# Scalability Analysis of Mini-Cluster Jetson TX2 for Training DNN Applied to Healthcare

John A. GARCÍA. H.

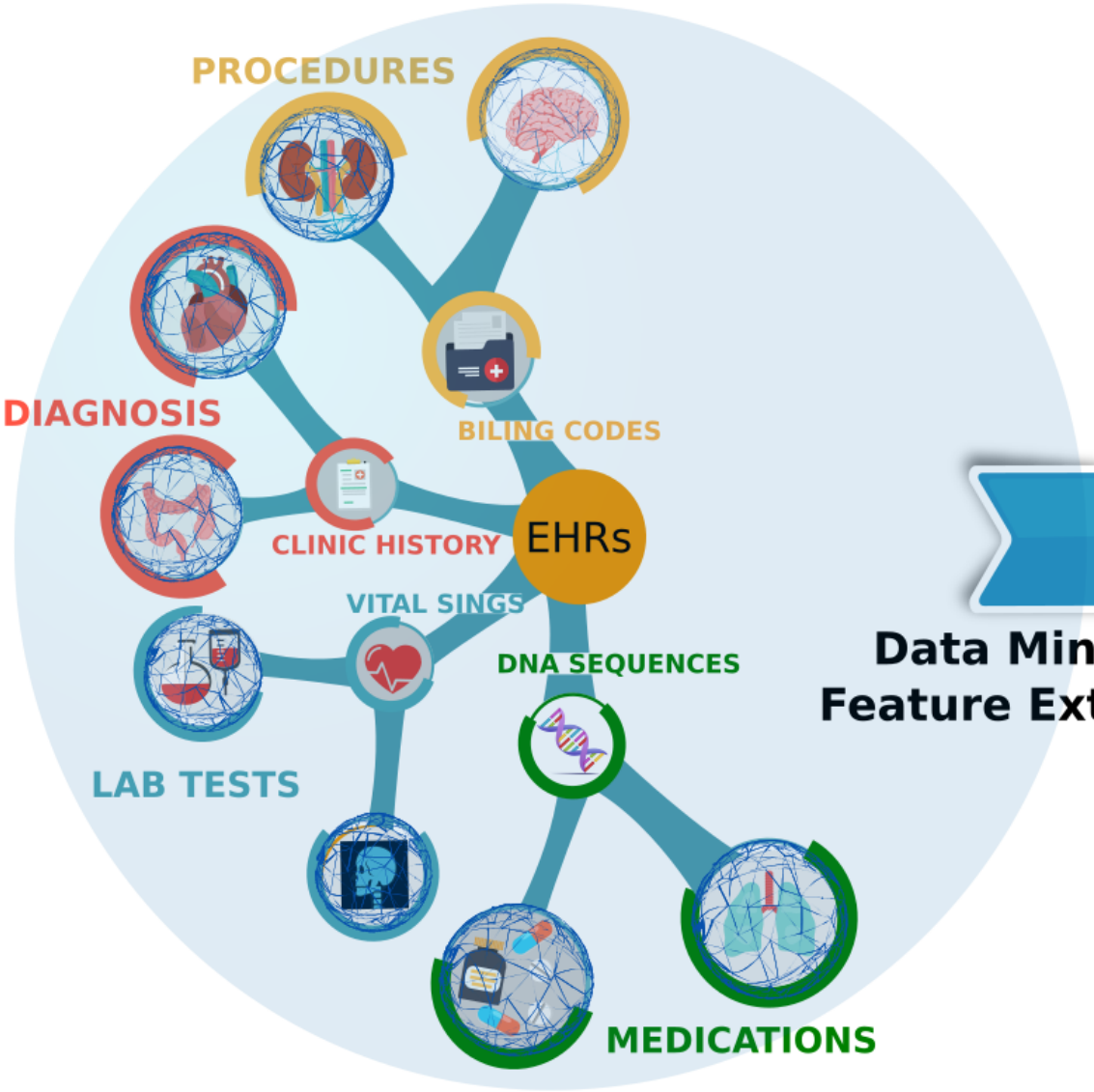
Frédéric PRECIOSO, Pascal STACCINI, Michel RIVEILL

**Université Côte d'Azur, CNRS**

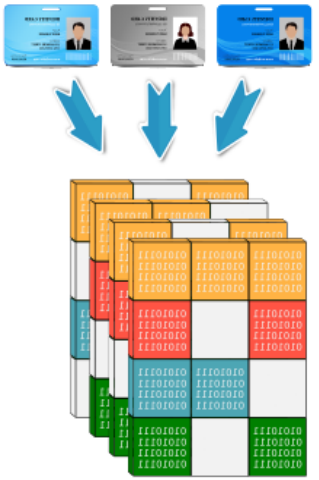
**Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis - I3S**




# Motivation: Health Care Decision-Making




**Data Mining & Feature Extraction**



*General Patient Phenotype Representation*

**Predictive Modeling**

# Challenges

- *A common challenge in healthcare today is that physicians have access to massive amounts of data on patients, but have short time to analyze all of them.*

- *One limitation is that hospitals without robust computational systems for processing, storing and drawing conclusions requires to outsource the clinical tasks and that is a risk for privacy clinical data.*

***Developing a Green Intelligence Medical System to derivate a patient representation for predict general medical targets and improving the computational resources usage.***

# DiagnoseNET

***Green Intelligence Medical System***

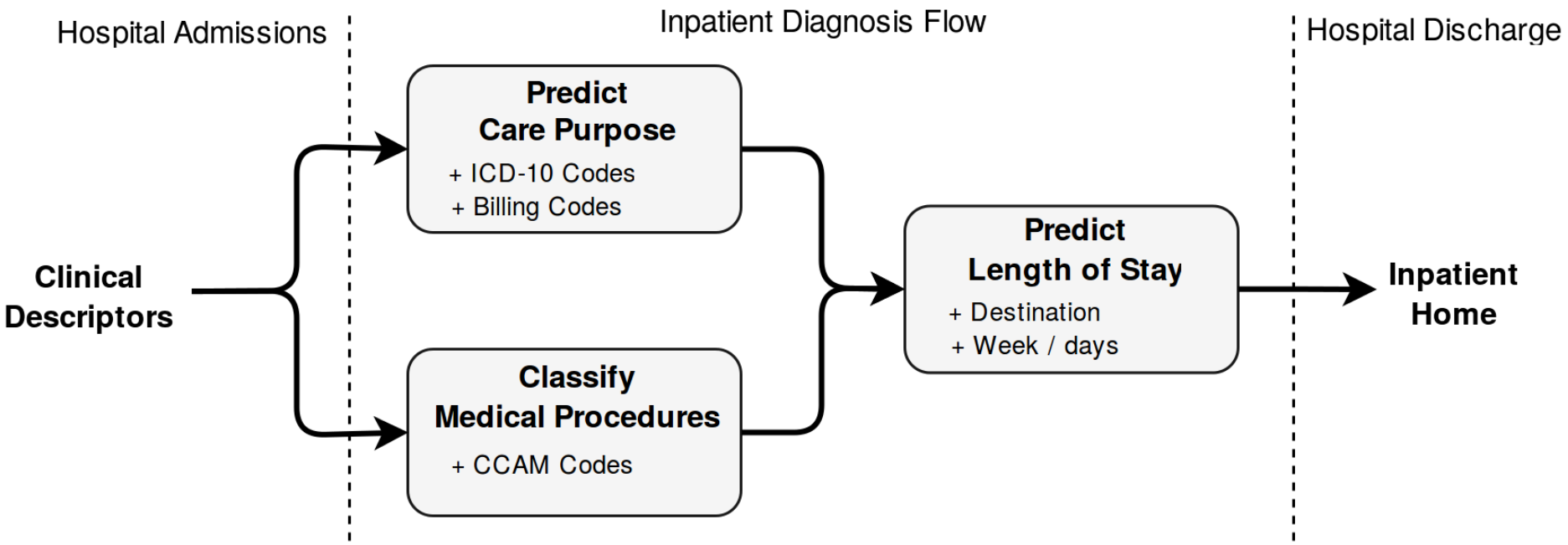
Provides three high-level features:

1) A framework to build full Deep Neural Networks (DNN) workflow;

2) A distributed processing for training DNN on Jetson TX2 Mini-Clusters;

3) An energy-monitoring tool for workload characterization.

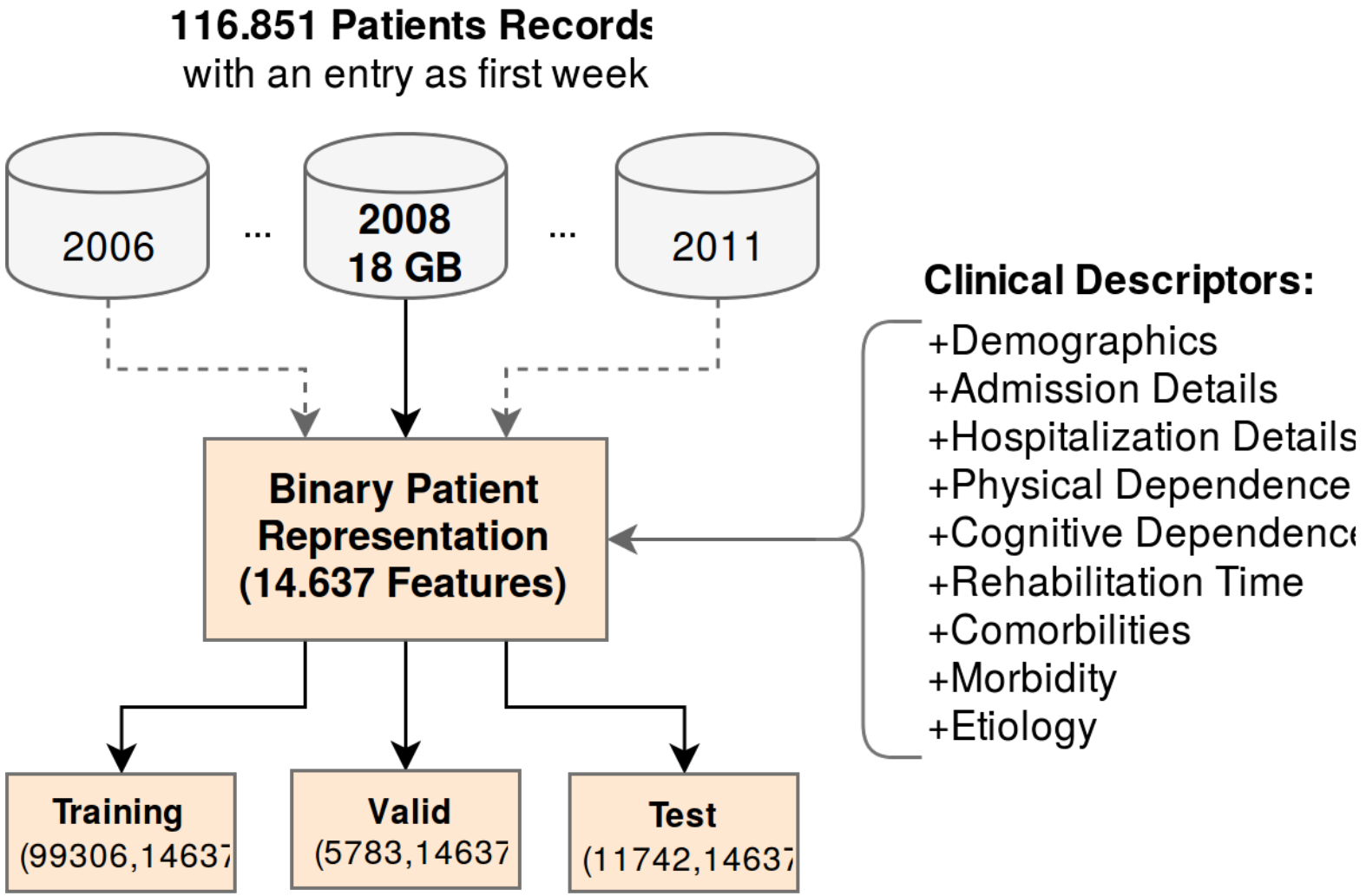
# Case Study: Predict the Medical Future of Hospitalized Patients




| | Diagnosis-related Group | ICD-10 Codes | Definition |
|---|---|---|---|
| **Patient 1** | Morbidity Principal | R402 | Unspecified coma |
| | Etiology | I619 | Nontraumatic intracerebral hemorrhage, unspecified |
| *Medical Target* | Care Purpose | Z515 | Encounter for palliative care |
| *Label used* | Clinical Major Category | 20 | Palliative care |
| **Patient 2** | Morbidity Principal | R530 | Neoplastic (malignant) relate fatigue |
| | Etiology | C20 | Malignant neoplasm of rectum |
| *Medical Target* | Care Purpose | Z518 | Encounter for other specified aftercare |
| *Label used* | Clinical Major Category | 60 | Other disorders |

# PMSI-PACA Clinical Dataset

*As Input we are using **result features** that describe the patient clinical descriptors to predict the medical targets.*



# DiagnoseNET

***Green Intelligence Medical System***

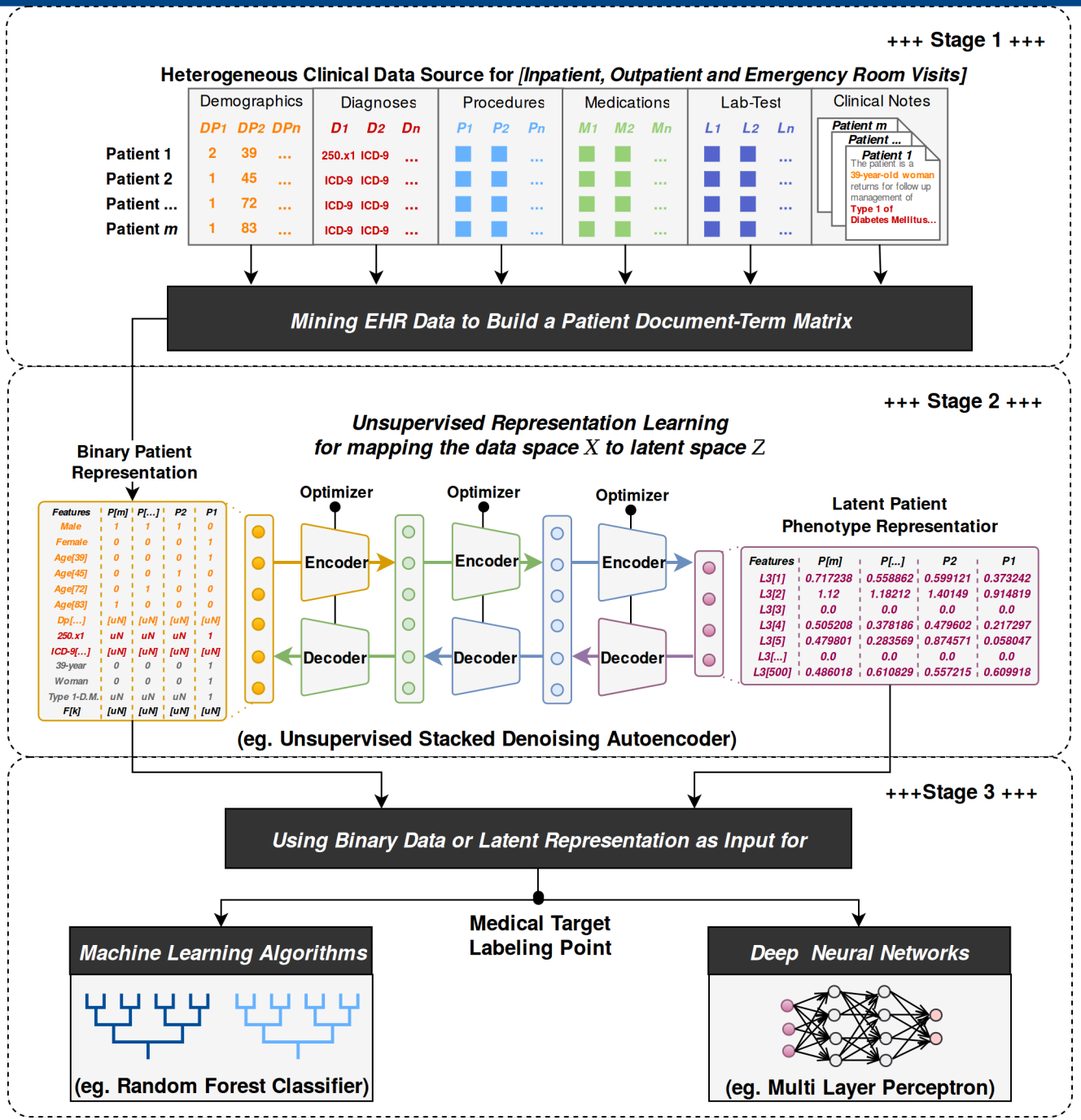
Provides three high-level features:

1) A framework to build full Deep Neural Networks (DNN) workflow;

2) A distributed processing for training DNN on Jetson TX2 Mini-Clusters;

3) An energy-monitoring tool for workload characterization.

# DiagnoseNET: Framework to automize the Patient Phenotype Representation






3. Adapted from: **Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records**. By Riccardo Miotto et al. SCIENTIFIC REPORTS, 2016.

# Mining Electronic Health Records To Build A Patient Entity-Term Matrix

# Data-mining: Feature Extraction From Electronic Health Records

## Serialized each patient record in a clinical document architecture schema

| Patients | x1_demographics | | | x4_physical_dependance | | | x7_related_diagnoses | | |
|---|---|---|---|---|---|---|---|---|---|
| | gender | ... | age | feeding | ... | displacement | Das1 | ... | Das 3 |
| Patient 1 | 2 | ... | 61 | 4 | ... | 2 | Z431 | ... | Z501 |
| Patient 2 | 2 | .. | 65 | 4 | ... | 2 | J459 | ... | F322 |
| .... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Patient m | 1 | ... | 95 | 1 | ... | 2 | C259 | ... | F322 |

## Build a binary patient phenotype representation from their features selected

| Patients | x1 :demographics | | | x4 :physical_dependance | | | x7 :related_diagnoses | | |
|---|---|---|---|---|---|---|---|---|---|
| | [1 :male] | [2 :female] | 60-74 | [4 :Assistance] | | [2 :normal_transfer] | Z431 | ... | F322 |
| Patient 1 | 0 | 1 | 1 | 1 | ... | 1 | 1 | ... | 0 |
| Patient 2 | 0 | 1 | 1 | 1 | ... | 1 | 0 | ... | 1 |
| .... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Patient m | 1 | 0 | 0 | 0 | ... | 1 | 0 | ... | 1 |

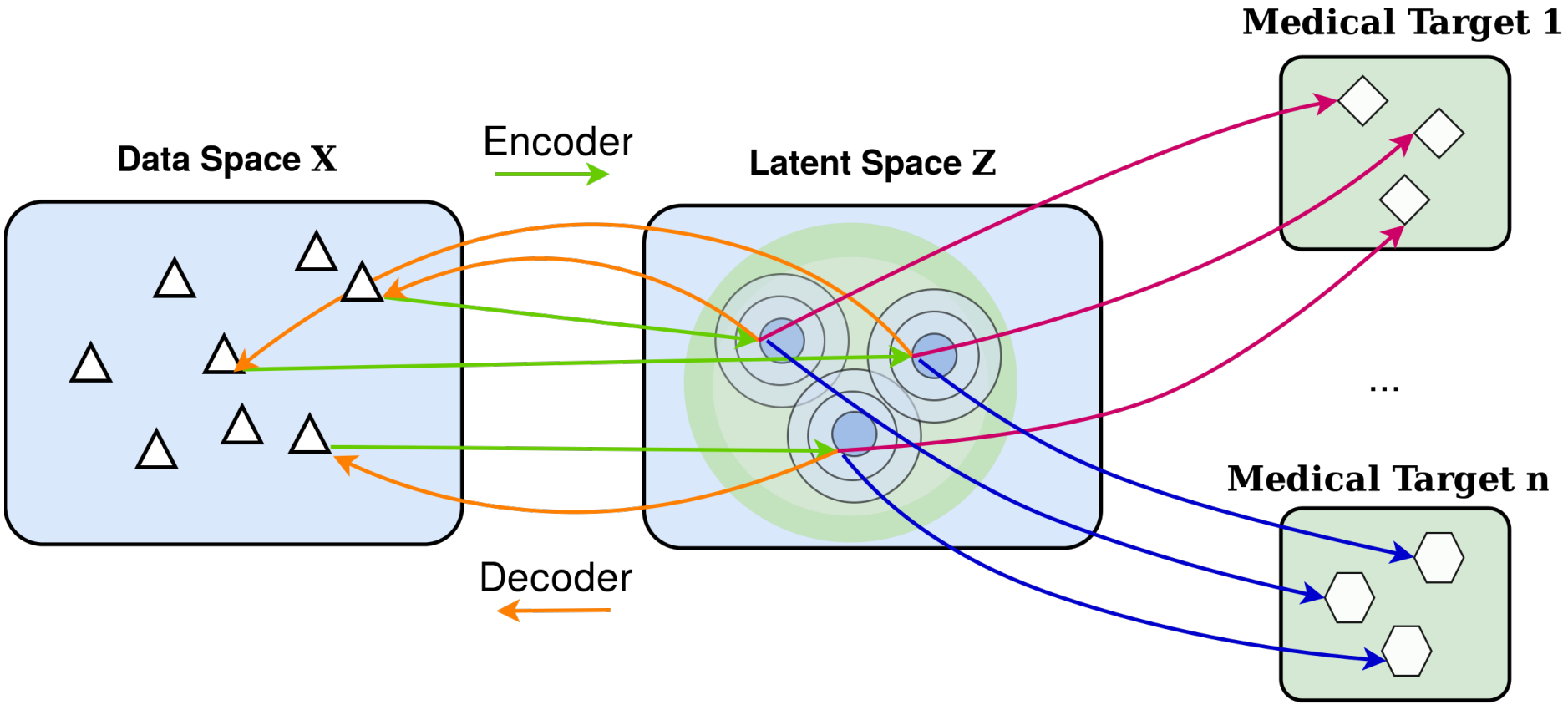
# Unsupervised Patient Phenotype Representation

# Methodology: Unsupervised Patient Phenotype Representation

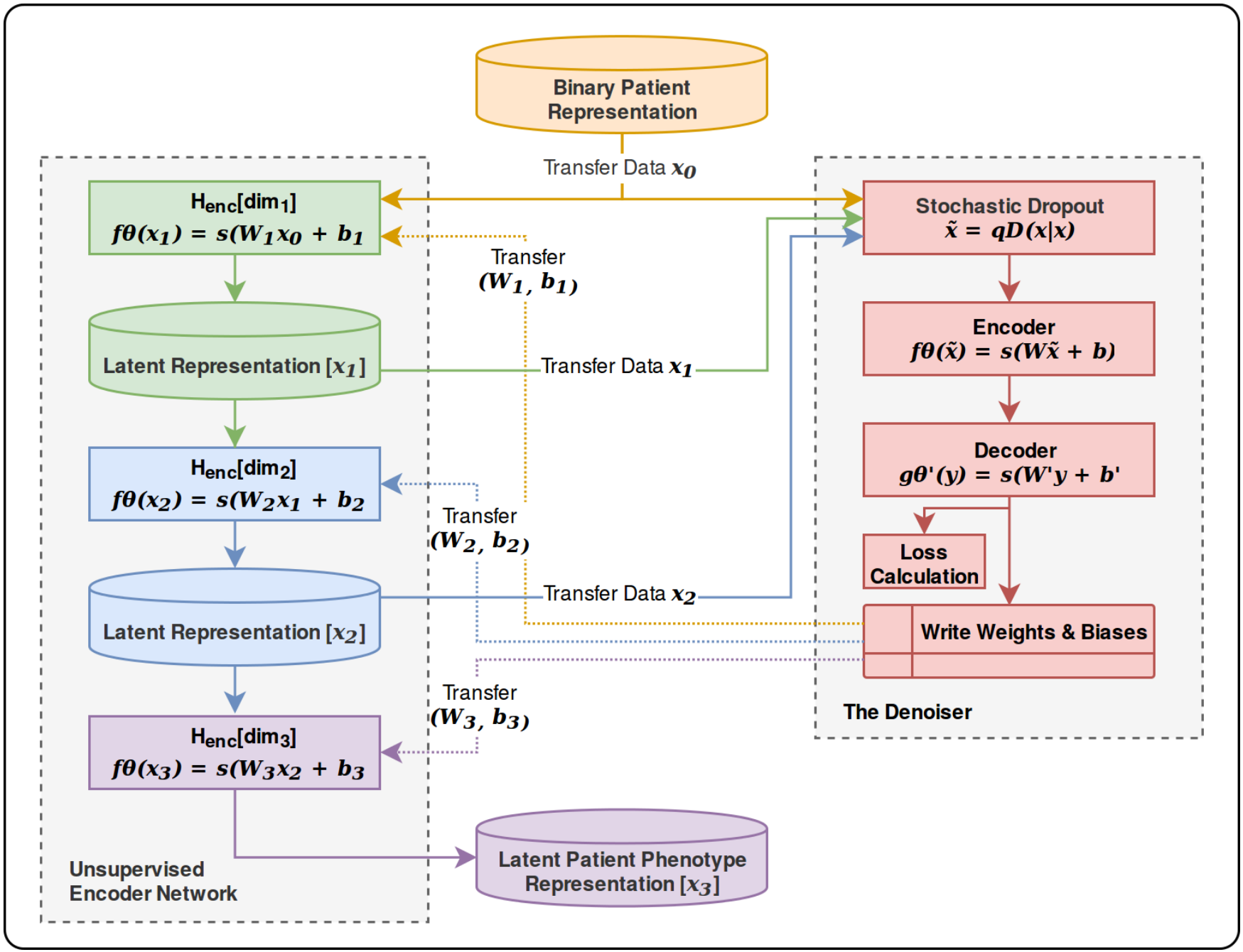
**The task:**

**+** From a binary patient representation **{X}** derive a latent patient representation **{Z}.**

**+** Using the general representation plus a supervised learning algorithms for predict different medical targets.



# Unsupervised Learning Representation





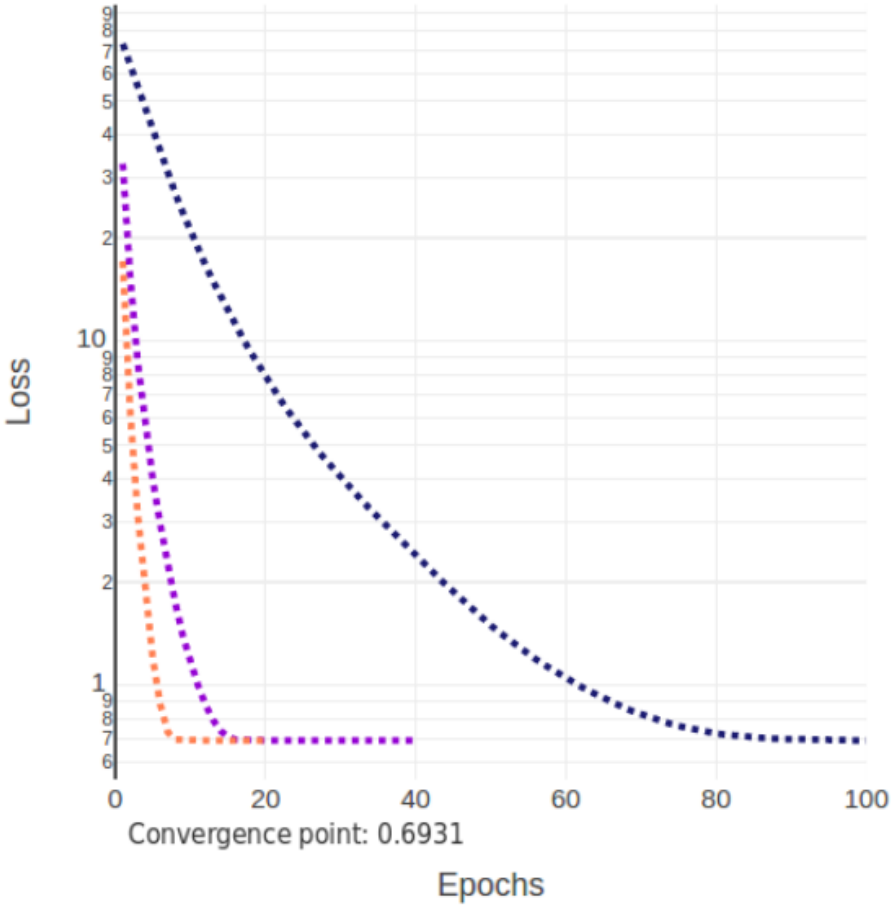

4. Adapted from **Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.** By Pascal Vincent et al. 2010

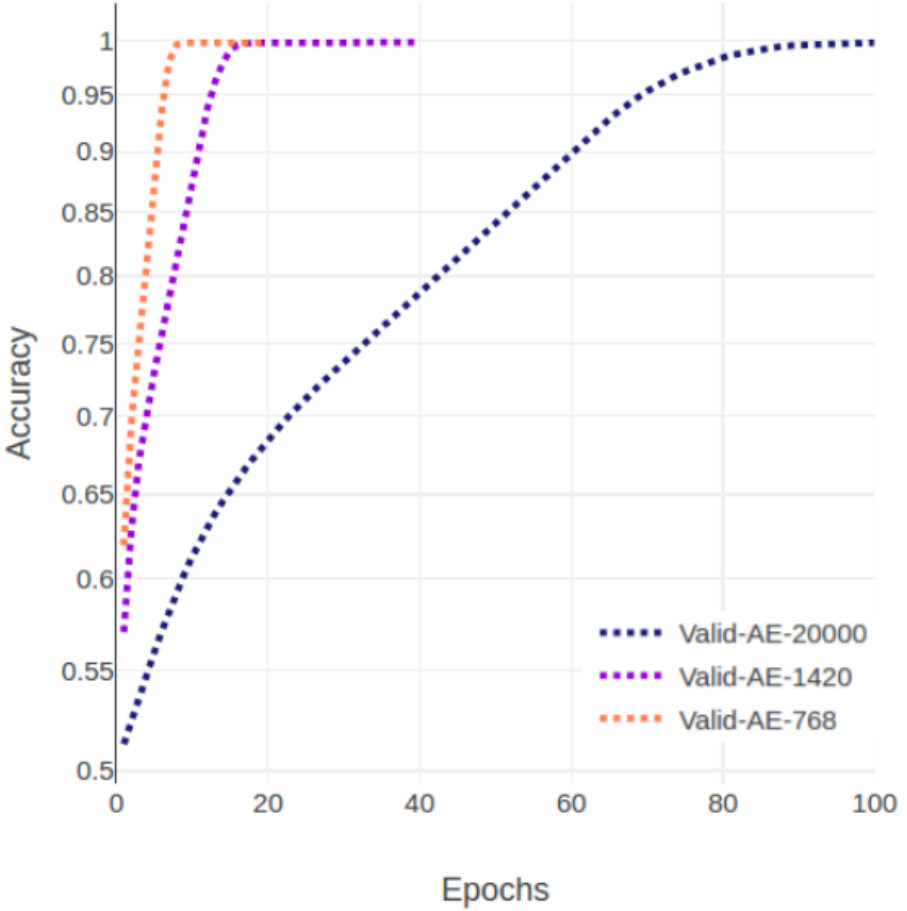
# Experiment Analysis

***1) Number of Gradient Updates as Factor to Early Model Convergence.***

# 1) Number of Gradient Updates as Factor to Early Model Convergence

- *Network convergence using batch partitions of [20000, 1420, 768] records to generate [4, 59, 110] gradient updates by epoch respectively.*

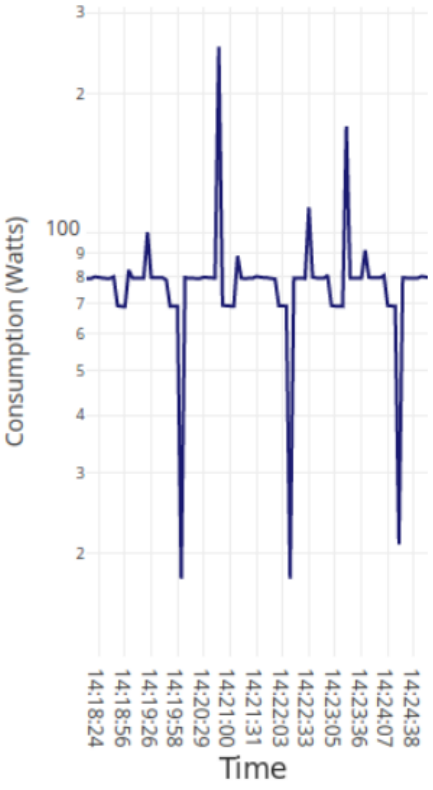



| | 1-Layer1 | 2-Layer2 | 3-Layer3 | 4-Activation_funct | 5-GD_Optimizer | 6-Learning_rate | 7-Dropout-rate |
|---|---|---|---|---|---|---|---|
| **0** | 2048 | 2048 | 768 | relu | adam | 0.0001 | 0.5 |

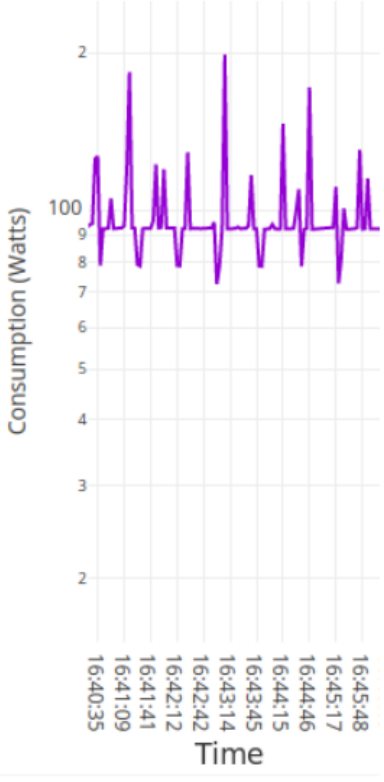
# Power consumption in a window of 6 minutes

***BF: 20.000***
***100 Epochs***
***EC: 137.65 Kj***



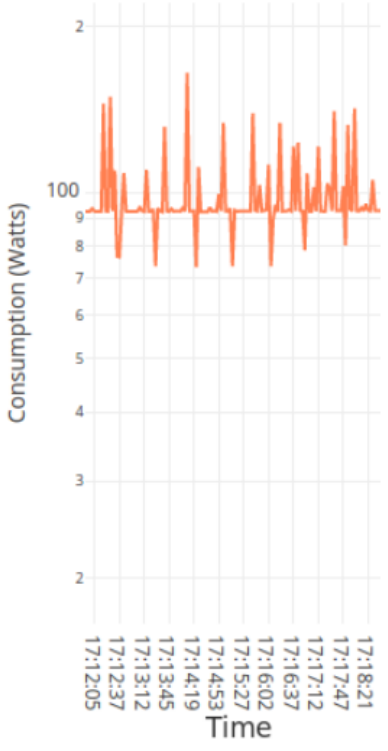

63.35 *Watts* in average to process 68 gradient updtaes in 17 epochs.

***BF: 1420***
***40 Epochs***
***EC: 41.26 Kj***




86.61 *Watts* in average to process 885 gradient updtaes in 15 epochs.

***BF: 768***
***20 Epochs***
***EC: 21.87 Kj***



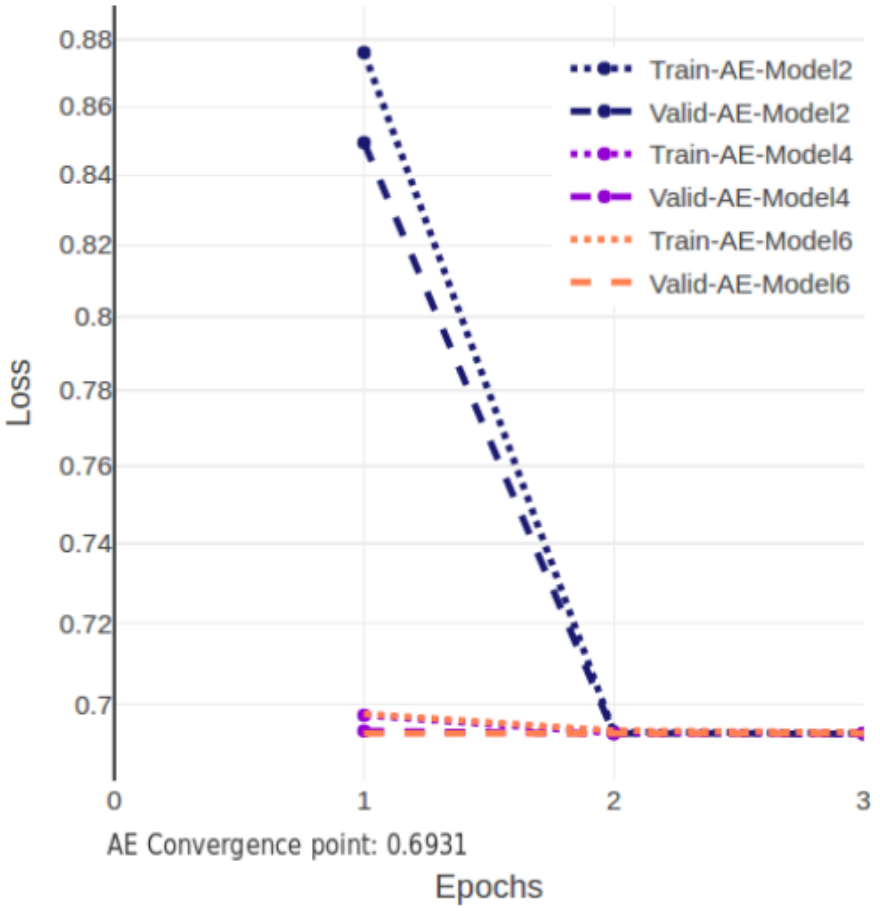

82.21 *Watts* in average to process 1540 gradient updtaes in 14 epochs.

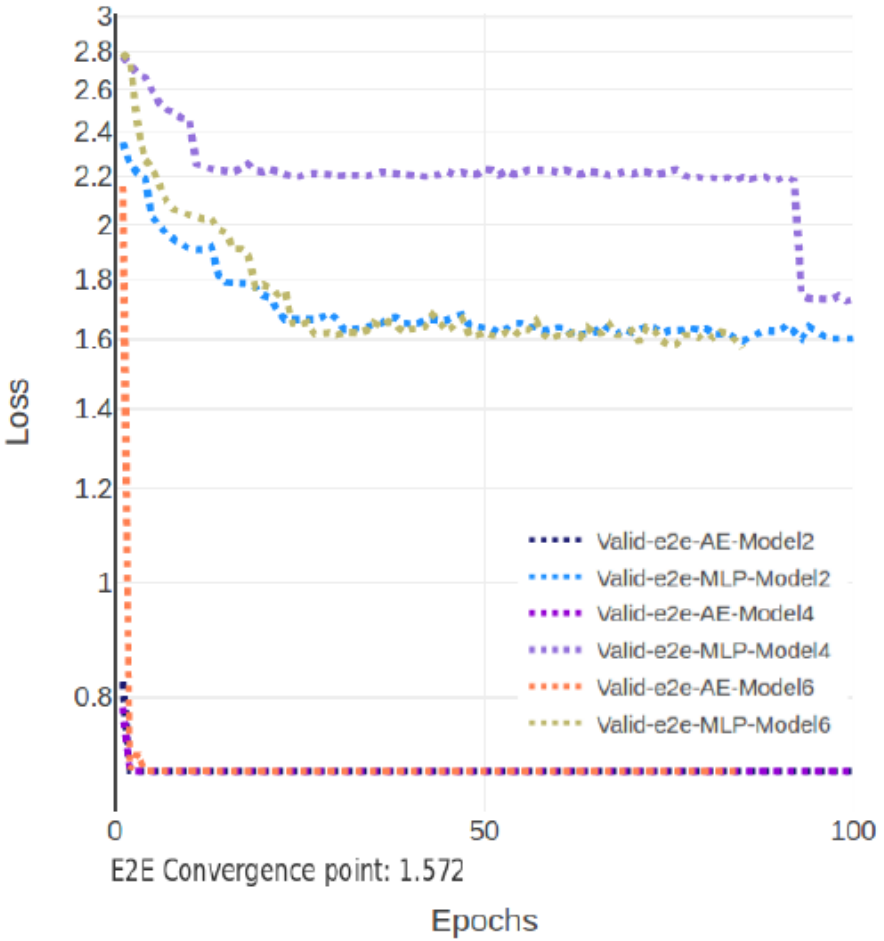
# Experiment Analysis

***2) Model Dimensionality as Factor to Generate Quality Latent Representation***

# Model Dimensionality as Factor to Generate Quality Latent Representation

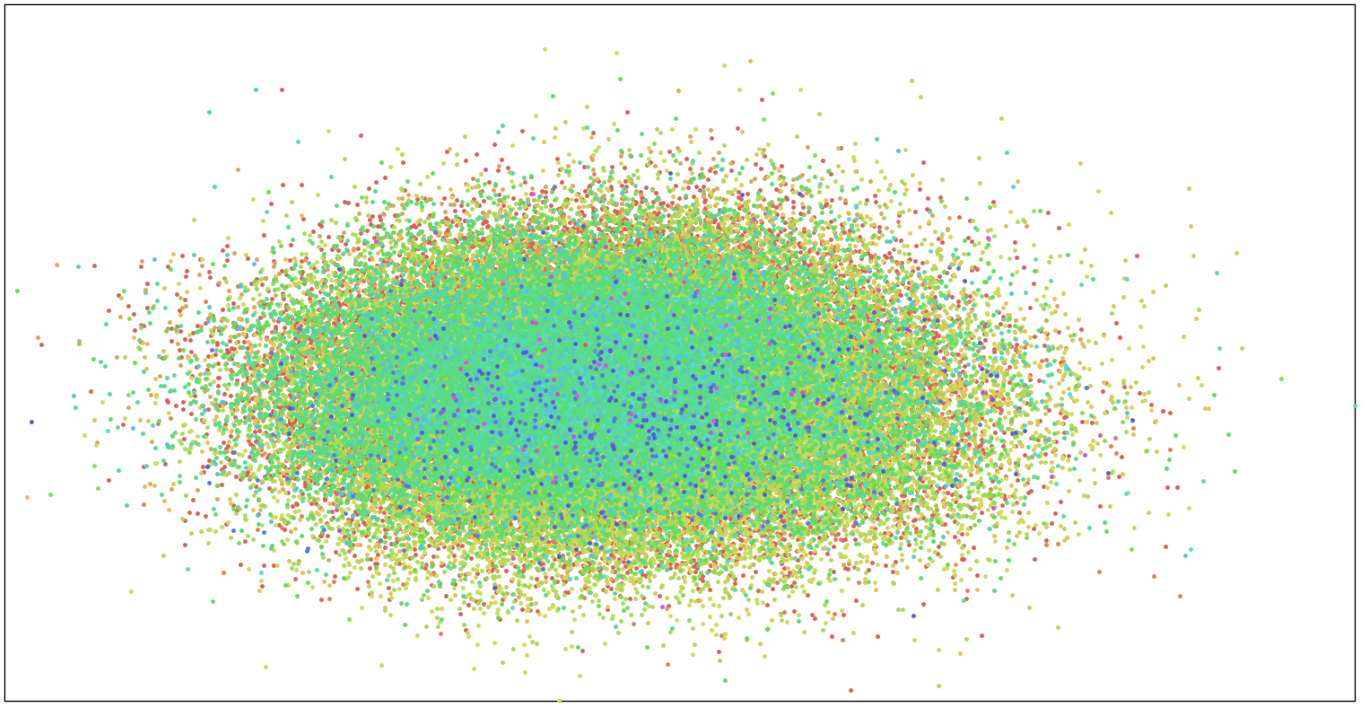
- *Comparison of different model dimensionality using relu as function to generate the latent representation.*



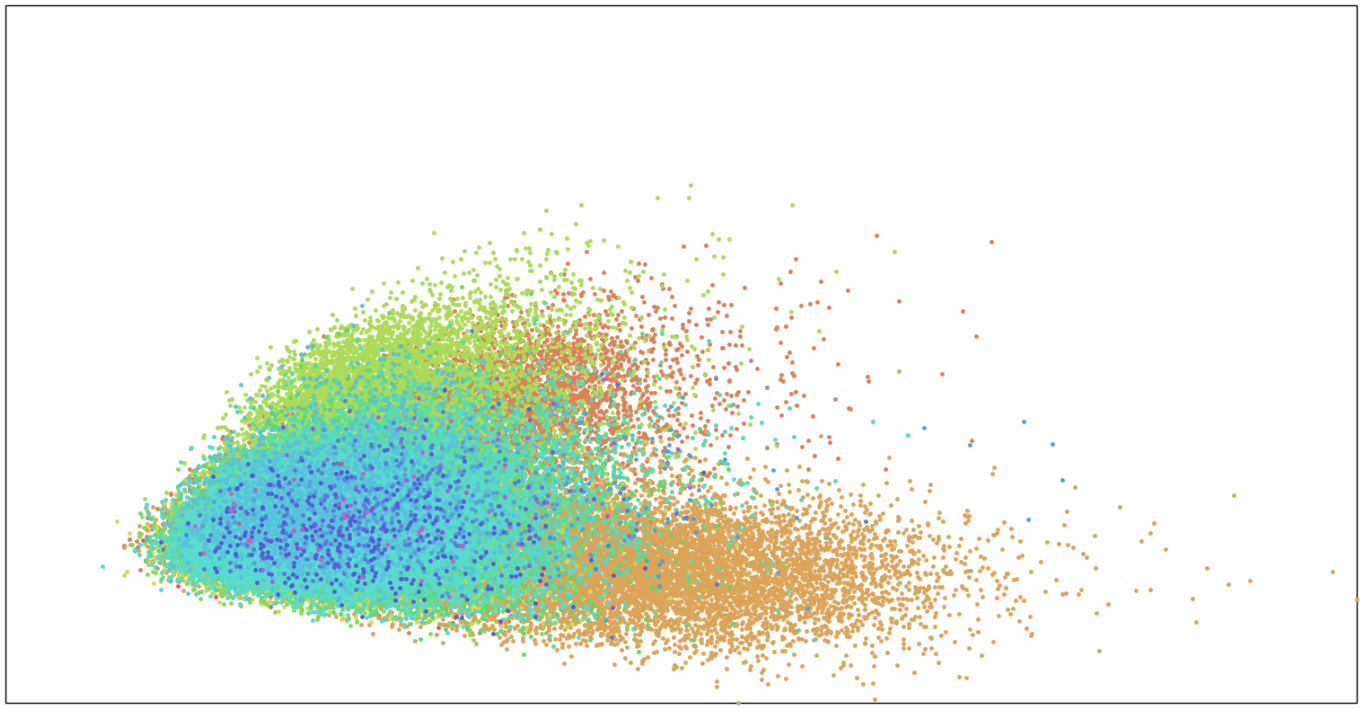



# Model Dimensionality as Factor to Generate Quality Latent Representation

***Autoencoders:***



***End to End:***

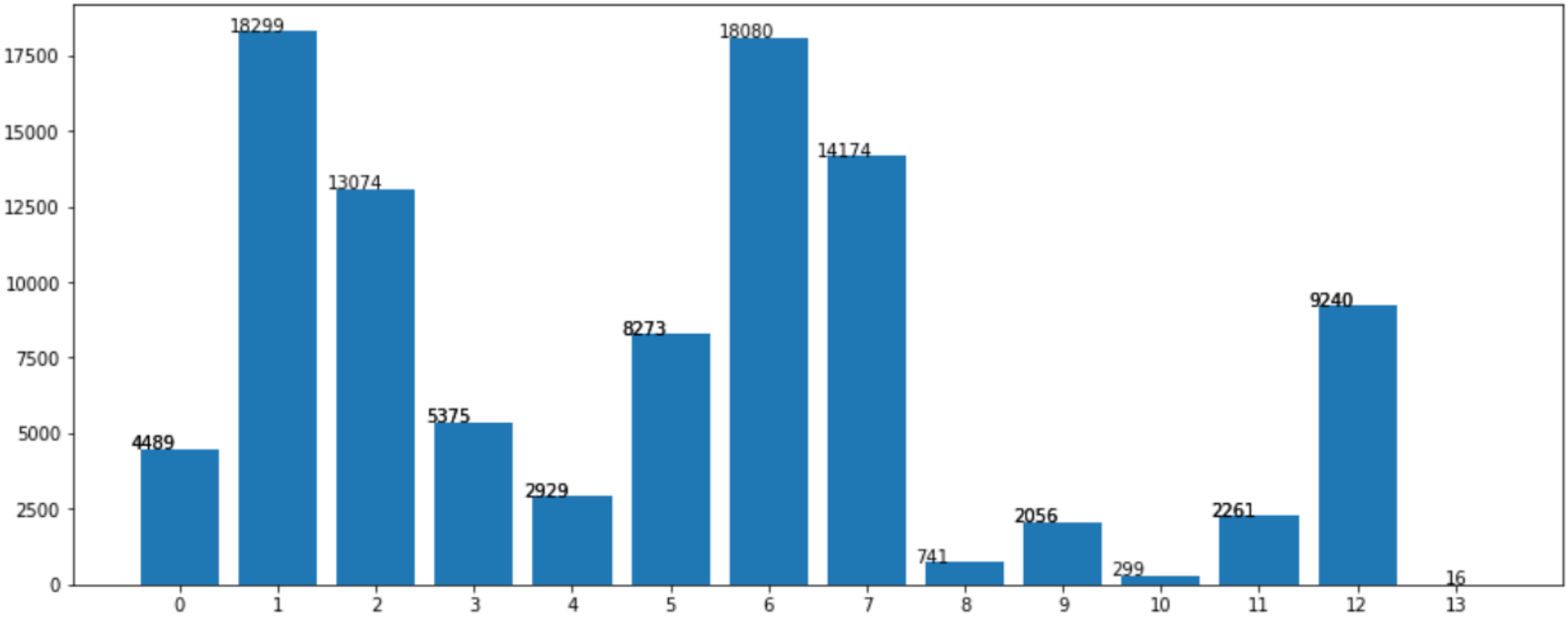


# Supervised Learning

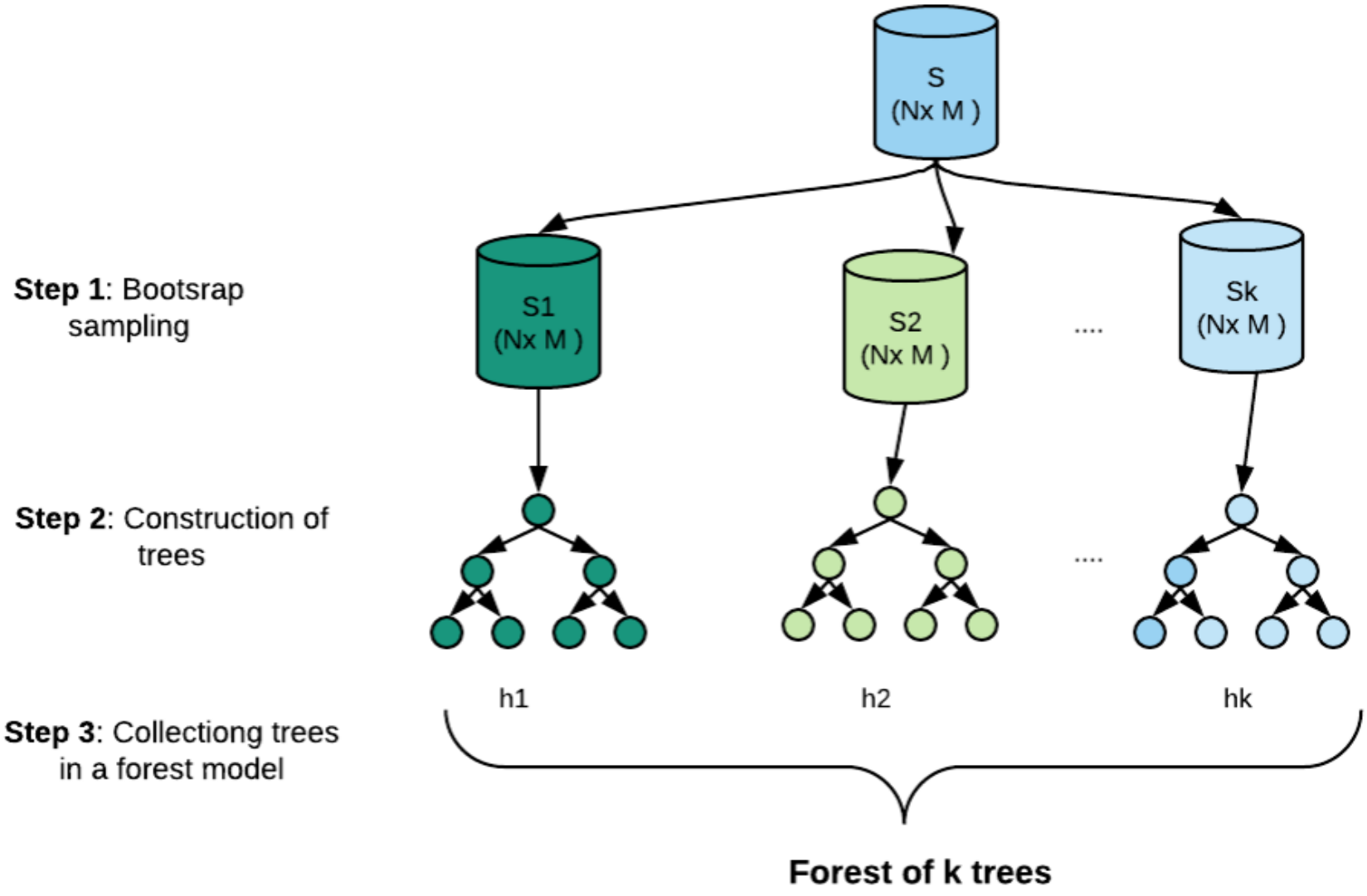
# Medical Target 1: Care Purpose Description Labels at ICU

| Position | Description |
|---|---|
| 0 | Other situations |
| 1 | Proceedings of Medical Cardiovascular / Respiratory Care |
| 3 | Proceedings of Neuro-Muscular Medical Care |
| 4 | Proceedings of Medical Care Mental Health |
| 5 | Proceedings Sensory and Skin Medical Care |
| 6 | Proceedings of Rheumatics / Orthopedic Medical Care |
| 7 | Proceedings of Post-Traumatic Medical Care |
| 8 | Proceedings of Medical Amputations |
| 9 | Palliative care |
| 10 | Placement expectation |
| 11 | Rehabilitation |
| 12 | Proceedings of Nutritional Medical Care |
| 13 | Grouping impossible |

# Medical Target 1: Class Distribution for Training Data

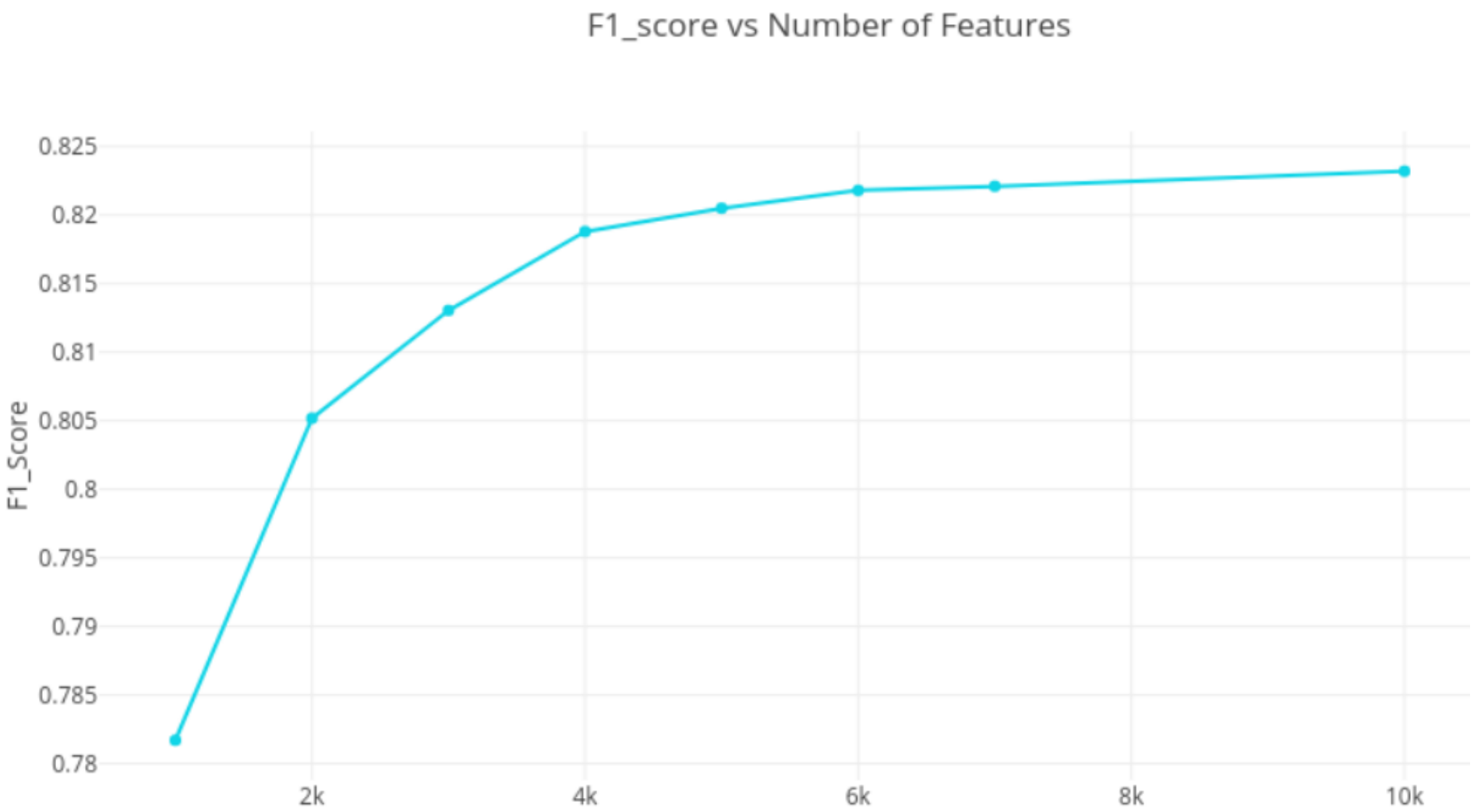


# Machine Learning Algorithm: Random Forest



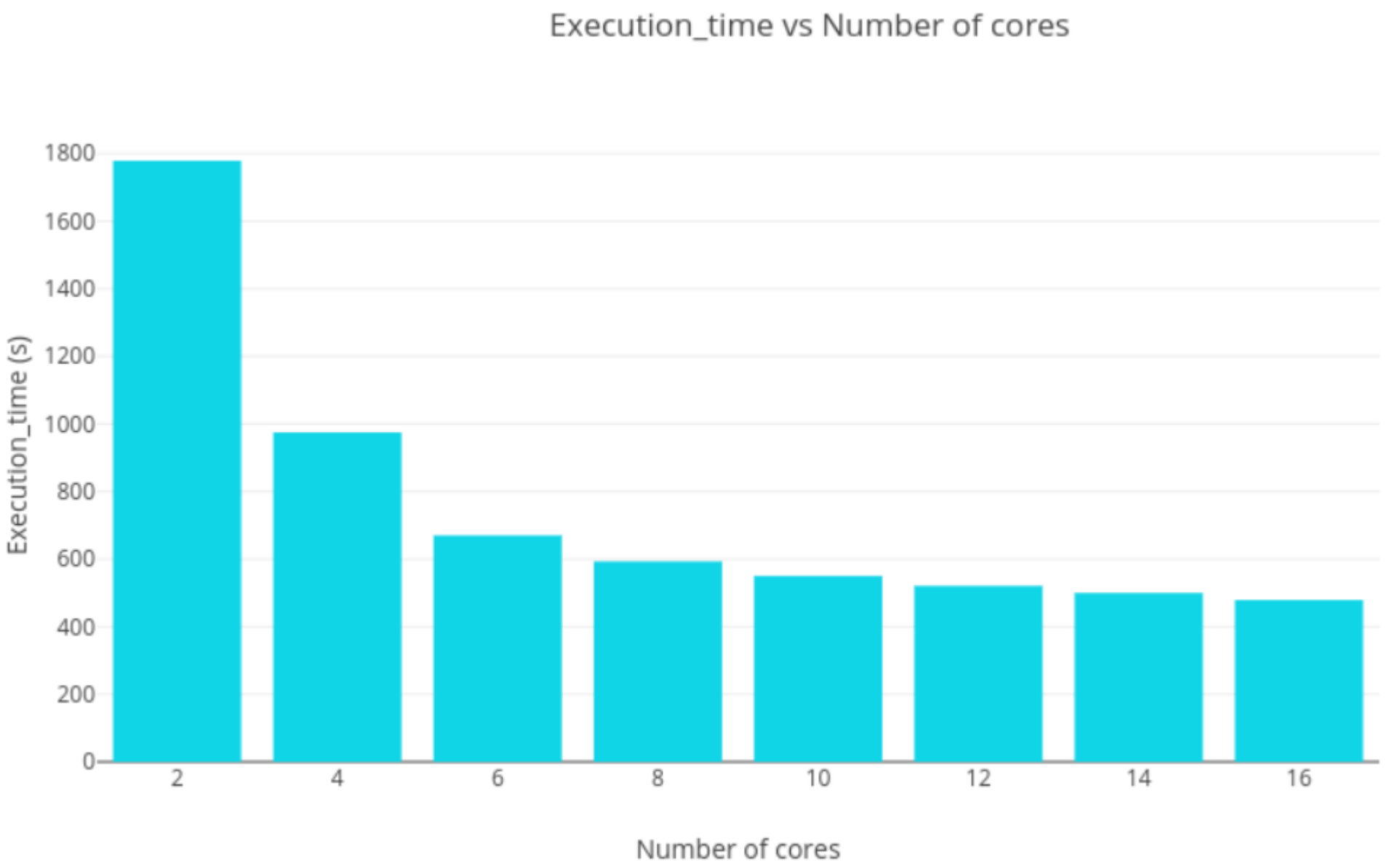

**Ensemble Algorithm Based on Decision Tree Model**

# Random Forest: F1-Score for Different Number of Features Scales

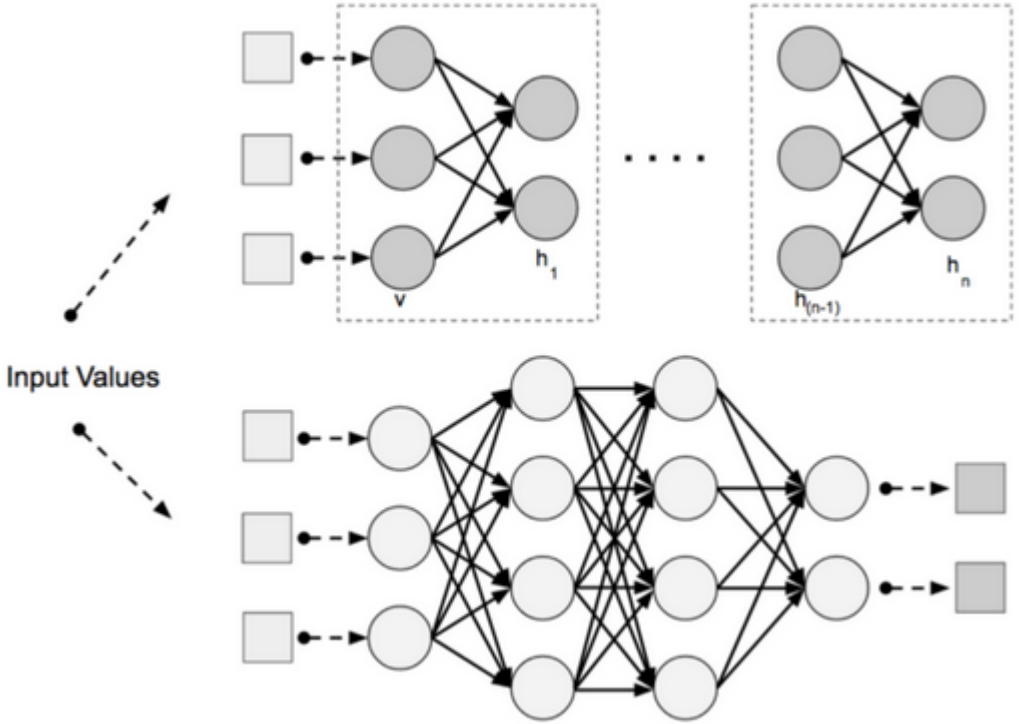


# Random Forest: F1-Score for Different Number of Features Scales

# Random Forest: F1-Score for Different Number of Features Scales



# Deep Learning Network: Feed-forward Multilayer Perceptron



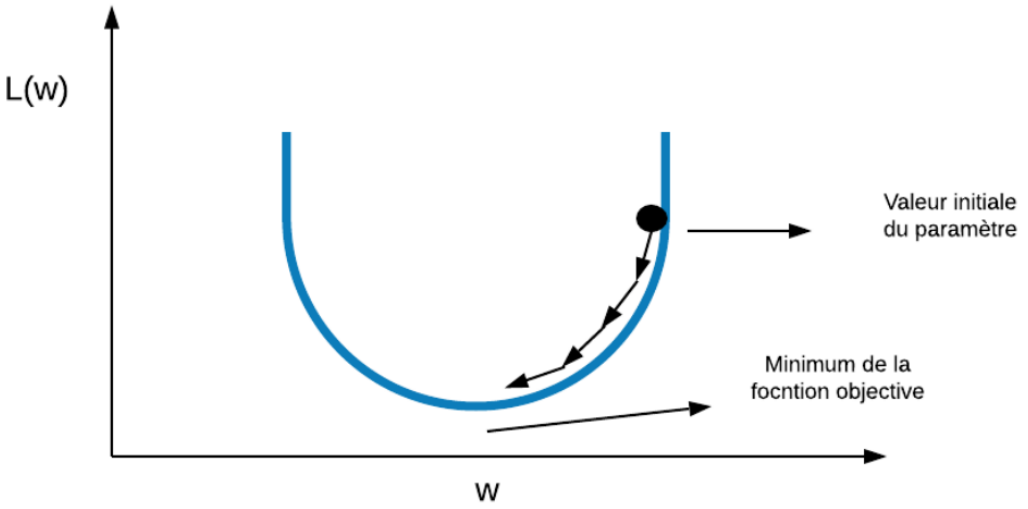

Hidden Layers (*hi)*

Output *(Y'i)*

**Cross Entropy Loss Function**

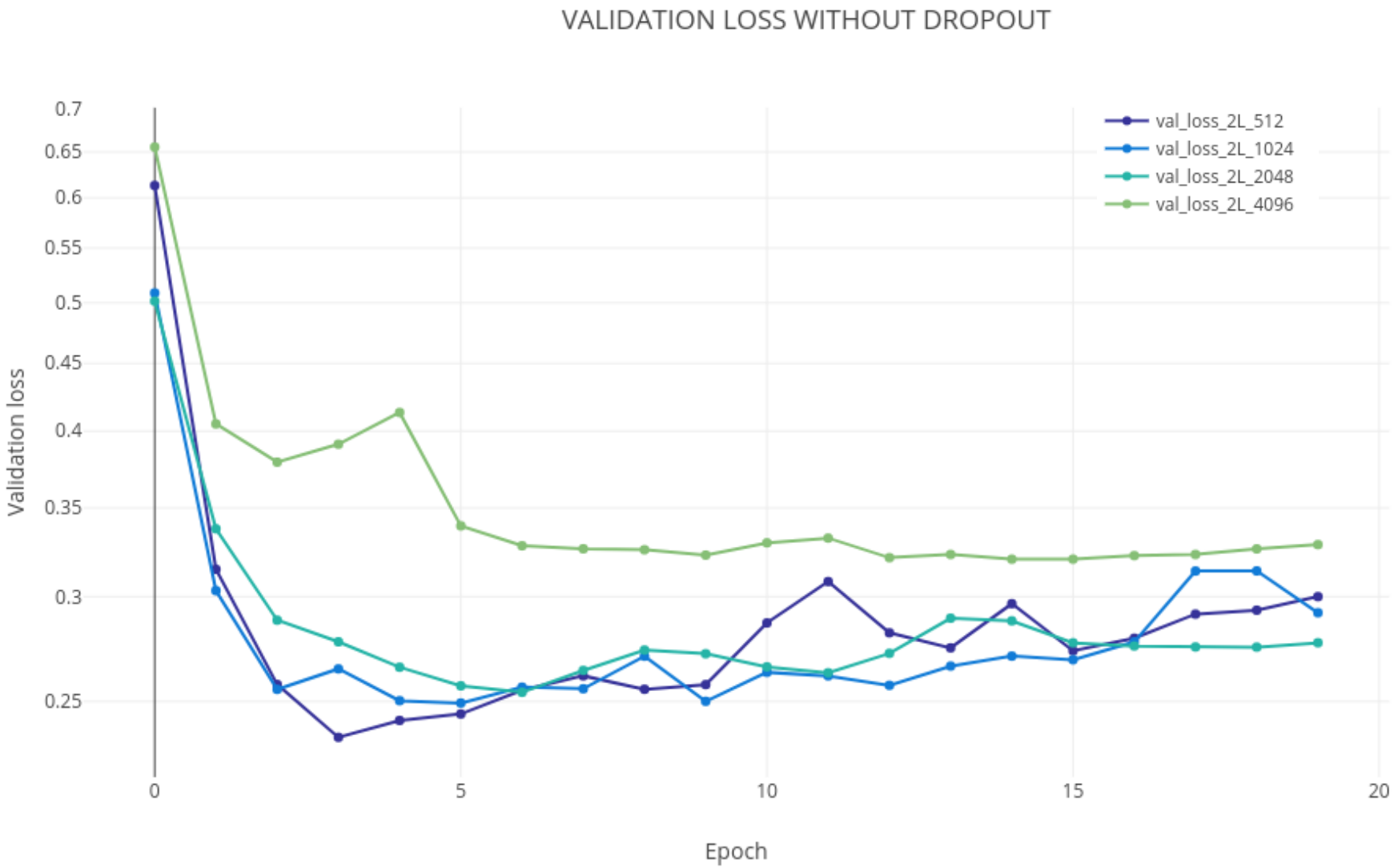
$$L = -\sum_{i=0}^{n} y_i \log(y_i')$$

$$h_i = f(\sum_{j=1}^{n} w_{ij} x_j + b_{ij})$$

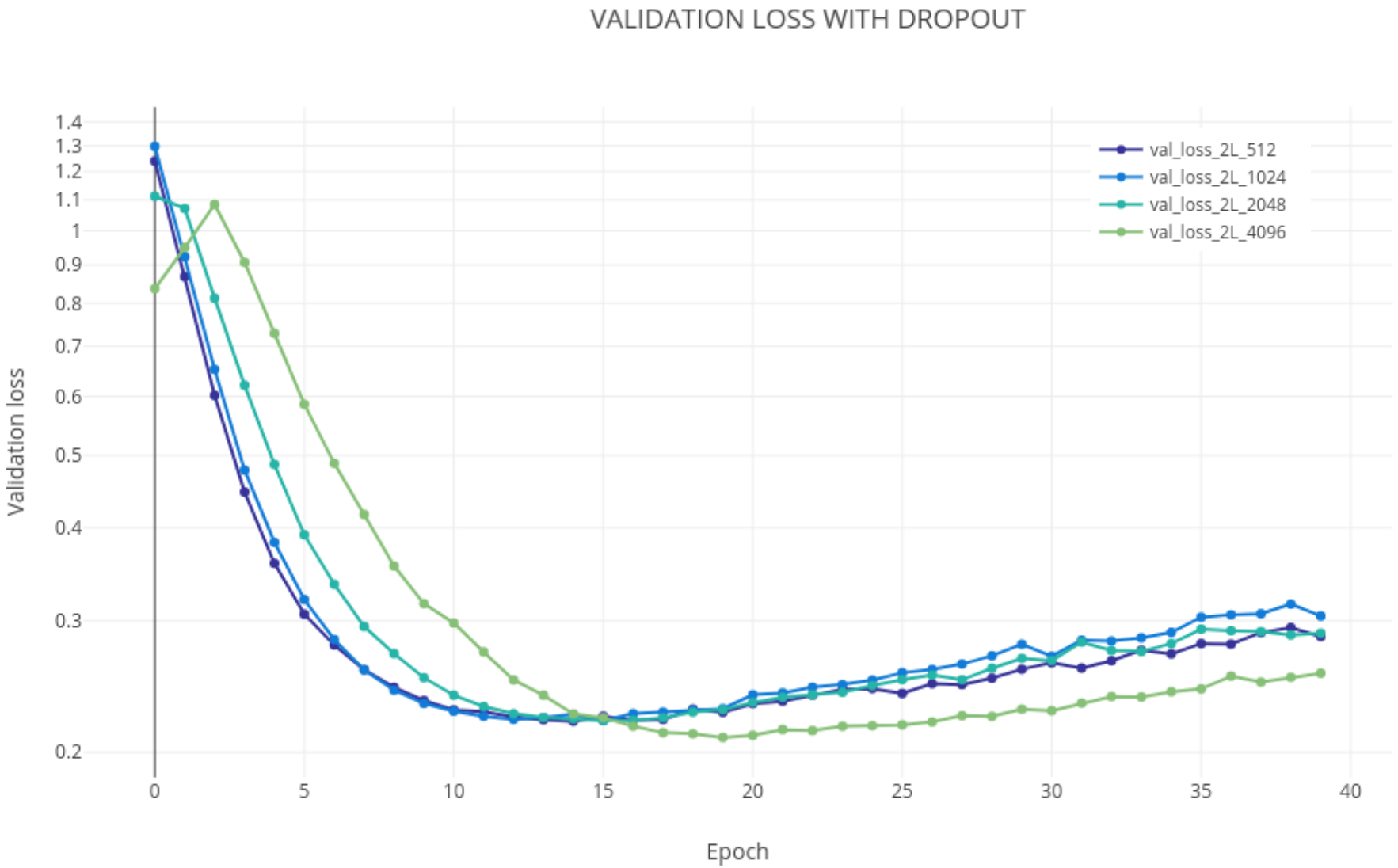



**Gradient Descent Optimization**

# Experiments: Feed-forward Multilayer Perceptron



# Experiments: Feed-forward Multilayer Perceptron



# Stretching the 4096 Neurones over Deep Architectures

## In 2, 4, 8 and 16 hidden layers

| Number of units = 4096 | F1 score | execution time (sec) | energy consumption Kj |
|---|---|---|---|
| 2 layers - 4096 units | 0.92 | 1108 | 238.06 |
| 4 layers - 2048 units | 0.85 | 934 | 161.74 |
| 8 layers - 1024 units | 0.72 | 793 | 124.04 |
| 16 layers - 512 units | 0.75 | 693 | 90.74 |

# Resuls F1 score for different stretching configurations

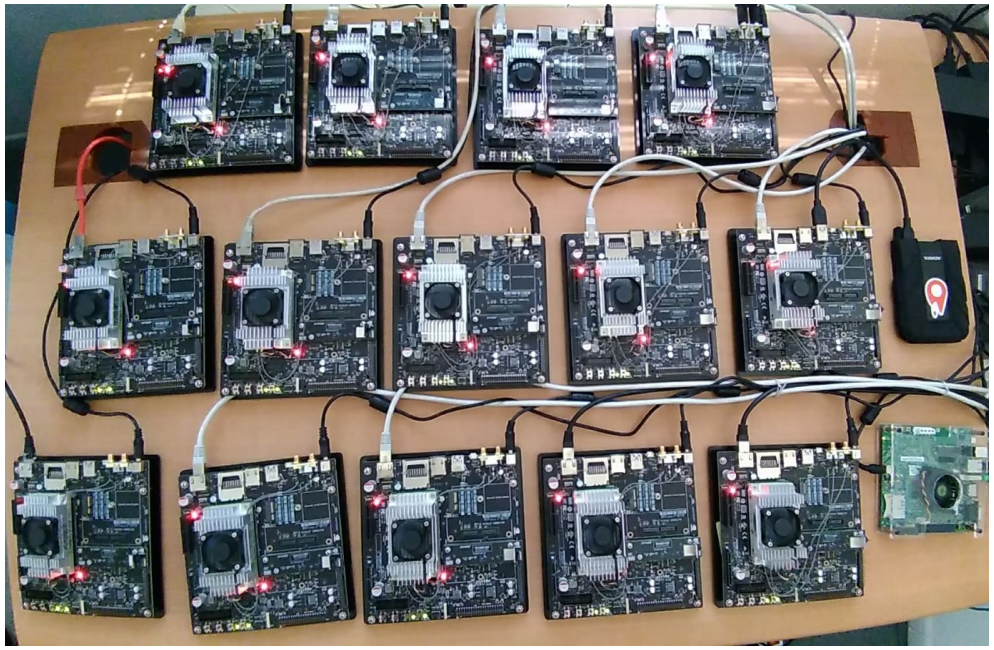

| Architecture | F1 score | execution time (sec) | energy consumption Kj |
|---|---|---|---|
| 256 units - 2 layers - 128 units | 0.92 | 686 | 59.97 |
| 2048 units - 8 layers - 256 units | 0.91 | 654 | 66.49 |
| 8192 - 2 layers - 4096 units | 0.92 | 1108 | 238.06 |

# Performance Results By class to Classify the Medical Target 1

| Classes | True positives | False positives | False negatives | precision | recall | f1 score | occurence de la classe |
|---|---|---|---|---|---|---|---|
| 0 | 424 | 54 | 122 | 0.89 | 0.78 | 0.83 | 546 |
| 1 | 2089 | 136 | 59 | 0.94 | 0.97 | 0.96 | 2148 |
| 2 | 1382 | 98 | 79 | 0.93 | 0.95 | 0.94 | 1461 |
| 3 | 598 | 72 | 34 | 0.89 | 0.95 | 0.92 | 632 |
| 4 | 211 | 73 | 153 | 0.74 | 0.58 | 0.65 | 364 |
| 5 | 861 | 136 | 141 | 0.86 | 0.86 | 0.86 | 1002 |
| 6 | 2086 | 96 | 105 | 0.96 | 0.95 | 0.95 | 2191 |
| 7 | 1574 | 115 | 74 | 0.93 | 0.96 | 0.94 | 1648 |
| 8 | 76 | 9 | 10 | 0.89 | 0.88 | 0.89 | 86 |
| 9 | 101 | 74 | 122 | 0.58 | 0.45 | 0.51 | 223 |
| 10 | 36 | 1 | 3 | 0.97 | 0.92 | 0.95 | 39 |
| 11 | 275 | 31 | 20 | 0.90 | 0.93 | 0.92 | 295 |
| 12 | 1088 | 44 | 16 | 0.96 | 0.99 | 0.97 | 1104 |
| 13 | 0 | 2 | 3 | 0.0 | 0.0 | 0.0 | 3 |

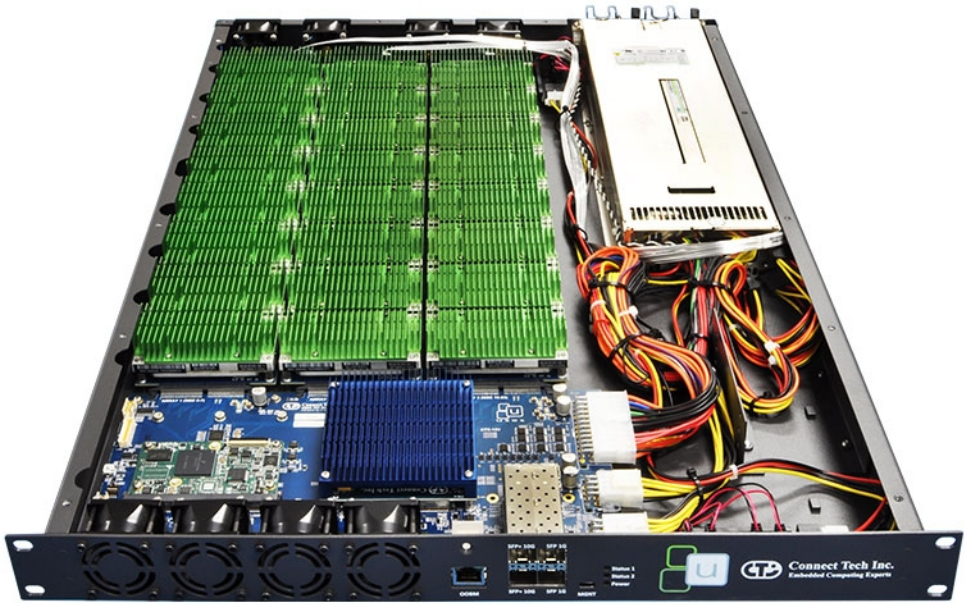
# Distributed Processing for Training DNN on Jetson TX2 Mini-Clusters

# Computational Resources



**Mini-Cluster Jetson TX2**

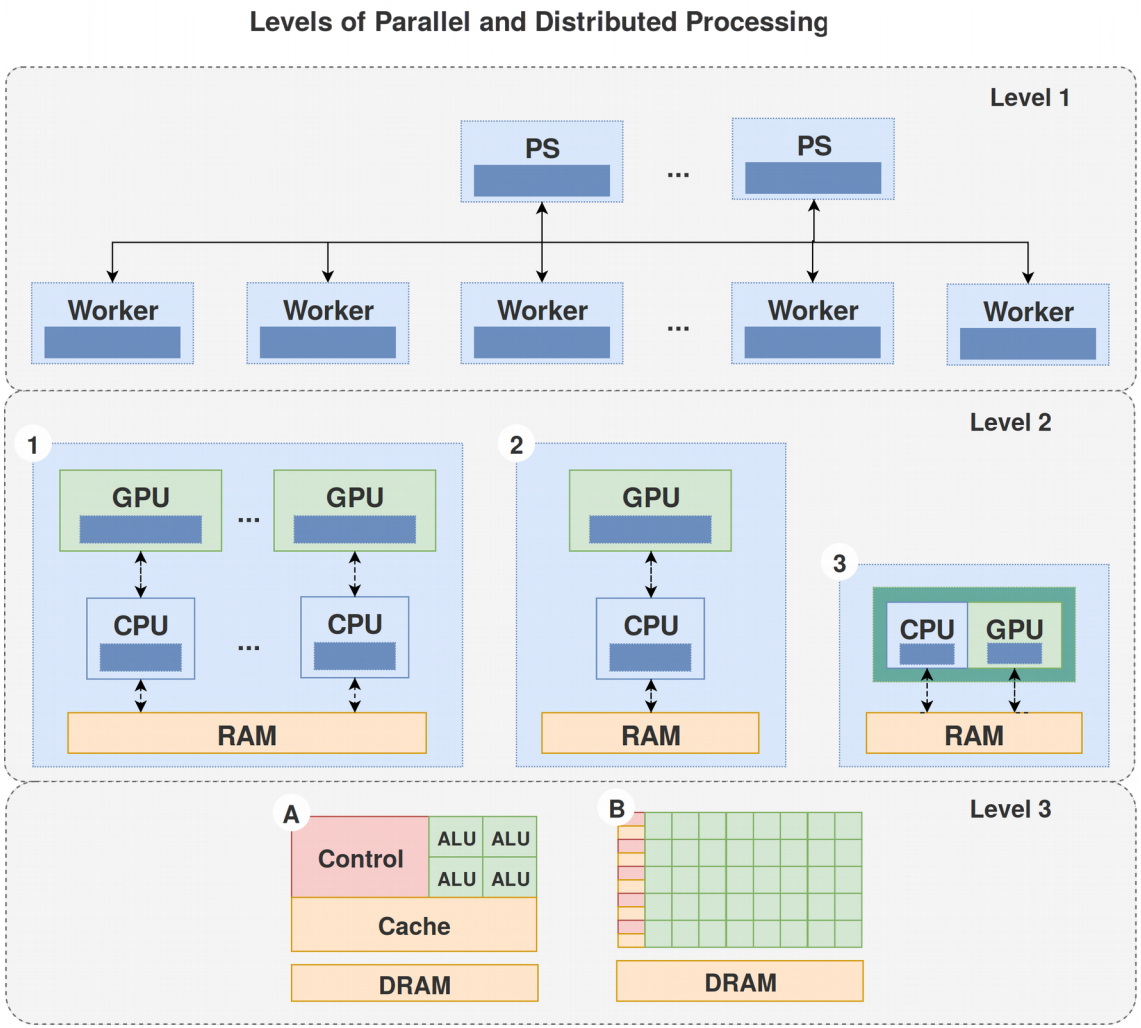
(Distributed Memory)




**Array Node with 24 Jetson TX2**

(Hybrid Memory)

# Develop DiagnoseNET for Training Large-Scale DNN on Distributed Systems




DiagnoseNET

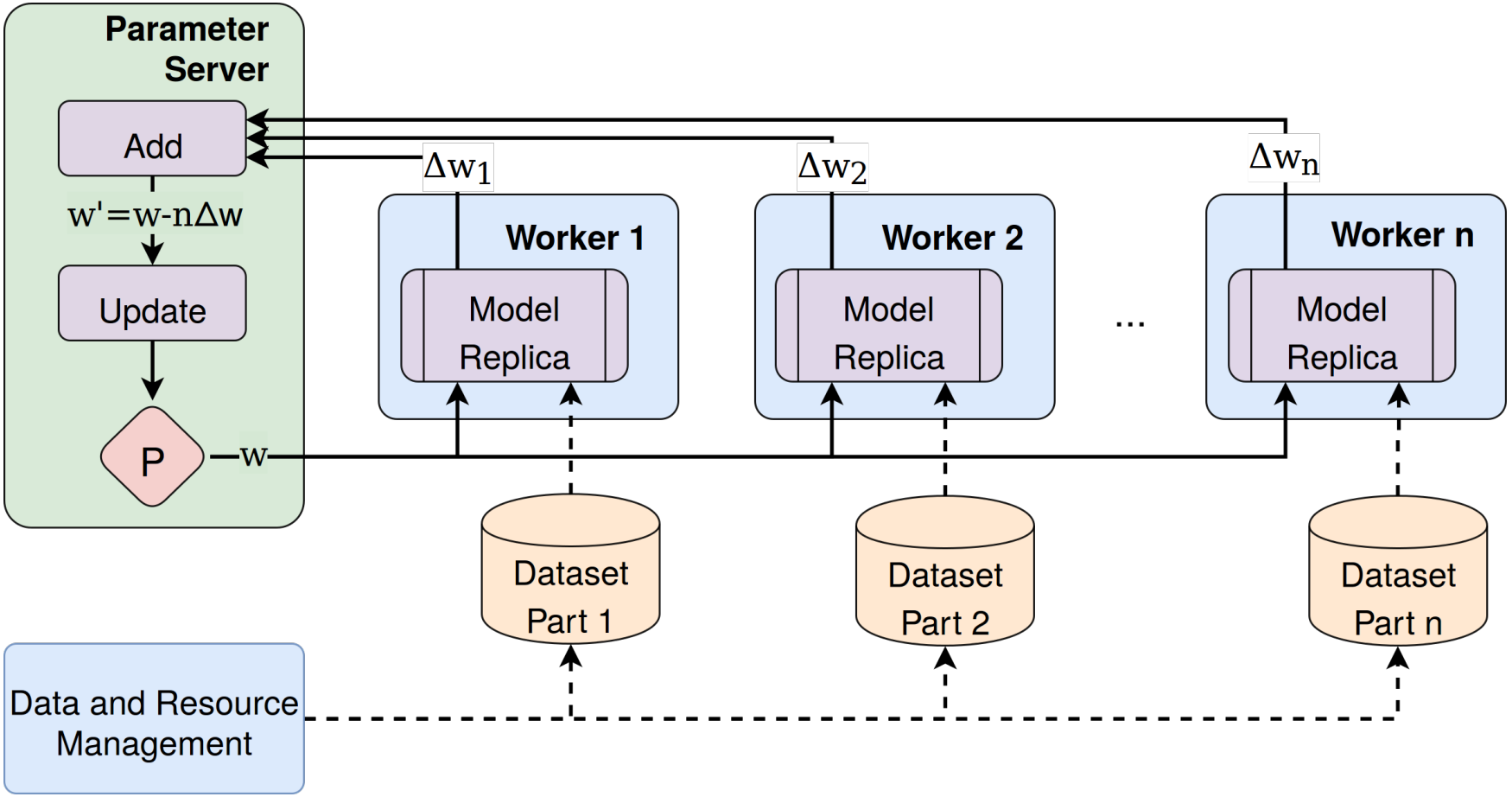
Cross-platform Library

- User Interface
- DiagnoseNET Analysis Interface
- DiagnoseNET DNN Models
- DiagnoseNET Parallel & Distributed Training
- DiagnoseNET Data Management & Resource Manager
- Training Libraries
- Inference Libs
- Python Client
- C++ Client
- C API
- Distributed Master
- Dataflow executor
- Networking Layer
  - GRPC
  - RDMA
  - MPI
- Kernel Implementations cuDNN
  - Relu
  - Conv2
  - MatMul
- enerGyPU Monitor-Tool
- GCC, CUDA, Protobuf, Bazel, ...



6. Adapted from Snap Machine Learning. By IBM Research et al. 2018
7. Adapted from TensorFlow Architecture. By Google Research.

# Task-Based Data Parallelism: Synchronous



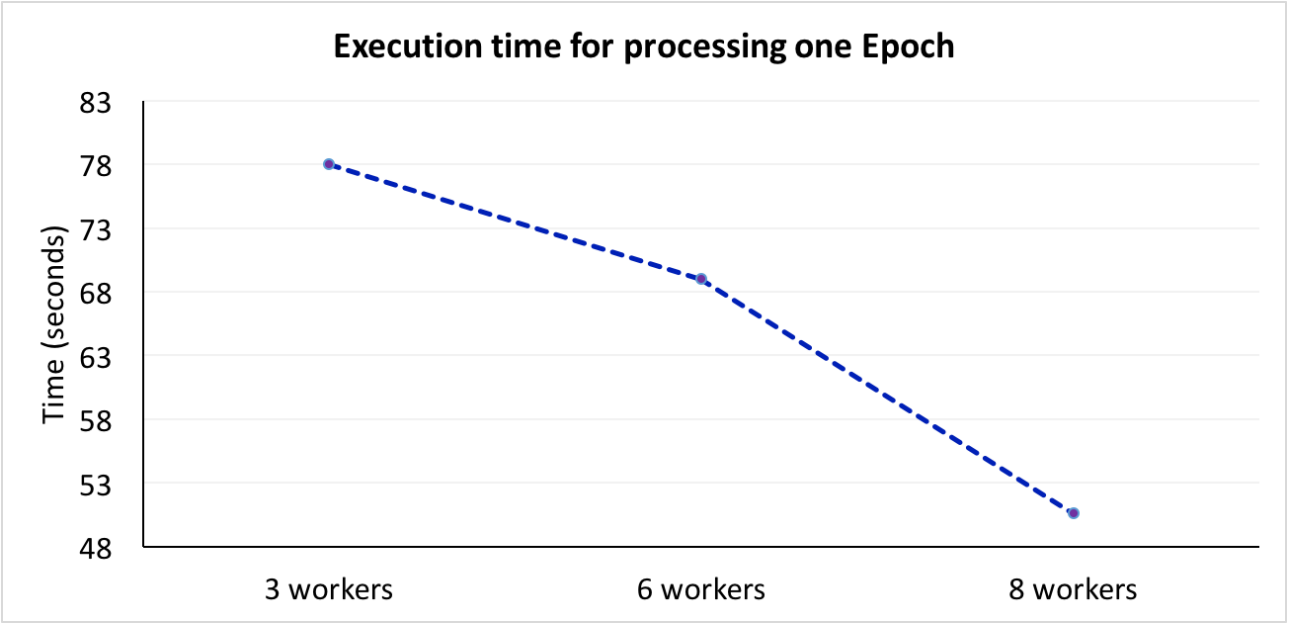

**(Data Size):**

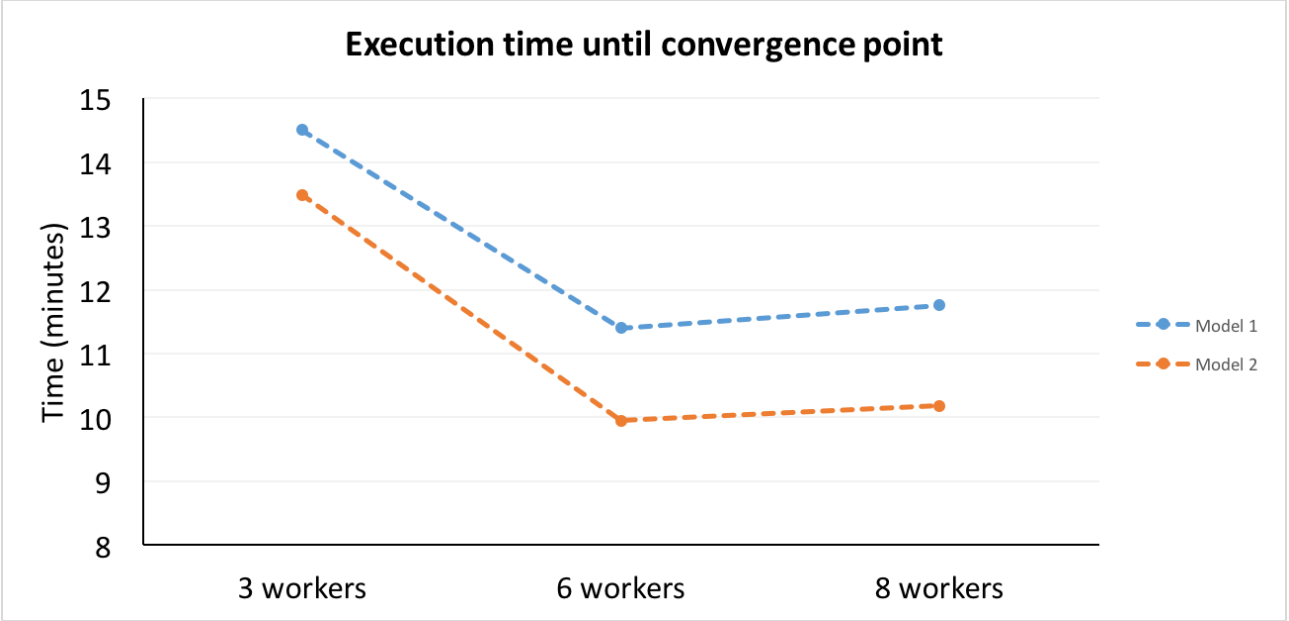
+ Setting the number of workers and micro batch.
+ Fine-tuning DNN hyperparameters.
+ Speeds up the training.
+ I/O Intensive.

8. Adapted from TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. By Google Research. 2015.

# 2) Preliminary Results to Scale the Unsupervised Representation Learning

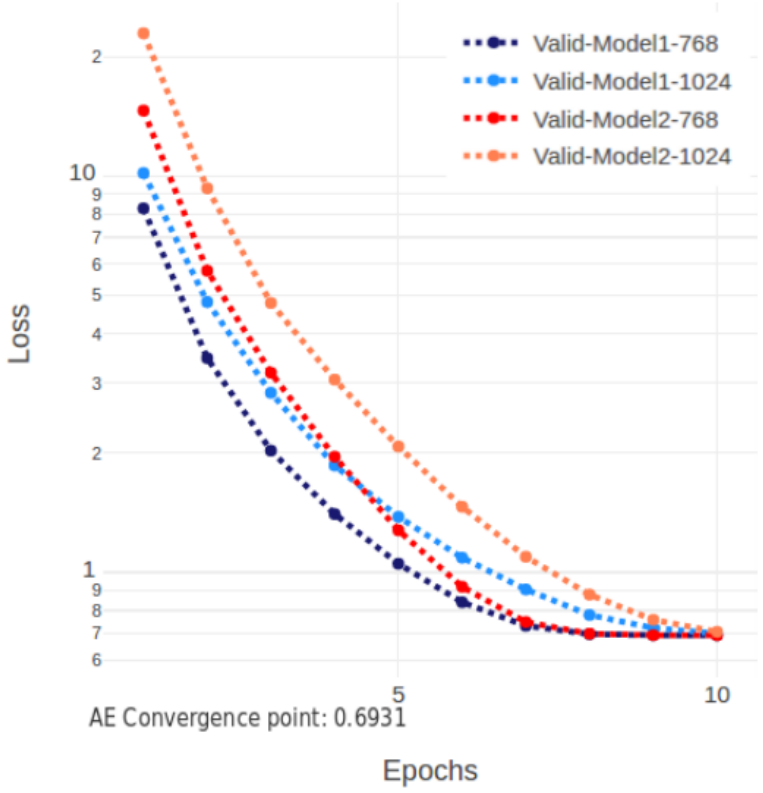
*Preliminary results using: 10.000 records and 11.466 features.*



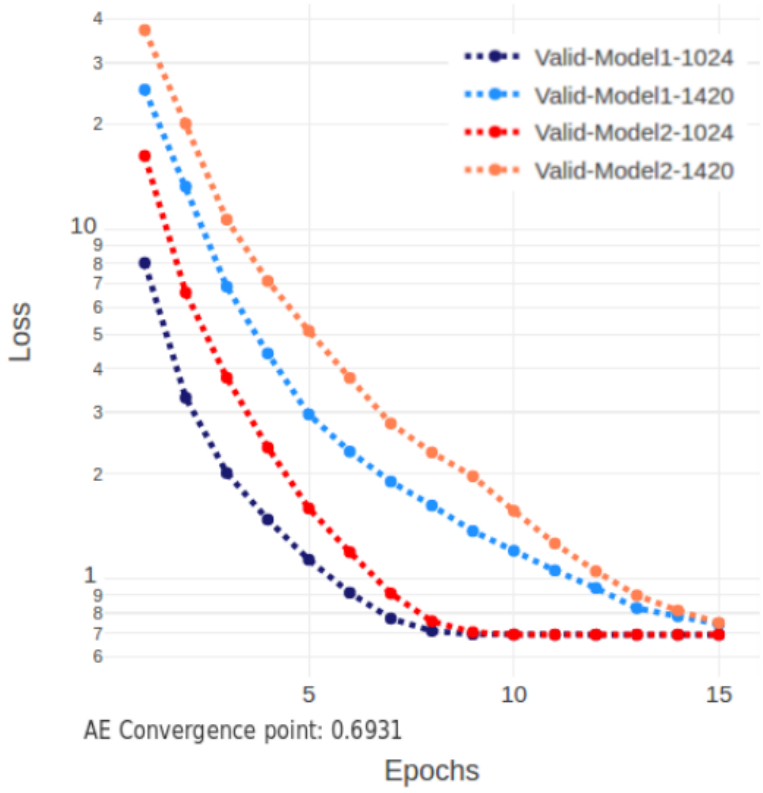



# 3) Number of Workers and Task Granularity as Factor to Early Model Convergence

- *Early convergence comparison between different groups of workers and task granularity for distributed training with 10.000 records and 11.466 features.*



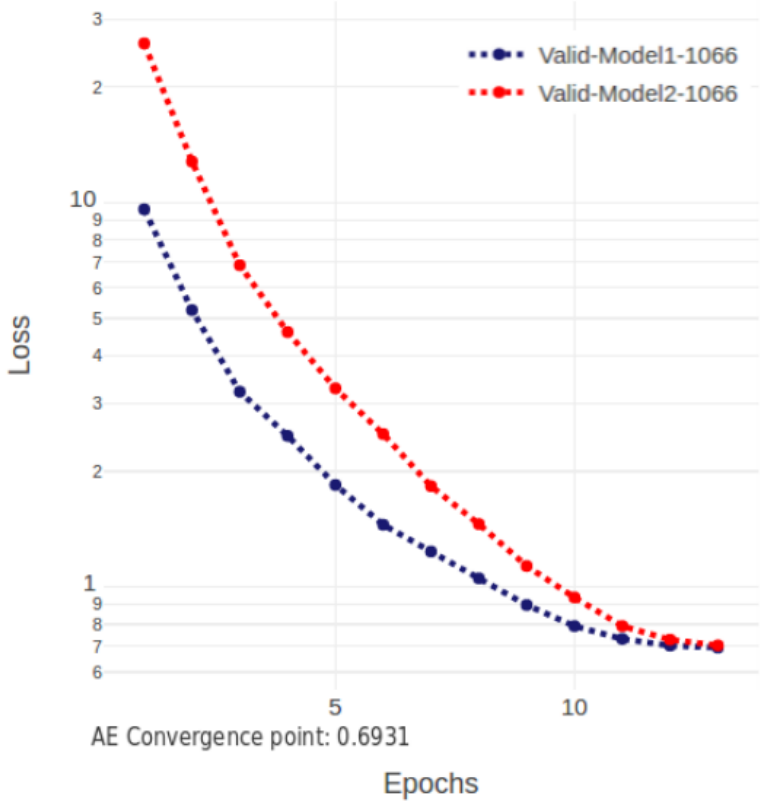

1.30 mins in avergange for processing one epoch on 1 PS 3 workers.




1 min in avergange for processing one epoch on 1 PS 6 workers.

# 3) Number of Workers and Task Granularity as Factor to Early Model Convergence

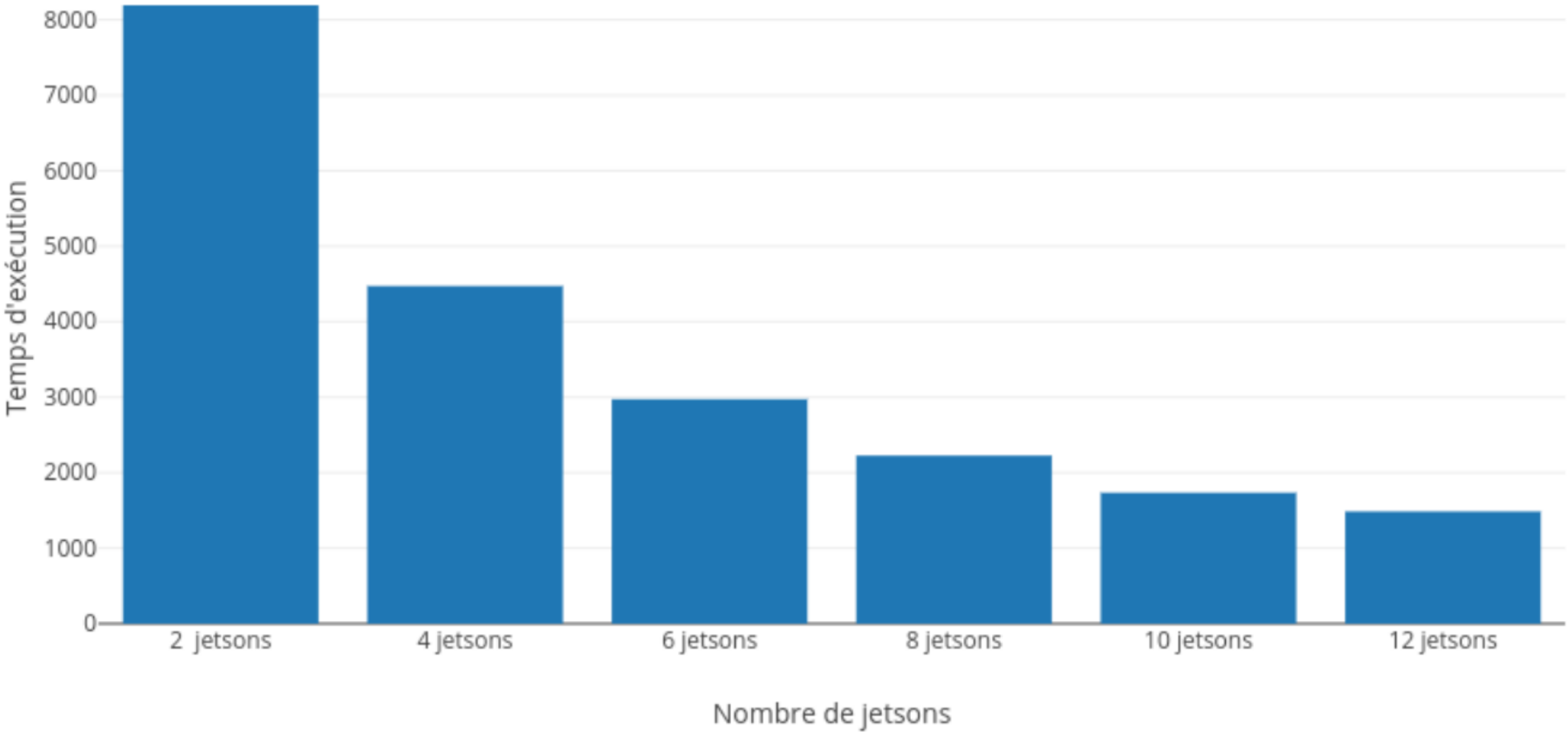
- *Early convergence comparison between different groups of workers and task granularity for distributed training with 10.000 records and 11.466 features.*



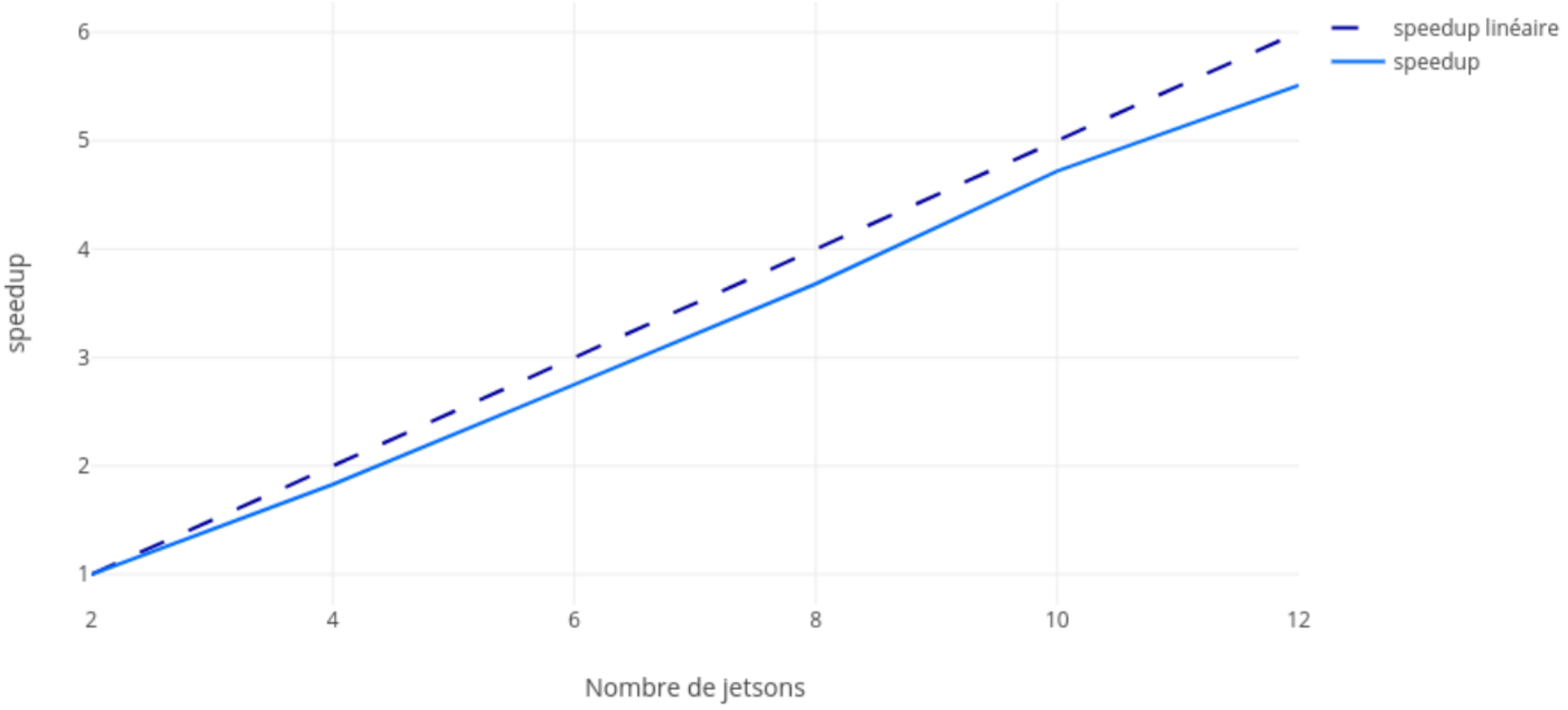

50.6 secs in avergange for processing one epoch on 1 PS 8 workers.

# Preliminary Results to Scale the Feed-forward Multilayer Perceptron

F1-score of 8-Layers Model and 256 neurons per layer
on a cluster of 2, 4, 6, 8, 10 and 12 Jetsons



# Preliminary Results to Scale the Feed-forward Multilayer Perceptron



# Conclusions

- Use the unsupervised embedding stage to create a new lower dimensional patient representation, reduces the number of sparse features to classify at stage 3. In which, the execution time for training is minimized by 41% with regard to BPPR and the precision to classify the first medical target is almost equal.

- Use small batch partitions with larger number of gradient updates allows an early model convergence and minimizes energy consumption.

- The future work is focused on evaluating the different DNN approach using the different platform such as, cluster Jetson TX2 (distributed memory), a multiGPU Node with 8 GPUs (Share memory) and the array Node with 24 Jetson TX2 (Hybrid memory).

# DiagnoseNET: Green Intelligence System Criteria

**1) Select optimal computational resources and make good mapping of task granularity for training one model in less time and less power consumption give a mini-batch size factor.**

**2) Minimize the number of different trained models to converge the optimal generalization-accuracy model.**

**3) Management the queue of models to training and determine optimal combination of computational resources to use in each model training.**

# Gracias Por su Atención

# Scalability Analysis of Mini-Cluster Jetson TX2 for Training DNN Applied to Healthcare

John A. GARCÍA. H.

Frédéric PRECIOSO, Pascal STACCINI, Michel RIVEILL

**Université Côte d'Azur, CNRS, I3S**

**Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis - I3S**