IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

ENS | PSL

Impact de la classe de documents dans l'extraction automatique d'environnements mathématiques dans la littérature scientifique

Antoine Gauquier

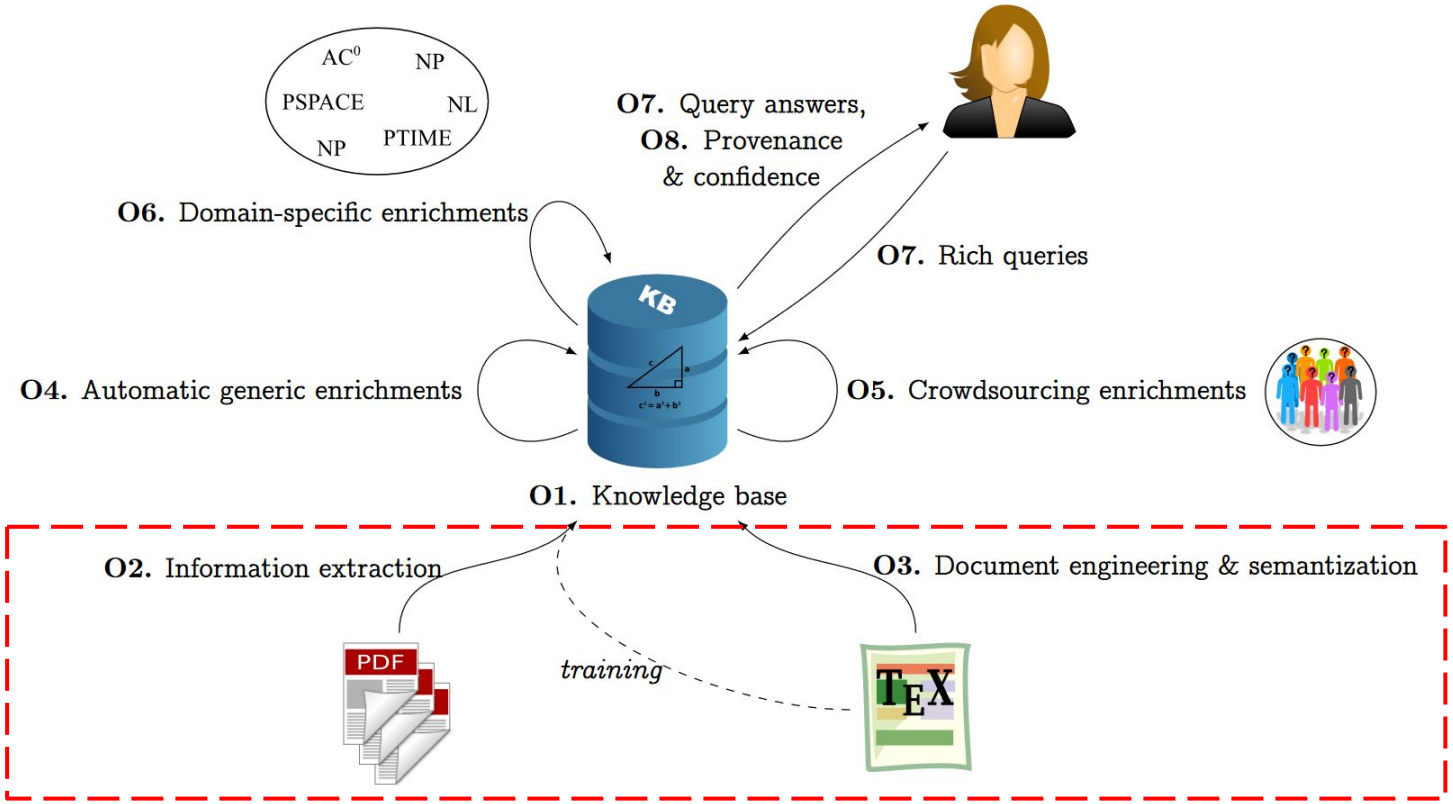
FISE 2023

2022-2023

Soutenance de Projet de Fin d'Études (PFE) – Mardi 11 juillet 2023

Introduction

TheoremKB



Source : https://pierre.senellart.com/talks/sinfra-20191213.pdf

Introduction

Impact de la **classe de documents** dans l'extraction automatique d'environnements mathématiques dans la littérature scientifique

LaTeX → `\documentclass`

- Association for Computing Machinery (ACM)
- American Astronomical Society (AAS)
- American Mathematical Society

Structurent le rendu PDF des articles

. . .

Introduction

Association for Computing Machinery

American Astronomical Society

American Mathematical Society

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

Sommaire

ENS | PSL

1. Inférence automatique de la classe de document
2. Extraction des environnements mathématiques dans la littérature scientifique
3. Étude de l'impact de la classe de document dans cette tâche d'extraction
4. Tâches encore en cours et futur de ce travail (conclusion)
5. Bilan des compétences, projet professionnel et tâches annexes au cours du stage

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

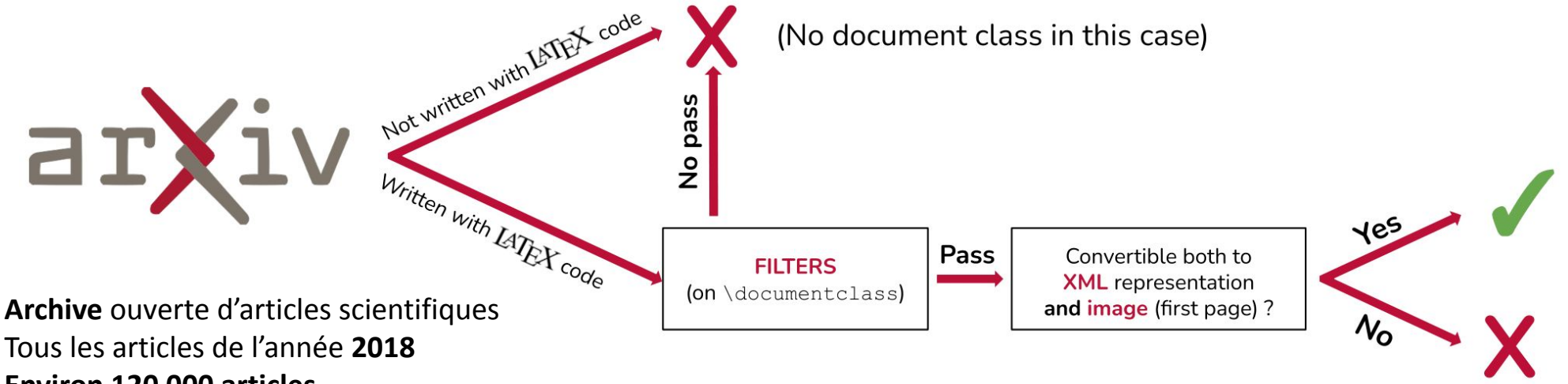
ENS | PSL

1. Inférence automatique de la classe de document

1. Inférence automatique de la classe de document

Problème à résoudre et données

Filtrage des articles



- **Archive** ouverte d'articles scientifiques
- Tous les articles de l'année **2018**
- **Environ 120 000 articles**

Un peu moins de 100 000 articles exploitables au final

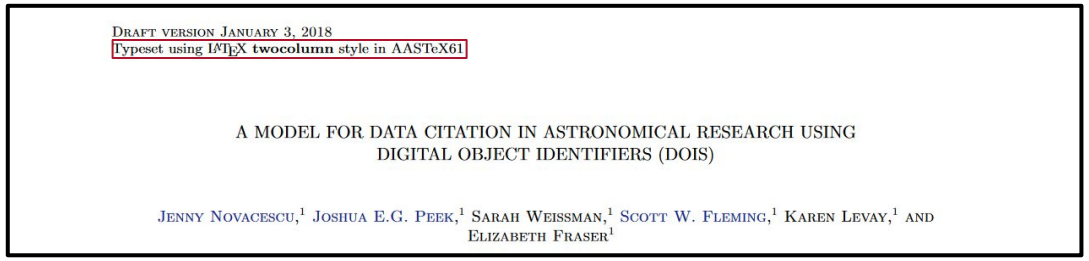
Filtrage des classes de document

Plus de 1200 → *Au moins un millième des données* → **44** → *Regroupement basé sur une distance d'édition* → **33**

1. Inférence automatique de la classe de document

Modélisation choisie (1)

Exemple d'élément caractéristique dans une des classes de document de l'AAS



Draft version January 3, 2018
Typeset using LaTeX **twocolumn** style in AASTeX61

A MODEL FOR DATA CITATION IN ASTRONOMICAL RESEARCH USING DIGITAL OBJECT IDENTIFIERS (DOIS)

Jenny Novacescu,[1] Joshua E.G. Peek,[1] Sarah Weissman,[1] Scott W. Fleming,[1] Karen Levay,[1] and Elizabeth Fraser[1]

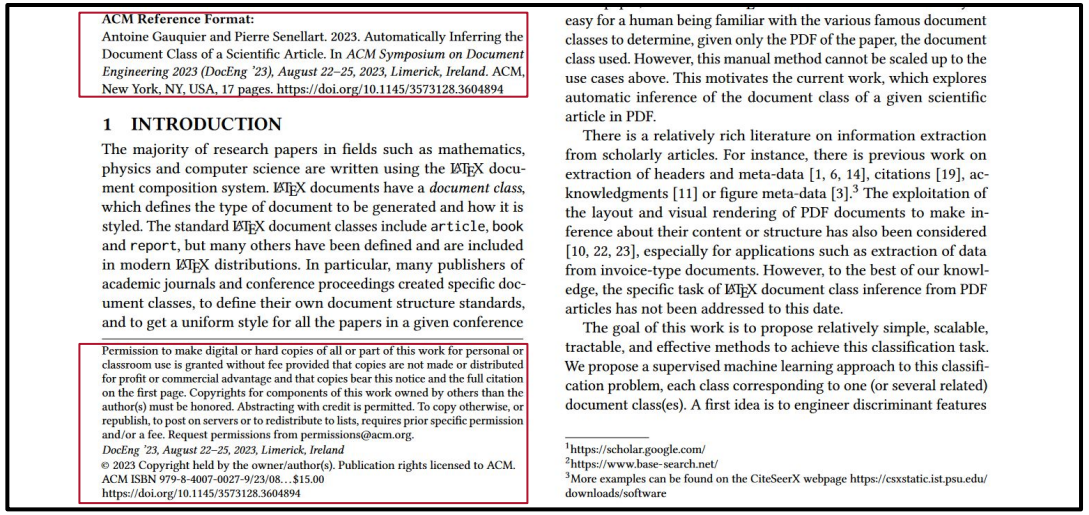
Source : https://arxiv.org/pdf/1801.00004.pdf

Exemple d'une image générée, fournie en entrée au CNN

256

Source : https://arxiv.org/pdf/1806.06252.pdf

Exemple d'élément caractéristique dans une des classes de document d'ACM



ACM Reference Format:
Antoine Gauquier and Pierre Senellart. 2023. Automatically Inferring the Document Class of a Scientific Article. In *ACM Symposium on Document Engineering 2023 (DocEng '23), August 22–25, 2023, Limerick, Ireland.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3573128.3604894

1 INTRODUCTION

The majority of research papers in fields such as mathematics, physics and computer science are written using the LaTeX document composition system. LaTeX documents have a *document class*, which defines the type of document to be generated and how it is styled. The standard LaTeX document classes include `article`, `book` and `report`, but many others have been defined and are included in modern LaTeX distributions. In particular, many publishers of academic journals and conference proceedings created specific document classes, to define their own document structure standards, and to get a uniform style for all the papers in a given conference

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DocEng '23, August 22–25, 2023, Limerick, Ireland
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0027-9/23/08…$15.00
https://doi.org/10.1145/3573128.3604894

easy for a human being familiar with the various famous document classes to determine, given only the PDF of the paper, the document class used. However, this manual method cannot be scaled up to the use cases above. This motivates the current work, which explores automatic inference of the document class of a given scientific article in PDF.

There is a relatively rich literature on information extraction from scholarly articles. For instance, there is previous work on extraction of headers and meta-data [1, 6, 14], citations [19], acknowledgments [11] or figure meta-data [3].[3] The exploitation of the layout and visual rendering of PDF documents to make inference about their content or structure has also been considered [10, 22, 23], especially for applications such as extraction of data from invoice-type documents. However, to the best of our knowledge, the specific task of LaTeX document class inference from PDF articles has not been addressed to this date.

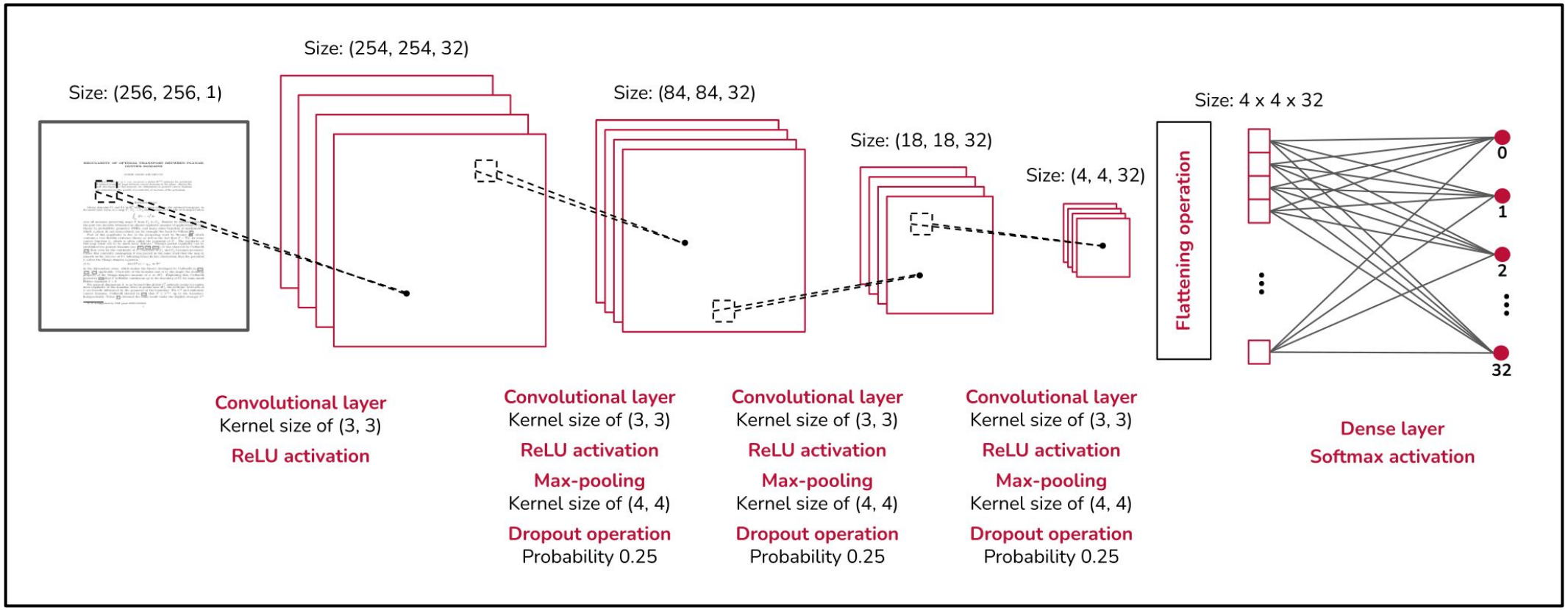
The goal of this work is to propose relatively simple, scalable, tractable, and effective methods to achieve this classification task. We propose a supervised machine learning approach to this classification problem, each class corresponding to one (or several related) document class(es). A first idea is to engineer discriminant features

[1] https://scholar.google.com/
[2] https://www.base-search.net/
[3] More examples can be found on the CiteSeerX webpage https://csxstatic.ist.psu.edu/downloads/software

Source : https://github.com/AntoineGauquier/inferring_document_class_of_scientific_article/blob/main/article.pdf

1. Inférence automatique de la classe de document

Modélisation choisie (2)



1. Inférence automatique de la classe de document

Résultats obtenus

Comparaison de performances entre notre architecture et quelques architectures de l'état de l'art

Architecture	F1-Score moyen (macro)	Nombre de paramètres	FLOPS (en milliards)
Notre architecture	*92.31 %*	***38 177***	*1.36*
ResNet50V2	92.28 %	23 632 417	9.13
NASNetMobile	91.31 %	4 304 597	1.50
EfficientNetV2B0	**93.43 %**	4 091 844	**0.80**

Conclusion :

➔ 100 fois moins de paramètres
➔ Efficacité quasi-similaire
➔ Nombre d'opérations à l'inférence légèrement supérieur

$$\text{F}_1\text{-score}_i = 2\frac{\text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i}$$

Ce travail a abouti à une publication scientifique : Antoine Gauquier, Pierre Senellart. Automatically Inferring the Document Class of a Scientific Article. *DocEng 2023 - 23rd ACM Symposium on Document Engineering*, Aug 2023, Limerick, Ireland.

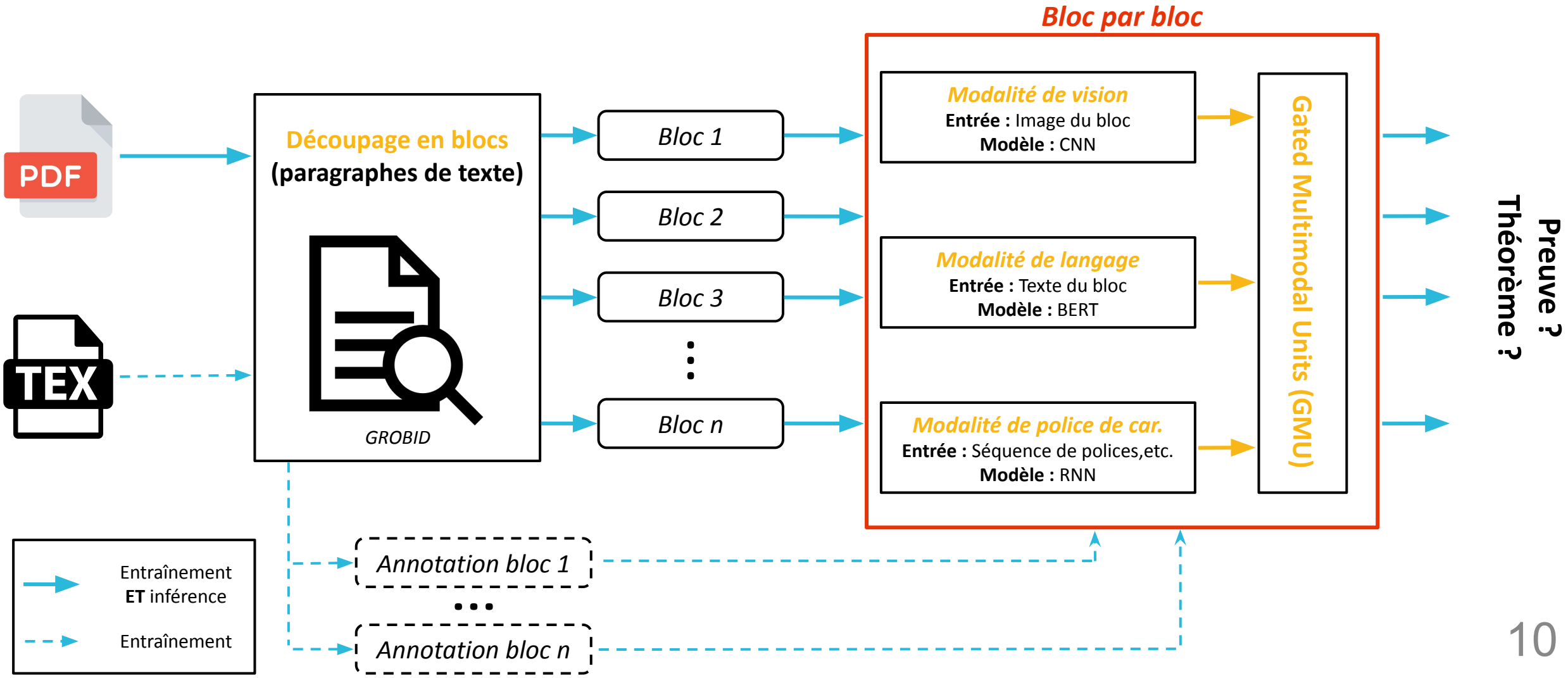
IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

ENS | PSL

2. Extraction des environnements mathématiques dans la littérature scientifique

2. Extraction des environnements mathématiques dans la littérature scientifique

Présentation du modèle multimodal (travail de Shrey Mishra)



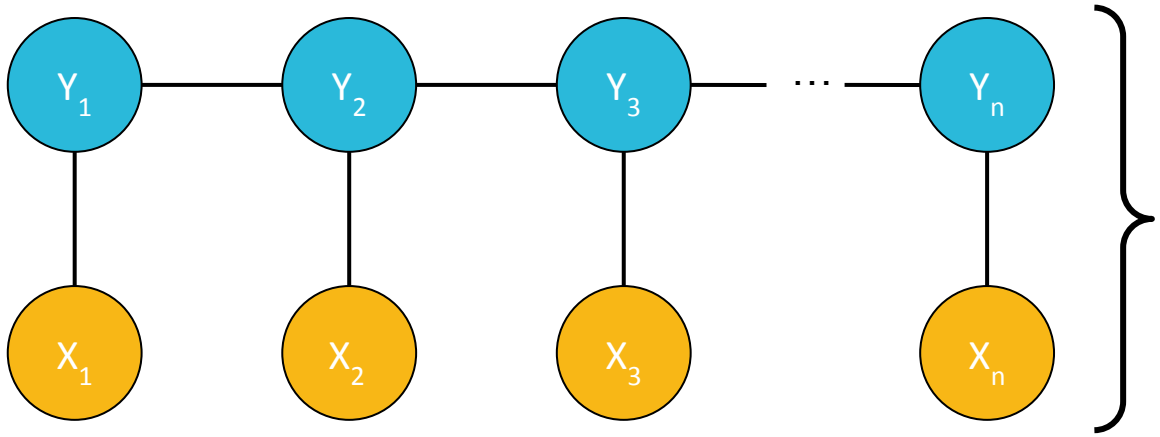
2. Extraction des environnements mathématiques dans la littérature scientifique

Approche séquentielle (ma contribution)

Modélisation probabiliste : champs conditionnels aléatoires (Conditional Random Fields – **CRF**)

Types d'environnement des blocs

Observations (embeddings + 4 variables)



Propriété de Markov

Linear-chained CRF

Observations X selon les modélisations

Modalité de vision	Embeddings CNN	1 280 variables
Modalité textuelle	Embeddings BERT	768 variables
Modalité de « font »	Embeddings RNN	128 variables
Approche multimodale	Embeddings GMU	768 variables

\+ 4 variables relatives au séquencement des blocs

2. Extraction des environnements mathématiques dans la littérature scientifique

Résultats obtenus

Modélisation	F1-Score moyen (macro)		Précision (micro)	
	Baseline	*CRF*	*Baseline*	*CRF*
Modalité de vision	60.33 %	**69.82 %**	69.43 %	**74.13 %**
Modalité textuelle	72.33 %	**80.52 %**	76.45 %	**82.70 %**
Modalité de « font »	45.00 %	**50.49 %**	**65.00 %**	52.20 %
Approche multimodale	73.87 %	**83.01 %**	76.86 %	**84.38 %**

Conclusion :

➔ L'approche séquentielle **améliore** (presque) toutes les **modélisations**

➔ Elle augmente la **robustesse** du modèle (capacité à **détecter les classes d'intérêt**)

➔ Amélioration significative, malgré une **approche simple**

Ce travail a abouti à la soumission d'un article : Shrey Mishra, Antoine Gauquier, and Pierre Senellart. Multimodal machine learning for extraction of theorems and proofs in the scientific literature. *Under review*, May 2023.

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

ENS | PSL

3. Étude de l'impact de la classe de document dans cette tâche d'extraction

3. Étude de l'impact de la classe de document dans cette tâche d'extraction

Fusion des modèles

Modèle architecture GMU
Tâche : *Classification des blocs de texte*

?

Modèle architecture CNN
Tâche : *Inférence de la classe de document*

Preuve ? Théorème ?

Architecture GMU

- ➔ ***Variables :*** Embeddings de dimension 768 en sortie du modèle
- ➔ Varie **d'un bloc à l'autre**

Architecture CNN

- ➔ ***Variables :*** Embeddings de dimension 288 en sortie du modèle **OU** « one-hot encoding » de la classe de document prédite
- ➔ **Constant au sein d'un même PDF**

3. Étude de l'impact de la classe de document dans cette tâche d'extraction

Architectures (1)

Bloc par bloc

Embeddings GMU
768 variables

Embeddings CNN
288 variables

« One-hot encoding » des classes de document prédites
33 variables

Variables de séquencement
4 variables

→ **Conditional Random Fields** *(CRF)*

1060 variables d'entrée
OU
805 variables d'entrée

→ **Preuve ?**
Théorème ?

Avantages

- ➔ Poids des modèles **figés** (« frozen »)
- ➔ Entraînement **peu coûteux** (pas de GPU)

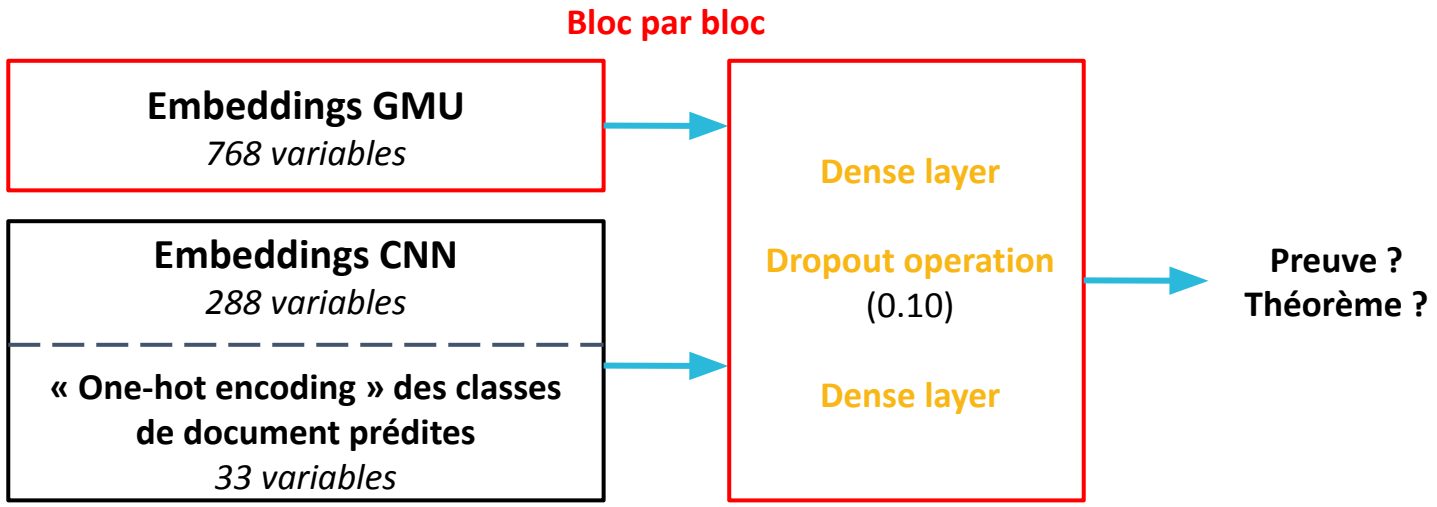
Inconvénients

- ➔ Ne capture pas l'**interaction** entre les modélisations (variables **présentées au CRF indépendamment**)
- ➔ Impossible d'identifier l'impact sur la tâche d'extraction **sans le séquencement**

Architecture	F1-Score moyen (macro)	Précision (micro)
Architecture GMU (CRF) de base	*83.01 %*	*84.38 %*
Architecture CRF à 1060 variables	**83.13 %**	84.49 %
Architecture CRF à 805 variables	83.11 %	**84.52 %**

3. Étude de l'impact de la classe de document dans cette tâche d'extraction

Architectures (2)



Avantages

- ➔ Poids des modèles **figés** (« frozen »)
- ➔ Capture *en partie* l'interaction entre les modélisations
- ➔ Est indépendant de la prise en compte du séquencement (et donc **modélisation avec séquencement également possible**)

Inconvénients

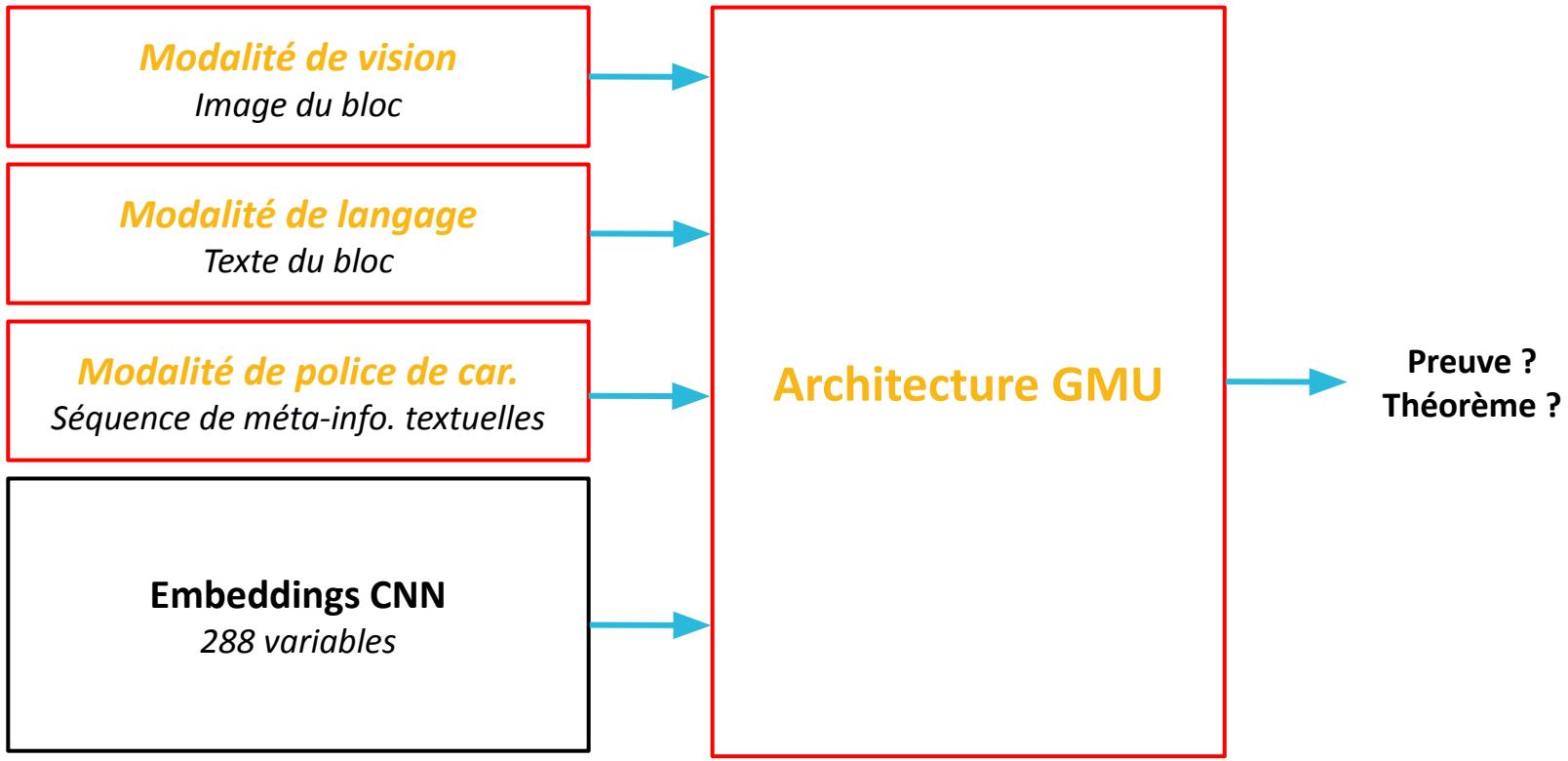
- ➔ Est *légèrement* coûteux en ressources GPU

Architecture[2]	F1-Score moyen (macro)	Précision (micro)
Architecture GMU de base	*73.87 %*	*76.86 %*
Connection embeddings GMU et CNN	76.11 %	78.85 %
Connection embeddings GMU + OHE	**76.12 %**	**78.92 %**

3. Étude de l'impact de la classe de document dans cette tâche d'extraction

Architectures (3)

Bloc par bloc



Avantages

- ➔ Capture **l'intégralité** de l'interaction entre les modélisations
- ➔ Indépendant du séquencement
- ➔ En théorie, la **modélisation la plus complète**, mais …

Inconvénients

- ➔ … aussi **la plus coûteuse** en GPU et en temps (ré-entraînement complet d'une architecture complexe)
- ➔ Question de sens sur la **fusion d'embeddings représentant des informations différentes**

Résultats ?

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

ENS | PSL

4. Tâches encore en cours et futur de ce travail (conclusion)

4. Tâches encore en cours et futur de ce travail (conclusion)

Tâches restantes avant la fin du stage

- ➔ Finalisation des **expérimentations restantes**.
- ➔ **Rédaction** du **rapport** de projet de fin d'études.
- ➔ Aide à la **préparation du Git** sur les travaux de la partie 2 et d'une **démonstration du système** (deadline début août).

Futur de ce travail et conclusion

- ➔ Quelques pistes d'amélioration à explorer : **changement** de **modélisation du séquencement** (avec du **deep learning** : Bi-LSTM, auto-encodeur : résoudrait potentiellement la redondance des informations de classe de document). **Apports des parties 1 et 2 très importants** en ce sens, même si les résultats de la partie 3 peuvent paraître « faibles ».
- ➔ Performances assez hautes : suffisamment robuste pour commencer à **extraire le contenu** des environnements mathématiques identifiés (objectif raisonnable avec une des architectures présentées ci-dessus).
- ➔ **Suite logique :** peupler une **base de connaissance** avec les contenus extraits pour **passer à l'étape suivante du projet TheoremKB**.

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

ENS | PSL

5. Bilan des compétences, projet professionnel et tâches annexes au cours du stage

5. Bilan des compétences, projet professionnel et tâches annexes au cours du stage

Projet professionnel et apports personnel

- ➔ Ce stage m'a conforté dans ma volonté de poursuivre dans la recherche en informatique.
- ➔ Confirmation de mon intérêt pour l'intelligence artificielle, les sciences et la gestion de données.
- ➔ Poursuite de mes études en **thèse** à partir de septembre 2023 :
  - ◆ **Sujet :** « Construction intelligente d'un entrepôt de données multimodales et hétérogènes, avec traçabilité des données »
  - ◆ **Encadrement : Pierre Senellart** (maître de stage actuel, École normale supérieure) et **Ioana Manolescu** (maîtresse de stage de CI2, Inria, laboratoire d'informatique de l'École Polytechnique).

Compétences techniques et relationnelles développées

- ➔ *Soft skills :* **Organisation**, **autonomie** mais aussi **travail en équipe**, échange d'idées et **discussions** sur des sujets scientifiques, **équipe interculturelle**.
- ➔ Utilisation du super-calculateur Jean-Zay (IDRIS, CNRS) – **Gestion de jobs** via « Slurm », réservation de **ressources GPU**, scripts.
- ➔ Perfectionnement de **compétences de développement** : **Python** (Keras, Tensoflow, crfsuite, Pandas, …), en particulier au **format « scripts »**.
- ➔ Compétences en **rédaction scientifique** (**écriture** rigoureuse d'un **article scientifique**, présentation de résultats, **anglais technique**).

Tâches annexes

Nelly Barret, Antoine Gauquier, Jia-Jean Law, and Ioana Manolescu. PathWays: entity-focused exploration of heterogeneous data graphs. In *20th European Semantic Web Conference (ESWC 2023)*, Hersonissos (Crete), Greece, May 2023. Demonstration paper.

Nelly Barret, Antoine Gauquier, Jia-Jean Law, and Ioana Manolescu. Exploring heterogeneous data graphs through their entity paths. In *27th European Conference on Advances in Databases and Information Systems (ADBIS 2023)*, Barcelona, Spain, September 2023.

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

ENS | PSL

Merci pour votre attention !

Des questions ?