Institut Thématique Multi-Organismes
Technologies pour la santé

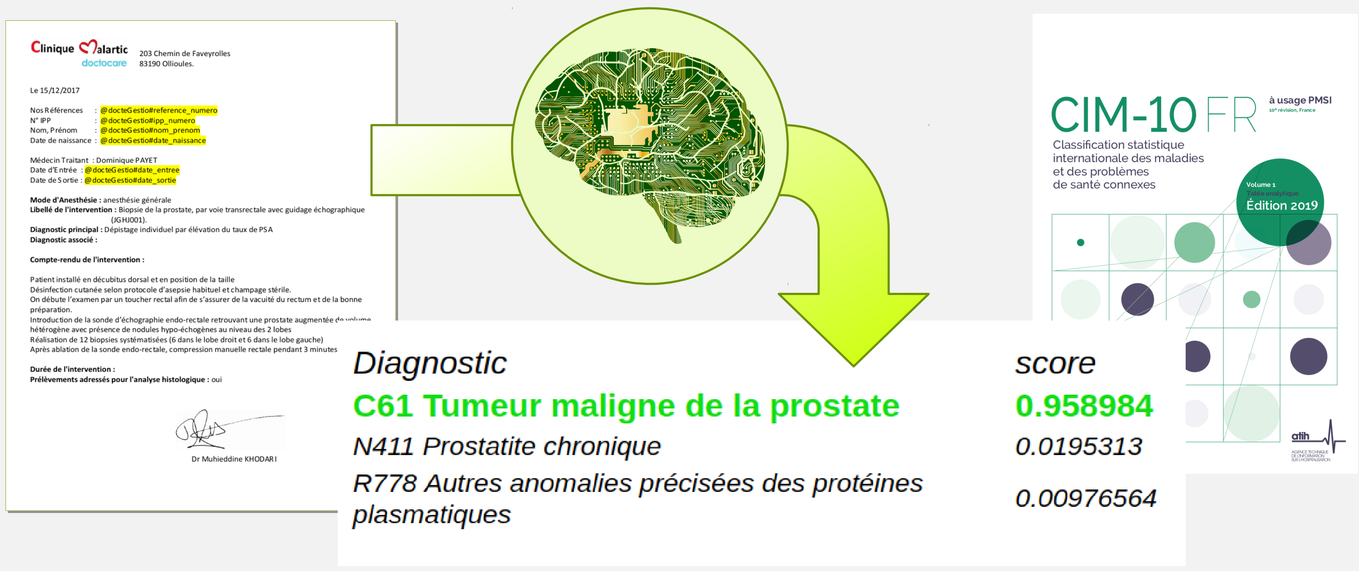
Doctegestio

Analyse sémantique et codification automatique de textes médicaux par l'IA

Romain Farel, Romaric Besançon CEA-LIST Contact : romain.farel@cea.fr

Enjeux

- Simplifier le traitement administratif des dossiers des patients pour permettre aux médecins hospitaliers de se concentrer sur leurs activités à plus forte valeur ajoutée
- Accélérer et fiabiliser la chaîne de facturation des actes médicaux.
- Proposition
  - Aide à la codification automatique des dossiers patients à partir de l'analyse du texte libre des documents de ces dossiers



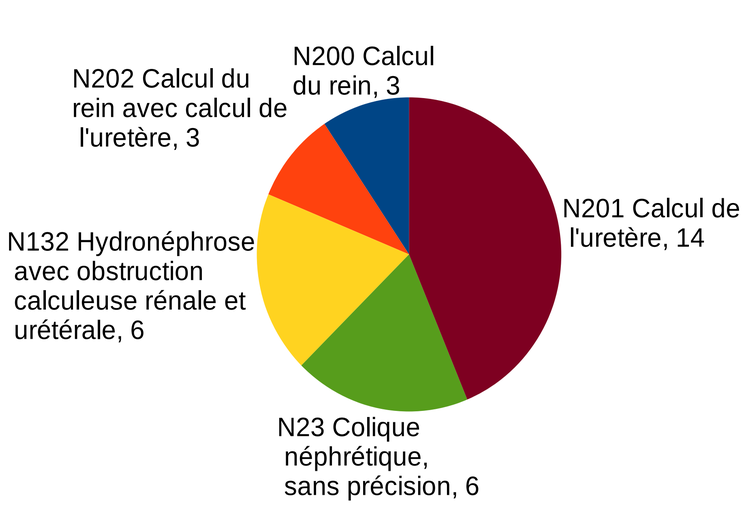
Apprentissage automatique de modèles pour la classification de textes

- Classification des diagnostics **au niveau global des dossiers**
  - les intitulés des diagnostics de la CIM-10 ne sont pas toujours dans les textes

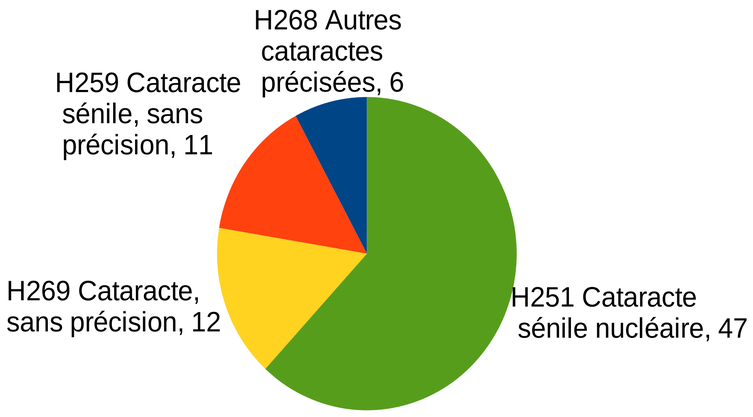
«dépistage individuel par élévation du taux de PSA » → C61 Tumeur maligne de la prostate

  - les descriptions de diagnostics sont ambiguës vis-à-vis des classifications

« colique néphrétique gauche »



« cataracte œil gauche »



Modèles d'apprentissage automatique

- Représentations des textes sous formes de vecteurs de traits : modèles vectoriels classiques (*tf-idf*) ou représentations distribuées (*word embeddings*)
- Apprentissage automatique de modèles de classification :
  - repérage des traits et combinaison de traits les plus discriminants pour catégoriser le texte : modèles statistiques classiques (*SVM, Random Forest, XGBoost*), ou modèles neuronaux (*CNN, Fasttext*)

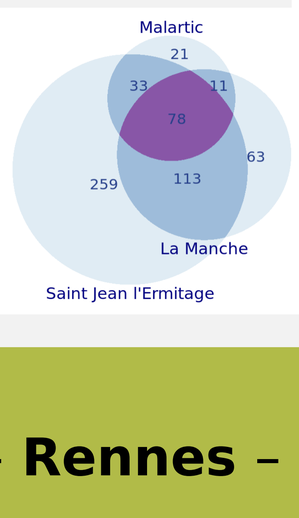
Évaluation

- Données
  - environ 110 dossiers patients provenant de trois cliniques (Malartic, La Manche, Saint-Jean L'Ermitage)
  - 2267 diagnostics représentés
  - 725 diagnostics avec une fréquence > 10

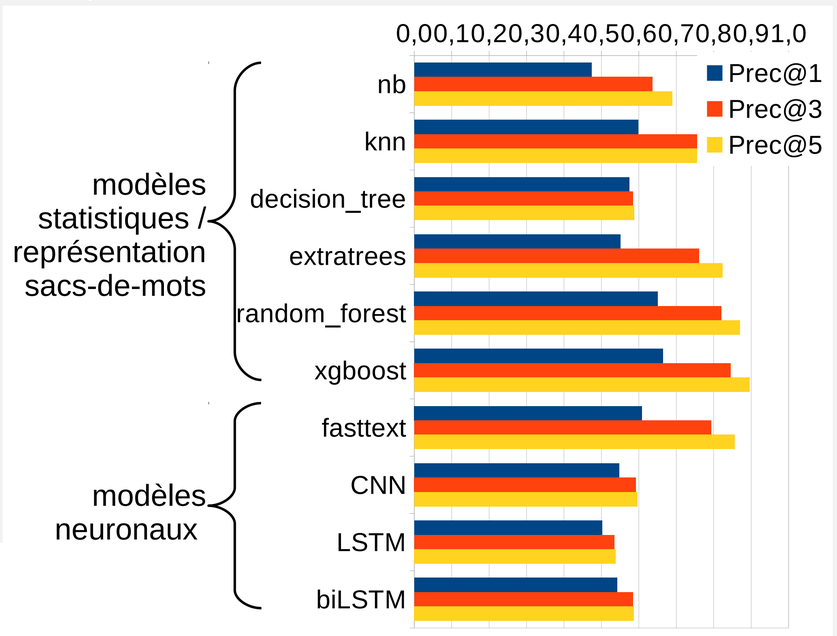
Précision sur 725 diagnostics
60 % à la première réponse
80 % dans les 5 premières

- **Difficulté** : adaptation des modèles appris d'une clinique à l'autre, en particulier à cause de la variabilité des diagnostics

Appris sur \ testé sur	Malartic	La Manche	Saint Jean	Malartic+ La Manche	Malartic+ Saint Jean	La Manche+ Saint Jean	Tous
Malartic	0,602	0,068	0,038	0,260	0,108	0,044	0,101
La Manche	0,185	0,615	0,093	0,531	0,101	0,241	0,237
Saint-Jean	0,329	0,208	0,528	0,240	0,515	0,474	0,466
Malartic+La Manche	0,624	0,623	0,111	0,623	0,165	0,242	0,273
Malartic+Saint Jean	0,625	0,230	0,531	0,343	0,538	0,479	0,488
La Manche+Saint Jean	0,333	0,629	0,532	0,560	0,518	0,551	0,539
Tous	0,636	0,632	0,534	0,633	0,542	0,554	0,559



Étude des performances des différents modèles de classification



→ Amélioration par l'apprentissage de modèles plus ciblés (par clinique, par service, par sous-catégorie CIM10)