atmosphere-eubrazil.eu @AtmosphereEUBR

ATMOSPHERE

Adaptive, Trustworthy, Manageable, Orchestrated, Secure Privacy-assuring Hybrid, Ecosystem for REsilient Cloud Computing

Machine Learning Pipelines on Medical Imaging

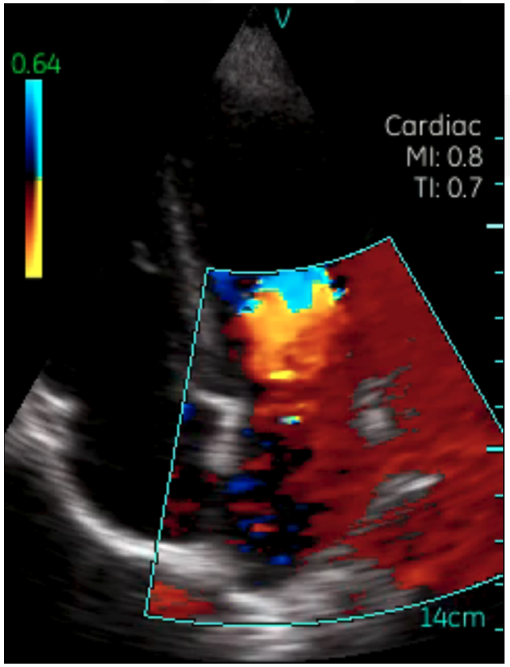
Ignacio Blanquer (UPV), Eduardo Camacho-Ramos, Ana Jiménez-Pastor, Ángel Alberich-Bayarri (QUIBIM), Walter Dos Santos, Prof. Wagner Meira Jr. (UFMG)

Co-funded by the European Commission
Horizon 2020 - Grant #777154

European Commission

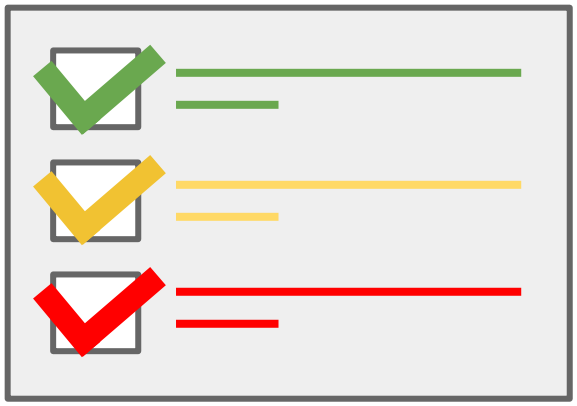
The problem

- The Rheumatic Heart Disease (RHD) is a disease that can be easily treated in its early stages, but may produce enormous damage to the heart if remains untreated, including severe sequelae and death.
- The challenge is to process a large set of medical images, along with additional metadata and clinical information, efficiently and securely, to extract features that could be used to assist and even automate diagnosis.
- Data comes from the PROVAR Echocardio data
  - 4.021 studies (4.035 Normal + 180 Borderline + 26 Definite)
  - 59.018 240×320 MP4 videos of 1-3 seconds.
  - To be classified into three categories according to the WHF criteria: Normal, Borderline and Definite RHD.
- Challenges:
  - Unbalance of the cases.
  - Noise and low quality of the echocardio images.
  - No information on the view.



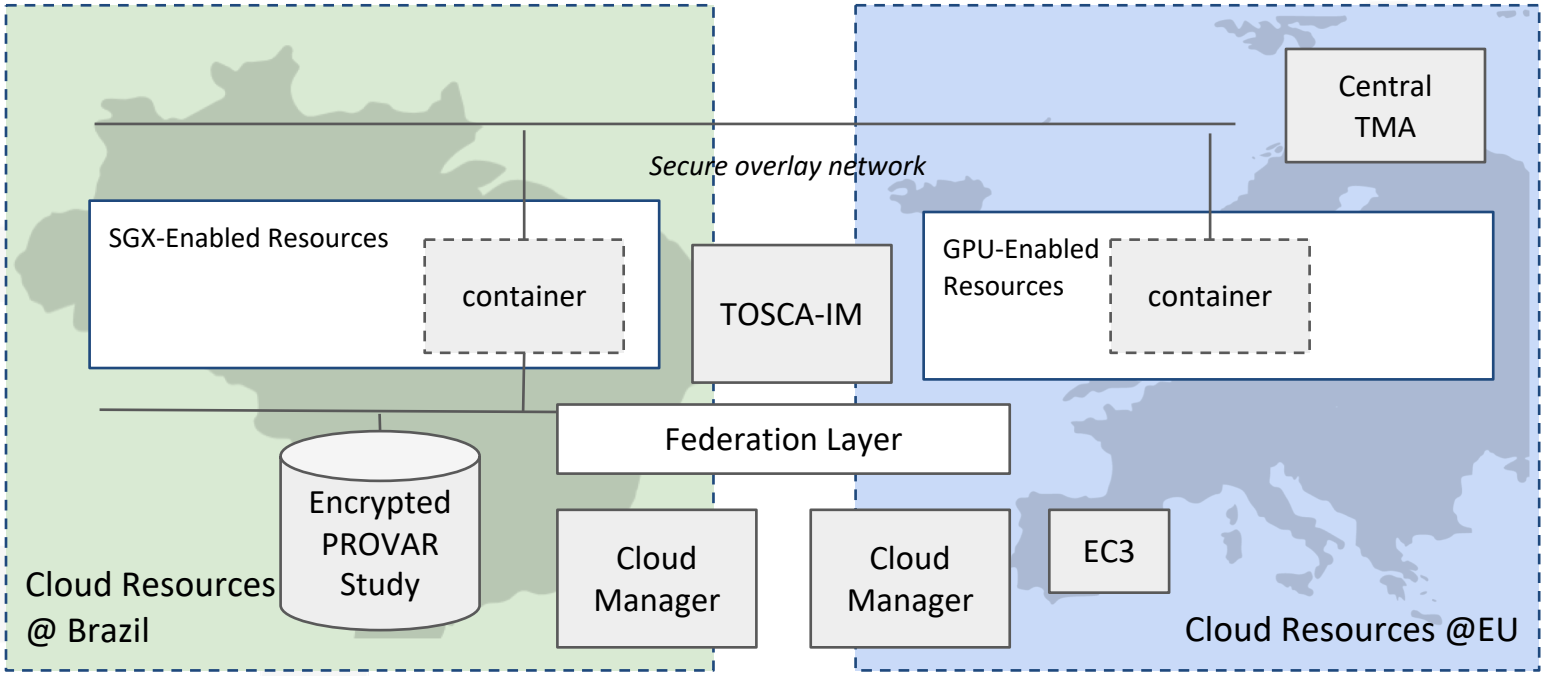
Requirements

- Sensitive data should remain in the Brazilian geographical boundaries and confidentiality should be preserved.
- Computing requires accelerators and may not be available within the boundaries where the sensitive data is located.
- Parallel execution should be provided.
- Repeatability and reproducibility should be a main goal.
- Flexible and dynamic environment.
- Simplified interfaces for non-ICT experts.



Intercontinental infrastructure

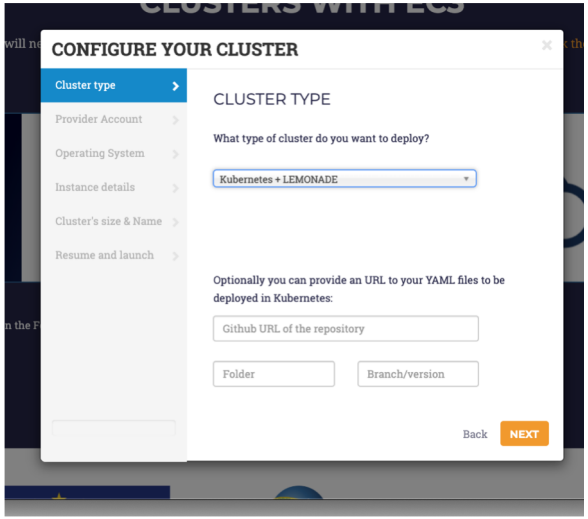
- The underlying infrastructure is a federated cloud
  - Using fogbow (www.fogbowcloud.org) on OpenStack and OpenNebula.
  - With a Federated Network to provide a coherent network space among nodes.
  - Heterogeneous resources: SGX-enabled and GPU nodes.
- Using EC3(1) and Infrastructure Manager(2) to deploy a virtual infrastructure.

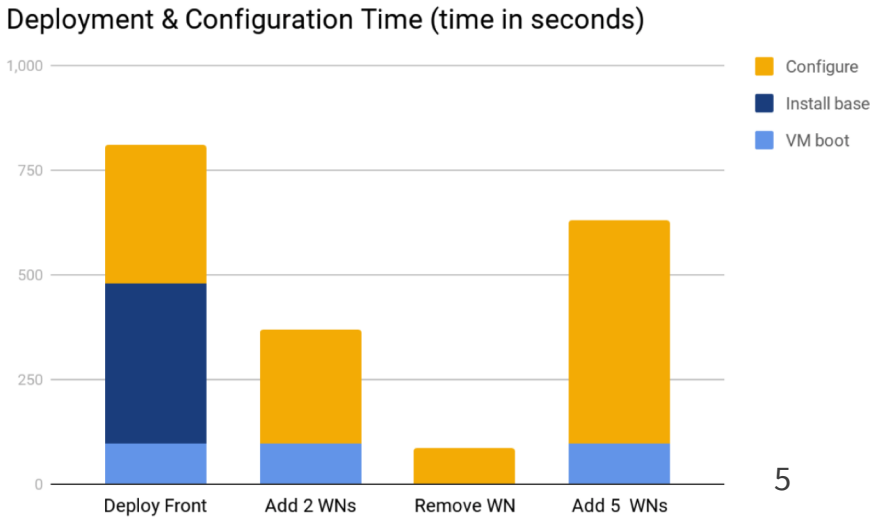


(1) https://marketplace.eosc-portal.eu/services/elastic-cloud-compute-cluster-ec3
(2) https://marketplace.eosc-portal.eu/services/infrastructure-manager-im

Deployment of the virtual infrastructure

- The virtual infrastructure is managed by an elastic Kubernetes cluster spawn over the federated network
  - Containers and services are accessible from both sites but only through the federated network.
  - Resources are properly tagged (SGX and GPU capabilities and Brazil / Europe) so K8s applications are placed in the correct resource.
  - Infrastructure is described as code(3).
- K8s Front-end is deployed and nodes are being powered on as the applications are deployed, creating the request for specific resources.

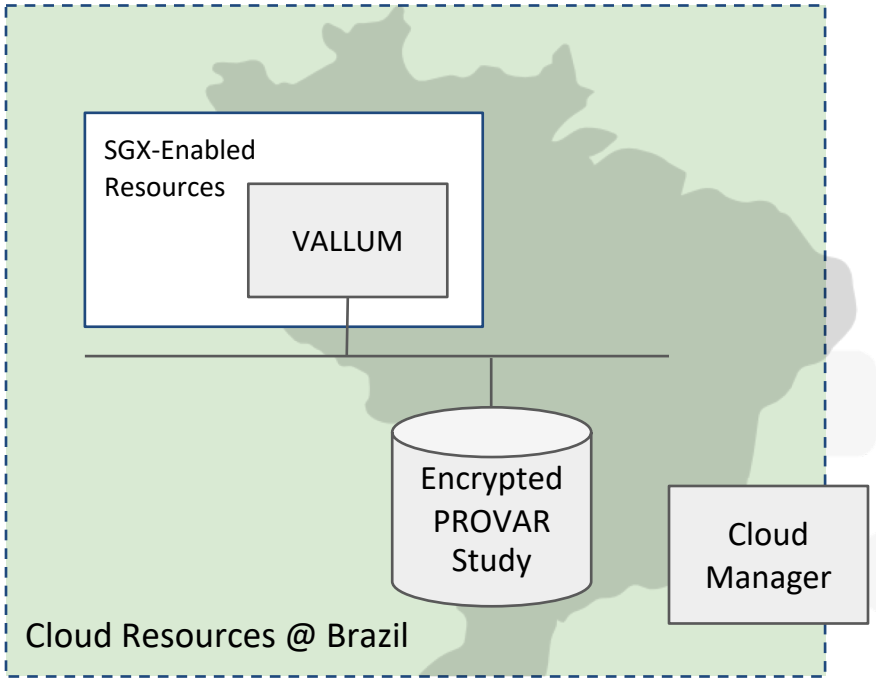




(3) https://github.com/grycap/ec3/tree/atmosphere

Secure storage at Brazilian side

- A secure storage is deployed at the Brazilian side
  - It uses Vallum[4], a service that provides on-the-fly annonymisation based on policies.
  - It masks (or blurs) the fields that are marked as sensitive to different profiles of users.
  - It relies on an HDFS filesystem for the files and on SQL databases for the structured data.

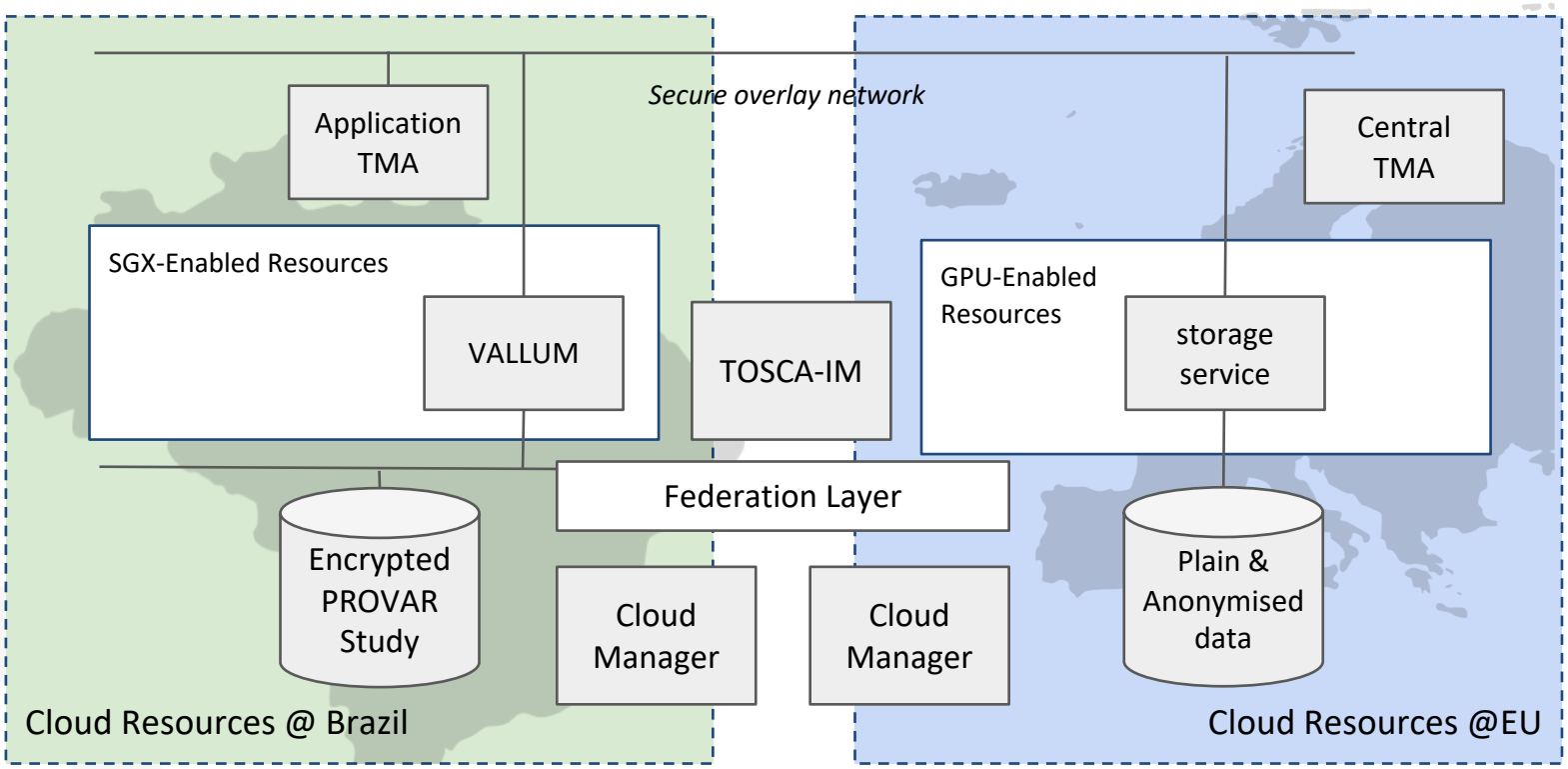


- It runs the data anonymisation and sensitive data access on enclaves running on SGX-enabled containers, so they can securely run even in untrusted cloud resources
  - Data remains encrypted in disk.

(4) https://www.atmosphere-eubrazil.eu/vallum-framework-access-privacy-protection

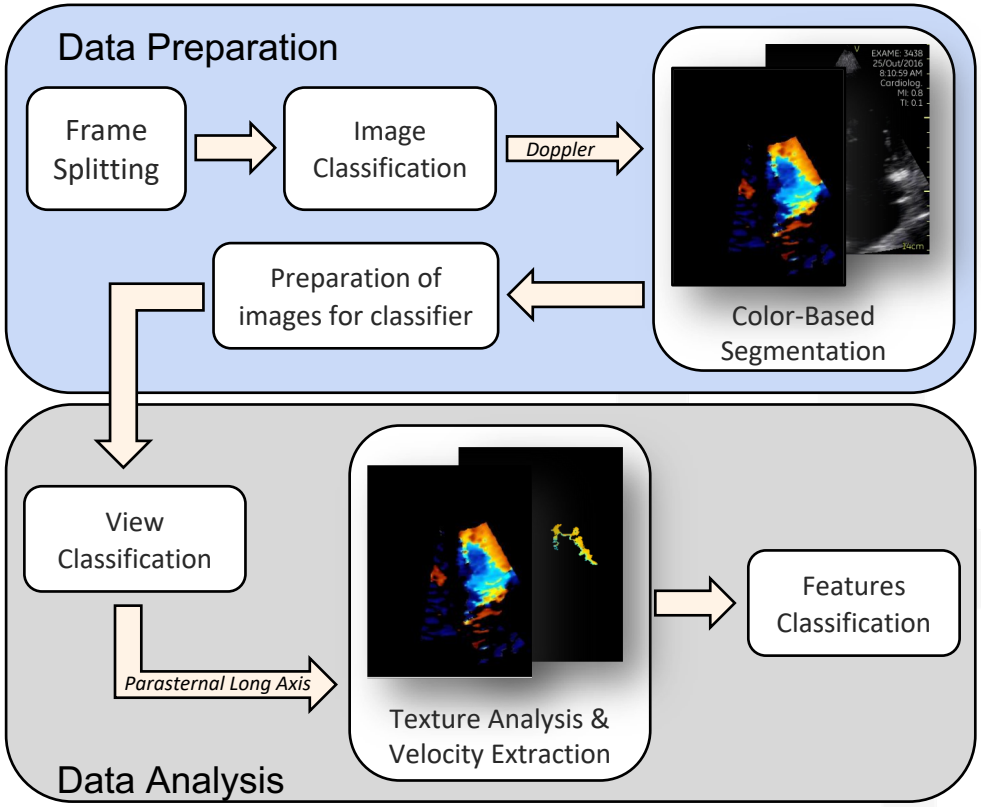
Anonymised Data

- Data is requested to Vallum from external users, but they will only access to partially anonymised data
  - Anonymised data (~1TB) is copied where the computing accelerators are placed.



Building the models for the Estimation pipeline.

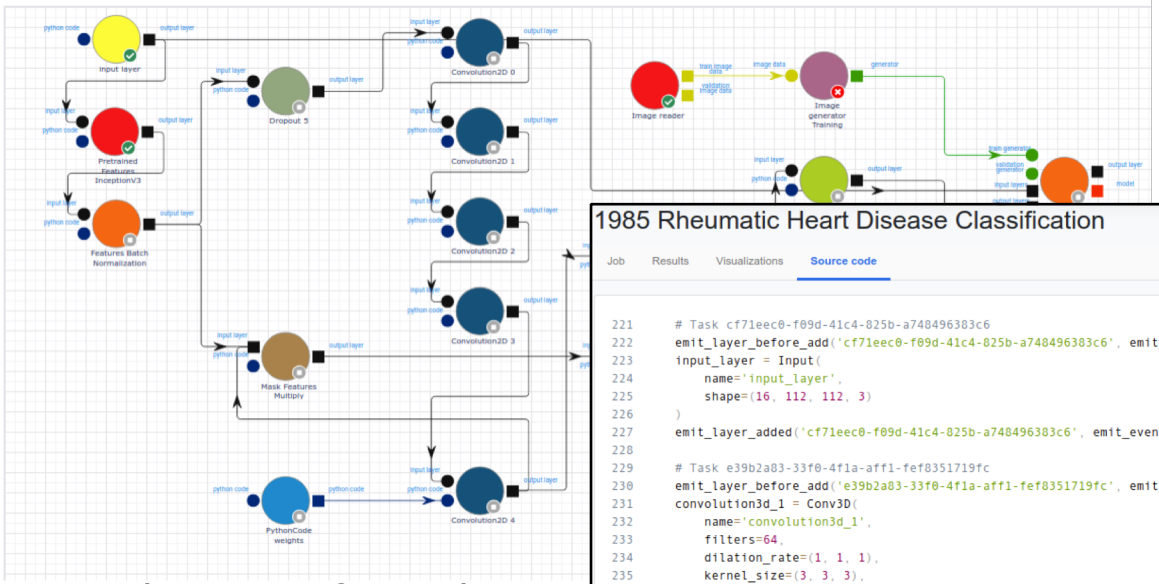
- Videos are split into frames and classified by color inspection.
  - A color-based segmentation using k-means clustering extracts the color pixels from the Doppler images
- Images are classified according their acquisition view using a CNN
  - Parasternal long axis view has proven to be relevant to obtain an accurate classification.
- First & second order texture analyses characterize the images by the spatial variation of pixel intensities.
  - Besides texture features, blood velocity information is also obtained.



- Finally, all the extracted features are classified through machine learning techniques in order to differentiate between RHD positive and healthy subjects.

Coding the pipeline: LEMONADE

- The pipeline is developed using LEMONADE[5]
  - LEMONADE provides a GUI and a Machine Learning librarie to develop data analytics pipelines.
  - Pipelines can be run interactively or transformed into executable code.
  - Code can be interactively run or further embed into services to be exposed for production.
- A model building pipeline and an estimation pipeline are developed.



1985 Rheumatic Heart Disease Classification

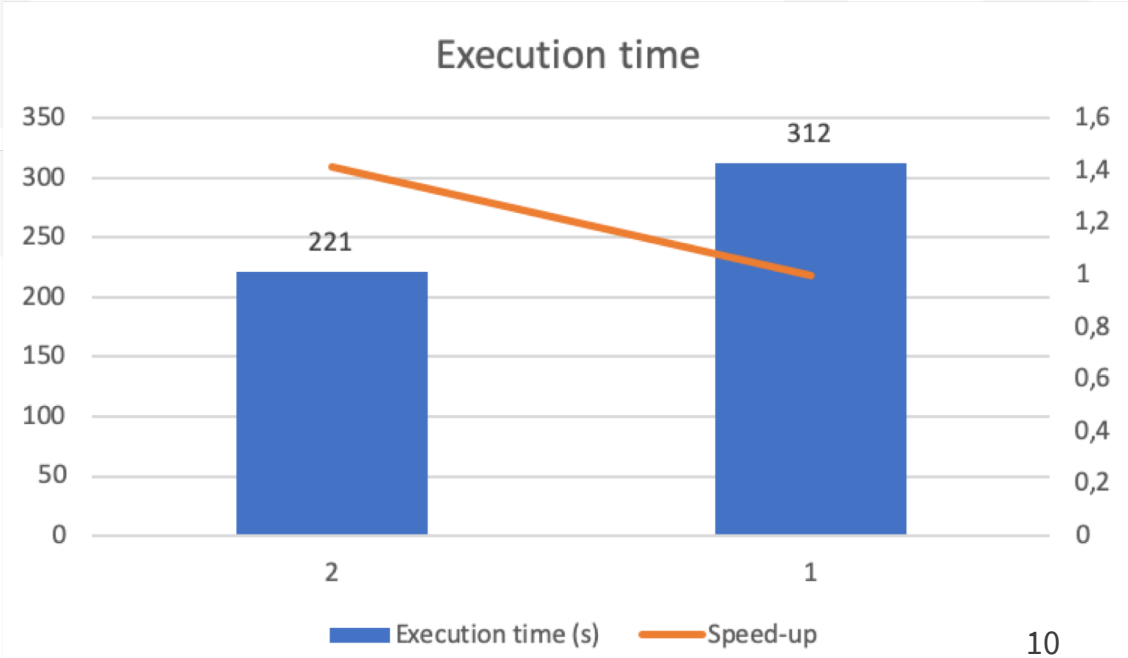
Job | Results | Visualizations | Source code

```
# Task cf71eec0-f09d-41c4-825b-a748496383c6
emit_layer_before_add('cf71eec0-f09d-41c4-825b-a748496383c6', emit_event)
input_layer = Input(
    name='input_layer',
    shape=(16, 112, 112, 3)
)
emit_layer_added('cf71eec0-f09d-41c4-825b-a748496383c6', emit_event)

# Task e39b2a83-33f0-4f1a-aff1-fef8351719fc
emit_layer_before_add('e39b2a83-33f0-4f1a-aff1-fef8351719fc', emit_event)
convolution3d_1 = Conv3D(
    name='convolution3d_1',
    filters=64,
    dilation_rate=(1, 1, 1),
    kernel_size=(3, 3, 3),
    use_bias=True,
    padding='same',
    activation='linear'
)(input_layer)
convolution3d_1.trainable = True
emit_layer_added('e39b2a83-33f0-4f1a-aff1-fef8351719fc', emit_event)

# Task ffbc14e3-1785-4333-bb57-e73633047d49
emit_layer_before_add('ffbc14e3-1785-4333-bb57-e73633047d49', emit_event)
maxpooling3d_1 = MaxPooling3D(
    name='maxpooling3d_1',
    pool_size=(1, 2, 2),
    strides=(1, 2, 2),
    padding='valid',
    trainable=True
)(convolution3d_1)
emit_layer_added('ffbc14e3-1785-4333-bb57-e73633047d49', emit_event)

# Task 7caa5f93-ef38-49ab-bdd4-1df83ed97f4b
emit_layer_before_add('7caa5f93-ef38-49ab-bdd4-1df83ed97f4b', emit_event)
convolution3d_2 = Conv3D(
    name='convolution3d_2',
    filters=128,
    dilation_rate=(1, 1, 1),
    kernel_size=(3, 3, 3),
    use_bias=True,
```

[5] https://www.atmosphere-eubrazil.eu/lemonade-live-exploration-and-mining-non-trivial-amount-data-everywhere

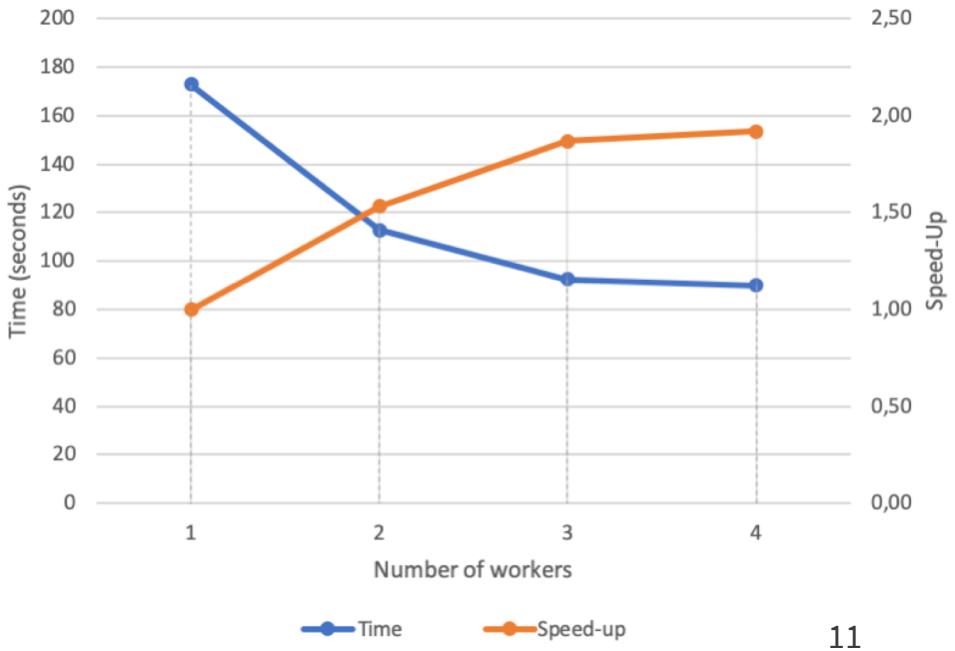
Model building in parallel

- Model building can run in parallel using MPI and Horovod
  - The model is build with keras using fp16 compression for the reduction operations.
  - Experiments have been used with 1 and 2 working nodes equipped with a TESLA V100 GPU connected through PCI Passthrough to the working nodes and the containers which run the processes.
  - Execution time shows a reduction with the addition of a second GPU but the speed-up is limited by the penalty of using an overlay network.



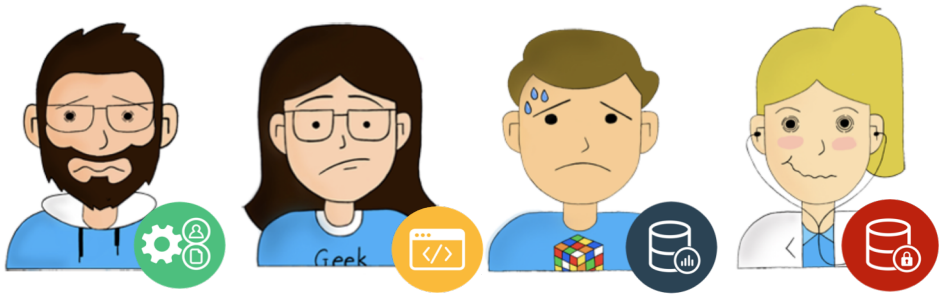
Running the estimation pipeline

- An experiment has been performed for the classification of 8 patients on 1, 2 and 4 virtual compute nodes
  - Job code is extracted and executed through Jupyter on an ipyhton cluster that shares the filesystem.
  - Each node is a Kubernetes Pod executing a Docker container in a different Virtual Machine to reduce resource contention.
  - Speed-up is moderated (up to 2) but usability is high.



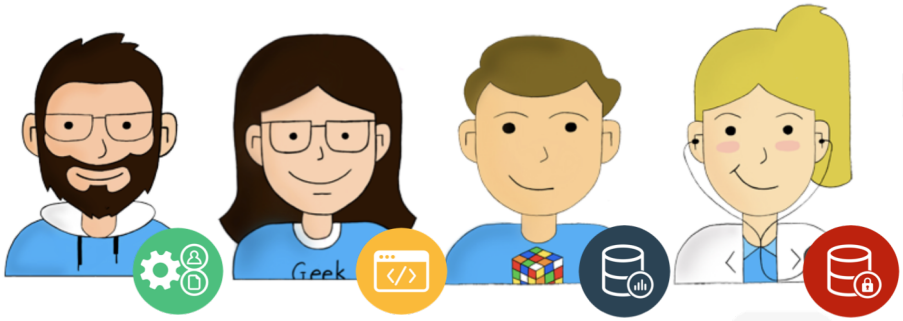
Conclusions

Before



- Need to manually configure the environment.
- Lack of reproducibility.
- Qualitative appraisal of the trustworthiness.
- Manual analysis of GDPR/LGDP risks
- Need to trust on the storage provider.
- Anonymisation level is qualitative.

After



- Applications templates for complex & distributed applications.
- Provide a repeatable way to deploy the whole application.
- Quantitative measure of trustworthiness
- Self-assessment of GDPR/LGDP.
- Trustable storage environment even on an untrusted provider.
- Quantitative anonymisation level.