Data integration: a perpetually evolving challenge for new research perspectives

HDR Defense

Francesca Bugiotti

CentraleSupélec, LISN, CNRS, Paris-Saclay University

LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

2012

PhD
Roma Tre

2013

Post Doc
Inria

2015

Assistant professor
CentraleSupelec

2023

Promotion
CentraleSupélec

2025

HdR defense

Supervision and Projects

- ## PhD students and Post-Docs

Molood Arman discussed 2023, Shwetha Salimath 3rd year, Quentin Bruant 3rd year, Jyotishka Das starting, Yuchen Tao starting, Adnan El-Moussawi 2021, Charles Ndungu-Ndegwa 2025

- ## Projects

Vrailexia (2021), Remission RHU (2024), GeoTS (2024), BMP trajectory Analyses (2023), IT4Energies (2021), Proclaim (2019), B-Graph (2018), NOAM (2016), Estocada (2015), SOS (2014), MATRIX-EXL (2014), MIDST (2013)

- ## Industry Collaborations

Genvia, SLB, Transvalor, Tissium, Vires, Dalkia, Generali, Solinum, Central Bank of Italy, Consip, ISA

- ## Academic Collaborations

Roma Tre, Nanterre University, Inria, CEA, TU Berlin, University of Oulu, Nairobi University, Tuscia University, Cordoba University

Data integration:
a perpetually evolving challenge for new research perspectives

Problem

- Data is the key engine or the output of almost any kind of application
- Ideally, we want to give applications the possibility to access any kind of data, stored according to any possible model and format

Challenges:

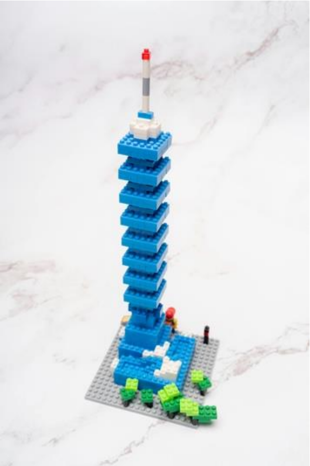
- Distributed Data
- Heterogeneous data

Solution:

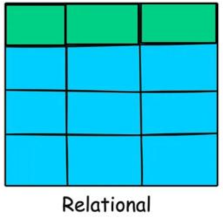
- Integrate data

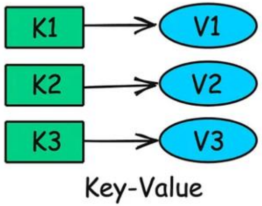
Data Integration

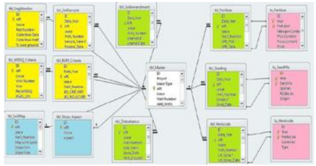
Data Integration

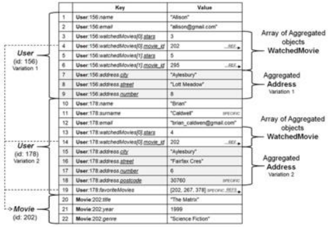


Data Integration

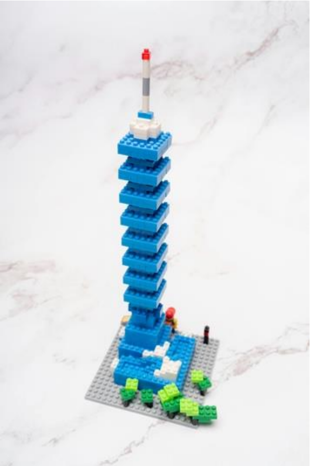






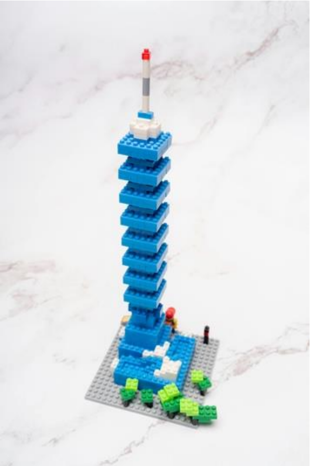


Data Integration

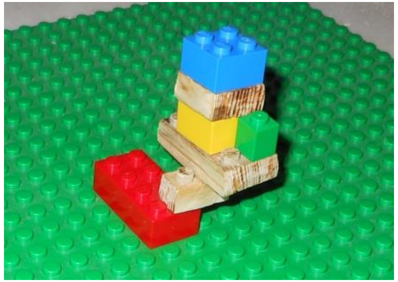


?

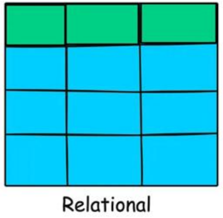
Data Integration

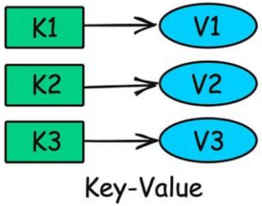


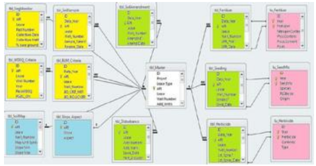
?

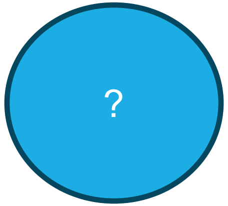


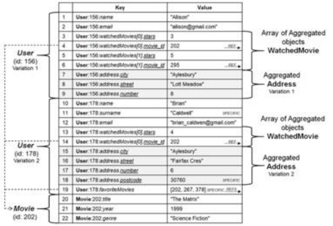
Data Integration

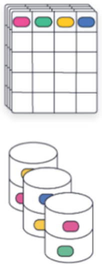












ID	Country	City	Building	Area	Pin code
1	India	Noida	Galaxy Tower	Sector 62	110078
2	India	Gurgaon	NULL	Sector 69	110040
3	US	New York	NULL	Manhattan	NULL

Key	Value
India	{"Galaxy Tower sector 58, Noida, 110078", "Sector 69, Gurgaon, 110040"}
US	"Manhattan, New York"

Data Integration Approaches

- Meta model data integration
- Semantic data integration
- Structural data integration
- Software-delegating data integration

Data Integration Approaches

- Model data integration
  - All data belongs to a unified schema in a Target Metamodel

Data Integration Approaches

- Semantic data integration
  - A general domain ontology represents all the concepts

Data Integration Approaches

- Structural data integration
  - Data integration occurs at the physical storage level

Data Integration Approaches

- Software-delegating data integration
    - Off-the-shelf software is used for integration

Early Contributions

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

CentraleSupélec

Main research areas and contributions

- Metamodel Data Integration
- Graph Data Integration and Large Language Models
- Data preparation and analysis for Time Series in the Energy Domain

Main research areas and contributions

- **Metamodel Data Integration**

- Graph Data Integration and Large Language Models

- Data preparation and analysis for Time Series in the Energy Domain

Metamodel Data Integration

- **Collaborators**: Paolo Atzeni, Luigi Bellomarini, Luca Cabibbo, Jesus Camacho-Rodriguez Marco De Leonardis, Adnan El-Moussawi, Moditha Hewasinghage, Zoi Kaoudi, François Goasdoue, Ioana Manolescu, Riccardo Torlone, Nacéra Seghouani, Stamatis Zampetakis

- **Projects:** NOAM, Estocada, SOS, MIDST, MATRIX-EXL

- **Papers**: Linked Data Management 2022, DEXA 2020, ER 2018, CIDR 2015, ER 2014, EDBT 2013

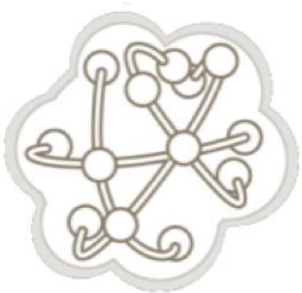
Metamodel Data Integration

Metamodel Data Integration

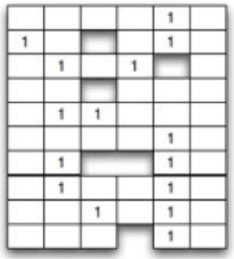
- *NoSQL datastores:*
  - new generation of distributed database systems
  - large data sets distributed over many servers

NoSQL data models

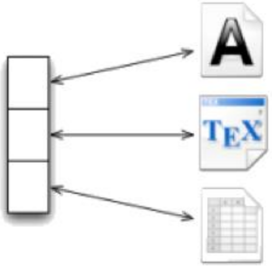
Key-Value



Graph



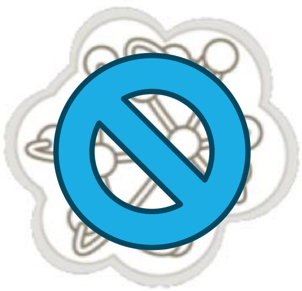
Column-Family



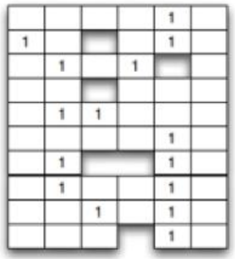
Document

NoSQL data models

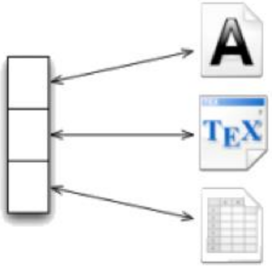
Key-Value



Graph



Column-Family



Document

Metamodel Data Integration – Goals

- **NoAM** (**NoSQL Abstract Model**) – an abstract and system-independent data model for NoSQL databases
    - commonalities of the various data models
    - abstractions to balance the differences and variations
    - general and flexible structure

Metamodel Data Integration - NoAM

- Looking for the *"smaller" data access unit called* **entry**

Example of entries:

- a column
- a field
- an individual key-value pair

Metamodel Data Integration - NoAM

- Identifying the *collections* of data access units

Example of collection:

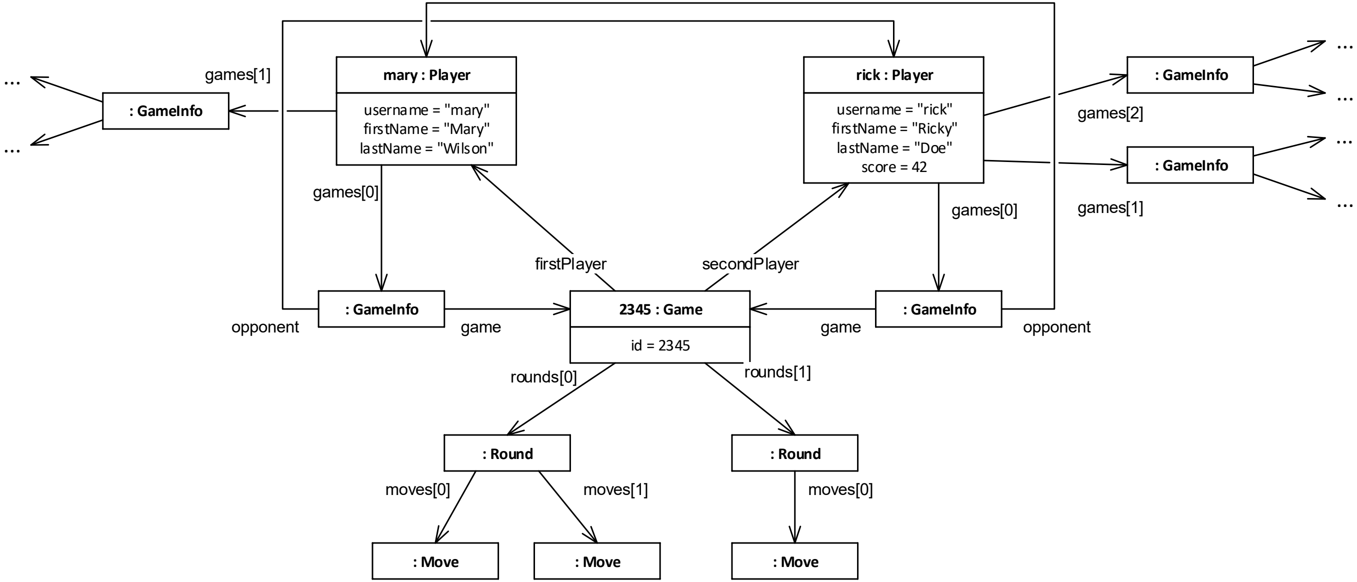
  - a table
  - a document
  - a collection of key-values

Metamodel Data Integration - NoAM

- The **NoAM abstract data model**
    - a **database** is a set of collections – each collection has a distinct name
    - a **collection** is a set of blocks – each block is identified in its collection by a block key
    - a **block** is a non-empty set of entries
    - each **entry** is a pair (*ek*,*ev*)
        - *ek* is the *entry key* – unique within its block
        - *ev* is a value (either a scalar or a complex value), called the *entry value*

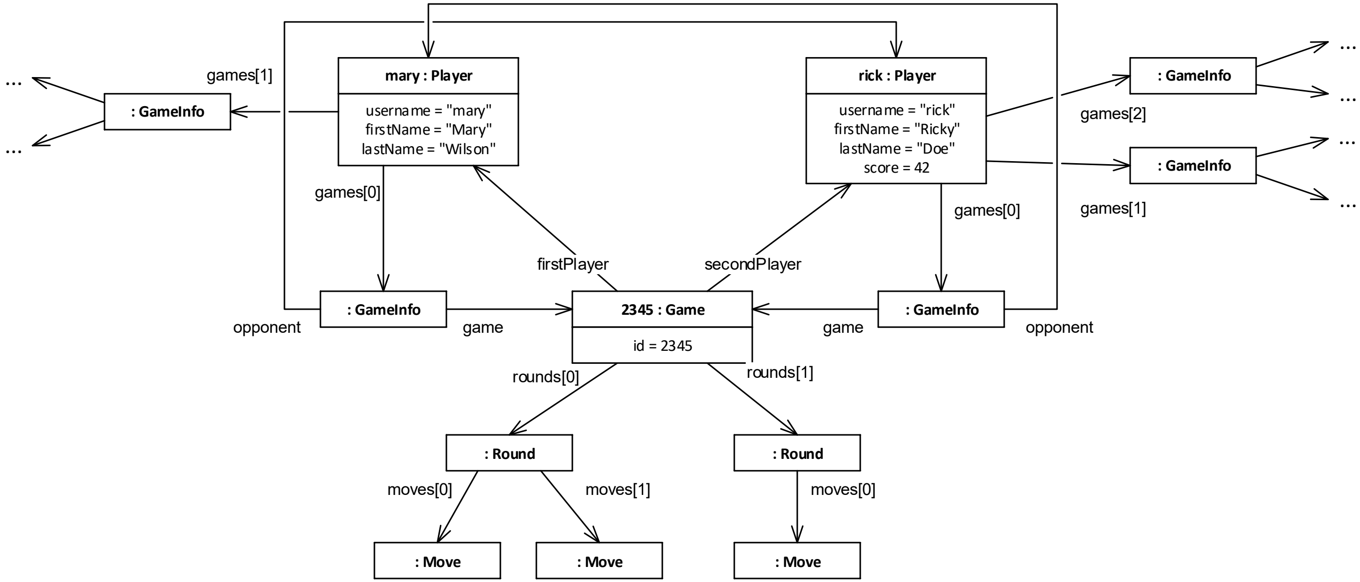
Noam in Action: A running example

- Consider a fictious online, web 2.0 game – e.g., some variant of Ruzzle – which should manage various application objects, including players, games, rounds, and moves



Noam in Action: A running example

- We start by considering application objects…

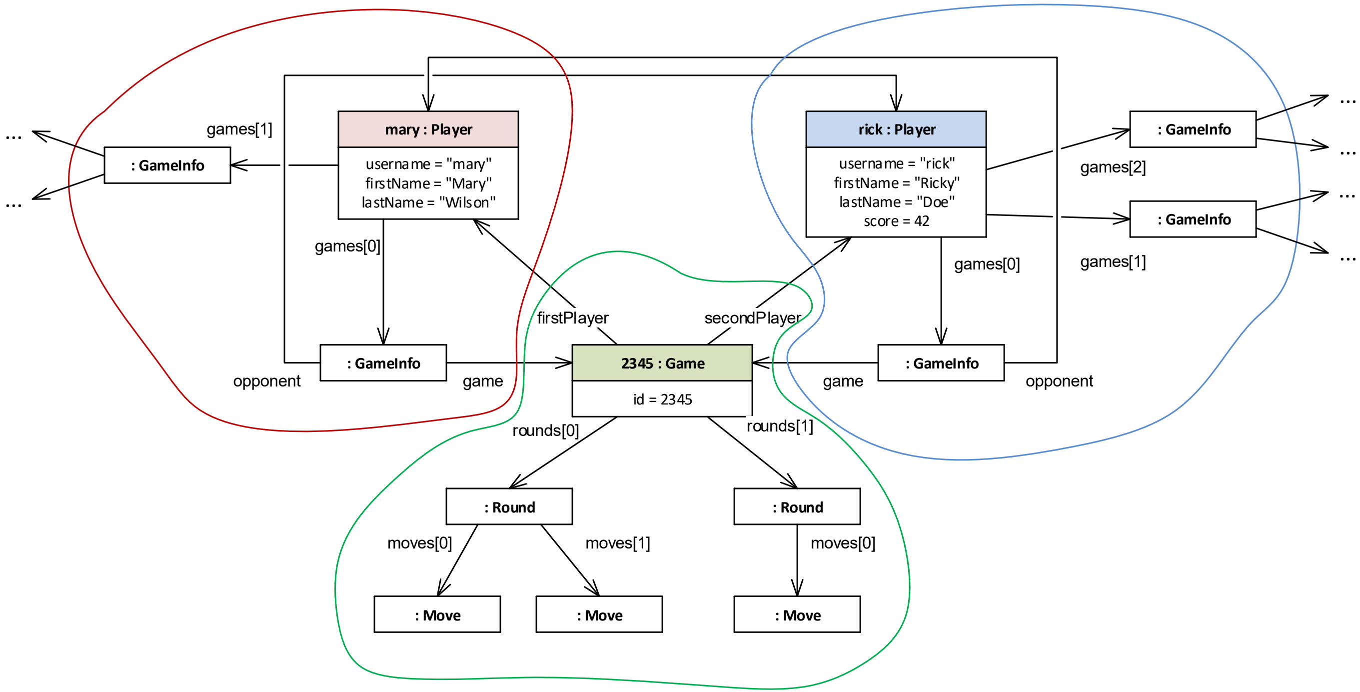


LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

CentraleSupélec

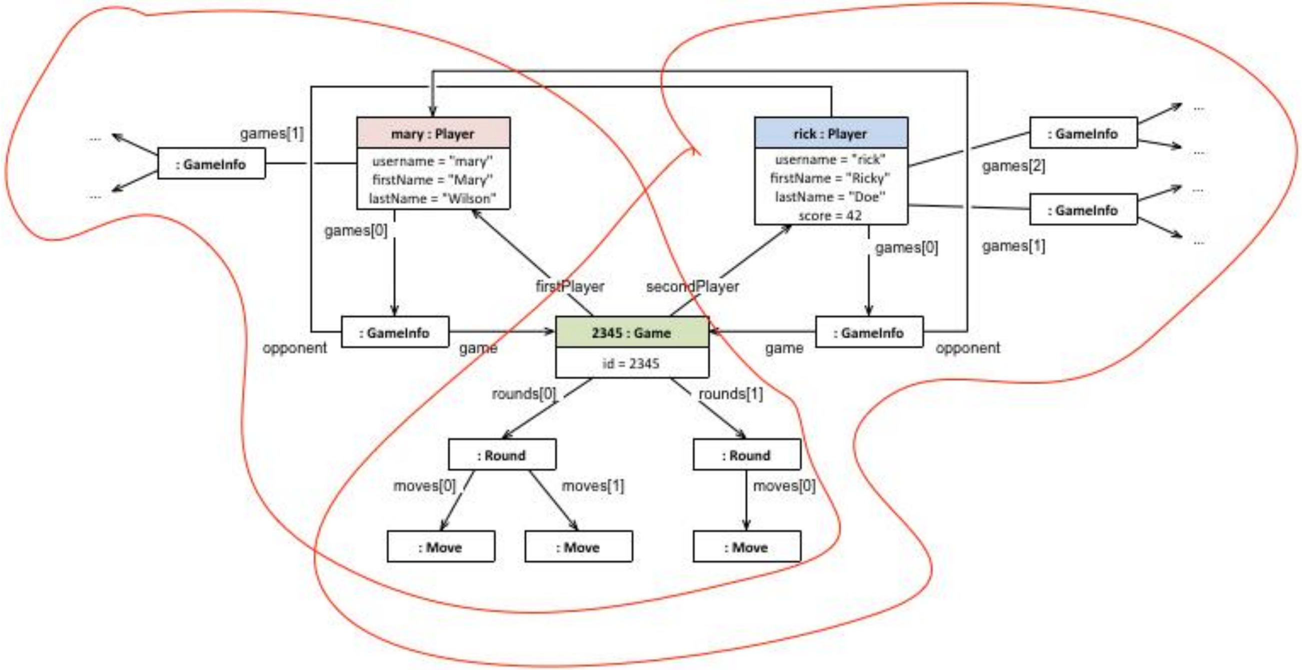
Noam in Action: A running example

- … we group them in aggregates (*decisions needed!*) …



Noam in Action: A running example

- … we group them in aggregates (*decisions needed!*) …



Noam in Action: A running example

- … we consider aggregates as complex-value objects…

```
Player:mary : ⟨
    username : "mary",
    firstName : "Mary",
    lastName : "Wilson",
    games : {
        ⟨ game : Game:2345, o
        ⟨ game : Game:2611, o
    }
⟩
```

```
Player:rick : ⟨
    username : "rick",
    firstName : "Ricky",
    lastName : "Doe",
    score : 42,
    games : {
        ⟨ game : Game:2345, opponent : Player:mary ⟩,
        ⟨ game : Game:7425, opponent : Player:ann ⟩,
        nny ⟩
```

```
Game:2345 : ⟨
    id : "2345",
    firstPlayer : Player:mary,
    secondPlayer : Player:rick,
    rounds : {
        ⟨ moves : … , comments : … ⟩,
        ⟨ moves : … , actions : … , spell : … ⟩
    }
⟩
```

Noam in Action: A running example

- … we partition these complex values: entries, blocks, and collections…

Player:mary : ⟨
username : "mary",
firstName : "Mary",
lastName : "Wilson",
games : {
⟨ game : Game:23
⟨ game : Game:26
}
⟩

Player:rick : ⟨
username : "rick",
firstName : "Ricky",
lastName : "Doe",
score : 42,
games : {
⟩

Game:2345 : ⟨
id : "2345",
firstPlayer : Player:mary,
secondPlayer : Player:rick,
rounds : {
⟨ moves : … , comments : … ⟩,
⟨ moves : … , actions : … , spell : … ⟩
}
⟩

Noam in Action: A running example

- … and represent them into NoAM (*consequence of decisions*) …

Player

mary

username	“mary”
firstName	“Mary”
lastName	
games[0]	
games[1]	

rick

username	“rick”
firstName	“Ricky”
lastName	“Doe”
score	42
games[0]	⟨ game : Game:2345, opponent : Player:mary ⟩
	ann ⟩
	johnny ⟩

Game

2345

id	2345
firstPlayer	Player:mary
secondPlayer	Player:rick
rounds[0]	⟨ moves : … , comments : … ⟩
rounds[1]	⟨ moves : … , actions : … , spell : … ⟩

Storage in NoSQL systems

- … and finally we map the intermediate representation to the data structures of the target datastore (*the approach specifies how*)

table **Player**

<u>username</u>	firstName	lastName	score	games[0]	games[1]	games[2]	…
mary	Mary	Wilson		{…}	{…}		
rick	Ricky	Doe	42	{…}	{…}	{…}	

table **Game**

<u>id</u>	firstPlayer	secondPlayer	rounds[0]	rounds[1]	rounds[2]	…
2345	Player:mary	Player:rick	{…}	{…}		

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
CentraleSupélec

Storage in NoSQL systems

- … and finally we map the intermediate representation to the data structures of the target datastore (*the approach specifies how*)

key	value
/Player/mary/-/username	mary
/Player/mary/-/firstName	Mary
/Player/mary/-/lastName	Wilson
/Player/mary/-/games[0]	{ “game” : “Game:2345”, “opponent” : “Player:rick” }
/Player/mary/-/games[1]	{ “game” : “Game:2611”, “opponent” : “Player:ann” }
…	…
/Games/2345/-/id	2345
/Games/2345/-/firstPlayer	Player:mary
/Games/2345/-/secondPlayer	Player:rick
/Games/2345/-/rounds[0]	{ … }
/Games/2345/-/rounds[1]	{ … }
…	…

ORACLE NOSQL DATABASE

Entry per Aggregate Object (EAO)

- An aggregate object is represented by a single entry
- The entry value is the whole complex value – the entry key is empty

mary	ε	⟨
		username : "mary",
		firstName : "Mary",
		lastName : "Wilson",
		games : {
		⟨ game : Game:2345, opponent : Player:rick ⟩,
		⟨ game : Game:2611, opponent : Player:ann ⟩
		}
		⟩

Entry per Top-level Field (ETF)

- An aggregate object is represented by multiple entries – a distinct entry for each top-level field of the complex value
- The entry value is the field value – the entry key is the field name

mary

username	“mary”
firstName	“Mary”
lastName	“Wilson”
games	{ ⟨ game : Game:2345, opponent : Player:rick ⟩, ⟨ game : Game:2611, opponent : Player:ann ⟩ }

Entry per Atomic Value (EAV)

- An aggregate object is represented by multiple entries – a distinct entry for each atomic value in the complex value
- The entry value is the atomic value – the entry key is the “access path” to the atomic value

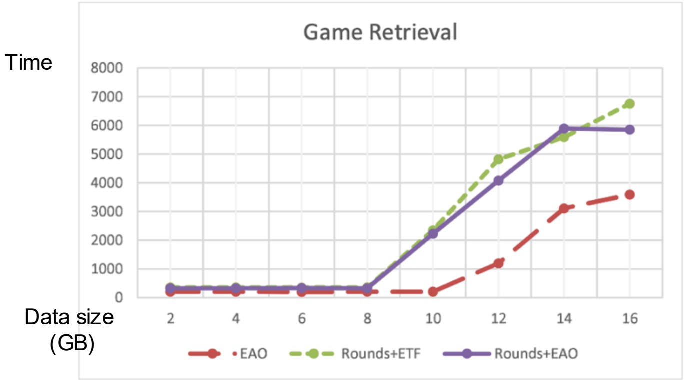
mary

username	“mary”
firstName	“Mary”
lastName	“Wilson”
games[0].game	Game:2345
games[0].opponent	Player:rick
games[1].game	Game:2611
games[1].opponent	Player:ann

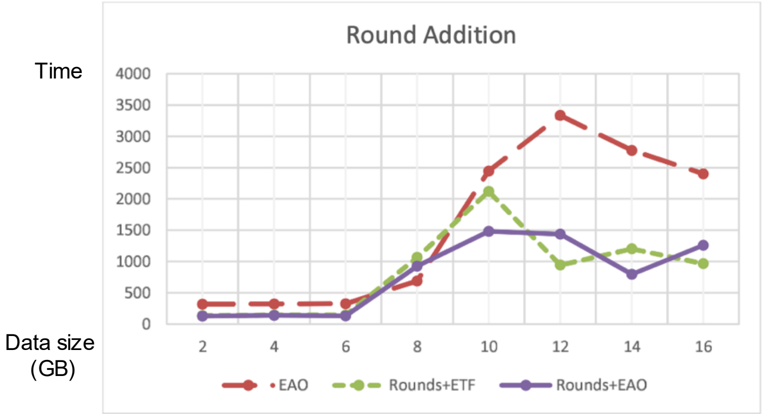
NOAM Implementation

- **ONDM** (**Object-NoSQL Datastore Mapper**) is a framework that provides application developers with:
  - a uniform access towards a variety of NoSQL datastores
  - the ability to map application data to different data representations, in a flexible way
- Main features of ONDM
  - object-oriented API, based on Java Persistence API (JPA)
  - transparent access to various NoSQL datastores – such as Oracle NoSQL, Redis, MongoDB, CouchBase, and Cassandra

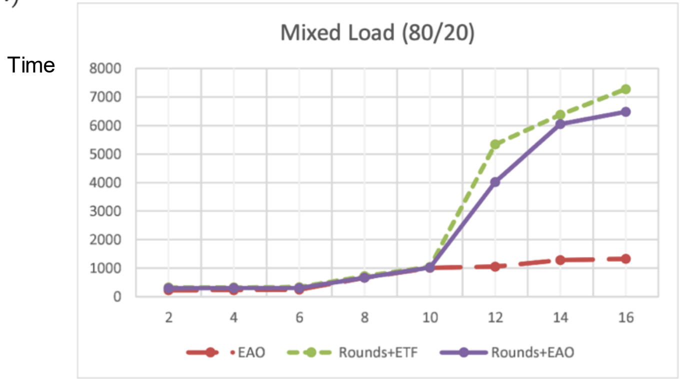
NOAM Experiments



(a) Game Retrieval (*msec per op.*)



(b) Round Addition (*msec per op.*)



Data size (GB)

(c) Mixed Load (*msec per op.*)

NOAM Conclusions

- NOAM: First abstract data model for NoSQL databases

- Aggregate partitioning has an impact on the performance of the various operations:
  - In general, when using a NoSQL database, decisions on the organization of data are required
  - These decisions are significant, as the data representation affects major quality requirements, such as scalability, performance, and consistency

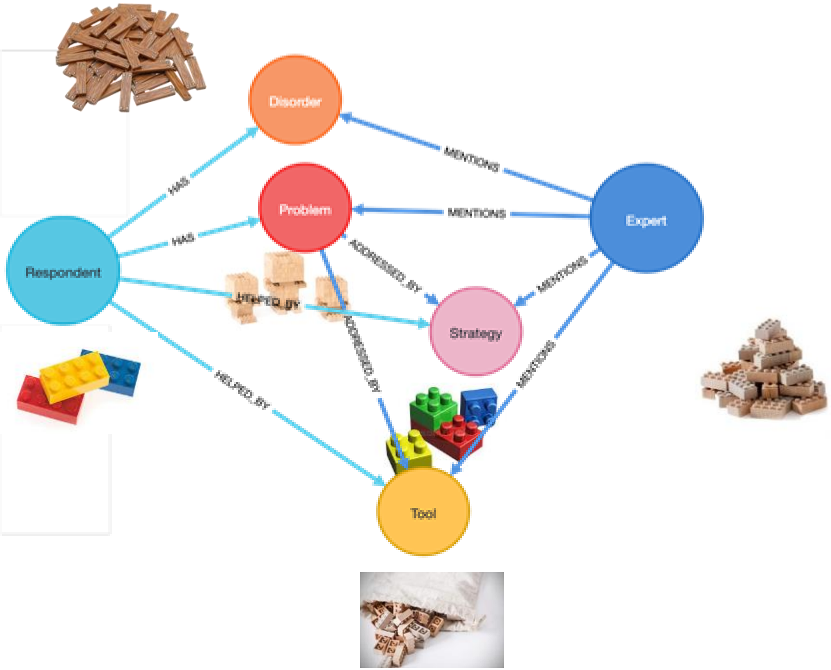
Main research areas and contributions

- Metamodel data integration
  - **Papers**: Linked Data Management 2022, DEXA 2020, ER 2018, CIDR 2015, ER 2014, EDBT 2013
- **Graph Data Integration and Large Language Models**
- Data preparation and analysis for Time Series in the Energy Domain

Graph data integration and LLM

- **Collaborators**: Karim El Hage, Yasmina Hobeika, Victor Hong, Ruining Ma, Adel Remadi, Salahidine Lemaiko, Bernard Quinio, Antoine Hafouche
- **Projects**: Vrailexia
- **Papers**: BigData 2023, J. Glob. Inf. Manag 2023, DKE 2024,

Graph data integration and LLM



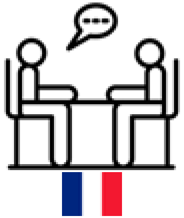
Graph data integration and LLM

104 columns

Mixture of data types

Incorrect /Missing info

Subjective



BQ : Donc comme tu le vois, j'ai démarré l'enregistrement

Alors voilà donc première chose, est ce que tu peux me dire un peu qui tu es ce que tu fais en ce moment et surtout en rapport avec la dyslexie, c'est à dire quels sont tes voilà tes contacts et sur ce sujet-là.

EXP3 : Je suis XX, je suis docteur en neurosciences. Actuellement post-doc au laps, un labo qui s'occupe du développement de l'enfant.

Et je me suis spécialisée dans l'apprentissage scolaire. Enfin, apprentissage scolaire principalement au départ, la lecture. Et maintenant je me spécialise plutôt dans les mathématiques.

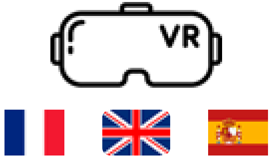
Bon, j'ai commencé mes études sur la lecture et aussi, je me suis intéressée sur certains projets sur la dyslexie, mais maintenant c'est vrai que je me suis plutôt spécialisée dans tout ce qui est apprentissage des mathématiques donc apprentissage simple mais aussi dyscalculie.

BQ : Alors donc, maintenant première question, comme je te l'ai dit, c'est vraiment général, donc sens toi libre de d'aller où tu veux dans tes réponses.

Quelles sont les caractéristiques des apprenants, dyslexiques ou plus généralement des Dys, selon ce que tu as envie de dire, que l'on doit prendre en compte pour les aider dans leur apprentissage.

EXP3 : Alors là caractéristique la plus importante pour moi, c'est de leur laisser le temps.

Ils ont souvent, ils ont une difficulté à lire et puis donc pour les dyscalculique, c'est une difficulté à comprendre les chiffres et que ça leur demande un temps plus long pour faire le quelque chose qui est automatique pour nous et ils utilisent souvent des stratégies de compensation et donc cette stratégie de compensation, bien qu'elle soit utile cela prend souvent plus de temps que nous. Nous on prend l'autoroute de A à B et eux prennent les petites routes. Du coup, même s'ils vont arriver au point B, ça leur demande plus de temps. Et aussi plus de capacités cognitives. C'est un temps aussi qui doit être calme, donc ce qu'on leur donne et si on leur autorise plus de temps à lire ou plus de temps à comprendre une équation mathématique, ce temps a besoin d'être calme et être un temps dédié à ça.



Two Different Tests

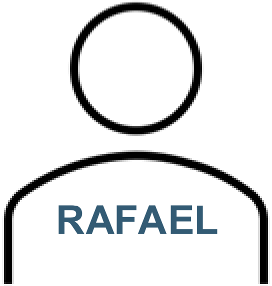
Mixture of data types

Semi-Objective

Graph data integration and LLM

Dyslexia affects 5-17% of the Population

Effort for educators to realize benefits of evidence-based intervention



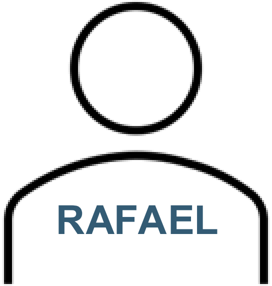
No Access to Information to prepare for Higher Education

Context & Motivation

PROBLEMS
- TEXT COMPREHENSION
- FOCUS ONLINE CLASSES
- TAKING EXAMS

DISORDERS
- DYSLEXIA
- DYSCALCULIA
- DYSORTHOGRAPHIA

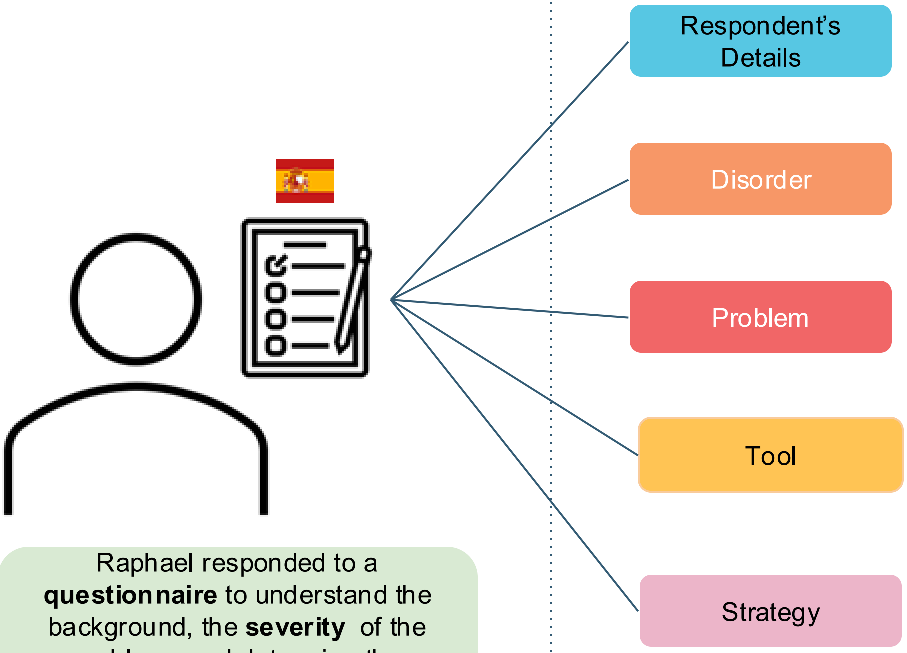
BACKGROUND
- FINISHED HIGH SCHOOL
- REPEATED A GRADE
- DIAGNOSED WHEN TEEN



TOOLS
- AUDIOBOOKS
- CLASS SLIDES
- NOISE HEADPHONES

STRATEGIES
- MAKE CONCEPT MAPS
- DIFFERENT COLOR WORDS
- RECORD LESSON

Walkthrough of Solution



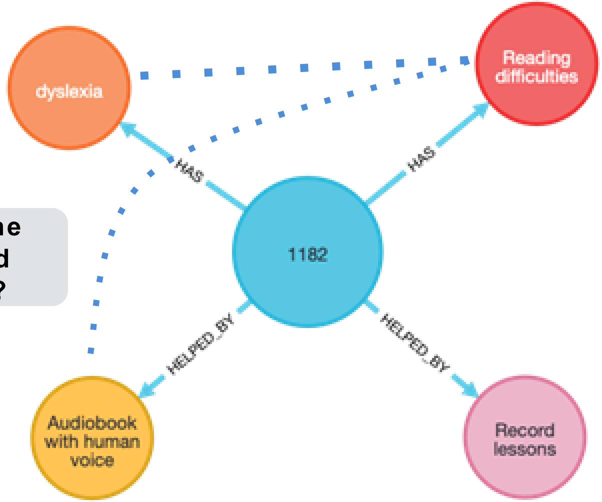
Raphael responded to a **questionnaire** to understand the background, the **severity** of the problems and determine the **usefulness** of some tools and strategies

Is this star-shaped representation enabling us to **draw conclusions**?

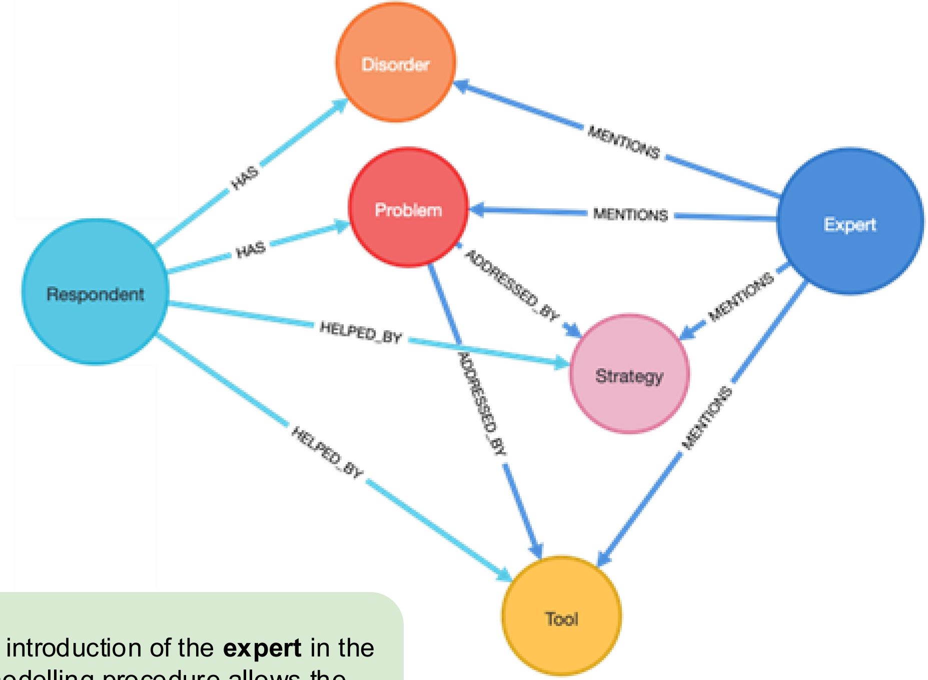
Questions can be categorized and **modelled** into entities, which can then be related using a **graph implemented** on Neo4j

Walkthrough of Solution

A French neuroscientist, Amélie, was **interviewed** and found in her study and research that dyslexia caused a **difficulty to read**: an issue which could be alleviated by **audio recordings of text**



Why the dotted lines?



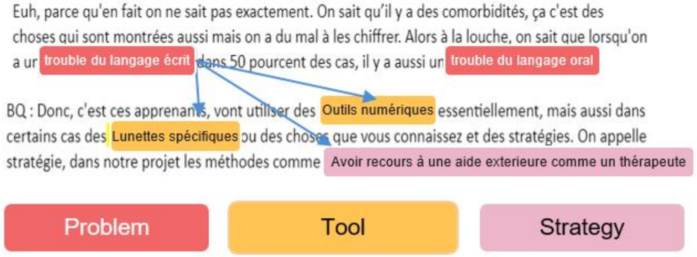
The introduction of the **expert** in the modelling procedure allows the creation of **causality relationships** between the different entities

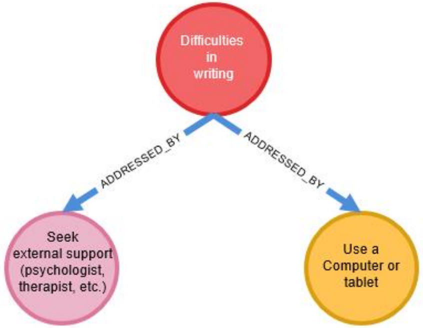
Name Entity Recognition problem

STEP 1
Raw text chunking

STEP 2
Node and Relationship extraction

STEP 3
Integration into database





Inefficient, not precise…

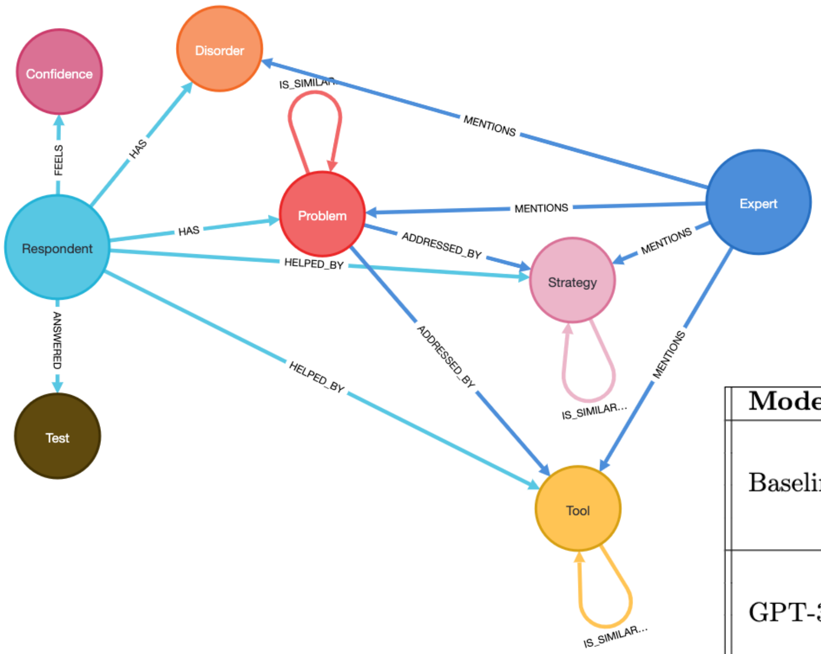
So How can we create Value with Such Graph Database?

- Use LLM and the first version of GPT to align entities?

Scaling the example, how can these recommendation actually be presented to students in a meaningful manner?

A personalized list of **Tools** and **Strategies** can be recommended to address the most severe **Problems** of Dyslexic students through **graph navigation of the different sources**

GPT and LLMs in action



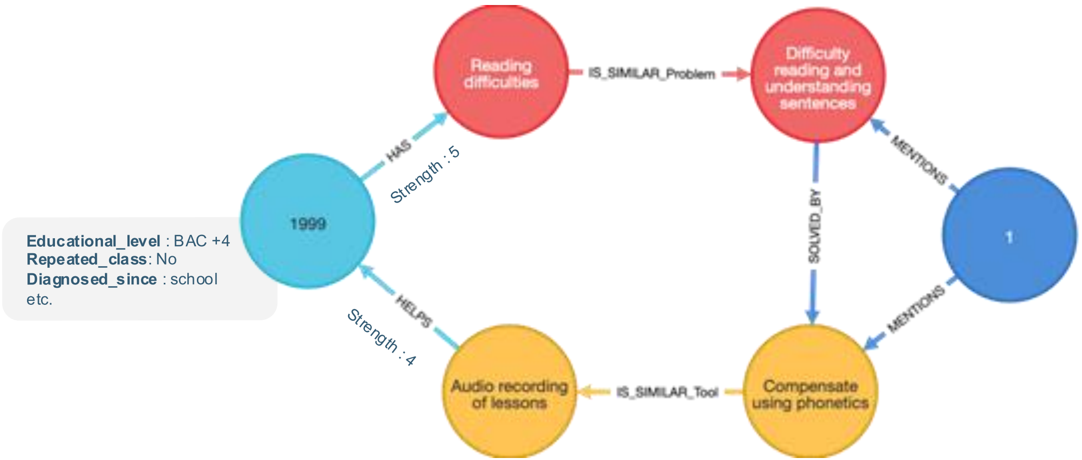
Entity	Recall	Precision	F1-score
All	75.41	69.84	72.49
Nodes	89.84	75.11	81.80
Relationships	52.87	58.64	55.34

Model	**Node Type**	**Precision**	**Recall**	**F1-score**
Baseline: Clustered Embeddings	Problem	**91.67**	29.72	44.40
	Tool	36.00	24.32	23.03
	Strategy	**85.71**	16.22	27.27
	Total	59.09	23.42	33.54
GPT-3.5-Turbo	Problem	63.83	**81.08**	**71.42**
	Tool	**63.33**	**51.35**	**56.72**
	Strategy	80.77	**56.76**	**66.67**
	Total	**67.96**	**63.06**	**65.42**

Hybrid: Recommending through Experts, Ranking through Respondents

- The recommender system leverages the graph database's insights and a ranking model to predict customized suggestions of learning strategies and tools

Graph database insights



Identify the most significant problems → Use the experts' causal explanations → Identify relevant strategies and tools →

Ranking algorithm

Extract features from database ↓ Neural Network model ↓ Rank strategies and tools by their usefulness score ↓

Generate customized recommendations

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
CentraleSupélec

The Result

- Three most severe problems of one respondent and is recommended specific tools by experts to address each problem
- Display the Top 5 recommended tools, some problems have not yet been recommended more than even 1 tool!
- More Experts → More Refined Results

Most Severe Problems	**Recommended Tools**
Reading Difficulties	1) Use a special font for easy reading 2) Use Audio Books 3) Numerical tutor (e.g., Siri) to which it is possible to query verbal explanations on challenging concepts 4) Words written in different colors
Difficulties to focus during online courses	1) A clearer presentation of the study material
Difficulties to understand complex or rare words	1) Register courses 2) Underline text with different colors 3) Conceptual sketches made by oneself 4) Repeat the studied contents 5) Summaries prepared by oneself

Conclusions and perspectives

- Work on better modelling of the the different entities and relationships
  - usability, efficiency, scalability, and flexibility
- Better explore the integration of different structures and
  - Complexity, heterogeneity
- Rank tools and strategies

Conclusions and New perspectives

Fondation université PARIS-SACLAY

- Financed project to continue exploring the research
- Automate the integration of the data collection process using the graph databases
- Use LLMs to conduct NER of transcripts
- Weight the similarity between entities coming from different sources
- Align the Entities also with the connections

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

CentraleSupélec

Main research areas and contributions

- Metamodel data integration
  - **Papers**: Linked Data Management 2022, DEXA 2020, ER 2018, CIDR 2015, ER 2014, EDBT 2013
- Graph Data Integration and Large Language Models
  - **Papers:** BigData 2023, J. Glob. Inf. Manag 2023, DKE 2024
  - **New financed project**
- **Data preparation and analysis for Time Series in the Energy Domain**

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

CentraleSupélec

Data preparation and analysis for Time Series in the Energy Domain

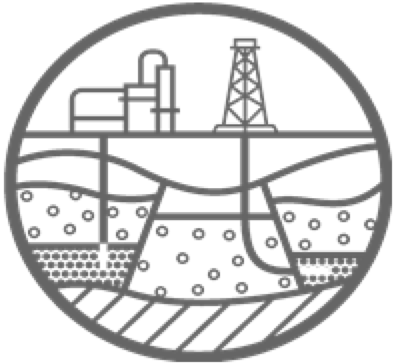
GENVIA

cea

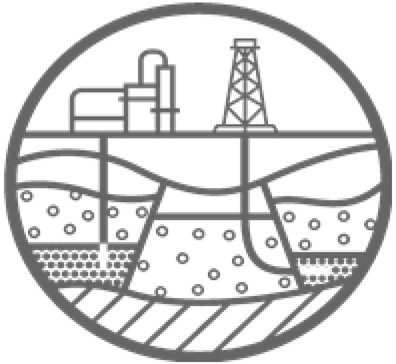
slb

- **Collaborators**: **PhD Molood Arman**, Yutao Chen, René Gómez Londoño, Sohaib Ouzineb, **PhD student Shwetha Salimath,** Nacéra Seghouani, Sylvain Wlodarczyk
- **Projects:** Proclaim, GeoTS
- **Papers:** CAiSE Forum 2020, DS 2022, KDD 2025, ADBIS 2025

Data preparation and analysis for Time Series in the Energy Domain



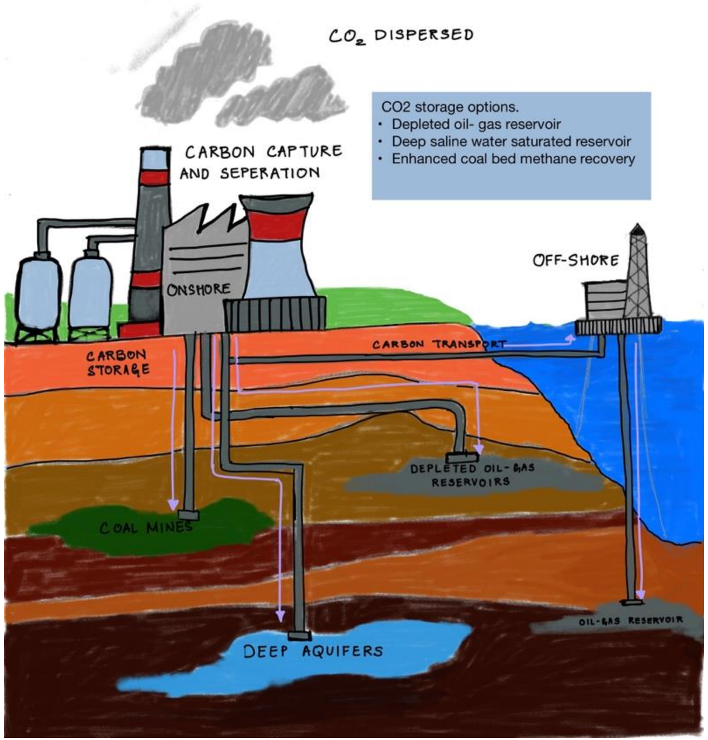
Data preparation and analysis for Time Series in the Energy Domain



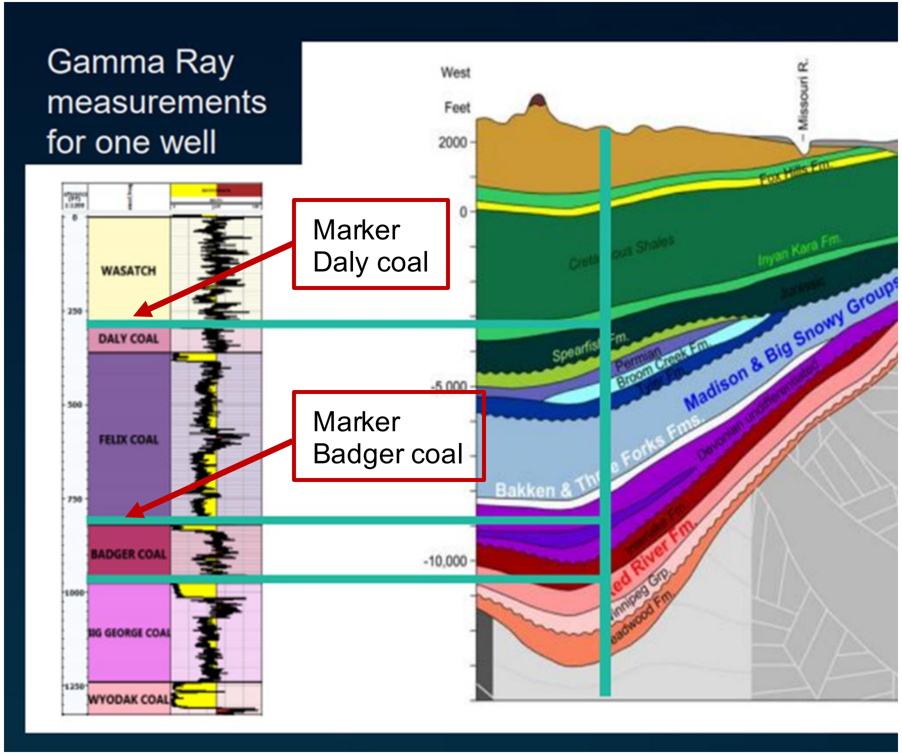
id	eventTime	latitude	longitude	eventType
672	8/27/2018 14:44	43.6392	-79.3818	9
711	8/27/2018 21:18	43.6408	-79.3956	8
732	8/27/2018 23:53	43.6408	-79.3956	6
733	8/27/2018 23:57	43.6408	-79.3956	7
736	8/28/2018 0:20	43.6408	-79.3956	9
737	8/28/2018 0:20	43.6408	-79.3956	8
762	8/28/2018 3:36	43.6408	-79.3956	9
763	8/28/2018 3:38	43.6408	-79.3956	8
764	8/28/2018 3:43	43.6408	-79.3956	8
765	8/28/2018 3:47	43.6408	-79.3956	8

Research problem: Carbon Capture Storage

- CCS involves capturing CO2, transporting it, and storing it in deep geological formations to prevent it from entering the atmosphere
- Reassessment of seal integration and storage potential
- Geological analysis and monitoring by studying subsurface rock properties and correlating formations for accurate reservoir modeling

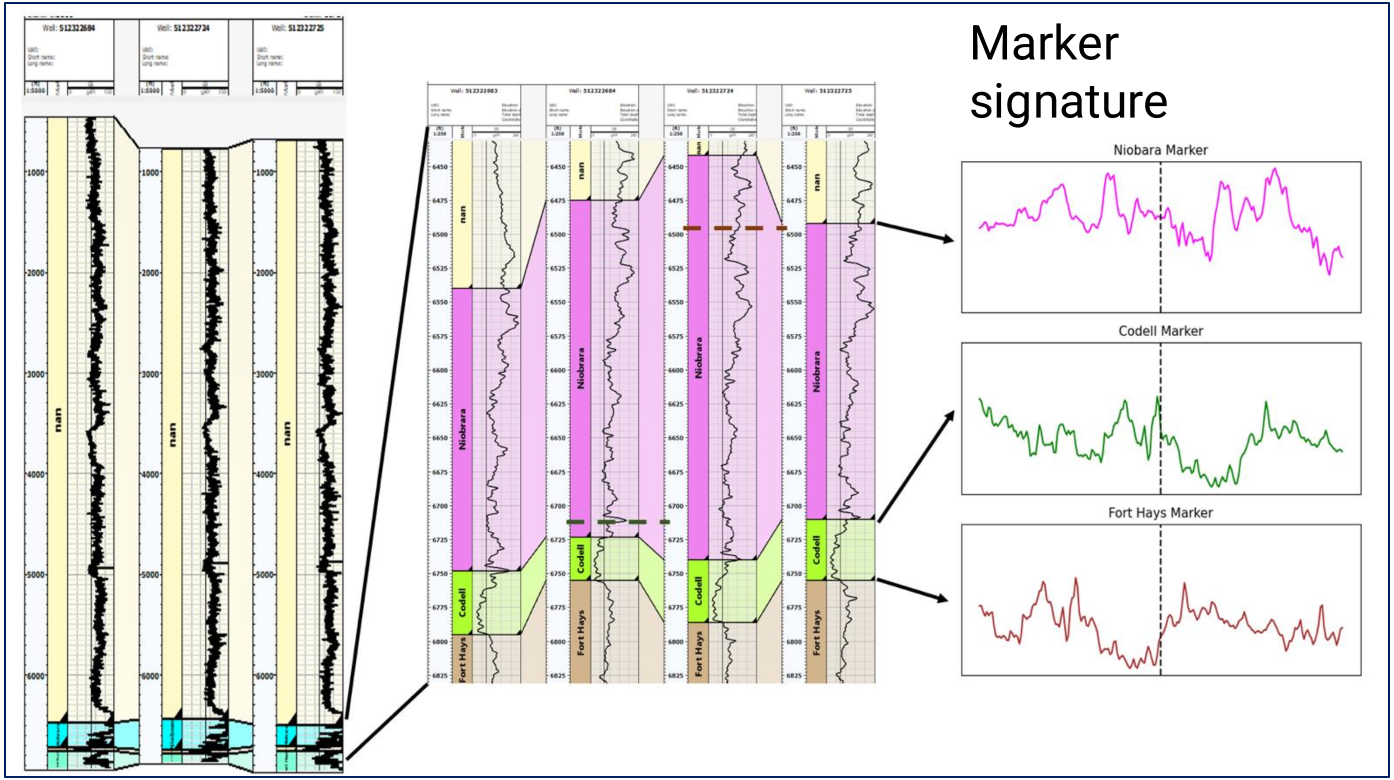


Problem Statement

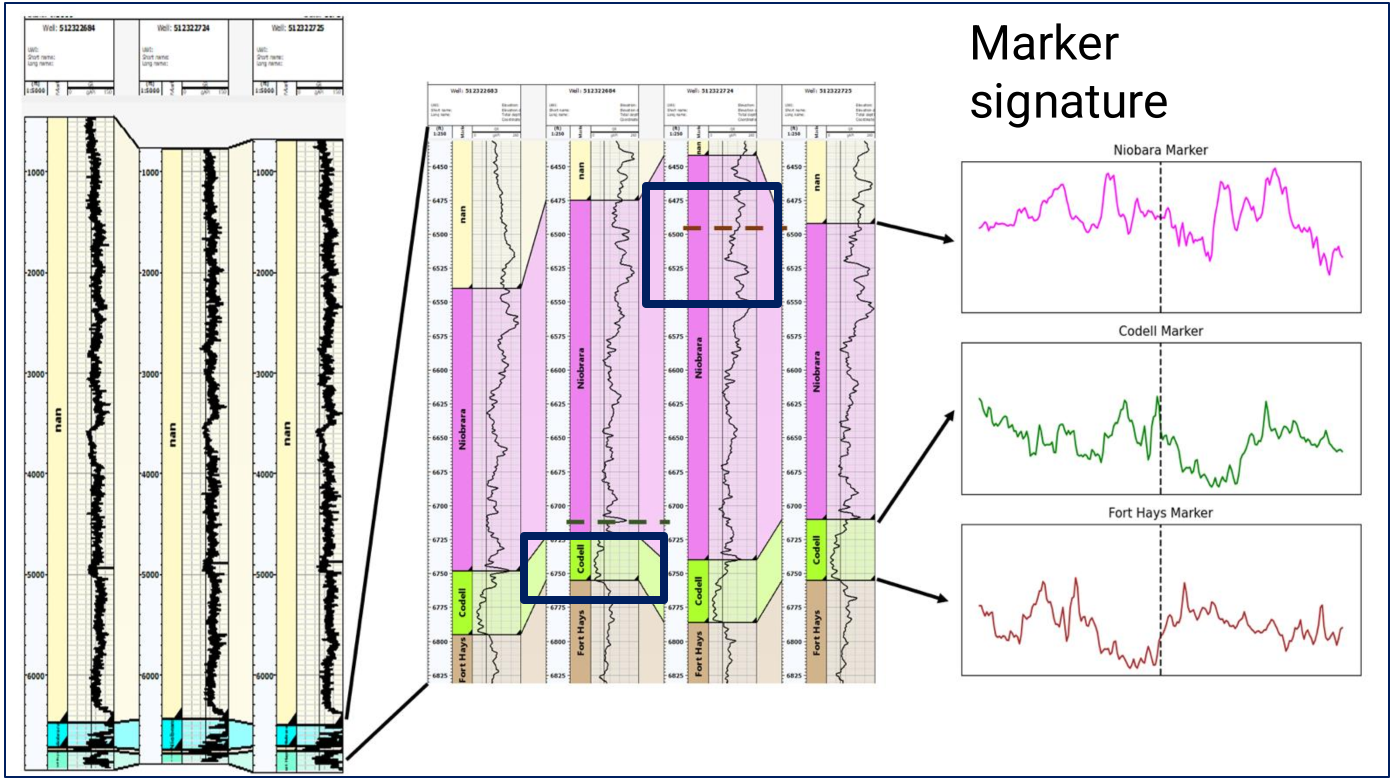


- Geologists use mud logs and the rocks extracted during borehole drilling to study formation characteristics
- Tedious and time-consuming
- Finding an efficient way to extract information from wireline logs using deep learning would save time and resources

Well log Data - a lot and heterogeneous time series



Well log Data - a lot and heterogeneous time series



Problem

- **Well Correlation**
    - Industrial baseline with dynamic time warping distance (DTW).
    - Minimum spanning tree to find pairs and then DTW.
    - Autoencoders and bidirectional LSTM for correlating neighboring wells.

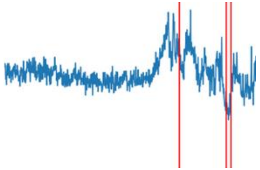
Challenges with DTW for well log data

- Bad alignment of the wells and local shifts in marker signatures
- Depth incoherent signature pattern
- Each marker prediction is independent of the other
- Since only one marker can be processed at a time, it is a time-consuming process

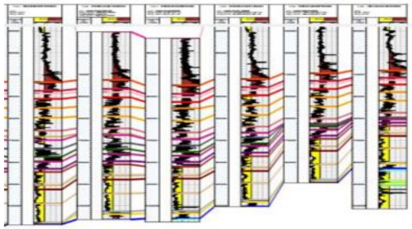
Dataset

USGS science for a changing world

WellDatabase

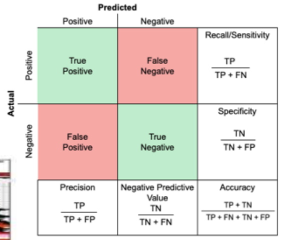


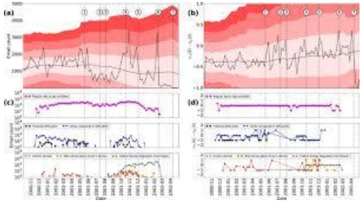
GeoTS



Evaluation

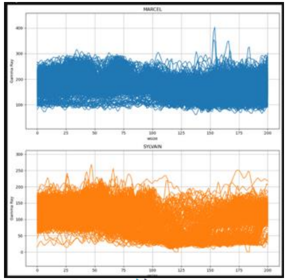
Accuracy, Recall and Visualizations

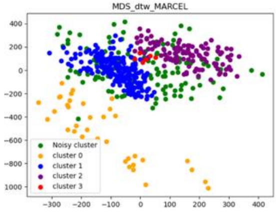




Data Processing

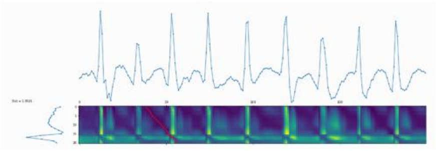
Cleaning, Clustering and Template extraction

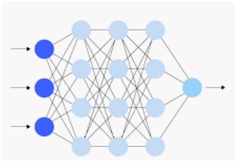




Models

Industrial baseline and deep learning models





Data Processing

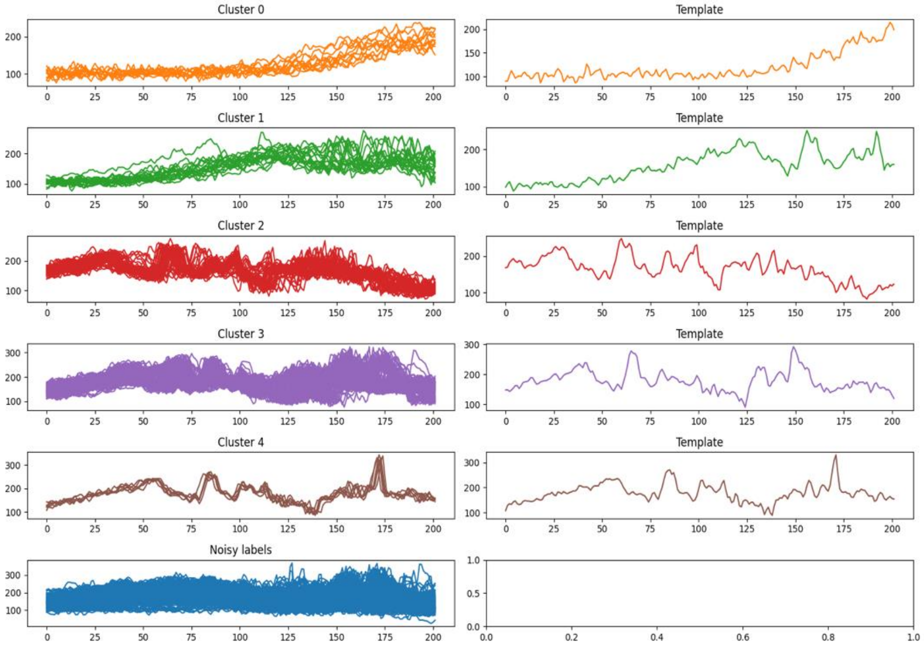
- **Signature Extraction**: This step involves extracting the signature of a formation from the training log data with a specified window size

- **Clustering**: The DTW distance matrix containing the DTW distances between all pairs of extracted signatures is used for clustering

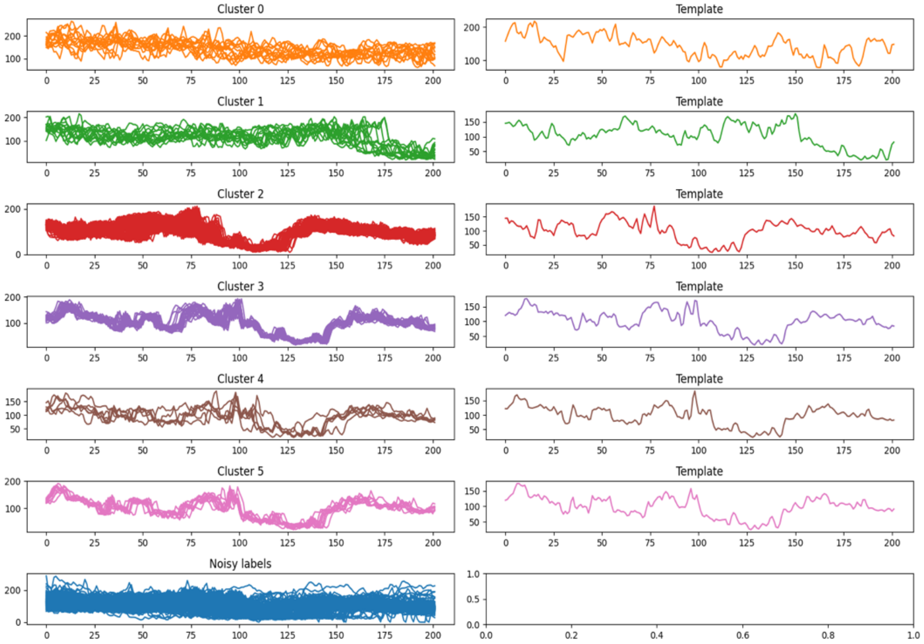
- **HDBSCAN** clustering algorithm is used. We analyze signature templates representing a cluster of similar signatures for a particular formation

Clustering result

Niobrara

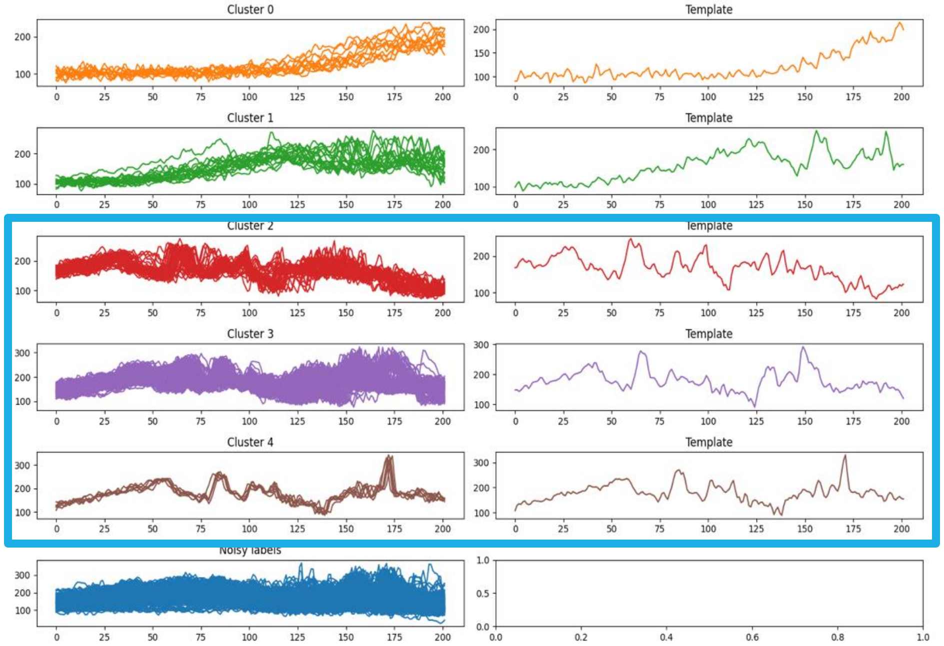


Codell

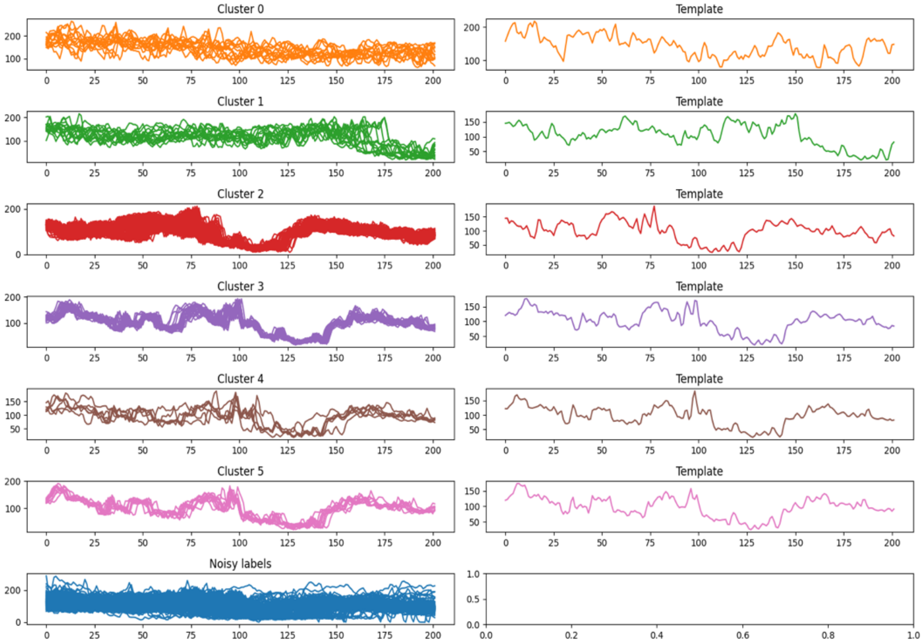


Clustering result

Niobrara

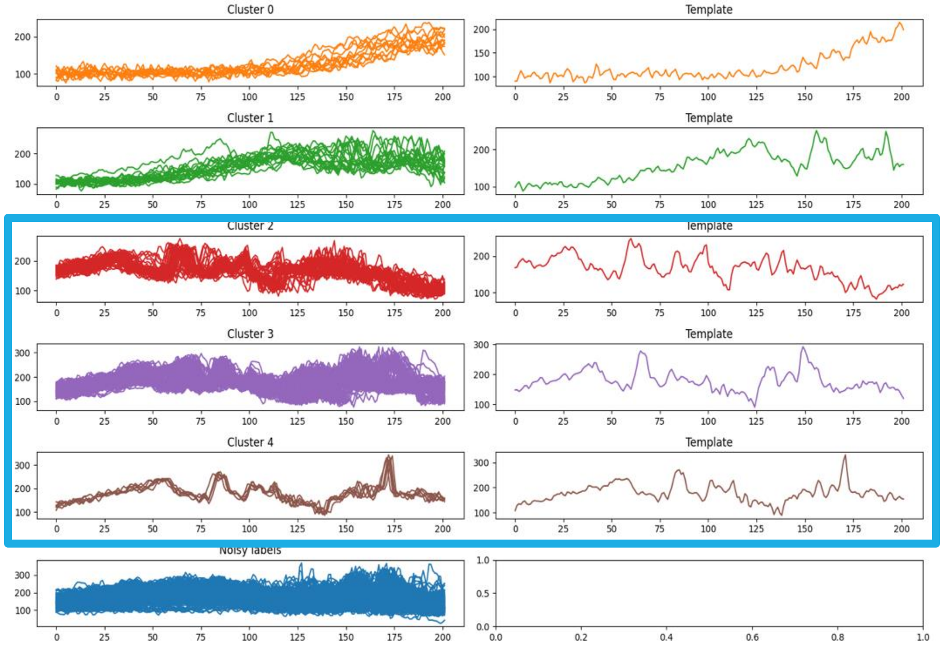


Codell

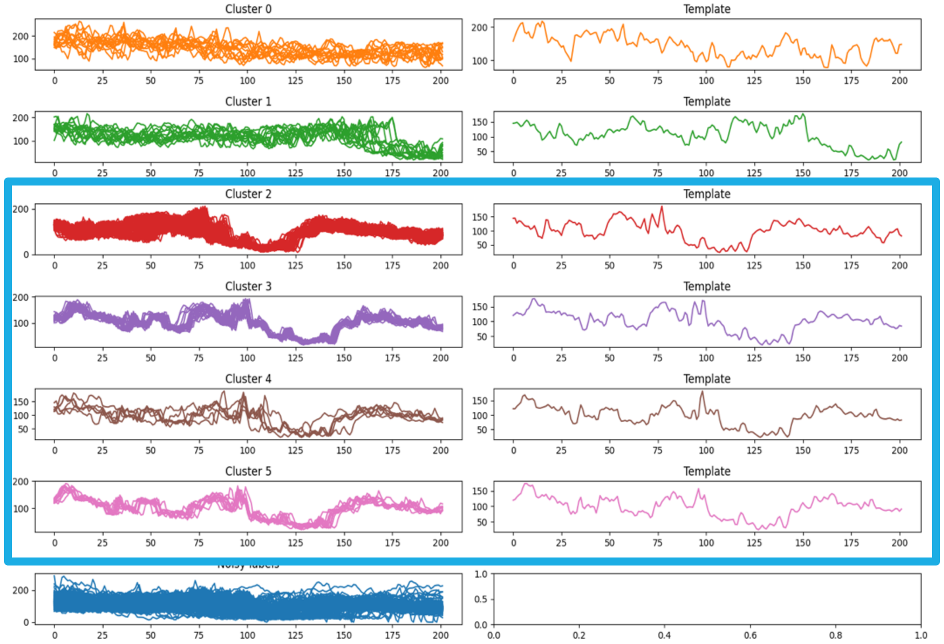


Clustering result

Niobrara

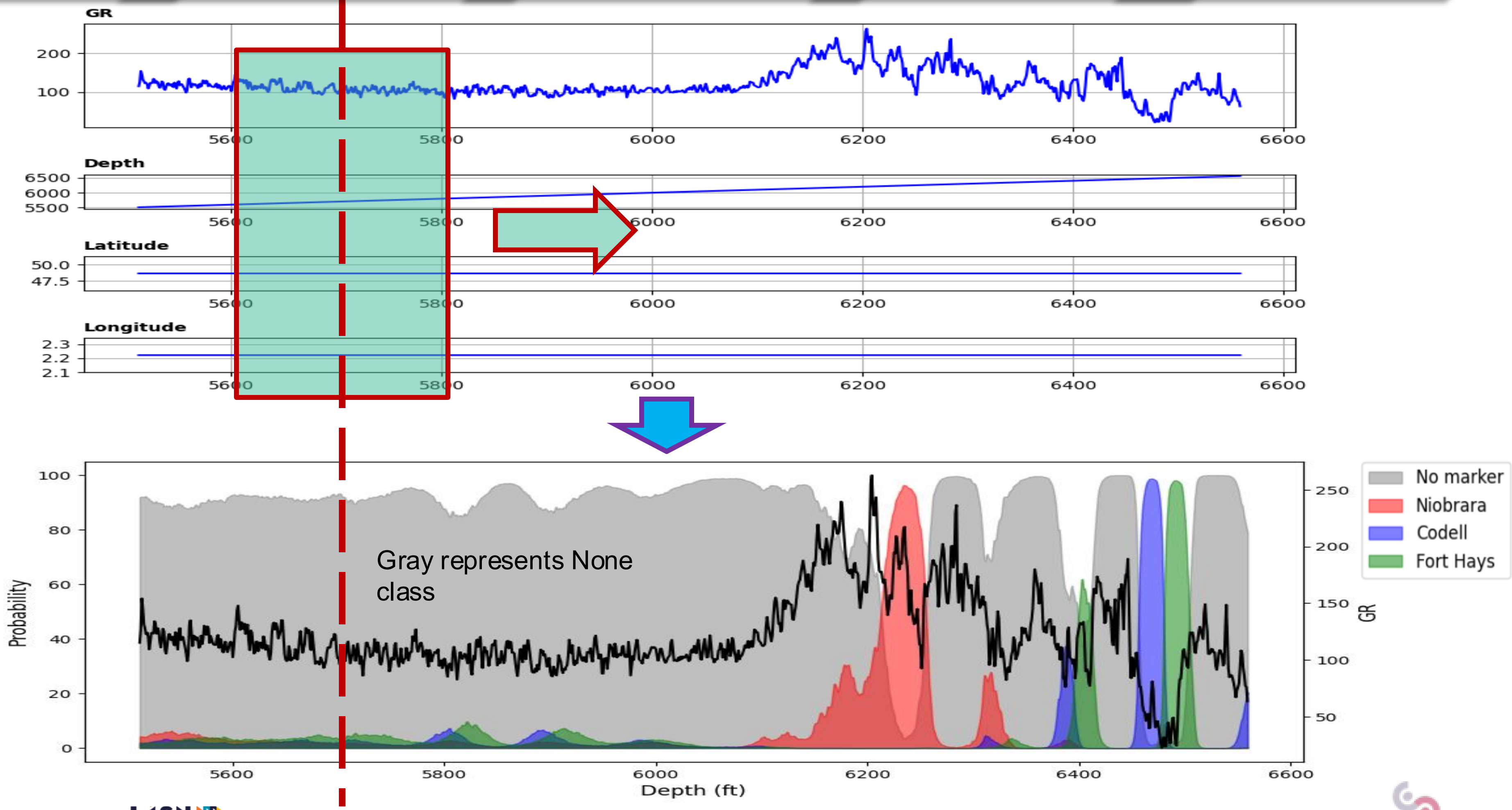


Codell



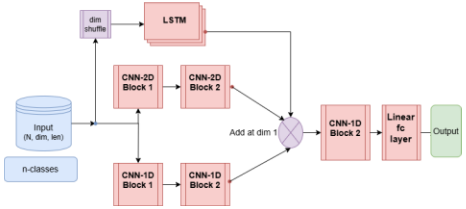
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

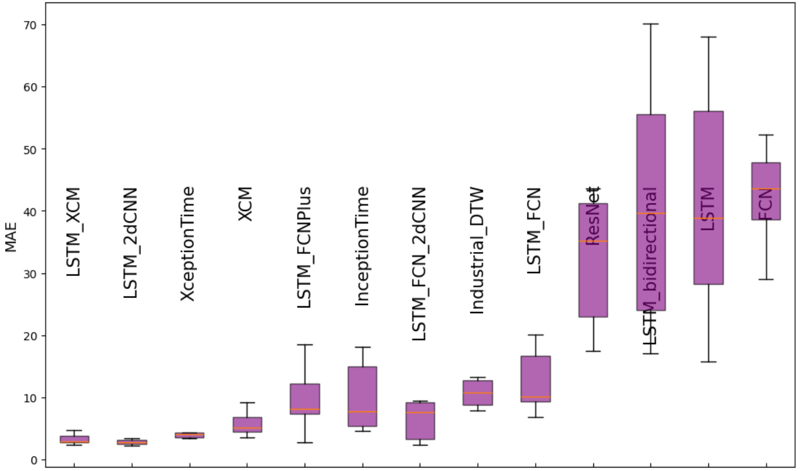
CentraleSupélec

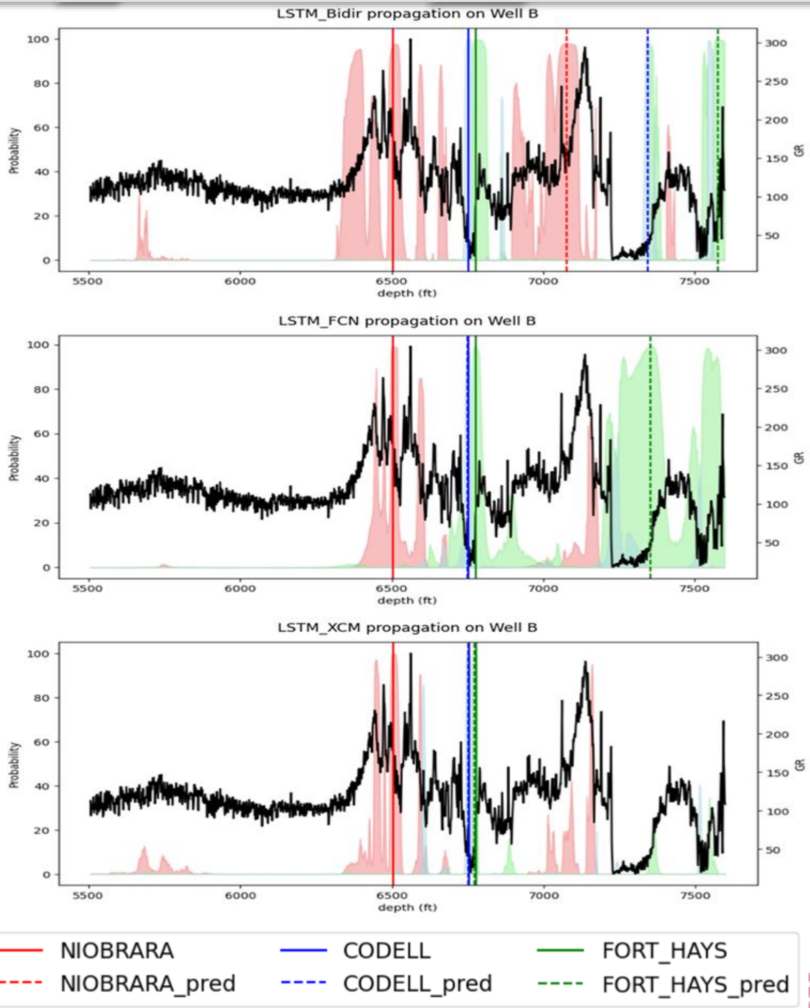

GR
Depth
Latitude
Longitude
Probability
Depth (ft)
GR
No marker
Niobrara
Codell
Fort Hays
Gray represents None class

Maximum absolute error

LSTM-XCM

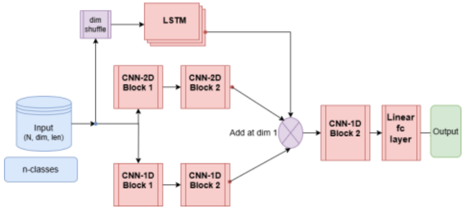


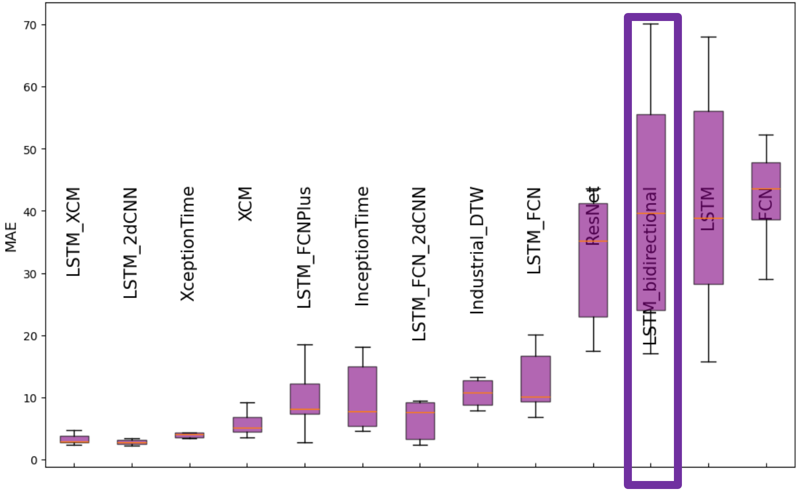


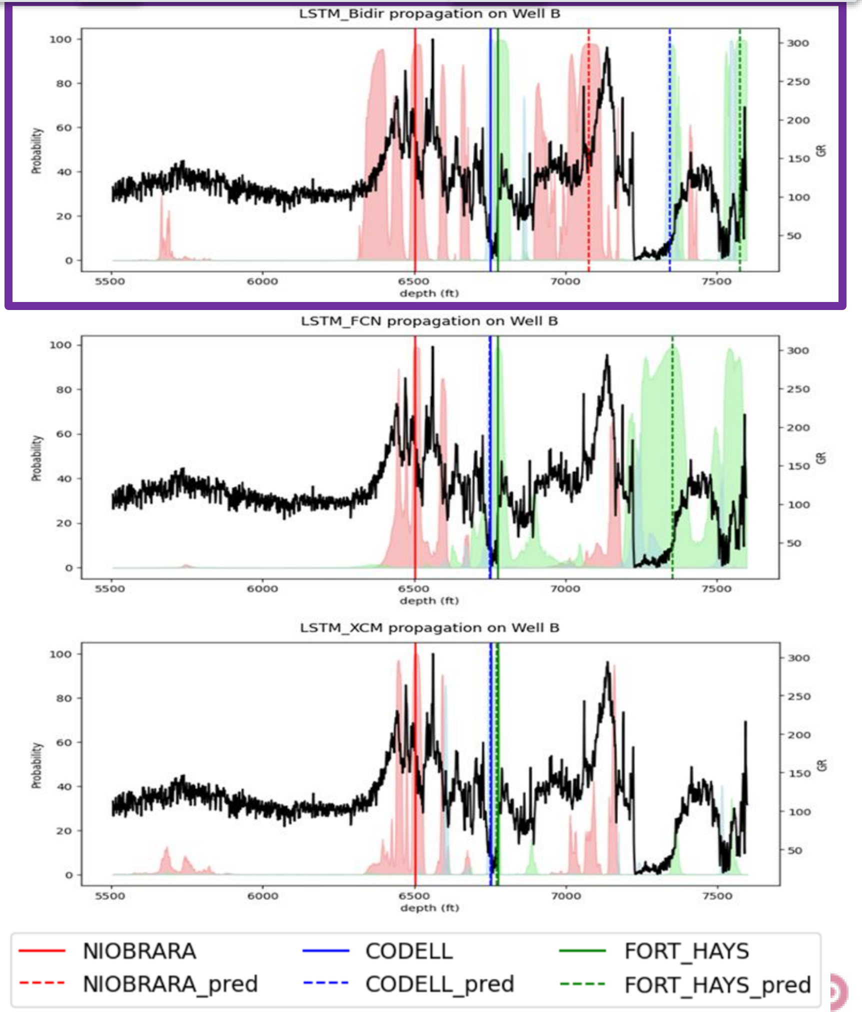


Maximum absolute error

LSTM-XCM

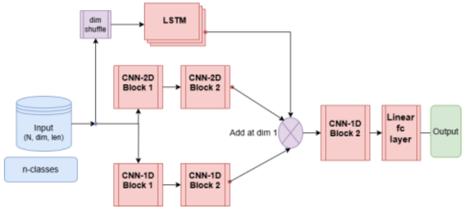


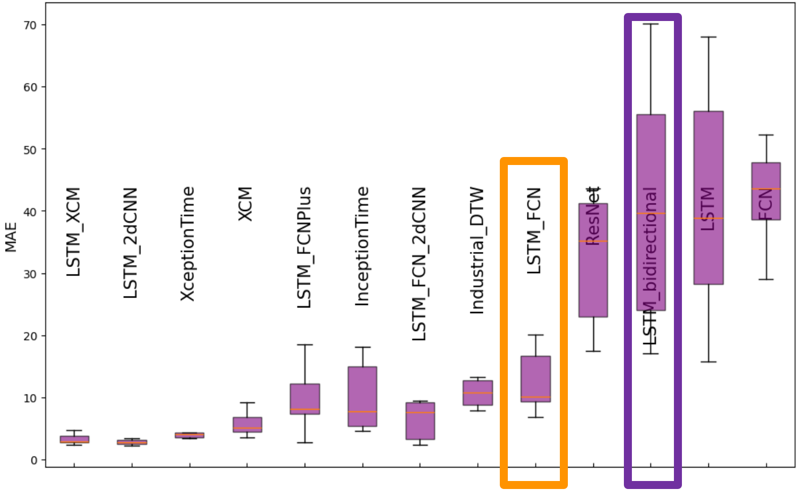


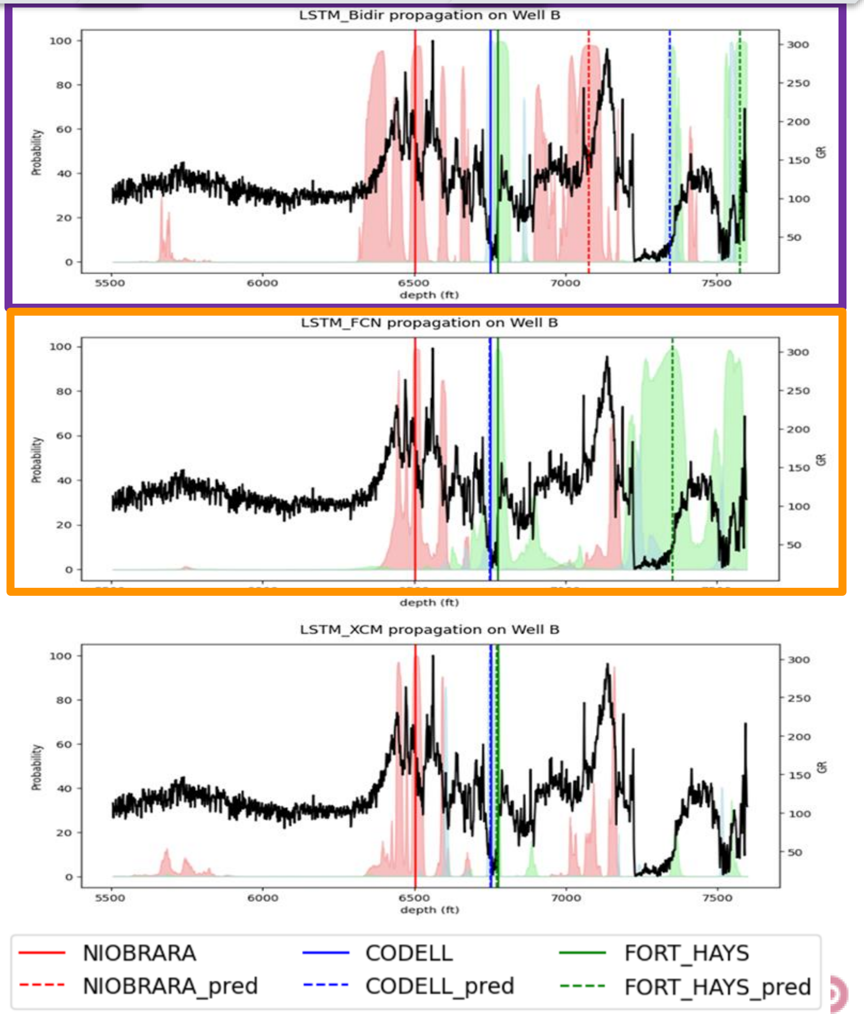


Maximum absolute error

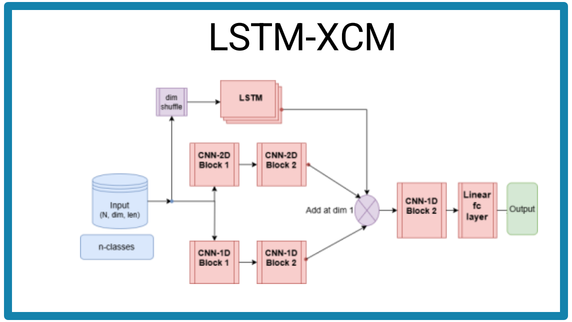
LSTM-XCM

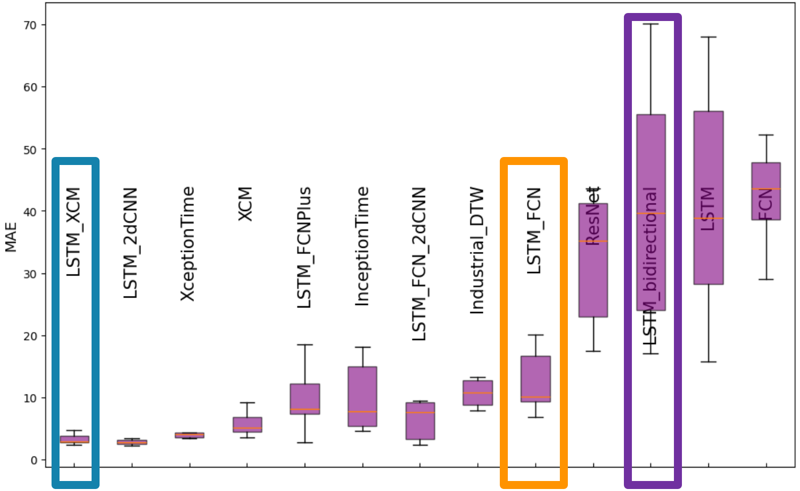


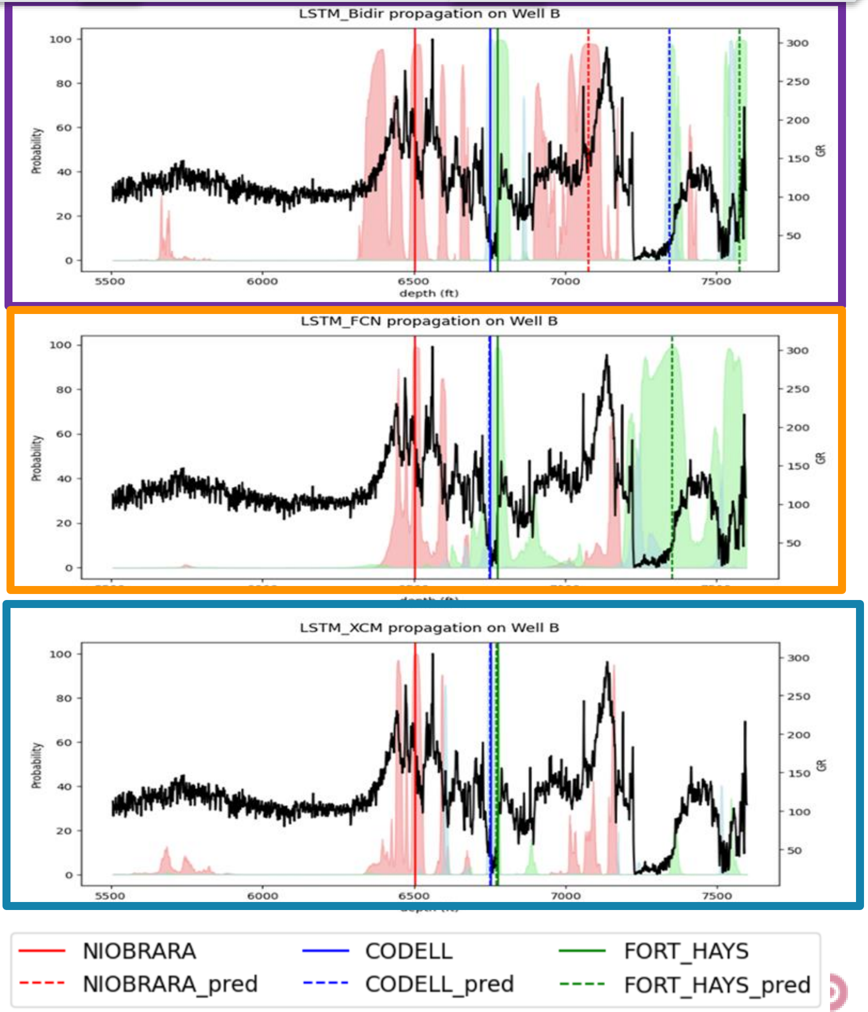




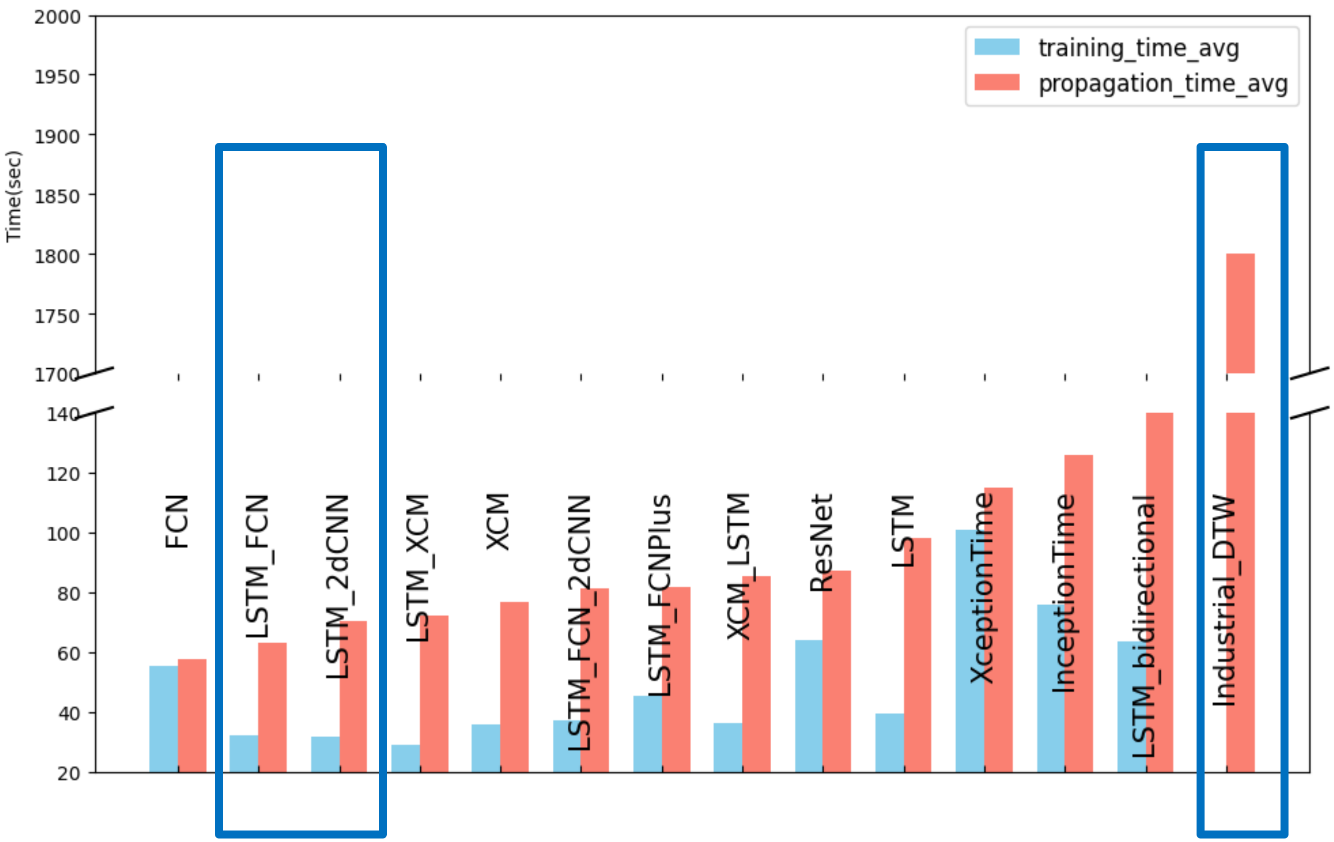
Maximum absolute error







Time Efficiency



LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

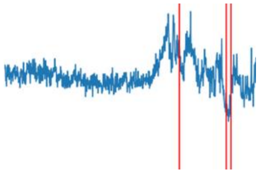
CentraleSupélec

Enhancing and enriching time series analysis

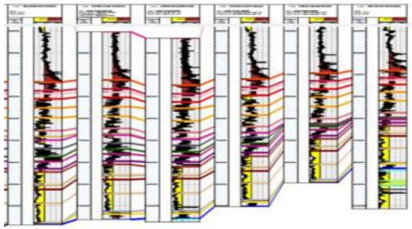
Dataset

USGS science for a changing world

WellDatabase



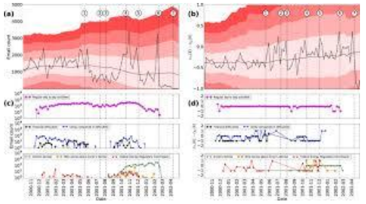
GeoTS



Evaluation

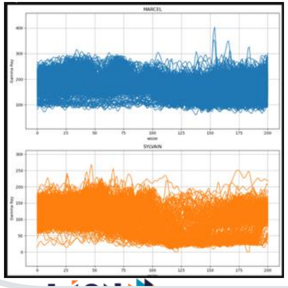
Accuracy, Recall and Visualizations

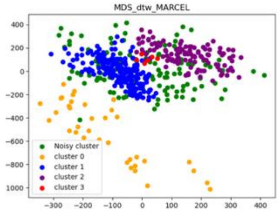
		Predicted Positive	Predicted Negative	
Actual	Positive	True Positive	False Negative	Recall/Sensitivity $\frac{TP}{TP + FN}$
	Negative	False Positive	True Negative	Specificity $\frac{TN}{TN + FP}$
		Precision $\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{TP + FN + TN + FP}$



Data Processing

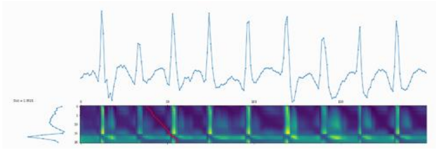
Cleaning, Clustering and Template extraction

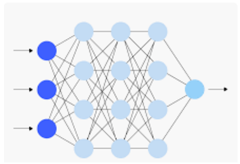




Models

Industrial baseline and deep learning models





Problem context - Reports/surveys

Experts require structure and unstructured data for theoretical data control and analysis

- **Data acquisition**
  - Wells drilled a long time ago with historical log data
  - Different tools/sensors from different service providers
  - Well sample analysis described in reports
- **Data assessment**
  - Data quality and Interpretation done manually by petrophysicists/geologists based on reports

- Retrieval-Augmented Generation (RAG) techniques
- Automate the process by exploring agentic RAGs

2. Stratigraphy an. leoenvironment Results

2.1 Cenozoic

2.1.1 Pleistocene to Pliocene
850 to 970 feet (thickness more than 120 feet)

No samples were available from the interval between sea bottom and 850 feet.

Paleontology

The benthonic foraminiferal assemblage contains mainly species which are at present still living; typical Pliocene forms are nearly absent (only single specimens of Cassidulina cf. pliocarinata and Cibicides lobatulus grossa were found, which could be reworked). This microfauna suggests a Pleistocene or uppermost Pliocene age for these deposits.

Paleoenvironment

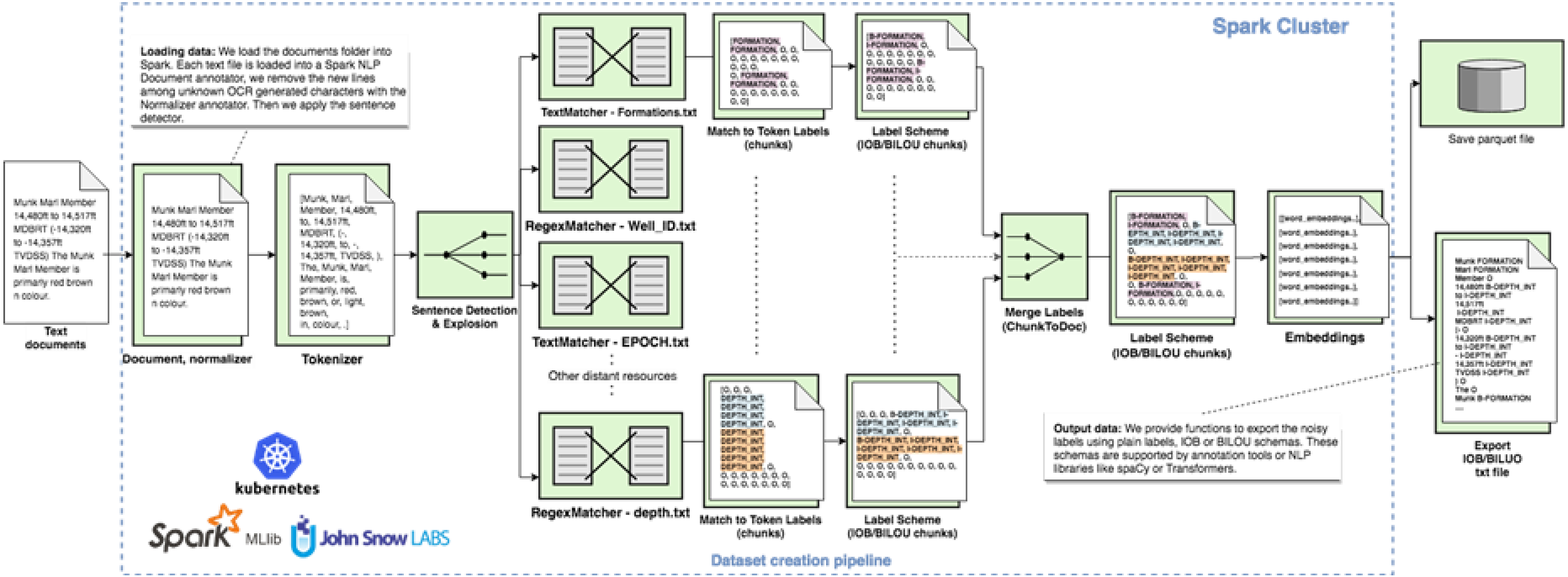
The benthonic foraminiferal assemblage, the near absence of planktonic foraminifera and the occurrence of frequent shell fragments suggest shallow marine (inner neritic) environment.

```
In [33]: data
```

Out[33]:

	Area Abbreviation	Area Code	Area	Item Code	Item	Element Code	Element	Unit	latitude	longitude	...	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009
0	AF	2	Afghanistan	2511	Wheat and products	5142	Food	1000 tonnes	33.94	67.71	...	3249.0	3486.0	3704.0	4164.0	4252.0	4538.0
1	AF	2	Afghanistan	2805	Rice (Milled Equivalent)	5142	Food	1000 tonnes	33.94	67.71	...	419.0	445.0	546.0	455.0	490.0	415.0
2	AF	2	Afghanistan	2513	Barley and products	5521	Feed	1000 tonnes	33.94	67.71	...	58.0	236.0	262.0	263.0	230.0	379.0
3	AF	2	Afghanistan	2513	Barley and products	5142	Food	1000 tonnes	33.94	67.71	...	185.0	43.0	44.0	48.0	62.0	55.0
4	AF	2	Afghanistan	2514	Maize and products	5521	Feed	1000 tonnes	33.94	67.71	...	120.0	208.0	233.0	249.0	247.0	195.0
5	AF	2	Afghanistan	2514	Maize and products	5142	Food	1000 tonnes	33.94	67.71	...	231.0	67.0	82.0	67.0	69.0	71.0
6	AF	2	Afghanistan	2517	Millet and products	5142	Food	1000 tonnes	33.94	67.71	...	15.0	21.0	11.0	19.0	21.0	18.0
7	AF	2	Afghanistan	2520	Cereals, Other	5142	Food	1000 tonnes	33.94	67.71	...	2.0	1.0	1.0	0.0	0.0	0.0
8	AF	2	Afghanistan	2531	Potatoes and products	5142	Food	1000 tonnes	33.94	67.71	...	276.0	294.0	294.0	260.0	242.0	250.0
9	AF	2	Afghanistan	2536	Sugar cane	5521	Feed	1000 tonnes	33.94	67.71	...	50.0	29.0	61.0	65.0	54.0	114.0
10	AF	2	Afghanistan	2537	Sugar beet	5521	Feed	1000 tonnes	33.94	67.71	...	0.0	0.0	0.0	0.0	0.0	0.0

Dataset creation – First temptative



Enhancing and enriching time series analysis

- Reinforcement learning from human feedback (RLHF)
  - Integrating physical constraints into models

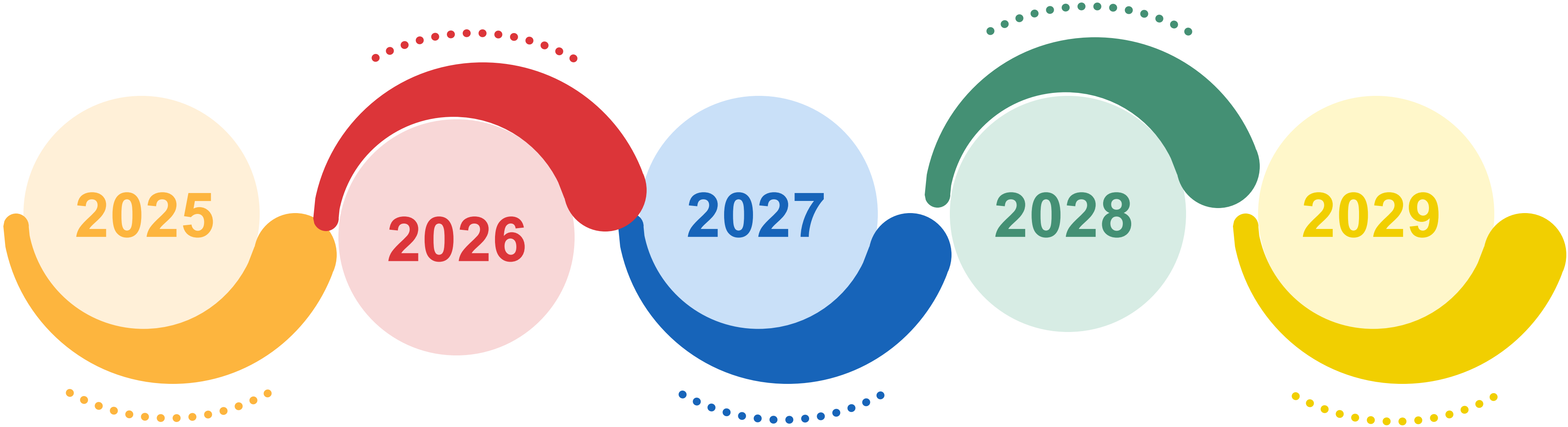
Enhancing and enriching time series analysis

- Agentic Artificial Intelligence
  - How will autonomous systems interact with data?
  - Specialized petrophysical interpretations
  - Going back also to the report interpretation
  - Integrate human continuous feedback

- CIFRE PhD thesis will start in the next months

Main areas and contributions

- Metamodel data integration
  - **Papers**: Linked Data Management 2022, DEXA 2020, ER 2018, CIDR 2015, ER 2014, EDBT 2013
- Graph Data Integration and Large Language Models
  - **Papers:** BigData 2023, J. Glob. Inf. Manag 2023, DKE 2024
  - **New financed project**
- Data preparation and analysis for Time Series in the Energy Domain
  - **Papers:** CAiSE Forum 2020, DS 2022, KDD 2025, ADBIS 2025
  - **New financed thesis to explore agentic AI**


2025
2026
2027
2028
2029

Axes

- Data modeling - DataFrames integration
- Smart cities and data integration
- **Global database for health**
- Data for Physics

Global database for health

- **Collaborators**: Paul-Henry Cournede, Jyotishka Das, Aaron Mamann
- **Project:** RHU Remission

RHU Remission

Using fresh tissues (blood and tumor)

- source of biomarkers
- adapt new immunotherapy strategies to the biology of patients and their cancer

Global database for health

Personalized Treatments

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

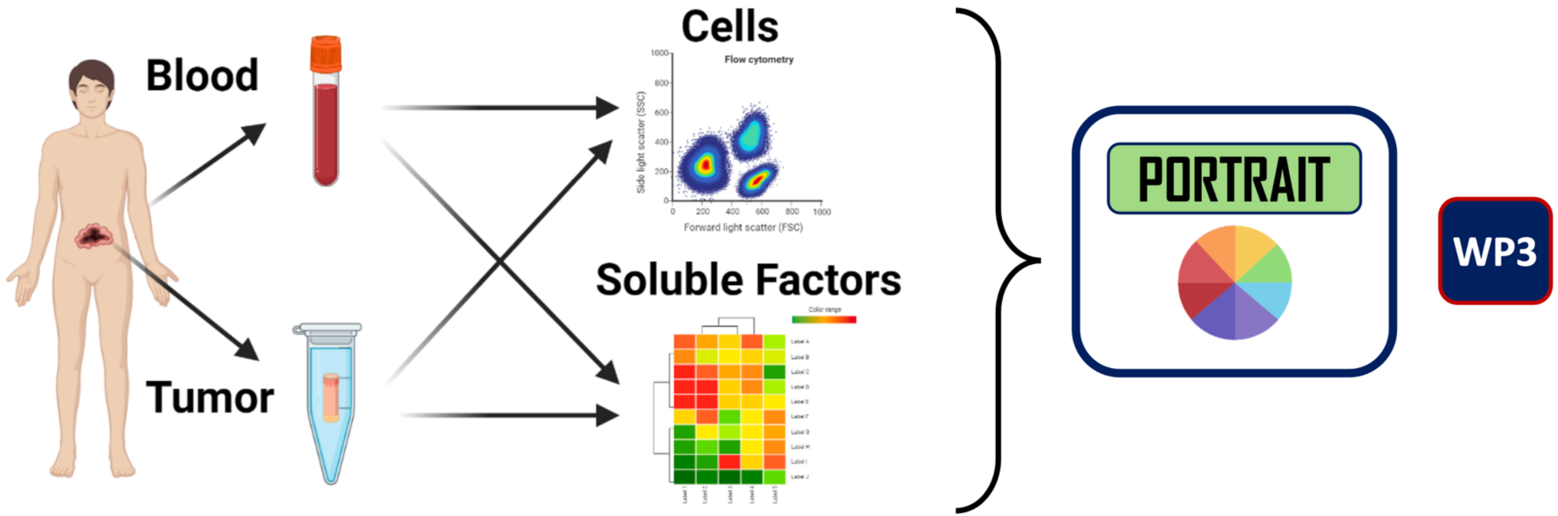
CentraleSupélec

Global database for health

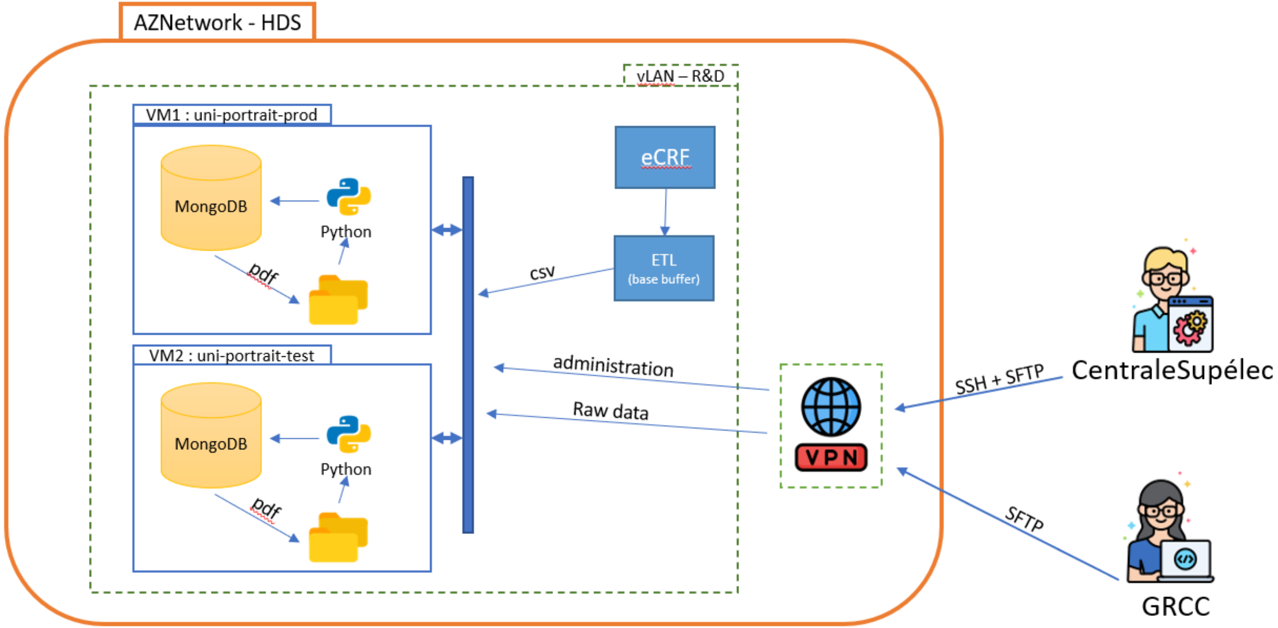
- Enhancing immunotherapy clinical trials
  - Continuing the identification of predictive biomarkers
  - Developing expertise and capabilities at Gustave Roussy in the analysis of fresh tissues
  - Building a national bioclinical research platform

- Integrating Different data

Profile in Onco-immunology for a Rapid Treatment Research Adapted to your Immunity and your Tumor

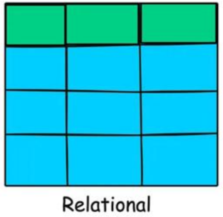


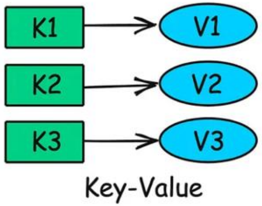
Database

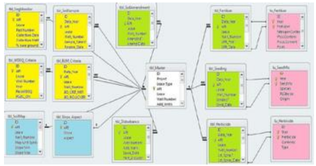


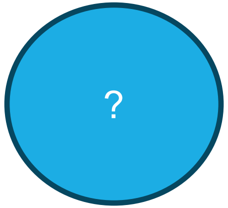
Phd Thesis will start in the next months

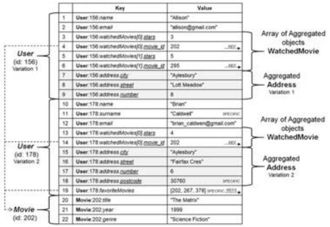
Data Integration: a perpetually evolving challenge for new research perspectives

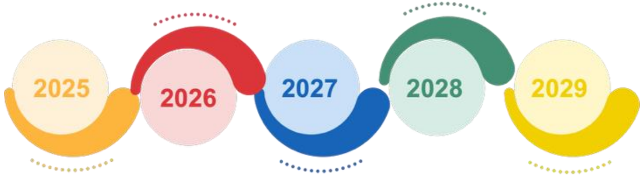












Summary

- ## Papers

11 Journals,

21 International Conferences,

3 Demonstrations,

10 National Conferences

BigData 2023, DKE 2024, J. Glob. Inf. Manag, Linked Data Management, ER 2018, DEXA 2020, CIDR 2015, ER 2014, EDBT 2013, Energies 2023, EPS-HEP 2025, CAiSE Forum 2020, DS 2022, KDD 2025, Adbis 2025, …

- ## PhD students and Postdocs

Molood Arman, Shwetha Salimath, Quentin Bruant, Jotyshka Das, Yuchen Tao, Adnan El-Moussawi 2021, Charles Ndungu-Ndegwa 2025

- ## Master Students

Andrés Gomez, Konstatinos Mira, Lin Siying, Antony Joseph, Akshay Tayde, Moditha Hewasingage, Suela Sais, Abdellah Oumida, Pallavi Katihalli-Manjegowda

- ## Projects

Vrailexia, Remission RHU, GeoTS, BMP trajectory Analyses, IT4Energies, B-Graph, Proclaim, NOAM, Estocada, SOS, MIDST, MATRIX-EXL

- ## Industry Collaborations

Genvia, SLB, Transvalor, Tissium, Vires, Dalkia, Generali, Solinum, Central Bank of Italy, Consip, IS

- ## Academic Collaborations

Roma Tre, Nanterre University, Inria, CEA, TU Berlin, University of Oulu, Nairobi University

- ## Committees

Handiversité, Dasc, Adbis, BDA

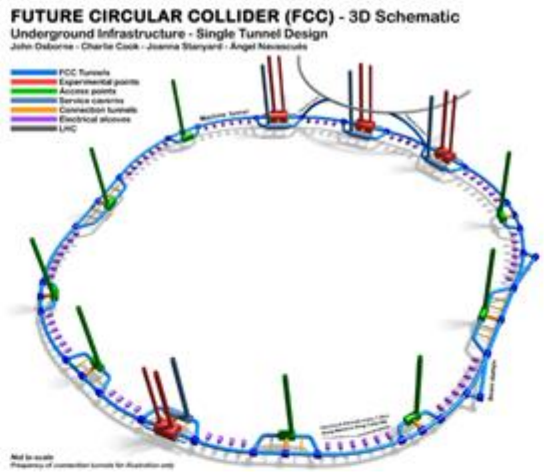
BDA, EGC, MAB-KG, DS,

Data for Physics

- Quentin Bruant, Antoine Chance, Barbara Dalena, Valerie Gautard, Andrés Gomez, Adnan Ghribi, Hugo Le Corre, Jacqueline Keintzel, Yasmina Nasr, Charles Ndungu-Ndegwa , Yukiyoshi Ohnishi , Rogello Thomas Garcia, Jonathan Piscart, Leonardo Vitileia,

- **Papers:** Energies 2023, EPS-HEP 2025

Data for Physics



International FCC collaboration (CERN as host lab) to study:

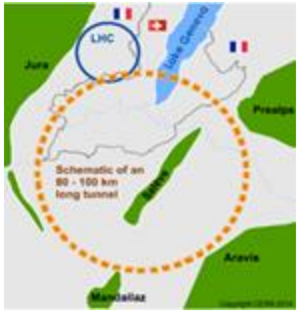
- *pp*-collider (*FCC-hh*), main emphasis, defining infrastructure requirements
- ~100 km tunnel infrastructure in Geneva area, site-specific

$^{+}$*e*$^{-}$ collider (*FCC-ee*), as a potential first step

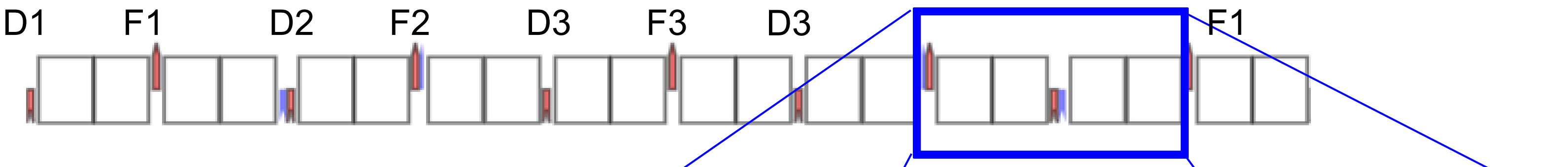
- HE-LHC with *FCC-hh* technology
- *p-e* (*FCC-he*) option, IP integration, e^{-} from ERL

Summary documents provided to EPPSU SG

- FCC-integral, FCC-ee, FCC-hh, HE-LHC
- Accessible on http://fcc-cdr.web.cern.ch/

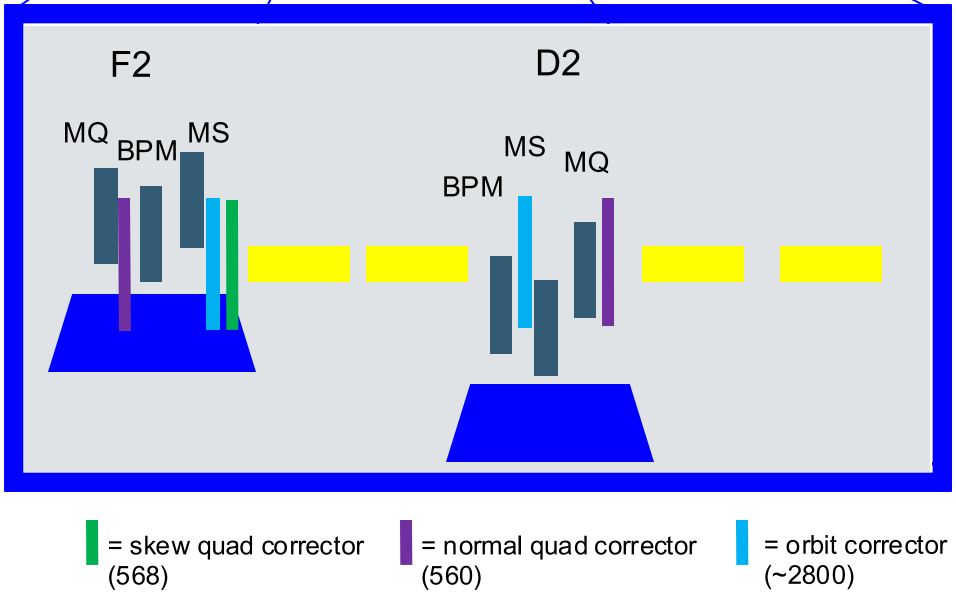


Errors and Correctors



Error type	σ value
Dipole relative field error	10^{-3}
Quadrupole relative field error	**2×10^{-4}**
Sextupole relative field error	**2×10^{-4}**
Main dipole roll error	300 µrad
Offset quadrupoles	200 µm (girder) + 50 µm
Main Quadrupoles roll	**300 µrad**
Offset BPMs	200 µm (girder) + 50 µm
Offset sextupoles	200 µm (girder) + 50 µm

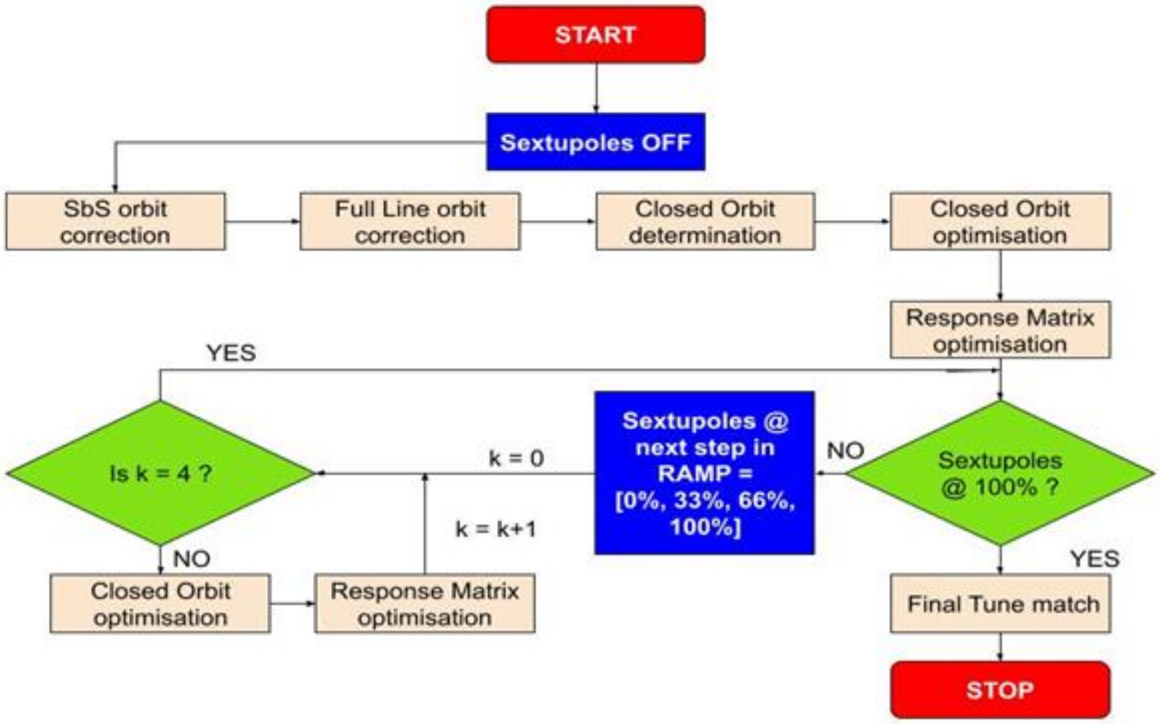
Errors are randomly distributed in arcs
(PDF=Truncated gaussian @3σ).



Correction Strategy

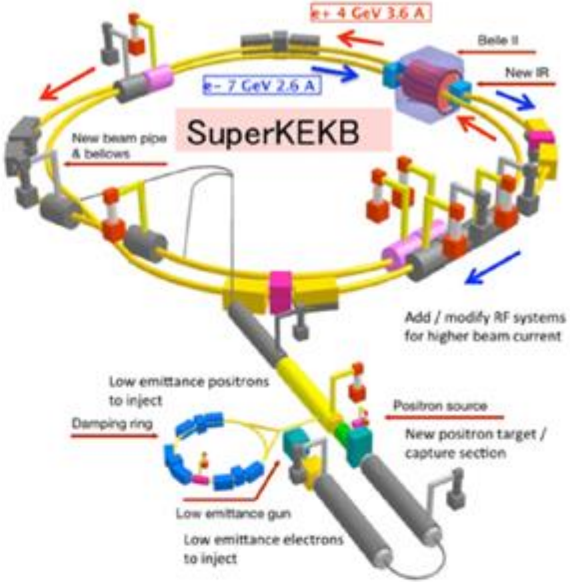
- Sextupole RAMP ensures a limited effect due to the interplay between sextupole strength and imperfections

- 100 seeds simulated



SuperKEKB

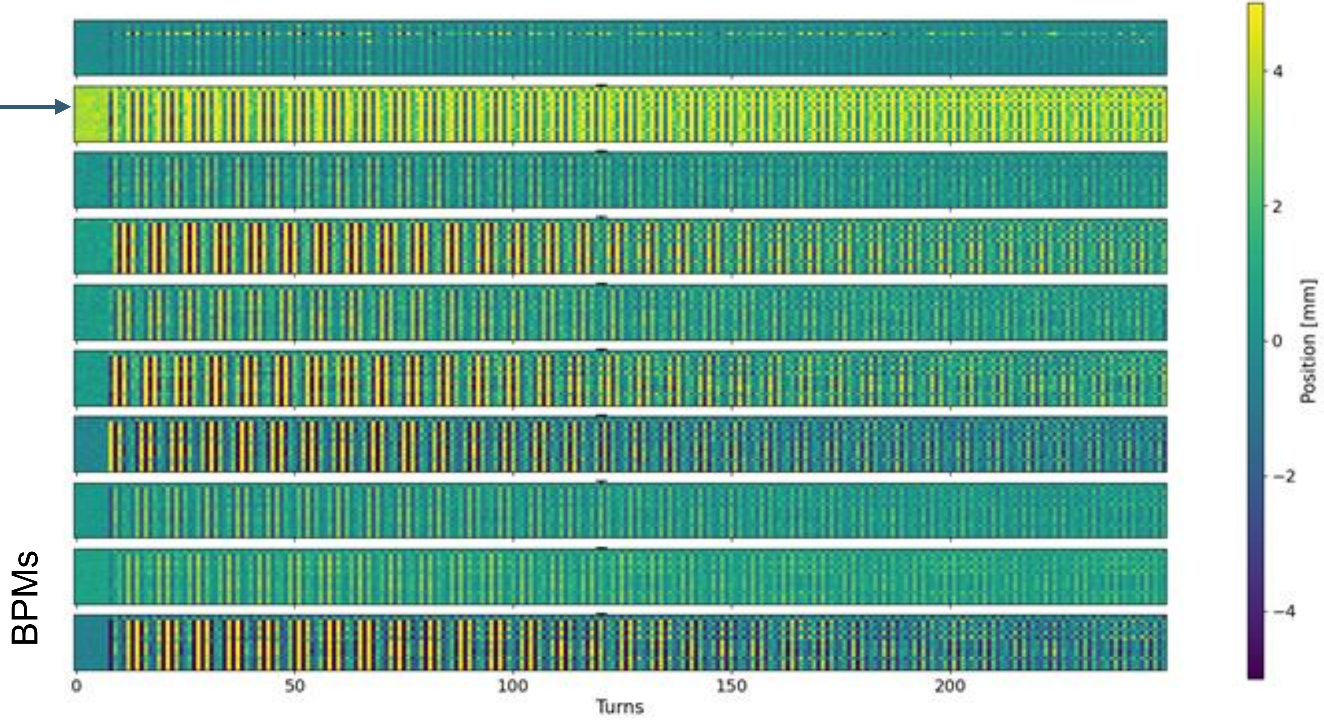
- Largest existing e+/e- collider
  - ~3km long
  - small-scale FCCee
- Provides proof of principle of several concepts and design choices.



SuperKEKB data

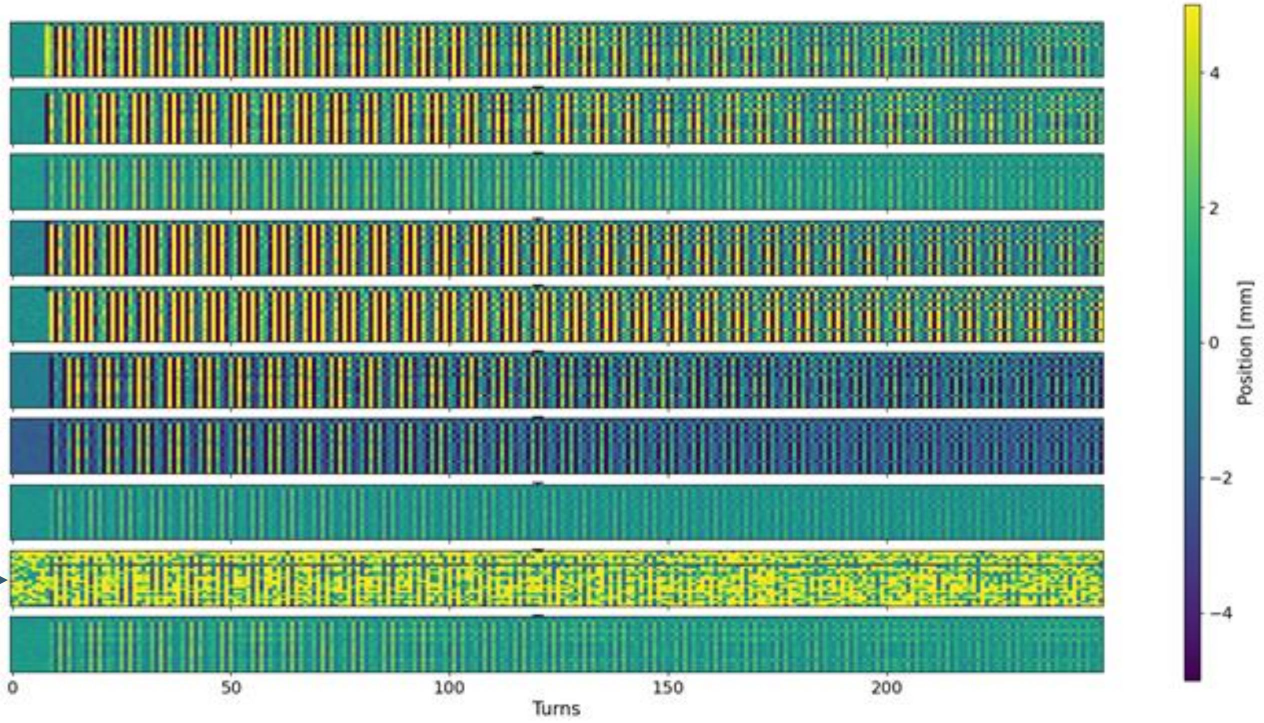
Objective: Find faulty Beam Position Monitors(BPMs)

Shifted BPM



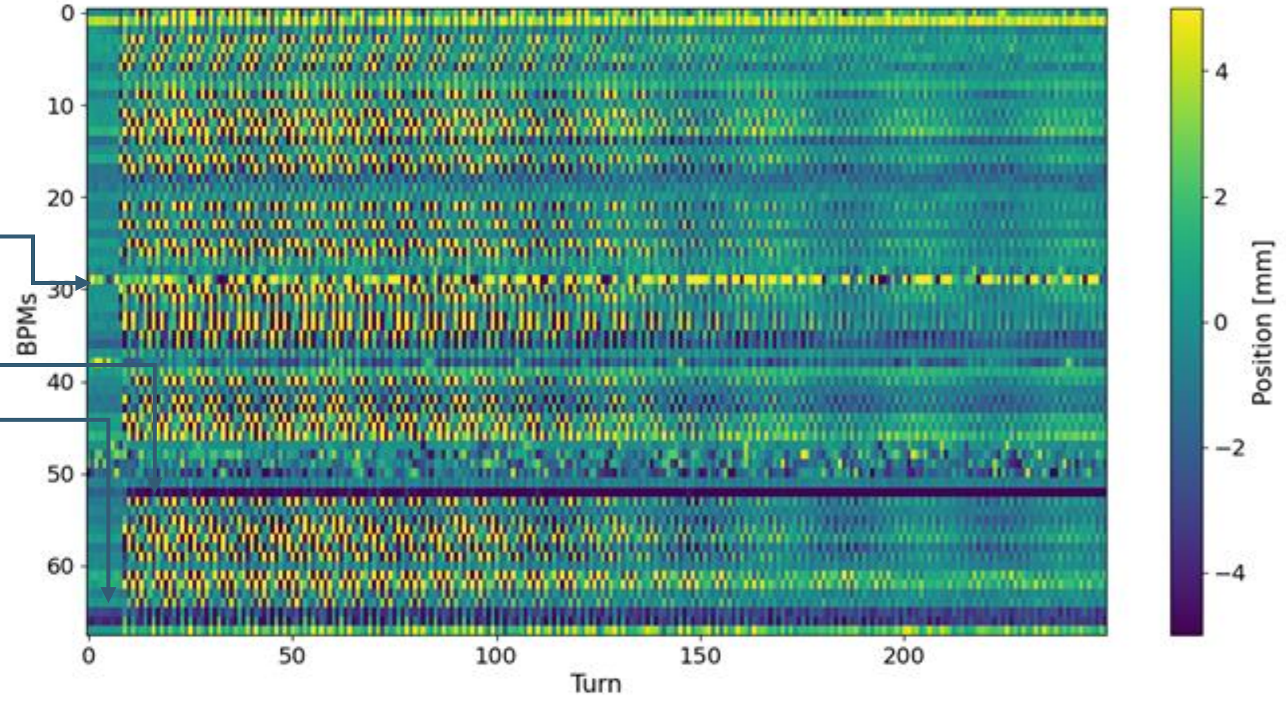
SuperKEKB data

Faulty BPMs



SuperKEKB data

Faulty BPMs

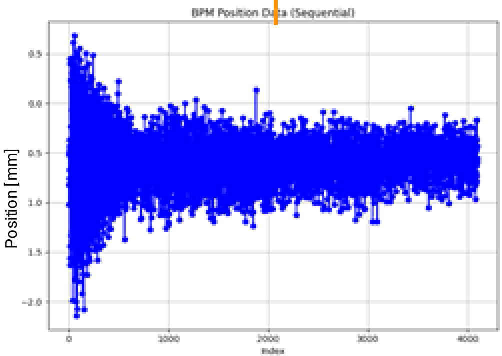


Motivations et défis

Problem	More than 1000 BPM in the FCC rings -> A lot of data but noisy
Impact	Faulty BPMs -> Not good reconstriction of optic functions
Data research Objectif	Machine Learning algorithms for the detection of faulty BPMs and denoising of the others BPMs
Phisics Research objective	Understand the dynamic of the orbits ->

BPM data – standard process

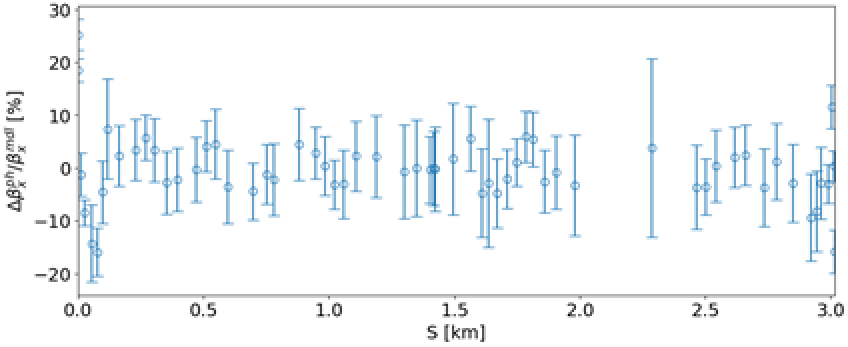
ACCELERATOR → BPM data (gen. sdds format) → Synchronization → SVD cleaning

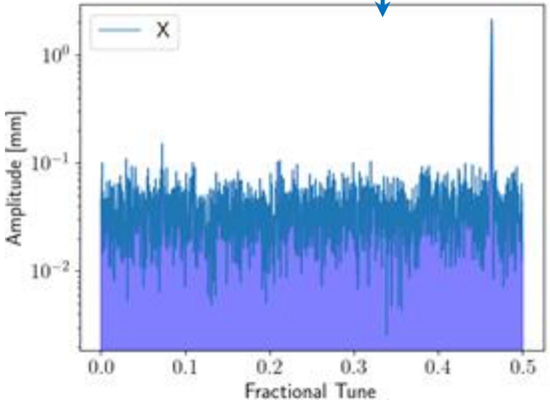


If high singular intrinsic errors → BPM cuts → Harmonic analysis

Else → Harmonic analysis

Harmonic analysis → Optics parameters reconstruction

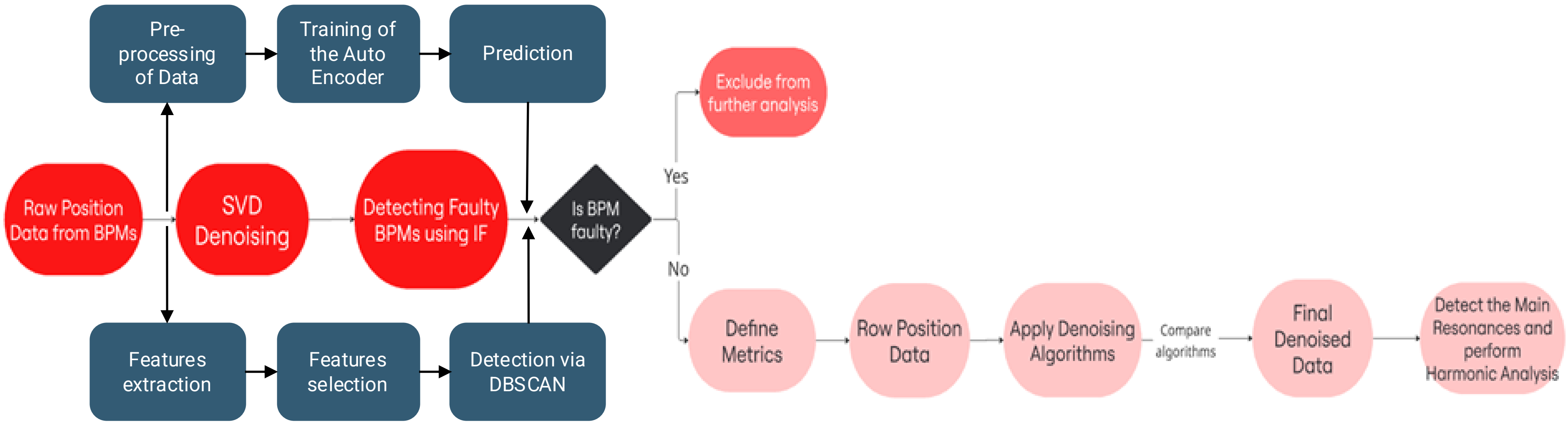




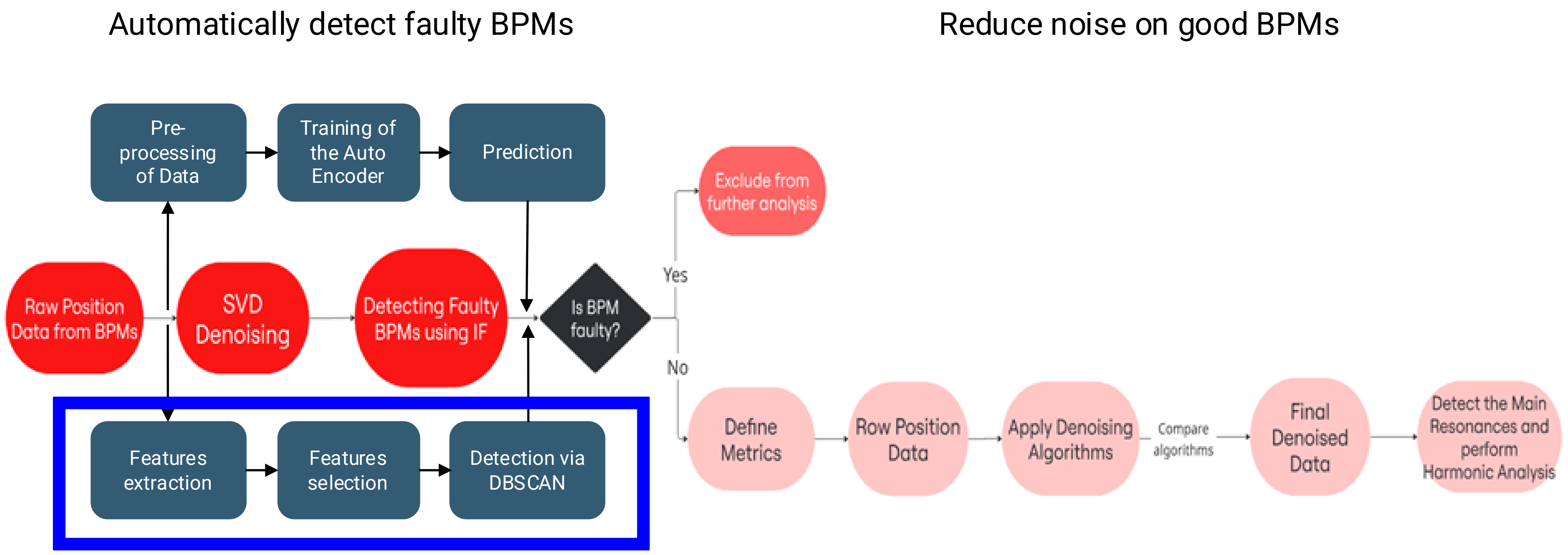
New approach

Automatically detect faulty BPMs

Reduce noise on good BPMs



New approach



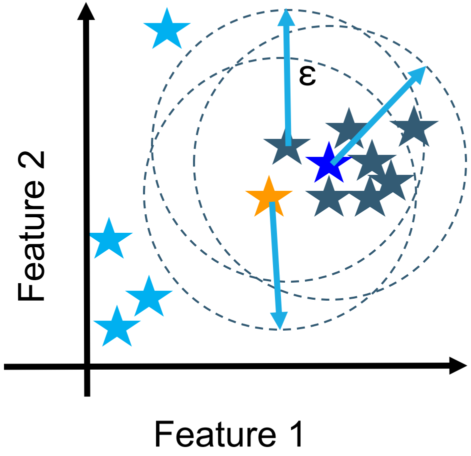
Anomaly detection

★ : outlier out of ε range from other points

★ : first point considered in the cluster

★ : point in the cluster

★ : point in the cluster w/ no neighbour



- Clustering algorithm: DBSCAN (Density-Based Spatial Clustering)
  - Hyperspace of the statistical features from Multivariate Time Series (MTS)
- Two dimensions:
- ε distance-based algorithm

Processus suivi

- ➢ The features extracted by the library Time2Feat
- ➢ PCA decomposition with explainability with variance (thresh. 90%)

➢ 1 mesure, all the BPMs

For the high-energy ring (electrons)

Dataset:
HER_2024_02_06_IK_H_Vinjkick/
HER_2024_02_06_16_22_57.data

Features : 37/201

For the low-energy ring (positrons):

Dataset:
LER_2024_02_06_IK_H_Vinjkick/
LER_2024_02_06_17_02_14.data

Features: 31/207

First Results

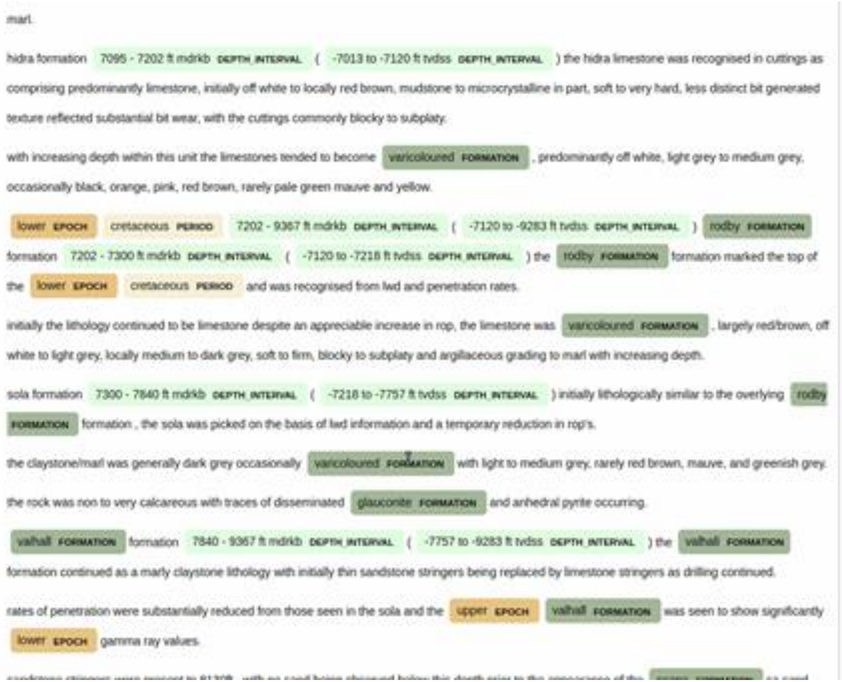
- HER:
  5/6 of faulty BPM are identified
- LER:
  3/6 of faulty BPM are identified
- The LER has more variability in data

BPM faulty HER (DBSCAN)	BPM alredy labeled as faulty HER	BPM faulty LER (DBSCAN)	BPM alredy labeled as faulty LER
MQEAE35	MQEAE35	MQEAP35	MQEAP35
MQD3E18	MQD3E18	MQW2ORP	MQW2ORP
MQEAE20	MQEAE20	MQEAP29	MQEAP29
MQD3E8	MQD3E8		MQD3P8
MQR2ORE	MQR2ORE		MQEAP10
	MQD3E23		MQD3P23
MQEAE25		MQEAP32	
MQEAE33		MQEAP33	
MQD3E29		MQI6P	
		MQEAP38	
		MQEAP44	
		MQD3P29	

Perspectives

- Design FCCee-HEB:
  - Study to lower the number of correctors (autocorrelation, NN-based methods, …)
  - Implement a method considering tapering for ttbar operation (@182GeV)
  - Compute the Dynamic Aperture with the errors implemented
- Anomaly detection and denoising of TbT-BPMs:
  - Building a truth table for the algorithm DBSCAN
  - Understand the feature selection process

Named Entity Recognition System



creating new dictionaries and running the pipeline again

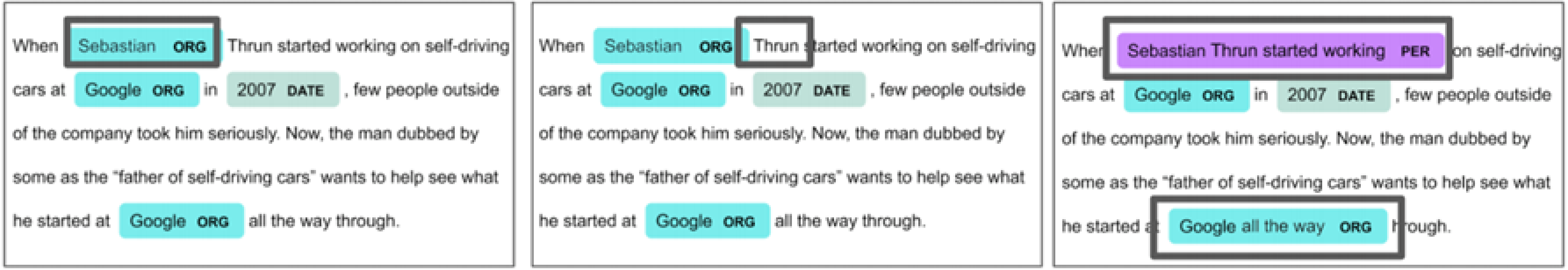
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

CentraleSupélec

Noisy training set - Data annotation

Errors in labels are common, even when annotated by humans:

Name-Entity Recognition



PER (person) misclassified as *ORG* (organization)

PER (person) annotation missing

Imprecise boundaries

Source: [Abid 2020]

Method overview

1. We create a **noisy training set** in a semi-automatic way **avoiding manual data annotation**
2. Use **Deep Neural Networks approach** (DNNs)

3. **We use transfer learning to create a NER model:**
   - Using pre-trained language models to learn contextual representation in our domain-specific corpus
   - During training we use regularization to avoid learning the noisy examples -> and two training steps
4. **Evaluate using human-reviewed data sets**

NER in energy domain

III STRATIGRAPHY

All depths are referenced to KB PERSON (80 DATE ' above MSL ORG)

Quaternary - Tertiary GPE

Lower Miocene - Recent PERSON

Seabed 434 CARDINAL ' - 1460" (1026 DATE ' thick)

Samples ORG dumped to sea bed.

HORDALAND GROUP ORG

Utsira Formation ORG

1460 DATE ' (top not seen)-2224' (764 ' Thick WORK_OF_ART)

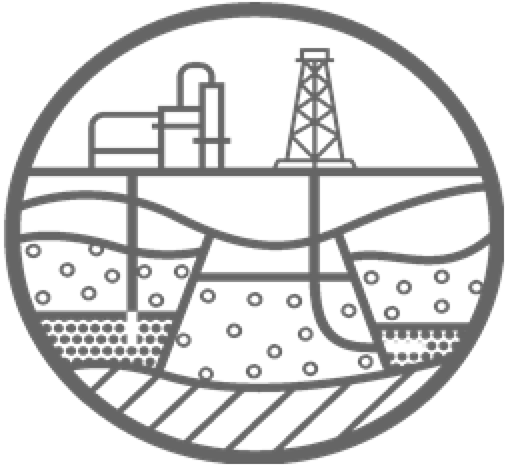
HORDALAND GROUP ORG

Lark Formation

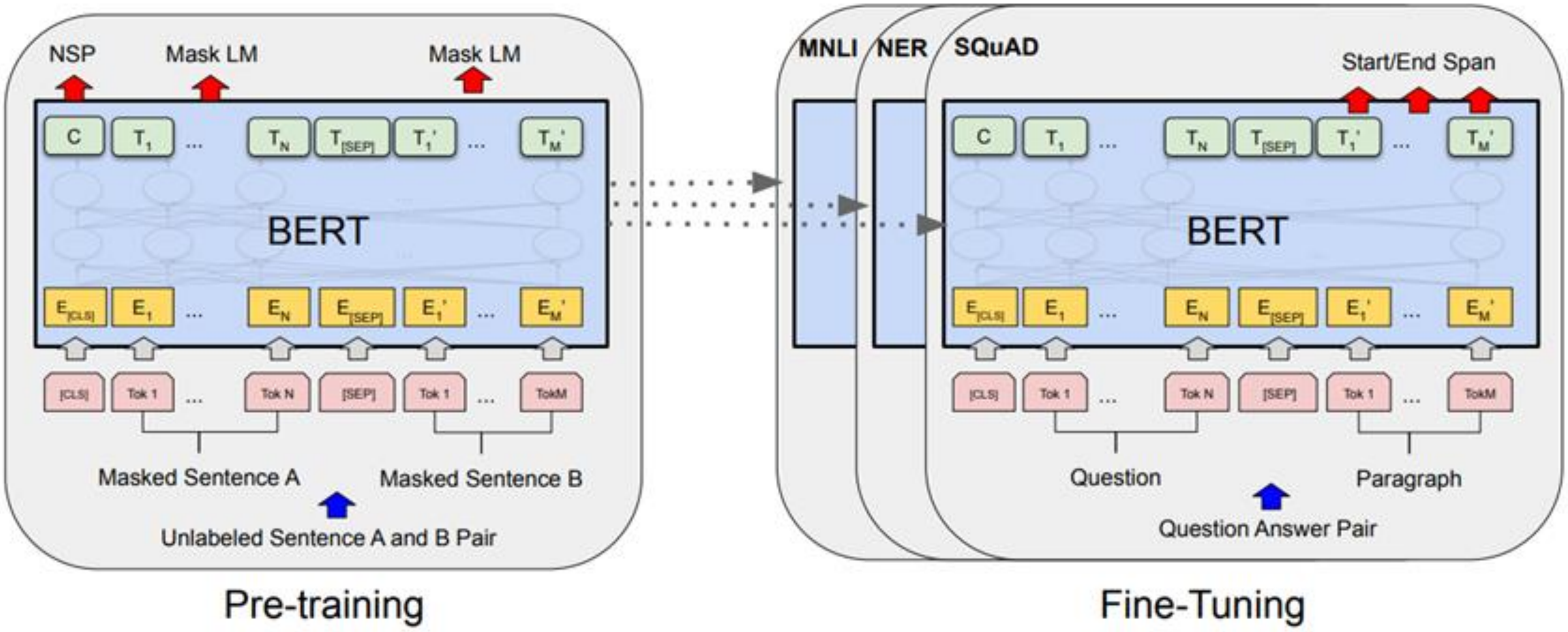
2224 DATE ' - 3104' (880' Thick WORK_OF_ART)

The Lark Formation WORK_OF_ART is developed largely in argillaceous facies comprising light grey, grey-green and occasionally brownish grey, soft, argillaceous siltstones. These are frequently pyritic, glauconitic, micaceous and calcareous. Locally these grade to claystones of similar character, which

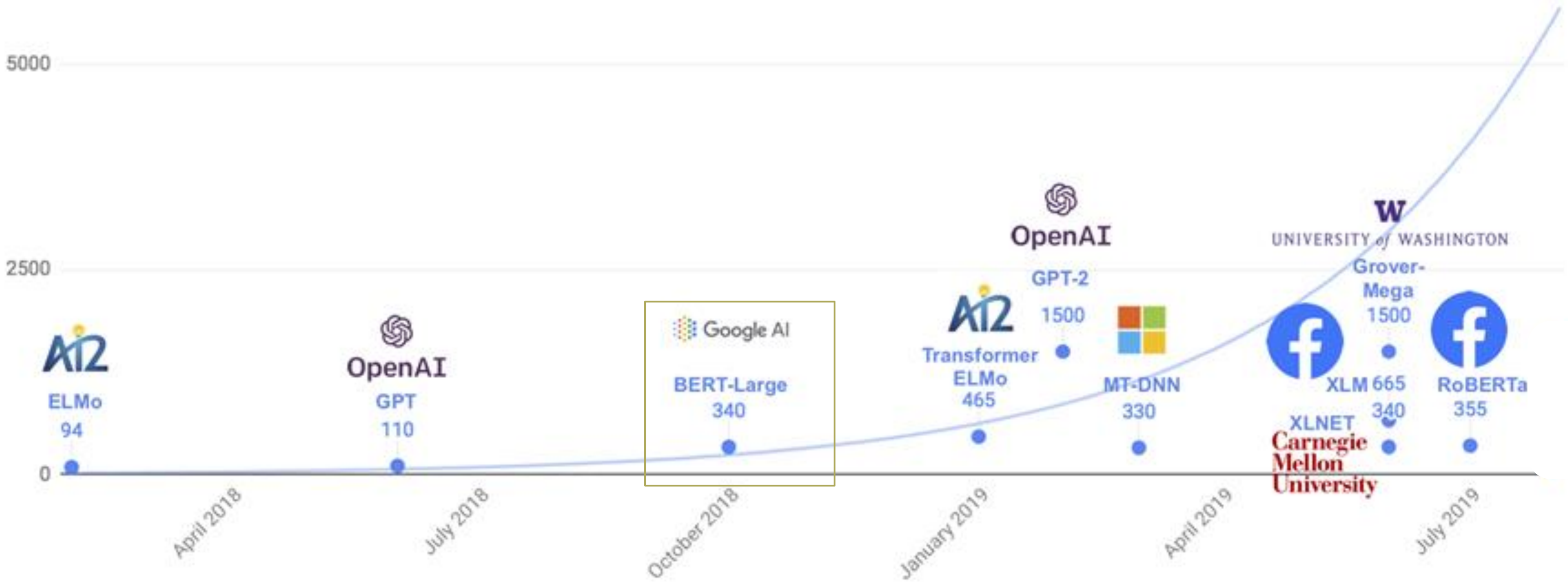
NER is a domain-specific task!

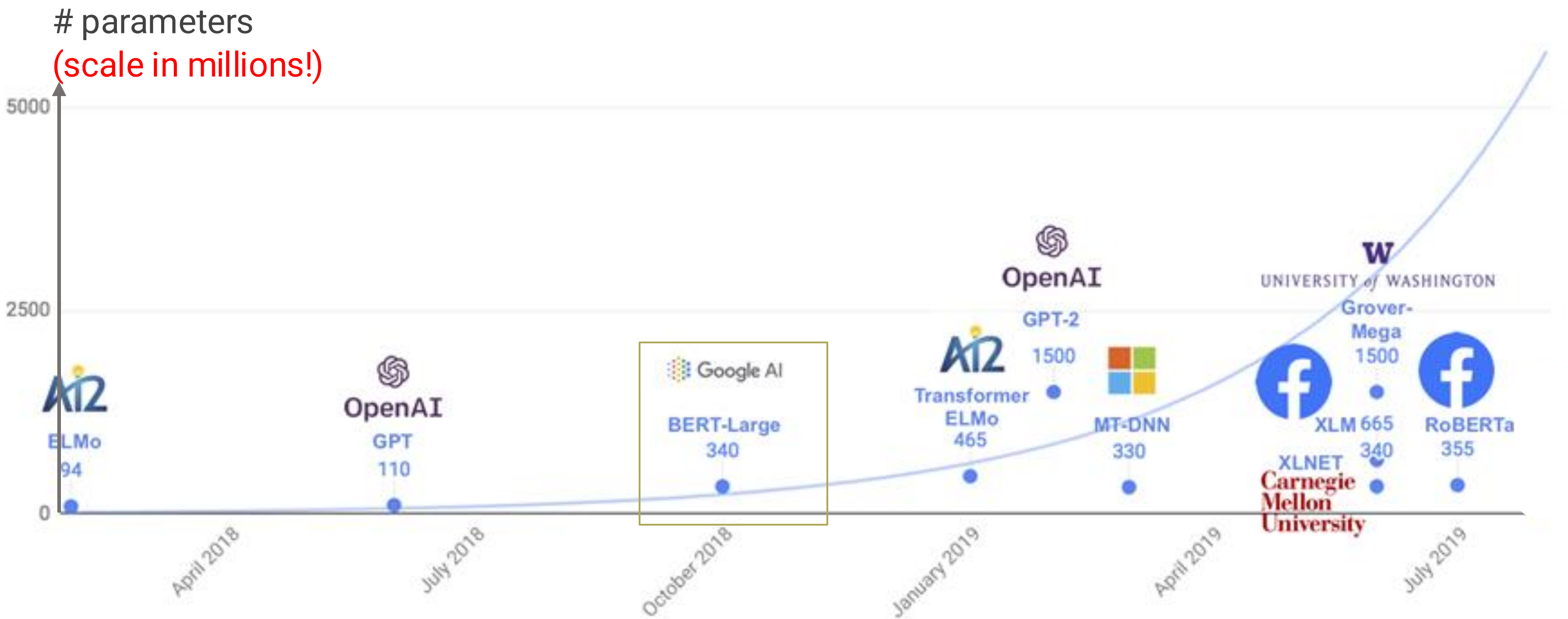


BERT - Bidirectional Encoder Representations from Transformers



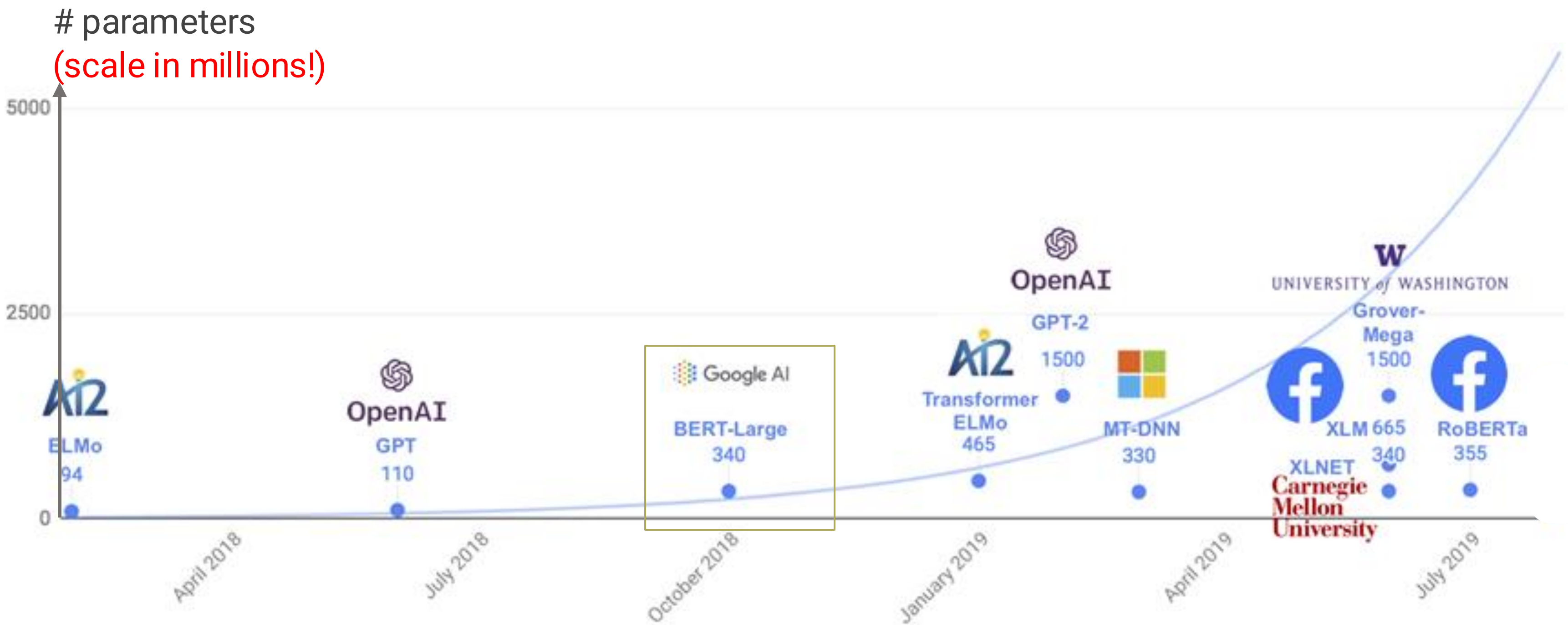
Common NLP training setting: Use a pre-trained model and fine-tune to the downstream task


5000
2500
0
ELMo
94
OpenAI
GPT
110
Google AI
BERT-Large
340
Transformer
ELMo
465
OpenAI
GPT-2
1500
MT-DNN
330
UNIVERSITY of WASHINGTON
Grover-
Mega
1500
XLM 665
340
XLNET
Carnegie
Mellon
University
RoBERTa
355
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019



parameters
(scale in millions!)
5000
2500
0
ELMo
94
OpenAI
GPT
110
Google AI
BERT-Large
340
Transformer
ELMo
465
OpenAI
GPT-2
1500
MT-DNN
330
UNIVERSITY of WASHINGTON
Grover-
Mega
1500
XLM 665
340
XLNET
Carnegie
Mellon
University
RoBERTa
355
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019

Knowledge Distillation



Knowledge distillation: model compression method in which a small model is trained to mimic a pre-trained, larger model (or ensemble of models). https://arxiv.org/abs/1910.01108

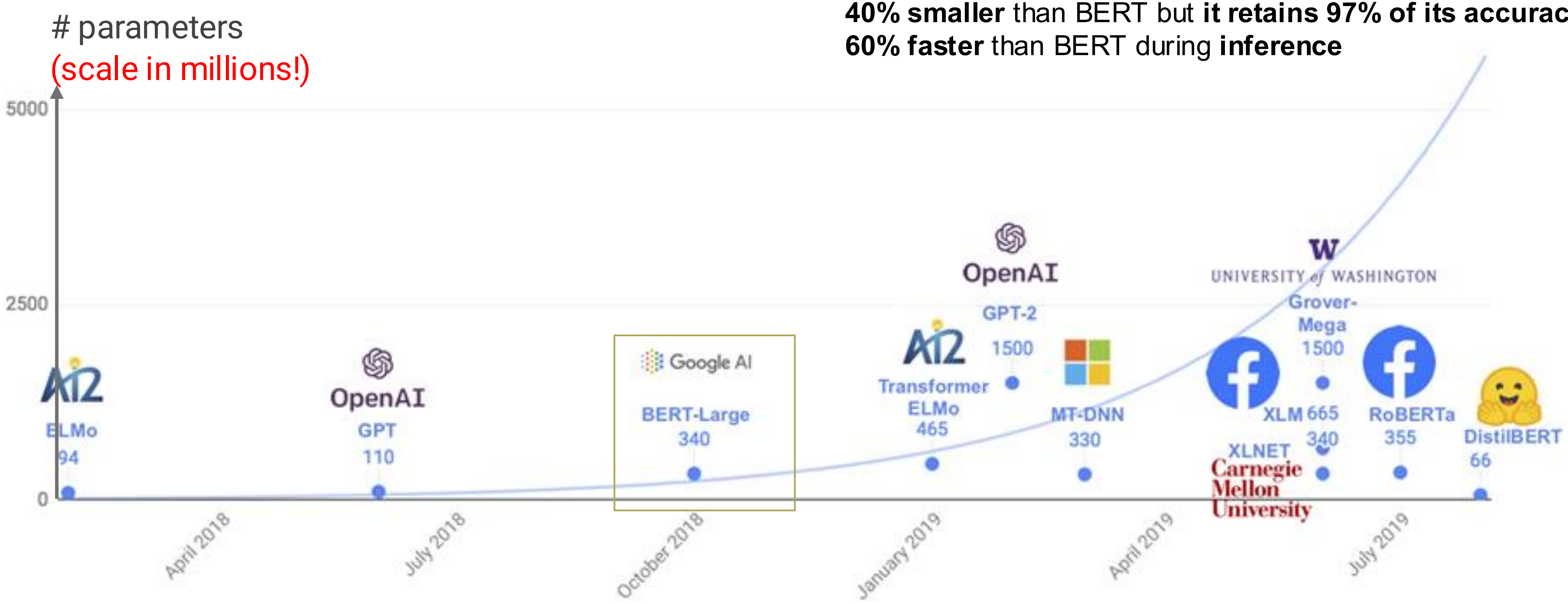
Knowledge Distillation

parameters
(scale in millions!)

Knowledge distillation: model compression method in which a small model is trained to mimic a pre-trained, larger model (or ensemble of models). https://arxiv.org/abs/1910.01108

Knowledge Distillation

DistilBERT:
40% smaller than BERT but **it retains 97% of its accuracy**
60% faster than BERT during **inference**



Knowledge distillation: model compression method in which a small model is trained to mimic a pre-trained, larger model (or ensemble of models). https://arxiv.org/abs/1910.01108

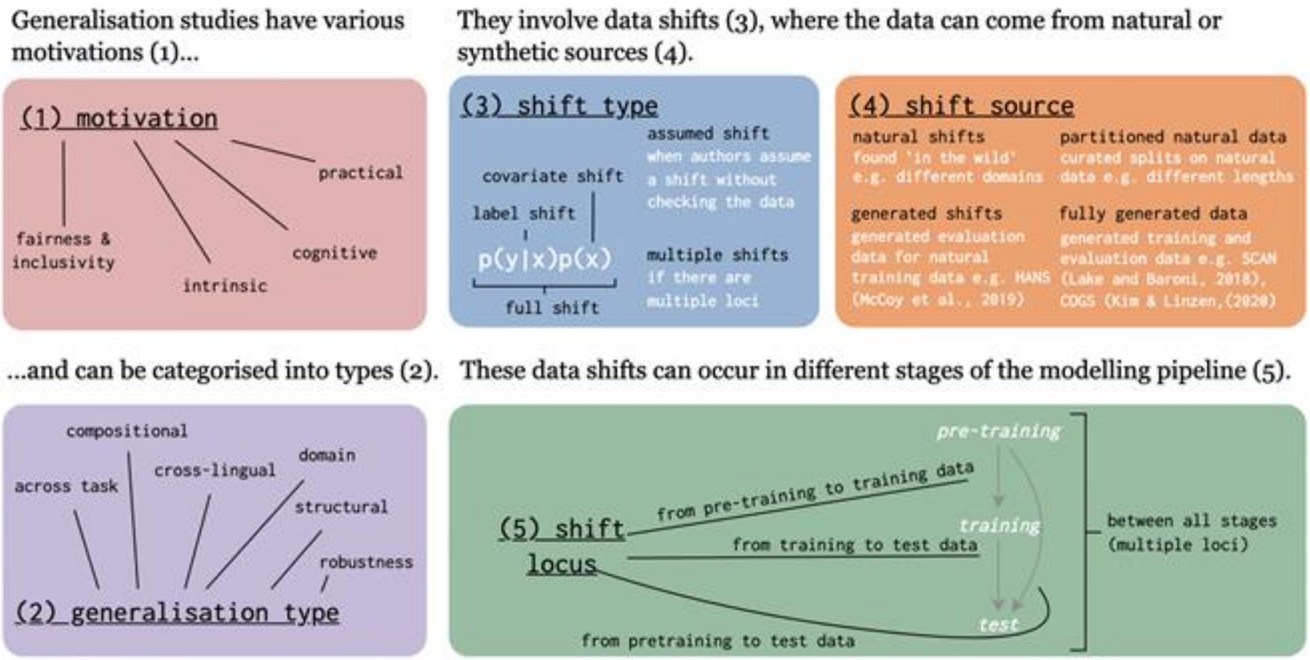


Figure 1: A graphical representation of the NLP generalisation taxonomy we present in this paper. The taxonomy consists of five different (nominal) axes, that describe the high-level *motivation* of the work (§2); the *type* of generalisation the test is addressing (§3); what kind of *data shift* occurs between training and testing (§4), and what the *source* and *locus* of this shift are (§5 and §6, respectively).

State-of-the-art generalisation research in NLP: a taxonomy and review

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Rita Frieske, Ryan Cotterell, Zhijing Jin

Named Entity Recognition

Named Entity Recognition

London is the capital and most populous city of **England** and the **United Kingdom**.

Geographic Entity — Geographic Entity — Geographic Entity

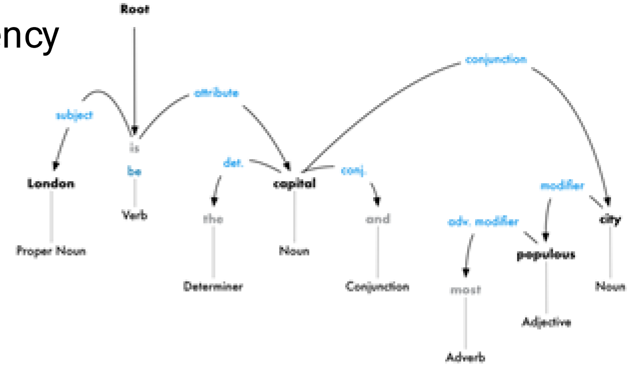
Natural Language Processing Pipeline

Input text → **Sentence Segmentation** → Tokenization → Part of Speech Tagging → Dependency Parsing → NER → Coreference resolution → Parsed text

POS

London	is (be)	the	capital	and	most	populous ...
Proper Noun	Verb	Determiner	Noun	Conjunction	Adverb	Adjective

Dependency



Co-Reference

London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.

Source: https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e

Scope - Defined Named Entities

```
UPPER CRETACEOUS
2,371ft to 3,811ft MDRT
Chalk Group (Undifferentiated)
(-2,154ft to -3,425ft TVDSS)
The Chalk Group was identified by an initial drop in ROP and a marked downward step in GR values.
CHALK: white, very soft, amorphous, mostly dispersing in mud, very little removed by shakers, leaving
abundant Chert fragments and fossil debris in samples (Echinoid spines, sponge spicules, bryozoa,
forams).
CHERT: off white, bluish grey to yellowish brown translucent, very hard splintery conchoidal shards,
cryptocrystalline, occasionally coated with white opaque silicified chalk.
ROP: 10 - 128 ft/hr (Average: 66 ft/hr).
Gas: 0.014% (Average: 0.012%)
3,811ft to 3,816ft MDRT
Plenus Marl Formation
(-3,425ft to -3,429ft TVDSS)
The Plenus Marl Formation was identified by a sudden change in drilled cuttings and by a distinctive
Gamma response, known to be a regional feature.
CLAYSTONE: dark greenish grey to black, firm, blocky, crumbly and earthy in part, common carbonaceous
specks, moderately calcareous.
```

Conoco (U.K.) Ltd.

Lyell Field Dipmeter Study

(Well 3/2-1A)

Interval 11354-11376ft. (logged depth). Crevasse splay sandstones interbedded with overbank mudstones (core description). This interval includes two fining upwards sandstone packages, see Figure 11 circa 11360ft. and 11370ft. These comprise small crossbeds and wavy and ripple laminated units which are too small to be resolved by dipmeter. The overbank mudstones include coals at the top of each unit circa 11355ft. and 11365ft. with associated rooted horizons. The upper sand (11359-11365ft.) shows an upward decrease in dip magnitude in the range 200-350, which although consistent with the magnitude range of cross bedding, core evidence indicates that individual bed thickness here are generally less than 6 inches and therefore most of the dips seen are probably bed boundaries.

- ✓ Well ID
- ✓ Period
- ✓ Age
- ✓ Epoch
- ✓ Formation
- ✓ Depth Interval
- ✓ Interval (without depth reference)

Overview

- We don't have training data
  - Annotating data is a labor-intensive task (AI bottleneck)
  - We have external resources: dictionaries and regular expressions (search patterns)
  - We can create labels using external resources -> noisy training set
- Extend to new entities -> We can't follow traditional approach
  - Avoid hand-crafted grammar rules

✓ We decided to use Deep Neural Networks (DNNs)

- But how to mitigate the effect of noisy labels?
  - ✓ Contextual representation (pre-trained language models) -> good examples have consistent context -> they will be closer in the embedding space
  - ✓ Take advantage of DNN regularization to avoid learning the noisy examples

Noisy training set - Regular Expressions

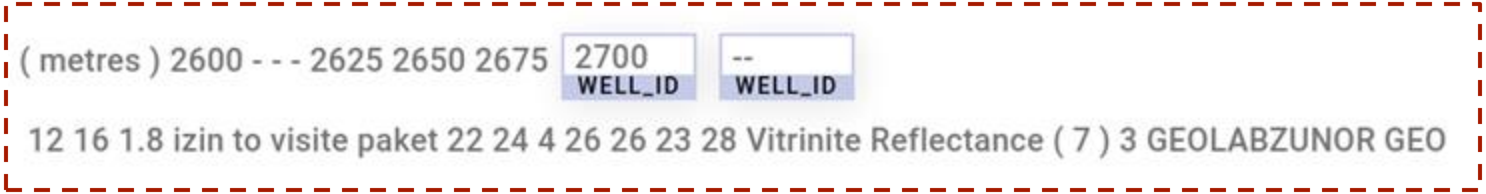
RegExp for Well IDs

[1-9][0-9]{0,2}([\-_/]{1}([0-9]{1,3}[a-z]{0,1}){1,3}){1,5}

154/3-1
well 154/3-1
well: 154/3-1
15/30-8
30/05-03

> *Some people, when confronted with a problem, say*
> ***«I know, I will use regular expressions»***
> *Now they have two problems.*
>
> Jamie Zawinski
> Netscape and Mozilla Cofounder

Noisy example



Other noisy examples include **dates**, **page descriptors**, and **coordinates**

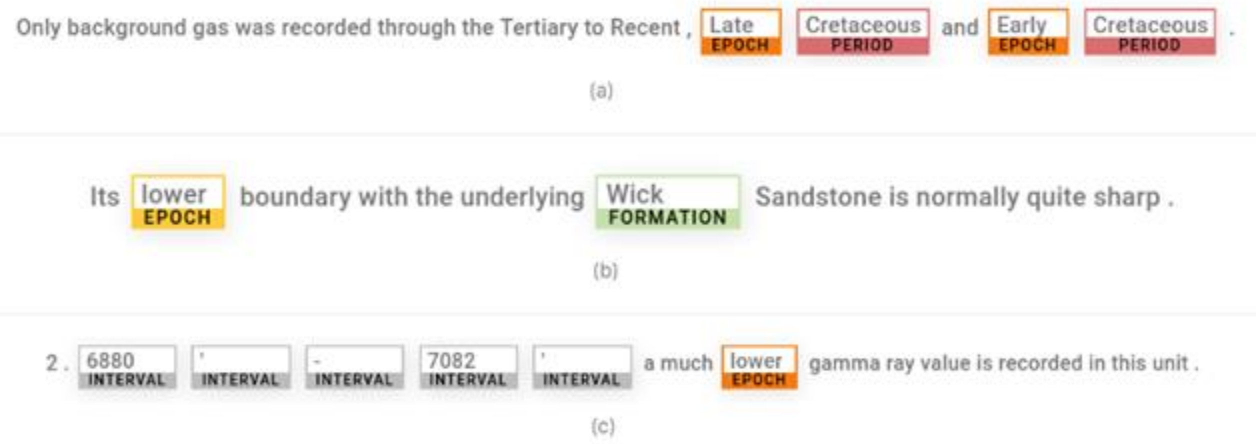
Noisy training set - Dictionaries

EPOCH Dictionary
(geological time)

Mesozoic Era
- Cretaceous Period
  - Late (Upper) Epoch
  - Early (Lower) Epoch
- Jurassic Period
  - Late (Upper) Epoch
  - Middle Epoch
  - Early (Lower) Epoch
- Triassic Period
  - Late (Upper) Epoch
  - Middle Epoch
  - Early (Lower) Epoch

Paleozoic Era
- Permian Period
  - Lopingian Epoch
  - Guadalupian Epoch
  - Cisuralian Epoch
- Carboniferous Period
  - Pennsylvanian Epoch*
  - Mississippian Epoch*

Challenges
1. Not always complete (generalization)
2. Polysemy problem: Terms in the dictionary are used with other meaning



The drilling process started late -> hopefully the word **late** is not referring to geological time

Learning contextual representation - Language Models

"The service was poor, but the food was…"

Example by Sebastian Ruder. More interesting & interactive examples: https://pudding.cool/2019/04/text-prediction/

Learning contextual representation - Language Models

"The service was poor, but the food was…"

delicious

tasteless

horrible

yummy

Example by Sebastian Ruder. More interesting & interactive examples: https://pudding.cool/2019/04/text-prediction/

Learning contextual representation - Language Models

"The service was poor, but the food was…"

delicious

tasteless

horrible

yummy

Example by Sebastian Ruder. More interesting & interactive examples: https://pudding.cool/2019/04/text-prediction/

Learning contextual representation - Language Models

"The service was poor, but the food was…"

delicious

tasteless

horrible

yummy

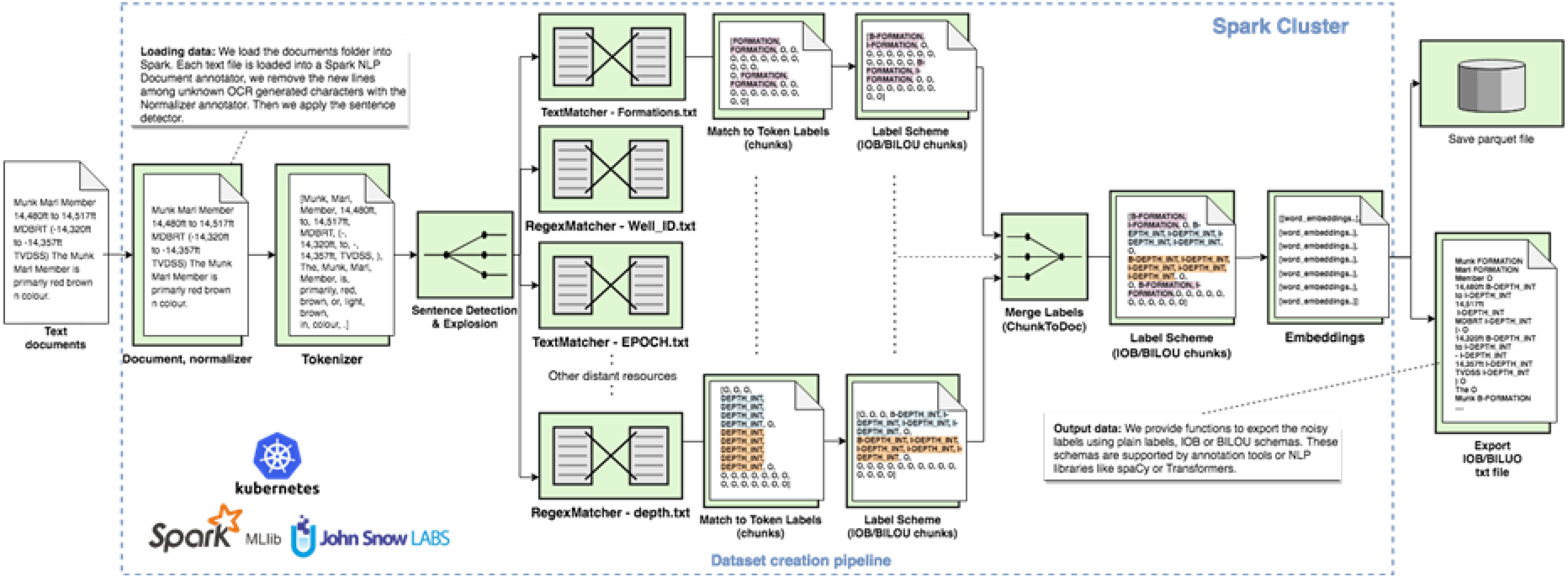
The model has to memorize what words are used to describe food

Identify that 'but' introduces a contrast: the new adjective has the opposing sentiment of 'poor'

- ✓ **It help us to learn the fundamentals of language!**
- ✓ **We don't require labels to solve this problem**

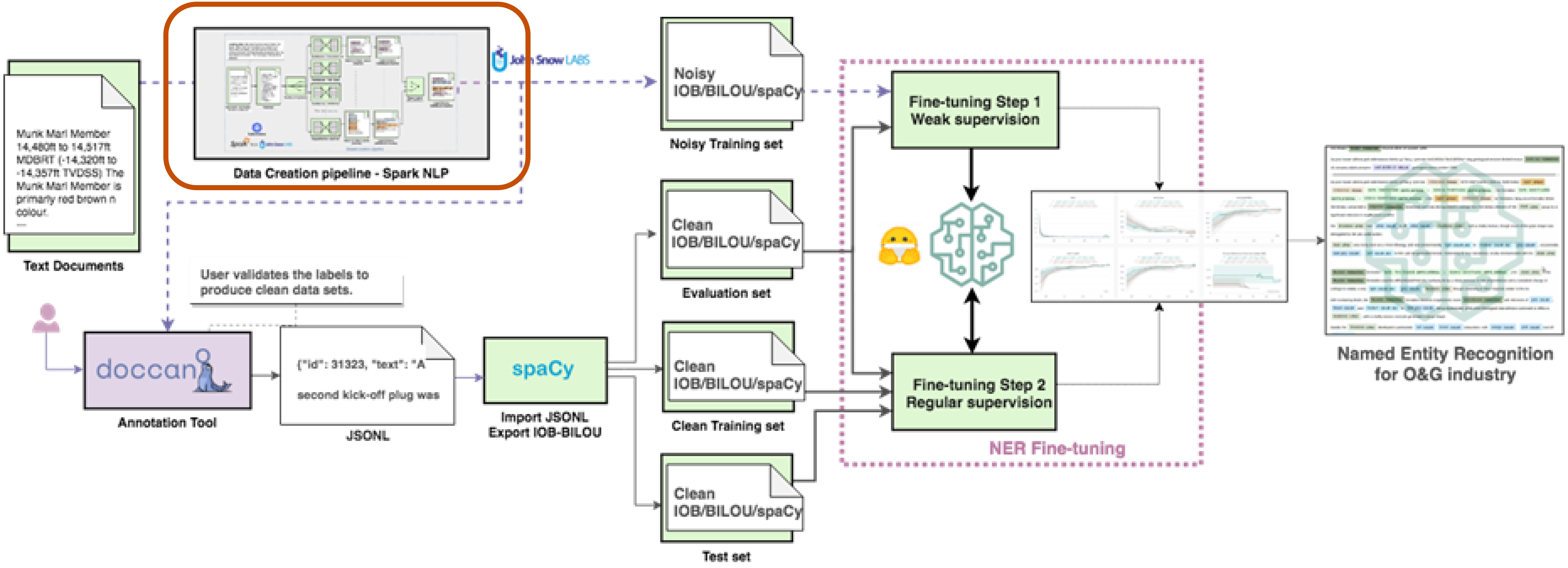
Example by Sebastian Ruder. More interesting & interactive examples: https://pudding.cool/2019/04/text-prediction/

Project Implementation – Dataset creation

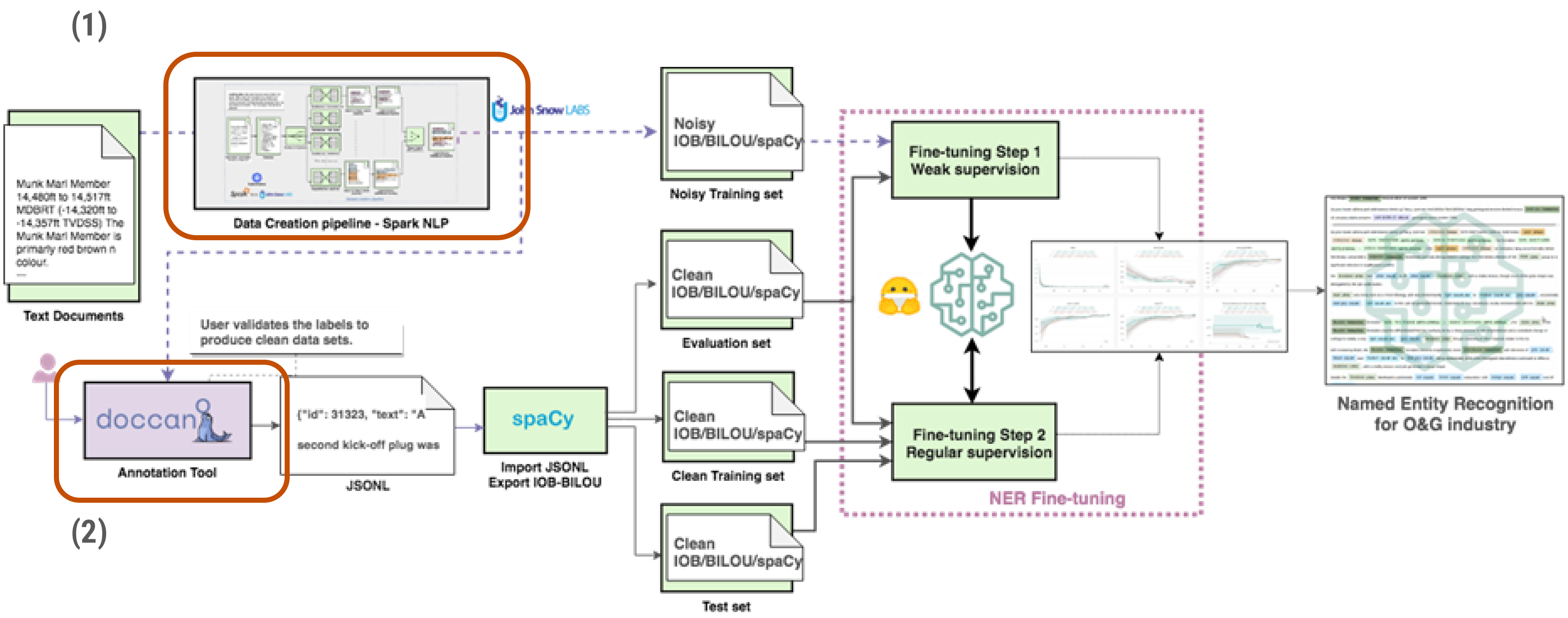


Implementation

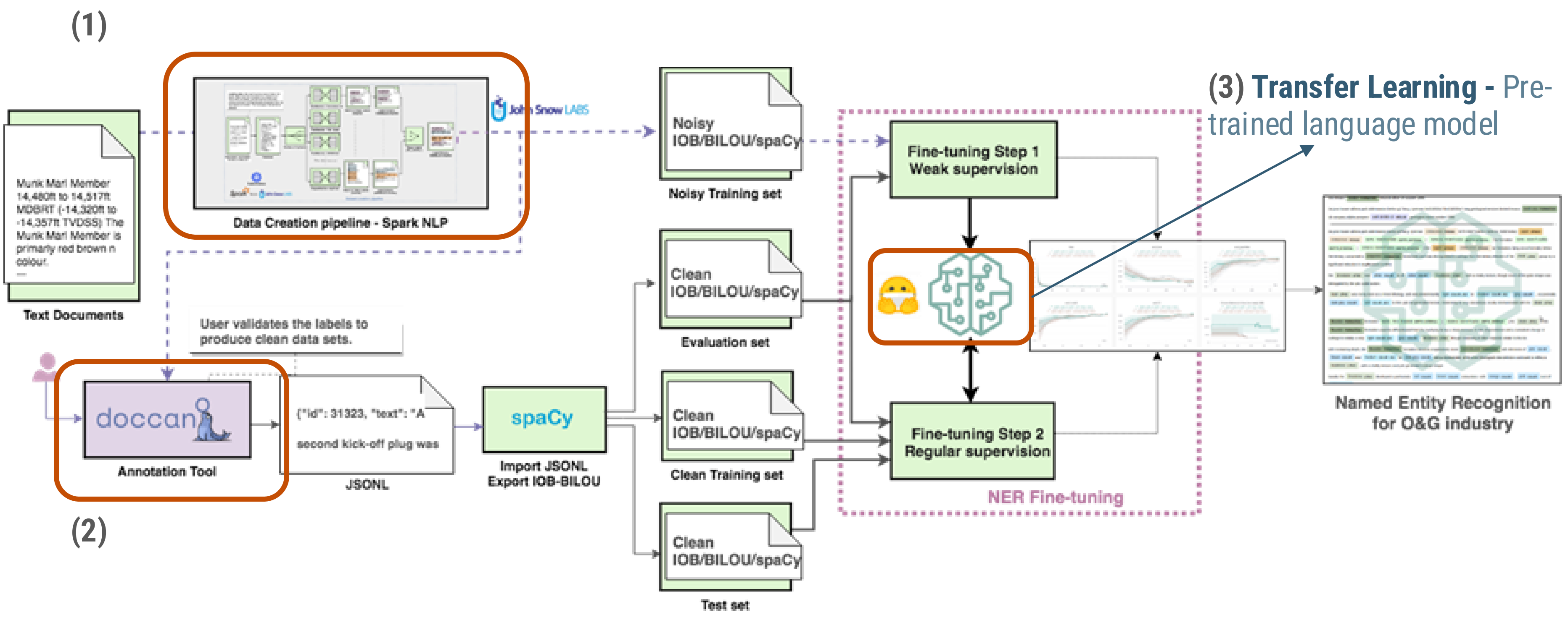
(1)



Implementation



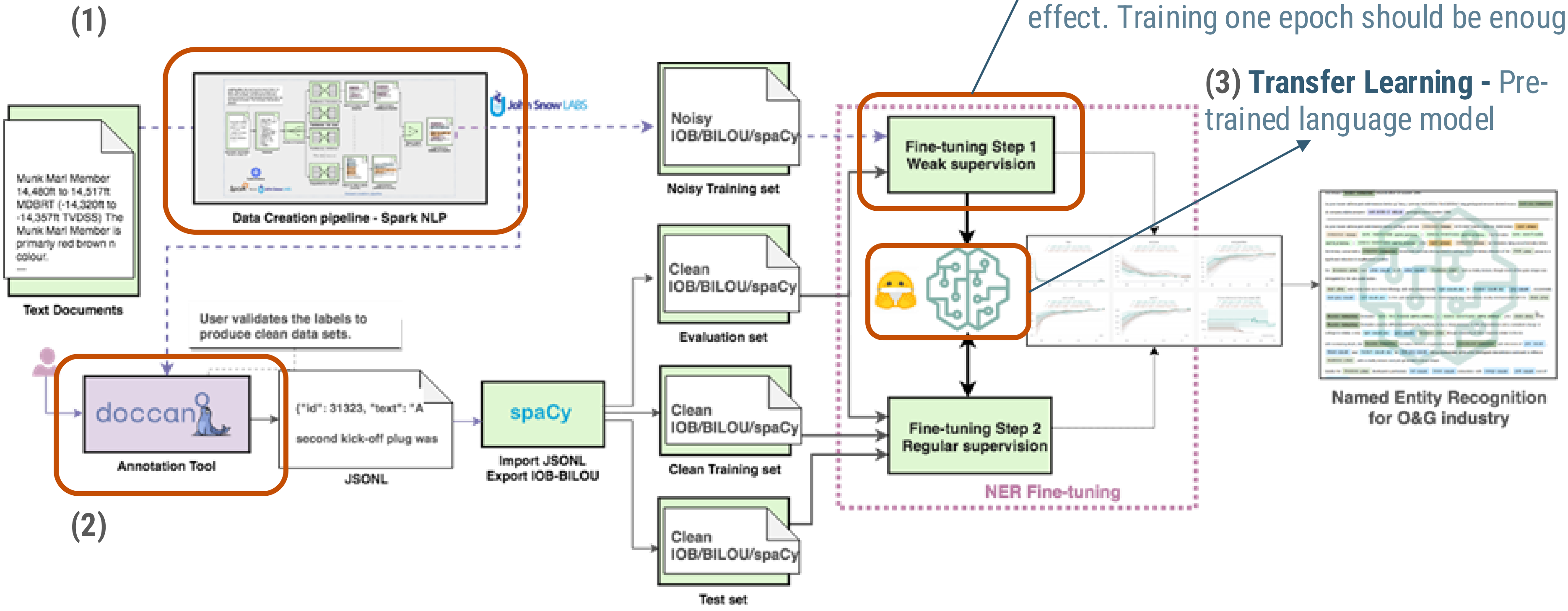
Implementation



Implementation

(4) **Unreliable data:** Group training data with large **Batch Size** to diminish the noise effect. Training one epoch should be enough

(3) **Transfer Learning -** Pre-trained language model



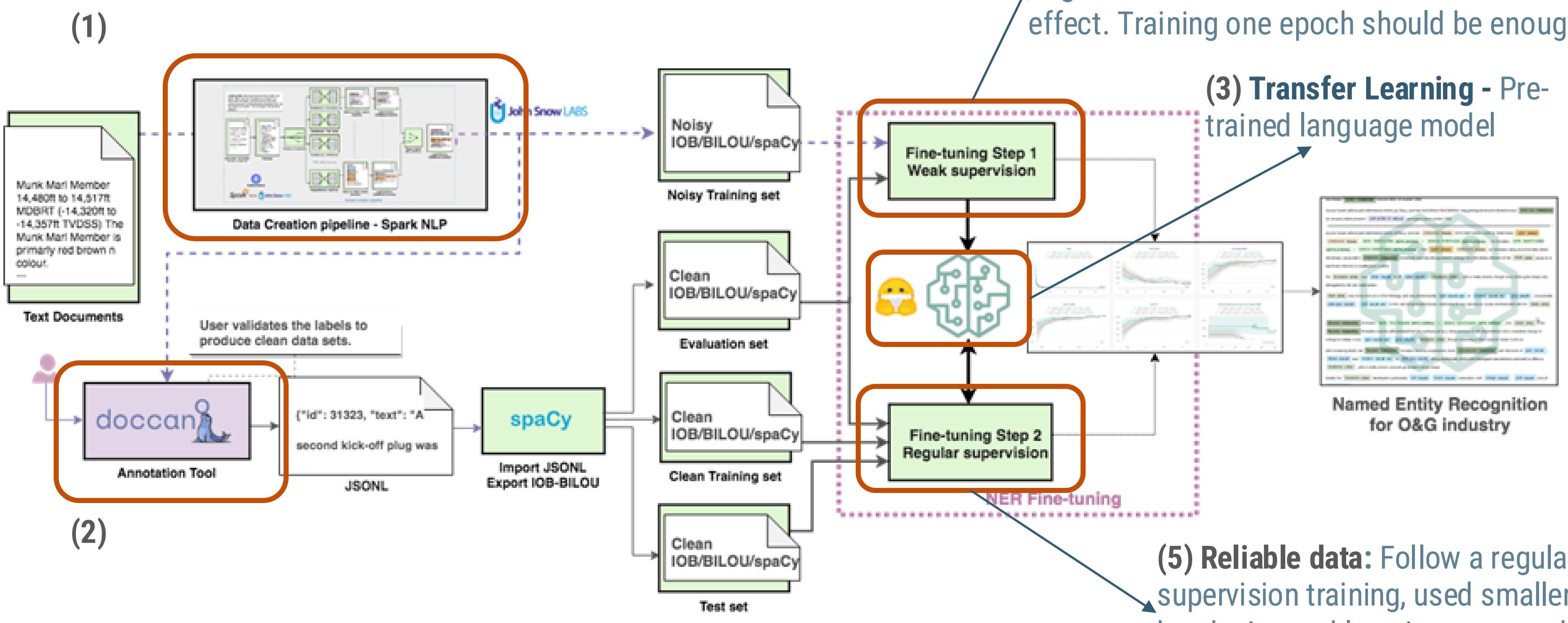
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

CentraleSupélec

Implementation

(4) **Unreliable data:** Group training data with large **Batch Size** to diminish the noise effect. Training one epoch should be enough

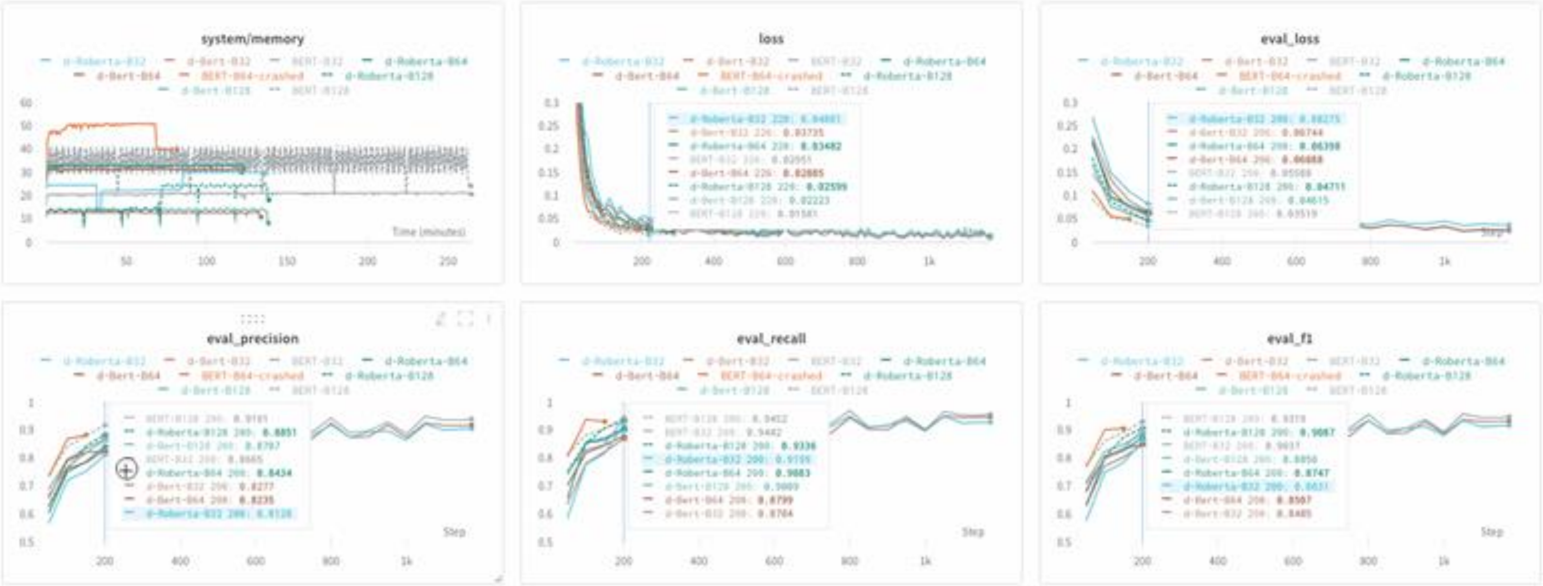
(3) **Transfer Learning -** Pre-trained language model



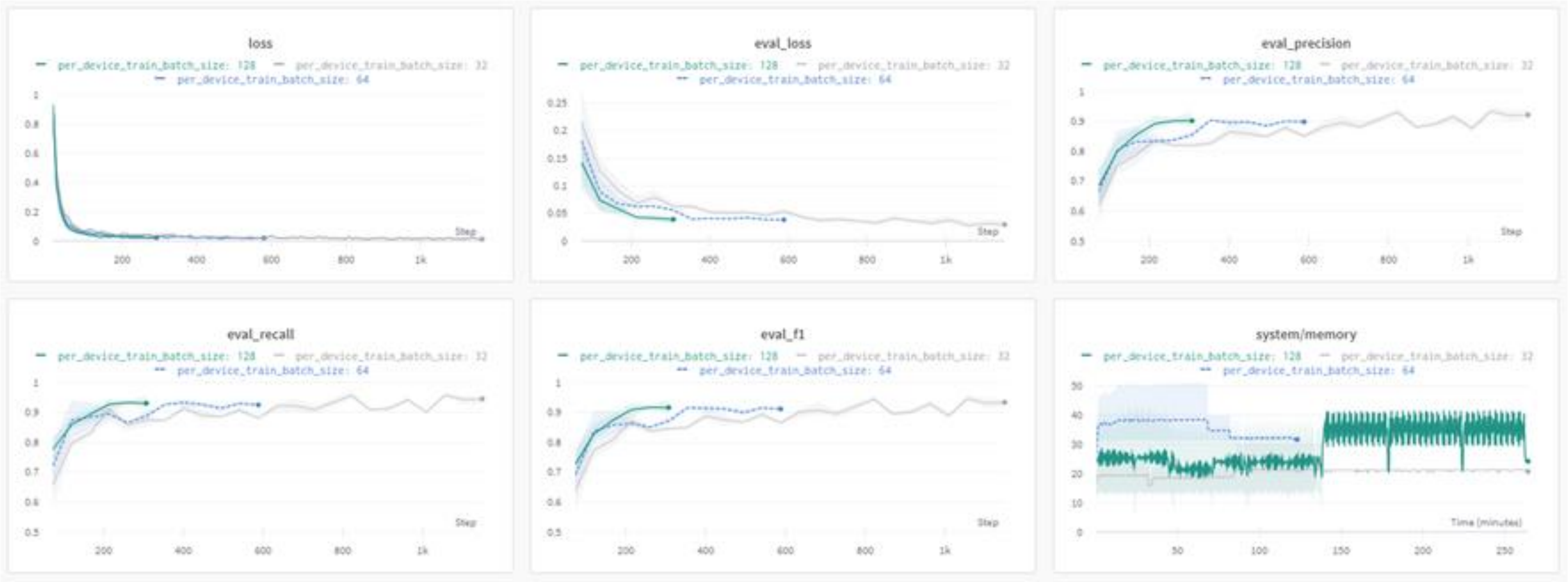
(5) **Reliable data:** Follow a regular supervision training, used smaller batch size and learning rates and train during more epochs

Training & Evaluation

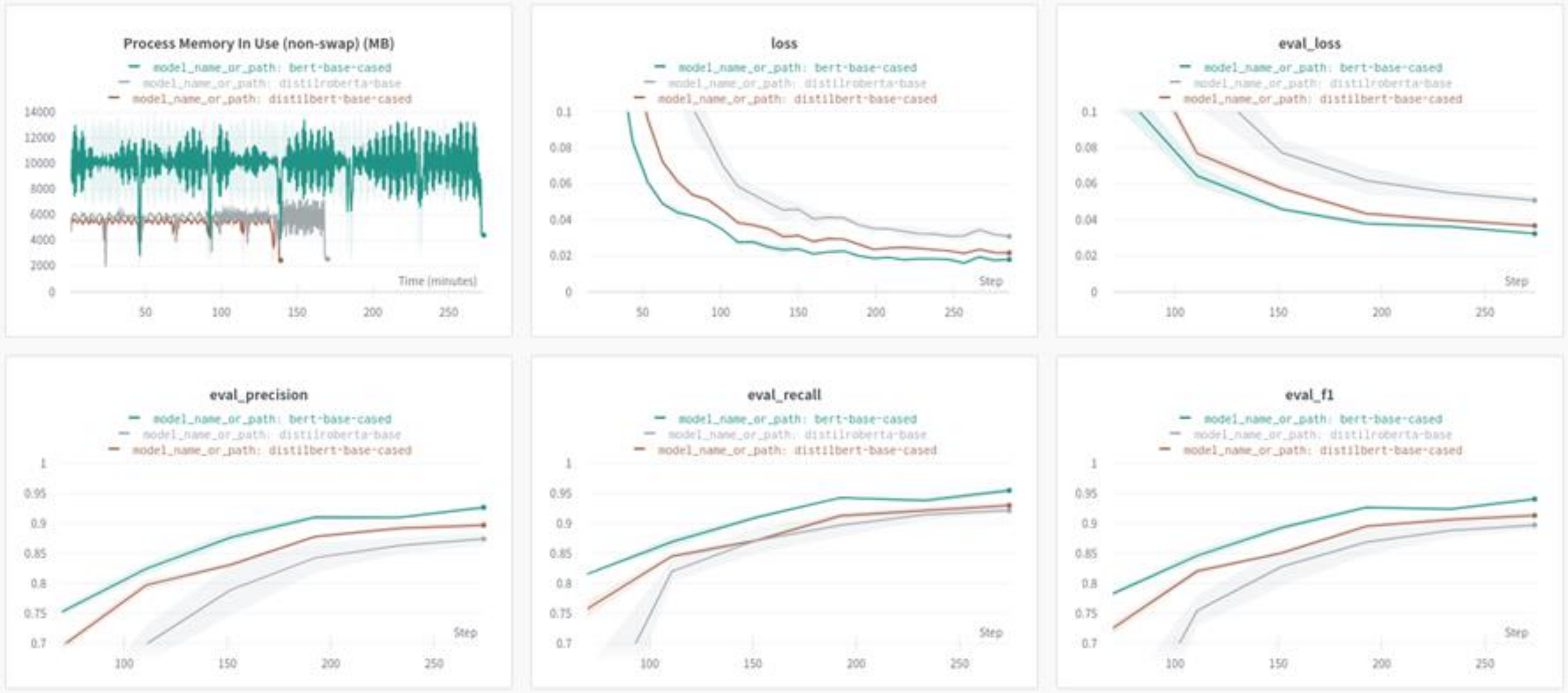
- BERT (Google), DistilBERT (Hugging Face) & DistilRoBERTa (Facebook, Hugging Face)
- We followed training recommendations to mitigate the noise effect [Rolnick 2017] [Abid 2020]
- We use CoNLL sequence evaluation (named-entity level)



Fine-tuning Step 1 – Use large Batch Size to mitigate noise effect



Fine-tuning Step 1 – Multiple runs shows consistent behavior



Bert Version	Named Entity	Step 1			Step 2			Single-Step			Supp
		P	R	F1	P	R	F1	P	R	F1	
	DEPTH_INT	0.99	0.99	0.99	0.98	0.98	0.98	0.96	0.98	0.97	92
	FORMATION	0.9	0.86	0.88	0.84	0.9	0.87	0.90	0.86	0.88	381
	WELL_ID	0.46	0.48	0.47	0.9	0.96	0.93	0.62	0.64	0.63	345
Distilled Bert	AGE	0.97	0.97	0.97	0.96	0.97	0.97	0.97	0.98	0.97	280
	PERIOD	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	166
	INTERVAL	0.92	0.97	0.95	0.93	0.97	0.95	0.93	0.96	0.94	189
	EPOCH	0.89	0.98	0.93	0.89	0.97	0.93	0.91	0.99	0.94	360
	Micro avg	**0.84**	**0.86**	**0.85**	**0.91**	**0.96**	**0.93**	**0.87**	**0.89**	**0.88**	**1813**
	DEPTH_INT	0.98	0.98	0.98	0.95	0.98	0.96	0.92	0.98	0.95	92
	FORMATION	0.92	0.87	0.89	0.85	0.91	0.88	0.90	0.87	0.89	381
	WELL_ID	0.46	0.48	0.47	0.91	0.96	0.94	0.72	0.74	0.73	345
Bert	AGE	0.96	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97	280
	PERIOD	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	166
	INTERVAL	0.93	0.95	0.94	0.92	0.96	0.94	0.95	0.96	0.96	189
	EPOCH	0.91	0.99	0.94	0.9	0.98	0.94	0.91	0.99	0.94	360
	Micro avg	**0.85**	**0.86**	**0.85**	**0.91**	**0.96**	**0.94**	**0.89**	**0.91**	**0.90**	**1813**

Table 4: Results for test set with a batch size of 64. During Step 1 the models are trained only with the noisy set, in Step 2 the resulted model from Step 1 is fine-tuned again using a small cleaned training set. Single-Step is the fine-tuned results training the models with the noisy and cleaned labels as a single training set

Performance – Improvements in fine-tuning step 2

Token	Finetuning St1	Finetuning St2
Well	B-WELL_ID	B-WELL_ID
13/22a	B-WELL_ID	I-WELL_ID
-	B-WELL_ID	I-WELL_ID
C29X	O	I-WELL_ID
Wellsite	O	O
Geological	O	O

....

Performance – Improvements in fine-tuning step 2

Token	Finetuning St1	Finetuning St2
Well	B-WELL_ID	B-WELL_ID
13/22a	B-WELL_ID	I-WELL_ID
-	B-WELL_ID	I-WELL_ID
C29X	O	I-WELL_ID
Wellsite	O	O
Geological	O	O

....

```
RESEARCH O
SHELL O
14/286 B-WELL_ID
- I-WELL_ID
2 O
biostratigraphy O
```

```
RESEARCH O
SHELL O
14/286 B-WELL_ID
- I-WELL_ID
2 I-WELL_ID
biostratigraphy O
```

Performance – Improvements in fine-tuning step 2

Token	Finetuning St1	Finetuning St2
Well	B-WELL_ID	B-WELL_ID
13/22a	B-WELL_ID	I-WELL_ID
-	B-WELL_ID	I-WELL_ID
C29X	O	I-WELL_ID
Wellsite	O	O
Geological	O	O
....		

```
RESEARCH O
SHELL O
14/286 B-WELL_ID
- I-WELL_ID
2 O
biostratigraphy O
```

```
RESEARCH O
SHELL O
14/286 B-WELL_ID
- I-WELL_ID
2 I-WELL_ID
biostratigraphy O
```

```
54.2 O
ft O
/ O
hr O
4 B-WELL_ID
- O
5-WT I-WELL_ID
- O
S O
```

```
54.2 O
ft O
/ O
hr O
4 B-WELL_ID
- I-WELL_ID
5-WT I-WELL_ID
- O
S O
```

Performance – Improvements in fine-tuning step 2

Token	Finetuning St1	Finetuning St2
Well	B-WELL_ID	B-WELL_ID
13/22a	B-WELL_ID	I-WELL_ID
-	B-WELL_ID	I-WELL_ID
C29X	O	I-WELL_ID
Wellsite	O	O
Geological	O	O
....		

```
RESEARCH O
SHELL O
14/286 B-WELL_ID
- I-WELL_ID
2 O
biostratigraphy O
```

```
RESEARCH O
SHELL O
14/286 B-WELL_ID
- I-WELL_ID
2 I-WELL_ID
biostratigraphy O
```

```
54.2 O
ft O
/ O
hr O
4 B-WELL_ID
- O
5-WT I-WELL_ID
- O
S O
```

```
54.2 O
ft O
/ O
hr O
4 B-WELL_ID
- I-WELL_ID
5-WT I-WELL_ID
- O
S O
```

```
Through O
Well O
Location O
13/21/ B-WELL_ID
12/25-N B-WELL_ID
( O
Projected O
) O
```

```
Through O
Well O
Location O
13/21/ B-WELL_ID
12/25-N I-WELL_ID
( O
Projected O
) O
```

Conclusions

- We **successfully** built a **Named Entity Recognition System for the Oil&Gas Industry**

- We built a **distributed NLP pipeline for weak data labelling**, extensible to **new named-entities** and suitable for other domains with similar characteristics

  - Our project implementation required new features in Spark NLP, now they are available in the open-source library

- We used a **two-step fine-tuning approach** that shows to be effective in **improving** the **prediction** capacity in hard-to-learn named entities. It also shows **promising results removing False Positives**.

Data modelling - LLM and graphs

- Explore how multiple prompting steps with human feedback can improve the query process.

    - push the feedback of the experts

    - construct a fine-tuning data integration and query answering approach.

Preparing data for Artificial Intelligence

- DataFrames integration
  - Tabular data structures that do not strictly belong to a schema or a database.

- Flexible framework for integrating and processing DataFrames across platforms like Spark, R, and Pandas
- Metamodel for representing relationships within distributed data sources
  - Schema evolution
  - Natural language query

Enhancing and enriching time series analysis – Carbon storage

- Reinforcement learning from human feedback (RLHF).
  - Integrating physical constraints into models
- Agentic Artificial Intelligence
  - How will autonomous systems interact with data?
  - Specialized petrophysical interpretations
  - Going back also to the report interpretation
  - Integrate human feedback
- PhD thesis will start in the next months

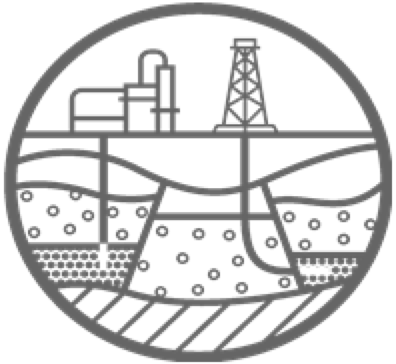
Data preparation and analysis for Time Series in the Energy Domain

- **Collaborators**: **PhD Molood Arman**, Yutao Chen, René Gómez Londoño, Sohaib Ouzineb, **PhD student Shwetha Salimath,** Nacéra Seghouani, Sylvain Wlodarczyk

- **Projects:** Proclaim, GeoTS

- **Papers:** CAiSE Forum 2020, DS 2022, KDD 2025, ADBIS 2025

Data preparation and analysis for Time Series in the Energy Domain

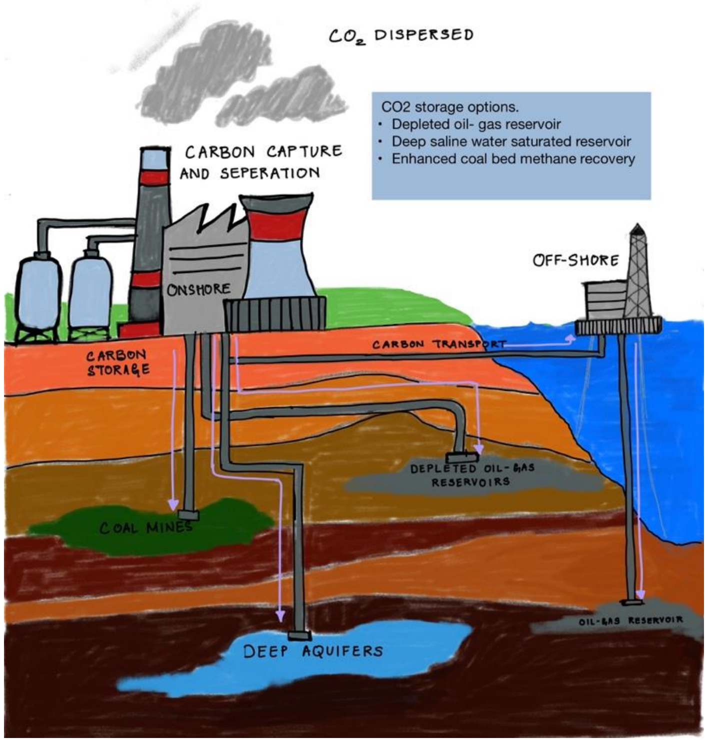


Research problem: Carbon Capture Storage

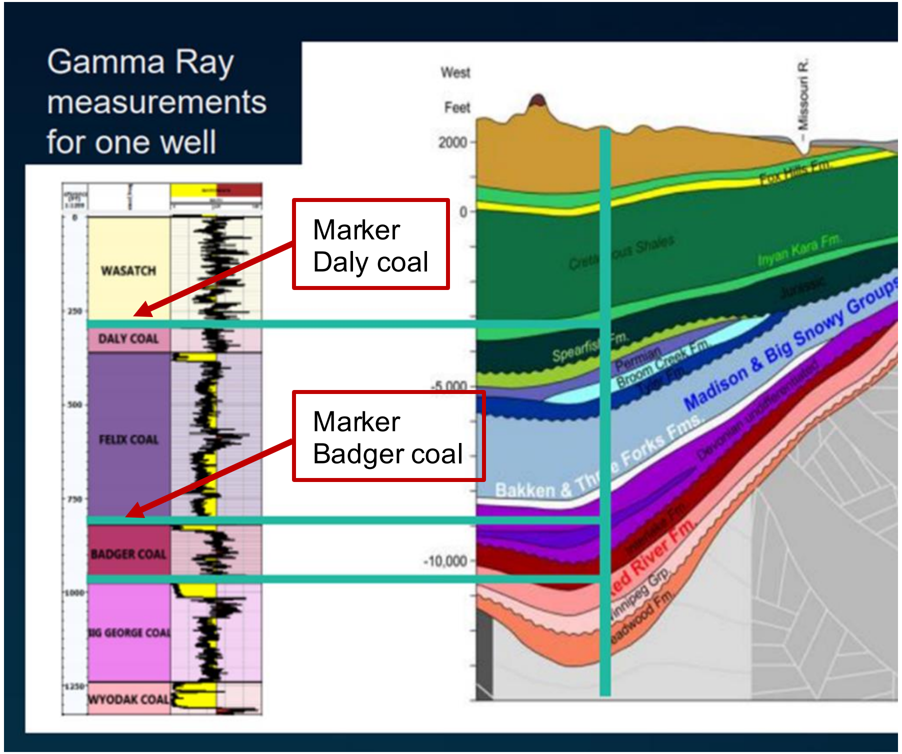
- CCS involves capturing CO2, transporting it, and storing it in deep geological formations to prevent it from entering the atmosphere

- Reassessment of seal integration and storage potential

- Geological analysis and monitoring by studying subsurface rock properties and correlating formations for accurate reservoir modeling

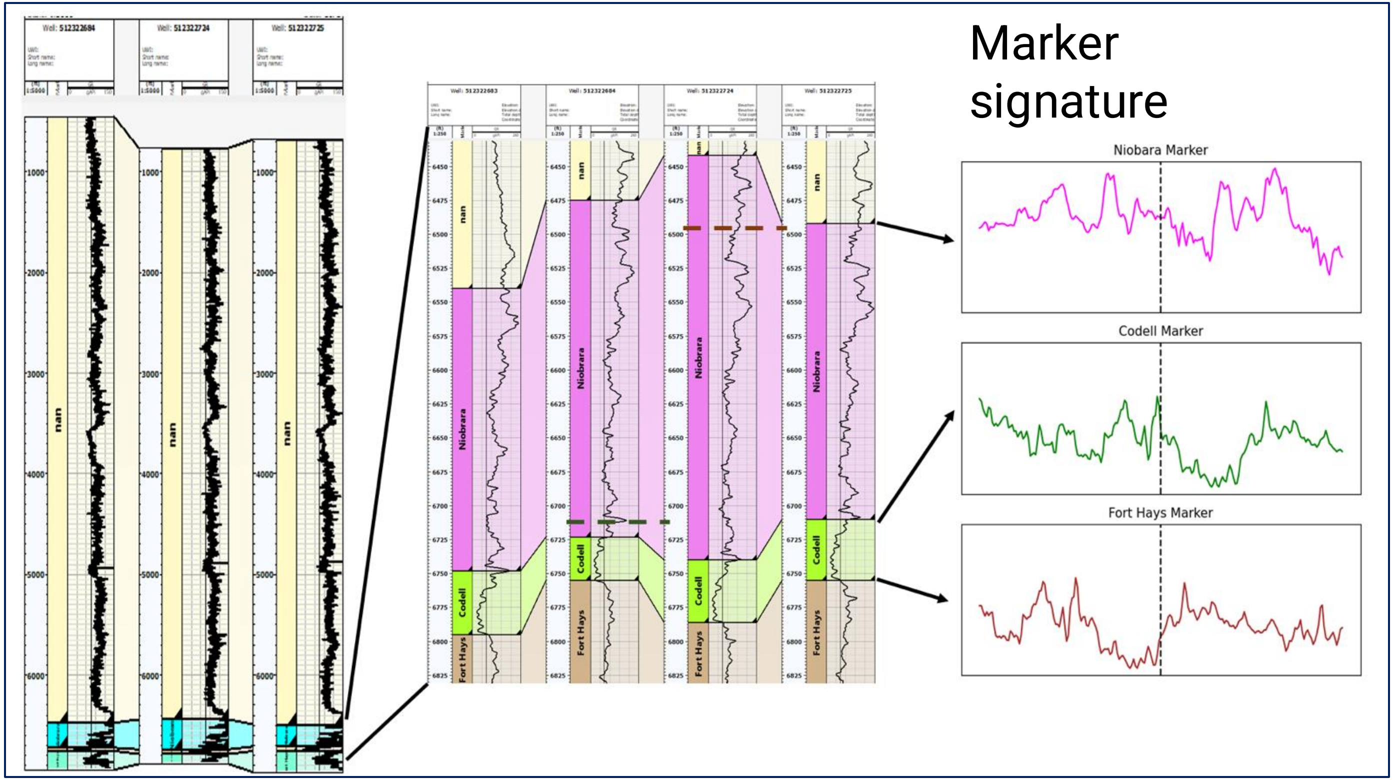


Problem Statement



- Geologists use mud logs and the rocks extracted during borehole drilling to study formation characteristics
- Tedious and time-consuming
- Finding an efficient way to extract information from wireline logs using deep learning would save time and resources

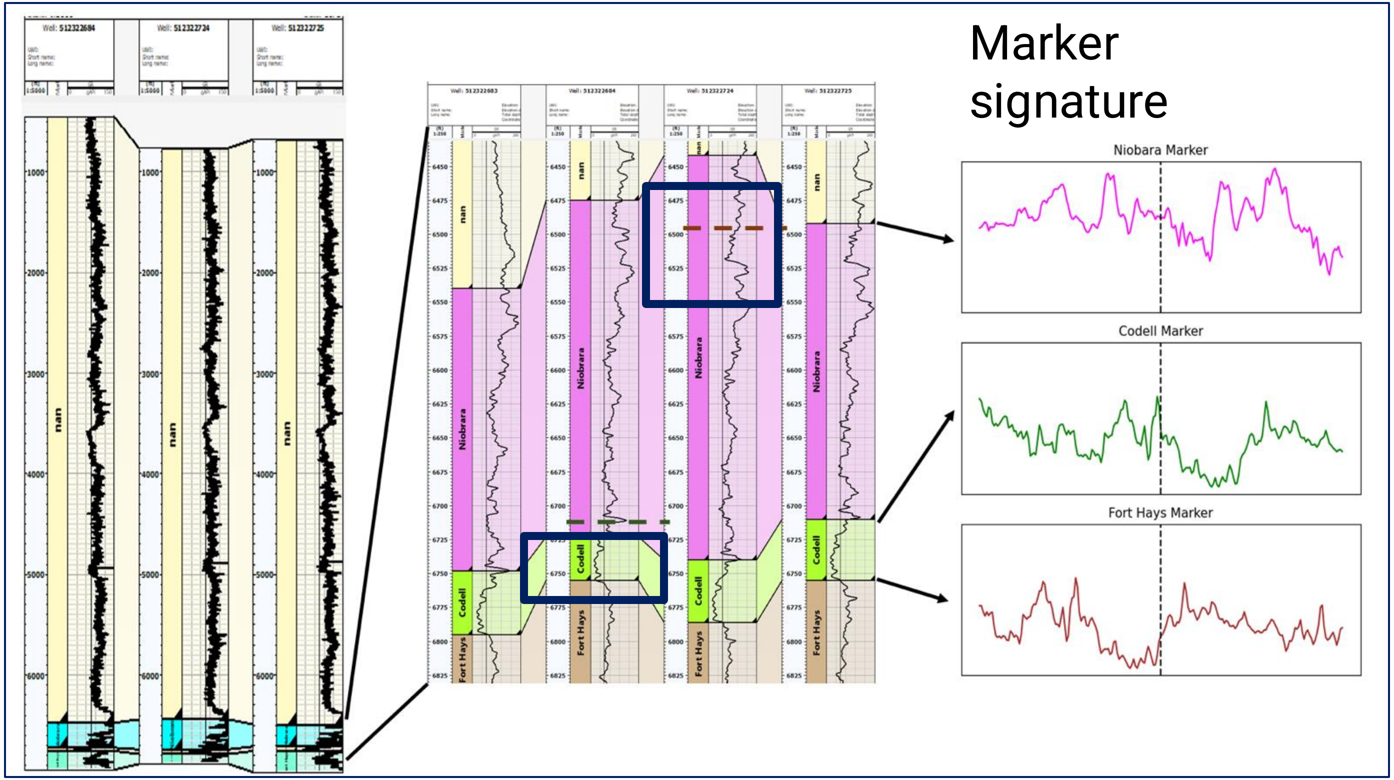
Well log Data - a lot and heterogeneous time series



LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

CentraleSupélec

Well log Data - a lot and heterogeneous time series



LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
CentraleSupélec

Problem

- **Well Correlation**
    - Industrial baseline with dynamic time warping distance (DTW).
    - Minimum spanning tree to find pairs and then DTW.
    - Autoencoders and bidirectional LSTM for correlating neighboring wells.

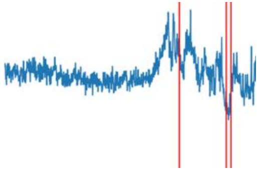
Challenges with DTW for well log data

- Bad alignment of the wells and local shifts in marker signatures
- Depth incoherent signature pattern
- Each marker prediction is independent of the other
- Since only one marker can be processed at a time, it is a time-consuming process

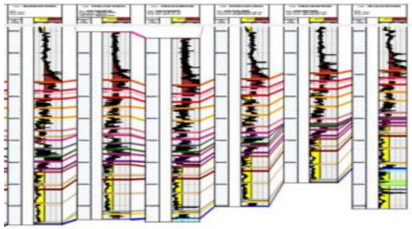
Dataset

USGS science for a changing world

WellDatabase

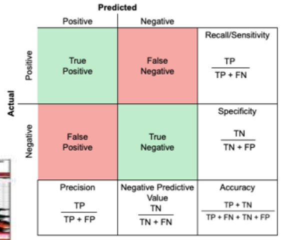


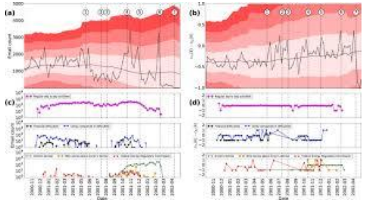
GeoTS



Evaluation

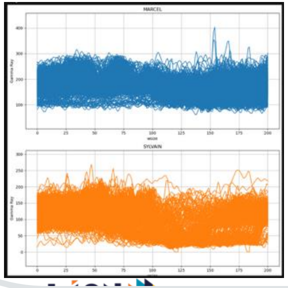
Accuracy, Recall and Visualizations

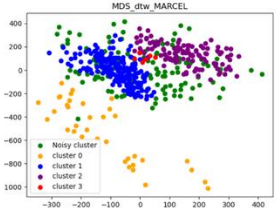




Data Processing

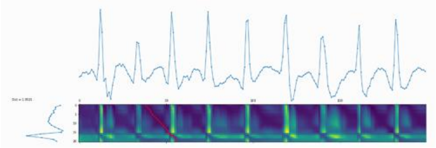
Cleaning, Clustering and Template extraction

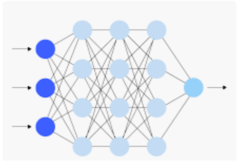




Models

Industrial baseline and deep learning models





Data Processing

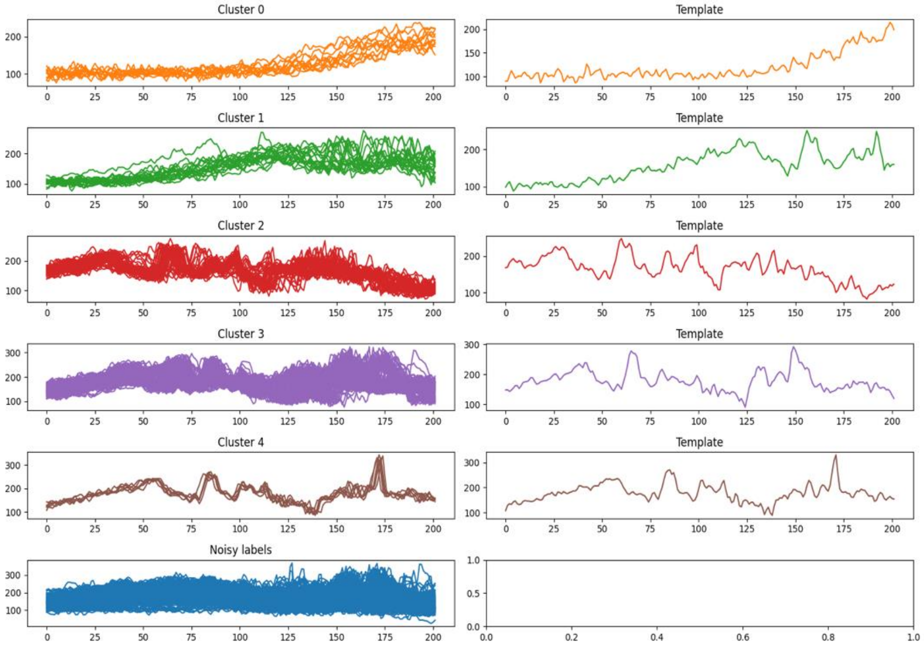
- **Signature Extraction**: This step involves extracting the signature of a formation from the training log data with a specified window size

- **Clustering**: The DTW distance matrix containing the DTW distances between all pairs of extracted signatures is used for clustering

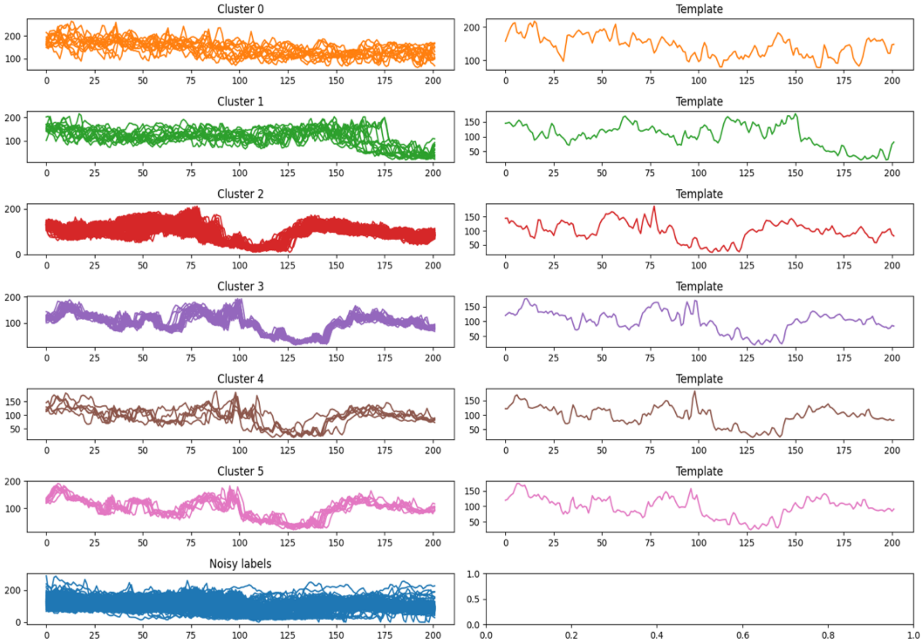
- **HDBSCAN** clustering algorithm is used. We analyze signature templates representing a cluster of similar signatures for a particular formation

Clustering result

Niobrara

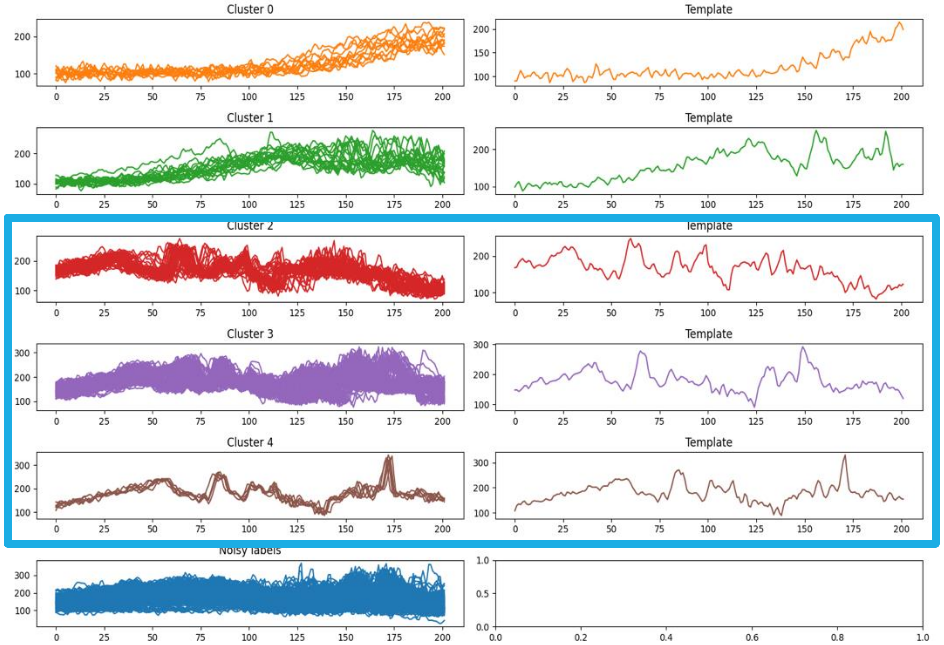


Codell

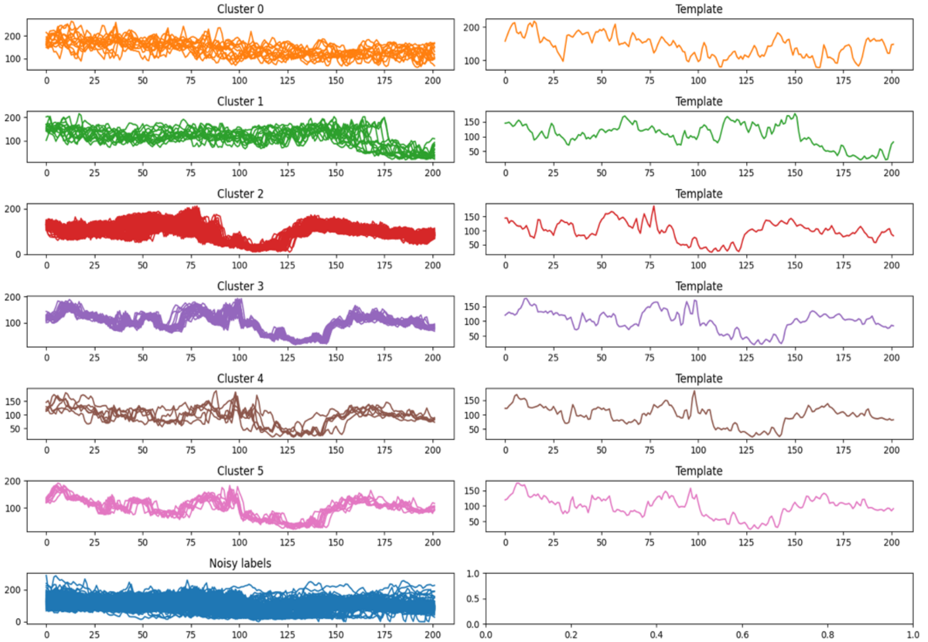


Clustering result

Niobrara

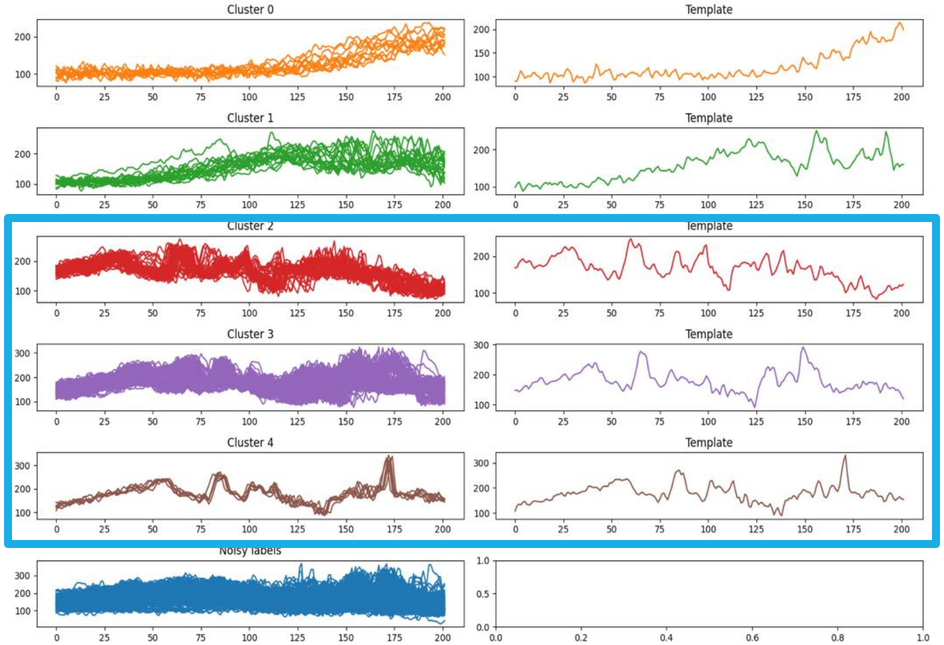


Codell

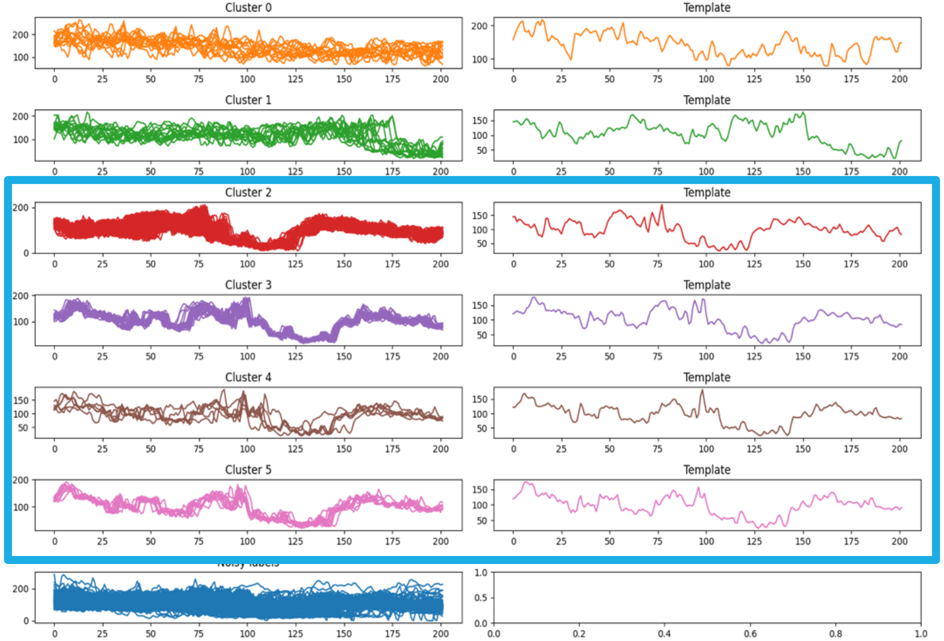


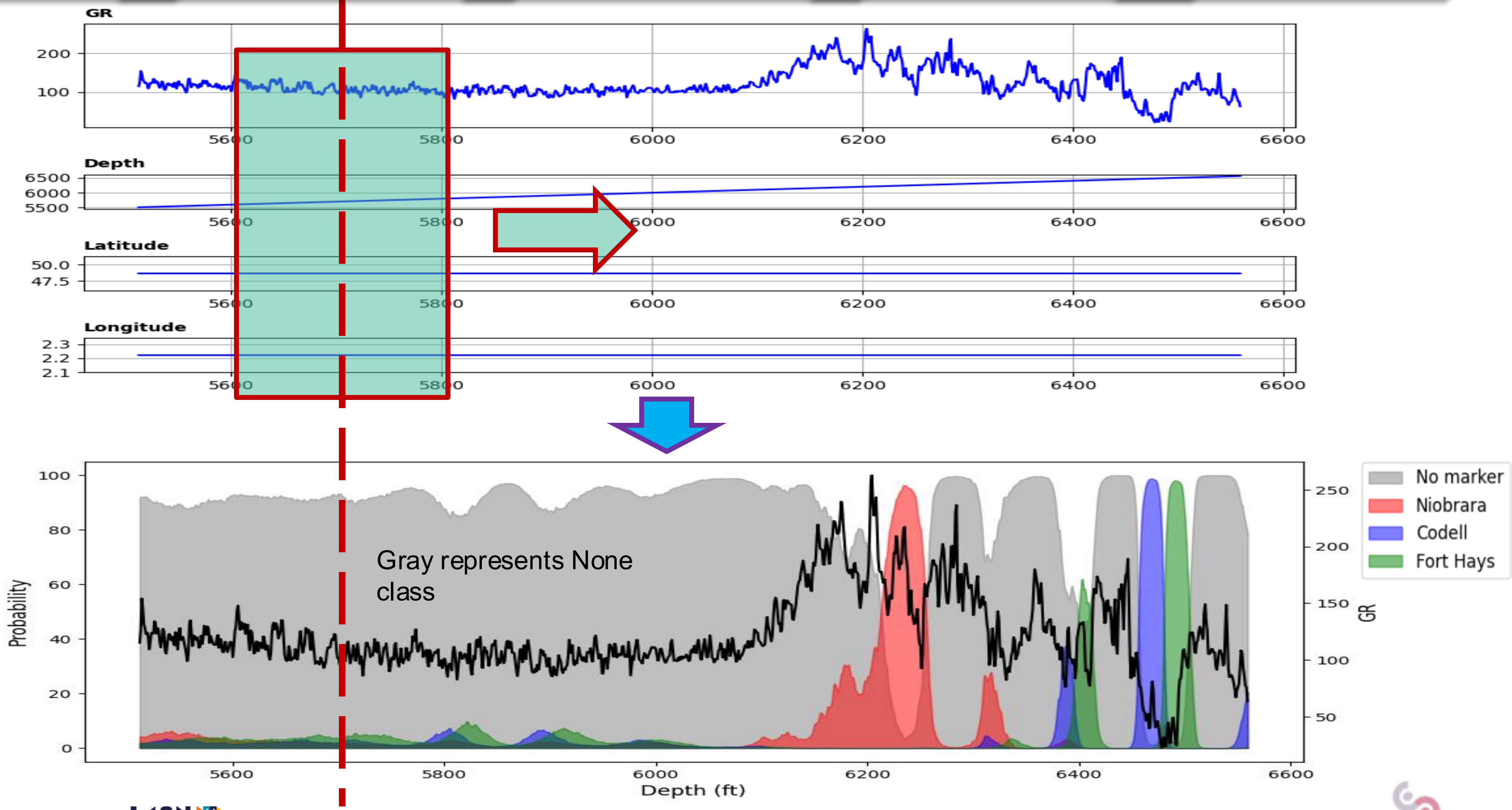
Clustering result

Niobrara



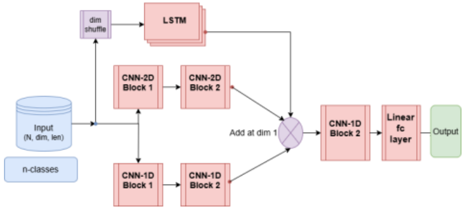
Codell

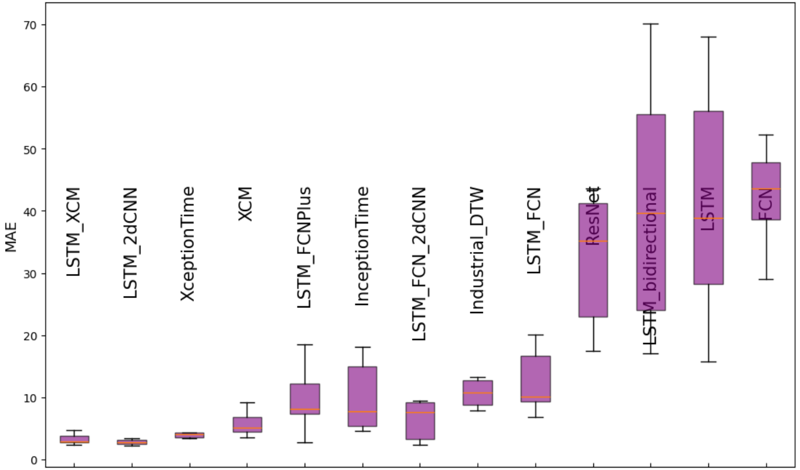


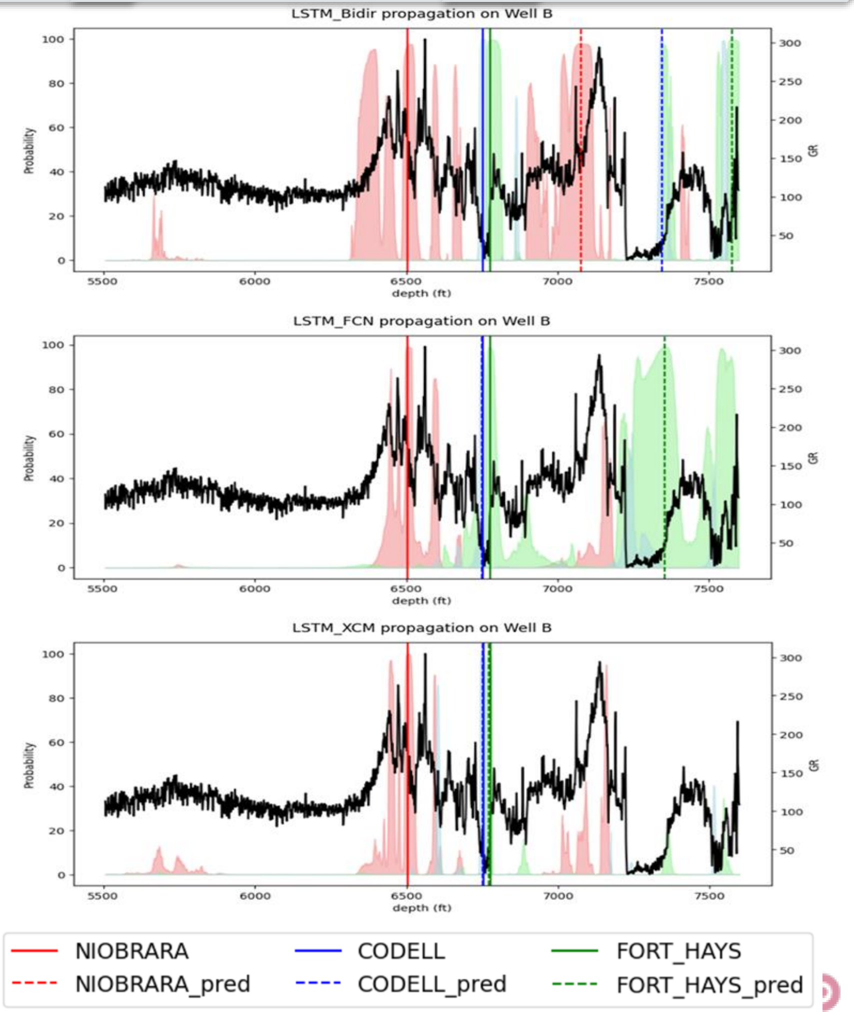

GR
Depth
Latitude
Longitude
Gray represents None class
No marker
Niobrara
Codell
Fort Hays
Probability
GR
Depth (ft)

Maximum absolute error

LSTM-XCM



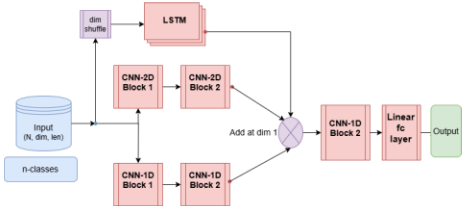


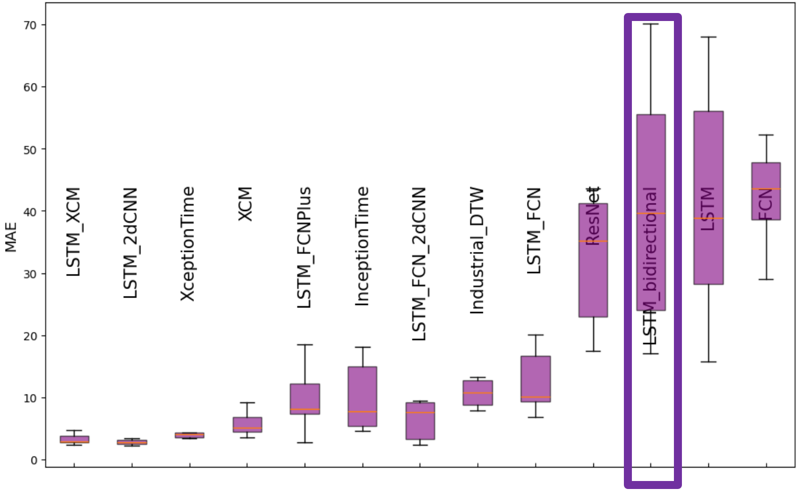


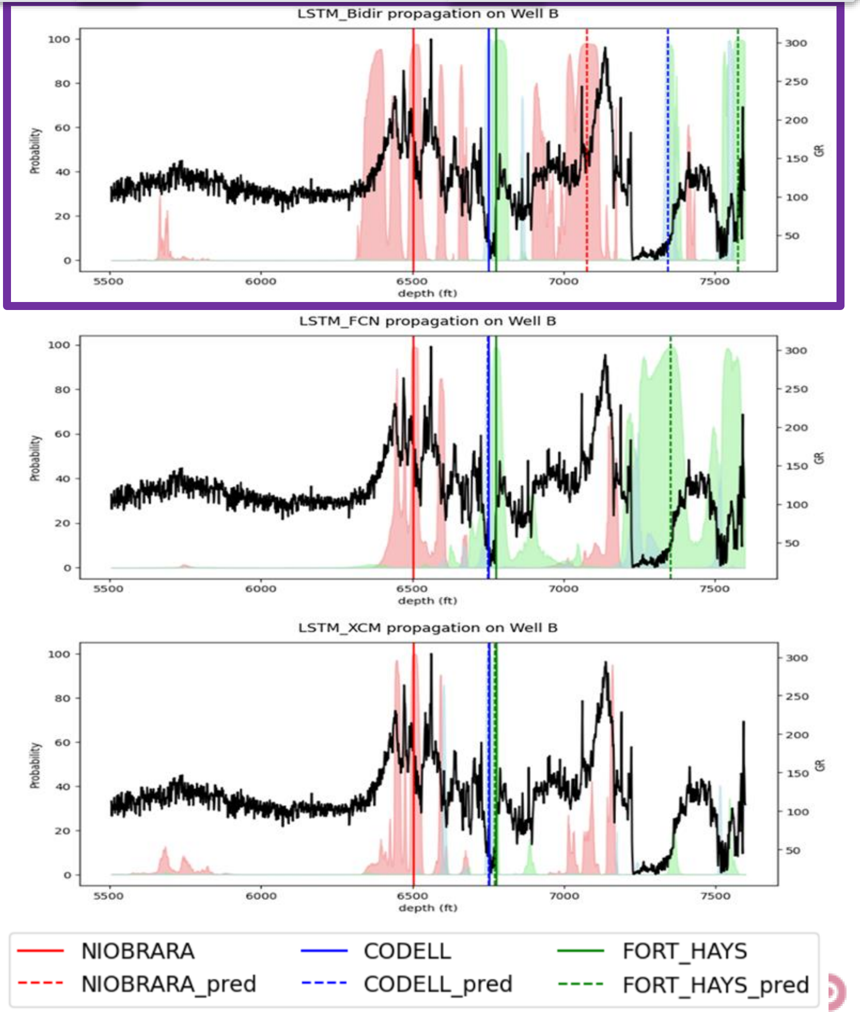
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
CentraleSupélec

Maximum absolute error

LSTM-XCM

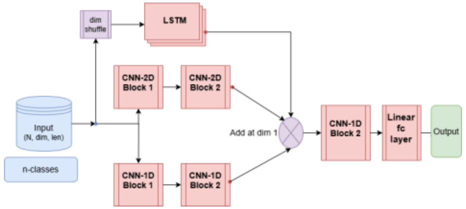


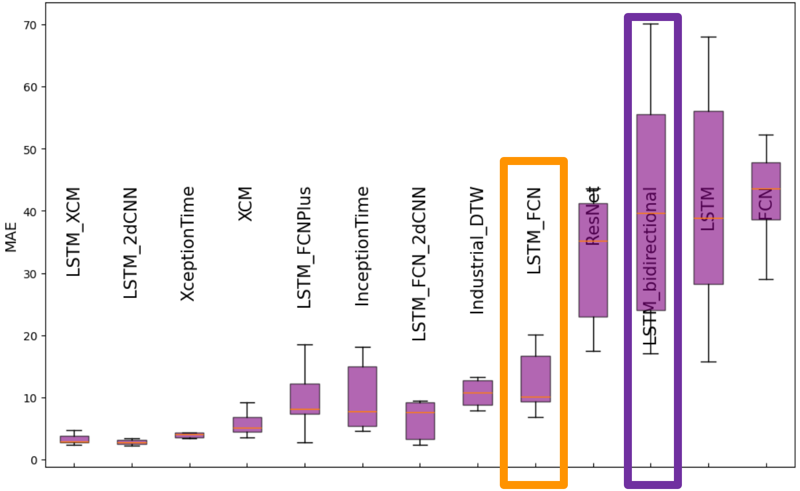


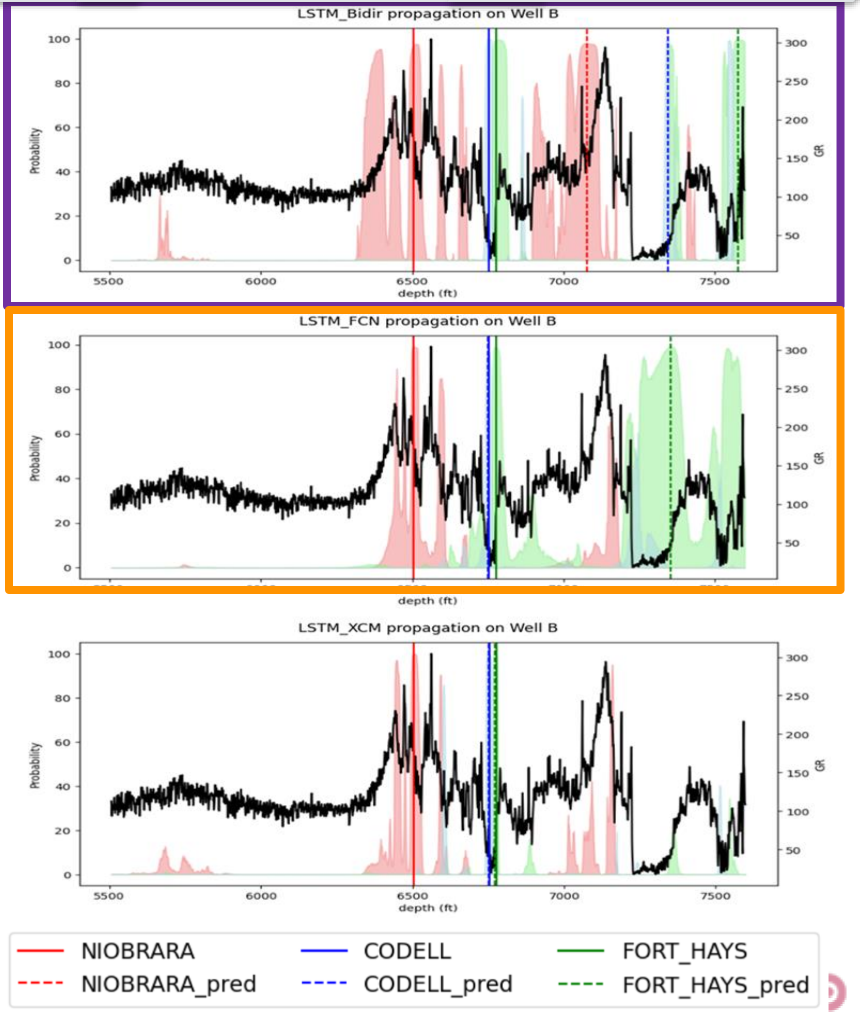


Maximum absolute error

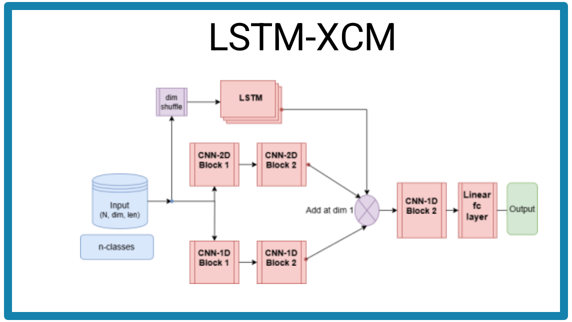
LSTM-XCM

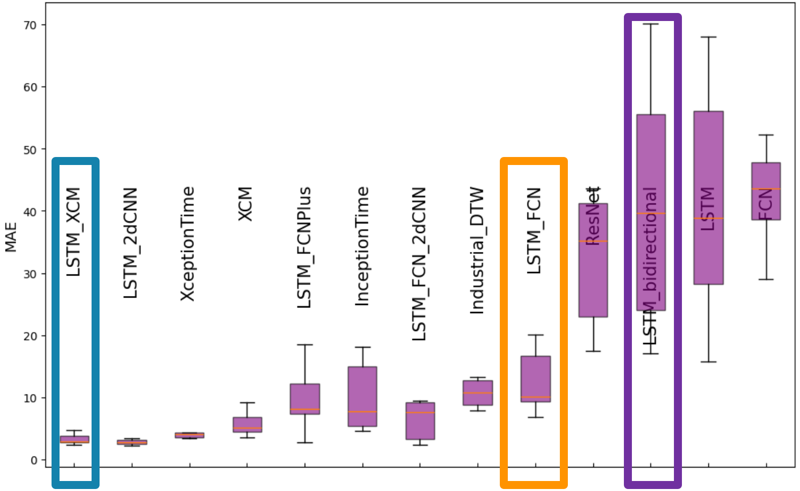


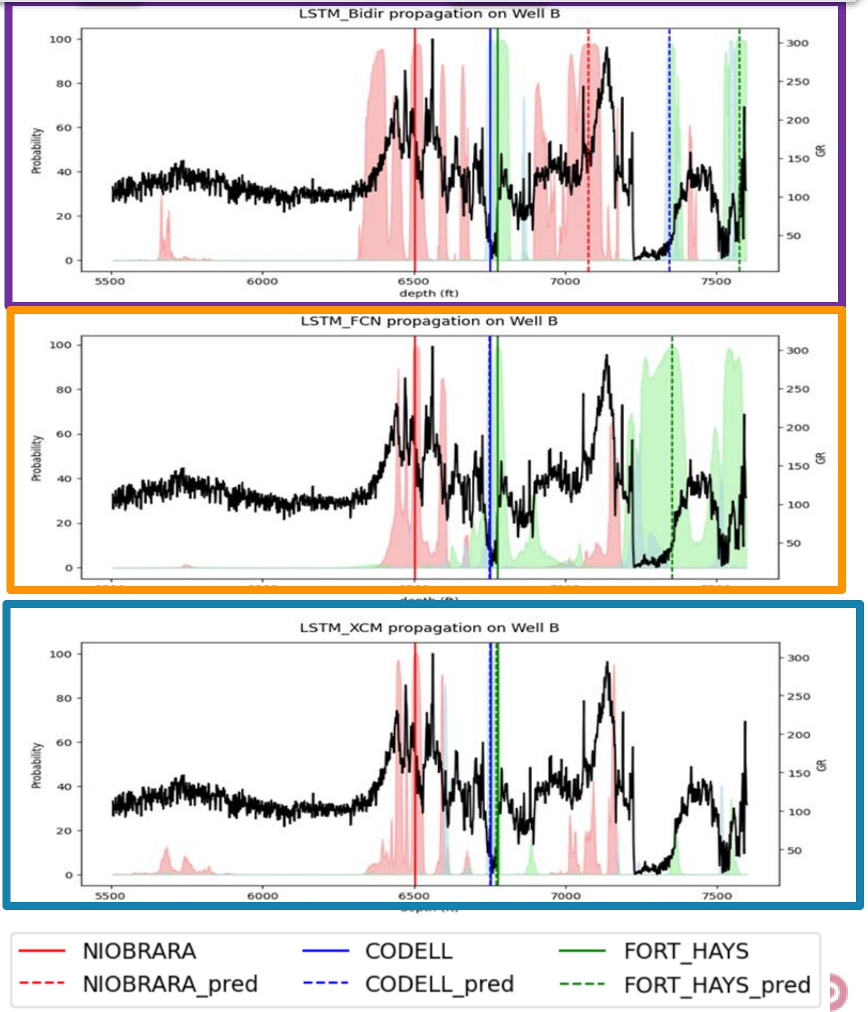




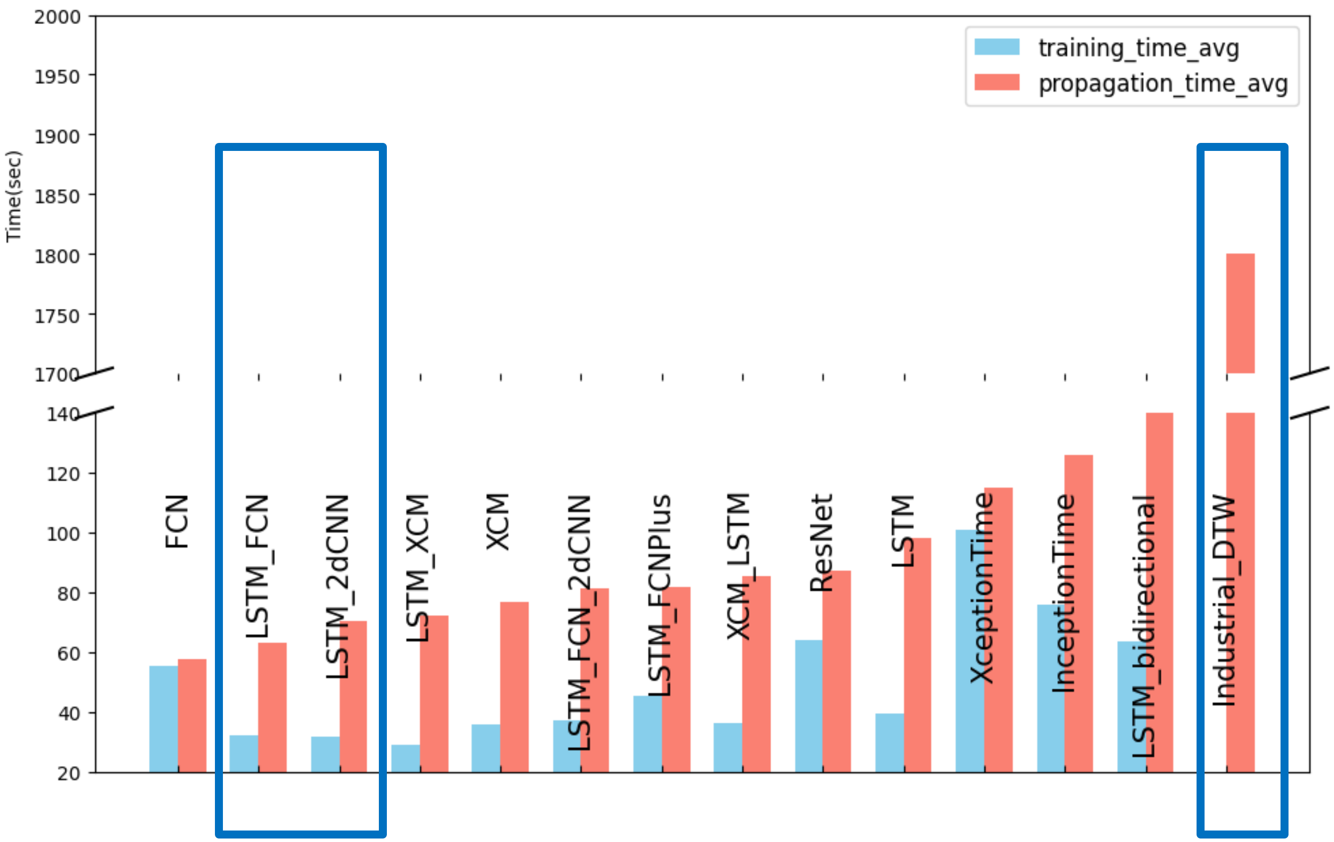
Maximum absolute error







Time Efficiency

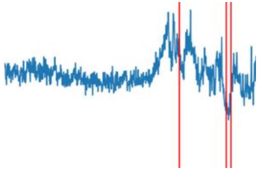


Enhancing and enriching time series analysis

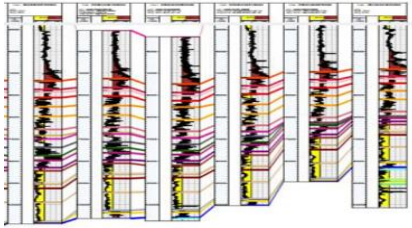
Dataset

USGS science for a changing world

WellDatabase

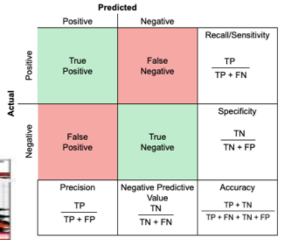


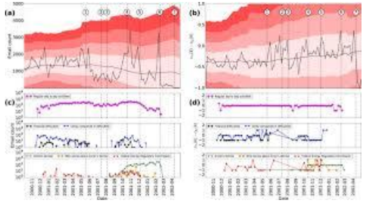
GeoTS



Evaluation

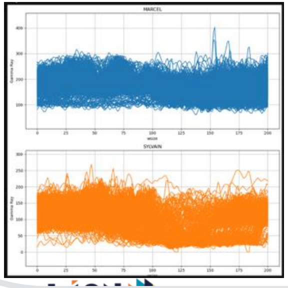
Accuracy, Recall and Visualizations

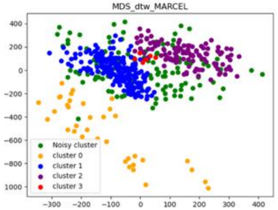




Data Processing

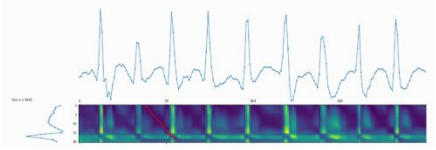
Cleaning, Clustering and Template extraction

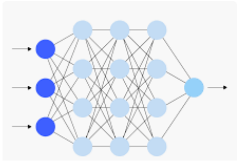




Models

Industrial baseline and deep learning models





Problem context - Reports/surveys

Experts require structure and unstructured data for theoretical data control and analysis

- **Data acquisition**
  - Wells drilled a long time ago with historical log data
  - Different tools/sensors from different service providers
  - Well sample analysis described in reports
- **Data assessment**
  - Data quality and Interpretation done manually by petrophysicists/geologists based on reports

- Retrieval-Augmented Generation (RAG) techniques
- Automate the process by exploring agentic RAGs

2. Stratigraphy an... leoenvironment Results

2.1 Cenozoic

2.1.1 Pleistocene to Pliocene
850 to 970 feet (thickness more than 120 feet)

No samples were available from the interval between sea bottom and 850 feet.

Paleontology

The benthonic foraminiferal assemblage contains mainly species which are at present still living; typical Pliocene forms are nearly absent (only single specimens of Cassidulina cf. pliocarinata and Cibicides lobatulus grossa were found, which could be reworked). This microfauna suggests a Pleistocene or uppermost Pliocene age for these deposits.

Paleoenvironment

The benthonic foraminiferal assemblage, the near absence of planktonic foraminifera and the occurrence of frequent shell fragments suggest shallow marine (inner neritic) environment.

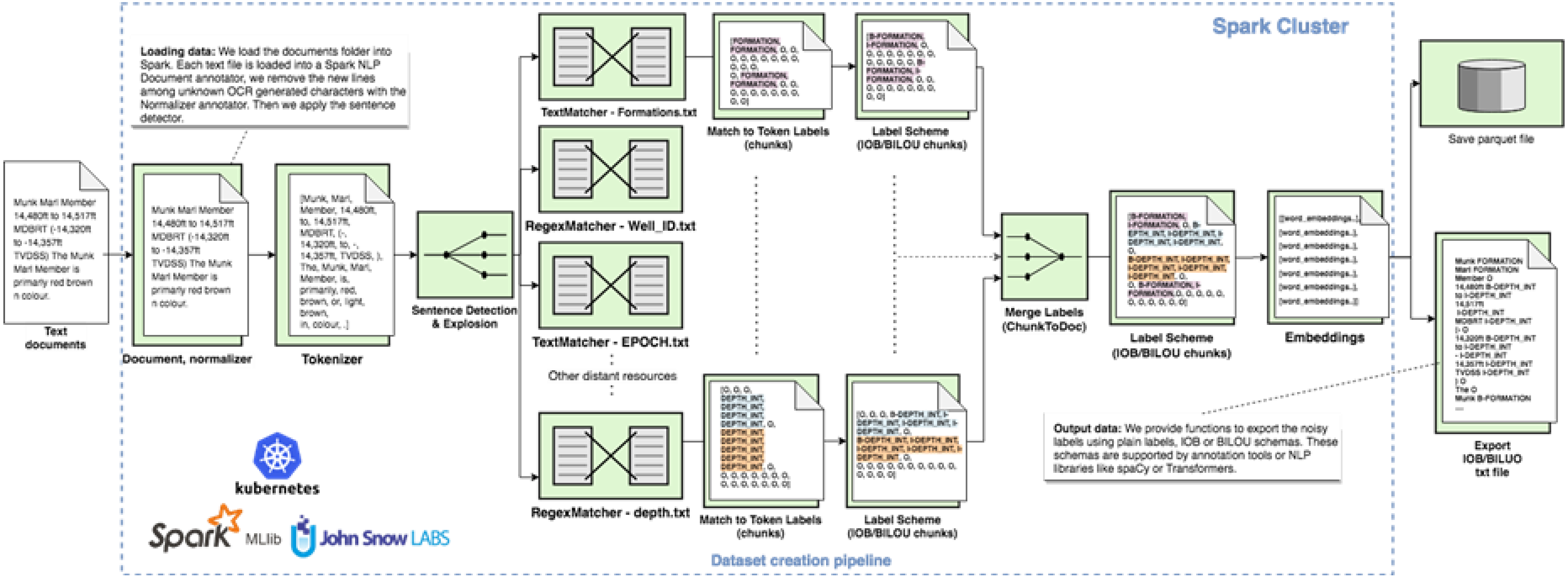
```
In [33]: data
```

Out[33]:

	Area Abbreviation	Area Code	Area	Item Code	Item	Element Code	Element	Unit	latitude	longitude	...	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009
0	AF	2	Afghanistan	2511	Wheat and products	5142	Food	1000 tonnes	33.94	67.71	...	3249.0	3486.0	3704.0	4164.0	4252.0	4538.0
1	AF	2	Afghanistan	2805	Rice (Milled Equivalent)	5142	Food	1000 tonnes	33.94	67.71	...	419.0	445.0	546.0	455.0	490.0	415.0
2	AF	2	Afghanistan	2513	Barley and products	5521	Feed	1000 tonnes	33.94	67.71	...	58.0	236.0	262.0	263.0	230.0	379.0
3	AF	2	Afghanistan	2513	Barley and products	5142	Food	1000 tonnes	33.94	67.71	...	185.0	43.0	44.0	48.0	62.0	55.0
4	AF	2	Afghanistan	2514	Maize and products	5521	Feed	1000 tonnes	33.94	67.71	...	120.0	208.0	233.0	249.0	247.0	195.0
5	AF	2	Afghanistan	2514	Maize and products	5142	Food	1000 tonnes	33.94	67.71	...	231.0	67.0	82.0	67.0	69.0	71.0
6	AF	2	Afghanistan	2517	Millet and products	5142	Food	1000 tonnes	33.94	67.71	...	15.0	21.0	11.0	19.0	21.0	18.0
7	AF	2	Afghanistan	2520	Cereals, Other	5142	Food	1000 tonnes	33.94	67.71	...	2.0	1.0	1.0	0.0	0.0	0.0
8	AF	2	Afghanistan	2531	Potatoes and products	5142	Food	1000 tonnes	33.94	67.71	...	276.0	294.0	294.0	260.0	242.0	250.0
9	AF	2	Afghanistan	2536	Sugar cane	5521	Feed	1000 tonnes	33.94	67.71	...	50.0	29.0	61.0	65.0	54.0	114.0
10	AF	2	Afghanistan	2537	Sugar beet	5521	Feed	1000 tonnes	33.94	67.71	...	0.0	0.0	0.0	0.0	0.0	0.0

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
CentraleSupélec

Dataset creation – First temptative



Enhancing and enriching time series analysis

- Reinforcement learning from human feedback (RLHF)
  - Integrating physical constraints into models

Enhancing and enriching time series analysis

- Agentic Artificial Intelligence
  - How will autonomous systems interact with data?
  - Specialized petrophysical interpretations
  - Going back also to the report interpretation
  - Integrate human continuous feedback

- CIFRE PhD thesis will start in the next months

Main areas and contributions

- Metamodel data integration
  - **Papers**: Linked Data Management 2022, DEXA 2020, ER 2018, CIDR 2015, ER 2014, EDBT 2013
- Graph Data Integration and Large Language Models
  - **Papers:** BigData 2023, J. Glob. Inf. Manag 2023, DKE 2024
  - **New financed project**
- Data preparation and analysis for Time Series in the Energy Domain
  - **Papers:** CAiSE Forum 2020, DS 2022, KDD 2025, ADBIS 2025
  - **New financed thesis to explore agentic AI**