Valorisation de connaissances internes et gain de performances Projet Innosuisse Collective Memory

Hatem Ghorbel - AI-days 2024- Lausanne

- PICC Solution SA
- Haute école Arc Ingénierie
- Haute école de gestion Arc (HEG Arc)

Certifié ISO 9001

Projet Innosuisse Collective Memory: Transforming intelligence to actionnable knowledge

- 2023-2024
- PICC Solution SA
- Haute école Arc Ingénierie
- Haute école de gestion Arc (HEG Arc)

Problem: knowledge management

"Managing organizational knowledge effectively is a prerequisite for securing competitive advantages in the global marketplace",
V. D. Kalkan

Problem

- Knowledge workers spend one working month each year searching for the right information
- Employee turnover causes permanent loss of unrecorded knowledge
- Existing solutions fail to keep people motivated to share and update knowledge
- 90% of employees want better opportunities to share knowledge

- **122 million knowledge workers recreate existing solutions on a daily basis**
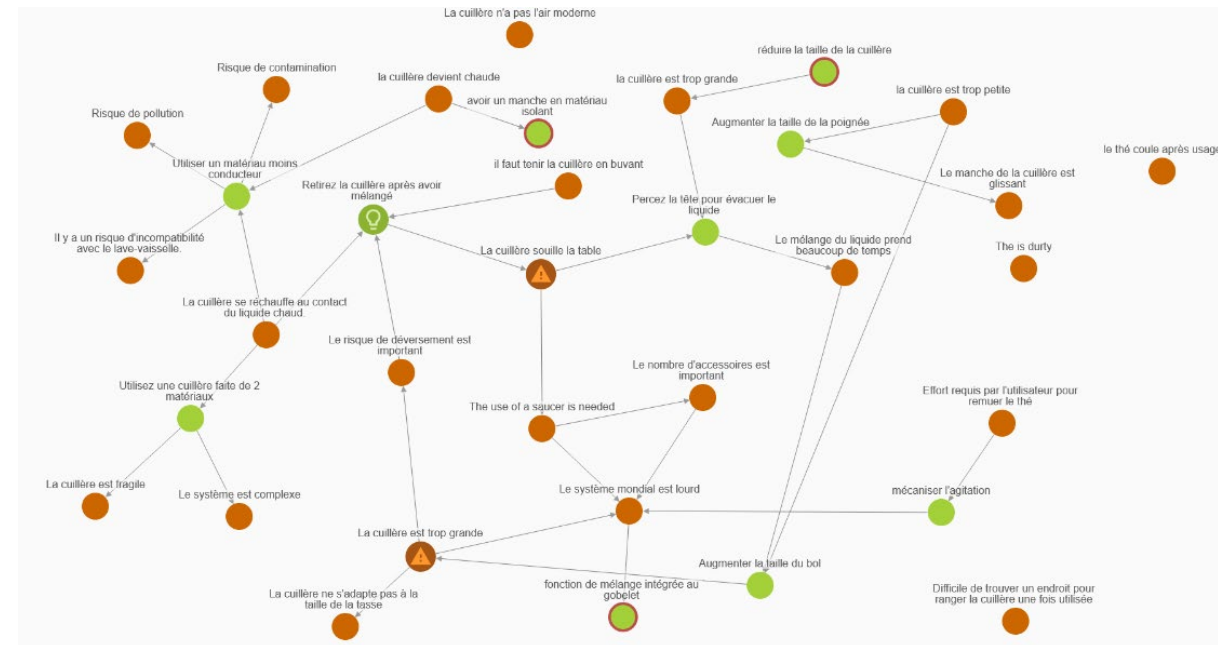- **Knowledge management comes up short in 82% of cases according a study recently conducted**
- Ref : https://dzone.com/   Ref : https://www.starmind.ai/ Ref : https://www2.deloitte.com/
  Ref : OECD

PICC Solution

Actionnable knowledge

Knowledge modeling

Textual segment

Problem | Solution | Action | Invalid

Contamination par des particules de poussière métallique dans l'air

Les sacs peuvent être contaminés par la poussière lorsqu'ils sont posés sur le sol.

Les capsules sont contaminées par la poussière

Risque de contamination des produits finis

Installation de la boîte AMG avant le conditionnement

Les capuchons de décharge augmentent le risque de chocs électriques pour les opérateurs en raison de la contamination par les poussières.

Knowledge modeling: Problem-solution structure

- **A problem** is seen as a circumstance or state of affairs that presents a challenge or difficulty. It signifies an obstacle or an aspect of a situation that hinders the realization of a desired outcome.

- **A solution** is a means or method proposed to address and resolve a problem. Solutions are proactive measures intended to bring about (positive) change, alleviate issues, or prevent the recurrence of problems.

- **An action** is defined as an activity, something that has been done or performed and that is neither a solution nor a problem.

Knowledge decomposition

Problem	Solution
"The destruction of forests not only diminishes crucial habitats for countless species but also reduces the planet's capacity to absorb carbon dioxide, exacerbating the effects of climate change."	→ "Implementing sustainable land management practices and undertaking large-scale reforestation efforts are critical to counteract deforestation." → "This includes promoting responsible logging practices, protecting existing forests, and restoring degraded areas to enhance biodiversity and carbon sequestration."

Methodology: Annotation guideline

- **Problem** : automatique Extraction of problems/solutions from textual spans?
- **Solution**: Machine learning approach

- **Methodology**

Open Source

Data Labeling

Label Studio

1. **Manual annotation**

- Make the annotation as objective as possible: annotation guideline
  - Two annotators annotate a same set of texts
  - Annotation agreement should go beyond 80% (kappa ratio)

➡ A golden stardard dataset for ML is ready

Human challenges

- Human annotation / correction of weak labels
  - Subjectivity
    - annotations guidelines
    - Inter-annotation agreement
    - Golden stardard
  - Motivation
    - Gamification
    - User profiling
    - Notification
    - User experience (UX)

Scientific challenge

2. Text extractio nfrom documents (PDF, word, ppt, XLS, HTML)
3. Text segmentation
    – Sentences, ok but very short/long sentences…
4. Text categorizations
    – Classification into problem, solution, action, invalid
    – Classical NLP tasks: NER, topic modeling, quantifier extraction
5. Problem solution matching
    – A problem could generate an other problem
    – A solution could match to more than one problem
6. Domain specificity
    – Warm start using transfert learning (LLM)
    – Adaptation by iteration using active learning

Module IA NLP

Texte Extraction
Word, PowerPoint, PDF

Segmentation

Classification
problème/solution
✔ Bert Classifier

NER
Spacy
✔

Topic Modeling
BERTopic
✔

Extraction de Quantifiables
✔ Quantulum3

Matching problèmes et Solutions
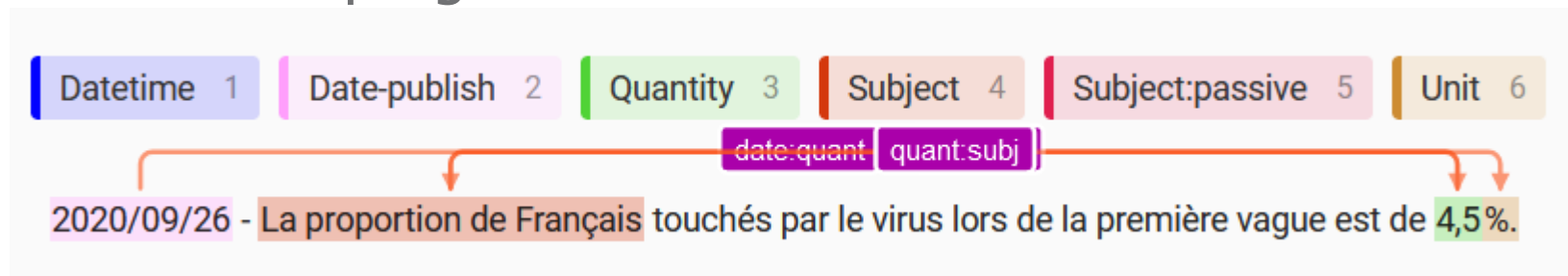
Active learning

Problem/solution classification

- A first ML model is developped to classify a sentence in 3 classes
    - Problem
    - Solution
    - Invalid
- Dataset:

    - *Already segmented sentences*

    - *Dataset of 4000 annotated sentences*
- Embeddings using XML-RoBERTa (multilingual model)
- Fine-tuned with additional dense network layer
- F1_score : 82% on test-unseen data

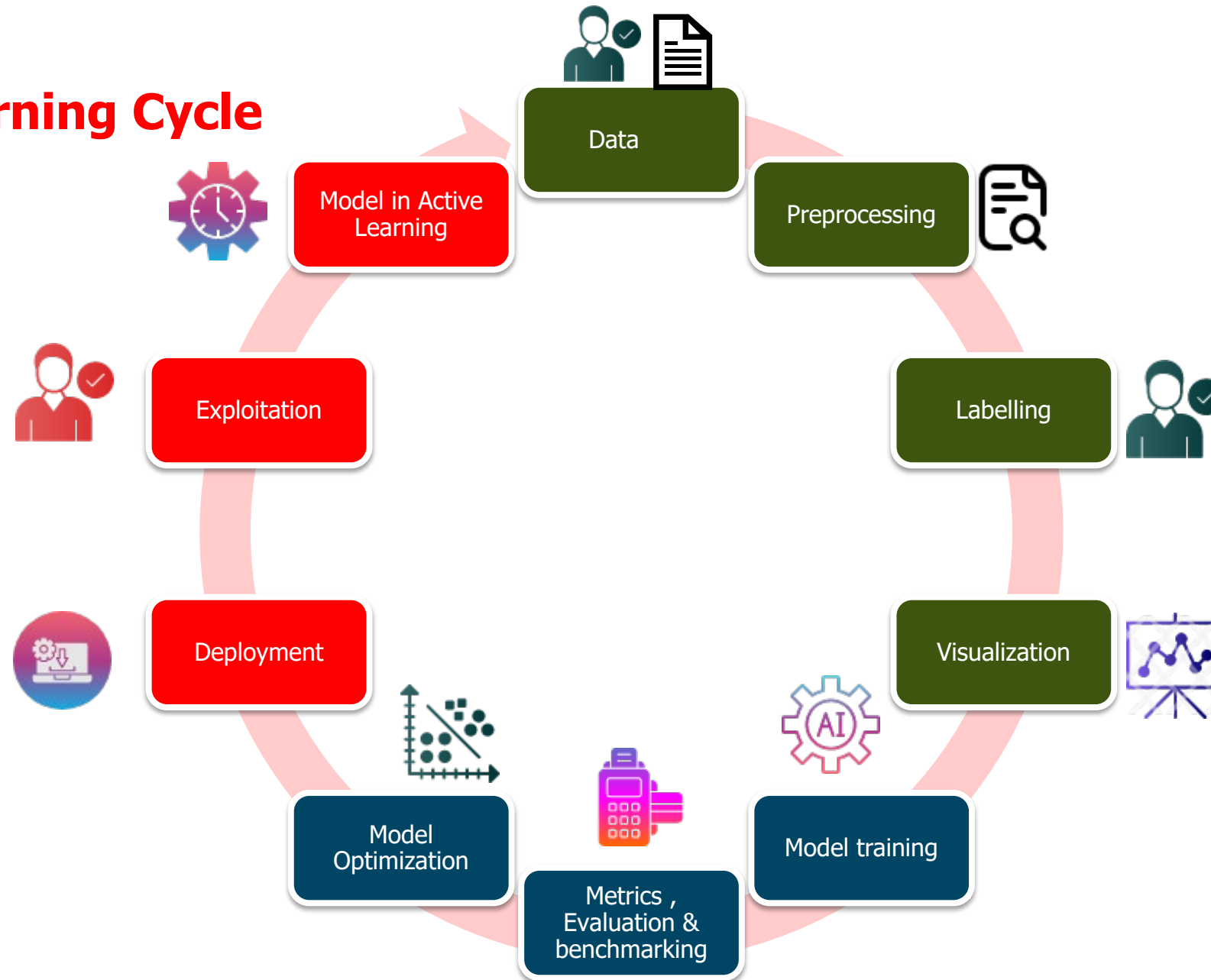# Extraction of quantifiers

- Using **quantulum3** a Python library for information extraction of quantities, measurements and their units from unstructured text.

  - based on their k-nearest neighbours in their GloVe vector

    >>> parser.parse('Sound travels at 0.34 km/s')
      Unit="kilometre per second"
      Entity="sound speed"
      Quantity = 0.34

- Multilingual solution in progress
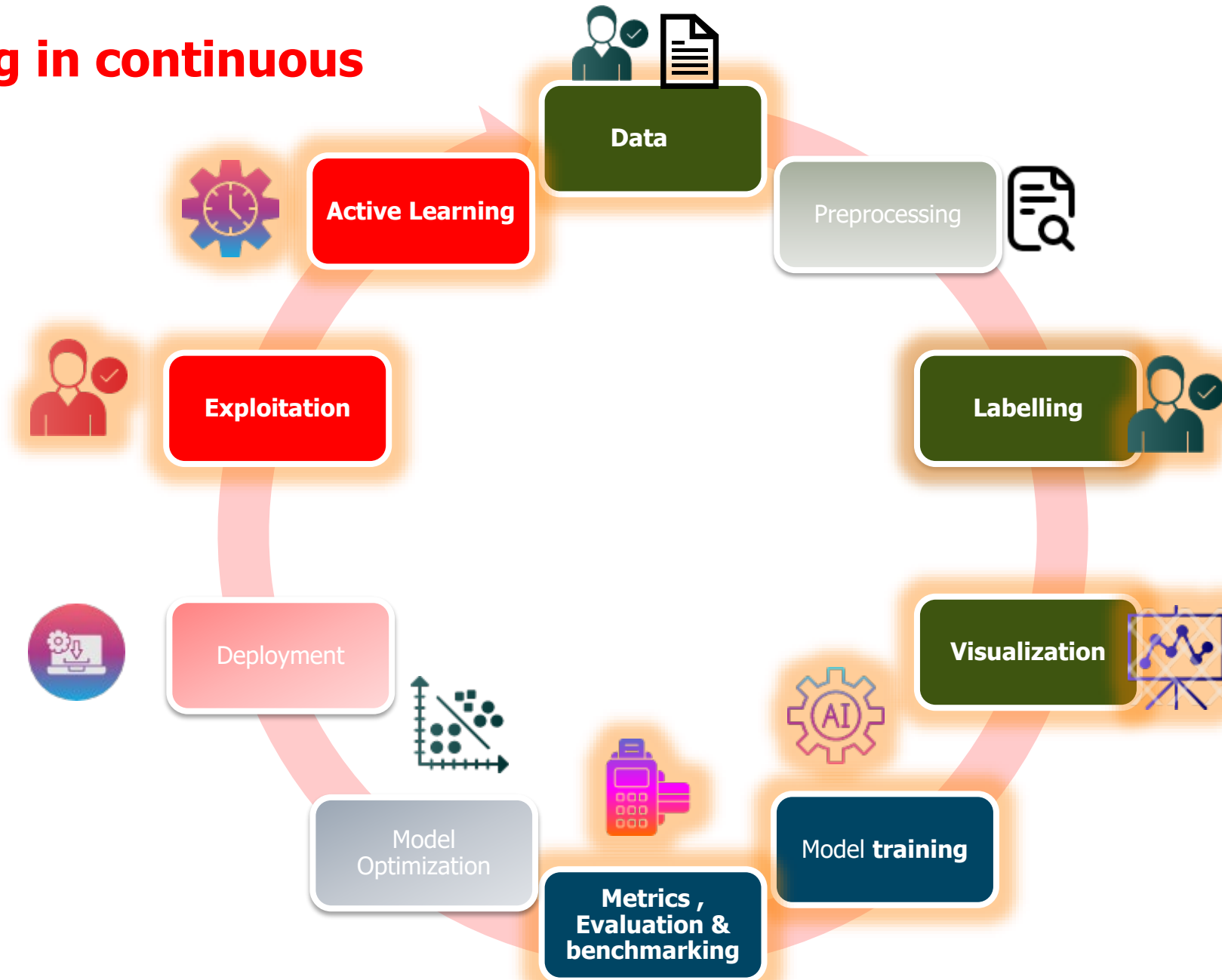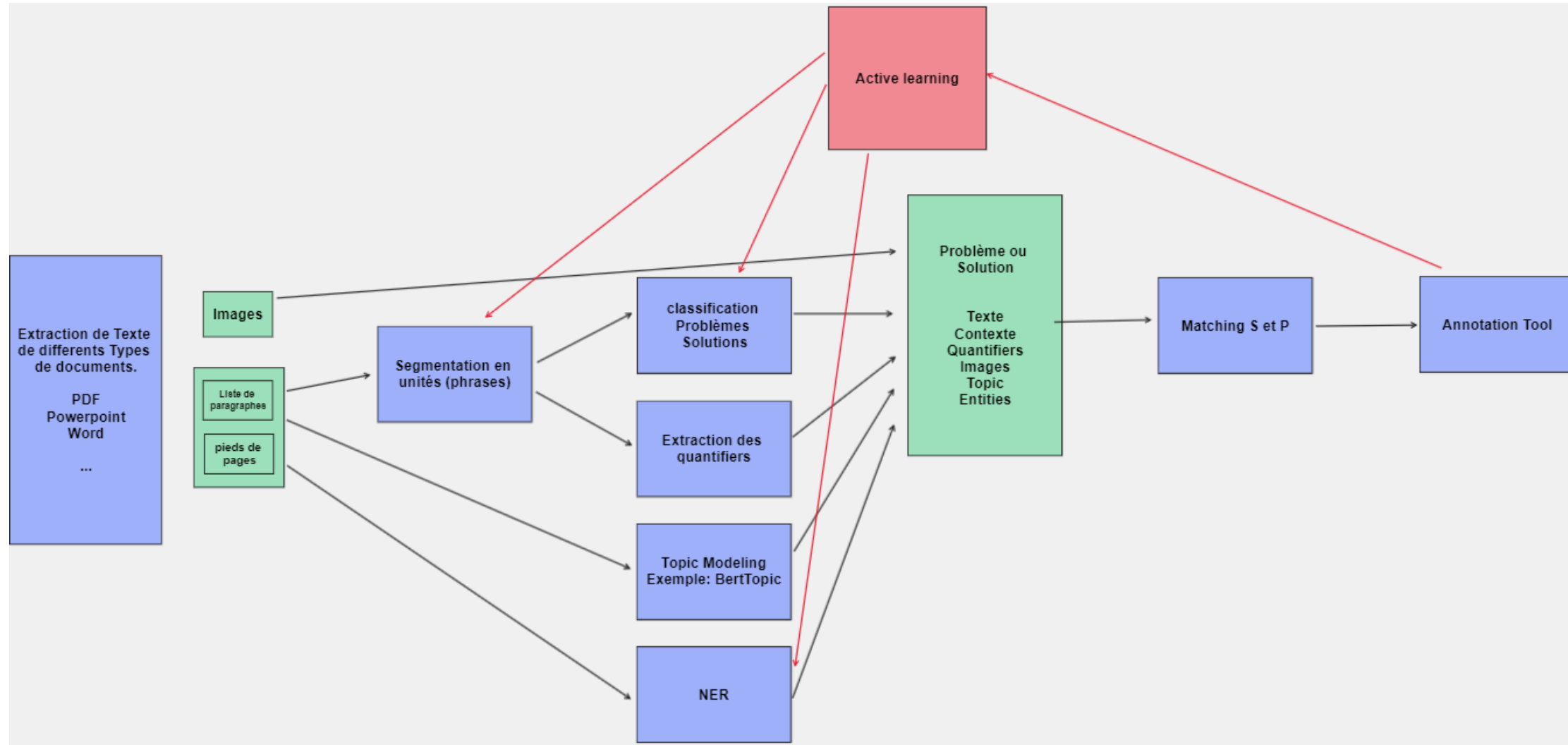
Machine Learning Cycle

- Data
- Preprocessing
- Labelling
- Visualization
- Model training
- Metrics, Evaluation & benchmarking
- Model Optimization
- Deployment
- Exploitation
- Model in Active Learning

Machine Learning in continuous

Architecture

Conclusions and next steps

- Problem/solution pattern is well adapted to the technical and industrial knowledge

- The pattern could variate from one field to an other -> need to use fundamental models to fine-tune

- Matching Problem/solution is an endless process: new problems and new solutions to problems could raise -> ML model continuously learn from new incoming data (active learning)

- Human in the loop for fine-tuning and correcting weak labeling

- LMM : why not problem scalability

Thank you
Questions?