

# Application Of AI (Knowledge Graph Embeddings) Industrial Use Cases Formal Knowledge Integration in Machine Learning Model For Industry

**Industrial PhD Student**, **40** Mouloud IFERROUDJENE

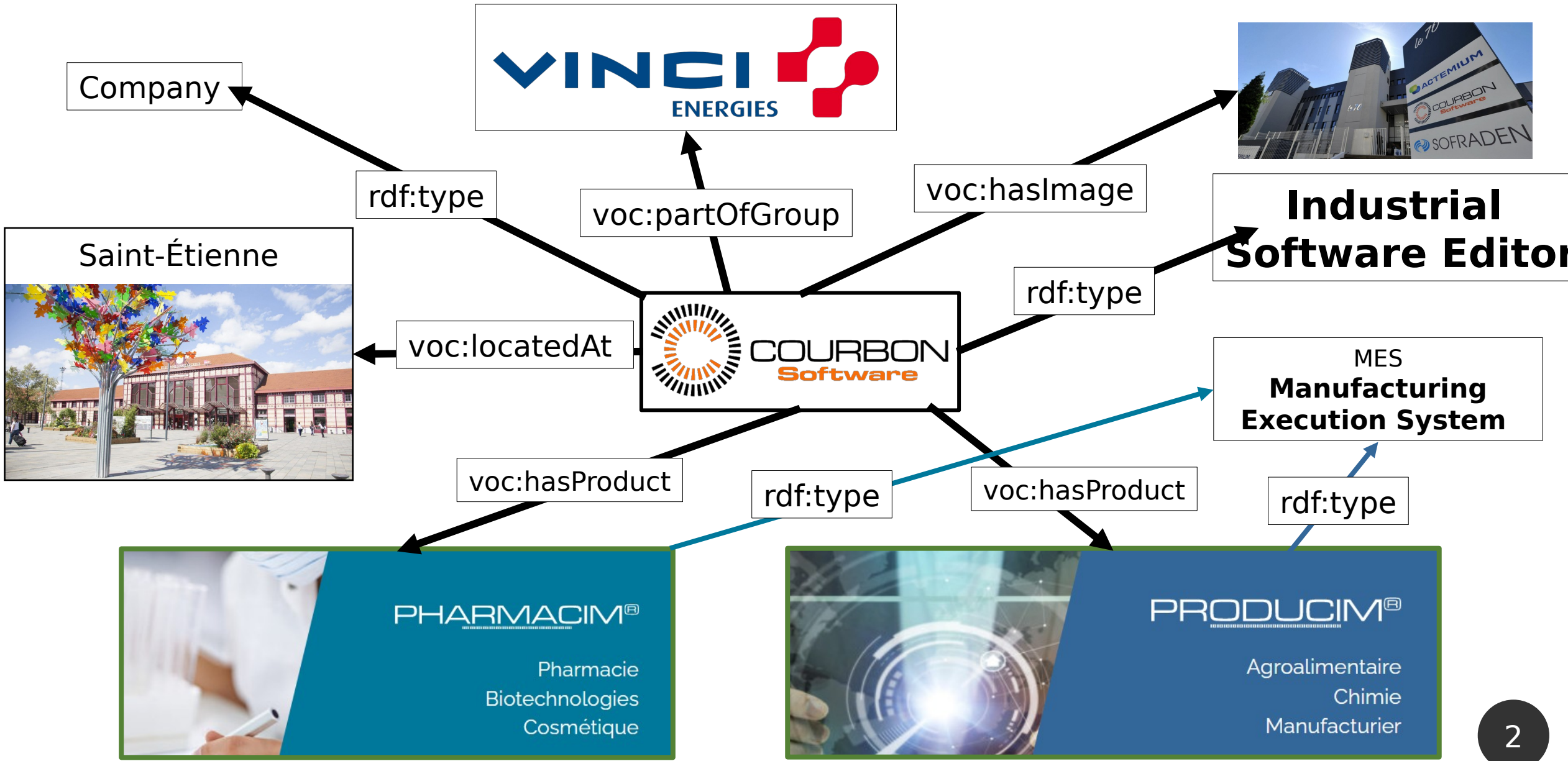
**Supervised by** Antoine ZIMMERMANN, Victor CHARPENAY, Thierry



LABORATOIRE D'INFORMATIQUE,  
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES



# COURBON Software (CSO)





# What is a MES ?

MES Stands for

- **Manufacturing execution system**

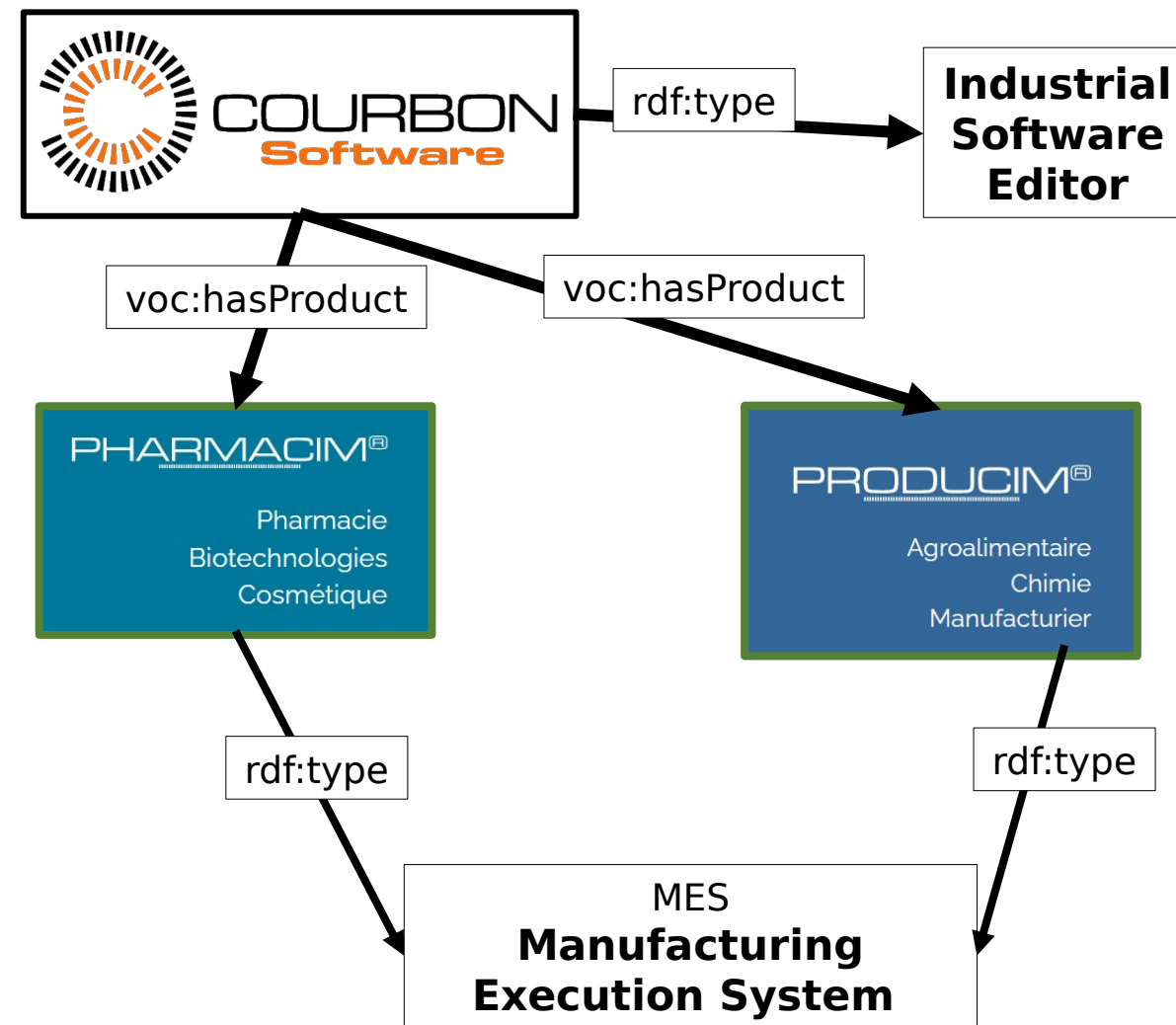
**MES objectives :**

- Ensure the proper execution of manufacturing operations
- Improve production efficiency

**MES Functionalities :**

- Product traceability
- Quality control
- Production monitoring
- Scheduling, Etc.

**Data acquisition => Lot of DATA from different sources**



**CONTEXT**

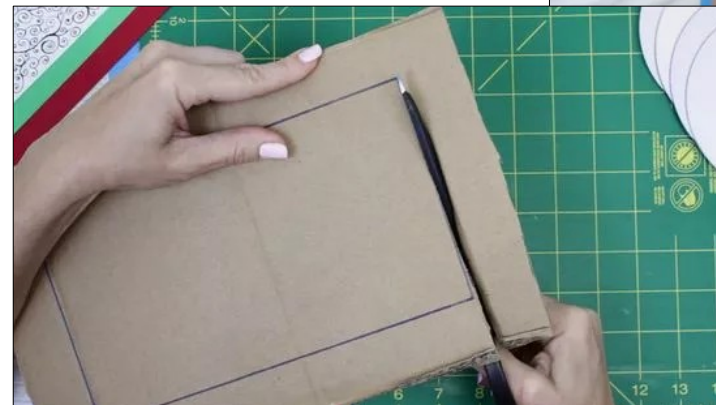


# Illustration Example ( toy car production )

Imagine you have a toy factory where you make toy cars !!

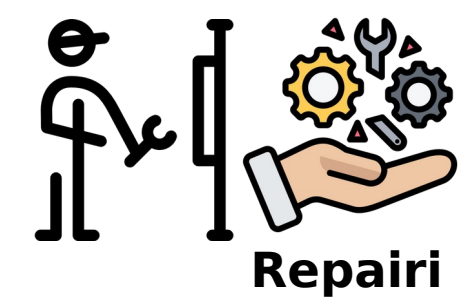
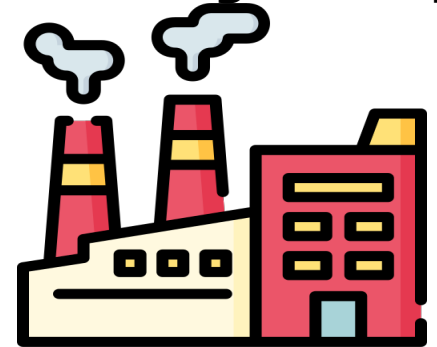
the past, you had to do everything by hand:

- Make the cars
- Paint them
- And package them



# Illustration Example ( car toy production )

g factories (e.g., cars production)

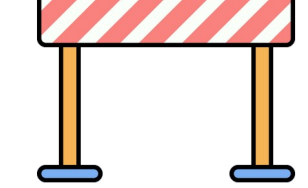


Repairing



The factory stop

production

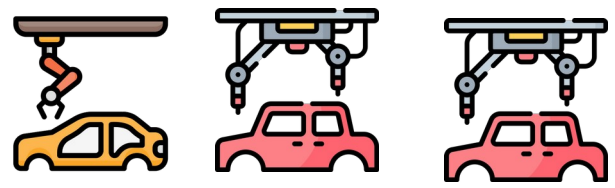


Lose

Money

assembly line and mass production

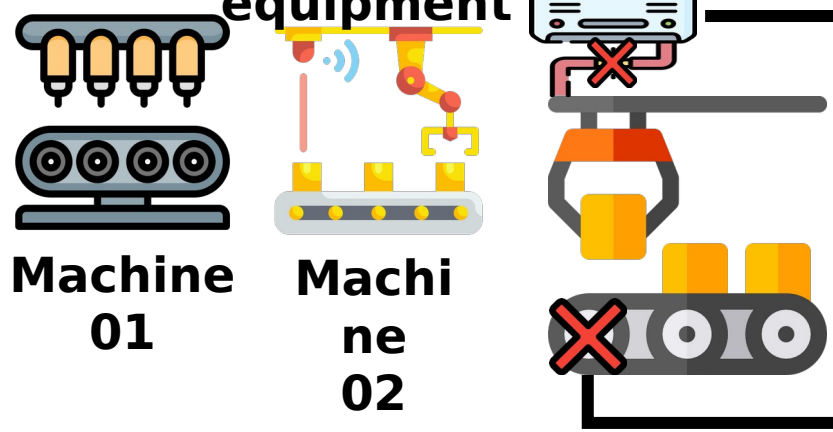
Industry 2.0



Sometimes !!  
machine  
break down

Oops!  
Oops!  
Oops!

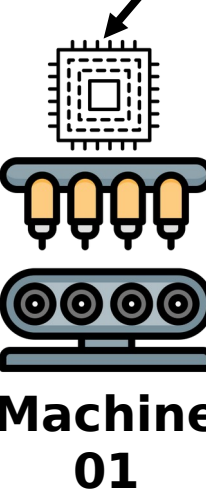
Have a lot of machines and  
equipment



What if we install chips  
to monitor our  
machines'  
status ?



Industry 3.0

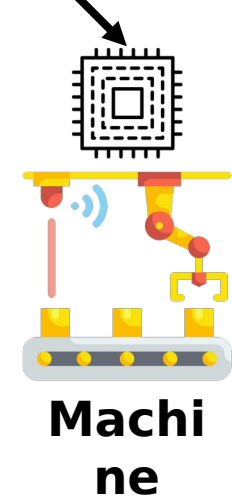


Machine  
01



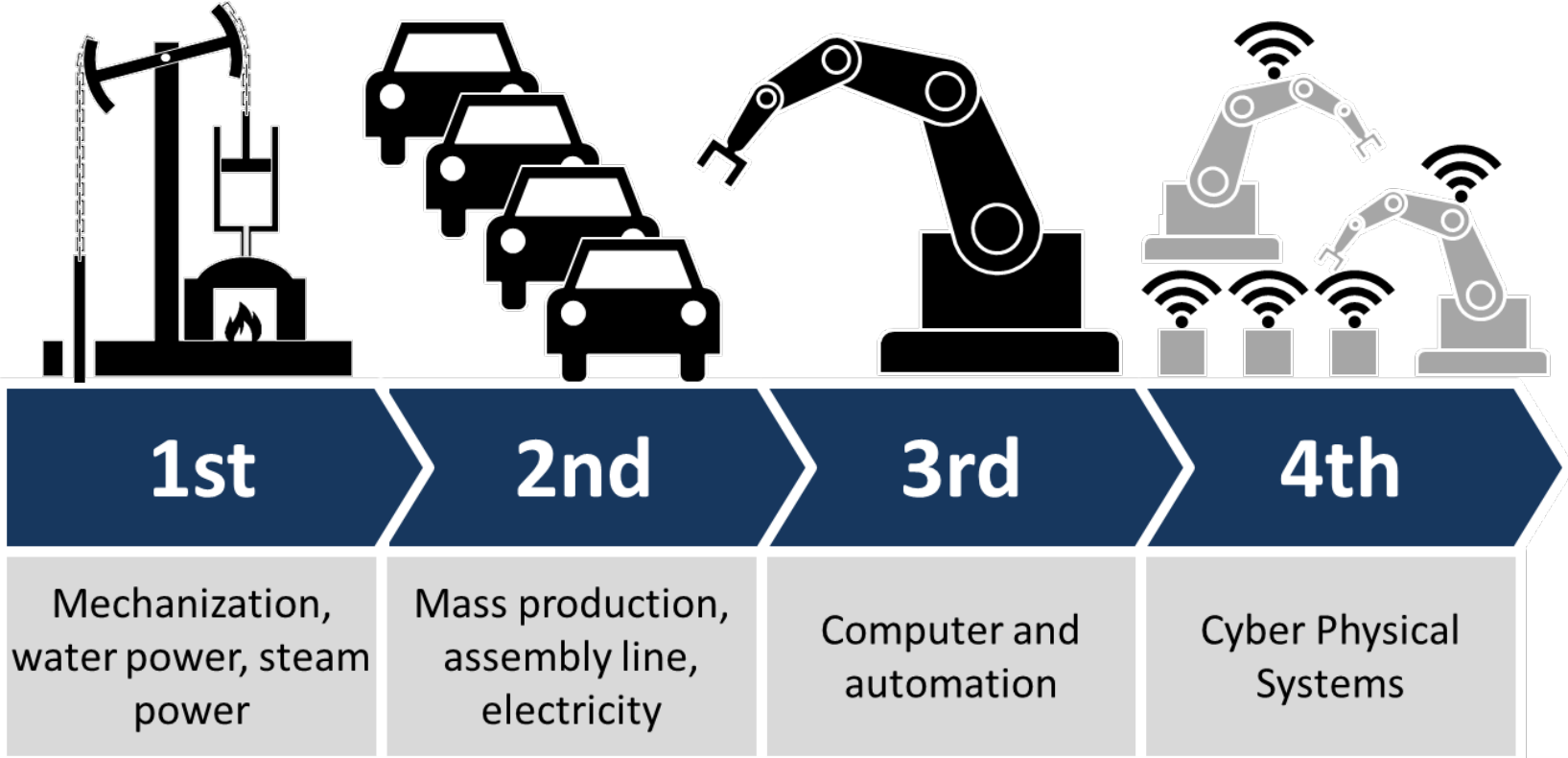
Communication  
Industry 4.0

Big Data

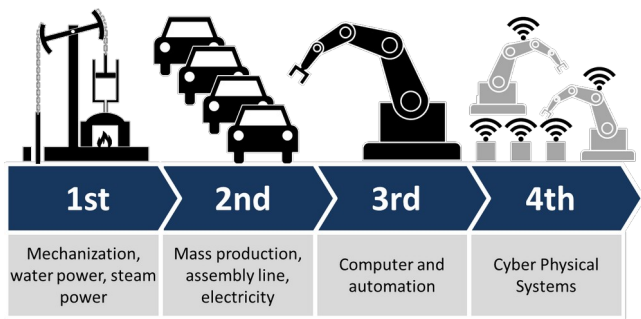
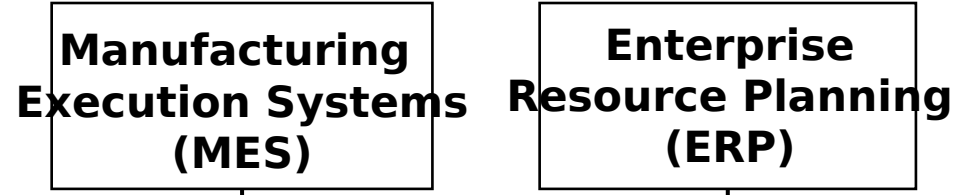


Machi  
ne

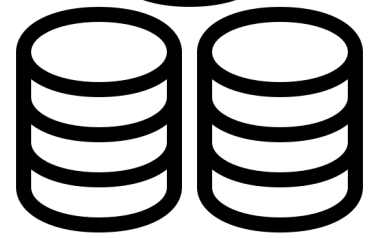
# The Fourth Industrial Revolution (Industry 4.0)



## COURBON Software



+



### AI applications in the industry :

- early detection of rejects
- **Predictive Maintenance**
- quality control
- industrial prognosis

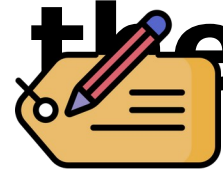
**Industry 4.0**

**Big data**

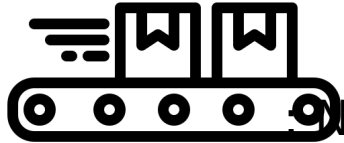
**Intelligent industrial system<sup>etc.</sup>**



# Labeling of the cause of the error



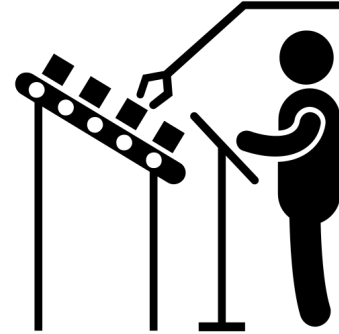
- Workers labels by hand the cause of failures of every equipment, e.g.,



Noticed **High Speed** of the conveyor



Noticed **Low energy** of the conveyor



- Every data about the status of machine is collected
  - Every issues noticed are reported and data are labeled
- ⇒ Extract insight from data to **solve** the **prevent future** the problems.
- ⇒ The manager (or workers) gain domain-specific knowledge and expertise.

Create Database



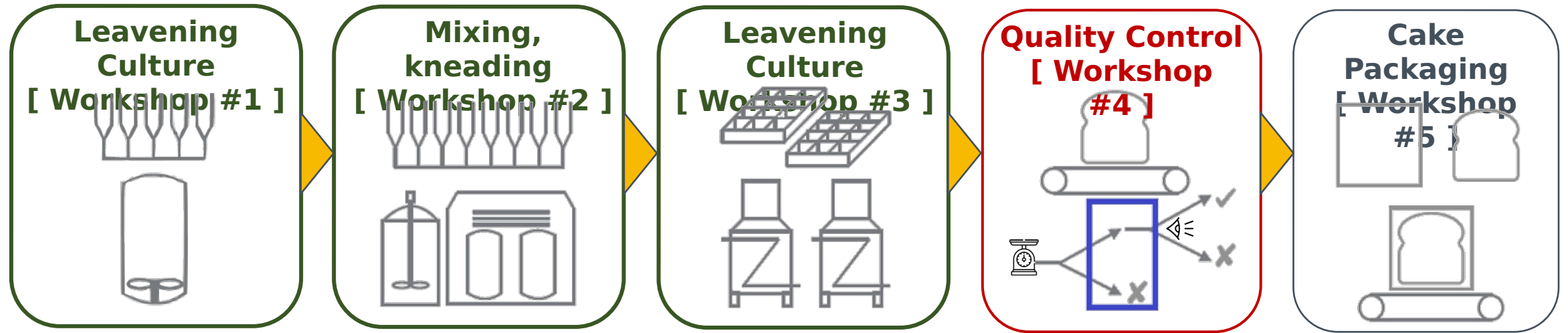
Machine N°	DateTime	Product lot N°	Issue label
M1	24022023T19:00:02	152 (Material)	0
M2 (Conveyor)	24022023T16:02	12 (Toy)	Low Energy



# Use Case - Quality Control in Cake Factory



## Example of Industrial Use Case



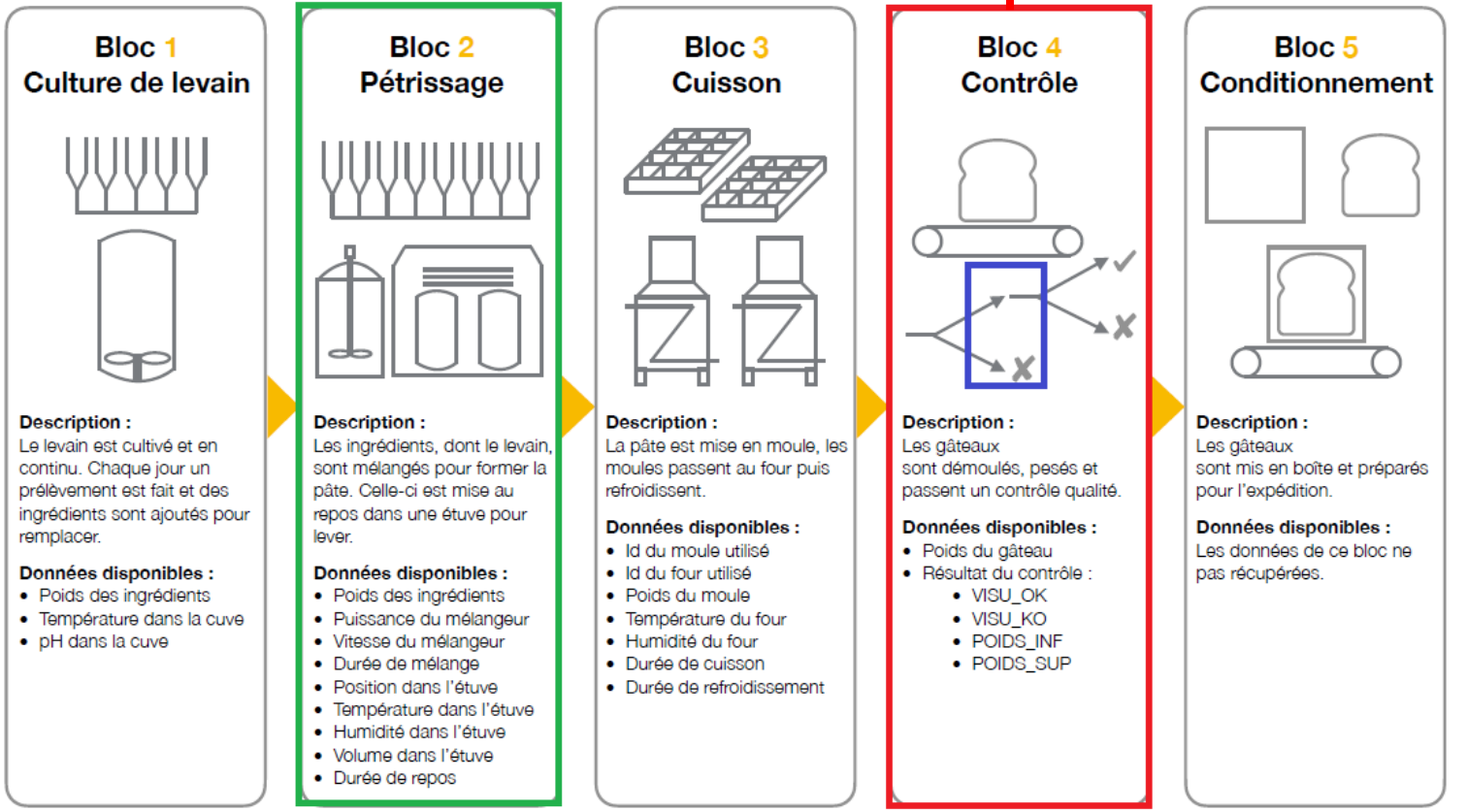
Description of an example **of Sequential Production-Line**



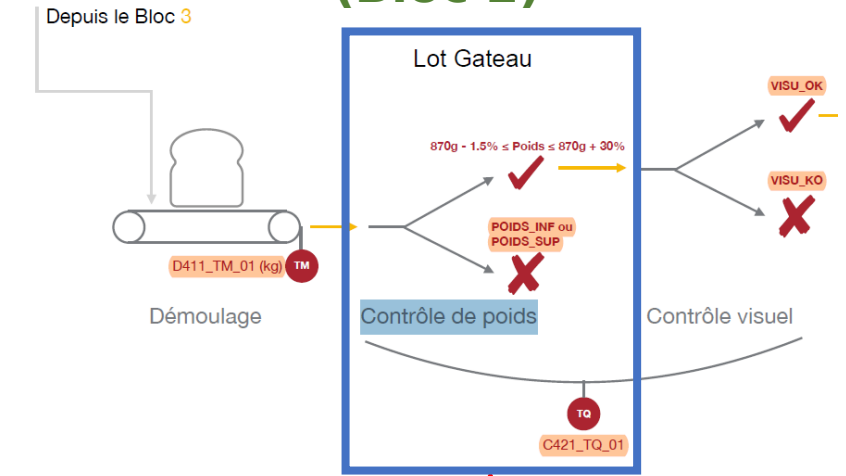
# Use Case - Quality Control in Cake Factory

A reminder of OICAKE production line schema

## DESCRIPTION GÉNÉRALE DE OICAKE



We are interested in classification problem of **weight control** using data collected in **mixing phase (Bloc 2)**



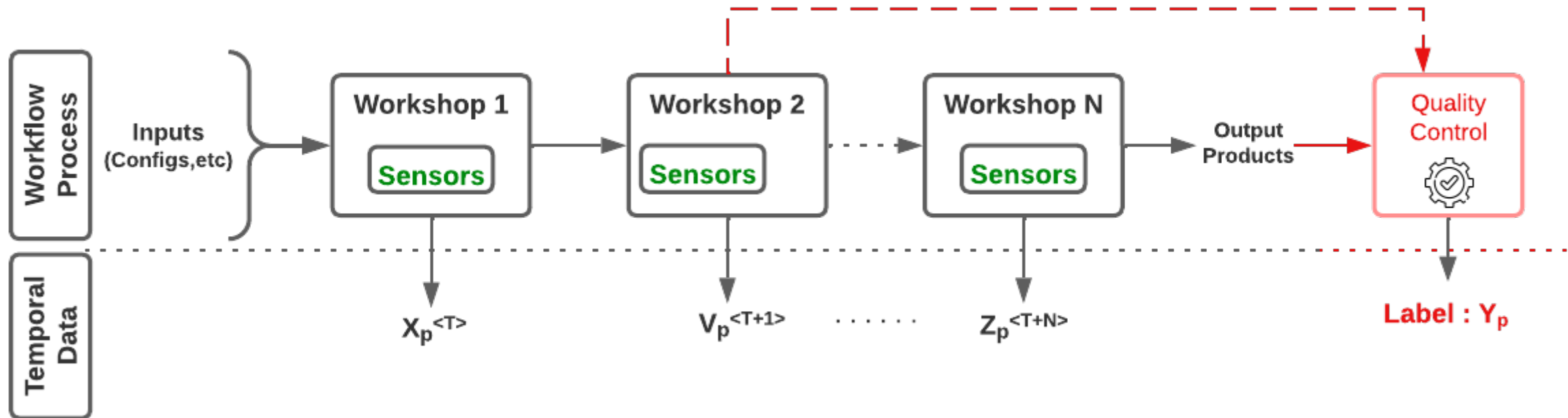


# Use Case - Quality Control in Cake Factory



## Example of Industrial Use Case

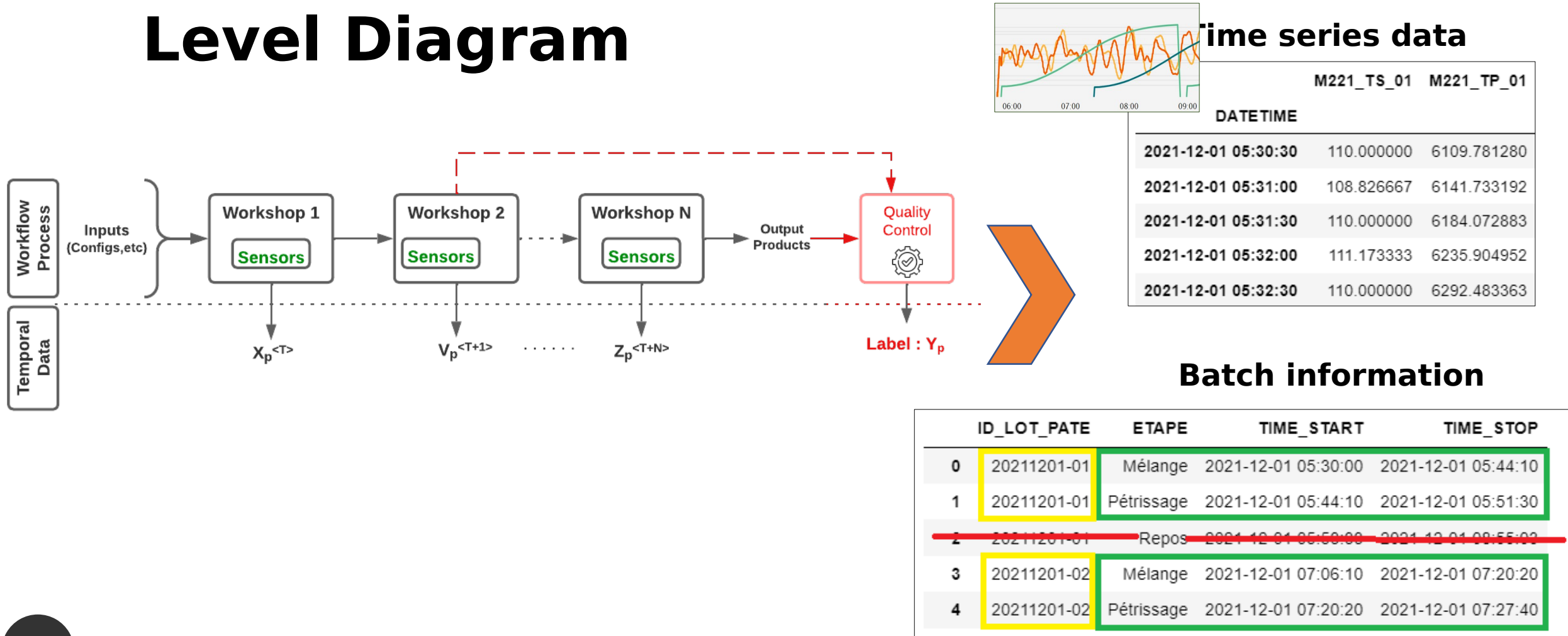
### Quality Control / Predictive Maintenance





# Use Case - Quality Control in

# Cake Factory Workflow in Production line - Multi-Level Diagram





# Train ML Model for Quality Control

## Time Series Data

	M221_TS_01	M221_TP_01
DATE TIME		
2021-12-01 05:30:30	110.000000	6109.781280
2021-12-01 05:31:00	108.826667	6141.733192
2021-12-01 05:31:30	110.000000	6184.072883
2021-12-01 05:32:00	111.173333	6235.904952
2021-12-01 05:32:30	110.000000	6292.483363



## Batch information

	ID_LOT_PATE	ETAPE	TIME_START	TIME_STOP
0	20211201-01	Mélange	2021-12-01 05:30:00	2021-12-01 05:44:10
1	20211201-01	Pétrissage	2021-12-01 05:44:10	2021-12-01 05:51:30
2	20211201-01	Repos	2021-12-01 05:50:00	2021-12-01 06:55:00
3	20211201-02	Mélange	2021-12-01 07:06:10	2021-12-01 07:20:20
4	20211201-02	Pétrissage	2021-12-01 07:20:20	2021-12-01 07:27:40

## Merge Data : Batch Identifier with the temporal data

(INPUT DATA)

	M221_TS_01	M221_TP_01	ID_LOT_PATE	ETAPE
DATE TIME				
2021-12-01 05:30:30	110.000000	6109.781280	20211201-01	Mélange
2021-12-01 05:31:00	108.826667	6141.733192	20211201-01	Mélange
2021-12-01 05:31:30	110.000000	6184.072883	20211201-01	Mélange
2021-12-01 05:32:00	111.173333	6235.904952	20211201-01	Mélange
2021-12-01 05:32:30	110.000000	6292.483363	20211201-01	Mélange
...	...	...	...	...
2022-03-01 12:13:30	118.854275	8742.818066	20220301-05	Pétrissage
2022-03-01 12:14:00	119.876393	8765.277318	20220301-05	Pétrissage
2022-03-01 12:14:30	120.888755	8787.267457	20220301-05	Pétrissage

## Wight Quality Control per ID\_PATE (TARGET DATA)

ID_LOT_PATE	QC_wight_err
20211201-01	1
20211201-01	1
20211201-01	0
20211201-01	1
20211201-02	0
...	...
20211201-02	0
20211201-02	0

Only Zeros

ID_LOT_PATE	QC_wight_err
20211201-01	1
20211201-02	0
20211201-03	0
20211201-04	1
20211201-05	0
20211202-01	1
20211202-02	0
20211202-03	1
20211202-04	0
20211202-05	0



# Train ML Model for Quality Control

## OICake

### Model Training & Evaluation

**(1) Gradient Boosting Classifier**  
(~ with 50 estimators)

**Train Precision : 97.2% Train Recall: 96%**  
**Test Precision : 96.9% Test Recall : 85.1%**

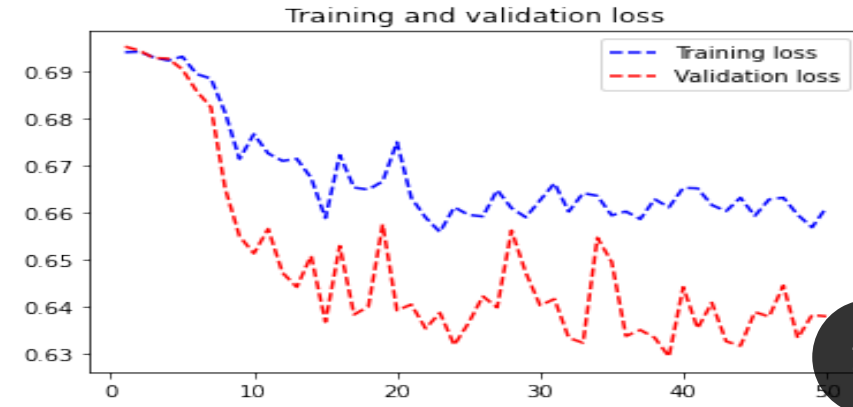
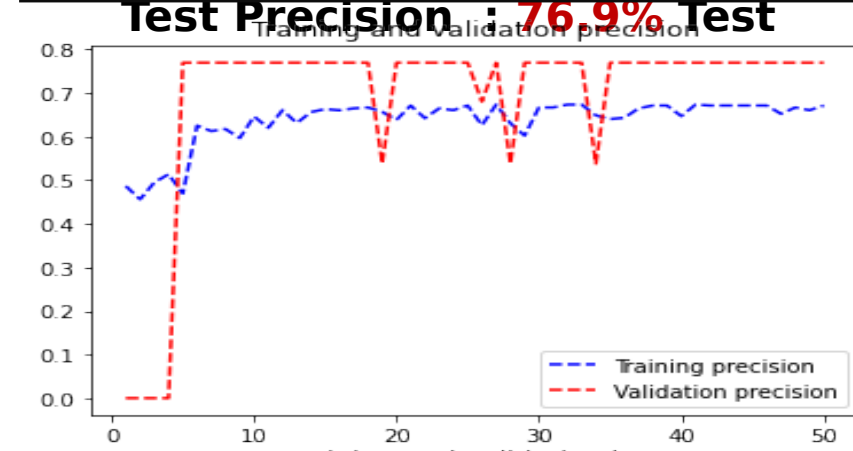
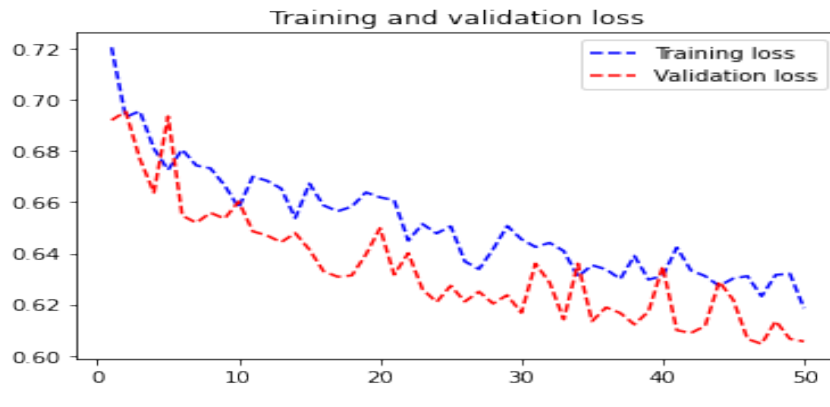
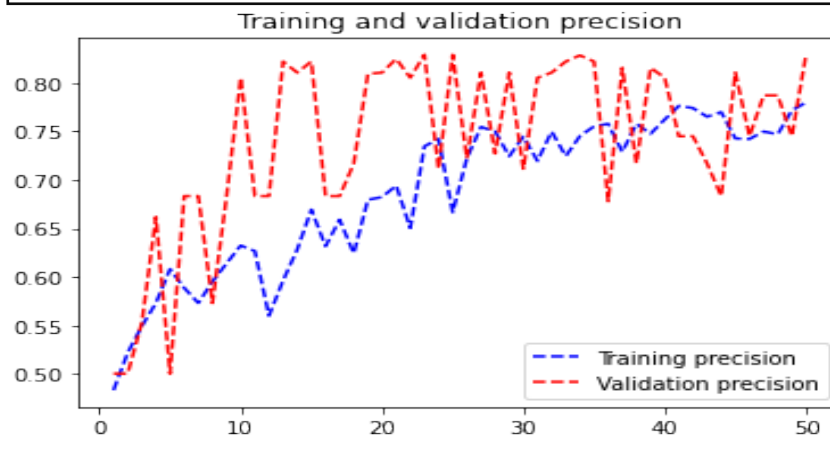
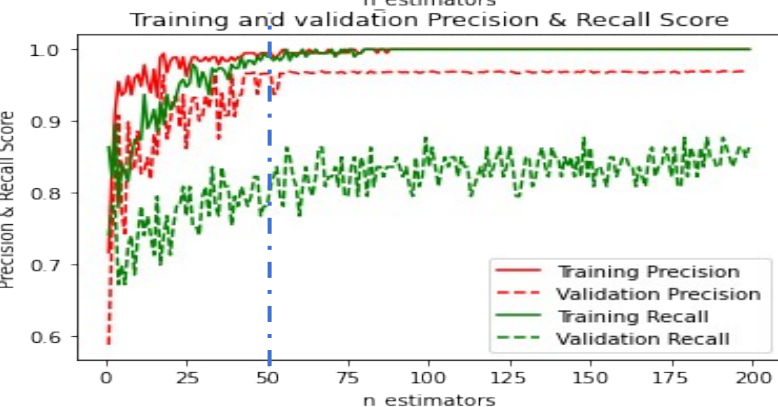
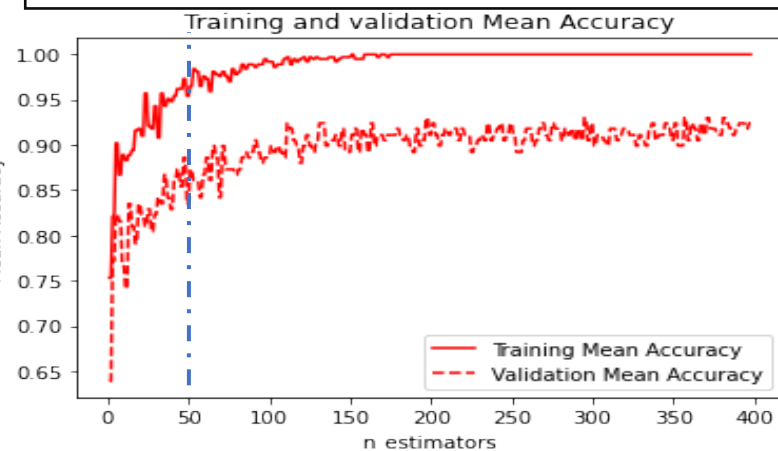
**(2) Deep Neural Network**

**Train Precision : 83.3% Train Recall: 40.8%**  
**Test Precision : 82.9% Test Recall : 43%**

**(3) LSTM**

**Train Precision: 67.1% Train Recall: 49.5%**

**Test Precision : 76.9% Test**



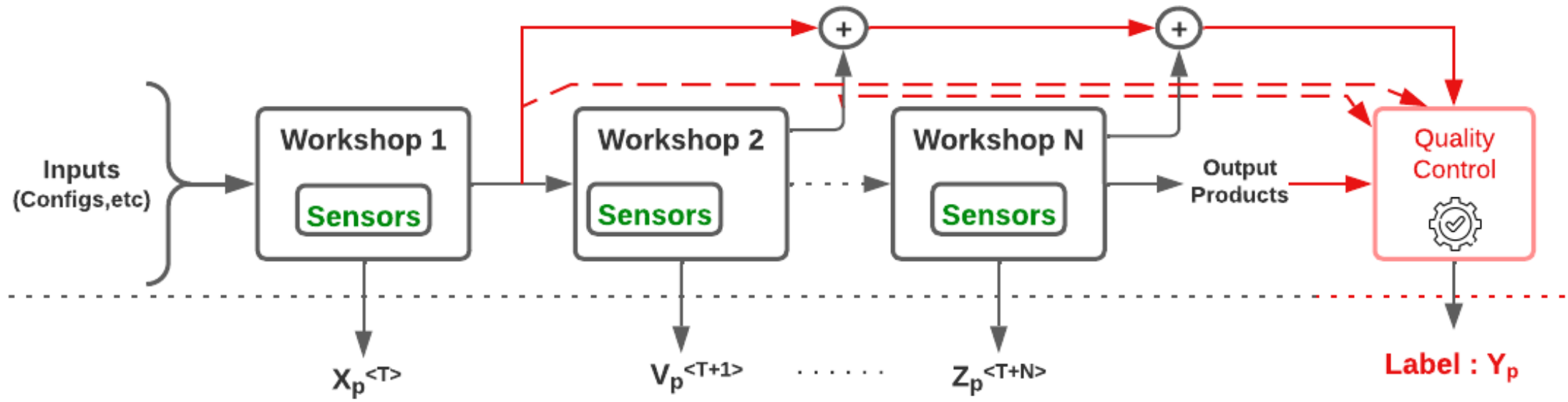


# Use Case - Quality Control in Cake Factory



## Example of Industrial Use Case

### Quality Control / Predictive Maintenance

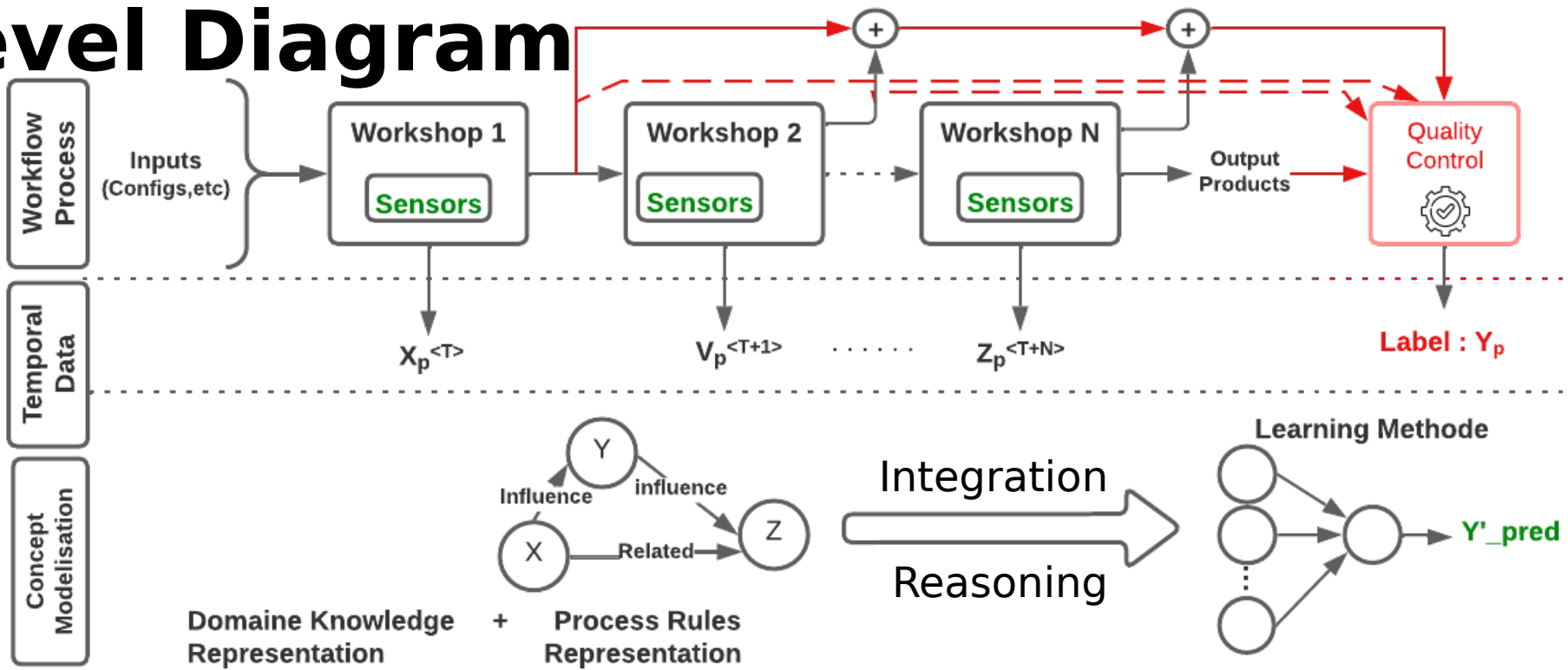






# Use Case - Quality Control in

# Cake Factory Workflow in Production line - Multi-Level Diagram



information about the process

Expert Knowledge

What is a **Concept** ?  
What knowledge reasoning skill we need to integrate to our model ?



# RESEARCH QUESTION

How to **Integrate** Domain-specific (industry 4.0) **knowledge** into **Machine Learning** to enhance its performance in **downstream tasks** ?

➤ **RQ1** : How to **integrate numerical data** from **heterogeneous** observations to apply machine learning to downstream tasks? => **C1 [ DATA INTEGRATION & HETEROGENEITY ]**

➤ **RQ2** : How to learn **the implicit knowledge** embedded in the industrial process and reason on it using machine learning (ML) models? => **C1 [ DATA INTEGRATION & HETEROGENEITY ] & C2 [ Explicability ]**



## Problems & Challenges

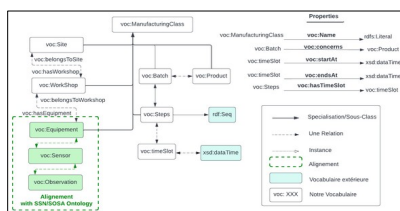
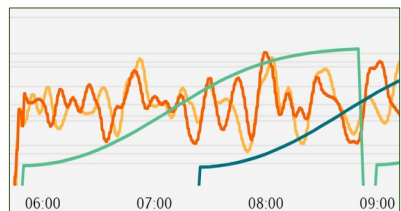
➤ **C1 [ Data integration and heterogeneity ]** : The overabundance and heterogeneity of available data **limits application of AI techniques** in the industry.

➤ **C2 [ Explicability ]** : ML/AI trained on raw data produces **black-box models** which lack **explicability**.



# Industrial Applications (Overview)

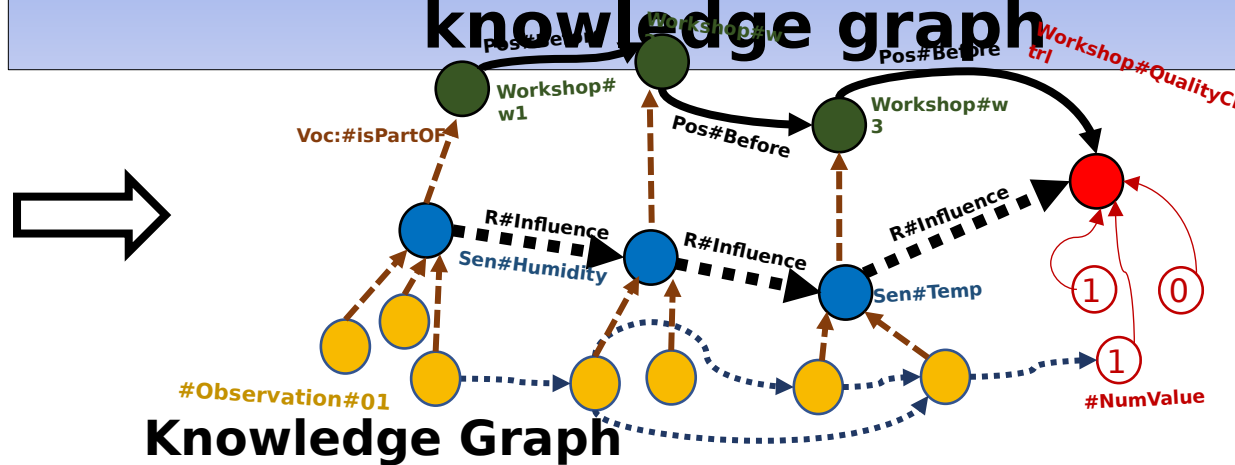
## (1) Input Data



Process Data

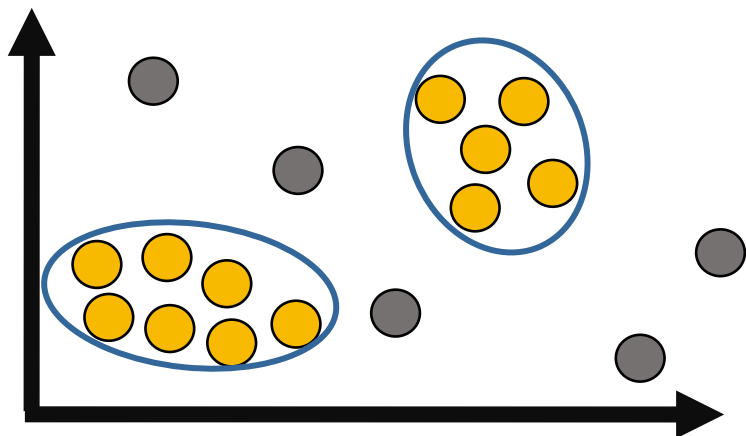
Ontology

## (2) Integrate temporal data in knowledge graph



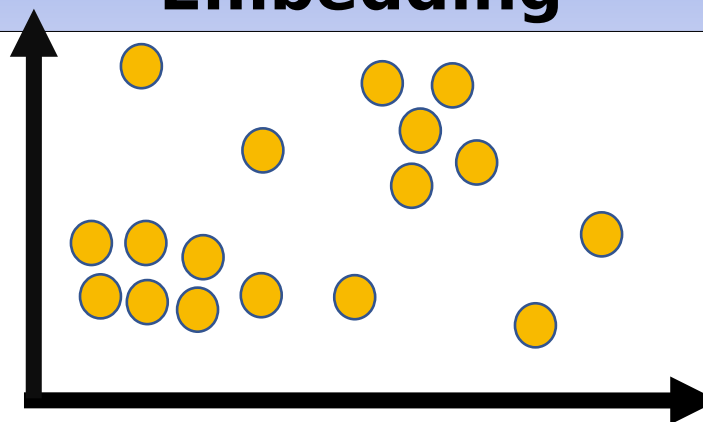
Knowledge Graph

## (4) Downstream Task



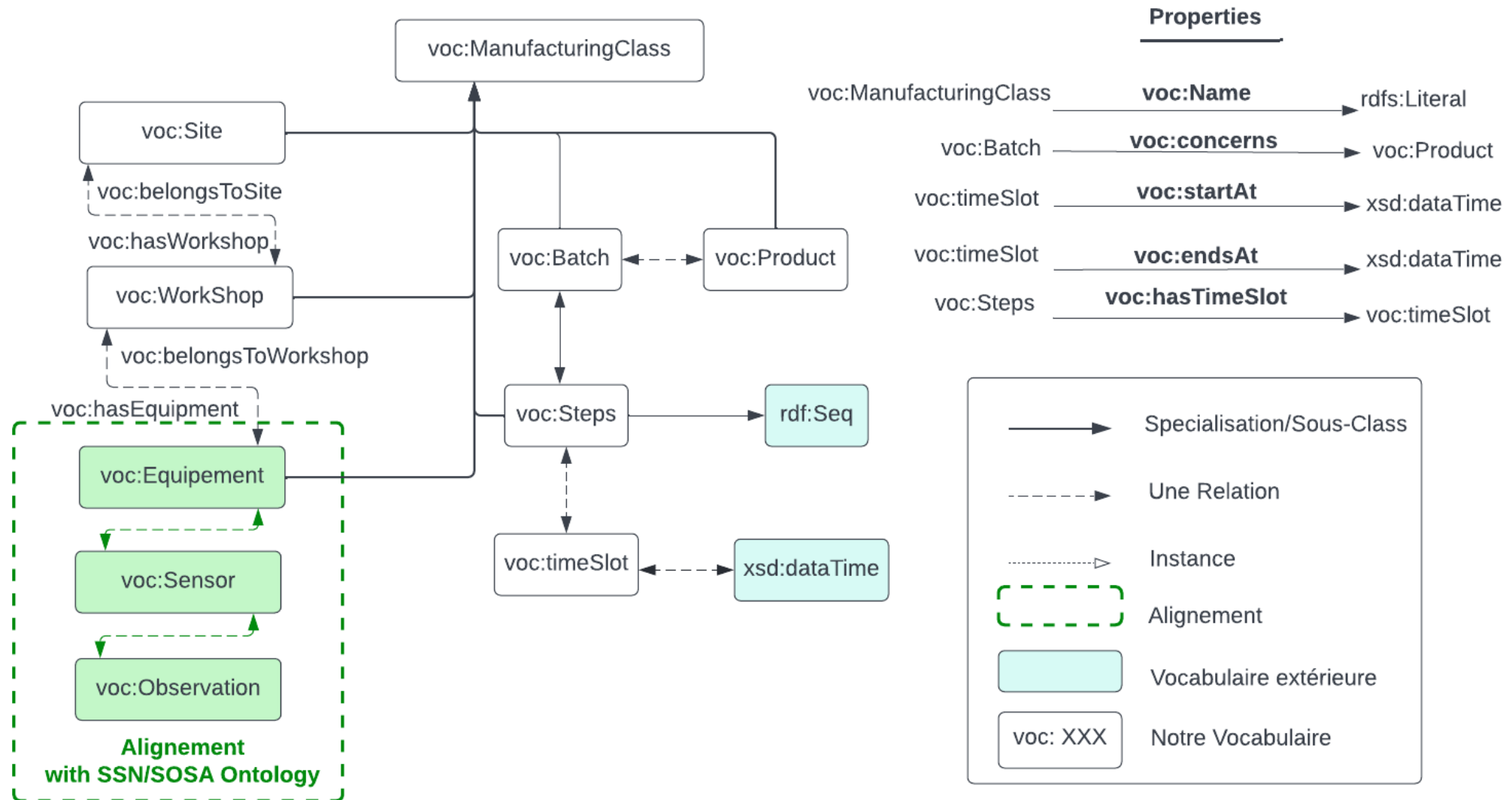
Classification

## (3) Generate Knowledge Graph Embedding



# Industrial context - Modeling example

## Proposed Hierarchical Ontology for the "OICake" (cake factory) manufacturing process





# My Posters / Join Me To Check My WORK



## On the Contribution of Domain-Specific Knowledge to Machine Learning

IFERROUDJENE Mouloud [CIFRE PhD Student]  
Supervised by Victor CHARPENAY, Antoine ZIMMERMANN, Thierry LAVELLE

Une école de l'IMT

**Context**

The industry 4.0: Our research work is situated in the domain of Cyber Physical and modern industrial systems, where the latter are expected to be interconnected and provide the needed technology to exchange and share data between its equipment (Sensors) to make critical decisions.  
Big Data: Large amount of data is collected from MES by leveraging enterprise software in Application Level and the integrated sensors in the Hardware level (Manufacture Machines).  
AI Applications in the Industry: Early Waste Detection, Quality control, Industrial prognosis, etc.

**Keywords**

- Neuro-Symbolic AI
- Knowledge Graph
- Deep Learning
- Semantic Web
- Explainability
- Industry 4.0

**Description of an example of Production-Line**

**Proposed Approach**

(1) Input Data  
Fig.1. Temporal data (Raw Data) collected from sensors

(2) Integrate temporal data in knowledge graph  
Fig.2. Proposed Ontology for a Production Line Aligned with S2K [2] and SARET [2]

(3) Generate New Data Representation

(4) Downstream Task

**Problems & Domain Challenges**

C1 [DATA INTEGRATION]: The overabundance and heterogeneity of available data limits application of AI techniques in the industry.  
C2 [EXPLAINABILITY]: ML/AI trained on raw data produces black-box models which lack explainability.

**Research Questions**

Global  
How to Integrate Domain-specific (industry 4.0) knowledge into Machine Learning to enhance its performance in downstream tasks?

Specific Questions

RQ1: How to integrate numerical data from heterogeneous observations to apply machine learning to downstream tasks?  
RQ2: How to learn the implicit knowledge embedded in the industrial process and reason on it using machine learning (ML) models?  
RQ3: How to integrate numerical data into a knowledge graph, in order to apply machine learning to downstream tasks?

**Related Work**

Clinical Embedding of Patients (CLEP) [5]  
An Evaluation of Knowledge Graph Embeddings for Autonomous Driving Data: Experience and Practice [4]

[1] Provide Explanations based on the knowledge graph embedding of Screens Instances -> Solve C2  
[2] Doesn't provide any method to explain the Knowledge graph embedding for enhancing the performance of the used Deep Learning e.g., CNN for Image Segmentation

[3] Propose an integration method and try to answer the question in C1.  
[4] I don't provide explainability C2, as they are not embedding patients' data, so I consider them as not explainable.

**Bibliographical References**

[1] Wang, M., Liu, L., & Wang, X. (2021). A survey on knowledge graph embeddings for link prediction. *Knowledge-Based Systems*, 220, 106884.  
[2] Borlini, A., Giacinto, G., Giacinto, A., Motta, J., & Valentini, G. (2019). Transferring embeddings for missing information data. Advances in neural information processing systems, 32.  
[3] Tsochantzidis, A., & Chen, C. (2014). Joint Observation vector learning for knowledge base and text completion. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).  
[4] Borlini, A., Giacinto, G., Giacinto, A., Motta, J., & Valentini, G. (2019). Transferring embeddings for missing information data. Advances in neural information processing systems, 32.  
[5] Yang, B., Yu, W., He, X., Gao, J., & Ding, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. In Proceedings of the AAAI conference on artificial intelligence (AAAI).

## FB15k-CVT: A Challenging Dataset for Knowledge Graph Embedding Models

IFERROUDJENE Mouloud, Victor CHARPENAY, Antoine ZIMMERMANN, Thierry LAVELLE

Une école de l'IMT

### FB15k-237: A Dataset Built From Freebase

For KG Completion And transductive Link prediction [1]

Freebase

Used For

Knowledge Base (General Facts) -> Knowledge Graphs (Triples Store) -> Subsets of KG (Datasets)

FB15k [2]

- Subset of Freebase
- Created by Antoine Bordes et al
- 592,213 triples
- 14,951 entities
- 1,345 relationships

FB15k-237 [3]

- Most used dataset evaluation KGEMs
- Created by Toutanova and Chen
- Excludes redundant relations
- Excludes direct training links (Held-out triples), No Near-duplicate and Inverse relations
- 272,115 triples, 237 relationships

Goal: Making the task more realistic Non-trivial inference

### FB15k-237: Disadvantageous Design

Additional preprocessing steps

Two-hop composed relationships  
144 Relations aren't atomic  
Not all paths have been preserved  
There's NO Freebase n-ary relations

prefix fb: http://rdf.freebase.com/ns/

- fb:om.0cn180 fb:award\_nominee.award\_nominations:fb:award\_nomination.award fb:om.0cqh40
- fb:om.02358 fb:award\_nominee.award\_nominations:fb:award\_nomination.award fb:om.0cqh40 ; fb:award\_nominee.award\_nominations:fb:award\_nomination.award fb:om.0cqh40 ;
- fb:award\_nominee.award\_nominations:fb:award\_nomination.award fb:om.02g5h5
- fb:om.02358 fb:award\_nominee.award\_nominations:fb:award\_nomination.award fb:om.0cqh40

fb:om.0cn180 (Phyllis Smith), fb:om.02358 (Calista Flockhart) and fb:om.02g5h5 (Peter MacNicol) are movie actors.  
All nominated to the same award "Screen Actors Guild";  
C. Flockhart and P. MacNicol were nominated to the same edition.  
BUT: P. Smith and C. Flockhart were not nominated at the same edition.  
Hence: The last triple seems to have a high plausibility, which is NOT

### Research Questions

RQ1. How KGEM models can distinguish between these facts and learn what is correct? In n-ary relation correctly modeled?  
RQ2. How does the lack of atomic relations in FB15k-237 affect the performance of KGEM models in real world applications?

### Contribution: FB15k-CVT The Proposed Dataset

Freebase (Last update) (in August 2015)

Reintroducing CVT To represent n-ary Relation.

Entity Filtering (FB15k-237 topics) -> Select Relations (decomposed + direct relations in FB15k-237)

### Fair Evaluation: Link Prediction to Path Prediction

+ Same Train/Valid/Test entities split as in FB15k-237 dataset + Split two-hop relations, Adding intermediate nodes (CVTs) + All Topics in different subject should appear in training set + Not all CVTs appears in training set + Not all CVTs are learned by KGEM Models + Baseline KGEMs: TransE [4] and DistMult [5] + Experiment: Path prediction (PP) on FB15k-CVT

### Adapting Score Functions

Where? Modifying scoring function in the evaluation phase  
Why? Not taking CVT representations into account  
Can it learn the space representation of intermediate nodes

TransE

$$\|TransE(e_1, r_1, r_2, e_2)\| = \|(e_1^* + \vec{r}_1 - \vec{r}_2) - (e_2^* - \vec{r}_2)\|$$

### Conclusions

- Our paper addresses the limitations of current KGEM evaluation datasets and introduces FB15k-CVT.
- Our dataset FB15k-CVT is created from Freebase and FB15k-237. It is an exact subset of Freebase.
- We argue that the reintroduction of CVTs in the dataset offers new challenges for emerging KGEMs.
- Preliminary evaluation results shows that the overall performances of the two models significantly decrease in the presence of CVTs.

### References

[1] Wang, M., Liu, L., & Wang, X. (2021). A survey on knowledge graph embeddings for link prediction. *Knowledge-Based Systems*, 220, 106884.  
[2] Borlini, A., Giacinto, G., Giacinto, A., Motta, J., & Valentini, G. (2019). Transferring embeddings for missing information data. Advances in neural information processing systems, 32.  
[3] Tsochantzidis, A., & Chen, C. (2014). Joint Observation vector learning for knowledge base and text completion. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).  
[4] Borlini, A., Giacinto, G., Giacinto, A., Motta, J., & Valentini, G. (2019). Transferring embeddings for missing information data. Advances in neural information processing systems, 32.  
[5] Yang, B., Yu, W., He, X., Gao, J., & Ding, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. In Proceedings of the AAAI conference on artificial intelligence (AAAI).

## FB15k-CVT: A Challenging Dataset for Knowledge Graph Embedding Models

Mouloud Iferroudjene<sup>1,2</sup>, Victor Charpenay<sup>1</sup> and Antoine Zimmermann<sup>1</sup>

<sup>1</sup>Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, Saint-Etienne, France  
<sup>2</sup>COURBON Software, 70 rue de la Montat - CS 60327, 42015 Saint-Etienne, France

### Abstract

Knowledge Graphs (KGs) are an essential component of neuro-symbolic AI. KG Embedding Models (KGEMs) are used to represent elements of a KG (its entities and relations) in a vector space, to enable efficient processing and reasoning over knowledge. Most KGEMs are evaluated against datasets derived from the Freebase KG: FB15k and FB15k-237. In this paper, we identify limitations in these datasets with respect to Compound Value Types (CVTs), which are nodes introduced in Freebase as a substitute for n-ary relations. In FB15k and FB15k-237, CVTs have been removed, thereby eliminating valuable information. To evaluate whether KGEMs can learn semantically accurate representations of entities and relations in Freebase, we introduce here a new dataset named FB15k-CVT, which reintroduces the deleted CVT nodes. In a preliminary evaluation, we assess the limitations of baseline KGEMs (TransE, DistMult) in the presence of CVTs. The evaluation suggests that KGEMs based on tensor decomposition are more promising than translational models but, most of all, it calls for further experiments with KGEMs that can answer conjunctive queries or that preserve logical entailment.

### Keywords

Knowledge Graphs, Neurosymbolic AI, Knowledge Graph Embeddings Models, FB15k-237

## 1. Introduction

Knowledge graphs (KGs) have become an essential component of neuro-symbolic AI research. A KG is a uniform source of information in which physical-world entities are represented as vertices of a directed edge-labeled graph. In the context of representation learning, edge labels of a KG are called relations, and its edges are called facts or triples [1].

KGs can be leveraged in a great variety of AI applications. Over the past decade, many KG Embedding Models (KGEMs) have been developed for that purpose [1, Sec. 4.2]. By representing entities and relations as numeric structures in a vector space, KGEMs provide a way to integrate both symbolic and sub-symbolic knowledge, enabling efficient processing and reasoning over complex and heterogeneous data. Most KGEMs are evaluated against datasets that are derived from Freebase<sup>1</sup>, a (now archived) public KG containing millions of entities and billions of facts.

In KGEM research, the most notable datasets derived from Freebase are FB15k and FB15k-237. FB15k is a subset of Freebase that includes 15k entities selected among the most frequent entities

NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning, Certosa di Pontignano, Siena, Italy  
mouloud.iferroudjene@emse.fr (M. Iferroudjene); victor.charpenay@emse.fr (V. Charpenay); antoine.zimmermann@emse.fr (A. Zimmermann)  
© 2023 Copyright by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).  
CEUR Workshop Proceedings (CEUR-WS.org)  
<sup>1</sup>Freebase data dumps https://developers.google.com/freebase, accessed 20 March 2023.

## Membres :

- Valentin GUIEN
- Mouloud IFERROUDJENE
- Frederic HAYEK
- Aurelien MOMBELLI

# PhD students' seminar



Introduction to  
Knowledge Graph  
Embedding  
Models (KGEMs)

Presented by :  
IFERROUDJENE Mouloud

05/02/2024

PhD students'  
seminar

" Entre nous, parlons science "

LIMOS



LIMOS

LABORATOIRE D'INFORMATIQUE,  
DE MODELISATION ET D'OPTIMISATION DES SYSTEMES



agaetis

Measuring Data Quality in Classification  
Applications

Roxane Jouseau, Sébastien Salva,  
and Chafik Samir  
06/03/2024

PhD students'  
seminar

Knowledge  
Representation  
for Ecosystems of  
Digital Twins

LIMOS  
CLERMONT  
AUVERGNE  
ÉCOLE DOCTORALE  
DES SCIENCES POUR L'INGÉNIEUR



Samuele Burattini - Alma Mater Studiorum University of Bologna

The End

Thank You For Your  
Attention



Any  
Questions ?