LLM-based Entity Extraction Is Not for Cybersecurity

Maxime Würsch, EPFL & CYD Campus

EEKE Workshop

26.06.2023

Context

- Software present everywhere
  - Contains vulnerabilities
- Technologies evolves fast
  - Increase security gap
- Need to stay up to date
  - Reduce risk of attack

- Common approach is bibliometrics search
  - Using entities extraction
  - Comparing through embedding space
- Emergence of LLM-based entities extraction
  - Need evaluation about performance

Aim

- Measure performance of entities extractors
  - Compare between LLM-based and not
  - Similitudes and differences between models

- Relevance of the method to classify documents

- Are LLM-based entity extractors suited for scientific bibliometrics ?

Large Language Model (LLM)

- Attention since late 2022, with conversational agent's public trials
- Term come from ELMo LLM in 2018
  - At least 100M parameters and 1B tokens
- Now goes to more than trillions of parameters
- Small LLMs
  - Less resources demanding
  - Weaker version of larger model
  - Reduce version of failure modes

Methods (dataset information)

- Comes from arXiv, until late 2022, Computer Science (cs) category
- Subset of the cs category
- Keep only English text
    - XLM-RoBERTa model
- Remove preamble and references

arXiv selected cs listings

cs.CR	Cryptography and Security
cs.NI	Networking and Internet Architecture
cs.CC	Computational Complexity
cs.LO	Logic in Computer Science
cs.DS	Data Structures and Algorithms
cs.IT	Information Theory
cs.CL	Computation and Language
cs.AI	Artificial Intelligence

Methods (models)

- 4 major types
- Document segmented to fully fit the attention windows
- At most 100 entities extracted
  - Select by highest confidence

Model Name	Refs	Entities/Doc	Type
spaCy Large*^{p}	[17]	99.3 ± 6.93	Noun Extractor
spaCy Transformer^p	[17]	99.3 ± 6.97	
Yake*^{p}	[5]	19.9 ± 1.97	Keyphrase Extractor
KeyBERT^p	[15]	99.3 ± 7.25	
KBIR kpcrowd	[23, 25]	96.9 ± 14.6	
KBIR inspec	[23, 37]	76.4 ± 27.7	
BERT-base-uncased	[11]	44.7 ± 24.0	
BERT-base-uncased	[11]	43.3 ± 23.3	NER+CON R
XLM-RoBERTa-base Onconotes 5	[40, 18]	36.4 ± 23.4	NER+NUM
ELECTRA-base conll03	[8, 38]	39.9 ± 25.0	NER
BERT-large-cased conll03	[11, 38]	41.7 ± 24.9	
BERT-large-uncased conll03	[11, 38]	33.5 ± 23.3	
DistilBERT-base-uncased conll03	[39, 38]	37.7 ± 24.8	
RoBERTa-large conll03	[24, 38]	28.7 ± 21.1	
XLM-RoBERTa-large conll03	[14, 38]	26.0 ± 19.5	
BERT COCA-docusco	[11, 20]	99.6 ± 6.11	TokC

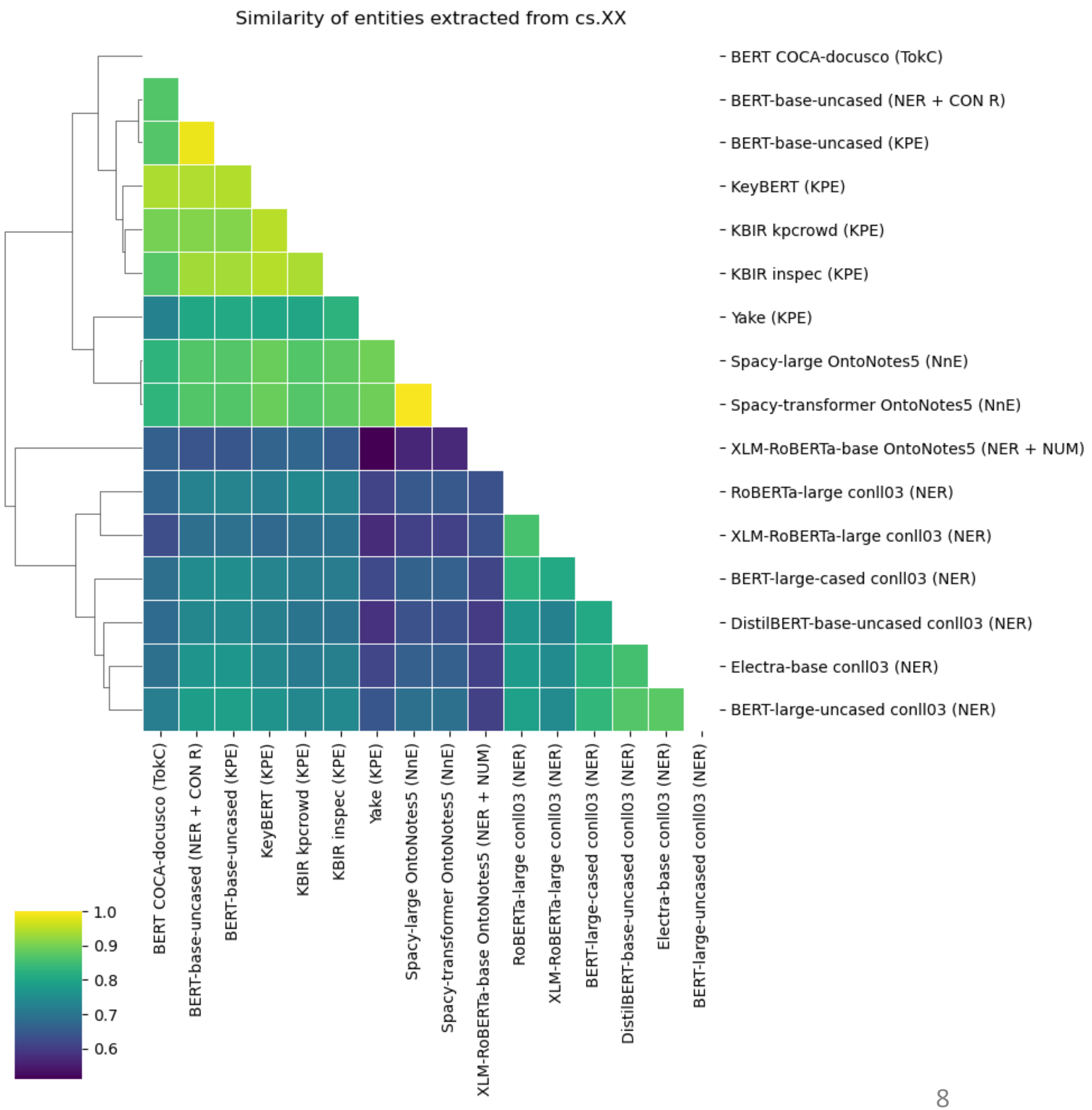
Methods (visualisations)

- Hierarchical clustering
  - Embedded with SpaCy
  - Average cosine distance
  - Identify similitude between extractor

- 2D Projection
  - Subsample data: reduce processing time and number of point
  - 6 embeddings:
    - SpaCy, GloVe, Fasttext, Word2Vec, BERT-large, GPT-2
  - 4 low-dimensional projection
    - Linear, spectral, t-SNE, UMAP
  - Show if themes can be detected in an unsupervised way

Results and Discussion

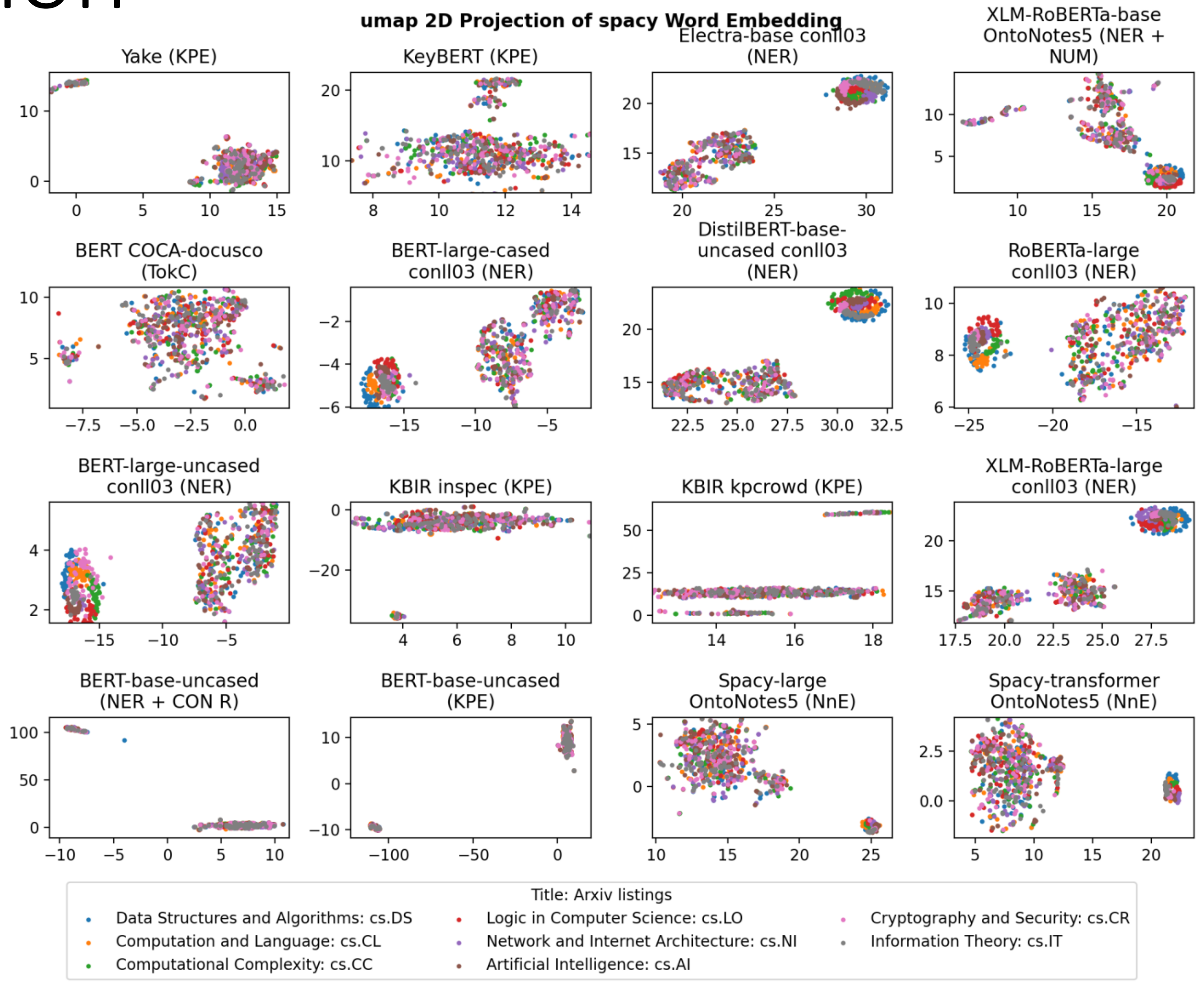
- Performance manly define by architecture and fine-tuned dataset
- Dataset not based on scientific texts
  - Conll03
- => Not suited for scientific articles



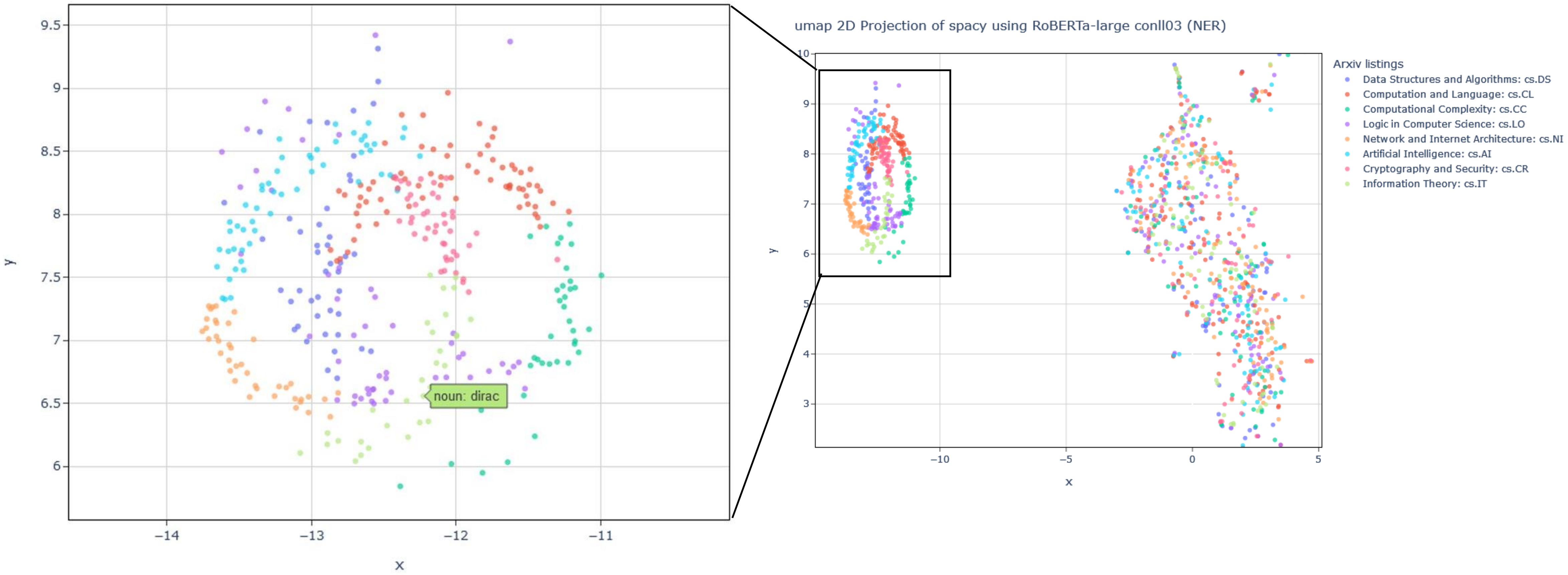
Results and Discussion

- Cosine similarity of embedding do not perform well to cluster themes
  - Even with 2D embedding algorithm that tend to overfit

- Exception with NER

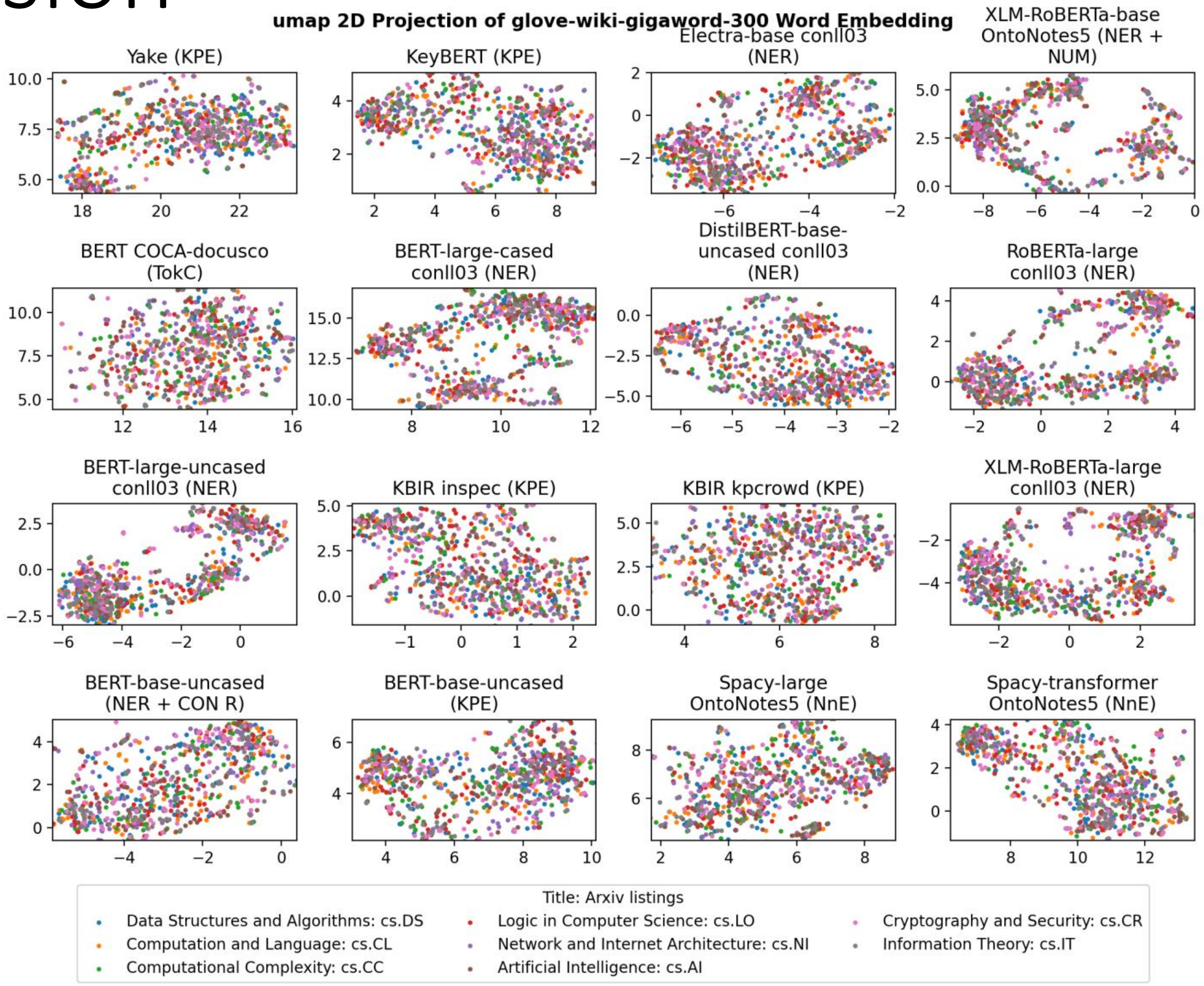


Umap projection of Spacy using RoBERTa-large conll03 (NER)



Results and Discussion

- Cosine similarity highly dependent of embedding space
- Important change with different embedding and algorithm



Conclusion

- LLM-based entity extraction seems not suited for concept-oriented bibliometrics in scientific article
- Work only on arXiv cs category
- Nouns extraction seems more robust

Maxime **Würsch**
CYD Intern

Andrei **Kucharavy**
Research associate UAS

Dimitri **Percia David**
Assistant professor UAS

Alain **Mermoud**
Head of Technology Monitoring Team

Thanks for your attention !

Questions?