ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Outils de traitement des comptes-rendus cliniques dans les entrepôts de données de santé

Inria-Almanach

17 février 2023

Perceval Wajsbürt

Romain Bey

Présentation

EDS-NLP: partager des algorithmes de NLP clinique

EDS-PDF: intégrer et mettre en qualité les documents cliniques

EDS-PSEUDO: dé-identifier les textes cliniques

Romain Bey
Directeur du service Science des Données,
Pôle Innovation et Données,
Direction des Services Numériques,
AP-HP

Perceval Wajsbürt
Data scientist,
Equipe Science des Données,
Pôle Innovation et Données,
Direction des Services Numériques,
AP-HP

CHU d'Ile de France

Directeur général: M. Nicolas Revel
Présidente du conseil de surveillance: Mme. Anne Hidalgo

8,3M de prises en charge

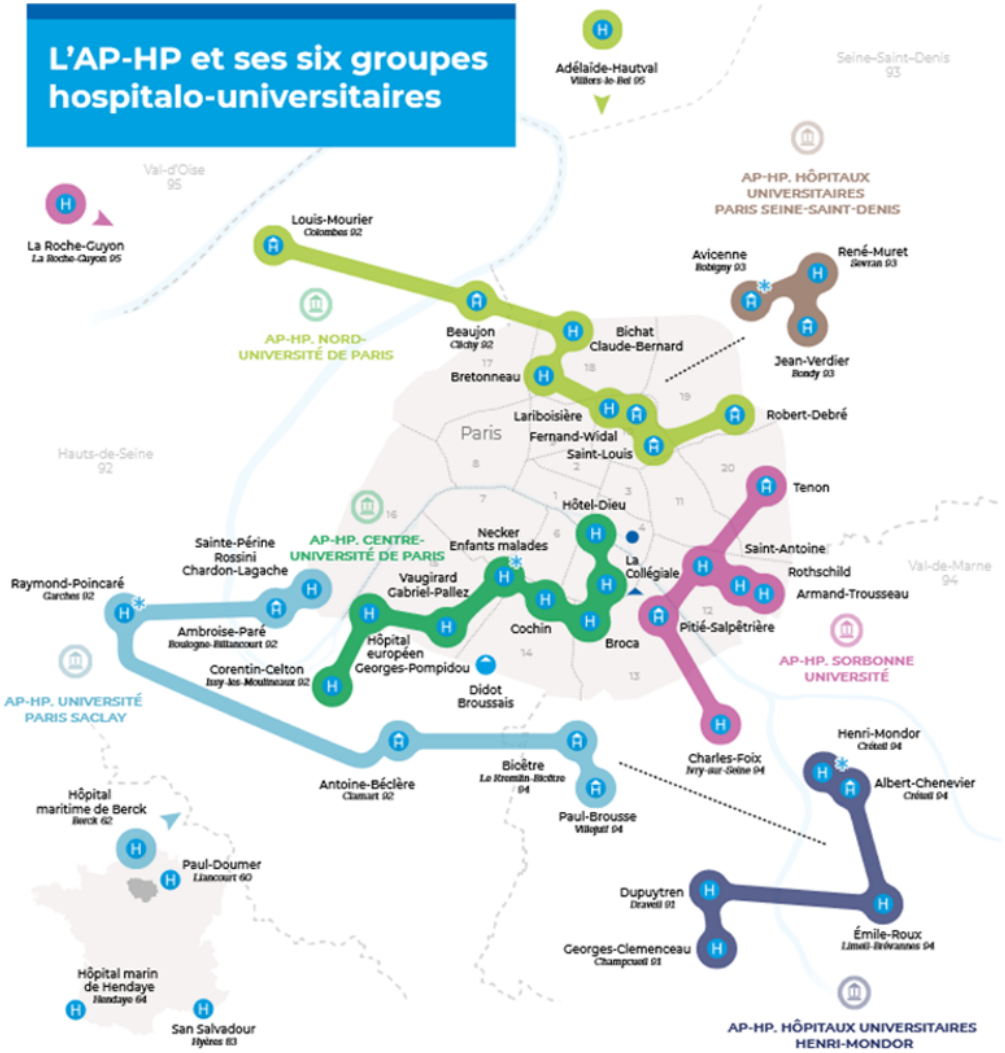
- **1,4M** de séjours en médecine, chirurgie, obstétrique
- **5,2M** de consultations externes
- **1,5M** de passages aux urgences

39 hôpitaux
20 098 lits
54 blocs chirurgicaux (**315** salles d'opération)

100 000 professionnels

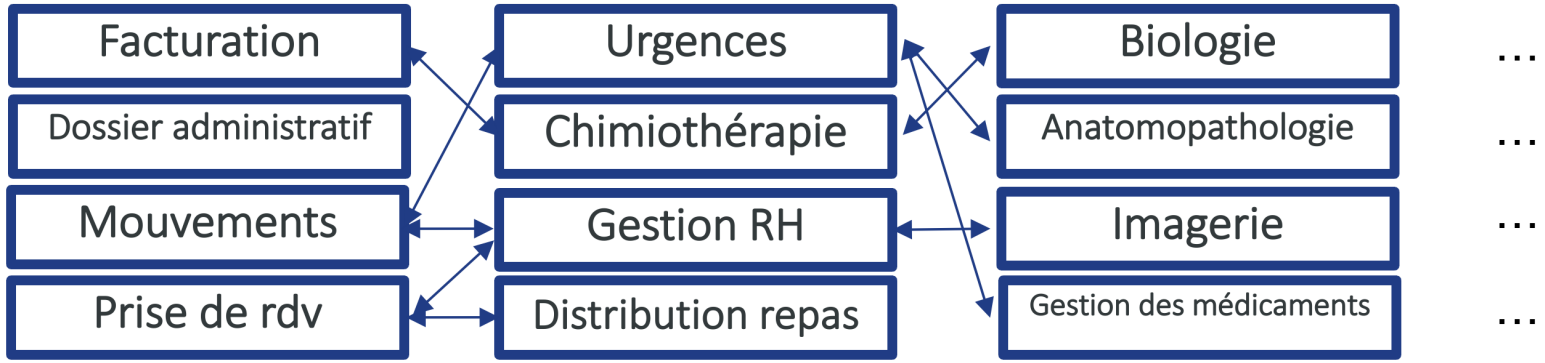
- **13 220** médecins
- **2000** bénévoles auprès des patients et des familles

7,5Md€ de budget



La direction des systèmes numériques (DSN)

- 800 professionnels
- 800 applications répertoriées



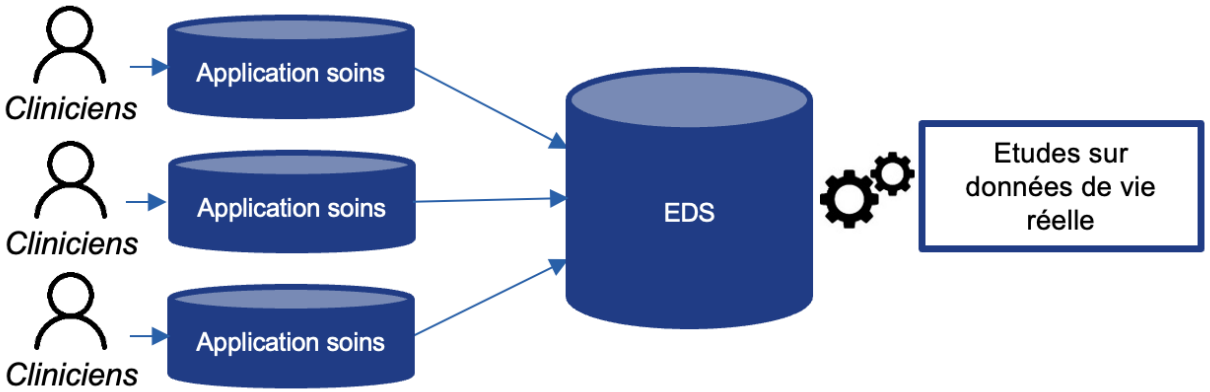
- Depuis 2012, déploiement d'un Dossier Patient Informatisé commun aux 39 hôpitaux
- Depuis 2019, unique serveur identités
- Adaptation continuelle du système d'information aux besoins métiers
- Standardisation en cours des données et API

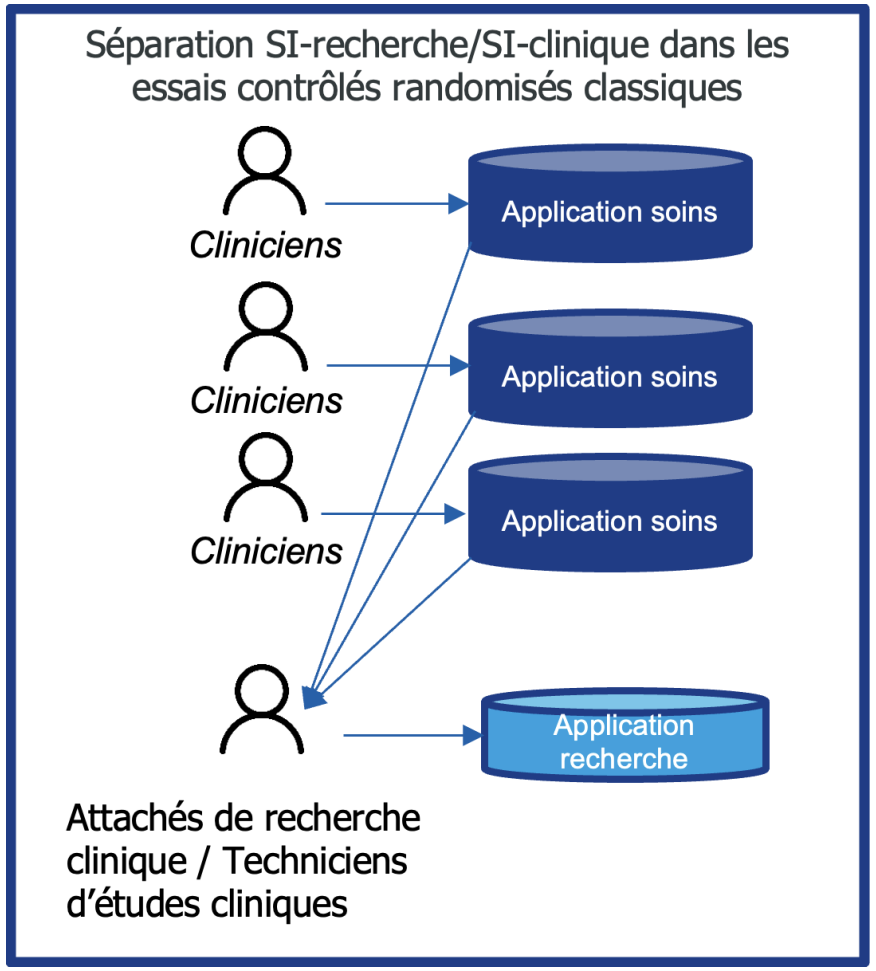
Une base de données

L'Entrepôt des Données de Santé (EDS) est une **base de données** contenant les données de vie réelles collectées par les logiciels de l'AP-HP

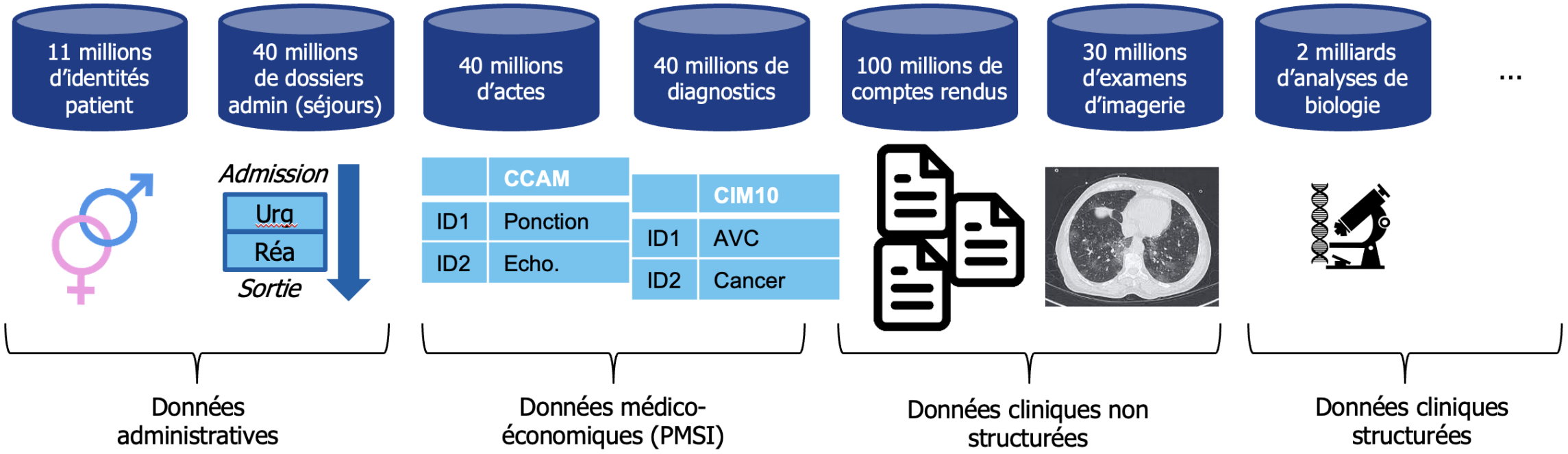
Finalités: recherche, innovation et pilotage

Le SI-recherche est directement connecté au SI-clinique





Volumétrie des données de l'EDS

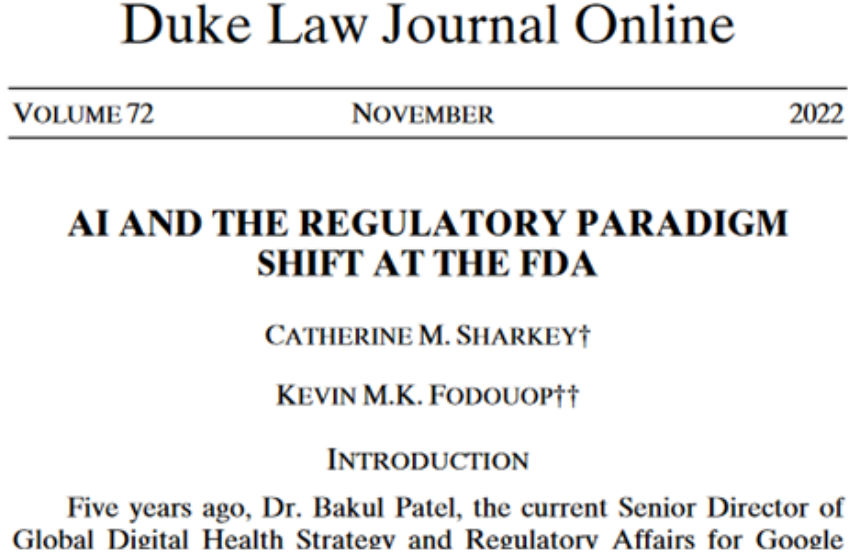


- **Nombreuses données non structurées**: données nécessitant un prétraitement algorithmique avant de fournir des variables utilisables dans des modèles statistiques classiques
- Les informations cliniques sont **souvent enregistrées sous forme textuelle** ce qui permet d'exprimer des nuances et qui est souvent plus rapide à collecter
- Les informations textuelles sont partagées avec le patients, entre les professionnels de santé et entre les applications métier principalement sous forme de **PDFs**

NLP clinique: Exemples

Quelques exemples d'utilisation du NLP pour:

- Recherches rétrospectives sur données
  - ex: impact de facteurs de risques
- Constitution de jeux de données pour l'entraînement de modèles d'IA en computer vision/signal processing
  - ex: algorithme de diagnostic à partir d'un scanner
- Optimisation d'essais cliniques
  - ex: étude de faisabilité, repérage de patients éligibles
- Suivi post-marketing des médicaments
  - ex: augmentation d'hospitalisations liées à certaines complications
- Surveillance sanitaire
  - ex: COVID-19, santé mentale
- Pilotage du système de soin
  - ex: Indicateurs de qualité (proportion de patients ayant bénéficiés d'un test particulier)



Duke Law Journal Online

VOLUME 72 NOVEMBER 2022

AI AND THE REGULATORY PARADIGM SHIFT AT THE FDA

CATHERINE M. SHARKEY†

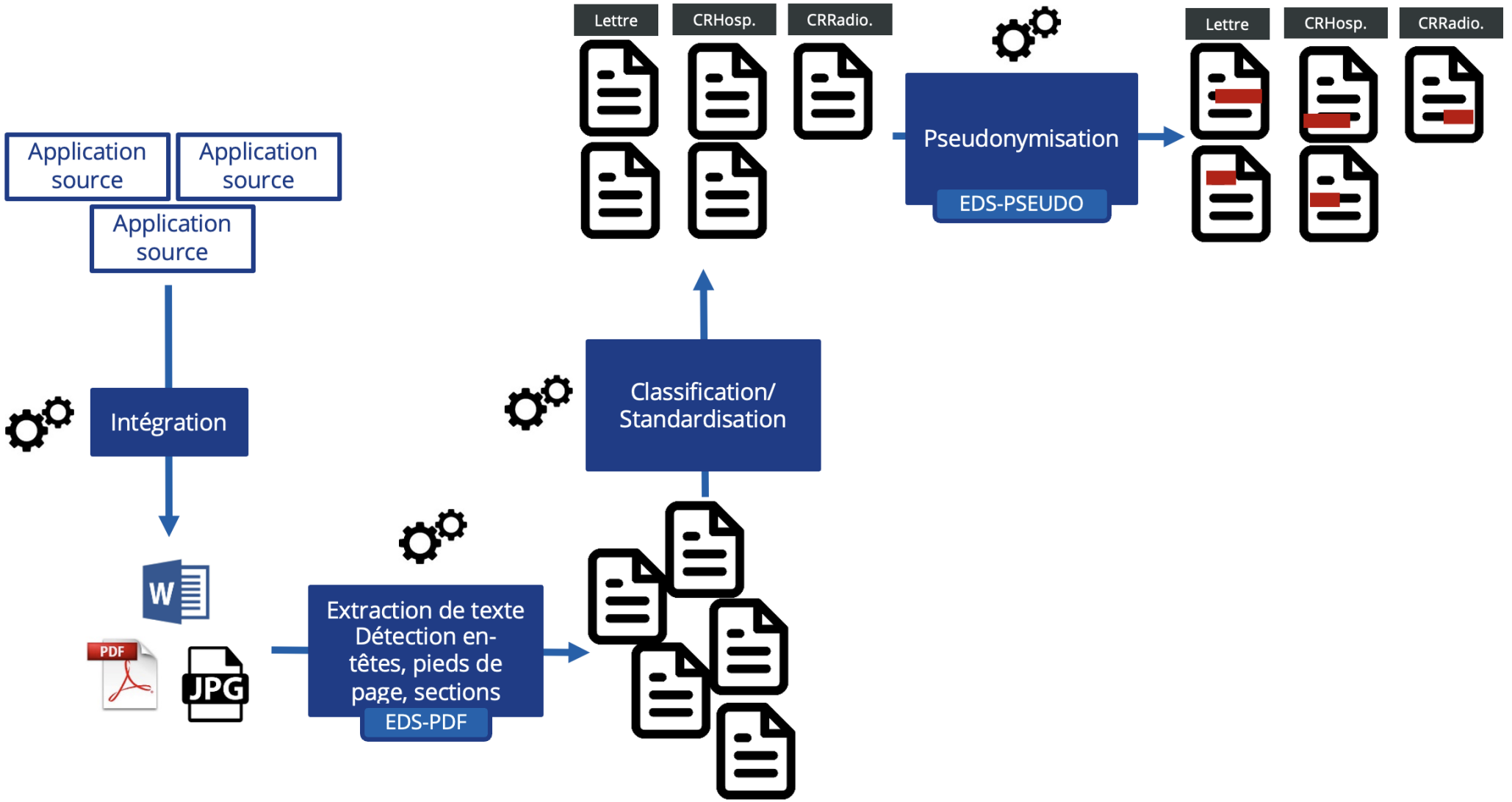
KEVIN M.K. FODOUOP††

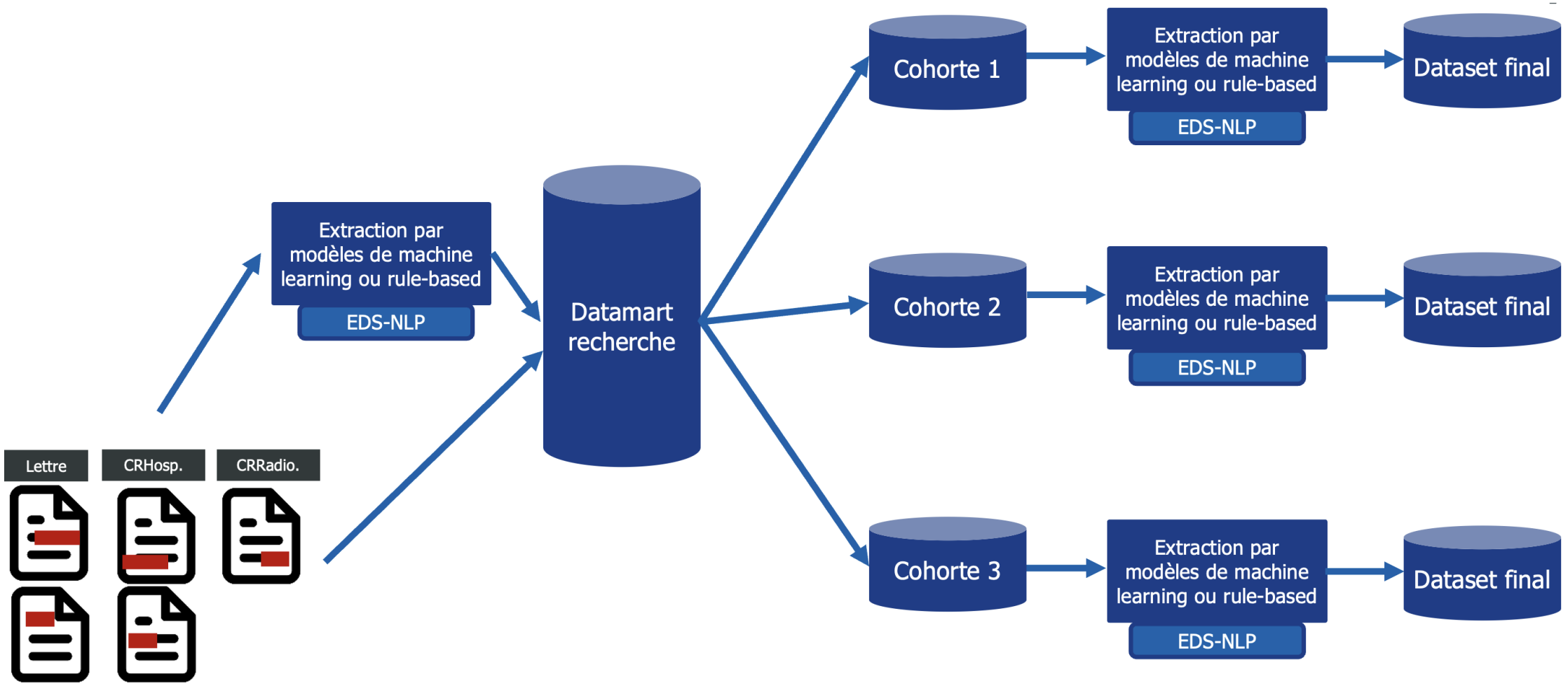
INTRODUCTION

Five years ago, Dr. Bakul Patel, the current Senior Director of Global Digital Health Strategy and Regulatory Affairs for Google

Le NLP en santé est prometteur, mais pour pleinement bénéficier au système de santé il est nécessaire de gagner la confiance des utilisateurs en créant des outils, méthodologies et processus vérifiés et pérennes

Préparation initiale des documents





Extraction de variables structurées en amont des projets de recherche… (équipe DSN)

… ou dans l'espace de travail de chaque projet de recherche (chaque équipe de recherche)

EDS-NLP

Contexte

Constats :

- Les outils NLP sont de plus en plus utilisé par les **data-scientists, biostatisticiens, cliniciens-chercheurs qui ne sont pas experts en NLP**
- Nombreux algorithmes d'**intérêt transversal** : diabète, score pronostic, prescription, tabagisme ...
- **Approches rule-based** souvent performantes mais consommatrices en temps-clinicien
- **Évolution régulière** des algorithmes pour tenir en compte des cas spécifiques et évolutions du SI
- La **reproductibilité** des études est une composante importante de leur qualité
- Algorithmes publiés sont **principalement anglophones**

Objectifs de EDS-NLP:

Mettre à disposition une bibliothèque scientifique de NLP clinique francophone qui soit robuste, versionnée, et co-développée par une communauté regroupant divers acteurs de la recherche, de l'innovation et du pilotage

Difficultés et besoins

- Documents difficiles à exploiter en raison de la diversité des formulations, lexiques, etc:

> "Le patient n'est pas diabétique"
> "Le patient est fumeur" / "... 15 paquets-années ..."
> "Patient fait 1m50" / "Taille: 150cm"

- Pollution récurrentes (ex: consentement / information patient)
- Prise en compte des sauts de lignes dans le découpage en phrase
- Extraction de terminologies, dates, mesures...
- Extraction de médicaments prescrits (surtout en dehors de l'APHP)
- Extraction de comorbidités/facteurs de risque mentionnées dans les CRH (diabète, insuffisance rénale, VIH, tabagisme, violences subies, etc.)

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Proposition

La bibliothèque EDS-NLP:

- est basée sur spaCy : cadre simple

```
import spacy

nlp = spacy.blank('eds')
nlp.add_pipe('eds.normalizer')
nlp.add_pipe('eds.covid')
nlp.add_pipe('eds.negation')
```

- composée majoritairement de briques à base de règles: *facile* à partager, améliorer, résultats reproductibles
- est versionnée, documentée, testée en ligne, license BSD 3 Clauses :

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Fonctionnalités

Une trentaine de briques prêtes à l'emploi et plus à venir

- pretraitements: découpage en phrase, section, nettoyage du texte
- extracteurs génériques: regex, lexiques
- extracteurs spécifiques: médicaments, cim10, scores TNM, adicap, charlson, ...
- extracteurs de dates, mesures
- qualificateurs: négation, parenté, hypothèse
- architectures de ML: NER imbriqué

Divers connecteurs entrée/sortie

- BRAT
- Pandas / Spark (OMOP)
- parallélisation simplifiée

Et tous les travaux de la communauté autour de spaCy !

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Extension & collaboration

Configuration des briques

```
nlp.add_pipe('eds.regex', config={
  "regex": {
    "smoker": [r"\bfume", "clope"]
  }})
```

Nouveaux composants

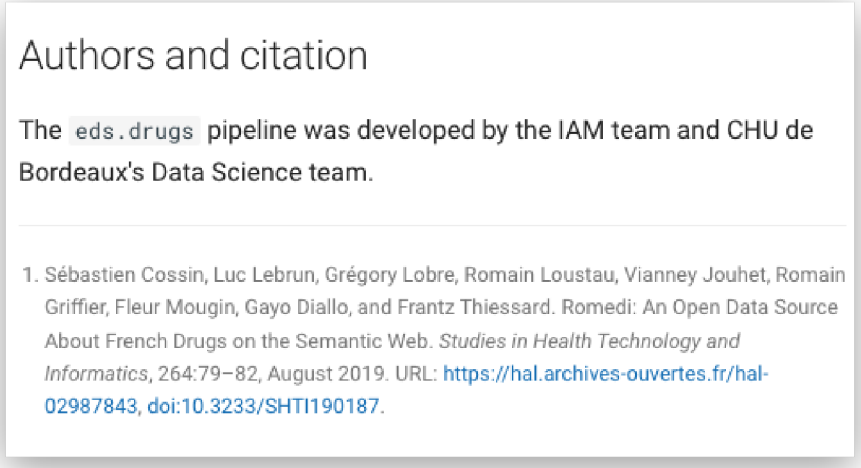
```
@register("smoker")
class Smoker:
  def __call__(self, doc): ...
```

Amélioration du code

Fichiers `patterns.py` / code des composants

Collaboration par GitHub

- issues, suggestions de features
- revues de code systématiques
- batterie de tests avec CI
- reconnaissance de la paternité



Authors and citation

The `eds.drugs` pipeline was developed by the IAM team and CHU de Bordeaux's Data Science team.

1. Sébastien Cossin, Luc Lebrun, Grégory Lobre, Romain Loustau, Vianney Jouhet, Romain Griffier, Fleur Mougin, Gayo Diallo, and Frantz Thiessard. Romedi: An Open Data Source About French Drugs on the Semantic Web. *Studies in Health Technology and Informatics*, 264:79–82, August 2019. URL: https://hal.archives-ouvertes.fr/hal-02987843, doi:10.3233/SHTI190187.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Pourquoi spaCy ?

- **Interopérabilité**: format unique, mais extensible de document
- **Rapide**: implémentation partielle en Cython
- **Extensibilité**: principe des entry-points
- **Communauté**: visualiseurs, autres composants, documentation...

```
$ pip install edsnlp
# J'installe mon projet hors edsnlp
$ pip install ma_pipeline
```

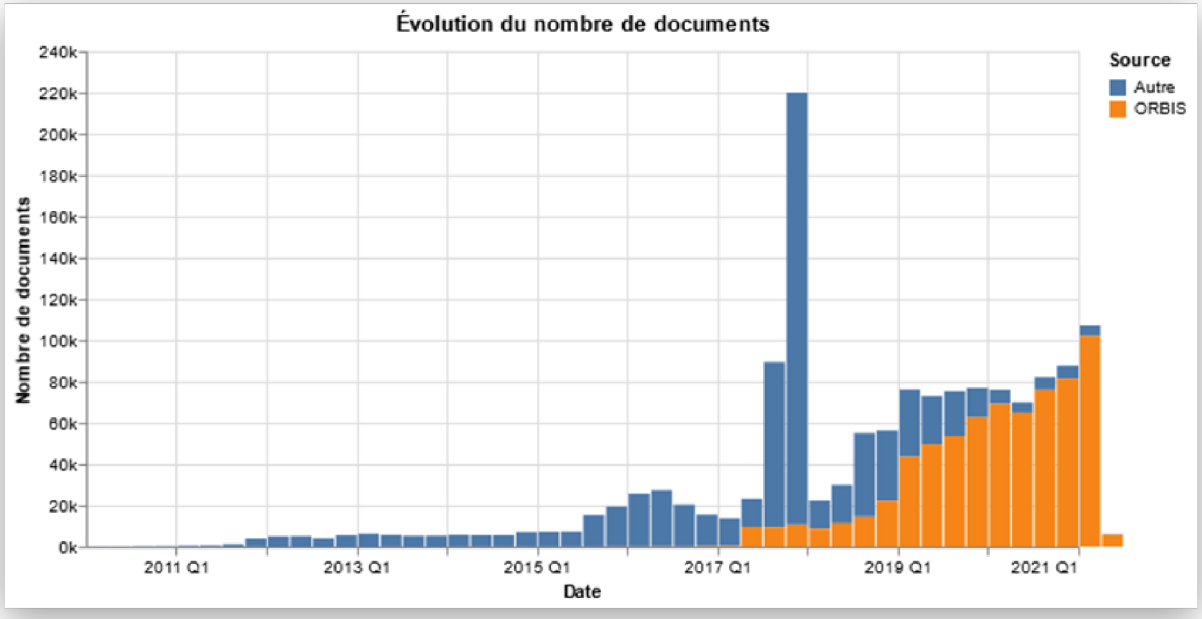
```
import spaCy

nlp = spacy.blank('eds')
nlp.add_pipe('eds.normalizer')
# directement disponible ↓
nlp.add_pipe('ma_pipeline')
```

EDS-PDF

Contexte

- Documents principalement issus des applications sources en **PDFs de formats variés**
- À l'EDS de l'APHP, l'extraction des textes depuis les PDF et la pseudonymisation ont lieu **avant la mise à disposition aux projets de recherche** (contrainte d'architecture)
- Les **métadonnées du PDF** (type de document, date), ne sont pas toujours fiables, contrairement aux données contenues dans le PDF



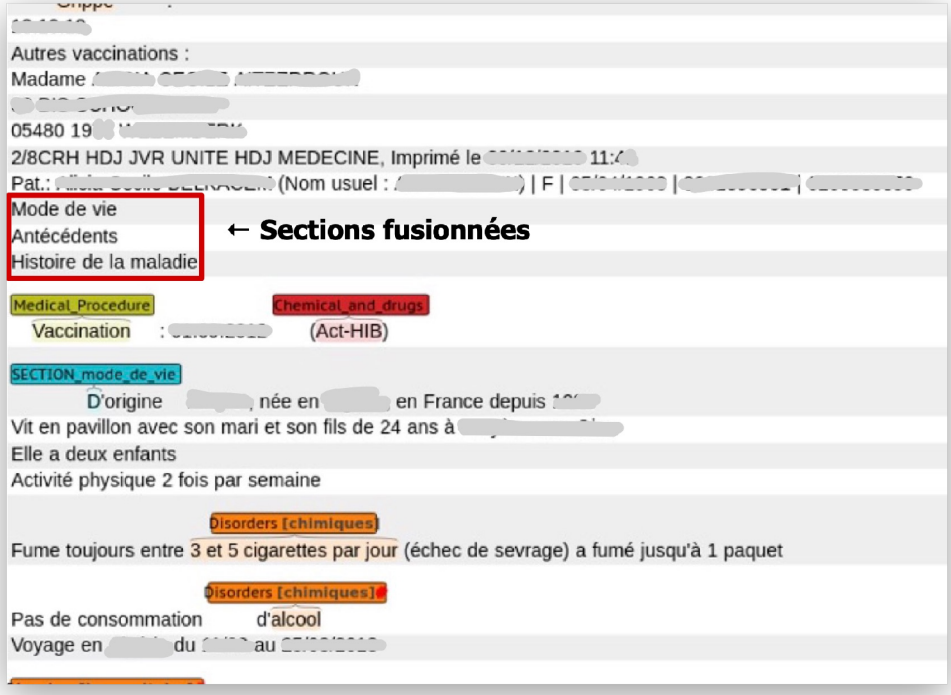
Constat

❌ Extraction "naïve" des PDFs conduit à:

- fusion du corps de texte et des entêtes
- titres de section souvent groupés
- bruitage & perte de la structure originale
- lecture difficile

✅ Une extraction avancée peut:

- pré-decouper l'entête pour enrichir les métadonnées
- filtrer beaucoup d'entités identifiantes
- extraire les sections & styles (gras/italique)



Objectifs

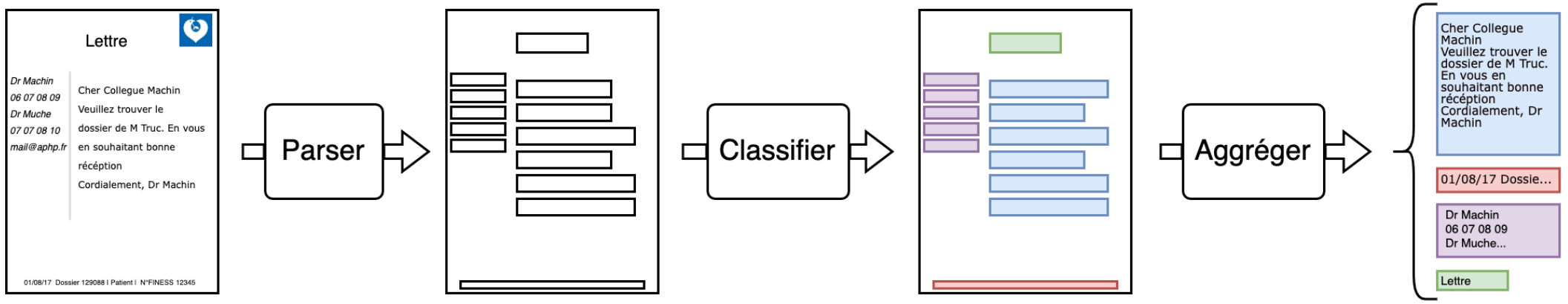
- Extraire les textes des PDF en **s'adaptant automatiquement** aux divers formats
- **Mise en production** pour l'intégration quotidienne des documents (200k par jour)
- Extraire les **métadonnées** du document (type, date & service d'édition, ID d'examens)
- **Open-sourcer** les développement réalisés pour ouvrir le projet à des collaborations
- Permettre de **ré-entrainer / améliorer** le modèle et **reproduire** les résultats

Modélisation

> Grande majorité des PDFs est "parsable", i.e. non scans

Stratégie

1. avec parseur PDF (pdfminer / mupdf / poppler): extraire les lignes
2. avec modèle de classification, les étiquetter
3. les aggréger par type pour obtenir les textes finaux



Architecture

Embedding des lignes

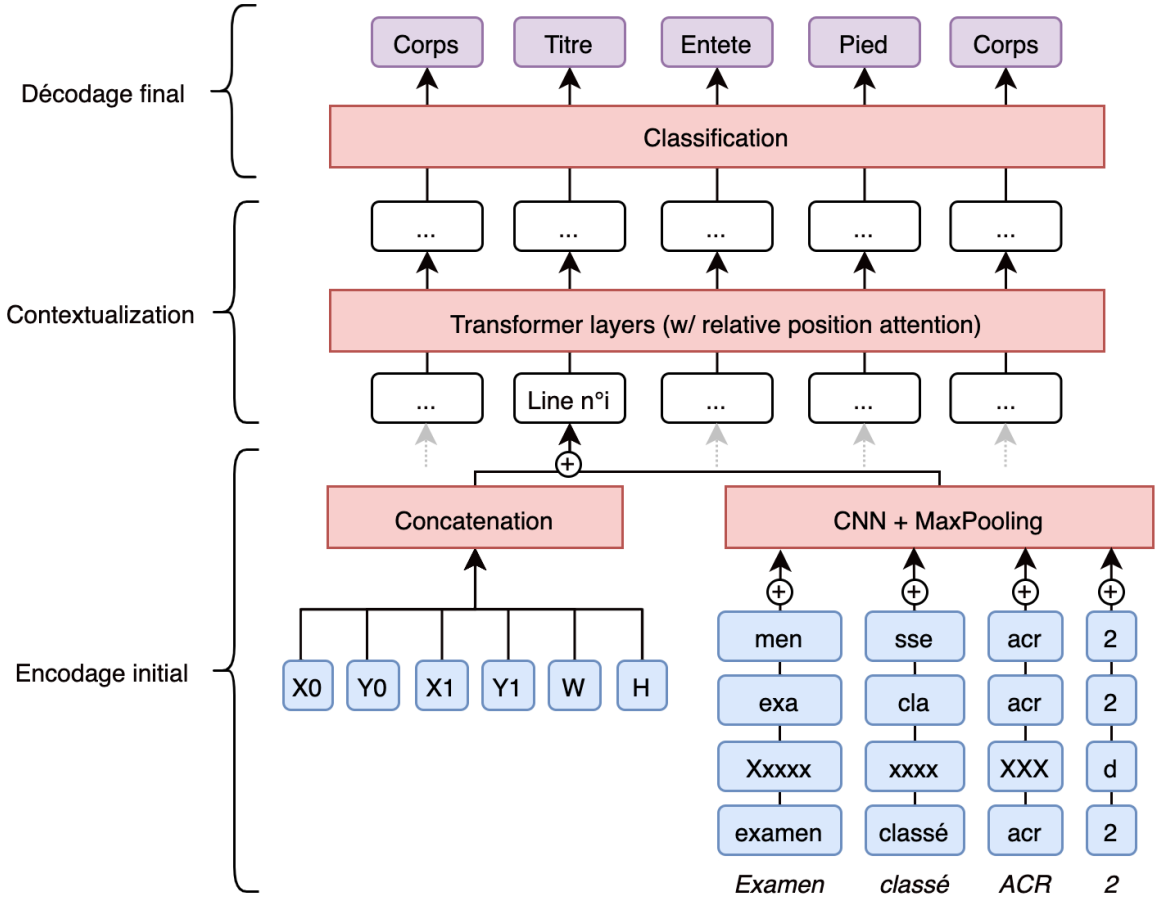
- Features géométriques: taille, position
- Features de texte: préfixe, suffixe, casse, ...

Contextualization

- Couche de self-attention par page
- Attention avec positions relatives [1] (distances x et y entre 2 lignes)

Classification finale

- Couche linéaire simple



[1] Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-Attention with Relative Position Representations.

Harmonisation : problème

Même avec les techniques précédentes, le modèle peut commettre des erreurs "gênantes":

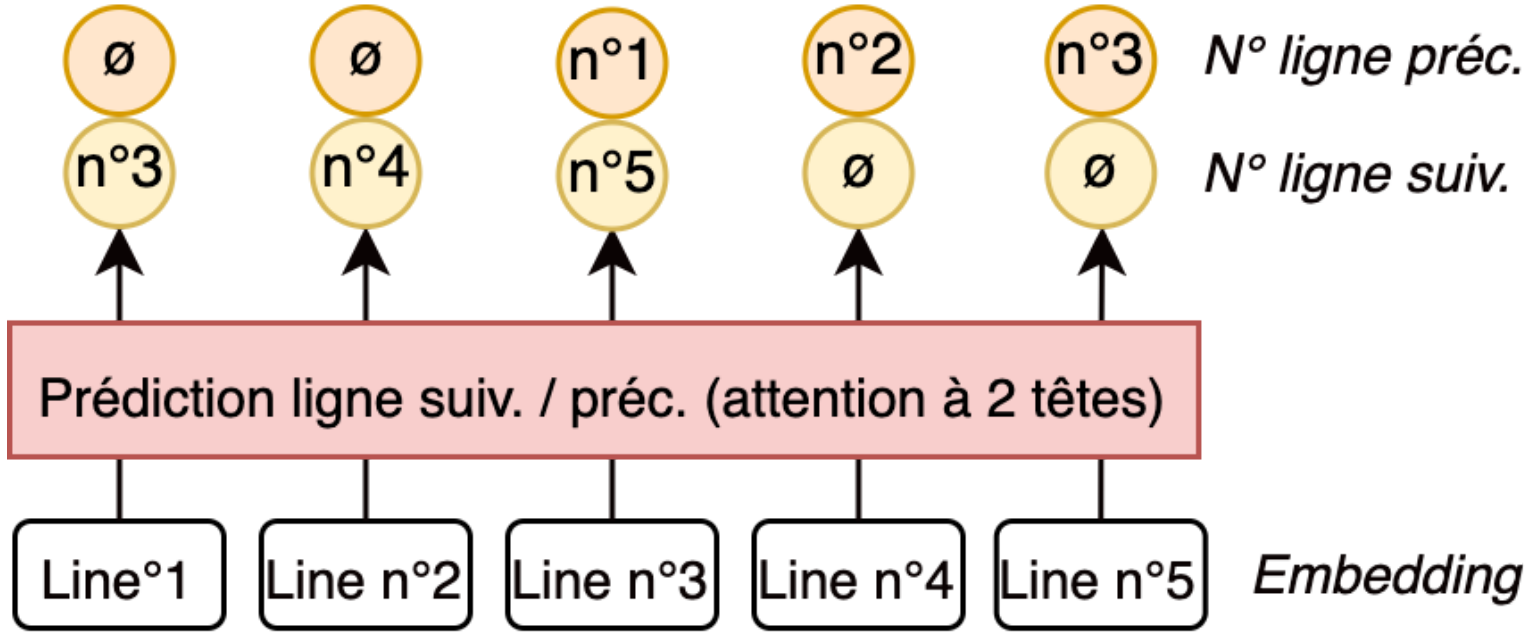
Harmonisation: solution

> Si deux lignes se suivent au sein d'un même bloc, elles doivent avoir le même type

1. Pour chaque ligne, on prédit son type (cf slide prec.)

2. Pour chaque ligne, on prédit également *quelle ligne la suit*, et *quelle ligne la précède*

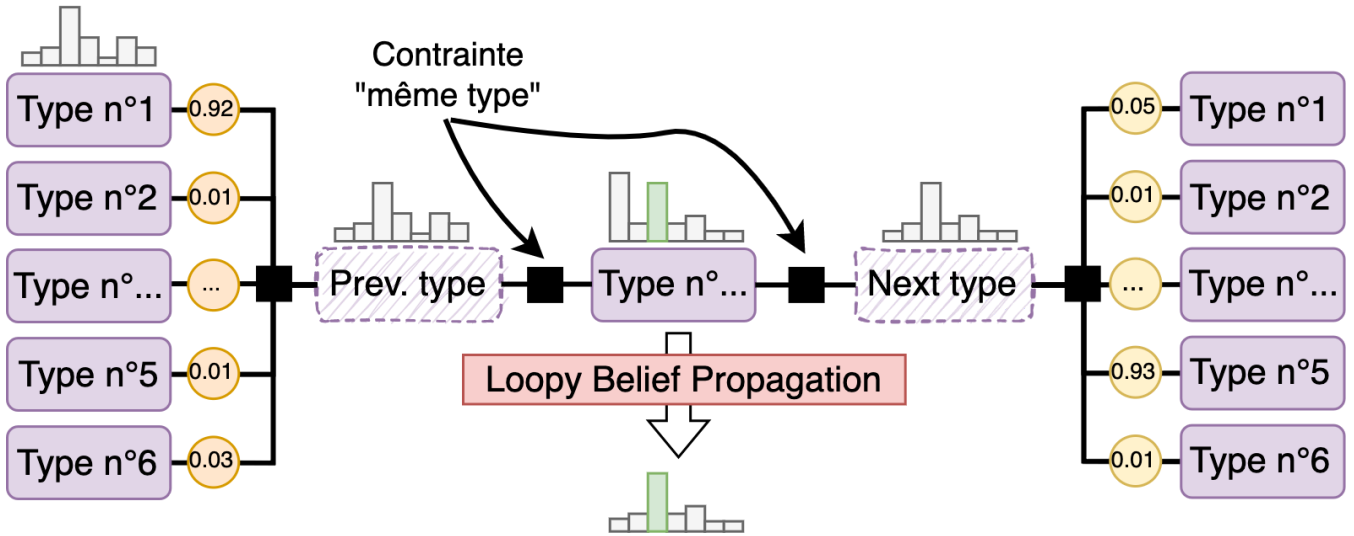
3. On obtient ainsi une estimation du type suivant et du type précédent de chaque ligne



Harmonisation: solution

> Si deux lignes se suivent au sein d'un même bloc, elles doivent avoir le même type

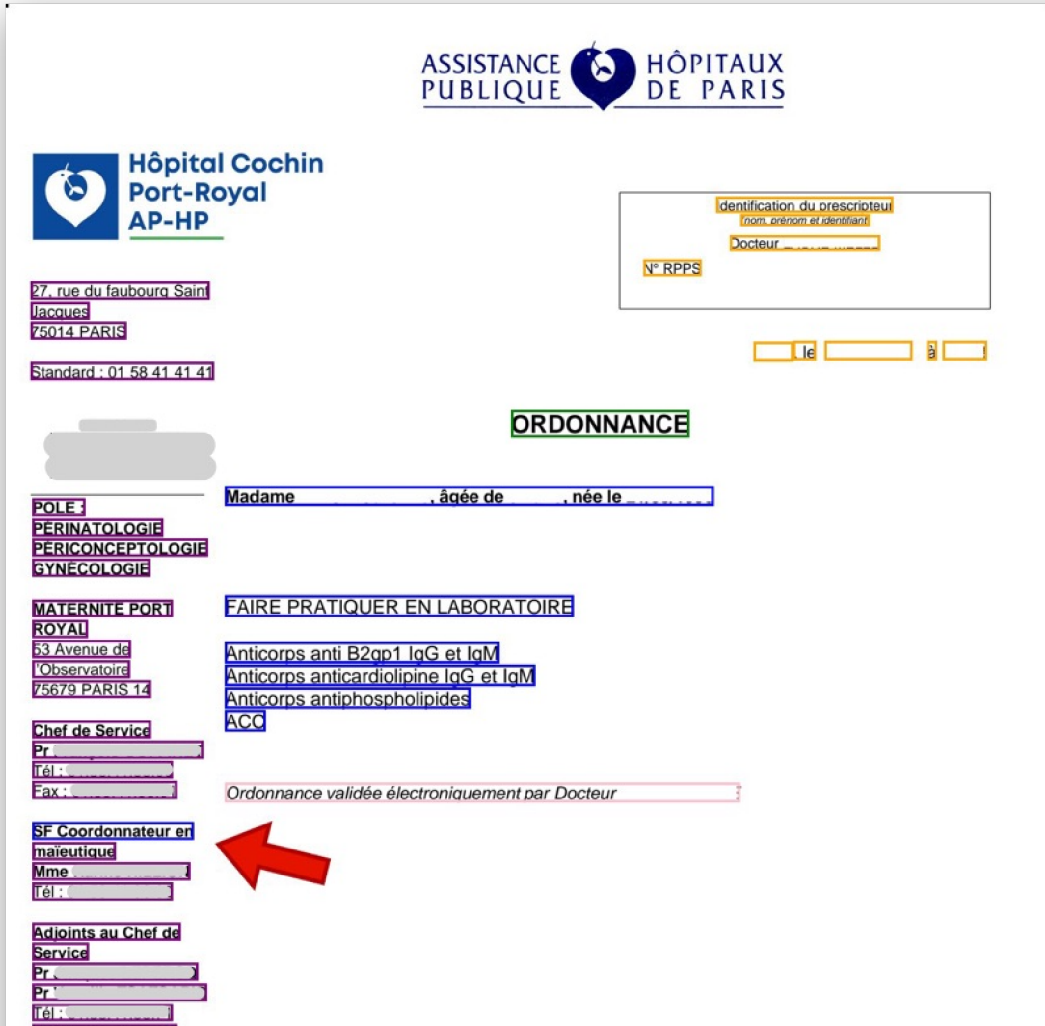
4. On contraint une ligne à être typée comme la suivante et précédente
5. Pour appliquer cette contrainte en tenant compte des incertitudes du modèle, dans chaque page, on construit un CRF (Conditional Random Field) généralisé
6. Graphe complet: pas de solution exacte ⇨ approx par Loopy Belief Propagation [2]



[2] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.

Harmonisation: solution

Avant harmonisation



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Hôpital Cochin
Port-Royal
AP-HP

dentification du prescripteur
(nom, prénom et identifiant)
Docteur
N° RPPS

27, rue du faubourg Saint Jacques
75014 PARIS

Standard : 01 58 41 41 41

le à

ORDONNANCE

Madame , âgée de , née le

POLE :
PERINATOLOGIE
PERICONCEPTOLOGIE
GYNECOLOGIE

MATERNITE PORT ROYAL
53 Avenue de l'Observatoire
75679 PARIS 14

Chef de Service
Pr
Tél :
Fax :

SF Coordonnateur en maïeutique
Mme
Tél :

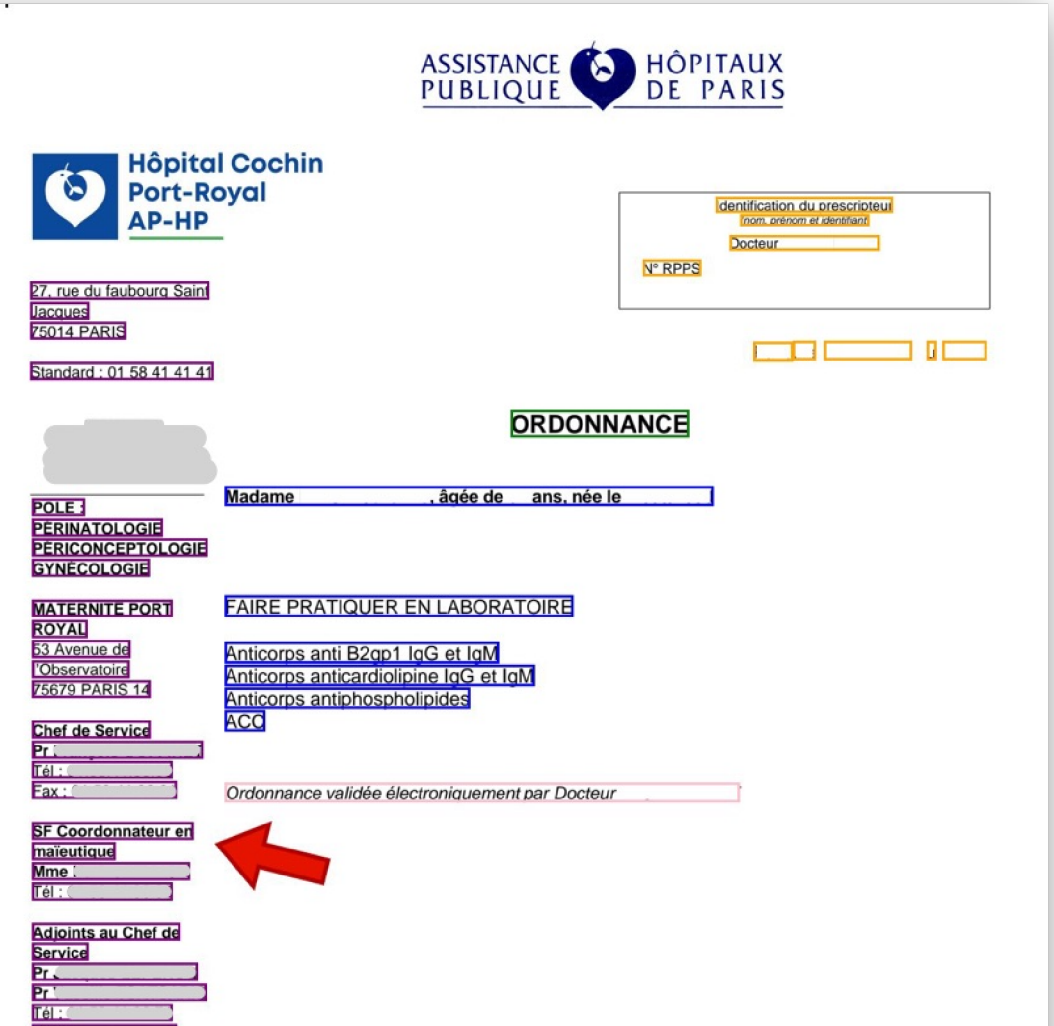
Adjoints au Chef de Service
Pr
Pr
Tél :

FAIRE PRATIQUER EN LABORATOIRE

Anticorps anti B2gp1 IgG et IgM
Anticorps anticardiolipine IgG et IgM
Anticorps antiphospholipides
ACO

Ordonnance validée électroniquement par Docteur

Après harmonisation



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Hôpital Cochin
Port-Royal
AP-HP

dentification du prescripteur
(nom, prénom et identifiant)
Docteur
N° RPPS

27, rue du faubourg Saint Jacques
75014 PARIS

Standard : 01 58 41 41 41

ORDONNANCE

Madame , âgée de ans, née le

POLE :
PERINATOLOGIE
PERICONCEPTOLOGIE
GYNECOLOGIE

MATERNITE PORT ROYAL
53 Avenue de l'Observatoire
75679 PARIS 14

Chef de Service
Pr
Tél :
Fax :

SF Coordonnateur en maïeutique
Mme
Tél :

Adjoints au Chef de Service
Pr
Pr
Tél :

FAIRE PRATIQUER EN LABORATOIRE

Anticorps anti B2gp1 IgG et IgM
Anticorps anticardiolipine IgG et IgM
Anticorps antiphospholipides
ACO

Ordonnance validée électroniquement par Docteur

Résultats

Performances du modèle complet

- Transformer
- Attention + positions relatives
- CRF

⇨

Type	Precision	Recall	F1-score
body	98.2 ± 0.3	97.6 ± 1.0	97.9 ± 0.4
footer	90.7 ± 0.8	87.8 ± 1.7	89.2 ± 0.9
header	91.9 ± 0.6	95.5 ± 0.7	93.6 ± 0.4
left_note	96.8 ± 2.4	97.8 ± 0.5	97.3 ± 1.2
page	97.2 ± 5.6	90.7 ± 1.5	93.7 ± 2.8
pollution	95.7 ± 1.0	92.3 ± 0.7	93.9 ± 0.6
signature	87.4 ± 3.8	79.5 ± 2.0	83.2 ± 2.3
title	94.5 ± 2.1	81.4 ± 0.0	87.4 ± 0.9
macro avg	94.1 ± 1.0	90.3 ± 0.4	92.0 ± 0.6
accuracy	96.5 ± 0.4	96.5 ± 0.4	96.5 ± 0.4

Impact sur la performance des différentes techniques évoquées ⇩

Modèle	Micro-accuracy	Macro-accuracy	Body f1-score	Body recall
Full model	96.5 ± 0.4	92.0 ± 0.5	97.9 ± 0.4	97.6 ± 1.0
- CRF	96.1 ± 0.7 (-0.4)	90.9 ± 2.3 (-1.1)	97.4 ± 0.7 (-0.4)	96.6 ± 1.4 (-1.0)
- Rel. position attn	94.8 ± 0.6 (-1.3)	88.9 ± 2.0 (-2.0)	96.6 ± 0.6 (-0.8)	96.2 ± 1.3 (-0.3)
- Transformer	92.0 ± 0.5 (-2.9)	85.2 ± 1.6 (-3.7)	95.2 ± 0.4 (-1.4)	95.3 ± 0.6 (-0.9)

EDS-Pseudo

Contexte

Constat

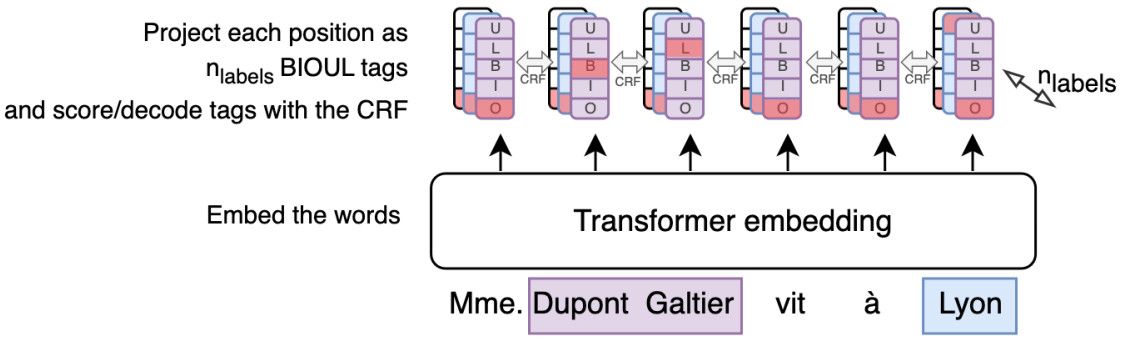
- Les textes de l'EDS contiennent de nombreuses **informations identifiantes**
- La CNIL demande aux EDS de **suivre l'état de l'art** pour retirer le plus de ces infos
- Il a été montré que le **machine learning** permet de meilleures performances de pseudonymisation

Enjeux

- **Développer** un algorithme de pseudonymisation des textes cliniques à l'état de l'art
- **Mise en production** pour l'intégration quotidienne des textes cliniques à l'EDS
- **Valider** ses performances ⇨ assurer une transparence vis-à-vis des **risques résiduels**
- **Open-sourcer** les développement réalisés pour ouvrir le projet à des collaborations

EDS-Pseudo

1. **Deep learning**: *Transformer + CRF*

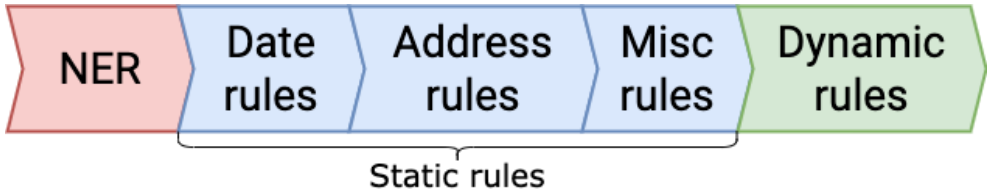


2. **Règles**

- statiques (regex ou +)
- dynamiques: recherche à partir des données structurées

3. **Fusion**: ML ou Règles

Implémentation



Avec EDS-NLP 🎉

```python
import spacy

nlp = spacy.blank('eds')
nlp.add_pipe('nested_ner') # à entrainer
nlp.add_pipe('pseudonymisation-dates')
nlp.add_pipe('pseudonymisation-adresses')
nlp.add_pipe('pseudonymisation-rules')
nlp.add_pipe('structured-data-matcher')
```

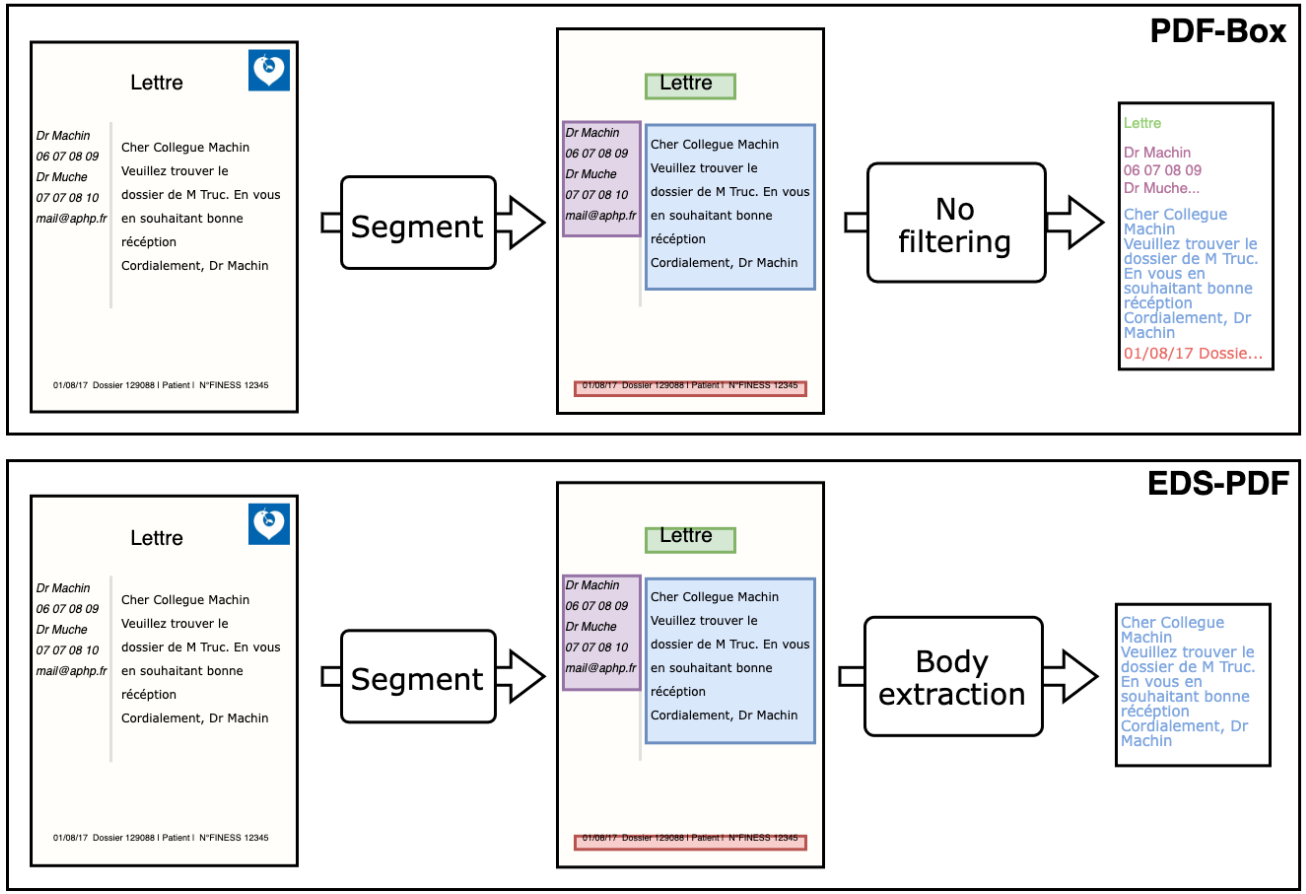
Extraction des textes

Répartition document:

- Train: documents post-2017
- Test: sampling aléatoire

Répartition extraction PDF:

- 90% edspdf
- 10% pdfbox



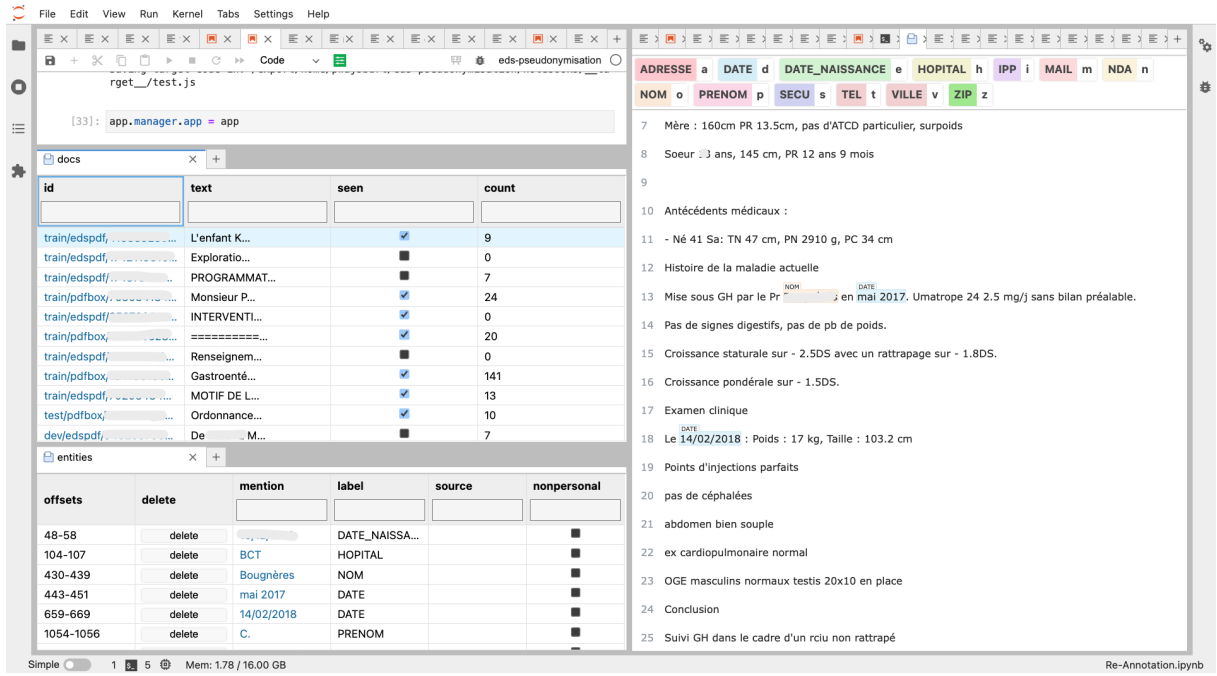
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Annotation

Campagne en deux étapes:

- phase une en dec 21 - jan 22
- vérification en dec 22 - jan 23

	ENTS	DOCS
train/edspdf	27135	3025
train/pdfbox	16071	348
dev/edspdf	1615	200
dev/pdfbox	967	22
test/edspdf	3491	348
test/pdfbox	16793	348



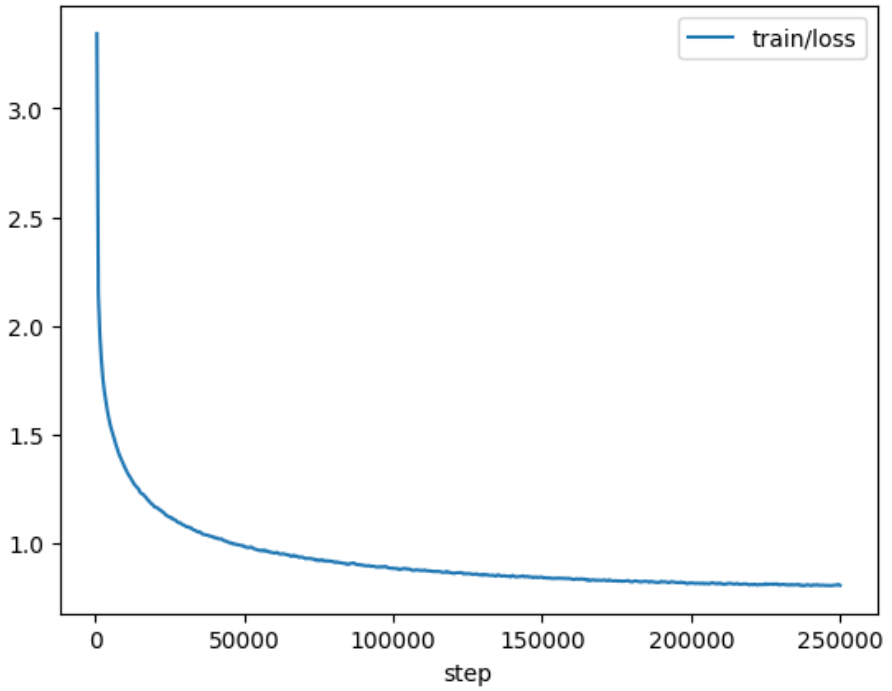
2e phase d'annotation (metanno)

C'est dur d'obtenir des annotations de qualité !

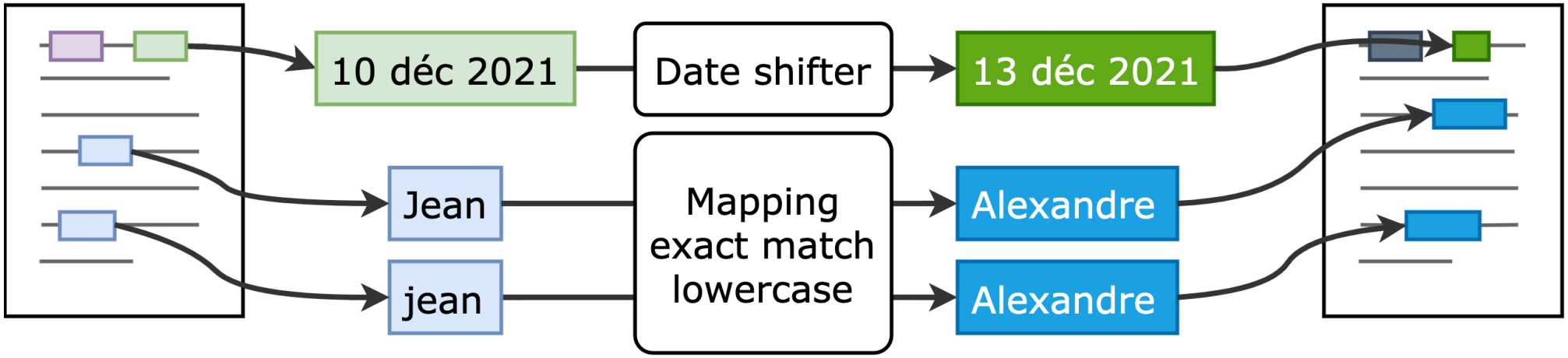
Il faut voir chaque doc 2-3 fois, vérifier, re-vérifier...

Finetuning de BERT

- Voir Dura, B., Jean, C., Tannier, X., Calliger, A., Bey, R., Neuraz, A., & Flicoteaux, R. (2022). Learning structures of the French clinical language:development and validation of word embedding models using 21 million clinical reports from electronic health records
- Ici, finetuning "rapide": 250000 steps soit une epoch sur 28M de CR bruts (non pseudonymisés) avec *Whole-Word Masking*
- 8 GPUs V100 pendant 36h sur 140Go de textes



Remplacement dans les textes



- Remplacement par exact match en minuscules
- ⚠️ Il faut décaller les indices si détection d'entités en amont

Résultats

Métriques

- Niveau du token (pas exact match)
- Redact: caviardage en % de mots
- Full: % de docs avec redact à 100%

Observations

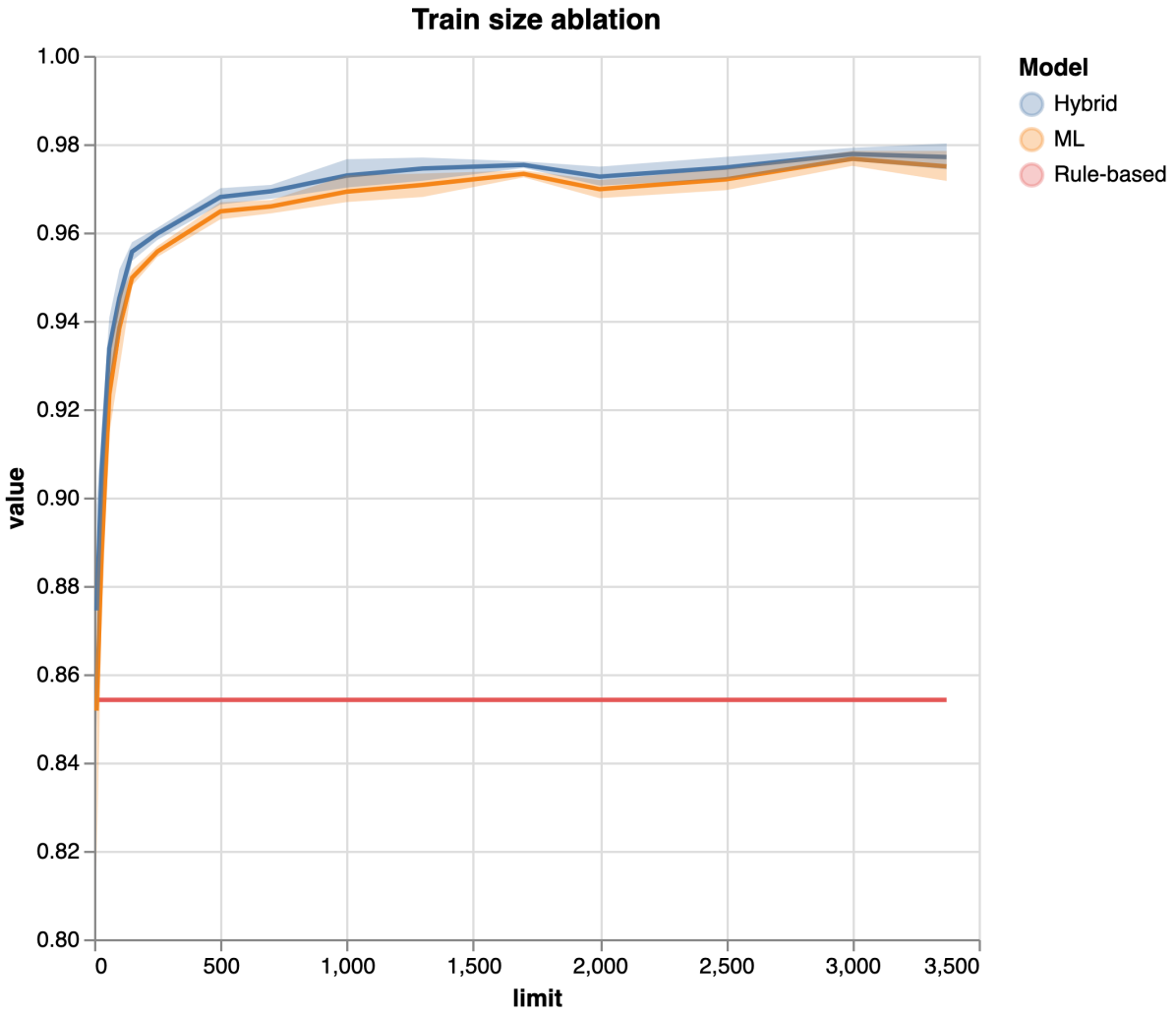
- 99% en métrique F1 🎉
- et 99.4 en redact !
- 86.2% de docs entièrement caviardés
- VISIT ID difficile car formats divers et source dans les données structurées

Label	P	R	F1	Redact	Full
ADDRESS	99.0	98.4	98.7	98.5	98.4
BIRTHDATE	98.2	98.2	98.2	99.8	99.7
CITY	98.0	98.8	98.4	98.8	98.2
DATE	99.7	99.3	99.5	99.6	95.4
EMAIL	98.9	99.9	99.4	99.9	99.9
FIRSTNAME	98.8	98.4	98.6	99.4	97.4
LASTNAME	98.6	98.6	98.6	99.6	97.2
NSS	88.0	98.9	93.1	100.	100.
PATIENT ID	99.0	94.0	96.4	98.2	99.1
PHONE	99.6	99.7	99.7	99.7	99.0
VISIT ID	91.5	89.4	90.4	90.4	98.3
ZIP	99.9	99.9	99.9	99.9	99.9
ALL	99.0	98.9	99.0	99.4	86.2

Performances du modèle hybride

Si on a moins de docs ?

- Modèle immédiatement meilleur que les règles
- On voit qu'on peut s'arreter autour de 1500 documents
- Mais la performance croît avec le nombre de documents



Combien de documents faut-il ?

- ⚠️ Analyse à prendre avec des pincettes
- Certaines entités sont plus difficiles à apprendre que d'autres
- Entités "regexables" les plus rapides: `ZIP`, `EMAIL`, `PHONE`, `ADDRESS`
- Entités ambigües les plus lentes: `FIRSTNAME`, `LASTNAME`, `VISIT ID` et `PATIENT ID`
- Difficulté semble plus fonction du #entités que #docs

Quantités pour 98% de score max

label	docs	ents	score
ZIP	10	2	98.0
EMAIL	248	15	97.6
ADDRESS	122	25	97.1
PHONE	100	33	97.8
CITY	147	46	96.8
DATE	10	48	97.7
NSS	1275	54	96.6
BIRTHDATE	449	103	97.4
VISIT ID	2000	177	91.2
PATIENT ID	1559	207	96.2
FIRSTNAME	108	269	96.8
LASTNAME	141	348	96.7

Incidence de l'extraction PDF

Rappel: EDS-PDF retire environ 80% des entités identifiantes

PDF extraction	P	R	F1	Redacted	Full
edspdf	99.1 ± 0.1	98.8 ± 0.1	98.9 ± 0.1	99.2 ± 0.1	**93.1 ± 1.0**
pdfbox	99.1 ± 0.0	98.9 ± 0.2	99.0 ± 0.1	99.4 ± 0.1	**75.7 ± 3.0**

On observe :

- la même performance selon les métriques micro-moyennées (P, R, F1, Redact)
- mais le nombre de documents entièrement caviardés gagne *18.6%* !

Incidence du Transformer

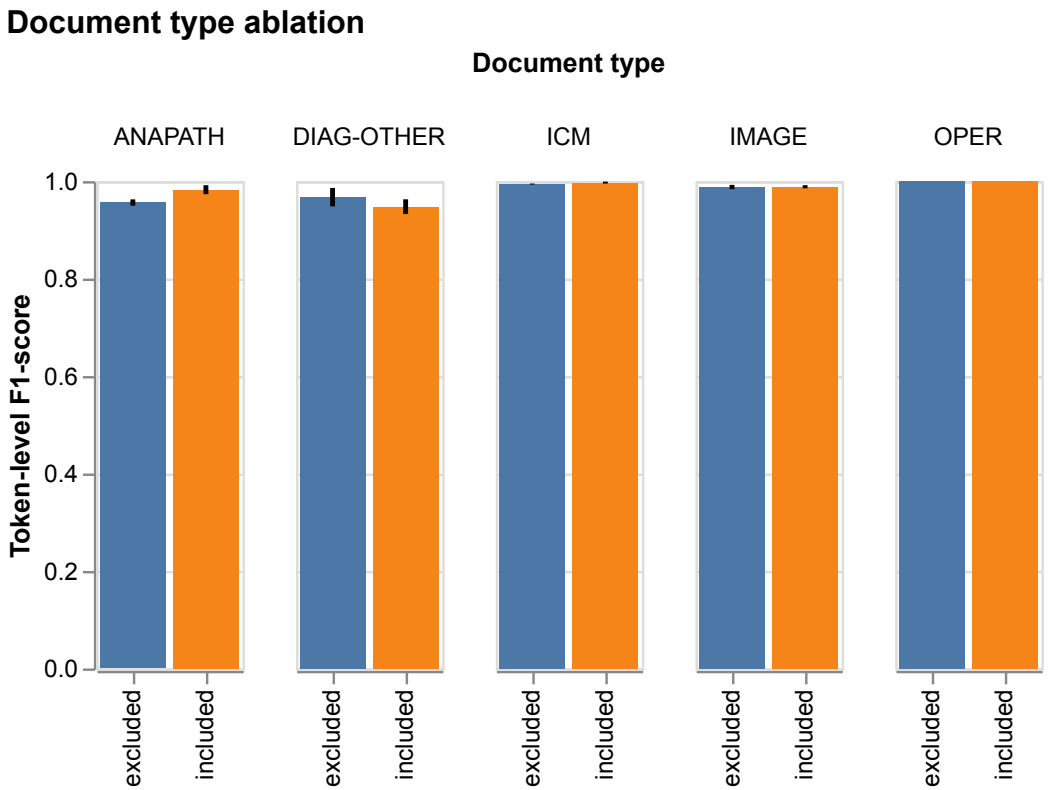
- **camembert**: Pré-entrainement sur corpus général français (OSCAR/FR)
- **finetuned**: Finetuning de Camembert sur 28M CR originaux, 1 epoch
- **scratch**: Pré-entrainement sur 28M CR pseudonymisés, ? epochs

Model	P	R	F1	Redact	Full
finetuned	97.8	97.7	97.8	98.2	75.5
camembert	96.8	96.9	96.8	97.4	68.9
scratch	97.3	97.2	97.3	97.6	69.0

- Finetuned est significativement meilleur
- Et plus rapide à entrainer que *scratch*

Si on a oublié des documents dans le jeu d'entrainement ?

- On retire certains documents du train (*excluded*) et on regarde la performance sur des documents dans le jeu de test



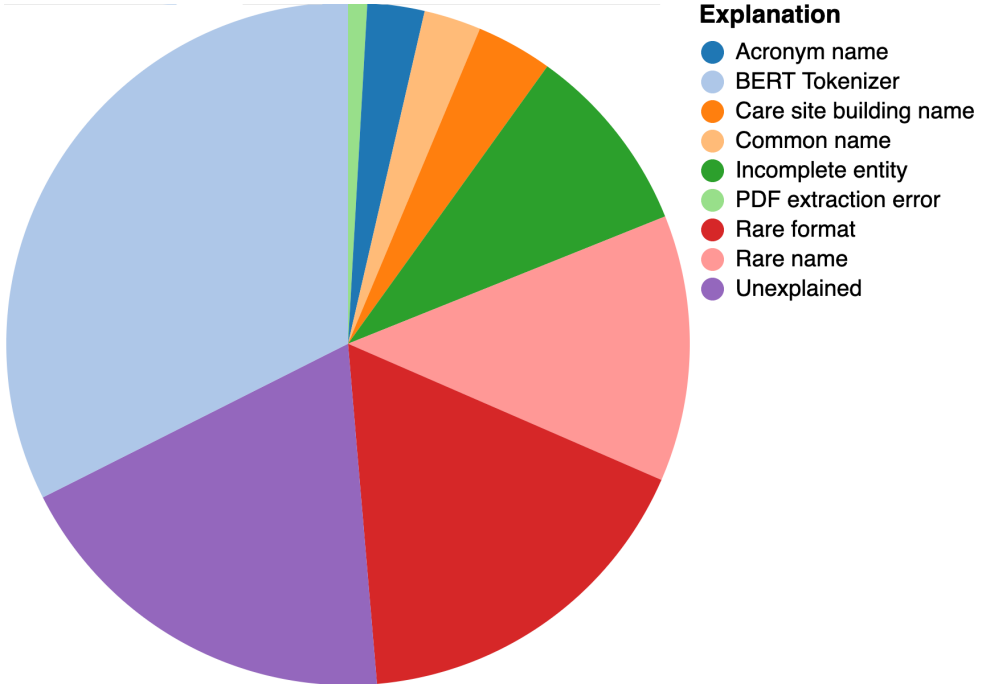
- Relativement robuste si il manque un type de document
- Dans le doute, il vaut mieux avoir un dataset d'entrainement diversifié

- Le NLP est en **forte croissance** à l'entrepôt des données de santé de l'AP-HP
- Des **craintes subsistent** quant à la qualité des études utilisant des données issues du NLP – un travail de consolidation technologique, de validation statistique et d'appropriation est nécessaire
- Les algorithmes de NLP sont nécessaires à **plusieurs niveaux des chaînes de traitement**, étant ainsi soumis à des contraintes techniques, réglementaires et organisationnelles différentes
- Plusieurs **projets open sources** ont été initiés et des communautés variées sont en cours de formation

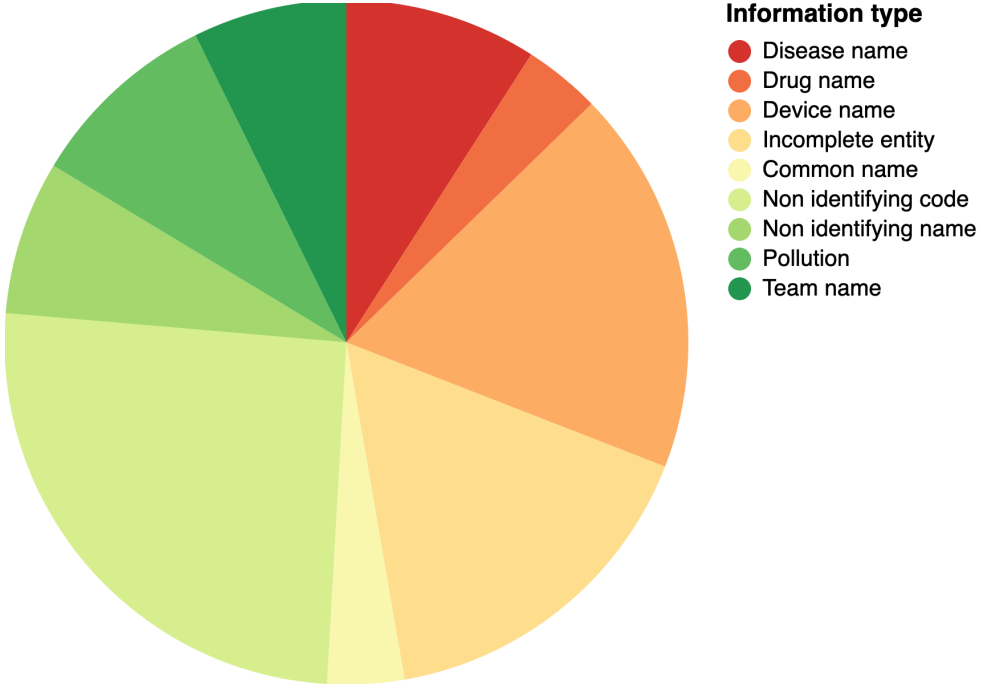
Merci !

Annexes

Quelles erreurs ?



Entités complètement manquées



Entités complètement inventées

- Parmis les *rares* entités complètement inventées (qqs dizaines): 30% sont graves
- Tokenizer responsable d'un tiers des *rares* entités complètement manquées

Performance par type et effet de la fusion règles — ML

Type	RB		ML		Hybrid	
PDF	edspdf	pdfbox	edspdf	pdfbox	edspdf	pdfbox
RCP	99.3	86.5	99.7	98.8	99.7	98.8
CR-URGE	83.3	80.3	99.6	99.2	99.6	99.1
CRH-HOSPI	93.5	83.1	99.2	99.2	99.2	99.2
CR-CONS	93.1	89.1	99.1	99.1	99.3	99.2
CR-ANAPATH	99.0	83.6	98.1	96.6	98.1	96.6
CR-IMAGE	88.1	90.0	98.9	97.8	98.9	97.9
CR-OPER	100.0	90.9	100.0	99.3	100.0	99.4
ALL	92.7	86.1	98.9	99.0	99.0	99.0