Synthetic populations for personalized policy

Jiri Hradec

Jiri.Hradec@ec.europa.eu

Margherita Di Leo

Margherita.Di-Leo@ext.ec.europa.eu

Joint
Research
Centre

Public policy design and evaluation

Challenge: Improve policy design and evaluation by identifying priority target

Vulnerable groups

Under-represented groups

Minorities

Examples

- Policy support to access healthcare and education
- Support to energy transition
- Elderly people living alone
- Single parents with small children
- ...

Limitations of univariate figures

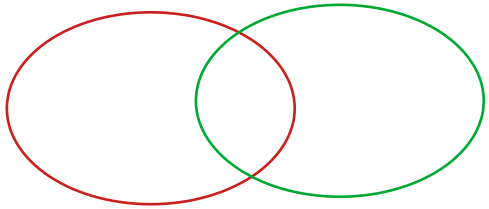
women

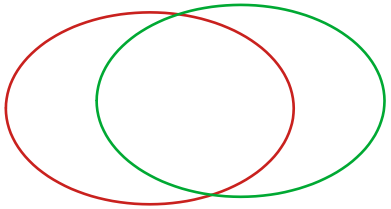
university educated people

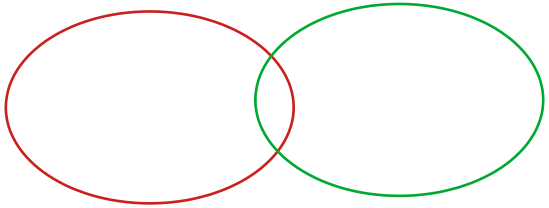
Limitation of statistics released at area-level aggregates:

People living in the same spatial units all share the same characteristics

Infinite different scenarios...







Multivariate distributions in spatial modelling...

… improve the representation by means of **high-quality aggregates for specific use cases**

… pose **privacy concerns**: combination of features in low density areas prone to give away identity of citizens

Personalized Policy

The objective

Design policy to target those who need it the most

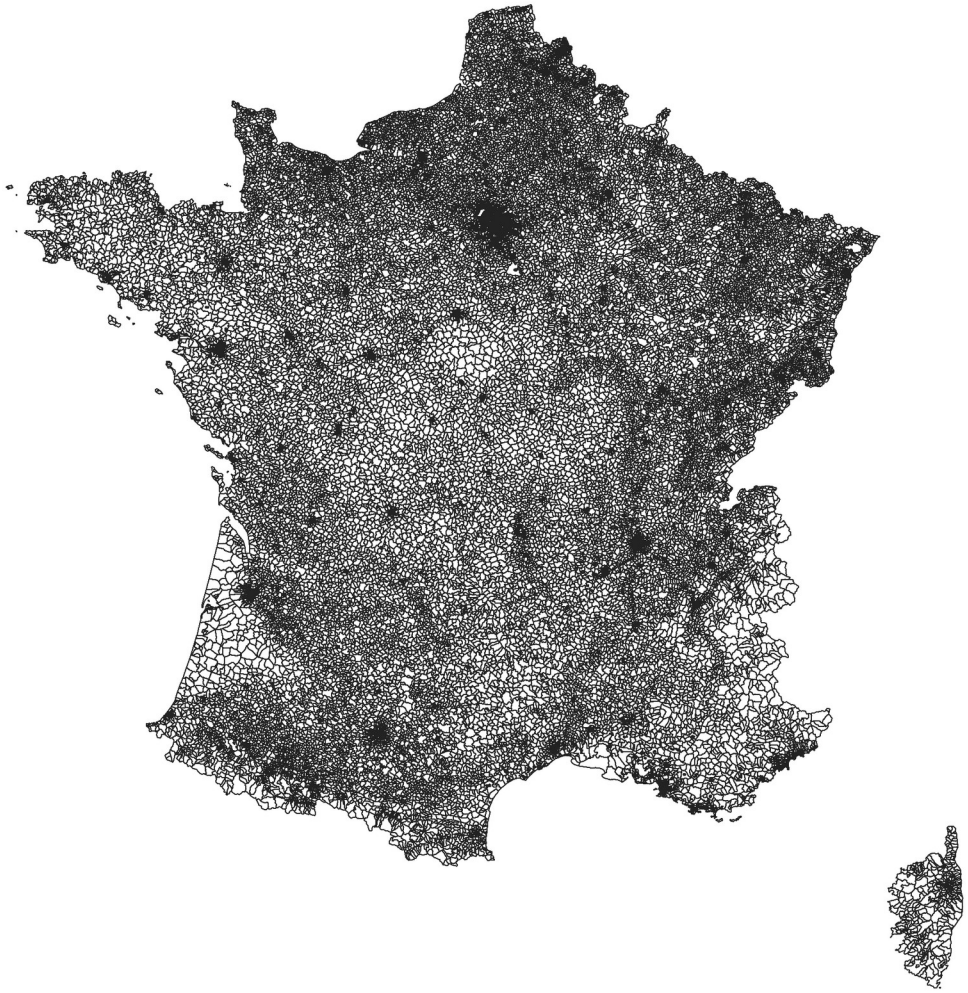
The idea

Use profiling technologies similar to those used by big commercial platforms to personalise services to customers

> Develop personalised profiles without having to deal with real personal data: create a **probabilistic synthetic population** from disaggregated official statistics

- Synthetic population as a baseline to be used as an input to spatial models.
- Flexible enough to be successively enriched and updated whenever more data becomes available.
- Population to carry all possible information until a model is chosen and the relevant features can be selected accordingly.

Case study: France



France population (Source: UN)
64,668,000 (2016)

Population density
118.1 per sq km

Surface area
551,500 sq km

~35k census units (IRIS)

Each IRIS constitutes a “micro-neighborhood”, made up of a set of contiguous and homogeneous blocks, bringing together 2,000 inhabitants or more.
Each IRIS constitutes a basic municipal sector, a homogeneous geographic and demographic “micro-district”, clearly and durably defined.
This “elementary link” is used to collect statistical and demographic data. These data are then analyzed and the results published by INSEE (French Statistical Office).

Data sets description and preparation

Population census data provided by the **French National Institute of Statistics and Economic Studies INSEE**
Data model is a set of tables linked to each other by means of keys
The tables we used:

INDCVI (Individus localisés au canton-ou-ville)
Table containing the characteristics of the individuals, such as age, sex, level of education, household composition, etc. Each record contains characteristics of an individual and a weight (IPONDI) that gives a measure of the frequency with which every record (profile of the individual) is found in the population in a certain area. Records also contain the 5-digit code of the census area. Small census areas are grouped into larger areas and the attributes are given at a coarser resolution.

LOGEMT (Logement)
Every record corresponds to an ordinary dwelling described by its location and characteristics (category, type of construction, comfort, surface area, number of rooms, etc.), and the socio-demographic characteristics of the household residing there.

Data sets description and preparation

MOBPRO (Mobilités professionnelles)
MOBZELT (Fichier Activité professionnelle des individus (localisation à la zone d'emploi du lieu de travail))
These 2 tables provide information about professional mobility. Each record corresponds to an individual described according to the characteristics of their professional category, trips to workplace, main socio-demographic characteristics, as well as those of the household to which they belong.

MOBSCO (Mobilités scolaires)
Mobility related to education. Each record corresponds to an individual described according to the characteristics of their trips to an education institute, main socio-demographic characteristics, as well as those of the household to which they belong.

BPE (Base Permanente des équipements)
Location of points of interest such as shopping, education, healthcare, leisure, sport etc.

Data sets description and preparation

Additional datasets

- Map of the **census units** used by the INSEE
- **Cadastral data** cross linked with geographic data form the French Geographic Institute (**IGN**) and **OpenStreetMap** to create a detailed map with the distribution of dwellings by type
- Location of **educational establishments** extracted from the Ministry of Education
- Location of **economic activities** obtained by cross referencing the data from MOBZELT which covers 64 different economic activities with the buildings for the OpenStreetMap database

Method

Because of the **large amount** and **size** of data, the method has been designed from the beginning to run in parallel

We prototyped on a sample census unit, and later extended the method to all the units running the calculations in parallel for each census unit

According to this design, in our workflow, the 5 digit **code** of the census unit is given as input

IRIS units include 2,000+ inhabitants.

“Special” cases:

Metropolitan areas of Paris, Lyon and Marseille, for which we used the ARM code (*Arrondissement*) in place of the IRIS.

Units with **less than 200** inhabitants: one IRIS represents 2+ units to reach at least 200 inhabitants

33k+ “**unmapped communes**”: for units with number of inhabitants between 200 and 2k, stats are given at coarser resolution: IRISes are grouped, statistics are given at “*Cantons ou ville*” level

Work Flow

List of census units

INDCVI

LOGEMT

Combined MOBPRO MOBZELT

MOBSCO

CENSUS units

Buildings

Economic activities

Educational establishments

POIs

Enrichment of people profile with workplace data

Adding attributes to students

Individuals unweighting

House – Family mapping

Treating Small IRISes

INDCVI table contains a weight (IPONDI) that gives the “frequency” with which every record is found.
The profile of an individual is characterised by a set of n features:

$INDCVI = \langle f_1, f_2, ..., f_n \rangle$

The MOB table (combination MOBPRO+MOBZELT)

$MOB = \langle f_{wm1}, f_{wm2}, ..., f_{wmM} \rangle$

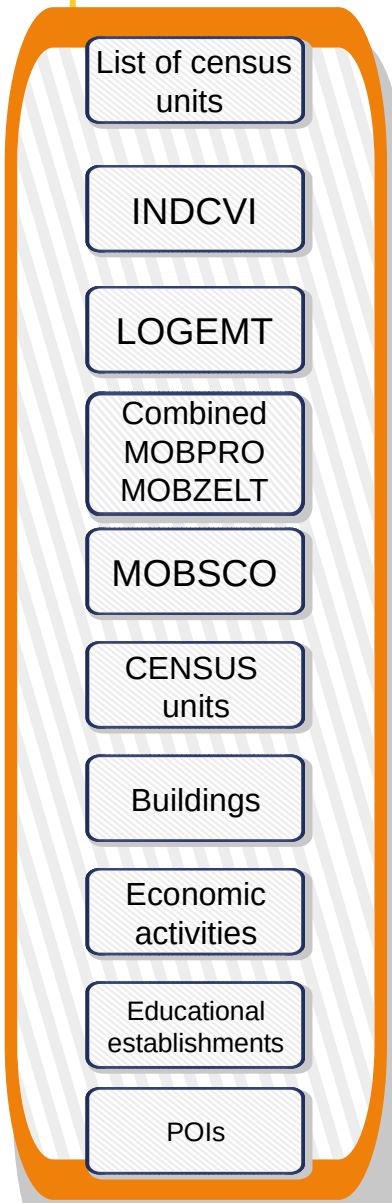
$merging\_keys = INDCVI \cap WORK\_MOB$

$added\_attributes = WORK\_MOB \text{ XOR } INDCVI$

$INDCVI' = INDCVI \cup added\_attributes$

Based on DCLT attribute and professional category, associate the location of workplace

Work Flow



Enrichment of people profile with workplace data

Adding attributes to students

Individuals unweighting

House – Family mapping

Treating Small IRISes

merging_keys = INDCVI' ∩ MOBSCO

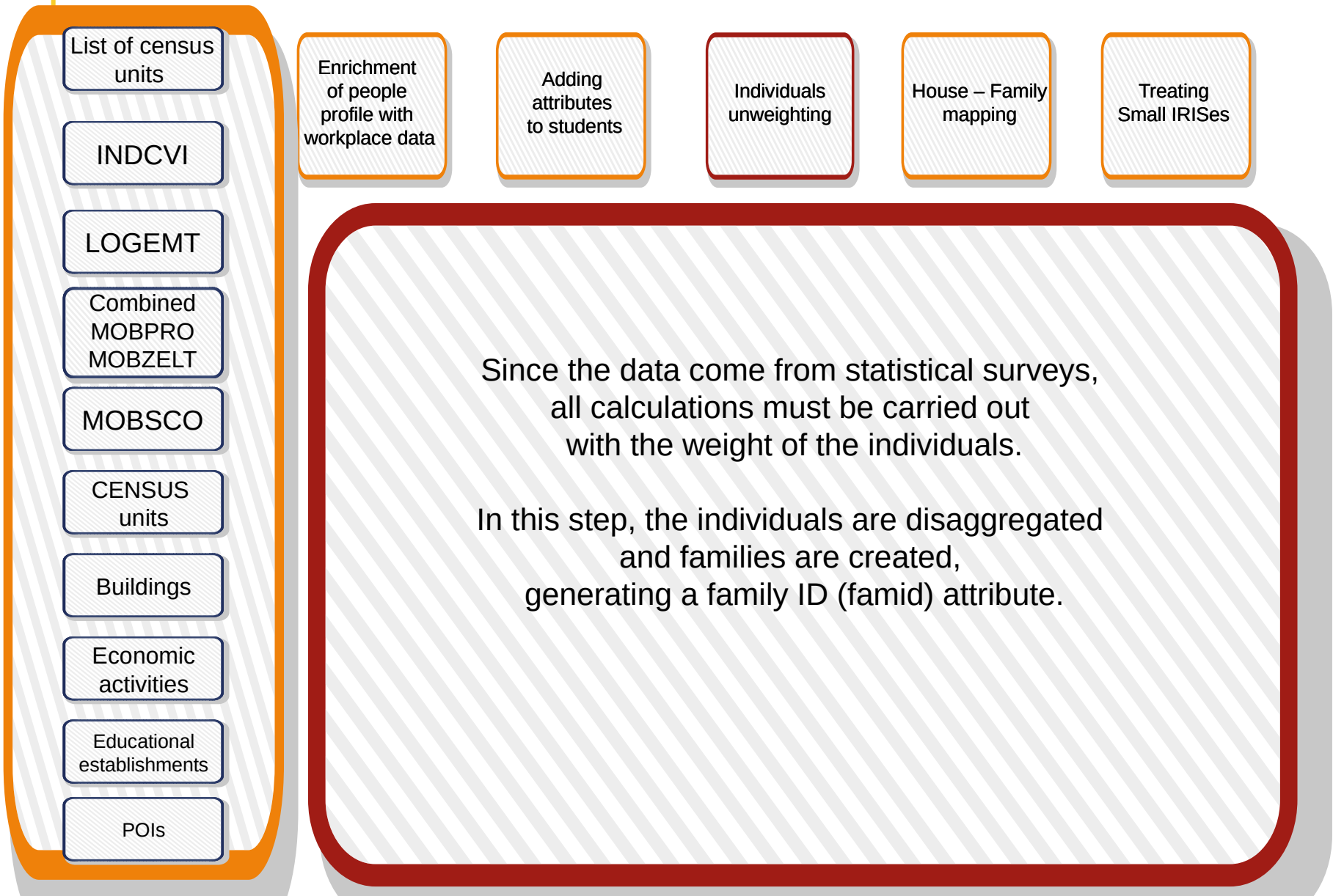
added_attributes = MOBSCO XOR INDCVI'

INDCVI'' = INDCVI' ∪ added_attributes

additional merging keys were also generated from other variables

Using DCETUF (location of school) combined with age,
it is possible to shortlist a set of educational institutes.
The selection among the candidates is operated on a random basis

Work Flow


List of census units
INDCVI
LOGEMT
Combined MOBPRO MOBZELT
MOBSCO
CENSUS units
Buildings
Economic activities
Educational establishments
POIs
Enrichment of people profile with workplace data
Adding attributes to students
Individuals unweighting
House – Family mapping
Treating Small IRISes
Since the data come from statistical surveys, all calculations must be carried out with the weight of the individuals.
In this step, the individuals are disaggregated and families are created, generating a family ID (famid) attribute.

Work Flow

List of census units

INDCVI

LOGEMT

Combined MOBPRO MOBZELT

MOBSCO

CENSUS units

Buildings

Economic activities

Educational establishments

POIs

Enrichment of people profile with workplace data

Adding attributes to students

Individuals unweighting

House – Family mapping

Treating Small IRISes

Map families into houses.
Combinatorial optimization problem:
Variable Size Multiple Knapsack.

Trade-off between precision and computational intensity.
Better precision only possible when the input data add up useful information.

Any additional attribute to houses, e.g. year when built, would make people positioning much more precise.
In lack of better information,
we assumed larger families in larger housing surfaces.

Work Flow

- List of census units
- INDCVI
- LOGEMT
- Combined MOBPRO MOBZELT
- MOBSCO
- CENSUS units
- Buildings
- Economic activities
- Educational establishments
- POIs

Enrichment of people profile with workplace data

Adding attributes to students

Individuals unweighting

House – Family mapping

Treating Small IRISes

IRISes are defined small if cover less than 200 people.

Incorporate adjacent census areas.

The Parallel Processing

The Challenge:

Modelling

~64+ million people
~35 million households
~10 million houses
~35k jobs

Very large data sets (e.g. 12GB)

Including travel to:

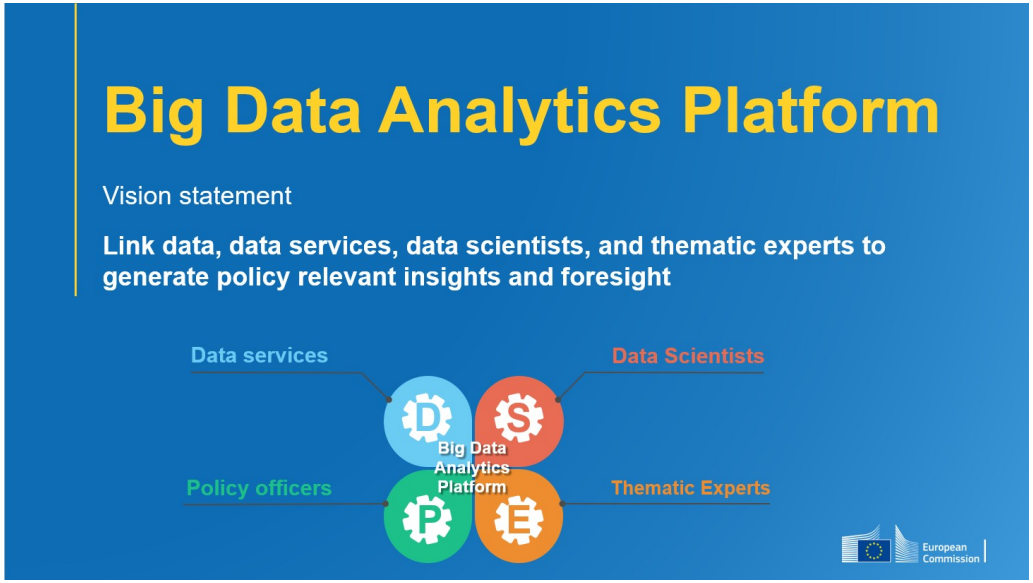
- Work
- Study places
- Shopping
- Leisure
- Healthcare
- Sport
- Etc.

Scripts in Bash and Python

Libraries:
Numpy, Pandas, geoPandas, Shapely

The Parallel Processing

The computations were performed in batch processing on the **JRC Big Data Analytics Platform (BDAP)**, that uses **HTCondor** as a job scheduler and Docker Universe set up.



~35k jobs
one for each French commune
each job taking 1 CPU

20 servers of 40 CPUs each and 1TB RAM

Machine set shared with other users

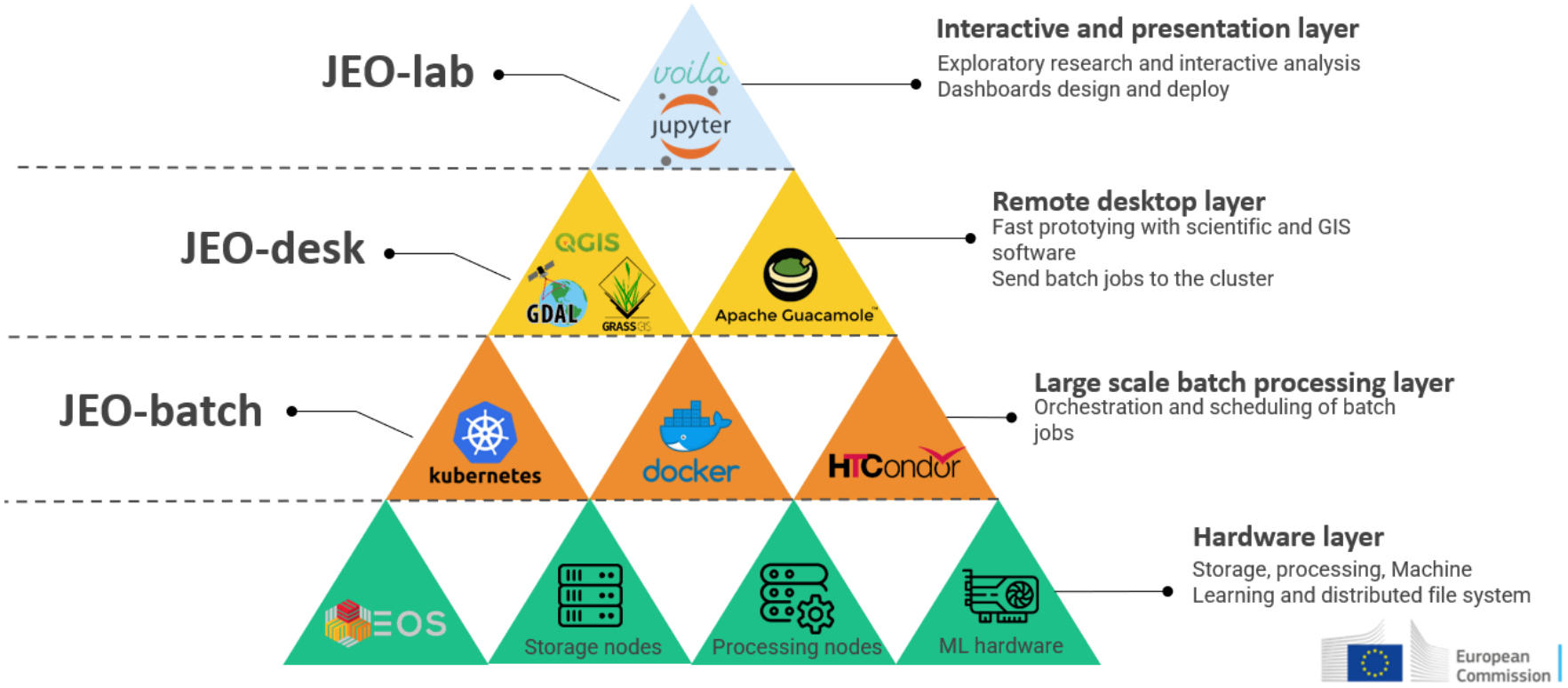
Relatively unlimited storage space

Data storage based on CERN EOS distributed filesystem

BDAP home page:
https://jeodpp.jrc.ec.europa.eu/services/shared/home/

BDAP reference publication:
https://doi.org/10.1016/j.future.2017.11.007

JRC Big Data Platform main services



Building blocks

Security

- Encrypted protocol (https)
- Multi-factor authentication

Hardware

- Petabyte scale data hub
- Co-located with computing cluster

Web-based access

- JupyterLab
- Remote data science desktop
- Distributed computing
- Development environments

Interactivity

- Novel interactive data analysis
- Exploratory visualisation tools

Machine Learning

- Specialized hardware
- Artificial Intelligence and Deep Learning modelling

Dissemination

- Web-based data dissemination
- Visualisation services
- Secure file transfer (ftps)

European Commission

Lessons learned

1) **Metadata overhead:** Writing several output files in the same folder, or writing several small files too quickly, not suitable for CERN EOS filesystem

Solution:
Run batches of ~6k jobs

Write output of each job on physical disk and move output at the end

Lessons learned

2) Dealing with **very large** (~12GB) **CSV files in input**.

Opening in Pandas, even for subsetting the relevant input for each job, would require memory demand of 200 GB in Condor submit file

Machines were seldom allocated to our jobs (35k jobs!)

Approach #1

Open files once, subset them once for all jobs.

Drawback: inconvenient, also due to challenge 1)

Approach #2

Find a tool for subsetting that reads file without loading

Lessons learned

Tested:

CSVgrep (CSVtools https://ctan.org/pkg/csvtools)

XSV (https://github.com/BurntSushi/xsv)

AWK (https://en.wikipedia.org/wiki/AWK)

Best speed
Inferior memory requirement

Thank you