Intelligence Artificielle & Machine Learning pour la modélisation de séries temporelles et de signaux

Introduction

Laurent Oudre
laurent.oudre@ens-paris-saclay.fr

Diplôme ARIA
ENS Paris Saclay
2026-27

Plan du cours

1. Présentation du cours

2. Signaux et séries temporelles

3. Machine Learning pour les séries temporelles

4. Plan du cours

Plan du cours

1. Présentation du cours

2. Signaux et séries temporelles

3. Machine Learning pour les séries temporelles

4. Plan du cours

Quelques informations pratiques

- Quatre séances de cours (10h)
- Projet tutoré en Python sur données réelles (20h)

`http://www.laurentoudre.fr/signalml.html`

Le projet

- Trouver sur Internet une base de données de séries temporelles / signaux correspondant à deux classes différentes
- Réussir (à l'aveugle) à construire des descripteurs pertinents permettant de séparer les signaux, en se basant sur les outils vus dans le cours et sur l'expertise métier recueillie dans des articles de recherche
- Exemples de données :

`http://www.timeseriesclassification.com/dataset.php`
`https://www.data.gouv.fr/fr/`
`https://physionet.org/about/database/`

Plan du cours

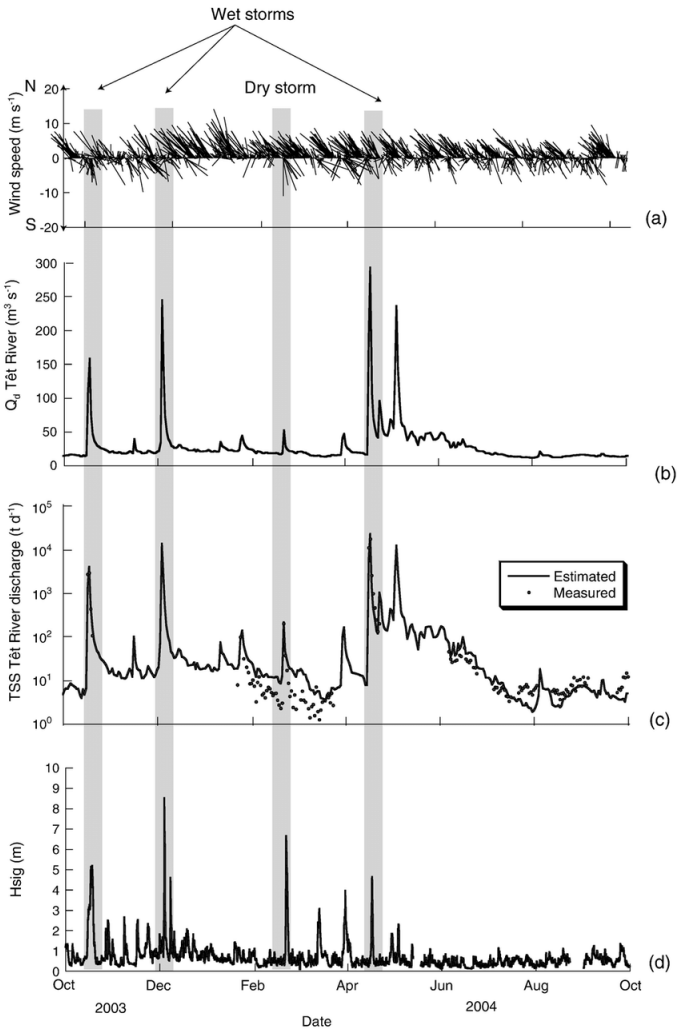
1. Présentation du cours

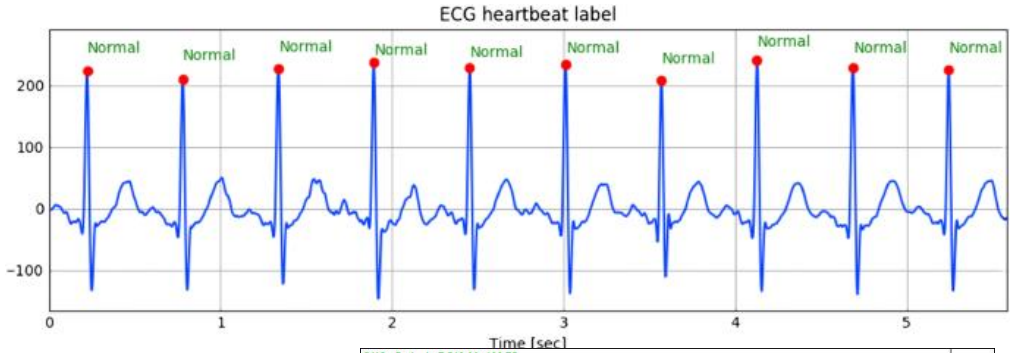
2. Signaux et séries temporelles

3. Machine Learning pour les séries temporelles

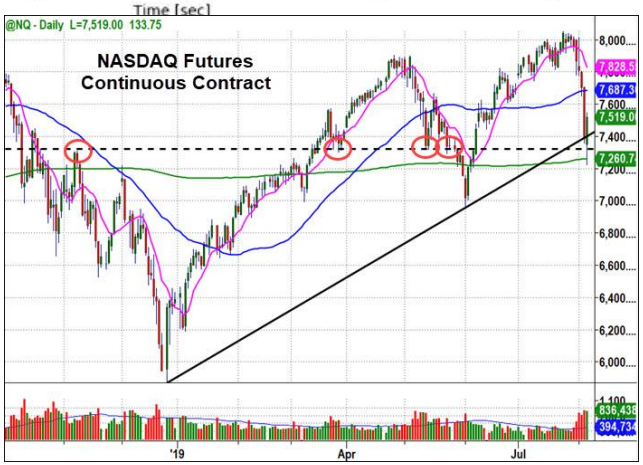
4. Plan du cours

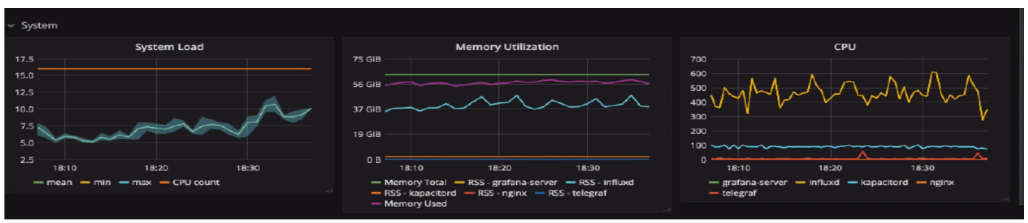
Les séries temporelles sont partout !





Météorologie,
Finance,
Santé,
Monitoring,
Épidémiologie,
Réseaux de capteurs…



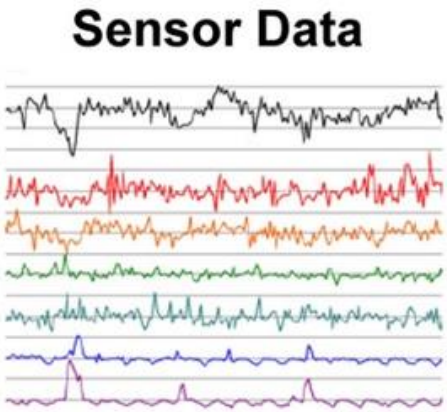


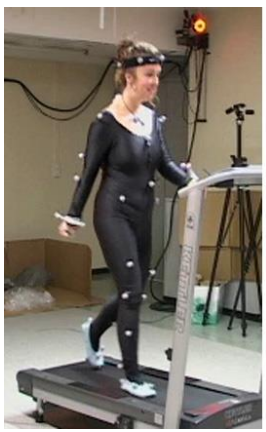
Signaux multivariés

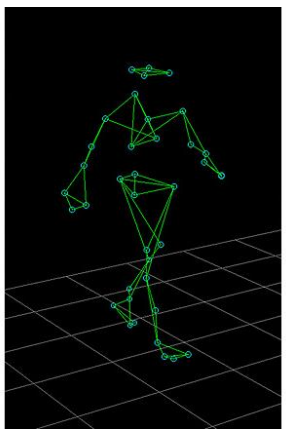
- Certains signaux peuvent correspondre à deux ou plusieurs mesures simultanées
  ex : position de la souris (x, y) enregistrée en fonction du temps
- On parle dans ce cas de **signaux multivariés** et on les représente sous forme de matrices de taille $d \times N$ (où d est le nombre de dimensions)
- Dans ce cas, des liens forts peuvent exister entre les différentes variables mesurées, qui peuvent être prises en compte pour le traitement
- Dans le cours, sauf cas particulier, on supposera souvent que l'on traite des signaux univariés

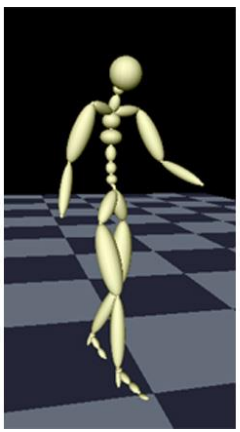
Séries univariées et multivariées

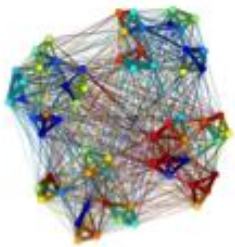
Trajectoires 2D/3D,
Séries temporelles multivariées,
Données multimodales sur des réseaux de capteurs...

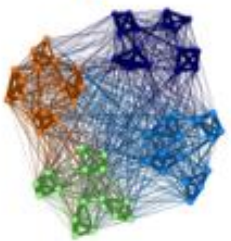


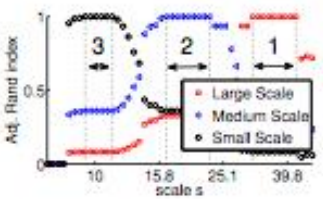










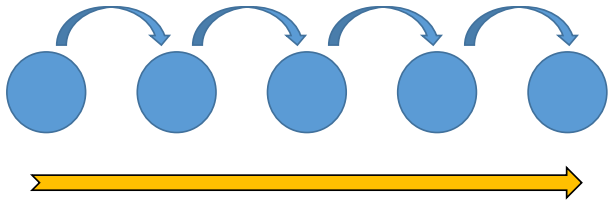


Les séries temporelles sont complexes

- Données potentiellement massives (ex : son qualité CD : 44100 échantillons par seconde)
- Données multivariées, multimodales, hétérogènes
- Données bruitées, données manquantes
- Souvent liées à un contexte applicatif : le *data scientist* n'est pas formé pour comprendre les données
- Manque d'annotations et de vérités terrain

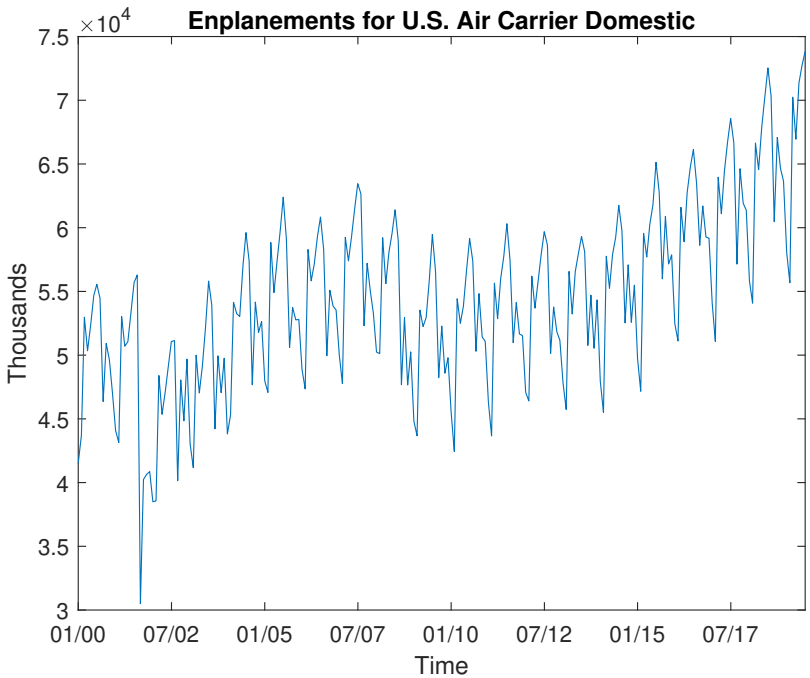
Importance du temps

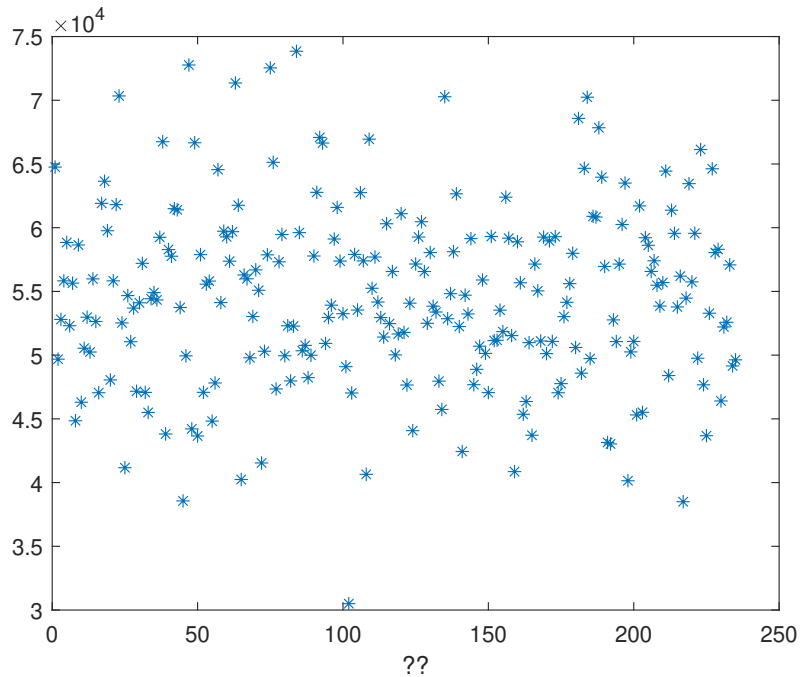
- Quelle est la différence entre des données classiques et des données temporelles ? Notion de séquentialité et de chronologie



- Chaque échantillon correspond à la mesure d'un phénomène à un instant donné
- La prise en compte du temps est cruciale pour la compréhension de l'évolution du phénomène

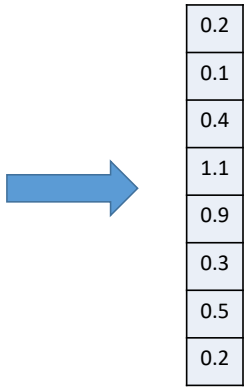
Importance du temps

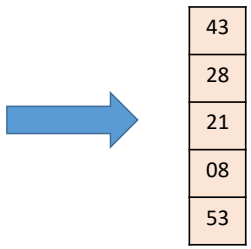




La même série temporelle... mais où les temps ont été mélangés

Temps et ML





Male
Young
British accent
« Hello »

New user
Likely to buy a car

- La plupart des algorithmes de ML ne prennent pas en compte le temps de façon native
- Comment utiliser cette information temporelle pour extraire des caractéristiques et des motifs qui peuvent être utilisés au sein d'un algorithme de ML ?

Deux visions : physique et statistique

- Il existe plusieurs façons de voir une série temporelle $x[n]$
  - **Vision physique.** La série temporelle correspond à une version discrétisée d'un phénomène physique $x(t)$: on cherchera alors à retrouver les propriétés de $x(t)$ à partir de $x[n]$. Il s'agit d'une approche **déterministe** correspondant classiquement à ce qui se fait en **traitement du signal**.
  - **Vision statistique.** La série temporelle correspond à une réalisation d'un phénomène aléatoire $X[n]$: on cherchera alors à retrouver les propriétés statistiques de $X[n]$ à partir de $x[n]$. Il s'agit d'une approche **probabiliste** correspondant à ce qui se fait en **traitement statistique du signal**.
- Bien entendu ces deux visions sont liées, et afin de bien explorer les données il est utile d'adopter les deux points de vue

Annotations et vérité terrain

- Contrairement aux images (ex : classification de chats et chiens), l'annotation de séries temporelles demande souvent une expertise, qui est dure à capturer
- Contexte typique :
  - Données bruitées et non consolidées
  - Quelques signaux annotés avec des étiquettes floues (annotations hyper-spécialisées qui ne se transforment pas aisément en labels de classes)
  - Un expert-métier avec des dizaines d'années d'expertise mais incapable de traduire ceci dans des annotations compatibles avec le ML

Comment utiliser le ML dans ce contexte ?

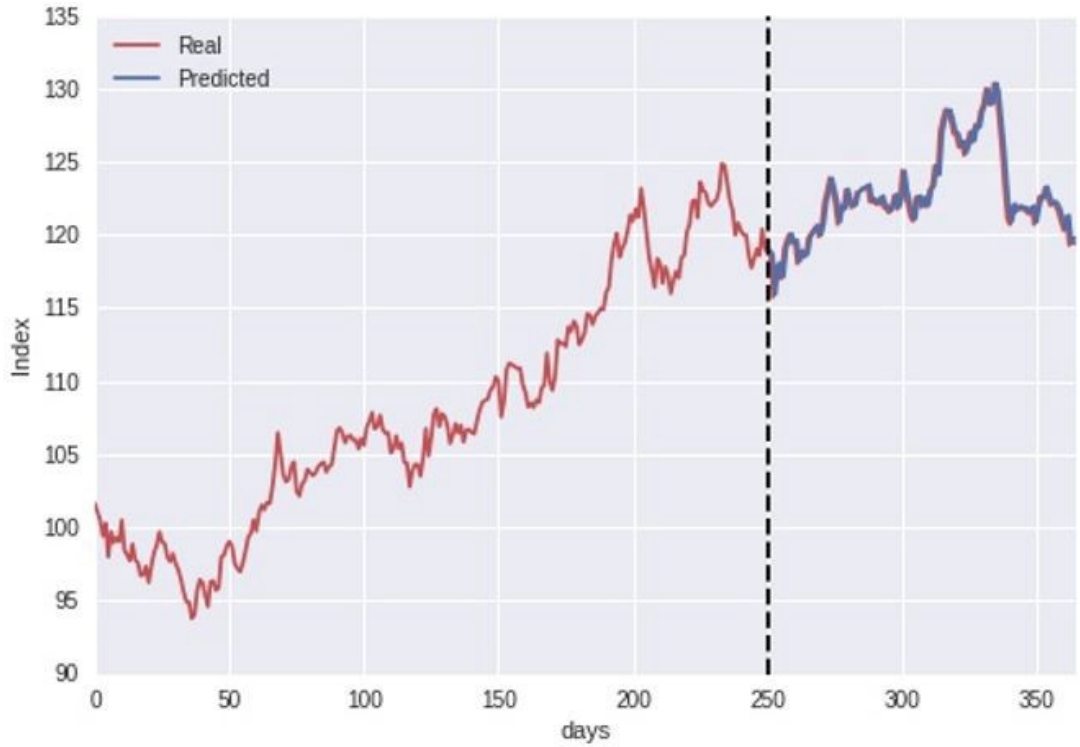
Deep learning : la solution ultime ?

Le deep learning permet effectivement d'atteindre l'état de l'art en termes de performances sur plusieurs tâches, **MAIS**...

- De bonnes performances n'impliquent pas une bonne compréhension des données (voir slide suivant)
- Le DL est une boîte noire qui n'offre parfois qu'une satisfaction partielle aux experts métier qui ne peuvent pas interpréter les résultats
- Bien que certains réseaux prennent en compte le temps (ex LSTM), il reste difficile d'utiliser des informations distantes temporellement
- Le DL est assez mauvais lorsqu'il y a peu de données/annotations

Un joli contre-exemple (1/4)

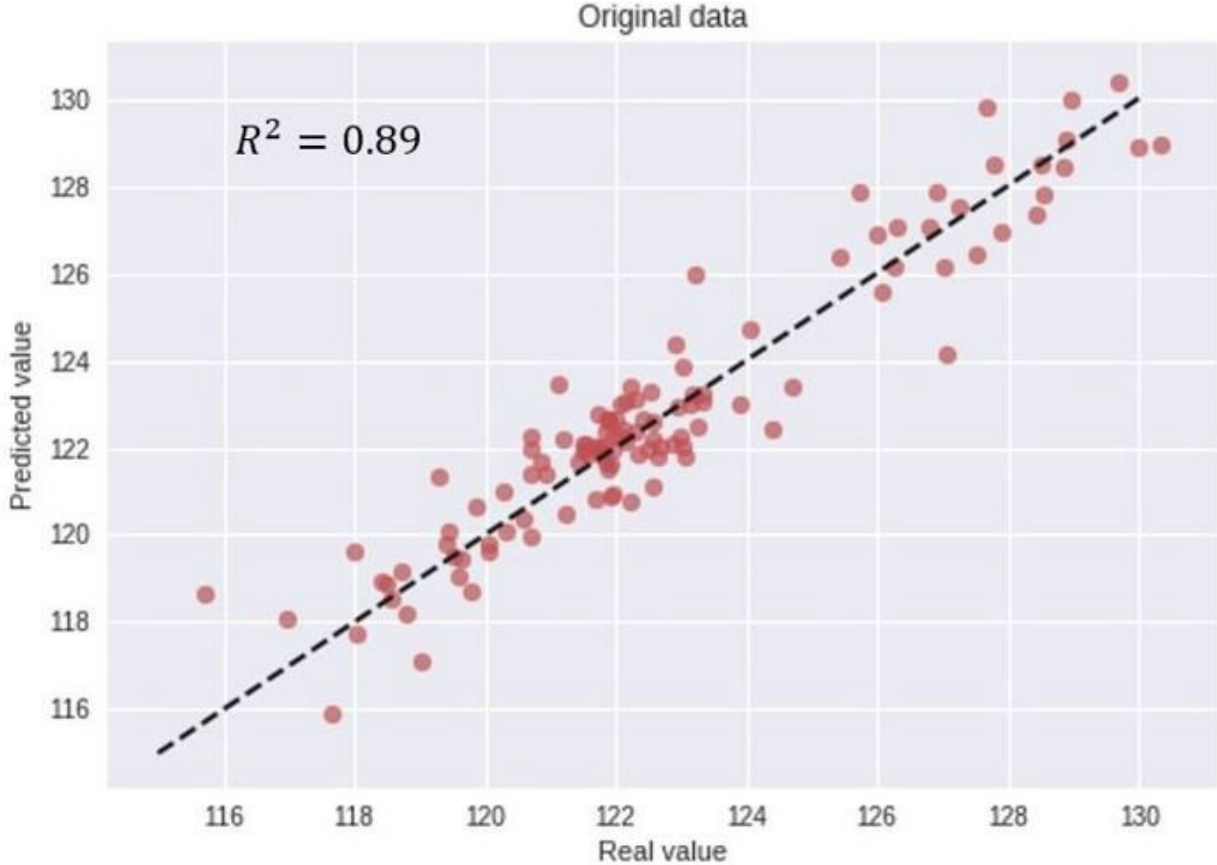
Blog post from V. Flovik
`https://towardsdatascience.com/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424`



Prédiction de la valeur d'une action grâce à un réseau LSTM : utilisation des 250 premiers jours comme données d'entraînement. La prédiction semble excellente !

Un joli contre-exemple (2/4)

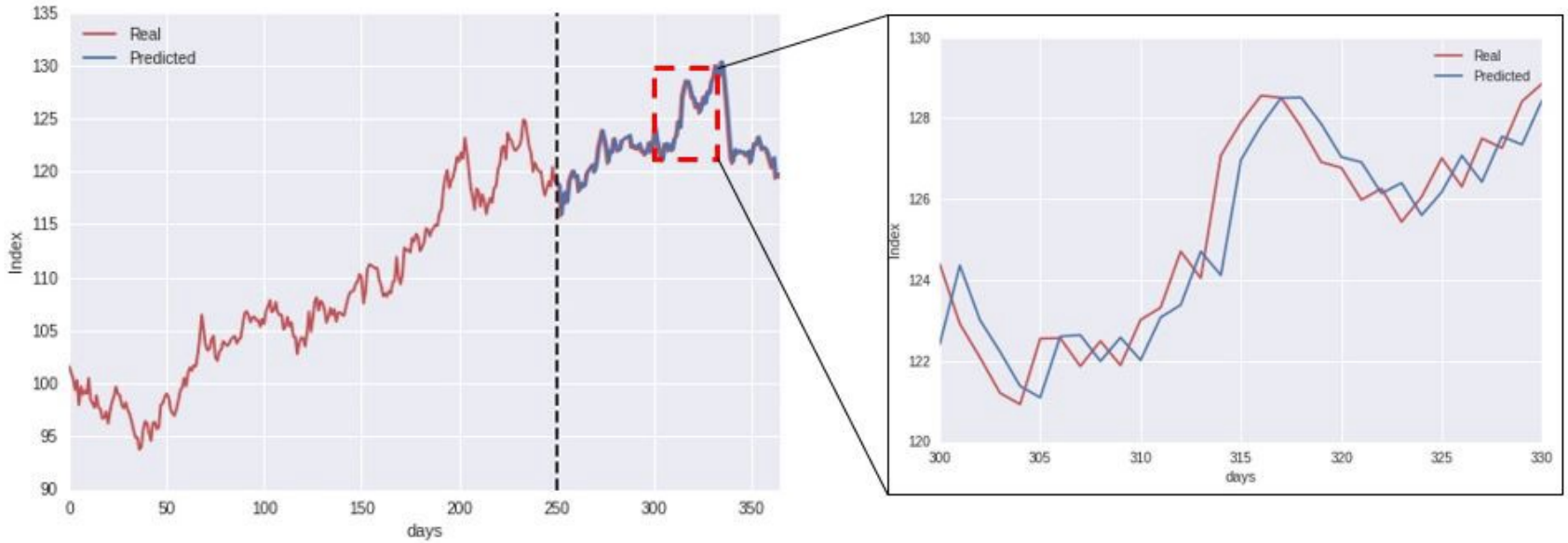
Blog post from V. Flovik
`https://towardsdatascience.com/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424`



Excellentes performances également ! Prédiction et RMSE de bonne qualité !

Un joli contre-exemple (3/4)

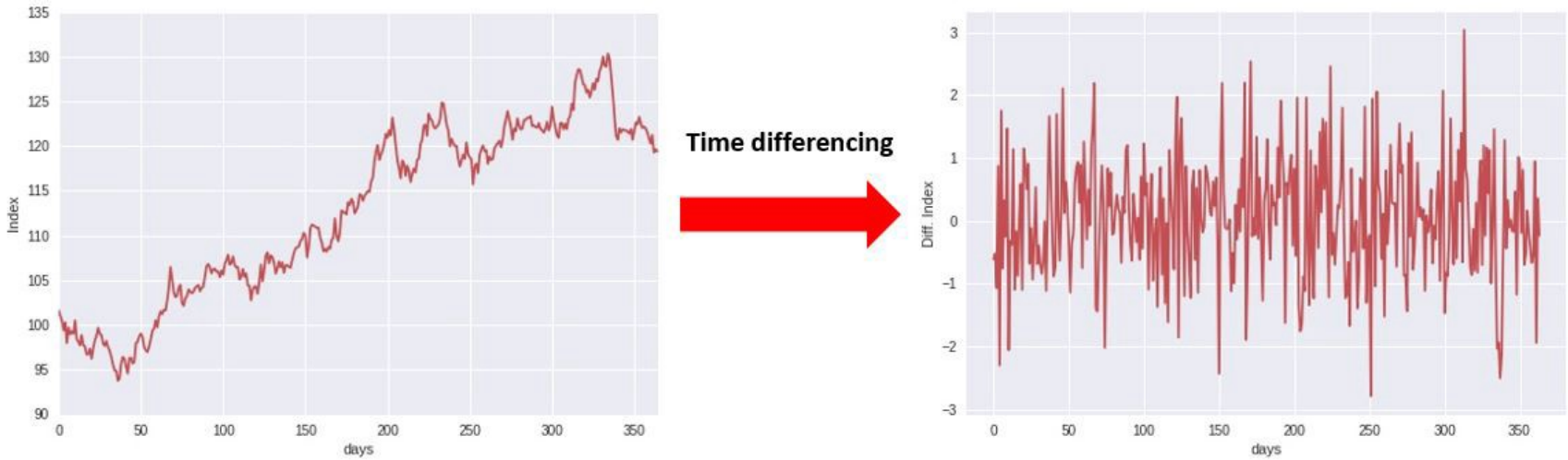
Blog post from V. Flovik
https://towardsdatascience.com/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424



En réalité, LSTM ne faisait que recopier l'échantillon précédent...

Un joli contre-exemple (4/4)

Blog post from V. Flovik
`https://towardsdatascience.com/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424`



En réalité, il s'agissait d'une marche aléatoire : il était impossible de prédire quoi que ce soit. Ceci aurait pu être détecté immédiatement grâce à une exploration consciencieuse...

Plan du cours

1. Présentation du cours

2. Signaux et séries temporelles

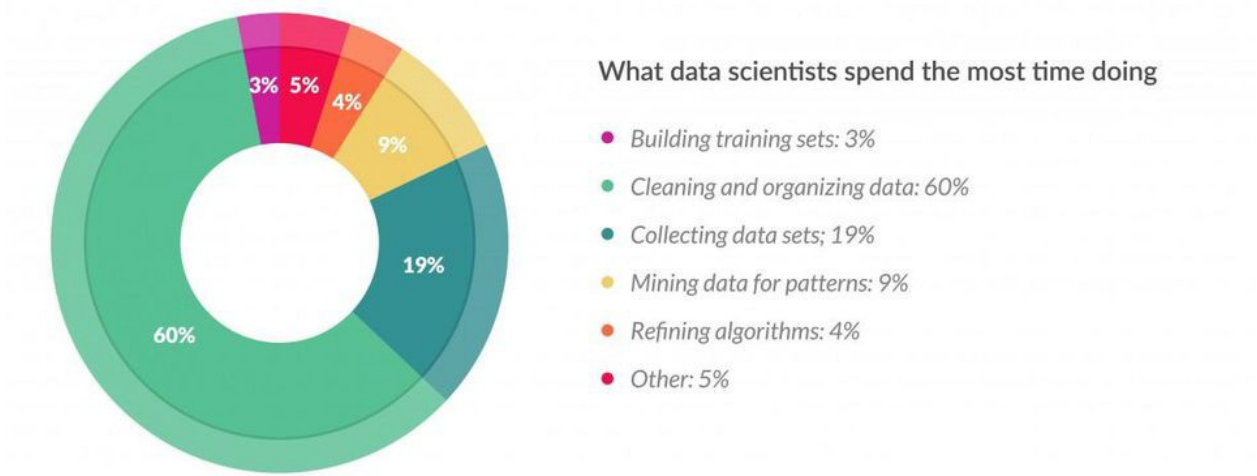
3. Machine Learning pour les séries temporelles

4. Plan du cours

Qu'est-ce que le Machine Learning ?

- Le Machine Learning ne consiste pas (uniquement) à obtenir les meilleures performances possibles en lançant des packages de Machine Learning en Python
- Le Machine Learning a également pour but de comprendre les données, d'interagir avec des experts métier, et d'améliorer la connaissance
- Il n'y a pas d'intelligence artificielle sans intelligence humaine
- Appliquer des méthodes perfectionnées de Machine Learning ne dispense pas d'une phase d'exploration préliminaire... **et le Machine Learning peut également aider à cela !**

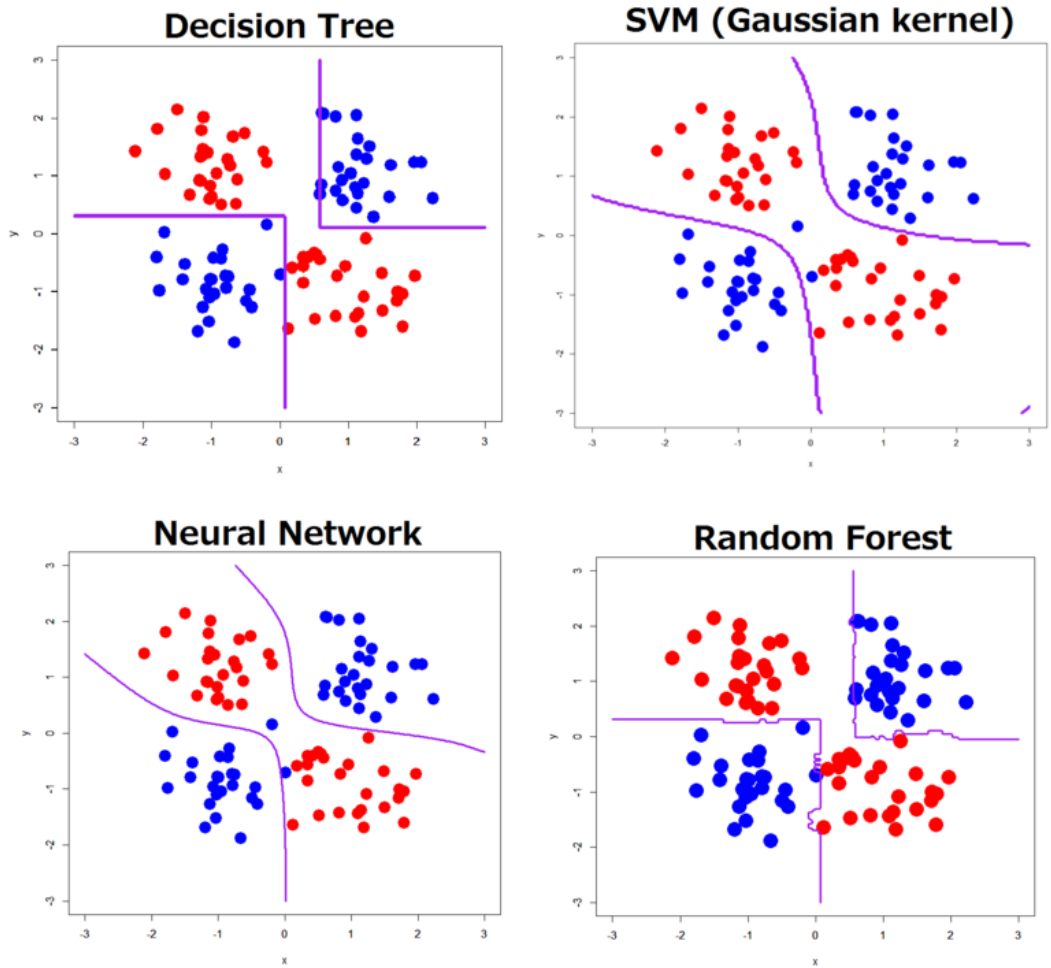
Comprendre les données : une tâche complexe et chronophage



Source : https ://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

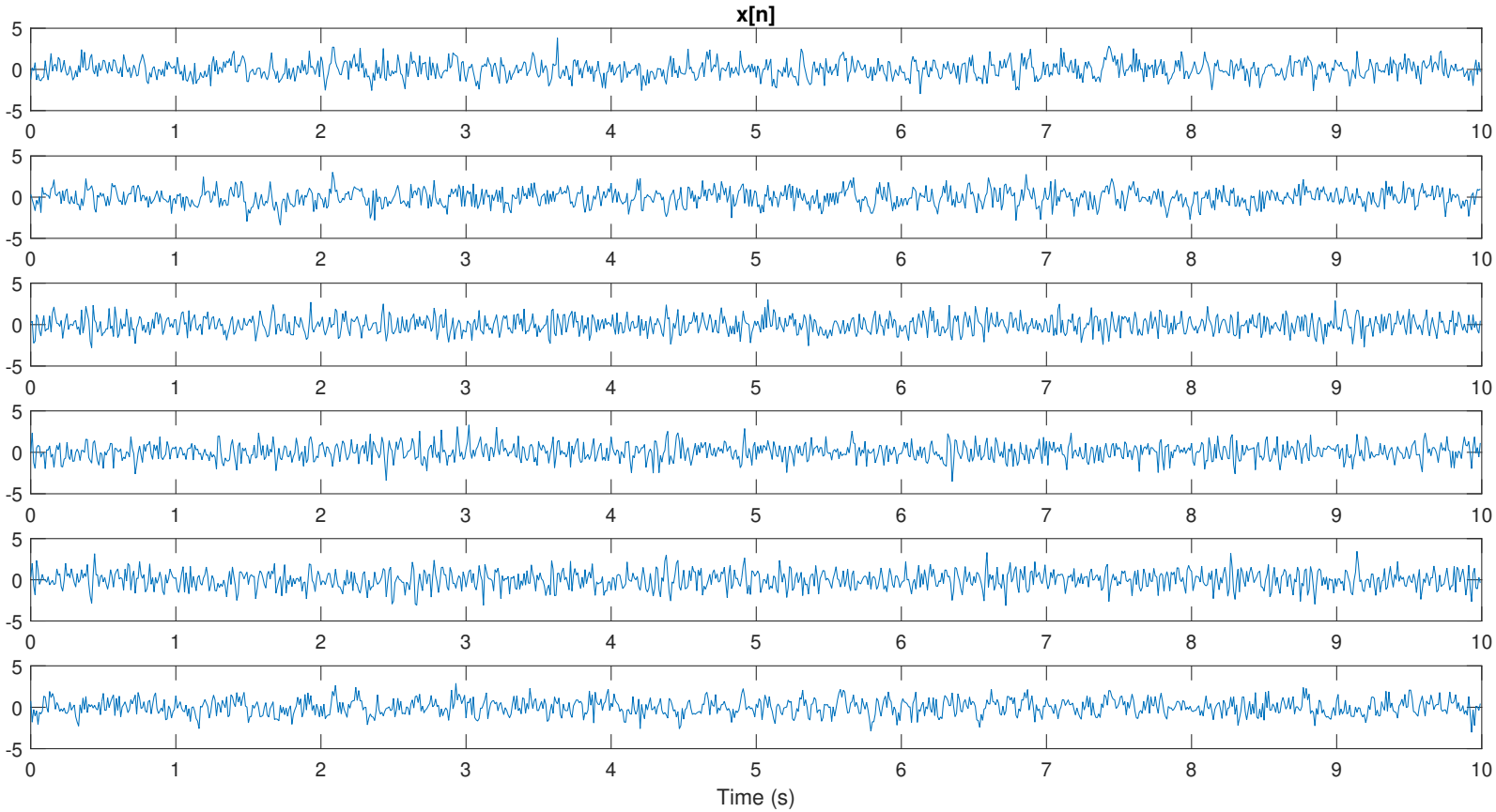
- Comprendre les données pour en extraire l'information nécessaire
- Comprendre ce que l'on fait, pourquoi on le fait, comment on le fait : interprétabilité

Représentation et complexité



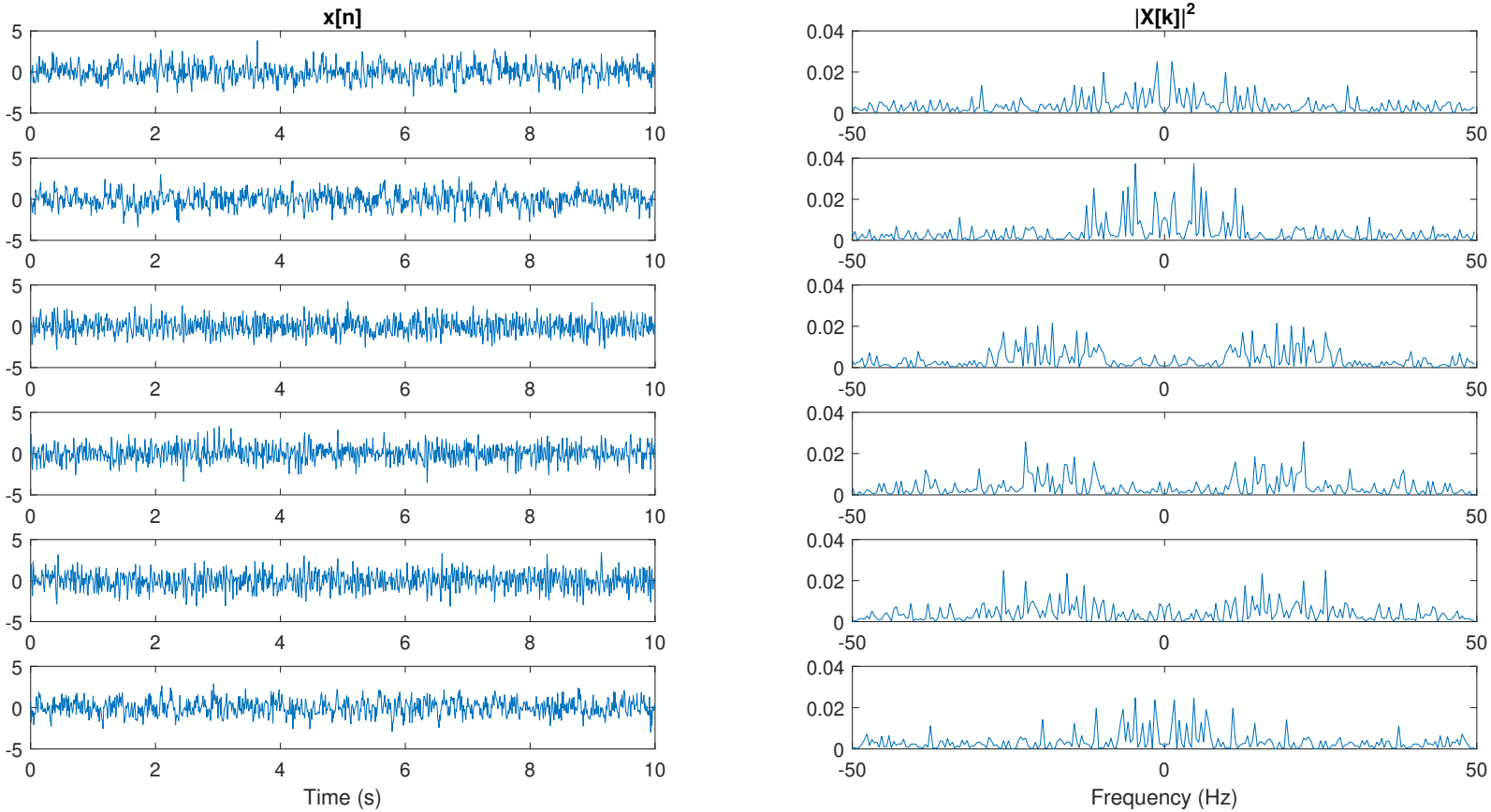
La complexité des données peut être prise en compte par des algorithmes de ML toujours plus sophistiqués : est-ce toujours la meilleure solution ?

Importance de la représentation



Deux classes de signaux ?

Importance de la représentation



Tâche triviale si on travaille dans le bon domaine de représentation

Principales tâches de ML pour les séries temporelles

- **Prédiction :** Prédire les valeurs futures d'une série temporelle
- **Complétion/interpolation :** Retrouver des échantillons manquants ou perdus dans une série temporelle
- **Classification :** Attribuer une étiquette de classe à une série temporelle ou à une sous-série
- **Régression extrinsèque :** Prédire une variable externe continue à partir d'une série temporelle (contrairement à la classification, la cible est numérique et non une étiquette de classe)
- **Partitionnement :** Former des groupes de séries temporelles ayant les mêmes propriétés
- **Recherche par contenu/indexation :** Étant donnée une série temporelle source, retrouver la série temporelle la plus proche dans une grande base de données, au sens d'une distance donnée
- **Prototypage :** Trouver une ou plusieurs séquences représentatives résumant un ensemble de séries temporelles similaires
- **Segmentation/détection de ruptures :** Trouver des changements abrupts significatifs dans une série temporelle
- **Détection d'anomalies :** Trouver des événements anormaux dans une série temporelle
- **Extraction de motifs :** Trouver des événements répétitifs dans une série temporelle

Tâches *cachées*

- **Comprendre les données :** savoir d'où elles viennent, comment elles ont été acquises, quelles sont leurs caractéristiques
- **Modéliser les données :** modèles physiques/statistiques ou basés sur de la connaissance métier
- **Améliorer les données :** trouver les bons espaces de représentation où les phénomènes d'intérêt sont visibles, consolider les données (débruitage, suppression de tendance, détection et suppression de valeurs aberrantes...)
- **Extraire de l'information à partir des données :** trouver des motifs répétitifs et des caractéristiques pertinentes

Plan du cours

1. Présentation du cours

2. Signaux et séries temporelles

3. Machine Learning pour les séries temporelles

4. Plan du cours

But du cours

Intelligence Artificielle & Machine Learning pour la modélisation (et la compréhension) de séries temporelles et de signaux

- Focus sur les *tâches cachées* : comprendre, modéliser, améliorer et extraire
- Algorithmes de ML interprétables et reproductibles : boîtes blanches
- Méthodes permettant de créer des ponts entre les *data scientists* et les experts métier
- Toutes ces méthodes permettent ensuite de résoudre facilement toutes les tâches classiques en ML : prédiction, classification, partitionnement...

Plan du cours

- Séance 1 : Quelques notions de traitement du signal (en vidéo)
- Séance 2 : Quelques notions de traitement statistique du signal (en vidéo)
- Séance 3 : Modèles pour les séries temporelles
- Séance 4 : Apprentissage de représentation et de dictionnaire
- Séance 5 : Pré-traitements des séries temporelles
- Séance 6 : Extraction et sélection de caractéristiques
- Séance 7 : Détection et extraction de motifs
- Séance 8 : Détection de ruptures et d'anomalies