Fonds européen de développement régional (FEDER)
Europäischer Fonds für regionale Entwicklung (EFRE)

interreg Oberrhein Rhin Supérieur

Quels jeux de données pour la prédiction d'anomalies dans l'industrie 4.0 ?

Mouhamadou Saliou Diallo, Sid Ahmed Mokeddem, Agnès Braud, Gabriel Frey, Nicolas Lachiche

{ms.diallo, mokeddem, agnes.braud, g.frey, nicolas.lachiche}@unistra.fr

Gestion et Analyse des données Spatiales et Temporelles (GAST-2020)

HOCHSCHULE FURTWANGEN UNIVERSITY HFU

ENGEES
ÉCOLE NATIONALE DU GÉNIE DE L'EAU ET DE L'ENVIRONNEMENT DE STRASBOURG

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES STRASBOURG

cnrs dépasser les frontières

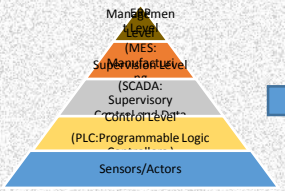
UNIVERSITÉ DE STRASBOURG

HALFBACK

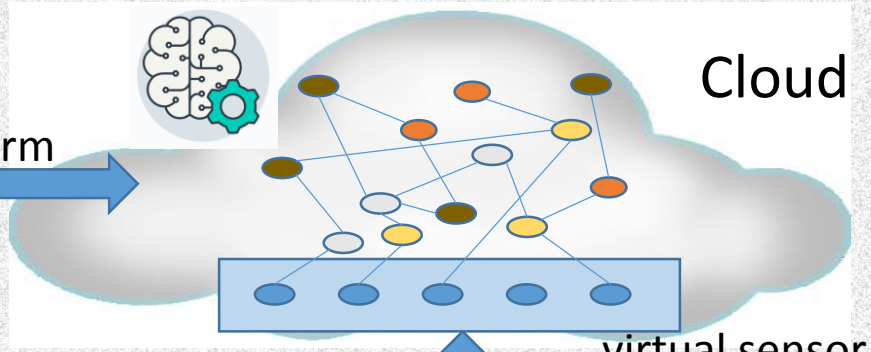
Industrie 4.0 Infrastructure

Modernisation des usines:

- en ce basant sur la science
- et la technologie dites intelligentes

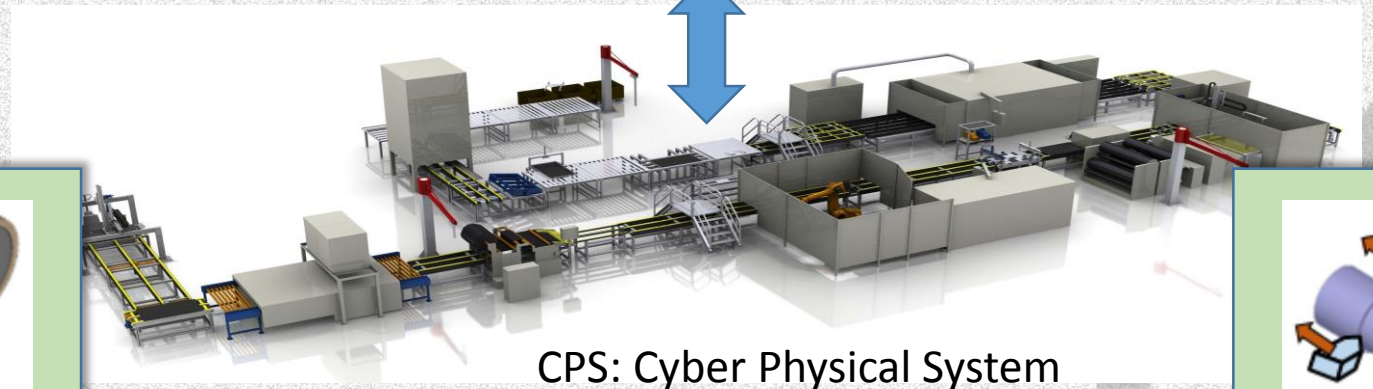


transform

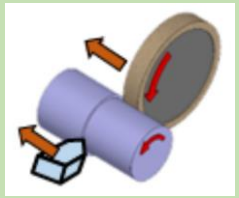


virtual sensor/actors

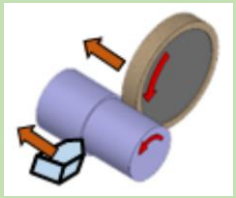
Cloud Gateway (Edge Gateway, IoT Gateway)



CPS: Cyber Physical System



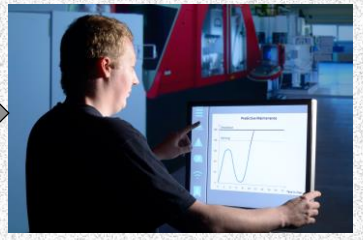
Machine A

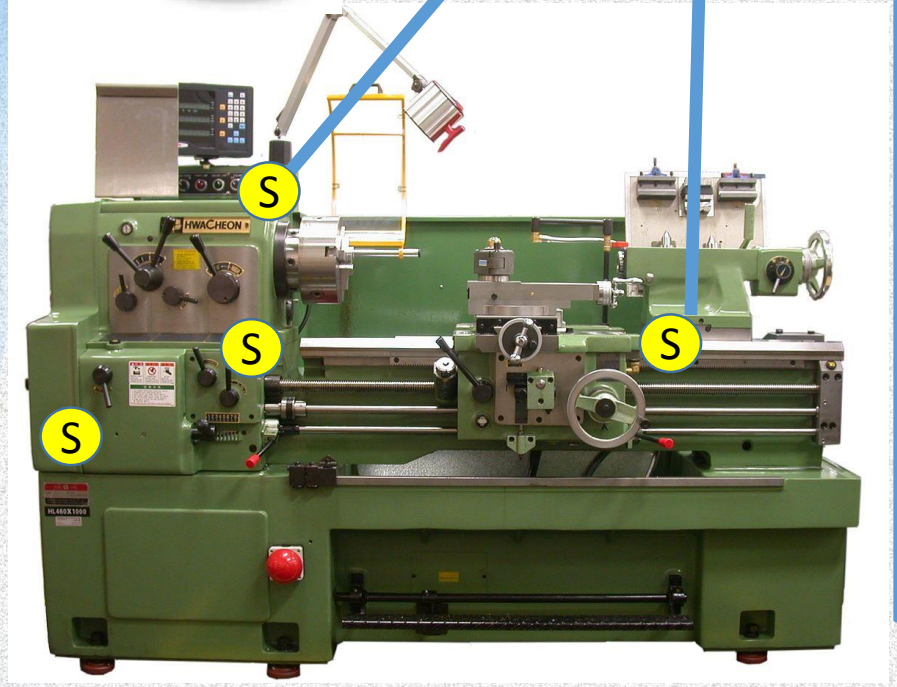


Machine n

HALFBACK

Collecte de données → Préparation des données → Analyse de données





- Les données sont collectées en utilisant les capteurs sur les machines et outils
- Des informations supplementaires seront collectées à partir de l'environnement de production;
- Du produit lui-meme ainsi que de l'experience de l'opérateur de la machine.

- Des algorithmes de Big Data pour:
  - Apprendre le processus
  - Apprendre de l'expérience des operateurs

Ceci permettra à l'entreprise d'agir avant l'arret du processus de production et d'optimiser la plannification en reorganisant la ligne de production

HALFBACK

Approches pour la gestion de maintenances (Susto et al., 2015)

Approches	Caractéristiques
Maintenance réactive	Simple, Peu efficace: ➢ Temps d'intervention important, ➢ Temps d'arrêt des équipements important
Maintenance préventive	Défaillances sont généralement évitées, Augmentation des coûts d'exploitation: ➢ Interventions inutiles sont souvent effectuées, ➢ Utilisation inefficace des ressources.
Maintenance prédictive	➢ Basée sur des données historisées, ➢ Maintenance effectuée sur la base d'une estimation de l'état d'un équipement. ➢ Détection à l'avance des anomalies en attentes ➢ intervention à temps utile avant une défaillance. ➢ Utilise des outils de prédiction basées sur des données historisées.

HALFBACK

Objectif principal

- Caractéristiques des jeux de données pour la prédiction d'anomalies dans l'industrie 4.0
- Donner des exemples de jeux de données adaptées
- Donner des exemples de jeux de données inadaptés.
- Montrer un exemple de mise en oeuvre
- Proposer plusieurs perspectives de recherche.

HALFBACK

Défis de la prédiction d'anomalies

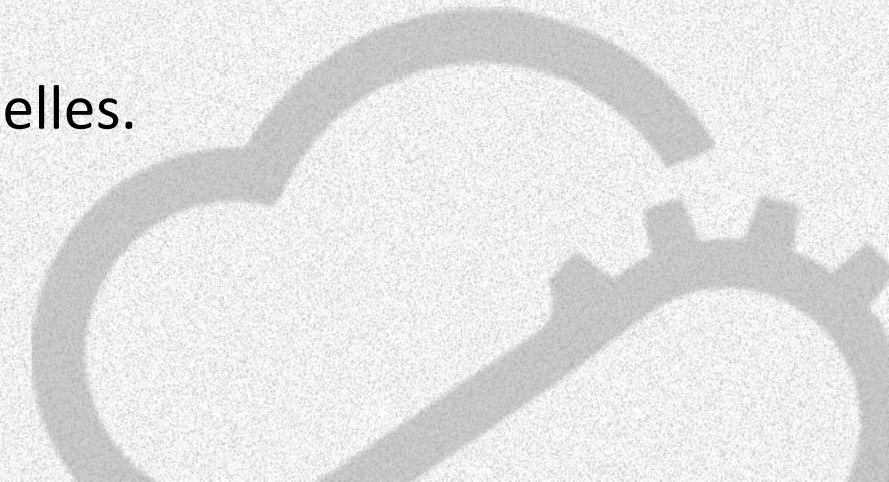
- **Les masses de données:**
  - Les capteurs génèrent automatiquement une quantité de données qui atteint rapidement l'ordre du Go.
- **Le déséquilibre des données:**
  - Les cas d'anomalies sont beaucoup plus rares-heureusement!- que les cas normaux.
- **La diversité des données :**
  - On doit souvent apprendre avec peu d'exemplaires
  - d'une même machine,
  - voire d'une même famille (mais de puissances différentes, par exemple).

HALFBACK

Caractéristiques nécessaires

- Nous ne faisons pas de différence entre
  - Flux de données,
  - Données temporelles
  - et données séquentielles

- Les données ont une estampille temporelle
  - qui permet de les ordonner en données séquentielles.



HALFBACK

Caractéristiques nécessaires

- La collecte des données s’arrête lorsqu’une anomalie survient
- On peut supposer que la collecte des données reprendra lorsque la panne/anomalie aura été corrigée,
- et constituera une nouvelle séquence.
- Ainsi chaque séquence d’apprentissage se termine par une panne.

HALFBACK

Caractéristiques nécessaires

- **<u>Apprentissage Artificiel:</u>**
  - Nous nous plaçons dans le cadre d'un apprentissage supervisé,
    - une étiquette est associé à chaque individu.
  - Nous considérons que chaque instant correspond à un individu.
  - **Problème:**
    - Apprendre un modèle qui prédit
    - La valeur de cette étiquette
    - Pour un nouvel individu,
    - Qui est un autre instant dans notre cas.

HALFBACK

Caractéristiques nécessaires

- **<u>Apprentissage Artificiel:</u>**
- Etiquette qualitative /Catégorielle :
  - Technique de classification (supervisée)
  - Définir étiquette comme un booléen
  - Indiquant si la panne va se produire dans moins de temps qu'un seuil alpha
  - Alpha est le temps nécessaire pour planifier la maintenance et réorganisation la production

HALFBACK

Caractéristiques nécessaires

- **Apprentissage Artificiel:**
- Etiquette quantitative/numérique:
  - Technique de régression
  - Définir étiquette comme le temps le restant avant la panne
  - Remaining Useful Life (RUL) en anglais.

HALFBACK

Caractéristiques nécessaires

En conclusion:

- Nous recommandons qu'un jeu de données soit constitué de séquences :
  - Toutes comparables entre elles:
    - c'est-à dire ayant une même représentation
    - incluant une description de leurs points communs et de leurs différences,
    - et permettant d'apprendre un modèle à partir de n'importe quel sous-ensemble de ces séquences
    - et de pouvoir l'appliquer aux séquences restantes ;
  - Où chaque séquence se termine par une anomalie,
    - Permettant de déterminer la valeur de l'étiquette
    - Pour chacun des instants de la séquence.

Contre Exemple

Secom:

- Concerne un processus de fabrication de semi-conducteurs
- 591 attributs, 1567 exemples.
- 104 exemples se terminent par une panne.
- On ne peut pas utiliser les 1463 autres "exemples" dans une approche supervisée
- De plus ces séquences sont courtes:
  - 18 estampilles temporelles en moyenne
  - voire 3 pour la plus courte

Contre Exemple

Li-ion Battery Aging Datasets :

- Données sur le vieillissement de batteries lithium-ion
- Chaque séquence est composée de cycles de charge et décharge.
- 5 à 7 attributs sont mesurées en fonction de l’état courant du cycle.
- Mais seulement 4 batteries sont testées.
- Cela fournit trop peu d’exemples sur lesquels apprendre et tester.

Exemples

Backblaze:(septembre 2019)

- Statistiques sur des disques de 4 fabricants et sont de plusieurs modèles
- Des données sur 112 864 disques durs, 6 078sont tomb és en panne.
- On ne peut pas utiliser les 106 786 autres disques qui n'ont pas encore failli.
- Model ST4000DM000 de Seagate:
  - Il y a 3724 exemplaires en panne
  - Cela fournit un jeu de données homogène,
  - où les disques sont décrits par les mêmes attributs.
  - Souvent utilisé dans la littérature

Backblaze Lifetime Hard Drive Annualized Failure Rates

For hard drive models in service as of September 30, 2019
Reporting period April 2013 - September 2019 inclusive

MFG	Model	Drive Size	Drive Count	Avg. Age	Drive Days	Drive Failures	AFR*
HGST	HMS5C4040ALE640	4TB	2,707	42.0	11,420,392	161	0.51%
HGST	HMS5C4040BLE640	4TB	12,641	35.6	18,409,871	233	0.46%
HGST	HUH728080ALE600	8TB	1,001	22.3	746,311	16	0.78%
HGST	HUH721212ALE600	12TB	1,560	4.8	183,560	4	0.80%
HGST	HUH721212ALN604	12TB	10,849	6.1	1,923,518	25	0.47%
Seagate	ST4000DM000	4TB	19,330	47.3	50,839,992	3,724	2.67%
Seagate	ST6000DX000	6TB	886	53.9	2,821,207	83	1.07%
Seagate	ST8000DM002	8TB	9,839	36.3	10,910,157	316	1.06%
Seagate	ST8000NM0055	8TB	14,416	26.8	11,856,443	386	1.19%
Seagate	ST10000NM0086	10TB	1,200	24.3	897,426	14	0.57%
Seagate	ST12000NM0007	12TB	37,116	15.4	17,458,380	1,102	2.30%
Toshiba	MD04ABA400V	4TB	99	52.3	225,739	5	0.81%
Toshiba	MG07ACA14TA	14TB	1,220	11.9	441,195	9	0.74%
		Totals	112,864		128,134,191	6,078	1.73%

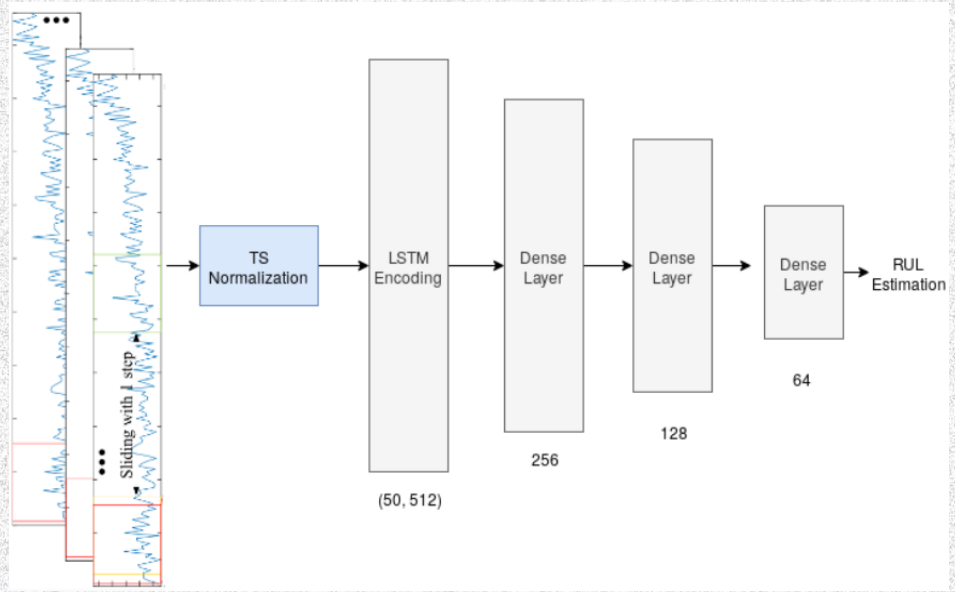
* AFR - Annualized Failure Rate

BACKBLAZE

HALFBACK

Exemple: Mise en Œuvre avec Turbofan

- **100** séquences se terminant par une panne.
- Ces séquences ont des longueurs comprises entre **128** et **362**, avec une moyenne de **206** instants.
- On dispose des valeurs relevées par **24** capteurs à chaque instant. Chaque instant a été étiqueté par la durée de vie restante avant la panne.
- **70** séquences ont été utilisées pour construire un modèle à l'aide de LSTM.
- Les **30** séquences restantes servent de jeu de test.



HALFBACK

Les défis du jeu de données proposé par Bosch

Contexte:

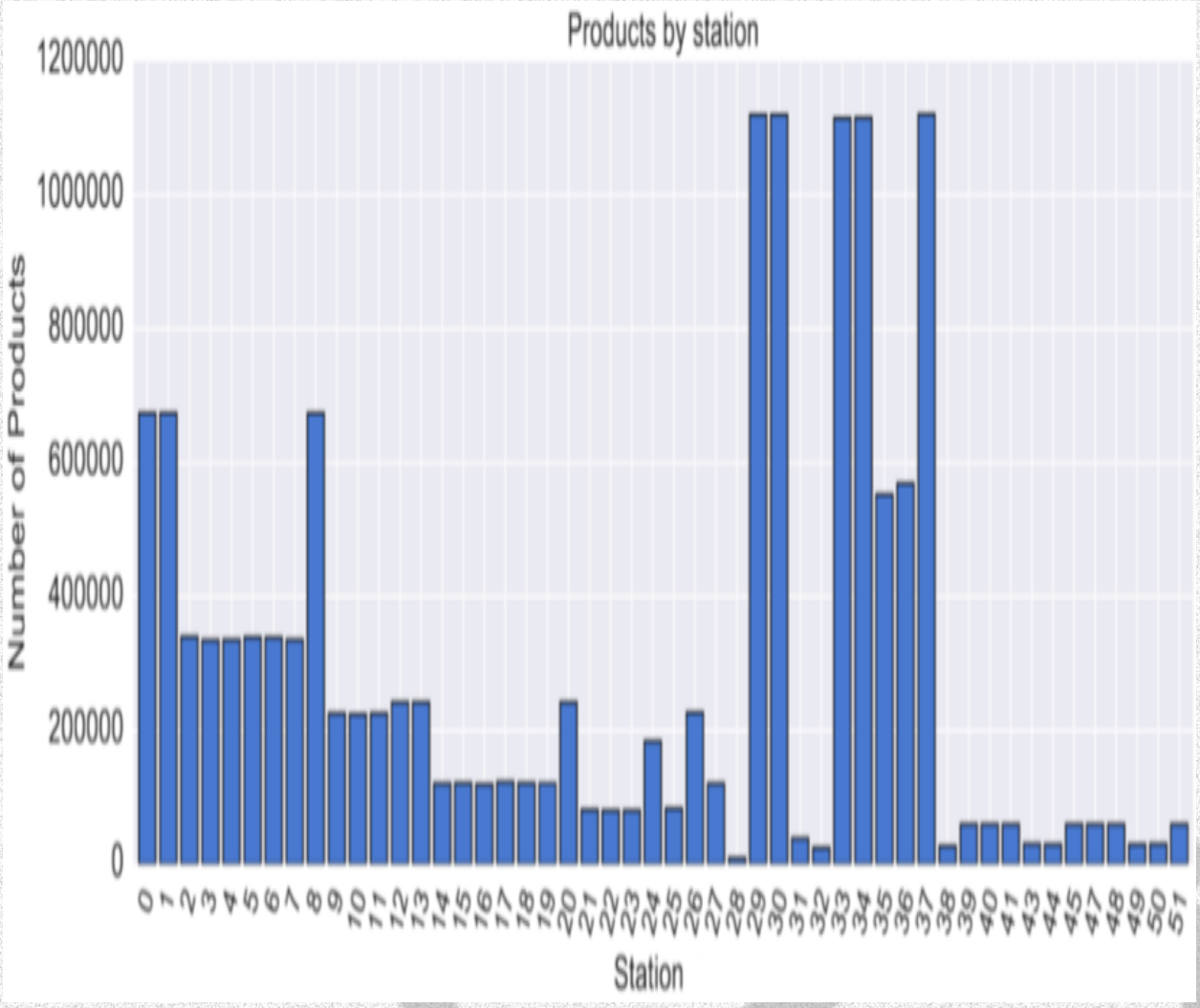
- Prédiction de la qualité des pièces produites.
- Le jeu de données d’ entrainement est relatif à **1 184 687** produits,
- décrits par **968** valeurs numériques des différents capteurs
- et **1156** estampilles temporelles correspondant aux instants de passage par les différents capteurs.
- Les données ne sont pas plus détaillées par l’entreprise.
- On sait seulement qu’il y a **51** stations sur un total de **4** lignes de production.

HALFBACK

Les défis du jeu de données proposé par Bosch

Problématique:

- Le jeux de données proposé est de grande taille **(14,3 Go)**.
- Chaque opération sur une telle masse de données est difficile
- Il n’y a que **6 879** produits défectueux, soit **0,58%** des produits.
- Les données sont extrêmement déséquilibrées.
- De nombreuses valeurs sont manquantes.



HALFBACK

Conclusion et perspectives

- Pour apprendre à prédire des pannes:
  - Il faut des séquences en nombre suffisant et de durée suffisante.
  - Il est utile aussi que les séquences se terminent par une panne
    - afin de pouvoir étiqueter chacun des instants par rapport à la fin de la séquence.
  - On peut imaginer des séquences de test, voire d'entraînement, qui ne se terminent pas par une panne,
    - à condition que chaque instant soit étiqueté,
    - donc que l'on connaisse la fin de la séquence.

HALFBACK

Perspectives

En cours

- Quand faut il prédire la panne?
  - La période où l'on souhaite prédire la panne est évidemment celle où il faut planifier la maintenance.
  - Sa valeur exacte est indiquée par les experts,
    - en prenant en compte le temps qu'il faut pour organiser la maintenance.
- En classification supervisée, il faut distinguer les faux positifs des faux négatifs.
- En régression, on peut distinguer une surestimation d'une sous-estimation.
- En effet, il est plus grave de prédire et donc planifier une maintenance trop tard que plus tôt. Cet aspect a été étudié dans le cas général par exemple par (Hernández-Orallo, 2013) mais pas dans le cas particulier des séries temporelles.

HALFBACK

Wissenschaftsoffensive

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

Fonds européen de développement régional (FEDER)
Europäischer Fonds für regionale Entwicklung (EFRE)

Merci de votre Aimable Attention: