Topic modeling and classification of scientific disciplines

Radim Hladík[1,2]
Yann Renisio[3]

[1] Centre for Science, Technology, and Society Studies (CSTSS) @ IP Czech Academy of Sciences
[2] visiting scholar at Centre Georg Simmel @ EHESS

[3] Centre for Research on social InequalitieS (CRIS) @ CNRS/Sciences-Po

Motivation

- between 2006 and 2020, more than 300k Ph.D. theses submitted at French universities
- no controlled vocabulary for the variable “discipline”
  - 23057 unique labels for “discipline”
    - 14538 labels appear only 1x
- regular expression and fuzzy matching can only go so far
- to analyze the data on French doctoral degrees, disciplines have to be reliably inferred

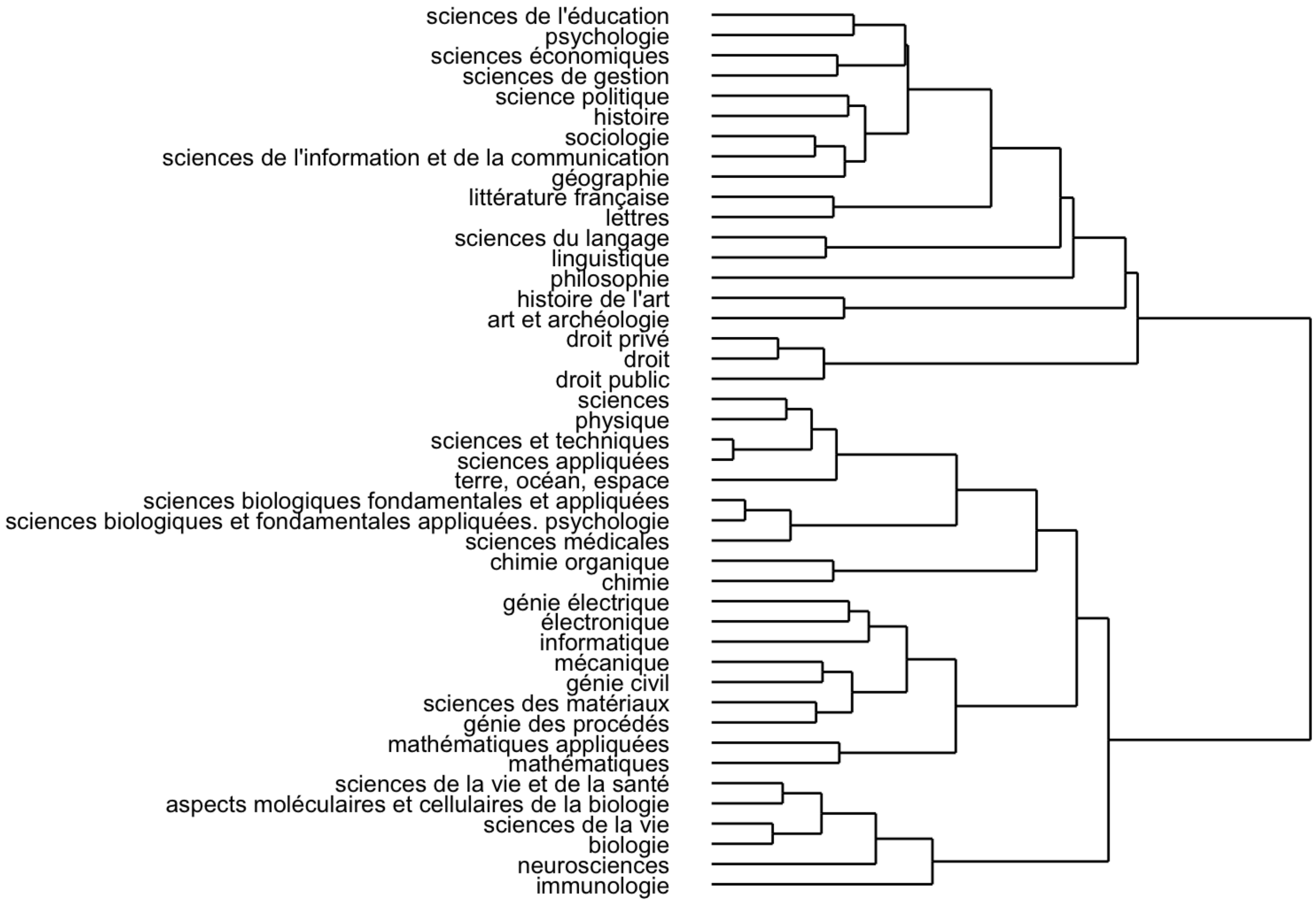
Data

- abstracts (+ title, keywords) of ~285k of French doctoral theses and their disciplinary labels
- preprocessing
  - lemmatization (UDPipe)
  - removal of stopwords and non-alphabetical characters
  - compounding frequent bi-grams and tri-grams
- topic modeling with TopSBM
  - 7 levels of topic hierarchy
  - 2043 topics at level 1 (most nuanced)

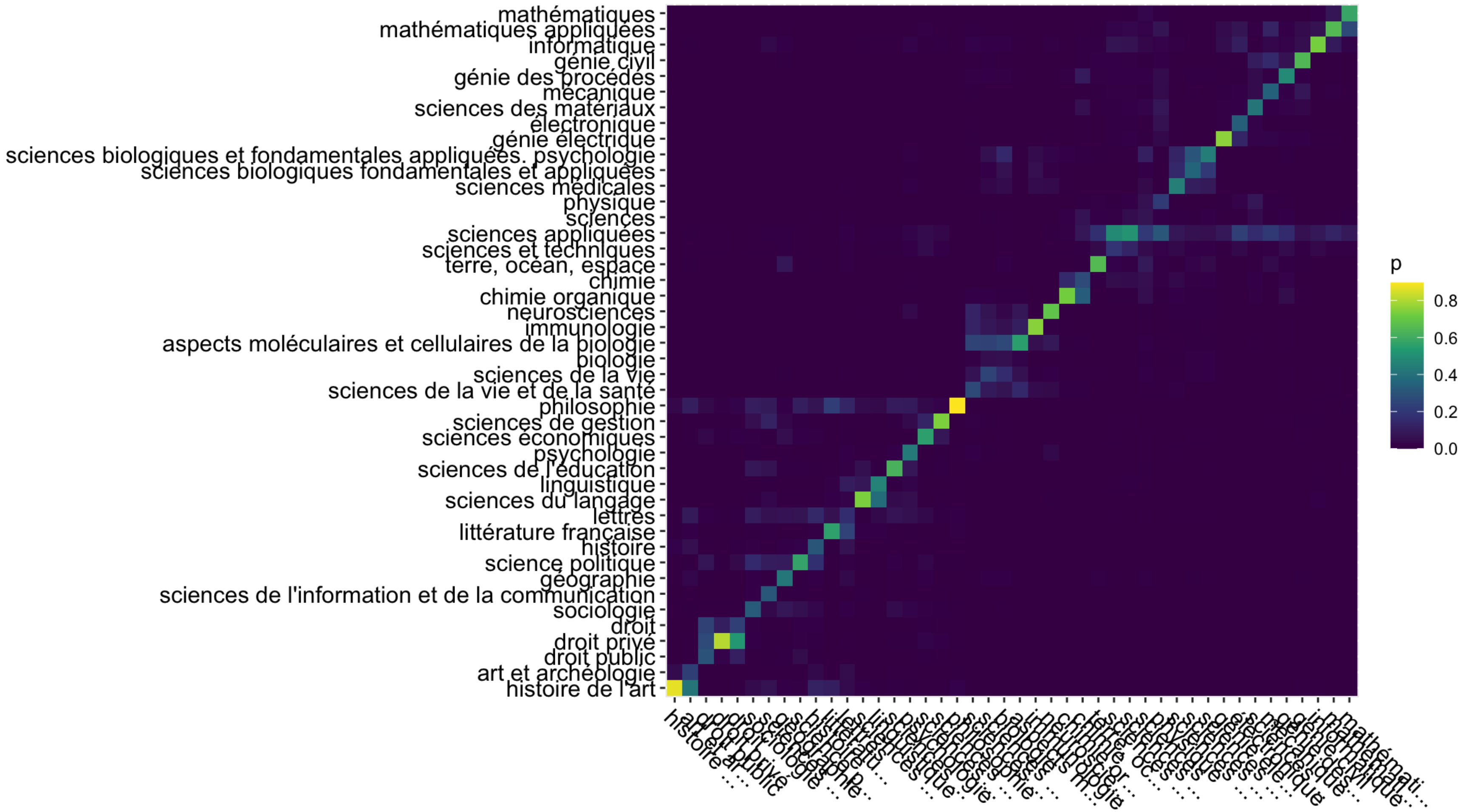
Method

- identify most frequent discipline labels ($n > 1000$)
  - 44 labels, ~146k documents
- hiearchical clustering
  - mean topic vectors for disciplines
- create “training” data subset ($n = 14601$)
  - 10% of documents for each disciplines
  - reference topic vectors
    - mean topic vectors for disciplines from the “training” subset
- assign the least surprising discipline to each document
  - for each document in the “test” subset, assign the discipline with the smallest Kullback–Leibler divergence from the reference topic vector

Clustering of disiciplines in topic space



The most frequent disciplines

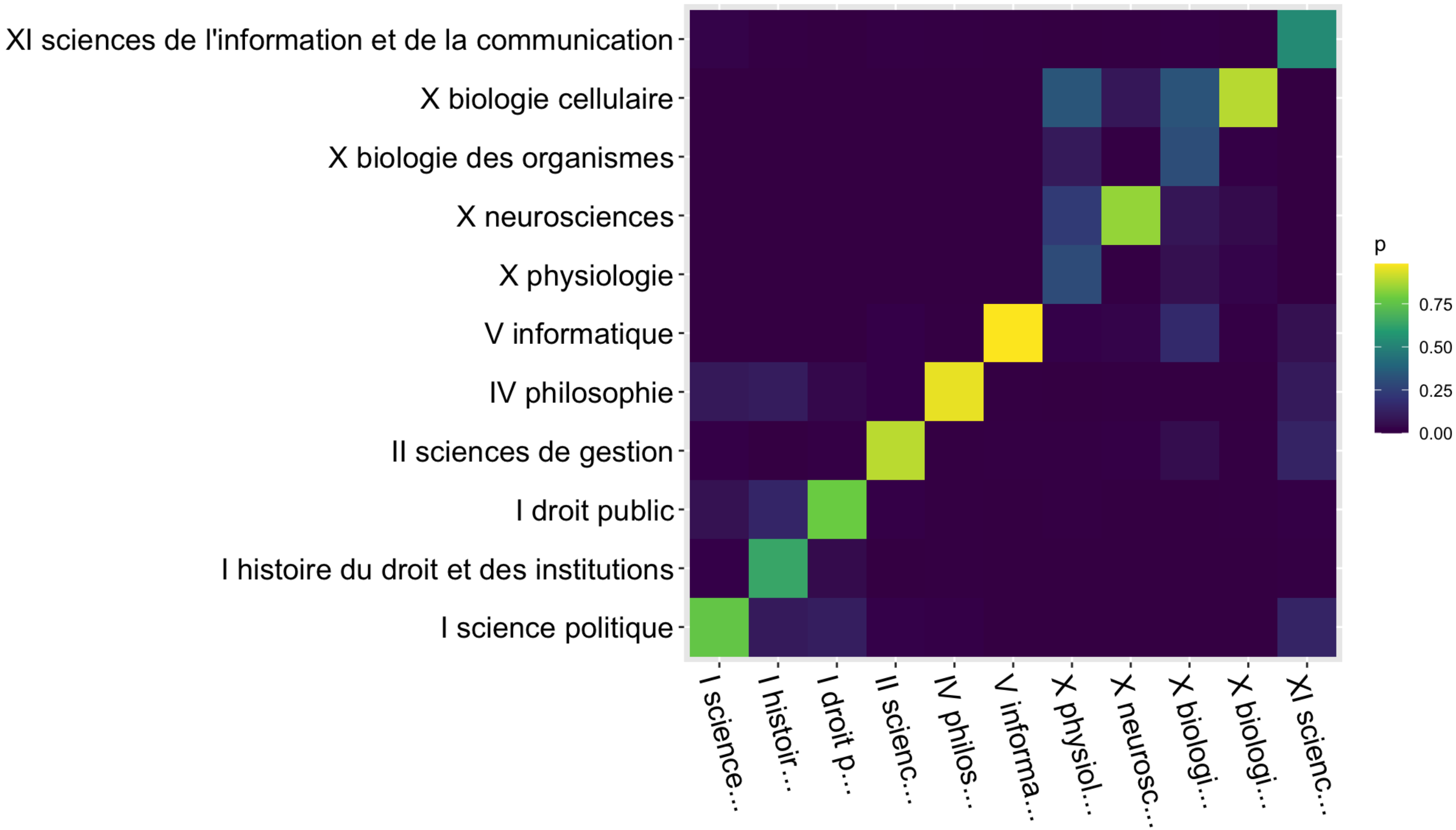


accuracy = 47%

Nested disciplinary nomenclature

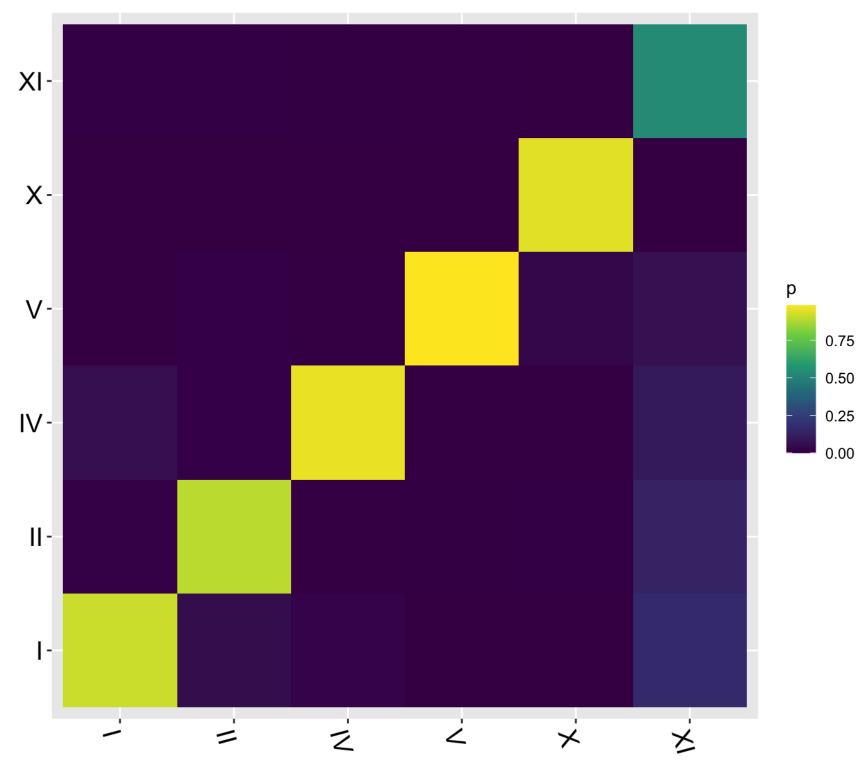
- Conseil National des Universités
  - advisory and administrative body for oversight of researchers' careers
  - organized around 81 disciplinary sections in 11 official groups plus medical and pharmaceutical researchers
- 11 CNS sections have an exact match in the theses dataset ($n = 28986$)
  - the 11 sections come from 6 groups
- classification improves as we move upward in the disciplinary hiearchy
  - 86% accuracy at sections level, 91% accuracy at groups level

Nested disciplines



Disciplinary groups

- I Group (Droit, économie, gestion)
  - science politique, histoire du droit et des institutions, droit public
- II Group (Droit, économie, gestion)
  - sciences de gestion
- IV Group (Lettres et sciences humaines)
  - philosophie
- V Group (Sciences)
  - informatique
- X Group (Sciences)
  - physiologie, neurosciences, biologie des organismes, biologie cellulaire
- XI Group (Lettres et sciences humaines)
  - sciences de l'information et de la communication



Further directions

- how many documents are required for good reference topic vectors?
- how fine-grained topic models need to be?
  - and does the chosen algorithm make a difference?
- does the approach generalize to other types of scholarly documents?
- can topic distributions help with author disambiguation?
  - especially when assigning new documents to existing author publication profiles

Conclusion

- topic models contain signal about disciplines
- topic models can be successfully used in collections where citation data are missing
- do we even need disciplinary labels when we have topic models?
  - or should we draw a sharper distinctions between social and discursive manifestation of disciplinary boundaries?

Acknowledgements

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS