University of British Columbia Okanagan / Irving K. Barber Faculty of Science / Department of Computer Science, Mathematics, Physics and Statistics

Exploring Latent Topics in Data Science

Dr. Irene Vrbik, Daniel Krasnov

Motivation

Data Science is a relatively new discipline which is constantly evolving to keep up with the rapid development of contemporary technologies. While we expect common modalities to arise in Data Science curriculum, we anticipate a level of variability that would not necessarily be present in more established disciplines.

Objectives

We aim to discover overarching themes and topics in Data Science curricula to better understand commonalities and highlight differences between undergraduate programs in Data Science. With this information we can:

- Develop and shape our Program and Course level outcomes
- Establish “core” courses that should be required in UBCO’s Data Science program
- Identify gaps or deficiencies within our own undergraduate program to inform course revision and creation

Methods

Topic Modeling:

- Unsupervised learning method
- Clusters words from documents into *K* “topics”
- Captures latent structure of text

Latent Dirichlet Allocation[1] (LDA)

- A popular topic modeling method
- Three-level hierarchical Bayesian model
- Each document is treated as a mixture of *K* topics
- Each topic is treated as a mixture of words

We employed the `ldatuning` package[2] to tune the model and find optimal values of *K*.

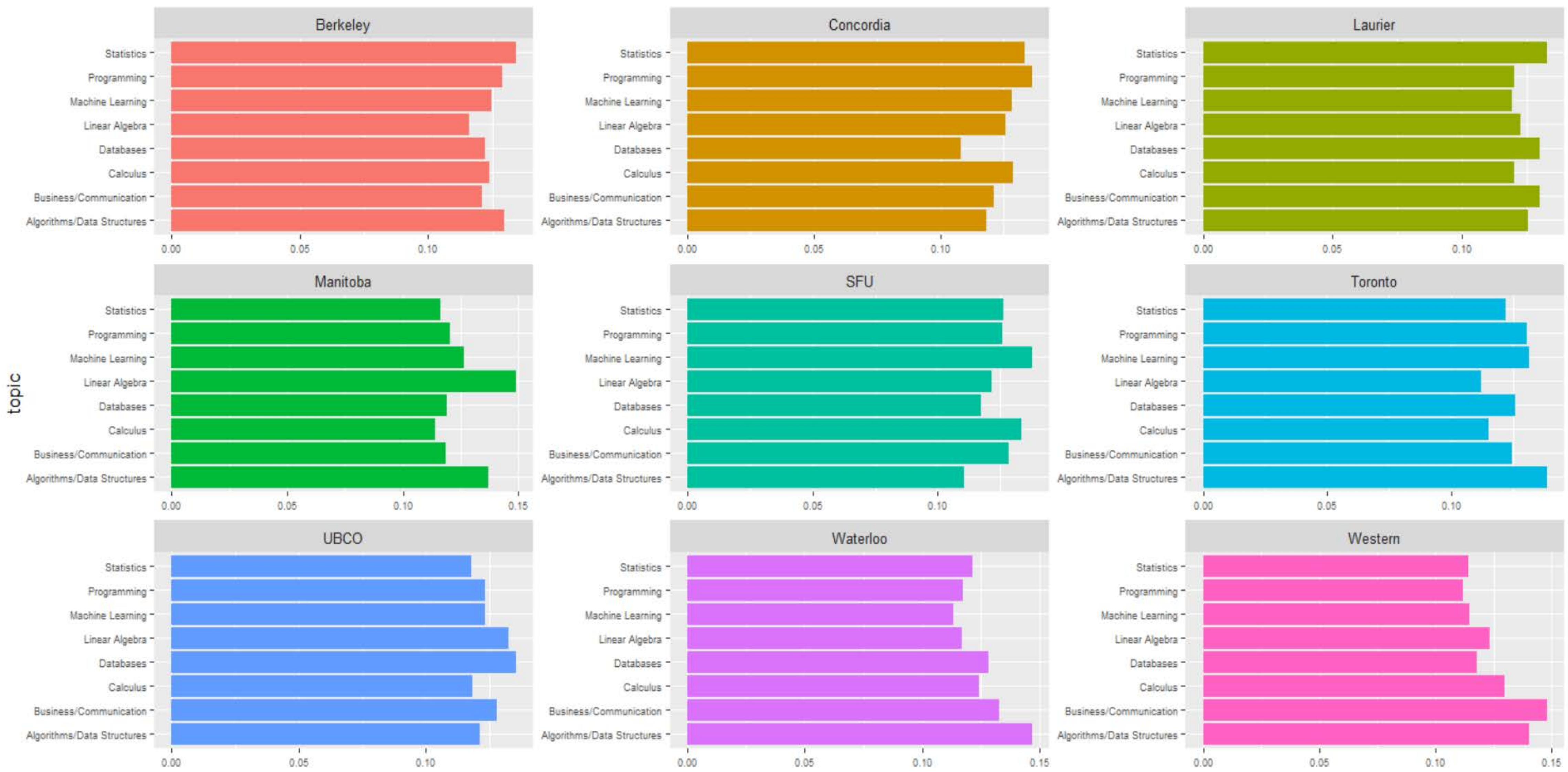
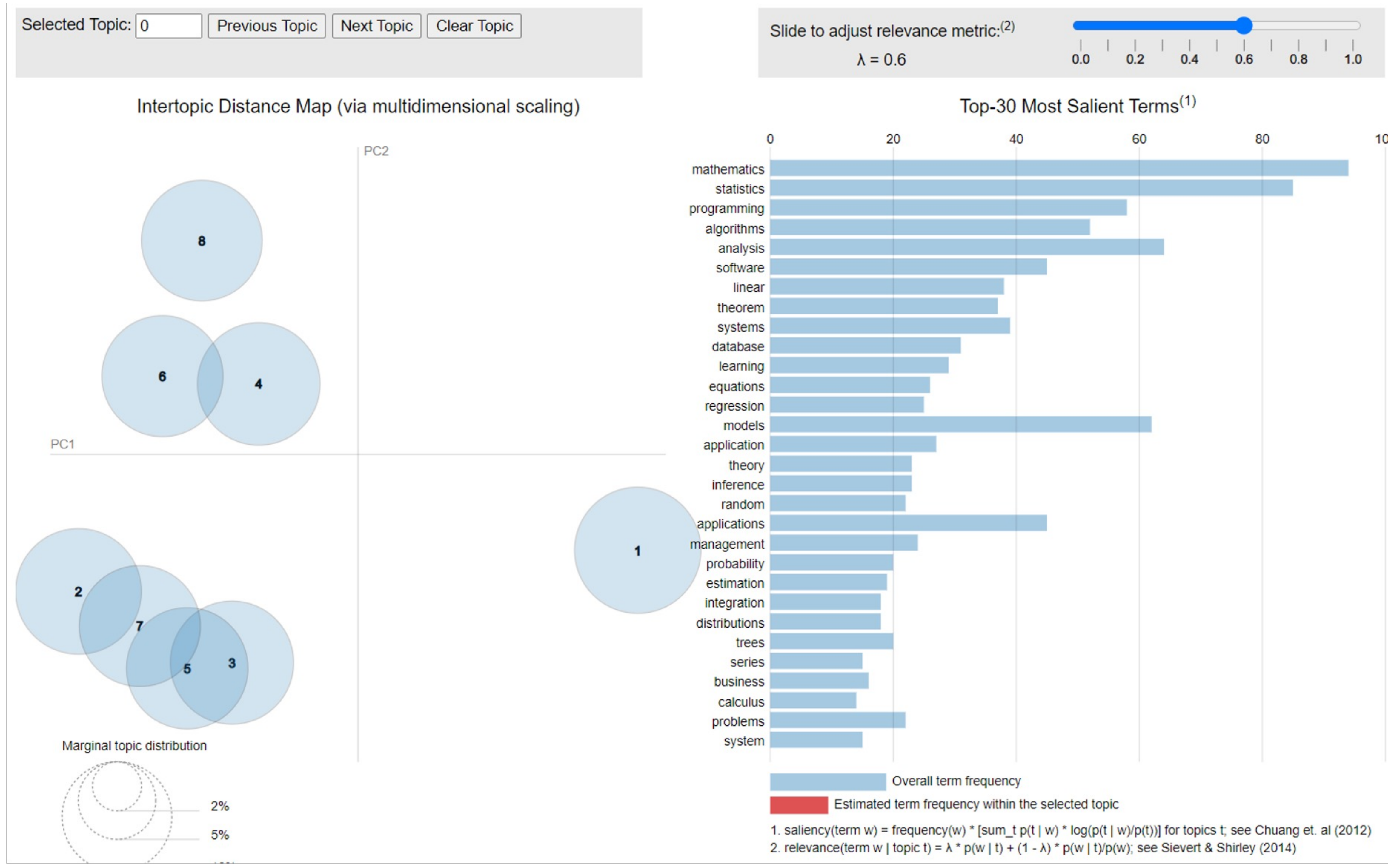


Figure 1: University-Topic Composition *K* = 8 *The relative proportion of each topic at each University*



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 2: Visualization of the fitted LDA model using LDAvis for *K* = 8

A visualization of the LDA 8-topic model fitted to our corpus of Data Science course descriptions. An interactive version of this visualization created using LDAvis[3] *can be found here*

Data

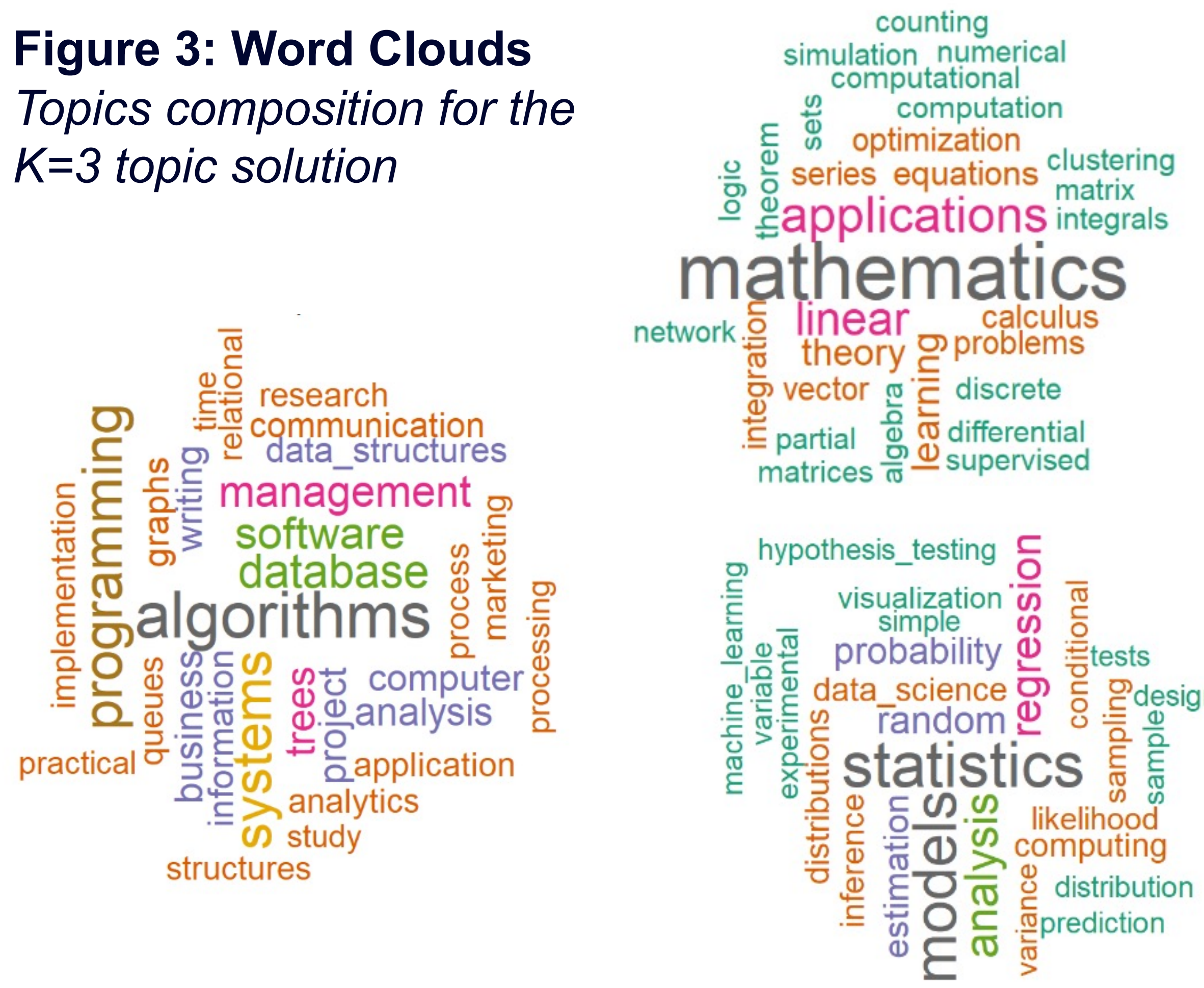
Our data consists of course descriptions scrapped from undergraduate Data Science programs across the United States and Canada. Common NLP text preprocessing techniques were employed before fitting the LDA model.

Results

- The *K*=3 topic solution uncovered three themes which appear to map to: Programming, Statistics, and Mathematics (see Figure 3)
- The *K*=8 topic solution produced clusters which we identified as: Statistics, Programming, Machine Learning, Linear Algebra, Databases, Calculus, Business Communication and Algorithms (Figure 1)

Figure 3: Word Clouds

Topics composition for the K=3 topic solution



Reference / Bibliography

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3 (Jan), 993-1022.
2. Nikita M (2020). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.2
3. Sievert, C., Shirley, K., & Davis, L. A method for visualizing and interpreting topics. *In Proceedings of Workshop on Interactive Language Learning, Visualization, and Interfaces, Association for Computational Linguistics* (pp. 63-70).

Acknowledgement

We gratefully acknowledge the financial support for this project provided by UBC Vancouver (or UBC Okanagan) students via the Aspire-2040 Learning Transformations Fund.