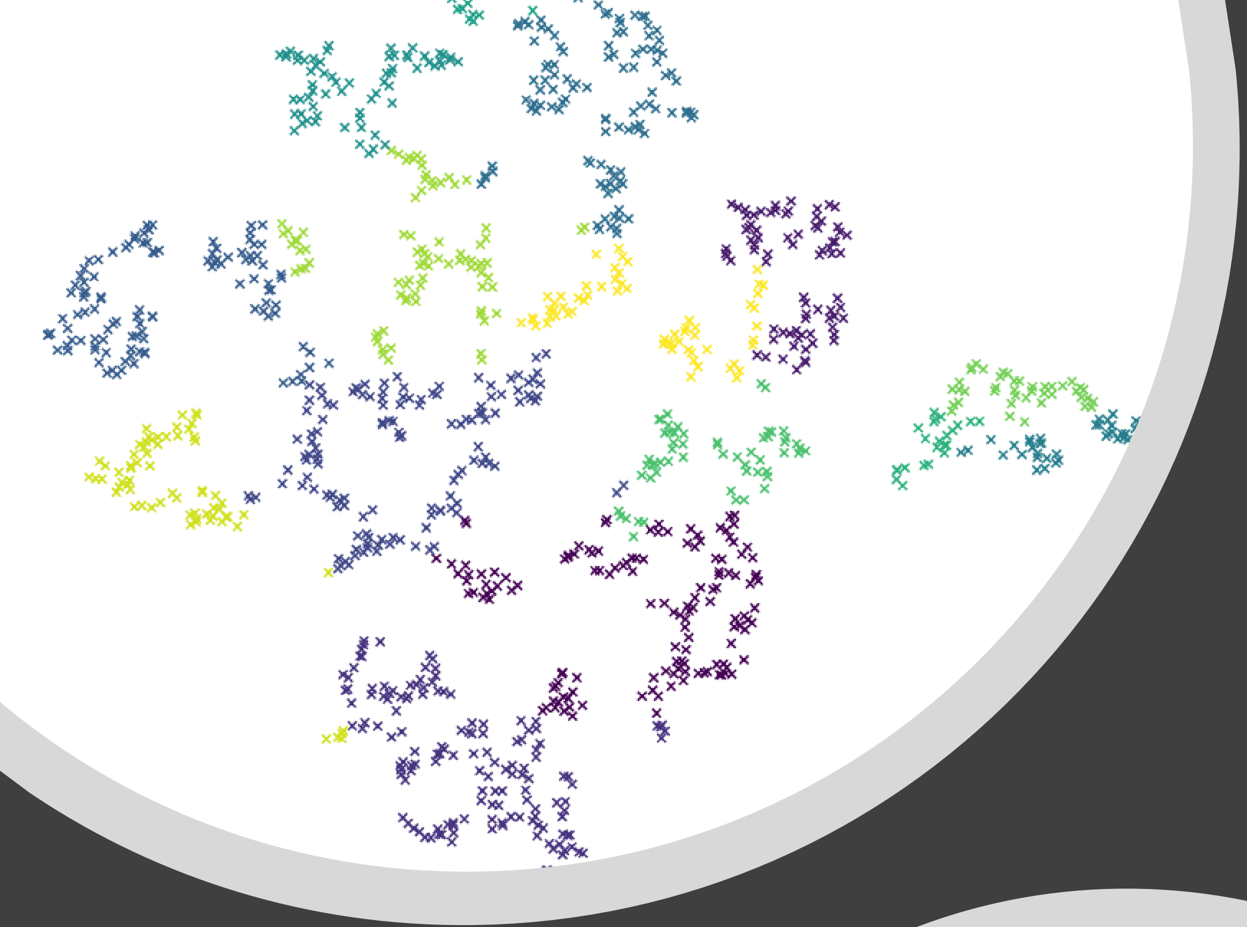

Scientific paper clustering – ITA WS 20/21

Daniela Fichiu, Christian Homeyer, Jessica Kächele, Jonas Reinwald

THERE ARE TOO MANY PAPERS!

- How to keep track of the flood of papers?
- How to organize this in a better way?

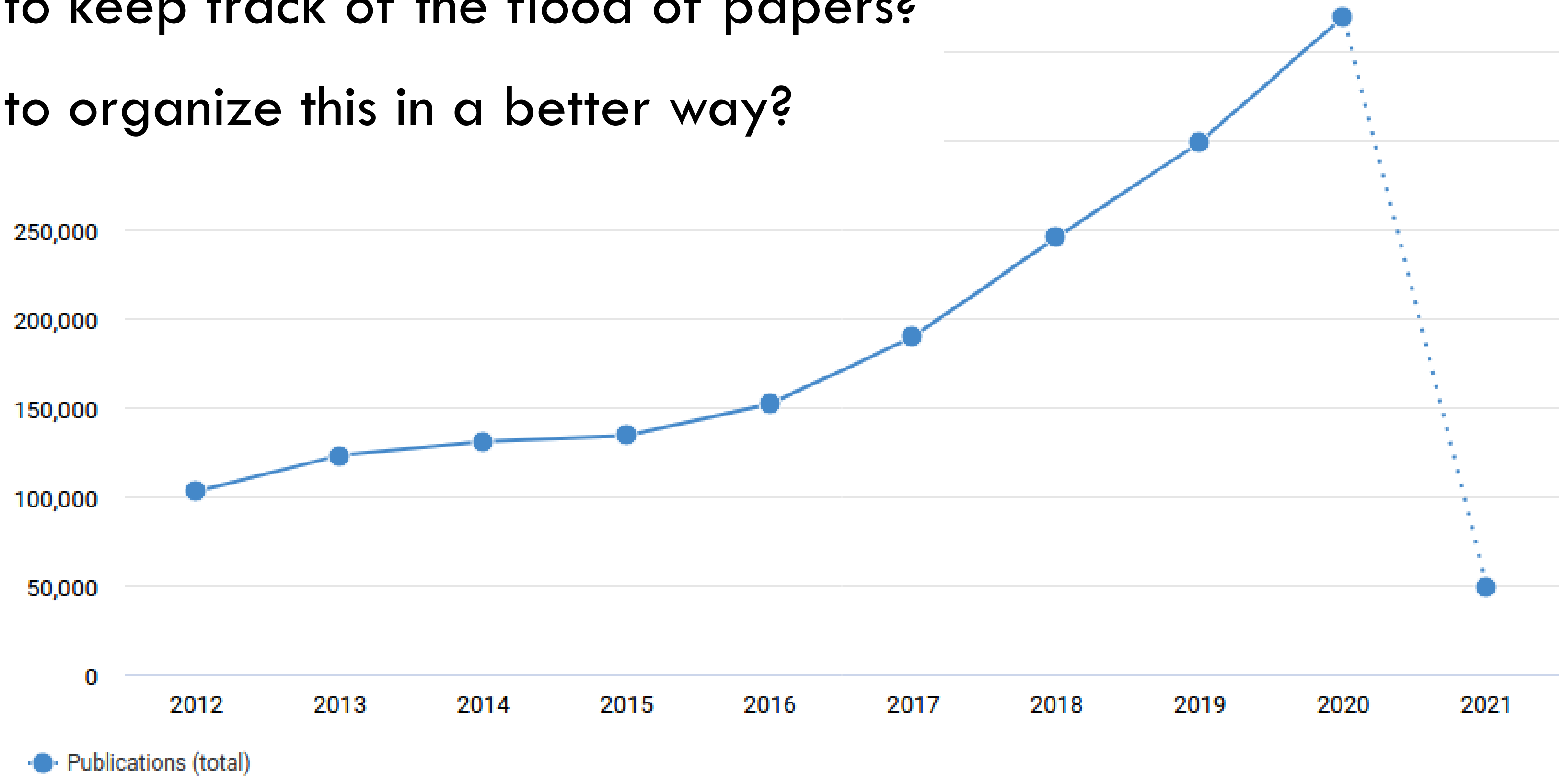

Fig. 1: Publications with key word „machine learning“, source: apps.dimensions.ai

IDEA

- Every paper is submitted with a standardized abstract and a set of keywords

- Use clustering algorithm on abstracts to group papers

- Exploit keywords to create ground-truth / supervision

PIPELINE

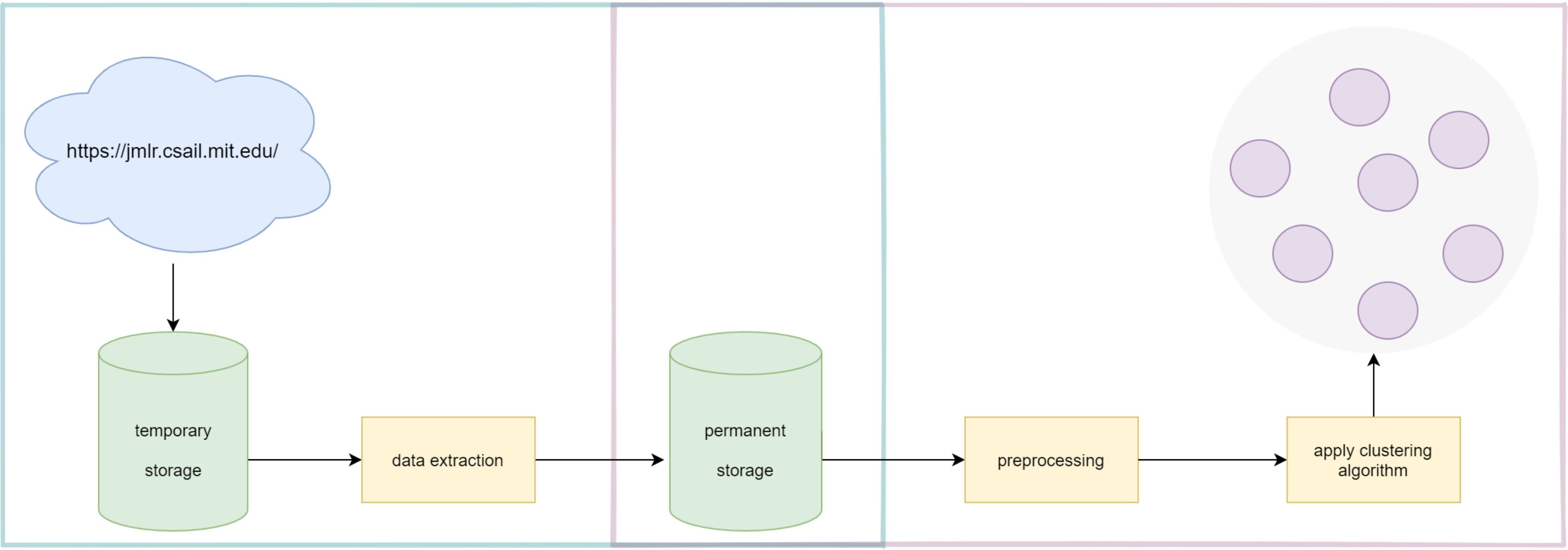

Fig. 2: Overview of the project pipeline

DATA SOURCE

Information from a research paper we were interested in

- Keywords → Evaluation
- Abstract → Cluster
- Title
- Authors
- Link to paper (Ref)

Statistics

- Journal of Machine Learning: peer-reviewed open access scientific journal covering machine learning
- Papers organized in 21 (currently 22) volumes - between 60 and 250 papers/volume

DATA SCRAPING PIPELINE

- Initial Solution: scrape with script (http requests) and then extract information using Regex → Problem: Find one regex to suit them all

- Solution: Grobid - machine learning software for extracting information from scholarly documents

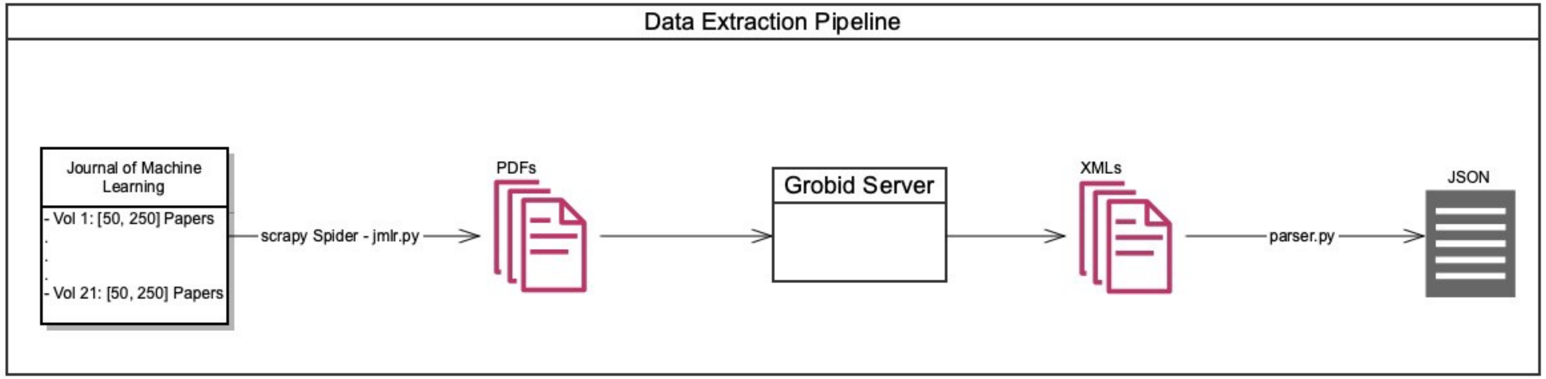

Fig. 2: Complete pipeline from online paper source to extracted information in well-structured json files

DATASET

- 2230 Research Papers
- 20 without an abstract
- 159 without no keywords
- Average no. of keywords per research paper is ~4.70

Journal of Machine Learning Research 1 (2000) 49-75 Submitted 5/00; Published 10/00

Dependency Networks for Inference, Collaborative Filtering, and Data Visualization

David Heckerman HECKERMA@MICROSOFT.COM
David Maxwell Chickering DMAX@MICROSOFT.COM
Christopher Meek MEEK@MICROSOFT.COM
Robert Rounthwaite ROBERTRO@MICROSOFT.COM
Carl Kadie CARLK@MICROSOFT.COM
Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA

Editor: Leslie Pack Kaelbling

Abstract

We describe a graphical model for probabilistic relationships—an alternative to the Bayesian network—called a dependency network. The graph of a dependency network, unlike a Bayesian network, is potentially cyclic. The probability component of a dependency network, like a Bayesian network, is a set of conditional distributions, one for each node given its parents. We identify several basic properties of this representation and describe a computationally efficient procedure for learning the graph and probability components from data. We describe the application of this representation to probabilistic inference, collaborative filtering (the task of predicting preferences), and the visualization of acausal predictive relationships.

Keywords: Dependency networks, Bayesian networks, graphical models, probabilistic inference, data visualization, exploratory data analysis, collaborative filtering, Gibbs sampling

Fig. 3: Example of a typical paper and its structure, source: https://www.jmlr.org/papers/volume1/heckerman00a/heckerman00a.pdf

CLUSTERING

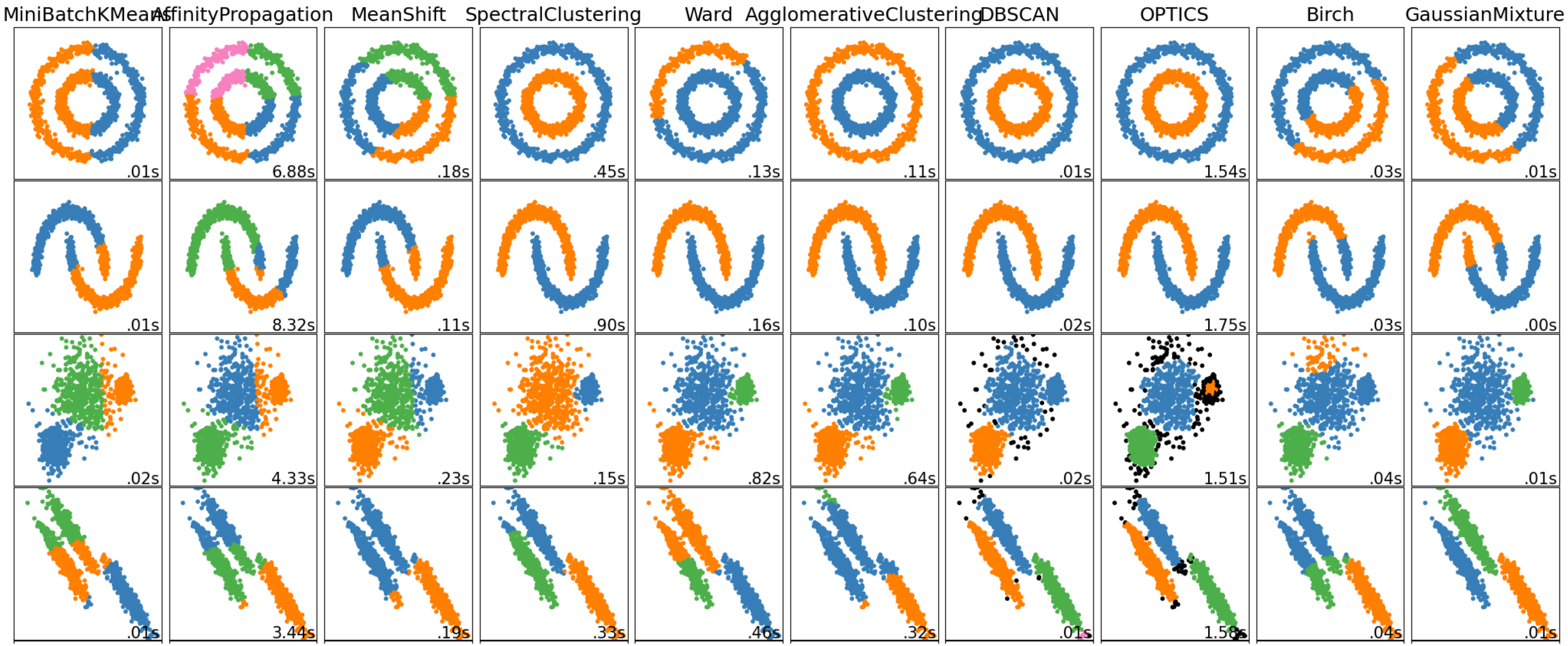

Fig. 4: Different clustering methods from sklearn, source: https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods

EVALUATION

Alternatives to ground-truth evaluation

- Silhouette Coefficient
- Calinski-Harabasz Index
- Davies Bouldin Index

Create ground-truth from keywords

- Abstract can contain nonrelevant information
- Keywords are (hopefully) distilled truth
  - they describe the paper contents in the least number of words

EVALUATION (GROUND-TRUTH)

Process

- Split the keywords into words and preprocess them
  - Example: learning to rank, Bayesian inference, neural networks → learn, rank, bayesian, inference, neural, network
- Create bag-of-words corpus
- Cluster using DBSCAN
  - Try different thresholds for DBSCAN parameter “eps”
- Manually check clustering results

Problems

- We have a very unbalanced dataset
- Ground-truth itself is biased
- Task is hard → results are not ideal / objective

RESULTS

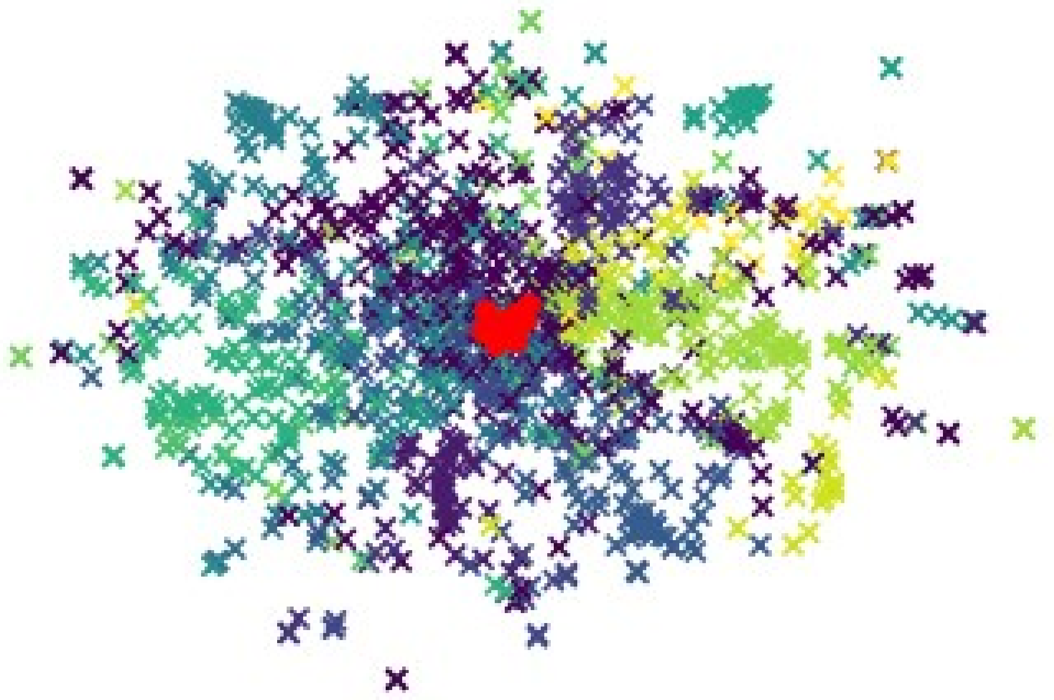

Fig. 5: Clustering visualization with KMeans

Algorithm	Silhouette Score
KMeans	**0.01027**
Spectral Clustering	0.01166
Gaussian Mixture	0.0115
Birch	0.00165
Affinity Propagation	0.03049
Agglomerative Clustering	0.00047
DBSCAN	-0.00971
OPTICS	-0.00323

Table 1: Silhouette scores of different cluster algorithms

RESULTS (DIMENSIONALITY REDUCTION)

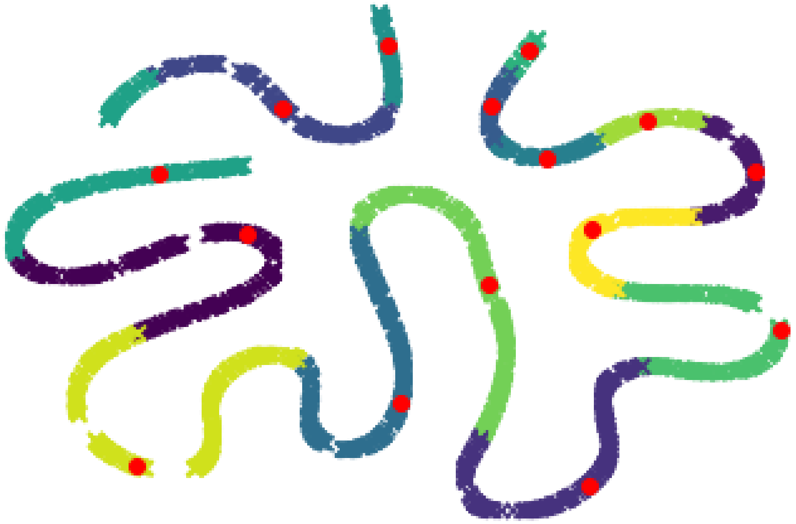

Fig. 6: Clustering visualiztion with LSA and KMeans

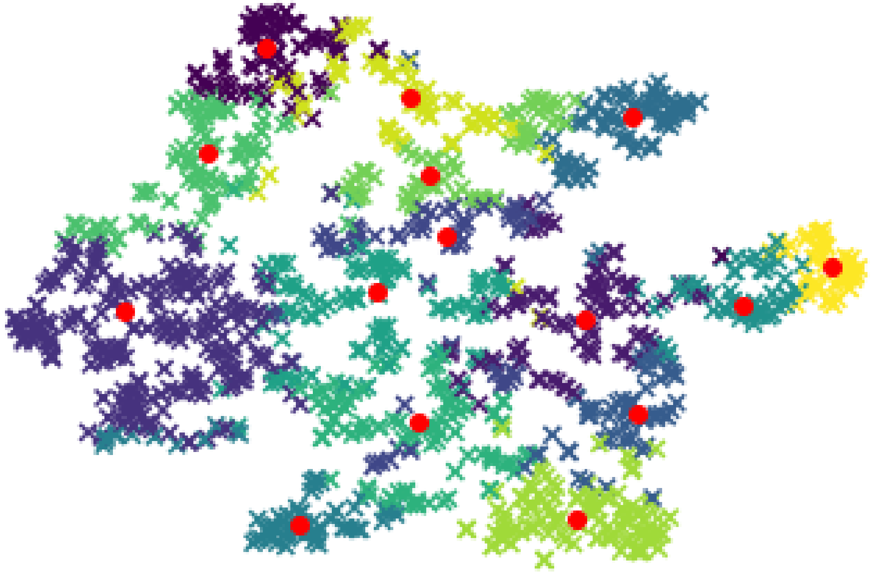

Fig. 7: Clustering visualiztion with Spectral Embedding and KMeans

KMeans	w/o dimensionality reduction	LSA	Spectral Embedding
Silhouette Score	0.010273	0.520976	0.340198

Table 2: Silhouette scores of KMeans with and without performing dimensionality reduction before clustering

RESULTS (GROUND-TRUTH)

KMeans	w/o dimensionality reduction	LSA	Spectral Embedding
purity	0.7150	0.7136	0.7123
adjusted_rand_score	0.0042	0.0054	0.0081
adjusted_mutual_info_score	0.0627	0.0619	0.0673
precision	0.7344	0.7699	0.7325
recall *	0.0050	0.0123	0.0050
f1 *	0.0047	0.0189	0.0047
accuracy *	0.0050	0.0123	0.0050

* usually only used for classification tasks, included for completeness

Table 3: Evaluation scores for KMeans from comparing ground-truth against obtained clusters

CONCLUSION

- We created a dataset of about 2300 machine learning papers
- Clustering of abstract and keywords
  - Built a scraping, processing, clustering, evaluation pipeline
  - Compared several unsupervised clustering algorithms

- This project could be useful to organize a large body of papers
  - How cool would it be to identify trends in papers?
- We could create a keyword proposal pipeline by mapping clusters to keywords