# Typological Feature Prediction with Matrix Completion

University of Colorado **Boulder**

Annebeth Buis
anne.buis@colorado.edu

Mans Hulden
mans.hulden@colorado.edu

## Motivation

- **Typological features are informative to many cross-lingual tasks in NLP**. Unfortunately typological data is often sparse or incomplete and generating it is costly. For example, in the World Atlas of Language Structures (WALS) the value of 80% of language-feature combinations is undefined.

## WALS and Preprocessing

- **WALS** (World Atlas of Language Structures, Dryer and Haspelmath) contains 144 features for 2,679 languages.
- **WALS contains features in 11 domains**: phonology, sign languages, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, complex sentences, lexicon and other. Meta-features (e.g., isocodes, language family, genus, etc.) were not included in the experiments.
- **The original WALS matrix contains categorical feature values**, which were binarized before running matrix completion. We excluded 214 languages for which only 1 feature value has been recorded in WALS.
- **No additional preprocessing** or excluded features.

## Matrix Completion

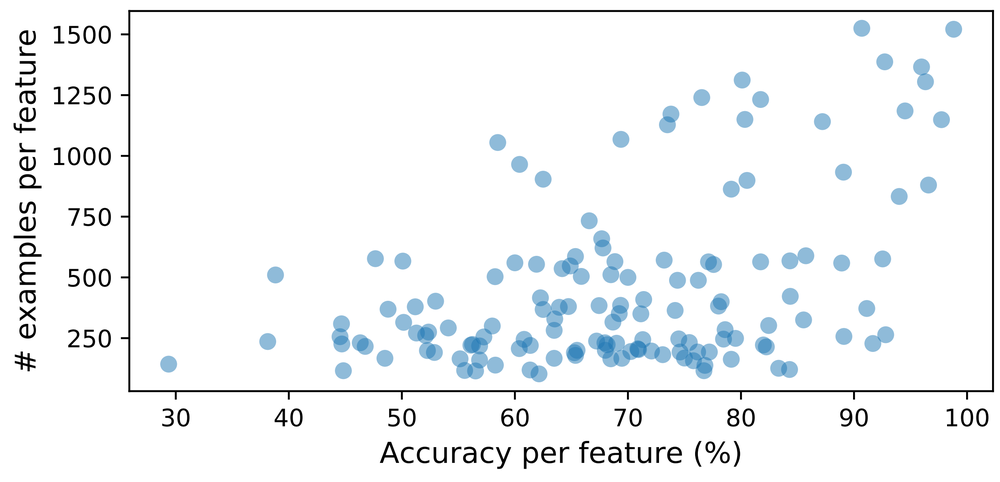
- Matrix completion methods have been used extensively with **sparse matrices** and **are able to learn more holistic patterns in the data** than individual local predictors (such as our logistic regression baseline).
- **IterativeSVD** (based on Troyanskaya et al., 2001) learns a low-rank approximation of the original matrix by using Singular Value Decomposition (SVD).

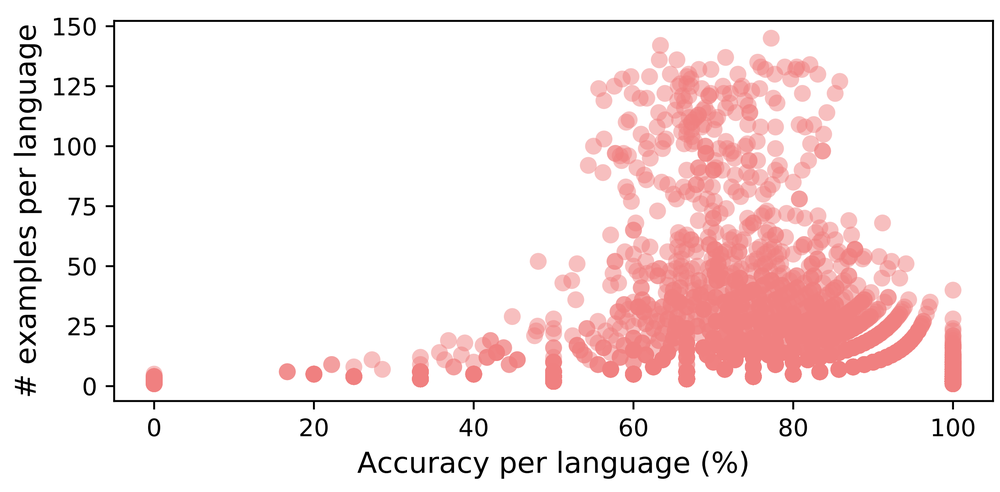
## Experiments

- **Leave-one-out-cross-validation** to predict each language × feature-combination that is currently in WALS.
- Results are compared against a majority class baseline and a logistic regression classifier.
- To test for robustness, 2 additional experiments:
  1. **Leaving out features in the same domain** (e.g., Phonology, Morphology, etc.)
  2. **Leaving out languages with the same language family** (e.g., Indo-European, Afro-Asiatic, etc.)

## Results

| | Accuracy | Method |
|---|---|---|
| Georgi et al. (2010) | 65.5% | Language clustering |
| Takamura et al. (2016) | 75.5% | Logistic regression |
| *without language family* | 73.0% | Logistic regression |
| Murawaki (2017) | 74.5% | Bayesian model |
| Baseline 1 | 53.1% | Majority class |
| Baseline 2 | 65.7% | Logistic regression |
| Matrix completion | 74.3% | IterativeSVD |
| *without domain* | 61.6% | IterativeSVD |
| *without language family* | 71.2% | IterativeSVD |




- **Prediction accuracy per language** does not improve with more examples.
- **Feature accuracy** improves when more examples are available.

## Conclusion

- **Matrix completion outperforms the baselines on the WALS data and performs on par with previous work.** Matrix completion requires minimal preprocessing and can easily be used with any typological database.
- **Our work has shown that treating WALS as a matrix is an effective approach.** This should be further explored in future work.

Data and code are available at:
`github.com/annebeth/wals-matrix-completion`.

## References

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393.

Yugo Murawaki. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 69–76.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.