EUROGEO WORKSHOP 2022

The Present & Future of Data Cubes

Moderated by **Tom Hengl**

and **Carson Ross** of OpenGeoHub

Learn more here:

OpenGeoHUB
Connect · Create · Share · Repeat

OPEN EARTH MONITOR

ATHENS 7-9 DECEMBER 2022

HELLENIC REPUBLIC
MINISTRY OF
DEVELOPMENT AND INVESTMENTS

GSRI
GENERAL SECRETARIAT FOR
RESEARCH AND INNOVATION

European Commission

SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

- Please connect to slido.com and post your questions throughout the presentations to discuss at the end of the session.
- We will also have questions and polls to respond to throughout, and we will look at and share results after all presentations are completed.

HELLENIC REPUBLIC
MINISTRY OF DEVELOPMENT AND INVESTMENTS
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

What is your field as a data user / producer?

- EU Policy
- Large international institution (i.e. UN)
- Private Sector
- Academia & Research
- Civil Society
- Other

https://ec.europa.eu/eusurvey/runner/OEM_FAIR-Survey_D01_Copy

☑ Save a backup on your local computer (disable if you are using a public/shared computer)

Survey on FAIR geospatial data

Fields marked with * are mandatory.

Disclaimer

The European Commission is not responsible for the content of questionnaires created using the EUSurvey service - it remains the sole responsibility of the form creator and manager. The use of EUSurvey service does not imply a recommendation or endorsement, by the European Commission, of the views expressed within them.

OPEN EARTH MONITOR

HELLENIC REPUBLIC
MINISTRY OF DEVELOPMENT AND INVESTMENTS
GSRI
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
GEO
European Commission

Online registrations for
the virtual event

EuroDataCube.com

Grega Milcinski

Sinergise

Slides available upon request to:
grega.milcinski@sinergise.com

SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

Which of the following data services do you access at least once a year?

a. Google Earth Engine
b. Microsoft Planetary computer
c. Amazon AWS Open Data
d. EDC Open Data
e. Vito Copernicus Land Service API
f. Sinergise Sentinel Hub API
g. NASA REST API
h. Other

F.A.I.R. information cube

Stefan Jetschny

NILU – Climate and environmental research institute, Norway

FAIRiCUBE

NILU

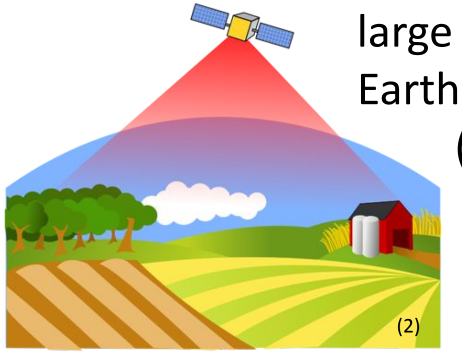
Learn more here:

Overview

- Motivation
- FAIRiCUBE Objectives
- Use cases

«Deliver the power of ***data cubes and ML*** *to decision/policy makers and data scientists»*

Motivation

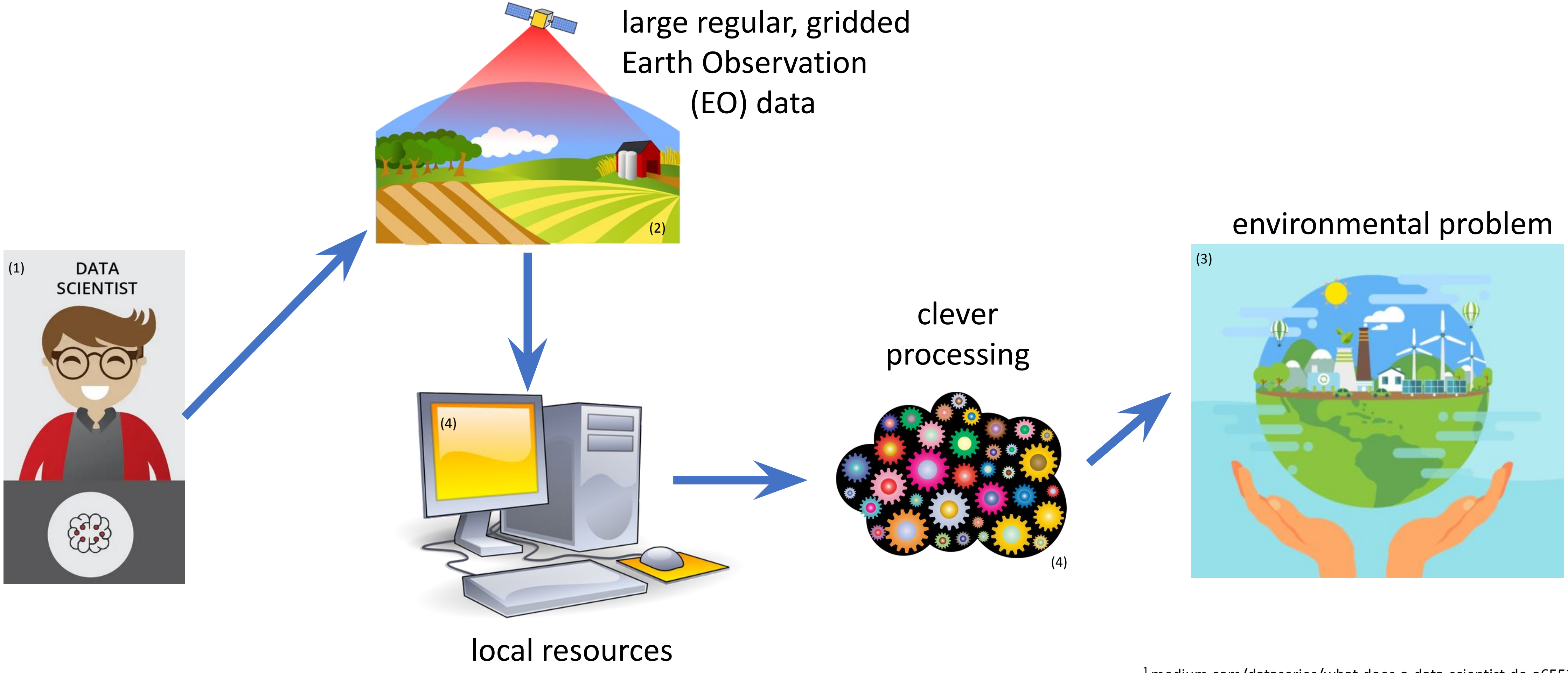


large regular, gridded Earth Observation (EO) data

environmental problem

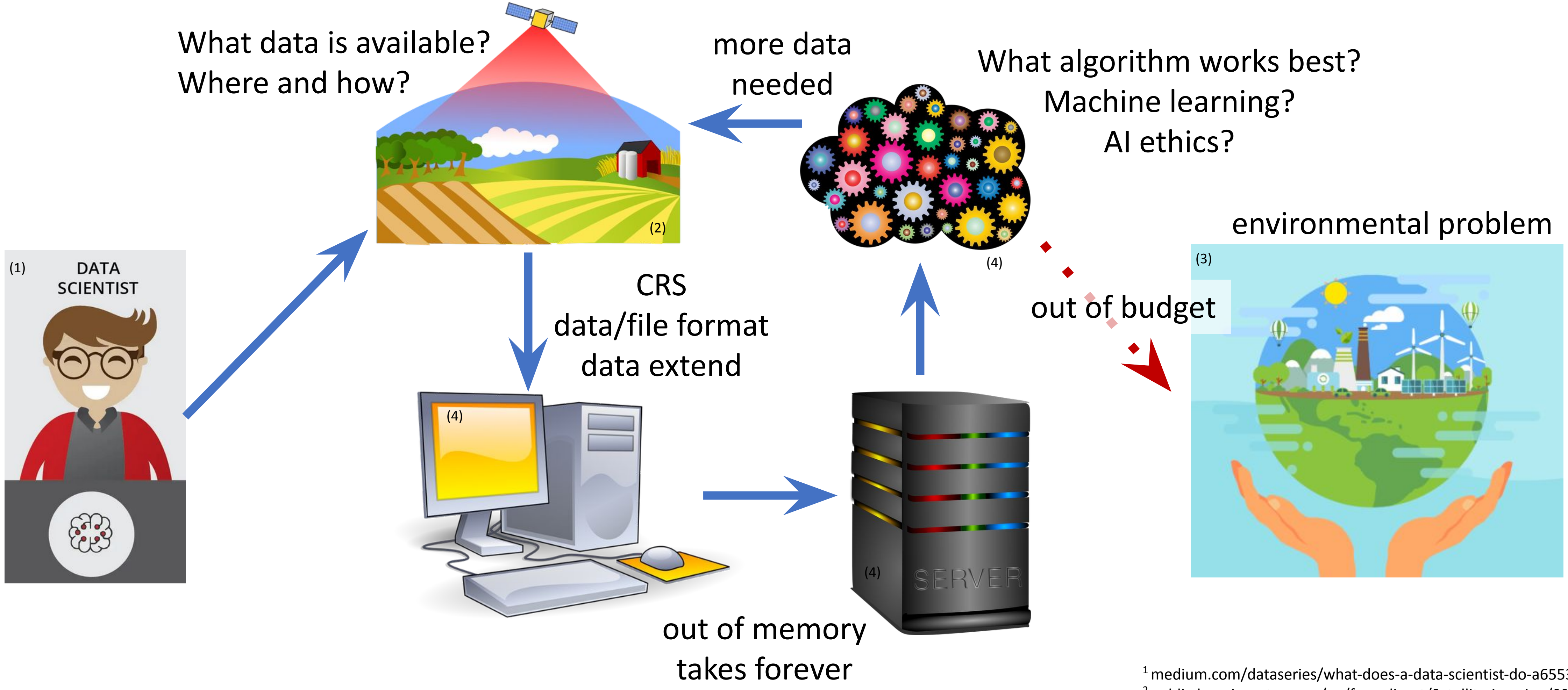
[1] https://medium.com/dataseries/what-does-a-data-scientist-do-a6553dc720f
[2] https://publicdomainvectors.org/en/free-clipart/Satellite-imaging/82710.html
[3] https://www.mindler.com/blog/careers-to-save-the-environment/

Motivation – Ideal case



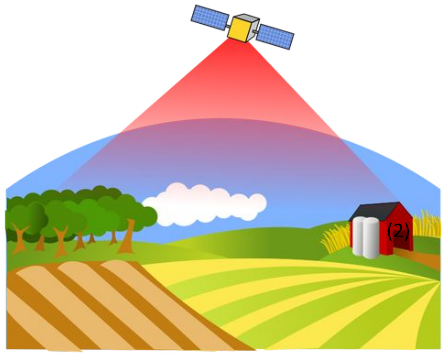
[1] medium.com/dataseries/what-does-a-data-scientist-do-a6553dc720f
[2] publicdomainvectors.org/en/free-clipart/Satellite-imaging/82710.html
[3] www.mindler.com/blog/careers-to-save-the-environment/
[4] www.rawpixel.com

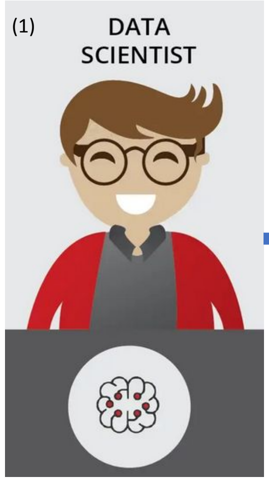
Motivation – Real case



[1] medium.com/dataseries/what-does-a-data-scientist-do-a6553dc720f
[2] publicdomainvectors.org/en/free-clipart/Satellite-imaging/82710.html
[3] www.mindler.com/blog/careers-to-save-the-environment/
[4] www.rawpixel.com

Motivation – FAIRiCUBE case





FAIRiCUBE

FAIRiCUBE Hub

environmental problem

[1] medium.com/dataseries/what-does-a-data-scientist-do-a6553dc720f
[2] publicdomainvectors.org/en/free-clipart/Satellite-imaging/82710.html
[3] www.mindler.com/blog/careers-to-save-the-environment/
[4] www.rawpixel.com

FAIRiCUBE Objectives

- data catalogue of pre-gridded, pre-aligned, pre-referenced EO data
    - create your own custom data cube

- data storage & compute resources

- data processing catalogue (including Machine Learning)

- community platform sharing
    - data layers, data products
    - processing steps (e.g. Python Notebooks)
    - ML models & AI ethics
    - documents

- if all this works: F.A.I.R. ✓

FAIRiCUBE Objectives

The **core objective** of FAIRiCUBE is to enable players from beyond classic Earth Observation (EO) domains to **provide, access, process, and share gridded data and algorithms** in a FAIR and TRUSTable manner.

The project's goal is to **leverage the power of Machine Learning** (ML) operating on multi-thematic datacubes for a broader range of **governance and research institutions** from diverse fields, who at present cannot easily access and utilize these potent resources.

Project duration: 07/2022 – 06/2025
kick-off held 01.07.2022, technical kick off 13/14.10. 2022 ✓

FAIRiCUBE Partners

- data cubes are mature, performance, scalability and capability is proven
- platforms are well established

- data cube platform providers are project partners
  - Jacobs University / rasdaman (EarthServer), Germany
  - EOX (EuroDataCube), Austria
- research institutes
  - **<u>NILU</u>**, Climate and environmental research institute, Norway
  - Wageningen university and research, Holland
  - Museum of Natural History, Vienna, Austria
- environmental SMEs
  - space4environment, Luxembourg
  - 4sfera, Spain
  - epsilon, Italia

Use case «demonstrators»

- 4 Use Cases being executed on the FAIRiCUBE Hub
  - addressing **EU green deal** action items
  - urban / regional focus
  - Barcelona / Vienna / Oslo / Luxembourg
  - using different scale length, areas, objectives, technical platforms
  - How can we make this data available through the FAIRiCUBE Hub

1) Urban adaptation to **climate change**

2) Spatial and temporal assessment of neighborhood building stock (urban, **circular economy**)

3) **Biodiversity** and **agriculture** nexus (regional)

4) **Biodiversity** occurrence cubes (urban / regional, **climate change**)

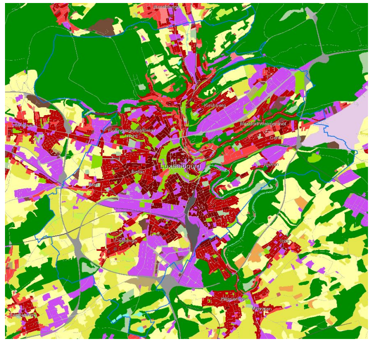
FAIRiCUBE

EOX
rasdaman raster data manager
JACOBS UNIVERSITY
space 4 environment
WAGENINGEN UNIVERSITY & RESEARCH
nhm naturhistorisches museum wien
NILU
4sfera INNOVA
EPSILON Italia

Urban adaptation to climate change

space 4 environment

- Cities face a lot of challenges combating the impacts of climate change <u>“If you can’t measure it, you can’t manage it”</u> basis for all actions are reliable and accessible data and information of high quality
- Creation of dashboards with indicators describing status and trends of parameters of single cities, comparable across Europe since based on European data
- Collect local data and indicators based on stakeholder exchange; harmonise them with other data
- Use of simulations/models/scenarios/machine-learning to simulate possible outcomes and the consequences thereof
- We try to provide a “tool kit” to cities to make better informed decisions by having as much data as possible at their fingertips and being able to simulate their decision-making process:
    - What happens if one parameter changes intentionally by decisions?
    - Which parameter has the strongest influence on improving the quality of life?



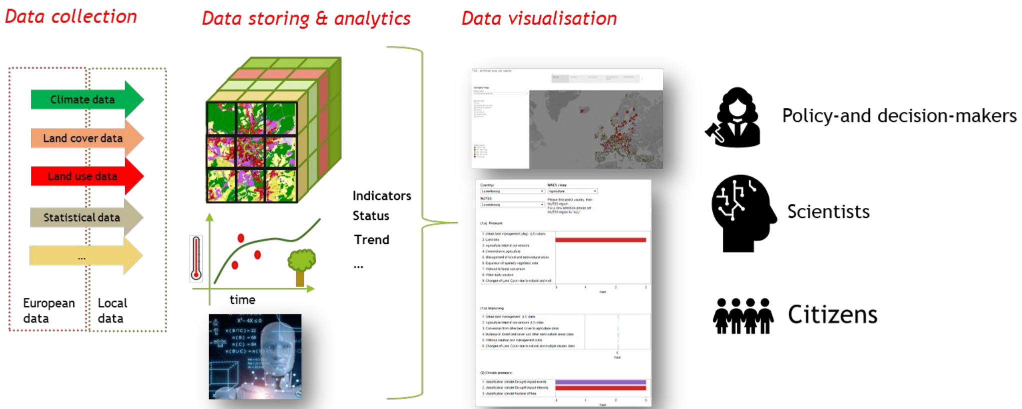


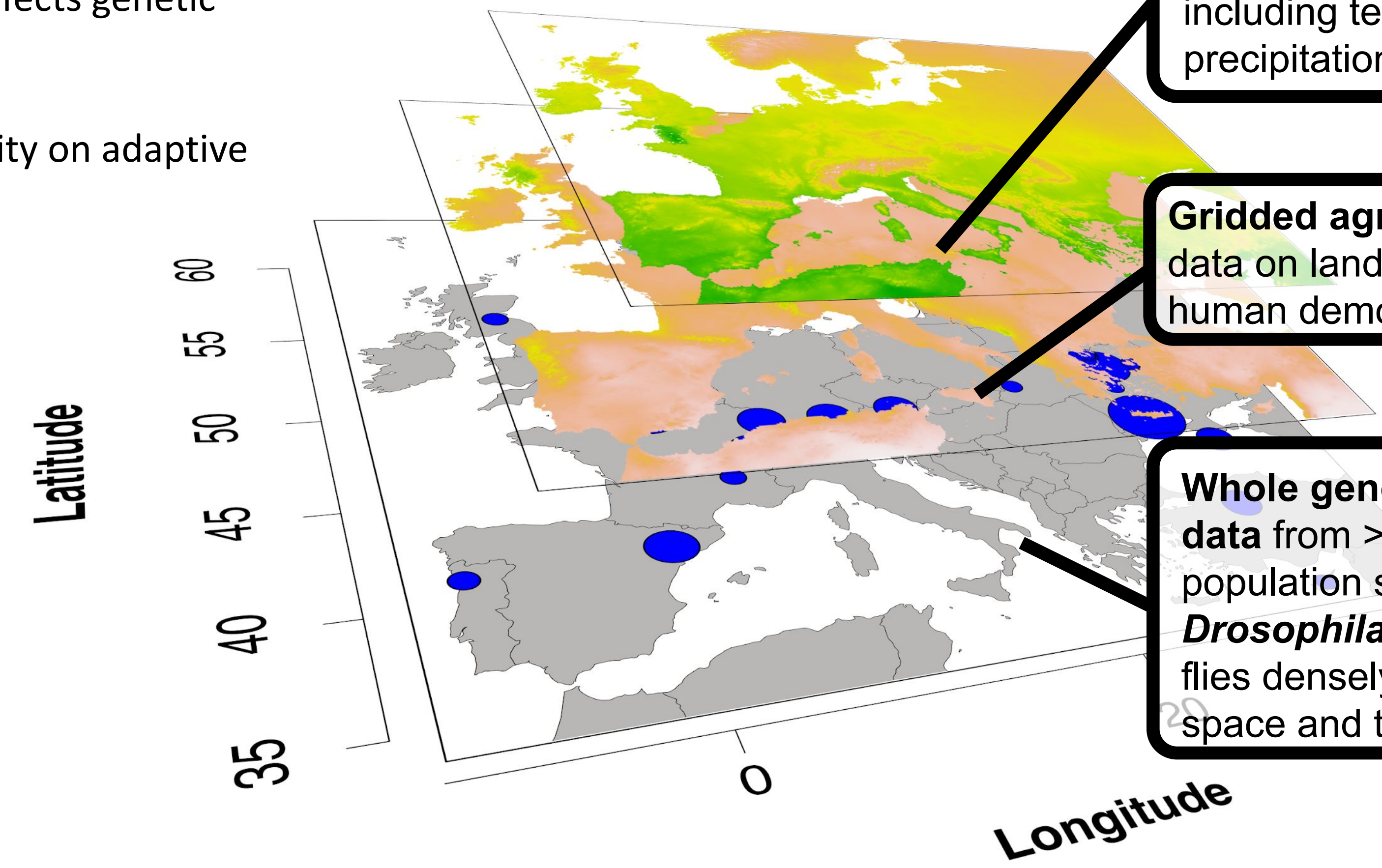
Image sources: Urban Atlas map of Luxembourg from https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018; pedestrian zone in Guangzhou/China from https://urban.uw.edu/news/accessible-pedestrian-routing-tools-expand-to-three-washington-cities/

Linking climatic and genetic variation for biodiversity inference

phm naturhistorisches museum wien

DROSEU EUROPEAN DROSOPHILA POPULATION GENOMICS CONSORTIUM

- API to intersect quantitative genomic and spatial datasets
- To investigate how climatic variation affects genetic diversity
- To assess the influence of human activity on adaptive evolution



Gridded climatic data, including temperature, precipitation, seasonality, etc.

Gridded agricultural data and data on land-use, traffic and human demography

Whole genome sequencing data from > 170 European population samples of ***Drosophila melanogaster*** fruit flies densely sampled through space and time

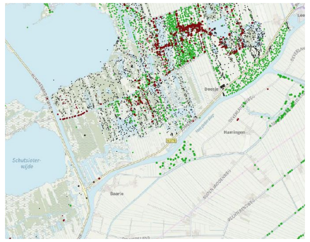
Drosophila melanogaster

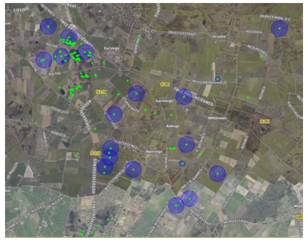
as a genomic and ecological model organism

WAGENINGEN UNIVERSITY & RESEARCH

Agriculture and Biodiversity Nexus

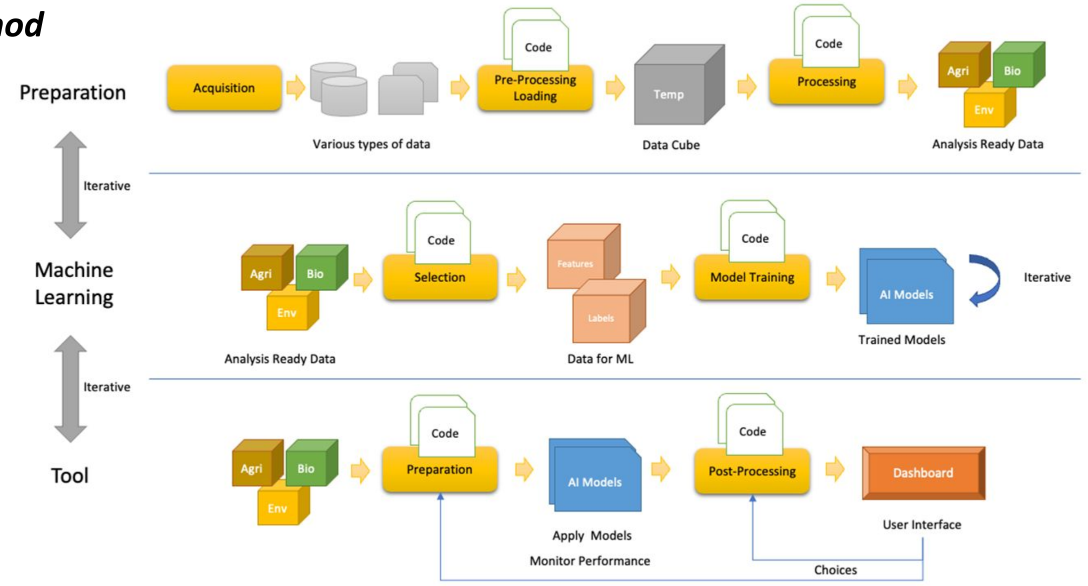
- Clarify links between biodiversity and agriculture by detecting patterns in the data through the use of **explainable machine learning;**
- Provide a step forward in making more precise **estimates of biodiversity** in a spatial context;
- Improve the **data challenges** by the use and sharing of prepared **data and occurrence cubes;**
- Add to the **current knowledge base** by cooperating with **domain experts;**
- Ultimately assist stakeholders in selecting more **nature-inclusive practices promoting biodiversity**





Method

s



Data

- **Biodiversity –** species occurrences (NDFF, GBIF, statistical), existing RS derived products (HR-VPP,...)
- **Agricultural –** crop parcel registration, agricultural nature management, farm management activities, .
- **Environmental –** remote sensing (signals, markers, indices), land use/cover, terrai soil, weather, ...

Building stock model at neighbourhood level

- Buildings are the cornerstone of our living; however, they have faced multidimensional challenges, for instance:
    - They consume tremendous amount of natural resources to be built and generate enormous construction waste at the end of their life cycle,
    - They consume more than 1/3 of final energy use and emit more than 1/3 of global energy-related greenhouse gas emissions,
    - They are energy inefficient and at the same have faced the aging problem,
    - They are at risk of being unable to cope with the negative effects of climate change
- Answering such challenges require sufficient knowledge to benefit energy and resource efficiency at reduced impacts to environment
- This use case aims at developing models based on data compliant with F.A.I.R-data definition to estimate (a) material use intensity and energy performance and (b) associated greenhouse gas emissions of building stocks (LOD2 or above)
- The work uses machine-learning models, energy modelling and life cycle assessment to estimate circularity and energy performance and in-service materials at a neighbourhood level

DOI: 10.1016/j.compenvurbsys.2016.04.005

Many thanks for your attention

Finansiert av
Den europeiske union

FAIRiCUBE

This project has received funding from the European Commission under Horizon Europe grant No 101059238.

Learn more here:

www.fairicube.eu / #FAIRiCUBE / FAIRiCUBE@ LinkedIn

SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

What are the main bottlenecks of using Earth Observation data for data science projects in your organization?

a. Data Access
b. Data licenses
c. Scalability of workflows
d. Data formats
e. Data not analysis ready (contains clouds, gaps, etc.)
f. Too many EO systems
g. Complexity of produced Machine Learning knowledge
h. Computing complexity
i. Ethics or GDPR issues

The ENVISION Data Cube for Big Satellite Image Time-Series to Support AI based on National-scale agriculture monitoring

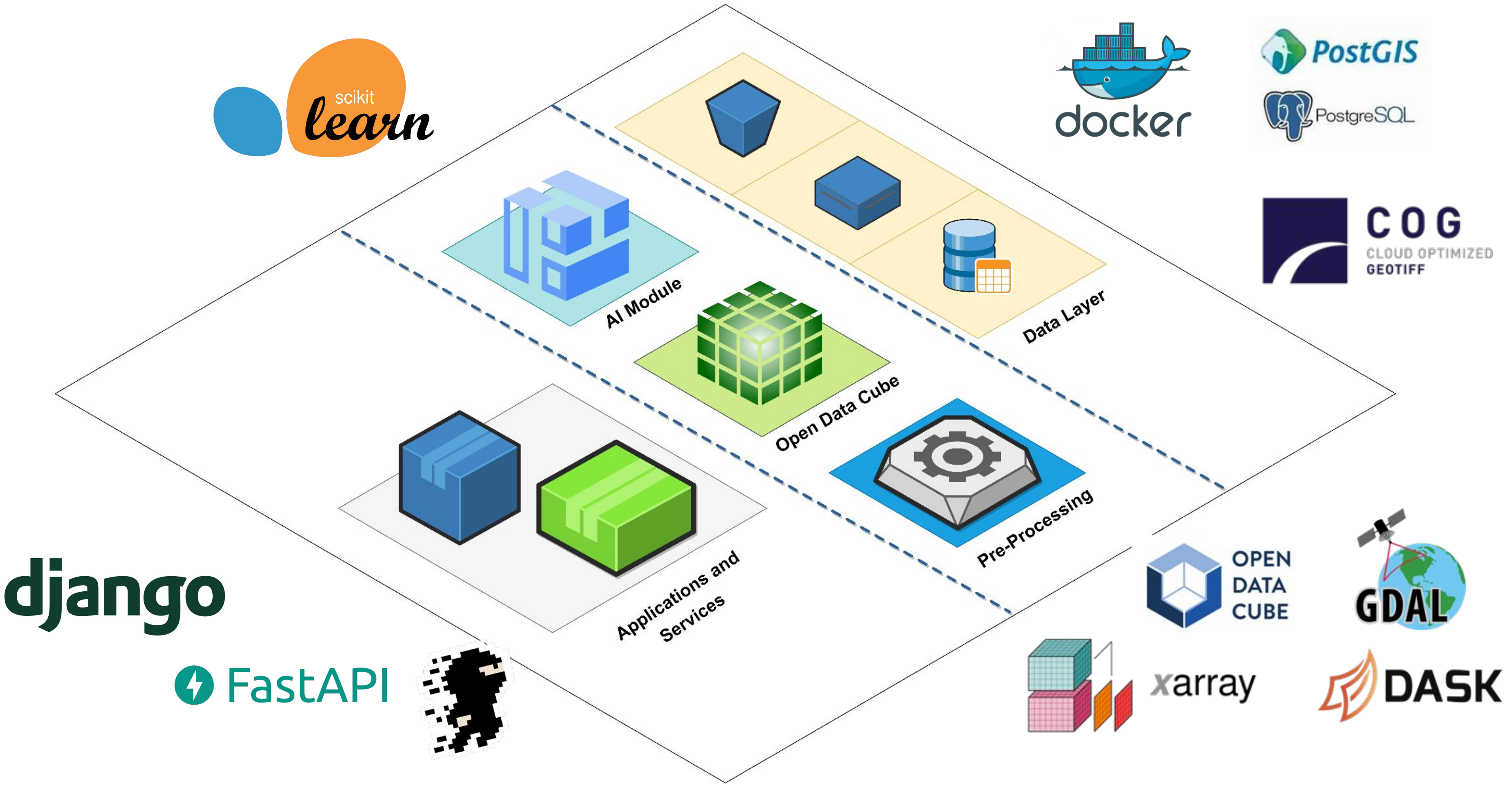
Thanassis Drivas

NOA

The ENVISION Data Cube

- The **modernization of the Common Agricultural Policy** (CAP) requires the **large scale and frequent monitoring** of agricultural land.
- In that direction, we have developed the Agriculture monitoring Data Cube (ADC), which is an **automated**, **modular**, **end-to-end framework** for **discovering**, **pre-processing** and **indexing** optical and Synthetic Aperture Radar (SAR) images into a multidimensional cube
- Currently, Data Cube hosts **3 years of data for both Cyprus and Lithuania at national scale** (~10TB) and is being automatically populated by new data
- Functionalities and services are build **on-the-top of this cube** open the door to the effortless and reliable **decision-making**

The ENVISION Data Cube - High Level Architecture



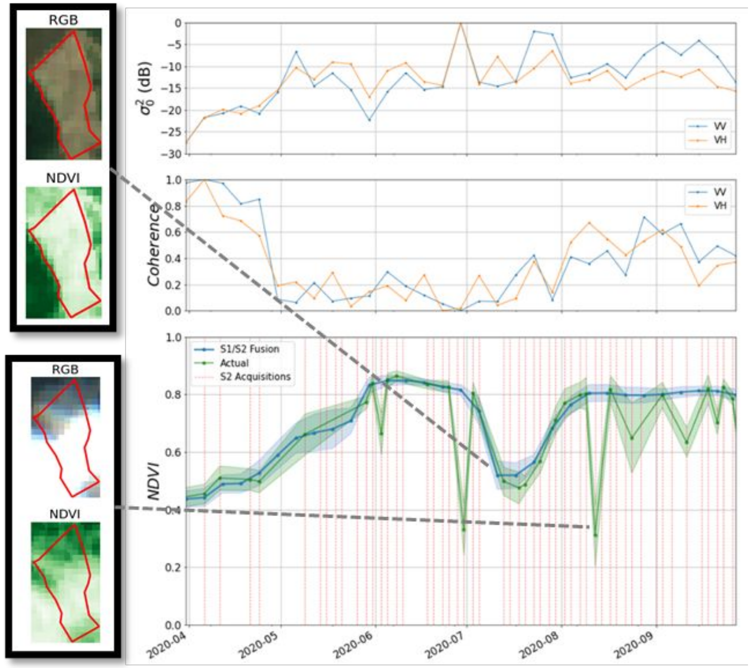
The ENVISION Data Cube - Main Modules

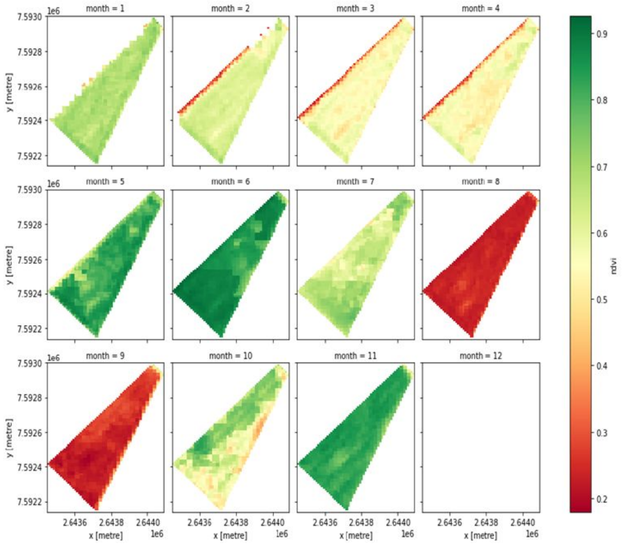
1. **Data Discovery** using cron jobs to search for new ingested products into eodata catalogue of CreoDIAS
2. **Data Preprocessing**
   a. Generation of **Sentinel-1 Backscatter** and **Coherence** using **SNAPPY**.
   b. Generation of resampled and reprojected **Sentinel-2 COGs** along with cloud masks using **Sen2cor** and **FMask**.
3. **Automated indexing** of new products into the cube
4. Metadata of products and status of each step is written into a **geospatial database**.

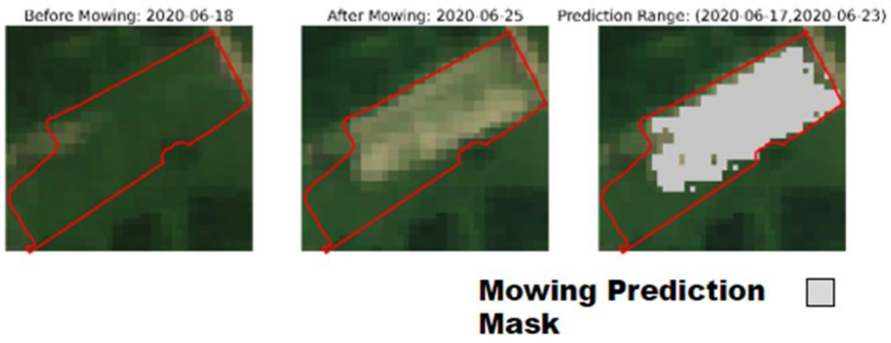
CREODIAS
powered by CloudFerro

Current usage of ENVISION Data Cube: High Level

1. **Cloud-free Mosaics**: Median, Geomedian etc.
2. **Indices**: NDVI, NDMI, NBR, NDWI, SAVI, dNDVI etc.
3. **Feature space creation** for crop classification
4. Monitoring **soil erosion**
5. **Grassland mowing events** detection
6. Identification of **stubble burning**
7. **Analytics**





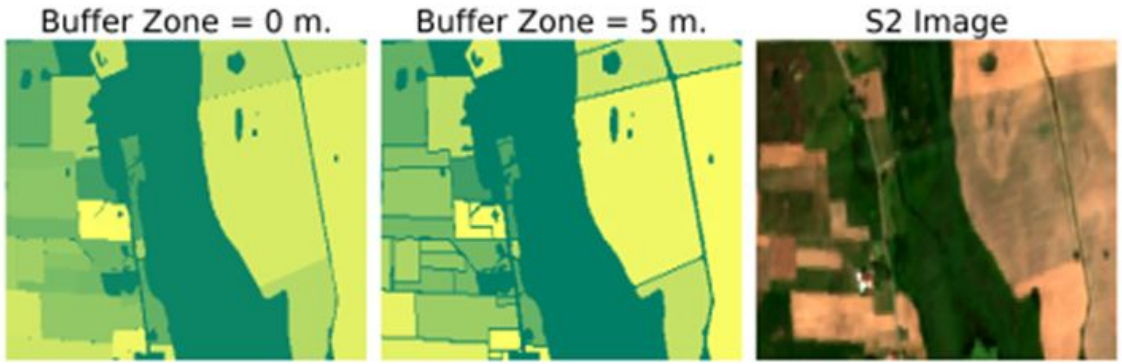


Visualization can be extracted for any s1/s2 product, any aggregated statistic measurement, either on the pixel or on the parcel level and for a specific parcel, a crop type or a crop family. **It plays a crucial role to the validation of results generated by the rest of back- end processes**

Current usage of ENVISION Data Cube: High Level

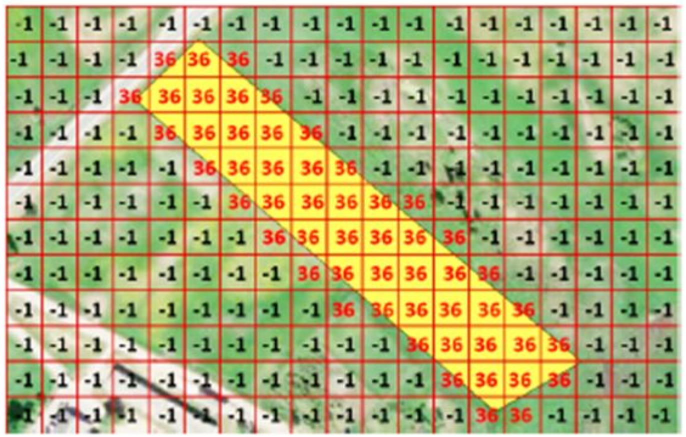
Current usage of ENVISION Data Cube: Spatial Buffers

1. As **LPIS** of each country and year **is rasterized and indexed** into the Data Cube, one can use our **inward** buffer functionality which can alleviate the adverse consequences of **mixed pixels**.
2. This way, we can extract only the representative information from the rest of the pixels encompassed inside the parcels' buffered geometries
3. The Sen2Cor scene classification product has s**uboptimal recall for the cloud and shadow** classes.We additionally provide an adjustable **outward** buffer that can be applied on the less-than-perfect cloud and shadow mask products in order to reduce some of the noise.

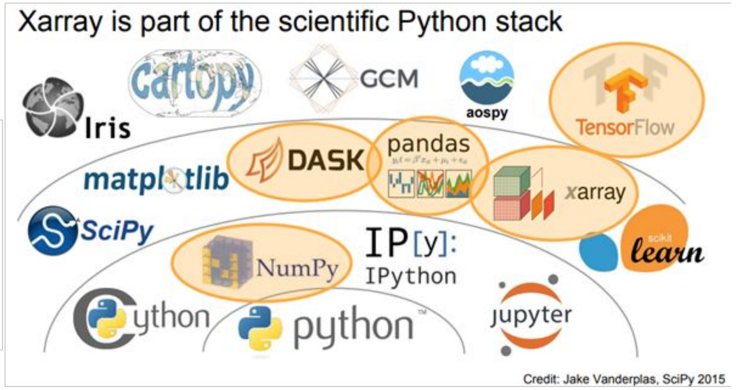


Current usage of ENVISION Data Cube: Zonal Statistics

1. ENVISION demands the statistics computation and the execution of geospatial queries for **millions of parcels.** To address this need **LPIS is rasterized** and indexed also as an additional product in ENVISION Data Cube
2. This addition enables the **fast and parallel computation** of zonal statistics per parcel using the **groupby** function, which groups the xarray dimensions **based on the respective IDs**
3. It is observed that for a **large number of parcels, groupby is far more computationally economic** opening the room for the effortless and efficient scaling up.

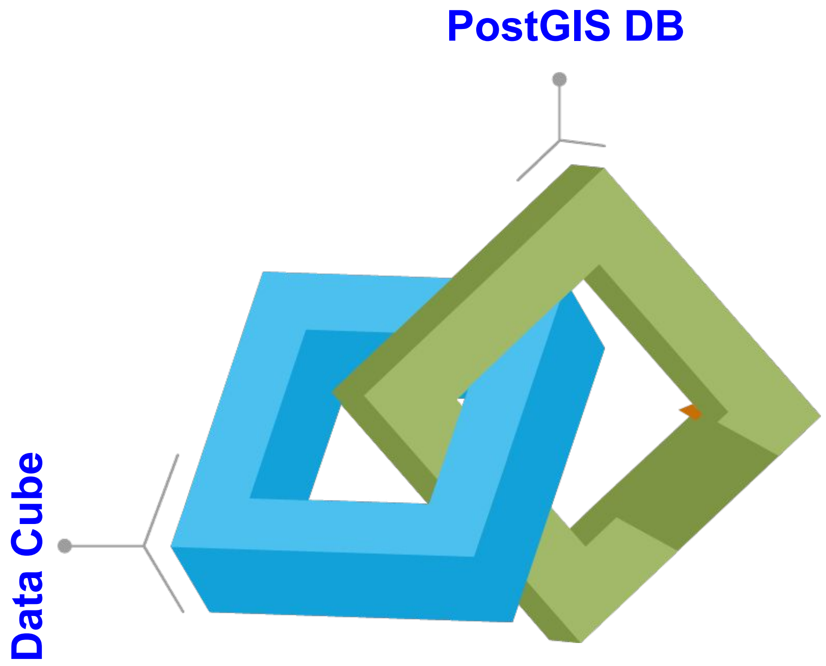


# parcels	groupby	serial querying
1 k	69 sec	250 sec
10 k	71 sec	40 min
100 k	150 sec	400 min

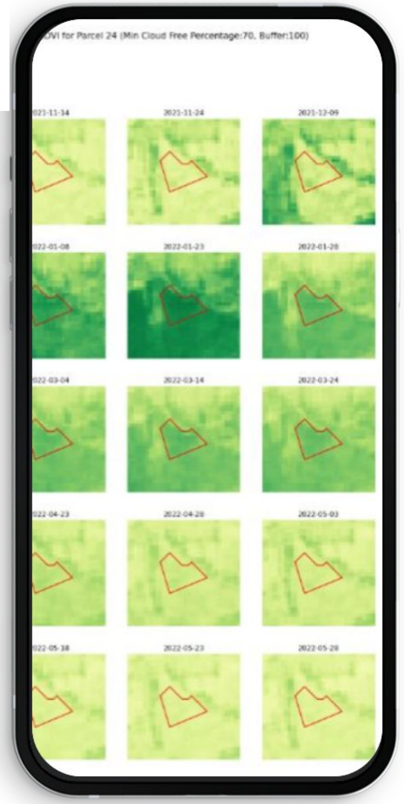


Scalable country-specific knowledge

1. The **results of the downstream services** that Data Cube supports are used to **update a PostgreSQL/PostGIS database.**
2. The database contains **aggregated results per parcel**, which can then be easily accessed, enabling a **back-and-forth communication with the Data Cube**.
3. Thus, we have a Data Cube that includes and keep on being **dynamically populated** by Sentinel-1 and Sentinel-2 products.
4. The Cube also hosts **auxiliary geospatial data** (e.g., LPIS) that are used to enable the provision **high level data products** (e.g., crop classification) **that in turn populate the cubes**.
5. This way, we end up with **country specific knowledge bases for CAP monitoring**



On-demand Access to Data



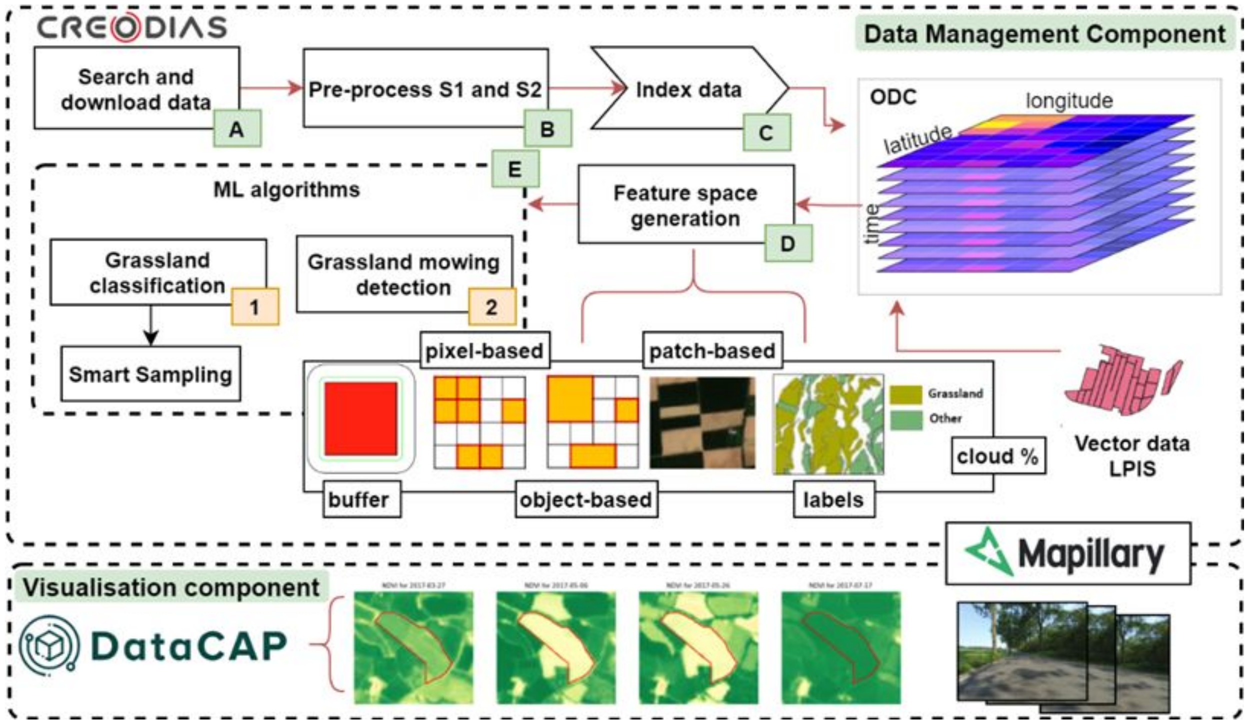
- **ENVISION Data Cube offers two APIs** on top of it.
- The first one accepts requests that include parcel id, time range and vegetation index / band.
    - Once the request is accepted, the back-end system **generates time-series GeoTIFFs** / **plots** and pushes the results to **user's repository.**
- The second API can be exploited by **Data scientists and developers** that may need **directly access to the data** instead of getting images.
    - Data is offered via the **xpublish**. Xpublish lets you easily publish Xarray Datasets via a REST API.
    - Under the hood, Xpublish is using **FastAPI** and **Dask.**

Data or/and products are transferred to organisations' internal systems and thus decision-makers have additional tools for enhancing the validation process.

NPA and CAPO use these APIs

DataCAP: A web app connected to the Data Cube

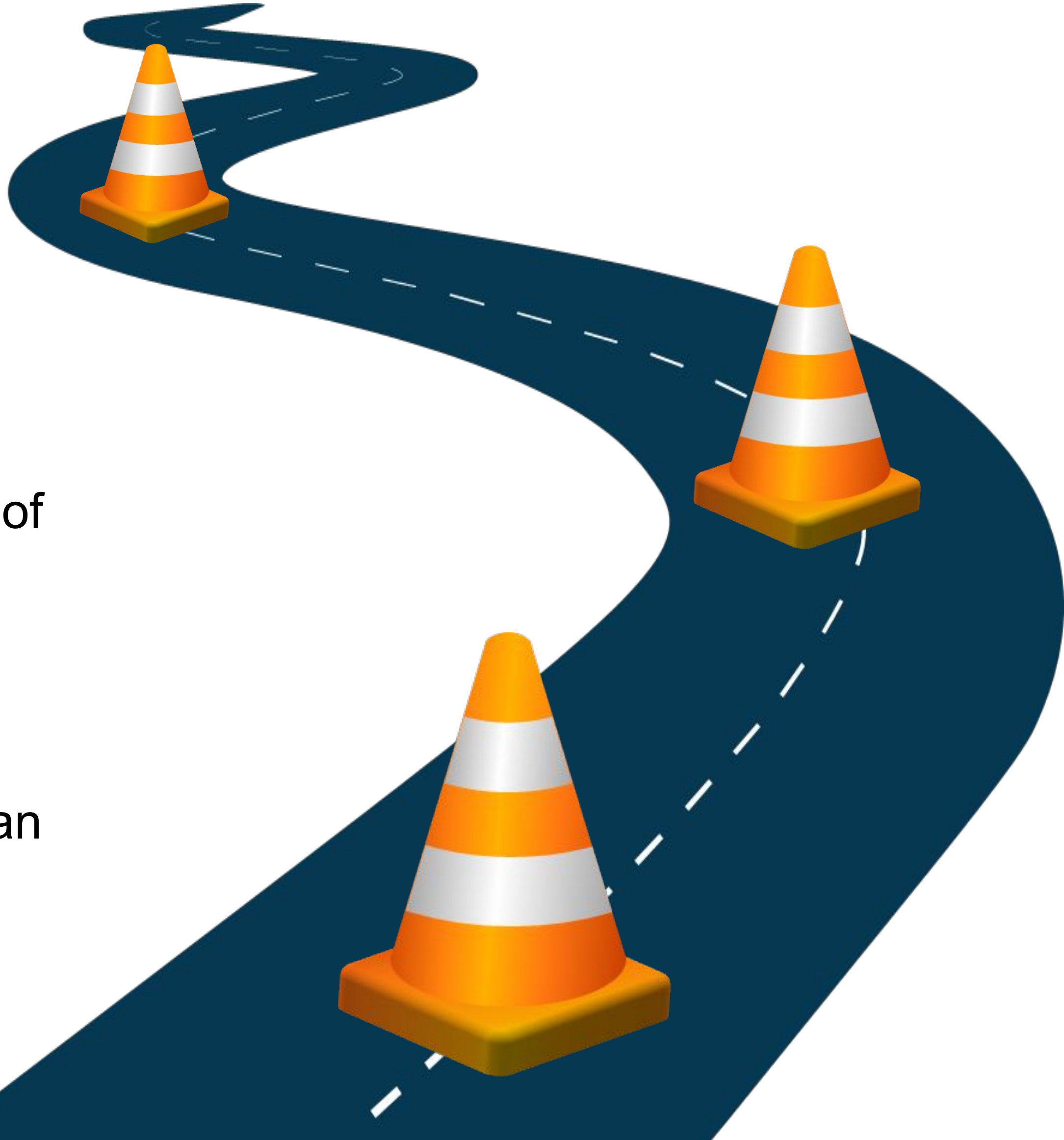
- Having data and parcels indexed in the Data Cube, opens the room from **various applications** to be implemented.
- A **Django web application** has been developed serving **crop classification and grassland mowing results** for a series of parcels.
- In addition, it allows **user to enter visualization parameters** so the requested satellite images can be plotted.
- At the same time, **Street Level Images are also visualized** to enhance the inspection of crop classification and grassland mowing results.



http://62.217.82.91/

Lessons Learnt

- Generation of ARD is not a **simple process**
- **Be close to the data** seems a great idea for such kind of applications
- **Data Preprocessing** is a major step before the indexing process
- **Rasterization of shapefiles** improves the time complexity of zonal statistics
- However, the **number of objects** included in the zonal statistics calculation can still remain a **bottleneck**.
- **Xpublish** is an efficient tool for publishing xarray data via an API



SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

Would you be interested in installing the ENVISION tools / services / data cube instance for your work?

a. Yes
b. No
c. Maybe
d. No opinion

HELLENIC REPUBLIC
MINISTRY OF
DEVELOPMENT AND INVESTMENTS
GENERAL SECRETARIAT FOR
RESEARCH AND INNOVATION
European
Commission

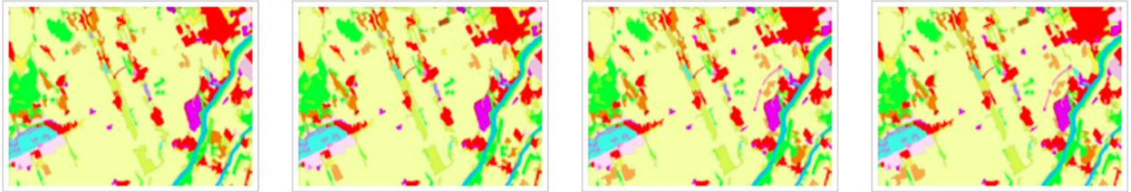
EcoDataCube.eu
Open Environmental Data Cube based on GLAD Landsat and Sentinel-2 L2A imagery

Leandro Parente

OpenGeoHub

Introduction & context

2018-EU-IA-0095



CLC 1990 CLC 2000 CLC 2006 CLC 2012

100 m resolution only
partially harmonized
inter-periods missing

Sentinel 2 time-series
Landsat 7/8 time-series
TANDEM-x topographic data
ALOS radar images

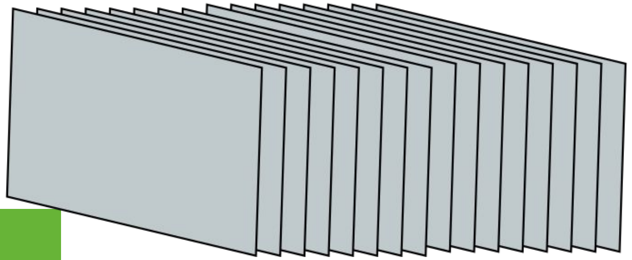
30 m resolution
full space-time coverage

Geo-harmonizer (2019-2022)

cellular automata,
urban growth models...

Model training
(Machine Learning based)

Spatiotemporal model of land cover dynamics



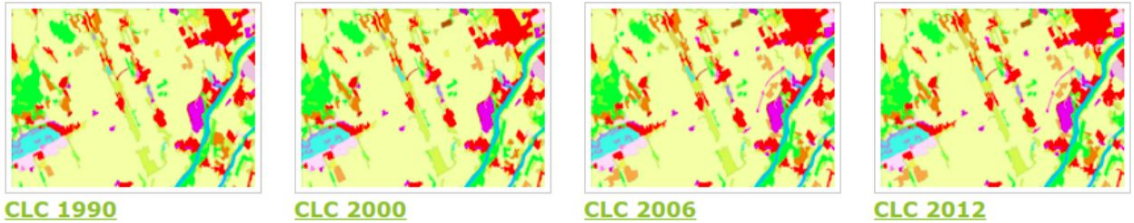
Complete, harmonized
time-series of land cover images
2000, 2001, 2002, ... 2020

30 m resolution
full space-time coverage
fully harmonized / seamless

Help improve **land cover mapping** (CLC) by automate predictions at higher spatial resolution (10—30m)

Introduction & context

2018-EU-IA-0095



Sentinel 2 time-series
Landsat 7/8 time-series
TANDEM-x topographic data
ALOS radar images

Geo-harmonizer (2019-2022)

100 m resolution only
partially harmonized
inter-periods missing

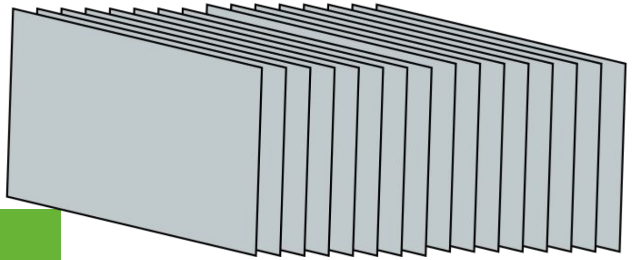
cellular automata,
urban growth models...

30 m resolution
full space-time coverage

Model training
(Machine Learning based)

Spatiotemporal model of land cover dynamics

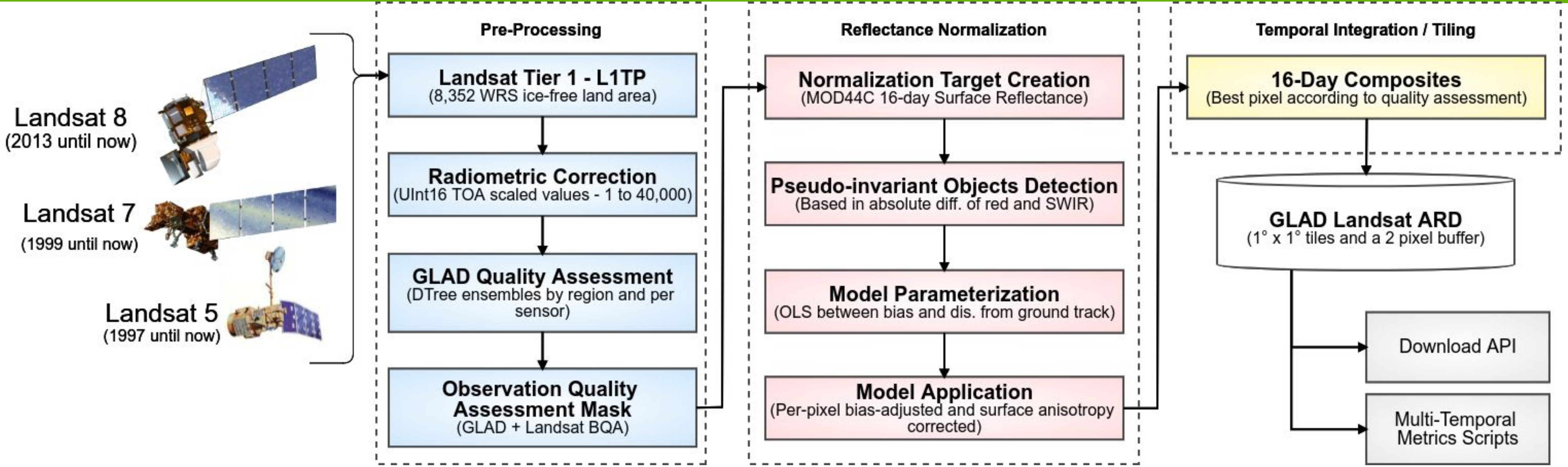
Complete, harmonized
time-series of land cover images
2000, 2001, 2002, ... 2020



30 m resolution
full space-time coverage
fully harmonized / seamless

Help improve **land cover mapping** (CLC) by automate predictions at higher spatial resolution (10—30m)

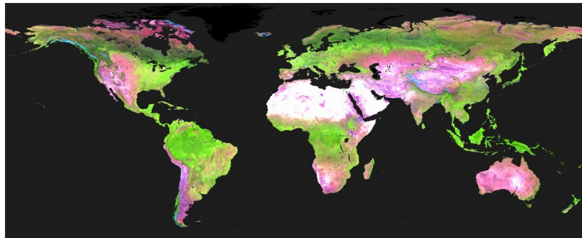
GLAD Landsat ARD



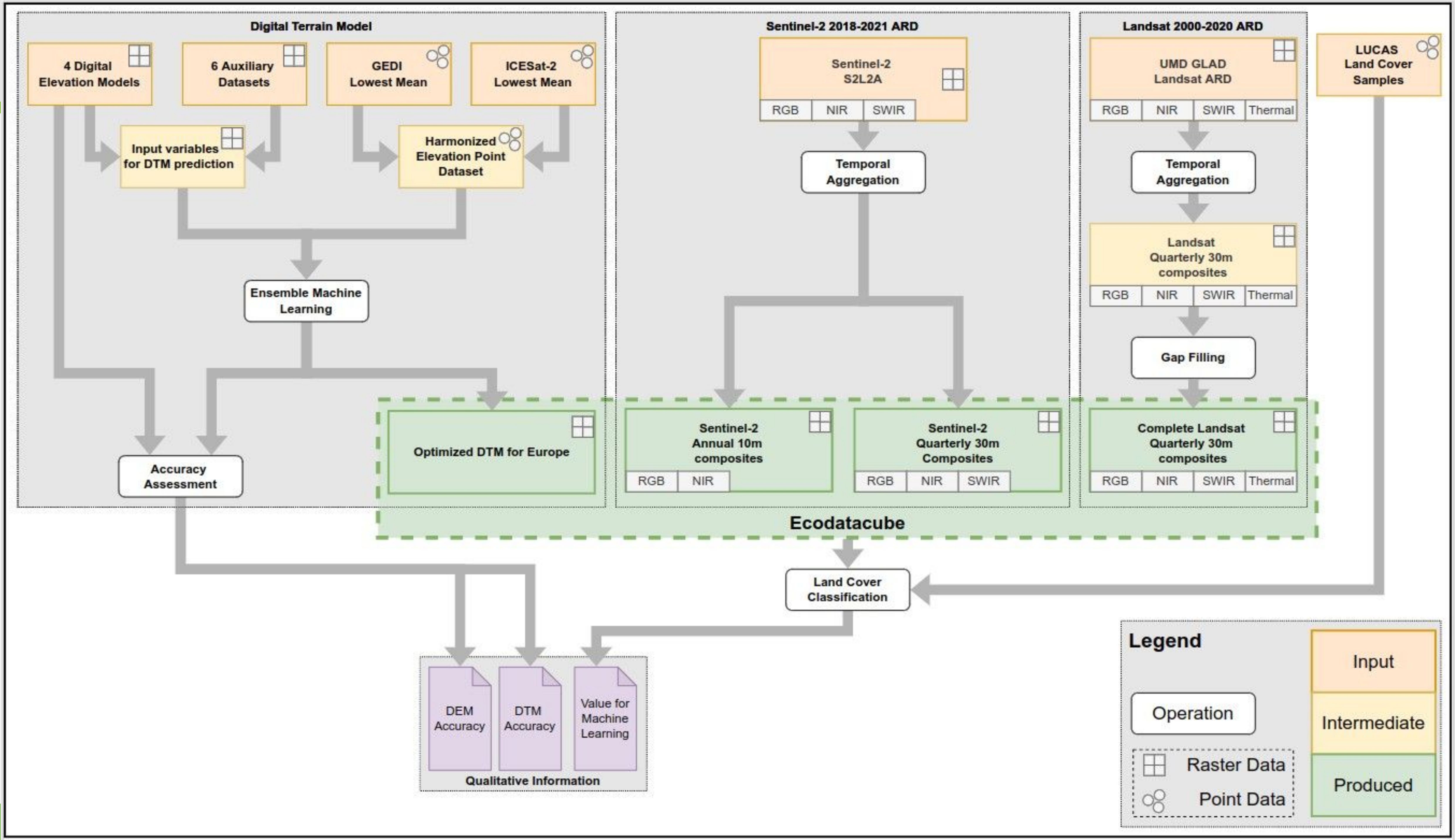
Limitations - It's not suitable for:

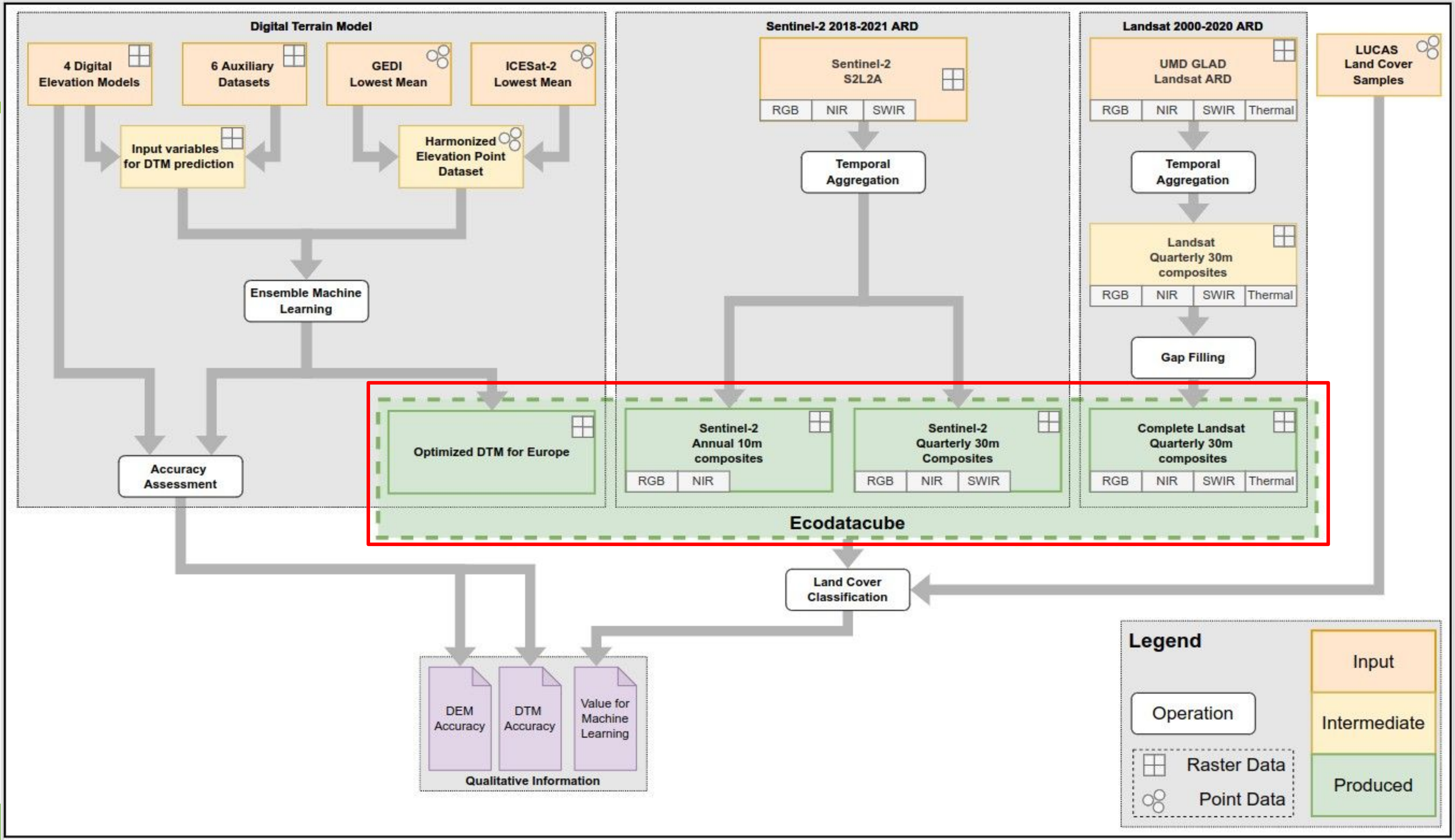
- Real-time land cover monitoring (1-month delay)
- Winter time image processing above 30N and below 45S Latitude
- Precise analysis of land surface reflectance
- Water quality assessment or any other hydrology applications beyond surface water extent mapping

GLAD Global Land Analysis & Discovery



http://dx.doi.org/10.3390/rs12030426


Digital Terrain Model
4 Digital Elevation Models
6 Auxiliary Datasets
GEDI Lowest Mean
ICESat-2 Lowest Mean
Input variables for DTM prediction
Harmonized Elevation Point Dataset
Ensemble Machine Learning
Accuracy Assessment
Optimized DTM for Europe
Sentinel-2 2018-2021 ARD
Sentinel-2 S2L2A
RGB
NIR
SWIR
Temporal Aggregation
Sentinel-2 Annual 10m composites
RGB
NIR
Sentinel-2 Quarterly 30m Composites
RGB
NIR
SWIR
Landsat 2000-2020 ARD
UMD GLAD Landsat ARD
RGB
NIR
SWIR
Thermal
Temporal Aggregation
Landsat Quarterly 30m composites
RGB
NIR
SWIR
Thermal
Gap Filling
Complete Landsat Quarterly 30m composites
RGB
NIR
SWIR
Thermal
LUCAS Land Cover Samples
Ecodatacube
Land Cover Classification
DEM Accuracy
DTM Accuracy
Value for Machine Learning
Qualitative Information
Legend
Input
Operation
Intermediate
Raster Data
Point Data
Produced


Digital Terrain Model
4 Digital Elevation Models
6 Auxiliary Datasets
GEDI Lowest Mean
ICESat-2 Lowest Mean
Input variables for DTM prediction
Harmonized Elevation Point Dataset
Ensemble Machine Learning
Accuracy Assessment
Sentinel-2 2018-2021 ARD
Sentinel-2 S2L2A
RGB
NIR
SWIR
Temporal Aggregation
Landsat 2000-2020 ARD
UMD GLAD Landsat ARD
RGB
NIR
SWIR
Thermal
Temporal Aggregation
Landsat Quarterly 30m composites
RGB
NIR
SWIR
Thermal
Gap Filling
LUCAS Land Cover Samples
Optimized DTM for Europe
Sentinel-2 Annual 10m composites
RGB
NIR
Sentinel-2 Quarterly 30m Composites
RGB
NIR
SWIR
Complete Landsat Quarterly 30m composites
RGB
NIR
SWIR
Thermal
Ecodatacube
Land Cover Classification
DEM Accuracy
DTM Accuracy
Value for Machine Learning
Qualitative Information
Legend
Input
Operation
Intermediate
Raster Data
Point Data
Produced

Landsat gapfilling

Time span
2000 — 2020

Q1: Dec. 2 — Mar. 20

Q2: Mar. 21 — Jun. 24

Q3: Jun. 25 — Sep. 12

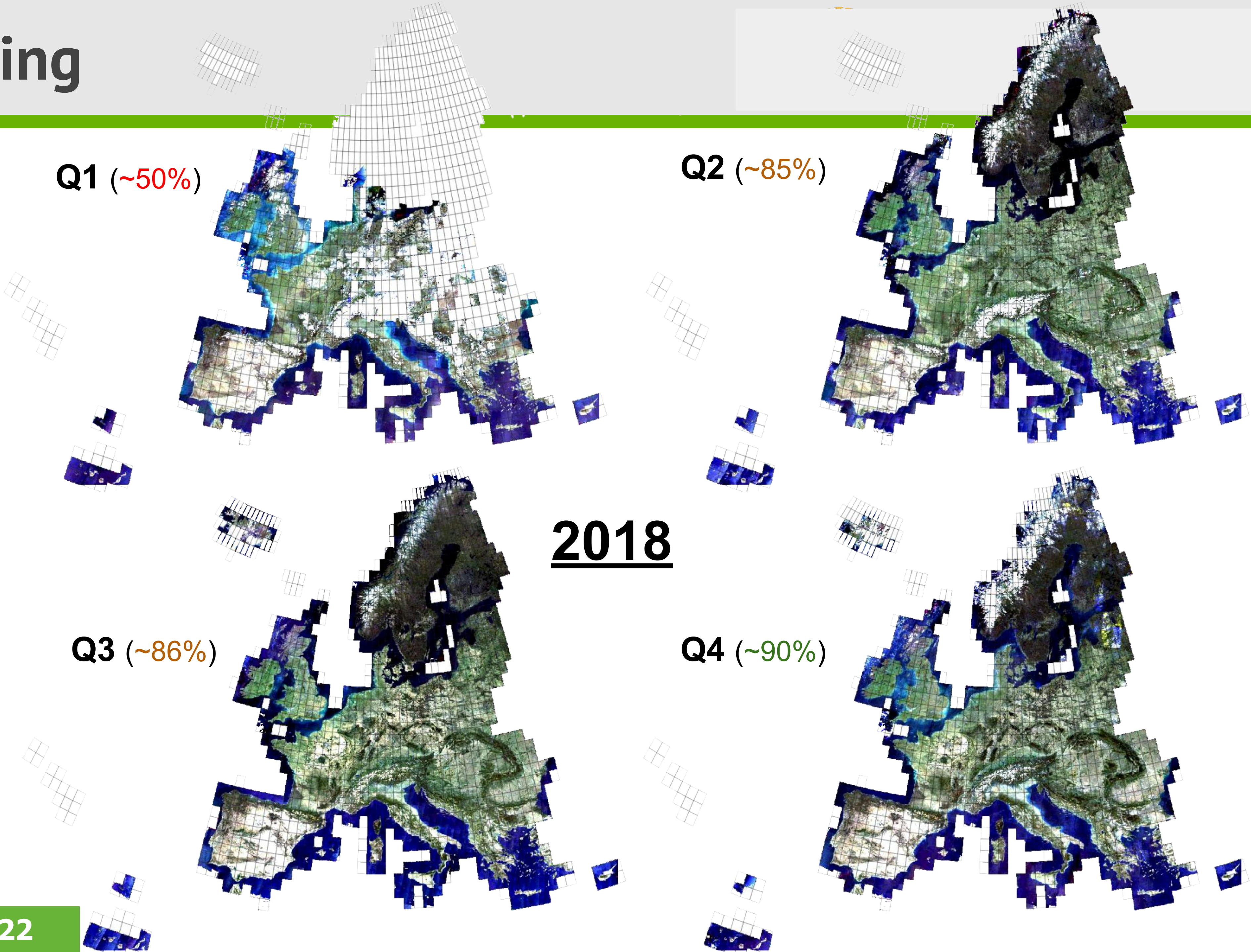
Q4: Sep. 13 — Dec. 1

↓

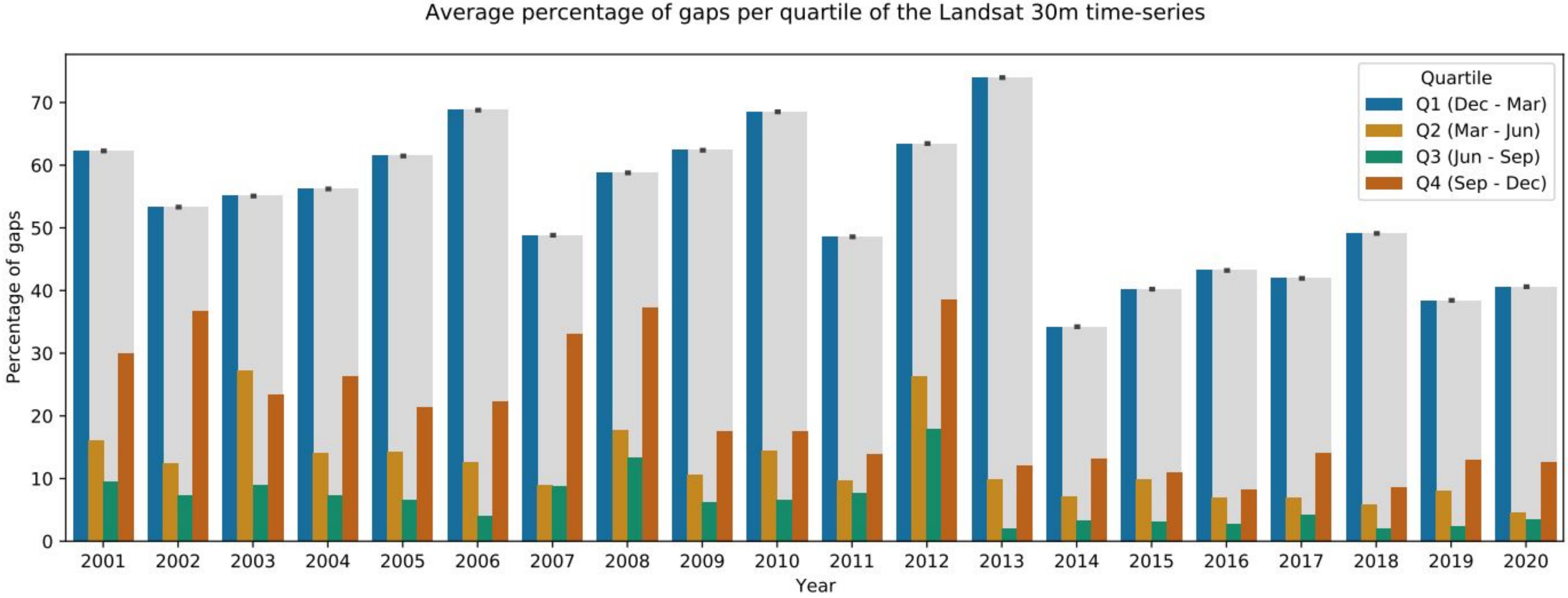
84
time steps / layers

↓

5,2 millions pixels
for each layer



Landsat gapfilling

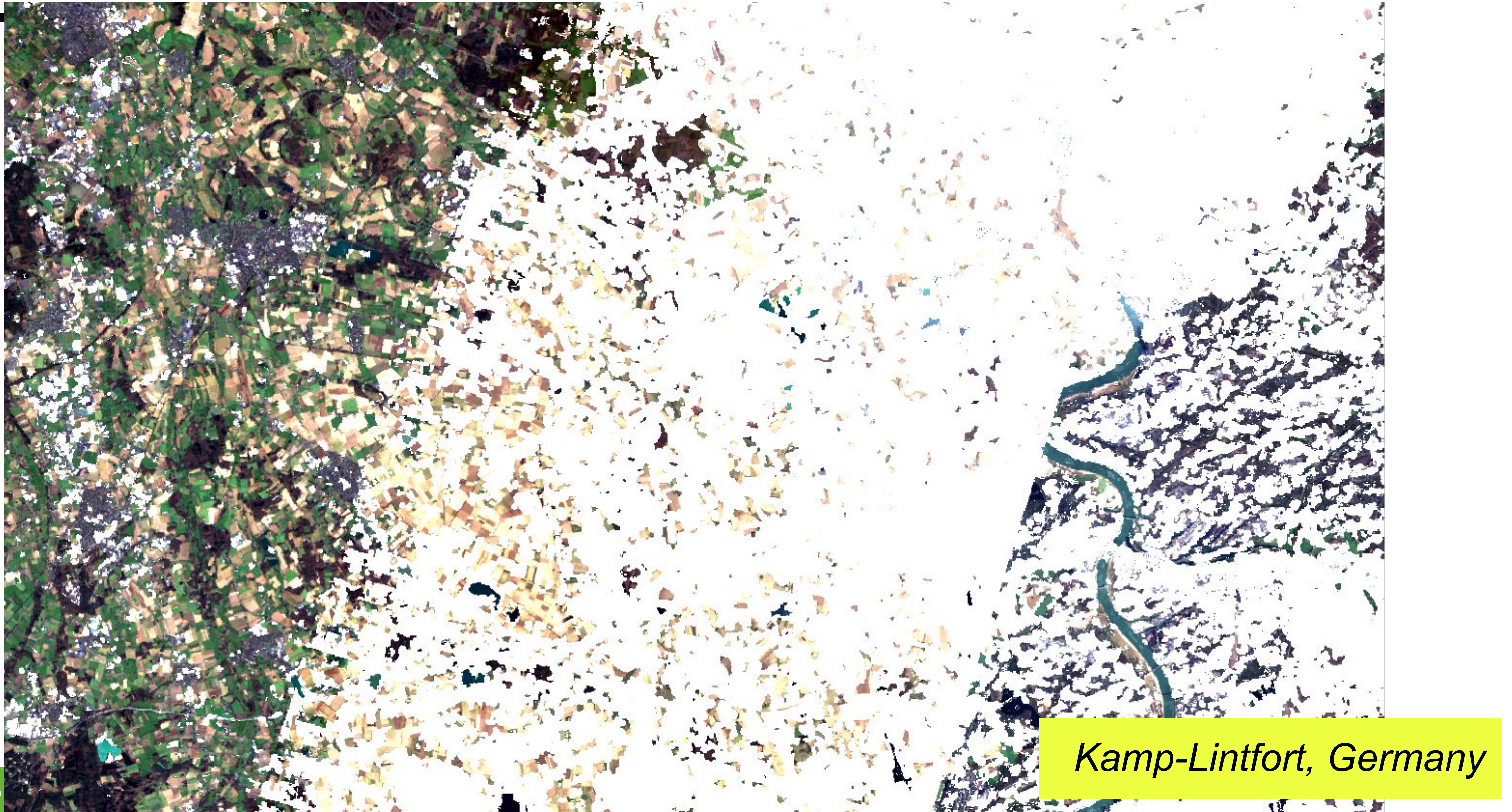


EUMAP LIBRARY

Temporal Moving Window Median (TMWM)

https://eumap.readthedocs.io/en/latest/_autosummary/eumap.gapfiller.TMWM.htm

Landsat gapfilling - Dec. 2 — Mar. 20 / 2018



Landsat gapfilling - Dec. 2 – Mar. 20 / 2018

Kamp-Lintfort, Germany

Landsat and Sentinel data cube (consistent & gapfilled)

30m resolution almost 400GB of analysis-ready **Sentinel-2** data: P25, P50 and P75 for all bands for 2016 to 2021.

30m resolution **Landsat** ARD 10TB of data available as COGs via our Wasabi service.

10m resolution **Sentinel-2** mosaics (120GB per image!) are also available via https://EcoDataCube.eu.

Manuscript submitted

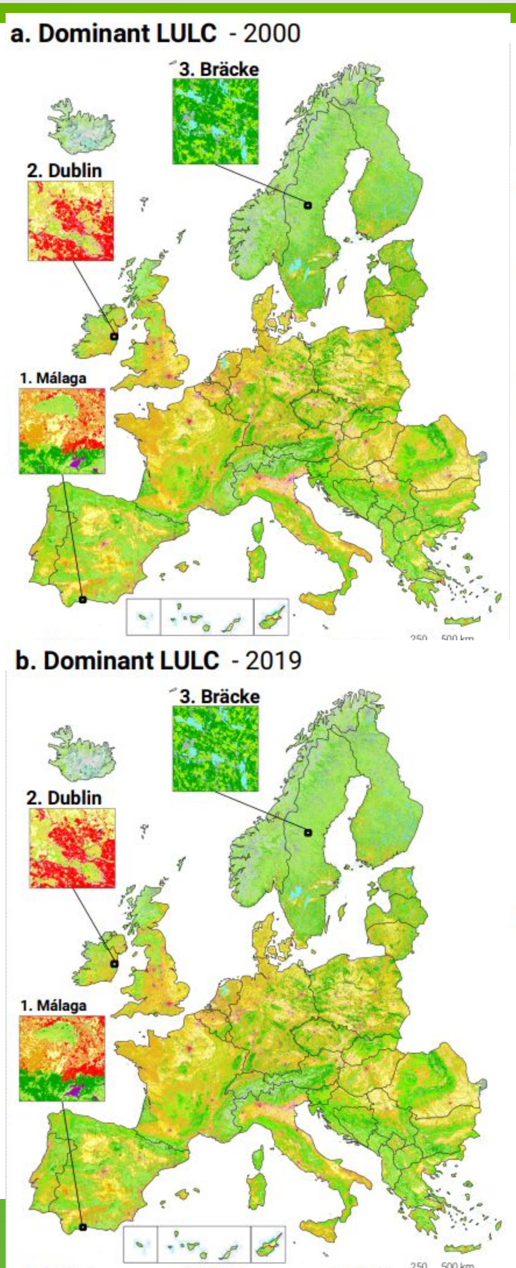
Landsat and Sentinel data cube (consistent & gapfilled)

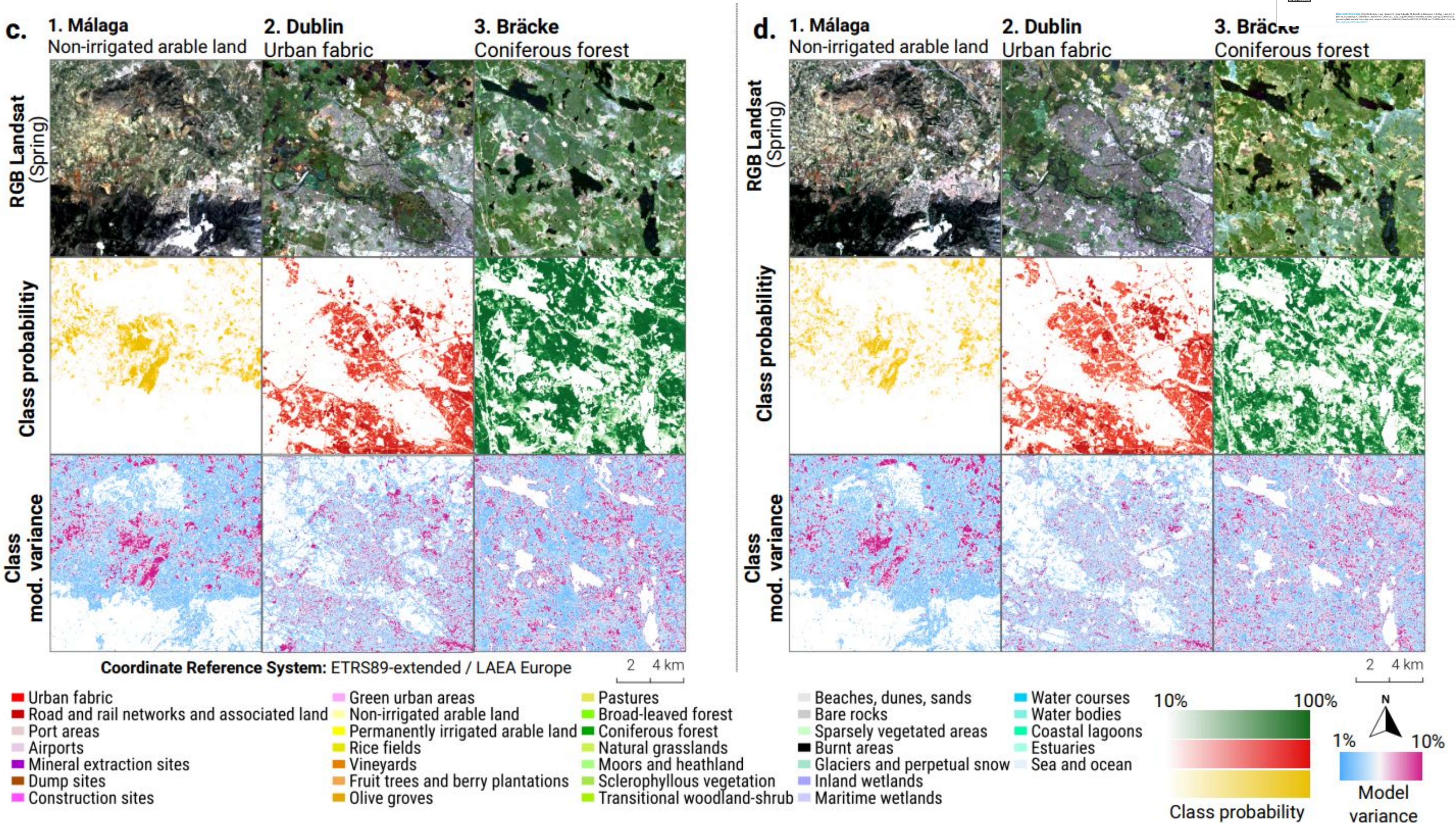
10m resolution mosaics (120GB per image!)

30-m Land use / land cover mapping

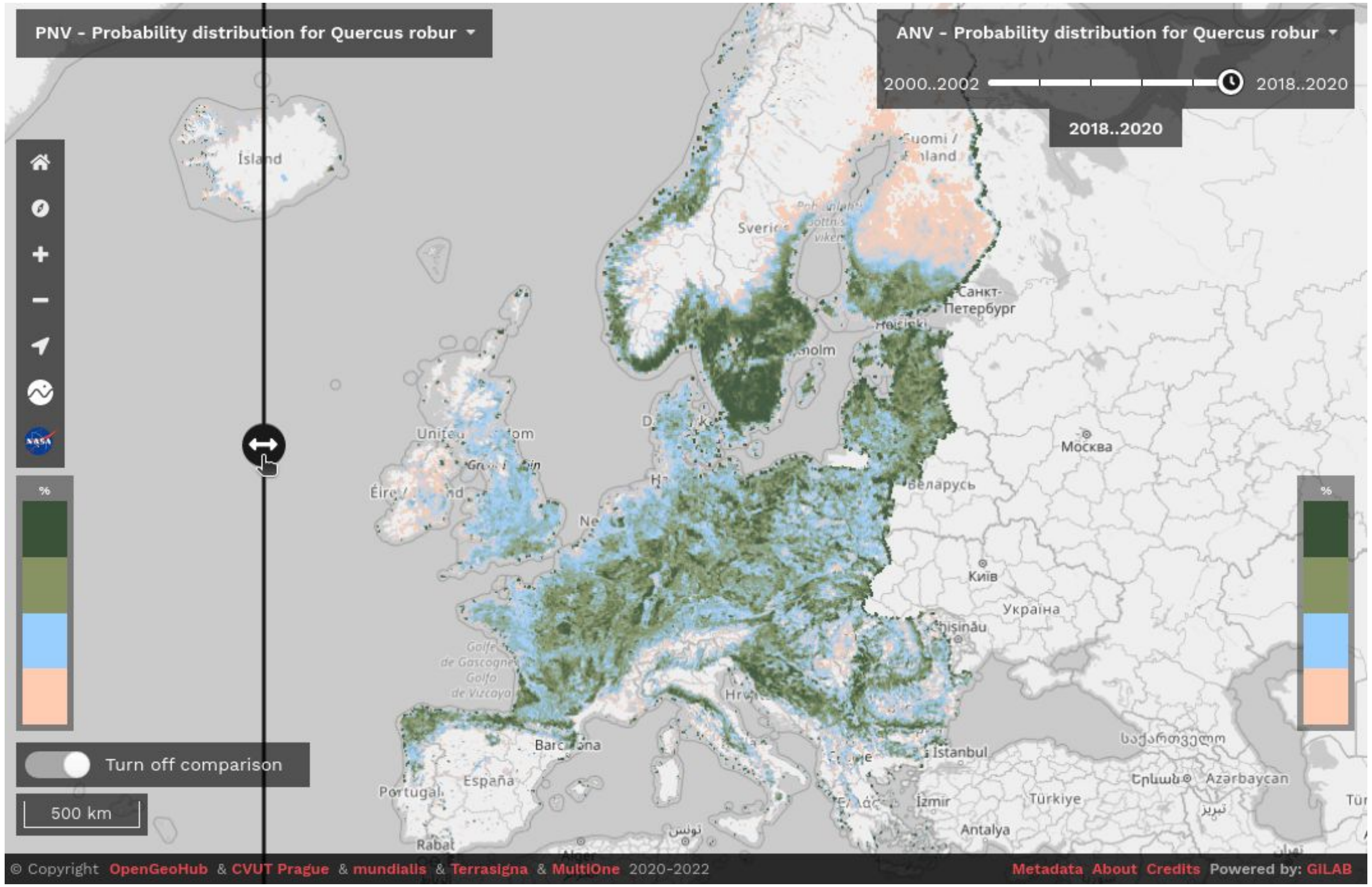
33 CLC classes (2000—2020)

https://doi.org/10.7717/peerj.13573



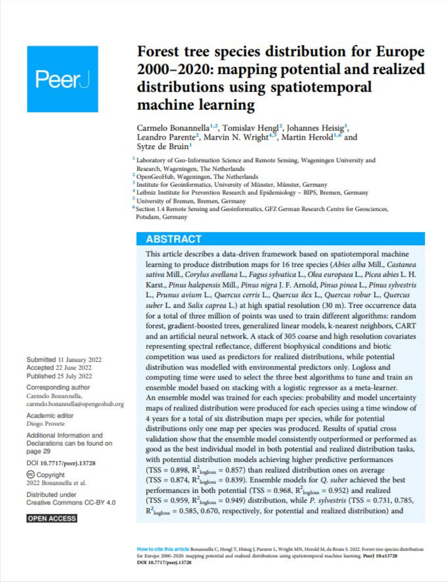


30-m Forest tree species distribution



Quercus Robur (potential and realized)

- **16 species** (<u>potential</u> and <u>realized</u> distribution)
- 2000 - 2020, with time steps:
  - 2000 - 2002
  - 2002 - 2006
  - 2006 - 2010
  - 2010 - 2014
  - 2014 - 2018
  - 2018 - 2020



PeerJ

Forest tree species distribution for Europe 2000–2020: mapping potential and realized distributions using spatiotemporal machine learning

Carmelo Bonannella[1,2], Tomislav Hengl[2], Johannes Heisig[3], Leandro Parente[2], Marvin N. Wright[4,5], Martin Herold[1,6] and Sytze de Bruin[1]

[1] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University and Research, Wageningen, The Netherlands
[2] OpenGeoHub, Wageningen, The Netherlands
[3] Institute for Geoinformatics, University of Münster, Münster, Germany
[4] Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
[5] University of Bremen, Bremen, Germany
[6] Section 1.4 Remote Sensing and Geoinformatics, GFZ German Research Centre for Geosciences, Potsdam, Germany

ABSTRACT

This article describes a data-driven framework based on spatiotemporal machine learning to produce distribution maps for 16 tree species (*Abies alba* Mill., *Castanea sativa* Mill., *Corylus avellana* L., *Fagus sylvatica* L., *Olea europaea* L., *Picea abies* L. H. Karst., *Pinus halepensis* Mill., *Pinus nigra* J. F. Arnold, *Pinus pinea* L., *Pinus sylvestris* L., *Prunus avium* L., *Quercus cerris* L., *Quercus ilex* L., *Quercus robur* L., *Quercus suber* L. and *Salix caprea* L.) at high spatial resolution (30 m). Tree occurrence data for a total of three million of points was used to train different algorithms: random forest, gradient-boosted trees, generalized linear models, k-nearest neighbors, CART and an artificial neural network. A stack of 305 coarse and high resolution covariates representing spectral reflectance, different biophysical conditions and biotic competition was used as predictors for realized distributions, while potential distribution was modelled with environmental predictors only. Logloss and computing time were used to select the three best algorithms to tune and train an ensemble model based on stacking with a logistic regressor as a meta-learner. An ensemble model was trained for each species: probability and model uncertainty maps of realized distribution were produced for each species using a time window of 4 years for a total of six distribution maps per species, while for potential distributions only one map per species was produced. Results of spatial cross validation show that the ensemble model consistently outperformed or performed as good as the best individual model in both potential and realized distribution tasks, with potential distribution models achieving higher predictive performances (TSS = 0.898, $R^2_{logloss}$ = 0.857) than realized distribution ones on average (TSS = 0.874, $R^2_{logloss}$ = 0.839). Ensemble models for *Q. suber* achieved the best performances in both potential (TSS = 0.968, $R^2_{logloss}$ = 0.952) and realized (TSS = 0.959, $R^2_{logloss}$ = 0.949) distribution, while *P. sylvestris* (TSS = 0.731, 0.785, $R^2_{logloss}$ = 0.585, 0.670, respectively, for potential and realized distribution) and

Submitted 11 January 2022
Accepted 22 June 2022
Published 25 July 2022

Corresponding author
Carmelo Bonannella, carmelo.bonannella@opengeohub.org

Academic editor
Diogo Provete

Additional Information and Declarations can be found on page 29

DOI 10.7717/peerj.13728

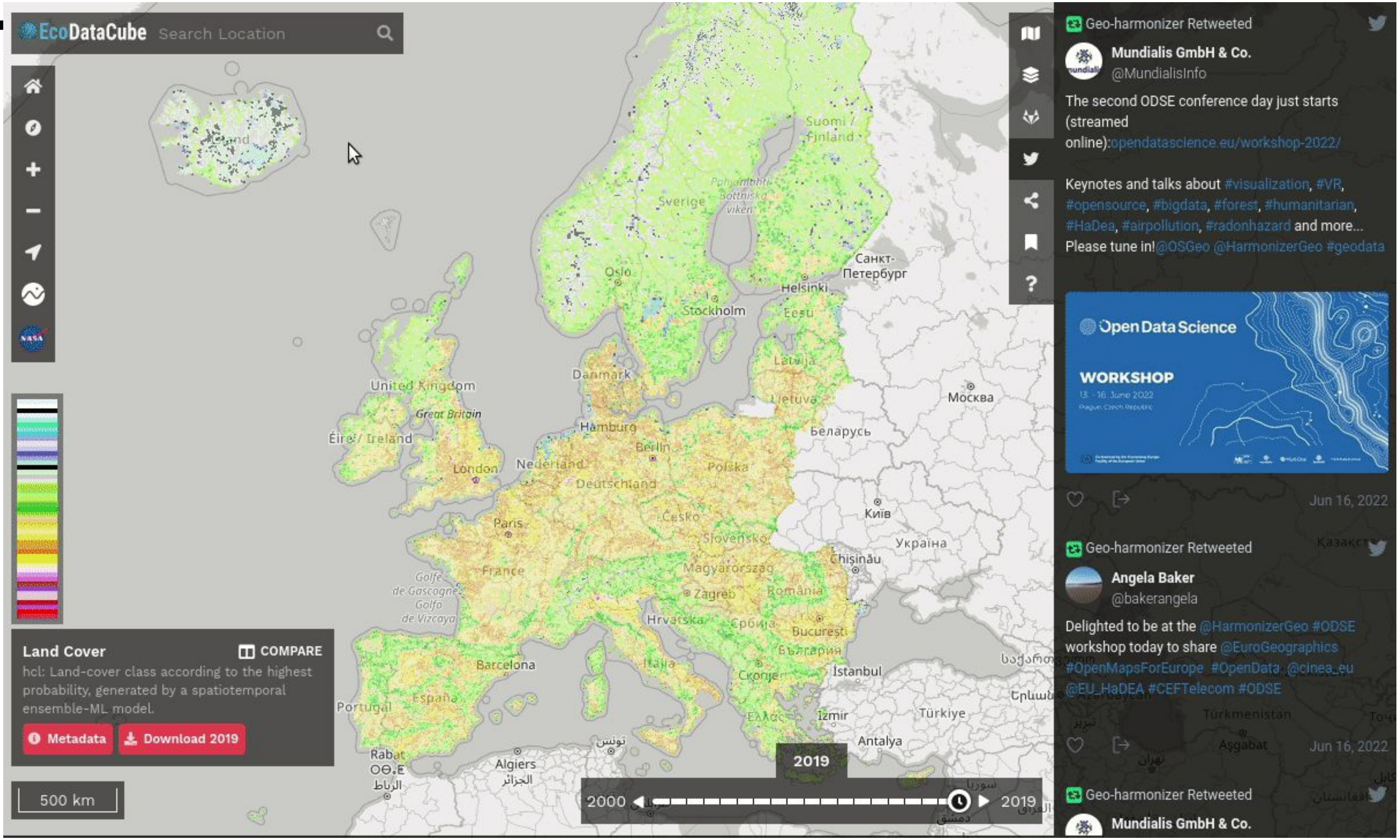
How to cite this article Bonannella C, Hengl T, Heisig J, Parente L, Wright MN, Herold M, de Bruin S. 2022. Forest tree species distribution for Europe 2000–2020: mapping potential and realized distributions using spatiotemporal machine learning. PeerJ 10:e13728
DOI 10.7717/peerj.13728

http://dx.doi.org/10.7717/peerj.13728

<u>Probabilities</u> + *<u>uncertainties</u>*

Data portal https://ecodatacube.eu

Open Geo HUB
Connect • Create • Share • Repeat



Data catalog

A complete comprehensive raster database with all layers imported and documented

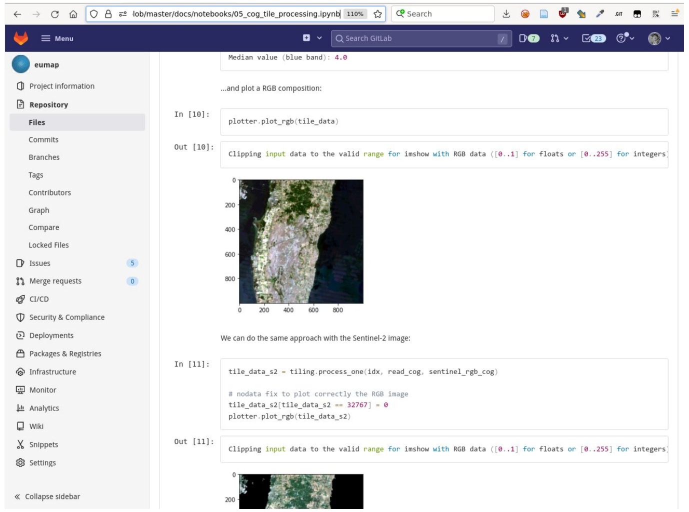
STAC catalog:
https://stac.EcoDataCube.eu

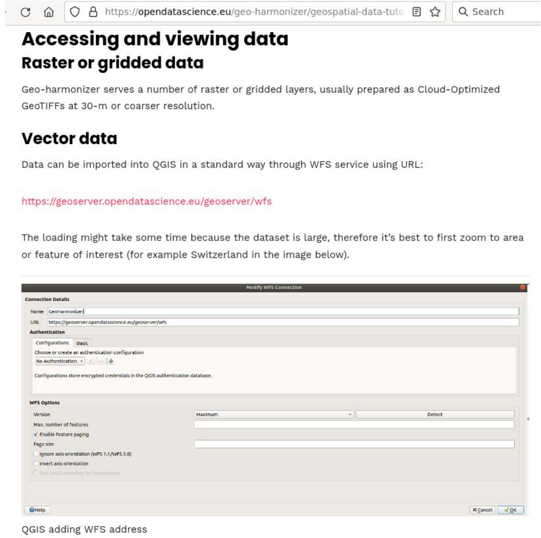
Online tutorials:

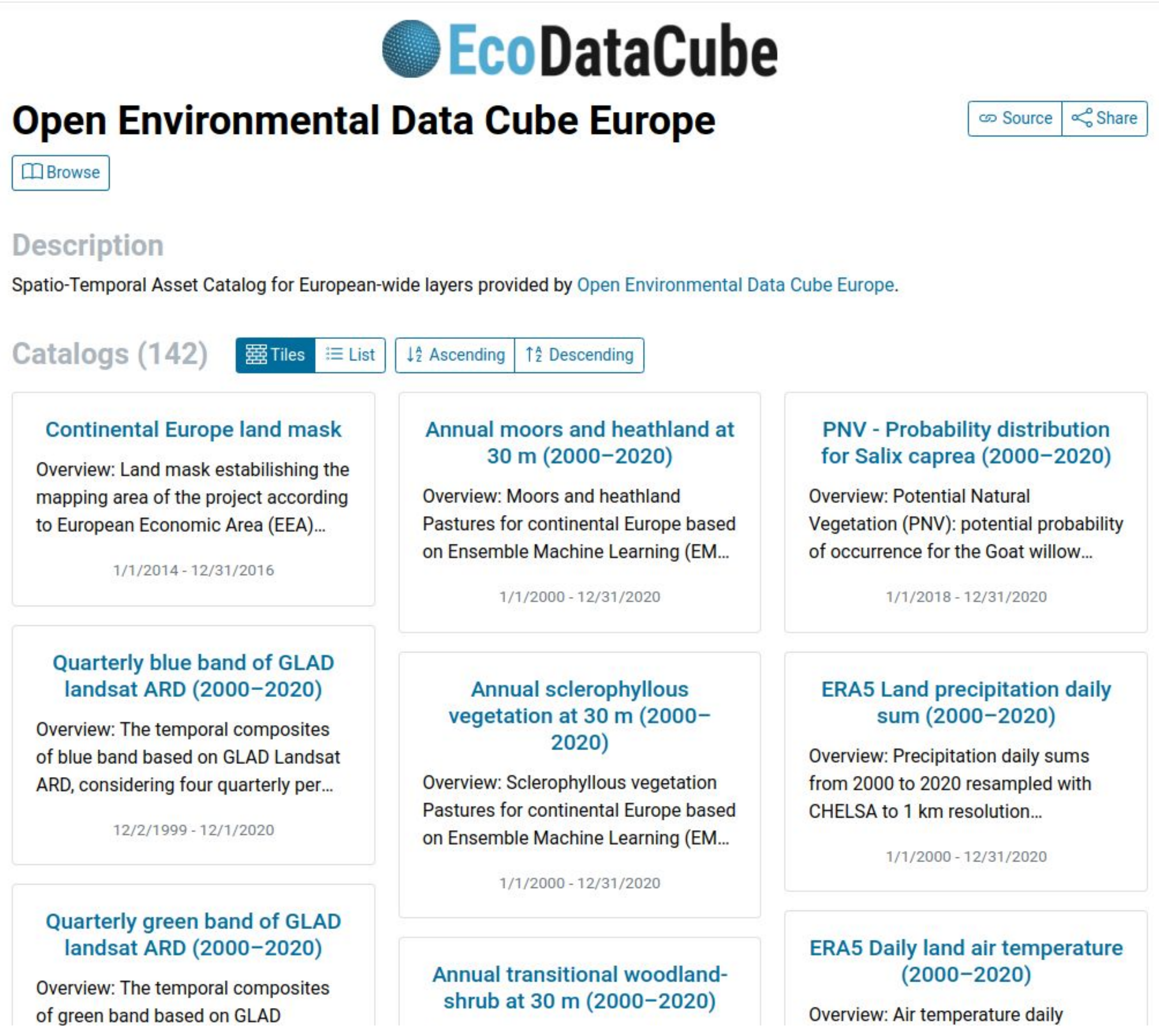
https://opendatascience.eu/geo-harmonizer/geospatial-data-tutorial/

Jupyter notebooks:

https://eumap.readthedocs.io/en/latest/notebooks/10_stac.html







Future of data cubes

- COG + STAC seems to be a robust solution for decentralized (multiarray/EO) data sharing;
- Geoparquet + PMTiles (+Geonetwork) seems to work for sharing point / polygon data;
- **Any cloud-solution you can integrate seamlessly into QGIS is a winner!**
- Still, there is **no free meal**: to overlay thousands of points on remote COG's is still a cumbersome (*e.g.* max number of simultaneous queries is often <10);
- To build your own web-gis best use: Rshiny, xcube viewer, G3W, geemap, leafmap or similar;

(from our perspective)

SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

How do you prefer accessing multiarray data?

a. Point queries, download CSV
b. Download tiles
c. API
d. GDAL
e. R code
f. Python code
g. Zip files

AD4DG:
Datacubes for Biodiversity in e-Shape

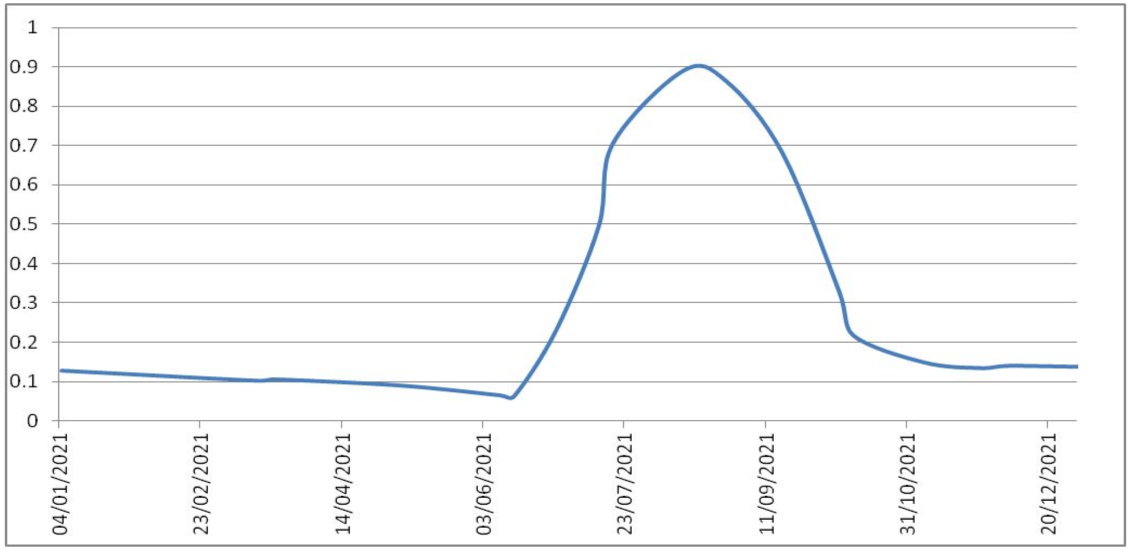
Joan Maso

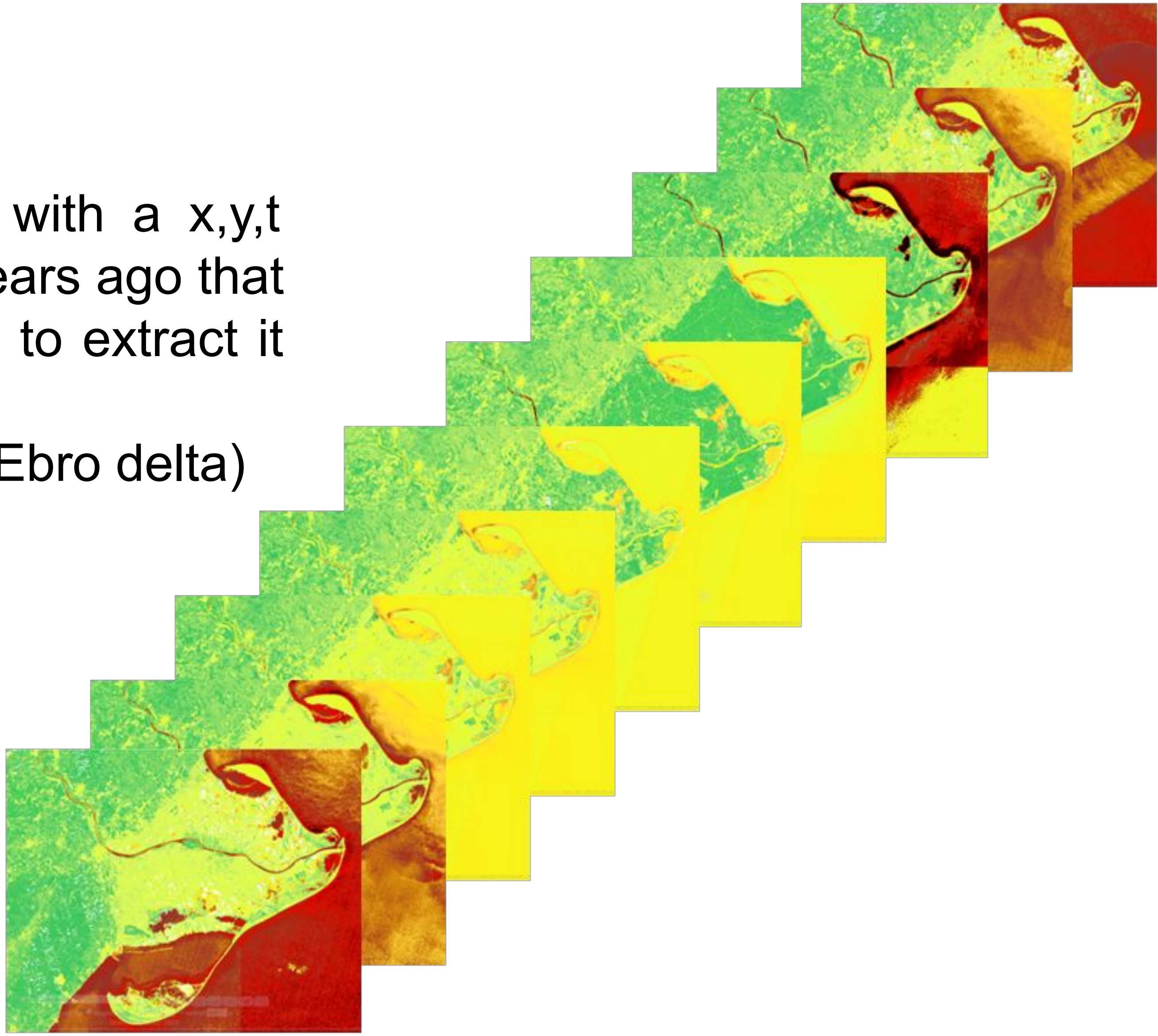
CREAF

e-shape

Pilot: mySPACE

Phenology can be observed and analyzed with a x,y,t datacube representing NDVI values. Some years ago that was done with MODIS but now it is possible to extract it from Sentinel 2.
(example of the dynamics of rice fields in the Ebro delta)

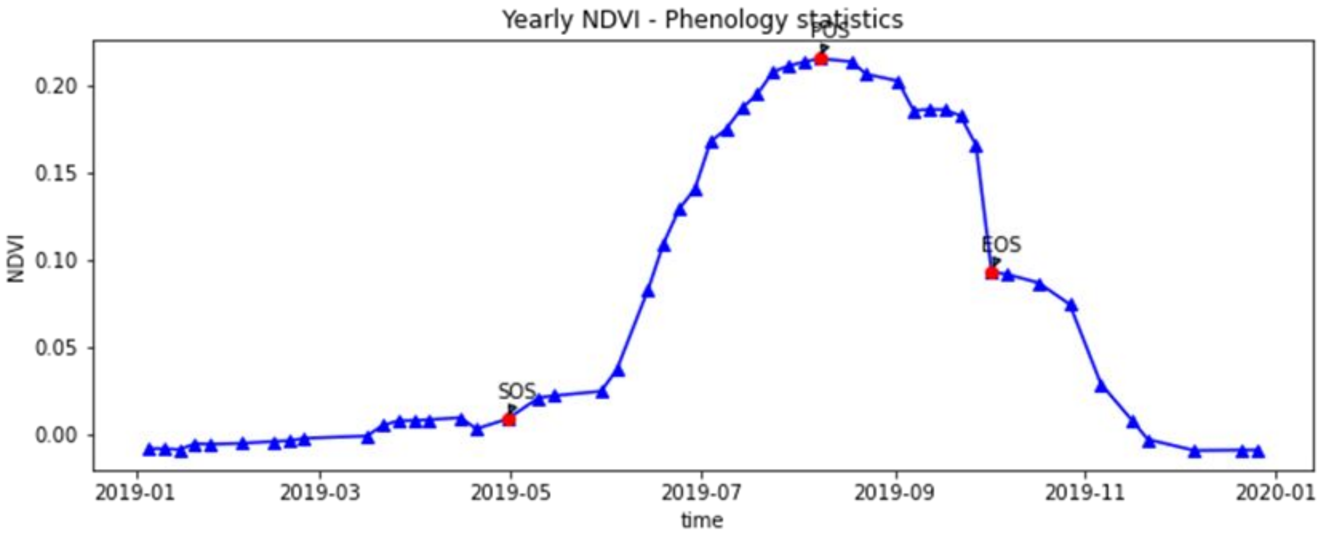


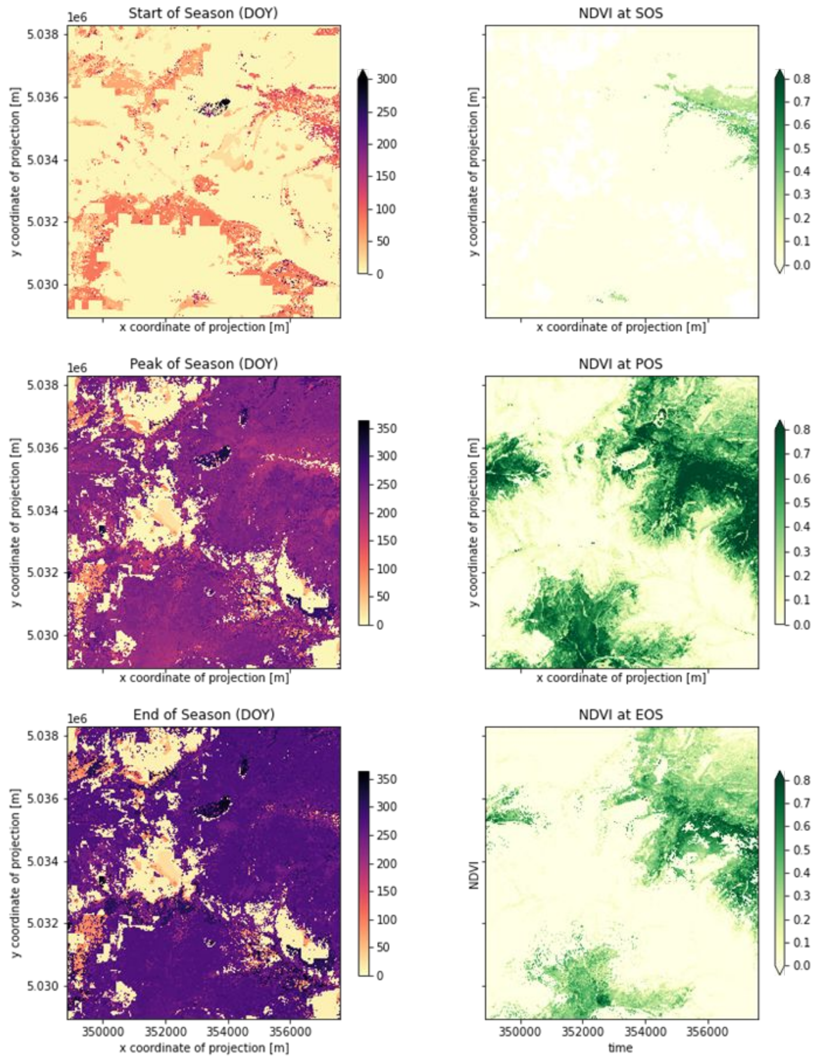


e-shape

Pilot: mySPACE

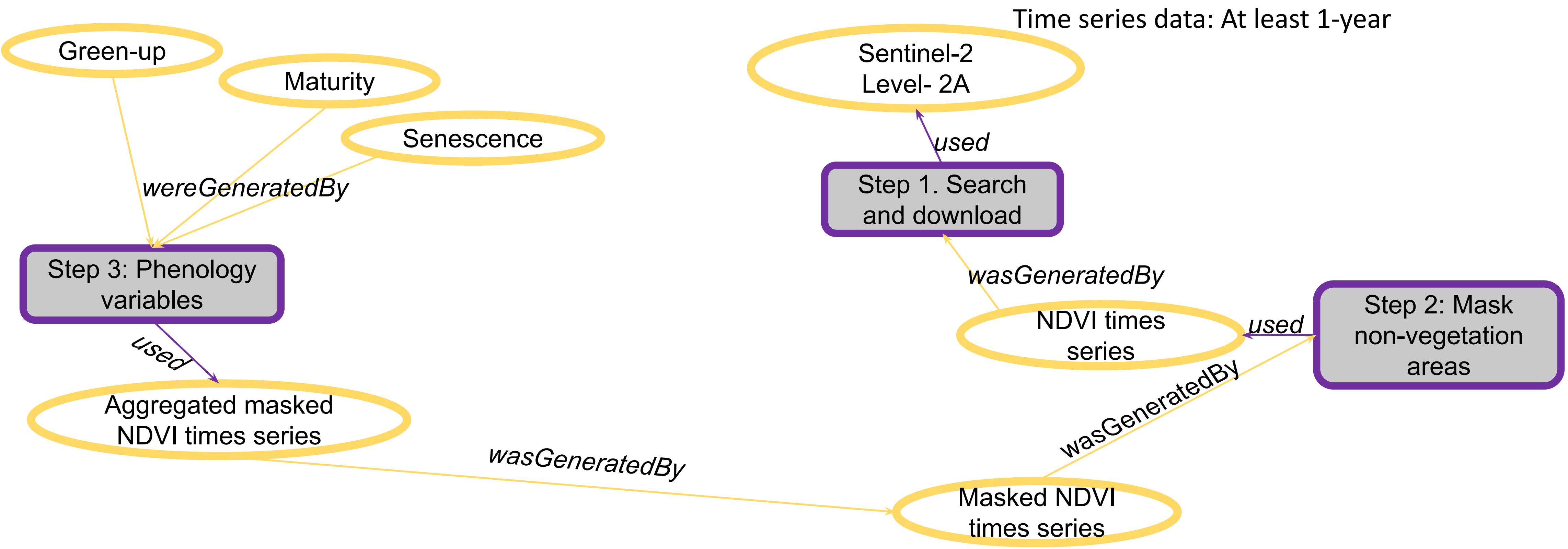
Phenology metrics describe specific stages on the seasonal trajectory e.g., start of the season (SOS), the peak of the season (POS), length, end of the season (EOS), etc. This variables characterize the phenology trends for an ensemble of Protected Areas.





How to derive phenology from Sentinel-2?

- We produce specific phenology products with careful considerations for each site



Alternative infraestructures:

OPEN DATA CUBE

jupyter

CREODIAS
WWW.CREODIAS.EU

sobloo
WWW.SOBLOO.EU

mundi WEB SERVICES
WWW.MUNDIWEBSERVICES.COM

ONDA
WWW.ONDA-DIAS.EU

WEkEO by COPERNICUS
WWW.WEKEO.EU

jupyter

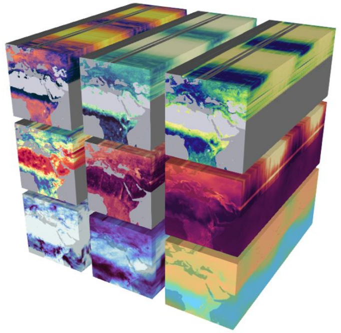
ESSI Lab VLAB

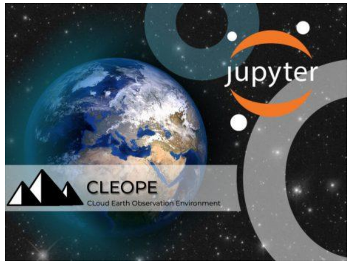
CO

docker

jupyter

Direct access to products in an xarray.Dataset





OData API

Harmonised Data Access (HDA) API

S3 Bucket

Search and download Sentinel-2 granule files

We don't have the data available. Query and download satellite EO data is needed e.g, from Copernicus Open Access Hub, Google Earth Engine…

Lessons learnt

- Xarray package models the datacube and allows for extracting phenology products
- Creating the xarray is tricky as this will depend on how to access or download the Earth Observation data. This is different in each infrastructure, and needs to be adapted for each environments.
  - In some cases data needs to be downloaded (that requires a lot of space and time)
  - In some others needs to be accessed by a different API (that requires transform the data model into a common xarray model)
- All in Python with Jupyter notebooks
- Documented in GitHub

https://www.ad4gd.eu/

AD4GD

Our mission

Design and develop Integrate standard data sources (e.g. Insitu, RS, CitSci, IoT, AI) in the **Green Deal Data Space**, in collaboration with other groups, to connect data and demonstrate with concrete examples that climate change zero pollution, biodiversity general problems can be solved

AD4GD

Our proposal

- Interoperability
  - Encourage implementation of the new OGC APIs including STA
  - Improve semantic interoperability by means of Essential Variables
  - Improve data quality descriptions
- Defragmentation
  - By applying Data sharing and FAIR principles
  - Reusing existing cyberinfrastructures such as EOSC and GAIA-X
- Based on datacubes?
  - *Not really*, but we need a dialog with the sister projects

Demonstrations

- In Water quality lakes (In the city of Berlin)
- In Biodiversity Corridors (In the Metropolitan Area of Barcelona)
- In Air Quality (with low cost sensors)

SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

What should be the priority for the DataCube evolution? (a few key words)

Harmonia

Anastasios (Tasos) Temenos

National Technical University of Athens

ICCS

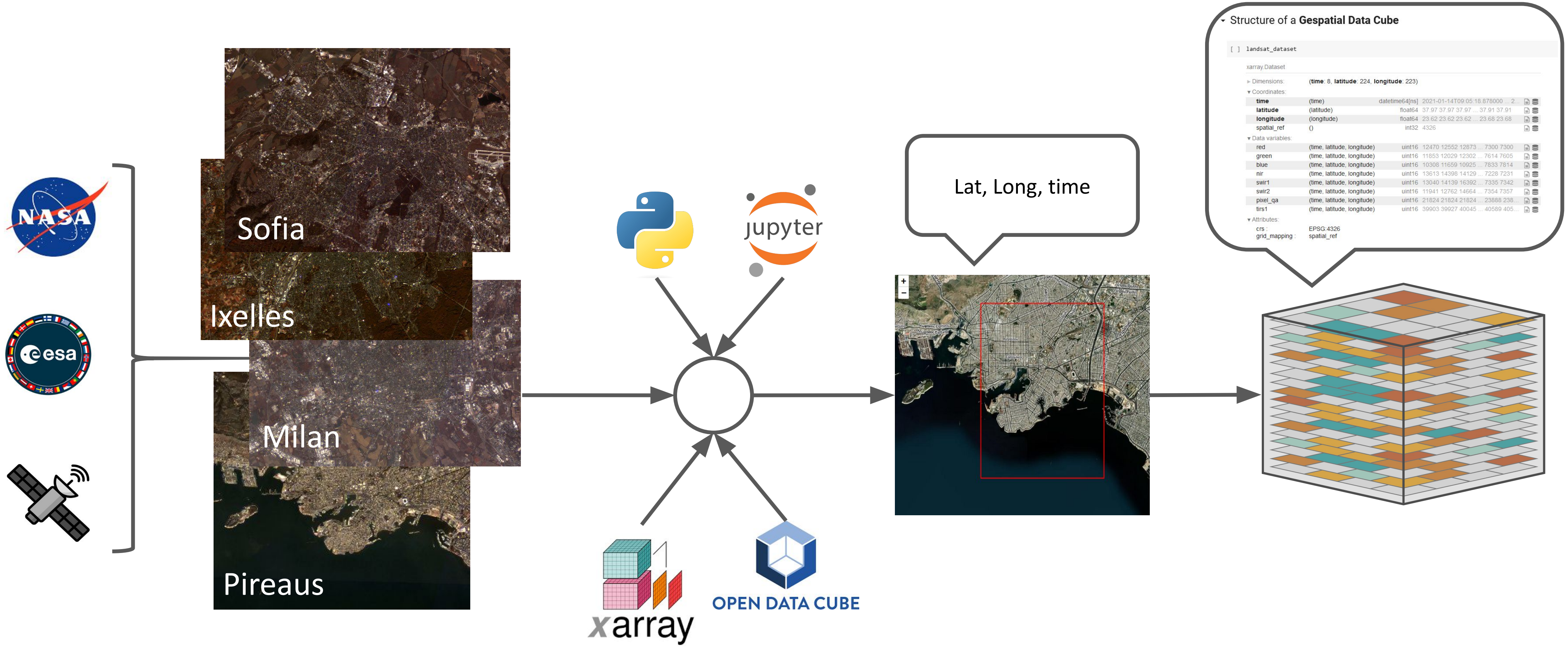
Goal of the presentation

- ✔ Datacube construction concept
- ✔ Presentation of a sample application
- ✔ Datacube extension
- ✔ Deployment using docker containerization
- ✔ Datacube applications:
    1. Spectral products
    2. Water Observations from Space (satellite)
    3. Vegetation phenology changes
    4. Vegetation change using NDVI
    5. ERA5 Climate Data - Rainfall and Air Temperature
- ✔ CNN models for Cloud removal task

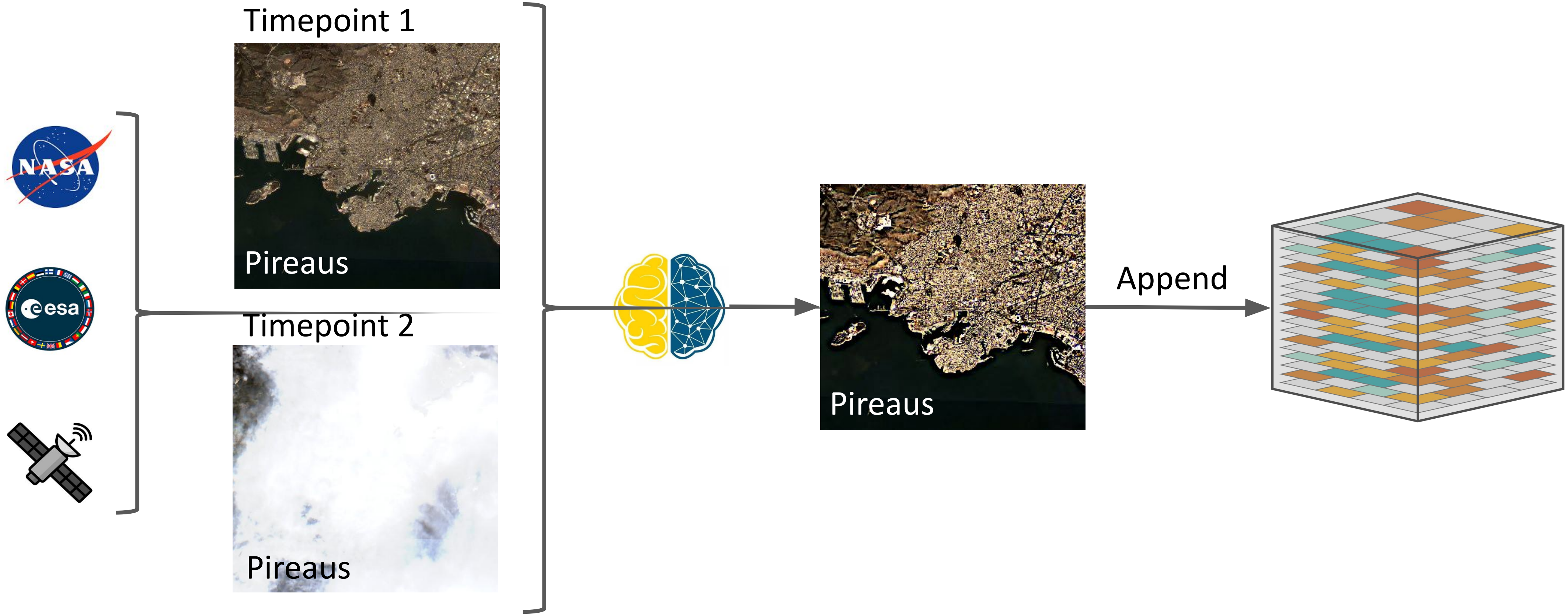
Novelties

- ✔ Storage of heterogeneous data satellite images, time series data, etc.
- ✔ Utilization of both spatial and temporal information
- ✔ Datacube dynamical extension aiming to store not only the raw data but also the generated information
- ✔ Flexible retrieval of complex information (i.e. spectral response of specific pixel on a satellite image

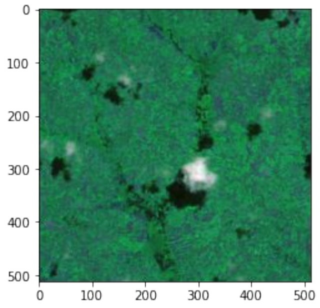
Data Cube initial construction

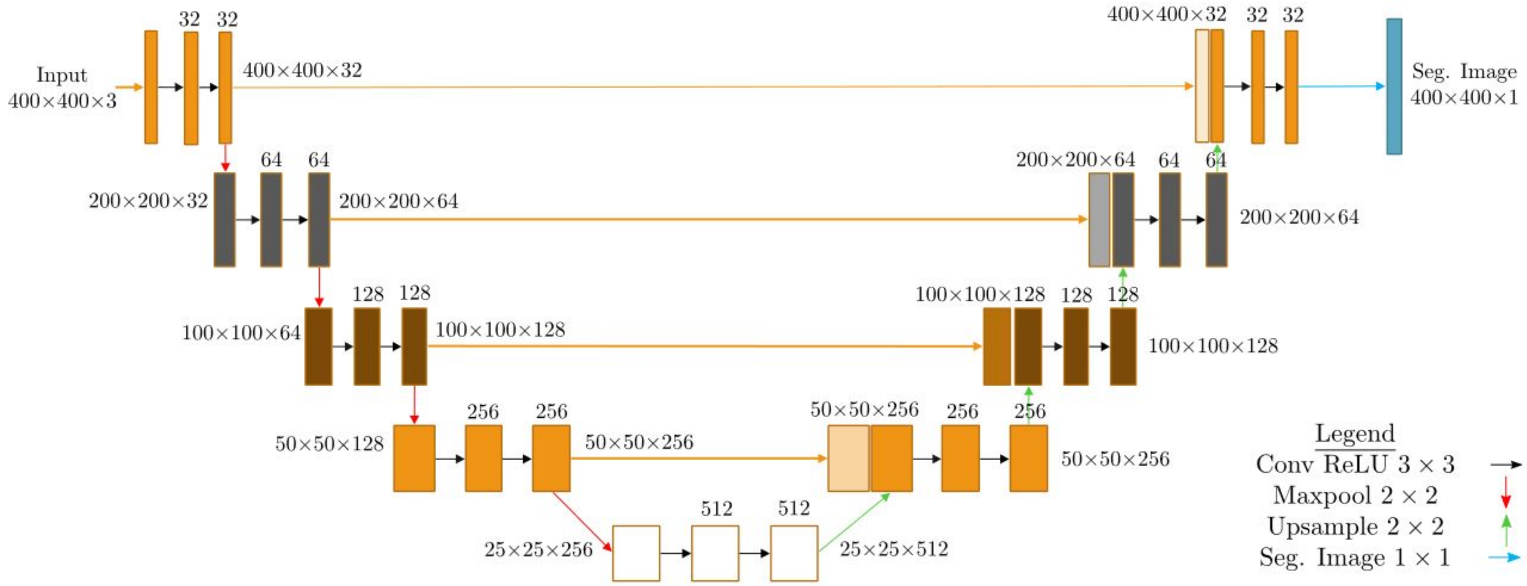


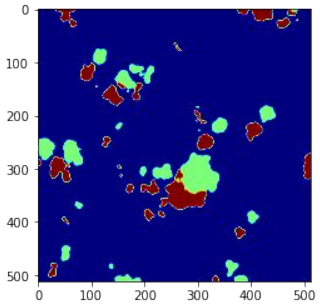
Cloud and shadow removal using AI techniques



U-NET for Cloud and Shadow removal



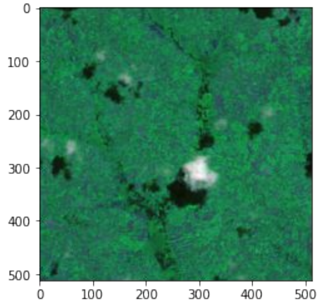




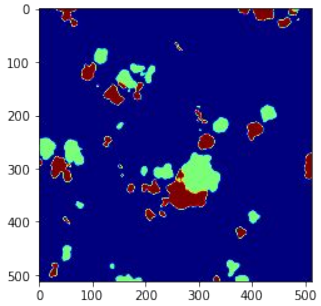
Temenos, A., Temenos, N., Doulamis, A., & Doulamis, N. (2022). On the Exploration of Automatic Building Extraction from RGB Satellite Images Using Deep Learning Architectures Based on U-Net. *Technologies*, *10*(1), 19.

Data Cube extension

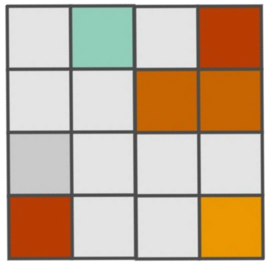
Satellite image

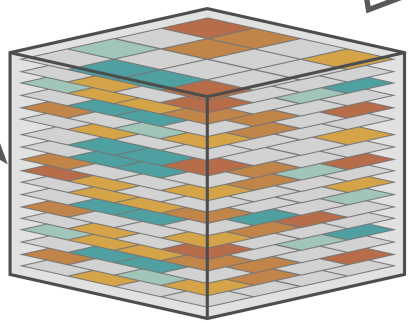


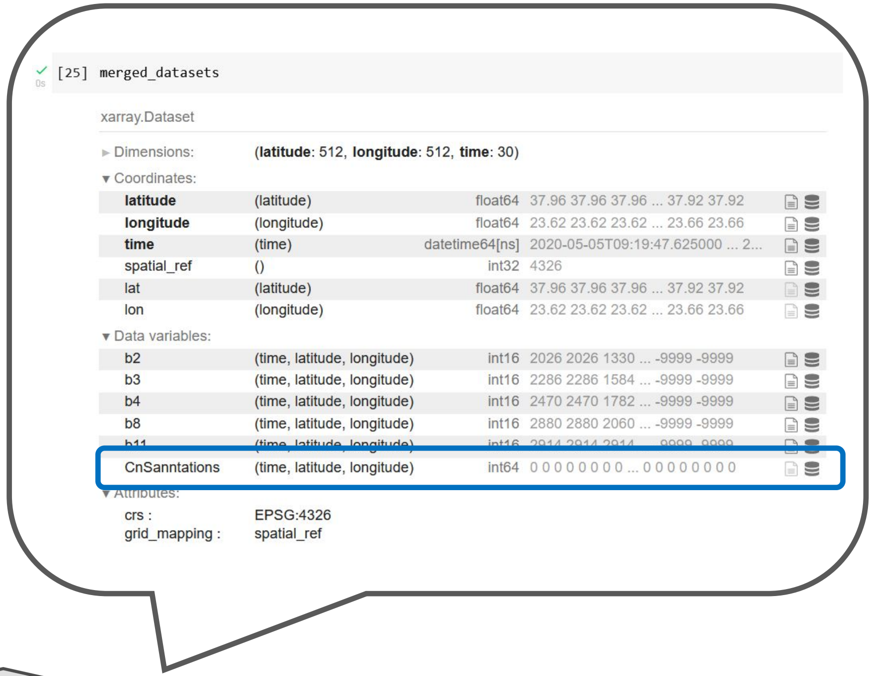
Model predictions



New layer



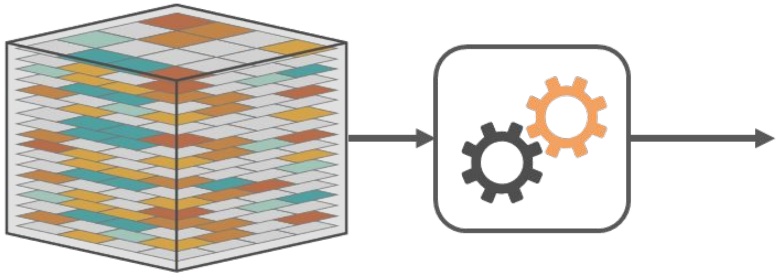


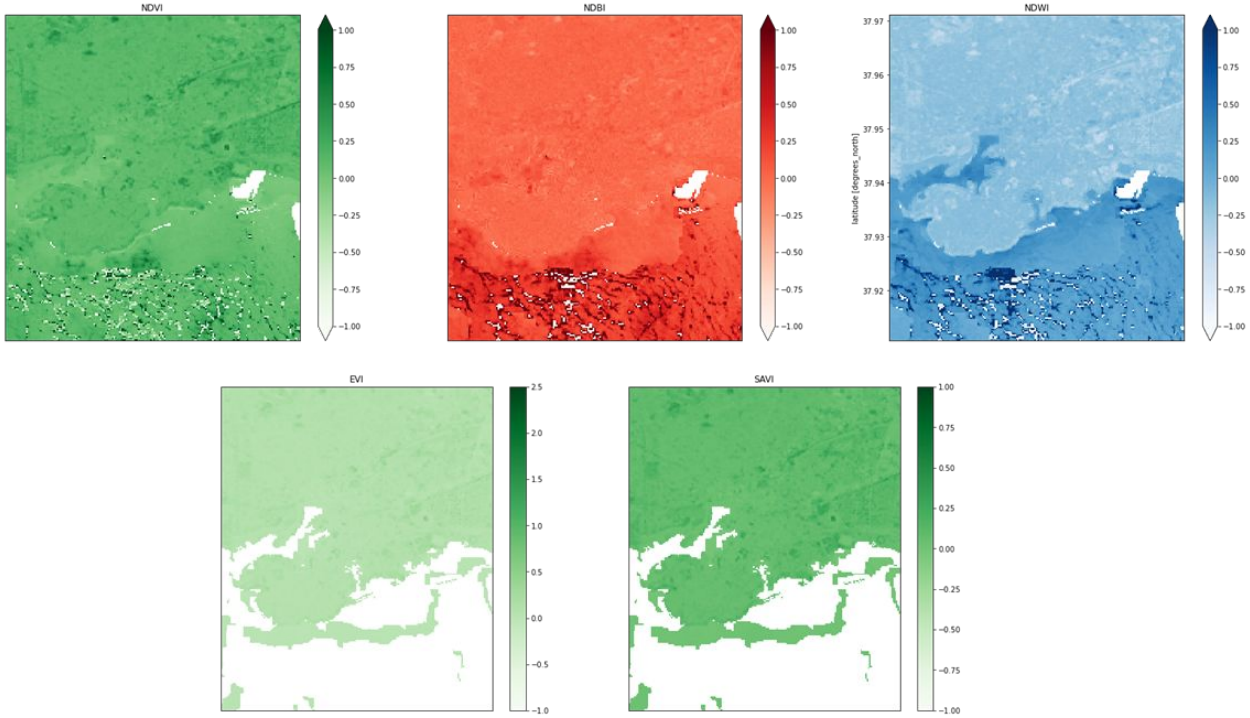


Datacube spectral info exploitation

NDVI NDBI NDWI EVI and SAVI map construction

Use of distinct spectral bands in an aim to generate various remote sensing indexes that illustrate certain areas of interest map.

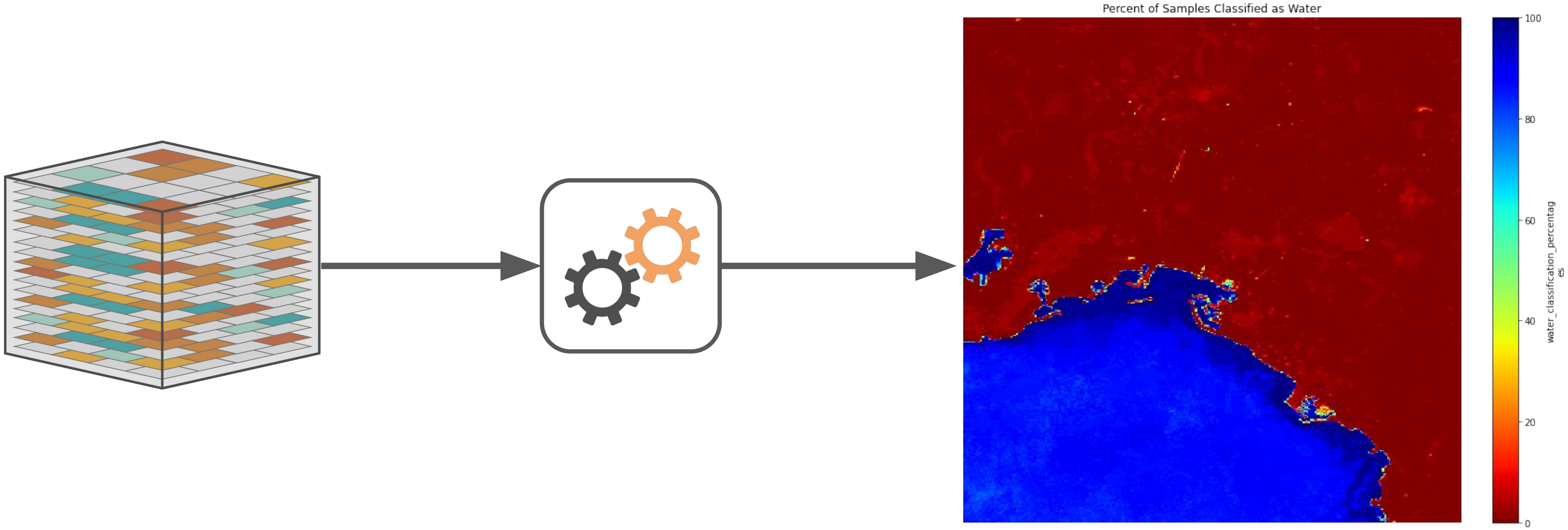




Datacube temporal info exploitation

Water observations from Space

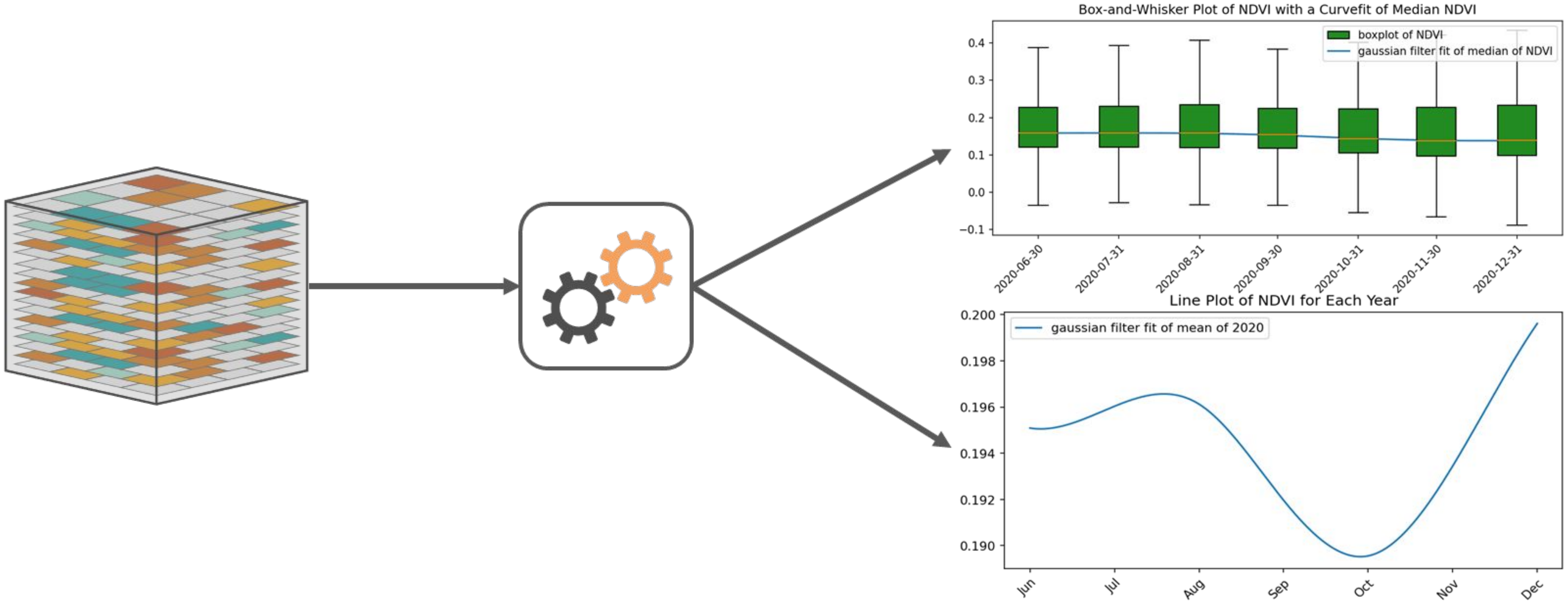
The selected data is used to create a time series water product that reflects the percentage of time each pixel contains water over the entire time frame. If a pixel represents cloud, it is not considered in the analysis due to some Cloud Statistics analysis previously performed.



Combination of spatial and temporal info

Vegetation phenology changes

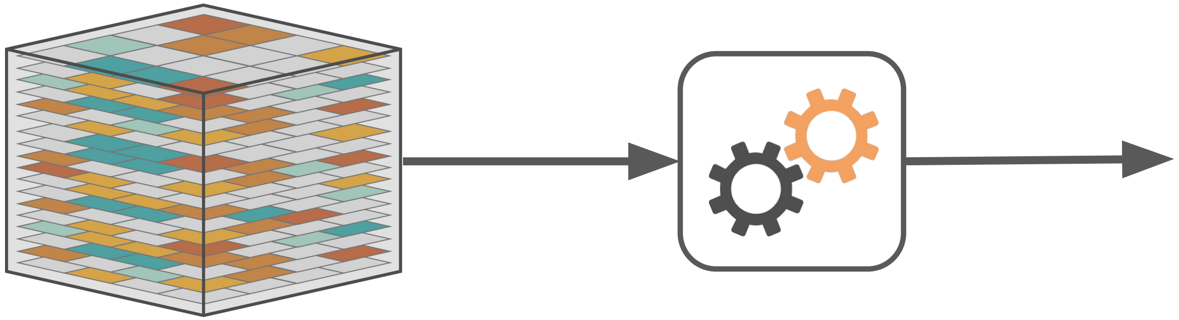
This application calculates the vegetation phenology changes using Landsat-8 data. Aiming to detect changes, the algorithm uses Normalized Difference Vegetation Index (NDVI) which is a common proxy for vegetation growth and health. The outputs can be used to assess differences in agriculture fields over time or space and also allow the assessment of growing states such as planting and harvesting.

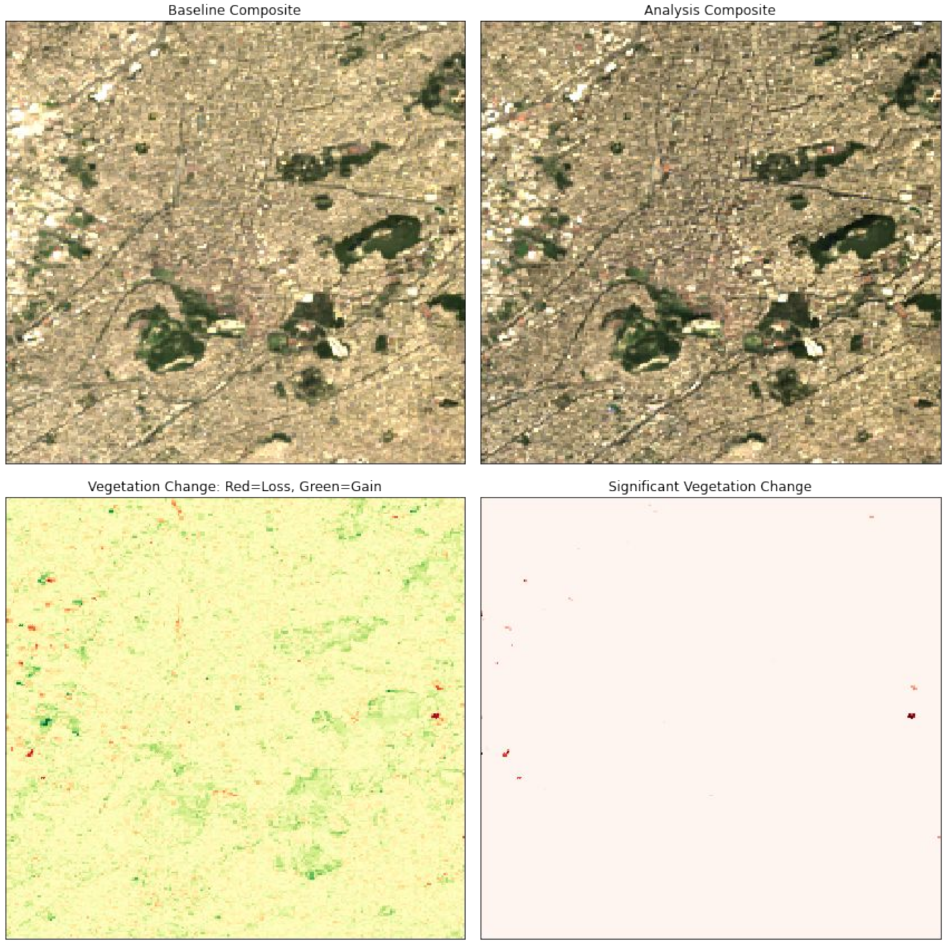


NDVI related, spatio-temporal application

Vegetation change using NDVI

The selected data is used to calculate changes in the Normalized Difference Vegetation Index (NDVI) which is consistent with vegetation change. The algorithm identifies a "baseline" and "analysis" time period and then compares the spectral index in each of those time periods. Significant changes in NDVI (vegetation greenness) are coincident with land change, as long as the comparisons are done between similar time periods (seasons or years)

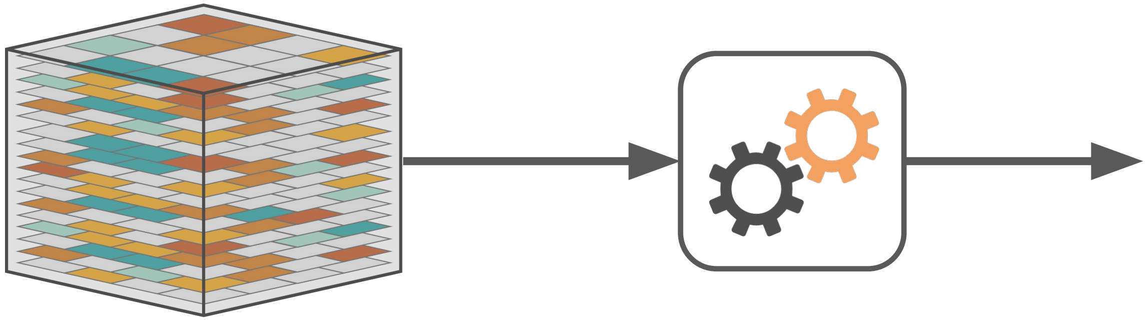


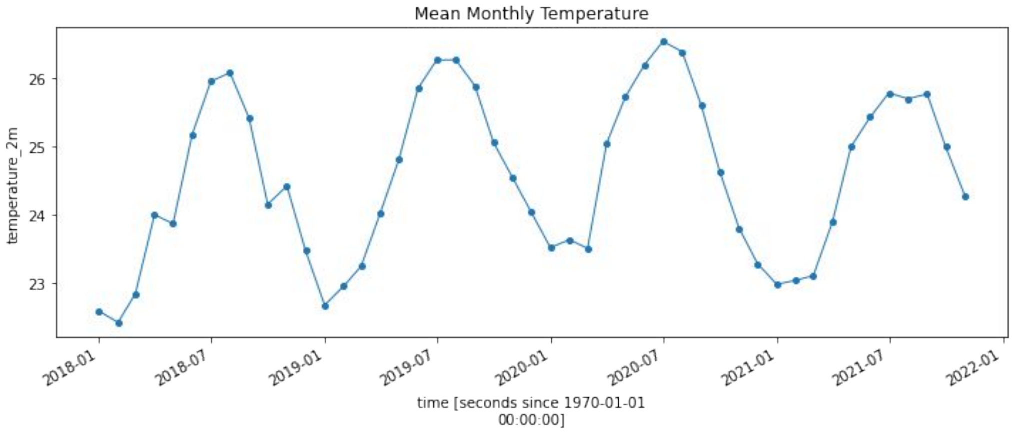


ERA5 related, spatio-temporal application

Monthly temperature estimation

The product produces data that goes several decades back in time, providing an accurate description of the climate of the past. The data presented here is a subset of the full ERA5-Land dataset post-processed by ECMWF. Monthly-mean averages have been pre-calculated to facilitate many applications requiring easy and fast access to the data, when sub-monthly fields are not required. The spatial resolution is 11.132 km. The dataset starts in 1981.





THANK YOU FOR YOUR ATTENTION

Thank you

Any questions?

SLIDO

Event Code: 3767288

lorem ipsum lorem ipsum

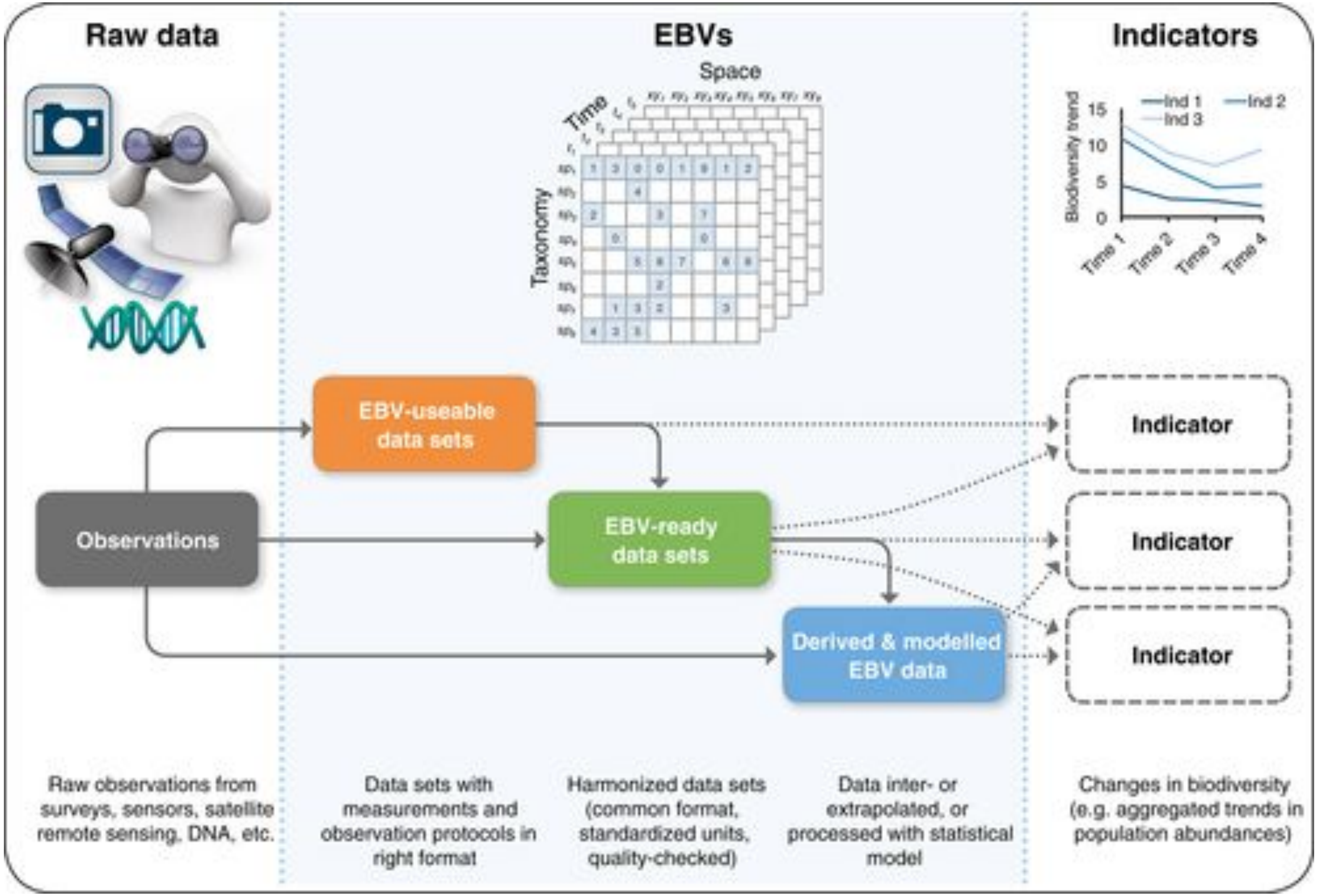
https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

B^3 (B-Cubed)

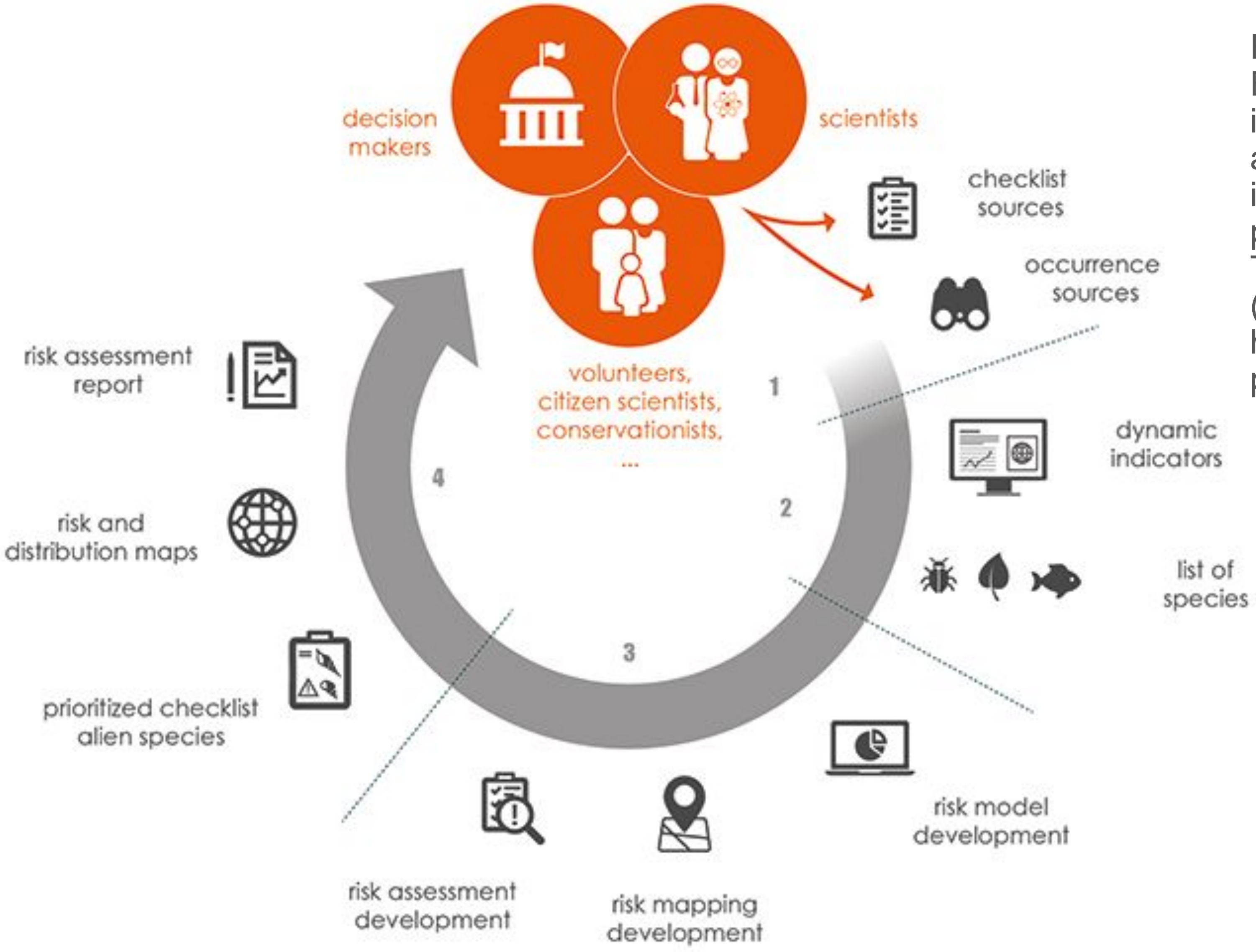
Quentin Groom on behalf of the B^3 consortium

Meise Botanic Garden

BIODIVERSITY
BUILDING
BLOCKS FOR
POLICY MAKING

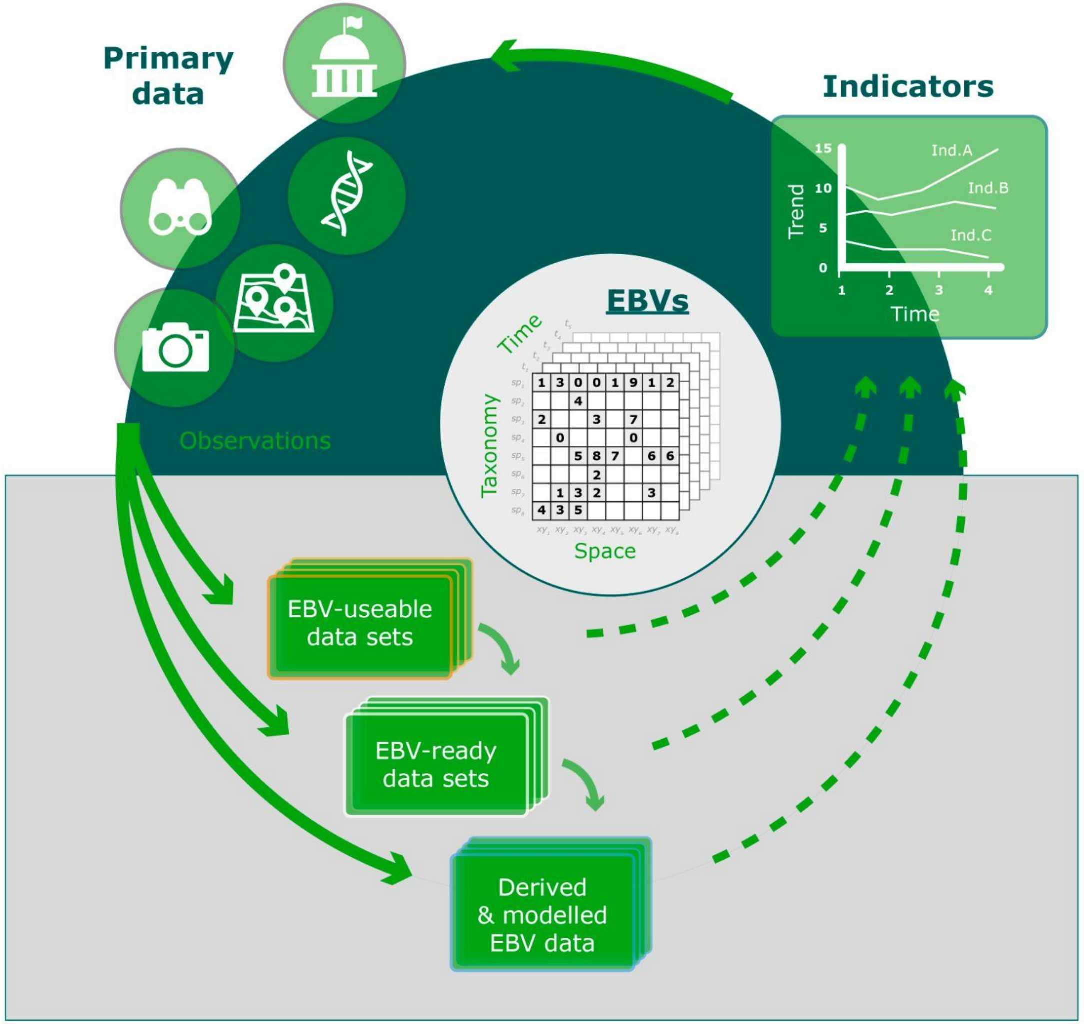


From: Kissling et al. (2018) Biol Rev, 93: 600-625. https://doi.org/10.1111/brv.12359



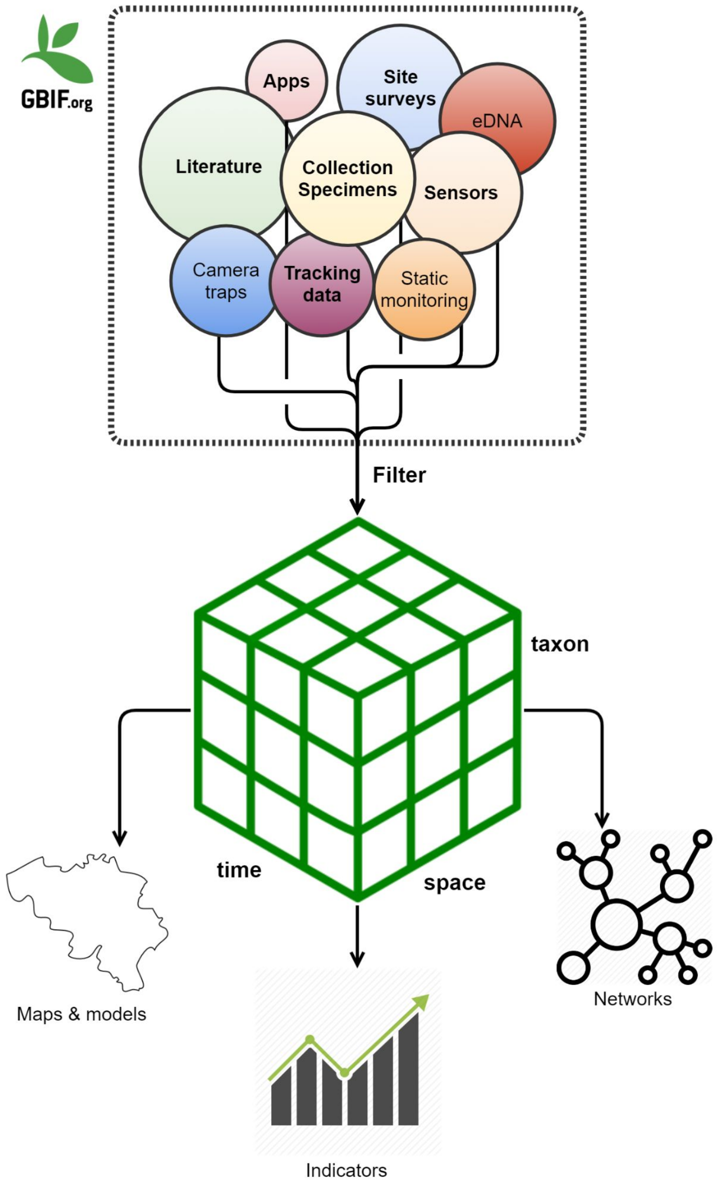
From: Groom et al. (2019) Empowering citizens to inform decision-making as a way forward to support invasive alien species policy. Citizen Science: Theory and Practice, 4 (1), 33. 11, pp. https://doi.org/10.5334/cstp.238

EUROGEO WORKSHOP 2022



FAIR

- Findable
  - Individually referenceable cubes
- Accessible
  - In the cloud close to compute, data and storage
- Interoperable
  - In formats suitable for use with other environmental data
- Reusable
  - Provence to the original data is preserved

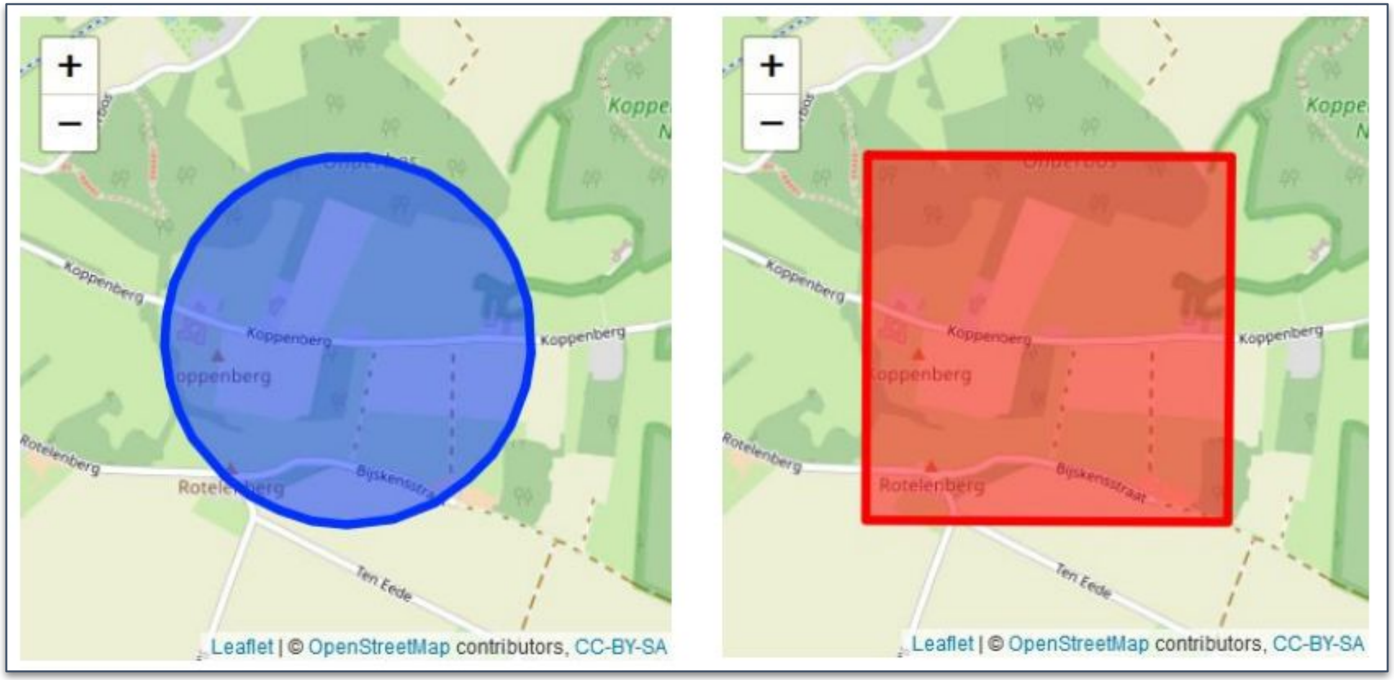


Fitting round pegs into square holes

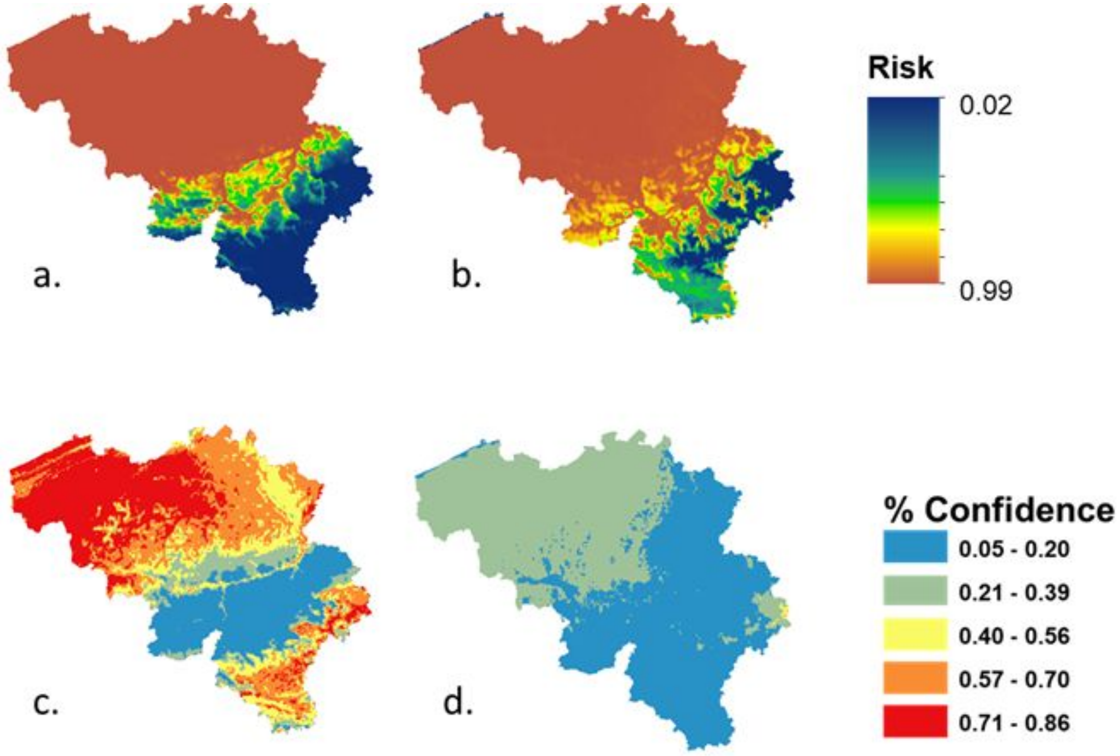
Occurrence cubes: a new paradigm for aggregating species occurrence data

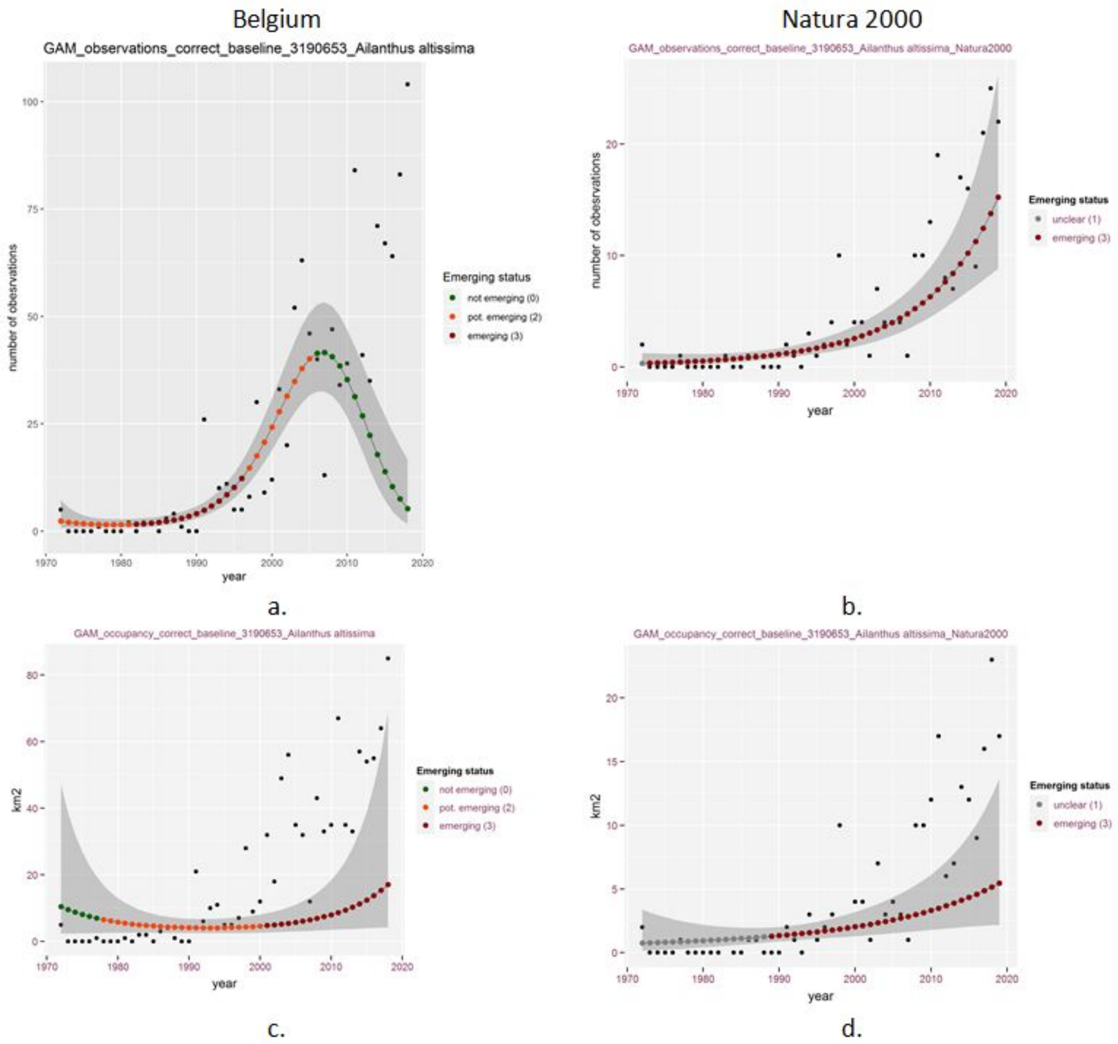
Damiano Oldoni, Quentin Groom, Tim Adriaens, Amy J.S. Davis, Lien Reyserhove, Diederik Strubbe, Sonia Vanderhoeven, Peter Desmet

bioRxiv 2020.03.23.983601; doi: https://doi.org/10.1101/2020.03.23.983601



Assigning observations to grid cells


a.
b.
Risk
0.02
0.99
c.
d.
% Confidence
0.05 - 0.20
0.21 - 0.39
0.40 - 0.56
0.57 - 0.70
0.71 - 0.86


Belgium
GAM_observations_correct_baseline_3190653_Ailanthus altissima
number of observations
year
Emerging status
not emerging (0)
pot. emerging (2)
emerging (3)
Natura 2000
GAM_observations_correct_baseline_3190653_Ailanthus altissima_Natura2000
number of observations
year
Emerging status
unclear (1)
emerging (3)
a.
b.
GAM_occupancy_correct_baseline_3190653_Ailanthus altissima
km2
year
Emerging status
not emerging (0)
pot. emerging (2)
emerging (3)
GAM_occupancy_correct_baseline_3190653_Ailanthus altissima_Natura2000
km2
year
Emerging status
unclear (1)
emerging (3)
c.
d.

Partners

Meise Botanic Garden, Belgium

Global Biodiversity Information Facility, Denmark

INBO – Research Institute for Nature and Forest, Belgium

University of Bologna, Italy

University of Giessen, Germany

Ovidius University of Constanța, Romania

South African National Biodiversity Institute, South Africa

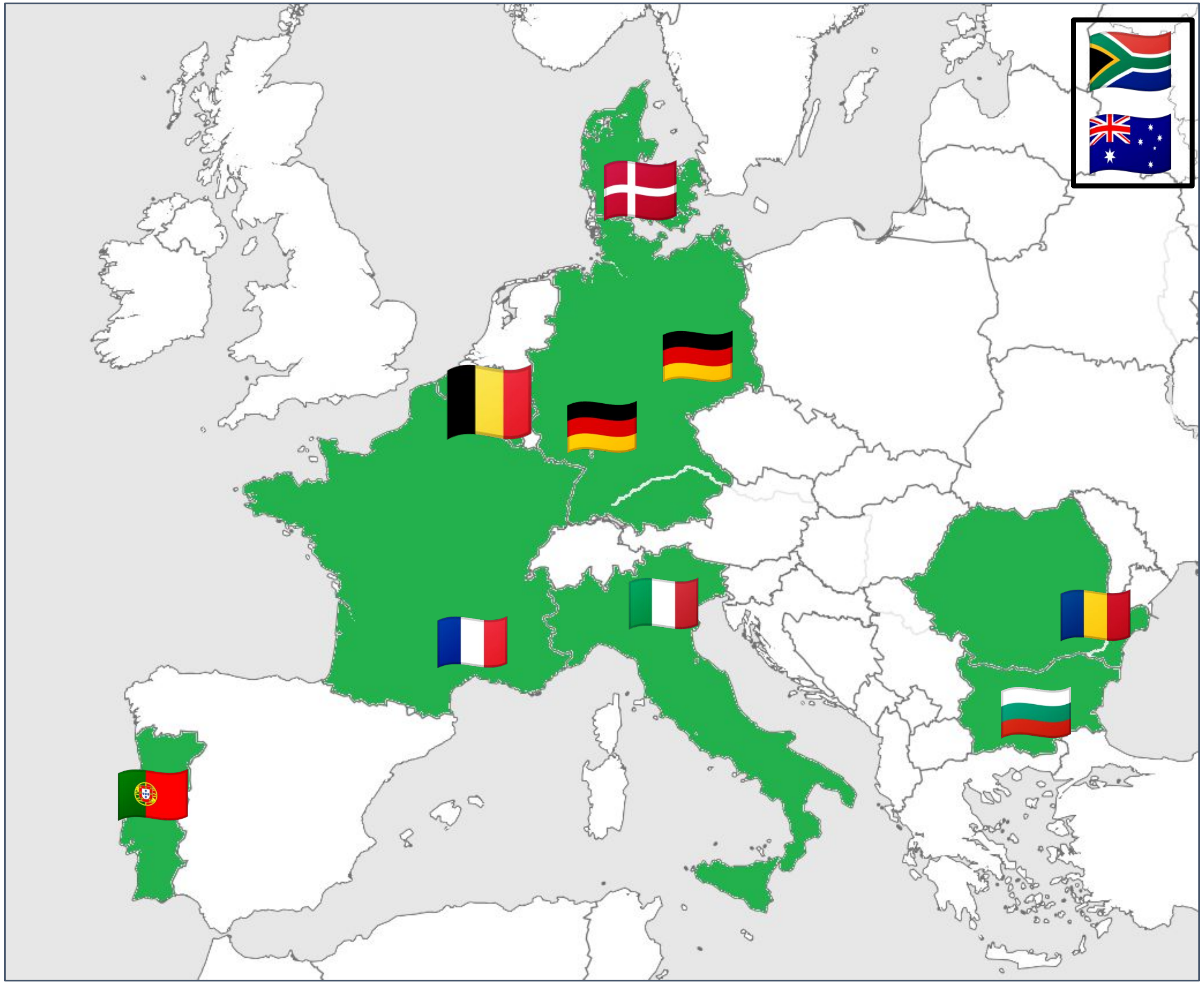
Stellenbosch University, South Africa

Pensoft Publishers, Bulgaria

Martin Luther University Halle-Wittenberg, Germany

National Institute for Research in Digital Science and Technology, France

University of Aveiro, Portugal



BIODIVERSITY BUILDING BLOCKS FOR POLICY MAKING

Let us know what we can do for you!

SLIDO

Event Code: 3767288

https://app.sli.do/event/bzc8pB6ffxSZwHnVpwiPQj

What additional data from biodiversity observations would you like to see aggregated together into biodiversity cubes (B3 project)?

a. sex ratios
b. life stage
c. behavior
d. reproductive condition
e. individual count
f. degree of establishment
g. basis of record

HELLENIC REPUBLIC
MINISTRY OF DEVELOPMENT AND INVESTMENTS
GSRI
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

Discussion Forum Q&A

Tom Hengl

OpenGeoHub

Next: Lunch at the Roof Garden
(8th Floor)
13:00 - 14:30

See full schedule
Here: