

Analyzing trends in cognitive science with a dynamic topic model

Zhi-Wei Li

Alexander Rich

Anselm Rothe

Introduction

Topic models have been used successfully to capture structure in large text corpora and make these corpora more human-understandable (Blei, Ng & Jordan 2003). However, traditional topic models do not capture the temporal ordering of documents, and the way topics change over time. Dynamic topic models address this issue by allowing topic mixing proportions and the distribution of words in topics to change as a function of time (Blei & Lafferty 2006). In this project, we applied dynamic topic modeling to 18 years of the Proceedings of the Cognitive Science Society to understand and visualize how the research topics of the society have changed over its history.

Related Work

Topic models have been used to understand the structure of academic publishing in the past. The archives of the journal *Science* were used as a test dataset in the original dynamic topic model paper (Blei & Lafferty 2006), as well as in other topic model extensions (e.g., Blei & Lafferty 2007). Cohen Priva & Austerweil (2015) applied topic modeling to the field of cognitive psychology, using the archives of the journal *Cognition*. However, they used a static topic model, and visualized the change of topics over time in an ad-hoc manner using the empirical frequency of words within each topic. Thus, this project represents the first attempt to directly model changes in the field of cognitive science over time.

Problem Definition and Algorithm

In standard topic modeling, the goal is to infer a fixed set of latent topics underlying a corpus of documents. In the underlying generative model, each of K topics is associated with a distribution β_k over a fixed vocabulary. For each document, topic proportions θ are drawn from a Dirichlet with hyperparameter α . Finally, for each word in the document, topic assignments for the word are chosen according to $Z \sim Mult(\theta)$, and the actual word is chosen according to $W \sim Mult(\beta_z)$.

In the standard topic model, the time at which a document was published would have no effect on word distributions of its underlying topics. In the dynamic topic model, we relax the assumption that $\beta_{1:K}$ are fixed over all time points. Instead, we replace β_k with $\beta_{t,k}$, denoting the word distribution of topic k at time t .

To model the drifting of β over time, we represent β in an unconstrained space and assume that the terms of beta drift over time according to Gaussian noise,

$$\beta_{t,k} | \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I)$$

Then, when drawing words for a given document, words are chosen according to $W \sim Mult(\pi(\beta_{t,z}))$, where π maps the unconstrained parameters back into a multinomial space,

$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}$$

Once the generative model has been defined, variational inference is used to infer β , as well as θ for each document.

We note that in the original topic modeling the prior distribution of topics α is inferred from the data along with β , and in the dynamic topic model α is additionally allowed to drift over time. However, we found that in the Blei lab’s public implementations of the standard topic model (LDA) and the dynamic topic model (DTM), α was fixed, presumably to decrease computational complexity. Thus our analyses assume a fixed α across topics and over time.

Experimental Evaluation

Data

We used python to scrape 7821 pdf files from the Cognitive Science Conference archives, representing submissions from 2000 to 2017. In general, each submission is a 6 page paper. We converted the pdfs to text using an automated pdftotxt utility. We tokenized the text based on whitespace, and removed lines in which few tokens were english words, because these lines tended to contain equations. We also removed words that were less than four characters long and were not in a standard english dictionary, as these tended to be caused by errors in the pdf parser. Finally, we removed tokens that occurred in fewer than 36 documents (i.e., 2 documents per year on average), as well as tokens that occurred in more than 50% of documents. Our final vocabulary contained 9710 words.

Methodology

We used the Blei lab’s C implementation of the dynamic topic model (<https://github.com/blei-lab/dtm>) to fit the DTM to our data set. We assumed 20 topics, and chose a fixed α of .1, and a topic drift rate of $\sigma^2 = .005$. The model used for our qualitative results was fit using all 7821 pdf files from all 18 years. To evaluate the DTM compared to a standard topic model, we fit a series of models in which we trained the model up to a given year, and then inferred the log-likelihood of the data for the given year. Using the Blei lab’s topic model C implementation (<https://github.com/blei-lab/lda-c>), we compared this to a standard topic model trained up to the given year, and a standard topic model trained only on the prior year. We conducted this procedure for the years 2002, 2005, 2008, 2011, 2014, and 2017.

As an aside, while attempting to fit the DTM up to a held out year we discovered an error in the C code that prevented this feature from functioning. We were able to fix the bug, and it has now been incorporated into the DTM repository (<https://github.com/blei-lab/dtm/commit/0774574142a8a38aed480cb78ddb8d6182eaa4a9>)

Quantitative Results

The results of the DTM and LDA models on a heldout year are displayed in Figure 1. Interestingly, we find that while both models trained on all prior data perform much better than a model trained on only the previous year’s data, the DTM consistently performs slightly worse than LDA. This contrasts with the results of Blei and Lafferty (2006), who found that the DTM outperformed LDA on the Science corpus. Despite its slightly worse performance, we continue to analyze the fit DTM model below, due to its interesting temporal properties. Possible reasons for its disappointing performance are presented in the discussion.



Figure 1: Performance of the DTM relative to LDA trained on all previous years and LDA trained on only the prior year. Higher values of log likelihood are better. We find that DTM performs slightly worse than LDA trained on all prior years, but much better than LDA trained on only the previous year.

Qualitative Results

Using the DTM’s inferred topic-year-word distributions (β) and topic-document distributions (θ), we created several visualizations to understand how the field of Cognitive Science has changed over the last two decades.

Figure 2 shows the overall proportion of each of the 20 topics over the last 18 years. Since the Blei lab DTM assumes fixed topic proportions (α), these proportions were estimated empirically by averaging the topic proportions in all documents in a given year, and then smoothing the curve using a LOESS regression. Some topics have remained fairly stable over the years. Others have become much more or less popular. The topic with top word “probability”, for example, which has common terms strongly associated with Bayesian statistics, has more than doubled in popularity to become the most popular topic. Decision making has become much more popular as well. The topic focused on neural networks, in contrast, has decreased in popularity from its earlier heyday, but is starting to show a resurgence.

Figures 3, 4, and 5 show the similarity structure of different topics, projected into two dimensions using multi-dimensional scaling (Figures 3 and 4) or t-SNE (Figure 5). The structure of the MDS solution seemed not to change much over 18 years, suggesting the relation of topics to each other was fairly stable over time. In the t-SNE visualization, documents tend to cluster closely based on their highest-proportion topic.

Figures 6, 7, 8, and 9 show the evolving word distributions of the top words of four topics over time, as determined by the inferred β parameters. Some topics, such as the topic on Bayesian and probabilistic modeling (Figure 7), have kept stable word distributions over time, even while the topic itself has become much more popular (see Figure 2). Others have seen more variability. For example, the frequency of the word game in the decision making topic peaked around 2011, likely due to a surge in popularity of game theory in cognitive science (see Figure 8). Its frequency has since sharply decreased.

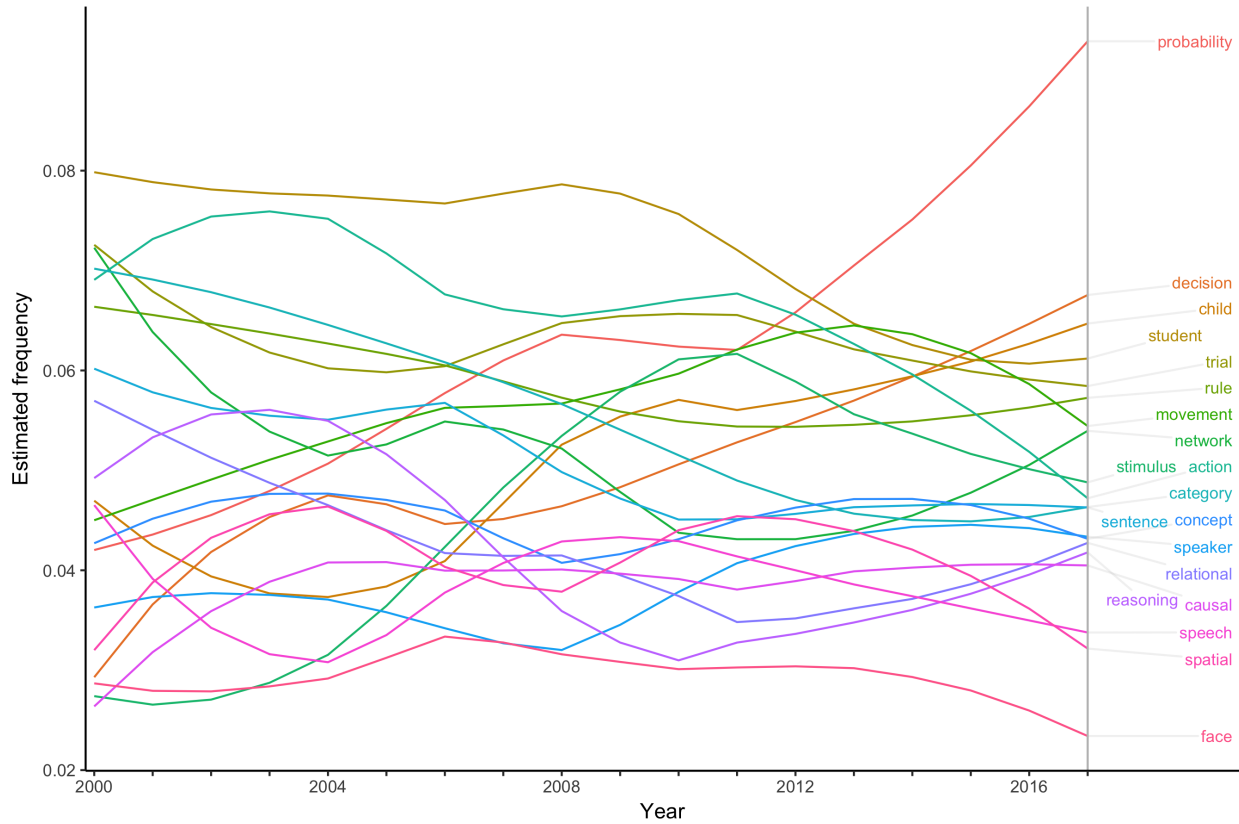


Figure 2: Popularity of each topic over the last 18 years. Label indicates the top word for the topic.

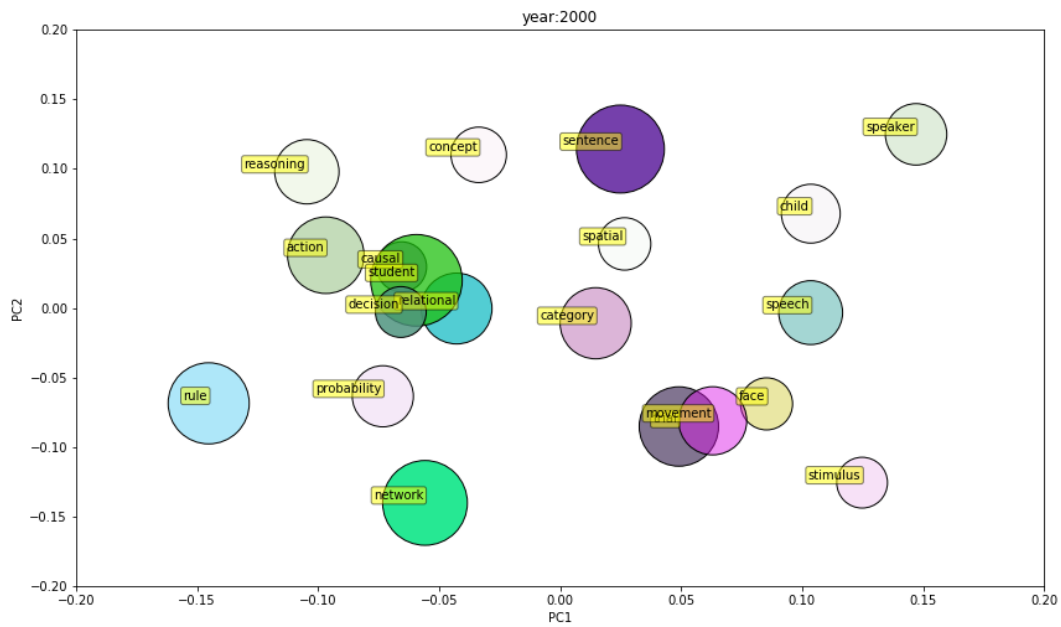


Figure 3: Similarity of topics in 2000, visualized in two dimensions using Multi-Dimensional Scaling.

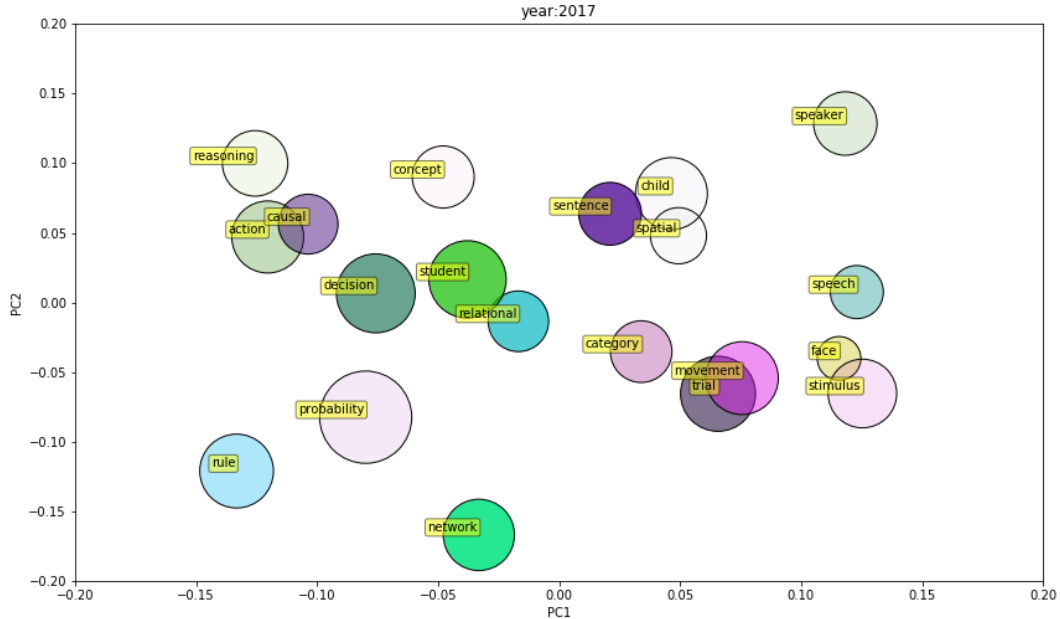


Figure 4: Similarity of topics in 2017, visualized in two dimensions using Multi-Dimensional Scaling.

Discussion

Our dynamic topic model of cognitive science appeared to reflect a variety of trends within the field. It captured the main subfields (decision-making, categorization and reasoning; probabilistic and connectionist modeling; developmental and education; perception, etc), as well as changes in the field, such as the growing popularity of Bayesian models of cognition. However, there is major room for improvement in our analysis.

The most pressing issue is the fact that standard LDA achieved slightly higher log-likelihood on held-out data than the DTM. This might have occurred because topics within the field were too stable over the 18 years for which we collected data for the assumption of drifting topics to be necessary. In Blei and Lafferty’s (2006) analysis of Science, they show that the topics in physics or neuroscience changed drastically over a century, which is hardly surprising. In contrast, a paper on decision making now might contain many of the same words as one 18 years ago. A useful follow-up might be to tighten the DTM’s prior on the drift rate of word frequencies over the years, and test if this improves model performance.

Another more ambitious future step would be to infer the prior α on topic mixing proportions by topic and year. This would allow us to infer the shifting popularities of topics within the model, rather than estimating them after the fact by averaging the topic proportions within each year (see Figure 2). The changing popularity of topics might be more pronounced in cognitive science than the changing vocabulary within topics, so this might give the DTM an advantage over standard LDA.

Conclusions

We presented the first dynamic topic model of the field of cognitive science. Our model was fit to over 7000 papers spanning the last 18 years of the Proceedings of the Annual Conference of the Cognitive Science Society, the field’s premier conference. While the DTM performed slightly worse than LDA as measured by log likelihood, it captured a variety of interesting historical trends in the conference. We look forward to

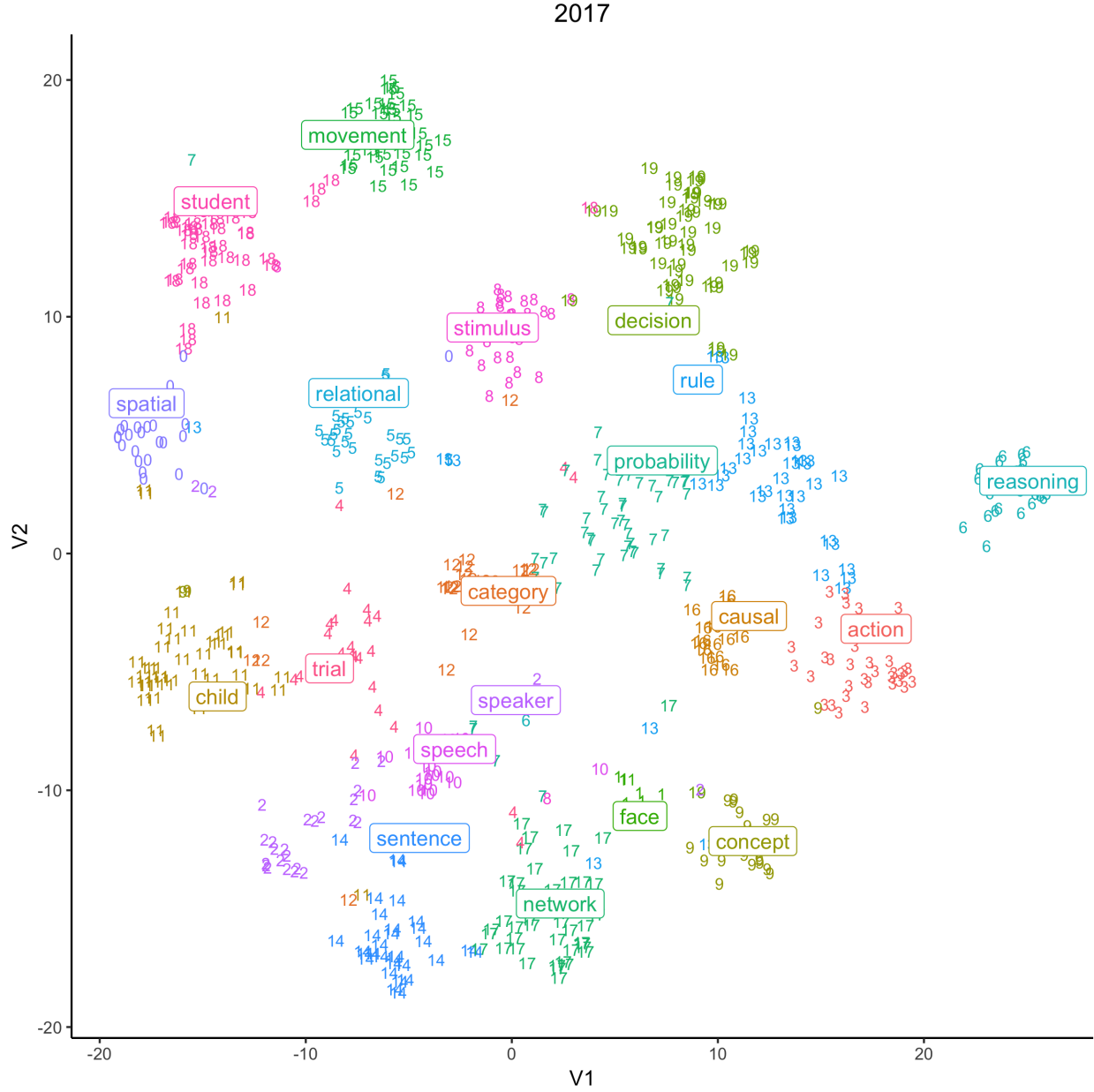


Figure 5: Similarity of documents in 2017 based on their topic distribution, visualized in two dimensions using t-SNE. The color and number of the document indicate it's highest-proportion topic.

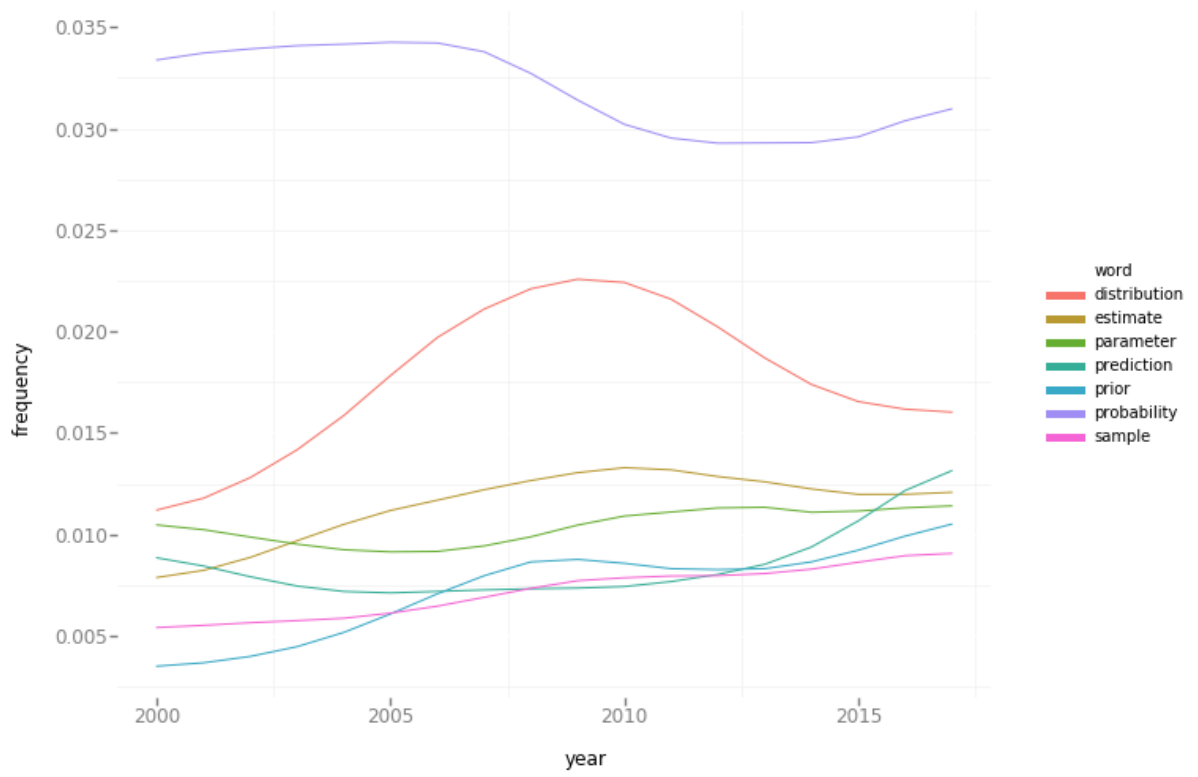


Figure 6: Trends in top words of topic centred on probabilistic and Bayesian modeling.

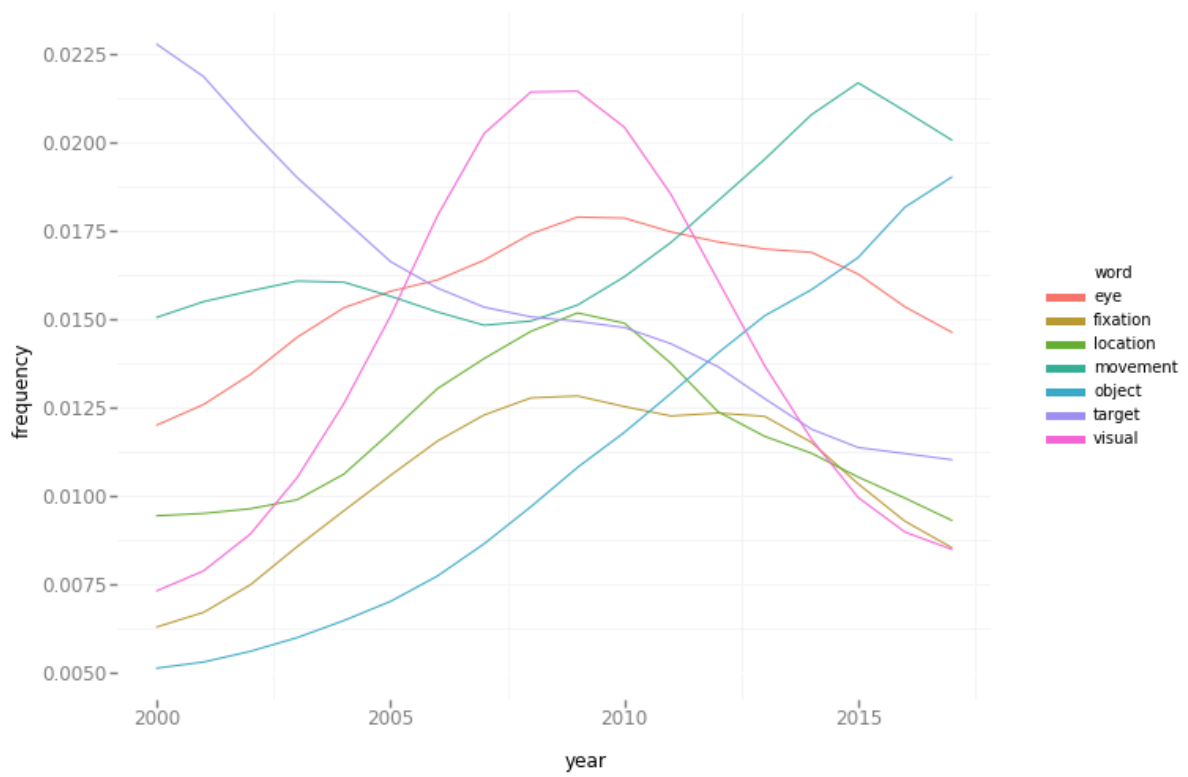


Figure 7: Trends in top words of topic centred on visual perception.

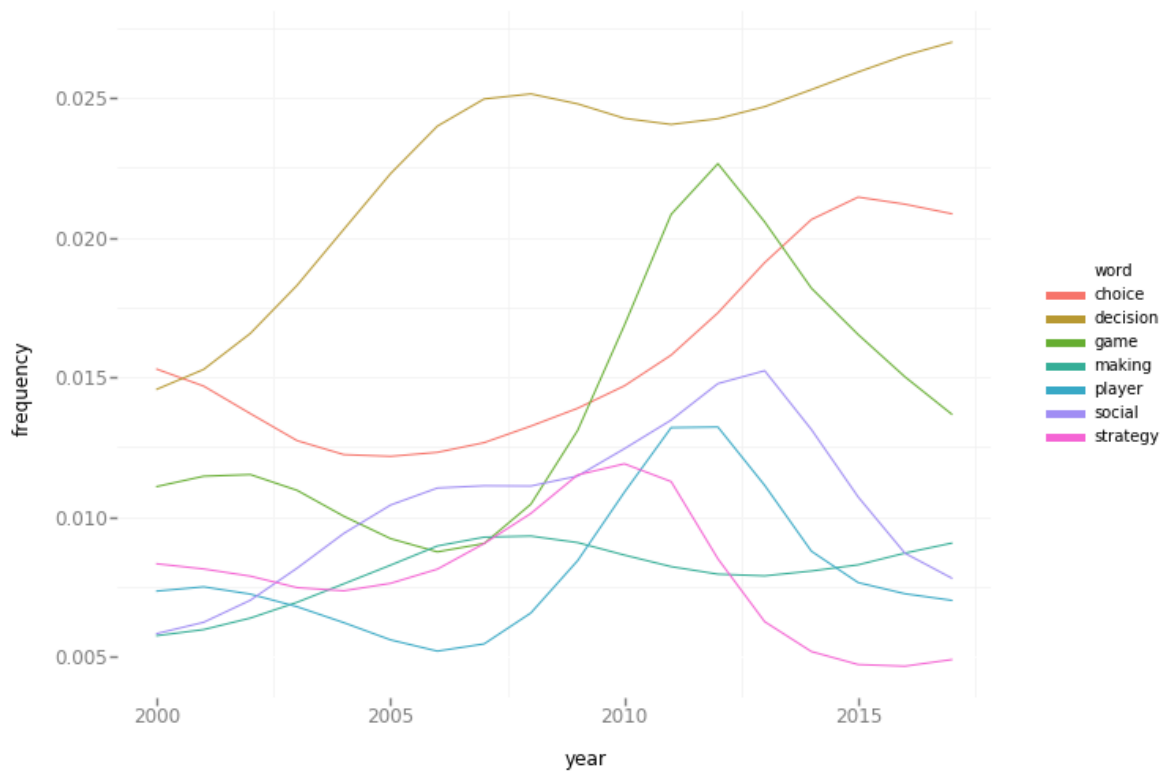


Figure 8: Trends in top words of topic centred on decision making.

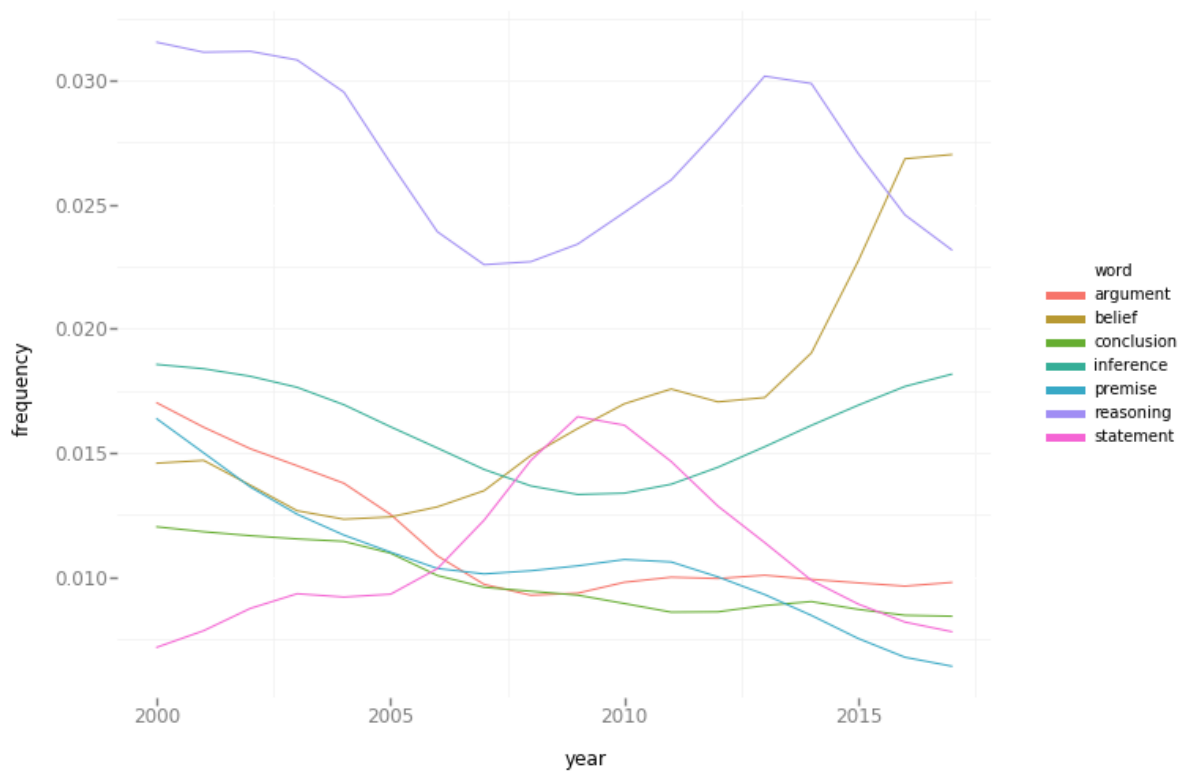


Figure 9: Trends in top words of topic centred on reasoning.

extending our analysis to further understand the past and present of our field and perhaps to predict its future.

Bibliography

Blei, D., & Lafferty, J.D. (2006). Dynamic topic models. Proceedings of the 23rd international Conference on Machine Learning. ICML '06, 113-120.

Blei, D., & Lafferty, J.D. (2007). A correlated Topic Model of Science. The Annals of Applied Statistics. Vol 1., No. 1, 17-35.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research. 993-1022.

Cohen Priva, U., & Austerweil, J.L. (2015). Analyzing the history of Cognition using Topic Models.