platform.

The **Earth System Data Cube (ESDC)** developed by Earth System Data Lab project, seeks to be a service to the scientific community to facilitate access and exploitation of multivariate data sets in Earth Sciences to actually understand the interactions between the Earth's subsystems.The core part of the ESDC is the data in analysis-ready form, together with tools and methods to generate, access, and exploit the ESDC. A data cube essentially consists of screened, or Analysis Ready Data (ARD), with the dimensions "latitude", "longitude", "time", "variable". Further dimensions can be added as a result of an analysis. Currently ESDC supports a common spatio-temporal grid [2], DeepCube will advance to create Data Cubes where information layers are stored in heterogeneous spatio-temporal resolution. ESDC is committed to open source computations, and open data usage. Dynamic resource allocation and rapid scalability of ESDC are its cornerstones for data analysis on the cloud.

The **Semantic EO Data Cube** [3] enables the semantic enrichment of ESDC. The Semantic Data Cube allows users to query metadata, EO data, other Linked Open Data (LOD), and information/knowledge extracted from the data using a semantic query language, thus creating new value chains. In a semantic data cube [3], at least one categorical interpretation exists for each observation in an image (i.e., each pixel). EO data co-exists with its interpretation and can also be queried using the same high-level query language that is used for querying interpretations. For example, a user can query the reflectance values of certain bands in an image (e.g., for calculating an index) and in the same query also refer to an interpretation of these values. Semantic data cubes are also enriched by other kinds of data (e.g., other kinds of geospatial data such as OSM). In this way, the querying possibilities for a user become even larger. DeepCube will develop the first semantic data cube technology internationally by extending the geospatial ontology-based data access system Ontop-spatial [4].

Hopsworks is an open-source **Data-Intensive AI** platform for developing and operating end-to-end ML pipelines at scale. Hopsworks provides first-class support for popular open-source frameworks for distributed data processing, data engineering and data science. In addition, Hopsworks supports DL on large volumes of data, such as those produced by the Copernicus program, using distributed training. Distributed training uses many GPUs and data-parallel model training to reduce the time required to train models by adding more GPUs. Hopsworks leverages Apache Spark to make distributed training easier for programmers. However, modern approaches to distributed training require developers to rewrite their code when moving from using a single GPU to hyperparameter tuning (using lots of GPUs) to distributed training. DeepCube will develop a comprehensive new framework that unifies single-host training, hyperparameter tuning, and distributed training. We will also expand Hopsworks to support model-parallel training, as well as API support for distributed semi-supervised learning and self-supervised learning. As such, DeepCube will build a state-of-the-art and the most feature complete framework for distributed DL.

DeepCube will extend **Sextant** [5], a web based and mobile ready platform for **visualizing**, exploring and interacting with linked geospatial data. Sextant is a user-friendly application that allows both domain experts and non-experts to take advantage of semantic web technologies, creating thematic maps by combining spatio-temporal information with other data sources, e.g. industrial intelligence, socio-economic

data, etc., allowing visual analytics based on big Copernicus data. In addition, DeepCube will develop user interfaces offering multiple ways of visualisation and filtering of social media data, detected locations and visual concepts, allowing analytics on top of them.

3. APPLICATIONS

3.1. Forecasting localised drought impacts in Africa

Climate change will lead to an accumulation and intensification of various climate extremes [6]. Drought and heat waves, as experienced repeatedly in the last decade, are expected to become more frequent in the future, as the corresponding persistent weather situations become more and more probable. The effects on various sectors are substantial, as could be seen, for example, from the effects on agriculture, inland waterways, and consequently nutrition and energy supply.

There are two significant gaps that will be addressed by DeepCube: the first one relates to lack of methods for assessing, in fine resolution, drought impact at the local level. This requires downscaling from meteorological scales to sub-km level using satellite data. The second gap is a lack of understanding of memory effects considering ecosystem dynamics, after a drought event. A better understanding will be achieved with so-called hybrid dynamic models [1], which model the system partly with physical equations, partly with ML.

3.2. Climate induced migration in Africa

In the current context of climate change, extreme heat waves, droughts and floods are not only impacting the biosphere and atmosphere but the anthroposphere too. Human populations are forcibly displaced, which are now referred to as climate-induced migrants. On the agenda of the United Nations Framework Convention on Climate Change, for instance, there is an item dedicated to migration, displacement and human mobility. The problem has obvious environmental, societal and economic implications, in both adaptation and mitigation to climate change, as well as for assistance to their home states. Modeling, anticipating, characterizing and understanding the severity of migration flows and the direct and latent factors are of paramount relevance.

There is a growing number of media reports assuming the link of climate change, conflicts, and forced migration. However, there is little empirical evidence supporting that climate change and migration are interrelated [7]. At present, there is no theoretical approach to adequately represent the causal mechanisms through which climate change induces human displacement and migration flows. This will be the first time that advanced causal inference schemes are developed to investigate the climate-induced migration in Africa.

Therefore, DeepCube will identify the main environmental and socio-economic drivers of human mobility and develop models able to reproduce and forecast migration flows, apply causal discovery methods to gain a deeper understanding of the characteristics of the climate-induced migration flows and establish the causal relationships of environmental and socio-economic drivers with human mobility in sub-Saharan Africa.

3.3. Fire hazard forecasting in the Mediterranean

Climate change is playing an increasing role in determining wildfire regimes, with future climate variability expected to enhance the risk and severity of wildfires in many biomes including Southern Europe [6]. Fire hazard forecasting systems linked with the operational authorities (Civil Protection, Fire Brigade/Service etc.), would increase their preparedness and enhance the emergency response capacity in a changing climate.

DeepCube will identify the climatic, vegetation status and anthropogenic drivers that impact the most fire proneness based on multivariate historical data analysis on the Mediterranean. Based on these insights, the application will use AI bound by an ecosystem modeling [1] to model short and mid-term fire hazard and make more accurate and with less uncertainty future predictions using EO data time-series analysis. XAI techniques (permutation analysis, visualization of features-heatmap activations, and clustering activations) will be used to open-up the DL box and gain trust on what the model has learnt. Finally fire hazard forecasts will be combined with LOD to assess fire risk for assets (population, environment, economic activity) on the ground.

3.4. Global volcanic unrest detection & alerting

Interferometric Synthetic Aperture Radar (InSAR) can systematically provide ground deformation estimations over volcanic areas, see 6-day repeat pass cycle of Sentinel-1AB. Fringes detected in Sentinel-1 wrapped interferograms over volcanic areas indicate the onset of deformation, which is usually due to magma chamber fill-in at depth. Such activity is considered as precursor for a potential eruption.

Having the work by Anantrasirichai et al. [8], as a starting point, DeepCube will research DL architectures that can automatically detect the presence of ground deformation triggered by volcanic unrest, within wrapped interferograms, towards establishing a volcanic deformation alert service, covering several volcanoes globally.

3.5. Automated infrastructure monitoring with InSAR

SAR-derived information is used to produce millimetric-precision ground surface deformation maps. Thanks to Sentinel-1 SAR revisit time, new deformation maps can be delivered to end-users on a regular basis [9], showing average deformation rates (mm/yr) of Persistent Scatter (PS) “points” and their displacement time series. Each information layer is made of hundreds of thousands of measurement points,

and can be used for detecting significant instabilities on critical infrastructures thus contributing to plan and optimize mitigation actions.

However, no automated processes are in place to robustly detect hotspots, i.e. zones for which displacement time series show a significant change in trend motion. In addition, for zones experiencing these changes, no indication is given to end-users about possible reasons and driving mechanisms. DeepCube will attempt to link any deformation hotspots to a possible reason for trend change, using DL on InSAR data and sparse in-situ geodetic measurements for training and fusion.

3.6. Copernicus services for sustainable tourism

Tourism is one of the pillars of the modern economy. It constitutes more than 10% of global GDP with a CAGR of 3+%. The number of international tourists is forecasted to rise to 1.8 Billion in 2030, making it crucial to find efficient ways to handle this growth, preserve the fragile destinations and adapt to the increasing demand over limited hospitality infrastructures. Additionally, more than 65% of European travellers have declared that they are striving to make their travels more sustainable but do not find the right information or the possibility to assess their environmental footprint.

DeepCube will create a new commercial service, by producing a pricing engine for tourism packages, which incorporates the environmental dimension. The goal is to calculate a suite of price coefficients for a travel agency to apply to its packages, considering environmental impact automatically, utilizing Copernicus and data (water quality degradation, marine pollution, air pollution), product characteristics (ecological potential), and supply and demand information coming from social media streams. The application will be set-up as a reinforcement learning problem and a prototype will be developed for the northeast coast of Brazil nearby the Lencois national park.

4. CONCLUSIONS

We see DeepCube as a showcase of the Digital Twin Earth potential, by 1) delivering the DeepCube platform as a technology enabler for the deployment of end-to-end AI pipelines on big EO data regardless of the underlying cloud infrastructure, and 2) designing and testing new AI architectures to address significant scientific questions related to Climate Change and generating business value via the joint analysis of EO with industrial data.

The DeepCube platform consists of mature, high technology readiness level, components. This interoperable platform will be a DeepCube legacy which could be deployed in different cloud environments. The platform will provide novel solutions for all phases on an EO-based AI pipeline, from data ingestion, to big data organisation (data cubes), feature engineering, semantic annotation, distributed DL, semantic reasoning and visualisation.

In addition, DeepCube will test a hybrid modeling approach for geophysical parameters estimation, enhanced through XAI for "physics-aware" AI applications. DeepCube will also perform causality analysis to understand and interpret patterns, cause and effects on diverse datasets, including satellite, social media and socio-economic data. Finally, it will employ for the first time AI on complex Sentinel-1 SAR data, an archive of the order of PBs, currently the richest asset that remains underexploited. We expect that the first concrete results will be shared by the end of 2021.

DeepCube will deliver to the community Data Cubes with ARD and training datasets allowing to capture hidden trends for key environmental variables. These cubes will be made available for reuse by June 2021.

REFERENCES

[1] Reichstein, M., Camps-Valls, G., Stevens, B. et al. "Deep learning and process understanding for data-driven Earth system science", Nature, 566, 195–204, doi: 10.1038/s41586-019-0912-1, 2019

[2] Mahecha, M. D., Gans, F., Brandt, G., et al. "Earth system data cubes unravel global multivariate dynamics", Earth Syst. Dynam., 11, 201–234, doi: 10.5194/esd-11-201-2020, 2020.

[3] Augustin, H., Sudmanns, M., Tiede, D., Lang, S., Baraldi, A. "Semantic Earth Observation Data Cubes", Data, 4, 102. doi: 10.3390/data4030102, 2019.

[4] Bereta K., Xiao G., Koubarakis M. "Ontop-spatial: Ontop of geospatial databases", Journal of Web Semantics, 58, doi: 10.1016/j.websem.2019.100514, 2019.

[5] Nikolaou C., Dogani K., Bereta K., Garbis G., Karpathiotakis M., Kyzirakos K., Koubarakis M., "Sextant: Visualizing time-evolving linked geospatial data", Journal of Web Semantics, 35, 1, 35-52, doi: 10.1016/j.websem.2015.09.004, 2015

[6] Shukla P.R., Skea J., Calvo Buendia E., et al. "IPCC, 2019: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems", 2019.

[7] Brzoska M., Fröhlich C., "Climate change, migration and violent conflict: vulnerabilities, pathways and adaptation strategies, Migration and Development", 5:2, 190-210, doi: 10.1080/21632324.2015.1022973, 2016.

[8] Anantrasirichai N., Biggs J., Albino F., Bull D., "A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets", Remote Sensing of Environment, 230, doi: 10.1016/j.rse.2019.04.032, 2019.

[9] Raspini F., Bianchini S., Ciampalini A., et al., "Continuous, semi-automatic monitoring of ground deformation using Sentinel-1 satellites" Scientific Reports, 8:7253, doi: 10.1038/s41598-018-25369-w, 2018.