# Predicting Structured Metadata from Text

We describe a framework to **predict structured metadata values** from text in order to improve the quality and quantity of experimental metadata. We demonstrate our framework using gene expression metadata from the **GEO database**.

## Approach

We use **Latent Dirichlet Allocation (LDA)** to discover latent topics present in unstructured metadata. Topic models **drastically reduce document description length** as compared to traditional methods and they produce **semantically meaningful** features as **latent topics.**

We **compare** the performance of a **Support Vector Machine (SVM)** trained with **LDA** or **TF-IDF** against the **majority classifier**.

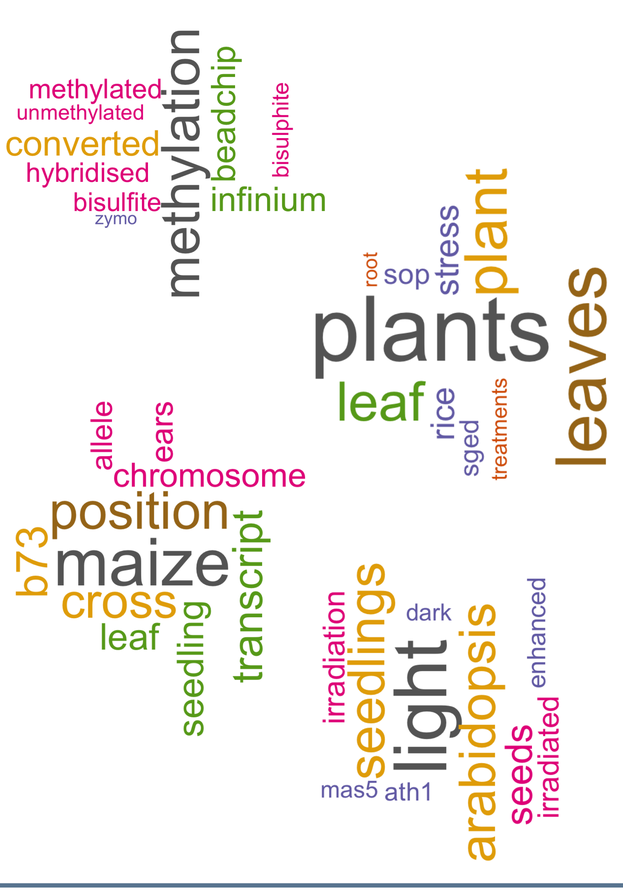
## Conclusions

Unstructured metadata elements contain information which can be successfully exploited using either LDA or TF-IDF for predicting structured metadata elements well beyond the majority classifier.

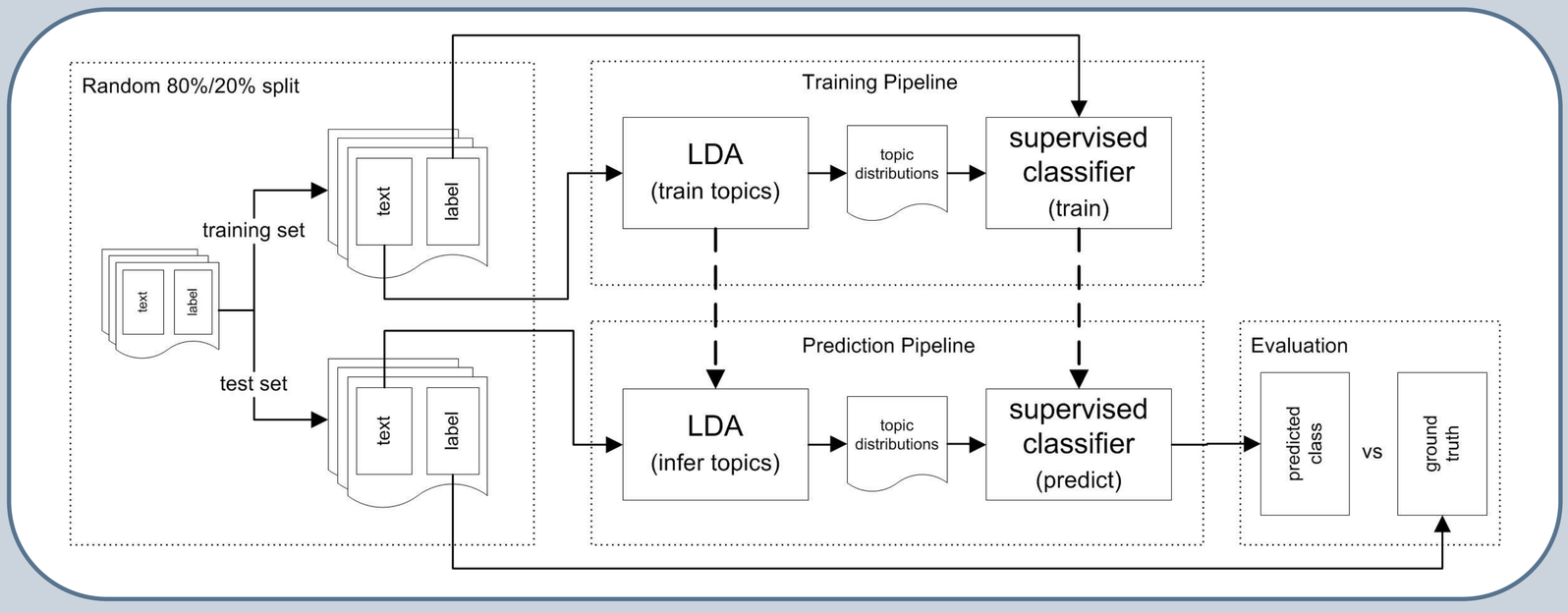
**Limitations:**

- Not all classes were predicted with equal accuracy.
- Our results are limited to a subset of structured metadata from GEO, infrequently used (<0.1% of the dataset) values are excluded.
- More work is required to understand the applicability to other metadata.
- LDA is a parametric topic model (predefined number of topics).

## GEO Sample Topics

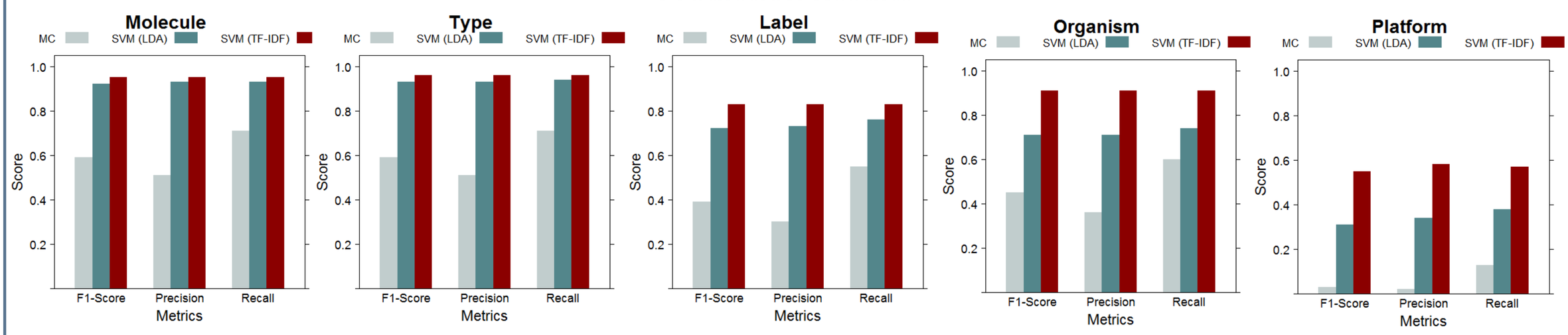



## Experimental Setup




**Experimental setup:** Preprocessing and train/test set split, training the LDA model as well as the supervised classifier, inferring the per-document topic distributions and predicting the classes, and evaluation of the predictions. The setup for the classifiers using TF-IDF features is analogous (with TF-IDF values for document representation).

## Prediction Results




Weighted class averages for precision, recall and F1-Score for each structured element. Results are reported for **linear SVM with LDA features**, **linear SVM with TF-IDF features** and for the **majority classifier (MC)** baseline.


**Contact:** Lisa Posch, Maryam Panahiazar, **Michel Dumontier**, Olivier Gevaert

**Email:** lisa.posch@gesis.org, {marypan, micheld, olivier.gevaert}@stanford.edu