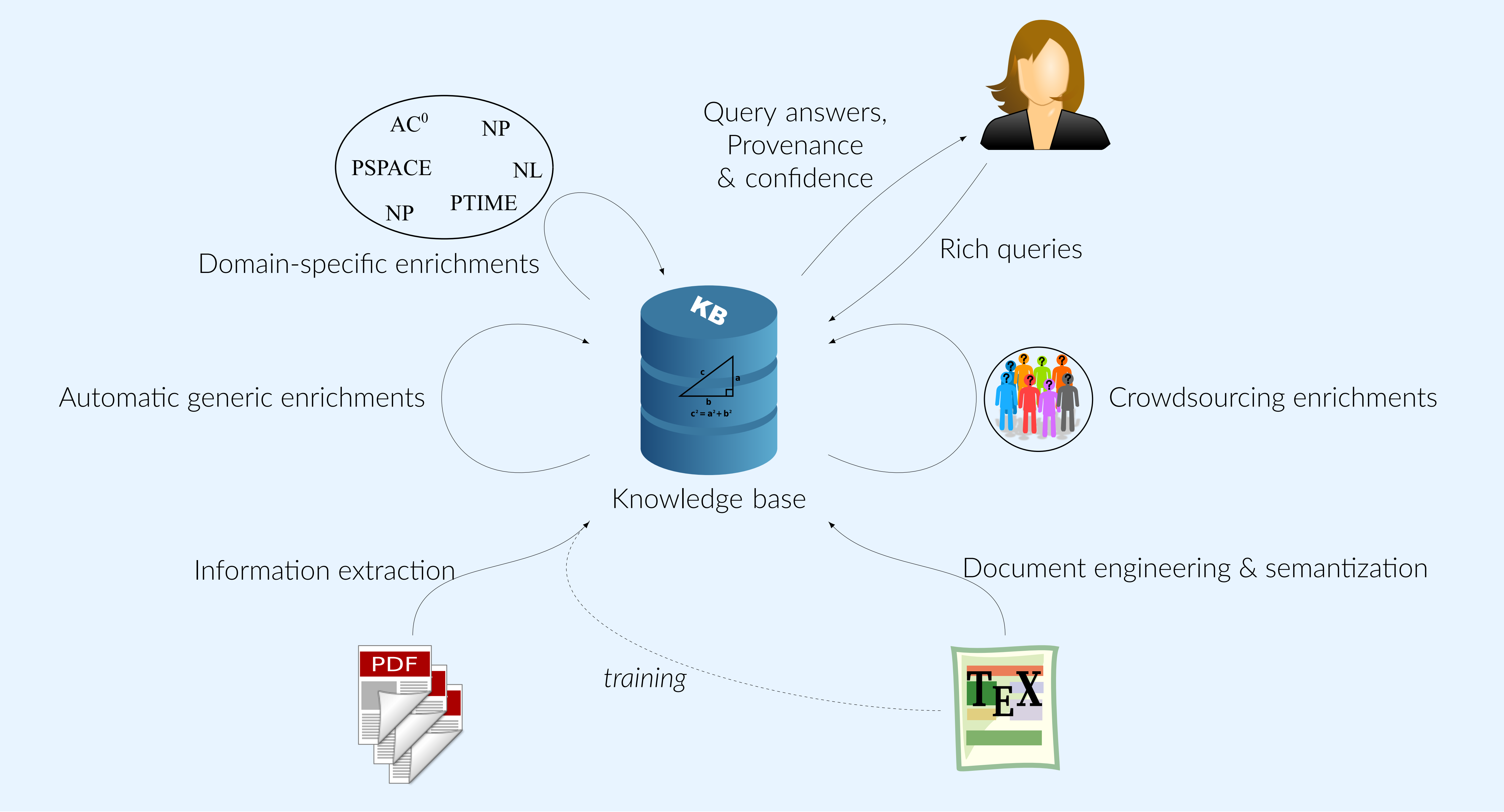
First Steps in Building a Knowledge Base of Mathematical Results

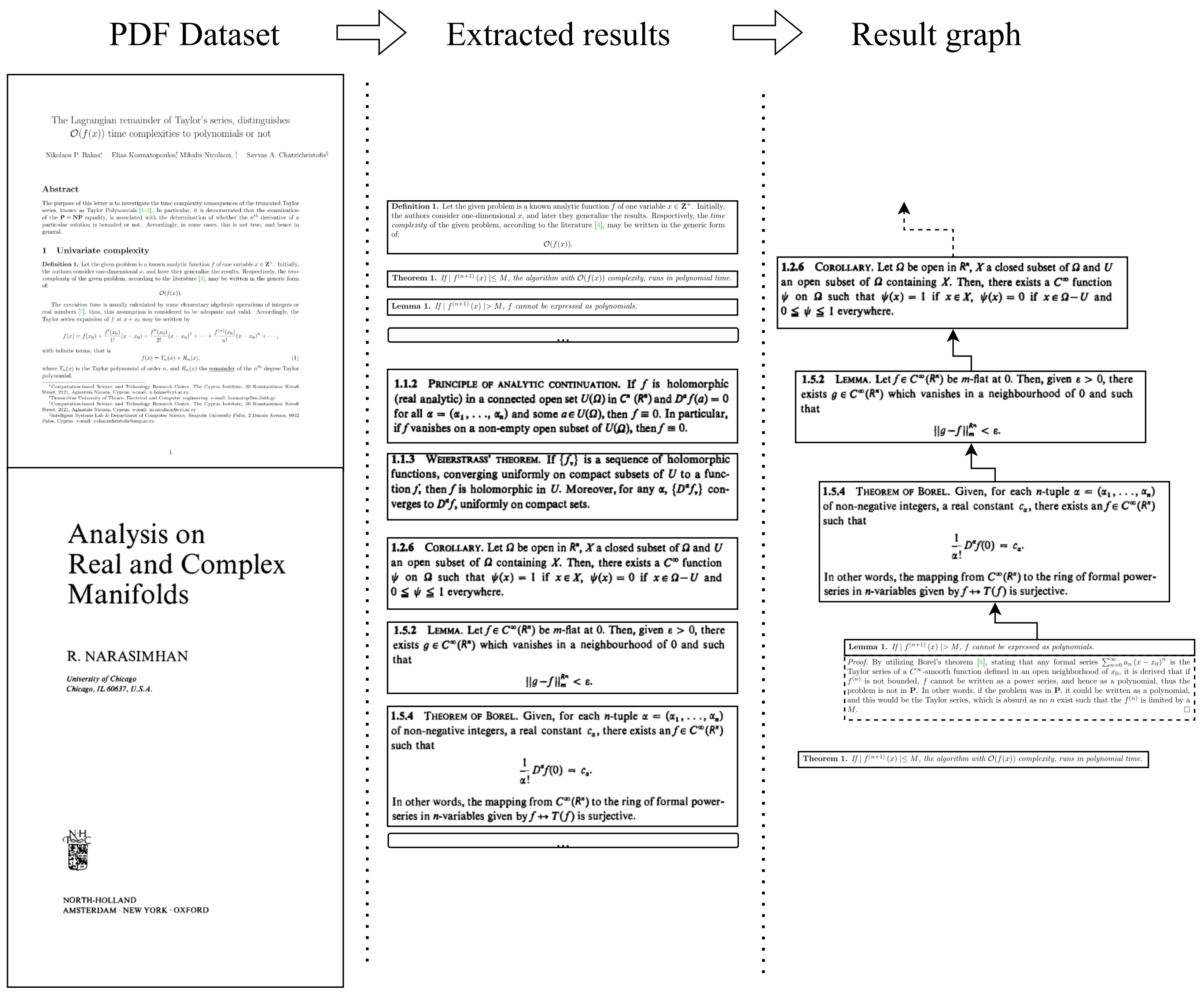
Shrey Mishra   Yacine Brihmouche   Théo Delemazure   Antoine Gauquier   **Pierre Senellart**

ENS | PSL   cnrs   Inria   CRITEO   Dauphine Université Paris | PSL   institut universitaire de France

TheoremKB Project



First steps: Extraction + Linking



Extraction: Problem & Methodology [4]

Classify whether a paragraph of text is part of a mathematical statement (theorem, definition, etc.), part of a proof, or neither (basic text).

We train (deep learning) classifiers from an automatically labeled dataset from arXiv:

Font (LSTM) Use the sequence of fonts assigned to each character of a paragraph

Vision (EfficientNet) Use the bitmap rendering of the PDF

Text (custom-trained BERT-like LM) Use the text extracted from the PDF

Multimodal (GMU) Integration of features from 3 modalities

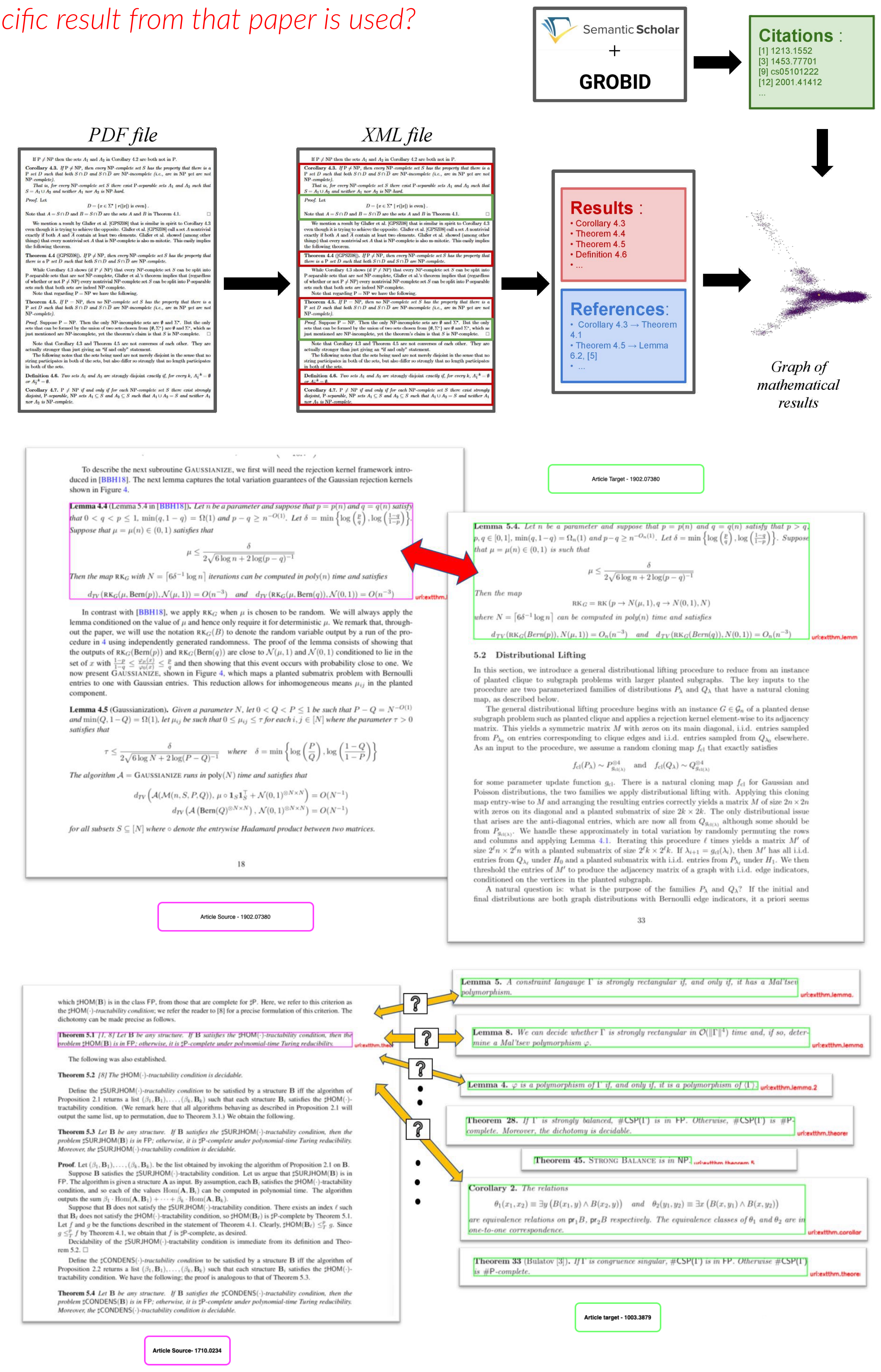
Sequence model On each unimodal and multimodal model, also take into account the sequence of labels

Extraction: Preliminary Results

Modality	Seq. approach	#Batches	#Params (M)	Accuracy (%)	Mean $\mathbf{F}_1$ (%)
Dummy	–	–	–	59.41	24.85
Line-based [5]	–	–	110	57.31	55.71
Font	–	11	2	64.93	45.48
	CRF	11+1	2	71.50	64.51
Vision	–	9	53	69.44	60.33
	CRF	9+1	53	74.63	70.82
Text	–	20	124	76.45	72.33
	CRF	20+1	124	83.10	80.99
Multimodal	–	10	185	76.86	73.87
	CRF	10+1	185	**84.19**	**82.91**

Linking: Problem & Methodology [2, 1]

Within the proof of a theorem, identify which result is used (and therefore which result the theorem depends on): from which paper does the result come from? which specific result from that paper is used?



References

[1] Yacine Brihmouche. TheoremKB : une base de connaissance des résultats mathématiques. Master's thesis, Paris IX Dauphine, September 2022.

[2] Theo Delemazure. A Knowledge Base of Mathematical Results. Master's thesis, Ecole Normale Supérieure (ENS), September 2020.

[3] Shrey Mishra, Yacine Brihmouche, Théo Delemazure, Antoine Gauquier, and Pierre Senellart. First Steps in Building a Knowledge Base of Mathematical Results. In *Proc. SDP*, Bangkok, Thailand, August 2024.

[4] Shrey Mishra, Antoine Gauquier, and Pierre Senellart. Multimodal machine learning for extraction of theorems and proofs in the scientific literature. *CoRR*, abs/2307.09047, 2023.

[5] Shrey Mishra, Lucas Pluvinage, and Pierre Senellart. Towards extraction of theorems and proofs in scholarly articles. In Patrick Healy, Mihai Bilauca, and Alexandra Bonnici, editors, *DocEng '21: ACM Symposium on Document Engineering 2021, Limerick, Ireland, August 24-27, 2021*, pages 25:1–25:4. ACM, 2021.