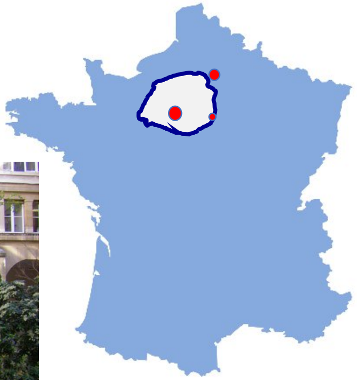
The challenges of industrializing Natural Language Processing for health data

Xavier Tannier
xavier.tannier@sorbonne-universite.fr

Montréal, April 8, 2025

Merci !

Research in Computer Science for Health



Sites AP-HP
- Hôpital Tenon
- Hôpital Avicenne
- Hôpital Charles Foix
- Hôpital La Pitié Salpetrière
- DSN AP-HP

And our own site in Hôpital Trousseau

Disciplinary fields of research

Knowledge Engineering

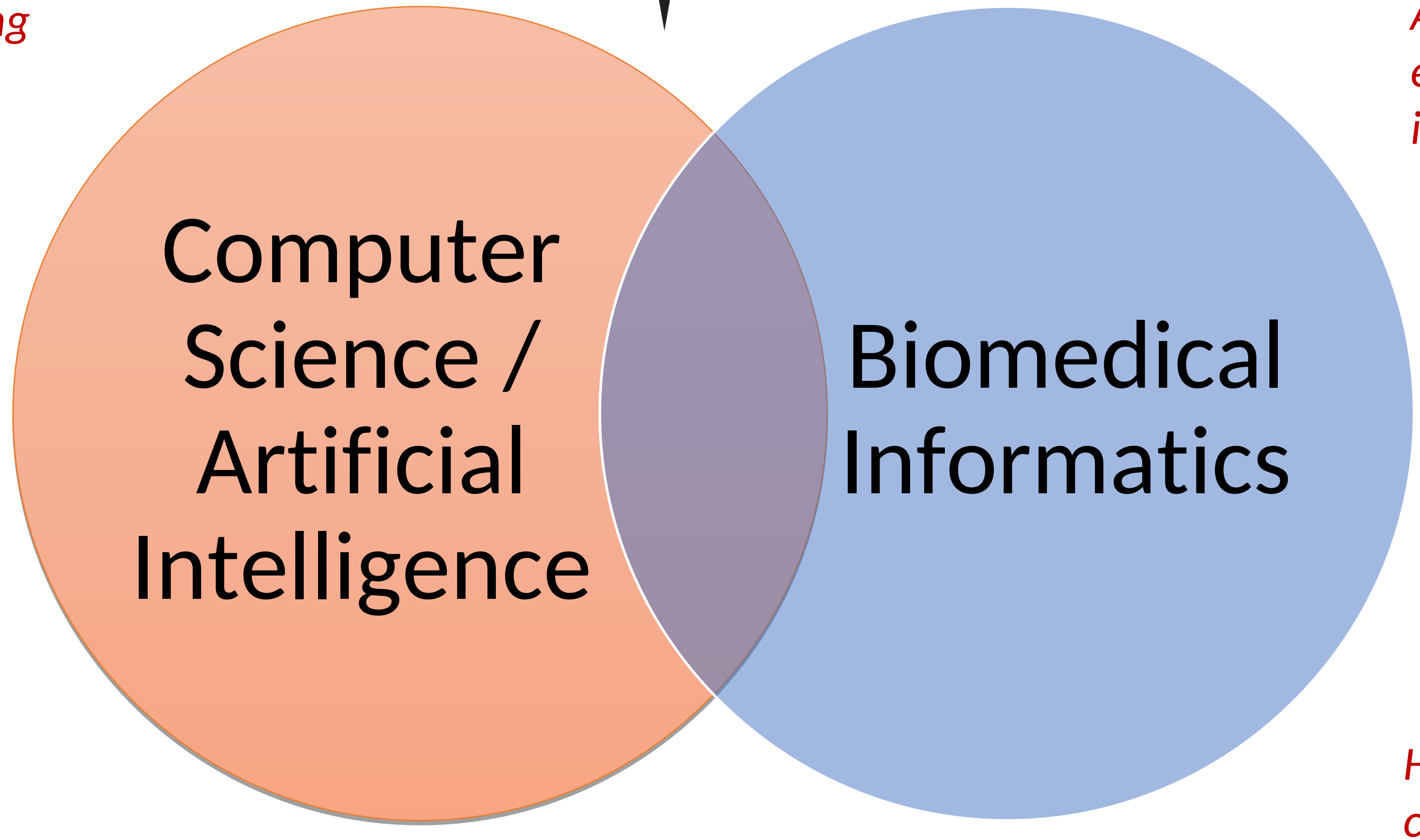
Machine Learning

Natural Language Processing

Ontologies

Formal reasoning

Hybrid Symbolic / numerical AI



Annotation and semantic enrichment of data, interoperability

Data and knowledge visualization

Decision support systems

Health data warehouses, clinical research informatics

Themes

M3P — Drug representation: pharmaceutical knowledge, prescription, pharmacovigilance

Catherine DUCLOS
Marie-Christine JAULENT

EDIRC — Clinical data warehouses & cohorts / Secondary use of healthcare data

Christel DANIEL

ADES — Decision support for care, management, public health and healthcare research

Emmanuelle KEMPF
Brigitte SÉROUSSI

RIIS — Research in informatics and health informatics

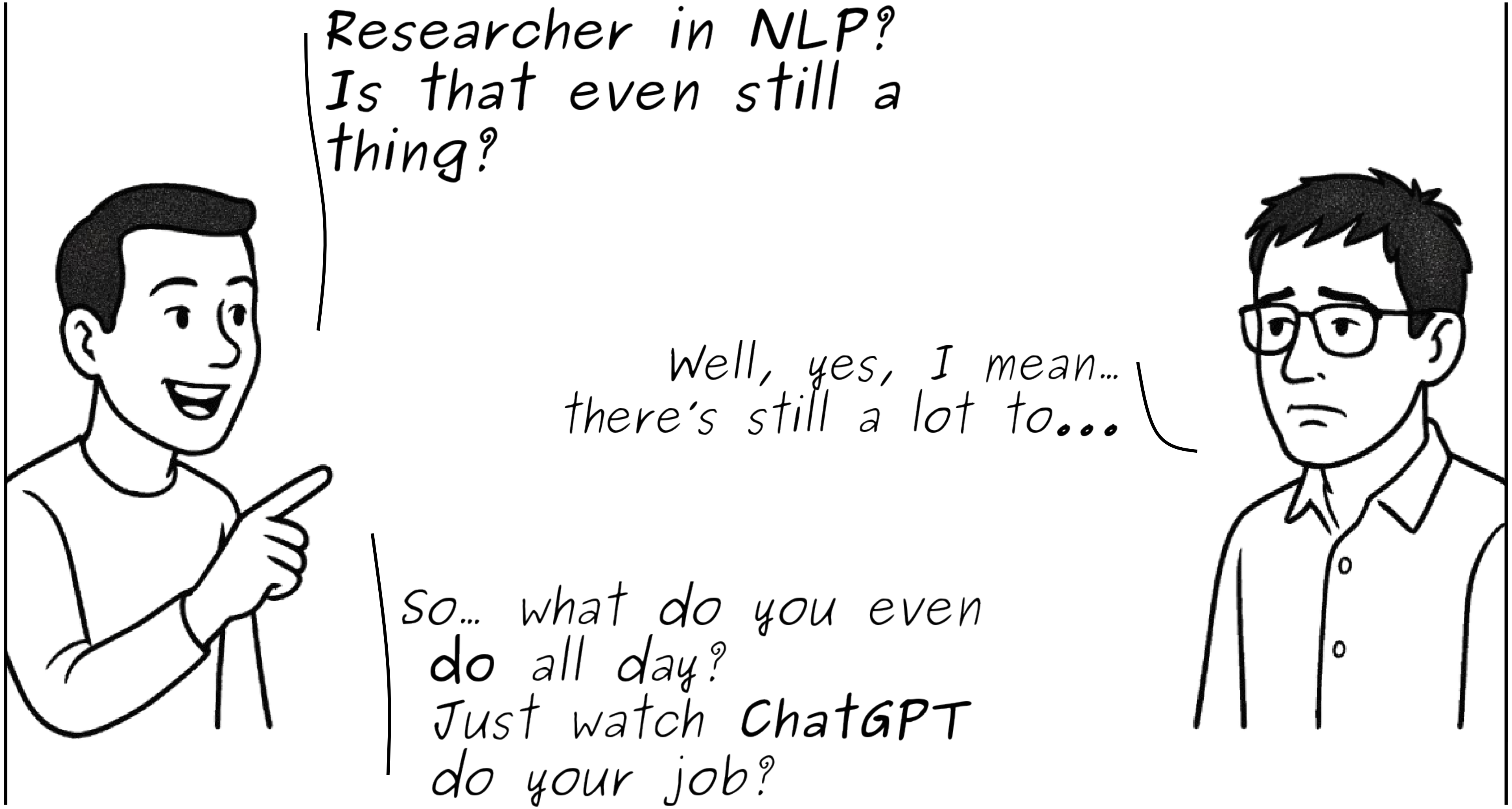
Jean-Baptiste LAMY
Chan Le Duc

Who am I?

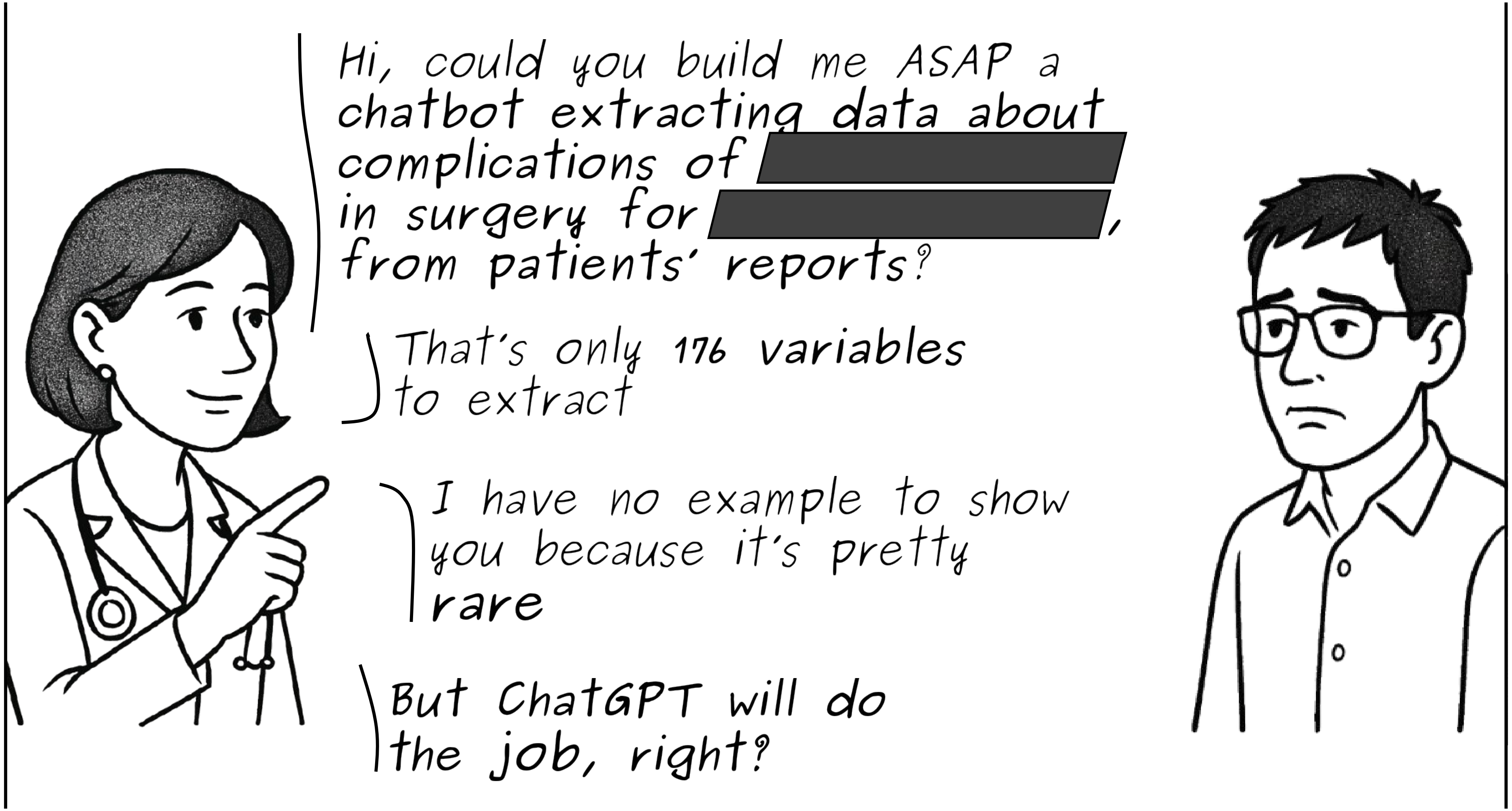
Xavier Tannier

- Professor in Computer Science at Sorbonne Université
- Researcher at Limics
- Natural Language Processing, Information extraction, machine learning… applied to health data

Being an NLP researcher in 2025

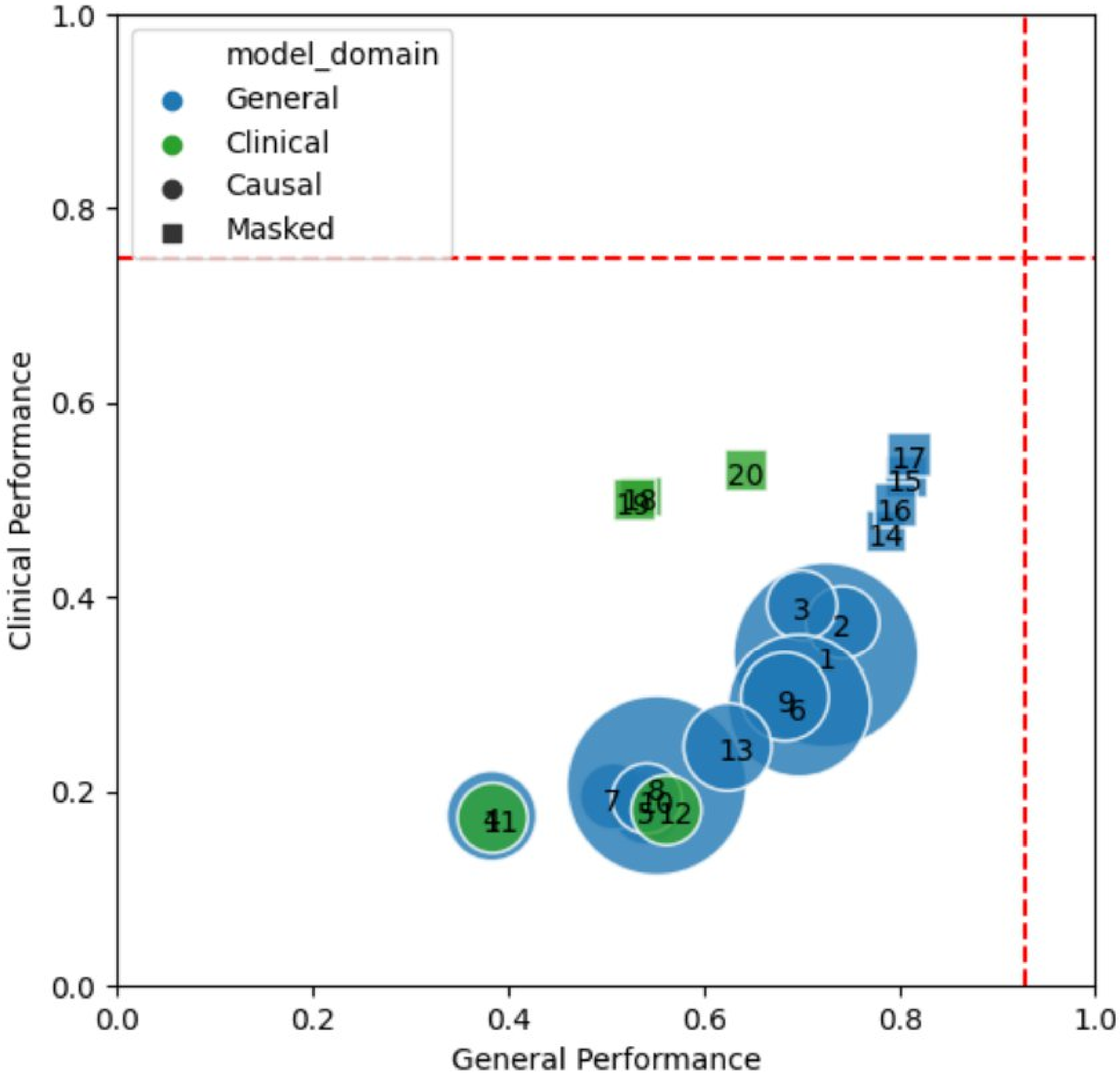


Being a ***clinical*** NLP researcher in 2025



… And yet (1/2)

In few-shot clinical entity recognition: masked language models outperform generative model prompting
(Naguib et al., 2024)

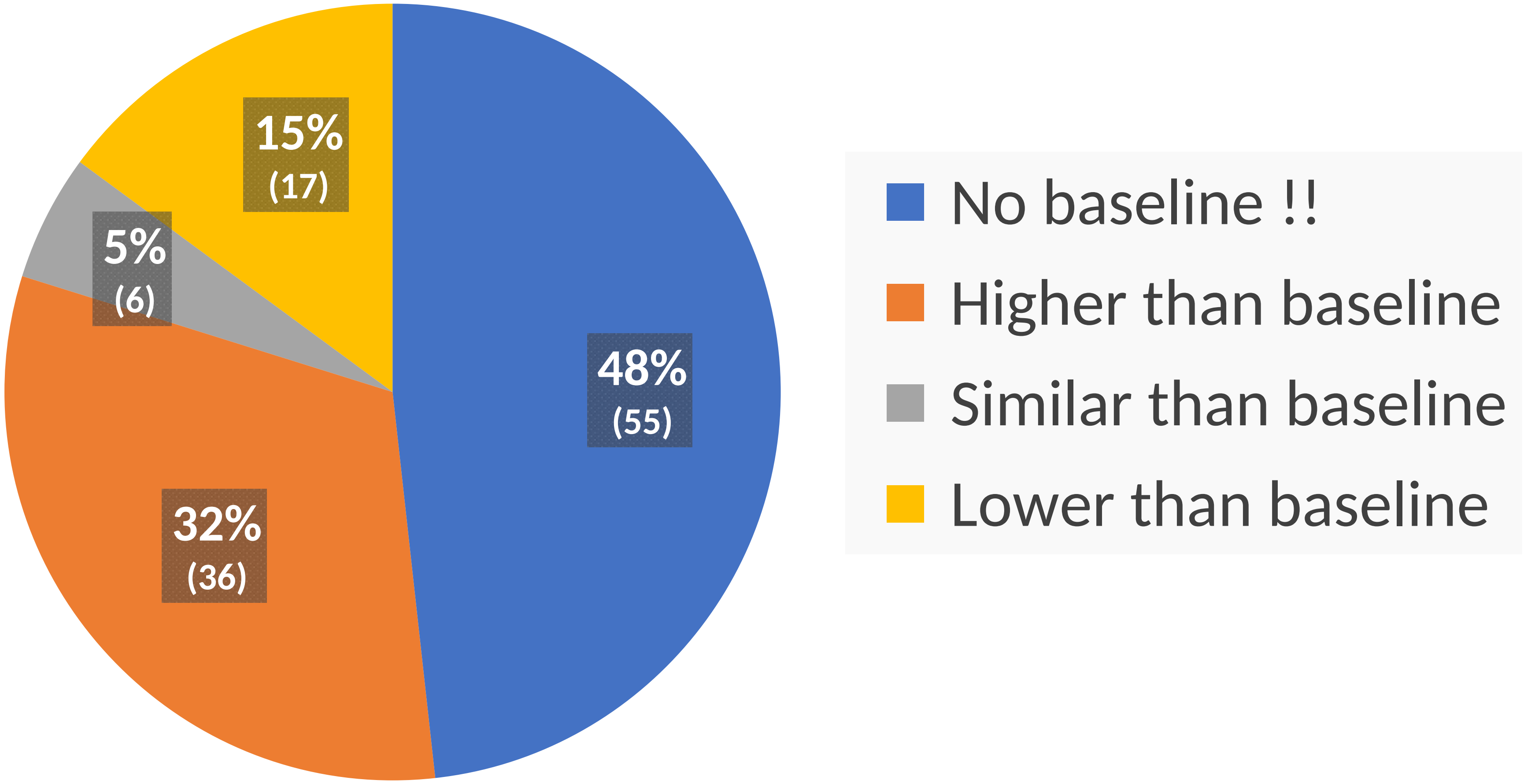


Marco Naguib, Xavier Tannier, Aurélie Névéol.
Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting.
in *Findings of EMNLP 2024*.

… And yet (2/2)

Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C.
Prompt Engineering Paradigms for Medical Applications: Scoping Review.
Journal of Medical Internet Research. September 2024.

Comparison of LLM prompt engineering versus other approaches, through a systematic literature review
(Zaghir et al., 2024)

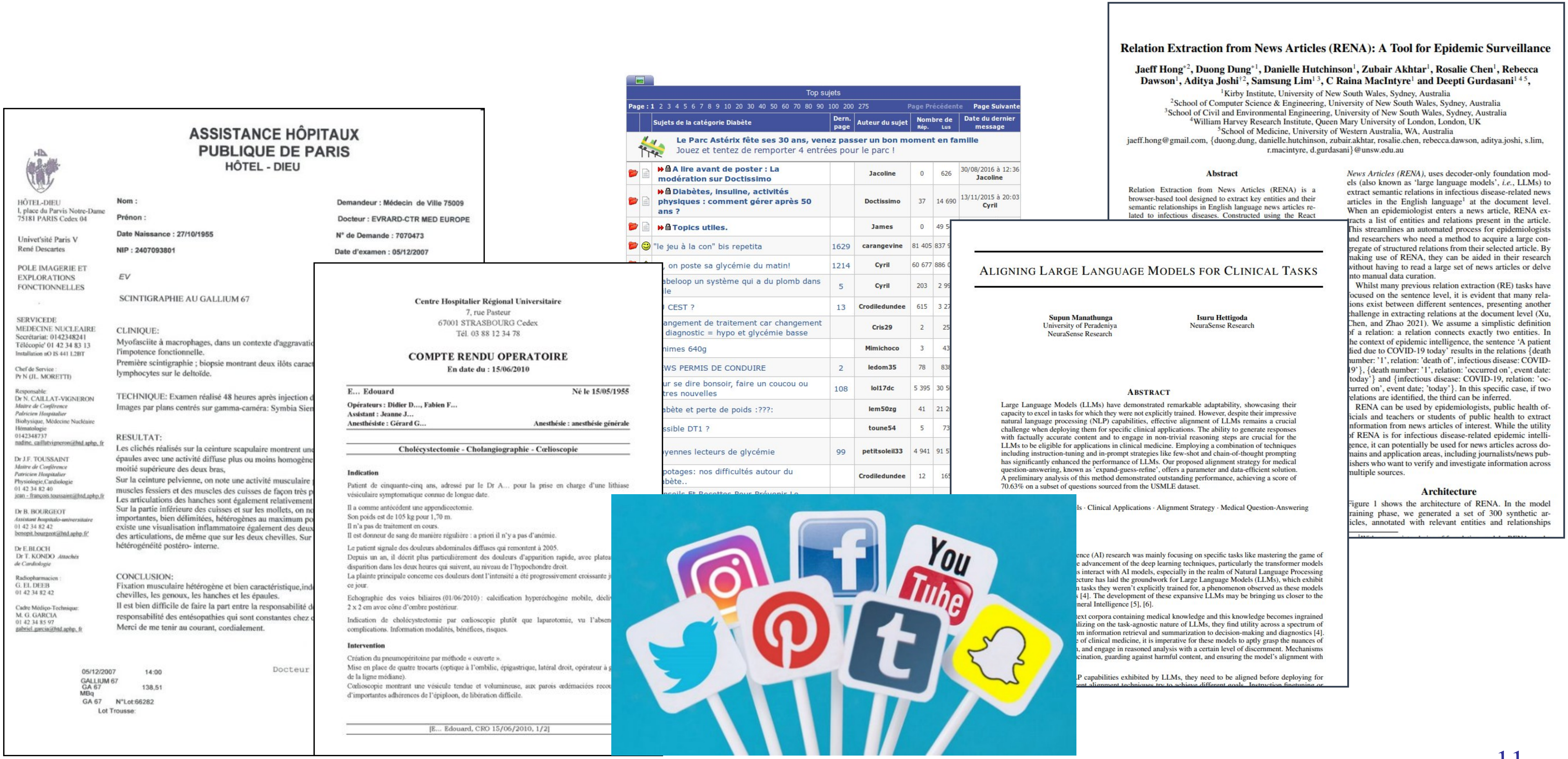


Natural Language Processing

&

Healthcare data

Why Natural Language Processing?



NLP tasks

9 blood cultures the 26/6/15 to *Staphylococcus aureus* methicilline sensitive.

BACTERIA RESISTANCE

VASCULAR DISEASE

Patient with arterial ulcers of the lower limbs, followed at St Joseph (Dr Wyliana) with skin graft in October 2015

PROCEDURE ANATOMY MEDICAL DEVICE

Contraception with intrauterine devices

Patient with no evident sign of cranial traumatism [fact:neg] [C0018674] (Craniocerebral Trauma)

Patient with history of bariatric surgery [fact:patient history] [C1456587] (Bariatric Surgery)

localization

Suspicious masses are observed in the right breast.

localization

The exam is rated ACR 2 on both right and left.

localization

NLP tasks

MAMMOGRAPHY:

SHAPE LESION

There is a 1.8 cm round mass with a circumscribed

LATERALITY LOCALIZATION

margin in the left breast in the anterior depth central

SHAPE LESION

to the nipple. There also is a 1.4 cm oval mass with an

LATERALITY

obscured margin in the left breast in the anterior

depth of the inferior region.

LOCALIZATION

SCORE

ASSESMENT: BI-RADS Category 3

SCORE	LATERALITY	LESION	LATERALITY	LOCALIZATION	SHAPE
BI-RADS 3	LEFT	mass	LEFT	QSI	1.8
BI-RADS 1	RIGHT	mass	LEFT	QI	1.4

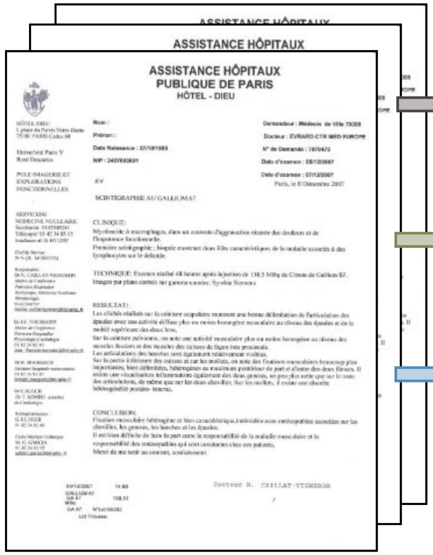
NLP tasks


ASSISTANCE HÔPITAUX
PUBLIQUE DE PARIS
HÔTEL - DIEU

Reason for admission = low-kinetic fracture?

Family history?

Endogenous or exogenous Cushing's syndrome ?
(complications due to glucocorticoid excess)


ASSISTANCE HÔPITAUX
PUBLIQUE DE PARIS
HÔTEL - DIEU

Diabetic patient?

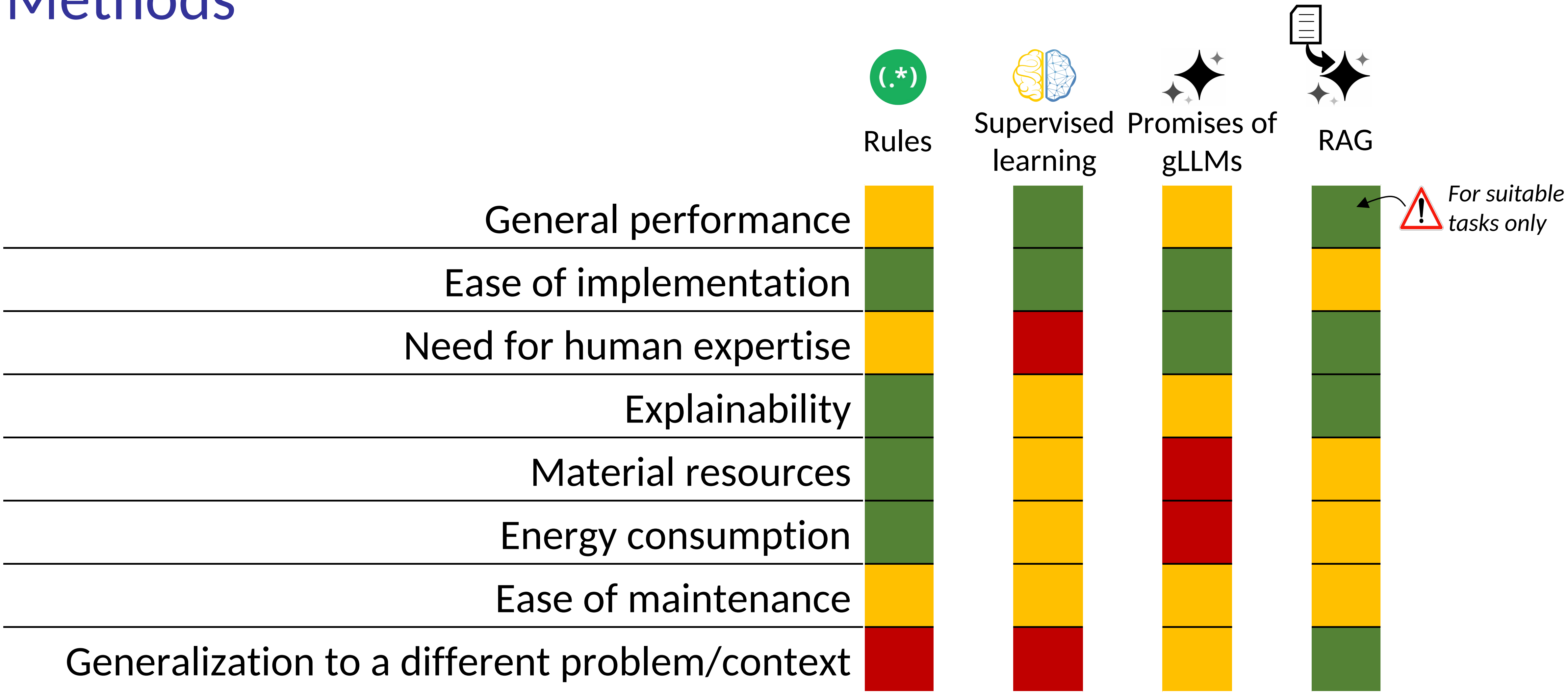
Smoker?

Able to make his/her own decision?

Objectives

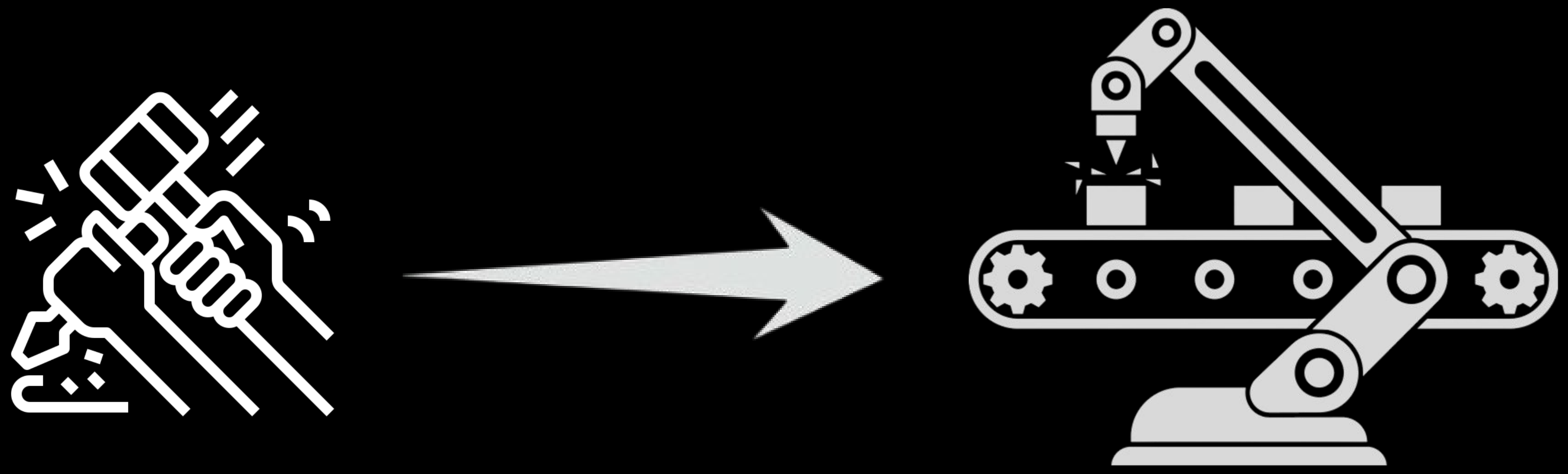
- Selecting patients matching criteria
- Pseudonymising patient records
- Selecting similar patients
- Structuring data
- Summarizing and visualizing patient's history
- Indexing & Retrieving
- Pre-filling forms
- …

Methods



see backup slides for more details...

The challenges of industrializing NLP for health data



Disclaimer

LIMICS

LISN
LABORATOIRE INTERDISCIPLINAIRE
DES SCIENCES DU NUMÉRIQUE

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé

iPLesp
Institut Pierre Louis d'Épidémiologie et de Santé Publique
Pierre Louis Institute of Epidemiology and Public Health

UNIVERSITÉ DE GENÈVE

The challenges of industrializing NLP for health data

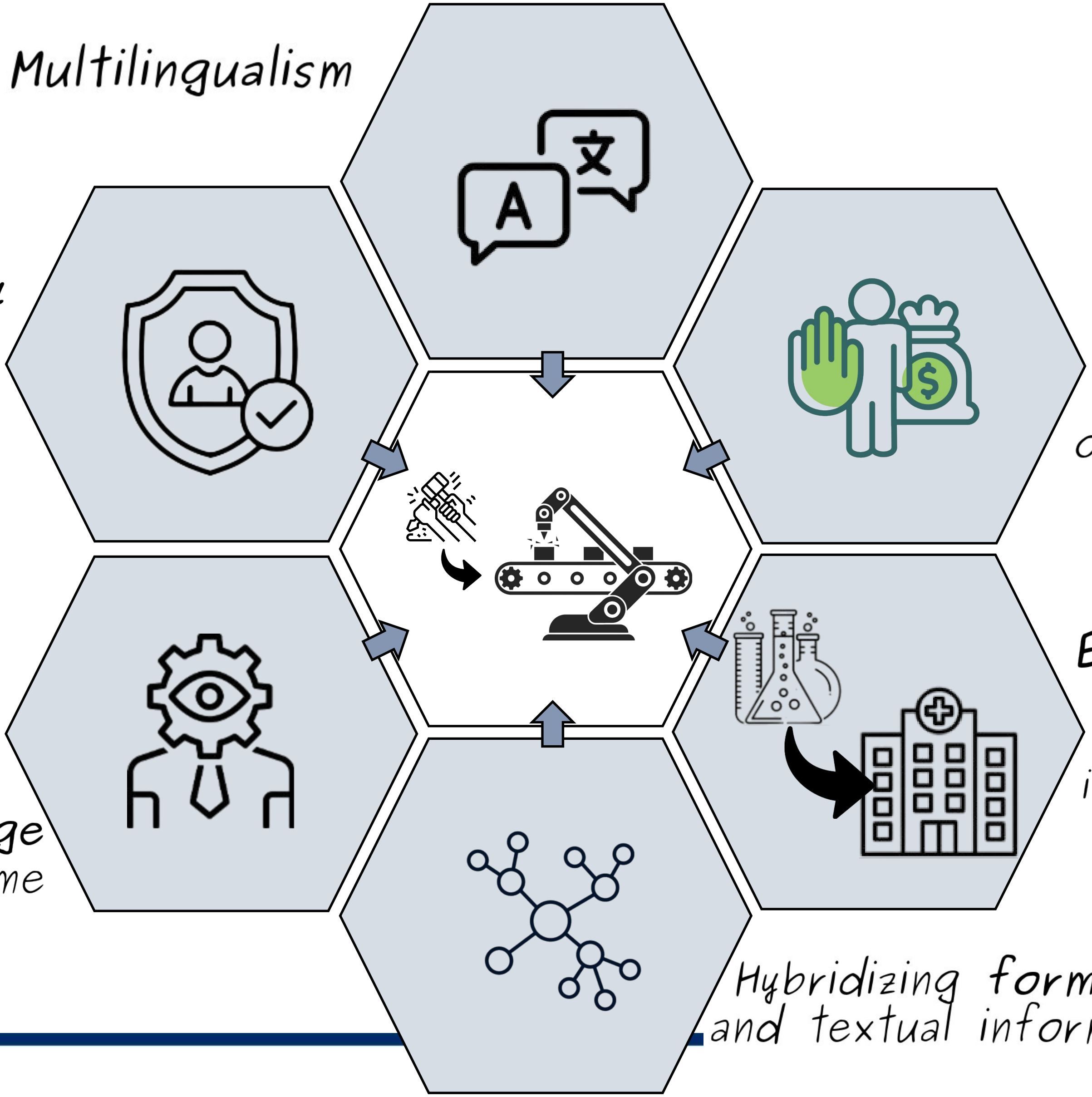
Multilingualism

Privacy

Pseudonymization

Minimizing **resources**

Bring out the big guns only when necessary



Saving human time

– Using weak or **distant supervision**

– Using **human knowledge** without using human time

Evaluation in clinical settings

Interoperability with hospital information systems

Hybridizing **formal knowledge** and textual information

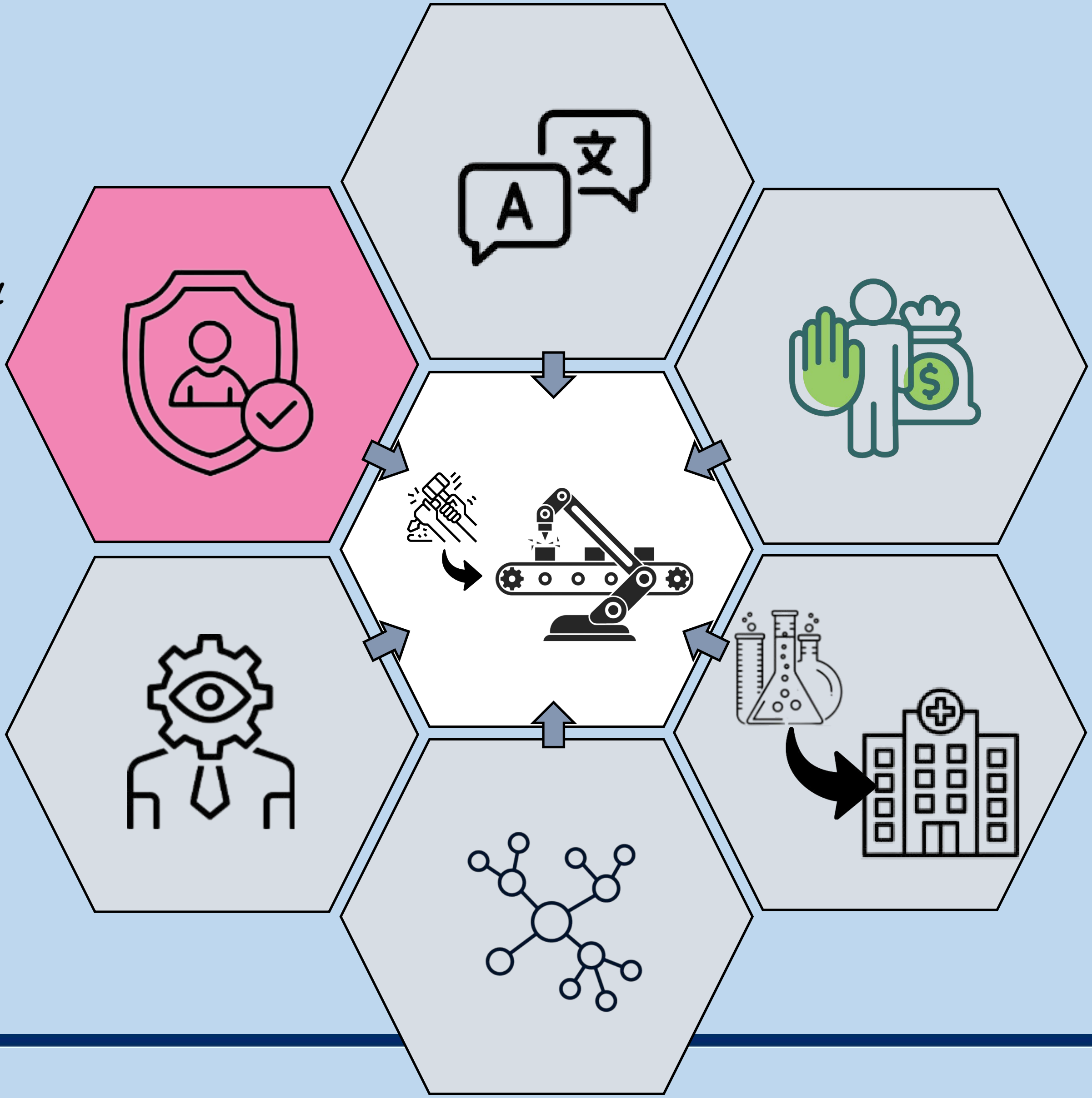
Example 1: Pseudonymization

Privacy

Pseudonymization

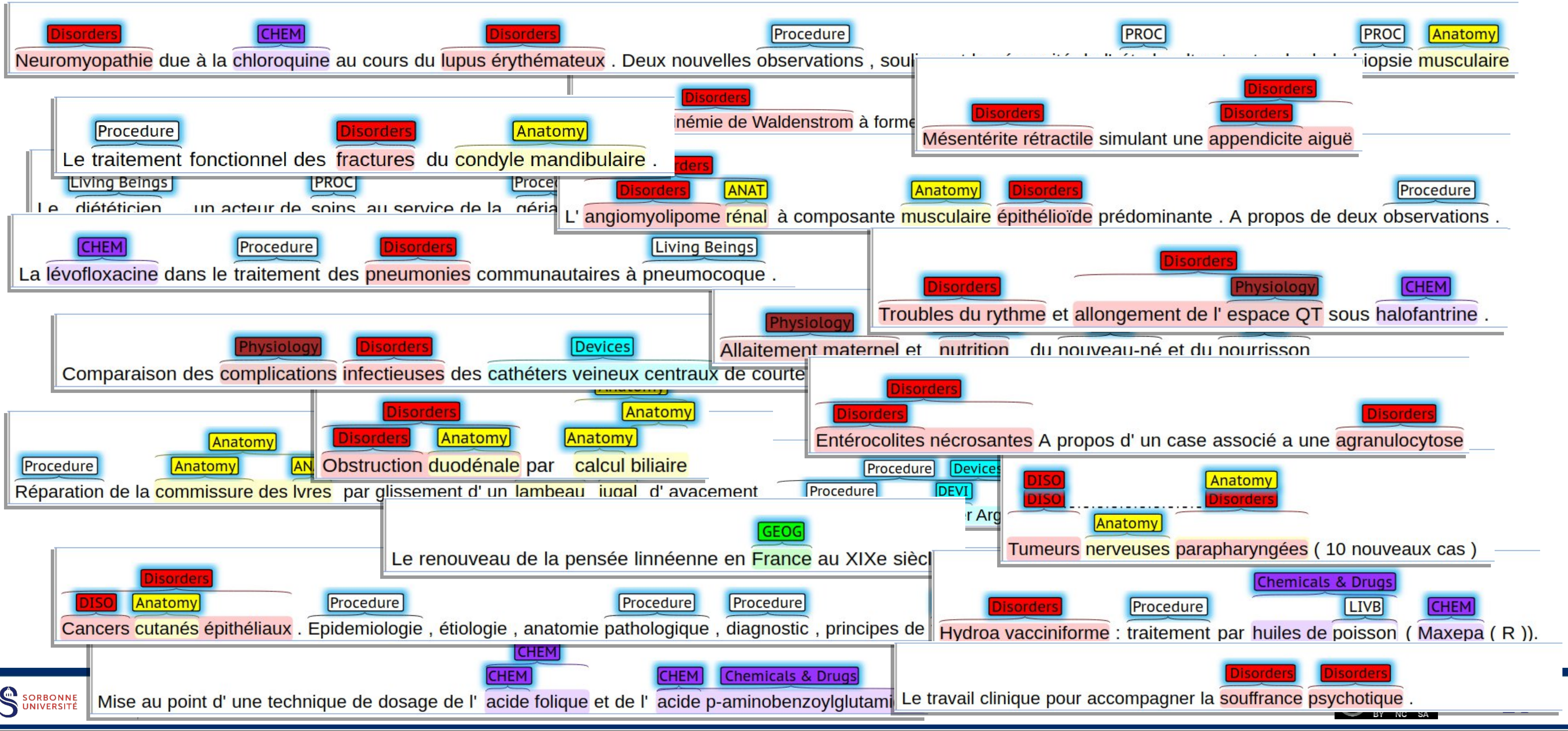
with

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé



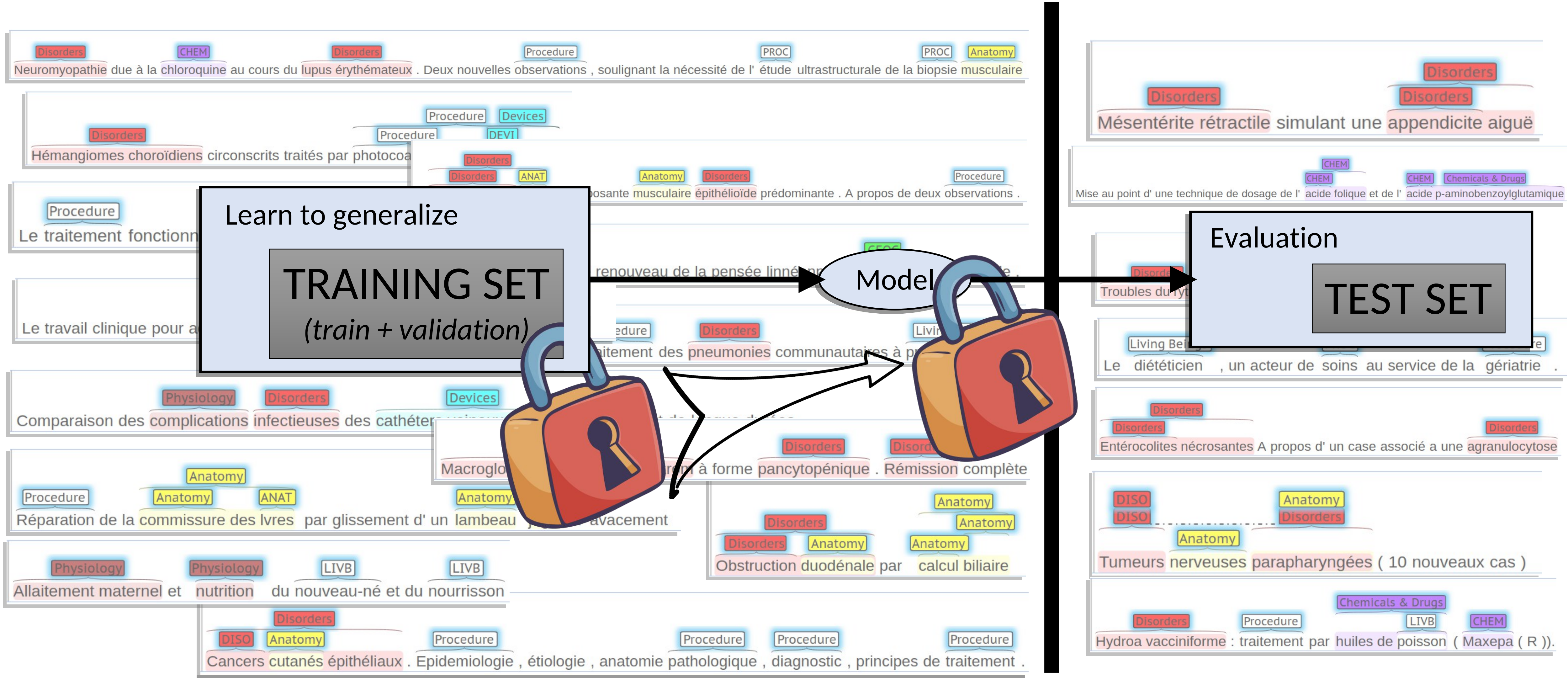
Supervised learning systems

Névéol, A., Grouin, C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014).
The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.



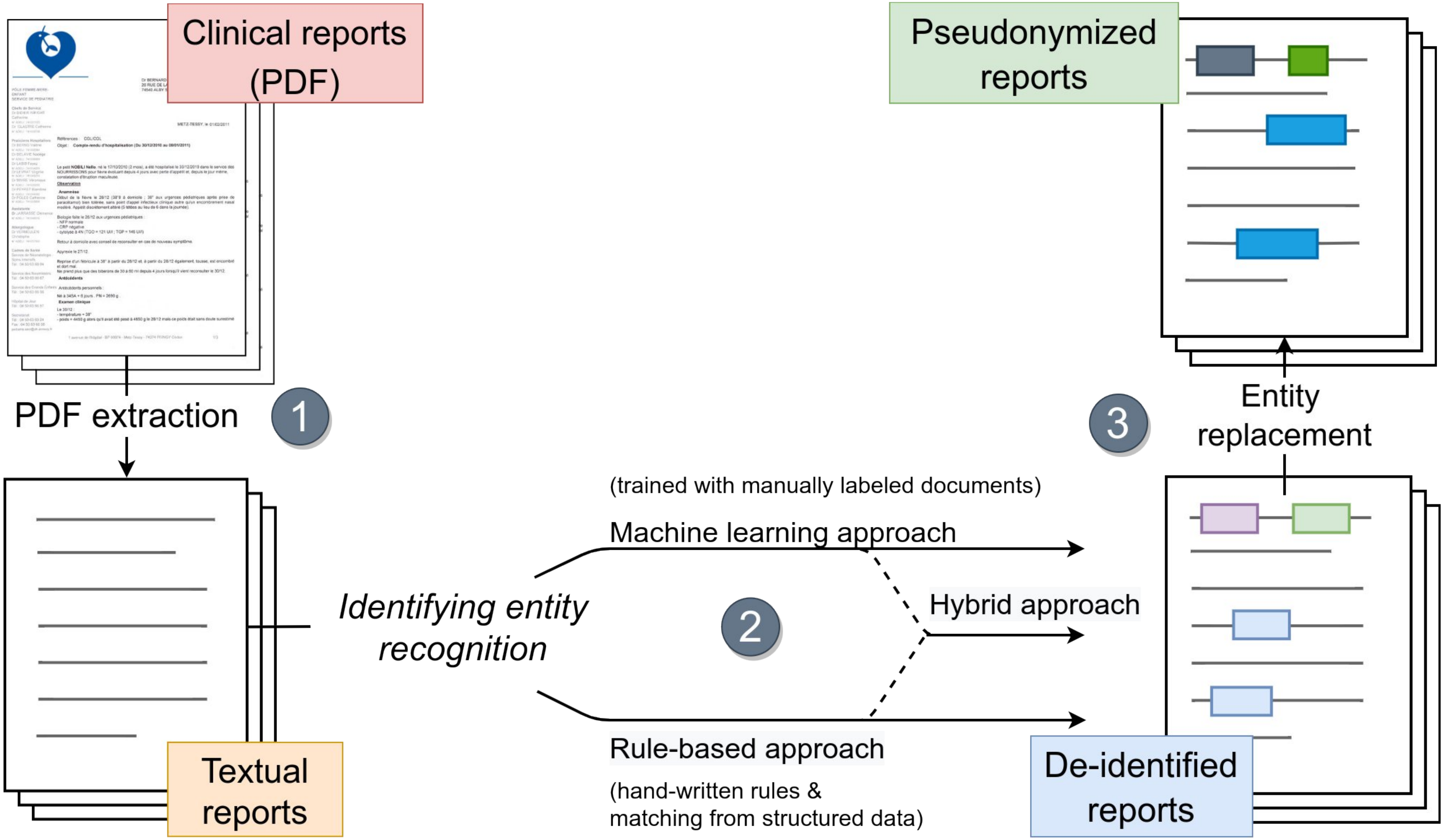
Supervised learning systems

Névéol, A., Grouin, C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014).
The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.



Pseudonymisation

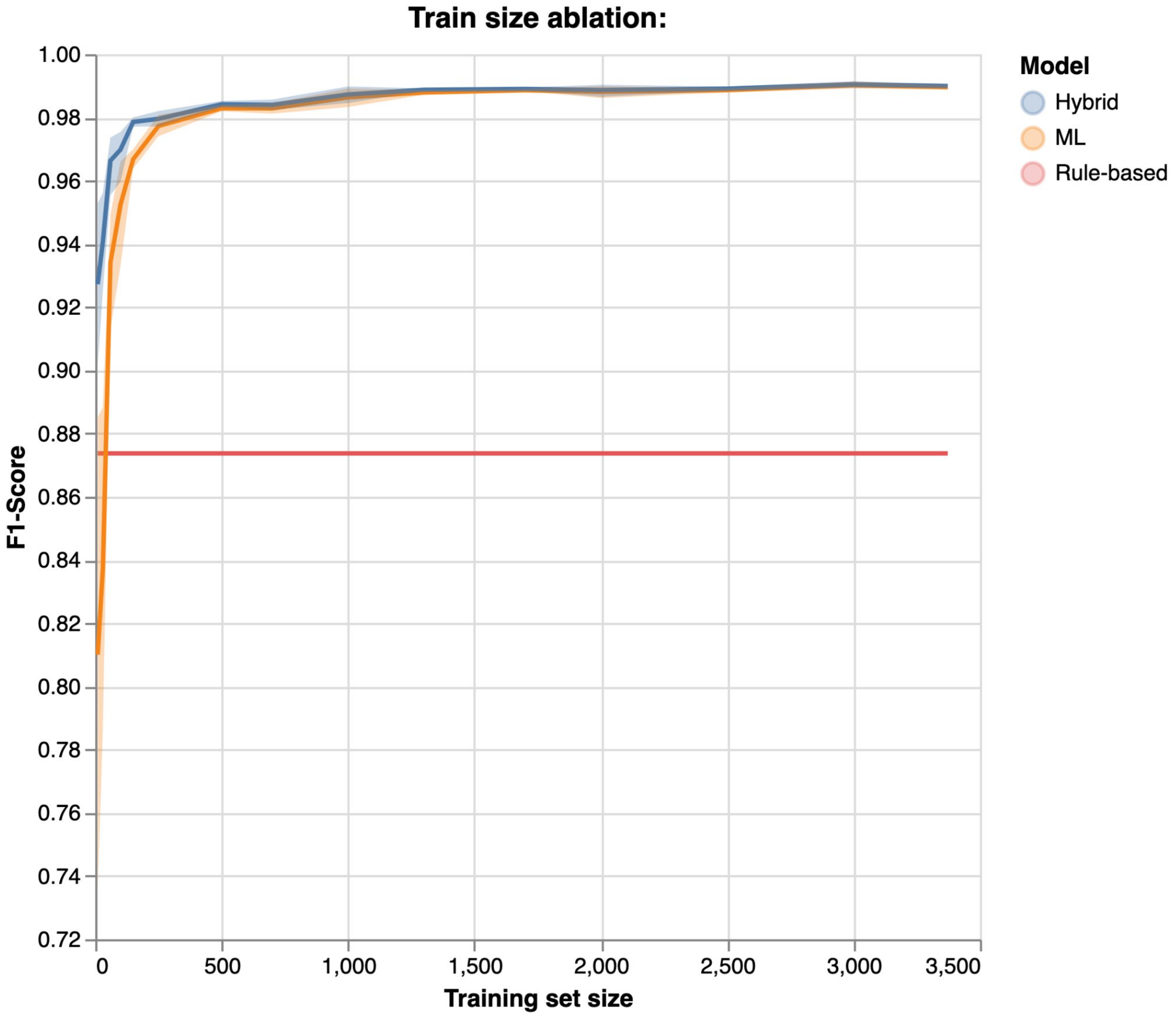
Tannier et al. **Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse.** *Methods of Information in Medicine*. 2024



Pseudonymisation

A **cookbook** to build your own pseudonymization pipeline :

- Training set size
- Document types
- Language models
- Methods (only ML or ML+rules)



Tannier et al. **Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse.** *Methods of Information in Medicine*. 2024

Pseudonymisation

EDS-Pseudo, an AP-HP library for pseudonymisation:

- Code to train your pipeline on your own data
- A **public model trained on fake data** (handwritten + GPT3 augmentation)

Label	Description
`ADRESSE`	Street address, eg `33 boulevard de Picpus`
`DATE`	Any absolute date other than a birthdate
`DATE_NAISSANCE`	Birthdate
`HOPITAL`	Hospital name, eg `Hôpital Rothschild`
`IPP`	Internal AP-HP identifier for patients, displayed as a number
`MAIL`	Email address
`NDA`	Internal AP-HP identifier for visits, displayed as a number
`NOM`	Any last name (patients, doctors, third parties)
`PRENOM`	Any first name (patients, doctors, etc)
`SECU`	Social security number
`TEL`	Any phone number
`VILLE`	Any city
`ZIP`	Any zip code

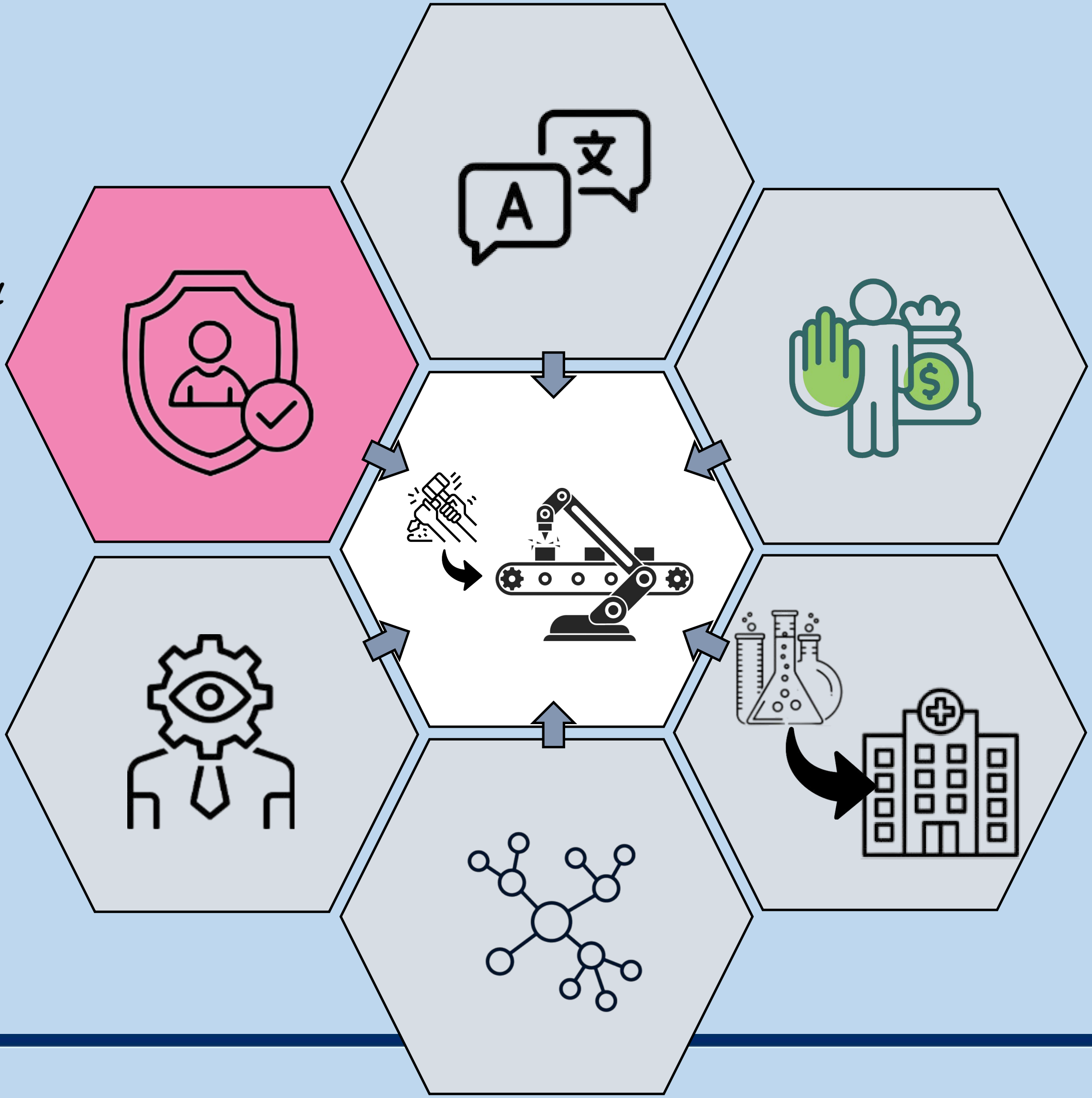
https://github.com/aphp/eds-pseudo

Example 2: Privacy-preserving models

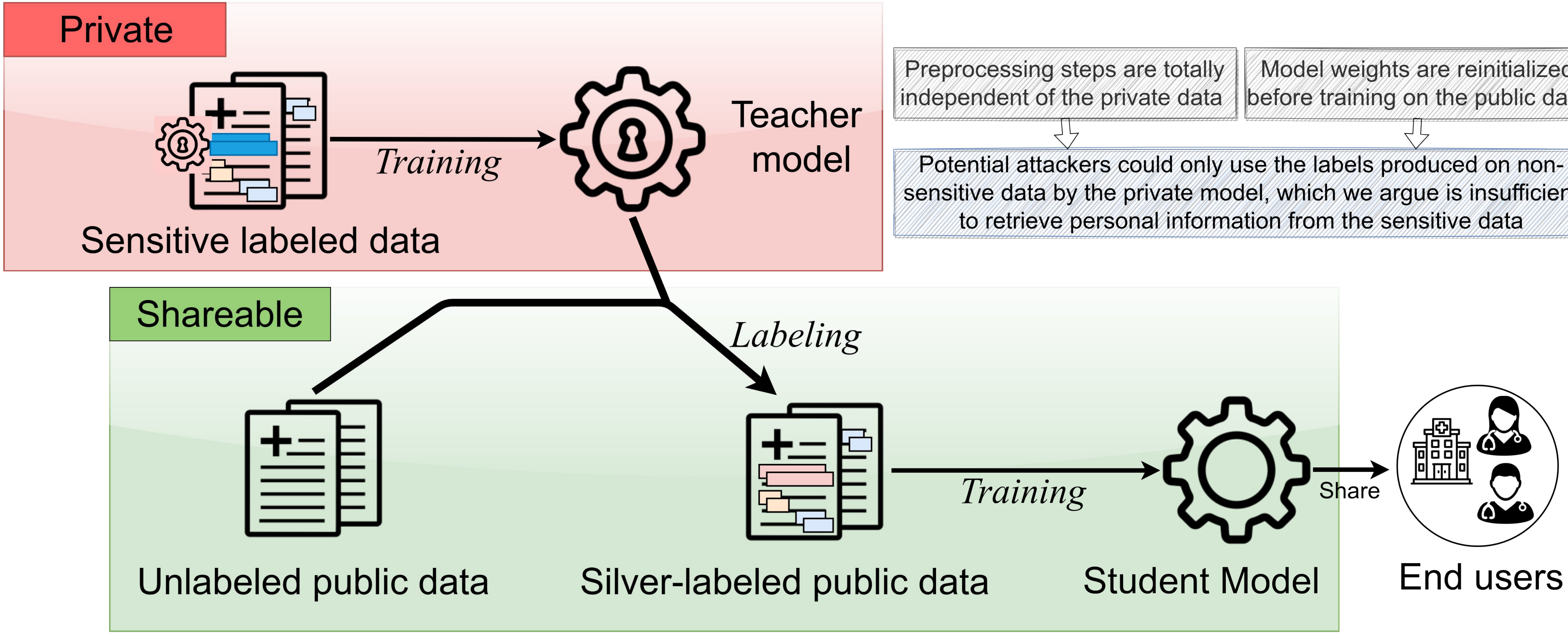
Privacy

Pseudonymization

with



Privacy-preserving mimic models



Preprocessing steps are totally independent of the private data

Model weights are reinitialized before training on the public data

Potential attackers could only use the labels produced on non-sensitive data by the private model, which we argue is insufficient to retrieve personal information from the sensitive data

Privacy-preserving mimic models

- Task : medical concepts in French
- Private data : MERLOT (L. Campillos et al., 2018)
- Public data : DEFT (R. Cardon et al., 2020)

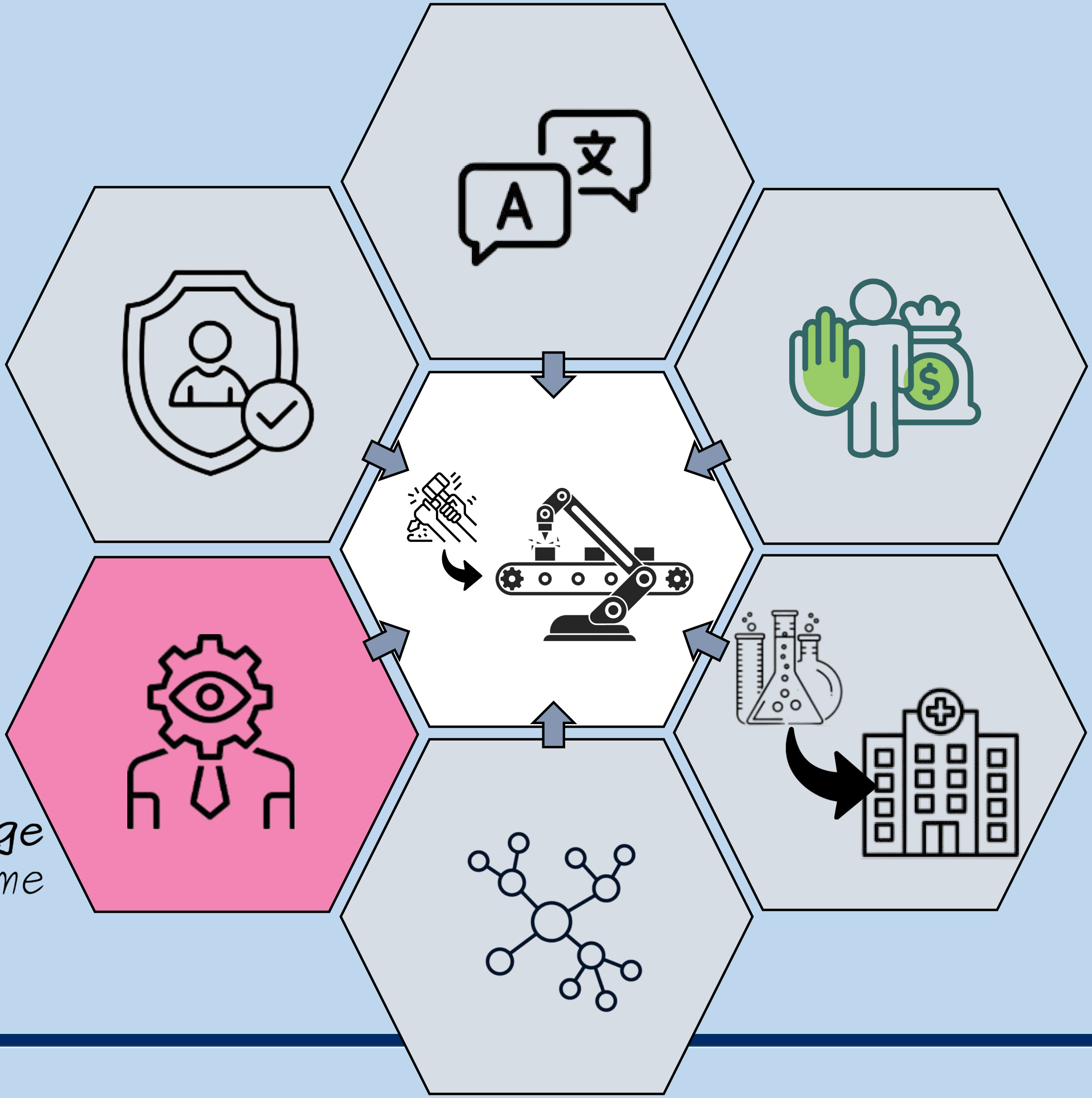
	Precision	Recall
Dictionary-based model (UMLS)	0.25	0.17
Student (public) model	0.60	0.74
Teacher (private) model	0.86	0.86

baseline

skyline

Can be used for preannotation or bootstraping a noise-robust model

Example 3: Weak / distant supervision



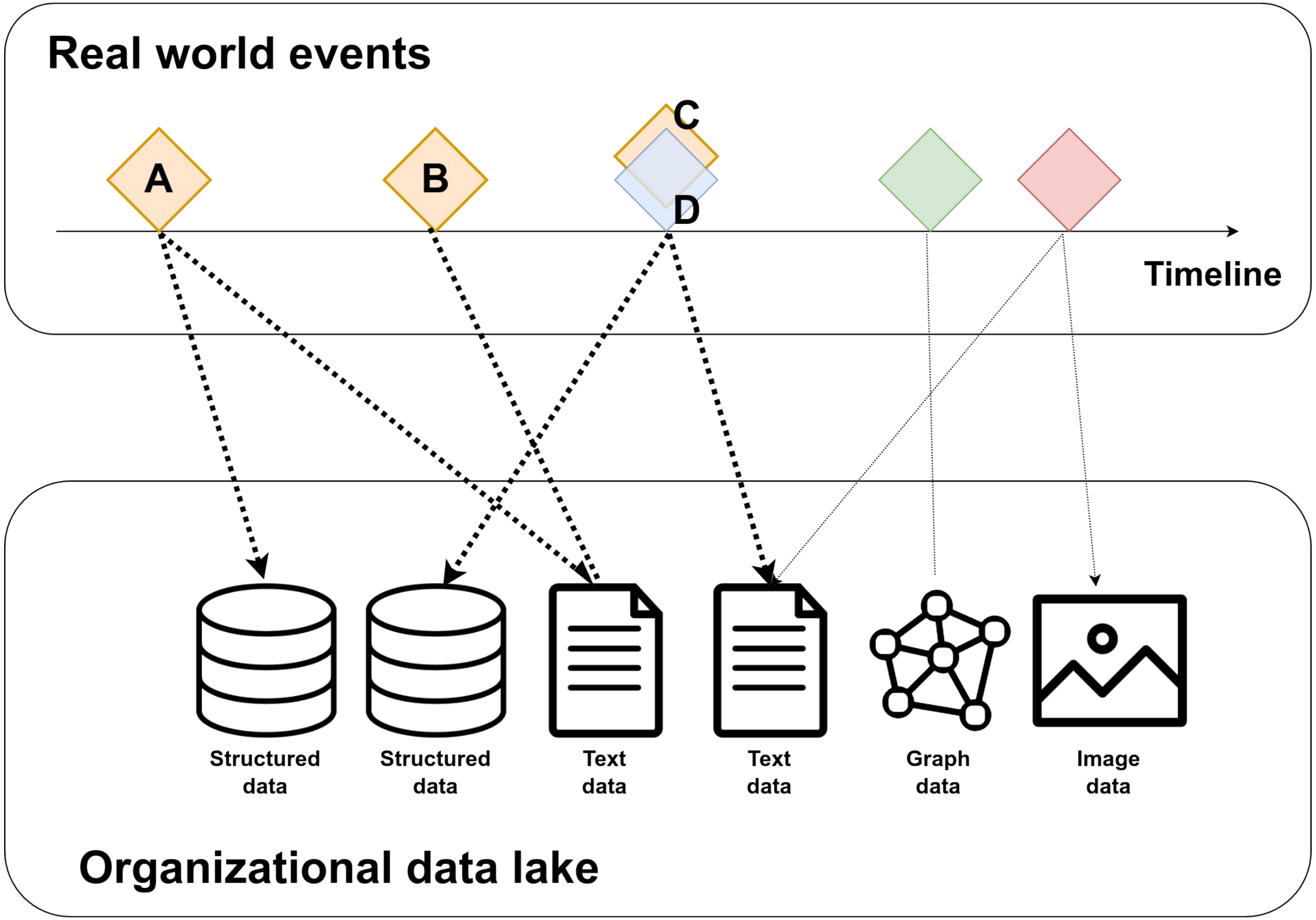
Saving human time

– Using weak or distant supervision

– Using human knowledge without using human time

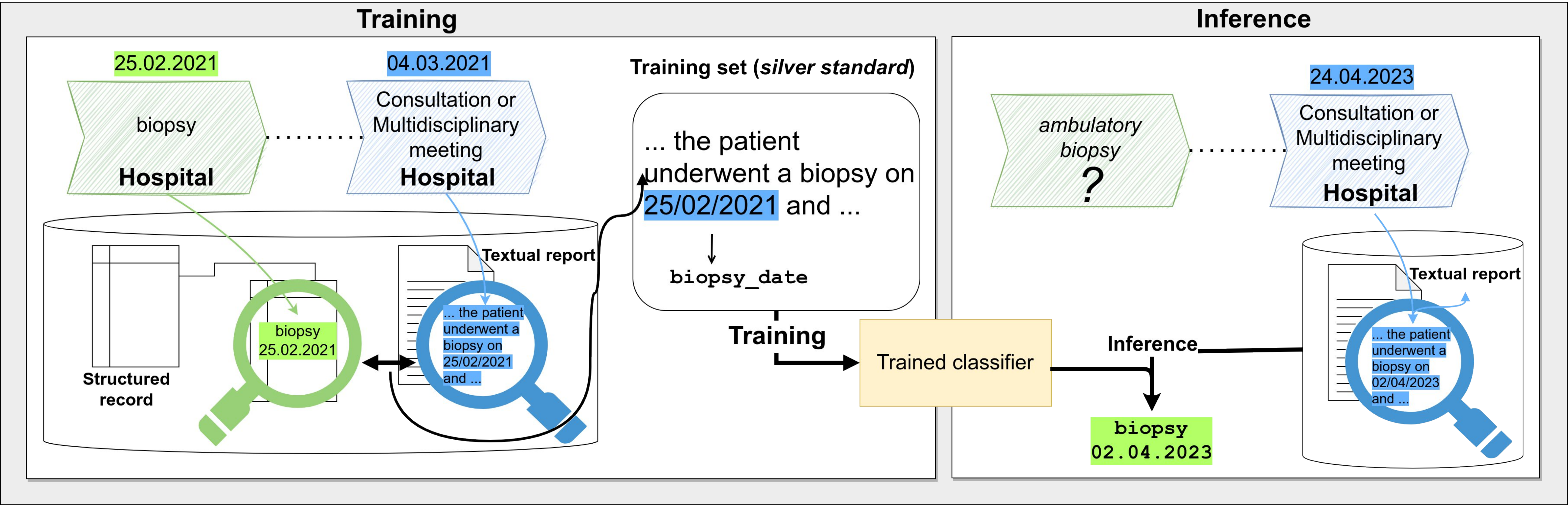
Weak / distant supervision

Information redundancy in an clinical data lake / data warehouse



Weak / distant supervision

How to get distant labels



Weak / distant supervision

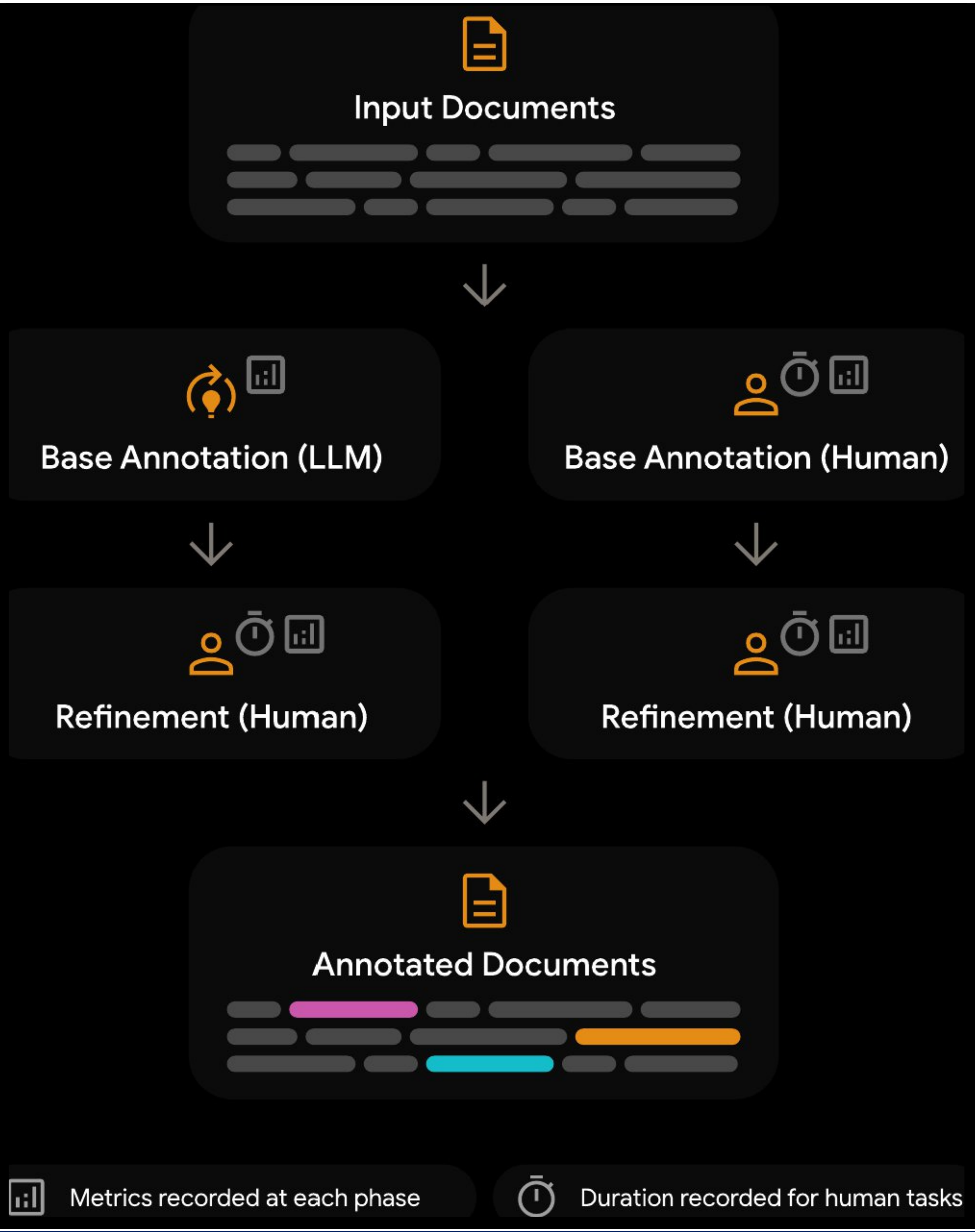
Dealing with noisy labels introduced by the distant supervision

		Precision	Recall
baseline	Regular cross-entropy	0.38	0.90
three different ways to deal with noisy labels	Robust Loss Function (NCE-RCE)	0.61	0.90
	Sample Selection (O2U - CE)	0.56	0.93
	Label Refurbishment (LRT)	0.64	0.90
skyline (supervised)	Expert label dataset (supervised learning)	0.69	0.93

LLM-based supervision

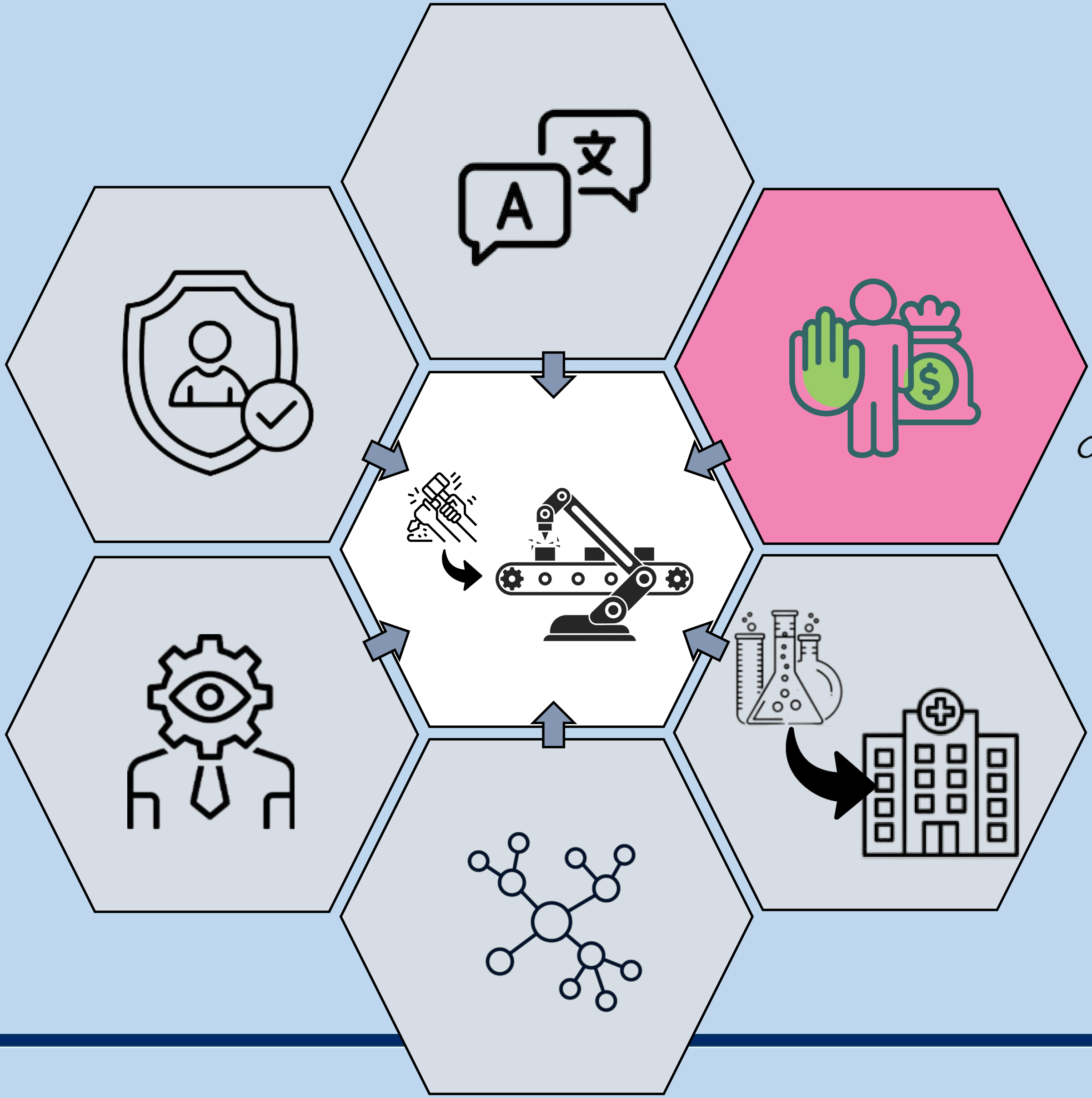
Example:

LLMs Accelerate Annotation for Medical Information Extraction.
Goel at al., *Google Research*



Goel at al. **LLMs Accelerate Annotation for Medical Information Extraction**.
Machine Learning for Health (ML4H) 2023

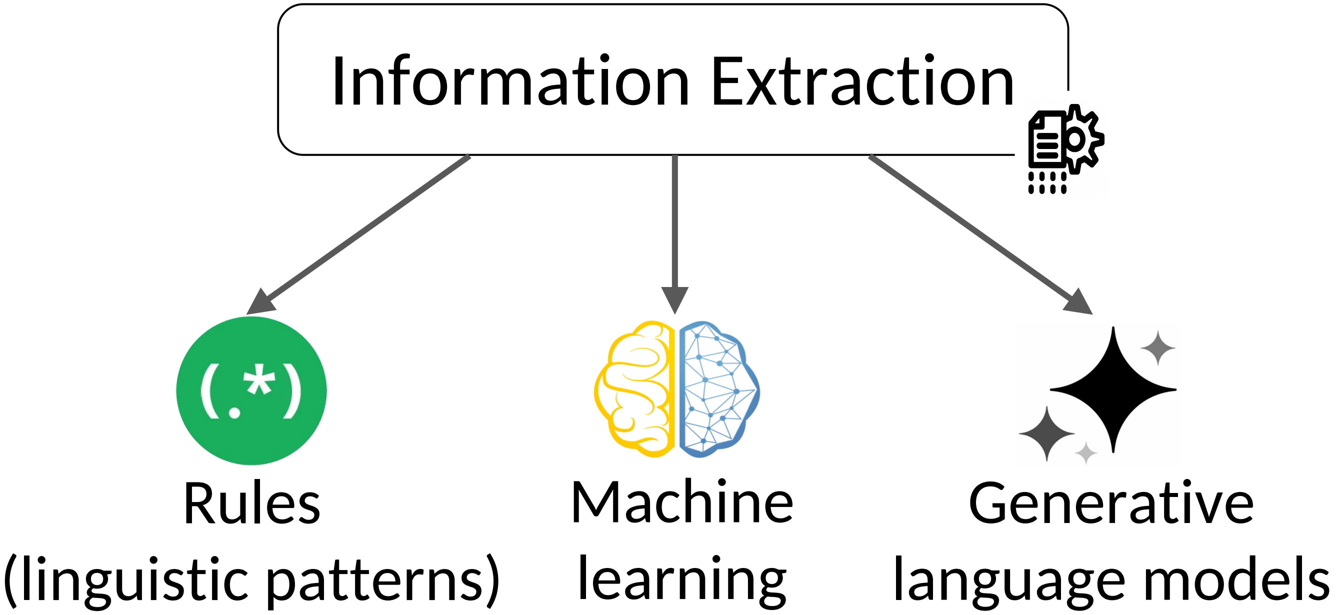
Example 4: Choose the lightest methods



Minimizing **resources**

Bring out the big guns only when necessary

Choose the lightest methods (REST)



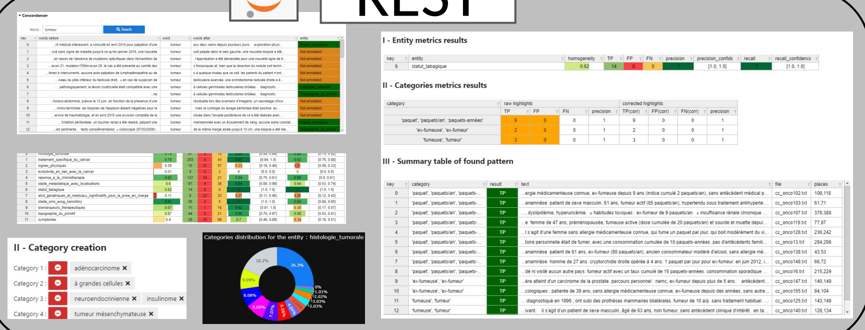
Sustainability | *Transferability* | *Interpretability* | *Human cost (expert, IT)*

Choose the lightest methods (REST)

Information Extraction

- Rules (linguistic patterns)
- Machine learning
- Generative language models

REST



A tool to visualize and predict feasibility of a rule-based system for an information extraction task

https://github.com/GuiGuiBazin/REST-interface

Choose the lightest methods (REST)

Quick-and-dirty annotation by expert
(“*highlighting*”)

Choose the lightest methods (REST)

Concordancer

Word : tumeur

Seach

key	words before	word	words after	entity
0	...nt médical intéressant, a consulté en avril 2010 pour palpation d'une	tumeur	aux deux seins depuis plusieurs jours. exploration physi...	Signes_physiques
1	...ncé sans signe de maladie jusqu'à ce qu'en janvier 2016, une nouvelle	tumeur	soit palpée dans le sein gauche. une nouvelle biopsie a été...	Not annotated
2	...en raison de l'absence de mutations spécifiques dans l'échantillon de	tumeur	. l'approbation a été demandée pour une nouvelle ligne de tr...	Not annotated
3	...-exon 21, mutation t790m-exon 20. le cas a été présenté au comité des	tumeur	s thoraciques et, bien que la résection du nodule soit techn...	Not annotated
4	...ômes b intercurrents. aucune auto-palpation de lymphadénopathie ou de	tumeur	s à quelque niveau que ce soit. les parents du patient n'ont...	Not annotated
5	...iveau du pôle inférieur du testicule droit. » en cas de suspicion de	tumeur	testiculaire avancée, une orchidectomie radicale droite a é...	Not annotated
6	... pathologiquement, la lésion cicatricielle était compatible avec une	tumeur	à cellules germinales testiculaires brûlées. diagnostic ...	Histologie_tumorale
7	...ne	tumeur	à cellules germinales testiculaires brûlées. diagnostic	Topographie_du_primitif

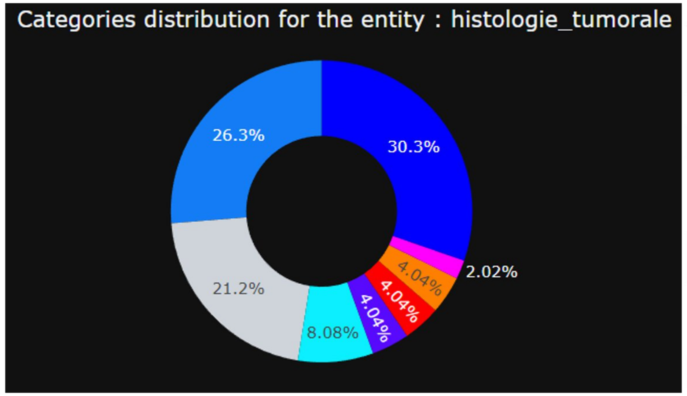
Choose the lightest methods (REST)

Intra-entity linguistic homogeneity

entity	homogeneity
histologie_tumorale	0.79
traitement_specifique_du_cancer	0.79
signes_physiques	0.29
evolutivite_en_lien_avec_le_cancer	0.01
reponse_a_la_chimiotherapie	0.82
stade_metastatique_avec_localisations	0.6
statut_tabagique	0.62
atcd_geriatriques_et_medicaux_significatifs_pour_la_prise_en_charge	0.11
stade_oms_ecog_karnofsky	0.91
biomarqueurs_therapeutiques	0.67
topographie_du_primitif	0.57
symptomes	0.4

Choose the lightest methods (REST)

Precision and coverage
of existing annotations



	category	total_annotations_number
2	'carcinome'	30
1	'adénocarcinome'	26
0	category?	21
4	'mélanome'	8
3	'tumeur germinale', 'tératome', 'germinales'	4
5	'lymhome', 'leucémie', 'blastique'	4
6	'mésenchymateux', 'neuroendocrine', 'mésenchymateuse'	4
7	'sarcome'	2

I - Entity metrics results

key	entity	homogeneity	TP	FP	FN	precision	precision_confide	recall	recall_confidence
5	stade_metastatique_avec_localisations	0.6	45	26	38	0.63	[0.49, 0.76]	0.54	[0.42, 0.66]

II - Categories metrics results

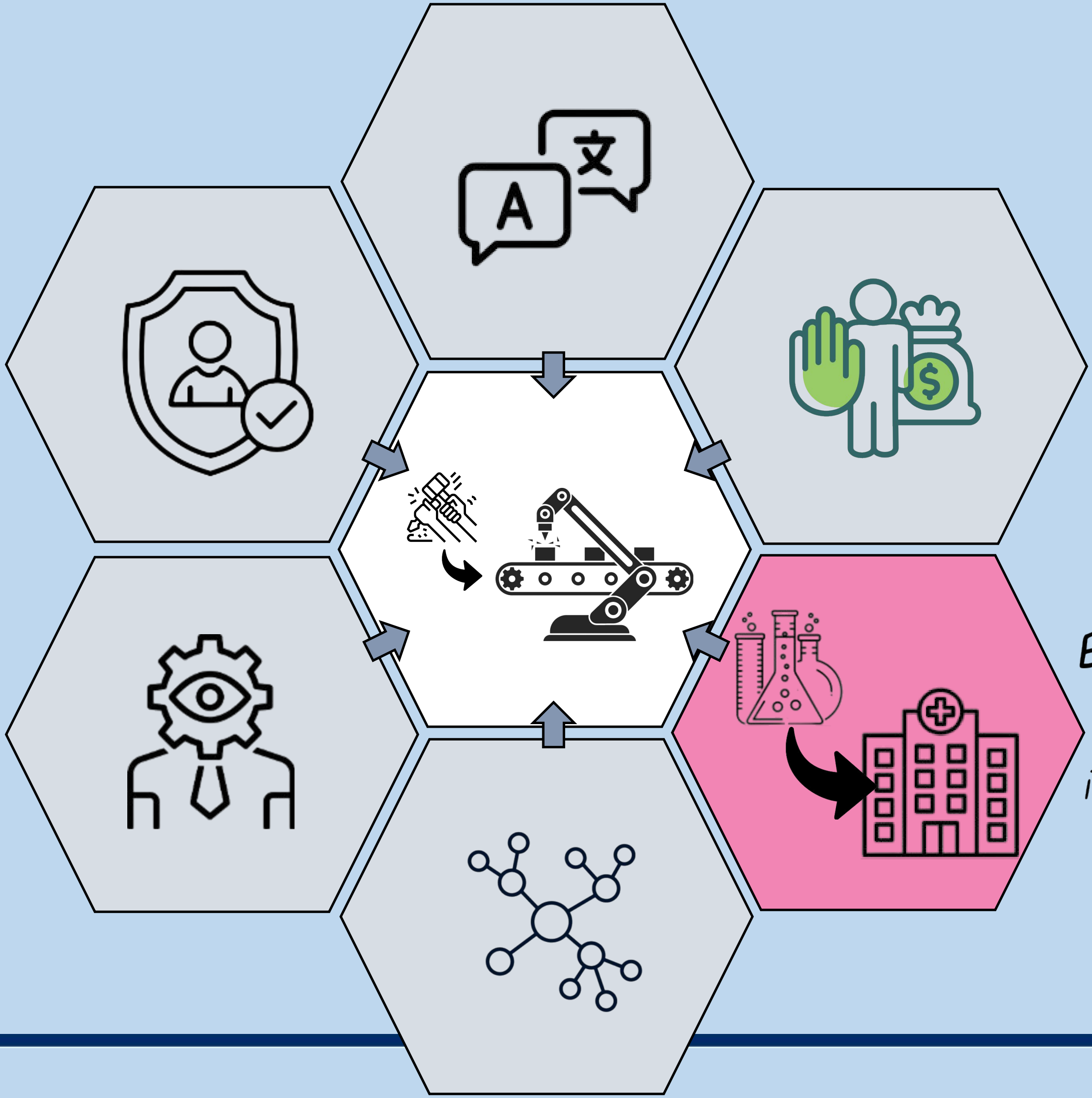
category	raw highlights				corrected highlights			
	TP	FP	FN	precision	TP(corr)	FP(corr)	FN(corr)	precision
'métastase', 'métastatique'	22	11	0	0.67	22	11	0	0.67
'stade iv', 'stade 4', 'stade IV'	9	2	0	0.82	9	2	0	0.82
'm0', 'm1'	3	1	0	0.75	3	1	0	0.75
'implants péritonéaux', 'implant péritonéal'	3	0	0	1	3	0	0	1
'carcinose'	1	8	0	0.11	1	8	0	0.11
't.n.m.', 'pt.n.m.', 'ct.n.m.', 'pt..n.m.'	7	4	0	0.64	7	4	0	0.64

Choose the lightest methods

More generally, try to

- Choose the lightest methods
- Choose the lightest models
- Choose scalable methods
- Estimate your energy consumption / carbon impact

Example 5: Evaluation in clinical settings



with

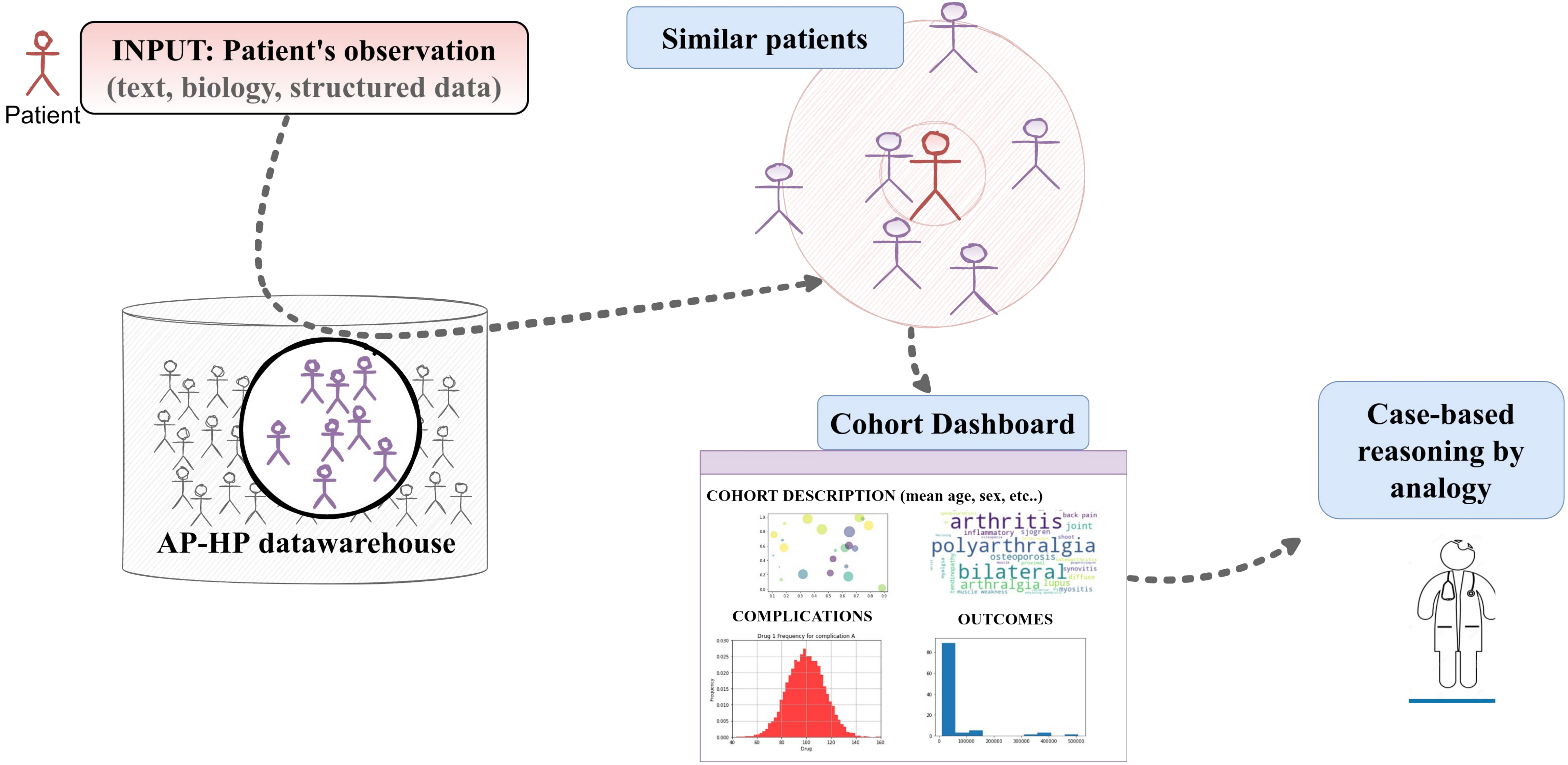
LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé

Evaluation in clinical settings

Interoperability with hospital information systems

Use case 1, similar patient selection

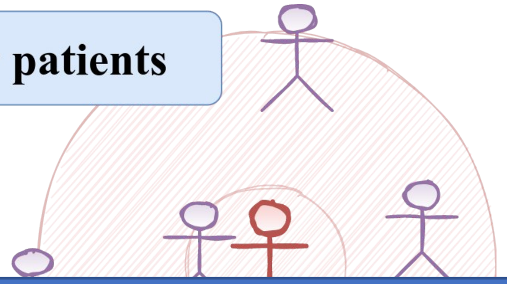


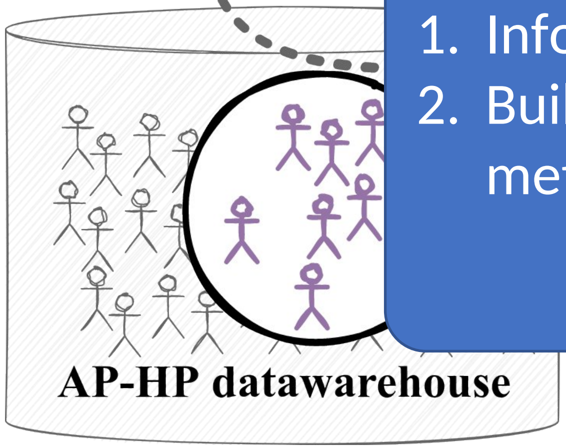
Use case 1, similar patient selection

Patient

INPUT: Patient's observation (text, biology, structured data)

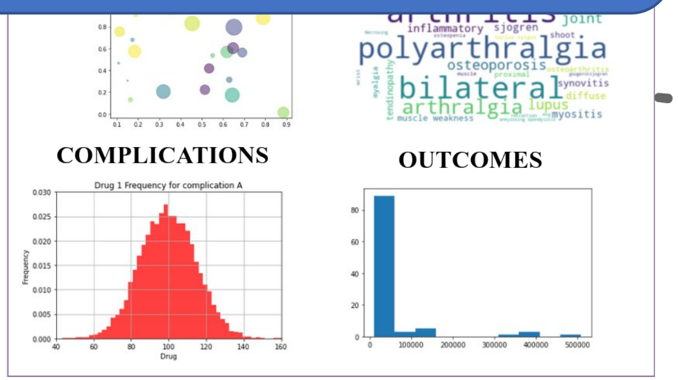
Similar patients





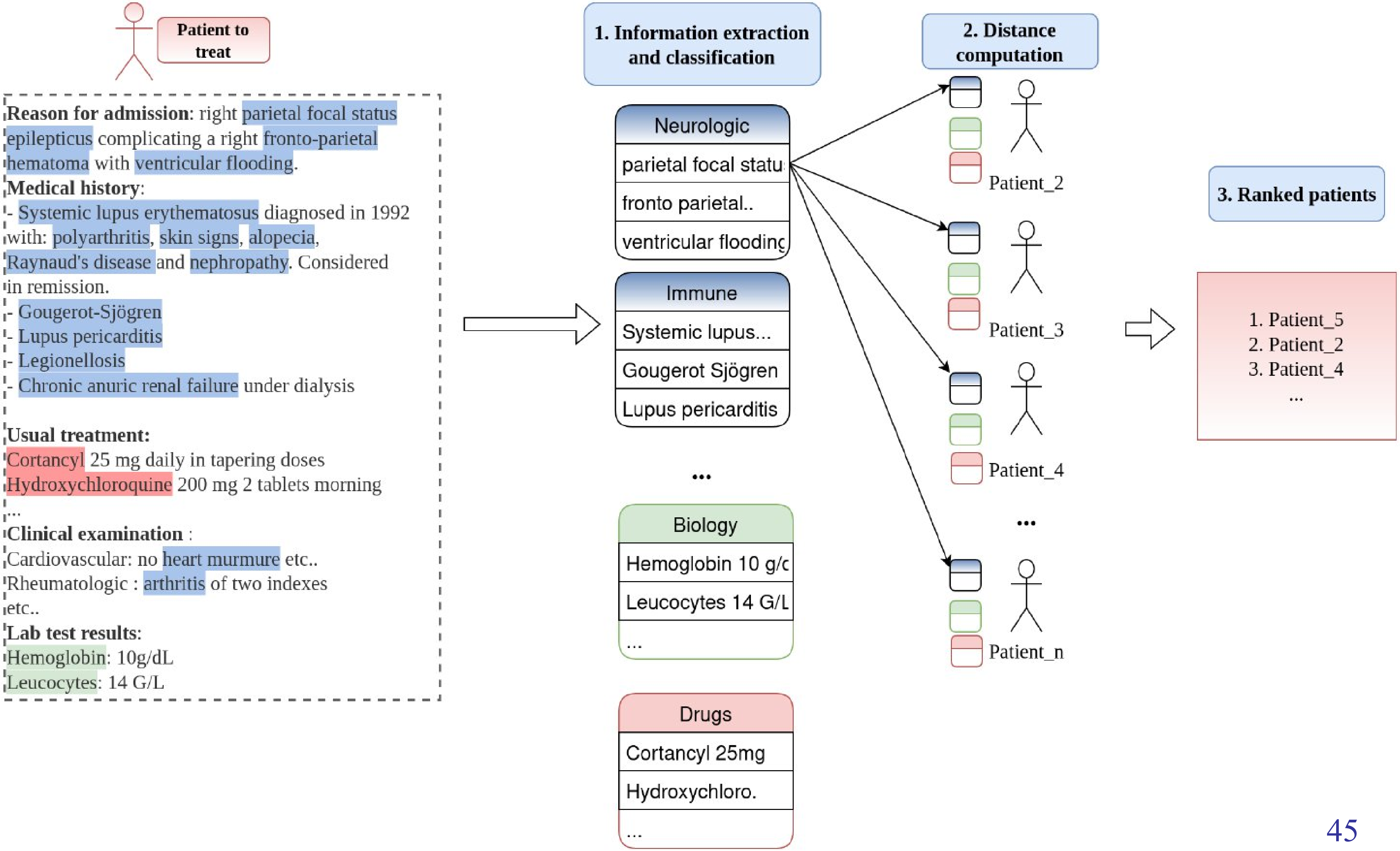
Methods :

1. Information retrieval / patient selection
2. Building clinically relevant similarity metrics

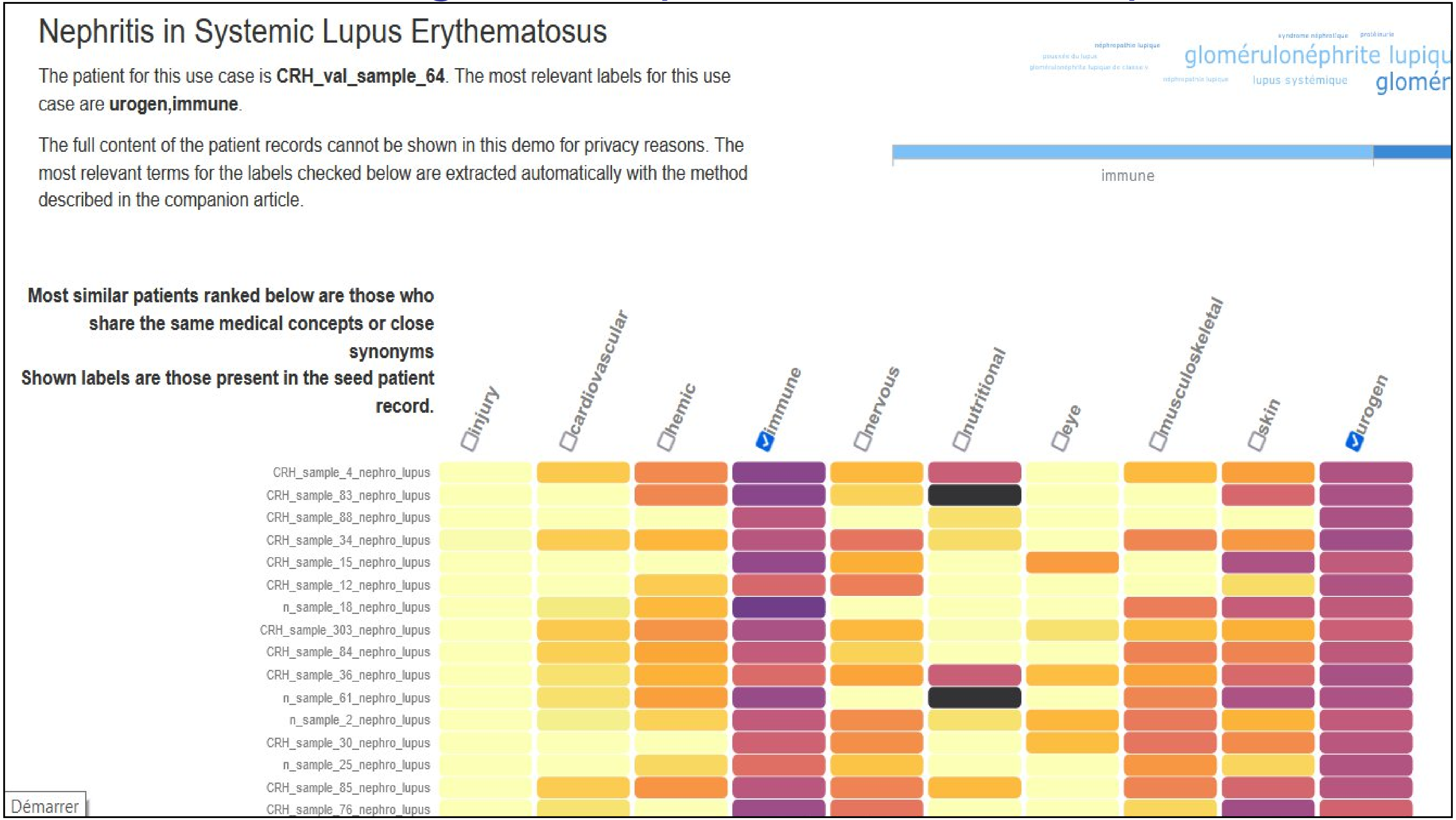


Case-based reasoning by analogy

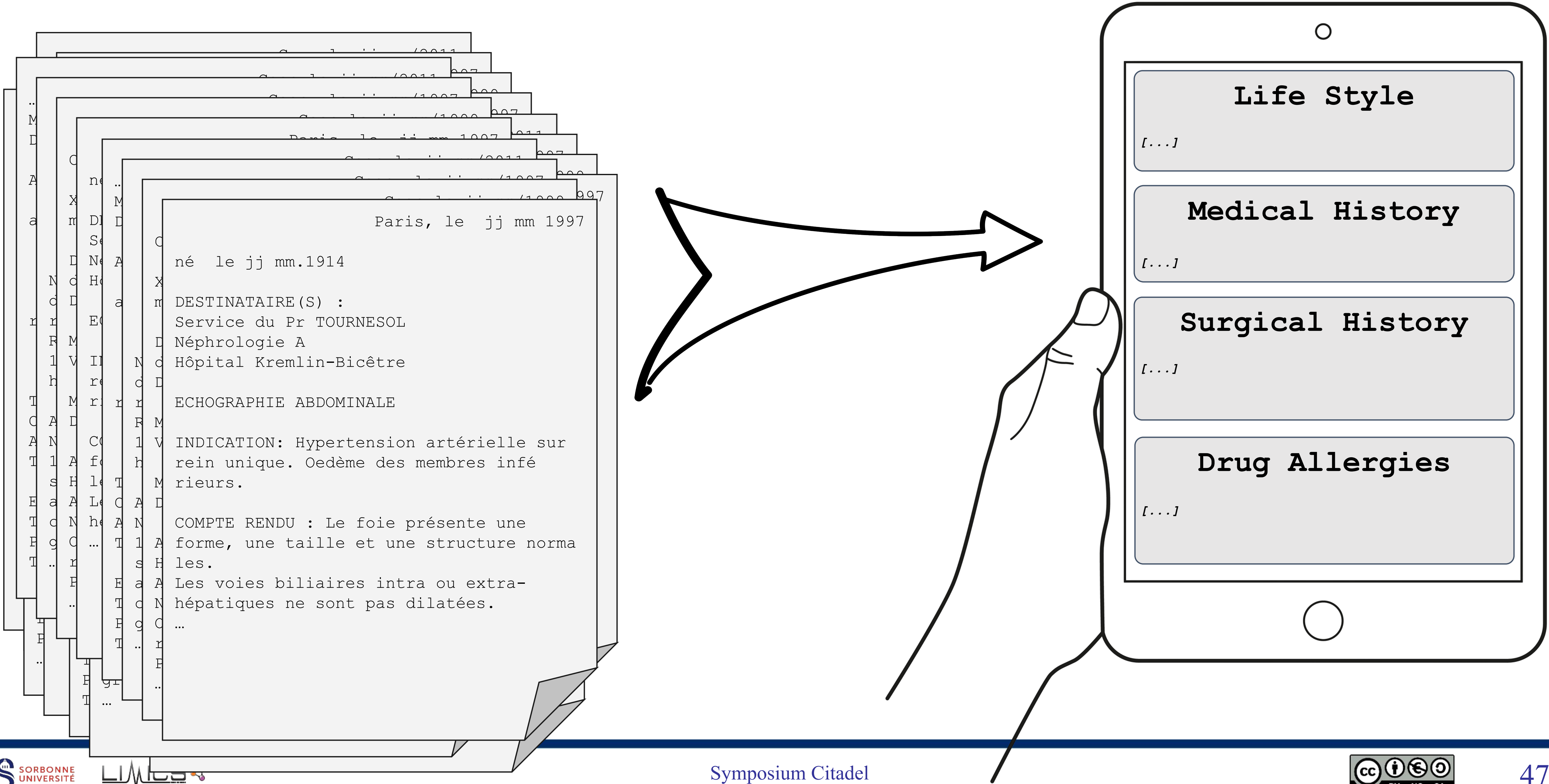
Use case 1, Building clinically relevant similarity metrics



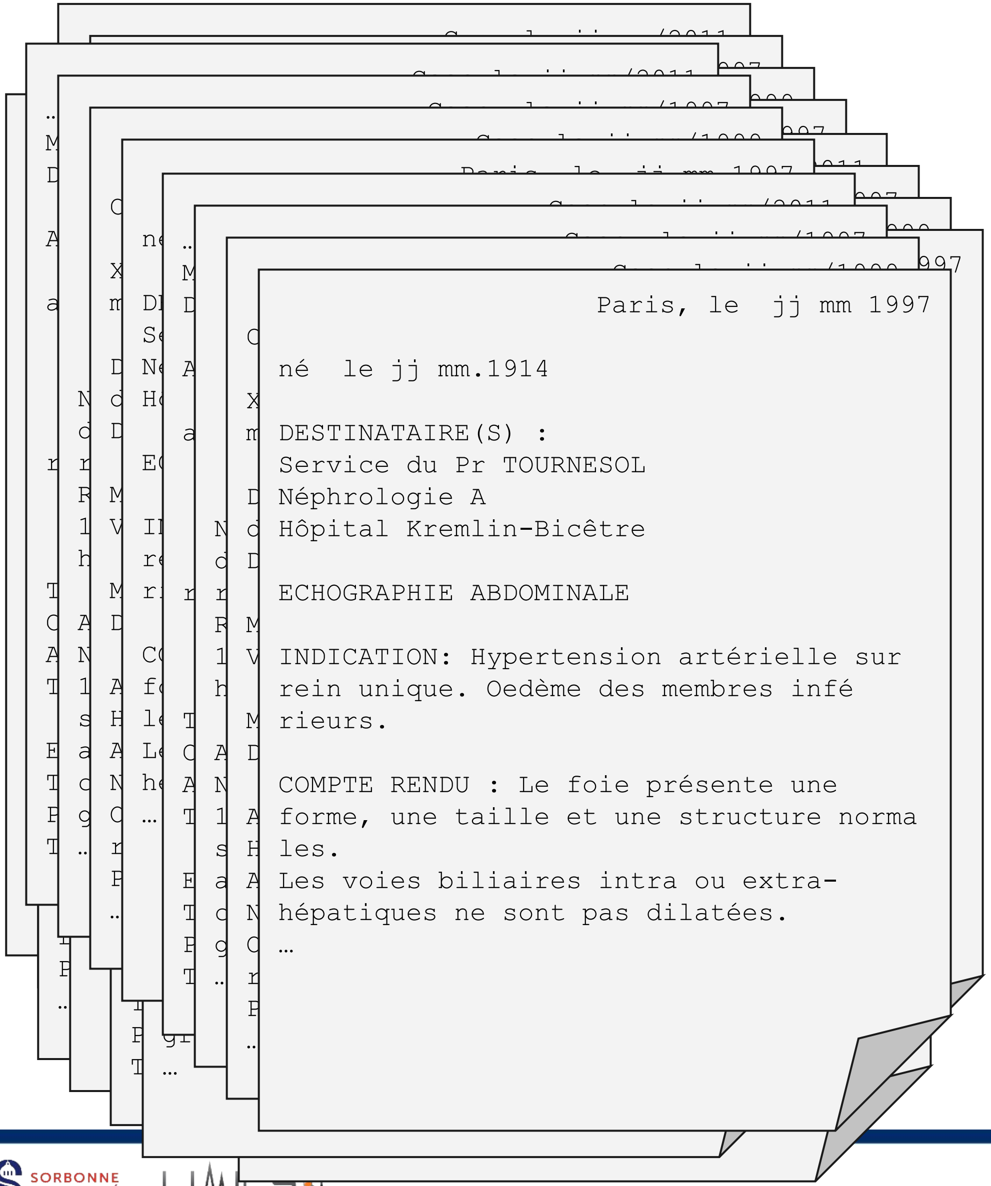
Use case 1, Building clinically relevant similarity metrics



Use case 2, ER patient summary



Use case 2, ER patient summary



Life Style

Patient divorcé
3 enfants : 19, 21 et 24 ans
Vit seul en appartement
Complètement autonome pour les actes de la vie quotidienne (juin 2024)
Performans Status 0 (juin 2024)
Travail comme conseiller en communication
Pas d'intoxication tabagique
Consommation éthylique : 1-2 verre de vin par jour
Dernier poids connu : 83 kg ;
Dernière taille connue : 175 cm

Surgical History

- **Résection recto-sigmoïdienne (janvier 2024)**
  Pour adénocarcinome rectal infiltrant
  Réalisée par le Pr Larzus en 1er opérateur
  Par coelioscopie puis laparo-convertie
  Anastomose colo-sus-anale mécanique
  Compliquée d'une embolie pulmonaire post opératoire (*cf antécédent médicaux*)
- **Appendicectomie** (dans l'enfance)
- **Cure de hernie inguinale droite** (date inconnue)

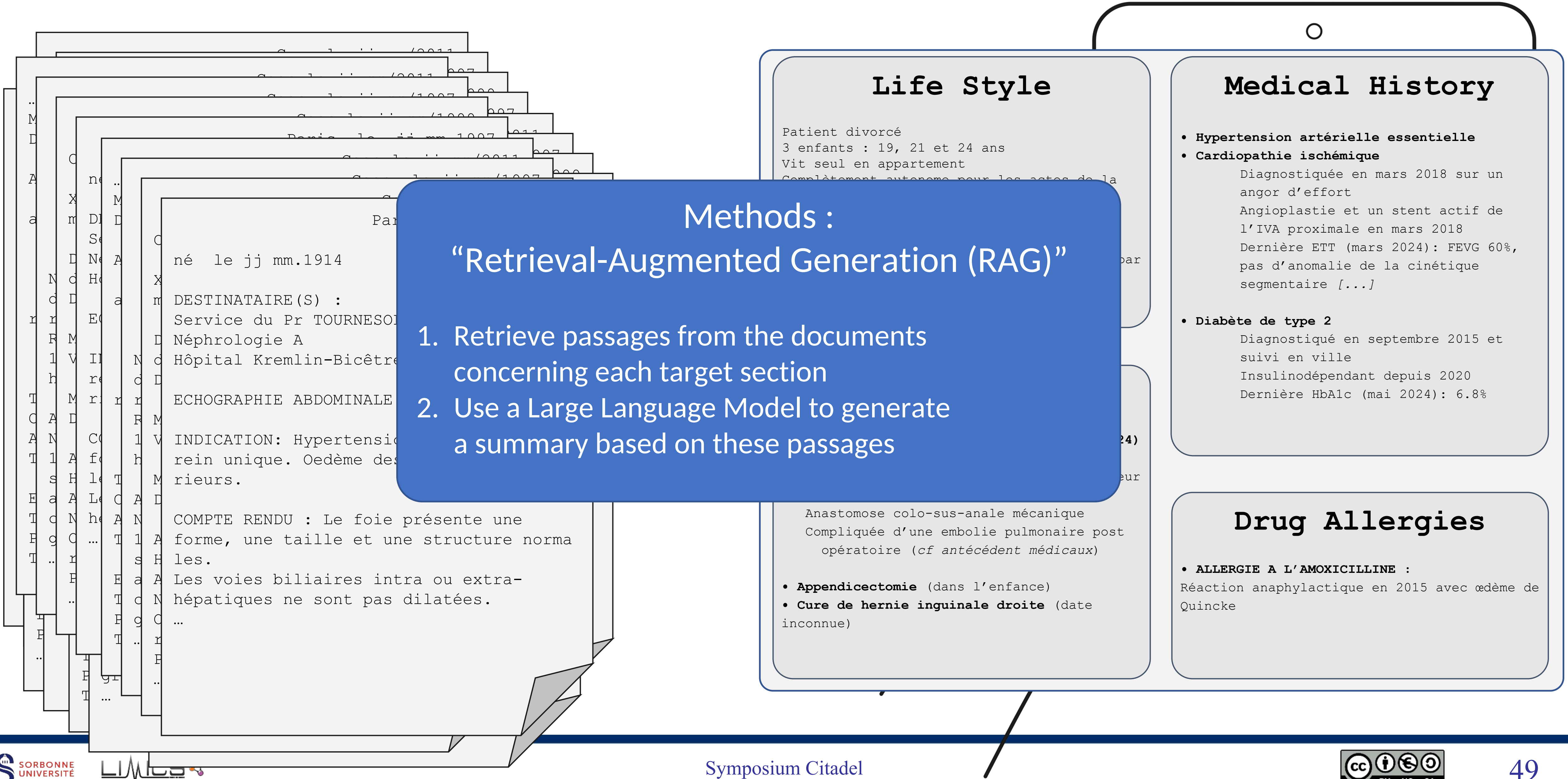
Medical History

- **Hypertension artérielle essentielle**
- **Cardiopathie ischémique**
  Diagnostiquée en mars 2018 sur un angor d'effort
  Angioplastie et un stent actif de l'IVA proximale en mars 2018
  Dernière ETT (mars 2024): FEVG 60%, pas d'anomalie de la cinétique segmentaire *[...]*
- **Diabète de type 2**
  Diagnostiqué en septembre 2015 et suivi en ville
  Insulinodépendant depuis 2020
  Dernière HbA1c (mai 2024): 6.8%

Drug Allergies

- **ALLERGIE A L'AMOXICILLINE** :
Réaction anaphylactique en 2015 avec œdème de Quincke

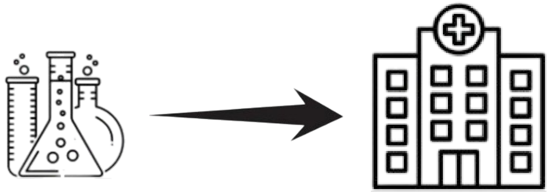
Use case 2, ER patient summary



Methods :
"Retrieval-Augmented Generation (RAG)"

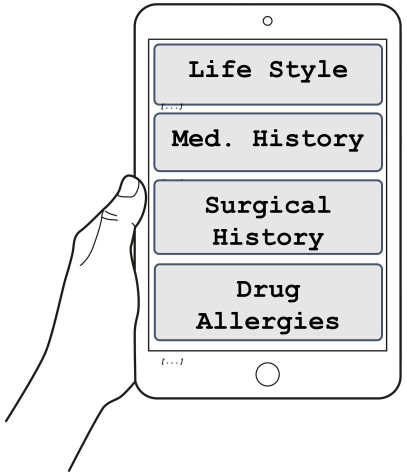
1. Retrieve passages from the documents concerning each target section
2. Use a Large Language Model to generate a summary based on these passages

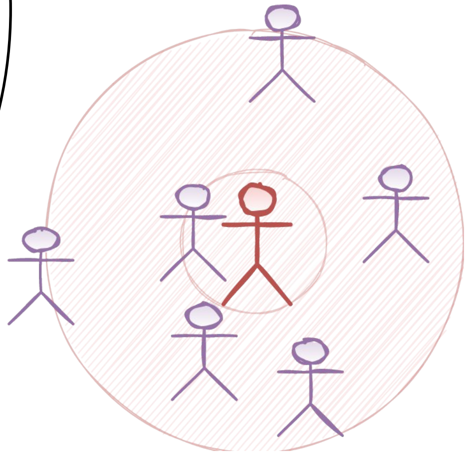
Evaluation in clinical settings



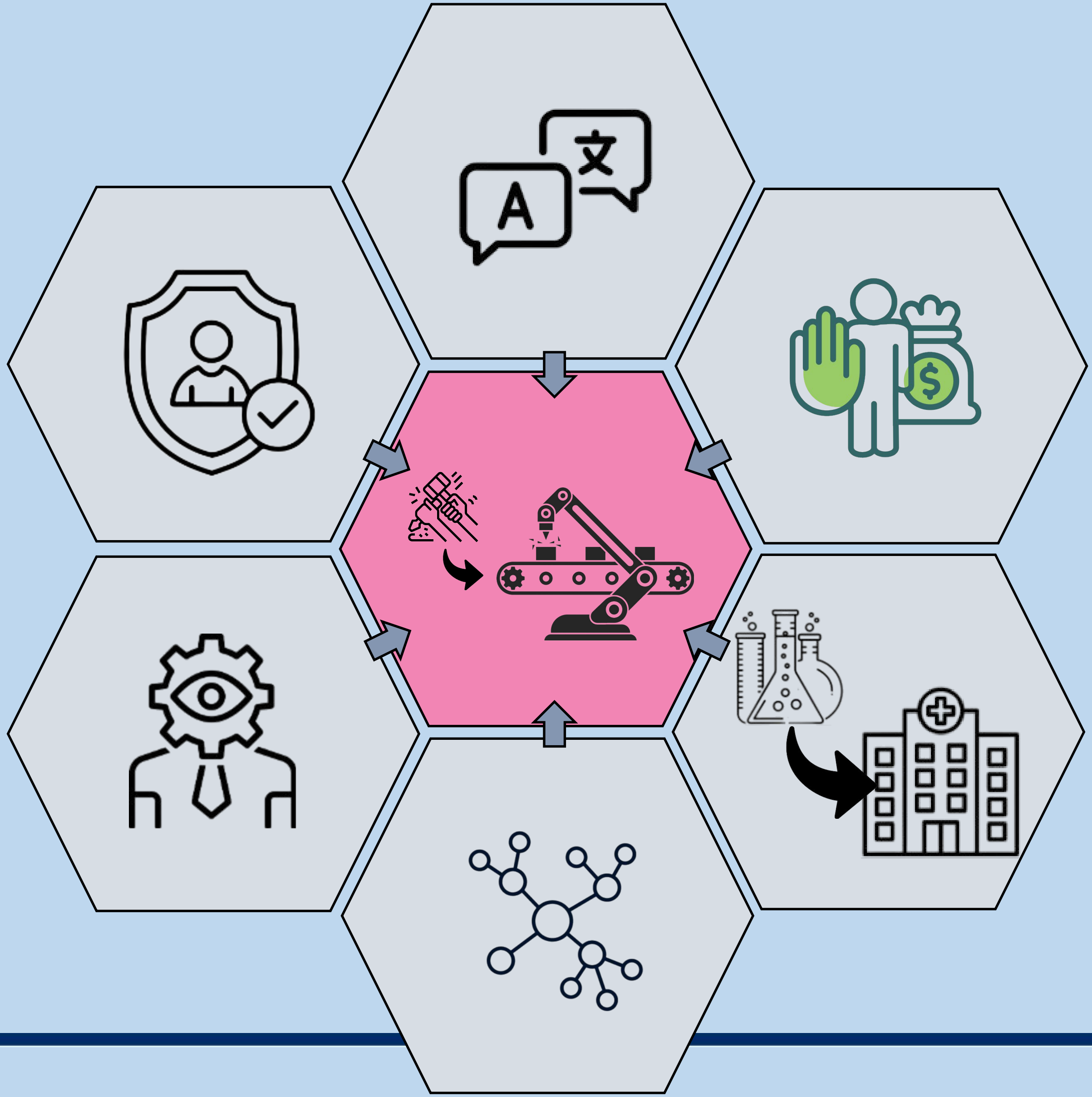
- Is it usable?
- Is it useful?
- Is it so useful that patients are given better care?
- Is it never harmful?
- Is the result a “medical device”?
- Who is responsible for missed information?
- …

- Interoperability issues (plug the system into the information system)
- Privacy issues (keep data and computing power inside the hospital)
- Computing resources issues (big data is… big)





Example 6: edsnlp *et al.*, the AP-HP “EDS” software suite



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé

The AP-HP "EDS" software suite

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé

- Named entity recognition
- Normalization to terminologies (concept unique identifiers)
- Patient classification

EDS-NLP

https://github.com/aphp/edsnlp

EDS-NLP is a collaborative NLP framework that aims primarily at extracting information from French clinical notes. At its core, it is a collection of components or pipes, either rule-based functions or deep learning modules. These components are organized into a novel efficient and modular pipeline system, built for hybrid and multitask models. We use spaCy to represent documents and their annotations, and Pytorch as a deep-learning backend for trainable components.

EDS-NLP is versatile and can be used on any textual document. The rule-based components are fully compatible with spaCy's components, and vice versa. This library is a product of collaborative effort, and we encourage further contributions to enhance its capabilities.

Check out our interactive demo !

The AP-HP "EDS" software suite

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé

EDS-PDF

https://github.com/aphp/edspdf

EDS-PDF provides a modular framework to extract text information from PDF documents.

You can use it out-of-the-box, or extend it to fit your specific use case. We provide a pipeline system and various utilities for visualizing and processing PDFs, as well as multiple components to build complex models:complex models:

- 📄 Extractors to parse PDFs (based on pdfminer, mupdf or poppler)
- 🎯 Classifiers to perform text box classification, in order to segment PDFs
- 🧩 Aggregators to produce an aggregated output from the detected text boxes
- 🧠 Trainable layers to incorporate machine learning in your pipeline (e.g., embedding building blocks or a trainable classifier)

Visit the 📖 documentation for more information!

The AP-HP “EDS” software suite

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé

https://github.com/aphp/eds-pseudo

EDS-Pseudo

The EDS-Pseudo project aims at detecting identifying entities in clinical documents, and was primarily tested on clinical reports at AP-HP's Clinical Data Warehouse (EDS).

The model is built on top of edsnlp, and consists in a hybrid model (rule-based + deep learning) for which we provide rules (`eds-pseudo/pipes`) and a training recipe `train.py` .

We also provide some fictitious templates (`templates.txt`) and a script to generate a synthetic dataset `generate_dataset.py` .

The AP-HP "EDS" software suite

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé

eds-scikit

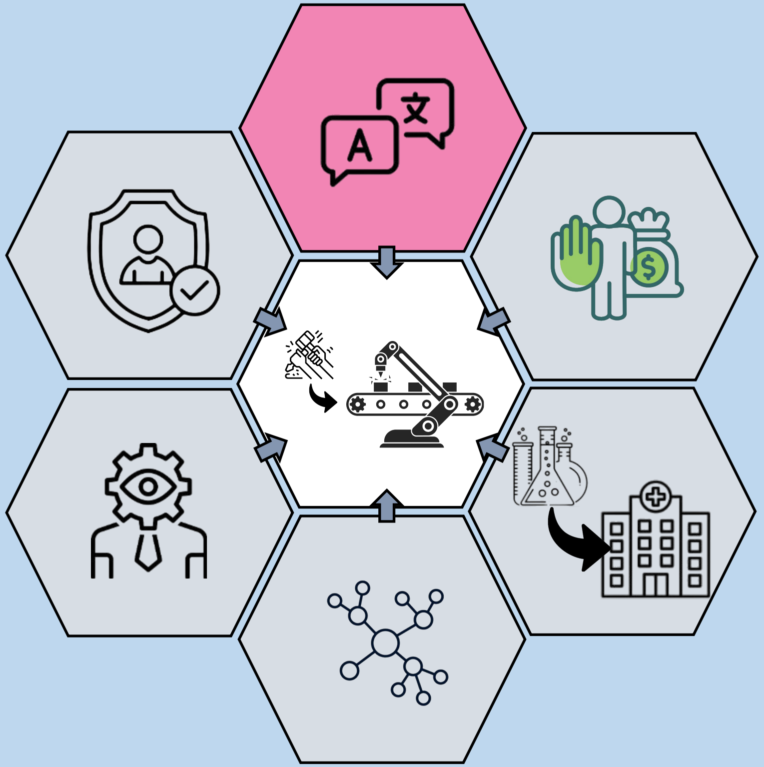
tests passing | docs passing | pypi v0.1.8 | python >= 3.7.1 | < 3.8 | code style black | coverage 94% | DOI 10.5281/zenodo.11636968

eds-scikit is a tool to assist data scientists working on the AP-HP's Clinical Data Warehouse. It is specifically targeted for OMOP-standardized data. It main goals are to:

- Ease access and analysis of data
- Allow a better transfer of knowledge between projects
- Improve research reproducibility

https://github.com/aphp/eds-scikit

And also…



with

iPLesp

Institut Pierre Louis d'Épidémiologie et de Santé Publique
Pierre Louis Institute of Epidemiology and Public Health

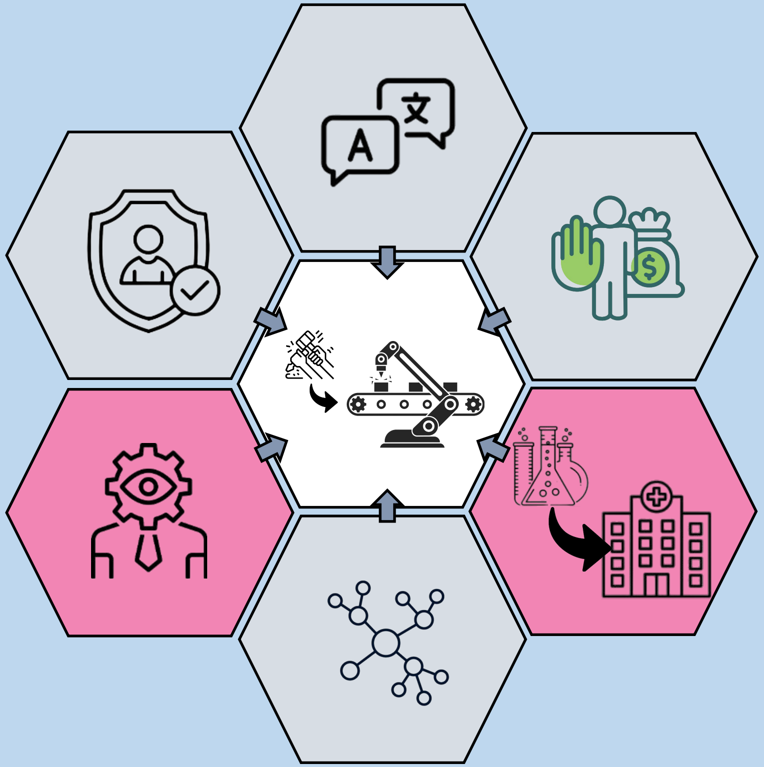
Translation

Is it a good idea to translate into English to benefit from the richer resources of that language?

(spoiler: it's complicated)

Gérardin et al. **Impact of translation on biomedical information extraction: an experiment on real-life clinical notes**. JMIR Medical Informatics, 2024

And also…



with

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

UNIVERSITÉ DE GENÈVE

Generative LLMs

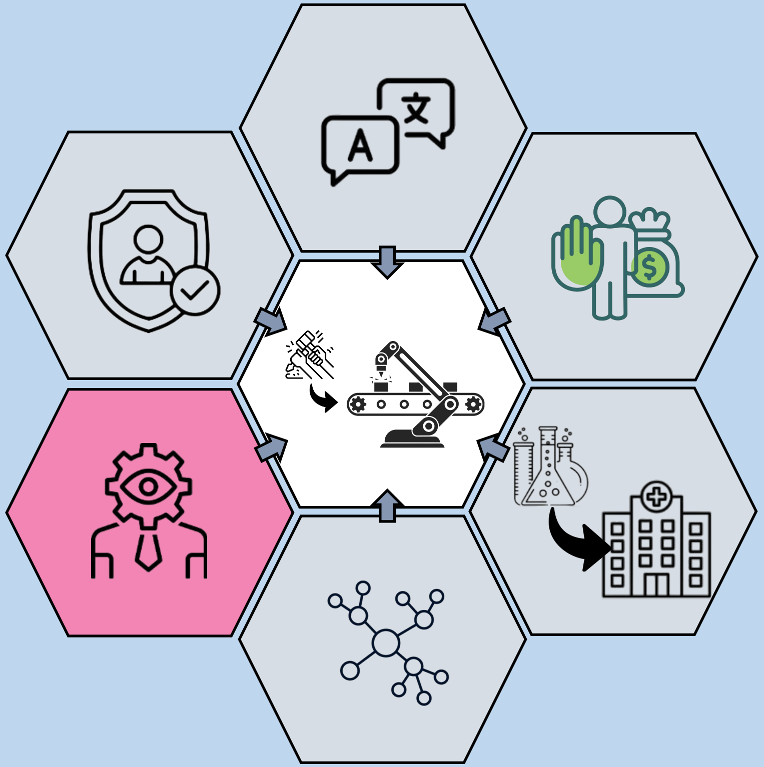
Performance evaluation of generative LLMs in different configurations

Naguib et al. **Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting**. EMNLP 2024

Zaghir et al. **Prompt Engineering Paradigms for Medical Applications: Scoping Review**. JMIR, 2024

Tai et al. **Clinical trial cohort selection using Large Language Models on n2c2 Challenges**. Submitted 2025

And also…



with

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

Active learning & Few-shot

Can we

- optimize the choice of annotated documents
- pre-annotate
- learn from few samples

to save human time?

Naguib et al. **Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting**. EMNLP 2024

Naguib et al. **Stratégies d'apprentissage actif pour la reconnaissance d'entités nommées en français**. TALN 2023


Thank you!
Questions ?

Backup slides…

Methods

1. **Rule-based systems**
2. Supervised learning systems
3. Generative, large language models
4. Retrieval-Augmented Generation

Rule-based systems

1. Terminological approach

VASCULAR DISEASE

- *intermittent claudication*
- *arteriopathy of the limbs*
- *arteritis of limbs*
- *arterial ulcer*
- *venous ulcer*
- *arterial insufficiency*
- *gangrene*
- *acute limb ischemia*
- *acute peripheral ischemia*
- *aortic aneurysm*

(Approximate) term matching in the documents

VASCULAR DISEASE

Patient with arterial ulcers of the lower limbs, followed at St Joseph with skin graft in October 2015

Rule-based systems

1. Terminological approach
2. Additional rules

BACTERIA RESISTANCE

- *<BACTERIA> + .* sensitive*
- *<BACTERIA> + .* resistant*
- *<BACTERIA> + sensitive to methicillin*
- …

9 blood cultures positive on 6/26/15 for methicillin-sensitive Staphylococcus aureus.

BACTERIA RESISTANCE

Rule-based systems

1. Terminological approach
2. Additional rules
3. Trickier rules

Emmanuelle Kempf et al., 2023

secondary, progressive, fractured or suspected osteocondensation	*ost[eé]ocondensa. {1,20} (suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image). {1,20}os. {1,30} (suspect|secondaire|[ée]volutive)|os. {1,30} (l[ée]sion|anomalie|image). {1,20} (suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image). {1,20}L[I,Y]tique|(l[ée]sion|anomalie|image). {1,20}condensant. {1,20} (suspect|secondaire|[ée]volutive)|fracture. {1,30} (suspect|secondaire|[ée]volutive)|((l[l[ée]sion|anomalie|image|nodule). {1,80} (secondaire))|((l[l[ée]sion|anomalie|image|nodule)s.{1,40}suspec?ts?).*
post-operative anatomopathological tumor stage (pTNM)	*([ycpP]{1,2}\s? (T([01234x]|is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo01234\+][abcdx]?) *\s? (M[o01]? [\+x]?)?)|((T([01234x]|is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo01234\+][abcdx]?) \s? (M[o01]? [\+x]?)?)*

Rule-based systems: Pros & Cons

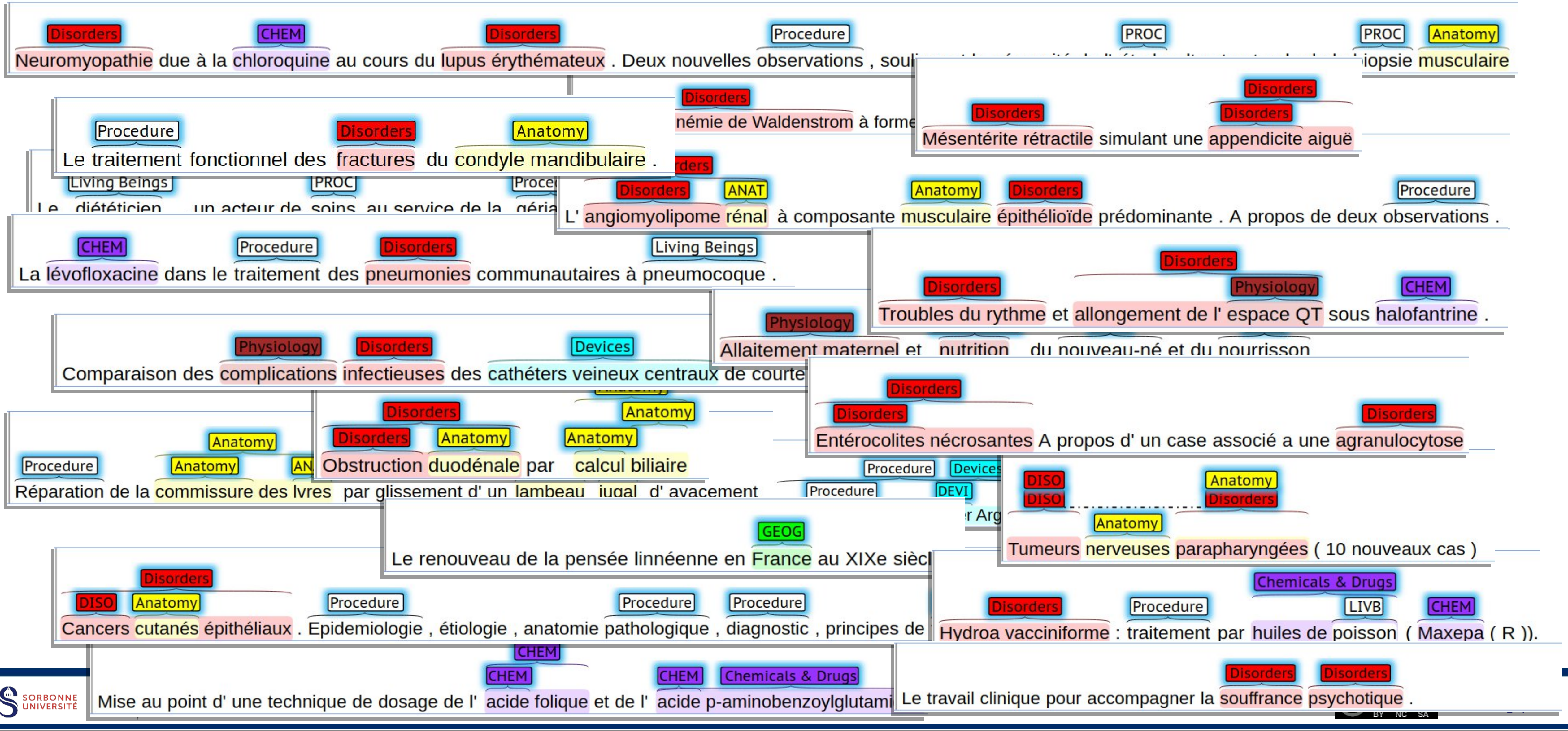
	Rules
General performance	
Ease of implementation	
Need for human expertise	
Explainability	
Material resources	
Energy consumption	
Ease of maintenance	
Generalization to a different problem/context	

Methods

1. Rule-based systems
2. **Supervised learning systems**
3. Generative, large language models
4. Retrieval-Augmented Generation

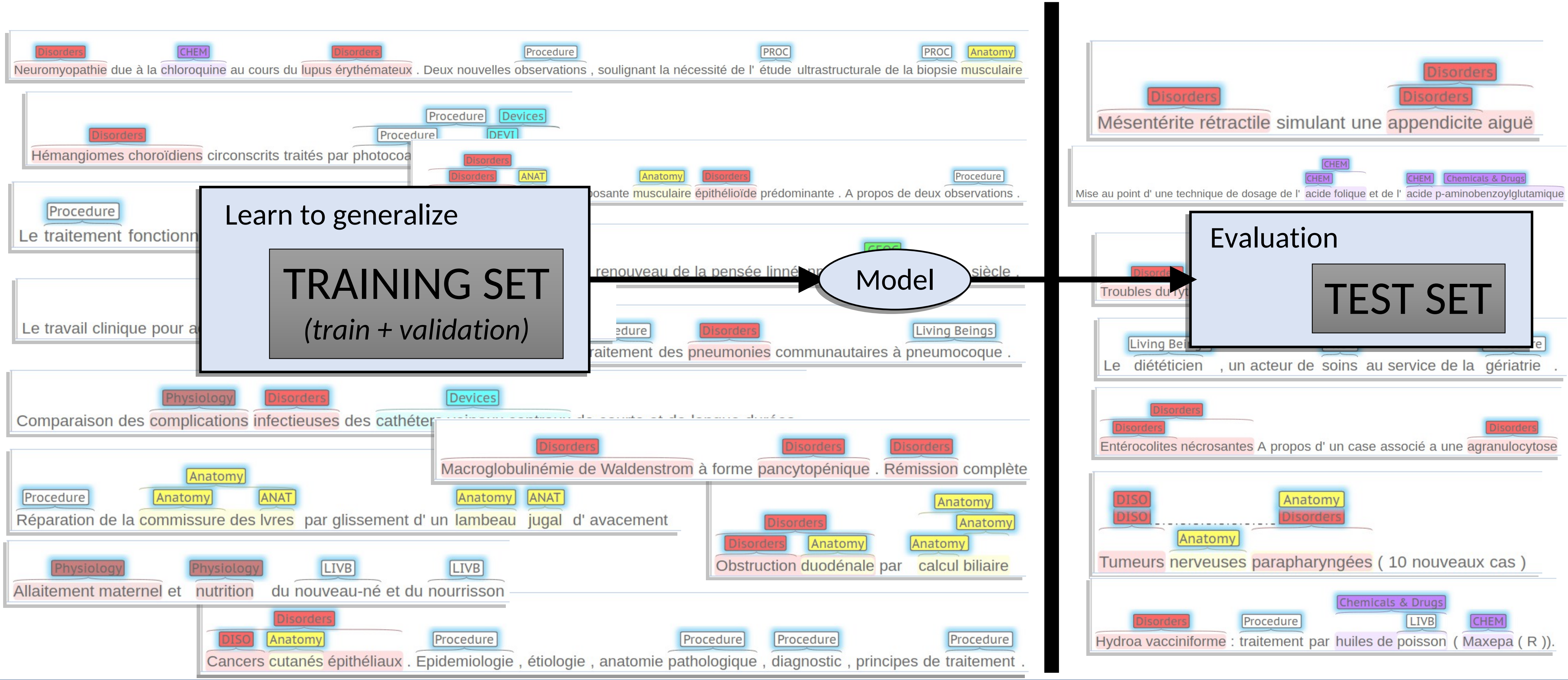
Supervised learning systems

Névéol, A., Grouin, C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014).
The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.

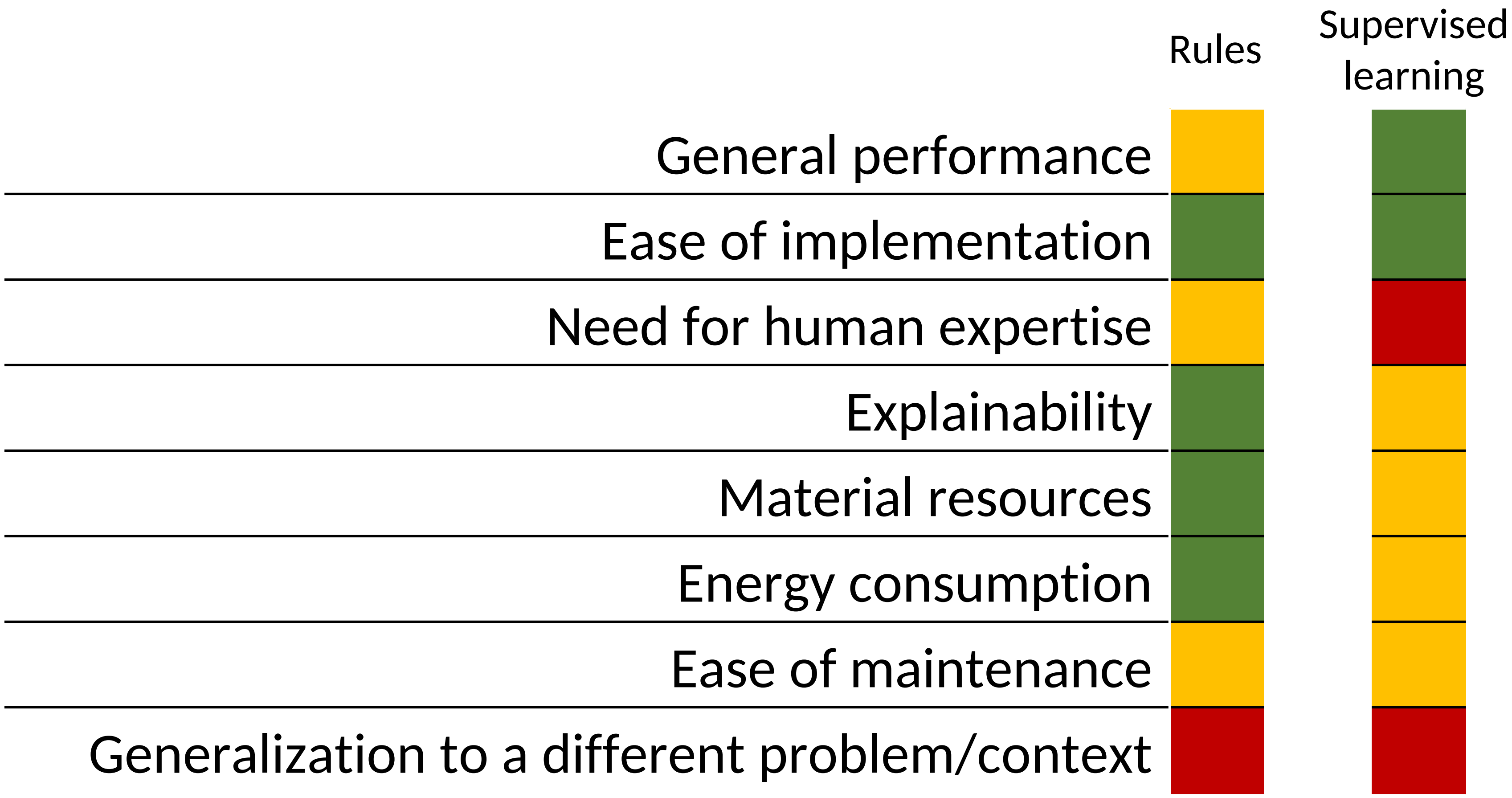


Supervised learning systems

Névéol, A., Grouin, C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014).
The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.



Supervised learning systems: Pros & Cons



Methods

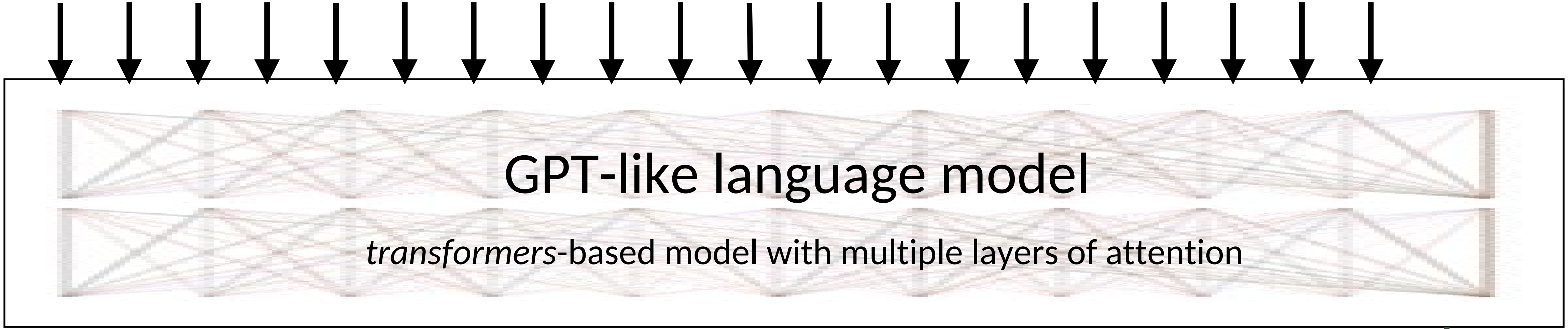
1. Rule-based systems
2. Supervised learning systems
3. **Generative, large language models**
4. Retrieval-Augmented Generation

Large Language Models

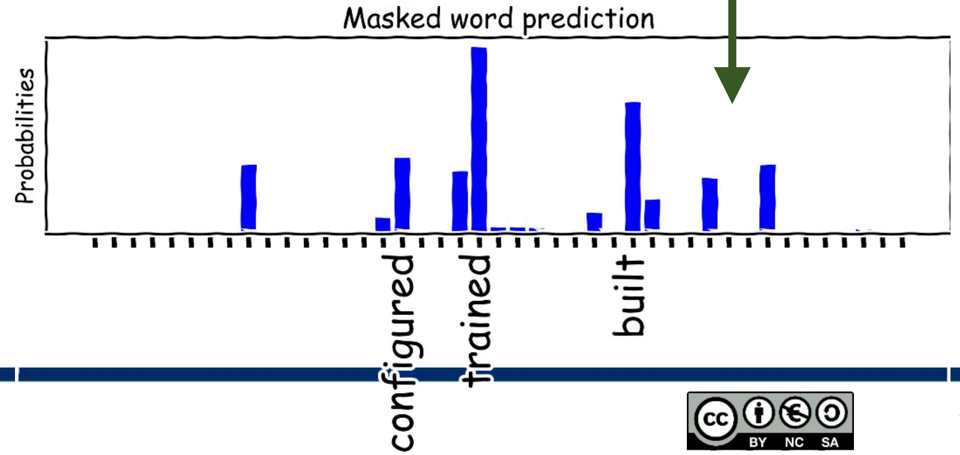
- Generative large language models (LLMs) are trained to produce human-like text.
- They simulate human understanding by predicting and generating text based on the input they receive.
- They can be finetuned for different tasks, e.g rewriting a prompt in another styles.

Auto-regressive language models (*decoder-style*, *e.g. GPT*)

```
User: What is a large language model?
Assistant: A large language model is a type of artificial intelligence (AI)
model designed to understand and generate human-like language. These models
are
```

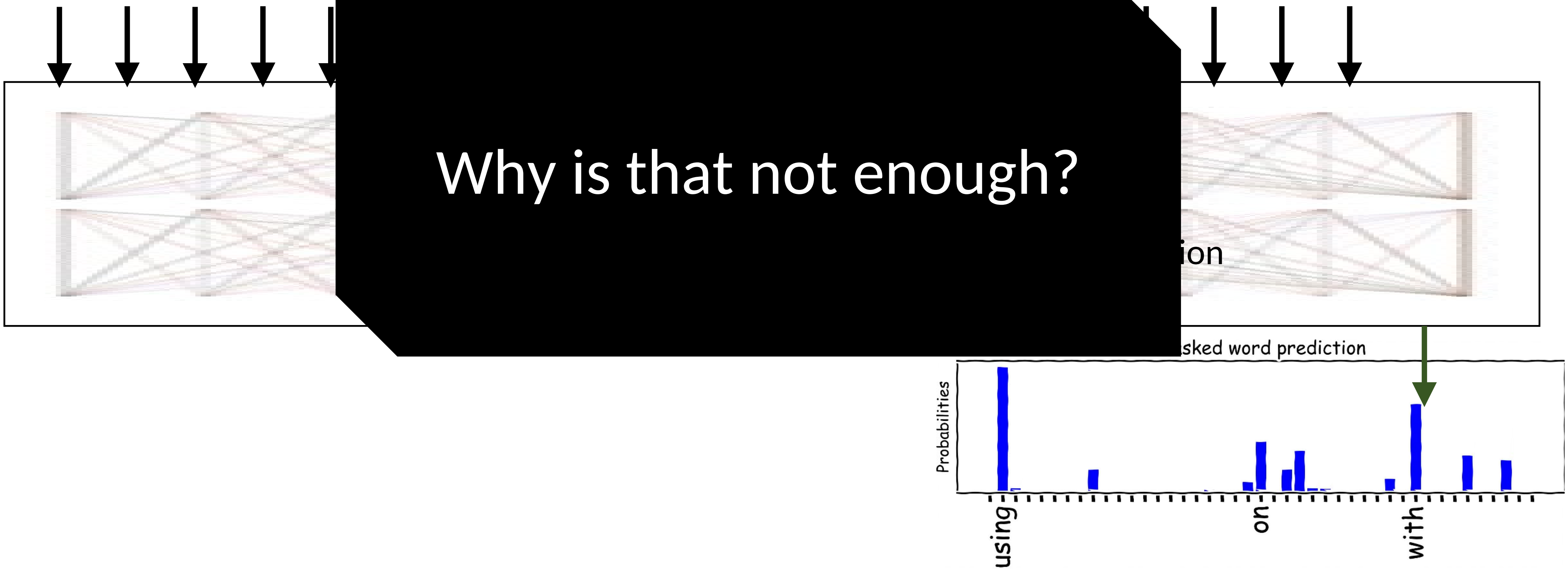


Next tokens are randomly selected from the predicted distribution, introducing variability in the generated output



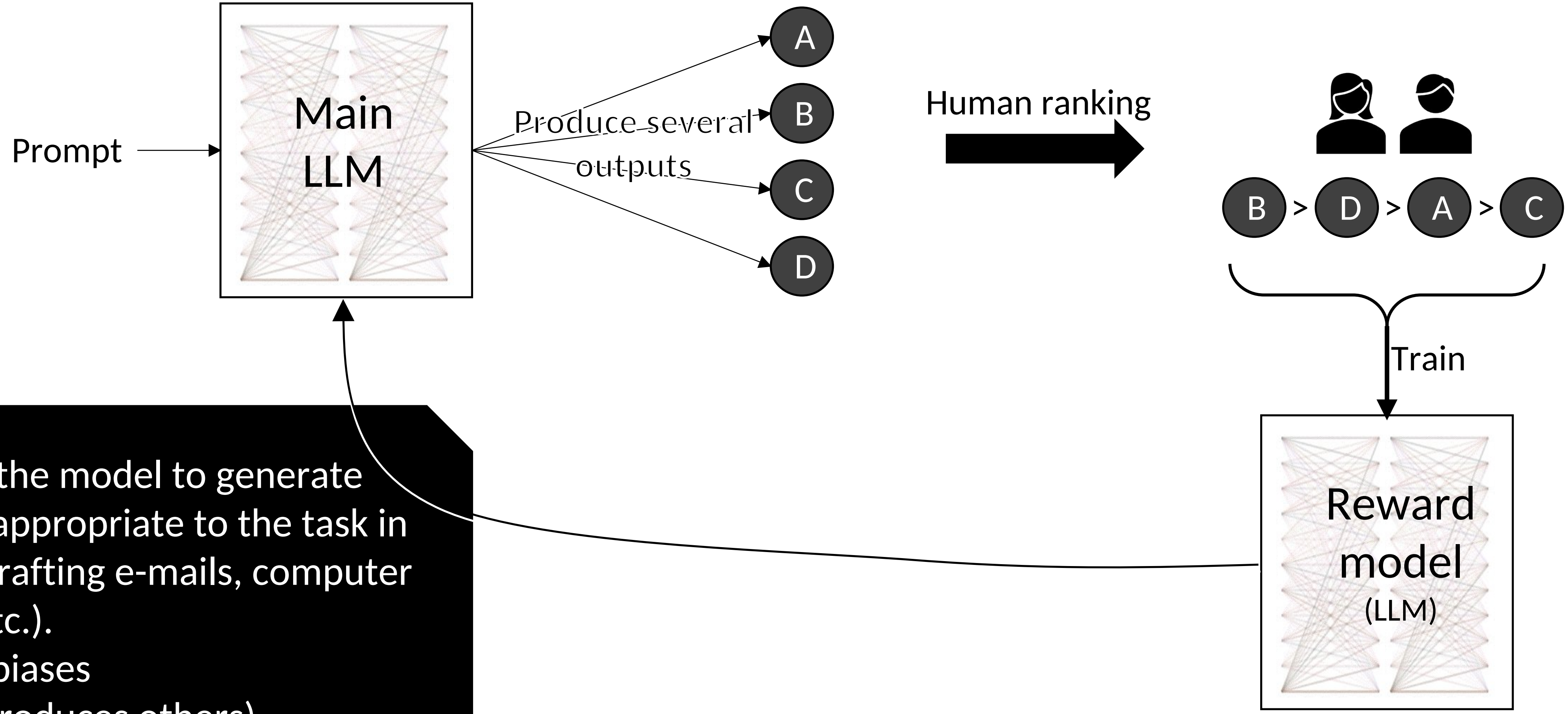
Auto-regressive language models (*decoder-style*, *e.g. GPT*)

```
User: What is a large language model?
Assistant: A large language model is a type of artificial intelligence (AI)
model designed to understand and generate human-like language. These models
are built
```



Why is that not enough?

Reinforcement Learning with Human Feedback



- Adapts the model to generate results appropriate to the task in hand (drafting e-mails, computer code, etc.).
- Avoids biases
(and introduces others)

Why are LLMs so good?

- LLMs are good at producing relevant, well-written and convincing information, thanks to:
  - Representation learning
  - Contextual understanding and attention mechanisms
    (capture long-range dependencies and relationships in the data)
  - Scalability and massive amount of training data
    (wide range of linguistic nuances and topics)
  - Massive finetuning with human feedback
    (produce the right kind of results for the task prompted by the user)

- LLMs are NOT good at:
  - Factuality / knowledge
  - High precision
  - Humor, creativity, originality

How can we use LLMs?

- Direct prompt: *"How do genetic mutations in the ALMS1 gene contribute to the pathophysiology of Alström Syndrome?"*
  ► only knowledge from the pretraining

- Prompt with persona: ***"You're an assistant specialized in research on rare diseases****. How do genetic mutations in the ALMS1 gene contribute to the pathophysiology of Alström Syndrome?"*
  ► allows to guide the answer and its style

- Prompt with document: ***"Given this document****, how do genetic mutations in the ALMS1 gene contribute to the pathophysiology of Alström Syndrome?"*
  ► knowledge from the pretraining + supporting document

How can we use LLMs?

- Few-shot prompting: *“*

  Classify the following rare diseases into their appropriate categories: Genetic Disorder or Neurodegenerative Disorder

  Input: "Alström Syndrome"

  Output: "Genetic Disorder"

  Input: "Batten Disease"

  Output: "Neurodegenerative Disorder"

  Input: "Huntington's Disease"

  Output: "Neurodegenerative Disorder"

  Input: "Marfan Syndrome"

  Output: "Genetic Disorder"

  Input: "Parkinson's Disease"

  Output: "Neurodegenerative Disorder"

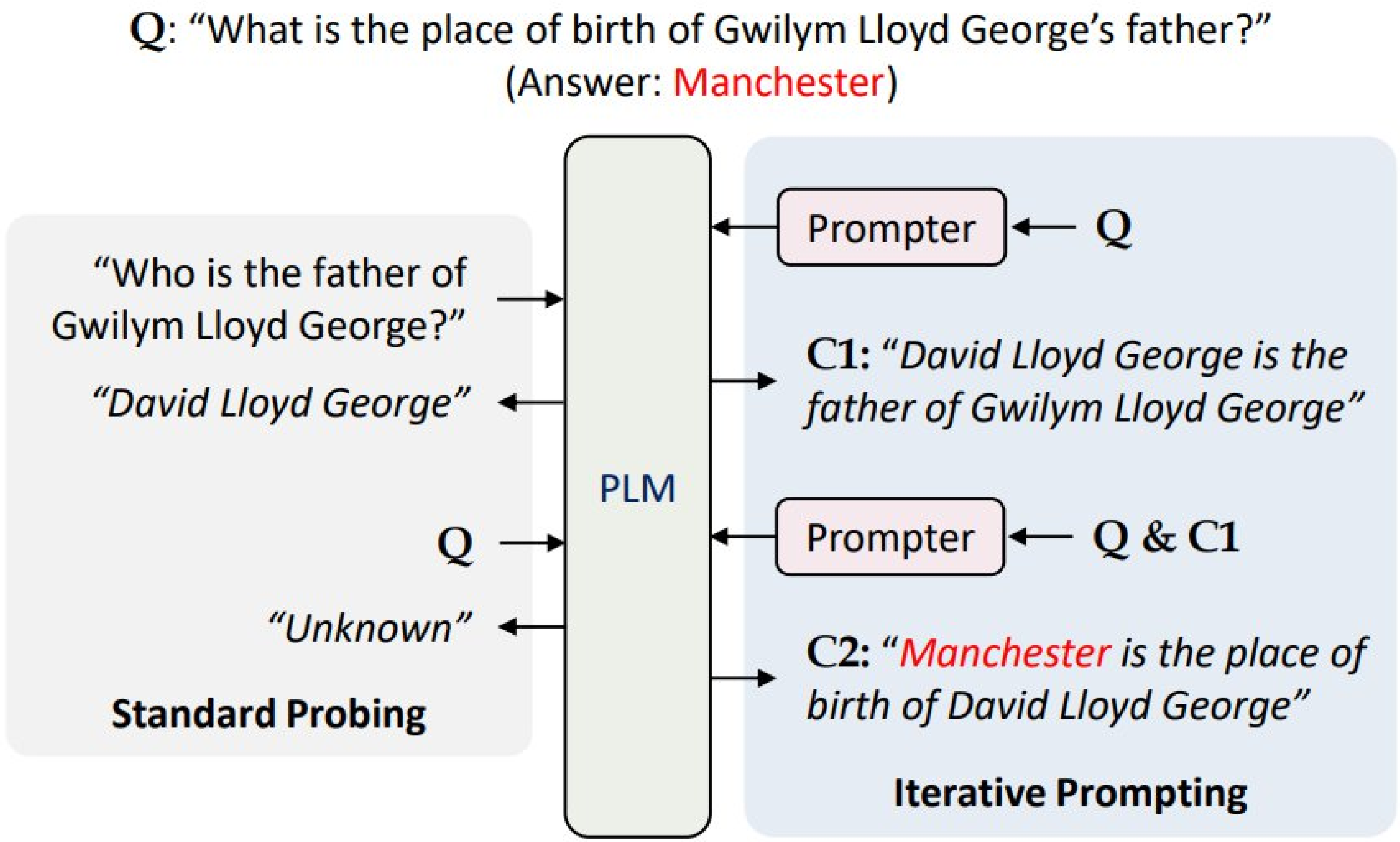
  Input: "Cystic Fibrosis"

  Output:

”

How can we use LLMs?

- Prompt chaining:



Wang, B., Deng, X., & Sun, H. (2022). Iteratively Prompt Pre-trained Language Models for Chain of Thought. *Conference on Empirical Methods in Natural Language Processing*.

How can we use LLMs?

- Chain of thoughts:

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

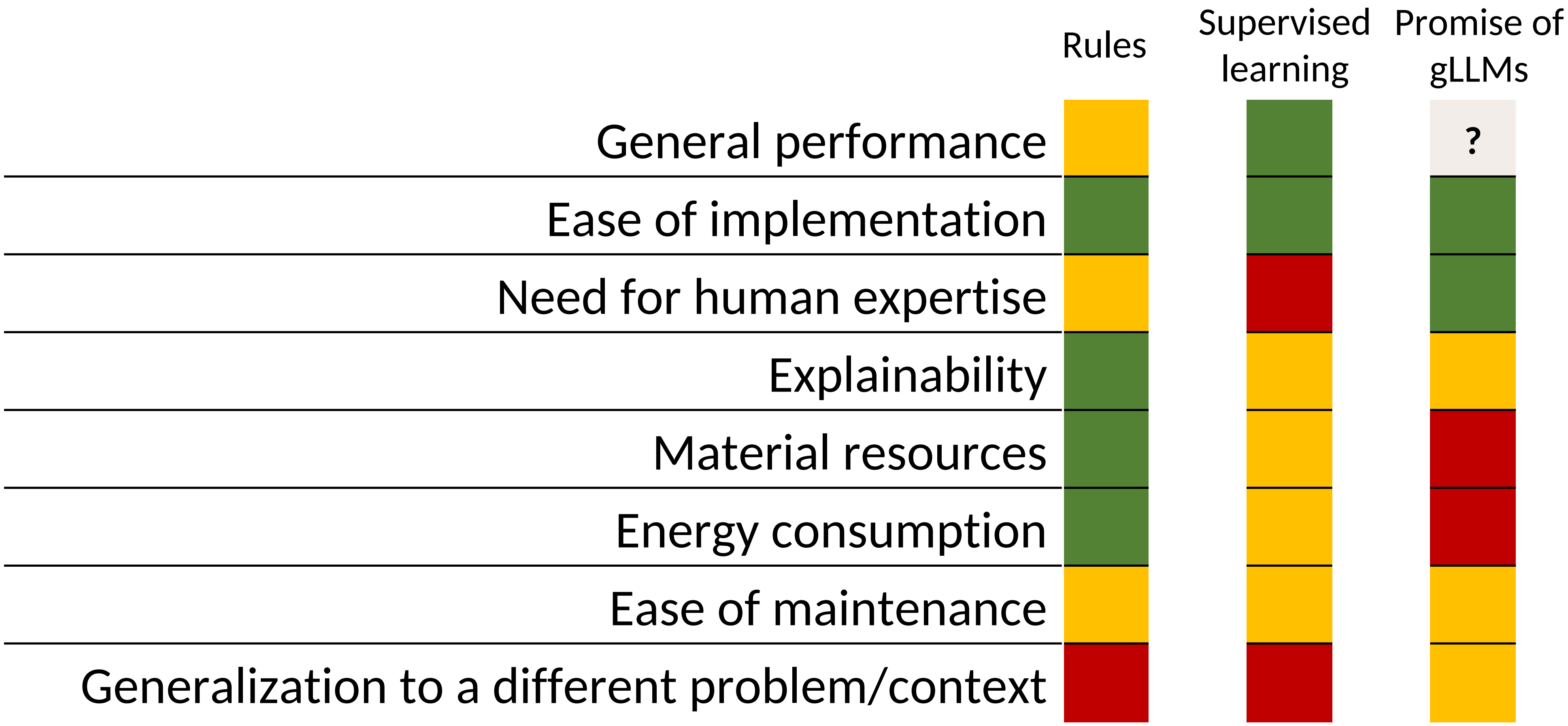
A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *Conference on Neural Information Processing Systems*.

How can we use LLMs?

- Self-Consistency
- Generated Knowledge Prompting
- Tree of Thoughts (ToT)
- Automatic Prompt Engineer (APE)
- Active-Prompt
- Directional Stimulus Prompting
- Program-Aided Language Models
- … (see for example https://www.promptingguide.ai/)

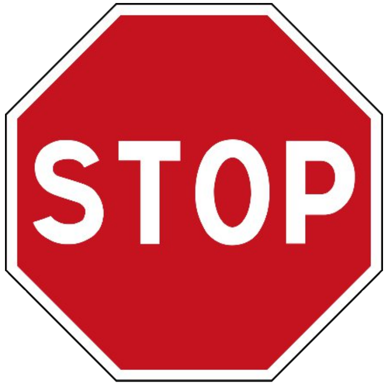
Generative LLMs: Pros & Cons



A thorough evaluation of these models
is always necessary!

Whatever the method (rules, ML, LLMs…),
build a test set for the evaluation

Do not prompt public or API-based LLMs
with sensitive data!

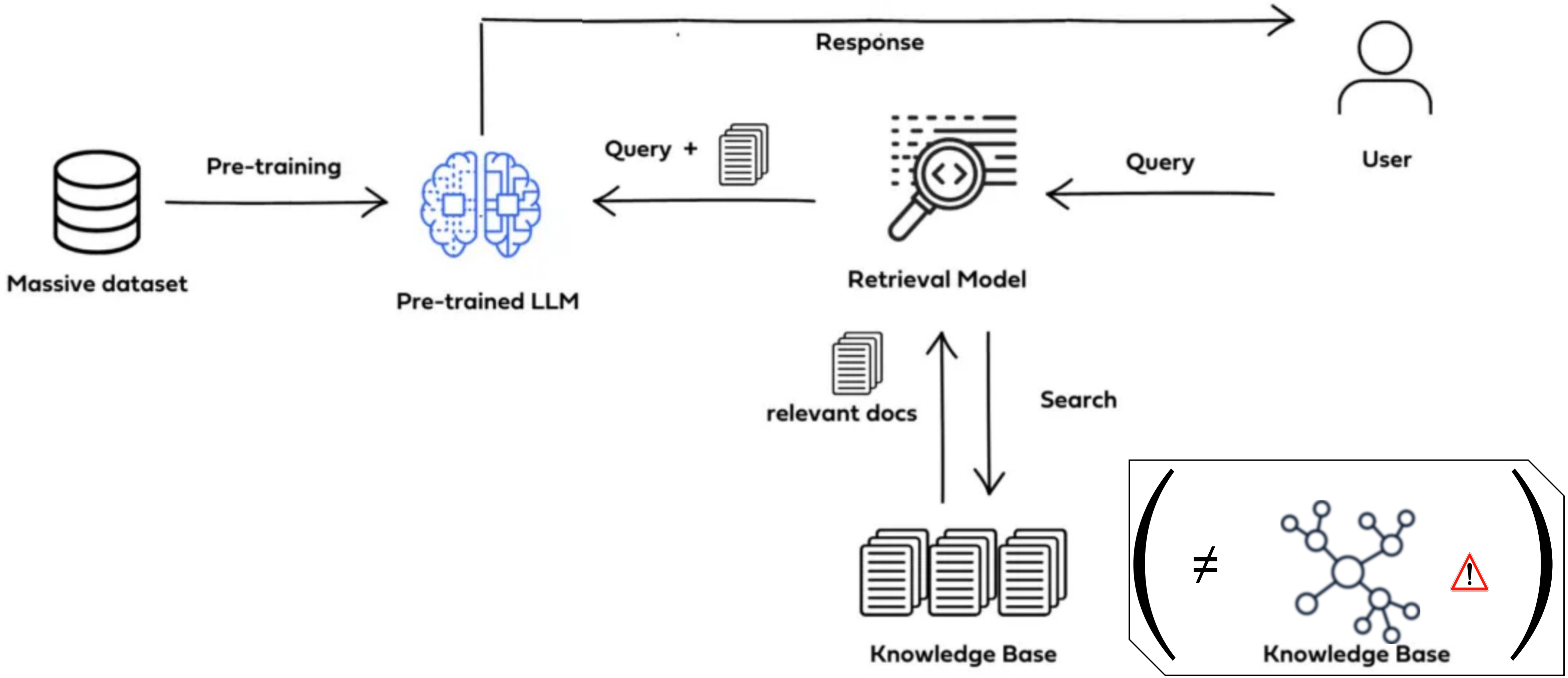


Methods

1. Rule-based systems
2. Supervised learning systems
3. Generative, large language models
4. **Retrieval-Augmented Generation**

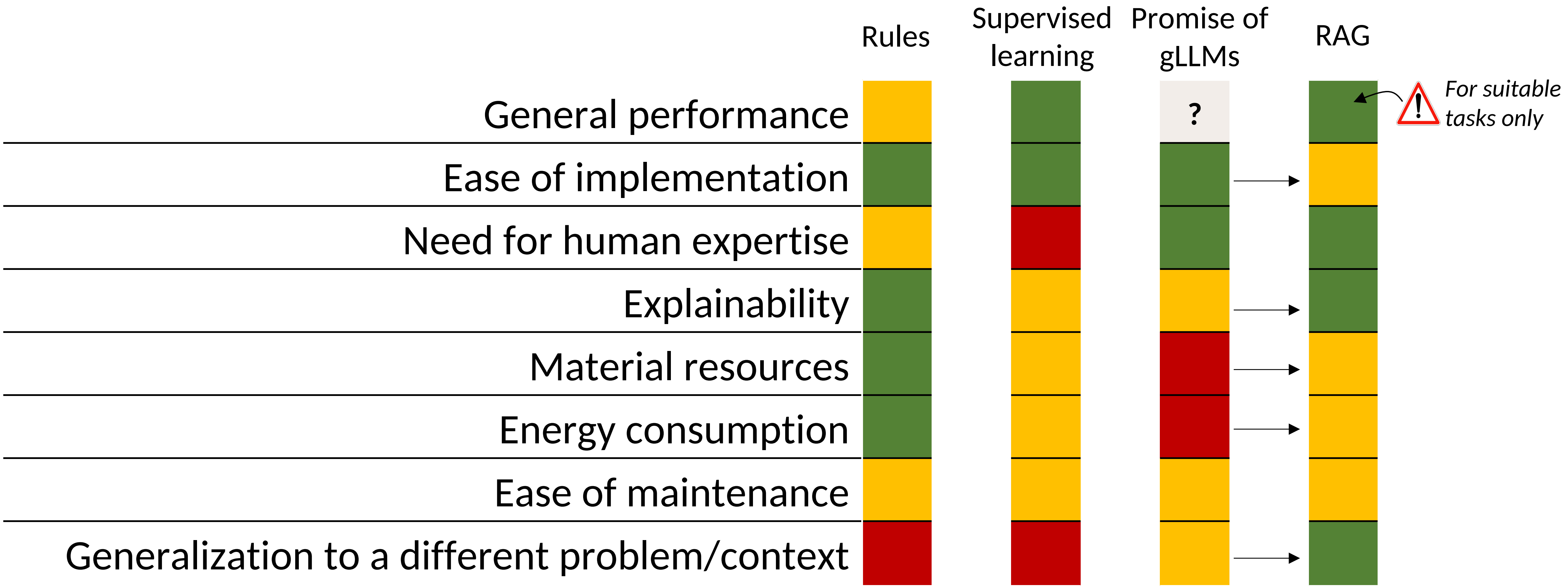
The new search engine

Retrieval-Augmented Generation



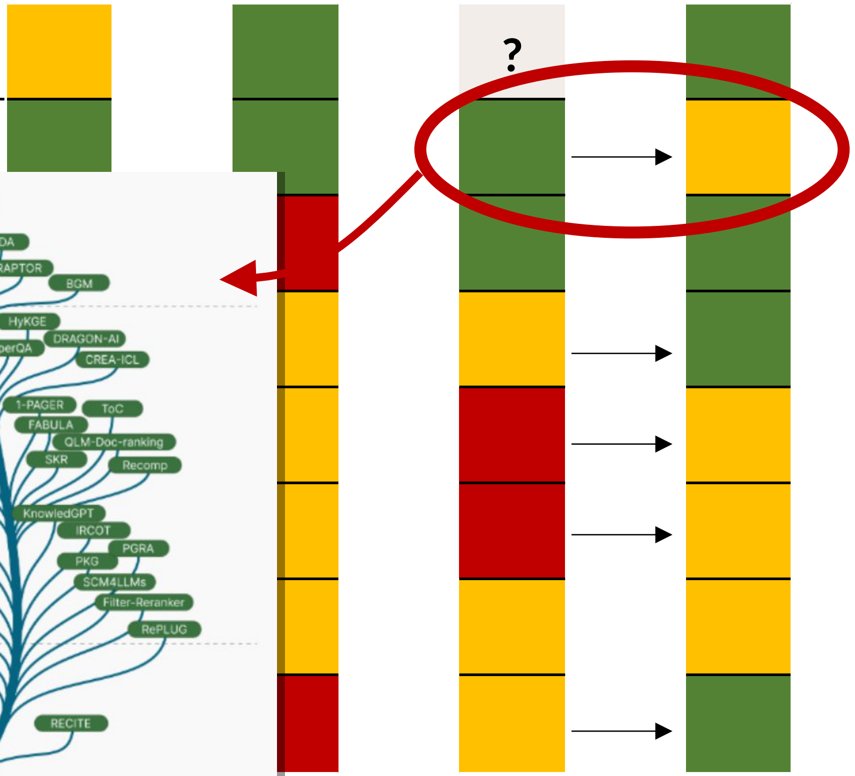
https://medium.com/@krtarunsingh/introduction-to-retrieval-augmented-generation-rag-and-its-transformative-role-in-ai-c07e35da7f01

RAG: Pros & Cons



RAG: Pros & Cons

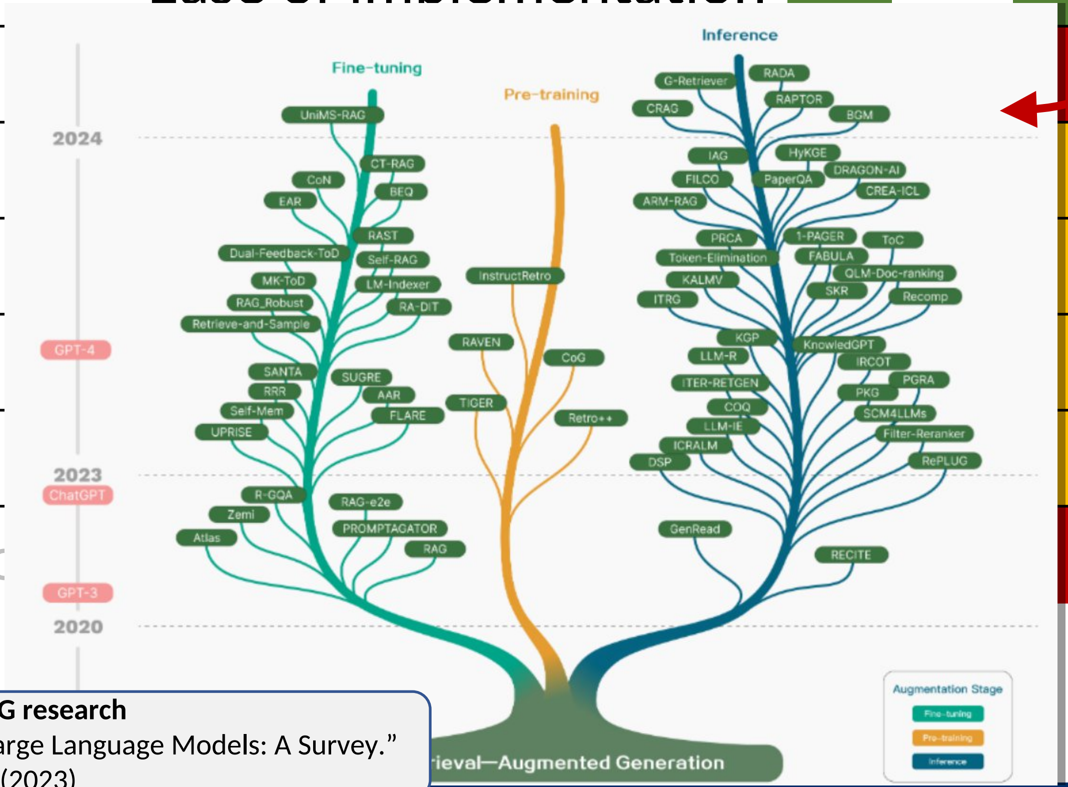
	Rules	Supervised learning	Promise of gLLMs	RAG



General performance

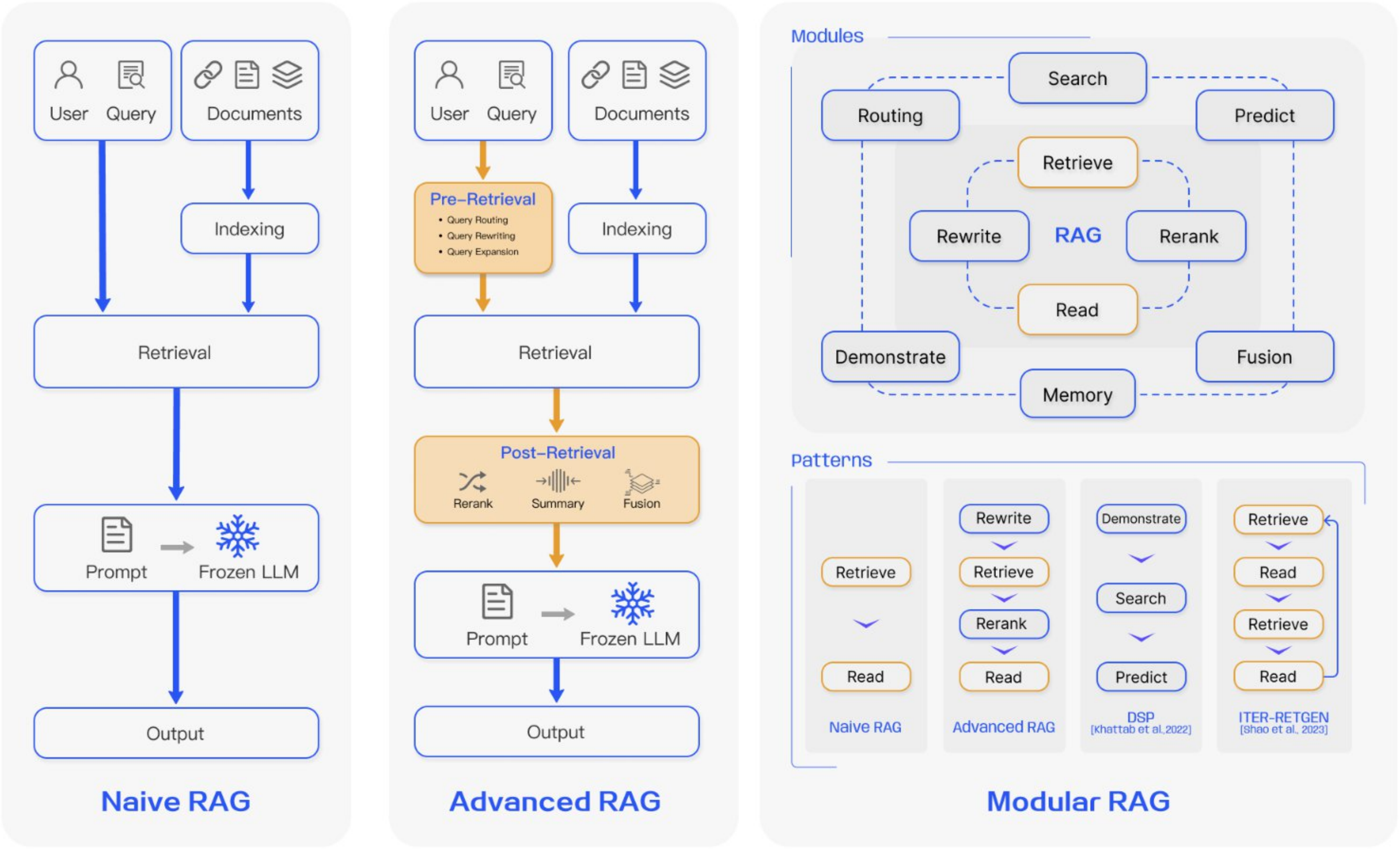
Ease of implementation

Generalization to

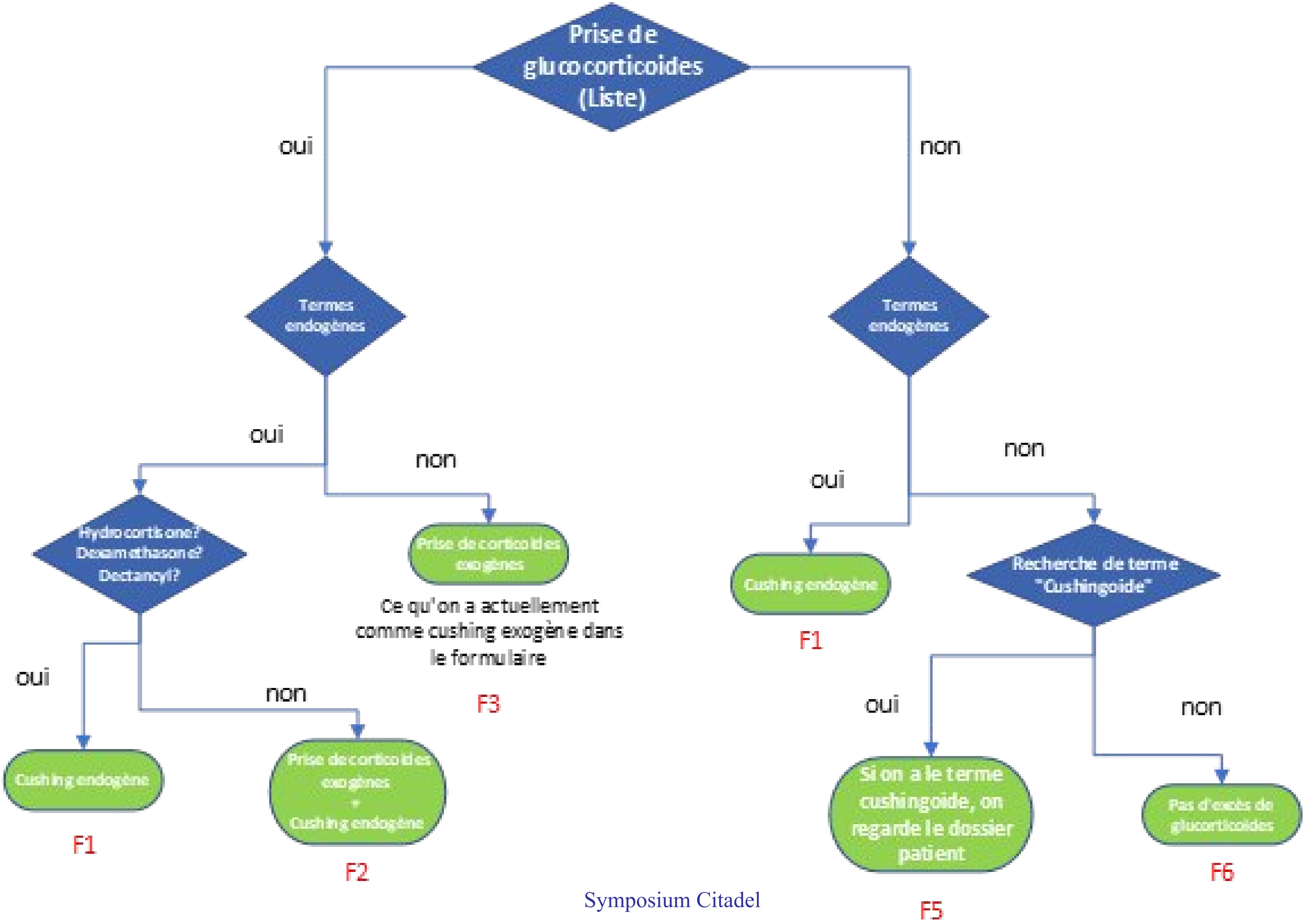


The technology tree of RAG research
Gao et al. "Retrieval-Augmented Generation for Large Language Models: A Survey." *ArXiv* abs/2312.10997 (2023)

RAG



Gao et al. "Retrieval-Augmented Generation for Large Language Models: A Survey."
ArXiv abs/2312.10997 (2023)


Prise de glucocorticoides (Liste)
oui
non
Termes endogènes
Termes endogènes
oui
non
oui
non
Hydrocortisone? Dexamethasone? Dectancyl?
Prise de corticoides exogènes
Ce qu'on a actuellement comme cushing exogène dans le formulaire
F3
Cushing endogène
F1
Recherche de terme "Cushingoide"
oui
non
oui
non
Cushing endogène
F1
Prise de corticoides exogènes + Cushing endogène
F2
Si on a le terme cushingoide, on regarde le dossier patient
F5
Pas d'excès de glucorticoides
F6