Classification automatisée des thèses par IA

Tommy Arrault, Benjamin Bastian, Marianne Cinot
Rémi Sureau, Grégoire Pau
DMG Rennes, POPS

Thomas Morel, Jérôme Nguyen-Soenen
DMG Nantes, POPS

Pas de conflit d’interêt

Contexte

- Vaste ressource pour la recherche en soins primaire
  - 3.500 thèses de MG par an
  - Sudoc > 44.000 thèses MG en ligne
  - Mal annotée, sous-exploitée

- Objectifs
  - Évaluer l'utilisation de LLM pour la classification de thèses
  - Pour l'aide à la validation de projet de thèse
  - Pour cartographier les thèmes abordés en MG

Classification de thèses par ChatGPT

- Extraction de 4578 résumés de thèse de médecine (MG ou pas)

BU Angers 1809

BU Rennes 1397

BU Nantes 1372

- Classification par ChatGPT

```
Pour la thèse suivante :
1. Est-ce une thèse de médecine générale (réponds par oui ou non) ?
2. Est-ce que la méthode est quantitative, qualitative ou autre ?
3. Quels sont les 3 codes CISP-2 les plus adaptés pour classer cette thèse ?
4. Quels sont les 3 mots décrivant le mieux cette thèse ?
```

Consommation : 3.5 heures, 6.5 euros, 13.7 kg de CO_2 (1 requête = 3 g CO_2, Bashir 2024)

Validation par annotation manuelle

- Annotation manuelle de 360 thèses par double codage et consensus

MG ou pas ?

Exactitude 91.7%

Divergences : médecine sport, HAD, urgences, interdisciplinarité…

➔ 51.8% thèses de MG

Validation par annotation manuelle

- Annotation manuelle de 360 thèses par double codage et consensus

MG ou pas ?

Exactitude 91.7%

Divergences : médecine sport, HAD, urgences, interdisciplinarité…

➔ 51.8% thèses de MG

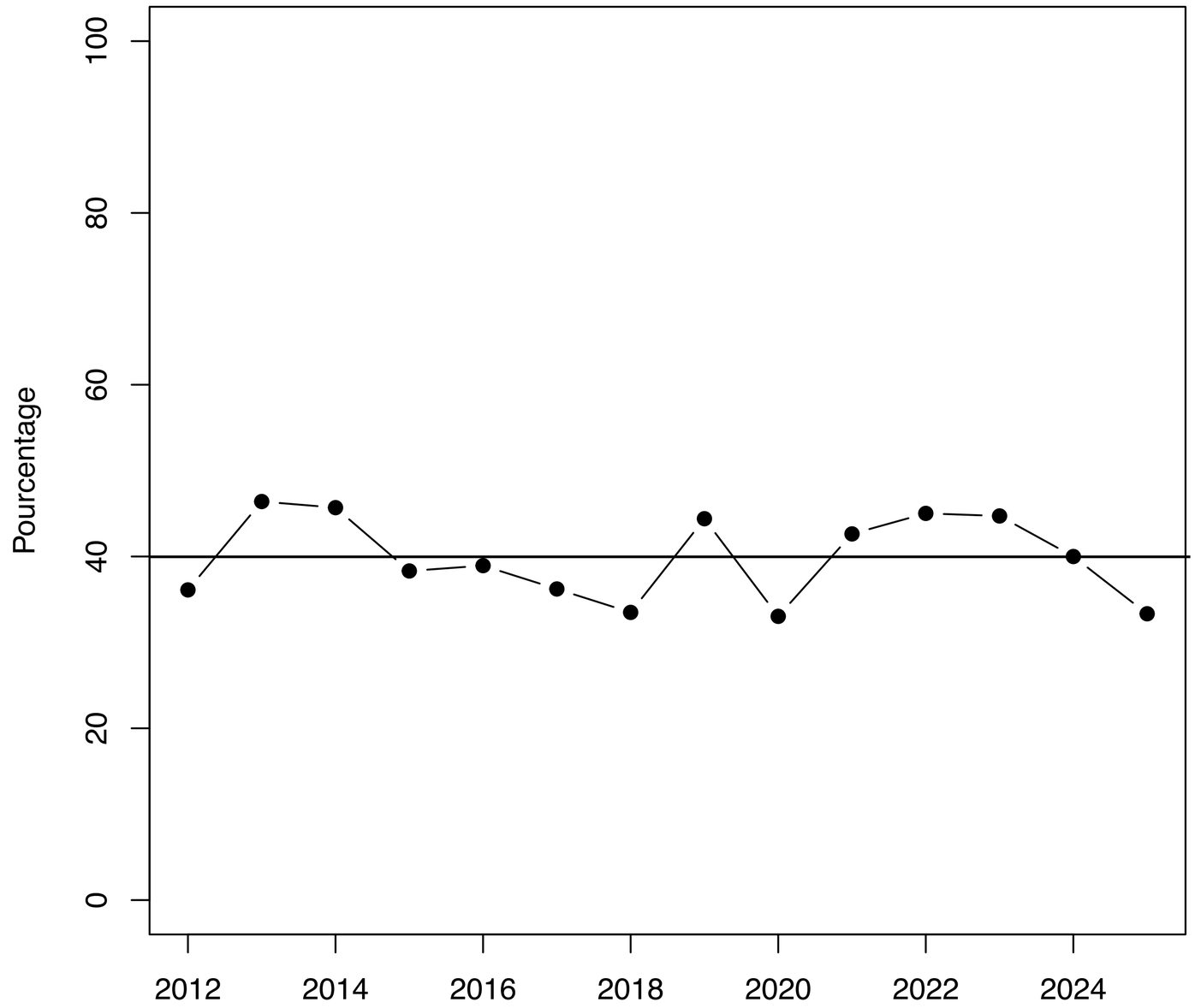
Quanti, quali ou autre ?

Sensibilité 98.1%

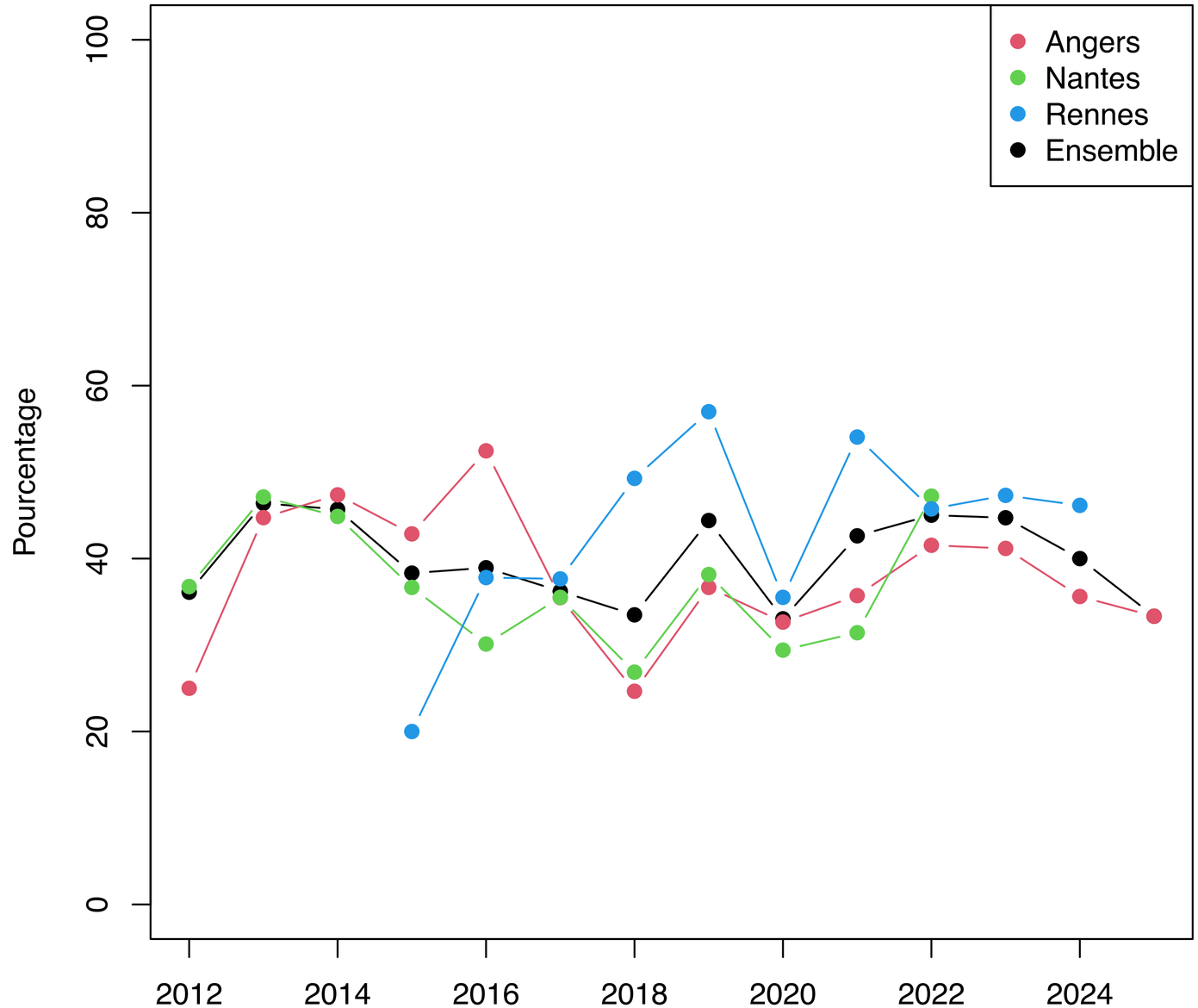
Divergences : méthode non renseignée, mixte, ambigüe…

➔ 40.2% thèses de MG quali

40.2% de thèses de MG utilisent une méthode qualitative



40.2% de thèses de MG utilisent une méthode qualitative



Rennes : 45.5%**

Angers : 38.2%

Nantes : 36.9%

Validation par annotation manuelle

- Annotation manuelle de 360 thèses par double codage et consensus

MG ou pas ?

Exactitude 91.7%

Divergences : médecine sport, HAD, urgences, interdisciplinarité…

➔ 51.8% thèses de MG

Quanti, quali ou autre ?

Sensibilité 98.1%

Divergences : méthode non renseignée, mixte, ambigüe…

➔ 40.2% thèses de MG quali

CISP-2

Exactitude 67.1%

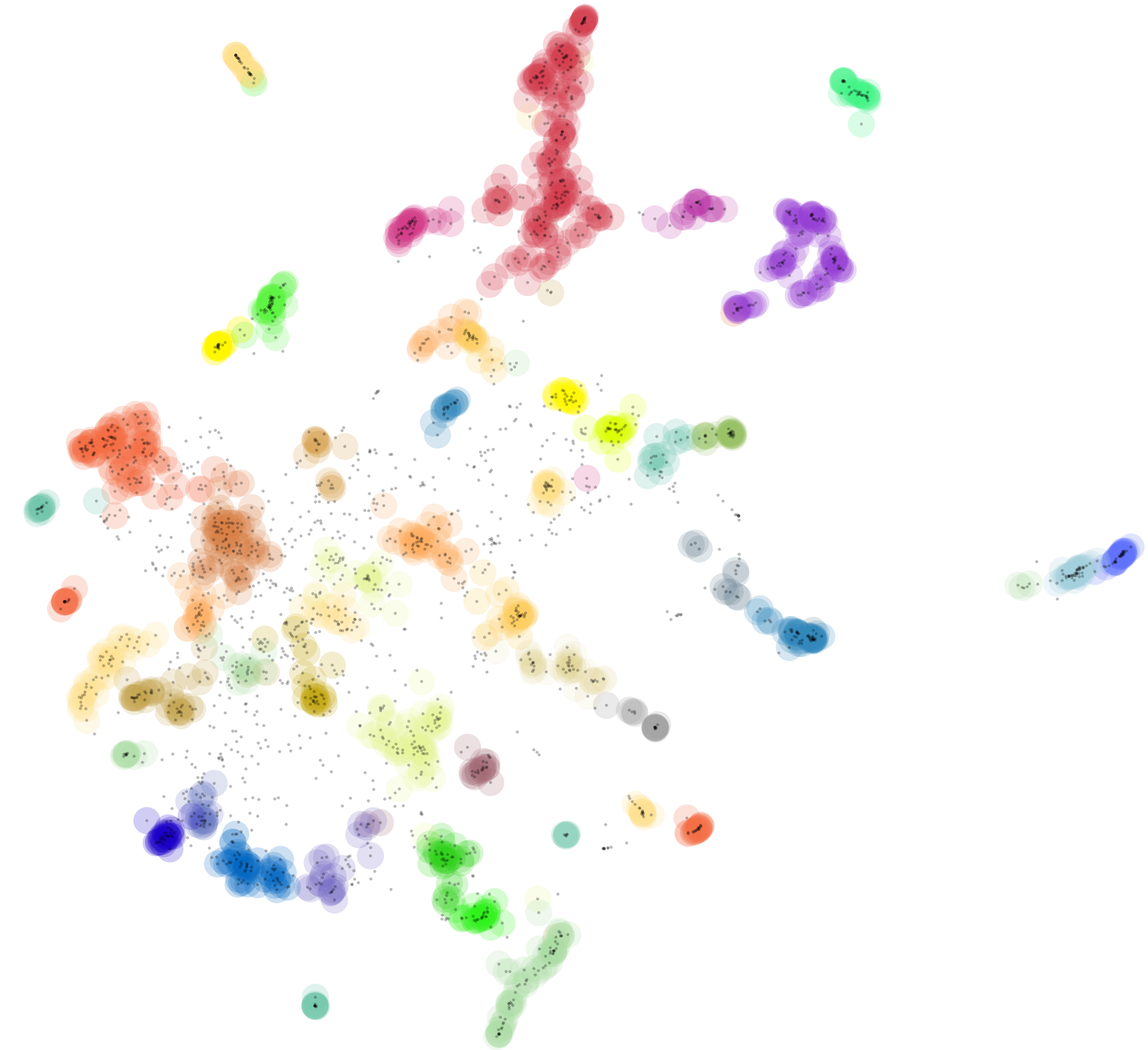
Divergences : beaucoup de thèses non-cliniques, approximants sémantiques

Espaces sémantiques et métrique

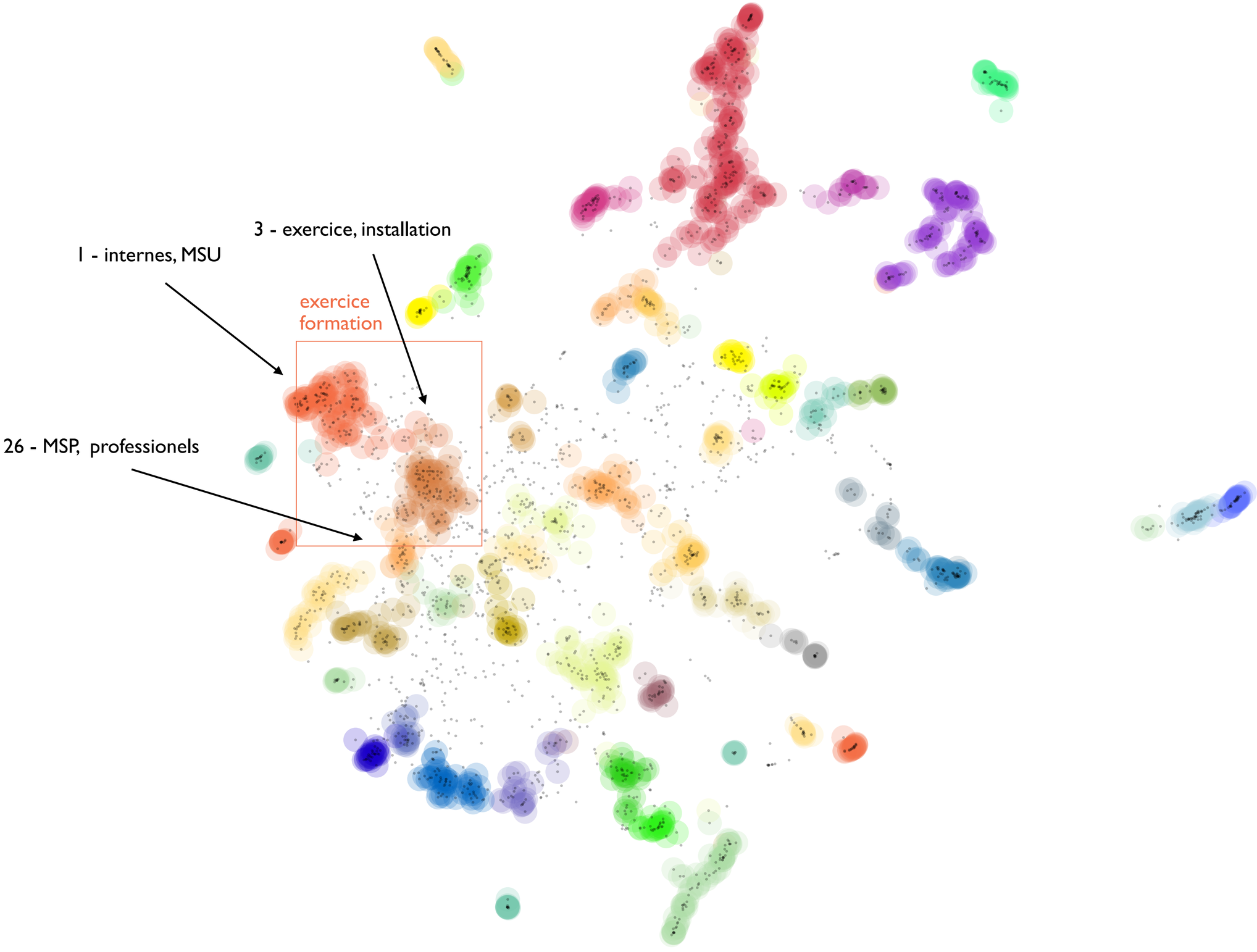
Mise en place d'un programme de formation sur la sécurité des soins à destination des internes

Évaluation des connaissances des patients sur la vaccination HPV

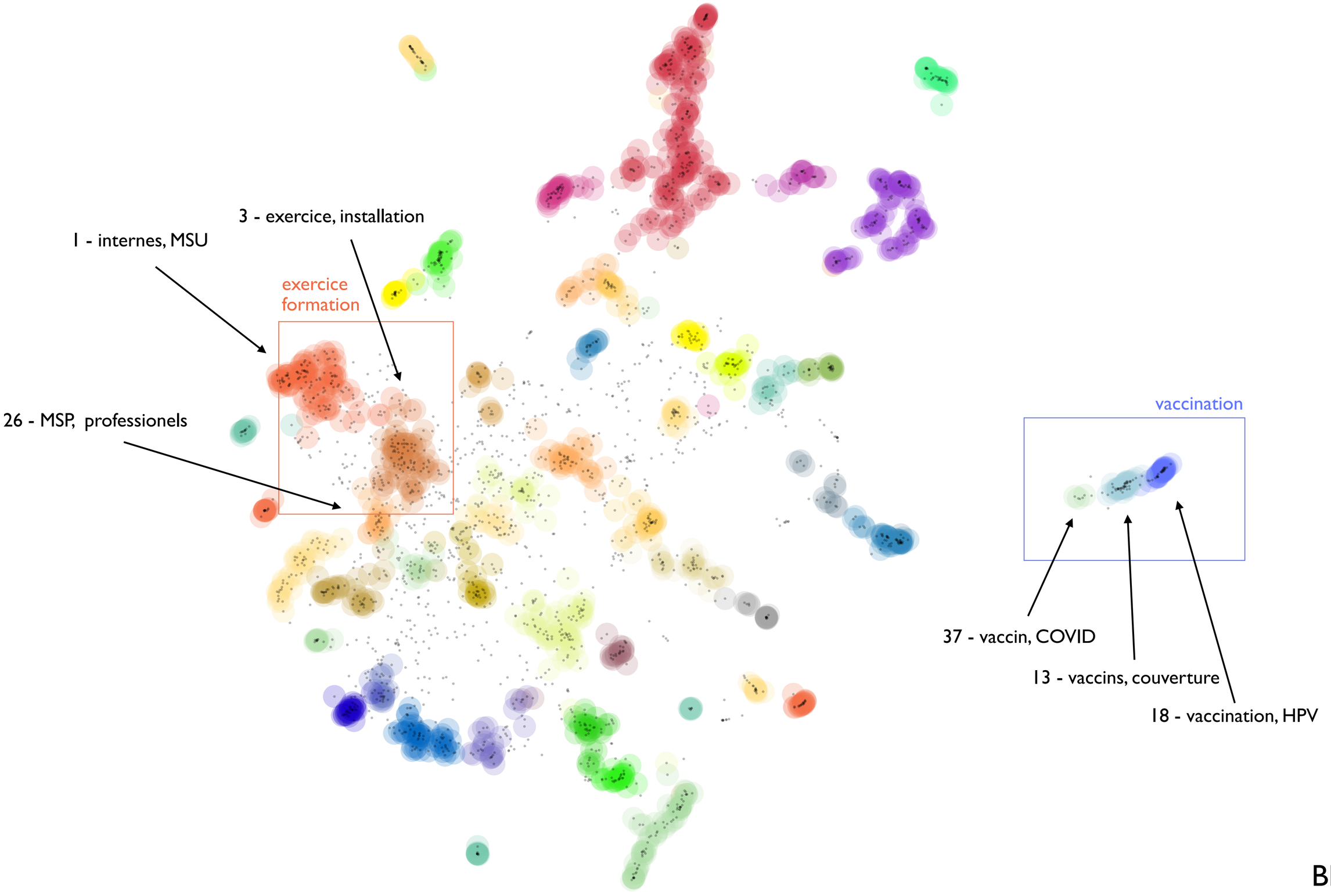
Étude des déterminants à la réticence de la vaccination contre la grippe



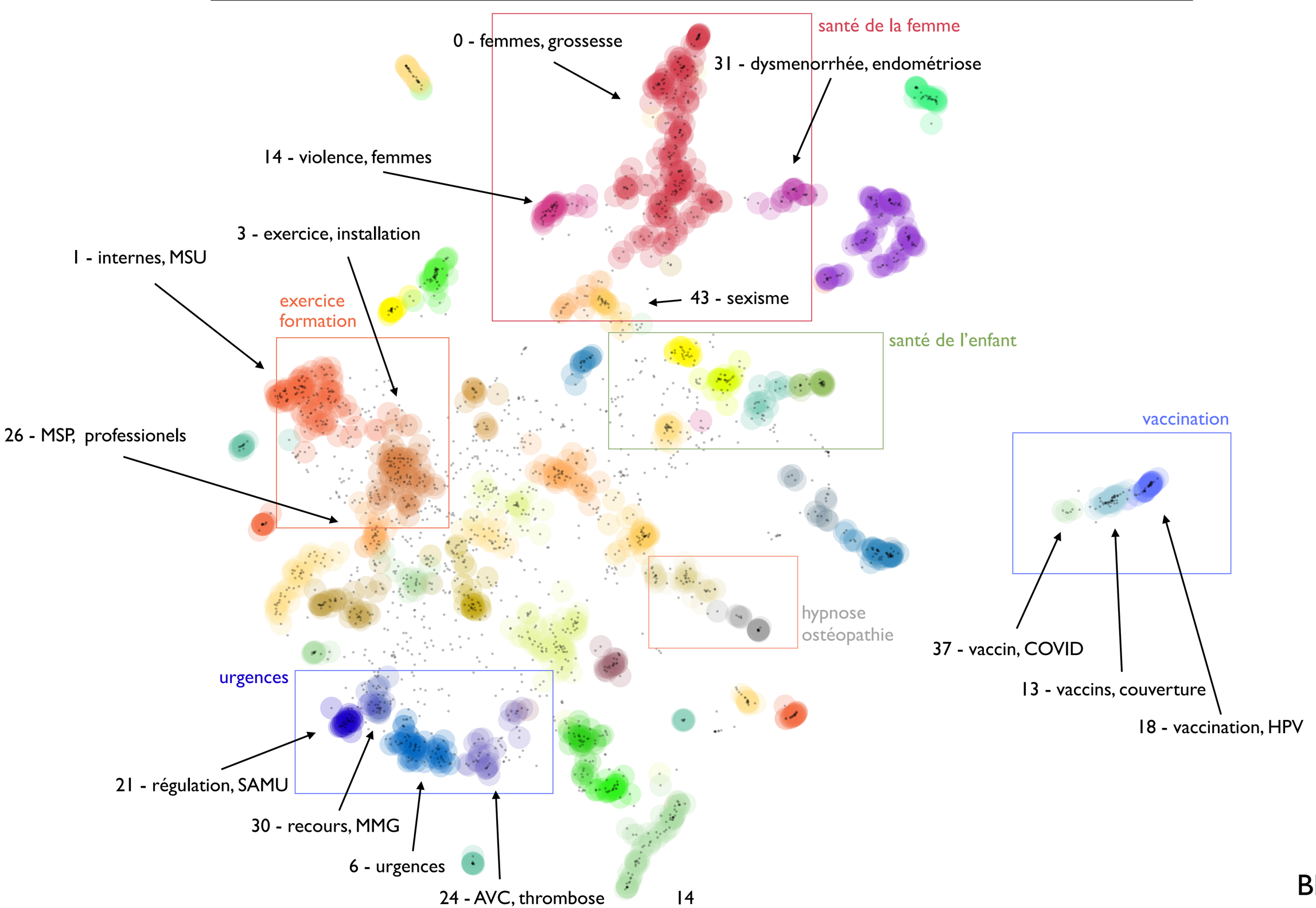
BERTopic


3 - exercice, installation
1 - internes, MSU
exercice
formation
26 - MSP, professionels

BERTopic


3 - exercice, installation
1 - internes, MSU
exercice
formation
26 - MSP, professionels
vaccination
37 - vaccin, COVID
13 - vaccins, couverture
18 - vaccination, HPV

BERTopic


santé de la femme
0 - femmes, grossesse
31 - dysmenorrhée, endométriose
14 - violence, femmes
3 - exercice, installation
1 - internes, MSU
43 - sexisme
exercice
formation
santé de l'enfant
vaccination
26 - MSP, professionels
hypnose
ostéopathie
37 - vaccin, COVID
13 - vaccins, couverture
18 - vaccination, HPV
urgences
21 - régulation, SAMU
30 - recours, MMG
6 - urgences
24 - AVC, thrombose

BERTopic

Conclusion

- Validation de l'approche de classification de thèse par LLM
- Pour l'aide à la validation de fiche de thèse
- Pour les choix des sujet, sur- et sous-représentation
- Pour la recherche en pédagogie
- A étendre aux 44.000 thèses de médecine de Sudoc ?

Merci de votre attention !