

DataOps et apprentissage machine au service du bâtiment intelligent

Thèse de doctorat en informatique défendue par Louis
CLOSSON

6 Mai 2024



Plan

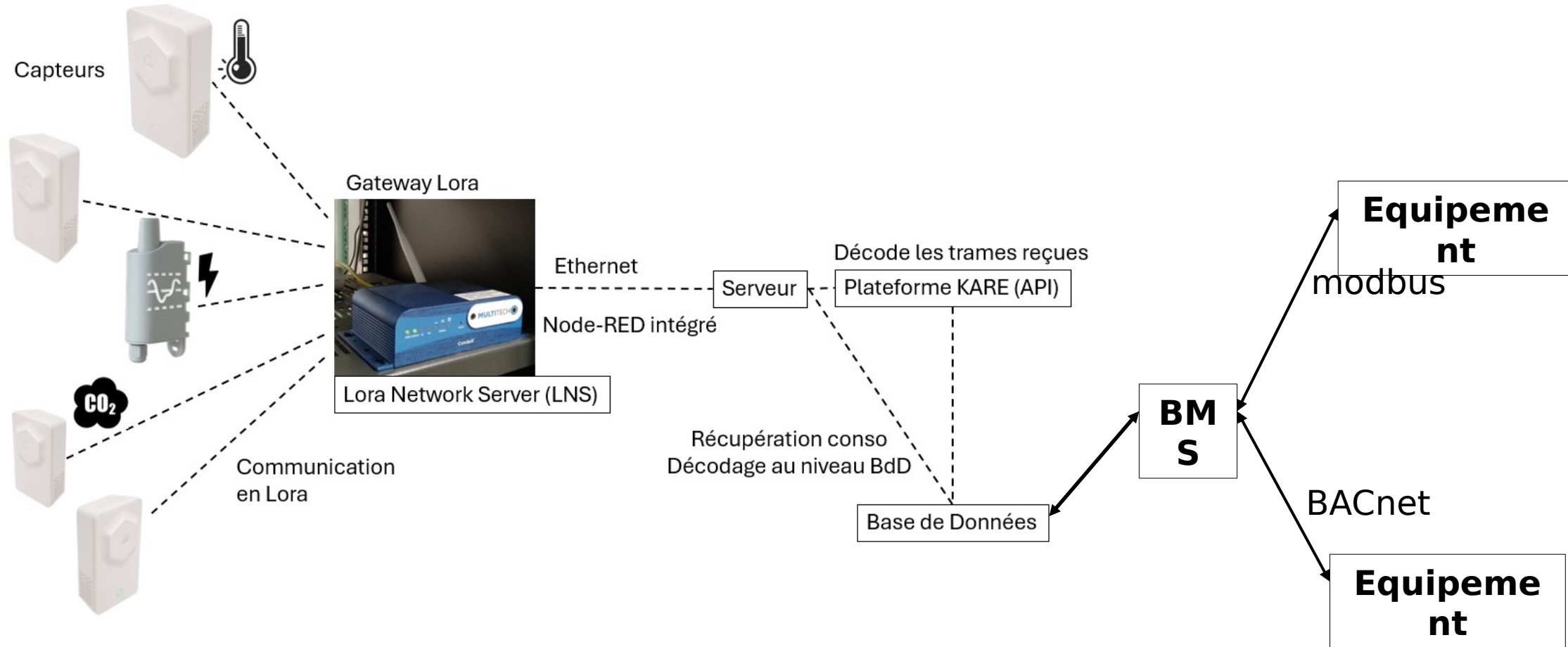
- I. Contextes du bâtiment intelligent
- II. Détection d'anomalies et regroupement
- III. Prévion de données
- IV. Apprentissage par transfert
- V. Conclusion
- VI. Perspectives

Impact environnemental du bâtiment

- 34% de la consommation / 37% des émissions de CO2 (mondial, 2021)
 - en constante augmentation
- 50% du parc tertiaire antérieur à 2000 pour les pays de la Loire (1)
 - Beaucoup de rénovations
 - Parc des logements plus ancien que le parc tertiaire
- Une multitude d'acteurs
- Une législation exigeante : décret BACS, décret tertiaire ...

[\(1\) Rapport restitution DREAL Pays-de-la-Loire](#)

Le bâtiment intelligent - surveillé et contrôlé



L'apprentissage machine supervisé

- Apprentissage Machine = statistiques

Technique	Type
Analyse en composante principale	IA non supervisée
Machine à vecteur de support	IA non supervisée (paramètres)
Descente de gradient	IA non supervisée
Arbre de décision	IA supervisée
Le Perceptron	IA supervisée

$y = f()$ avec f fonction non linéaire

L'Apprentissage profond

- Apprentissage Profond = Apprentissage itératif
- Apprentissage supervisé =
 - En entraînement : entrées et sorties données (et y)
 - En test : seuls des sont disponibles
- Importance des données (quantité et qualité)
 - ChatGPT $\Rightarrow y$ est un texte crédible / usuel d'internet
 - Succès DeepSeek = notation partielle du **raisonnement**

Contributions et démarches

Construction d'un jeu de données augmenté

- Etude de méthodes de regroupement et détection d'anomalies
Contribution 1

Instrumentation d'un bâtiment tertiaire

- Etude d'algorithmes de prévision de données
Contribution 2
- Etude sur potentiel de l'apprentissage par transfert
Contribution 3

II. Détection d'anomalies et regroupement

Dans le bâtiment intelligent,

Des comportements typiques à identifier ?

Données - Bâtiment GreEn-Er (ENSE3 et G2Elab)

- De nombreuses types de CVC et types de salles
- Instrumenté à la construction
 - 350 capteurs de température et humidité
 - 74 capteurs de CO2



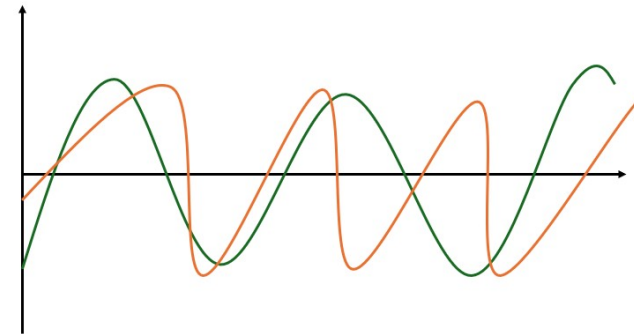
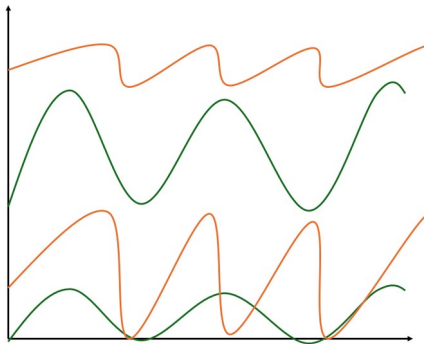
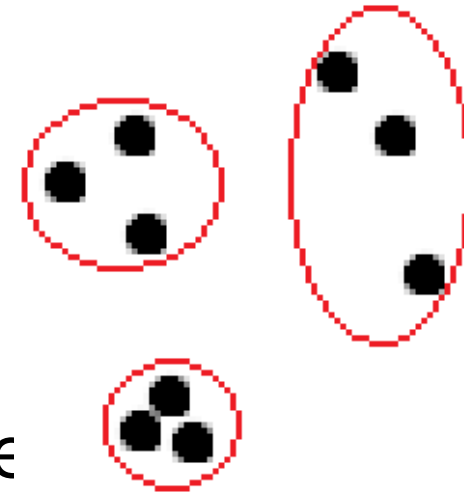
Source: ENSE3



Source: [GreEn](#)
[-ER living lab: A green building with energy aware occupants](#)

Regroupement - Principe

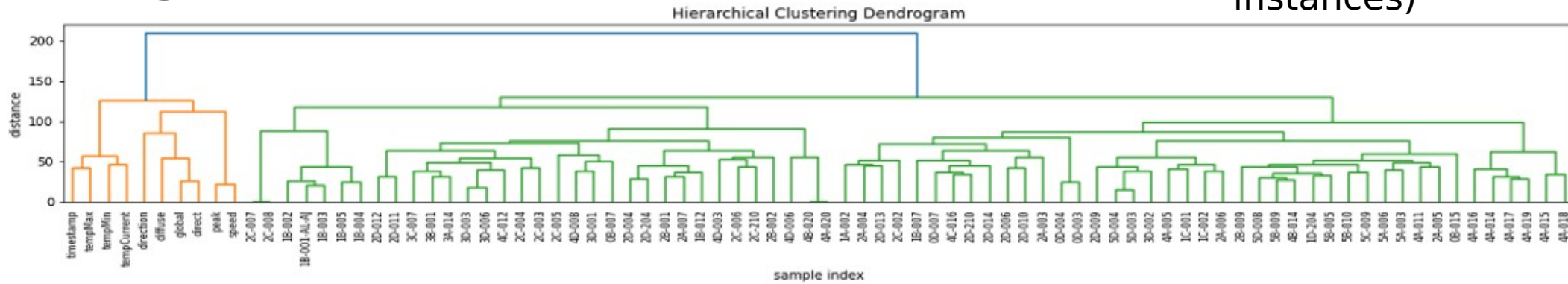
- Faire des groupes de capteurs pertinents
 - distance euclidienne en 2D
 - autre définition pour séries temporelles



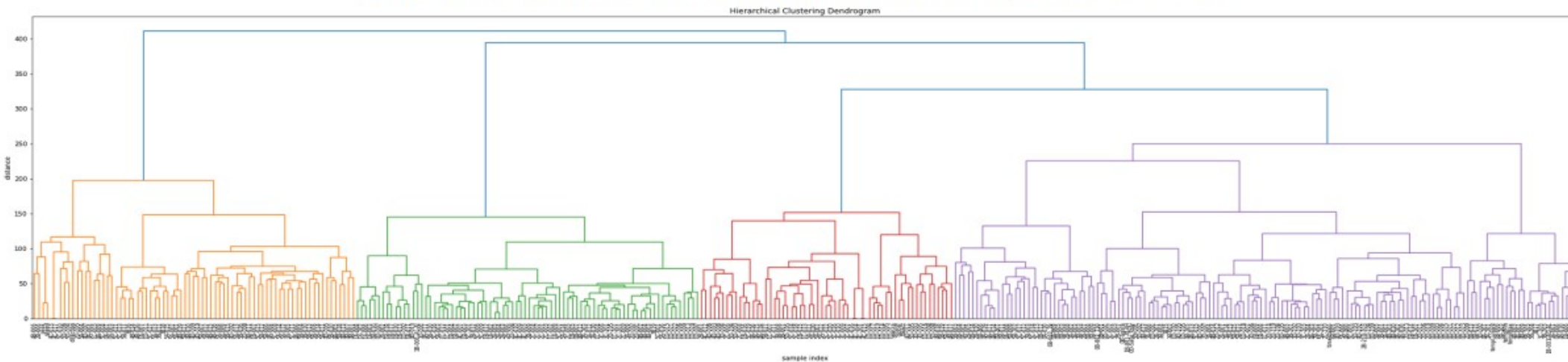
- Distance entre courbes :
 - Données normalisées = distance euclidienne
 - Combien de groupes = pas de choix fait

Regroupement - résultats

Historique d'un an (38000 instances)



(a) Dendrogramme du regroupement des capteurs de CO2



(b) Dendrogramme du regroupement des capteurs de températures

Regroupement - résultats

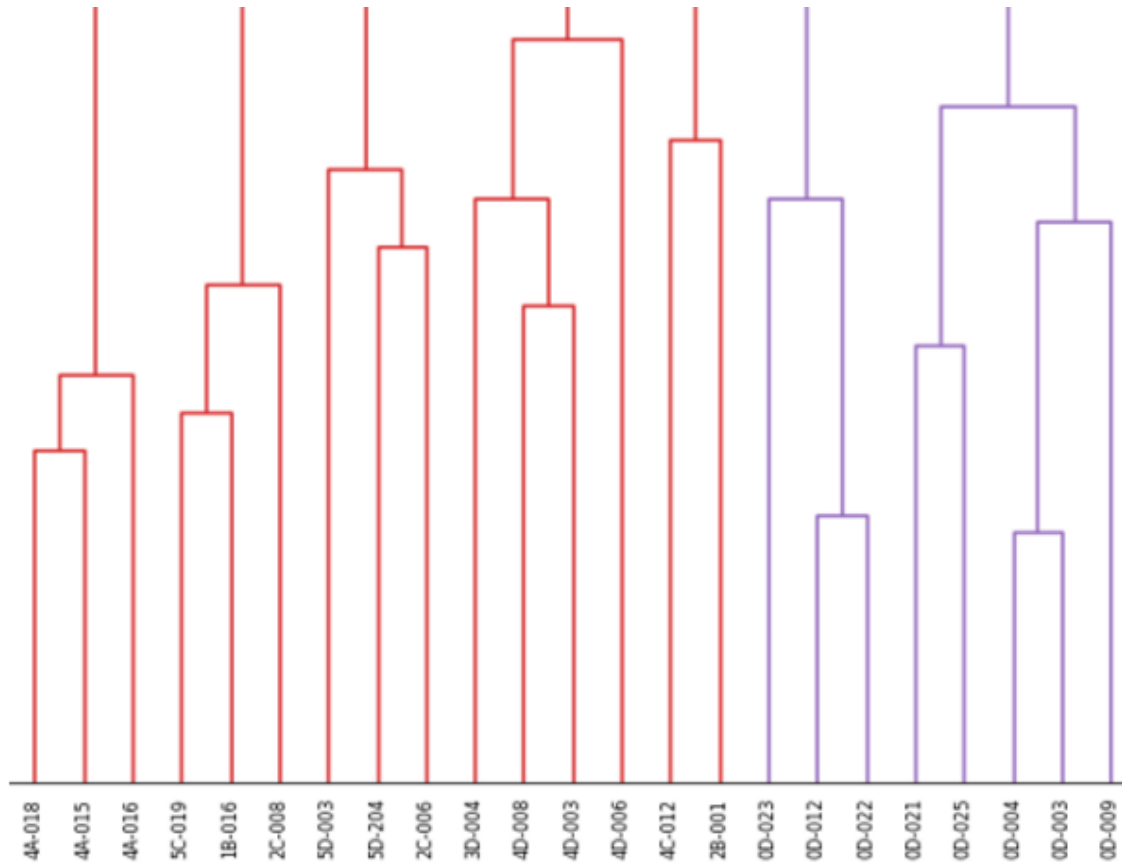


FIGURE 3.3 – *Zoom sur frontière*

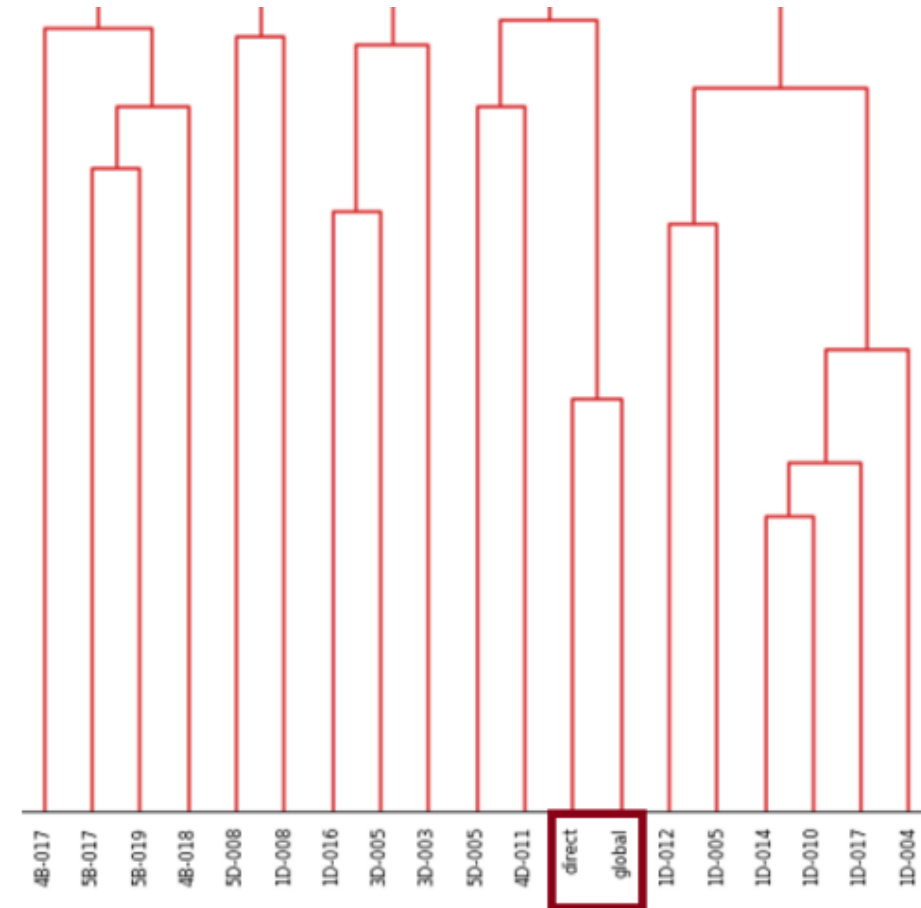


FIGURE 3.4 – *Zoom sur l'ensoleillement*

Détection d'anomalies - principe

- Pas d'information réelle $\Rightarrow$ anomalies statistiques
- 38 000 instances de temps (ici : 16 grandeurs relatives au fluides)

Algorithme	Types d'anomalies
LOF	anomalies locales basée sur plus proches voisins
HDBSCAN	anomalies de densités de points
iForest	isolation par frontière (paramétrique)

Détection d'anomalies - évaluation

- Algorithmes multiples pertinents ?
Oui – Indice de Jaccard de 0,3% et 0,8% des points anormaux
- Détection pertinente pour effectuer un regroupement temporel ?
Oui, pour la sélection de 7 clusters totaux
 - Réduction des diamètres moyens des groupes, d'un facteur:
 - 4 pour distance euclidienne,
 - 1,5 pour mahalanobis
 - Peu de changement sur les centres des groupes

= anomalies ponctuelles

III. Pr vision de donn es

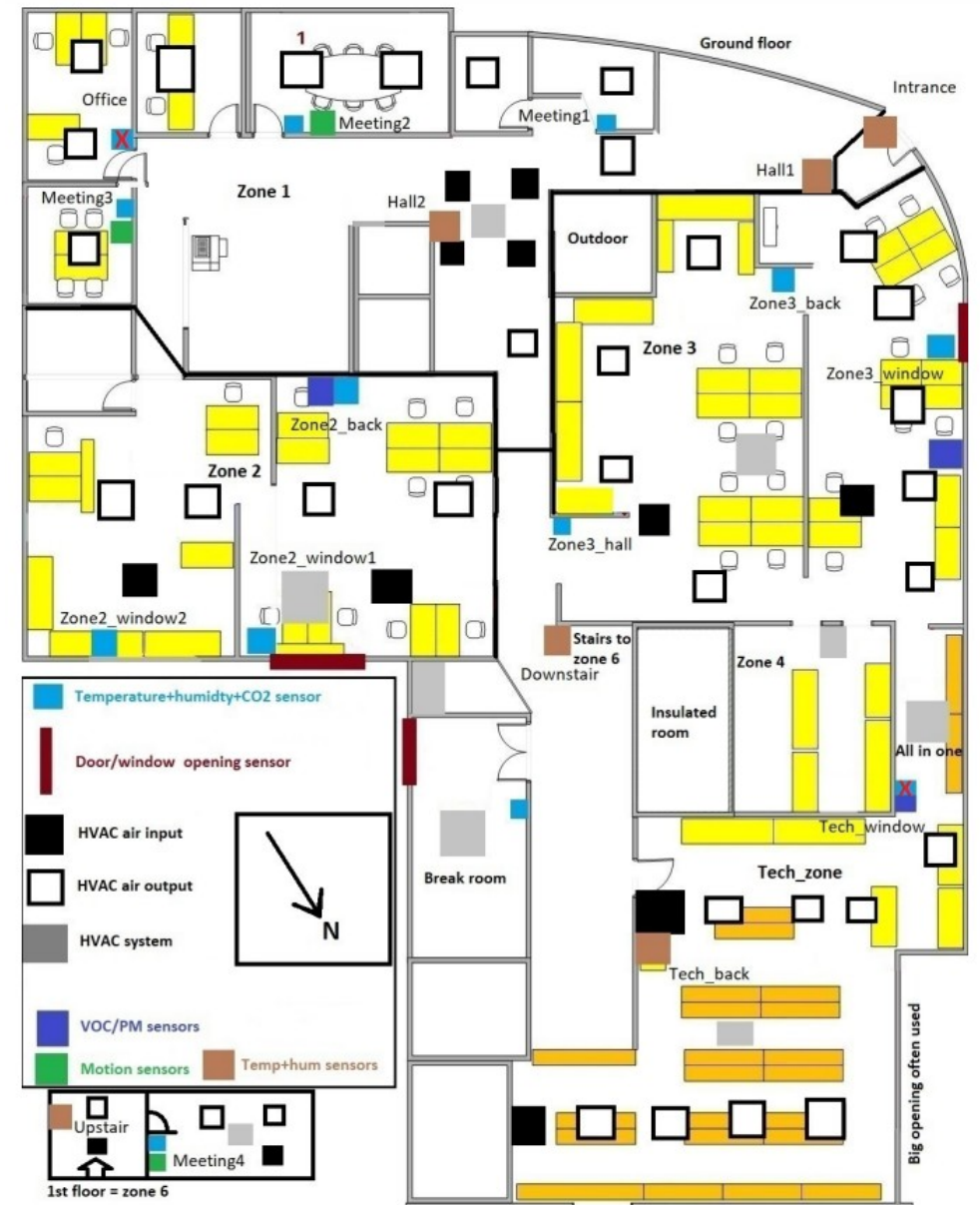
Mod liser le fonctionnement du b timent
Pour en am liorer l'usage

Donn es - B timent instrument 

Contr le manuel mural :

- Consigne variable
- Pas de retour consigne

- B timent tertiaire sans gestion centrale
- Capteurs connect s en LoRa
- 5 CVC instrument es sur 7



Motivations

Meilleure gestion :

Mod lisation de syst mes

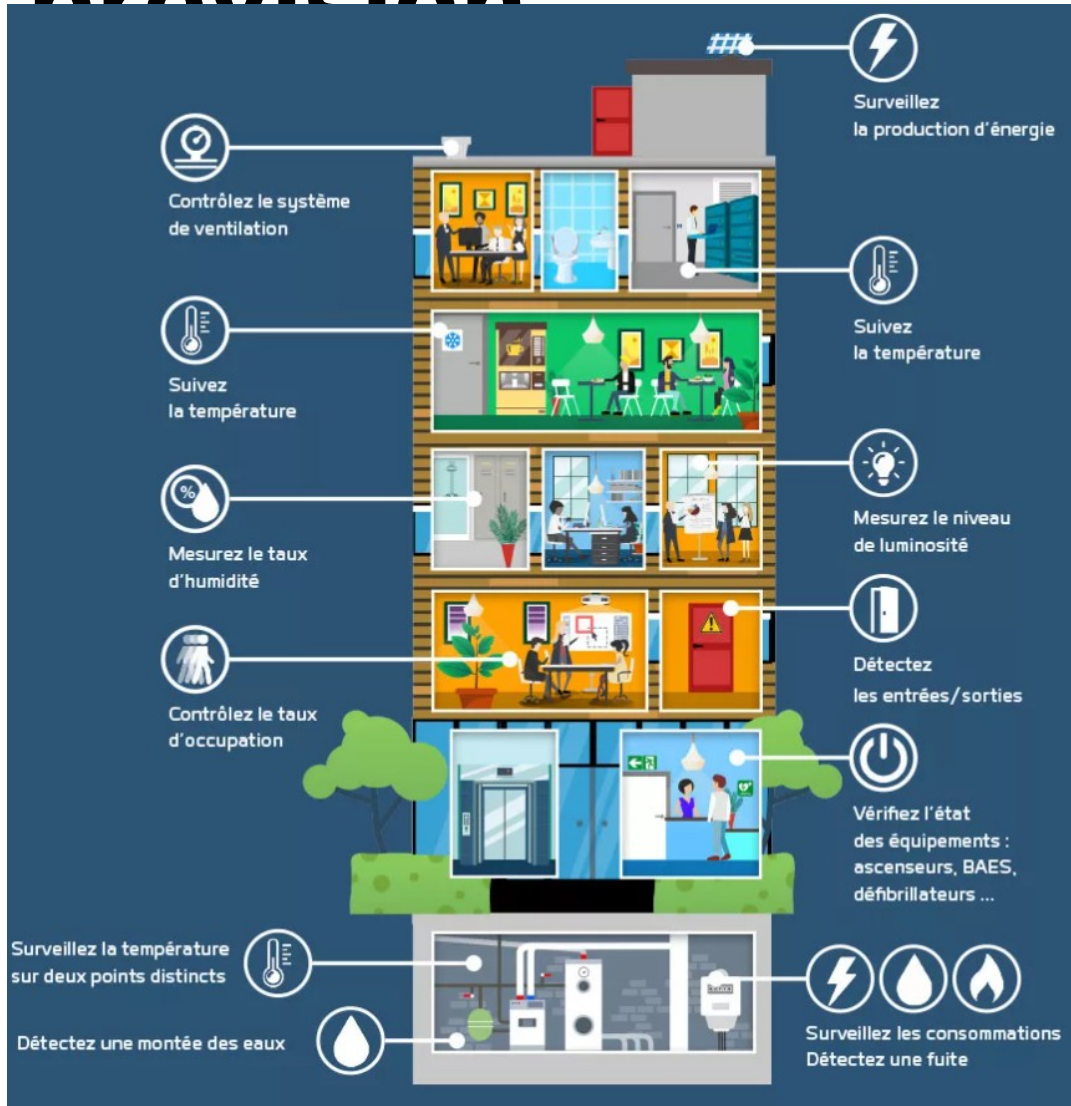
Contr le horaire $\Rightarrow$ Contr le dynamique $\Rightarrow$ Contr le pr dictif

D tection d'anomalie rapide

Maintenance pr dictive

Du bâtiment physique à la

prédiction



?



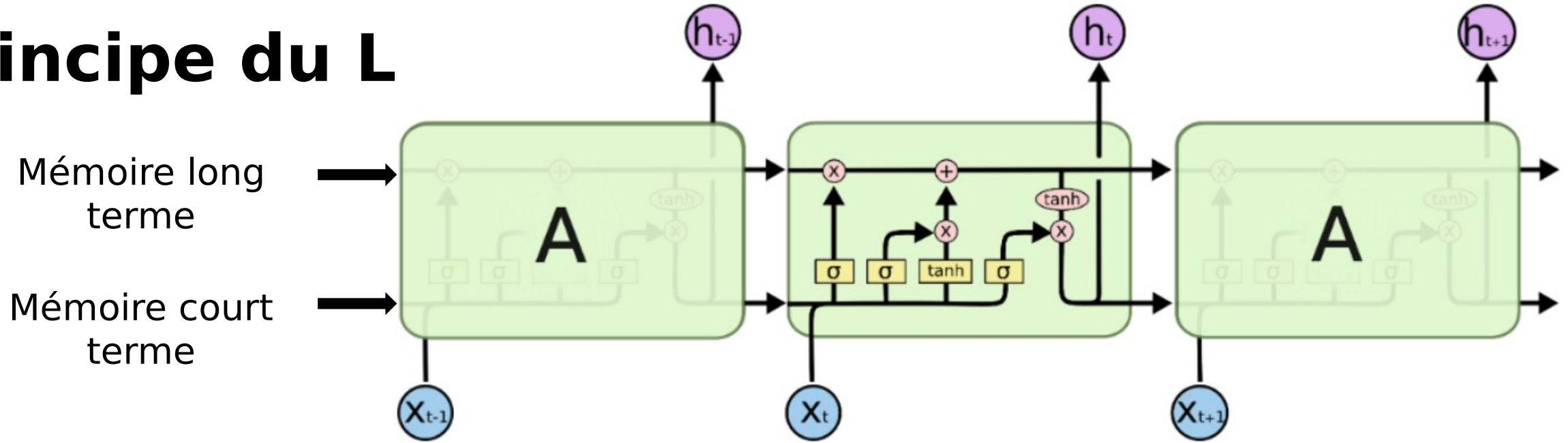
**Modèle prédictif
multivarié**

Compétition de prévision de séries temporelles toute catégorie (M4) $\Rightarrow$ podium pour réseaux récurrents

Grand gagnant : **Long Short Term Memory (LSTM)**

[M4 Forecasting Competition](#)

Principe du L



Chaque nouvelle donn e change les 2 m moires

Qu'est-ce que l'algorithme apprend ?   Les cases jaunes

Comment  valuer les performances ?

M trique : le ratio de r duction du RMSE – Root Mean Square Error

$$score = \frac{RMSE_{delay} - RMSE}{RMSE_{delay}}$$

- Compare   la situation sans algorithme
- Donne le gain net de pr cision avec IA et sans algorithme
- Utilisable pour comparer   des interpolations simples

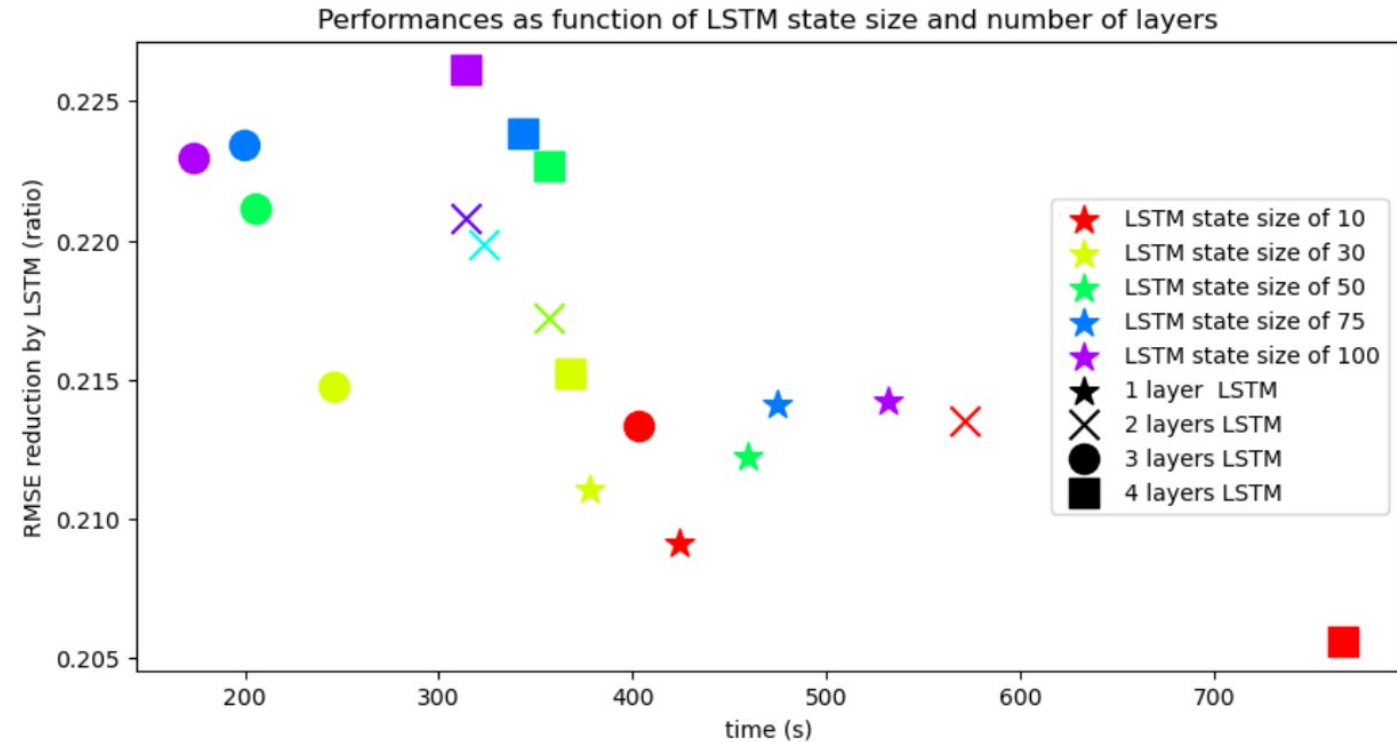
Quelle taille pour le réseau LSTM ?

Essais de différentes tailles et de LSTM et de nombres de couches superposées

Observations : Il y a une taille optimale.

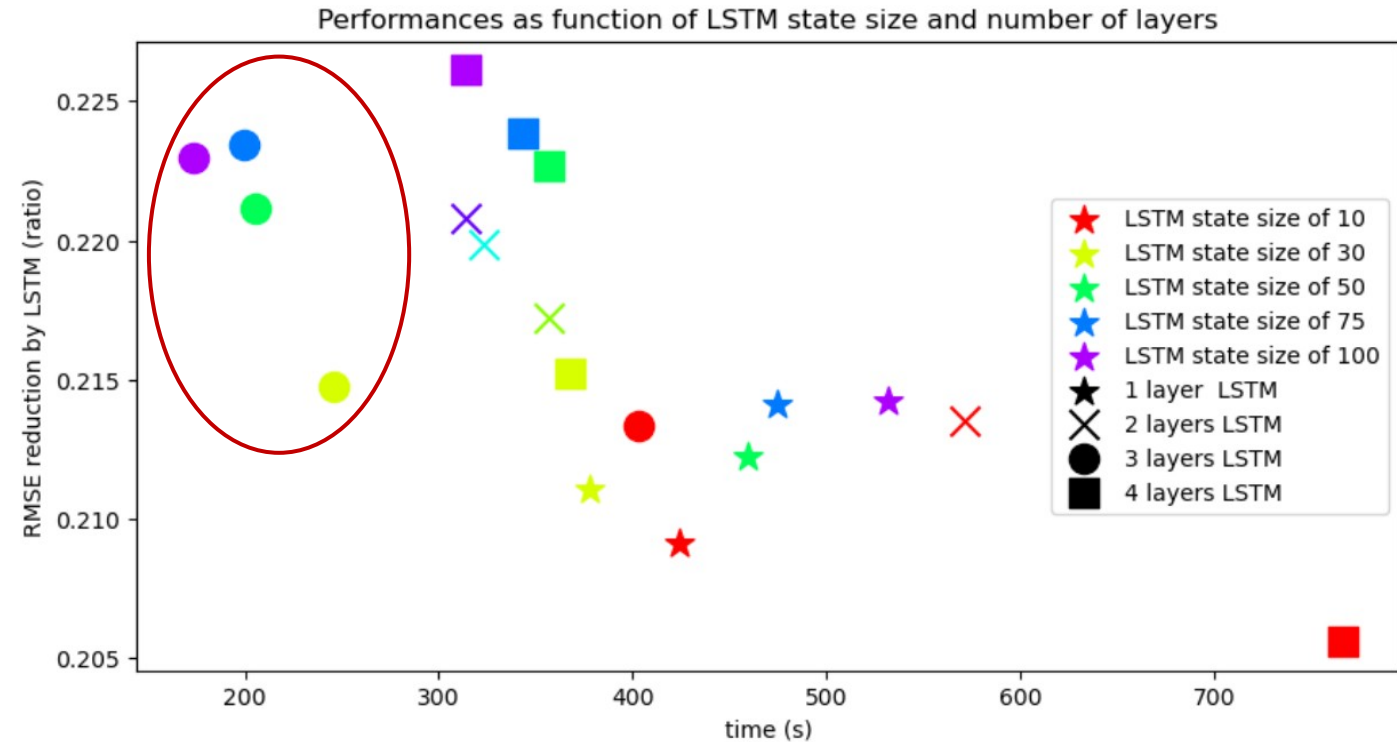
Compromis entre

- Apprentissage réguliers grâce à assez de paramètres
- Trop de paramètres et calculs inutiles (ralentit la convergence)



Quelle taille pour le réseau LSTM ?

Essais de différentes tailles et de LSTM et de nombres de couches superposées



Observations : Il y a une taille optimale.

Compromis entre

- Apprentissage réguliers grâce à assez de paramètres
- Trop de paramètres et calculs inutiles (ralentit la convergence)

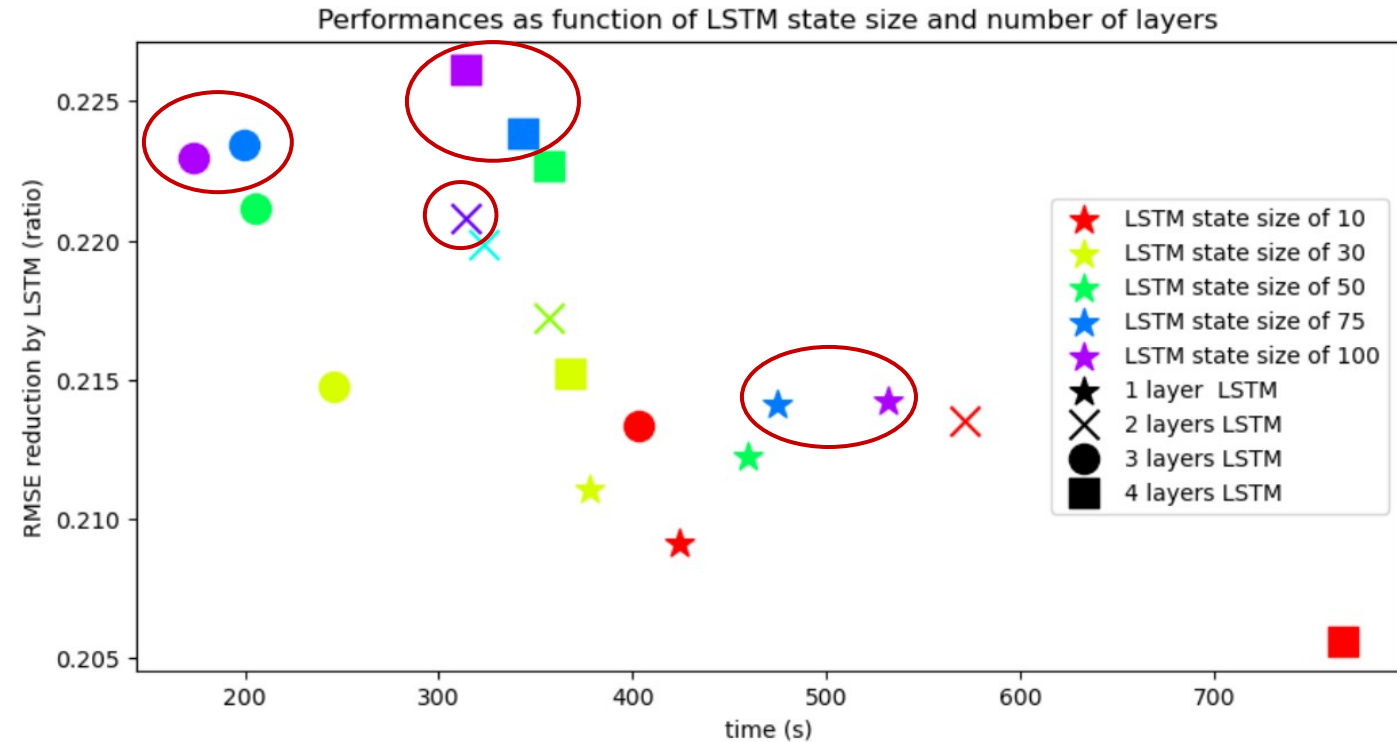
Quelle taille pour le réseau LSTM ?

Essais de différentes tailles et de LSTM et de nombres de couches superposées

Observations : Il y a une taille optimale.

Compromis entre

- Apprentissage réguliers grâce à assez de paramètres
- Trop de paramètres et calculs inutiles (ralentit la convergence)

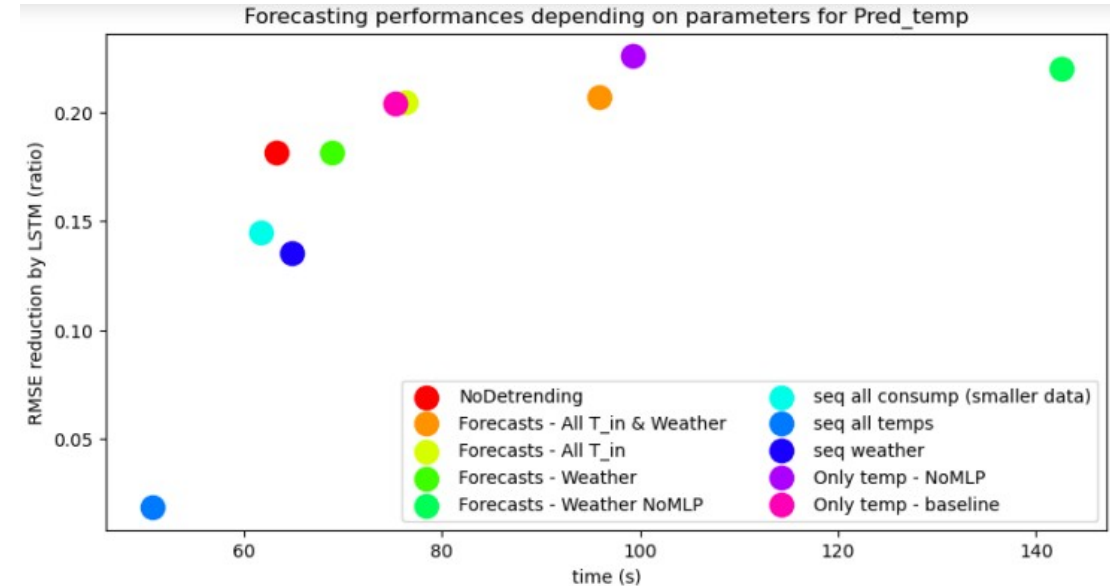


Quels types de données exploiter ?

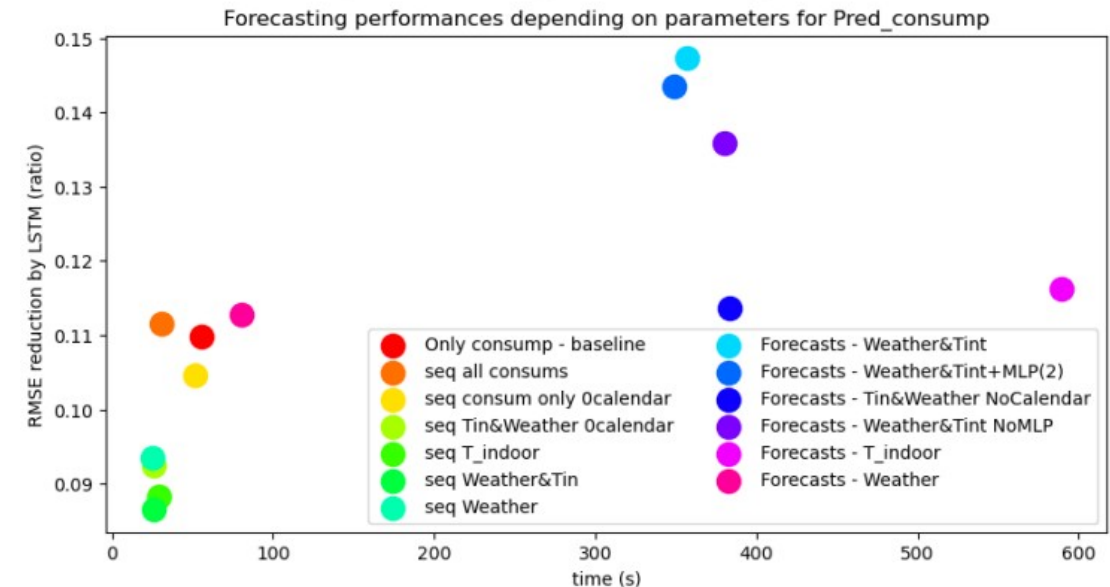
Essais de différentes combinaisons de variables, pour prédire température et consommation des CVC

Observations :

- +efficace = +long
- Températures plus pertinentes pour conso



(a) Performances de la prédiction de températures



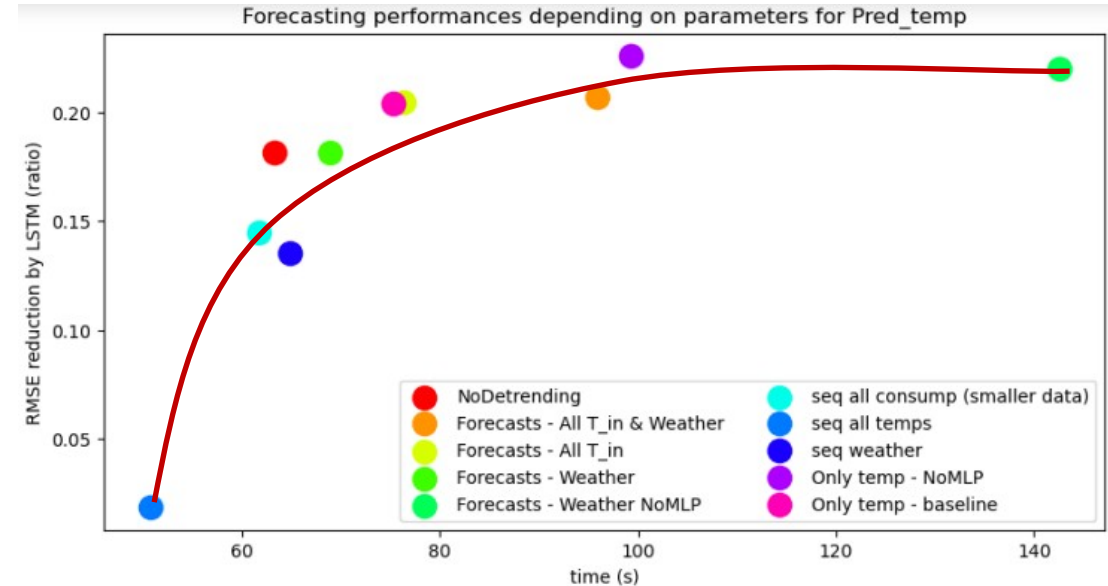
(b) Performances de la prédiction de consommation

Quels types de données exploiter ?

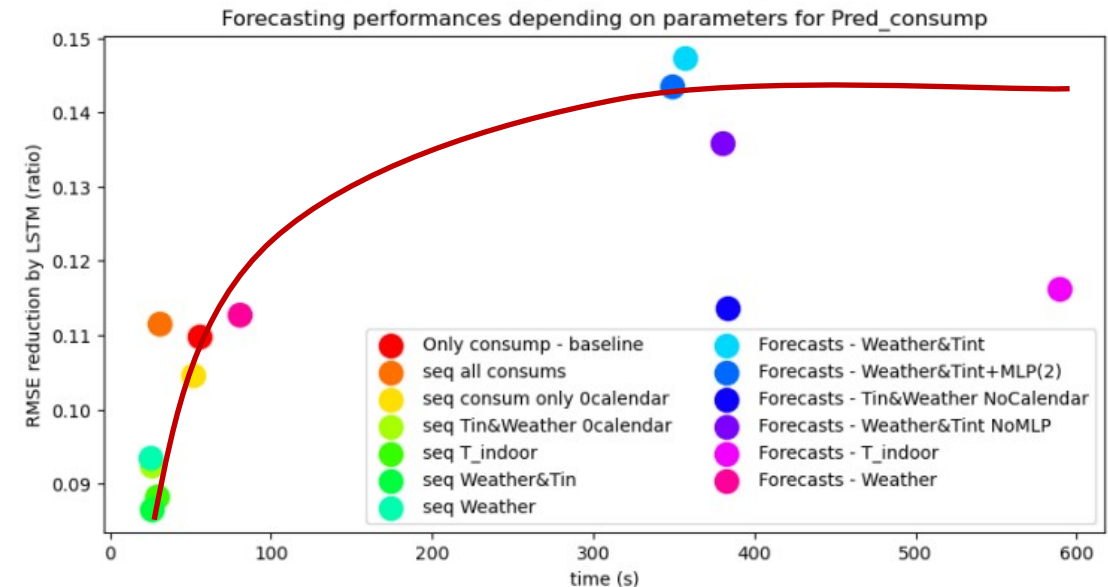
Essais de différentes combinaisons de variables, pour prédire température et consommation des CVC

Observations :

- +efficace = +long
- Températures plus pertinentes pour conso



(a) Performances de la prédiction de températures



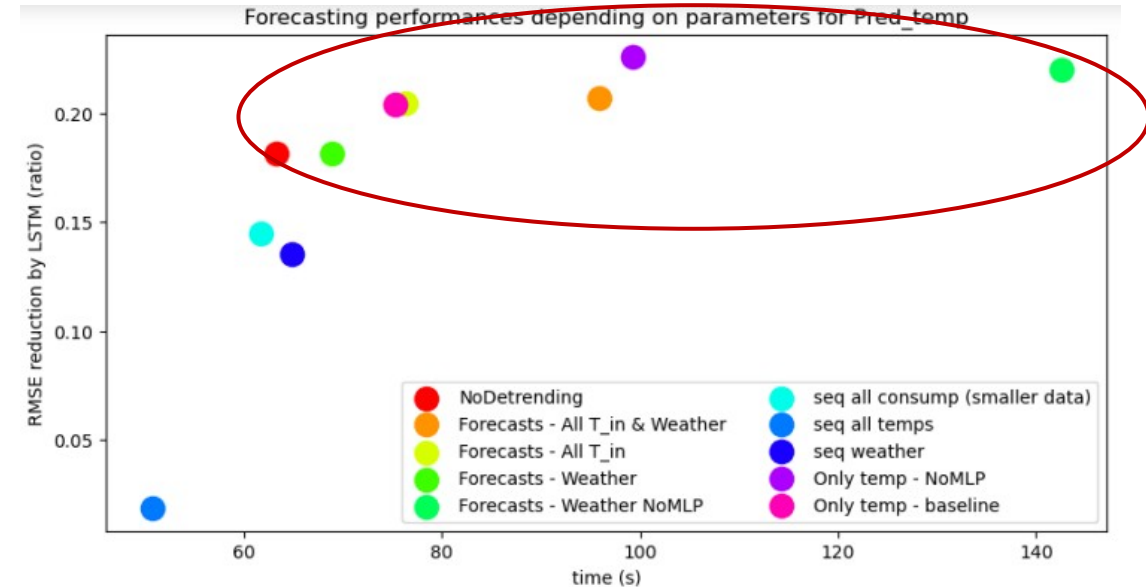
(b) Performances de la prédiction de consommation

Quels types de données exploiter ?

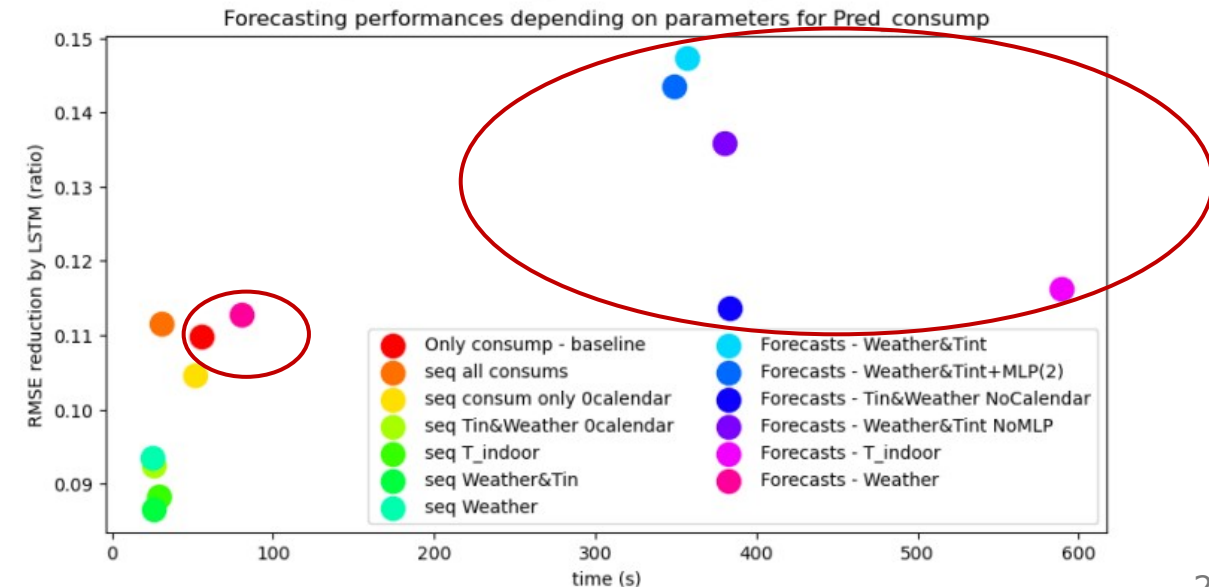
Essais de différentes combinaisons de variables, pour prédire température et consommation des CVC

Observations :

- +efficace = +long
- Plusieurs réseaux mieux qu'un unique réseau



(a) Performances de la prédiction de températures



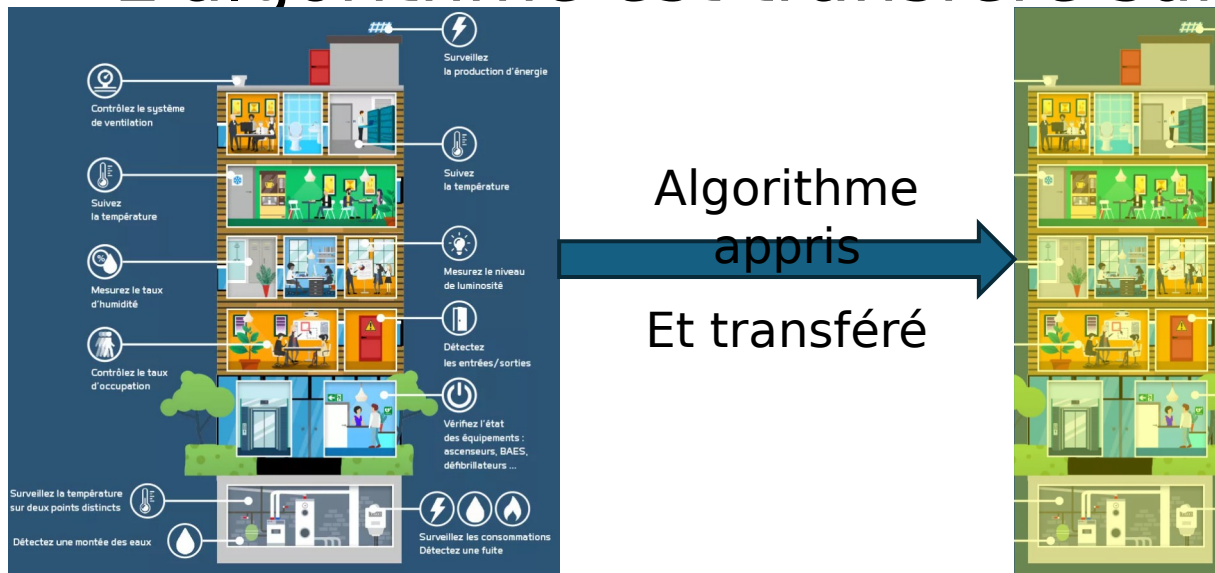
(b) Performances de la prédiction de consommation

IV. Apprentissage par transfert

Comment mieux exploiter les nombreuses données issues de l'instrumentation des bâtiments ?

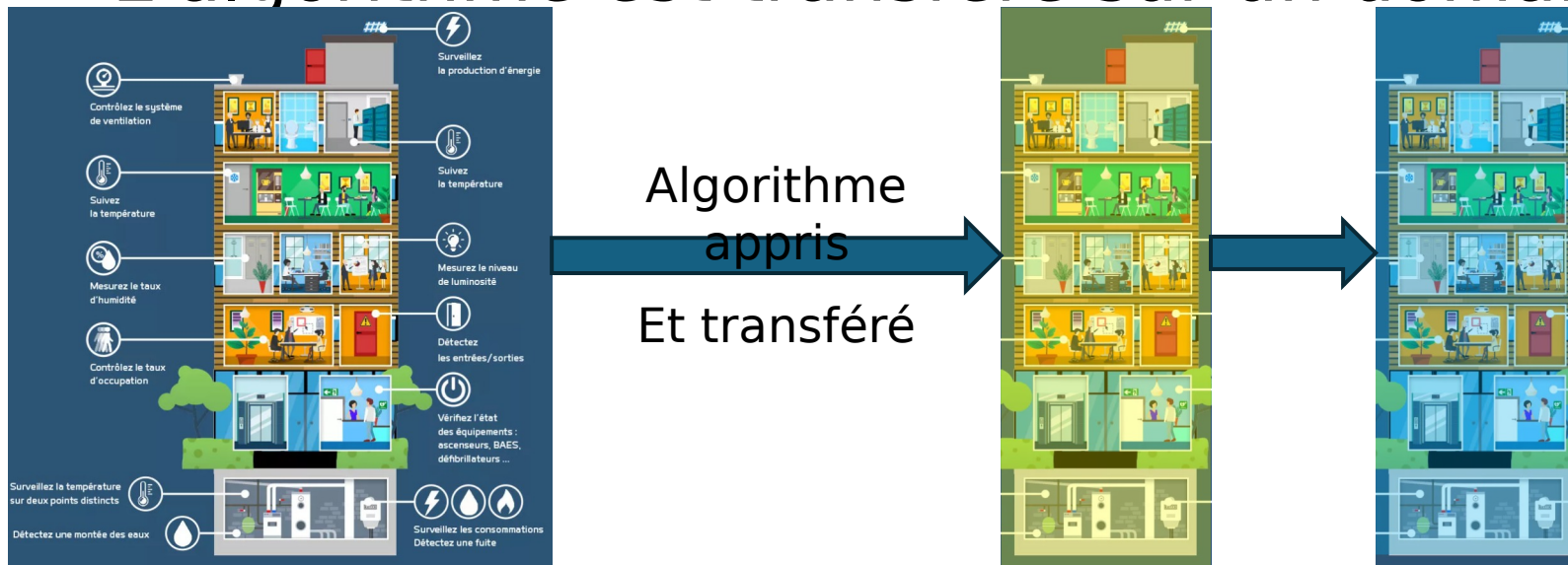
Motivation et principes

- Réduire les coûts d'apprentissage et durées d'observation
- En capitalisant les connaissances déjà apprises
 - Un apprentissage est fait sur un domaine source
 - L'algorithme est transféré sur un domaine cible



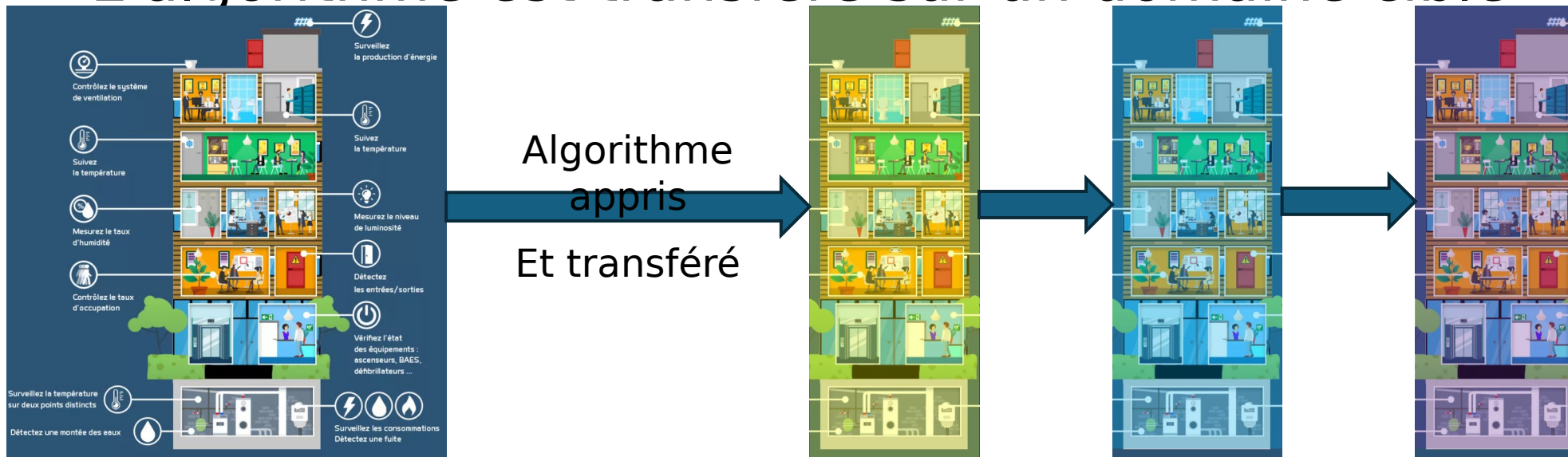
Motivation et principes

- Réduire les coûts d'apprentissage et durées d'observation
- En capitalisant les connaissances déjà apprises
 - Un apprentissage est fait sur un domaine source
 - L'algorithme est transféré sur un domaine cible



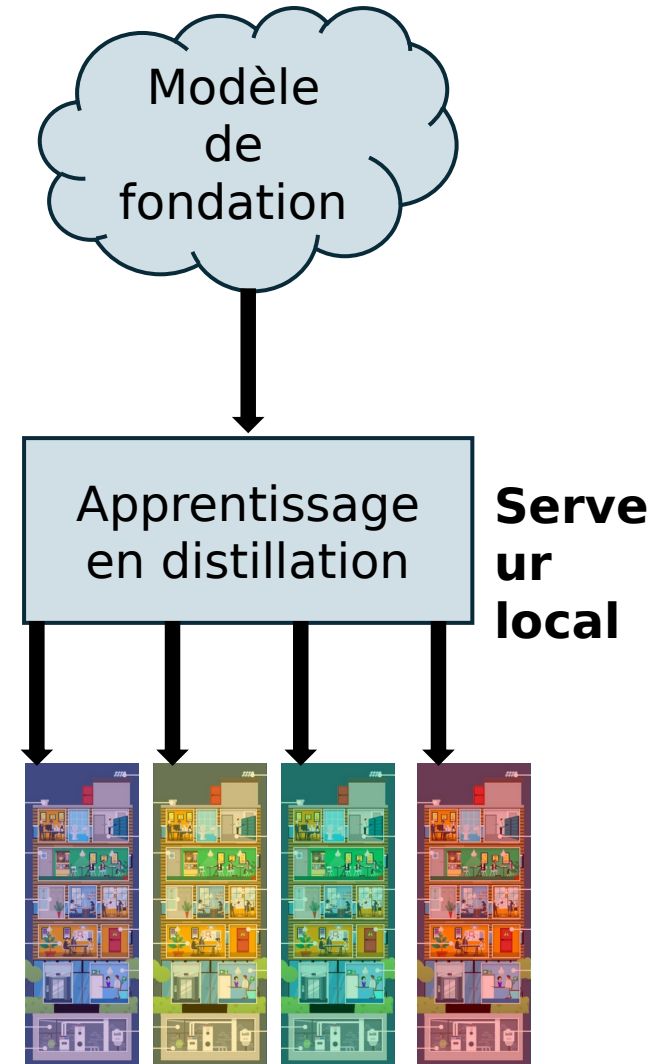
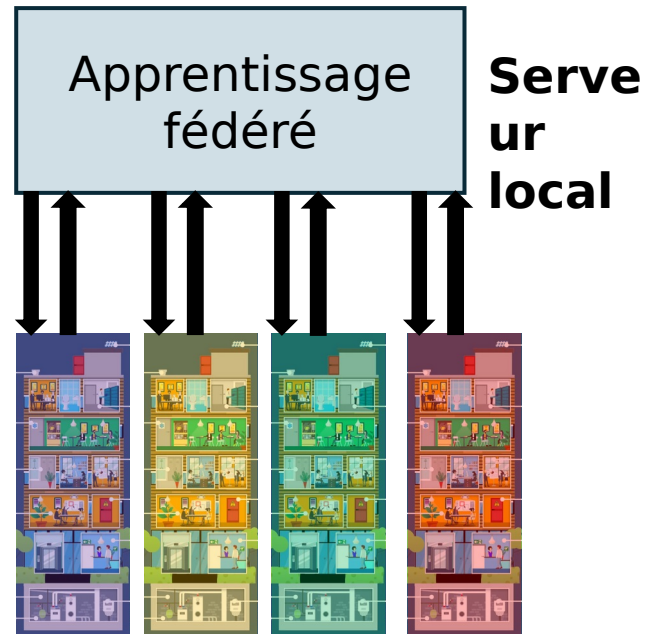
Motivation et principes

- Réduire les coûts d'apprentissage et durées d'observation
- En capitalisant les connaissances déjà apprises
 - Un apprentissage est fait sur un domaine source
 - L'algorithme est transféré sur un domaine cible



Motivation et principes

Transfert pouvant être effectuée sur le edge



Des algorithmes utilisables partout ?

$$Metric = \frac{RMSE_{InfAlgo}(Y_{ExpAlgo})}{RMSE_{ExpAlgo}(Y_{ExpAlgo})}$$

Y	Données utilisées pour l'inférence
ExpAlgo	Abscisse
InfAlgo	Ordonnée



On **infère** l'algorithme entraîné sur « ordonnée » aux données de « abscisse »

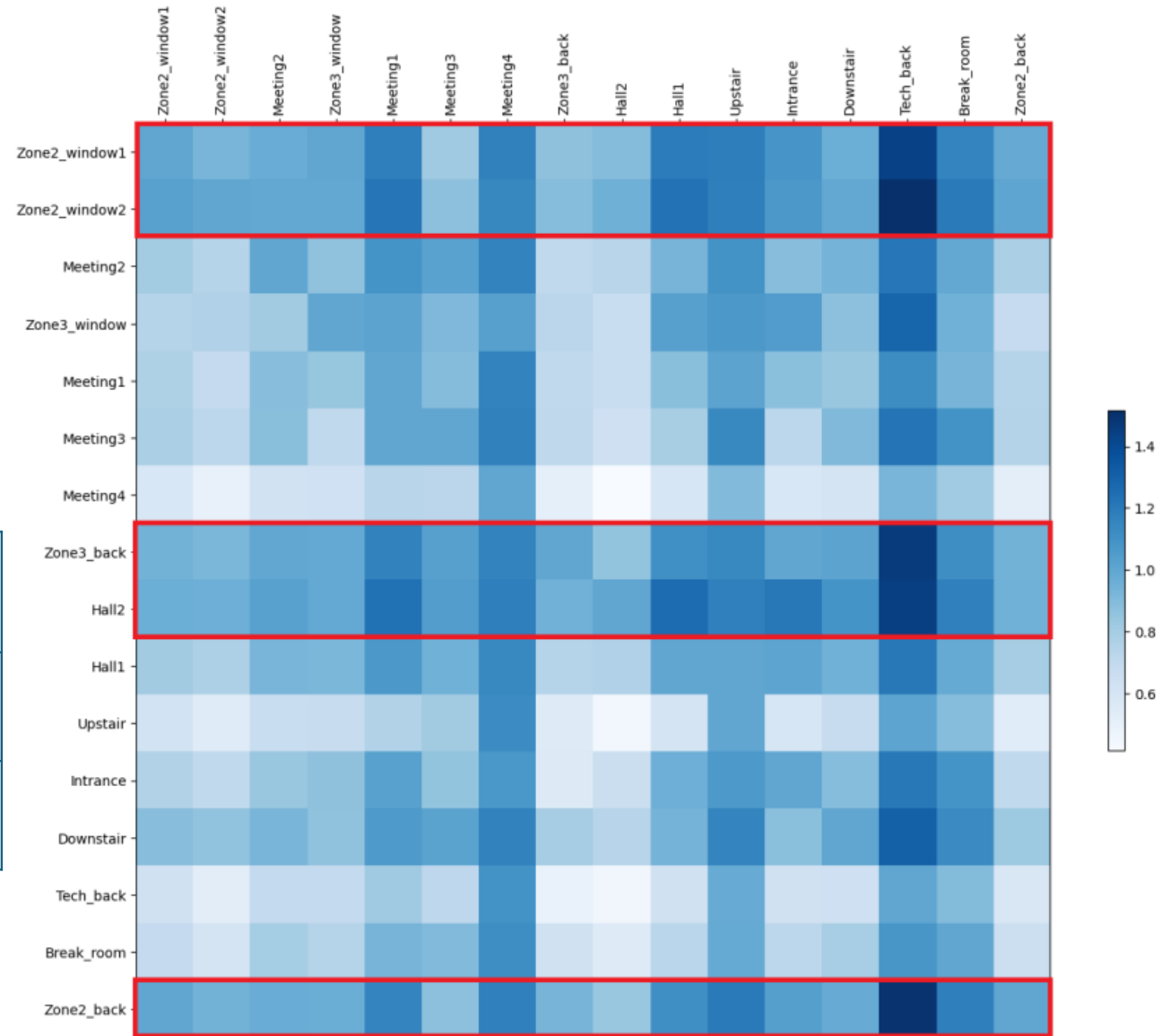
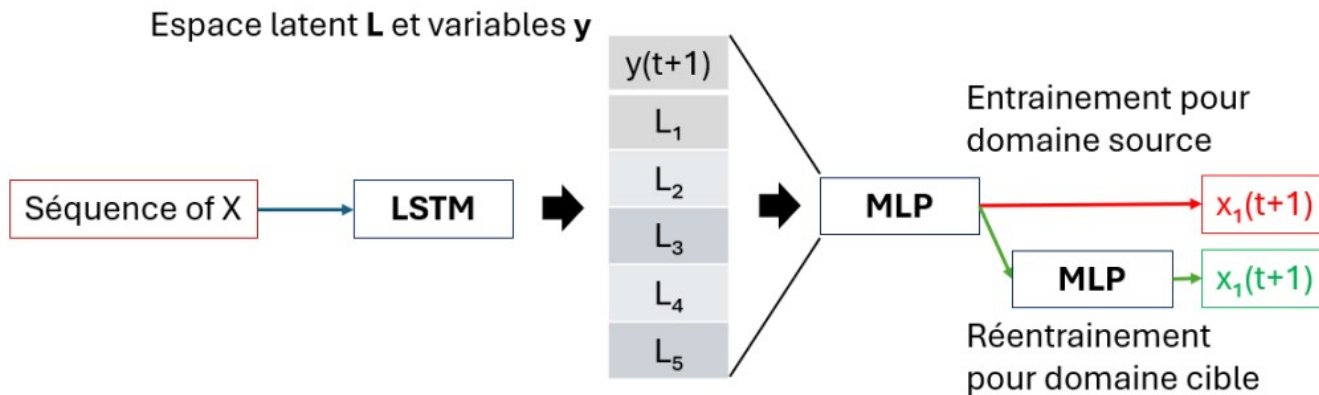
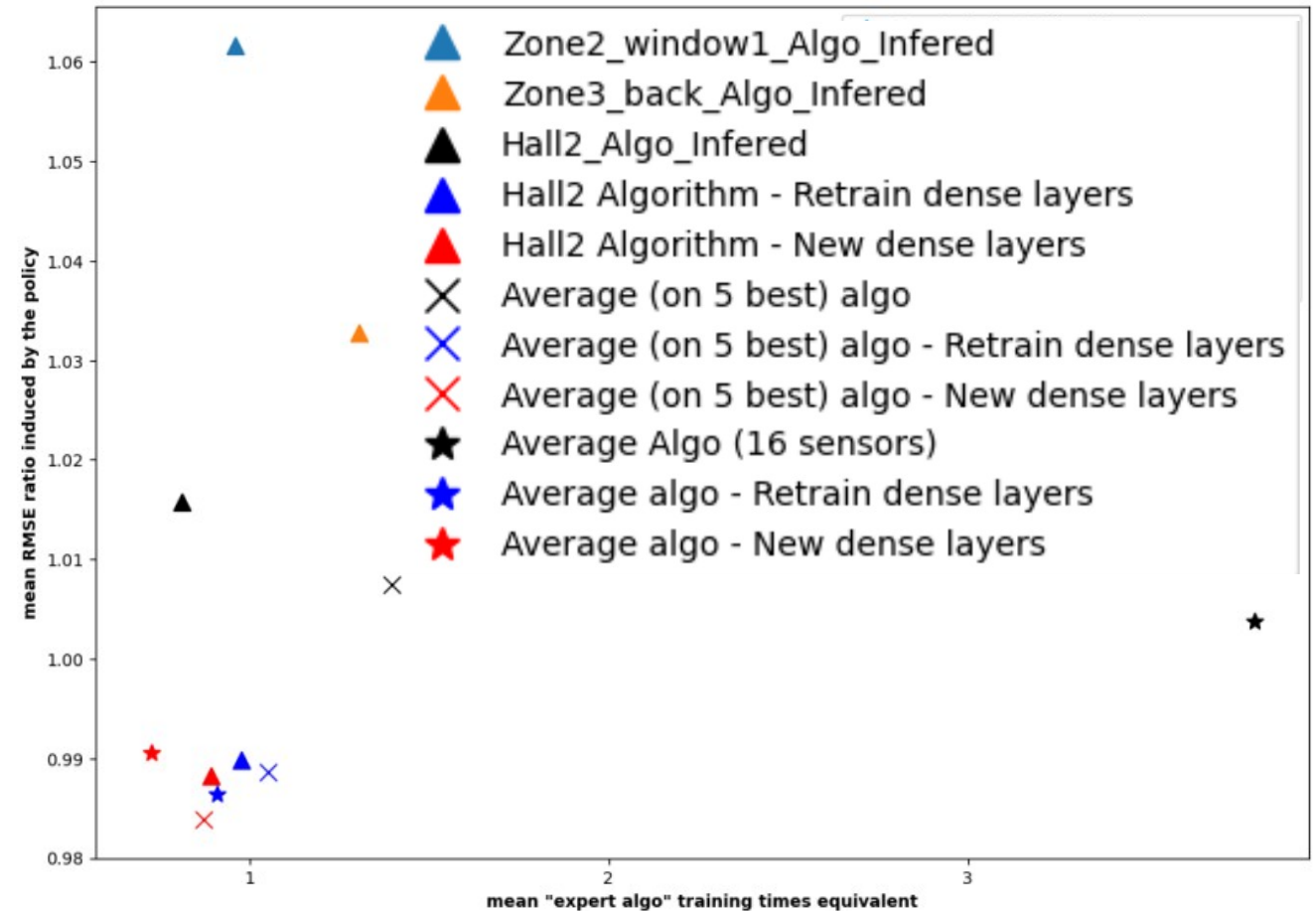


FIGURE 5.4 – Matrice de transférabilité

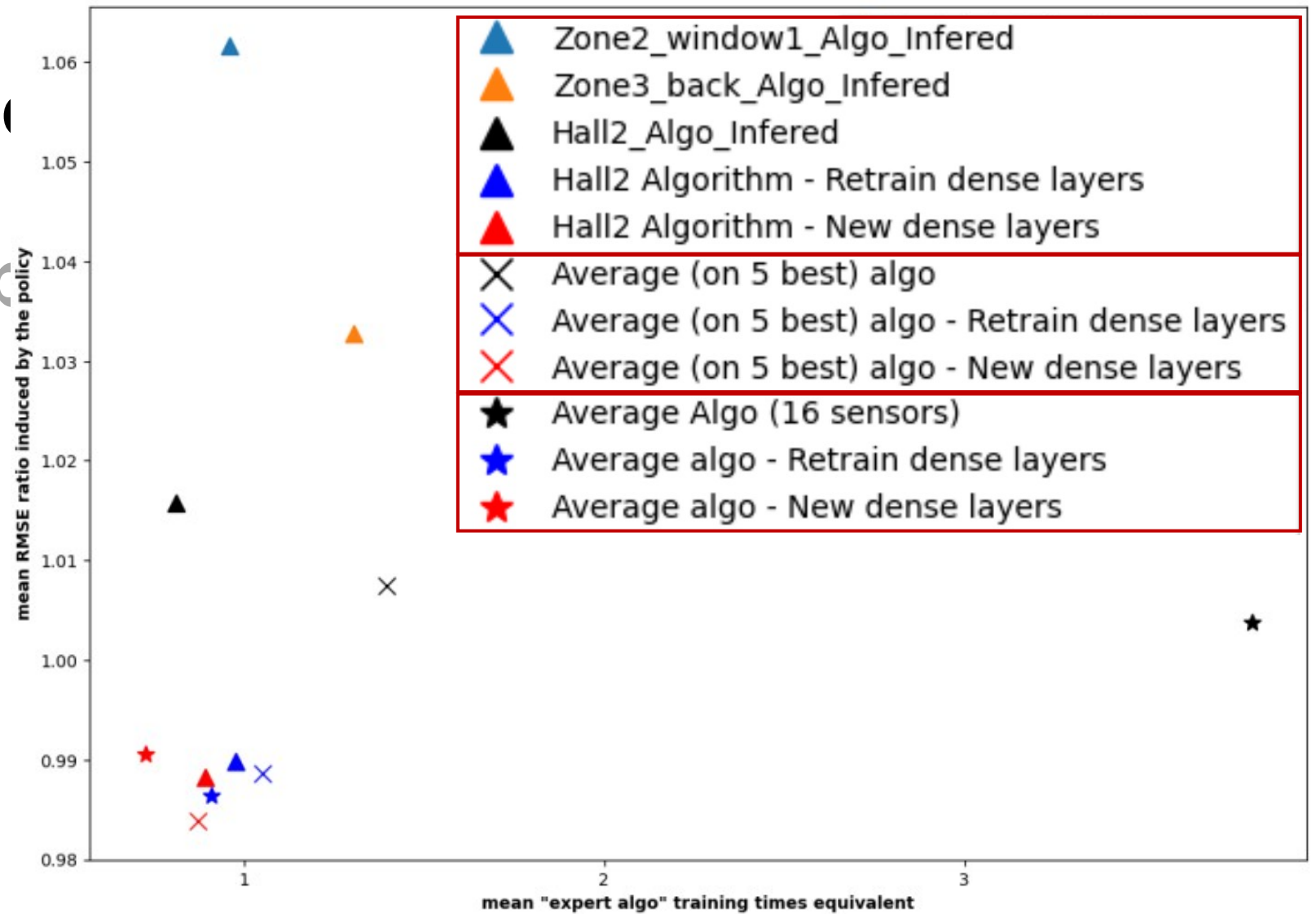
**Quelles données exploiter ?
Comment adapter l'algorithme ?**



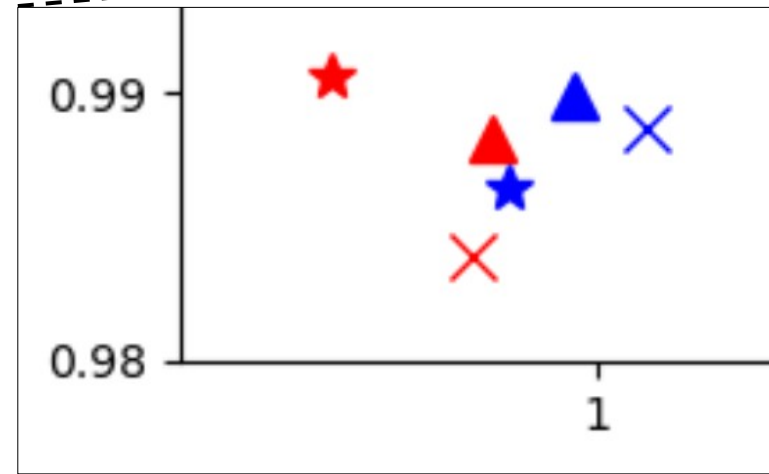
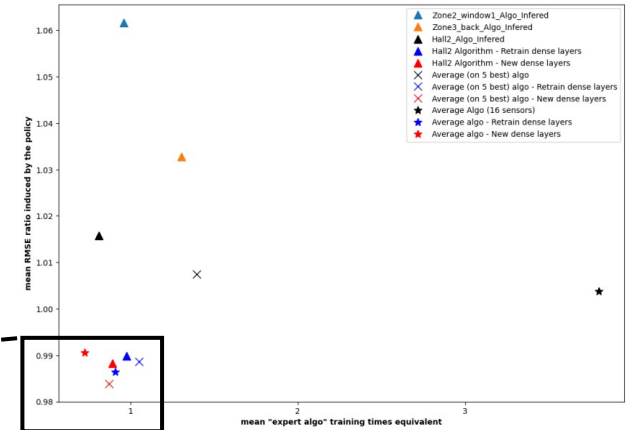
Algorithme moyen et nouveau réseau
➤ 25% de réduction de temps d'entraînement

**Quelles données
exploiter ?**

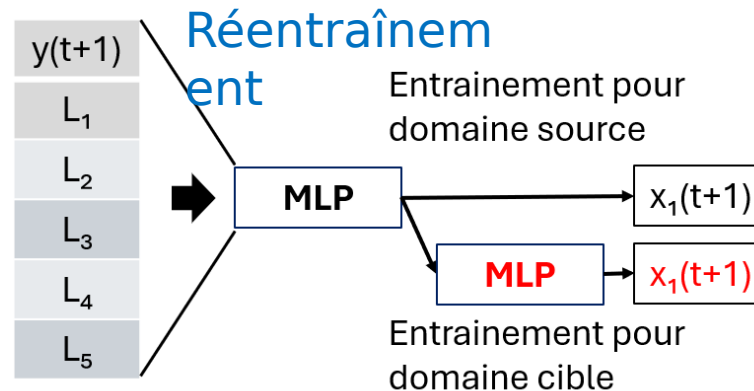
**Comment adapter
l'algorithme ?**



Quelles données exploiter ? Comment adapter l'algorithme ?



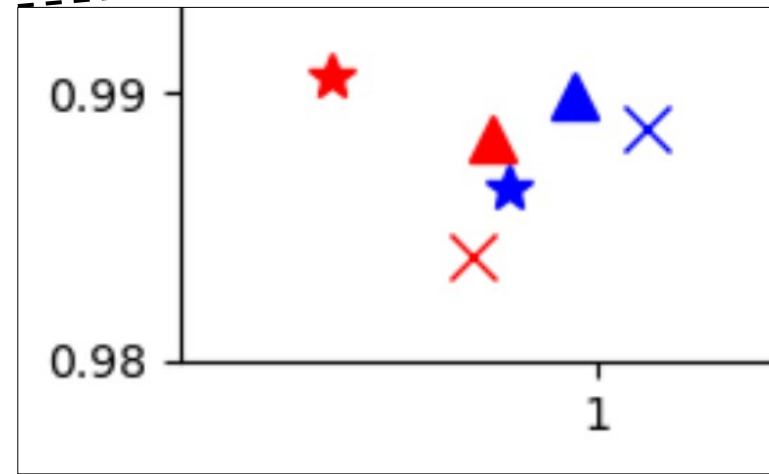
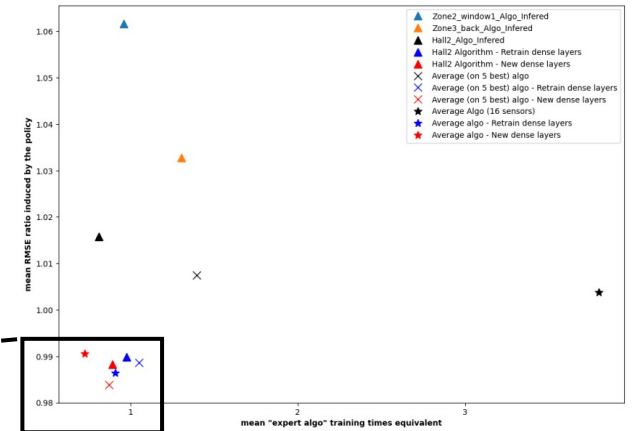
Espace latent L et variables y



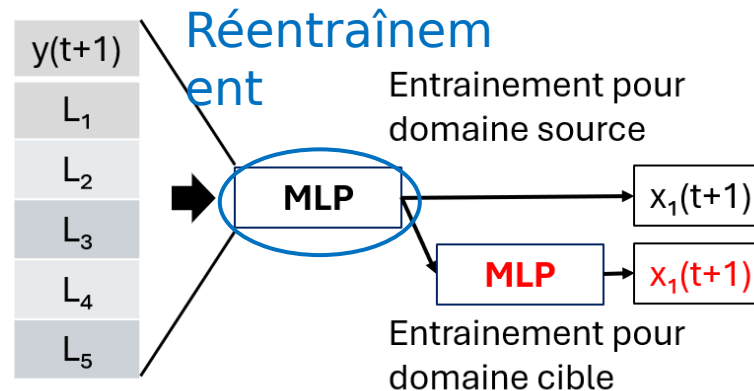
Algorithme moyen et nouveau réseau

- 25% de réduction de temps d'entraînement

Quelles données exploiter ?
Comment adapter l'algorithme ?



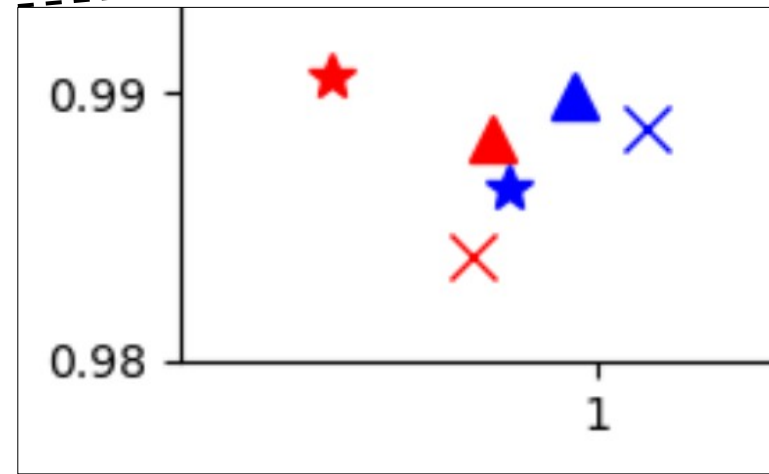
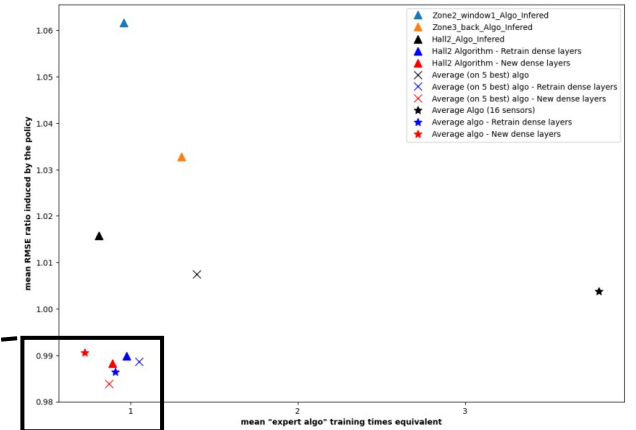
Espace latent L et variables y



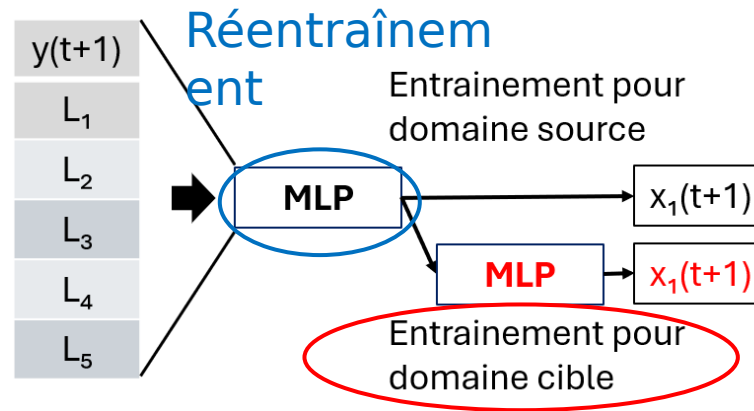
Algorithme moyen et nouveau réseau

- 25% de réduction de temps d'entraînement

Quelles données exploiter ? Comment adapter l'algorithme ?



Espace latent L et variables y



Algorithme moyen et nouveau réseau

- 25% de réduction de temps d'entraînement

Anticiper les performances de l'IA

Utilisation des **données seules** : comparer des séries temporelles

Expérience :

- Une métrique est calculée pour mesurer la **similarité** entre chaque série temporelle
- Un **classement** est effectué (lignes puis colonnes). On regarde

Métrique	MARE-transmit	MARE-expert	durée(s)
corrélacion de Pearson	3.0	7.1	0.3
métrique DTW	4.5	5.9	51
distances entre MP	2.1	8.0	1189

Processeur:

AMD Ryzen 7 4800HS with
Radeon Graphics 2.90 GHz

GPU: GEFORCE RTX 3080
max-Q

Pour conclure

Pertinence des regroupements et anomalies détectées

- Connaissances experte requise

Efficacité démontrée des LSTM

- Compromis entre simplicité et flexibilité
- Pour l'architecture des réseaux, et les grandeurs utilisées

Efficacité des politiques de transfert

- Pour réduire le temps d'apprentissage
- Pour réduire la durée d'observation

Pistes de recherche entamées

Utilisation des jeux de données supplémentaires obtenus

- Données de calendrier
- Campagnes de mesures UGA

Modélisation de bâtiments (BIM=Building Information Modélisation)

- Utilisation des plans de masse
- Vers des jumeaux numériques

Prise en compte des différents niveaux d'usage et de complexité

- Services techniques de la ville de La Mure
- Direction du patrimoine de l'UGA
- Gestionnaires Eiffage du bâtiment GreEn-Er

Perspectives sur l'IA pour le bâtiment intelligent

Contrôlabilité

Frugalité

Localité

Résilience et capitalisation

- temps, espace, variables disponibles
- apprentissage continu

Convivialité et explicabilité

- collectivités et entreprises
- usagers

Merci de votre attention



Rappel des contributions :

Towards a Methodology for the Characterization of IoT Data Sets of the Smart Building Sector - Louis Closson, Christophe Cérin, Didier Donsez et Denis Trystram. In 2022 IEEE International Smart Cities Conference (ISC2), pages 1-7, September 2022. ISSN : 2687-8860

Design of a Meaningful Framework for Time Series Forecasting in Smart Buildings - Louis Closson, Christophe Cérin, Didier Donsez et Jean-Luc Baudouin. Information, vol. 15, no. 2, page 94, February 2024. Number : 2 Publisher : Multidisciplinary Digital Publishing Institute.

Towards Online Machine Learning Libraries for Embedded Systems - C. Cérin, M. Sow, D. Donsez and L. C., 2024 IEEE 10th World Forum on Internet of Things (WF-IoT), Ottawa, ON, Canada, 2024, pp. 790-797, doi: 10.1109/WF-IoT62078.2024.10811431.

Deep Transfer Learning in Smart Building and links with Edge Computing - IEEE COMPSAC2025 - Accepté

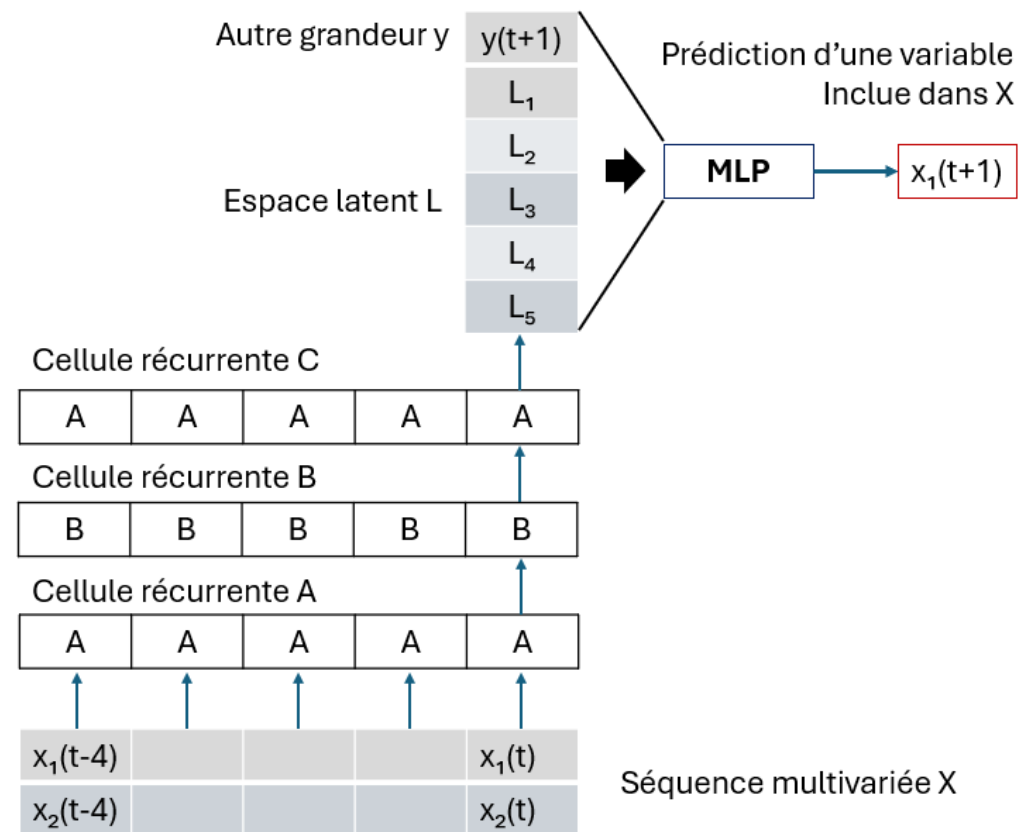
A smart tertiary building dataset - Louis Closson, Didier Donsez, Jean-Luc Baudouin, Denis Trystram et Christophe Cérin. <https://github.com/CampusIoT/datasets/blob/main/BuildPred/README.md>, 2023

Point technique

Nombre de paramètres

Layer (type)	Output Shape	Param #	Connected to
input_5 (InputLayer)	[(1008, 7, 1)]	0	[]
lstm_4 (LSTM)	(1008, 7, 100)	40800	['input_5[0][0]']
lstm_5 (LSTM)	(1008, 7, 100)	80400	['lstm_4[0][0]']
input_6 (InputLayer)	[(1008, 38)]	0	[]
lstm_6 (LSTM)	(1008, 100)	80400	['lstm_5[0][0]']
dense_6 (Dense)	(1008, 38)	1482	['input_6[0][0]']
concatenate_2 (Concatenate)	(1008, 138)	0	['lstm_6[0][0]', 'dense_6[0][0]']
dense_7 (Dense)	(1008, 10)	1390	['concatenate_2[0][0]']
dense_8 (Dense)	(1008, 1)	11	['dense_7[0][0]']

=====
 Total params: 204,483
 Trainable params: 204,483
 Non-trainable params: 0



Rôle dans Flexitron

Mettre en place des projets d'IA en interne :

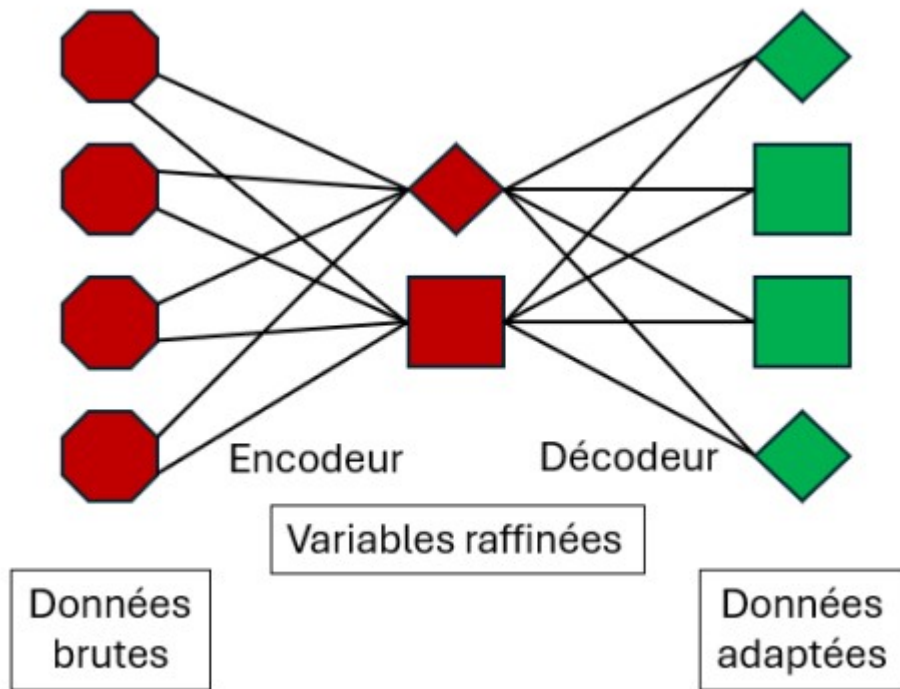
- Utilisation d'API
- Aide au support, au code

Créer de nouvelles solutions :

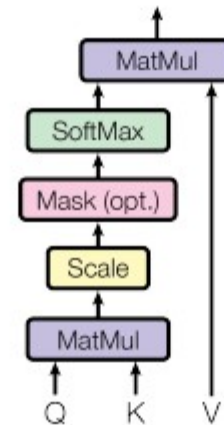
- Pour la gestion de parc photovoltaïque
- Pour la gestion du bâtiment

Point technique - Transformers

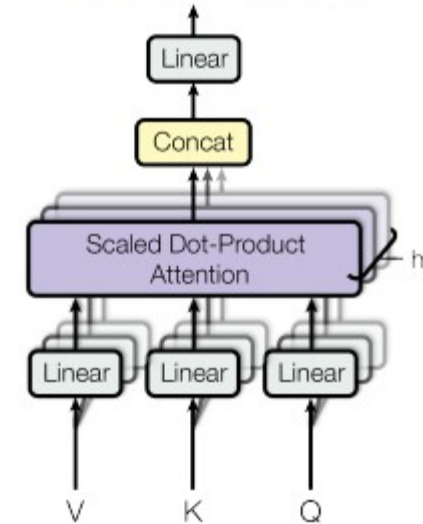
Encodeur - Décodeur



Scaled Dot-Product Attention

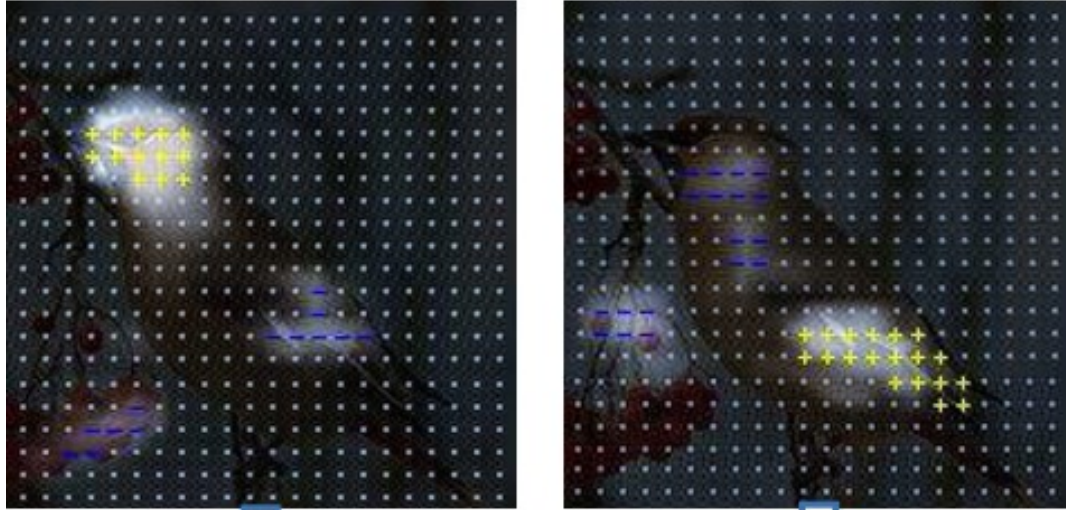


Multi-Head Attention



Attention Is All You Need,
31st Conference on Neural Information Processing
Systems (NIPS 2017), Long Beach, CA, USA

Point technique - Transformers



(a) head



(b) wing

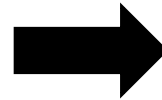
H. Zheng, J. Fu, T. Mei and J. Luo, "Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 5219-5227, doi: 10.1109/ICCV.2017.557.

keywords: {Birds;Head;Feature extraction;Convolution;Visualization;Training;Proposals},

Point technique

Complétion des données

Estampil le	Grandeu r 1	Grandeu r 2	Grandeu r 3
8	X	X	X
11	X		X
19		X	X
25		X	X
30	X		



Estampil le	Grandeu r 1	Grandeu r 2	Grandeu r 3
8	X	X	X
10	INTERP	INTERP	INTERP
11	X	INTERP	X
19	INTERP	X	X
20	INTERP	INTERP	INTERP
25	INTERP	X	X
30	X	NULL	NULL



Estampil le	Grandeu r 1	Grandeu r 2	Grandeu r 3
10	INTERP	INTERP	INTERP
20	INTERP	INTERP	INTERP

Pas de remplissage de plus de ½ heure

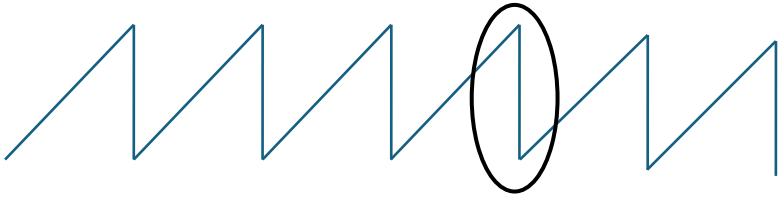
Pas régulier :

- Environ 1 minute puis 5 minutes pour GreEn-Er
- Environ 10 minutes pour le bâtiment d'Adeunis

III. Technical points

3. Time encoding

Time encoding before :



Order relationship problems

Solution

0h	1h	...		Heure h		...	22h	23h
0	0	0		1		0	0	0

Lundi	Mardi	Mercr edi	Jeudi	Vendre di	Same di	Dimanc he
0	0	0	0	1	0	0

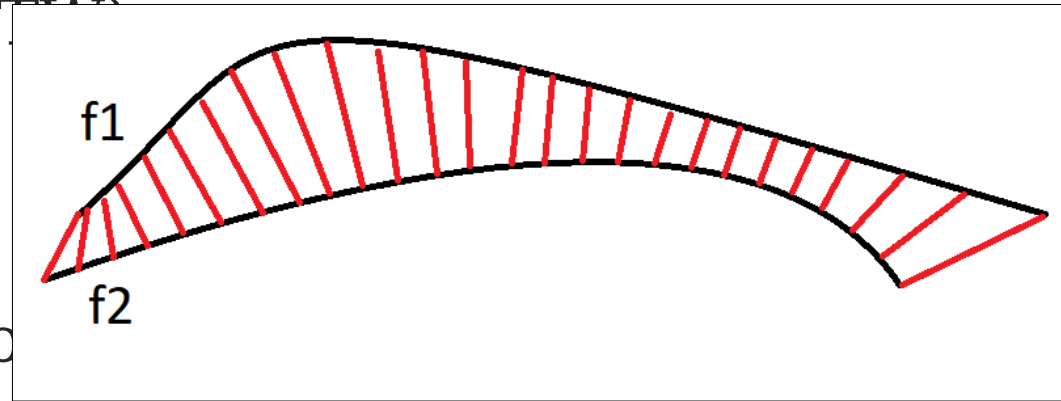


Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW)

Déphasage = ok

Fréquences variables = ok

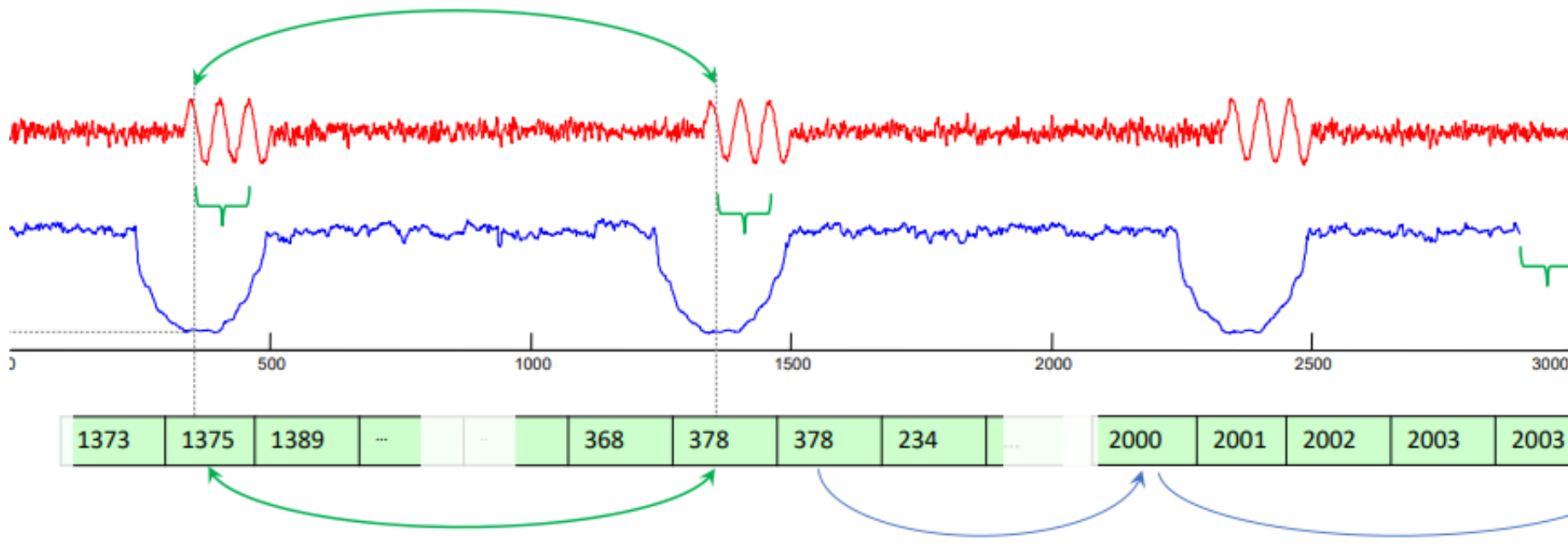


Matrix Profile - [UCR Matrix Profile Page](#)

Note that the pointers in the **matrix profile index** are not necessarily symmetric.

If **A** points to **B**, then **B** may or may not point to **A**

An interesting exception, the two smallest values in the MP must have the same value, and their pointers must be mutual. This is the classic *time series motif*.



Tutoriel 1 – Eamonn Keogh

Point ...

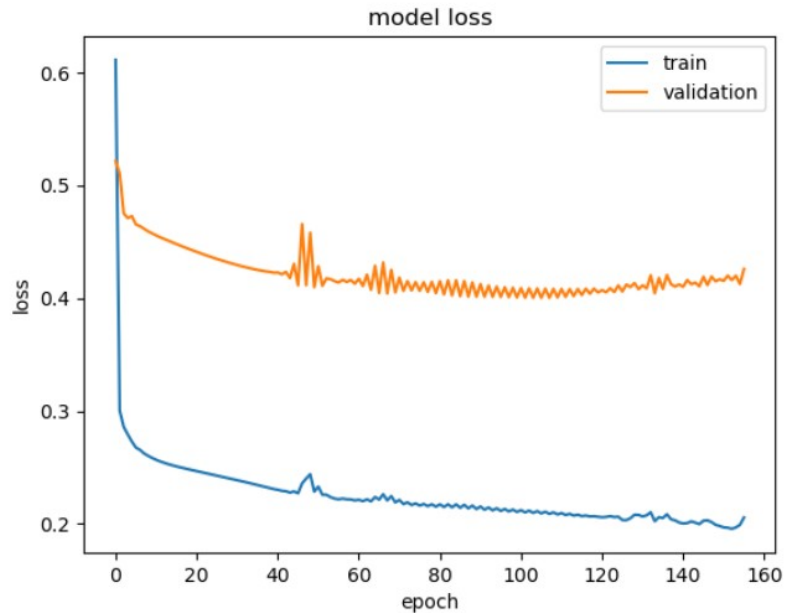
Législation européenne

Rôles dans la récupération des données

Projets de bout en bout

Cyber sécurité (cloud, IoT)

Data forecast – technical part: process



Evolution of RMSE in train and validation, over the training process for an history of 30 minutes
(All done using `keras tensorflow`)

After each model update during training : one inference on each dataset to monitor the training process.

Best model found in 106 epochs.
50 epochs are used to be sure we reached our best model over the training process

Like a student learning with exercices :

- ❑ **Orange** : he looks at other dataset, thus the **dynamics and raisonnement**
- ❑ **Blue** : the same exercice again and again, thus the **answers to this exercice only**