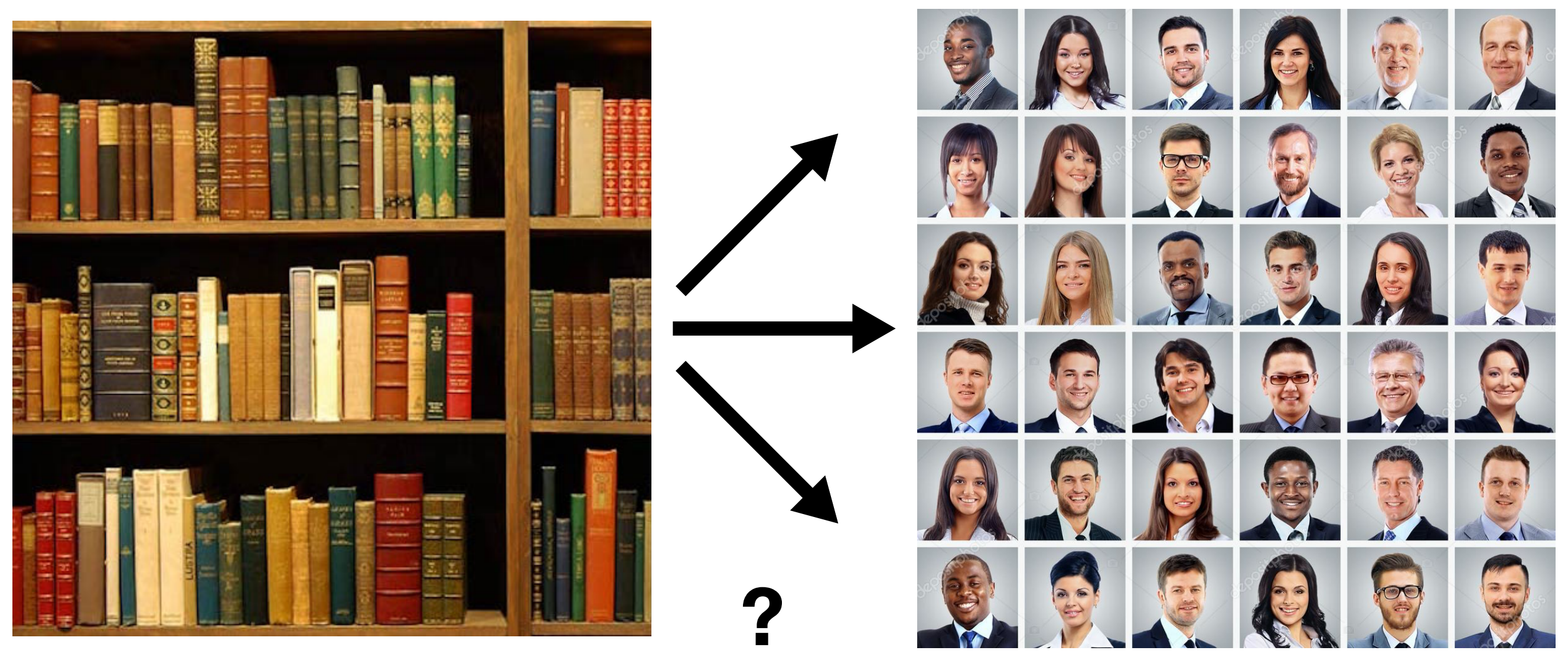
Authorship Identification with Support Vector Machines

Austin Hoover
University of Tennessee, Knoxville

Overview

- Basic task: match each document with the correct author

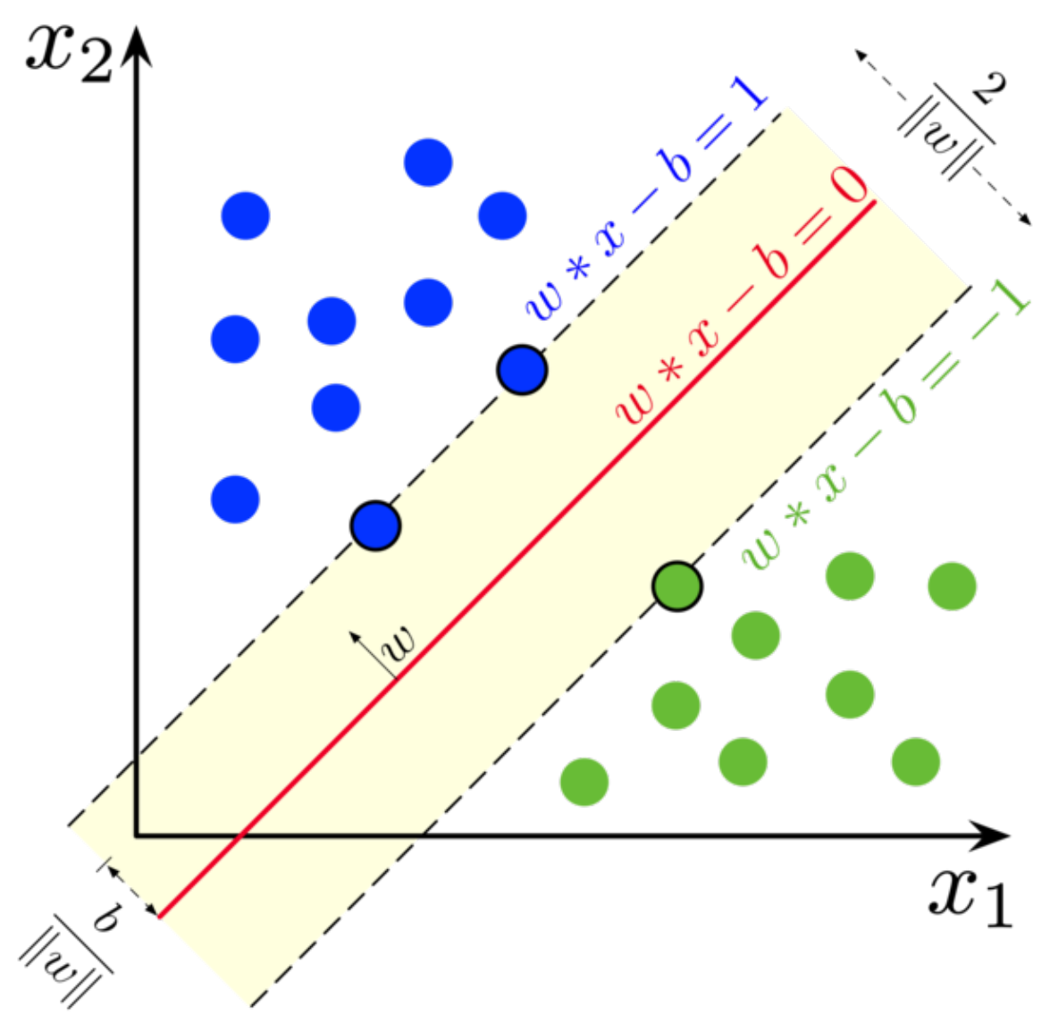


- Dataset
  - 5000 documents, 50 authors
  - Average document length < 4 paragraphs
  - All authors share common subject area
- Character n-grams

a_string $\xrightarrow{n=4}$ | a_st | _str | stri | trin | ring |

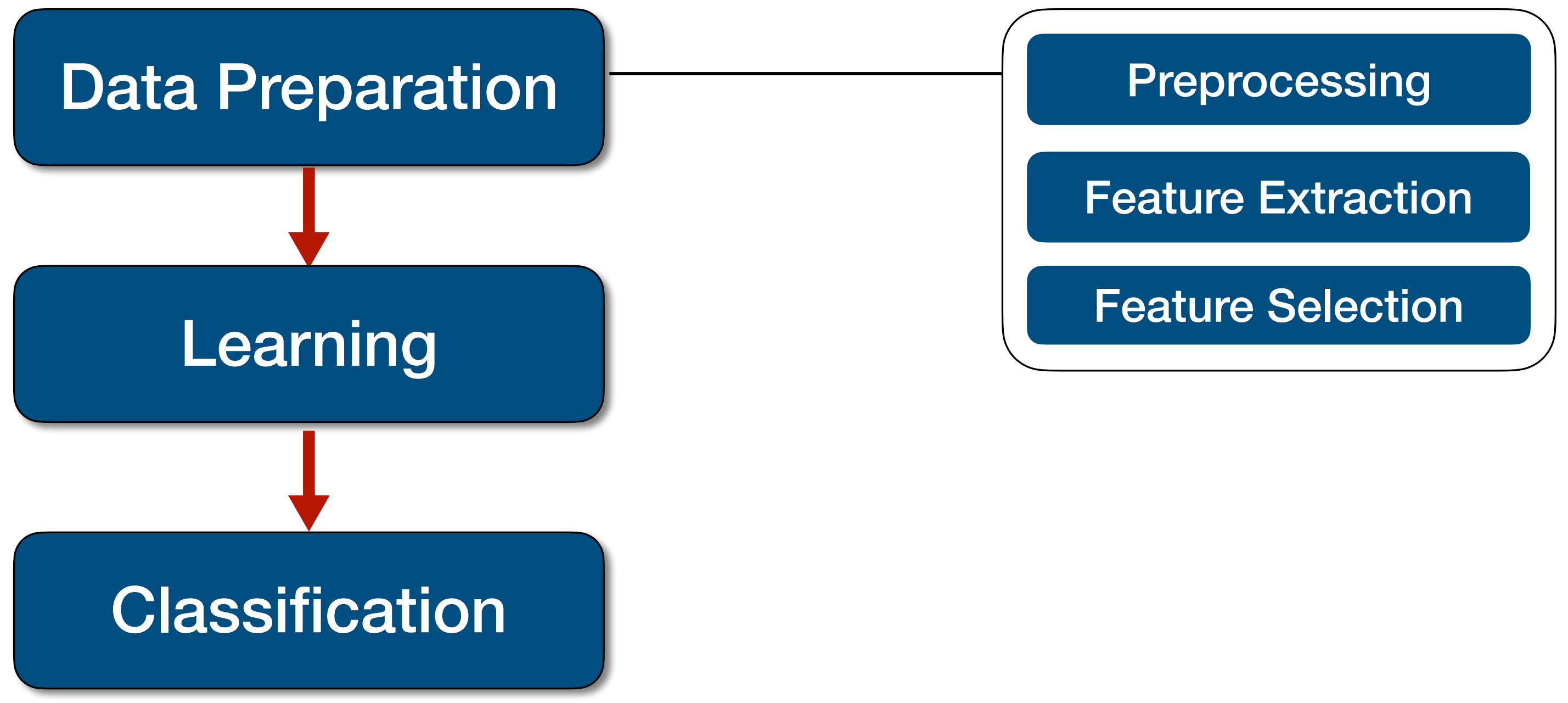
Tools

- Feature Selection Methods
  - Mutual Information — $I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \dfrac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$
  - χ^2
  - Anova F-value
- Support Vector Machine (SVM)
  - Find optimal separating plane
  - Scales to arbitrary number of dimensions



- Multi-class classification
  - *One-vs-all:* train n_{class} classifiers
  - For each document, choose class with “best” separating plane.

Method



- Which feature selection method is best?
- What is the optimal number of features?
- Which n-gram length is best: 3, 4, 5, or multi-length?
- Can the performance be improved?
  - Scaling / normalizing
  - Preprocessing text

Results

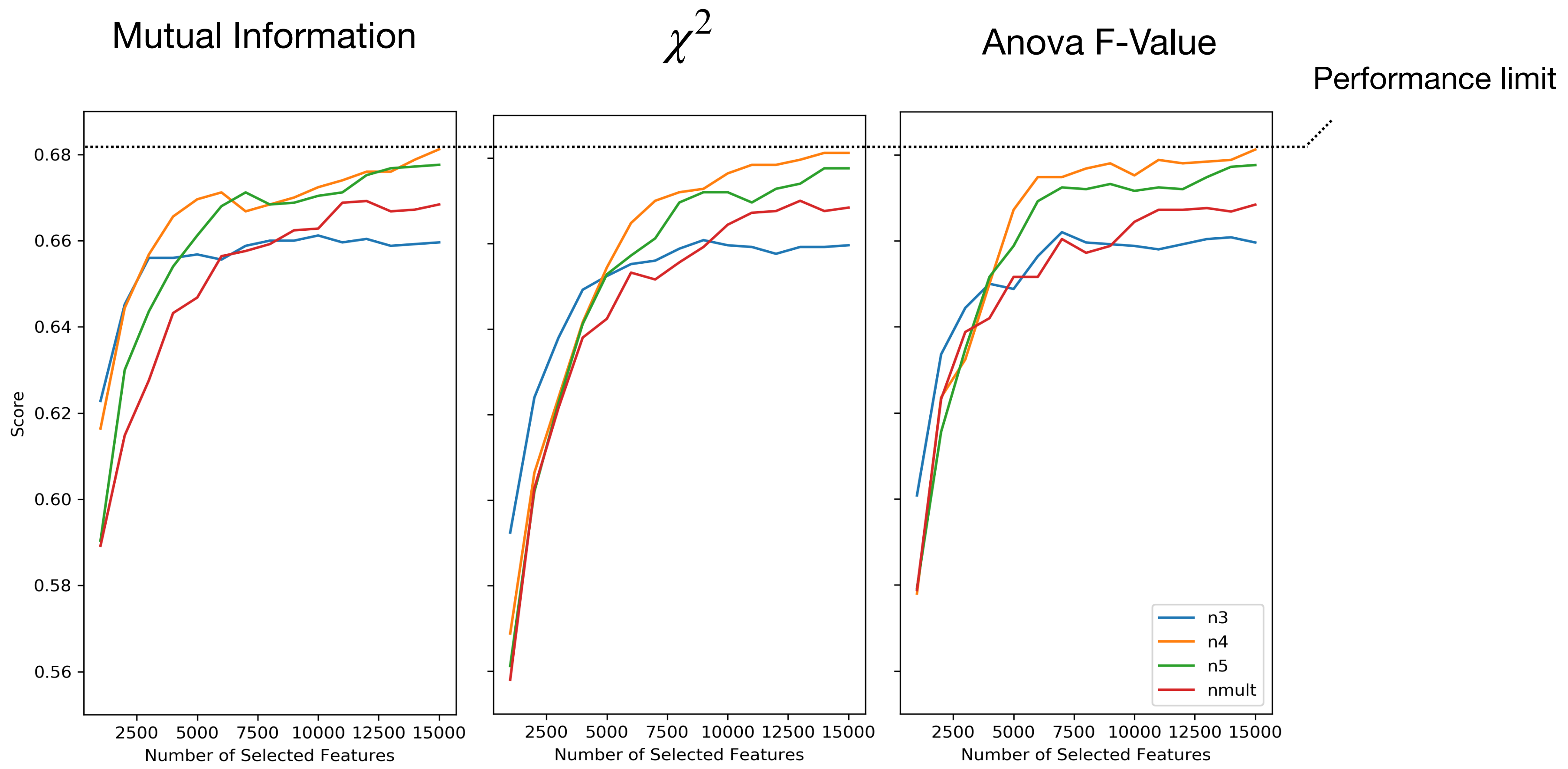


Fig 1: *Average prediction success vs number of selected features.*

Feature selection methods are compared across columns. 3-grams perform well at low feature numbers, while 4-grams are best as the dimensionality increases.

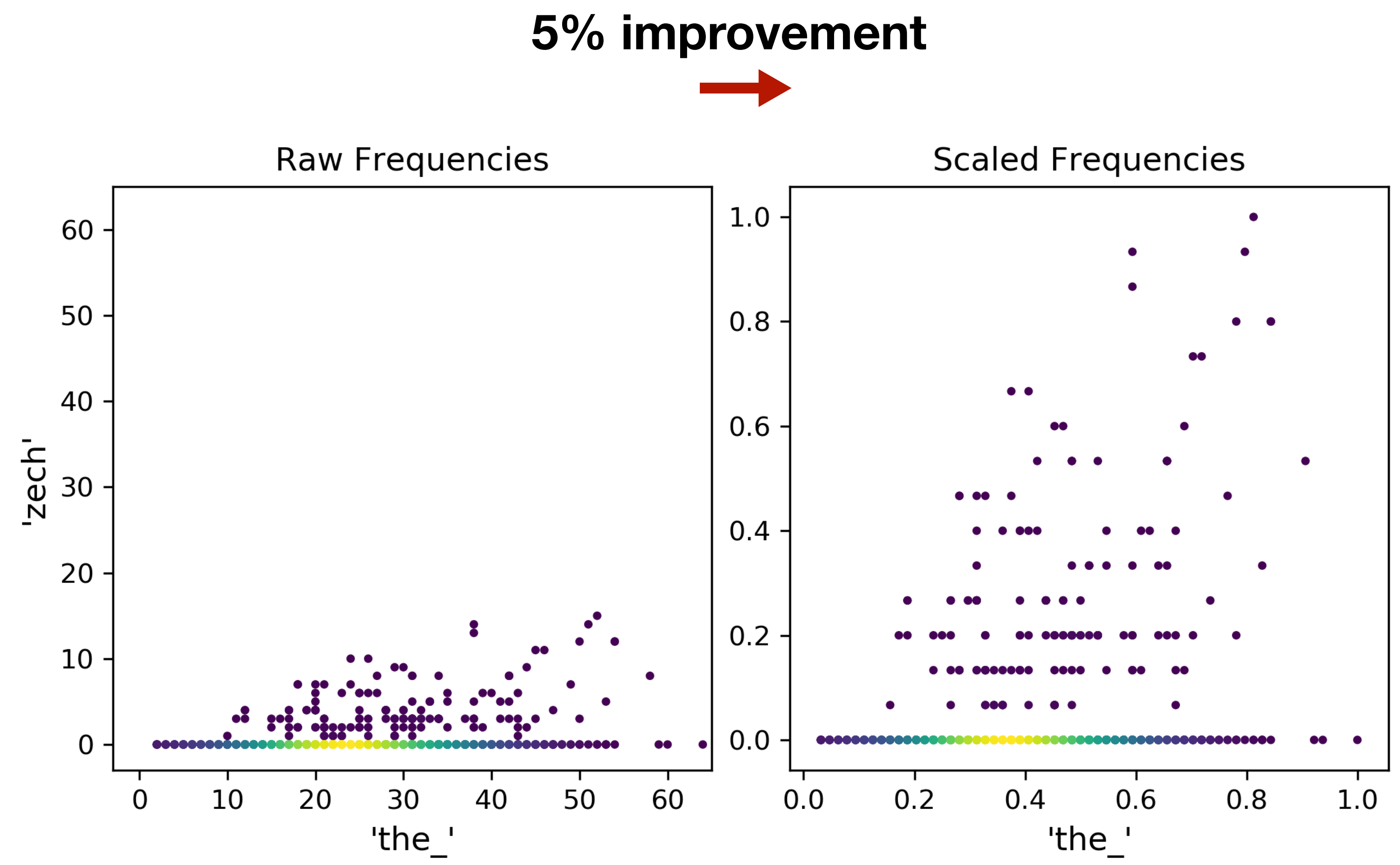


Fig 2: *N-gram frequencies for “the_” and “zech”.*

One author writes frequently about the Czech Republic. With scaling, the SVM may have less difficulty identifying this cluster.

Conclusions

- Optimal feature set
  - Mutual information is best performing feature selection method
  - 4-grams outperform 3, 5, and multi-length n-grams
  - Accuracy ~ log (number of features).
- Improvements
  - Scaling features improves accuracy
  - Max accuracy ~ 75%

References

- Houvardas, John and Efstathios Stamatatos. *N-Gram Feature Selection for Authorship Identification.* AIMSA (2006).
- Holmes, D.: *The Evolution of Stylometry in Humanities Scholarship.* Literary and Linguistic Computing, 13:3 (1998) 111-117.
- Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. *Authorship Attribution with Support Vector Machines.* Applied Intelligence 19, 1-2 (May 2003), 109-123. DOI:https://doi.org/10.1023/A:1023824908771