FAIRiCUBE

COMPLEMENTING EO DATA WITH FINE-GRAINED IN-SITU OBSERVATIONS

STEFANIA MORRONE, KATHI SCHLEIDT, STEFAN JETSCHNY

eurac research
Institute of Atmospheric Pollution Research
National Research Council of Italy
European Commission

BACKGROUND

- Discussions with non-EO colleagues on what all could be done by combining in situ observations with the wealth of EO products, if only we:
  - could gain an easy overview of what data is available, as well as necessary storage requirements;
  - understood how to merge different grids;
  - understood the potential of Machine Learning techniques;
  - could estimate the processing resources required for data analysis;
  - could integrate existing point and vector data.

Like a child in front of a candy store, not finding a way past the glass pane …

BREAKING THE GLASS PANE

Deliver the power of data cubes & ML to decision makers and data scientists merging spatio-temporal dimensions with thematic dimensions

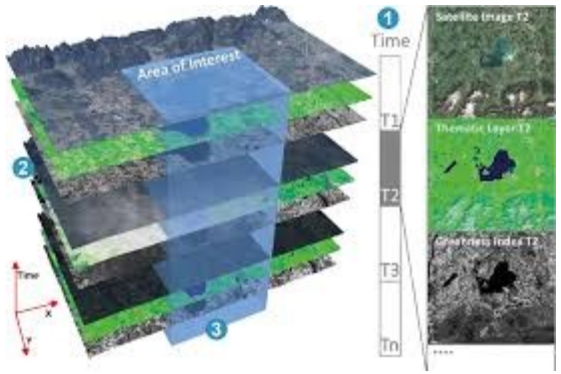
Findable

Accessible

Interoperable

Reusable

Data Cubes



- from collection of existing tools and services to integrated platform
- data catalogue of pre-gridded, pre-aligned, pre-referenced EO data
- data processing catalogue (including a/p resources & ML)
- data storage & compute resources
- possibility to create own custom data cube
- meta data pipeline (data, processing steps, …)
- community platform sharing

ABOUT

Horizon Europe Project - Duration: 07/2022 – 06/2025

Consortium:

- 3 research institutes

  NILU, Norway - **Wageningen University**, Netherlands - **Museum of Natural History**, Vienna, Austria

- 3 SMEs in the environmental & geomatic field

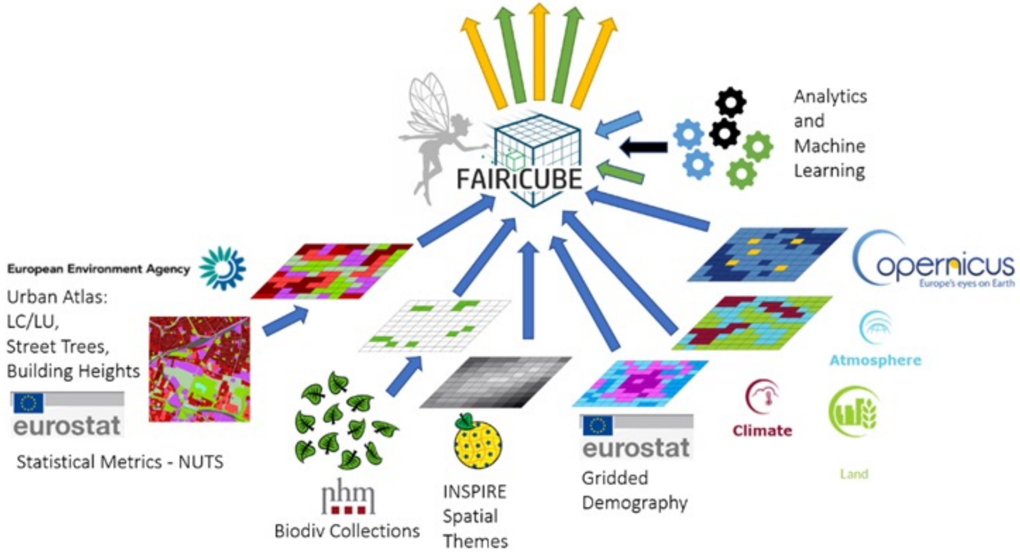
  space4environment, Luxembourg - **4sfera,** Spain - **Epsilon Italia,** Italy

- 2 technical SMEs

  EOX, Austria - **rasdaman,** via **Constructor University,** Germany

MISSION & OBJECTIVES

- Enable players from beyond classic EO domains to provide, access, process, and share gridded data and algorithms in a FAIR and TRUSTable manner.
- Create a common marketplace for data, algorithms, ML models

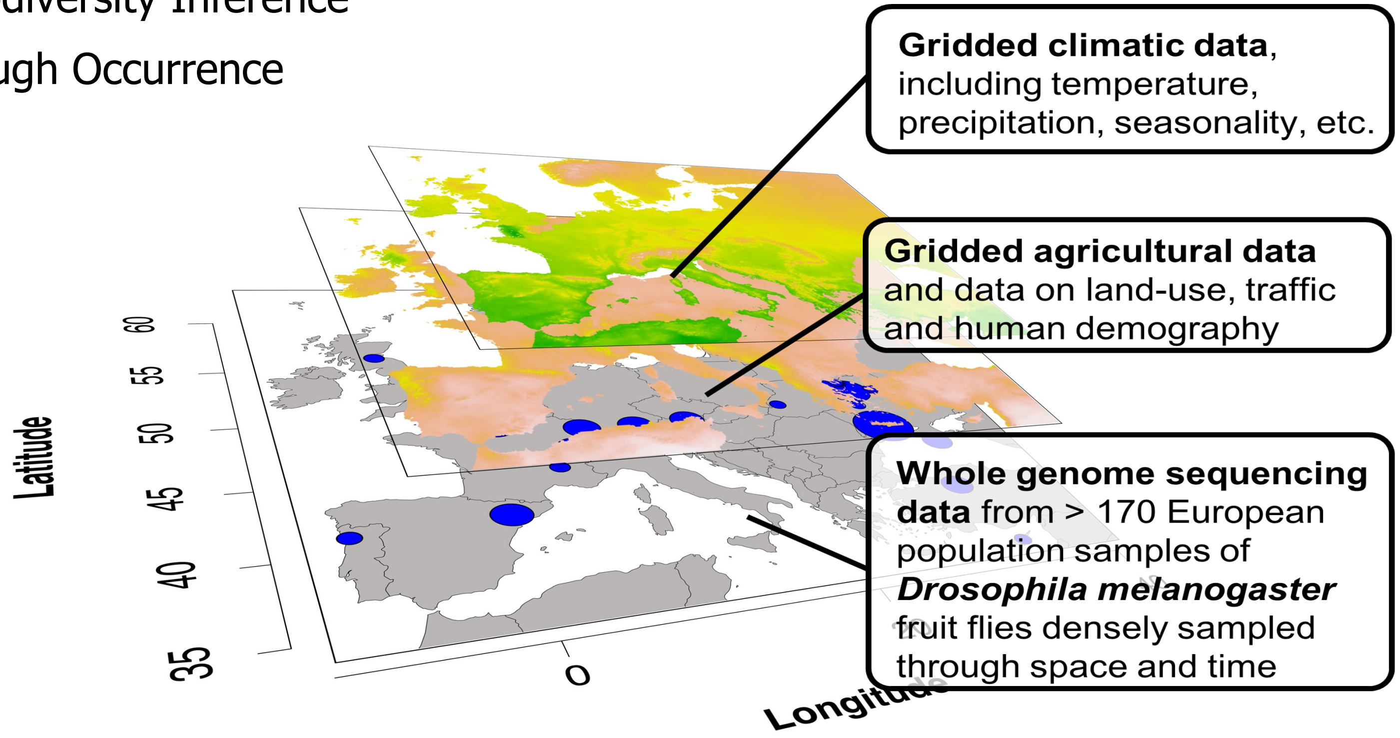


- Establish FAIRiCUBE Hub – an integrated platform for FAIR spatial earth observation data ingestion, analysis and ML
- Demonstrate FAIRiCUBE Hub by running 5 use cases addressing EU green deal actions (climate change, circular economy, biodiversity,..)
- Collaborate with major communities working on data cubes (Euro Data Cube, EarthServer)
- Extend the usability & visibility of EO data
- Provide insights to the creation of the GDDS

USE CASES

Urban and regional focus:

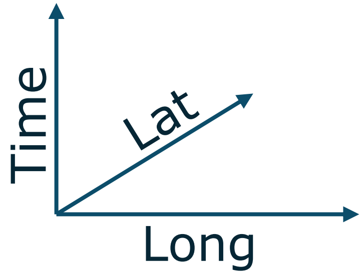
- Urban adaptation to climate change
- Spatio-temporal assessment of neighborhood building stock
- Biodiversity and agriculture nexus
- Linking Climatic and Genetic Variation for Biodiversity Inference
- Validation of Phytosociological methods through Occurrence Cubes



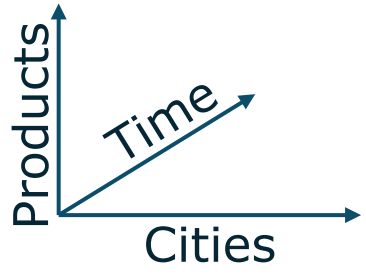
DATA CUBES

- A *data cube* refers to a multi-dimensional data structure, i.e., data within a data cube is explained by specific dimensional values.
- Separation between spatio-temporal cubes (Coverages) and OLAP cubes
- Both are required by our use cases, merging spatio-temporal dimensions with thematic dimensions such as:
  - Species taxon
  - Genomic variance
  - Land cover types

Coverage

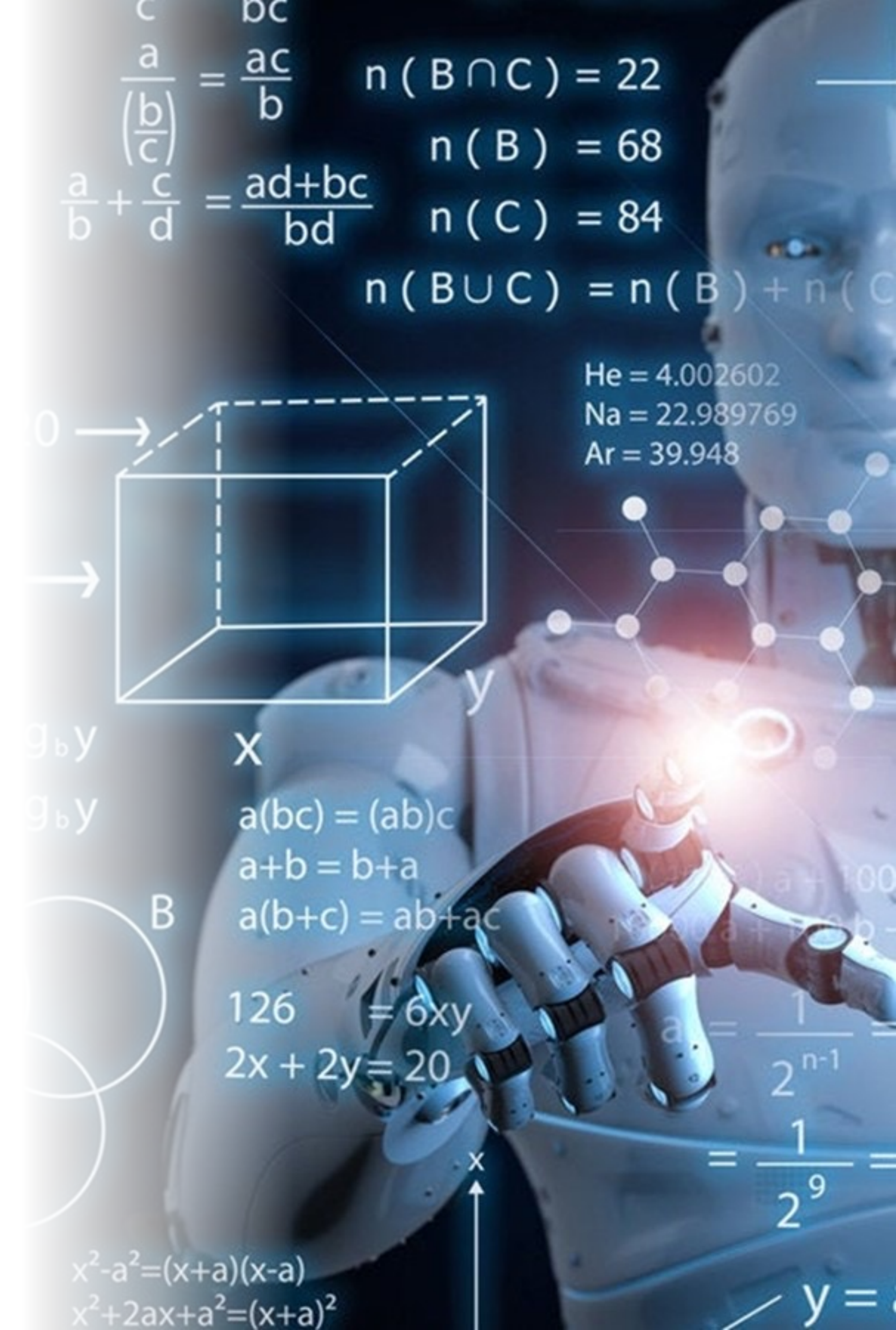


OLAP Cube



MULTIDIMENSIONAL SPATIO-TEMPORAL DATACUBES

- New suite of ISO Coverage Standards
  ISO 19123 Schema for coverage geometry and functions
  - ISO 19123-1:2023 - Part 1: Fundamentals
  - ISO 19123-2:2018 - Part 2: Coverage implementation schema
  - ISO 19123-3:2023 - Part 3: Processing fundamentals
- Focus is on multi-dimensional gridded (“raster”) coverages
  - Supports grid topologies whose axes are aligned with the axes of the CRS
  - Axes can also be referenceable, e.g. categorical lists, such as land cover or species
- rasdaman array database supports ISO 19123 specifications

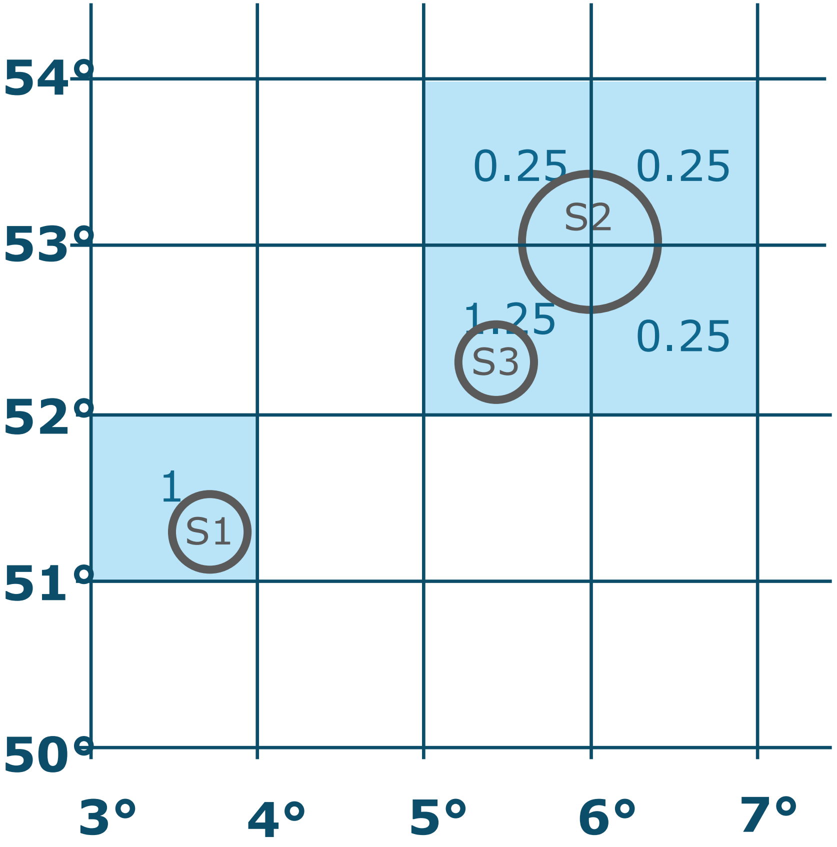


rasdaman
raster data manager

CONSTRUCTOR
UNIVERSITY

OCCURRENCE CUBES

- Species dimension in addition to spatiotemporal dimensions
- Underlying spatiotemporal dimensions aligned with Copernicus sources
- Takes uncertainty into account

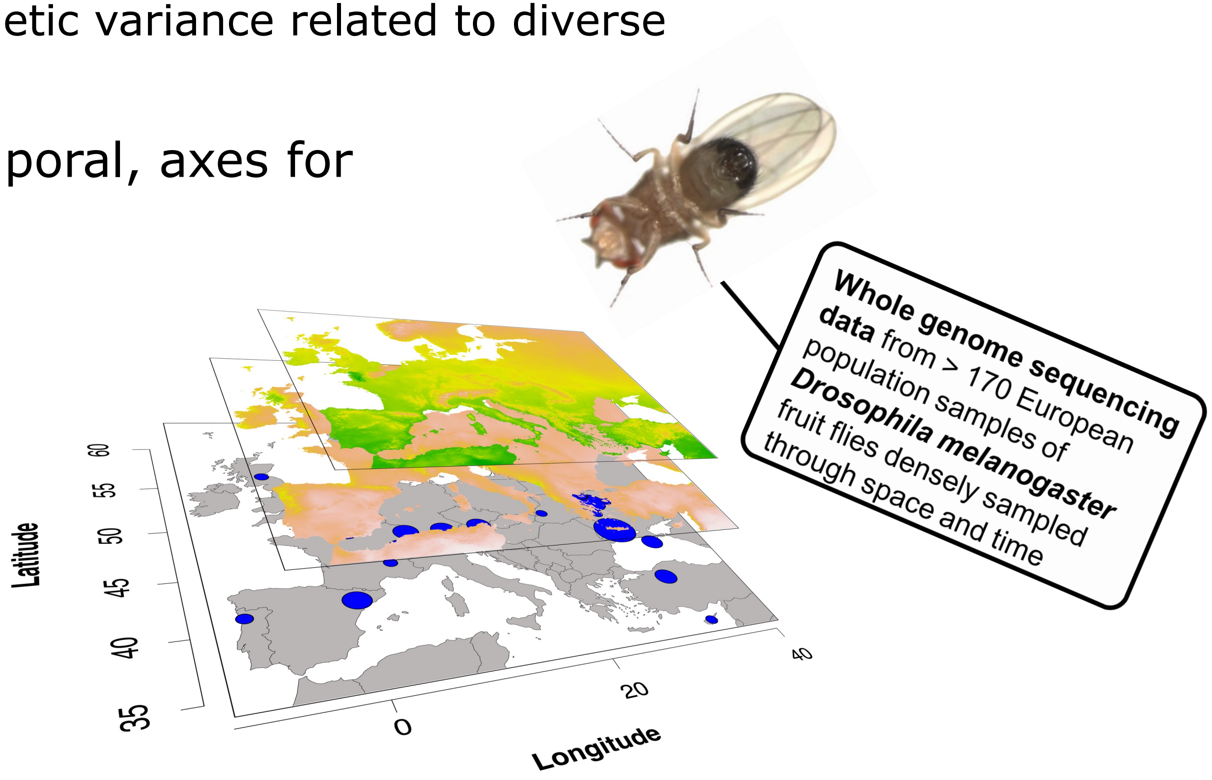


WAGENINGEN
UNIVERSITY & RESEARCH

naturhistorisches museum wien

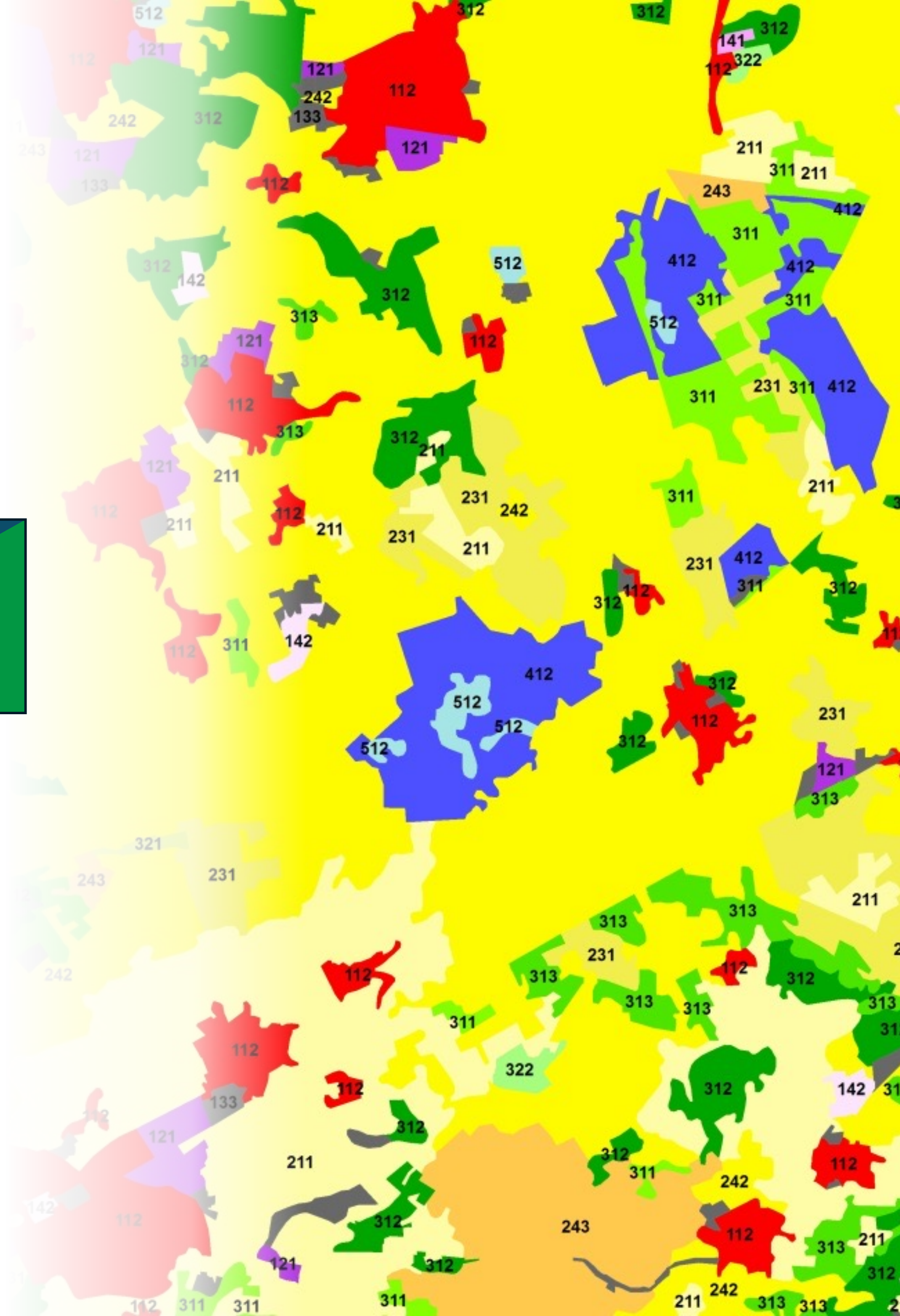
GENOMICS CUBES

- Leverage data from DrosEU - European Drosophila Population Genomics Consortium
  - Sequenced DNA data from 100s of *Drosophila melanogaster* populations
  - Better understand genetic variance related to diverse environmental factors
- In addition to spatiotemporal, axes for
  - Chromosome
  - Position
  - Nucleotide



LAND COVER ALTERNATIVE

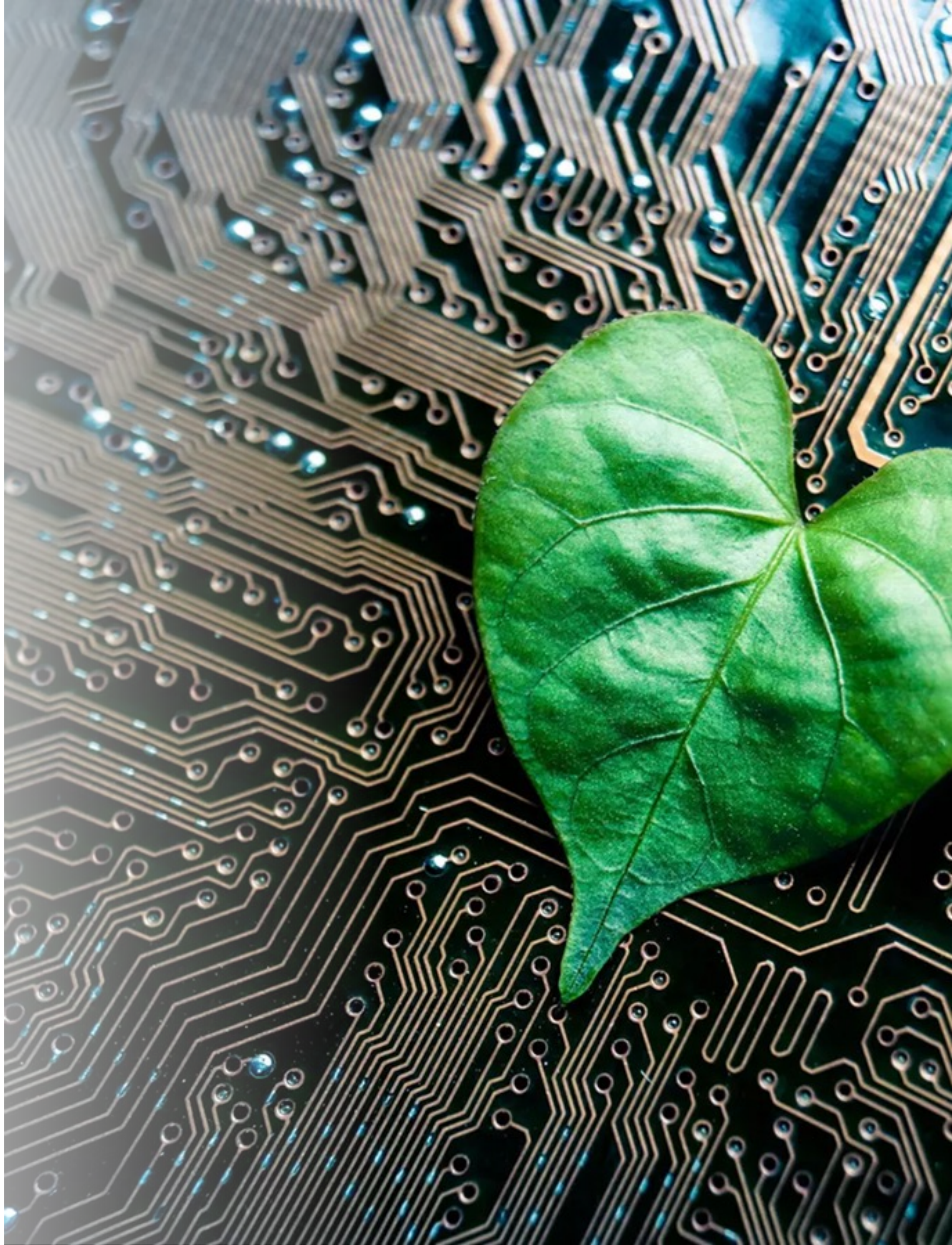
- Loss of information when converting from vector to raster
- Example:
  - Current grids only allow for one land cover type per cell
  - Through addition of land cover dimension, percentage of land cover type per cell can be provided
  - Allows further processing to access full information



X-DOMAIN ISSUES

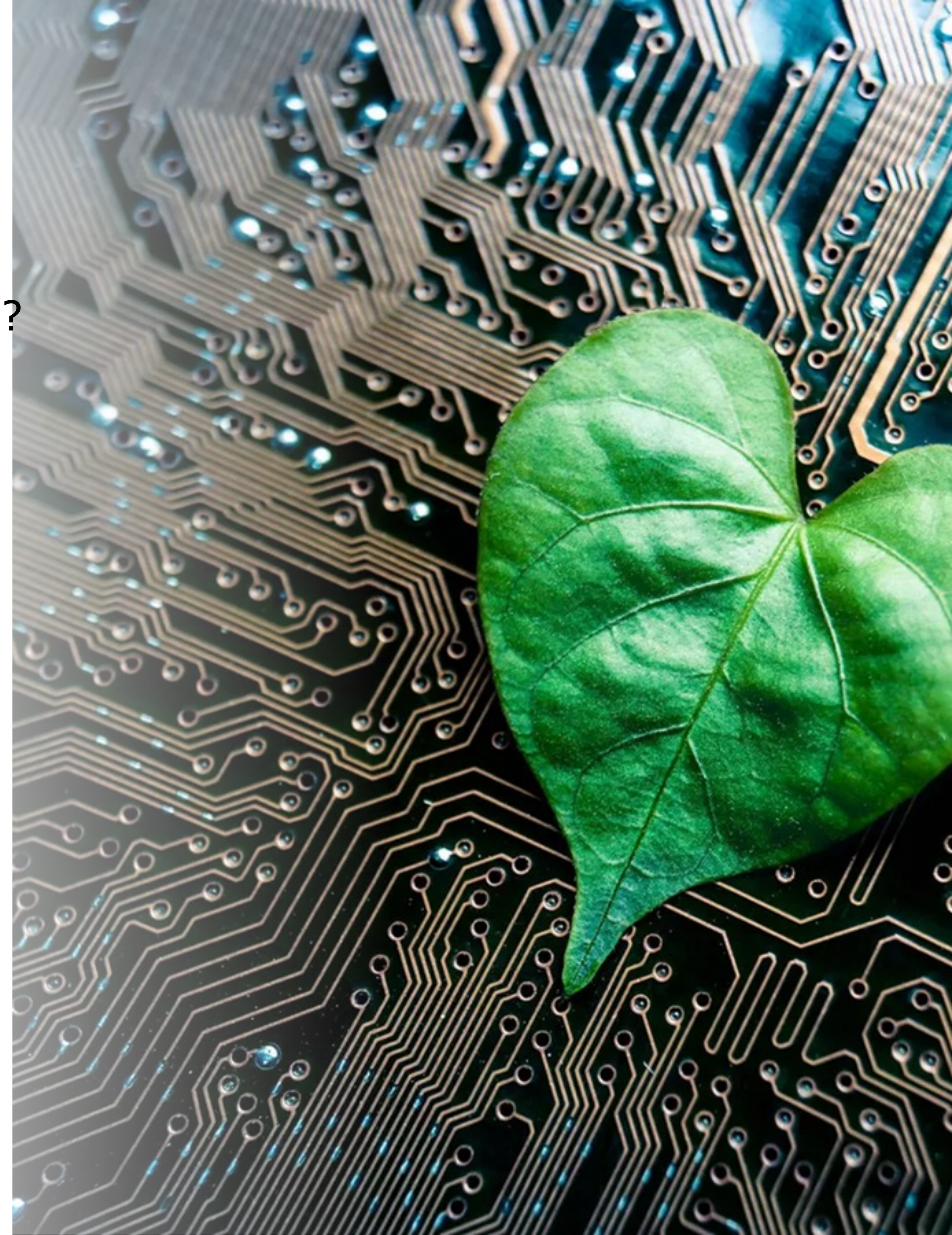
- CRS: Domain and ML experts usually unaware, think in "Google Coordinates"
- Domain Axes: how to integrate domain dimensions with spatiotemporal dimensions?
- AI/ML learning approaches: what provides dependable understandable results?
  - Requirement for understandable AI when supporting scientific research!
- Resource requirements: how much data storage and processing resources are required to reach the target?
- Metadata concepts:

  - Datasets: not much support for analysis ready data (ARD)
  - Analysis/Processing Resources: STAC encoding emerging, but not complete



X-DOMAIN ISSUES: GRIDS

- How to align different grids, some geodetic, some projected?
- Understanding different grid approaches:
  - Value in corner (which corner?)
  - Value in centre
  - Value is cell/pixel
- Different types of data require different resampling approaches:
  - Qualitative (categorical) data
    - Nominal (NON natural ordering): classification codes, telephone number, …
    - Ordinal (natural ordering): Hurricane scale, Richter magnitude scale, ..
  - Quantitative Data type
    - continuous data: temperature, slope, elevation, …
    - discrete date: number of rain days, population



CONTRIBUTIONS TO GDDS

- Standardizing data models supporting ARD
  - Utilizing OGC/ISO standards
  - Including comprehensive metadata
- Dynamic access to data and processing via APIs
  - Subsetting, e.g., give me surface temperature with this resolution and CRS
  - WCPS queries and User Defined Functions (UDF), e.g., calculate vegetation index out of LC, EL, OI layers
  - Encapsulate trained models in UDF called via WCPS
- Support in resource estimation (both memory & processing), required to correctly scope a project proposal
- Knowledge base providing support on diverse aspects of gridded ARD and ML/AI processing

CONCLUSIONS

- Current geospatial datacubes overfocus on spatiotemporal dimensions (alternatively ignore / poorly support thematic dimensions)
- For well founded research, as well as unlocking potential of deap learning and AI, we require both spatiotemporal and thematic dimensions
  - ISO 19123 suite of standards enable best of breed
- Standardization of data models and analysis/processing routines enables far more efficient utilization of these resources
- Information on potential pitfalls when applying ML to geospatial datacubes hard to come by, leading to inadvertent mistakes, must be transparently available

Many thanks for your attention!

QUESTIONS ?

s.morrone@epsilon-italia.it kathi@datacove.eu sjet@nilu.no